EAS 595:
Project Report

NYC Uber Data Analysis

- Anish C Shah
- Martin Jia
- Yashas J Shamaraju

EAS 595

Project Report Group 8

NYC Uber Taxi Data Probability Analysis

Abstract:

With people using For-Hire Vehicles (FHVs) more than ever, it could give us various meaningful insights and patterns for its usage. We explored, analyzed and visualized the NYC Uber Data and now based on this analysis we have tried to predict the time duration needed for a person from an arbitrary pickup place to reach his destination.

Data Description:

Early in 2017, the NYC Taxi and Limousine Commission (TLC) released a dataset about Uber's ridership between September 2014 and August 2015. This dataset contains features such as destination, trip distance, and duration that were not available in other sets released before. The data comprises one complete year of trips, with a total of about 31 million entries.

Proposed Analysis:

1. Analyze the trip duration to determine the probability distribution.

2. Use the Training data to find the parameters of probability distribution and validate it using the test data.

3. Estimate the start time from an arbitrary place to reach the destination at 2:00 pm ?

Analysis Methods:

Quantitative Data Analysis Methods using Python in Jupyter and probability distribution model to find the p.d.f of the time duration.

Methodology:

Time duration in the interval of 5min is plotted against the number of trips taken in each time interval (of 5min). Since our cleaned set of data points consist of time duration of >10 hours, we considered time duration from 0 to 600 min in a step size of 5 min. Before any further analyses we should also note the following

1. Each trip duration is independent of the other.
2. Number of trips is sum of these independent variables.

This provided an insight that the distribution is Normal or Gaussian.

On plotting this data, we found this to be true.

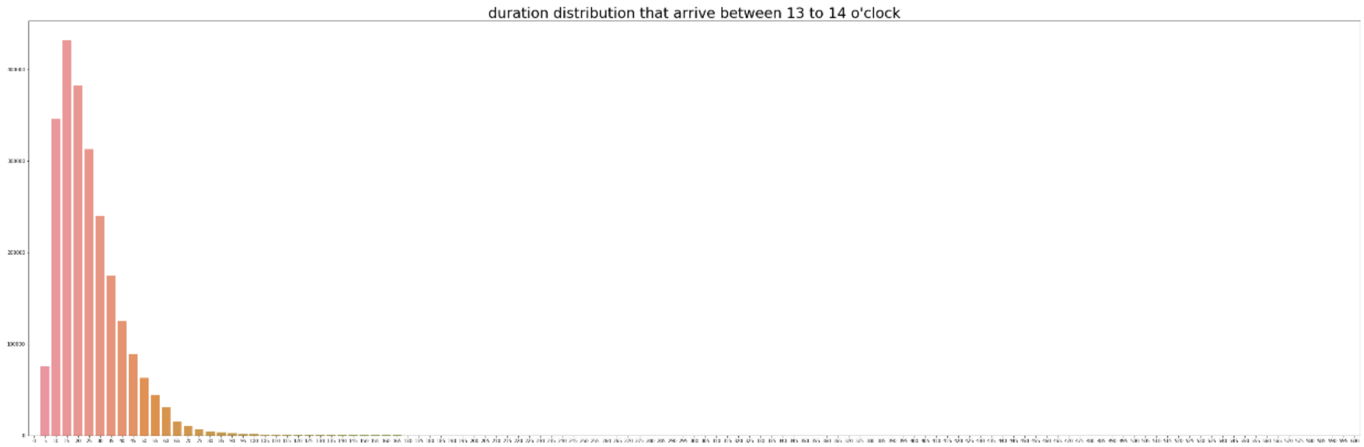
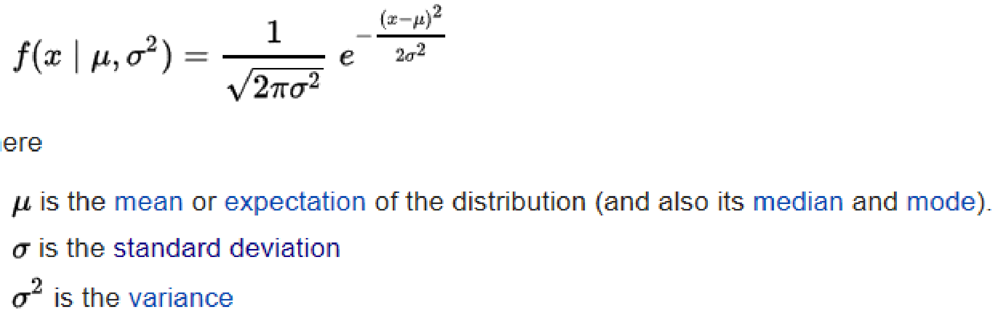


Figure 1: plot of time duration vs Number of Trips

This is a normal distribution that has pdf of the form:



$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where

- μ is the mean or expectation of the distribution (and also its median and mode).
- σ is the standard deviation
- σ^2 is the variance

Also this data is skewed to the left making it log-normal distribution.

Further to find the parameters of this distribution we did curve fitting using the data provided. As these parameters will be helpful in answering the third question, we have selected the specific data set whose start_time is between 1:00 pm to 2:00 pm. This helps to reduce the data used from 31 million (30,925,736) to around 2 million (2,369,128) data. To find the parameters of this curve we have divided the reduced data set in the ratio 0.8:0.2. Making training data = 1896355 data set and testing data = 472773.

Using the function norm.fit() in python we were able to fit the training data set to obtain the mean and the standard deviation of the normal deviation.We validated this by using the test data.

Later using the mean (expected time duration we found the answer to problem of the start time to reach a destination)

Results:

On fitting the graph, we found the mean = 22.8240 and Standard deviation = 15.9432. These are the parameters required to plot the pdf of the curve.

We validated the curve with test data and found the mean square error to be around 0.14%

The probability distribution function of the curve is :

$$P(x) = \frac{1}{\sqrt{2\ \pi\ \sigma^2}}\ e^{\frac{-(x-\mu)^2}{2\ \sigma^2}}$$

$$P(x) = \frac{1}{\sqrt{2\ \pi\,(15.94)^2}}\ e^{\frac{-(x-22.82)^2}{2\ .(15.94)^2}}$$

The CDF of the normal distribution is given:

$$p = F(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} e^{\frac{-(t-\mu)^2}{2\sigma^2}}\,dt$$

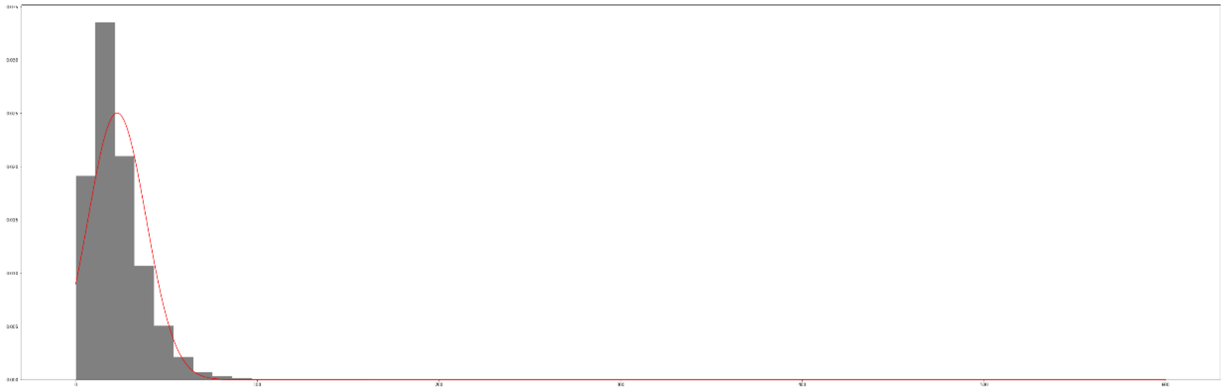


Figure 2: Fitted data with test data histogram

Conclusion:
Since the Expected value or the mean of the pdf was 22.82 min we found out that in order to be at the destination at 2:00 pm, a person has to start his journey at least 23min before 2:00pm.

References:

1. https://www.kaggle.com/dotman/data-exploration-and-visualization/notebook
2. http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
3. Dataset Link: https://s3.amazonaws.com/nyc-tlc/misc/uber_nyc_data.csv