**COMP 1800**

**Data Visualization: Coursework**

Khassan Nazhem

001151725

Dr. Christopher Walshaw

March 26, 2021

In the age of Big Data, the demand for the visualization is increasing drastically. Data visualization is a tool that helps make sense of the amounts of data through graphical representation. This can be applied in many careers such as education, finance, biology to name the least. Nowadays, it has become one of the most useful professional skills to learn. It allows scientists and users to arrange the data into a form of coherent and handy information. This information makes it easier to interpret the study’s trends, highlights, and outliers through applying different representation types like bar graphs, line plots, and heat maps. Each of these types can be helpful to gain various insights into the given data.

This report will focus on a dataset that includes the company's 40 stores, each identified by a unique 3 letter code. The main targets are large and medium sized stores, shedding light on their trends and highlights. Nevertheless, the report will also identify new stores that have been opened or closed during the year. The most common trend discovered in visualization is how the data shifts through time. This trend generally gives an insight into whether the feature under study is changing positively or negatively. Moreover, it determines how relevant the change is with respect to the frequency of customers visitors throughout the year. It is difficult to determine the correlation between elements without visualizations because such correlations are critical to understand and perceive the cause of existing trends. Hence, data visualization provides precise analysis to identify correlations between the given features. For example, brands tend to discover which audience to target with their current released products. So, they analyze the relationship between the locations and the number of sales to identify the target audience. Finally, data visualization has an impact on determining metrics such as value and risk computation. Such calculations require many different variables; the graphs presented are color-coded to show whether the elements are valuable or risky. With that being said, the given data and the visuals need to work in coherence to create the art of combining analysis with presentations.

To begin with the investigation for the company’s data, **Figure 1** presents a broad understanding of how each of the five features vary between the 40 stores. In this case, a bar graph plotting is essential to compare the variations under different circumstances. The features include the number of customer visits, marketing expenses, store size, staff number, and overhead costs. So, through this bar graph, we can distinguish between trends and outliers faster than a table of numerical data. The y-axis holds the values of the features, whereas each bar presents a different store.

**Figure 1** suggests that there are several patterns to be discovered. The marketing expenses subplot shows that there at least four stores that are putting a high budget on marketing their products, such observations can create further speculations. To follow up with the number of customers that are visiting the given stores, it is seen that there are also at least four stores that have a very high number of customers. Is that related to the marketing expenses? Which stores are they? There also seems to be eight other stores having a medium range of marketing expenses and customer visits. Furthermore, the bar chart proposes a positive correlation between the store size and the number of employees working at that specific store. These stores need further investigation in order to ascertain whether there exists a positive correlation between these variables or not.

This process is proceeded by separating the stores into three different criteria based on the number of customers visiting, leading to three-segmented sections: high, medium, and low. These categories will be useful for further investigation. High stores include: RAH, SGA, SMM & QSN. Medium stores include: PAA, RGS, QMD, OSG, NAQ, PGL, OMV & MUY. The rest are low sized stores.

Graphical user interface, diagram, application

Description automatically generated

**Figure 1**. Bar charts of all features and stores

**Figure 2** presents a line graph for the medium stores. It includes eight different stores with 30 days rolling average of customer visits. Since we are working with large amounts of data, a rolling average is applied to avoid the unnecessary accumulation of line plots of data, making the line plot more visible. This line graph displays the frequency of customer visits during that year, where the y-axis holds the number of visits, and the x-axis represents the time series.

Following up with **Figure 2**, it only sheds light on two stores which are QMD and OSG. Such aspect is a key for these two stores because it was apparent that QMD kept gaining visitors with a major ups and downs throughout the year, whereas OSG had minor ups and downs and a semi-constant number of customers. Knowing that both QMD and OSG put 16000 on their marketing expenses, perhaps one of the other features like store size show more effectiveness rather than marketing expenses.

Graphical user interface, chart

Description automatically generated

**Figure** **2**. Graph for medium visitor’s range stores

**Figure 3** also illustrates a line graph for the medium stores. It includes twenty-eight different stores with 30 days rolling average of customers visits. A rolling average was also applied to this dataset to avoid unnecessary accumulation of line plots. This line graph displays the frequency of customer visits during that year, where the y-axis holds the number of visits, and the x-axis represents the time series.

Through **Figure 3**, the reader can extract the stores that have closed at a certain point since there was a sudden decrease of customer visits to zero. Additionally, the graph also shows the new stores that opened during the year, where it was observed that the number of customers gradually increased over time. The stores that closed are XSV and ZMS. The stores that opened are NMO, AEI, NGB, YYO, MAJ and ZSD. This data can be significant for the company’s sales department to create certain claims behind the reason to why the stores are closing and why the new opening stores are successful.

Graphical user interface, chart

Description automatically generated

**Figure 3.** Graph for low visitor’s range stores

Since the company cares about understanding the high visits stores and how they brought the corresponding value, a box plot was applied on high visits data to introduce diversity and a clearer vision as shown in **Figure 4**. Similar to a line graph, a boxplot can handle a large number of data as well. Box plots show the median, which points to the midpoint of the visits data for each store, presented as a yellow line in the center of the box. Further, the upper and lower qualifier present the minimum and maximum values of the store visits. The data that falls outside the minimum and maximum are known as outliers.

Based on **Figure 4**, the first conclusion is that all of the top 4 stores are normally distributed except for ‘RAH’ because it has a low range. This is due to the fact that all of the store’s medians are centered in the box between the 1st and 3rd quartile. RAH has the highest median out of all four stores, making a median of 1050 visits, whilst the median for SGA, SMM and QSN are 950, 930 and 900 respectively. Regardless, SGA and QSN seem to have higher values and range of visits than RAH, where SGA has a range between 700 and 1200 yet QSN has a range of 600-1150 and RAH ranges between 850 and 1100.

Chart, box and whisker chart

Description automatically generated

**Figure 4**. High visits store boxplot

**Figure 5** presents four autocorrelation subplots. The properties of customer visits over time series for the stores with a high number of visitors (RAH, SMM, SGA, QSN). Are they positively correlated, negatively correlated, or independent from one another? The horizontal axis shows the “Lag” size between the peaks of the time series. The correlation peaks in a time series and the existing lags aid in measuring with a one-day lag and then with a two-day. These discovered peaks are referred to as seasonality.

For instance, in **Figure 5**, the focus is on the spikes that rise above the dashed lines. Such spikes are usually statistically significant, which is evidence of autocorrelation. What makes it interesting is that all of the four top stores present a peak every seven days, thus presenting a weekly distribution (seen on x-axis:7-14, 14-21, and so on). Such seasonality could be occurring because of several reasons like weekends, holidays, and sales. Besides, this means that the store’s numbers of visiting customers are correlated with each other. To elaborate more, when the number of visitors rises, the lines seen on the graphs tend to also rise, and vice versa.

Chart

Description automatically generatedDiagram

Description automatically generated with medium confidence

**Figure 5.** High stores autocorrelation plots (seasonality)

After studying the store’s number of customers visiting, it is important to understand the reasons why some stores are more successful than the others. The scatter plots in **Figure 6** are used to identify whether there is a correlation between all of the given features for all the stores. The scattered dots represent values for two different features that are compared together, reaching a total of 10 scatter subplots. The dot’s alignment over the x-axis and y-axis signifies the value for an individual feature. In addition, its diagonal presents a bar plot for every feature, so it presents 5 bar plots. Each bar plot consists of the frequency of the feature’s values, including number of customers visits, marketing expenses, store size, staff number and overhead costs.

In **Figure 6**, most of the plots present a linear increase in the dots, rather than randomly scattered ones. With that being said, the linear increase in **Figure 6** shows that there is a positive correlation between some of the elements in the study. These correlations exist between marketing expenses and customer visits, store size, and staff number. Moreover, the observation shows that there is no correlation between overhead costs and all of the other elements since the points are randomly scattered. Finally, having no decreasing motion of dots shows that there are no negative correlations among the plots.

Chart, calendar

Description automatically generated

**Figure 6.** All features’ correlations

In order to validate the points mentioned before and following up with the correlation scatter plots, **Figure 7** presents radar charts used to display the properties of each store. All five features are equally spaced along with the circles. The inside of the circle is segmented into five different parts, ranging from 0, 0.2, 0.4, 0.6, 0.8, and 1. Hence, it is important to mention that the values used in a radar plot are normalized so that they would in the range of 0-1: the higher the value, the closer it is to 1.

The reason behind choosing only the top stores is to focus on what the company cares about, which includes understanding how these stores managed to excel in comparison to the others. Through observing RAH, it is obvious that the budget spent on marketing expenses is paying off, where both of the values almost reached the maximum (1). In addition, the ratio size of RAH (0.8) is as twice as the staff number (0.4). Even though the marketing expenses for SGA (0.7) were less than the ones of RAH (1), SGA still managed to have a high number of customers visiting their store (0.9). This could be due to the large store size; SGA seems to have one of the largest stores among the others, and again the number the ratio for store size (1) is as twice as the staff number (0.5). This means there is a pattern that makes a store more successful. This pattern includes having high marketing budget, large store size where the number of employees should be half of the store’s size ratio.

Diagram

Description automatically generated

**Figure 7.** Radar plots of high visiting stores

After recognizing that the store size also has an impact on the daily visits of the stores, a bubble graph would be suitable to elaborate on this point. **Figure 8** is a bubble chart of a scatter plot for the marketing expenses and the store size. Along the x-axis lies marketing expenses, and store size lies on the y-axis. Further, to express the value of the third variable which is the number of visits, the data is presented through a bubble instead of a dot; the value of customers visits per store determines the size of the bubble. The darker the color of the bubble is (closer to purple), the higher the number of visits for that particular store.

It is apparent from **Figure 8** that there exists a positive correlation between the three mentioned features. The bubble size is increasing as store size and marketing expenses are increasing. This means that the number of customers visiting per store are also increasing as both of the other elements increase. This claim can be validated when zooming at the medium and small stores (using the Hollo plots on the notebook). Medium stores are able to adapt to market budgeting, thus leading to a high number of customers because they have an acceptable store size. However, in the small stores, it is seen that even though the number of stores is slightly increasing, most of the stores manage to have a constant number of visitors. Moreover, increasing the marketing strategy hardly increases the number of visiting customers in the small stores. Logistically speaking, this is because smaller stores are uncapable to fit a high number of customers simultaneously. Ultimately, applying a marketing budget on a large store can increase the sales (number of visitors) at a much higher rate than in a small store.

Chart, bubble chart

Description automatically generated

**Figure 8**. Store Size vs Marketing Expenses (vs Customer Visits) Bubble graph

The presentation ensures high readability and scalability for this project’s development through following the best idiomatic practices. Starting with identifying the target, the company who handled the data is interested in knowing what trends and observation the data hides, with great focus on the large and medium sized stores. Since the data presented is already clean and meets the requirements for data analysis, we can instantly start with applying the visualizations to extract required information. Through importing Python packages for data science such as Pandas and Hollo Views & Seaborn, the code is reproducible and efficient since it avoids repetition of unwanted code. To achieve efficiency, the code is integrated with for loops to append similar multiple plots. Further, the required libraries imports, and data sets are extracted at the beginning of the code blocks. Since the audience are not experts in reading visualization, it is imperative to direct their attention to specific visual cues and trends; the chosen graphs and charts are mostly effectual and proficient to convey the essential information in the store’s dataset. Moreover, the representation supports labels and caption to explain how to read the given datasets. The notebook satisfies all instructions as it includes markdowns on how to use a block of code where each block of code includes data outputting in texts and figures. Such consistencies and clarities can be achieved through providing static chart’s axis point lines, minimizing clutter and noises as well as making use of a wide range of colors to allow readers to understand data better.

Ultimately, the company is focusing on large and medium sized stores, thus this study presents major findings and trends in those stores. However, there are several points that counter the low sized stores. Of the small stores, XSV and ZMS closed during the year. Furthermore, the stores that opened were NMO, AEI, NGB, YYO, MAJ and ZSD. Another point focuses on QMD and OSG: these two stores have a different number of customers even though they both spent 16000 on their marketing expenses, where QMD had more visitors that OSG. When it comes to the high sized stores, RAH has the highest median out of all four stores but SGA seems to have higher values and range of visits than RAH. In addition, all of the four top stores present weekly seasonality. After plotting the bar plots, a positive correlation was observed between marketing expenses and customer visits, as well as in store size and staff number. Besides, there is no correlation between overhead costs and all of the other elements. There were no negative correlations observed among the mentioned features. Therefore, positive trends are present when stores employ a high marketing budget and increase their store size. Having a marketing budget on a large store can therefore increase the sales (by increasing the number of visitors) at a much larger scale than in a small store. Finally, it is recommended that the number of employees should be half of the store’s size ratio.