Data Glacier

Your Deep Learning Partner

# Exploratory Data Analysis

## G2M Case Study

**16/06/2022**

**Ramya Mercy Rajan**

**Batch : LISUM 10:30 , Data Science**

# Agenda

| S.No | Index |
|---|---|
| 1 | Executice Summary |
| 2 | Exploratory Data Analysis |
| 3 | Multivariate Analysis |
| 4 | Profit Analysis |
| 5 | Areas to Investigate - Company requirement |
| 6 | Some Findings based on Analysis |
| 7 | Hypothesis Testing |
| 8 | Outlier Detection and Removal |
| 9 | Label Encoding |
| 10 | Training and Testing |
| 11 | Applying Model |
| 12 | Recommendation |

Data Glacier
Your Deep Learning Partner

# Executive Summary

- **About Project :**

  The purpose of this Project is making suggestion on "Investment" for a Company XYZ in United States of America by analyzing and investigating the records of two Cab Companies based on the informations provided.

- **The Client : XYZ**

  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.

# Executive Summary

- **About Project :**

  The purpose of this Project is making suggestion on "Investment" for a Company XYZ in United States of America by analysing and investigating the records of two Cab Companies based on the information's provided.

- **The Client : XYZ**

  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, as 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.

- **Project delivery :**

  Provided with multiple data sets that contains information on 2 cab companies. Each file (data set) provided represents different aspects of the customer profile. XYZ is interested in using actionable insights to help them identify the right company to make their investment.

# Executive Summary

- **Provided 4 individual data sets.**
- **Time period of data is from 31/01/2016 to 31/12/2018.**
- **Below are the list of datasets which are provided for the analysis:**
  1) **Cab_Data.csv** – this file includes details of transaction for 2 cab companies.
  2) **Customer_ID.csv** – this is a mapping table that contains a unique identifier which links the customer's demographic details.
  3) **Transaction_ID.csv** – this is a mapping table that contains transaction to customer mapping and payment mode
  4) **City.csv** – this file contains the list of US cities , their population and the number of cab users.
- **Analysis made :**

  Create multiple hypothesis

- **Areas to Investigated :**

  1. Which company has maximum cab users at a particular time period?
  2. Does margin proportionally increase with increase in number of customers?
  3. What are the attributes of these customer segments?

# Executive Summary

**Topic covered as per requirement**

1. Review the Source Documentation
2. Understand the field names and data types
3. Identify relationships across the files
4. Field/feature transformations
5. Determine which files should be joined versus which ones should be appended
6. Create master data and explain the relationship
7. Identify and remove duplicates
8. Perform other analysis like NA value and outlier detection

# Exploratory Data Analysis

- **Data Sets :** 4 individual datasets are given

1. ***Cab_Data.csv*** *(*359392 Rows and 7 Columns)

   This file contains the Cab Companies Information.

   a) Number of Cab Companies-2
   b) Names of Cab Companies - Yellow Cab and Pink Cab
   c) Date of Travel
   d) City
   e) Kilometres Travelled
   f) Price Charged by each company
   g) Cost of Trip

# Exploratory Data Analysis

2. ***Customer_ID.csv*** *(*49171 Rows and 4Columns)

   This file contains the customer's demographic details.

   a) Customer ID
   b) Gender
   c) Age
   d) Income

3. ***City.csv*** *(20* Rows and 3Columns)

   This file contains list of US cities, their population and number of cab users.

   a) US cities
   b) Population
   c) Number of cab users

# Exploratory Data Analysis

***4. Transaction_ID.csv*** *(440098* Rows and 3Columns)

This file contains list of

a) Transaction ID

b) Customer ID

c) Payment_Mode

- **Data Preprocessing-** There was no Null Values and Duplicate values in the four datasets.
- **Outliers** was found in the feature Price Charged and rectified.
- **Additional Features**
  - ➢ The feature "Date of Travel" was not in a proper format and was corrected using datetime.timedelta() function. A new "Date" column was created after correction of the old feature "Date of Travel".
  - ➢ The Profit column was created by finding the difference between Price Charged and Cost of Trip. Profit was considered as my target variable.
  - ➢ Age Groups column was created using class intervals format(Bins) in order to know the age groups performance on other features.
- **Statistical measure** was checked on every datasets in order to know the Count, Mean, Std, Min and Max.

# Multivariate Analysis

- Multivariate analysis was made to check the relationship between Profit and other features .

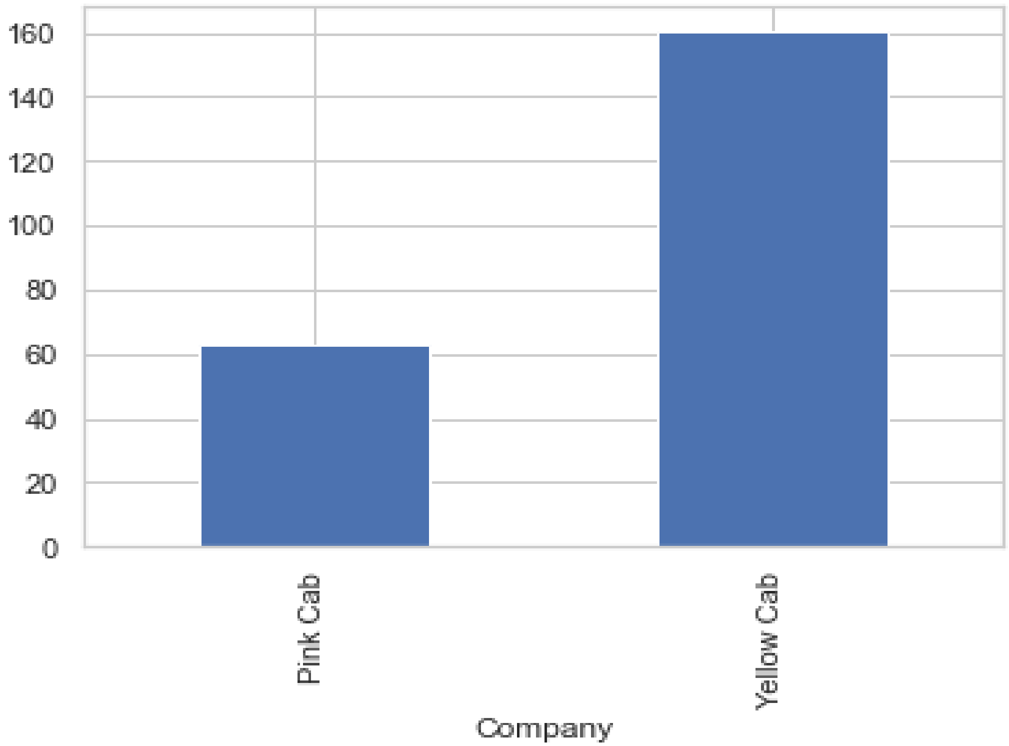
```
In [45]: df.groupby(['Company','Price Charged','Cost of Trip','Age','KM Travelled'])['Profit'].mean()

Out[45]: Company     Price Charged  Cost of Trip  Age  KM Travelled
         Pink Cab    15.60          21.3840       34   1.98              -5.7840
                     15.75          24.7800       32   2.10              -9.0300
                     16.38          19.3800       43   1.90              -3.0000
                     16.53          19.2000       57   1.92              -2.6700
                     16.76          19.9820       18   1.94              -3.2220
                                                                           ...
         Yellow Cab  1981.05        556.9092      37   41.81           1424.1408
                     1993.83        594.7200      60   47.20           1399.1100
                     2013.95        580.6080      64   43.20           1433.3420
                     2016.70        571.4280      37   43.29           1445.2720
                     2048.03        584.0640      18   46.80           1463.9660
         Name: Profit, Length: 359386, dtype: float64
```

# Profit Analysis based on Companies

Observing Graph of Company vs Profit

- Noticed Yellow Cab have more Profit

In [46]:
```
graph_1 =df.groupby(['Company'])['Profit'].mean().plot(kind='bar')
```



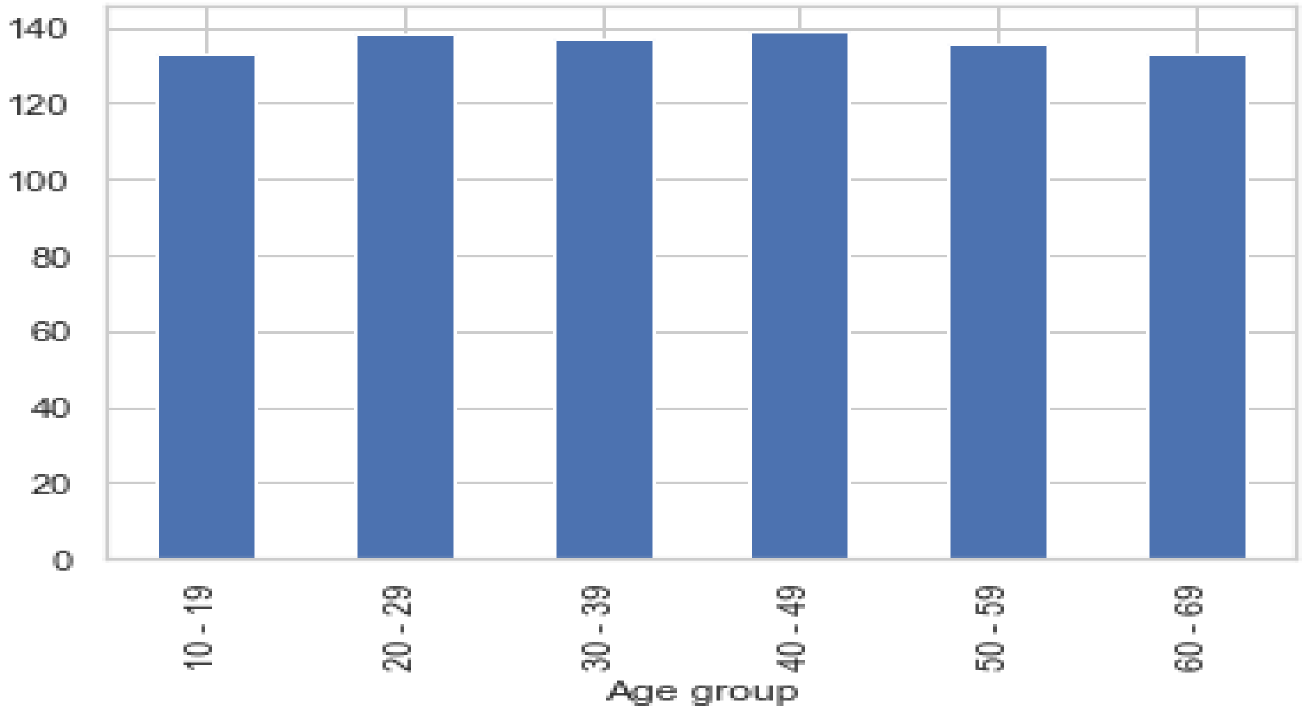
# Profit Analysis based Age

## Observing Graph of Age vs Profit

## - Noted Age groups from 20-49 creates more Profit

In [47]:

```
graph_2 = df.groupby(['Age group'])['Profit'].mean().plot(kind='bar')
```



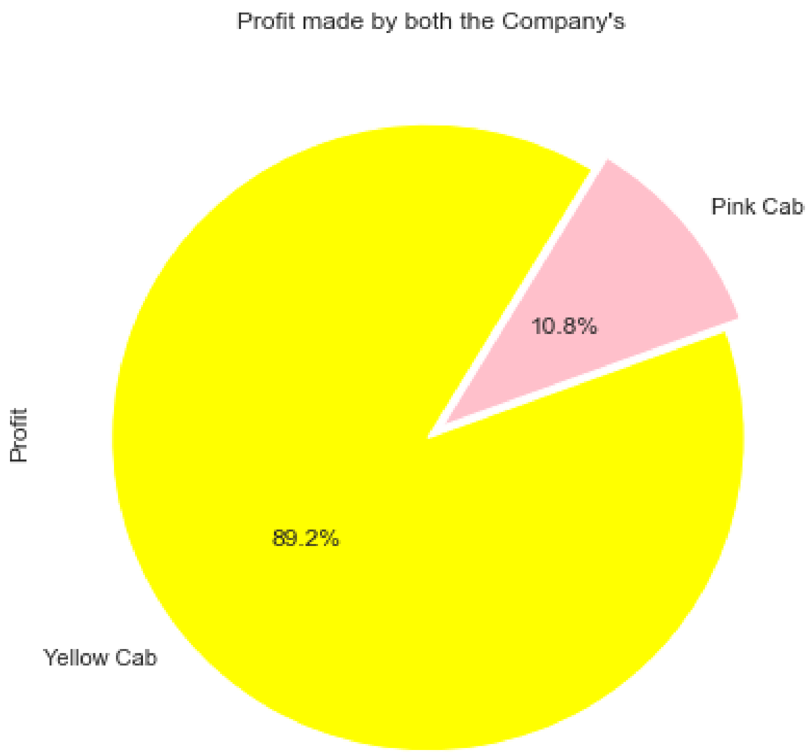
# Profit Analysis based on both Cabs

## Profit per cabs

In [48]:

```
plt.title("Profit made by both the Company's ")
df.groupby('Company')['Profit'].sum().plot(kind='pie',y='Company',startangle=20,colors = ("pink", "yellow") ,figsize=(15,7),autopct='%1.1f%%',explode=
```

Out[48]: <AxesSubplot:title={'center':"Profit made by both the Company's "}, ylabel='Profit'>



# Areas to investigate as per Case Study requirement:

## Q.1) Which company has maximum cab users at a particular time period?

In [52]:
```
df.groupby([df['Date'].dt.year,'Company'])['Customer ID'].count()
```

Out[52]:
```
Date  Company
2016  Pink Cab      25080
      Yellow Cab    82239
2017  Pink Cab      30321
      Yellow Cab    98189
2018  Pink Cab      29310
      Yellow Cab    94253
Name: Customer ID, dtype: int64
```

## In 2016, 2017 and 2018, Yellow cab has maximum customers

# Areas to investigate as per Case Study requirement:

- Which company made maximum Profit at a particular time period?

```
In [53]: df.groupby([df['Date'].dt.year,'Company'])['Profit'].mean()

Out[53]: Date  Company
         2016  Pink Cab       68.321819
               Yellow Cab    169.347821
         2017  Pink Cab       67.070839
               Yellow Cab    168.817057
         2018  Pink Cab       53.229689
               Yellow Cab    143.416122
         Name: Profit, dtype: float64
```

# Areas to investigate as per Case Study requirement:

## Q.2) Does margin proportionally increase with increase in number of customers?

In [55]:
```
df.groupby([df['Date'].dt.year,'Company'])['Income (USD/Month)'].mean()
```

Out[55]:
```
Date  Company
2016  Pink Cab      15124.241587
      Yellow Cab    15034.414900
2017  Pink Cab      15058.789123
      Yellow Cab    15069.108974
2018  Pink Cab      15003.528420
      Yellow Cab    15031.072146
Name: Income (USD/Month), dtype: float64
```

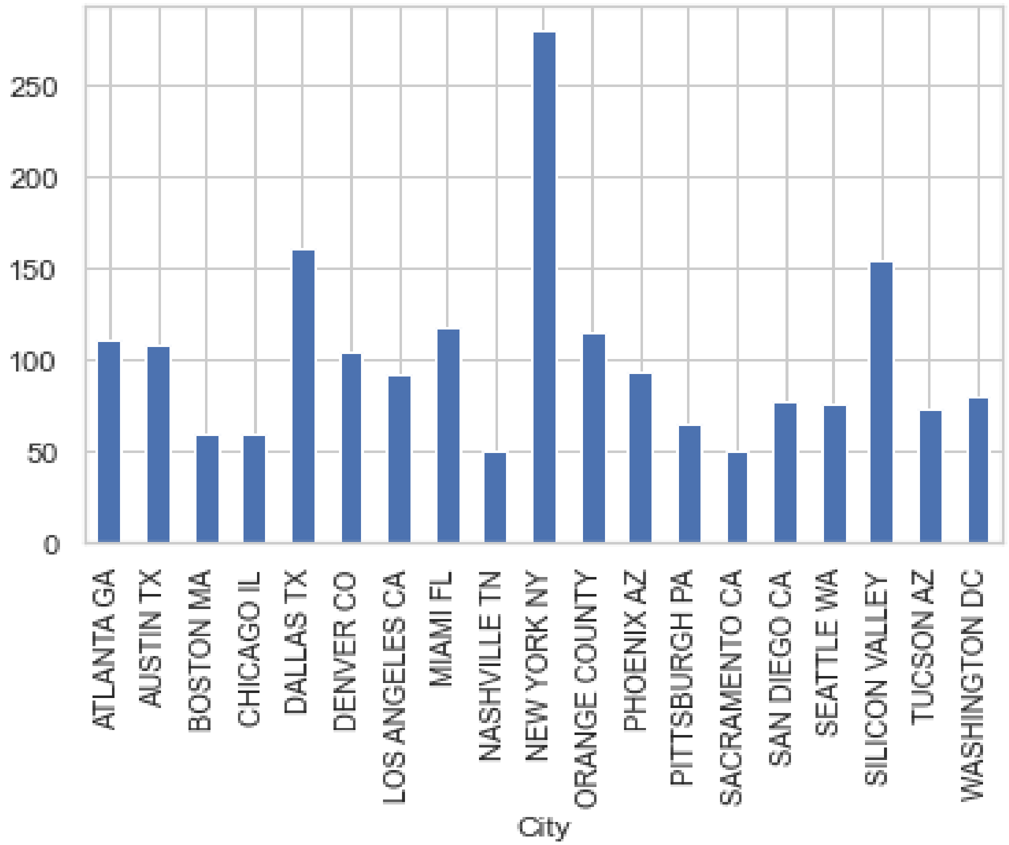
- Outcome: Yellow Cab Company has got higher Income and Customers as compared to Pink Cab Company

# Areas to investigate as per Case Study requirement:

### Q.3) What are the attributes of these customer segments?

In [56]:

```
graph_3 = df.groupby(['City'])['Profit'].mean().plot(kind='bar')
```




Rides to New York made maximum Profit

# Areas to investigate as per Case Study requirement:

b) Yellow Cab Travelled more Kilometers as compared to Pink Cab. So Kilometers Travelled by each cab and the availability of more cabs can be an important factor of getting more Customers.

In [61]:

```
df.groupby([df['Date'].dt.year,'Company'])['KM Travelled'].count()
```

Out[61]:

```
Date  Company
2016  Pink Cab      25080
      Yellow Cab    82239
2017  Pink Cab      30321
      Yellow Cab    98189
2018  Pink Cab      29310
      Yellow Cab    94253
Name: KM Travelled, dtype: int64
```

# Areas to investigate as per Case Study requirement:

c) The Cost of Trip is higher for Yellow Cab as compared to Pink Cab because Yellow Cab had more trips and travelled more kilometers.

In [62]:
```
df.groupby([df['Date'].dt.year,'Company'])['Cost of Trip'].count()
```

Out[62]:
```
Date  Company
2016  Pink Cab      25080
      Yellow Cab    82239
2017  Pink Cab      30321
      Yellow Cab    98189
2018  Pink Cab      29310
      Yellow Cab    94253
Name: Cost of Trip, dtype: int64
```

# Areas to investigate as per Case Study requirement:

d) Yellow cab company has much more income than pink cab company, since it has more customers.

In [63]:
```
df.groupby([df['Date'].dt.year,'Company'])['Income (USD/Month)'].count()
```

Out[63]:
```
Date  Company
2016  Pink Cab      25080
      Yellow Cab    82239
2017  Pink Cab      30321
      Yellow Cab    98189
2018  Pink Cab      29310
      Yellow Cab    94253
Name: Income (USD/Month), dtype: int64
```

# Areas to investigate as per Case Study requirement:

e) When compared with Age group in the year 2016 , 2017 and 2018 people aged between 20-39 might have used more cab services as compared to other Age groups.

In [65]:

```
df.groupby([df['Date'].dt.year,'Age group'])['Profit'].count()
```

Out[65]:

```
Date   Age group
2016   10 - 19       6606
       20 - 29      34704
       30 - 39      33884
       40 - 49      13885
       50 - 59      11445
       60 - 69       6795
2017   10 - 19       8208
       20 - 29      41717
       30 - 39      40004
       40 - 49      16899
       50 - 59      13554
       60 - 69       8128
2018   10 - 19       7623
       20 - 29      40009
       30 - 39      38847
       40 - 49      16233
       50 - 59      13088
       60 - 69       7763
Name: Profit, dtype: int64
```

# Some Findings based on above Analysis:

1) Rides to which city made more Profit?

-New York NY

2)Which City is in high demand for cab users?

-New York NY

3)Which Company has more Users?

-Yellow Cab with 302,149

4)Which City has highest Population?

-New York Ny with 8,405,837

# Hypothesis Testing: Payment Mode and Profit-Yellow Cab

## - Hypothesis Testing based on Payment Mode and Profit

In [67]:

```
cash = df[(df['Payment_Mode']=='Cash')&(df.Company=='Yellow Cab')].groupby('Transaction ID').Profit.mean()
card = df[(df['Payment_Mode']=='Card')&(df.Company=='Yellow Cab')].groupby('Transaction ID').Profit.mean()
print(cash.shape[0],card.shape[0])
_, p_value = stats.ttest_ind(cash.values,card.values,equal_var=True)

print('P value is ', p_value)

if(p_value<0.05):  # alpha value is 0.05 or 5%
    print("we are rejecting null hypothesis and it says that there is a difference regarding Payment Mode and Yellow Cab")
else:
    print("We are accepting null hypothesis that there is no difference noted regarding Payment Mode and Yellow Cab")
```

```
109896 164785
P value is  0.2933060638298729
We are accepting null hypothesis that there is no difference noted regarding Payment Mode and Yellow Cab
```

# Hypothesis Testing: Payment Mode and Profit-Pink Cab

In [68]:

```
cash = df[(df['Payment_Mode']=='Cash')&(df.Company=='Pink Cab')].groupby('Transaction ID').Profit.mean()
card = df[(df['Payment_Mode']=='Card')&(df.Company=='Pink Cab')].groupby('Transaction ID').Profit.mean()
print(cash.shape[0],card.shape[0])
_, p_value = stats.ttest_ind(cash.values,card.values,equal_var=True)

print('P value is ', p_value)

if(p_value<0.05):  # alpha value is 0.05 or 5%
    print("we are rejecting null hypothesis and it says that there is a difference regarding Payment Mode and Pink Cab")
else:
    print("We are accepting null hypothesis that there is no difference noted regarding Payment Mode and Pink Cab")
```

```
33992 50719
P value is  0.7900465828793288
We are accepting null hypothesis that there is no difference noted regarding Payment Mode and Pink Cab
```

# Hypothesis Testing: Age and Profit-Yellow Cab

## - Hypothesis Testing based on Age and Profit

In [69]:

```
x = df[(df.Age <= 40)&(df.Company=='Yellow Cab')].groupby('Transaction ID').Profit.mean()
y = df[(df.Age >= 40)&(df.Company=='Yellow Cab')].groupby('Transaction ID').Profit.mean()
print(x.shape[0],y.shape[0])
_, p_value = stats.ttest_ind(x.values,y.values,equal_var=True)

print('P value is ', p_value)

if(p_value<0.05):  # alpha value is 0.05 or 5%
    print("we are rejecting null hypothesis and it says that there is a difference regarding Age and Yellow Cab")
else:
    print("We are accepting null hypothesis that there is no difference noted regarding Age and Yellow Cab")
```

```
201029 82454
P value is  0.44246196729249976
We are accepting null hypothesis that there is no difference noted regarding Age and Yellow Cab
```

# Hypothesis Testing: Age and Profit-Pink Cab

In [70]:

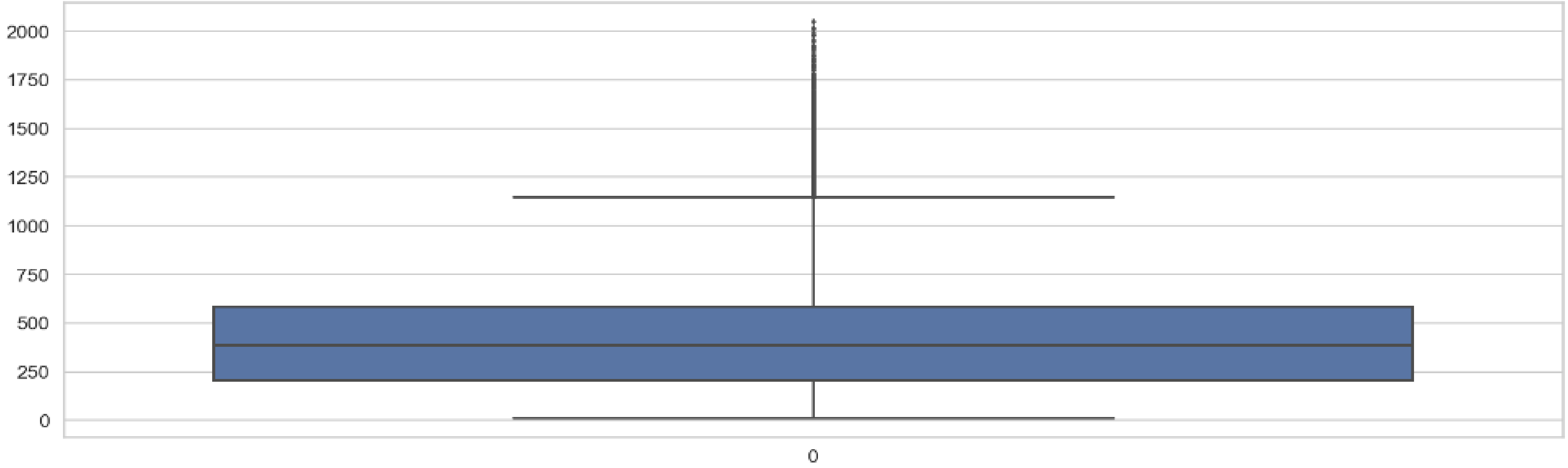
```
x = df[(df.Age <= 40)&(df.Company=='Pink Cab')].groupby('Transaction ID').Profit.mean()
y = df[(df.Age >= 40)&(df.Company=='Pink Cab')].groupby('Transaction ID').Profit.mean()
print(x.shape[0],y.shape[0])
_, p_value = stats.ttest_ind(x.values,y.values,equal_var=True)

print('P value is ', p_value)

if(p_value<0.05):  # alpha value is 0.05 or 5%
    print("we are rejecting null hypothesis and it says that there is a difference regarding Age and Pink Cab")
else:
    print("We are accepting null hypothesis that there is no difference noted regarding Age and Pink Cab")
```

```
62109 25336
P value is  0.09093510590632374
We are accepting null hypothesis that there is no difference noted regarding Age and Pink Cab
```

# Outliers detected and solved : Price Charged

`Outlier Detection on Price Charged.` - **Noticed there are Outliers**

In [72]:
```
plt.figure(figsize=(30,10), facecolor='w')
sns.boxplot(data=df['Price Charged'])
plt.show()
```



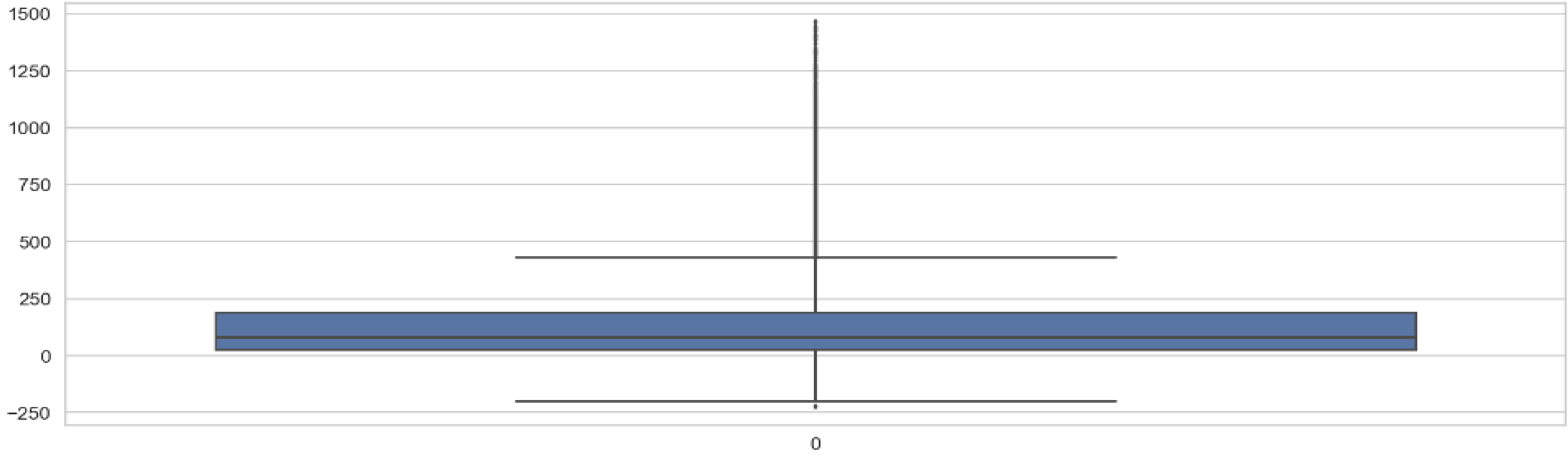

In [73]:
```
len(np.where(df['Price Charged']>1200)[0])/len(df)
```

Out[73]: 0.012139947466832873

# Outliers detected and solved : Profit

`Outlier Detection on Profit.` - Noticed there are Outliers

In [74]:
```
plt.figure(figsize=(30,10), facecolor='w')
sns.boxplot(data=df['Profit'])
plt.show()
```




In [75]:
```
len(np.where(df['Profit']>450)[0])/len(df)
```

Out[75]:
```
0.059261196687739 29
```

# Outliers detected and solved : Profit and Price Charged

In [79]:
```
df['z_profit']=np.abs(stats.zscore(df.Profit))
print("Removing Outliers: ", len(df[df['z_profit']>3]))
```

```
Removing Outliers:  7244
```

In [80]:
```
df = df[df['z_profit']<3]
```

In [81]:
```
df['z_price_charged']=np.abs(stats.zscore(df['Price Charged']))
print("Removing Outliers: ", len(df[df['z_price_charged']>3]))
```

```
Removing Outliers:  579
```

In [82]:
```
df = df[df['z_price_charged']<3]
```

# Label Encoding done before applying Model

## Label Encoding

In [87]:
```
from sklearn.preprocessing import LabelEncoder
lb= LabelEncoder()
```

In [88]:
```
lb.fit(data['Company'])
```

Out[88]:
```
LabelEncoder()
```

In [89]:
```
classes= list(lb.classes_)
```

In [90]:
```
classes
```

Out[90]:
```
['Pink Cab', 'Yellow Cab']
```

In [91]:
```
data['companies']= lb.fit_transform(data['Company'])
```

In [92]:
```
data.head(5)
```

Out[92]:

| | Company | City | KM Travelled | Price Charged | Cost of Trip | Profit | Payment_Mode | Gender | Income (USD/Month) | Population | Users | Age | companies |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Pink Cab | ATLANTA GA | 30.45 | 370.95 | 313.6350 | 57.3150 | Card | Male | 10813 | 814,885 | 24,701 | 28 | 0 |
| 1 | Yellow Cab | ATLANTA GA | 26.19 | 598.70 | 317.4228 | 281.2772 | Cash | Male | 10813 | 814,885 | 24,701 | 28 | 1 |
| 2 | Yellow Cab | ATLANTA GA | 42.55 | 792.05 | 597.4020 | 194.6480 | Card | Male | 10813 | 814,885 | 24,701 | 28 | 1 |
| 3 | Pink Cab | ATLANTA GA | 28.62 | 358.52 | 334.8540 | 23.6660 | Card | Male | 9237 | 814,885 | 24,701 | 27 | 0 |
| 4 | Yellow Cab | ATLANTA GA | 36.38 | 721.10 | 467.1192 | 253.9808 | Card | Male | 9237 | 814,885 | 24,701 | 27 | 1 |

0 is for Pink Cab and 1 is for Yellow Cab

# Steps before applying Model

Dropping Features ["Gender","Payment_Mode","City"] so that I have only numerical values for modeling

In [95]:
```
data.drop(["Gender","Payment_Mode","City",], axis=1, inplace=True)
```

In [96]:
```
data.head()
```

Out[96]:

| | Company | KM Travelled | Price Charged | Cost of Trip | Profit | Income (USD/Month) | Population | Users | Age | companies |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Pink Cab | 30.45 | 370.95 | 313.6350 | 57.3150 | 10813 | 814,885 | 24,701 | 28 | 0 |
| 1 | Yellow Cab | 26.19 | 598.70 | 317.4228 | 281.2772 | 10813 | 814,885 | 24,701 | 28 | 1 |
| 2 | Yellow Cab | 42.55 | 792.05 | 597.4020 | 194.6480 | 10813 | 814,885 | 24,701 | 28 | 1 |
| 3 | Pink Cab | 28.62 | 358.52 | 334.8540 | 23.6660 | 9237 | 814,885 | 24,701 | 27 | 0 |
| 4 | Yellow Cab | 36.38 | 721.10 | 467.1192 | 253.9808 | 9237 | 814,885 | 24,701 | 27 | 1 |

# Steps before applying Model

## String values noted while Normalizing

In [99]:

```
data.Population = data.Population.str.replace(',','').astype(float)
data.Users = data.Users.str.replace(',','').astype(float) # string values noted while normalizing
```

## Splitting Data into Training and Test Set

In [100…

```
X = data[['KM Travelled', 'Price Charged', 'Cost of Trip',
        'Profit','Age', 'Income (USD/Month)',
        'Population', 'Users', ]]
y = data['Profit']
```

# Steps before applying Model

## Normalizing

In [106…

```
scaler = MinMaxScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

In [107…

```
from sklearn import datasets, linear_model, metrics
reg = linear_model.LinearRegression()

# train the model using the training sets
reg.fit(X_train, y_train)

# regression coefficients
print('Coefficients: ', reg.coef_)
```

```
Coefficients:  [ 6.20007014e-12  1.58726582e+03 -9.37039551e+02 -3.21844275e+02
  6.08291195e-13  3.12638804e-13 -3.41060513e-13  7.10542736e-14]
```

# Steps before applying Model

## Cross Validation

In [108…

```
def get_cross_val(model, X_train, y_train, X_valid, y_valid):

    # Fit on train, predict on validation
    clf = model
    clf.fit(X_train, y_train)
    y_pred = clf.predict(X_valid)

    # Cross validation score over 10 folds
    scores = cross_val_score(clf, X_train, y_train, cv=10)
    print("Cross validation over 10 folds: ", sum(scores)/10.0)

    return y_pred
```

# Applying Model

## Applying Linear Regression Model

In [109…

```
Model = 'Linear Regression'
lin_predicted = reg.predict(X_test)
lin_acc_score = reg.score(X_train, y_train)
print("LinearRegression:",lin_acc_score*100,'\n')
get_cross_val(reg, X_train, y_train, X_test, y_test)
```

```
LinearRegression: 100.0

Cross validation over 10 folds:  1.0
```

Out[109…

```
array([113.902 , 118.9316,  55.    , ...,  27.682 , 246.5196, 146.3204])
```

# Recommendation

From the data it is noted that the rides to New York have got more users which in turn have increased the income and profit of the firm.

It is well noted that Yellow cab is an outstanding performer as compared to Pink cab.

The demand for Yellow cab was higher as compared to Pink cab as it travelled longer kilometers as compared to pink cab and while taking the case of the availability for more cab was there even though many of yellow cab have gone for longer trips.

# G2M Case Study

Data Glacier
Your Deep Learning Partner

Thank You