COMP 1800
DATA VISUALIZATION
STUDENT NAME: SAI SANDEEP CHITTIBOINA
STUDENT ID: 001203994
COURSE LEADER: DR CHRIS WALSHAW

Table of Contents:

List of Figure:

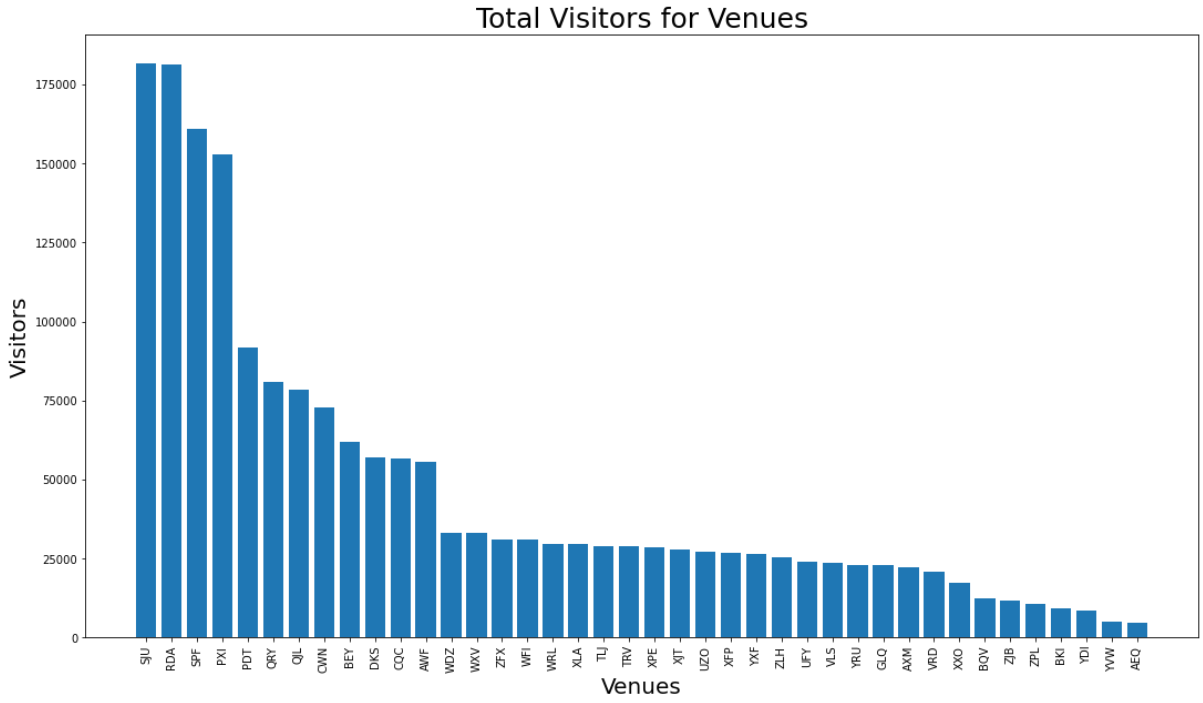
INTRODUCTION OF DATA VISUALIZATION:

Generally our eyes are habituated to give more attention to the colours, patterns, diagrams and pictures. Data visualization is an art of representing the data by using the data which we have texts and numbers into a visual representation this helps us to quickly identify the trends and the outliners of the data. The main goal of Data visualisation is to make it easy to understand to everyone easily.

Some of the most commonly used visualisation techniques in data position are bar chart, heat map, scatterplot, line graph, box plot, pie charts etc

The uses of data visualisation are, Analysing the data in a better way this helps the stakeholders to concentrate on the things which required for them, this helps to take decisions quickly, makes a story of insight, helps to identify the Errors quickly and all of this helps to discover the latest trends and outliners to acquire maximum gains.

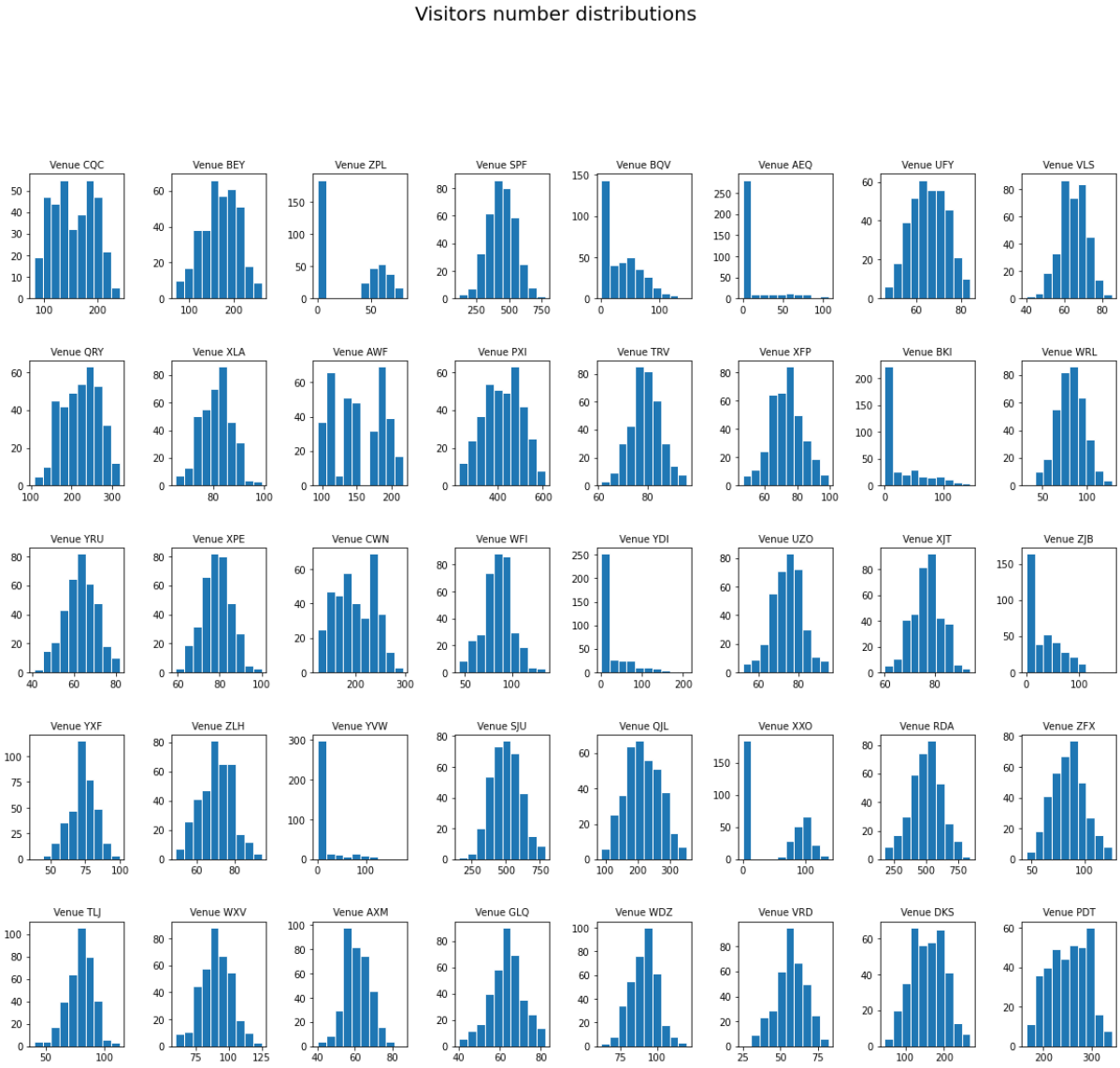
VISUALIZATION 1:



From the (figure 1) on the X axis the venues are plotted and on the Y axis the number of visitors were mentioned in the year 2019 this is a bar graph of the total number of visitors for every individual venues and these venues are sorted in descending order from the highest to the lowest.

From the graph we can say that the total number of visitors range from:
4614(AEQ) to 181794 (SJU)

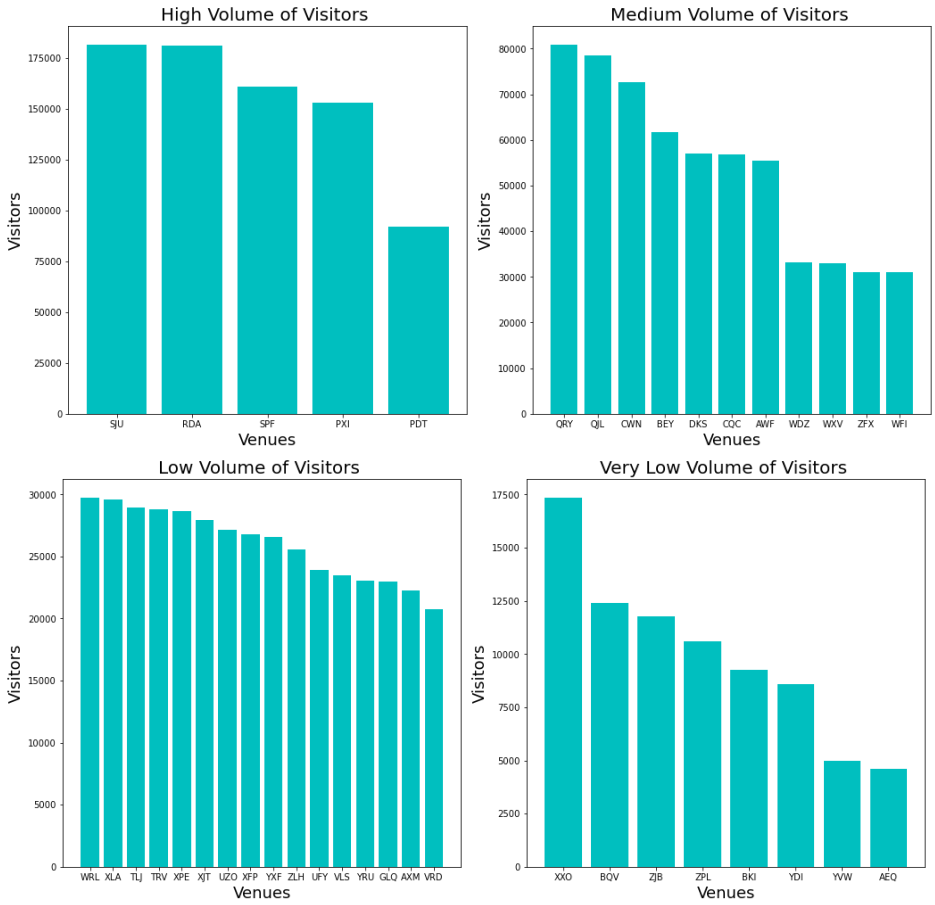
VISUALIZATION 2:



The is Histogram of all 40 venues with its respective visitors numbers, this gives as a clear idea among the venues which one has the maximum amount of visitors histogram is just a bar chart it gives the different heights using the Bins widths.

By looking at this graphs we can say that ‘AEQ’ ‘YVW’ ‘YDI’ ‘BKI’ and ‘ZPL’ when compare it with other graphs these show a very low number of visitors.

VISUALIZATION 3:



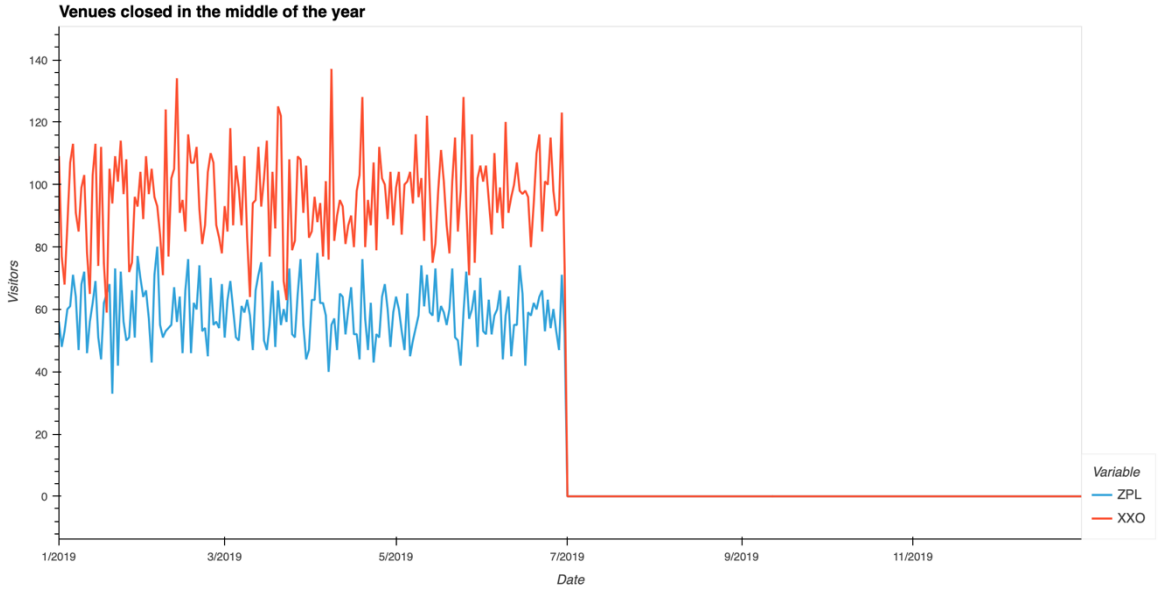
From the (figure:3) the total number of visitors has been classified into four different categories:

Highest volume of visitors = 'SJU' 'RDA' 'SPF' 'PXI' 'PDT'
Medium volume of visitors= 'ORY' 'QJL' 'CWN' 'BEY' 'DKS' 'CQC' 'AWF' 'WDZ' 'WXV' 'ZFX' 'WFI'
Low volume of visitors 'WRL' 'XLA' 'TLJ' 'TRV' 'XPE' 'XJT' 'UZO' 'XFP' 'YXF' 'ZLH' 'UFY' 'VLS' 'YRU' 'GLQ' 'AXM' 'VRD'
Very low volume of visitors 'XXO' 'BQV' 'ZJB' 'ZPL' 'BKI' 'YDI' 'YVW' 'AEQ'

By classifying the total number of visitors like this we can identify on which group we need to particularly concentrate on, this helps to analyse on the venues easier.

If it's more than 85000 visitors I considered it as high-volume visitors venues
If it's more than 30000 and less than 85000 visitors I considered it as medium-volume visitors venues
If it's more than 20000 and less than 30000 visitors I considered it as low-volume visitors venues
If it's more less than 20000 visitors I considered it as very low-volume visitors venues

VISUALIZATION 4:

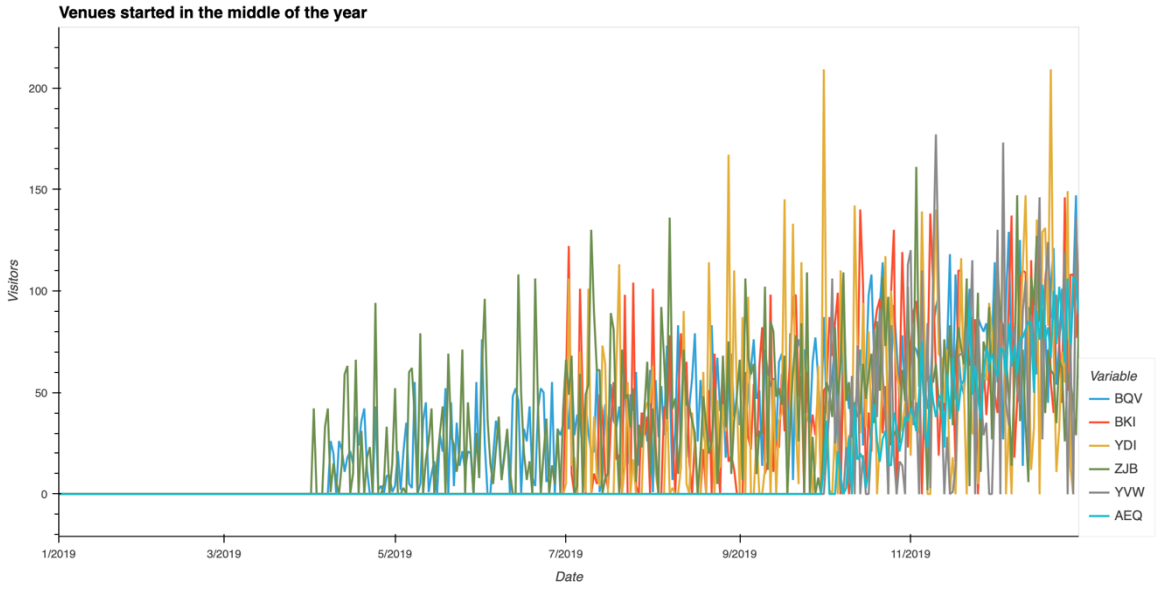


The above figure 4 is an interactive version of line plot this provides us with some tools like box zoom, hover, pan, reset and to save these tools will help us to get a clear idea and what are the outcomes of the visualisation.

Using this interactive line plot we can say on which particular date the venues are closed in the middle of the year

By this graph we can say that 'ZPL' & 'XXO' are closed on the same dates 01-07-2019.

VISUALIZATION 5:



The above figure 5 is an interactive version of line plot this provides us with some tools like box zoom, hover, pan, reset and to save these tools will help us to get a clear idea and what are the outcomes of the visualisation.

Using this interactive line plot we can say on which particular date the venues are started in the middle of the year

“BQV” started on 08-04-2019

“AEQ” started on 01-10-2019

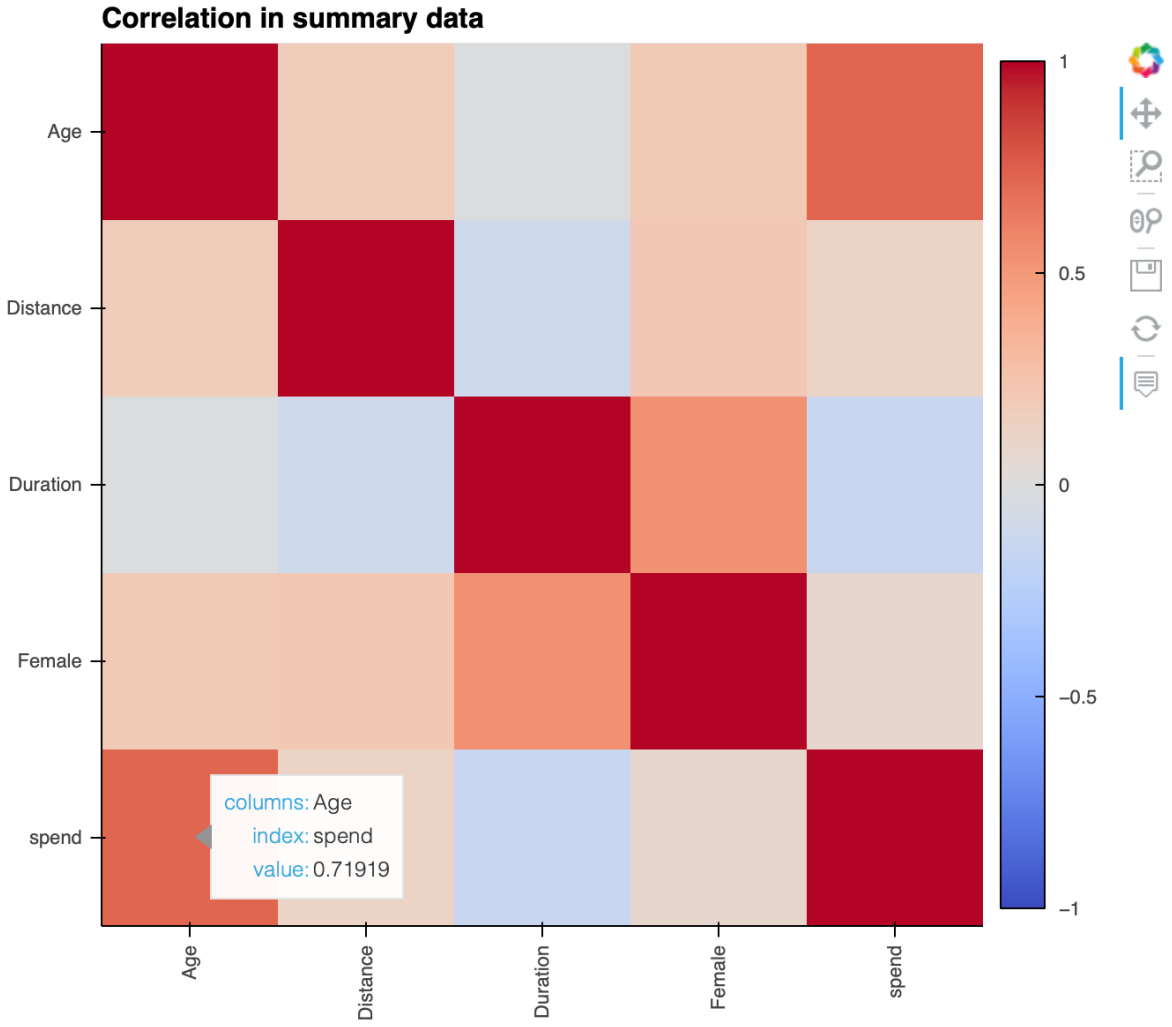
“BKI” started on 01-07-2019

“YDI” started on 01-07-2019

“ZJB” started on 02-04-2019

“YVW” started on 02-10-2019

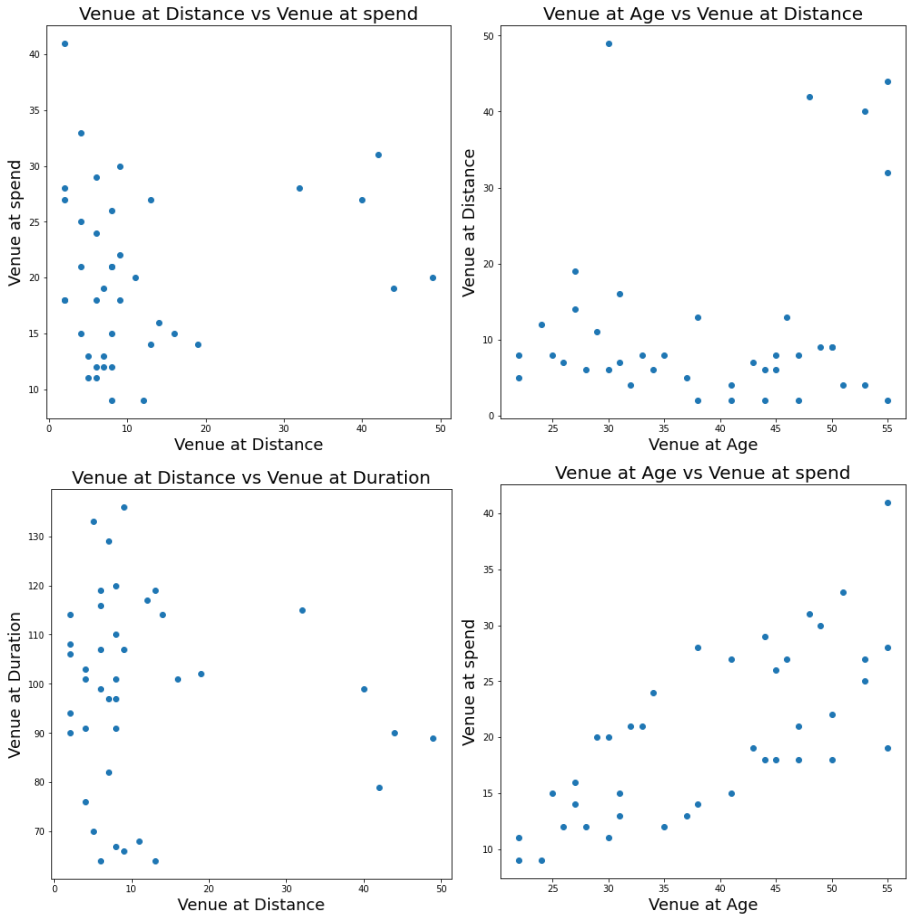
VISUALIZATION 6:



Summary data means all data like age distance duration female ratio spent amount in the venues is summarised for analysing the correlation between them using a heat map by which we can say whether this relation is positively correlated or negatively correlated.

The positive correlation is when the courier data is nearby 1.00 and the negative correlation is when the value is near -1.00 the above heat map is an interactive heat map so when we place a cursor on the heat map vehicle get the column Index and value by looking at the value we can say that whether the column in the index are positively correlated or negatively correlated to each other.

VISUALIZATION 7:



The scatterplot is used to show the relationship between two numeric variables it is most commonly used to identify the correlation between the variables it can be described in many ways positive or negative or it might be strong or weak or it can be linear and non-linear.

The above scatter plots are plotted to see the relation between:

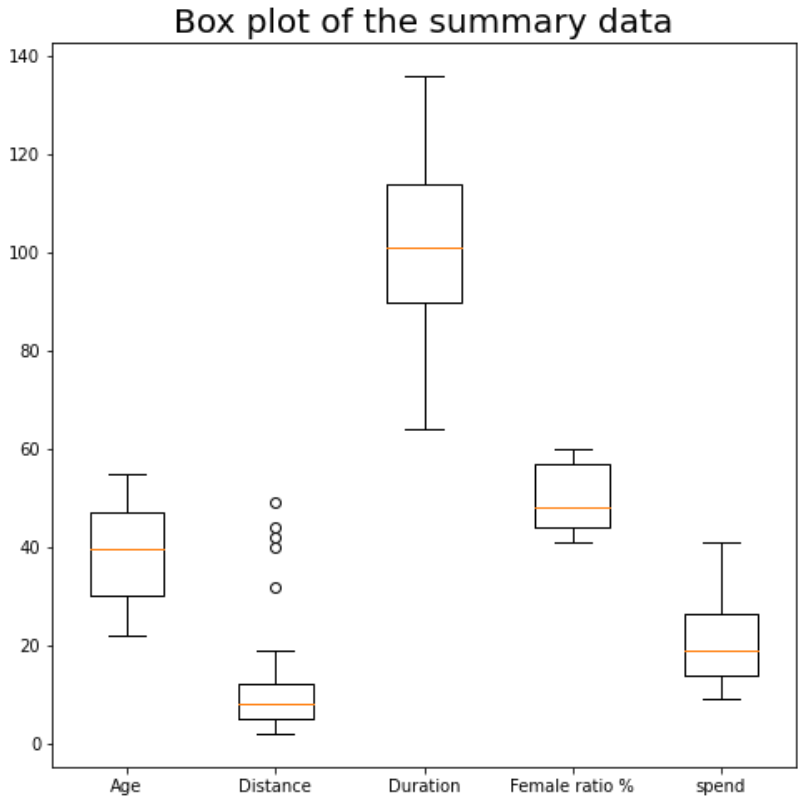
Distance of venues & Amount spend at the venues.(the visitors who came from short distance spend more money at the venues)

Age of people at venues & Distance of venues.(all age visitors are willing to go for the nearest venues which are there for them)

Distance of venues & Duration spend at the venue.(the venues with the short distance are seems that the people are spending more time in them)

Age of people at venues & Amount spend at the venues.(here we can say that young people are spending less amount when compare to older once)

VISUALIZATION 8:



The box plots are the graphical representation of the six things minimum, Q1, median, Q3, maximum and outliers. The box plots are separated into 25% approximately of the data in the set this helps to identify quickly the mean value the dispersion of the data set.

By looking at the above box plot we can say thee mean of the age of people who visited to venues is 40 and the mean distance is approximately 8 and the mean duration spend at the venues is 1hr 40 mins and the mean amount spend at the venues is approximately 20 pounds.

CRITICAL REVIEW:

From the data visualization I have learnt to analyse the data with the help of bar graphs, line plots, histograms, heatmaps, box plots and many other ways. For the corelation between the data I have learned to use of heatmap and scatter plots to find them weather they are positivity or negatively corelated or we can identify if they are linear or non-linear or mix or min. And by using the interactive versions of the graphs we can get more detail information regarding the data set by using the tools which are there in it such as zoom in, zoom out, box zoom, hover, undo, redo, save as and many more.

By using the line charts and the bar graphs I just got to know how to put a data set in a visualised Manner in to get the insights of the data by just looking at the graphs and charts. This will definitely help to analyse the behaviour of the data whether it is correlated and its outlines and its seasonal behaviour by the change in time. By overall we can identify the trend lines where we are getting the highest number of visitors to the venues and the venues who are getting with the medium volume of the visitors and we can say how the other attributes of the data like age distance spent amount duration female ratio are happening at these venues so in will affect the change in the volume of the visitors to the venues.

CONCLUSION:

In this report I have analysed the data on the number of visitors visiting for the venues where I found the volume of the visitors are like this.

Highest volume of visitors: 'SJU' 'RDA' 'SPF' 'PXI' 'PDT'; Medium volume of visitors: 'ORY' 'QJL' 'CWN' 'BEY' 'DKS' 'CQC' 'AWF' 'WDZ' 'WXV' 'ZFX' 'WFI'; Low volume of visitors: 'WRL' 'XLA' 'TLJ' 'TRV' 'XPE' 'XJT' 'UZO' 'XFP' 'YXF' 'ZLH' 'UFY' 'VLS' 'YRU' 'GLQ' 'AXM' 'VRD'; Very low volume of visitors: 'XXO' 'BQV' 'ZJB' 'ZPL' 'BKI' 'YDI' 'YVW' 'AEQ'

So after dividing them into 4 groups I got to know that some of the venues were opened in the middle of the year they weren't there from the beginning of the year. I found them after plotting them in a line plot they are "BQV" started on 08-04-2019 "AEQ" started on 01-10-2019 "BKI" started on 01-07-2019 "YDI" started on 01-07-2019 "ZJB" started on 02-04-2019 "YVW" started on 02-10-2019 there are some venues which are closed in the middle of the year they are 'ZPL' & 'XXO' closed at 01-07-2019 this seems that both of the venues are closed on the same day.

And after looking into the summary data I understood that the distance is one of the major aspect for the visitors if the distance is getting more people are not willing to come to the venues, and the young people are not spending that much amount at the venues when compare to the old once.

Reference:

Tableau (2020). *Data Visualisation beginner's guide: a definition, Examples and Learning Resources*. [online] Tableau Software. Available at: https://www.tableau.com/en-gb/learn/articles/data-visualization.

Business Insights - Blog. (2019). *Data Visualization Techniques for All Professionals | HBS Online*. [online] Available at: https://online.hbs.edu/blog/post/data-visualization-techniques.