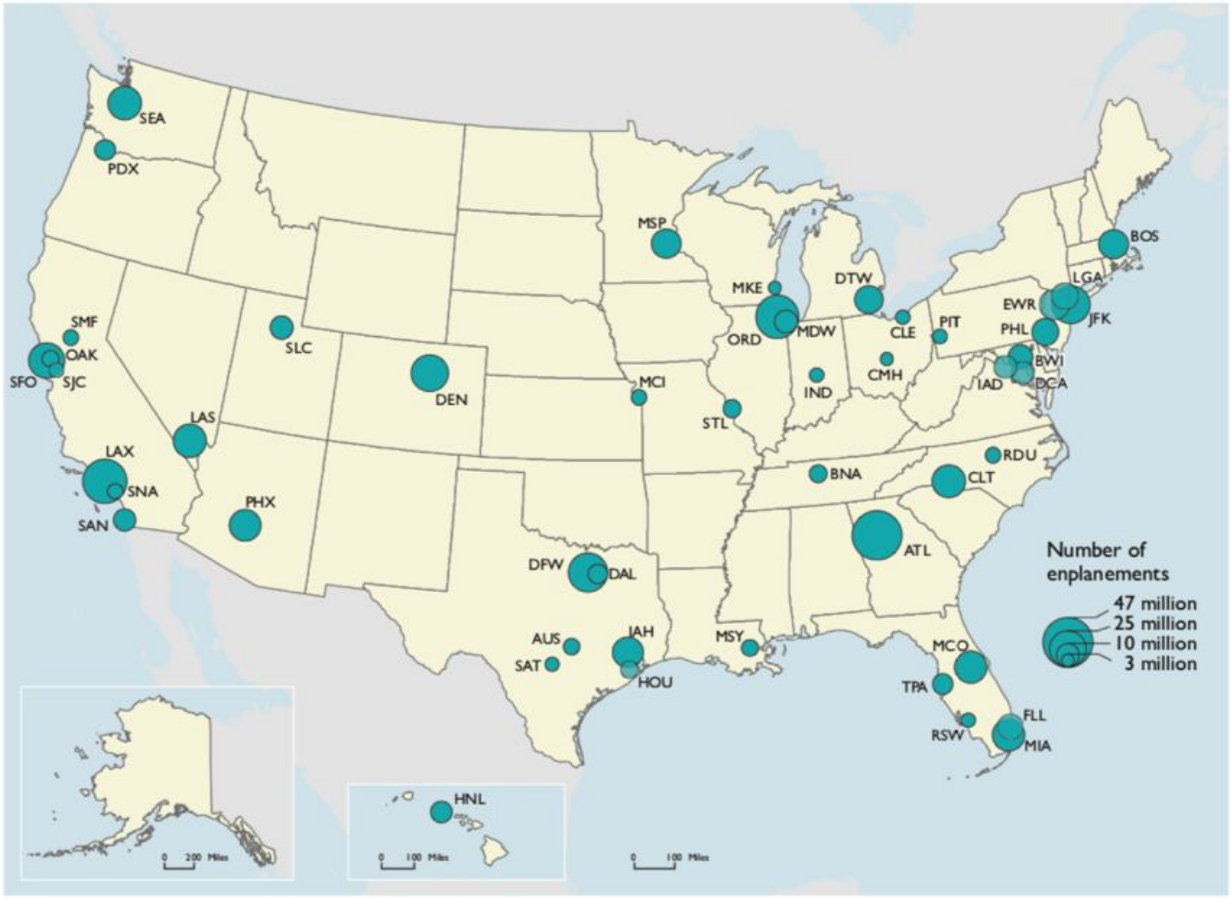
LGA Travelling Time Prediction

Xurui Chen(xc1454@nyu.edu), Yunhe Cui(yc3420@nyu.edu), Yuchen Ding(yd1402@nyu.edu)

Background and Introduction

Transportation planning for airport commute has gained increasing attention in recent years. According to the Transportation Statistics Annual Report 2017 (Dept of Transportation, 2018), US airports handled about 5.6 million commercial airline flights in 2016. Figure 1 shows the passenger boarding at the top 50 US airports in 2016. About 15 years ago, 65% of airport trips are made by private vehicles (Humphreys & Ison, 2005).



NOTES: Includes passengers enplaned on U.S. carrier scheduled domestic service.

SOURCE: U.S. Department of Transportation, Bureau of Transportation Statistics, Office of Airline Information, *T-100 Market Data*, available at www.transtats.bts.gov as of May 2017.

Figure 1. Passengers Enplaned on U.S. Flights at the Top 50 US Airports: 2016

However, things have been completely changed in recent years. In some large international airports (for example, SFO), more than 75% of airport trip is made through rideshare and taxi service (Marshall, 2019). In New York City, although there is a well-developed metro system, some people still prefer using service and taxi service because of taking a subway with luggage could be painful experiences. However, in metropolitans like New York City, traffic conditions

change quickly, and the traveling times from and to airports varies to a large extent. Depart early will enable the travelers to go to airport leisurely, but people might have to wait for a long period of time at the boarding gate. Moreover, we should admit that people(especially business people who need to fly frequently) generally do not have a long time budget for airport travel. Thus, predict and adjust traveling time accordingly becomes a need for people using ground transportation to commute between the city and airport.

In this study, we will mainly focus on trips from New York City boroughs to LaGuardia Airport (LGA). We will use machine learning models (Ridge, Random Forest, Neural network with MLP classifier and Gradient Boosting) to conduct a prediction of traveling time and the prediction result could help the travelers to better plan for their trips. The input data for our study are weather, yellow cab travel time data, traffic accident record (elaborations are in Data part).

Literature Review

We noticed that some researchers are interested in this topic, and a range of machine learning techniques are applied in those researches. In the literature, approaches for real-time travel time prediction had been proposed. However, travel time for the urban network is highly dependent on random fluctuation in travel demands, which makes it hard to predict. Lee, Tseng, and Tsai (2008) created a knowledge-based real-time travel time prediction model which contains real-time and historical travel time predictors. Using data mining and transforming to travel time prediction rules, they discovered traffic patterns from the raw data of location-based services.

Researches about urban taxi travel time have been mostly based on GPS data. The analysis conducted in 2013 by Gao, Zhu, Wan, and Wang focused on the urban travel time pattern (Gao, et. al, 2013). They proposed a method to recognize popular paths between origin-destination pairs and cluster the same trajectories together. They treated the travel time data as time series and used Weighted Moving Average (WMA) to extract the normal travel time pattern. The model is validated through the taxi GPS data of Beijing. The study focused on large scale travel time estimation (Zhan, Hasan, Ukkusuri, and Kamga, 2013) used GPS taxi data from New York City, which provides locations of origins and destinations, travel times, fares and other information of taxi trips. The Levenberg–Marquardt (LM) method is used to evaluate a test network from Midtown Manhattan and the result proved that they can efficiently estimate hourly average link travel times.

The paper discussed taxi time prediction at Charlotte airport (Lee, Malik, Zhang, Nagarajan, and Jung, 2015) used Linear Optimized Sequencing, a discrete-event fast-time simulation, and a data-driven analytical method with machine learning techniques. The researchers also discussed how the operational complexity at Charlotte Airport affects prediction accuracy and improved this model before applying to an airport scheduling algorithm in a real-time environment.

Data and Feature Engineering

Data name, data source link, and the data glance are shown in Appendix A while the processed data is shown in Appendix B. For this project, we used taxi zone shapefile as our base data. That is, other spatial-related data (weather, taxi, accident) will be merged/spatial joined with it and our data output will be a single chart with location-id serves as an index.

As for NYC weather data, we selected hourly humidity(%RH), temperature(℉), station pressure (Hg), visibility (mile), and average wind speed (mph) from the raw dataset. However,

there is no wind speed record after 13 Oct 2018 in our primary weather dataset, so we found the second wind speed dataset and merged them together. Other than that, there are still some sporadic data missing in hourly weather record rows, we decided to fill in the data with the 1-hour prior record rather than the average daily record because the weather change is much closely related to the nearby hours than to the daily average.

We filtered the trips with LGA as a destination based on the location ID of LGA region (taxi zone = 138) from taxi data (refer to Yellow Cab Travel Time data) and only keep columns with pickup time, drop-off time, trip distance, and pick up location ID information from the 17 columns of the original file. Then we removed the outliers and abnormal values to smooth our data. According to the glance of data (Figure 2), we defined the research trip time range as 5-150 min and distance range as 1-25 miles and removed data that fell out of either of these two ranges. We also applied a spatial join between traffic accident data and taxi zone shapefiles. A choropleth map is attached below (Figure 3). We could see that Queens and Brooklyn has a much larger number of car accidents. The frequent accident will possibly lead to delays in travel time. To simplify the future analysis, we create Day ID (0~364, int) and hour ID (0~8759, int) based on the pick-up time of each trip rather than using timestamp. The last three taxi travel features we created were the average the trip time for 1 hour/2 hour/1 day prior to pick-up time for a certain location id, and finally, we merged the averages with taxi travel data based on the location id.

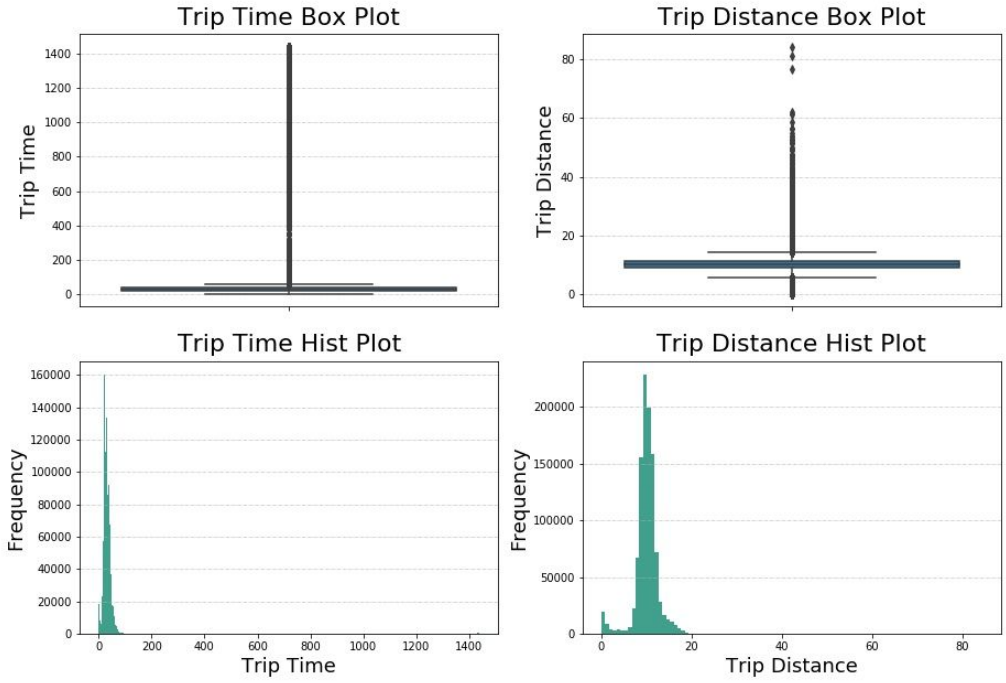


Figure 2. Histogram and box plots of taxi trip time and distance

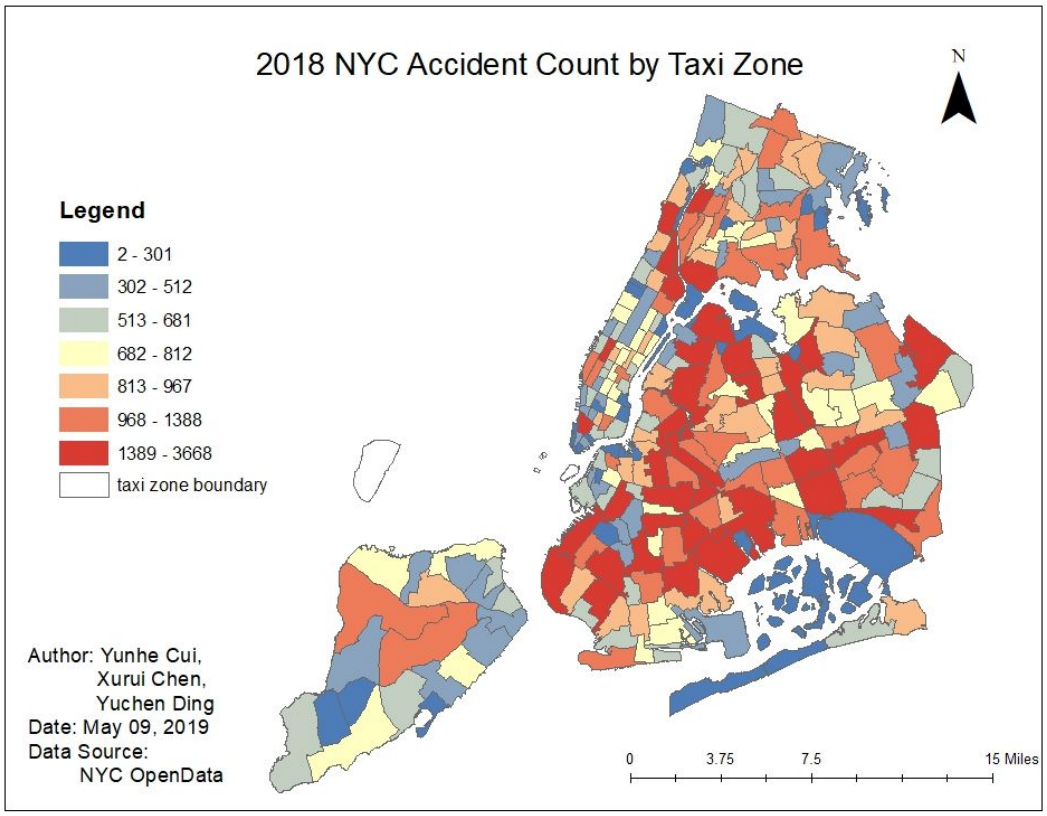


Figure 3. 2018 car accident count by taxi zone, NYC

We then used the holiday data as a reference file for taxi data to create the holiday-related feature for each trip. Based on the pick-up time, we created a series of binary variables to see whether this trip happened in weekdays (Mon-Fri)/ Friday/holiday/1 day prior to holiday/ 1 day after the holiday. Those variables enable us to further probe in the holiday/weekend.

In terms of vehicle accident data, we only kept date, latitude, and longitude for each accident and neglected other more detailed descriptive columns. After cleaning NA values, we spatial joined with taxi zone shapefile to getting location id reference and counted the accident for each location. We also created a binary column to indicate whether or not one location has accidents for each day. Finally, we merged the aggregated vehicle accident data with taxi data and output our processed dataset (data sample is shown in Appendix B).

Model

For our project, we used four models: Ridge Regression Model, Random Forest, Gradient Tree Boosting and Neural network (Multi-Layer Perceptron).

Ridge Regression Model

When we using least squares to calculate the regression model, if there is high multicollinearity between features, the least square will be extremely sensitive to the noise in input data. Moreover, since the formula for linear regression is $y = w^T x$, the parameters will

become extremely large if the multicollinearity exists and a minor change in the input variable will lead to a large change in y value. To limit the parameter w, a punishment parameter is added to the original function. This parameter is called L2 regularization-loss function (Function 1) and α, in this case, is a parameter which should be 0 or larger.

$$J_w = \min_w \{||Xw - y||^2 + \alpha ||w||^2\}$$

Function 1, L2 Regularization-loss function

Random Forest

Random forest is an ensemble learning method using bootstrap aggregating (also called bagging). It learns a large set of models, decision trees, which not only selects a random sample with replacement from the training set but also uses a random subset of features. And the final prediction is an unweighted average of each individual predictors. Also, this sample and feature bagging method could alleviate the multicollinearity problem of the dataset, which is a more suitable method of our dataset. Even random forest is generally a more accurate model, it lacks interpretability, requires a longer time to tune parameters (computationally expensive) and needs a large dataset to train this model.

Gradient Tree Boosting

Gradient Tree Boosting is an ensemble learning method which produces a prediction model based on the ensemble of decision trees. It fits the model to the residuals left by fitting all previous models. The Gradient Boosting method is skilled in solving data unbalance problem as it focuses step by step on training data and strengthens the impact of the minority class (Ravanshad, 2018). However, it is hard to tune parameters and takes even longer time than Random Forest to converge. Besides, it is more sensitive to outliers and abnormal.

Neural Network (Multi-Layer Perceptron):

Multi-layer Perceptron (MLP) is a feedforward artificial neural network consisting of at least three layers of nodes: an input layer, a hidden layer, and an output layer. The hidden layer of MLP shows below. MLP uses backpropagation to train the neural network model (Figure 4).

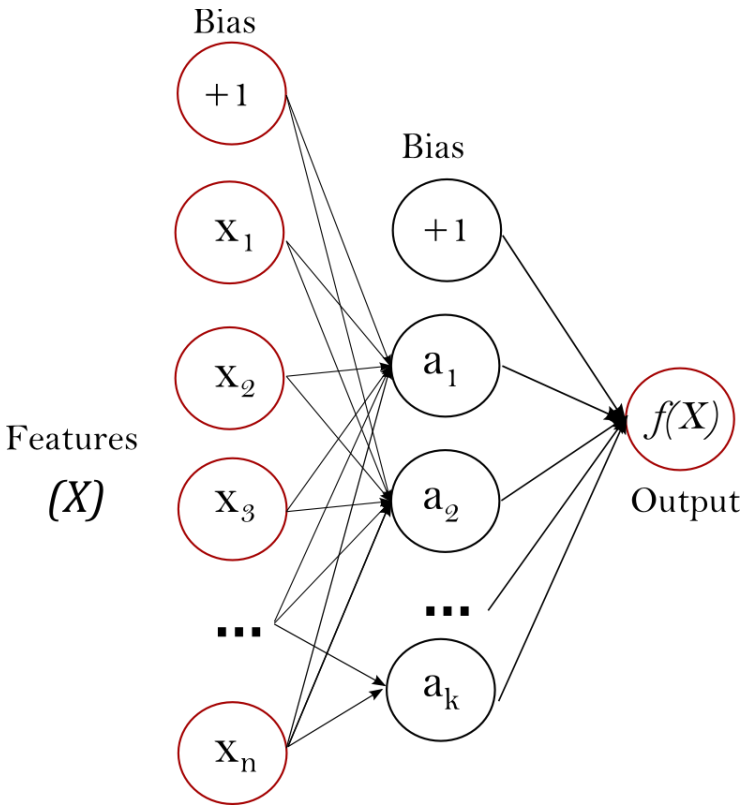


Figure 4. One hidden layer MLP (scikit learn, n.a)

Model Choice Discussion, Evaluation and Conclusion

Feature multicollinearity will influence the regression model performance as it might result in very sensitive coefficients, which will ruin the model robustness. The multicollinearity in our data is shown in Figure 5. From the correlation matrix, we could see that the features that have strong positive/negative correlation are correlated in daily life, for example, the average trip times one hour prior is correlated to two-hour ones since it is the travel time for the same route with minor changes caused by other features.

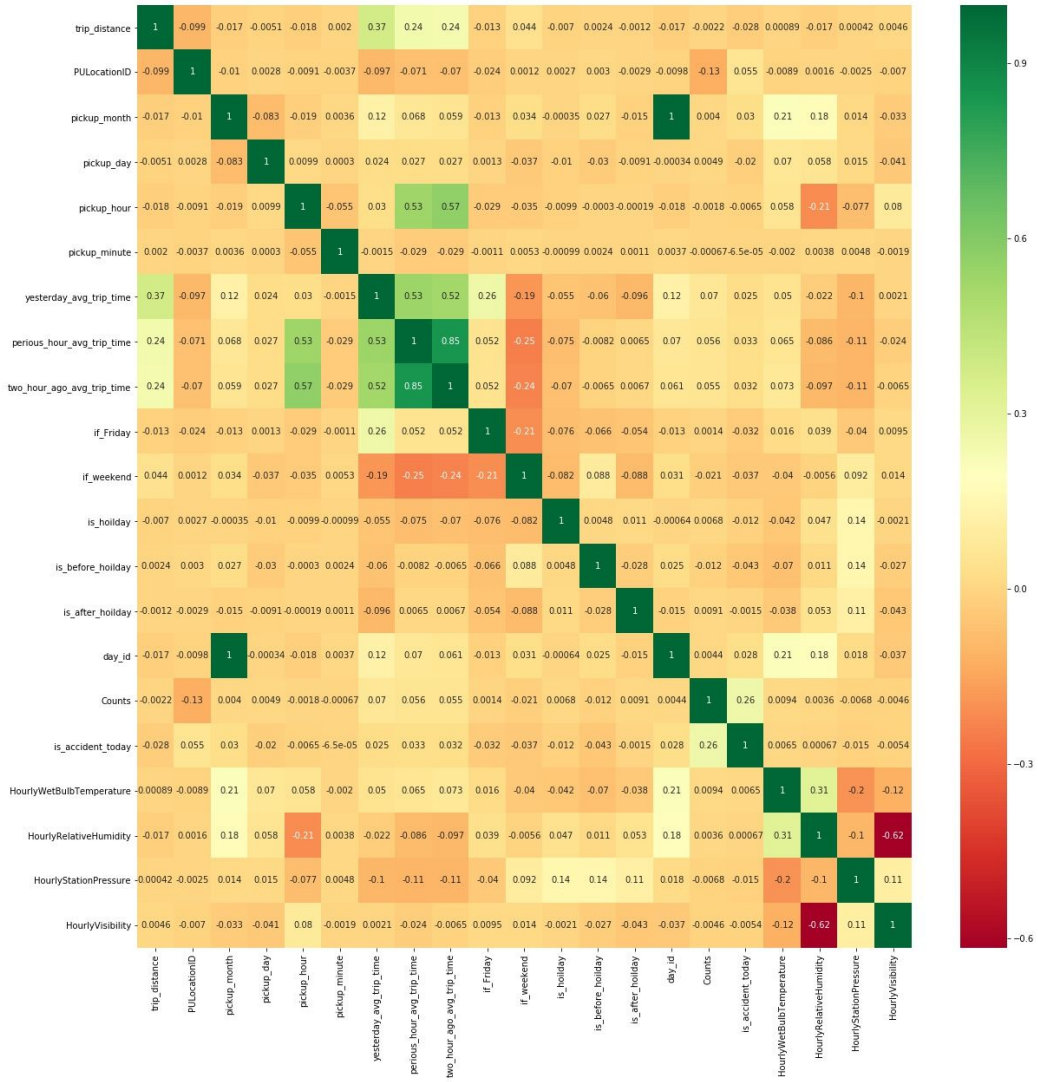


Figure 5. Feature correlation matrix

Principal Component Analysis (PCA) is a widely used statistical procedure in exploratory data analysis and for predictive model constructions to reduce multicollinearity. However, in our project, we have plenty of dummy features and the application of PCA will lead to loss of feature meaning and even result in a worse model prediction output. So that we decided to modify our model choice and chose the one that will be slightly influenced by multicollinearity. We measured the model performance using runtime, Mean Squared Error(MSE) and R^2 . The formulae of those two are shown in Function 2a and 2b.

$$MSE = \frac{SSE}{n} = \frac{1}{n}\sum_{i=1}^{m} w_i(y_i - \hat{y}_i)^2$$

Function 2a

$$R^2 = 1 - \frac{\Sigma(y-\hat{y})^2}{\Sigma(y-\bar{y})^2}$$

Function 2b

Model	Runtime	MSE	R^2
Ridge Regression	0.701670919	62.91962761	0.21412059
Random Forest	115.1541575	15.22999904	0.877939582
Gradient Tree Boosting	400.1240345	52.93717812	0.379856958
Neural Network (MLP)	7001.793479	50.75832576	0.431815082

Table 1, Run time, MSE and R^2 for four project models

The model performance features are shown in Table 1 above. We could notice that Random Forest performs far better than other models. The reasonable explanation is that, GBTR is very sensitive to abnormal values and our dataset inevitably has extreme values (such as extreme weather, extreme high holiday traffic, etc) and Neural Network has high computational and time complexity, the limited model tuning we finished might not lead to the best model, so the model performance is worse than Random Forest.

In conclusion, from both operation time and prediction quality perspective, Random Forest is the most "economical" regression prediction model with a significantly low MSE (15.23000) and much higher R^2 (0.87794). Moreover, due to short operation time, we could tune more parameters in Random Forest get a better model performance within the same time limit.

Discussion and Future Work

Although we got good prediction outcome for most of our models (within ± 5 minutes error), we should admit that, compared to the real-world situation, our model is oversimplified and our datasets are biased to some extent. The first concern is that we used Yellow Taxi data as our major object of analysis. However, as the booming of the transportation network companies such as Uber and Lyft, increasing numbers of travellers will choose to request a ride with them rather than take a taxi ride, thus, travelling analysis with taxi data is no longer enough for us to sketch the (non-bus, individual) ground transportation picture and conduct related analysis. In this model, we merely considered the count of accidents in terms of traffic condition change. However, the influence of different accidents on traffic is obviously not identical. The time of the accident happened, the severity level of traffic accident and length of solving time all have an impact on the level of accident influence. Moreover, the change in pedestrian number could also influence the traffic condition, for example, more pedestrian across the street at a stop sign will lead to longer passing time for vehicles and less smooth traffic in general. In the future, with Uber/Lyft trip data(with no sensitive traveler data), using much detailed factor analysis, we could give more practical and accurate predictions.

From the technical aspect, the high computational complexity of models is a challenge for us when we constructed, trained and tuned our model. Also, the sklearn package is not intended for large-scale applications. So when we use larger dataset and apply more complex models later, we should rewrite the model in pyspark and use High Performance Computer (HPC) to shorten operation time and enable parallel processing so that it will be more efficient in adjusting parameters.

Collaboration

All group members contributed to the whole process of this project, but each of us is more concentrated in one part of the project. Yunhe was responsible for data collection and report writing, Xurui was leading in data cleaning, feature engineering, and modeling while Yuchen contributed more on literature review, report polishing and presentation preparation.

References

Gao. Mengdan, Zhu. Tongyu, Wan. Xuejin, and Wang. Qi, Analysis of Travel Time Patterns in Urban Using Taxi GPS Data, August. 2013, Retrieved May 08. 2019, from https://ieeexplore.ieee.org/abstract/document/6682115

Humphreys, I., Ison, S., Changing Airport Employee Travel Behavior: The Role of Airport Surface Access Strategies. In Transport Policy, 2005, Vol.12, pp. 1-9.

Lee. Hanbong, Malik. Waqar, Zhang. Bo, Nagarajan. Balaji, and Jung. C. Yoon, Taxi time prediction at Charlotte Airport using fast-time simulation and machine learning techniques, June 18, 2015, Retrieved May 08, 2019, from https://arc.aiaa.org/doi/pdf/10.2514/6.2015-2272

Lee. Wei-Hsun, Tseng. Shian-Shyong, and Tsai. Sheng-Han, A knowledge based real-time travel time prediction system for urban network, April 2009, Vol. 36, Issue 3, Part 1, Retrieved May 08. 2019, from https://www.sciencedirect.com/science/article/pii/S0957417408001875

Marshall. Aarian, Airports cracked Uber and Lyft – Time for cities to take note, November 11, 2018, Retrieved April 26th, 2019, from https://www.wired.com/story/uber-lyft-ride-hail-airport-traffic-cities/

Ravanshad. Abolfazl, Gradient boosting vs Random Forest, April 27th. 2018, Retrieved May 07. 2019, from https://medium.com/@aravanshad/gradient-boosting-versus-random-forest-cfa3fa8f0d80

U.S. Department of Transportation, Bureau of Transportation Statistics, 2018, Transportation Statistics Annual Report 2017, Retrieved April 26th, 2019, from https://www.bts.dot.gov/sites/bts.dot.gov/files/docs/browse-statistical-products-and-data/transportation-statistics-annual-reports/215041/tsar-2017-rev-2-5-18-full-layout.pdf

Scikit learn, 1.17.1 Neural network models (supervised), n.a, Retrieved May 08. 2019, from https://scikit-learn.org/stable/modules/neural_networks_supervised.html

Zhan. Xianyuan, Hasan. Samiul, Ukkusuri. Satish, and Kamga. Camille, Urban link travel time estimation using large-scale taxi data with partial information, August 2013, Vo1 33, pp 37-49, Retrieved May 08. 2019, from https://www.sciencedirect.com/science/article/pii/S0968090X13000740

Appendix

A: Data source and raw data glance

Data Name and Format	Data source
Weather (2018); CSV	https://openweathermap.org API https://www.wunderground.com/history/daily/us/ny/new-york-city/KLGA
Yellow Cab Travel Time (2018); CSV	https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Holiday (2018); CSV	https://bsc.ogs.ny.gov/sites/default/files/BSC_Announcement_NYS_Holiday_Announcement_2017_2018.pdf
Traffic Accident (2018); CSV	https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95 API
Taxi zone NYC; shapefile	https://data.cityofnewyork.us/Transportation/NYC-Taxi-Zones/d3c5-ddgc

Table 2. Data source

VendorID	tpep_pickup_datetim	tpep_dropoff_dateti	passenger	trip_distan	RatecodeI	store_and_	PULocatior	DOLocatio	payment_t	fare_amou	extra	mta_tax	tip_amoun	tolls_amou	improvem	total_amount
2	2018/1/1 0:57	2018/1/1 1:43	2	9.29	5	N	100	138	1	70	0	0	0	0	0.3	70.3
1	2018/1/1 0:26	2018/1/1 0:51	2	9.5	1	N	161	138	2	30	0.5	0.5	0	0	0.3	31.3
2	2018/1/1 0:53	2018/1/1 1:07	6	11.36	1	N	132	138	1	30.5	0.5	0.5	0.02	0	0.3	31.82
2	2017/12/31 23:29	2017/12/31 23:34	1	1.29	1	N	138	138	2	6	0.5	0.5	0	0	0.3	7.3
1	2018/1/1 0:41	2018/1/1 1:18	1	10.5	1	N	170	138	2	36.5	0.5	0.5	0	0	0.3	37.8
1	2018/1/1 0:10	2018/1/1 0:49	1	21.3	1	N	68	138	3	57.5	0.5	0.5	0	5.76	0.3	64.56
1	2018/1/1 1:18	2018/1/1 1:20	1	1	5	N	138	138	1	127	0	0	0	10.5	0.3	137.8
2	2018/1/1 1:18	2018/1/1 1:40	3	9.81	1	N	161	138	2	29.5	0.5	0.5	0	5.76	0.3	36.56
1	2018/1/1 1:44	2018/1/1 2:12	1	10.5	1	N	48	138	1	32.5	0.5	0.5	7.9	5.76	0.3	47.46
2	2018/1/1 2:29	2018/1/1 2:55	1	10.01	1	N	186	138	1	30.5	0.5	0.5	4.08	5.76	0.3	41.64
1	2018/1/1 2:23	2018/1/1 2:37	2	8.1	1	N	262	138	1	23	0.5	0.5	0.46	5.76	0.3	30.52
1	2018/1/1 2:50	2018/1/1 3:17	1	11	1	N	161	138	2	33	0.5	0.5	0	5.76	0.3	40.06
2	2018/1/1 2:48	2018/1/1 3:12	1	11.25	1	N	230	138	1	33.5	0.5	0.5	10.14	5.76	0.3	50.7
2	2018/1/1 2:41	2018/1/1 3:06	1	10.51	1	N	230	138	1	31	0.5	0.5	0	5.76	0.3	38.06
2	2018/1/1 2:40	2018/1/1 2:58	5	7.94	1	N	262	138	2	24	0.5	0.5	0	5.76	0.3	31.06
2	2018/1/1 2:48	2018/1/1 3:10	1	16.13	1	N	261	138	1	43	0.5	0.5	9.97	5.54	0.3	59.81
2	2018/1/1 2:05	2018/1/1 2:30	1	9.63	1	N	186	138	1	28.5	0.5	0.5	2	5.76	0.3	37.56
2	2018/1/1 2:34	2018/1/1 2:54	1	11.56	1	N	45	138	1	32.5	0.5	0.5	6.76	0	0.3	40.56
1	2018/1/1 2:36	2018/1/1 2:54	1	9	1	N	162	138	1	25.5	0.5	0.5	37	5.76	0.3	69.56
1	2018/1/1 3:54	2018/1/1 4:20	1	11.6	1	N	45	138	2	33.5	0.5	0.5	0	0	0.3	34.8
2	2018/1/1 3:35	2018/1/1 4:03	6	10.75	1	N	100	138	2	32	0.5	0.5	0	0	0.3	33.3
1	2018/1/1 3:55	2018/1/1 4:09	1	5	1	N	7	138	2	17	0.5	0.5	0	0	0.3	18.3
2	2018/1/1 3:47	2018/1/1 4:05	2	9.84	1	N	229	138	2	27	0.5	0.5	0	5.76	0.3	34.06
1	2018/1/1 3:50	2018/1/1 4:05	1	7.4	1	N	263	138	1	22	0.5	0.5	8.7	5.76	0.3	37.76

Figure 6a. Raw Weather Data

STATION	DATE	REPORT_TY	SOURCE	HourlyAltir	HourlyDew	HourlyDry	HourlyPrec	HourlyPres	HourlyPres	HourlyPres	HourlyRela	HourlySeal	HourlySky	HourlyStat	HourlyVisi	HourlyWet	HourlyWin	HourlyWin	HourlyWin	REM	REPORT_TY	SOURCE 1
7.25E+10	2018-01-0	FM-15	7	30.33	-4	9	0		0.01	5	55	30.31	CLR:00	30.16	10	7	300	21	9	MET11501	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.34	-4	9	0				55	30.32	CLR:00	30.17	10	7	290	18	10	MET09701	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.35	-4	8	0				57	30.33	CLR:00	30.18	10	6	VRB	20	10	MET09701	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.35	-5	8	0		-0.02	1	55	30.33	CLR:00	30.18	10	6	310		7	MET10001	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.37	-5	8	0				55	30.35	CLR:00	30.2	10	6	VRB	21	9	MET10401	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.37	-5	7	0				57	30.35	CLR:00	30.2	10	5	VRB		6	MET09401	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.38	-5	7	0		-0.02	1	57	30.36	CLR:00	30.21	10	5	300		8	MET11201	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.38	-4	7	0				60	30.36	CLR:00	30.21	10	5	VRB	17	6	MET09701	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.39	-3	8	0				60	30.37	CLR:00	30.22	10	6	290		6	MET09401	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.4	-3	11	0		-0.03	3	53	30.39	CLR:00	30.23	10	8	VRB		8	MET10001	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.39	-4	12	0				48	30.37	CLR:00	30.22	10	9	310	20	9	MET09701	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.38	-4	14	0				44	30.36	CLR:00	30.21	10	11	300		8	MET09401	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.36	-5	15	0		0.05	8	40	30.34	CLR:00	30.19	10	11	300	21	10	MET11501	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.34	-5	17	0				37	30.32	CLR:00	30.17	10	13	VRB	17	9	MET10101	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.35	-5	18	0				35	30.33	CLR:00	30.18	10	13	VRB	20	3	MET09701	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.34	-6	18	0		0.01	6	34	30.32	CLR:00	30.17	10	13	310	18	9	MET10301	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.35	-5	18	0				35	30.33	CLR:00	30.18	10	13	VRB		7	MET09401	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.36	-6	18	0				34	30.34	CLR:00	30.19	10	13	290		9	MET09401	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.37	-4	17	0		-0.03	3	39	30.36	CLR:00	30.2	10	13	VRB		3	MET11201	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.39	-4	17	0				39	30.37	CLR:00	30.22	10	13	VRB	18	8	MET10401	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.36	-2	17	0				43	30.35	CLR:00	30.19	10	13	VRB		5	MET09401	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.36	0	17	0		0.01	8	47	30.34	CLR:00	30.19	10	13	270	20	10	MET10301	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.36	2	16	0				54	30.35	CLR:00	30.19	10	13	VRB	23	11	MET09701	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.36	3	15	0				59	30.35	CLR:00	30.19	10	12	280		3	MET10401	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.37	2	15	0		-0.01	3	56	30.35	CLR:00	30.2	10	12	280		10	MET11201	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.38	1	15	0				53	30.37	CLR:00	30.21	10	12	290	22	15	MET10801	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.39	1	14	0				56	30.37	CLR:00	30.22	10	11	290	22	13	MET10801	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.39	1	14	0		-0.02	0	56	30.37	CLR:00	30.22	10	11	280	20	11	MET11001	FM-15	7
7.25E+10	2018-01-0	FM-15	7	30.39	1	13	0				59	30.37	CLR:00	30.22	10	10	VRB		5	MET10101	FM-15	7

Figure 6b. Raw Yellow Cab Travel Data

borough	contributin	contributin	contributin	contributin	contributin	cross_stree	date	latitude	location	longitude	number_of	number_of	number_of	number_of	number_of	number_of	number_of	number_of	off_street_r	on_street_r	time	unique_key	vehicle_typ	vehicle_typ	vehicle_typ	vehicle_typ	vehicle_typ	zip_code
MANHATT	Other Vehicular						2018-06-18T00:00:00.000				0	0	0	0	0	0	0	0	EAST 47 S	LEXINGTO	8:59	4026083	Taxi					10017
	Passing or	Unspecified					2018-06-29T00:00:00.000				0	0	0	0	0	0	0	0		HARLEM R	10:10	4026115	Sedan					
	View Obst	Unspecified					2018-08-16T00:00:00.000				0	0	0	0	0	0	0	0			0:35	4026257	Tractor Tru	Tractor Truck Diesel				
BROOKLYN						351 G	2018-09-1	40.65239	{u'type': u'	-73.8681	0	0	0	0	0	0	0	0			13:12	4026324						11239
	Driver Inat	Unspecified					2018-08-2	40.77077	{u'type': u'	-73.9173	1	0	0	0	0	0	1	0	31 STREET	HOYT AVE	11:25	3969187	Bike	Sedan				
	Failure to '	Unspecified					2018-11-1	40.62682	{u'type': u'	-73.9267	0	0	0	0	0	0	0	0		FLATLAND	9:30	4026658	Station Wagon/Sport Utility Vehicle					
BROOKLYN	Unsafe Lar	Unspecified					2018-11-1	40.67591	{u'type': u'	-73.8949	0	0	0	0	0	0	0	0	VERMONT	ATLANTIC	12:00	4026695	Pick-up Tr	Sedan				11207
	Following '	Unspecified					2018-11-13T00:00:00.000				0	0	0	0	0	0	0	0		STATEN IS	20:10	4026759	Station Wa	Sedan				
	Backing Ur	Unspecified					2018-11-15T00:00:00.000				0	0	0	0	0	0	0	0	WASHING'	CROSS BR	19:00	4026913	Sedan	Station Wagon/Sport Utility Vehicle				
BROOKLYN	Unspecified					574 V	2018-11-1	40.66533	{u'type': u'	-73.8885	0	0	0	0	0	0	0	0			8:15	4026963	Sedan					11207
	Unspecifie	Unspecified					2018-11-15T00:00:00.000				0	0	0	0	0	0	0	0	WESTMINI	AVENUE H	8:00	4027076	Bus	Sedan				
	Passing To	Unspecified					2018-11-16T00:00:00.000				0	0	0	0	0	0	0	0		WILLIAMSE	8:54	4027098	Box Truck	Box Truck				
	Following '	Unspecified					2018-11-14T00:00:00.000				0	0	0	0	0	0	0	0		HENRY HU	13:30	4027122	Sedan	Sedan				
	Pavement	Unspecified					2018-11-1	40.66579	{u'type': u'	-73.7464	0	0	0	0	0	0	0	0		BELT PARK	5:20	4027230	Sedan	Sedan				
	Unspecified						2018-11-1	40.66599	{u'type': u'	-73.9453	0	0	0	0	0	0	0	0		BROOKLYN	10:20	4027248	Sedan					
	Unspecified						2018-11-1	40.68108	{u'type': u'	-73.883	0	0	0	0	0	0	0	0		ESSEX STR	9:13	4027351	Sedan					
BRONX	Driver Inat	Unspecified				1491 W	2018-11-1	40.83677	{u'type': u'	-73.8615	0	0	0	0	0	0	0	0			12:35	4027375	Station Wa	Station Wagon/Sport Utility Vehicle				10462
BROOKLYN	Unspecified						2018-11-1	40.67241	{u'type': u'	-73.8636	0	0	0	0	0	0	0	0	BLAKE AVE	ELDERTS L	14:29	4027380	Sedan					11208
BROOKLYN	Unspecified					310 LE	2018-11-1	40.65421	{u'type': u'	-73.949	0	0	0	0	0	0	0	0			12:30	4027385	Sedan					11226
	Unspecifie	Unspecified					2018-11-15T00:00:00.000				0	0	0	0	0	0	0	0	BROADWA	VANCORT	23:35	4027376	Sedan	Sedan				
BRONX	Unspecifie	Unspecified				1656 C	2018-11-1	40.83915	{u'type': u'	-73.8527	0	0	0	0	0	0	0	0			17:20	4027381	Sedan	Sedan				10462
BROOKLYN	Reaction t	Unspecified				166 H	2018-11-1	40.67972	{u'type': u'	-73.9093	0	0	0	0	0	0	0	0			12:30	4027384	Station Wa	Station Wagon/Sport Utility Vehicle				11233
	Pavement Slippery						2018-11-1	40.86561	{u'type': u'	-73.8011	0	0	0	0	0	0	0	0		PARK DRIV	20:34	4027388	Station Wagon/Sport Utility Vehicle					
	Tire Failure	Unspecified					2018-11-1	40.78896	{u'type': u'	-73.8236	0	0	3	0	0	0	3	0		WHITESTC	13:15	4027377	Bus	Sedan				
	Pavement Slippery						2018-11-1	40.81916	{u'type': u'	-73.9604	0	0	0	0	0	0	0	0		HENRY HU	10:20	4027432	Pick-up Truck					
	Driver Inat	Unspecified					2018-11-1	40.58425	{u'type': u'	-73.9638	0	0	1	0	0	0	1	0		BELT PARK	16:30	4027476	Station Wa	Sedan				
QUEENS	Unspecified						2018-11-1	40.73835	{u'type': u'	-73.9027	0	0	1	0	0	0	1	0	LAUREL HI	63 STREET	19:55	4027632	Sedan					11377
BRONX	Fell Asleep	Unspecified				1939 E	2018-11-1	40.85027	{u'type': u'	-73.8321	0	0	0	0	0	0	0	0			8:22	4027783	Sedan	Sedan				10461
	Unspecified						2018-11-1	40.86268	{u'type': u'	-73.9091	0	0	0	0	0	0	0	0		WEST FOR	9:45	4027818	Sedan					
QUEENS	Unsafe Lar	Unspecified					2018-06-16T00:00:00.000				0	0	0	0	0	0	0	0	DALNY RC	HILLSIDE A	20:00	3922384	Sedan	Bus				11432
							2018-06-1	40.66217	{u'type': u'	-73.9537	0	0	0	0	1	0	1	0		LEFFERTS	1:20	3922387						
	Unspecifie	Unspecified					2018-11-17T00:00:00.000				0	0	4	1	0	0	4	1		MAJOR DE	6:54	4027917	Station Wa	Flat Bed				
MANHATT	Passing or	Backing Unsafely					2018-01-2	40.86932	{u'type': u'	-73.9167	0	0	0	0	0	0	0	0	WEST 214	BROADWA	16:30	3837742	PASSENGE	BU				10034
	Backing Unsafely					2236 F	2018-05-0	40.62694	{u'type': u'	-74.1643	0	0	0	0	1	0	1	0			13:00	3896150						
QUEENS	Unspecifie	Unspecified				255-07	2018-11-1	40.65513	{u'type': u'	-73.7339	0	0	0	0	0	0	0	0			16:20	4027972	Station Wa	Station Wagon/Sport Utility Vehicle				11422
MANHATT	Unspecified						2018-02-0	40.82265	{u'type': u'	-73.9513	0	0	0	0	0	0	0	0	HAMILTON	WEST 140	13:00	4073290	Sedan					10031

Figure 6c. Raw Traffic Accident Data

B: Processed data glance

trip_time	trip_distanc	PULocatior	pickup_mo	pickup_day	pickup_hou	pickup_mir	yesterday_	perious_ho	two_hour_a	if_Friday	if_weekend	is_hoilday	is_before_h	is_after_hol	day_id	Counts	is_accident	HourlyWet	HourlyRela	HourlyStati	HourlyVisib	HourlyWindSpeed
24	11.52	100	1	2	4	49	23.12121	21	17	0	0	0	0	1	1	609	0	10	59	30.22	10	5
21	8.99	170	1	2	4	41	19.73016	20	17	0	0	0	0	1	1	1287	1	10	59	30.22	10	5
21	8.99	170	1	2	4	41	19.73016	20	17	0	0	0	0	1	1	1287	1	10	59	30.22	10	5
21	8.99	170	1	2	4	41	19.73016	20	17	0	0	0	0	1	1	1287	1	10	59	30.22	10	5
15	8.67	170	1	2	4	38	19.73016	20	17	0	0	0	0	1	1	1287	1	10	59	30.22	10	5
15	8.67	170	1	2	4	38	19.73016	20	17	0	0	0	0	1	1	1287	1	10	59	30.22	10	5
15	8.67	170	1	2	4	38	19.73016	20	17	0	0	0	0	1	1	1287	1	10	59	30.22	10	5
25	12.2	230	1	2	4	15	25.37282	20.5	17	0	0	0	0	1	1	822	1	10	59	30.22	10	5
23	7.99	48	1	2	4	39	26.04762	21	23	0	0	0	0	1	1	1393	1	10	59	30.22	10	5
23	7.99	48	1	2	4	39	26.04762	21	23	0	0	0	0	1	1	1393	1	10	59	30.22	10	5
23	7.99	48	1	2	4	39	26.04762	21	23	0	0	0	0	1	1	1393	1	10	59	30.22	10	5
23	7.99	48	1	2	4	39	26.04762	21	23	0	0	0	0	1	1	1393	1	10	59	30.22	10	5
27	9.86	48	1	2	4	31	26.04762	21	23	0	0	0	0	1	1	1393	1	10	59	30.22	10	5
27	9.86	48	1	2	4	31	26.04762	21	23	0	0	0	0	1	1	1393	1	10	59	30.22	10	5
27	9.86	48	1	2	4	31	26.04762	21	23	0	0	0	0	1	1	1393	1	10	59	30.22	10	5
27	9.86	48	1	2	4	31	26.04762	21	23	0	0	0	0	1	1	1393	1	10	59	30.22	10	5
19	9.9	48	1	2	4	48	26.04762	21	23	0	0	0	0	1	1	1393	1	10	59	30.22	10	5
19	9.9	48	1	2	4	48	26.04762	21	23	0	0	0	0	1	1	1393	1	10	59	30.22	10	5
19	9.9	48	1	2	4	48	26.04762	21	23	0	0	0	0	1	1	1393	1	10	59	30.22	10	5
19	9.9	48	1	2	4	48	26.04762	21	23	0	0	0	0	1	1	1393	1	10	59	30.22	10	5
24	9.45	161	1	2	4	45	22.79048	18	22	0	0	0	0	1	1	746	0	10	59	30.22	10	5
24	11.6	230	1	2	4	12	25.37282	20.5	17	0	0	0	0	1	1	822	1	10	59	30.22	10	5
24	9.84	48	1	2	4	14	26.04762	21	23	0	0	0	0	1	1	1393	1	10	59	30.22	10	5
24	9.84	48	1	2	4	14	26.04762	21	23	0	0	0	0	1	1	1393	1	10	59	30.22	10	5
24	9.84	48	1	2	4	14	26.04762	21	23	0	0	0	0	1	1	1393	1	10	59	30.22	10	5
24	9.84	48	1	2	4	14	26.04762	21	23	0	0	0	0	1	1	1393	1	10	59	30.22	10	5
19	11.7	170	1	2	4	47	19.73016	20	17	0	0	0	0	1	1	1287	1	10	59	30.22	10	5
19	11.7	170	1	2	4	47	19.73016	20	17	0	0	0	0	1	1	1287	1	10	59	30.22	10	5
19	11.7	170	1	2	4	47	19.73016	20	17	0	0	0	0	1	1	1287	1	10	59	30.22	10	5
23	10.93	48	1	2	4	59	26.04762	21	23	0	0	0	0	1	1	1393	1	10	59	30.22	10	5

Figure 7. Cleaned and Processed Dataset

C: Project Code link: https://github.com/ml-project-lga-commute-analysis