MEF UNIVERSITY

MARKET BASKET ANALYSIS ON RETAIL STORES OF ELECTRONIC DEVICES

Capstone Project

Feray Ece Topcu

İSTANBUL, 2018

MEF UNIVERSITY

MARKET BASKET ANALYSIS ON RETAIL STORES OF ELECTRONIC DEVICES

Capstone Project

Feray Ece Topcu

Advisor: Asst Prof. Serap Kırbız

İSTANBUL, 2018

MEF UNIVERSITY

Name of the project: Market Basket Analysis on Retail Stores of Electronic Devices
Name/Last Name of the Student: Feray Ece Topcu
Date of Thesis Defense: 10/09/2018

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

10/09/2018
Asst Prof. Serap Kırbız

I hereby state that I have examined this graduation project by Feray Ece Topcu 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.

10/09/2018

Director
of
Big Data Analytics Program

We hereby state that we have held the graduation examination of Feray Ece Topcu and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Asst. Prof. Serap Kırbız	……..
2. Prof. Dr. Özgür Özlük	……..

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.

Feray Ece Topcu 10/09/2018

Name Date Signature

EXECUTIVE SUMMARY

MARKET BASKET ANALYSIS
ON
RETAIL STORES
OF
ELECTRONIC DEVICES

Feray Ece Topcu

Advisor: Yard. Doç. Serap Kırbız

SEPTEMBER, 2018, 40 Pages

Market basket analysis is a technique that discovers the relationship between the pairs of products purchased together. It simply analyses the purchase coincidence with the products purchased among the sales transactions and explain what is purchased with what. This study presents a market basket analysis to discover the association rules between products within a dataset that is extracted from one of a leading retail company of electronic devices. The aim is to understand the purchasing behavior trends by examining which products are purchased together. The Apriori algorithm provides the opportunity to discover association rules. Market basket analysis is developed with R Language. R community has a library for association rules that's called "arules". Additionally, "arulesViz" and "plotly" libraries are used to visualize the output of analysis. Further steps to this study could be diversification of stores groups through machine learning algorithms and according to this classification, market basket analysis may apply on generated classes of the stores separately. In addition, the output of market basket analysis can be input for a recommendation engine.

Key Words: Market Basket Analysis, Association Rules, Apriori Algorithm, arules package, arulesViz package

ÖZET

ELEKTRONİK CİHAZ
PERAKENDE MAĞAZALARINDA
PAZAR SEPETİ ANALİZİ

Feray Ece Topcu

Advisor: Yard. Doç. Serap Kırbız

EYLÜL, 2018, 40 Sayfa

Pazar sepeti analizi, birlikte satın alınan ürünler arasındaki ilişkileri keşfeden bir tekniktir. Sepetler arasında, satın alınan ürünlerle satın alma tesadüflerini analiz eder ve ürünlerin hangi ürün ile beraber satın alındığını açıklar. Bu çalışma, elektronik cihazların önde gelen perakende şirketlerinden birinden alınan bir veri kümesindeki ürünler arasındaki ilişki kurallarını keşfetmek için bir pazar sepeti analizi sunmaktadır. Amaç, hangi ürünlerin birlikte satın alındığını inceleyerek satın alma davranış eğilimlerini anlamaktır. Apriori algoritması, ilişkilendirme kurallarını keşfetme fırsatı sunar. Pazar sepeti analizi, R Dili ile geliştirilmiştir. Birliktelik kurallarını ortaya çıkarmak için R topluluğu "arules" olarak adlandırılan bir kütüphaneye sahiptir. Ayrıca, analiz çıktılarını görselleştirmek için "arulesViz" ve "plotly" kütüphaneleri kullanılmaktadır. Bu çalışmanın diğer adımları, makine öğrenimi algoritmaları ile mağazaların gruplandırılması ve bu gruplandırmaya göre pazar sepeti analizinin her bir gruba ayrı ayrı uygulanması olabilir. Ek olarak, pazar sepeti analizinin çıktısı, bir öneri sistemine girdi olarak verilebilir.

Anahtar Kelimeler: Pazar Sepeti Analizi, İlişkisel Kurallar, Apriori Algoritması, arules paketi, arulesViz paketi

TABLE OF CONTENTS

1. INTRODUCTION

Because cross selling is one of the most significant part of e-commerce, e-commerce cannot be considered without a recommendation engine. Offering a product is also necessary for offline sales in stores to improve the selling process and profitability.

During this consumption age, shopping behavior of customers is extremely important to examine for retail companies in order to gain a competitive edge on the market. One of the bigger challanges for retail companies is knowing how to extract significant information while having a huge amount of sales data. Sales data can be used to understand customers' shopping behaviour which is especially useful when the company requires to discover product association and improve marketing strategy.

Market basket analysis is one of the key techniques used to discover associations between products (Berry and Linoff, 2004) . It simply analyses the purchase coincidence with the products purchased among the sales transactions and explain what is purchased with what. Market basket analysis works by looking for combinations of products that occur together in sale transactions.

To sum it up, for improvement it is as important to understand product association for offline sales as it is important for the e-commerce sale process. The market basket analysis is the technique for exploration of relation or correlations among a set of products and it is beneficial for retail companies to gain a competitive advantage on the market.

1.1. Market Basket Analysis: Literature Survey

Market basket analysis (also known as association rule mining) is one of the significant data mining methods (Berry and Linoff, 2004) particularly focusing on exploring purchasing patterns by extracting correlations of products from transactional sales data.

The main objective of market basket analysis is to explore what kind of products are usually purchased together, which ones are rarely purchased and what is the likelihood that a customer who has bought a given product will also buy another one. (Raorane et al., 2012).

Market basket analysis focuses on association rules (Agrawal and Srikant, 1994) in the form A → B, where A and B are sets of the so-called attributes. An association rule is one of the forms , where A is an "antecedent" (if part) and B is the "consequent" (then part). Here variables A and B are the item sets and the rule( ) means that customer who purchase an item set A are expected to purchase an item set B with the probability %c, where c is called confidence (Szymkowiak et al., 2018). For association rules like A → B, it is possible to define three important measures of significance: support (Berry and Linoff, 2004), confidence (Larose, 2005) and lift (Zhang, 2002).

These measures can be expressed in terms of probability as follows:

- Support of a rule expresses the probability of the co-occurrence of A and B; it indicates the proportion of transactions in the dataset of all transactions containing A and B:

$$\frac{n(A \cap B)}{N} = P(A \cap B).$$

where N denotes the number of all transactions, and n(x) is the number of transactions containing x.

- Confidence of a rule defines the probability of event A occurring given that event B has occurred and indicates the proportion of transactions containing A, which also contain B:

$$\frac{n(A \cap B)}{n(A)} = P(B \mid A).$$

- Lift is defined as the ratio of the confidence to the marginal probability of the consequent B; when the lift of a rule is greater than 1, it means that the purchase of A increases the probability of purchasing B:

$$\frac{\text{confidence}}{P(B)} = \frac{P(B \mid A)}{P(B)},$$

Although there are a lot of appliance of market basket analysis in literature, the related up-to-date works with this project are presented by Setiabudi et al. (2011) and Musungwini et al. (2014). Setiabudi et al. (2011) present a new hybrid methodology for association rules according to case study of a grocery supermarket. In the research, market basket analysis method is implemented, where it can analyze the buying habit of the customers. The testing is conducted in a supermarket. Searching for frequent itemsets performed by Apriori algorithm (Agrawal and Srikant, 1994) to get the items that often appear in the database and the pair of items in one transaction. Pair of items that exceed the minimum support will be included into the frequent itemsets are selected. Frequent itemsets that exceed the minimum support will generate association rules after decoding. One frequent itemset can generate association rules and find the confidence, which uses a hybrid-dimension association rules. Hybrid-dimension association rules are the multidimensional association rule with repetitive predicates, which contain multiple occurrences of some predicates. The test results show, the application can generate the information about what kind of products are frequently bought in the same time by the customers according to hybrid-dimension Association Rules criteria. Results from the mining process show a correlation between the data (association rules) including the support and confidence that can be analyzed. This information will give additional consideration for owners of Minimarket X to make the further decision. In addition, Musungwini et al. (2014) suggest to investigate the role of 4P's (product, place, price, promotion) in market basket analysis and establish how the concept can be applied as a tool for competitive advantage in the retail sector according to case study of grocery retail shops in Zimbabwe.

Market Basket Analysis was first proposed by Agrawal and Sikrant in 1994 as a data mining technique using association rules. Agrawal and Sikrant proposed two new algorithms; Apriori and AprioriTID. The Apriori algorithm is an iterative algorithm which searches frequent itemsets, which are representatives of sets of items that occur together in transactions. It is also assumed that the support of a frequent itemset is equal to or greater than a certain minimum support. Frequent itemsets are used to create association rules whose confidence is greater than or equal to a predefined minimum value. The algorithm is described in detail in the article by Agrawal and Srikant (1994).

1.2. Dataset

The dataset is shared by one of the biggest retail companies of electronic devices which has a huge amount of transactional sales data.

The dataset includes 1228221 order data that occurred during April-June 2018 period. There are no NA values on dataset; It has 1782991 rows and 15 columns explained as follow:

Table 1.2.1: Explanation of Columns

Column Name	Description	Data Type
Order ID	Unique order id.	Numeric
Line number	Line number for each product on order.	Numeric
Prod ID	Unique product id of the product on order line.	Numeric
Product Name	Product name given by company.	Character
Sales Price	Sales price of product on order line.	Numeric
Brand	brand of the product.	Factor
Cat1 Name	The top level category name of the product on order line.	Factor
Cat2 Name	Second level category name of the product on order line.	Factor
Cat3 Name	Root level category name of the product on order line.	Factor
Cat1 ID	Top level category ID of the product on order line.	Numeric
Cat2 ID	Second level category ID of the product on order line.	Numeric
Cat3 ID	Root level category ID of the product on order line.	Numeric
Region	The region of the store.	Factor
City	The city which the store is located.	Factor
Period	Time period when order is occurred. Stored as "YYYYMM" format.	Factor

The relationship of columns can be explained as follow; each order can be thought as a basket that includes at least one product. And each line of order holds one product and its details (brand, category, sales price and etc.), if there is more than one product, each product gets its own line number. One order may have one or more products and each product has its own line number. Because of this, there are 1782991 rows while there are 1228221 unique orders. Therefore, concatenation of Order ID and Line Number represents a unique transaction id. Region and City represents where the store of order is located. Lastly, Period shows when order has occurred, shown as "YYYYMM" format.

1.2.1. Exploratory data analysis

The dataset is examined on exploratory data analysis phase. The aim of this phase is emphasizing significant details on dataset which are important on association rule mining.

This dataset includes 1228221 unique order and 9384 product data. As shown on Figure 1.2.1.1; the maximum product number in a single order is 2000 which is an outlier. The general distribution of first 1000 orders which have the most product number by the number of products spread around 0-50 range. Although average product amount per each order is 50.715, for 1000 the average product amount for the whole dataset is 1.76.

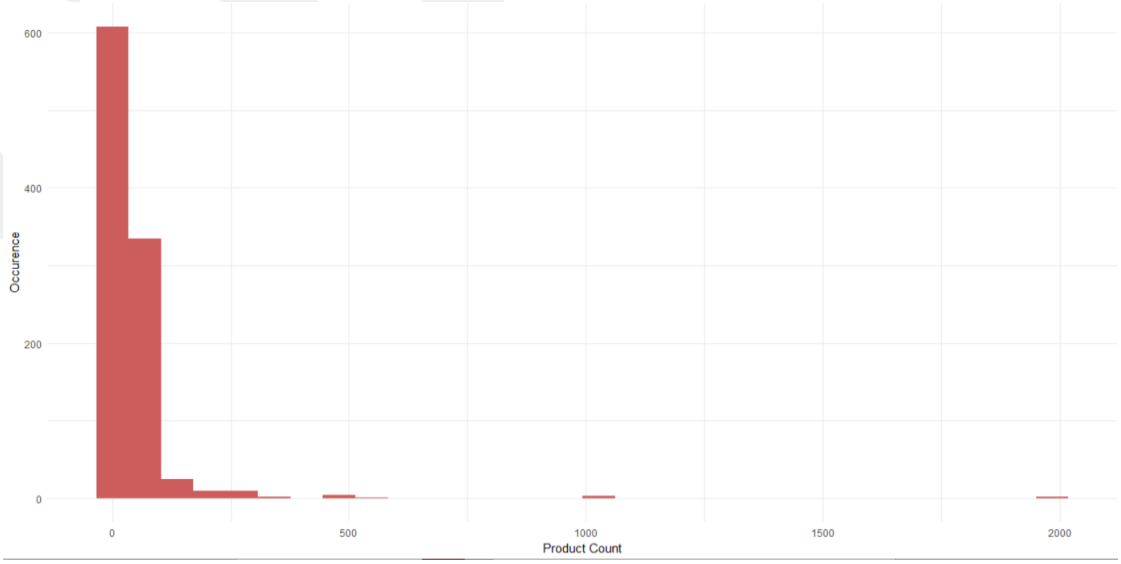

Figure 1.2.1.1: Histogram Product per 1000 Order

As shown on Figure 1.2.1.2; the maximum cost of order is around 1000000 Turkish Lira (TL). The average cost of order is 50048.31 TL for first 1000 order which have the most cost of order while it is 850.4317 TL for the whole of the dataset.

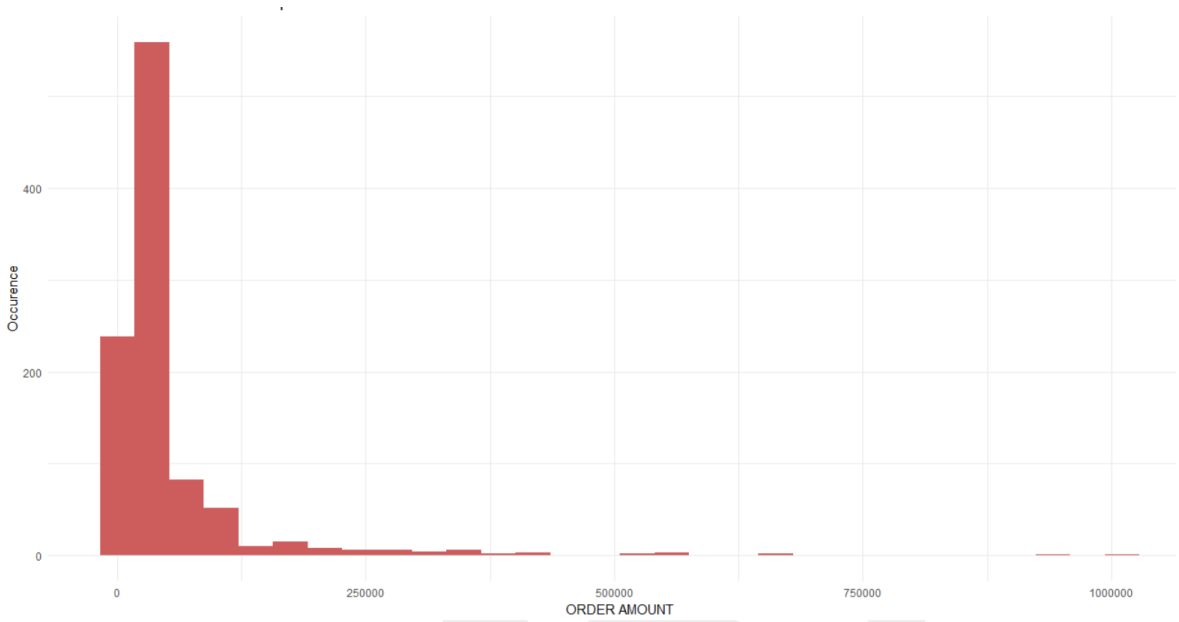

Figure 1.2.1.2: Histogram Order Amount per 1000 Order

As seen on Figure 1.2.1.3; while the total cost of orders increases throughout the given periods, the cost of order distribution is stable for each region during these three periods.

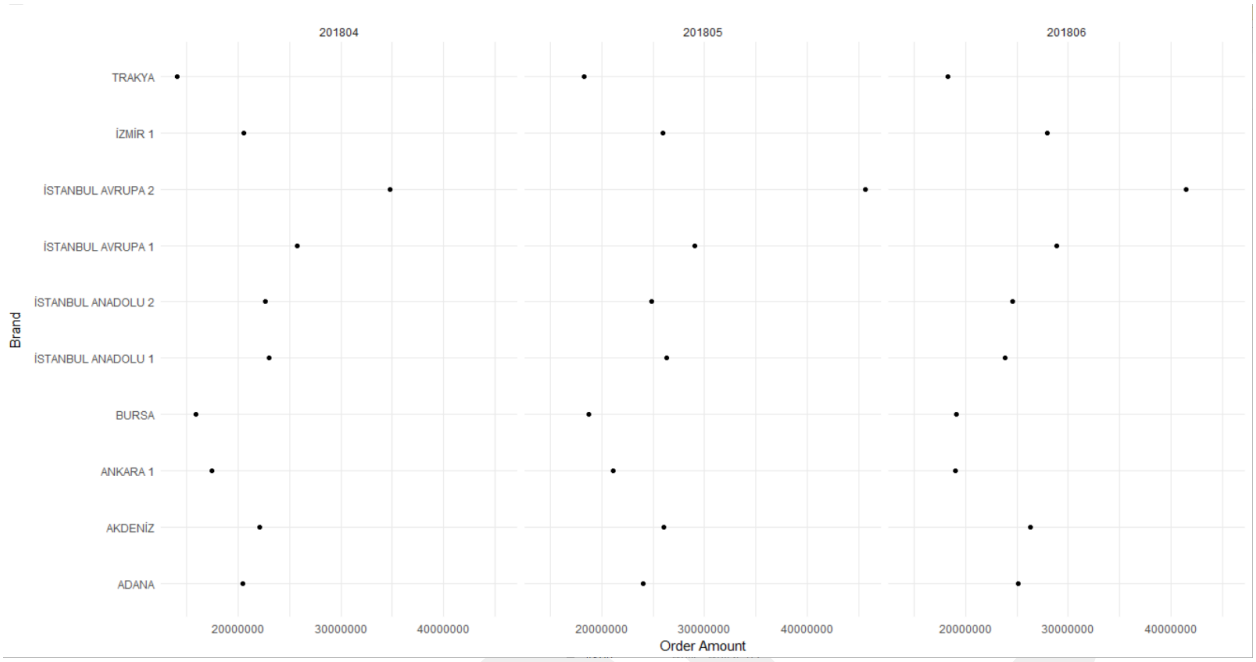

Figure 1.2.1.3: Order Amount due to Region and Period

Secondly, when product distribution on dataset is explored as seen on Table 1.2.1.4; the most purchased product is technologic insurance. Battery and smart phones are also on the 20 most purchased products list. The most significant result of the table is that the most purchased products are generally supportive products which are mainly cheap but rarely purchased alone. Therefore, according to the table below; it would be highly probable to expect the output of association rules to include some of these supportive products on both hand sides.

Table 1.2.1.4: Top 20 Products

Product Name	Count
1 YILLIK GENİŞLETİLMİŞ GARANTİ	87740
DURACELL BASIC KALEM PİL 10LU AA	12872
TOSHIBA 16GB TAŞINABİLİR BELLEK	11166
SAMSUNG G610 J7 PRIME GOLD AKILLI TELEFON	10993
APPLE 1M LIGHTNING TO USB KABLOSU (MD818ZM/A)	10930
SANDISK SDSQUNS-016G-GN3MN 80MB/S 16GB MICRO SD ANDROID HAFIZA KARTI	10472
Dayanıklı Cam Ekran Koruma iphone 6/7/8 (4.7 INCH)	10308
ININAL KART	9605
SSH-PANDA GLOBAL PROTECTION 1 KULLANICI 1 YIL	9271
TEKNOLOJİ DESTEK PAKETİ	9230
TTEC 2DK7510B MICRO USB BEYAZ DATA KABLOSU	8328
TTEC 2DK7508B IPHONE 5/6 BEYAZ DATA KABLOSU	8160
GOOGLE PLAY STORE 100	7875
HP CZ101AE 650 SİYAH MUREKKEP KARTUŞU	7475
IPHONE 6 32GB SPACE GRAY AKILLI TELEFON	7473
SEAGATE 1TB 2.5 USB 3.0 STEA1000400 EXPANSION PORTABLE SIYAH	7301
DAYANIKLI CAM EKRAN KORUMA GALAXY J7 PRIME / PRIME2	7165
DURACELL BASIC İNCE KALEM PİL 10LU AAA	6765
VARTA ALKALIN 4106229414 AA KALEM PİL	6697
TEC-SUPPORTS FIXED-3255-S EMNİYET KİLİTLİ 32-55 (82-140 CM) SABİT TV DUVAR ASKI APARATI MAX VESA:400*400 MM 50 KG SİYAH	6670

The cost of orders (cost of sales) for each region is examined and shown in the Table 1.2.1.5 below.

Table 1.2.1.5: Order Distribution amongst Regions

Region	Order Count
İSTANBUL AVRUPA 2	208616
İSTANBUL AVRUPA 1	156335
AKDENİZ	129581
İSTANBUL ANADOLU 1	127199
İSTANBUL ANADOLU 2	125537
İZMİR 1	117087
ADANA	109354
TRAKYA	99682
BURSA	91530
ANKARA 1	91349
ANKARA 2	77520
İZMİR 3	77208
TRABZON	71943
İZMİR 2	66202
GAZİANTEP	65270
KAYSERİ	58907
İZMİT	56971
SAMSUN	52700

In summary of exploratory data analysis; the maximum number of products in one order is 2000 where average is 1.76. Meanwhile the maximum cost of order is more than 1000000 TL and average cost of order is 850.4317 (TL). The cost of order distribution is stable for each region during three periods despite the increase in total cost each period. The most purchased products are generally supportive products which are cheap but rarely purchased alone. Additionally, smart phones are on the list of the most purchased products.

2. PROJECT DEFINITION

In this section, the objective and scope of the project will be discussed.

2.1. Problem Statement

Cross selling is one of the most significant parts of e-commerce. Thus, e-commerce cannot be considered without a recommendation engine. However, offering a product is also a necessity for offline sales in stores to improve selling process and profitability.

The company that shared their dataset requires to offer products automatically on its In-House sales application (In-house application is a software that is produced by a corporate entity for purpose of utilization inside the organization.). This sales application is used to create a sale workflow. From generating an order to printing a bill, sales consultants in all the stores need this native application. This sales software generates huge amounts of data from these orders. This formentioned data includes order, product, and customer information. The data specifically details category and price information. The application does not have any product recommendation module and the sale consultants need to remember all of the products and their details. This human based offering system is not reliable for cross selling and upselling. Therefore, the company needs to understand which product is purchased with which product using the native sales application's data to improve sale process.

In summary, the company needs an analysis of product association that includes which product is purchased with which product and the output of market basket analysis can solve this problem for them.

2.2. Project Objectives

The main objective of this project is to discover the association between the products and improve offline selling process with the output of the market basket analysis.

The main five steps of the project are as follows:

- Data exploratory analysis according to basket analysis strategy. In this step, dataset will be examined in detail with descriptive analysis methods.

- Secondly, data will be transformed into transactions for being input for Apriori Algorithm.

- After data transformation, Apriori algorithm will be executed with different values of parameter (support and confidence) to extract the best output. The output of the Apriori algorithm will be association rules between products.

- The best output of Apriori executions will be selected based on execution time and number of generated association rules.

- After the selection of the best Apriori algorithm according to its input parameters, the output of the algorithm -association rules- will be interpreted. The association rules will explain the relationship between products and the rules will be examined based on their metric value (confidence and lift value) to discover hidden patterns.

2.3. Project Scope

The scope of this project includes market basket analysis for offline sales of a retail company. Retail sales include electronic devices (phone, television etc.) and relevant products of electronic devices such as battery, pc mouse, additional warranty etc. Additionally, because of the huge amounts of data, market basket analysis on this work focuses solely on sales that had occurred between April 2018 and June 2018.

3. METHODOLOGY

The methodology of the project aims to discover association rules between products by Apriori algorithm with R Language (Chapman and Feit, 2016).

R is a programming language and a free software environment which is widely used for data analysis and it is supported by the R Foundation for Statistical Computing .(“What is R?”, n.d) It is a GNU package and the source code for R software environment is written primarily in C Language.

R is an integrated suite of software facilities for data analysis and graphical display. The capabilities of R are extended through user-created packages, which allow specialized statistical techniques, machine learning techniques, import/export capabilities and graphical devices.These packages are developed mostly in R. As of May 2018, there are more than 12500 additional packages available at the Comprehensive R Archive Network.

3.1. System Properties

Algorithm performance is measured by execution time, consequently, system properties are significant. This project executes on a personal computer which has 64-Bit Windows 10 OS with Intel® Core™ i7-7700HQ CPU @ 2.80GHz processor. Its installed memory capacity is 16.0 GB.

3.2. General Definition of Association Rules

Rule-based methods are a popular class of techniques in machine learning and data mining (Fürnkranz and Kliegr, 2011). They share the goal of finding regularities in data that can be expressed in the form of an IF-THEN rule. A → B is an example of IF-THEN rule where the realization of B depends on the realization of A's. Depending on the type of rule that should be found, it can be discriminated between descriptive rule discovery (Fürnkranz and Kliegr, 2011) which aims at describing significant patterns in the given dataset in terms

of rules, and predictive rule learning (Fürnkranz and Kliegr, 2011). Whereas descriptive rule discovery aims at finding individual rules that capture some regularities and patterns of the input data, the task of predictive rule learning is to generalize the training data so that predictions for new examples are possible.

In line with the problem statement and project objectives, a descriptive rule discovery approach rather than a predictive rule learning is preferred. The descriptive rule discovery algorithms aim to have the ultimate precision in detection of association rules which would not be possible with a predictive rule learning method such that support vector machines.

The general descriptive rule discovery approach to market basket analysis is using association rules algorithms. Association rules are widely used to analyze retail basket or transaction data, and are intended to identify strong rules in transaction data using measures of interestingness, based on the concept of strong rules (Han et al., 2012). These algorithms take transactional data and give the associations between products as output. The best property of this approach is the clarity and versatility of the results, which are in the form of rules about the products. There is an intuitive approach to an association rule because it expresses how tangible products and services group together ie. ”if a specific product is purchased with another specific product frequently, then the second product may be purchased with first product.” (Prasad and Mourya, 2013).

Association rules for market basket analysis is a technique to uncover how items are associated with each other. The three most commonly used methods to measure association is as shown below:

- **Support**: Support shows how popular an itemset is, as measured by the proportion of transactions in which an itemset appears.
- **Confidence**: Confidence how likely item Y is purchased when item X is purchased, expressed as {X -> Y}. This is measured by the proportion of transactions with item X, in which item Y also appears.
- **Lift**: Lift shows how likely item Y is purchased when item X is purchased, while considering the popularity of the item Y. A lift value greater than 1 means that item

Y is likely to be bought if item X is bought, while a value less than 1 means that item Y is unlikely to be bought if item X is bought (substitute products). Lift summarizes the strength of association between the products on the left and right-hand side of the rule; the larger the lift the greater the link between the two products. (Tan et al., 2004)

To explain the metrics of association rules clearly, the following example is given. Suppose that Table 3.2.1 shows the baskets and products for each basket. Table 3.2.2 indicates the calculation of metrics (support, confidence and lift) for each rule.

Table 3.2.1: Basket Samples

Baskets	Products
Basket 1	A,B,C
Basket 2	A,C,D
Basket 3	B,C,D
Basket 4	A,D,E
Basket 5	B,C,E

s

Table 3.2.2: Calculated Metrics for Association Rules

Rule	Support	Confidence	Lift
C-->A	2/5	1/2	5/6
A-->C	2/5	2/3	5/6
B&C-->D	1/5	1/3	5/9

With regards to the dataset that focuses on only three months, appearance of products on the dataset may have low frequency. Therefore, support threshold value should be less than what the literature suggests generating more association rules. Additionally; because of the same reason, confidence threshold may have less value than suggested on literature. Because of the formulation of the lift; when support threshold is less, lift metric may be extremely high for the association rules where it summarises the strength of association. Therefore, lift value vary according to dataset because it is used to compare the rules in a rule set.

3.3. Apriori Algorithm for Market Basket Analysis

3.3.1. Apriori algorithm

In computer science and data mining, Apriori is a classical algorithm for learning association rules. (“The Apriori Algorithm”, 2015) Apriori is designed to govern transactions. As is common in association rule mining, given a set of itemsets, the algorithm attempts to find subsets which are common to at least a minimum number C of the itemsets. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found.

Apriori works in two steps:

1. Systematically identify item sets that occur frequently in the data set with a support greater than a pre-specified threshold.
2. Calculate the confidence of all possible rules given the frequent item sets and keep only those with a confidence greater than a pre-specified threshold.

The thresholds at which to set the support and confidence are user-specified and are likely to vary between transaction data sets.

3.3.2. Library for Apriori Algorithm in R

Market basket analysis is developed with R Language. R community has a library for association rules that’s called “arules” (Hahsler et al., 2018). Additionally, to visualize the output of algorithms “arulesViz” (“Package arulesViz”, 2018) and “plotly” libraries will be used. Apriori algorithm from “arules” package takes two inputs: support and confidence and gives association rules as output.

3.3.3. Apply Apriori Algorithm in R

First, to apply Apriori Algorithm in R, prior to using the Apriori algorithm, data must be transformed into transactions such that all the items purchased together are shown in one row (Marafi, 2014). For this transformation; Order ID and Prod Name columns are enough. While the given dataset includes 1782991 rows, this transformation takes 757.91 seconds and 1226992 transactions (same value as the unique order number) are generated. Then, to simplify the process; these transaction set is written into market_basket.csv file. The writing process takes 2.72 seconds. Afterwards, this file is read in 385.64 seconds. Summary of transactions generated by reading file expresses that there are 1226992 rows (amount of unique orders on dataset) and 9384 columns (number of unique products in dataset). Most frequent item is "1 Yıllık Genişletilmiş Garanti" which is discovered on exploratory data analysis part and density is 0.0001443. Density represents the total number of items that are purchased divided by the total number of possible items in that matrix. Considering density is useful to select "support" threshold for Apriori algorithm. Additionally, summary of transactions indicates that there are 913708 transactions that includes just 1 item and 228669 transactions that includes 2 items, 60711 transactions that has 3 items and all the way up to the biggest transaction: 1 transaction that has 46 items.

After generating the transaction set, Apriori algorithm executes with the parameters: support=0.0001 and confidence=0.7. These values of parameters indicate thresholds for the expected output -association rules-. It takes 1.79 seconds and returns 117 rules that have a support of at least 0.01% and confidence of at least 70%. The rules are ordered by lift because higher lift shows the strength of association. According to summary of association rules generated by Apriori shown as Figure 3.3.3.1; a length of 2 products (2 products in one order) has the most rules, 60 rules. Length of 3 products has 46 rules and length of 4 products has 11 rules.

```
set of 117 rules

rule length distribution (lhs + rhs):sizes
 2  3  4
60 46 11

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   2.000   2.581   3.000   4.000

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.0001011   Min.   :0.7000   Min.   :  10.12   Min.   : 124.0
 1st Qu.:0.0001345   1st Qu.:0.8448   1st Qu.: 105.94   1st Qu.: 165.0
 Median :0.0002135   Median :0.9270   Median :1755.31   Median : 262.0
 Mean   :0.0005075   Mean   :0.9078   Mean   :2624.96   Mean   : 622.7
 3rd Qu.:0.0004214   3rd Qu.:1.0000   3rd Qu.:4366.52   3rd Qu.: 517.0
 Max.   :0.0043700   Max.   :1.0000   Max.   :9895.10   Max.   :5362.0

mining info:
 data ntransactions support confidence
   tr       1226993  0.0001        0.7
```

Figure 3.3.3.1: Summary of Rules generated by Apriori Algorithm (support=0.0001 and confidence=0.7)

Figure 3.3.3.2 shows the summary of the rules by confidence, support and lift. This figure illustrates the relationship between the different metrics. It has been shown that the optimal rules generate “support-confidence boundary”. Optimal rules reveal where support, confidence or both metrics are maximized. Although there is a boundary on graph, it is seen that there are less rules when compared the number of products which is over 9000. Therefore, different confidence and support values are tried to explore more rules by Apriori.

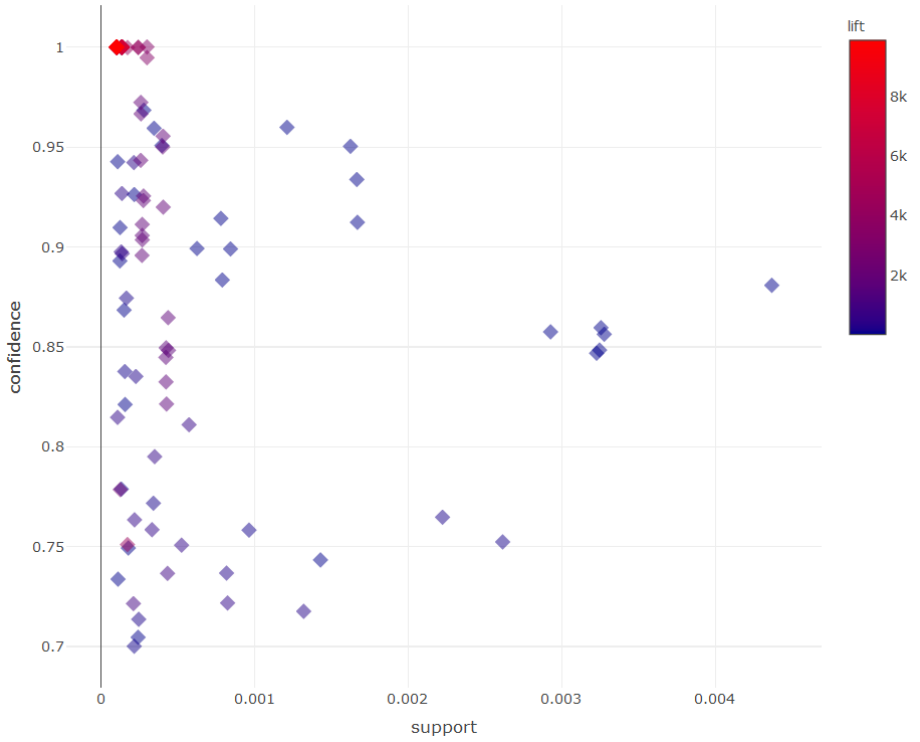

Figure 3.3.3.2: Distribution of Rules generated by Apriori Algorithm (support=0.0001 and confidence=0.7)

The first execution of Apriori generates few rules than expected. This “Few Rules Problem” causes Apriori to execute with the parameters given below: support=0.00001 and confidence=0.7. It takes 2.08 seconds and returns 913 rules that has a support of at least 0.001% and confidence of at least 70%. Then, support=0.00001 and confidence=0.6 are given as input to Apriori and this trial generates 1126 rules in 2.08 seconds which is more acceptable than previous trials due to business knowledge.

```
set of 1126 rules

rule length distribution (lhs + rhs):sizes
  2   3   4   5
282 668 167   9

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.250   3.000   2.914   3.000   5.000

summary of quality measures:
    support            confidence          lift               count
 Min.   :0.00001059   Min.   :0.6000   Min.   :    8.62   Min.   :  13.0
 1st Qu.:0.00001467   1st Qu.:0.7341   1st Qu.:  129.75   1st Qu.:  18.0
 Median :0.00002445   Median :0.8992   Median :  746.02   Median :  30.0
 Mean   :0.00008192   Mean   :0.8579   Mean   : 7868.18   Mean   : 100.5
 3rd Qu.:0.00004646   3rd Qu.:1.0000   3rd Qu.: 9435.25   3rd Qu.:  57.0
 Max.   :0.00437003   Max.   :1.0000   Max.   :94384.08   Max.   :5362.0

mining info:
 data ntransactions support confidence
   tr       1226993 0.00001        0.6
```

Figure 3.3.3.3: Summary of Rules generated by Apriori Algorithm (support=0.00001 and confidence=0.6)

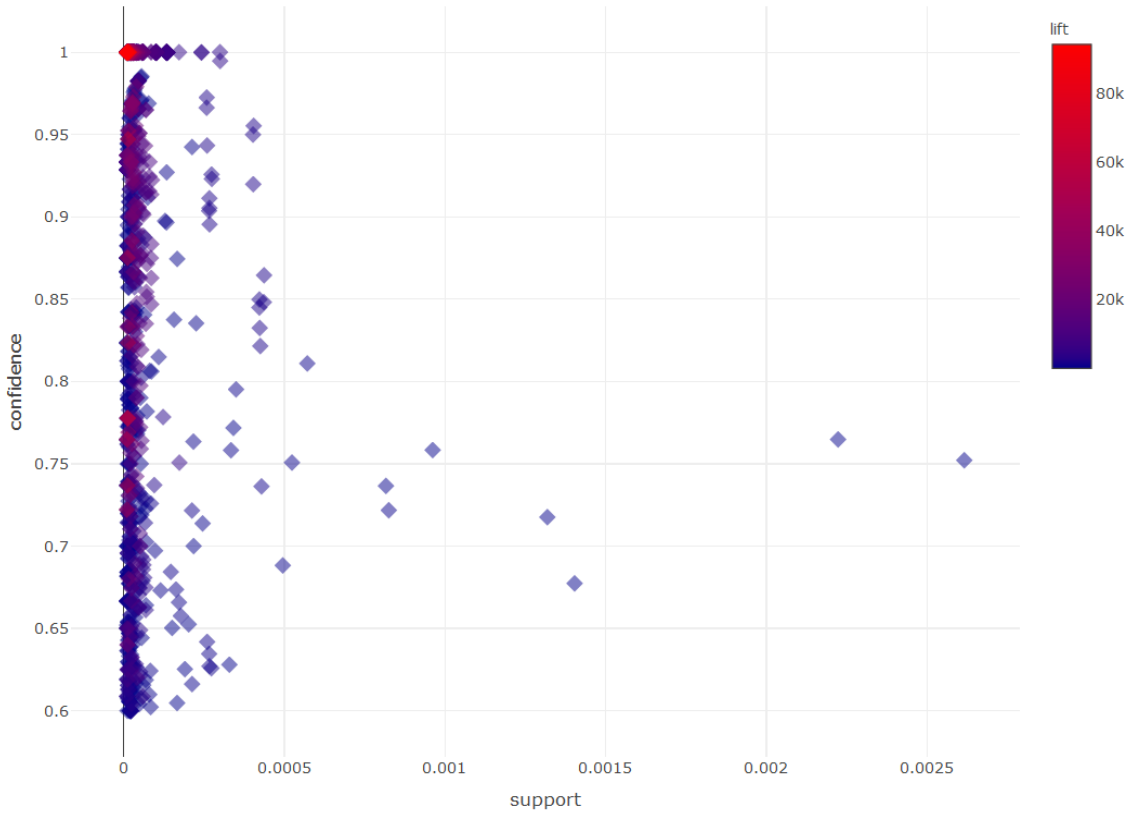

Figure 3.3.3.4: Distribution of Rules generated by Apriori Algorithm (support=0.00001 and confidence=0.6)

Figure 3.3.3.4 displays the distribution of 1126 rules among support and confidence. Decreasing support and confidence value causes stricter and more comprehensible "support-confidence boundary". Additionally, it is shown that on the figure, most of the rules are spread on [0.0001,0.0005] interval of support. When confidence is between 0.7 and 0.8, there are some significant outlier rules. These outlier rules may be generated because of a campaign such that "take this product and this one is free" in this period.

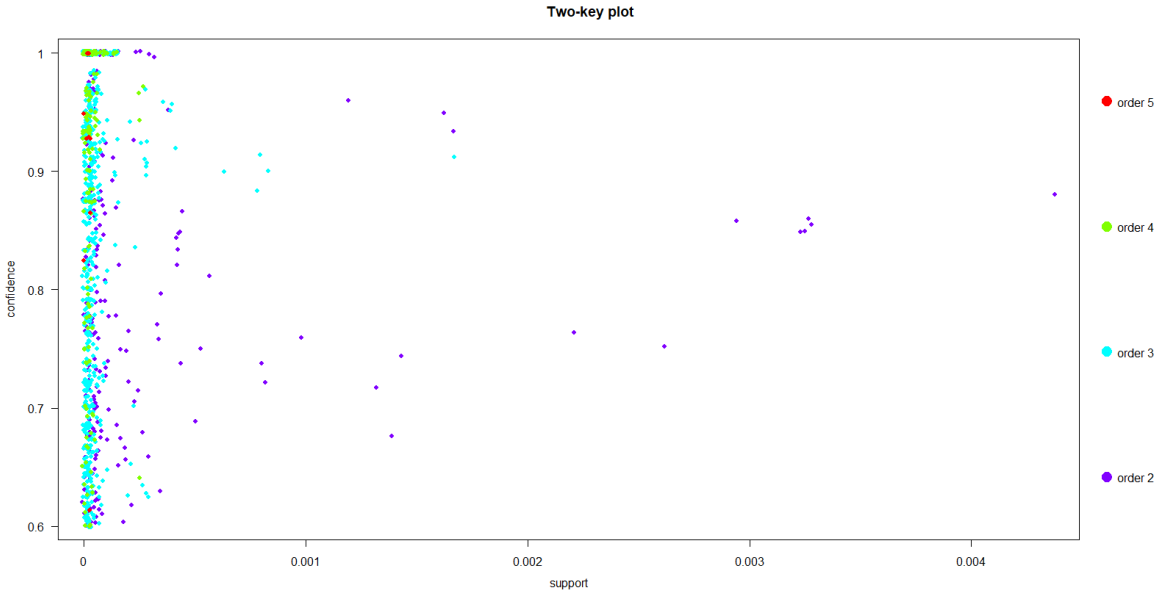

Figure 3.3.3.5: Generated Rules due to Number of Item in One Transaction (support=0.00001 and confidence=0.6)

As shown on Figure 3.3.3.5; the left side of support-confidence boundary includes all the different amounts of order.

4. EVALUATION & OUTCOMES

4.1. Outcomes of Apriori Algorithm applied on All Datasets

The purpose of this analysis is to generate a set of rules that link two or more products together. There are three significant metrics to interpret the output of Apriori; support, confidence and lift. Support shows how popular an item is, as measured by the proportion of transactions in which an item appears. Moreover, confidence expresses how likely an item on the right-hand side (RHS) of the rule is purchased when an item on the left hand side (LHS) is purchased. Confidence is measured by the proportion of transactions with LHS item, in which RHS item also appears. Therefore, higher confidence rules are the ones where there is a higher probability of items on the righthand side being a part of the transaction, given presence of the items on the left-hand side. In addition, lift explains how likely RHS item would be purchased when LHS item is also purchased, while controlling for how popular RHS item is. Each of the generated rules should have a lift greater than one for complementary products, while lift explains how strong the association between products is. If lift is equal to one, the products are not associated and if lift is less than 1, the products are substitutes.

The last execution of Apriori with parameters: support=0.00001 and confidence=0.6 generates 1126 rules in 2.08 seconds. The execution time is reasonable for a dataset as large as this. Furthermore, based on business knowledge and definition of metrics; support and confidence threshold are faultless. The dataset includes orders (baskets) of only three months, so, frequency of occurrence of an item or itemset may be low during this period for electronic items like the ones in this dataset. Therefore, support threshold (0.00001) and confidence threshold (0.6) are acceptable where support shows how popular an item and confidence express how likely an item on the righthand side (RHS) of the rule is purchased when item on left hand side (LHS) is also purchased. To sum up; when number of generating rules, support and confidence threshold and execution times are considered, the output of this Apriori execution will be interpreted as below.

4.2. Outcomes of Apriori Algorithm Based on Confidence of The Rules

As shown on Figure 3.3.3.3; the minimum value of lift when support threshold (0.00001) and confidence threshold (0.6) is 8.62 and maximum value is 94384.08 and mean is 7868.18. For any rule of generated ruleset has a lift value much greater than 1. It indicates that antecedent (LHS of the rule) and consequent (RHS of the rule) appear more often together than expected, this means that the occurrence of the antecedent has a positive effect on the occurrence of the consequent. As a result of that, all generated rules by Apriori shows products which purchased together, a.k.a, complementary products. The rules which have higher lift values indicate bundle products, which are forced to be purchased together. This analysis aims to find hidden associations -complementary- instead of bundle products. Because of this reason, generated association rules are mainly examined based on their confidence.

In addition, Figure 3.3.3.4 shows that most of the generated rules are spread on [0.0001,0.0005] interval of support but the distribution of rules vary due to confidence. Therefore, the rules are examined on the next chapter part by part due to their confidence interval and to make visualization easier, the top ten rules for each confidence interval are selected according to their confidence value.

4.2.1. The rules that have confidence between 0.6 and 0.7

Figure 4.2.1.1 is a graph-based visualization with items and top ten rules based on lift where confidence of the rules is between 0.6 and 0.7. On this graph type, the size of the bubbles represents the support and the color of the bubbles represent the lift value. The direction of arrow is from the left-hand side of the rule to the right-hand side of the rule.

On this confidence interval, association rules include mainly smart phones on left hand side and supportive products of smart phones on the right-hand side. The graph shows that when a smart phone is purchased, the phone case or additional warranty is purchased.

The important outcome of the graph is that brands of supportive products vary depending on the smart phone's brand.

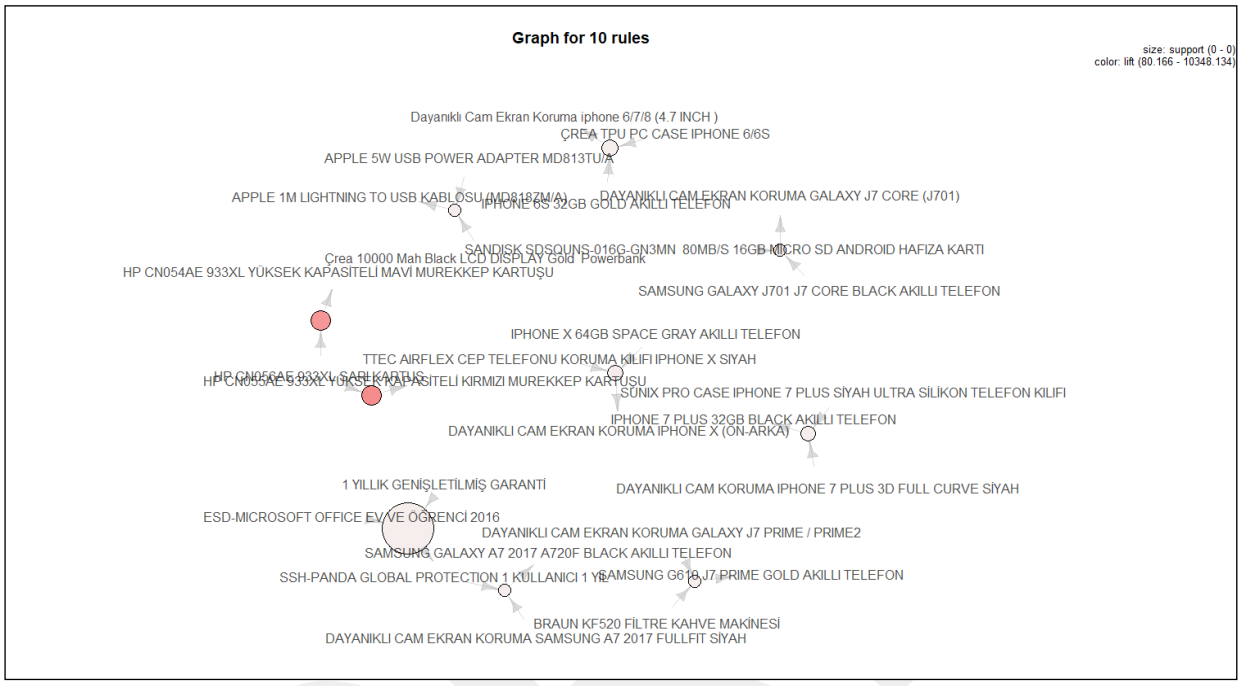

Figure 4.2.1.1: Graph of Generated Rules that have Confidence between 0.6 and 0.7

4.2.2. The rules that have confidence between 0.7 and 0.8

Table 4.2.2.1 is the tabular representation of the below graph. It is clearly seen that the rules that has a lift value of more than 100 including bundle products whereas the rules that has a lift value of less than 100 and a confidence interval between 0.7 and 0.8 are the hidden association rules between the products.

Table 4.2.2.1: Table of Rules that have a Confidence Value between 0.7 and 0.8

RULES	SUPPORT	CONFIDENCE	LIFT
{HP CZ110AE 655 MAVİ KARTUŞ} => {HP CZ111AE 655 KIRMIZI MUREKKEP KARTUŞU}	0.00005135	0.7974684	13403.94642
{PHILIPS GC9650/80 ELITE BUHAR KAZANLI UTU} => {PHILIPS ÜTÜ MASASI}	0.00034801	0.7951583	504.21377
{1 YILLIK GENİŞLETİLMİŞ GARANTİ,APPLE MQD32TU/A MacBook Air 13 i5 1.8GHz/8GB/128GB flash/HDG,SSH-PANDA GLOBAL PROTECTION 1 KULLANICI 1 YIL} => {ESD-MAC İÇİN MICROSOFT OFFICE EV VE ÖĞREN(	0.00002527	0.7948718	2311.14248
{DURACELL BASIC KALEM PİL 10LU AA,HP F6V24AE 652 RENKLI MUREKKEP KARTUŞU} => {HP F6V25AE 652 SİYAH MÜREKKEP KARTUŞU}	0.00001875	0.7931034	167.49266
{HUAWEI P SMART BLUE AKILLI TELEFON,LENOVO TAB 7 ESSENTİAL BLACK 7 1GB/8GB/WIFI TABLET} => {1 YILLIK GENİŞLETİLMİŞ GARANTİ}	0.00001875	0.7931034	11.39446
{FAKİR TRINITY BLENDER SET (KIRMIZI),PREMİER PTM 3505 TOST MAKİNESİ} => {PREMİER PTC 2016 ELEKTRİKLİ CEZVE}	0.00001549	0.7916667	314.46082
{APPLE 5W USB POWER ADAPTER MD813TU/A,APPLE MNHF2TU/A EARPODS 3.5MM BAĞLANTILI MİKROFONLU KULAKLIK} => {APPLE 1M LIGHTNING TO USB KABLOSU (MD818ZM/A)}	0.00003097	0.7916667	91.72516
{DAYANIKLI CAM EKRAN KORUMA GALAXY A6} => {1 YILLIK GENİŞLETİLMİŞ GARANTİ}	0.00006194	0.7916667	11.37382
{SAMSUNG GALAXY A6+ A605F BLUE AKILLI TELEFON} => {1 YILLIK GENİŞLETİLMİŞ GARANTİ}	0.00008639	0.7910448	11.36488
{HP 953XL F6U16AE CYAN YUKSEK KAPASITELI MUREKKEP KARTUSU (1.600 SAYFA)} => {HP 953XL F6U18AE SARI YUKSEK KAPASITELI MUREKKEP KARTUSU (1.600 SAYFA)}	0.00003994	0.7903226	14692.73143

As shown on Table 4.2.2.2; distribution of rules which has a confidence value between 0.7 and 0.8 is like the previous confidence interval. However, as seen on Figure 3.3.3.4, there are some outlier rules according to their high support values on this confidence interval. These outlier rules can be examined on Table 4.2.2.2. This outlier rules are found to be generated by a promotional campaign on that period. Therefore, there are some "bundle" products because of the promotional campaigns. For instance, when iron of a specific brand is purchased, ironing board of the same brand is also purchased. This rule has high support value due to bubble size and high lift value due to bubble color. So, irons and ironing boards are purchased together during this campaign period.

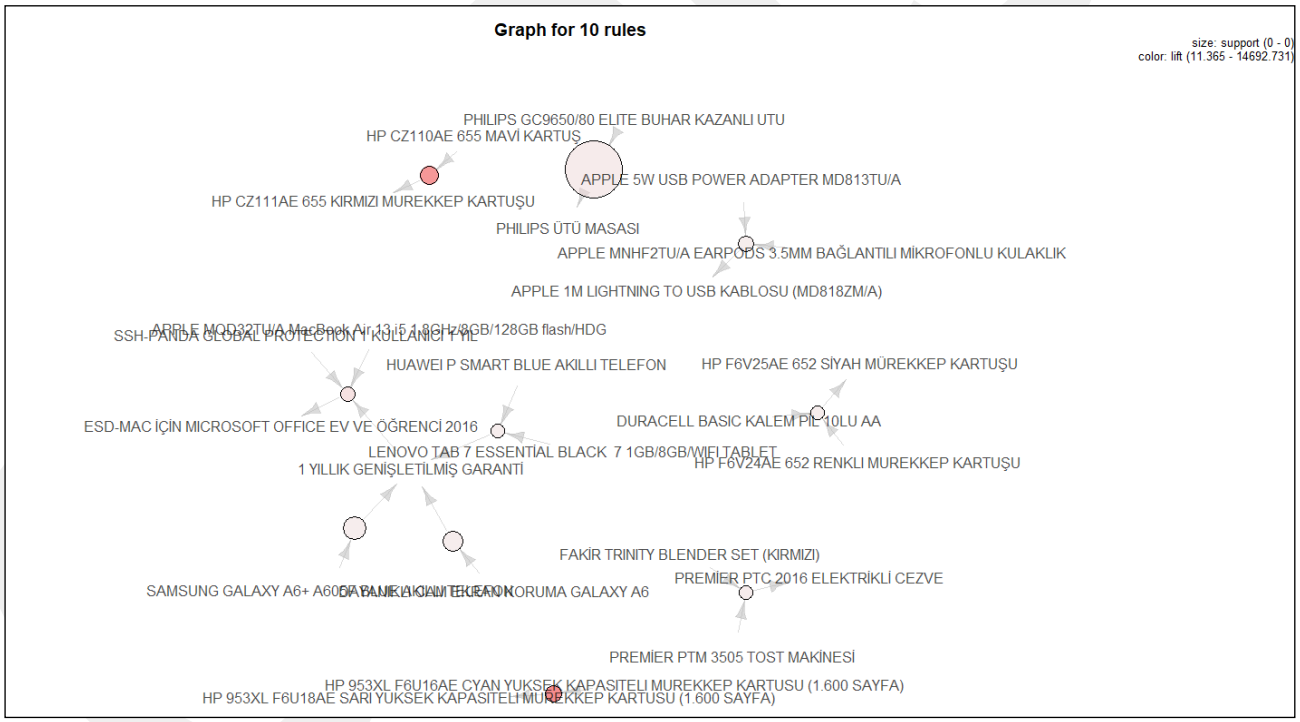

Figure 4.2.2.2: Graph of Generated Rules that have a Confidence value between 0.7 and 0.8

4.2.3. The rules that have confidence between 0.8 and 0.9

Table 4.2.3.1 shows that the rules that have less than 200 lift value are complementary products when their confidence is between 0.8 and 0.9 whereas other rules indicate bundle products.

Table 4.2.3.1: Table of Rules That have Confidence between 0.8 and 0.9

RULES	SUPPORT	CONFIDENCE	LIFT
{DAYANIKLI CAM KORUMA HUAWEI P20 LITE,HUAWEI P20 LITE BLACK AKILLI TELEFON} => {1 YILLIK GENİŞLETİLMİŞ GARANTİ}	0.00062511	0.8991794	12.91844
{DAYANIKLI CAM EKRAN KORUMA HUAWEI P SMART,HUAWEI P SMART GOLD AKILLI TELEFON} => {1 YILLIK GENİŞLETİLMİŞ GARANTİ}	0.00084271	0.8991304	12.91774
{1 YILLIK GENİŞLETİLMİŞ GARANTİ,DAYANIKLI CAM EKRAN KORUMA SAMSUNG A7 2017 FULLFIT SİYAH} => {SAMSUNG GALAXY A7 2017 A720F BLACK AKILLI TELEFON}	0.00012877	0.8977273	423.33016
{1 YILLIK GENİŞLETİLMİŞ GARANTİ,PREMİER PTC 2016 ELEKTRİKLİ CEZVE} => {PREMİER PTM 3505 TOST MAKİNESİ}	0.00013448	0.8967391	488.36779
{BRAUN KF520 FİLTRE KAHVE MAKİNESİ,PHILIPS GC8750/60 PERFECT CARE BUHAR KAZANLI ÜTÜ} => {PHILIPS ÜTÜ MASASI}	0.00002119	0.8965517	568.50785
{DAYANIKLI CAM EKRAN KORUMA GALAXY A6,SAMSUNG GALAXY A6 A600F BLUE AKILLI TELEFON} => {1 YILLIK GENİŞLETİLMİŞ GARANTİ}	0.00002119	0.8965517	12.88069
{HP 953XL F6U16AE CYAN YUKSEK KAPASITELI MUREKKEP KARTUSU (1.600 SAYFA),HP 953XL F6U17AE MAGENTA YUKSEK KAPASITELI MUREKKEP KARTUSU (1.600 SAYFA)} => {HP	0.00003505	0.8958333	16654.26105
{ÇREA TPU PC CASE SAMSUNG GALAXY J7 PRIME,SAMSUNG GALAXY J7 PRIME2 G611F GOLD TELEFON} => {DAYANIKLI CAM EKRAN KORUMA GALAXY J7 PRIME / PRIME2}	0.00003505	0.8958333	158.01915

As seen on Figure 4.2.3.2, when confidence of rules increases, the graph becomes more meaningful. Additionally, rules with more than two items are clearly seen on this confidence interval. Unrelated to support value and category of the products, the most purchased product, additional warranty (“1 Yıllık Genişletilmiş Garanti”), is seen on both side of the rules. This product is purchased if a smart phone and a phone case is purchased together. Therefore; although the brand of products differs on each rule, smart phone, additional warranty and phone case are complementary products according to generated association rules.

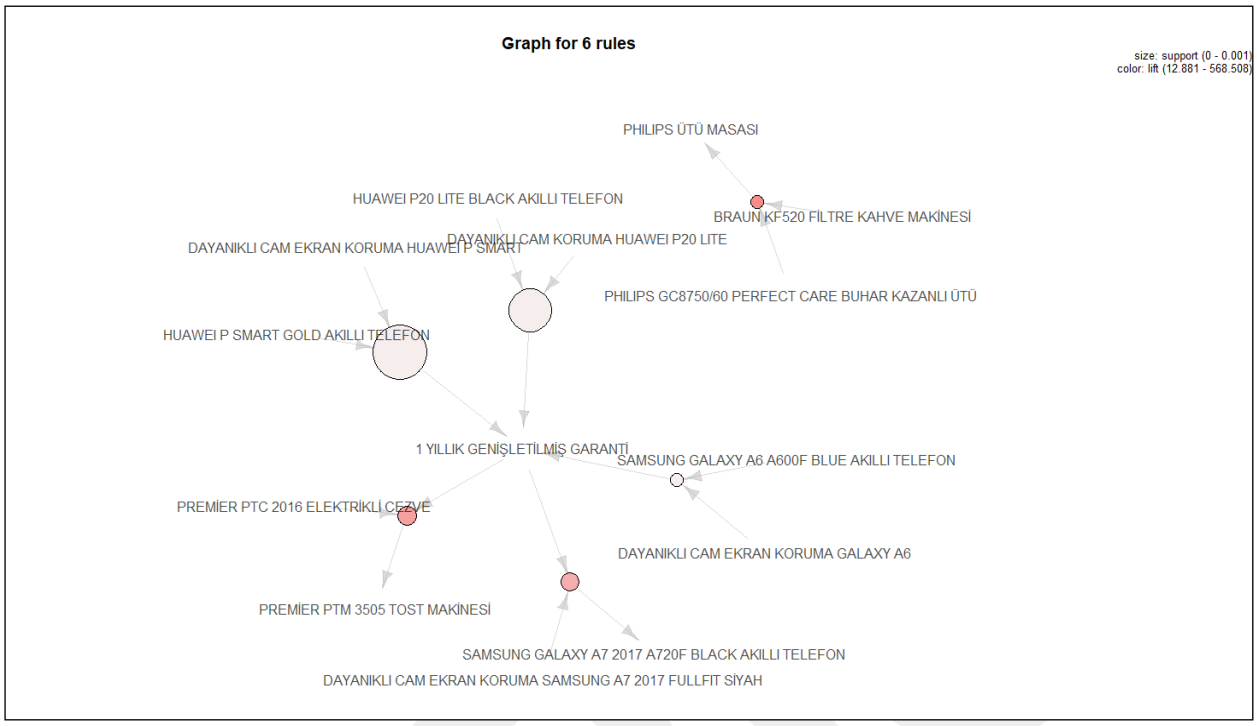

Figure 4.2.3.2: Graph of Generated Rules That have Confidence between 0.8 and 0.

4.2.4. The rules that have confidence between 0.9 and 1

It is clearly seen on Table 4.2.4.1; when confidence is between 90% and %100, the lift of the rules increases extremely. It points out complementary products where lift is less than 100 and confidence is between 0.9 and 1.

Table 4.2.4.1: Table of Rules That have Confidence between 0.9 and 1

rules	support	confidence	lift
{SAMSUNG F-FRAME40 FRAME} => {SAMSUNG FRH1438Z4 BAKIR BORU}	0.00029911	0.9945799	3325.18428
{HUAWEI P SMART GOLD AKILLI TELEFON,SSH-PANDA GLOBAL PROTECTION 1 KULLANICI 1 YIL} => {1 YILLIK GENİŞLETİLMİŞ GARANTİ}	0.00005461	0.9852941	14.15565
{DAYANIKLI CAM KORUMA HUAWEI P20 LITE,SSH-PANDA GLOBAL PROTECTION 1 KULLANICI 1 YIL} => {1 YILLIK GENİŞLETİLMİŞ GARANTİ}	0.00005216	0.9846154	14.1459
{SAMSUNG AR12MSFSCWKXSK DIŞ ÜNİTE KLİMA} => {SAMSUNG F-FRAME40 FRAME}	0.00004646	0.9827586	3267.85352
{SAMSUNG AR12MSFSCWKNSK İÇ ÜNİTE KLİMA} => {SAMSUNG F-FRAME40 FRAME}	0.00004646	0.9827586	3267.85352
{SAMSUNG AR12MSFSCWKNSK İÇ ÜNİTE KLİMA,SAMSUNG AR12MSFSCWKXSK DIŞ ÜNİTE KLİMA} => {SAMSUNG F-FRAME40 FRAME}	0.00004646	0.9827586	3267.85352
{CANON GI-490 BK SIYAH KARTUS 6000 SAYFA,CANON GI-490 M KIRMIZI KARTUS 7000 SAYFA,CANON GI-490 Y SARI KARTUS 7000 SAYFA} => {CANON GI-490 C MAVI KARTUS 7000 SAYFA}	0.00004564	0.9824561	11703.56123
{SAMSUNG AR12MSFSCWKXSK DIŞ ÜNİTE KLİMA,SAMSUNG F-FRAME40 FRAME} => {SAMSUNG FRH1438Z4 BAKIR BORU}	0.00004564	0.9824561	3284.6507
{SAMSUNG AR12MSFSCWKNSK İÇ ÜNİTE KLİMA,SAMSUNG F-FRAME40 FRAME} => {SAMSUNG FRH1438Z4 BAKIR BORU}	0.00004564	0.9824561	3284.6507
{SAMSUNG AR12MSFSCWKNSK İÇ ÜNİTE KLİMA,SAMSUNG AR12MSFSCWKXSK DIŞ ÜNİTE KLİMA,SAMSUNG F-FRAME40 FRAME} => {SAMSUNG FRH1438Z4 BAKIR BORU}	0.00004564	0.9824561	3284.6507

When Figure 4.2.4.2 is examined; bundle products are easily seen on the graph. The network on right-top of graph shows a bundle product of an air conditioner's different parts. The lift of each rule that generates this network almost has the same value. Therefore, these products cannot be purchased alone. The metrics are similar for the network on the bottom of the graph. Despite that, the network on the left of graph with rules that have less lift value indicating complementary products.

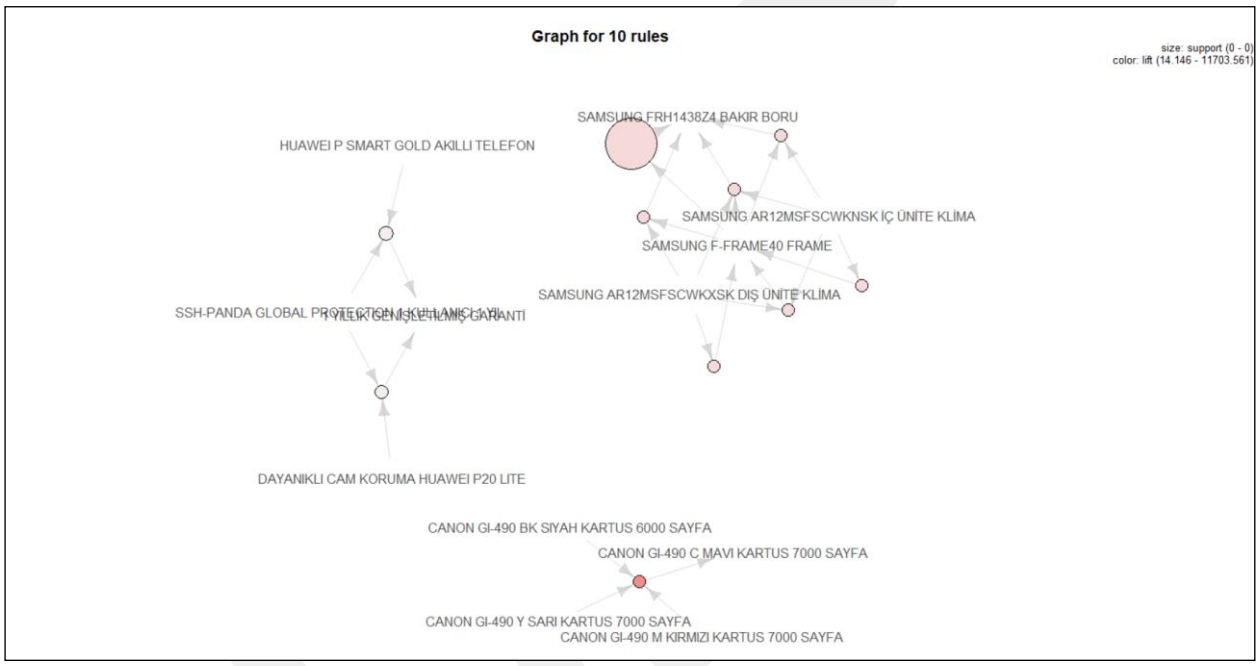

Figure 4.2.4.2: Graph of Generated Rules That have Confidence between 0.9 and 1

4.3. Outcomes of Apriori Algorithm Based on Lift of The Rules

Outcomes of Apriori Algorithm based on confidence of the rules shows that there is a lift threshold value for association rules. This threshold value represents the boundary between bundle products (which are forced to be purchased together) and complementary products. When generated association rules are examined according to their confidence, it is clearly seen that the rules that have a lift value of more than 200 indicate bundle products whereas the rules that have a lift value of less than 200 indicate complementary products. Thus, association rules generated by Apriori algorithm (executed with support threshold:

0.00001, confidence threshold=0.6) are split depending on their lift values to examine complementary products.

After splitting, 333 association rules are left to examine. Figure 4.3.1.1 displays these association rules. Bubbles are rules, their color represents their lift value and their size represents their confidence value. Blue sticks represent different products.

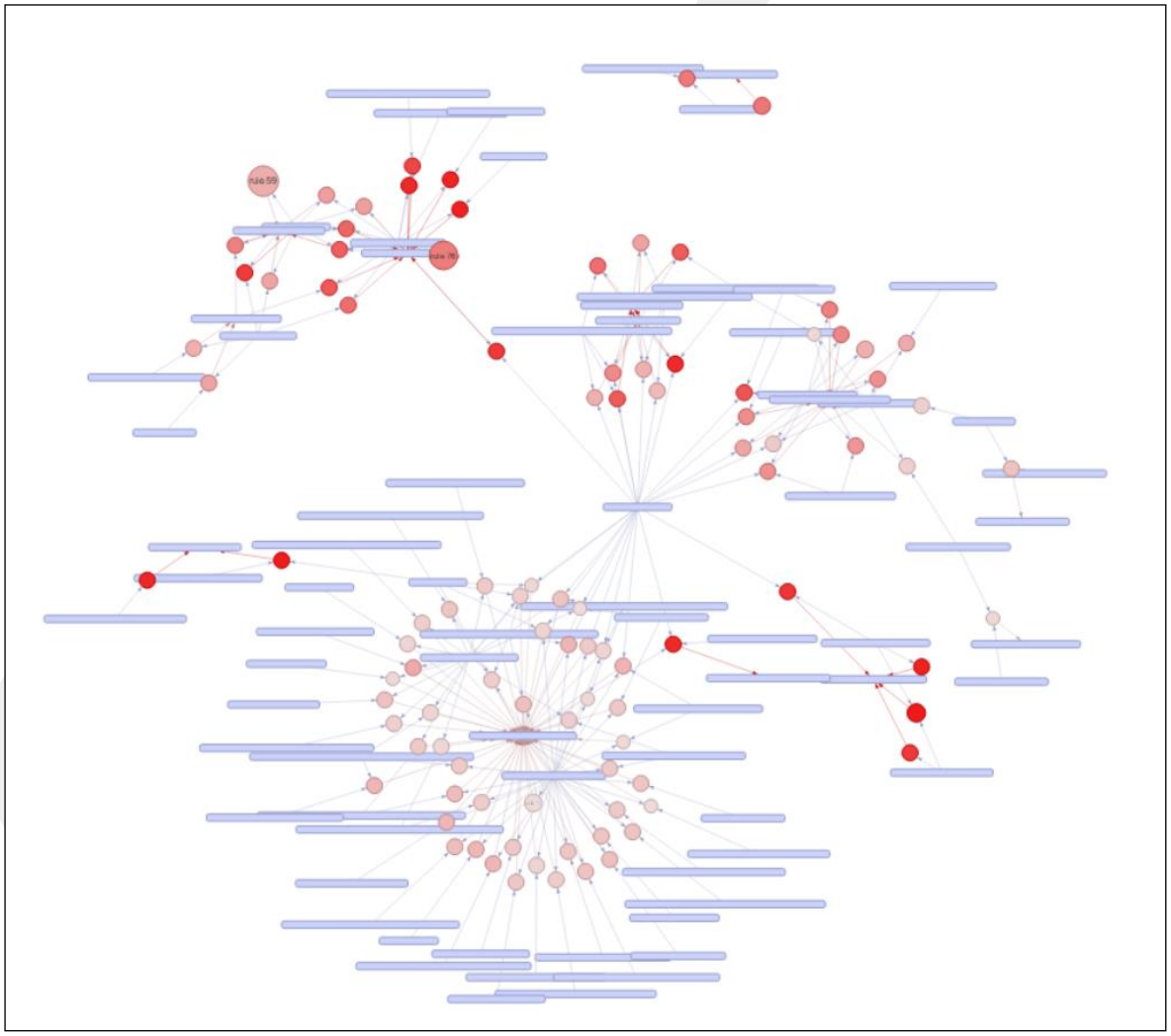

Figure 4.3.1.1: Graph of 333 Rules

Although the visualization is not detailed on Figure 4.3.1.1; the generated graph can be examined on R Studio. The first result of this graph is that the virtual products (products

which has no physical stock in the store such that additional warranty, antivirus, technological support) are the most reliable complementary product for electronic devices. Depending on product category in the basket, these supportive products should be offered while considering the association rules which include the main product.

Figure 4.3.1.2 is the part of Figure 4.3.1.1 and it examines the top ten rules based on their confidence value. As shown on Figure 4.3.1.2; if there is a smart phone or tablet in the basket, additional warranty must be offered. In addition, if there is additional warranty in the basket, phone case should be offered. Interestingly, if there is phone case and shaving machine together in the basket; additional warranty should be offered.

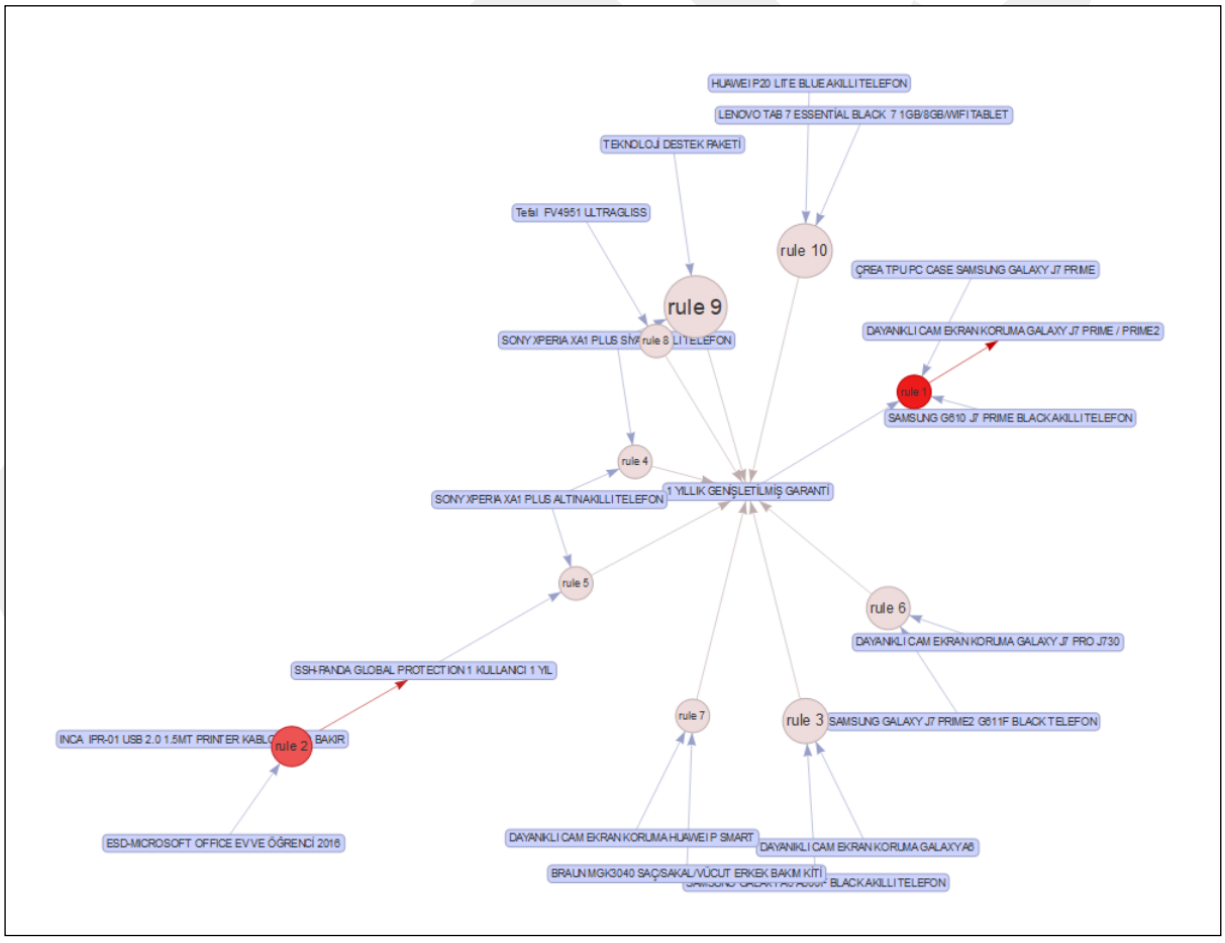

Figure 4.3.1.2: Graph of Top 10 Rules

Some of the main consequences of these rules can be ordered as follows:

- If there is a smart phone, notebook or phone case in the basket; additional warranty is the best choice to offer.
- If a smart phone and a Bluetooth earphone are in the basket together, protective cover for phone should be offered.
- If there is a notebook or a notebook case in the basket; anti-virus should be offered.
- If there is a notebook case and additional warranty in the basket together; anti-virus should be offered.
- If there is both a notebook and an antivirus in the basket, Microsoft Office should be offered.
- If a printer is in the basket, cartridge should be offered.
- If a printer and a battery are in the basket together, cartridge should be offered.

Therefore, complementary products due to 333 association rules can be summarized depending on their category are as follows:

- Smart Phone, Additional Warranty, Phone Case, Earphones, Power bank, Charge Cable
- Notebook, Notebook Case, Additional Warranty, Anti-Virus, Microsoft Office
- Printer, Colorful Cartridge, Black-White Cartridge, Battery

In summary of outcomes of Apriori algorithm applied on all datasets; support and confidence threshold for Apriori algorithm on this dataset are 0.00001 and 0.6, respectively. This execution of Apriori Algorithm generates 1126 rules in 2.08 seconds. The dataset includes orders (baskets) of only three months, so, frequency of occurrence of an item or itemset may be low during this period for electronic items like the ones in this dataset. Therefore, support threshold (0.00001) and confidence threshold (0.6) are acceptable. Secondly, when generated rules are examined according to their confidence in detail; a lift threshold is discovered to notice if item set in the rule is bundle or complementary. The rules with lift less than 200 indicate complementary products where the rules with lift more than

200 indicate bundle products. Bundle products are not the part of this project because they are the products which are forced to purchase together. Thus, the rules with lift less than 200 are examined separately to discover which product is purchased with which product consciously. These rules indicate the complementary products which are useful itemset.

5. DELIVERED VALUE AND FURTHER STEPS

5.1. Project's Delivered Value

In this study, association rules between products are discovered to improve sales process and cross selling. The sale dataset which includes 1228221 orders and about 9000 products is analyzed with Apriori Algorithm. The Apriori Algorithm is executed with different values of the parameters to find the best results according to execution time and number of generated association rules. After getting the best result that generates 1126 rules in 2.08 msec, the generated rules are examined seperately. The generated rules are firstly examined based on their confidence value, then lift value. The interpretation of the association rules due to their lift value shows that there are different association types between products. There is a lift threshold that split the generated rules according to their association types. The rules that have a lift higher than the lift thresold shows the bundle products and the rules that have a lift lower than the lift thresold indicates the complementary products. Bundle products are not the part of this project because they are the products which are forced to be purchased. Thus, the rules with lift less than the lift threshold are examined separately to discover which product is purchased with which product consciously. According to the rules which indicate complementary products, which products are purchased together is discovered and this output is usable to develop an automated offering system.

5.2. Social and Ethical Aspects

Through market basket analysis, a retail company can understand its products value, customer shopping behaviors and gain competitive advantage. This project does not include customers' personal information, It examines only orders (baskets) for understanding their purchasing behavior trends. Additionally, this project does not contain any competitor's data. Consequently, the output of this market basket analysis would increase understanding

of customers' purchasing behaviors trends without any personal information or competitor data.

5.3. Further Steps

This study assumes the traffic of all stores are equal. Before market basket analysis, clustering or classification methods can be applied on stores and according to this classification or clustering, market basket analysis may apply on generated clusters/classes of the stores separately (Chen et al., 2005).

Additionally, understanding the association of the products is the fundament behind the recommendation engine. This study can be the first step of a recommendation engine developed by machine learning methods. The output of this market basket analysis can be improved by increasing data volume and velocity and it can also be used as an input of a recommendation engine.

6. REFERENCES

- Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *In Proc. of the 20th VLDB Conference,* (pp. 487–499). Santiago.
- Chapman, C., & Feit, E. M. (2016). In *R for Marketing Research and Analytics* (pp. 339-360). Springer.
- Chen, Y.-L., Tang, K., Shen, R.-J., & Hu, Y.-H. (2005). *Market Basket Analysis in a Multiple Store Environment.* National Central University.
- Fürnkranz, J., & Kliegr, T. (2011). *A Brief Overview of Rule Learning.*
- Hahsler, M., Buchta, C., Gruen, B., Hornik, K., & Johnson, I. (2018, April 7). Retrieved July 30, 2018, from cran.r-project.org: https://cran.r-project.org/web/packages/arules/arules.pdf
- Han, J., Pei, J., & Kamber, M. (2012). In *Data Mining: Concepts and Techniques* (p. 17). New York: Morgan Kauffman Publishers.
- Linoff, G., & Berry, M. J. (2011). Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. In *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management* (pp. 535-542). Indiana: Wiley Publishing.
- Marafi, S. (2014, March 19). *Market Basket Analysis with R.* Retrieved August 30, 2018, from www.salemmarafi.com: http://www.salemmarafi.com/code/market-basket-analysis-with-r/
- Musungwini, S., Zhou, T. G., Gumbo, R., & Mzikamwi, T. (2014). *The Relationship Between (4ps) & Market Basket Analysis. A Case Study Of Grocery Retail Shops In Gweru Zimbabwe.*
- (2018, April 24). Retrieved August 05, 2018, from cran.r-project.org: https://cran.r-project.org/web/packages/arulesViz/arulesViz.pdf
- Prasad J, P., & Mourya, M. (2013). *A Study on Market Basket Analysis Using a Data Mining Algorithm.*
- Setiabudi, D. H., Budhi , G. S., Purnama , I. J., & Noertjahyana, A. (2011). *Data mining market basket analysis' using hybrid-dimension association rules, case study in Minimarket X.* IEEE.

- Szymkowiak, M., Klimanek, T., & Józefowski, T. (2018). *Applying Market Basket Analysis To Official Statistical Data.*
- Tan, P.-N., Kumar, V., & Srivastava, J. (2004). *Selecting The Right Objective Measure For Association Analysis.*
- *The Apriori Algorithm.* (2015, October 22). Retrieved June 28, 2018, from en.wikibooks.org: https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Frequent_Pattern_Mining/The_Apriori_Algorithm
- *What is R?* (n.d.). Retrieved August 01, 2018, from https://www.r-project.org/about.html

APPENDIX A

```
#Load Libraries and Set envirment:
library(arules)  # association rules
library(arulesViz)  # data visualization of association rules
library(RColorBrewer)  # color palettes for plot
library(dplyr)
library(ggplot2)
library(corrplot)
library(dummies)
library(tidyverse)
library(ggplot2)

#set env:
setwd("C:/Users/ecetp/Downloads/MEF/Capstone/R Codes")
Sys.setlocale("LC_ALL","Turkish")
options(scipen=999)

####### Read Data:
data <- read.delim("sales_2018_456_vs2.txt",header=TRUE,sep=";",quote = "\"",stringsAsFactors = TRUE,encoding = "ANSI")
str(data)

#clean data$prod_name from double quotes and quotes ( to read transactions efficiently later):
data$PROD_NAME<-sapply(data$PROD_NAME, function(x) gsub("\"", "", x))
data$PROD_NAME<-sapply(data$PROD_NAME, function(x) gsub("\'", "", x))
data$PROD_NAME<-sapply(data$PROD_NAME, function(x) gsub(",", ".", x))

length(unique(data$PROD_NAME))
length(unique(data$ORDER_ID))

#is there any null value?
summary(is.na(data))
#convert period into factor:
data$PERIOD <- as.factor(data$PERIOD)
unique(data$PERIOD)
#order by ordeR_ID:
data <- data[order(data$ORDER_ID),]

#########EDA#########
####### Take sample and play it:
# take sample:
ex <- data
str(ex)

#convert into num
```

```
ex$ORDER_ID <- as.numeric(as.character(ex$ORDER_ID))

#The most number of products:
t <- ex %>% group_by(ex$ORDER_ID) %>%
  dplyr::summarize(count = n()) %>%
  arrange(desc(count))
t <- head(t,1000)
mean(t$count)

qplot(t$count, geom="histogram",
      xlab="Product Count",
      ylab="Occurence",main="Distribution Of Product per Order-Top 1000",
      fill=I("indian red")) + theme_minimal()

#The most cost of order:
t <- ex %>% group_by(ex$ORDER_ID) %>%
  dplyr::summarize(ord_amount=sum(SALES_PRICE)) %>%
  arrange(desc(ord_amount))

t <- head(t,1000)
mean(t$ord_amount)

qplot(t$ord_amount, geom="histogram",
      xlab="ORDER AMOUNT",
      ylab="Occurence",main="Distribution Of Order Amount-Top 1000",
      fill=I("indian red")) + theme_minimal()

#Regional dist of Cost:
t <- ex %>% group_by(REGION) %>%
  dplyr::summarize(ord_amount=sum(SALES_PRICE)) %>%
  arrange(desc(ord_amount))

t %>%
  ggplot(aes(x=reorder(REGION,ord_amount),  y=ord_amount))+  ggtitle("Regional
Distribution of Order Amount" ) +
  geom_bar(stat="identity",fill="indian red")+
  coord_flip() + theme_minimal()

#Periodical dist of Amount:
t <- ex %>% group_by(PERIOD) %>%
  dplyr::summarize(ord_amount=sum(SALES_PRICE)) %>%
  arrange(desc(ord_amount))
t %>%
  ggplot(aes(x=PERIOD, y=ord_amount))+
  geom_bar(stat="identity",fill="indian  red")+  ggtitle("Periodical  Distribution  of
Order Amount" ) +
  coord_flip() + theme_minimal()
```

```
###period region and ord amount
t <- ex %>% group_by(PERIOD,REGION) %>%
  dplyr::summarize(ord_amount=sum(SALES_PRICE)) %>%
  arrange(PERIOD,desc(ord_amount),REGION) %>% top_n(10,ord_amount)
qplot(x = ord_amount, y=REGION ,data = t,
          ylab = "Region",
          xlab = "Order Amount") +
  theme_minimal() +
  facet_wrap(~PERIOD)

# Top 10:
library(dplyr)
tmp <- ex %>%
  group_by(PROD_ID,PROD_NAME) %>%
  dplyr::summarize(count = n()) %>%
  arrange(desc(count))
tmp <- head(tmp, n=20)
View(tmp)
tmp %>%
  ggplot(aes(x=reorder(PROD_NAME,count), y=count))+
  geom_bar(stat="identity",fill="indian red")+
  coord_flip() + theme_minimal()

#Brand dist of Amount:
t <- ex %>% group_by(BRAND) %>%
  dplyr::summarize(ord_amount=sum(SALES_PRICE)) %>%
  arrange(desc(ord_amount))
t <- head(t,10)
t %>%
  ggplot(aes(x=BRAND, y=ord_amount))+
  geom_bar(stat="identity",fill="indian red")+ ggtitle("Distribution of Order Amount
vs BRAND-Top 10" ) +
  coord_flip() + theme_minimal()

# Period  and brand vs Order Amount
t <- ex %>% group_by(PERIOD,BRAND) %>%
  dplyr::summarize(ord_amount=sum(SALES_PRICE)) %>%
  arrange(PERIOD,desc(ord_amount),BRAND) %>% top_n(10,ord_amount)
p0 <- qplot(x = ord_amount, y=BRAND ,data = t,
          ylab = "Brand",
          xlab = "Order Amount") +
  theme_minimal() +
  facet_wrap(~PERIOD)

####Market Basket Analysis######
##Create Transactions:
library(plyr)
ex <- ex %>% select (ORDER_ID,PROD_NAME)
```

```
time0 <- proc.time()
itemList <- ddply(ex,c("ORDER_ID"),
                  function(df1)paste(df1$PROD_NAME,
                                     collapse = ";"))
dim(itemList)

time1 <- proc.time() - time0
cat("Prepare itemList lasts: ", time1)

#we dont need order_id any more:
itemList$ORDER_ID <- NULL
colnames(itemList) <- c("items")

#write transactions into a file:
time0 <- proc.time()
write.csv(itemList,"market_basket123.csv",quote = FALSE, row.names = FALSE) #
row.names=TRUE olduÄŸunda satÄ±r no koyar.
time1 <- proc.time() - time0
cat("write transaction lasts: ", time1)

#read transactions from the file:
time0 <- proc.time()
tr <- read.transactions('market_basket123.csv', format = 'basket', sep=';')
time1 <- proc.time() - time0
cat("read transaction lasts: ", time1)

###### Algorithms:
######Apriori:
time0 <- proc.time()
rules <- apriori(tr, parameter = list(supp= 0.00001 , conf=0.6))
rules <- sort(rules, by='lift', decreasing = TRUE)
summary(rules)

time1 <- proc.time() - time0
cat("Apriori: ", time1)

#convert the output rules into dataframe:
df <- as(rules,"data.frame")
View(df)
inspect(rules[1:10])

#plot the rules:
library(plotly)
plotly_arules(rules, jitter = 10,
              marker = list(opacity = .5, size = 10, symbol = 2),
              colors = c("red", "dark blue"))
```

```
plotly_arules(rules, jitter = 10, method="two-key plot",
          marker = list(opacity = .5, size = 3, symbol = 2))

topRules <- rules[1:10]
plot(topRules, method="graph", control=list(type="items"))

#Generate subset of the rules based on confidence of rules:
subsetx <- subset(rules, confidence >= 0.9 & confidence <1)
retail.hi <- head(sort(subsetx, by="confidence"), 10)
inspect(retail.hi)
df <- as(retail.hi,"data.frame")
View(df)
table(df)
plot(retail.hi, method="graph", control=list(type="items"))

## without the rules which dont indicate the bundle products:
subsetx <- subset(rules, lift <= 200)
ruleExplorer(subsetx)
```