Abdullah Bin Zubair

abdullahbinzubair570@gmail.com | az5645a@gre.ac.uk

DATA EXPLORATION REPORT

## Introduction

Data Visualization is the technique or process of communicating complex information from a dataset using simple charts and plots. There are numerous mathematical, mainly statistical, concepts involved in the process. In this age there are many platforms and languages in which you can implement data visualisations but the most common language is Python. Python gives support of many complex mathematical functions as there are numerous highly useful open source libraries available for it. Some of them are Numpy, Pandas, Seaborn, Hvplots etc.

The process begins with importing or creating the dataset. After some pre-processing such as indexing and column names finalizing, the real game begins. Most analysts/developers/scientists segment their data so that they can run different visualisations on different parts of their data. The visualisations can start from simple bar chart to line plot, correlation matrix, histogram, box plot, autocorrelation plot, scatter plot, bubble plot etc. Data Visualization has the power to tell data-driven stories while allowing people to see patterns and relationships found in data.

DATA: The data used is provided by Chris Walshaw. You can check out his Youtube channel by clicking this link: https://www.youtube.com/user/cwalshaw . He taught me Data Visualization course when I was doing my Masters, I think he probably has more industry experience than my age.

## Visualization

**Number 1**

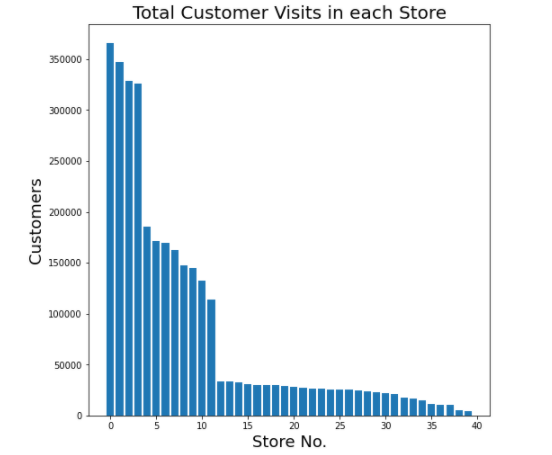


Figure 1 – Total Customers vs Store

**Justification:** Bar graphs can summarize huge dataset in a very simple manner. This is used often for segmentation of dataset and to get a generic overview of the data. This is not the same thing as Histogram as Histograms measure frequency of specific data occurrence (Histogram is coveerd later in this report).

**Description:** We can clearly see the 3 segments in our dataset.

1. There are 4 stores which exceed 200,000 customers and all the remaining stores have customer figures less than 200,000. These are the high volume stores which means that these stores have the most amount of customers throughout the year. In the following report, ‘High Volume Store Customers’ refer to those stores’ data which were visited the most. Same is the case for Medium and Low Volume Store Customers. High Volume Customer Stores: ['RAH','SGA','QSN','SMM']
2. Then there is second segment of Medium Volume Store Customers which starts from around 50,000 to 200,000. There are 8 stores in this range and this set has the medium amount of customer visits throughout the year. Medium Volume Customer Stores: ['PAA','RGS','QMD','OSG','NAQ','PGL','OMV','MUY']
3. Lastly we have the set of stores for below 50,000 customers throughout the year. This is our third segment. Low Volume Customer Stores: ['WMB','EFN','WYG','TSE','ENY','BTB','YGY','TAP','XML','UGJ','VSM','UMU','BZM','CNQ','CFG','DTJ','VYZ','WGR','ATT','DZT','NMO','XSV','AEI','ZMS','YYO','NGB','ZSD','MAJ']

**Note**: X-axis in Figure 1 are numbered stores instead of store initials as given in database. The stores are numbered in this way because the initials were making bar chart very untidy and it never looked good to the eye. We can see from this that there are 40 different stores in this dataset.

**Number 2**

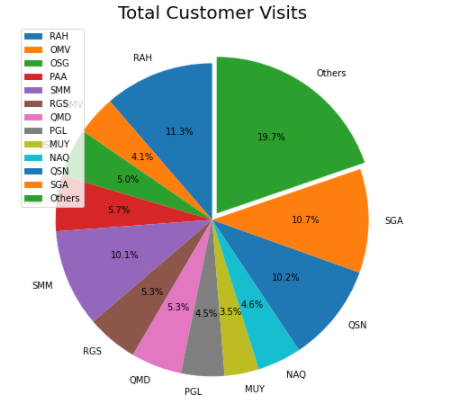


Figure 2 – Pie Chart of customer share among stores

**Justification:** Pie charts are a very effective way of visualizing each entity’s proportion in the dataset. In our case we have different stores as entities so we will find the stores with most or least proportion of customers.

**Description:** We can clearly see that more than 40% (10.7 + 10.3 + 11.3 + 10.2) of customer visits are on just 4 stores and the rest of customers are distributed among the other 36 stores. All the low volume stores have been grouped into one category ‘others’ for this visualisation as the individual visits for these stores were far less than the high or medium volume customer stores.

**Number 3**

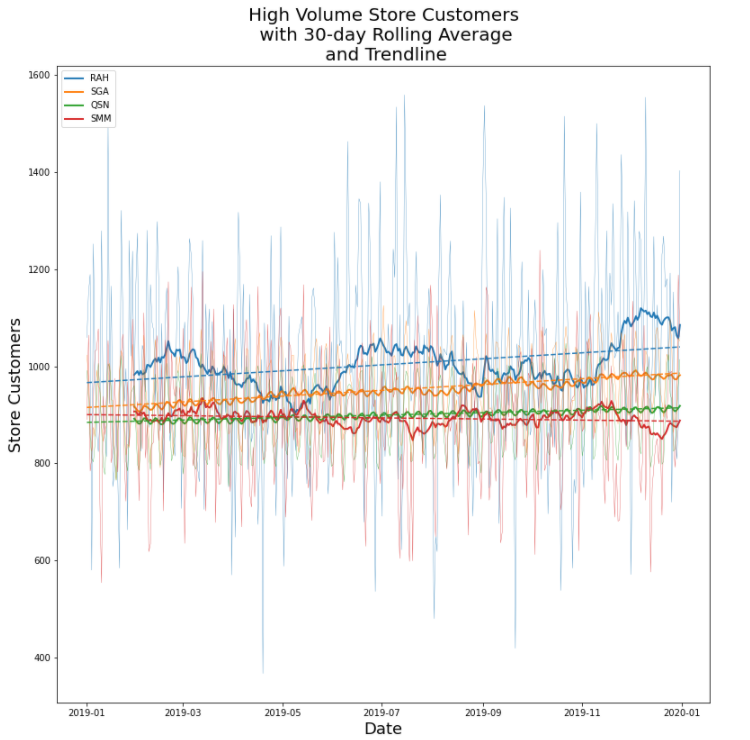


Figure 3 - Rolling average and trend line for customer data

**Justification:** When 50% of data is about 4 stores, it serves as an inspiration to look deeper inside those stores. That is why I have used line plots to their maximum potential for further investigate the 4 high volume stores. There are 3 different line plots for each store in this visualisation:

1. Raw customer data – These are the faded lines which are least prominent and are spreading/fluctuating the most on the chart. This line represent the actual number of customers.
2. Rolling average line – In this case I have plotted 30 day rolling average which remarkably smooths the raw data and gives a clear indication of customer visit changes throughout the year. The key on top left of figure 2 is for 30-day-rolling-average-line plots.
3. Trend line – This is a simple indication of upward or downward trend for each store’s customer visits. In figure 2, the dotted lines are trend lines.

**Description:** We can see that customer visits to stores ‘RAH’ and ‘SMM’ is quite fluctuating and there is no particular pattern which they follow. However this is not the case with ‘SGA’ and ‘QSN’. Both of these store’s rolling average lines are pretty neat and follow a consistent pattern throughout the year. This can be vital sign for the seasonality in these stores as we can see both of these continue to have similar kind of ups and downs throughout the time series. We can also see that SGA and QSN rolling average lines are very similar to each other which births the idea of correlation in developer/analyst/scientist’s mind. We will further investigate the seasonality and correlation in below visualisations.

**Number 4**

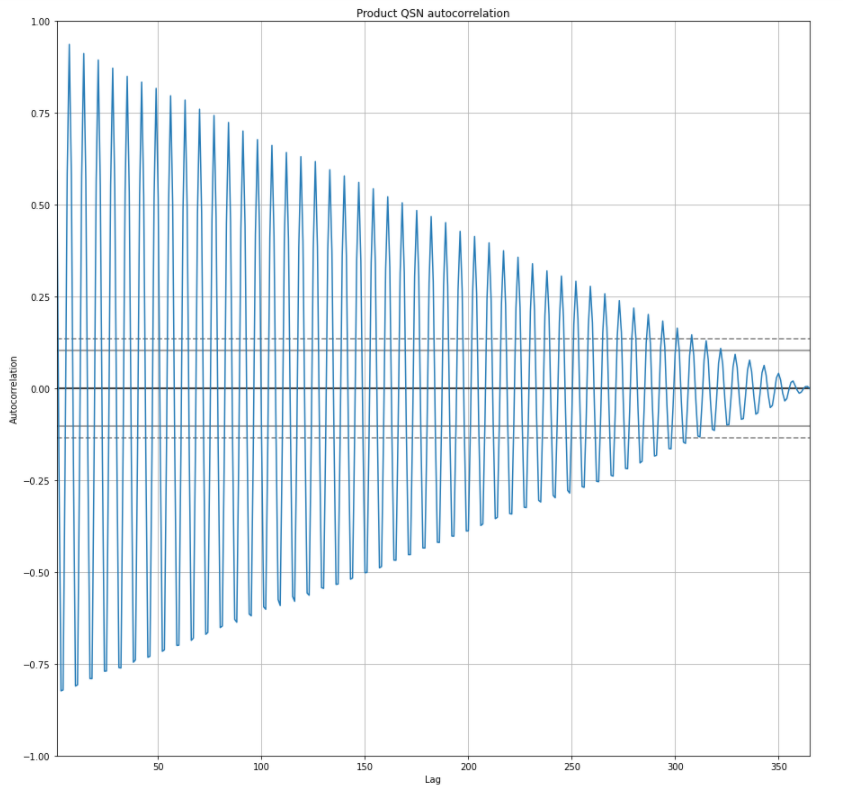


Figure 4 - Autocorrelation plot for QSN

**Justification:** From figure 3 we got a clue that there can be some sort of seasonality in SGA and QSN. So that is why we are further plotting an autocorrelation plot for QSN only (had to choose one) which can help us determine the seasonality in customer visits for this store. The time reading from start of second bump lets the user know about the season time frame.

**Description:** We can clearly see the line plots making consecutive bumps after every 7 days. This might be hard to notice from figure 4 but as I saw it in my Colab result window, it is very clear that there is a seasonality pattern for 7 days for this store as the second crest forms on 7th day and then repeats itself after each 7 day interval.

**Number 5**

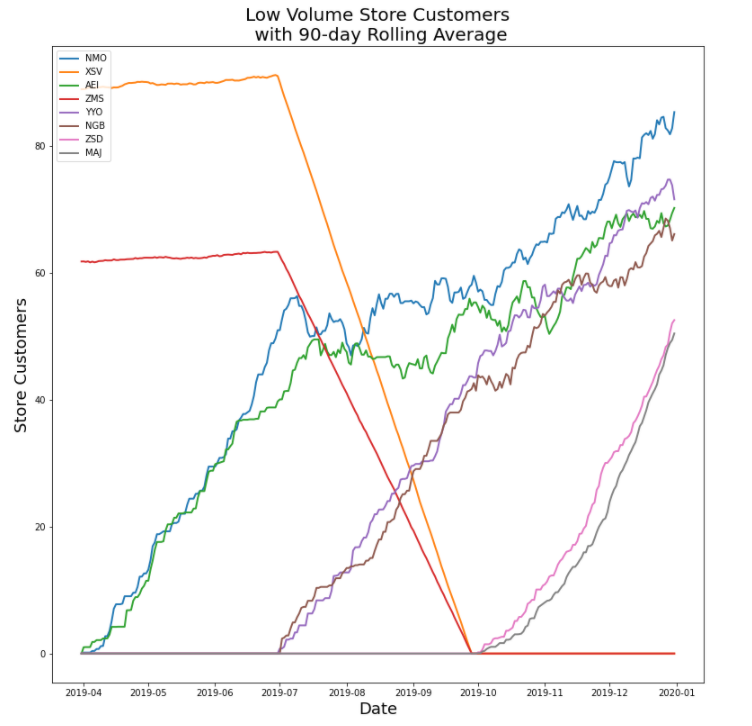


Figure 5 - 90-day-rolling-average lines for a specific set of stores

**Justification:** After visualizing few plots on my own, I realized that there was some vital information to be discovered in the data of low volume customer stores. For this segment I have chosen 90 day rolling average lines as they make relatively smooth lines. I noticed that some of the store lines were coming to a complete 0 and some were starting from 0. This made me further cut down my stores for analysis in this visualisation and finally I found the set of low volume customer stores which were making a vital difference.

**Description:** Beginning from 0,we can see that the stores AEI and NMO have consistently increased the number of customers visiting the store from start of the year 2019. We can also see the opening of 2 stores namely NGB and YYO from the month of July, 2019. We can also see that two stores – XSV and ZMS have completely halted their operations and reduced the customer visits to 0 till late September, 2019. This might be the case where these 2 stores have been closed or shut down due to some major reason. But as seen from the figure 5, two new stores - MAJ and ZSD have begun their operations from start of October, 2019 and have rapidly increased the customer visits till end of year.

**Number 6**

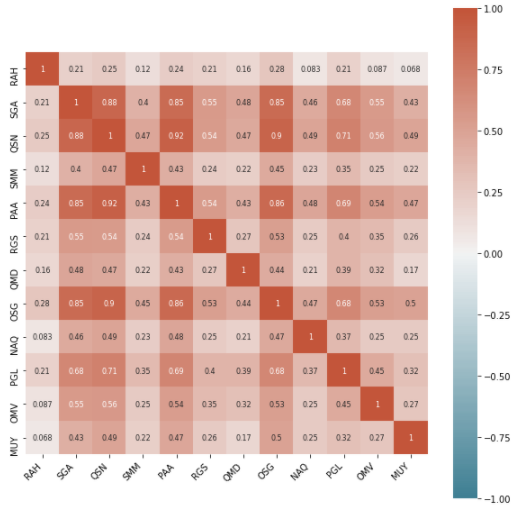


Figure 6 - Correlation between high and medium volume customer stores

**Justification:** We could see from figure 3 that SGA and QSN had very similar looking rolling average line. This indicated that we should create a correlation matrix and visualize/note the correlation coefficients in order to know how each store customer visits are related to other.

**Description:** From figure 6 we can see that none of the stores are negatively correlated to any other store. For this implementation, we are only interested to know the stores which have coefficient greater than or equal to 0.85. Correlated store pairs: SGA and OSG (0.85), QSN and OSG (0.90), SGA and PAA (0.85), QSN and PAA (0.92), SGA and QSN (0.88).

**Note**: These highly correlated stores need further investigation but as I have not used the summary\_data and visualized the distributions, I will be halting this EDA line and will move towards the summary\_data.

**Number 7**

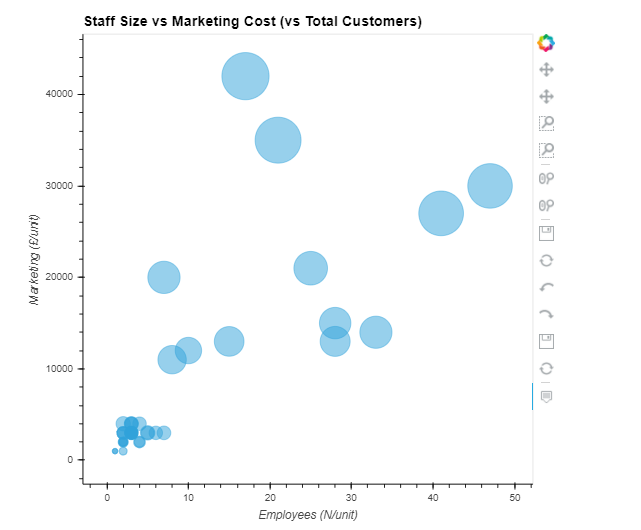


Figure 7 - Employees vs Marketing Cost (vs Customers)

**Justification:** Bubble plot is a very handy visualisation especially when user wants to analyze 3 variables effectively.In bubble plot, the axis represent 2 distinct dimensions and the third one is represented by the bubble size of data point. In this particular example we are taking staff size on x-axis, marketing cost on y-axis and number of total customers as the bubble.

**Description:** We can see that as number of staff increases in the store and as marketing cost increases, the size of bubble also enlarges. So far I did not find any anomalies with this general trend, RAH and SGA have far less employees but as the marketing cost have been so high, they are performing in terms of customer visits. There is a proper balance in the dataset which makes sense, as the company increased the marketing cost the customer visits increased. And as the customer visits increased, the number of employees in that store also increased (in case of SMM and QSN). All these factors imply on each other and there can be many possible reasons for each change.

**Number 8**

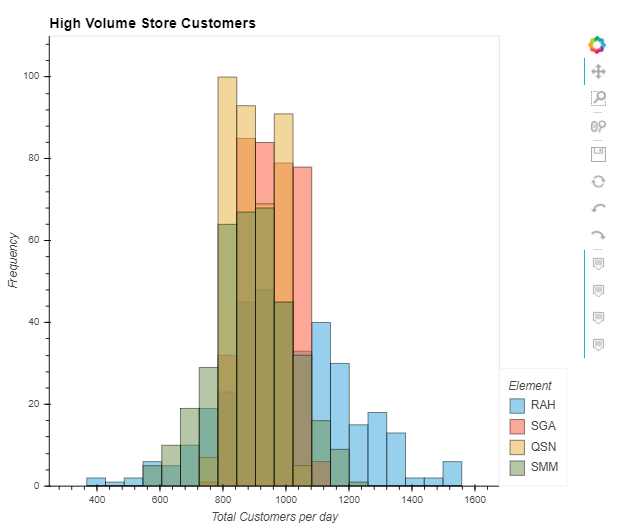


Figure 8 - Frequency distribution of customer classes

**Justification:** This Histogram represents the frequency of customer classes throughout the year. This means that each bin (class) represents 60 values (in figure 8) and if the visits each day ends among those values, the frequency is increased by 1 on that day for that particular bin. This is very useful when user needs to visualize the distribution of its data throughout the time series.

**Description:** In figure 8 I have included the High Volume Customer Stores so that they can be easily visualized. This is an interactive visualization (you can interact with it if you run on Google Colab the submitted notebook) and user can easily hide some of the stores data and analyze the others and do much more with it using hvplot’s HTML interface. Among these 4 we can see that RAH’s frequency distribution is most spread, which means there were good and bad days both with majority being normal. Whereas SGA and QSN’s distribution mostly comprise of same class, it is much narrow, it’s not flattened like RAH’s distribution has. This means that SGA’s and QSN’s have been serving a constant number class of customers all year.

## Critical Review

Throughout the implementation of all the visualisations, it was made sure that the approach is efficient and effectively passes on the valuable insight/information. The code written is very clean and easy to understand. There is no excessive or over use of any code blocks which were not meant to be over used. The implementation follow a very descriptive approach as at every point I have introduced a new idea, Implemented it, visualized it, taken insight from it and further investigated the output/findings of previous visualisation. In the notebook I have not discussed the findings and the thought process for the whole data exploration. In the report everything is very detailed. I have documented the whole process in simple and standard terminology and use of language is very direct. Personally I believe that there is so much more that can be explored from this dataset and the above 8 visualizations can serve the basis for the research. I am aware that this data set is not organic, its synthetic, and that is why there are some very obvious correlations, patterns and trends forming. Keeping everything in mind, the approach followed has been completely from the point of exploring some company’s actual enterprise data.

## Summary of conclusions

* Dataset is clearly divided into 3 segments – High, Medium and Low volume customer stores.
* There are 40 different stores’ data and some of them have closed their operations in 2019 while others have opened their store.
* Many of High and Medium volume customer stores are highly correlated (positive).
* More than 40% of customers are served by top 4 stores.
* There is strong evidence of seasonality in many stores. I have mentioned the ones in 2 stores from top 4 ones in my report above. I believe that there are more seasonality patterns in this data set besides these 2.
* Almost all the stores have increased the customers visits except for few which halted their operations.
* Few stores have begun their operations during the year which is why some stores’ customer visits have rose from 0 during the year.
* As the marketing cost increases, the store also attracts more customers (might be other way around).
* As the customer visit increases, the store also increases their staff size (might be other way around).
* Among top 4 stores, the frequency distribution of customers is most spread for RAH and it is most narrow for QSN.