**COMP1800 DATA VISUALIZATION**

**Richard Raja**

**001370307**

**MSc Data Science**

Contents

[Introduction 3](#_Toc170868300)

[Visualizations 4](#_Toc170868301)

[1. Bar chart 4](#_Toc170868302)

[Description 4](#_Toc170868303)

[Justification 4](#_Toc170868304)

[2. Box plot 5](#_Toc170868305)

[Description 5](#_Toc170868306)

[Justification 6](#_Toc170868307)

[3. Rolling average 6](#_Toc170868308)

[Description 7](#_Toc170868309)

[Justification 7](#_Toc170868310)

[4. Autocorrelation 8](#_Toc170868311)

[Description 8](#_Toc170868312)

[Justification 8](#_Toc170868313)

[5. Histplot 9](#_Toc170868314)

[Description 9](#_Toc170868315)

[Justification 9](#_Toc170868316)

[6. Interactive heatmap 10](#_Toc170868317)

[Description 10](#_Toc170868318)

[Justification 11](#_Toc170868319)

[7. Heat Map 11](#_Toc170868320)

[Justification 12](#_Toc170868321)

[Description 12](#_Toc170868322)

[8. Interactive line plot 13](#_Toc170868323)

[Description 13](#_Toc170868324)

[Justification 13](#_Toc170868325)

[Critical Review 15](#_Toc170868326)

[Conclusion 15](#_Toc170868327)

# Introduction

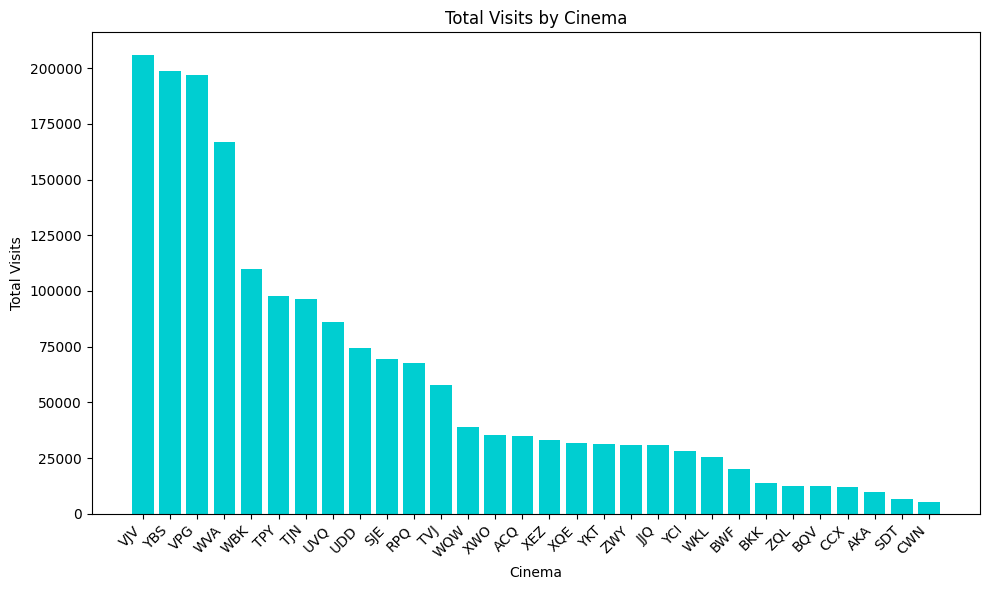
Data visualization serves as a bridge between raw data and meaningful insights. As humans, we process visual information more efficiently than rows of numbers or textual descriptions. Visualizations enable us to discover patterns. Whether it's identifying seasonal trends in sales data or spotting anomalies in medical records, visualizations help us recognize patterns that might otherwise remain hidden.

There's a wide array of visualization techniques, each suited to two different data types and objectives. For example, line charts are ideal for showing trends over time. Line charts connect data points headlines. They are commonly used for stock market analysis weather patterns call mom and tracking metrics like website traffic etc. Other types of visualizations are scatter plots, neat maps and bar charts.

Crisco cinema is a cinema chain in the UK. The stakeholders are more interested in high volume and moderate volume cinemas. We must carry out our exploration on the provided data with the help of data visualization techniques.

# Visualizations

## Bar chart



*Figure 1*

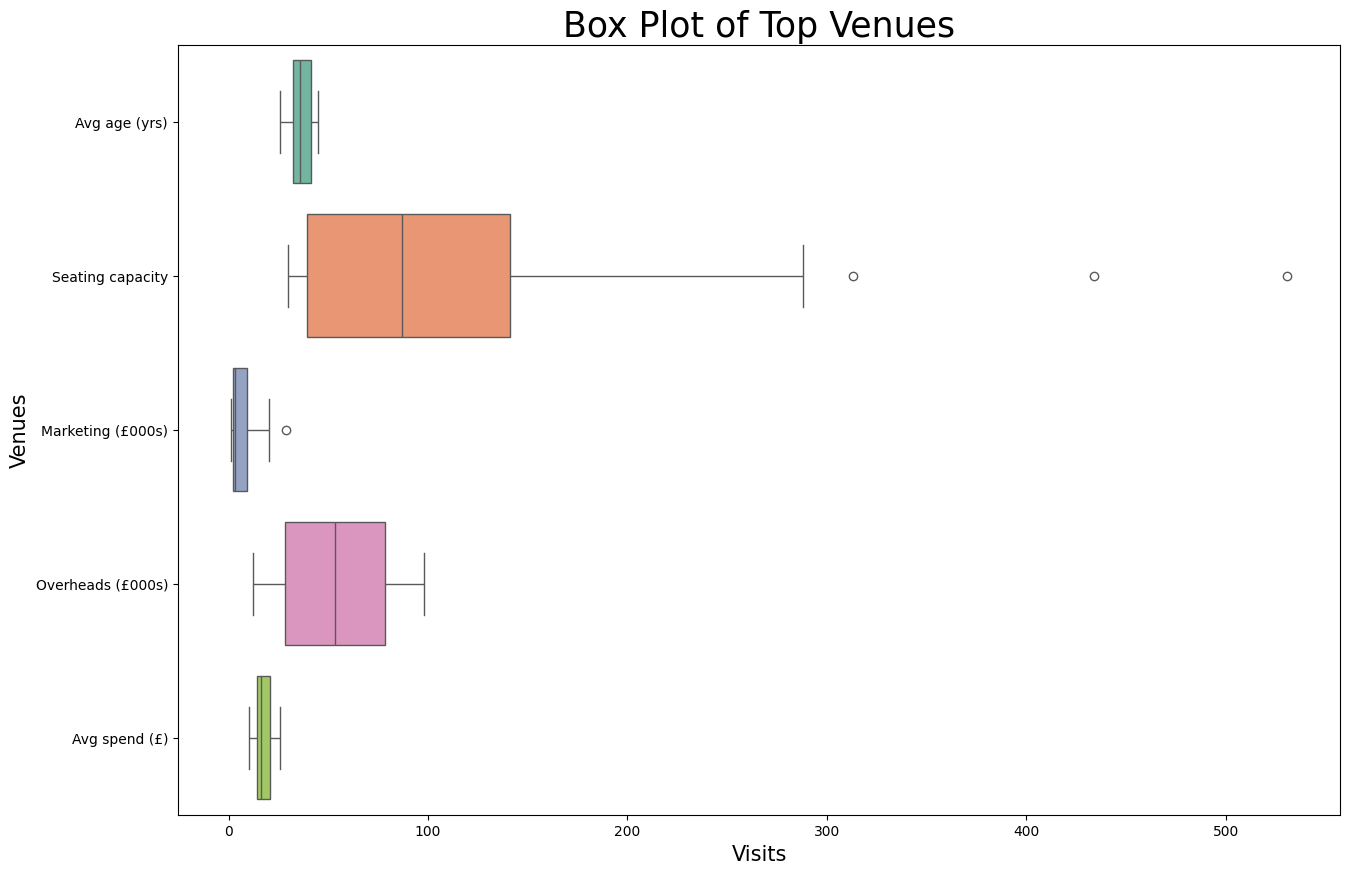
## Description

Bar charts are a fundamental type of data visualization that display categorial data using rectangular bars. Each bar represents a category such as product names, time periods, or group. Bar charts are commonly used for comparisons. You can easily compare values across different categories, for example, visualize sales performance for various products or compare average temperatures in different cities.

## Justification

As the stakeholders are more interested in high and moderate volume cinemas, I decided to represent the total visit count of each venue in the form of a bar chart. We can clearly distinguish the three categories (high moderate and low volume) based on height of the bar. We can clearly observe that 4 venues, with a count of more than 150 thousand customers are in the high-volume category, while threshold from moderate level cinemas starts from about 11,000 to about 6000 visits.

## Box plot



*Figure 2*

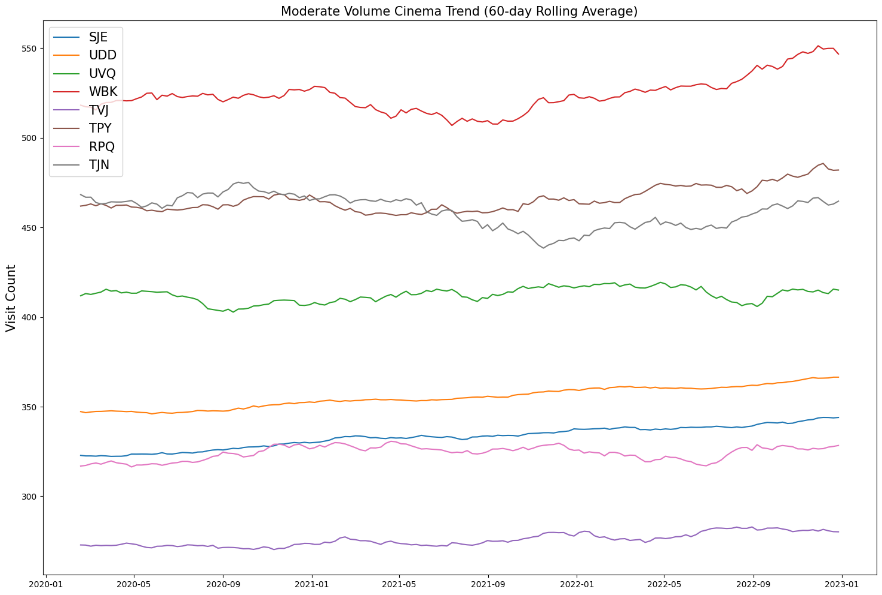
## Description

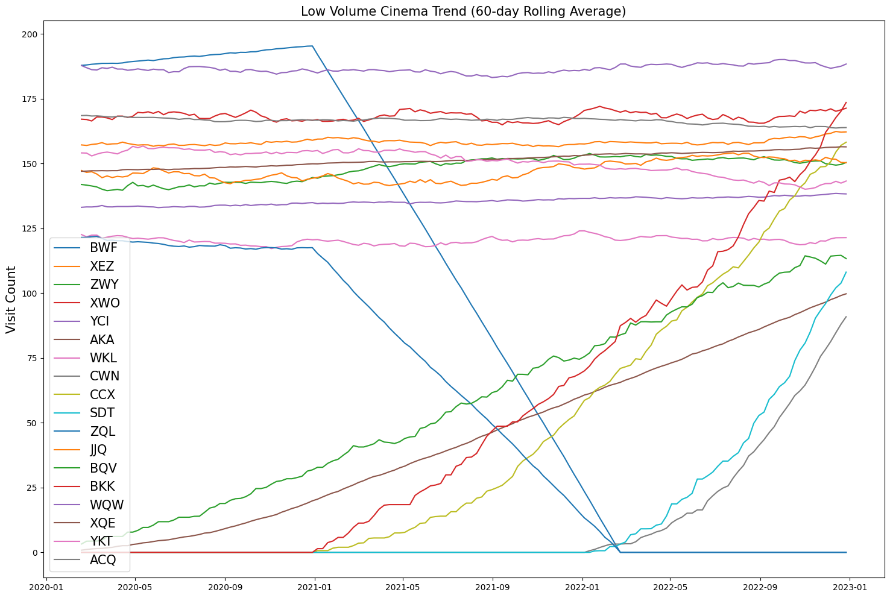
Box plot also known as box and whisker plot, is a powerful visual tool used to represent the distribution of a data set. It provides a concise summary of key statistics, including the median, what else, and potential outliers. The box represents the middle 50% of the data, which spans from the first quartile to the third quartile. The line inside the box represents the median, which marks the midpoint of the data.

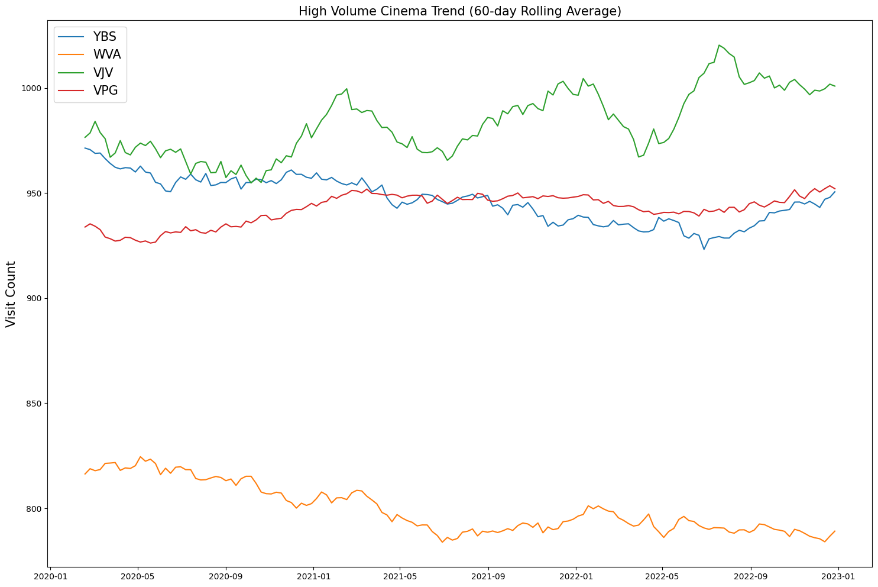
## Justification

I added a box block to understand the ranges of values for each attribute. As we can see, the seating capacity attribute has the widest range. The hollow circles represent the outliers. Although the attributes are not correlated with each other, this type of chart was essential for representing the ranges of those attributes.

## Rolling average







*Figure 3*

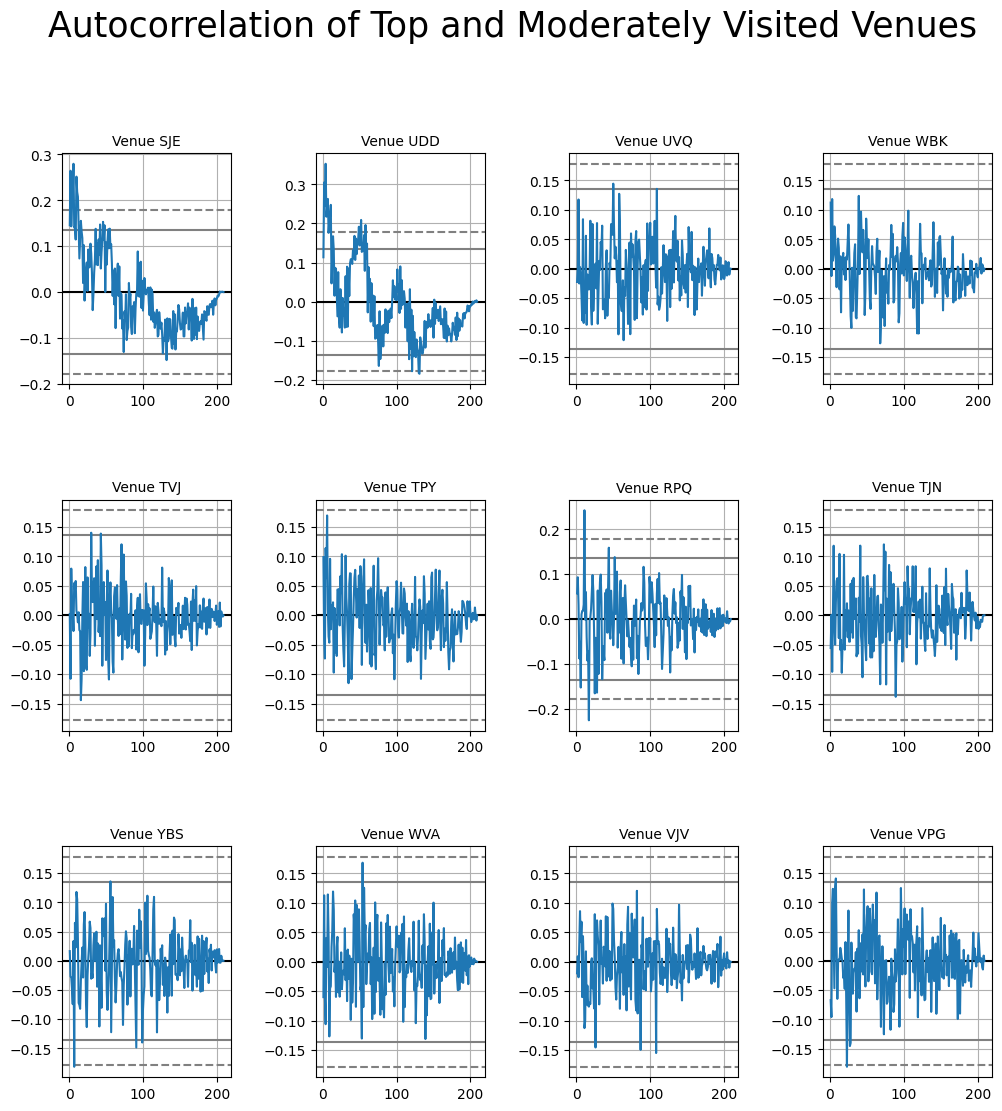
## Description

A rolling average, also known as a moving average, is a statistical method used to analyze and smooth out data by creating a series of averages from different subsets of the full data set. The calculation involves taking the average of a fixed subset of data points and then incrementally shifting the subset forward, including the next data point and excluding the first, to calculate the next average. This process continues, providing a smooth data curve that makes it easier to observe trends and cycles.

## Justification

With the help of the chart, we can understand the trend of each cinema in their respective categories. For example, in the high-volume cinema category, the venue VJV shows the highest fluctuation. We also observe that in this venue, the ninth month brings the least customer count. The moderate category does not provide any meaningful insights, however in the low category we observe some the key points. The sharp incline in some venues shows that these venues were opened in the given data frame, meaning no data was provided for the earlier part of the time frame. On the other hand, the sharp decline shows the venue was shut down during the time frame.

## Autocorrelation



*Figure 4*

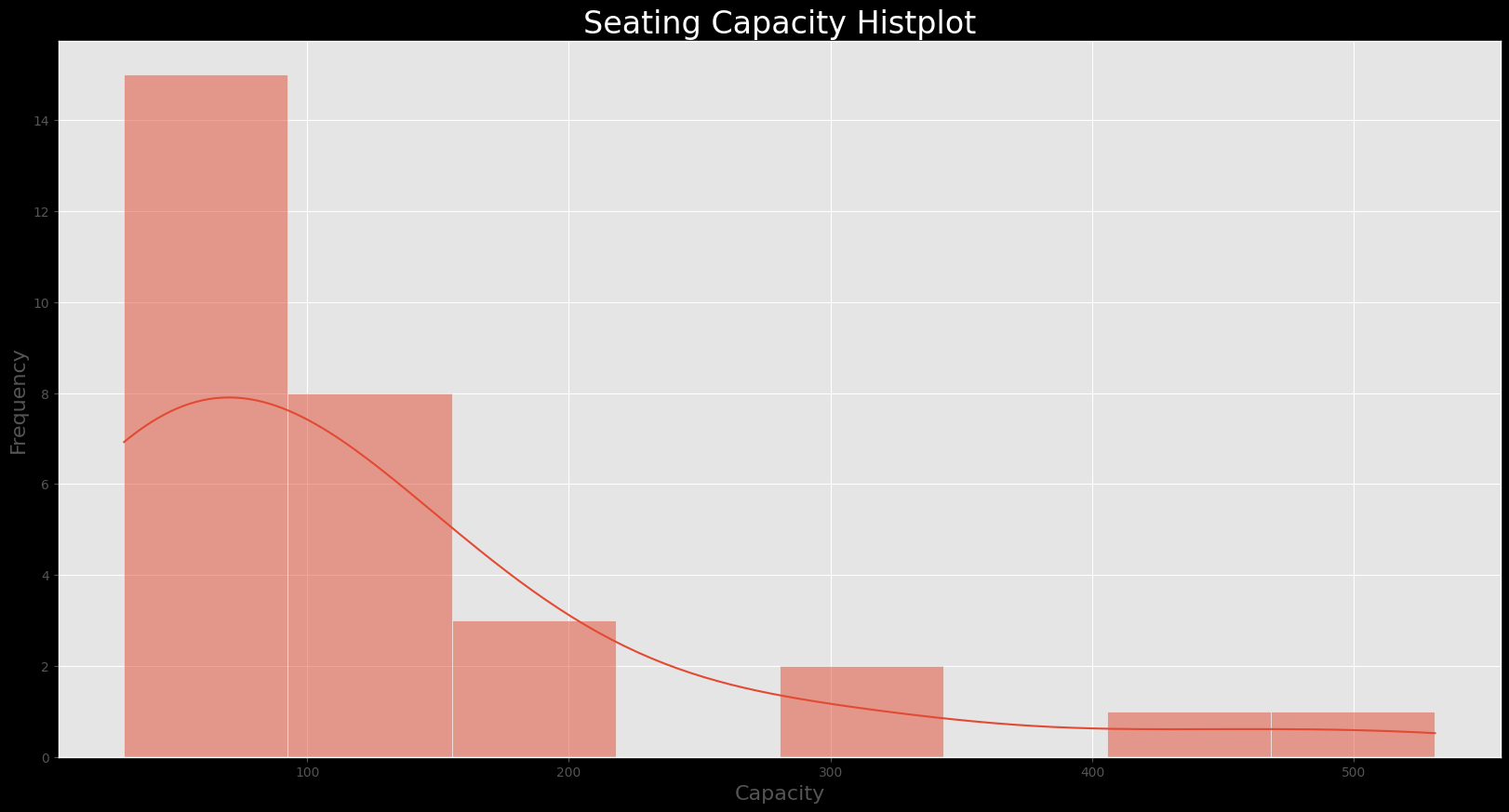
## Description

Autocorrelation measures the degree of similarity between a time series and its lag versions. In other words, it quantifies how correlated a data point is with its previous values at different time intervals. A high correlation at a specific lag suggests repeating pattern or seasonality while low autocorrelation indicates randomness. Autocorrelation plots have identified friends, periodicity, and potential lag effects in time dependent data.

## Justification

By looking at the chart, we can observe that after 50 days the pattern repeats itself. This observation is primarily for the venue SJE and the venue UDD. The venues are also quite different from 120 days earlier. Other venues do not provide us any meaningful insight.

## Histplot



*Figure 5*

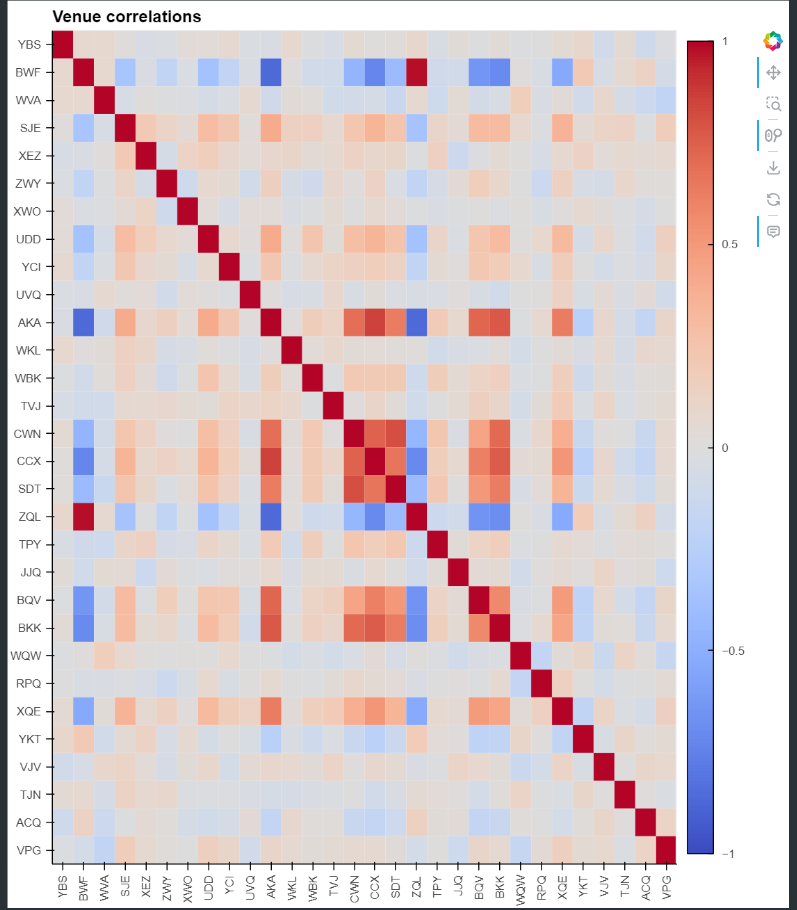
## Description

Histplot function in Seaborn library creates histograms, which are graphical representation of data distributions. A histogram groups numbers into ranges, and dollar bars indicate that more data falls within those ranges. It displays the shape and spread of continuous sample data. With his plot, you can customize bin sizes, colors, and even stack histograms.

## Justification

As we saw earlier, the seating capacity attribute had the widest range of numbers. To further explore this, I used histogram to understand the categories within that range. As we can observe, the seating capacity of less than one hundred is the most common.

## Interactive heatmap



*Figure 6*

## Description

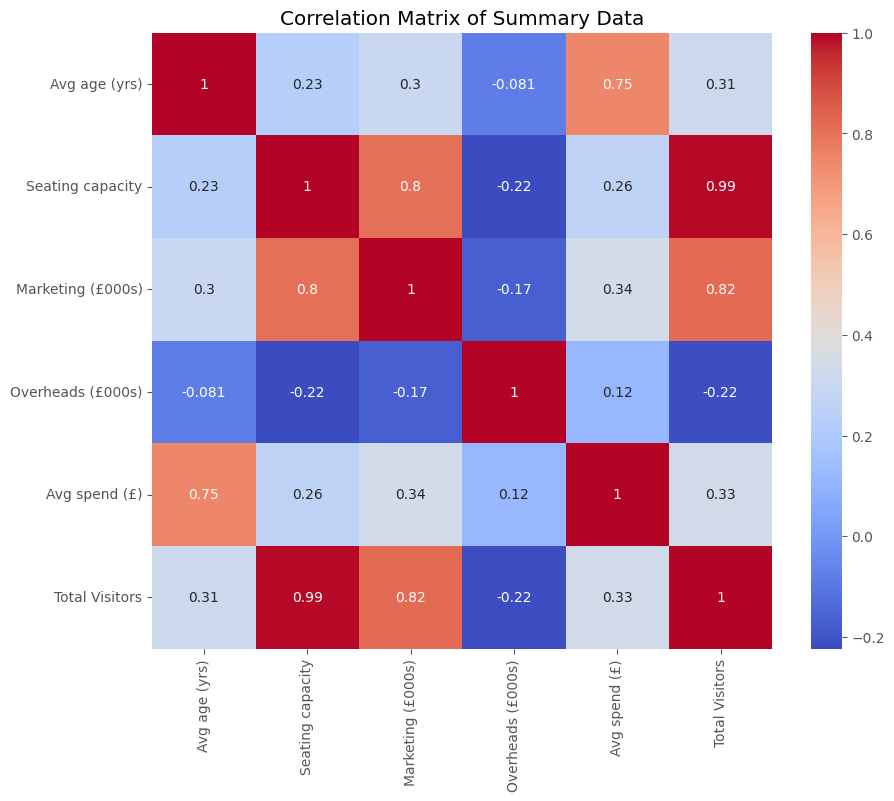
A heat map is a graphical representation of data where individual values are represented as colors within a matrix. It is commonly used to visualize the distribution and relationships between two variables. Heat maps are particularly useful for displaying patterns, correlations, and variations in data, such as identifying hotspots or clusters.

## Justification

I wanted to explore the correlation between venues from all over the categories, so I used a heat map. The lighter color represents no correlation while the darker color, red, represents a stronger correlation. This visualization is particularly important for further exploration of the metrics that make these venues correlate.

The chart is interactive, which means when the user hovers over a cell, he/ she can get the precise information.

## Heat Map



*Figure 7*

## Justification

Heatmaps are powerful data visualization tools that use color gradients to represent values within a matrix or dataset. They offer an intuitive way to explore complex data, making patterns and relationships more apparent.

Each cell in a heatmap corresponds to a specific value, and the color intensity reflects the magnitude of that value. By visually encoding information, heatmaps allow us to identify trends, correlations, and outliers efficiently.

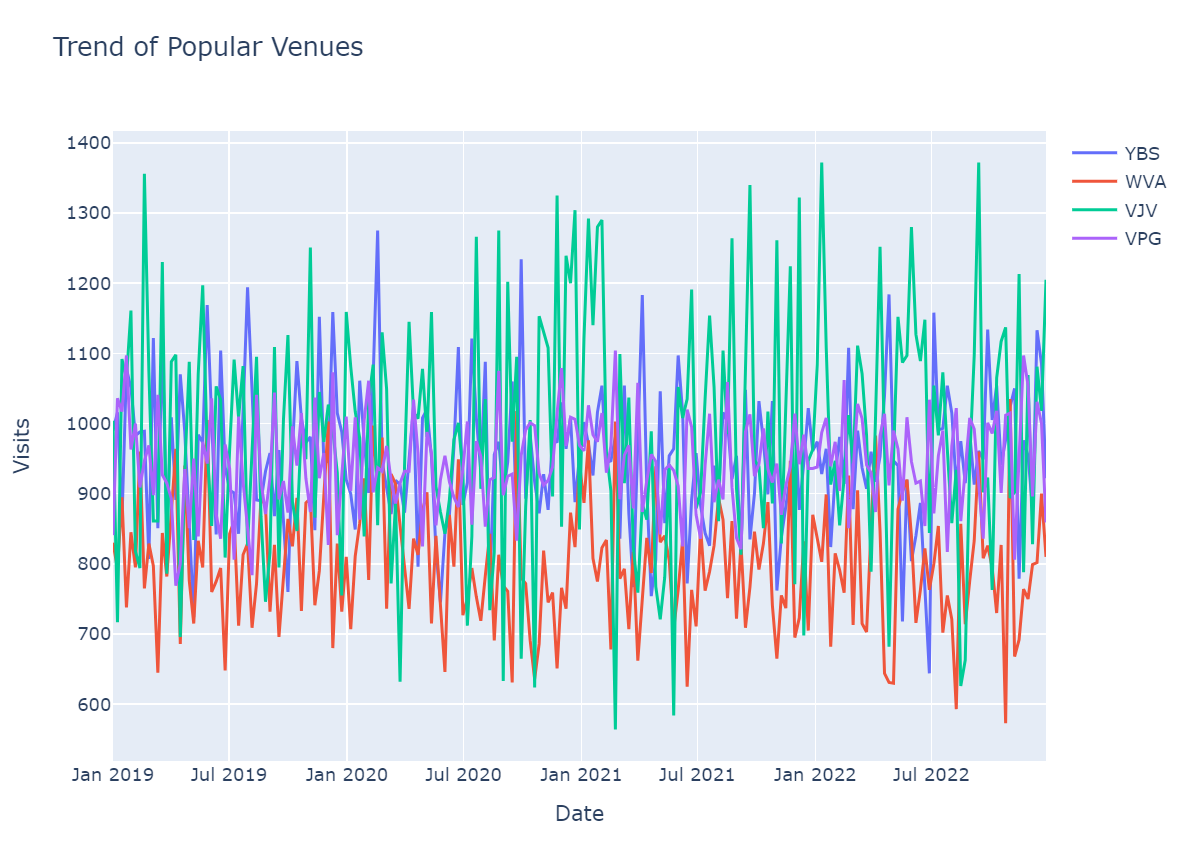
Their versatility makes them valuable across various domains, including biology, finance, and machine learning.

## Description

I started by computing a new column in the cinema data dataframe, named "total visitors," which resulted from multiplying the seating capacity by the average age specific to each cinema. This provided an estimate of the total visitors based on these two factors. Following this, I utilized the seaborn library to generate a heatmap. The correlation coefficients were displayed in the heatmap cells, and I utilized a warm color scheme to enhance clarity.

The cells shaded in dark red indicate the strongest correlations. Notably, we observe a strong correlation between seating capacity and visitor numbers, suggesting that customers often prefer more spacious venues. Additionally, seating capacity demonstrates the least correlation with average spending at the venue.

## Interactive line plot



*Figure 8*

## Description

Line plot is a type of graphical representation used to display data trends over time or across ordered categories. It consists of points connected by straight lines, where each point represents a data value corresponding to a specific time or category. Line plots are effective for illustrating changes in data continuously or sequentially, making them ideal for showing trends, patterns, and relationships within the data set. They are particularly useful for highlighting trends over time, such as increasing or decreasing trends, cyclic patterns, or fluctuations.

## Justification

To get the exact count of customers for a specific week, the user can hover over the line, above the time frame and extract useful information. This chart also allows users to quickly see the high/ low points.

## Critical Review

This course taught me how to transform raw data into meaningful insights. I was able to strengthen my skills in Python and learn useful libraries such as pandas, matplotlib, seaborn etc. I was also able to learn techniques for data cleaning, data manipulation and data visualization in this course. Data visualization is a very in demand skill. Today's world highly depends upon data. As the world of AI is progressing, we must learn to take advantage of the abundant data.

ChrisCo company has venues all across the country. The study conducted for Crisco, a company operating numerous cinemas in the UK, provides valuable insights into customer behavior and marketing strategies. Although it offers significant advantages, the study also identifies areas where improvements can be made. The depth of this analysis enhances the value of the visualization, yet there remains potential for further exploration to uncover specific findings that could enhance the company's marketing initiatives.

## Conclusion

Data visualization was implemented to study the data set of Crisco. I was able to convey complex insights to beginners with the help of visually appealing charts. Some key data points were extracted, such as there is high fluctuation in visit counts of high volume cinemas, shutting down of cinemas with low volume and also establishment of new cinemas. With the data provided, basic insights were noted, but for full length report on the cinema chain, more data such as ticket price, daily visit count, should be provided.