
# Exploratory Data Analysis

## G2M CASE STUDY

**18th August 2023**

# Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

# Executive Summary

XYZ is a private company in the United States. Due to the spectacular rise in the Cab Industry in recent years and the presence of numerous significant players in the market, it is planning to invest in the Cab Industry, and as part of their Go-to-Market (G2M) strategy, they want to understand the market before making a final choice.

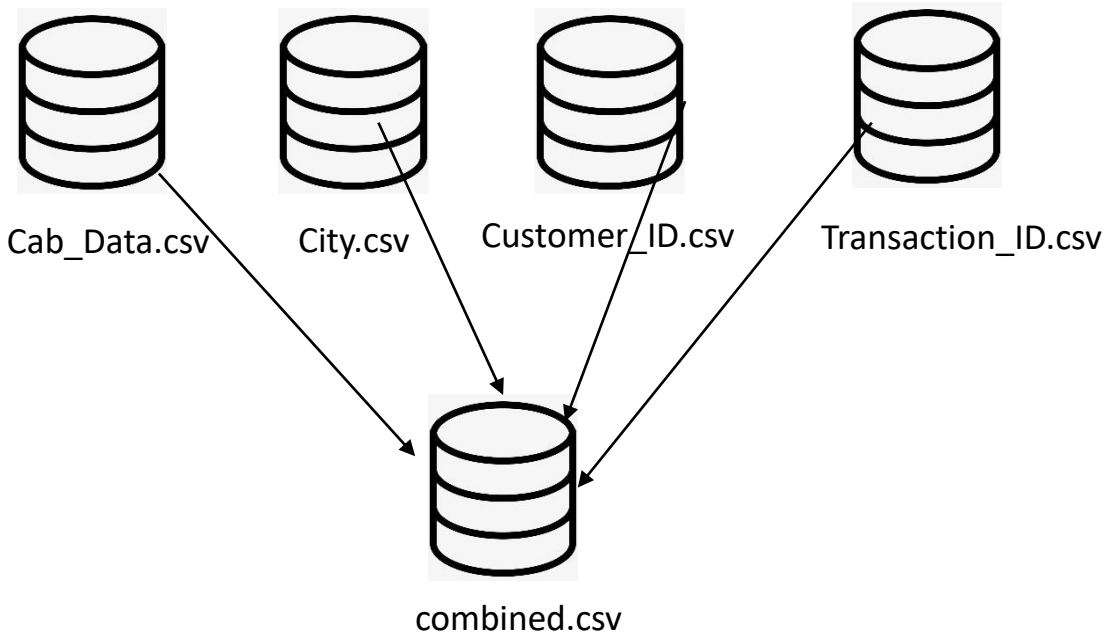
Objective: Provide insights into the market and identify the best cab company to invest in.

This analysis has been divided into the following parts:

1. Exploratory Data Analysis (EDA)
2. Data Visualization and Relationships
3. Statistical Hypothesis Testing
4. Data Interpretation and Conclusions

# Data Exploration

- 20  features (including 3 derived features)
- Timeframe of the data: 2016-01-31 to 2018-12-31
- Total data points : 355,032




**Assumptions:**

- The variables in the dataset follow a normal distribution for statistical tests.
- The assumption of homoscedasticity holds, indicating equal error variances in statistical analyses.
- Data preprocessing includes removal of missing values and outliers.
- Dataset is complete and representative of the specified time period.

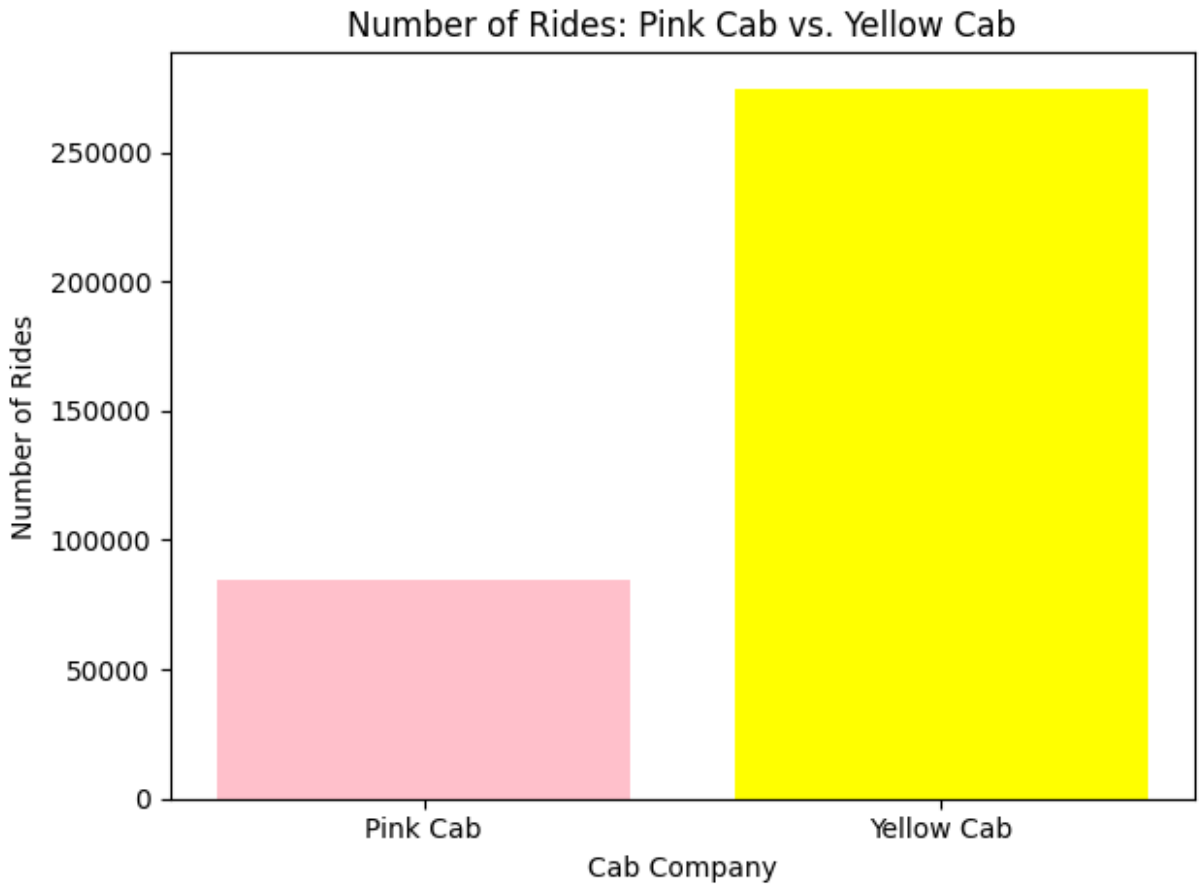
# Approach

- Implement deduplication validation to ensure data accuracy using Pandas drop_duplicates() method.
- One-hot encode categorical variables before conducting statistical tests.
- Conduct exploratory data analysis (EDA) to understand data distribution and relationships.
- Perform statistical tests for hypothesis validation.
- Utilize data visualization techniques to present insights effectively.

# Exploratory Data Analysis (EDA)

The next few slides depict the EDA techniques used, including the results and insights derived from them.

# Market Share




Two-Sample Proportion Test Results:

Z-statistic: -448.14217516061643
P-value: 0.0

Reject the null hypothesis: There is a significant difference in market share between Pink Cab and Yellow Cab.

# Profits Comparison

T-Test Results:

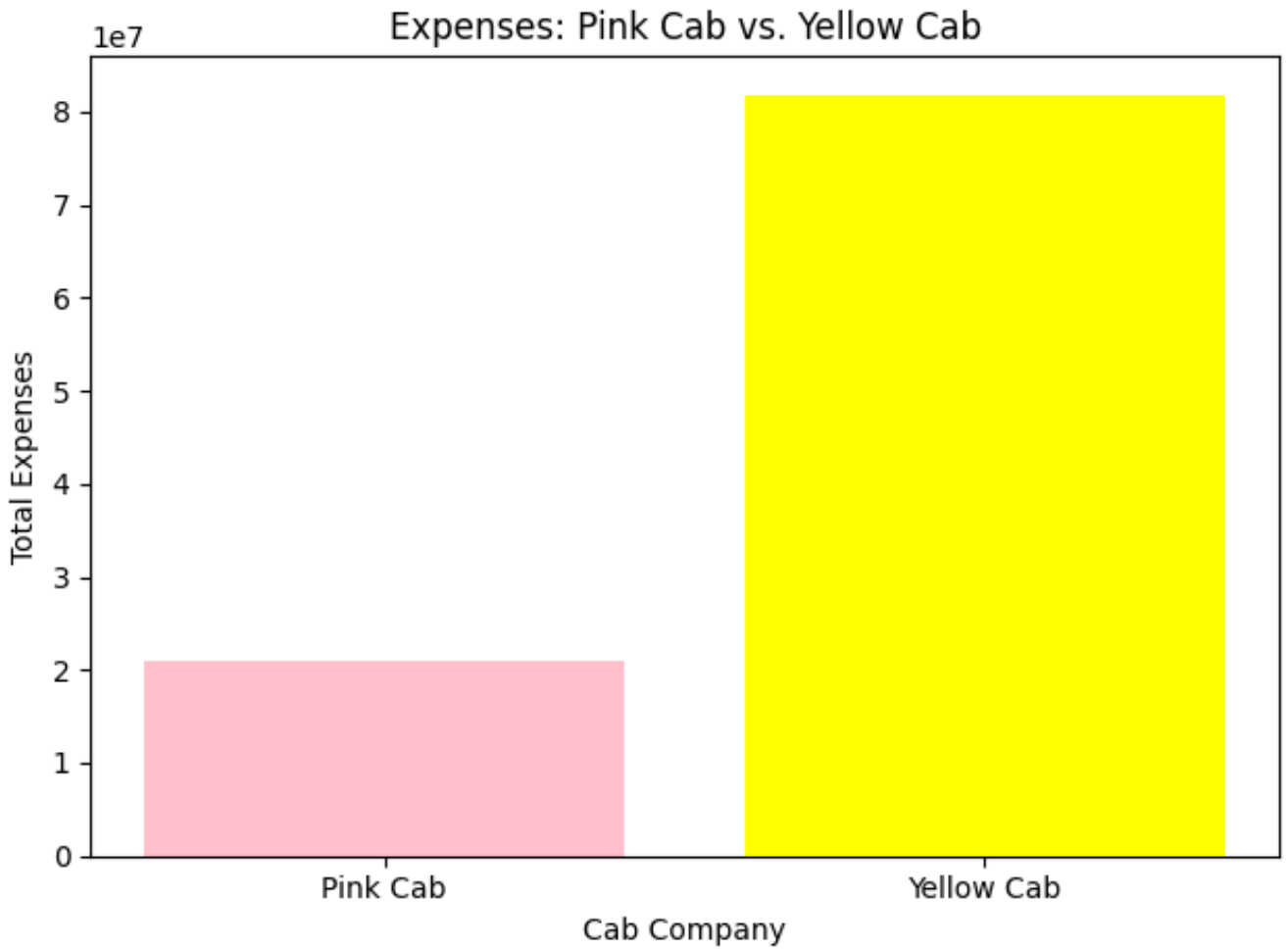
T-Statistic: -230.99551452746311

P-value: 0.00

Reject the null hypothesis: There is a significant difference in profits between Pink Cab and Yellow Cab.

**Yellow Cab had higher average profits.**

# Expenses Comparison



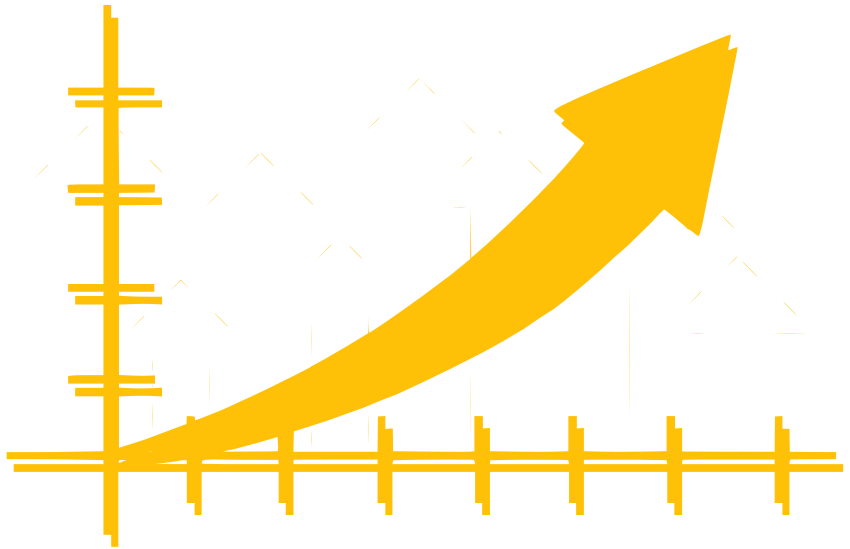

Difference in expenses:
**60812591**

Percentage difference:
**119%**

The expenses of Pink Cab are significantly less than Yellow Cab.

# Profit Under Equal Expenses

- A statistical test revealed that the profits of the Pink Cab company would increase significantly if its expenses were equal to that of Yellow Cab company.



Paired t-test Results:

T-Statistic: 533.3993527784962

P-value: 0.0

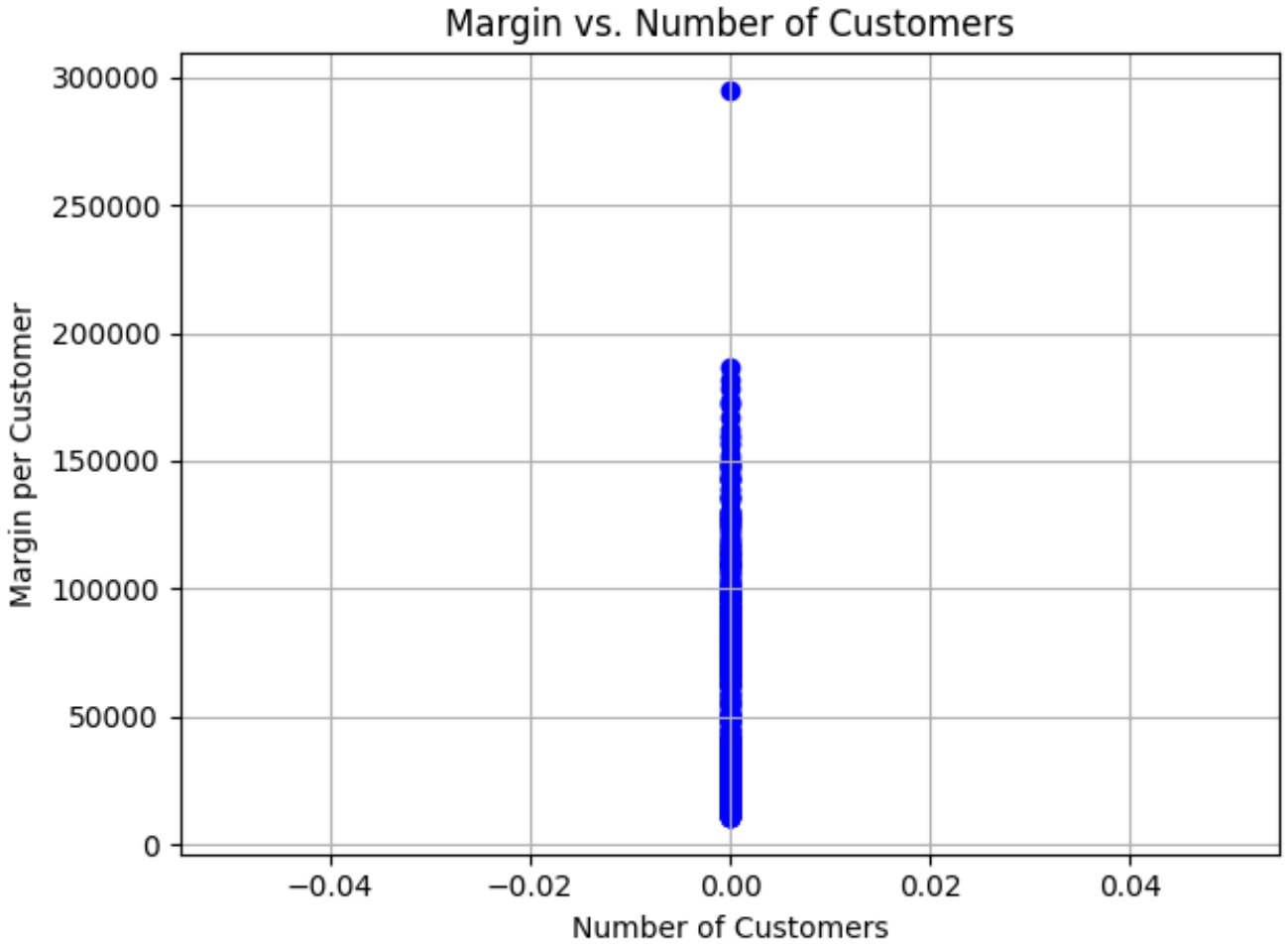
Reject the null hypothesis: **The profits of Pink Cab will increase after 119% increase in expenses**.

# Profit Comparison If Expenses Were Equal

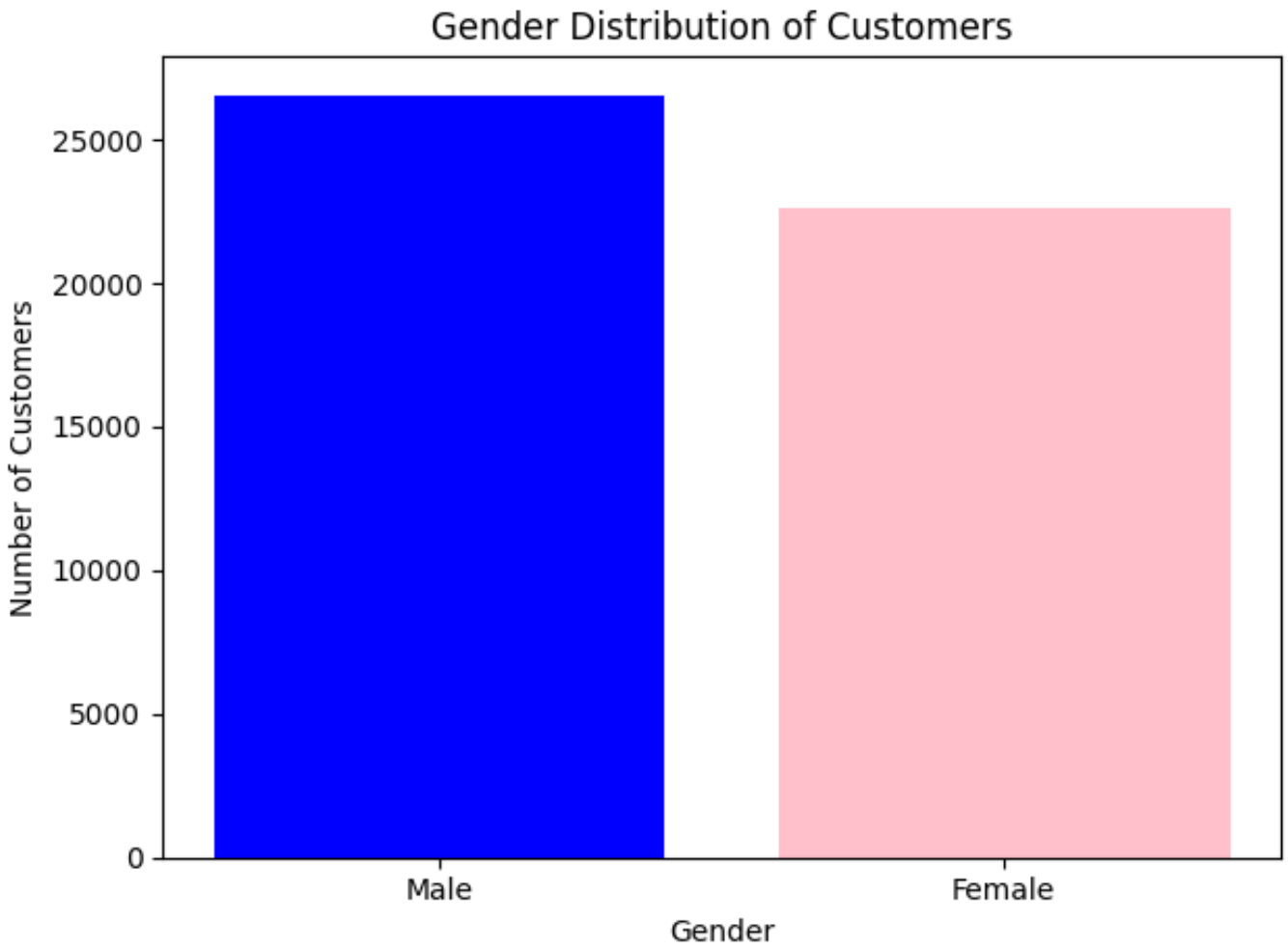
- However, even after the predicted profit increase of the Pink Company if expenses were equal, the current profits of the Yellow Company are greater than the predicted value by $42707020.

```
Predicted Pink Cab Profits: $1313352.9492900013

Current Yellow Cab Profits: $44020373.17080001

Difference: $42707020.22151001
```

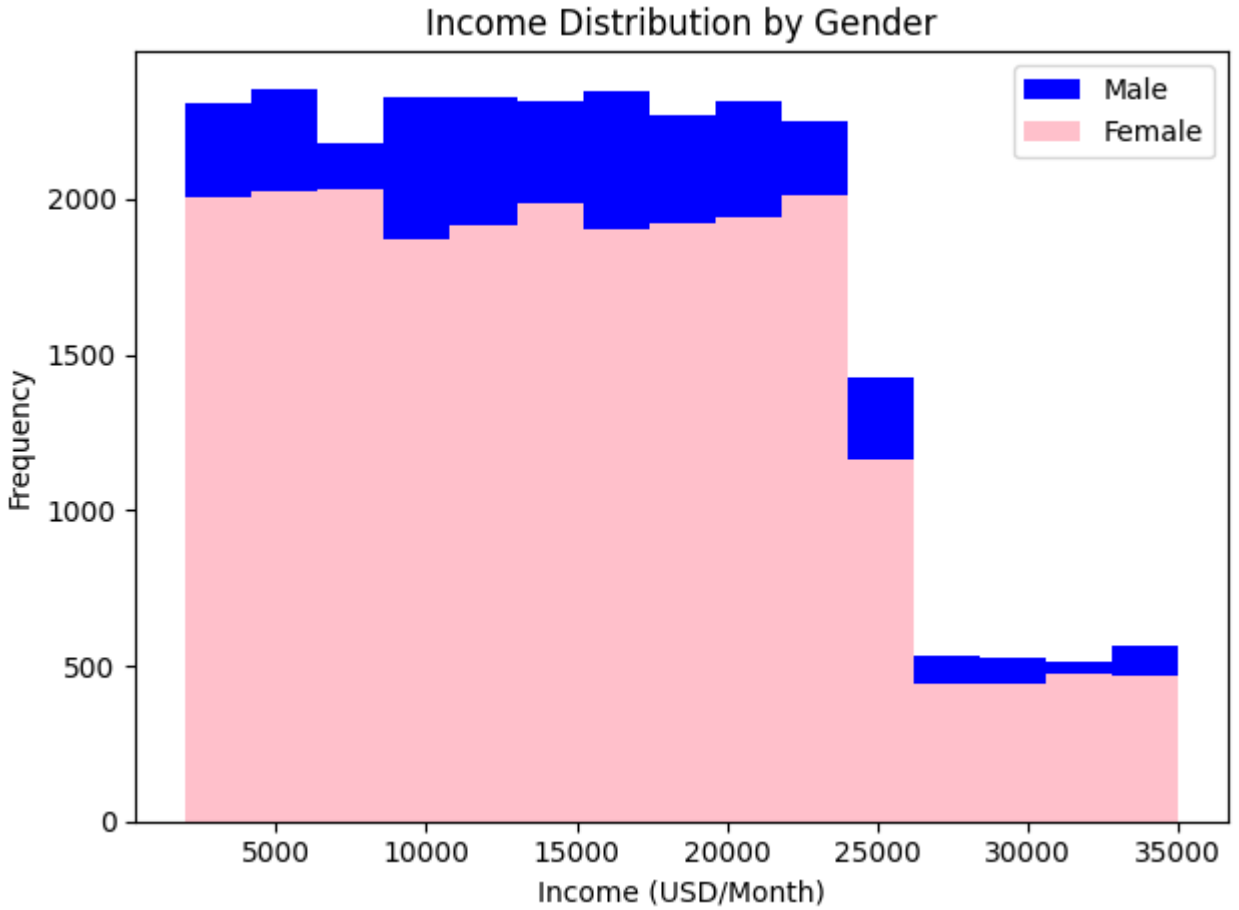
# Relationship Between Margin & Number of Customers



# Gender Distribution of Customers in the Cab Industry (Both Companies)



# Customers' Income Distribution by Gender



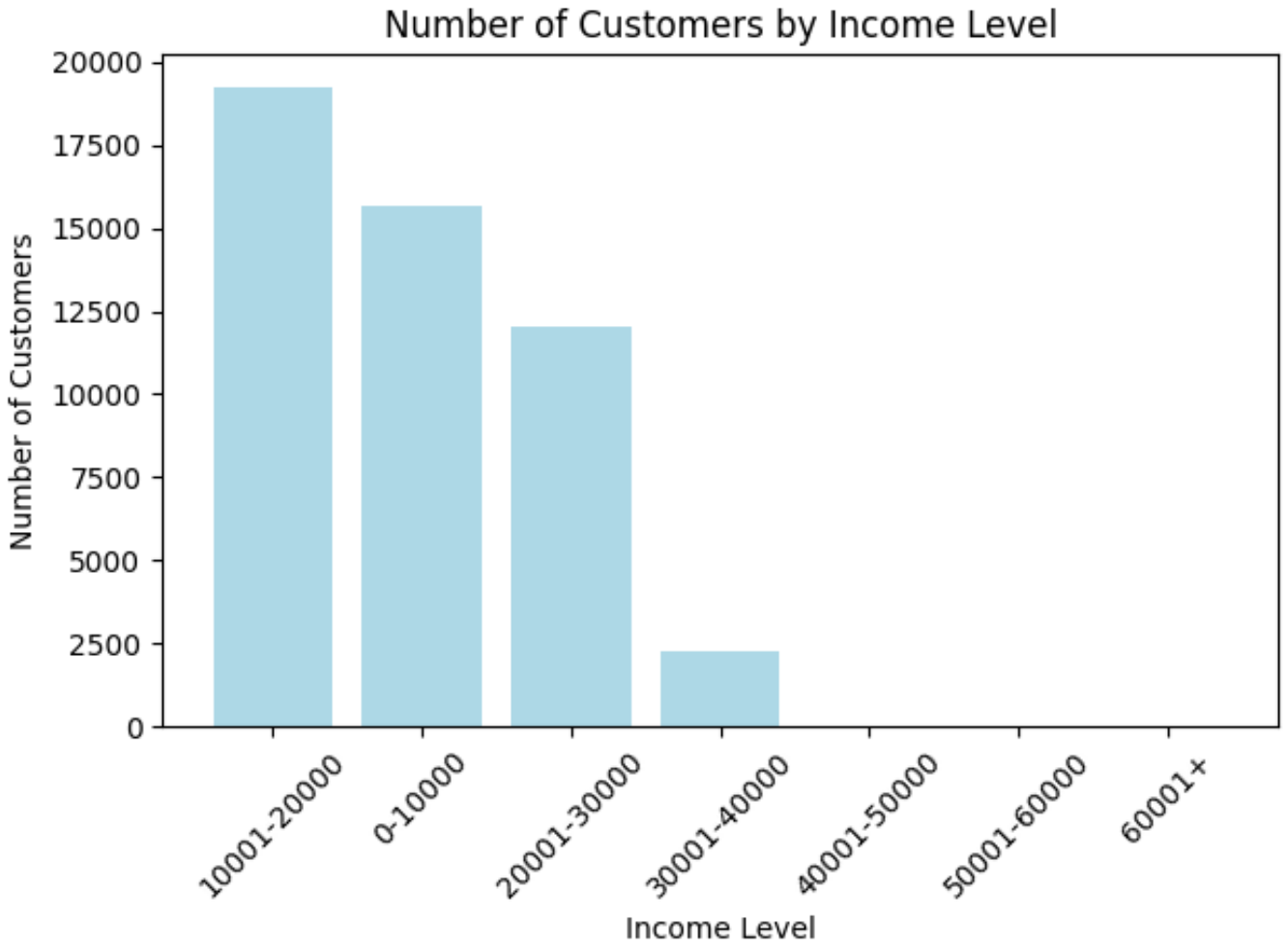

T-Test Results:

T-Statistic: 0.7557972721208138

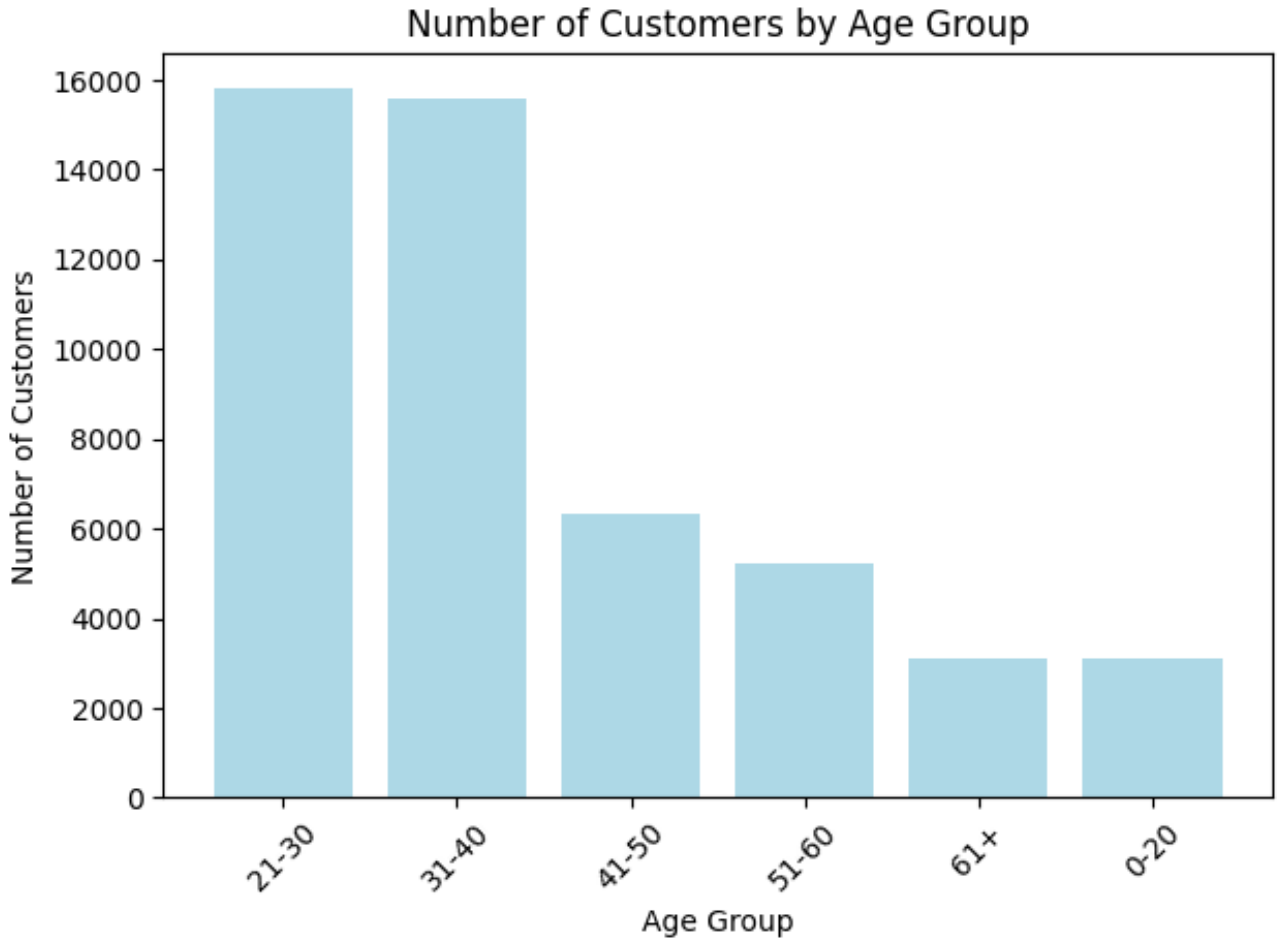
P-value: 0.44977437321752045

Fail to reject the null hypothesis: No significant effect of gender on income.

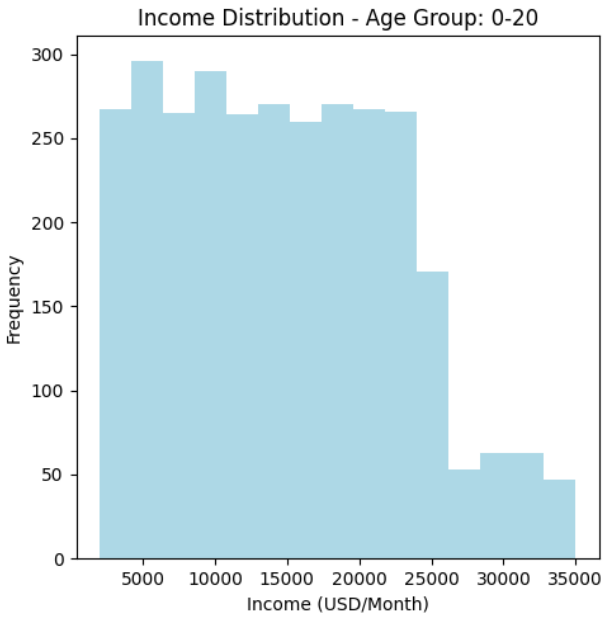
# Income Distribution of Customers in the Cab Industry (Both Companies)

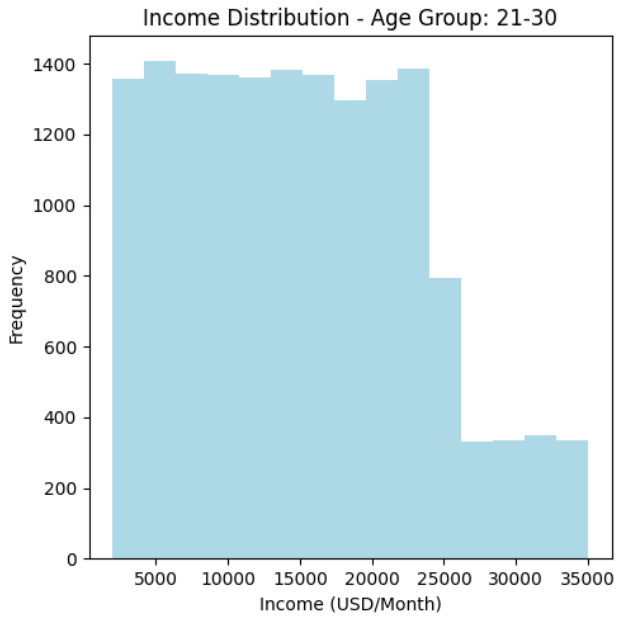


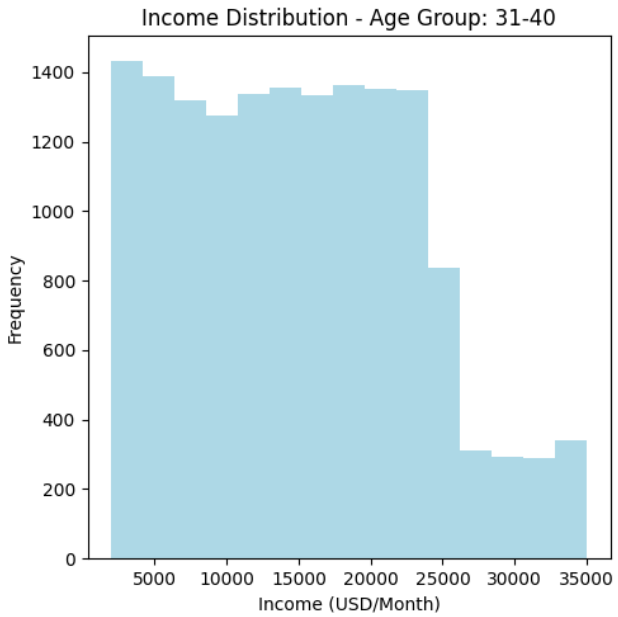
# Age Distribution of Customers in the Cab Industry (Both Companies)

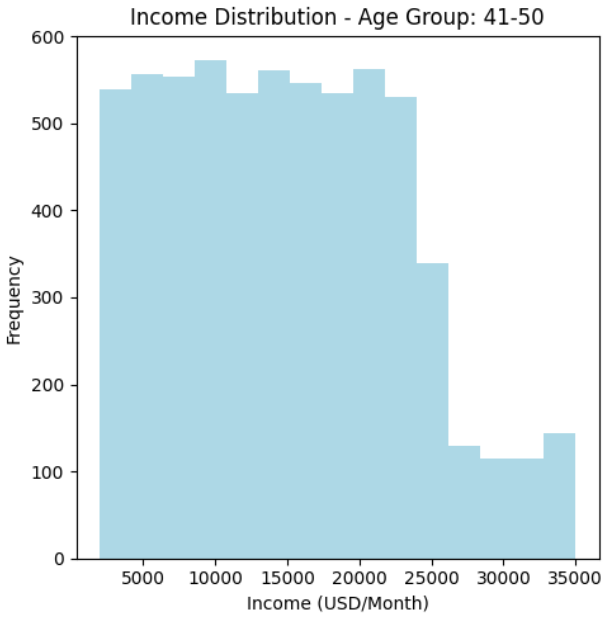


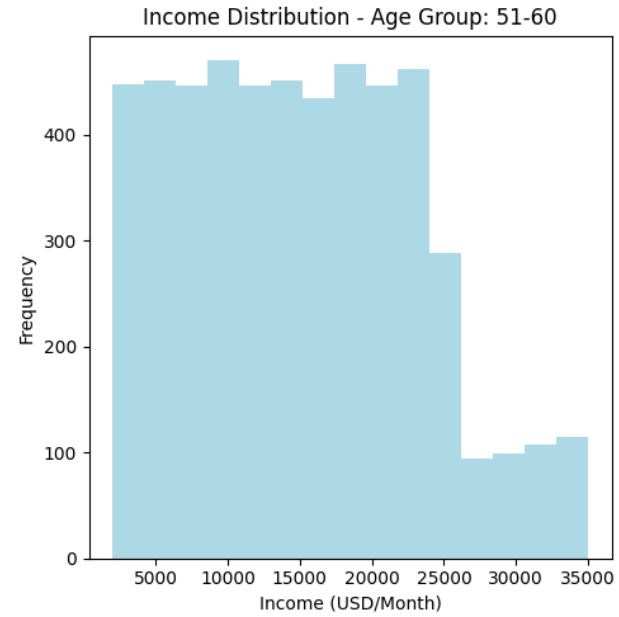
# Customers’ Income Distribution by Age Groups

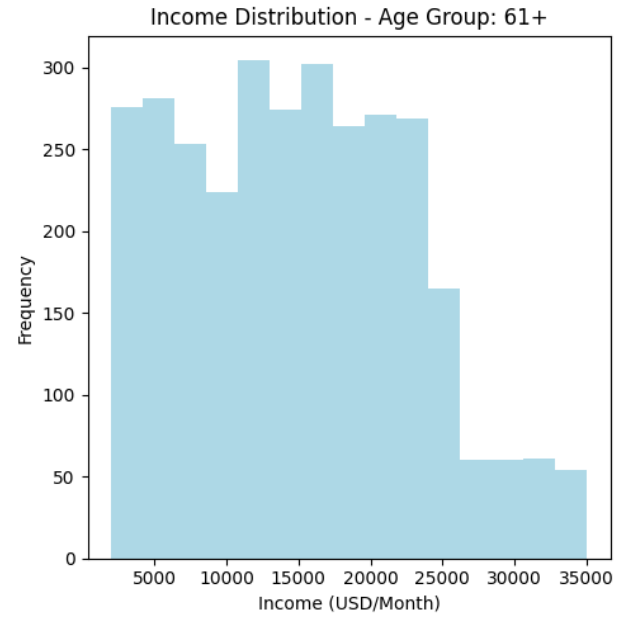


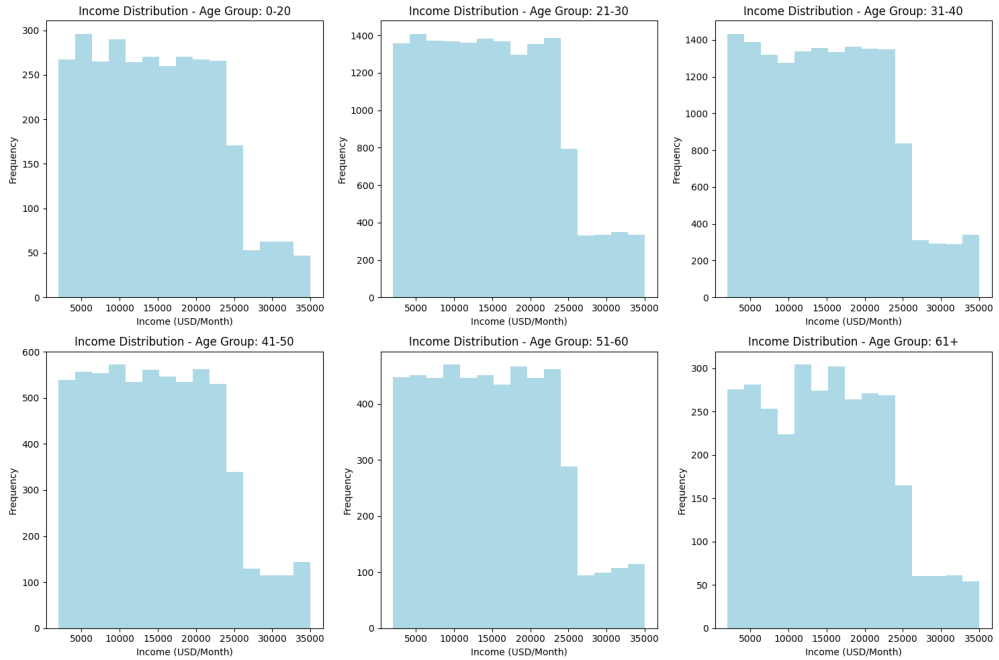
# Customers' Income Distribution by Age Groups

ANOVA Results:

F-statistic: 0.595463697918255

P-value: 0.7034827226856755

Fail to reject the null hypothesis: No significant impact of age on income level.

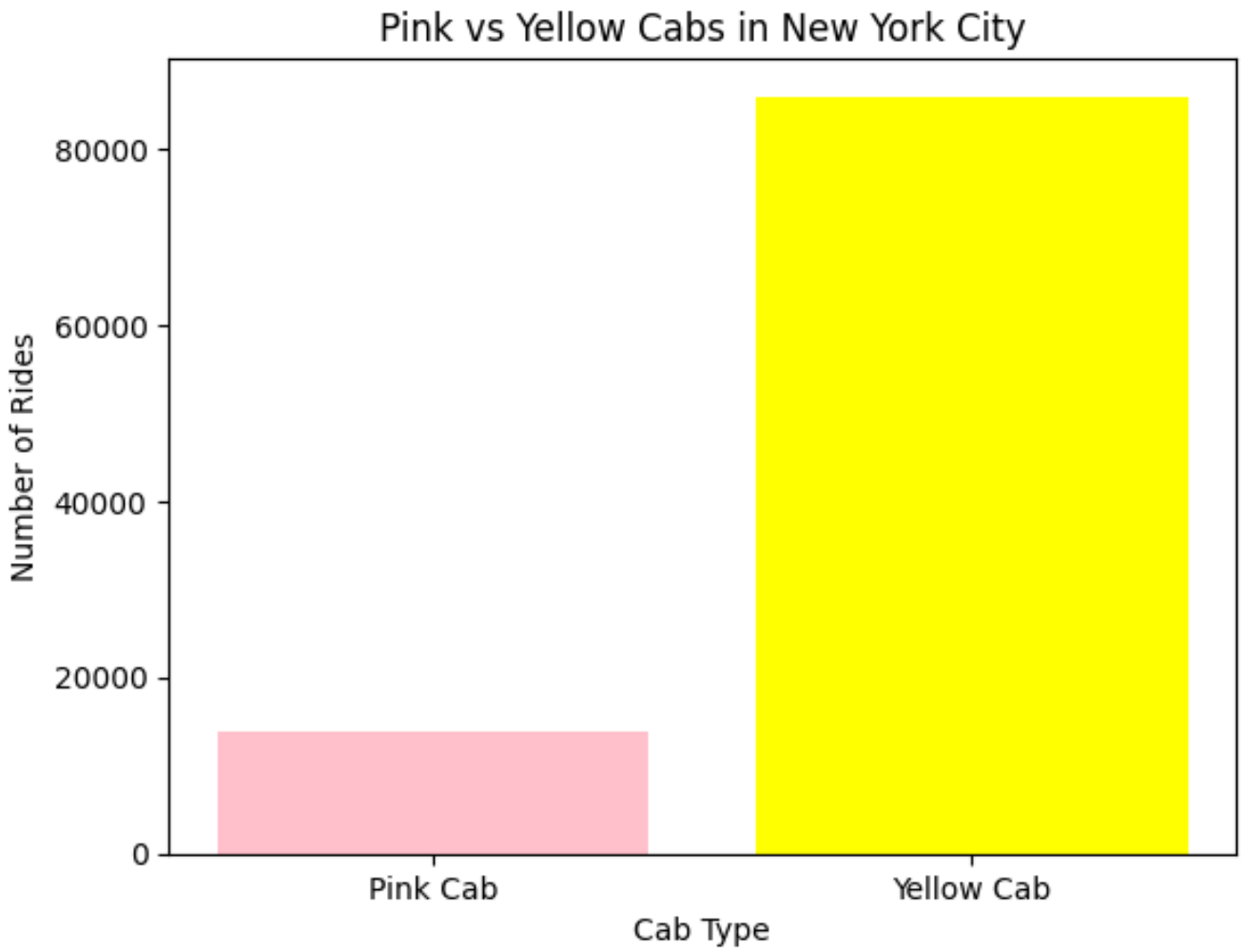


# Profitability Across Cities

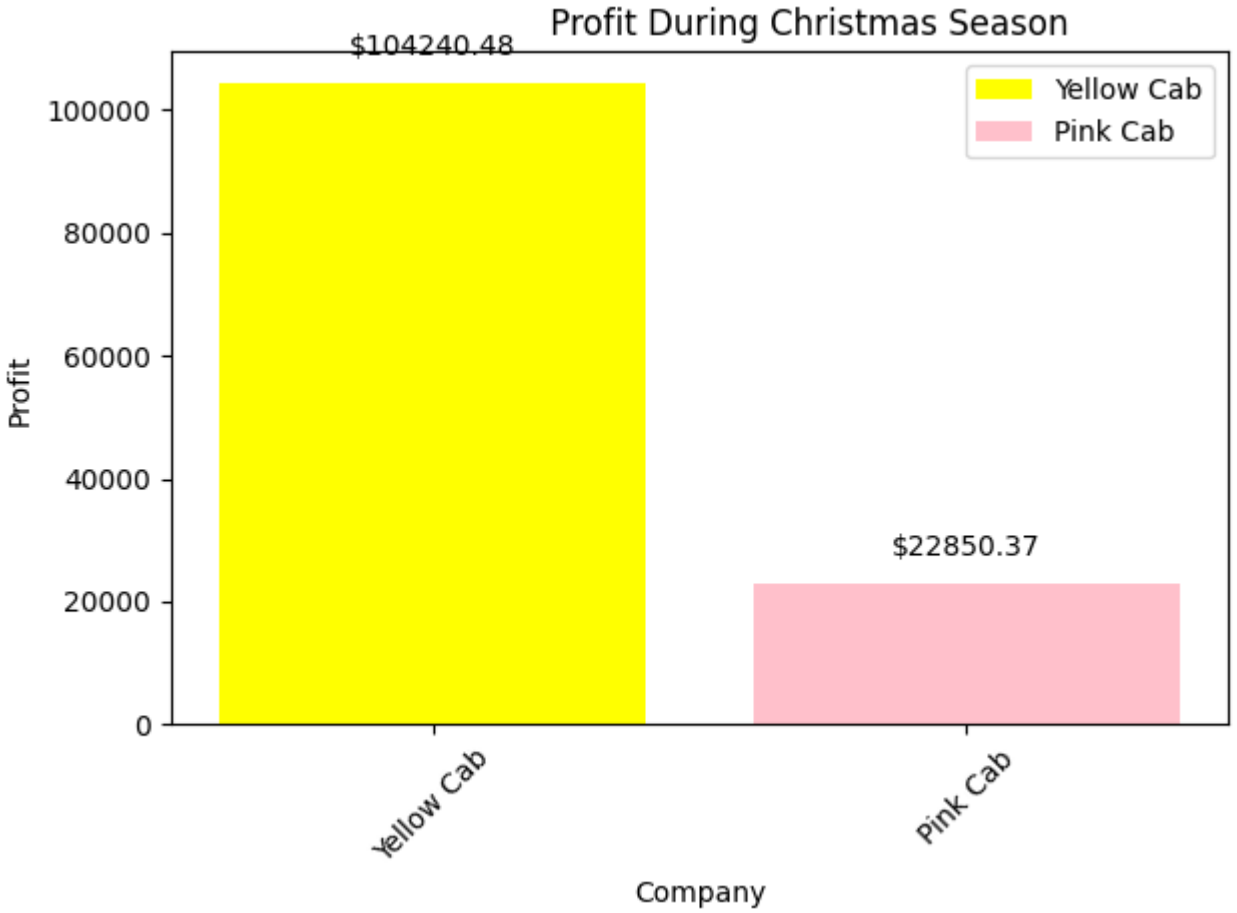
```
ANOVA Results:
F-statistic: 9528.597969224551
P-value: 0.00
Reject the null hypothesis: There
is a significant difference in
profitability across different
cities.
```

- Utilized the Python function idxmax() to pinpoint NEW YORK NY as the most profitable city, showcasing an impressive mean profit of $279.95

# Proportion of Pink vs Yellow Cabs in the Most Profitable City



# Seasonal Profit Comparison




- The company with the maximum profit during Christmas is Yellow Cab with a profit of $104240.48.

# EDA Summary

In the Data Exploration phase, datasets were examined, and key features summarized. Assumptions were made regarding data distribution and terminology. Diagrammatic representations were used to highlight relationships between variables.

Key findings from the analysis include:

1. **Profitability Analysis:** Yellow Cab outperforms Pink Cab in terms of profitability, exhibiting a significantly higher average profit. Despite Pink Cab's lower expenses, adjusting for equal expenses still shows Yellow Cab maintaining superior profitability.
2. **City-wise Profits:** Different cities exhibit varying profitability levels. Notably, New York City stands out with the highest profits for all cabs. The prevalence of Yellow Cabs in New York City indicates potential for sustained profitability.
3. **Market Share:** Yellow Cab commands a larger market share compared to Pink Cab, indicating a stronger presence and market positioning. This difference suggests that Yellow Cab is better poised for growth and sustained performance.
4. **Seasonal Profits:** Yellow Cab demonstrates higher profits during Christmas, indicating its ability to capitalize on peak seasons. Its profit during this period notably exceeds Pink Cab's performance.

The analysis involved a comprehensive range of statistical tests, including t-tests, ANOVA, and proportions z-test, to validate findings and draw robust conclusions.

# Recommendations

It is recommended to invest in yellow cab due to analysis based on the following points:

•**Profits:** There is a significant difference in profits between Pink Cab and Yellow Cab. Yellow Cab has higher average profits.

•**Cost and Profit Prediction:** Although the pink cost had 119% less expenses than the Yellow Cab, increasing the expenses and forecasting the increase in profits after equal expenses still revealed they would be lower than Yellow Cab's current profits by $42707020.

•**City-wise Profits and Distribution:** There is a significant difference in profitability across different cities, with New York City having the highest profits for all cabs. There are more Yellow Cabs than Pink Cabs in New York City, increasing their likelihood for sustained profitability.

•**Market Share:** There is a significant difference in market share between Pink Cab and Yellow Cab, with Yellow Cab having a noticeably higher share of the total rides. This difference suggests that Yellow Cab has a stronger presence in the market compared to Pink Cab.

•**Seasonal Profits:** The company with the maximum profit during Christmas is Yellow Cab with a profit of $104240.48, notably higher than the Pink Cab.

In conclusion, the analysis of various aspects of the cab industry leads to the confident recommendation of investing in Yellow Cab.

# Thank You