STAT433 Final Project Report: Uber Pickups in NYC

Caitlin Bolz, Matt Voss, Steven Xia

The group's GitHub repository is located at this link. The group's shiny application is hosted on a shinyapps.io website and can be found at this link.

Introduction: In recent years, the emergence of the ridesharing application, Uber, has drastically altered the transportation industry. Just a decade ago, the streets of Manhattan were filled with yellow taxi cabs. Today, there are nearly four times as many ridesharing vehicles on the road than taxis. As Uber has led the taxi industry into a rapid decline, the group is interested in exploring its usage patterns across New York City. Specifically, the group is interested in determining the specific factors (such as time, weather, location, etc.) that have the most influence on the number of Uber pickups in NYC. This knowledge is important because it will not only help future ridesharing customers better coordinate their trip experiences, but will also assist Uber and other ridesharing companies in developing more efficient, accurate algorithms and software tailored to optimize the amount of trips purchased from their application. Based on the data analysis and visualizations, the group reasonably concludes that the two factors that have the greatest impact on Uber usage levels in NYC during May 2014 are Borough Code and Hour .

Data: The group collected over 624,000 rows of data from two main sources. The first subset of data originates from the NYC Taxi and Limousine Commission (TLC), and consists of 4.5 million pickups from April to September 2014 as well as 14.3 million pickups from January to June 2015. However, due to limited computer processing power, the group decided to only use Uber pickup data from May 2014. For each unique Uber trip in the first data source, there exists several spatial and temporal variables, such as 'Hour', 'Day', 'Longitude', 'Latitude', and 'Burough'. The group's first data source can be found at this link.

The group's second data source consists of weather information in NYC from 2012 to 2017. This weather data was originally collected by David Beniaguev of Hebrew University in Telaviv, Israel. Each entry in this dataset represents a single hour from a given day. Because we are only interested in the Uber trips that occurred in May 2014, we also took a subset of this second data source, only looking at weather information within May 2014. For each hour within this dataset, there is an extensive amount of weather variables such as 'Temperature', 'Sky Description', 'Pressure', 'Wind Speed', and 'Humidity'. The group's second data source can be found at this link.

After the group obtained a large 2012-2017 weather dataset and cut out only the entries from 2014, this one year of NYC weather data was joined with the initial Uber pickup dataset. After inspecting our joined dataset we then found several empty entries within the 'Humidity' variable, originally from the weather dataset. Thus, we removed all NA values. Lastly, the group's joined dataset was filtered for only entries during the month of May as the group had specified this timeframe.

Methods: The group was initially curious about the amount of Uber usage during weekdays as compared to weekends. The initial analysis of our data included a bar graph that visualized the number of Uber pickups for each day within NYC during the month of May 2014. Next, to determine the strength of each predictor, a linear model was fitted. This was followed by checking for multicollinearity to confirm that each predictor was truly independent. The group then proceeded to calculate Root Mean Squared Error for Prediction (RMSEP), Mean Squared Error of Prediction (MSEP), and R^2 correlation coefficient. Lastly, the top 5 principal components of the dataset were calculated and visualized, and a final model was created for predicting the number of pickups using our spatial, temporal, and weather variables.

Results

Figure 1

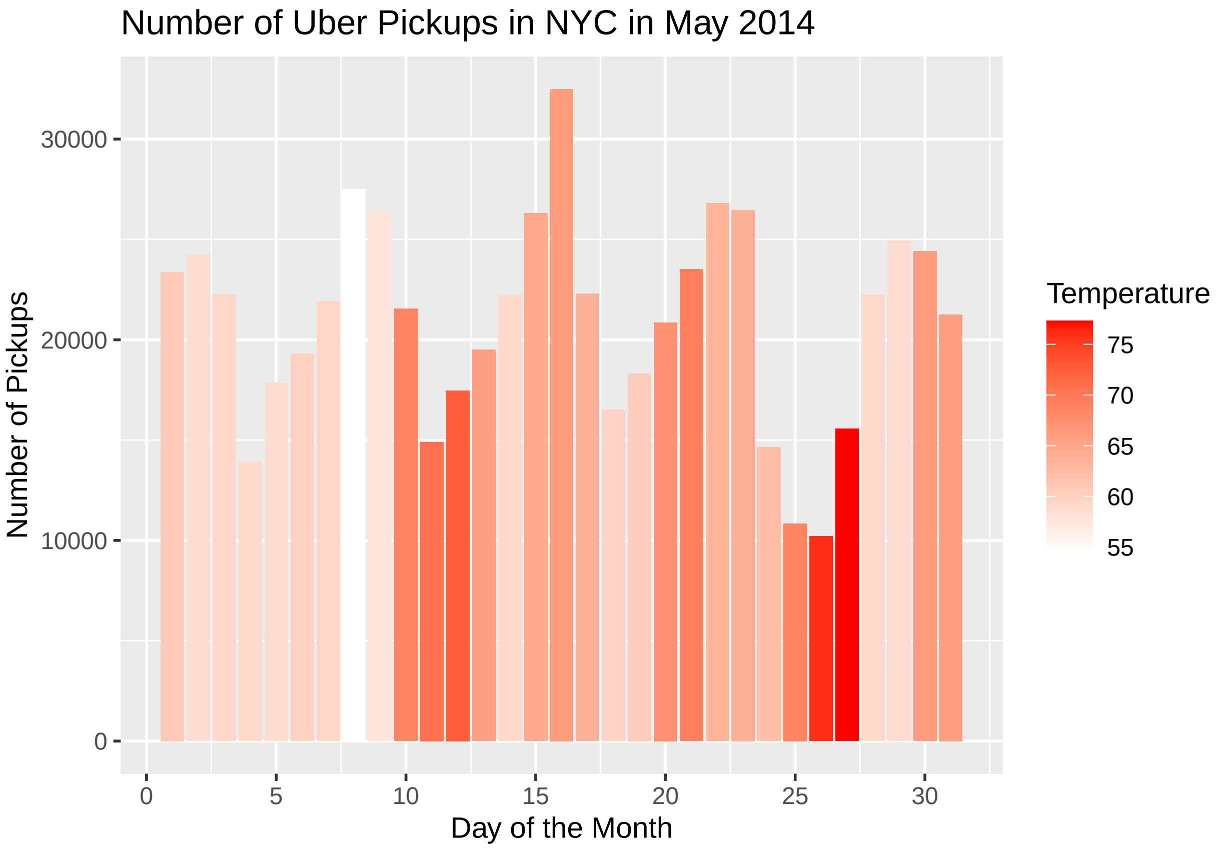
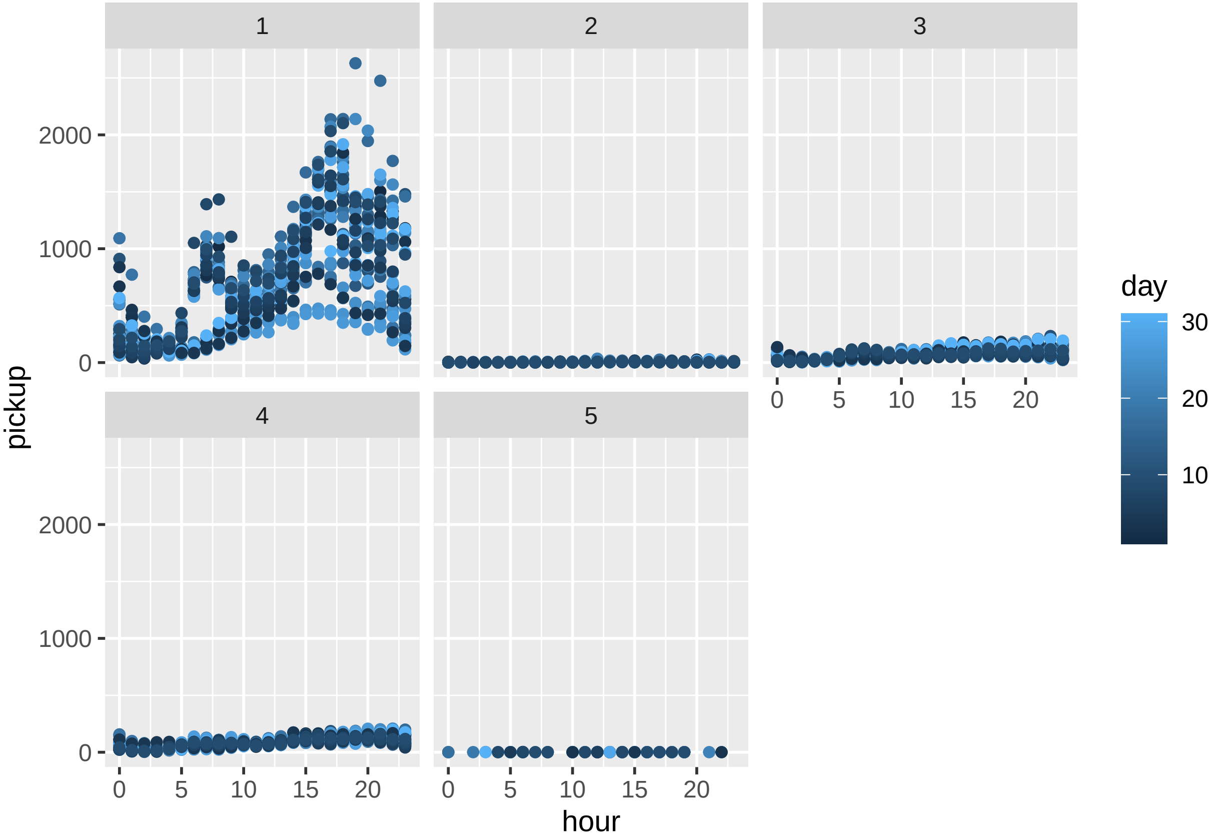


Figure 2

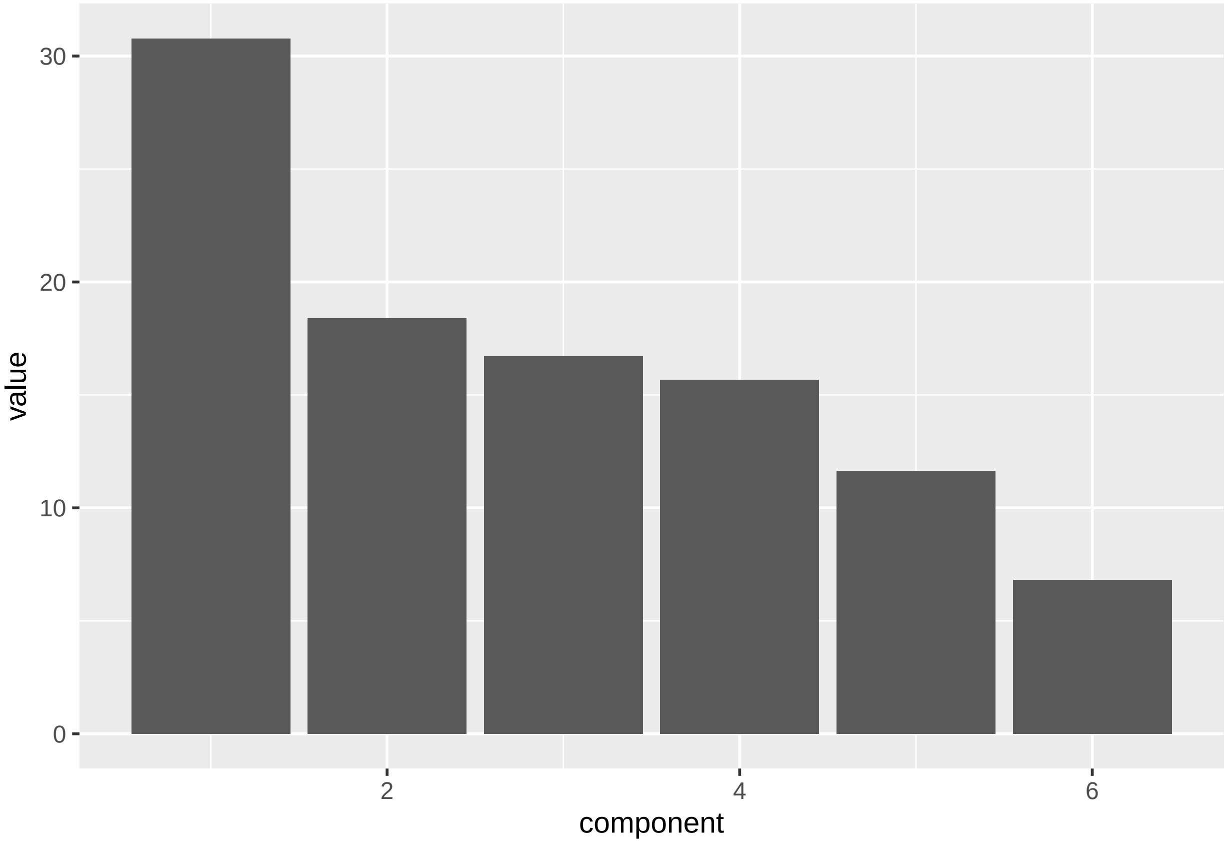


Linear Model

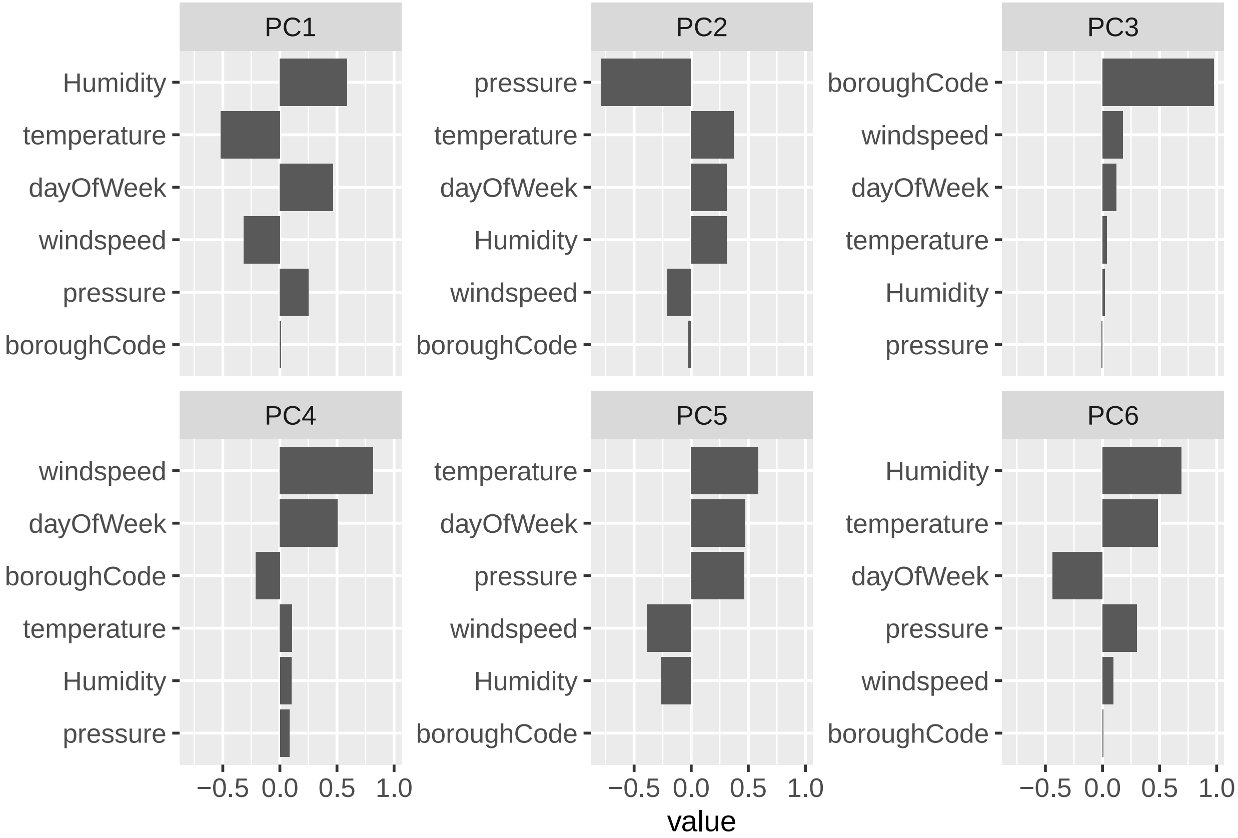
```
##
## Call:
## lm(formula = pickup ~ ., data = Ytilda)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -500.40 -203.78  -22.15  124.12 2032.16
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -32.6900  1157.2427  -0.028   0.9775
## boroughCode -163.9124     4.5978 -35.650  < 2e-16 ***
## hour          12.6379     0.8946  14.126  < 2e-16 ***
## day           -0.4854     0.6536  -0.743   0.4577
## dayOfWeek     14.6037     3.1328   4.662 3.28e-06 ***
## temperature    0.4749     0.8896   0.534   0.5935
## pressure       0.3987     1.1186   0.356   0.7216
## windspeed     -1.5750     2.9464  -0.535   0.5930
## Humidity       0.5090     0.3047   1.671   0.0949 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 299.4 on 2978 degrees of freedom
## Multiple R-squared:  0.3437, Adjusted R-squared:  0.3419
## F-statistic: 194.9 on 8 and 2978 DF,  p-value: < 2.2e-16
```

Principal Components Analysis

Plot of Variance



Plot of Top 6 Components



After mapping the plot of variance, the group determined that six Principal Components is the most prudent number of components to both reduce dimensionality and variance for the model. Next, the group constructed component graphs to understand how the original factors impacted each component. We interpret the components of our Principal Components by looking at the linear coefficients of the variables used to define them. For example, we can see that the first Principal Component positively captures humidity and day of the week and negatively captures the temperature and wind speed. From here, one can see which factor impacts the value of the principal component the most. Those near zero had little effect on the individual principal component and those near the extremes had negative or positive effects on the component based on its dependent axis location. As observed from the Principal Components Analysis, one can see that PC1 is mostly about temperature and humidity, PC2 is mostly pressure, PC3 is mostly borough code, PC4 is mostly wind speed and day of the week. PC5 is mostly temperature, and similar to PC1, PC6 is related to both humidity and temperature.

Conclusion

The group determined that Hour and Borough Code are the most influential factors in determining the number of Uber pickups for May 2014 in NYC. The group notices the high p-values in the linear regression for Hour and Borough Code and have also noticed the third principal component was mostly influenced by Borough Code. From Figure 1, the group also notices that there exist clear patterns of fluctuation in the number of Uber pickups between each day of the week. This pattern is consistent across all four weeks in May 2014. In Figure 2, there is a considerable difference in the graphs based on the borough and can see in Borough Code 1 the massive increase in pickups later in the day or early in the morning.