NYC TAXI TRIP DATA

A Data Science Project

Aviv Lahat, Haya Riesel, Miriam Goldstein

The Hebrew University of Jerusalem
Introduction to Data Science Course

Project Description: In NYC, it's very likely that one's form of transportation will be in a yellow cab. Observing taxi traffic can provide city planners, citizens and taxi companies with information in diverse aspects, such as pricing, hourly and season patters, insights on where to direct taxi drivers and more.

We aim to analyze these aspects and draft a recommendation for city planners based on NYC taxi trip data that spans over several years.

Git link: https://github.com/a-lahat/NYC-Taxis.git

The Data: From the NYC OpenData (https://opendata.cityofnewyork.us/)

1. 2017-2020 Taxi Trips
   From each year we queried the necessary rows and columns for the specific questions. Each year is ~8 GB of data, about 100M taxi rides, where each row is a single ride and the corresponding columns are pick-up and drop-off time, passenger count, trip distance, etc.

2. NYC Taxi Zones – 4 KB GeoJSON file
   New York is divided into 263 zones, each for which the json file gives information on area, id, borough, geometry, etc.

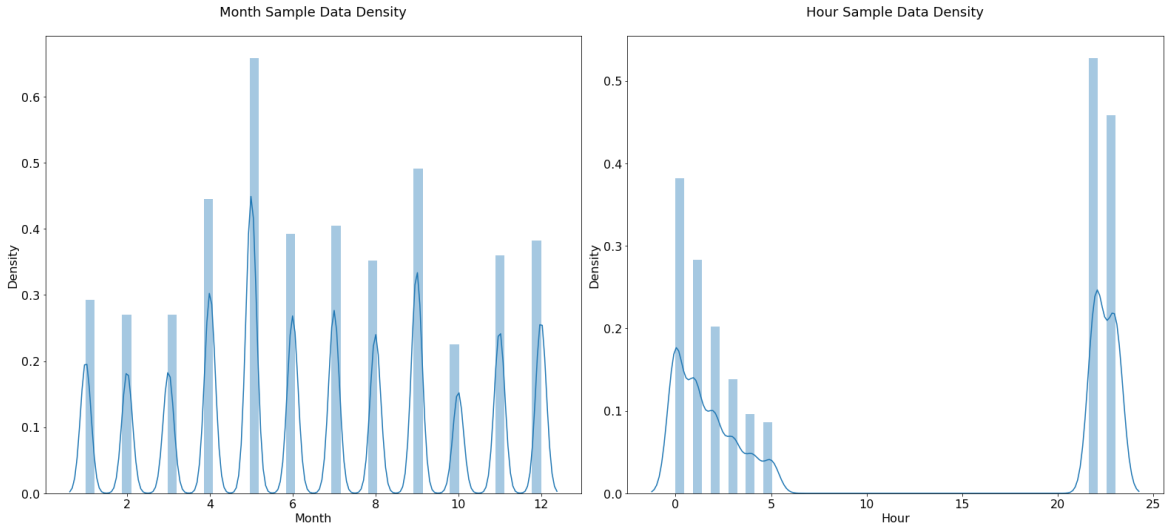
3. DOF: Summary of Neighborhood Sales by Neighborhood Citywide by Borough
   The Department of Finance (DOF) maintains records for all property sales in New York City, including sales of family homes in each borough. This list is a summary of neighborhood sales for Tax Class 1, 2 and 3 Family homes.
   About ~6,500 rows of real estate sales by year and neighborhood

The Solution:

Part 1: Hey Party People!

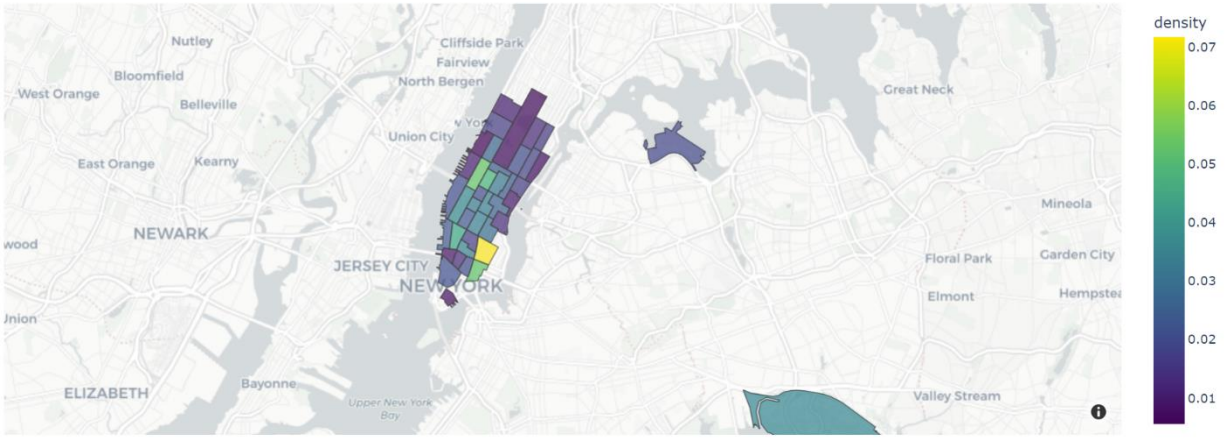
When planning new neighborhoods in the city, city planners and different groups of citizens like families, young couples, students, etc. want to know where they should hunt for housing. Having information on the night life is a crucial consideration. To visualize this question, we mapped out taxi pick-up hot spots from 10pm to 5am, which we label as night-life hours.

For this part, we sampled 10M out of 83M records from the taxi trips taken in 2019. Sanity check: sample data spreads across all months in the hour slots that we defined.



The interactive map shows areas in NYC with at least 50K taxi pick-ups during night life hours, and is colored by how popular it is for a taxi to pick-up riders from an area in NYC compared to all other areas.

Night life pick-up locations density map



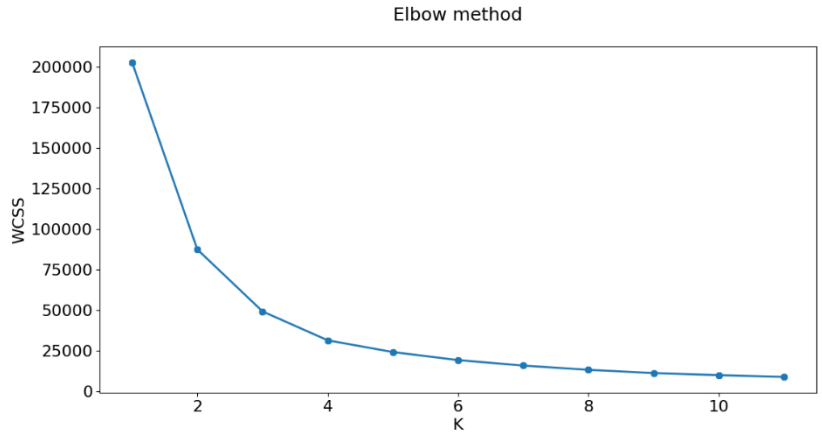
Interactive map link: https://www.cs.huji.ac.il/w~a_lahat1/NYCTaxi.html

For this next part we will focus on neighborhoods in Manhattan with more than 50K taxi pick-ups in this time frame. From our data, if you want to live in the center of night life, the recommendation is that you should lean towards East Village, Clinton East or Lower East Side where 20% of taxi rides started here during the night-life hours. Otherwise, for example, if you're a family with children, you'd want something quieter and the recommended areas are Upper East Side or Clinton West where only 2% of taxi rides were there throughout the night-life hours.

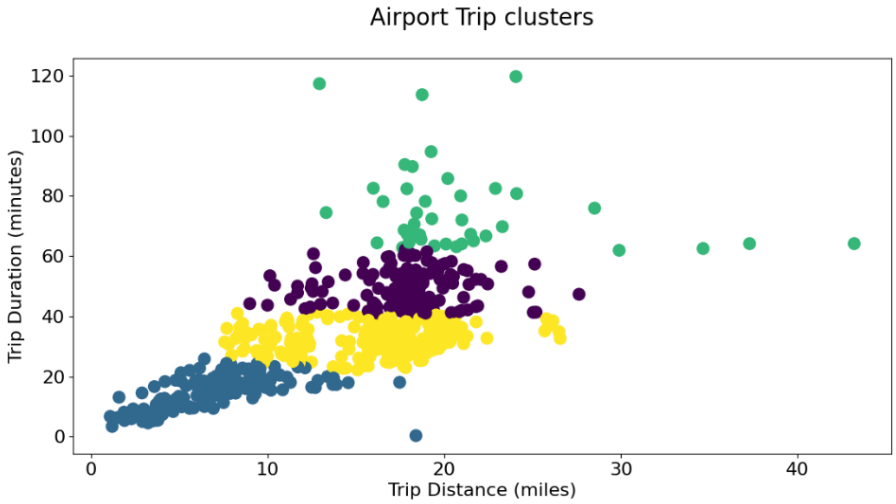
Part 2: Where should I land?

We want to use clustering in order to test if the airport a person lands in will affect other factors of their trip home, for example, duration, distance, time of day etc. We took a representative sample of trips home from JFK, Newark and LaGuardia Airport by sampling 20 rides in each hour of the day. We defined two features to cluster upon - trip duration and trip distance, and checked what other features are similar in each resulting clustered group.

First, to find the best k to cluster by, we used the Elbow method and chose k to be 4.

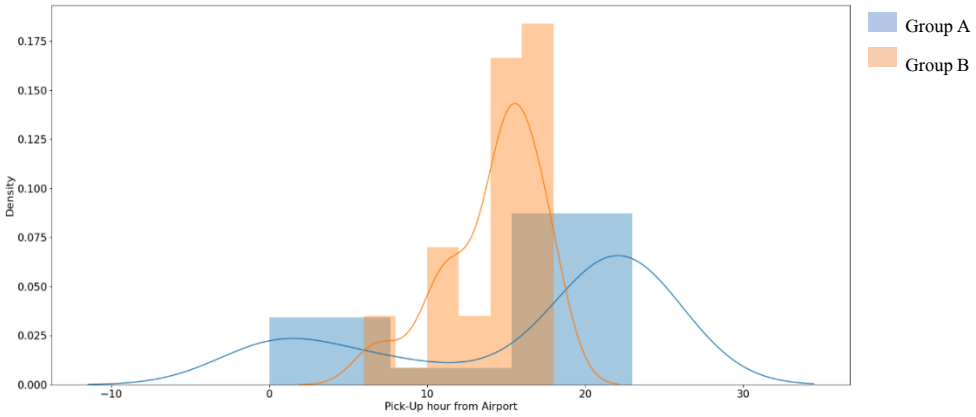


Next, we used K-Means algorithm for the clustering.



One feature that correlates nicely with the resulting clusters is the hour of the taxi trip. Let us define group A as the taxi trips with a trip duration between 20 and 40 minutes, and group B as the trips with a trip duration between 40 and 60 minutes. Focusing on trip distances between 15 and 20 miles, we can see that group B was picked up from the airport mostly during rush hours between 17:00-20:00, and group A during other hours.

Pick-Up hours density plot for clusters 2 + 3

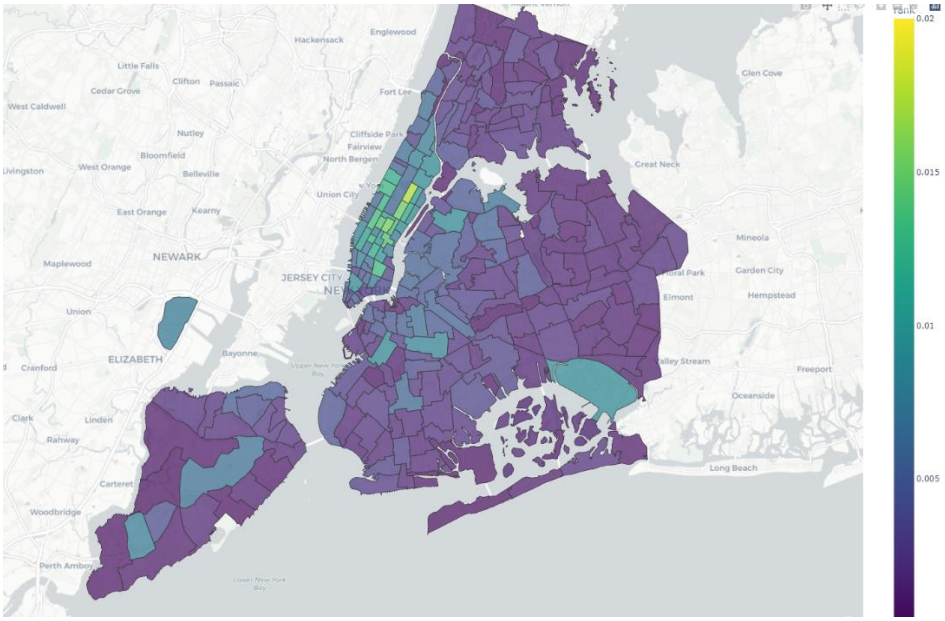


Part 3: PageRank to CityRank

Another approach we used to find the popular zones in New York City was with the PageRank algorithm.

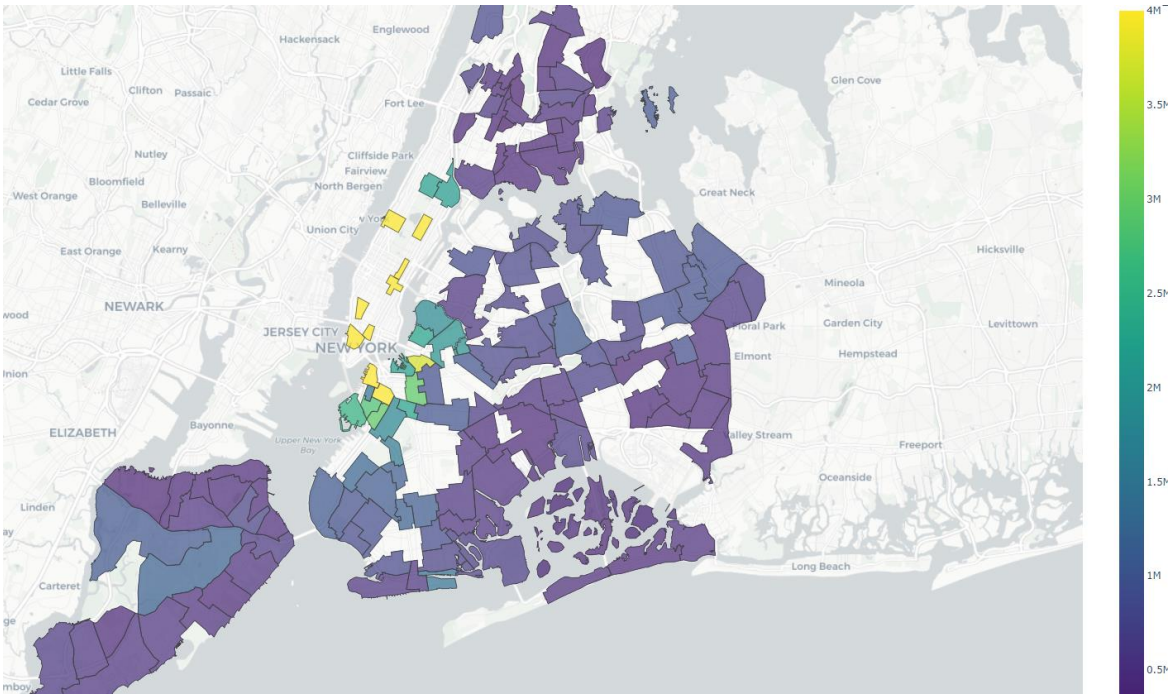
To find these areas, we revised the algorithm so that the nodes are the different zones as defined by the taxi data, and the directed and weighted edges are the percentage of rides between each two zones.

We ran the revised version, nicknamed CityRank on 5M taxi rides from 2017 and mapped out the algorithm output in this visualization.



To evaluate our ranking, we used a separate data base of NYC Real Estate and hypothesized that central zones in New York would have corresponding higher demand in real estate. We looked at the median price of a family estate in NYC and saw that there exists a correlation between the zones ranked highly by the CityRank algorithm and the Real Estate demand for that area.

This map visualizes the differences in median prices of each neighborhood:



The cross between the two sources databases was done by the names of the neighborhoods. Neighborhood names is a non-injective value and there can be a neighborhood that is listed slightly differently between the two databases. To do this we used a function that calculates the degree of match between two words and did a crossover only when the score crossed a certain threshold. The function works so that it calculates the edit distance.

For an addition evaluation, we also checked multiple years, and similar findings were found there as well.

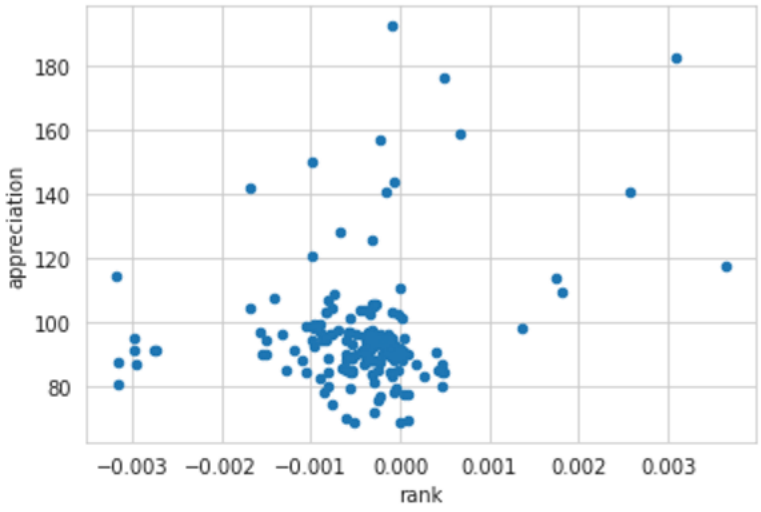
Continuing this train of thought, it is interesting to see if the areas where there is an increase in CityRank along the years are also increasing in the value of real estate in the area.

We calculated the CityRank of the years 2019 and 2017. We took 2 years distant enough years so that if a significant enough change occurs in that period, we would detect it.

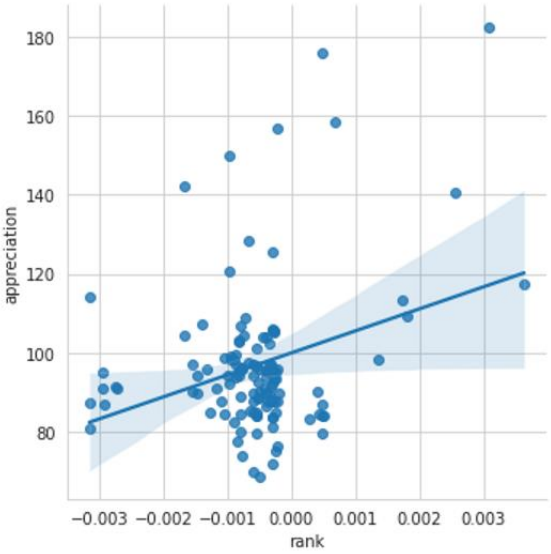
Since real estate has risen significantly over the years, we would like to calculate the percentage of increase in New York in general and treat it as an estimate of whether there has been an increase or decrease in real estate. (It is possible that the price of the property increased, but it increased less than the general rate of increase in real estate, so this is considered a decrease in the value of the real estate.) In addition, it will be seen whether the rank of the area decreased or not.

We calculated the appreciation (calculation of the percentage increase compared to the initial price in percent) and the difference between the ranks (the difference can be negative or positive). We have interpreted the points on the graph and indeed you can see the change:

(Note that most of the points with a positive ranking are above an appreciation of 80)



The neighborhoods whose ranking did not rise or fall do not interest us (because they may have been central enough before so there is no increase in rank, but in appreciation there will be a drastic increase because in central areas there is a very high increase in price). Let's just look at the points whose ranking went up or down to see what happened there:



Indeed, it can be seen that there is an uptrend in the graph as the rating rises.

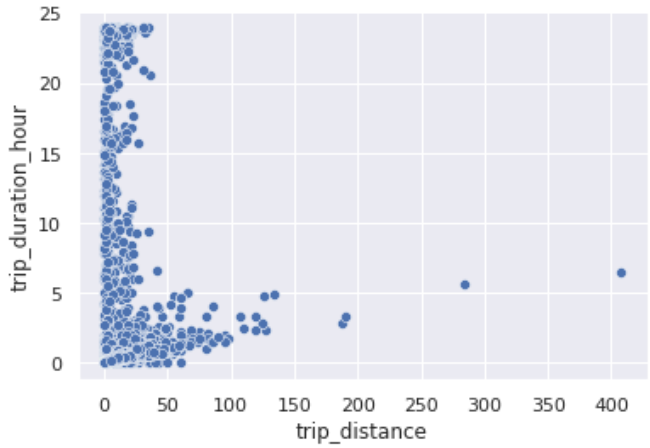
Part 4: Prediction of taxi trip duration in NYC using linear regression.

A problem a taxi service might face is efficiently assigning the cab drivers to passengers in a way that maximizes profit. One of the main issues is determining the duration of the current trip so it can predict when the cab will be free for the next trip. We will attempt to predict the duration of a taxi trip in NYC.

The data we used for this problem is only from December 2020 (this was for runtime optimization) however the model as is can be used on larger time frames.

Our implementation is building a machine learning model with using linear regression

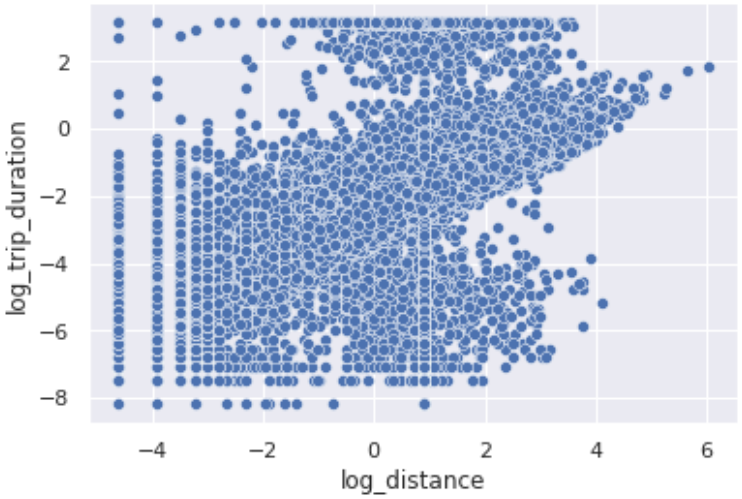
First, we preformed some analysis and transformation of some of the features to identify invalid data points. By plotting the distribution of this distance feature against the trip duration value.



We found several data points with values much beyond 200km and many values with a trip distance of 0 km. Possibly these could be rows depicting cancelled rides, however they cannot be part of our final model. To help us in identify data points where time of travel and the distance of the trip don't match up, we created a feature called speed.

By observing the distribution of the distance variable against the trip duration in hour feature we see several data points where the distance is < 50 km and the time taken to be >10 hours. This is very unlikely as the average speed is 7.5 km/hour [1]. We then use log transform on these columns and plot the distribution again.

[1] https://www.latimes.com/nation/la-na-new-york-traffic-manhattan-20180124-story.html#:~:text=The%20average%20speed%20of%20a,mph%20%E2%80%94%20barely%20faster%20than%20walking.



Here we see that the log transformed value of trip duration and distance has a somewhat linear relationship. But still there are some data points where the duration value does not change even with the change in distance. Therefor we decided to drop the rows beyond log_trip_duration > 2.
We cannot use all the features in the data to build our model as this would make the model too complex, so we created a new data frame and selected only the features which had some effect on the target variable - the trip duration. We dropped certain features as they were transformed to other features or not relevant. We assumed that some features have high correlation with other features and some are not correlated at all. Following this assumption, we first created a model with the mean of trip duration as the prediction. Then we created a base line model with all the features as is. Next, we will choose the features which are positively correlated with trip_duration and create the third model using only those features.

We split our data into 2 parts. The first part we used to train our data and the 2nd part we used for testing, with an 80: 20 ratio. We used K-Fold cross validation for base model and we used Root Means Squared Error as the evaluation metric.

Model Evaluation:

First, we created a model using the mean value as the predicted value for each test data point. This we used as a baseline method to compare to.

The second model uses only the distance feature as the only predictor columns and build a linear regression model. RMSE on Train Data: 392.4

CV Score: Mean - 392.1 | Std - 15.89 | Min - 359.9 | Max - 424.7

RMSE on Test Data: 387.9

The coefficient is [[133.04143287]]

As we can see, all the values are much less than the mean prediction RMSE, so our model using linear regression worked better than the baseline mean prediction.

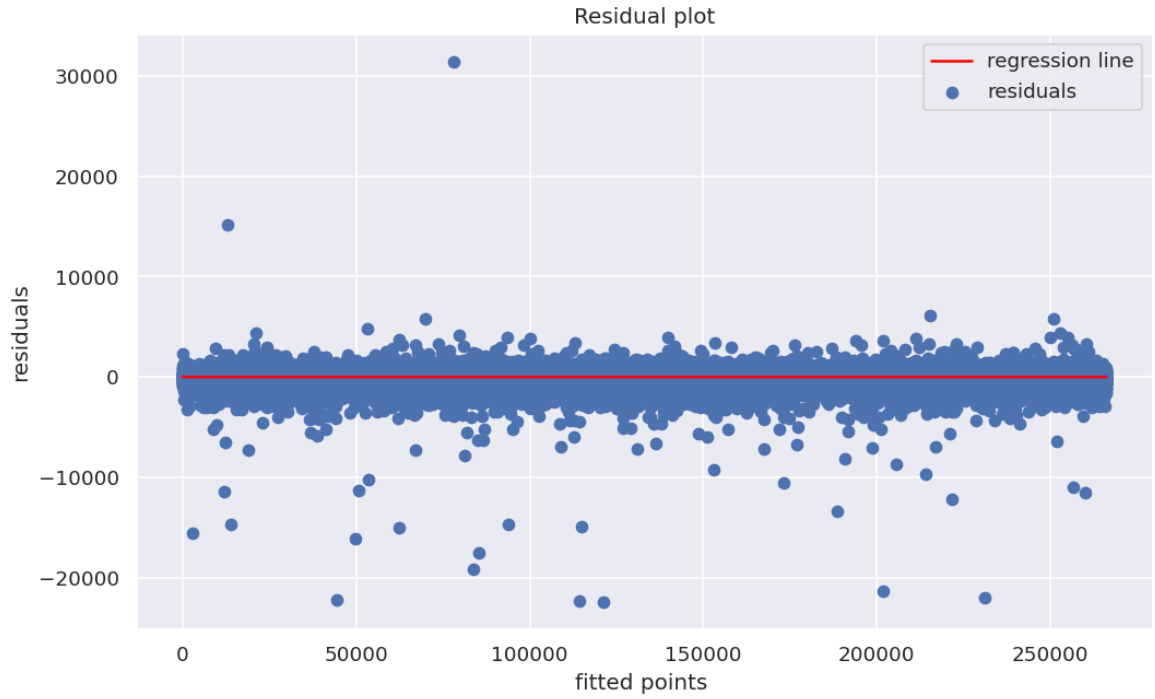
With Model 3 using all the features with positive correlation to the trip duration, we got even better results:

RMSE on Train Data: 390.2

CV Score: Mean - 390 | Std - 15.94 | Min - 357.6 | Max - 422.2

RMSE on Test Data: 386

Finally, we plotted the residuals and found the distribution of the residuals is Homoscedastic, meaning the assumption of Linear Regression holds true, as can be seen in the graph.



Evaluation: The main part of our evaluation was making sure our sample data was representative of the whole taxi trip data for the year we used in each part.

To ensure this, we created the distribution plots featured throughout the report that show the sample data spreads across the months, hours, days etc. needed for each different part of the project.

For example: in Part 1 we focused on night life hours between 10pm to 5am, and the corresponding distribution plot shows that indeed the data is taken from these hours and across all months of 2019.

Impediments: We would have wanted to run these algorithms on the whole collection of taxi data spanning across multiple years instead of using representative samples of a specific year for each part of the project. This would have required large cloud storage, therefor for the framework of this project we were content with samples and checked that they were representative of the population.

Model impediments: We encountered records in the data that did not add up, we found ways to clean up our data by filtering out passenger counts of 0 or more than 6, we discarded values of 0 distance drives and replaced drives that were impossibly long with the mean trip duration all off this improved performance of the model.

Future Work: The taxi trip data set has a number of features that can add useful data to a prediction model for NYC citizens. Incorporating it with other NYC data bases such as weather, population spreads and more would result in a very informative model.

Conclusion: In our project we were able to check multiple directions on the same data set. The database was trustworthy and mostly clean data. Some of our hypotheses were correct and proven throughout this project, and others failed. This was a great learning experience with GeoJSON files and visualizing through geographical maps. All in all, NYC is in good hands.