# EAS 501 PROJECT

# Chicago Taxi Trip Analysis


| NAMES | UBIT NAME |
|---|---|
| ADVAIT KULKARNI | advaitan |
| SUHIT DATTA | suhitdat |
| VARAD TUPE | varadsha |

## Introduction

- The data was extracted from Chicago Data Portal
- Main dataset contains total of 113 Million records
- The data set used for analysis is from April - July 2017

## Data Description

### Taxi Trip Data

| Column Name | Description | Type |
|---|---|---|
| Trip ID | A unique identifier for the trip. | Plain Text |
| Taxi ID | A unique identifier for the taxi. | Plain Text |
| Trip Start Timestamp | When the trip started, rounded to the nearest 15 minutes. | Date & Time |
| Trip End Timestamp | When the trip ended, rounded to the nearest 15 minutes. | Date & Time |
| Trip Seconds | Time of the trip in seconds. | Number |
| Trip Miles | Distance of the trip in miles. | Number |
| Pickup Census Tract | The Census Tract where the trip began. For privacy, this Census Tract is not shown for some trips. | Plain Text |
| Dropoff Census Tract | The Census Tract where the trip ended. For privacy, this Census Tract is not shown for some trips. | Plain Text |
| Pickup Community Area | The Community Area where the trip began. | Number |
| Dropoff Community Area | The Community Area where the trip ended. | Number |
| Fare | The fare for the trip. | Money |
| Tips | The tip for the trip. Cash tips generally will not be recorded. | Money |
| Tolls | The tolls for the trip. | Money |

| Column Name | Description | Type |
|---|---|---|
| Extras | Extra charges for the trip. | Money |
| Trip Total | Total cost of the trip, the total of the previous columns. | Money |
| Payment Type | Type of payment for the trip. | Plain Text |
| Company | The taxi company. | Plain Text |
| Pickup Centroid Latitude | The latitude of the center of the pickup census tract or the community area if the census tract has been hidden for privacy. | Number |
| Pickup Centroid Longitude | The longitude of the center of the pickup census tract or the community area if the census tract has been hidden for privacy. | Number |
| Dropoff Centroid Latitude | The latitude of the center of the dropoff census tract or the community area if the census tract has been hidden for privacy. | Number |
| Dropoff Centroid Longitude | The longitude of the center of the dropoff census tract or the community area if the census tract has been hidden for privacy. | Number |

After cleaning the data, the volume of the data has been reduced to 1.6 million records.

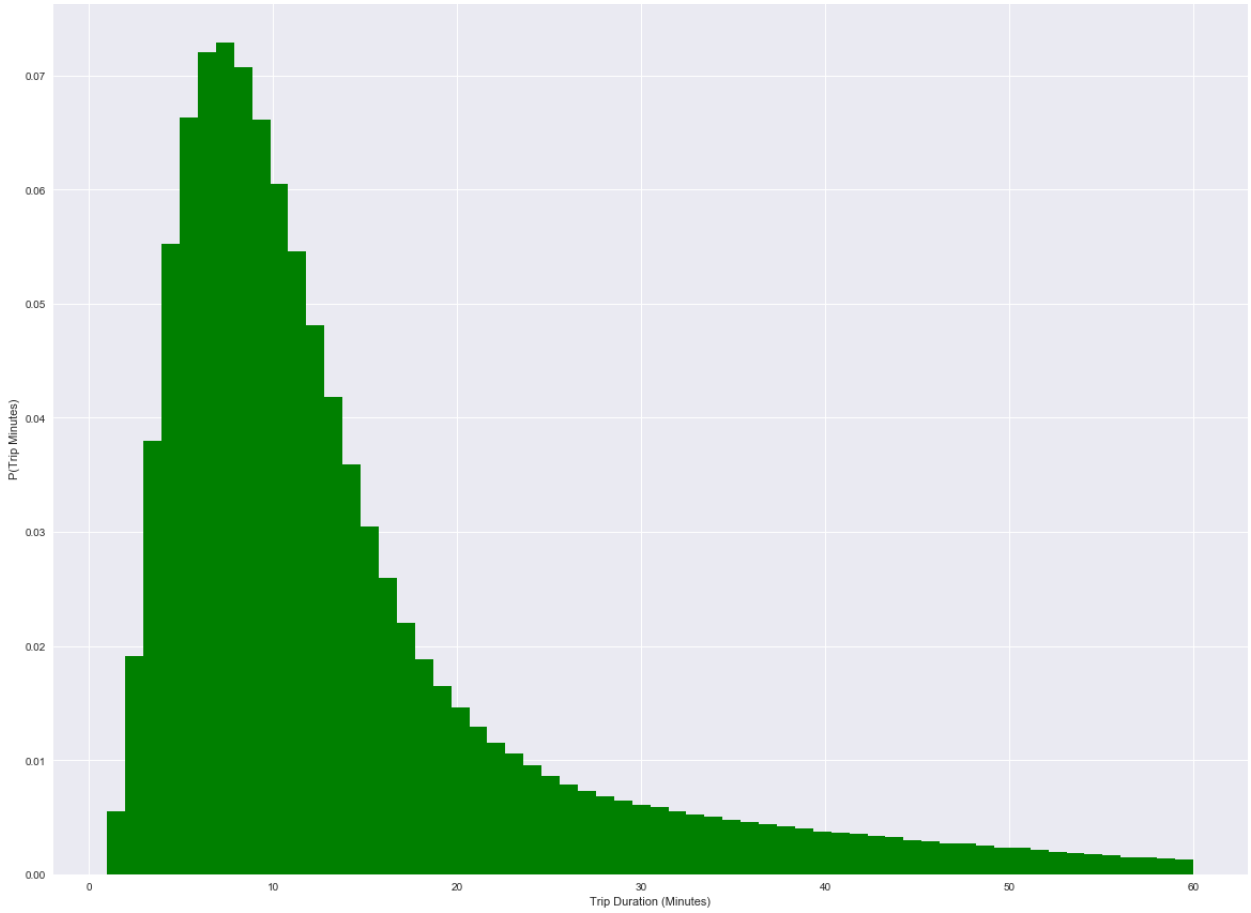
Assumption**:** The trip duration is a value between 1 and 60 minutes both inclusive.

**Q1) If you had to fit a probability distribution to the trip duration, how would you do it and why?**

In order to analyze the data, we extracted the data pertaining to the trip duration column and made a new table in MySQL. Then we extracted the data in Python using pandas library.

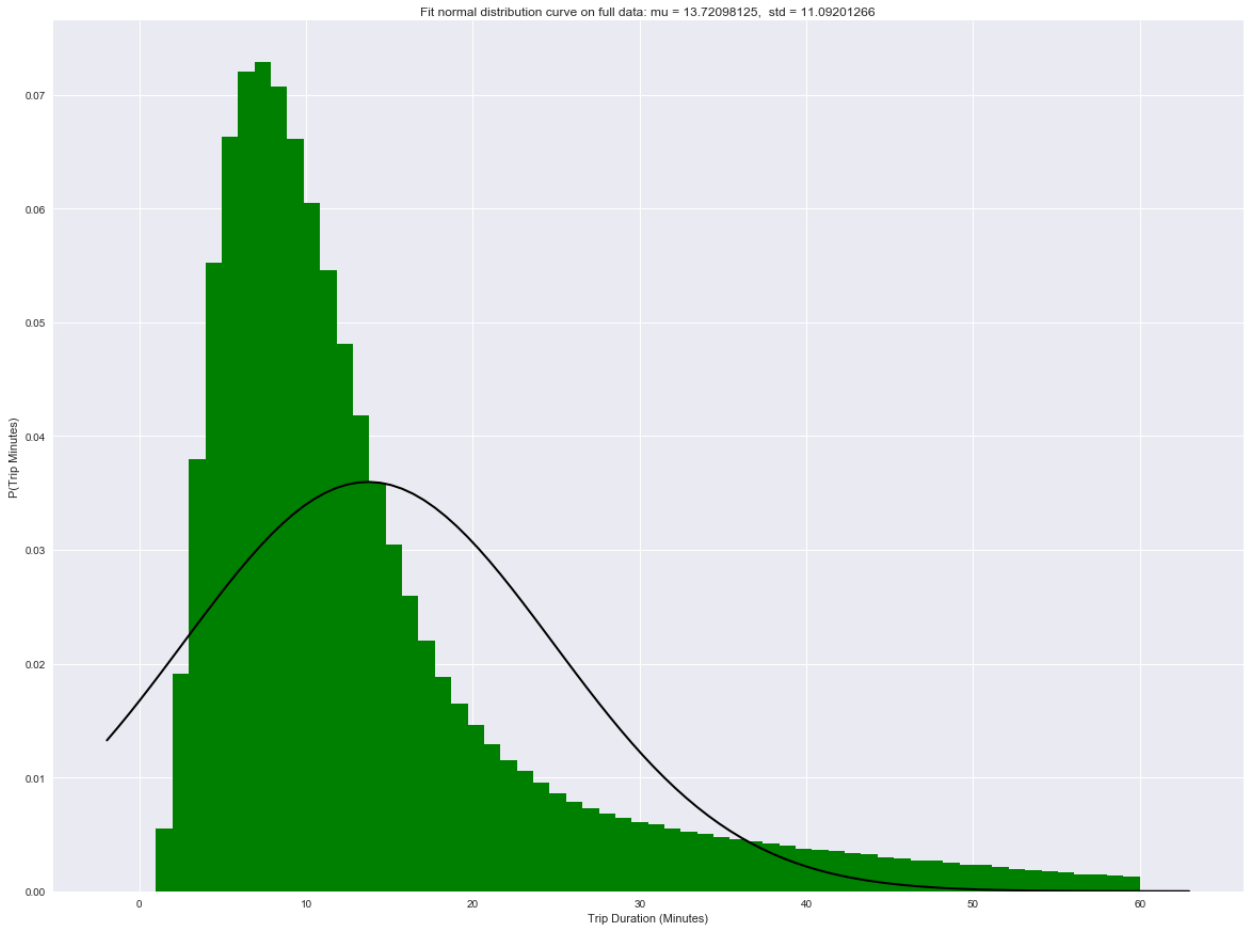
The data consists of Trip duration which is varying from 1 minutes in duration to 60 minutes.

The next step is to analyze the data. The best way to analyze the data is to plot a Histogram of the data points related to Trip Duration. The visual representation is as shown below:




The Histogram plot shows that there is a high bulge in the data during the Trip Duration values between 5 to 15 minutes. This kind of plot is strikingly similar to a Gaussian Distribution. The only difference in this case it is Right-Skewed that is it displays Positive Skewness.

We fit a Normal Curve on this plot and we obtain the following:




The values of Mean and Standard Deviation are as follows for the fitted curve.

Mean : 13.72098125

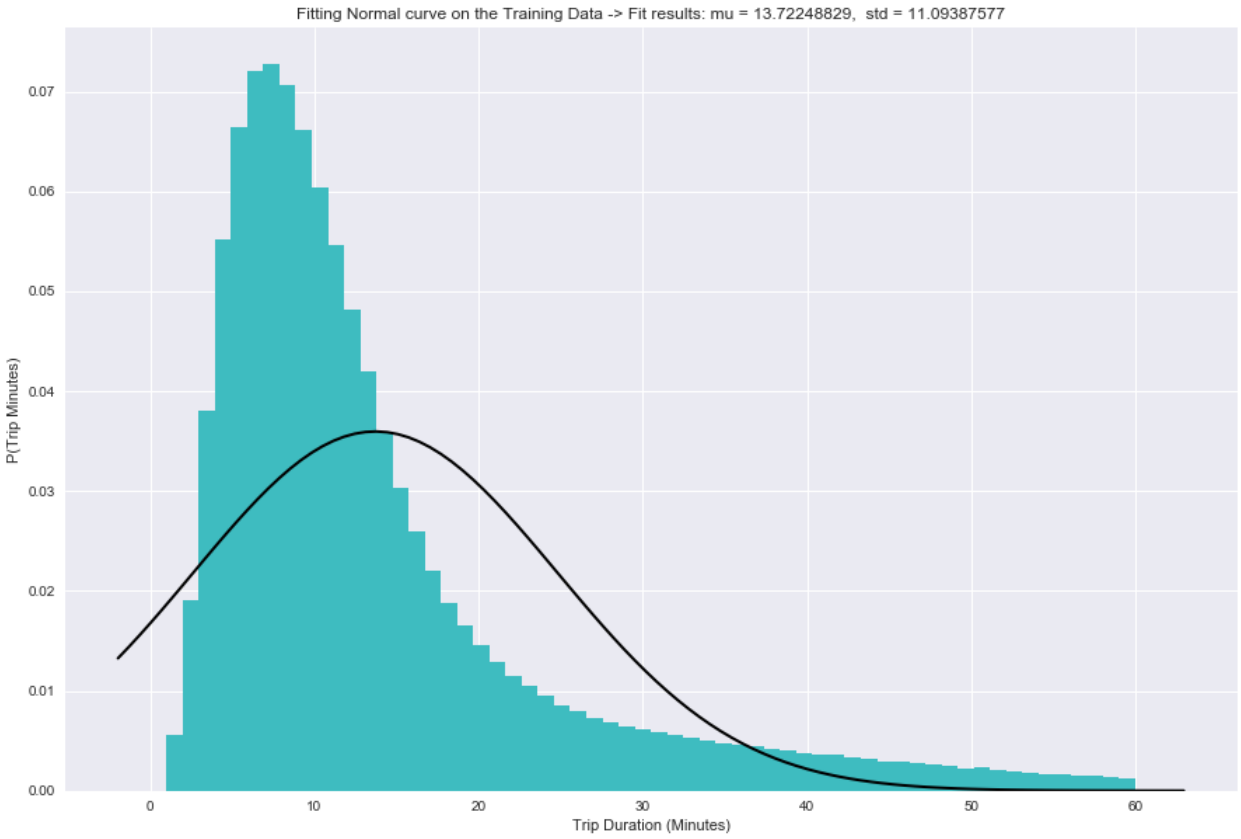
Standard Deviation : 11.09201266

Variance : 123.0327448496

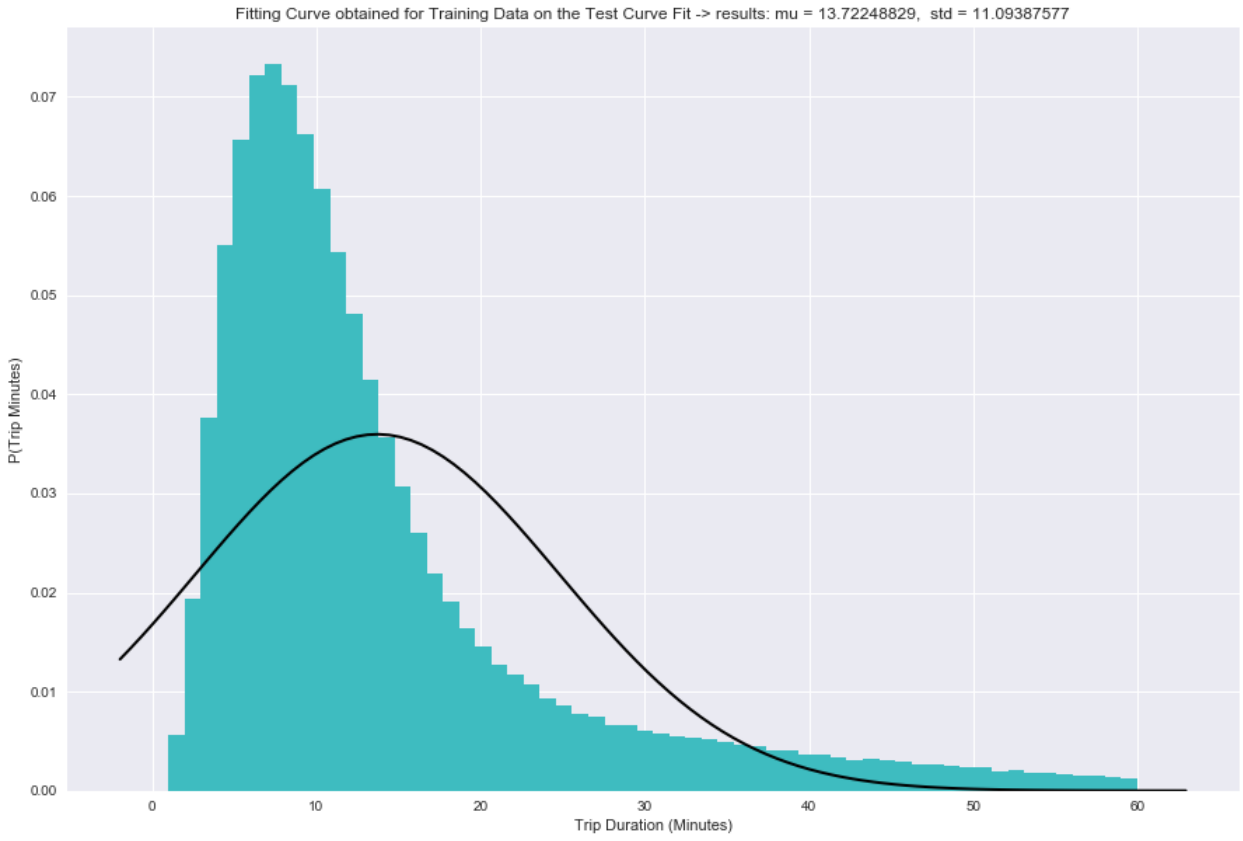
**Q2) Divide the data you are using into two parts (Training and Testing), and analyze it using the distribution you mentioned in Q1. Basically you have to learn the parameters from the training set and analyze the applicability of the fitted model on the testing set.**

The data obtained in the first part is divided into two parts: Training Set and the Test Set in the ratio 80% and 20%.

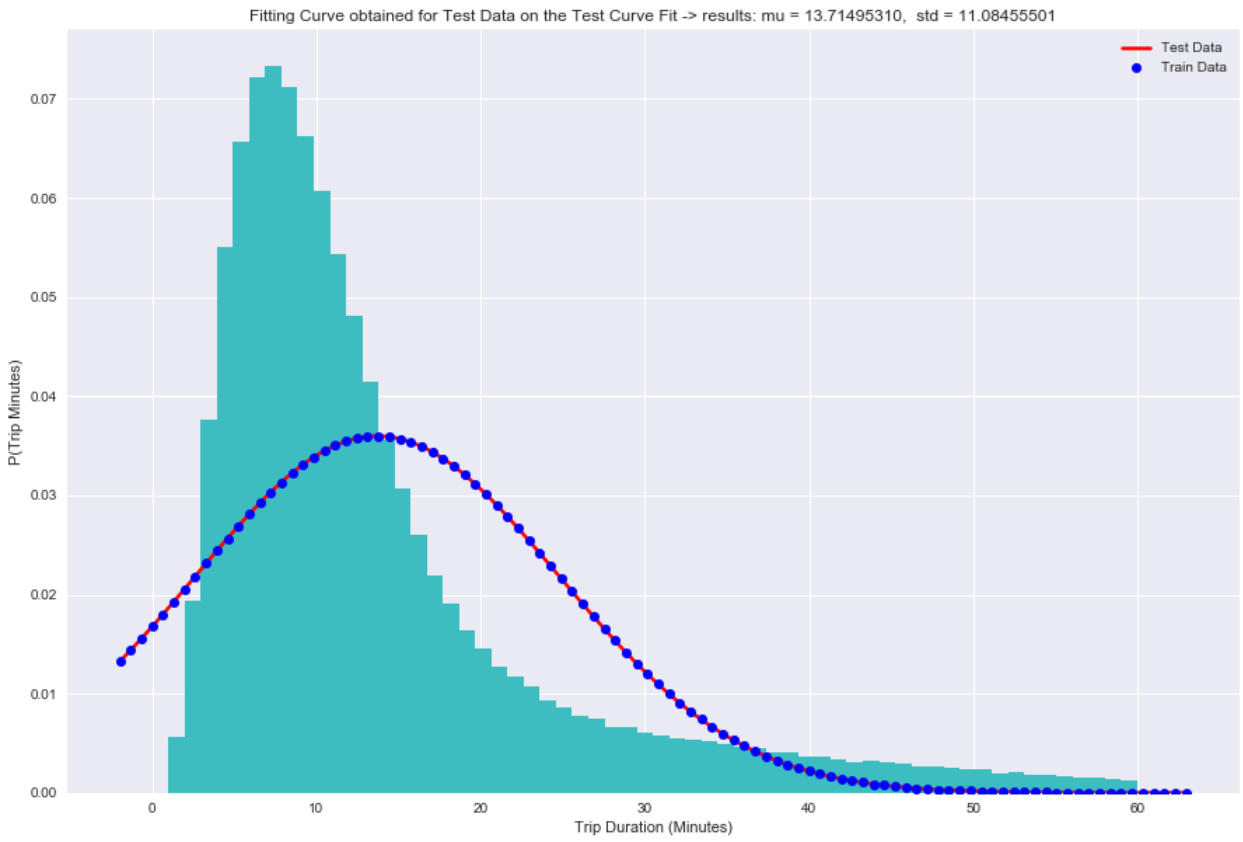
Following is the Histogram of the Training Data along with the fitted Normal Curve.



The Normal curve used to fit the Training Data is plotted along with the Test Data as follows:



In order to compare the fit of the training curve on the Test Data, the curve used to fit the Training Data and the Test Data are plotted in conjunction with the Test Data as follows :




The red line represents the Fit for the Test Data and the Blue Dotted Line represents the fit for the Training Data. Both these lines are plotted over the Histogram of the Test Data.

As can be seen in the figure, the fits for the Training Data and the Test Data are almost overlapping. This shows that the fit for the Training Data nearly fits the Test Data Set.

The following are the parameters of the Dataset:

| | **Mean** | **Standard Deviation** |
|---|---|---|
| Training Data (80% of Data) | 13.722488291950164 | 11.09387577373475 |
| Test Data (20% of Data) | 13.714953098683321 | 11.084555005296062 |

It can be seen that the Mean and the Standard Deviation parameters are very close to each other for the Training Data and the Test Data.

**Q3) Given you have to reach a certain location (any area code of your choice) at 2 pm what should be the estimated start time?**

For this problem, we assume that the area we have to reach has the code 8, as that area has the maximum drop-offs.

The aim is to find the time we have to reach there by 2 pm.

To find the estimated start time, we consider all the instances where the pickup was a certain area and drop-off area was one with code 8(near South Side). This area has one of the largest number of Drop-offs.

The process that we used is as follows:

Due to limited data for some areas, like those with Area Codes 9 and 10 which have very few instances (~ 2 or 3), the Average was considered.

Also Mode has been used for those area codes with a considerable number of rides to area 8.

The threshold value considered for the cutoff was at least 10 trips for the Mode.

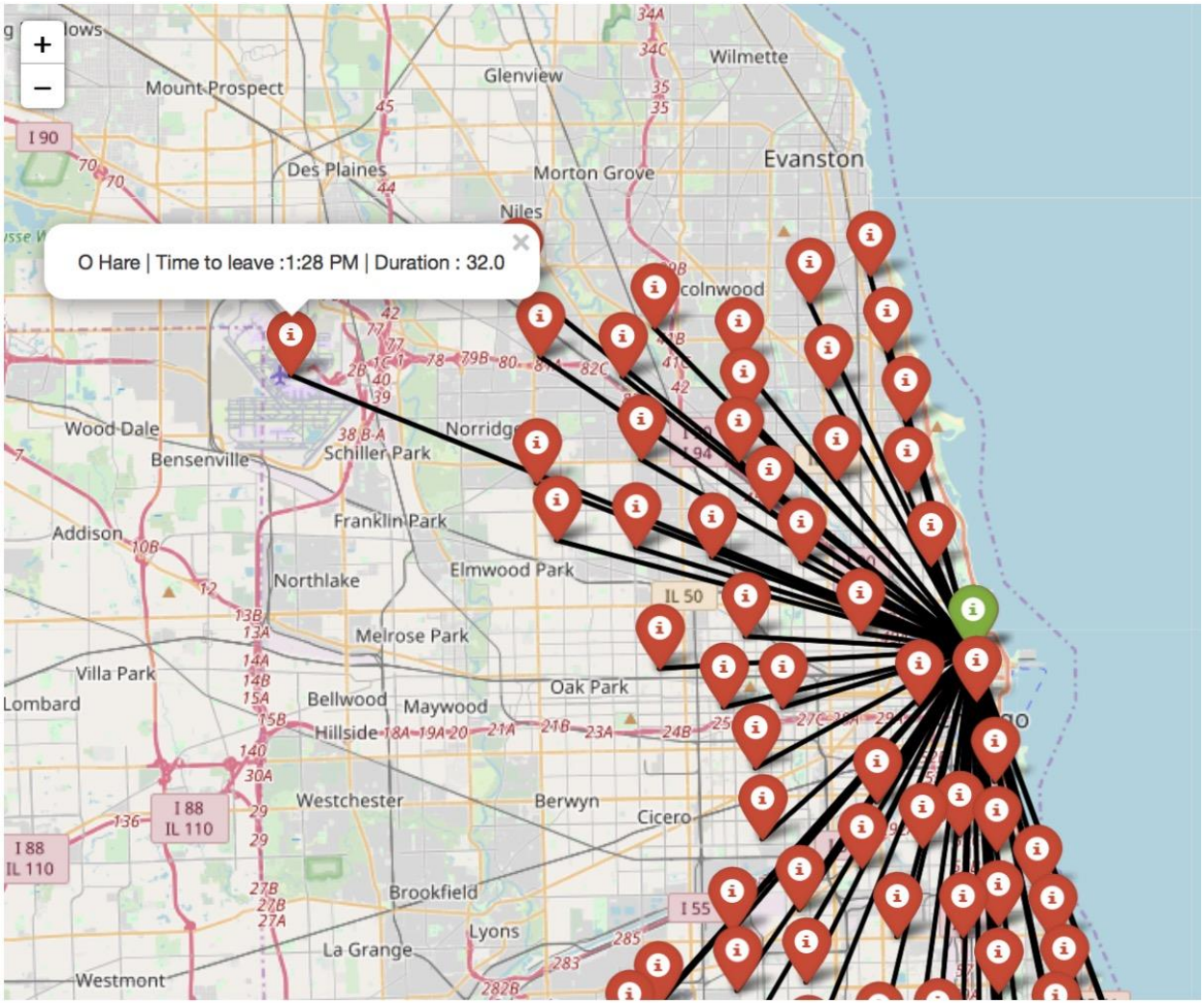
The following is the approximate Leaving Time for each Community Area to Near North Side (Community Area 8) :

| **Area** | **Community_Name** | **Leaving Time** | **Time to Reach (Min)** | **Method** | **Nrow/Freq** |
|---|---|---|---|---|---|
| 1 | Rogers Park | 1:41 PM | 19 | Mode | 21 |
| 2 | West Ridge | 1:34 PM | 26 | Average | 133 |
| 3 | Uptown | 1:49 PM | 11 | Mode | 218 |
| 4 | Lincoln Square | 1:39 PM | 21 | Mode | 22 |
| 5 | North Center | 1:43 PM | 17 | Mode | 38 |
| 6 | Lake View | 1:49 PM | 11 | Mode | 1400 |
| 7 | Lincoln Park | 1:51 PM | 9 | Mode | 1916 |
| 8 | Near North Side | 1:55 PM | 5 | Mode | 28085 |

| | | | | | |
|---|---|---|---|---|---|
| 9 | Edison Park | 1:30 PM | 30 | Average | 2 |
| 10 | Norwood Park | 1:34 PM | 26 | Average | 18 |
| 11 | Jefferson Park | 1:34 PM | 26 | Average | 39 |
| 12 | Forest Glen | 1:28 PM | 32 | Average | 10 |
| 13 | North Park | 1:32 PM | 28 | Average | 9 |
| 14 | Albany Park | 1:44 PM | 16 | Mode | 42 |
| 15 | Portage Park | 1:34 PM | 26 | Average | 28 |
| 16 | Irving Park | 1:42 PM | 18 | Mode | 12 |
| 17 | Dunning | 1:31 PM | 29 | Average | 4 |
| 18 | Montclare | 1:24 PM | 36 | Average | 14 |
| 19 | Belmont Cragin | 1:30 PM | 30 | Average | 16 |
| 20 | Hermosa | 1:35 PM | 25 | Average | 14 |
| 21 | Avondale | 1:47 PM | 13 | Mode | 11 |
| 22 | Logan Square | 1:49 PM | 11 | Mode | 120 |
| 23 | Humboldt Park | 1:41 PM | 19 | Average | 52 |
| 24 | West Town | 1:52 PM | 8 | Mode | 914 |
| 25 | Austin | 1:37 PM | 23 | Average | 3 |
| 26 | West Garfield Park | 1:45 PM | 15 | Average | 2 |
| 27 | East Garfield Park | 1:45 PM | 15 | Average | 9 |
| 28 | Near West Side | 1:50 PM | 10 | Mode | 5733 |
| 29 | North Lawndale | 1:38 PM | 22 | Average | 10 |
| 30 | South Lawndale | 1:35 PM | 25 | Average | 16 |
| 31 | Lower West Side | 1:42 PM | 18 | Mode | 11 |
| 32 | The Loop | 1:53 PM | 7 | Mode | 17207 |
| 33 | Near South Side | 1:45 PM | 15 | Mode | 1383 |
| 34 | Armour Square | 1:45 PM | 15 | Mode | 49 |
| 35 | Douglas | 1:41 PM | 19 | Mode | 24 |
| 36 | Oakland | 1:41 PM | 19 | Average | 19 |
| 37 | Fuller Park | 1:29 PM | 31 | Average | 8 |
| 38 | Grand Boulevard | 1:36 PM | 24 | Average | 26 |
| 39 | Kenwood | 1:44 PM | 16 | Mode | 38 |
| 40 | Washington Park | 1:33 PM | 27 | Average | 6 |
| 41 | Hyde Park | 1:42 PM | 18 | Mode | 101 |
| 42 | Woodlawn | 1:35 PM | 25 | Average | 14 |
| 43 | South Shore | 1:36 PM | 24 | Average | 38 |
| 44 | Chatham | 1:25 PM | 35 | Average | 11 |
| 45 | Avalon Park | 1:31 PM | 29 | Average | 4 |
| 46 | South Chicago | 1:27 PM | 33 | Average | 5 |
| 49 | Roseland | 1:25 PM | 35 | Average | 1 |
| 50 | Pullman | 1:35 PM | 25 | Average | 2 |
| 52 | East Side | 1:16 PM | 44 | Average | 1 |
| 53 | West Pullman | 1:25 PM | 35 | Average | 1 |
| 54 | Riverdale | 1:07 PM | 53 | Average | 1 |

| | | | | | |
|---|---|---|---|---|---|
| 56 | Garfield Ridge | 1:28 PM | 32 | Mode | 443 |
| 57 | Archer Heights | 1:35 PM | 25 | Average | 4 |
| 58 | Brighton Park | 1:33 PM | 27 | Average | 10 |
| 59 | McKinley Park | 1:39 PM | 21 | Average | 11 |
| 60 | Bridgeport | 1:39 PM | 21 | Average | 102 |
| 61 | New City | 1:32 PM | 28 | Average | 13 |
| 62 | West Elsdon | 1:27 PM | 33 | Average | 20 |
| 63 | Gage Park | 1:17 PM | 43 | Average | 3 |
| 64 | Clearing | 1:24 PM | 36 | Average | 9 |
| 65 | West Lawn | 1:24 PM | 36 | Average | 3 |
| 66 | Chicago Lawn | 1:25 PM | 35 | Average | 4 |
| 67 | West Englewood | 1:28 PM | 32 | Average | 1 |
| 68 | Englewood | 1:28 PM | 32 | Average | 10 |
| 69 | Greater Grand Crossing | 1:34 PM | 26 | Average | 9 |
| 71 | Auburn Gresham | 1:28 PM | 32 | Average | 2 |
| 76 | O Hare | 1:28 PM | 32 | Mode | 1018 |
| 77 | Edgewater | 1:47 PM | 13 | Mode | 98 |

The image below shows a sample point such as O Hare. On clicking the sample point, we obtain the estimated start time and the duration. This can be better visualized in the Python notebook relevant to this Project. This can be done for the other points as per the Visualization.




Github Link :
https://github.com/varadtupe/ChicagoTaxiTrip/blob/master/Probability/Chicago%20Taxi%20Probability%20Project.ipynb

Nbviewer Link :

http://nbviewer.jupyter.org/github/varadtupe/ChicagoTaxiTrip/blob/master/Probability/Chicago%20Taxi%20Probability%20Project.ipynb