LOCATION AND CLUSTER BASED SALES CHANNEL POTENTIAL ANALYSIS IN RETAIL

Capstone Project

Birtuğ BİLGİN

İSTANBUL, 2021

MEF UNIVERSITY

LOCATION AND CLUSTER BASED SALES CHANNEL POTENTIAL ANALYSIS IN RETAIL

Capstone Project

Birtuğ BİLGİN

Advisor: Prof. Dr. Adem KARAHOCA

İSTANBUL, 2021

MEF UNIVERSITY

Name of the project: Location and Cluster Based Sales Channel Potential Analysis in Retail
Name/Last Name of the Student: Birtuğ BİLGİN
Date of Thesis Defense: 20/06/2021

I hereby state that the graduation project prepared by Birtuğ Bilgin has been completed under my supervision. I accept this work as a "Graduation Project".

28/06/2021
Prof.Dr. Adem KARAHOCA

I hereby state that I have examined this graduation project by Birtuğ Bilgin which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

28/06/2021

Director
of
Information Technology Program

We hereby state that we have held the graduation examination of Birtuğ Bilgin and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Prof.Dr. Adem KARAHOCA	……..
2. ……..	…….

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Birtuğ Bilgin Date Signature

EXECUTIVE SUMMARY

LOCATION AND CLUSTER BASED SALES CHANNEL POTENTIAL ANALYSIS in RETAIL

Birtuğ Bilgin

Advisor: Prof.Dr. Adem KARAHOCA

JUNE, 2021, 34 pages

This analysis project was conducted on the need to obtain new analysis and inferences for the existing traditional sales channels of company, which wants to progress in line with its omni-channel goals. In order to reach the customer with the same level of service in all channels it is necessary to analyze the dynamics of the channel well. In this project, I aimed to make sense of demographic data with the linear model and future selection model and to transform it into meaningful information that will guide sales strategies. Especially for diffusion strategies, in addition to traditional methods, data-based location analysis and analysis of sales weights of existing points are required. With the information to be provided, new dealer opening processes will also be based on data.

Key words: omnichannel, clustring, retail, feature selection

ÖZET

PERAKENDE SEKTÖRÜNDE

LOKASYON VE KÜMELEME BAZLI SATIŞ KANALI POTANSİYEL ANALİZİ

Birtuğ BİLGİN

Proje Danışmanı: Prof Dr. Adem KARAHOCA

HAZİRAN,2021, 34 sayfa

Bu projede, başta omni-channel strateji hedefleri doğrultusunda ilerlemek isteyen perakende sektöründe faaliyet gösteren firmasının geleneksel satış kanalları için analiz ve çıkarımlar elde etme ihtiyacı üzerinden yola çıkılmıştır. Müşteriye tüm kanallarda aynı seviyede hizmet verebilmek için kanalın dinamikleri iyi analiz edilmelidir. Bu projede demografik verilerin lineer model ve özellik önceliklendirme algoritmaları kapsamında ve kümeleme yöntemleri ile anlamlı sonuçlar elde etmek ve iş birimlerine çıktılar oluşturmayı amaçladım. Makine öğrenmesi modellerini de bu amaçla uyarlamak istedim. Özellikle yayılım stratejileri kapsamında geleneksel yöntemlere ek olarak veriye dayalı lokasyon analizi ve mevcut noktaların satış ağırlıklarının analizi gerekmektedir. Çalışmanın çıktıları yayılım stratejilerine de faydalı olacaktır.

Anahtar Kelimeler: çoklu kanal, kümeleme, perakende, nitelik seçimi

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

This analysis study is in line with the omnichannel vision for 'X' company; will benefit the vision of having equal information flow in all channels; was made for the purpose of dealer segmentation study.

It should be noted that this 'X' company continues its commercial activities in the online and offline channels in the retail consumer electronics sector. Also in the offline channel; consists of its own dealer points and a traditional sales channel affiliated with a separate commercial company. The need of the company is the need to apply the strategy that it has in its own online channel and that it follows in its own stores, to the dealer distribution channel with analytical outputs.

It should be defined as managing a high-volume retail is based on the algorithms working in the background and these algorithms based on accurate data. The working power of the models is directly related to the fluency and correct scoring of the data. There are prerequisites such as "managing the logistics infrastructure, creating the right product penetration in the field, ensuring the availability of the right product for the right customer, and examining the stock availability and accuracy parameters cumulatively for the integration of the online sales channel and the offline sales channel". The high level of competition and customer satisfaction in the retail sector will be more manageable in the offline channel as well as in the online channel.

Clustering outputs of the segmentation study will provide the right product delivery to the right point in the market. It will provide the infrastructure of the common campaign creation strategy for the segments formed within the channel and provide the ability to make data-based automatic decision-making in business rules. The data on the basis of location analytics will also enable location analysis and it will provide a preliminary data of information pool to determine a sales point expansion strategy for existing potential regions.

Roadmap, creation of dataset of sales points in company 'X'; examining the data, selecting features for demographic data, creating parameters with location based clusters, graphing the results with the data pool outputs of the clustering study based on the K-Means algorithm and accessing meaningful outputs for business teams.

2. BUSINESS UNDERSTANDING

Company 'X', which is the subject of the case, provides services in retail industry both online and offline. For this purpose, sectors will be discussed from different aspects.

Due to the structure of the retail industry, directly appeals to people and consumers. It is related to many sectors. For this reason, their size is determined by each of the streams of the scope substeps.

Observing trends and adapting developments effectively come to the fore at the most effective point of orientation to the consumer through the right channel. The retail industry is open to growth, extremely lively and variable and offers a structure in transformation with technological developments. It is one of the areas where developments and consumer trends are felt the most.

In the researches, the volume and numerical situations of the sector are stated as follows; "The market has reached a size of 25 trillion dollars as of 2019. Although it seems to be growing slower than the world economy in general, we see that the trend has changed since 2016." [12]

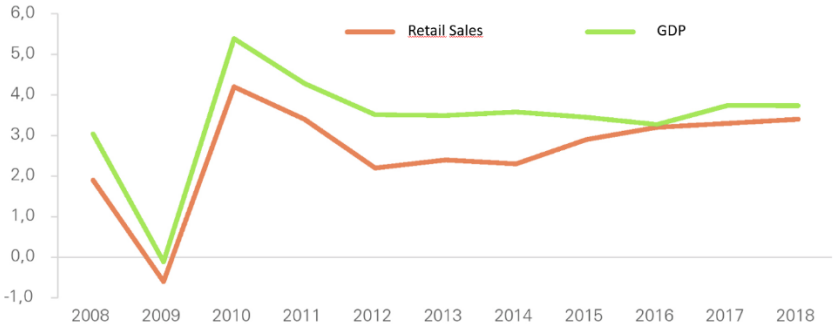

Figure 1: Change of Sales and GPD Ratio of Turkey by Years [12]

According to the data of the Turkish Informatics Industry Association (TÜBİSAD), Turkey e-commerce volume grew by 42% as of the end of 2018, reaching 59.9 billion TL. The value 31.5 billion TL of this volume part seems to belong to the retail sector. Due to the pandemic living conditions experienced in 2020, it is seen that this rate has moved much higher in the e-commerce channel. These indicators show that the online channel with potential should merge with the offline channel.

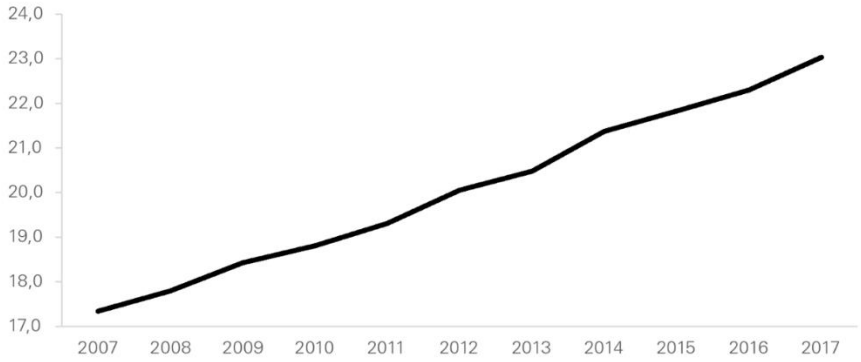

Figure 2: Retail Growth Rates of Turkey per Year [12]

The constant transformation challenge in the retail industry affects profit margins and creates pressure on companies to take part in many channels. In particular, the shift of consumer demand from location-based markets to online media and the increase in consumers access to the same service and the ease of comparison between channels in this process increase competition among institutions. At this point, keeping the profit margins at the desired rates again necessitates the joint management of the channels and the approach to the consumer with the same service quality through all channels.

Level of transformation, the competition in the market is directly under the following headings;

- Customers' habits and tendencies towards shopping vary; established with brands. The relationship called interaction and customer experience affects deeply. Although it is certain that retail companies with classical organizations will protect their assets; strategies to increase its market share push it to enrich its classical services. Brands not only sell products to their customers, but also enable them to experience their brands and products in creative ways with shopping using augmented reality and peer technologies, and experience enhancing customized products.
- The development of technology and the shift of competition to this field in the short term particularly beneficial to consumers. Nowadays; even with social media applications, product placement, experience creation, visual similar product search and price comparison services are offered. These data not only enable consumers to shop in more rational ways but also enable them to compare their experience and sales channels.

Companies that carry out sales organization through different channels, especially in the geography where they operate. It becomes mandatory to reach customers with data-based and equal service in all channels of sales organizations. The point where brands benefit is that they have a data pool for optimizing sales channels with various artificial intelligence and machine learning techniques of data obtained from all channels reaching the consumer.

- Using artificial intelligence technologies establishment of this infrastructure brings the advantage of being a pioneer for all brands that ensure the data-based development of their channels and aim to establish a structure that grows around the concept of omni-channel. The power of multi-channell is critical to transform and track the footprints left by customers in digital channels and to use retail channels at the point of contact with customers.

  While companies offer an end-to-end experience to their customers in both online and offline sales processes, they are challange to follow their customers well and to provide equal quality service at all contact points in the light of the omni-channel concept.

  Service quality comes to the fore at this point. Especially in the consumer durables and electronics market, the fact that customers can now have more information and access than they have ever had makes the concept of satisfaction critical.

  The impact of a customer who had a bad experience is now really great as opposed to operating only in offline channels, as in the 90's. For this reason, brands vary from the quality of the products they sell to the delivery performance; they have to perform many background tasks flawlessly, starting from their own sales organization such as the dealer network to the quality of their after-sales services. This highlights the ability to achieve the service quality determined in all channels and the effect this creates.

2.1. E-Commerce with Retail

Although the electronic commerce medium is growing rapidly in our country, it lags far behind developed economies in terms of penetration rate. This also reveals the magnitude of the potential available for brands. The rate of internet access and the number of people using smart devices has increased considerably. This shows that brands can maintain their commercial presence even in channels that do not have sales points, and the chance of direct

access to customers. Especially with the effect of the pandemic, we see that the rates are even higher.

It is clearly seen that the curfews experienced during the pandemic period and the long-term closure of sales points such as shopping malls also accelerate the online shopping tendencies of consumers. Retail companies that see this as a chance and develop strategies in this direction, fast delivery to the consumer in sectors such as textile and white goods and consumer electronics. It aims to provide a quick response to the trade flow and customers by turning to methods such as e-commerce deliveries through the dealer network.

At this point, maintaining the turnover of the sales points has been an important point for the points to survive. Especially if your sales network is established with a third-party business partner, thats are dealers also requires companies to implement similar models in order to protect their business partners. Dealers with low sales potential can be directed to sales via online channels and marketplaces and sales teams can work to penetrate the right products for the target audience in their locations. The concept of penetration comes to the fore especially in terms of providing brand awareness. For this reason, the outputs should be evaluated by considering the concept of calibalization.

If you maintain your presence in the market with a third-party dealership network in your sales network. As with your own sales points and e-commerce channel, you should have live sales tracking, instant stock tracking and customer traffic data tracking and match them in the database. In order for smart algorithms to serve the omni-channel approach, it is necessary to provide all this data flow and optimize with appropriate models. In particular, stock tracking is the building block for your strategies to direct your customers to your dealer through your central e-commerce site. This will not be enough. However, that can be model local marketing campaigns, sales strategies, channel expansion initiatives and analysis of sales performance against potential with high accuracy.

2.2. Situation of White Good Sectors in Retail

It is known that the white goods and durable consumer goods sector is dominant in the retail sector. Distribution and growth rates in the sector progress in parallel.

The situation of the white goods sector is mentioned in the researches of the relevant institution as follows; ‘’Turkey's white goods industry is Europe's largest and the world's second

largest production base after China. The sector made a significant contribution to Turkey's industrial power by producing a total of 29.1 million units in product groups consisting of refrigerators, freezers, washing machines, dryers, dishwashers and ovens in 2020.'' [13]

Digital transformation takes place in a way that covers the entire value chain. Not only products but also production processes are affected by digitalization. In this research, interactions are discussed especially in terms of sales organizations.

Structures as separate and optimized cells turn into fully interconnected, automated and optimized manufacturing processes. In this way, while providing high efficiencies. The traditional relations between suppliers, manufacturers and customers even between machine and human are changing. Similar efficiency studies are also carried out in domestic retail sales organizations. High levels of stock in the market and increase in stock/month ratios cause additional costs. It also affects the penetration and calibalization status of the products. Therefore, the retail market should also be managed based on data.

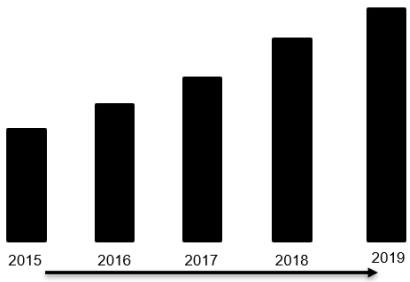

Figure 3: World White Good Sector Growth Year to Year

The sector has created a serious economy both in terms of volume and logistics mobility. Research shows that; It stands out as a pioneer in the white goods sector, especially in certain products, both in production and sales channels. ‘’The Turkish White Goods Industry is one of the leading sectors of the Turkish economy with its competitive power and brands, producing high added value in production and exports. The size of the global white goods market, which consists of refrigerators, freezers, washing machines, dryers, dishwashers and ovens, is 247 billion USD and 493 million units as of 2019, according to retail data.'' [13]

When we examine the statistics, we see that there is a growth trend parallel to consumption values. Again, when we approach with the figures stated in the report, the sector shows growth both in terms of volume and total economy.

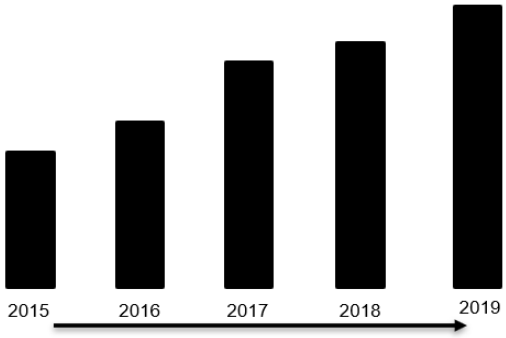

Figure 4: White Good Sales Pieces Comparison by Years[13]

Company 'X', which is the subject of our analysis, operates directly in the white goods and consumer electronics sectors. For this reason, it is necessary to mention the metrics of the white goods sector in our country. The white goods sector is one of the locomotive sectors of Turkey with its production power. Turkey, which is the leader after China, has also come to the position of supporting directly and sub-industry organizations with millions of production between 2016-2020.

Figure 5: White Good Sector Sales of Turkey[13]

The global success of the Turkish white goods industry is directly proportional to the industry's production power based on high export volume. As of 2020, the white goods industry exported 22 million units in total, in 6 product groups consisting of refrigerators, deep freezers, washing machines, dryers, dishwashers and ovens. This export performance has made the white goods industry one of the leading industries that contribute to Turkey's foreign trade balance and reduce foreign dependency. “The white goods sector, which has a high export intensity, exported 76% of its total production in 2020. The export rate of the sector has maintained its high pace throughout the year.”[13]

“The domestic market sales of the Turkish white goods industry totaled 7.8 million units as of 2020, in 6 product groups consisting of refrigerators, freezers, washing machines, dryers, dishwashers and ovens. According to the domestic market sales data for 2020, washing machines reached 27%, refrigerators 26% and dishwashers 19%. These three product groups correspond to 72% of the total domestic market sales.” [13]

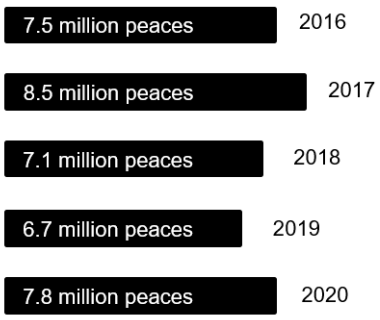

Figure 6: White Goods Sector Domastic Sales of Turkey

“The Turkish e-commerce sector reached 14.6 billion USD in 2019 with a growth of 18% on USD basis. 7.9 billion USD, which corresponds to 54% of the total market size, is realized by the retail industry. The Turkish e-commerce retail market, on the other hand, constitutes 6.2% of the total retail sector in Turkey as of 2019. This rate, which is 12.3% in developed countries, sheds light on the future in terms of the growth performance of the Turkish e-commerce retail market. According to BCG's 'Consumer Device Industry Trends' report, the global consumer devices market internet retail has grown by 12% on a 5-year basis. As a result of this growth, the e-commerce channel has become the largest channel after specialist retailers. Hyper-markets follow the e-commerce channel”. [13]

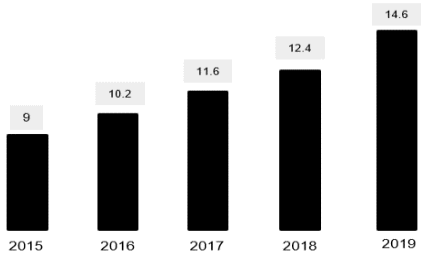

Figure 7: E-Commerce Market Size of Turkey per Years(Billion USD**)**

The factors contributing to the development of internet retailing in the world can be listed as follows: low prices applied on the Internet, easy comparison of products, ease of delivery of products.

Improvement in enabling consumers to make payments securely, consumers can find a large number of products on the internet, this methods that provide ease of payment (Installation options, etc.). “The e-commerce retail market in Turkey achieved a growth rate of 12.2% on annual averages in the 2015-2019 period and 21.5% in 2019 on a USD basis.” [13]

3. DATA PREPARATION AND MODEL DEVELOPMENT

Clustering is one of the most common exploratory data analysis techniques used to gain an intuition about the structure of data. It can be defined as the task of identifying subgroups in data because data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. The expectation is to be able to identify relationships analytically outside of the organizational approach.

If we examine it under the title of machine learning methods; K-Means algorithm is an unsupervised learning and clustering algorithm.

Techniques used in data mining can be divided into models according to the type and the intended use of the results obtained. These models can be grouped under two headings. These are predictive and descriptive models.

Descriptive models extract relationships from the dataset. Data mining techniques used in descriptive models are clustering, summarization, association rules, and ordered sequences.

Predictive models; on the other hand, develop a model from situations with known results and obtain new results from data sets with unknown results. Here, this preferred purpose in adapting it to the existing case cluster analysis is the process of grouping information in a data set according to certain proximity criteria.

The method subject to the model is unsupervised learning. Unsupervised machine learning finds all kinds of unknown patterns in data. Unsupervised methods help you find features that can be useful for categorization. The requirements in this analysis study, in addition to the data pool owned by 'X' company, by obtaining demographic data sets from TUIK and other external data sources. In the role of the relevant location and categorical data is identifying invisible relationships between features and applying them to business rules. This happens in real time within the run, so all input data is analyzed and labeled in the presence of learned.

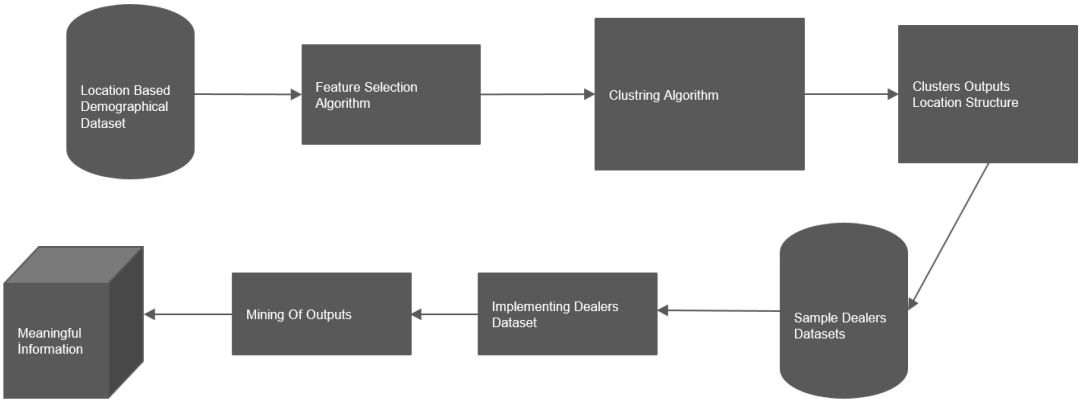

Figure 8: Main Structure of Development Steps

The sub-method model will focus on is the location-based clustering method. It clusters the objects given here by measuring their distance from some random or specific objects on an n-dimensional plane. Therefore, these methods are also known as distance-based methods.

According to the working mechanism of the k-means algorithm, k objects are randomly selected to represent the center point or mean of each cluster. The remaining objects are included in the clusters with which they are most similar, taking into account their distance from the mean values of the clusters. Then calculating the average value of each cluster; new cluster centers are determined and the distances of the objects to the center are examined again. The algorithm continues to repeat until there is no change.

3.1. Analysing The Dealer Datasets

The study was started with the data preparation steps in the data set. Here, if there is a deficiency in the data of the relevant sales point, it is removed from the data set. (For the sample to be equal in the entire data set, 12 months of sales data is required for the partnership of the model.)

Again, physical properties were examined in terms of missing value. The data is in the data set both categorically and numerically. Encoding is essential for categorical data to be processed in a cluster algorithm. In this way, all data is expressed in numerical values. Data type commonality is ensured.

Label encoding and one hot encoding methods are available in the literature as encoding methods. Here, the appropriate preference for the data set was determined. One Hot Encoding takes the column with categorical variables and splits it into multiple columns. Label encoding works without seperate blocks. At this stage, we will convert such categorical data into numerical variables with the label encoder so that it can be understood in the model.

After Label encoding, assign the data set to an array numerically. It is ready to work in order to perform operations related to the new dataframe we have created.

Table 1: Encoded Dealers Data Table

```
array([[0.75      , 0.13513514, 0.26724138],
       [0.75      , 0.83783784, 0.50287356],
       [0.75      , 0.09459459, 0.55747126],
       ...,
       [0.        , 0.06756757, 0.56609195],
       [0.75      , 0.06756757, 0.        ],
       [0.        , 0.32432432, 0.        ]])
```

Clustering work plays a clustering algorithm by selecting features that have physical separations of sales points (for example; store physical area, total turnover per year). The structure we observe here is that there is no regular distribution in the data set. Therefore, we will expand our dataset using demographic data. After the feature selection steps, we will evaluate our sample locations together.

Location_ID	Area	Dealer_Type	Sales_Score
14553	10	3	93
14087	62	3	175
14586	7	3	194
14043	50	0	290
14891	71	3	260

Figure 9: Header View Of Numericized Dealer Data Table

The categorical data in the data set was converted into numerical data. At the same time, the related painting was shared with the figure. After this point, a data set was prepared for labeling the sales points with IDs according to the location based data modeling outputs.

3.2.Feature Selection For Linear Modeling And Cluster Analysing With Locational Based Demographic Dataset

To start the application, we download the necessary libraries in the notebook. Our approach is to get to know the first-stage dataset, focus on feature selection, and carry out clustering studies on parametrically meaningful features.

There are more than 70 features in the dataset. In this data set based on district location. There are numerical values such as socio-economic status, per capita income level, income level per household, number of ATMs in the region. As the first step, we select the features that the business teams concentrates on and that they want to include in the decision processes. Then we work with the feature selection algorithm on these features**.** The studies included in the analysis progressed on the basis of features. Related titles are labeled as score_ID for data privacy reasons.

It is quite valuable for AIC (Akaike information criterion) feature selection. In the case of a linear model, validating the data and observing the significant score can often result in better model selection than traditional training/validation/test model selection methods.

While observing the AIC score in the linear model, it is important to reach the minimum score that will minimize information loss. The score is observed in the outputs of the steps. While removing the parameters with low significant value from the model, the loop is terminated at the point where the value starts to increase. The analysis continues by taking our features onto the dataframe with the minimum AIC score obtained as a result of the trials.

Table 2: AIC Score and Model Outputs Step1

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.723 on 32145 degrees of freedom
Multiple R-squared:  0.8085,    Adjusted R-squared:  0.8083
F-statistic:  4680 on 29 and 32145 DF,  p-value: < 0.00000000000000022

175935.747161626
```

It defines our dataset to the program. R script related analysis is starting to work. This is the step that covers all the features in the model. Multiple R-squared value: 0.8005. At this point, we need to look at the summary detail on the basis of all in order to make feature selection. By assigning score to the significant value of the summary table. It gives preliminary information about the features that should be extracted from the data set.

Table 3: Score_ID's Outputs Step1

```
Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) -1.330131   0.314920  -4.224    0.000024097505635 ***
Score_ID1   -0.019874   0.005631  -3.529             0.000418 ***
Score_ID2   -0.018137   0.001030 -17.606 < 0.0000000000000002 ***
Score_ID5   -0.280593   0.016462 -17.045 < 0.0000000000000002 ***
Score_ID14  -0.079491   0.021328  -3.727             0.000194 ***
Score_ID15  -2.431117   0.075484 -32.207 < 0.0000000000000002 ***
Score_ID17  -0.106435   0.003727 -28.558 < 0.0000000000000002 ***
Score_ID18  -0.320730   0.018140 -17.681 < 0.0000000000000002 ***
Score_ID19   0.580443   0.024496  23.695 < 0.0000000000000002 ***
Score_ID32  -0.056606   0.006296  -8.991 < 0.0000000000000002 ***
Score_ID33   0.023299   0.012481   1.867             0.061932 .
Score_ID34   0.053810   0.004450  12.093 < 0.0000000000000002 ***
Score_ID35   0.025276   0.002394  10.560 < 0.0000000000000002 ***
Score_ID39   1.148821   0.025807  44.516 < 0.0000000000000002 ***
Score_ID40   1.583706   0.054336  29.147 < 0.0000000000000002 ***
Score_ID41  -0.061077   0.014819  -4.121    0.000037743344726 ***
Score_ID42   0.035422   0.013637   2.598             0.009393 **
Score_ID58   0.842466   0.010825  77.826 < 0.0000000000000002 ***
Score_ID61   0.050216   0.001958  25.651 < 0.0000000000000002 ***
Score_ID62  -0.081258   0.011369  -7.147    0.000000000000904 ***
Score_ID63  -0.068490   0.015226  -4.498    0.000006875549466 ***
Score_ID64   0.031422   0.012494   2.515             0.011907 *
Score_ID66  -0.064252   0.004543 -14.143 < 0.0000000000000002 ***
Score_ID67  -0.036093   0.011595  -3.113             0.001854 **
Score_ID68   0.114869   0.018603   6.175    0.000000000670167 ***
Score_ID69  -0.054492   0.013225  -4.120    0.000037916645466 ***
Score_ID70   0.071564   0.008177   8.751 < 0.0000000000000002 ***
Score_ID71  -0.022757   0.013834  -1.645             0.099980 .
Score_ID72  -0.015119   0.017898  -0.845             0.398272
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the second step, we will do the exclusion of the list for the score_ID72 variable and run the model again. This feature has been chosen because the Pr(>|t|) value is higher than the Score_ID71. Here we run the linear model for the second time.

Table 4: Score_ID's Outputs Table Step2

```
Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept) -1.315359   0.314433  -4.183    0.000028809968236 ***
Score_ID1   -0.019701   0.005628  -3.501             0.000465 ***
Score_ID2   -0.018122   0.001030 -17.595 < 0.0000000000000002 ***
Score_ID5   -0.281342   0.016438 -17.115 < 0.0000000000000002 ***
Score_ID14  -0.077631   0.021214  -3.659             0.000253 ***
Score_ID15  -2.430303   0.075477 -32.199 < 0.0000000000000002 ***
Score_ID17  -0.106579   0.003723 -28.627 < 0.0000000000000002 ***
Score_ID18  -0.321004   0.018137 -17.699 < 0.0000000000000002 ***
Score_ID19   0.580910   0.024490  23.720 < 0.0000000000000002 ***
Score_ID32  -0.056819   0.006291  -9.032 < 0.0000000000000002 ***
Score_ID33   0.022293   0.012424   1.794             0.072753 .
Score_ID34   0.054502   0.004374  12.461 < 0.0000000000000002 ***
Score_ID35   0.025340   0.002392  10.592 < 0.0000000000000002 ***
Score_ID39   1.148895   0.025807  44.519 < 0.0000000000000002 ***
Score_ID40   1.585301   0.054302  29.194 < 0.0000000000000002 ***
Score_ID41  -0.063903   0.014436  -4.426    0.000009609357597 ***
Score_ID42   0.035226   0.013635   2.584             0.009783 **
Score_ID58   0.842324   0.010824  77.822 < 0.0000000000000002 ***
Score_ID61   0.050269   0.001957  25.691 < 0.0000000000000002 ***
Score_ID62  -0.082412   0.011287  -7.302    0.000000000000291 ***
Score_ID63  -0.070132   0.015101  -4.644    0.000003427734474 ***
Score_ID64   0.031569   0.012492   2.527             0.011506 *
Score_ID66  -0.064227   0.004543 -14.138 < 0.0000000000000002 ***
Score_ID67  -0.036131   0.011594  -3.116             0.001834 **
Score_ID68   0.114377   0.018593   6.151    0.000000000776675 ***
Score_ID69  -0.055939   0.013114  -4.266    0.000019987106977 ***
Score_ID70   0.071534   0.008177   8.748 < 0.0000000000000002 ***
Score_ID71  -0.022440   0.013829  -1.623             0.104658
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When we observed the new results, the AIC score decreased compared to the previous step. (175934.747161626>175933.3863535) According to our approach, we continue to make feature selections to reduce the AIC score parameter. It will repeat the next step by removing the Score_ID71 parameter from the table because it still gets a low score.

Table 5: Score_ID's Outputs Table Step3

```
Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept) -1.320776   0.314423  -4.201    0.000026688680499 ***
Score_ID1   -0.019496   0.005626  -3.465             0.000531 ***
Score_ID2   -0.018097   0.001030 -17.572 < 0.0000000000000002 ***
Score_ID5   -0.283424   0.016388 -17.294 < 0.0000000000000002 ***
Score_ID14  -0.079841   0.021170  -3.771             0.000163 ***
Score_ID15  -2.426476   0.075442 -32.163 < 0.0000000000000002 ***
Score_ID17  -0.106362   0.003721 -28.586 < 0.0000000000000002 ***
Score_ID18  -0.321229   0.018137 -17.712 < 0.0000000000000002 ***
Score_ID19   0.580460   0.024489  23.703 < 0.0000000000000002 ***
Score_ID32  -0.057290   0.006284  -9.116 < 0.0000000000000002 ***
Score_ID33   0.022478   0.012423   1.809             0.070404 .
Score_ID34   0.054200   0.004370  12.403 < 0.0000000000000002 ***
Score_ID35   0.025251   0.002392  10.557 < 0.0000000000000002 ***
Score_ID39   1.148844   0.025807  44.516 < 0.0000000000000002 ***
Score_ID40   1.583310   0.054290  29.164 < 0.0000000000000002 ***
Score_ID41  -0.063430   0.014434  -4.395    0.000011137473579 ***
Score_ID42   0.035271   0.013635   2.587             0.009692 **
Score_ID58   0.842564   0.010823  77.850 < 0.0000000000000002 ***
Score_ID61   0.050149   0.001955  25.647 < 0.0000000000000002 ***
Score_ID62  -0.082428   0.011287  -7.303    0.000000000000288 ***
Score_ID63  -0.070767   0.015096  -4.688    0.000002774110942 ***
Score_ID64   0.031944   0.012490   2.557             0.010547 *
Score_ID66  -0.063834   0.004537 -14.071 < 0.0000000000000002 ***
Score_ID67  -0.035639   0.011591  -3.075             0.002109 **
Score_ID68   0.114600   0.018593   6.163    0.000000000720156 ***
Score_ID69  -0.056128   0.013114  -4.280    0.000018729416465 ***
Score_ID70   0.069449   0.008076   8.600 < 0.0000000000000002 ***
---
```

In the new step, we see that our AIC score parameter is 175933.73596727. This feature is not removed as the AIC score is in an upward trend compared to the previous step; We end the loop in this step. Draw a heatmap on our new model.

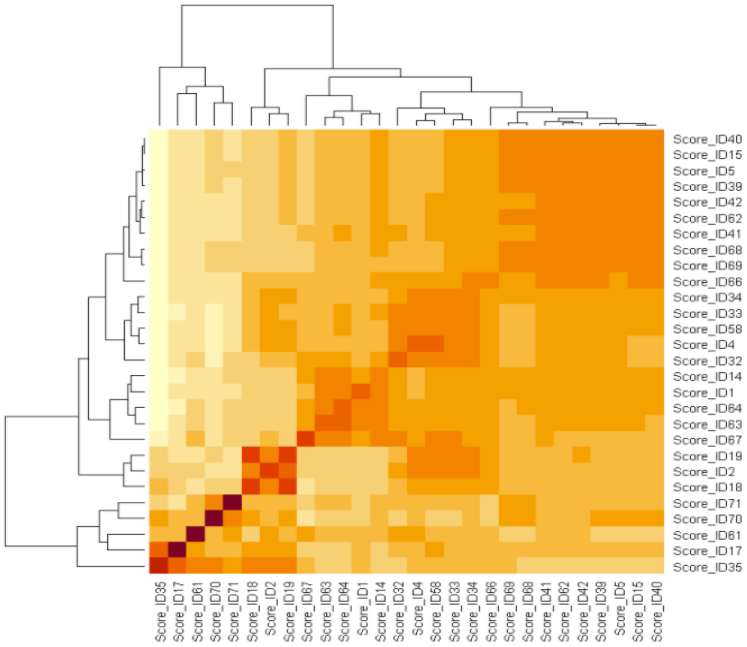

Figure 10: Heatmap of Selected Features

After this stage, continue the model with the clustering stage. The goal will be to create and monitor existing potential groups with relevant features in the building dataset.

3.3. K-Means Clustring

The K-means algorithm is an unsupervised learning clustering algorithm, as mentioned in the literature explanation section. Focusing on features, it creates clusters with similar features in spatial dimension. The algorithm puts statistically similar records in the same group. An element is allowed to belong to only one set. Here, this aimed to make clustering work based on location and analyze the outputs.

```
In [18]: # function to compute total within-cluster sum of square
wss <- function(k) {
  kmeans(na.omit(data_modeling3), k, iter.max=30 )$tot.withinss
}

# Compute and plot wss for k = 1 to k = 15
k.values <- 1:15

# extract wss for 2-15 clusters
wss_values <- map_dbl(k.values, wss)

plot(k.values, wss_values,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Total within-clusters sum of squares")
#3 clusters seem to be sufficient for clustering.
#8 clusters might be better if more info is required.
```

Figure 11: K-Means Algoritm Steps in R

In order to determine the ideal number of clusters refer to the elbow method. Here is the Elbow method according to each K value of the points, the sum of the square of their distance from the cluster center is calculated. According to these values, for each K value get the graph mix; for interpretation, the elbow point on the graph where the difference between the totals starts to decrease is chosen as the most appropriate K value.

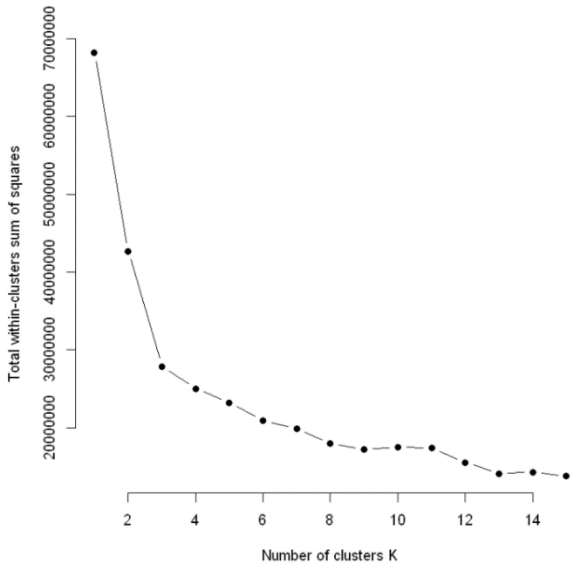

Figure 12: Drawing Outputs of Algorithm Defined Elbow Method

We are looking for an elbow point to determine the ideal number of clusters we come across. Here, we see the elbow structure for the 3 clusters most prominently. Although this observe a small bend at 8 points, we are advancing the location cluster analysis over 3 clusters to be useful in product segment analysis.

```
# Compute k-means with k = 3
km.res3 <- kmeans(data_modeling3, 3, nstart = 25)
#Mean values of Features for Clusters
aggregate(data_modeling3, by=list(cluster=km.res3$cluster), mean)
```

cluster	Score_ID1	Score_ID2	Score_ID5	Score_ID14	Score_ID15	Score_ID17	Score_ID18	Score_ID19	Score_ID32	...	Score_ID62	Score_ID63	Score_ID64
1	0.07434474	0.00000	0.2670252	0.1002067	0.1580375	49.44110	0.09499908	0.07351198	0.2425888	...	0.1387666	0.1454095	0.1574340
2	0.42142745	54.94447	1.5115929	0.2513194	1.2087541	51.49070	2.71883058	2.11588824	0.3253250	...	0.9924473	0.1750599	0.2101796
3	5.06161576	74.88710	11.2814879	2.0728091	8.9994666	51.73869	4.13388835	4.09086728	15.9819440	...	8.5860900	3.0876160	4.8068020

Figure 13: Score_ID Tables with Clusters Head in Phyton Script

For cluster analysis, we label our dataset for 3 clusters. The number of clusters determined here is labeled separately for each point, and printed the cluster output in main population rating file.

3.4. Integrated Data Set for Model

At this stage, the phyton script is passed. Location_ID and cluster labels were made on the data set and the output was taken. In the next step, analytical outputs about the clusters will be obtained by using the data in the 'Dealer Cluster Data' and 'Product Segment Data' tables. In this way, numerical balances and differences within the relevant clusters will be observed, and sales count evaluations will be made on a segment basis.

4. EXPORTING AND INTERPRETING OUTPUTS TO DEALER MASTER DATASET PROPERTIES

Up to this step of the analysis, examined the main data regarding our dealers in the sample. Then search for meaningful outputs for clustering by performing location analysis in the light of demographic data. In this level of the analysis, it will map the selling points in our current sample with the points labeled clusters.

Next, it will examine the sales turnover values for these three sample clusters and the segment information regarding the products they main sales data set. Based on this information, analyze the situation for the sales points in the clusters and develop recommendations.

In the master data files of the 'X' company, first obtain the sales data of the 'Y' year for the selected sample set on a product basis. (Based on the principle of data privacy, the date range of the data set could not be shared here.) Calculation of the rates will convey information about the sales trends of the points and clusters.

Then, calculate the product segment information for each of the three clusters and compare their effects on sales and amount ofturnover. It will comment on the target customer group based on the sales movements within the clusters. As stated at the beginning of our approach, outputs are extremely important in order to draw inferences about sales points, to set proposal, campaign and product targets, and to direct points such as online channel sales and integration to different areas.

At this step, a basic category from the electronic product group was chosen to examine. (The product category name and details are hidden due to data privacy and security policy.) First, match the product segment information of the sales data we base on for the relevant 'Y' year in our master data file, on the basis of product code. Then, create our main block data set by calling the relevant location cluster information from the population ratings file.

Then, using the pivot method; graph the average turnover data and total amount of turnover data for these three clusters. (Within the scope of data privacy and security policy, calculations are transferred proportionally and units and numbers are hidden.)

- Chart of totals sales; it allows us to comment on these three clusters. In the related product category, we see the dominance of the C3 cluster class in terms of turnover. It is followed by C1 and C2 clusters, respectively.

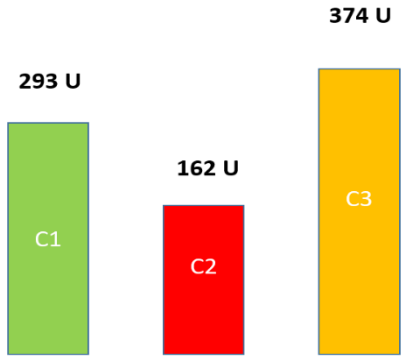

Figure 14: Scaled Total Sales per Clusters

- Looking at the cumulative turnover is insufficient because the number of samples in the cluster is different. For this reason, plot the average sales data of the clusters with a graph. When examine the second graph, we see that the average value of the C1 cluster which is in the second place according to the total sales amount is the highest. They are followed by C3 and C2 clusters respectively.

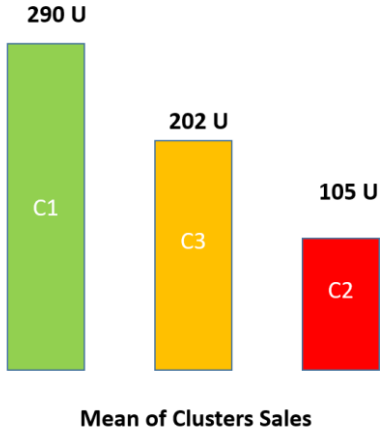

Figure 15: Scaled Mean Dealers Sales per Clusters

Here try to see why the ranking is not the same in both graphs with additional data. Since the segments addressed by the points are different; the product groups they sell to are also

different according to their respective target audiences. For this reason they do not see the same ranking as the product quantities are different within the segments. This shows that the clusters also differ in terms of target audience. Again in explain of additional data, they examine the sales weights in the clusters graphically for the relevant product category.

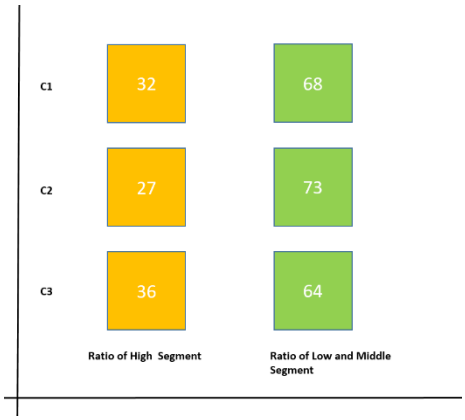

Figure 16: Ratio of Product Segments per Clusters

When we examine the graph, we see that the C3 cluster has the highest sales rate in high segment products. The sales rate in the C1 cluster, on the other hand, has a high rate almost as in the C3 cluster. For this reason, while the C1 cluster is in the second place in the total ranking, it is in the 1st place in the average comparison.

Again, while the C2 cluster is in the second place in the total ranking in the weighted average calculation, it is in the last place. These inferences show us that clusters differ in terms of sales habits and the locations they address.

Output-based information and decision sets are suggested as follows;

- Brand positioning should be reconsidered for the spread of high-end products in all regions. A dynamic product segment model can be created according to the model and the demographics of the regions.

- A business model can be created to increase turnover by integrating online channels for dealers located in locations with low average turnover.

- Local marketing campaigns can be created to increase turnover in locations with a low average turnover.

- Store concepts that will increase the user experience can be developed in order to increase online channel success and spread in regions with high sales of high-end products.

- With the omni-channel concept, smaller volume product delivery interim warehouses can be created in regions with low sales potential. Here the segment pivot table can be used while calculating the stock quantities and penetrations of the products.

- Sales teams can prioritize the locations in the cluster with high turnover potential for store expansions. Similar campaigns can be organized for dealers in the same clusters. Here, it is recommended to analyze the sales and development strategies over the areas in the clusters.

- This study shows us that different clusters appeal to different sales potentials and product segments. Within the location expansion strategy, new point determination preferences should be made in parallel with the company's brand and product positioning. This analysis can be used.

5. CONCLUSION

Working through the existing dealer sales channel of 'X' company which is in the process of omni-channel transition; focused on a more homogeneous separation of the channel, examining the sales potentials within the channel independent of the organization, and seeing the current potential with an approach based on product segments. The data sets of the 'X' company were used by scaling and visualizing, and location-based demographic data set was used for location analytics. (Locations and titles included are hidden for data security and privacy.)

Diffusion strategy; positioning with physical stores in the right locations for 'X' company, which wants to provide peer customer satisfaction in all its existing channels. In regions with low sales potential, it also provides outputs to appeal to customers through its online sales channel with a lower cost service. These outputs are aimed at increasing the performance of the physical sales channel as well as providing strategy outputs in order to position new stores in regions with existing potential. At the same time, reviewing product segment information it also provides homework outputs to business teams in order to reorganize the product groups to be penetrated according to the potential clusters in the field.

As a data scientist and analyst approach, feature-based analysis and selection of the location-based dataset with the AIC linear methodology increases the performance of the data set in your hand. When this performance increase is tested with the sample data set; It was seen that the cyclical and physical differences were distinguishable by numbers. It was discovered that when the cluster formation was made using the Elbow method and the K-Means model, the data set was made functional and had analytical results for the business team.

When we evaluate the outputs; 384 physical sales dealers in the sample were analyzed based on their existing data, and then the relevant points were pivoted with potential groups. The turnovers of the sales points; product segment information that they mainly sell the location data set was a good sample set to make sense of and generalize.

The outputs were examined with a business background the recommendations in the fourth section were created. In order for the model to yield higher analytical results, the data set should be expanded. A comprehensive analysis should be made with numerical data such as credit information and stock ratios regarding the existing dealer channel. In this way, a separate mathematical model can be created in which target-oriented turnover analysis and physical store inferences can be made in the process of creating a new store.

In this analysis, linear model and cluster models were used and relevant outputs were examined in order to correctly position the existing dealer sales channel of the business units in the omni-channel organization by analyzing the existing data.

REFERENCES

[1] Sulekha Goyat ," The Basis Of Market Segmentation: A Critical Review Of Literatüre", European Journal of Business and Management www.iiste.org ISSN 2222-1905 (Paper) ISSN 2222-2839 (Online) Vol 3, No.9, 2011

[2] Marina Meireles Pereira and Enzo Morosini Frazzon, "Towards A Predictive Approach For Omni-Channel Retailing Supply Chains", IFAC PapersOnLine 52-13, 844–850, 2019

[3] Subir Bhattacharya and Kushal Saha, "Look Before You Leap: Economics Of Being An Omnichannel Retailer", OPERATIONS AND SUPPLY CHAIN MANAGEMENT Vol. 13, No.3 , pp. 256 - 268 ISSN 1979-3561, EISSN 2759-9363, 2020

[4] İpek Kazançoğlu and Keti Ventura and Çağlar Aktepe,"Omnı-Channel Applıcatıons In Retaılıng: Challenges And Barrıers For Logıstıcs Operatıons", UİİİD-IJEAS, 16. UİK Özel Sayısı :219-236 ISSN 1307-9832, 2017

[5] Stanislava GROSOVA and Magda CISAROVA and Ivan GROS, "B2b Segmentatıon As A Tool For Marketıng And Logıstıc Strategy Formulatıon", INTELEKTINĖ EKONOMIKA INTELLECTUAL ECONOMICS 2011, No. 1(9), p. 54–64, 2011

[6] Jayant Tikmani and Sudhanshu Tiwari and Sujata Khedkar, "An Approach To Customer Classification Using K-Means", International Journal of Innovative Research in Computer and Communication Engineering Vol.3, Issue 11, November 2015

[7] Malgorzata Weso Lowska, "A Study On Feature Selection Based On Aıcc And İts Application To Microarray Data", Universitat Polit`ecnica de Catalunya, June 2009

[8] Ceyhun Çelik and Hasan Şakir Bilge, " Feature Selectıon Wıth Weıghted Condıtıonal Mutual Informatıon", Journal of the Faculty of Engineering and Architecture of Gazi University Cilt 30, No 4, 585-596, 2015 Vol 30, No 4, 585-596, 2015

[9] D3M Labs, "Customer Segmentation: Rules-based vs. K-Means Clustering", .[Online]. Avaliable: Https://Www.D3mlabs.De/2019/09/11/Customer-Segmentation-An-İntroduction/

[10] Özlem Toplu Yılmaz and Onur Bayram, "E-Trade And E-Export Of Turkey İn Covıd-19 Pandemic Period", Kayseri University Journal of Social Sciences, Vol 2, No: 2, 37-54, December 2020

[11] Nicolien Teunissen , “Exploring The Opportunity To Combine Customer Segments İn A B2b Market”, MSc. Business Administration - Purchasing and Supply Management Master Thesis

[12] Kpmg Turkey, Retail Market Research Report 2019, .[Online]. Avaliable: https://assets.kpmg/content/dam/kpmg/tr/pdf/2019/03/sektorel-bakis-2019-perakende.pdf

[13] White Goods Manufacturers’ Association of Turkey, Sector Report, 2020 .[Online]. Avalible: http://www.turkbesd.org/userfiles/files/T%C3%9CRKBESD%20Beyaz%20E%C59Fya%20Sekt%C3%B6r%20Raporu%20.pdf

[14] Şevket Ay, “K-Means Algoritm” Oct. 2019, [Online]. Avaliable: Https://Medium.Com/Deep-Learning-Turkiye/K-Means-Algoritmas%C4%B1-B460620dd02a

[15]Alexandre Zajic, “Introduction to AIC — Akaike Information Criterion”, Dec,2019 .[Online].Avaliable:Https://Towardsdatascience.Com/İntroduction-To-Aic-Akaike-İnformation -Criterion-9c9ba1c96ced

[16] Meral Candan Cetin and Aydin Erar, “Variable Selection With Akaike Information Crıterıa: A Comparative Study”, Hacettepe Journal of Mathematics and Statistics Volume 31, 89–97,2002

APPENDIX

```
#Future Selection Linear Model R Script;
#install.packages("factoextra")
#install.packages("class")
#install.packages("NbClust")
library("purrr")
#install.packages("tidyverse")
library("readxl")
library("factoextra")
library("class")
library("NbClust")
options(scipen=999)
set.seed(123)

#importing location based file
df = read_excel('main location datas.xlsx')
head(df)

data = df[,c(1,2,3,6,15,16,18,19,20,33,34,35,36,40,41,42,43,59,61,62,63,64,65,67,68,69,70,71,72,73,5)]

#Removing Total Row
data = data[-nrow(data),]

#NA values are assumed as zero. (Score_ID1)
table(is.na(data))
apply(is.na(data), 2, any)
colnames(data)[apply(is.na(data), 2, any)]
data[is.na(data)] = 0
```

```
data_modeling = data[,-1]

#Normalization
normalize <- function(x) {
  return (round((x - min(x)) / (max(x) - min(x)) * 100, 3)) }
data_modeling <- as.data.frame(lapply(data_modeling, normalize))

str(data_modeling)

linear_model = lm(Score_ID4~., data = data_modeling)
plot(linear_model, which=1)

summary(linear_model)
AIC(linear_model)
#Lower AIC value defined is indicates a better fit.

#Removing insignificant feature from data set
data_modeling2 = data_modeling[,-18]
linear_model2 = lm(Score_ID4~., data = data_modeling2)
AIC(linear_model2) #AIC value decreased.
summary(linear_model2)

#Removing insignificant feature from data set
data_modeling3 = data_modeling2[,-28]
linear_model3 = lm(Score_ID4~., data = data_modeling3)
AIC(linear_model3) #AIC value decreased
summary(linear_model3)
```

```
#Removing insignificant feature from data set
data_modeling4 = data_modeling3[,-27]
linear_model4 = lm(Score_ID4~., data = data_modeling4)
AIC(linear_model4) #AIC value increased, so we continue with 'data_modeling3' data set.
summary(linear_model4)

correlation = data.frame(cor(data_modeling3))
correlation = round(correlation,3)
heatmap(cor(correlation))

#Data set with instance names
data_set = data[,-c(18,28)]
head(data_set)

# function to compute total within-cluster sum of square
wss <- function(k) {
  kmeans(na.omit(data_modeling3), k, iter.max=30 )$tot.withinss
}

# Compute and plot wss for k = 1 to k = 15
k.values <- 1:15

# extract wss for 2-15 clusters
wss_values <- map_dbl(k.values, wss)

plot(k.values, wss_values,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Total within-clusters sum of squares")
```

```
#3 clusters seem to be sufficient for clustering.

# Compute k-means with k = 3
km.res3 <- kmeans(data_modeling3, 3, nstart = 25)
#Mean values of Features for Clusters
aggregate(data_modeling3, by=list(cluster=km.res3$cluster), mean)

#Results
data_set <- cbind(data_set, Cluster_3 = km.res3$cluster)

write.csv2(data_set, file = "Location_Segment.csv")

-Phyton Script;
#!/usr/bin/env python
# coding: utf-8

# For data loading and manipulation
import pandas as pd
import numpy as np

# For visualization/EDA
import matplotlib as mpl
import matplotlib.pyplot as plt
import seaborn as sns
sns.set(color_codes=True)
sns.set(style="white")
get_ipython().run_line_magic('matplotlib', 'inline')
from matplotlib import cm
import xlrd
```

```
# For modeling/machine learning
from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_samples
from sklearn.preprocessing import OneHotEncoder
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import LabelEncoder

# Adding Model.2
import warnings
warnings.filterwarnings('ignore')

df= pd.read_excel('Dealer Data.xlsx', dtype={'Location_ID': np.int64,'Dealer_Type': np.str})
df.info()

le = LabelEncoder()

df['Area']= le.fit_transform(df['Area'])
df['Dealer_Type']= le.fit_transform(df['Dealer_Type'])
df['Sales_Score']= le.fit_transform(df['Sales_Score'])

df.head()

Dealer_std=df[['Dealer_Type','Area','Sales_Score']]

scaler = MinMaxScaler()
```

```
scaler.fit(Dealer_std)
scaler.transform(Dealer_std)

km = KMeans(n_clusters=2,
        init='k-means++',
        n_init=10,
        max_iter=300,
        random_state=0)
kmeans = km.fit(scaler.transform(Dealer_std))
y_km = km.fit_predict(scaler.transform(Dealer_std))
Dealer_std['cluster'] = kmeans.labels_

sns.lmplot(x="Area", y="Sales_Score", hue="cluster", data=Dealer_std, fit_reg=False)



#!/usr/bin/env python
# coding: utf-8

# For data loading and manipulation
import pandas as pd
import numpy as np

# For visualization/EDA
import matplotlib as mpl
import matplotlib.pyplot as plt
import seaborn as sns
sns.set(color_codes=True)
sns.set(style="white")
get_ipython().run_line_magic('matplotlib', 'inline')
```

```
from matplotlib import cm
import xlrd

# For modeling/machine learning
from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_samples
from sklearn.preprocessing import OneHotEncoder
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import LabelEncoder
```

Phyton Script_2;

```
import warnings
warnings.filterwarnings('ignore')

df= pd.read_excel('Dealers Cluster Data.xlsx', dtype={'Location_ID': np.int64,'Dealer_Type': np.str})
df.info()

df.groupby('Cluster3').size()

Sales = df.groupby(by = ['Cluster3'])['Sales_Score'].sum()

print(Sales.head())

Sales = df.groupby(by = ['Cluster3'])['Sales_Score'].mean()

print(Sales.head())
```

```
df2= pd.read_excel('Product Segment Data.xlsx', dtype={'Sales_Score_Count': np.float64,})
df2.info()

Product = df2.groupby(by = ['Dealer_Cluster'])['Sales_Score_Count'].sum()

print(Product.head())

Product = df2.groupby(by =
['Dealer_Cluster','Product_Segment'])['Sales_Score_Count'].sum()

print(Product)
```