Data Intake Report

Name: G2M-insight-for-Cab-Investment-firm
Report date: 31 July
Internship Batch: LISUM35
Version:<1.0>
Data intake by:Janesh Hasija
Data intake reviewer: Data Glacier
Data storage location:

Tabular data details:

Total number of observations	359392
Total number of files	1
Total number of features	7
Base format of the file	csv
Size of the data	20,1 MB

Tabular data details: City_Data

Total number of observations	20
Total number of files	1
Total number of features	3
Base format of the file	csv
Size of the data	759 bayt

Tabular data details: Transacition_Data

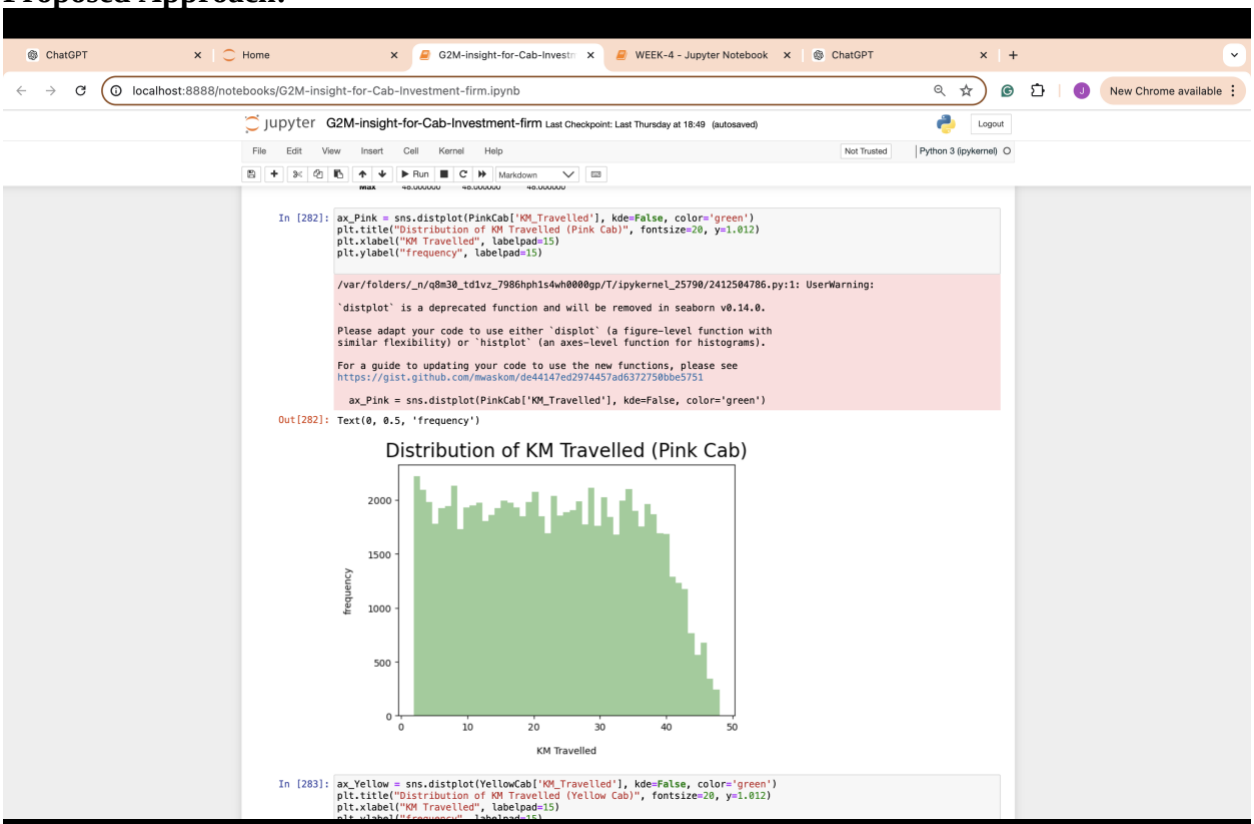
Total number of observations	440098
Total number of files	1
Total number of features	3
Base format of the file	csv
Size of the data	8,58 MB

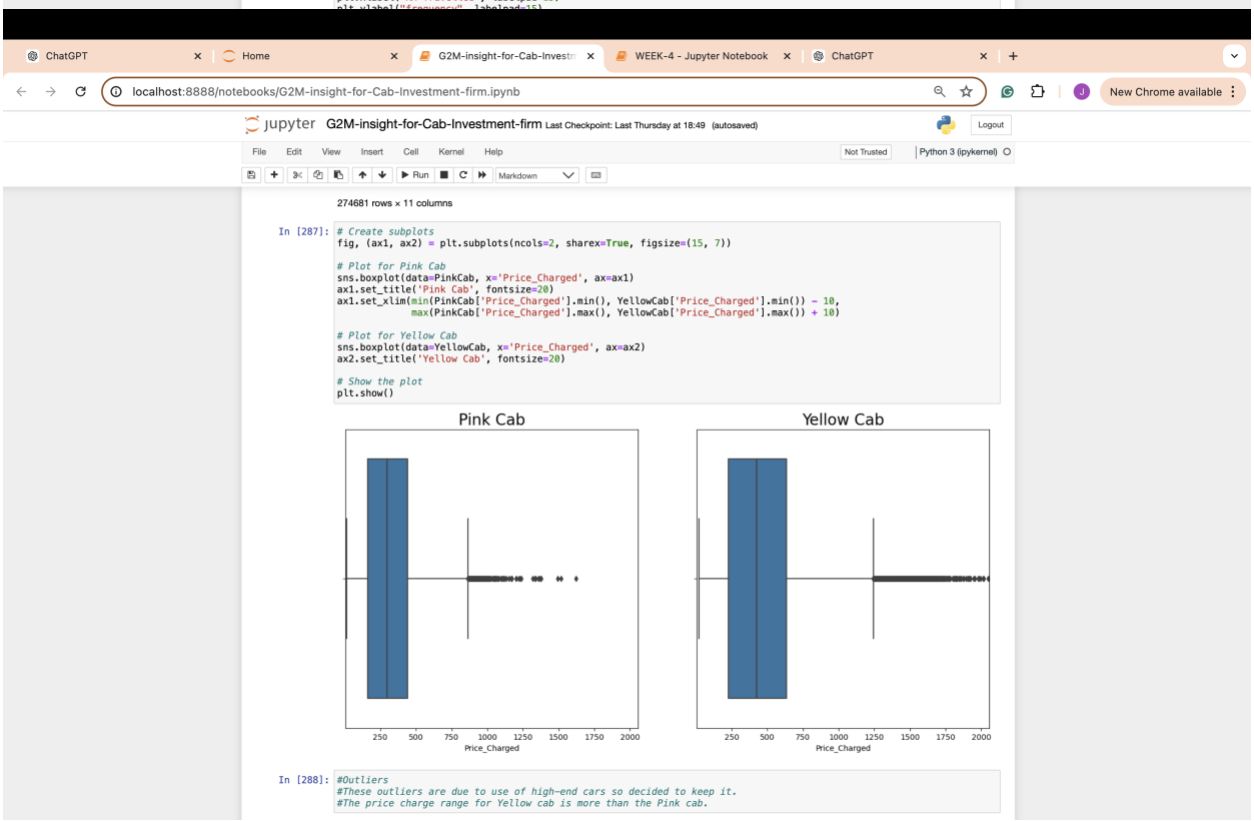
Tabular data details: Customer ID_Data

Total number of observations	49171
Total number of files	1
Total number of features	4
Base format of the file	csv
Size of the data	1,00 MB

Note: Replicate same table with file name if you have more than one file.

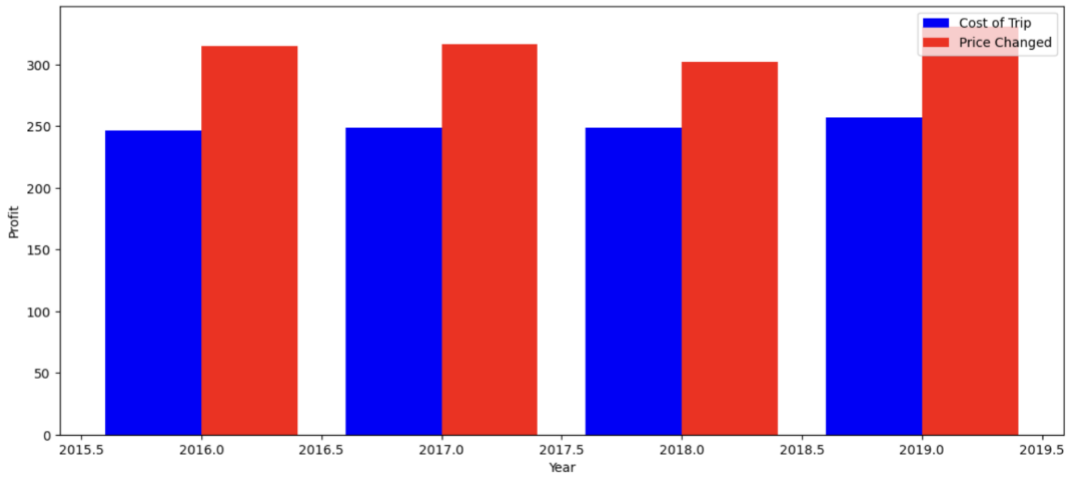
Proposed Approach:

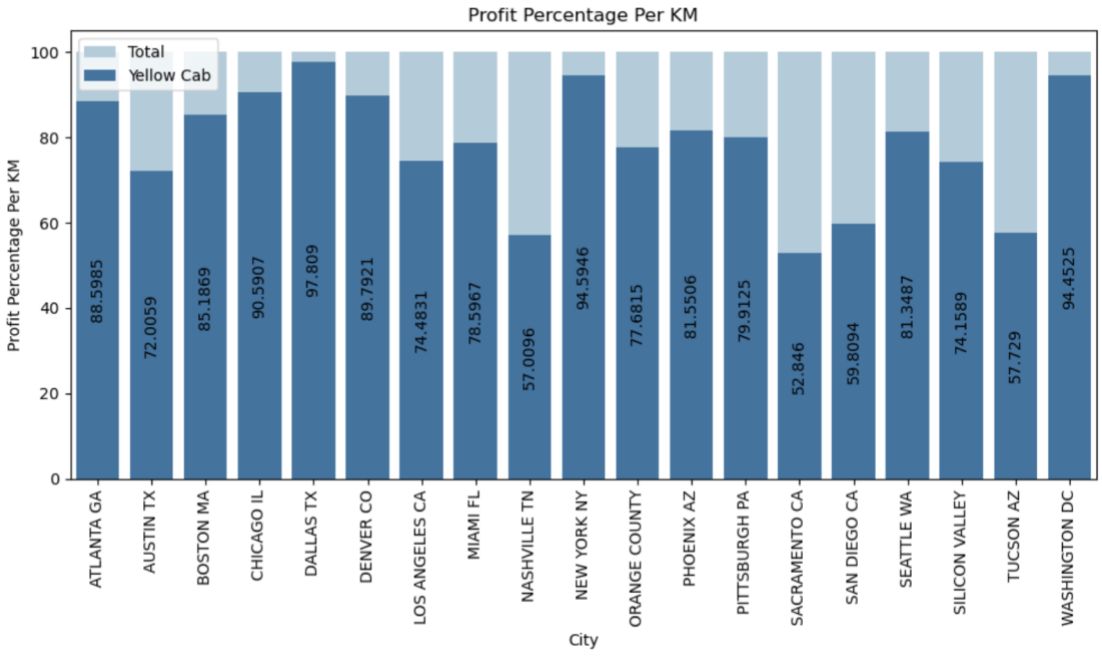




In [299]:
```
plot1 = cab_data[cab_data.Company=='Pink Cab'].groupby('Year').Transaction_ID.count()
# plot1 = PinkCab.groupby('Year')['Transaction_ID'].count()
plot3 = cab_data[cab_data.Company=='Pink Cab'].groupby('Year').Price_Charged.mean()
plot4 = cab_data[cab_data.Company=='Pink Cab'].groupby('Year').Cost_of_Trip.mean()

plt.figure(figsize=(14,6))
ax = plt.subplot(111)
ax.bar(plot1.index-0.2, plot4.values, width=0.4, color='blue', align='center',label='Cost of Trip')
ax.bar(plot1.index+0.2, plot3.values, width=0.4, color='red', align='center',label='Price Changed')
plt.ylabel('Profit')
plt.xlabel('Year')
plt.legend()
plt.show()
```





Hypothesis 1:

H0:Gender has not effect on company profit H1:Gender has effect on company profit

In [582]:
```
dict={
    "Profit":"mean"
}

group1= data.groupby((data.Company== 'Pink Cab')& (data.Gender=='Male')).agg(dict)
group2= data.groupby((data.Company== 'Pink Cab')& (data.Gender=='Female')).agg(dict)

t_stat,p_value=ttest_ind(group1, group2, equal_var=True)

print('t_statistics:', t_stat, '\np_value:', p_value)


if p_value<0.05:
  print("Reject Null Hypothesis (H0)")
elif p_value>=0.05:
  print("Reject Alternative Hypothesis(H1)")


dict={
    "Profit":"mean"
}

group1= data.groupby((data.Company== 'Yellow Cab')& (data.Gender=='Male')).agg(dict)
group2= data.groupby((data.Company== 'Yellow Cab')& (data.Gender=='Female')).agg(dict)

t_stat,p_value=ttest_ind(group1, group2, equal_var=True)

print('t_statistics:', t_stat, '\np_value:', p_value)


if p_value<0.05:
  print("Reject Null Hypothesis (H0)")
elif p_value>=0.05:
  print("Reject Alternative Hypothesis(H1)")
```

```
t_statistics: [0.02786795]
p_value: [0.98029821]
Reject Alternative Hypothesis(H1)
t_statistics: [-0.08386648]
p_value: [0.94080144]
Reject Alternative Hypothesis(H1)
```

Hypothesis 2:

H0:Payment mode has not effect on company profit H1:Payment mode has effect on company profit

In [585]:
```
dict={
    "Profit":"mean"
}

group1= data.groupby((data.Company== 'Pink Cab')& (data.Payment_Mode=='Cash')).agg(dict)
group2= data.groupby((data.Company== 'Pink Cab')& (data.Payment_Mode=='Card')).agg(dict)

t_stat,p_value=ttest_ind(group1, group2, equal_var=True)

print('t_statistics:', t_stat, '\np_value:', p_value)


if p_value<0.05:
  print("Reject Null Hypothesis (H0)")
elif p_value>=0.05:
  print("Reject Alternative Hypothesis(H1)")


dict={
    "Profit":"mean"
}

group1= data.groupby((data.Company== 'Yellow Cab')& (data.Payment_Mode=='Cash')).agg(dict)
group2= data.groupby((data.Company== 'Yellow Cab')& (data.Payment_Mode=='Card')).agg(dict)

t_stat,p_value=ttest_ind(group1, group2, equal_var=True)

print('t_statistics:', t_stat, '\np_value:', p_value)


if p_value<0.05:
  print("Reject Null Hypothesis (H0)")
elif p_value>=0.05:
  print("Reject Alternative Hypothesis(H1)")
```

```
t_statistics: [-0.03834307]
p_value: [0.97289731]
Reject Alternative Hypothesis(H1)
t_statistics: [0.17875066]
p_value: [0.8746019]
Reject Alternative Hypothesis(H1)
```

Hypothesis 3:

H0:Age has not effect on company profit H1:Age has effect on company profit

In [587]:
```
dict={
    "Profit":"mean"
}

group1= data.groupby((data.Company== 'Pink Cab')& (data.Age<=50)).agg(dict)
group2= data.groupby((data.Company== 'Pink Cab')& (data.Age>50)).agg(dict)

t_stat,p_value=ttest_ind(group1, group2, equal_var=True)

print('t_statistics:', t_stat, '\np_value:', p_value)


if p_value<0.05:
  print("Reject Null Hypothesis (H0)")
elif p_value>=0.05:
  print("Reject Alternative Hypothesis(H1)")


dict={
    "Profit":"mean"
}

group1= data.groupby((data.Company== 'Yellow Cab')& (data.Age<=50)).agg(dict)
group2= data.groupby((data.Company== 'Yellow Cab')& (data.Age>50)).agg(dict)

t_stat,p_value=ttest_ind(group1, group2, equal_var=True)

print('t_statistics:', t_stat, '\np_value:', p_value)


if p_value<0.05:
  print("Reject Null Hypothesis (H0)")
elif p_value>=0.05:
  print("Reject Alternative Hypothesis(H1)")
```

```
t_statistics: [0.13340069]
p_value: [0.90608835]
Reject Alternative Hypothesis(H1)
t_statistics: [-0.50355161]
p_value: [0.66456451]
Reject Alternative Hypothesis(H1)
```

Hypothesis 4:

H0:The variable Year has not a positive correlation with mileage H1:The variable Year has a positive correlation with mileage

In [590]:
```
agg_dict = {
    'KM_Travelled': 'mean'
      }
group1 = data.groupby((data.Year==2018)&(data.Company=='Pink Cab')).agg(agg_dict)
group2 = data.groupby((data.Year==2017)&(data.Company=='Pink Cab')).agg(agg_dict)

t_stat,p_value=ttest_ind(group1 , group2, equal_var=True)

print('t_statistics:', t_stat, '\np_value:', p_value)

if p_value<0.05:
  print("Reject Null Hypothesis (H0)")
elif p_value>=0.05:
  print("Reject Alternative Hypothesis(H1)")

group1 = data.groupby((data.Year==2018)&(data.Company=='Yellow Cab')).agg(agg_dict)
group2 = data.groupby((data.Year==2017)&(data.Company=='Yellow Cab')).agg(agg_dict)

t_stat,p_value=ttest_ind(group1 , group2, equal_var=True)

print('t_statistics:', t_stat, '\np_value:', p_value)

if p_value<0.05:
  print("Reject Null Hypothesis (H0)")
elif p_value>=0.05:
  print("Reject Alternative Hypothesis(H1)")
```

```
t_statistics: [-0.66243408]
p_value: [0.57581736]
Reject Alternative Hypothesis(H1)
t_statistics: [-0.42772975]
p_value: [0.71050085]
Reject Alternative Hypothesis(H1)
```

Hypothesis 5:

H0:The variable Income has not a positive correlation with mileage H1:The variable Income has a positive correlation with mileage

In [596]:
```
agg_dict = {
    'KM_Travelled': 'mean'
      }
group1 = data.groupby((data.Income<=15048.822937071498)&(data.Company=='Pink Cab')).agg(agg_dict)
group2 = data.groupby((data.Income>15048.822937071498)&(data.Company=='Pink Cab')).agg(agg_dict)

t_stat,p_value=ttest_ind(group1 , group2, equal_var=True)

print('t_statistics:', t_stat, '\np_value:', p_value)

if p_value<0.05:
  print("Reject Null Hypothesis (H0)")
elif p_value>=0.05:
  print("Reject Alternative Hypothesis(H1)")

group1 = data.groupby((data.Income<=15048.822937071498)&(data.Company=='Yellow Cab')).agg(agg_dict)
group2 = data.groupby((data.Income>15048.822937071498)&(data.Company=='Yellow Cab')).agg(agg_dict)

t_stat,p_value=ttest_ind(group1 , group2, equal_var=True)

print('t_statistics:', t_stat, '\np_value:', p_value)

if p_value<0.05:
  print("Reject Null Hypothesis (H0)")
elif p_value>=0.05:
  print("Reject Alternative Hypothesis(H1)")
```

```
t_statistics: [-0.75691848]
p_value: [0.52811574]
Reject Alternative Hypothesis(H1)
t_statistics: [0.19035288]
p_value: [0.86660314]
Reject Alternative Hypothesis(H1)
```

In [597]:
```
data.to_csv('/Users/jhasija9/Documents/Data Glacier Internship/Week-2/DataSets/master_data.csv',index = False)
```