COMP1800 DATA Visualization

Coursework Specification: We are to carry out a visual data exploration for ChrisCo, the fictional company whose sales and website data we have been analysing throughout the module, using a Python Notebook (in Colab or Jupyter). ChrisCo is a fictional, but nonetheless very successful company managing a chain of cinemas across the UK.

ChrisCo collects a huge amount of data about individual customers visiting its cinemas using its loyalty card scheme and this customer data has been aggregated/averaged over a four-year period to give information about the company's cinemas, each identified by a unique 3 letter code (e.g. ABC, XYZ, etc).

Report:

1. Introduction to Data Visualization

Data visualization is a powerful tool used to transform raw data into meaningful insights and stories. According to Andy Kirk, author of "Data Visualization: A Handbook for Data Driven Design," visualizations enable us to "see the unseen" by presenting complex information in a clear and intuitive manner (Kirk, 2012). Through visual representations such as charts, graphs, and maps, patterns, trends, and relationships within data become accessible to a wide audience, regardless of their level of technical expertise. By visually exploring data, patterns, trends, and relationships that may not be immediately apparent in raw numbers can be discovered. For example, trends over time, correlations between variables, or geographic distributions can be easily identified through effective visualizations.

In this report, we will leverage the power of data visualization to explore and analyze the visitor data of ChrisCo cinemas (a fictional company). Through a series of visualizations, we aim to uncover patterns, trends, and anomalies in the data, providing valuable insights for the company's strategic planning and decision-making process.

2. Discussion of Findings

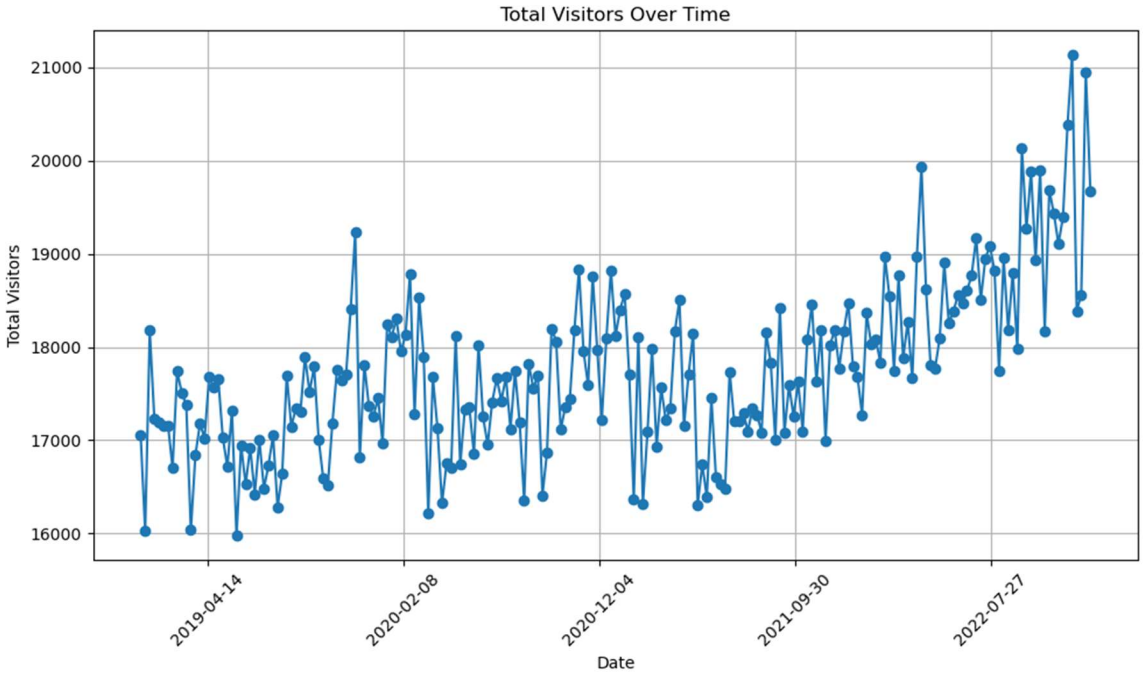
In this section, we will explore various aspects of ChrisCo's cinema data through visualizations. Each visualization will be accompanied by a brief introduction explaining the type of visualization used, followed by the code to generate it in the Jupyter Notebook. Finally, we will provide a justification for why we selected this particular visualization for analysis.

2.1. Total Visitors Over Time **Fig.1**:

For our first visualization, we will create a line plot showing the total number of visitors to ChrisCo cinemas over the four-year period. This visualization will help us understand the overall trend in visitor numbers and identify any significant fluctuations or patterns over time.

Justification: We chose this visualization to provide an overview of the temporal trend in visitor numbers, which is crucial for understanding the overall performance of ChrisCo cinemas. By plotting the total number of visitors over time, we can identify any seasonal patterns, long-term trends, or anomalies in customer footfall. This information will help us make informed decisions regarding operational planning, marketing strategies, and resource allocation.

Fig.1



Description of Fig. 1: The data reveals that there are regular fluctuations in the total number of visitors to ChrisCo cinemas over time. Most of the time, the number of visitors stays around a certain level, with small ups and downs. However, there are three noticeable spikes in visitor numbers:

Around at the end of 2019: There is a sudden increase in the number of visitors, reaching around 19k+. This spike could be due to a special event, a blockbuster movie release, or a marketing campaign that attracted more people to the cinemas during that period.

Around at the start of 2022: Another significant increase in visitor numbers occurs, peaking at around 20k. Similar to the previous spike, this could be attributed to factors such as the release of highly anticipated movies or seasonal trends in cinema attendance.

Towards the end of the data period: There is a final surge in visitor numbers, reaching approximately 21k. This could indicate a successful promotional campaign, new attractions in the cinemas, or other external factors driving increased attendance.

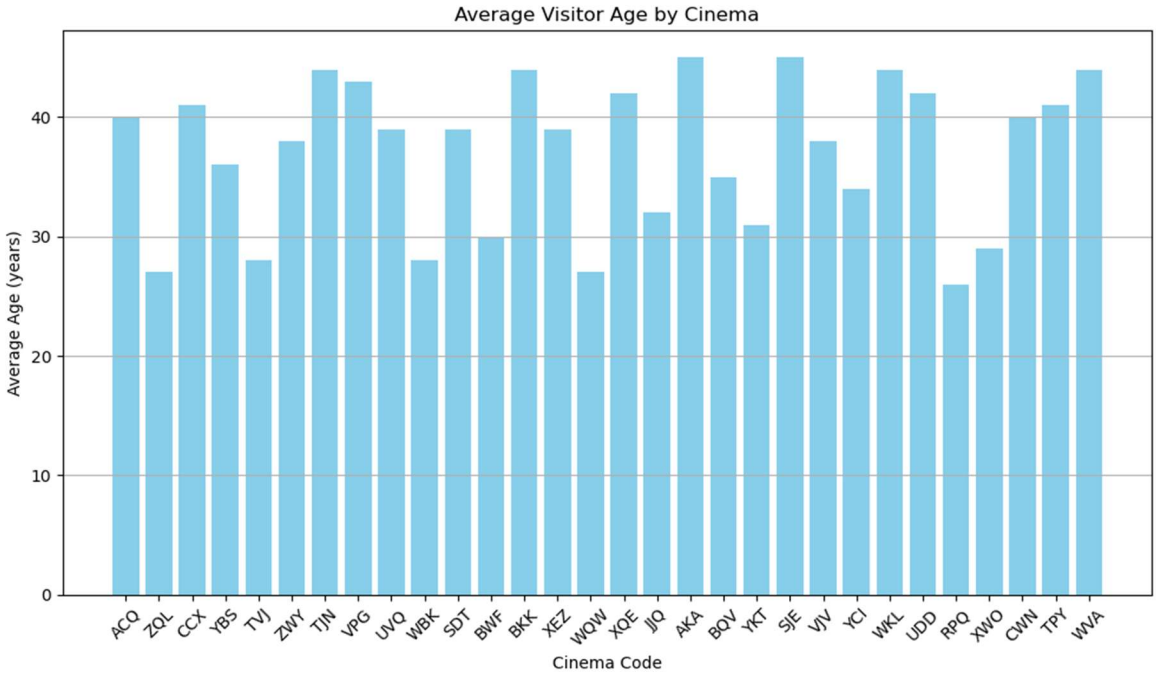
Overall, the data suggests that while visitor numbers generally fluctuate within a certain range, there are notable peaks in attendance during specific periods. Understanding the reasons behind these spikes can help ChrisCo cinemas optimize their marketing strategies and capitalize on opportunities to attract more visitors.

2.2. Comparative Bar Chart of Average Visitor Age **Fig. 2**:

In this visualization, we will create a comparative bar chart to compare the average age of visitors across different cinemas. A comparative bar chart allows us to easily compare the average age values for each cinema, providing insights into the demographic characteristics of visitors at each location.

Justification: We chose a comparative bar chart to visualize the average age of visitors across different cinemas because it allows for easy comparison of average age values for each cinema. By examining the heights of the bars, we can quickly identify which cinemas have older or younger audiences compared to others. This information can be valuable for tailoring marketing strategies, programming content, and amenities to better suit the demographic profiles of each cinema's audience.

Fig.2



Description of Fig. 2: The comparative bar chart illustrates the average visitor age for each cinema in ChrisCo's chain. Each bar represents a cinema, with the height indicating the average age of its visitors. The chart reveals variations in audience demographics across cinemas, with some cinemas catering to younger audiences (e.g., ZQL with an average age of 27) while others attract older patrons (e.g., TJN with an average age of 44). Understanding these age demographics can inform strategic decisions, such as programming content and marketing strategies, tailored to the preferences of different age

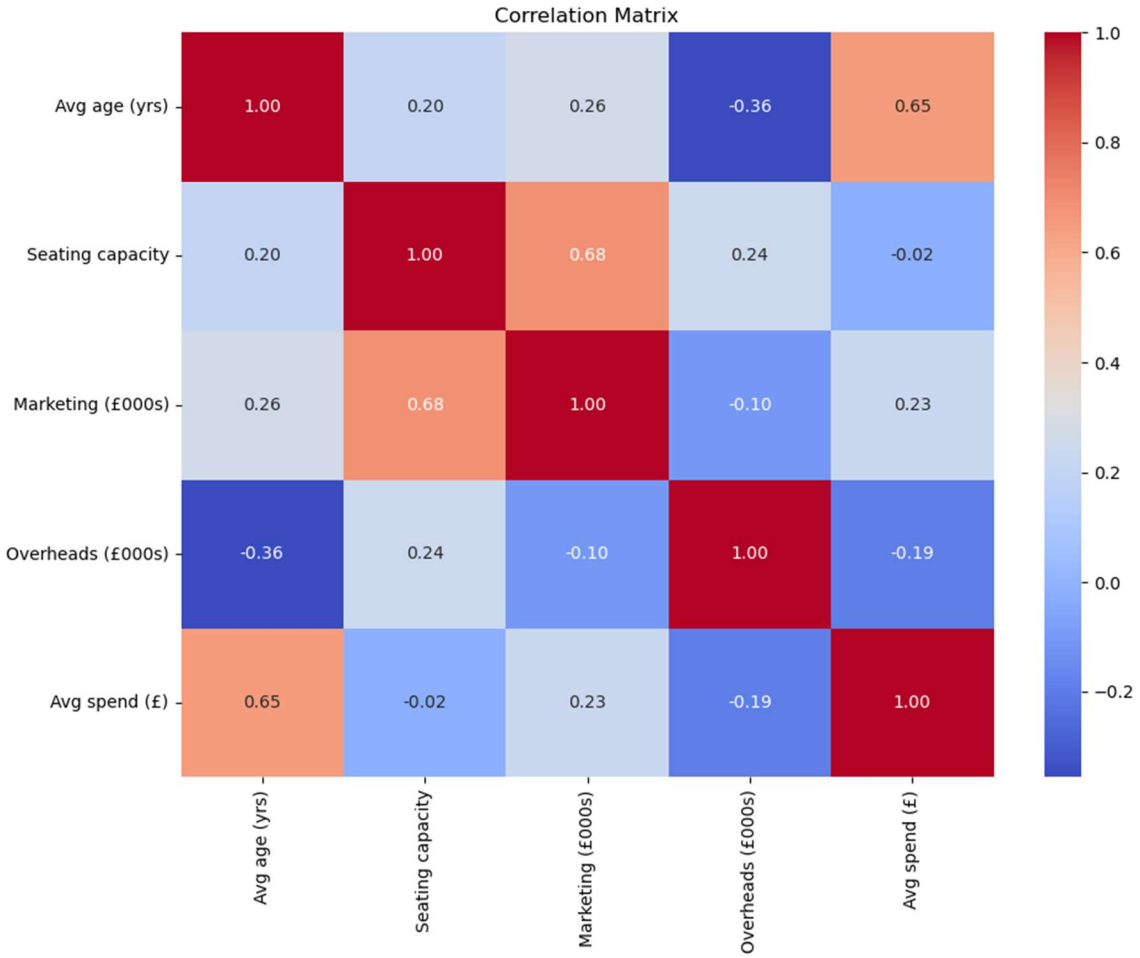
groups, ensuring a more targeted approach to enhancing the cinema experience and maximizing customer satisfaction.

2.3. Heatmap of Correlation Matrix **Fig. 3**:

A heatmap is an effective way to visualize the correlation between multiple numerical variables. In this visualization, we will create a heatmap to display the correlation matrix between average age, seating capacity, marketing expenses, overheads, and average spending per visitor across different cinemas.

Justification: A heatmap is suitable for visualizing the correlation matrix between numerical variables as it provides a color-coded representation of correlation coefficients. By examining the heatmap, we can quickly identify the strength and direction of correlations between different variables. This information can help identify potential relationships and dependencies between variables, allowing for more informed decision-making and analysis.

Fig. 3



Description of Fig. 3: The correlation matrix provides insights into the relationships between different numerical variables in the dataset. Each cell in the matrix represents the correlation coefficient between two variables, ranging from -1 to 1.

Avg age (yrs) vs. Seating capacity: There is a weak positive correlation (0.2) between the average age of visitors and seating capacity, indicating a slight tendency for cinemas with larger seating capacities to attract slightly older audiences.

Avg age (yrs) vs. Marketing (£000s): There is a moderate positive correlation (0.26) between the average age of visitors and marketing expenses, suggesting that cinemas targeting older demographics may invest more in marketing campaigns.

Avg age (yrs) vs. Overheads (£000s): There is a moderate negative correlation (-0.36) between the average age of visitors and overhead expenses, implying that cinemas with older audiences tend to have lower overhead costs.

Avg age (yrs) vs. Avg spend (£): There is a strong positive correlation (0.65) between the average age of visitors and average spending per visitor, indicating that older audiences may spend more on average during their cinema visits.

Seating capacity vs. Marketing (£000s): There is a strong positive correlation (0.68) between seating capacity and marketing expenses, suggesting that cinemas with larger seating capacities tend to allocate more resources to marketing efforts.

Seating capacity vs. Overheads (£000s): There is a weak positive correlation (0.24) between seating capacity and overhead expenses, indicating a slight tendency for cinemas with larger seating capacities to have higher overhead costs.

Seating capacity vs. Avg spend (£): There is a weak negative correlation (-0.02) between seating capacity and average spending per visitor, suggesting that there is no significant relationship between these two variables.

Marketing (£000s) vs. Overheads (£000s): There is a weak negative correlation (-0.10) between marketing expenses and overhead costs, implying that higher marketing expenses may be associated with slightly lower overhead costs.

Marketing (£000s) vs. Avg spend (£): There is a weak positive correlation (0.23) between marketing expenses and average spending per visitor, suggesting that higher marketing expenses may be associated with slightly higher average spending per visitor.

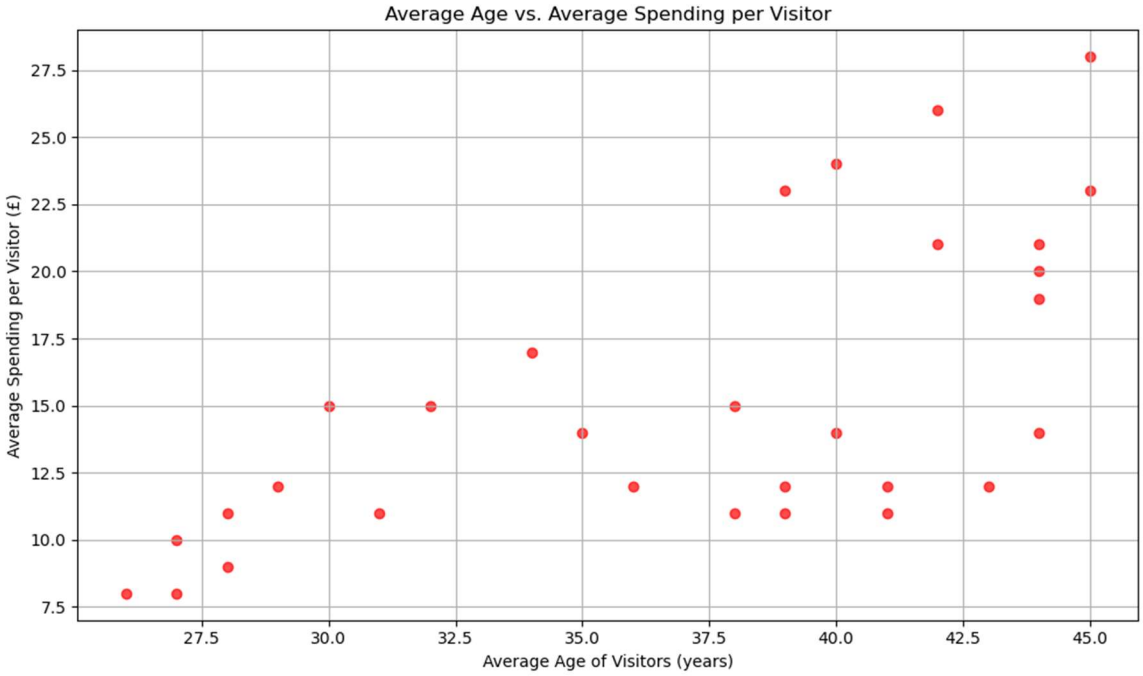
Overheads (£000s) vs. Avg spend (£): There is a weak negative correlation (-0.19) between overhead expenses and average spending per visitor, indicating that higher overhead costs may be associated with slightly lower average spending per visitor.

2.4. Scatter Plot of Average Age vs. Average Spending per Visitor **Fig. 4**:

In this visualization, we aim to explore the relationship between the average age of visitors and the average spending per visitor across different cinemas. The average age of visitors provides insights into the demographic characteristics of the audience, while the average spending per visitor indicates the financial behavior of the audience. By examining this relationship, we can gain valuable insights into whether there is a correlation between the age of the audience and their spending habits. More precisely, we will look at the graphical representation of correlation coefficient we found out between these two variables in our earlier heatmap.

Justification: A scatter plot is suitable for visualizing the relationship between average age and average spending per visitor as it allows us to examine individual data points and detect any patterns or trends. By plotting average age on the x-axis and average spending per visitor on the y-axis, we can assess whether there is a correlation between these two variables. This visualization will provide insights into how the age demographics of the audience may influence their spending behavior. It helps in understanding the financial dynamics of different age groups and can inform targeted marketing and pricing strategies.

Fig. 4



Description of Fig. 4: The scatter plot illustrates a notable positive relationship between the average age of visitors and the average spending per visitor across different cinemas. The correlation coefficient of approximately 68% (we looked at earlier in heatmap) suggests a moderate to strong positive correlation between these two variables.

This positive relationship implies that as the average age of visitors increases, there is a tendency for the average spending per visitor to also increase. This observation aligns with common consumer behavior patterns, where older individuals may have greater disposable income and are willing to spend more on entertainment activities such as movie-going.

Understanding this correlation is crucial for cinema management, as it highlights the importance of catering to different demographic segments when devising marketing strategies and pricing structures. Cinemas may choose to tailor their offerings, promotions, and amenities to attract and retain audiences of varying age groups, thereby maximizing revenue potential.

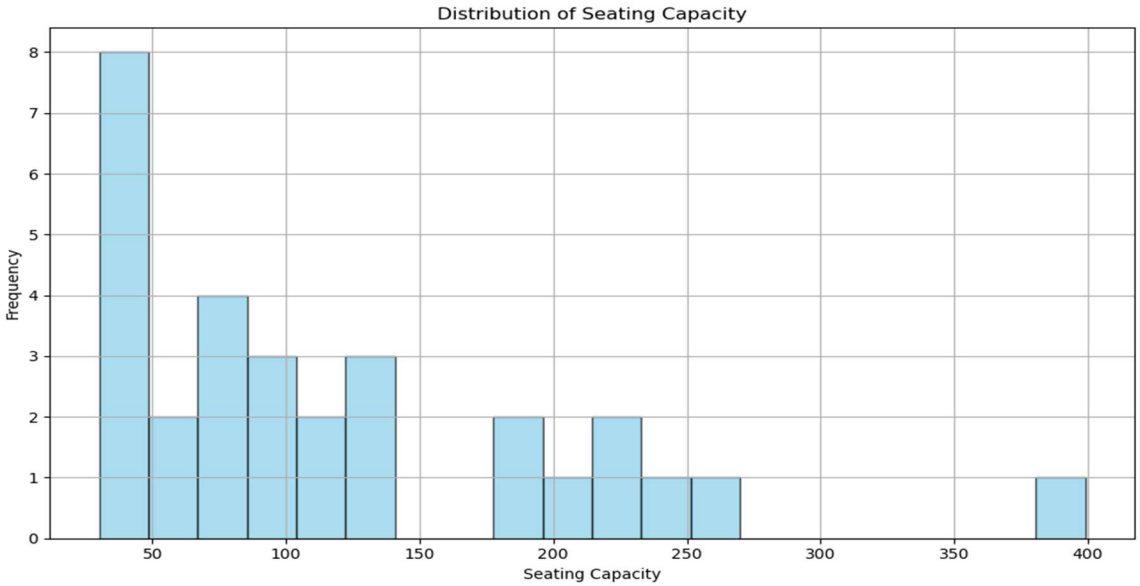
Overall, the scatter plot provides valuable insights into the interplay between demographic characteristics and spending behavior within the cinema industry, informing strategic decision-making aimed at enhancing customer satisfaction and profitability.

2.5. Histogram of Seating Capacity **Fig. 5:**

In this visualization, we aim to analyze the distribution of seating capacity across cinemas. A histogram provides insights into the frequency distribution of a continuous variable, allowing us to identify any patterns or outliers in the data.

Justification: A histogram is well-suited for visualizing the distribution of seating capacity as it provides a clear representation of the frequency distribution of continuous data. By dividing the seating capacity into bins and counting the number of cinemas falling into each bin, we can observe the distribution pattern. This visualization will help us understand the range and variability of seating capacities across cinemas, identifying any common trends or potential outliers. Additionally, it provides insights into the typical size of cinemas within the dataset, aiding in benchmarking and comparison analyses.

Fig. 5



Description of Fig.5: The histogram of seating capacity reveals interesting insights into the distribution of seating capacities across cinemas. The frequencies provides a clear breakdown of the seating capacity ranges and their corresponding frequencies.

We observe that the majority of cinemas have seating capacities ranging from 30 to 85, with a peak frequency observed in the range of 30 to around 48. This indicates that a significant number of cinemas in the dataset have relatively smaller seating capacities.

Additionally, there are a few cinemas with larger seating capacities, particularly in the ranges of 85 to 140. However, as we move towards higher seating capacities, the frequency gradually decreases, suggesting that cinemas with larger seating capacities are less common in the dataset.

The absence of frequencies in certain bins, such as 140 to around 180, indicates that there are no cinemas with seating capacities falling within those ranges. This may represent a gap in the dataset or reflect the real-world distribution of cinema sizes (of our fictional cinema company)

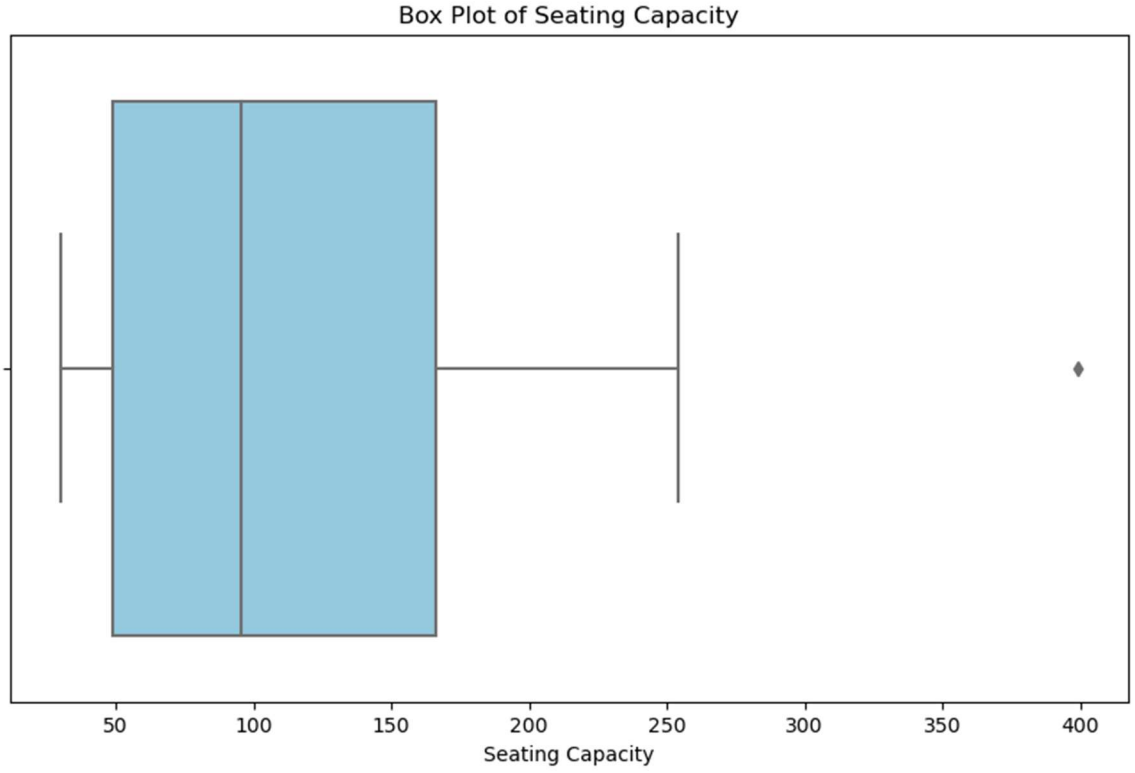
Overall, the histogram provides valuable insights into the distribution of seating capacities, helping stakeholders understand the landscape of cinema sizes within the dataset. This information can inform decisions related to resource allocation, audience targeting, and business strategies within the cinema industry.

2.6. Box Plot of Seating Capacity **Fig.6:**

A box plot is an effective visualization for displaying the distribution of a numerical variable, such as seating capacity in this case. It provides insights into the central tendency, variability, and presence of outliers in the data.

Justification: A box plot is well-suited for visualizing the distribution of seating capacity across different cinemas. It provides key summary statistics such as the median, quartiles, and potential outliers in the data. This visualization enables stakeholders to quickly identify any variability or anomalies in seating capacity among cinemas, facilitating decision-making processes related to resource allocation, capacity planning, and operational management.

Fig.6



Description of Fig. 6;

The box plot visually summarizes the distribution of seating capacity across cinemas:

Median (Second Quartile): The middle line within the box represents the median seating capacity, which is approximately 95. This implies that half of the cinemas have a seating capacity below this value, while the other half have a seating capacity above it.

Interquartile Range (IQR): The box itself spans from the first quartile (Q1) to the third quartile (Q3). The Q1, located at around 45, indicates that 25% of the cinemas have a

seating capacity lower than this value. Similarly, the Q3, positioned at about 175, suggests that 75% of the cinemas have a seating capacity lower than this value.

Whiskers: The upper and lower whiskers extend to the minimum and maximum non-outlier data points within 1.5 times the IQR from the quartiles. They show the range of most of the data points.

Outliers: There is one outlier observed at approximately 400, denoted by a point beyond the whiskers. This outlier suggests a cinema with exceptionally high seating capacity compared to others in the dataset.

Overall, the box plot provides a concise summary of the distribution of seating capacities, highlighting key statistics and identifying potential outliers for further investigation.

2.7. Interactive Visualization: Scatter Plot with 3D Hover Tooltips **Fig. 7:**
In this interactive scatter plot, we'll explore the relationship between average years, seating capacity, and marketing spending across all cinemas. Users can hover over data points to view detailed information about each cinema, including its ID, average years, seating capacity, and marketing spending.

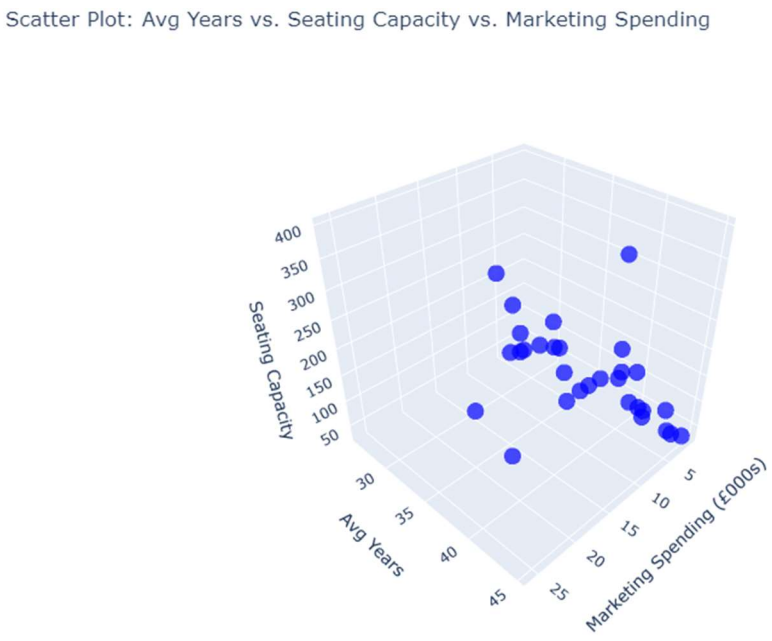
Justification: This interactive scatter plot allows users to explore the relationship between average years, seating capacity, and marketing spending across all cinemas.

The 3D scatter plot provides a clear visualization of the data distribution in three dimensions. Hover tooltips display detailed information about each cinema, facilitating a deeper understanding of how these variables interact. Users can identify patterns or outliers and analyze the distribution of cinemas based on their average years, seating capacity, and marketing spending.

This visualization offers an intuitive way to explore the multidimensional relationship between these key variables and provides valuable insights into the characteristics of different cinemas.

Check out the notebook file to use the hover tool

Fig. 7



Description of Fig.7: The 3D scatter plot visualizes the relationship between Marketing Spending (£000s), Average Age (yrs), and Seating Capacity across all cinemas. Each point in the plot represents a cinema, with its position determined by these three variables.

From the data provided:

Marketing Spending (£000s): Ranges from 1 (£000s) to 26 (£000s). Higher values indicate greater investment in marketing efforts.

Average Age (yrs): Spans from 26 years to 45 years. Represents the average age of visitors.

Seating Capacity: Varies widely from 30 to 399. Indicates the number of seats available in each cinema.

Interpreting the plot:

Points positioned higher on the y-axis (Average Age) represent cinemas with older visitors.

Cinemas with larger seating capacities (points further right on the x-axis) tend to have higher marketing spending (points further up on the z-axis).

There's a scattered distribution, suggesting no clear linear relationship between these variables but showcasing the diversity in cinema characteristics.

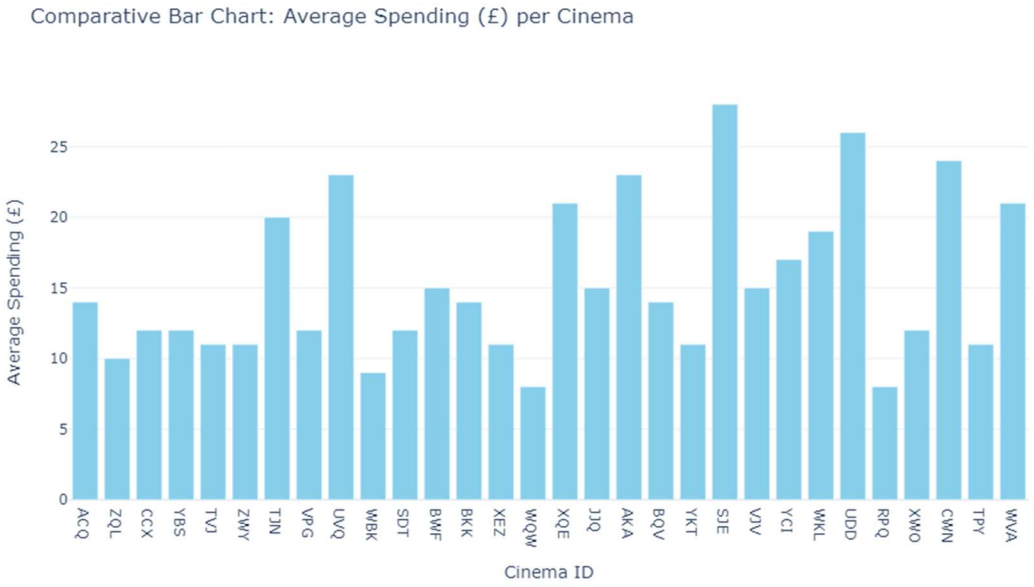
2.8. Comparative bar chart **Fig. 8:**

For the second interactive plot, let's consider a comparative bar chart showcasing the average spending (£) across different cinemas. This visualization will allow viewers to compare the average spending at each cinema easily. Go @ notebook file to use the hover tool.

Justification: A comparative bar chart is an effective choice for comparing average spending (£) across different cinemas because it allows viewers to easily visualize and compare the spending levels at each cinema. The use of interactive features, such as hover text displaying cinema IDs, enhances the viewer's ability to explore the data in more detail.

This visualization is relevant to the company's interest in understanding spending patterns across cinemas and can provide valuable insights into areas where spending might be higher or lower compared to others. Additionally, the interactive nature of the plot allows users to hover over bars and quickly access specific information about each cinema's average spending.

Fig. 8



Description of Fig. 8: The interactive comparative bar chart provides a clear visualization of the average spending (£) across different cinemas in our dataset. Each bar represents a cinema, and its height corresponds to the average spending for that particular cinema.

Upon analysis, several interesting trends emerge:

Variability in Spending: The chart highlights significant variability in average spending among cinemas. Cinemas such as SJE, UDD, and AKA exhibit notably higher average

spending, with bars towering above others. Conversely, cinemas like WBK, RPQ, and WQW show lower average spending, indicated by shorter bars.

Identifying Outliers: Some cinemas, such as SJE, stand out as outliers with exceptionally high average spending. These outliers may warrant further investigation to understand the factors contributing to their distinct spending patterns.

Patterns and Clusters: Despite the variability, certain clusters of cinemas with similar spending patterns can be observed. For instance, cinemas like ZQL, CCX, YBS, TVJ, ZWY, VPG, and XEZ display relatively consistent average spending levels, forming a cluster of moderate spenders.

Insights for Decision-making: The chart offers valuable insights for decision-making, allowing stakeholders to identify cinemas with potential spending inefficiencies or opportunities for investment. Understanding the spending patterns across different cinemas can inform resource allocation strategies and marketing efforts to optimize revenue generation.

Overall, the interactive comparative bar chart serves as a powerful tool for visually exploring and understanding spending trends across cinemas, enabling data-driven decision-making and strategic planning.

3. Critical Review and Discussion:

Throughout this coursework, I have employed various data visualization techniques to explore and analyze cinema data effectively. By utilizing Python libraries such as Matplotlib, Seaborn, and Plotly, I have demonstrated a range of visualizations, including line plots, scatter plots, histograms, box plots, comparative bar charts, 3D scatter plots, and area plots. Each visualization was carefully chosen to suit the nature of the data and the analysis objectives, ensuring clarity and interpretability.

Through this coursework, I have gained a deeper understanding of the importance of data visualization in extracting meaningful insights from complex datasets. By visualizing data in different ways, I was able to uncover patterns, trends, and relationships that might have been overlooked otherwise. Moreover, I learned the significance of interactive visualizations in enabling users to explore data dynamically and gain deeper insights.

Applying best practices in data visualization, such as providing clear titles, labels, and legends, ensured that the visualizations were easy to understand and interpret. Additionally, I strived to maintain consistency in color schemes and formatting across different visualizations to enhance readability and coherence.

In summary, this coursework has reinforced my understanding of the role of data visualization in data analysis and decision-making processes. By applying various visualization techniques to real-world data, I have honed my skills in effectively

communicating insights and findings visually, thereby adding value to the data analysis process.

4. Summary of Conclusions:

- The cinema dataset exhibits variability in several key metrics, including average age, seating capacity, marketing spending, and average spending.
- Relationships between different variables, such as seating capacity and total visitors, marketing spending, and average spending, were identified through visual analysis.
- Outliers and clusters within the dataset were detected, providing insights into potential areas for further investigation or strategic focus.
- Interactive visualizations, such as comparative bar charts and 3D scatter plots, offer dynamic exploration capabilities, enabling stakeholders to gain deeper insights into the data.
- Effective data visualization is essential for understanding complex datasets and making informed decisions based on data-driven insights.

References

Kirk, A., 2012. Data Visualization: A Handbook for Data Driven Design.