Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

<G2M insight for Cab Investment firm>

<14/12/2022>

Agenda

Executive Summary

Problem Statement

Approach

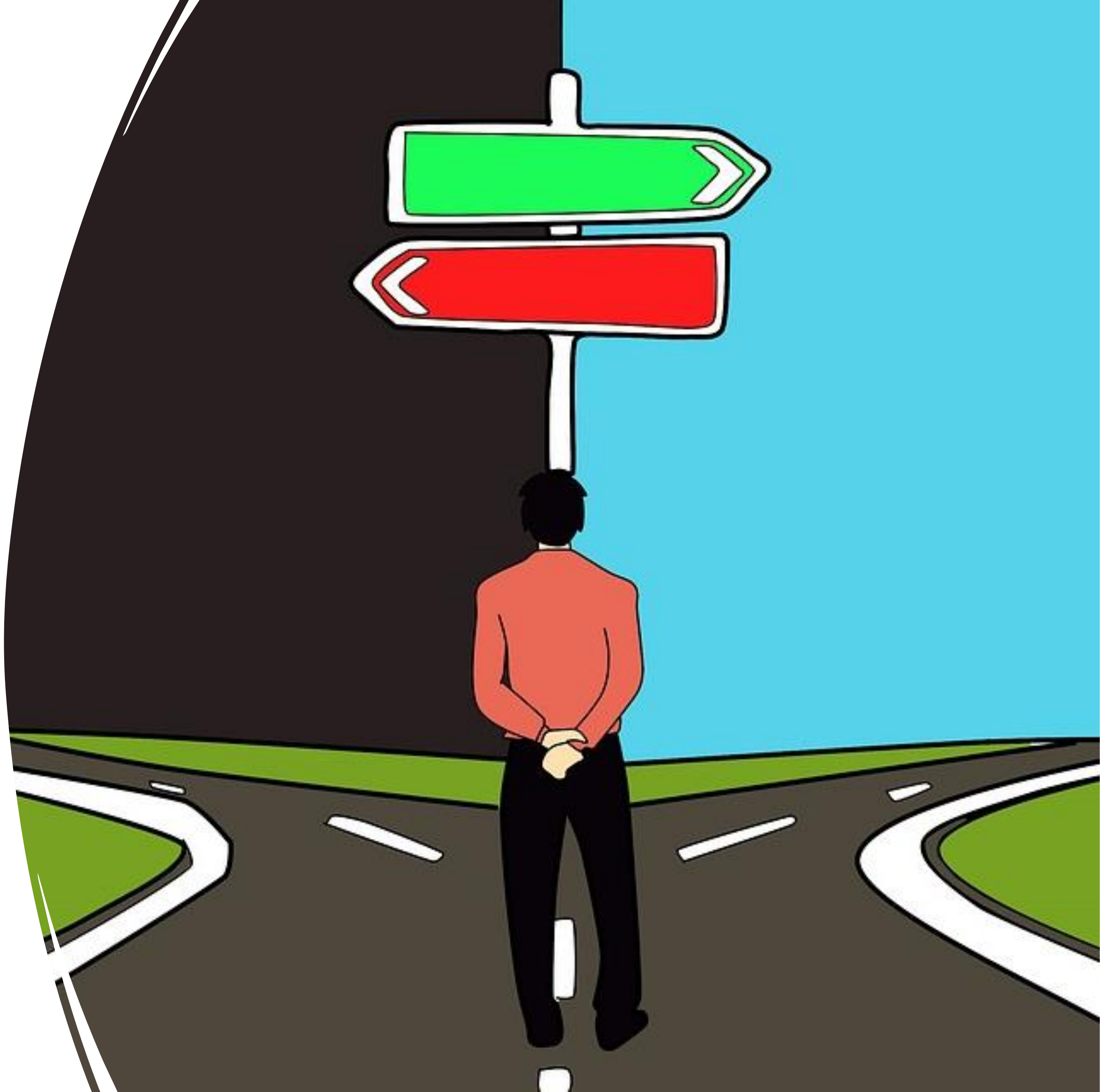
EDA

EDA Summary

Recommendations

Executive Summary

- This project aims to help XYZ’s Executive team analyse and inspect how is the market is going before taking the final decision on identifying the right company to invest in the foreseeable future.



Problem Statement

- 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.

Approach

- I used some Explanatory Data Analysis(EDA) and statistical techniques to analyse and forecast profit for each Pink and Yellow cab companies in the US industry using the 4 provided csv files
- The 4 csv files are combined into one dataframe that contains all relations of all files together. Transaction_ID and Customer_ID csv files are merged using the surrogate key Customer_ID. Cab_Data csv file is merged with merged table above using the Customer_ID. Then, it is merged with the city attribute in the City_csv file.
- I created 3 csv files: ProfitAnalysis, YearlyProfitAnalysis, CityWiseProfitAnalysis to inspect data extracted from original datasets clearly.
- There are total of: 15 columns/features(including 8 derived features),

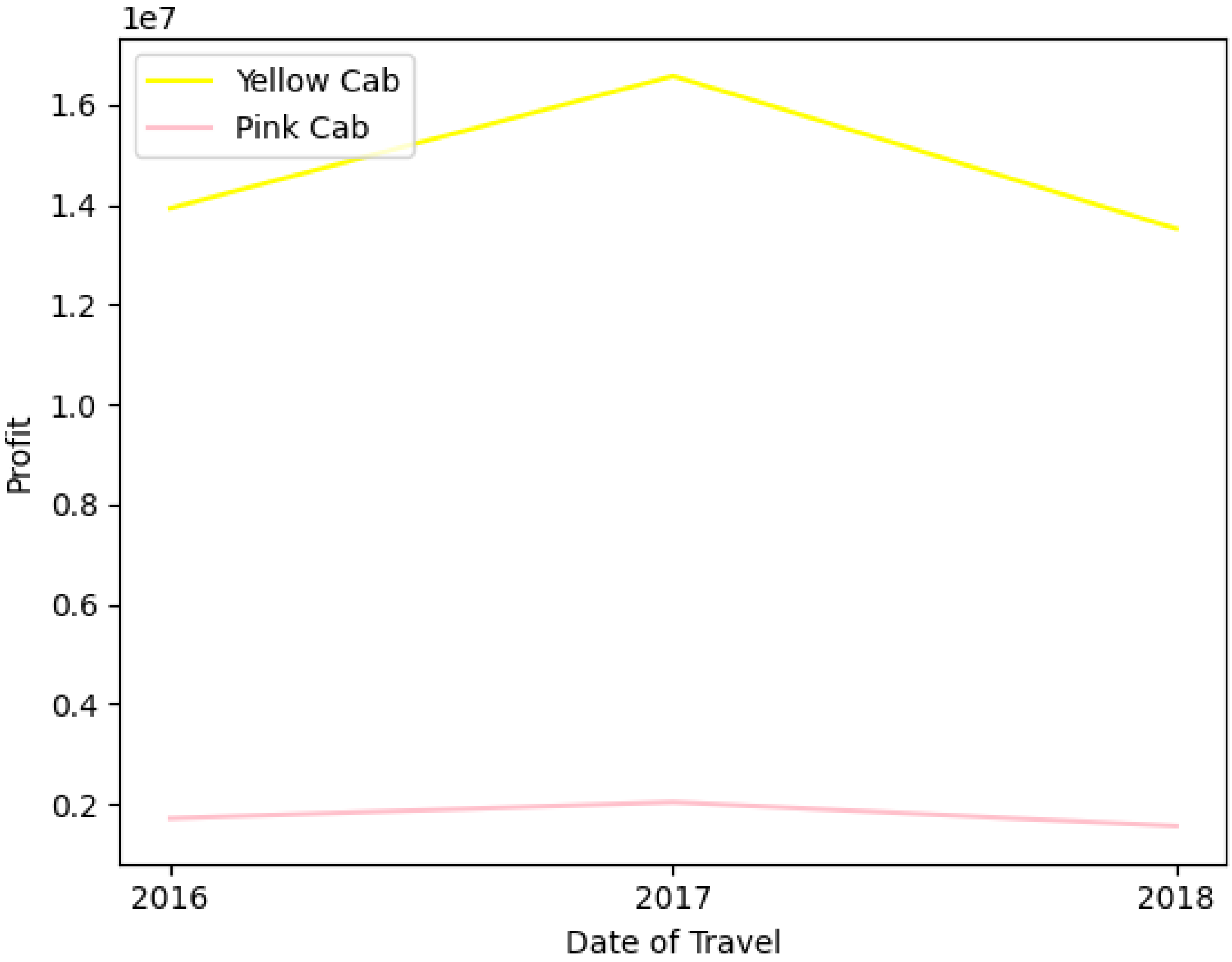
359,392 total data points, Timeframe of the data: 2016-01-31 to 2018-12-31

Assumptions

- Outliers are present in the Profit feature but due to unavailability of trip duration details, we can neglect this as an outlier.
- Outliers are calculated using the Interquartile range formula. Anything below or below upper and lower bounds is considered as an outlier.
- Profit is calculated using Cost of Trip and Price Charged features keeping all other factors constant.
- We assumed that the users feature includes total number of users from all cab companies(including yellow and pink cab) per city.

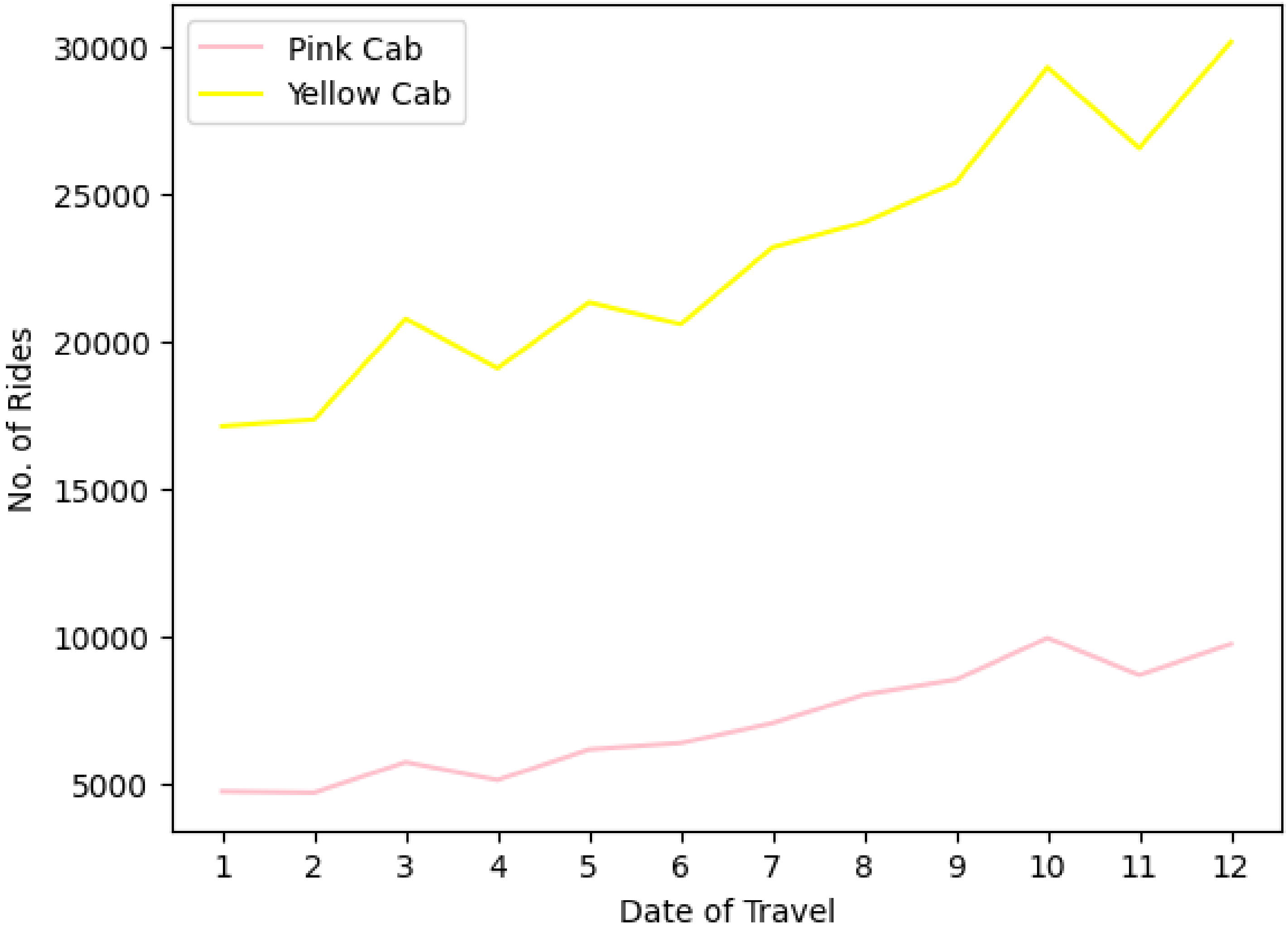
EDA:
1.Yearly Profit Analysis of Yellow and Pink Cab Company

- Both companies reached their peak profit at the year of 2017, however Yellow Cab company has a higher peak and overall profit than Pink Cab in all years.



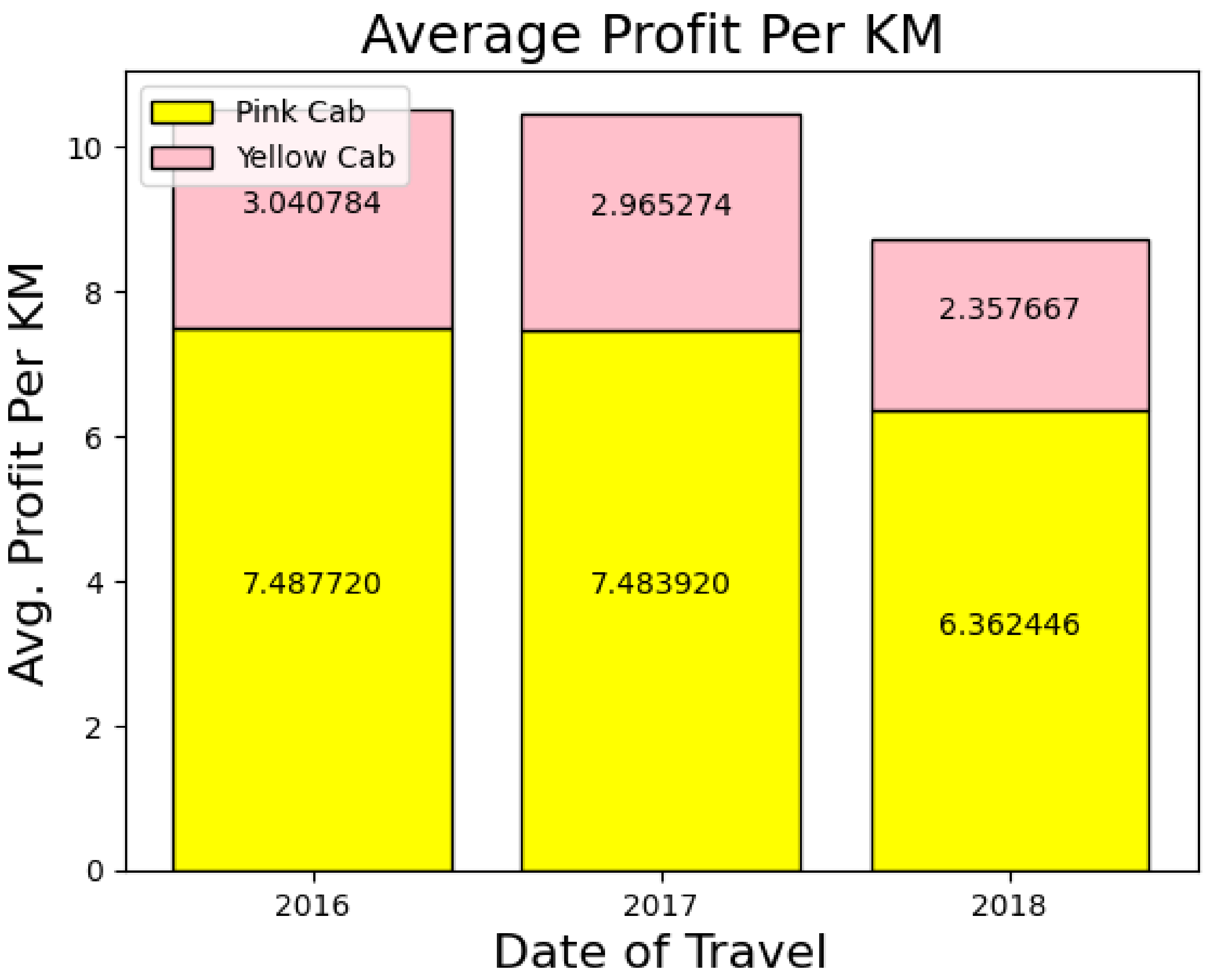
EDA:
2.Monthly No. of Rides Analysis

- As we can see both companies have roughly the same pattern . There is an exponential rise in number of rides conducted by both companies between the months 2-3 and 9-10 in all years between the time data frame. This concludes that there is a high customer demand during these months. It could be due to start of new academic semester and people need more transportation during these specific times. Pink Cab has retained roughly same no. of rides for 4 consecutive months.



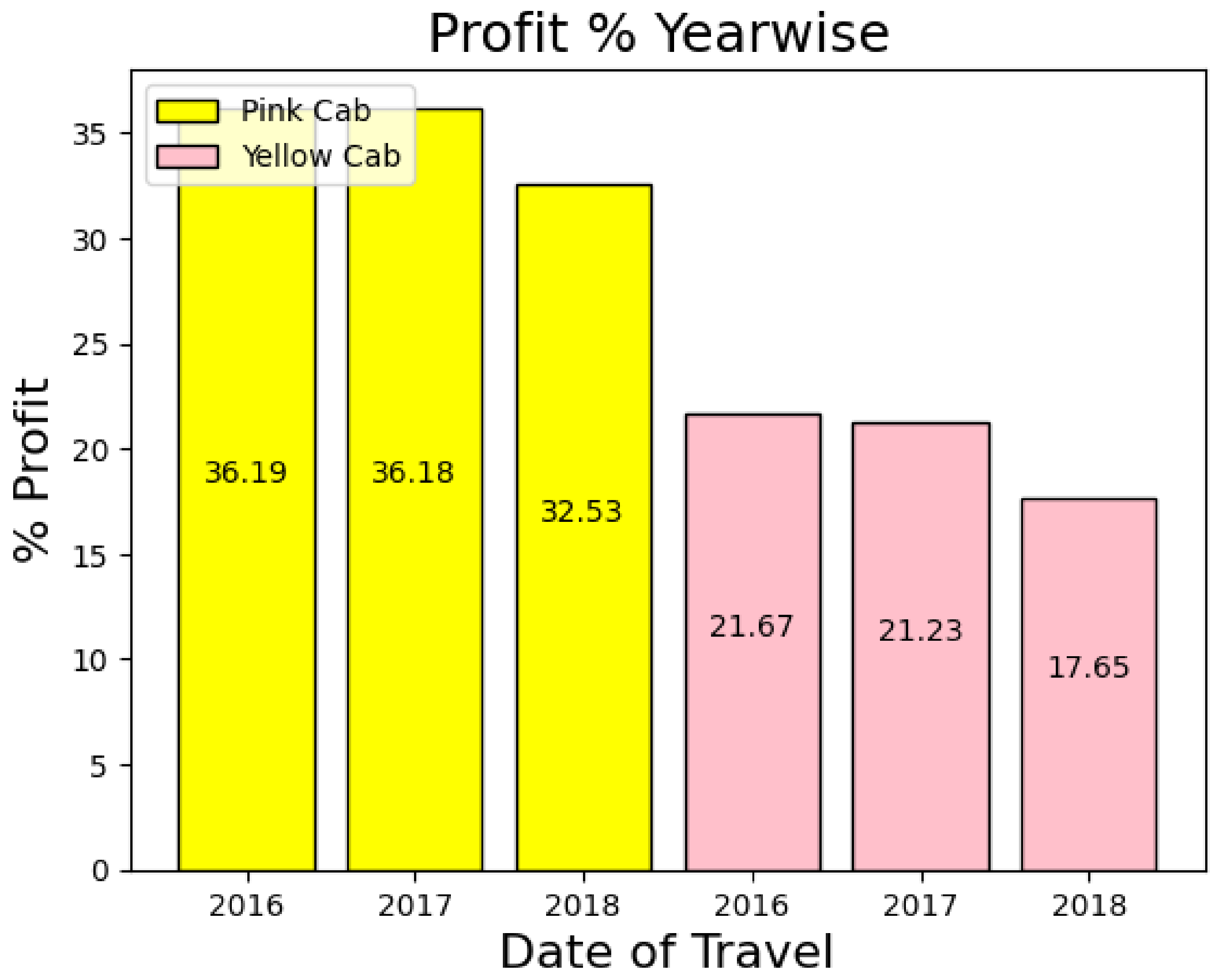
EDA:
3.Average Profit per KM Analysis

- The average Profit per km of Yellow cab is higher than Pink Cab(roughly 2-3 times). Both companies experience a decline in the avg profit per km as years pass by. This could be due to the riding distance increases much faster than the profit attained by each company.



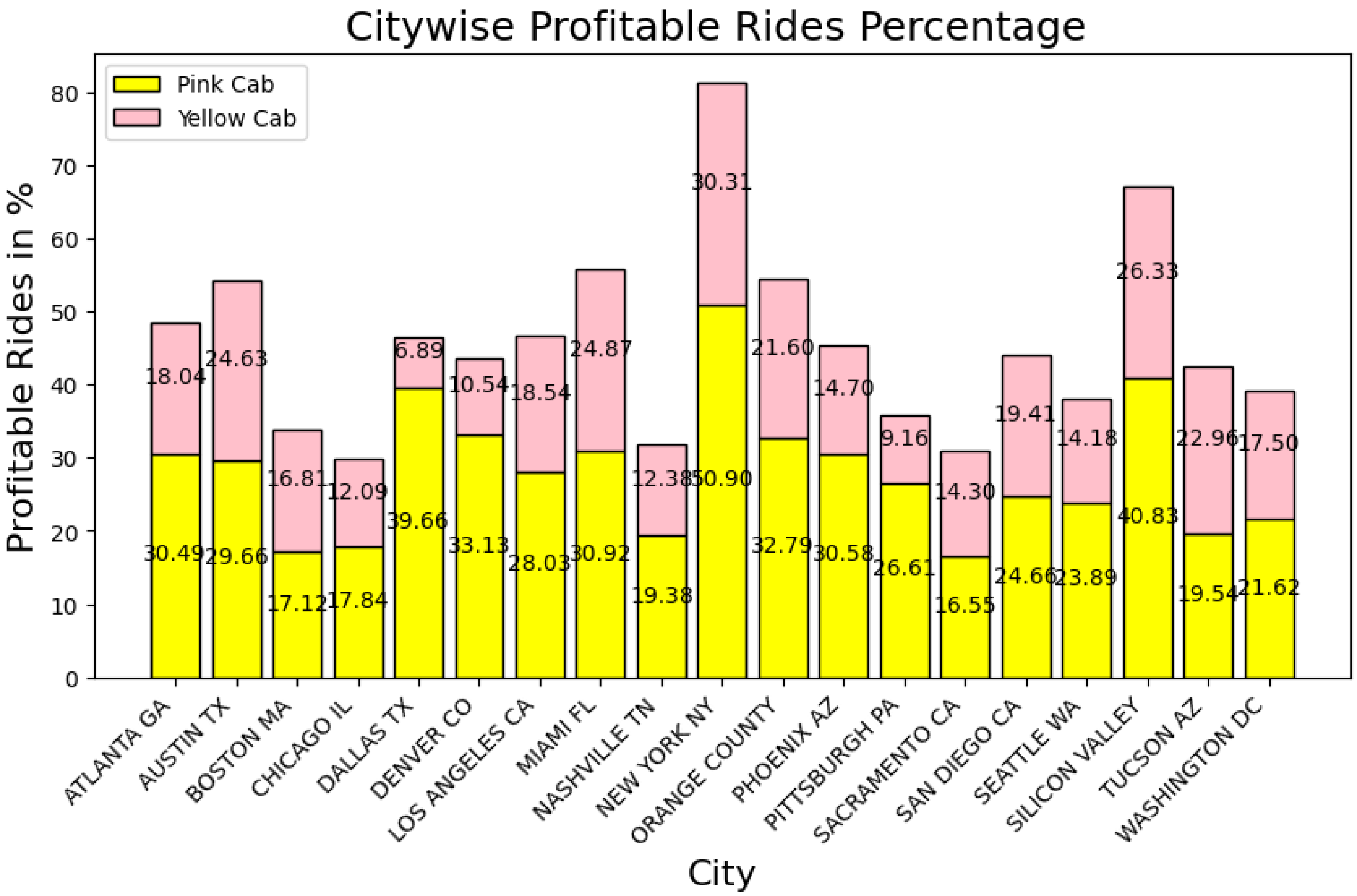
EDA:
4. Profit Percentage Year-wise Analysis

Both companies experience roughly the same decline rate in profit percentage yearly. The Yellow Cab company again is obviouly higher!



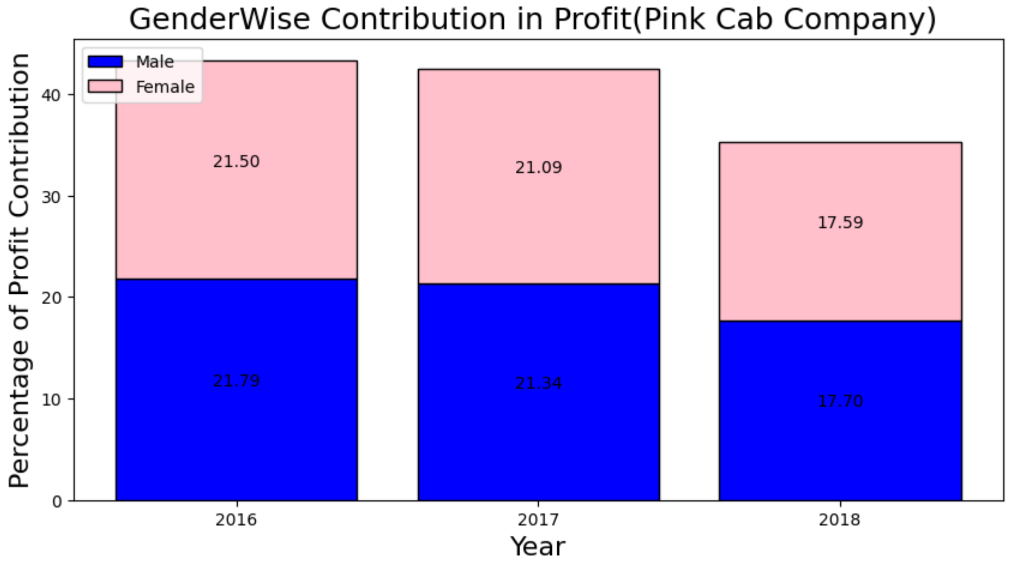
EDA:
5. City-Wise Profitable Rides Percentage Analysis

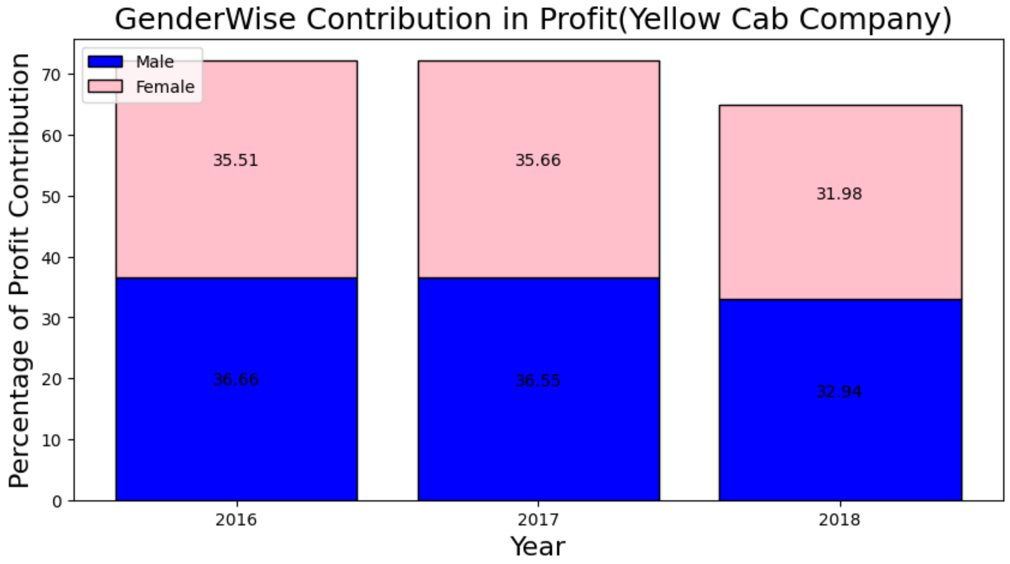
- Pink Cab outperforms Yellow Cab company in only 1 city out of 19 served cities. Pink Cab has roughly same profitable rides as that of its competent in 7 cities(difference is <=5%). This indicates that Pink Cab company is trying to increase its profit in some cities. Yellow Cab has astonishgly exceeded profit % by more than 6 times that of Pink Cab.



EDA:
6. Gender-wise Contribution in Profit

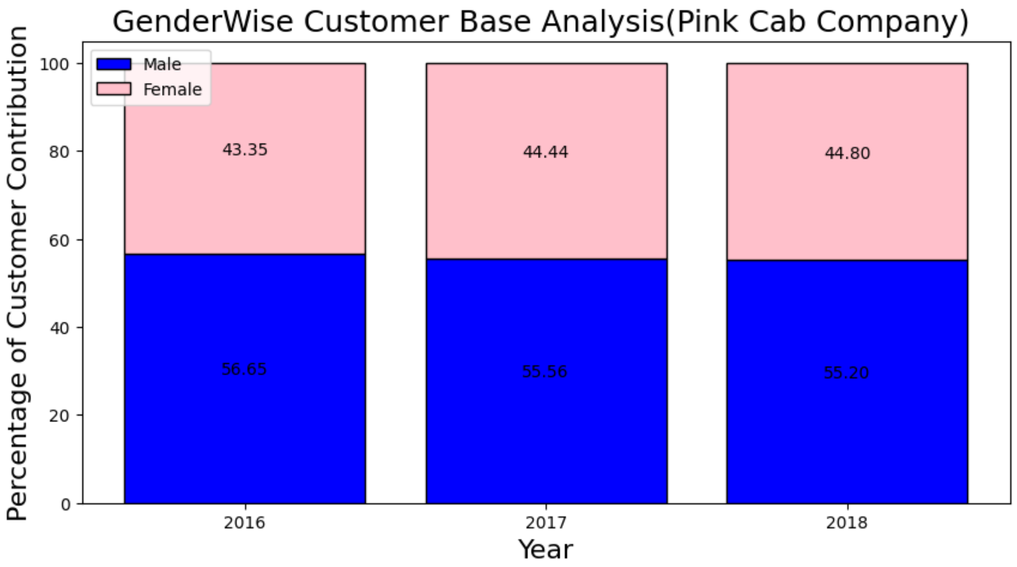
- The Male users has slightly higher contribution in profit in both companies. Overall, there is no significant difference in profit contribution between both genders.

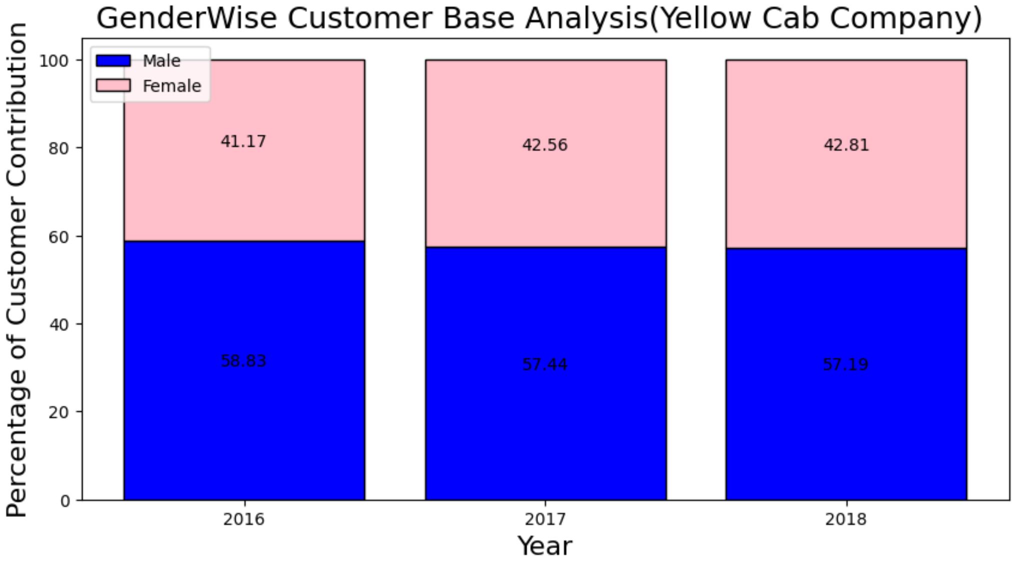




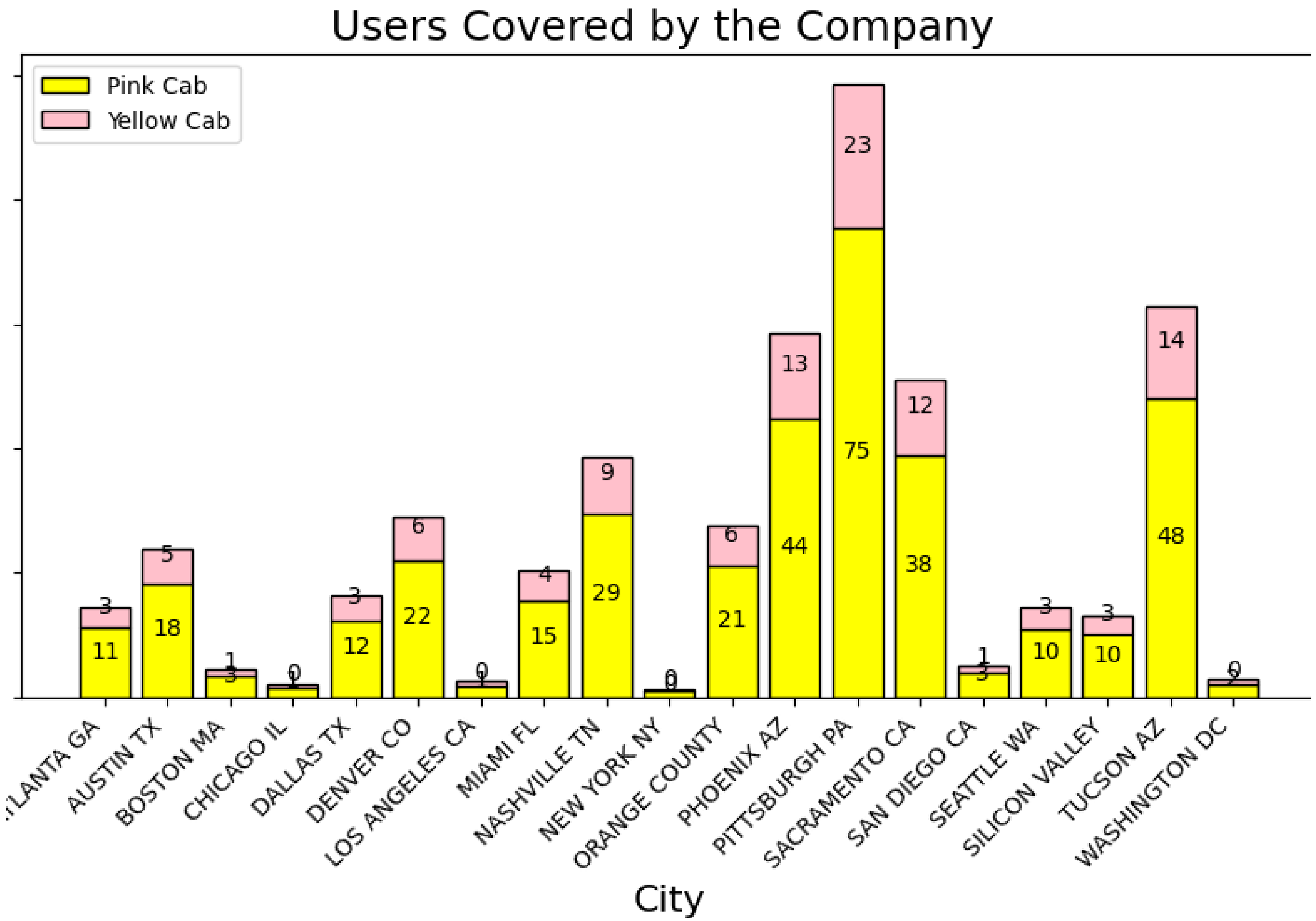
EDA:
7. Gender-wise Customer base Analysis

- This shows the percentage of male and female customers served by both companies. Also, we can see that percentage of male customers are slightly higher than the other gender. Overall, there is no significant differences. This could be reason why male customers contribute to greater portion of profit in both companies.





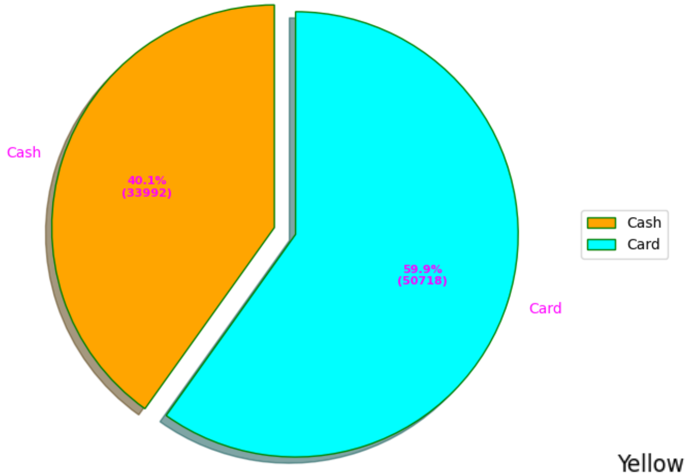
EDA:
8. Users Covered by both company per city

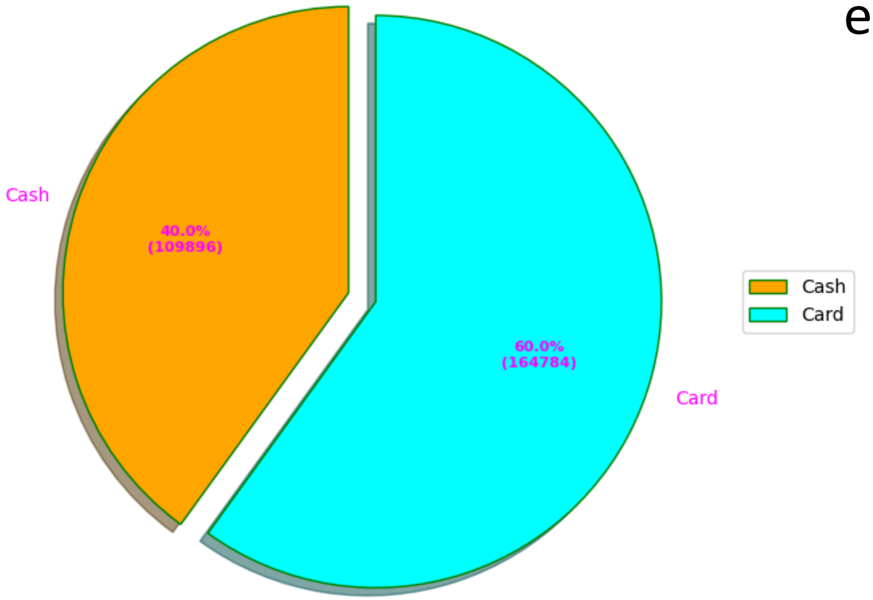


- This shows ratio of users served by both companies in all cities against all other users present in the city.
- We can see that there is a huge difference in customer reach from both companies. Yellow Cab has approximately the same low customer reach as Pink Cab in only 6 cities.

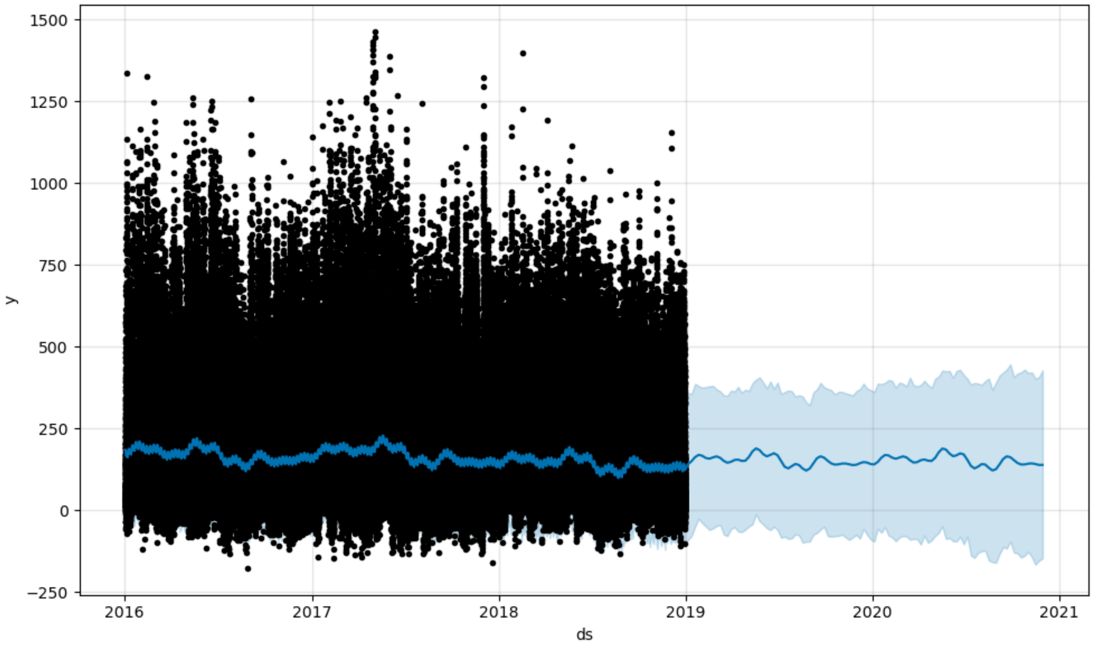
EDA:
9. Payment mode percentage Analysis

- There are greater number of users paying by card in both companies as compared to other payment methods. There is no significant difference between card and cash users in both companies this indicates that payment mode has no effect on company’s profit.

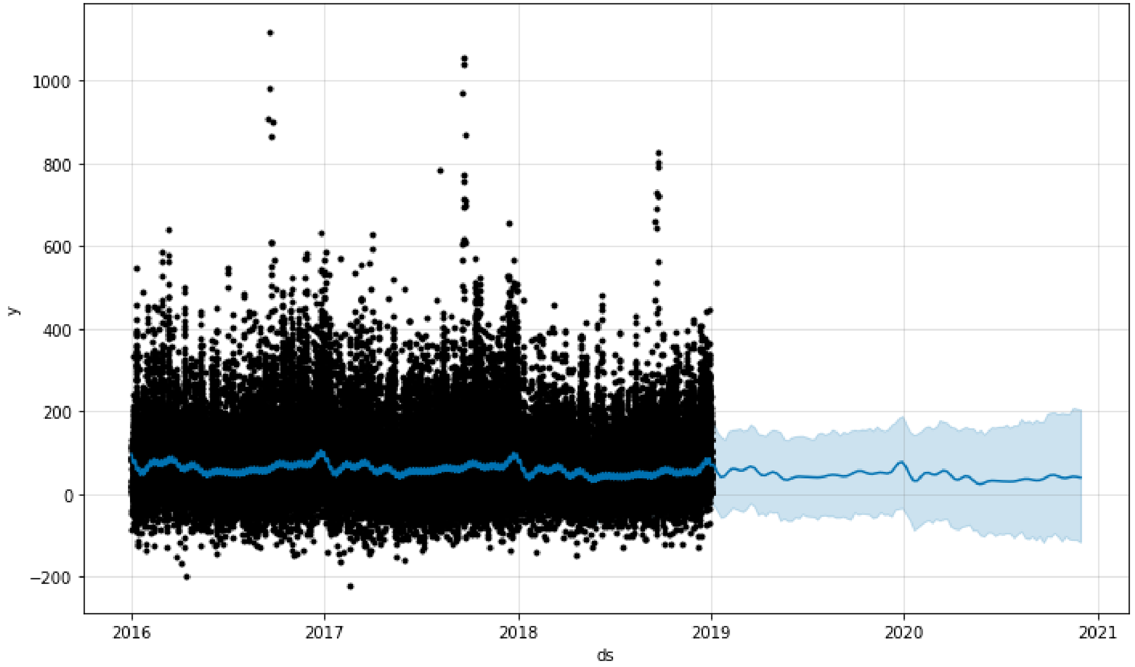

Pink Cab Commpany Payment Mode Percentage


Yellow Cab Commpany Payment Mode Percentage

EDA:
10. Profit Forecasting

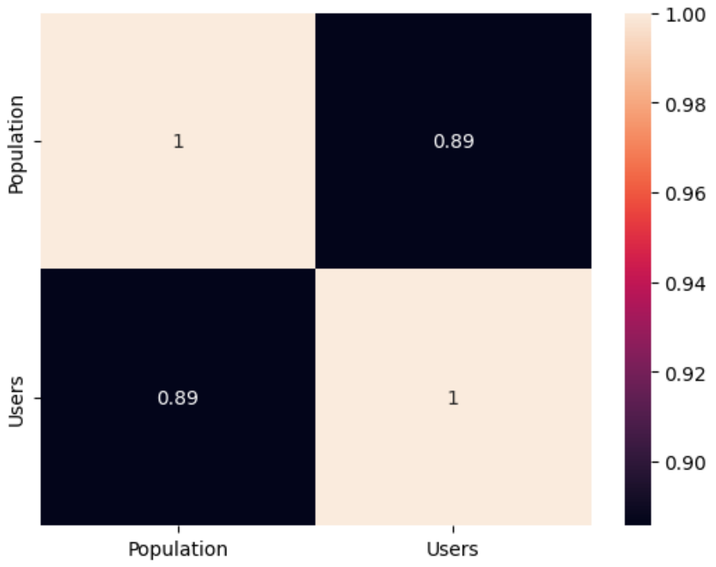


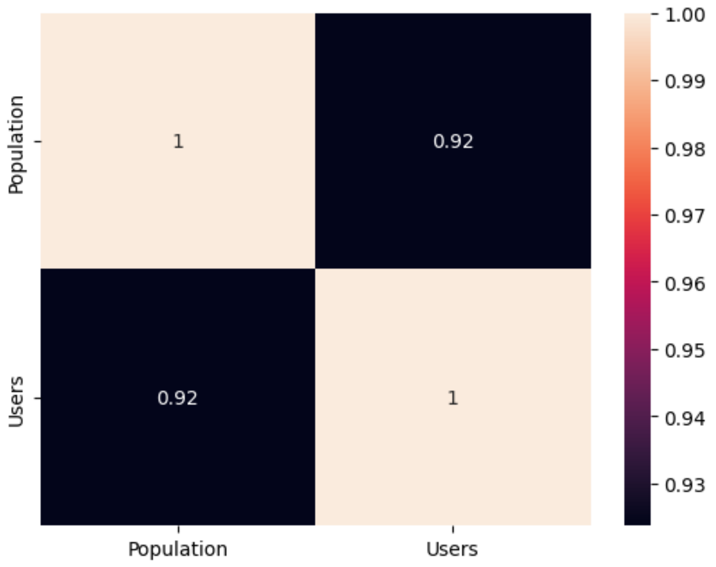
Yellow Cab



Pink Cab

Heat Map Correlation between users and Population

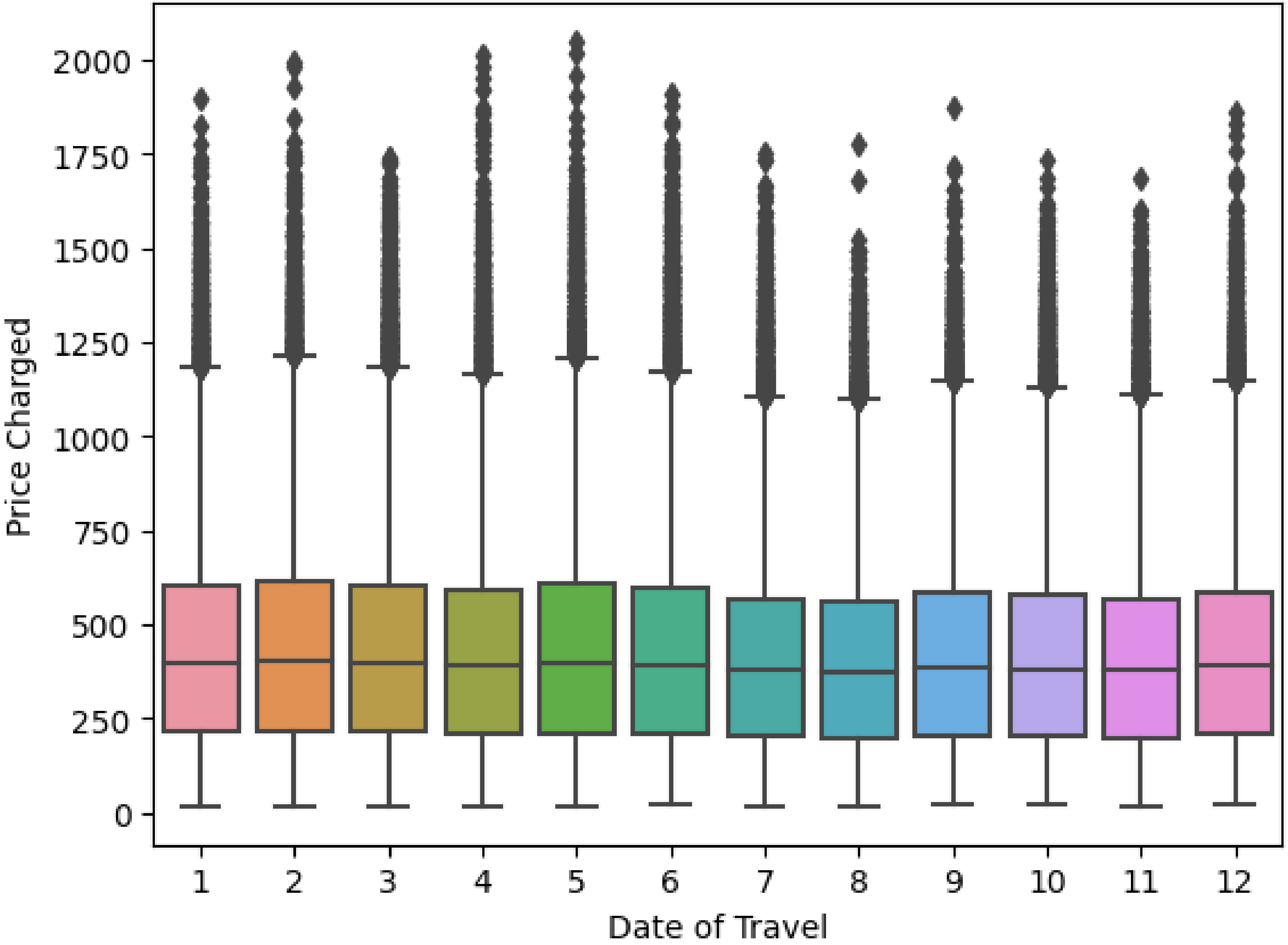




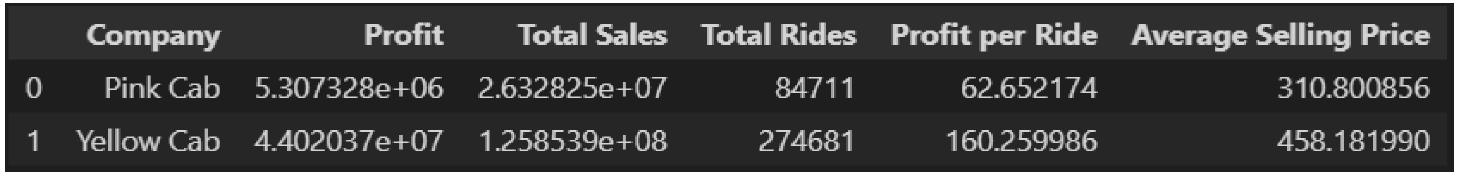
- Pink Cab

Yellow Cab

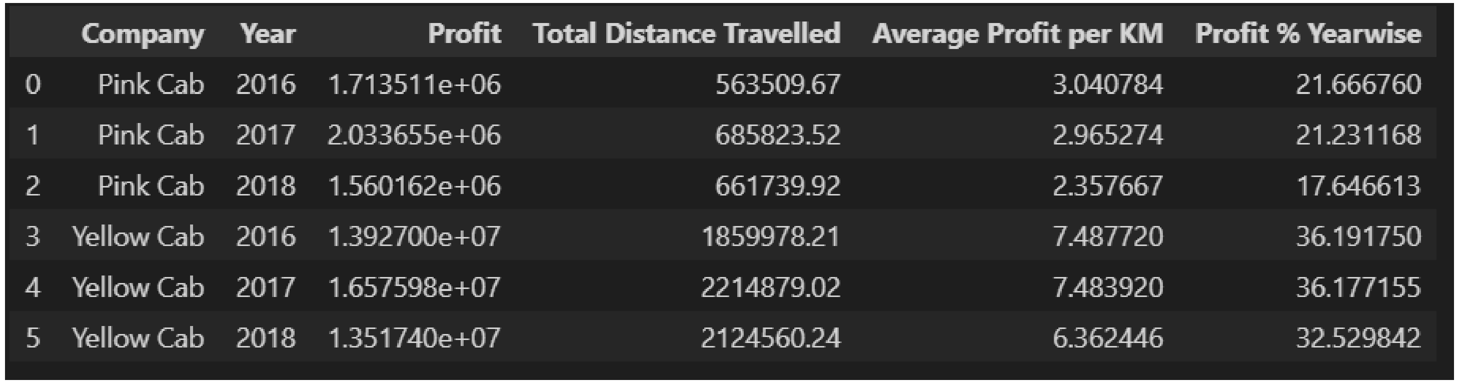
Outliers Box plot



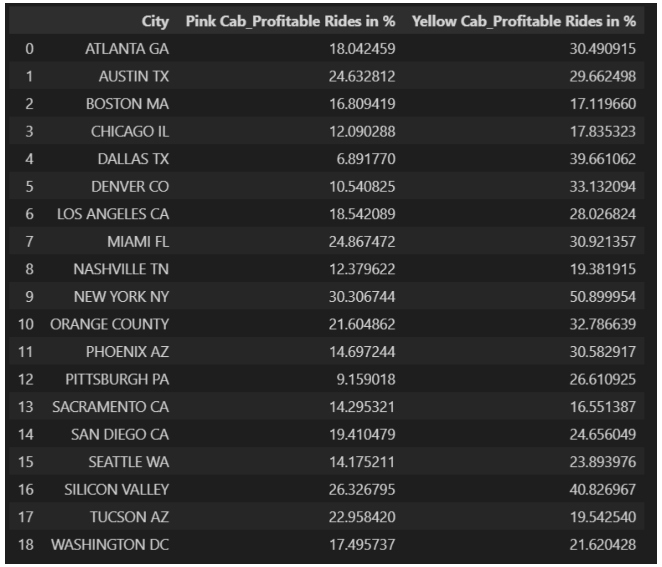
Created CSV Files



	Company	Profit	Total Sales	Total Rides	Profit per Ride	Average Selling Price
0	Pink Cab	5.307328e+06	2.632825e+07	84711	62.652174	310.800856
1	Yellow Cab	4.402037e+07	1.258539e+08	274681	160.259986	458.181990



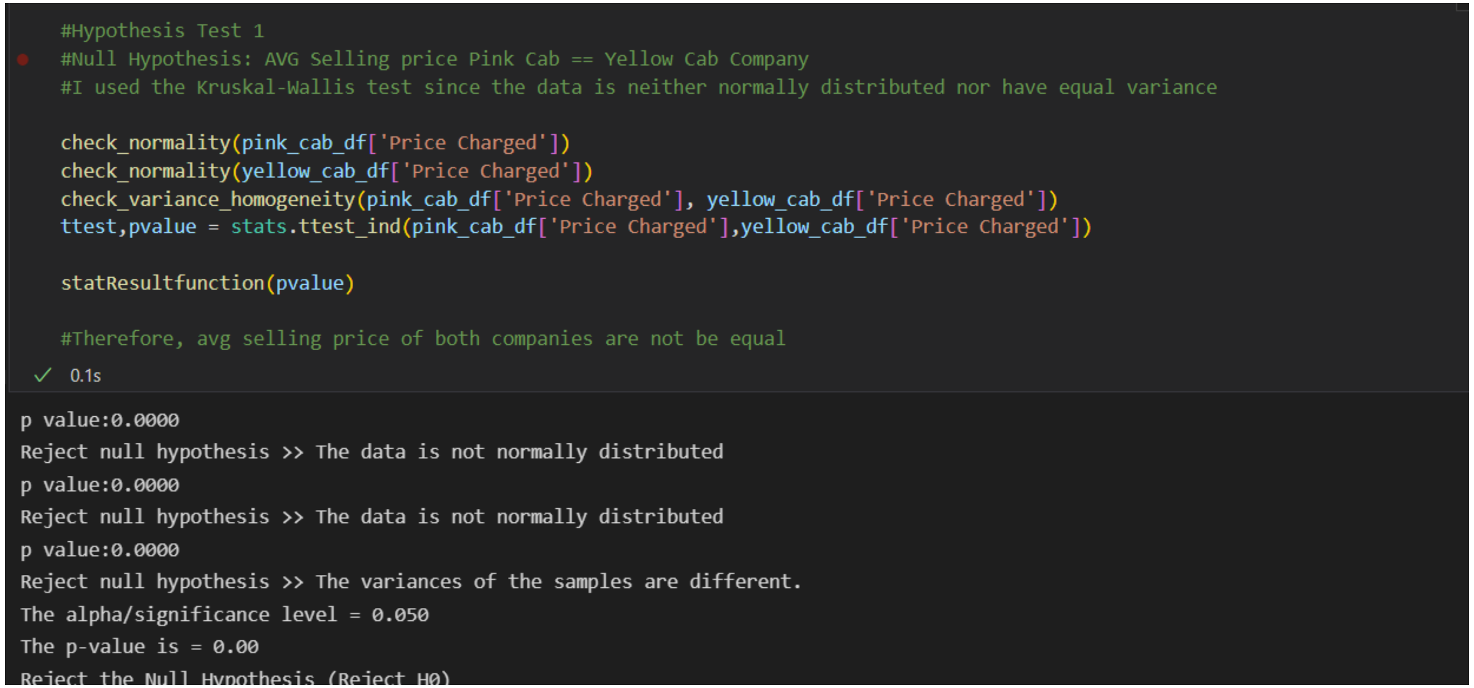
	Company	Year	Profit	Total Distance Travelled	Average Profit per KM	Profit % Yearwise
0	Pink Cab	2016	1.713511e+06	563509.67	3.040784	21.666760
1	Pink Cab	2017	2.033655e+06	685823.52	2.965274	21.231168
2	Pink Cab	2018	1.560162e+06	661739.92	2.357667	17.646613
3	Yellow Cab	2016	1.392700e+07	1859978.21	7.487720	36.191750
4	Yellow Cab	2017	1.657598e+07	2214879.02	7.483920	36.177155
5	Yellow Cab	2018	1.351740e+07	2124560.24	6.362446	32.529842



	City	Pink Cab_Profitable Rides in %	Yellow Cab_Profitable Rides in %
0	ATLANTA GA	18.042459	30.490915
1	AUSTIN TX	24.632812	29.662498
2	BOSTON MA	16.809419	17.119660
3	CHICAGO IL	12.090288	17.835323
4	DALLAS TX	6.891770	39.661062
5	DENVER CO	10.540825	33.132094
6	LOS ANGELES CA	18.542089	28.026824
7	MIAMI FL	24.867472	30.921357
8	NASHVILLE TN	12.379622	19.381915
9	NEW YORK NY	30.306744	50.899954
10	ORANGE COUNTY	21.604862	32.786639
11	PHOENIX AZ	14.697244	30.582917
12	PITTSBURGH PA	9.159018	26.610925
13	SACRAMENTO CA	14.295321	16.551387
14	SAN DIEGO CA	19.410479	24.656049
15	SEATTLE WA	14.175211	23.893976
16	SILICON VALLEY	26.326795	40.826967
17	TUCSON AZ	22.958420	19.542540
18	WASHINGTON DC	17.495737	21.620428

Hypothesis Tests

- **Test1**, Is there difference between Avg. Selling Price of both cabs?

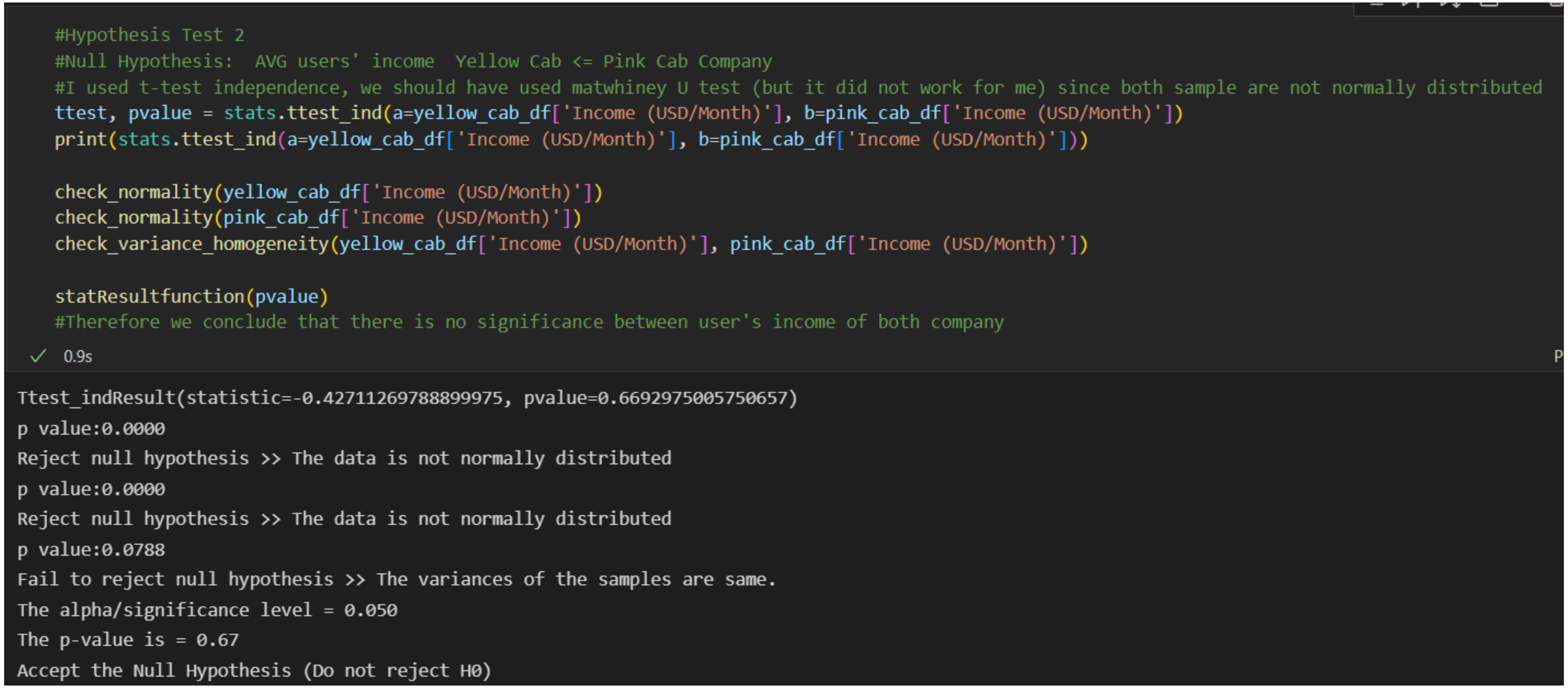


```
#Hypothesis Test 1
#Null Hypothesis: AVG Selling price Pink Cab == Yellow Cab Company
#I used the Kruskal-Wallis test since the data is neither normally distributed nor have equal variance

check_normality(pink_cab_df['Price Charged'])
check_normality(yellow_cab_df['Price Charged'])
check_variance_homogeneity(pink_cab_df['Price Charged'], yellow_cab_df['Price Charged'])
ttest,pvalue = stats.ttest_ind(pink_cab_df['Price Charged'],yellow_cab_df['Price Charged'])

statResultfunction(pvalue)

#Therefore, avg selling price of both companies are not be equal
```

✓ 0.1s

```
p value:0.0000
Reject null hypothesis >> The data is not normally distributed
p value:0.0000
Reject null hypothesis >> The data is not normally distributed
p value:0.0000
Reject null hypothesis >> The variances of the samples are different.
The alpha/significance level = 0.050
The p-value is = 0.00
Reject the Null Hypothesis (Reject H0)
```

Test2, Is there difference between Avg. user income of both cabs?

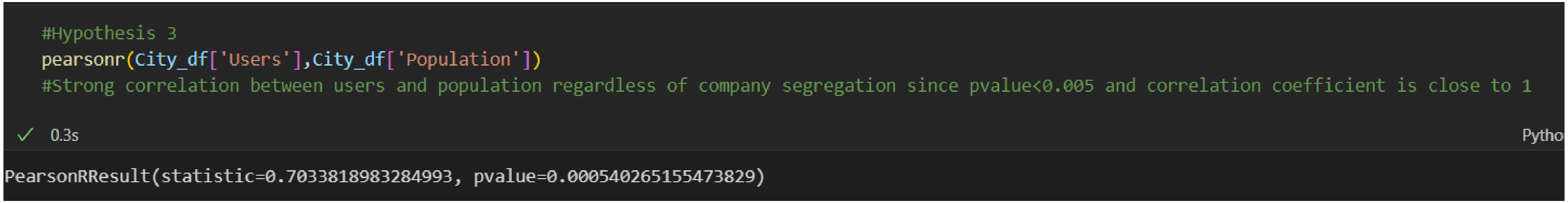


```
#Hypothesis Test 2
#Null Hypothesis:  AVG users' income  Yellow Cab <= Pink Cab Company
#I used t-test independence, we should have used matwhiney U test (but it did not work for me) since both sample are not normally distributed
ttest, pvalue = stats.ttest_ind(a=yellow_cab_df['Income (USD/Month)'], b=pink_cab_df['Income (USD/Month)'])
print(stats.ttest_ind(a=yellow_cab_df['Income (USD/Month)'], b=pink_cab_df['Income (USD/Month)']))

check_normality(yellow_cab_df['Income (USD/Month)'])
check_normality(pink_cab_df['Income (USD/Month)'])
check_variance_homogeneity(yellow_cab_df['Income (USD/Month)'], pink_cab_df['Income (USD/Month)'])

statResultfunction(pvalue)
#Therefore we conclude that there is no significance between user's income of both company
```

✓ 0.9s

```
Ttest_indResult(statistic=-0.42711269788899975, pvalue=0.6692975005750657)
p value:0.0000
Reject null hypothesis >> The data is not normally distributed
p value:0.0000
Reject null hypothesis >> The data is not normally distributed
p value:0.0788
Fail to reject null hypothesis >> The variances of the samples are same.
The alpha/significance level = 0.050
The p-value is = 0.67
Accept the Null Hypothesis (Do not reject H0)
```

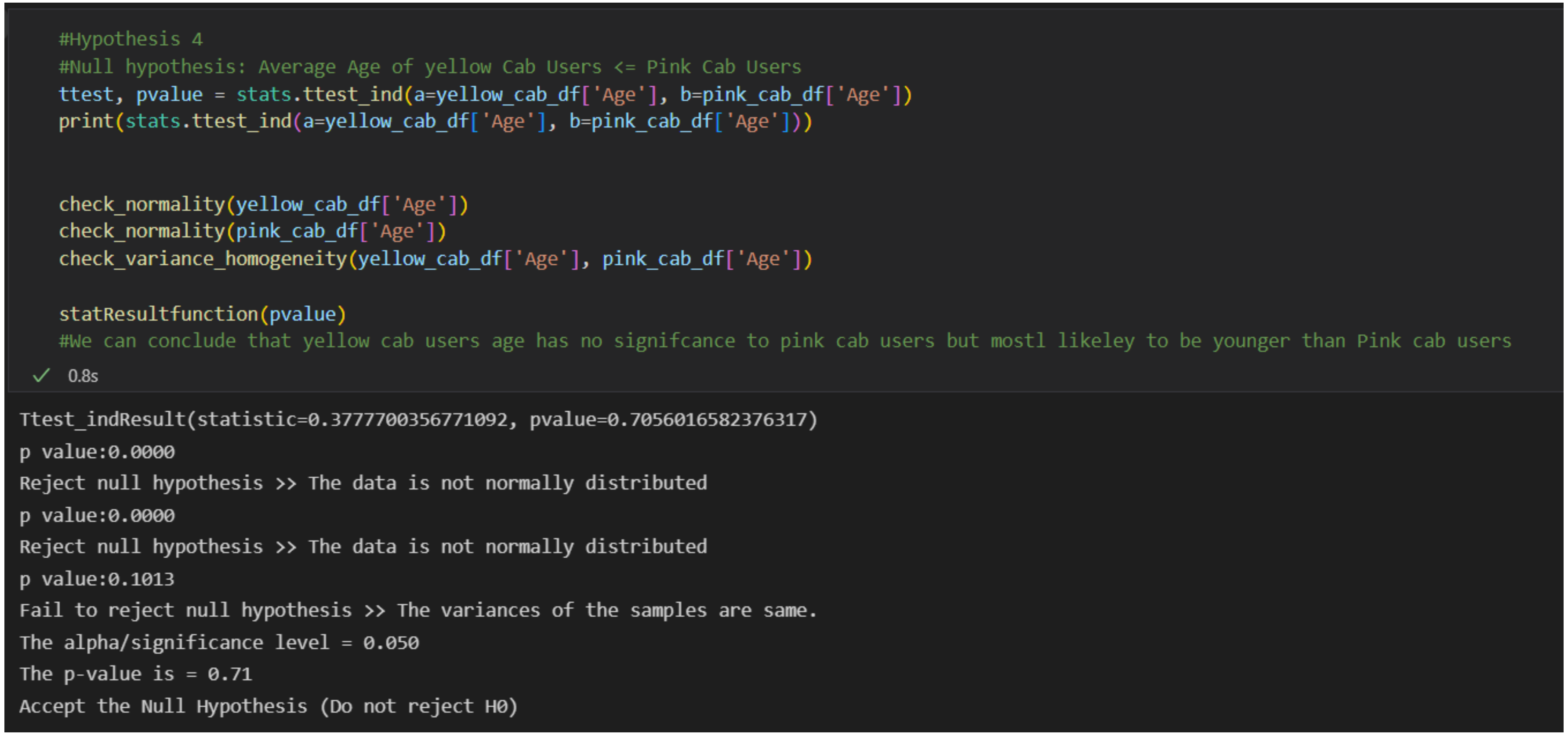
Test3, Is there any relation between users and population per city?



```
#Hypothesis 3
pearsonr(City_df['Users'],City_df['Population'])
#Strong correlation between users and population regardless of company segregation since pvalue<0.005 and correlation coefficient is close to 1
```

```
PearsonRResult(statistic=0.7033818983284993, pvalue=0.000540265155473829)
```

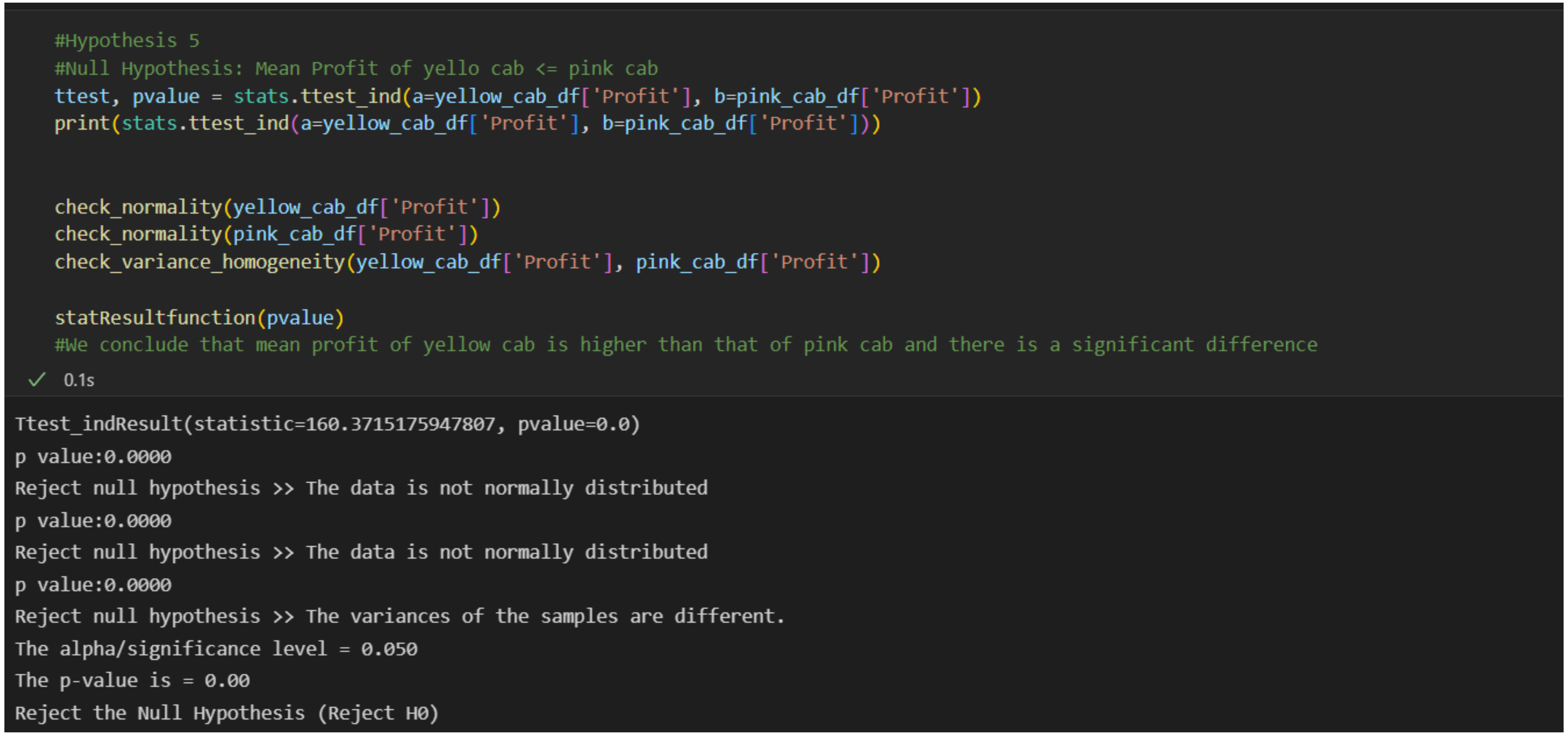
Test4, Is there difference between Avg. user age of both cabs?



```
#Hypothesis 4
#Null hypothesis: Average Age of yellow Cab Users <= Pink Cab Users
ttest, pvalue = stats.ttest_ind(a=yellow_cab_df['Age'], b=pink_cab_df['Age'])
print(stats.ttest_ind(a=yellow_cab_df['Age'], b=pink_cab_df['Age']))


check_normality(yellow_cab_df['Age'])
check_normality(pink_cab_df['Age'])
check_variance_homogeneity(yellow_cab_df['Age'], pink_cab_df['Age'])

statResultfunction(pvalue)
#We can conclude that yellow cab users age has no signifcance to pink cab users but mostl likeley to be younger than Pink cab users
```

```
Ttest_indResult(statistic=0.3777700356771092, pvalue=0.7056016582376317)
p value:0.0000
Reject null hypothesis >> The data is not normally distributed
p value:0.0000
Reject null hypothesis >> The data is not normally distributed
p value:0.1013
Fail to reject null hypothesis >> The variances of the samples are same.
The alpha/significance level = 0.050
The p-value is = 0.71
Accept the Null Hypothesis (Do not reject H0)
```

Test5, Is there difference between the Mean Profit of both cabs?



```
#Hypothesis 5
#Null Hypothesis: Mean Profit of yello cab <= pink cab
ttest, pvalue = stats.ttest_ind(a=yellow_cab_df['Profit'], b=pink_cab_df['Profit'])
print(stats.ttest_ind(a=yellow_cab_df['Profit'], b=pink_cab_df['Profit']))


check_normality(yellow_cab_df['Profit'])
check_normality(pink_cab_df['Profit'])
check_variance_homogeneity(yellow_cab_df['Profit'], pink_cab_df['Profit'])

statResultfunction(pvalue)
#We conclude that mean profit of yellow cab is higher than that of pink cab and there is a significant difference
```

```
Ttest_indResult(statistic=160.3715175947807, pvalue=0.0)
p value:0.0000
Reject null hypothesis >> The data is not normally distributed
p value:0.0000
Reject null hypothesis >> The data is not normally distributed
p value:0.0000
Reject null hypothesis >> The variances of the samples are different.
The alpha/significance level = 0.050
The p-value is = 0.00
Reject the Null Hypothesis (Reject H0)
```

Yes, the answer is visualised in slide no.8

Is there any seasonality in number of customers using the cab service?

Recommendations

- Customer Reach: Yellow Cab company has a high customer reach in 13 cities as compared to Pink Cab
- Average Profit per KM: It can be concluded that Yellow Cab has outperformed Pink Cab by approx. 3 times average Profit per km after the time range.
- Ride count and Profit Forecasting: Both companies are facing loss in profit; however, after the year 2019 yellow cab experiences profit rise unlike the other cab . Pink Cab experienced a constant un-change per year in ride count for a long period of time as compared to other cab.
- City-wise profit analysis: Yellow Cab has higher profit percentage that is covered from larger portion of cities in the US as compared to other cab.
- Average Selling Price: Yellow Cab has higher average selling price as compared to Pink Cab due to its widely availability.

Therefore, based on the above points we can conclude that Yellow Cab company is the best recommended option to invest in for the foreseeable future.

Thank You

Data Glacier

Your Deep Learning Partner