Exploratory Data Analysis

G2M Insight for Cab Investment Firm

Kutay Selçuk

21/08/2022

Agenda

- Executive Summary
- Problem Statement
- Approach
- EDA
- EDA Summary
- Recommendations

Executive Summary

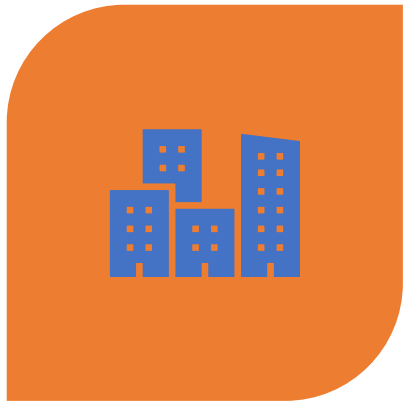
XYZ IS A PRIVATE FIRM IN US, AND THEY WANT TO INVEST IN CAB INDUSTRY. THEY ARE TRYING TO FIND OUT WHICH CAB COMPANY IS MORE PROFITABLE FOR THEM, PINK CAB COMPANY OR YELLOW CAB COMPANY.

INVESTIGATION WILL GO THROUGH THE MASTER DATA WHICH IS CREATED FROM OTHER DATA SETS, AND I WILL ANALYZE DATA AND TRY TO FIND OUT RELATIONS BETWEEN DATA SETS.

RESULTS WILL BE SHOWN AS GRAPHS ACCORDING TO RELATIONS.

THERE WILL HYPOTHESIS TESTING AND RECOMMENDATIONS FOR INVESTMENT IN THE END.

Problem Statement

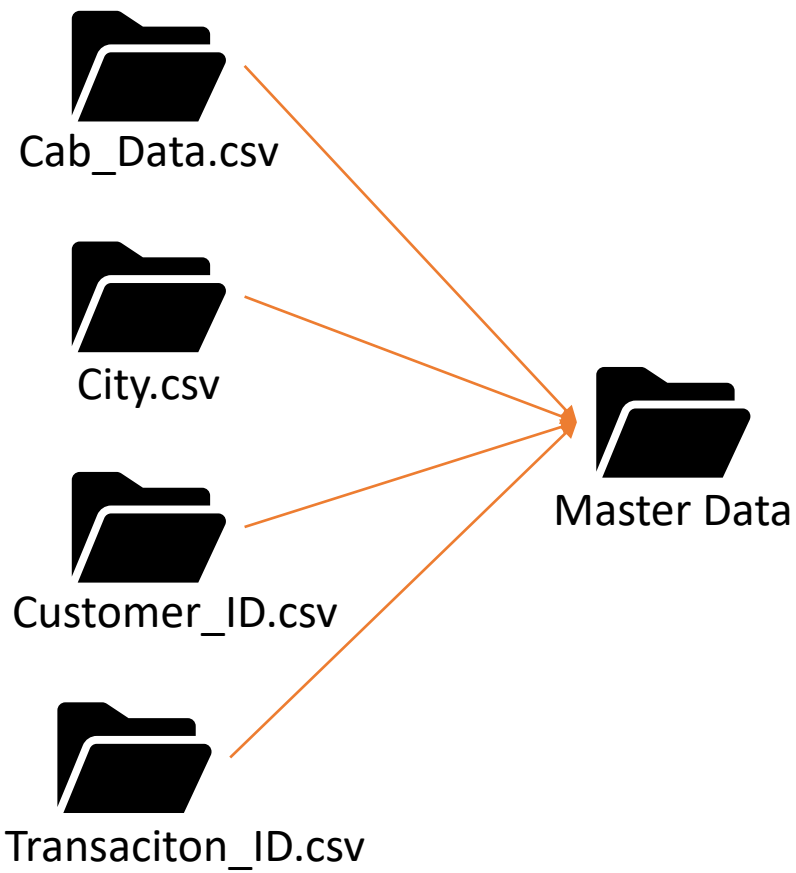


XYZ IS A PRIVATE FIRM IN US. DUE TO REMARKABLE GROWTH IN THE CAB INDUSTRY IN LAST FEW YEARS AND MULTIPLE KEY PLAYERS IN THE MARKET, IT IS PLANNING FOR AN INVESTMENT IN CAB INDUSTRY AND AS PER THEIR GO-TO-MARKET(G2M) STRATEGY THEY WANT TO UNDERSTAND THE MARKET BEFORE TAKING FINAL DECISION.

THERE ARE TWO DIFFERENT OPTIONS FOR THIS INVESTMENT, AND THIS PROJECT IS COMPERING THESE OPTIONS IN DATA DRIVEN WAY.

PROJECT'S IMPORTANCE IS COMING FROM NOT JUST VISUALIZATION BUT ALSO SEARCHING AND CREATING RELATIONS BETWEEN DATA SETS. PROJECT ASKED THE MOST IMPORTANT QUESTIONS AND COMPILED ANSWERS FOR RECOMMENDATIONS.

Approach



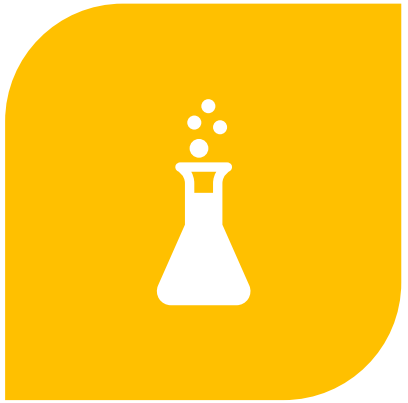
Information and Assumptions

- 01-01-2016 and 31-12-2018 is the time interval of data sets.
- Time column separated as day, month and year as well.
- There aren’t duplicated rows or N/A values in rows.
- “Price Charged – Cost of Trip” is used to calculate profit and Profit column is added to master data.

EDA

DATA INFORMATION

GRAPHS, EXPLORATION AND ANALYSIS



HYPOTHESIS TESTING

```
customer_id.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 49171 entries, 0 to 49170
Data columns (total 4 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   Customer ID         49171 non-null  int64 
 1   Gender              49171 non-null  object
 2   Age                 49171 non-null  int64 
 3   Income (USD/Month)  49171 non-null  int64 
dtypes: int64(3), object(1)
memory usage: 1.5+ MB
```

```
transaction_id.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 440098 entries, 0 to 440097
Data columns (total 3 columns):
 #   Column          Non-Null Count   Dtype 
---  ------          --------------   ----- 
 0   Transaction ID  440098 non-null  int64 
 1   Customer ID     440098 non-null  int64 
 2   Payment_Mode    440098 non-null  object
dtypes: int64(2), object(1)
memory usage: 10.1+ MB
```

```
cab.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 359392 entries, 0 to 359391
Data columns (total 7 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
dtypes: float64(3), int64(2), object(2)
memory usage: 19.2+ MB
```

```
city.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20 entries, 0 to 19
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   City        20 non-null     object 
 1   Population  20 non-null     float64
 2   Users       20 non-null     float64
dtypes: float64(2), object(1)
memory usage: 608.0+ bytes
```

```
master_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 359392 entries, 0 to 182928
Data columns (total 17 columns):
 #   Column              Non-Null Count   Dtype         
---  ------              --------------   -----         
 0   City                359392 non-null  object        
 1   Population          359392 non-null  float64       
 2   Users               359392 non-null  float64       
 3   Transaction ID      359392 non-null  int64         
 4   Date of Travel      359392 non-null  datetime64[ns]
 5   Company             359392 non-null  object        
 6   KM Travelled        359392 non-null  float64       
 7   Price Charged       359392 non-null  float64       
 8   Cost of Trip        359392 non-null  float64       
 9   Customer ID         359392 non-null  int64         
 10  Payment_Mode        359392 non-null  object        
 11  Gender              359392 non-null  object        
 12  Age                 359392 non-null  int64         
 13  Income (USD/Month)  359392 non-null  int64         
 14  Year                359392 non-null  int64         
 15  Month               359392 non-null  int64         
 16  Day                 359392 non-null  int64         
dtypes: datetime64[ns](1), float64(5), int64(7), object(4)
memory usage: 49.4+ MB
```

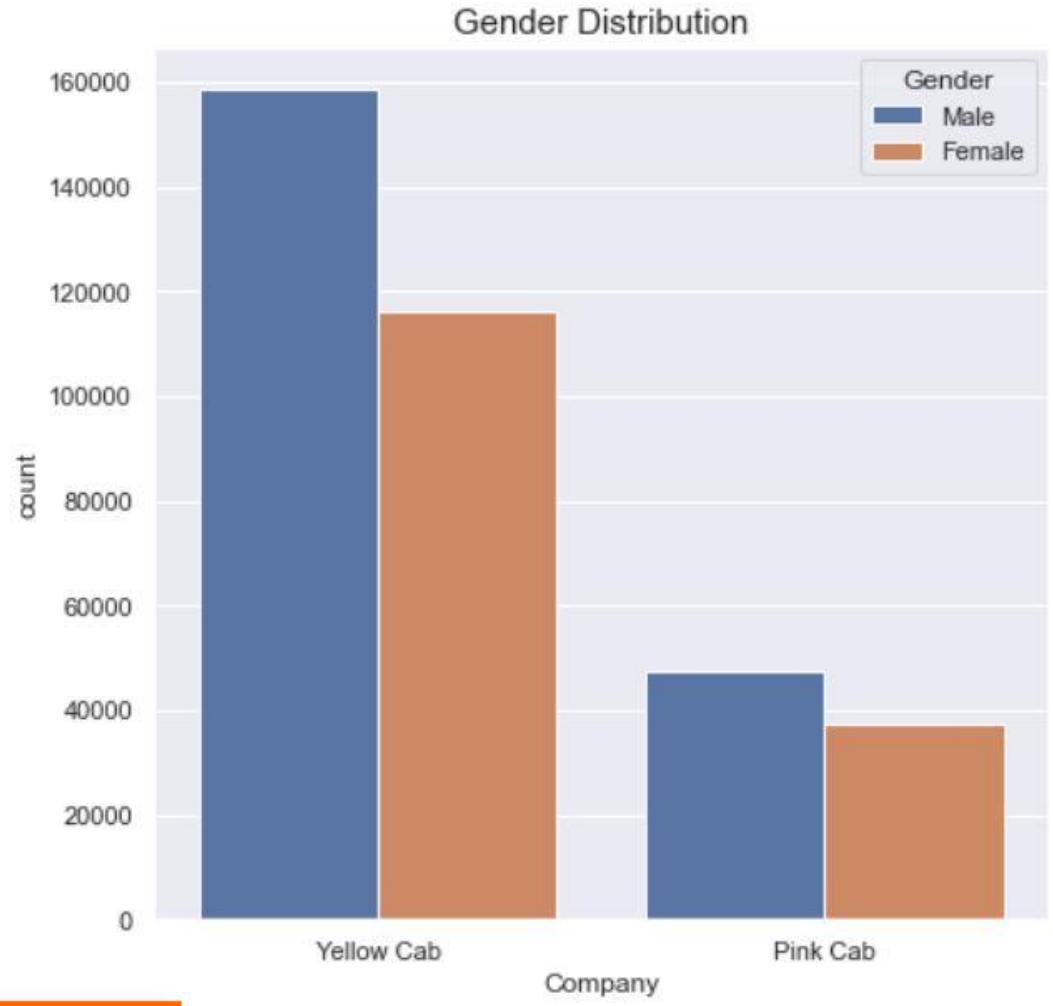
Data Information

- These are the information of data sets after rearrangements which are reading files, checking for nulls and duplications, changing data types of some data and merging in master data.

Correlations

- There are positive correlations between “Price Charged” and “KM Travelled” and “Price Charged” and “Cost od Trip”.

Gender Distributions

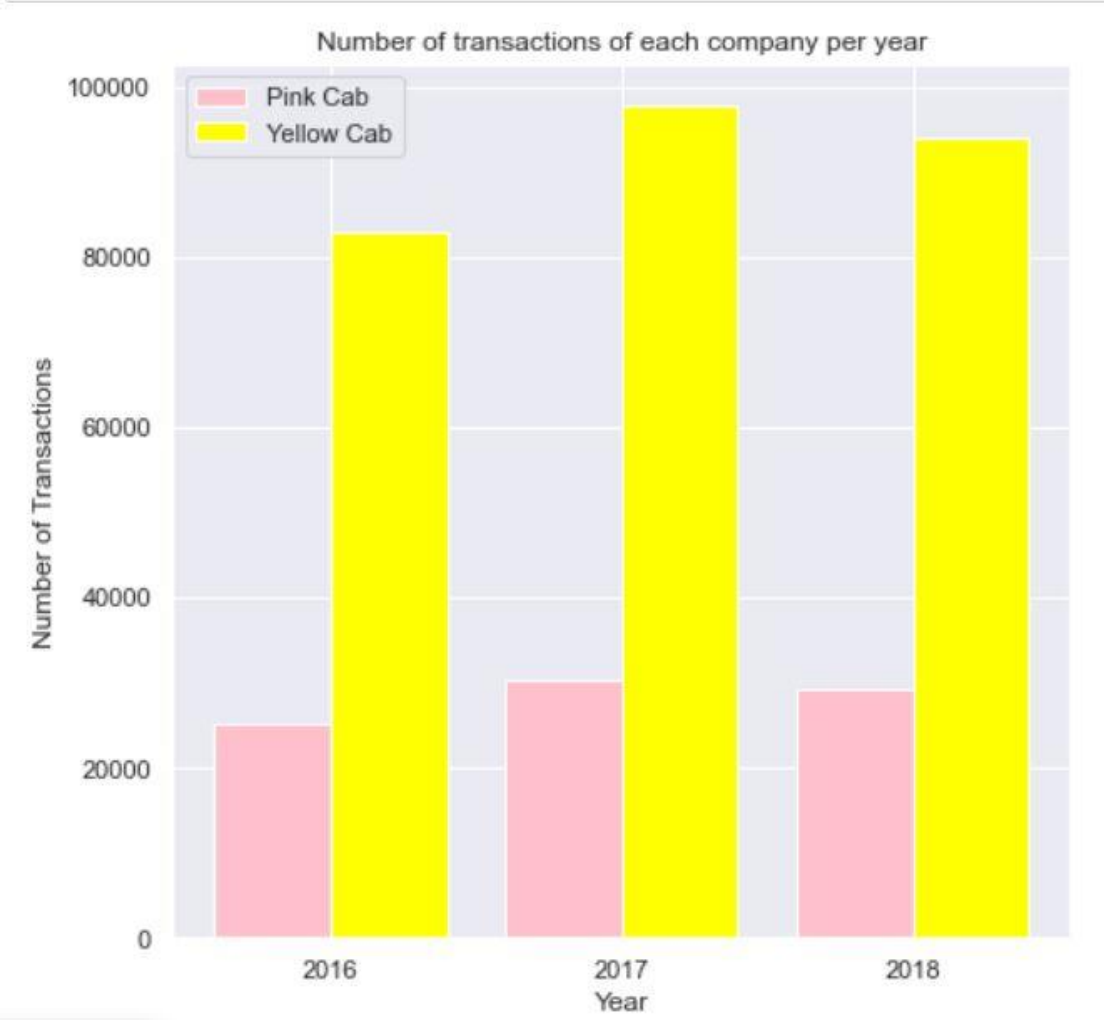


- Both genders use Yellow Cab more than Pink Cab as a result of general user counts.

Payment Preferences

- Users prefer using card more than cash for both companies.

Transaction Counts per Year

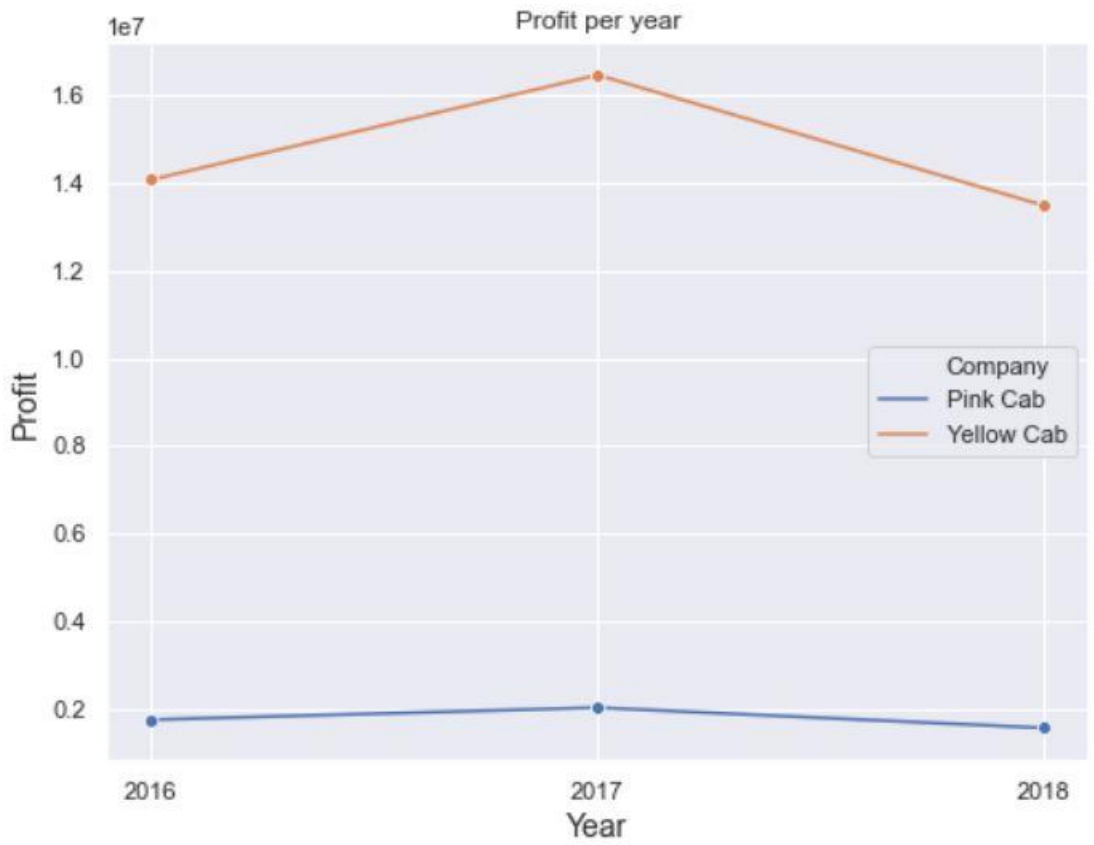


- Yellow Cab was preferred more than Pink Cab for all of three years.

Transaction Counts per Month

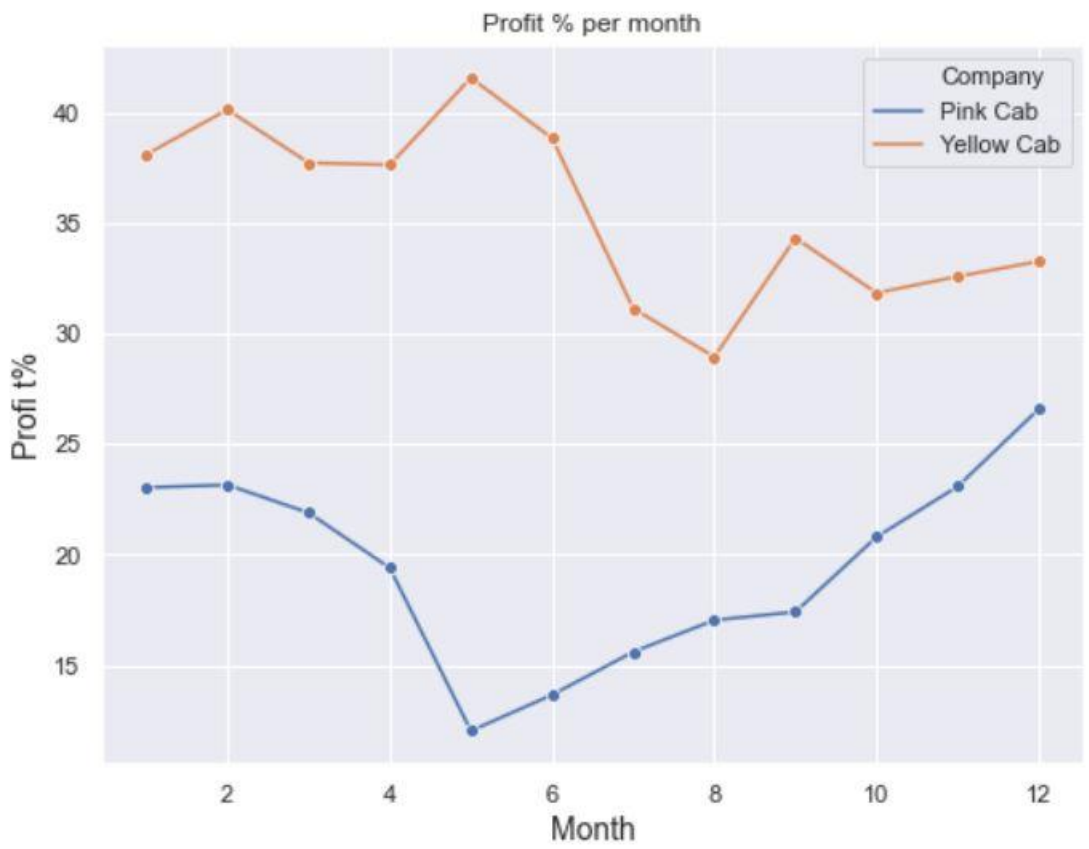
- Yellow Cab was preferred more than Pink Cab for all of months. Both companies had more users throughout end of year which are holiday session for U.S. but also gap were getting bigger as well.

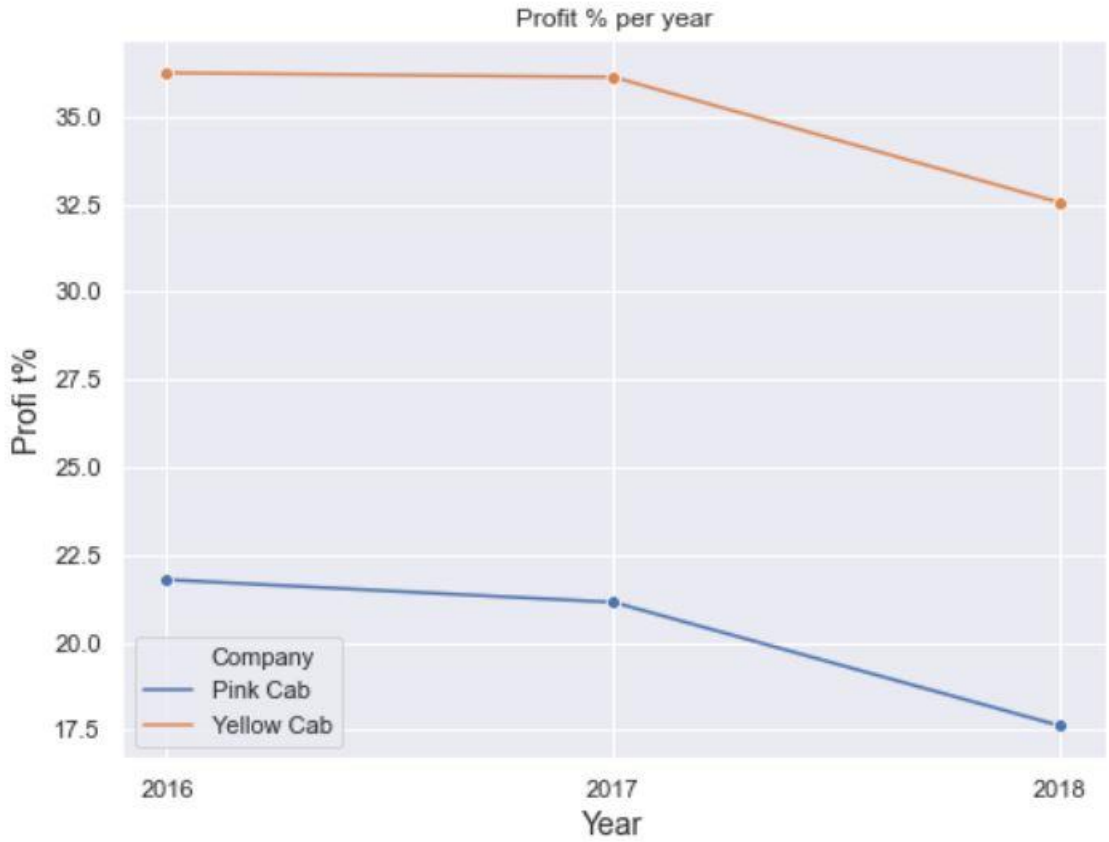
Profits per Months and Years



Yellow Cab made more profit than Pink Cab for all the months and years.

Profit % per Months and Years





Yellow Cab was more profitable than Pink Cab for all the months and years.

Hypothesis 1

H0 : There is no difference regarding Gender in both cab companies.

H1 : There is difference regarding Gender in both cab companies.

In [54]:
```
#Pink Cab

data_1 = master_data[(master_data.Gender == 'Male')&(master_data.Company == 'Pink Cab')].groupby('Transaction ID').Profit.mean()
data_2 = master_data[(master_data.Gender == 'Female')&(master_data.Company == 'Pink Cab')].groupby('Transaction ID').Profit.mean(

_, p_value = stats.ttest_ind(data_1.values,
                             data_2.values,
                             equal_var=True)

print('P value is ', p_value)
```

```
P value is  0.11515305900425798
```

We accept null hypothesis.

In [55]:
```
#Yellow Cab

data_1 = master_data[(master_data.Gender == "Male")&(master_data.Company == 'Yellow Cab')].groupby("Transaction ID").Profit.mean(
data_2 = master_data[(master_data.Gender == "Female")&(master_data.Company == 'Yellow Cab')].groupby("Transaction ID").Profit.mea

_, p_value = stats.ttest_ind(data_1.values,
                             data_2.values,
                             equal_var=True)

print("P value is ", p_value)
```

```
P value is  6.060473042494144e-25
```

We reject null hytpothesis.

- It seems like women preferred Yellow Cab.

Hypothesis 2

H0 : There is no difference regarding payment methods in both cab companies.

H1 : There is difference regarding payment methods in both cab companies.

In [56]:
```
data_1 = master_data[(master_data.Payment_Mode == "Card")&(master_data.Company == "Pink Cab")].groupby("Transaction ID").Profit.r
data_2 = master_data[(master_data.Payment_Mode == "Cash")&(master_data.Company == "Pink Cab")].groupby("Transaction ID").Profit.r

_, p_value = stats.ttest_ind(data_1.values,
                             data_2.values,
                             equal_var = True)

print("P value is ", p_value)
```

```
P value is  0.7900465828793288
```

We accept null hypothesis.

In [57]:
```
data_1 = master_data[(master_data.Payment_Mode == "Card")&(master_data.Company == "Yellow Cab")].groupby("Transaction ID").Profit
data_2 = master_data[(master_data.Payment_Mode == "Cash")&(master_data.Company == "Yellow Cab")].groupby("Transaction ID").Profit

_, p_value = stats.ttest_ind(data_1.values,
                             data_2.values,
                             equal_var = True)

print("P value is ", p_value)
```

```
P value is  0.29330606382985325
```

- It seems like there is no difference regarding payment methods in both cab companies.

Hypothesis 3

H0 : There is no difference regarding income level in both cab companies.

H1 : There is difference regarding income level in both cab companies.

We accept null hypothesis.

In [58]:
```
master_data["Income (USD/Month)"].median()
```

Out[58]:
```
14685.0
```

In [59]:
```
data_1 = master_data[(master_data["Income (USD/Month)"] <= 14685)&(master_data.Company == "Pink Cab")].groupby("Transaction ID").
data_2 = master_data[(master_data["Income (USD/Month)"] > 14685)&(master_data.Company == "Pink Cab")].groupby("Transaction ID").F

_, p_value = stats.ttest_ind(data_1.values,
                             data_2.values,
                             equal_var = True)

print("P value is ", p_value)
```

```
P value is  0.07500814329070742
```

We accept null hypothesis.

In [60]:
```
data_1 = master_data[(master_data["Income (USD/Month)"] <= 14685)&(master_data.Company == "Yellow Cab")].groupby("Transaction ID"
data_2 = master_data[(master_data["Income (USD/Month)"] > 14685)&(master_data.Company == "Yellow Cab")].groupby("Transaction ID"

_, p_value = stats.ttest_ind(data_1.values,
                             data_2.values,
                             equal_var = True)

print("P value is ", p_value)
```

```
P value is  1.2025645018067682e-07
```

We reject null hypothesis. It looks like Yellow Cab company is prefered by high income levels.

- It seems like Pink Cab was more preferred by high income levels.
- I used median of the income levels data which is 14,685 to separate low and high income levels.

EDA Summary

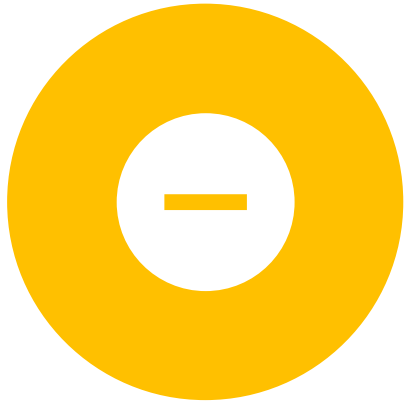
- Both genders used Yellow Cab.
- Payments made by card more than cash.
- Yellow Cab had more transactions for all months and years.
- Yellow Cab made more profit for all months and years.
- Yellow Cab was more profitable according to costs for all moths and years.
- Women prefers Yellow Cab mostly.
- There was no difference between payment methods for both companies.
- It seems like Pink Cab preferred by high income levels more than Yellow Cab, but Yellow Cab was profitable clearly. Also, we don’t know actual distribution of income levels, therefore it is hard to define something positive for Pink Cab in this hypothesis.

Recommendation

According the EDA I recommend Yellow Cab for investment.

YELLOW CAB HAS MORE TRANSACTION COUNT, SO I CAN SAY IT IS MORE POPULAR THAN PINK CAB.

YELLOW CAB FITS WITH OBJECTIVE TO MAKE PROFIT. ALSO, IT IS MORE PROFITABLE ACCORDING TO PROFIT-COST RATIO THAN PINK CAB.



THERE ARE NO POSITIVE SIGNS TO CHOOSE PINK CAB, AS WELL.

Thank You