Chicago’s Bike Sharing System Data Analysis

Bootcamp Data Science Project – Abdulaziz Abdullah Binlaksar

Project Objective

The objective of this project is leverage the power of Data science in order to understand the patterns and trends in the Chicago's bike sharing system in order to provide the stakeholders with actionable insights that could assist them in decision making and increase their profits

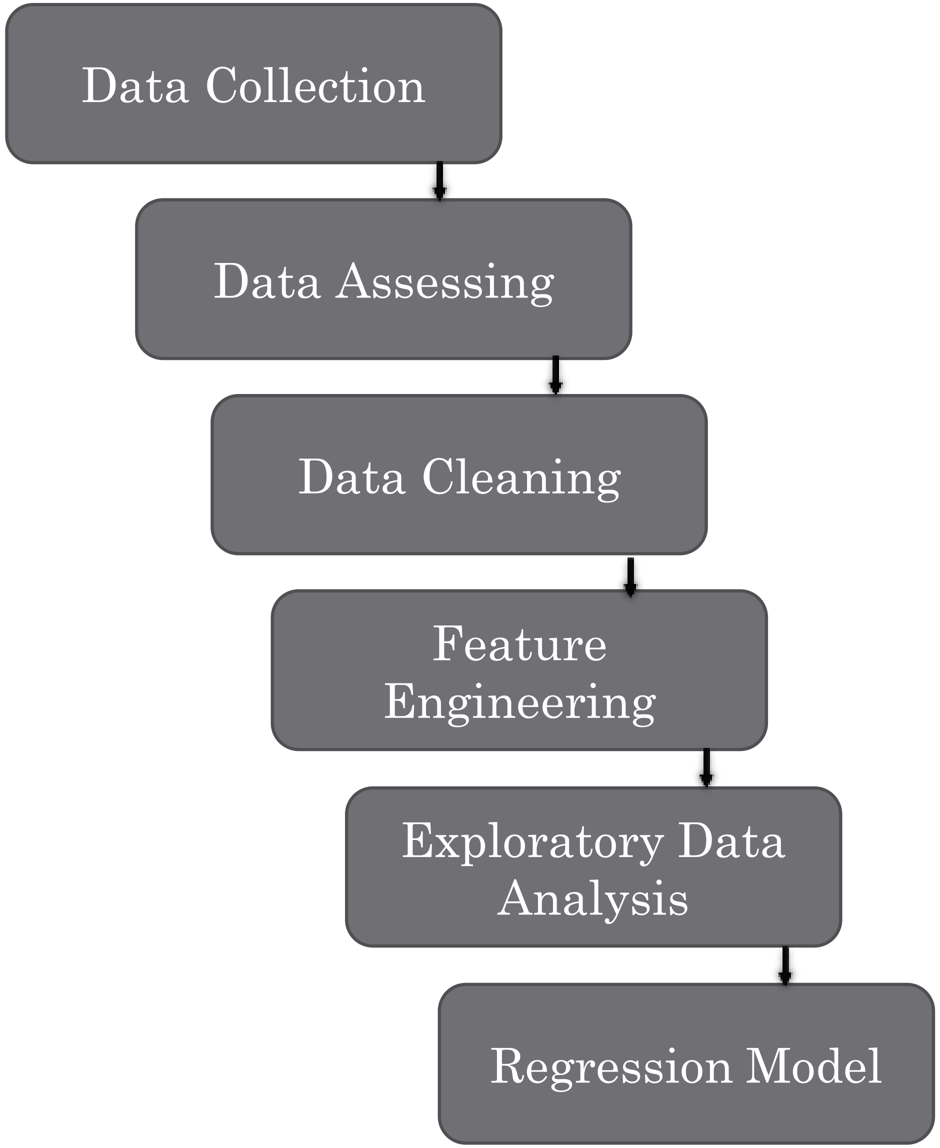
Data

The data is provided by savvy company on their website, I chose the data of the 4th quarter of 2019, which consists of 700000 row and 11 features

Data source:

https://divvy-tripdata.s3.amazonaws.com/index.html

Methodology



Data Collection

Data Assessing

Data Cleaning

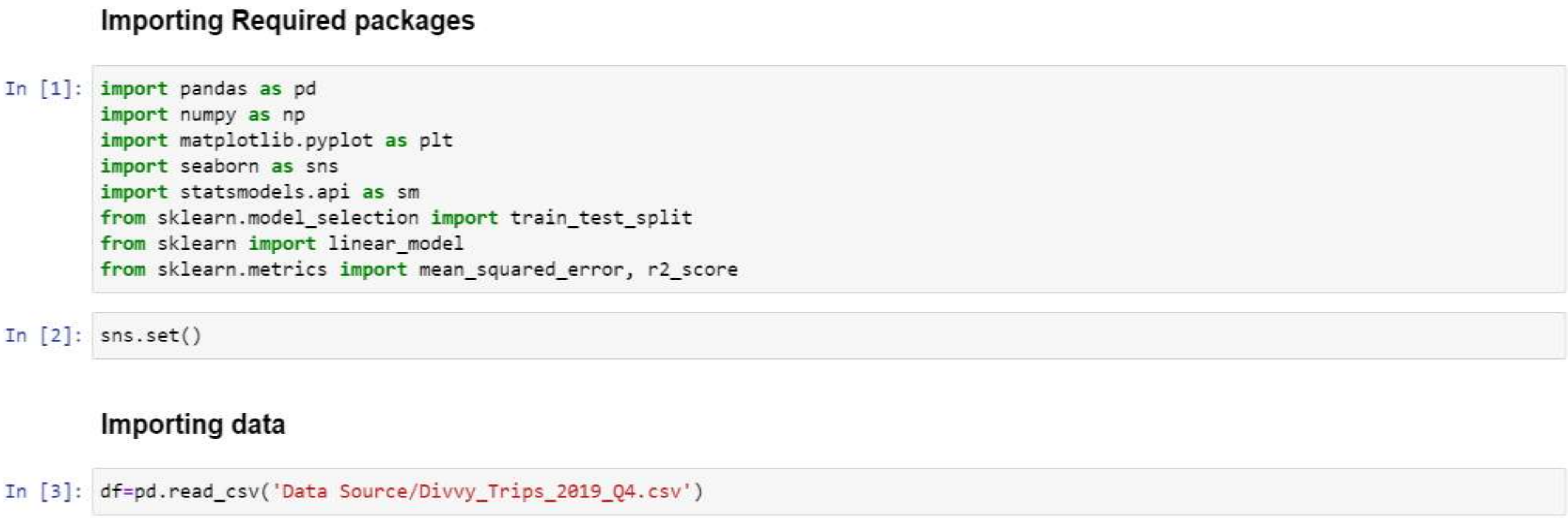
Feature Engineering

Exploratory Data Analysis

Regression Model

Data Collection

Data downloaded directly from the website of the company and it was loaded into data frame using pandas package in Jupyter notebook



Importing Required packages

```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
        import statsmodels.api as sm
        from sklearn.model_selection import train_test_split
        from sklearn import linear_model
        from sklearn.metrics import mean_squared_error, r2_score
```

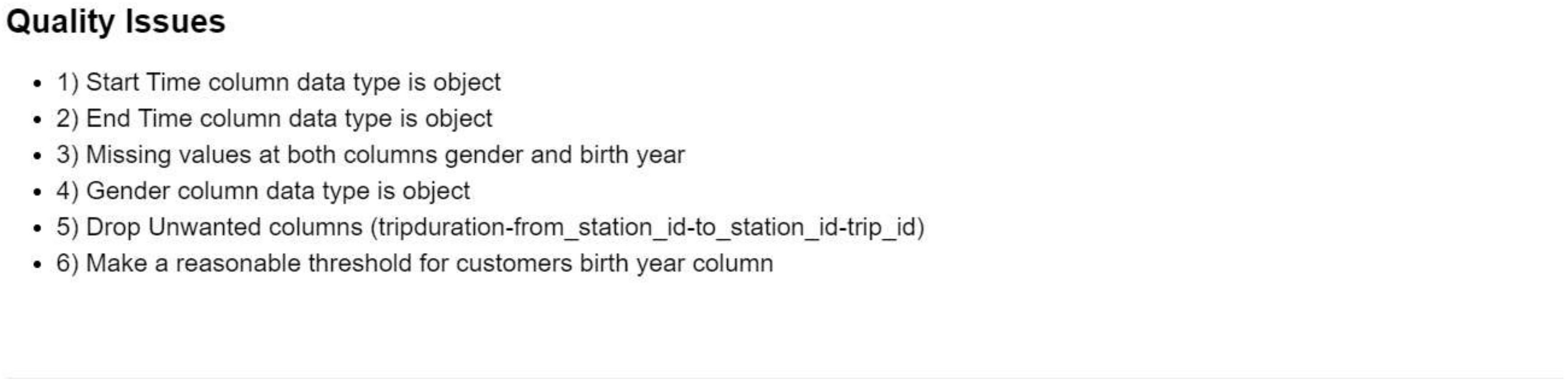
```
In [2]: sns.set()
```

Importing data

```
In [3]: df=pd.read_csv('Data Source/Divvy_Trips_2019_Q4.csv')
```

Data Assessing and Cleaning

The second step is to assess the data for any issue and to clean it before starting the analysis

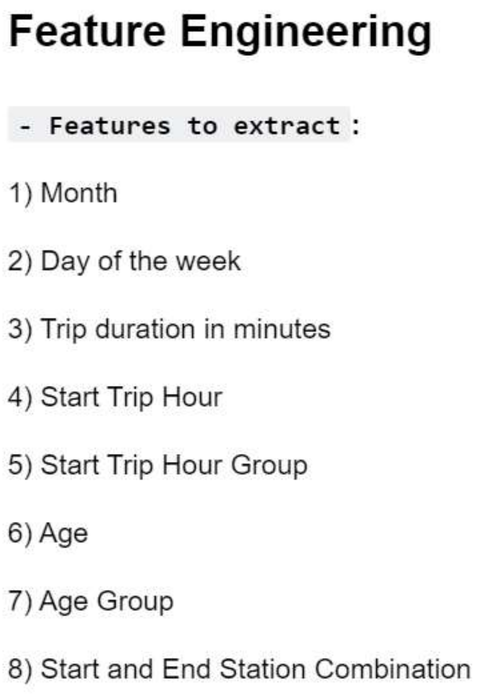


Quality Issues

- 1) Start Time column data type is object
- 2) End Time column data type is object
- 3) Missing values at both columns gender and birth year
- 4) Gender column data type is object
- 5) Drop Unwanted columns (tripduration-from_station_id-to_station_id-trip_id)
- 6) Make a reasonable threshold for customers birth year column

Feature Engineering

Feature Engineering is the process of creating new columns from existing columns in order to benefit our analysis



Feature Engineering

`- Features to extract`:

1) Month

2) Day of the week

3) Trip duration in minutes

4) Start Trip Hour

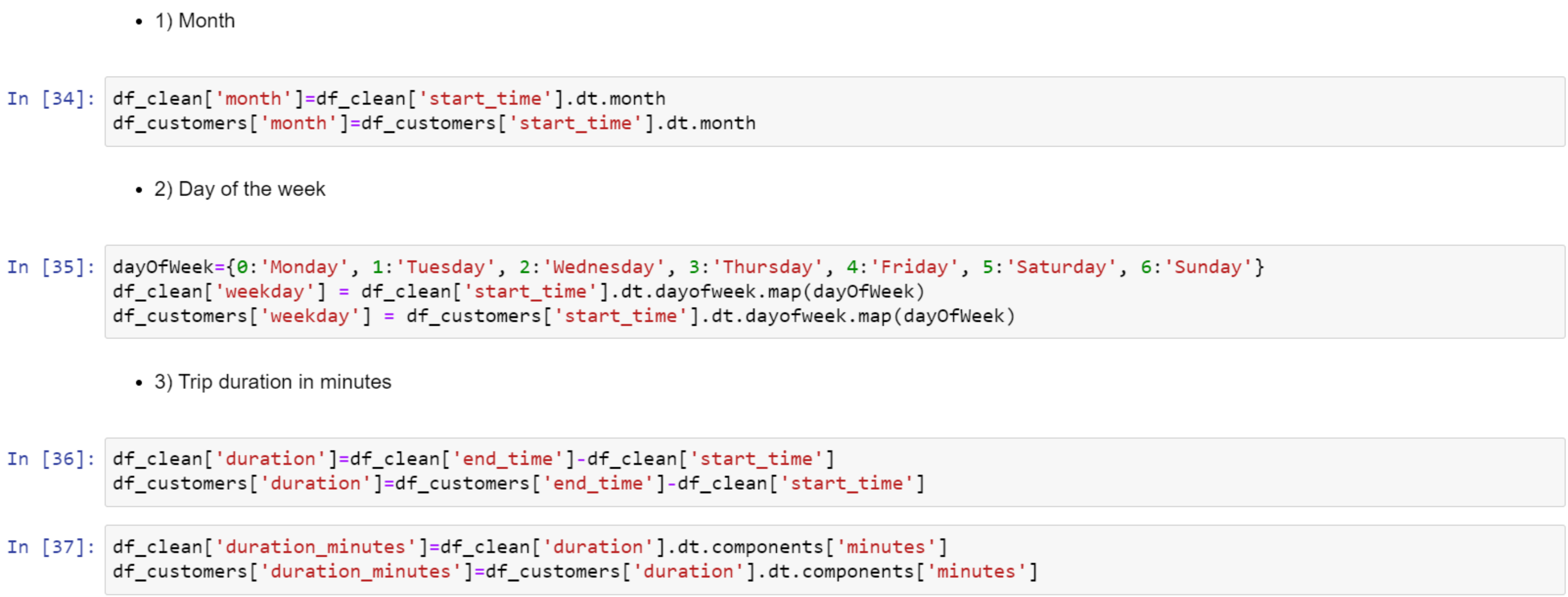
5) Start Trip Hour Group

6) Age

7) Age Group

8) Start and End Station Combination

Feature Engineering



- 1) Month

```
In [34]: df_clean['month']=df_clean['start_time'].dt.month
         df_customers['month']=df_customers['start_time'].dt.month
```

- 2) Day of the week

```
In [35]: dayOfWeek={0:'Monday', 1:'Tuesday', 2:'Wednesday', 3:'Thursday', 4:'Friday', 5:'Saturday', 6:'Sunday'}
         df_clean['weekday'] = df_clean['start_time'].dt.dayofweek.map(dayOfWeek)
         df_customers['weekday'] = df_customers['start_time'].dt.dayofweek.map(dayOfWeek)
```

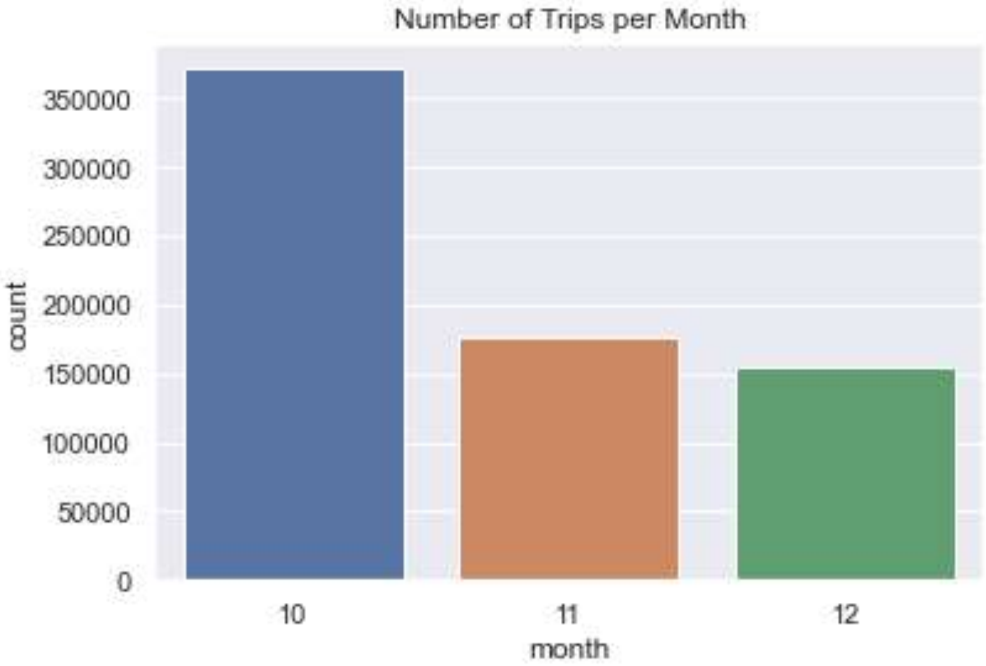
- 3) Trip duration in minutes

```
In [36]: df_clean['duration']=df_clean['end_time']-df_clean['start_time']
         df_customers['duration']=df_customers['end_time']-df_clean['start_time']
```

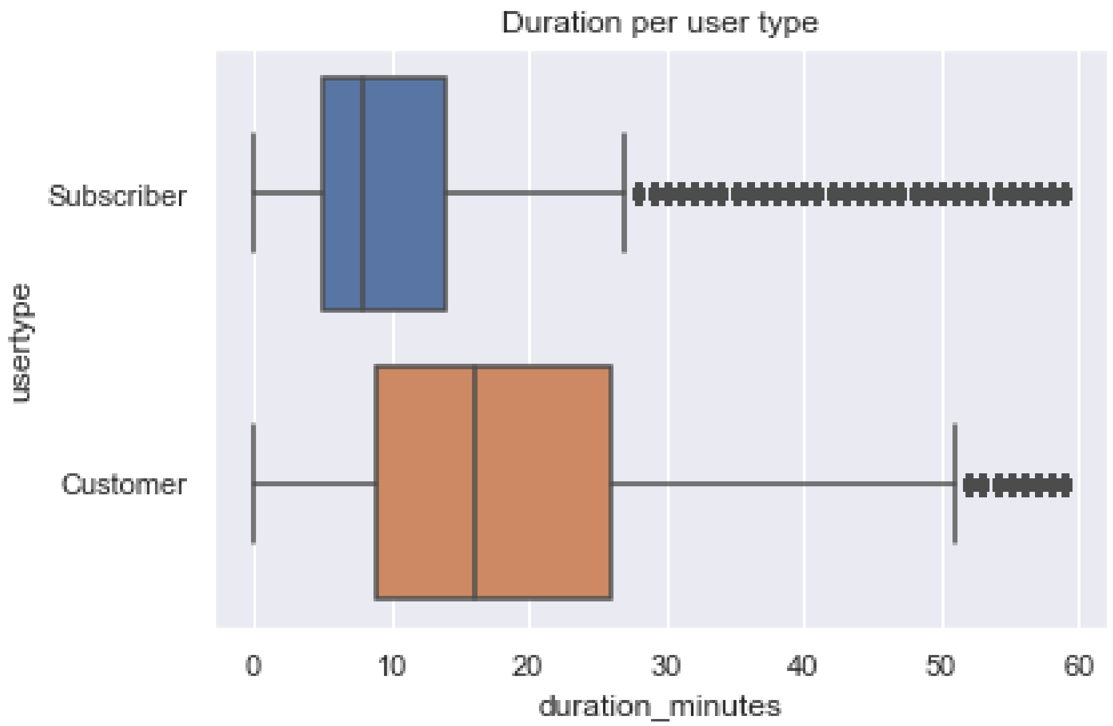
```
In [37]: df_clean['duration_minutes']=df_clean['duration'].dt.components['minutes']
         df_customers['duration_minutes']=df_customers['duration'].dt.components['minutes']
```

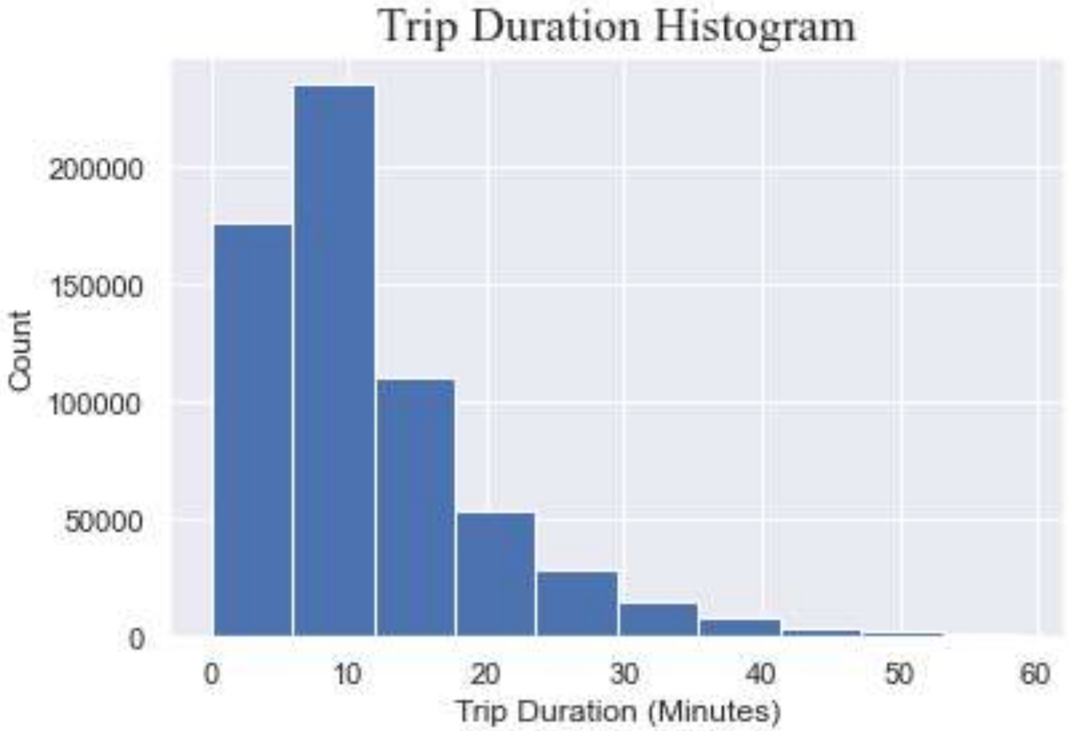
Exploratory Data Analysis

Here we search the data for any relationship or any trends or patterns, different views and visuals were used at this step in order to explore the data



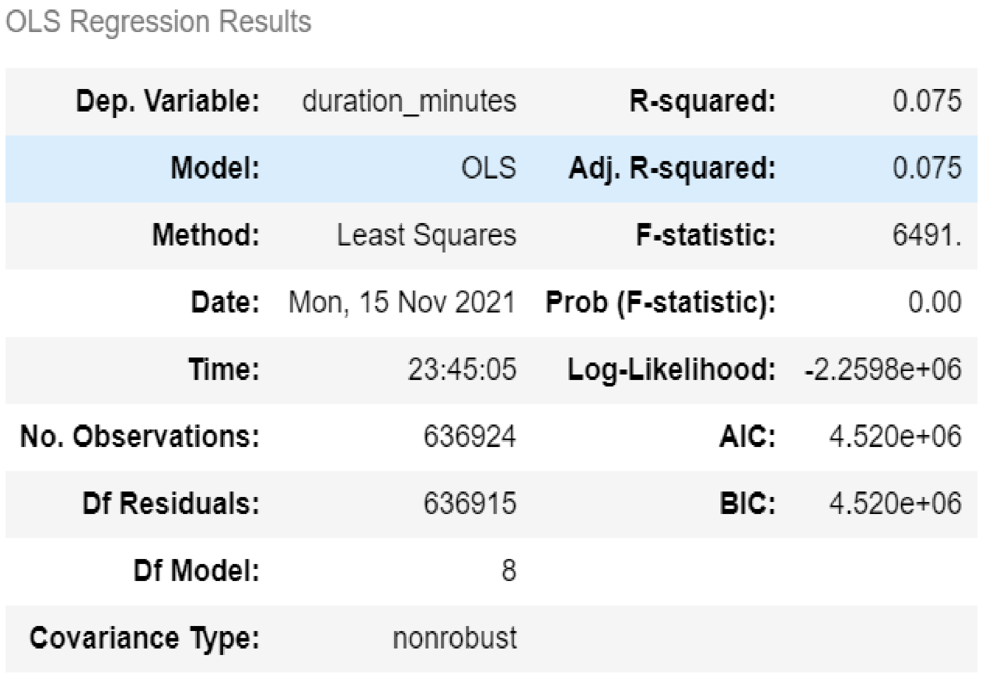
Exploratory Data Analysis





Regression

We tried to use the data we have in order to make a supervised machine learning model to predict the trip duration using the features we have, but we find that these features don’t explain the variability of the trip duration variable



OLS Regression Results

Dep. Variable:	duration_minutes	**R-squared:**	0.075
Model:	OLS	**Adj. R-squared:**	0.075
Method:	Least Squares	**F-statistic:**	6491.
Date:	Mon, 15 Nov 2021	**Prob (F-statistic):**	0.00
Time:	23:45:05	**Log-Likelihood:**	-2.2598e+06
No. Observations:	636924	**AIC:**	4.520e+06
Df Residuals:	636915	**BIC:**	4.520e+06
Df Model:	8		
Covariance Type:	nonrobust		