UNIVERSITY of GREENWICH

Course Title: Data Visualization COMP1800

Student Name: Mohamed Al-Kaisi

Student ID: 000931504

Contents

Table of Figures

ChrisCo Company

Introduction

Data visualization defined in What is data visualization, and why is it important? 2020 as a practice of translating information into a visual context using graph or map as representation. Representing information in a visual context makes it easier to for the human brain to understand and pull insight from the data. The goal of data visualization is to make it easier for identifying patterns, trends, and outliers in large datasets. Visualization is one of the steps in the data science process done after the data collected to visualize for a conclusion to be made. The practice of visualization helps businesses to understand which factor affect customer behavior, pinpoint areas that need to be improved or need more attention and making the stakeholder understand when and where to place a specific product and predict sales volumes.

Findings

The data received from ChrisCo company contains Daily Customers, Marketing, overhead, size, and staff files. All the data files given observe information about the 40 stores that

Chrisco Company owns. The aims to explore the data provided to us at the start and focus for a start on the main file of data, which was the Daily Customer containing the number of customers, attended their stores over a year.

1. Line chart

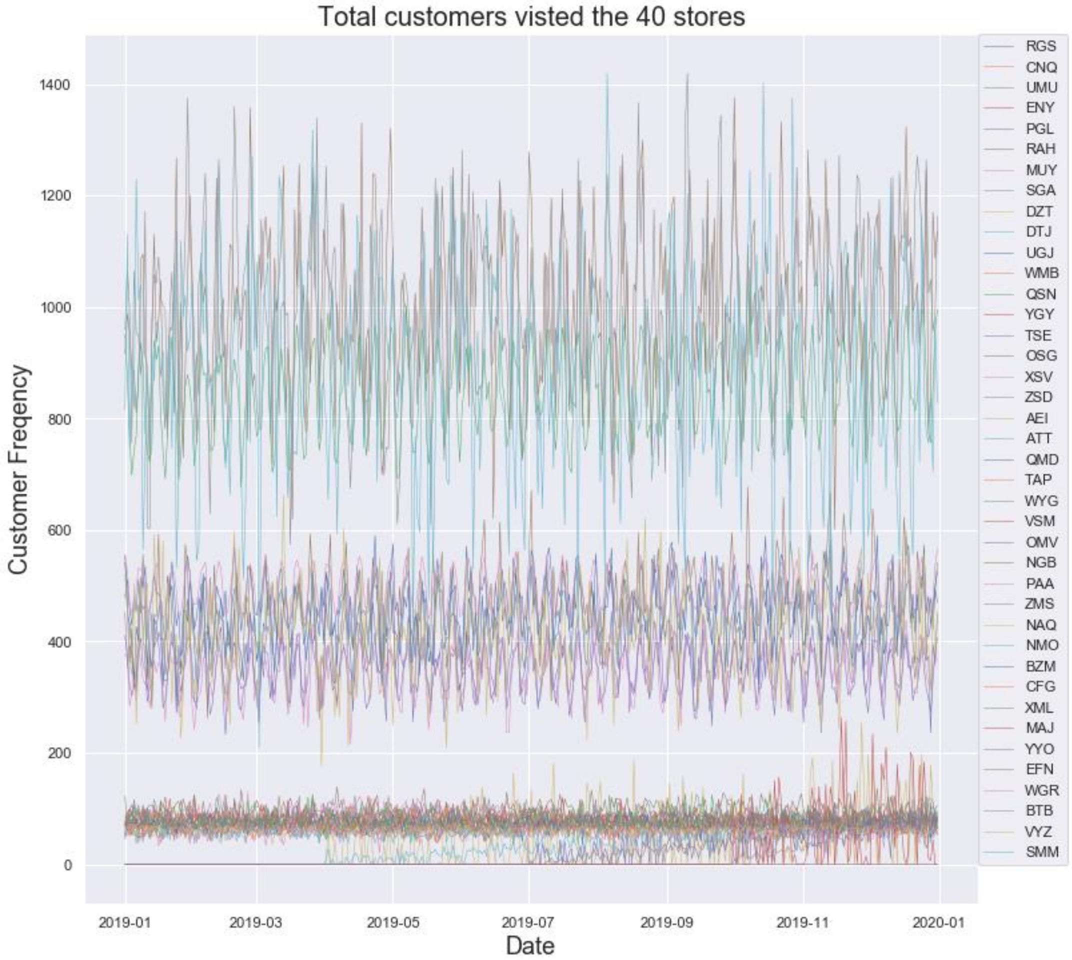
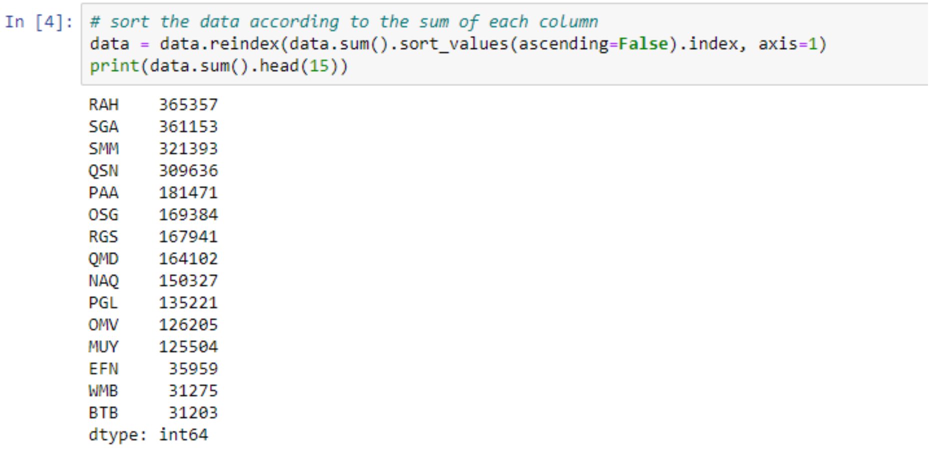


Figure 1 Showing time series plot of all the daily customers who visited 40 stores

Figure 1-line plot used as the go-to plot for visualizing time-series data such as measuring several points in time and allows us to show trends along time. The daily customer CSV file contains date and customers attended each store daily, so it was an obvious choice to use a line plot for showing time-series data to identify seasonality and trend.

The figure 1-line plot chart showing a display of all the 40 stores in terms of the daily customers attended. From the plotline above, we can understand three different types of volume customers visited different stores. Approximately the plotline shows from 600 and upwards are the highest attendance of customers to which specifics stores that conclude a colour for each store, as seen in the plot. The second observation was that from 200 to 600 would be classed as medium volume stores, and lastly, from 0 to 200 approximately would be class smallest stores with customer volume frequency. After identifying the three types of customers volumes attending the 40 stores from Figure 1, but there was too much noise to be able to locate the store names. There is another way of identifying the type of volumes but more clearly to identify the names of stores as well buy sorting and summing the values for each store to help identify three types such as High, medium, and small. Figure 2 shows all the summed values of stores and sorted in order. The values starting from 300000 and upwards tend to be the highest. The values starting 100000 up to 300000 classed to be the medium volume of customers. The values from 35959 and downgraded to be the smallest size but the highest of the smallest size. Figure 2 showing only the most top 15 using a head function, which means there is more than three smallest sizes of volumes than showing in fig 2.



```
In [4]: # sort the data according to the sum of each column
        data = data.reindex(data.sum().sort_values(ascending=False).index, axis=1)
        print(data.sum().head(15))
```

```
RAH    365357
SGA    361153
SMM    321393
QSN    309636
PAA    181471
OSG    169384
RGS    167941
QMD    164102
NAQ    150327
PGL    135221
OMV    126205
MUY    125504
EFN     35959
WMB     31275
BTB     31203
dtype: int64
```

Figure 2 Summed and sorted values to identify different type of customer volumes

Highest stores

- RAH
- SGA
- SMM
- QSN

Medium stores

- PAA
- OSG
- RGS
- QMD
- NAQ
- PGL
- OMV
- MUY

Smallest stores and there are more but in fig only showing the highest of smallest stores

- EFN
- WMB
- BTB

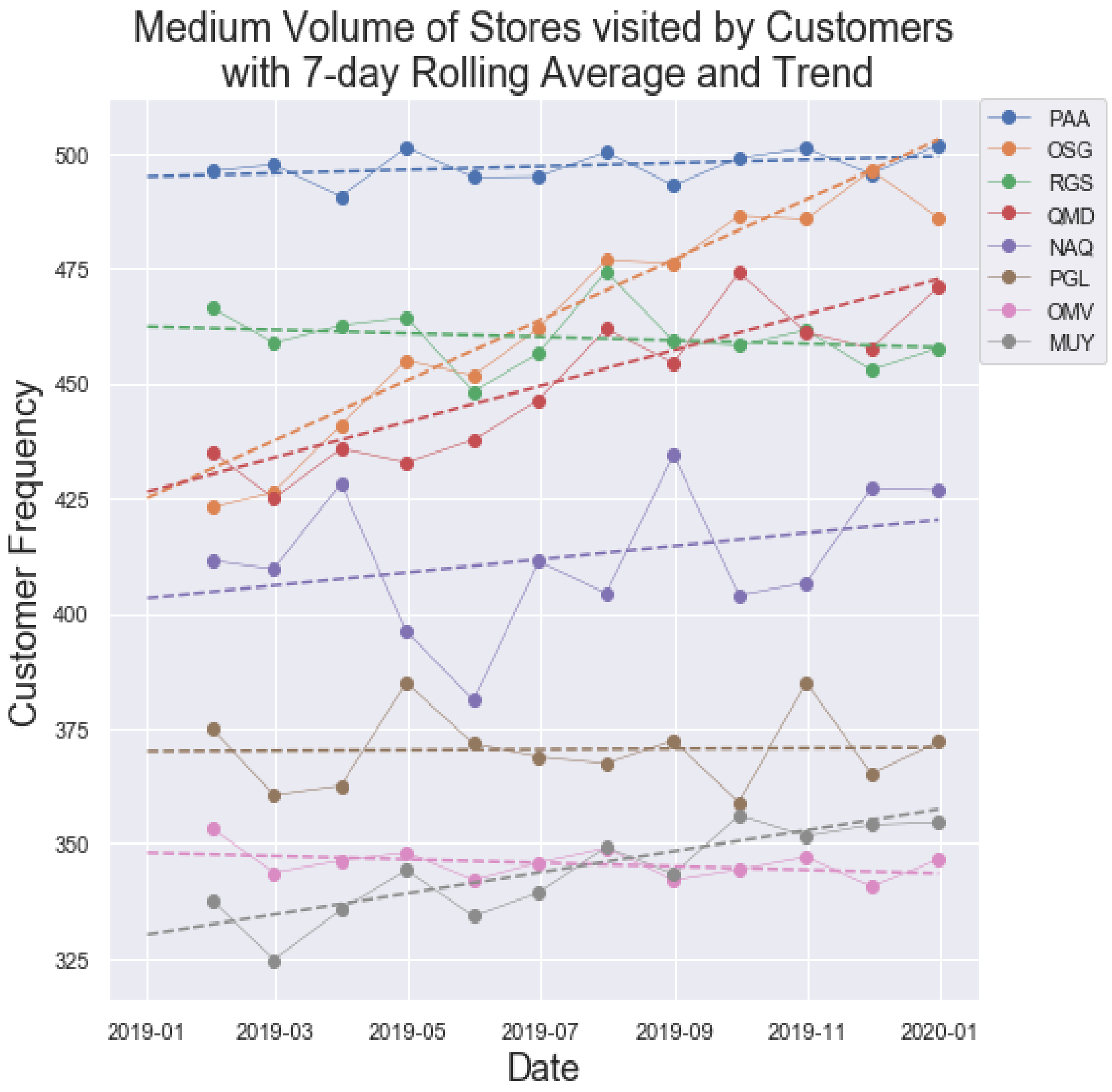


Figure 3 Showing time series data by showing trend line and rolling average for medium stores

Fig 1 chart showed loads of noise as it was showing all the 40 stores, but identifying and classifying stores in volumes type now, we can show fewer data to understand it more by only using the average. Rolling average helps to shows, for example, monthly instead of plotting daily data to remove some of the noise in fig1. Figure 3 rolling average has been applied to the data and only displaying data for the medium stores. Another method used in figure 3 to make sense of the data of medium stores by adding a trendline method, which helps to understand whether the straight line for each store is increasing over time or decreasing. Figure 3 shows the medium store's data using a monthly rolling average and showing the trend of each of the selected stores. The trend gives us a clear idea of how the stores whether customers increases throughout the year or decrease. All the medium stores were showing customers' growth as months moving apart from OMV and RGS, showing a bit decrease with monthly movement.

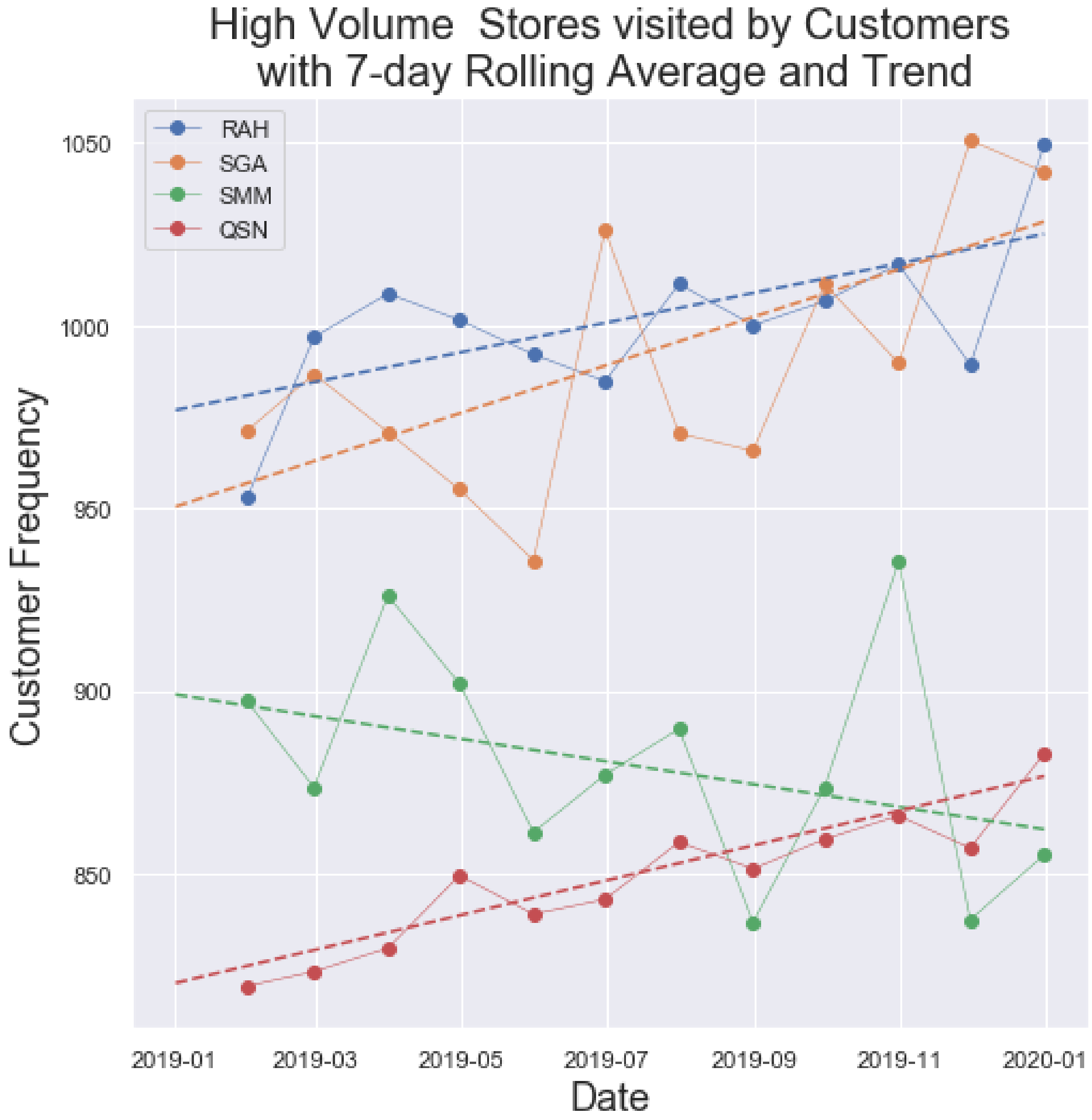


Figure 4 showing time-series data by showing a trend line and rolling average for high volume stores

The highest stores all showing to have increasing customers as month moving to look at the trend line or can be preferred as linear regressions. However, only SMM stores are showing to lose customers as months moving. The monthly rolling average and trend help to understand that customers attending will increase mostly, but some stores will decrease throughout the year. For ChrisCo company to improve customer store attraction, then possibly need also to understand how customers' weekly attendance in terms is what has the most presence and the lowest presence of the week.

2. Box Plot

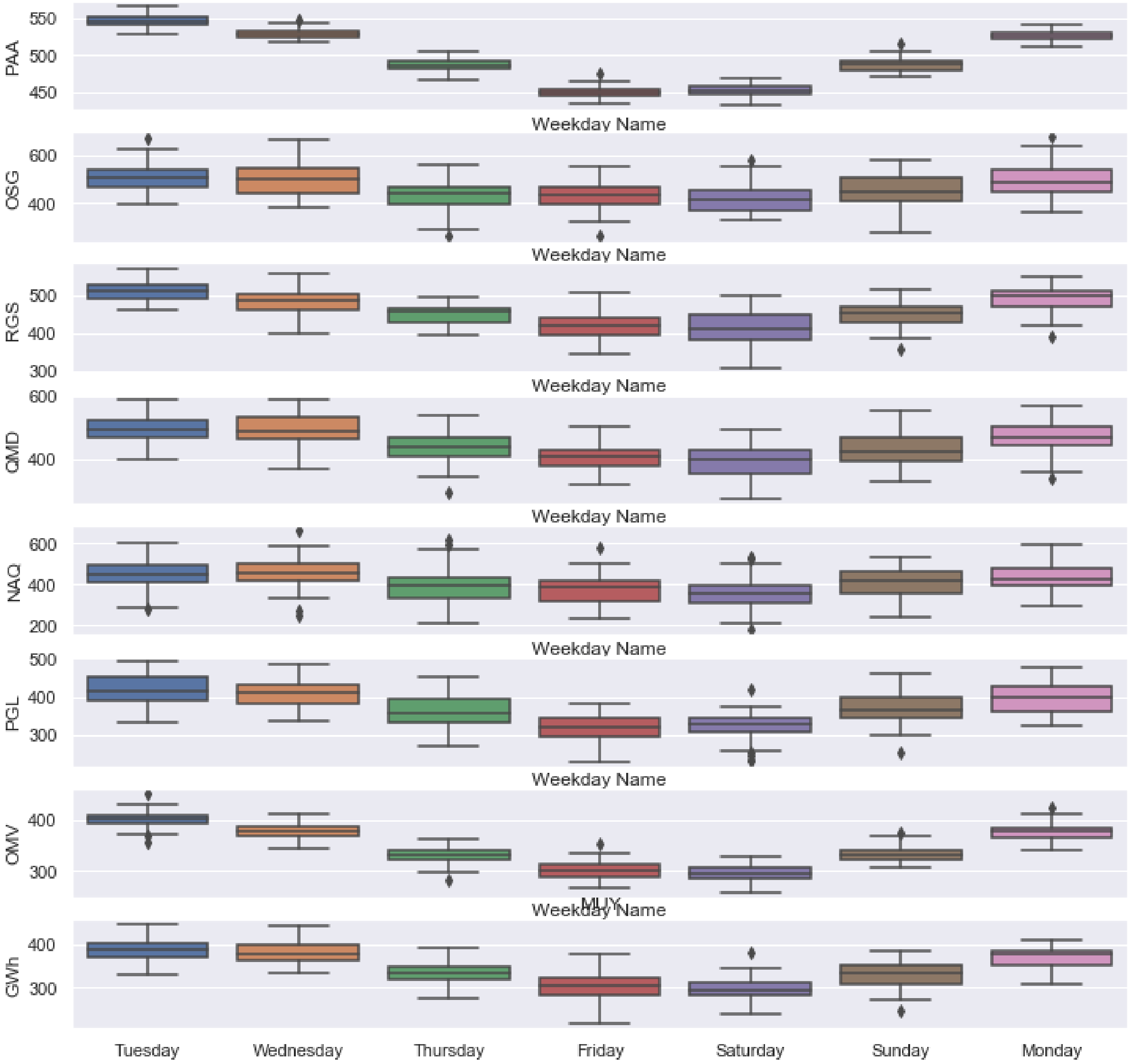


Figure 5 Box plot showing distribution customers over a week attending medium stores

Boxplot used in this aspect used to visualize the distribution of customers over a week instead of months. Using a Box plot helps to visualize more than one variable, such as the maximum value, median, and minim value, which in our aspect gave us an overall idea of what happened each day. Looking at the box plot fig 5, we can see declining mostly on Fridays, and some have the lowest customer every Saturday all for the medium stores.

3. Box Plot

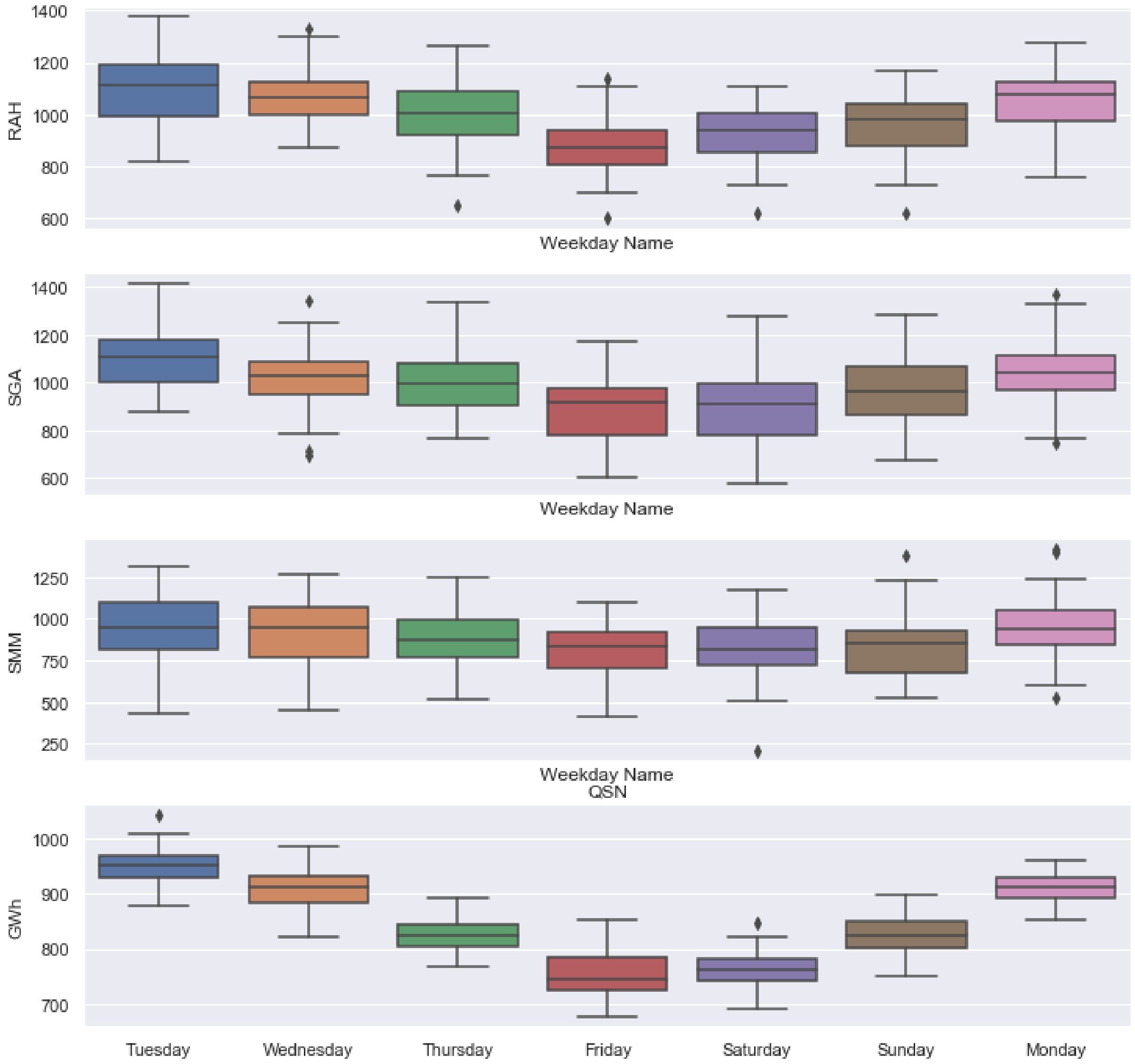
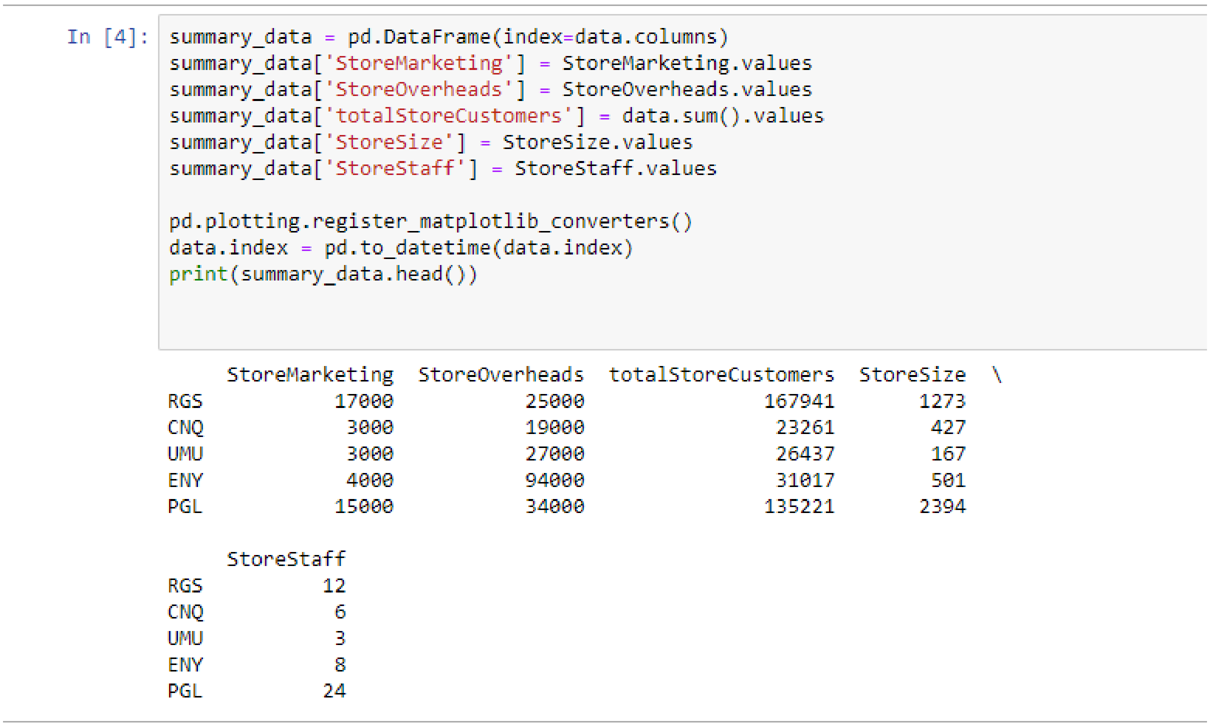


Figure 6 Box plot showing distribution customers over a week attending highest stores in terms of customers volumes

Figure 6 box plot showing the distribution of customers over a week for the highest stores. The plot shows that Friday and Saturday tend to be the least number of customers visiting the most top stores.

We have discovered from the daily customer file several things such as some of the stores started with the highest customers visiting then throughout the year's customers seeing declined—most of the store's customers increasing as the year the going. We also identified the stores into three types, such as small, medium, and high depends on the number of customers visited. The next step is to Merage the rest of the files together as independent variables to be able to visualize all the components for each store. Figure 7 shows that all the variables added together which each contain a value for all the 40 stores such as like how much spent of marketing, overhead, staff and the size for store RAH.



```
In [4]: summary_data = pd.DataFrame(index=data.columns)
        summary_data['StoreMarketing'] = StoreMarketing.values
        summary_data['StoreOverheads'] = StoreOverheads.values
        summary_data['totalStoreCustomers'] = data.sum().values
        summary_data['StoreSize'] = StoreSize.values
        summary_data['StoreStaff'] = StoreStaff.values

        pd.plotting.register_matplotlib_converters()
        data.index = pd.to_datetime(data.index)
        print(summary_data.head())
```

```
     StoreMarketing  StoreOverheads  totalStoreCustomers  StoreSize  \
RGS           17000           25000               167941       1273
CNQ            3000           19000                23261        427
UMU            3000           27000                26437        167
ENY            4000           94000                31017        501
PGL           15000           34000               135221       2394

     StoreStaff
RGS          12
CNQ           6
UMU           3
ENY           8
PGL          24
```

Figure 7 Adding all the files together as all has information about the 40 stores

Figure 7 showing all the other files about the 40 stores merged with daily customers to be able to visualize all the data together for further insights. In this case, classified as independent variables and the regular customers, all the values summed into one value for each store. Before in daily customer file, we had a year, data showing the number of customers visiting each store daily, but to fit in this aspect with the other variables, we need to make some value the values for each store into one. Aggregation function used in the element called sum (), which helps to add all the values together for each store.

After adding all the variables, the next task running an investigation to find out if there is a mutual relationship between two or more variables, which know as correlation relationship. From the correlation formed, we can understand when there is a positive or negative correlation between the variable. However, if two variables have two correlation between each other does not mean one causes the other. In our case, this will be the start of the investigation. Visualizing correlation is usually done using a scatter plot.

4. Scatter Graph

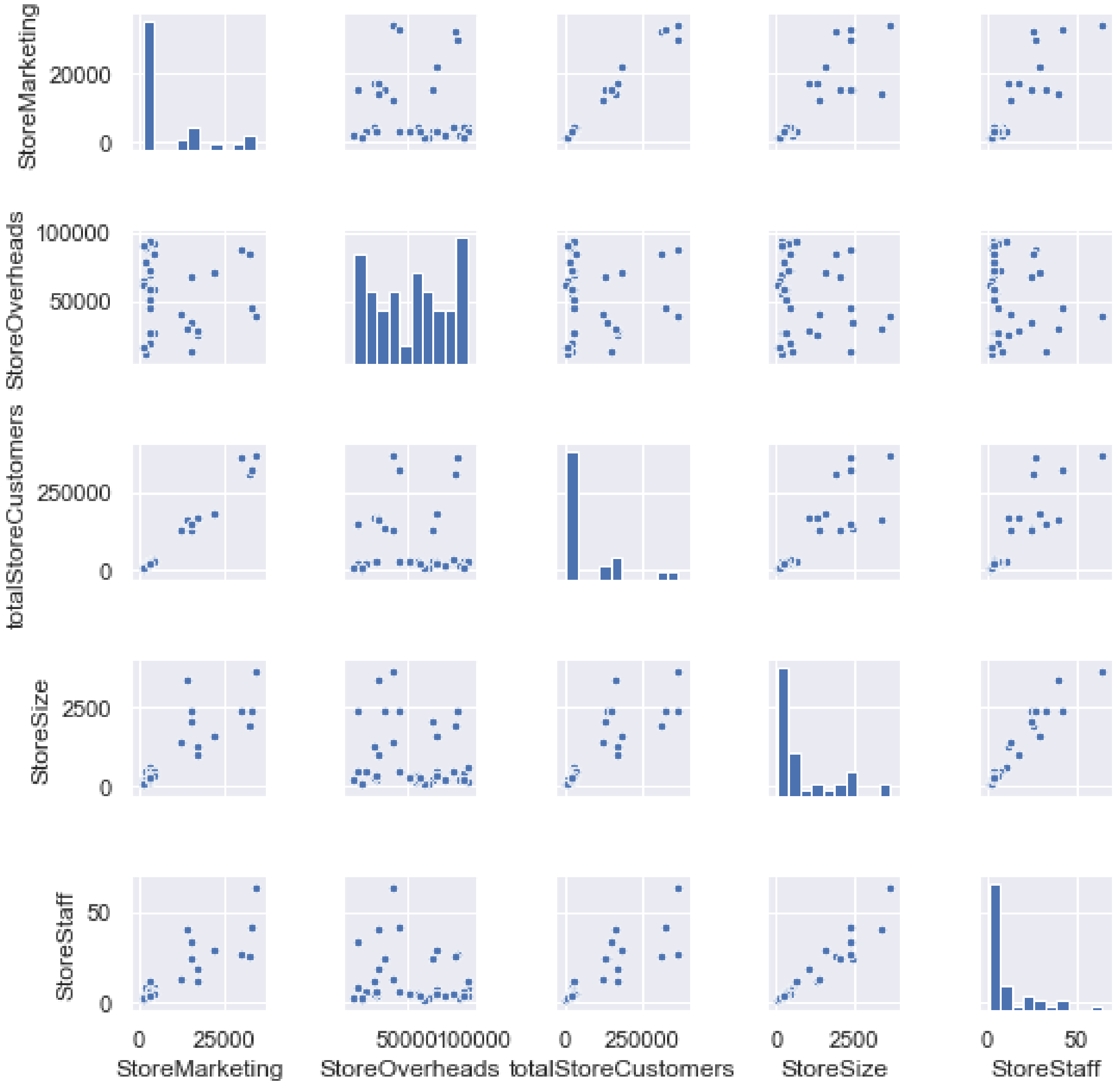


Figure 8 Subplot scatter graph comparing the variables against each other

A Scatter subplot included in this report to explain the correlation relationship between the variables and to help us discover more information about the stores. The subplot scatter graph helped compare each variable against each other all in one plot, which makes it easier to find any relationship.

Store Marketing

- Store marketing have a low positive correlation with store staff and store size
- Store marketing has a high positive correlation with total customer size
- Store marketing does not correlate with store overhead

Store overhead

- Store overhead does not correlate with any other variables

Total Store Customers

- Total store customer has a low positive correlation with store size.
- Total store customer has a low positive correlation with store staff.
- Total store customer does not correlate with overheads
- Full store customer has a perfect positive correlation.

Store size

- Store size have a high positive correlation with store staff
- Store size low positive correlation with total store customer.
- Store size does not correlate with overheads
- Store size have a low positive correlation with store marketing

Store Staff

- Store staff have a high positive correlation with store size
- Store staff have a low positive correlation with store marketing and total store customers

5. Heatmap

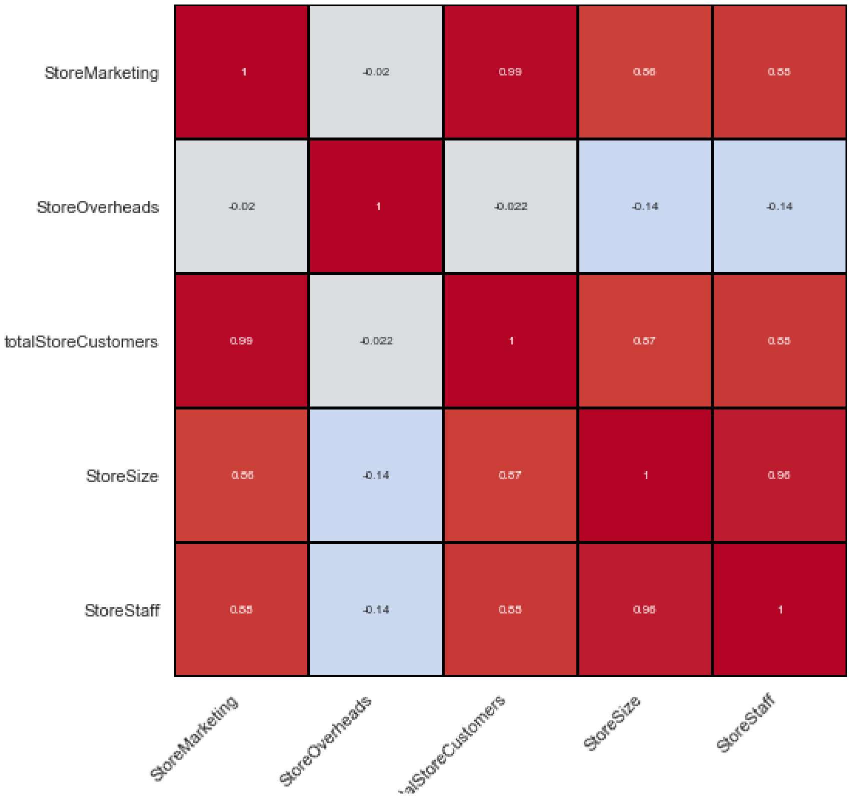


Figure 9 The heatmap plot show the highest correlation between the variables

Scatter graph in figure 8 can be tedious, and error-prone comparing each store to each other can be long and challenging to fit all the stores. However, using heatmap gives you a definite answer in terms of which variables are highly correlated. Using heatmap can make it more efficient to shows the variables that highly coded using the number system and color-coded. The blue represents cold (-1), and the red represents hot(+1). The heatmap showing that the store size and store staff are highly correlated with 0.94 can estimate that the amount of staff can work in-store will depend on the size of the store. Store marketing and total customers have the highest correlation of 0.99.

6. Bubble plot

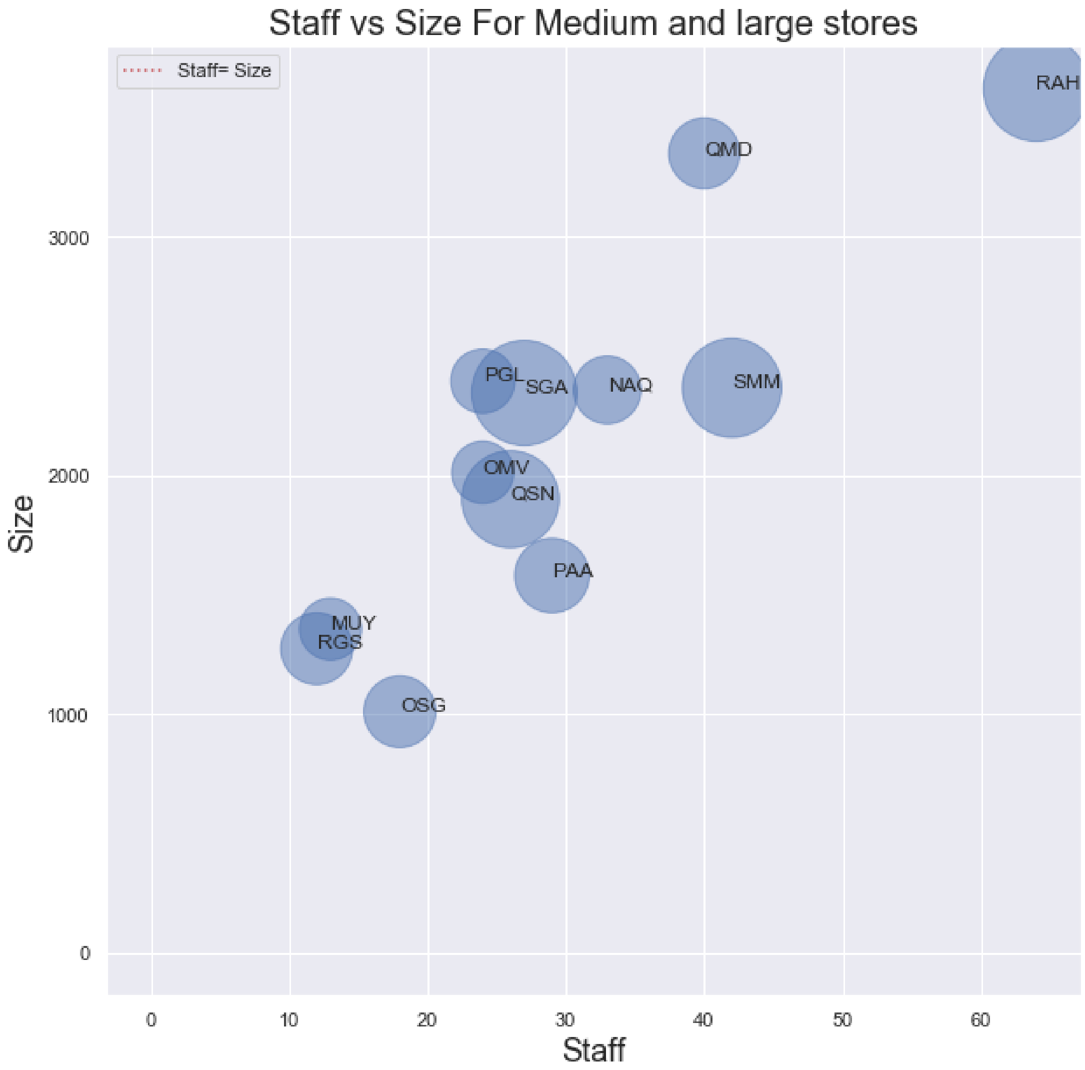


Figure 10 ***Bubble plot for medium and large stores comparing size vs staff vs total customers***

A bubble plot is an enhanced Scatter graph that enables us to visualize three variables instead of two likes in a standard scatter graph. Fig 9 identified that staff and size have a high correlation, so in this plot wanted to know more than if both depend on each other, and the result shows that both do.

7. Interactive bubble plot

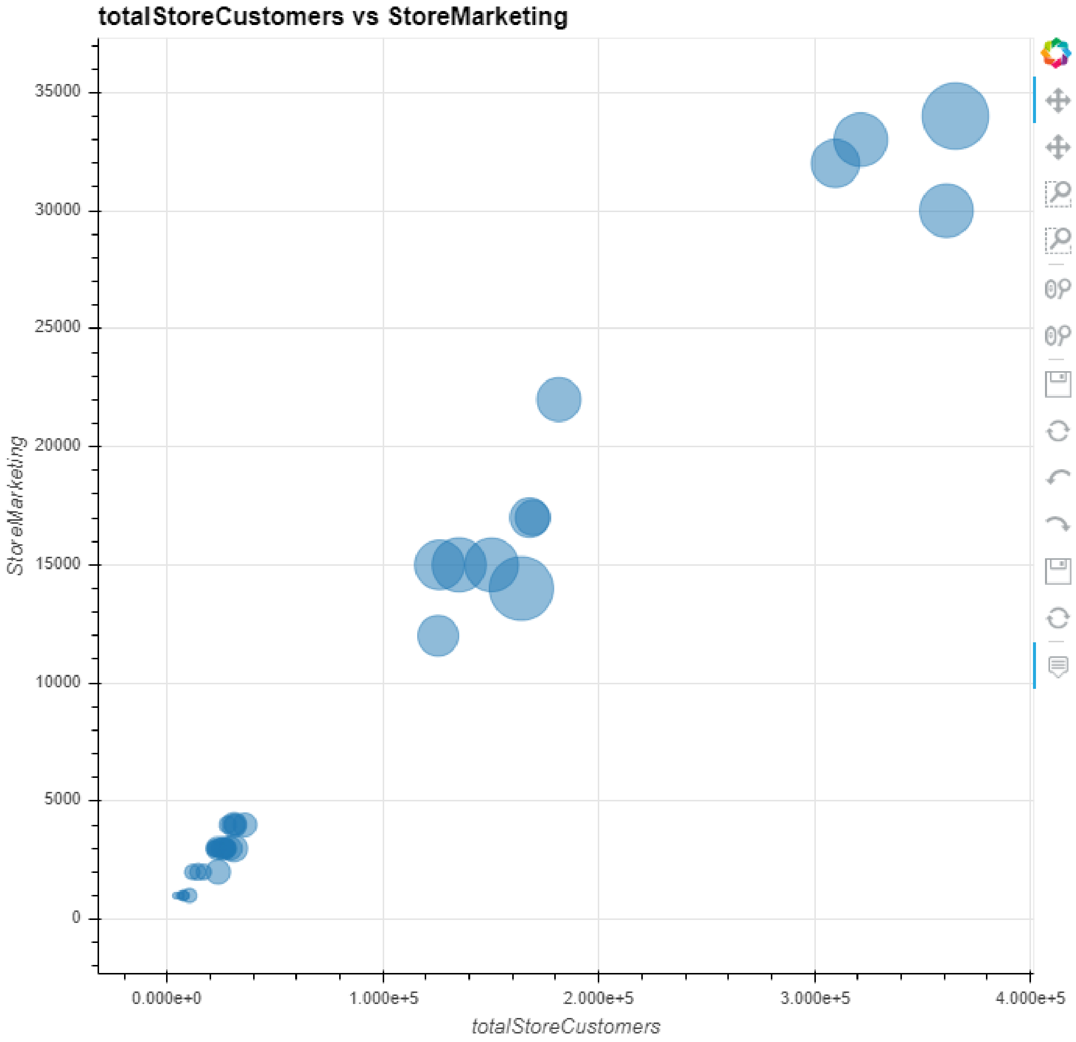


Figure 11 Interactive bubble plot comparing total store customer variable against the marketing variable

A bubble plot is an enhanced Scatter graph that enables us to visualize three variables instead of two likes in a standard scatter graph. Fig 11 showing a comparison between the store marketing and total store customers to find a correlation. The bubble plot shows a perfect correlation between the store marketing and the total customers, which what also demonstrated in the heatmap, a correlation of 0.99. The plot also can help to see that the highest marketing spent on stores it will increase store customers. The plot also shows stores with three different volumes small, medium, and high, which proves my point that stores spent less marketing had the least number of customers, and the highest amount on marketing received the most top customers.

8. Line interactive plot

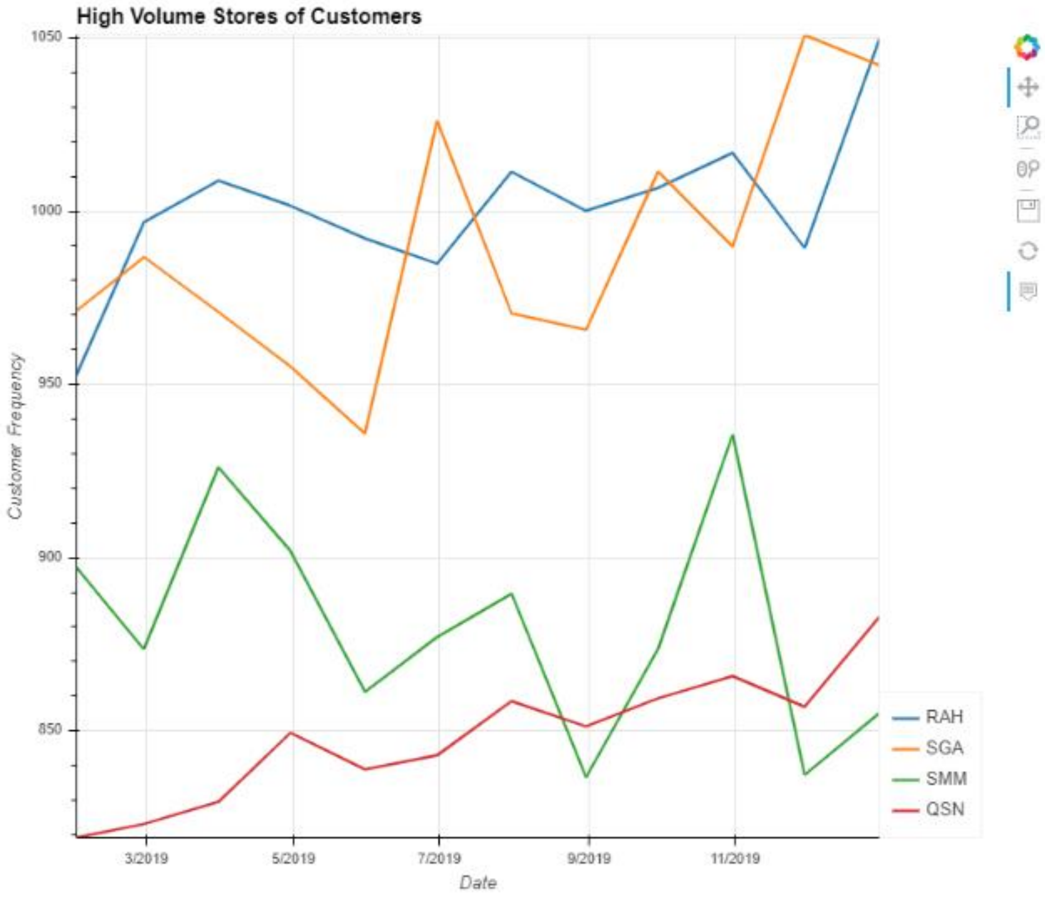


Figure 12 Interactive line plot displaying data for the highest stores

The line plot above showing how the data of daily customers changed every month for the highest volume stores in term of customers visited. Line plot is most suitable for graph for data that includes data as it enables to see the changes of data over time. The method used on the figure 12 was rolling average in order to remove the noise from data in order to get a better insight and it was the insight that for RAH,SGA and QSN showing that throughout the year customers are increasing but for SMM showing the customers decrease at the end of the year. Choosing the line plot for the interactive because in the interactive when point a peak or trough you can understand more information about how much customers attended that day and the exact day for a peak or trough.

A critical review of your work, with a discussion of how best practices were demonstrated and applied.

Starting the project with reading all the CSV files to understand the variables included in the data was the best practice. After reading all CSV using excel software, it helped me to realize what new information we can get from the data given by ChrisCo, but it was not 100% efficient yet.

My first reading on the data, including all files it would be useful to understand how the company works, such as how to decide to amount of money, need to spend for each store for the marketing? How many staff required for each store can be determined, and lastly, how to attract more customers into a store? After following each week's presentation and completing each week's lab on time, selecting a suitable visualization was more comfortable to determine for the Chrisco project.

The methodology was to look at each week's lab and practice the same methods with the project as we still at the data exploration stage, which practically means that do as many as you can of visualization to understand the data at the start. After doing the first lab of time series analysis, we used the file with date included, which was daily customers. We identified three types of customers volumes to stores, such as highest stores, medium stores, and small stores. After the identification, it became clear my understanding of the data at the start when reading all the CSV files. Still, it was time to prove the concept using visualization techniques in a gentle exploratory way using charts.

Another methodology used was storytelling, such as finding a similar answer using different types of visualization charts; for example, we identified a highly correlated correlation relation between each store variable by comparing it with each other using a scatter plot. A Scatter plot showed a great result, but it was not 100% clear if it was a strong correlation. Using the heatmap after gave us a clear answer of which variables are highly correlated, and it was store marketing vs. the total customers, then store staff vs. the size.

After determining the highly correlated variable then the decision to identify which one of these affect other such as whether store marketing effect the increase of customers or the amount of staff determined on the size of the store. The use of a Bubble plot helped to establish an answer to our question because it helps to visualize more than two variables. The use of interactive plot helped to design an exciting visualization that enables us to understand more data as it can give us entirely when clicking the mouse on the graph for example inline plot figure 12 when you click on any peak you can get a full detailed of how many customers visited with day visited on.

Conclusions

Chrisco company project was an exciting project to work on to help them understand their data by visualizing for them fascinating information found during data exploration using data science techniques. The result from the data that Chrisco can increase attract more customers into visiting their stores by doing more marketing about each store. Another exciting fact figured from the data that the decision on the amount of staff to have on each store is depended on the size of each store. Lastly, Store Overhead showed no interesting data or any relation to the rest of the data.

References

What is data visualization and why is it important?, (2020) *SearchBusinessAnalytics*, [online] Available at: https://searchbusinessanalytics.techtarget.com/definition/data-visualization (Accessed 18 March 2020).