Data Glacier

Your Deep Learning Partner

G2M Case Study

Virtual Internship

Furkan Ay

18-Sep-2024

Content

Background/Problem

Data Insights/EDA

Hypothesis testing

Background –G2M case study

- 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.

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

The analysis has been divided into four parts:

- Data Understanding
- Visualization of the correlations
- Hypothesis testing

Datasets

- **Cab_Data.csv-** File includes the data on the different cab companies
- **Customer_ID.csv**- File includes data on several details of the cab customers
- **Transaction_ID.csv**- File includes data on the transaction details
- **City.csv**- File includes data on different US cities and data on their cab users

Coding in Profit

➔ Added profit to make better inferences

```
cab_df['Profit'] = pd.DataFrame(cab_df["Price Charged"] - cab_df["Cost of Trip"])
cab_df.head()
```

	Transaction ID	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip	Profit
0	10000011	42377	Pink Cab	ATLANTA GA	30.45	370.95	313.635	57.315
1	10000012	42375	Pink Cab	ATLANTA GA	28.62	358.52	334.854	23.666
2	10000013	42371	Pink Cab	ATLANTA GA	9.04	125.20	97.632	27.568
3	10000014	42376	Pink Cab	ATLANTA GA	33.17	377.40	351.602	25.798
4	10000015	42372	Pink Cab	ATLANTA GA	8.73	114.62	97.776	16.844

Merging Dataset

```
df= cab_df.merge(transaction_df, on= 'Transaction ID').merge(customer_df, on ='Customer ID').merge(city_df, on = 'City')
df.head(4)
```

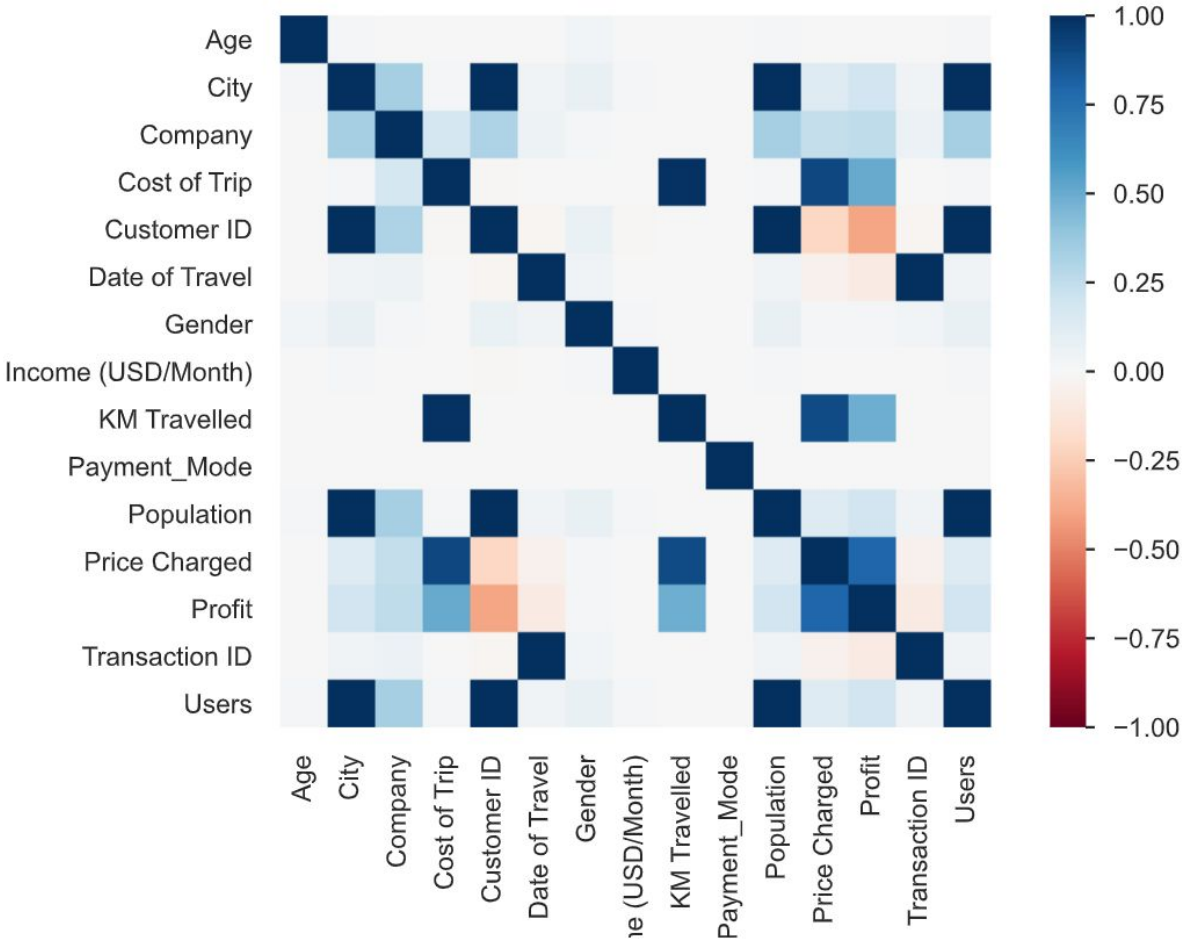
2]

Python

```
RangeIndex: 359392 entries, 0 to 359391
Data columns (total 15 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   Transaction ID      359392 non-null  int64
 1   Date of Travel      359392 non-null  int64
 2   Company             359392 non-null  object
 3   City                359392 non-null  object
 4   KM Travelled        359392 non-null  float64
 5   Price Charged       359392 non-null  float64
 6   Cost of Trip        359392 non-null  float64
 7   Profit              359392 non-null  float64
 8   Customer ID         359392 non-null  int64
 9   Payment_Mode        359392 non-null  object
 10  Gender              359392 non-null  object
 11  Age                 359392 non-null  int64
 12  Income (USD/Month)  359392 non-null  int64
 13  Population          359392 non-null  object
 14  Users               359392 non-null  object
dtypes: float64(4), int64(5), object(6)
memory usage: 41.1+ MB
```

➔ Merged all the datasets into one dataframe to compare and analyze better

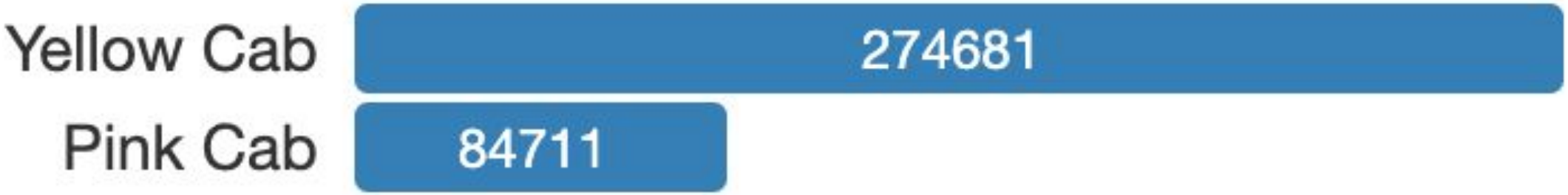
Analysis of Correlation



➔ 6 Highly Correlated Categories

1. City and Customer ID
2. Cost of trip and KM traveled
3. Date of Travel and Transaction ID
4. Price charged and Cost of trip
5. Population and City
6. Profit and Cost of trip

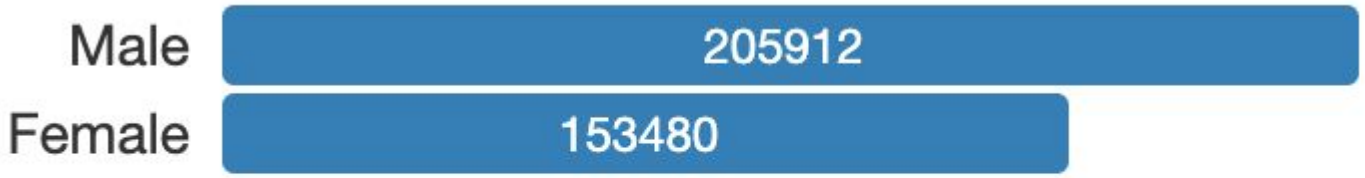
Distribution Cab Users



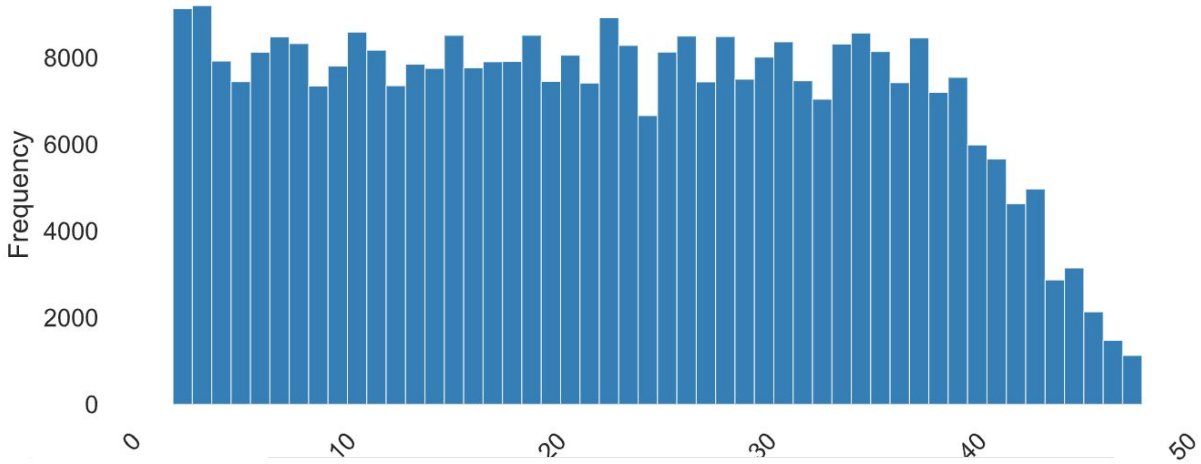
➔ From the graph above we can tell that there are more Yellow cab users than there are Pink cab users

Distribution Of Gender

- From the graph we can tell that there are more male users than female users
- This is true for both cab companies



Distribution of KM Traveled



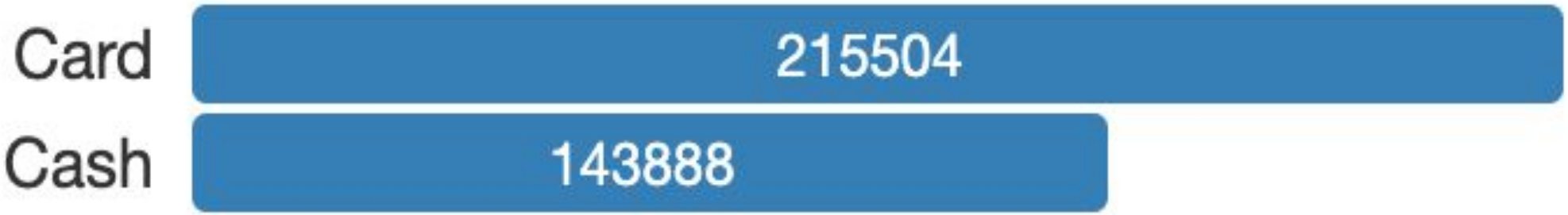
Value	Count	Frequency (%)
1.9	339	0.1%
1.92	375	0.1%
1.94	329	0.1%
1.96	383	0.1%
1.98	374	0.1%
2	362	0.1%
2.02	341	0.1%
2.04	358	0.1%
2.06	346	0.1%
2.08	369	0.1%

Minimum 10 values | Maximum 10 values

Value	Count	Frequency (%)
48	366	0.1%
47.6	381	0.1%
47.2	378	0.1%
46.8	737	0.2%
46.41	380	0.1%
46.4	356	0.1%
46.02	385	0.1%
46	336	0.1%
45.63	344	0.1%
45.6	704	0.2%

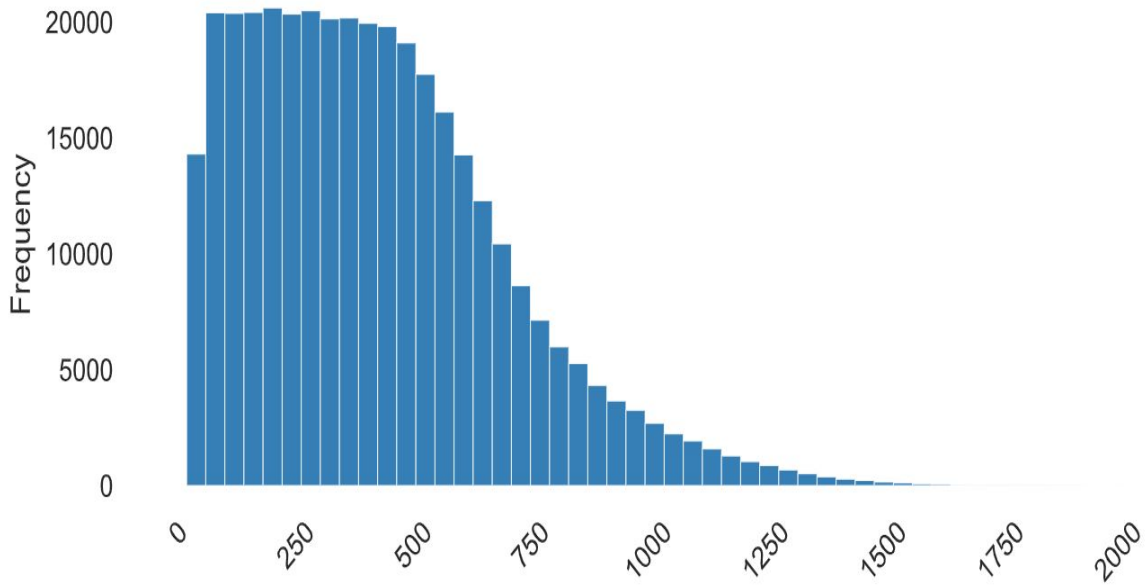
- From These graphs we can tell that the common range of KM traveled would be from 2.10km to 45 km
- We can see that 49 km is the maximum of extreme values
- We can see that

Distribution Of Payment Mode



- ➔ From the figure we can tell that users use card than pay for it with cash
- ➔ This si the case for both companies

Distribution of price charged



Value	Count	Frequency (%)
15.6	1	< 0.1%
15.75	1	< 0.1%
16.38	1	< 0.1%
16.53	1	< 0.1%
16.76	1	< 0.1%
17.03	1	< 0.1%
17.11	1	< 0.1%
17.21	1	< 0.1%
17.27	1	< 0.1%
17.46	1	< 0.1%

15.75

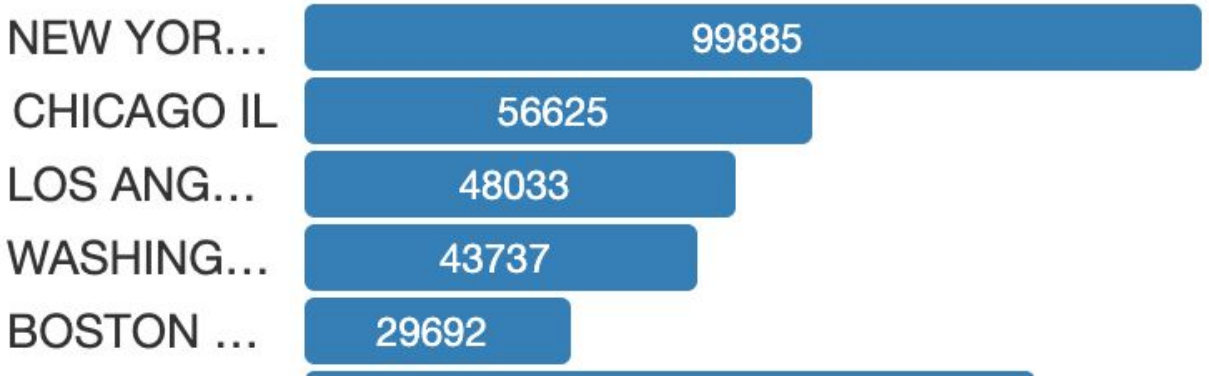
Value	Count	Frequency (%)
2048.03	1	< 0.1%
2016.7	1	< 0.1%
2013.95	1	< 0.1%
1993.83	1	< 0.1%
1981.05	1	< 0.1%
1978.79	1	< 0.1%
1957.1	1	< 0.1%
1947.91	1	< 0.1%
1925.92	1	< 0.1%
1920.59	1	< 0.1%

2013.95

- From the graphs we can see that the range of the price charged is 17.50 to 1920 depending on the distance traveled
- From the extreme values we see that the max is 2048.03
- The minimum for the extreme values is 15.6

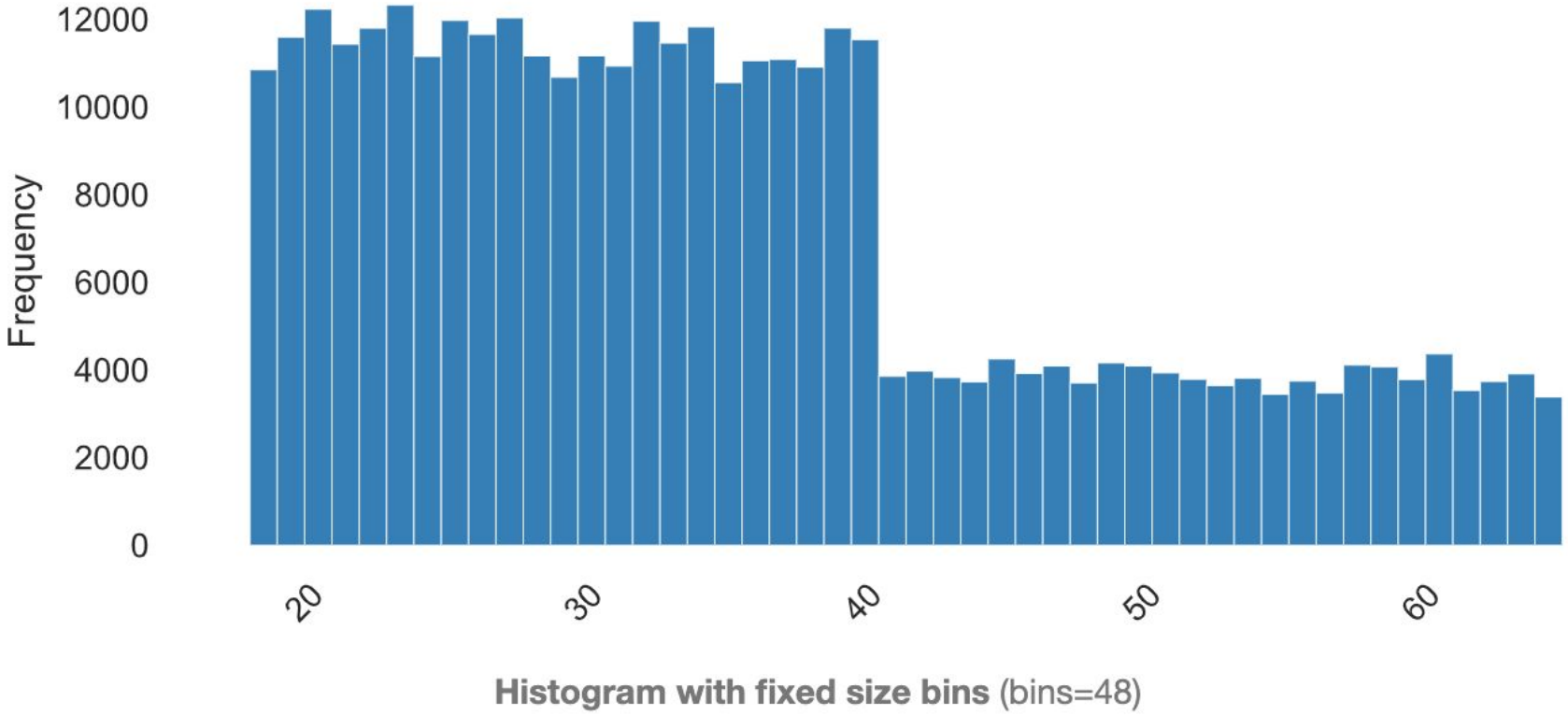
Distribution of Cities

Value	Count	Frequency (%)
NEW YORK NY	99885	27.8%
CHICAGO IL	56625	15.8%
LOS ANGELES CA	48033	13.4%
WASHINGTON DC	43737	12.2%
BOSTON MA	29692	8.3%
SAN DIEGO CA	20488	5.7%
SILICON VALLEY	8519	2.4%
SEATTLE WA	7997	2.2%
ATLANTA GA	7557	2.1%
DALLAS TX	7017	2.0%



➔ From the graphs it is seen that the cab users are mostly in New York,Chicago,Los Angeles, Washington DC and Boston.

Distribution of Age



➔ From this we can see that there is a younger distribution in cab drivers

Hypothesis 1

H0= One gender makes more profit than the other H1= Both genders have the same impact on the company profit

```
df_male = df[(df['Gender'] == 'Male') & (df['Company'] == 'Yellow Cab')]
df_female = df[(df['Gender'] == 'Female') & (df['Company'] == 'Yellow Cab')]

profit_male = df_male['Profit'].values
profit_female = df_female['Profit'].values

t_stat, p_value = ttest_ind(profit_male, profit_female, equal_var = True )
print('t_statistics:', t_stat, 'p_value:', p_value)


if p_value < 0.05:
    print("Hypthesis H0 rejected")
else:
    print("Hypothesis H1 rejected")

df_male = df[(df['Gender'] == 'Male') & (df['Company'] == 'Pink Cab')]
df_female = df[(df['Gender'] == 'Female') & (df['Company'] == 'Pink Cab')]

profit_male = df_male['Profit'].values
profit_female = df_female['Profit'].values

t_stat, p_value = ttest_ind(profit_male, profit_female, equal_var = True )
print('t_statistics:', t_stat, 'p_value:', p_value)


if p_value < 0.05:
    print("Hypthesis H1 rejected")
else:
    print("Hypothesis H0 rejected")
```

```
t_statistics: 10.315494207195322 p_value: 6.060473042494056e-25
Hypthesis H0 rejected
t_statistics: 1.5754642478511207 p_value: 0.11515305900425798
Hypothesis H0 rejected
```

➔ From the hypothesis H0 benign rejected we see that gender has the same impact

Hypothesis 2

H0= Younger than 30 makes more profit H1 = Older than 30 makes more profit

```
df_young = df[(df['Age'] <30 ) & (df['Company'] == 'Yellow Cab')]
df_old = df[(df['Age'] >=30 ) & (df['Company'] == 'Yellow Cab')]

profit_young = df_young['Profit'].values
profit_old = df_old['Profit'].values

t_stat, p_value = ttest_ind(profit_young, profit_old, equal_var = True )
print('t_statistics:', t_stat, 'p_value:', p_value)


if p_value < 0.05:
    print("Hypthesis H0 rejected")
else:
    print("Hypothesis H1 rejected")


df_young = df[(df['Age'] <30 ) & (df['Company'] == 'Pink Cab')]
df_old = df[(df['Age'] >=30 ) & (df['Company'] == 'Pink Cab')]

profit_young = df_young['Profit'].values
profit_old = df_old['Profit'].values

t_stat, p_value = ttest_ind(profit_young, profit_old, equal_var = True )
print('t_statistics:', t_stat, 'p_value:', p_value)


if p_value < 0.05:
    print("Hypthesis H0 rejected")
else:
    print("Hypothesis H1 rejected")
```

```
t_statistics: 1.675783219486658 p_value: 0.09378180031137037
Hypothesis H1 rejected
t_statistics: 1.8177525977759004 p_value: 0.06910548546568591
Hypothesis H1 rejected
```

➔ From H1 being rejected we know that H0 is the correct hypothesis meaning younger people make more profit

Hypothesis 3

H0 = Pink Cabs have more profit H1= Yellow cabs have more profit

```
df_yellow = df[(df['Company'] == 'Yellow Cab')]
df_pink = df[(df['Company'] == 'Pink Cab')]

profit_yellow = df_yellow['Profit'].values
profit_pink = df_pink['Profit'].values

t_stat, p_value = ttest_ind(profit_yellow, profit_pink, equal_var = True )
print('t_statistics:', t_stat, 'p_value:', p_value)


if p_value < 0.05:
    print("Hypthesis H0 rejected")
else:
    print("Hypothesis H1 rejected")
```

```
t_statistics: 160.37151759478058 p_value: 0.0
Hypthesis H0 rejected
```

➔ From hypothesis H0 being rejected it can be shown that yellow cabs make more profit

Recommendation

- ➔ Because the Yellow cabs have more users seen from my EDA and they generate more profit seen from my hypothesis testing I recommend that the XYZ firm invests in the Yellow cab company as there is more advantages when compared to the pink cabs

Thank You