Group Final Report - IDS 702

Elisa Chen, Ahmed Ibrahim, Genesis Qu, Pomelo Wu

2022-12-02

Abstract

New York City is known for heavy traffic congestion, and with the rise of ride-sharing services in the past decade, it's becoming increasingly important to understand how ride pickup services impact the traffic flow in New York City. In this report, we aim to understand whether the number of Uber pickup rides vary depending on the weather conditions, and whether Uber services increased the likelihood of traffic collision deaths in New York City. We are going to work with data on Uber pickup rides in New York City between the months January - June in 2015 as well as weather condition dataset sourced from the National Oceanic and Atmospheric Administration, which contains daily data on temperature, precipitation, snow depth, and wind strength. We also obtained data on traffic collision deaths in New York City sourced from NYPD. Based on our findings, we learned that the weather, day of the week, and Borough explain over 90% of the variation within the model. With the exception of temperature, weather conditions are less significant for determining the number of Uber pickup rides contrary to our initial belief.

Introduction

In this report, we would in particular like to research the following two questions: Q1) do weather conditions have an impact on the number of Uber pickup rides in New York City? and Q2) Did the introduction of Uber in 2015 increase the likelihood of traffic collision deaths in New York City (yes / no)? We will be building a predictive model to estimate the number of Uber rides given the weather conditions and traffic collision information on a given day to estimate how these factors influence the traffic flow in New York City. The data for Uber rides was provided by Fivethirtyeight who obtained it from NYC Taxi & Limousine Commission (TLC) (Fivethirtyeight, 2015), the weather data was obtained from the National Oceanic and Atmospheric Administration (Weather. National Oceanic and Atmospheric Administration, 2015), and the data for Traffic collisions was provided by NYPD obtained from Kaggle (NYPD, 2017).

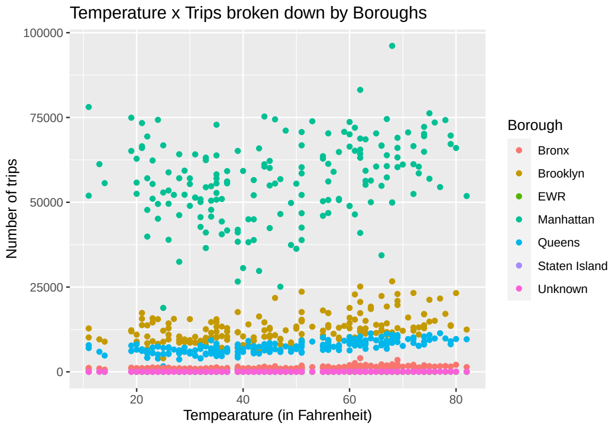
Methods

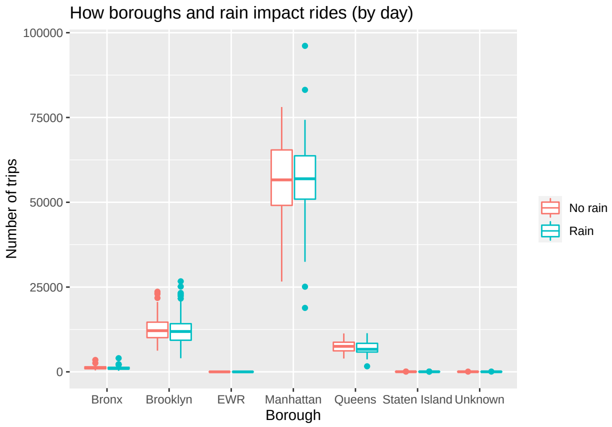
Q1

Data After aggregating the raw Uber data at the date and Borough level, we have 1160 observations with 50 variables. Please see appendix Table 1.1 for more details about the descriptive statistics of the dataset, and Table 1.2 for a data dictionary for the weather variables. We joined the Uber dataset with the weather dataset and aggregated at the day level for days in January - June 2015.

First thing we observe that regardless of the temperature of the day, Uber rides are significantly more popular in Manhattan than any other Borough. The number of pickups seem to remain fairly constant within a Borough with a slight increase in number of pickups during hotter days (65+Fahrenheit).

We can also see that the average number of pickups doesn't vary significantly between rainy and non-rainy days. This potentially signals that weather conditions might not be a good indicator for predicting number of pick-up rides on a given day. On the other hand, factors like Borough seem to be strongly correlated with the number of Uber pick-ups.





We did not observe any gaps in our weather or Uber dataset and thus didn't need to impute any missing data.

Models We chose a Multiple Linear Regression model to make predictions on how many Uber rides there would be on a given day. We chose this model because the outcome variable is continuous and we have multiple predictors to take into consideration. Additionally, the model is easy to interpret and offers robust assumptions for our data to fit against.

We conducted a priori analysis using correlation coefficients and exploratory data analysis to select a subset of predictors to use as independent variables in the regression model. Because our research question asks whether adverse weather has an impact on the day's number of Uber rides, we created a new variable named weekday, based on the recorded date, for which day of the week the ride occurred. We also recorded the precipitation variable into a categorical variable with two levels: one for a day with no rain, and the other for rainy days. Therefore, the final selection of independent variables are: day of the week, whether the day had rain, amount of snow, wind strength, average temperature, and borough.

We used the VIF function and found no multicollinearity issues among the predictors. We also did not transform the variables because their distributions were approximately normal. Please see table 1.3 for more details about VIF values.

The equation of the model is:

$$Rides = \beta_0 + \beta_1 Day + \beta_2 Wind + \beta_3 Rain + \beta_4 Snow + \beta_5 Temp + \beta_6 Borough$$

Below is the summary of our final model:

Figure 1.1

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2091.97	961.31	-2.18	0.03
weekdayMonday	-1794.20	614.33	-2.92	0.00
weekdayTuesday	-370.60	596.10	-0.62	0.53
weekdayWednesday	315.75	606.82	0.52	0.60
weekdayThursday	1692.48	609.39	2.78	0.01
weekdayFriday	2439.82	602.57	4.05	0.00
weekdaySaturday	2850.48	595.39	4.79	0.00
rainrain	185.75	396.43	0.47	0.64
SNOW	-63.84	187.20	-0.34	0.73

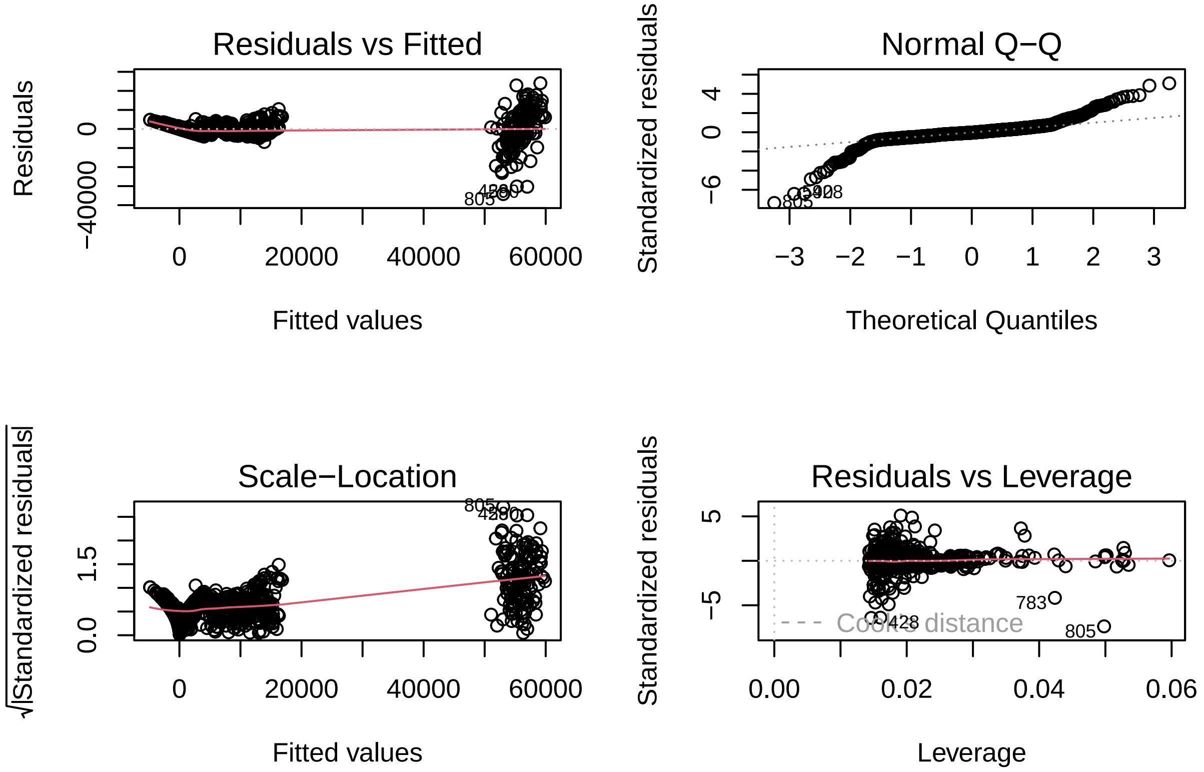
	Estimate	Std. Error	t value	Pr(>|t|)
AWND	-22.48	40.62	-0.55	0.58
BoroughBrooklyn	11555.48	582.35	19.84	0.00
BoroughEWR	-1350.58	762.21	-1.77	0.08
BoroughManhattan	54624.68	570.87	95.69	0.00
BoroughQueens	6232.60	598.07	10.42	0.00
BoroughStaten Island	-1214.27	570.04	-2.13	0.03
BoroughUnknown	-1247.54	577.56	-2.16	0.03
TAVG	59.28	9.70	6.11	0.00

Table 2: Fitting linear model: trips ~ weekday + rain + SNOW + AWND + Borough + TAVG

Observations	Residual Std. Error	R^2	Adjusted R^2
870	4757.72	0.95	0.95

Model Assessment The model residuals seem randomly distributed around 0. The residuals appear approximately normal. And there does not seem to be any alarming leverage points. We do observe separation within our dataset having values either in the lower ranges (<20k) or in the higher ranges (>50k) as illustrated by the "Residuals vs Fitted" graph. This is likely due to the stark difference in population density across the Boroughs in NYC. We'd expect Boroughs like Manhattan that is more densely populated to have significantly more pickup rides than Queens. While we're slightly concerned about the separation, we do not think it'll largely impact the overall findings of our report, and we therefore conclude that the multiple linear regression assumptions are reasonably met for our model.

Figure 1.2



The model performed well in the training set, with an R^2 value of 0.947 and a Root Mean Squared Error of 4711.01. It does not seem to have overfit and performs just as well in the testing data set, with a Root Mean Squared Error of 4619.93.

Q2

Data

Models For this question, we would like to examine contributory factors to the likelihood of people getting killed in traffic collisions. We are interested in building a predictive model to find out whether factors such as daily uber trips, variations in weather pattern or the number of daily traffic collisions contribute to the probability of people getting killed in traffic collisions. Though our outcome variable, people getting killed, is a discrete variable, this incident is most likely to either takes place or not. In other words, we will categorize at least one or more person killed as Yes or 1, and no deaths in a particular collision as No or 0. Therefore, we are treating our outcome variable as dichotomous and decided to apply logistic regression model.

Variable Selection Our response variable in the data set is people getting killed from the traffic collision (Yes/No) and is labelled in the data set as `killed_YN` which is a binary variable of Yes and No. Our explanatory variables include no. of daily uber trips per Borough, `ubertrips`, daily traffic collisions aggregated on borough level, `trafficcol`, and the five boroughs of New York City, `Borough`. To understand the influence of variation in weather patterns, we have also included Average daily wind, `AWND`, average daily temperature, `TAVG`, average daily snow, `SNOW`, and whether it rained on a given day, `rain`, which is a boolean data type. We used a priori variable selection approach to include the above variables in our model.

Our data is collected between January 1, 2015 through June 30, 2015 and we used random sampling to split the dataset into 70/30 ratio between train and test data, where we will be training our model on the train data set and analyzing it's out of sample accuracy using the test data set.

Model Assessment We created a model fitting equation that explains the probability of people getting killed in a traffic collision in terms of the above explanatory variables:

$$P[Y_i = 1|x_i] = \pi_i$$

$$logit(\pi_i) = ln(\pi_i/1 - \pi_i) = \beta_0 + \beta_1 Borough + \beta_2 ubertrips + \beta_3 trafficcol + \beta_4 AWND + \beta_5 rain + \beta_6 SNOW + \beta_7 TAVG$$

Table 3: Logistic Regression Models

	Dependent variable:
	killed_YN
BoroughBROOKLYN	−0.42 (0.90)
BoroughMANHATTAN	−0.98 (2.20)
BoroughQUEENS	−0.35 (0.70)
BoroughSTATEN ISLAND	−1.05 (0.85)
ubertrips	0.0001 (0.001)
trafficcol	0.01** (0.01)
AWND	0.01 (0.04)
rainrain	−0.11 (0.34)
SNOW	−0.78 (0.61)
TAVG	0.01 (0.01)
Constant	−3.90*** (0.86)
Observations	634
Log Likelihood	−172.91
Akaike Inf. Crit.	367.82
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4: Odds ratio and confidence intervals

	Odds ratio	2.5%	97.5%
(Intercept)	0.0201696	0.0035402	0.1037883
BoroughBROOKLYN	0.6591188	0.1094529	3.8546563
BoroughMANHATTAN	0.3753290	0.0043594	25.4352526
BoroughQUEENS	0.7019414	0.1781148	2.7568869
BoroughSTATEN ISLAND	0.3483625	0.0485932	1.5822428
ubertrips	1.0001259	0.9986433	1.0016191
trafficcol	1.0132222	1.0015277	1.0249652
AWND	1.0149823	0.9407563	1.0926858
rainrain	0.8982315	0.4522597	1.7039420
SNOW	0.4562871	0.0579017	0.9561358
TAVG	1.0108782	0.9937766	1.0288315

The tables above show the summary of the logistic regression model(table 1) for "killed_YN" as the outcome variable, and table 2 shows the `Odds ratio` and `confidence interval` of the coefficients in our model(table 2).

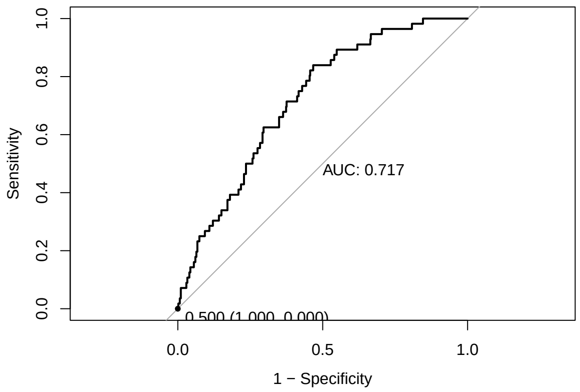


Table 5: Confusion Matrix(in sample)

	Not Killed	Killed
Pred not killed	578	56
Pred killed	0	0

Table 6: Confusion Matrix(out of sample)

	Not Killed	Killed
Pred not killed	247	24
Pred killed	0	0

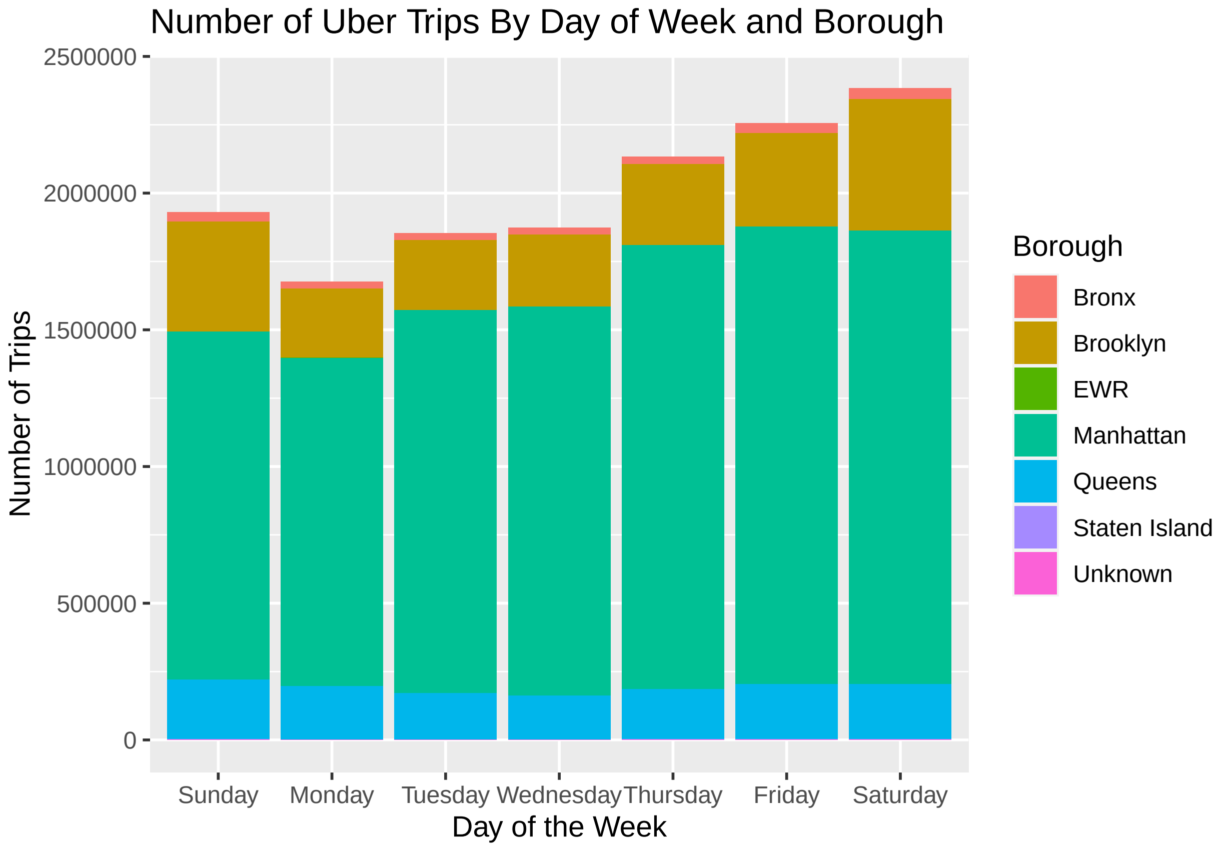
Roc curve and Confusion Matrices

The ROC plot above shows the performance of our logistic regression model on the sensitivity, specificity scale along with the Area under the curve(AUC) value. Our model has an AUC value of 0.717 which implies that our model has a greater chance of correctly predicting true samples(True positive) than incorrectly predicting (false positives). In other words, the proportion of correctly classified "killed" or "not killed" is higher than the proportion of incorrectly classified "killed"/"not killed". In addition, the confusion matrix (1) which shows the in sample predictive power of the model shows that our model accurately predicted 578 "not killed" while only predicting 56 deaths inaccurately as "not killed". This gives our an accuracy of 91.17%. The confusion matrix(2) below is created using the test data set which depics the performance of our model when predicting out of sample dataset using the 0.5 cut off probability. The out of sample accuracy is slightly lower, 91.14%, however still good enough to accurately predicting "killed" or "not killed" in the data set.

We noticed in both in sample and out of sample dataset, we have a sensitivity of 1 and specificity of 0. This is because our data set contains a very small proportion of "killed" compared to the large number of "not killed" in the traffic collision, which realistically makes a lot of sense. Because of setting the probability cut off at 0.5, our model is predicting all the outcomes as "not killed", which is a limitation in our dataset.

Results

Based on our results, we learned that the weather, day of the week, and Borough explain over 90% of the variation within the model. With the exception of temperature, weather conditions are less significant for determining the number of Uber pickup rides contrary to our initial belief. As illustrated by the EDA, the pickup Borough has the most explanatory power in terms of predicting the number of pick-ups. On average, Manhattan and Brooklyn have 54,625 and 11,555 more Uber Rides than that of Bronx respectively as illustrated in Figure 1.1 in the Models section. We also learned that on Mondays, on average we'd expect ~1795 less Uber pick-ups compared to Sunday. On Thu, Fri and Sat we'd expect there to be more Uber pick-ups compared to that of Sunday on average. Please see below graph that illustrates the number of pickups by weekday by Borough:



Supported by our model summary, the above graph illustrates that we have the most pickups during the weekends with majority of the rides taking place in Manhattan.

We noticed all the boroughs are negatively correlated with the outcome variable, with `Bronx` being the reference state. Compared to `Bronx`, for every collision that happens in `Brooklyn`, the odds of getting killed is reduced by 35%, while in `Manhattan`, `Queens`, `Staten Island`, the odds of getting killed is lowered by 63%, 30% and 66%. By contrast, we can notice that traffic collisions in `Brooklyn` and `Queens` have a higher likelihood of resulting in death compared to `Manhattan` and `Staten Island`. Variables `Average Wind`, `Daily traffic collisions` and `Average daily temperature` all have a slight influence on traffic collision deaths, such that for every increase in traffic collision, or unit increase in average daily temperature or wind, the odds of getting killed in the traffic collision increases by 1%. In contrast, when there is rain in New York city, the odds of traffic collision deaths get reduced by 10%. This makes sense because when there is rain, people tend to be less likely on the streets and the motorists tend to be drive slower reducing the risks of getting killed in the traffic collision. We also noticed that the number of daily uber trips does not have any impact on the traffic collision deaths. To be precise, for every increase in the number of daily uber trips, the odds of getting killed increases by 0.01%, which can be considered negligible.

In our model, we noticed that the variable daily traffic collision (`trafficcol`) is statistically significant with our response variable, `killed_YN`. This is partially due to the fact that we only roughly 900 observations in our dataset and the dataset is constrained within the first six months of the year 2015. This definitely possesses a limitation on the predictive power of using this model to predict traffic collision deaths. In addition, this also suggests that we need to further look into independent variables which could explain our model better.

Conclusion

In this analysis we investigated whether weather conditions impact the number of Uber pick-up rides and whether Uber pick-up rides increase the likelihood of traffic collision deaths in NYC. Factors like Pick-up Borough, and day of the week have the largest impact on the number of Uber rides in NYC. Contrary to our initial belief, with the exception of temperature, weather conditions do not significantly impact the number of Uber rides on a given day. <POMELO - ADD>

Limitations

During our model assessment, we observed separation of low and high values within the data, which could potentially be mitigated with data transformations or by simply creating two separate models for low populated areas (Bronx, Queens, Staten Island) and high populated areas (Manhattan and Brooklyn). Additionally, our dataset does not capture seasonality as we do not have data for the months of August - December. We also did not control for other factors like public transit availability and Borough population that could also influence the number of pickup rides in a given area. We observed a small degree of multicollinearity for our logistic regression model, which could be mitigated by limiting the number of variables in the model. The dataset with traffic collisions deaths was highly imbalanced due to the nature of data, which resulted in low model accuracy.

Future Work

In the future, it would be interesting to expand the research by collecting more data on other months and years, and including more information in our datasets such as public transit availability and population to account for factors that also influence the number of pickup rides. We would also want to create synthetic datasets using techniques like SMOTE to account for imbalanced datasets.

Appendix

Table 1.1

	trips	PRCP	SNOW	TAVG
	Min. : 1	Min. :0.0000	Min. :0.000	Min. :11.00
	1st Qu.: 36	1st Qu.:0.0000	1st Qu.:0.000	1st Qu.:32.00
	Median : 1516	Median :0.0000	Median :0.000	Median :46.00
	Mean :12162	Mean :0.1063	Mean :0.243	Mean :46.97
	3rd Qu.:11048	3rd Qu.:0.0300	3rd Qu.:0.000	3rd Qu.:62.00
	Max. :96118	Max. :1.6200	Max. :5.800	Max. :82.00

Table 1.2

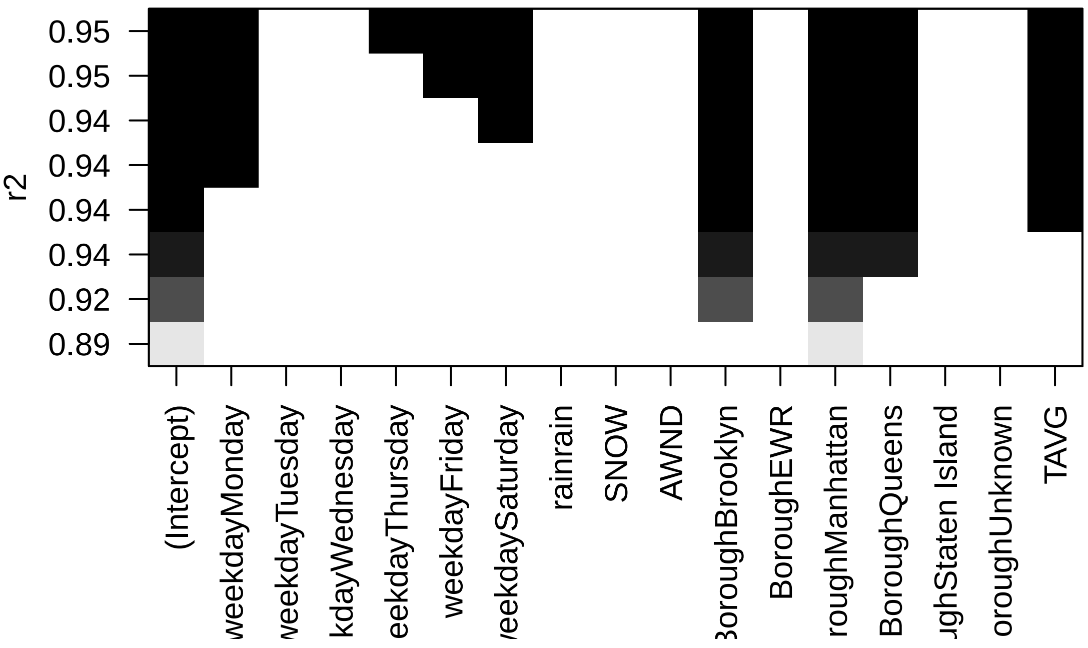
WT03	Thunder
WESF	Water equivalent of snowfall
WT04	Ice pellets
PRCP	Precipitation
WT05	Hail (may include small hail)
WT06	Glaze or rime
WT08	Smoke or haze
MDSF	Multiday snowfall total
SNWD	Snow depth
WT09	Blowing or drifting snow
DASF	Number of days included in the multiday snow fall total (MDSF)
WDF2	Direction of fastest 2-minute wind
WDF5	Direction of fastest 5-second wind
WT10	Tornado or funnel cloud
PGTM	Peak gust time
WT11	High or damaging winds
TMAX	Maximum temperature
DAPR	Number of days included in the multiday precipitation total (MDPR)
WSF2	Fastest 2-minute wind speed
WSF5	Fastest 5-second wind speed
SNOW	Snowfall
TOBS	Temperature at the time of observation
AWND	Average wind speed
WT01	Fog
WESD	Water equivalent of snow on the ground
WT02	Heavy fog
PSUN	Daily percent of possible sunshine for the period
TAVG	Average Temperature
TMIN	Minimum temperature
MDPR	Multiday precipitation total (use with DAPR and DWPR
if available)	
TSUN	Total sunshine for the period

Table 1.3

	x
weekdayMonday	1.7006
weekdayTuesday	1.7137
weekdayWednesday	1.6946
weekdayThursday	1.7560
weekdayFriday	1.6709
weekdaySaturday	1.6875
rainrain	1.3134
SNOW	1.2846
AWND	1.1783
BoroughBrooklyn	1.6880
BoroughEWR	1.3448
BoroughManhattan	1.7300
BoroughQueens	1.6347
BoroughStaten Island	1.7346
BoroughUnknown	1.7010
TAVG	1.1701

In addition to a priori variable selection, we also performed a backward selection to complement our decision-making. It appears that for the most part our Priori variable selection aligns with the backward selection.

```
## Subset selection object
## Call: regsubsets.formula(trips ~ weekday + rain + SNOW + AWND + Borough +
##     TAVG, data = train_data, method = "backward")
## 16 Variables  (and intercept)
##                       Forced in Forced out
## weekdayMonday            FALSE      FALSE
## weekdayTuesday           FALSE      FALSE
## weekdayWednesday         FALSE      FALSE
## weekdayThursday          FALSE      FALSE
## weekdayFriday            FALSE      FALSE
## weekdaySaturday          FALSE      FALSE
## rainrain                 FALSE      FALSE
## SNOW                     FALSE      FALSE
## AWND                     FALSE      FALSE
## BoroughBrooklyn          FALSE      FALSE
## BoroughEWR               FALSE      FALSE
## BoroughManhattan         FALSE      FALSE
## BoroughQueens            FALSE      FALSE
## BoroughStaten Island     FALSE      FALSE
## BoroughUnknown           FALSE      FALSE
## TAVG                     FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: backward
##          weekdayMonday weekdayTuesday weekdayWednesday weekdayThursday
## 1  ( 1 ) " "           " "            " "              " "
## 2  ( 1 ) " "           " "            " "              " "
## 3  ( 1 ) " "           " "            " "              " "
## 4  ( 1 ) " "           " "            " "              " "
## 5  ( 1 ) "*"           " "            " "              " "
## 6  ( 1 ) "*"           " "            " "              " "
## 7  ( 1 ) "*"           " "            " "              " "
## 8  ( 1 ) "*"           " "            " "              "*"
##          weekdayFriday weekdaySaturday rainrain SNOW AWND BoroughBrooklyn
## 1  ( 1 ) " "           " "             " "      " "  " "  " "
## 2  ( 1 ) " "           " "             " "      " "  " "  "*"
## 3  ( 1 ) " "           " "             " "      " "  " "  "*"
## 4  ( 1 ) " "           " "             " "      " "  " "  "*"
## 5  ( 1 ) " "           " "             " "      " "  " "  "*"
## 6  ( 1 ) " "           "*"             " "      " "  " "  "*"
## 7  ( 1 ) "*"           "*"             " "      " "  " "  "*"
## 8  ( 1 ) "*"           "*"             " "      " "  " "  "*"
##          BoroughEWR BoroughManhattan BoroughQueens BoroughStaten Island
## 1  ( 1 ) " "        "*"              " "           " "
## 2  ( 1 ) " "        "*"              " "           " "
## 3  ( 1 ) " "        "*"              "*"           " "
## 4  ( 1 ) " "        "*"              "*"           " "
## 5  ( 1 ) " "        "*"              "*"           " "
## 6  ( 1 ) " "        "*"              "*"           " "
## 7  ( 1 ) " "        "*"              "*"           " "
## 8  ( 1 ) " "        "*"              "*"           " "
##          BoroughUnknown TAVG
## 1  ( 1 ) " "            " "
## 2  ( 1 ) " "            " "
## 3  ( 1 ) " "            " "
## 4  ( 1 ) " "            "*"
## 5  ( 1 ) " "            "*"
## 6  ( 1 ) " "            "*"
## 7  ( 1 ) " "            "*"
## 8  ( 1 ) " "            "*"
```



References

Fivethirtyeight. (n.d.). Uber trip data from NYC's Taxi & Limousine Commission. GitHub. Retrieved November 29, 2022, from https://github.com/fivethirtyeight/uber-tlc-foil-response

NYPD. (2017, March 9). Vehicle collisions in NYC, 2015-Present. Kaggle. Retrieved November 29, 2022, from https://www.kaggle.com/datasets/nypd/vehicle-collisions

Weather. National Oceanic and Atmospheric Administration. (n.d.). Retrieved November 29, 2022, from https://www.noaa.gov/weather