Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

G2M insight for Cab Investment firm

18-Apr-2024

Team : NLP

Location : Jordan

By:Mohammad Bani Abed Alghani

Agenda

- Executive Summary
- Problem Statement
- Approach
- EDA
- EDA Summary
- Recommendations

Description :

- XYZ is a private equity firm in US. Due to remarkable growth in the Cab Industry in last few years and multiple key players in the market, it is planning for an investment in Cab industry.

- Provide actionable insights to help XYZ firm in identifying the right company for making investment.
- Cab Companies:

1. Yellow Cab
2. Pink Cab

- The Analysis include :

1. Data Understanding,
2. Data Visualization,
3. Creating multiple hypothesis

Data Preparation:

- There are 4 datasets:

1. Cab_Data.csv – this file includes details of transaction for 2 cab companies.

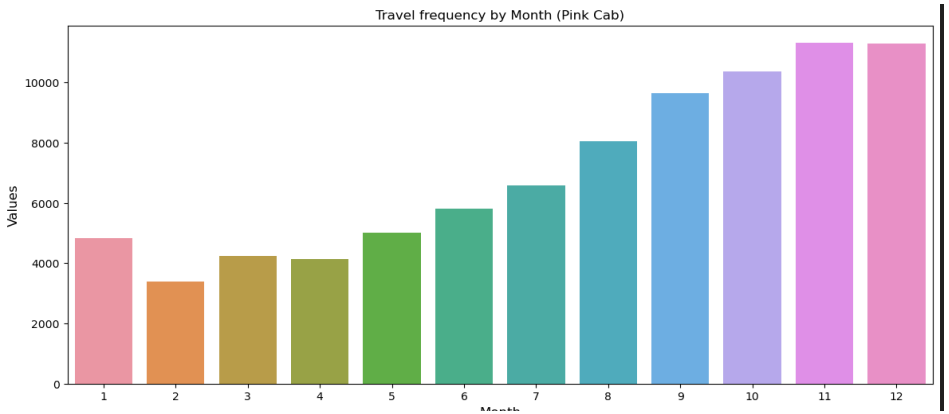
2. Customer_ID.csv – this is a mapping table that contains a unique identifier which links the customer's demographic details.

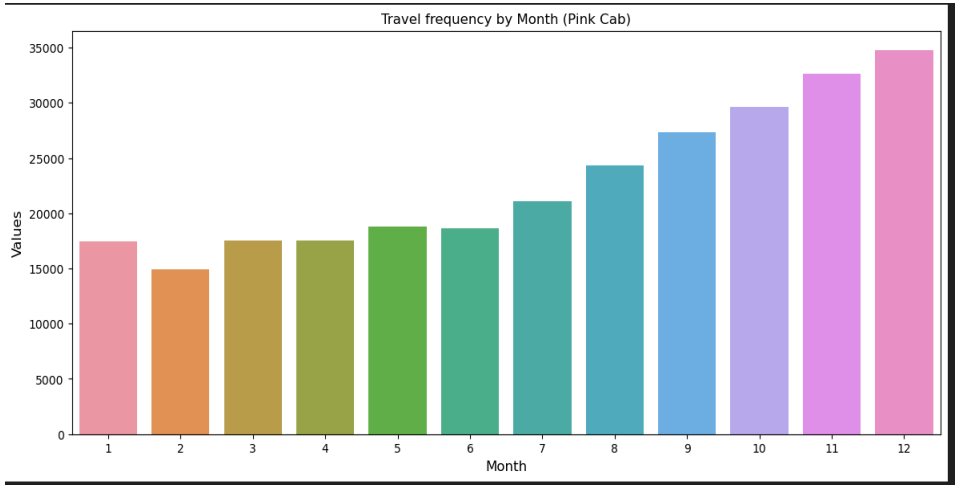
3. Transaction_ID.csv – this is a mapping table that contains transaction to customer mapping and payment mode.

4. City.csv – this file contains list of US cities, their population and number of cab users

EXPLORATORY DATA ANALYSIS (EDA)

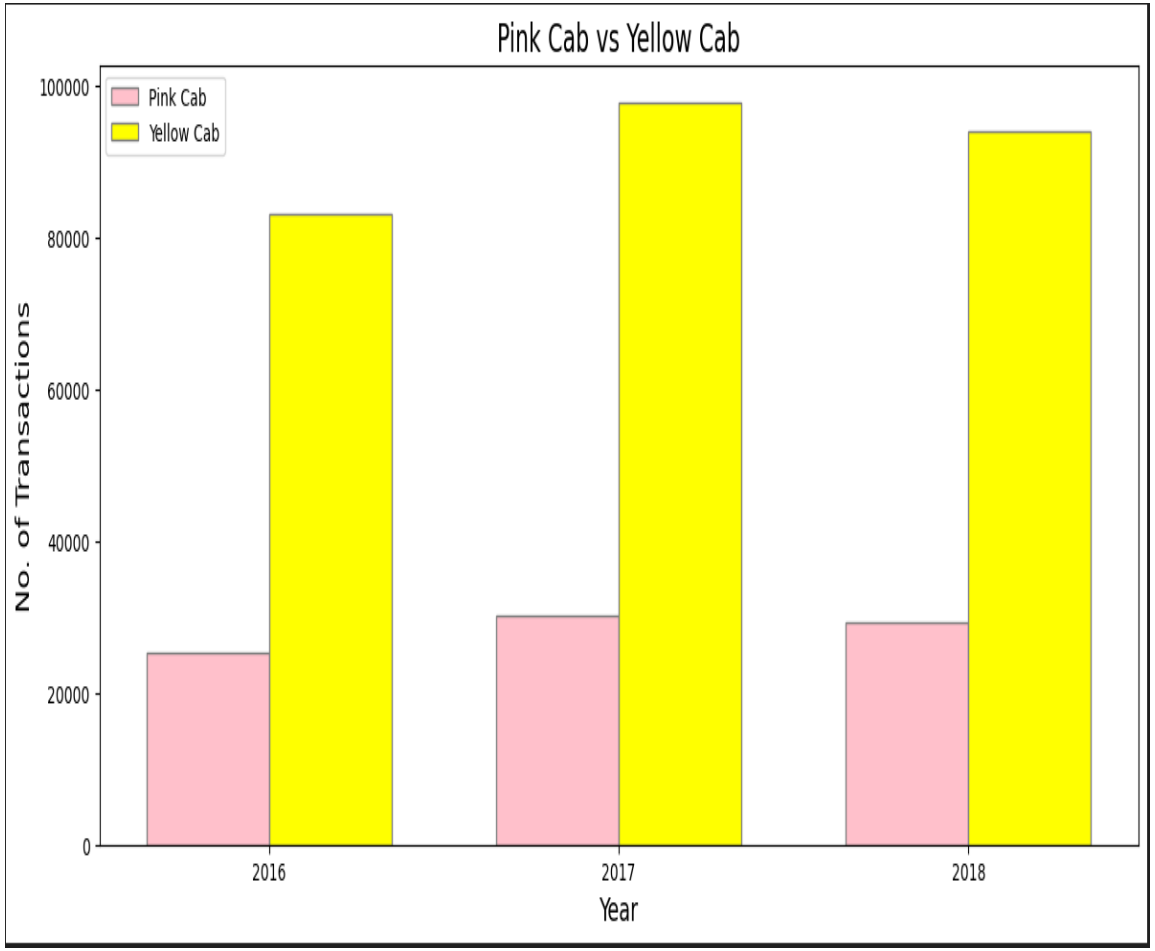
Travel frequency by Month (Pink & Yellow Cab)





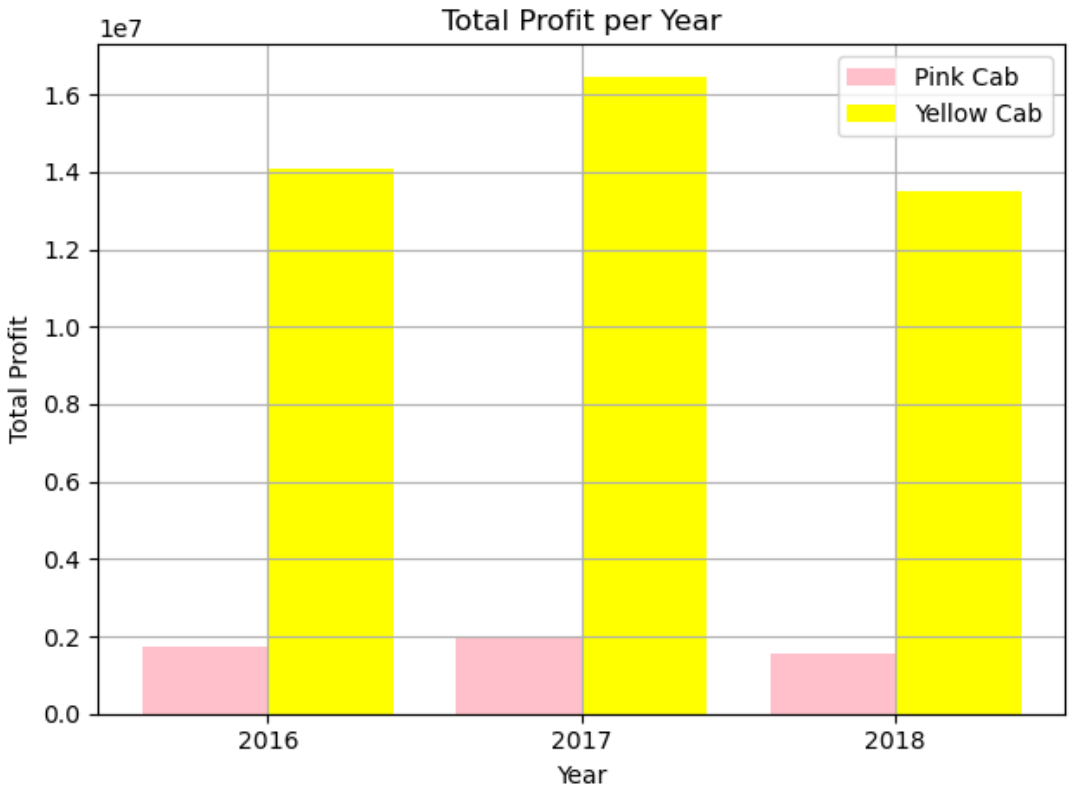
Yellow Has A Higher Travel in All the Months Assuming IT Has A better Services in the Car Such As A good (Ac) For Example or It has A modern Cars

The Transaction per Year Based On Caps



We See Also The Number Of Transaction In the Yellow Cab is Higher than Pink Cab .

Total Profit per Year



- **We See That The Profit Increase in 2017 in both Yellow and Pink cab**
- **We See Also That The Yellow Cab Profit is Higher Than Pink Cab**
- **We see Also Both Got Decrease in 2018**

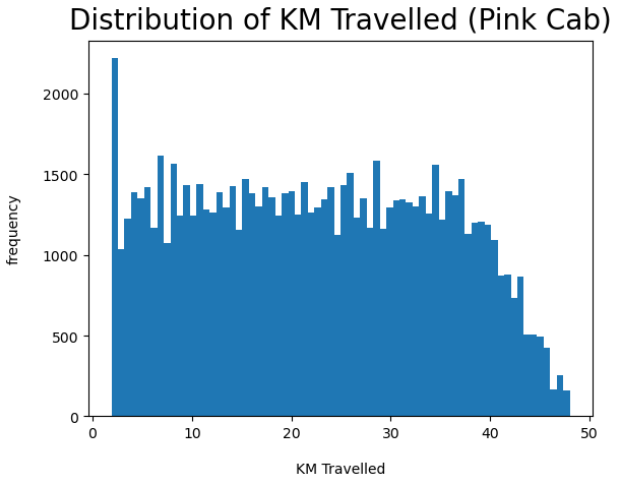
But An Assumption : In 2018

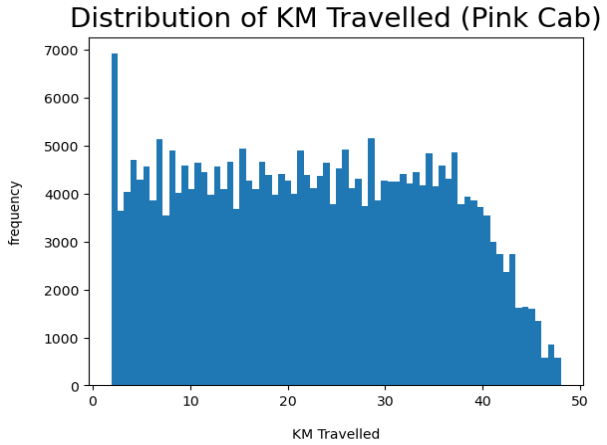
"The decrease in American cab industry profits in 2018 can be attributed to the rise of ride-sharing services like Uber and Lyft, which offer competitive pricing and innovative technology, attracting consumers away from traditional taxis. Consumers increasingly prefer the convenience and flexibility of ride-sharing apps due to quicker response times, cleaner vehicles, and real-time ride tracking. Additionally, traditional taxi companies face higher operating costs due to regulatory challenges, including licensing fees and insurance requirements. These companies also lag in adopting new technologies, resulting in less efficient operations and a poorer customer experience compared to ride-sharing services. Furthermore, ride-sharing platforms offer more flexible working arrangements for drivers, leading many taxi drivers to switch, further reducing the profitability of traditional cab companies."

In 2017 It increases Because :

1) Economic Growth
2)Increased Demand
3)Pricing Strategies

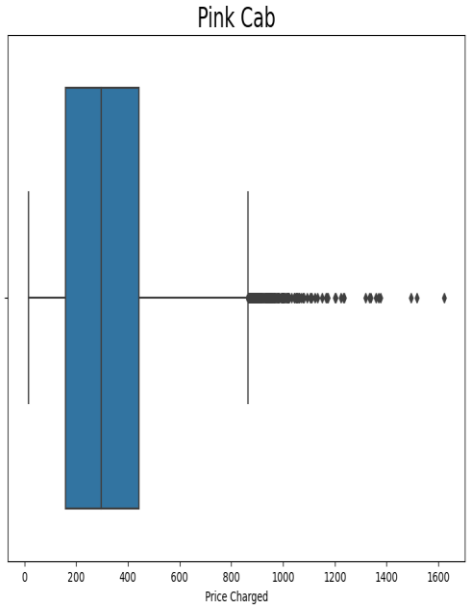
Distribution of KM Travelled for both Cabs:

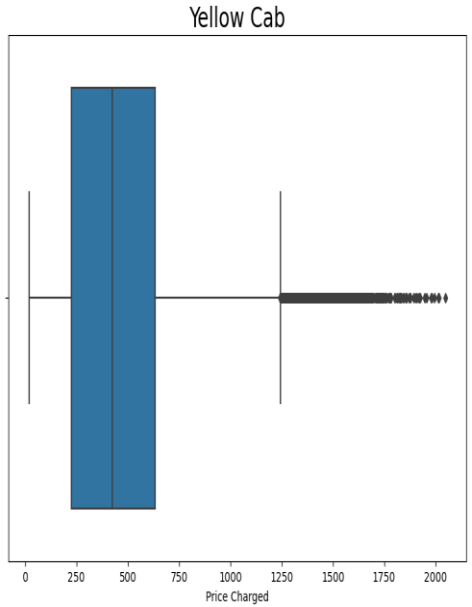




From the above graphs, we can see that for both Pink and Yellow Cab most of the rides are in the range of approximately 2 to 48 KM.

Distribution of Price Charged for both Cabs:





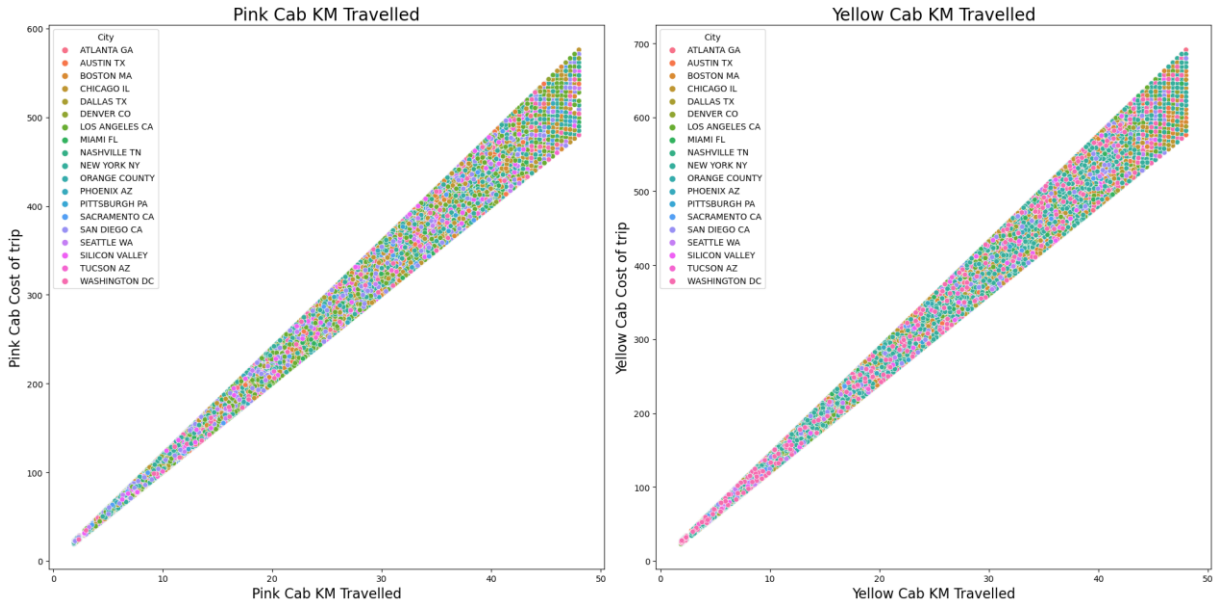
We just see the outliers of The Prices it means that :

1) The Cars Prices in the Yellow cab is more higher than the Pink which indicate of other services maybe .

2) The Type of The Cars in the yellow cap are a gasoline based cars so that indicate of the higher prices

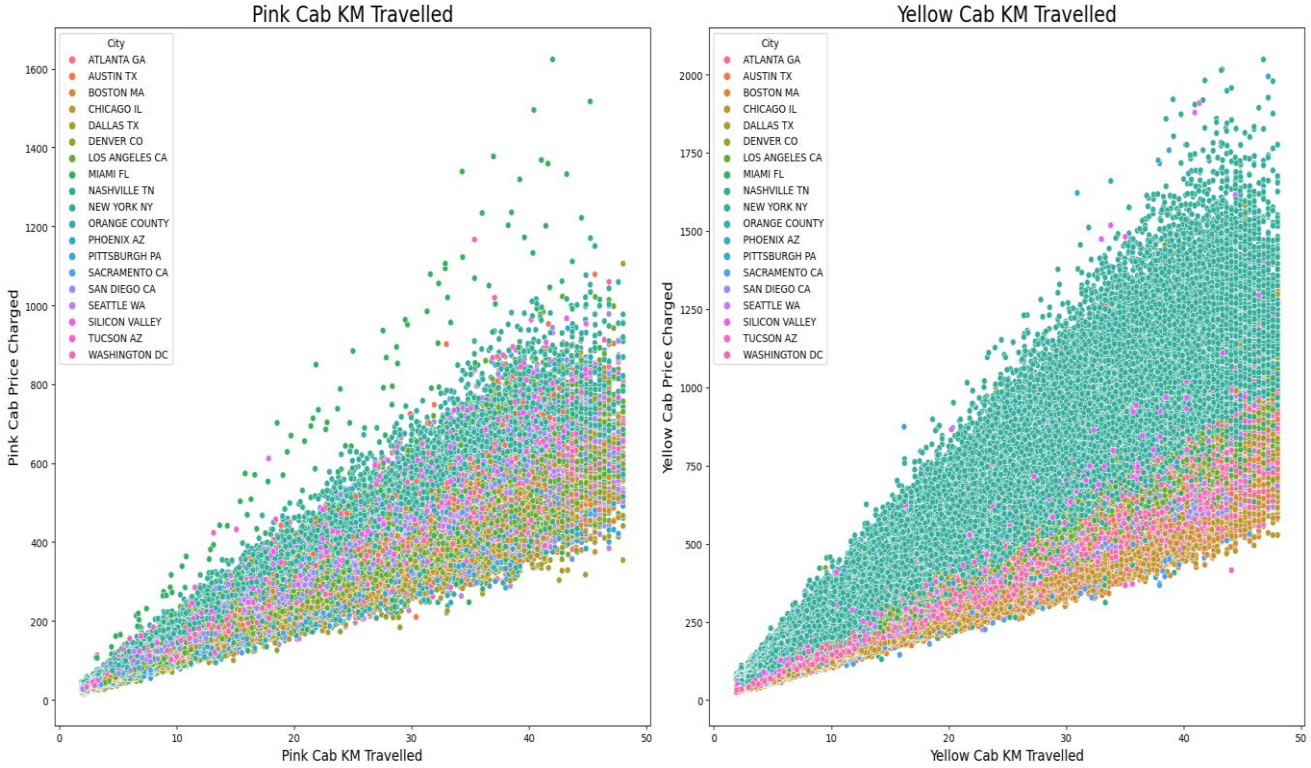
3) If Assumption 2 is wrong That indicates of an a discounts or based on the city rules in A yellow cars sometimes that's one of the reason of changing the prices

KM Travelled vs Price



We See That The Longer The Distance the Higher the price , in the yellow cab still highest Cost

Which City Has The heights population ?

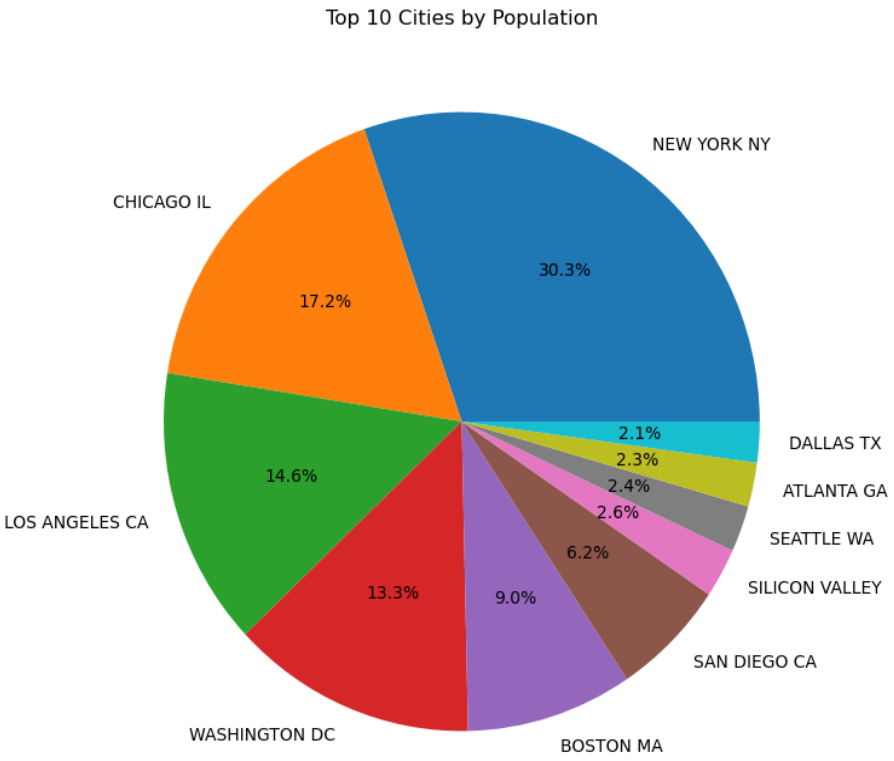


From This Graph we Find that :

1) New York city is the heights among all the cities in the Yellow cab and Pink cab indicate for a higher population

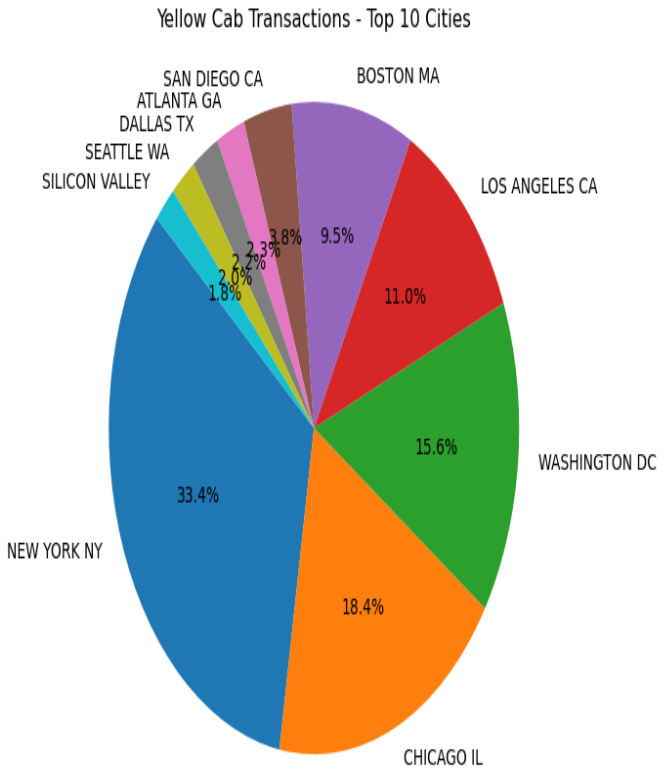
2) there is a strong positive correlation between price and km travelled

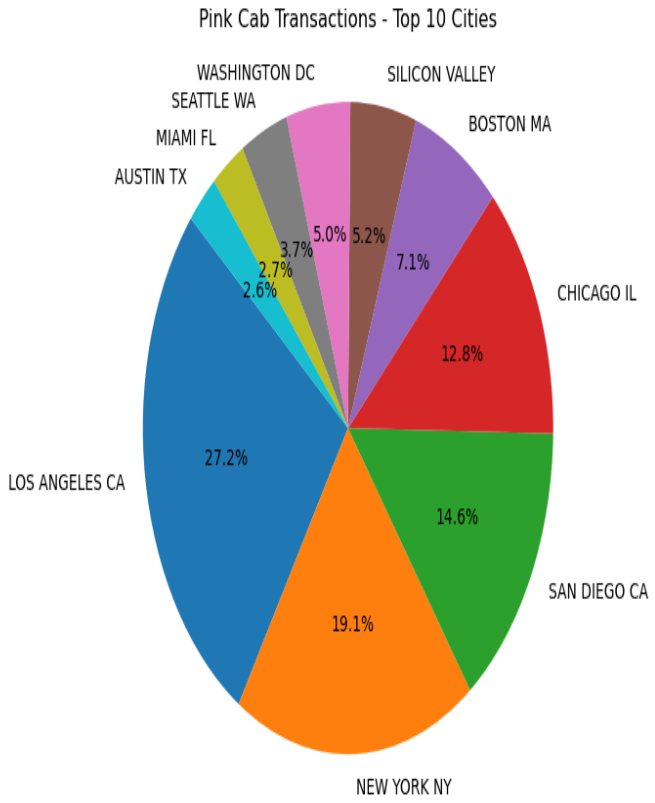
Top 10 Cities by Population



We see that the NY has the heights population and Its The most city that used the cap
We see also that Chicago the 2nd one has 17.2 OF Population that means both cities have a great economic , health care centers ,etc....

Pink and Yellow Cab Transactions - Top 10 Cities





From The Figure we see that :

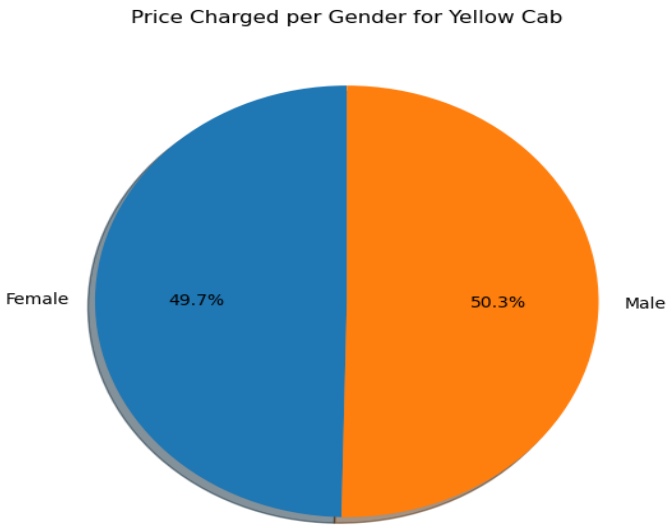
1)New York city has the Highest Transaction in Yellow cap

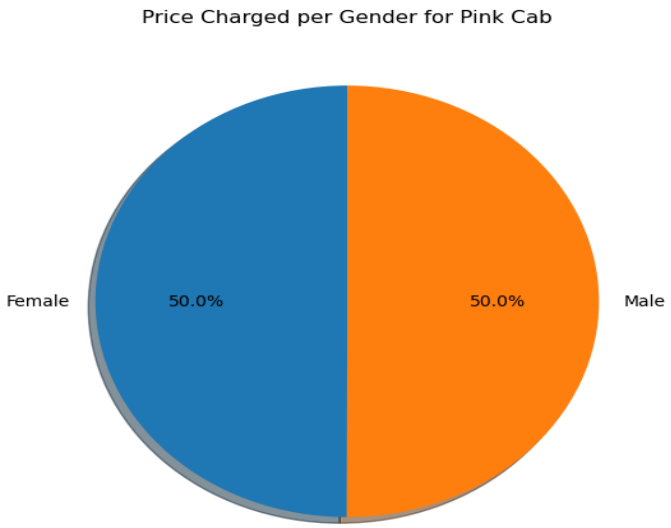
2)Los Angeles city has the Highest Transaction in pink cap

3) Both Has Large population The Reason maybe Because LA has promote the pink cap better than the yellow cap , and NY promote the Yellow cap better than the pink cap

4) Another Reasons LA is Cheaper in pink cab than the Yellow cab , Or The prices in pink suitable to everyone in there cities

Price Charged per Gender for Pink & Yellow Cab

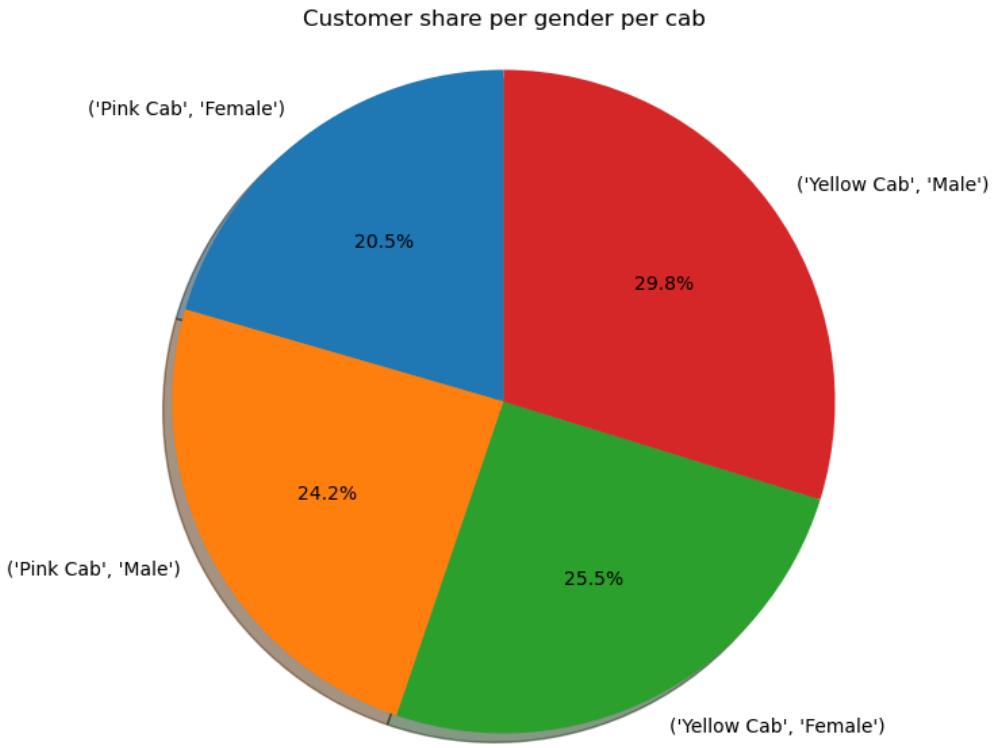




1) From This Figure we see That the Yellow cab Takes From The Mail More than the Female : Indicate that the discounts is for the Girls is much better than the boys or it seems that the girls go in a short distances than the boys.

2) The Pink charge both Girl And Boys the same

Customer share per gender per cab



WE Can see That :
1) Male are the most used for the cap cars , Because He Travel a lot , from time to time
2) Girls are less Move than the mans
3) Yellow cap is Used most

(EDA) Summary

Pink	Yellow
Rides are in the range of approximately 2 to 48 KM.	Rides are in the range of approximately 2 to 48 KM.
Price Charge range from 150 to 450 dollars.	Price Charge range from 250 to 600 dollars.
In December which is the holiday season, no. of travels was around 11000.	In December which is the holiday season, no. of travels was around 35000.
Transaction per year: • 2016: 20000 – 40000 • 2017: 20000 – 40000 • 2018: 20000 – 40000	Transaction per year: • 2016: 80000 – 100000 • 2017: 80000 – 100000 • 2018: 80000 – 100000
All the cities have the same increase in price charge with increase in distance.	In New York City the Price charged for Yellow Cab is more in comparison to the other cities
Pink Cab charges same for both Male and Female Customers.	Yellow Cab charge less from Female Customers.
Female customers are around 20.5% out of the total Customers.	Female customers are around 25.5% out of the total Customers.
Profit Margin is low each year (2016- 2018) compared to Yellow Cab.	Profit Margin is high each year (2016- 2018) compared to Pink Cab.
Pink Cabs increase margins with increase in number of Transactions.	Yellow Cab decrease Margins with the increase in Transaction.

Recommendation

- Transaction per year: For Yellow Cab Transaction per year from 2016 to 2018 is almost double than Pink Cab.

- Margin per Gender: For Yellow Cab there is difference in Margin between Male and Female Customers due to which Female Customer percentage is higher in Yellow Cab in comparison to Pink Cab.

- Profit Margin: For Yellow Cab the Profit Margin is higher per year from 2016 to 2018 in comparison to Pink Cab.

- Margin per Age: In Yellow Cab there is difference in Margin for people older than 50 yrs., whereas in Pink Cab there
- is no difference in Margin of all age group.

- Yellow Cab decreases Margins with the increase in Transaction, hence for Yellow Cab the travel frequency during
- the Month of December which is the holiday season is 3 times more than Pink Cab.

- Customers for Yellow Cab is highest in New York City which has the highest Cab Users of 28%.

- On the basis of the above points, Yellow Cab is recommended for investment.

Hypothesis Testing

1) Gender Impact on Profit

This analysis compares the mean profit between male and female customers using a two-sample t-test.

Hypotheses:

- **H0:** There is no significant difference in mean profit between male and female customers.
- **HA:** There is a significant difference in mean profit between male and female customers.
- **Test:** Conduct a two-sample t-test to compare the mean profit of male and female customers.

```
profit_male = df[df['Gender'] == 'Male']['Profit']
profit_female = df[df['Gender'] == 'Female']['Profit']
t_statistic, p_value = ttest_ind(profit_male, profit_female)
alpha = 0.05
print("T-statistic:", t_statistic)
print("P-value:", p_value)
if p_value < alpha:
    print("Reject null hypothesis: There is a significant difference in profit between male and female customers.")
else:
    print("Fail to reject null hypothesis: There is no significant difference in profit between male and female customers.")
```

2] ✓ 0.0s

```
T-statistic: 12.70131593950125
P-value: 5.921884821326977e-37
Reject null hypothesis: There is a significant difference in profit between male and female customers.
```

2) Effect of Payment Mode on Profit:

Hypotheses:

- **H0:** There is no significant difference in profit between customers who pay by card and those who pay by cash.
- **HA:** There is a significant difference in profit between customers who pay by card and those who pay by cash.
- **Test:** Perform a t-test or Mann-Whitney U test to compare the profit distributions for card and cash payments.

```
profit_card = df[df['Payment_Mode'] == 'Card']['Profit']
profit_cash = df[df['Payment_Mode'] == 'Cash']['Profit']
t_statistic, p_value = ttest_ind(profit_card, profit_cash)
alpha = 0.05
print("Results of t-test:")
print("T-statistic:", t_statistic)
print("P-value:", p_value)
if p_value < alpha:
    print("Reject null hypothesis: There is a significant difference in profit between customers who pay by card and those who pay by cash.")
else:
    print("Fail to reject null hypothesis: There is no significant difference in profit between customers who pay by card and those who pay by cash.")
```

✓ 0.1s

```
Results of t-test:
T-statistic: -0.7630743349932244
P-value: 0.4454195660215633
Fail to reject null hypothesis: There is no significant difference in profit between customers who pay by card and those who pay by cash.
```

- **H0:** There is no correlation between age and profit.
- **HA:** There is a significant correlation between age and profit.
- **Test:** Calculate the Pearson correlation coefficient between age and profit.

pearsonr Function :

The pearsonr function is a part of the scipy.stats module in Python. It is used to calculate the Pearson correlation c
p-value for testing non-correlation between two variables.

```
age = df['Age']
profit = df['Profit']
correlation_coefficient, p_value = pearsonr(age, profit)
alpha = 0.05
print("Pearson correlation coefficient:", correlation_coefficient)
print("P-value:", p_value)
if p_value < alpha:
    print("Reject null hypothesis (H0). There is a significant correlation between age and profit.")
else:
    print("Fail to reject null hypothesis (H0). There is no significant correlation between age and profit.")
```

5] ✓ 0.0s

```
Pearson correlation coefficient: -0.005092963667618835
P-value: 0.002264105651669707
Reject null hypothesis (H0). There is a significant correlation between age and profit.
```

4) Income Level and Profitability:

- **H0:** There is no association between income level and profit.
- **H1:** There is a significant association between income level and profit.
- **Test:** Use ANOVA or Kruskal-Wallis test to compare profit across different income groups.

```
f_statistic, p_value = f_oneway(low_income, medium_income, high_income)

print("ANOVA Test:")
print("F-statistic:", f_statistic)
print("p-value:", p_value)

if p_value < 0.05:
    print("Reject the null hypothesis. There is a significant association between income level and profit.")
else:
    print("Fail to reject the null hypothesis. There is no significant association between income level and profit.")
```

98] ✓ 0.0s

```
ANOVA Test:
F-statistic: 13.973536610620892
p-value: 8.542915321759085e-07
Reject the null hypothesis. There is a significant association between income level and profit.
```

5)Effect of City Population on Profit:

- **H0:** There is no significant difference in profit between cities with different populations.
- **HA:** Profit varies significantly between cities with different populations.
- **Test:** Perform ANOVA or Kruskal-Wallis test to compare profit among cities with different population sizes.

```
f_statistic, p_value_anova = f_oneway(*city_groups.values())
print("ANOVA Test Result:")
print("F-statistic:", f_statistic)
print("p-value:", p_value_anova)
alpha = 0.05
if p_value_anova < alpha:
    print("\nReject the null hypothesis (ANOVA): There is a significant difference in profit among cities with different population sizes.")
else:
    print("\nFail to reject the null hypothesis (ANOVA): There is no significant difference in profit among cities with different population sizes.")
```

[402] ✓ 0.0s

```
ANOVA Test Result:
F-statistic: 9528.597969224551
p-value: 0.0

Reject the null hypothesis (ANOVA): There is a significant difference in profit among cities with different population sizes.
```

6)Relationship between Distance Traveled and Profit:

- **H0:** There is no correlation between the distance traveled and profit.
- **HA:** There is a significant correlation between the distance traveled and profit.
- **Test:** Calculate the Pearson correlation coefficient between distance traveled and profit.

```
pearson_corr, p_value = pearsonr(df['KM Travelled'], df['Profit'])
alpha = 0.05
print("Pearson correlation coefficient between distance traveled and profit:", pearson_corr)
print("P-value:", p_value)
if p_value < alpha:
    print("Reject null hypothesis: There is a significant correlation between distance traveled and profit.")
else:
    print("Fail to reject null hypothesis: There is no significant correlation between distance traveled and profit.")
```

[404] ✓ 0.0s

```
Pearson correlation coefficient between distance traveled and profit: 0.4627681978971101
P-value: 0.0
Reject null hypothesis: There is a significant correlation between distance traveled and profit.
```

7)Profitability Over Time:

- **H0:** There is no significant difference in profit across different years or months.
- **HA:** Profit varies significantly across different years or months.
- **Test:** Perform ANOVA or Kruskal-Wallis test to compare profit across different years or months.

```
profit_by_time = [df[df['Year'] == year]['Profit'] for year in df['Year'].unique()]
alpha = 0.05
```

05] ✓ 0.0s

```
f_statistic, p_value_anova = f_oneway(*profit_by_time)
print("Results of ANOVA test:")
print("F-statistic:", f_statistic)
print("P-value (ANOVA):", p_value_anova)

if p_value_anova < alpha:
    print("Reject null hypothesis: Profit varies significantly across different years or months (according to ANOVA).")
else:
    print("Fail to reject null hypothesis: There is no significant difference in profit across different years or months (according to ANOVA).")
```

06] ✓ 0.0s

```
Results of ANOVA test:
F-statistic: 852.8806371524433
P-value (ANOVA): 0.0
Reject null hypothesis: Profit varies significantly across different years or months (according to ANOVA).
```

Models

```python
reg = RandomForestRegressor(n_estimators=100, random_state=42)
reg.fit(X, y)

# Determine feature importances
importances = reg.feature_importances_

# Sort indices of features by importance
indices = np.argsort(importances)[::-1]

# Select the top 6 features
selected_features_indices = indices[:6]
selected_features = X.columns[selected_features_indices]

print("Top 6 selected features:", selected_features)
```

20] ✓ 4m 41.0s

```
Top 6 selected features: Index(['Cost of Trip', 'Profit', 'KM Travelled', 'Income (USD/Month)', 'Month',
       'Year'],
      dtype='object')
```

```python
linear_reg = LinearRegression()
linear_reg.fit(X_train,y_train)
predect=linear_reg.predict(X_test)
score=linear_reg.score(X_test,y_test)
print(f"score of the random forest:{score}")
mse = mean_squared_error(y_test, predect)
rmse = np.sqrt(mse)
print("Root Mean Squared Error (RMSE): {:.4f}".format(rmse))
```

✓ 0.1s

```
score of the random forest:1.0
Root Mean Squared Error (RMSE): 0.0000
```

```python
decision_tree = DecisionTreeRegressor(max_depth=None, random_state=42)
decision_tree.fit(X_train,y_train)
decision_tree_prediction=decision_tree.predict(X_test)
score=decision_tree.score(X_test,y_test)
print(f"score of the random forest:{score}")
mse = mean_squared_error(y_test, predect)
rmse = np.sqrt(mse)
print("Root Mean Squared Error (RMSE): {:.4f}".format(rmse))
```

✓ 1.9s

```
score of the random forest:0.9999586879203661
Root Mean Squared Error (RMSE): 0.0000
```

```python
random_forest = RandomForestRegressor(n_estimators=100, random_state=42)
random_forest.fit(X_train,y_train)
random_forest_prediction=random_forest.predict(X_test)
score=random_forest.score(X_test,y_test)
print(f"score of the random forest:{score}")
mse = mean_squared_error(y_test, predect)
rmse = np.sqrt(mse)
print("Root Mean Squared Error (RMSE): {:.4f}".format(rmse))
```

✓ 3m 20.6s

```
score of the random forest:0.9999928592648589
Root Mean Squared Error (RMSE): 0.0000
```

Thank You