Catalogue Retail Forecasting – An Empirical Evaluation of Linear Regression and K-Nearest Neighbours sensitivity to preprocessing

Rodriguez Calderon, Carlos Eduardo

This dissertation is submitted for the degree of Master of Research

October 2018

Management School

Declaration

This thesis has not been submitted in support of an application for another degree at this or any other university. It is the result of my own work and includes nothing that is the outcome of work done in collaboration except where specifically indicated. Many of the ideas in this thesis were the product of discussion with my supervisor Dr Sven Crone.

Acknowledgements

This dissertation was supported by the Peruvian National Scholarship and Academic Loan Program (PRONABEC). I thank Dr Sven Crone who provided insight and expertise that greatly assisted this dissertation work

Contents

List of Tables

List of Figures

List of Abbreviations and Acronyms

k-NN: k-Nearest Neighbour (machine learning technique)

MAE: Mean Absolute Error

MAPE: Mean Absolute Percentage Error

RMSE: Root-mean-square error.

sMAPE: Symmetric Mean Absolute Percentage Error

WAPE: Weighted Absolute Percent Error

List of Appendices

1 Introduction

Selling retail products through catalogues is an established retail channel, dating back to the 18th Century and still growing in volume and revenue despite online retail (Hall, 2007). Catalogues are the primary instrument to present products to customers. Product pictures are displayed with different offers across sales periods, with different marketing strategies like including models or celebrities with the product, or applying different offers for the same product across time, like discounts or offers type BOGOF (buy one, get one free) or set of products that cannot be sold separately. However, despite its prominence and growth, catalogue retail has not received a lot of attention in academic research in general, and next to none in forecasting in particular.

Early papers by Chambers and Eglese (1986), which addressed particular catalogue forecasting challenges in new product forecasting or within-season forecasting, were not followed up until recently Boada et al (2011, 2017), applying standard linear regression models with log-transforms for catalogue forecasting. However, these studies fail to compare forecast accuracy across available algorithms, limiting a discussion to log-regression, nor of the impact of data preprocessing such as log-transforms, outlier correction, or feature selection on the result. More importantly, they

fail provide a reliable empirical evaluation, using but in sample metric of goodness-of-fit instead of out-of-sample errors, squared error metrics of RMSE instead of more robust sMAPE, and no comparison against benchmark methods. While this substantial omission may seem surprising, recent research by Ma and Fildes (2017) confirm a general gap in literature on retail forecasting, which is further supported by this author's personal experience having worked as a forecasting analyst in a catalogue retail company. Consequently, this dissertation seeks to address the gap in literature and provide an empirical evaluation of the effect on data preprocessing on two major algorithms: linear regression and k-nearest neighbours, which have not yet been applied to catalogue forecasting. As a dataset, we use an empirical dataset from a direct selling company, a particular form of retail using catalogues and consultants working as sales representatives to attract consumer demand, with data spanning three catalogue periods of a prestigious cosmetic company. Therefore, our contribution is twofold, by providing a first valid empirical design and evaluation of catalogue forecasting, and by conducting a first comparing of algorithms of statistics to machine learning to testing their sensitivity to different of data preprocessing.

This paper is divided in seven chapters, first introducing retail catalogue sales, its role and relevance, its relationship with direct selling and how the catalogue becomes an important tool with unique properties to drive sales and also for this situation to predict sales. Next we discuss the lack of forecasting literature in this area. The third chapter develops the experimental evaluation, describing how the dataset will be transformed and splitted to train a model, also how the error measures are going to be applied over forecasted results compared with test data. The fourth chapter will include the experimental results, the accuracy of each tested method, focusing on different results

for Regression and k-NN. The fifth chapter will present the conclusions for this dissertation, followed by the references and the appendices.

2 Forecasting Demand from Retail Catalogues

2.1 Properties of Catalogue retail

2.1.1 Retail Catalogues in the Retail Industry

Retail remains to be a major sector in developed and developing economies, ranging from 14% of total GDP in the USA and 16.5% in the UK to 20%-40% in tourist-oriented island economies (United Nations Statistics Division, 2013). The retail sector continues to grow, with India showing growth from 8.4% to 18.7% and China from 7.3% to 11.5%, with China becoming the largest retail market in the world in 2016 (Millward, 2016). In 2011, the Central European grocery market was worth nearly €107bn, and growing 2.8% from the previous year, and employing a significant amount of the labour force (PMR, 2012). Consequently, retail remains an area of research, with numerous academic journals dedicating research to the topic, including forecasting.

Retailing, the process of selling consumer goods or services to customers, is traditionally achieved through multiple channels of distribution. With historic evidence of trade dating back 10,000s of years, retail and its channels have undergone major transitions (Findlay & Sparks, 2002). Recent developments see the growth in revenues affecting the main retail channels differently, most notably local retail shopping, non-local retail shopping, Internet shopping, television shopping, and catalogue shopping, fuelled by changing online consumer behaviour. Amongst these main channels, catalogue retail is one of the traditional yet still growing channels. Traditionally using

mail order, catalogue retail emerged during the 19th century with improvements in transport and postal services. In 1861, Welsh draper Pryce Pryce-Jones sent catalogues to clients who could place orders for flannel clothing which was then despatched by post, extending his client base across Europe (Goldstein, 2013). In 1872 US retailer Montgomery Ward also devised a catalogue sales and mail-order system listing 163 items for sale (devising the catch-phrase "satisfaction guaranteed or your money back") (Hevrde, 2017). In the 1890s, US retailers Sears and Roebuck also started using mail order with great success.

Whilst ordering products from catalogues may sound like an antiquated approach from the 1950s, many remain successful. In today's UK economy, the largest general goods retailer is Argos, a catalogue merchant which (similar to Littlewoods, Next, and others) also owns local retail shops. Some large UK catalogue retailers including Shop Direct, which had no shops and sold only by postal orders, developing their businesses around offering revolving credit which has lost some of its market niche. Today Shop Direct's brands remain strong with Very.co.uk, Littlewoods.com, Very Exclusive and Littlewoods Ireland which consolidate mayor UK retailers Additions Direct, Abound, Choice, Great Universal, Index, Isme, K&Co, Kays, Marshall Ward and Woolworths (The Telegraph, 2014). Furthermore, numerous new entrepreneurs have chosen a catalogue line of retailing, most notably clothing retailers Boden, The White Company, Artigiano or Howies, or furniture retailers OKA and Lombok, selling everything from wine to childrens' clothing, bed linen and ethnic furniture in an upmarket niche of high-end wares using a combination of printed catalogues and the internet, and they are all seeing sales growth exceeding that of traditional high street retailers (Hall, 2007).

In the US, leading strategy consultants Forrester anticipate a stabilization of online sales in the next 5 years, expecting to plateau at around 10% of total US retail sales, whilst "the mature US catalogue industry will see sales growth of 8% year over year, and the DMA projects that it will remain at that rate over the coming five years, with economic factors causing some near-term downward adjustments". (Forrester, 2009, p. 5)

Consequently, catalogue retail remains a relevant economic sector in general, and a pertinent channel in retailing, with its unique differentiators warranting analytical solutions just as any other relevant sector.

2.1.2 Direct Selling using Catalogues

Retailers who employ direct selling, using sales representatives making direct customer contact and acquiring (non-anonymous) orders traditionally employ catalogues catalogue as a product's showroom for their particular sales channel, adding the differentiators of direct sellers as the main actors to perform sales in addition to the catalogue. The World Federation of Direct Selling Associations (WFDSA, 2016) describes direct selling as "a retail channel used by top global brands and smaller, entrepreneurial companies to market products and services to consumers."

As a channel of sales, two main characteristics of this process that can illustrate the previous definition (Ongallo, 2007) are the way the sale is performed, by a face-to-face relationship that can take place at the customer's workplace or home in many of the cases. In addition, the way transactions are developed, by demonstrating a product. Thus, the seller adopts some titles such as "beauty consultant," "dealers," "counsellors." Direct selling has different variants, by covering a zone and trying sales door to door, by making appointments in customer's address or workplace, by demonstrating the product in special tasting meetings or by trying to locate products in trips.

For the particular case of the cosmetic industry, direct selling is the most important sales channel, despite the Internet. The largest direct-selling company in the world, with annual sales of over US$6 billion, depends on its own independent sales team. Sales are not made through department stores but its beauty consultants through catalogues. Nevertheless, the door-to-door sales model would be at risk due that three-quarters of American women now work outside the home (Kumar, et al., 2006).

Another concept, related to direct selling, is the concept of direct marketing (Michael Baker, 2018), where the experience for the customer is turned into direct response advertising by maintaining a customer database. With this concept, it is possible to differentiate essential customers from regular customers and to prepare special offers by combining direct marketing and Pareto's principle. This differentiation between customers is provided by the brand and how direct selling is delivered to the customer. It is usual to use a strategy face-to-face using a catalogue, every sale is delivered to each customer by separate, and usually the seller is the one who delivers the product to the customer. This strategy is usual for cosmetics companies like Avon, Yanbal, and Oriflame (Ongallo, 2007).

Direct sales claim to be focused on customer more than sales by itself. It is said that: experiences are an essential factor during the buying process (Puccinelli, et al., 2009), every demonstration or meeting with final customers and their consultants must be a memorable experience, for cosmetics this experience can be achieved by using catalogues instead of products.

Other variants of direct selling to target customers traditionally employ catalogues (Ongallo, 2007), include Party Plans, using meetings taking place inside family houses where the host present the product to all the guests, Door-to-door, where the seller visits

different workplaces, and multilevel networks, where an agent network is developed to recruit memberships, reporting to the main seller who is in charge of the group. It should be noted though that other forms of direct selling do not utilise catalogues, such as call centre sales, where the customer calls or is called for sales to be made, or direct factory outlets that combine the previously explained concept of direct marketing with showrooms for a specific group of customers with special offers, but without catalogues.

2.1.3 Unique Properties of Catalogues

In catalogue retail, the catalogue itself is a unique instrument for indirect and direct selling, and its merits and benefits warrant understanding.

A catalogue is a tool to present products to potential customers, a booklet with hundreds of products with different prices, promotions, and discounts for a fixed period that is renewed after it ends. A catalogue is an essential tool, a portable showroom for sellers and for demand planners.

An ordinary catalogue handles a selection of features, an introduction for a particular brand, the internal catalogue code, the regular price, and reduced price; this offer discount can vary from one period to another. An example of how all this information is presented is available in Figure 1, where it is possible to visualise different cosmetics, prices, offers and some phrases to catch the attention of the product, also a model to add some personality to the brand or product.

Figure 1: a sample of an Oriflame catalogue

Sales periods would vary depending on the company, some companies work with a period of sales of 20 days, so it allows these to have 18 periods of sales per year and 18 catalogues for each period of sales (Belcorp, 2015) and (Oriflame Sweden, 2018) during the year which is required to be printed six months in advance or more.

Part of the business intelligence and marketing strategy can be found inside the catalogue, marketing variables related to price, discount, and the way the product is presented across the pages or if it will be absent for certain periods. As a catalogue is printed in advance, a company can use future catalogue designs as an input to predict demand. It is expected that the main constituents in a catalogue will drive demand, and as such require careful consideration in marketing and thus estimating future demand.

To discuss the importance of catalogue-based sales (Blattberg & Deighton, 1996), a factor to be considered is the ceiling, i.e. the maximum number of clients the company can obtain in a certain period. Individual companies like insurance companies find a catalogue retailer has a better option to acquire new customers that others only based on direct mail. In the same study, it is declared that individual companies, like IBM,

have turned out the old famous sales force they used to have, to apply "less expensive retention tools."

A catalogue, which is a sort of magazine focused on specific brands, have some features showing "must-have" items in individual sections, like front or back covers. In fashion (Iqani, 2012) for instance, issues are highlighted in covers, featured by a celebrity. The purpose of using covers with celebrities or models is to allow an ordinary woman to copy a celebrity's style and 'get the look' by wearing the item offered by the star (or model).

If efficiency and convenience are benefits for the customer, retailers like locating in remote, low-cost locations can obtain some benefits; saving on store operating expenses and workforce associated. Nevertheless, catalogue retailing is criticised for being inflexible, with limited service and effective only when it is narrowed to a specific set of products (Pride & Ferrell, 2012).

Another feature that is available in catalogues is the use of smell. Catalogues can carry samples of perfume that can be released by lifting or peeling off the odour-impregnated paper. This feature enables the reader to watch and sniff simultaneously (Cook, 2001).

2.2 Forecasting Literature on Catalogue Retailing

2.2.1 General Research in Catalogue Retailing

Despite its prevalent economic relevance, and distinct properties in marketing and supply chain logistics which differentiate it from traditional on-site, TV or online retailing, only few academic studies have considered the channel of catalogue retailing (and its many variants, e.g. catalogue retail, catalogue merchants, mail-order retail etc).

Research in marketing focusses largely on consumer perception, from the role of consumer attitudes towards technology as predictors of online versus offline catalogue usage (Noble & Oconnor, 1986), the impact of computer interfaces on consumer involvement with print versus online catalogues (Griffith, et al., 2001), the overall effect of consumer shopping experiences and perceived values (Mathwicka, et al., 2002), or the the influence of catalogue versus store shopping on customer satisfaction and risk (Festervand, et al., 1986). Despite using analytical frameworks, little insight into marketing drivers that might influence forecasting are identified from this line of research.

Few analytical papers focus on the supply chain challenges in catalogue retail, such as the organisation of sustainable supply chains in fashion catalogues (de Brito, et al., 2008), and forecasting returns for reverse logistics (Masmoudi, 2011). A second group of research focusses on pricing policies in the context of management science in general and revenue management in particular, see e.g. (Kashyap, 1995) analysing price changes in retail catalogues to determine stickiness. Azuma et al. (2016) propose rules for a cumulative deposit rate for payment in Japanese mail order catalogues, but despite its suggestion in the title without consideration of the forecasting methods.

2.2.2 Forecasting Research in Catalogue Retailing

Despite supply chain decisions as well as revenue management and the resulting fixed or dynamic pricing heavily dependent on forecasts of future demands, forecasting should play a prominent role in catalogue retail research. However, literature on Forecasting methods or methodologies using retail catalogues or direct selling is scarce.

A structured literature review at the ISI web of knowledge (searching for "(catalog* or catalogue* or mail-order) AND (forecast* or predict*)") reveals only

Early work by Michael (1971) applied computer simulation for forecasting catalog sales. In the 1980s, Chambers and Eglese assess the ability to use preview exercises to forecast demand for new lines in mail-order catalogues (Chambers & Eglese, 1986) and to forecast demand during the selling season (1988)However, the methods are not directly applicable to standard products reoccurring in catalogues prior to printing the catalogue.

More recently, two research studies by Boada et al. have been published to assess forecasting in cosmetics catalogues which is in line with our later experiment: (Boada & Mallorca, 2011) and (Boada, 2017), and two conference papers (Millan & Boada, 2010) and (Boada, 2017b), which we review in detail.

In a first study (Millan & Boada, 2010), the authors focus on the importance of an accurate forecast, like inventory excess, or non-attended demand for a large worldwide cosmetic company called Avon. The problem to predict demand for a cosmetic company is presented as a multidisciplinary problem, directly related to the supply chain, it is described that demand can be explained with mathematical models (80%) and expert criteria (20%). In their study applying linear regression variants, they infer that demand is influenced by 31 different variables and 52 models are used for their products combining four years of data. Model details are not provided, but the determination coefficient is calculated between 72.97% and 85.54%, in addition to MAPE and WAPE being used as error measurement for forecast accuracy. The authors mention that error prediction improves above 10% and no more than 85%.

In a second study (Boada, 2017) applies another linear regression model in conjunction with a Bayesian dynamic linear model. Marketing variables are described for the model as a number of sellers (representatives), advertising in the catalogue (yes/no), discount, price, category, type, and size. A feature included in this study is that judgemental forecasting is required for certain products by detecting situations that can lead to underestimation or overestimation. For instance, situations that include a free product offer (by buying another cheaper product) or combined offers including two or more products that cannot be sold separately. The Bayesian model includes external factors that can influence demand, like new possible competitors, length of the period of sales, product redesign, and the Country risk. In this case, demand forecast is adjusted by using additional models to correct previous forecast by focusing on categories instead of individual products.

A similar approach, working with linear regression models, can be found for a different company, also based in catalogues and direct selling in a previous dissertation work, using only internal data and catalogue information also with logarithmic demand transformation and combining data for different products into a single model by using dummy variables for differentiation between them (Rodriguez-Calderon, 2017).

It is notable that all authors apply a Logarithmic transformation of demand in these studies, going back to early studies related to mail-order catalogue sales for fashion products and periods that last half a year (Green & Harrison, 1973). However, beyond details on traditional logarithmic transformations, other steps in the data preprocessing such as data sampling, variable scaling or encoding of ordinal or nominal variables, missing value imputation, feature selection and transformation, are largely neglected. What is more noticeable, that none of the papers considers more contemporary

algorithms from machine learning and artificial intelligence, e.g. k-nearest neighbours, decision trees, neural network, or support vector regressions, or ensembles thereof such as random forests, which are widely published in other areas of forecasting research. And finally, the design of the empirical evaluation of the aforementioned studies is found lacking, not comparing accuracy results against adequate forecasting benchmark algorithms, or indeed applying adequate out-of-sample error metrics in the first place instead on relying only on in-sample, goodness of fit statistics such as R-squared, despite the apparent challenges in overfitting limited dataset.

To conclude, the literature on catalogue retail forecasting is next to non-existent, which is surprising considering the relative importance and unique challenges of catalogue retailing. However, recent research by Fildes (2017) has also indicated that research in forecasting for retail is underdeveloped so that we may presume no bias or gaps in our literature review. Nonetheless, the lack of papers results in a significant gap in research on forecasting for catalogue retail, which this study seeks to close.

3 Experimental Design

3.1 Empirical Dataset

In order to assess different approaches to forecasting we will conduct an empirical evaluation using a real world dataset. The provided dataset belongs to an important cosmetic company with presence in Europe, North, and South America. It included 1765 records and 27 variables. Each record containing information regarding every product offered for the Spanish market in the first three periods of 2014. The information describes each offered product, the SKU code, its classification (category, brand, segment, and description) and the marketing strategy that should be visible in the catalogue. For an entire description of the dataset, please look at Table 1

Table 1: dataset composition

Field	Description
Product Code	SKU code (1160 available products)
Long Description	Description of the product
Market	Country: Spain
Campaign	From 2014.01 to 2014.03
Category	Seven different categories: Accessories, Colour Cosmetics, Fragrances, Other category, Skin care, Toiletries and Wellness
Product Status	Internal code with five different values: "D","L","N","X", "Z"
Brand	101 available brands for three periods.
Segment	90 different segments available.
Catalogue Driver	Yes/No
Model Wearing	Constant: No
New Splash	New, Second New or Not New Splash
Scratch and Sniff	Yes/No
Focus	Constant: Not Visual Focus
Catalogue Section	Five available positions inside the catalogue: back cover, category sections, ending section, middle spread or platform

Field	Description
Offer Percentage	% discount over normal product price: 0, 30, 35, 40, 45, 50, 60, 70, 75, 80, 100
Offer	No offer or one of six different offers that can be applied: combined offer, discontinued sales out, gift with purchase, purchase with purchase, set, straight discount
Source of sales	Constant: standard catalogue
Current forecast	Previous forecast
In Catalogue Phantoms	1 or 2
Requested quantity	Actual demand for the product in that period
Active consultants	A "consultant" is the person who executes the sale to a final customer. Active consultants are part of internal jargon, used by the company to identify which of them are continuously selling products of the company for two, three, or more periods without gaps. This definition would vary from one company to another but the purpose is to know how many sellers are available for each period.

To understand how data is classified, it is necessary to identify how the records are organised according to the hierarchy established by the company. By looking at the records of the data, it is possible to define three different aggrupations for the products: brand, category, and segment. The order how this classification is applied consider in a first place the brand and then the Category, Segment and finally SKU. To illustrate this classification, it is possible to show products from different categories and segments that belong to a same single brand, see Table 2.

Table 2: sample of classification for Christmas brand

Brand	Category	Segment	Product code	Long description
Christmas	Fragrances	Female Fragrance	26840	Fairy City Lights Eau de Toilette
Christmas	Skin Care	Hand Care	24709	Brazil Nuts Hand Cream
Christmas	Skin Care	Hand Care	30012	Christmas Wish Hand Cream
Christmas	Skin Care	Hand Care	30017	Fairy City Lights Hand Cream
Christmas	Toiletries	Bar Soap	24710	Brazil Nuts Soap Bar
Christmas	Toiletries	Bar Soap	30011	Christmas Wish Soap Bar
Christmas	Toiletries	Bath Additives	30010	Berry Christmas Bubble Bath

Brand	Category	Segment	Product code	Long description
Christmas	Toiletries	Deodorant	30016	Fairy City Lights anti-perspirant 24h deodorant
Christmas	Toiletries	Shower Products	30015	Fairy City Lights Shower Cream

3.2 Data exploration

In this subsection, every variable is analysed by separate and then some connections and dependences between variables are revealed.

Demand (Requested Quantity)

Summary statistics for this variable reports zero as its minimum value, median 121, and the mean 321.4; the interquartile range is 310, and the maximum value is 11912. Distribution of values is visible in Figure 2. In addition, there are 41 products (see Annex 01) with zero demand (nine for 2014-C01, 14 for 2014-C02 and 18 for 2014-C03).

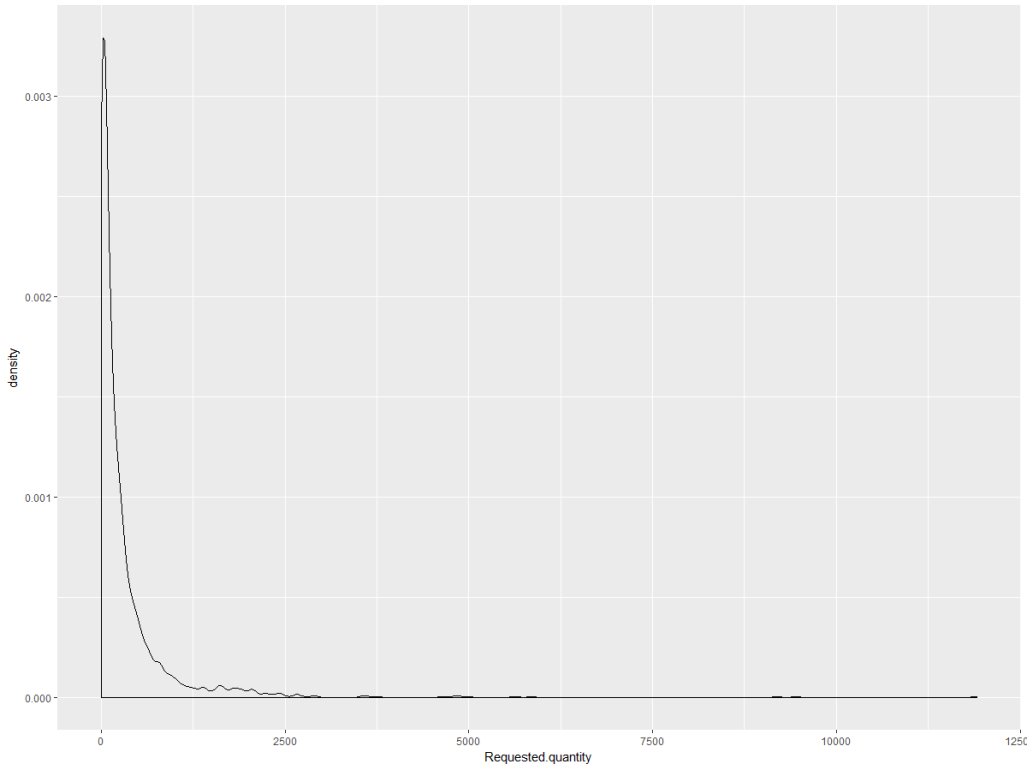


Figure 2: Density plot for requested quantity

As small values are most frequent and larger values above 2000 units almost rare, it is also possible to adjust the density of this variable to an exponential distribution by using mean = 121 (see Figure 3).

As larger values (greater than 2000 units) are less frequent, there are two different choices; one option is to consider these values as outliers (it is observed that 40 observations have a demand greater than 2000 and 19 greater than 2500) or to use variable transformation to include the whole dataset without filtering large values.

If the logarithmic transformation is taken into account, another problem would be related to cases that are not possible to be treated. It is possible to detect up to 41 cases where demand is not significant. These records would be included by adding a small amount to allow the evaluation of the logarithm; another option is to treat them

separately to explain what generates zero demand. After applying the logarithmic transformation, it is possible to observe a different density for transformed values (See Figure 3).

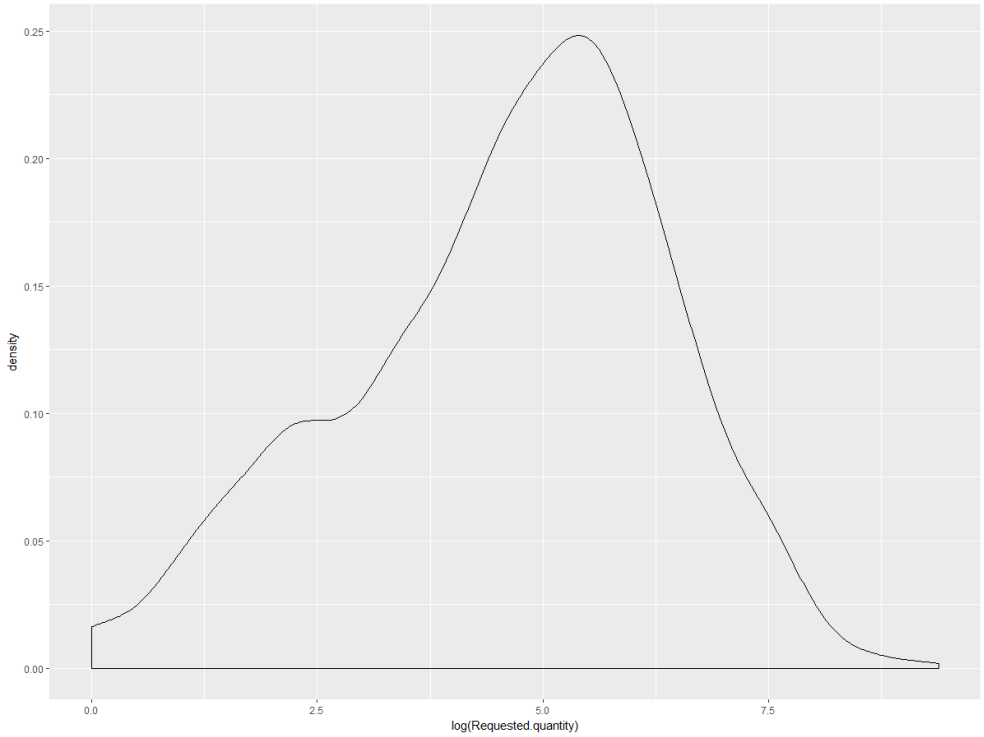


Figure 3: histogram for logarithm of demand

In addition to products' demand, the dataset includes 20 variables (columns) related to these sales. These variables are associated to each product offered in the catalogue in one of three different periods (campaigns), and they describe how the product is offered in the catalogue, some variables are merely descriptive, like product code or description, others are static since they have only a fixed value for the entire dataset. After removing these variables, it is possible to work with the remaining set and then to use some classifiers to understand data and then their connection between themselves and the demand.

In this subsection, every variable non-static or descriptive will be explored. Variables not considered are Product Description, Market (Spain), Model Wearing (Not Model Wearing), Focus (Not visual focus), and Source of Sales (Standard Catalogue).

Campaign

In cosmetics a campaign is a period of sales, consultants offer products to their clients by using a catalogue. A catalogue/campaign normally lasts around 20 days.

In the dataset, the campaign can adopt three different values. The first three campaigns of 2014. The dataset includes 1765 records, 544 products are available for 2014-C01, 605 for 2014-C02 and 616 for 2014-C03 (see Figure 19 at Annex 01).

1160 different products are offered for the three campaigns. Only 113 products (9.7%) are present in the three periods (see Figure 20 at Annex 01).

Category

Seven categories are available for each product: Skin Care, Colour Cosmetics, Toiletries, Accessories, Fragrances, and Wellness.

From the 1160 products available, the largest proportion is related to Colour Cosmetics. A complete distribution of products by Category is available in Table 3

Table 3: Products per Category

Category	Number of Products	Percentage
Colour Cosmetics	466	40%
Toiletries	251	22%
Skin Care	180	16%
Accessories	141	12%
Fragrances	96	8%
Wellness	15	1%

Category	Number of Products	Percentage
Other Category	11	1%

Product Status

Product Status is another product attribute available, it has no related documentation, and it can adopt five different values. For each of the 1160 product, more than 50% of products are tagged with Product Status = "N."

As it is possible to assume that N could stand for "New Products," it is possible to create a crossed table between the times a product is repeated in the three campaigns and the status. Products with status "D" and "N" are available the three campaigns (79%), products with status "L" only two campaigns (18%) and "X," and "Z" just one campaign (less than 3%). This association is confirmed by a Chi-square test = 98.78194 and a p-value = p = 7.571816e-18. Results and Chi-square contribution is available at

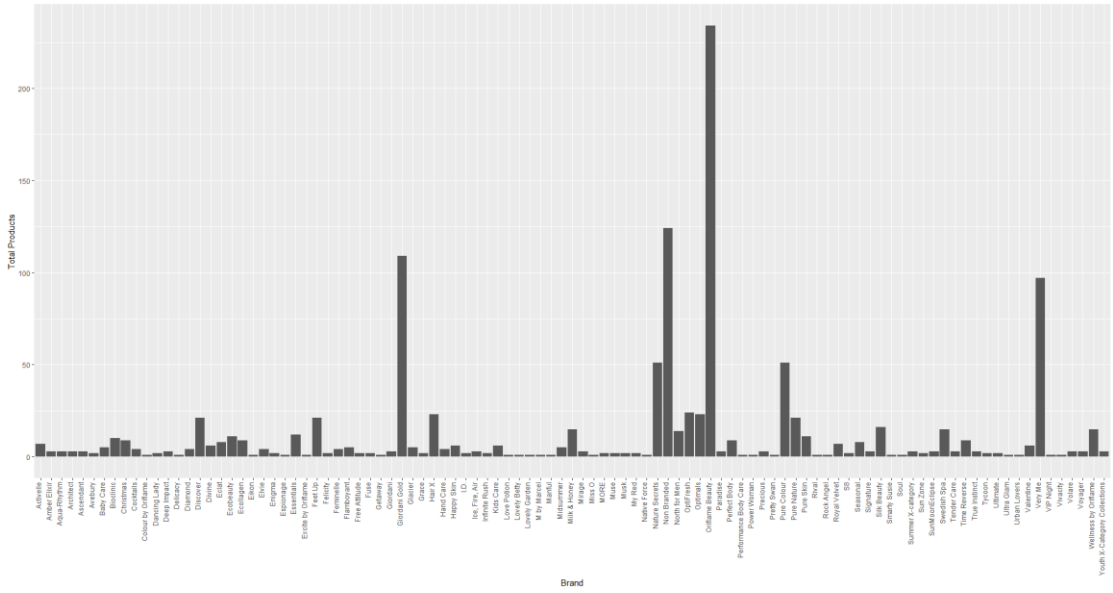


Figure 21: products per brand

Table 9 at Annex 01.

Brand

A brand is also an attribute of the product; a brand can handle different categories and segments. Inside the dataset, it is possible to detect 101 different brands with a median of three products per campaign (mean = 11.49). Nevertheless, there are Brands like 'Oriflame Beauty' that can hold up to 234 products (20% of total 1160).

Segment

Similar to the brand, another product attribute is the Segment, 90 different segments are available for each product. "Lipstick" as a segment holds 137 products (11.8% of 1160 products), nevertheless the median is 5.5, and the mean is 12.89. 45 segments hold less than five products each (see Figure 22 at Annex 1).

Catalogue Driver

This variable can adopt only two different values, and it is related to the way a product is exhibited inside a particular catalogue. There are two products that are catalogue driver for certain campaign and not for the others (code products 21353 in 2014-C01 and 24181 in 2014-C02).

There are only eight products (0.45%) that are catalogue driver in the dataset, two for 2014-C01, and three for the other campaigns. It is expected that being catalogue driver will raise demand for this catalogue.

New Splash

This is another attribute for an offered product in a particular catalogue/campaign. It can adopt three different values: not new splash, new splash, or second new splash. 216 records (12%) can be found as new splash or second splash. For products that have recorded one splash and second splash across campaigns, it can be found that a second splash occurs only after a new splash, and then the proportion of second splash (78 / 4%) is significantly lower than products with only new splash (138 / 8%) see Figure 23

Scratch and Sniff

Thi attribute is available for specific categories like Fragrances, Skincare or Toiletries. From a direct observation in the catalogue, it is a feature that can be opened to smell the scent of the product by opening a sticker or scratching a small square region in a catalogue page.

It can be activated for any campaign; nevertheless, the use is restricted to only twenty products (1%) in the three campaigns (maximum eight products in 2014-C02).

Catalogue Section

In this dataset, catalogue sections are divided in five sections where a product can be located inside the catalogue: Back Cover, Category Sections, Ending Section, Middle Spread, and Platform.

Products are usually located in Category Sections (81.42%) and just a few products in Back Cover (eight products, less than 1%). Sections with limited space are expected to boost demand. E.g. Back Cover. For a detailed distribution of products by sections look Figure 4 and Table 10 at Annex 1.

Offer Percentage

Offer Percentage is a numeric variable that can adopt values between 30% to 100% and 0% if there is no offer involved. This variable seems to be related to the next variable: Offer. As these values can differ from one campaign to another, an analysis per campaign is developed with a boxplot. The median is 30% at any campaign. For the first two campaigns, products with no discount (0%) or with a discount over 60% are considered outliers (See Figure 5). Nevertheless, in 2014-C03 there are no outliers considered, 50% of values centred on the median are varying from 0% to 40%, then the whisker covers all possible values beyond 40%.

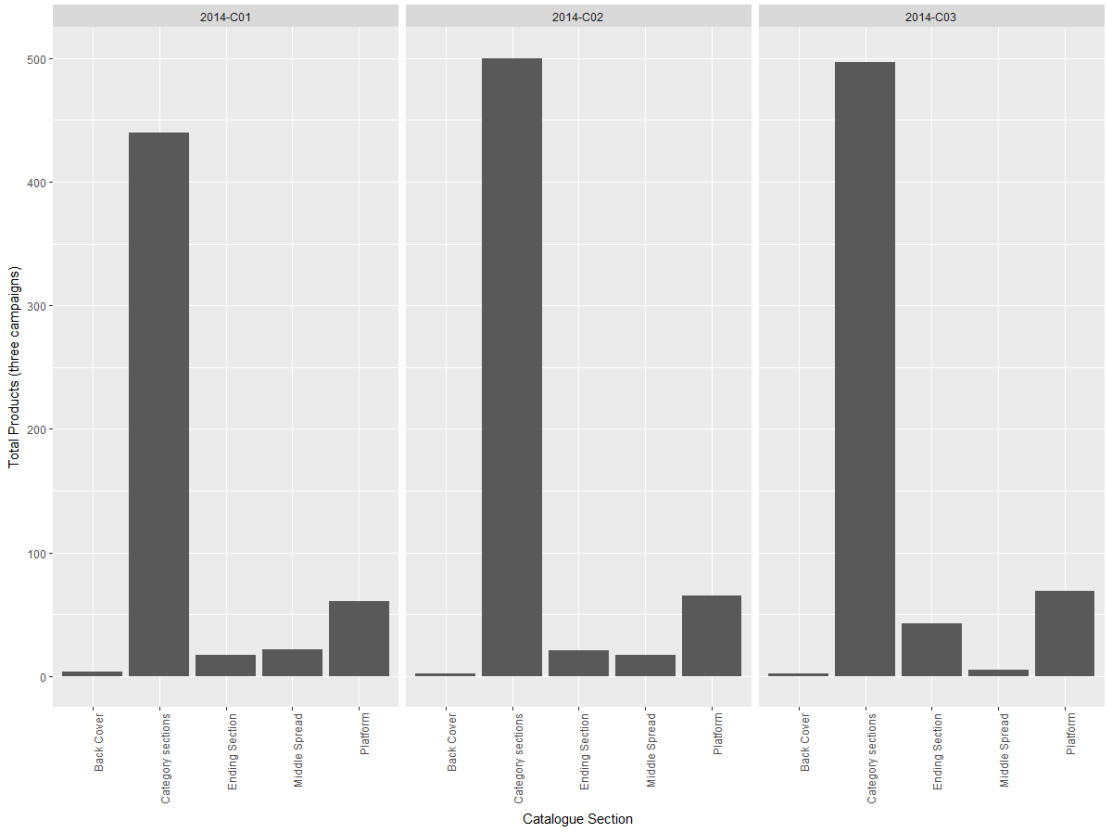


Figure 4: products per Catalogue Section

It is expected more products with no offer for 2014-C03, and then the next step will be to validate how effective is the percentage to increase demand.

Offer

Products are offered with different strategies. The last variable we discussed was Offer Percentage where the most common value was 30%. In the following section, a connection between offer and offer percentage will be analysed. Here we will try to find the most usual offer applied to the products. As the offer is different for each campaign, it is expected to find certain offer more used than others.

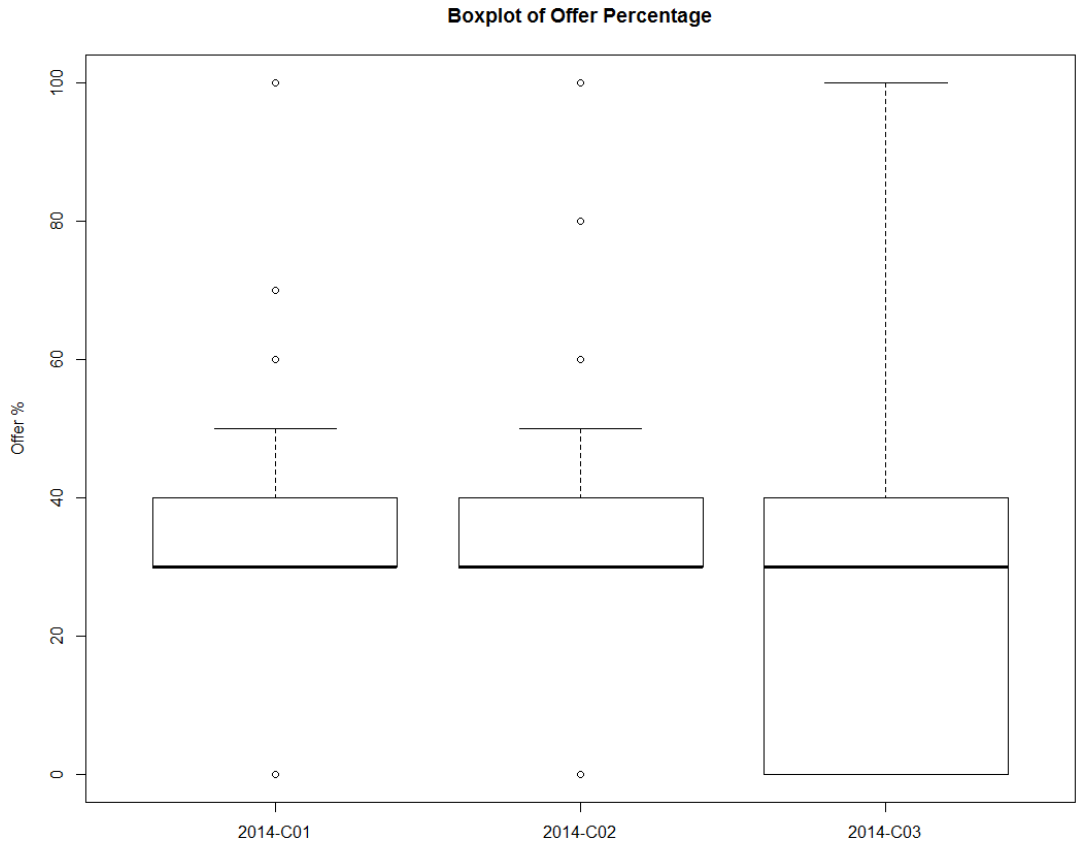


Figure 5: Boxplot of Offer Percentage by Campaign

The most used offer is a "Straight discount" despite the campaigns (64.8% of total records), then "No Offer" (23.4%), for a detailed composition of offers by campaign see Figure 6.

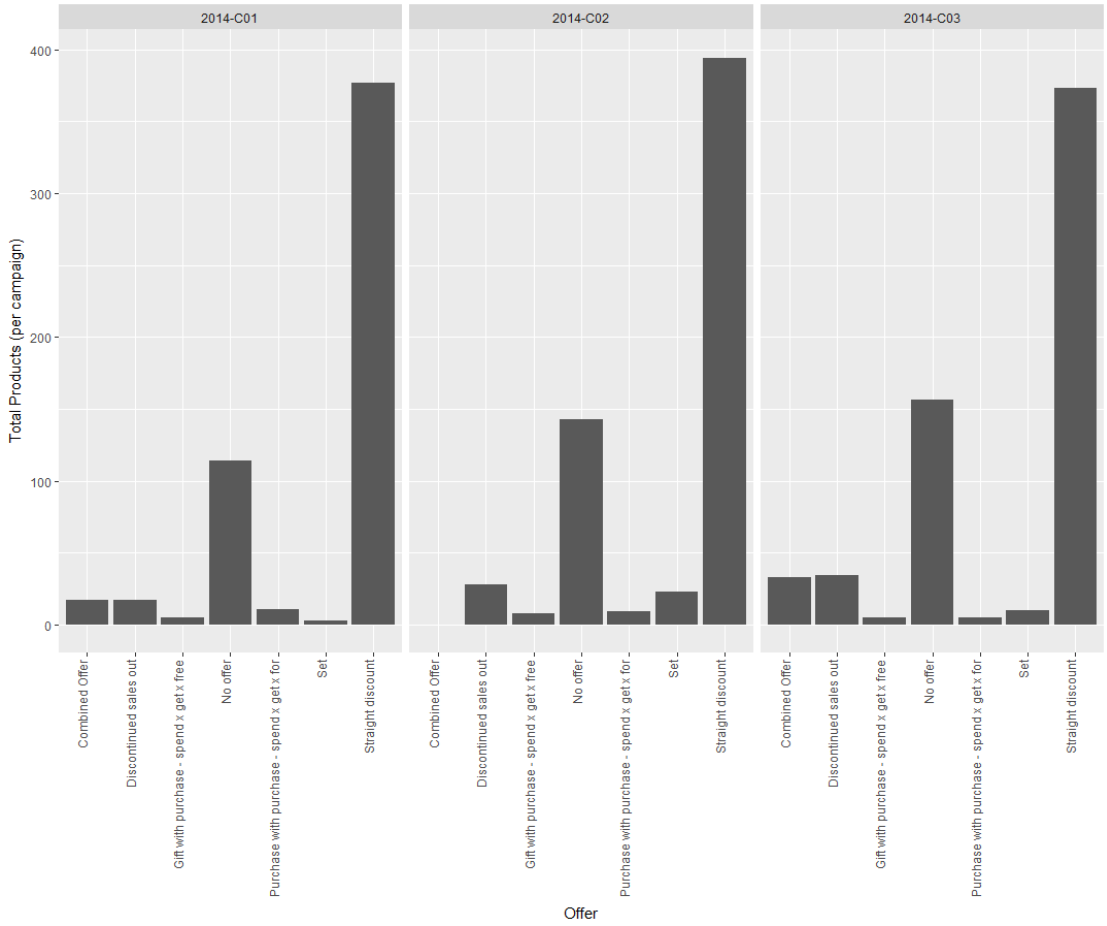


Figure 6: Offer used for products by campaign

Active Consultants

As a final variable to consider, it is possible to work with the number of active consultants. As described in **Error! Reference source not found.** (see Table 1), the number of active consultants is related to the number of sellers they have that orders frequently in a year. This value is normally recorded after each period of sales (campaign).

The numbers of active consultants for the three registered campaigns are 12282, 12001, and 12662 respectively. This value is uncertainty for future campaigns, but it is supposed that it could be forecasted and used as an input if necessary.

Relations between independent variables

After analysing variables by separate, some relations are expected to appear. For instance, a connection between variables related to offers, then a possible dependence between categories, product status.

Some of these relations will be discussed in this section.

A first connection to analyse is the one between Offer and Offer percentage. This relation turns evident when the ranges of percentages are analysed by Offer. Straight discounts are usually 30% and no more than 50%, "No offer" means 0% and "Gift with purchase" is always 100%. A combined offer is equivalent to 50%. "Purchase with purchase" and "Discontinued sales," have the same median: 60%, nevertheless the ranges are slightly different, "Purchase with purchase" could adopt values over 40% and "Discontinued sales out" could not exceed 60%. For an extended visualisation, see Figure 7.

Regarding the product status, where there is no information available, one option was to find a relationship between categories and the status. As a result it is possible to assure that there are no exclusive categories (see Figure 24 at Annex 1) for each status except Status "X" which, for this sample, would be totally related to Wellness products (see Figure 25 at Annex 1). There is also no direct relation between status and the applied offer (see Figure 26 at Annex 1).

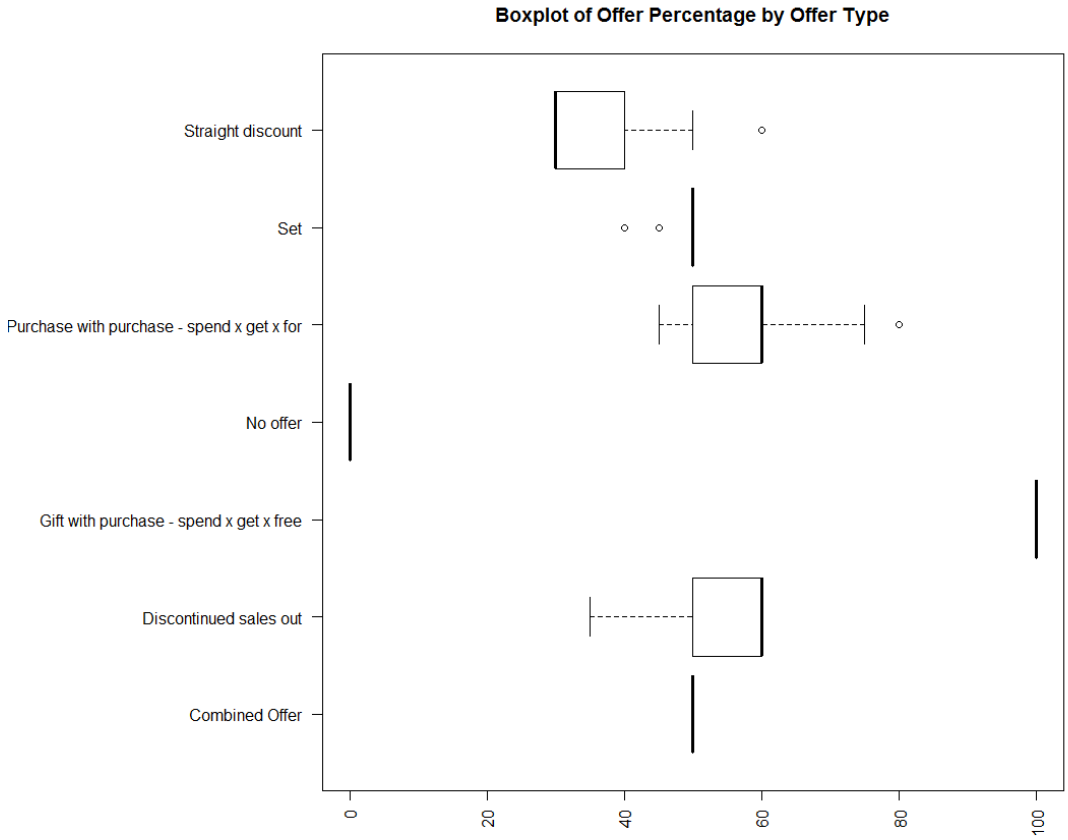


Figure 7: Different offer percentages for different offer types

Before modelling, it is expected that demand is influenced by sales, but there is another option we discussed previously, and it is related to the section inside the catalogue. If the position inside the catalogue would increase the demand, it is necessary to confirm that this position is relevant by itself and not by the associated offer in that position.

The only offer that applies for a specific section is a straight discount for the back cover (see Figure 27 at Annex 1), other sections can combine at least three different offers.

In a different view, by looking at the proportions for Offer percentage for each section (see Figure 28 at Annex 1), certain offer percentages look to be restricted for certain

sections. For instance, it is possible to find offers with 45% to 60% discount on the back cover and no gifts (100%) also at the ending section.

Another possible relation would be present for the location of the product inside the catalogue and the category. From Figure 29 (see Annex 1) it is clear that products are usually presented in category sections. Colour cosmetics, skin care and toiletries are located on the back cover (and this would be related to profit expectations or how profitable are these categories).

Relations between independent variables and the dependant variable

In this subsection, it will be discussed how certain variables or attributes can contribute to increase or decrease demand. In the case of continuous variables, it will be possible to contrast if demand is being affected by changes in the values.

As there are only two continuous variables, offer percentage and active consultants, it is possible to visualise the current relationship by using scatterplots.

Firstly, the relation between offer percentage and demand could be expected in a way that increases in offer percentage could lead to higher sales at any time. Nevertheless, this is not the case as demand falls after 60% of discount (See Figure 30 at Annex 1).

In a different perspective, it is possible to split the same relation by looking every Offer by separate; this could show the effectiveness of each change in discount depending on the offer applied at that time. There it is possible to see that Discontinued sales out increases slightly demand when offer percentage is increased. The straight discount could shoot up sales when the higher discount is applied. In contrast, Purchase by Purchase seems to be less effective if the offer percentage is risen (See Figure 8).

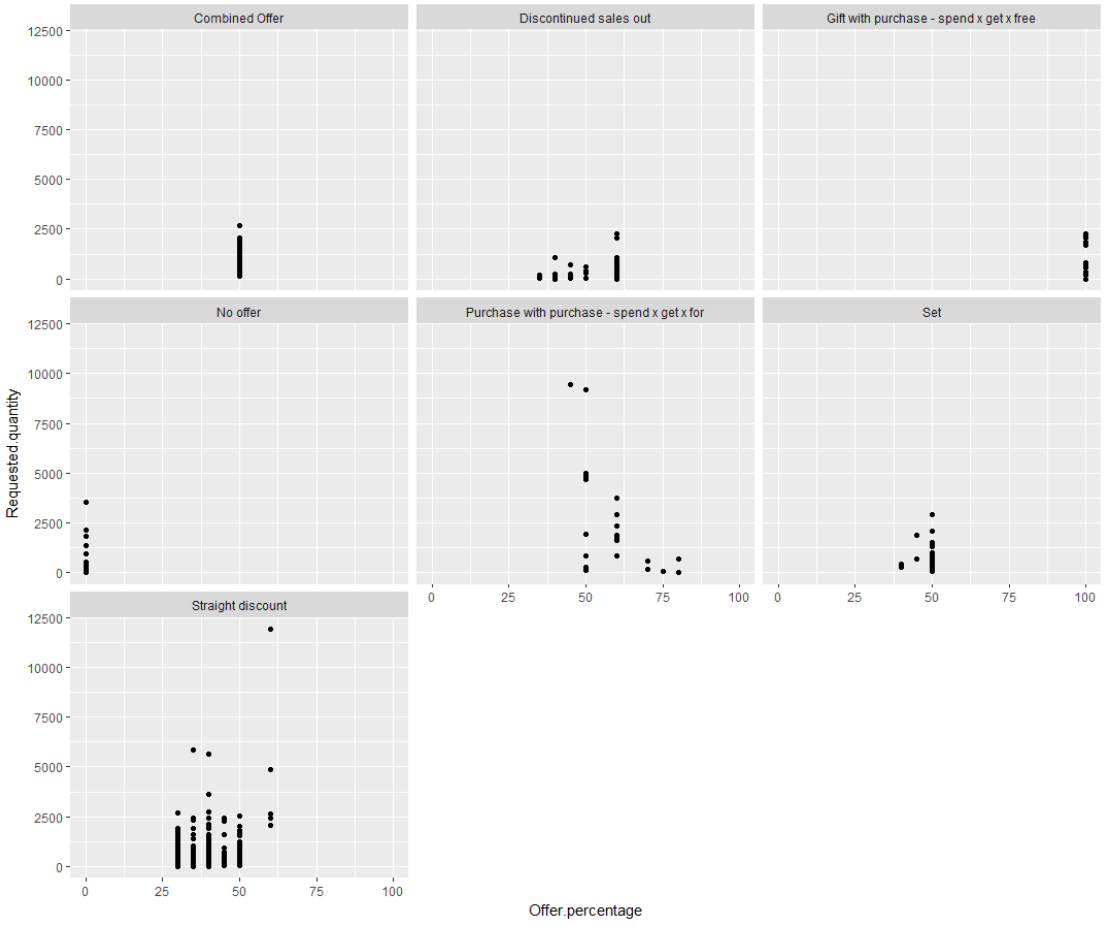


Figure 8: scatterplot of Offer percentage vs Requested quantity (by Offer)

Regarding the number of active consultants, it would appear that the higher the consultants, the lower the demand is, but considering there are only three campaigns (and just three different values for active consultants); a larger sample with more historical data would be necessary to confirm this.

At this time, the next step is to find out how other categorical variables can affect demand or how the demand could react after changing the values of these variables. As the demand can adopt larger values these are possible to be considered as outliers in future sections. In some cases, a logarithmic transformation would be required.

Density plots can explain if demand is sensitive to certain features. These will be used in this subsection.

To see the effectiveness of New Splash (see Figure 32), a density plot can show that a new splash has larger demand than a second new splash, this can be observed for Colour Cosmetics, Fragrances, Skin Care and Toiletries. It is expected that New Splash more effective than a 2nd new splash or if this is not present.

A similar analysis can be performed for Scratch and Sniff (see Figure 33 at Annex 1), as expected, it can be confirmed that demand's mean and median are increased when "Scratch and Sniff" is activated. Nevertheless, this feature is available only for specific categories (Fragrances, Skin Care, and Toiletries).

A similar situation is observed for Catalogue Driver, primarily for Skin Care, demand is more significant when the catalogue driver is enabled (see Figure 34 at Annex 1).

To discuss the most effective product position inside the catalogue, demand density can be divided by catalogue sections. As it is visible in Figure 9, Back Cover is the most useful section to ensure a larger demand, in second place Ending Section and in the third Platform. However, the kurtosis level reduces between sections. In a different view (see Figure 35 at Annex 1), medians can be compared between sections; a boxplot without transforming demand can confirm what is described from the density plot for the same variable. A two-sample Wilcoxon test confirms demand differences.

Finally, it is expected that a higher discount should result in higher sales, but this is only visible when the discount is analysed by the offer, each offer presents a different behaviour and reaction when discounts are applied. Some discounts applied under certain offers are faster to increase sales (Discontinued sales out) than others are (e.g.

Straight Discount). Purchase by purchase, on the other hand, seems to have a negative trend when offer percentage is increased (See Figure 10).

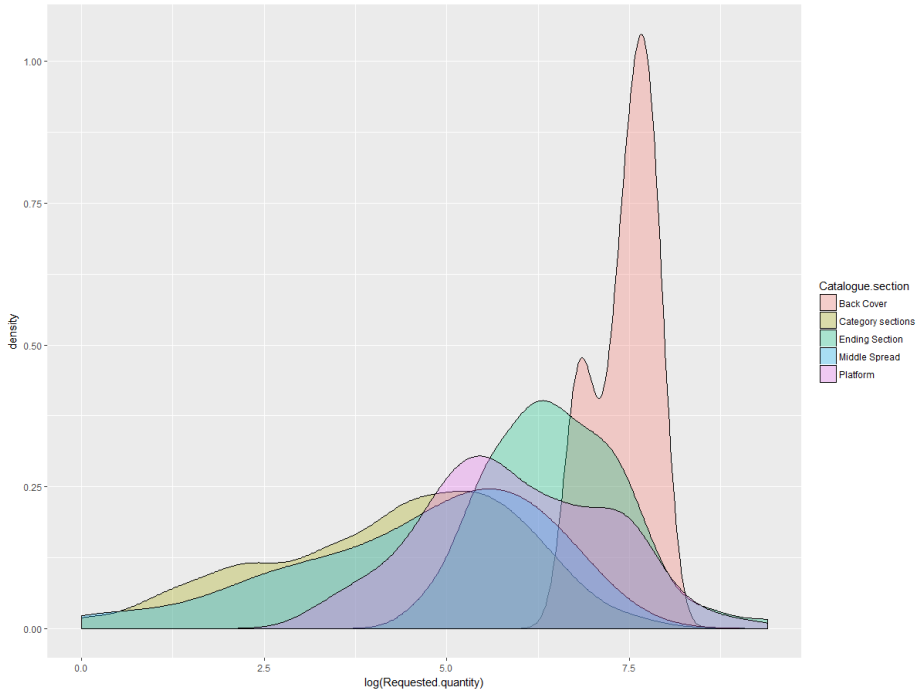


Figure 9: density plot demand by catalogue section

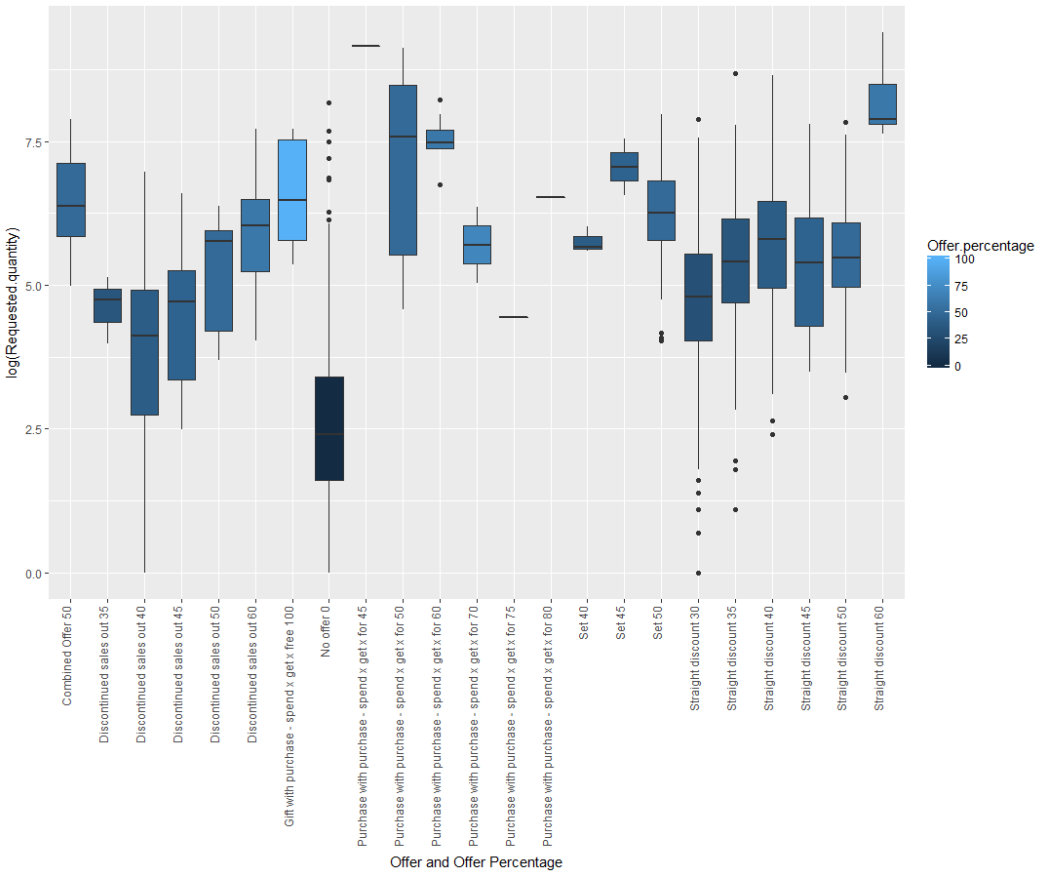


Figure 10: boxplot by Offer and Offer Percentage

Correlation plots

In order to conclude data exploration by looking at connections between variables, some plots were prepared to contrast dependant variable behaviour against independent ones.

Four charts are available at Annex 01 to describe existing relation between variables.

From Figure 36, a negative correlation is present regarding the periods and active consultants (demand decreases as consultants and periods increase).

This also happens for products with status D and N. It is also evident the relationship between consultants and periods, however, there are only three different values for each so this relation could be discussable.

Category Sections (Figure 37) are influencing demand according to the chart, Back Cover and Ending Sections when active can increase demand, the opposite is observed for category sections.

Regarding the offer percentage (Figure 38), it is visible a trend from the loess smooth between Demand and Straight Discount. Regarding categories, Skin Care products are more likely to expect large demand over periods (see Figure 39).

3.3 Experimental Design

3.3.1 Data Sampling and Encoding

The provided dataset considers only thee sales campaigns from sequential selling periods. In order to prepare a valid model and to evaluate out-of-sample performance it is necessary to split data. We use the first two campaigns as a training set and the third one for testing purposes. Whilst potentially introducing sequential biases, this method allows and efficient "out-of-time" evaluation as in a possible real situation with limited data. Also, working with only two periods will create a correlation between periods and number of active consultants since only two different values will be found for each campaign. We split the dataset into training data, considering the first two periods (campaigns) 2014-C01 and 2014-C02 with 1149 records (65%) and test data including the last period 2014-C03 with 616 records (35%). This dataset sampling will be applied to each algorithm, where we expected an impact for statistical as well as machine learning algorithms when changing the size of the training and test set (Karan, et al., 2014).

The number of products per campaign it is almost similar, 544, 605 and 616 for 2014-C01, 2014-C02 and 2014-C03 respectively. Then it is expected that predictions for

2014-C03 could not use history of the same product for some of them. It is also necessary to remember that not all products are offered every period of sales.

3.3.2 Data Pre-processing

Data pre-processing is designed to ensure that every explanatory variable will receive equal importance during the training process (Dawson & Wilby, 2001). Data pre-processing techniques include data transformation, rescaling or standardisation (Wu, et al., 2009) and can have a significant effect on predictive accuracy (Crone et al., 2010).

The provided dataset contains 1765 records and 27 fields, some of them with show repetitive or combined values, many of alphanumerical type. As the dataset is not available to be used directly for modelling, we are required to encode the variables by feature transformation of values into suitable scales as well as imputing missing values and removing outliers, analysing and excluding fields with constant or merely descriptive values by feature selection. Also certain results, metric and previous forecasting calculations are omitted as the used method is not provided.

Thus, this dataset is encoded by transforming all categorical variables using a binary encoding of variables in order to prepare input data for both regression and k-nearest neighbour models equally. Variables Category, Product Status, New Splash, Catalogue Sections, and Segment are transformed into binary (dummy) variables and the rule of (n-1) values (Ord & Fildes, 2013, p. 279) to avoid multicolinearity. Category Section is transformed into six dummy variables (Other Category is omitted), Product Status into four ("X" Status is not considered), New Splash into two binary variables (excluding "Not new splash) and Catalogue Sections into four (excluding "Middle Spread"). Segment will provide 89 variables which will be tested if are necessary to be included if a significant improvement on the model is detected. During the encoding, two high

cardinality variables are detected: Brand and Segment, with 101 and 90 different values respectively. Brand will be not considered because of its large cardinality.

A significant change during the encoding will be related to a combination of related and dependant variables. As the discount (offer percentage) is not directly correlated to demand but the offer applied, it was decided to separate the discount into six different variables with the discount used for each applied offer.

As a result, we derive 28 variables; demand will be used with logarithmic transformation in addition to models with its original value (see Annex 1). An extended version includes 89 extra dummy variables (related to segment).

For this empirical evaluation demand will be used as provided and then transformed into logarithm to develop linear regression models, as the demand can include zero values, calculation of logarithm will be replaced by the equivalent of log(0.1) for only these occurrences. Also as the traditional method for rescaling features for machine learning algorithms (in particular k-Nearest Neighbours) is min-max normalisation which is performed substracting the minimum value and then dividing by the range of the variable (maximum – minimum value) (Lantz, 2013). Nevertheless, using provided R functions from caret package (Kuhn, 2018), the pre-processing selected for the training data will be the one used by default: centring and scaling. This is performed to give equal importance to each variable despite their different range.

The list of variables used for modelling can be observed at Annex 2, the complete list of variables considers Demand, Active Consultants, Period Number as continuous variables, and then categorical (yes/no) variables like Catalogue Driver, and Scratch and Sniff. In Catalogue Phantoms with finite values from one to three, which is

assumed as a continuous one and then four dummy variables for categories and six dummy variables for each offer percentage. Then New Splash and Second Splash is presented separately in two additional variables, four more dummy variables for product status and then six final dummies for category. Consequently, 28 variables are available at first for linear regression models, in addition 90 extra variables that can be added if segments are included in the model. As these 90 variables are dummy variables related to the different values a Segment can adopt, they are not introduced in the early models to avoid overfitting.

3.3.3 Forecasting Algorithms

Regression methods and recently using machine learning algorithms commonly tackle forecasting numeric data. It is observed a scarcity of literature related to demand forecasting or retail forecasting using this method.

3.3.3.1 Linear Regression for Catalogue Forecasting

The multiple regression model, as described by (Ord & Fildes, 2013) considers K explanatory variables $X_1, X_2, \dots, X_K$ and assumes that the dependant variable Y is linearly related to them through the following model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K + \varepsilon$$

Where β_0 is denoted as the intercept and all β_j are the slope for each X_j.

Certain assumptions have to be considered before using these models, these assumptions that (1) the expected value has the form: $E(Y/X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K$, (2) the difference between observations and expected values are the random errors $Y = E(Y/X) + \varepsilon$; (3) the error have zero means, (4) the errors for different observations are uncorrelated with another and with other explanatory variables, (5) the

error terms come from distributions with equal variances, (6) the errors are drawn from a normal distribution.

This method, according to a personal experience, is mostly used by practitioners around the world, mostly to support the demand planning process (Christoph Kilger, 2008). Some applications include forecasting freight demand (Fite, et al., 2002), others could be combined with time series models (seasonal ARIMA) for instance, for food retailers (Arunraj & Ahrens, 2015).

In this work, to refine the model, AIC criterion is used automatically by using R-functions from glmnet package to perform stepwise variable evaluation. The stepwise procedure is performed after using all the variables, backwards by automatic final selection and when used in forward direction, the initial model to begin the process is one that considers all discount variables (per offer), then using both directions is only to observe what direction coincides with the previous stepwise executions (backwards or foreward).

3.3.3.2 k-Nearest Neighbours (k-NN) for Catalogue Forecasting

k-Nearest Neighbours (k-NN) is a special technique due its simplicity to implement. Introduced by Fix and Hodges (1951) for classification when data distribution is unknown, is currently used also to define clusters according to a knowledge base by distance calculations and by using a selected number of closer (near) neighbours to classify (by voting) or to predict a value (by averages).

To describe the algorithm we can adapt a version from a Machine Learning book (Lantz, 2013). Let us consider the following training data: $D = \{(x_1, y_1), \dots, (x_N, y_N)\}$ where N is the number of input/output pairs. y_i is the output/label and x_i is a vector with D

attributes or dimensions (discrete or continuous). Let us define x_{im} as the m-th feature of x_i. Regarding the output y_i, it can be discrete, i.e. $y_i \in \{1, \dots, C\}$ for classification; a discrete variable or continuous, i.e. a real value $y_i \in \mathbb{R}$ for regression purposes.

The objective of the method is to calculate the output **y** for a new element **x** (a test point). k-NN attempts to look at the K most similar training examples assigning the majority class label (majority voting) for classification or assigning the average value (for regression). To work with the algorithm k-NN it is required one parameter K (the number of nearest neighbours to look for) and a distance function $d(x_i, x_j)$ to calculate how close or distant are two different observations (x_i, x_j) to each other.

There are two common ways to determine the distance, a first option is to use Euclidean distance in a D-dimensional space, i.e. $d(x_i, x_j) = \sqrt{\sum_{m=1}^{D}(x_{im} - x_{jm})}$ another option is to apply Mahalonobis distance.

Mahalonobis distance can be defined in the following way: $d_{ij}^2 = (x_i - x_j)^T M(x_i - x_j)$ where x_i and x_j are the vectors of observations of featured variables for the *i-th* and *j-th* units, respectively, and M is a square, positive definite matrix. When M is the identity matrix, Euclidean distance results.

Distance calculations depend on the scale of each feature, a common procedure before calculating any distance is to normalise every variable. Normalisation is performed in two different ways: by using z-score standardisation or by subtracting the minimum value and then dividing by the range of this variable (the distance from the minimum to the maximum). Choosing a metric allow excluding features which have little or no

relevance (Ripley, 1996, p. 197). Another option is to assign weights to features $d(x_i, x_j) = \sum_{m=1}^{D} w_m d(x_{im}, x_{jm})$ to rescale each component.

According to (Al-Qahtani & Crone, 2013), independent works leaded to an application of k-NN for Time Series forecasting (Yakowitz, 1987), (Cleveland, 1979), (Stone, 1977). Nevertheless, it is possible also to connect the theory with chaotic time series predictions (Farmer & Sidorowich, 1987) because of the idea of a state vector with a constant time delay; idea that is supported by (Fiordaliso, 1998) and (Gimeno Illa, et al., 2004).

Predictions for time series using k-NN as a method have been found as a proposal from different authors. First, an application and description of the method for wastewater treatment plant (Gimeno Illa, et al., 2004). Second, a more exhaustive work with a definition of the algorithm is applied to forecast electricity (Al-Qahtani & Crone, 2013) and recently a more extended methodology including recommendations for selecting the number of neighbours and variables (lags) also for time series (Martínez, et al., 2017). Other works also combine other techniques, like Differential Evolution, for Time Series Forecasting (De La Vega, et al., 2014).

It is important to note that k-NN could be combined with other techniques for Time Series forecasting, this is the case with Differential Evolution (which would be more reliable than ARIMA) evaluated in chaotic time series (De La Vega, et al., 2014). Long-term predictions of time series are also feasible by combining the technique with other methods, like genetic algorithms tested over financial datasets after pre-processing to reduce features (Sovilj, et al., 2010).

Despite the lack of massive applications, it is interesting to see the technique being applied to time series in competitions like NN3, just by using the method or combining it with support vector regression (Crone, et al., 2011).

From a different perspective, k-NN is used to predict energy consumption, nevertheless the algorithm is applied as a classifier, not for regression (Wahid & Kim, 2016). There's a patented system and method (Zhang, et al., 2014) including this technique in combination to time series algorithms; this is not applicable for the analysed situation as there are no time series models involved because of the reduced amount of campaigns.

3.3.3.3 Benchmark Algorithms for Catalogue Forecasting

Previous studies in catalogue forecasting did not evaluate accuracy against benchmark algorithms, which is an important omission. In order to have a base to compare the efficiency for each method, we propose a method based on averages to calculate within categories. As the dataset is nested by Brand, Category and Segment, averages from the training set are calculated for each category and segment separately. These averages are used to create a Naïve Forecast for the test data by looking for the associated value to the product segment and if it is not available because the segment was not available in training data, then looking for the respective average for the associated category for this product. We consider this to be a representative benchmark since the representative variables of the brand, category and segment are used, but no additional data is considered to refine this

This approach is deemed more powerful than a Naïve forecast, an approximation to the naïve or random-walk method (Hyndman & Koehler, 2006) which is based only on looking the most recent observation for the same product.

3.3.4 Selected Error Metrics

A further limitation of previous studies was the use of a single, non-robust error metric of squared errors, commonly used in regression modelling but largely disputed in forecasting research (see e.g. Tashman (2000)).

Therefore we propose to use three different measures: RMSE (root of mean squared errors) and MAE (mean absolute error) in order to link our results to previous research findings, and sMAPE (symmetric MAPE). A definition for these metrics is provided by (Hyndman & Koehler, 2006): let Y_t denote the observation at time t and F_t denote the forecast of Y_t. Then the forecast error is defined as $e_t = Y_t - F_t$. The most used scale-dependant measures are based on the absolute error or squared errors: Root Mean Square Error (RMSE) = $mean(e_t^2)$, Mean Absolute Error (MAE) = $mean\ (|e_t|)$. In addition, a measure based on percentage errors (Makridakis, 1993) is introduced: Symmetric Mean Absolute Percentage Error (sMAPE) = $mean(200|Y_t - F_t|/(Y_t - F_t))$ as an alternative to classic Mean Absolute Percentage Error (MAPE) = $mean(100|Y_t - F_t|/Y_t)$ which is largely increased when actual values (Y_t) are close to zero.

These error measures are calculated for training and test sets separately, where test results are calculated over models constructed using in sample data only, and errors on training and test are provided. Forecast errors are computer for each item in training and test dataset and then averaged using Mean and Median to derive mean sMAPE, median sMAPE and so forth. Of all measure we propose to utilise sMAPE, a scaled error metric to allow error summation and comparison across many products with different levels of sales.

4 Experimental Results

4.1 Empirical Accuracy without Preprocessing

The results provided by each model execution considered 2 different algorithms, Linear Regression (LR) and k-Nearest Neighbours (kNN) across different data pre-processing of variable selections, notably enter, forward, backward, and stepwise, with and without data transformation of logarithms of metric scaled variables and with or without outlier removal. Altogether, 18 variants of models were created, and compared to the Naïve benchmark method for accuracy. The results across the multiple error measures RMSE, MAE, sMPAE are presented as mean and median metrics in the table below. In-sample metrics of R-squared and Std.Errors also also provided for Regression models, as they are unavailable for kNN models.

Table 4: empirical Accuracy of Baseline Models

Model (# var)	R2 Adj	Std. Error	In-sample Mean			Out-sample Mean		
			RMSE	MAE	sMAPE	RMSE	MAE	sMAPE
Naïve			690.77	301.98	1.00	531.45	315.26	1.01
LR-ND-Enter(28)	0.46	541.70	535.28	245.06	1.00	582.16	320.02	1.08
kNN-ND Enter(28)			489.01	175.26	0.69	529.61	271.38	0.89
			In-sample Median			**Out-sample Median**		
			RMSE	MAE	sMAPE	RMSE	MAE	sMAPE
Naïve			137.00	137.00	0.97	161.00	161.00	0.97
LR-ND-Enter(28)			116.49	116.49	0.89	146.38	146.38	1.08
kNN-ND Enter(28)			70.50	70.50	0.59	112.05	112.05	0.84

Considering both models of linear regression and k-nearest neighbours without any data preprocessing first, we observ that kNN outperforms both Naïve benchmark and also the Linear regression models, both in in-sample error of mean sMAPE 0.69 versus 1.00 and 1.00 respectively, but more importantly on out of sample errors of mean sMAPE of 0.89 versus 1.08 and 1.01 respectively. An identical ranking is observed on median sMAPE as well as mean and median RMSE and MAE, making the results rather robust. We further note that Linear Regression without variable transformation and selection does not outperform the Naïve benchmark, something that would not have been observed in previous studies of Linear Regression where benchmark methods were omitted.

4.2 Empirical Accuracy by Variable Scaling

In analysing residuals for the parameric linear regression models (see analysis below) it should be observed that the residuals violate most regression assumptions, indicating the need for variable transformations. Therefore we transform metrically scaled variables using the logarithm (LD), both for Regression and kNN, although kNN do not normall yequire such variable transformation as they are assumption free.

For both algorithms, linear regression and kNN the logarithm transform decreases forecast errors in-sample and out-of sample (after retransforming predictions of course), from 1.08 on LR-ND to 0.83 LR-LD and 0.89 kNN-ND to 0.85 kNN-LD out of sample and similar magnitudes in sample. However, improvements for Linear Regression are much more substantial that for kNN, as we would have expected given the lack of assumptions for kNN – however, they are still measurable on mean and median accuracy improvements across multiple erro metrics. This is a nove insight, that kNN might benefit from adequate variable transformations to a certain extent, just not as

much as linear regression would. It should also be noted that the improvement in accuracy from log-transforms for Regression exceed the difference between the algorithms in the LD experiments, indicating that the preprocessing may be more relevant than the choice of algorithms to drive accuracy.

Table 5: Empricial Results of Variable Transformations

Model (# var)	R2 Adj	Std. Error	In-sample Mean RMSE	In-sample Mean MAE	In-sample Mean sMAPE	Out-sample Mean RMSE	Out-sample Mean MAE	Out-sample Mean sMAPE
Naïve			690.77	301.98	1.00	531.45	315.26	1.01
LR-ND-Enter(28)	0.46	541.70	535.28	245.06	1.00	582.16	320.02	1.08
LR-LD-Enter(28)	0.58	1.34	595.68	205.22	0.75	852.14	300.91	0.83
			In-sample Mean			**Out-sample Mean**		
			RMSE	**MAE**	**sMAPE**	**RMSE**	**MAE**	**sMAPE**
Naïve			137.00	137.00	0.97	161.00	161.00	0.97
kNN-ND Enter(28)			489.01	175.26	0.69	529.61	271.38	0.89
kNN-LD Enter (28)			537.77	173.90	0.68	521.98	252.44	0.85

4.3 Empirical Accuracy by Variable selection

Next, we observer the differences in forecasting accuracy depending on the different variable selection, employing forward selection, backward selection and stepwise selection in addition to enter, where all 28 variables are entered into the model irrespective of their statistical significance (for LR). It becomes apparent that the selection of variables using stepwise improves accuracy both for LR and for KNN significantly, but interacts with logistic transforms.

Table 6: Overall Results across Models and Preprocessing

Model (# var)	R2 Adj	Std. Error	In-sample Mean			In-sample Median			Out-sample Mean			Out-sample Median)		
			RMSE	MAE	sMAPE	RMSE	MAE	sMAPE	RMSE	MAE	sMAPE	RMSE	MAE	sMAPE
Naïve			690.77	301.98	1.00	137.00	137.00	0.97	531.45	315.26	1.01	161.00	161.00	0.97
LR-ND-Enter(28)	0.46	541.70	535.28	245.06	1.00	116.49	116.49	0.89	582.16	320.02	1.08	146.38	146.38	1.08
LR-ND-Backw (22)	0.46	541.00	535.76	244.19	1.01	118.69	118.69	0.91	577.61	318.73	1.06	150.45	150.45	1.01
LR-ND-Fwd (18)	0.46	541.40	537.15	242.53	1.05	113.06	113.06	0.97	574.21	313.27	1.13	135.80	135.80	1.13
LR-ND-Stepw (22)	0.47	541.00	535.76	244.19	1.01	118.69	118.69	0.91	577.61	318.73	1.06	150.45	150.45	1.01
LR-LD-Enter(28)	0.58	1.34	595.68	205.22	0.75	62.06	62.06	0.68	852.14	300.91	0.83	76.31	76.31	0.76
LR-LD-Backw (22)	0.58	1.34	578.00	203.61	0.75	60.73	60.73	0.68	926.31	312.09	0.81	77.73	77.73	0.75
LR-LD-Fwd (22)*	0.58	1.34	589.28	205.60	0.75	60.59	60.59	0.68	823.54	301.54	0.81	75.88	75.88	0.75
LR-LD-Stepw (22)	0.58	1.34	578.00	203.61	0.75	60.73	60.73	0.67	926.32	312.10	0.81	77.74	77.74	0.75
LR-LD-Stepw-W/O (21)	0.61	1.06	551.66	202.37	0.70	63.26	63.26	0.64	974.30	308.99	0.80	78.88	78.87	0.75
LR-LD-Stepw-W/O+ Segment52	0.69	0.94	546.28	190.81	0.63	51.68	51.68	0.57	1071.89	306.52	**0.77**	78.25	78.25	**0.74**
kNN-ND Enter(28)			489.01	175.26	0.69	70.50	70.50	0.59	529.61	271.38	0.89	112.05	112.05	0.84
kNN-ND Backw (22)			495.93	186.14	0.75	83.68	83.68	0.66	522.67	262.64	0.86	103.78	103.78	0.82
kNN-ND Fwd (18)			591.91	296.12	1.85	147.23	147.23	1.90	528.17	271.21	0.87	120.64	120.63	0.82
kNN-LD Enter (28)			537.77	173.90	0.68	54.96	54.96	0.58	521.98	252.44	0.85	88.56	88.56	0.77
kNN-LD Backw (22)			544.98	179.95	0.71	55.90	55.90	0.61	524.20	254.35	0.83	89.45	89.45	0.76
kNN-LD Fwd (22)*			550.53	186.58	0.72	56.20	56.20	0.63	524.20	251.57	0.83	79.54	79.54	0.77
kNN-LD-Stepw W/O (21)			551.27	183.99	0.67	59.03	59.03	0.58	525.75	258.78	0.83	94.05	94.05	0.78
kNN-LD-Stepw-W/O+Segment52			551.41	175.04	**0.62**	53.53	53.53	**0.55**	530.73	265.53	0.89	100.43	100.41	0.87

LR= Linear Regression, kNN =k-Nearest Neighbour model, ND=Non-transformed demand, LD=log transformed demand, Procedure: including Enter = All variables, Backwards, Forward, Stepwise = both directions. Segment: when 90 extra dummy variables for each segment are included.

Overall, knn are outperformed by properly specified LR models, but over all transformation KNN show a much more robust performance than LR, indicating their robustness to data conditions and less reduced need for variable transformations and data preprocessing.

The performance for models, including stepwise procedure or kNN calculates a result before 50 seconds, so both are applicable in practice.

We also note that kNN are somewhat prone to overfitting, so in selecting on validation data we would have chosen a kNN model over a LR, given the lowest error of 0.55 over 0.57, but at the cost of higher out of sample errors of 0.87 versus 0.74. However, both algorithms significantly outperform the benchmark method, providing the empricial evidence of the value of using analytical forecasting approaches form statistics and machine learning in catalogue retailing.

4.4 Residual Analysis of Regression

4.4.1 Linear regression model with non-transformed demand

To develop the first regression model, a first summarised dataset with only 28 variables is selected. Demand values are used as provided and then by applying a logarithmic transformation.

Then the training set will be evaluated in two ways: using all the variables, then using Akaike Information Criteria (AIC) to execute a stepwise regression with three different procedures (backwards, forward and both directions). To compare these models, the Std. Error, in the sample and out of sample error measures are calculated. Also the coefficients signs are evaluated to assure coherence in the model.

Initially, working with provided demand, Stepwise procedure with backwards direction ended with 22 variables returning the same provided model by automatic procedure using both directions. When using stepwise with forward direction, it started with six variables (offer percentage) and then it increased to 18.

So, in these executions, it is not observed a significant improvement after using the stepwise procedure in any direction. sMAPE for both, in-sample and out-of-sample is over 100%. Regarding the calculated coefficients and the sign, from the last model it is observed that the returned intercept is negative and catalogue section are always positive, with larger values for the back cover, product status is always negative and categories contribute to demand but wellness, offers are only positive for Straight Discounts and Discontinued products. When watching the residuals (see Figure 11), and Cook's distances, it is possible to detect at least three outliers (see Table 7). These outliers provide evidence that a linear model by itself is not sufficient to predict larger demand as Cook's distance would suggest to remove them to improve current model performance (46.56% of variability explanation).

Table 7: outliers detected by the stepwise procedure

Record #	Actual Demand	Predicted Demand
100	9452	3900
433	9202	3150
553	11912	2937

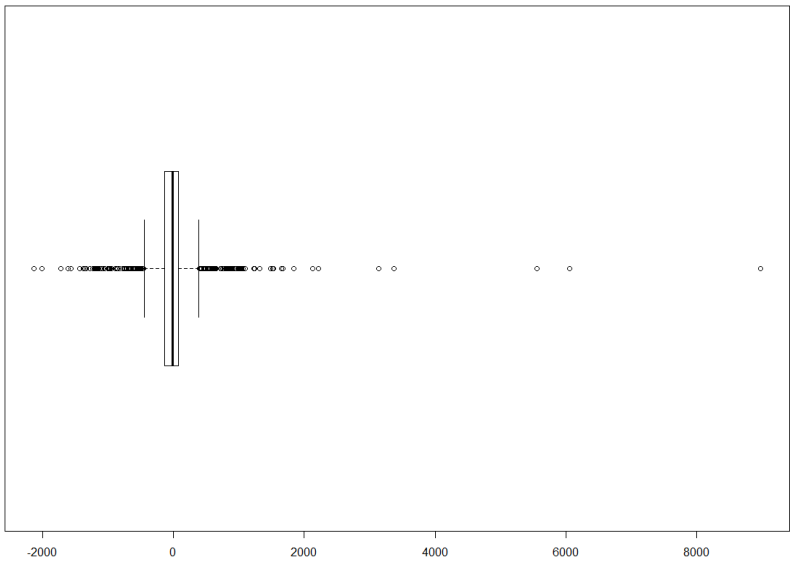


Figure 11: residual boxplot for linear model with stepwise procedure (demand not transformed)

Outputs from R executions and residual plots are available at Annex 3.

4.4.2 Linear regression model with log-transformed demand

A second stage considers the same dataset; just transforming demand into a logarithm, then a log-linear model is proposed, and it is expected this will be able to predict large values (above 5000 units).

Then to prepare log-linear models, a logarithmic transformation is applied over demand after adding a minimum value to avoid the exclusion of 41 records that have demand equal to zero (see Annex 3).

In this case 22 variables are used by any procedure, the category that can generate larger demand is Skin Care and the opposite is presented for Wellness that could return the minimum demand. Regarding catalogue location, Back Cover and Ending Sections are presented as the best locations to increase demand. Product Status (D, N and L) appears

to restrict demand for these products. Offers and percentage, in this model, all are positive and their coefficients reveals that a Straight Discount and Combined Offer and Discontinued products are most effective when combining different discounts. In this model, all coefficients related to discount percentages are positive which follows expectations, as discount is always the most important driver for demand.

An improved R^2-Adj (close to 60%) can reveal a significant improvement to explain demand through this model, sMAPE goes below the naïve model, nevertheless there are certain finding in residual analysis that are discussed and lead to improve the model.

If outliers were present at previous models (without demand transformation) because they were unable to capture larger values, in this case, the opposite occurs. From residual plots (see Figure 12, Annex 5 also) it is observed that certain residuals (outliers in fact) shows a linear relation with fitted values. However Cook's distance suggests no major improvements over the model after removing these outliers.

The three new detected outliers (for log-linear models) can be observed in Table 8.

Table 8: outliers detected by stepwise log-linear model

Record #	Actual Value	Predicted
965	0	171
1624	0	113
1706	0	269

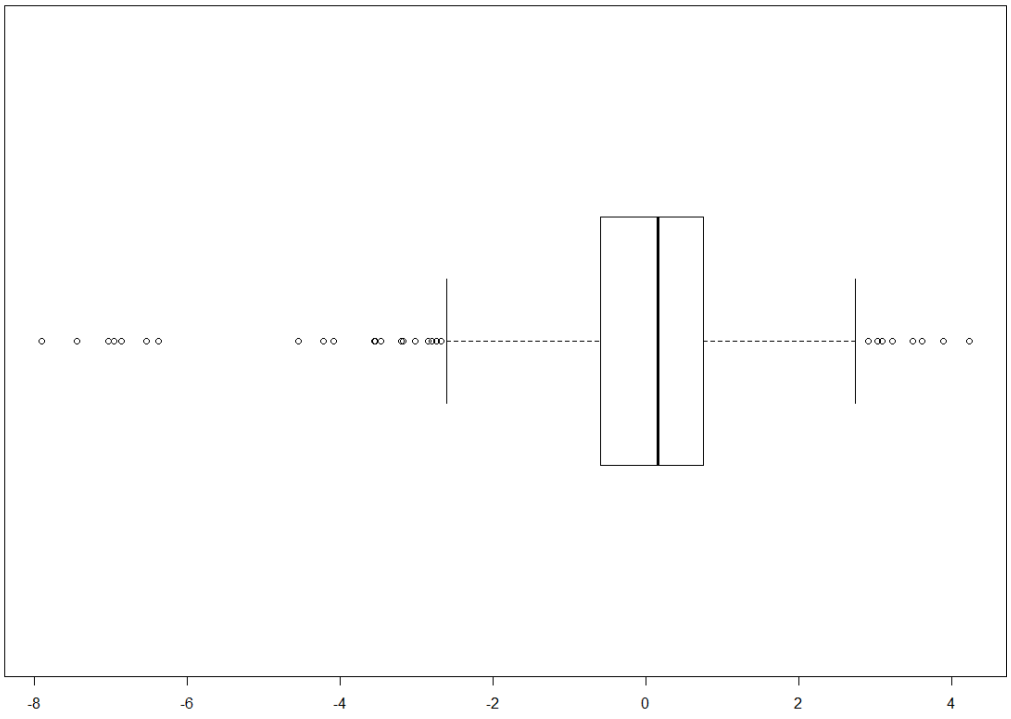


Figure 12: residual boxplot for log-linear model with a stepwise procedure

Then two additional models are proposed to enhance performance, despite the Cook's distances in this case do not reveal a significant improvement over the model after removing outliers. First, by removing from training data, records with demand equal to one or less than one which performs a better fit for small and large values (see Figure 12) and then in a second version, an extended model with more dummy variables (including Segment) and the proposed filter to evaluate if their inclusion could be significant to increase model robustness.

Increasing the number of dummy variables by including segments seems not to provide a significant improvement since the model could lead in overfitting and the RMSE is increased for the out-of-sample. Just the sMAPE is slightly reduced by only 0.04 and Q-Q Plots show a better adjustment (see Figure 14) but the risk of overfitting is present.

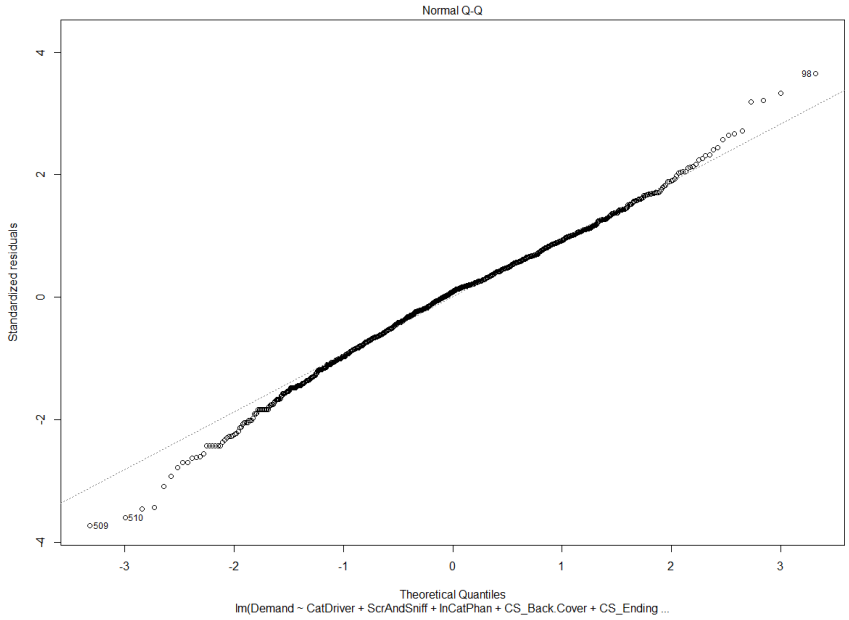


Figure 13: Q-Q Plots for linear regression models excluding small demand (less than 2)

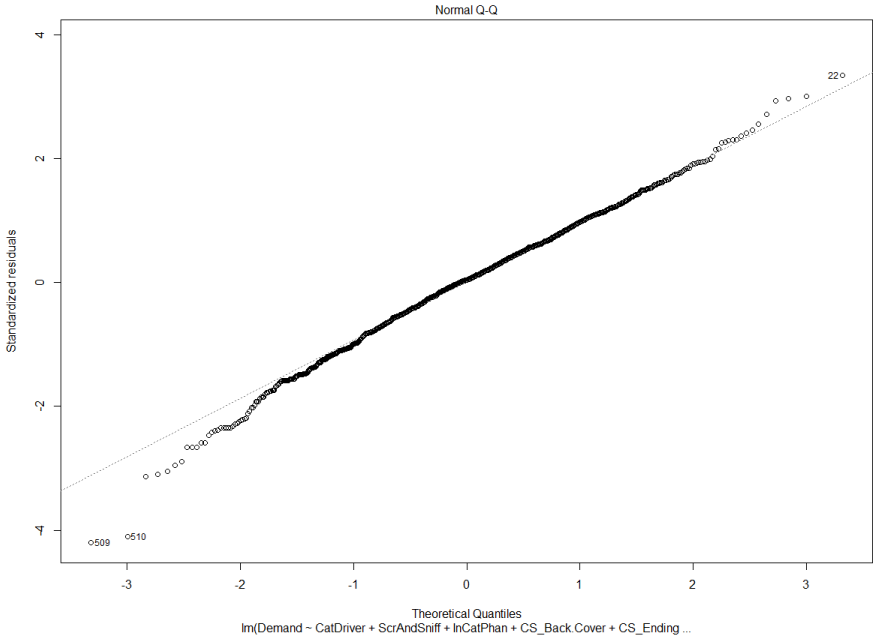


Figure 14: Q-Q Plots for linear regression models excluding small demand (less than 2) and adding dummy variables for segment

4.5 Residual Analysis of k-NN

As the k-Nearest Neighbours Technique is based on an average calculation of the "k" Nearest Neighbours (more similar records). Caret R-package includes the algorithm to be used for prediction. However the "k" parameter has to be selected before any prediction.

As the parameter "k" is essential before using k-NN; the selected method to perform this calculation will be repeated cross-validation, thus, the number of neighbours (k) is returned after several executions and re-samples. R caret package will be used for this purpose (function train, trainControl and predict).

In similar studies, however, for this particular situation, it is possible to approximate the experiment to a real case, where with available data it is necessary to predict a new period at least.

To calculate the value of k, five repeated ten-fold validations are used over prepared training data. To get more confidence about the selected number of neighbours, the procedure will be repeated twice by choosing two different metrics: RMSE and MAE.

For the dataset with non-transformed demand, the selected number of neighbours (k) is set to 5 for the full training set. This value is selected for both chosen metrics (see Figure 15 and Figure 16.).

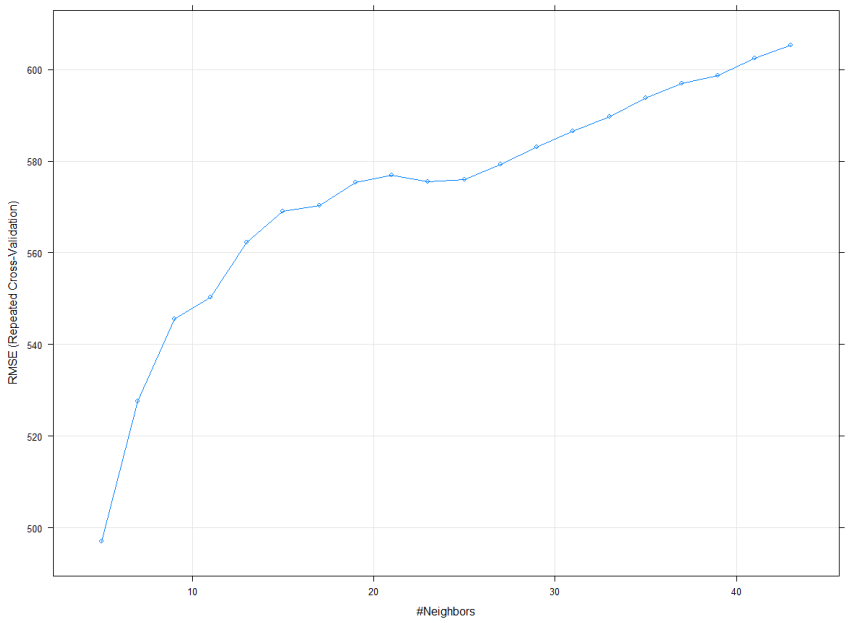


Figure 15: repeated cross-validation result (original demand), optimising RMSE

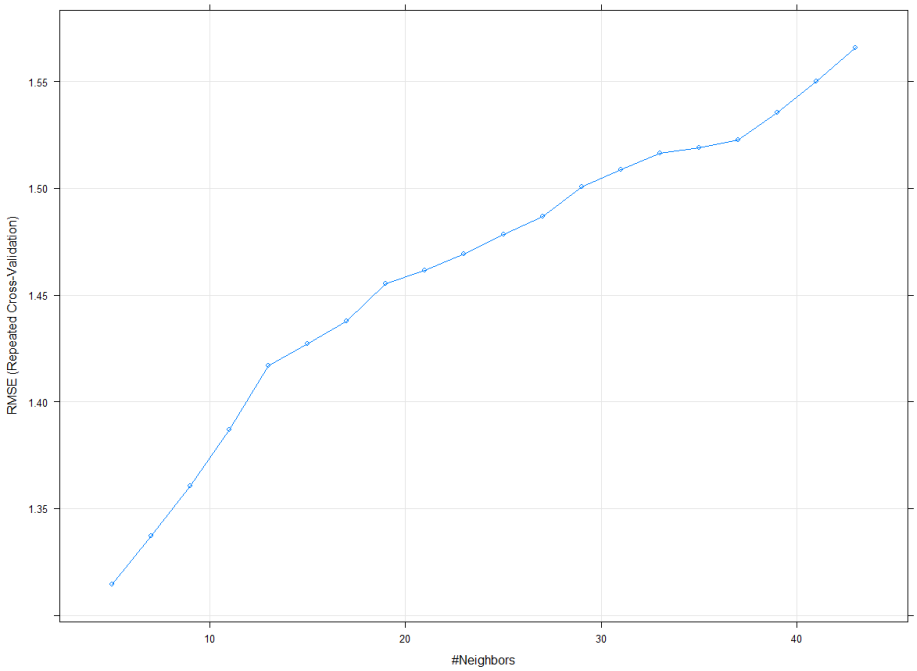


Figure 16: repeated cross-validation result (log-demand), optimising RMSE

Regarding residuals, the Q-Q Plot for k-NN (see Figure 17), it is possible to observe that there's still a problem to predict small values which could be considered as outliers.

These outliers can be removed from training data; however, error metrics are not significantly improved after removal when testing. And the risk of removing zeroes would increase the error since an specific number of neighbours is required anyway to perform a prediction and no zero demand evidence into history could generate significant forecasts when the expected value according to data should be zero or a value close to zero.

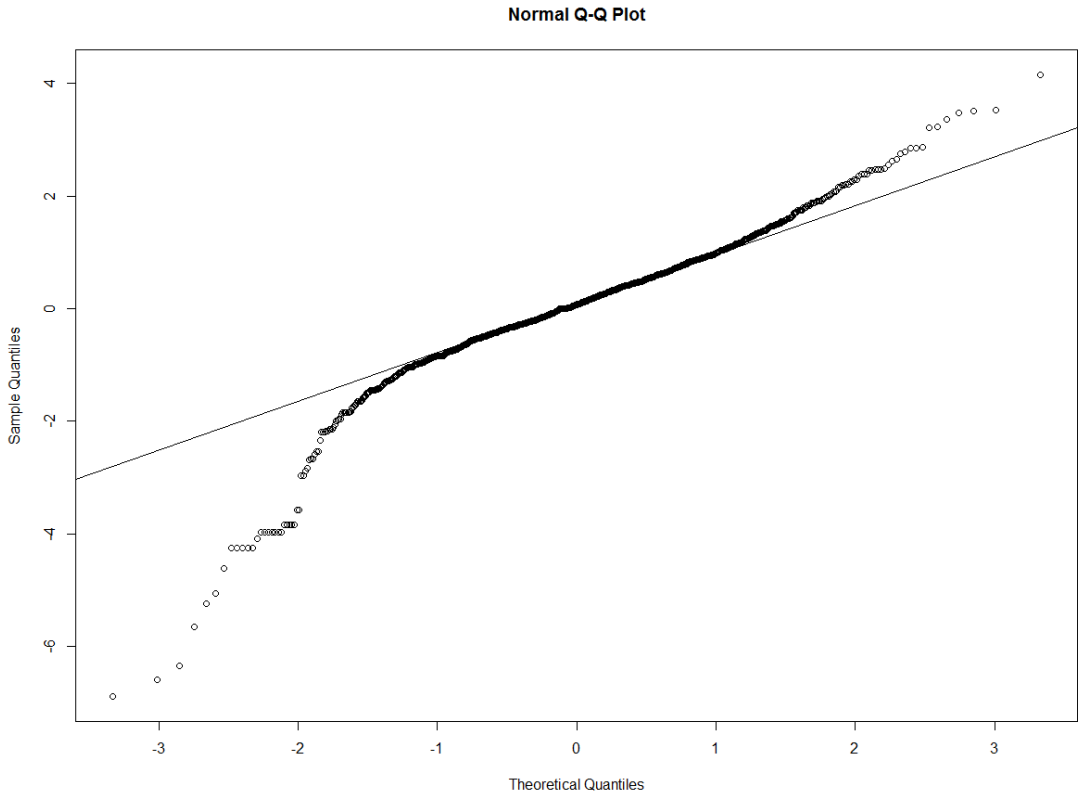


Figure 17: Q-Q Plot for residuals of k-NN Model

A complete summary of executed models is presented at Table 6. A particular behaviour is observed in the Out-of-sample predictions, in all cases, except the Naïve proposed model or k-NN, error metrics are increased, as it is expected as the data is new. Nevertheless, a significant reduction is observed for prediction errors for RMSE and MAE when using k-NN or Naïve.

By adding segments or excluding outliers, it is observed that apart from the risk of model overfitting, there is no significant improvement but the smallest sMAPE value when using the log model without outliers and including segments.

Excluding outliers for the log model, improves MAE and sMAPE results, but this reduction of data increases the RMSE as these outliers can repeat in the future. As there are only three periods to evaluate, it is not recommended to exclude data.

Despite linear regression models present a better result at sMAPE for this particular case, kNN with a reduced RMSE and MAE appears to be a promising technique for prediction, here again it is also clear that including segments, and possibly brands as an input would not significantly reduce errors.

5 Conclusion

In this dissertation two methods, one classic and a novel one are applied to a particular problem for retail forecasting where academic literature is not extended at this time. In addition this piece of work highlights the relevance of data exploration and preprocessing, which are significantly important before modelling.

From data exploration and prior studies it is suggested that catalogue sales, possibly influenced by the usual channel where they exist, direct selling, presents an exponential behaviour; this observation is important, as it will require data transformation before model fitting and this transformation, encoding and preprocessing will possible improve prediction but it will generate outliers.

The relevance of data exploration allows the researcher to perceive connections and behavioural changes in variables influenced by other (e.g. demand by category would vary depending the position of the product inside the catalogue) and also how marketing policies can affect demand (e.g. scratch and sniff are highly correlated to large demand, but it cannot be used more than eight times per campaign). Furthermore, there are soft

concepts, like catalogue design, page distribution that could affect any model prepared by using this data and it could be important to include in a future work.

Regarding modelling, it is clear that a trade-off between simplicity and error accuracy is present, including high cardinality variables, like segments, brand or product could lead into overfitting with just small or no improvement over forecast accuracy.

In addition it is observed how two completely opposite methods (a "classical" statistical linear regression and k-NN "the lazy learner") can complement and bring improved results for skewed demand, in one side the linear regression model can suggest the most relevant variables which when used by k-NN with a small number of neighbours (k=5) can improve the forecast significantly. For this particular example, kNN cannot improve significantly error measures by itself, if R packages can allow to execute training with k-fold validation and cross validation. It is visible that results are acceptable when working with selected variables from linear regression models.

kNN predictions shows a particular behaviour for the provided dataset as they can be enhanced when using a refined dataset without outliers. Nevertheless, this omission could lead into a bad forecast. If outliers are removed, it would increase error for predictions as the knowledge has been supressed before. A proposed option could include a two-stage prediction system, first analysing what factors could create very large demand (above 4000 units) or non-significant demand, if these cases are observed before using any regression model, it is possible to anticipate any model deficiency because of this data dissection. An attempt to predict zero demand can be found in Annex 6, by using Decision Trees. Nevertheless, this is out of scope for this dissertation and it has to be included in a future work. An additional recommendation, for a future

researcher, would be to evaluate performance for other machine learning algorigthms, as Random Forest, Neural Networks of Gradient Boost (XGBoost).

Finally, this dissertation reveals that despite a lack of academic literature and a short-term dataset, current and traditional techniques can create synergies to deliver an improved forecast. Very rudimentary techniques, as the proposed for the naïve method, are still used at companies when a model cannot be fitted, then it gives also the opportunity to improve significantly essential company processes such as Demand Planning.

6 References

A. M. Findlay, L. S., 2002. *Retailing: The evolution and development of retailing.* 1st ed. London: Taylor & Francis.

Al-Qahtani, F. H. & Crone, S. F., 2013. *Multivariate k-nearest neighbour regression for time series data — A novel algorithm for forecasting UK electricity demand.* Dallas, 2013 International Joint Conference on Neural Networks (IJCNN).

Arunraj, N. S. & Ahrens, D., 2015. A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting. *International Journal of Production Economics,* Volume 170, pp. 321-335.

Azuma, H., Suzuki, N., Matsuodani, T. & Tsuda, K., 2016. *Proposal For A Cumulative Deposit Rate Prediction Method For Payment After Delivery In The Mail Order Business.* Atlanta, Proceedings 2016 Ieee 40th Annual Computer Software And Applications Conference Workshops (COMPSAC).

Belcorp, 2015. *¿Qué es el Programa Brillante Belcorp?*. [Online] Available at: http://comunidad.somosbelcorp.com/t5/Historias-de-%C3%A9xito/Qu%C3%A9-es-el-Programa-Brillante-Belcorp/td-p/7596 [Accessed 22 08 2018].

Blattberg, R. C. & Deighton, J., 1996. Manage Marketing by the Customer Equity Test. *Harvard Business Review,* Volume July-August, pp. 136-144.

Boada, A. J., 2017b. Demand Forecast Systems. Automated Forecast Practical Case In Catalog Sales Companies. *Revista Perspectiva Empresarial,* 4(1), pp. 23-41.

Boada, A. J., 2017. Sistema de proyección de la demanda. Caso práctico de predicción automatizada en empresas de venta por catálogo.. *Perspectiva Empresarial,* 4(1), pp. 23-41.

Boada, A. J. & Mallorca, R., 2011. Planificación de demanda, en empresas con estilo de venta por catálogo. *Revista Lasallista de Investigación,* 8(2), pp. 124-135.

Chambers, M. & Eglese, R., 1986. Use Of Preview Exercises To Forecast Demand For New Lines In Mail-Order. *Journal Of The Operational Research Society,* 37(3), pp. 267-273.

Chambers, M. & Eglese, R., 1988. Forecasting Demand For Mail Order Catalog Lines During The Season. *European Journal Of Operational Research,* 34(2), pp. 131-138.

Christoph Kilger, M. W., 2008. Demand Planning. In: *Supply Chain Management and Advanced Planning* . Berlin, Heidelberg: Springer, pp. 133-160.

Cleveland, W. S., 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association,* Volume 74, pp. 829-836.

Cook, G., 2001. *The Discourse of Advertising.* Second ed. Oxon: Routledge.

Crone, S. F., Hibon, M. & Nikolopoulos, K., 2011. Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction. *International Journal of Forecasting,* 27(3), pp. 635-660.

Dawson, C. W. & Wilby, R. L., 2001. Hydrological modelling using artificial neural networks. *Progress in Physical Geography: Earth and Environment,* 25(1), pp. 80-108.

de Brito, M. P., Carbone, V. & Meunier Blanquart, C., 2008. Towards a sustainable fashion retail supply chain in Europe: Organisation and performance. *International Journal of Production Economics,* 114(2), pp. 534-553.

De La Vega, E., Flores, J. & Graff, M., 2014. *k-Nearest-Neighbor by Differential Evolution for Time Series Forecasting.* Tuxtla Gutiérrez, Mexico, Nature-Inspired Computation and Machine Learning. MICAI.

Farmer, J. D. & Sidorowich, J. J., 1987. Predicting chaotic time series. *Physical Review Letters,* 59(8), pp. 845-848.

Festervand, T. A., Snyder, D. R. & Tsalikis, J. D., 1986. Influence Of Catalog Vs. Store Shopping And Prior Satisfaction On Perceived Risk. *Journal Of The Academy Of Marketing Science,* 14(4), pp. 28-36.

Fildes, R., 2017. Research into Forecasting Practice. *Foresight: The International Journal of Applied Forecasting,* Winter(44), pp. 39-46.

Findlay, A. M. & Sparks, L., 2002. *Retailing: The evolution and development of retailing.* 1st ed. London: Taylor & Francis.

Fiordaliso, A., 1998. A nonlinear forecasts combination method based on Takagi–Sugeno fuzzy systems. *International Journal of Forecasting,* 14(3), pp. 367-379.

Fite, J. T. et al., 2002. Forecasting freight demand using economic indices. *International Journal of Physical Distribution & Logistics Management,* 32(4), pp. 299-308.

Fix, E. & J. L. Hodges, J., 1951. *Discriminatory analysis, nonparametric discrimination: consistency properties.* Randolph Field, Texas.: Tech. Rep. 4, USAF School of Aviation Medicine.

Forrester, 2009. *US Online Retail Forecast, 2008 To 2013.* [Online] Available at: https://www.forrester.com/report/US+Online+Retail+Forecast+2008+To+2013/-/E-RES53795
[Accessed 05 October 2018].

Gimeno Illa, J. M., Béjar Alonso, J. & Sànchez Marré, M., 2004. Nearest-Neighbours for Time Series. *Applied Intelligence,* 20(1), pp. 21-35.

Goldstein, J., 2013. *101 Amazing Facts about Wales.* 1st ed. s.l.:Andrews UK Limited.

Green, M. & Harrison, P. J., 1973. Fashion Forecasting for a Mail Order Company Using a Bayesian Approach. *Operational Research Quarterly (1970-1977),* pp. 193-205.

Griffith, D. A., Krampf, R. F. & Palmer, J. W., 2001. The Role Of Interface In Electronic Commerce: Consumer Involvement With Print Versus On-Line Catalogs. *International Journal Of Electronic Commerce,* 5(4), pp. 135-153.

Hall, J., 2007. *A catalogue of success stories for retailers.* [Online] Available at: https://www.telegraph.co.uk/finance/migrationtemp/2805221/A-catalogue-of-success-stories-for-retailers.html
[Accessed 05 October 2018].

Hevrde, J., 2017. *Montgomery Ward's First Catalog.* [Online] Available at: http://www.chicagotribune.com/news/nationworld/politics/chi-chicagodays-firstcatalog-story-story.html
[Accessed 05 October 2018].

Hyndman, R. J. & Koehler, A. B., 2006. Another look at measures of forecast accuracy. *International Journal of Forecasting,* 22(1), pp. 679-688.

Iqani, M., 2012. *Consumer Culture and the Media - Magazines in the Public Eye.* First ed. New York: Palgrave Macmillan.

Karan, M. et al., 2014. *The impact of training data tailoring on demand forecasting models in retail.* Opatija, Croatia, IEEE, pp. 1473-1478.

Kashyap, A. K., 1995. Sticky Prices: New Evidence From Retail Catalogs, The Quarterly Journal Of Economics. *The Quarterly Journal Of Economics,* 110(1), p. 245–274.

Kuhn, M., 2018. *https://www.rdocumentation.org/packages/caret/versions/6.0-80/topics/preProcess.* [Online] Available at: https://www.rdocumentation.org/packages/caret/versions/6.0-80/topics/preProcess
[Accessed 04 10 2018].

Kumar, S., Massie, C. & Dumonceaux, M. D., 2006. Comparative innovative business strategies of major players in cosmetic industry. *Industrial Management & Data Systems,* 106(3), pp. 285-306.

Lantz, B., 2013. *Machine Learning with R.* Birmingham: Packt Publishing Ltd.

Lantz, B., 2013. *Machine Learning with R. Packt Publishing.* [Online] Available at: https://app.knovel.com/hotlink/toc/id:kpMLR0000G/machine-learning-with/machine-learning-with
[Accessed 23 05 2018].

Makridakis, S., 1993. Accuracy measures: Theoretical and practical concerns. *International Journal of Forecasting,* 9(4), pp. 527-529.

Martínez, F., Frías, M. P., Pérez, M. D. & Rivera, A. J., 2017. A methodology for applying k-nearest neighbor to time series forecasting. *Artificial Intelligence Review,* pp. 1-19.

Masmoudi, M., 2011. *Forecasting Returns In Reverse Logistics: Application To Catalog And Mail-Order Retailing.* Metz, France, International Conference on Industrial Engineering and Systems Management.

Mathwicka, C., Malhotrab, N. K. & Rigdonc, E., 2002. The Effect Of Dynamic Retail Experiences On Experiential Perceptions Of Value: An Internet And Catalog Comparison. *Journal Of Retailing,* 78(1), pp. 51-60.

Michael Baker, S. H., 2018. *The Marketing Book.* Seventh ed. London: Routledge.

Michael, G., 1971. Computer Simulation Model For Forecasting Catalog Sales, Journal Of Marketing Research. *Journal Of Marketing Research,* 8(2), pp. 224-232.

Millan, A. & Boada, A., 2010. *Predicción de la demanda de productos en empresas de venta directa – aplicación de regresión múltiple y series temporales en la psicología del consumo.* Caracas, II Congreso Venezolano de Psicología, I Congreso Venezolano de Psicología Positiva y I Congreso Venezolano de Ciencias de la Educación.

Millward, S., 2016. *Asia's ecommerce spending to hit record $1 trillion this year – but most of that is China.* [Online]
Available at: https://www.techinasia.com/asia-ecommerce-spending-1-trillion-dollars-2016
[Accessed 05 October 2018].

Noble, G. & Oconnor, S., 1986. Attitudes Toward Technology As Predictors Of Online Catalog Usage. *College & Research Libraries,* 47(6), pp. 605-610.

Ongallo, C., 2007. *El Libro de la Venta Directa.* First Edition ed. Madrid (España): Diaz de Santos.

Ord, K. & Fildes, R., 2013. *Principles of Business Forecasting.* International ed. Ohio: South-Western, Cengage Learning.

Oriflame Sweden, 2018. *Catalogue Oriflame UK 12-2018 (17/08 - 06/09).* [Online]
Available at: https://uk.oriflame.com/products/digital-catalogue-current?pageNumber=1&catalogue=2018012
[Accessed 22 08 2018].

PMR, 2012. *Grocery retail in Central Europe.* 1st ed. Krakow: 2012.

Pride, W. M. & Ferrell, O., 2012. *Marketing.* 16th ed. Mason, OH: South-Western Cengage Learning.

Puccinelli, N. M. et al., 2009. Customer Experience Management in Retailing: Understanding the Buying Process. *Journal of Retailing,* 85(1), pp. 15-30.

Ripley, B., 1996. *Pattern Recognition and Neural Networks.* First ed. Melbourne: Cambridge University Press.

Rodriguez-Calderon, C. E., 2017. *Metodología basada en modelos econométricos para predicción de demanda en una industria cosmética.* [Online] Available at: http://cybertesis.uni.edu.pe/handle/uni/3458?mode=full

Sovilj, D. et al., 2010. Neurocomputing OPELM and OPKNN in long-term prediction of time series using projected input data. *Neurocomputing,* Volume 73, p. 1976–1986.

Stone, C. J., 1977. Consistent nonparametric regression. *The annals of statistics,* Volume 5, pp. 595-620.

Tashman, L. J., 2000. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting,* 16(4), pp. 437-450.

The Telegraph, 2014. *Shop Direct to launch Very Exclusive website as it accelerates luxury strategy.* [Online] Available at: Shop Direct to launch Very Exclusive website as it accelerates luxury strategy
[Accessed 05 October 2018].

United Nations Statistics Division, 2013. *UN National Accounts Main Aggregates Database.* [Online] Available at: https://unstats.un.org/unsd/snaama/Introduction.asp [Accessed 16 May 2014].

Wahid, F. & Kim, D., 2016. A Prediction Approach for Demand Analysis of Energy Consumption Using K-Nearest Neighbor in Residential Buildings. *International Journal of Smart Home,* pp. 97-108.

WFDSA, 2016. *What is Direct Selling?.* [Online] Available at: https://wfdsa.org/about-direct-selling/ [Accessed 2018 08 19].

Wu, C. L., Chau, K. W. & Li, Y. S., 2009. Predicting monthly streamflow using data-driven models coupled with data-preprocessing techniques. *Water Resources Research,* 45(W08432), pp. 1-23.

Yakowitz, S., 1987. NEAREST-NEIGHBOUR METHODS FOR TIME SERIES ANALYSIS. *Journal of time series analysis,* Volume 8, pp. 235-247.

Zhang, B. et al., 2014. United States of America, Patent No. US 2014/0039979 A1 .

7 Appendices

Annex 1: additional charts for exploratory analysis

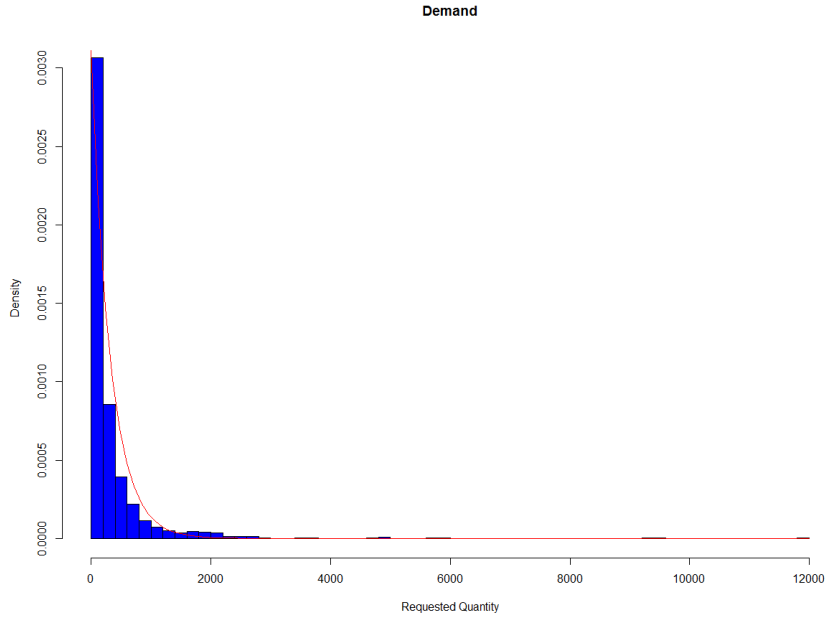


Figure 18: density for the requested quantity (exponential curve with λ=1/121)

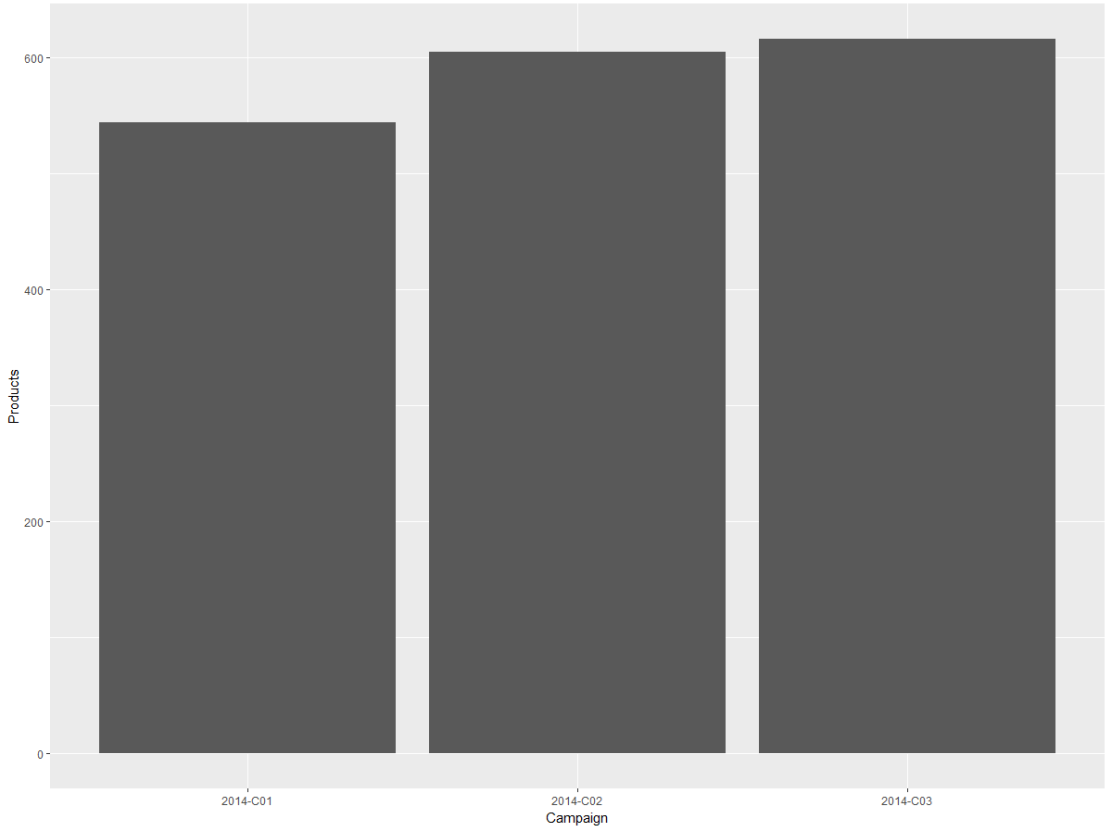


Figure 19: products per campaign

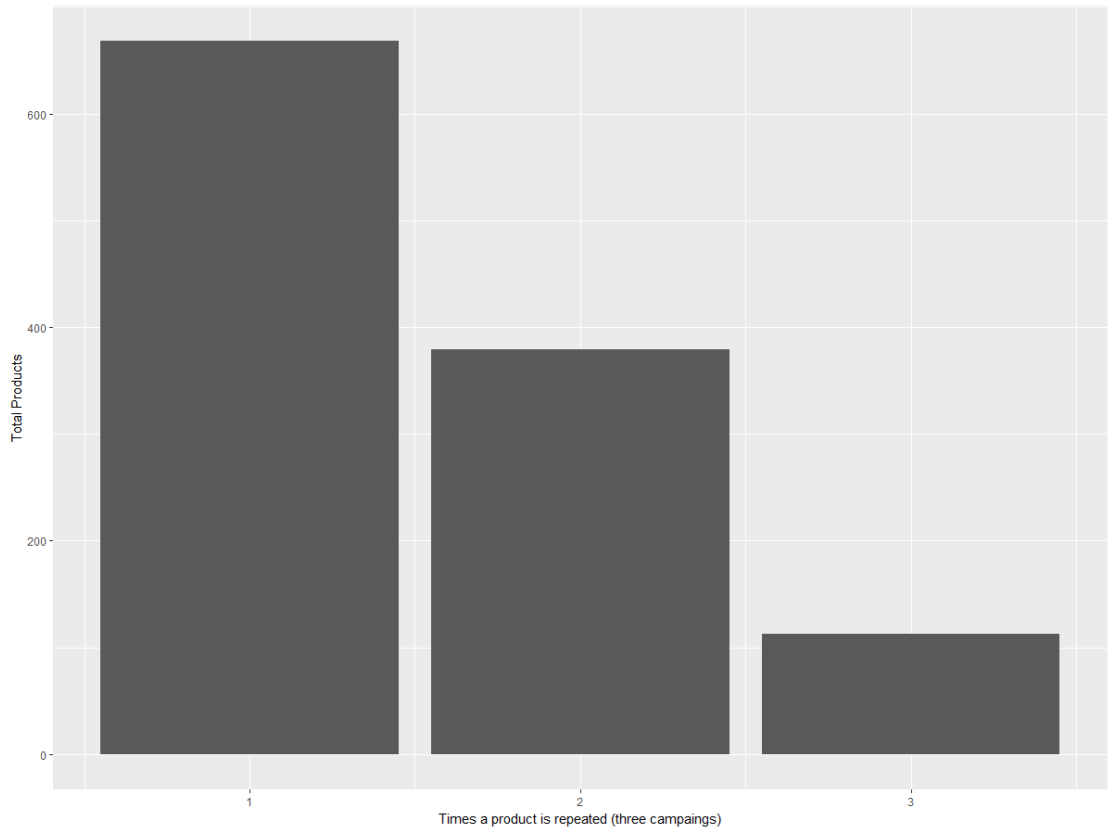


Figure 20: times a product is repeated in the three campaigns

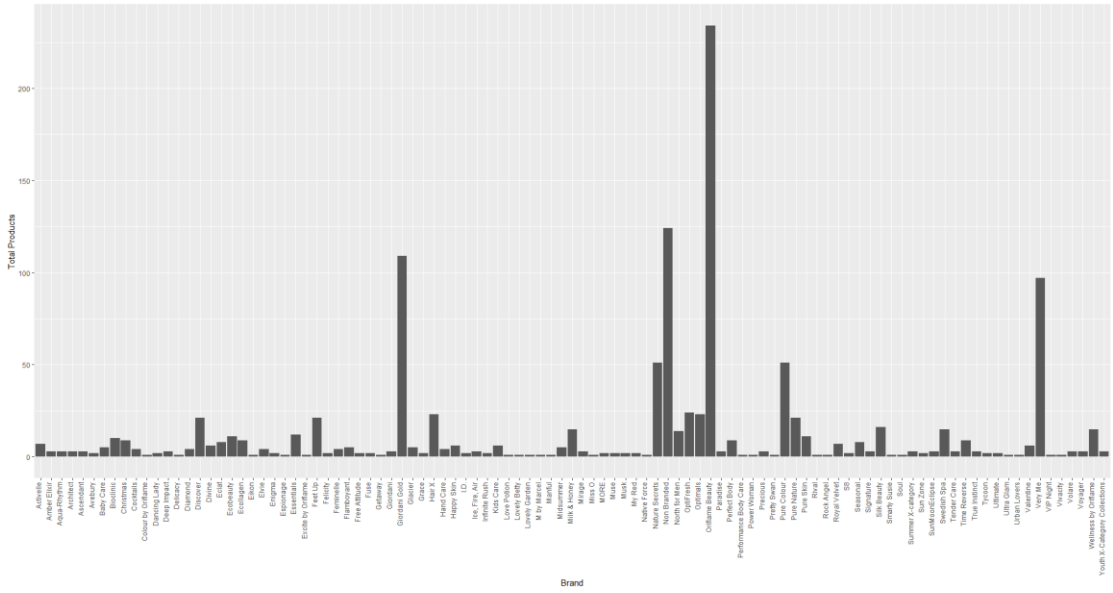


Figure 21: products per brand

Table 9: Cross-table between times a product is repeated and its status (including count, chi-square contribution, row per cent, column per cent and total per cent)

	Status					
Times repeated	**D**	**L**	**N**	**X**	**Z**	**Row Total**
1	131	173	335	3	26	**668**
	3.527	22.419	4.479	0.937	8.122	
	19.61%	25.90%	50.15%	0.45%	3.89%	**57.59%**
	48.88%	82.38%	51.30%	100.00%	100.00%	
	11.29%	14.91%	28.88%	0.26%	2.24%	
2	104	37	238	0	0	**379**
	3.086	14.565	2.848	0.98	8.495	
	27.44%	9.76%	62.80%	0.00%	0.00%	**32.67%**
	38.81%	17.62%	36.45%	0.00%	0.00%	
	8.97%	3.19%	20.52%	0.00%	0.00%	
3	33	0	80	0	0	**113**
	1.82	20.457	4.222	0.292	2.533	
	29.20%	0.00%	70.80%	0.00%	0.00%	**9.74%**
	12.31%	0.00%	12.25%	0.00%	0.00%	
	2.85%	0.00%	6.90%	0.00%	0.00%	
Column Total	**268**	**210**	**653**	**3**	**26**	**1160**
	23.10%	**18.10%**	**56.29%**	**0.26%**	**2.24%**	

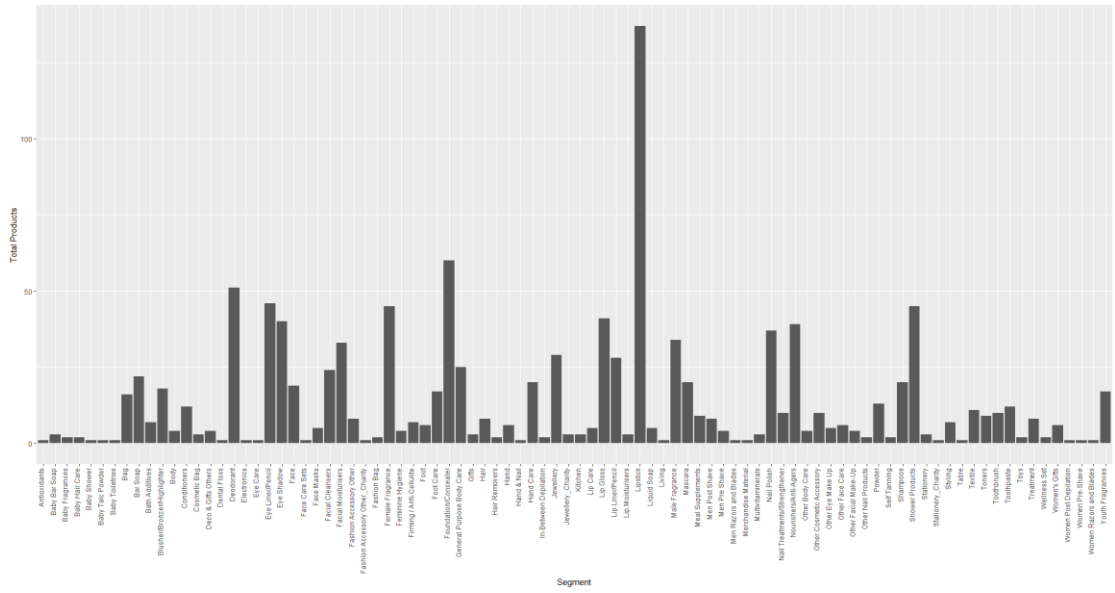


Figure 22: total products per segment

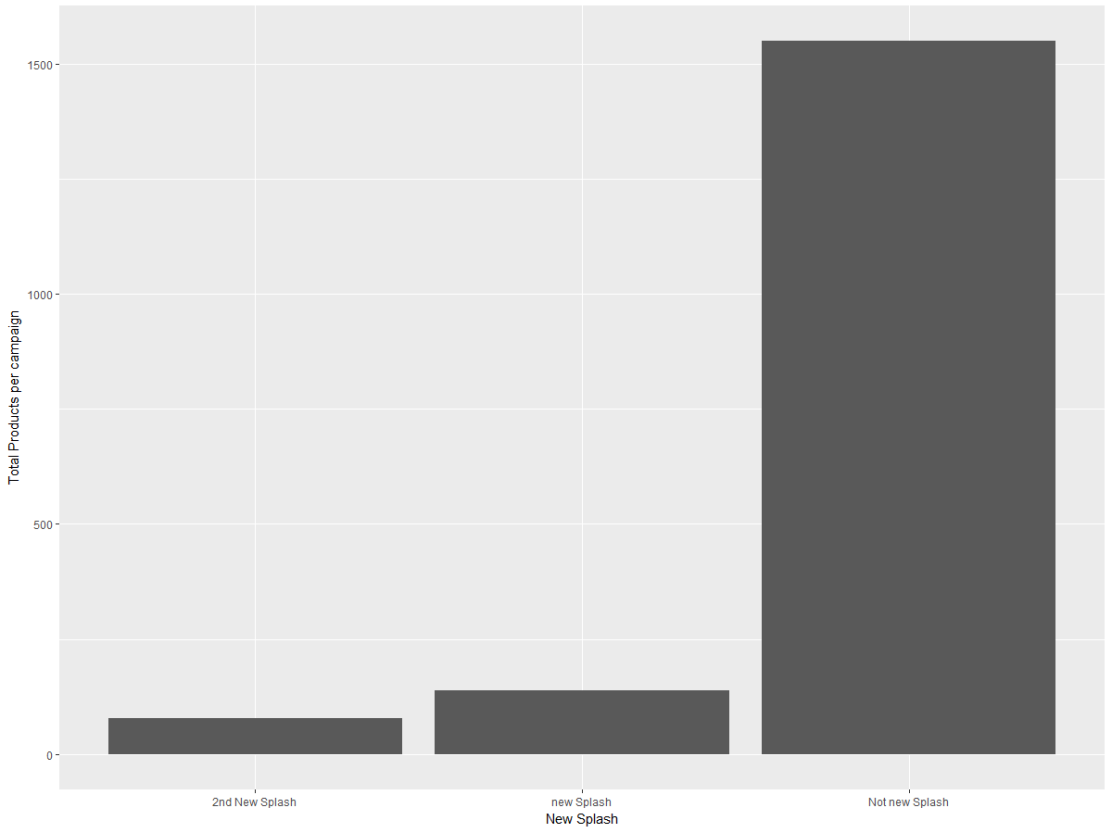


Figure 23: New Splash

Table 10: products located by section and campaign

	Status					
Times repeated	**Back Cover**	**Category sections**	**Ending Section**	**Middle Spread**	**Platform**	**Row Total**
2014-C01	4	440	17	22	61	544
	0.74%	80.88%	3.13%	4.04%	11.21%	30.82%
	50.00%	30.62%	20.99%	50.00%	31.28%	
	0.23%	24.93%	0.96%	1.25%	3.46%	
2014-C02	2	500	21	17	65	605
	0.33%	82.65%	3.47%	2.81%	10.74%	34.28%
	25.00%	34.80%	25.93%	38.64%	33.33%	
	0.11%	28.33%	1.19%	0.96%	3.68%	
2014-C03	2	497	43	5	69	616
	0.33%	80.68%	6.98%	0.81%	11.20%	34.90%
	25.00%	34.59%	53.09%	11.36%	35.39%	
	0.11%	28.16%	2.44%	0.28%	3.91%	
Column Total	8	1437	81	44	195	1765
	0.45%	81.42%	4.59%	2.49%	11.05%	

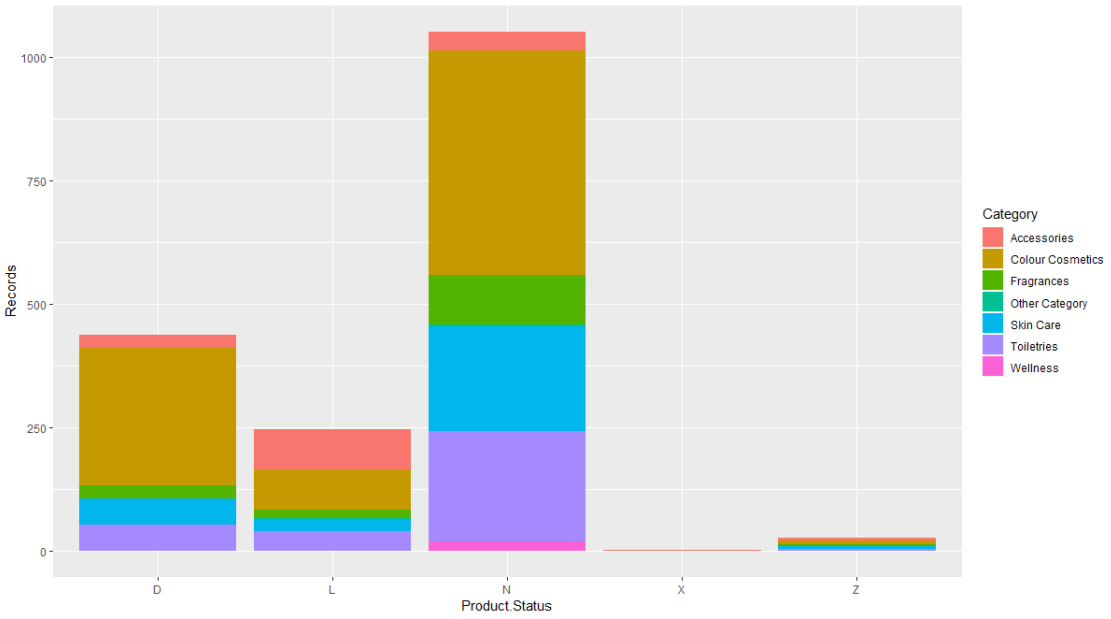


Figure 24: records by product status and category

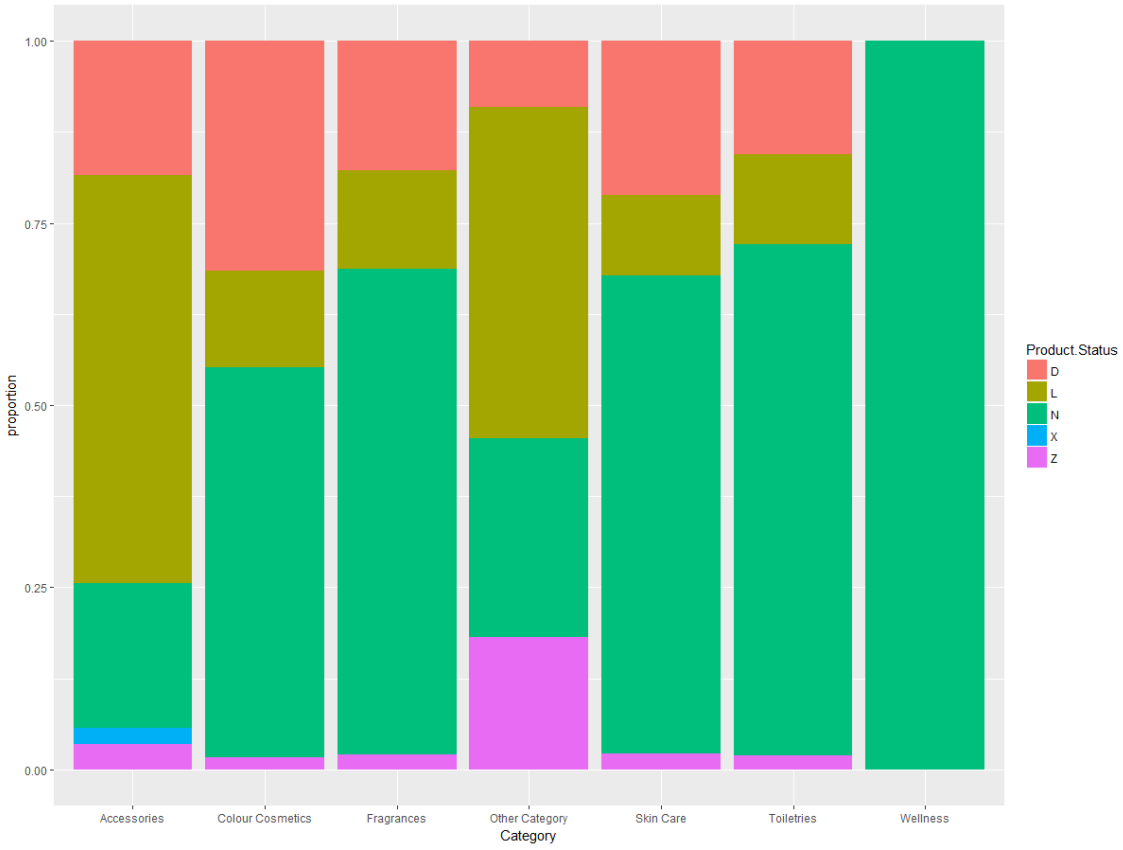


Figure 25: proportion of categories by status

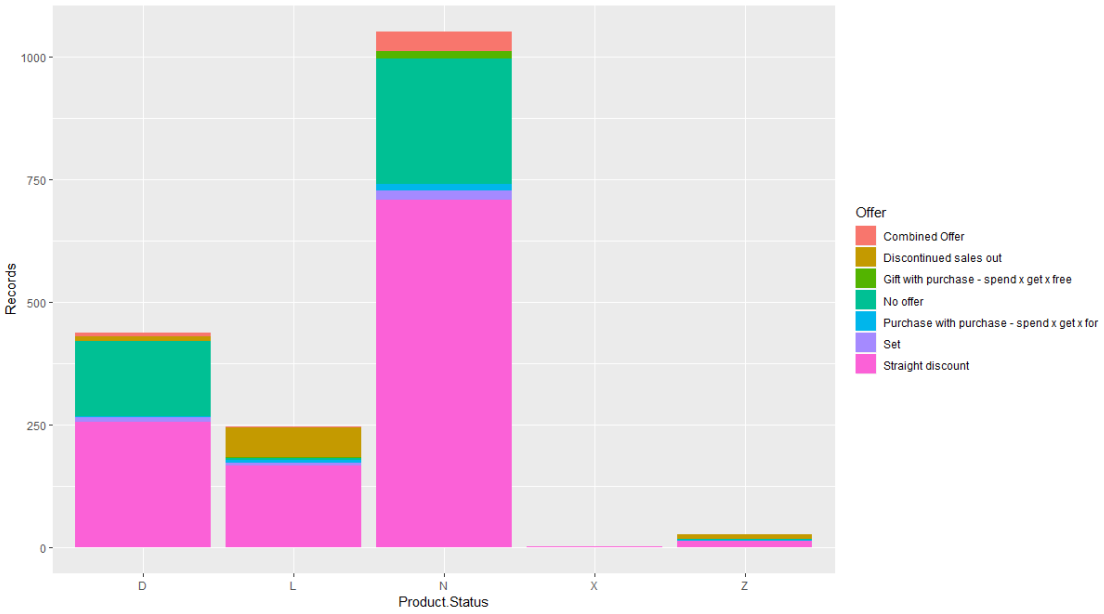


Figure 26: proportion of records by product status and offer applied

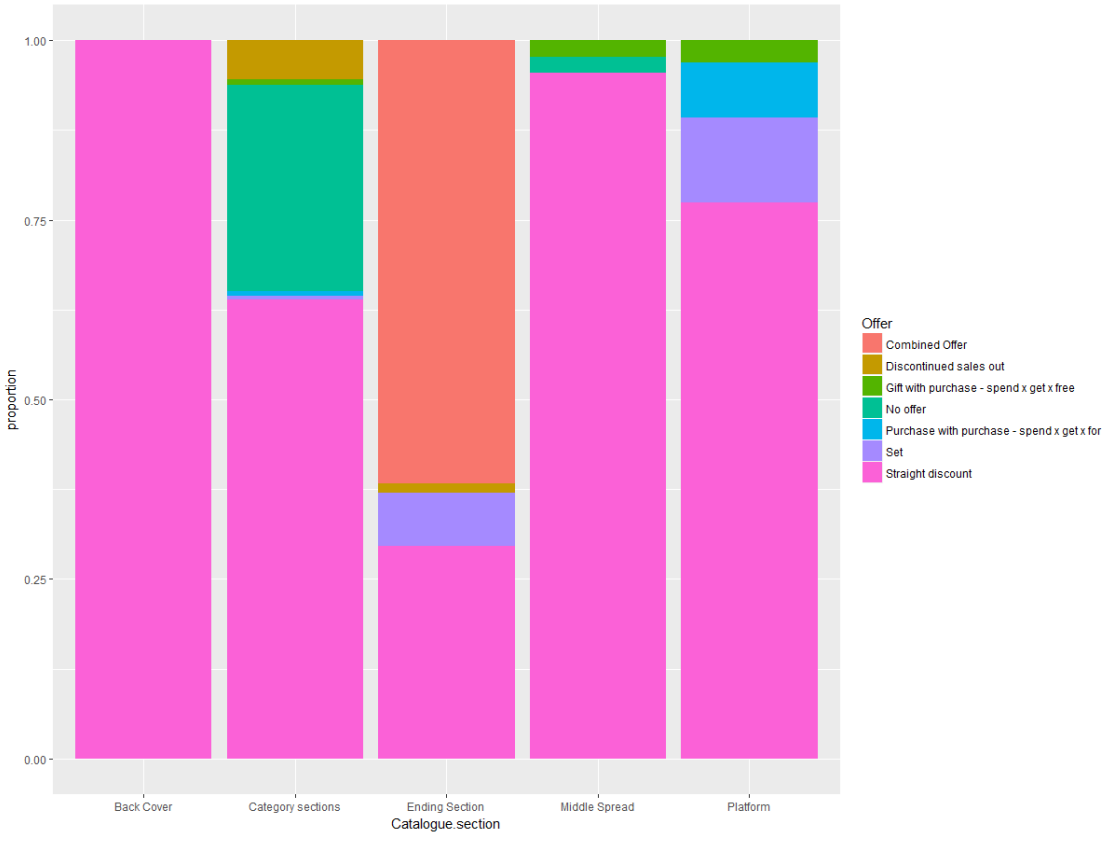


Figure 27: offer types applied over catalogue sections.

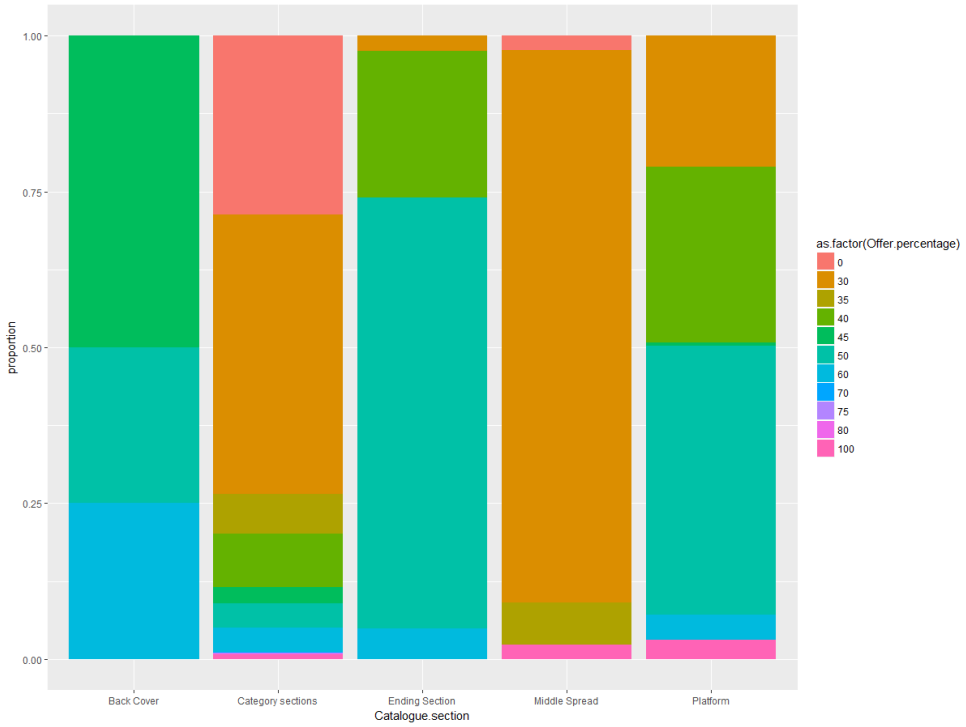


Figure 28: offer percentages applied across different catalogue sections

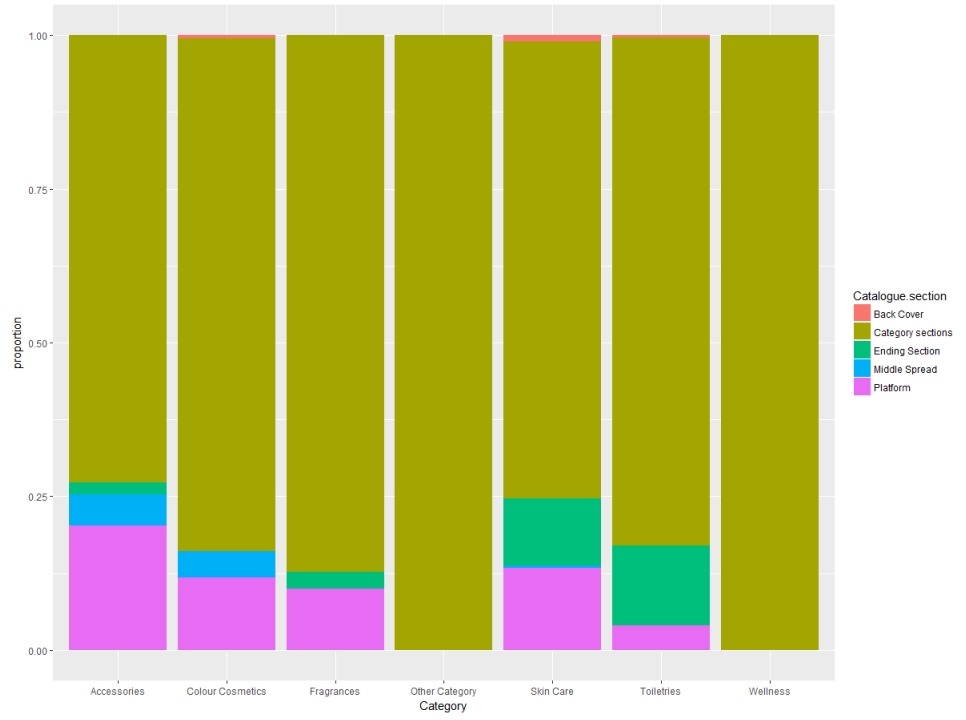


Figure 29: the proportion of products by category and section

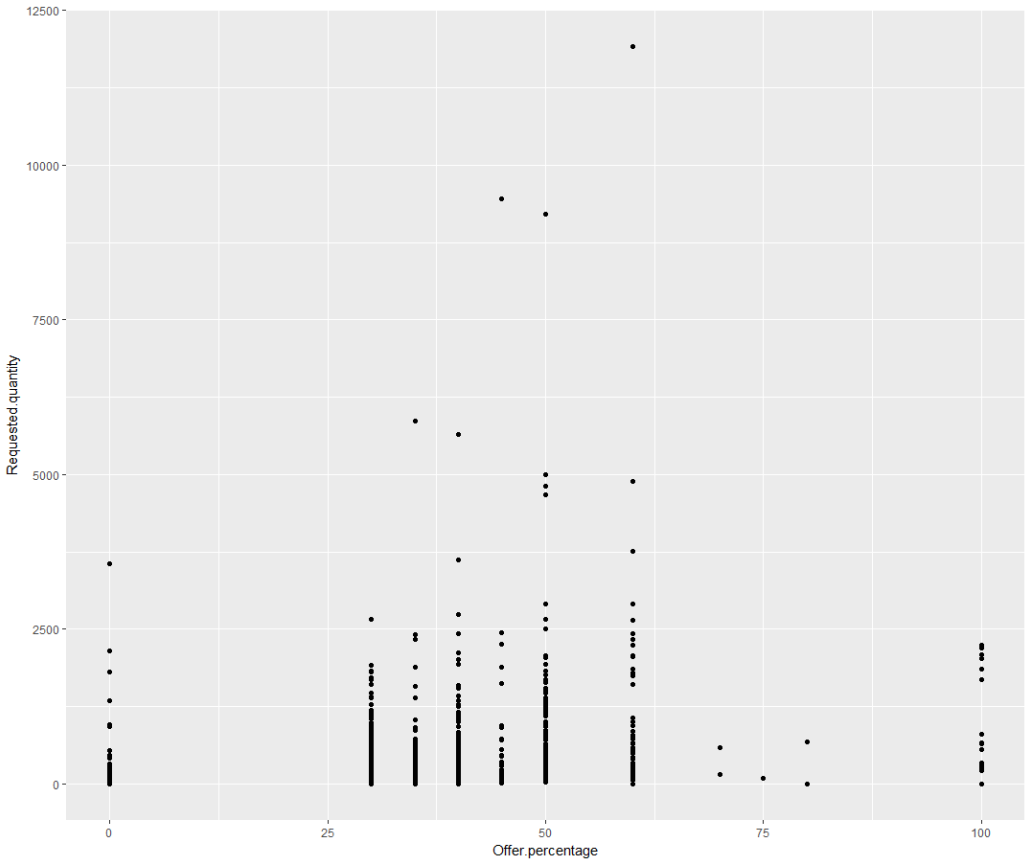


Figure 30: scatterplot of Offer percentage vs. Requested quantity

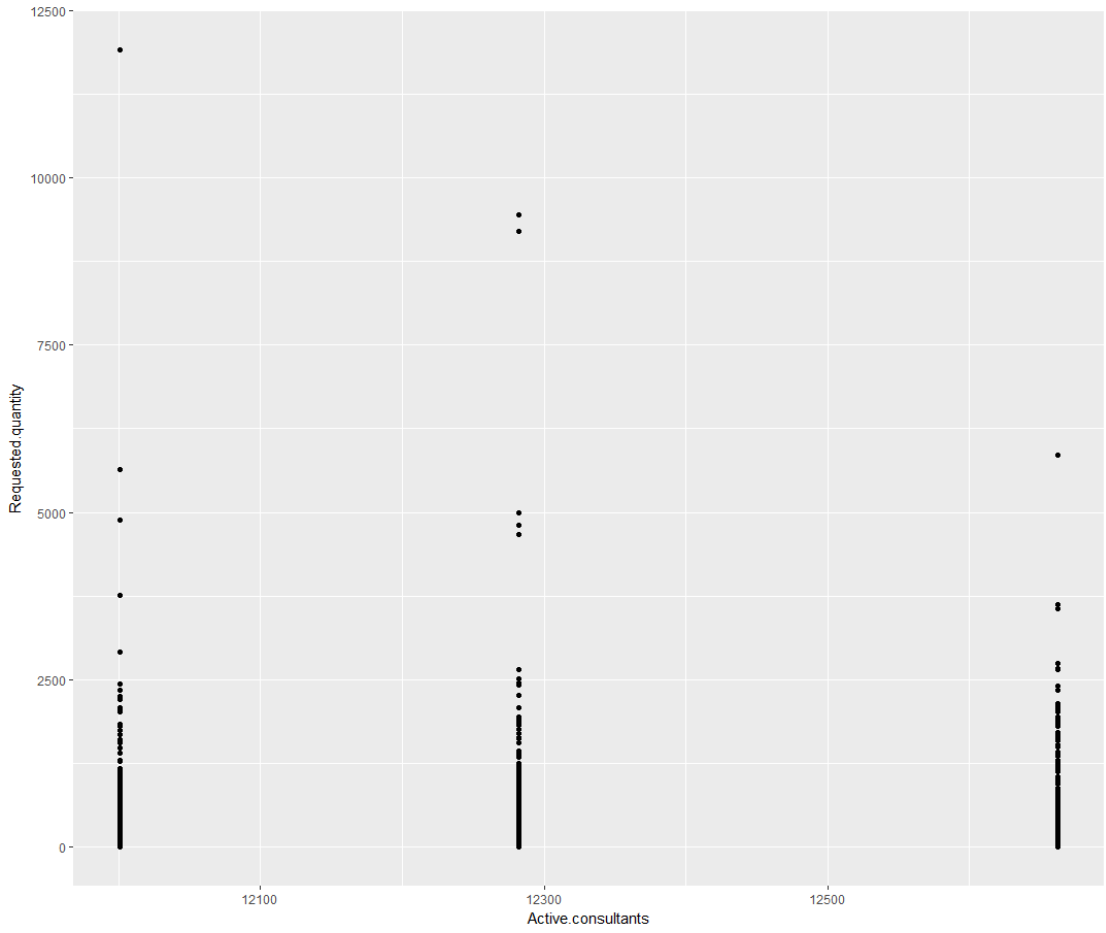


Figure 31: scatterplot of number of Active consultants vs Requested quantity

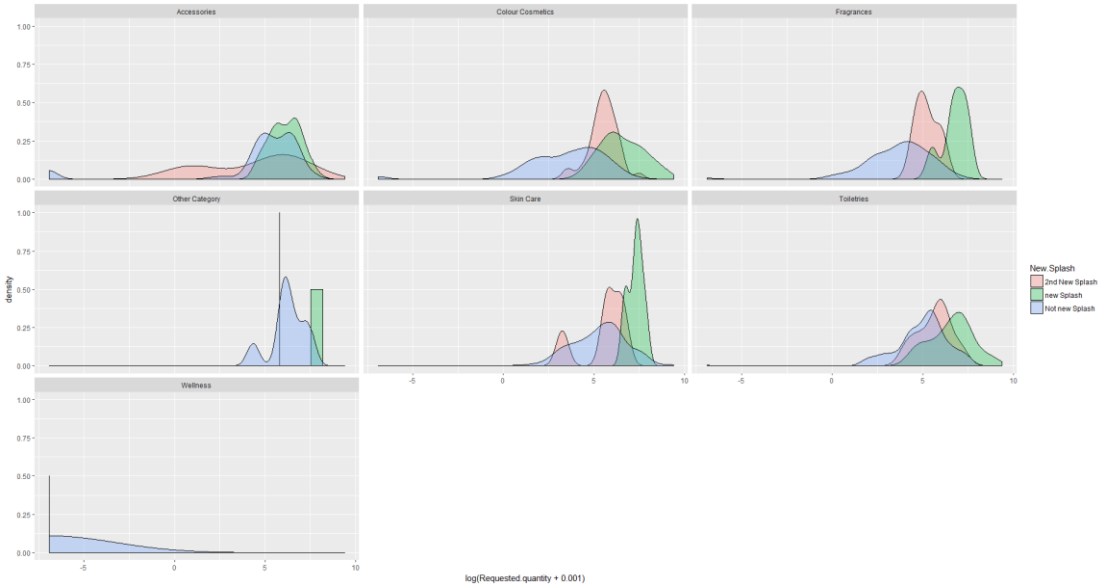


Figure 32: demand density by category and New Splash type.

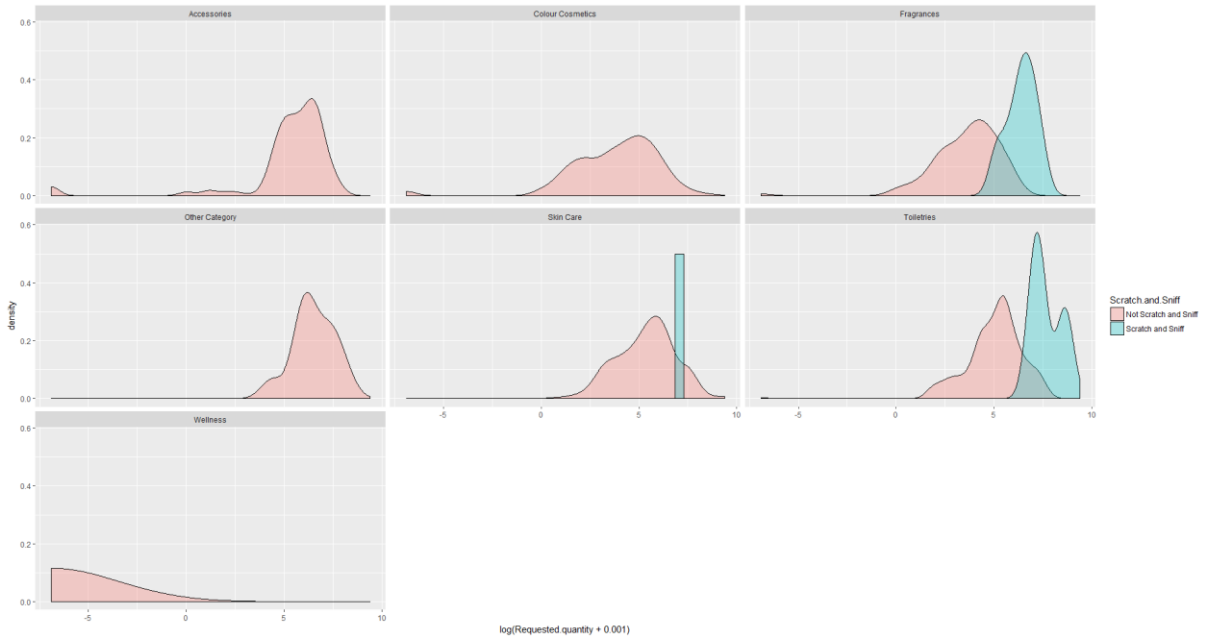


Figure 33: demand density by category and w/o Scratch and Sniff

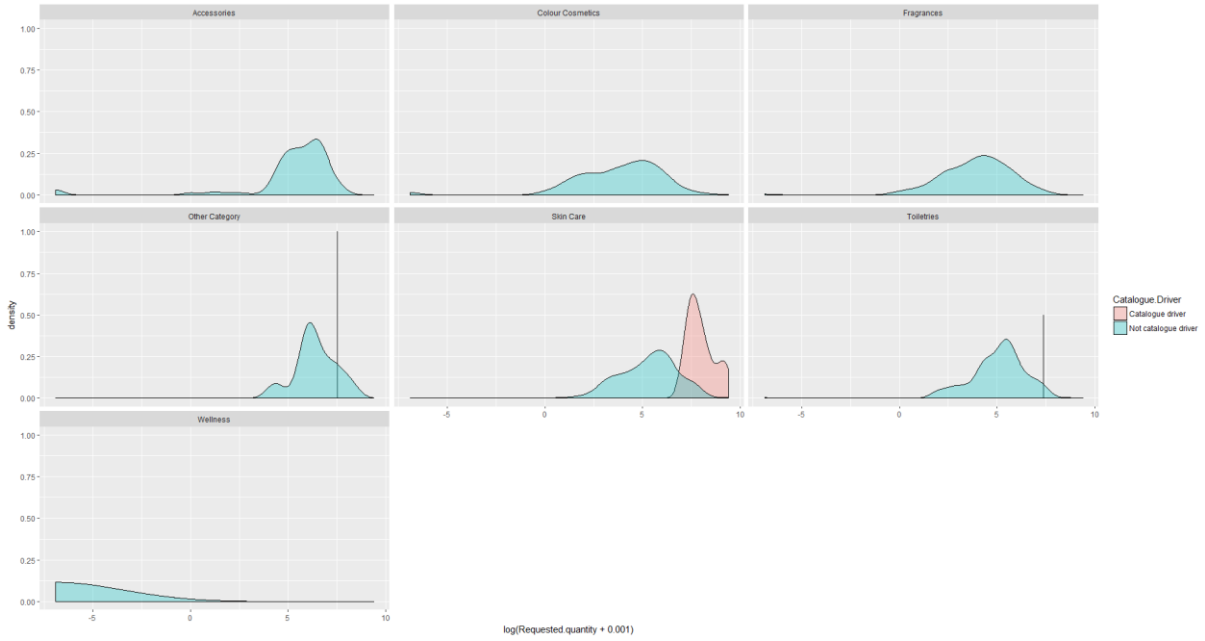


Figure 34: demand density by the category w/o catalogue driver.

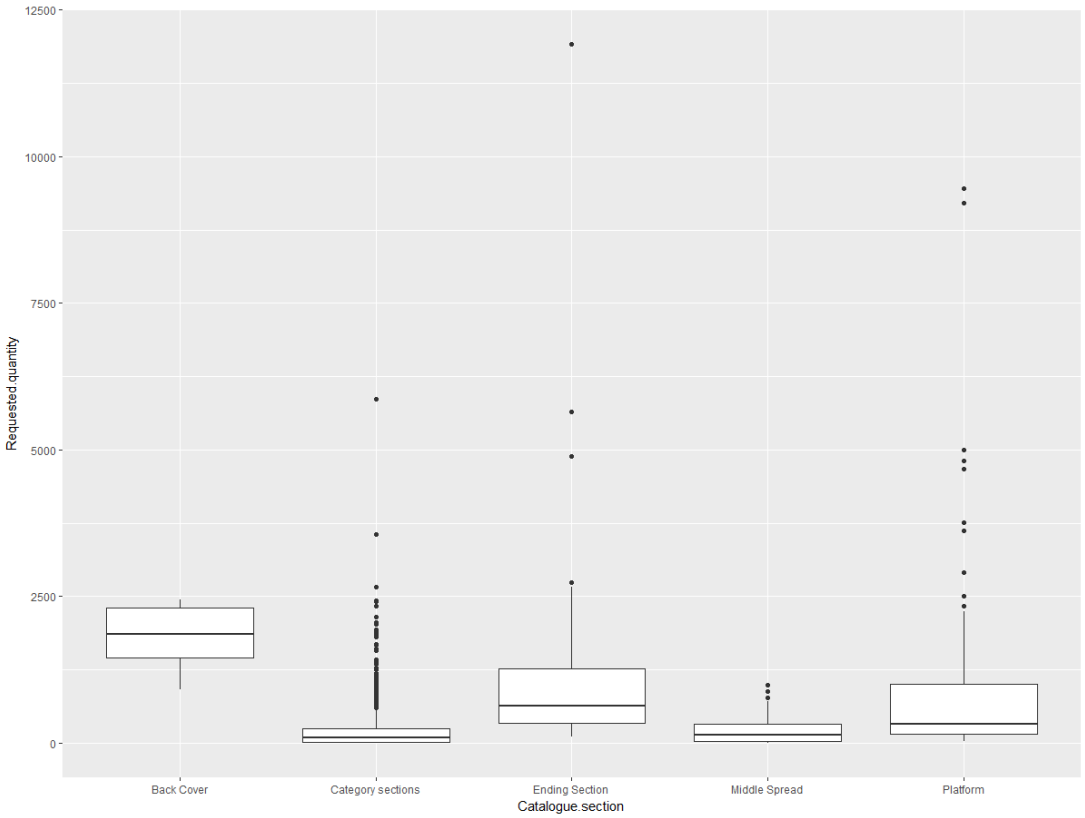


Figure 35: Box plot of demand by catalogue section

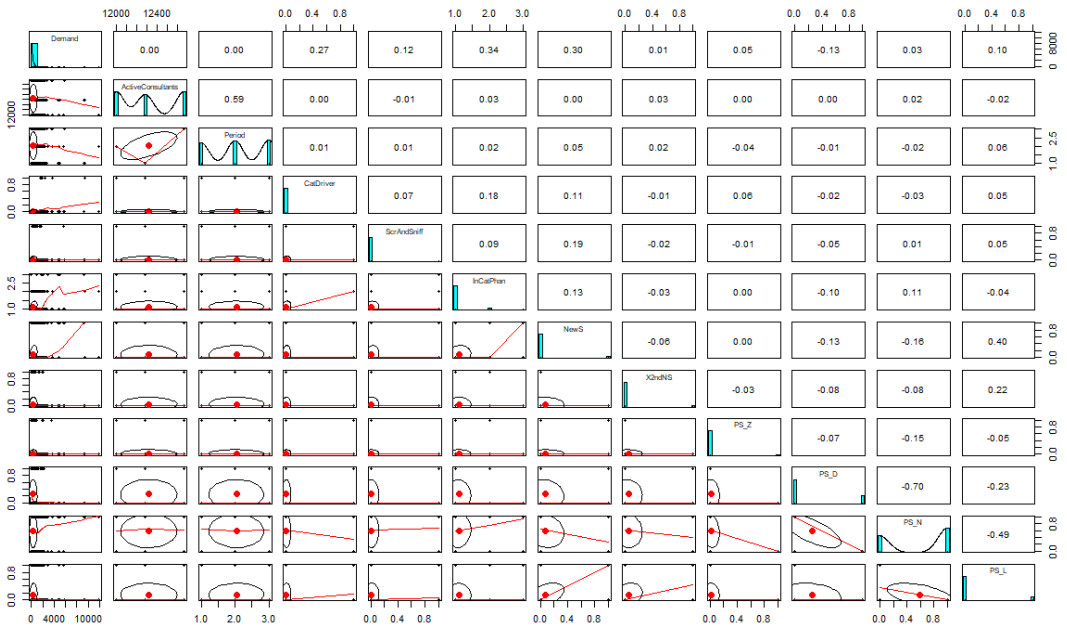


Figure 36: Correlation between variables (including product status)

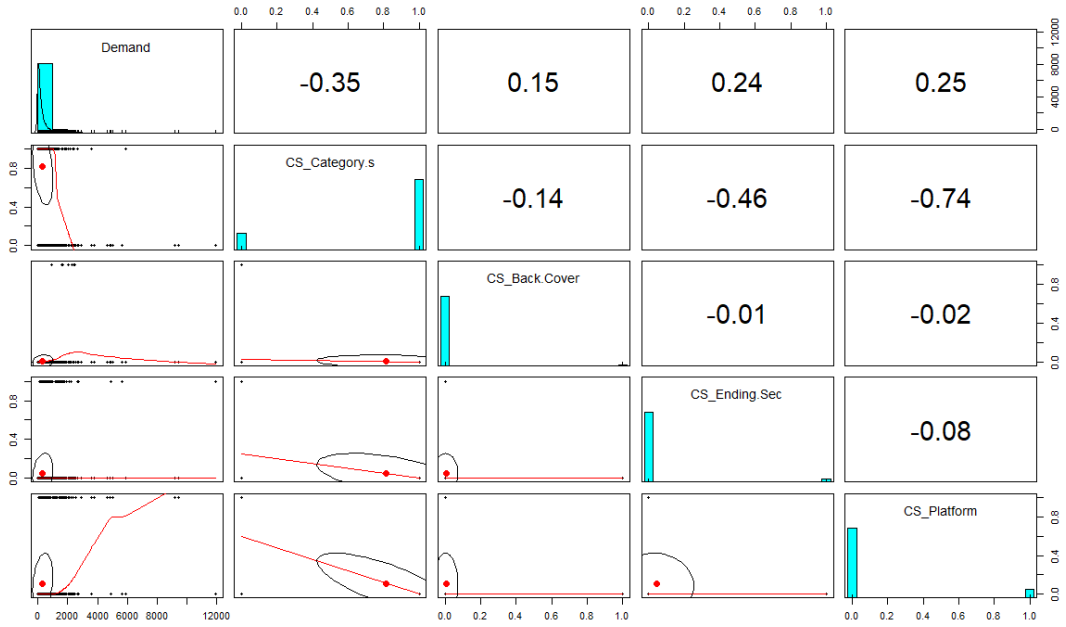


Figure 37: Correlation between demand and category sections

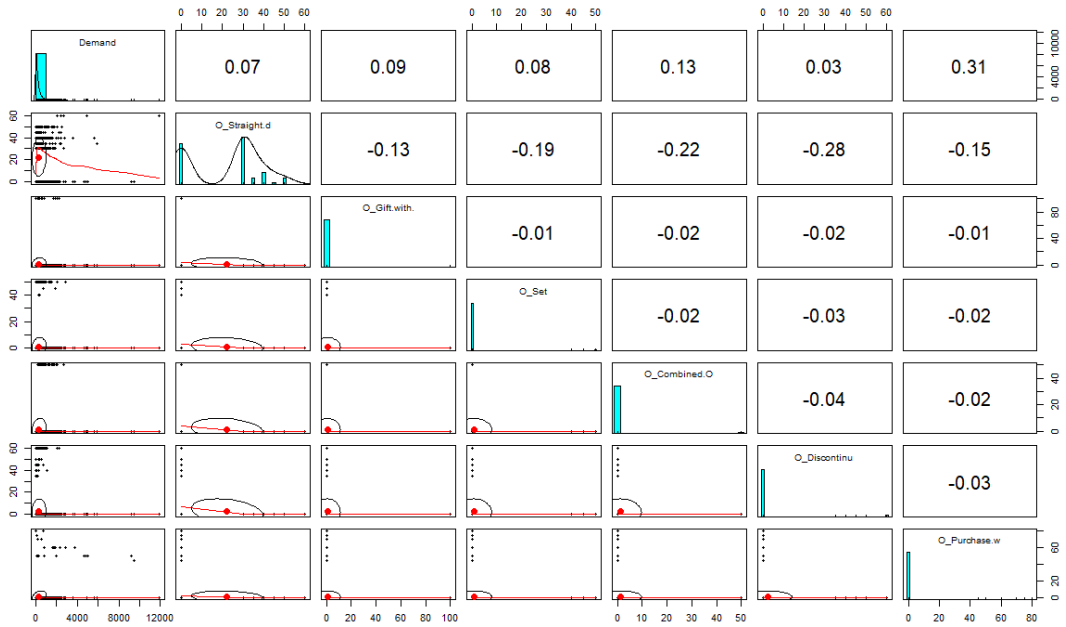


Figure 38: Correlation between demand and offer percentage by offer

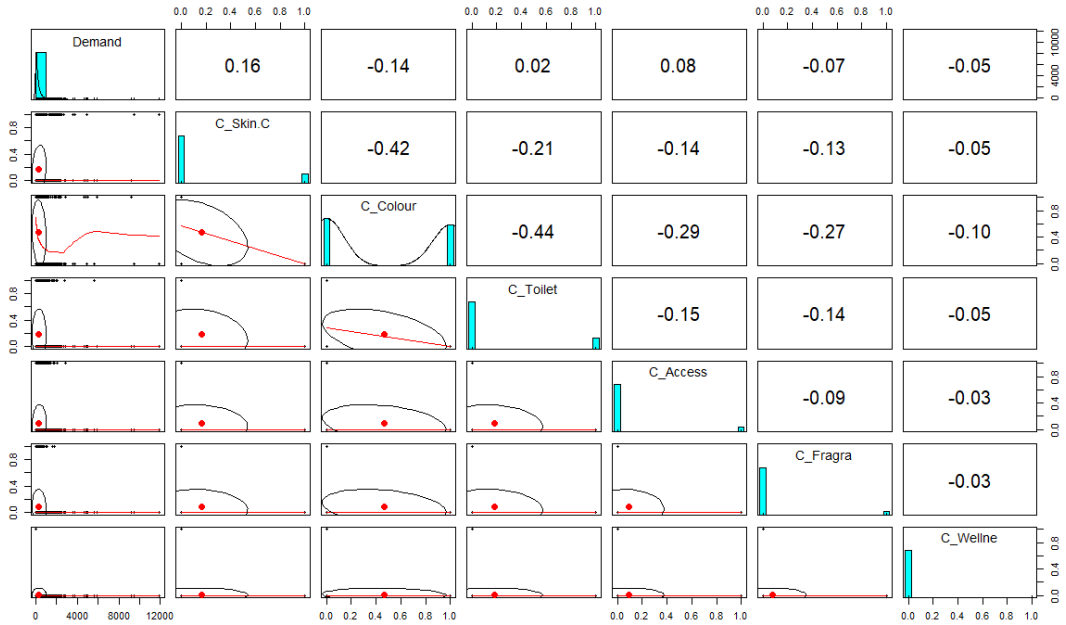


Figure 39: Correlation between demand and category

Annex 2: Encoded dataset (R output)

```
'data.frame':  1765 obs. of  28 variables:
 $ Demand           : int  3562 1889 907 2737 2199 1892 1607 2093 139
2 2017 ...
 $ ActiveConsultants: int  12662 12282 12282 12662 12001 12662 12001
12662 12001 12001 ...
 $ Period           : int  3 1 1 3 2 3 2 3 2 2 ...
 $ CatDriver        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ ScrAndSniff      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ InCatPhan        : int  1 2 1 1 2 2 1 2 1 1 ...
 $ CS_Category.s    : int  1 1 0 0 0 1 1 0 1 0 ...
 $ CS_Back.Cover    : int  0 0 1 0 0 0 0 0 0 0 ...
 $ CS_Ending.Sec    : int  0 0 0 1 0 0 0 0 0 0 ...
 $ CS_Platform      : int  0 0 0 0 1 0 0 1 0 1 ...
 $ O_Straight.d     : int  0 35 45 40 0 0 30 0 35 40 ...
 $ O_Gift.with.     : int  0 0 0 0 100 0 0 100 0 0 ...
 $ O_Set            : int  0 0 0 0 0 45 0 0 0 0 ...
 $ O_Combined.O     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ O_Discontinu     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ O_Purchase.w     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ NewS             : int  1 0 0 0 0 0 0 0 0 0 ...
 $ X2ndNS           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ PS_Z             : int  1 0 0 0 0 0 0 0 0 0 ...
 $ PS_D             : int  0 1 0 0 0 0 1 0 0 0 ...
 $ PS_N             : int  0 0 1 1 1 1 0 1 1 1 ...
 $ PS_L             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ C_Skin.C         : int  0 1 0 0 1 1 0 1 0 1 ...
 $ C_Colour         : int  0 0 1 0 0 0 1 0 0 0 ...
 $ C_Toilet         : int  0 0 0 1 0 0 0 0 1 0 ...
 $ C_Access         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ C_Fragra         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ C_Wellne         : int  0 0 0 0 0 0 0 0 0 0 ...
```

Annex 3: Linear regression models with original demand (R output)

All variables

```
Call:
lm(formula = Demand ~ ., data = trd)

Residuals:
    Min      1Q  Median      3Q     Max
-2147.6  -133.3   -10.7    86.0  8933.3

Coefficients: (1 not defined because of singularities)
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.492e+03  1.492e+03  -1.000  0.31764
ActiveConsultants  3.277e-02  1.193e-01   0.275  0.78357
Period                    NA         NA      NA       NA
CatDriver          2.285e+03  2.905e+02   7.866 8.59e-15 ***
ScrAndSniff        4.544e+02  1.630e+02   2.787  0.00540 **
InCatPhan          1.219e+03  1.115e+02  10.932  < 2e-16 ***
CS_Category.s     -3.901e+01  9.132e+01  -0.427  0.66932
CS_Back.Cover      1.417e+03  2.432e+02   5.825 7.46e-09 ***
CS_Ending.Sec      1.223e+03  1.524e+02   8.029 2.47e-15 ***
CS_Platform        3.293e+02  1.071e+02   3.075  0.00215 **
O_Straight.d       4.920e+00  1.198e+00   4.108 4.28e-05 ***
O_Gift.with.      -4.671e+00  1.861e+00  -2.510  0.01222 *
O_Set             -2.226e+01  3.481e+00  -6.396 2.34e-10 ***
O_Combined.O      -2.285e+01  4.219e+00  -5.416 7.45e-08 ***
O_Discontinu       8.335e+00  1.885e+00   4.420 1.08e-05 ***
O_Purchase.w      -2.518e+00  3.441e+00  -0.732  0.46448
NewS               4.279e+02  8.196e+01   5.221 2.12e-07 ***
X2ndNS             1.246e+02  8.732e+01   1.427  0.15373
PS_Z              -4.429e+02  3.427e+02  -1.292  0.19653
PS_D              -4.409e+02  3.220e+02  -1.369  0.17116
PS_N              -4.072e+02  3.215e+02  -1.267  0.20553
PS_L              -4.367e+02  3.237e+02  -1.349  0.17749
C_Skin.C           5.280e+02  2.627e+02   2.010  0.04467 *
C_Colour           3.294e+02  2.618e+02   1.258  0.20852
C_Toilet           2.983e+02  2.633e+02   1.133  0.25748
C_Access           2.198e+02  2.657e+02   0.827  0.40817
C_Fragra           1.349e+02  2.657e+02   0.508  0.61184
C_Wellne          -1.640e+03  3.527e+02  -4.649 3.73e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 541.7 on 1122 degrees of freedom
Multiple R-squared:  0.4763,   Adjusted R-squared:  0.4642
F-statistic: 39.25 on 26 and 1122 DF,  p-value: < 2.2e-16
```

Stepwise - Backward:

```
Call:
lm(formula = Demand ~ CatDriver + ScrAndSniff + InCatPhan + CS_Back.C
over +
    CS_Ending.Sec + CS_Platform + O_Straight.d + O_Gift.with. +
    O_Set + O_Combined.O + O_Discontinu + NewS + X2ndNS + PS_Z +
    PS_D + PS_N + PS_L + C_Skin.C + C_Colour + C_Toilet + C_Wellne,
    data = trd)

Residuals:
    Min      1Q  Median      3Q     Max
-2125.9  -130.7    -9.5    82.3  8974.9

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -872.469    326.880  -2.669 0.007715 **
CatDriver      2155.995    259.698   8.302 2.90e-16 ***
ScrAndSniff     419.023    157.112   2.667 0.007762 **
InCatPhan      1168.110     82.677  14.129  < 2e-16 ***
CS_Back.Cover  1447.614    227.320   6.368 2.78e-10 ***
CS_Ending.Sec  1259.260    122.769  10.257  < 2e-16 ***
CS_Platform     372.475     63.417   5.873 5.61e-09 ***
O_Straight.d      5.101      1.172   4.351 1.48e-05 ***
O_Gift.with.     -4.136      1.707  -2.424 0.015527 *
O_Set           -21.230      2.965  -7.159 1.46e-12 ***
O_Combined.O    -21.570      3.902  -5.528 4.01e-08 ***
O_Discontinu      8.023      1.819   4.411 1.13e-05 ***
NewS            437.628     79.512   5.504 4.60e-08 ***
X2ndNS          133.126     86.650   1.536 0.124730
PS_Z           -491.038    338.400  -1.451 0.147042
PS_D           -489.593    317.122  -1.544 0.122903
PS_N           -457.265    315.917  -1.447 0.148058
PS_L           -476.198    320.041  -1.488 0.137049
C_Skin.C        365.293     58.933   6.198 7.99e-10 ***
C_Colour        165.619     48.661   3.404 0.000689 ***
C_Toilet        132.147     57.791   2.287 0.022401 *
C_Wellne      -1723.206    206.833  -8.331 2.30e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 541 on 1127 degrees of freedom
Multiple R-squared:  0.4754,   Adjusted R-squared:  0.4656
F-statistic: 48.63 on 21 and 1127 DF,  p-value: < 2.2e-16
```

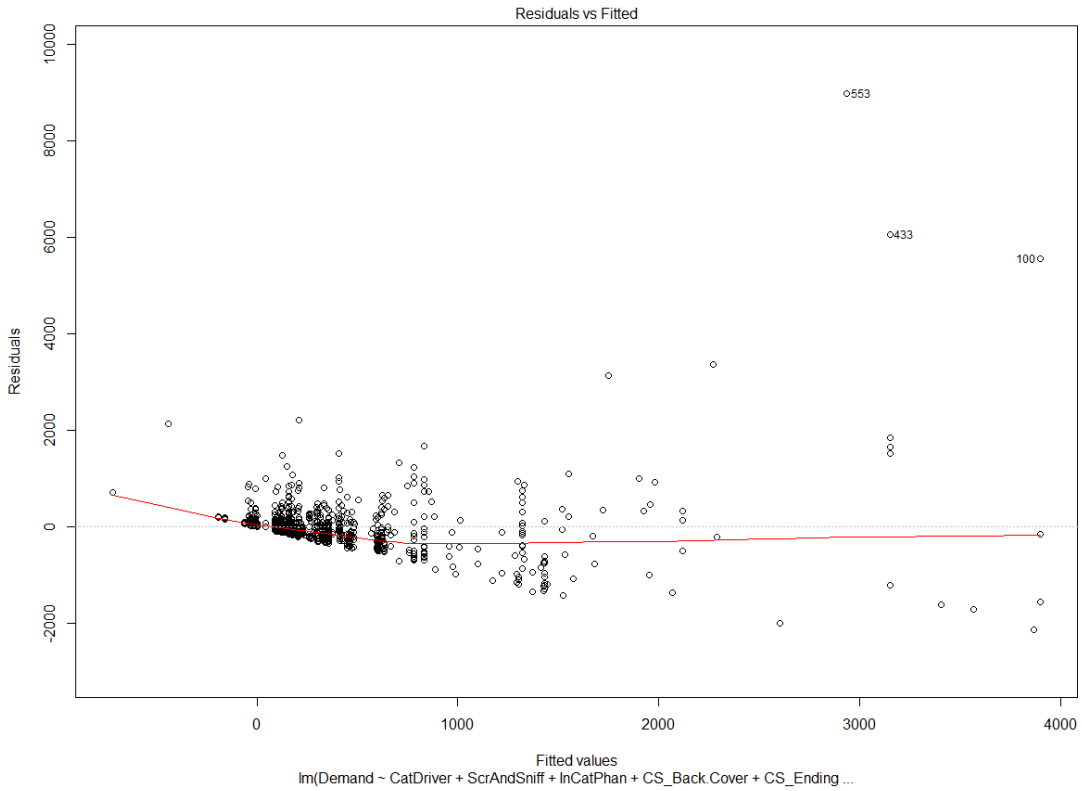
Stepwise - Forward:

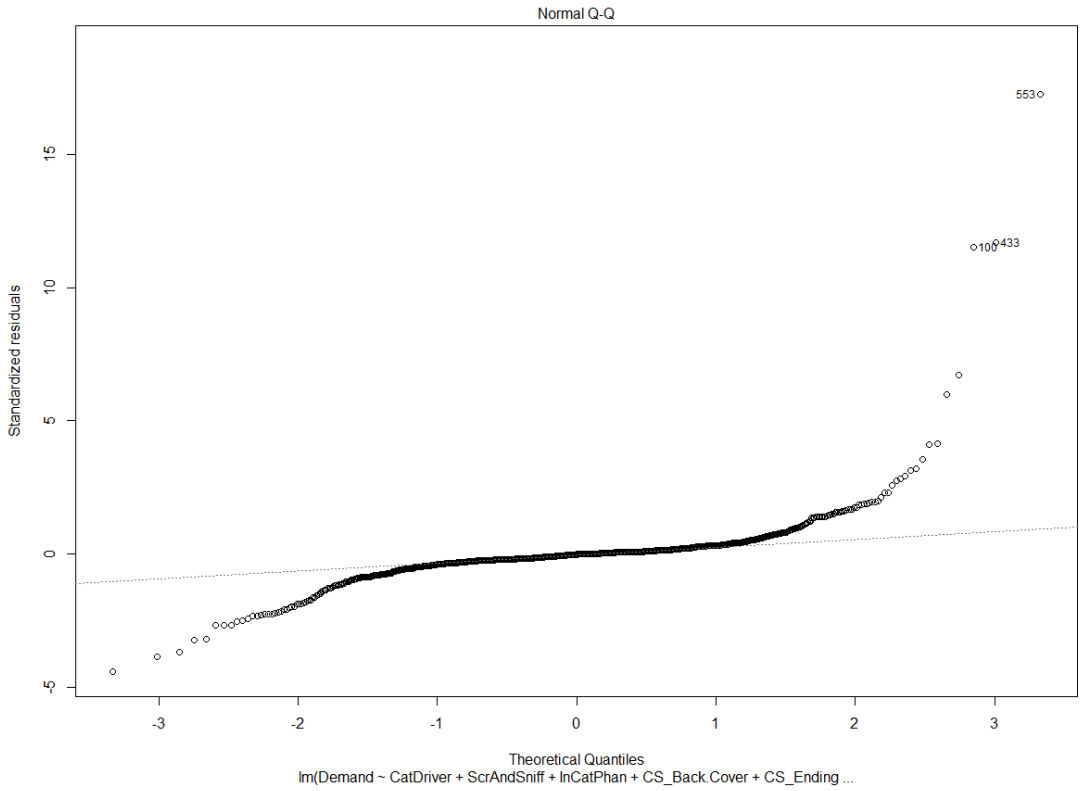
```
Call:
lm(formula =
    Demand ~ O_Straight.d + O_Gift.with. + O_Set + O_Combined.O +
    O_Discontinu + O_Purchase.w + CS_Category.s + InCatPhan +
    C_Wellne + CatDriver + CS_Ending.Sec + NewS + C_Skin.C +
    CS_Back.Cover + CS_Platform + ScrAndSniff + C_Fragra, data = trd)

Residuals:
    Min      1Q  Median      3Q     Max
-2114.0  -130.9    -5.9    73.1  8916.1

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1220.750    145.260  -8.404  < 2e-16 ***
O_Straight.d      5.040      1.147   4.396 1.21e-05 ***
O_Gift.with.     -4.972      1.839  -2.704  0.00696 **
O_Set           -22.895      3.411  -6.713 3.02e-11 ***
O_Combined.O    -23.121      4.152  -5.569 3.21e-08 ***
O_Discontinu      7.110      1.595   4.458 9.10e-06 ***
O_Purchase.w     -3.554      3.386  -1.049  0.29418
CS_Category.s   -51.395     89.398  -0.575  0.56548
InCatPhan      1255.872    109.342  11.486  < 2e-16 ***
C_Wellne      -1992.315    235.458  -8.461  < 2e-16 ***
CatDriver      2258.975    282.894   7.985 3.43e-15 ***
CS_Ending.Sec  1186.293    148.215   8.004 2.97e-15 ***
NewS            372.317     68.773   5.414 7.54e-08 ***
C_Skin.C        216.179     45.057   4.798 1.82e-06 ***
CS_Back.Cover  1443.839    241.140   5.988 2.86e-09 ***
CS_Platform     303.088    105.304   2.878  0.00407 **
ScrAndSniff     476.390    161.851   2.943  0.00331 **
C_Fragra       -169.281     60.351  -2.805  0.00512 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 541.4 on 1131 degrees of freedom
Multiple R-squared:  0.4726,   Adjusted R-squared:  0.4647
F-statistic: 59.63 on 17 and 1131 DF,  p-value: < 2.2e-16
```

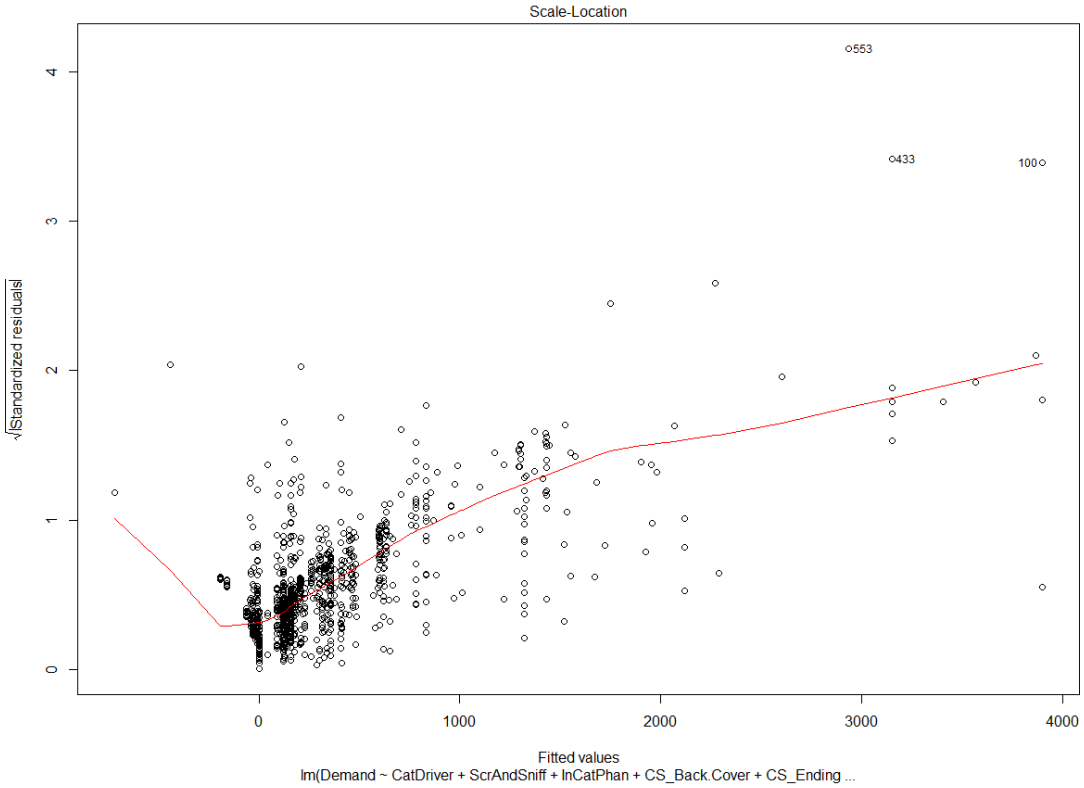
Stepwise – Both

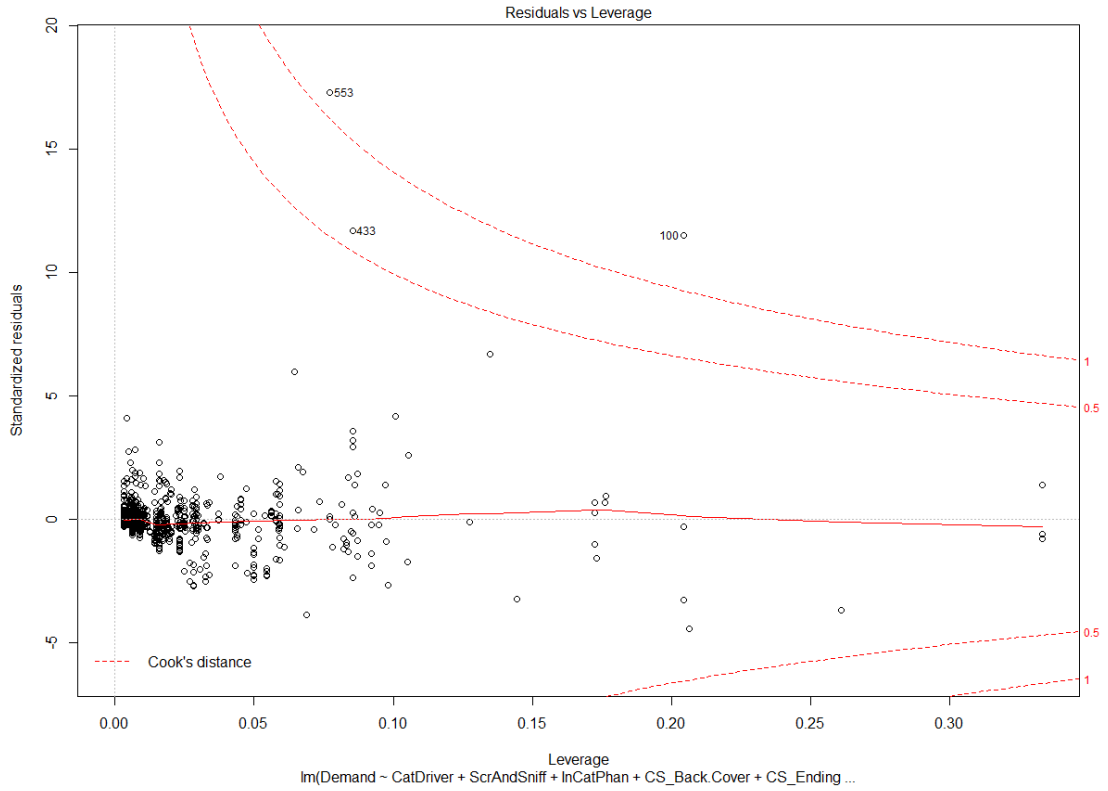
```
Call:
lm(formula =
Demand ~ CatDriver + ScrAndSniff + InCatPhan + CS_Back.Cover +
    CS_Ending.Sec + CS_Platform + O_Straight.d + O_Gift.with. +
    O_Set + O_Combined.O + O_Discontinu + NewS + X2ndNS + PS_Z +
    PS_D + PS_N + PS_L + C_Skin.C + C_Colour + C_Toilet + C_Wellne,
    data = trd)

Residuals:
    Min      1Q  Median      3Q     Max
-2125.9  -130.7    -9.5    82.3  8974.9

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -872.469    326.880  -2.669 0.007715 **
CatDriver      2155.995    259.698   8.302 2.90e-16 ***
ScrAndSniff     419.023    157.112   2.667 0.007762 **
InCatPhan      1168.110     82.677  14.129  < 2e-16 ***
CS_Back.Cover  1447.614    227.320   6.368 2.78e-10 ***
CS_Ending.Sec  1259.260    122.769  10.257  < 2e-16 ***
CS_Platform     372.475     63.417   5.873 5.61e-09 ***
O_Straight.d      5.101      1.172   4.351 1.48e-05 ***
O_Gift.with.     -4.136      1.707  -2.424 0.015527 *
O_Set           -21.230      2.965  -7.159 1.46e-12 ***
O_Combined.O    -21.570      3.902  -5.528 4.01e-08 ***
O_Discontinu      8.023      1.819   4.411 1.13e-05 ***
NewS            437.628     79.512   5.504 4.60e-08 ***
X2ndNS          133.126     86.650   1.536 0.124730
PS_Z           -491.038    338.400  -1.451 0.147042
PS_D           -489.593    317.122  -1.544 0.122903
PS_N           -457.265    315.917  -1.447 0.148058
PS_L           -476.198    320.041  -1.488 0.137049
C_Skin.C        365.293     58.933   6.198 7.99e-10 ***
C_Colour        165.619     48.661   3.404 0.000689 ***
C_Toilet        132.147     57.791   2.287 0.022401 *
C_Wellne      -1723.206    206.833  -8.331 2.30e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 541 on 1127 degrees of freedom
Multiple R-squared:  0.4754,   Adjusted R-squared:  0.4656
F-statistic: 48.63 on 21 and 1127 DF,  p-value: < 2.2e-16
```

Model plots over final model with stepwise procedure (both directions)






Scale-Location
553
433
100
√|Standardized residuals|
0
1
2
3
4
0
1000
2000
3000
4000
Fitted values
lm(Demand ~ CatDriver + ScrAndSniff + lnCatPhan + CS_Back.Cover + CS_Ending ...


Residuals vs Leverage
553
433
100
Standardized residuals
20
15
10
5
0
-5
1
0.5
0.5
1
Cook's distance
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Leverage
lm(Demand ~ CatDriver + ScrAndSniff + lnCatPhan + CS_Back.Cover + CS_Ending ...

Annex 4: List of products with no demand

Product code	Long description	Campaign	Category	Brand	Catalogue section	Offer percentage	Offer
25385	Natural Balance Bar Choco	2014-C03	Wellness	Wellness by Oriflame	Category sections	0	No offer
25386	Natural Balance Bar Super Berries	2014-C03	Wellness	Wellness by Oriflame	Category sections	0	No offer
22791	WellnessPack woman	2014-C02	Wellness	Wellness by Oriflame	Category sections	0	No offer
22793	WellnessPack man	2014-C02	Wellness	Wellness by Oriflame	Category sections	0	No offer
22794	Multivitamin & Mineral woman	2014-C02	Wellness	Wellness by Oriflame	Category sections	0	No offer
22795	Multivitamin & Mineral man	2014-C02	Wellness	Wellness by Oriflame	Category sections	0	No offer
25385	Natural Balance Bar Choco	2014-C02	Wellness	Wellness by Oriflame	Category sections	0	No offer
25386	Natural Balance Bar Super Berries	2014-C02	Wellness	Wellness by Oriflame	Category sections	0	No offer
25385	Natural Balance Bar Choco	2014-C01	Wellness	Wellness by Oriflame	Category sections	0	No offer
25386	Natural Balance Bar Super Berries	2014-C01	Wellness	Wellness by Oriflame	Category sections	0	No offer
24693	Natural Balance Soup Tomato & Basil	2014-C01	Wellness	Wellness by Oriflame	Category sections	0	No offer
24694	Natural Balance Soup Asparagus	2014-C01	Wellness	Wellness by Oriflame	Category sections	0	No offer
25414	Swedish Beauty Complex Plus	2014-C01	Wellness	Wellness by Oriflame	Category sections	0	No offer

Product code	Long description	Campaign	Category	Brand	Catalogue section	Offer percentage	Offer
15447	Natural Balance Shake natural strawberry	2014-C03	Wellness	Wellness by Oriflame	Category sections	0	No offer
15448	Natural Balance Shake natural vanilla	2014-C03	Wellness	Wellness by Oriflame	Category sections	0	No offer
22138	Natural Balance Shake natural chocolate	2014-C03	Wellness	Wellness by Oriflame	Category sections	0	No offer
24695	Botanical Infusion Revitalise	2014-C03	Wellness	Wellness by Oriflame	Category sections	0	No offer
25032	Botanical Infusion Relax	2014-C03	Wellness	Wellness by Oriflame	Category sections	0	No offer
27039	Multivitamins & Minerals essentials	2014-C03	Wellness	Wellness by Oriflame	Category sections	0	No offer
23834	Eternal Gloss - Timeless Red	2014-C02	Colour Cosmetics	Oriflame Beauty	Category sections	0	No offer
26570	Lip Impact Crayon - Pink Impact	2014-C03	Colour Cosmetics	Oriflame Beauty	Category sections	0	No offer
22591	Pure Colour Mono Eye Shadow - Taupe Metal	2014-C01	Colour Cosmetics	Pure Colour	Category sections	0	No offer
22907	Oriflame Beauty Studio Artist Foundation - Porcelain	2014-C03	Colour Cosmetics	Oriflame Beauty	Category sections	0	No offer
18460	Oriflame Beauty Triple Core Lipstick - Amazing Peach	2014-C02	Colour Cosmetics	Oriflame Beauty	Category sections	0	No offer
23868	Pure Colour Mono Eye Shadow - Pearly Gold	2014-C02	Colour Cosmetics	Pure Colour	Category sections	0	No offer

Product code	Long description	Campaign	Category	Brand	Catalogue section	Offer percentage	Offer
26573	Lip Impact Crayon - Deep Berry	2014-C03	Colour Cosmetics	Oriflame Beauty	Category sections	0	No offer
26665	Pure Colour Mono Eye Shadow - Copper Plum	2014-C02	Colour Cosmetics	Pure Colour	Category sections	0	No offer
26907	Colour Drop Lipstick - Melting Pink	2014-C01	Colour Cosmetics	Oriflame Beauty	Category sections	0	No offer
22552	Oriflame Beauty Wonder Colour Lipstick - Violet Fairy	2014-C02	Colour Cosmetics	Oriflame Beauty	Category sections	0	No offer
26664	Pure Colour Mono Eye Shadow - Sheer Lavender	2014-C02	Colour Cosmetics	Pure Colour	Category sections	0	No offer
26571	Lip Impact Crayon - Striking Rose	2014-C03	Colour Cosmetics	Oriflame Beauty	Category sections	0	No offer
22752	Giordani Gold Jewel Lipstick - Warm Coral	2014-C03	Colour Cosmetics	Giordani Gold	Category sections	0	No offer
24214	Cocktails & the City Flirty Bella Shower Gel	2014-C01	Toiletries	Cocktails	Category sections	100	Gift with purchase - spend x get x free
22436	Air Eau de Toilette	2014-C02	Fragrances	Ice, Fire, Air	Category sections	0	No offer
21659	Very Me Lovebirrrds Eye Pencil - Sweet Green	2014-C03	Colour Cosmetics	Very Me	Category sections	60	Discontinued sales out
20438	Very Me Clickit Eyeliner - Grey	2014-C03	Colour Cosmetics	Very Me	Category sections	35	Straight discount
22959	Oriflame Paper Bag - Big	2014-C03	Accessories	Non Branded	Category sections	60	Discontinued sales out

Product code	Long description	Campaign	Category	Brand	Catalogue section	Offer percentage	Offer
26824	Silver Cosmetic Bag	2014-C02	Accessories	Bioclinic	Category sections	80	Purchase with purchase - spend x get x for
26974	Vivienne Purse	2014-C03	Accessories	Non Branded	Category sections	60	Discontinued sales out
26975	Vivienne Cape	2014-C03	Accessories	Non Branded	Category sections	60	Discontinued sales out
25480	Classic Chic Umbrella	2014-C01	Accessories	Non Branded	Category sections	60	Discontinued sales out

Annex 5: Log-linear regression models (R outputs)

All Variables included

```
Call:
lm(formula = Demand ~ ., data = trd)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7104 -0.5944  0.1531  0.7662  4.2627

Coefficients: (1 not defined because of singularities)
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        5.8726502  3.7031920   1.586  0.11306
ActiveConsultants -0.0002152  0.0002960  -0.727  0.46730
Period                    NA         NA      NA       NA
CatDriver          1.9871223  0.7208899   2.756  0.00594 **
ScrAndSniff        0.8965763  0.4045051   2.216  0.02686 *
InCatPhan          0.8593549  0.2766977   3.106  0.00195 **
CS_Category.s     -0.1518993  0.2266117  -0.670  0.50280
CS_Back.Cover      0.8379219  0.6035259   1.388  0.16530
CS_Ending.Sec      0.8229473  0.3781130   2.176  0.02973 *
CS_Platform        0.3377171  0.2657013   1.271  0.20398
O_Straight.d       0.0658756  0.0029718  22.167  < 2e-16 ***
O_Gift.with.       0.0203339  0.0046183   4.403 1.17e-05 ***
O_Set              0.0395653  0.0086370   4.581 5.15e-06 ***
O_Combined.O       0.0515434  0.0104680   4.924 9.75e-07 ***
O_Discontinu       0.0560396  0.0046786  11.978  < 2e-16 ***
O_Purchase.w       0.0217056  0.0085387   2.542  0.01115 *
NewS               1.5552989  0.2033739   7.647 4.39e-14 ***
X2ndNS             0.4057200  0.2166699   1.873  0.06139 .
PS_Z              -0.4821300  0.8504161  -0.567  0.57087
PS_D              -1.4111423  0.7990193  -1.766  0.07765 .
PS_N              -1.0888447  0.7976833  -1.365  0.17252
PS_L              -1.4969432  0.8031281  -1.864  0.06260 .
C_Skin.C           0.3727909  0.6518834   0.572  0.56753
C_Colour          -0.6481330  0.6495453  -0.998  0.31858
C_Toilet          -0.2041507  0.6534182  -0.312  0.75477
C_Access          -0.3940210  0.6592912  -0.598  0.55020
C_Fragra          -1.0902513  0.6594256  -1.653  0.09854 .
C_Wellne          -6.7940454  0.8752421  -7.762 1.87e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.344 on 1122 degrees of freedom
Multiple R-squared:  0.5918,   Adjusted R-squared:  0.5824
F-statistic: 62.57 on 26 and 1122 DF,  p-value: < 2.2e-16
```

Stepwise - Backward

```
Call:
lm(formula = Demand ~ CatDriver + ScrAndSniff + InCatPhan + CS_Back.C
over +
    CS_Ending.Sec + CS_Platform + O_Straight.d + O_Gift.with. +
    O_Set + O_Combined.O + O_Discontinu + O_Purchase.w + NewS +
    X2ndNS + PS_D + PS_N + PS_L + C_Skin.C + C_Colour + C_Fragra +
    C_Wellne, data = trd)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8972 -0.5983  0.1578  0.7522  4.2280

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.386761   0.408574   5.842 6.76e-09 ***
CatDriver      2.130501   0.702659   3.032 0.002484 **
ScrAndSniff    0.946610   0.401619   2.357 0.018594 *
InCatPhan      0.875307   0.274510   3.189 0.001469 **
CS_Back.Cover  0.983193   0.564183   1.743 0.081662 .
CS_Ending.Sec  1.004929   0.304793   3.297 0.001007 **
CS_Platform    0.448829   0.155935   2.878 0.004074 **
O_Straight.d   0.066112   0.002924  22.607  < 2e-16 ***
O_Gift.with.   0.020543   0.004605   4.461 8.98e-06 ***
O_Set          0.040512   0.008565   4.730 2.53e-06 ***
O_Combined.O   0.050579   0.010335   4.894 1.13e-06 ***
O_Discontinu   0.056355   0.004521  12.466  < 2e-16 ***
O_Purchase.w   0.020529   0.008446   2.431 0.015226 *
NewS           1.511284   0.196391   7.695 3.07e-14 ***
X2ndNS         0.363414   0.212922   1.707 0.088136 .
PS_D          -0.957271   0.303379  -3.155 0.001645 **
PS_N          -0.621742   0.296314  -2.098 0.036105 *
PS_L          -1.048707   0.317898  -3.299 0.001001 **
C_Skin.C       0.613232   0.129399   4.739 2.42e-06 ***
C_Colour      -0.392885   0.102177  -3.845 0.000127 ***
C_Fragra      -0.856161   0.163939  -5.222 2.10e-07 ***
C_Wellne      -6.570436   0.591389 -11.110  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.342 on 1127 degrees of freedom
Multiple R-squared:  0.591,    Adjusted R-squared:  0.5834
F-statistic: 77.55 on 21 and 1127 DF,  p-value: < 2.2e-16
```

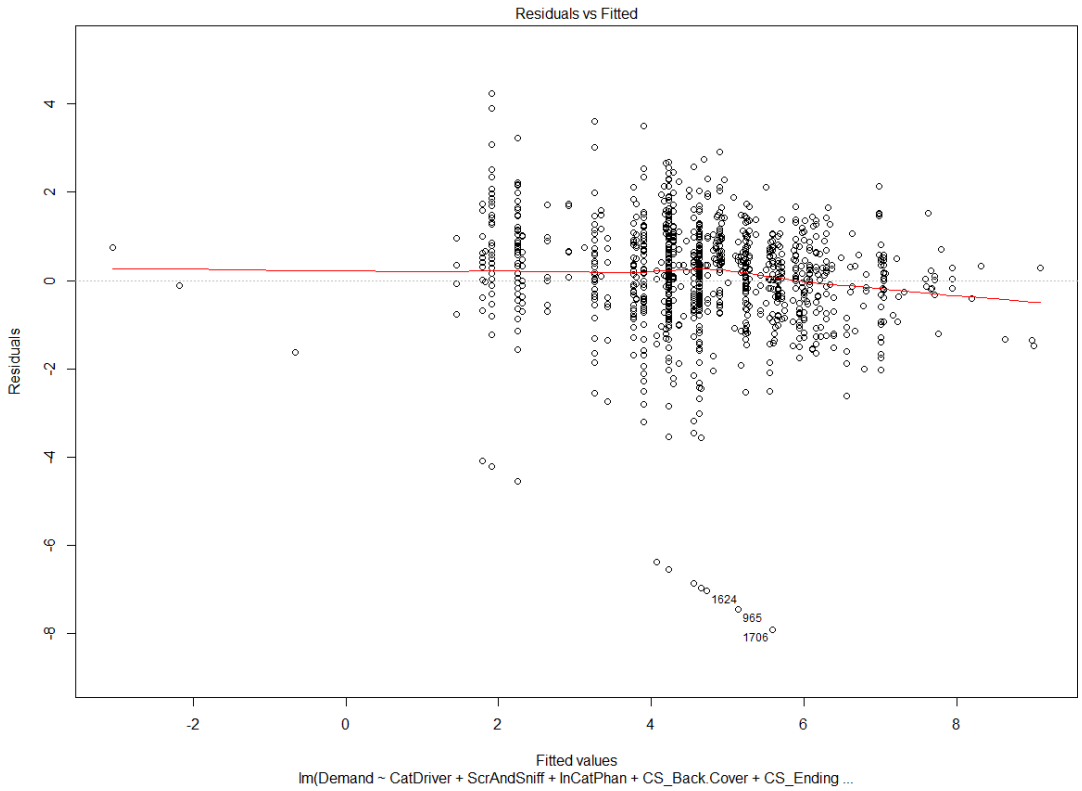
Stepwise – Forward

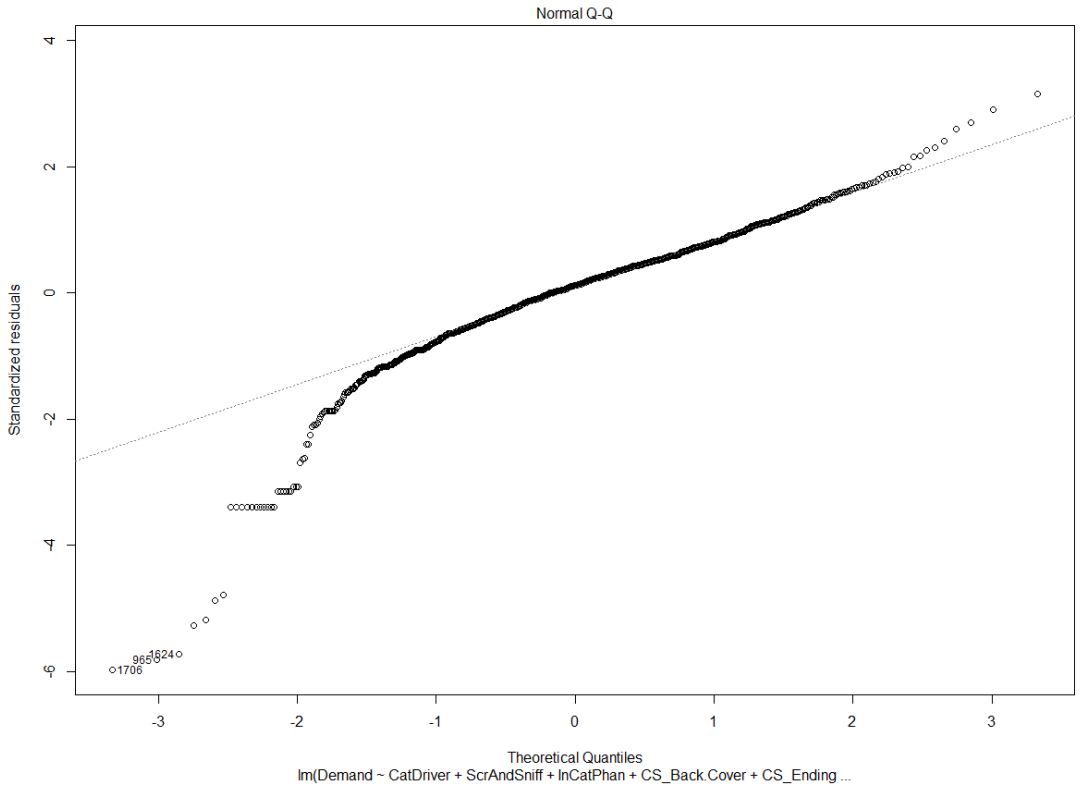
```
Call:
lm(formula = Demand ~ CatDriver + ScrAndSniff + InCatPhan + CS_Back.C
over +
    CS_Ending.Sec + CS_Platform + O_Straight.d + O_Gift.with. +
    O_Set + O_Combined.O + O_Discontinu + O_Purchase.w + NewS +
    X2ndNS + PS_D + PS_N + PS_L + C_Skin.C + C_Colour + C_Fragra +
    C_Wellne, data = trd)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8972 -0.5983  0.1578  0.7522  4.2280

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.386761   0.408574   5.842 6.76e-09 ***
CatDriver      2.130501   0.702659   3.032 0.002484 **
ScrAndSniff    0.946610   0.401619   2.357 0.018594 *
InCatPhan      0.875307   0.274510   3.189 0.001469 **
CS_Back.Cover  0.983193   0.564183   1.743 0.081662 .
CS_Ending.Sec  1.004929   0.304793   3.297 0.001007 **
CS_Platform    0.448829   0.155935   2.878 0.004074 **
O_Straight.d   0.066112   0.002924  22.607  < 2e-16 ***
O_Gift.with.   0.020543   0.004605   4.461 8.98e-06 ***
O_Set          0.040512   0.008565   4.730 2.53e-06 ***
O_Combined.O   0.050579   0.010335   4.894 1.13e-06 ***
O_Discontinu   0.056355   0.004521  12.466  < 2e-16 ***
O_Purchase.w   0.020529   0.008446   2.431 0.015226 *
NewS           1.511284   0.196391   7.695 3.07e-14 ***
X2ndNS         0.363414   0.212922   1.707 0.088136 .
PS_D          -0.957271   0.303379  -3.155 0.001645 **
PS_N          -0.621742   0.296314  -2.098 0.036105 *
PS_L          -1.048707   0.317898  -3.299 0.001001 **
C_Skin.C       0.613232   0.129399   4.739 2.42e-06 ***
C_Colour      -0.392885   0.102177  -3.845 0.000127 ***
C_Fragra      -0.856161   0.163939  -5.222 2.10e-07 ***
C_Wellne      -6.570436   0.591389 -11.110  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.342 on 1127 degrees of freedom
Multiple R-squared:  0.591,    Adjusted R-squared:  0.5834
F-statistic: 77.55 on 21 and 1127 DF,  p-value: < 2.2e-16
```

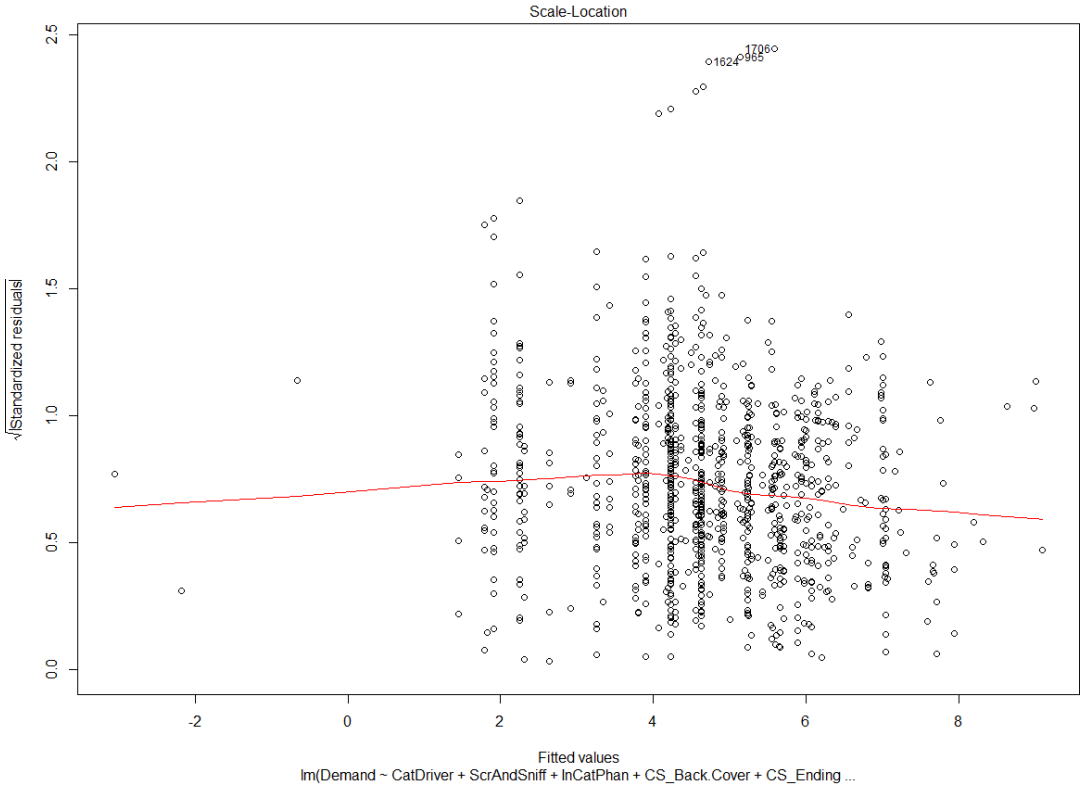
Stepwise – Both

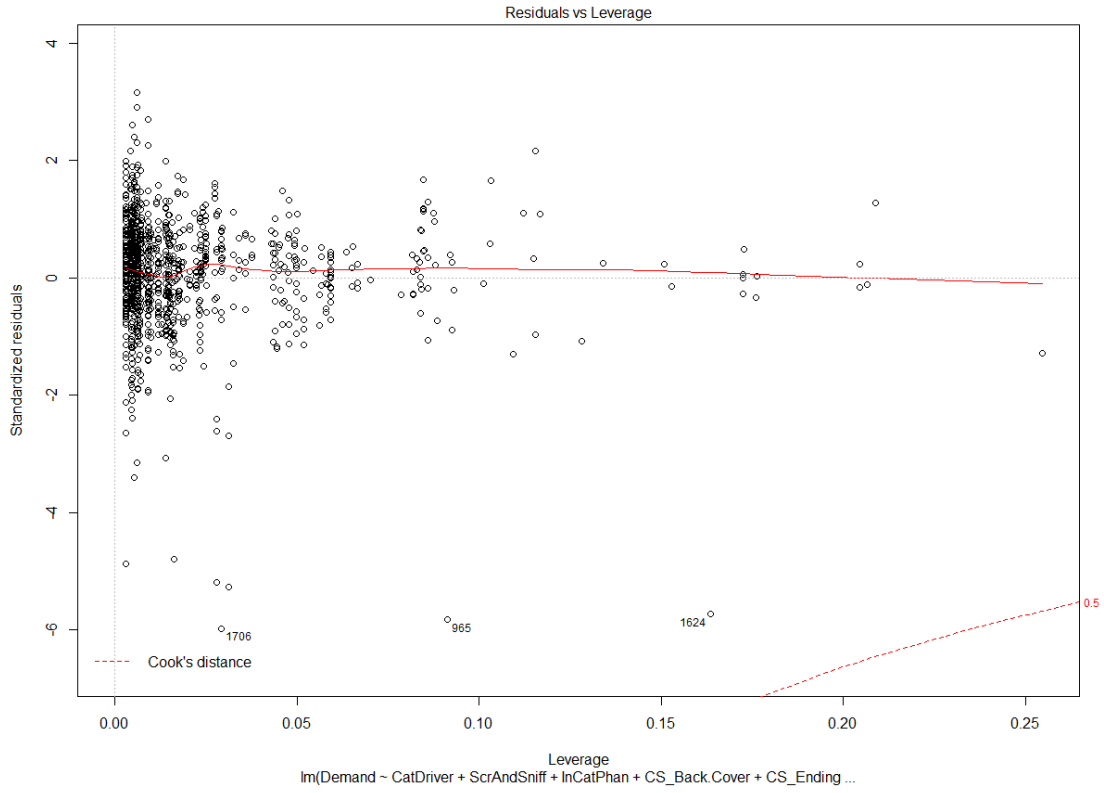
```
Call:
lm(formula = Demand ~ CatDriver + InCatPhan + CS_Ending.Sec +
    CS_Platform + O_Straight.d + O_Gift.with. + O_Set + O_Combined.O
+
    O_Discontinu + NewS + X2ndNS + PS_Z + PS_D + C_Skin.C + C_Colour
+
    C_Toilet + C_Access + C_Fragra + C_Wellne, data = trd)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2337 -0.6384  0.1435  0.7807  4.6141

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.458391   0.506802   6.824 1.39e-11 ***
CatDriver      2.272784   0.546463   4.159 3.42e-05 ***
InCatPhan      1.109578   0.218748   5.072 4.54e-07 ***
CS_Ending.Sec  0.816482   0.332068   2.459  0.01408 *
CS_Platform    0.369697   0.145805   2.536  0.01135 *
O_Straight.d   0.064540   0.002784  23.186  < 2e-16 ***
O_Gift.with.   0.027224   0.004782   5.693 1.56e-08 ***
O_Set          0.037426   0.008017   4.669 3.37e-06 ***
O_Combined.O   0.037473   0.009202   4.072 4.96e-05 ***
O_Discontinu   0.038852   0.003583  10.844  < 2e-16 ***
NewS           1.415042   0.172498   8.203 5.90e-16 ***
X2ndNS         0.836115   0.192601   4.341 1.54e-05 ***
PS_Z           0.753841   0.348582   2.163  0.03077 *
PS_D          -0.302411   0.093443  -3.236  0.00124 **
C_Skin.C      -1.283598   0.460224  -2.789  0.00537 **
C_Colour      -2.343625   0.453509  -5.168 2.77e-07 ***
C_Toilet      -1.808064   0.460235  -3.929 9.03e-05 ***
C_Access      -1.967960   0.464459  -4.237 2.44e-05 ***
C_Fragra      -2.535208   0.467937  -5.418 7.27e-08 ***
C_Wellne      -8.777726   0.659866 -13.302  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.359 on 1215 degrees of freedom
Multiple R-squared:  0.5927,   Adjusted R-squared:  0.5863
F-statistic: 93.06 on 19 and 1215 DF,  p-value: < 2.2e-16
```

Model plots over final log-linear model with stepwise procedure (both directions)






Scale-Location
√|Standardized residuals|
Fitted values
lm(Demand ~ CatDriver + ScrAndSniff + lnCatPhan + CS_Back.Cover + CS_Ending ...
1706
965
1624


Residuals vs Leverage
Standardized residuals
Leverage
lm(Demand ~ CatDriver + ScrAndSniff + lnCatPhan + CS_Back.Cover + CS_Ending ...
Cook's distance
0.5
1706
965
1624

Annex 6: decision trees to identify drivers of zero demand or values above 4000 units – an attempt to predict outliers.

After excluding outliers, it was possible to develop a prediction model for demand. Nevertheless, there are two situations out of scope. First, products without demand and more important, the reasons that could lead a demand above 4000 units.

For the first subproblem, it is possible to formulate a case to be tackled with a decision tree; data is rearranged with a new variable that can detect if the demand is zero or not. Then by using all product classification and variables, it is possible to use a decision tree to predict what factors could lead to zero demand.

As a result, the decision rule would infer that Wellness category is more likely to get zero demand (see Figure 40).

Figure 40: decision tree for Demand=0

On the other hand, after evaluating cases with a demand larger than 4000 units a decision tree suggests that “Purchase with purchase - spend x get x for” is an effective offer when it is combined with two segments: foundations and nourishers. More than 50% of the records with larger demand, belongs to this combination (Figure 41).

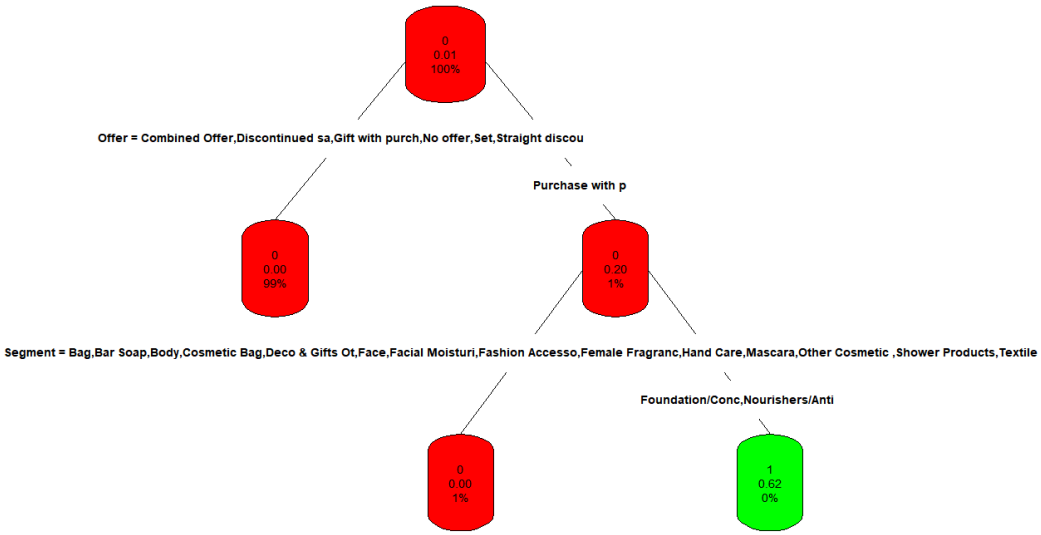


Figure 41: decision tree for larger demand cases (over 4000 units).

Annex 7: R-Scripts

Data splitting:

```
trainingData <- d7a[d7a$Period<3, ] # model training data

testData  <- d7a[d7a$Period>2, ]   # test data
```

Demand transformation for Log linear models:

```
trd <- trainingData[trainingData$Demand > 1, 1:28]  # optional

#to exclude outliers

trd <- trainingData

trd$Demand[trd$Demand == 0 ]

trd$Demand[trd$Demand > 0 ] <- log(trd$Demand[trd$Demand > 0 ])

trd$Demand[trd$Demand == 0 ] <- log(0.1)
```

Routines for stepwise procedure

```
m1min <- lm(Demand ~ O_Straight.d+  O_Gift.with.+     O_Set+
       O_Combined.O+    O_Discontinu+     O_Purchase.w, data=trd)

formulam1 <- formula(m1)

#m1b <- step(m1,direction = "backward") #backward forward and #both

#m1b <- step(m1min,direction = "forward", scope=formulam1) #backward
forward and #both

m1b <- step(m1,direction = "both")
```

Error calculation for non-transformed demand:

```
sqrt(mean((exp(m1ptrd)-exp(trd$Demand))^2)) # in sample RMSE

mean(abs(exp(m1ptrd)-exp(trd$Demand))) # in sample MAE
```

```
sMAPE(exp(trd$Demand), exp(m1ptrd)) # in sample sMAPE

sqrt(mean((exp(m1ptst)-tst$Demand)^2)) # out sample RMSE

mean(abs(exp(m1ptst)-tst$Demand)) # out sample MAE

sMAPE(tst$Demand, exp(m1ptst)) # out of sample sMAPE
```

Error calculation for log models:

```
sqrt(mean((exp(m1ptrd)-exp(trd$Demand))^2)) # in sample RMSE

mean(abs(exp(m1ptrd)-exp(trd$Demand))) # in sample MAE

sMAPE(exp(trd$Demand), exp(m1ptrd)) # in sample sMAPE

sqrt(mean((exp(m1ptst)-tst$Demand)^2)) # out sample RMSE

mean(abs(exp(m1ptst)-tst$Demand)) # out sample MAE

sMAPE(tst$Demand, exp(m1ptst)) # out of sample sMAPE
```

Linear models

```
m1d1 <- lm(Demand~.,data=d1tr)


pr_m1d1 <- predict(m1d1,d1, type="response")

pr_insample_m1d1 <- predict(m1d1,d1tr, type="response")

pr_outsample_m1d1 <- predict(m1d1,d1ts, type="response")


Accuracy(d1ts$Demand, pr_outsample_m1d1 ,d1tr$Demand, digits=2)


m1d1a <- step(m1d1,direction = "backward")

pr_insample_m1d1a <- predict(m1d1a,d1tr, type="response")
```

```
pr_outsample_m1d1a <- predict(m1d1a,d1ts, type="response")


m1d1min <- lm(Demand ~ 1, data=d1tr)


formulam1d1 <- formula(m1d1)


m1d1b <- step(m1d1min,direction = "forward", scope=formulam1d1)
summary(m1d1b)
pr_insample_m1d1b <- predict(m1d1b,d1tr, type="response")
pr_outsample_m1d1b <- predict(m1d1b,d1ts, type="response")




m1d1c <- step(m1d1,direction = "both")
summary(m1d1c)
pr_insample_m1d1c <- predict(m1d1c,d1tr, type="response")
pr_outsample_m1d1c <- predict(m1d1c,d1ts, type="response")
```

Model training (KNN) both metrics with repeated cross validation:

```
repeats = 5
numbers = 10
tunel = 20
x <- trainControl( method = "repeatedcv" , repeats = repeats,
```

```
                    number = numbers ,
                    classProbs = FALSE ,
                    summaryFunction = defaultSummary,
                    verboseIter = TRUE)
model1 <- train(Demand~. , data = trainingData, method = "knn",
                preProcess = c("center","scale"),
                trControl = x,
                metric = "MAE", #first metric
                tuneLength = tunel, verboseIter = TRUE)

model1 <- train(Demand~. , data = trainingData, method = "knn",
                preProcess = c("center","scale"),
                trControl = x,
                metric = "RMSE", # second metric optional
                tuneLength = tunel, verboseIter = TRUE)
```