Caleb, Fidel, Ryan

Northwestern Data Science Bootcamp

June 20, 2023

**UBER PICKUPS DATA ANALYSIS**

1. **Intro**

Our project takes a look at Uber pickup data in New York City from January 1, 2015 to June 30, 2015. We decided to use a data set that we found on Kaggle and see how different factors impact pickups. The questions we decided to ask are the following:

* How does precipitation impact pickups?
* Does snow have a different effect than rain?
* How does wind speed affect pickups?
* Does relative temperature affect pickups?
* How does each borough affect pickups?
* What impact does the day of the week have on pickups? Weekend vs Weekday?
* How does an average day compare to an average holiday?

1. **Analysis**

Our data set gave us three different types of rain precipitation and data on snowy days. The three types of rain data that it gave were days with 1 hour of rain, days with 6 hours of rain, and days with 24 hours of rain. We took the total number of pickups per each day with that type of precipitation and then averaged them out. What we found was increased periods of rain led to a greater number of average pickups per day. Days with no precipitation still had a greater average number of pickups compared to days with rain. What was surprising was how big of a difference snow played into the average number of pickups. Snowy weather had the greatest number of average pickups per day. Something that would help give our data even greater accuracy would be if we had individual driver and rider data. We would be able to see if more uber rides are requested for shorter periods of time during inclement weather or if people are using uber to and from locations during inclement weather versus days of good weather.

The data provided pickup timestamps for everyday by the hour for 6 months. We decided to parse the data to view it on a day to day level to see if the day affected the number of pickups. We decided to group them by day and sum their total. We decided to use an Anova test to view if there was any correlation between the day and pickups. We found that there is a correlation amongst the day of the week and pickups. We found that Saturday has the highest amount of pickups. We wanted to look as well to see if there is a difference between Weekday vs Weekend. We performed another Anova test to find if there’s any sort of correlation as well. We found that there was a strong correlation between weekend and weekday and the amount of pickups. When looking at the averages we found that the weekends had an overall higher average but when looking at the overall totals there was a higher amount of trips taken on the weekday. Another aspect we were focusing on was if holiday’s affect the amount of trips taken. We found that the average were similar but looking at the overall trips there was more data available for non-holidays.

Another part of the dataset that we wanted to analyze was the impact of wind speed and the number of pickups per day. To do this we consolidated the dataset to show all pickups, not broken down by borough, and combine all pickups per day and that day’s average wind speed. We then plotted this on a scatter plot to show the relation. The plot had a r-value of -0.22577964333481676 which shows a weak correlation. We can now draw the conclusion that wind speed does not have a significant effect on the number of pickups per day.

We also wanted to look at the effects of temperature on the number of pickups but wanted to look at the real feel temperatures, but our dataset didn’t provide enough information to do this. We found an API that had historical data for weather in New York City for the same time period, 01/01/2015 to 06/30/2015, and used that weather data to merge with our current dataset. This scatter plot showed an r-value of 0.3642127572146948 which is also a weak correlation. We can conclude that the real feel temperature does not have a significant impact on the number of pickups.

Lastly, we wanted to compare the number of pickups within each borough to see which ones were most popular. Manhattan was by far the most common for pickups with Brooklyn and Queens close for second and third most. If Uber were to have any surge pricing for locations, these would be the best boroughs to target.