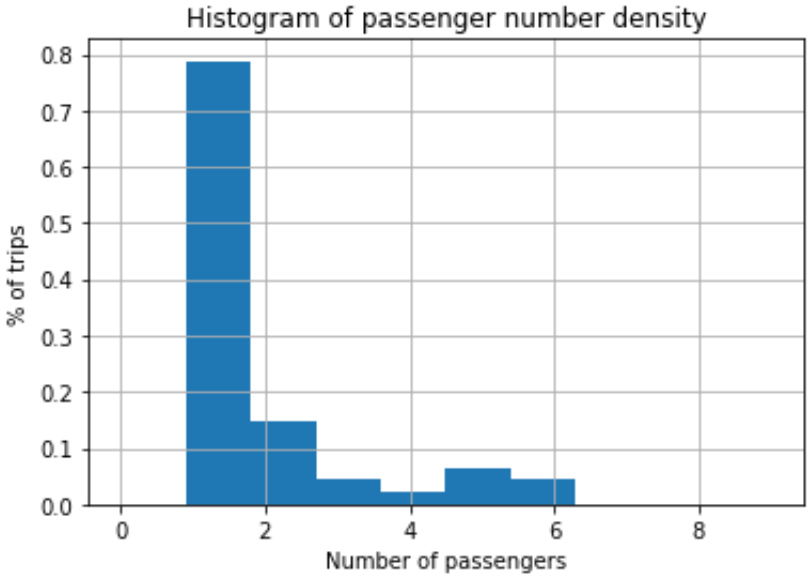
# New York Taxi Case Study

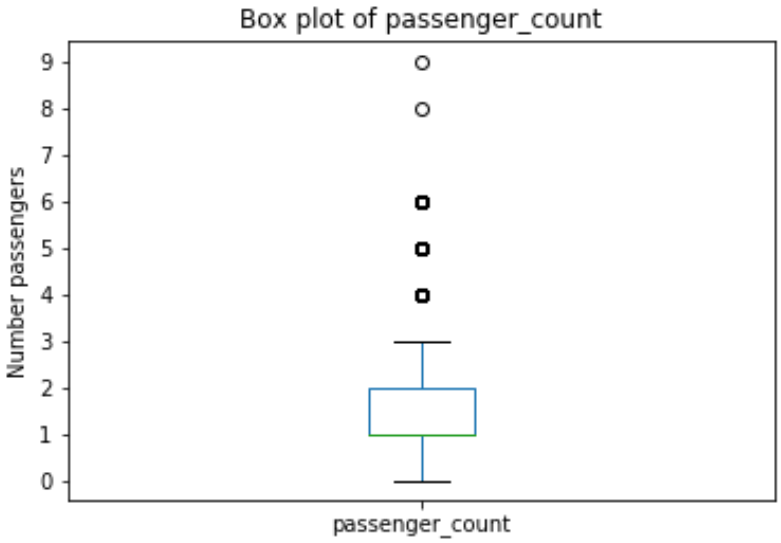
Tony Moriarty for CBA - 5/5/2018
Notebooks referenced are at:
https://github.com/desultir/nyctaxi

## 1. What is the distribution of number of passengers per trip?

*For working, see "describe data.ipynb"*

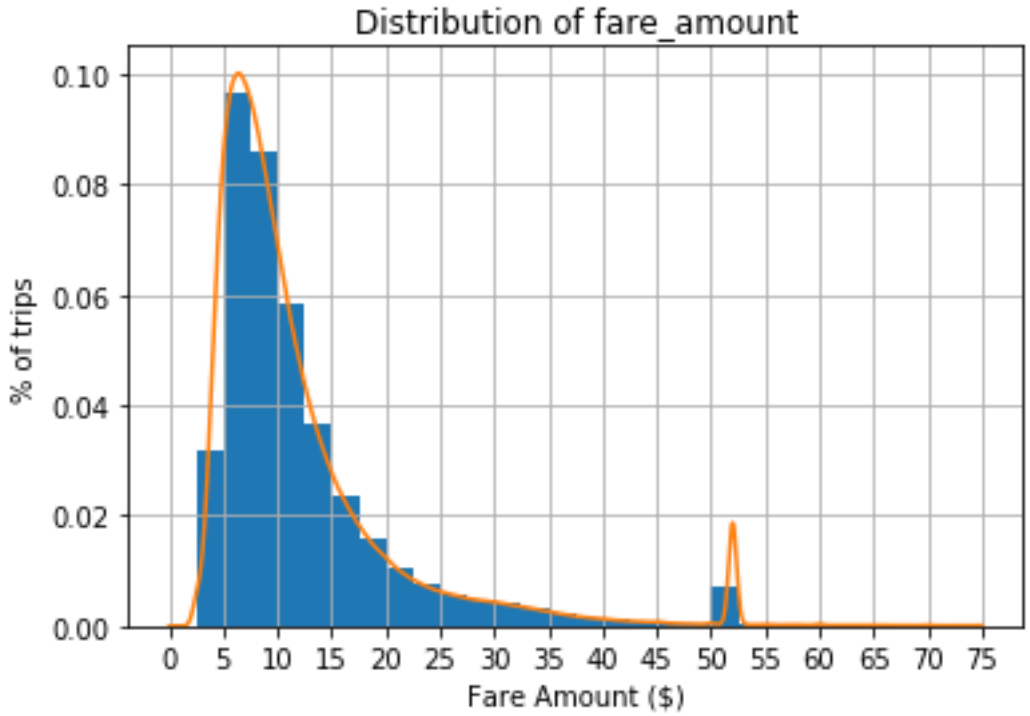



The mean is 1.71 but the median is 1. The 75$^{th}$ percentile is 2.

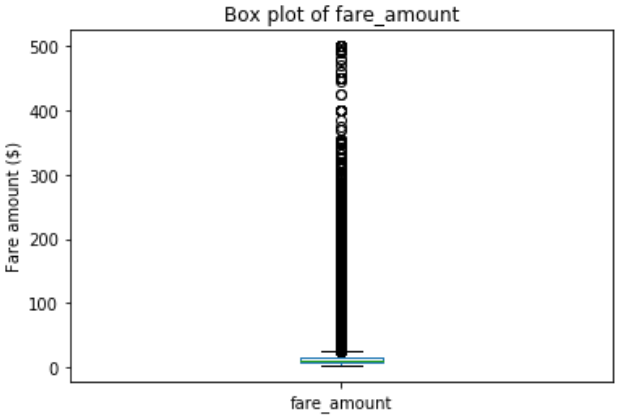



Of interest are the 229 trips with 0 passengers. This is likely incomplete data as time, distance, and in some cases latitude and longitude are mostly zero for at least 75% of these records.

## 2. What is the distribution of fare amount?

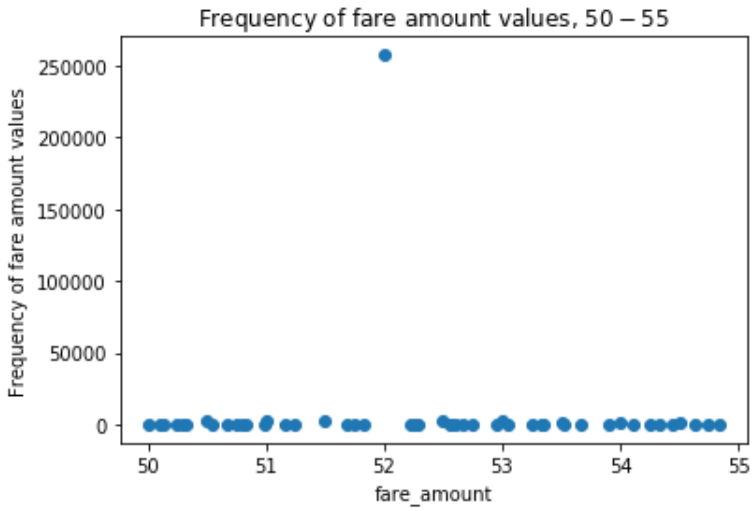



From the histogram, it seems to be a long-tailed distribution, perhaps an exponential distribution. Both KDE and the histogram itself pick up an interesting bump in the $50-$55 bin.

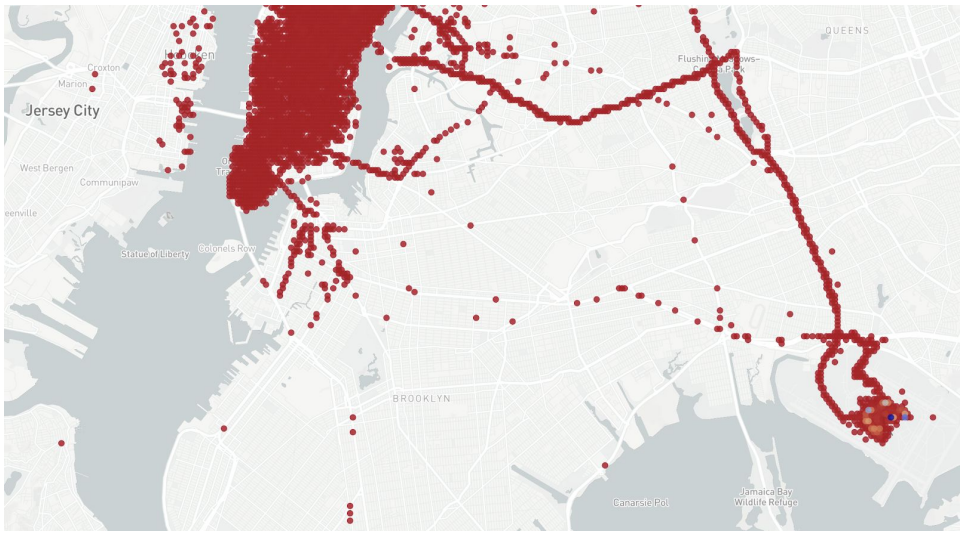



*The long tail of fare_amount*

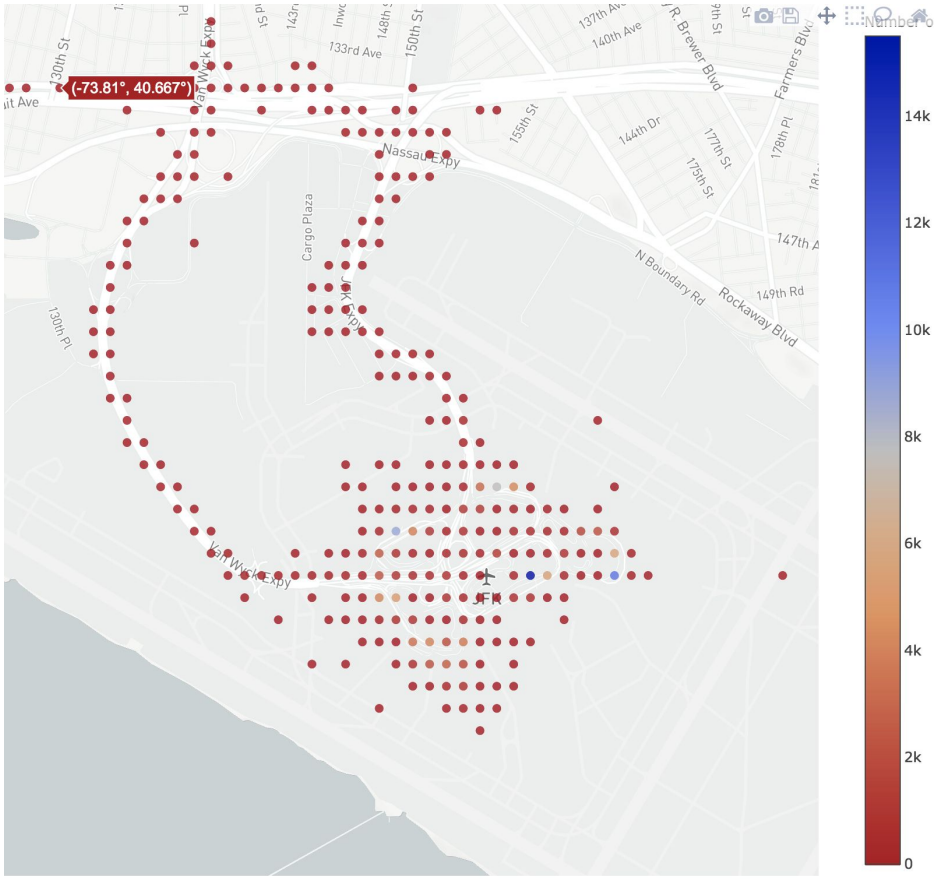
Examining this data range more closely uncovers nearly 250,000 trips with fare of exactly $52.00.




Plotting the $52 fares on a map of NYC shows a reasonably flat distribution across Manhattan, with some interesting heat around JFK.




*$52 fares in Downtown Manhattan and Brooklyn NYC*

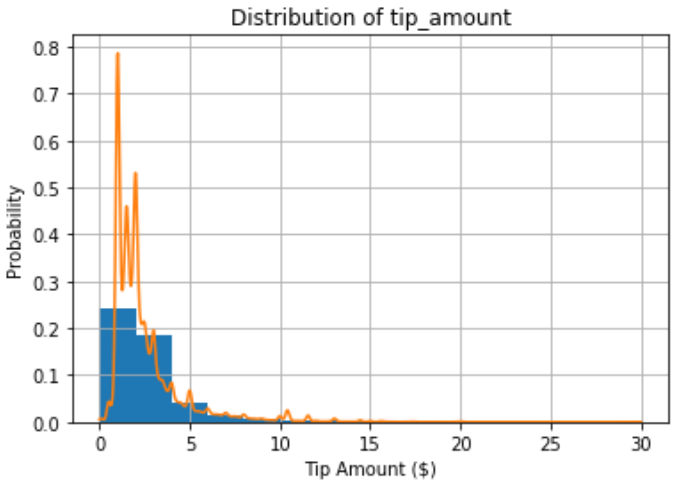



*$52 fares from JFK airport*

Note the blue dots - these are 15,000 pickups all with exactly the same fare in exactly the same spot - the taxi rank at JFK. JFK airport website discusses this flat “$52 to manhattan” fare. https://www.airport-jfk.com/taxi.php

## 3. What is the distribution of tip amount?

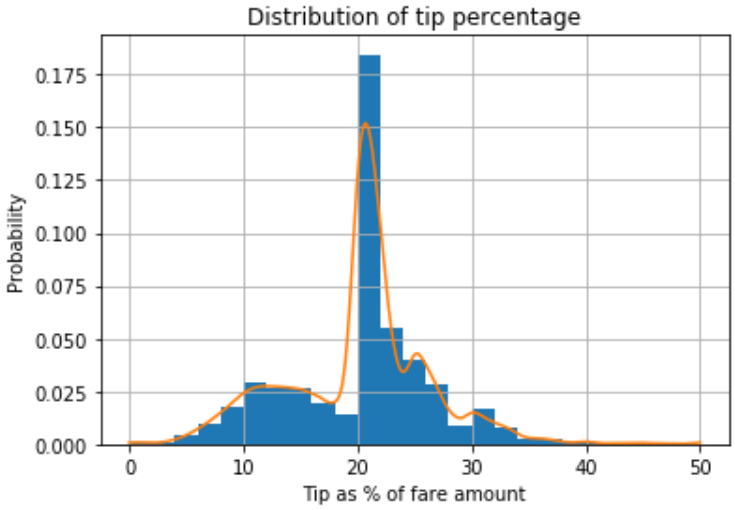
7,220,943 of the 15,101,772 trips in the dataset have tip of $0.00. This is a caveat of the data collection method – only credit card tips are represented; cash tips are not. These trips should essentially be treated as a ‘null’ and excluded from analysis.




*Tip amount with KDE density function*

The bandwidth selected by the automatic scipy gaussian_kde function is too small as the density function exhibits too much variance in the 0-5 bin. It does still tell the story of most tips being in this range.

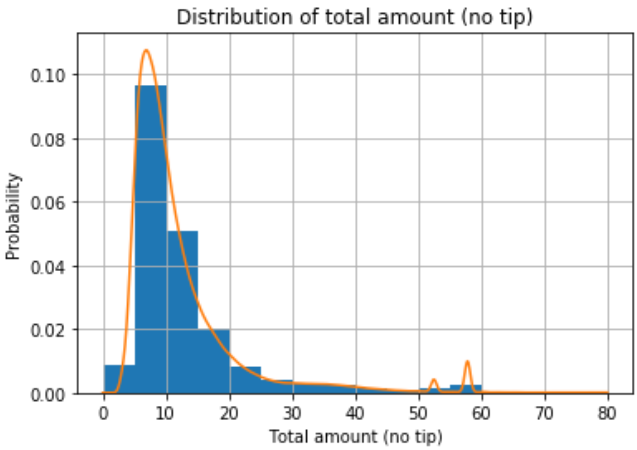
What's more informative is tip as a percentage of fare_amount. As expected, the mode of the distribution is at 20%.




*Tip as percentage of fare amount*

## 4. What is the distribution of total amount?

Due to the large amount of missing data for tips, "total_amount" is really drawn from two underlying distributions – those where the tip was paid by credit card and those where it was paid in cash. As such, I will examine them separately.




*Distribution of total amount for fares with no tip*

As with the fare-amount graph, there is the bump between 50 and 60 which is the $52 flat-rate JFK fare plus tolls.

## 5. What are top 5 busiest hours of the day?

For pickups, the busiest hours are 18:00 through 23:00 (19:00-20:00 being the busiest).

For dropoffs, the busiest are the same 5 hours, meaning summing pickups and dropoffs and counting both as activity gives the same conclusion.

| | Activity |
|---|---|
| **19** | 1924447 |
| **20** | 1842894 |
| **18** | 1828371 |
| **21** | 1769995 |
| **22** | 1712795 |

*Total activity by hour in which it occurred*

3:00 to 5:00 show the least activity.

## 6. What are the top 10 busiest locations of the city?

Taking locations to be the Taxi neighbourhoods as defined by the New York Taxi Commission (http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml) and binning trips by pickup and dropoff location, the busiest pickup locations are:

| pickup_neighbourhood | Pickups |
|---|---|
| Upper East Side South | 566047 |
| Midtown Center | 526084 |
| Upper East Side North | 510157 |
| Midtown East | 506596 |
| Union Sq | 505786 |

For dropoffs, the busiest are:

| dropoff_neighbourhood | Dropoffs |
|---|---|
| Midtown Center | 604857 |
| Upper East Side North | 514560 |
| Upper East Side South | 506748 |
| Murray Hill | 494526 |
| Times Sq/Theatre District | 486968 |

Defining busiest as the sum of total activity in the neighbourhood (dropoffs + pickups) the 5 busiest are:

| pickup_neighbourhood | Activity |
|---|---|
| Midtown Center | 1130941 |
| Upper East Side South | 1072795 |
| Upper East Side North | 1024717 |
| Murray Hill | 999437 |
| Midtown East | 984616 |

## 7. Which trip has the highest standard deviation of travel time?

Again using Taxi neighbourhoods from the New York Taxi Commission, the trips with the highest std-dev are as follow:

| pickup_neighbourhood | dropoff_neighbourhood | trip_time_in_secs_stddev | trip_time_in_secs_med |
|---|---|---|---|
| Whitestone | JFK Airport | 5598.687822 | 1080.0 |
| South Ozone Park | Briarwood/Jamaica Hills | 4941.262187 | 3734.0 |
| East New York/Pennsylvania Avenue | OutsideNYC | 4356.484879 | 3080.5 |
| College Point | Jackson Heights | 3908.555520 | 690.0 |
| Inwood | Woodside | 3595.187663 | 1500.0 |

Getting from Whitestone to JFK, a trip with a median travel time of only 18minutes, has a std deviation of 5598 seconds, or 93 minutes. This would appear to be a data errors, as there are only 3 trips on this route. Two took roughly 18 minutes, and one took 3 hours. Perhaps the driver neglected to log off the meter.

### 8. Which trip has most consistent fares?

There are 313 neighbourhood combinations with exactly 1 fare (giving a std dev. of zero.). 5481 combinations have never seen a single fare.

Of the trips with a significant number of fares and a std.dev of zero, Hamilton Heights to JFK has the most fares (93). This would be due to the $52 fixed fare to JFK.

## Open Questions

### 1. In what trips can you confidently use respective means as measures of central tendency to estimate fare, time taken, etc.

*For working, see "describe data.ipynb"*

In essence for central tendency we must compare mean to the median, or assess the skewedness of the distribution. For exponential or gamma distributions, the mean and the median will not necessarily be close and median is a better measure of central tendency.

To assess the usefulness of mean as a measure of central tendency, we will assess how many standard deviations the median is away from the mean for each column. As we want them as close as possible, lower is better. For trip_time, the following trips (with a significant number of fares > 2000) all had a very close median and mean:

| | | trip_time_in_secs | t |
|---|---|---|---|
| pickup_neighbourhood | dropoff_neighbourhood | | |
| Upper East Side North | Garment District | 0.001860 | |
| JFK Airport | Upper East Side South | 0.006878 | |
| | Lincoln Square East | 0.009252 | |
| Upper East Side North | Midtown South | 0.020752 | |
| Upper West Side South | Union Sq | 0.025638 | |
| East Harlem North | Central Harlem | 0.029904 | |
| East Harlem South | Central Harlem | 0.030989 | |
| JFK Airport | Penn Station/Madison Sq West | 0.036724 | |
| | Midtown North | 0.040757 | |
| Manhattan Valley | Lincoln Square East | 0.041830 | |
| Morningside Heights | Central Harlem | 0.042916 | |
| Lincoln Square East | Lincoln Square West | 0.048109 | |
| Upper West Side South | Bloomingdale | 0.048679 | |
| Midtown North | Little Italy/NoLiTa | 0.049188 | |
| Upper West Side North | Lincoln Square East | 0.050288 | |
| Upper East Side South | East Chelsea | 0.051069 | |
| Upper East Side North | Union Sq | 0.052132 | |
| Penn Station/Madison Sq West | Upper East Side South | 0.052942 | |
| SoHo | Upper East Side South | 0.053007 | |
| JFK Airport | Clinton East | 0.053554 | |
| Upper West Side South | Upper West Side North | 0.054923 | |
| Lenox Hill East | Garment District | 0.055554 | |
| Morningside Heights | Central Harlem North | 0.056934 | |
| Upper West Side North | Union Sq | 0.057578 | |

*Most consistent fares by trip time; distance between mean and median in std.dev*

For trip_distance we run a similar calculation:

| | | trip_distance |
|---|---|---|
| pickup_neighbourhood | dropoff_neighbourhood | |
| Clinton East | Lower East Side | 0.000140 |
| | Yorkville East | 0.000179 |
| SoHo | Midtown North | 0.000331 |
| Lincoln Square East | Midtown Center | 0.000480 |
| Morningside Heights | Lincoln Square East | 0.000613 |
| LaGuardia Airport | West Village | 0.000719 |
| Upper West Side South | Midtown Center | 0.000755 |
| | Flatiron | 0.000797 |
| JFK Airport | Forest Hills | 0.001020 |
| West Village | Sutton Place/Turtle Bay North | 0.001059 |
| Meatpacking/West Village West | Midtown North | 0.001079 |
| Union Sq | Lenox Hill West | 0.001197 |
| Upper East Side North | Lincoln Square East | 0.001218 |
| East Village | West Chelsea/Hudson Yards | 0.001292 |
| West Chelsea/Hudson Yards | Murray Hill | 0.001660 |
| Lincoln Square East | Yorkville East | 0.001679 |
| Midtown North | East Village | 0.001749 |
| Clinton East | Lenox Hill East | 0.001766 |
| Times Sq/Theatre District | World Trade Center | 0.002202 |
| SoHo | Midtown Center | 0.002208 |
| Midtown East | West Village | 0.002396 |
| Greenwich Village North | Sutton Place/Turtle Bay North | 0.002524 |
| Midtown South | Lenox Hill West | 0.002558 |

*Most consistent fares by trip distance; difference between mean and median in std.dev*

Finally we examine fare_amoun. JFK is over-represented in this table due to the flat $52 fare:

| pickup_neighbourhood | dropoff_neighbourhood | fare_amount |
|---|---|---|
| Midtown North | Manhattan Valley | 0.000234 |
| Greenwich Village South | Kips Bay | 0.000408 |
| East Village | JFK Airport | 0.000793 |
| Lincoln Square West | Morningside Heights | 0.000855 |
| Lincoln Square East | Lincoln Square West | 0.001102 |
| Central Harlem | East Harlem North | 0.001658 |
| Morningside Heights | Upper West Side South | 0.001871 |
| Murray Hill | Lenox Hill East | 0.002708 |
| Midtown Center | JFK Airport | 0.003036 |
| Central Harlem | Morningside Heights | 0.003249 |
| Upper West Side North | Washington Heights South | 0.003747 |
| Financial District North | Midtown South | 0.003850 |
| Clinton East | Yorkville East | 0.004270 |
| Lower East Side | Flatiron | 0.004913 |
| Upper East Side South | West Chelsea/Hudson Yards | 0.004990 |
| Lincoln Square East | Bloomingdale | 0.005037 |
| Upper East Side North | East Harlem North | 0.005112 |
| Clinton East | JFK Airport | 0.005211 |
| Upper East Side North | Garment District | 0.005365 |
| Gramercy | Lincoln Square East | 0.005571 |

*Most consistent fares by fare amount; difference between mean and median in std.dev*

## 2. Can we build a model to predict fare and tip amount given pick up and drop off coordinates, time of day and week?

*For working, see "RFR on fares.ipynb"*

Tip amount and fare amount were taken as two 'target' variables and a Random Forest regression run to estimate them. Binning the geo-spatial aspect into "neighbourhoods" allows the RF to treat pickup and dropoff location as discrete neighbourhoods; these can be one-hot encoded to ensure distinction between them.

4-fold cross validation was performed using shuffling. The performance on the held-out test set was measured using RMSE, and was remarkably stable across the 4 folds. The mean RMSE on fare prediction was 12.75; the std dev 0.16. The mean RMSE on tip prediction was 2.87; the std 0.01.

This RMSE on fare prediction ensures that the model can tell a driver, to with plus or minus $12, what fare they will receive from a journey. This is an acceptable limit on longer journeys, and the assumption is that for shorter journeys the driver will be able to estimate.

The tip estimation is a much more useful feature; incorporating elements of “pickup_neighbourhood”. Examining the feature importances of the RF; the distance was by fare the most important feature which is logical… however time of day and day of week were also given a medium amount of importance.

## 3. If you were a taxi owner, how would you maximize your earnings in a day?

*For working, see “describe data”, “SARIMA model”, “GP model” and “RFR on fares”*

There are two factors to earnings - demand and revenue. Demand is important as an empty taxi in an area with low demand is not just an opportunity cost, but an actual cost (petrol, wages).

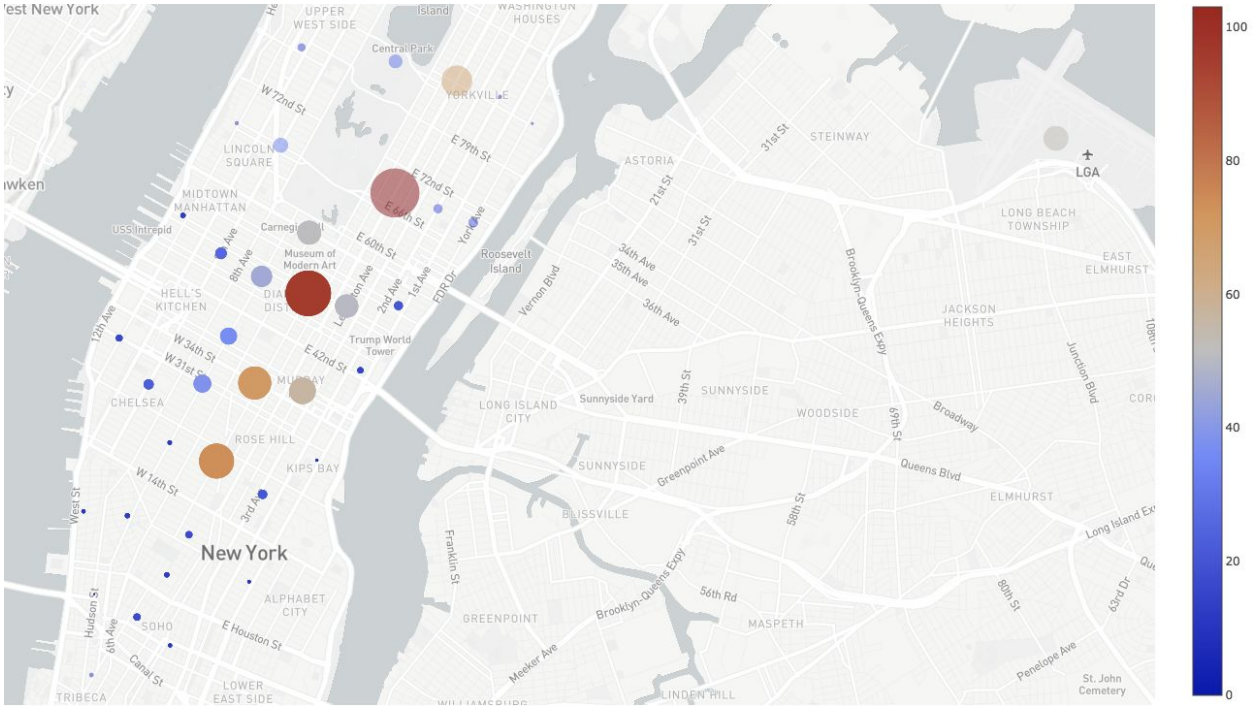
**Demand**

To model demand, to approaches were taken - a time series approach (SARIMAX model) modelling the demand time series by neighbourhood, and a Gaussian Processes model modelling demand by neighbourhood in 1 hour windows.

The SARIMA model had order chosen by inspection of ACF, PACF and grid search. A daily and weekly seasonal component was identified - the final model chosen using minimization of AIC along with Q-statistic had order 1,0,1 and seasonality 1,0,1. This model was tested using time-series cross validation and gave an RMSE of 26.64 (units are demand in number of pickups).

The Gaussian Processes model is a non-parametric Bayesian model - the advantage being that Bayesian models return the confidence or credibility of their predictions. Features were engineered to emulate the time-series aspect such as “dayofweek” and “hour_of_day” to encourage learning of seasonality.

4-fold cross validation was run, leaving one week out at a time. The convergence of a GP model can be assessed by examining the standard deviation across different runs - 3 of 4 weeks converged well giving a std-dev of 4.79 and RMSE of roughly 4.0. Leaving the second week out did fail to converge - the std-dev and RMSE were four times higher. Running more iterations may have helped convergence here.




*Predicted Demand by neighbourhood; GP model; 4pm Friday*

The GP model was selected as the primary model for modelling demand.

**Revenue**

The RF model from question 2 can be refitted to give fares based purely on pickup_neighbourhood, hour of the day, and day of week. Re-fitted using 4-fold cross validation, the RFR had mean RMSE of $8.82 with std.dev 0.07. For tip prediction, it had RMSE of $2.58 with std.dev of 0.01. These very stable results are not extremely accurate but ideally enough for the tiebreaking purposes outlined below.

**Conclusion**

Given the above two models, an ensemble could be built to identify nearby areas of high demand. This model could be provided through an interface to incorporate drive time - if it will be the next time bin by the time the driver arrives that forecast can be provided.

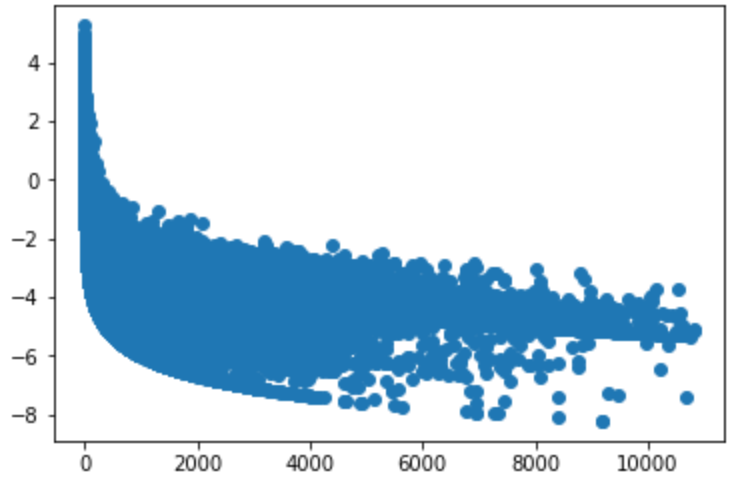
In the event of comparable nearby options, the ensemble will weight higher those neighbourhoods with higher predicted fare/tip.

## 4. If you were a taxi owner, how would you minimize your work time while retaining the average wages earned by a typical taxi in the dataset?

*For working see describe data and kde on trip_time notebooks*

Minimizing work time while holding earning constant creates a new target variable - dollars per second driven. As with question 3, the GP model should be consulted to find nearby areas of high demand as, if downtime is considered work, it's an even higher cost given this scenario.

To characterise this target variable, we plot its log against trip time in seconds. The following graph shows that there is a definite trend - shorter trips are more profitable (per second).



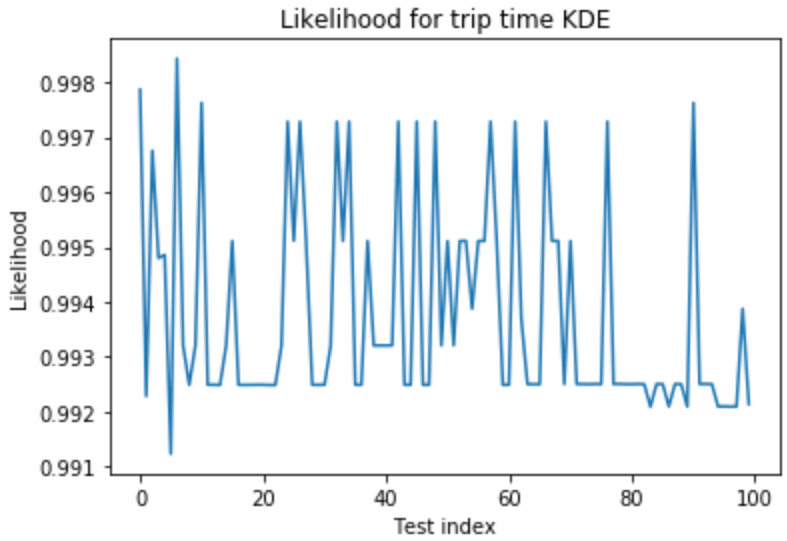

The goal is then recast as building a model for average trip time by pickup neighbourhood by time of day.

Two KDE models were built on the entire dataset, one on “["pickup_hour", "pickup_dayofweek"]” - probability of a pickup in a certain location at a certain time; the other on ["pickup_hour", "pickup_dayofweek", "trip_time_in_secs"] - probability of a certain trip length being picked up in a certain location at a certain time.

The probability of a given trip_time given a certain area at a certain time can then be calculated as:

$$p(\text{trip\_time}|\ \text{location, time\_of\_day}) = \frac{p(\text{trip\_time, location, time\_of\_day})}{p(\text{location, time\_of\_day})}$$

as per Baye’s law.




This model converged with high likelihood on the held out test data, however perhaps due to the scale of the data it gave high probability to almost any trip time. This model needs retraining, however that will be have to held as future work.

Conclusion

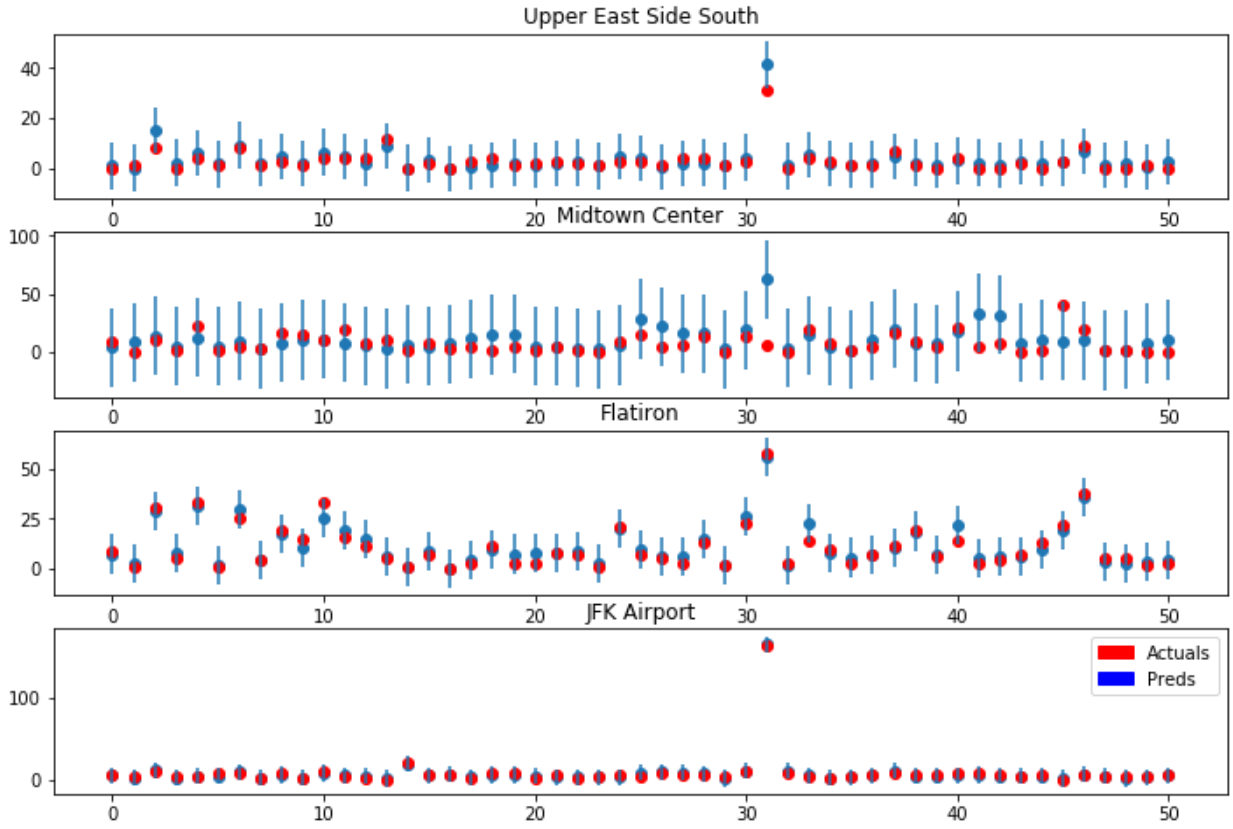
The taxi driver wishing to minimize their work time and maximise earnings should still consult the demand model first as downtime is still the largest cost to earnings. However, when deciding between multiple local maxima, they should consult the second model to find the area with the likely shortest trips.

## 5. If you run a taxi company with 10 taxis, how would you maximize your earnings?

Maximizing earnings for a taxi company can take multiple forms. Ideas involved classification problems identifying disputed/voided fares, using outlier detection to identify drivers defrauding the company, and better trip planning.

Given this trip-related dataset, the third approach was attempted. The issue of downtime is still the most pressing issue - getting cabs where they're needed is the key to customer engagement and therefore profitability. At the taxi network level, secondary effects can also be considered.

The GP model from question two was augmented to model not only demand in a neighbourhood, but which neighbourhoods those customers were likely to want to go to. The overall demand for a neighbourhood can still be calculated using the conditional probability, however the secondary demand is the interesting part of this model.




*Destination neighbourhoods for each of 4 pickup locations, at midnight on a sunday.*
The above charts were drawn using 4-fold cross validation with one week held out, on the held out data.

Each column of the above chart is a destination neighbourhood; the 4 subplots represent 4 chosen pickup neighbourhoods at midnight on a sunday. The choice of Bayesian model gives not only the prediction mean, but std.dev as well (shown as error bars). This gives the additional benefit of telling not only the prediction but the certainty.

It can be seen that the model for midtown at this timeslot is not well defined - the model has low certainty and indeed predicts the highest demand incorrectly. Conversely, the demand for JFK has much smaller error bars and higher certainty.

Conclusion

A taxi routing system based on this secondary-demand modeling software would allow smart planning of routes - taxis would be sent where they were likely to get fares that kept them busy in high-demand areas. Thus, downtime would be minimized.

A taxi which received a fare to go outside the optimal zone could quickly get back en-route after dropoff.

The fare modelling from previous questions could be used to augment this model, however the true power of this model is in the potential to keep cabs in high demand areas.

# Section 2: Personal Achievements

Inventor with 5 USPTO granted patents:
https://patents.google.com/?inventor=anthony+moriarty&oq=anthony+moriarty

Masters of Data Science thesis work:
Relating Social Media to Crime using NLP (in-progress)