Alex Vdovyn

Alex.vdovyn@gmail.com

NYC Taxis ANalysis DOcumentation

Capital One Data Science Challenge

Contents

[Introduction 1](#_Toc489216946)

[How to Run Script 1](#_Toc489216947)

[Libraries Used 1](#_Toc489216948)

[Questions 1](#_Toc489216949)

[Question 1 1](#_Toc489216950)

[Question 2 3](#_Toc489216951)

[Question 3 4](#_Toc489216952)

[Question 4 6](#_Toc489216953)

[Question 5 (Option A) 7](#_Toc489216954)

[Additional Notes 7](#_Toc489216955)

# Introduction

This documentation contains relavent information, description on how to use and execute script **(main.py)** for NYCTaxis Analysis Challenge.

# How to Run Script

To run script, please open Python version 3.6.

The script was fully tested on Windows OS. Latest version of Anaconda was used to run this code.

Installation can be found here: <https://www.continuum.io/downloads>

If you have any question regarding execution of the script, please contact: Alex Vdovyn, [alex.vdovyn@gmail.com](mailto:alex.vdovyn@gmail.com)

# Libraries Used

* Pandas – to read and load file,
* matplotlib – to plot all supporting graphs,
* numpy – to work with arrays ad perform analysis,
* datetime – for conversion of attributes to datetime data type,
* math – to build supporting function to haversine: to calculate distance between coordinates
* sklearn – for normalization of data during analysis, building predictive model
* stats – to perform ANOVA testing

# Questions

## Question 1

1. Programmatically download and load into your favorite analytical tool the trip data for September 2015.

We importing **pandas** library to read csv file from the link.

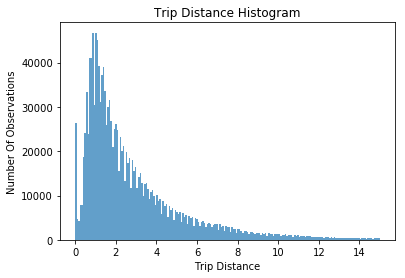
1. Report how many rows and columns of data you have loaded.

Following number of rows were loaded: **1494926**, number of columns: **21.**

## Question 2

1. Plot a histogram of the number of the trip distance ("Trip Distance").

After initial investigation, it was found that **Trip\_distance** variable is positive skewed distribution therefore we fit x-axis to following range: [0-15].



1. Report any structure you find and any hypotheses you have about that structure.

To get basic summary statistics about the structure of **Trip\_distance** variable *describe()* function was used:

**Mean:** 2.96

**Standard Deviation**: 3.07

**Median (50-th percentile**): 1.98

**Min:** 0

**Max:** 603.1

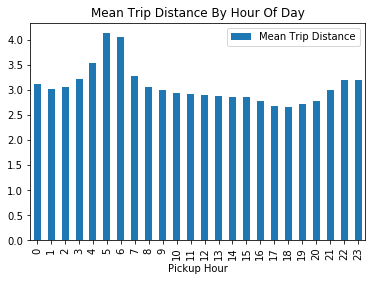
Mean (2.96) is greater than median (1.98) which proves hypothesis above that **Trip\_distance** has a right skewness.

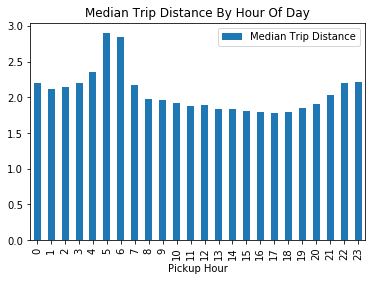
## Question 3

1. Report mean and median trip distance grouped by hour of day.

We will take **lpep\_pickup\_datetime** field to get hour information.

To extract hour attribute we will use **datetime** function from **datetime** module.





From the graphs above, the longest mean and median trips are between 5-6 AM.

1. We'd like to get a rough sense of identifying trips that originate or terminate at one of the NYC area airports. Can you provide a count of how many transactions fit this criteria, the average fair, and any other interesting characteristics of these trips.

Based on Wikipedia article, there are three airports in NYC area: JFK, LaGuardia and Newark Liberty International Airports: <https://en.wikipedia.org/wiki/Aviation_in_the_New_York_metropolitan_area>

Since Green Taxis have specifics to operate only in Bronx, Queens, Brooklyn and Staten Island, we will not include Newark Airport into our analysis.

To identify trips that originate or terminate at one of the NYC area airport, we will need get geographical coordinates (longitude and latitude) of airports above:

NYC area Airports Coordinates

|  |  |  |
| --- | --- | --- |
| Airport | Longitude | Latitude |
| JFK | -73.7787443 | 40.6398262 |
| LaGuardia | -73.8751358 | 40.7773237 |

We will use coordinates from Lat-Long website: <http://www.lat-long.com/>

Also, we will identify trip that originated or terminated at one of the airports if the distance from airport coordinates to the drop-off\drop-on coordinates is less than airport radius.

To find distance between two geographical coordinates, we will define function: *haversine.*

Based on Port Authority NJ website (<https://www.panynj.gov/airports/jfk-facts-info.html>) the area of airports as follows:

The area of JFK airport is approximately 4930 acres (or ~19.9 km2).

To find the radius of airport, we will use formula for circle:

Thus,

So, radius of JFK airport is approximately, 2.52km.

We will assume the radius of LaGuardia airport would have about the same.

We will define function to determine if trip was originated or terminated in one of the airports called: *is\_NYC\_airoport\_trip*

Finally, we will array as field which identifies if trip was originated or terminated at one of the airports.

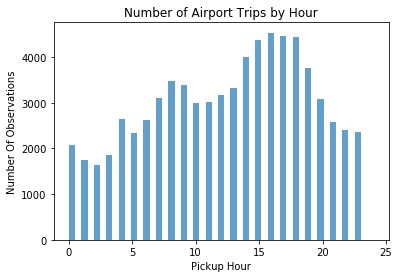
To do that we will call function *is\_NYC\_airport\_trip(v),* where *v*,- is vectors of Drop-off and Pick-up coordinates for each trip.

Technically, green cabs cannot pick up passengers at airports (<https://en.wikipedia.org/wiki/Boro_taxi>) unless is pre-arranged trip, but we will include this information for completeness.

Some other interesting characteristics of airport vs non-airport trips, please see below:

|  |  |  |
| --- | --- | --- |
| Characteristic | Airport Trips | Non-Airport Trips |
| Average Fair Amount | $19.1 | $12.2 |
| Average Distance | 5.5 miles | 2.8 miles |
| Average Tip Amount | $2.1 | $1.1 |
| Average Number of Passengers | 142 for 100 trips | 136 for 100 trips |

Histogram of trips by hour:



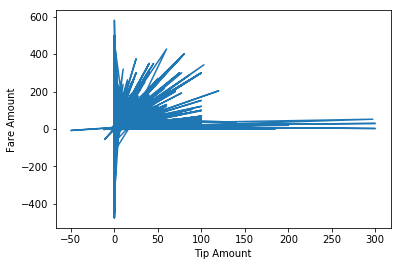
It is clear from histogram above that more airport trips are taken between 2-7PM.

## Question 4

1. Build a derived variable for tip as a percentage of the total fare.

We define percentage of the total fare as follow:

1. Build a predictive model for tip as a percentage of the total fare.



There are 2595 trips for which Tip Amount is greater than Fare Amount.

For further analysis following features are selected:

**Categorical Attributes:**

VendorID, Store\_and\_fwd\_flag, Payment\_type, Trip\_type, Passenger\_count, pickup\_hour, is\_airport\_trip

**Numeric Attributes:**

Trip\_distance, Fare\_Amount, Extra, MTA\_tax, Tolls\_amount, improvement\_surcharge

We will exclude following records from the data-set:

* Tip Amount less than ‘0’
* Tolls Amount less than ‘0’
* Fare Amount is at least $2.5. The initial fare amount
* Trip time should be greater than ‘0’ (in seconds)
* Trip distance should be greater than ‘0’
* Amount of tip should be less than Fare Amount
* Exclude bogus trips which are greater than 200 miles and fare is relatively small Fare amount (~$10)

Number of records in data-set after cleaning is: 1461625, which is around 97% of original records.

Build regression model using **sklearn** library.

## Question 5 (Option A)

1. Build a derived variable representing the average speed over the course of a trip.

We define average speed over the course of a trip as follows:

1. Can you perform a test to determine if the average trip speeds are materially the same in all weeks of September? If you decide they are not the same, can you form a hypothesis regarding why they differ?

We will define one-way analysis of variance (ANOVA) test to determine whether there are any statistically significant difference between the means of 5 groups (weeks) of average speed variable.

Our Null hypothesis is that mean average speeds among all weeks are not significantly different.

We will reject this Null hypothesis and accept Alternative: at least one of the mean average speeds are significantly different from others.

Based on performed test: **p-value** ~ 0.0094 is less than 0.05. Therefore we reject Null hypothesis and conclude that at average speed for at least one of the week is significantly different.

One of the possible reasons is Labor Day weekend on September 7.

1. Can you build up a hypothesis of average trip speed as a function of time of day?

Based on performed test: **p-value** ~ 0.00 is less than 0.05. Therefore we reject Null hypothesis and conclude that at average speeds for trip hours during the day are statistically different.

## Additional Notes

To further enhance analysis, I would include Mahalanobis function or clustering technique to identity outliers in the data and exclude them from the data-set.

In addition, to get a better sense of data I would build visualization and align geographical coordinates of the trips to actual map.

Finally, I would load and perform comparison analysis for green vs yellow cabs to identity some interesting distinguishing features of two types of cabs.