Data Glacier

Your Deep Learning Partner

G2M Case Study

Virtual internship

Shiva Ramezani

05/21/2023

Agenda

- Problem Statement
- Assumption
- Approach
- EDA
- Summary
- Recommendations

Problem Statement

- XYZ, a US-based private equity firm, intends to invest in the Cab Industry due to its significant expansion and the presence of numerous influential players. The aim is to assist XYZ in selecting the appropriate company for investment by offering practical guidance.
- **Objective**: Provide actionable insights to help XYZ firm in identifying the right company for making an investment.

Assumption

- **Data Integrity:** It is assumed that the provided datasets, including Cab_Data.csv, Customer_ID.csv, Transaction_ID.csv, and City.csv, are reliable and accurately represent the relevant information for analysis.
- **Data Merging:** The datasets have been merged based on common columns such as "City" and "Company" to combine relevant information from different sources. It is assumed that the merging process has been performed correctly, and the resulting merged dataset accurately represents the combined information.
- **Missing Data:** It is assumed that missing data, if any, has been handled appropriately. This includes strategies such as imputation or excluding incomplete data points, as deemed suitable for the specific analysis.
- **Currency and Units:** The financial figures in the datasets, such as "Price Charged," "Cost of Trip," and "Profit," are assumed to be in the same currency. Additionally, the units of measurement for distance (e.g., "KM Travelled") and population-related features (e.g., "Population," "Users") are assumed to be consistent throughout the datasets.
- **Time Frame:** The analysis is based on the available data within the specified time frame. It is assumed that this time frame provides a representative sample for understanding the patterns and trends in the data.

Assumption

- **Data Preprocessing:** Before proceeding with the analysis, ensure the dataset is clean and well-structured. Handle missing values, remove duplicates if any, and validate the data consistency and integrity. Convert relevant columns to appropriate data types, such as converting object type to datetime type for date-related columns.
- **Exploratory Data Analysis (EDA):** Perform exploratory data analysis to gain initial insights into the dataset. This includes descriptive statistics, data visualization, and identifying any outliers or anomalies. Explore the relationships between variables, uncover patterns, and understand the distribution of key features.
- **Customer Segmentation:** Utilize demographic features such as age, gender, and income to segment customers into meaningful groups. Analyze the preferences and usage patterns of each segment to identify the customer segments that contribute the most to the business in terms of revenue or a number of rides. Tailor marketing strategies and promotional campaigns specific to each segment to enhance customer satisfaction and loyalty.
- **Payment Mode Analysis:** Investigate the distribution of payment modes used by customers and examine any variations across different companies or cities. Determine if there is a correlation between payment modes and factors like profit or customer satisfaction. Identify the most common payment modes and explore strategies to optimize the payment options offered to customers.
- **Travel Patterns:** Analyze the date of travel to identify seasonal trends or patterns in cab usage. Determine if there are specific days of the week or months with higher demand. Visualize the trends using line plots or heatmaps to understand the variations in cab usage throughout the year. Use these insights to optimize resource allocation, pricing strategies, and marketing campaigns during peak demand periods.
- **City and Company Relationships:** Investigate the relationship between cities and companies in terms of cab usage. Analyze factors such as profit, number of rides, and user distribution across different cities and companies. Identify the cities where specific companies have a higher presence and assess their market share. Leverage these insights to target specific cities and optimize operations and marketing efforts accordingly.

EDA

Four files have been used:

- Cab_data.csv
- City_data.csv
- Customer_data.csv
- Transaction_data.csv

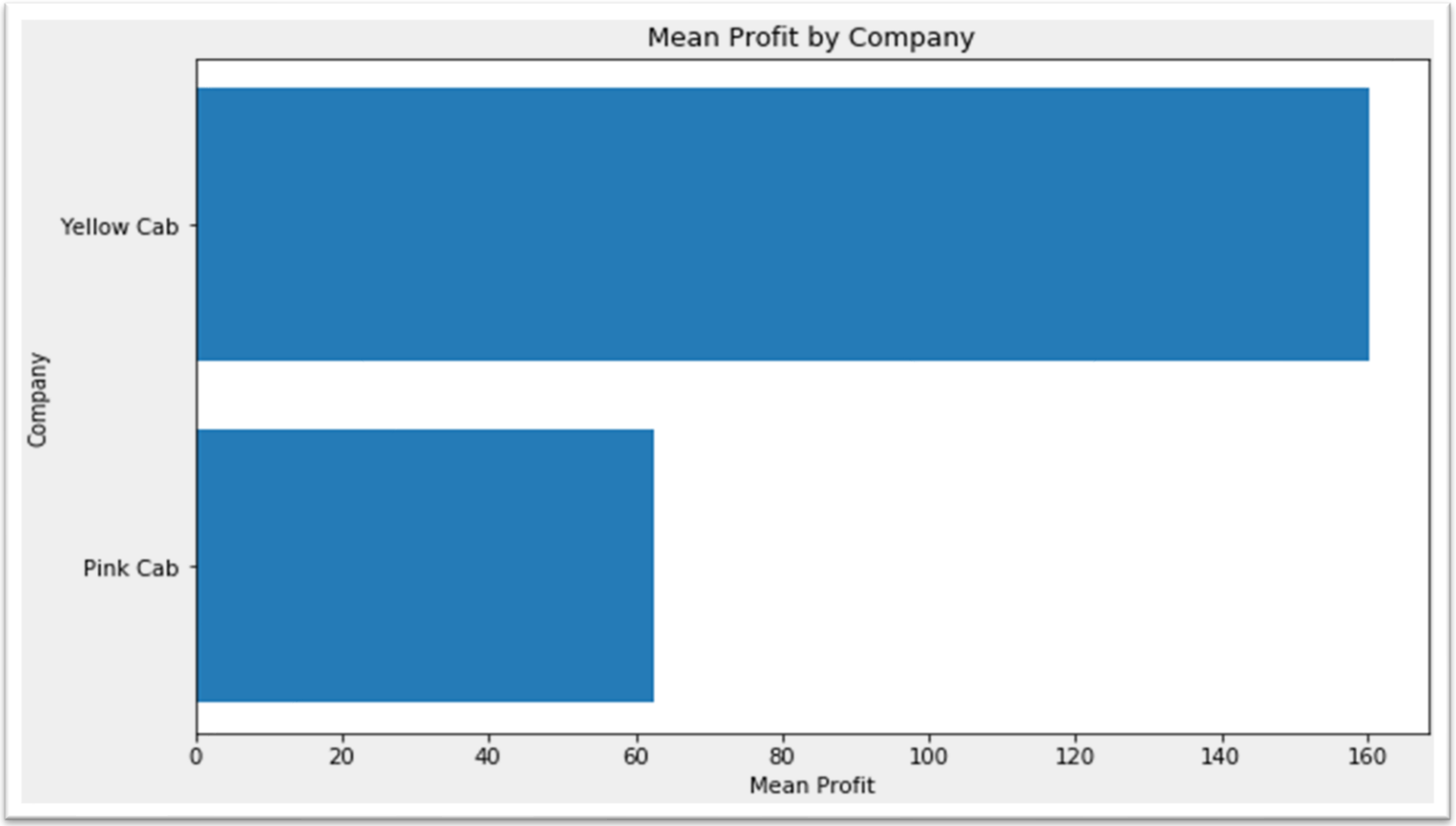
Three merged files were created:

- Merged_data = Cab_data & City_data
- Merged_data2 = Customer_data & Transaction_data
- Merged_data3 = Merged_data & Merged_data2

EDA - Cab data

- Importing cab dataset
- Checking the data types and changing the travel date to DateTime type.
- Checking for null values. None were found.
- Creating two new columns for profit and profit per kilometer.
  - Profit = Price charged – cost
  - Profit per km = profit / km traveled
- Aggregating the mean of the profit for each cab company and visualizing them.
- Aggregating the mean of the profit per KM for each cab company and visualizing them.
- Plotting the distribution of prices charged by different companies.
- Visualizing the relationship between the aggregated values of KM Travelled and Profit for each company in a scatter plot
- Visualizing the mean profit for each city in a bar plot, sorted in ascending order.

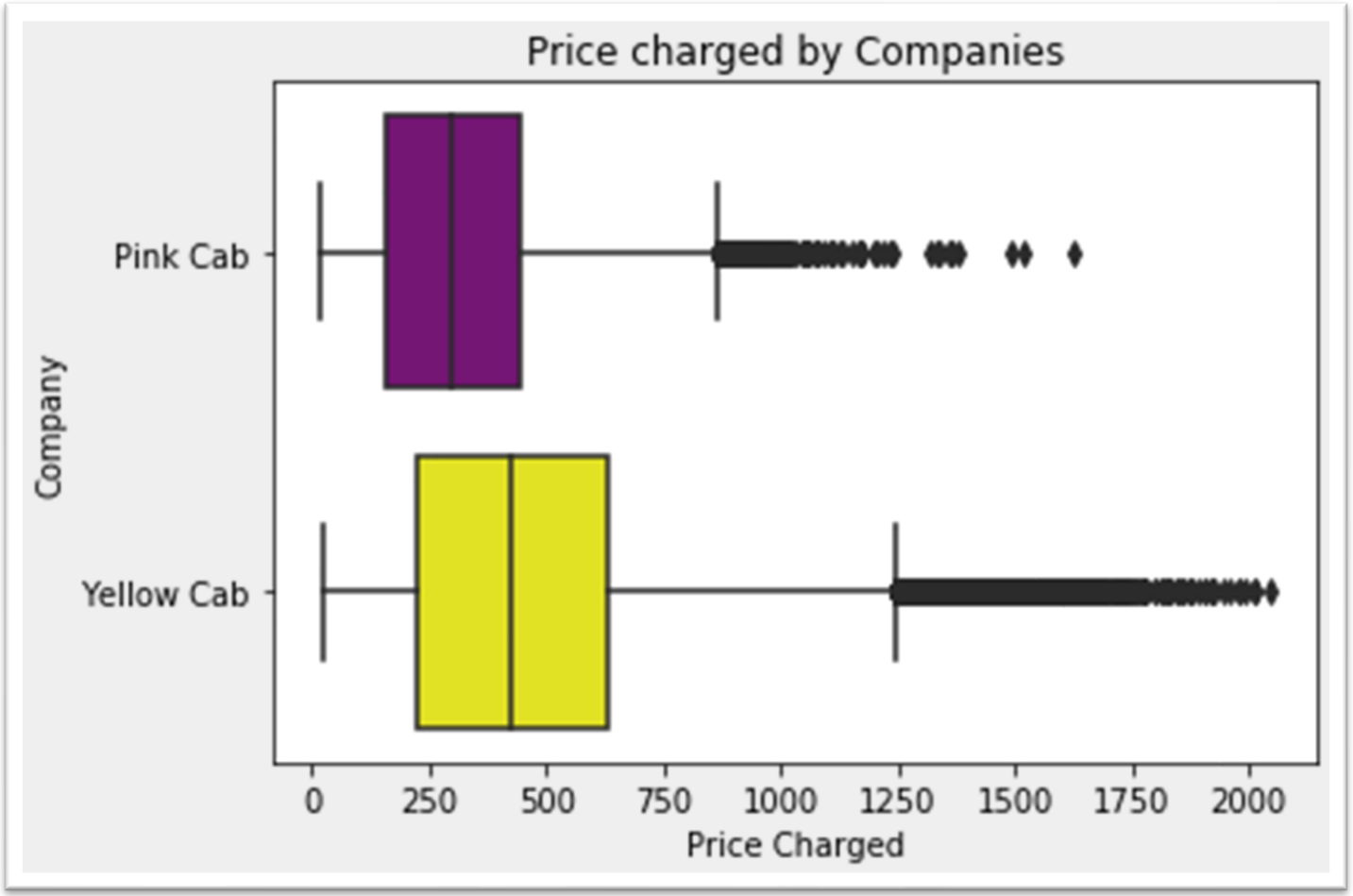
EDA - Cab data



The mean profit of each company in a horizontal bar plot is plotted which allows us to easily compare and analyze the profitability of different companies.

Result: The Yellow Cab Company has a higher mean profit.

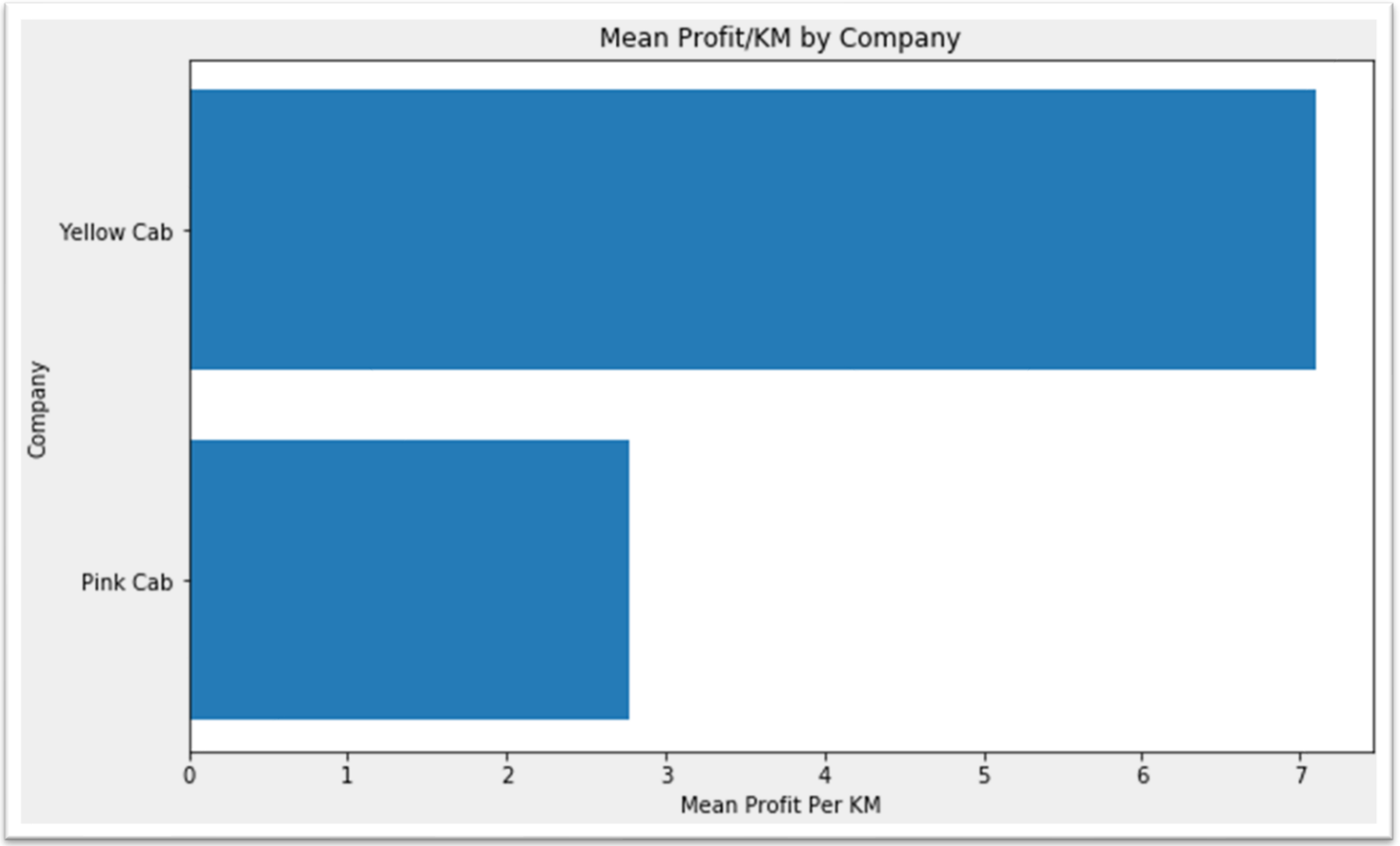
EDA - Cab data



This plot displays a box plot for each company, showing the distribution of the "Price Charged" variable. The box represents the interquartile range (IQR), with the median marked as a horizontal line inside the box. The whiskers extend to the minimum and maximum values within a certain range (usually 1.5 times the IQR). Any data points outside this range are considered outliers and are shown as individual points on the plot.

Result: That Yellow Cab Company has marginally higher prices than the Pink Cab Company.

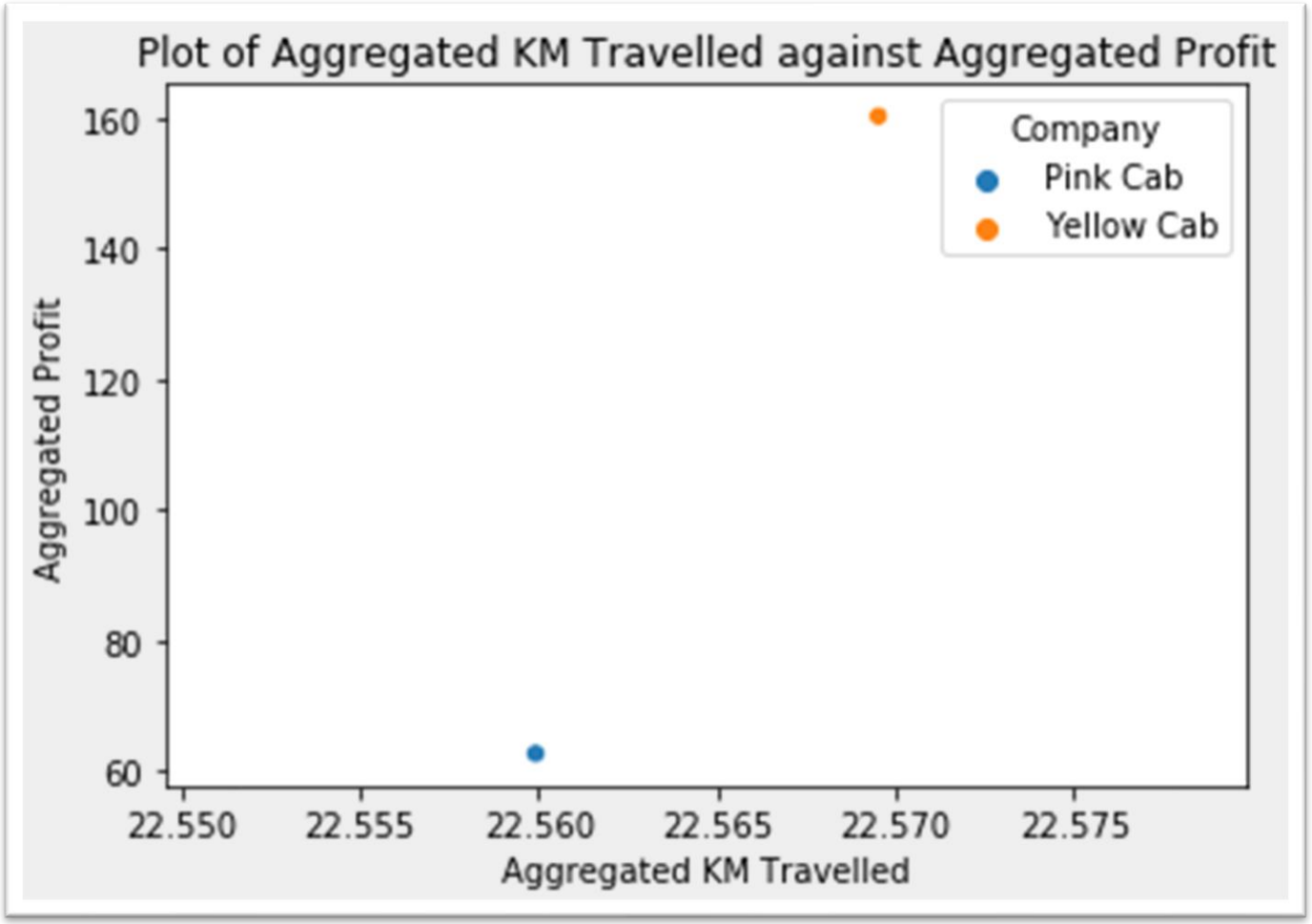
EDA - Cab data



This plot visualizes the mean profit per kilometer for each company in a horizontal bar plot, allowing you to compare and analyze the profitability efficiency of different companies in terms of profit earned per kilometer traveled.

Result: The Yellow Cab company has more profitability per KM!

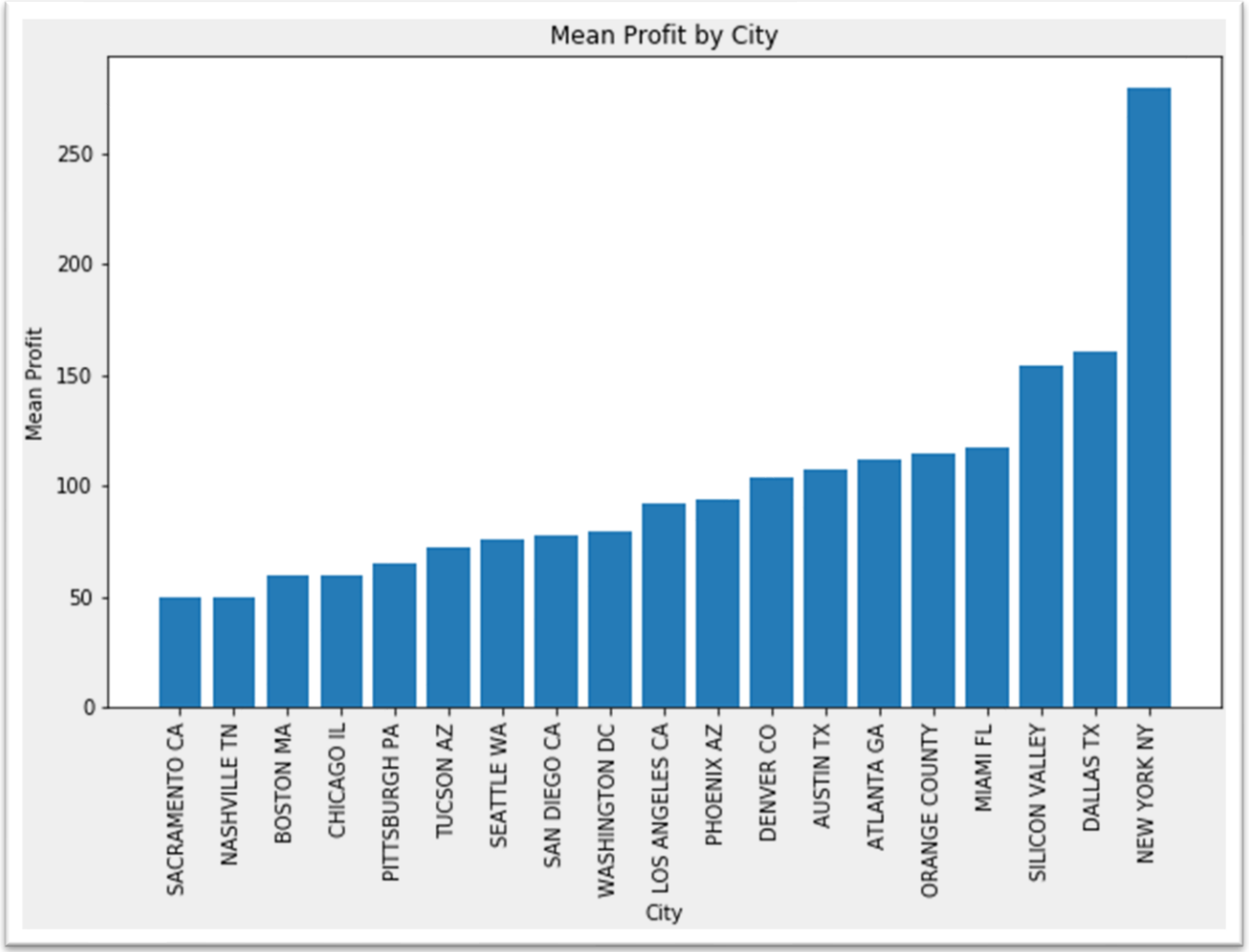
EDA - Cab data



This plot visualizes the relationship between the aggregated values of 'KM Travelled' and 'Profit' for each company in a scatter plot. The use of different colors and company names allows for easy identification and comparison of data points belonging to different companies.

Result: Yellow cab has a higher result again.

EDA - Cab data



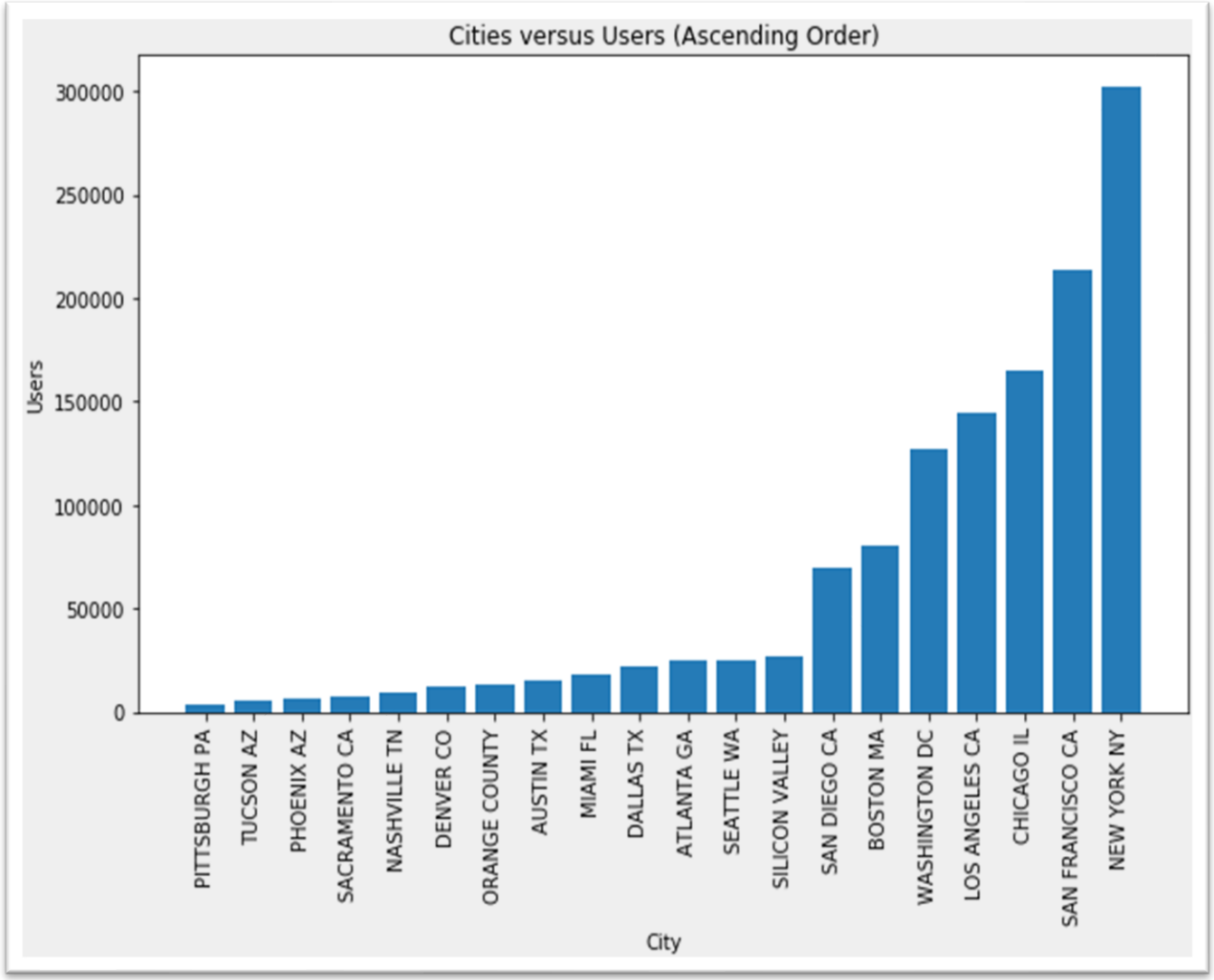
This visualizes the mean profit for each city in a bar plot, sorted in ascending order. The plot allows for an easy comparison of mean profits among different cities, providing insights into the profitability across various locations.

Result: The cab industry in New York gives more profit than all other cities, so the investment advice could be investing in cabs in New York.

EDA - City data

- Importing city dataset.
- Checking the data types and changing the Population and Users to float data type.
- Checking for null values. None were found.
- Creating a new column for User per Population.
  - User per Population = Users / Population
- Visualizing the number of users for each city in a bar plot.

EDA - City data



This visualizes the number of users for each city in a bar plot, with the cities sorted in ascending order of the number of users. The plot allows for easy comparison of user counts among different cities, providing insights into the distribution of users across various locations.

Result: New York has the most cab users among others. After New York, San Francisco, and Chicago are the next. This plot approves the city distribution plot from Cab dataset.

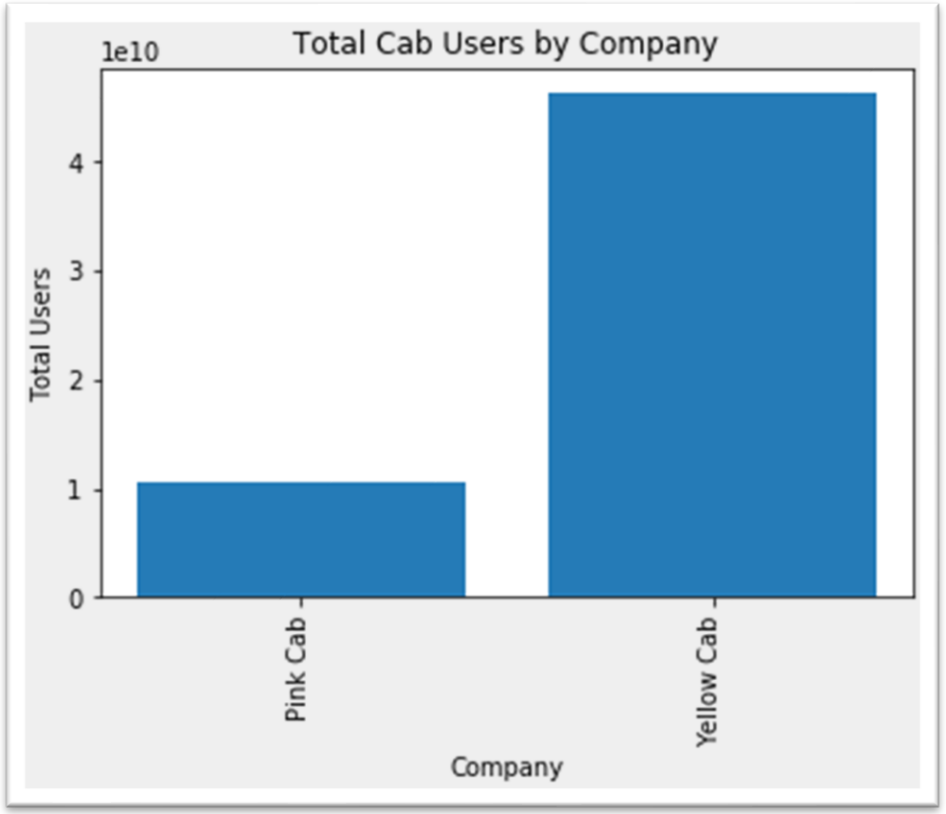
EDA - Customer & Transaction data

- Importing Transaction dataset.
- Checking the data types.
- Checking for null values. None were found.

- Importing Customer dataset.
- Checking the data types.
- Checking for null values. None were found.

EDA - Merged data

- Merging Cab & City datasets.



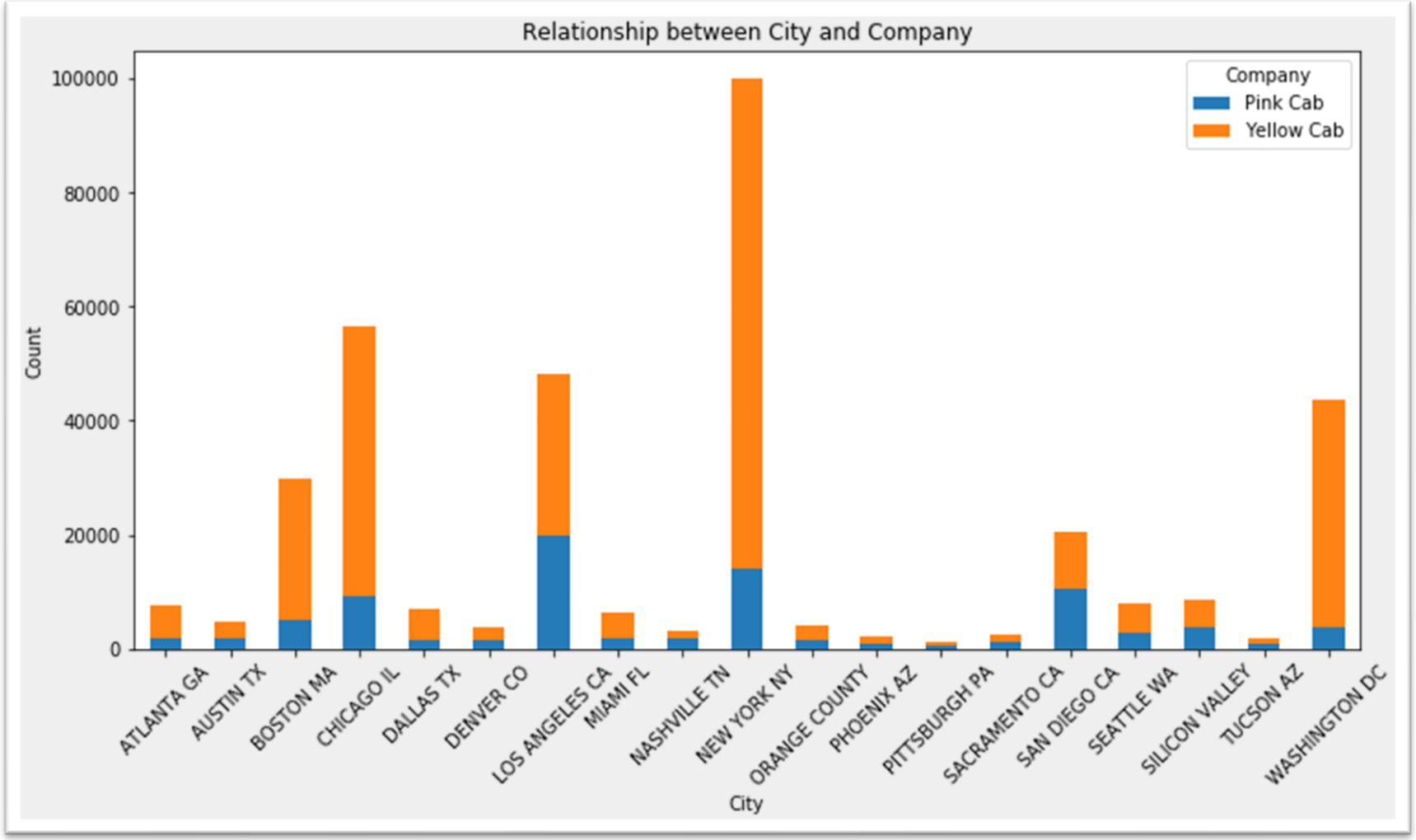
This visualizes the total number of cab users for each company in a bar plot. It provides insights into the distribution of cab users among different companies and allows for easy comparison of user counts. Additionally, it identifies the company with the maximum number of cab users, which can be useful for analyzing market dominance or customer preferences.

Result: Yellow cab company has considerably more users than Pink cab.

EDA - Merged data

- Merging Customer & Transaction datasets.
- Afterwards, the last two merged datasets were merged to make the master merged dataset.
- The mean age of both cab company users was calculated by the GroupBy function:
  - Pink Cab 35.322414
  - Yellow Cab 35.341112
  - Result: The average age of the users of both companies is the same (35).
- The mean income of both cab company users was calculated by the GroupBy function:
  - Yellow Cab 15045.669817
  - Pink Cab 15059.047137
  - Result: The average income of the users of both companies is the same (15000).
- The mode gender of both cab company users was investigated by the GroupBy function:
  - Pink Cab Male
  - Yellow Cab Male
  - Result: for both companies, most of the cab users are male.

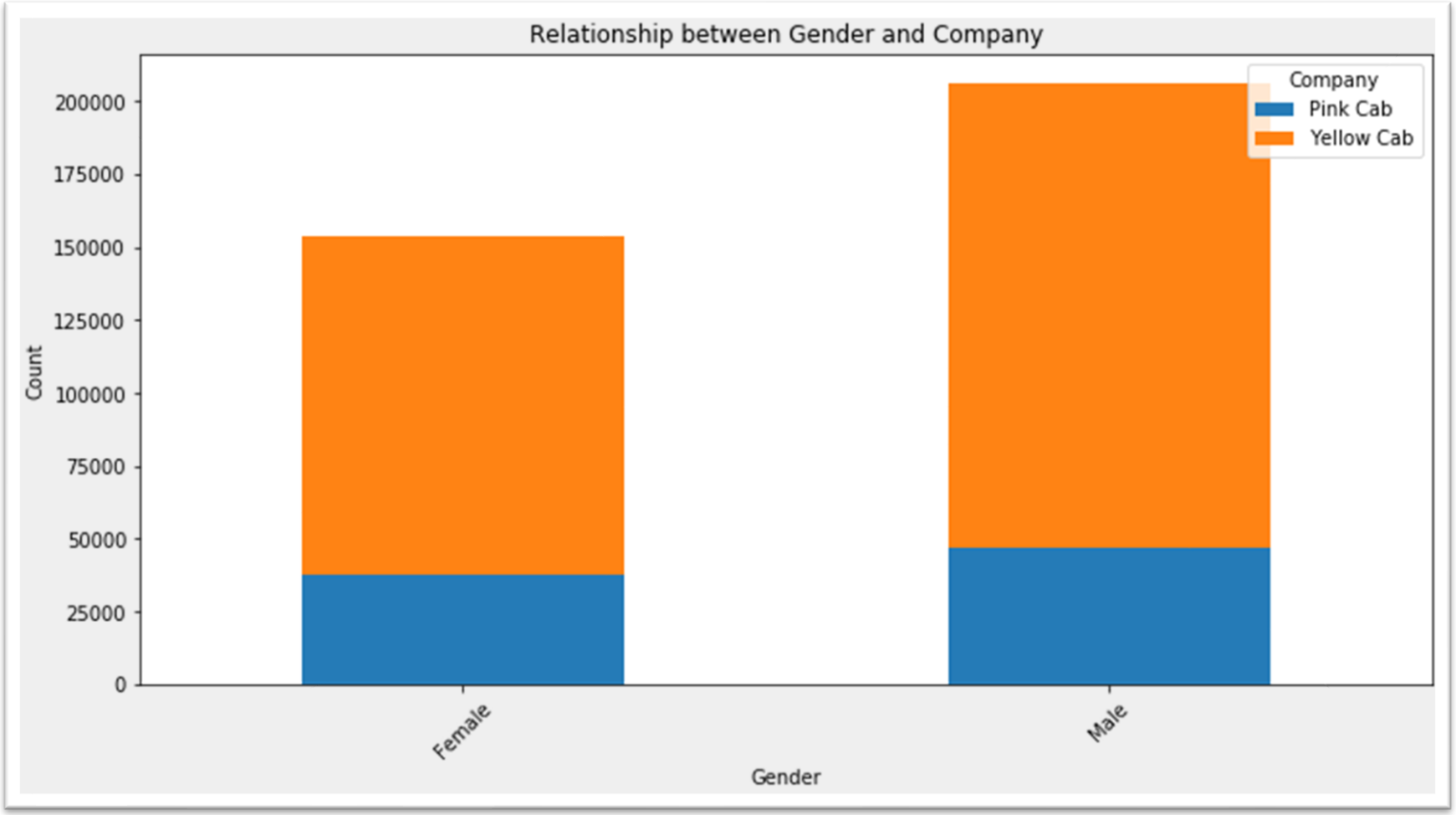
EDA - Merged data



This visualizes the relationship between cities and companies in terms of the count of occurrences. The stacked bar plot provides insights into the distribution of companies within each city, showing the relative presence of different companies in different cities.

Result: In most of the cities, the users of the Yellow Cab are more than the Pink Cab. For a few cities where the total number of cab users is less, the distribution of Yellow and Pink is ultimately equal.

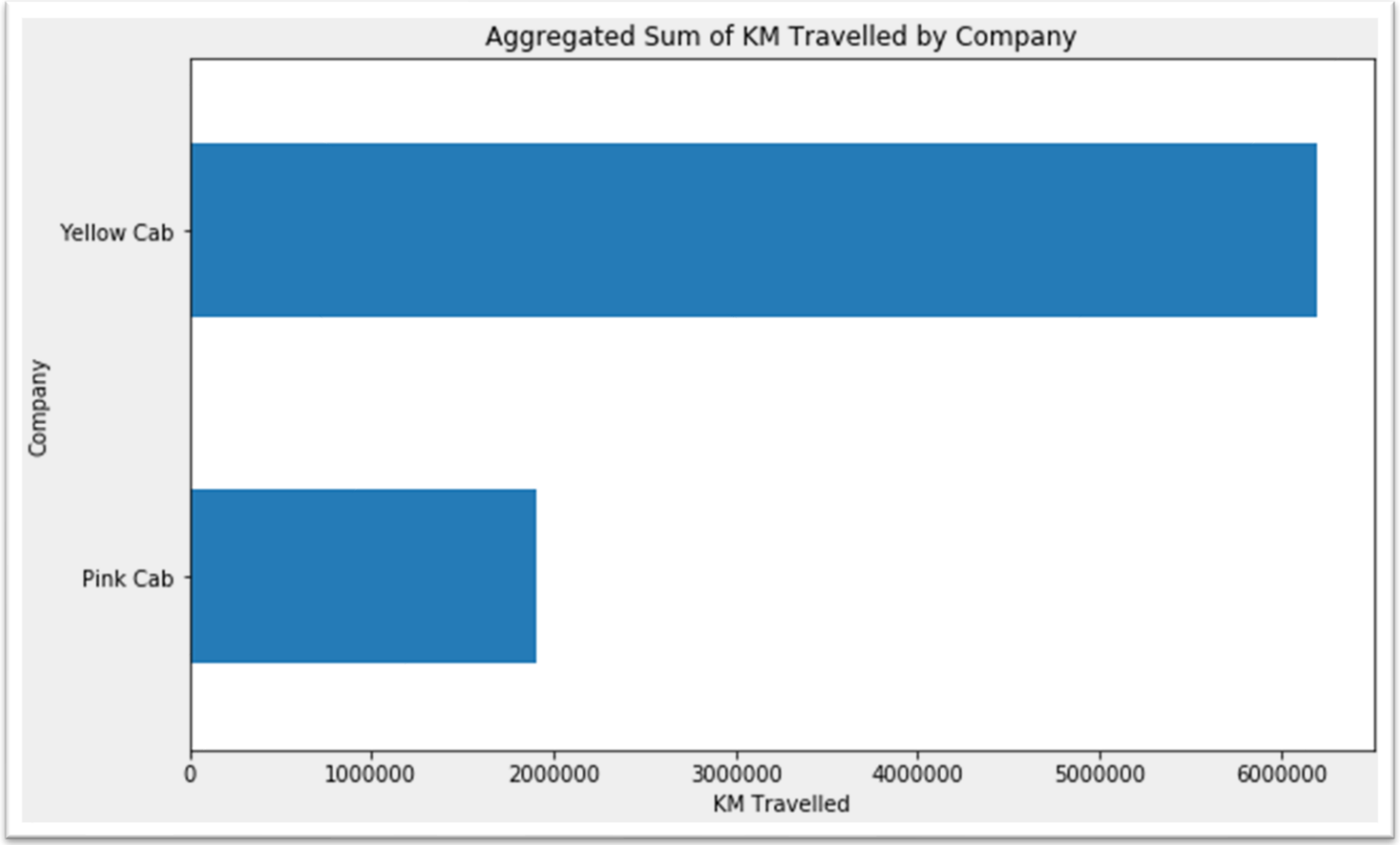
EDA - Merged data



This visualizes the relationship between gender and companies in terms of the count of occurrences. The stacked bar plot provides insights into the distribution of companies within each gender category.

Result: For both cab companies, the number of male users is considerably more than for females. Overall, the Yellow cab users are more than the Pink cab users.

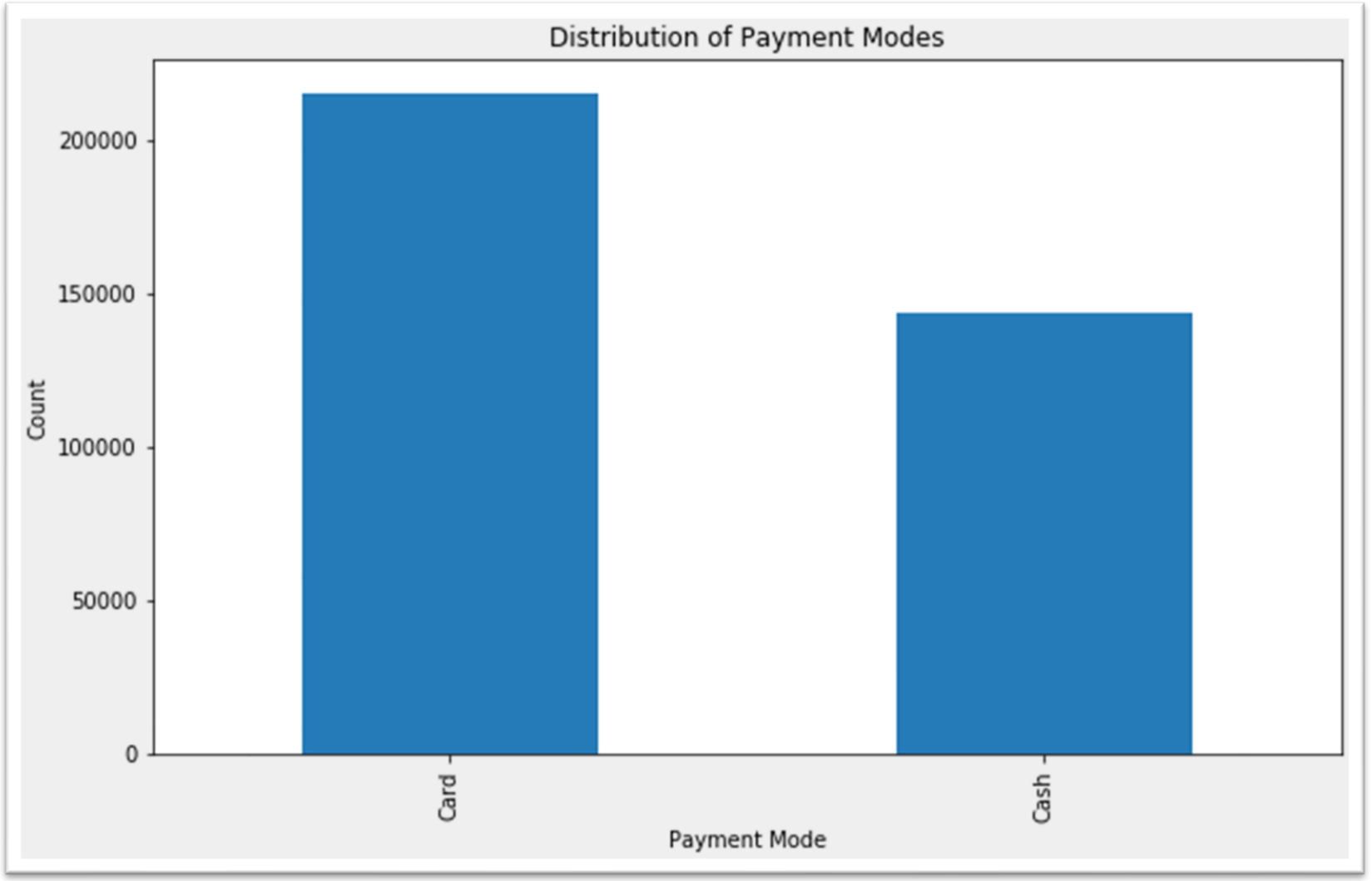
EDA - Merged data



This visualizes the aggregated sum of 'KM Travelled' by company. The horizontal bar plot provides a comparison of the total distance traveled by each company.

Result: The total Kilometers traveled by the Yellow cab company is much more than Pink cab.

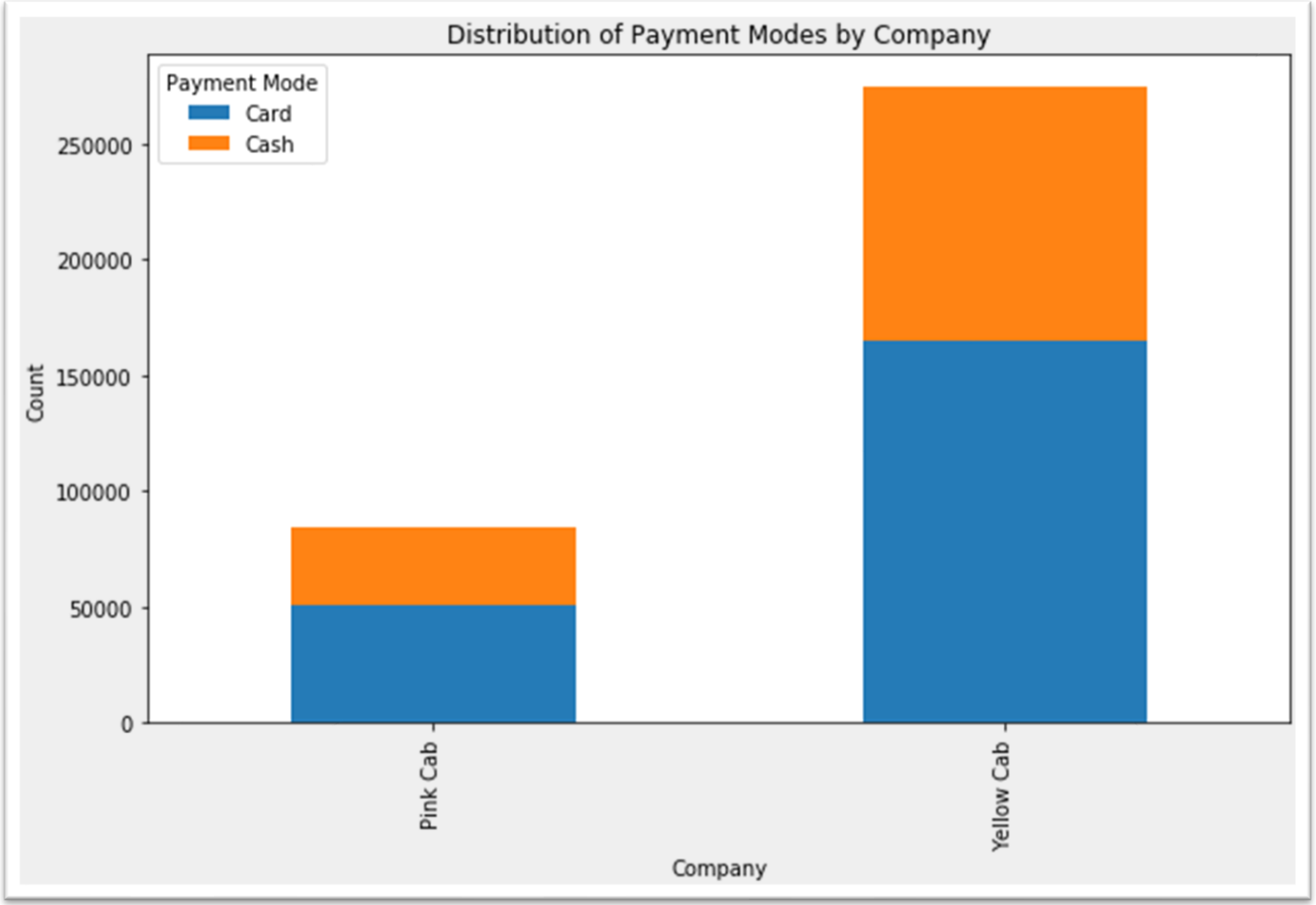
EDA - Merged data



The code helps you visualize the distribution of payment modes. The bar plot provides a clear comparison of the frequency or count of each payment mode.

Result: most of the cab users use the card as a payment method.

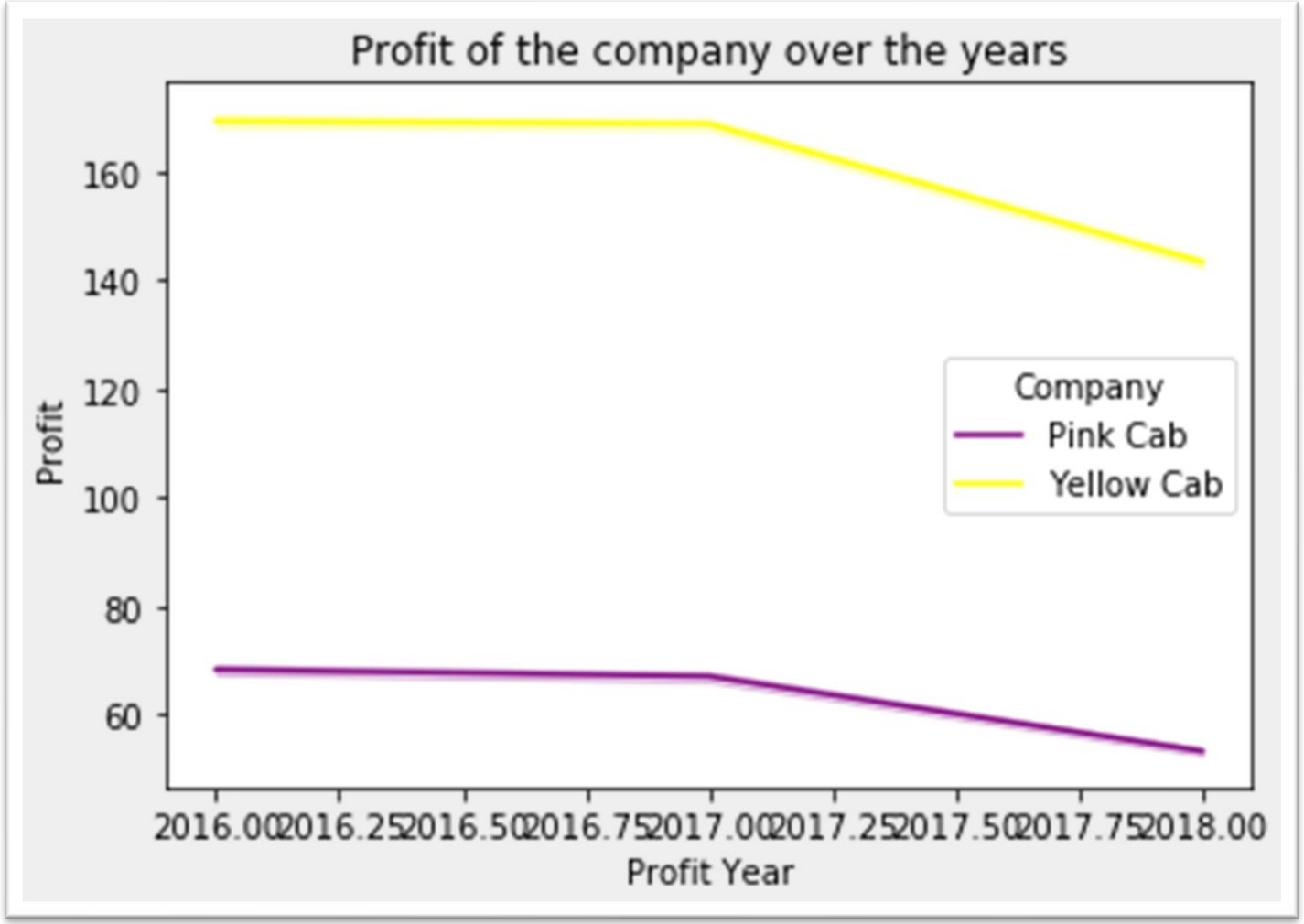
EDA - Merged data



This visualizes the distribution of payment modes by company. The stacked bar plot provides insights into the preferred payment modes for each company and allows for comparison between companies.

Result: Same result as the previous figure. most of cab users use cards instead of cash. The same pattern works for Yellow Cab.

EDA - Merged data



The resulting plot shows the trend of profit for each company over the years, with each company represented by a different colored line. It helps visualize how the profit of each company has evolved or changed over time.

Result: Profit of the both companies has reduced over the past few years.

Summary

- Through the analysis of the merged dataset, we have discovered several important findings, including the company with the maximum number of cab users, mean profits and profit per kilometer by company, mean profit by city, and user distribution across cities. These findings provide valuable information for strategic decision-making and business growth.
- The Yellow Cab Company has a higher value for all the above features.
- New York has the most cab users.
- Therefore, investing in Yellow Cab Company is recommended.

Recommendation

- **Targeted Marketing:** Utilize customer segmentation analysis to tailor marketing strategies and promotional campaigns for different customer segments. Focus on segments that contribute the most to the business in terms of revenue or number of rides. This can include personalized offers, loyalty programs, and targeted advertising to enhance customer satisfaction and increase customer retention.
- **Payment Options Optimization**: Optimize the available payment options based on the preferences identified in the payment mode analysis. Ensure a seamless and secure payment experience for customers by offering popular payment modes and exploring options for digital wallets or contactless payments. Monitor customer feedback and satisfaction to continuously improve the payment process.
- **Seasonal Demand Planning:** Leverage the insights from the analysis of travel patterns to optimize resource allocation and pricing strategies during peak demand periods. Identify the specific days of the week or months with higher demand and plan driver availability and fleet management accordingly. Implement dynamic pricing strategies to maximize revenue during high-demand periods.
- **Geographical Expansion:** Consider expanding operations to cities where specific companies have a lower presence or market share. Utilize the analysis of user distribution across cities to identify potential growth opportunities. Conduct market research and feasibility studies to assess the viability and potential success of expanding into new cities.
- **Customer Experience Enhancement**: Continuously monitor and analyze customer feedback and satisfaction metrics to identify areas for improvement. Focus on enhancing the overall customer experience by providing reliable and comfortable rides, ensuring prompt customer support, and offering personalized services based on customer preferences.
- **Competitor Analysis:** Conduct a thorough analysis of competitors in the market to understand their strategies, market share, and customer base. Identify areas

Thank You