COMP1800 – DATA VISUALISATION

Report

Mahima Shrestha

001348194

MSc. Computer Science

Table of Contents

1.Introduction to Data Visualisation:

Data Visualisation refers to modifying plain data into meaningful ones with the help of graphical representations. Data Visualisation involves data exploration, explanation and presentation with the help of visual tools such as charts, graphs and plots. The earliest history of data visualisation can be traced in 1854, by Dr John Snow. Although first developed and used by statisticians and mathematicians data visualisation today has been integral part of any business as it helps gain important insights from what might be considered as mundane information. There are many libraries available in different programming language such as ggplot in R, Matplotlib/Pandas in Python, etc. that ease data visualisation processes. Data Visualisation primarily helps in channeling the insights, patterns and trends in the data that help business owners or any organizations to take better decisions. However, we must select the right visualisation option to show the information we want for instance, we could select Barchart to show the corresponding object's values but we cannot expect it show any correlations between the objects.

2.Discussion of the findings:

Here the raw data is segmented into two different data frames, known as weeklyvisitors, which keep a track of number of visitors visiting the cinemas and the summary dataframe which keeps track of all the other parameters that influence the cinemas. Below are 8 plots each followed by a justification of why the plot was chosen and description of the visualisation.

A.Visualisation No.1: Bar Plot

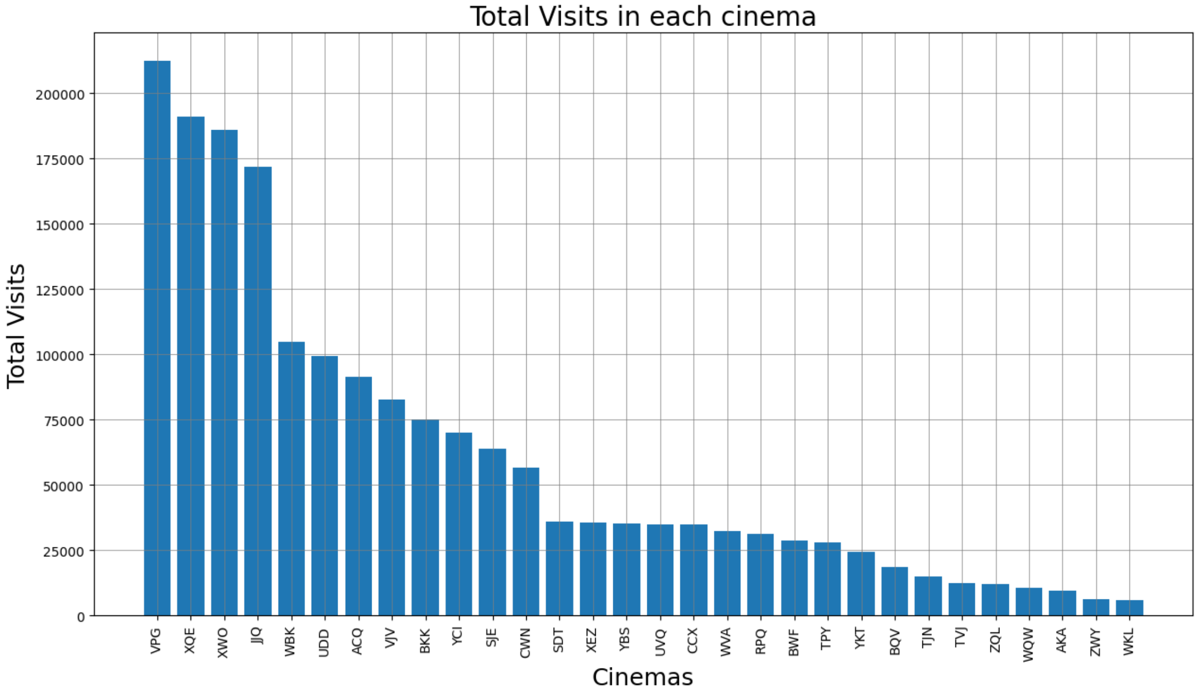


Figure 1 Bar Plot showing the total number of visits for each cinema in sorted order

Justification: Before deriving any insights, it is very important to have an overview of total visits in each cinema throughout the four years. Bar plot is one of the most convenient ways that help us visually grasp the number of visits to each cinema, point out the range of visits, and most importantly, compare each cinema based on the total number of visits. As we are required to segment cinemas into High, Medium and Low volumes based on the visits, with the help of a Bar plot, segregation of data can be made easily according to required categories.

Description: From the above image, we can clearly distinguish that there are four cinemas (VPG, XQE, XWO and JJQ) which have a staggering amount of number of visits. These can be labelled as High Volume Cinema. Similarly, there are eight cinemas whose visits lie in the range: (150k to 50k) which are labelled as medium-volume cinemas. Similarly, cinemas with visits from the range of 25k to 50k are labelled as low-volume cinemas. Additionally, cinemas with less than 25k are considered as Very Low Volume cinemas.

B. Visualisation No.2 : Line plot for very-low volume cinemas

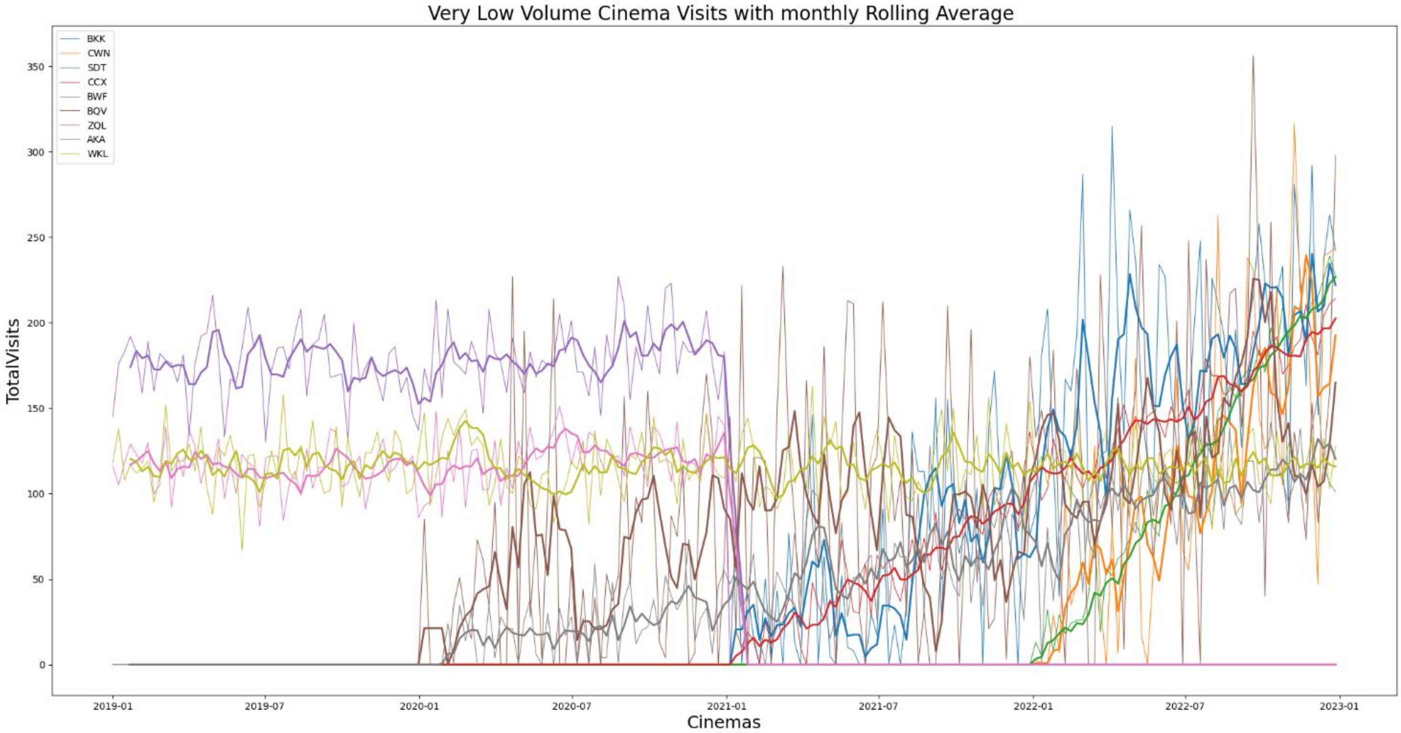


Figure 2 Line plot for very low volume cinemas with a monthly rolling average

Justification: An important task for us is to find out which cinemas were possibly opened or closed in the later stages of the four years. Line plot is a very effective way of showing this as it plots the data according to available time series. The data above has a huge range of fluctuations (i.e., each cinema above has witnessed a different range of cinema visits each week), therefore in this graph, rolling averages month-wise (with a period of four weeks) have been used to smooth out the noise, such that we can focus on the thick line which shows the monthly average for each cinema.

Description: From the figure we can jump to conclusion that cinema BWF and ZQL closed from starting of 2021. Cinema BQV opened in late 2019 whereas Cinema AKA opened in starting of 2020. Similarly, cinemas CCX and BKK opened from the starting of 2021 and CWN and SDT opened on the same time in late 2021.

C. Visualisation No. 3 Box plot

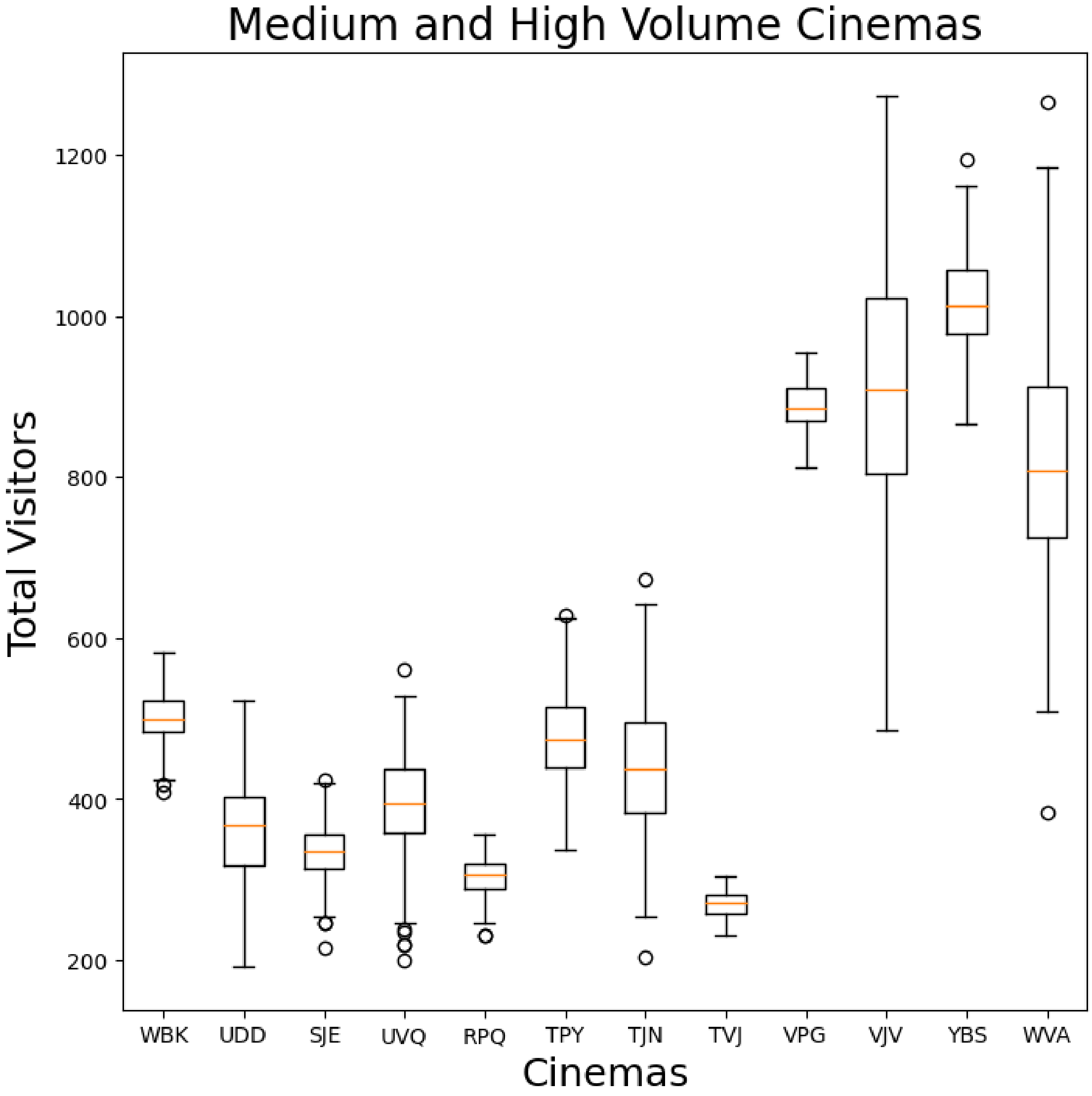


Figure 3 Box plot for medium and high volume cinemas

Justification: Boxplot helps to determine the range of data, indicating to what extent the spread exists for number of visitors visiting each cinema. It also helps us to determine the medians, quartile and any outliers that may exist for a given data.

Description: The boxplot above indicates that VJV, a high-volume cinema has the widest spread of data, indicating that this cinema witnesses a huge range of total customer numbers each week, whereas TVJ, a medium-volume cinema has a similar range of total customer numbers visiting each week over the period. Also, cinema WBK is positively skewed which means it usually has less numbers of customers visiting it apart from occasional spikes. Most of the medium-volume cinemas have extreme variations, known as outliers. However, YBS and WVA (the high volumes) each have only one outlier.

D. Visualisation No.4: Line plot with trendlines

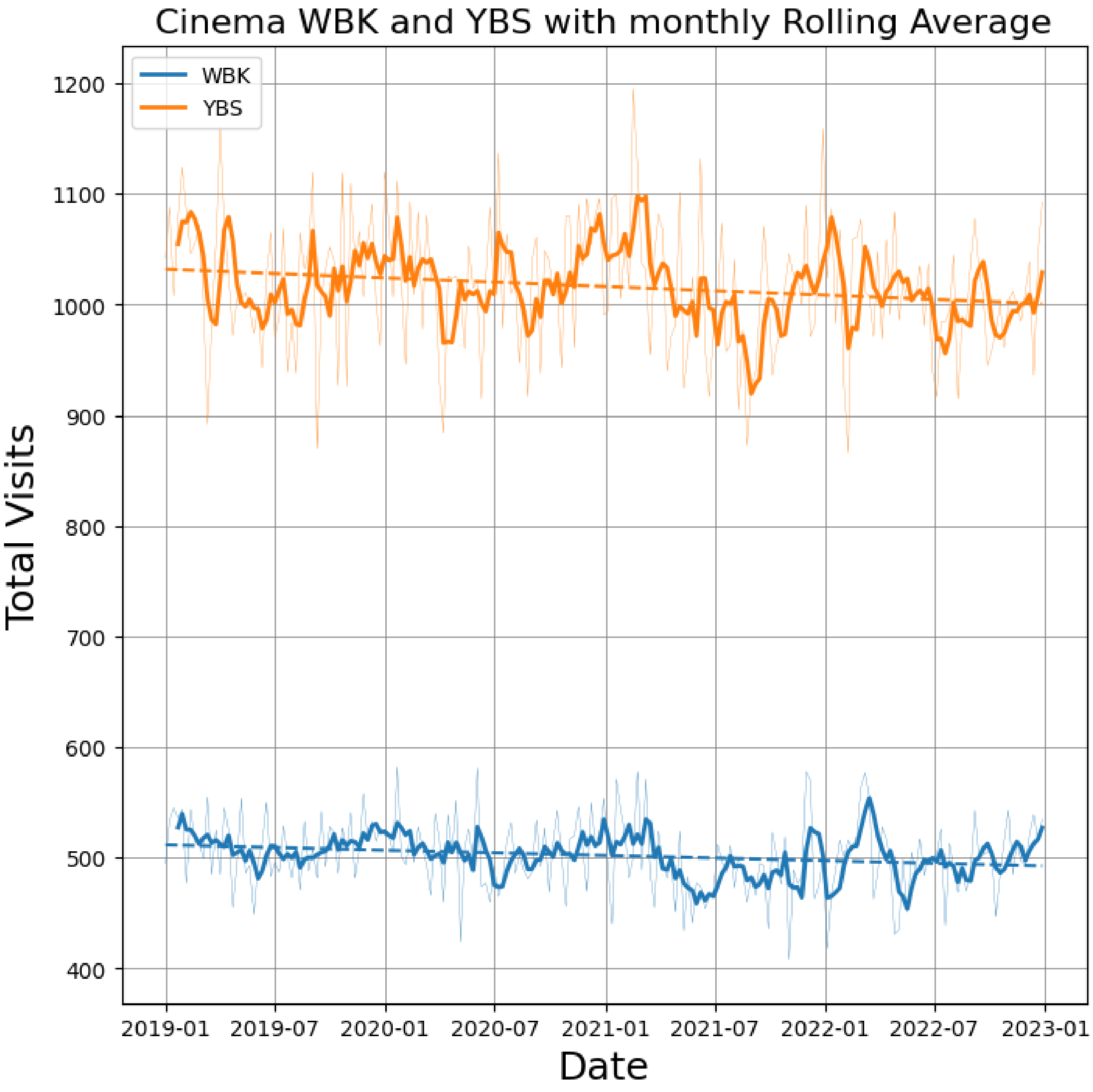


Figure 4 Line plot with trendline for Cinemas WBK and YBS

Justification: I have used Line plot again with trend line for one medium volume cinema WBK and one high volume cinema YBS as it helps us to visualize what is the overall trend for these cinemas. It also helps to compare any similarity in seasonal behaviours of the two cinemas over time.

Description: WBK is a medium-volume cinema whereas YBS is a high-volume cinema, so focusing on these two cinemas, we can point out that although they have vast differences in terms of total visits every month, both have fairly similar trends. For instance, the trend line (overall cinema visits) for both cinemas is slightly decreasing, however, the total number of visitors at the start and end of the period is the same for both cinemas. They both have witnessed a sudden spike in visitors during the first quarter of the year 2022 and 2021. Similarly, during mid-January of 2020, both these cinemas have a similar increase of visitors.

E. Visualisation 5: Autocorrelation for seasonality

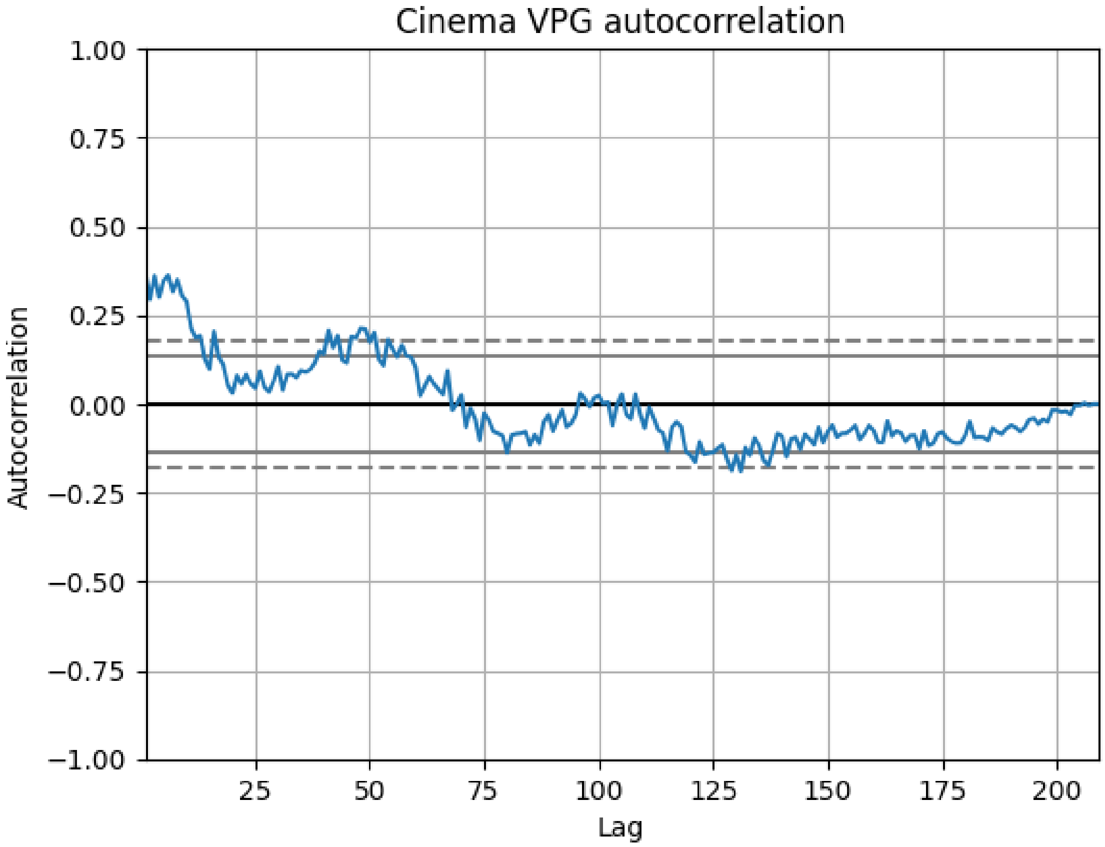


Figure 5 Autocorrelation plot for cinema with significant seasonality

Justification: In an attempt to find some insights regarding seasonality in cinemas, the autocorrelation plot was used. We can derive if there is a pattern in terms of cinema goers over the period. The gist here is to focus on the representation outside the dashed line for a strong statical significance.

Description: Out of all the cinemas, the high-volume cinema VPG only has some seasonality during the early phases of the period recorded. Here, we can see the secondary peak is at 49 implying that the pattern repeats after 49 weeks. It means that there is a significant increase in the total number of visitors for the cinema after a gap of 49 weeks.

F. Visualisation No. 6. Interactive heat map.

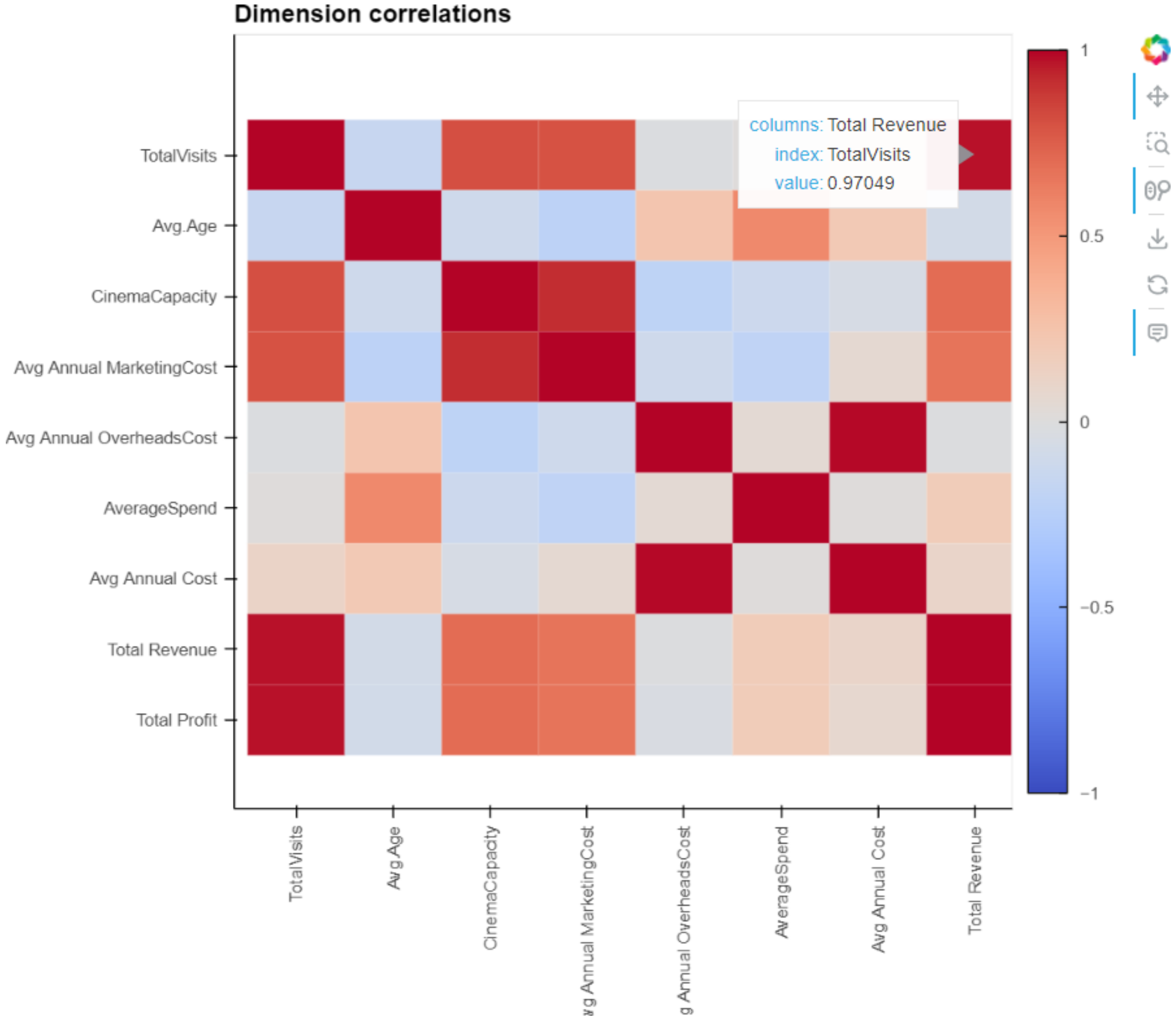


Figure 6 Interactive Heatmap to show correlation between dimensions of summary data

Justification: A heatmap is used to show the correlation between various parameters of summary data. Correlation refers to the fluctuation of two or more variables with respect to each other. In the above diagram, dark red colours represent strong positive correlations (close to value +1) whereas dark blue colours represent strong negative correlations (close to value -1). The heat map is a 9X9 boxes, each dimension being checked across all the other dimensions.

Clarification: The total number of visitors to each cinema has a strong positive correlation with Cinema capacity and Marketing Cost, it is obvious that to have more cinema capacity, more amount of money needs to be spent on marketing. However more cinema capacity leads to more visitors which in turn leads to more revenue and more profit.

G. Visualisation No. 7: Radar plot

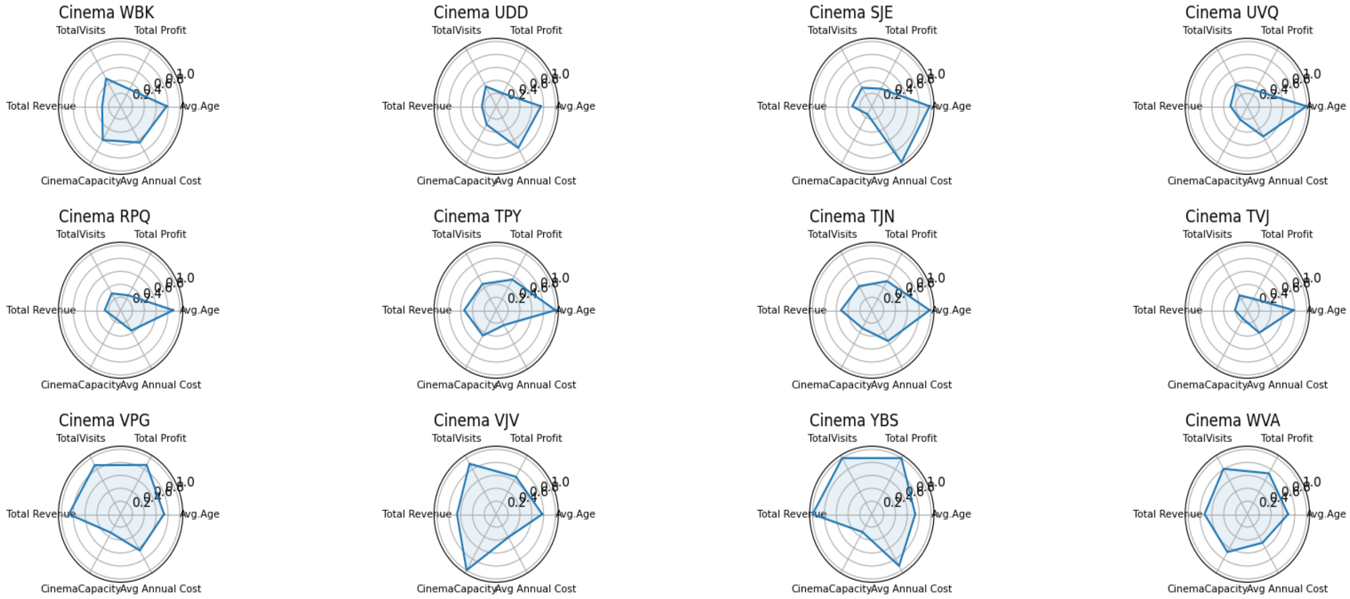
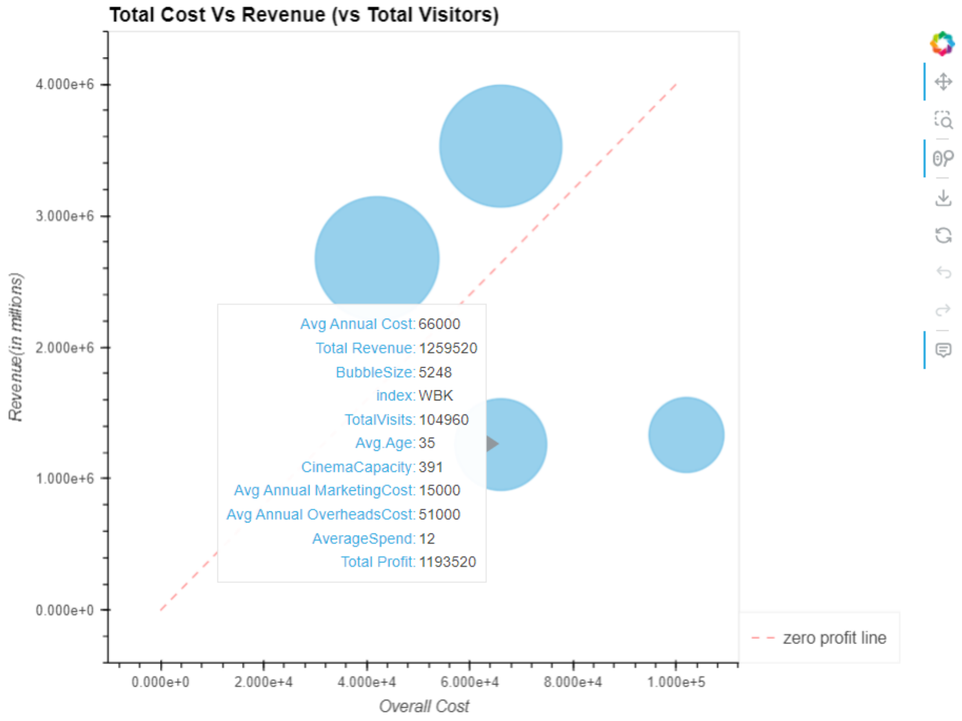


Figure 7 Radar plot for High and Medium Volume Cinemas

Justification: To compare each Cinema based on various parameters such as Cinema Capacity, Total Revenue, Total Profit, Total Cost and Total Visits radar plot is chosen. Here, radar plot helps to understand how each cinema is performing based on these parameters, additionally, we can also do a comparative analysis between various high and medium-volume cinemas since they are our major focus areas.

Description: Amongst the six different parameters; Total Profit, Total Visits, Total Revenue, Cinema Capacity, Average Age and Avg Annual Cost, age is considered a neutral measure. Here, Total Revenue, TotalVisits and Total Profit is considered as a positive measure as the company wants more of them whereas Cinema Capacity and Avg Annual Cost are considered negative measures as the company wants to cut them down (considering more cinema capacity requires more marketing and overhead cost). So, cinemas with the most volume on the top and least on the bottom are the company's most desired ones. Here, cinema VPG and VJV are one of the best performing cinemas whereas cinemas like SJE and WBK yield more expenditure as it has more volume on the bottom and less on top.

H. Visualisation No. 8: Interactive bubble plot



Avg Annual Cost: 66000
Total Revenue: 1259520
BubbleSize: 5248
index: WBK
TotalVisits: 104960
Avg.Age: 35
CinemaCapacity: 391
Avg Annual MarketingCost: 15000
Avg Annual OverheadsCost: 51000
AverageSpend: 12
Total Profit: 1193520

Figure 8 Interactive bubble plot for two high volume and two medium volume cinemas

Justification: This interactive bubble plot of revenue vs overall cost vs total customer visitors is chosen to justify the conclusions from above radar plot, here we can compare how these cinemas WBK and SJE (medium volume) yield less profit to the company as compared to VJV and VPG (high volume cinemas). Additionally, the size of bubble represents the total number of visitors in each cinema. The bubbles represent number of cinema visitors.

Description: Here, a red dashed line is used to segregate the graph into two halves, the upper half represent area where Revenue is more than overall cost. Whereas, the bottom half of the graph represents Cost more than Revenue. Clearly, VPG is one of the best performing cinemas which gave more revenue in limited cost while WBK yielded much less revenue for similar overall cost. Out of these four, SJE can be considered least performing as it has huge overall cost with less revenue generated and less cinema visitors as compared to WBK.

3. Critical Review

This module was a significantly organized module as each week we had the opportunity to learn a different visualization technique in a structured way. In this module, understanding of visual representations was made and the impact it can make on business scenarios was discussed. We learned to discover patterns, trends, variations, and correlations with the help of visualizations that may exist in data that normally could have been ignored or overlooked. These methods were applied in the coursework, starting off by creating dataframes, segmenting cinemas based on visitor volumes and applying necessary visualisation like line plot with trend lines to analyse the overall performance of cinemas with respect to time. Box plot was used to determine the spread of the data and identify outliers. Through heatmap, the correlation between different factors such as Cinema Capacity, Marketing Costs, and Revenue generated was justified. This helped determine how it has overall impact on the profitability of each cinema. This was further clarified with the help of radar plot and bubble plot which gave descriptive and clear information along with a comparative analysis of each cinema. While the report consists of only 8 visualizations, the Google Colab file has proof of rigorous testing and exploration of the data by applying different visualization techniques on each data frame to find out any seasonality, correlations, trends, patterns, outliers etc., as demanded by our coursework. Additionally, use of interactive visualisation techniques were used that enhanced simple pictorial visualisations.

4. Conclusions

Throughout this process of visualisation, the two separate data frames *weeklyvisitors* and *summary_df* containing different information were created which gave different insights on the data. WeeklyVisitors dataframe: This dataframe contained the information about total visitors visiting each cinema over the period of time.

- The highest volume cinemas are: VPG, VJV, YBS and WVA whereas the medium volume cinemas are WBK, UDD,SJE, UVQ, RPQ, TPY, TJN and TVJ.
- The very low volume group contained cinemas that were either opened or closed in the later stages of the recorded time.
- The high-volume cinema VJV had the most spread of data which indicates on some days, huge number of people come to the cinema whereas on some days it witnessed very less number of visitors. Medium-volume cinemas had most outliers.
- Out of all the cinemas, only VPG showed some level of seasonality in the first year of the recorded time.
- Strangely, there is similarity in trend of a medium volume cinema WBK and high volume cinema YBS. They both showed familiar peaks and troughs in and around same points of time.

The summary_df: This dataframe contained the other dimensions like TotalCapacity, Marketing Cost, Spending, Revenue etc.

- The Total number of visitors had strong dependencies on both Cinema Capacity and Marketing Cost, specifically, the higher the cinema capacity, the higher the marketing/overhead cost required which however had a notable impact on increasing the revenue.
- High volume cinemas such as VJV, VPJ proved to be more profitable than medium-volume cinemas as they were able to generate more profit on limited costs whereas medium-volume cinemas such as WBK and SJE had comparatively less revenue but more cost.