QTM 2000 Name: Angelina Cho

Case Studies in Business Analytics

Professor Mathaisel

Fall 2021

Final Project

Individual Assignment

The following pledge must be on your exam cover sheet and signed.

I pledge my honor that I have neither received nor provided unauthorized assistance during the completion of this work. Please Initial: AC

The objective of this project is to gain insight into customer purchases during "Black Friday".
The analytical method used is Association Rules (arules). For this project you are being asked to write an "Executive Summary" to accomplish that objective using the dataset and its accompanying R script provided. Your summary should be brief, but it must include:
Introduction; Analysis; and Conclusions. The analysis is divided into two parts: 1. Exploratory Data Analysis (EDA); and 2. Association Rules.

Part 1.	50 points
Part 2.	50 points
	100 points

Introduction

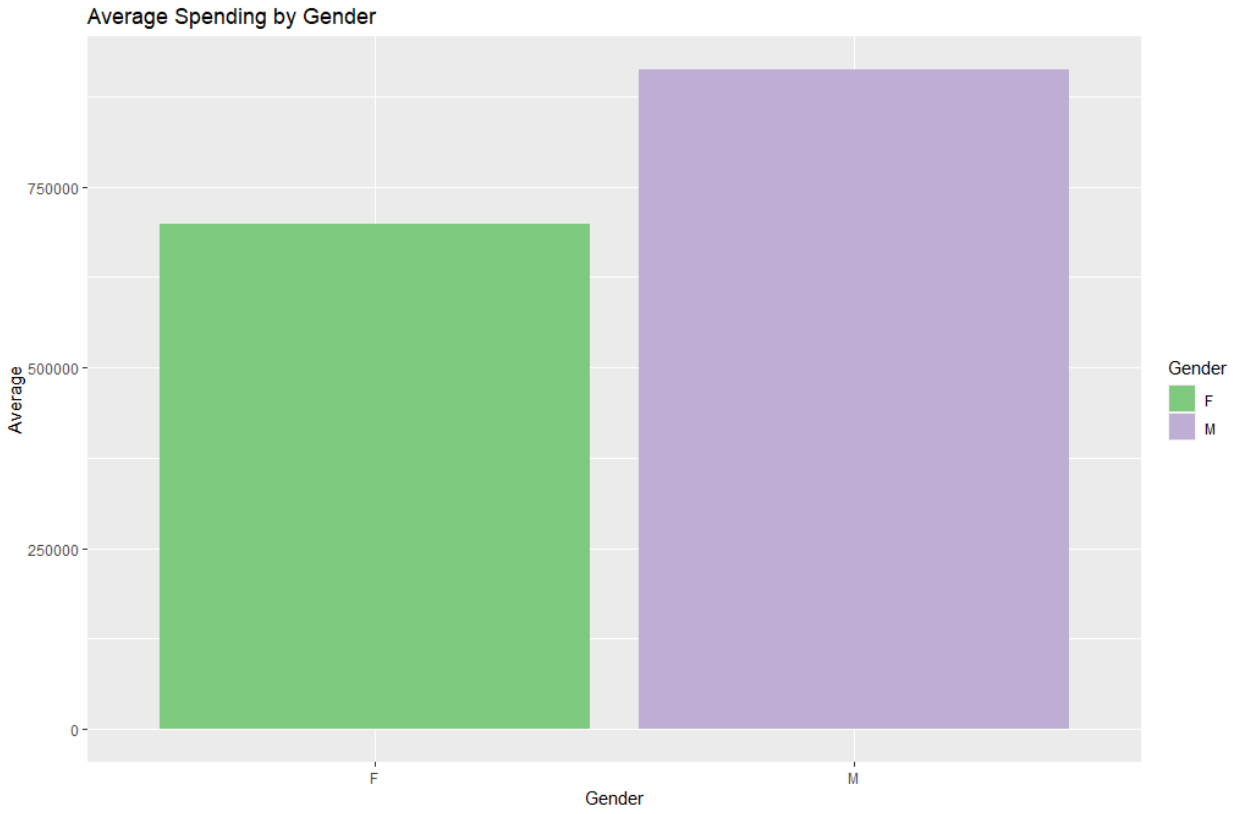
The purpose of this executive summary includes using Association Rules (arules), also known as Market Basket Analysis, to gain insight on customer purchases during "Black Friday". The Association Rules is used by a lot of companies that rely on recommendation engines, such as Amazon, Netflix, Spotify, etc. It is also used by companies trying to understand the products that are likely to be purchased together, which is, essentially, what we are trying to find out from looking at the Black Friday dataset. This enables retail stores to up-sell products that the customers potentially are interested in, by finding the logical rule between the purchase of one product/service (Antecedent) and another purchase of product/service (Consequent). It is done with no supervision (unsupervised learning) and in an unguided manner with no specific goal or objective apart from gaining insights on the data. Association Rules work in a way that it identifies the strengths of association between pairs of products that were purchased together, and identifies patterns of co-occurrence, which is when two or more purchases take place together. It then creates a "If-Then" scenario, where "If item A (Antecedent) is purchased then item B (Consequent) is likely to be purchased too" situation. The rule itself is derived from the probability of co-occurrence in the data. Both the Antecedent and Consequent can be sets of elements, for example A = {Beer, Wine, Scheltzer} and B = {Cups, Utensils, Cutlery}. However, A and B need to be disjoint, meaning the item sets should have no elements that are in common. Then, we look for the "support", "confidence", and "lift" of the data. Support is the frequency that the item sets occur together. We translate the frequencies of occurrence (Support) into likelihood or the probability, which is the confidence by finding the conditional probability of the event. We would, then, use the Lift Ratio to compare our confidence values to the benchmark confidence value to find the strength of an Association rule. In general, the larger the lift ratio, the greater the strength of the association is between the Antecedent and the Consequent. Therefore, as the assignment says, this project is divided into two parts: exploring the Black Friday.csv data, and utilizing Association Rules to find possible correlations between customers' purchases.

This case is focused on 550,000 observations on Black Friday shoppers in a retail store. It contains both numerical and categorical data types, as well as some missing values. The Black Friday dataset contains 12 different columns (variables) with a mix of data types. We will be using Exploratory Data Analysis (EDA) to group the observations for our visualizations. The reason behind it is because each row of the data represents a transaction made by a specific customer. This same customer may also purchase another product or make another transaction. Therefore, it is important for us to group all transactions by a specific User_ID to get a sum of all purchases made by one single customer, which we will soon see in the executive summary. We also need to be aware that there isn't a Data Dictionary for many of the attributes in this dataset, for example Product_ID, Occupation, City_Category, Marital

Status, and Purchase. It would have been much more useful if we had the name of an item and its Product_ID, and we would have gained much more information by knowing the actual product. However, it would not affect our Exploratory Data Analysis.

Figure 1 Average Spending by Gender

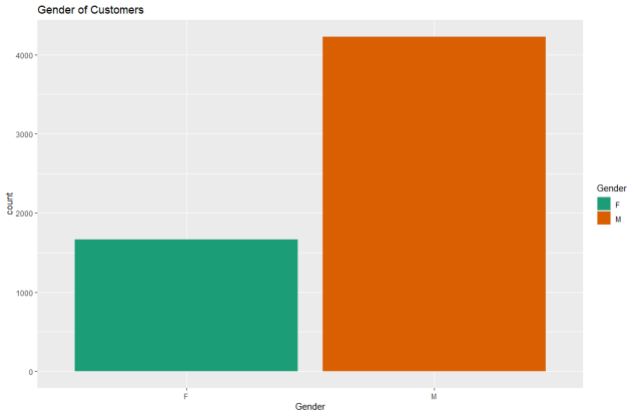
	Gender	Purchase	Count	Average
	<chr>	<dbl>	<int>	<dbl>
1	F	1164624021	1666	699054.
2	M	3853044357	4225	911963.



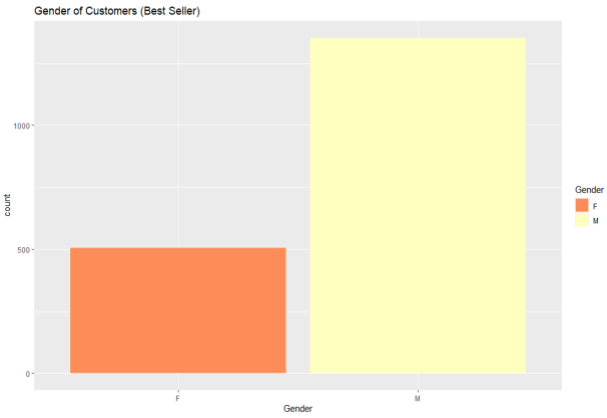
The table above shows the actual number of purchases and the amount made in all transactions for each gender, and concludes with an average of the purchase between female and male. It shows that for this retail store, the male made a higher dollar amount of purchases ($911,963), on average, compared to the female population ($699,054). It means that this store attracts male customers more than it attracts female customers.

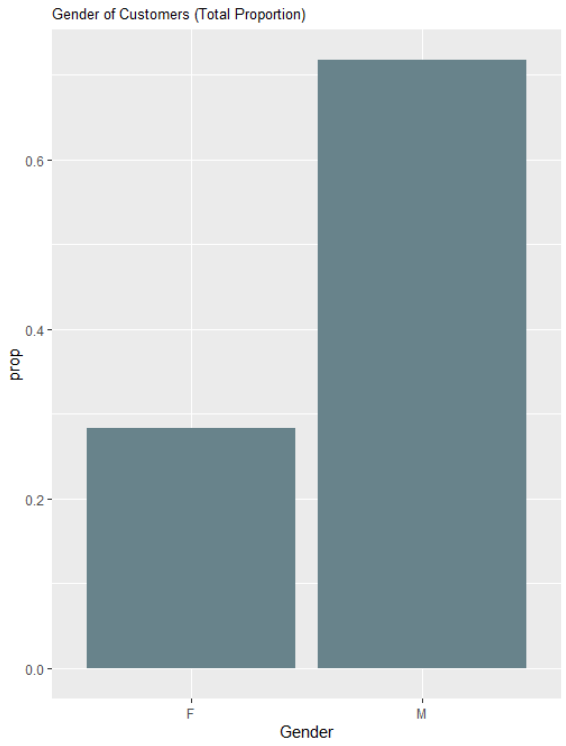
Figure 2 Gender of Customers (Total Proportion vs. Best Seller Proportion)

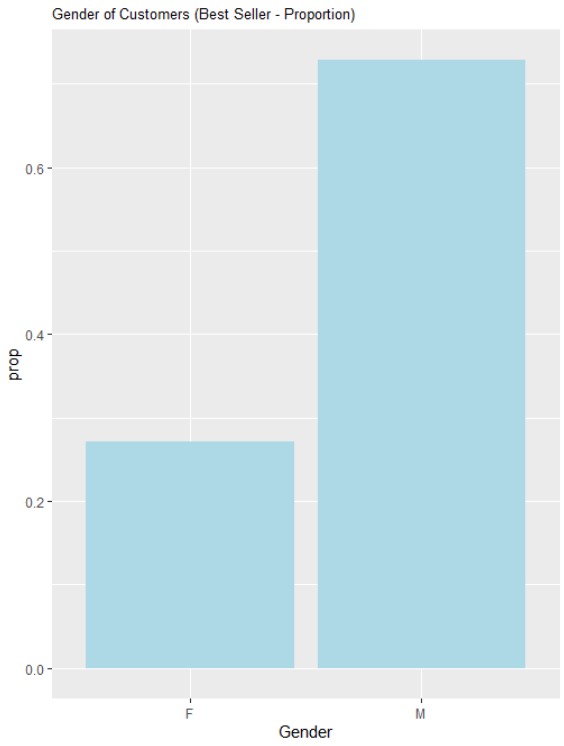
```
# Groups:   User_ID [6]
  User_ID Gender
    <int> <chr>
1 1000001 F
2 1000002 M
3 1000003 M
4 1000004 M
5 1000005 M
6 1000006 F
```




```
  Product_ID    n
1  P00265242 1858
2  P00110742 1591
3  P00025442 1586
4  P00112142 1539
5  P00057642 1430
```





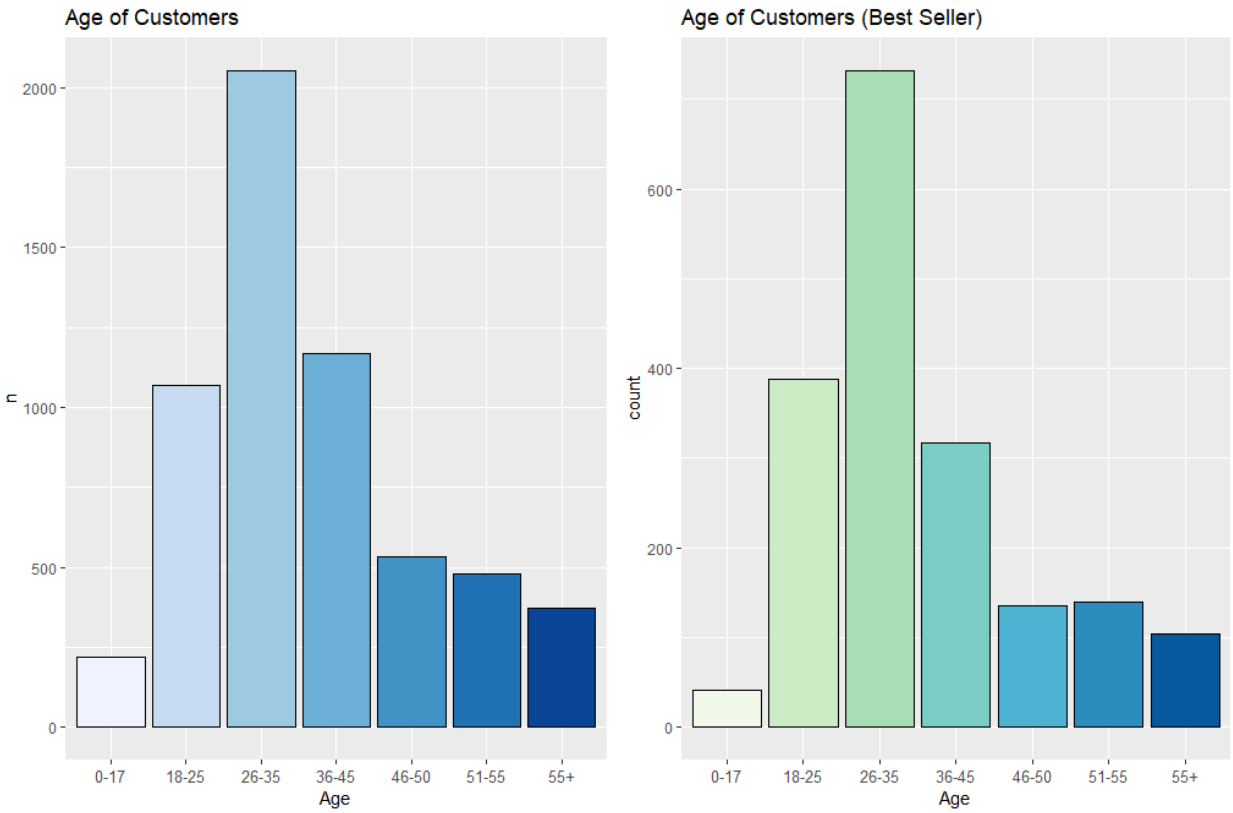


The gender variable is a categorical variable that shows the two different sex of customers, female and male. This figure simply shows how the users are split according to gender, and this is helpful to the retail store because they can modify their store's layout, decoration preference, product selection, etc., depending on the gender proportion of the shoppers. For example, a study published in the Clothing and Textiles Research Journal shows that the involvement and physical environment of stores are included as antecedents of shopping experience satisfaction for females. This shows a difference in shopping preference and values derived from shopping between the genders, which can bring additional sales if the information is used correctly. The first table shows a few observations of sorting the users into different genders. The visualization is the final result derived from the data. The retail store has a distinctly higher proportion of males than females making purchases on Black Friday, with a ratio of approximately 4200:1700 (M:F), respectively.

The second table and visualization displays the purchase according to gender for the top selling products. The products are no longer being grouped by Product_ID, since we want to see the duplicate purchases for the same products, in case people are buying two or more quantities of the same product. The table identified the top five best selling products, with the top (P00265242) having 1858 quantities sold, followed by (P00110742) 1591, (P00025442) 1586, (P00112142) 1539, and (P00057642) 1430 for the second, third, and so on. Unfortunately, as the Introduction mentioned, we do not have the key to reference the item name in order to know what it actually is. The visualization illustrates the frequencies of purchases made by different genders for the top best selling product. The female population made a total of approximately 500 purchases of the top best selling product, while the male population made a total of approximately 1200 purchases of the top best selling product. The ratio of male purchases to female purchases for the top best selling product is approximately 1200:500.

The distribution and proportion between gender and best seller and gender and total purchases is similar, with male being the dominant population making the purchases. We can see that the proportion for the gender and best seller graph for female to male is around 0.3:0.7, and the proportion for the gender and total purchases depicts a similar picture of 0.3:0.7 for female to male.

Figure 3 Age of Customers (Total Population and Best Selling Product)



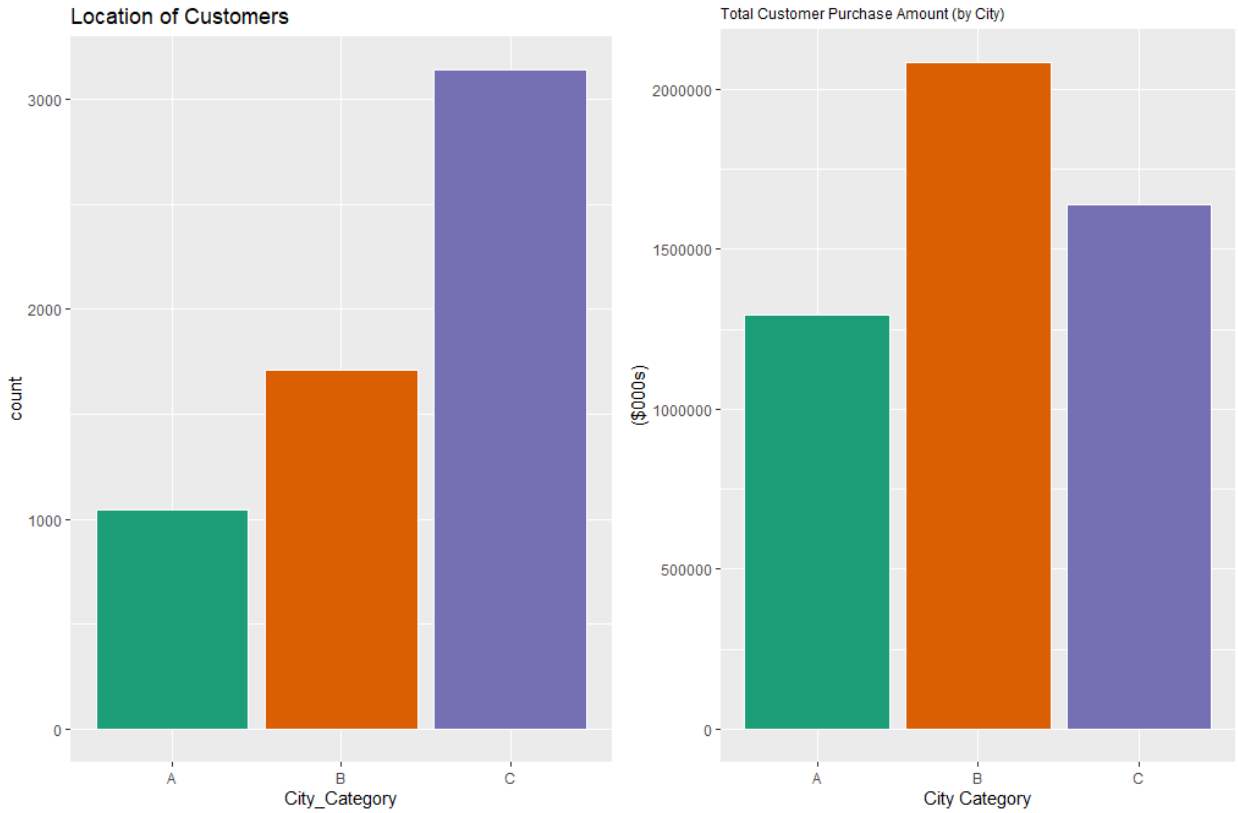
The Age variable is the age of customers split into bins/groups. The bar graph on the left shows the number of purchases made by different age groups of customers. It is skewed to the right, meaning the majority of the customers are younger. The mean for this distribution should be larger than the median, as this graph is skewed to the right. The median age group seems to be 26-35, while the mean age group is slightly higher than that. The age group that has the most number of purchases is 26-35, with more than 2000 purchases made. Age group of 36-45 has the second highest number of purchases of around 1200, followed by 18-25 with around 1200 purchases made. Age group of 0-17 has the lowest number of purchases, which makes sense as they often don't have the purchasing power yet, financially speaking. Age groups above age 46 have significantly lower numbers of purchases. It may be because Black Friday no longer appeals to them anymore.

The bar graph on the right is a similar graph, depicting the distribution of age for the top best seller product. It has a similar distribution compared to the total purchase graph on the left. The mean for this distribution is also slightly higher than the median, as this graph is also skewed to the right. Age group 26-35 is still leading with more than 700 purchases. It is, then, followed by the 18-25 age group, which is different from the total purchase graph, with around 400 purchases. The 36-45 age group is in third with around 300 purchases. This means that the best selling product appeals more to the age group

of 26-35, regarding Black Friday events. The best selling product does not seem to be appealing to the population between 46 and above because their number of purchases average at around 130 purchases.

The two graphs painted a similar picture. This means that Black Friday and the best selling product has a common characteristic that attracts the age group of 26-35. It is also likely that this age group just has more shopping sprees than all other age groups, as people tend to lean towards wants rather than needs for Black Fridays.

Figure 4 Location of Customers (Purchase Frequency and Purchase Amount $)



The visualization on the left shows the number of customers in City A, City B, and City C. The City Category is a categorical variable that represents three different cities. We are not sure where each location represents exactly due to lack of key in the Data Dictionary. However, it does give us an insight as to which city has the most sales. It appears that City C has the most number of customers compared to the other two cities, with slightly more than 3000 customers. City B has significantly less number of customers of around 1700, which is also almost less than half of City C. City A has the least number of customers of around 1000 only, which is less than one-third of the number of customers in City C.

The visualization on the right shows the total amount of customer purchase according to City A, City B, and City C. It depicts a different story than the graph on the left, which shows the number of

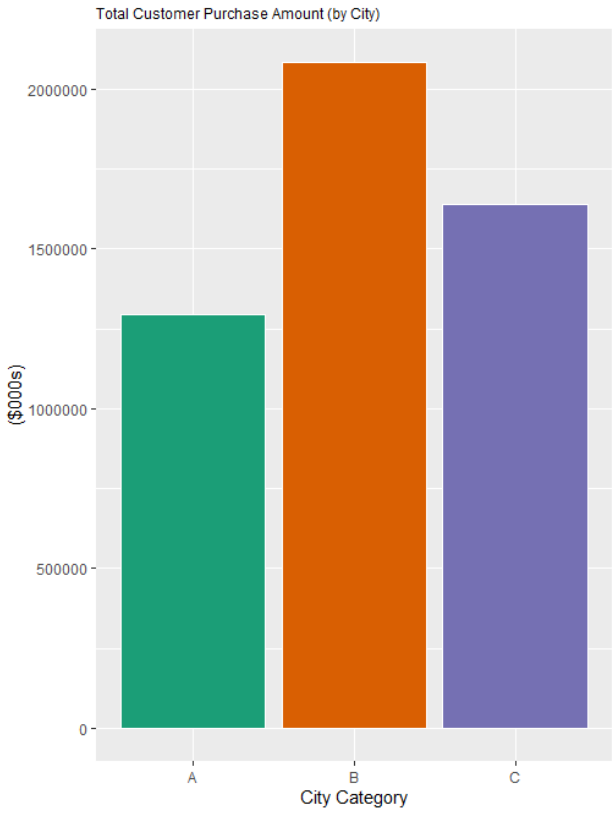
customers in different cities. It shows that City B's customers spent the most at the retail store of more than $2 billion. It is then followed by City C's customers with a total of around $1.6 billion, with a gap of $400 million compared to City B's. City A's customers spent the least at the retail store. Their total amount of purchase is around $1.3 billion, with a gap of $300 million compared to City C's.

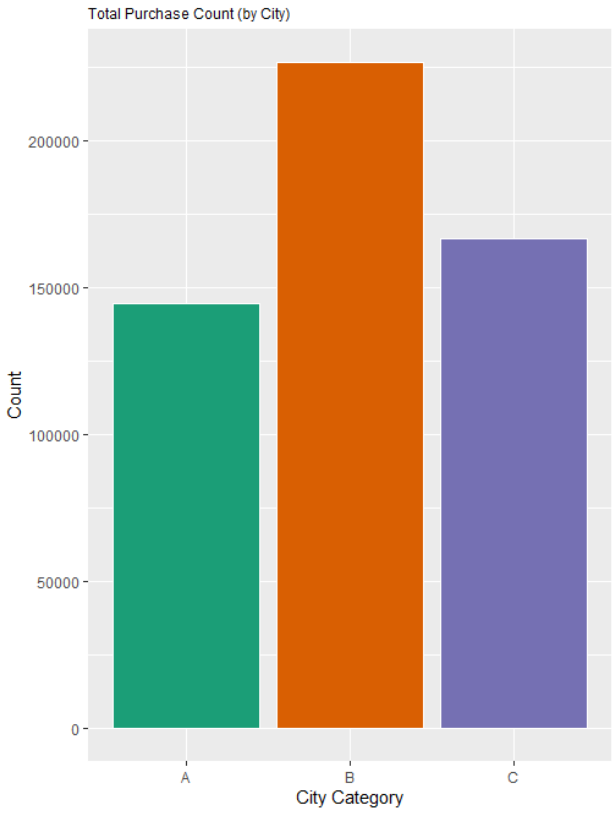
The difference between the two graphs show that there is a likelihood that City B has higher purchasing power compared to other cities. It is also likely that they have different cultures, which then results in different purchasing preferences. For example, Costco is more popular in the U.S. because they like to buy in bulk, while Eastern countries are less likely to buy any items in bulk, leading to different purchasing decisions and behaviors.

Figure 5 Total Customer Purchase Amount ($) by City and Total Purchase Number by City

	User_ID	City_Category	n
	<int>	<chr>	<int>
1	1000001	A	34
2	1000002	C	76
3	1000003	A	29
4	1000004	B	13
5	1000005	A	106
6	1000006	A	46

	City_Category	CountOfPurchases
	<chr>	<int>
1	A	144638
2	B	226493
3	C	166446

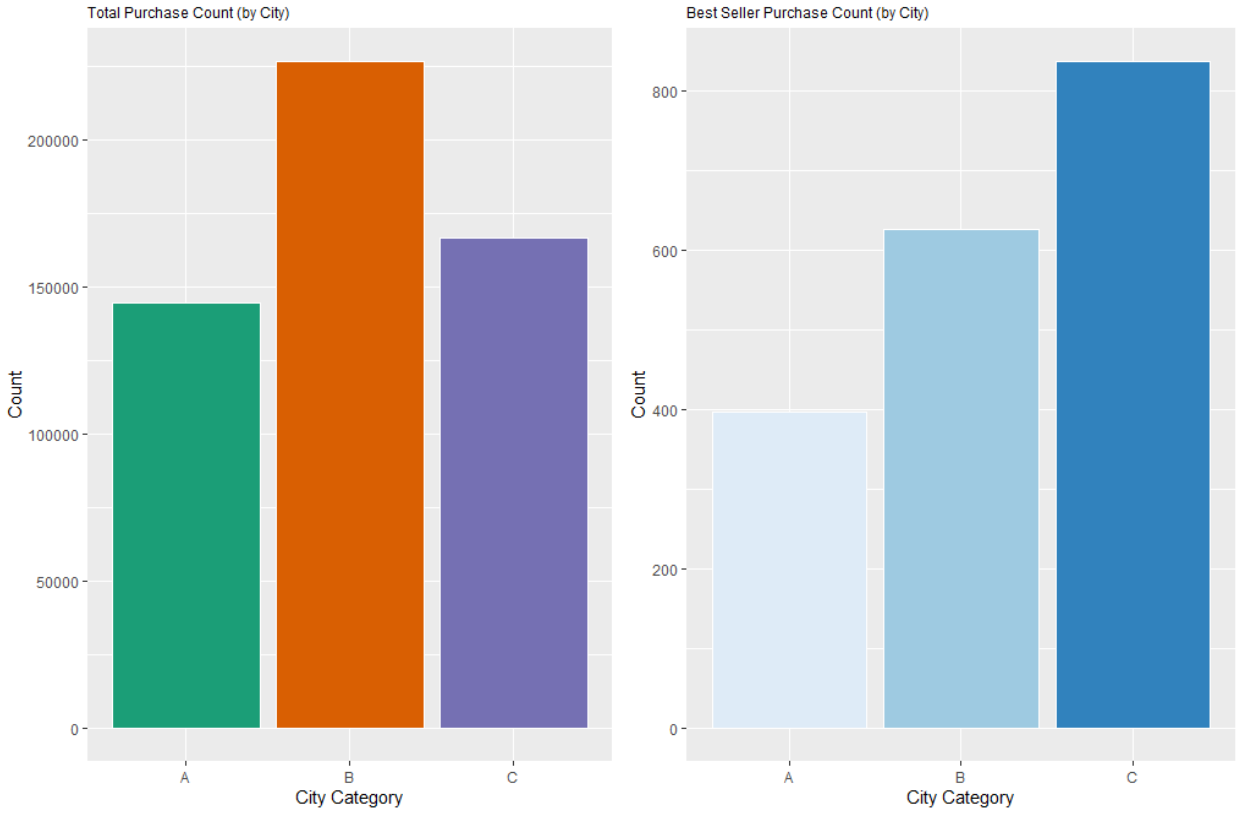




The figure on the left shows the total amount of customer purchase made by customers according to the cities. We found this data by getting the total number of purchases for each corresponding User_ID, and then we find the location of a certain customer and organize it into different City Categories, combining with other information. The total number of purchases by city depict a similar picture as to the total amount of purchases in dollars by city. This is because as the number of purchases increase, the total amount of purchase in dollars also increases. This time the total amount of purchase by city graph is compared side-by-side with total number of purchases by city. The table shows that City B has the most number of purchases of 226493, followed by City C with 166446 and City A with 144638.

Figure 6 Total Purchase Number by City and Best Seller Purchase Number by City

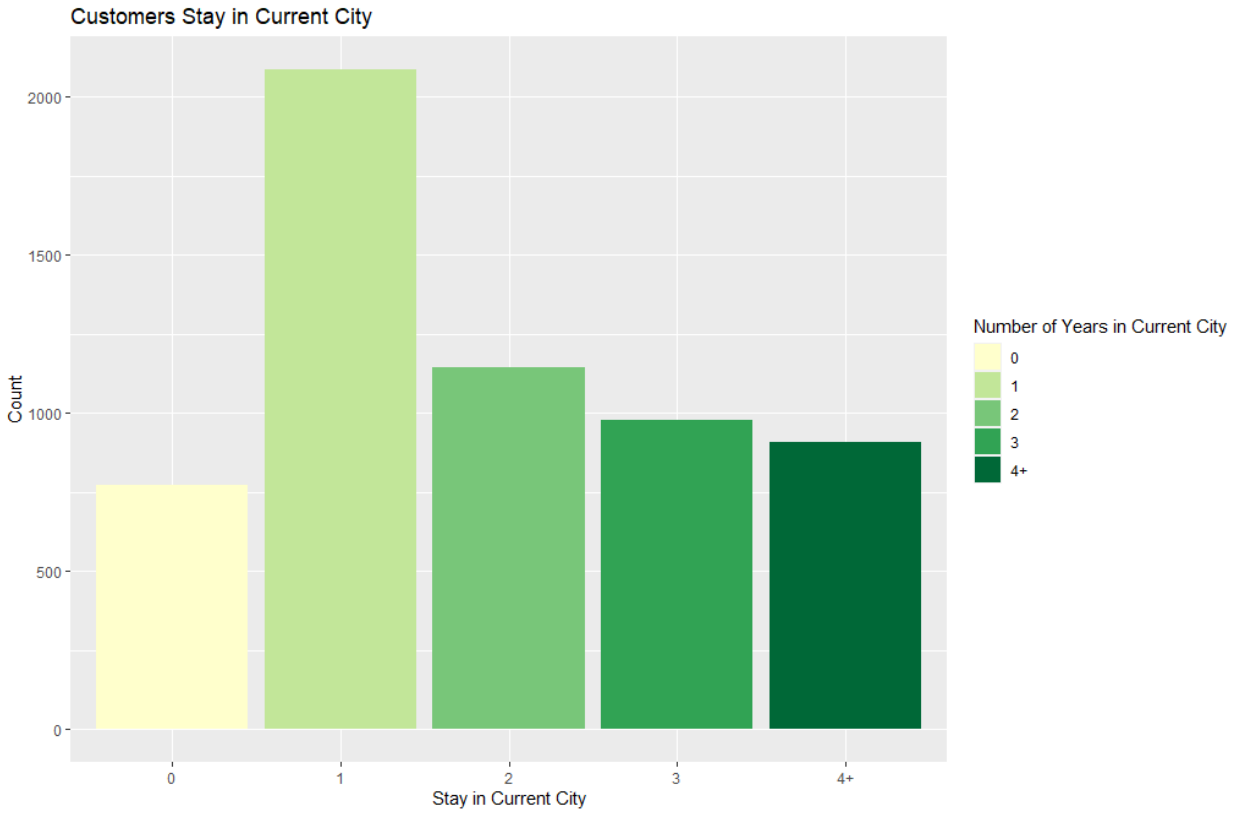
	City_Category	n
1	A	396
2	B	626
3	C	836



The visualization on the left, which is the total number of purchases according to the City Category, is used again to compare side-by-side with the number of purchases for the best selling product (P00265242) according to the City Category. The best selling product has the most sales in City C, which is surprising as it differs from the total sales by City Category. City C has 836 sales for their best selling

product, followed by City B with 626 and City A with 396 purchases, which is less than half of City C's sales.

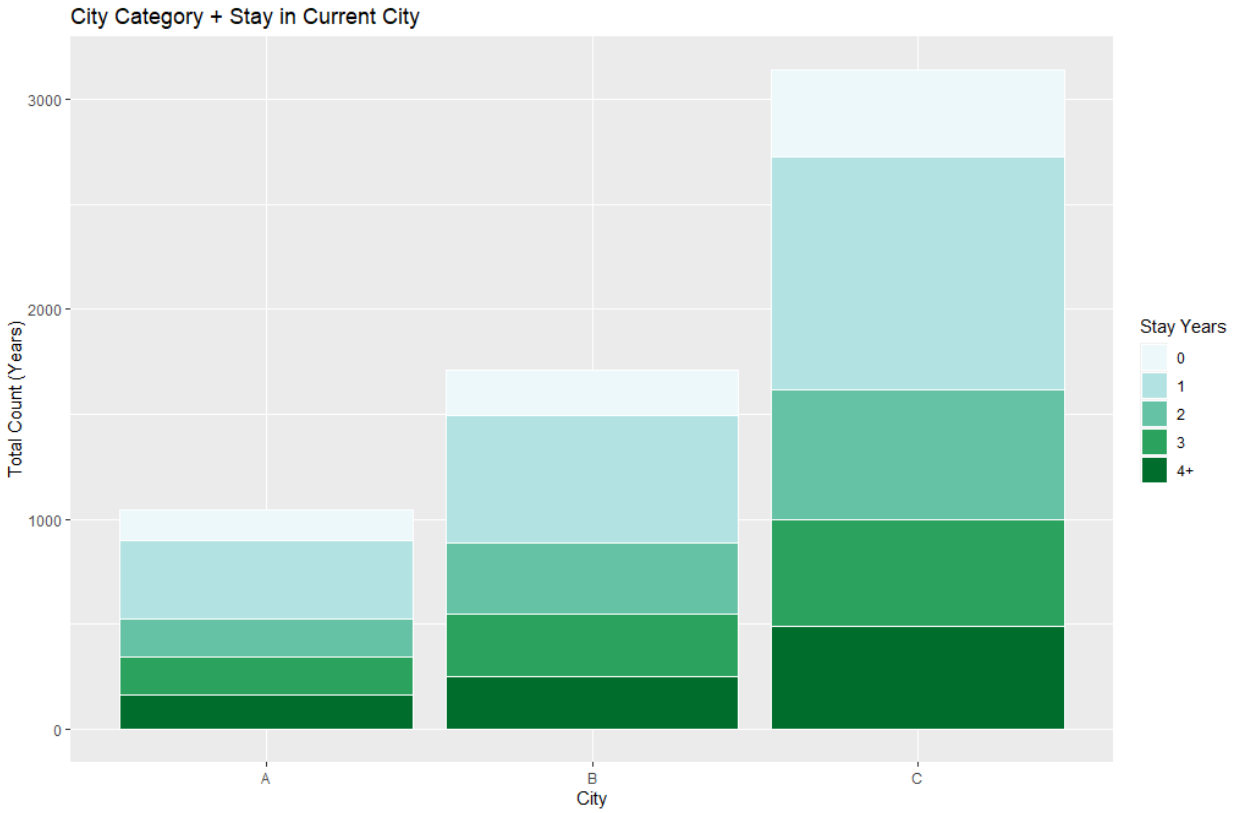
Figure 7 Customers Stay in Current City



The visualization shows the number of years the customers stayed in their corresponding city. Most of the customers (more than 2000) only stayed one year in their current city. The second in line is 2 years, which means that a lot of customers also only have stayed in their current city for 2 years (1200). The third is that customers stayed in their cities for 3 years, around 1000 customers. The fourth is customers that stayed more than 4 years in their current city, around 900. The least is customers that have stayed less than 1 year in their current city, around 750.

Figure 8 Customers Stay in Current City (Stacked Bar Chart)

	City_Category	n
	<chr>	<int>
1	A	1045
2	B	1707
3	C	3139



This graph took the information from the previous figure and sorted it into each City Category using a stacked bar graph. It shows us a clearer distribution in terms of the time customers stayed in different cities and whether or not there is a correlation. The table shows that there are the most number of people staying in their current city in City C, with 3139 customers. It is followed by City B with 1707 customers, which is only a little more than half of the amount of customers in City C that stayed in their current city. City A has the least number of customers that stayed in their current city, with 1045 customers. City C has the most number of customers that stayed more than 4 years, around 500, which is more than half of the customers of the retail store that stayed more than 4 years. There are also around 500 customers that stayed for 3 years in their current city for City C, which is about half of the customers of the retail store that stayed for 3 years. The fact that City C has the highest number of customers also contributes to the fact that it has a higher proportion of all types of Stay in current city customers (years). However, this graph does give an insight on which city has customers that are willing to stay there for long-term. In a city where customers are willing to stay, it is important to build a good reputation and brand name so that the customers would come back to the retail store for other products.

Figure 9 Distribution of Purchase Amount

	User_ID	Purchase_Amount
	<int>	<int>
1	1000001	333481
2	1000002	810353
3	1000003	341635
4	1000004	205987
5	1000005	821001
6	1000006	379450

	User_ID	Purchase_Amount
	<int>	<int>
1	1004277	10536783
2	1001680	8699232
3	1002909	7577505
4	1001941	6817493
5	1000424	6573609
6	1004448	6565878

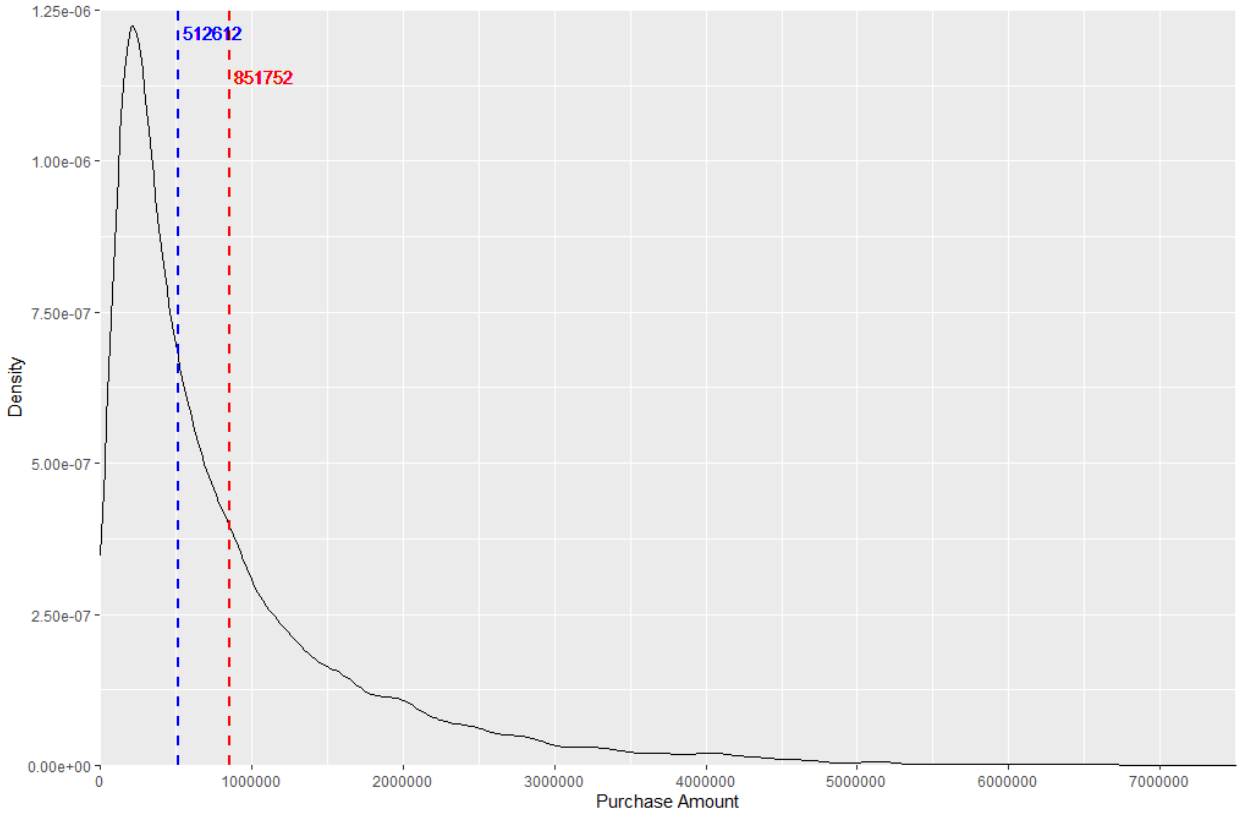
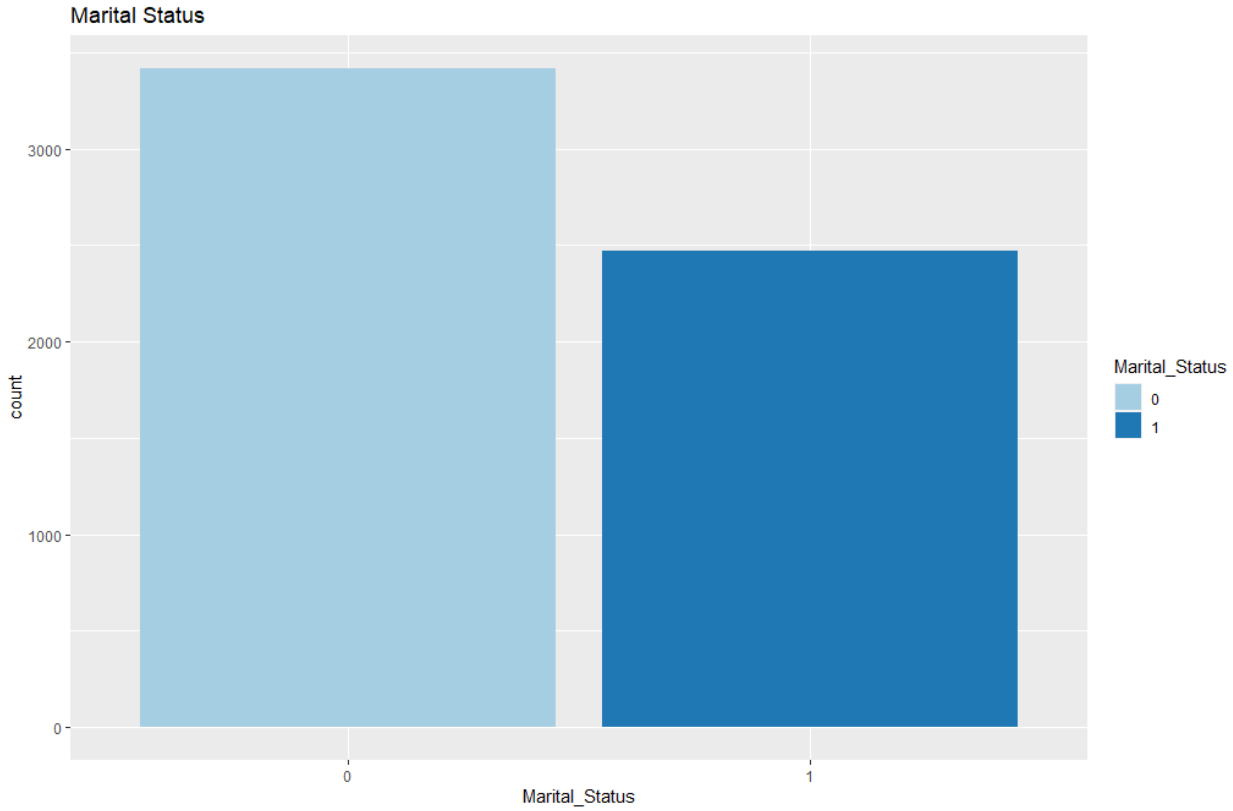


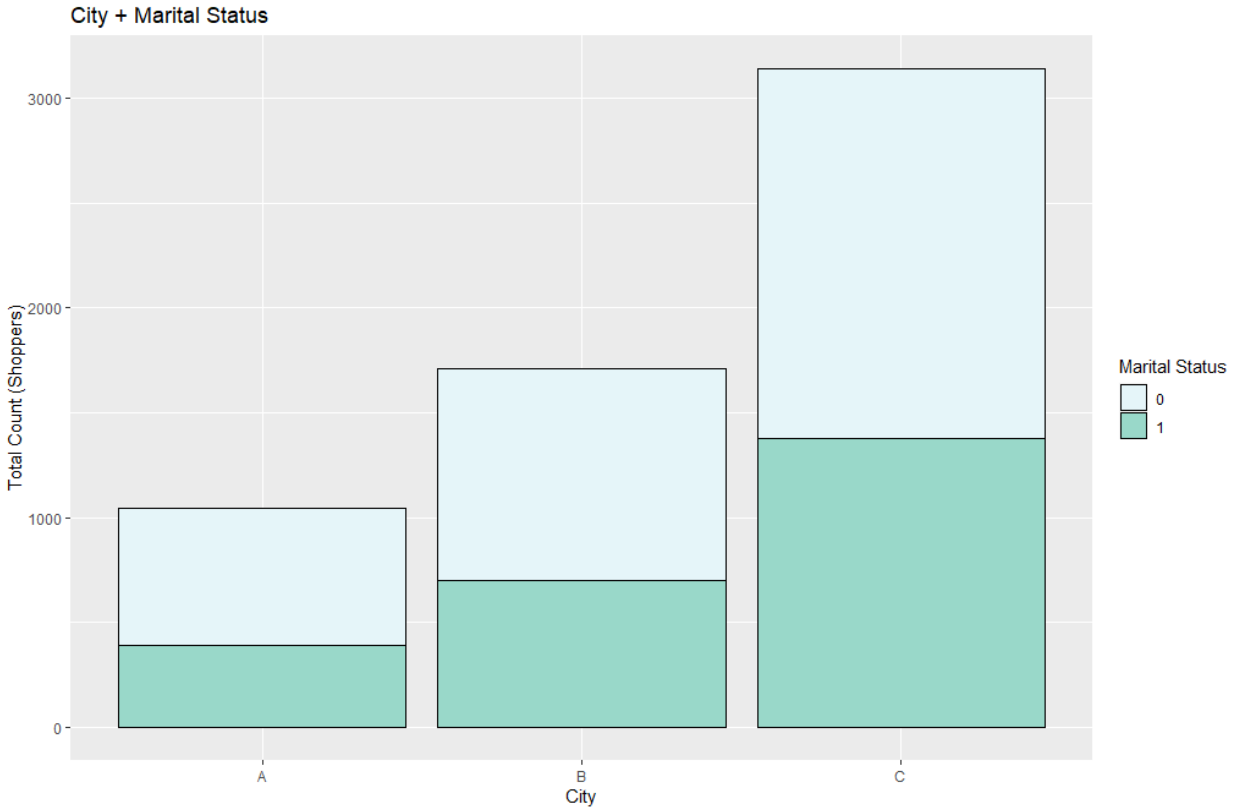
Figure 9 is a density plot that shows the distribution of purchase amount. We found this by computing the total purchase amount by User_ID, then we sorted and found the top spenders of the retail store and grouping customers by different purchase amounts. The purchases are not normally distributed. The graph is skewed to the right, which means the mean is higher than the median. We can prove this by looking at the red dashed line, representing mean ($851752), and the blue dashed line, representing the median ($512612). The highest number of similar purchase amounts rests at around $300000 in accordance with the entire customer base. The number of purchase amounts are significantly lowered after the value of $450000. This makes sense because not a lot of people have a purchasing power that high.

Figure 10 Marital Status of Customers



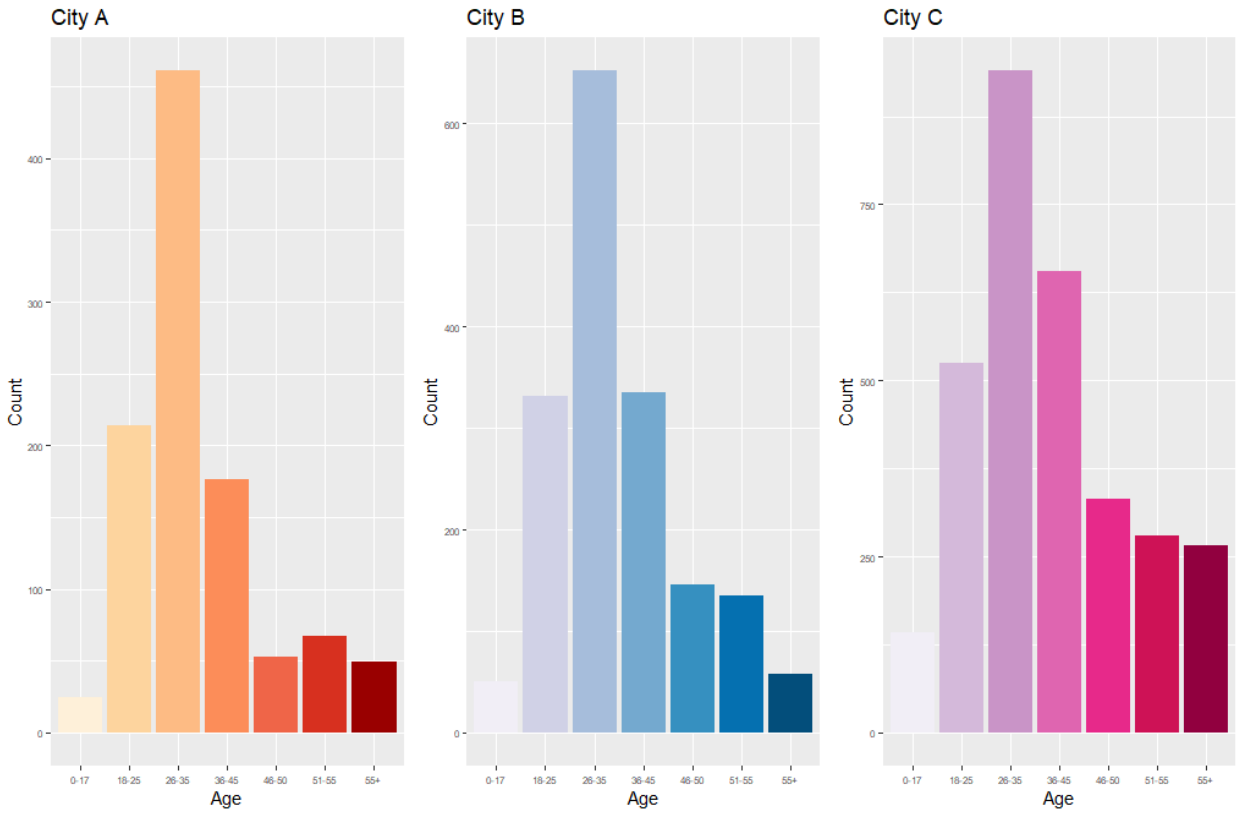
In order to produce this graph, we are changing the data type of Marital Status from numeric variable to categorical type. [1] "character" indicates that the variable is changed from numerical values of 1's and 0's to strings of 1's and 0's. If we look back at the assignment, we can see that we do not have a description in the data dictionary for the marital status variable. Usually it would be better if we make clear the representation of the variable by asking the source, but, in this case, we will assume that 1 = married and 0 = single. Looking at the graph, we can see that the retail store's customers are primarily made of single customers. However, the gap between the number of single customers (around 3400) and the number of married customers (around 2500) did not vary significantly. The ratio of single customers to married customers is around 1.36:1. This means for every single customer that comes to the store there will be 1.36 married customers that will come to the retail store, metaphorically. The reason that the number of single customers are higher is likely due to the fact that married customers are usually better at budgeting and buying only what they need for the family, while the single customers usually just buy what they need for themselves, resulting in giving in on Black Friday promotions and sales.

Figure 11 Total number of Customers by City and Marital Status



This figure shows the total number of customers in each city according to their marital status. We have to consider the fact that there are more customers in City C, followed by City B and then City A. This will affect the number of customers for different types of marital status. Therefore, we need to investigate further by looking at the ratio of single to married customers in each city. For City C, the ratio of single customers to married customers is around 1.21:1, meaning for every married customer out there for our retail store, there will be 1.21 single customers. For City B, the ratio of single customers to married customers is around 1.43:1. This means that for every married customer out there, there will be 1.43 single customers. For City A, the ratio of single customers to married customers is around 1.63:1, meaning that for every married customer out there, there will be 1.63 single customers. This shows that the proportion of single customers is higher in City A, followed by City B, then City C. This is important information for the retail store because the store can change their layout and decoration according to the information mentioned above. For example, City A's retail store can be marketed in a way that it attracts single customers more than it attracts married customers, so that we can strengthen the relationship with the single customers.

Figure 12 Age Distribution for each City



This figure shows the age distribution of customers for each City. The graph on the most left shows the distribution of City A customers according to different Age groups. The second and third graph shows the same information, but for City B and City C, respectively. It seems like all three cities have similar age distribution. City A has a peak of number of customers (around 460) at age group 26-35, which is also the same for age group 26-35 for City B (around 660) and City C (around 900). City A has a similar age distribution compared to the total population. City B is also somewhere similar. However, City C has a relatively different age distribution compared to the total population age distribution, with a higher number of customers at age above 46. For City A, the second highest age group is 18-25 (around 220) and the third highest is 36-45 (around 175). For City B, the second highest age group is 36-45 (around 330), and the third highest is 18-25 (around 325). For City C, the second highest age group is 36-45 (around 645), and the third highest is 18-25 (around 510). This information is important because the store can change their marketing strategy according to which age group they are targeting.

Figure 13 Top Shoppers on Black Friday

	User_ID	n
1	1001680	1025
2	1004277	978
3	1001941	898
4	1001181	861
5	1000889	822
6	1003618	766

	User_ID	n	Purchase_Amount
1	1001680	1025	8699232
2	1004277	978	10536783
3	1001941	898	6817493
4	1001181	861	6387899
5	1000889	822	5499812
6	1003618	766	5961987

	User_ID	n	Purchase_Amount	Average_Purchase_Amount
1	1001680	1025	8699232	8487.056
2	1004277	978	10536783	10773.807
3	1001941	898	6817493	7591.863
4	1001181	861	6387899	7419.163
5	1000889	822	5499812	6690.769
6	1003618	766	5961987	7783.273

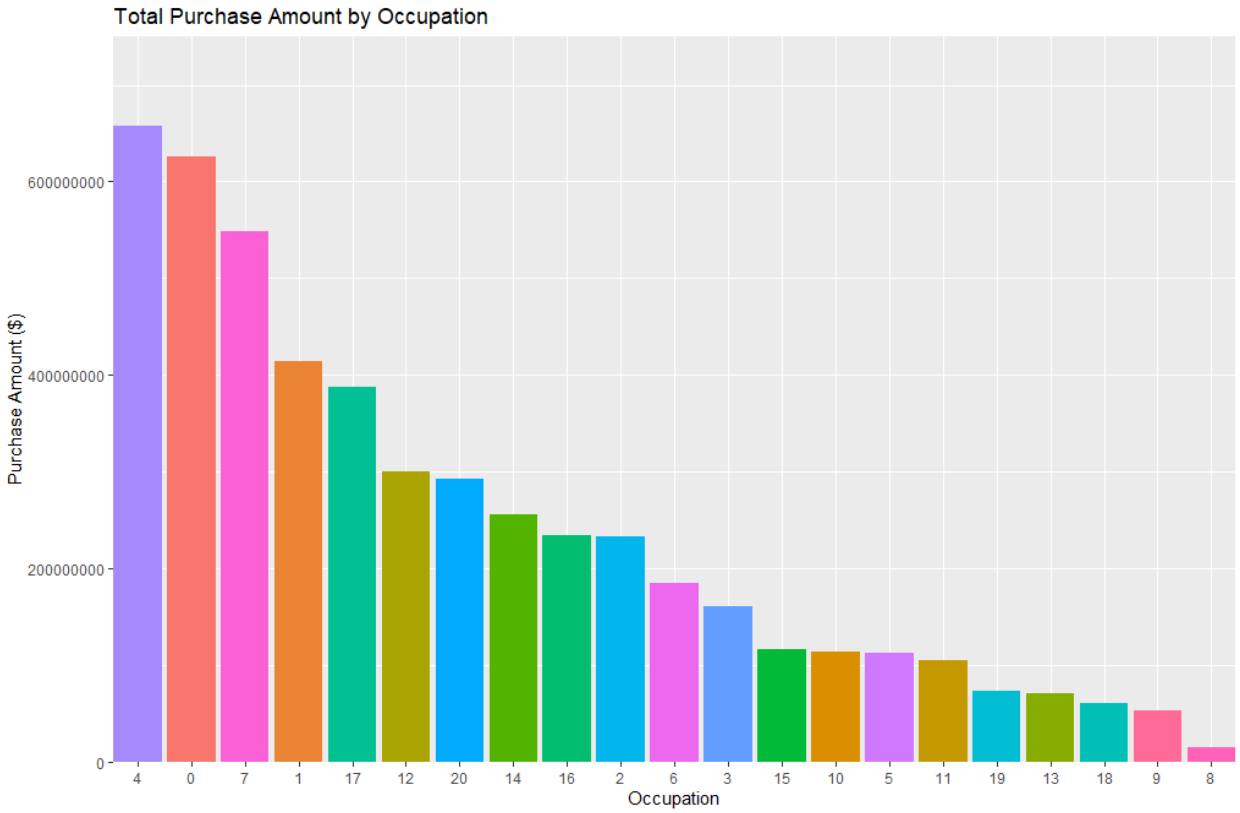
	User_ID	n	Purchase_Amount	Average_Purchase_Amount
1	1005069	16	308454	19278.38
2	1003902	93	1746284	18777.25
3	1005999	18	330227	18345.94
4	1001349	23	417743	18162.74
5	1000101	65	1138239	17511.37
6	1003461	20	350174	17508.70

The top left table shows the top 6 customers that have the highest number of total purchases. The table right next to it shows the top 6 customers that have the highest purchasing amount. The graph on the second row shows the average purchasing amount for the top 6 purchasing amount customers. The graph on the last row shows the top 6 customers that have the highest average purchasing amount in the retail store. These are all EDA that helps perform initial investigations on data so as to discover patterns.

Figure 14 Occupation Distribution of Customers

	User_ID	Occupation	Purchase_Amount
	<int>	<int>	<int>
1	1000001	10	333481
2	1000002	16	810353
3	1000003	15	341635
4	1000004	7	205987
5	1000005	20	821001
6	1000006	9	379450

	Occupation	Purchase_Amount
	<chr>	<int>
1	4	657530393
2	0	625814811
3	7	549282744
4	1	414552829
5	17	387240355
6	12	300672105



The table on the top left shows the customers with their occupation and their purchasing amount in dollars. In order to produce the Occupation distribution, we need to group together the total purchase amount for each Occupation by totaling the purchase amount from customers in the same occupation. We then convert Occupation to a character data type ([1] "character"), which is strings of numbers, indicating different occupations. The top 6 occupations with the highest total amount of purchase ($) are occupation 4 ($657530393), 0 ($625814811), 7 ($549282744), 1 ($414552829), 17 ($387240355), and 12 ($300672105). Unfortunately, we don't have the key to occupation in the data dictionary provided, so we couldn't classify the customers accordingly.

Figure 15 Sparse Matrix

```
transactions as itemMatrix in sparse format with
 5892 rows (elements/itemsets/transactions) and
 10539 columns (items) and a density of 0.008768598

most frequent items:
P00265242 P00110742 P00025442 P00112142 P00057642   (Other)
     1858      1591      1586      1539      1430    536489
```

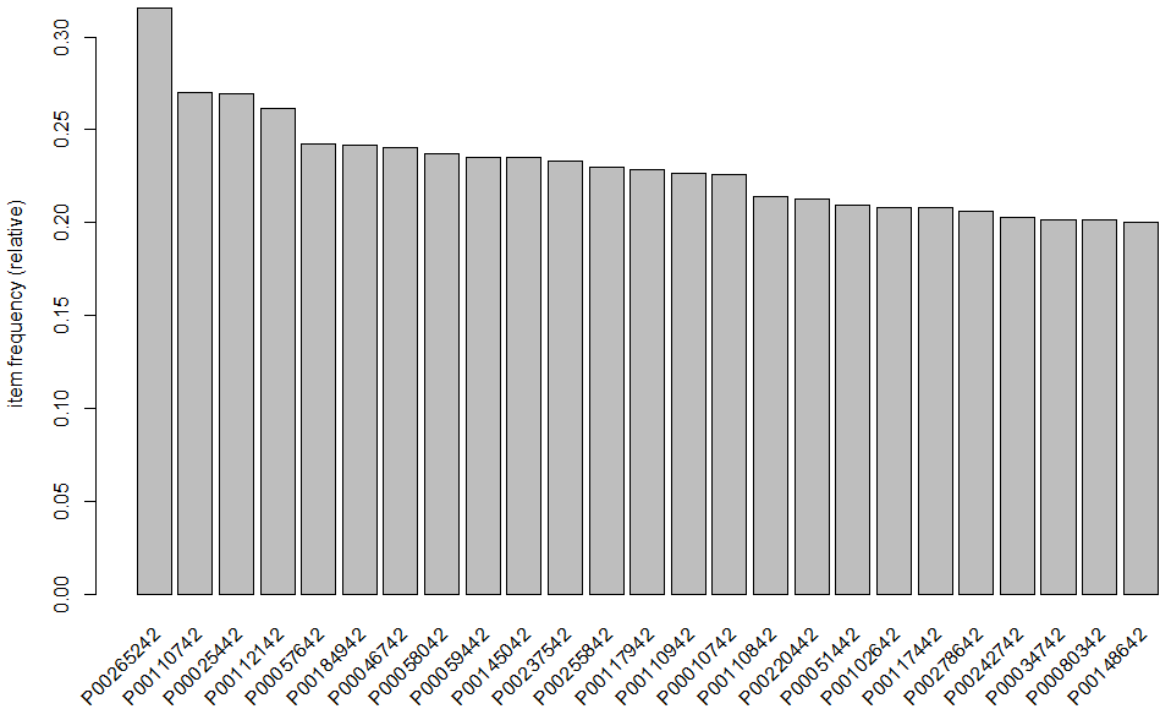
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.00   26.00   54.00   92.41  115.00 1026.00

includes extended item information - examples:
   labels
1 1000001
2 1000002
3 1000003
```

Starting from this figure, we are starting our process of Association Rules. In order for the Apriori algorithm. Apriori machine learning algorithms are used to make Association Rules regarding customer purchases. The arules package that we installed in R-Studio was developed specifically to deal with Association Rules and Frequent Itemset mining, which is what we are trying to achieve in this case. Before we begin our analysis, we must retrieve the necessary data from the original dataset and apply the correct formatting. For example, we are removing the User_ID row created earlier for EDA, as we no longer need it anymore. We also need to convert the dataset into the correct format. We need to convert the customers_products table into a sparse matrix and into binary numerical data types (0 and 1), as Apriori doesn't take strings or texts as inputs. We will be allocating a column for each individual product, and if a customer contains that certain product, it will be marked with a 1, otherwise, it will be marked as a 0. We are also importing the table as a csv file, and reading it using arules function of "read.transactions()" to get the sparse matrix, figure 15.

Looking at the second line of Sparse Matrix, we can see that there are 5892 elements/itemsets/transactions and 10539 items. We also get a density of 0.008768598 in our matrix. Recall that the Antecedent and Consequent together are called an item set, and there are 5892 itemsets and 10539 items altogether. The density tells us that we have 0.9% non-zero values (1) in our sparse matrix and 99.1% zero (0) values. Also, as we discovered in our EDA, the most frequent items that customers purchased will be useful in our association rules findings. It also shows that the most frequently bought items are P00265242, P00110742, P00025442, P00112142, and P00057642. The average number (mean) of items each customer purchased is 92.41. The interquartile range is 89, with the first quartile at 26 and the third quartile at 115. The minimum number of items bought is 6 and the maximum number of items bought is 1026. There are quite a few customers that have purchased over 1000 items, so it would be much more useful if we look at the median value (54) of items purchased instead of the mean because it can be heavily affected by outliers.

Figure 16 Product Frequency Plot



This visualizes the most frequently bought items, which is included in the arules package. This plot is limited to showing only the top 25 products, by choice because there are a lot of products.

Figure 17 Association Rules (Apriori)

```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target  ext
        0.8    0.1    1 none FALSE            TRUE       0   0.008      1     10  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 47

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[10539 item(s), 5892 transaction(s)] done [0.84s].
sorting and recoding items ... [2099 item(s)] done [0.03s].
creating transaction tree ... done [0.01s].
checking subsets of size 1 2 3 4 5 6 done [35.77s].
writing ... [7 rule(s)] done [0.72s].
creating S4 object  ... done [0.40s].
```

We produced the Association Rule model by setting parameters. The first parameters we will set are the support and confidence. The support value is the frequency of a specific item within the dataset.

Our support value will be the minimum number of transactions necessary divided by the total number of transactions. Setting the support value means that we are setting a minimum number of transactions necessary for our logical rules to take effect. In this case, our absolute minimum support count is 47. We have a total number of unique customer transactions of 5892 and a total of 10539 items or products. By looking at our dataset, let's assume that we want to choose a product which was purchased by at least 50 different customers. The support value can, therefore, be found with (50/5892) = .008486083.

The second value we need to find is confidence. Recall from Introduction, the confidence value determines the likelihood of how often a rule is to be found true. It is found by using conditional probability of buying the Consequent given that the Antecedent was first purchased. The minimum strength of any rule is a limit we placed when setting our minimum confidence value. We can start by using 80% as our minimum confidence value, as it is our default confidence value in the apriori() function in R-Studio. We can later adjust the parameters according to the results. For example, we can look at our Product_IDs and match it with recognizable names from the source of data, the confidence value can, then, be changed into something more relevant towards the results. However, we don't have the knowledge of the product names, so we will start with a value and then lower the confidence to see the different resulting logical rules. Maxtime = 0, will allow the apriori algorithm to run until completion with no time limit. At the second to last line we see that 7 rules are created in accordance with our specified parameters.

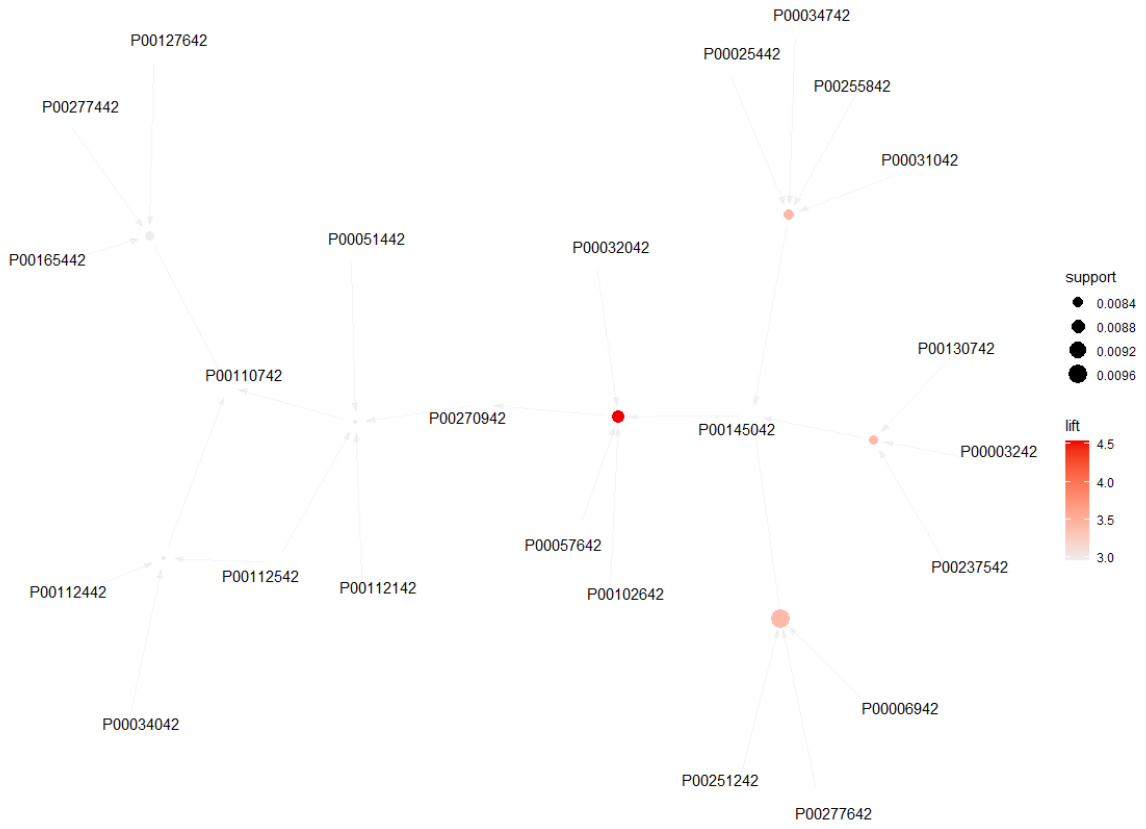
Figure 18 Seven Rules sorted by Lift

```
    lhs                                                rhs         support     confidence coverage
[1] {P00032042, P00057642, P00102642, P00145042} => {P00270942} 0.008655804 0.8793103  0.009843856
[2] {P00025442, P00031042, P00034742, P00255842} => {P00145042} 0.008486083 0.8064516  0.010522743
[3] {P00003242, P00130742, P00237542}            => {P00145042} 0.008316361 0.8032787  0.010353021
[4] {P00006942, P00251242, P00277642}            => {P00145042} 0.009674134 0.8028169  0.012050238
[5] {P00034042, P00112442, P00112542}            => {P00110742} 0.008146640 0.8135593  0.010013578
[6] {P00127642, P00165442, P00277442}            => {P00110742} 0.008316361 0.8032787  0.010353021
[7] {P00051442, P00112142, P00112542, P00270942} => {P00110742} 0.008146640 0.8000000  0.010183299
    lift     count
[1] 4.540663 51
[2] 3.433246 50
[3] 3.419738 49
[4] 3.417773 57
[5] 3.012880 48
[6] 2.974807 49
[7] 2.962665 48
```

There are 7 association rules that have been created according to the parameters we set. We ranked it from the highest lift to the lowest out of the seven. For rule 1, we can see that the Antecedent set of elements includes {P00032042, P00057642, P00102642, P00145042}. It is a group of items which the algorithm has pulled from the dataset. The rhs is the Consequent set, which is the value predicted by Apriori to be purchased with items in the "lhs" category. It also shows the support, confidence, and coverage values for the logical rule. If we look at rule 1, the likelihood of customers who bought

{P00032042, P00057642, P00102642, P00145042} products will also buy P00270942 (RHS) is 87.9%, with the support of 0.0087, which is the frequency that the events occur together. The similar can be said for Rule 2 to Rule 7. The lift value gives us the independence/dependence of a rule. It takes the confidence value and its relationship to the entire dataset into account. Rule number 1 has the highest lift value, in this case, 4.54. It means that if the customer purchased P00032042, P00057642, P00102642, and P00145042 products, the confidence in him/her purchasing a P00270942 goes up 4.54 times. In general, the larger the lift ratio is, compared to 1, the greater the strength of the association between the LHS (Antecedent) and the RHS (Consequent). The count is the number of times a rule occurred in our Black Friday dataset.

Figure 19 Visualization of the Association Rules



This figure visualizes our association rules in figure 16 by utilizing the arulesViz package. The arrows pointing from items to rule indicate LHS (grouped) items and arrows pointing from rules to items indicate the RHS (rule) items. The size of the circles indicate the support, larger circles represent higher support value. The gradient color of the circle represents the lift values, darker circles represent higher lift

values. Therefore, we can see that the logical rule with arrows pointing towards circles, {P00032042, P00057642, P00102642, P00145042}, is the association rule with highest lift, as we can also tell from the dark gradient circles. The rule with the highest support is rule 4, with a LHS {P00006942, P00251242, P00277642} with RHS {P00145042}, and a support of 0.0097. If customers bought products P00006942, P00251242, and P00277642, the likelihood of them buying product P00145042 is around 80.28%. Rule 4's lift is 3.42, meaning if the customer purchased P00006942, P00251242, and P00277642, the chances/confidence of them also buying P00145042 goes up by 3.42 times.

Figure 20 Modified Association Rules (Apriori)

```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target  ext
       0.75    0.1    1 none FALSE            TRUE       0   0.008      1     10  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 47

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[10539 item(s), 5892 transaction(s)] done [0.87s].
sorting and recoding items ... [2099 item(s)] done [0.06s].
creating transaction tree ... done [0.01s].
checking subsets of size 1 2 3 4
```

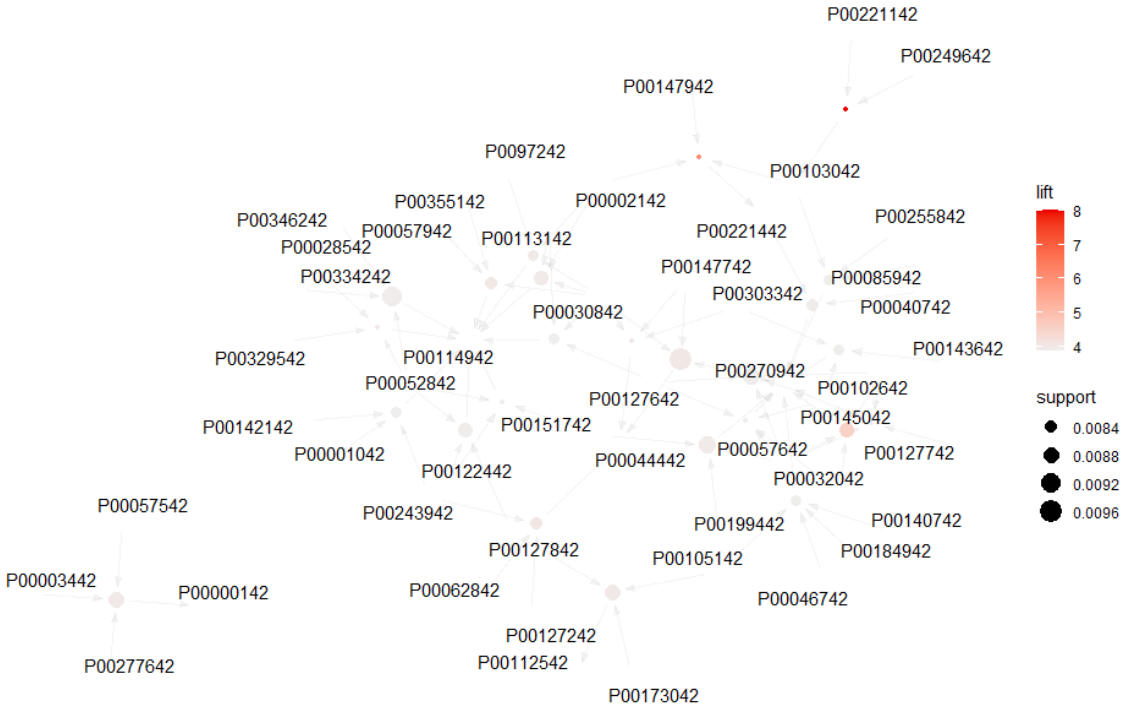
After looking at the result, we modified some of the parameters for the algorithm. Again , we don't have the key for Product_ID, so it's harder to make sense of this process. Therefore, we will only use the algorithm to modify once more. This time, we are decreasing our confidence value to 75% and keeping our old support value (0.008). Now that minimum confidence value is 75%, we have a total of 171 rules. This is significantly higher than our previous results, which only produced 7 association rules. Everything else remains the same as before.

Figure 21 Top 6 Association Rules after Modification

```
    lhs                                                rhs         support     confidence coverage
[1] {P00221142, P00249642}                          => {P00103042} 0.008146640 0.7619048  0.010692464
[2] {P00002142, P00103042, P00147942}               => {P00221442} 0.008146640 0.7500000  0.010862186
[3] {P00032042, P00057642, P00102642, P00145042} => {P00270942} 0.008655804 0.8793103  0.009843856
[4] {P00062842, P00127242, P00243942}               => {P00044442} 0.008486083 0.7575758  0.011201629
[5] {P00030842, P00057942, P00355142}               => {P00114942} 0.008486083 0.7936508  0.010692464
[6] {P00030842, P00147742, P00303342}               => {P00044442} 0.008146640 0.7500000  0.010862186
    lift     count
[1] 8.030667 48
[2] 6.045144 48
[3] 4.540663 51
[4] 4.061544 50
[5] 4.024260 50
[6] 4.020928 48
```

To examine the result easier, we are limiting the number of association rules we are looking at from 171 to 6. We can see that we have a completely new set of rules and the rule with highest lift value has also changed. Rule number 1 is an Antecedent (LHS) set of {P00221142, P00249642}, which means the customers who bought items P00221142 and P00249642 will also purchase item P00103042 (Consequent) approximately 76% (confidence) of the time, given a support of 0.008, which is the frequencies of this association rule happening. The lift of this rule is significantly high, with a value of 8.03. This means that if the customers bought P00221142 and P00249642, the chances/confidence that they will also buy P00103042 increases by 8.03 times.

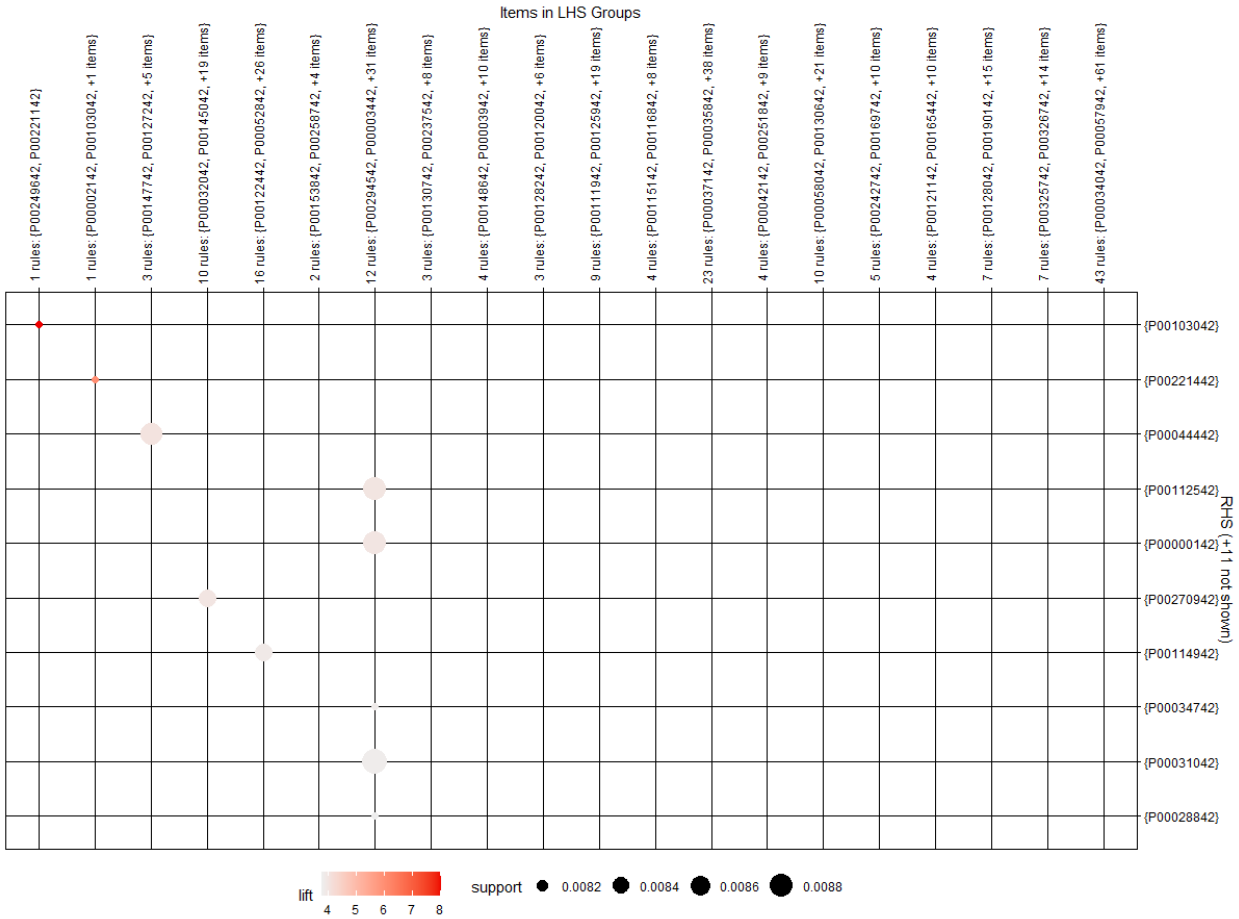
Figure 22 Visualization of the Association Rules after Modification



We have a total of 171 rules in this visualization, which also makes it much harder to interpret. Therefore, we will create a matrix that depicts a similar result, but clearer interpretation. This is done by using the “grouped” method.

Figure 23 Association Rules Matrix

```
Available control parameters (with default values):
k         =  20
aggr.fun          =  function (x, ...)  UseMethod("mean")
rhs_max   =  10
lhs_label_items   =  2
col       =  c("#EE0000FF", "#EEEEEEFF")
groups    =  NULL
engine    =  ggplot2
verbose   =  FALSE
```



The items on the top of the graph are the LHS, and the items listed on the right hand side of the graph are the corresponding RHS. Again, the size of the circles represents the support value of the association rules, and the gradient of the color represents the lift of the association rules. The association rule that has the largest circle is third from the left lhs and third from the top rhs {P00044442}. There are 3 association rules that have the same exact lhs. The rule that has the highest lift value is rule that is on the most left and top, with lhs of {P00249642, P00221142} and rhs of {P00103042}.

Conclusion

The aim of this assignment is to use the Association Rules to find the association rules between products purchased by the customers during Black Friday. In the first part of the executive summary, I used the Exploratory Data Analysis (EDA) to explore the observations using visualizations by grouping.

Figure 1 shows the average spending according to gender of the customers of the retail store on Black Friday. The male population made a higher dollar amount of purchases ($911,963) compared to the female population ($699,054). It is likely that the products are more appealing towards the male population. Figure 9 shows the distribution of purchase amount for the entire population. The mean is $851752, and the median is $512,612. The peak purchase amount is around $300,000. The average purchase amount of male population is higher than the mean and the median for the retail store on Black Friday. The average purchase amount of the female population is higher than the median, but lower than the mean for the retail store on Black Friday. This further proves that the products are likely to be more appealing towards male population rather than the female population.

Figure 2 shows the number and proportion of customers by gender for total population and best selling product. The retail store attracts more male customers than female customers, as the number of male customers are greater than the number of female customers. Number of male customers is significantly higher in all 3 graphs, where the male population is greater than the female population. The distribution and proportion of gender is similar for both total population and best selling products, with a ratio of female to male 0.3:0.7.

Figure 3 shows the age distribution for both total population and best selling product. Again, both graphs depict a similar picture. Figure 12 also shows a similar age distribution for City A, City B, and City C. This means that Black Friday and the best selling product may have similar or common characteristics that attract the population between the ages of 26 to 35. It is also likely that this age group population has more shopping sprees than all other age group populations, as people at this age tend to lean towards purchasing their wants more than their needs on Black Friday.

Figure 4 shows the number of customers in City A, City B, and City C. It also shows the total purchase amount according to different City Categories. Although City C has the most population, City B has the highest total purchase amount. It is likely that City B has higher purchasing power compared to other two cities. It is also likely that they have different cultures, which leads to different purchasing decisions and behaviors. Figure 5 shows the total number of purchases according to the City Categories. It displays a very similar picture compared to the graph of total purchase amount by City Categories. This makes sense as the number of purchases increase, the total amount of purchase in dollars will also be likely to increase, further proving the fact that City B has the highest purchasing power amongst the three cities observed. However, in figure 6 we also see a different picture, where City C has the highest number

of purchases with the best selling product. It is a surprising result because age distribution is the same for total population and best selling product, but it's not the same in terms of City Categories. It is likely that the best selling product is not one of the expensive products, therefore high sales doesn't result in higher dollar sales. It is important to know which city has the highest number of sales and purchasing amount because the retail store can, then, focus in City B to build better customer services and products to gain more customers that have higher purchasing power.

Figure 7 shows the number of customers that stay in their current cities arranged by different number of years. The most frequent value, in this case, is staying 1 year in their current cities, followed by 2 years, 3 years, 4 or more years, and, lastly, less than 1 year. It is not a surprising result, as it is hard to move around all the time, so not a lot of people are likely to do that, let alone when we are only looking at the customers from our retail stores. Figure 8 sorted the customers that stayed in their city by different cities. It is not surprising to see that City C has the highest proportion of all types of Stay in current city customers, as it does have the highest number of customers in general. However, this can still be an insightful piece of information, as it shows that City C's retail store has been doing really well by keeping their customers loyal. It is also important to keep building a good reputation and relationship with the customers, so that they would come back to the retail store for other products in the future.

Figure 10 depicts the number of customers with different marital status for the retail store. The ratio of single customers to married customers is around 1.36:1, meaning for every single customer that purchased at the store, there will be 1.36 married customers that will make a purchase at the store, metaphorically speaking. Figure 11 also demonstrates a similar relationship between the proportion of single customers to married customers. The higher proportion of single customers is likely due to the fact that married customers usually have kids or spouses that leads to better budgeting and usually buying only necessities for the family. While the single customers usually just buy whatever they find appealing, resulting in giving-in on Black Friday promotions and sales. The proportion of single customers is highest in City A, followed by City B, then City C. This is important in a way that retail stores can change their layout according to the type of customers they get and they can plan marketing strategies in a way that makes it more appealing towards the majority of their customers.

Figure 14 shows the occupation distribution of the customers in their retail store on Black Friday. The top 6 occupations with the highest total amount of purchase ($) are occupation 4 ($657530393), 0 ($625814811), 7 ($549282744), 1 ($414552829), 17 ($387240355), and 12 ($300672105). Unfortunately, we don't have the key to occupation in the data dictionary provided, so we couldn't classify the customers accordingly and not much insight can be gained by only knowing the identification of the occupations.

Starting from figure 15, the Association Rules analysis begins. There are 5892 transactions and 10539 products in our dataset. We also get a density of 0.00877 from our matrix. The density tells us that

we have 0.9% of non-zero values and 99.1% zero values. We also learned that the most frequently bought items in the retail store on Black Friday are P00265242, P00110742, P00025442, P00112142, and P00057642. Figure 16 visualizes the most frequently bought products with a bar graph. The average number of products bought by a customer is 92.41. However, it is more insightful to look at the median, which is 54 products because there are quite a few customers that have purchased over 1000 items, and those outliers can heavily affect the mean number of items bought.

Figure 17 shows that we created 7 association rules by setting the minimum number of transactions necessary, which is our minimum support value, which is 47. We also want our product to be bought by at least 50 customers, so the support value can be found with (50/5892) = 0.008. The second value we need to set is the confidence value. We started by using 80% as our minimum confidence value, as it is also a default confidence value in the Apriori. The maxtime = 0, allows us to run the Apriori without a time limit, so we won't be stopped from running too long. We will be adjusting the parameters the second time after seeing the results, to fit better with the result. Figure 18 shows the 7 association rules created by Apriori. Rule 1 has the highest number of lifts (4.54). The likelihood of customers who bought {P00032042, P00057642, P00102642, P00145042} products will also buy P00270942 (RHS) is 87.9%, with the support of 0.0087, which is the frequency that the events occur together. If the customer purchased P00032042, P00057642, P00102642, and P00145042 products, the confidence in him/her purchasing a P00270942 goes up 4.54 times. The amount of times Rule 1 appeared in the dataset is 51 times, meaning 51 exact same transactions happened in the retail store. Figure 19 shows the visualization of the result in figure 18.

Figure 20 is the modified association rules. We tweaked the parameters by decreasing the confidence value to 75% and keeping the support value the same (0.008). A total of 171 association rules shows up in the result. Figure 21 shows the top 6 association rules from the 171 rules we derived. Rule number 1 is an Antecedent (LHS) set of {P00221142, P00249642}, which means the customers who bought items P00221142 and P00249642 will also purchase item P00103042 (Consequent) approximately 76% (confidence) of the time, given a support of 0.008, which is the frequencies of this association rule happening. The lift of this rule is significantly high, with a value of 8.03. This means that if the customers bought P00221142 and P00249642, the chances/confidence that they will also buy P00103042 increases by 8.03 times. Figure 22 is the visualization of the 171 rules we derived by using Apriori in figure 20. Figure 23 is the better and clearer version of figure 22, and it is done by using the "grouped" method. The rule that has the highest lift value is rule that is on the most left and top, with lhs of {P00249642, P00221142} and rhs of {P00103042}.