Understanding Tourist Mobility Patterns in New York City and Predicting Taxi Demand for Tourist Hotspots

Akash Yadav, Siqi Huang, Chenjie Su & Yue Jin

New York University - Center for Urban Science and Progress

Abstract

New York City is amongst the most popular tourist places in the world. Every year, on an average, about 59 million people visit the city. This massive influx creates a huge stress on the current capacities of the city, especially public transportation services. The aim of this study is to understand if major tourist mobility patterns can be extracted from data sources including public transportation and hospitality datasets and if future demand for current capacities be estimated. The paper also attempts to highlight the coming of new-age urban tourism around New York city being driven by the usage of applications like Airbnb and Uber, and also discusses its impact in creating a dynamic shift in the urban structure of a city. To predict the demand for public services, the paper extends its applications to be applied for the tourism industry based on pre-existing literature and models.

Keywords: NYC, Tourism, Hotspots, Yellow Taxis, Demand Prediction

Introduction

New York City received a ninth consecutive annual record of approxiamately 65.2 million tourists in 2018, counting not just overnight visitors but anyone visiting for the day from over 50 miles away, including commuters[1]. Overall the city welcomed 37.9 million visitors who stayed overnight in 2018, of which 13.6 million were international. The yellow taxicabs is kind of a signal of this city. Grabbing a taxi can be ideal when tired feet, heavy luggage or shopping bags weigh tourists down.

Literature Review and Related Work

Mobility patterns play a central role in determining urban dynamics and structure within a particular city. Mobility patterns reveal various interactions people have with different parts of the city, spread across both spatially and temporally[2]. Our understanding of mobility patterns as an important indicator for identifying tourist zones originate from the same literature. We decided to focus our study on identifying tourist hotspots and consecutively predicting demand for taxi services, by understanding how modern mobility patterns have been heavily reliant on “on-demand” taxi services and “for-hire-vehicles”. The new age “on-demand” taxi services have not only been the most convenient and sought after modes of transportation for both international and national travellers but also has created a shift in ways people are accessing the tourism and hospitality industries[3].

Several analyses dealing with revealing hotspots have made use of clustering algorithms to group zones of similar characteristics together[4]. CrimeStat and London Accident Hotspot Analysis are some examples that have been successful at using K-Means clustering algorithm to determine areas of high criminal activity and accidents respectively[4]. There have been similar analyses that have been extended to identify hotspots for efficiently predicting taxi demand[5].

Taxi demand itself has been possible by making use of GPS trackers data that identify demand in each origin and destination point historically. There exists a vast amount of literature that make use of time-series, network based and various unsupervised learning algorithms in order to correctly predict demand for taxis within a city[6,7]. There have been studies that suggest that taxi demand is a function of the characteristic properties of the zone for which the demand is being predicted[8]. Residential areas, commercial zones, tourist attractions etc. have functional properties that determine the demand for taxis within them and need to be accounted for. There are gaps in the literature that identify the best ways to extract zone level characteristics to predict taxi demand and therefore, for the purposes of our research, it was decided to first extract these zones and then look at the demand for taxis within these zones.

Data Collection and Preprocessing

The dataset used in this study includes Yellow Taxi Trip Records and For Hire Vehicle Records made available by the NYC Taxi and Limousine Commision of New York City. For the purposes of identifying holiday periods and to characterise high tourist activity within these periods, our research specifically deals with December data obtained from these datasets, for the year 2018. These datasets has been further filtered for the holiday period, defined as the 3 weeks from the 18th of December to the 31st of December and for time periods between 11 AM to 4 PM, for the purposes of identifying our hotspots. These criteria were decided upon after clearly distinguishing mobility patterns across December and other months for all weeks and time periods. All other time periods were understood to be limited to gather information about dominant tourist mobility patterns.

Apart from taxi data, we have also made use of listings, prices and reviews data from the AirBnB data portal in order to understand some of the defining characteristics of tourist behavior in New York City (Table 1). This data has been utilised to study where people choose to stay during the vacation period (defined previously) and the factors that go into modeling this choice. Since check-in data was not readily available from these data sources; reviews data filtered for + 1 week of the above defined holiday period was used to make this analysis happen. The underlying assumption behind such filtering is that customers leave reviews for their stays as soon as they vacate, or within a week of vacating the place.

For the purposes of predicting taxi demand, features including weather data extracted from World Weather Online was used. The next feature makes use of Holiday Data that lists whether or not the said date for which the demand is being predicted is a holiday or not. Most time series analysis make use of lag variables to be able to accurately predict demand for taxis on a per hour basis. Our research has also incorporated such lag features in order to make our final dataset to train the demand for taxis in each of our tourist zones (Table 2).

Exploratory Data Analysis

Based on the datasets we selected, we carried out some exploratory data analysis to better understand the structure of our data.

Figure 1 highlights the places that people are traveling from airports during the holiday period alone. It can be seen that most people travel to Manhattan and places in Brooklyn and

Queens that are close to Manhattan. The most frequented taxi zone right after being picked up from the airport is Times Square while the second most frequented zone is JFK Airport, suggesting people rushing from the domestic airports to take connecting flights. The zones of drop-offs from airport very closely resembles the Check-In Density Map for Airbnb properties in Figure 2, confirming that tourists travel and stay in places that are in Manhattan or close to Manhattan.

Higher demand for places within Brooklyn and Queens could be because of the high prices in Manhattan and therefore people on a budget usually make a decision to stay as close to Manhattan as possible, if not within Manhattan. Figure 3 reveals the average prices of listings (standardised) for Airbnb properties in one year. We also created a heatmap in Figure 4 showing the prices of listings for Airbnb properties during the holiday period and contrasted it with prices during the rest of the year. The heatmap in Figure 5 shows the percentage increase of prices during the holiday periods. The price of listings increases by more than a 100% on average during the holiday season. Only in certain zones in Brooklyn and Queens, is it possible to find a place with a slightly less than 100% increase in prices and is also close to Manhattan. Prices over the holiday season therefore play an important role determining people's choices of where to stay.

Methods and Models

In this section, we briefly explain the models we selected which includes K-Means clustering algorithm, Linear Regression, Random Forest and XGBoost.

Clustering

For our research, we used K-Means clustering on taxi pickups and drop-offs within each taxi zone to understand if we could categorize each zone into high mobility and low mobility zones for the holiday period and extrapolate inferences about the resulting clusters. After preprocessing the datasets to remove outliers, NAs and redundant columns, we ran the clustering algorithm on Yellow Taxi Trips independently the first time and the second time after appending both Yellow Taxi and FHV datasets together. The clusters we found were representative of the tourist hotspots, the results of which will be discussed in the next section.

Prediction

For our prediction task, we intend to predict the next hour pick demand at Times Square taxi zone.

1. Naive Forecast, First Order Differencing and Scaling

In time series prediction task, a popular baseline model is using the value from the current time x(t) as the prediction for the next time step x(t+1). To make sure that our regression models do not just fit a noise, we establish this baseline model to compare the performance of more advanced algorithm. First order differencing method is a standard technique to deal with nonstationary time series data. The label variable becomes x(t)-x(t-1). Thus we are modeling the changes in response value rather than the response value itself. In our case the taxi demand's nonstationarity comes from the seasonality and the negative effect by the rise of FHV. So we believe it's useful to apply first order differencing in this case. Also since we are using both lag values that vary from 1 to over 1000 and one-hot encoded categorical features as our regressor, we apply MinMax scaler to reduce all value range to [0, 1] to help with the training process.

2. Linear Regression

Next, in order to predict taxi demand within these tourist zones, we first created a linear regression model. Filtering the dataset for each taxi zone, we wished to predict the count of trips within each zone by equating the dependant variable to parameters such as bad weather conditions, holidays, demand for taxis in an N=24 period lagged window, day of the week and hour of the day. Based on our literature review, we decided to make run the regression from the months of July to October as most papers we reviewed talked about how a continuous time series based models with some feature engineering is enough to appropriately predict the demand for taxi services in the subsequent months. Therefore, our training dataset came to be values for the months of July to October while data for the month of November became our valid dataset and data for the month of December became our test dataset.

3. Random Forest Regression

A Single regression tree splits each time based on a feature and splitting criterion. The region of training data is split so that the predictive power of tree is maximized. The predicted value takes the average value of response variables in the selected region (Figure 6).

The depth of a tree determines the number of regions it can split. However a deep tree often suffers from overfitting the data. Ensemble Method such as bagging: train multiple models on bootstrap samples of training data and average the output, can reduce the variance of the model. The problem of bagging is that many trees are highly correlated due to the same splitting based on strong predictors. Random Forest is a modified version of bagging that creates independent ensembles of decision trees by selecting a random set of features from the full set of features at each split. Thus, three hyperparameters need to be tuned: max number features, max depth of tree and number of trees.

4. Xgboost

Gradient boosting tree is another method to address the shortcoming of Decision tree. One can additively combine simple models to build a complex model. Each new sample model added to ensemble compensates for the weakness of the current ensemble. The parameters to tune for gradient boost tree are number of trees, learning rate and max depth (Figure 7).

Results

Clustering Results:

The results of k-means clustering suggest zones with prominent tourist activities within them as shown in Figure 8.

To decide on the number of k's, the silhouette score algorithm was run, giving us the highest resulting score of 0.84 for k=3 clusters. Based on the results, it can be inferred that the two Manhattan based clusters are zones of high and mid mobility patterns amongst tourists. The high mobility zones include places like Times Square, Midtown, Penn Station, Upper Manhattan along with other zones such as East Chelsea and East Village. All these places are comprised of numerous tourist attractions including museums, shopping centers, famous high-rises, and other places like Times Square and the High Line (Table 3).

The second cluster seems to indicate zones with mid tourist mobility. These zones include places like Central Park, Battery Park City, Greenwich Village, Little Italy, SoHo etc. These zones are filled with museums, parks, murals, skyscrapers and also the access points to

Liberty Island. It's also possible to extrapolate the tourist demographics from this data to include people traveling with families and certain offbeat explorers who wander to the not-so-famous parts of the city after exploring zones with high tourist activities. This cluster also includes the two airports JFK and LaGuardia.

The results obtained from running the clustering algorithm over both the yellow taxi and FHV datasets appended together seem to suggest a new pattern for tourist mobility within the dataset. The silhouette score algorithm for this dataset was the maximum at k=2 clusters.

The results obtained from the clustering image in Figure 9 seem to suggest "Millennial Tourist Hotspot Zones". These are zones found to be amongst the most frequented by Ubers and Lyfts along with the usual tourist spots within Manhattan. Amongst the list of places included, are names like Astoria, Williamsburg, Bushwick and Park Slope. These are places garnering a lot of social media buzz. The list of attractions in each of these zones include scenic parks, murals and graffiti and comedy clubs. The areas around Brooklyn and Queens also happen to be the ones where the density of AirBnB check-ins is amongst the highest, we can see that from figure 10. Both the density heatmap and the clustered map for tourist hotspots coincide, thus making it the millennial tourist hotspots (Table 4).

Linear Regression Diagnostics:

Linear Regression was a poor choice for modeling time series data, which is characterised by high seasonality and non-linearity. The results obtained suggested overfitting, with high p-values for the coefficients used. After diagnosing the model with a feature selection algorithm, we found bad weather conditions, holidays and rush hours to be important features to predict taxi demand with 75% accuracy for both our test and valid datasets. However, we chose to run other models in order to account for the non-linearity of the data.

Prediction Results

Regression Test Results evaluated by root mean squared error

	Linear Regression	Random Forest	Xgboost	Naive
Tuned	60.59	58.74	58.47	91.23
MinMax scaled and First order Difference	62.62	59.04	57.78	91.23
Selected Features	66.1	60.02	58.07	91.23

	Training RMSE	Test RMSE
Naive	91.23	91.23
Linear Regression	53.77	60.59
Random Forest	18.87	58.74
Xgboost	13.63	57.78

Although we see some improvement of RMSE score with complex model , the actual forecasting plot shows that these models did not learn to capture the anomaly events like New Years Eve in December (Figure 12). The result improves compared to the baseline as the models learn to fit a “regular day” better. We think either there is a better way to encode the anomaly events, or we need to find training dataset that includes patterns, for example, previous year’s December data.

Also from the training vs test RMSE table, we see the model that achieves better score also has strong overfitting. Again, simple model such as linear regression seems able to learn the regular pattern, but the training data seems not to capture the pattern in test datasets.

Conclusion

Through our analysis, the study was successful at finding hotspots with high and mid tourist mobility patterns within New York City. It has also been able to successfully highlight the prediction of demand for taxis within these zones during the holiday period. These results therefore suggest the variety of ways in which such context specific studies could be made from publicly available datasets to study changing trends within the tourism industry and how they model the ever-changing urban fabric of our cities.

References

[1] Alexandra, A. (Jan 2019). "NYC drew a record 65 million tourists". Retrieved from https://www.6sqft.com/in-2018-nyc-drew-a-record-65-million-tourists-and-that-number-will-keep-rising/

[2] Song, S., Xia, T., Jin, D., Hui, P., & Li, Y. (2019). UrbanRhythm: Revealing Urban Dynamics Hidden in Mobility Data. *arXiv preprint arXiv:1911.05493*.

[3] Dickinson, G. (May 2018). “From Ubercopters to Uberboats: how the ride-sharing app changed the way we travel”. Retrieved from https://www.telegraph.co.uk/travel/comment/where-can-you-get-uber-overseas-map/

[4] Chang, H. W., Tai, Y. C., & Hsu, J. Y. J. (2010). Context-aware taxi demand hotspots prediction. *International Journal of Business Intelligence and Data Mining*, *5*(1), 3.

[5] Chang, H. W., Tai, Y. C., Chen, H. W., Hsu, J. Y. J., & Kuo, C. P. (2008). iTaxi: Context-aware taxi demand hotspots prediction using ontology and data mining approaches. *Proc. of TAAI*.

[6] Dewi, A. & Silipo, R. (September 2019). “Time Series Analysis : A simple example with KNIME and Spark”. Retrieved from https://www.knime.com/blog/time-series-analysis-a-simple-example-with-knime-and-spark

[7] Yao, H., Wu, F., Ke, J., Tang, X., Jia, Y., Lu, S., ... & Li, Z. (2018, April). Deep multi-view spatial-temporal network for taxi demand prediction. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[8] Vanichrujee, U., Horanont, T., Pattara-atikom, W., Theeramunkong, T., & Shinozaki, T. (2018, May). Taxi Demand Prediction using Ensemble Model Based on RNNs and XGBOOST. In *2018 International Conference on Embedded Systems and Intelligent Technology &*

International Conference on Information and Communication Technology for Embedded Systems (ICESIT-ICICTES) (pp. 1-6). IEEE.

Appendix

Figures:

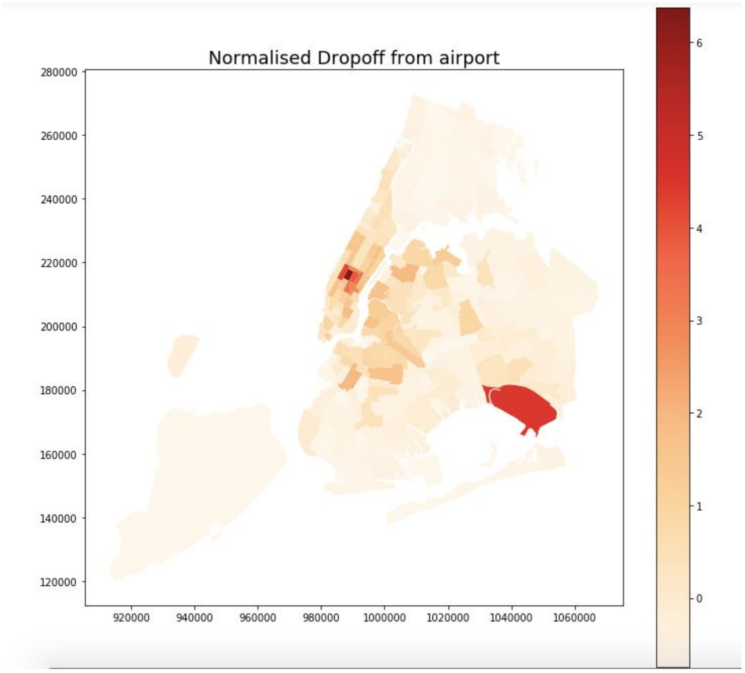

Figure 1. Normalised Drop-offs Location From Airport

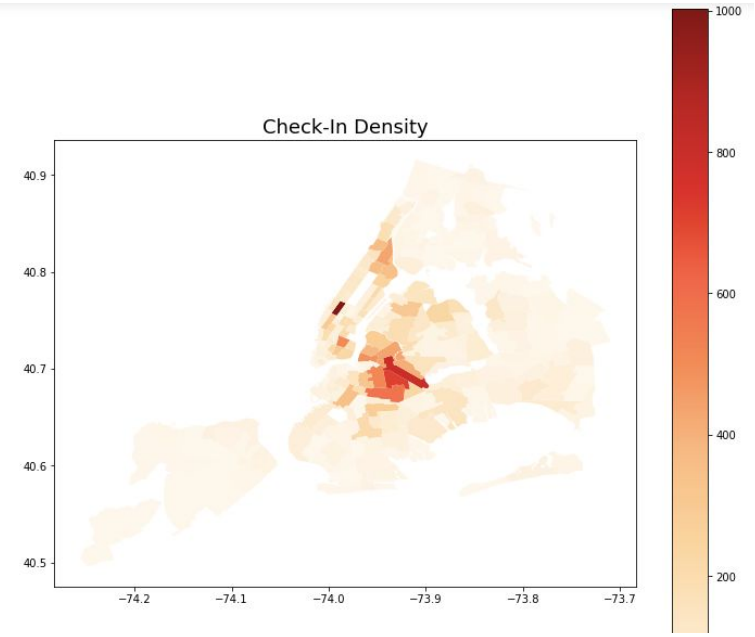

Figure 2. Check-in Density Map for Airbnb Properties

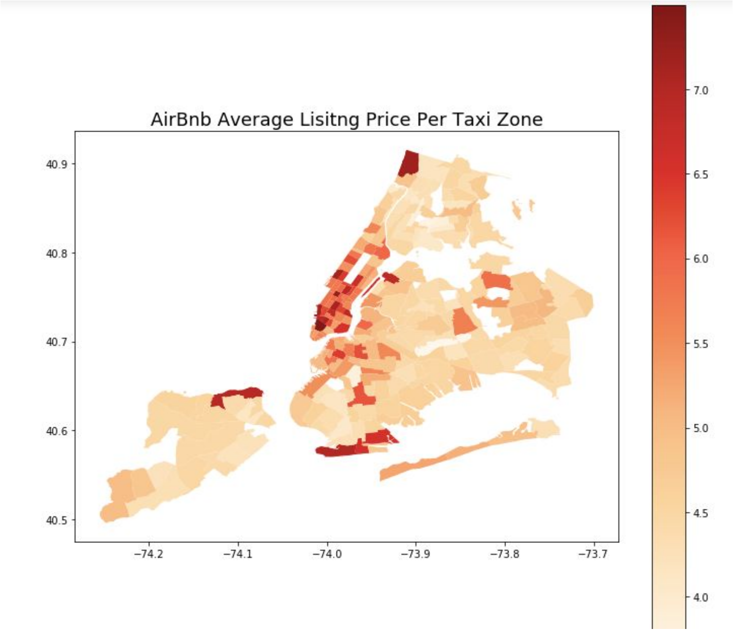

Figure 3. Airbnb Standardised Average Listing Price Per Taxi Zone

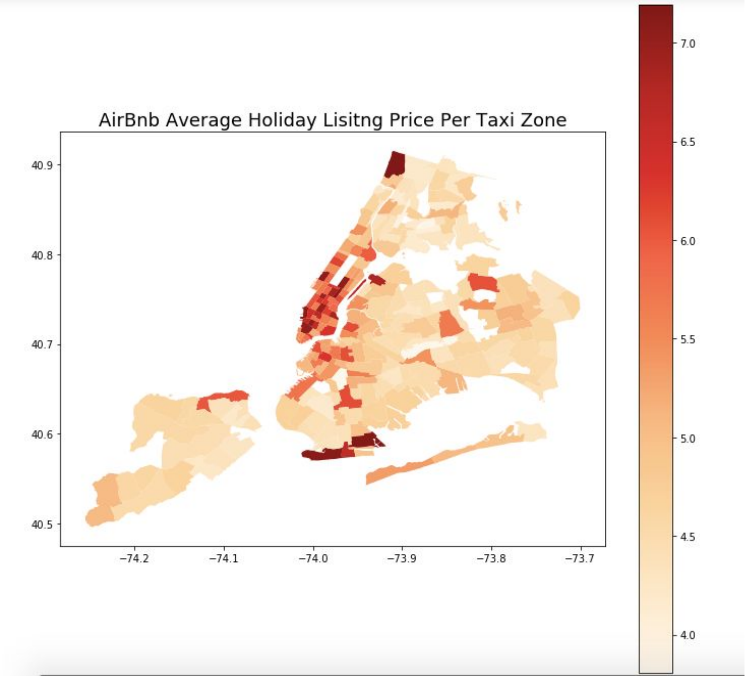

Figure 4. Airbnb Average Holiday Listing Price Per Taxi Zone

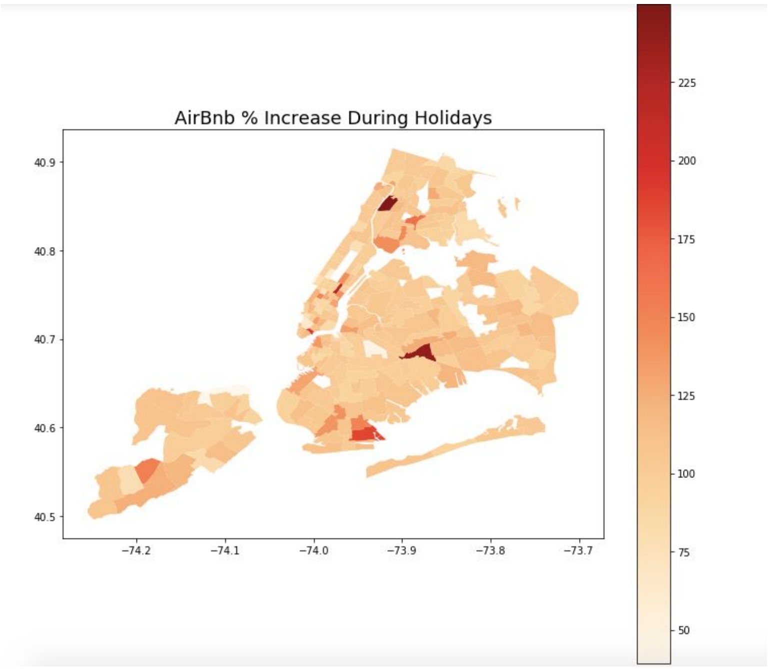

Figure 5. The Percentage Increase of Prices of Airbnb During Holidays

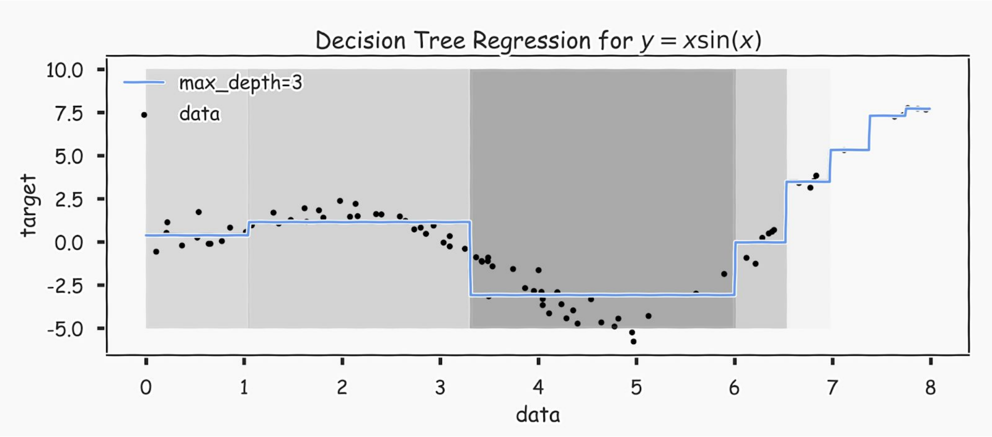

Figure 6. Decision Tree Regression for y = xsin(x)

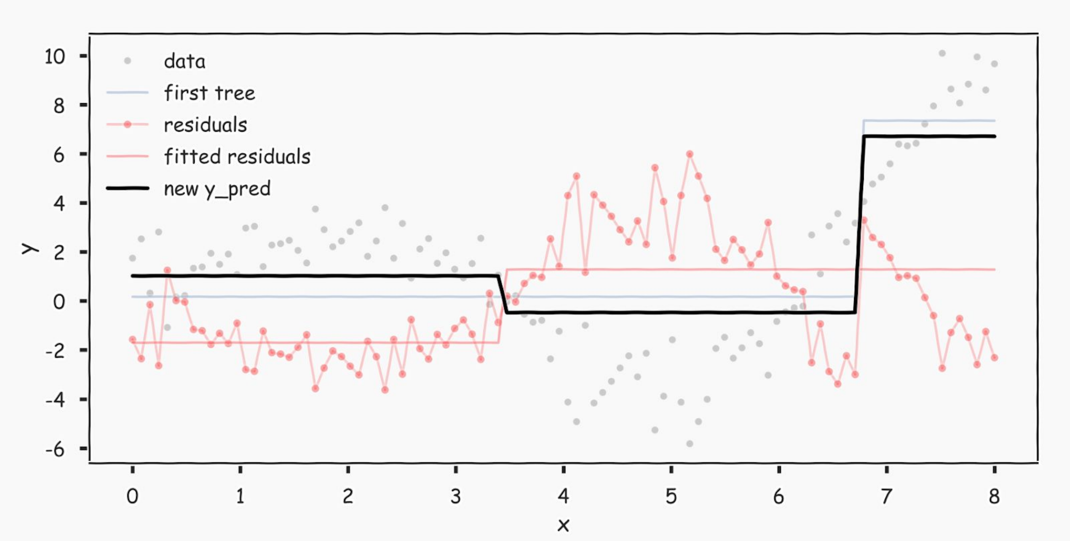

Figure 7. Introduction of Xgboost

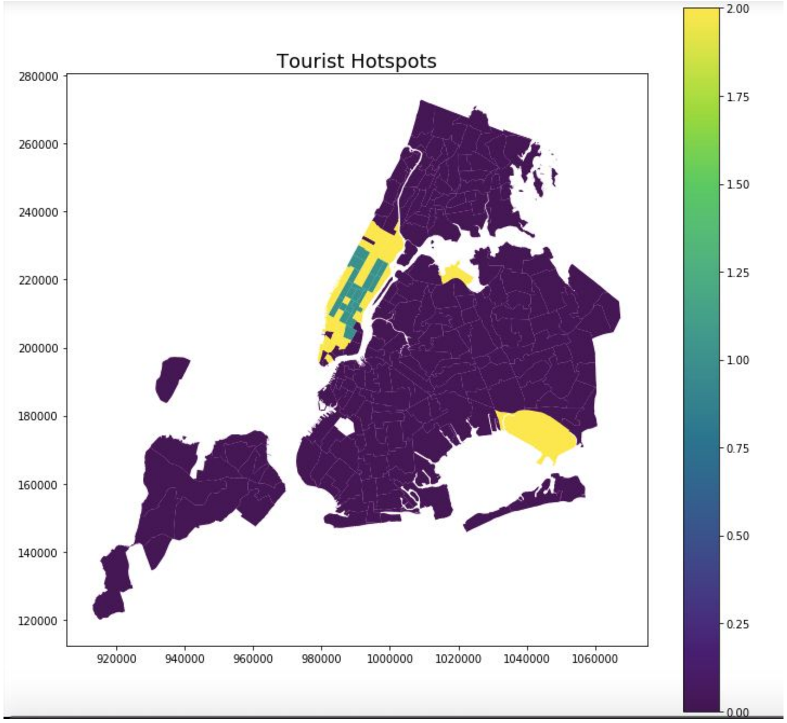

Figure 8. Tourist Hotspots with K-Means Clustering on Yellow Taxi Data

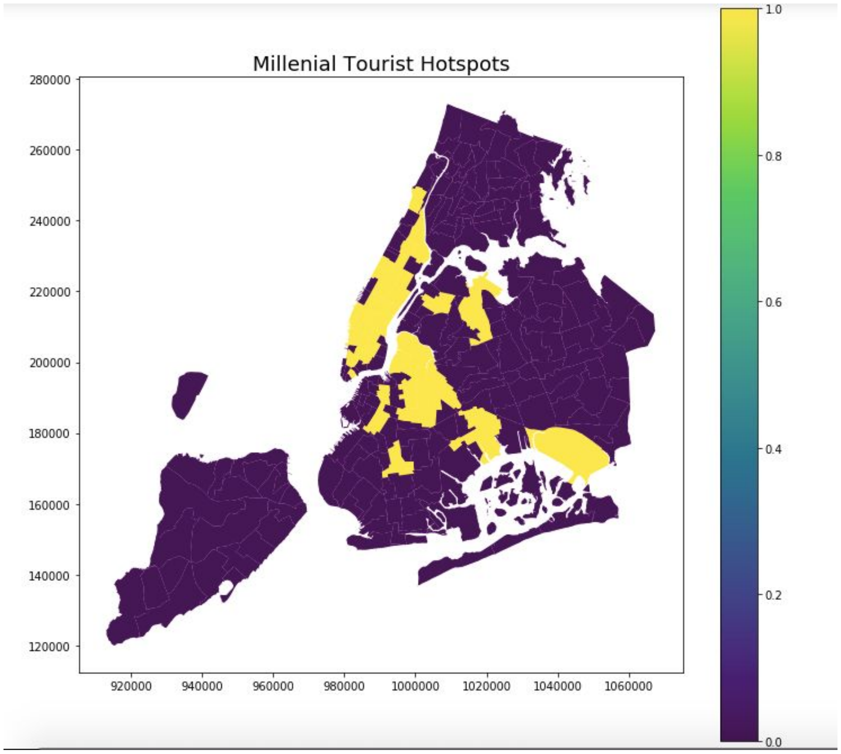

Figure 9. Millennial Tourist Hotspot Zones obtained through Yellow Taxi and FHV Data

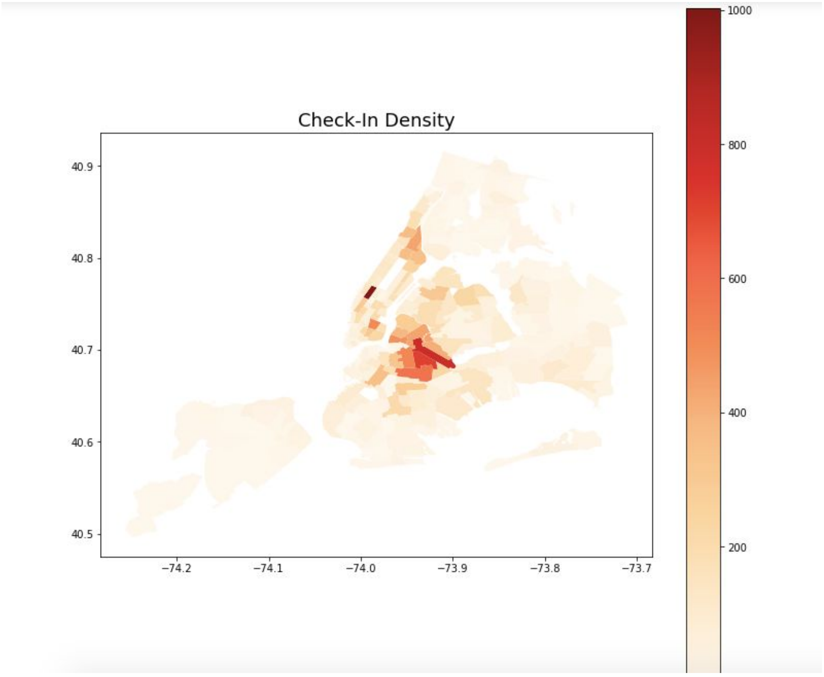

Figure 10. The Density of AirBnB Check-ins

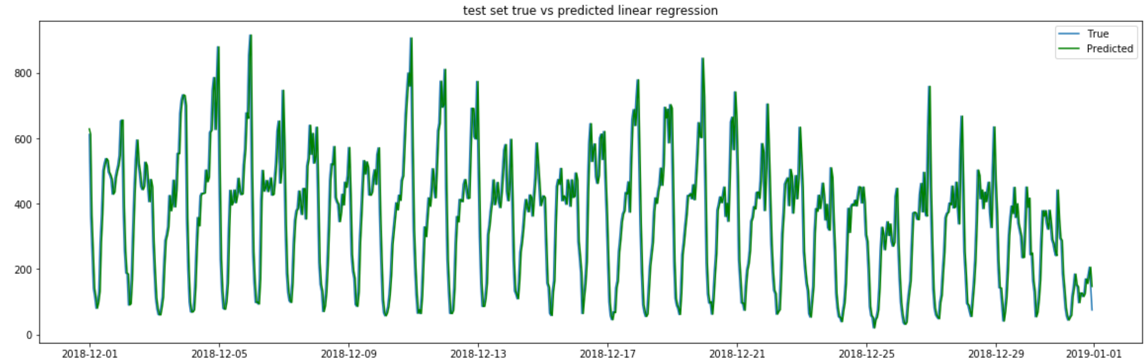

Figure 11. Naive Forecast Result for December

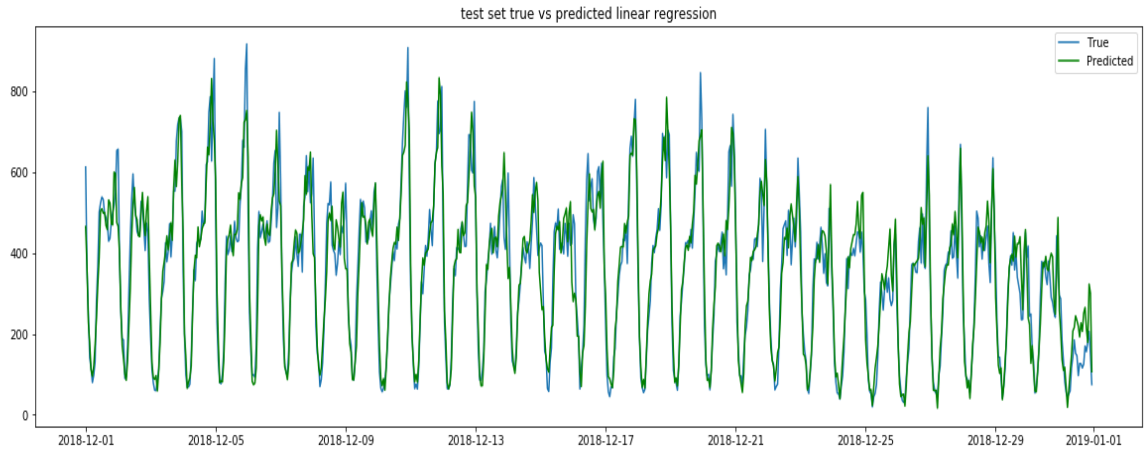

Figure 12. Regression Forecast Result for December

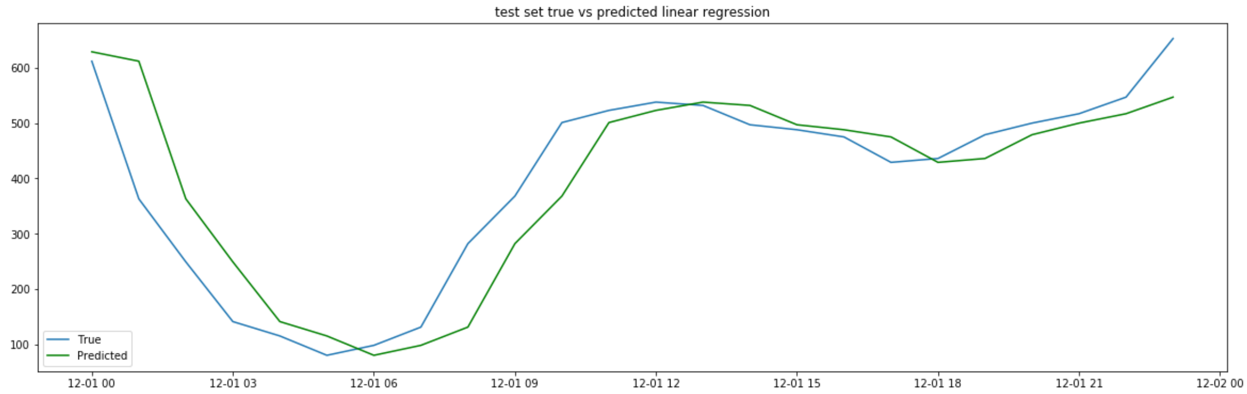

Figure 13. Naive Forecast Result for December 1st

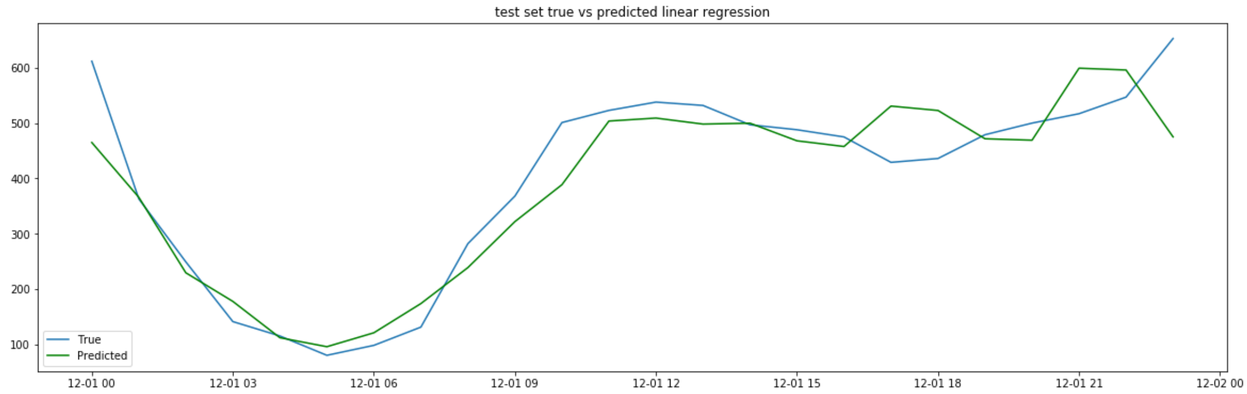

Figure 14. Regression Forecast Result for December 1st

Tables:

Features	**Feature Description**
Pickups	Pickups within each taxi zone across NYC
Drop-Offs	Drop-Offs within each taxi zone across NYC
Drop-offs filtered for pickups from airports	Feature to analyse mobility directly from the airport
Prices of Airbnb properties across the year	Price of each listing in a year averaged out for each taxi zone within NYC
Prices of Airbnb properties across the holiday period	Price of each listing in the holiday period, averaged out for each taxi zone in NYC
Number of Check-ins	Information about stays in Airbnb properties across NYC extracted from reviews dataset

Table 1. Description of Features for clustering and EDA

Features	**Feature Description**
Count of Trips	Pickups within each taxi zone across NYC
Hour of the Day	One Hot-Encoded feature for each hour of the day
Day of the Week	One Hot-Encoded feature for each day of the week
Bad Weather Condition	Weather data filtered for rainfall greater than 1.5 mm and snowfall greater than 2 inches
National Holiday	Whether the given date is a holiday or not

24 Lag Features	Time lagged variables for a (T-24) period, given current time T

Table 2. Description of Features for Prediction

Zone	**Location ID**	**Borough**	**Taxi Zone**	**Normalised Pickups**	**Normalised Dropoffs**	**Cluster**
Clinton East	48	Manhattan	48	3.2885188214339500	3.7081512275910500	1
East Chelsea	68	Manhattan	68	2.226414963818380	2.4563696014787400	1
East Village	79	Manhattan	79	2.581737483418240	2.512622427018680	1
Gramercy	107	Manhattan	107	2.01542928130647	2.05463067241436	1
Lenox Hill West	141	Manhattan	141	2.0677272366184500	2.6518481702300200	1
Lincoln Square East	142	Manhattan	142	3.6401156291778100	3.3884476691057500	1
Midtown Center	161	Manhattan	161	4.153711907993600	3.870971905959420	1
Midtown East	162	Manhattan	162	3.1926162384739900	2.8709216741383300	1
Midtown North	163	Manhattan	163	2.757121919306170	2.592782703413090	1
Midtown South	164	Manhattan	164	2.5519317885702500	2.5243417656728300	1
Murray Hill	170	Manhattan	170	2.567662571962250	3.094995429205740	1
Penn Station/Madison Sq West	186	Manhattan	186	3.5563561070818400	3.1610924992151700	1
Times Sq/Theatre District	230	Manhattan	230	3.8515152796737200	4.064887896223480	1
Union Sq	234	Manhattan	234	2.9561025302820900	2.3988667131490300	1
Upper East Side North	236	Manhattan	236	3.594441161609830	4.1856752132856200	1
Upper East Side South	237	Manhattan	237	4.075057991033630	3.895035614662620	1

Upper West Side North	238	Manhattan	238	1.585316545930650	2.325269266400940	1
Upper West Side South	239	Manhattan	239	2.869997189610120	2.959519874363730	1
Yorkville West	263	Manhattan	263	1.5985635214186400	2.095570228779540	1

Table 3. High Mobility Tourist Zones

Zone	**Location ID**	**Borough**	**Tazi Zone**	**Normalised Pickups**	**Normalised Drop Offs**	**Cluster**
Battery Park City	13	Manhattan	13	0.4910611748910900	0.6276839978846350	2
Central Harlem	41	Manhattan	41	-0.02667478376469880	0.4006413436915020	2
Central Park	43	Manhattan	43	1.7509037395305800	1.433505723744230	2
Clinton West	50	Manhattan	50	0.4371073476431120	0.8906659572838390	2
East Harlem North	74	Manhattan	74	0.005062761675288250	0.5481487528851140	2
East Harlem South	75	Manhattan	75	0.30201579553116700	0.8950411770480560	2
Financial District North	87	Manhattan	87	0.2381267366671930	0.34548232309262000	2
Flatiron	90	Manhattan	90	1.3344519476267500	1.0961450283533300	2
Garment District	100	Manhattan	100	1.7739479573065700	1.5480427268574900	2
Greenwich Village North	113	Manhattan	113	1.3116837085067600	1.128646660887520	2
Greenwich Village South	114	Manhattan	114	1.2088816591468000	0.8559767148675450	2
JFK Airport	132	Queens	132	2.8084539493221500	0.5142407997124300	2
Kips Bay	137	Manhattan	137	0.7738013079629750	1.0372358193851200	2
LaGuardia Airport	138	Queens	138	2.4756236901862800	0.31469952689437700	2
Lenox Hill East	140	Manhattan	140	1.1612753409868200	1.3388134674186700	2

Lincoln Square West	143	Manhattan	143	0.708118387835002	1.132709364954290	2
Little Italy/NoLiTa	144	Manhattan	144	1.0136267600268800	0.9142608924408680	2
Lower East Side	148	Manhattan	148	1.0190083438188800	0.9823893144836800	2
Manhattan Valley	151	Manhattan	151	0.5078958729070840	0.6944060992889490	2
Meatpacking/West Village West	158	Manhattan	158	0.7653839589549790	0.8726963046808030	2
Morningside Heights	166	Manhattan	166	0.07861107349925830	0.2868856298218530	2
SoHo	211	Manhattan	211	0.6587182084110220	0.7305016623437420	2
Sutton Place/Turtle Bay North	229	Manhattan	229	1.4706474143626900	1.8799343975431100	2
TriBeCa/Civic Center	231	Manhattan	231	1.2277861970827900	1.2658410520654700	2
UN/Turtle Bay South	233	Manhattan	233	0.7634521083629800	1.153022885288150	2
West Chelsea/Hudson Yards	246	Manhattan	246	0.7807007743629730	1.2108382893153100	2
West Village	249	Manhattan	249	1.8134129051145500	1.5177287042054100	2
Yorkville East	262	Manhattan	262	0.6629958775790210	1.2491214622522100	2

Table 4. Low Mobility Tourist Zones

Individual Roles and Contribution:

Akash Yadav: Data cleaning and running all clustering algorithms to identify tourist hotspots along with working on report and presentation

Chenjie Su: Did preliminary exploratory data analysis and extracting network features (not used in the report) and worked on the report

Siqi Huang: Feature engineering and ran all models for predicting taxi demand, along with working on report and presentation

Yue Jin: Ran diagnostics on linear regression, feature engineering on decision tree & random forest, and ran all the models for few different taxi zones identified as tourist hotspots