Final Project: Uber Usage
MIDS-W200 Section 001
Project 2: Data Analysis with Numpy and Pandas
The A-Team: Adam Yang, Armand Kok, Alla Hale

1. Background:

In recent years, ride sharing services, lead by Uber, have been rapidly growing in popularity. A big part of the rising popularity of these services is the convenience that allows users to travel on their own timelines, without the hassles of using their own vehicles. They also have an advantage over taxi services because they allow users to plan their trips more accurately by providing estimated arrival times and cost before commitment.

Data describing the use of these services can shed light on how people get around in a city. Analysis of when, where, and how people use these services can be used by rideshare providers to help optimize their operation, or by city administrators (e.g. city planners, transport authorities) to help minimize traffic or even improve public transportation systems.
Uber usage is quite high in New York City. This analysis will focus on Uber data from NYC in 2014. This data will be used to understand the effect of weather and other factors on Uber usage. Additionally, the variation in Uber usage will be analyzed by day, month, and time of day. The effects of special case variations, such as severe weather events or key holidays will also be identified.

2. Data Sources:

We have several data sets to combine for this project. The first is a data set that describes approximately 4.5 million Uber trips in NYC from April to September of 2014. This data set includes dates and times along with the location (latitude and longitude) of each Uber pickup.

The second data set describes the weather. This data set is much broader than needed for the Uber analysis, including weather for 36 cities from 2012 to 2017. The NYC weather data will be used for the same time period as the Uber data are available.

The third data set provides locations (latitude and longitude, and address) of public transit stops (subways and buses), that will be used to help answer questions around public transit as it relates to ride sharing.

Links to the data sets may be found in the appendix.

3. Data Preparation:

3.1 Uber Trip Data:

Before beginning analysis on the data, the data sets had to be cleaned and joined. The raw Uber pickup data in New York City were stored in separate CSV files for each month, containing the date and time, latitude, longitude, and base of each pickup. The first step would be to stitch them together into a single dataframe.

The raw data set provided by the source was limited to only the Uber pickups between April and the end of September. Individual trip information was needed for the exploration of time of day, hourly weather, and other factors influencing ride sharing usage. Additionally, ride sharing services other than Uber were excluded from this analysis because in 2014, Uber owned the majority of the market for ride sharing services (>90%).

3.2 Weather Data:

The weather data set was obtained from a separate source and needed to be merged with the Uber dataframe. The weather variables were contained in separate files, including information about multiple cities and a larger date range than required. Only the data for New York city, from April to September 2014 were required. Table 1 shows the different variables that were used as part of the analysis.

Additionally, the temperature data were converted from Kelvin to Celsius. Since the weather information was reported hourly, and the Uber rides were reported as they occurred, they had to be merged on the hour of the Uber Date/Time.

3.3 Transit Location Data:

The transit location data included the names of the transit stops, along with latitude and longitude of each stop. Most of these stops are subway stations or elevated train stations. No transit stops were excluded for this analysis.

3.4 Merged Data Set:

The primary data set used for analysis was the merged Uber and weather data along with a few other variables used to help with the analysis. These variables are shown in Table 1.

To analyze the data by the day of the week, a ‘dow’ column was added.

The final data set was approximately 6 months long, so an n_days column was also added to represent the number of times each “dow” occurred. This column was used to normalize the “trip counts” to “trip per day counts” for more useful analysis.

Table 1. *Variables used for data analysis*

Variable	**Description**
date_time_h	Date and time rounded to the nearest hour
date_time	Date and time of the Uber pickup, rounded to the nearest minute.
dow	Name of the day for the pickup (e.g. Monday, Friday, Sunday)
lat	The latitude of the Uber pickup
lon	The longitude of the Uber pickup
weather_description	Description of weather (e.g. sky is clear, broken clouds, drizzle)
wind_speed	Wind speed in m/s
humidity	% humidity
temperature	Temperature in celsius
holiday	Name of the holiday (NaN for regular days)

4. Sanity Check:

As part of the sanity check, most of the variables in the data set were evaluated.

4.1 Weather variables:

All weather variables (weather_descriptions, wind_speed, humidity, temperature) were complete, except for humidity. There were approximately 20,000 rows with no humidity values in the merged data set. Since the data set was greater than 4 million rows total, and humidity did not play a predominant role in the the analysis, the missing values were simply ignored.

4.2 Trip count:

As part of the sanity check, the monthly trip count in the data set was compared against another published study. The monthly trip count of the data set is very close to the one found in the study, therefore concluding that the trip count by month is reasonably accurate.

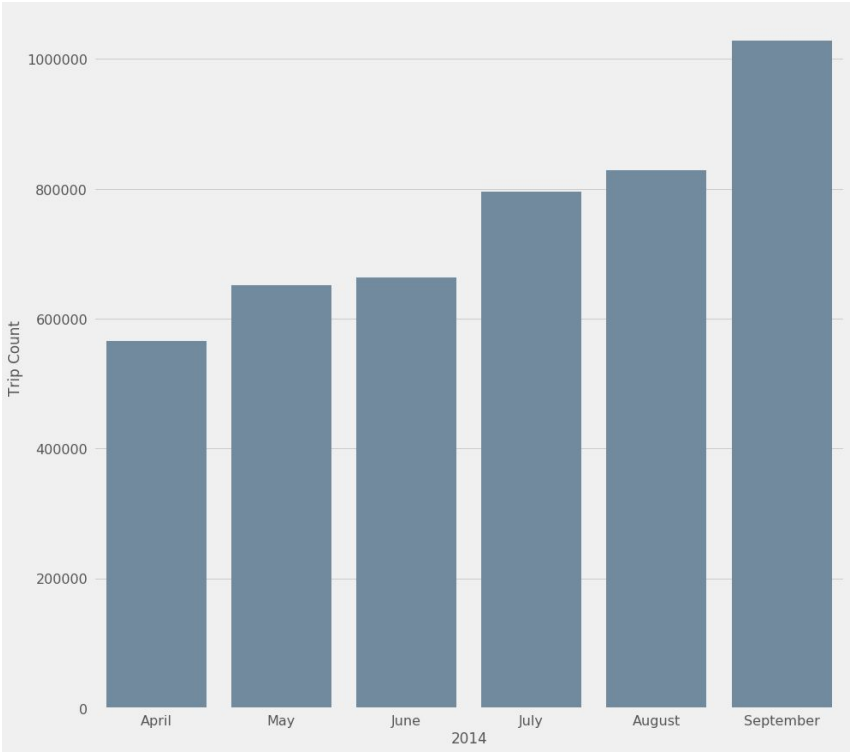


Figure 1. *Trip counts per month from April 2014 - September 2014*

4.3 “Day of Week” and Holidays data:

The Day of Week (dow) column was created from the Uber data set. Some quick sanity checks were done by looking up some random dates on a calendar app to see if the “dow” match. The same thing was done with the Holidays data.

4.4 Data Limitation:

Due to the nature of the data set, please note that there may be limitations in the analysis and findings in this report.

5. Data Analysis

5.1 What is the overall Uber usage?

The overall Uber usage from April to September of 2014 in NYC was examined by day (Figure 2). The number of rides, or pickups, per day shows an increasing trend over this time period, with strong weekly periodicity. From the beginning of April to the end of August, there seems to be a gradual increase in the number of Uber pickups in NYC. However, in September, the number of Uber pickups jumped significantly. Maybe this is in conjunction with school starting on September 4th in NYC.

On average, the number of pickups per day per month is shown in Table 2.

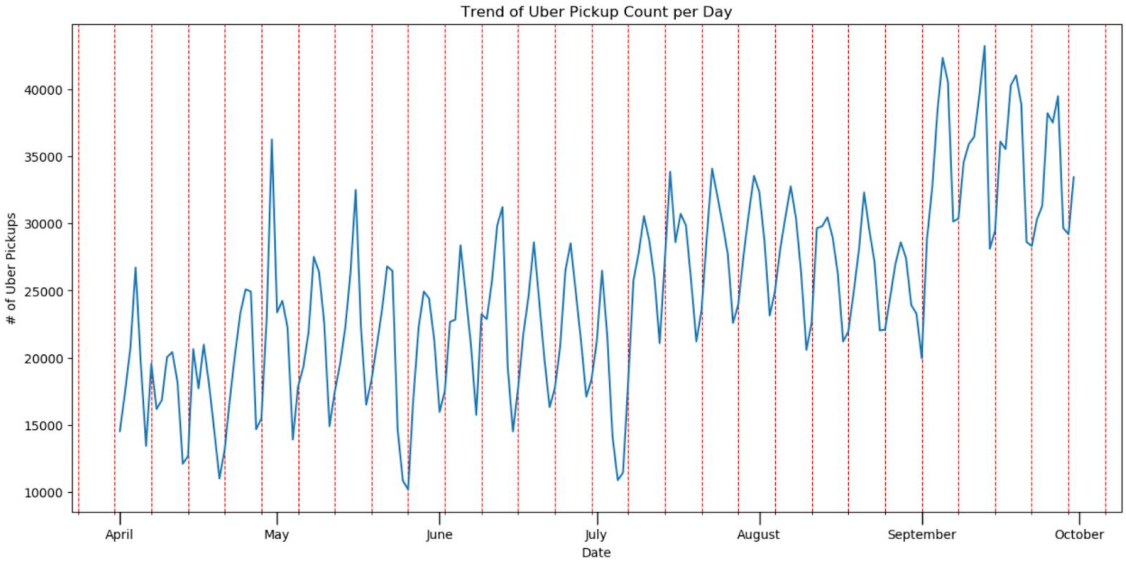


Figure 2. *Trend of Uber Pickup Count per Day. Vertical red lines represent Mondays. Uber usage appears to be increasing over time, with strong weekly periodicity.*

Table 2. *The Average Count of Pickups per Day by Month*

Month	Trips Per Day
April	18000
May	21000
June	22000
July	26000
August	27000
September	34000

5.2 How does Uber usage vary by day and time of the day?

From April to September of 2014 in NYC, the weekly periodicity was worth investigating further. The variation in Uber usage was analyzed by the day of the week (Figure 3). On the whole, Thursday was the most popular day of the week for Uber rides with an average of approximately 29,000 rides per day, and Sunday was the least popular with an average of approximately 19,000 rides per day.

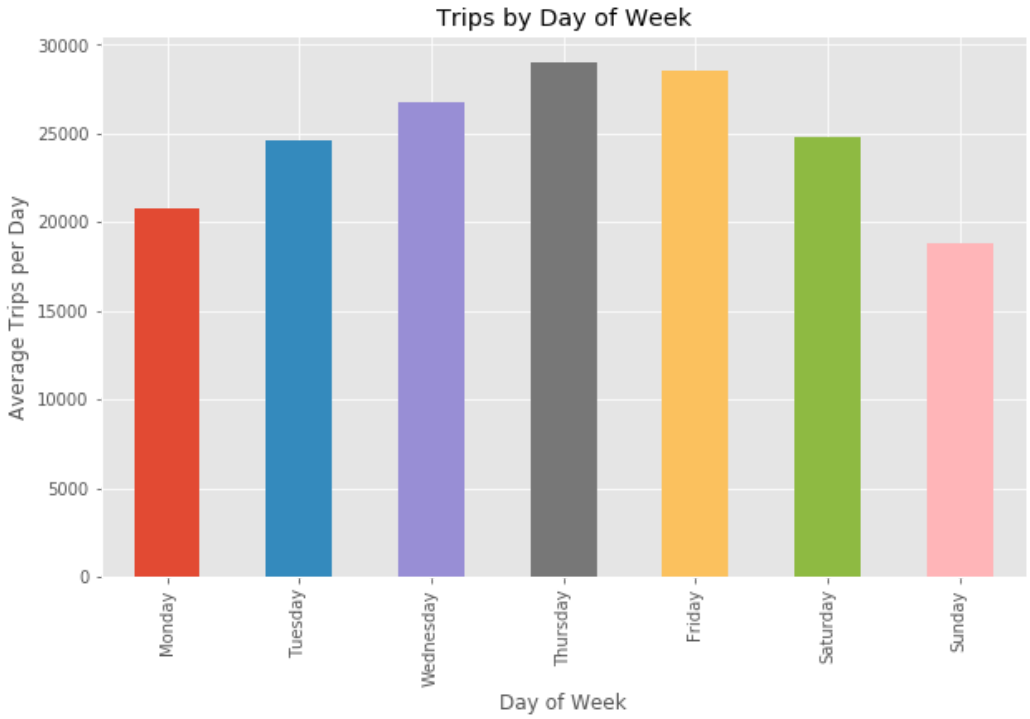


Figure 3. *Average Trips per Day by Day of the Week. Although overall usage varies over time, the weekly periodicity shows peak Uber usage on Thursdays and minimum usage on Sundays.*

The rides by time of the day were examined to fully characterize Uber usage during the day (Figure 4). There appear clear morning and afternoon rush hours at 7-8AM and 5PM. This is not surprising, since many people may use Uber to get to and from work every day.

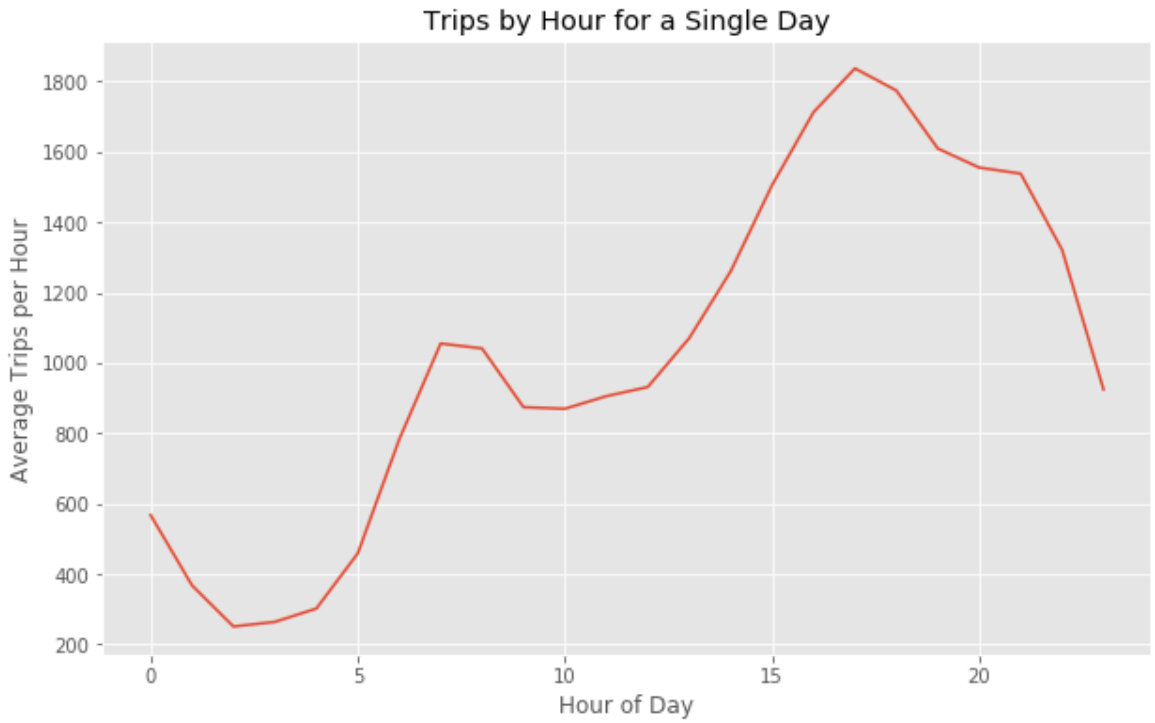


Figure 4. *Trips per Hour for a Single, Average Day. Uber usage shows clear morning and afternoon rush hours.*

However, when examining the rides by time of the day by day of the week, the differences between the days clearly indicate that Uber is used both for work and play (Figure 5). On Monday through Thursday, the morning and afternoon rush hour pattern is repeated. On Fridays, however, trip counts increase again in the late evening after 8PM and last into the early morning. This is likely because people are using Uber to go out. On Saturdays and Sundays, the morning starts much later, and does not have a peak, as people are likely not hurrying to work. Saturday evenings look much like Friday evenings, with people going out. Sunday evenings, however looks more like a weekday evenings, likely because New Yorkers are getting ready for the work week. Nothing here is shocking.

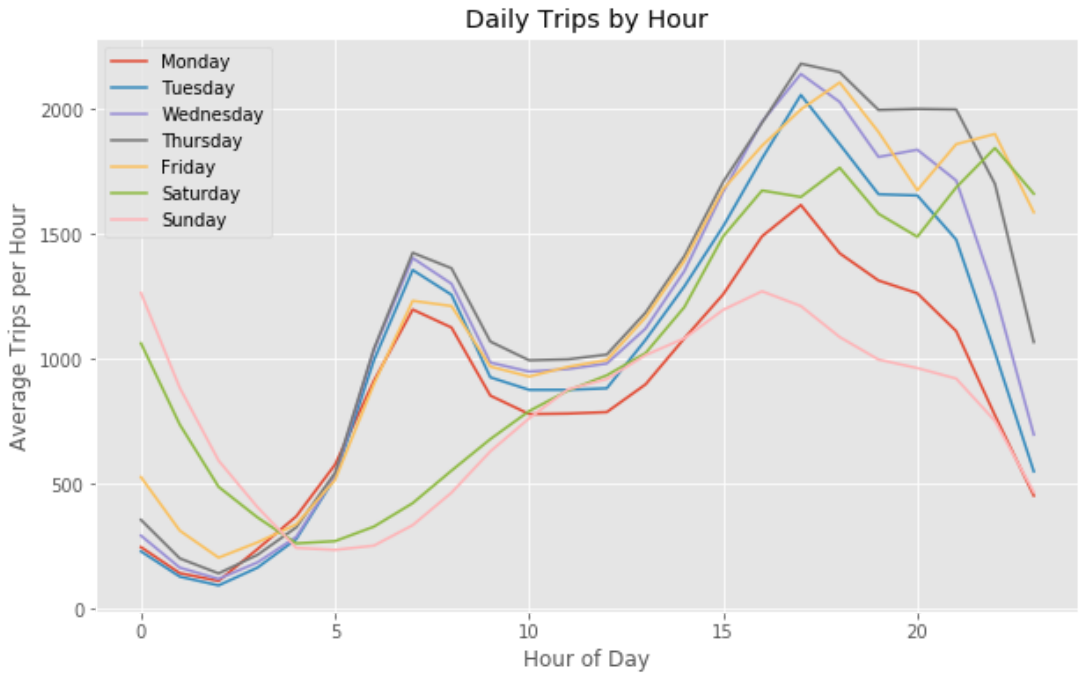


Figure 5. *Trips per Hour by Day. Uber is clearly used for both work and play.*

To simplify the story, weekends and weekdays can be separated as in Figure 6 to highlight the differences. Weekdays are for working. Weekends are for having a good time. This, however, loses the nuance shown in Figure 5, above.

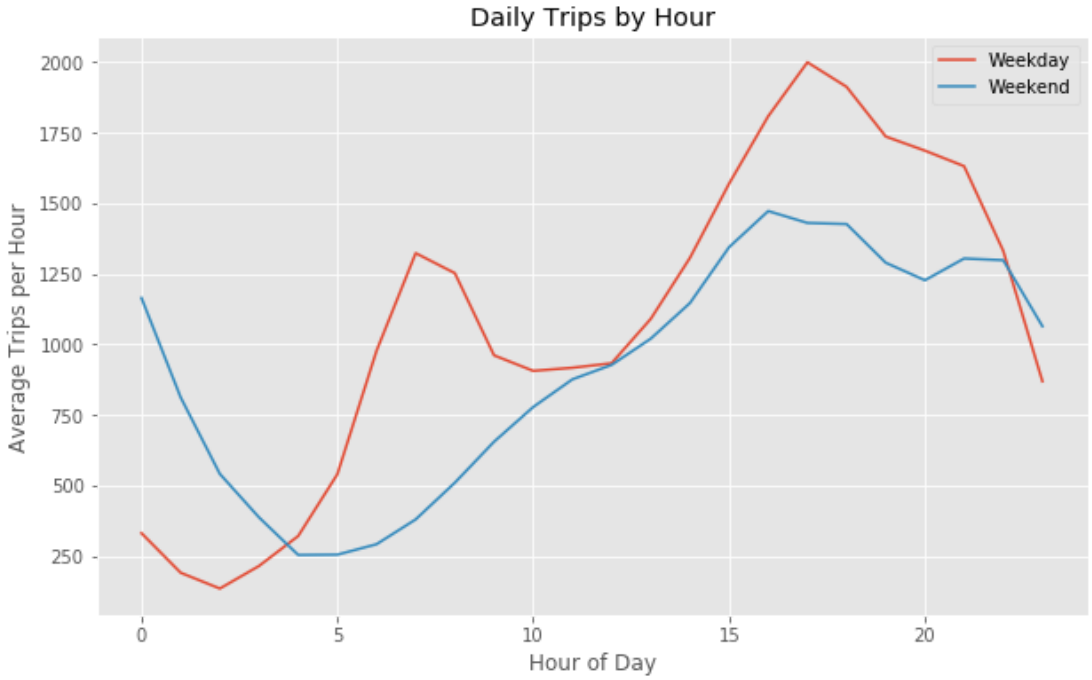
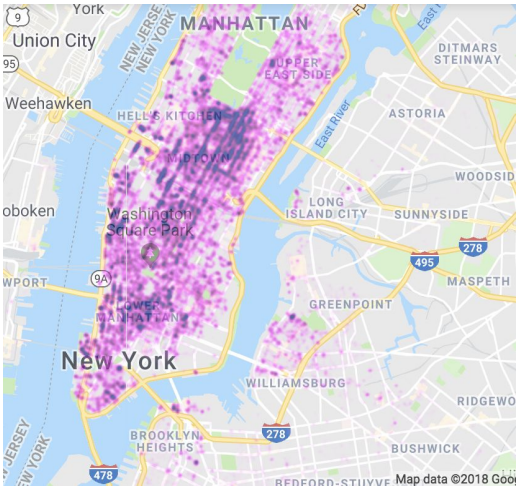


Figure 6. *Trips per Hour by Type of Day. Saturday and Sunday are weekends, all other days are weekdays.*

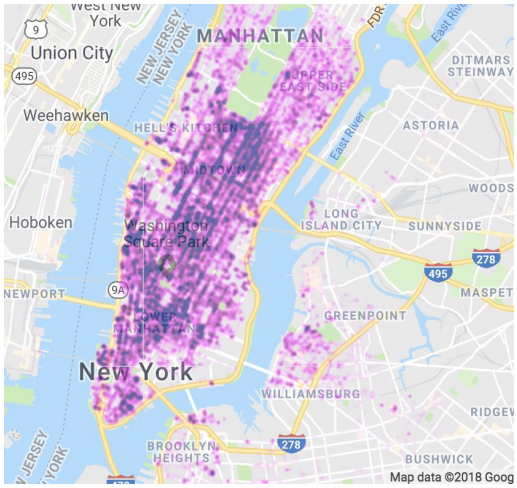
In addition to understanding *when* people are taking Uber rides, it is interesting to examine *where* people are taking Uber rides. This analysis is a bit limited, since we only have pickup locations. However, heatmaps for ride origins are shown for Mondays (Figure 7a), Fridays (Figure 7b), and Saturdays (Figure 7c) in September, below. Only days in September were used for this analysis since there was such a marked increase in ridership that month.

On Mondays, most traffic appears to be in Midtown, where many people work. Monday pickups are relatively light in Downtown, which is more of a party location. On Saturdays, however, the traffic is mostly in Downtown. Additionally, Saturdays show lots of pickups in Brooklyn Heights and Williamsburg, which are hip locations for people to unwind after long weeks. Saturday's heatmap also shows more pickups across the bridges in areas like Hoboken. This may be due to the fact that trains do not run frequently late in the evening, and people do not want to wait. Fridays are interesting, because the pickup heatmaps show elements of both weekday work, and weekend play.*

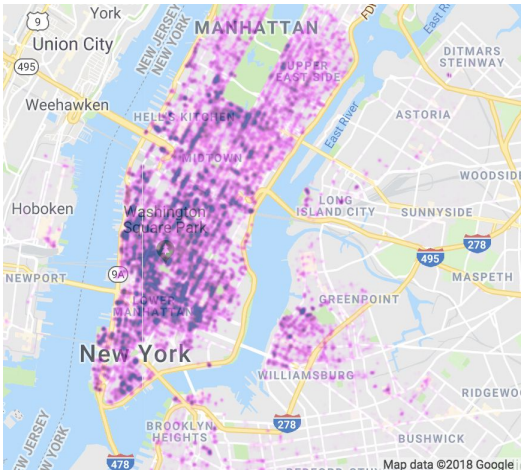
*NYC work and play distinctions courtesy of longtime NYC resident and expert, Danielle Adler.



(a) Mondays



(b) Fridays



(c) Saturdays

Figure 7. *Heatmap of Uber Rides on (a) Mondays, (b) Fridays, and (c) Saturdays In September. Dark red areas show more frequent pickups. (a) Most people are working on Mondays. (b) There is an even distribution between work and play on Fridays. (c) Many people are playing on Saturday. Work occurs mainly in Midtown and play in Downtown.*

Clearly, the day of the week affects Uber usage. It appears that people are using Uber as just another method of transportation in NYC.

5.3 How do key holidays impact Uber usage?

The variation in Uber usage on key holidays was noted by examining the overall time series of Uber rides per day (Figure 8). On holidays where people usually get time off from work, it is interesting to note that the number of rides seems to drop. It seems as if people in NYC use Uber for work more than during vacation times.

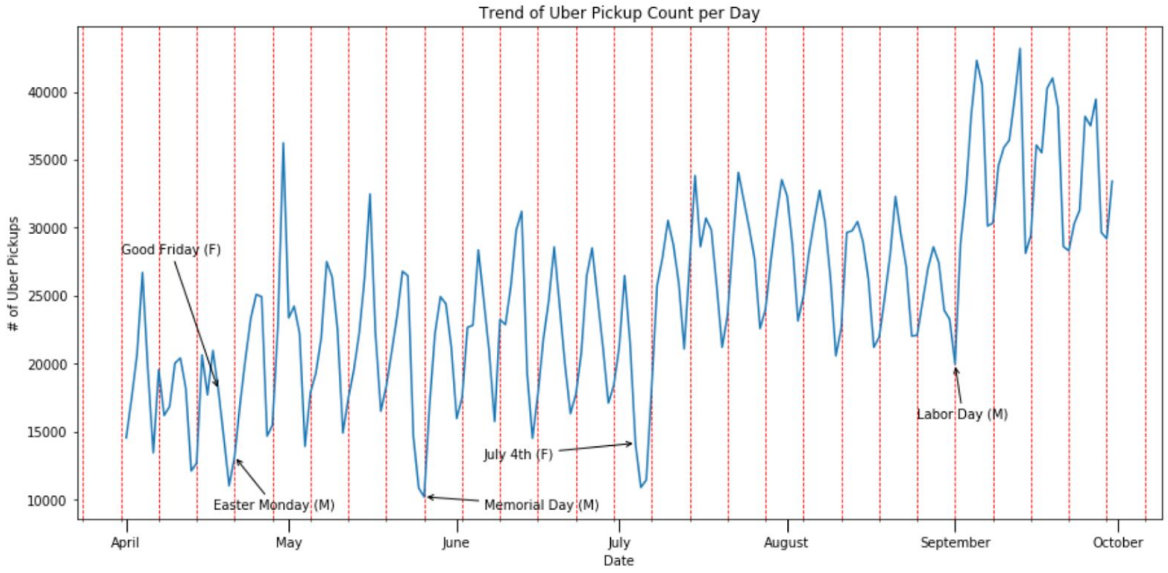


Figure 8. *This graph shows a trend of how many Uber pickups occured on each day in New York City from April 1st to Sept 30. Each of the red dotted lines highlights the Monday of each week. The Major holidays of this time period are labeled on the chart.*

Although for most weeks the Uber pickups peak on Thursdays, there is a different story during holiday weekends. During the Independence day weekend, for example, the number of rides peaked on Wednesday and started to drop from Thursday to Sunday (Figure 9). Contrary to what we would expect, in 2014, people in NYC didn't ride on Ubers when they have the day off. It is possible that this decline in ridership occurs for several reasons. People may be out of town or they may simply be staying at home. NYC also has a good transportation system so people may be riding the subway instead if they are not in a hurry to get anywhere.

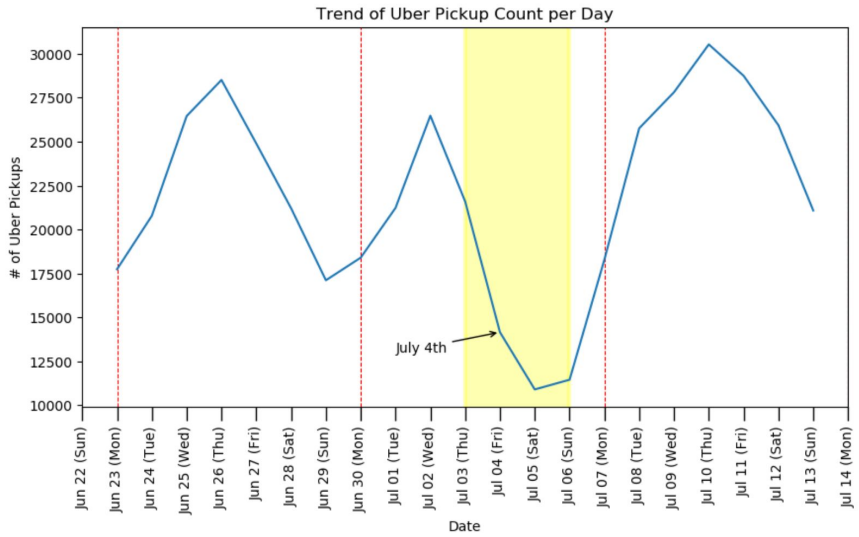


Figure 9. *Uber Pickup Count per day, July 4th Weekend. This graph is a zoomed in version of the previous plot. The red lines indicate the mondays of each week, and the yellow block highlights July 4th weekend where many people don't have to work.*

For the holidays falling between April and September of 2014, the overall number of Uber pickups is shown in Figure 10. It appears that there is no difference in overall Uber usage based on whether people get the day off of work or not.

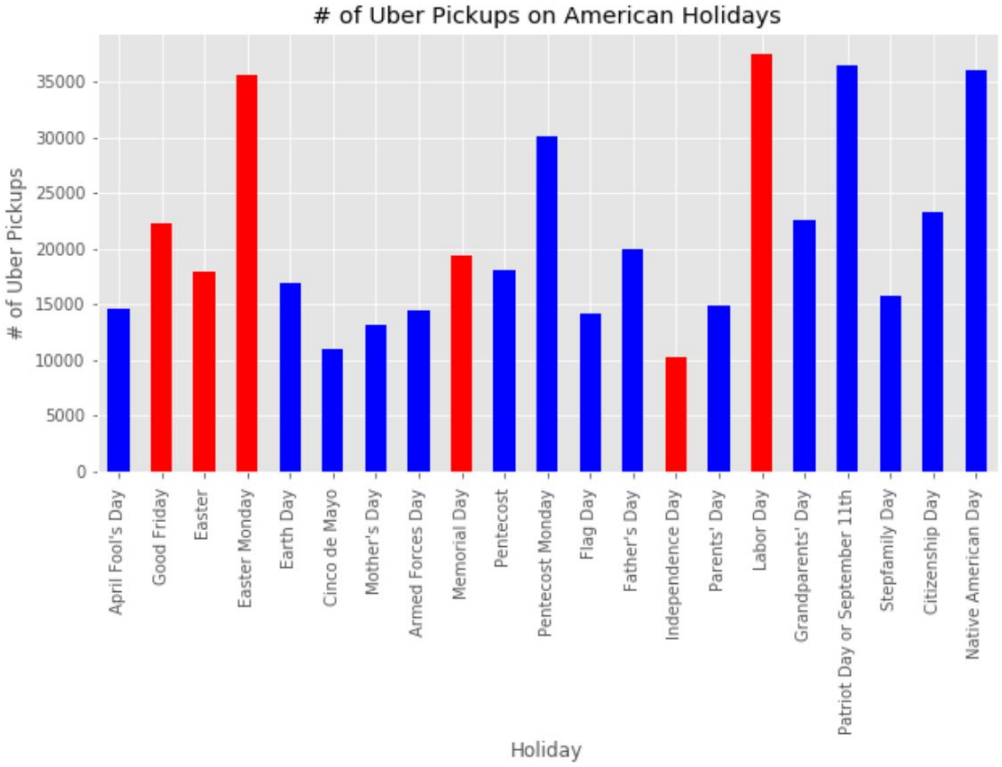
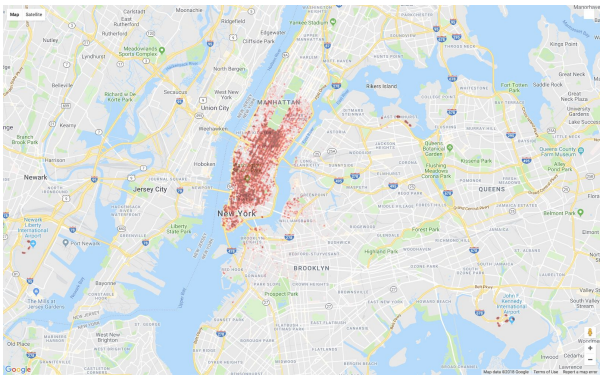
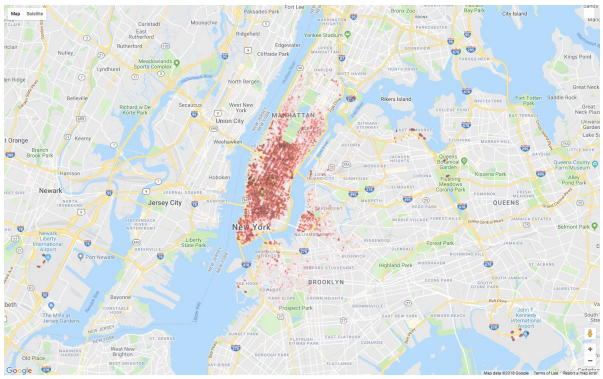


Figure 10. *Number of Uber PIckups on American Holidays. This chart shows all the holidays from April and September. Red bars indicate holidays that are celebrated by many.*

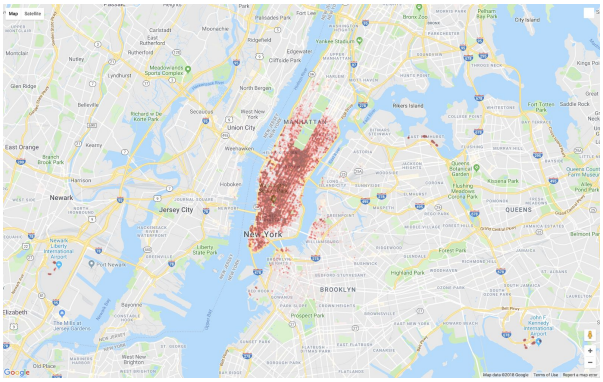
To further understand Uber usage during the holiday weekends, heatmaps for Uber rides during several key weekends are presented below. Memorial Day weekend (Figure 11a) and Labor Day weekend (Figure 11b) show little difference in where the pickups are located from a non-holiday weekend (Figure 11c). The overall count of rides is lower than a typical weekend, which is consistent with what was previously observed.



(a) Memorial Day Weekend



(b) Labor Day Weekend



(c) Plain Vanilla Weekend

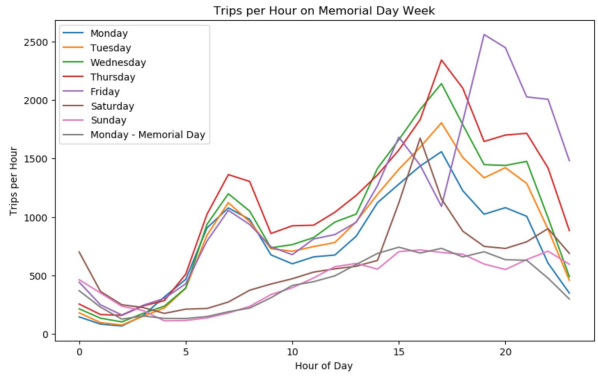
Figure 11. *Heatmaps of Uber pickups over several holiday weekends. No clear difference between pickup locations. Holiday weekends show a lower overall frequency.*

To examine Uber usage as a function of time on the major holidays, the trips per hour by day are shown in Figure 12. The Friday before Memorial day (Figure 12a) there is a peak at around 3PM, suggesting that people are leaving work early. The Uber rides then peak again at 8PM on Friday, probably because people are going out to enjoy the nightlife. When comparing Saturday, Sunday, and Monday of this week to the weekends on a non-holiday (Figure 5), it is noticeable that the Ubers are not nearly as active. People might have had their fun on Friday night and decided to spend time with family on the other days. There is a peak on the Saturday of the weekend at around 5-6PM. There might've been an event at that time, or people just all went out to have a nice meal.

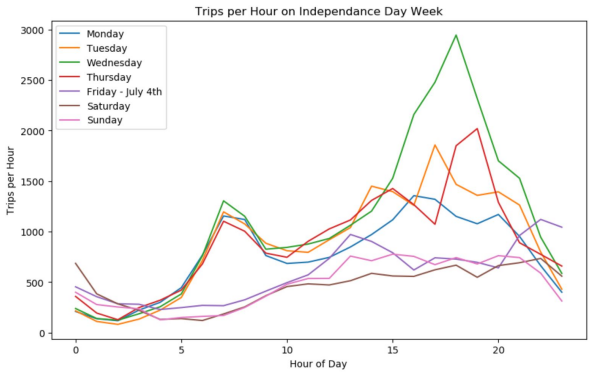
Similar to Memorial day weekend, people seemed to leave work around 3PM on the Thursday before Independence day (Figure 12c). Then the Uber rides peak again around 6-7PM

indicating that people are going out for dinner. Also like memorial day weekend, the Friday, Saturday and Sunday of this week seems to have much fewer Uber rides than the average weekend.

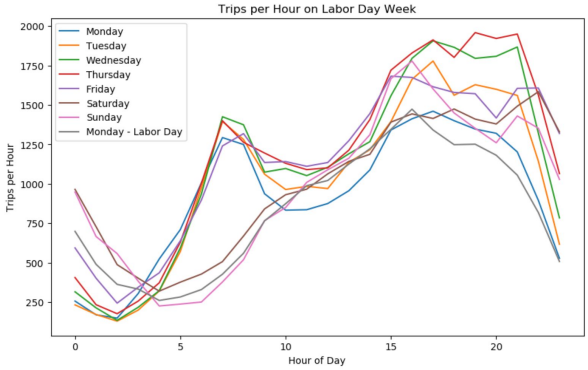
On the Friday before Labor Day weekend (Figure 12b), there is also a peak around 3PM. With all three holidays, people seem to leave work early. Unlike the previous two holidays however, people seem to use Uber a lot on the Saturday, Sunday and Monday of the weekend from 10AM onwards. For some reason, people seem to use Uber a lot more in the month of September, even on holiday weekends.



(a) *Trips per Hour by Day for Memorial Day Week.*



(b) *Trips per Hour by Day for Independence Day Week.*



(c) Trips per Hour by Day for Labor Day Week.

Figure 12. *Trips per Hour by Day for several Holiday Weekends. There are distinct differences in Uber pickup frequencies by hour for the different holiday weekends.*

The data from this analysis seem to suggest that people do not use Uber a lot during holidays where they have the day off. In fact, there are fewer Uber pickups than average. This suggests that people tend to use Uber for work over pleasure in New York City. The reason could be that people are choosing to stay one on the holidays, or they only use Uber when they are in a rush to get somewhere, such as work or a dinner reservation.

5.4 Do weather related variables impact Uber usage?

5.4.1 Temperature:

One of the goals of this analysis is to understand how weather impact uber usage. It is hypothesized that when it is warmer out, people tend to go out more. As one of the first steps a scatter plot was created to explore the relationships between temperature and number of rides (Figure 13).

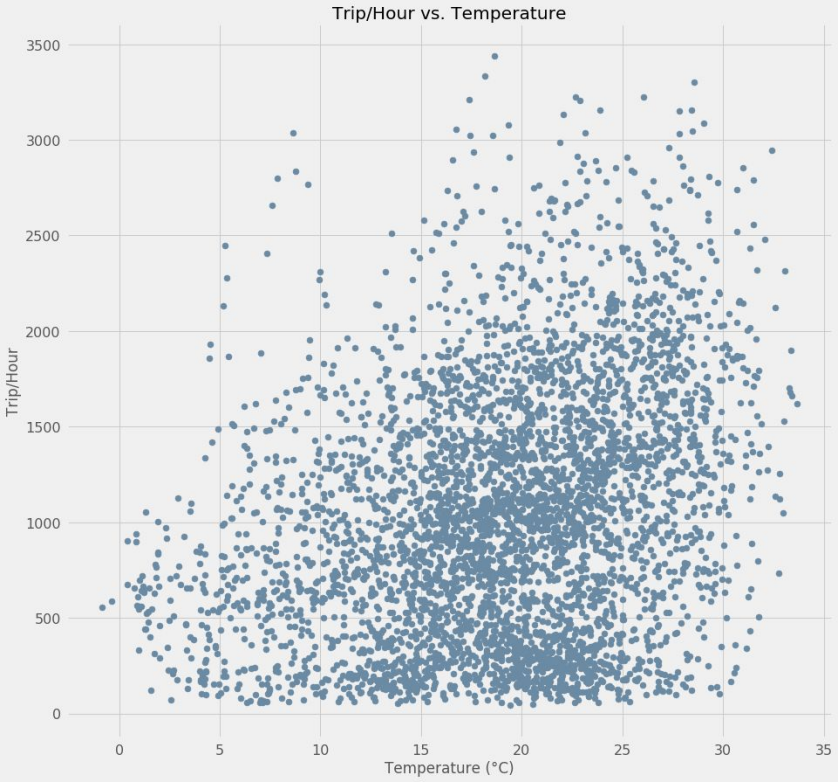


Figure 13. *Trips per Hour versus Temperature.*

5.4.2 Weather conditions:

Hourly description of the weather conditions (e.g. clear sky, raining, storming) are included in the data set. However, as the weather conditions are not evenly distributed (e.g. there are more hours with cloudy weather than thunderstorms), the counts must be normalized when trying to understand the impact of weather condition on Uber ridership.

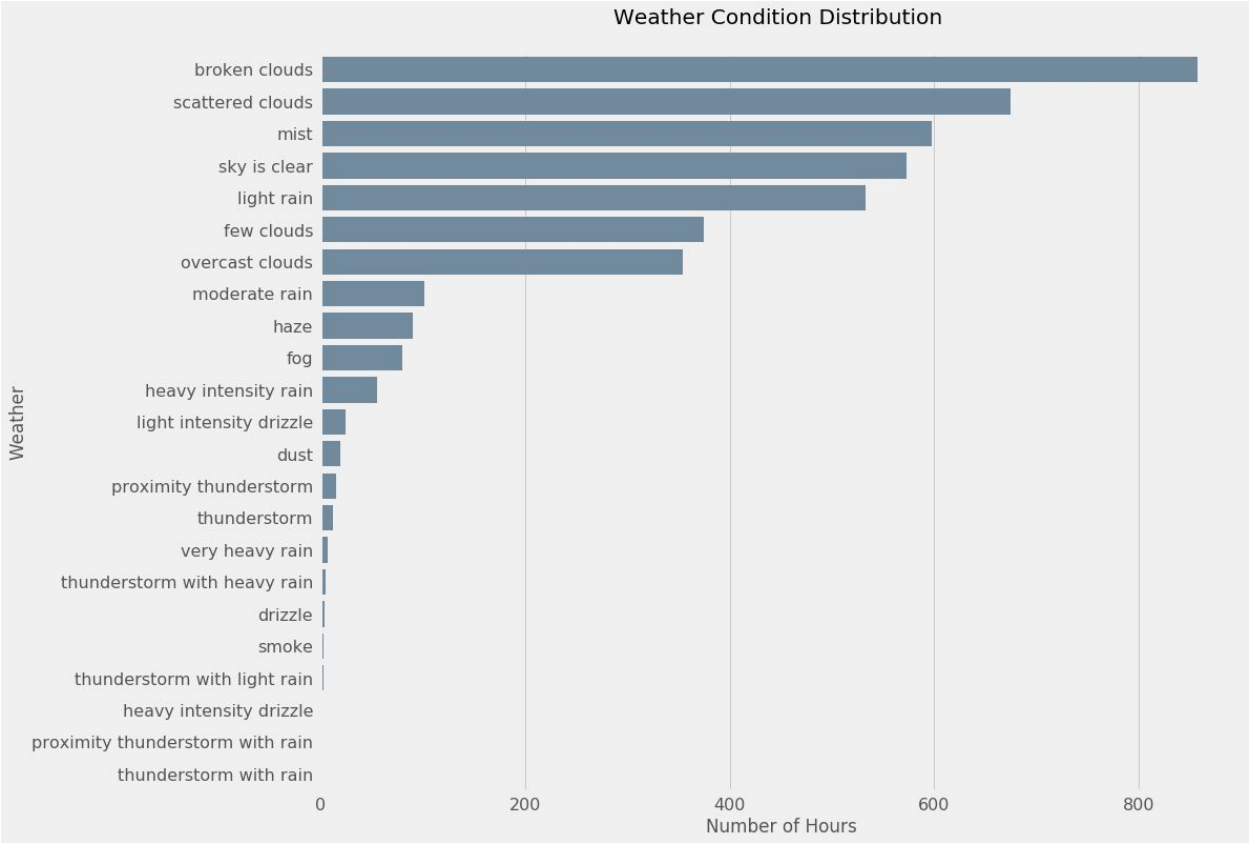


Figure 14. *Number of hours in the data set for each weather condition*

Had the number of trips been simply summed by weather conditions, the weather condition with the most trip would have been “broken clouds”, which would not clarify how weather conditions impact uber ridership. Hence, the number of trips per hour during each weather condition was calculated, and this exposed that there seems to be a higher rate of uber ridership when there in wet weather conditions.

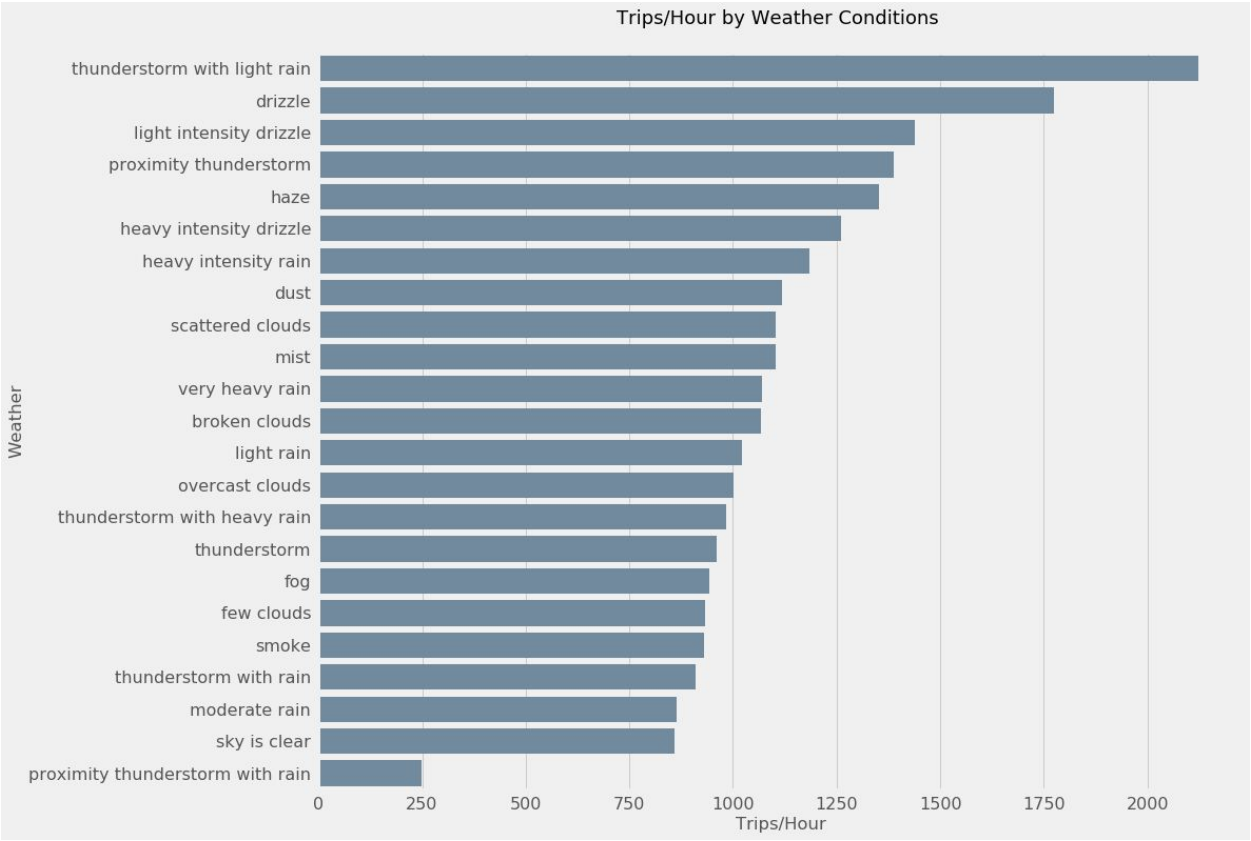


Figure 15. *Number of hours in the data set for each weather condition*

5.5 Factors Impacting the Number of Trips per Hour

5.5.1 Overview:

Since there are a variety of variables in the data set that individually have somewhat of an impact on uber ridership, a regression model was created to understand which factors have the most weight in terms of its impact to Uber ridership.

5.5.1 Ridge Regression

To get started, a few additional variables were created and one hot encoded the categorical variables. In addition, a correlation heat map was used to further understand the size of the effects of each variable on the number of trips per hour (Figure 16).

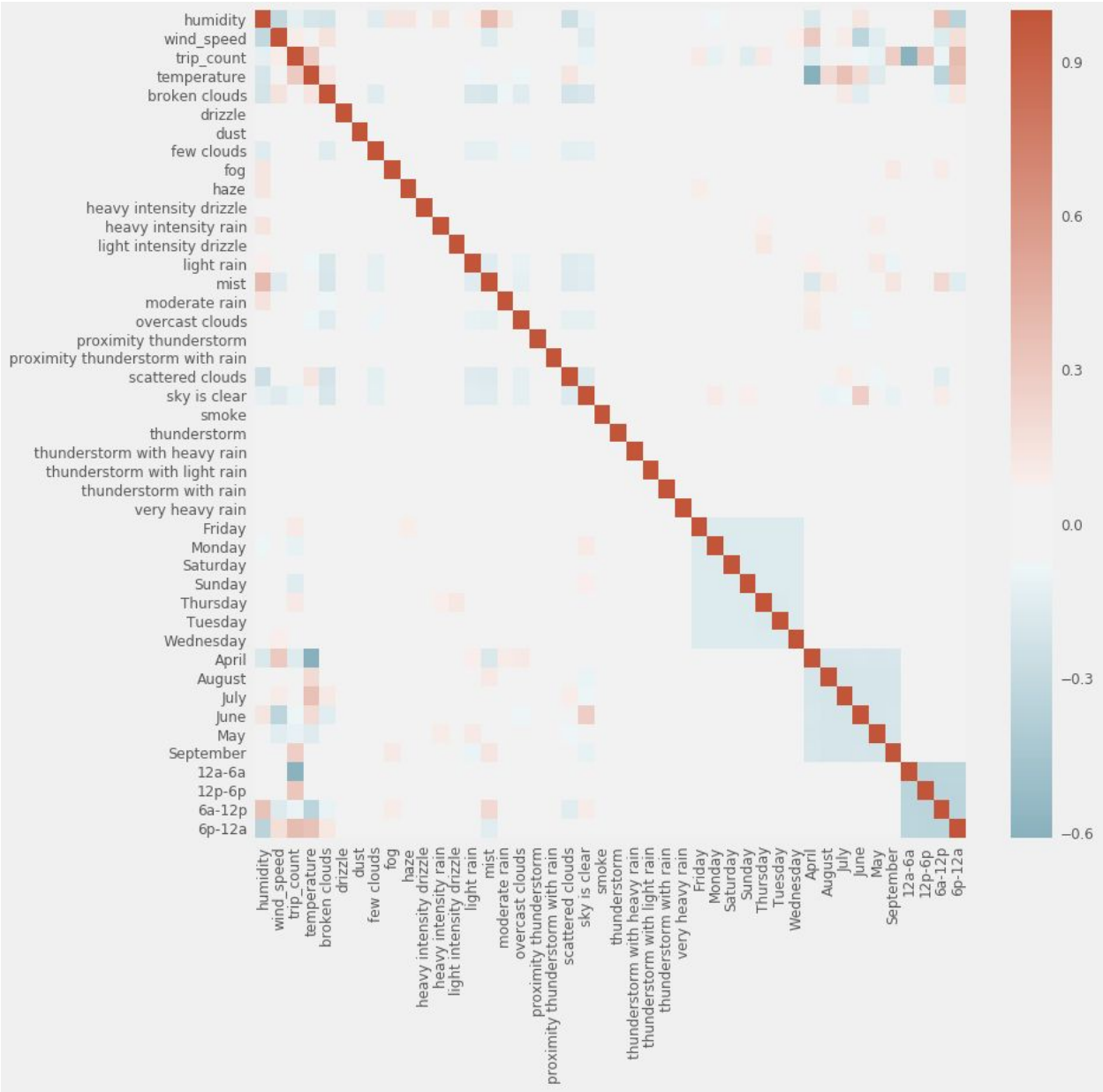


Figure 16. *Correlation Heatmap of variables used in the regression model*

Based on the correlation heatmap, it appears that there are quite a few variables that have small or medium sized effects on the number of trips per hour (denoted by trip_count), and because of this, a Ridge Regression helped illuminate how each of these variables affect affect the number of trips per hour.

The regression model had an R^2 of 0.62 on the training data set and an R^2 of .61 on the test data set, and produced weights for each variable, which can be seen in the following graph (Figure 17).

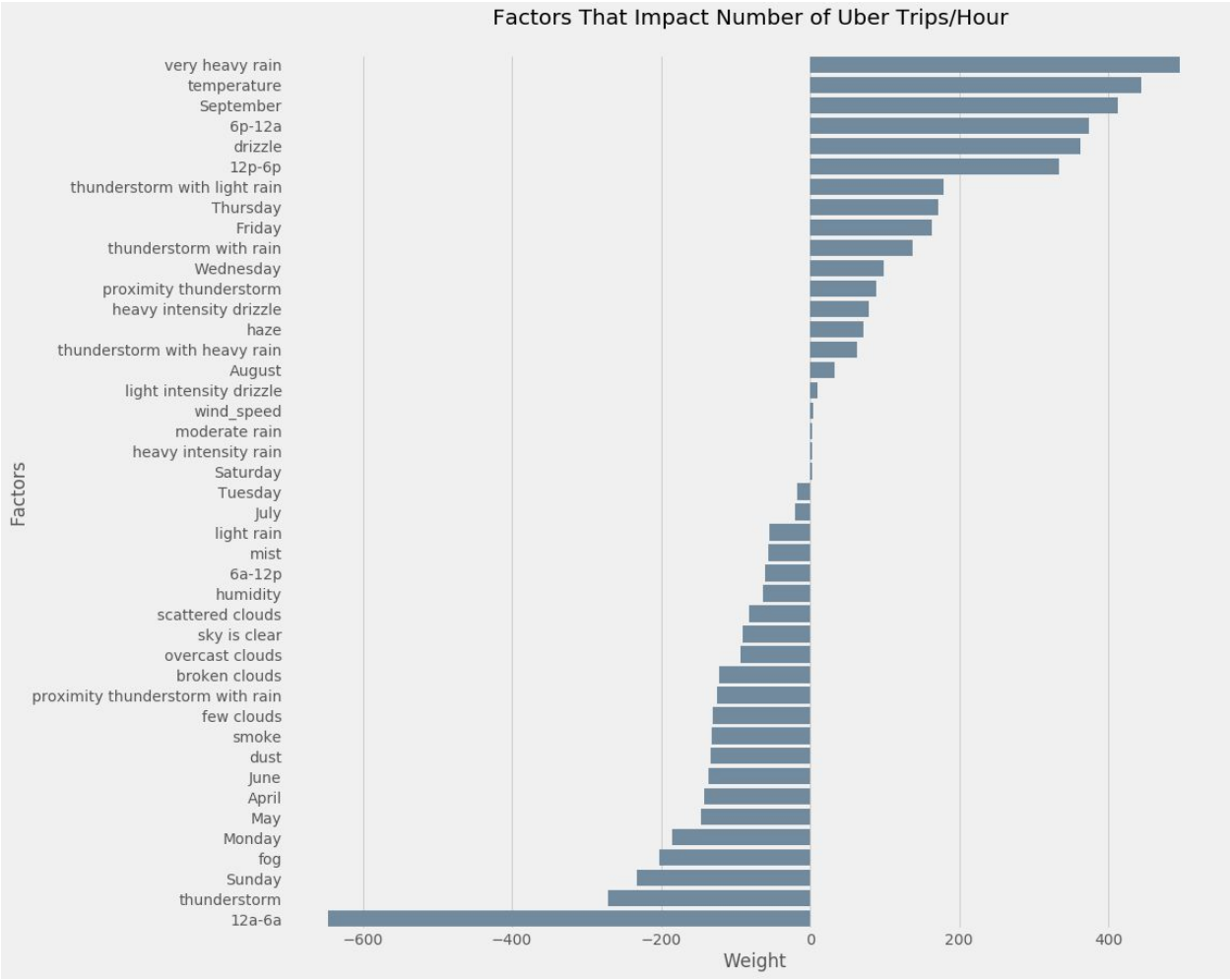


Figure 17. *Factors Influencing Number of Uber Trips per Hour*

Based on the weights, it looks like the number of trips per hour increases when there is wet weather, after 12PM, and whether the trip was in September.

5.6 Do Uber pickup data indicate that we are missing key public transit options?

Urban legend describes a wise architect that never included paved paths in his original designs, but would rather come back a year later and pave the trampled grass. In this way, the paths would be in just the right places. In an attempt to approach public transportation, “trampled” areas here are areas with lots of Uber pickups, but few public transit options. To this end, heatmaps of Uber pickups for Saturdays in September were overlaid with the locations of all public transit (mostly subway and elevated stops) options in NYC (Figure 18).

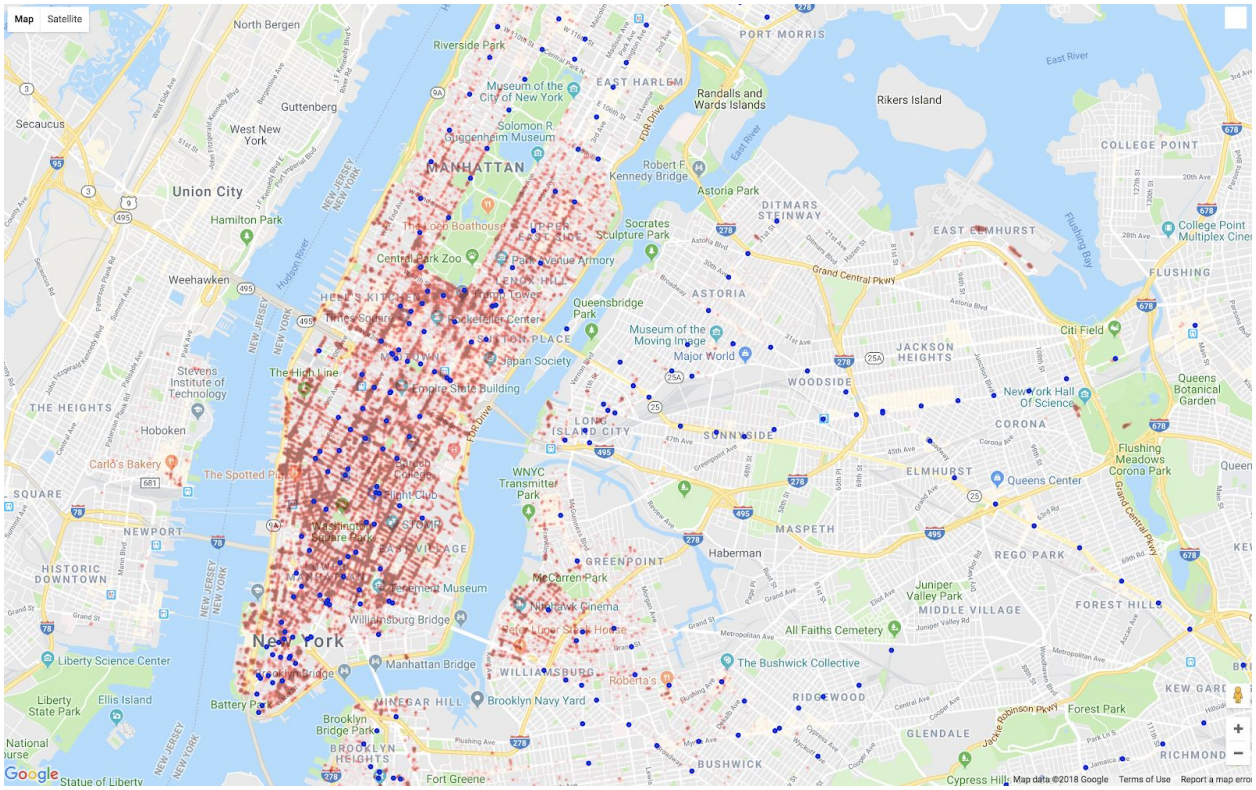


Figure 18. *Heatmaps of Uber pickups on Saturdays in September, overlaid with transit options, blue dots.*

It is clear that there are many public transit options in NYC. It is noteworthy that the densest areas for Uber pickups have public transit options nearby, with the possible exception of the coastal area between the 78 and the 495. It would be interesting to investigate more with full Uber ride details, including destinations.

6. Key Takeaways

Based on the analysis presented above, it is possible to state several conclusions. All of these conclusions, however, are limited because the data used to generate them only span 6 months of time. Ideally, a full year, or more, would be necessary to understand the seasonality of ride sharing usage. In this case, only Spring and Summer data are available, so no conclusions are drawn about snow and freezing temperatures.

People in NYC mostly use Ubers for the purpose of work, going to work around 7-8AM, and returning home around 5PM. More people use Uber to leave work than to get to work. Many New Yorkers do not own cars. In fact, only 44 % of people in New York did own cars in 2012. Uber is an alternative to owning a car.

Uber is also used to go out for evening activities. On Saturdays, Uber rides peak from 3PM to 12AM, and Fridays see a spike after work.

Uber usage showed a drastic increase in September 2014. This may have been caused for several reasons, although these are all speculative. Older children may be riding to school in Uber cars. Uber may have been gaining popularity. The warmer temperature in September may have caused people to go out more, or use Uber more as an alternative to walking.

People in NYC tend not to use Uber on holidays. The number of uber rides drops when people have the day off. Instead, people seem to take Uber when they are time-bound. For example, people are in a hurry when trying to get to and from work. On holidays, however, they may have a more leisurely schedule, which translates to walking or using the subway. The only times when uber rides peak on holidays are during dinner time, or at night. Perhaps at dinner time, people have dinner reservations to make. At night, they may be trying to get home quickly, or avoiding driving because of alcohol consumption.

People tend to use Uber more during bad weather or when it is hotter out. However, this is truly a limited conclusion based on the six-month data set. Because this set contains only Spring and Summer data, the full impact of really cold days is not examined.

The city public transportation seems to be in all of the right places. Uber usage does not pinpoint any specific locations that are clearly missing public transit options. In the future, it would be a useful analysis to compare public transit usage with Uber usage.

APPENDIX I

Project Proposal: Uber Usage
MIDS-W200 Section 001
Project 2: Data Analysis with Numpy and Pandas
The A-Team: Adam Yang, Armand Kok, Alla Hale

Background:
Ride sharing services are extremely popular; one does not have to ask many people before encountering someone who uses these programs frequently. These programs are convenient because they allow users to travel on their own timelines, without the hassles of using their own vehicles.

Data describing the use of these programs can shed light on how people get around in a city. Analysis of when, where, and how people use these services can be used by ride share providers to help optimize their operation, or by city administrators (e.g. city planners, transport authorities) to help minimize traffic or even improve public transportation systems.

Objective:
The objective of this project is to conduct an open-ended analysis of Uber usage data utilizing PyData packages (e.g. NumPy, Pandas) that are covered in class. This project will attempt to answer the following questions, while remaining open to additional discoveries as we iterate through the analysis:

1. Does weather impact Uber pickups?
2. What other factors impact Uber usage?
3. Does Uber usage vary by day, month, time of the day?
4. Do Uber pickup data indicate that we are missing key public transit options?
5. What does Uber usage look like on key holidays?

Data Sources:
We have several data sets to combine for this project. The first is a data set describing over 20 million Uber (and other for-hire vehicles) trips in NYC from 2014 to 2015. This data set includes dates and times along with the location (latitude and longitude) of each Uber pickup. For the other for-hire vehicles, the trip information varies by company, but can include day of trip, time of trip, pickup location, driver's for-hire license number, and vehicle's for-hire license number.

The second data set describes the weather. This data set is much broader than needed for the Uber analysis, including weather for 36 cities from 2012 to 2017. The NYC weather data will be used for the same time period as the Uber data are available.

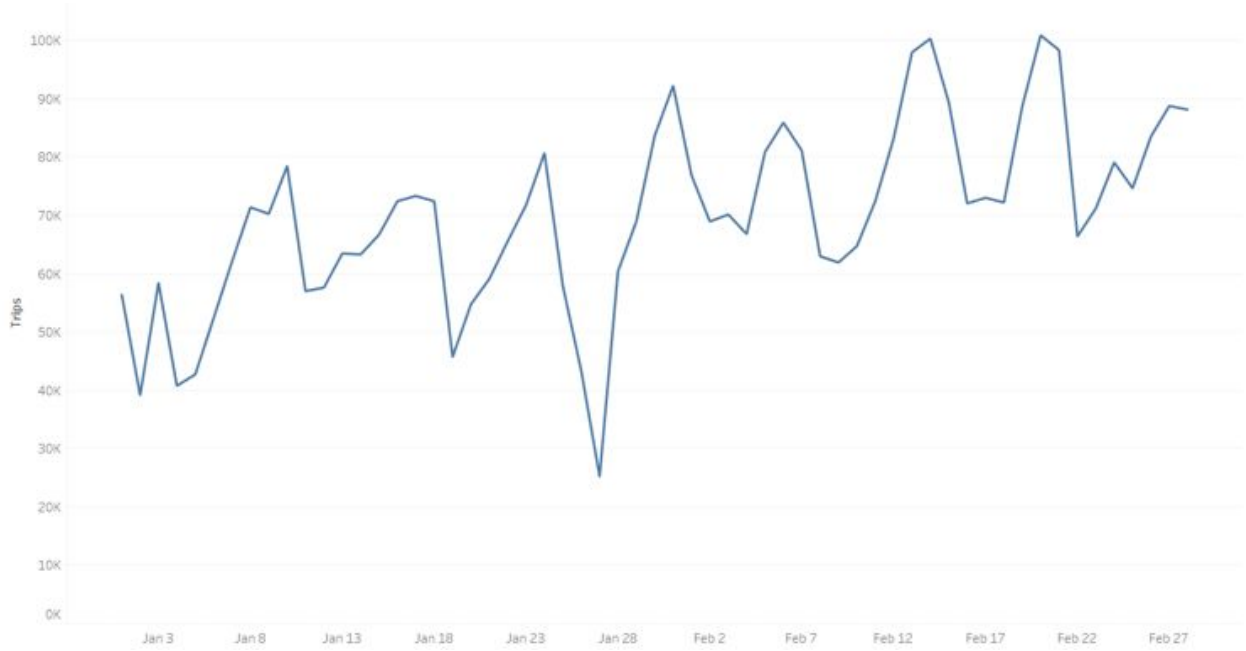
The third data set provides locations (latitude and longitude, and address) of public transit stops (subways and buses), that will be used to help answer questions around public transit as it relates to ride sharing.

Uber/for-hire vehicle data set is located here:
https://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city/data

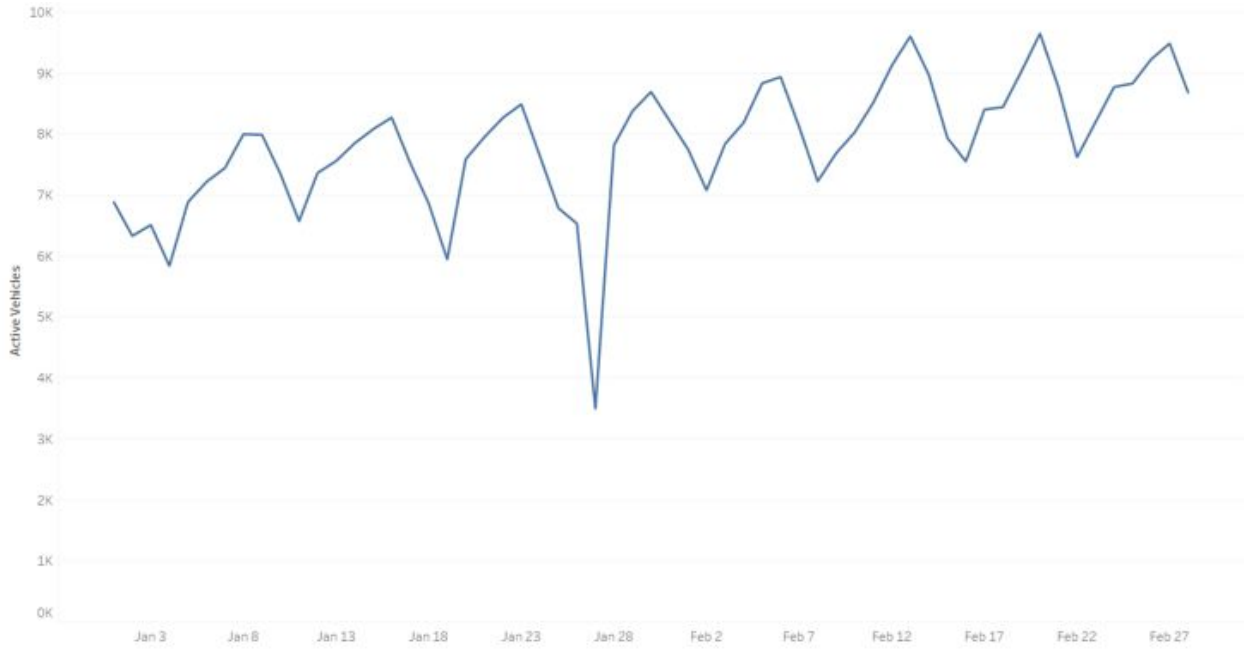
Weather data set is located here:
https://www.kaggle.com/selfishgene/historical-hourly-weather-data/data

Public transit data set is located here:
http://web.mta.info/developers/developer-data-terms.html#data

Initial Figures:
Number of trips (January 2015 – February 2015)



Number of active vehicles (January 2015 – February 2015)



Note: there was a major snowstorm on January 27th, 2015 which caused the large dip in trips and active vehicles