Group 52

ORIE 3120 Project: Data Visualization and Processing

For this project, we analyzed the dataset: "Uber Pickups in New York City". This directory contains data on over 20 million Uber pickups in New York City for the year 2014 to 2015, as well as trip-level data on other for-hire vehicle companies. Upon initially analyzing the dataset, we were prompted to answer the following questions:

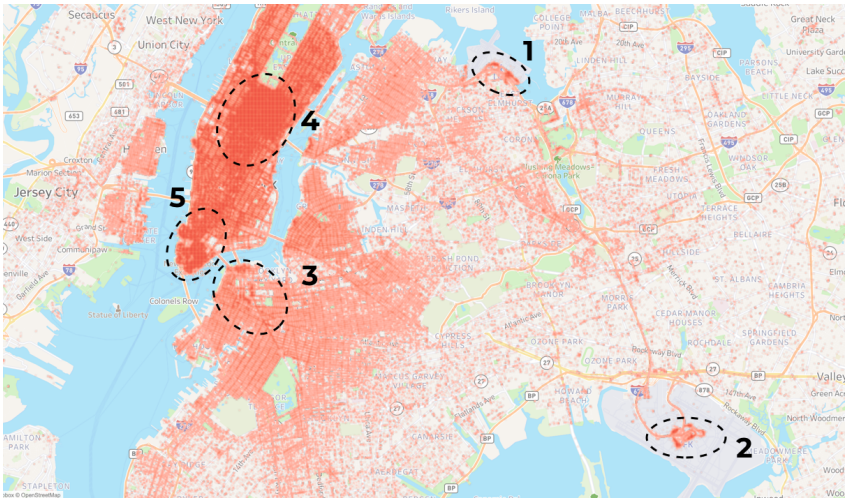
1) Where are the most popular locations for Uber rides to be called in NYC? How does this compare to the environment such as areas for mass transportation, tourist attractions, or residential areas?
2) Are there underlying time trends in Uber demand in NYC? What do these time trends reveal about Uber-hailing behavior in NYC?
3) How do changes in weather, e.g. precipitation and temperature, affect the number and frequency of Uber ride orders?

To answer the following questions and maintain consistency within our analyses, we chose to focus on the data concerning the 4.5 million Uber pickups during the April to September 2014. Each month in this period corresponds with a CSV file containing the date and time, starting latitude and longitude, and TLC base company code affiliated with each Uber pick up.

Question 1: *Where are the most popular locations for Uber rides to be called in NYC? How does this compare to the environment such as areas for mass transportation, tourist attractions, or area demographics such as income?*

Visualization, Reduced Heat Map:

All pickups from our dataset were placed as a heat map and overlaid landmarks and locations to help with future data analysis.



Analysis: We focused on the areas circled in the image below because they are relatively darker than the surrounding area. [Referencing areas 1 and 2] There are dark spots at the John F. Kennedy International Airport (JFK) and the LaGuardia Airport (LGA). [Referencing area 3] Downtown Brooklyn receives more pickups compared to the surrounding area. The streets of Manhattan receive as many or more Uber rides

than the airports and downtown areas surrounding. However, two areas in particular receive more trips. [Referencing area 4] This area is "Midtown Manhattan". This location has numerous tourist attractions including Times Square, the Empire State Building, the Chrysler Building, and 5th avenue which holds many popular shops. [Referencing area 5] This area is recognized as the "Financial District", which is home to the World Trade Center, Battery Park, access to the Statue of Liberty, as well as Wall Street and the New York Stock Exchange.

Conclusion: This heat map indicates popular locations by making them visually darker. Through analysis of these dark areas, we found that Uber rides are used most popularly around places of high commercial traffic such as shopping areas, tourist attractions, airports, and colleges, where parking may be inconvenient.

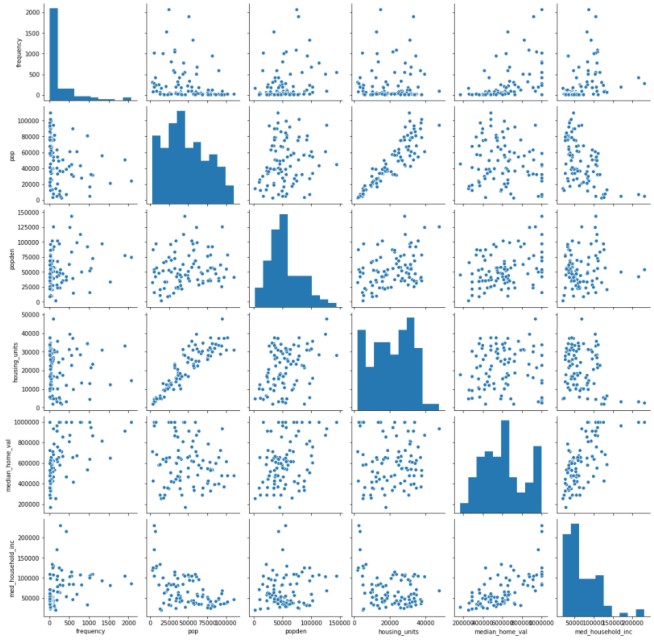
Data Analysis: Linear Regression, Model Selection, and Cross Validation:

To better understand where Uber pickups will be located, we used the uszipcode database to group the dataset by zip codes. Then, we used the demographics from that zipcode to help understand what variables most effectively predict an Uber pickup in that area.

Process: Because this dataset is very large to parse, a sample was taken from the total pickups. A zip code was appended to each datapoint using the uszipcode library.

	Date/Time	Lat	Lon	Base	zipcode
342971	2014-04-03 12:01:00	40.7214	-73.9950	B02682	10002
249861	2014-04-09 08:17:00	40.7201	-73.9552	B02617	11222
87006	2014-04-11 04:30:00	40.7626	-73.9248	B02598	11102
329478	2014-04-01 14:22:00	40.7204	-74.0054	B02682	10013
459890	2014-04-18 15:13:00	40.7631	-73.9718	B02682	10022

The data was first split into two different sets, train and test. Afterwards, the two sets were aggregated to see the total frequency of pickups per zip code.



The demographic information was appended to each dataset after aggregation, and used as the variables. A pairs plot gave insight to how some variables may affect pickup frequency as well as any other variable correlations.

Since there were several variables to use, a linear model was selected by minimizing the model AIC.

Conclusion: A zipcode's population size, number of housing units and median household income were statistically significant determinants of the frequency of Uber pick ups in that area. Additionally, the R-squared value of the model is about 0.5 for both the training and testing sets, meaning that it is a reasonable assumption that these variables affect Uber pickups and the model is well fit for the data. This relates to the prior visualization wherein we concluded that there was a high frequency of pickups where more people tended to be.

Test Set Model Summary

```
model,variables = minAIC(X_train_clean, y_train_clean)
model = sm.OLS(y_test_clean,X_test_clean[variables]).fit()
model.summary()
```

OLS Regression Results

Dep. Variable:	y	**R-squared (uncentered):**	0.519
Model:	OLS	**Adj. R-squared (uncentered):**	0.502
Method:	Least Squares	**F-statistic:**	31.59
Date:	Sun, 10 May 2020	**Prob (F-statistic):**	5.93e-14
Time:	23:04:18	**Log-Likelihood:**	-660.40
No. Observations:	91	**AIC:**	1327.
Df Residuals:	88	**BIC:**	1334.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
pop	-0.0109	0.003	-3.524	0.001	-0.017	-0.005
housing_units	0.0215	0.008	2.563	0.012	0.005	0.038
median_home_val	0.0006	0.000	4.932	0.000	0.000	0.001

Omnibus:	51.824	**Durbin-Watson:**	0.537
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	154.757
Skew:	2.021	**Prob(JB):**	2.48e-34
Kurtosis:	7.947	**Cond. No.**	159.

Train Set Model Summary

```
model = sm.OLS(y_train_clean,X_train_clean[variables]).fit()
model.summary()
```

OLS Regression Results

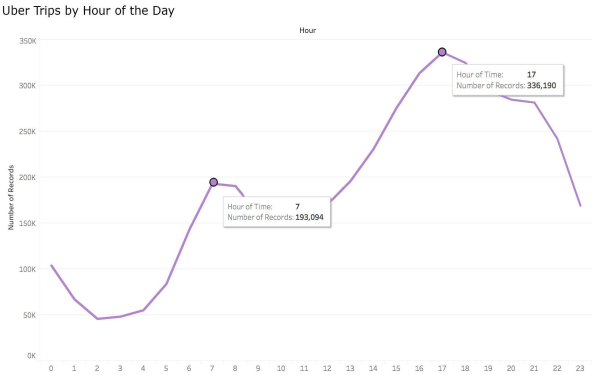
Dep. Variable:	y	**R-squared (uncentered):**	0.516
Model:	OLS	**Adj. R-squared (uncentered):**	0.499
Method:	Least Squares	**F-statistic:**	31.24
Date:	Sun, 10 May 2020	**Prob (F-statistic):**	7.63e-14
Time:	23:04:29	**Log-Likelihood:**	-660.57
No. Observations:	91	**AIC:**	1327.
Df Residuals:	88	**BIC:**	1335.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
pop	-0.0107	0.003	-3.348	0.001	-0.017	-0.004
housing_units	0.0209	0.009	2.444	0.016	0.004	0.038
median_home_val	0.0006	0.000	5.042	0.000	0.000	0.001

Omnibus:	53.525	**Durbin-Watson:**	0.385
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	166.760
Skew:	2.075	**Prob(JB):**	6.15e-37
Kurtosis:	8.173	**Cond. No.**	162.

Question 2: *Are there underlying time trends in Uber demand in NYC? What do these time trends reveal about Uber-hailing behavior in NYC?*

Visualization 1, Trips by Hour of the Day:
This line chart depicts the total number of Uber trips by hour of the day from April to September.



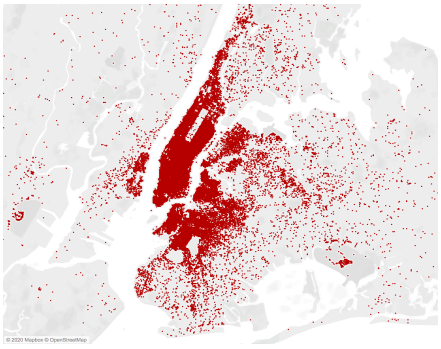
Analysis: The daily Uber-hailing behavior in NYC appears to be bimodal, with one distinct peak at 7AM and a second, significantly higher peak at 5PM. There is a sharp increase in the number of Uber trips between 4AM and 7AM.

Between 7AM and 5PM, there is a slight dip in the number of trips, followed by another sharp, constant increase from 12-5PM.

Visualization 2 and 3, Density Map Comparison of Uber Trips at Peak Hours:

These visualizations depict the number of trips (represented by size of the dots) by location, constricted to their respective time periods.

Map of 7AM Trips



Map of 5PM Trips

Analysis: It is evident from the comparison of the maps that Uber trips are more or less synonymous with the two peak hours, 7AM and 5PM. Furthermore, in areas of overlap, the density of 5PM Trips (on the right) is notably greater than the density of 7AM trips.

Conclusion: Overall, the time-trend appears to be that the number of Uber trips increases as the day progresses, with a smaller peak at 7AM and a bigger peak at 5PM. The hotspot areas for the Uber trips appear to be almost the same during both peaks, deferring only in their quantity. This consistency may be indicative of daily commutes to (at around 7AM) and from (at around 5PM) work or school.

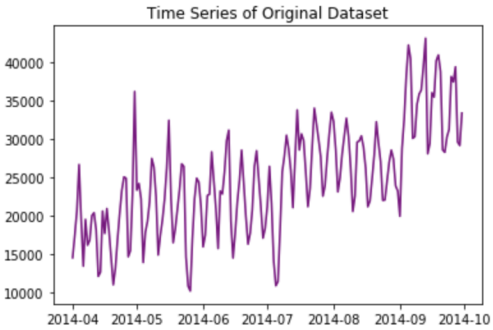
Data Analysis: Forecasting using ARIMA:

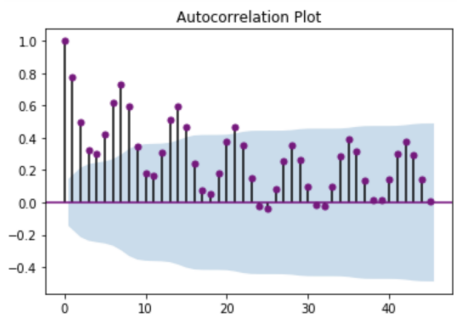
In addition to the time trends illuminated by our visualizations, we wanted to answer the additional question: "Can we use these time trends to forecast Uber demand in the future?" In this project, we decided to use the ARIMA model to forecast the demand for 2014 Uber trips in NYC.

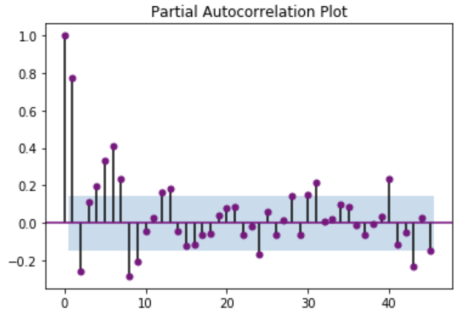
Process: The ARIMA (Auto-Regressive Integrated Moving Averages) model is a time series forecasting model that is similar to linear regression. However, its predictors depend on the following parameters:

- Number of Auto-Regressive terms (p): lags of the dependentvariable.
- Number of Moving Average terms (q): lagged forecast errors in the prediction equation
- Number of Differences (d): number of nonseasonal differences

Using the 2014 Uber Dataset, we needed to determine whether our data was stationary. We first plotted the original data, the autocorrelation plot (ACF) and partial autocorrelation (PACF) to determine parameters for the ARIMA model, and performed a significance test on the dataset to check whether the data is stationary.





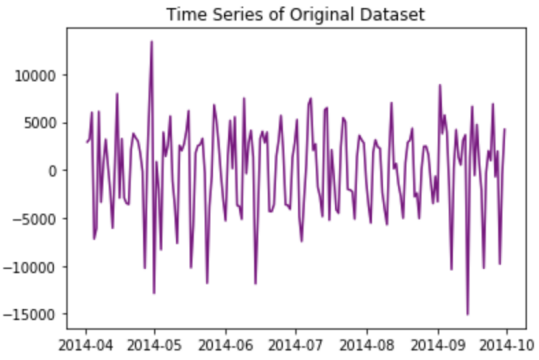


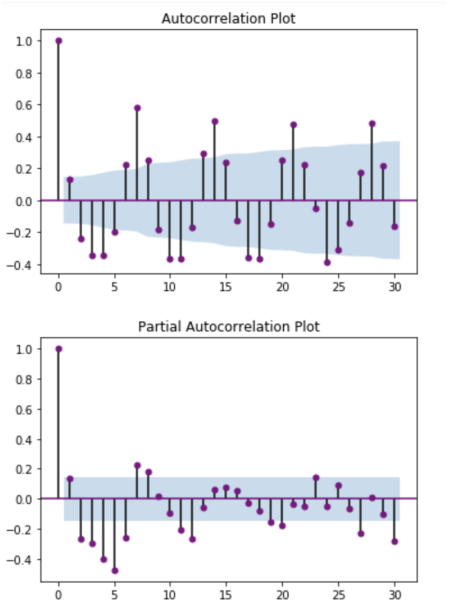
We obtained a **p-value of 0.8459** from the significance test. Since the p-value is not significant at the 95% confidence level, differencing (e.g. new value for the parameter d) is required. After setting the differencing of 1 period, e.g. d=1, another significance test shows the series produces a **p-value of 1.516e-08**, which is significant at the 95% confidence level. Therefore, we can assume that the series is stationary at the value of d=1.

Analyzing the ACF and PACF plots, we can see a spike every 7 days. We can assume that the seasonal MA component of 1 from the ACF plot, as this appears even more clearly in the ACF plot.

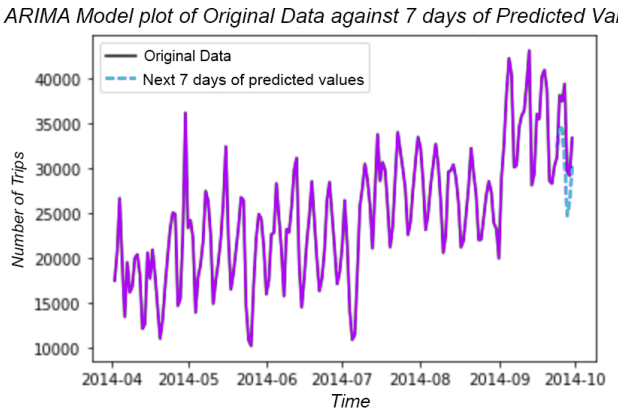
Next, we used RSME as a deciding factor to determine the best parameters for the ARIMA model. Our results showed that a relatively RMSE score of 5105.3 was for ARIMA(0,1,0)(0,1,2)[7]. These parameters resulted in the lowest AIC and BIC scores, showing us that this was the best fit for the data.

Result: Using ARIMA(0,1,0)(0,1,2)[7], we successfully forecasted the data for the next 7 days (in light blue below).





Conclusion: Analyzing our ARIMA model plot, we can see that the RMSE of 5103.3 results in an



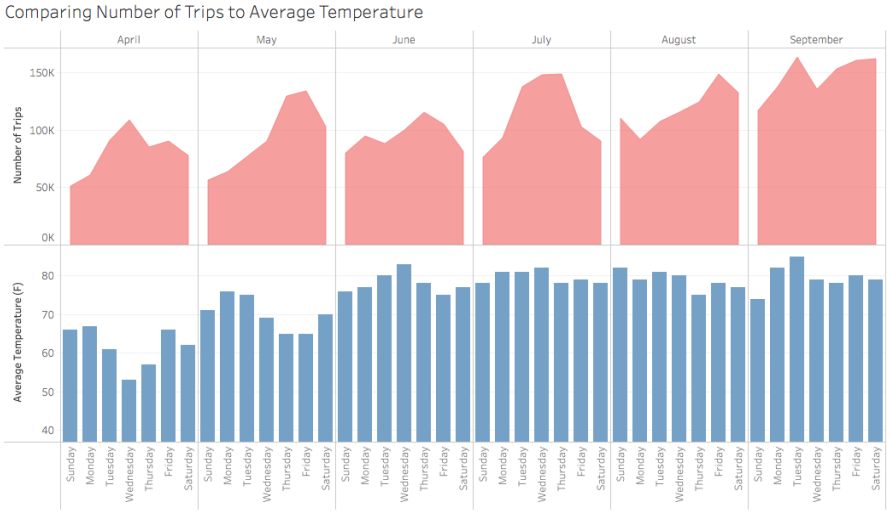
underprediction of the actual number of trips by around5000 trips for the next 7 days. However, its predictions for the trend and seasonality of the market demand is quite accurate. The predicted values dip and peak at the same dates as the original model. The forecasting may have been improved if our dataset provided a full year of data instead of afragment of a year. This may have resulted in some trends or seasonalities in the future that were not captured by the model and were thus not properly forecasted.

Question 3: *How do changes in weather, e.gn. precipitation and temperature, affect the demand for Uber trip ride orders?*

Group 52

We used an additional data set for the Weather in Central Park in 2014, from the National Centers for Environmental Information. For the months from April to September, 2014, each csv file contained information regarding the daily minimum, maximum and average temperature and the inches of rain & melted snow.

Visualization 1, Temperature: This visualization illustrates the number of Uber trips (red) and average temperature (blue) by weekday and month.



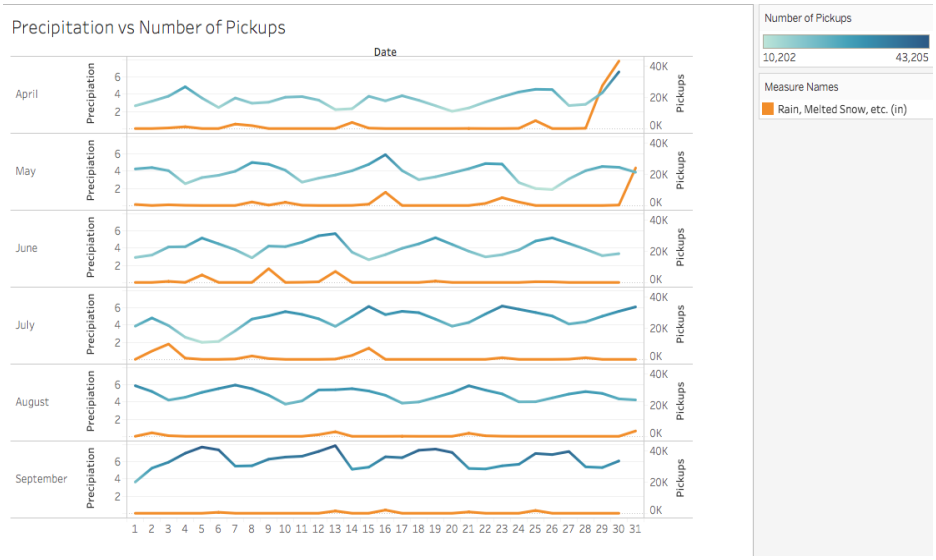
Analysis 1: Looking at the visualization, we can see that in April and May, the days with the lowest average temperature correspond to the peak in the number of Uber trips ordered. In June, July and August, there is less variation between average temperatures throughout the week, and not much variation between temperature and Uber trip orders per day. However, in general, the number of orders is greater. In September, the peak in average temperature on Tuesday corresponds to a peak in the number of orders.

Conclusion 1: During Spring (April & May), lower temperatures result in a higher number of Uber ride orders. During Summer (June, July, August), there are more ride orders, in comparison to during Spring, due to warmer weather. In September, when the temperature starts to become cooler, cooler temperature results in a higher number of orders.

Visualization 2, Precipitation: This visualization illustrates the relation between the number of pickups and average amount of precipitation (inches) per day.

Analysis 2: Many of the peak values in average precipitation correspond with peak values in pickups. However, this does not happen at all peaks.

Conclusion 2: There is a visual correlation between rainy days and Uber orders. As



precipitation increases, so do the number of pickups. People tend to use Uber when it rains.

Data Analysis: Linear Regression:

Dep. Variable:	frequency	**R-squared (uncentered):**	0.933
Model:	OLS	**Adj. R-squared (uncentered):**	0.928
Method:	Least Squares	**F-statistic:**	193.8
Date:	Mon, 11 May 2020	**Prob (F-statistic):**	3.96e-17
Time:	13:04:17	**Log-Likelihood:**	-298.45
No. Observations:	30	**AIC:**	600.9
Df Residuals:	28	**BIC:**	603.7
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
PRCP	4570.6007	1051.135	4.348	0.000	2417.448	6723.754
TMAX	279.8640	15.983	17.510	0.000	247.125	312.603

Omnibus:	0.283	**Durbin-Watson:**	0.651
Prob(Omnibus):	0.868	**Jarque-Bera (JB):**	0.346
Skew:	0.204	**Prob(JB):**	0.841
Kurtosis:	2.669	**Cond. No.**	67.9

To understand the relationship between temperature, precipitation and Uber pickup frequency, we applied a linear model to these variables.

Process: The adjusted R-squared value is 0.928 for this model, which indicates a strong relationship between temperature, precipitation and Uber pickup frequency. Both precipitation and temperature were also deemed statistically significant determinants of the frequency of Uber pickups.

Conclusion: There is a linear relationship between weather and the frequency of Uber pickups.