

**Oh, the Places You'll Go! The Effect of Transportation Accessibility on Housing Prices in
New York City**

Anna Berman, Lidia Azucena Morales Vasquez, Viggy Kumaresan, Sicong Zhao, Yifei Wang

Group F: Least Squares

Abstract

A variety of factors go into a property's value, such as the size of the unit, amenities, and safety. One such factor is transportation accessibility, a factor that becomes increasingly important in dense urban areas where personal vehicles are less accessible. However, it is difficult to quantify how influential different transportation methods are in determining the sales price of a housing unit. The following analysis explores the effects of accessibility to transportation on housing prices in New York City using subway, bike share, taxi and Uber data through the use of various machine learning techniques. Our findings suggest that a random forest approach produces the highest accuracy for property price prediction compared to alternative modeling techniques. Additionally, our model identifies features that are highly influential in predicting price, such as property square footage and age of the apartment, as well as property proximity to subway and bike share stations.

Introduction

For the first time in human history, more humans live in cities than in rural areas (Streikas, 2005). This urbanization movement has had dramatic effects on city life, including the rise of ride-sharing use through companies like Lyft and Uber and a growing increase in housing demand. This increase in housing demand has sparked continued housing construction in major cities, but it has also been correlated with an increase in housing prices. The dynamic urban real estate market offers opportunities for buyers and for sellers, but determining the value of a housing unit can be challenging. For sellers, optimizing the sales price on their property ensures that they are able to maximize their Return on Investment (ROI) from their housing unit. For buyers, determining the true value of a housing unit is useful for comparing different properties, as well as for price negotiation later in the process.

House is usually treated as a heterogeneous goods, defined by numerous utility bearing features (Fan, G.-Z., 2006). All these features go into a housing unit's sale value, such as size of the unit, amenities, proximity to landmarks, safety, ambience of neighborhood, etc. One factor that we will focus on in our analysis is the effect of transportation accessibility on housing prices. We hypothesize that transportation availability is an important factor when evaluating a housing unit, especially in dense urban cities like New York City. However, it is difficult to quantify how influential different transportation methods are in determining the sales price of a housing unit.

In our analysis, we plan to explore the effects of accessibility to transportation on housing prices in New York City. Specifically, we hope to integrate data around housing proximity to the Metropolitan Transit Authority (MTA) and bike share stops as well as Taxi Cab and Uber data in order to predict NYC housing prices and ultimately better understand the impact of each transportation channel on price. The results of this analysis have implications for the real estate industry in terms of more accurately pricing housing units in an extremely competitive market.

Additionally, this analysis may help inform policymakers on where to invest in specific transportation initiatives to potentially drive real estate growth. We also hope to learn more about urban mobility overall through analyzing the transportation data, and this information can be useful for future urban planning and smart city initiatives.

Background

In regards to housing price prediction, previous researchers have developed machine learning models using various neighborhood features, such as crime rate, local business presence, and access to highways (Jingyi Mu et al., 2014). In recent years, as more urban data has become available, researchers have begun to improve these models by integrating novel data sources, such as local census data and education profiles (Gao, 2019). This direction of research shows that as we examine more attributes that

affect a housing price, then we can truly start to develop models that mimic the logic of the underlying market.

The relation between transit stations and neighboring housing values is also not a novel research subject. Transportation accessibility has been shown to increase housing prices, due to the decrease in commuting costs for both tenants as well as landlords. On the other hand, there are also negative effects of transportation such as noise, traffic, pollution, crime, etc. that can have potentially negative effects on house value (Kilpatrick et Al., 2007).

Nowadays, there are more options for mobility in cities apart from subway transportation, services such as Uber, Lyft, Citi Bike are changing the way people transport daily. Because New Yorkers can now carpool into Manhattan for just a few dollars, many of residents are rethinking their priorities as homebuyers (Small, 2018). Buyers who do not live close to a subway station, have found alternatives in these services. This industry disruption does not mean that traditional public transportation is being replaced, but instead that a wider, holistic transportation system may begin to affect real estate property value.

Take bikeshares for example. Once introduced into a metropolitan area, individuals' transportation options and preferences are subject to change. Among Capital Bikeshare riders, 70% choose bikeshare to get to their destination as the quickest and easiest way. Moreover, bike sharing systems offer benefits such as decreased carbon dioxide emissions, traffic congestion, accident rates and construction costs (Sobolevsky et Al, 2018). Research has shown that the effects of bikeshare introduction do have an effect on property values. A study published in the journal Transport Policy found an average 2.7% increase property value in central Montreal after the city launched its bike sharing system (El-Geneidy, 2015).

Recently, researchers have turned to machine learning techniques to analyze urban geospatial data and specifically to examine the effects of transportation in urban areas. One area of research has been focused on using deep learning techniques to model traffic congestion through an Internet of Things based analysis (Ma et al, 2015). Most notably, researchers and engineers at Uber and Lyft use advanced machine learning methods to optimize pickups and dropoffs and provide reliable wait times to users (Turakhia, 2017). As this data collection continues to grow, the potential for future applications of these techniques will increase in magnitude and in accuracy, potentially leading to innovative solutions for urban areas. This movement has been referred to as the 'Smart City' movement (ICMA, 2016). Public officials have begun to realize the potential for technology and data to improve the lives of their city's residents, which has motivated the creation of open data portals in various cities in the United States, enabling technologists and data scientists to derive new insights and recommendations.

Data

In order to assess the impact of transportation accessibility on housing prices in New York City, New York, we chose to utilize several sources of data. Specifically, we chose to integrate property sale information with subway, taxi, Uber, and bike share data into our analysis. A description of the datasets used in our analysis is shown in Figure 1.

Figure 1. Primary Data Sources. A description of the five data sources that were integrated into our analysis.

Data Source	Description	Time Frame	Source
NYC Property Sales	The NYC Property Sales Data includes a record of every building or building unit sold in NYC in one year - approximately 85,000 records. This dataset contains the location, address, type, sale price, and sale date of building units sold. Housing unit sale prices ranged from tens of thousands up to \$2,210,000,000 with a heavy right skew.	September 2016 - September 2017	(New York City Department of Finance & Kaggle, 2017)
Uber TLC Foil	The Uber TLC Foil Dataset contains 4.5 million Uber pickups in New York City from April to September 2014, and 14.3 million more Uber pickups from January to June 2015. The dataset includes pickup location and date, and time.	April - September 2014 & January-June 2015	(NYC Taxi and Limousine Commission & FiveThirtyEight, 2015)
Taxi Duration	The 2016 Taxi Duration dataset includes over 2 million records of Taxi rides taken in NYC in 2016. The dataset includes pickup and dropoff date, time, and location as well as trip duration. This data is similar to the Uber data, since they are both collected and curated by the NYC Taxi and Limousine Commission (TLC).	January-December 2015	(NYC Taxi and Limousine Commission & Kaggle, 2017)
NYC Subway	This data file provides a variety of information on NYC subway station entrances and exits which includes but is not limited to: Division, Line, Station Name, Longitude and Latitude coordinates of entrances/exits.	April 2019	(NYC Transit Subway Entrance And Exit Data & Data.gov)
Citi Bike System	The Citi Bike Dataset includes data around bike share station locations in NYC, as well as pickup and dropoff time and locations. In 2016, more than 20 million NYC rides were	January-December 2016	(Citi Bike, 2016)

	collected and recorded for analysis.	
--	--------------------------------------	--

Taking a look at housing prices in NYC, we can see that the highest property prices seem to cluster around lower Manhattan and West Brooklyn with property sales decreasing with distance from this hub. In Figure 2 and 3 we can see subway and bike stations also cluster around this area. Specifically in terms of bike stations, we see that Citi Bike System stations are only located in lower Manhattan (Figure 4). This may be a potential challenge in interpreting our analysis as we do not want distance to the nearest bike station to become a proxy for being in this high priced area.

Figure 2. NYC property prices, bike and subway stations. Prices as well as subway and bike accessibility concentrate in similar areas.

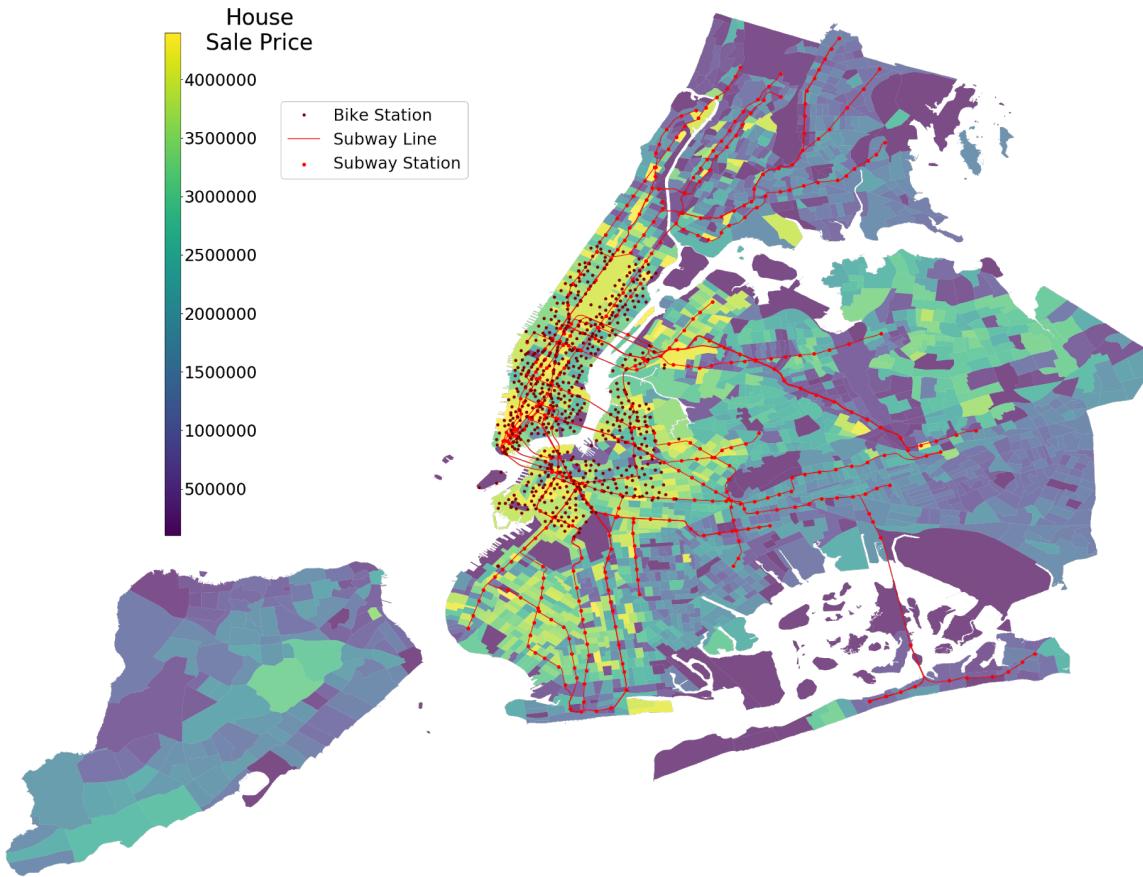


Figure 3. Brooklyn property prices, bike and subway stations. In southeast Brooklyn specifically, we see lower metro accessibility and lower house prices than in other parts of Brooklyn

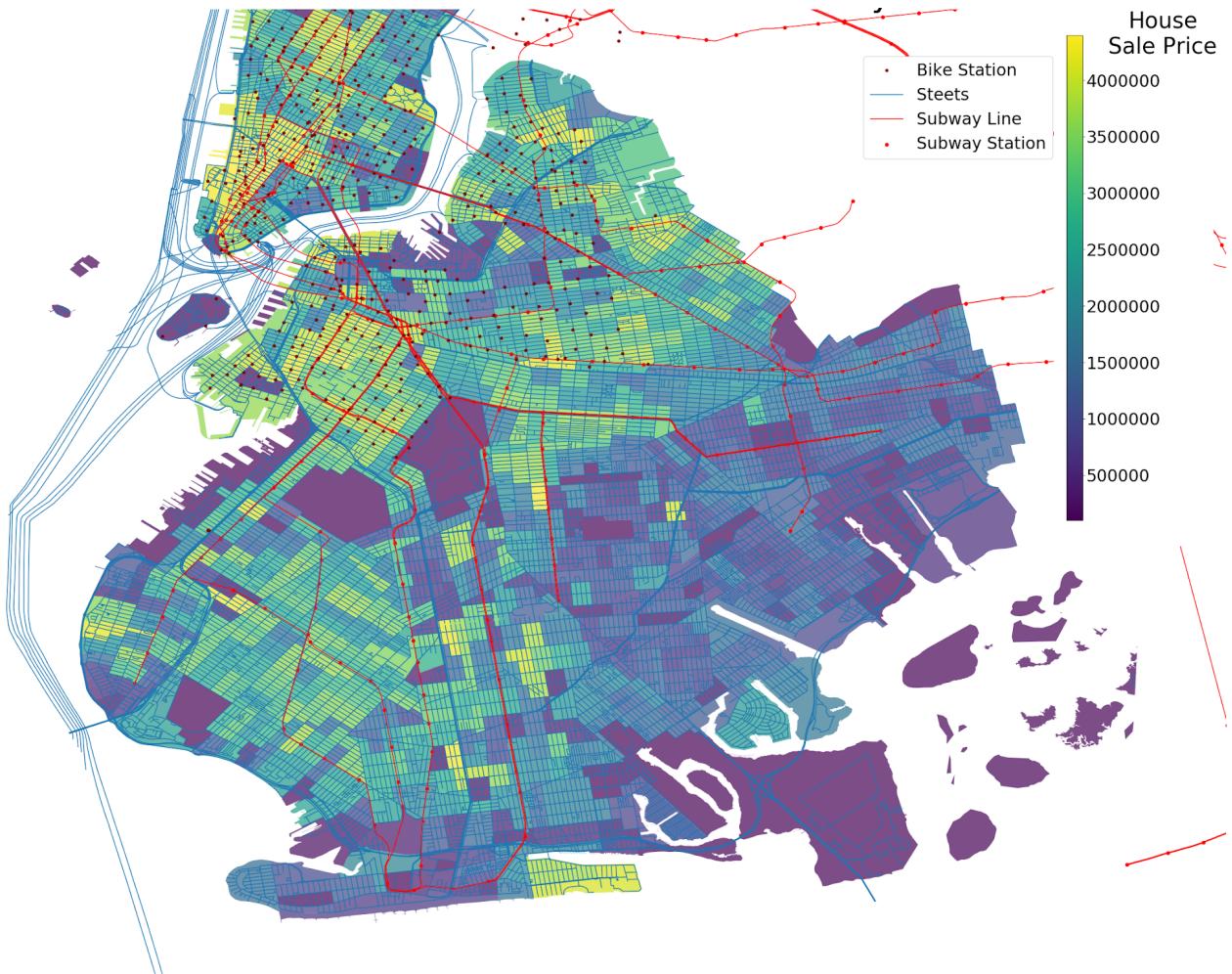
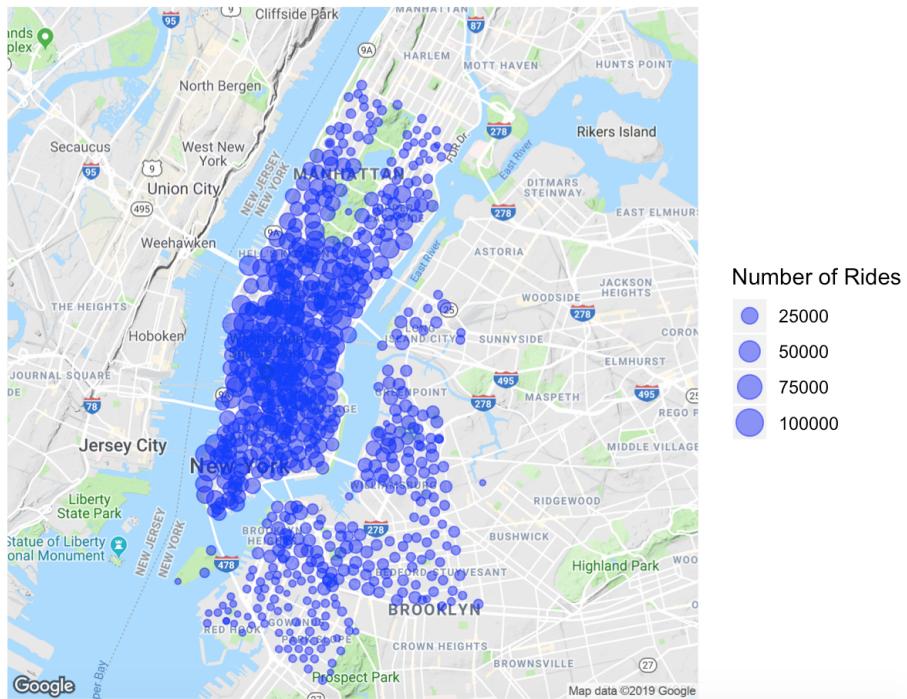


Figure 4. Citi Bike System usage in 2016. The size of the blue dots represents the number of rides originating from each station. Citi Bikes are only available in lower Manhattan and West Brooklyn. The majority of rides take place in lower Manhattan.



When looking at the Uber and Taxi data, we can see that Manhattan has the highest Uber and Taxi ride volume compared to any other borough (Figure 5 and 6). However, the difference between Manhattan and other boroughs is most pronounced in our Taxi data (Figure 6). Similar to our bike system data, we must be careful of interpreting our Taxi ride data when creating our models as we do not want Taxi rides to become a proxy for a property being located in Manhattan.

Figure 5. Uber pick up and drop off counts by borough. Manhattan has the highest number of Uber pick ups and drop offs followed by Brooklyn and Queens respectively.

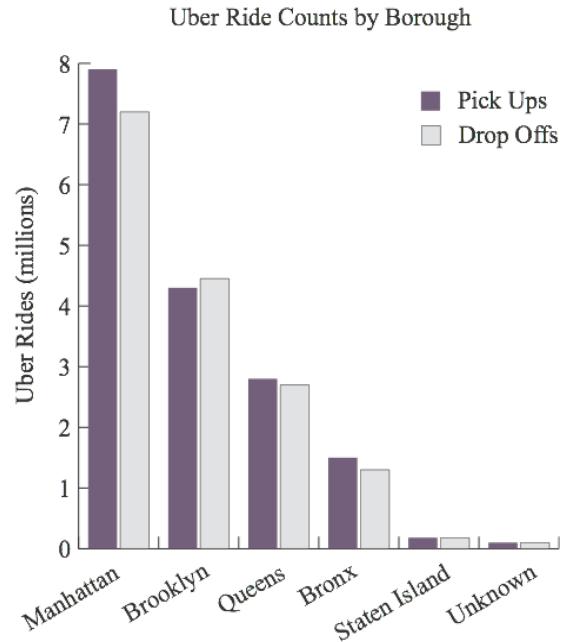
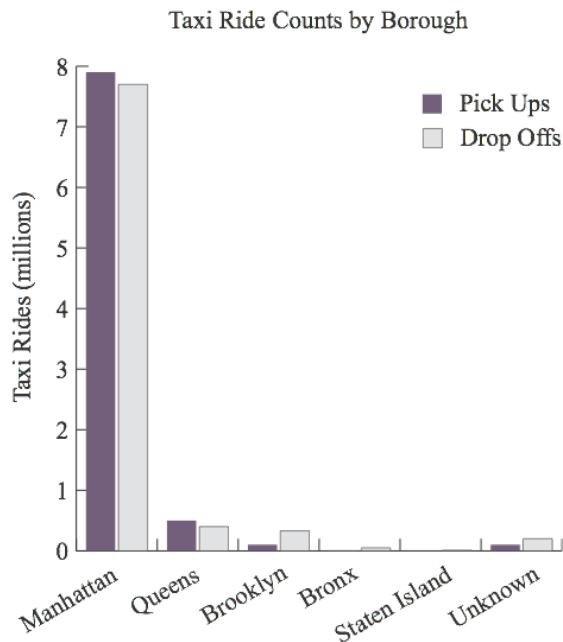
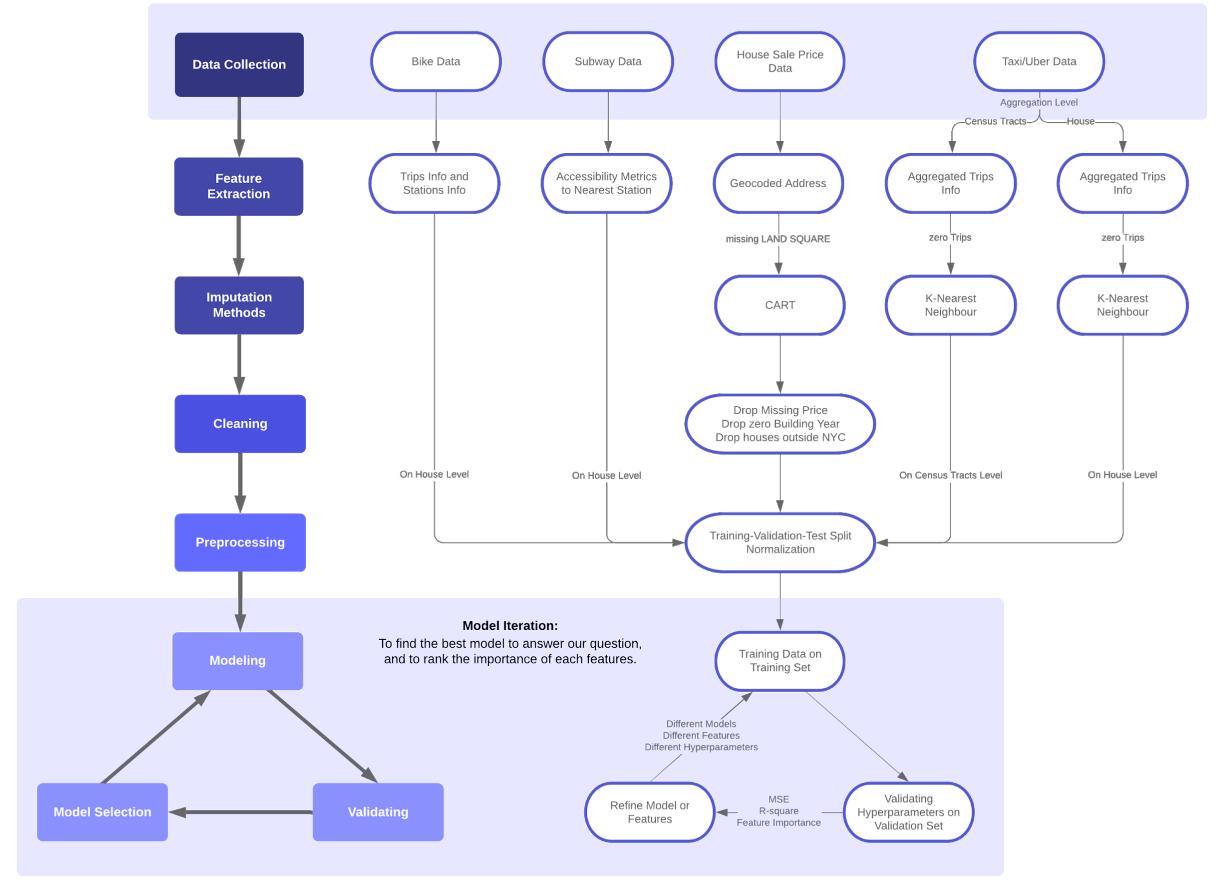


Figure 6. Taxi pick up and drop off counts by borough. Manhattan by far has the highest number of trips compared to any other borough.



Methods

Figure 7. Flowchart of analysis. Each box represents a major processing step in our analysis.



Data Preprocessing

NYC Property Sales

Integral to our research question was how to connect transportation information to each property in our NYC Property Sales dataset. The first step in creating this connection was to geolocate each property with latitude and longitude coordinates. This geolocation was implemented using the addresses of each property and the Google Maps API as well as manually entering 0.01% of all rows. Through this geolocation, we were able to identify several outliers in terms of location that appears outside of the five NYC boroughs. Due to the urban nature of our research question, we chose to not include these records in our analysis.

The NYC Property Sales dataset included several missing data around property square footage. To account for this missing data, we applied multiple imputation using CART and the mice package to impute square footage in approximately 31% of our data.

Beyond missing data, this dataset also included several nonsensical values, close to \$0, for property sale price. These sales ultimately involved deed transfers between parties such as parents transferring property ownership to their children. Again due to the nature of our research question, we chose to drop these records from our dataset and to focus only on traditional property sales. Dropping these observations decreased our data from around 84,000 observations to 56,000. We considered that imputation was not a good option in this case, since the dataset did not have enough information about the characteristics of the house and imputation had potential to skew our results.

Uber TLC Foil

For our Uber data, we had the latitude and longitude of the trip pickups, so we decided to map these latitudes and longitudes onto publicly available NYC census tract data. Since this data also contained missing values, we imputed using a nearest-neighbors approach: we mapped the data onto a 500mx500m grid, searching the nearby 8 grid blocks for values. If none of these blocks had values, then we expanded the grid radius and searched again until we had a value that we could use for a missing block. Once this was done for all blocks, we calculated a feature for the number of pickups in each block. We mapped the housing dataset onto the same census tracts, and then merged the files to integrate the Uber features.

Taxi Duration

For the taxi data, we had the latitude and longitude of both pickups and dropoffs, as well as additional trip information like trip duration, distance, and number of passengers. Using this data, we set a radius of 500m and filtered trips whose pickups and dropoffs are within that radius for each house in our dataset. Then, we were able to create features around pickup/dropoff count, average trip duration, average passenger count, and average trip distance for each house radius. We also created ‘speed’ features, by dividing trip distance by distance time, which serves as a proxy of traffic accessibility nearby. We also had to deal with missing data for these features, but due to the size of the dataset, we decided to still impute using a nearest-neighbors approach but just use the data from the 5 nearest house with valid values.

NYC Subway

For each property in the NYC Property Sales dataset we included a feature with the minimum Manhattan distance to a MTA subway station. Additionally, we also included included metrics for the number of subway stations walkable within both five and ten minutes. For a comprehensive list of features, see Appendix A.

Citi Bike System

For each property in the NYC Property Sales dataset we included a feature with the minimum Manhattan distance to a Citi bike station. Additionally, we also included aggregate features about the

nearest bike station such as the average time bikes are taken and returned to the specific station as well as demographic information about the riders who use the station. For a comprehensive list of features, see Appendix A.

Data partition

In order to both create as well as evaluate the performance of our model, we separated our original dataset into a training, a validation, and a test set. Out of the 56,000 properties, we split the dataset into 70% training, 20% validation, and 10% testing. We chose to allocate the majority of the data in the training set because each of our approaches required a significant amount of training data to result in accurate validation and test performance.

Metrics

During our model selection process, we attempted several modeling approaches. These initial models included linear regression, KNN, classification and regression tree (CART), random forest, and neural networks. We chose these models for their the ability to predict a continuous outcome.

For each modeling approach, we used Mean-Squared Error (MSE) as our performance metric, as this is a commonly used evaluation criterion for regression problems. Mean-squared error is calculated by squaring the distance between the predicted value and the true value for all data points, and then summing these distance values and dividing the sum by the number of observations:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

We also evaluated our performance on R-squared or the coefficient of determination. R-squared is a metric that measures the proportion of the variance in the dependent variable that is predictable from the independent variable.

Results

As discussed in the previous section, we chose to separate our original dataset into a single training, validation, and test set. The training set contained 70% of the properties, the validation set contained 20%, and the test set contained 10%.

To compare results across our five modeling approaches, we used Mean-Squared Error (MSE) and the coefficient of determination (R-squared). For each modeling approach, we also compared our results to an identical model without the addition of transportation accessibility (called the Baseline models). In this way, we are able to quantify the value of transportation features in prediction for each approach. The performances of each of our models are shown in Figure 8.

Figure 8. Performance comparison across model approaches. Comparing across our five model approaches, the Random Forest approach produced the best predictions on our test set.

Method	Means Squared Error (MSE)		Coefficient of Determination (R-Squared)	
	Property Features Only	Property Features + Transportation Features	Property Features Only	Property Features + Transportation Features
Linear Regression	3.85×10^{12}	4.37×10^{12}	-4.91	-5.72
K Nearest ¹ Neighbors (KNN)	5.27×10^{11}	4.00×10^{11}	0.189	0.386
Classification And Regression Tree (CART) ²	3.58×10^{11}	3.55×10^{11}	0.395	0.401
Neural Network ³	3.28×10^{11}	3.25×10^{11}	0.446	0.465
Random Forest	2.58×10^{11}	2.27×10^{11}	0.56	0.61

¹ KNN: K = 12 for model with only property features, K = 10 for model with property and transportation features

² CART: Max Depth = 10

³ Neural Network : 4 layers w/ 100 neurons, 1 output layer, 1 dropout layer, 100 epochs

Based on the results from our model comparisons, a Random Forest approach produces the most accurate predictions (Note: our linear regression models produced negative R-Squared values, indicating that the data does not follow a linear pattern). Random Forest models are especially useful for interpretation because they produce ranked feature importance scores to indicate which variables were most useful in property price predictions. Figure 9 and 10 outline the most predictive features from our property features model as well as our property and transportation features model.

Figure 9. Feature Importance for Property Features Only Model. Using only property features from the NYC Property Sales data, square footage and year built are the most predictive features in our baseline model.

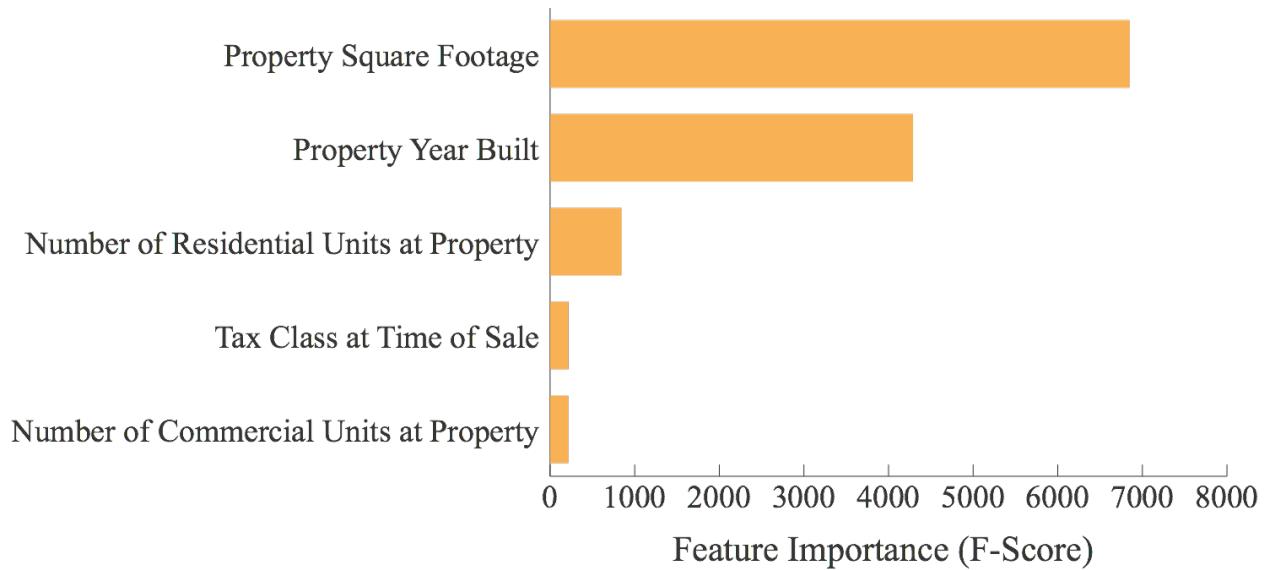
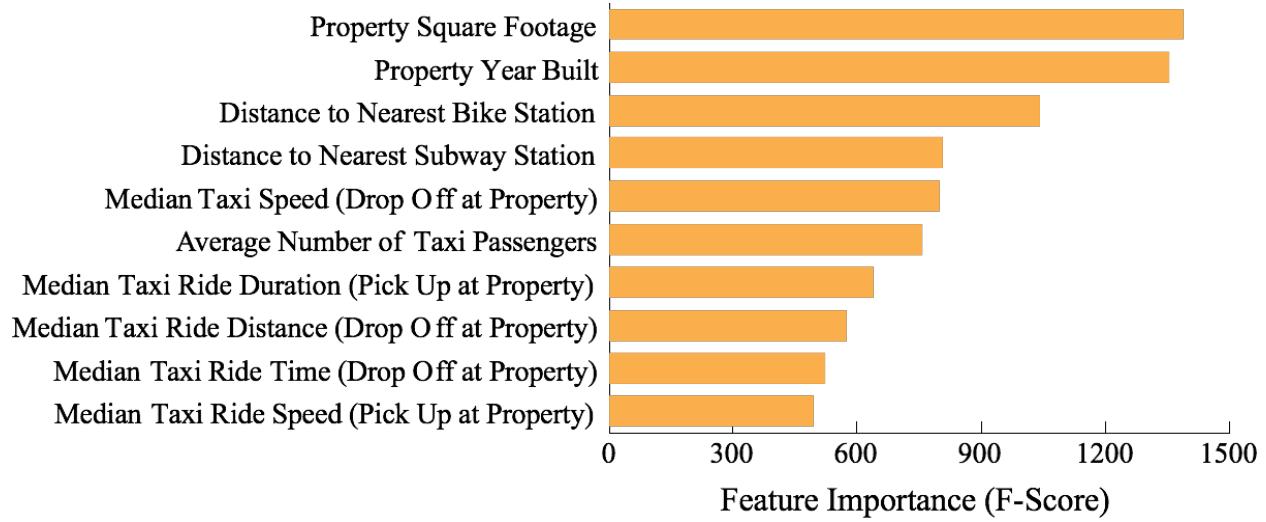


Figure 10. Feature Importance for Final (Property Features + Transportation Features) Model. By including transportation data from subways, bike shares, taxis and Uber, we still see that square footage and year built are the most predictive features in our baseline model. However, we also see that the nearest distance to a bike and subway stations are also important predictors.



Using this model, we see several instances of accurate predictions, poor predictions, and predictions that fall somewhere in between (Figure 11, 12, and 13).

Figure 11. Examples of accurate predictions. Our random forest model produced very accurate predictions for many properties such as these two in Brooklyn.

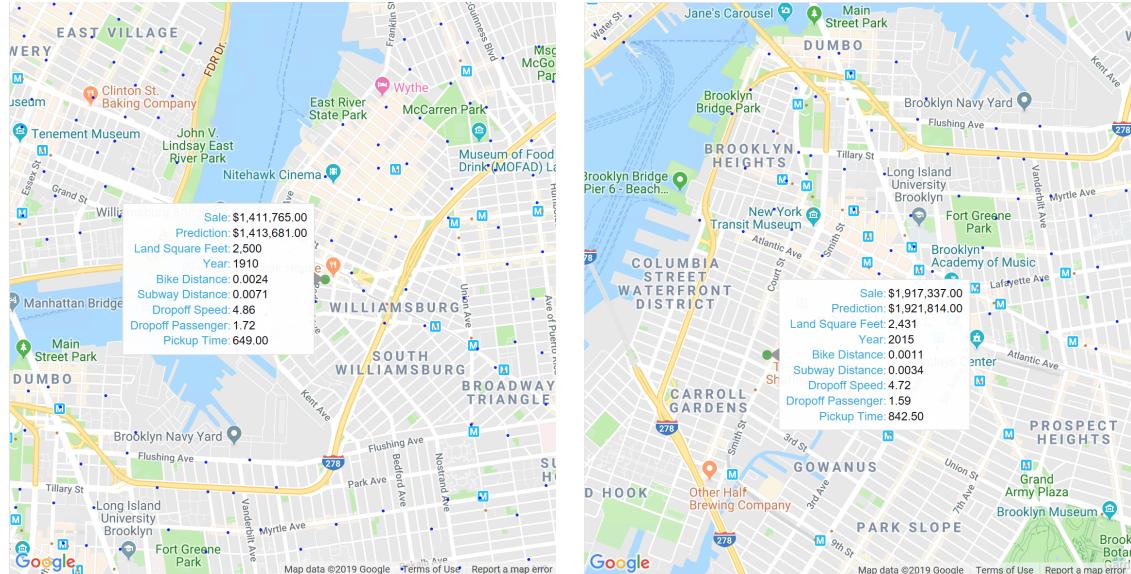


Figure 12. Examples of not accurate predictions. Our random forest model produced less accurate predictions for many properties such as these two in Lower Manhattan and Brooklyn.

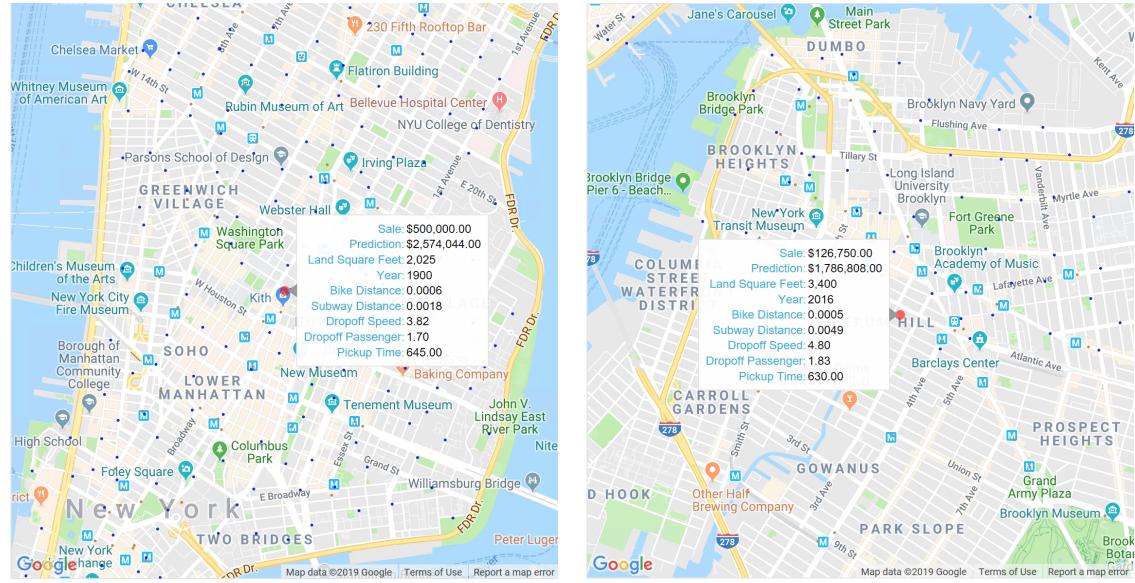
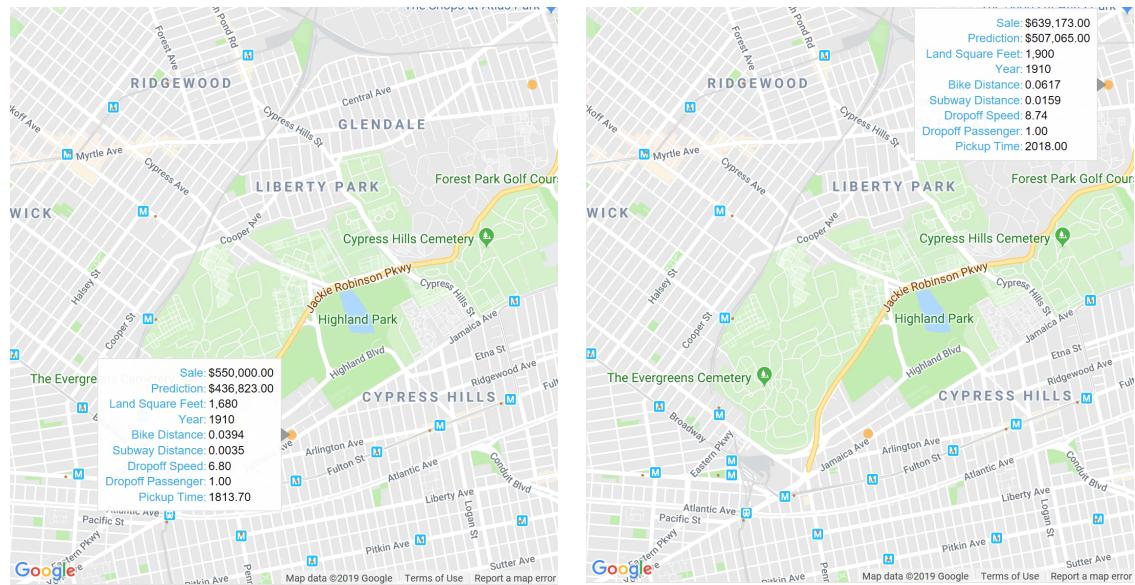


Figure 13. Examples of mediocre accurate predictions. Our random forest model produced relatively accurate predictions for many properties such as these two near Highland Park.



The housing sale price in NYC Property Sales dataset have a long tail on both sides; some houses are extremely expensive while others are extremely cheap. There are potentially some other hidden features contributing to the price but fail to be included in our model, such as luxury/poor amenities, great/poor directions or views, etc. Since the transportation features are determined mainly by house

location but not house features itself, our model tend to averaging the house price in certain areas, and thus fail to predict the outliers. That would result in overestimation on low price (compared to the area) house and underestimation on high (compared to the area) price house.

Conclusions

In studying the effect of transportation accessibility on housing price in New York City, we integrated data from several different sources and created features to predict housing price in a regression model. Our work in this paper highlights the use of these transportation features in housing price prediction; compared to a baseline model with only housing attributes (ex. Square footage, year the house was built, etc), our models with transportation features were able to achieve higher performance, as measured by MSE and R-squared values.

From the feature importance that we calculated, we can see that the most important factors in our model were still the square footage of the house and the year the house was built. But in addition to these two factors, we also see information added by bike, subway, and taxi/Uber data, showing that these factors are influential in determining the value of a property. However, we cannot determine a causal relationship, because often transportation infrastructure is set up in accordance with housing population. For example, ‘Distance to Nearest Bike Station’ was the third most important feature in our model, but we know that all of the bike stations are located in Manhattan and Brooklyn, so this feature might just be indicating that properties in this area have a higher sales price (which we already know as a broad fact).

We hope that our machine learning approach to predicting sales price has two main takeaways: first, that transportation is a key factor in determining the sales price of a house in New York City, and we can quantitatively calculate the effect of them on housing price. Second, that publicly available transportation/geographic data can be used in a useful manner to make inferences about social behavior. With these takeaways, we believe that future work can build off of this approach and create new novel methodologies, leading to even more accurate predictions. Ultimately, realtors, investors, sellers, and buyers can all use these models to help with their own real estate questions.

Roles

In general, all team members attended weekly meetings to discuss progress and future steps and helped wrote the final paper. In terms of specific roles:

- Anna was responsible for cleaning, aggregating, and integrating the Citi Bike dataset and for running the CART analysis.
- Yifei was responsible for cleaning, aggregating, and integrating the Taxi dataset and for creating the visualizations.
- Azucena was responsible for cleaning, aggregating, and integrating the MTA dataset and for running the random forest analysis.
- Viggy was responsible for cleaning, aggregating, and integrating the Uber dataset and for running the neural network analysis.
- Sicong was responsible for cleaning, aggregating, and integrating the Uber dataset, for running the linear and KNN analysis, and for creating the video animations.

References

- Citi Bike (2016). Systems Data. <https://www.citibikenyc.com/system-data>
- El-Geneidy, A., van Lierop, D., & Wasfi, R. (2016). Do people value bicycle sharing? A multilevel longitudinal analysis capturing the impact of bicycle sharing on residential sales in Montreal, Canada. *Transport policy*, 51, 174-181.
- Fan, G.-Z., Ong, S. E., & Koh, H. C. (2006). Determinants of House Price: A Decision Tree Approach. *Urban Studies*, 43(12), 2301–2315. <https://doi.org/10.1080/00420980600990928>
- Gao, G., Bao, Z., Cao, J., Qin, A. K., Sellis, T., & Wu, Z. (2019). Location-Centered House Price Prediction: A Multi-Task Learning Approach. *arXiv preprint arXiv:1901.01774*.
- ICMA (2016). The Smart City Movement. <https://icma.org/articles/smart-city-movement>
- Jingyi Mu, Fang Wu, and Aihua Zhang, “Housing Value Forecasting Based on Machine Learning Methods,” Abstract and Applied Analysis, vol. 2014, Article ID 648047, 7 pages, 2014. <https://doi.org/10.1155/2014/648047>.
- Kilpatrick, Throupe, Carruthers, & Krause (2007). The Impact of Transit Corridors on Residential Property Values.
- Ma, X., Yu, H., Wang, Y., & Wang, Y. (2015). Large-scale transportation network congestion evolution prediction using deep learning theory. *PloS one*, 10(3), e0119044.
- New York City Department of Finance & Kaggle (2017). NYC Property Sales. <https://www.kaggle.com/new-york-city/nyc-property-sales>
- NYC Taxi and Limousine Commission & FiveThirtyEight, (2015). Uber TLC Foil Response. <https://github.com/fivethirtyeight/uber-tlc-foil-response>
- NYC Taxi and Limousine Commission & Kaggle (2017). New York City Taxi Trip Duration. <https://www.kaggle.com/c/nyc-taxi-trip-duration/data>
- Data.gov (2019). NYC Transit Subway Entrance And Exit Data. https://catalog.data.gov/dataset/nyc-transit-subway-entrance-and-exit-data?fbclid=IwAR0ucXcUUfZF2YkJN8QJX06OMmeWaOWUm_vYJmt1TXCwLCUrH6KCD_aRDJU
- Small, E. (2014). Sayonara, subway: How ridesharing apps are changing the real estate calculus for brokers and developers. <https://therealdeal.com/2018/03/20/sayonara-subway-how-ridesharing-apps-are-changing-the-real-estate-calculus-for-brokers-and-developers/>
- Sobolevsky, S., Levitskaya, E., Chan, H., Postle, M., Kontokosta, C. (2018) Impact Of Bike Sharing In New York City

Strekas, T., 2005. New York as a Model for the Study of Urbanization. CUNY Institute to Nurture New York's Nature. <https://qcpages.qc.cuny.edu/nyn/model.html>

Turakhia, C., (2017). Engineering More Reliable Transportation with Machine Learning and AI at Uber. *Uber Engineering*. <https://eng.uber.com/machine-learning/>

Appendix A. Complete features involved in our analysis.

Feature	Description	Source
Latitude	Latitude of property	GoogleMaps API
Longitude	Longitude of property	GoogleMaps API
Borough	A digit code for the borough the property is located in; in order these are Manhattan (1), Bronx (2), Brooklyn (3), Queens (4), and Staten Island (5).	NYC Property Sales
Neighborhood	Neighborhood property is located in	NYC Property Sales
Building Class Category	Building classification at time of data collection as defined by the City of New York	NYC Property Sales
Residential Units	Number of residential units in property	NYC Property Sales
Commercial Units	Number of commercial units in property	NYC Property Sales
Land Square Feet	Number of Square Feet in property	NYC Property Sales
Year Built	Year property was built	NYC Property Sales
Tax Class at Time of Sale	Tax class of property at time of sale	NYC Property Sales
Building Class Category at Time of Sale	Building classification at time property sale as defined by the City of New York	NYC Property Sales
Sale Price	Price property was sold	NYC Property Sales
SubwayDist	Distance to the nearest MTA station	NYC Subway
numSub5min	Number of MTA stations within 5 minute walking radius	NYC Subway
numSub10min	Number of MTA stations within 10 minute walking radius	NYC Subway
bike_minDist	Distance to the nearest Citi bike share station	Citi Bike System
bike_nRides_start	Number of rides started from the nearest bike station in 2016	Citi Bike System

bike_durationMed_star	Median duration of rides started from the nearest bike station in 2016	Citi Bike System
bike_startTimeAvg	Mean start time of rides started from the nearest bike station in 2016	Citi Bike System
bike_startTimeSd	Standard Deviation start time of rides started from the nearest bike station in 2016	Citi Bike System
bike_endLatAvg	Mean latitude of bike drop off for rides started from the nearest bike station in 2016	Citi Bike System
bike_endLatSd	Standard Deviation latitude of bike drop off for rides started from the nearest bike station in 2016	Citi Bike System
bike_endLongAvg	Mean longitude of bike drop off for rides started from the nearest bike station in 2016	Citi Bike System
bike_endLongSd	Standard Deviation longitude of bike drop off for rides started from the nearest bike station in 2016	Citi Bike System
bike_females_start	Percentage of rides taken by females for rides started from the nearest bike station in 2016	Citi Bike System
bike_ageMed_start	Median age of riders or rides started from the nearest bike station in 2016	Citi Bike System
bike_nRides_end	Number of rides ended from the nearest bike station in 2016	Citi Bike System
bike_durationMed_end	Median duration of rides ended from the nearest bike station in 2016	Citi Bike System
bike_endTimeAvg	Mean start time of rides ended from the nearest bike station in 2016	Citi Bike System
bike_endTimeSd	Standard Deviation start time of rides ended from the nearest bike station in 2016	Citi Bike System
bike_startLatAvg	Mean latitude of bike drop off for rides ended from the nearest bike station in 2016	Citi Bike System
bike_startLatSd	Standard Deviation latitude of bike drop off for rides ended from the nearest bike station in 2016	Citi Bike System

bike_startLongAvg	Mean longitude of bike drop off for rides ended from the nearest bike station in 2016	Citi Bike System
bike_startLongSd	Standard Deviation longitude of bike drop off for rides ended from the nearest bike station in 2016	Citi Bike System
bike_females_end	Percentage of rides taken by females for rides ended from the nearest bike station in 2016	Citi Bike System
bike_ageMed_end	Median age of riders or rides ended from the nearest bike station in 2016	Citi Bike System
pickup_count	Number Uber rides picked up in property radius (500m)	Uber TLC Foil
pickup_trip_med_dist	Median distance traveled in Uber rides picked up in property radius (500m)	Uber TLC Foil
pickup_trip_med_time	Median time traveled in Uber rides picked up in property radius (500m)	Uber TLC Foil
pickup_trip_med_speed	Median speed traveled in Uber rides picked up in property radius (500m)	Uber TLC Foil
pickup_trip_avg_passenger	Median time traveled in Uber rides picked up in property radius (500m)	Uber TLC Foil
dropoff_count	Number Uber rides dropped off in property radius (500m)	Uber TLC Foil
dropoff_trip_med_dist	Median distance traveled in Uber rides dropped off in property radius (500m)	Uber TLC Foil
dropoff_trip_med_time	Median time traveled in Uber rides dropped off in property radius (500m)	Uber TLC Foil
dropoff_trip_med_speed	Median speed traveled in Uber rides dropped off in property radius (500m)	Uber TLC Foil
dropoff_trip_avg_passenger	Median time traveled in Uber rides dropped off in property radius (500m)	Uber TLC Foil
taxi_pu_time_sec_median	Median ride time for taxis dropped off in a specific tract (500m * 500m, roughly)	Taxi Duration

taxi_pu_v_median	Median velocity of taxi trips started in a specific tract (500m * 500m, roughly)	Taxi Duration
taxi_pu_distance_sum	Total ride distance for taxis picked up in a specific tract (500m * 500m, roughly)	Taxi Duration
taxi_pu_passenger_sum	Total passengers for taxis picked up in a specific tract (500m * 500m, roughly)	Taxi Duration
taxi_pu_passenger_mean	Median number of passengers for taxis picked up in a specific tract (500m * 500m, roughly)	Taxi Duration
taxi_pu_distance_median	Median ride distance for taxis picked up in a specific tract (500m * 500m, roughly)	Taxi Duration
taxi_pu_time_sec_sum	Total time ridden for taxis picked up in a specific tract (500m * 500m, roughly)	Taxi Duration
uber_trips	Number of uber trips started in a specific tract (500m * 500m, roughly)	Taxi Duration
taxi_do_v_median	Median velocity of taxi trips with their destinations in a specific tract (500m * 500m, roughly)	Taxi Duration
taxi_do_time_sec_median	Median ride time for taxis dropped off in a specific tract (500m * 500m, roughly)	Taxi Duration
taxi_do_distance_median	Median ride distance for taxis picked in a specific tract (500m * 500m, roughly)	Taxi Duration
taxi_do_passenger_sum	Total passengers for taxis dropped off in a specific tract (500m * 500m, roughly)	Taxi Duration
taxi_do_passenger_mean	Median number of passengers for taxis dropped off in a specific tract (500m * 500m, roughly)	Taxi Duration
taxi_do_time_sec_sum	Total time ridden for taxis picked in a specific tract (500m * 500m, roughly)	Taxi Duration
taxi_do_distance_sum	Total ride distance for taxis dropped off in a specific tract (500m * 500m, roughly)	Taxi Duration