ANALYZING UBER TRIPS OCCURED IN NEW YORK CITY

SUMMARY

Uber has become hugely popular in New York, and its trips outpaced yellow taxis for the first-time last year. There are about 65,000 vehicles affiliated with Uber in the city, which provide more than 400,000 trips per day, according to the Taxi and Limousine Commission. Objective of this study is to provide an analysis of Uber Rides in New York City.

DATA PREPARATION AND EXPLORATION

We required Uber data specific to New York City that aggregated key ride metrics including trip distance, trip duration, pick-up and drop-off locations and trip date. Data is retrieved from Amazon Web Services (AWS). Data set includes the trips occurred in New York city between September 2014 and September 2015. Data set consists of 31 million rows (1,5 GB size). Due to memory constraints, random sample of 100 thousand rows of data is used for analysis. 4078 rows of data is ignored because of missing values in the data set. No duplicate values are found in the dataset. In total, 4% of the records are ignored. 96% of original data is retained for data analysis.

DATA LIMITATIONS

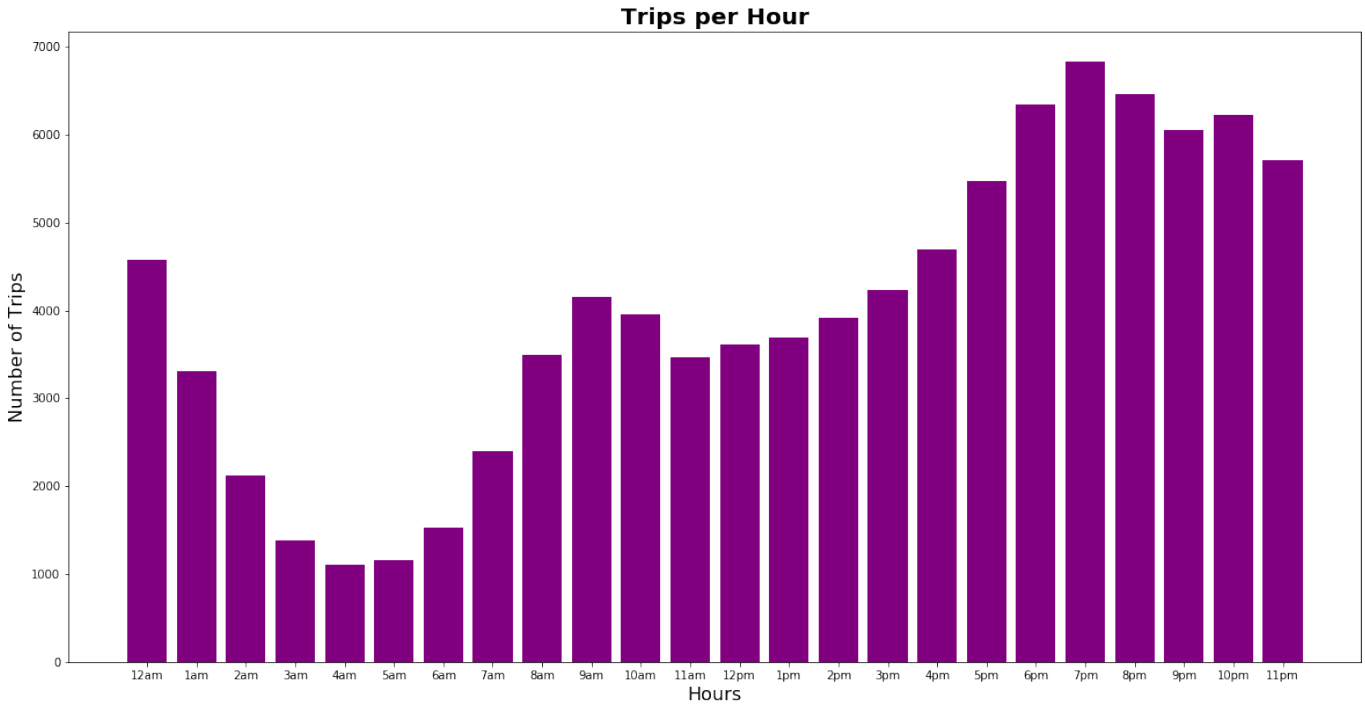
We do not have direct access to drivers to confirm that the entries in the data set are accurate with regard to distance, location and time. Additionally, we had to remove multiple outliers from the data set, because they were either impossibly short given the distance purportedly traveled or they were exceedingly long in duration, one of which exceeded 7 days. Due to memory constraints, we used 100,000 rows of data, which is relatively small given the 31 million data points in the source file.

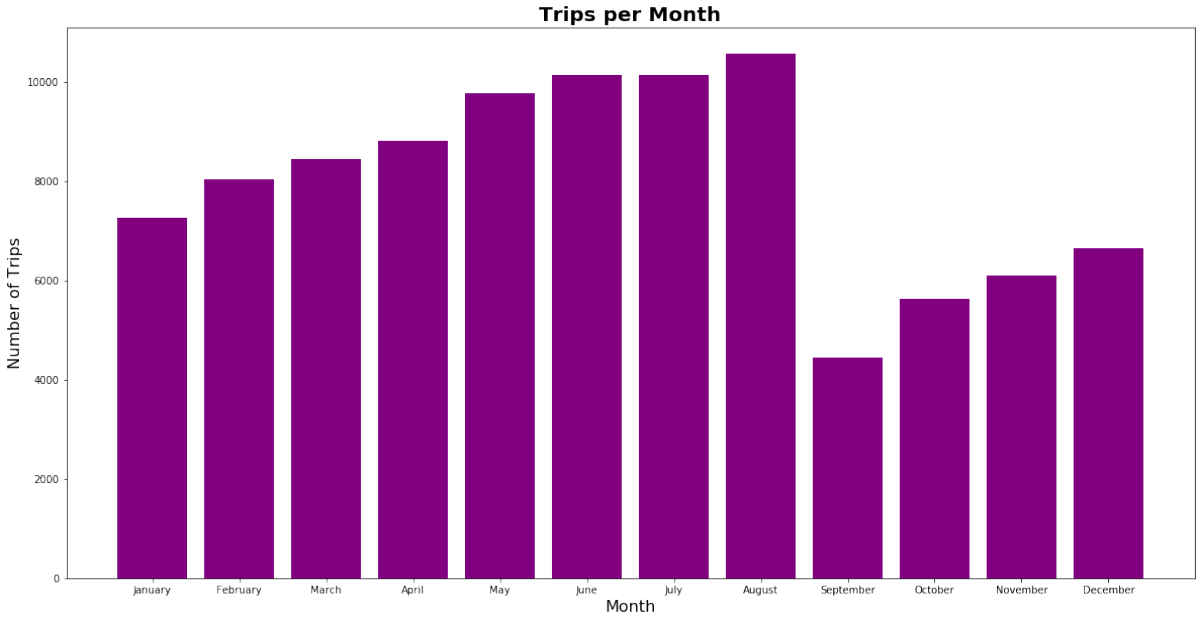
QUESTIONS & HYPOTHESES

-
- Do Uber rides occur more often after work?
  Hypothesis: There will be a higher prevalence of Uber rides after 5:00 PM in NYC
- Which date of the year are Uber rides requested more?
  Hypothesis: There will be more trips per day on Christmas and New Year's Day in NYC
- What is the relationship between trip distance and trip duration?
  Hypothesis: There is a positive correlation between trip distance and trip duration in NYC
  Hypothesis: 95% of the variance in trip duration is explained by trip distance in NYC
- What are popular pick-up and drop-off locations?
  Hypothesis: Some pick-up and drop-off locations will be more popular to others
- What is the relationship between staying in the same neighborhood and distance?
  Hypothesis: There is a strong association between distance traveled and whether you stay or leave the neighborhood
- What is the relationship between leaving the neighborhood and distance?
  Hypothesis: There is a negative correlation between Staying in the Neighborhood and the Distance traveled
- What is the difference of trip distance on New Year's Eve and average trip distance?
  Hypothesis: Trip distance on New Year's is longer than average trip distance in NYC

DATA ANALYSIS & DATA VISUALIZATIONS

When do Uber rides occur most?





Which dates of the year are Uber rides requested more?

The Most Trips Per Day

I want to find the most trips per day. My hypothesis is the most trips per day is Christmas or New Year's Eve.

How I Start

In my analysis I want to visualize the the number of trips per day for approximately one year: September 1, 2014 – September 1, 2015. I create a dataframe dateGroupDF.dates and grouped-by 'dates' and 'tripCount'. *The data I use for my portion of the analysis is Uberdata.csv.

How to plot the graph

I generate the line graph and emphasize the changepoints of holidays, weather and unique events. I use matplotlib.offsetbox import (OffsetImage, AnnotationBbox) to create the image box and matplotlib._png import read_png to load icons (.png file) to (x,y) position.

How I create the visualization

After the graph generates, I determine the max and min trip count by day. I want to ensure I plot these two dates with icons. However, these dates invalidate my hypothesis. The max date is June 27, 2015 and the min date is January 27, 2015. I research the changepoints and graph accordingly.

Summary

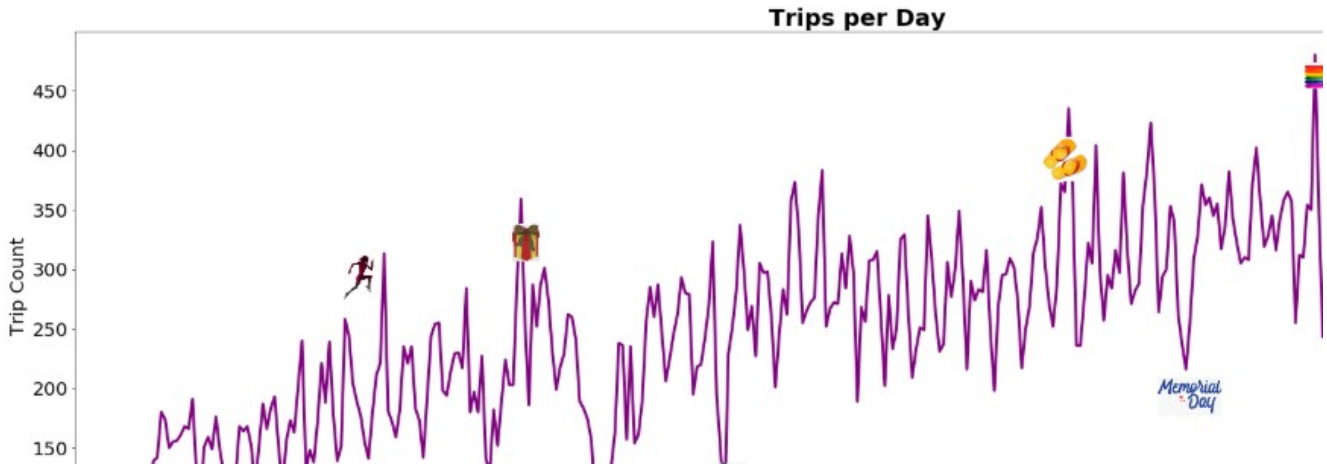
The hypothesis is invalid. The max number of trips in one day in 2015 is not Christmas or New Year's Eve. It is June 27, 2015. On June 26, 2015 the US Supreme Court legalized gay marriage. Uber NYC marched along with friends and family to the NYC Pride Parade on June 28, 2015.

dateGroupDF.loc[dateGroupDF.tripCount == dateGroupDF.tripCount.max()]

dates	trip count
6-27-2015	480

dateGroupDF.loc[dateGroupDF.tripCount == dateGroupDF.tripCount.min()]

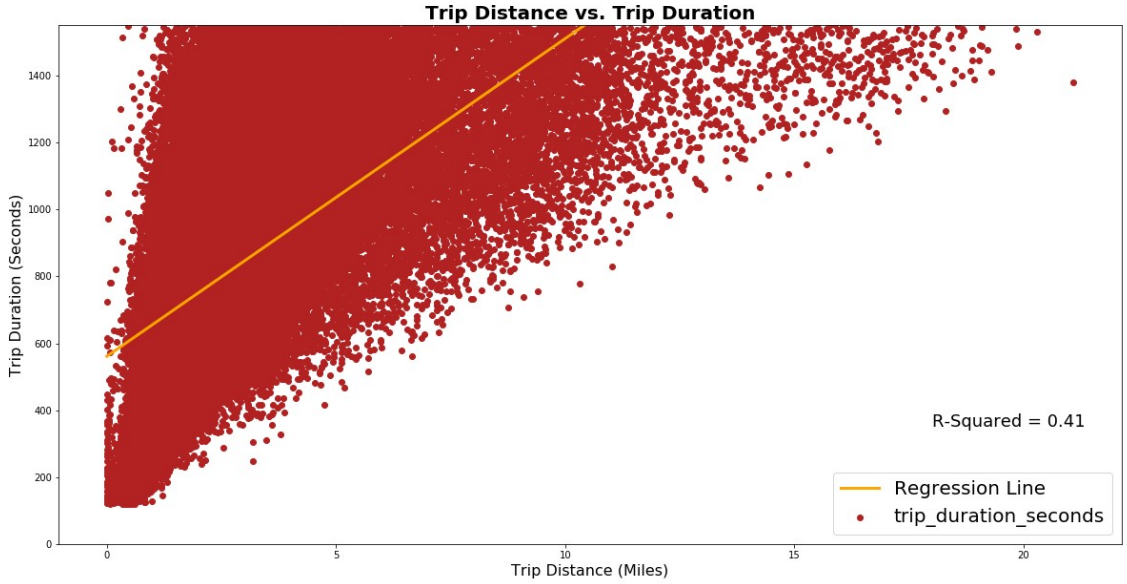
dates	trip count
1-27-2015	91



What is the relationship between trip distance and trip duration?

We hypothesized that trip duration and trip distance are highly correlated and estimated an R2 of at least 0.95. However, the data does not support this with an R2 of 0.41 calculated on a 100,000 entry set. Even so, the p-value of the dataset was found to be less than 0.05 (at 0.00) and is therefore significant and we reject the null hypothesis that trip duration and trip distance are not related.

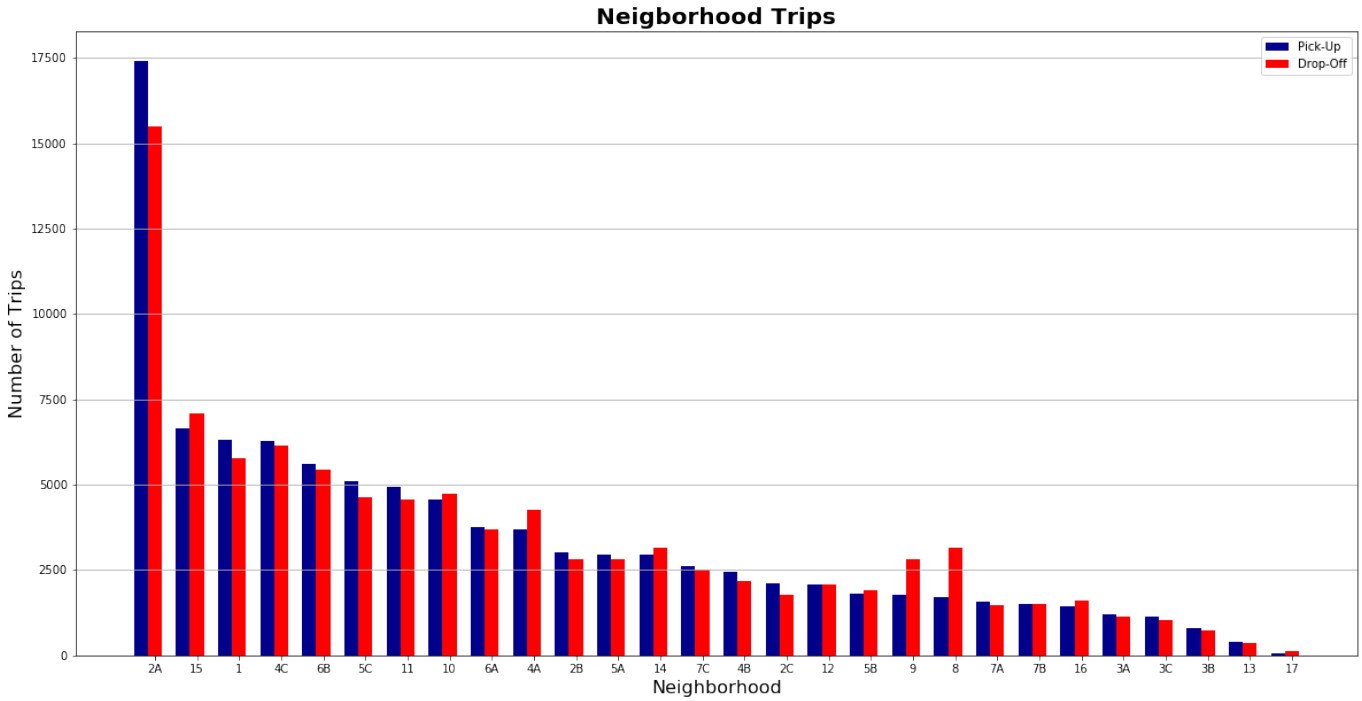
We induce that, particularly in New York City where the data hails, heavy traffic congestion in the nucleus of a city leads to a drawdown in the average distance per unit of time compared to more distal areas that allow drivers to travel greater distances at a more rapid cadence. Therefore, in certain areas of extreme traffic minimal distance was covered per unit of time, while others allow for a more steady and swift flow. Also, at the recommendation of our instructor, we capped the upper bound of duration at the third quartile to exclude substantive outliers, one particular point asserting that a trip lasted over 7 continuous days.



Additionally, we truncated the data to only include Uber trips that lasted longer than two minutes as many rides were logged as lasting zero seconds while simultaneously making headway with regard to distance. Assuming that teleportation is least likely to happen in a gridlocked New York City artery if anywhere, these points were also removed from our figure.

What are popular pick-up and drop-off locations?

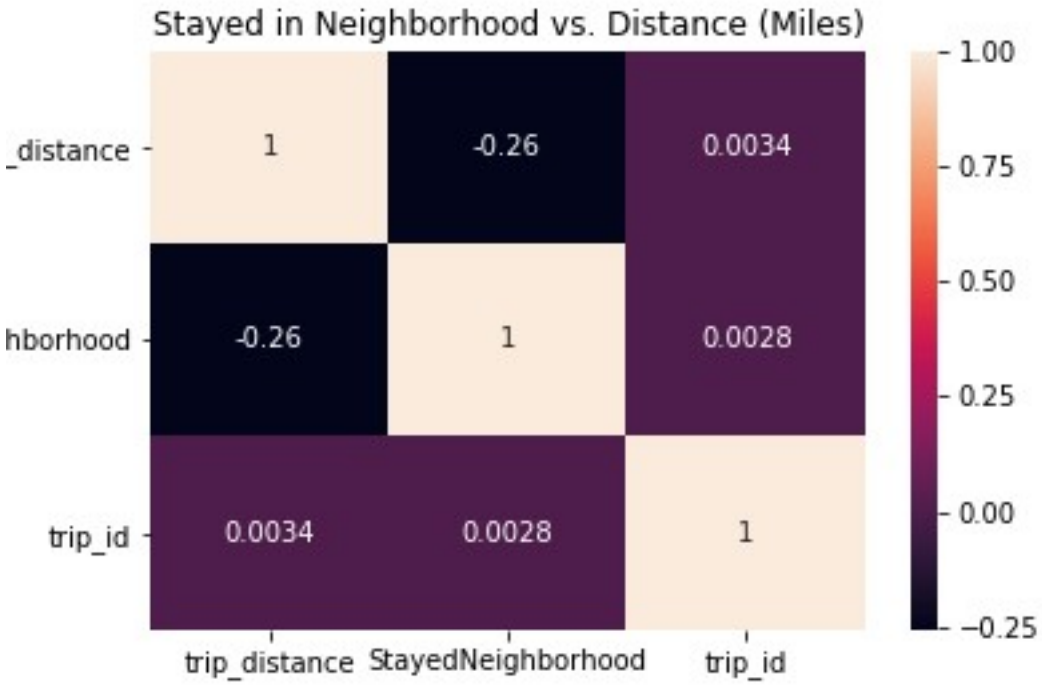
In our dataset, there are 28 unique pick-up and drop-off location labels (e.g. “4A”). We assume that those location labels refer to neighborhoods since there are trips started and ended at the same location (e.g. “4A to 4A”). We made a research to find out what these labels refer to, however, we failed to find any result. That is why we will not be able to provide detailed insights about locations/neighborhoods. When we analyze the data, we see that “2A” is the most popular neighborhood with the most number of trips among pick-up and drop off locations.



What is the relationship between leaving the neighborhood and distance?

There seem to be a slight positive correlation between “leaving the neighborhood” and “the distance traveled” which proves that there is some measure of dependence between the two variables. The value is not too close to 1 which signify that there is no strong positive correlation.

What is the relationship between staying in the same neighborhood and distance?



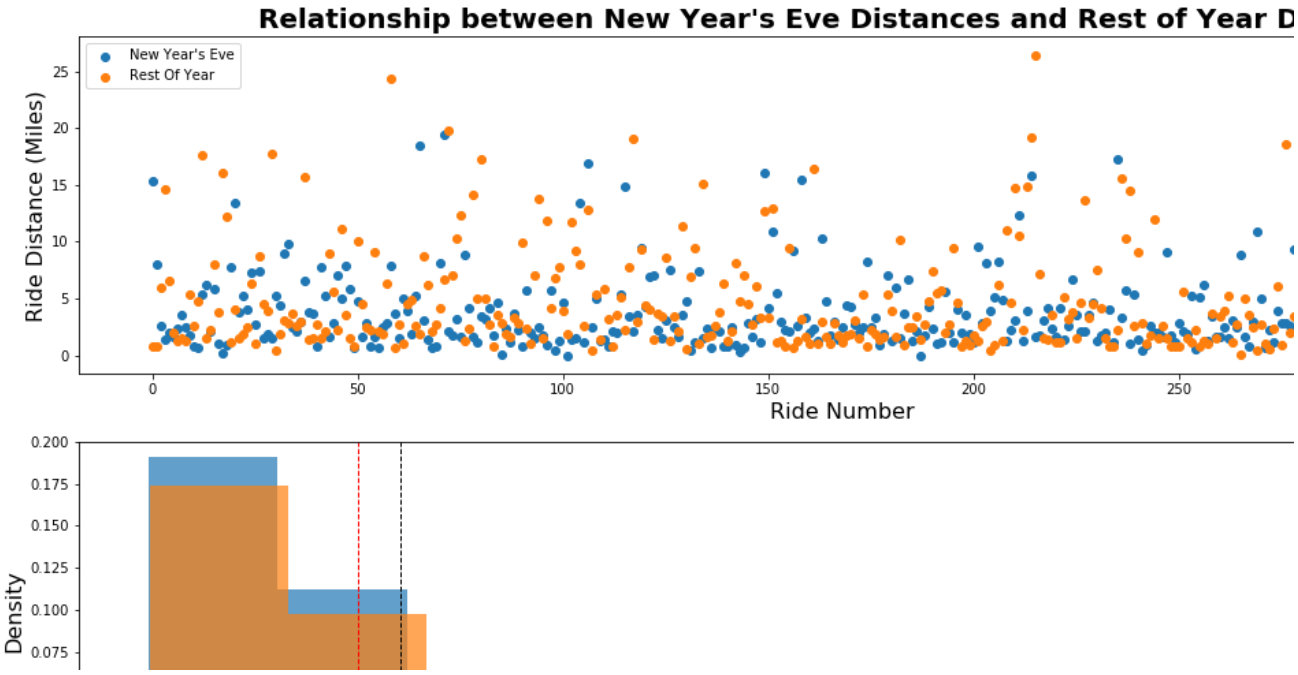
Conversely the relationship between "Staying in the Neighborhood" and the "Distance" has a negative correlation which signify the distance traveled is related to Staying in the neighborhood.

What is the difference of trip distance on New Year's Eve and average trip distance?

In our analysis, we would like to take a closer look to the days that people are off-work and more socialized. In those days, people are more likely to travel and possibly tend to pick up more Uber rides. We see that the number of trips occurred on New Year's Eve is 337, while the average daily number of trips occurred rest of the year is 263.

Since the number of trips on New Year's Eve more than average daily number of trips rest of the year, we chose New Year's Eve to analyze. Besides the number of trips taken in that day, we analyzed the uber trip distances occurred on New Year's Eve and the rest of the year. We would like to observe if people tend to travel to the far destinations that day or they prefer closer locations to enjoy their day.

We used independent samples t test to compare the average trip distances of New Year's Eve and the rest of the year, since variations of these two samples are close to each other. T test helped us to compare the means of these two independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different or not.



The null hypothesis (H_0) and alternative hypothesis (H_1) are as follows:
H_0: μ1 = μ2 ("Uber trip distances occurred on New Year's Eve and rest of the year are equal")
H_1: μ1 ≠ μ2 ("Uber trip distances occurred on New Year's Eve and rest of the year are not equal")
Our significance level is assumed as 0.05.

As a test result, p-value is 0.02. Since our p-value is less than significance level 0.01 ($p < \alpha$), we reject H_0. We have sufficient evidence to conclude that Uber trip distances occurred on New Year's Eve and rest of the year are not equal at 0.05 significance level.

CONCLUSIONS

We need data sets from 2016, 2017 and 2018 and compare the same holidays, weather and cultural events. If each year has the same min days holidays or weather events, my hypothesis will continue to be invalid. Further research is needed to determine why these days are lowest days for Uber trips per day.

We see that there is a higher prevalence of Uber trips between 7:00pm – 8:00pm and a higher prevalence of Uber trips in the month of August.

There is a positive correlation between trip distance and trip duration in NYC.

"2A" is the most popular destination with the highest number of trips as pick-up and drop-off location. Average trip distance on New Year's Eve is shorter than daily average trip distance rest of the year.