INSTITUTE FOR ADVANCEDCOMPUTING
AND SOFTWARE DEVELOPMENTAKURDI, PUNE

Documentation On

"Taxi Fare Prediction"

PG-DBDA MAY 2021

Submitted By:

Group No: 08

Mayuri Inde 1316

Snehal Kolikal 1351

Mr. Prashant Karhale
Centre Coordinator

Mr. Akshay Tilekar
Project Guide

Contents

List of Figures

1. INTRODUCTION

1.1 Problem Statement

Taxi Fare Prediction

1.2 Abstract

This research aims to study the predictive analysis, which is a method of analysis in Machine Learning. Many companies like Ola, Uber etc uses Artificial Intelligence and machine learning technologies to find the solution of accurate fare prediction problem. We are proposing this paper after comparative analysis of algorithms like regression and classification, which are useful for prediction modeling to get the most accurate value. This research will be helpful to those, who are involved in fare forecasting. In previous era, the fare was only dependent on distance, but with the enhancement in technologies the cab's fare is dependent on a lot of factors like time, location, number of passengers, traffic, number of hours, base fare etc. The study is based on supervised learning whose one application is prediction, in machine learning.

1.3 Introduction

New York City taxi rides paint a vibrant picture of life in the city. The millions of rides taken each month can provide insight into traffic patterns, road blockage, or large-scale events that attract many New Yorkers. With ridesharing apps gaining popularity, it is increasingly important for taxi companies to provide visibility to their estimated fare and ride duration, since the competing apps provide these metrics upfront. Predicting fare and duration of a ride can help passengers decide when is the optimal time to start their commute, or help drivers decide which of two potential rides will be more profitable, for example. Furthermore, this visibility into fare will attract customers during times when ridesharing services are implementing surge pricing. In order to predict duration and fare, only data which would be available at the beginning of a ride was used. This includes pickup and drop-off coordinates, trip distance, start time, number of passengers, and a rate code detailing whether the standard rate or the airport rate was applied. Linear regression with model selection, lasso, and random forest models were used to predict duration and fare amount.

1.4 Aims & Objectives

- We will using Taxi data for the year 2020 consist of 10Lakh rows with 21 columns. It includes trip records from all trips completed in taxis in NYC in 2020. Records include fields capturing pick-up and drop-off dates/times, trip distances, rate id code, payment types, and driver-reported passenger counts.

- We are predicting fare amount by taking the variable like pick-up and drop-off dates/times, trip distances, rate id code, payment types, total amount and passenger counts.

- Machine learning is applied to predict Fare amount We have used Linear Regression, Random Forest, and Polynomial Regression models.

2. OVERALL DESCRIPTION

2.1 Workflow of Project

The diagram below shows the workflow of this project.

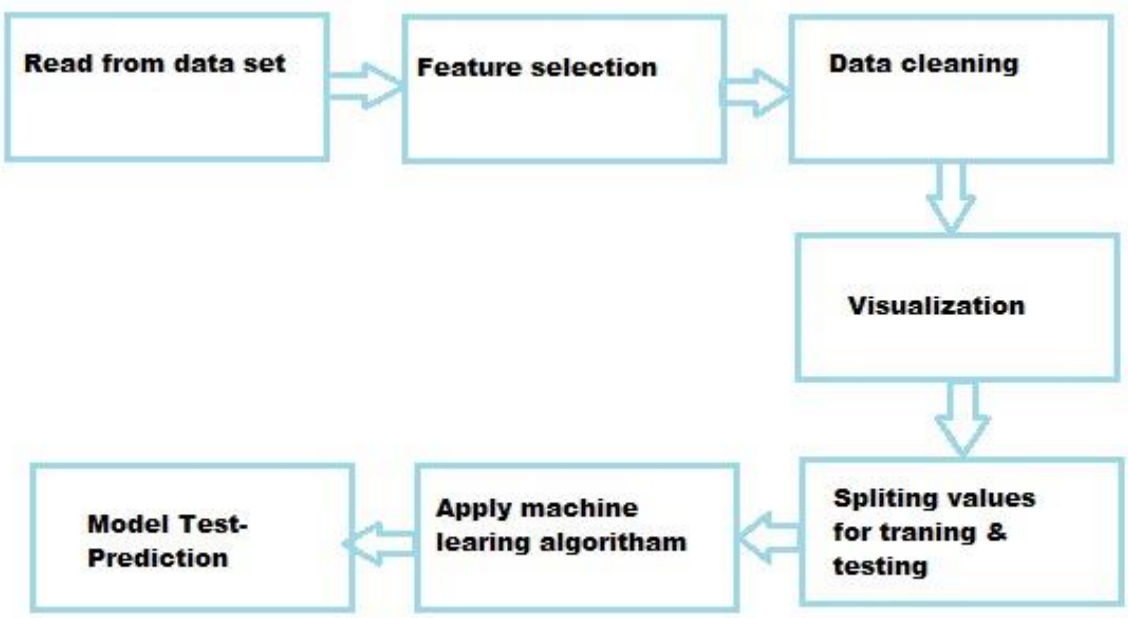

Figure 1 Workflow Diagram

2.2 Data Preprocessing and Cleaning

2.2.1 Data Cleaning

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the Data and then replacing, modifying, or deleting the dirty or coarse data.

2.2.2 Label encoding:

To make the data understandable or in human readable form, the training data is often labeled in words. Label Encoding refers to converting the labels into numeric form so as to convert it into the machine-readable form. Machine learning algorithms can then decide in a better way on how those labels must be operated.

2.2 Exploratory Data Analysis:

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with thehelp of summary statistics and graphical representations. Following are some plots we used to extract some useful information.

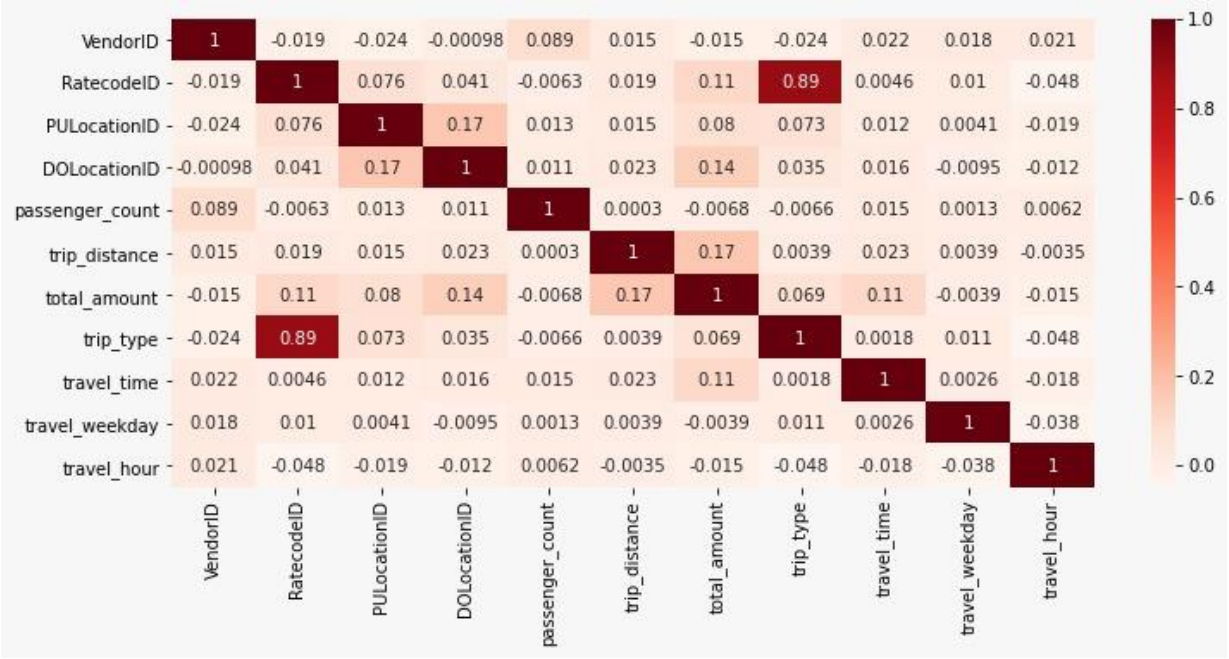

Figure 2 Correlation matrix

2.3.1 Visualization

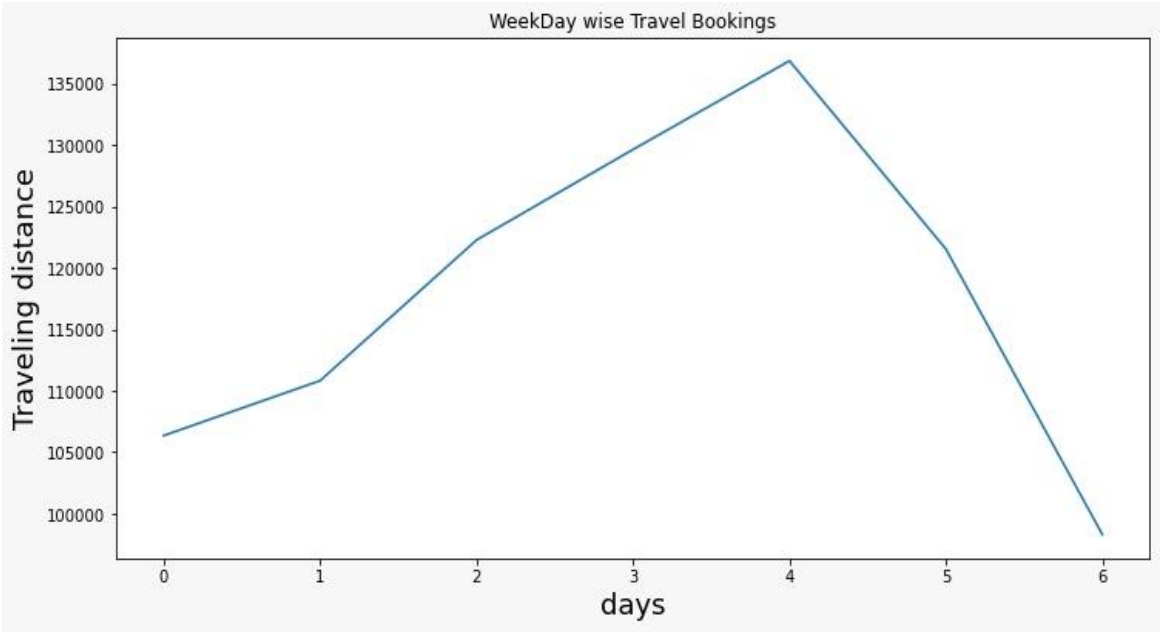

Figure 3 weekday wise Travel Booking

Exp: above graph show week day wise booking and we conclude that at Thursday peoples are travel more compare to other days.

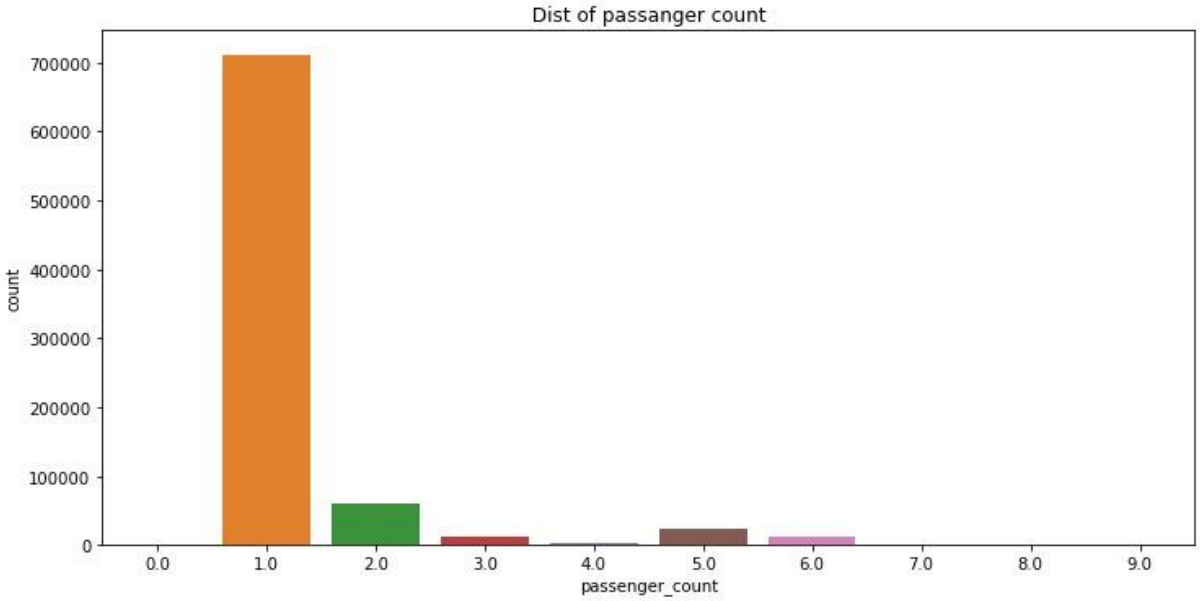

figure 3 Dist of passenger count.

Exp: passenger count graphed plot there are 0 to 9 people can travel in a taxi. From graph single person travels count is more

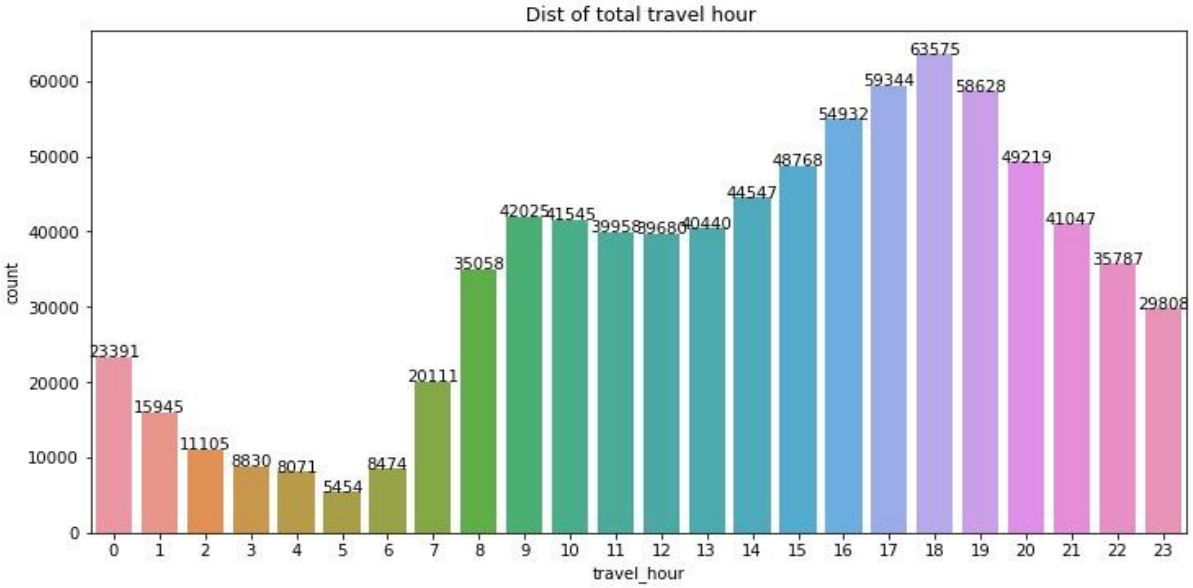

Figure 4 Dist of total travel hour.

Exp: From the above figure, we can conclude that the value of tip amount does have something to do with the hour of a day. We can see that 16:00-17:00 is the rush hour of a day. So, people probably give high tips because of heavy traffic

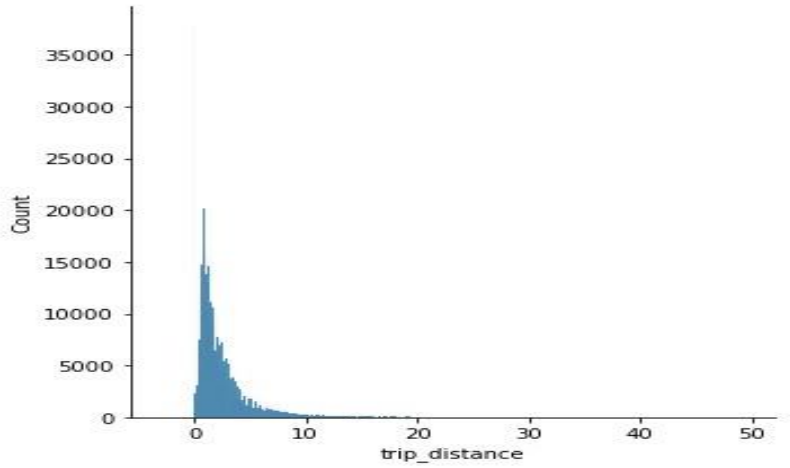

Figure 5 Trip distance and count

Exp: In reality, it makes sense that most taxi trips are short trips (around 1-2 miles). Because for long trips, it would be expensive to take a taxi and people probably would choose public transportations like metro or bus to save their money. While for short trips, People probably prefer to choose taxi to save their time.

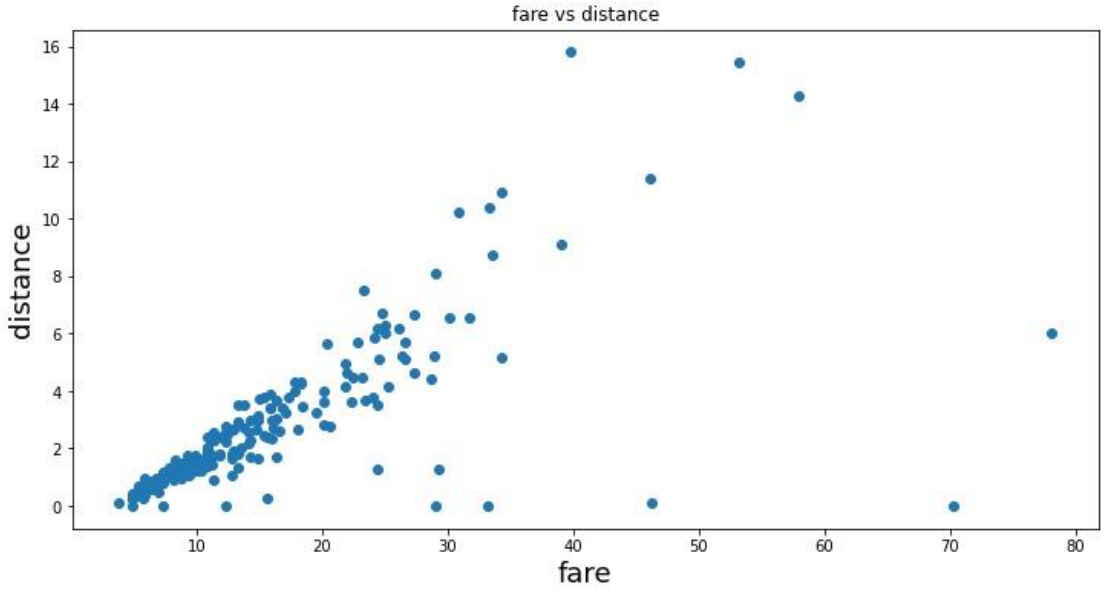

Figure 6 Fare vs distance

Exp: We predict the dependent variable using the known value of the independent variable. From graph as distance increases fare amount also increases. It is positive in natures.

2.1 Model Building

1. Train/Test split

One important aspect of all machine learning models is to determine their accuracy. Now, in order to determine their accuracy, one can train the model using the given dataset and then predict the response values for the same dataset using that modeland hence, find the accuracy of the model. A better option is to split our data into two parts. First one for training our machine learning model, and second one for testing our model.

- Split the dataset into two pieces: a training set and a testing set.
- Train the model on the training set.
- Test the model on the testing set, and evaluate how well our model did.

2. Advantages of train/test split

Machine learning consists of algorithms that can automate analytical model building. Using algorithms that iteratively learn from data, machine learning models facilitate computers to find hidden insights from Big Data without being explicitly programmed where to look. We have used the following three algorithms to build predictive model.

- Model can be trained and tested on different data than the one used for training.
- Response values are known for the test dataset, hence predictions can be evaluated
- Testing accuracy is a better estimate than training accuracy of out-of-sample performance.

1. Linear Regression

Linear regression is used to determine a mathematical relationship among a number of random variables. In other terms, MLR examines how multiple independent variables are related to one dependent variable.

```
from sklearn.linear_model import LinearRegression
lr=LinearRegression(normalize=True)
lr.fit(x_train,y_train)
y_predicted = lr.predict(x_test)
print("Linear Regression Score: ", lr.score(x_test,y_test))
err = mean_squared_error(y_test,y_predicted)
print('Mean Squared Err: ', err)
```

```
Linear Regression Score:  0.07357774616465473
Mean Squared Err:  126.8142601522355
```

2. Polynomial Regression

Polynomial Regression is a form of Linear regression known as a special case of Multiple linear regression which estimates the relationship as an nth degree polynomial. Polynomial Regression is sensitive to outliers so the presence of one or two outliers can also badly affect the performance

```
The Polynomial model performance for the training set
----------------------------------------
RMSE of training set is 5.431563990582172
R2 score of training set is 0.7847523177835105


The Polynomial model performance for the test set
----------------------------------------
RMSE of test set is 198.7856964871646
R2 score of test set is -287.6763131432656
```

3. Random Forest Regression

Random forests are one the most popular machine learning algorithms. Random forests consist of hundred decision trees which is default value, each of them built over a random extraction of the observations from the dataset and a random extraction of the features.

```
Random Forest Regressor Score:  0.07357774616465473
Random Forest Regressor RMSE:  4.369182593272292
```

5. Results and Finding

The polynomial Regression which gave us more R2 score and less Mean square error. So Polynomial model best for our system.

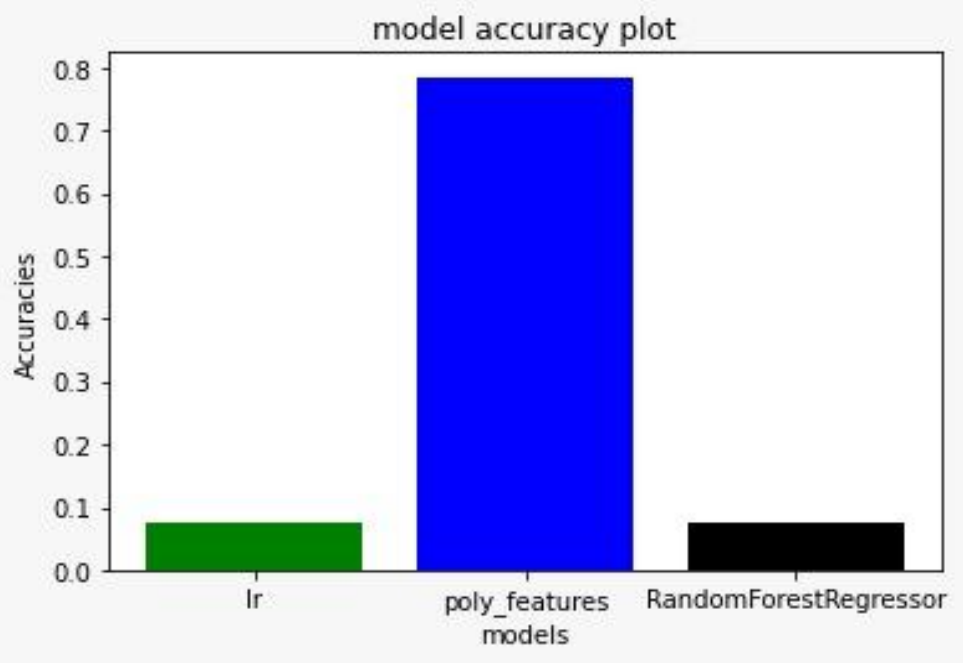

Figure 7 Model Accuracy plot

3. Requirements Specification

3.1 Hardware Requirement

- 500 GB hard drive (Minimum requirement)
- 8 GB RAM (Minimum requirement)
- PC x64-bit CPU

3.2 Software Requirement

Windows/Mac/Linux

Python-3.9.1

Anaconda/Spyder

Python Extension for VS Code

Libraries:

Numpy 1.18.2

Pandas 1.2.1

Matplotlib 3.3.3

Scikit-learn 0.24.1

Flask 1.1.2

3. Conclusion:

- We had successfully predict the Fare Amount using different Independent Variables, by using various Machine Learning algorithm.

- we concluded that the polynomial Regression which gave us more R2 score and less mean square error.

6. References

- https://www.scirp.org/pdf/AJOR_2019073016 16 0846.pdf

- Taxicab fact book http://www.nyc.gov/html/tlc/downloads/pdf/2014_taxicab_fact_book.pdf

- DATASET https://data.cityofnewyork.us/api/views/q5mz-t52e/rows.csv?accessType=DOWNLOAD