APSTA-GE 2047: Messy Data and Machine Learning

Final Project

Jui Nerurkar, Shreya Singhal

How can yellow taxi cab drivers maximize their profits in New York City?

Introduction:

With the recent advent of ridesharing apps, such as Lyft and Uber, New York City taxi cab drivers are finding it harder and harder to make ends meet. The medallion required to operate an NYC taxi, once valued at one-million dollars and highly sought-after, means that there are taxi drivers drowning in huge amounts of debt or paying to rent their taxi cabs on a daily or weekly basis. Now, more than ever, taxi drivers must make sure that they are operating their taxis when it will prove to be most lucrative for them.

However, in a city as dynamic and fast-paced as New York City, understanding the pattern of the infamous yellow taxi-rider is an arduous task. There is a substantial amount of literature that focuses on studying the effects of temporal factors on taxi ridership within New York City. Of course, it is rather intuitive that travel patterns are strongly dictated by say, the onset of rush hour or Friday night, but other temporal patterns may be less obvious, which is why they require additional investigation. For example, in [1], the authors find that taxi drivers have a better chance of getting long taxi trips at night, rather than during the day.

Additionally, there is a large amount of literature that explores the effects of weather on taxi ridership within New York City. For example, previous studies [2] show that although rain alone does not seem to affect overall daily ridership, snowfall, in general, has a negative impact on daily taxi ridership.

In order for taxi drivers to truly reap the benefits of the supply-and-demand nature of this industry, they must be attuned to the constant variation in taxi ridership, as well as the trip information readily available to them before the start of the journey. Investigating how the citizens of New York City consciously adjust their taxi use according to time and changes in weather, both short-term and long-term, has the potential to provide valuable insights that would allow taxi drivers to take trips that are most profitable to them personally, which is how our research question came to be. We are interested in studying how yellow taxi cab drivers, by paying attention to time, weather, and predetermined trip information, can maximize their monetary profits in New York City.

Data:

Our goal was to determine the effects of weather and time on taxi ridership, as well as the ways in which taxi drivers can use specific trip-related information available to them to maximize their monetary profits. Therefore, the final dataset used to construct the model merges data from two online sources: NYC Taxi & Limousine Commission and National Climatic Data Center.

As of now, data originating from the NYC Taxi & Limousine Commission (TLC) provides historical data on taxi services from January 2009 to June 2018. This data encompasses both yellow taxi and green taxi trip records; however, owing to the huge size of the dataset, for our purposes, trip records only involving yellow taxis were utilized. The trip records are contained in CSV files, each of which includes trip data from a particular month of the year.

As of now, data originating from the National Climatic Data Center (NCDC) provides historical data on weather and climate in the United States for the last 30 years. This data provides a comprehensive selection of variables pertinent to weather -- including, but not limited to, daily maximum/minimum temperature, snowfall, and precipitation level. The weather records are contained in CSV files, each of which includes daily weather data from a particular month of the year.

For the main analyses of the study, the data for yellow taxi rides and weather for the Central Park station was merged, by date, to create a dataset with each row representing one unique taxi ride. Due to limitations with respect to computational processing and time, the training dataset consisted only of randomly sampled 60000 taxi records and weather data from the entire first half of 2018 (January 2018 - June 2018), while the testing dataset consisted of randomly sampled 10000 taxi records and weather data from January 2017.

There are a substantial amount of variables contained in the original dataset, several of which were utilized either directly or indirectly. Indirect use comprised of extracting specific features from existing variables and using those features to create additional variables. For example, from the original TLC variables ‘pickup time/date’ and ‘drop-off time/date’, features such as ‘ride duration’, ‘hour of the day’, and ‘day of the week’ were created to provide more valuable, relevant information.

Data cleaning was an integral component of the process, as there were several values in the TLC dataset that did not align with the real-world operation of taxis or the geographical location of New York City. Firstly, we separated pick-up and drop-off dates from the ‘pick-up datetime’ and ‘drop-off datetime’ columns. We also deleted trips which, when calculated using the pick-up and drop-off time values, were documented to last longer than 24 hours, as, presumably, these trips would have spanned outside New York City, and we were not interested in such trips. For the same reason, we eliminated all values for which the variable ‘trip distance’ was greater than 50 miles, as our geographic location of interest does not span such a large area. Thereafter, we eliminated taxi trips indicating passenger counts equal to 0 or greater than 4, as this does not align with real-world taxi capacity. To better understand variation in taxi ridership based on time, we created a variable ‘weekday’, which encoded the day of the week on which the ride was taking place.

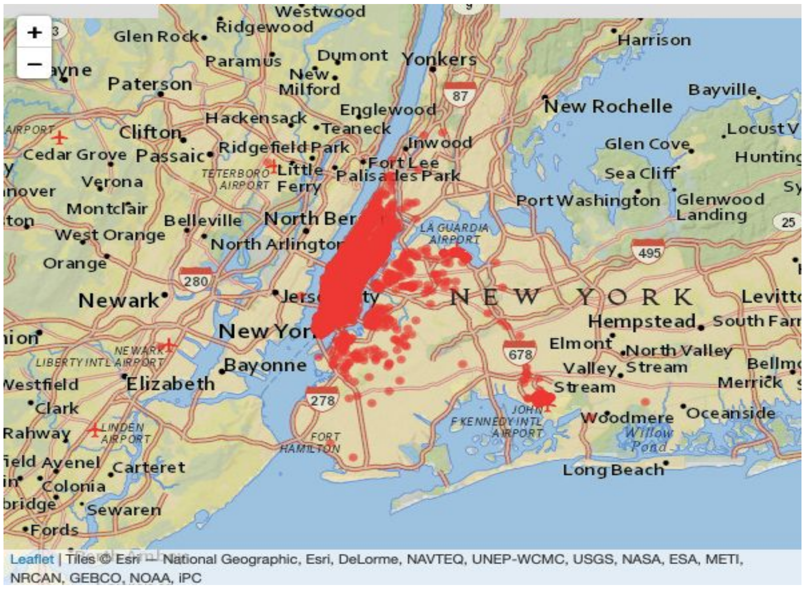
On the other hand, the NCDC dataset was fairly clean, and did not require a substantial amount of processing. Six variables - ‘maximum temperature’, ‘minimum temperature’, ‘average wind speed’, ‘precipitation’, ‘snowfall’, and ‘snow depth’ - were used to quantify weather.

Once we completed data cleaning, we merged the monthly taxi data with the monthly weather data. With this merge, we obtained six different datasets, each representing one month between January 2018 and June 2018. Each of these 6 datasets had 42 columns and approximately 8 million rows. As it would be computationally difficult to merge the data from all six months, we randomly sampled 10,000 rows from each dataset. Hence, we obtained a training data set consisting of 60,000 rows and 42 columns.

For the test dataset, we applied a similar cleaning procedure and sampled 10,000 rows from the cleaned dataset for January 2017.

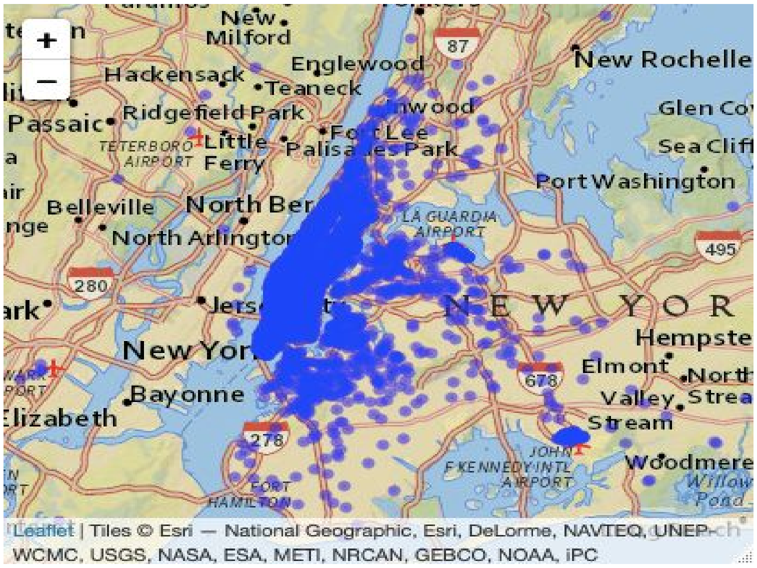
Preliminary Analysis:

Instead of jumping directly into the analysis, we first proceeded by acquainting ourselves with the TLC dataset, as it was rather large and contained many variables. We looked at the taxi trip records from January 2015. This dataset had the latitude and longitude coordinates for the pick-up and drop-off location of every ride documented for that month. Using the leaflet package from R, we created two maps, one for the pick-up coordinates of every ride in New York City during January 2015 and the other for the drop-off coordinates of every ride in New York City during January 2015.



Leaflet | Tiles © Esri — National Geographic, Esri, DeLorme, NAVTEQ, UNEP-WCMC, USGS, NASA, ESA, METI, NRCAN, GEBCO, NOAA, iPC

The first map shows the pick-up coordinates of all taxi rides in this particular month. It shows that the majority of taxi rides begin in the New York City borough of Manhattan. The map also shows that there is a higher concentration of taxi pick-ups near the two airports situated in New York City, LaGuardia and JFK. However, other than that, taxi pick-ups are extremely sparse in the remaining boroughs contained within New York City.

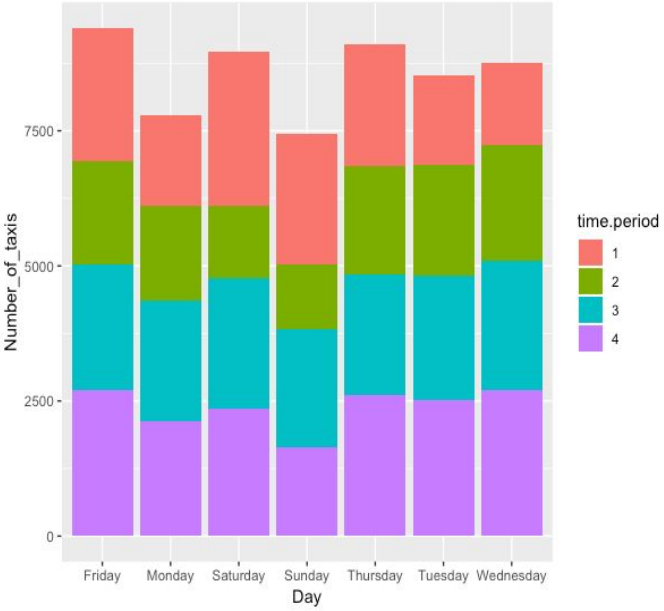


Leaflet | Tiles © Esri — National Geographic, Esri, DeLorme, NAVTEQ, UNEP-WCMC, USGS, NASA, ESA, METI, NRCAN, GEBCO, NOAA, iPC

The adjacent map depicts the drop-off coordinates of all taxi rides during this particular month. Similarly, taxi drop-offs seem to be highly concentrated in the New York City borough of Manhattan. However, this map shows that there are also indications of drop-off coordinates in Brooklyn and Queens. As expected, the drop off coordinates also seem to be concentrated near the two airports, LaGuardia and JFK.

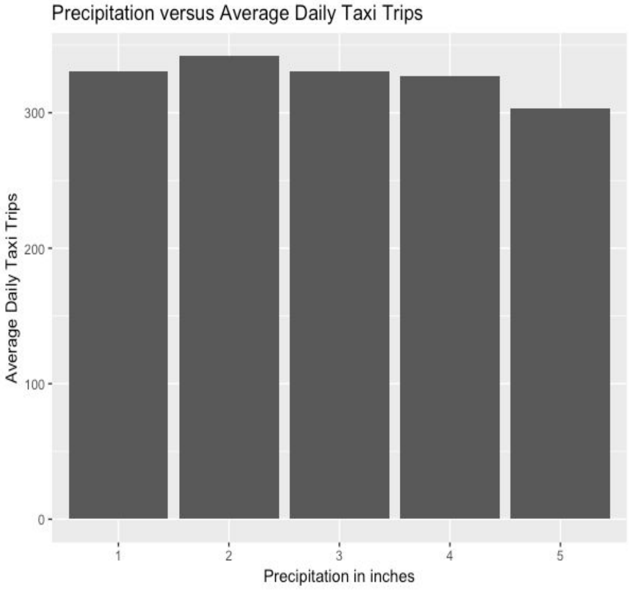
From these maps, we realized that the majority of the taxi rides represented by this dataset are taken within the New York City borough of Manhattan. Hence, we decided it would be most appropriate, and possibly more productive, to collect weather data originating from the station at Central Park. With this data, we strove to analyze the effect of weather conditions on the frequency of taxi rides.

Thereafter, we used the newly-merged dataset from January 2018 to June 2018 to analyze the various associations between the day of the week, time of the day, daily precipitation and daily snowfall and the frequency of taxi ridership.

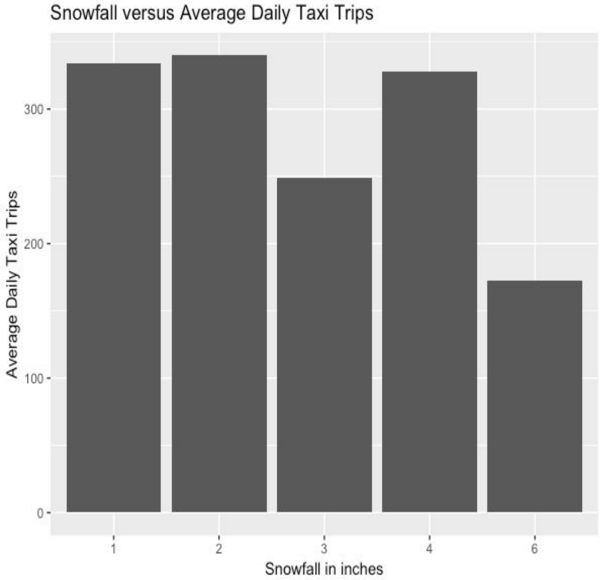


The graph to the left displays the total number of taxis hired between January 2018 and June 2018, based on the day of the week. The subdivisions of the bar graphs represent the number of taxis hired during the different time periods of the day. Time period 1 represents the time of the day between 12:00 AM and 6:00 AM. Time period 2 represents the time of the day between 6:00 AM and 12:00 PM. Similarly, time periods 3 and 4 represent the former and latter halves of the day, respectively. From the graph, we can say that Thursdays and Fridays see a higher number of taxi ridership compared to other days of the week. Time period 2, which could be characterized as rush-hour, sees the largest number of taxi riders on weekdays (i.e. Monday through Friday). Additionally, the number of taxi riders traveling during early morning increases as the weekend approaches, and is greatest on Fridays, Saturdays and Sundays.

Thereafter, we also analyzed the association between precipitation and the average amount of daily taxi trips taken. The graph below suggests that precipitation has no effect on the average number of taxi ridership, i.e. whether or not it will rain does not have a significant association with taxi ridership.



On the other hand, the graph to the right suggests a strong association between daily snowfall and taxi ridership. The bars numbered 1-6 represent the snowfall amounts to be 0 mm, 0-2 mm, 2-4 mm, 4-6 mm and >8 mm, respectively. Taxi ridership decreases significantly when the amount of snowfall ranges between 2 and 4 mm or is greater than 6 mm. Surprisingly, there is an increase in taxi ridership when the amount of snowfall is between 4 and 6 mm as compared to when the amount of snowfall is between 2 and 4 mm; however, taxi ridership is still less than when snowfall is less than or equal to 2 mm. This discrepancy could be due to the fact that there is an insufficient amount of data for taxi ridership on days with greater levels of snowfall.



Supervised Learning Method:

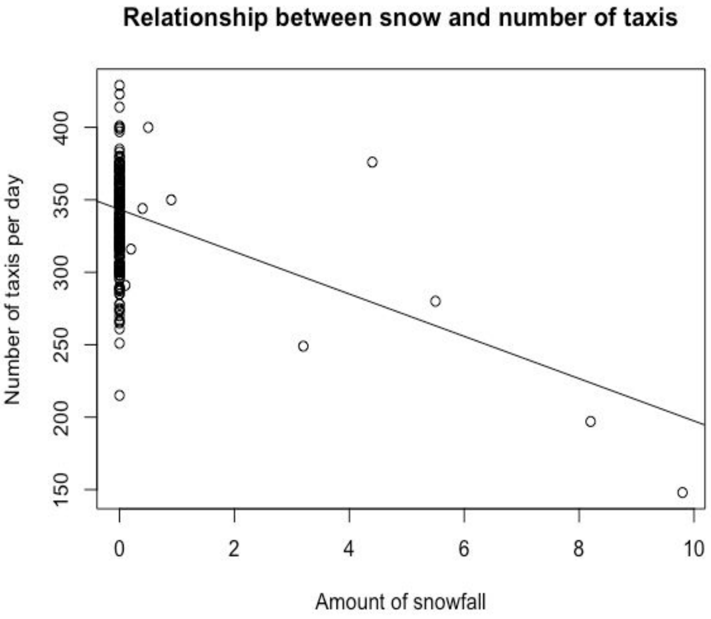
As our overall goal was to determine how to maximize taxi drivers' profits based on the data available to them, we decided to focus on which aspects of weather increased the demand of yellow taxis in New York City. However, as an added criterion, we decided to predict which trips were most likely to yield a tip of twenty percent or more, as the additional revenue from tips can prove to be very lucrative.

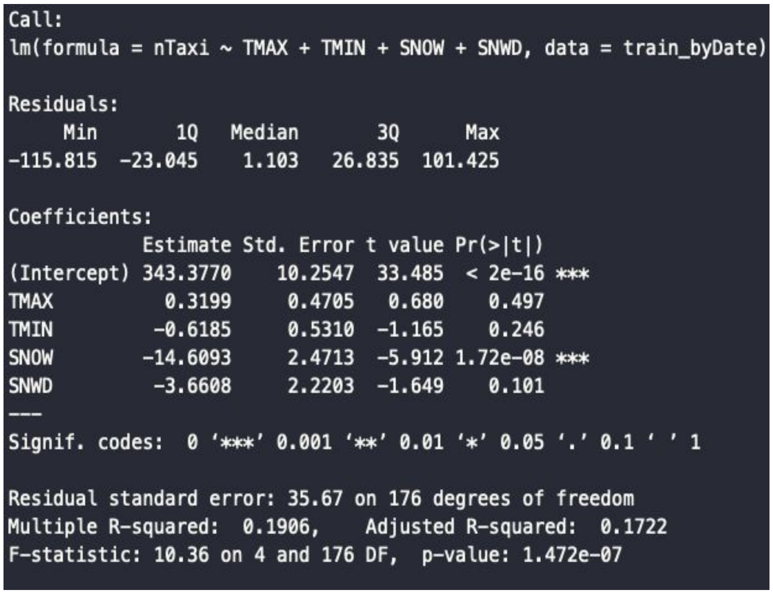
First of all, we decided to examine the relationship between the number of taxis hired daily and the accompanying weather conditions. We utilized a linear model, where the number of daily taxis was our response variable and daily ‘minimum temperature’, ‘maximum temperature’, ‘amount of snowfall’, and ‘snow depth’ were our predictors.

Thereafter, for predicting which trips would result in a tip for the driver, we used a Naive Bayes model with a binary variable for whether a tip of twenty percent or more was given as the response variable and ‘trip distance’, ‘rate code’, ‘pick-up location’, ‘pick-up hour’, ‘day’, and ‘month’ as our predictors. All of our predictors represented information that would be readily available to the taxi driver before the start of the trip.

Results:

The linear regression model had a p-value of less than 0.05 and hence, we observe a significant relationship between snowfall and the total number of yellow taxi rides taken on one particular day. However, there is a negative relationship, which can be observed in the linear model, i.e., it suggests that an increase in snowfall is associated with a decrease in the number of taxi rides taken in New York City. Nonetheless, there are only ten data points in the dataset in which snowfall is greater than 0 mm. Therefore, we believe that this analysis can only be confirmed by incorporating more data. Due to computational restrictions in the R software, it was not possible to merge all the data from January 2018 to June 2018 into one dataset.





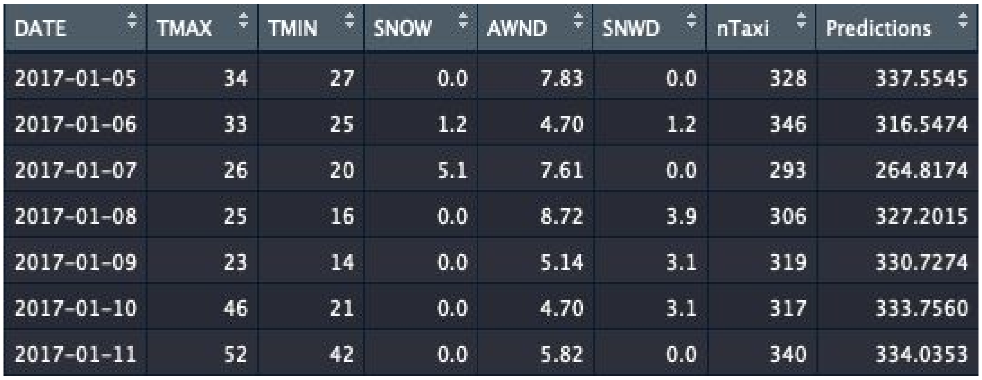
```
Call:
lm(formula = nTaxi ~ TMAX + TMIN + SNOW + SNWD, data = train_byDate)

Residuals:
     Min       1Q   Median       3Q      Max 
-115.815  -23.045    1.103   26.835  101.425 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 343.3770    10.2547  33.485  < 2e-16 ***
TMAX          0.3199     0.4705   0.680    0.497    
TMIN         -0.6185     0.5310  -1.165    0.246    
SNOW        -14.6093     2.4713  -5.912 1.72e-08 ***
SNWD         -3.6608     2.2203  -1.649    0.101    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35.67 on 176 degrees of freedom
Multiple R-squared:  0.1906,    Adjusted R-squared:  0.1722 
F-statistic: 10.36 on 4 and 176 DF,  p-value: 1.472e-07
```

For predicting whether a trip will yield a tip of twenty percent or more, we selected four features, which are based on information any taxi driver would know before the trip begins. The four features are as follows: whether the trip distance is greater than 10 miles, whether the trip is taken on Friday, Saturday or Sunday, whether the pick-up hour is between 6:00 PM and 6:00 AM, and the rate code corresponding to the trip, i.e., standard rate fare or airport rate fare. Using these features, we are able to predict whether the trip will yield a tip amount of twenty percent or higher, with approximately 71% accuracy.

Implications:

According to our analysis, there are discernible trends in taxi-riders' behavior, which is influenced by factors such as weather and time of day. By understanding these trends, taxi drivers can maximize their profits, as they can aim for availability at times when demand for taxis is highest. As per our first graph, there is a high demand for taxi ridership between 6:00 AM and 12:00 PM on weekdays, whereas there is a significantly greater demand for taxis between midnight and early mornings on weekends.

Furthermore, using the linear model we have created, we are able to predict the amount of taxi ridership in one day, using the knowledge we have on the daily weather conditions, as done in the table below. The decrease in the number of taxi rides when the level of snowfall is high could be attributed to the increasing significance of ridesharing competitors like Uber and Lyft, which have apps that can determine the real-time demand for rides. However, the alternative way to maximize profits is to minimize loss. The negative relationship between snowfall and taxi ridership could also suggest that on days with high snowfall, people tend to stay indoors and hence, instead of being on the road unnecessarily, taxi drivers can save fuel on those days and minimize



DATE	TMAX	TMIN	SNOW	AWND	SNWD	nTaxi	Predictions
2017-01-05	34	27	0.0	7.83	0.0	328	337.5545
2017-01-06	33	25	1.2	4.70	1.2	346	316.5474
2017-01-07	26	20	5.1	7.61	0.0	293	264.8174
2017-01-08	25	16	0.0	8.72	3.9	306	327.2015
2017-01-09	23	14	0.0	5.14	3.1	319	330.7274
2017-01-10	46	21	0.0	4.70	3.1	317	333.7560
2017-01-11	52	42	0.0	5.82	0.0	340	334.0353

loss. In the future, we can further improve these predictions by analyzing the location-wise demand within Manhattan for taxis on days with higher amounts of snowfall. This information is extremely important for taxi drivers, as on days with inclement weather, they can maximize their profits by being available in areas with higher demand.

Lastly, tips are an important addition to a taxi driver's income. As mentioned earlier in the paper, since the majority of taxi drivers do not have the financial means to buy their own medallion, they must pay to rent the taxi cars they drive on a daily or weekly basis, which ultimately becomes very costly. By looking for rides in which they will be more likely to earn a tip, taxi drivers can offset some of the costs associated with driving the taxi agency's cab.

Conclusions:

One of the important concerns while using data and machine learning methods is the potential breach of privacy of the people from whom data is collected. Most of the rideshare apps today are passenger-specific, and hence, the use of such data could give rise to privacy concerns. Instead, the approach we suggest in this paper is to find ways to predict demand for taxis without using passenger-specific personal data. Although the locations from which passengers are picked up and dropped off could contain hidden trends, the information in these datasets cannot be directly linked to any particular passenger. We conclude by stating that our analysis is an addition to the pre-existing machine learning and exploratory analyses conducted on the taxi datasets. In the future, we'd like to use this and possible additional datasets to explore potential ways to predict the exact location of high demand at any given time in Manhattan.

References

[1] Xu, L. (2017, February 1). *The Temporal and Weather Data Analysis on NYC Yellow Taxi Ridership Demands*. Retrieved from https://www.authorea.com/users/106322/articles/143566-the-temporal-and-weather-data-analysis-on-nyc-yellow-taxi-ridership-demands/_show_article#

[2] Schneider, T.W. (2015, November). *Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance.* Retrieved from http://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance/#taxi-weather