

NYC Uber Pickups with weather and holidays

-by Vishal Khatwani

# Introduction

Uber Technologies (Uber) is a mobile ride request company founded in 2009 by Travis Kalanick and Garrett Camp. Initially named “UberCab”, the company began offering black car services in San Francisco via an iPhone app. Beginning in May 2011, Uber expanded to a new city each month, moving into New York City, Chicago, and Washington D.C that year. The company also expanded internationally in December 2011, moving into Paris. After only its first six months of operation, Uber had already gained nearly 6,000 users and done close to 20,000 rides. The company expanded to its first international city, Paris, in December of 2011. As of February 2016, the company operates in over 377 cities worldwide. Uber has had a variety of economic impacts on the cities that it has moved into.

This paper will focus primarily on the city of New York how different factors like weather, geographic location and holidays affect the number of pickups for each day from each location. In this paper I tried to identify the peak times of the day when people book an Uber and in the kind of weather when people like to go out more.

# Overview of the study

This paper will focus primarily on the 6 main boroughs/county in the city of New York where most Uber rides are being booked. Throughout my analysis i focused on the various weather conditions and time in the day when most pickups take place or in simple words the effect of weather condition on the number of pickups.

The five counties I analyzed the data for were:

1. Bronx
2. Manhattan
3. Brooklyn
4. EWR(Newark)
5. Queens
6. Staten Island.

My goal was to analyze the effect of weather conditions like *Temperatures, Wind Speed, Dew Point, Snow Depth, Sea Level and visibility* on the number of pickups for a period of six months from 01/01/2015 to 30/06/2015 (by [FiveThirtyEight](https://inclass.kaggle.com/fivethirtyeight) via [kaggle.com](https://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city)).

In this study I compared weather people tend to book an Uber more on holidays or not? Does weather conditions like windy, cold, snow depth impact the number of pickups, and if it impacts weather it increases the number of pickups or decrease them? Thus I framed my hypothesis as follows:

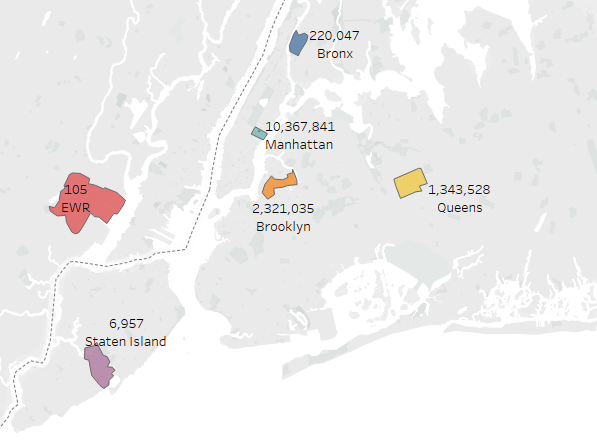
***H1: The number of pickups decrease as the weather conditions gets adverse. But at the same time the number of pickups increase on Holidays***

# Data

In order to conduct this analysis, I collected data from Kraggle.

<https://www.kaggle.com/yannisp/uber-pickups-enriched>

the data focuses 6 main boroughs/county in the city of New York where most Uber rides are being booked.



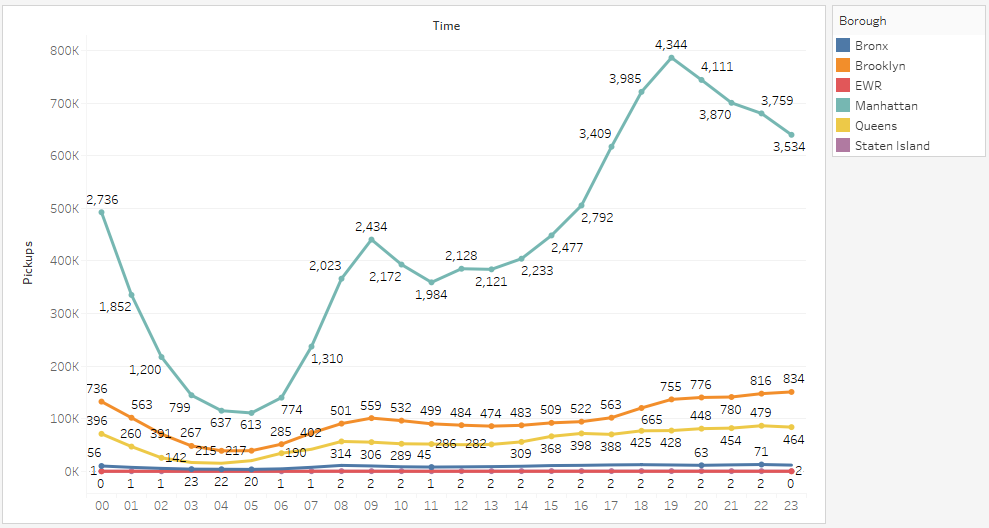
Manhattan being the busiest with the most number of pickups followed by Brooklyn then Queens, Bronx, Satten Island and EWR respectively.

The main dataset contained over 10 million observations of 4 variables which aggregated per hour and borough, and then joined with the rest of the datasets producing 29,101 observations across 10 variables. These are:

1. **pickup\_dt**: Time period of the observations.
2. **borough:** NYC's borough.
3. **pickups:** Number of pickups for the period.
4. **spd:** Wind speed in miles/hour.
5. **vsb:** Visibility in Miles to nearest tenth.
6. **temp:** temperature in Fahrenheit.
7. **dewp:** Dew point in Fahrenheit.
8. **slp:** Sea level pressure.
9. **sd:** Snow depth in inches.
10. **hday:** Being a holiday (Y) or not (N).

**Time of The day:**

Pickups from each region in the city of NYC is ought to show some relationship with the time i.e. in every region of the city the number of pickups depends on what time of the day it is. Consider the below visualization:

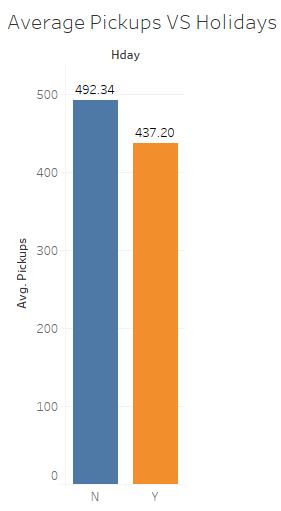


We can draw the below conclusions:

1. Manhattan has the most number of pickups for the entire days the busiest times being 09:00 AM and 06:00 PM to 11:59 PM.
2. Brooklyn is second in the number of pickups with the peak times being 09:00 AM and 06:00 PM to 11:59 PM.
3. Same is the case with Queens and Bronx
4. In EWR the average pickups are constant for the entire day except for early morning which is justified as it’s a busy airport.

**Holiday** (Being a holiday (Y) or not (N))

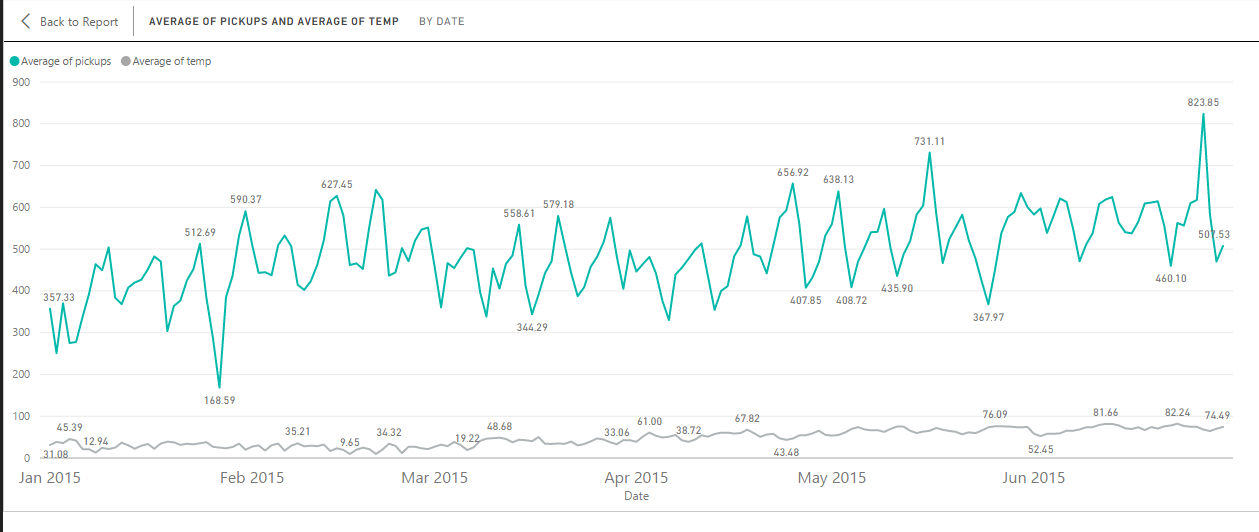
I used the variable Holiday to mark weather the observation was recorded on a holiday or not. If the observation was recorded on a holiday I marked it as “Y” and if it was not recorded on a holiday I marked it as “N”. Below were the results:



Based on the graph on the right we can easily summarize that the average pickups in MYC were more on workdays than they were on holidays, which can also mean that most of the people use Uber for their office commute.

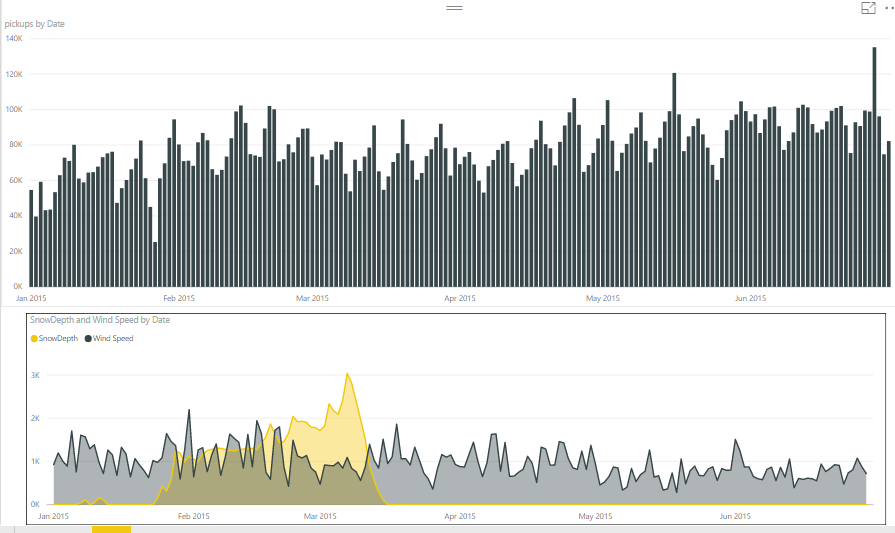
**Temperature (**temperature in Fahrenheit**)**

temperature is one of the most important variable that can help us determine whether the cause of the number of pickups. Based on the below visualization we can clearly see that the average number of pickups increase as the temperature becomes more pleasant.



**Wind Speed and Snow Depth**

Wind Speed and Snow Depth were also some of the very important variables in order to analyze the number of pickups, based on the below visualization we can draw some very important conclusions :



1. As the wind speed decreases the number of pickups increase we observe a linear decrease in the wind speed from April to June and a linear increase in the number of pickups for the same period
2. Also we can easily observe that as the snow depth is increased after a certain point the number of pickups decrease.
3. Also it is evident from the analysis that people tend to book more Uber cabs when the snow depth is 8 inches to 10 inches.

# Regression Model

Based on the hypothesis and the above noted observations i ran several regression models in order to check the effect of all the independent variables on our target variable i.e. **“Number of pickups”**. We estimated model, using linear least squares. If there was a significant effect of independent variable on our target, we expected to find the coefficient of to be positive or negative.

**Results:**

Based on the results from the regression model we found support for our alternate hypothesis to be partially true i.e. the weather and the borough/count has a significant effect on the number of pickups with their p value <0.05 but holidays does not have a significant effect on the number of pickups.

Below are the results from the regression model:

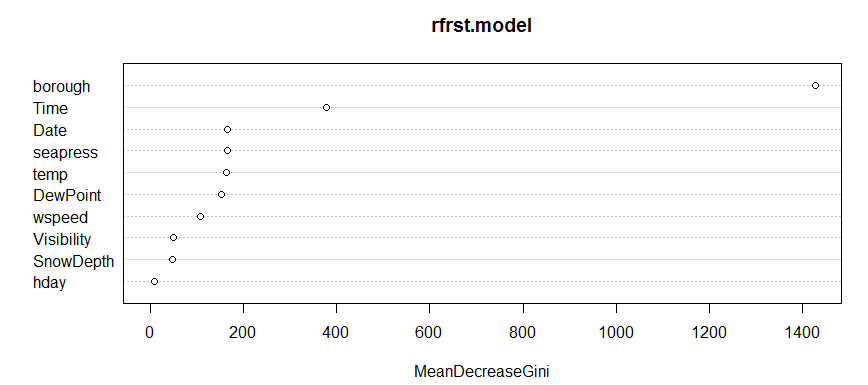
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Coefficients:** | | | | |
| **Intercept** | **Estimate** | **Std. Error** | **t value** | **Pr(>|t|)** |
| (Intercept) | 1.69E+02 | 3.79E+01 | 4.47E+00 | 7.89E-06 |
| Borough/County | 4.19E+01 | 3.74E+00 | 1.12E+01 | < 2e-16 |
| Wind Speed | 8.45E+00 | 1.83E+00 | 4.63E+00 | 3.73E-06 |
| Visibility | -1.99E+01 | 3.20E+00 | -6.23E+00 | 4.72E-10 |
| temperature | 1.14E+01 | 8.87E-01 | 1.29E+01 | < 2e-16 |
| Dew Point | -6.78E+00 | 8.32E-01 | -8.14E+00 | 4.12E-16 |
| Snow Depth | 8.29E+00 | 1.69E+00 | 4.90E+00 | 9.63E-07 |

Based on the above results we can say that other than visibility and dew point all the variables have a positive impact on the number of Pickups Borough/County, Wind Speed and Snow Depth has the most impact on the number of pickups and dew Point has a negative impact on the number of pickups. The above model is highly significant with a R-Squared value of **0.1319**.

# Random Forest Model

I ran a Random forest model over this dataset keeping my target variable as **“Number of pickups”** where the number of pickups weredivided into categories of “Low”, “Average”, “Above Average” and “High respectively”.I used 30% of my data to train the model and the remaining 70% to test my model. Below is the variable importance based on the model I built using 30% percent of the data.

|  |  |
| --- | --- |
| **Variable** | **Importance** |
| Borough/County | 1426.843142 |
| Time | 378.367727 |
| Date | 165.558683 |
| Sea Pressure | 164.437823 |
| temperature | 162.509676 |
| Dew Point | 153.000715 |
| Wind Speed | 107.283965 |
| Visibility | 49.805876 |
| Snow Depth | 47.929859 |
| Holiday | 7.858757 |



I tested my model with the rest of 70% data and I found an accuracy of **91%** that means that the random forest model I build correctly classifies 91% of the data correctly thus we can rely on the results I received from the model.

# Conclusion

This entire paper was focused to identify the factors that affect the number of Uber pickups in the New York City. Based on the analysis we can clearly state that the number of Uber pickups are affected by factors like Time of the day, Areas where the pickup took place and weather conditions like temperature, Wind Speed, Snow Depth, Visibility and Dew Point. Also some of the environmental conditions like Visibility and Dew Point have a negative impact on the number of pickups.

# References

* Kaggle : <https://www.kaggle.com/yannisp/uber-pickups-enriched>
* Tableau Public: <https://public.tableau.com/profile/vishal.khatwani8243#!/>
* Google: [www.google.com](http://www.google.com)
* YouTube: [www.youtube.com](http://www.youtube.com)
* Uber: https://www.uber.com/cities/new-york/