# New York City: Yellow taxi, Uber pick up analysis

Yogiraj Awati, Ashish Kalbhor, Rushikesh Badami, Sushil Thasale

In this project, we plan to analyze cab service dataset for Uber and yellow taxi in New York City for year 2014 and 2015. Our analysis will also include correlation of pickups with respect to weather data.

### Exploratory Analysis:

**Dataset**:

Uber data: April – Sept 2014 and Jan – June 2015

Yellow taxi cab – Jan 2015

Weather Report – Jan – June 2015

**Data dictionary:**

1. *Uber trip data from 2014*

There are six files of raw data on Uber pickups in New York City from April to September 2014. The files are separated by month and each has the following columns:

* Date/Time : The date and time of the Uber pickup
* Lat : The latitude of the Uber pickup
* Lon : The longitude of the Uber pickup
* Base : The [TLC base company](http://www.nyc.gov/html/tlc/html/industry/base_and_business.shtml) code affiliated with the Uber pickup

1. *Uber trip data from 2015*

Also included is the file uber-raw-data-janjune-15.csv This file has the following columns:

* Dispatching\_base\_num : The [TLC base company](http://www.nyc.gov/html/tlc/html/industry/base_and_business.shtml) code of the base that dispatched the Uber
* Pickup\_date : The date and time of the Uber pickup
* Affiliated\_base\_num : The [TLC base company](http://www.nyc.gov/html/tlc/html/industry/base_and_business.shtml) code affiliated with the Uber pickup
* locationID : The pickup location ID affiliated with the Uber pickup

The Base codes are for the following Uber bases:

B02512 : Unter B02598 : Hinter B02617 : Weiter B02682 : Schmecken B02764 : Danach-NY B02765 : Grun B02835 : Dreist B02836 : Drinnen

1. *Yellow cab taxi*

Major fields:

* VendorID : A code indicating the TPEP provider that provided the record. 1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.
* tpep\_pickup\_datetime : The date and time when the meter was engaged.
* tpep\_dropoff\_datetime : The date and time when the meter was disengaged.
* Passenger\_count : The number of passengers in the vehicle. This is a driver-entered value
* Trip\_distance : The elapsed trip distance in miles reported by the taximeter.
* Payment\_type : A numeric code signifying how the passenger paid for the trip. 1= Credit card 2= Cash 3= No charge 4= Dispute 5= Unknown 6= Voided trip
* Extra: Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges
* MTA\_tax: $0.50 MTA tax that is automatically triggered based on the metered rate in use.
* Total\_amount : The total amount charged to passengers. Does not include cash tips

1. *Weather data*

* Station id – weather station nid
* Station name
* Date
* Precipitation – Total precipitation
* Snow – wind
* Snow (in) – Snow for the day
* Tmax – maximum temperature
* Tmin – minimum temperature
* Sndp – snow depth

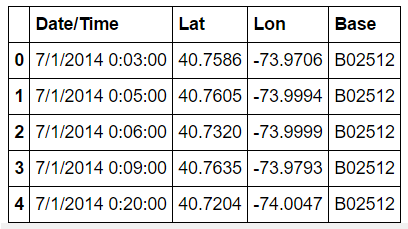
**Data Analysis:**

1. **Getting view of the data**

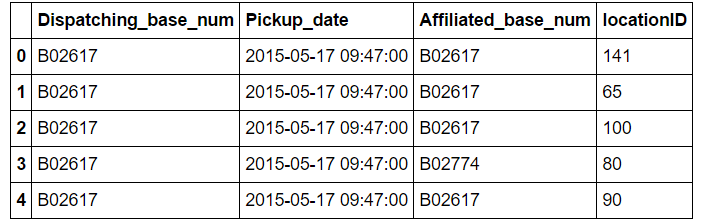
Command: uberrawdata.head()

This command retrieves top 5 rows of the data set.

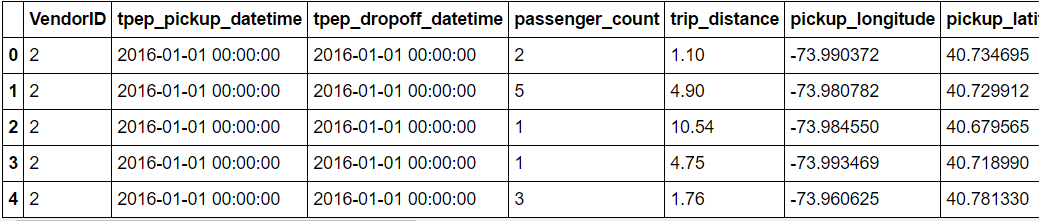
1. Uber 2014 data



1. Uber 2015 data



1. Yellow cab taxi



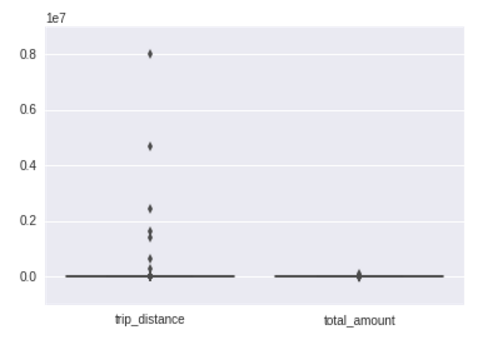
1. **Checking null values**

Command: uberrawdata.isnull().sum()

This command gives total null values per Column. We didn’t find any null values for any of the data set

1. **Identifying outlier**

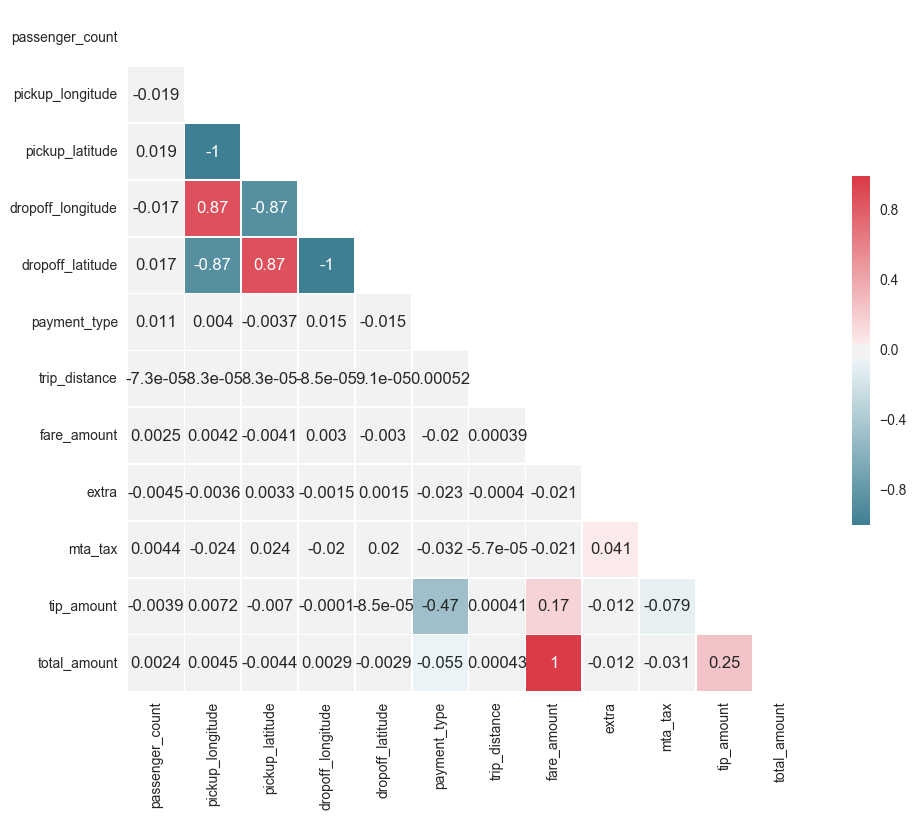
We plot boxplot for the Uber 2015 data.



We found that there do exist some outlier for trip distance values since some values are highly unrealistic. There are couple of points we observe from the graph where the trip distance was greater than 2 million!

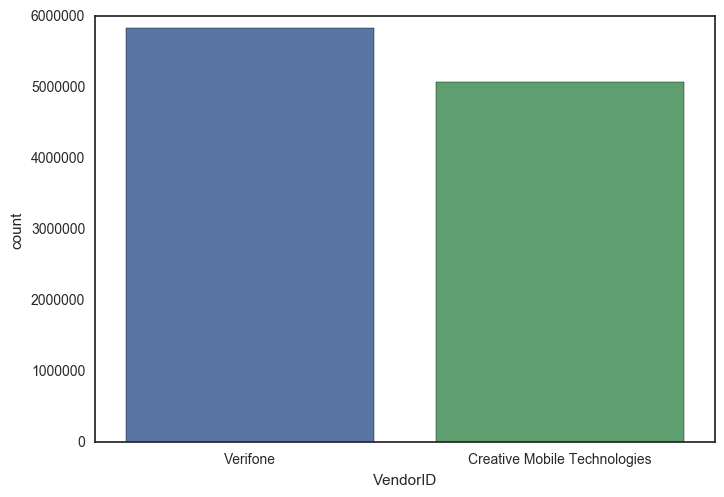
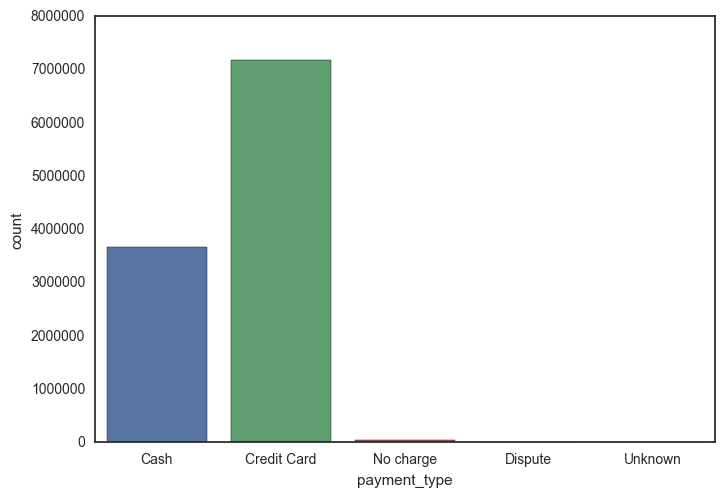
1. **Performing correlation analysis**

We performed plot a correlation graph.



We can clearly identify correlation between various parameters of the yellow cab taxi data set

1. Drop off longitude/latitude – pickup longitude/latitude are highly correlated.
2. Total amount mostly compromises of fare amount and tip amount do have some weight.
3. The extra category is influenced by mta\_tax component
4. **Bar plot**



Interesting observations can be made from bar plot. We find that there has been decent amount of digital payments compared to cash. Vendor Verifone has greater influence compared to Creative Mobile Technologies