ISyE 4031 - Regression Project
Members: Harrison Lyle, Jake Porter, Grant Cloud, Ryan Zweber, Charles Jeffrey Howard
Team Number: B9

Problem/System Description:

Uber is a prominent ride share service that picks up thousands of riders a day. Uber is especially prominent in New York City, where life and business move fast. To reduce cancellations and pickup wait times, Uber wants to predict the total number of Uber pickups for a given day in a New York City borough. Uber can use this regression model to decide how many drivers they need to staff in each borough each day in order to meet customer demands.

In order to predict total number of pickups per day, we will use weather data (wind speed, visibility, temperature, dew point, precipitation and snow depth), day type (weekday, weekend, holiday, weekend and holiday), borough, and month. More information on the variables can be found in the variables section.

Time Dependency:

The data points were collected at different times, thus they are time dependent. More information about the collection of data can be found in the source of data section below.

Variables:

1. borough: The Borough in New York City that the # of pickups occurred in
    - Qualitative (5 classes): Bronx, Brooklyn, Manhattan, Queens, Staten Island
2. pickups: Total number of uber pickups in that Borough on a given day
    - This is our response variable not a predictor (i.e. dependent variable)
    - Quantitative: Units is number of pickups
3. spd: Mean wind speed in New York City for the day the data point was recorded
    - Quantitative: Units are miles/hour
4. vsb: Mean visibility in New York City for the day the data point was recorded
    - Quantitative: Units are miles
5. temp: Mean temperature in New York City for the day the data point was recorded
    - Quantitative: Units are Fahrenheit
6. dewp: Mean dew point for New York City for the day the data point was recorded
    - Quantitative: Units are Fahrenheit
7. slp: Mean sea level pressure in New York City for the day the data point was recorded
    - Quantitative: Units are atmospheres
8. pcp: Total liquid precipitation in New York City for the day the data point was recorded
    - Quantitative: Units are inches
9. sd: Mean snow depth in New York City for the day the data point was recorded
    - Quantitative: Units are inches

10. dayType: Used to determine whether the day is during the week, on the weekend, and whether it is a holiday or not
    - Qualitative (4 classes): Wday (Weekday), Wend (Weekend), Hday (Holiday), WendHday (Weekend Holiday)
11. month: The month that the data point was recorded in
    - Qualitative (6 classes): Jan, Feb, Marc, Apr, May, Jun

Source of Data:

The data is sourced from Kaggle, and titled *NYC Uber Pickups Enriched*. This dataset is an enriched form of the *uber-tlc-foil-response* dataset published by fivethirtyeight. The original uber data is sourced from a freedom of information request to NYC's Taxi & Limousine Commission. The enriched dataset has additional weather data, the weather data is sourced from the National Centers for Environmental Information.

The downloaded *NYC Uber Pickups Enriched* dataset was then further modified with a self written python script. The original dataset had 29101 datapoints, one point for every hour for every day for every borough in New York City. The python script further modifies the dataset by calculating means and totals for each day of data recording. The script also creates the month class column and dayType class column. The final dataset is named *uber_nyc_clean.csv*

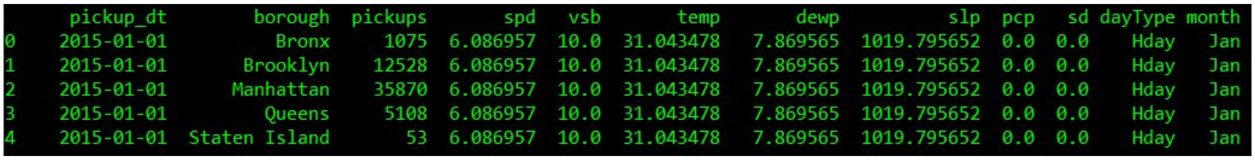
https://www.kaggle.com/yannisp/uber-pickups-enriched
https://github.com/fivethirtyeight/uber-tlc-foil-response
https://fivethirtyeight.com/
https://www.ncdc.noaa.gov/

Sample Size:

Our data originally had 29101 rows (data points) by 11 columns (variables) because the data was collected every hour of the 24 hour day for each of the 5 boroughs in New York over the course of 6 months. After cleaning the data with python we have a total of 905 data points (rows) and 11 columns (variables), in other words, we have one data point for the total number of uber pickups for each day for each of the 5 boroughs in New York over the course of 6 months.

Data Preview:

For a quick preview of the dataset, the first 5 data points are listed below.



```
     pickup_dt        borough  pickups       spd   vsb       temp      dewp          slp  pcp   sd dayType month
0   2015-01-01          Bronx     1075  6.086957  10.0  31.043478  7.869565  1019.795652  0.0  0.0    Hday   Jan
1   2015-01-01       Brooklyn    12528  6.086957  10.0  31.043478  7.869565  1019.795652  0.0  0.0    Hday   Jan
2   2015-01-01      Manhattan    35870  6.086957  10.0  31.043478  7.869565  1019.795652  0.0  0.0    Hday   Jan
3   2015-01-01         Queens     5108  6.086957  10.0  31.043478  7.869565  1019.795652  0.0  0.0    Hday   Jan
4   2015-01-01  Staten Island       53  6.086957  10.0  31.043478  7.869565  1019.795652  0.0  0.0    Hday   Jan
```