New York City Taxi Data Exploration and Ride Prediction

Kurt Nelson

1 Project Summary

Quick taxi pickup times and accurate ride duration predictions ease life in fast-pace cities like New York City (NYC). The project goal is to analyze taxi data collected from two separate carriers in NYC to identify useful patterns for taxi cab drivers, and to develop a prediction model for ride duration. The project is motivated by the Kaggle competition New York City Taxi Duration. Time series plots, joint probability mass functions, and heat maps are created and analyzed to learn about spatial and temporal trends in ride frequency, trip duration, and trip length. K-nearest neighbors is also employed on pickup and drop-off locations to investigate spatial connectivity between regions of the city. Finally, ride duration is predicted using gradient boosted trees. The initial training and test root mean squared logarithmic error was 0.4311 and 0.4413, respectively. This project is ongoing. The next stage of the model development is a parameter tuning study on the learning rate, maximum tree depth, and subsample size used to fit the gradient boosted trees. This document serves as project notes and will eventually be expanded into a full report. It is by no means an polished document.

2 Variables and Data Cleaning

The competition dataset is based on the 2016 NYC Yellow Cab trip record data made available in Big Query on Google Cloud Platform. The data was originally published by the NYC Taxi and Limousine Commission (TLC), and contains the following fundamental variables:

- id - a unique identifier for each trip
- vendor id: a code indicating the provider associated with the trip record
- pickup datetime: date and time when the meter was engaged
- dropoff datetime: date and time when the meter was disengaged
- passenger count: the number of passengers in the vehicle (driver entered value)
- pickup longitude: the longitude where the meter was engaged
- pickup latitude: the latitude where the meter was engaged
- dropoff longitude: the longitude where the meter was disengaged
- dropoff latitude: the latitude where the meter was disengaged
- trip duration: duration of the trip in section

From the fundamental variables, pickup and drop-off times are broken into month, day of week, and hour using the pandas "todatetime" function. The vincenty distance (method for computing distance between to locations specified by latitude and longitude) between pickup and drop-off locations is then computed using the geopy library. Final trip speed, which is a metric for traffic conditions is computed.

The data is then cleaned by accepting all data points lying between the bounds defined by the 0.1% and 99.9% quantiles for trip duration, distance, and speed. The bounds were selected by plotting marginal probability mass functions and selecting limits that removed unrealistic values. Marginal mass functions are shown in Figure 1, and statistics after removal of erroneous data points are shown in Figure 2. The small peak on the right tail of the trip duration mass function (middle panel) represents trips to and from JFK airport.

3 K-means Clustering and Spatially Connectivity

To examine the spatial connectivity of NYC, kmeans-clustering is applied with $k = 16$. The number of clusters was selected by plotting pickup and drop-off locations color coded by cluster number, and choosing the k-value that partitioned the data in a fashion that aligns with physical intuition of how NYC is geographically divide (Figure 3). If a more rigorous approach is desired, something like the "elbow method" can be applied.

A joint probability mass function of pickup and drop-off locations is shown in Figure 4. The most common ride is a pickup and drop-off in cluster 5 (Upper East Side), which is useful information for taxi cab drivers who prefer short trips. Airport pickups and drop-offs are best for drivers aiming for longer trips. For pickups heading to the airport, drivers should positions themselves in clusters 4 and 11. The most probable pickup location can be computed by row summation of the pickup/drop-off joint probability mass function. This can provide additional information for taxi cab drivers. Additional joint probability mass functions could be created for specific days of the week for further insight.

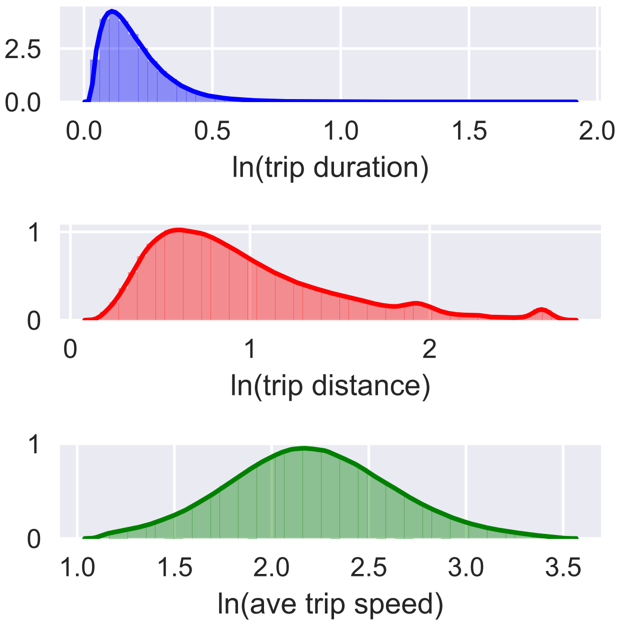
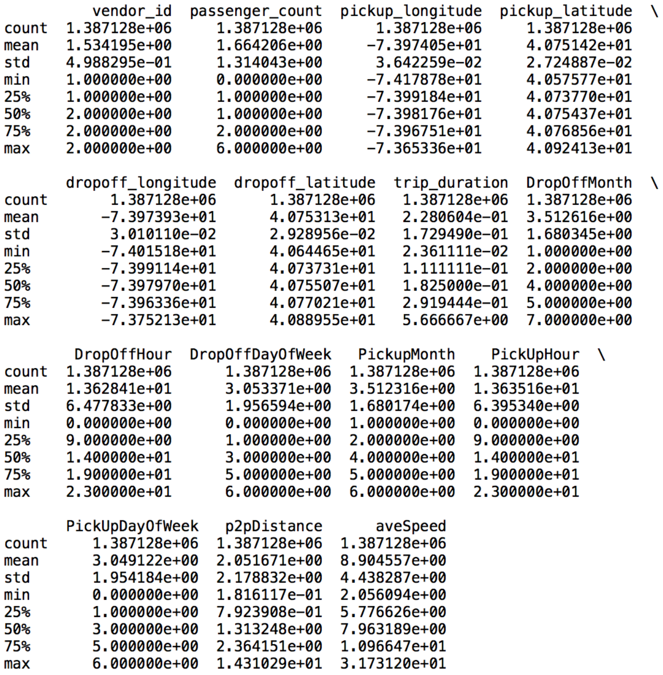


Figure 1: Marginal mass functions of the natural logarithm of trip (top panel) durations, (middle panel) distance, (bottom panel) and average speed after removing erroneous data.



```
          vendor_id  passenger_count  pickup_longitude  pickup_latitude  \
count  1.387128e+06     1.387128e+06      1.387128e+06     1.387128e+06
mean   1.534195e+00     1.664206e+00     -7.397405e+01     4.075142e+01
std    4.988295e-01     1.314043e+00      3.642259e-02     2.724887e-02
min    1.000000e+00     0.000000e+00     -7.417878e+01     4.057577e+01
25%    1.000000e+00     1.000000e+00     -7.399184e+01     4.073770e+01
50%    2.000000e+00     1.000000e+00     -7.398176e+01     4.075437e+01
75%    2.000000e+00     2.000000e+00     -7.396751e+01     4.076856e+01
max    2.000000e+00     6.000000e+00     -7.365336e+01     4.092413e+01

       dropoff_longitude  dropoff_latitude  trip_duration  DropOffMonth  \
count       1.387128e+06      1.387128e+06   1.387128e+06  1.387128e+06
mean       -7.397393e+01      4.075313e+01   2.280604e-01  3.512616e+00
std         3.010110e-02      2.928956e-02   1.729490e-01  1.680345e+00
min        -7.401518e+01      4.064465e+01   2.361111e-02  1.000000e+00
25%        -7.399114e+01      4.073731e+01   1.111111e-01  2.000000e+00
50%        -7.397970e+01      4.075507e+01   1.825000e-01  4.000000e+00
75%        -7.396336e+01      4.077021e+01   2.919444e-01  5.000000e+00
max        -7.375213e+01      4.088955e+01   5.666667e+00  7.000000e+00

        DropOffHour  DropOffDayOfWeek   PickupMonth    PickUpHour  \
count  1.387128e+06      1.387128e+06  1.387128e+06  1.387128e+06
mean   1.362841e+01      3.053371e+00  3.512316e+00  1.363516e+01
std    6.477833e+00      1.956594e+00  1.680174e+00  6.395340e+00
min    0.000000e+00      0.000000e+00  1.000000e+00  0.000000e+00
25%    9.000000e+00      1.000000e+00  2.000000e+00  9.000000e+00
50%    1.400000e+01      3.000000e+00  4.000000e+00  1.400000e+01
75%    1.900000e+01      5.000000e+00  5.000000e+00  1.900000e+01
max    2.300000e+01      6.000000e+00  6.000000e+00  2.300000e+01

       PickUpDayOfWeek   p2pDistance      aveSpeed
count     1.387128e+06  1.387128e+06  1.387128e+06
mean      3.049122e+00  2.051671e+00  8.904557e+00
std       1.954184e+00  2.178832e+00  4.438287e+00
min       0.000000e+00  1.816117e-01  2.056094e+00
25%       1.000000e+00  7.923908e-01  5.776626e+00
50%       3.000000e+00  1.313248e+00  7.963189e+00
75%       5.000000e+00  2.364151e+00  1.096647e+01
max       6.000000e+00  1.431029e+01  3.173120e+01
```

Figure 2: Data set statistics after removing erroneous data. This data should be put in a cleaner format for presentation purposes.

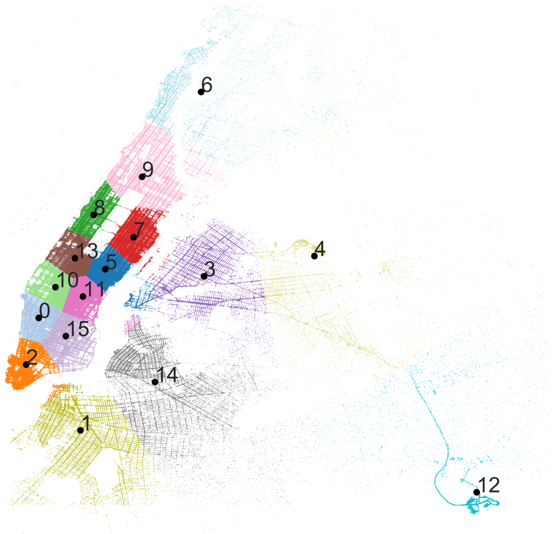


Figure 3: Scatter plot of trip pickup and drop-off locations color coded by cluster number. Black dots show cluster centroids.

Interestingly, commuters commonly travel between clusters 1 (Lower Manhattan) and 5 (Upper East Side). This information is valuable for city planers when deciding regions for increased public transportation.

4 Temporal trends

Daily-averages of ride duration, distance, and trip speed are shown in Figure 5. Several notable observations are:

- Ride duration is inversely correlated to ride distance and speed. People appear to take shorter trips when traffic is bad (slow speed), but the trip duration relatively long.
- Tuesday through Friday appear to have the worst traffic, shortest rides, and longest trip durations.
- People take longer taxi rides on Sunday than any other day. Surprisingly, rides are second longest on Monday. This observation makes me skeptical of the plots (possible shift of days), but it may be explained by more people traveling to and from the airport on Mondays. This prediction can be confirmed from joint probability mass function for pickups and drop-offs on Monday alone.

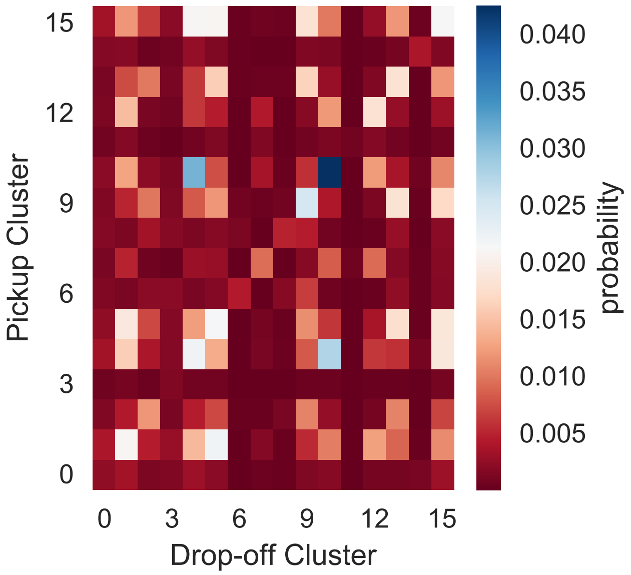


Figure 4: Joint probability mass function .

- Carrier 2 on average drives shorter distances and has slightly shorter ride durations. This trend might be a reflection of the regions within the city each carrier targets.

Hourly-averaged heatmaps of trip distance, duration, and speed are shown in Figure 6. Several notable observations are:

- People taking the longest-distance taxi rides tend to do so between 4-5 am on weekdays, and 5-7 am on weekends. However, in general rides are long at night likely because people are returning home after a long work day or night out.
- The long daily-averaged trip distance for Monday seen in Figure 5 appears to be from night owls spilling into Monday. Trip distances are long between 12-1 am on Monday.
- From roughly 9 am to 6 pm, traffic is bad (slow speed) during the week. The hourly-averaged speed is fastest from 5-6 am during the week, and 6-7 am on weekends.
- Long trip durations are correlated to bad traffic.
- Trip duration is again shown to be inversely correlated to trip distance and speed.

A joint probability mass function based on ride pickup time is shown in Figure **??**. This plot indicates when taxi services are most used. Several notable observations are:

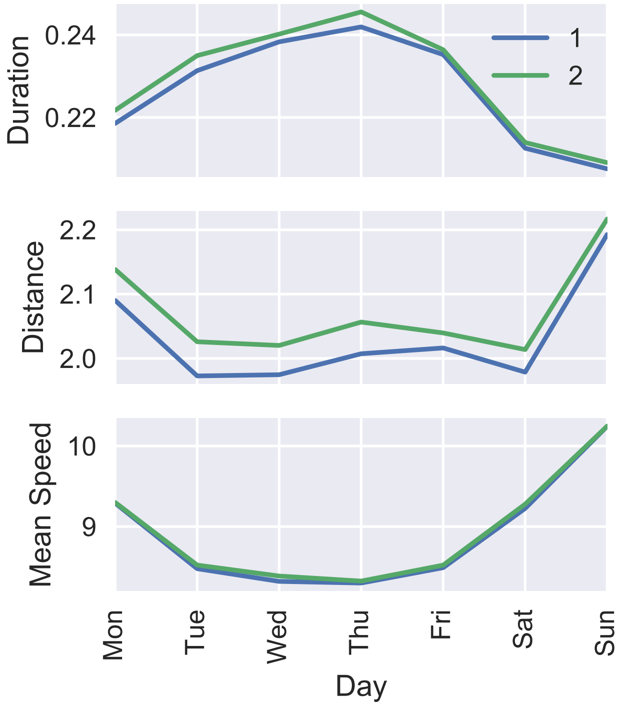


Figure 5: Daily-averaged (top panel) ride duration, (middle panel) distance, and (bottom panel) speed separated by carrier.

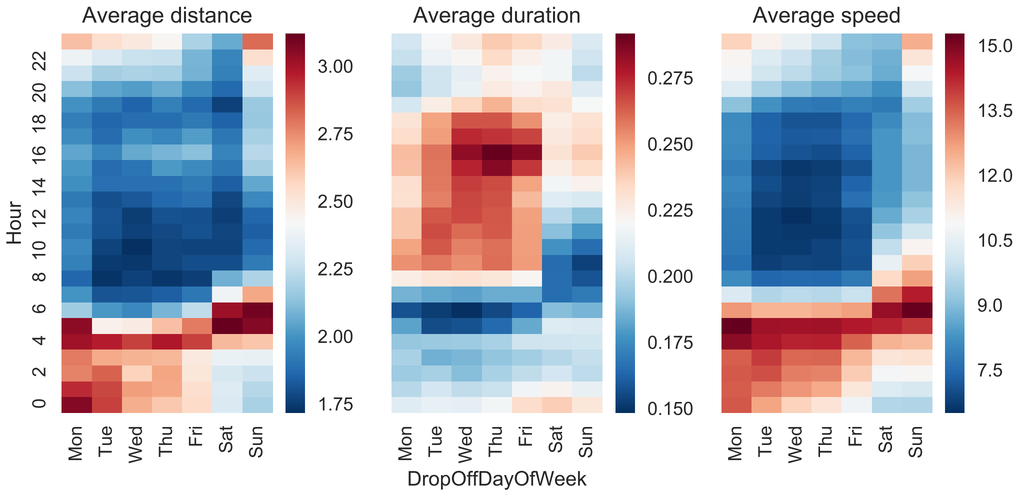


Figure 6: Heatmaps of hourly-averaged trip (top panel) distance, (middle panel) duration, and (bottom panel) speed.

- Taxi rides are frequently taken between roughly 5-9 pm during the week, with rides most frequent between 6-7 pm.
- There are relatively fewer taxi rides between 12-7 am during the week.
- During the weekend, people do not frequently take taxis between 4-9 am. The night owls however stay out later on the weekends, with ride frequency increasing between 12-3 pm. NYC is a late night city!
- Between roughly 5 pm and 12 am, taxi ride frequently goes up for Monday through Saturday.

The number of pickups and drop-offs for each cluster are plotted as a function of hour of day (Figure 8) and day of week (Figure 9). These plots should be recreated with legends indicating line labels. More observations will be noted once legends are created, however three quick comments:

- Generally, ride counts based on hour of day are correlated between clusters.
- From 6 pm to 5 am, the number of rides for all clusters monotonically decreases.
- Between approximately 5-8 am, the number of rides at all locations increase.
- Additional information on spatial and temporal correlations will be inferred once labels are created. Autocorrelation coefficients can also be computed to enhance the discussion.

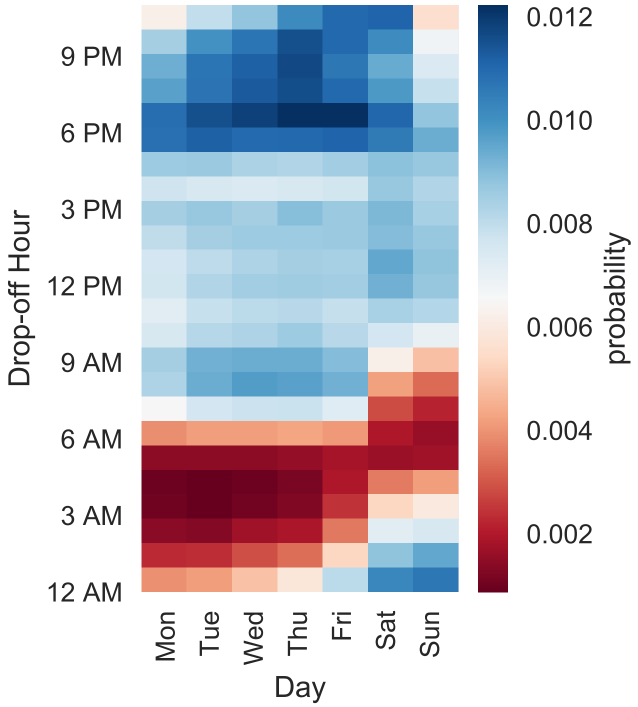


Figure 7: Joint probability mass function based on ride pickup times.

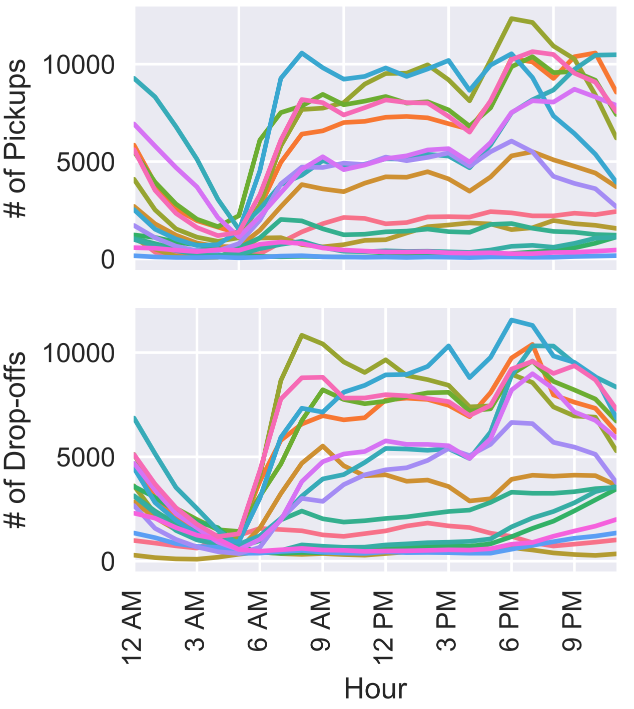


Figure 8: Number of rides as a function of hour of day for each cluster based on (top) pickup and (bottom) drop-off location.

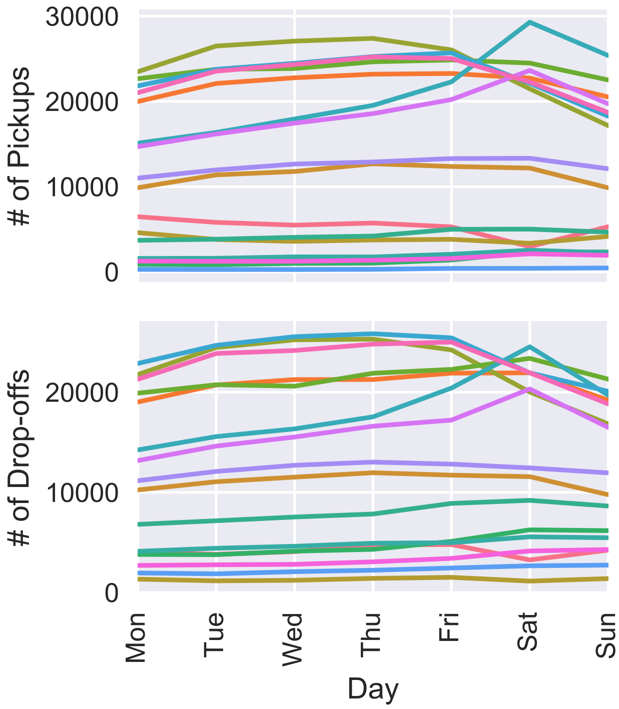


Figure 9: Number of rides as a function of hour of day for each cluster based on (top) pickup and (bottom) drop-off location.

5 Gradient-boosted Tree

Notes coming soon!