Venue Visitors Analysis

Contents

Table of figures

Abstract

Data visualization is a method of converting given data or information into pictorial or visual forms like maps or graphs which will make people understand the information easily and fastly, in this report we are going to graphically represent and bring out the conclusions to the data of a fictional company called Chrisco which has a very big amount of data which conations people visiting their venues using their loyalty card scheme.

Introduction to Data Visualization

People tend to look for images, relations, or patterns in any information or data. Data visualization is a process where data is present in various forms, a large amount of data is generated every second from various sources such as social media, Sensors, etc. Often, companies find it very hard to understand and process them to draw any conclusions. To overcome this problem, the data had to be changed into an easily understandable form like charts, graphs, and maps which are always created in the perspective of how a particular customer wants to view it and it can be made interactive which helps in getting more deeper information like focusing on parts whenever required, zooming in or out, etc.

Data Visualization is more effective when the data is huge. methods in data visualization make it easy to get conclusions and also allow to look into the report which is very easy to understand, that is it simplifies the way data is represented and also it is very easy to share with others

Various data visualization tools are present, which makes representing the data more attractive and easier, and faster for any person to understand.

With a clear understanding of this, we move into implementing these techniques into real-life datasets. (Chatterjee, 2019)

Discussions

1. Total number of visitors at each venue

The below plot was drawn to visualize the total number of visitors at each venue. The pie chart is used to visualize them.

The size of the slice indicates the portion of the corresponding venue. Each of the 40 venues is represented with different shades of colour. Also, the legend shows the venue Id for each colour. The percentage of each of them is also represented. This visualization is interactive and shows the venue Id and the total number of visitors hovering over the pie chart. We can clearly state by looking into the visualization that the venue 'RDA' has the maximum number of visitors for the year.

Figure 1 demonstrates the interactive nature of the plot.

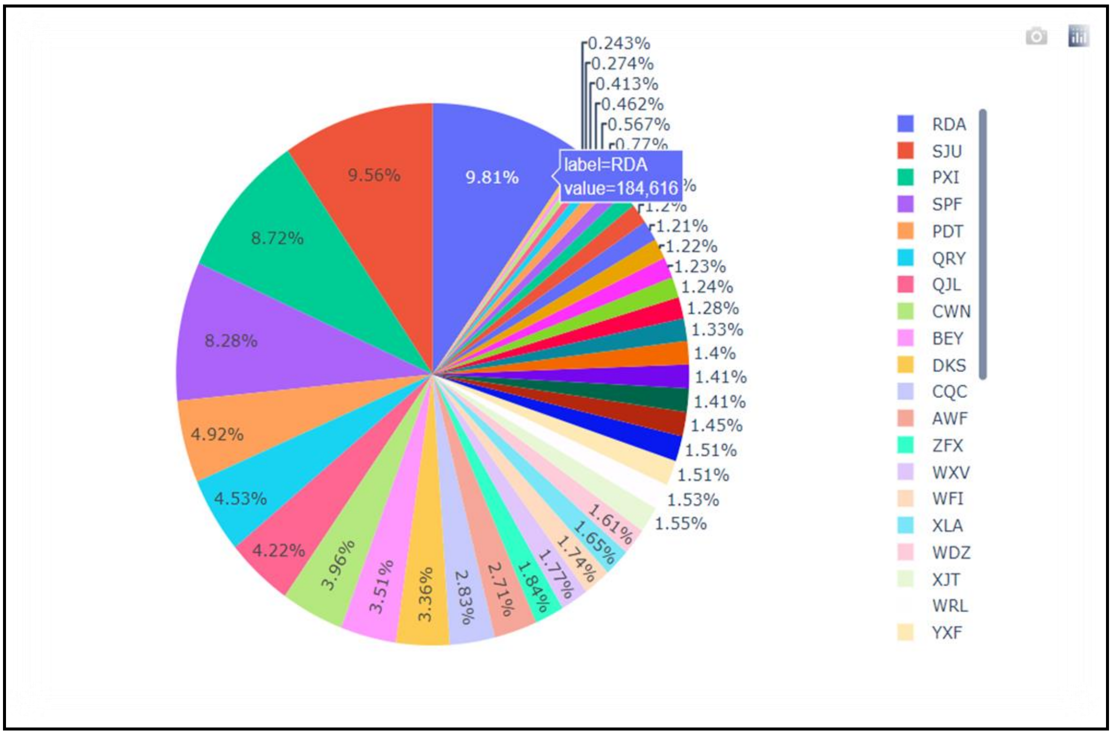

Figure 1: Total number of visitors at each venue

2. Segmentation of venues based on the total number of visitors

This bar chart is used to clearly segment the total number of visitors into 3 categories and distinguish each of them using separate colours.

This is a bar chart representing the total number of visitors at each venue by segmenting them into 3 categories as high, medium, and low. The categories and their ranges are:

a. High - More than 100,000 visitors

b. Medium - Between 40,000 and 99,999 visitors

c. Low – Less than 40,000 visitors

The High, Medium and Low are represented using teal, gold, and magenta colours respectively in Figure 2. Most of the low category venues are recently opened ones. Hence the total visitors count would be less.

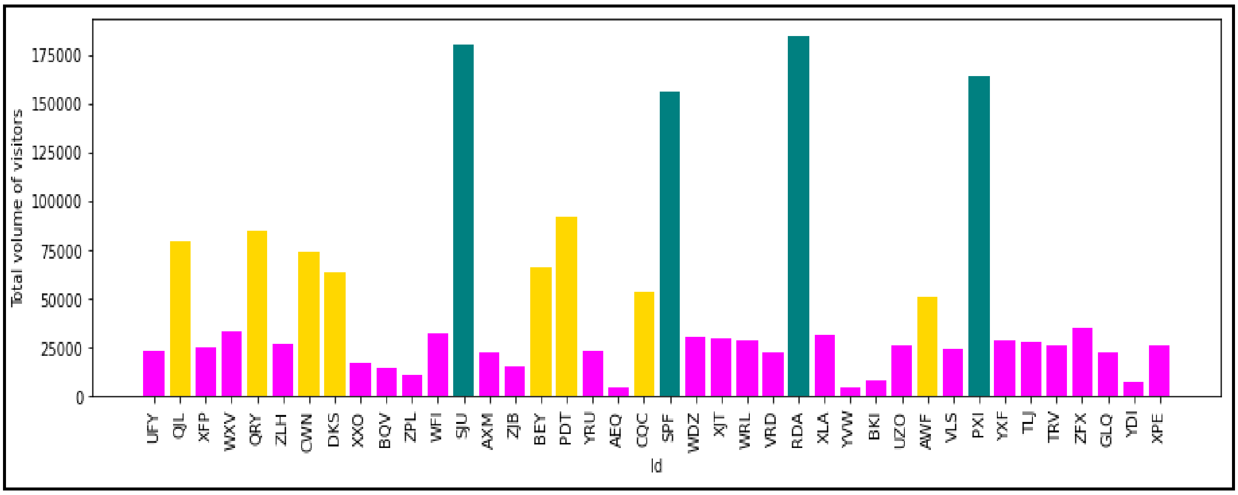

Figure 2: Segmentation of venues based on the total number of visitors

3. Variation in the number of visitors over the year for 'high' category venues

The boxplot is used to represent the distribution of a number of visitors on different days for the high category venues.

The upper limit tells the maximum number of visitors, and the lower limit shows the minimum number of visitors. The Line in the middle shows the median of the values. On hovering over the graph, the values and corresponding dates are displayed as it on an interactive graph. The dots represent the outliers, i.e., the drastic changes in the number of visitors on particular days. We can observe that at venue ‘SPF’, the number of visitors has decreased below 200 and at venue ‘PXI’, the number of visitors has increased above 700.

Figure 3 demonstrates the distribution and its interactive features.

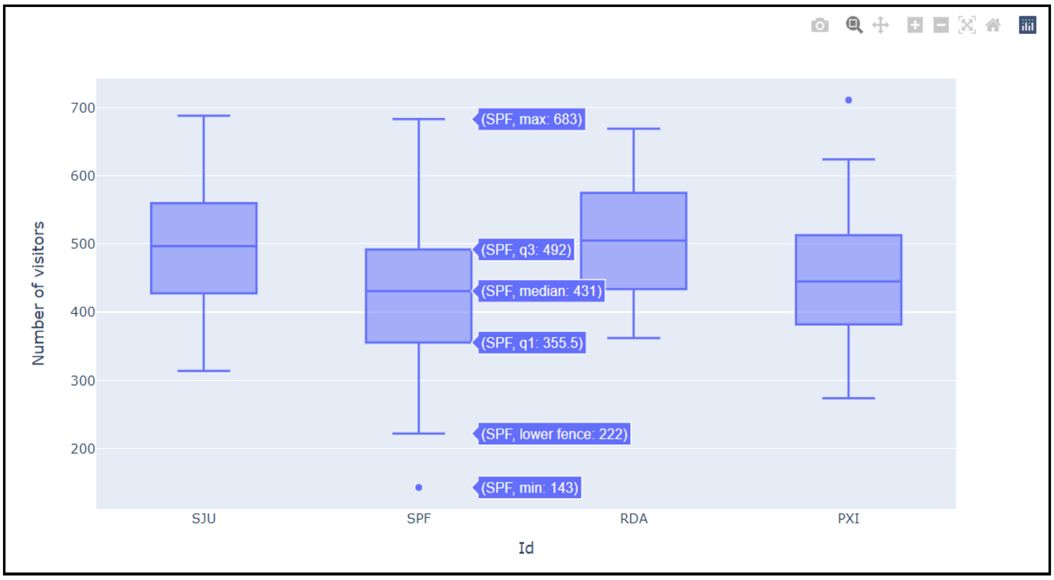

Figure 3: Variation in the number of visitors over the year for High category venues

4. Recently closed and opened venues

To get to the recently opened and closed shops, we need to analyze the number of visitors on each date. This is time-series data and can be best represented using line plots.

The number of visitors on each date is plotted using a line graph for all the 40 venues using 40 different plots. In case, the number of visitors remains 0, we can say that the venue is closed for that day. Similarly, if it is closed for a continuous number of days, it is permanently closed during that frame. We can see from the below plots that at venue ‘XXO’, the number of visitors remains 0 after July 2019. Hence, we can conclude that the venue was closed on July 1. Similarly, the number of visitors remained which remained at 0 until April 2 and started to increase at the venue ’BQV’. So, we can conclude that it is a recently opened venue. Also, the plots which never reach 0 are neither recently opened nor recently closed. From our observations, it can be concluded that:

a. Recently closed venues: XXO, ZPL

b. Recently opened venues: BQV, ZJB, AEQ, YVW, BKI, YDI

Figures 4,5,6 show examples of recently closed, opened, and neither recently opened nor recently closed venues respectively. Figure 6 shows the interactive feature of the plot (zoom and hover).

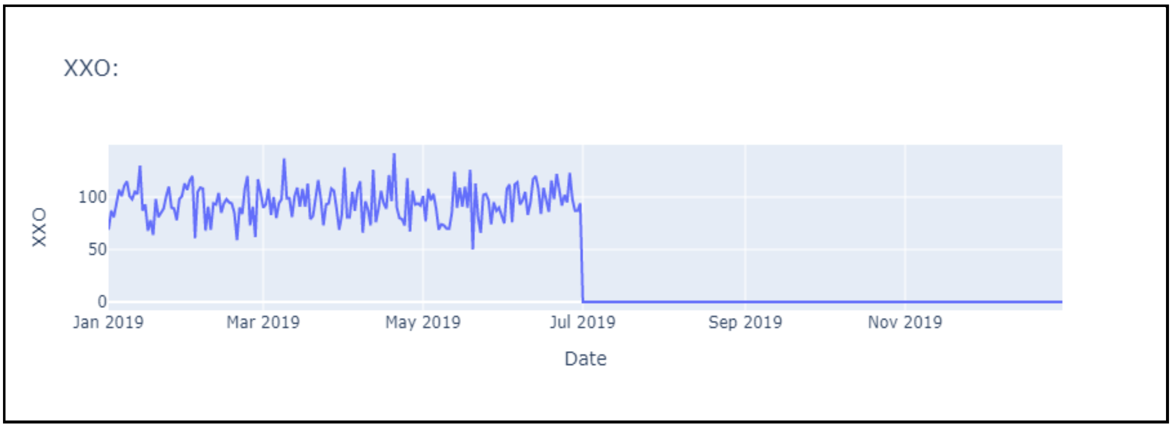

Figure 4: Recently closed venue

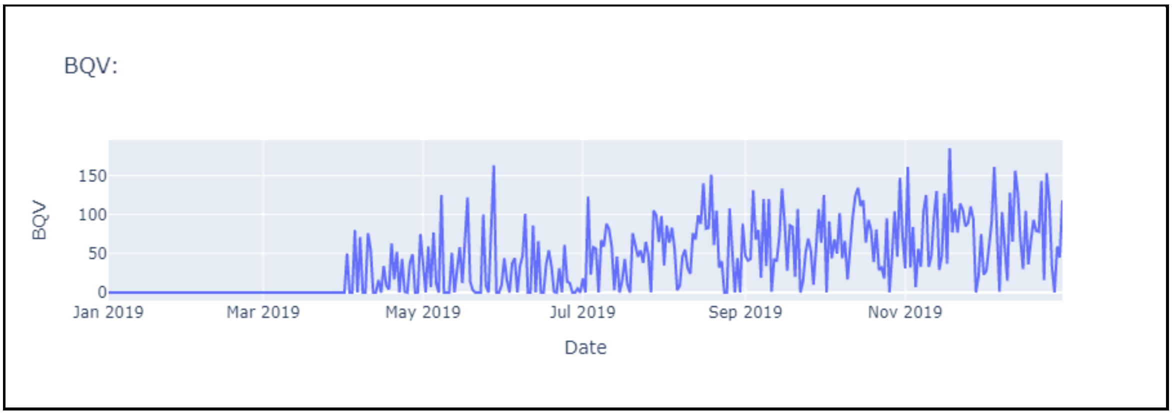

Figure 5: Recently opened venue

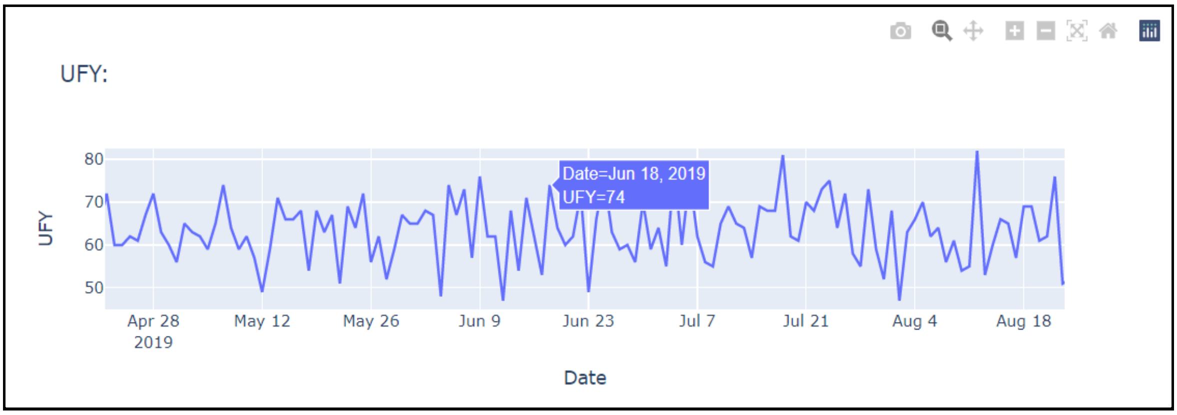

Figure 6: Open throughout the year

5. Correlations between summary data variables

This is a relationship heatmap used to picture how each attribute depends on one another in the summary data frame.

The heatmap contains values from -1 to 1. The index indicates the value of colour shade. The dark blue colour indicates +1 and a strong direct relationship. As the blue shade lightens, the value decreases and reaches white colour which indicates strong inverse proportion and the value becomes -1. From our heatmap, we can conclude that the total number of visitors at a venue is dependent on the maximum distance traveled by them to the venue as the value is +0.94.

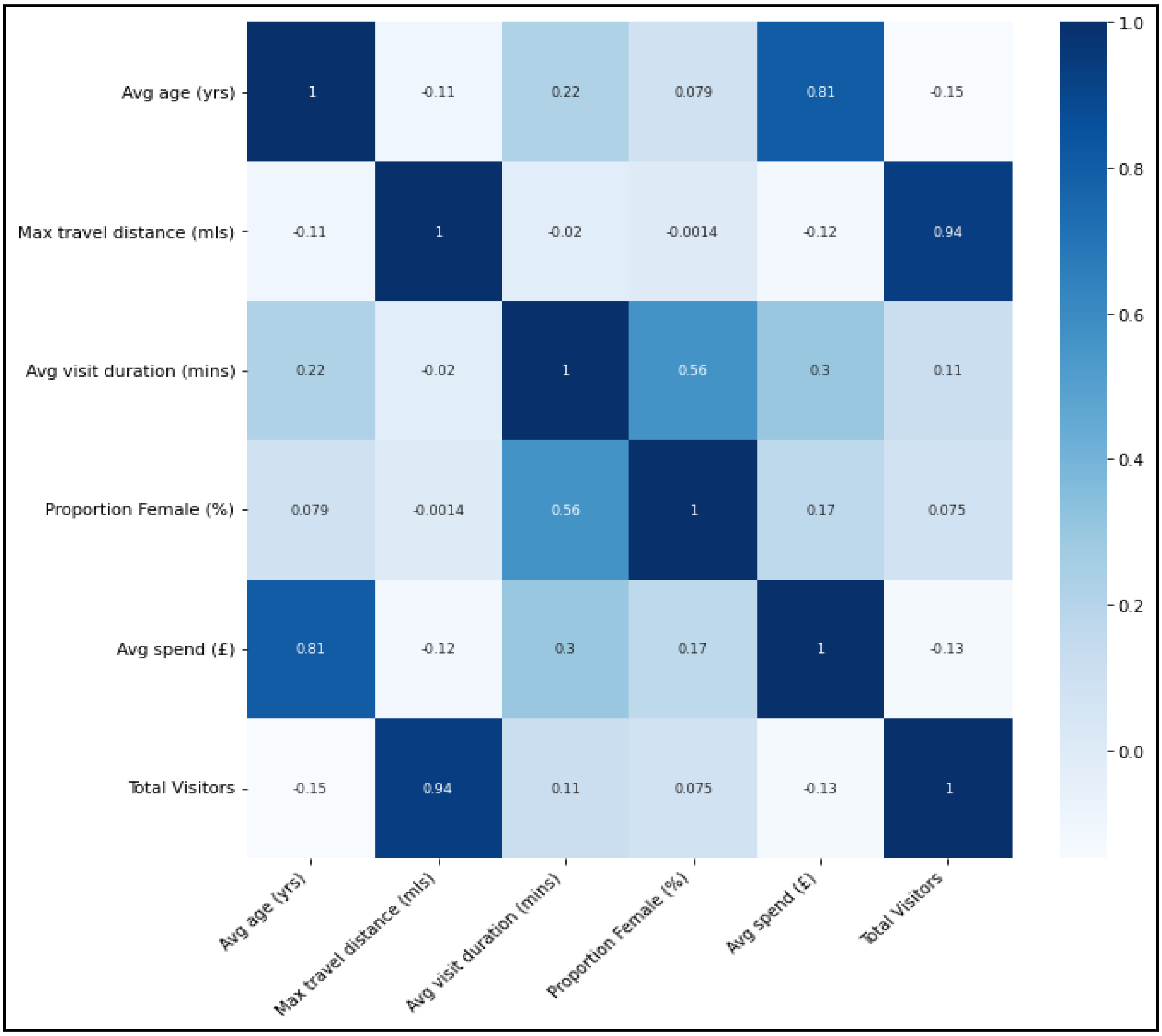

Figure 7: Correlations between summary data variables

6. Histogram for all the attributes of the summary data frame

This histogram was intended to show the number of venues in each distribution considering all the attributes of the summary data frame.

The histogram shows the distribution of venues based on the attributes in the summary data frame. From the graph, we can tell that visitors from most of the venues travel less than 40 miles. And only in 2 of the 40 venues, visitors travel more than 40 miles to the venue. The average age of visitors would most likely be less than 25 or more than 33. Only 1 venue has an average age of visitors between 25 and 33. Visitors spend 60 to 130 minutes on average across all 40 venues. There are 3 venues that have a proportion of female visitors below 45% and 3 venues have a proportion of female visitors above 55%.

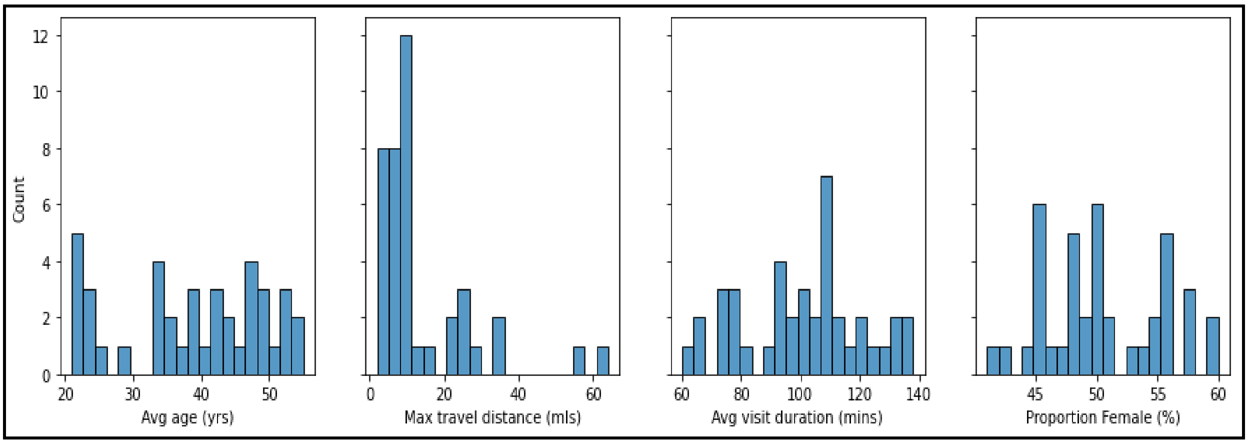

Figure 8: Histogram for all the attributes of summary data frame

7. Female v/s Non-female visitors

The donut has a clear representation of proportion of male and non-female visitors.

The donut chart is exploded. The pink color slice of the donut shows female visitors whereas blue color represents non-female visitors. As we can observe, the two slices of donut have almost same size which indicates they both are of almost equal proportions. The percentage label proves it by showing the exact numeric figures.

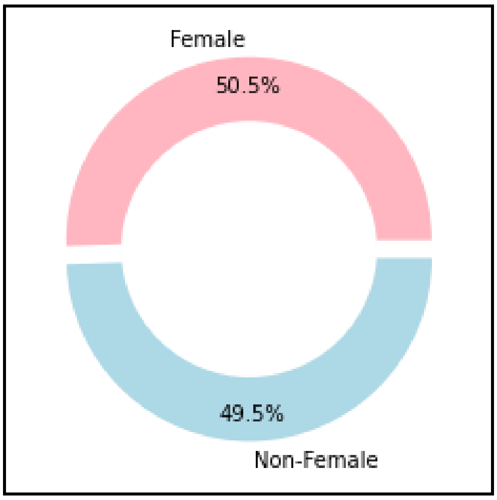

Figure 9: Female vs Non-female visitors

8. Total volume of visitors on different days

The below plot represents the time series data using bar graphs. The bar plot was chosen ahead of the line graph in this case because it has a lot of data points on the x-axis which is better represented using a bar plot.

This is an interactive plot that shows the total volume of visitors on different days. The graph follows a uniform pattern. This tells us that the number of visitors would be highest during the mid-week (Tues, Wed, Thurs), and by the weekend, the number of visitors would fall gradually (Friday, Saturday, Sunday). On Mondays, it would start to increase again.

Figure 10 represents this nature and its interactive features.

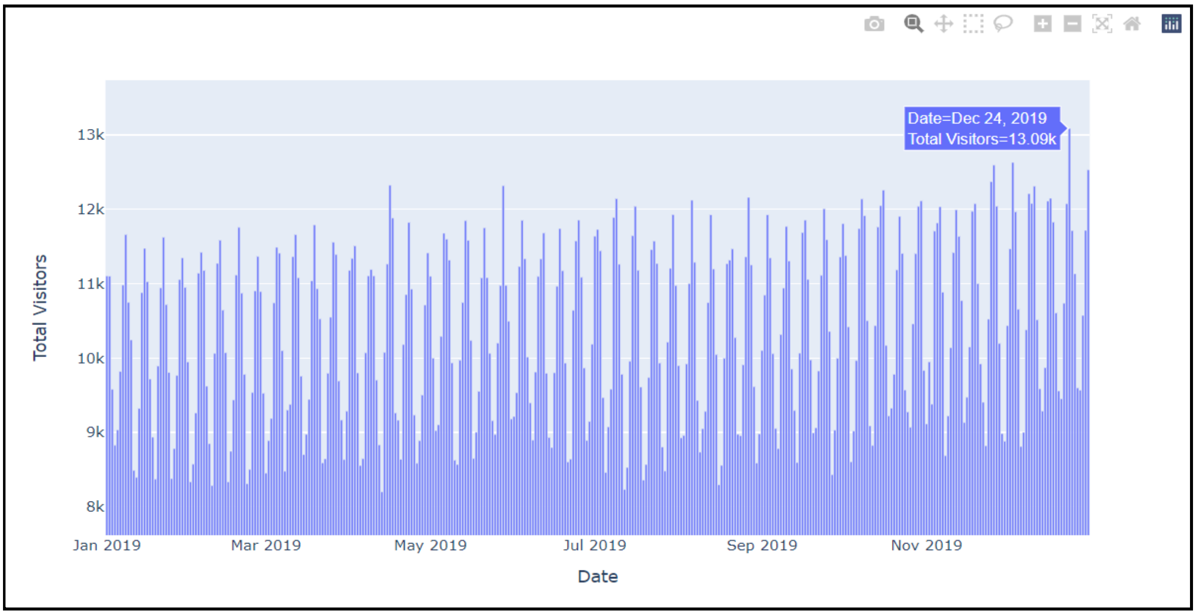

Figure 10: Total volume of visitors on different days

Critical Review

The data visualization course has made me understand multiple ways to manipulate and analyze the data by representing them through various graphs which is easy for a wide range of audience

The data visualizations are made using the Matplotlib and Seaborn libraries. All the code is written without declaring unnecessary variables or creating data frames unnecessarily. The data frame created at each cell is overwritten in the next cell for its specific application without creating another data frame. The interactive visualizations are plotted using the Plotly library. Plotly provides various interactive features for different plots.

Line plots which have a huge number of data points are difficult to analyze and bar plots can be used instead. The heatmap correctly represented the relationships between each of the attributes using correct shades.

Also, I tried to plot a scatter plot considering two separate dates from the Daily Visitors data frame. The scatter plot was drawn, but it didn't give proper insights from which fruitful conclusions could be drawn.

Summary

With the data provided by the ChrisCo company and the graphs formed using that data mainly gives the conclusion telling that

- The venue 'RDA' has the highest number of visitors and the venue 'YVW' has the lowest number of visitors.
- The month of December 2019 has the highest total volume of visitors.
- The ratio of male to female visitors overall is nearly 1:1, with the female population leading the male by 1% of the total.
- The maximum distance travelled by visitors is directly proportional to the number of visitors in the venue.
- The number of venues based on the total number of visitors is classified as:

  a. High: 4

  b. medium: 8

  c. Low: 28

The stores BQV, ZJB, AEQ, YVW, BKI, YDI fall under the category 'Low', as they are recently opened venues.

- The venues that opened and closed recently are:

  a. Recently opened: BQV, ZJB, AEQ, YVW, BKI, YDI

  b. Recently closed: XXO, ZPL

  Also, 'AEQ' is the most recent shop that opened on 3rd October 2019.

References

Chatterjee, M., 2019. *Introduction to Data Visualisation- Why is it Important?*. [online] Available at: <https://www.mygreatlearning.com/blog/introduction-to-data-visualisation-why-is-it-important/> [Accessed 4 April 2022].