1. Question 1

1.1 Preparing the dataset

First step was to exclude all the rows that include missing values. It is typically safe to remove missing data because the results will be unbiased. The test may not be as powerful, but the results will be reliable.

Next step was to split the column *InvoiceDate* to *InvoiceDate* and *InvoiceTime*. All the rows with negative quantities were also excluded from the dataset. Identification and removal of duplicated rows was also applied.

By observing the type of variables, we set as factor variable *Country*, as date *InvoiceDate* and as numeric *InvoiceNo.*

In order to continue with market basket analysis, we had to transform the data into proper format. We sorted transactions by *InvoiceNo* and then we transformed the data so that each row will correspond as one transaction. Same InvoiceNo and *InvoiceDate* will mean one transaction, so all items purchased in the same transaction will be in one cell. By transforming the data is such form we had the opportunity to create a statistic of how many transactions happened on each day during the time period given.

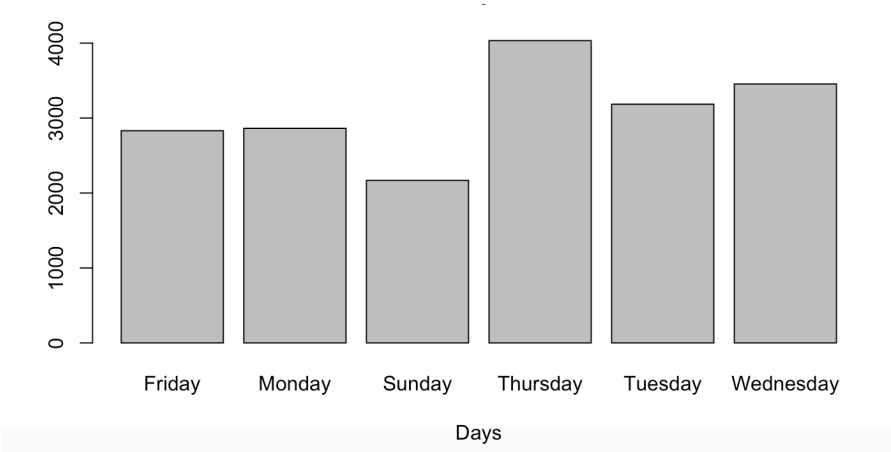

Figure 1.1: Total transactions on each day

As we can see from the bar chart Thursday was the most popular day of purchases. We also noticed that Saturdays have 0 purchases (possible reason is that the online – retail is not operating on Saturdays)

Last step on the data preparation was to read the final transaction data called *ItemList* into R-studio.

1.2 Summary information of the data-frame

From the element analysis distribution analysis, we found out that from all the recorded transactions, the most common number of items per transaction is 13, minimum item is 1 and the maximum items is 419.

The most bought item is WHITE HANGING HEART T-LIGHT HOLDER with 1760 purchases.

The transformation of baskets to sparse matrix results in a matrix of 7765 columns and 18536 rows. Column are all the items and rows are the transactions.

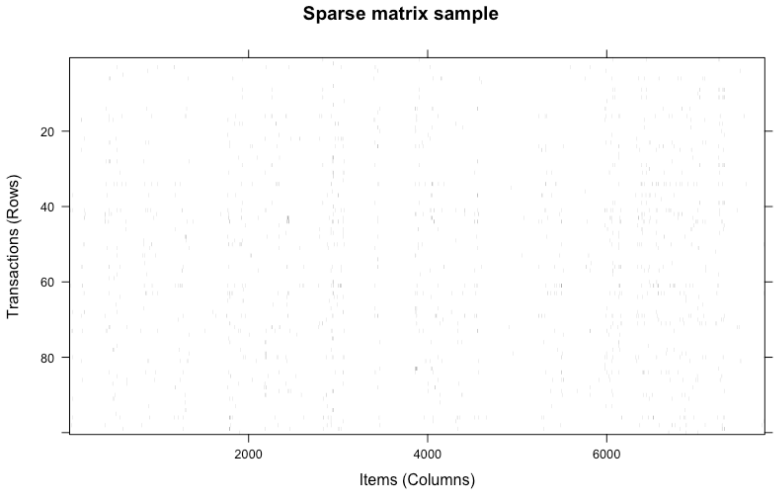

Figure 1.2: Small visualization of the sparse matrix

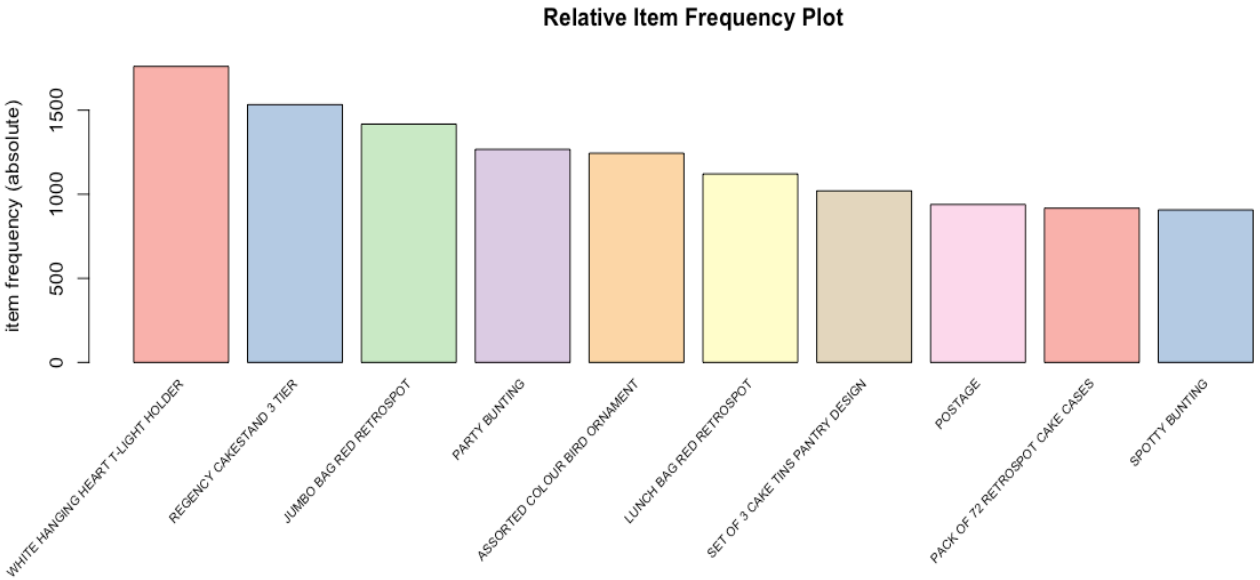

Figure 1.3: Frequency plot of the top 10 most bought items

1.3 Discover Association Rules

We used support = 0.001 and confidence = 0.8. As a result, we get 115245 rules. A length of 6 items has the most rules. A length of 2 items has the least rules. Since we have 115245 rules rules only the top 10 rules with the higher lift are shown below;

LHS		RHS
BILLBOARD FONTS DESIGN	→	WRAP
WRAP	→	BILLBOARD FONTS DESIGN
MIRRORED WALL ART LADIES	→	MIRRORED WALL ART GENTS
PARTY PIZZA DISH PINK POLKADOT, PARTY PIZZA DISH RED RETROSPOT	→	PARTY PIZZA DISH GREEN POLKADOT
PARTY PIZZA DISH BLUE POLKADOT, PARTY PIZZA DISH RED RETROSPOT	→	PARTY PIZZA DISH GREEN POLKADOT
DOG LICENCE WALL ART	→	BICYCLE SAFTEY WALL ART
FUNK MONKEY	→	ART LIGHTS
ART LIGHTS	→	FUNK MONKEY
PARTY PIZZA DISH GREEN POLKADOT, PARTY PIZZA DISH PINK POLKADOT	→	PARTY PIZZA DISH BLUE POLKADOT
PARTY PIZZA DISH PINK POLKADOT, PARTY PIZZA DISH RED RETROSPOT	→	PARTY PIZZA DISH BLUE POLKADOT

We can visualize the above information in the graph below.

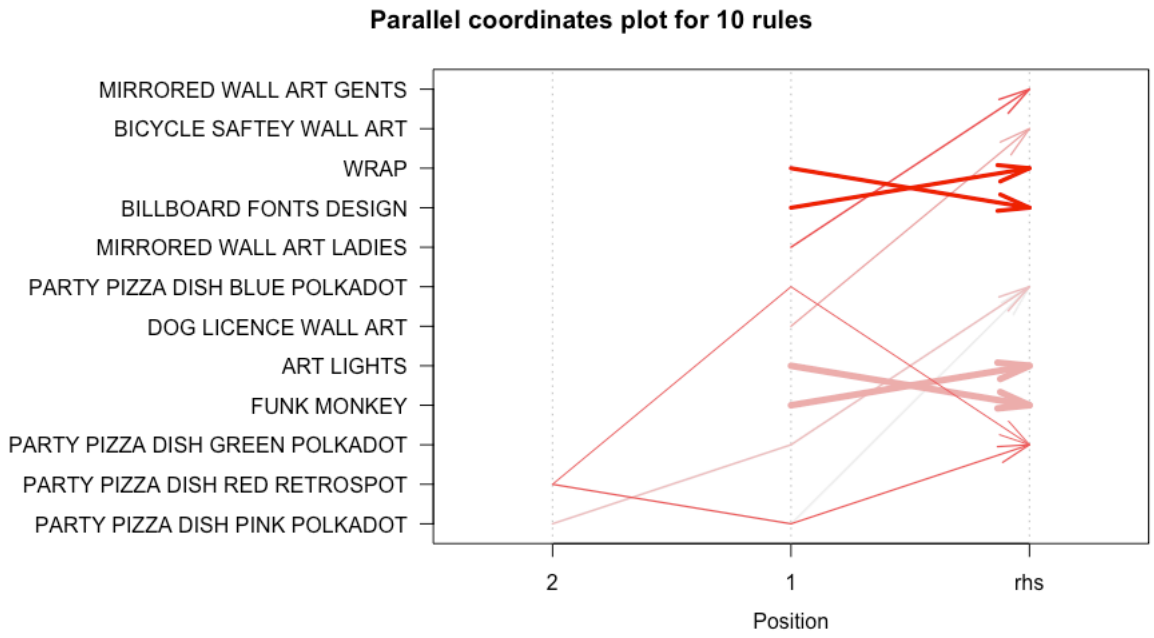

Figure 1.4: Visualization plot of the top 10 rules sort by lift

The positions are in the LHS where 2 is the most recent addition to our basket and 1 is the item we previously had. For example, when someone has in his/her basket PARTY PIZZA DISH RED RETROSPOT and PARTY PIZZA DISH PINK POLKADOT he/she is more likely to end up getting PARTY PIZZA DISH GREEN POLKADOT. The same logic applies to the rest of the arrow combinations.

1.4 Visualization of all the rules

Lastly in order to visualize all the rules with all their features (support, confidence, lift) we produced the following plot. Each dot represents one rule. The higher the lift, the darker the dot. As we can see most of the rules have low support level below 0.01

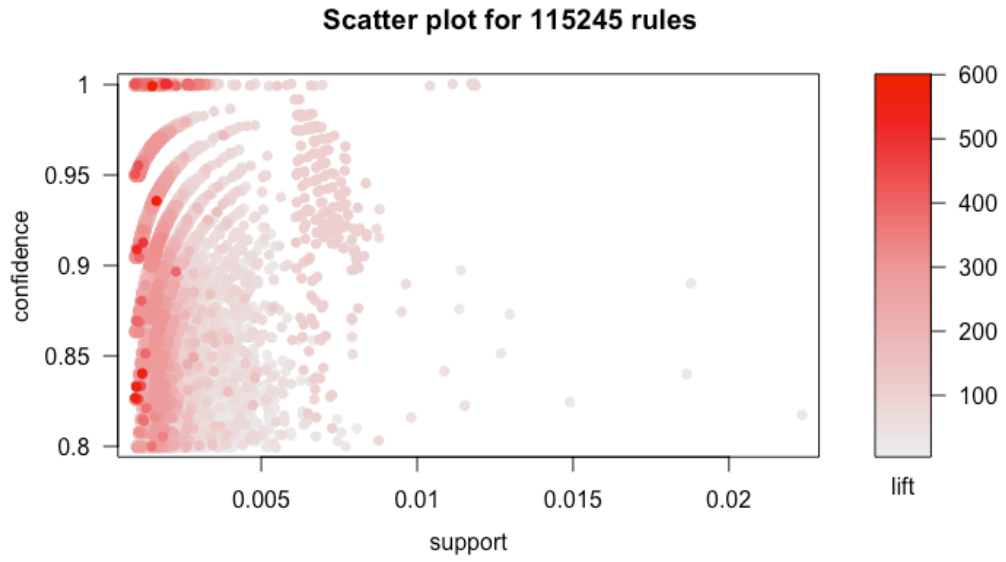

Figure 1.5: Visualization of the all the rules with the coloring based on lift

The only difference in that plot is that the last feature is order instead of lift. It uses order for coloring. In this plot we can see that the order 10 rules have the lowest support level. The general pattern obtained is that the higher the number of the order the smaller the support.

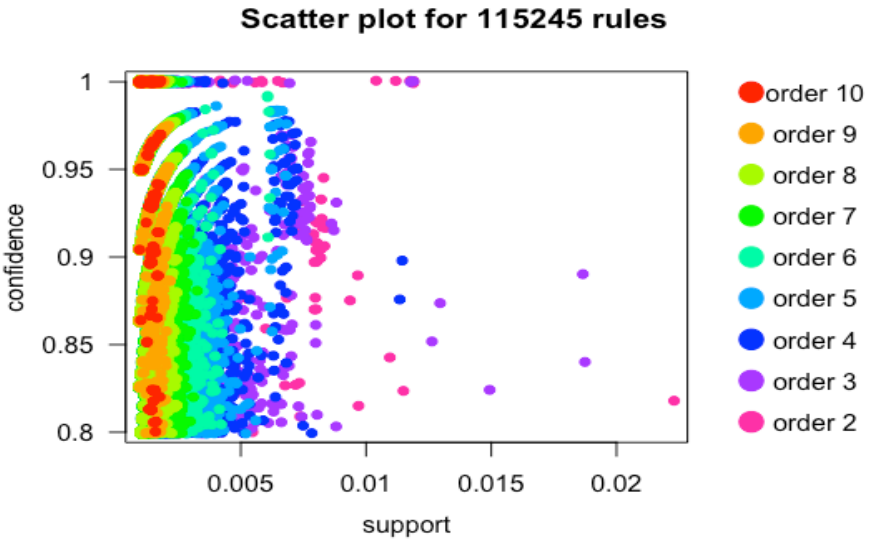

Figure 1.6: Visualization of the all the rules with the coloring based on order