Data Visualisation Coursework

Submitted by:- **Sohel Arsad Salam**
Student ID:- **001166291**

Subject:- **COMP1800 DATA VISUALISATION**
SCHOOL OF COMPUTING AND MATHEMATICAL SCIENCE

Q1. INTRODUCTION TO DATA VISUALISATION

In simple terms, the process of converting certain information into a visual representation, for instance a map or a graph, in order to make the data easier to comprehend and extract significant results is what is known as Data Visualisation.

Major purpose for data visualisation is to make the pattern recognition, trend recognition and outlier recognition simpler in massive datasets. Data Visualisation often goes by different terms such as Information Visualisation, Information Graphics, Statistical Graphics.

In Data Science, Data Visualisation is one of the many processes according to which the collected, processed, and modelled data needs to be visualised to draw significant conclusions. It is also a part of DPA (Data Presentation Architecture) discipline which strives to effectively and efficiently identify, locate, modify, process, and transmit data.

Almost every profession requires data visualisation. For instance, professions like teachers, computer scientists, CEOs use data visualisation almost everyday to display, research, and share information. Data Visualization plays a vital role when it comes to massive datasets and BigData projects.

For similar reasons, sophisticated analytics completely relies on visualisation. When we build a Machine Learning algorithm or advance predictive analytics, it's important to visualise the output to keep track of the conclusions and make sure the models are working as they are meant to be. And the most important part, visually represented algorithms are much easier to understand than any numerical outputs.

Advantages of Data Visualisation:

- Analysing large amounts of data in a clear and consistent manner with graphic representation.
- Easier to understand the facts, draw conclusions and view perspectives.
- Easier to identify developing trends and respond promptly.
- Seeing a graph, chart, or other visual representation makes the data easier to comprehend.
- Assisting decision-makers to understand the evaluation of business data better to make business decisions.

About the Provided DataSets:

Basically, we have been given six datasets of an infamous company named ChrisCo. The company manages a variety of venues all across the United Kingdom. Compiling the given datasets into two dataframes, one is of the daily visitors' data and the second one would be the summary data where all the datasets (Age, Distance, Duration, Gender, Spend) are compiled into one dataframe. After compiling the datasets into dataframe, we need to perform various data visualisations for conclusion and characteristics: such as correlations, seasonal behaviour, outliers, and so on.

Q2. Justification & Description of the 8 Data Visualisations

1. **Bar Graph:**

A Bar graph is a type of graph that uses a series of bars to display data on two axes. The x-axis divides the data into groups, with one bar for each category. It creates bars of different colours for every instance. The values for each category are shown on the y-axis. It has more applications when compared to pie charts, its versatility allows us to exhibit percentages, totals, counts, and a variety of other data. As long as you can categorise the contents of the x-axis in an acceptable fashion, it could be according to time or category. It is typically the preferable choice unless we have an unique need to use a pie chart.

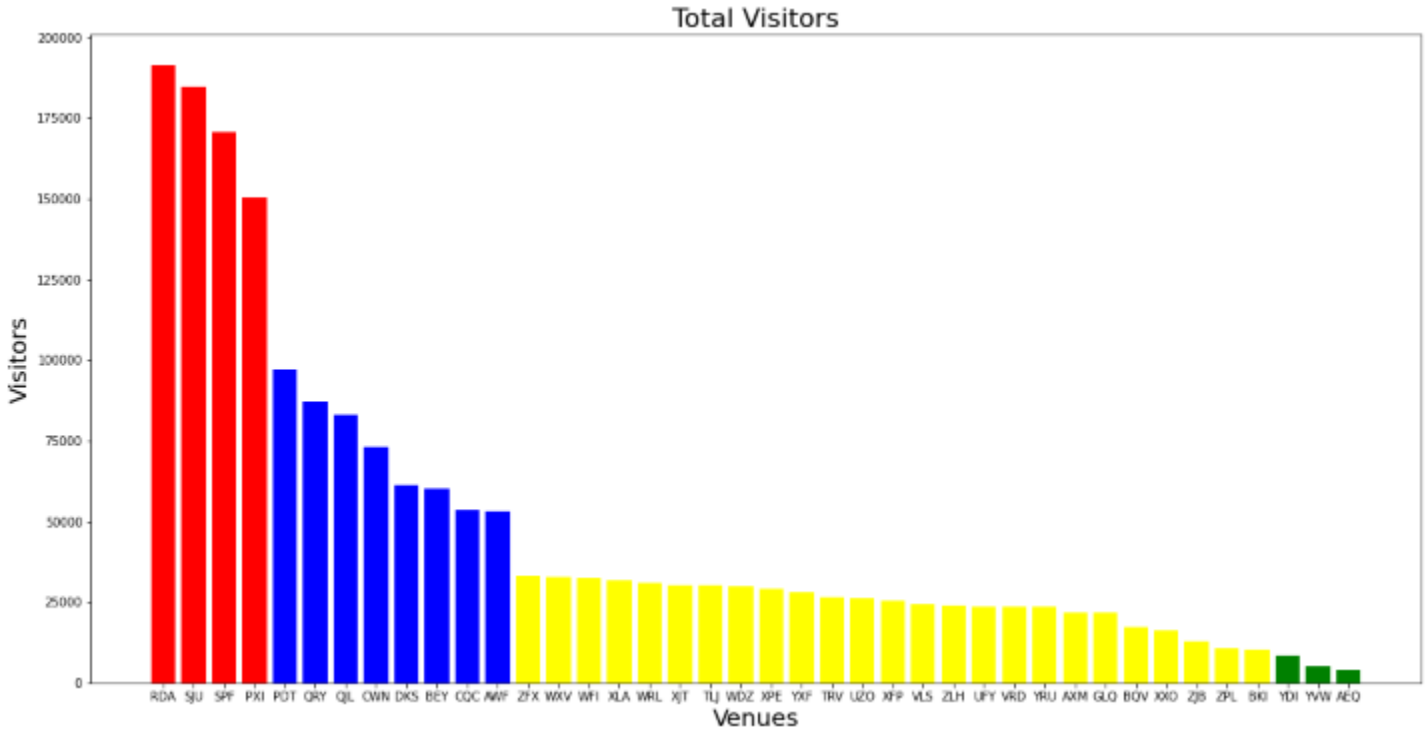

Figure 1: Bar Chart showing Visitors for all venues for year 2019

The above bar chart shows the total number of visitors for all the venues in the year 2019. As it is a large dataset. It needs to be segmented in order to make the visualisation more convenient. Segmenting the venues in four categories i.e. High, Medium, Low, and Very Low, each venue has got its individual bar and is being visualised in descending order showing the visitors from highest to lowest. Looking at the chart we can tell that RDA, SJU, SPF, and PKI have the highest number of visitors and the reason could be many things for instance, these venues have more amenities than the other venues and evidently more popular as well. Whereas the YDI, YVW, and AEQ venues have the least number of visitors throughout the year.

2. **Heat Map:**

Because of its precise nature we included the heatmap. Heat Map is a more detailed and easier to comprehend visualisation. Instead of creating a different graphics, the information is overlain on the map itself. The data is represented by colourful shapes known as isopleths that correlate to different values which makes it easier to identify any correlation between two categories. Heat Map visualises the correlation mathematically between -1 to 1 allowing users to draw more efficient conclusions.

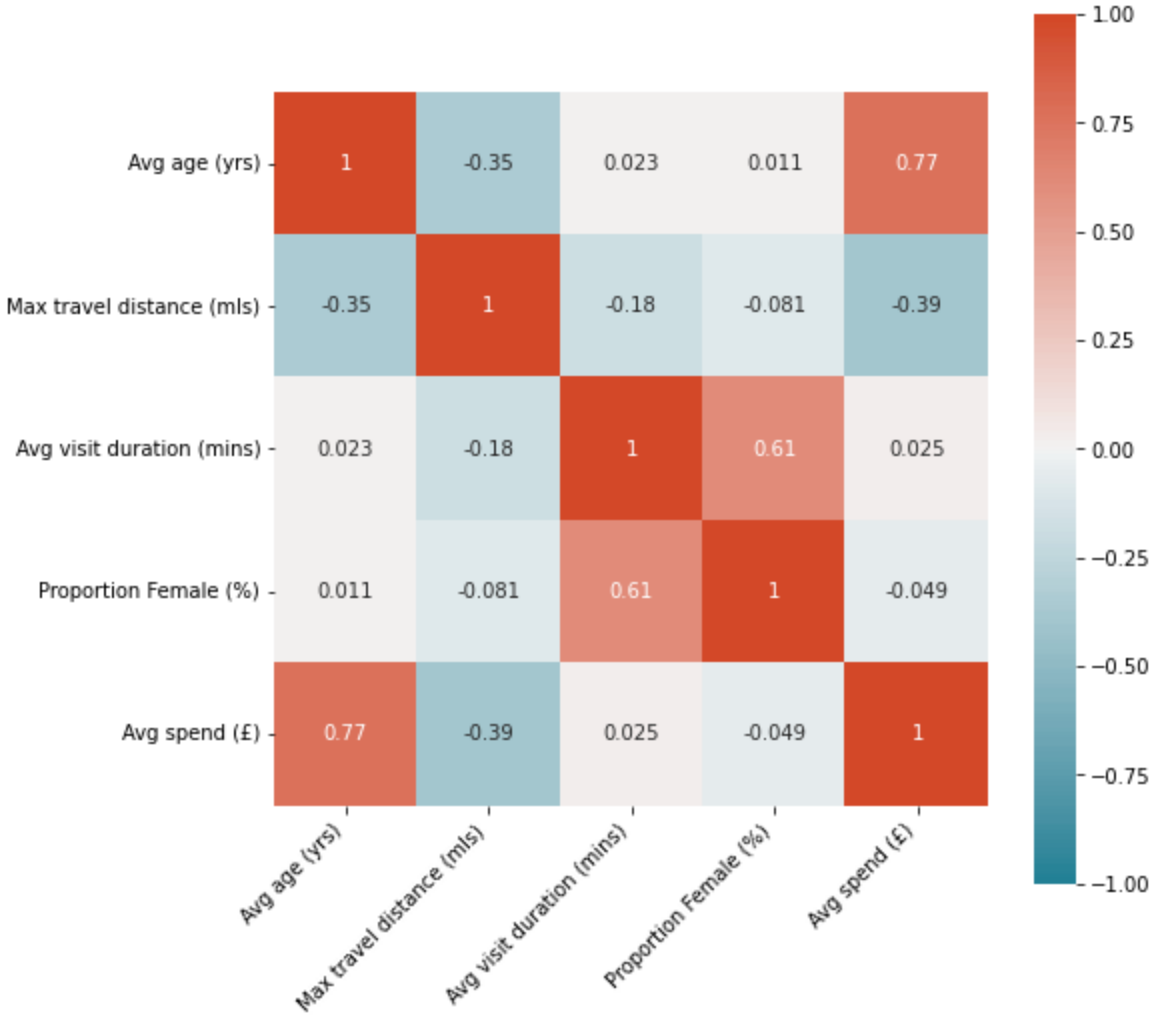

Figure 2: Heat Map showing correlation among the summarised categories.

The above Heat Map shows the correlation among the summarised data instances where we can tell how strongly the instances are correlated to each other. '1' is for perfect positive correlation, where the instances are directly positive proportional i.e. increasing together and '-1' is for perfect negative correlation where if one instance increases the other decreases. In figure 3, Avg age and Avg Spend have strong positive correlation suggesting that the more the average age of visitors, the more they tend to spend for the venues.

3. **Scatter Plot:**

A Scatter Plot uses dots to indicate data for different nominal values, the values for each data point are indicated by the positioning of each dot on the x and y axis. It is used to see how variables relate to each other and also to identify patterns. It is often used to identify correlation and clusters in the dataframe. Outliers can also be identified and fixed using a scatter plot.

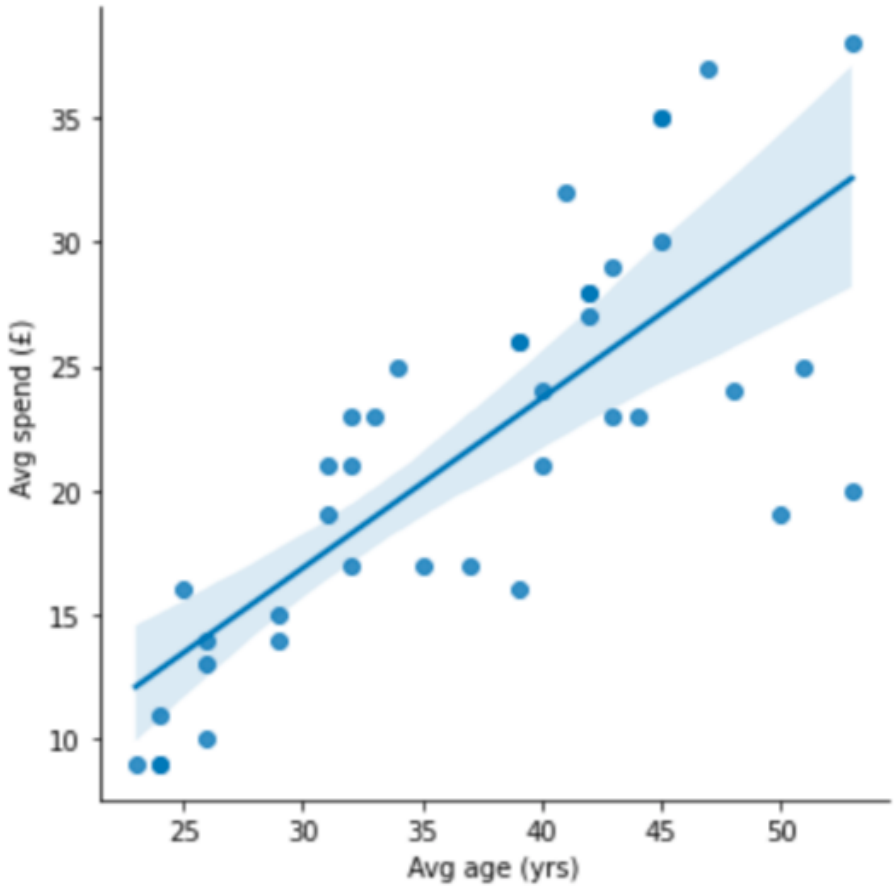

Figure 3: Scatter Plot for Average Spend on Average Age

Figure 3 shows a Scatter Plot with little dots where the numbers overlap for the venues. There's also a trend line representing the general movement of data which is obliquely to the right. It's a scatter plot of Average Spend on Average Age which further provides evidence for the analysis predicted by the heatmap earlier that Average Spend and Average Age are indeed strongly positively correlated. As the average age of visitors increases, the average spend also increases.

4. Line Graph:

A Line graph is needed to get some additional information on the visitors for the entire year. A Line Chart/Graph is very helpful to depict the progression of one or more nominal variables, straight line segments link the data points. The measurement points are organised and linked by straight line segments, kind of similar to scatter plot. Line Chart is very popular when it comes to depict a pattern in data across a time series and so the lines are frequently drawn in a chronological order.

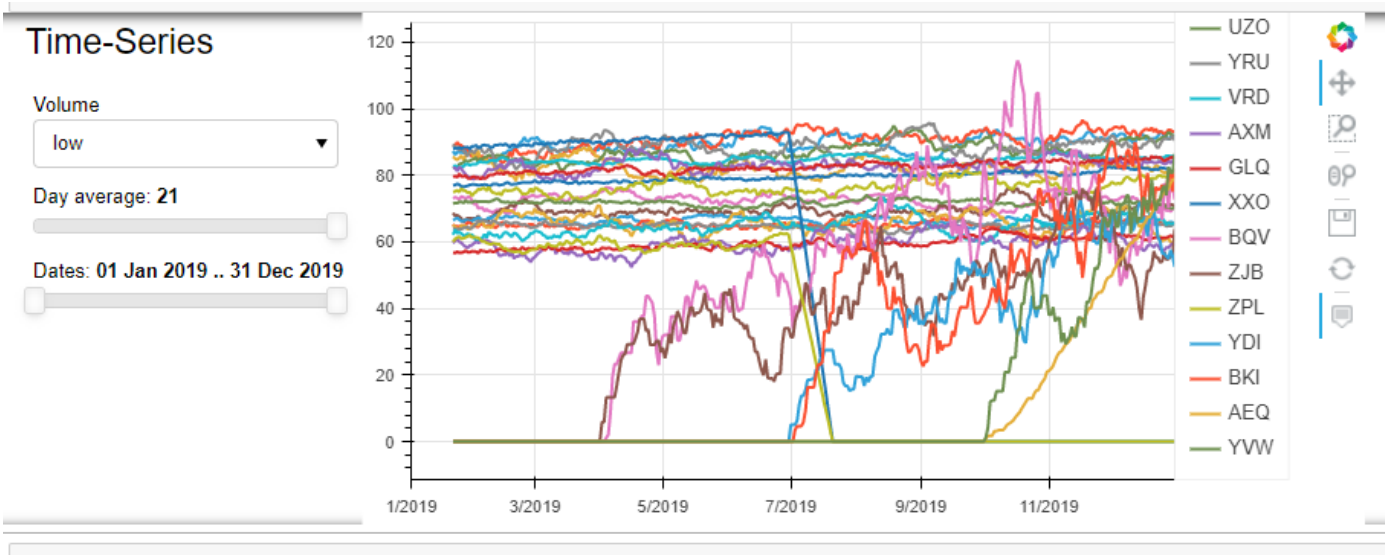

Figure 4: Line Graph showing venues with low volume sales with weekly trends.

In Figure 4 we can see the Line chart of the venues having low volume sales. The depiction of change in visitors throughout 2019 can be seen as all the venues are represented by a different colour. As it is an interactive visualisation, we can always play around with the attributes value and the rolling average and also manipulate the graphics. XXO and ZPL venues have no activities after 6 months. ZJB, BKI, BQV, YDL are the newly opened venues and have a high volume of visitors. Venues like AXM, VRD are the ones having a low volume of visitors.

5. **Box Graph:**

To make the visualisation more simple and clear to comprehend, a box graph is included.A box plot is useful to visualise the distribution of data, every single attribute has its own box plot. In each box, there are three quartiles and also the maximum and minimum values. The standard deviation and the mean could also be displayed as dashed lines. Outliers are shown as points in the box visualisation and are indicated by the spacing between different parts of the box, which represent the spread and deviation in the data. It can be done either vertically or horizontally.

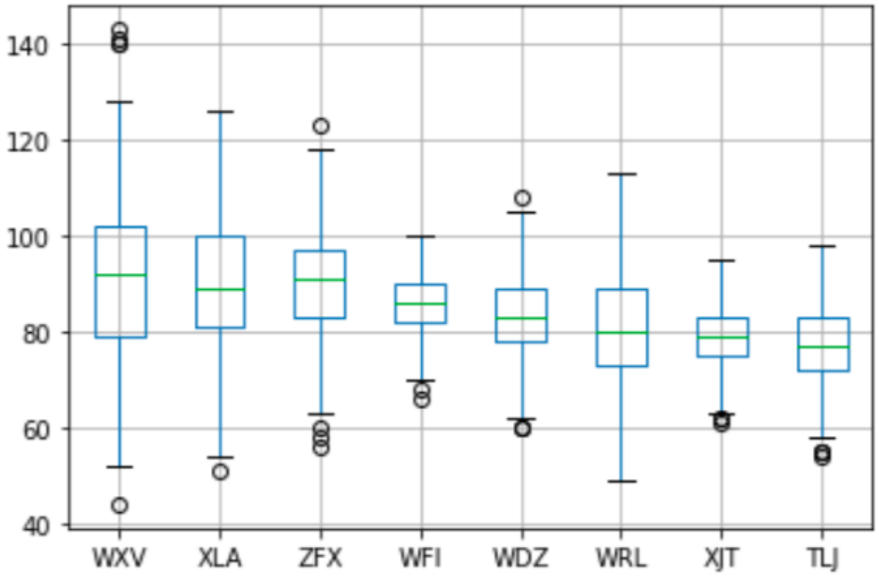

Figure 5: Box Plot showing the outliers of venues.

Figure 5 shows the outlier of different venues having low volume visitors, mostly the medium volume visitor venues have outliers on either lower or higher end. WXV venue has most of its outliers on the Q1 end whereas ZFX, WFI, XJT have their outliers on the Q2 end.

6. **Radar Chart:**

To visualise the summarised data frame entirely we are using a Radar chart. Radar charts are a type of multivariate data visualisation mostly used for plotting one or more sets of values against a common variable.Radar plot enables the user to have a comprehensive comparison of all the venues' statistical data rather than merely comparing among the individual categories and also creates extra linkages among the ignored categories. Radar plot is a contrasting and neat way to visualise the dataset making it more convenient to identify trends and relationships that might have been ignored.

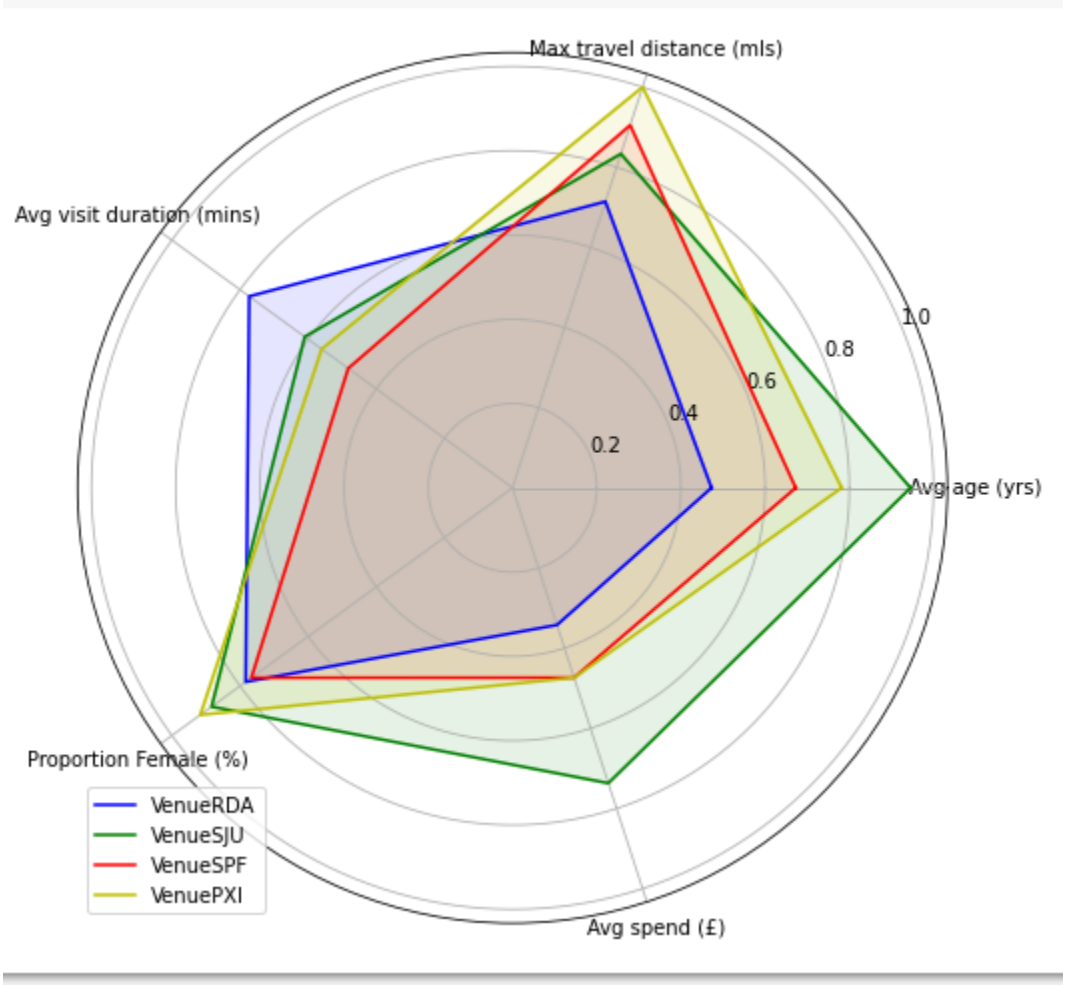

Figure 6: Radar Plot showing Venues with High Volume of Visitors.

In figure 5, the Radar plot shows the venues with a high volume of visitors throughout the year 2019. One can analyse that even here the average age and average spend are directly proportional as shown in the scatter plot and the heatmap. We can also see a relation between the distance and female proportion, as one increases so does the other. As the Distance increases, the average spend decreases. And with the increasing duration, the average age and average spend both decreases.

7. **Bubble Chart:**

This visualisation is included to further provide evidence for the correlation by plotting the attributes that correlates. A bubble chart is very similar to a scatterplot and is often used to depict correlation among three or more quantitative variables. Every individual data point is represented by a bubble.

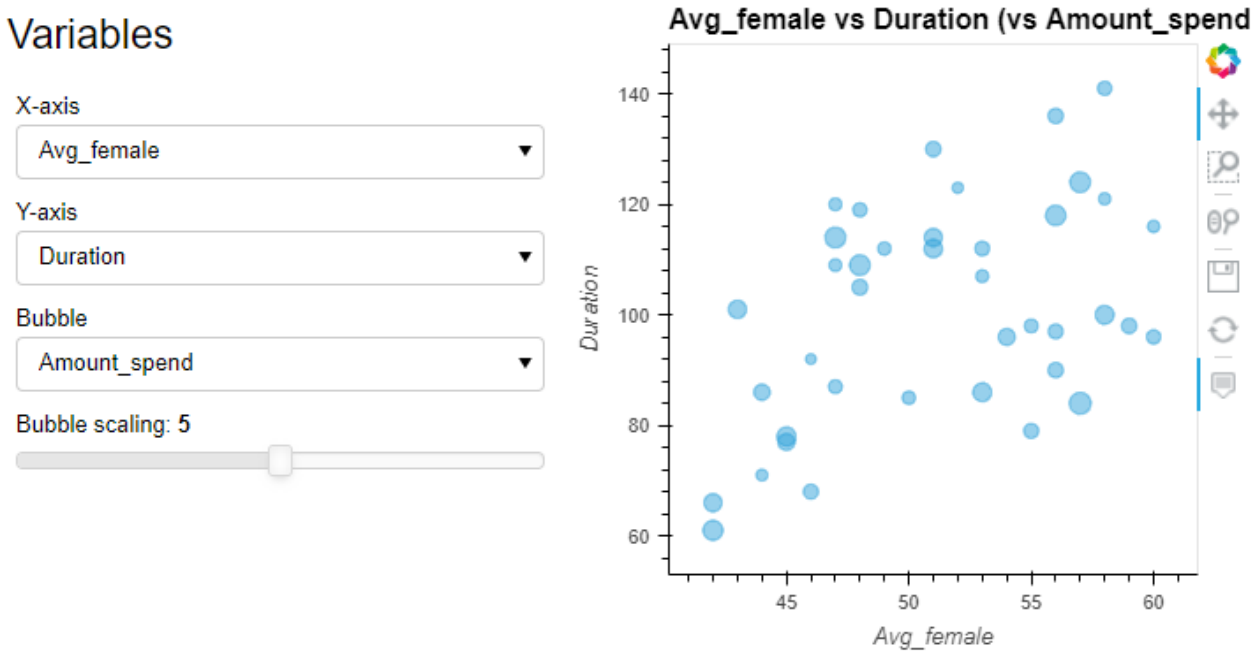

Figure 7: Bubble Plot shows visitors against female percentage.

In figure 7, the bubble plot shows the enlarged plots to represent the separate variables enabling all the three attributes for comparison. It also demonstrates the correlation we discovered in radar graph to be valid and that it is indeed a strong positive correlation. As the average female percentage and duration both compliment each other. It is also an interactive visualisation meaning we can always change the variables and can also manipulate the graphics as required.

8. **Histogram:**

A histogram shows the nominal data in ranges, with each bar representing the frequency with which integers fall into that range. It's very similar to a bar graph consisting of a series of vertical bars on the x-axis. It's often used to show how a piece of data looks in aggregate. It shows whether a dataset values are concentrated around a limited range or more evenly distributed.

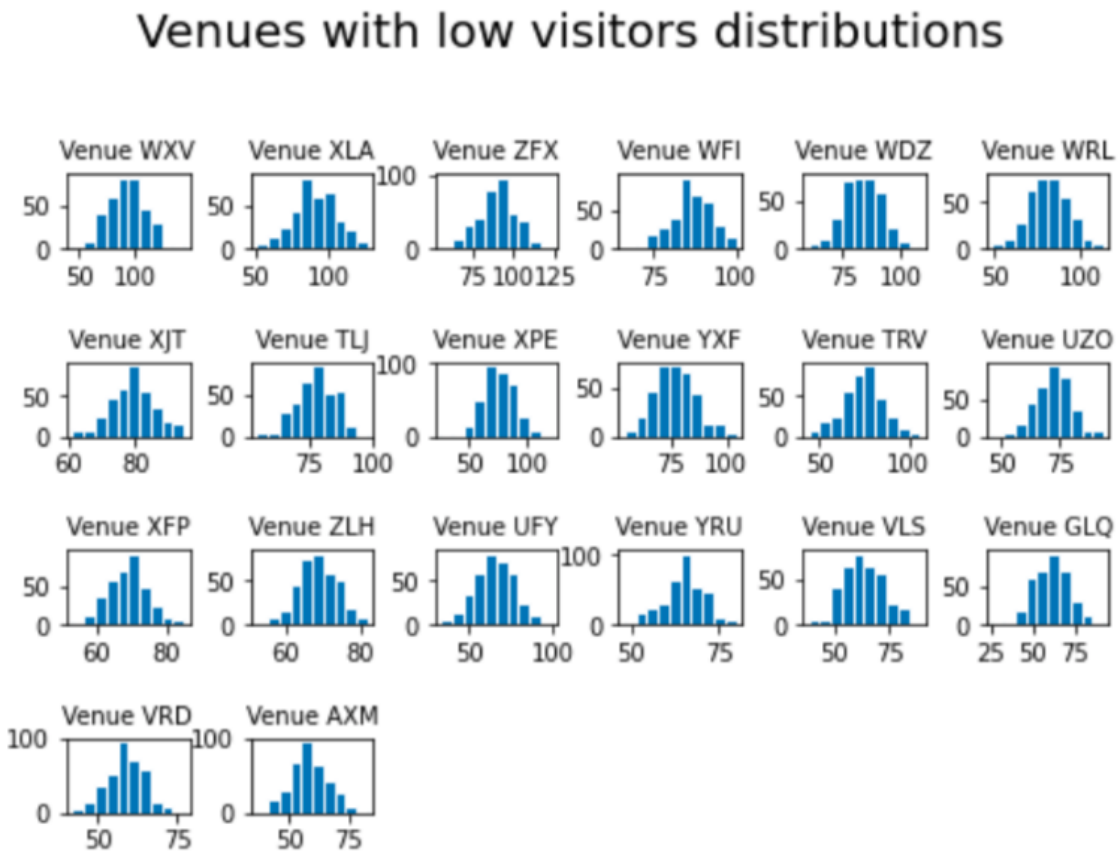

Figure 8:Histogram showing distribution of low volume visitors for venues.

In Figure 8, Histogram shows the distribution for all the venues with low volume visitors, and in almost all of them there is at least one empty column which suggests that there were no visitors whatsoever and also every graph has performed well at least once. For most of the venues the number of visitors increases thill the mid of the year and start falling after.

Q3. Critical Analysis

The module COMP1800 Data Visualisation introduced me to the core technologies and concept of data visualisation and enhanced my skills and prepared me for this field of my career. This programme taught me how to visualise a dataset and comprehend the various trends and patterns in the datasets. I have also learned how to use visual analytics with machine learning for data-intensive applications and which I will be using for my final project. Elaborating on my learning from this module, I can now identify and describe the basic elements of visualisation, determine which methods of visualisation are acceptable for distinct types of data, Create,test, and improve interactive visualisation systems and got an understanding of datasets, using different visualisation tools and approaches. Putting everything to work, I was able to complete my coursework efficiently and promptly. As the dataset was not that massive, I was able to get correlation of a few attributes and did a few visualisations to further provide evidence for the correlation like scatterplot and bubble graph. I also build a dashboard creating some interactive visualisations where one can always change the attributes and use it to get real-time use cases. While doing the data exploration, I verified the dataset for any missing values, duplicate attributes, and outliers. After completing the coursework successfully, I am confident enough to make a career in this field with an aim of further developing my skills.

Q4. Conclusion

After visualising the given datasets it has made clear how data visualisation can make everybody's life easier, be it a business or an individual like me. It made the dataset understandable by a lame man even if the person has no prior knowledge in this field. All the eight visualisations demonstrated in this coursework are very spot on to identify any trends or patterns in the dataset out of which two are interactive visualisation meaning the user can get real-time trends using those and own liking graphics. A human mind could only comprehend so much from a dataset that has numerical values in it but if the same dataset is visualised, it becomes more convenient to understand it. After segmenting the dataset into four parts i.e. High, Medium, Low, and Very Low I was able to identify the venues which have the highest and lowest visitors. The venues with high volume of visitors have more amenities and directly compliments all other factors hence have high visitors. The heatmap shows the strong positive correlation between some attributes meaning if one attribute increases, the other increases too. The scatterplot and radar graph only prove the correlation in the heatmap to be true giving the same results.