



Data Glacier

Your Deep Learning Partner

G2M Case Study

Virtual Internship

12-July-2022

Executive Summary

The main objective of this case study is to help XYZ company identify the right company to make its investment. In order to realize the case study's objective, a total of 359392 observations from 31/01/2016 to 31/12/2018 are examined. Demand, Seasonality, cost and profit analyses are performed using python libraries such as , pandas, numpy, matplotlib, and seaborn. In addition to that, Linear, Lasso, Ridge and Elastic Net models are developed to predict the price charged by cab companies. The findings imply that there is a high demand for Yellow cab company and its more profitable than its competitor pink cab company.

Background of The 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. as per their Go-to-Market(G2M) strategy they want to understand the market before taking final decision
- 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 exploration
- Cost and demand analysis
- Performing Profitability analysis and determined which company is more profitable
- Recommendations for investment

Data Exploration

Dataset:

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

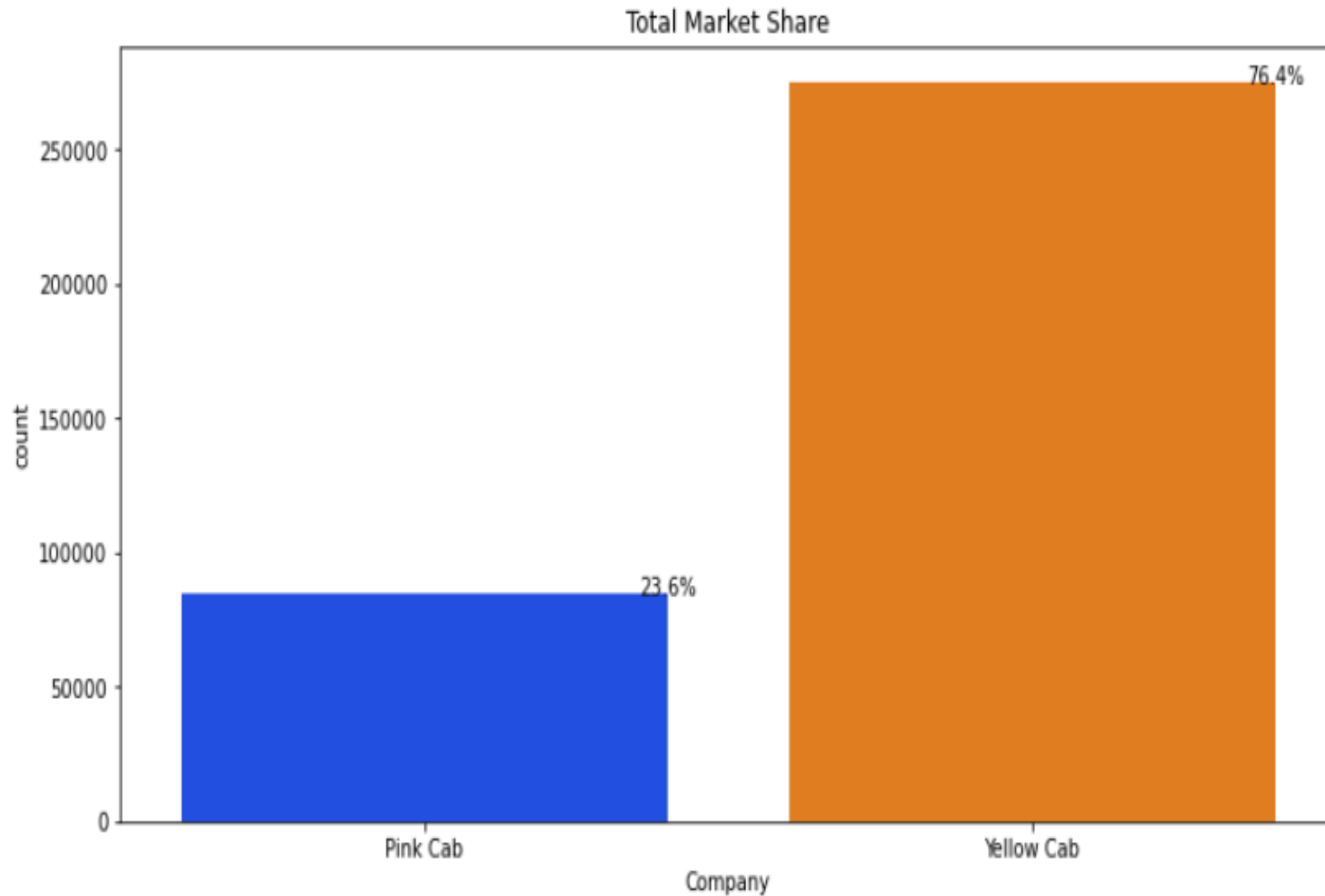
Customer_ID.csv – this is a mapping table that contains a unique identifier which links the customer's demographic details

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

- **Time period** : from 31/01/2016 to 31/12/2018
- **Features**: 20 features
- A total of 359392 examples

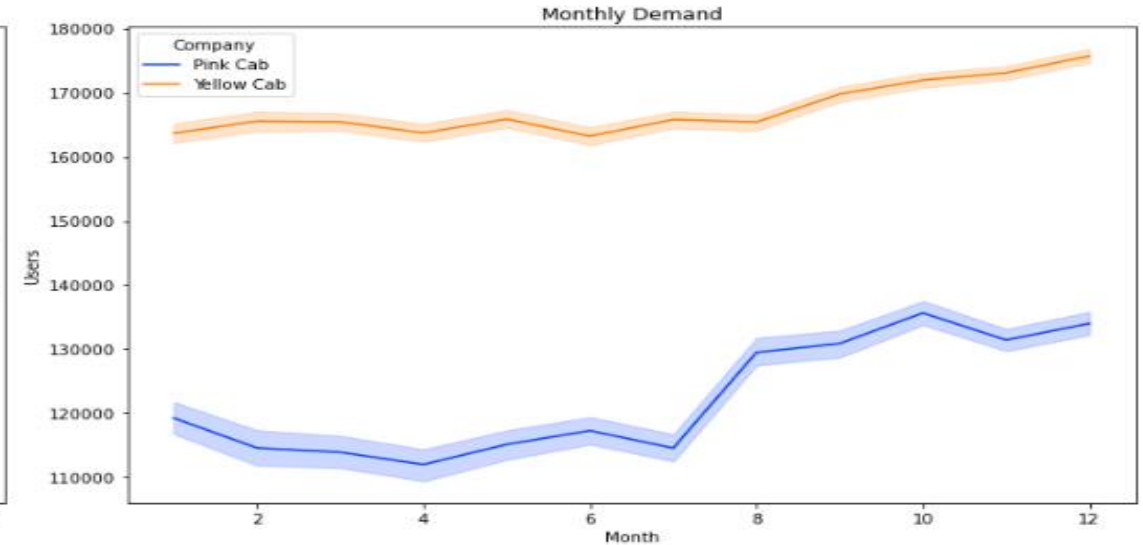
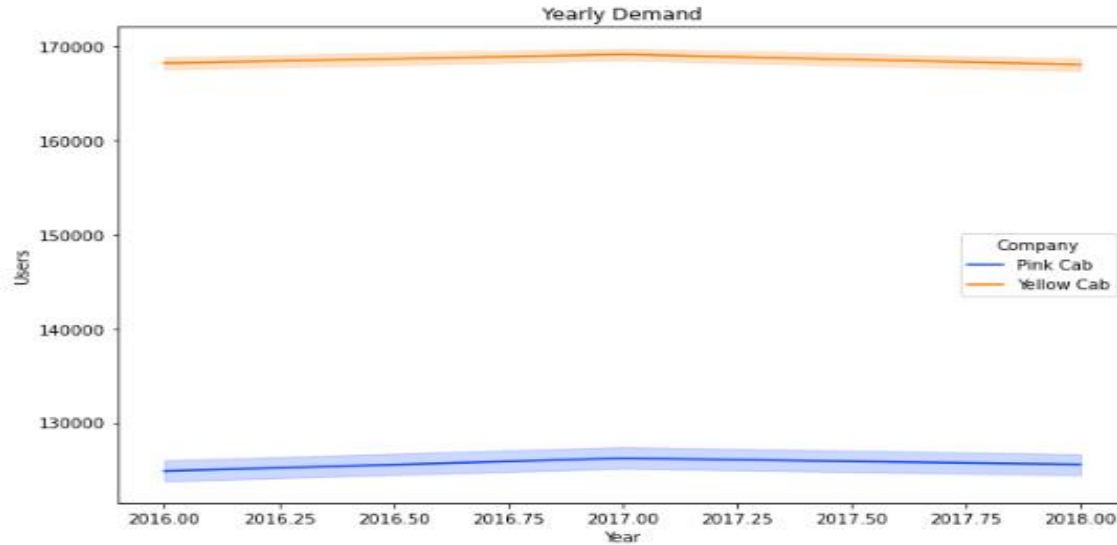
Market Share



		Users
Year	Company	
2016	Pink Cab	3157129931
	Yellow Cab	13949013152
2017	Pink Cab	3812654845
	Yellow Cab	16533576895
2018	Pink Cab	3669138612
	Yellow Cab	15793809907

- Yellow company has the largest market share (76.4%) which is 3 times higher than Pink company (23.6%)

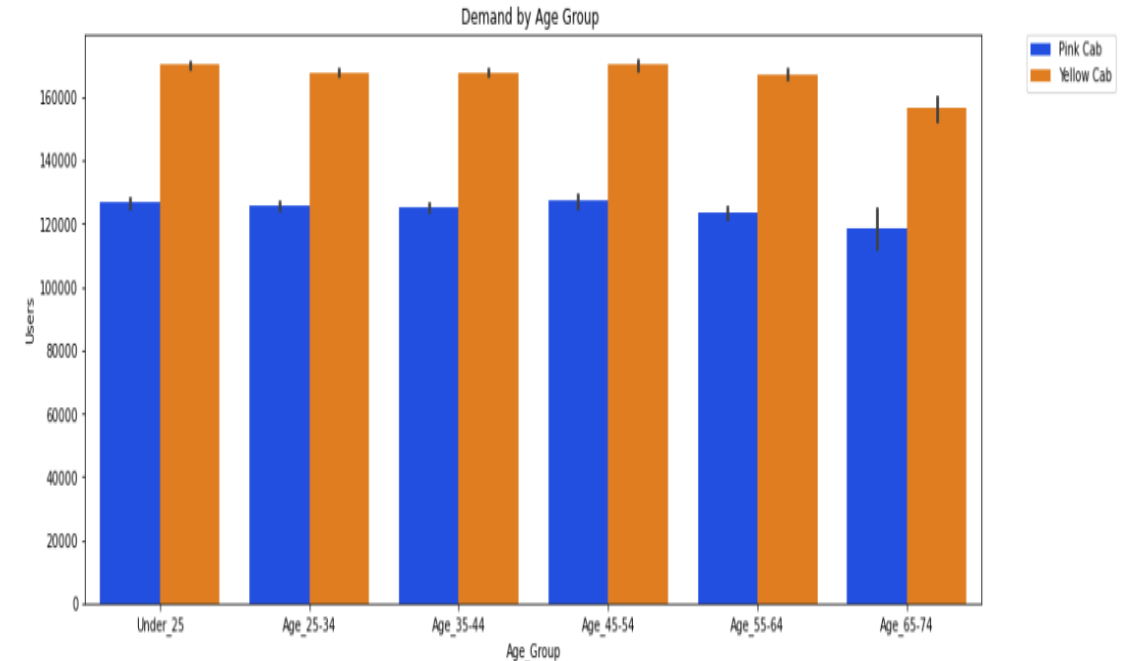
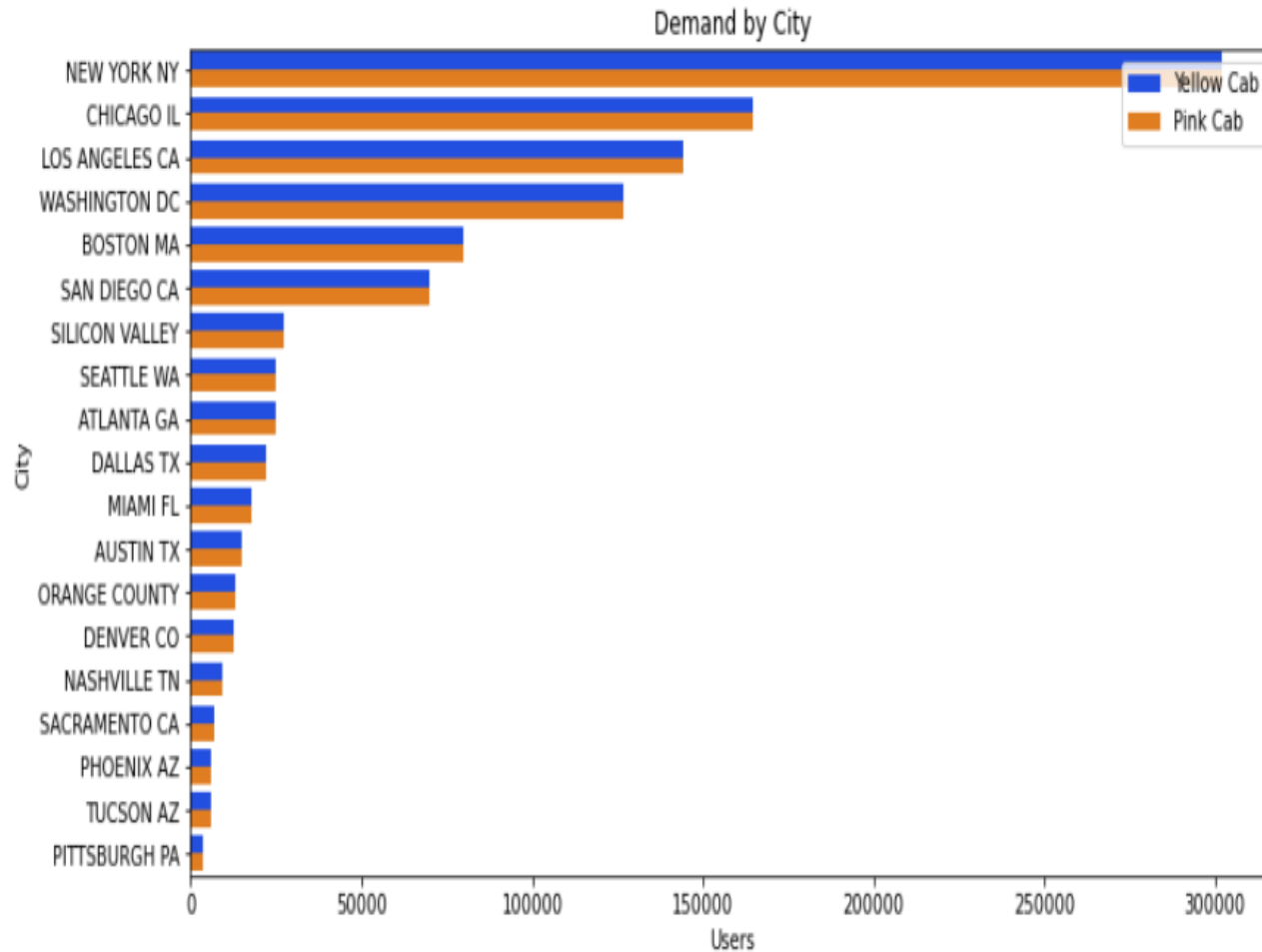
Demand and Seasonality Analysis



Users		
Seasons	Company	
Winter	Pink Cab	2478091038
	Yellow Cab	11444457368
Summer	Pink Cab	1523566709
	Yellow Cab	8888578758
Spring	Pink Cab	2480625190
	Yellow Cab	10563147151
Fall	Pink Cab	4156640451
	Yellow Cab	15380216677

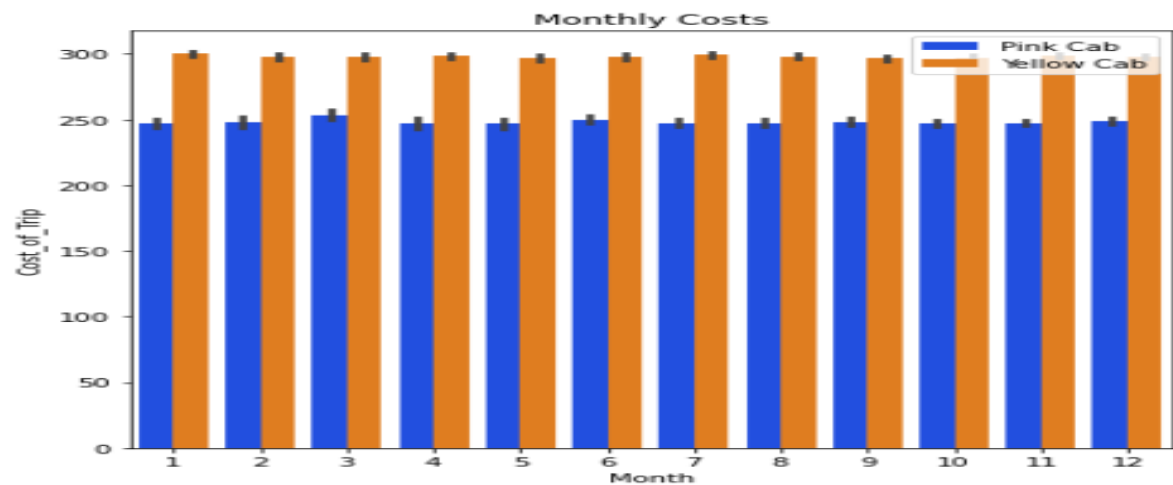
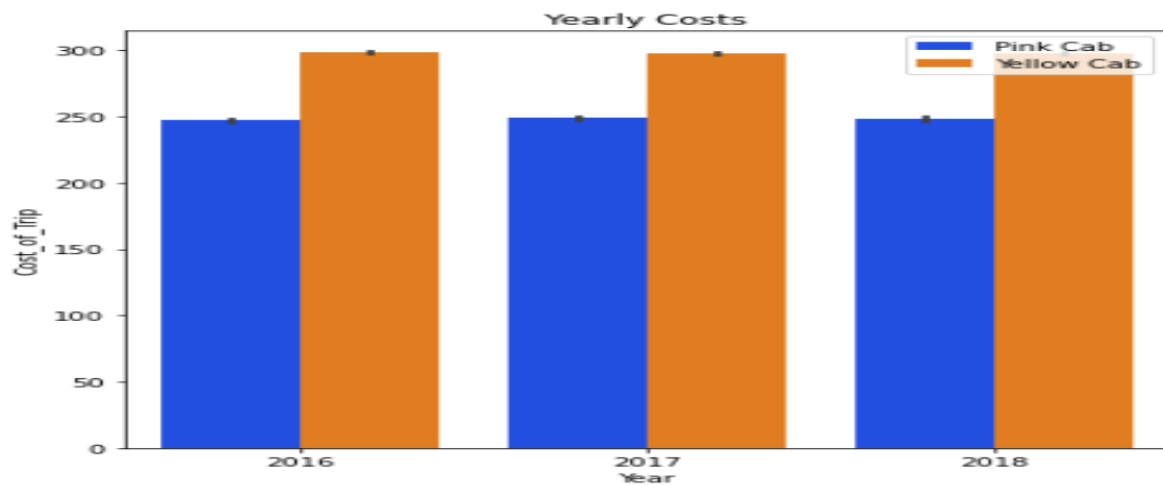
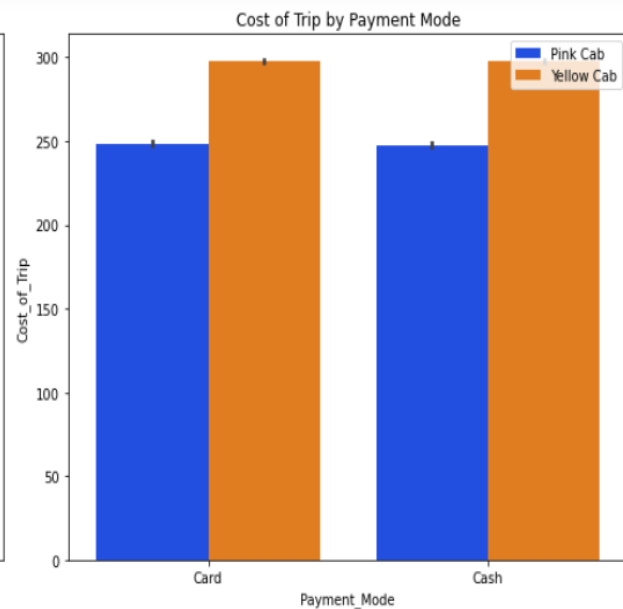
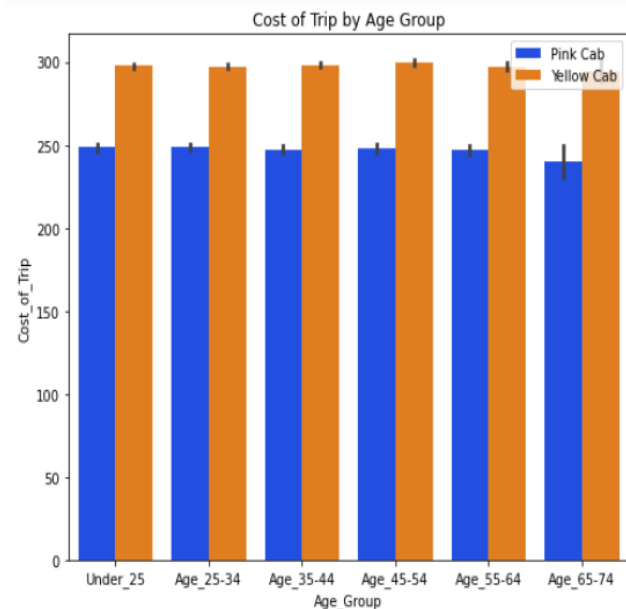
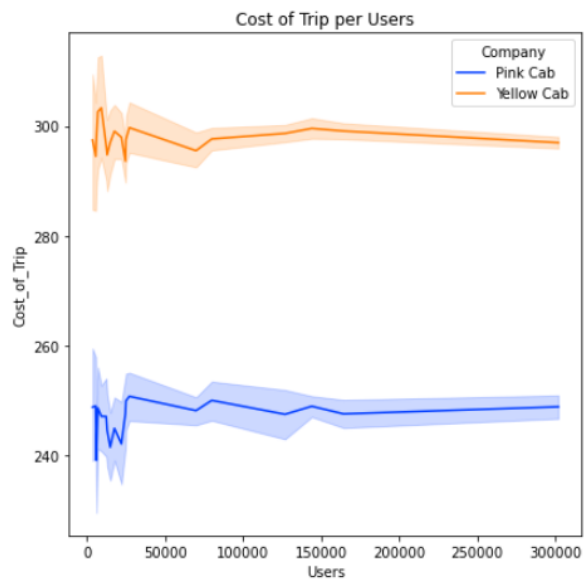
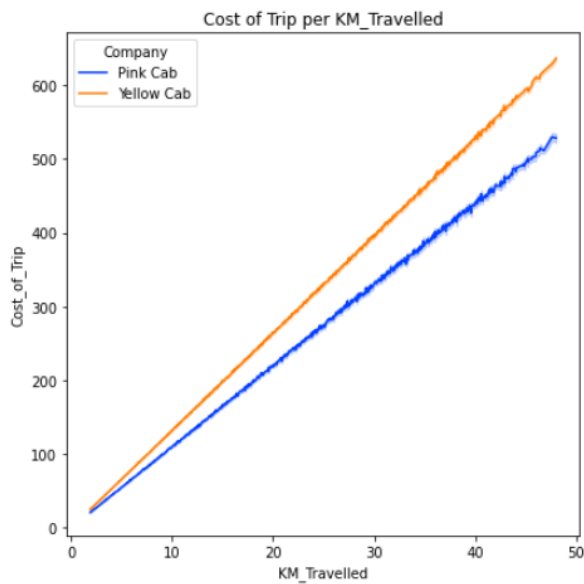
- the demand for cab service slightly decreased in year 2018 as compared to year the previous years.
- Demand for cab service is higher in the winter and lower in the summer.

Demand by Age Group and City

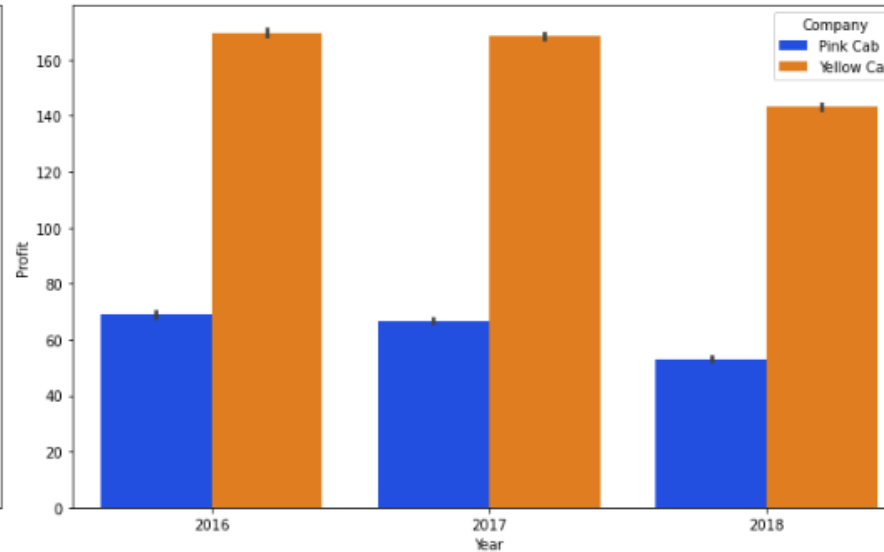
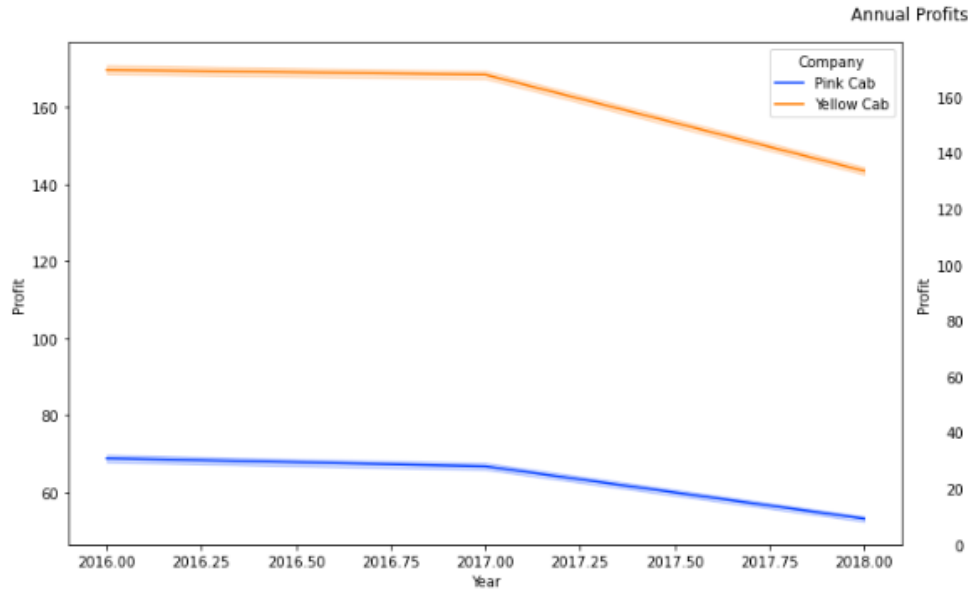


- demand for cab service was higher in big cities such as New York, Chicago and Los Angeles
- demand doesn't change(fairly the same) across different age groups

Cost Analysis

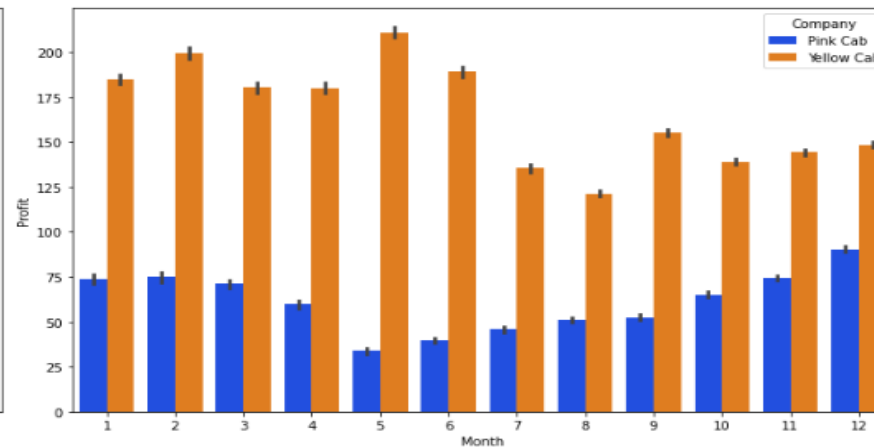
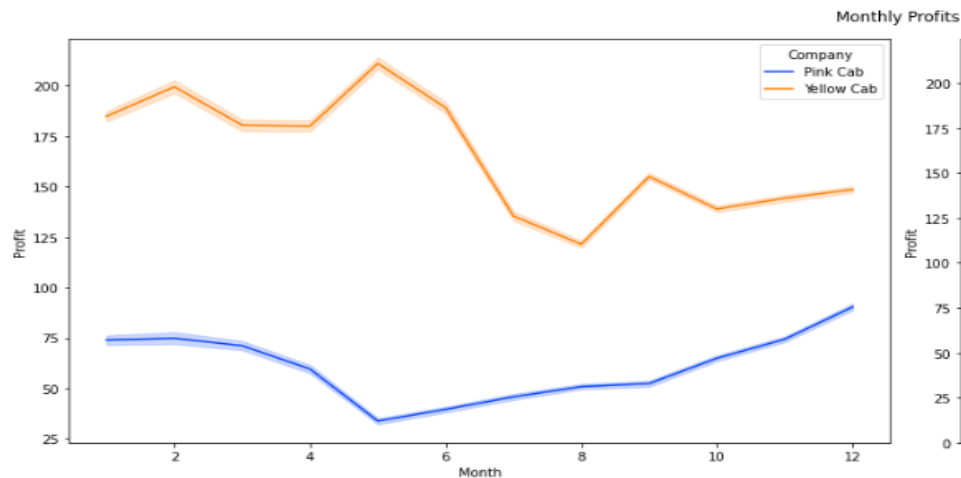


Profitability Analysis



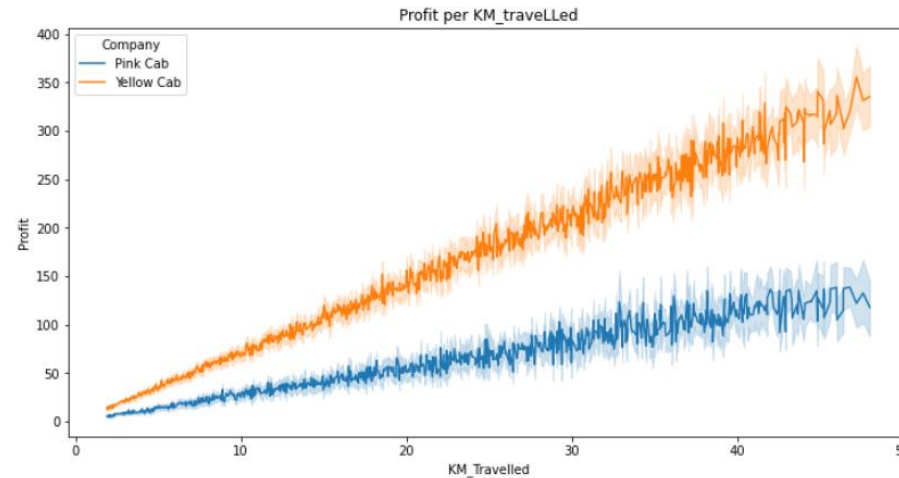
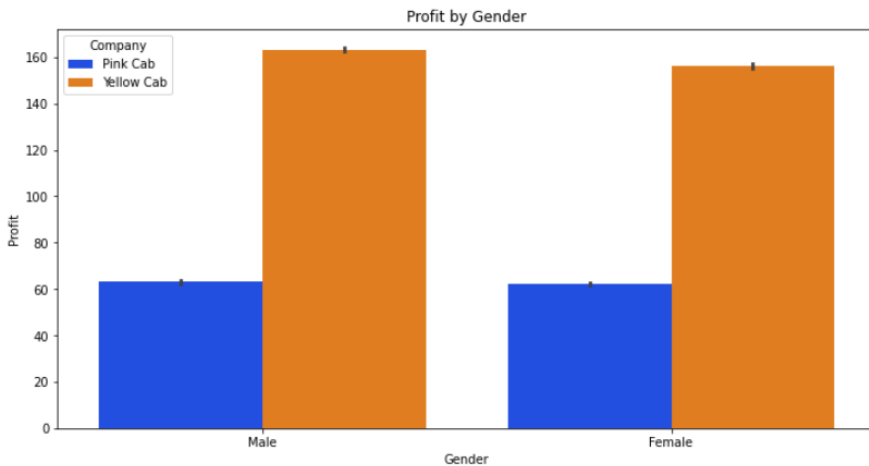
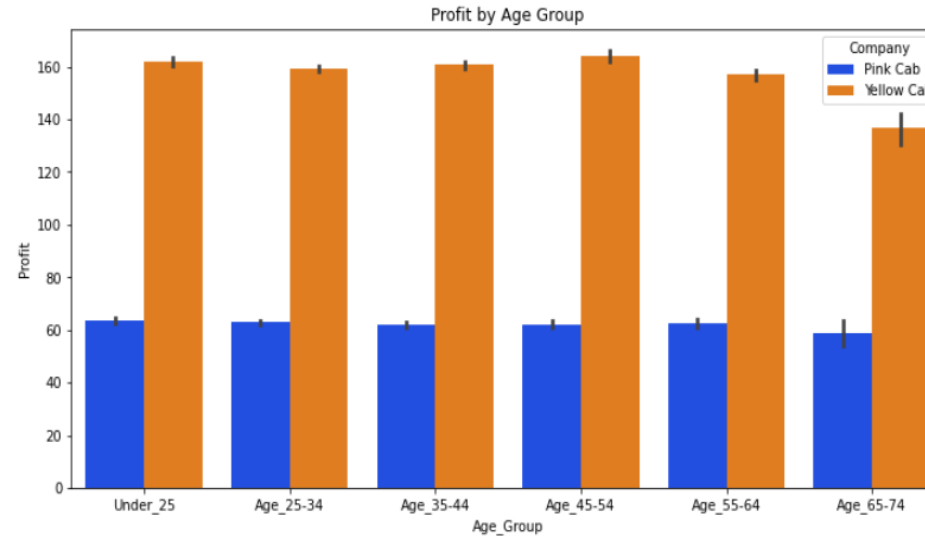
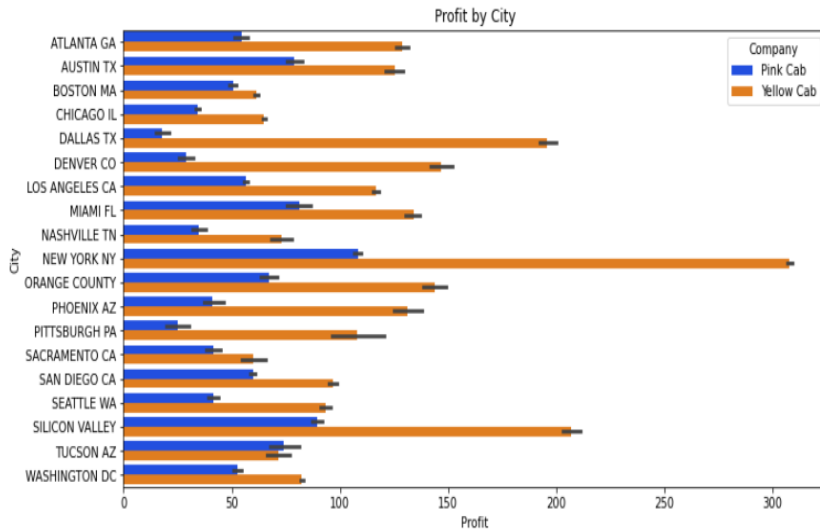
Average annual profits

		Profit
Year	Company	
2016	Pink Cab	68.810861
	Yellow Cab	169.693701
2017	Pink Cab	66.714142
	Yellow Cab	168.416894
2018	Pink Cab	53.124295
	Yellow Cab	143.450759



- Yellow cab company is more profitable than pink company
- profit decreased in 2018 for both companies
- profit varies based on season. Overall, profit was high in winter and low in summer

Profit per KM Traveled, City, Gender and Age group



- New York is the most profitable city for both companies
- Dallas Texas and Sacramento are the least profitable cities for Pink and Yellow cab Cos, respectively.
- Much of the profit come from customers between age 25- 44
- Profit per km traveled is higher for Yellow company

Linear Regression

Linear Regression

```
In [36]: 1 # Linear regression
         2 linear_model = LinearRegression()
         3
         4 linear_model.fit(scaled_X_train, y_train)
```

Out[36]: LinearRegression()

```
In [37]: 1 y_pred = linear_model.predict(scaled_X_test)
```

```
In [38]: 1 # mean absolute error
         2 mae = mean_absolute_error(y_test, y_pred)
         3 mae
```

Out[38]: 0.13846024556698114

```
In [39]: 1 # mean squared error
         2 mse = mean_squared_error(y_test, y_pred)
         3 mse
```

Out[39]: 0.030544861511077694

```
In [40]: 1 # measure goodness of fit
         2
         3 r_square = linear_model.score(scaled_X_train, y_train)
         4 r_square
```

Out[40]: 0.9529941083437189

Ridge Regression

Ridge Regression ¶

```
In [41]: 1 ridge_model= Ridge(alpha= 10.0)
          2 ridge_model.fit(scaled_X_train, y_train)
```

Out[41]: Ridge(alpha=10.0)

```
In [42]: 1 y_pred= ridge_model.predict(scaled_X_test)
```

```
In [43]: 1 # evaluate performance of the model
          2 mae= mean_absolute_error(y_test,y_pred)
          3 mae
```

Out[43]: 0.13846112631478924

```
In [44]: 1 # evaluate performance of the model
          2 mse= mean_squared_error(y_test, y_pred)
          3 mse
```

Out[44]: 0.030544346059260232

```
In [45]: 1 # measure goodness of fit
          2
          3 r_2= r2_score(y_test, y_pred)
          4 r_2
```

Out[45]: 0.9528349711945343

Lasso Regression

Lasso Regression

```
In [47]: 1 lasso_cv_model = LassoCV(eps=0.1,n_alphas=100,cv=5)
          2
          3 lasso_cv_model.fit(scaled_X_train, y_train)
          4
          5 y_pred= lasso_cv_model.predict(scaled_X_test)
```

```
In [48]: 1 # evaluate model performance
          2 mae= mean_absolute_error(y_test,y_pred)
          3 mae
```

Out[48]: 0.1772069402192157

```
In [49]: 1 # evaluate model performance
          2 mse= mean_squared_error(y_test, y_pred)
          3 mse
```

Out[49]: 0.049080555971406446

```
In [50]: 1 # measure goodness of fit
          2
          3 r_square= lasso_cv_model.score(scaled_X_train, y_train)
          4 r_square
```

Out[50]: 0.9244652443677643

Elastic Net

Elastic Net

```
In [51]: 1 # build elastic net model
          2 elastic_model = ElasticNetCV(l1_ratio=[.1, .5, .7,.9, .95, .99, 1],tol=0.01)
          3
          4 elastic_model.fit(scaled_X_train, y_train)
          5
          6 y_pred= elastic_model.predict(scaled_X_test)
```

```
In [52]: 1 # model performance
          2 mae= mean_absolute_error(y_test,y_pred)
          3 mae
```

Out[52]: 0.13853832800723023

```
In [53]: 1 # model performance
          2 mse= mean_squared_error(y_test, y_pred)
          3 mse
```

Out[53]: 0.030567504586597124

```
In [54]: 1 # measure goodnes of fit
          2
          3 r_square= elastic_model.score(scaled_X_train, y_train)
          4 r_square
```

Out[54]: 0.9529613705195545

Coefficients

```
In [63]: 1 coeff_df = pd.DataFrame(final_model.coef_, X.columns, columns=['Coefficient'])
          2 coeff_df
```

Out[63]:

	Coefficient
Cost_of_Trip	9.974561e-01
Age	3.998624e-04
Income_(USD/Month)	2.853799e-04
Population	4.306607e+10
Users	-6.604654e+10
Year	-5.079945e+01
Month	8.128779e-03
Day	-6.869461e-03
Company_Yellow Cab	1.552170e-01
City_AUSTIN TX	-2.639385e+10
City_BOSTON MA	1.268965e+11
City_CHICAGO IL	8.752365e+10
City_DALLAS TX	-1.346258e+10
City_DENVER CO	-4.207042e+10
City_LOS ANGELES CA	8.757297e+10
City_MIAMI FL	-4.349714e+10
City_NASHVILLE TN	-2.543179e+10
City_NEW YORK NY	6.488271e+10
City_ORANGE COUNTY	-5.251967e+10
City_PHOENIX AZ	-9.833935e+10
City_PITTSBURGH PA	-1.088453e+11
City_SACRAMENTO CA	-6.559724e+10
City_SAN DIEGO CA	6.176420e+10
City_SEATTLE WA	9.312367e+09
City_SILICON VALLEY	-9.377772e+09
City_TUCSON AZ	-8.571778e+10
City_WASHINGTON DC	1.368002e+11
Payment_Mode_Cash	-5.738274e-04
Gender_Male	3.468354e-03

Conclusions and Recommendation

Conclusions:

- Yellow cab company has large market share in 19 cities however, Pink cab Company has only in 2 cities. Yellow cab company's total market share triples Pink cab company's market share.
- demand is higher for yellow cab company in all years, although it slightly decreased for both companies in 2018.
- Demand for Yellow cab co. is higher in every season. Yellow cab company is preferred by all age groups
- Yellow cab is more profitable than Pink company. It generates more profit per user and km traveled.

Recommendation:

- Yellow cab company excels in all parameters used in this project, So i will recommend XYZ co to invest its money in **Yellow cab Company**.

Thank You



Data Glacier

Your Deep Learning Partner