Exploratory Data Analysis

By: Keith D. Smith

May 28, 2024

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

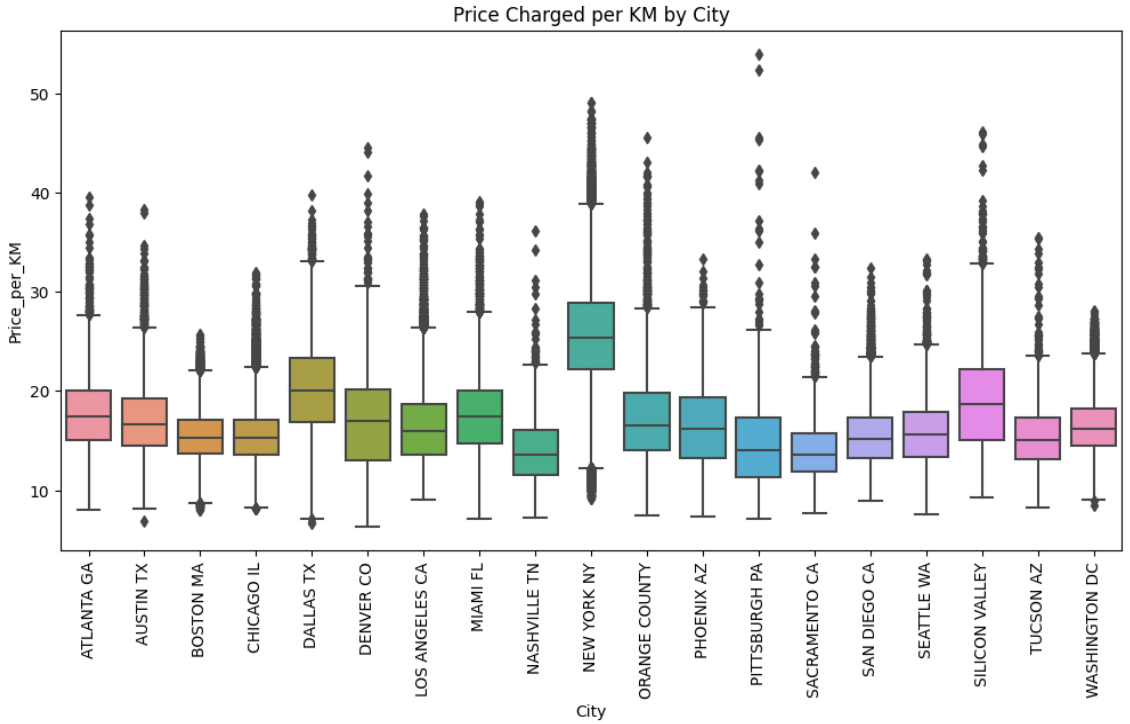
Problem Statements

1. The avg. Price/Km varies between different cities.
2. The Price/km is higher for one company compared to the other.
3. Higher-income customers travel further
4. The mode of payment varies by city
5. There is a difference in the avg. cost of trips between the two companies
6. The gender of customers influences the distance traveled and price charged
7. The cost of the trip is seasonal

Approach

1. One-way Anova
2. T-test
3. Pearson’s R
4. Chi-square Test
5. Anova

The Average Price charged per Km varies significantly between different cities

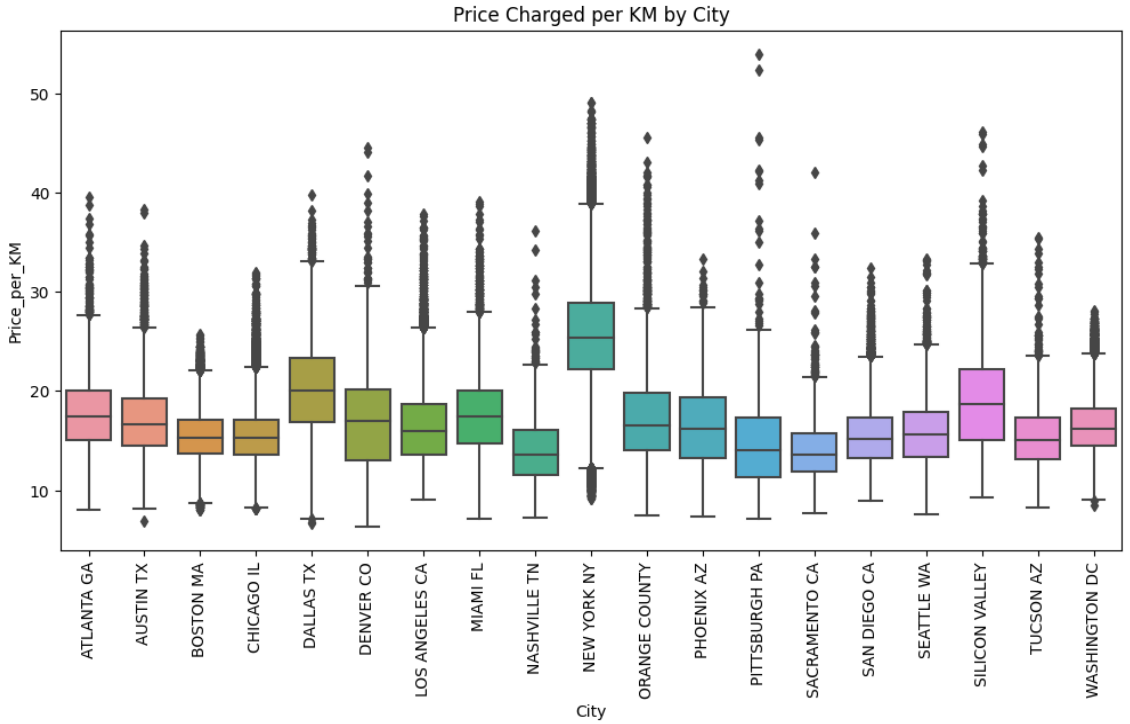


Test	P-Value
One-Way Anova	0.0

The p-value of .00 is less than the alpha of .05. There is a statistically significant difference in the average price charged per kilometer between different cities.

Data Glacier
Your Deep Learning Partner

The Average Price charged per Km varies significantly between different cities

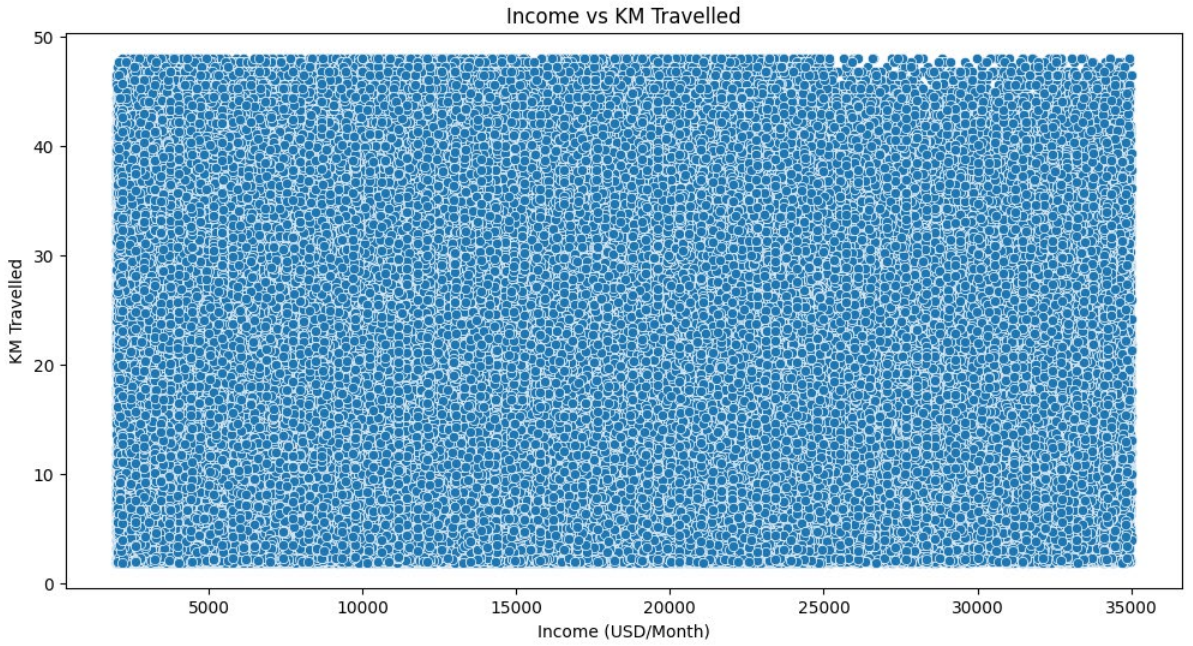


Test	P-Value
T-test	0.0

The p-value of .00 is less than the alpha of .05. This indicates a statistically significant difference in the price charged per kilometer between the two companies. .

Data Glacier
Your Deep Learning Partner

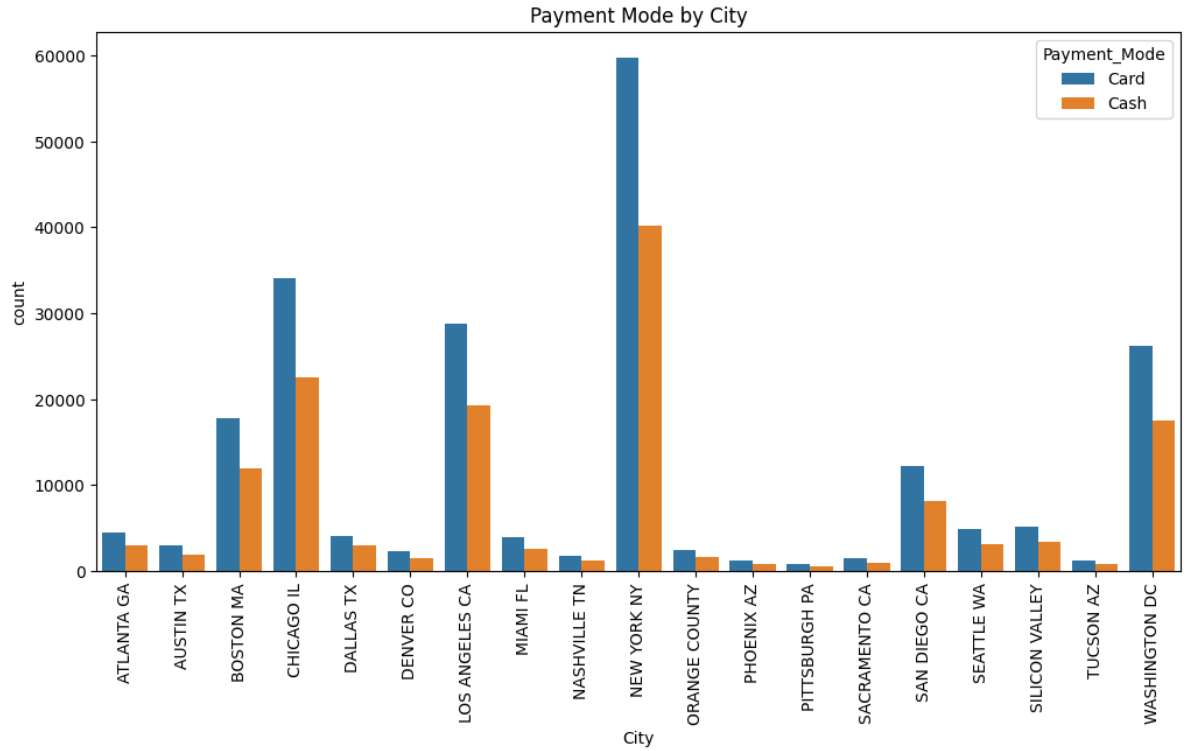
Higher income customers tend to travel longer distances



Test	Value
Pearson’s R	-0.005
There is not a linear relationship between the income of customers and the distance they travel.	

Data Glacier
Your Deep Learning Partner

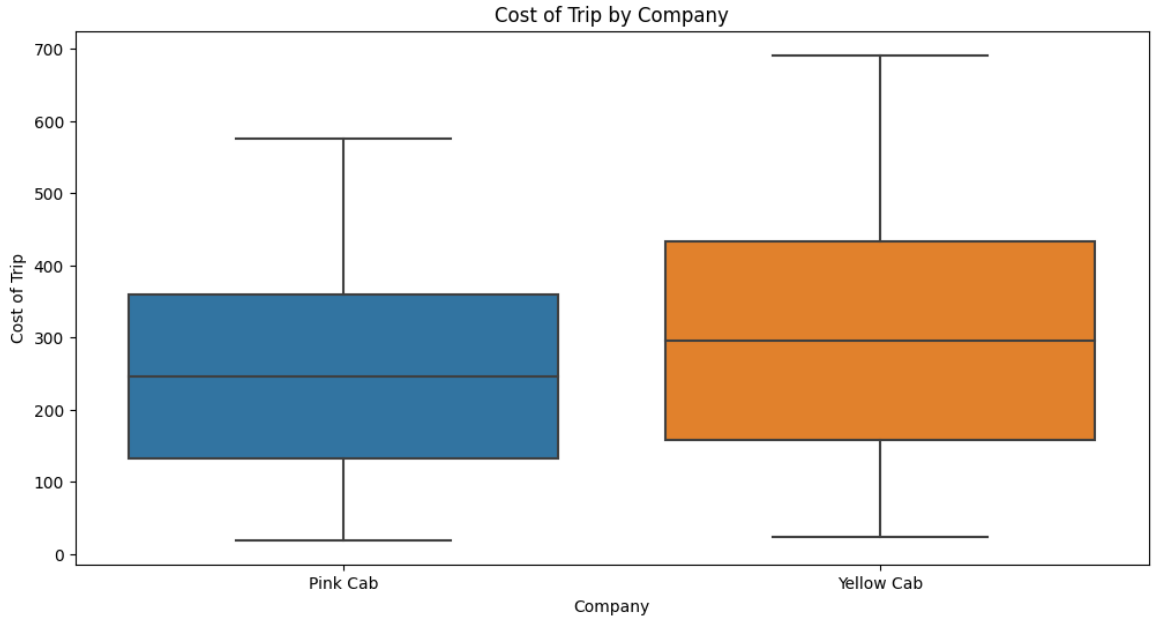
The mode of payment varies significantly by city



Test	P-Value
Chi-square	0.104
The p-value is 0.104, which is greater than 0.05. This indicates that there is no statistically significant difference in the mode of payment across different cities.	

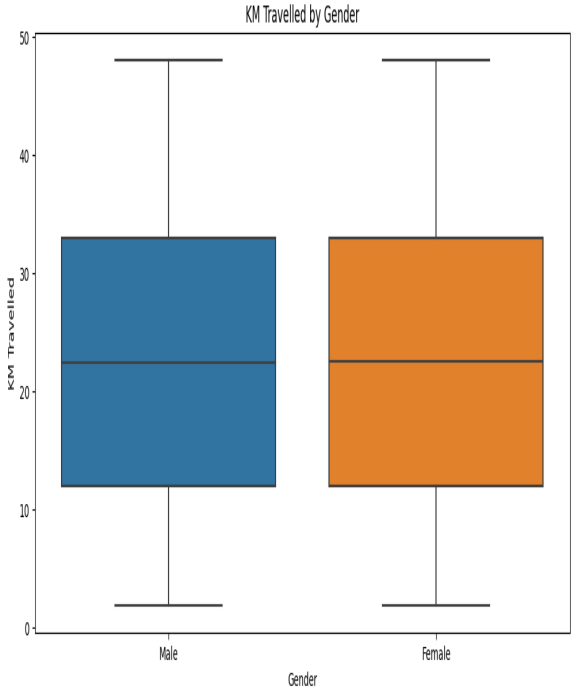
Data Glacier
Your Deep Learning Partner

There is a difference in the average cost of trips between the two companies



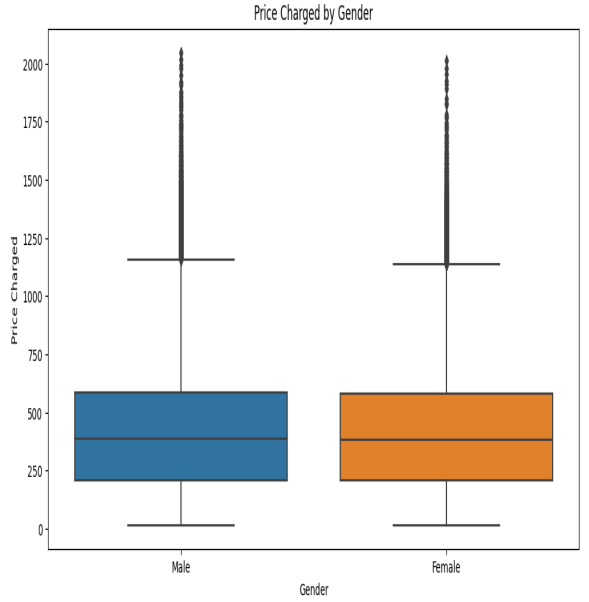
Test	P-Value
T-test	0.0
The p-value is 0.0, which is less than 0.05. This indicates a statistically significant difference in the average cost of trips between the two companies.	

The gender of customers influences the distance traveled and price charged



Test	P-Value
T-test	0.418

The p-value is 0.418, which is greater than 0.05. This indicates no statistically significant difference in the distance traveled between genders.

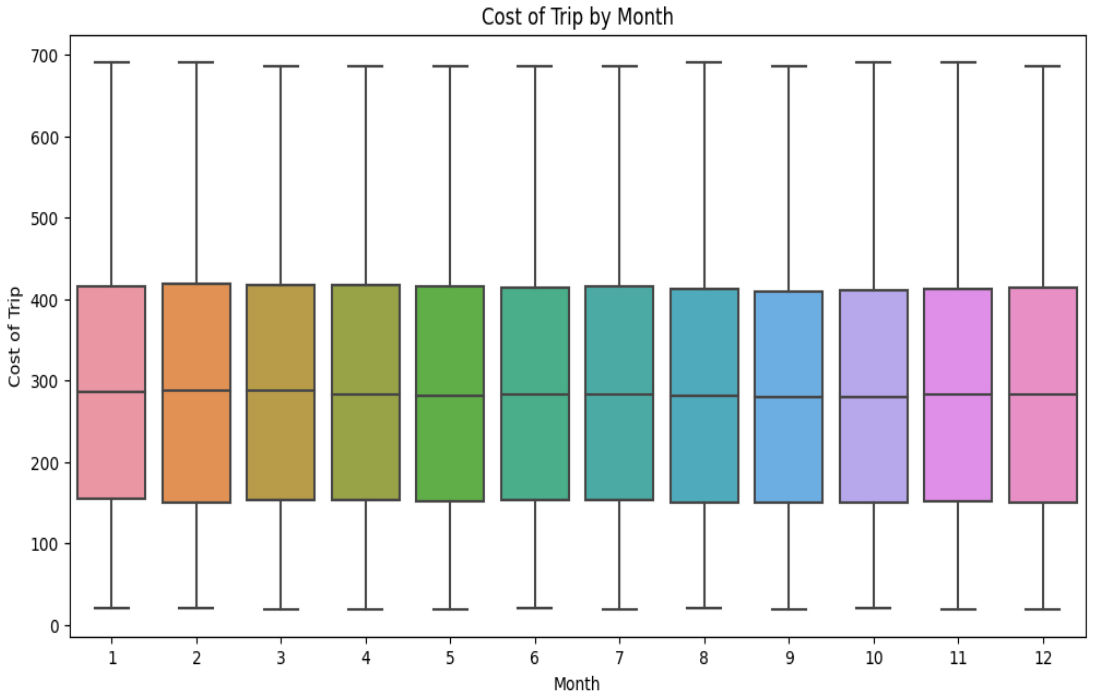


Test	P-Value
T-test	0.0

The p-value is very close to 0, which is less than 0.05. This indicates a statistically significant difference in the price charged between genders.

Data Glacier
Your Deep Learning Partner

The cost of the trip is seasonal



Test	P-Value
T-test	0.0
The p-value is very close to 0, which is less than 0.05. This indicates that there is a statistically significant seasonal variation in the cost of trips.	

Data Glacier
Your Deep Learning Partner

EDA Summary

1. No Missing Data
2. Outliers need to be clarified
3. Yellow Cab has a higher average cost per trip
4. Yellow Cab is more expensive per Km than Pink Cab
5. Cab revenue is seasonal by month
6. Gender has no association with the price charged
7. City has no association with mode of payment
8. Income has no correlation with distance traveled
9. City is associated with the avg. price per kilometer

Recommendations

1. Determine if outliers should be included in models.
2. Prepare data for time-series modeling.
3. Model data via interpretable models to determine revenue for Pink Cab and Yellow Cab.

Thank You