First Draft

Milica Cvetkovic(mcvetkovic) Keyu Hu(khu54) Steven Min (nmin) Xintong Shi(xshi242)

Group 11

December 1, 2020

1 Introduction

The data is called New York City Taxi Data, and it is from 2013. In total, there are about 14,500,000 trips per month, and 174,000,000 trips in a year. The data is stored in CSV format, organized by month. For each month, there are two separate CSV files – one for the taxi trip data and one for the taxi trip fare. The taxi trip data files include information such as car ID, driver ID, pick-up and drop-off date/time, pick-up and drop-off location, trip time, and trip distance. The taxi fare data files include information such as car ID, driver ID, payment type, fare amount, surcharge, tip, and total amount.

2 Statistical Questions

How does speed vary with time of day? - To answer the question, we used the distance formula: $speed = \frac{distance}{time}$. We grouped the data by each hour starting from 00:00h to 23:00h and found the average speed at each hour. Then, we made a table and plotted the findings. From the plot, we found that from 8:00-19:00, the speed is quite stable, and it is at the lowest. The average speed seems to increase over night, with the peak around 5:00, before it starts decreasing again.

Which area has the most taxi pick-ups? - We look at the pick-up longitudes and latitudes of the taxi trip data. We counted how many times each coordinate appeared and made a list of top 50 most common coordinates. This tells us the locations that the taxi-drivers are most likely to get customers. We made a table of the coordinates and created a heat map. Note for the reviewers, the heatmaps are not as dense as we worked only with the portion of our dataset before working with CHTC.

Can we predict trip time from pickup time, trip distance, and pickup and drop-off location (and/or other variables)?

Is it possible to predict tip amount given trip distance, pickup time, trip time and/or other variables in the dataset?

3 Model Overview and Statistical Analysis

The data is from the University of Illinois at Urbana-Champaign databank, and the dataset was made publicly available under the CC0 license. The data was obtained through a Freedom of Information Law (FOIL) request form the New York City Taxi Limousine Commission (NYCTL). We work with 12 trip data CSV files, which are around 1.7-1.8 GB each, and 12 trip fare CSV files, which are around 1 GB each. In total, there are about 33-34 GB.

To answer the first question, how the speed varies with time of day, we looked at six trip datasets from January to June which are 10.93 GB in total. We used R, specifically the dplyr package to efficiently manipulate our data. To do our calculation, we got the trip distance variable, which is already in our file, and we divided trip time in seconds variable by 3600 to covert to trip time in hour. This enabled us to find the speed in miles per hour. We found the speed at each hour of the day for January, and we repeated the same process for the five remaining months. Then, we averaged out the speed from January through June and made a table and a plot.

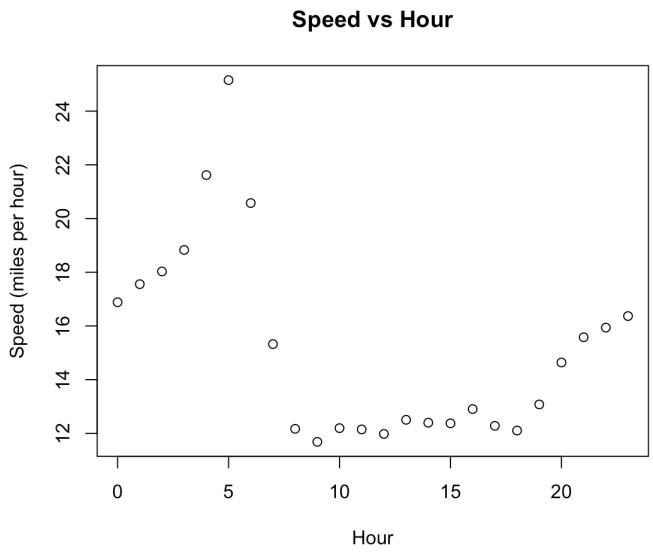

Figure 1: Plot average speed per hour

To answer the second question and to find the area that has the most taxi pick-ups, we investigated the busiest location by using the pick-up longitude and pick-up latitude variables as proxies. We used the same six months of data as for the Question 1. Using R and dplyr package, for January, we rounded the coordinates to four decimal places since the original dataset has up to six decimal places. Rounding the decimal to four decimal places make the data more general and reasonable. After that, we counted the coordinates based on how many times they appear in the dataset, and we made a list of top 50 coordinates. Then, we repeated the same process for the remaining five months. When looking at the top 50 coordinates, the (0,0) coordinate appeared in the lists of top 50 coordinates for each month, and these are obviously bad data points so we removed them. We combined the top 50 coordinates of January through June and plotted on a New York City map to indicate which areas have the most taxi pick-ups. The popular pick-up locations are shown by the map below, with the darker circles being more popular. From the map below, we see the airports being extremely popular pick up locations.

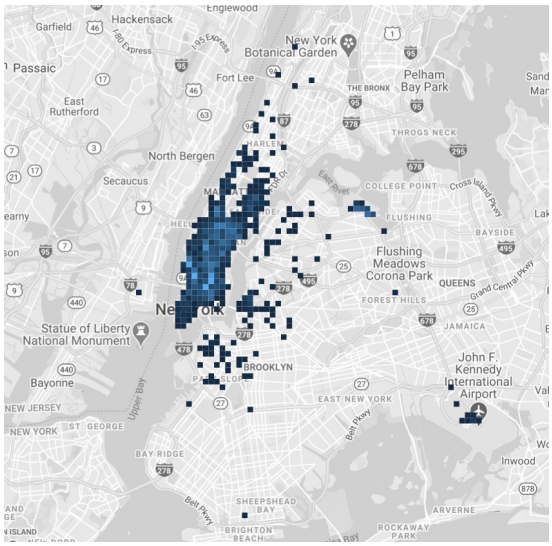

Figure 2: Heatmap

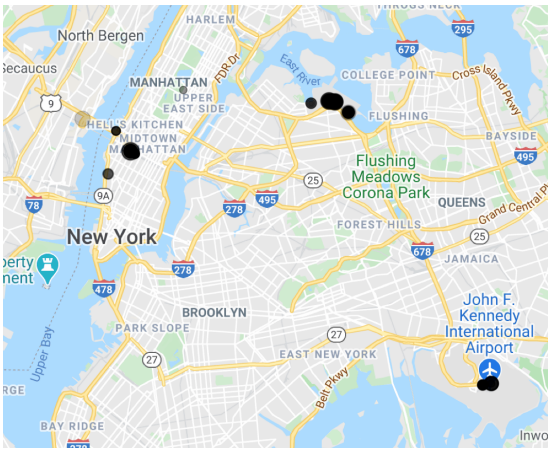

Figure 3: Heatmap

In general, a difficulty we encountered was figuring out bad data points. From skimming through the first few lines of data, it is hard to find out which type of errors could be present in our dataset.

To predict the tip amount, we built a multiple linear regression model. The tip amount is the response variable, and the explanatory variables include trip distance, trip time, vendor id and pickup time. The model summary shows that only the coefficients for trip distance and trip time are statistically significant. The adjusted r-squared is about 58%. Note to the reviewer, we are aware that this model is not very precise since we only used the portion of our

dataset. Once we used the CHTC, we will have a more accurate model. The current model formula:

$$tip = 0.4824856 + 0.3279138 \times tripDistance + 0.0130654 \times tripTimeInSec - 0.0516051 \times vendorId + 0.1556752 \times pickupTime \quad (1)$$

4 Conclusion

The results from investigating speed at each hour seems reasonable with the fastest speed being at 5 am, and then, it decreases as the day progresses due to traffic. The map of popular pick-up locations shows that the airports are popular pick-up locations, but more work could be done to expand the data points on the map such as rounding to different decimal places and taking more than 50 coordinates. For the portion of the data, our analysis seems reasonable, and we are planning on attempting to reproduce this using the large dataset. An additional visual in the future would be to add interactive heat map showing the change of popular pickup locations over time. We will try to test some additional model and compare the metrics to improve out predictions.

5 References

Brian Donovan and Daniel B. Work. "Using coarse GPS data to quantify city-scale transportation system resilience to extreme events." presented at the Transportation Research Board 94th Annual Meeting, January 2015.

Brian Donovan and Daniel B. Work "New York City Taxi Trip Data (2010-2013)". 1.0. University of Illinois at Urbana-Champaign. Dataset. http://dx.doi.org/10.13012/J8PN93H8, 2014.