Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

G2M insight for Cab Investment firm

21 November 2022

Introduction

Problem Case:

XYZ is a private 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 and as per their Go-to-Market(G2M) strategy they want to understand the market before taking final decision.

Data Set:

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

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

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

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

Solution:

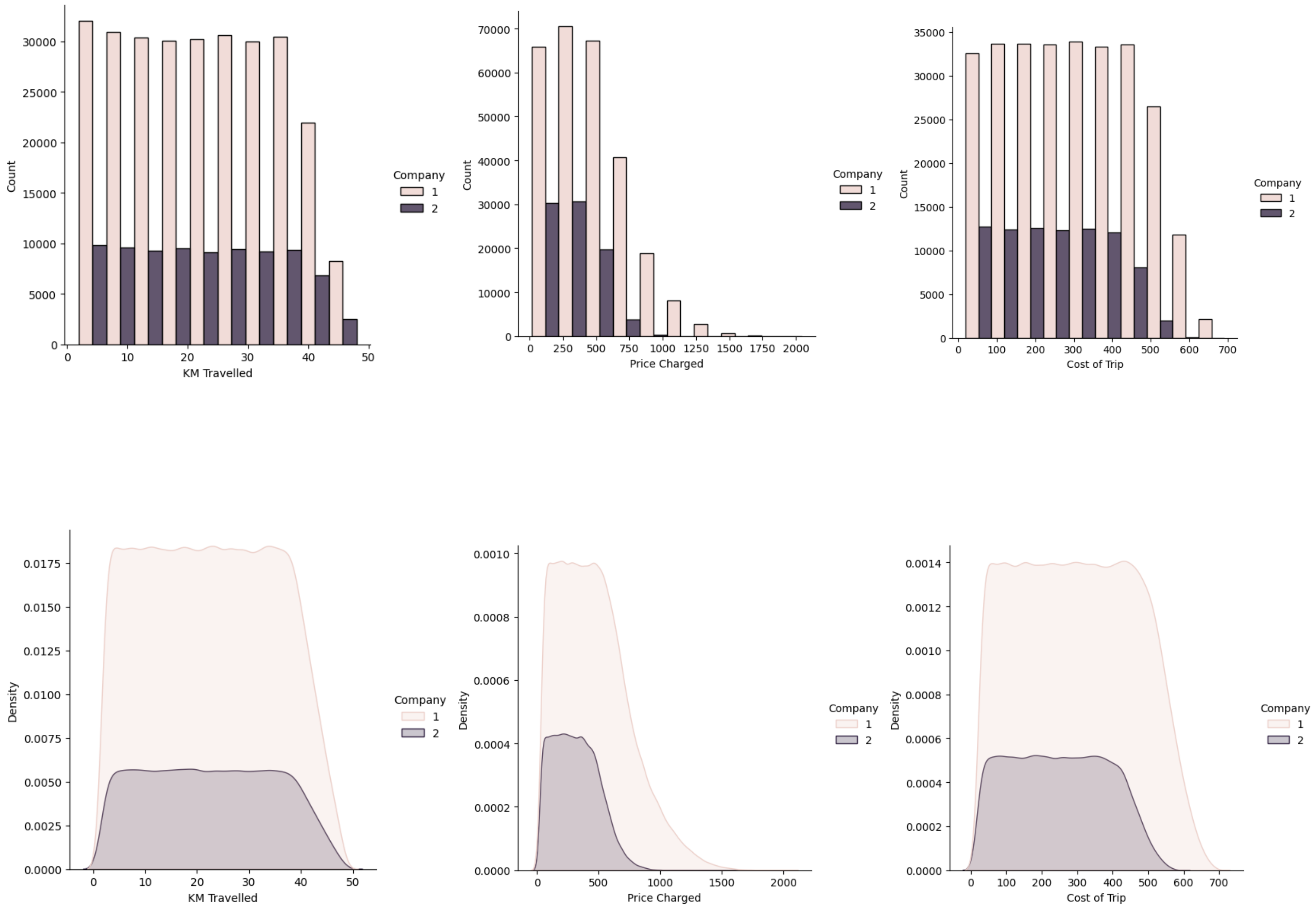
We have to provide an insight and analysis of these data for an investor to make their Investment Decision in the Cab Industry.

Company wise Analysis

Company 1: Yellow Cab
Company 2: Pink Cab

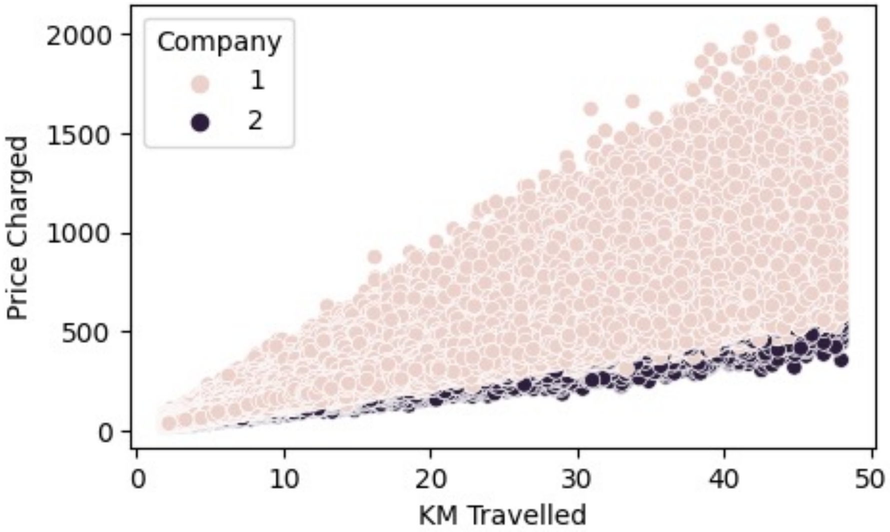
As we can see here clearly, the Yellow Cab is dominating with the highest KM Travelled, Cost of Trip, and the price that was changed.

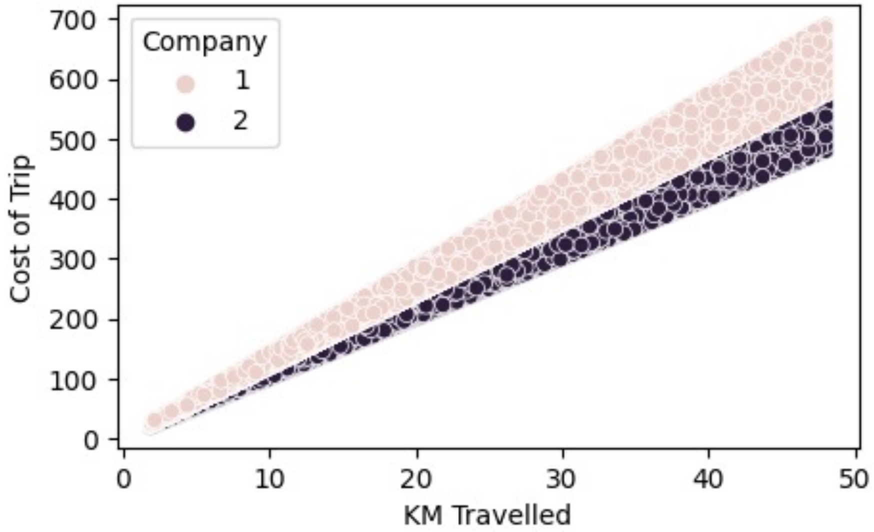
Further, we can see the KDE (kernel density estimator) distribution here, and both of the company's distribution looks similar.

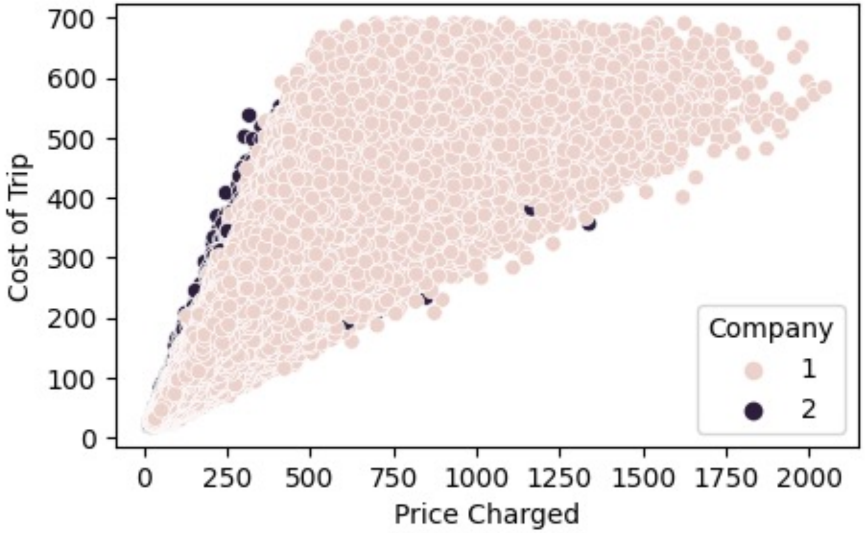


- Here are the results of scatter plots for both cabs with the charged price, cost of the trip, and the kilometers traveled.

  The kilometers that were traveled with the cost of the trip have a linear relationship as we can see on the top right plot.





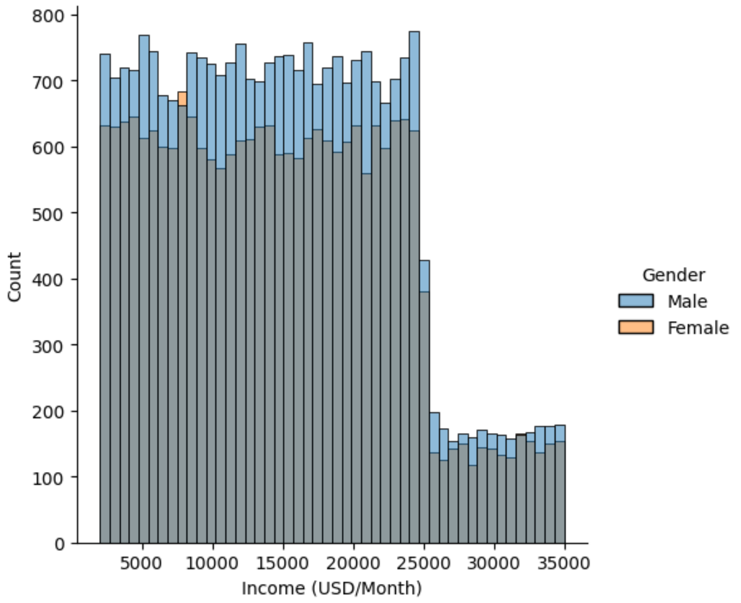


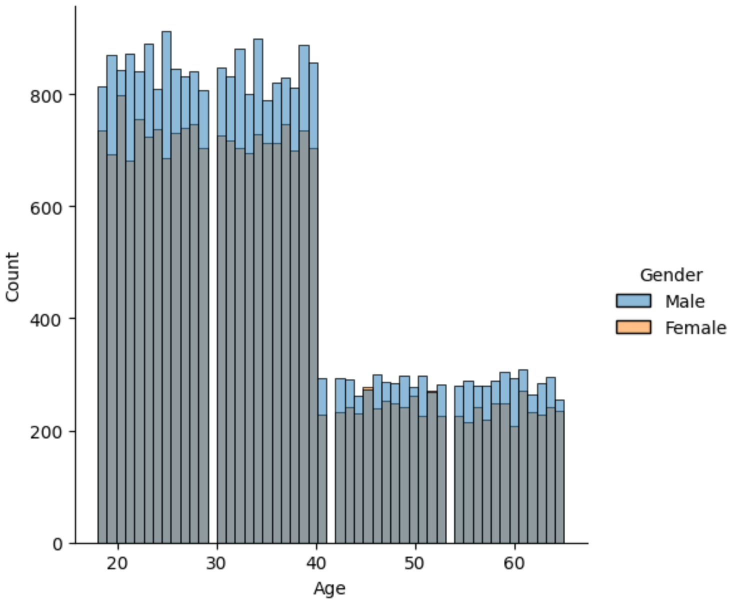
Data Glacier
Your Deep Learning Partner

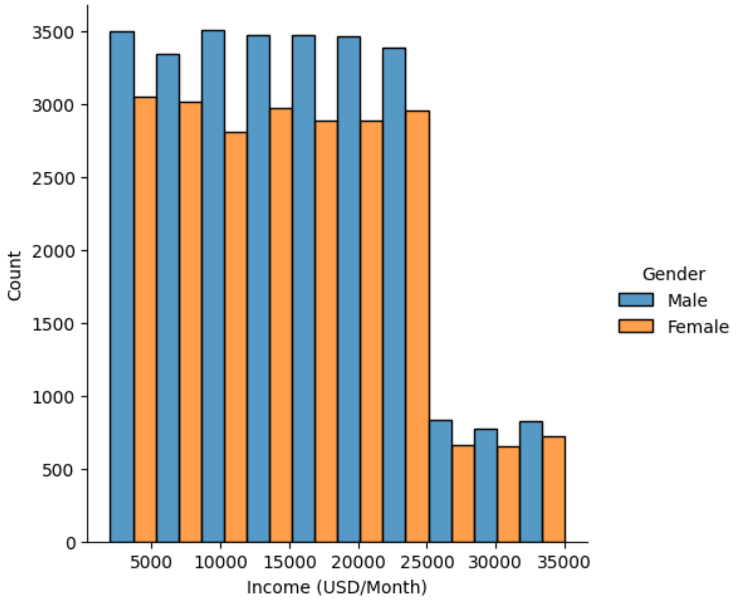
Gender Wise Analysis

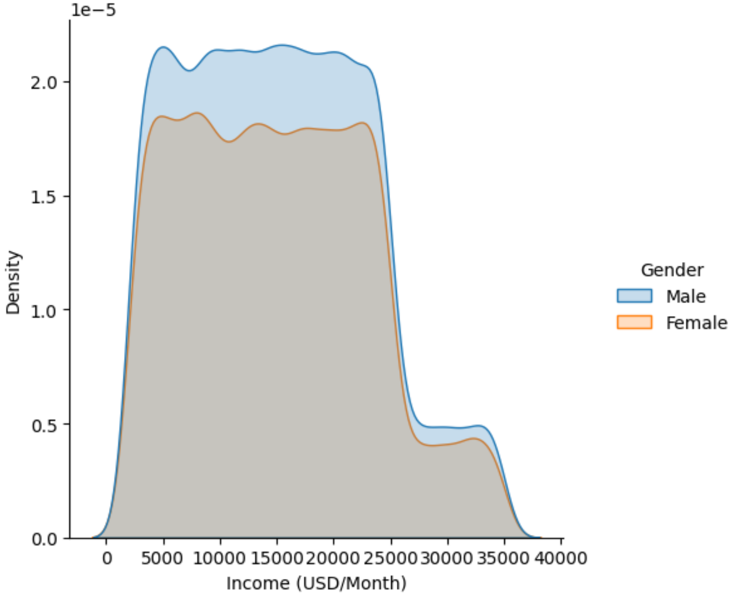
Here we can see that the number of passengers who traveled was more Males than Females.

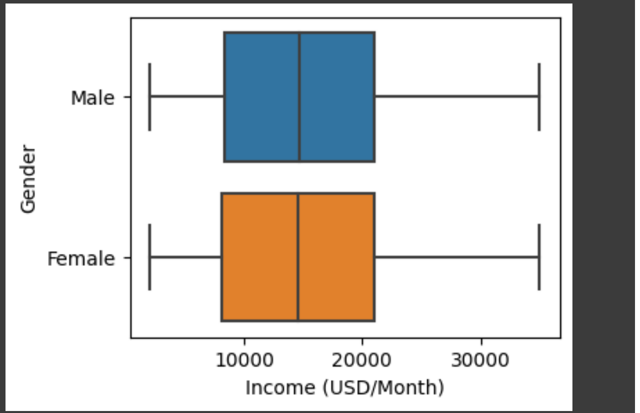
Results of Histogram plots and the distribution.









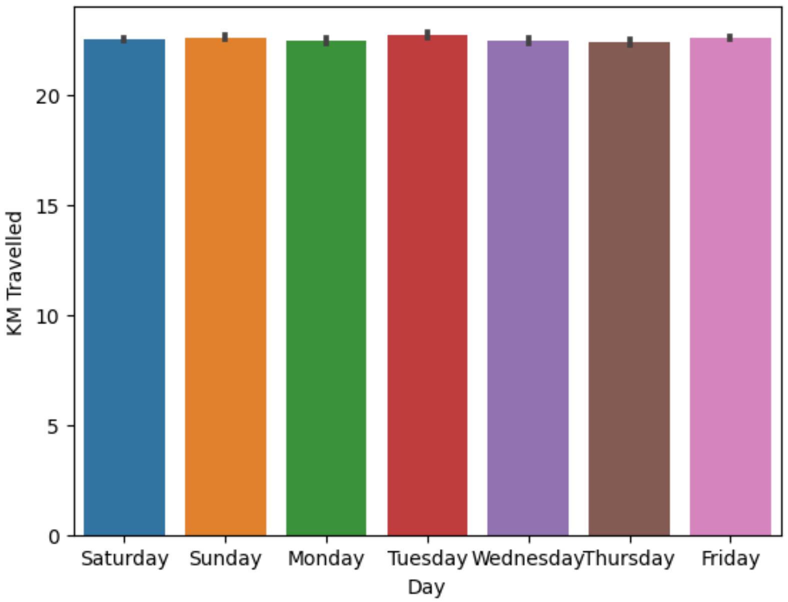


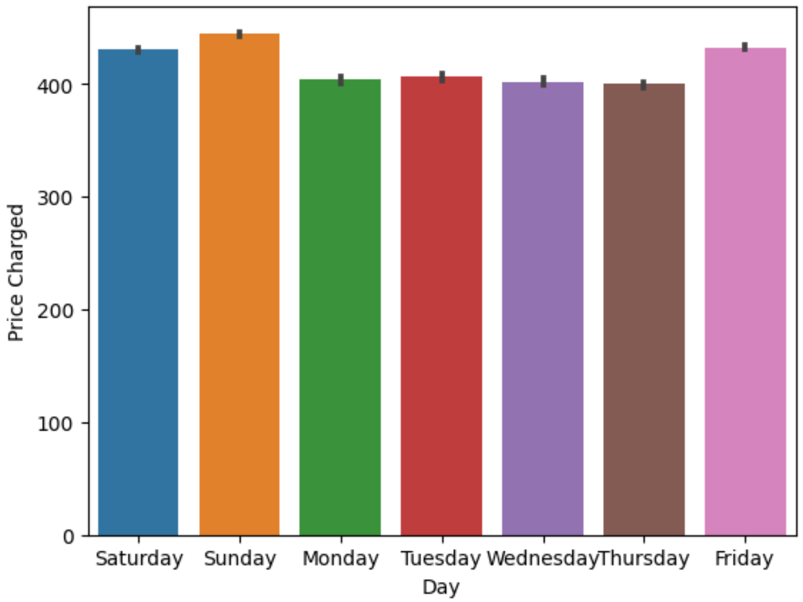
Data Glacier

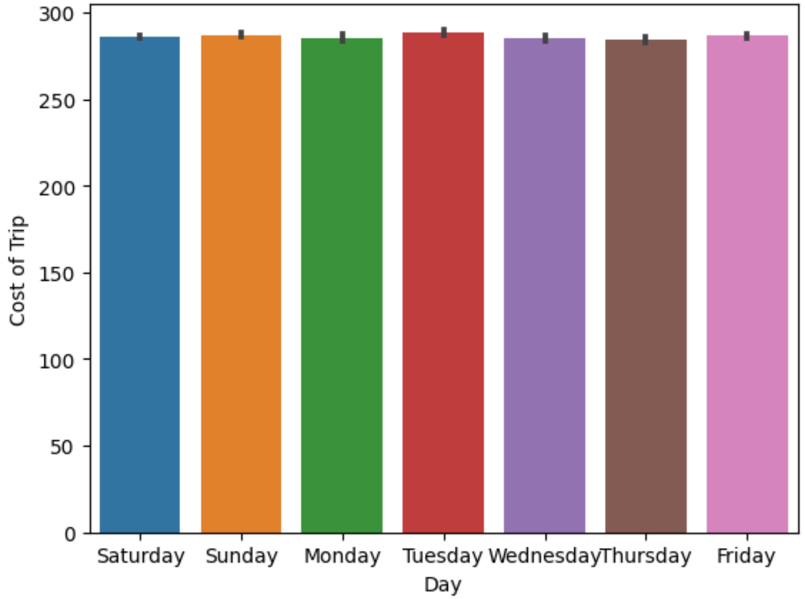
Your Deep Learning Partner

➢ Weekday Analysis

Price was being charged the highest on the weekends. Especially Sunday, as we can see in the plot.



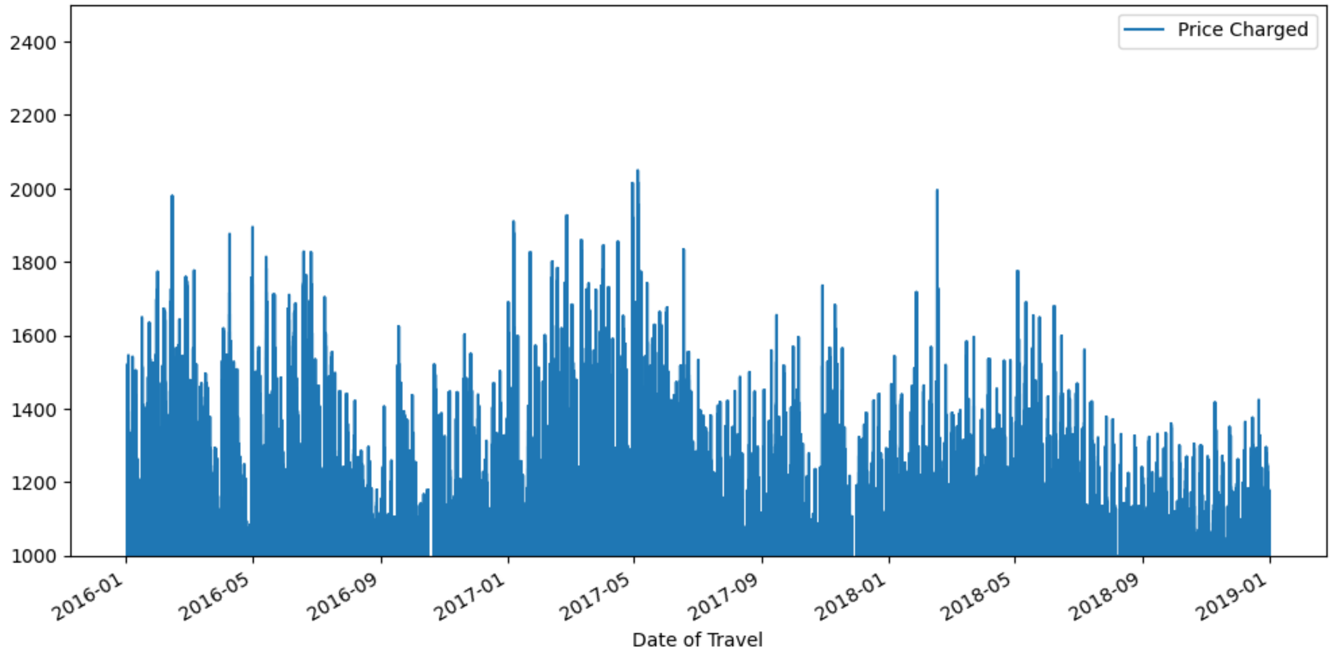


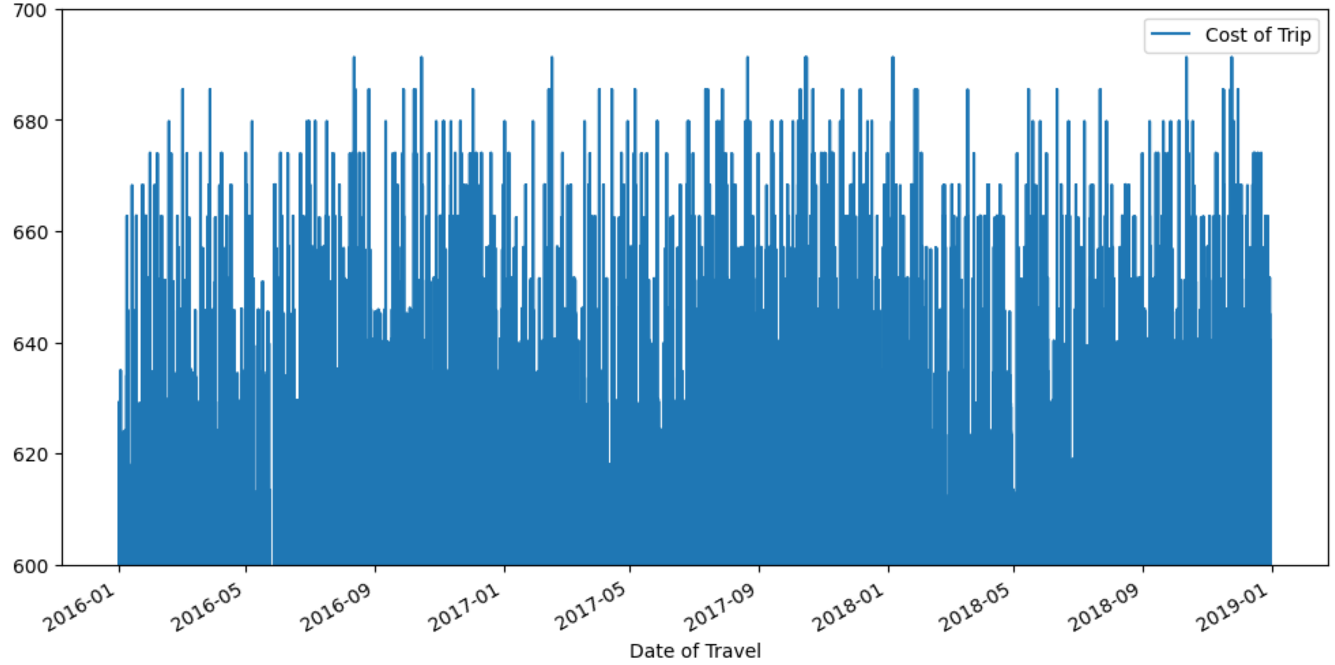


- Here we can see that the price that was charged for higher than the cost of the trip during the time series.

  We can also notice that the variation in cost for cab drivers was not that volatile but the prices during some time frames like 2017-05 to 2017-09 were volatile.

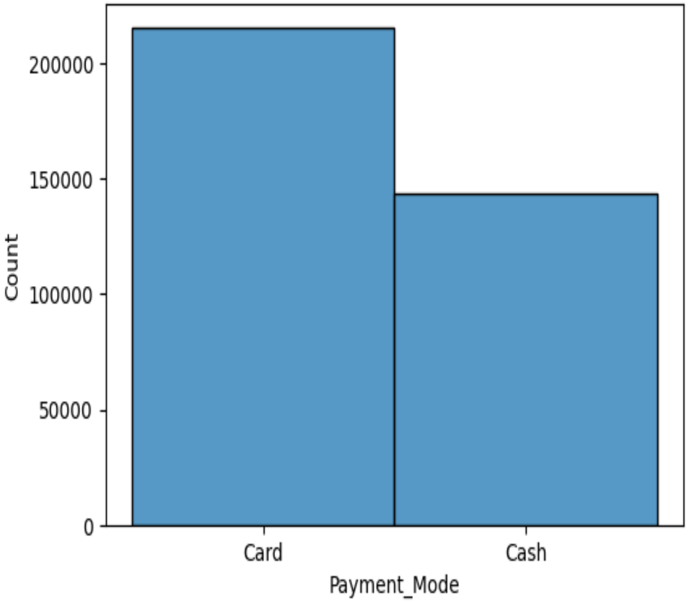
  Although the cost of the trip was not high, there was a spike in cab prices that were charged during 2018-03.

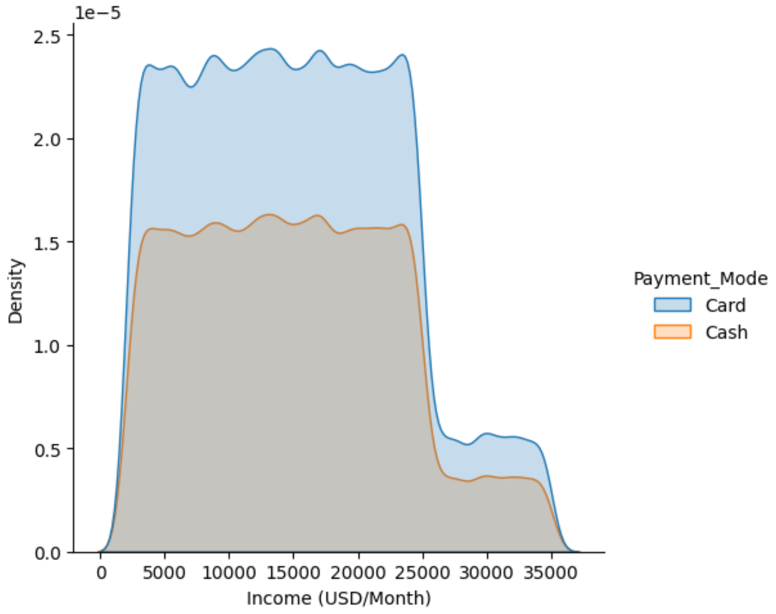


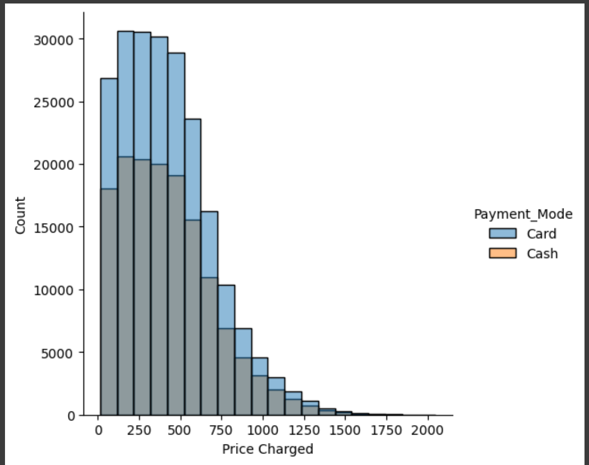


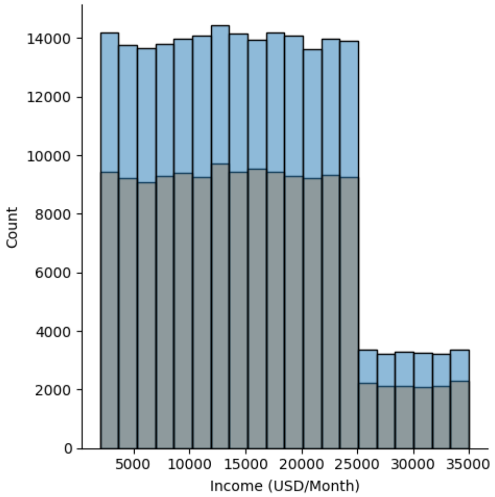
Payment Method: Card, Cash

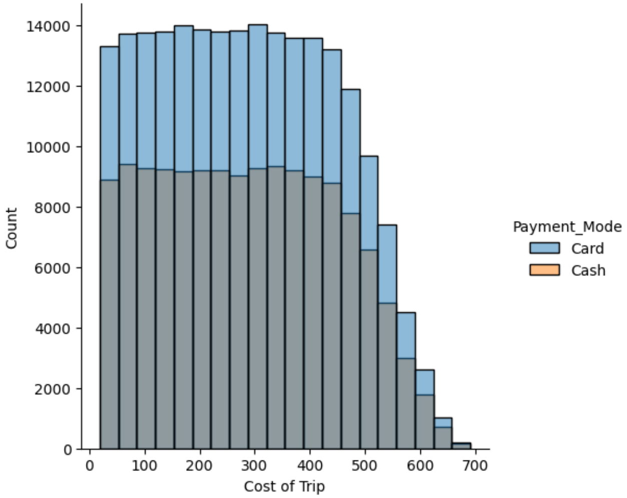
As we can see here, the transaction that was being made by Cash was higher than the transaction that was being done by Card. The distribution of all these plots tells us the cash users were higher in number than the card users.

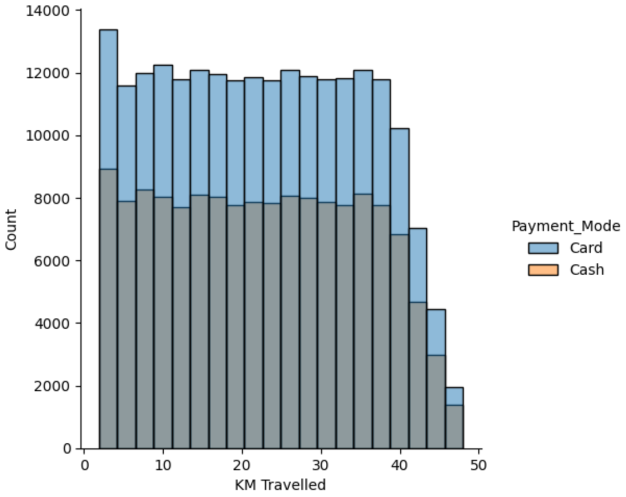








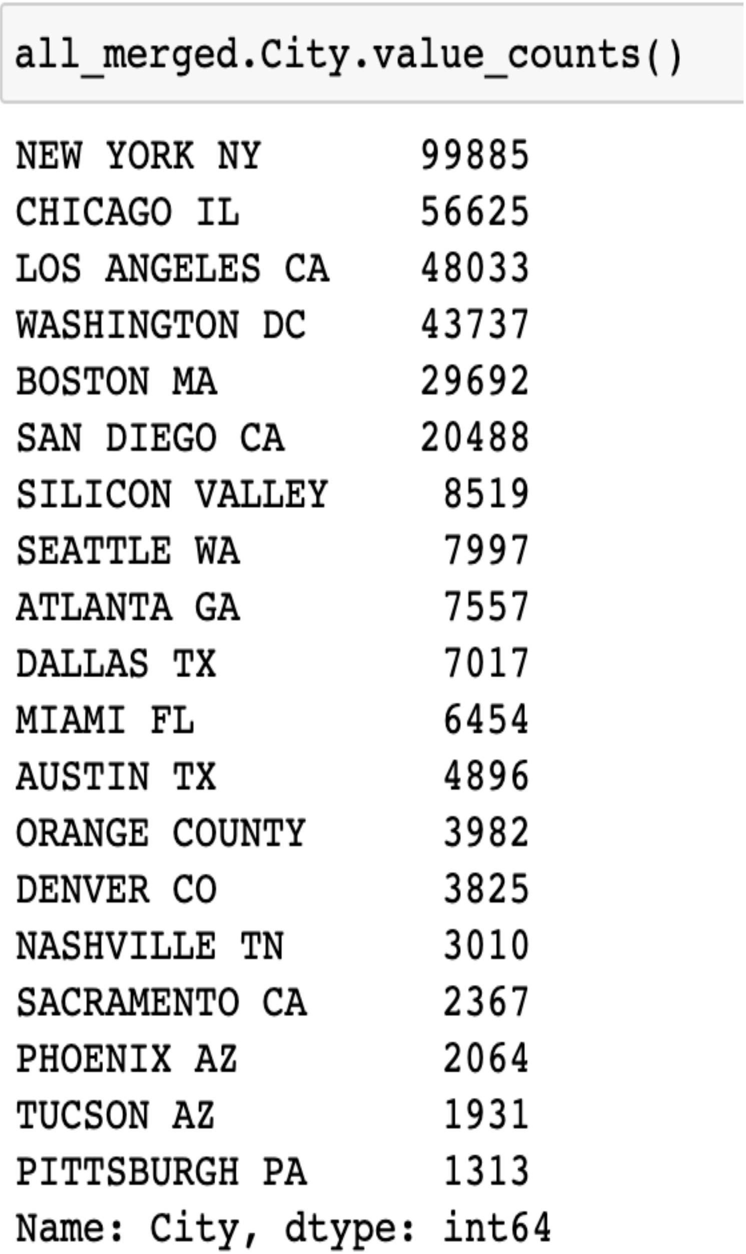




- The numbers of cab users in New York were the highest, and the cap between Chicago, which has the 2nd highest cab users, and NYC cab users is pretty big. Los Angeles and Washington have almost similar cab users. Boston and San Diego have a gap of about 9000 total users. Pittsburgh has the lowest number of cab users.

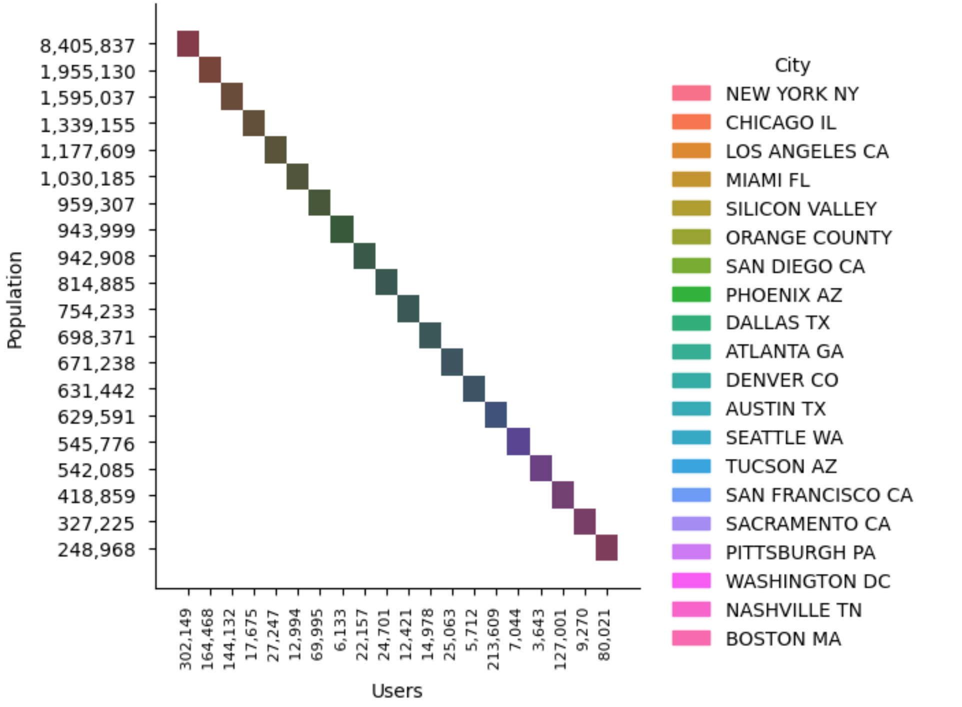
  Even Though Chicago's population is about 2 million, they have 165,000 cab users. On the other hand, New City's total population is about 8.5 million, but the total number of cab users is only 300,000. That is because NYC has a better transportation system like subways, which is why people prefer to take trains which are cheaper than cabs, and NYC cabs might take longer time than trains to reach the destination.

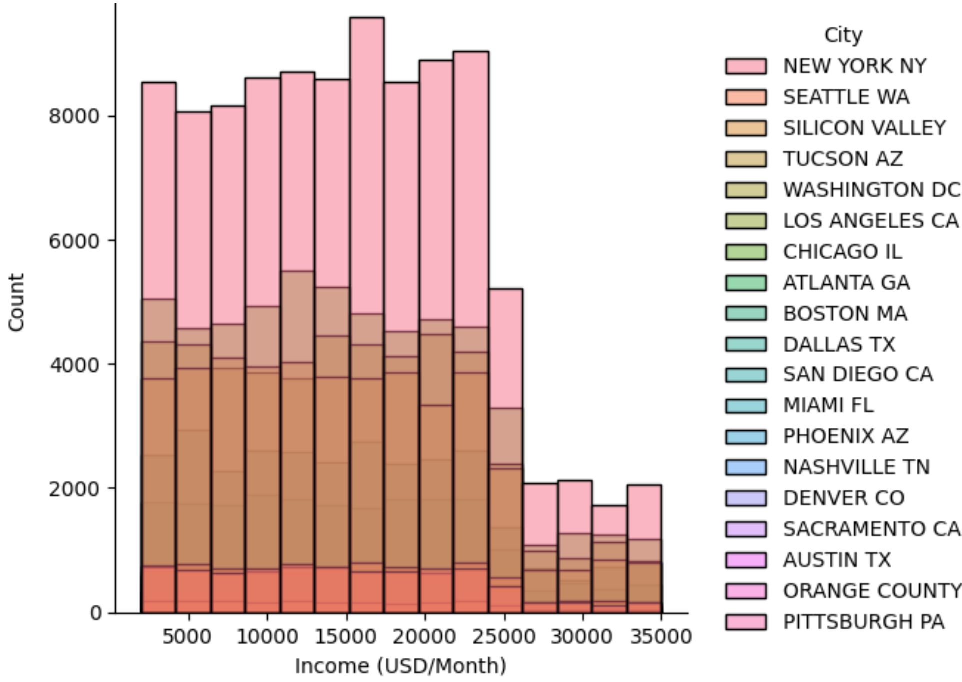
Data Glacier
Your Deep Learning Partner



```
all_merged.City.value_counts()
```

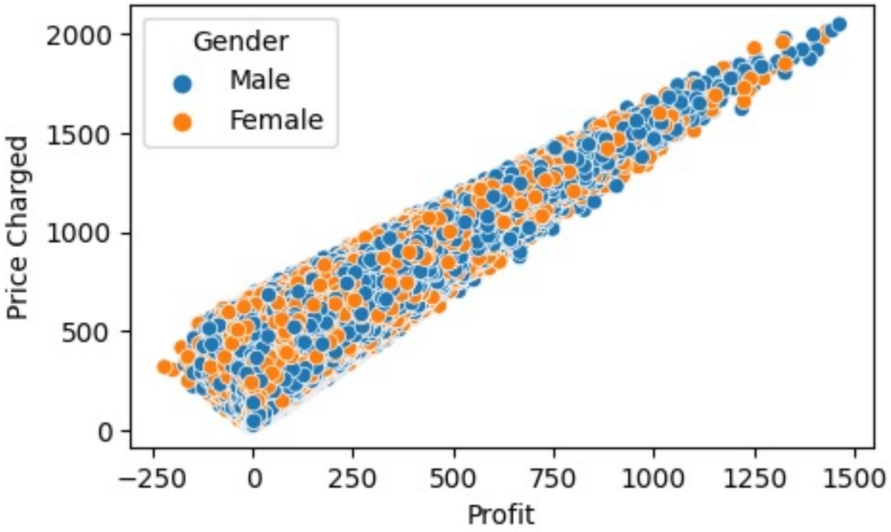
```
NEW YORK NY       99885
CHICAGO IL        56625
LOS ANGELES CA    48033
WASHINGTON DC     43737
BOSTON MA         29692
SAN DIEGO CA      20488
SILICON VALLEY     8519
SEATTLE WA         7997
ATLANTA GA         7557
DALLAS TX          7017
MIAMI FL           6454
AUSTIN TX          4896
ORANGE COUNTY      3982
DENVER CO          3825
NASHVILLE TN       3010
SACRAMENTO CA      2367
PHOENIX AZ         2064
TUCSON AZ          1931
PITTSBURGH PA      1313
Name: City, dtype: int64
```

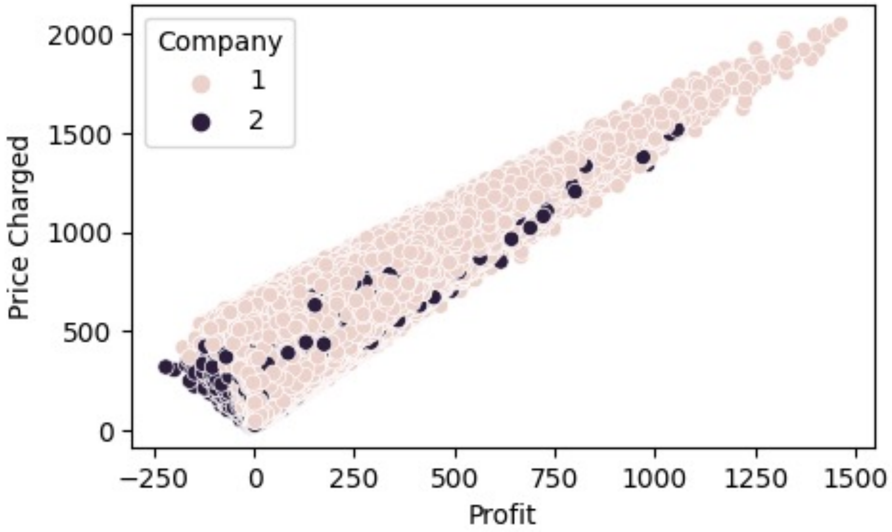


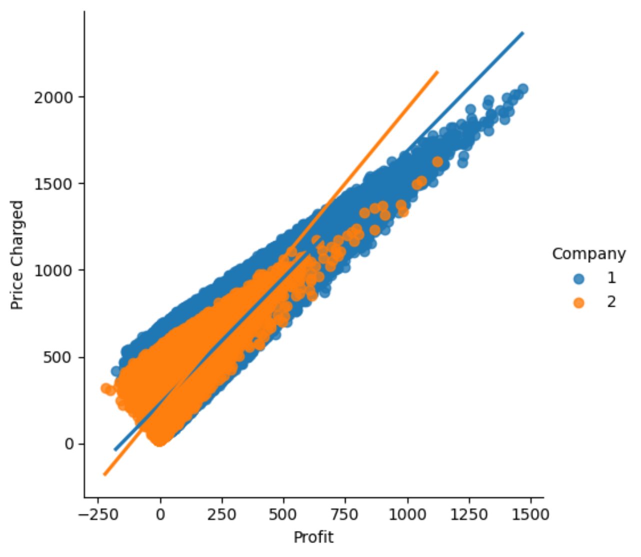


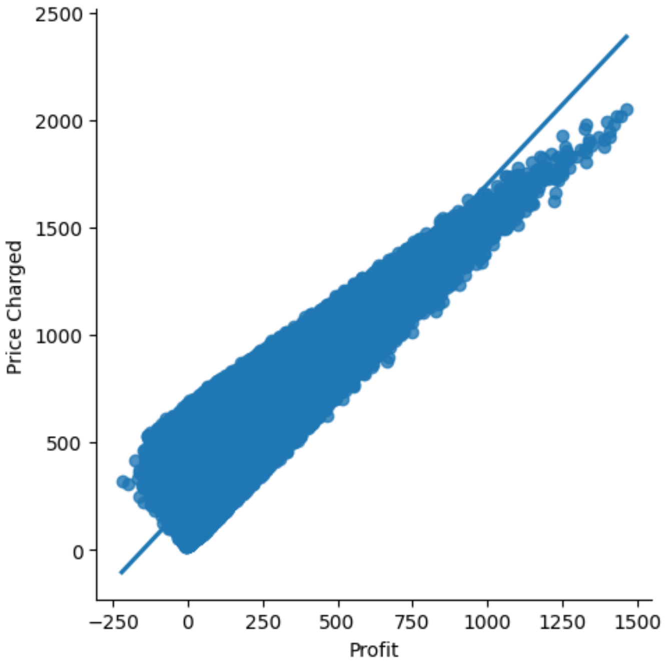
➢ Profit Analysis of both Cab Companies

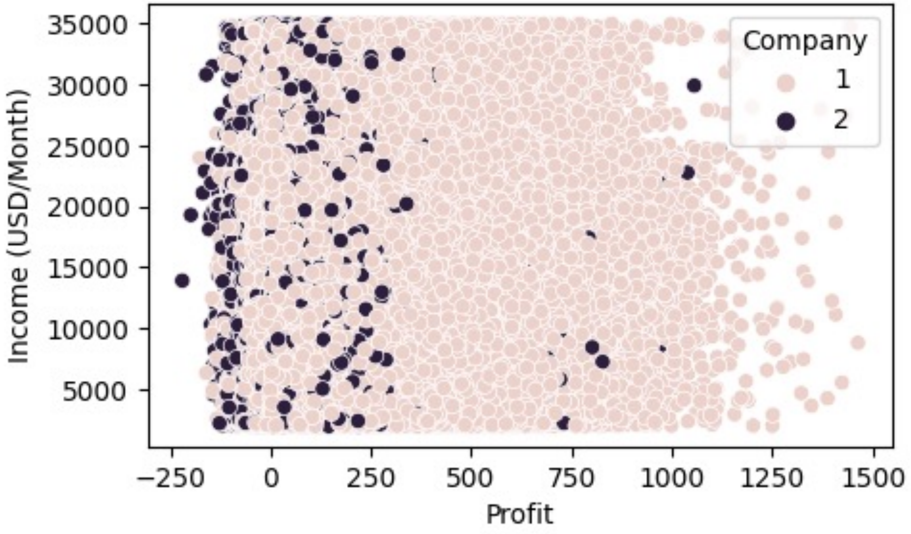
Pink Cab drivers have the majority of the profit compared to the price that was charged as we can see in all these plots. Below is the lmplot which is a scatter plot onto the FacetGrid. This plot combines a Regression Plot with multiple fits.











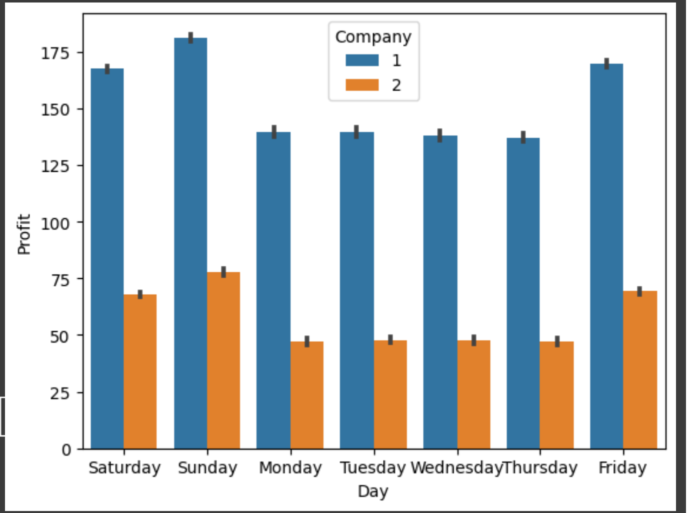
Data Glacier

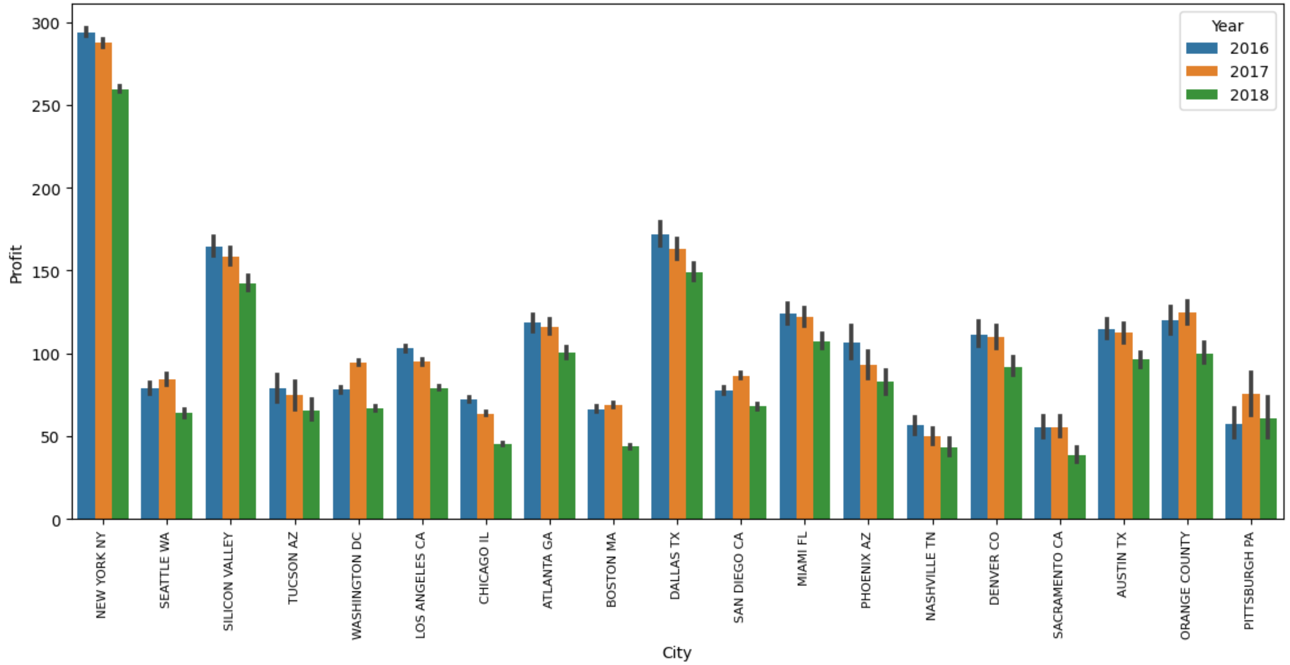
Your Deep Learning Partner

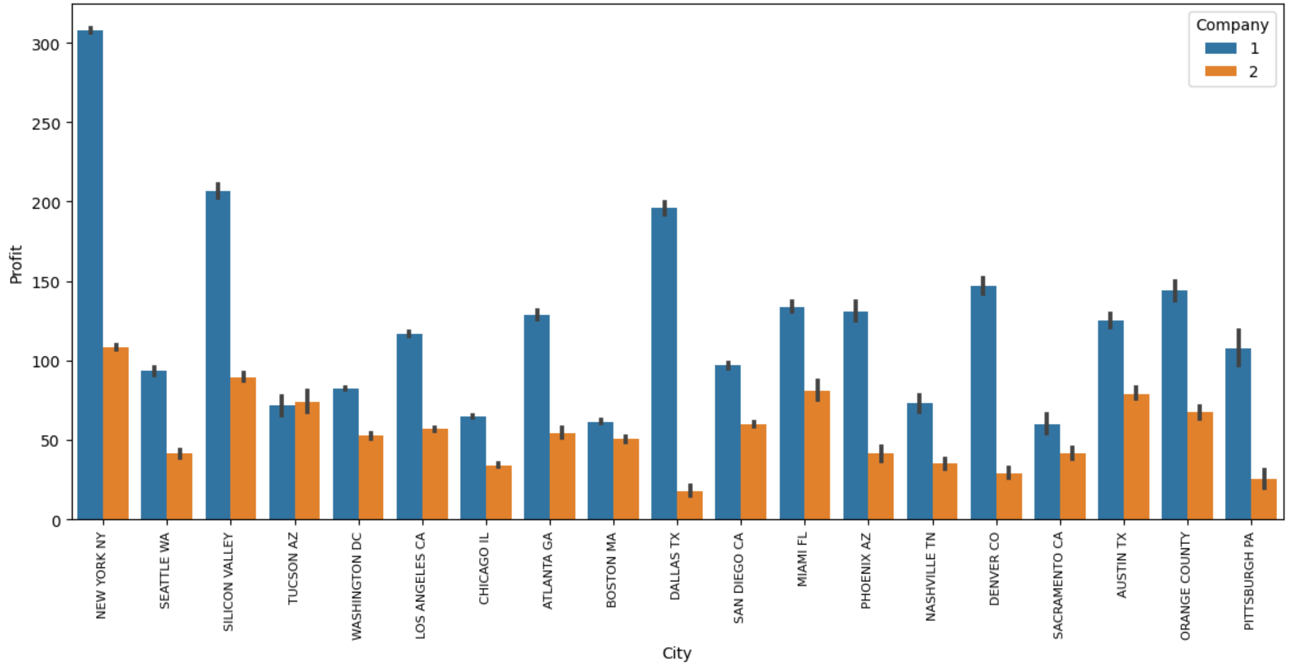
- Most of the cities has the highest profit in 2016 except Seattle, Washington, Boston, San Diego, Orange County, and Pittsburg. Profits for Dallas were significantly higher than most of the cities even though the cab users in Texas were less than in the other cities. Silicon Valley also has more profits than any other city.

  Yellow Company's profits were higher as we can see in the below plot during the weekends. The highest profits were on Sunday, and the lowest profits were on Thursday. In Dallas, the yellow company had huge profits compared to the pink company which has the lowest profits among all the other cities.

  We can clearly see here that Yellow Company is performing better.

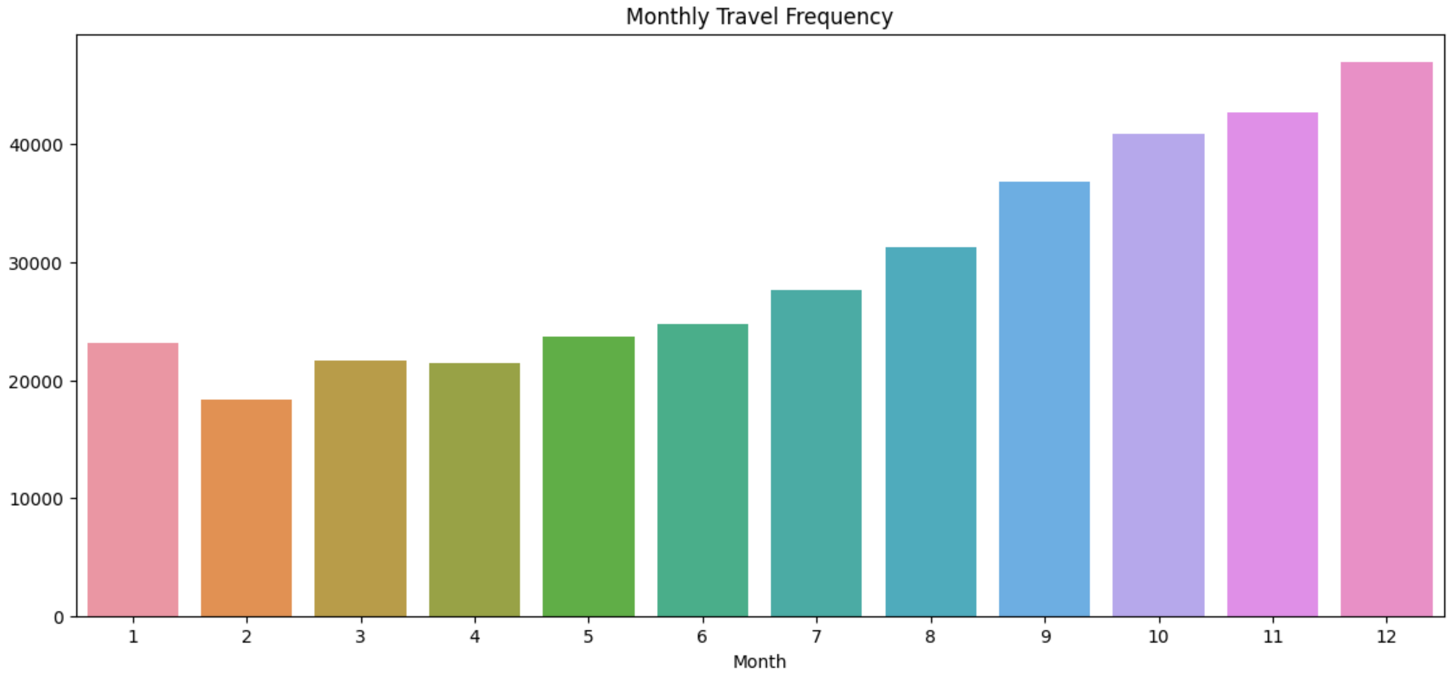


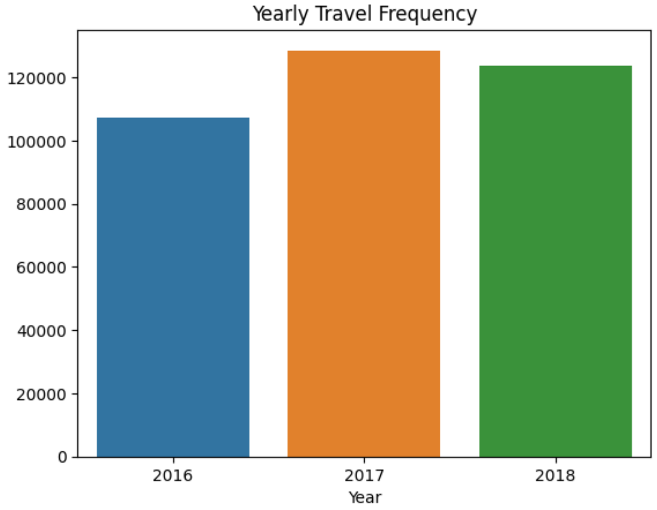




- The highest Frequency in which the users traveled with a cab was in December month, and the lowest was in February month.

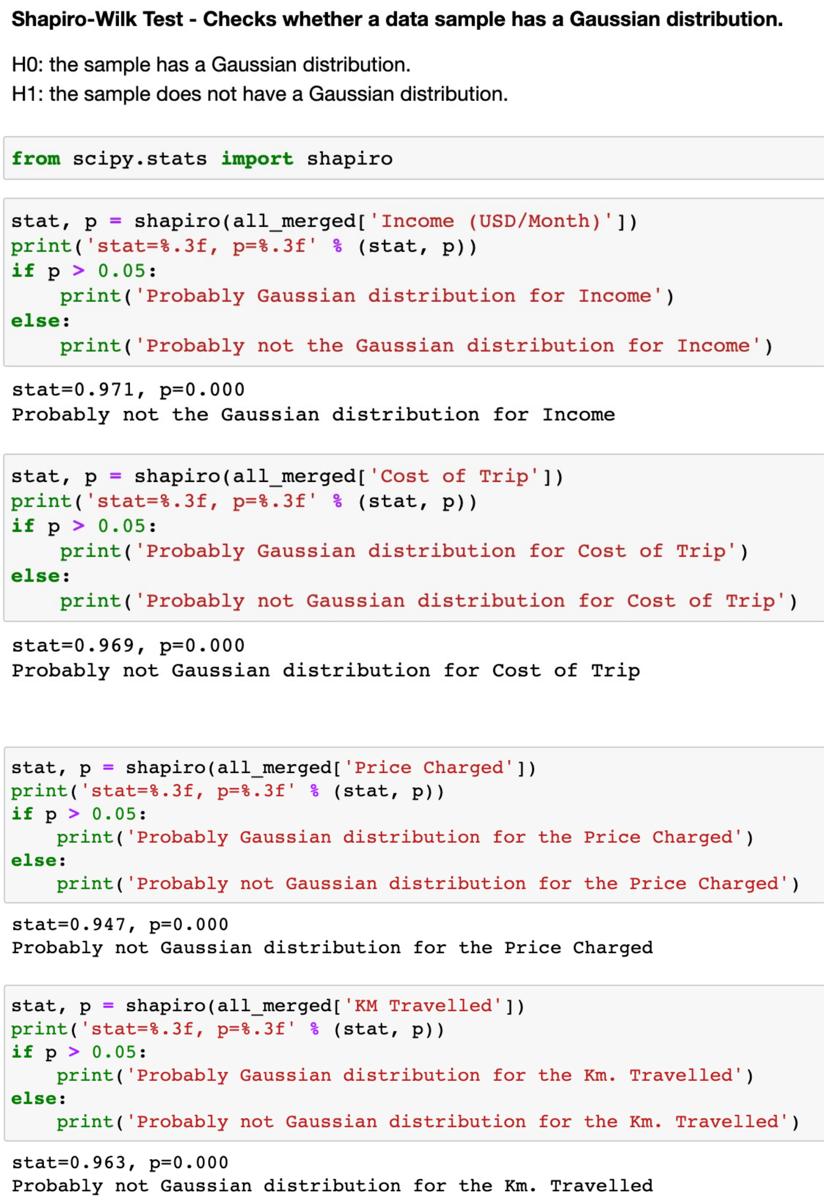
  The year 2017 had the highest travel frequency and 2016 had the lowest travel frequency. We can conclude that even though the travel frequency was lowest in 2016, the profits were highest among multiple cities.





Data Glacier
Your Deep Learning Partner

Hypothesis Testing Results



Shapiro-Wilk Test - Checks whether a data sample has a Gaussian distribution.

H0: the sample has a Gaussian distribution.
H1: the sample does not have a Gaussian distribution.

```
from scipy.stats import shapiro
```

```
stat, p = shapiro(all_merged['Income (USD/Month)'])
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('Probably Gaussian distribution for Income')
else:
    print('Probably not the Gaussian distribution for Income')
```

```
stat=0.971, p=0.000
Probably not the Gaussian distribution for Income
```

```
stat, p = shapiro(all_merged['Cost of Trip'])
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('Probably Gaussian distribution for Cost of Trip')
else:
    print('Probably not Gaussian distribution for Cost of Trip')
```

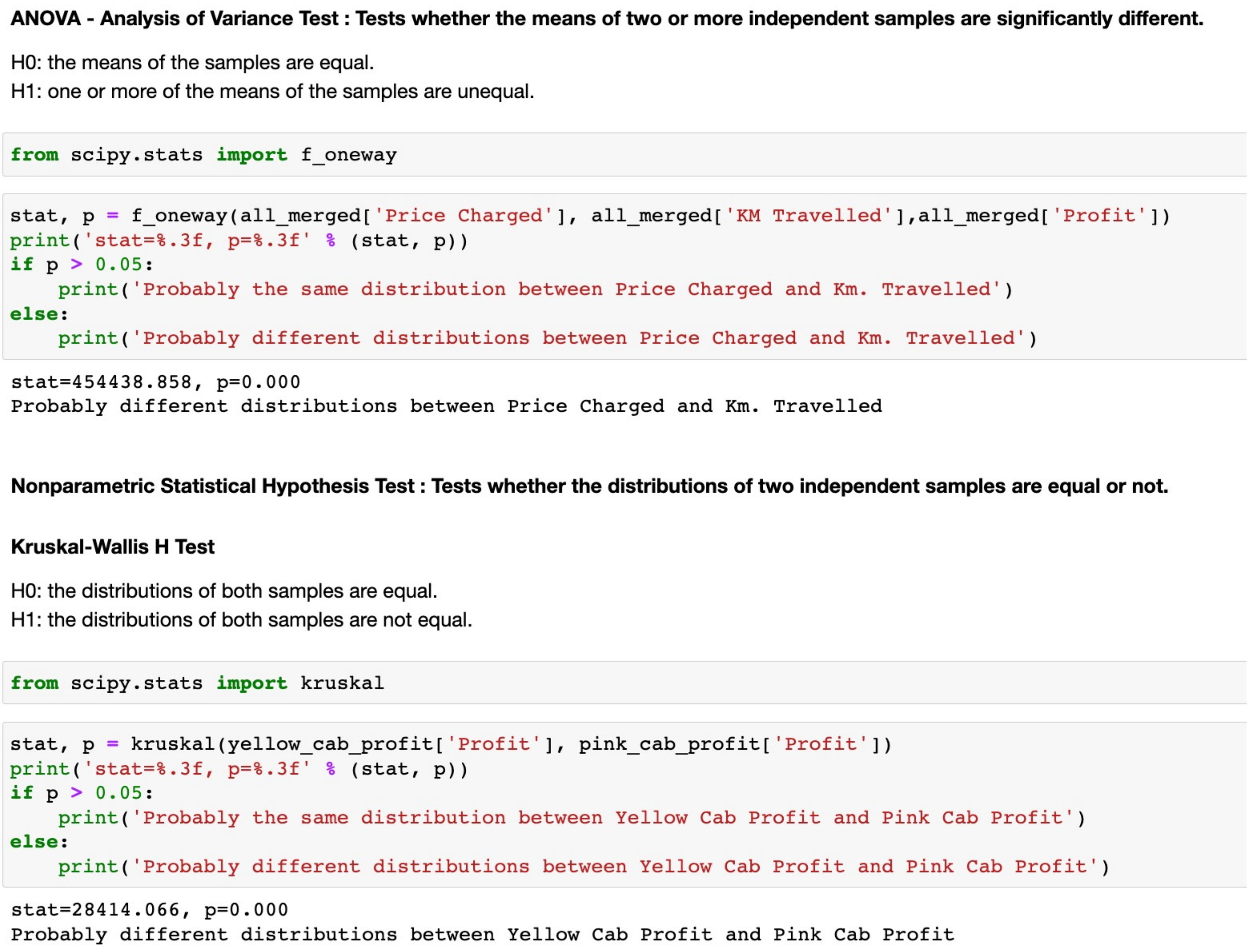
```
stat=0.969, p=0.000
Probably not Gaussian distribution for Cost of Trip
```

```
stat, p = shapiro(all_merged['Price Charged'])
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('Probably Gaussian distribution for the Price Charged')
else:
    print('Probably not Gaussian distribution for the Price Charged')
```

```
stat=0.947, p=0.000
Probably not Gaussian distribution for the Price Charged
```

```
stat, p = shapiro(all_merged['KM Travelled'])
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('Probably Gaussian distribution for the Km. Travelled')
else:
    print('Probably not Gaussian distribution for the Km. Travelled')
```

```
stat=0.963, p=0.000
Probably not Gaussian distribution for the Km. Travelled
```



ANOVA - Analysis of Variance Test : Tests whether the means of two or more independent samples are significantly different.

H0: the means of the samples are equal.
H1: one or more of the means of the samples are unequal.

```
from scipy.stats import f_oneway
```

```
stat, p = f_oneway(all_merged['Price Charged'], all_merged['KM Travelled'],all_merged['Profit'])
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('Probably the same distribution between Price Charged and Km. Travelled')
else:
    print('Probably different distributions between Price Charged and Km. Travelled')
```

```
stat=454438.858, p=0.000
Probably different distributions between Price Charged and Km. Travelled
```

Nonparametric Statistical Hypothesis Test : Tests whether the distributions of two independent samples are equal or not.

Kruskal-Wallis H Test

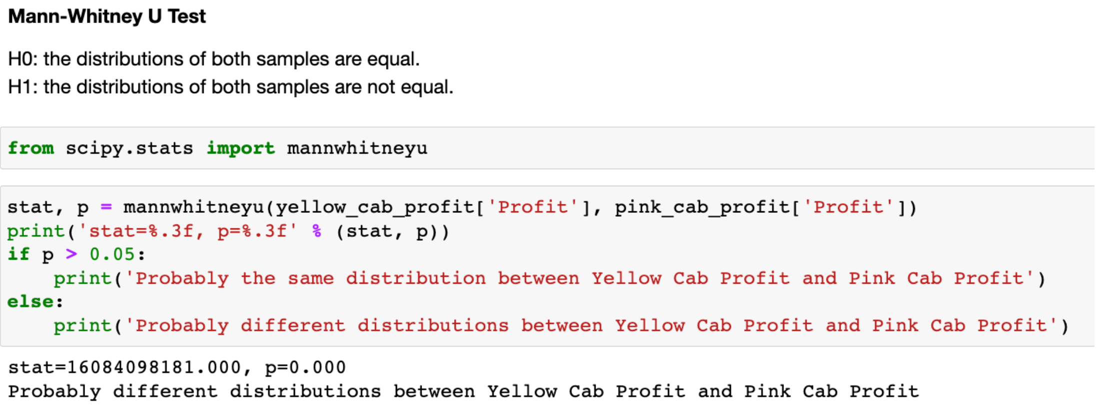
H0: the distributions of both samples are equal.
H1: the distributions of both samples are not equal.

```
from scipy.stats import kruskal
```

```
stat, p = kruskal(yellow_cab_profit['Profit'], pink_cab_profit['Profit'])
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('Probably the same distribution between Yellow Cab Profit and Pink Cab Profit')
else:
    print('Probably different distributions between Yellow Cab Profit and Pink Cab Profit')
```

```
stat=28414.066, p=0.000
Probably different distributions between Yellow Cab Profit and Pink Cab Profit
```

Using a Non-Parametric Statistical Test, we found that the distribution of Yellow Cab profits, and Pink Cab profits were different. And with all the previous analyses, we can say that investors should invest in Yellow cab, compared to the pink cab.



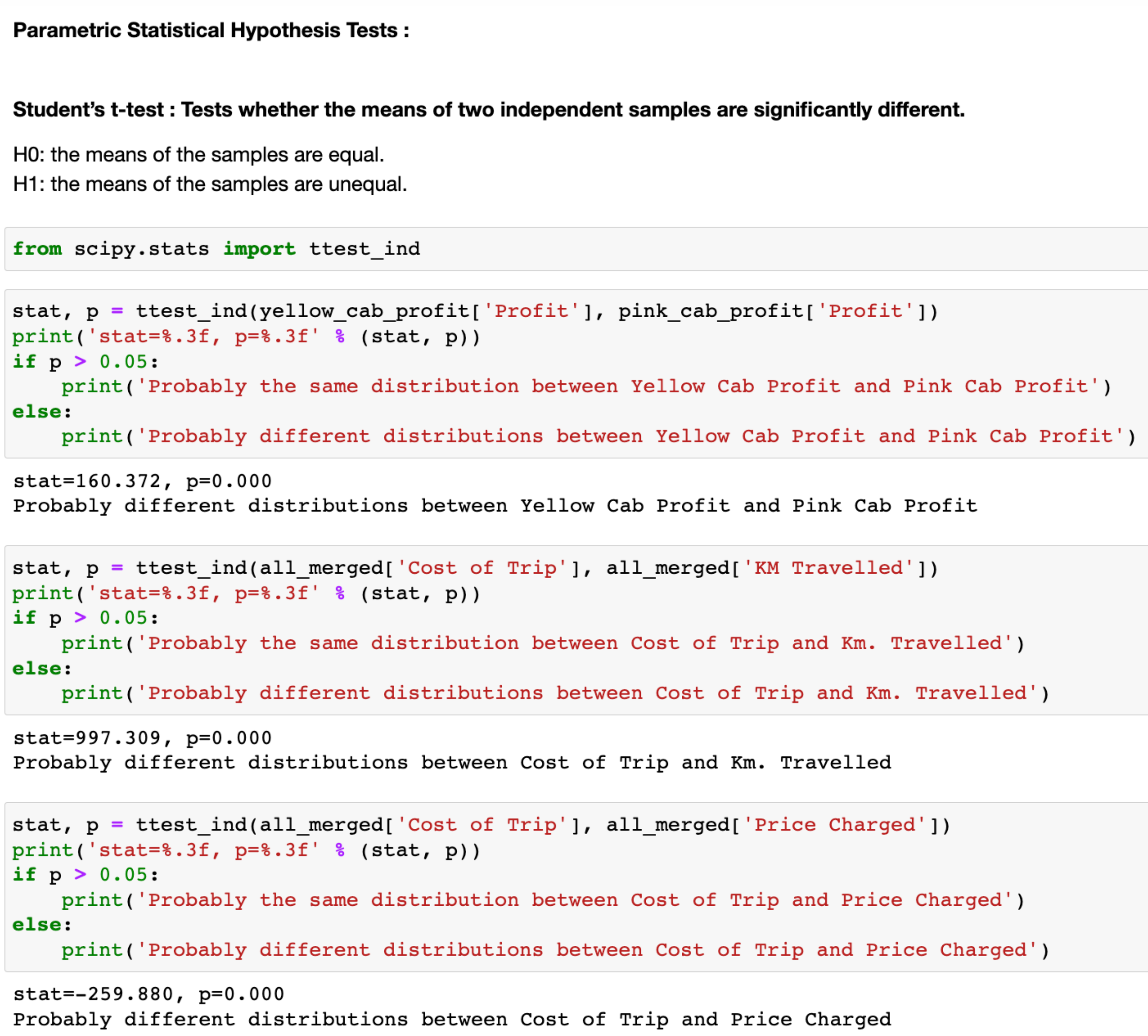
Mann-Whitney U Test

H0: the distributions of both samples are equal.
H1: the distributions of both samples are not equal.

```
from scipy.stats import mannwhitneyu
```

```
stat, p = mannwhitneyu(yellow_cab_profit['Profit'], pink_cab_profit['Profit'])
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('Probably the same distribution between Yellow Cab Profit and Pink Cab Profit')
else:
    print('Probably different distributions between Yellow Cab Profit and Pink Cab Profit')
```

```
stat=16084098181.000, p=0.000
Probably different distributions between Yellow Cab Profit and Pink Cab Profit
```



Parametric Statistical Hypothesis Tests :

Student's t-test : Tests whether the means of two independent samples are significantly different.

H0: the means of the samples are equal.
H1: the means of the samples are unequal.

```
from scipy.stats import ttest_ind
```

```
stat, p = ttest_ind(yellow_cab_profit['Profit'], pink_cab_profit['Profit'])
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('Probably the same distribution between Yellow Cab Profit and Pink Cab Profit')
else:
    print('Probably different distributions between Yellow Cab Profit and Pink Cab Profit')
```

```
stat=160.372, p=0.000
Probably different distributions between Yellow Cab Profit and Pink Cab Profit
```

```
stat, p = ttest_ind(all_merged['Cost of Trip'], all_merged['KM Travelled'])
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('Probably the same distribution between Cost of Trip and Km. Travelled')
else:
    print('Probably different distributions between Cost of Trip and Km. Travelled')
```

```
stat=997.309, p=0.000
Probably different distributions between Cost of Trip and Km. Travelled
```

```
stat, p = ttest_ind(all_merged['Cost of Trip'], all_merged['Price Charged'])
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('Probably the same distribution between Cost of Trip and Price Charged')
else:
    print('Probably different distributions between Cost of Trip and Price Charged')
```

```
stat=-259.880, p=0.000
Probably different distributions between Cost of Trip and Price Charged
```

Data Glacier

Your Deep Learning Partner

Thank You

Data Glacier
Your Deep Learning Partner

By Krutarth Haveliwala