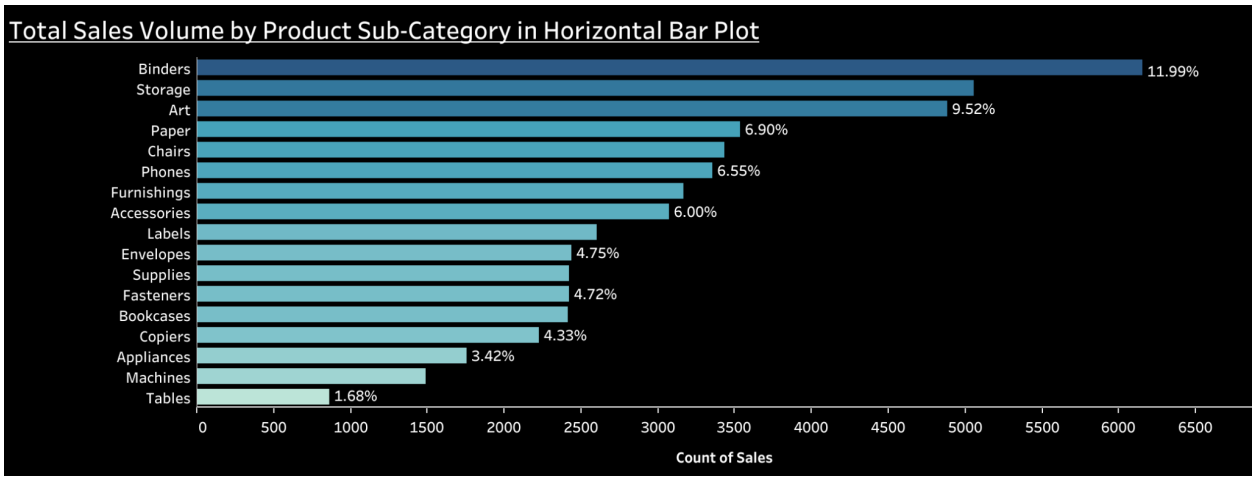
Assignment 3

Submitted by Anitha Balachandran (016684486)

Chapter 6: Visualizing Amounts

"Visualizing Amounts" means creating a visualization that emphasizes quantitative values when displaying a set of numbers. Below charts can be used to visualize amounts.
Dataset Used: Super Store business Dataset

1. Horizontal bar charts:

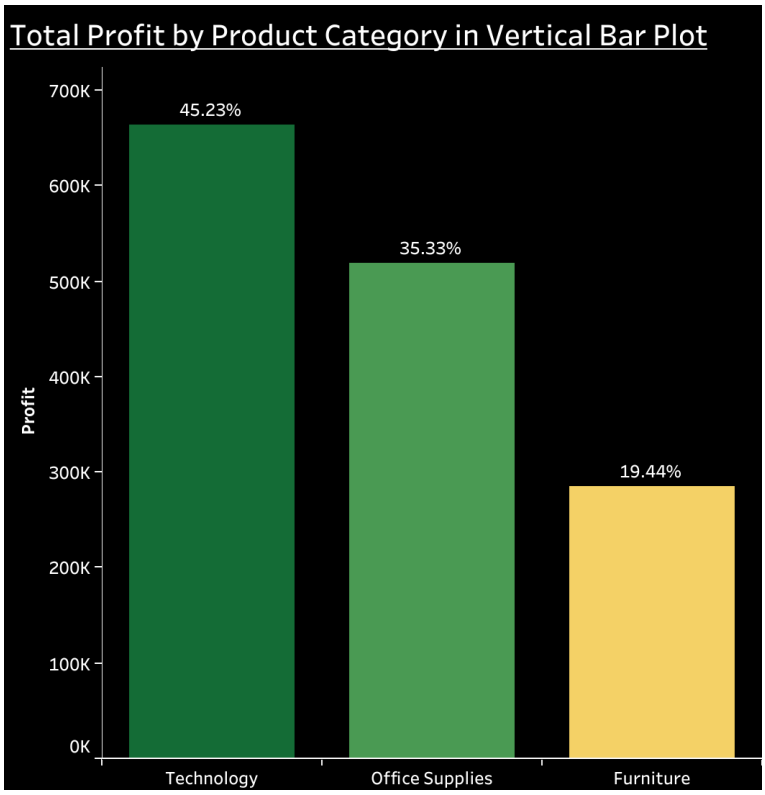


The above Horizontal Bar plot can be used to identify which sub-categories are the most popular and profitable for Super Store business, and which ones may need more attention or optimization.

Observations:

- The sub-category "Binders" has the highest sales volume, followed by "Storage" and "Art" which are highlighted with dark colors.
- The sub-category "Tables" has the lowest total sales volume, followed by "Machines" and "Appliances" which are highlighted by very light colors.

2. Vertical bar charts:



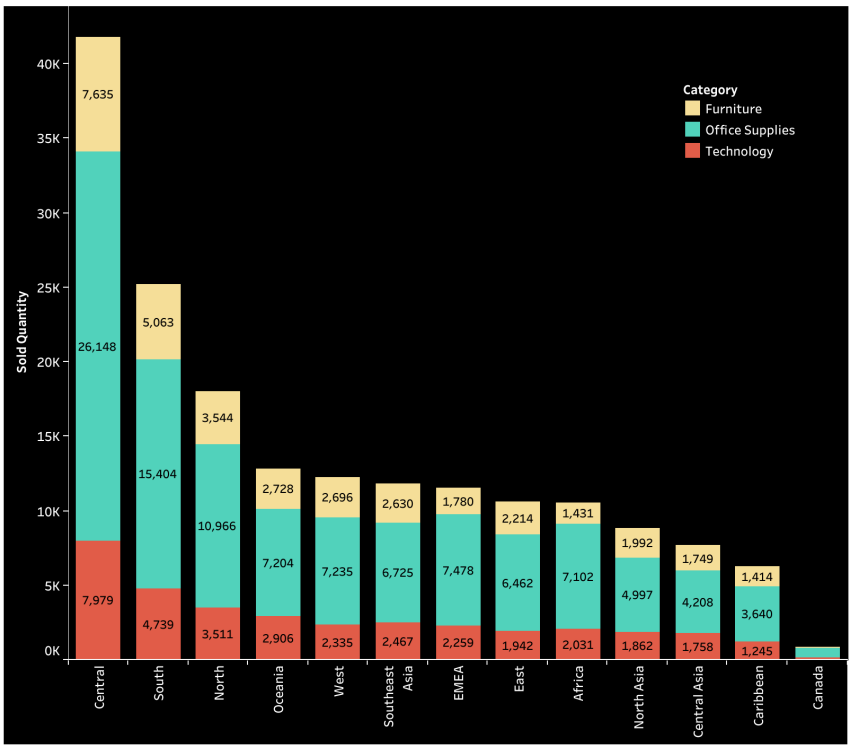
The above vertical bar plot highlights the importance of understanding the profitability of different product categories and subcategories in order to make informed business decisions about marketing, pricing, and product offerings.

Observations:

- The Technology category has the highest average profit margin compared to the other categories. This indicates that the products within this category are generally more profitable per unit sold, compared to the other categories.
- The Office Supplies category has the second-highest total profit, following the Furniture category.
- The Furniture category has the third-highest total profit, but has a much lower average profit margin compared to the Technology and Office Supplies categories. This could indicate that the company may want to re-evaluate its pricing or product mix within this category to improve profitability.

3. Stacked Bar Plot:

Total Quantity Sold by Product Category and Region in Stacked Bar Chart

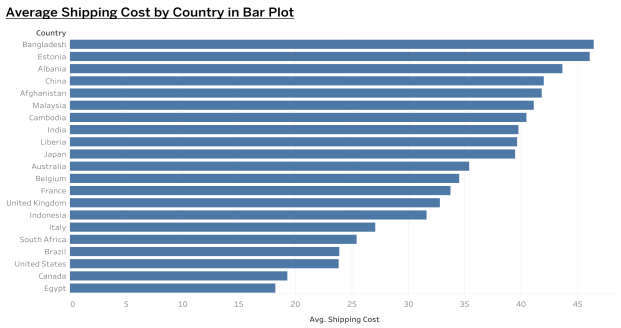


The above stacked bar plot for total quantity sold by product category and region highlights areas of focus for the company to optimize sales, adjust pricing strategies, and increase profitability.

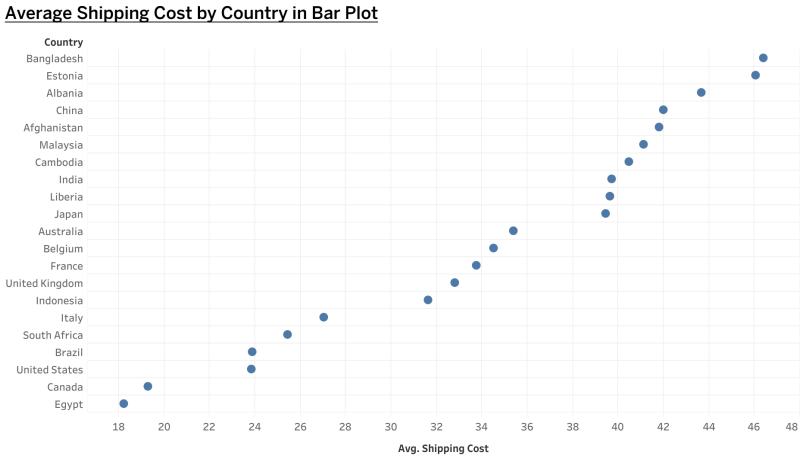
Observations:

- The highest total quantity sold across all regions is in the "Office Supplies" category, indicating it may be the most popular category among customers. This is followed by the "Furniture" and "Technology" categories, in that order,
- Despite having a higher profit margin(shown in second plot), the "Technology" category has a lower total quantity sold compared to the "Office Supplies" and "Furniture" categories, suggesting that customers may prefer purchasing items from these two categories over technology products.
- The region "Central" has the highest total quantity sold across all categories, followed by "South", "North", and "Oceania". "Caribbean" has the second-lowest total quantity sold across all categories, only slightly above "Canada".
- The lower quantity sold in the "Canada" region could be an area of concern for the company and may require further investigation into potential reasons for the lower sales in that area.

4. Need for Dot Plot:



Bars are too long and it doesn't clearly convey the comparison of avg shipping cost across countries as most of them have nearly the same length, cluttered and difficult to read .



By limiting the axis range to the interval from 18 to 48 Avg(ordered) Shipping Cost, the above Dot Plot highlights the below key observations of this dataset to provide a more clear and concise visualization of the data.

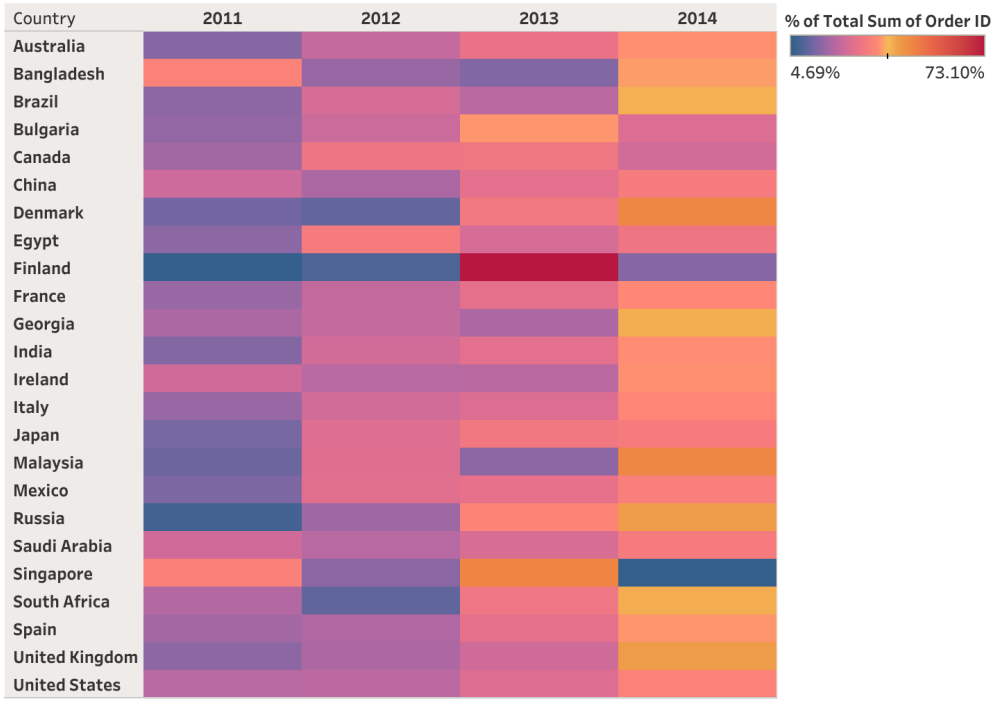
Observations:

- The average shipping cost varies significantly among different countries.

- ➢ Bangladesh has the highest average shipping cost, among all listed countries,followed by the United States and Mexico.
- ➢ Egypt, Canada and the United States have lower average shipping costs than all other countries.

5. Visualization using Heat Map:

Percentage of Orders by Year across Countries in Heat Map



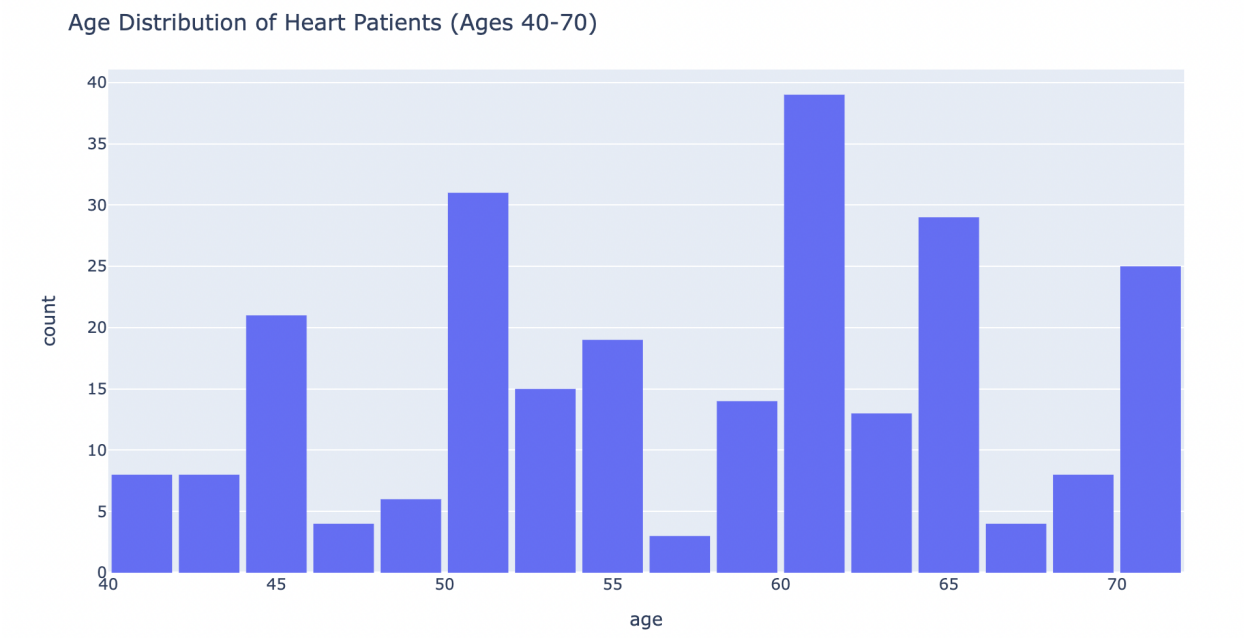
The above heat map provides a visual representation of the percentage of orders by year across different countries.

Observations:

- ➢ The majority of orders are concentrated in Finland, with the highest percentage of orders in 2013 and Finland at the same time had the lowest percentage of orders in 2011 and gradually increased in 2013 .
- ➢ In 2014, most of the countries had a neutral percentage of orders between 35% to 45% which is a good improvement for business compared to other years.

Chapter 7. Visualizing Distributions: Histograms and Density Plots

1. Age Distribution of Heart Patients using Histogram:



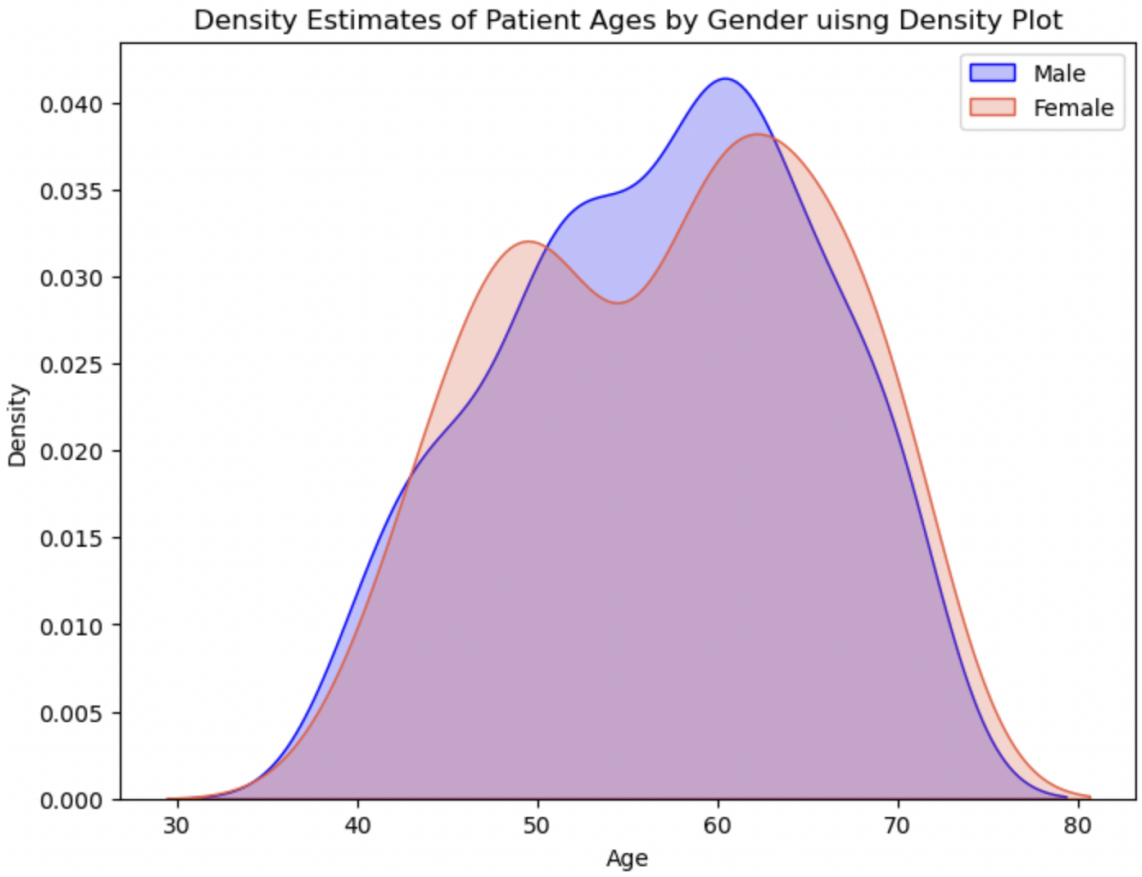
The below histogram provides insight into the age distribution of heart patients with Bin range: [40, 70] and Total patients: 247 which can be useful for identifying trends and patterns in the data.

Observations:

- The majority of heart patients in this age range are between 50-65 years.
- There is a noticeable dip in the number of heart patients in their mid-40s compared to the surrounding age groups.

Visualizing Multiple Distributions at the Same Time

2. Density Estimates of Patient Ages by Gender:



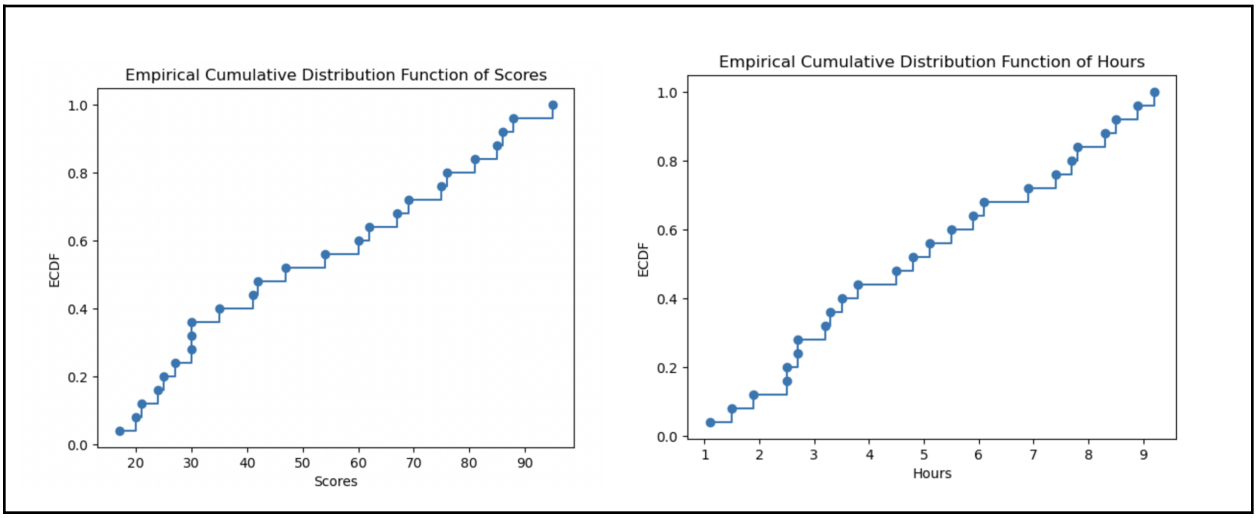
The above density plot shows the distribution of age for male and female patients separately and helps us understand the differences in the age distribution between the two genders.

Observations:

- The density plot for male patients shows a peak around age 55-60, while for female patients the peak is around 65.
- The distribution of age for female patients is wider and more spread out compared to female patients.
- The plot suggests that male patients are generally younger than female patients in this dataset.

Chapter 8. Visualizing Distributions: Empirical Cumulative Distribution Functions and Q-Q Plots

1. Empirical Cumulative Distribution Functions using step plot from matplotlib:



The above ECDF step plot explains the distribution of scores and hours studied by the students and the relationship between them, which can aid in setting appropriate grade boundaries and identifying students who may need extra help.
The first plot shows the ECDF of the scores obtained by students in an exam, while the second plot shows the ECDF of the hours studied by the same students for the exam.

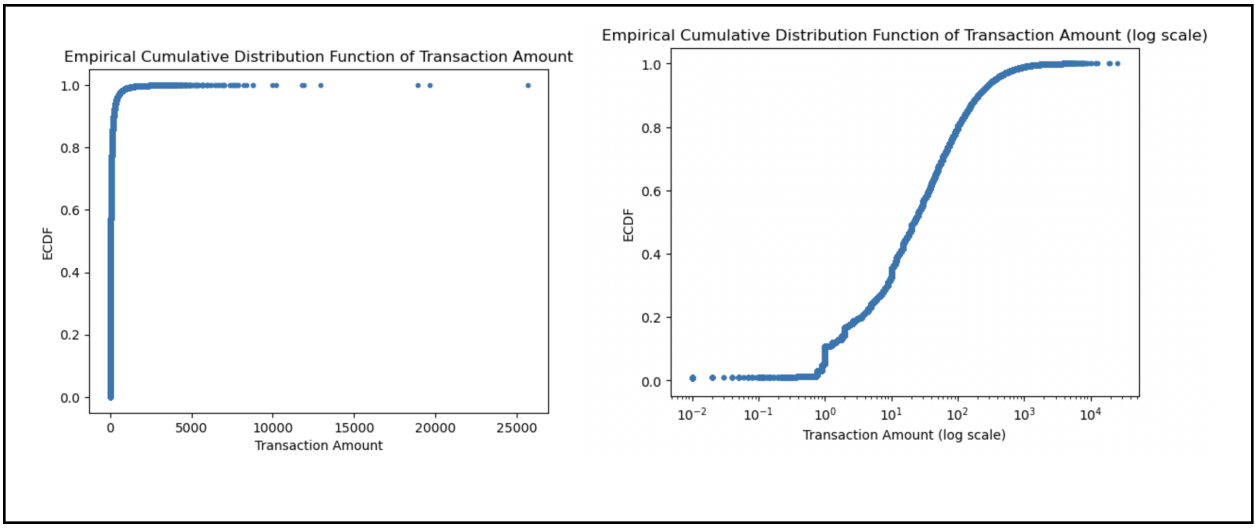
Observations:
From the ECDF plot of scores,

- ➢ Most students scored between 50 to 80 marks, with only a few students scoring above 80 marks.
- ➢ The distribution of scores appears to be slightly right-skewed, indicating that more data points with lower values and fewer data points with higher values.
- ➢ Approximately 60% of students scored less than or equal to 70 marks.

From the ECDF plot of hours studied,

- ➢ Most students studied between 2 to 5 hours, with only a few students studying for more than 6 hours.
- ➢ The distribution of hours studied appears to be slightly left-skewed, indicating that there are more data points with higher values and fewer data points with lower values.
- ➢ Approximately 80% of students studied less than or equal to 5 hours.

2. Empirical Cumulative Distribution Function of Highly Skewed Transaction Amounts:

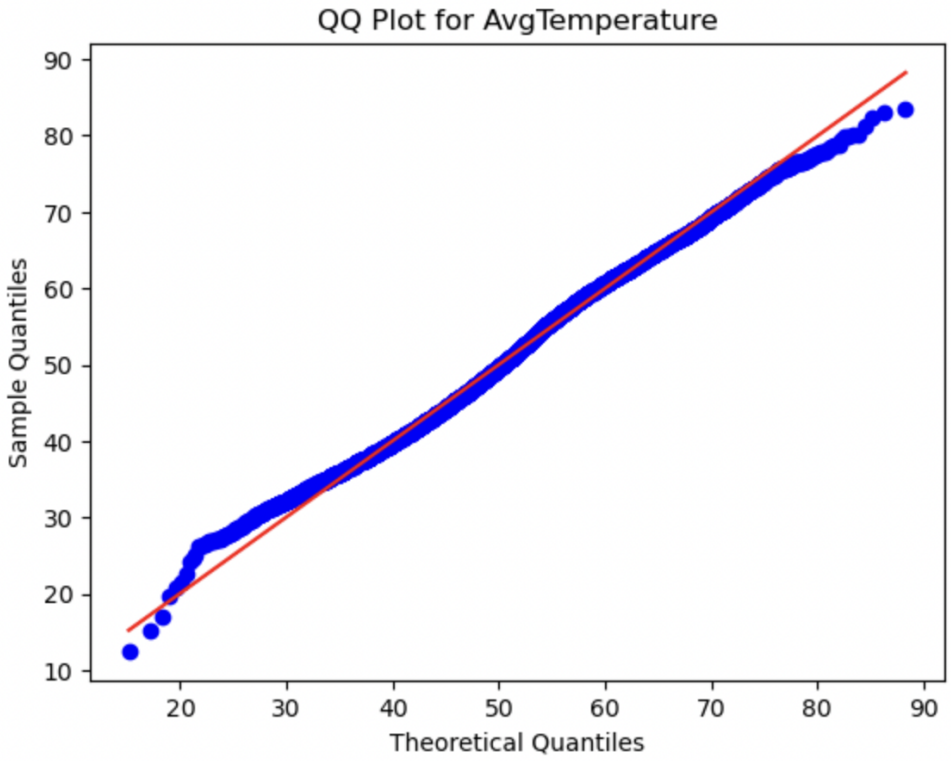


The above plot highlights the importance of identifying and dealing with highly skewed data, as and shows the comparison of distribution of transaction amounts on a logarithmic scale and regular scale.

Observations:

- The plot shows a steep increase in the ECDF at lower values of transaction amount, indicating that a large proportion of transactions have relatively small amounts as the majority of transactions are relatively small, with more than 90% of transactions being less than $1000.
- The plot also shows a long tail stretching out towards the higher values of transaction amount, indicating that there are a small number of transactions with very high amounts i.e. small number of transactions that are much larger, with the maximum transaction amount being close to $25,000.
- Comparing the two plots, we can see that the log-scale plot allows us to better visualize the distribution of the majority of transactions which are relatively small (less than $10,000), as they are more spread out and easier to distinguish. On the other hand, the regular-scale plot doesn't show much detail for these smaller transactions, making it difficult to see the shape of the distribution.

3. **Quantile-Quantile Plot for AvgTemperature**



The above QQ plot is a graphical method to check if the data follows a normal distribution or not for examining the distribution of average temperatures across different cities.

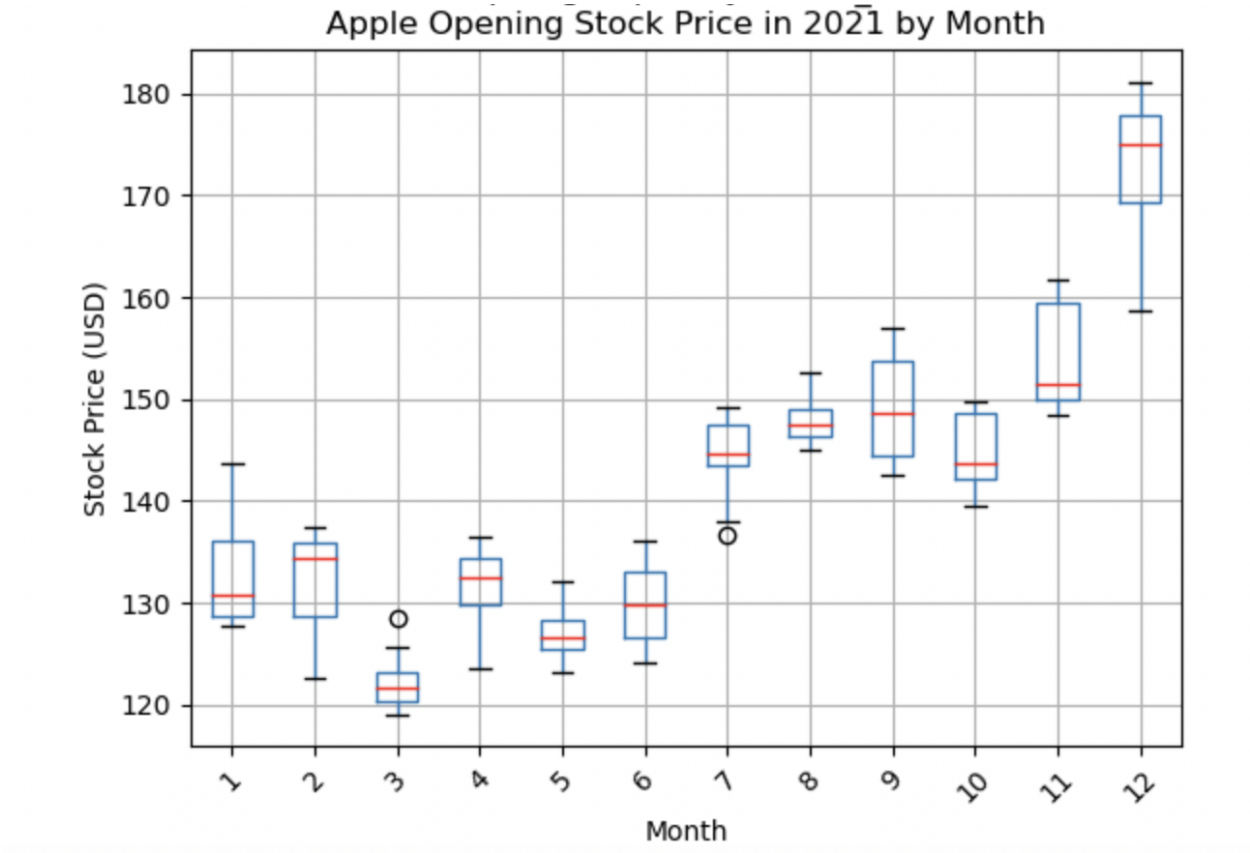
Observations:

- The plot shows that the data roughly follows a normal distribution, as the points are roughly in a straight line. However, there are some deviations from the straight line towards the tails, indicating that the distribution may not be perfectly normal.
- Generated a normal distribution with the same mean and standard deviation as the city temperature data, which is shown as a red line on the plot. The line intersects with the points at the center, indicating that the mean and standard deviation are a good fit for the data.

Chapter 9. Visualizing Many Distributions at Once

Visualizing Distributions Along the Vertical Axis

1. **Apple Stock Price Open in 2021 by Month using Box Plot:**



The above box plot shows the distribution of opening stock prices for Apple.inc in each month of 2021 and by analyzing the boxplot, investors can gain insights into the volatility of the stock price, identify periods of high or low performance, and make informed decisions about buying or selling Apple's stock.

Observations:

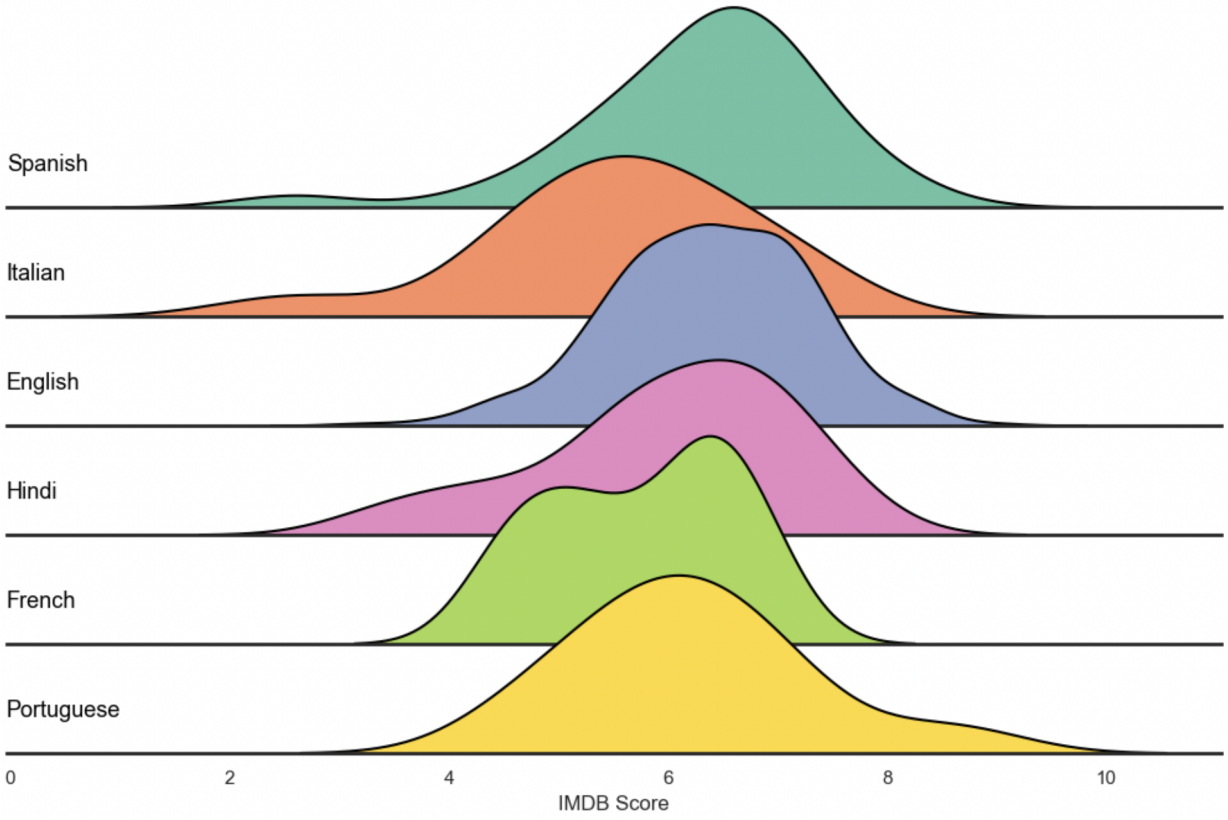
- The x-axis shows the month number, and the y-axis shows the stock price in USD
- The median stock price high for each month is indicated by the red line within the box, and the whiskers show the range of the data excluding any outliers.
- The box and whiskers in the months of December, November, September, June, February and January are relatively long, indicating that the data is spread out.
- The boxplot shows that the median opening stock price for Apple Inc. was around

130 USD for most of the months in 2021.

- ➢ The stock prices in March, May and August show less variability, with most of the data points falling within a narrow range.
- ➢ In the boxplot for the third month, there is an outlier above the upper whisker, which means there was an unusually high opening stock price on that day.
- ➢ In the boxplot for the seventh month, there is an outlier below the lower whisker, which means there was an unusually low opening stock price on that day.

Visualizing Distributions Along the Horizontal Axis

2. Distribution of IMDB Scores for Netflix Original Films in Different Languages using Ridgeline Plot:



The above Ridgeline plot provides a clear and informative visualization of the distribution of IMDB scores for Netflix original films in different languages.

Observations:

- The IMDB scores for Netflix original films in different languages are quite high, with most films scoring above 6.0.
- Films in Italian, Hindi and Spanish have a wider range of IMDB scores compared to films in other languages.
- For other languages such as French, English, and Portuguese, the distribution of IMDB scores is more concentrated around the mean which means that there are fewer movies in these languages that have very high or very low IMDB scores.
- The plot indicates that the distribution of IMDB scores for films in different languages is non-normal and each language's density plot has a different shape, suggesting that various factors affect the IMDB scores of films in different languages.