Data exploration for ChrisCo

This report represents the visual investigation on CrishCo Company's data, in which there are datasets of daily customer visits, Marketing cost, Store staff, Store size and Store overheads data pertaining to 40 stores.

Discussion of findings from data

Here I have concluded the 8 visualisations from the ChrisCo's Datasets. That are visually represented as follows.

1. Bar chart to show numbers of Customers visiting the particular store in year 2019

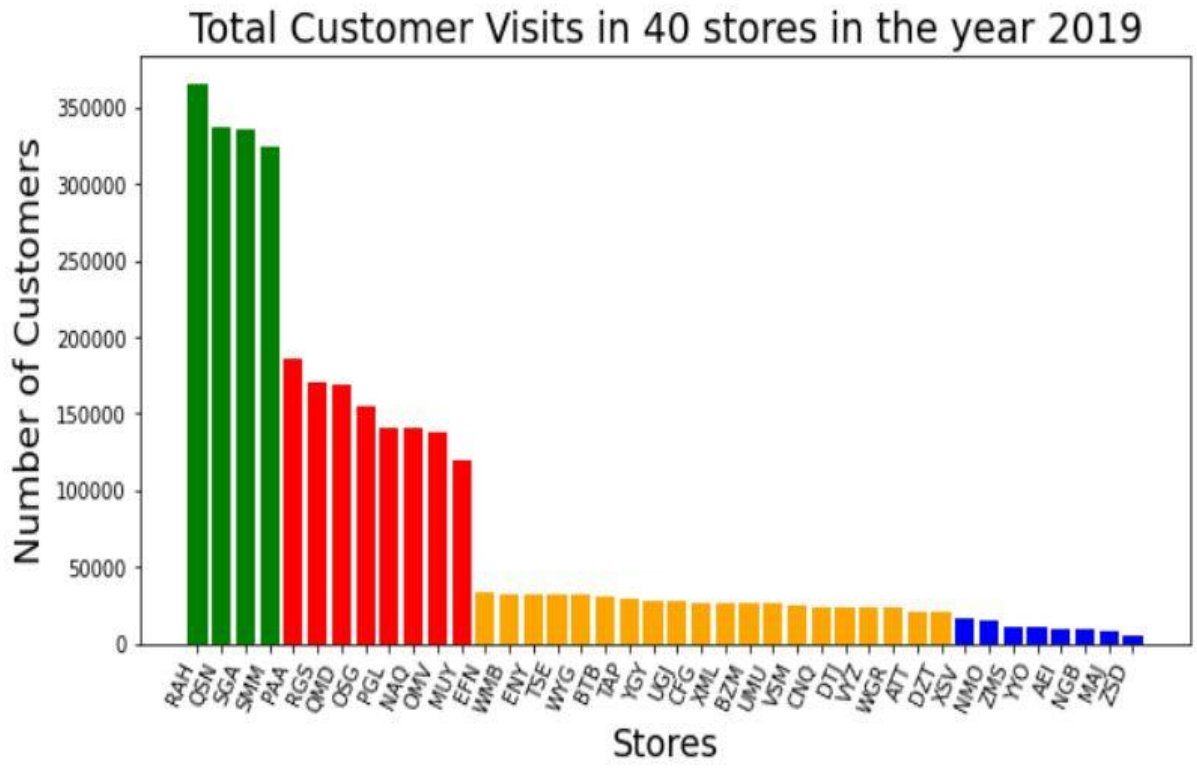

Figure 1 Bar Chart for Total customer visits in 40 stores in the year 2019

The purpose of including this Bar chart is to give overview of the ChrisCo's Stores Data. Using this bar-graph we can easily see the total number of customers in each column to analyse the number of customers that are visiting the particular store. Being a large dataset, its strongly needed to be segmented to analyse in a better way. So here, I have segmented the whole data in the same chart by differentiating it with different colours as different categories.

The Figure 1. Bar chart shows the Total customers visiting the store in the year 2019.It can be seen from the Bar chart that there are 40 stores on the X-axis and Y-axis represents total numbers

of customer visits in the year 2019. On x-axis stores have been segmented in the four categories, each indicated by different colours. The green colour shows the stores having the high volume of customer visits i.e., stores such as RAH, QSN, SGA and SMM (more than 300000 customers). The orange colour represents the other eight stores having medium volume customer visits i.e., stores such as PAA, RGS, QMD, OSG, PGL, NAQ, OMV and MUY (between 100000-200000 customers). The yellow colour shows the stores having low volume customer visits (between 20000-50000 customers). The blue colour shows the very low customer visits per year (less than 20000 customers).

2. Interactive Line Plot showing all 40 store's customer data

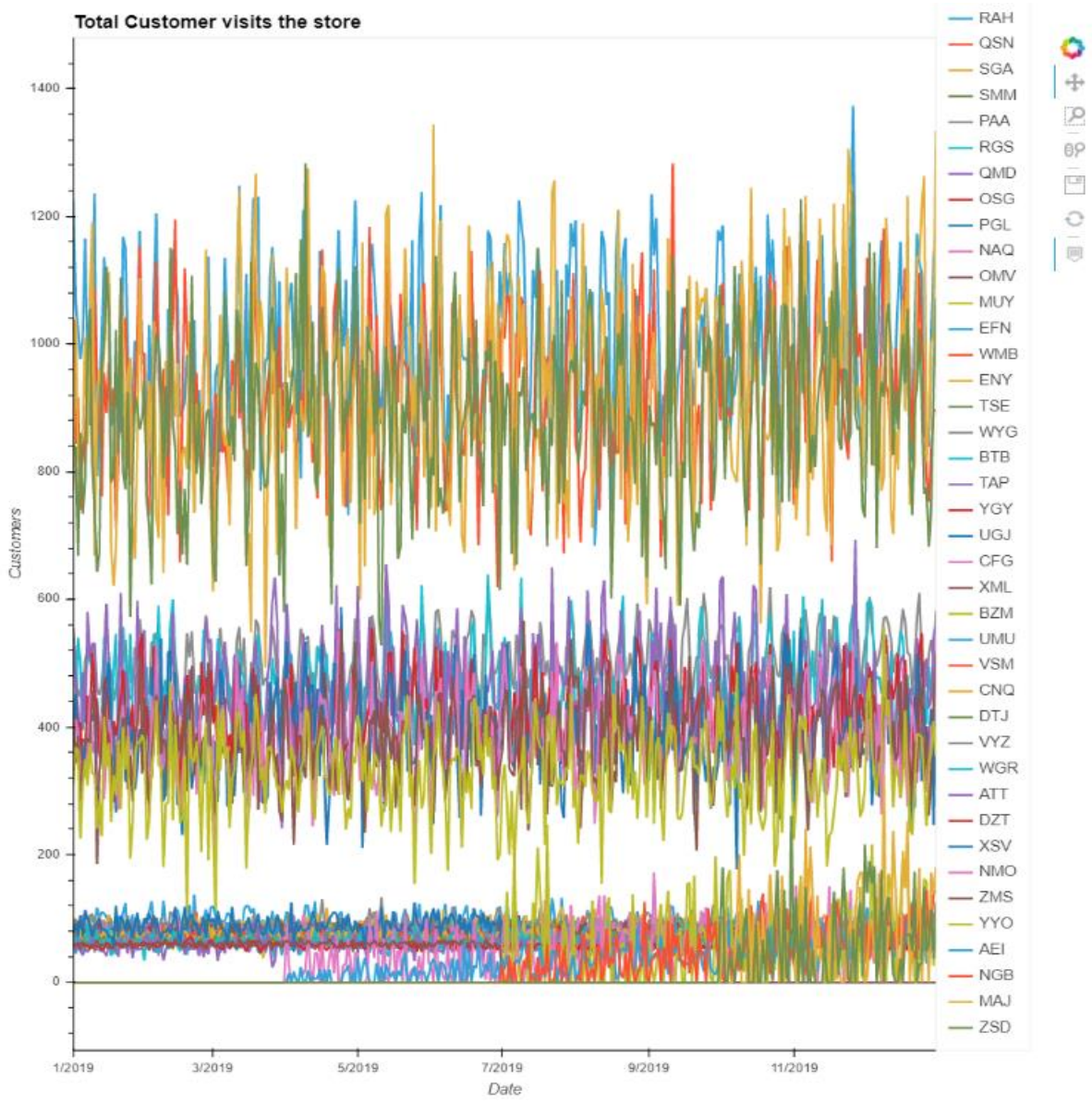

Figure 2 The Line Plot showing all store's customer data

The purpose of including this Interactive line plot is to show the data in minute of details. This chart helps to show daily data of the customer visits the stores i.e., 365 days data of the year 2019. So, with help of the interactive Line Plot, we can see the data of the daily customers visiting each store. Using Interactive Visualisation, we can analyse data in deep with help of given functionality such as pan, Zoom, wheel-zoom, save, refresh and hover button. so, we can have point-to-point detail about each datapoint by hovering on the line. We can zoom-in data to see minute of details. We can refresh the changes we have made into and also, we can save the graph.

The Figure 2. Line plot shows the data of the daily customer that visits each store in the year 2019. Where Y-axis shows the total number of Customers while X-axis shows the Date and the legend (right side of the line plot) shows the 40 stores name. By hovering on line, we can see the exact number of the customers of each store. It can be notice from the figure 2. the high and medium volume data seems normal so no missing data is there. All store's daily customer data is showing fluctuated pattern and there is too much noise in the data but we can see the huge difference of the data of the stores having too low customer visits. From figure, it can be notice that some store's customer data is missing for some months in the beginning of the year. So, to see that missing data I have made the separate Line plot of very low Volume customers data which is presented in the following Line plot.

3. Interactive visualisation showing summary data for stores having Very low volume customers

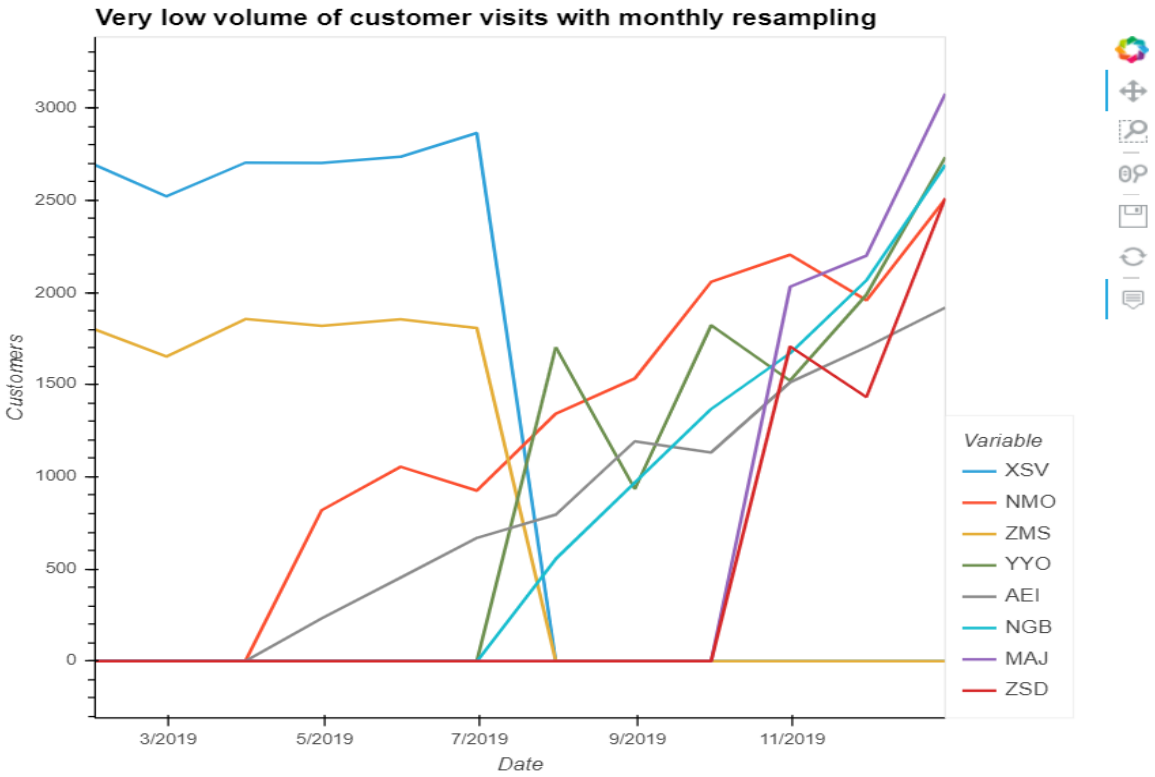

Figure 3 Interactive Line Plot shows summary of the stores having very low volume customer visits

Stores having very low volume customer visits data:

Date	XSV	NMO	ZMS	YYO	AEI	NGB	MAJ	ZSD
2019-01-31	2690	0	1799	0	0	0	0	0
2019-02-28	2521	0	1652	0	0	0	0	0
2019-03-31	2703	0	1855	0	0	0	0	0
2019-04-30	2702	819	1818	0	232	0	0	0
2019-05-31	2735	1054	1854	0	453	0	0	0
2019-06-30	2864	925	1806	0	669	0	0	0
2019-07-31	0	1342	0	1702	795	555	0	0
2019-08-31	0	1533	0	934	1192	968	0	0
2019-09-30	0	2057	0	1821	1132	1367	0	0
2019-10-31	0	2204	0	1521	1512	1670	2030	1707
2019-11-30	0	1957	0	1987	1703	2065	2199	1432
2019-12-31	0	2506	0	2733	1917	2690	3077	2511

I have used the Interactive visualisation line plot with monthly resampling of data of year 2019. Using Interactive Visualisation, we can analyse data in deep with help of given functionality such pan, Zoom, wheel-zoom, save, refresh and hover button. so, we can have point-to-point detail about each datapoint by hovering on the line and zooming option make easy to identify minute of details.

From the above Interactive visualisation line plot and table data, it can be notice that the stores having very low volume customer (less than 20000 customers) have a dramatic change. It can be noticed from the statistics as well as line plot that there are eight stores in which some stores have a zero data available for some months. Such as Store XSV and ZMS have zero data after six-months, so we can consider it as a store closed after the 6th-month. while, we can see that the Store NMO, YYO, AEI, NGB, MAJ and ZSD have zero data for some starting months of the year, so we can consider it as store open after the number of customers is available. However, in this very low volume customer data I have found the unexpected behaviour of the data i.e., anomalies, in the stores YYO, NGB, MAJ, ZSD that will be explain and proved by the upcoming Box-plot.

4. Correlation of summary data on Heatmap

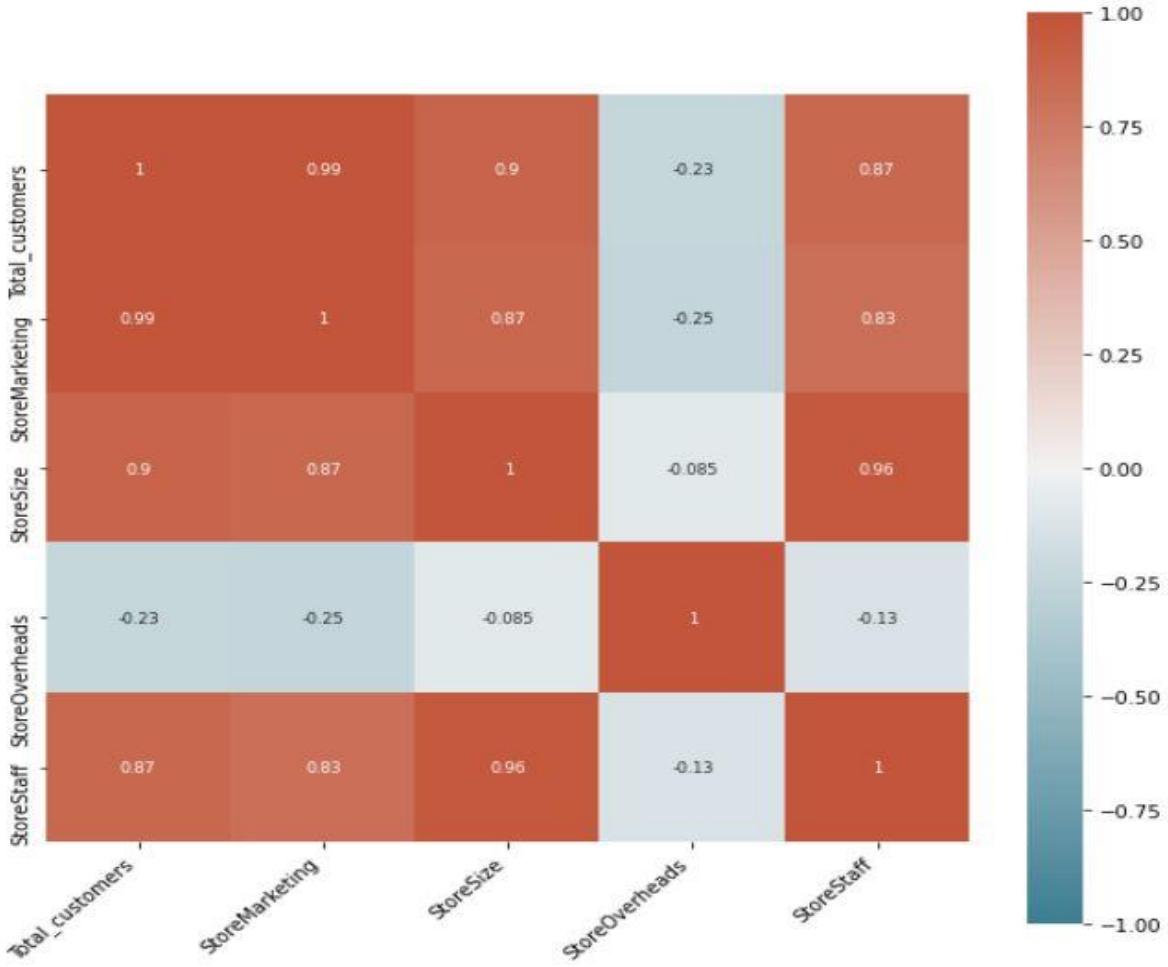

Figure 4 Correlation between Summary Data Variables related to each store

The above graph represents the heatmap to find out if there is any correlation that exists between Summary Data. Where there is one scale (right-side of the graph) has been given to identify what type of correlation is there between each variable. Dark brown and blue colour represents strong positive and strong negative correlation respectively. Lighter shades describe low correlations.

The figure 4. shows the correlation between summary data variables of each 40 stores. However, I have also tried to find out the correlation between each high and medium volume store's customer but there is no such strong or positive correlation between them. However, in the figure 4., we can see that the four variables i.e., total customers of each stores, Marketing cost of particular store, Store size and Store Staff are strongly positively correlated with each other except the Store overhead. Where, the store overhead represents the no correlation with other four variables. The total customers and marketing have a very strong positive correlation (0.99) i.e., if more marketing than more customer visits. Also, the store size, store staff and customers have a strong positive correlation too. It indicates if store size is big than staff members are also more and if stores are big than number of customers are also high.

5. Box-Whisker Plot for some stores having mixed (high, medium & very low) volume customers data

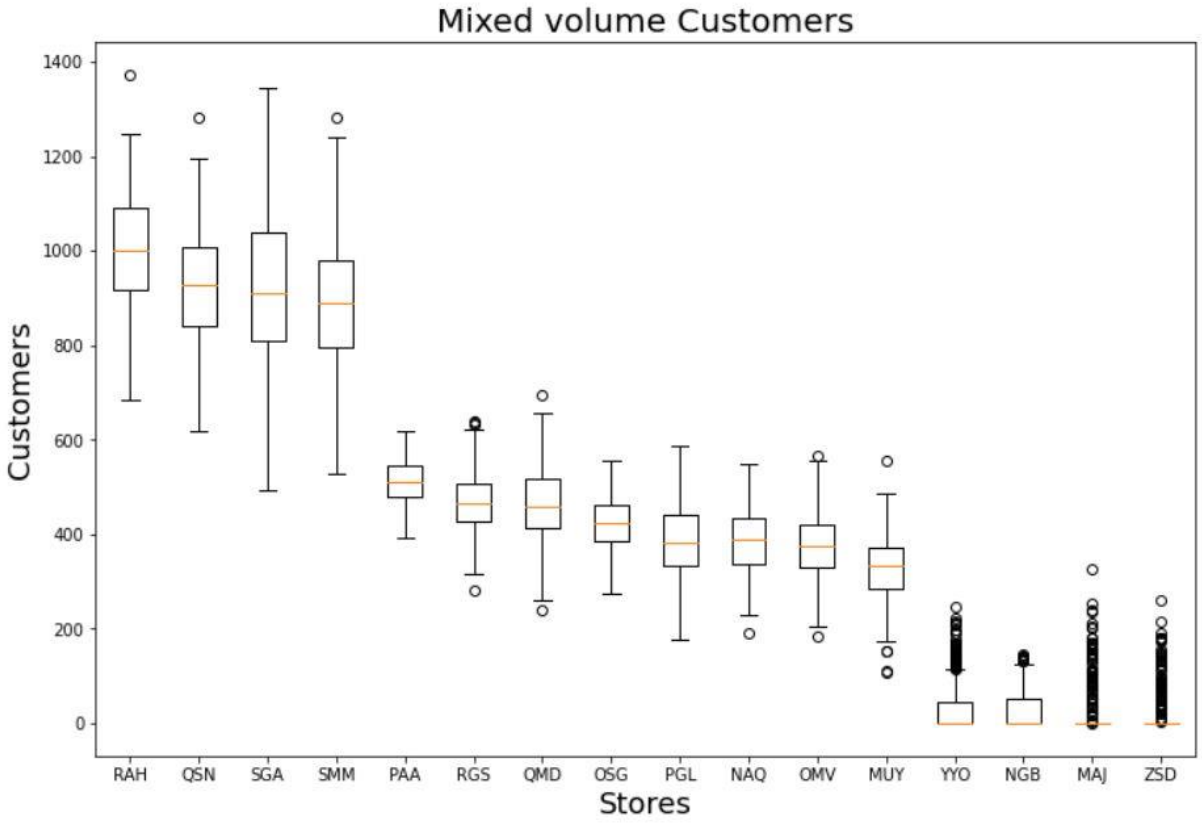

Figure 5 Box-whisker plot showing Mixed volume customers

The purpose of including the above Box-whisker plot is to detect the outliers/ anomalies presents in the data. It is a very simple but effective way to find outliers/anomalies among the available data.

The above figure 5., represents the mixed volume of data plotted in the Box-whisker plot. The box represents the 50% of the data, orange-line shows the median and upper & lower vertical line of box (whisker) shows the maximum 1.5time data, however the out-side of the whisker limit, there is a bubble that describes the outliers. Outlier here is an extremely different data from the most of the data points. The first four stores showed as a high volume of customers i.e., stores RAH, QSN, SGA, SMM, in which we can notice that these stores have a high volume of customers visit and also have more outliers except one store SGA, compare to the other medium volume customers. The following eight stores PAA, RGS, QMD, OSG, PGL, NAQ, OMV and MUY describes the medium volume customer data, in which the store PAA and OSG have the perfect data however the other medium volume stores RGS, QMD, NAQ, OMV and MUY have some outlier in the data. Lastly, the rest four stores YYO, NGB, MAJ and ZSD are some stores from very low volume data. I have included only four stores from very low data because I have found the anomalies in these stores only when I have tried the all-low volume data in a Box-plot to detect any anomalies that lies in the data or not, that I have mentioned in the discussion of

figure 3. These four stores having low volume customer have a great number of anomalies founded in the whole data and these stores are from the list of stores open in the middle of the year.

6. Line subplot- Decomposition for high volume customers to detect seasonality

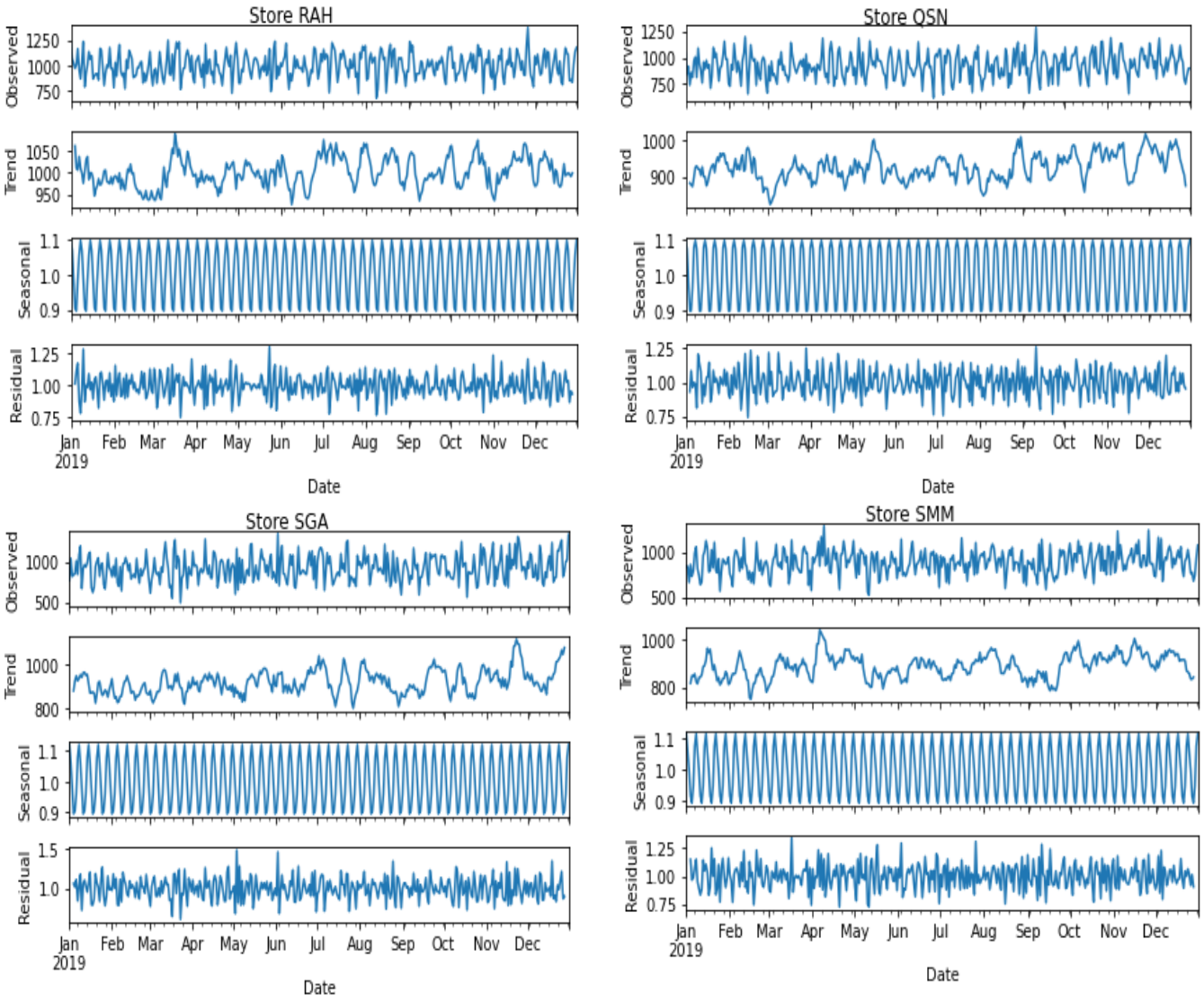

Figure 6 Showing the seasonality in stores having high volume of customer visits

The purpose to include this above chart Line subplot-decomposition is to find out any seasonality, patterns and noise from the data. This is the one way to find out the seasonality, trends and residual from the data.

The above figure 6., shows four high volume of customer visits stores. In which the time series have been decomposed into four parts observed, seasonality, trends and noise. Observed show the average values of the store's customer data. It can be noticed from the figure 6., that all stores have a same weekly seasonality. However, the trend(pattern) of the data is fluctuated daily. So, from the above figure we can conclude that the all stores have a one peak day i.e., 7^{th} day, when the customers are highly increasing on that day only.

7. Radar subplots for some store having mixed (high, medium and very low) volume of customers

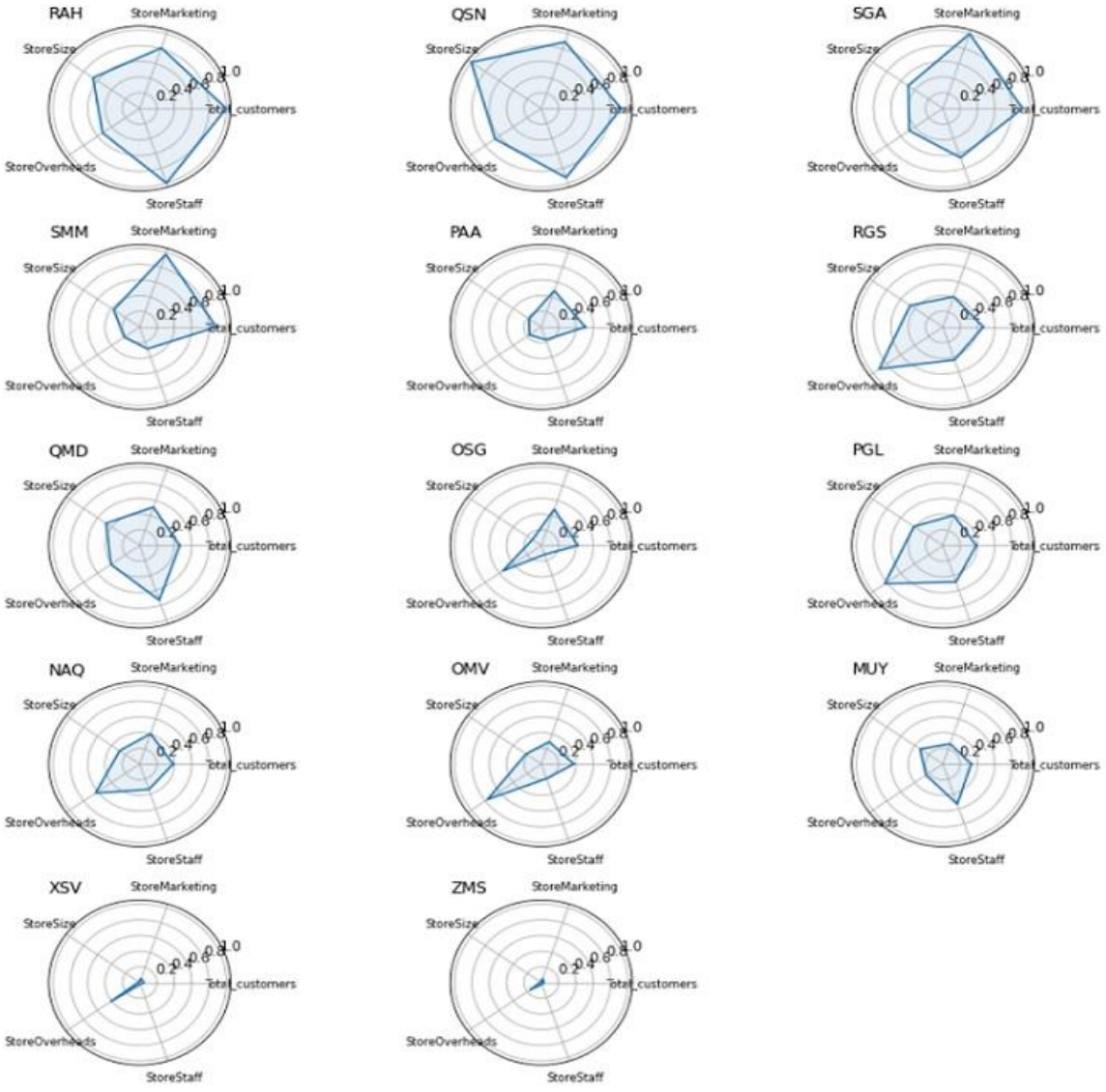

Figure 7 Radar Plot

The above plot figure 7., is to define the multivariate data of particular store. The reason behind to include this Radar plot is to compare each variable/entities of the stores on 2-D plane and based on that data-driven decision can be made easily.

Figure 7 shows radar subplots of different stores. Each radar-plot represents five entities in figure 7. Each spoke in the plot represent one of the entities. And the inner circles are created with respect to range of each normalized data point of the entities where each data points have been substituted by Max value from the data column that knowns as normalized data. Here, we can see that the stores have five entities i.e., total customers of the year, marketing cost, store size, store staff and store overheads. Each normalized data points together make a shape. It can be noticed from the figure that each Stores have great comparisons between each entity. We can notice the marketing cost for that year, the size of the stores, the staff members and the overheads for that store for the year 2019 in one single plot. High volume stores RAH, QSN, SGA and SMM looks normal however we can see some stores have a high overhead even though the store size is small and customer visits is low such stores are RGS, OSG, PGL, NAQ, OMV, XSV and ZMS. However, the store QMD is a small sized store, having low customer visits and low marketing cost but staff size is big. So, these are also known as an outlier.

8. Bubble plot for Store Marketing Vs Store Size (Vs Total customers)

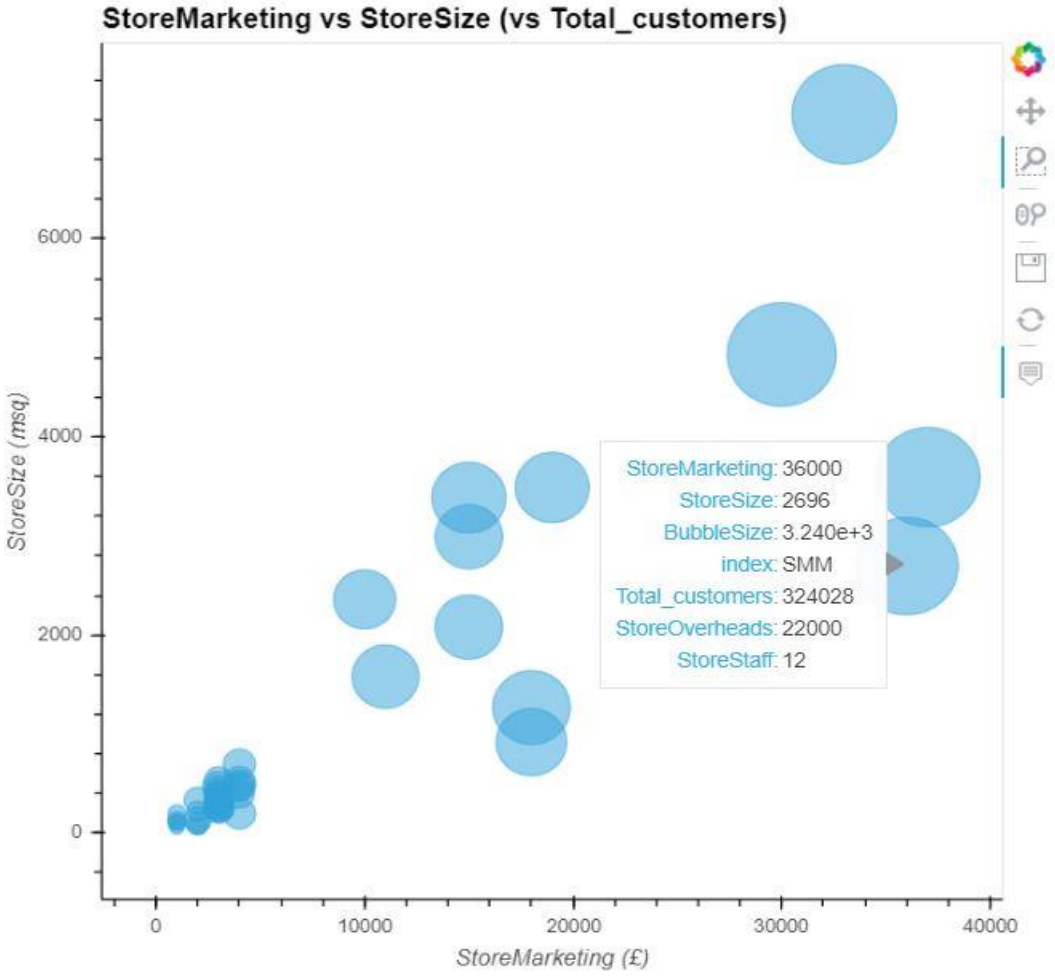

Figure 8 Bubble plot for Store marketing Vs Store size Vs Customer visits data

It is a three-dimensional plot. The purpose of the above figure 8 Bubble plot is to represents the three entities in a single plot.

The above figure 8 shows the Bubble plot associated with total customers, marketing and store-size. The x-axis represents the Store marketing cost and y-axis shows the store size. The size of the bubble is determined on basis of the total customer values. If there is less customer of the particular store, the bubble will be smaller. And bubble will be created on the interaction point of the marketing and store size data point. For example, It can be noticed from the above figure that the Store SMM have 324028 customers, and its marketing cost is 36000 of the year and Store size is 2696. So, we can see the size of the bubble of the customers is big.

Conclusion

From the above data exploration, following points can be concluded from the analysis of ChrisCo's datasets.

- ChrisCo company has the large data sets of its 40 stores, so to analyse each store's data, the data has been segmented in the four categories high, medium, low and very low.
- It can be concluded from the exploration of the data, that the stores having high and medium volume customer visits have a fluctuated pattern of the customers visits, and also the stores have a weekly seasonality detected in the data. However, because of fluctuation in number of customer visits there is so much noise has been found in the all 40 stores data. It can also be seen from the analysis that some of the high, medium and low volume customers stores have anomalies/outliers found in data such as RAH, QSN, SGA, SMM, RGS, QMD, NAQ, OMV, MUY, YYO, NGB, MAJ and ZSD.
- Also, there is no correlation have been found between each 40 store's customer data but there is strong correlation found between customer visits, store size and marketing cost and staff. i.e., bigger the store, higher the customers and higher the marketing cost for the both high and medium volume customer visits stores. Also, the bigger the store size more the staff. However, some stores have high overhead even though store size is small and less customer visits and also low marketing cost.
- On the other hand, it can be noticed from the very low volume chart that, some stores have a zero data available in some starting or after some months data. So, we can consider it as store closed (XSV and ZMS stores) or open (NMO, YYO, AEI, NGB, MAJ and ZSD stores) during the year.