Visual Exploration of New York Yellow Taxi Ridership

Geng Yuxiang

August 12, 2019

Abstract

Yellow taxis have long been the transportation of choice in New York[1]. In recent years, however, peer-to-peer ridesharing platforms has changed the market for taxi business. By offering competitive price and lower average waiting time[2], platforms like Uber and Lyft become major threats to the survival of taxi industry.

1 Introduction

In this study, I will explore attributes that are accosiated with yellow taxi in New York City and factors that affect taxi ridership via visualization, with the main goal of allowing taxi companies make better decisions in allocating taxis. I will start off by investigating simpler attributes including the distribution of trip count in general and compare difference in trip count between weekdays and weekends. Then I will move on to more complex analysis and explore rate of tip per minute by taxi zones and explain its various causes. Finally, I will analyse the effect of rain on ridership by investigating its correlation with trip count and rate of tip per minute.

2 Data and Tools

The dataset which I used is yellow taxi trip data from the entire 2017, which divides trip data according to corresponding taxi zones. However, different portions of data are used for each analysis and will be covered later in this report. The main visualization libraries I used are GeoPandas and Matplotlib.

3 Exploration

In this section, I will cover the investigations I went through and accosiated pre-processing, cleansing, visualization and find of analysis for each investigation.

3.1 Trip Count by Pick-up Location

I first investigated the number of pick-ups by taxi zones for 2017. I randomly selected 100K samples from each month in 2017 (1.2M in total) since it would be time-consuming and memory-consuming to visualize the entire year of 2017. And this also helps reduce bias from individual month. Now I visualized trip count by aggregating trip instances based on pick-up locations (PULocationID) and mapping to respective taxi zones.

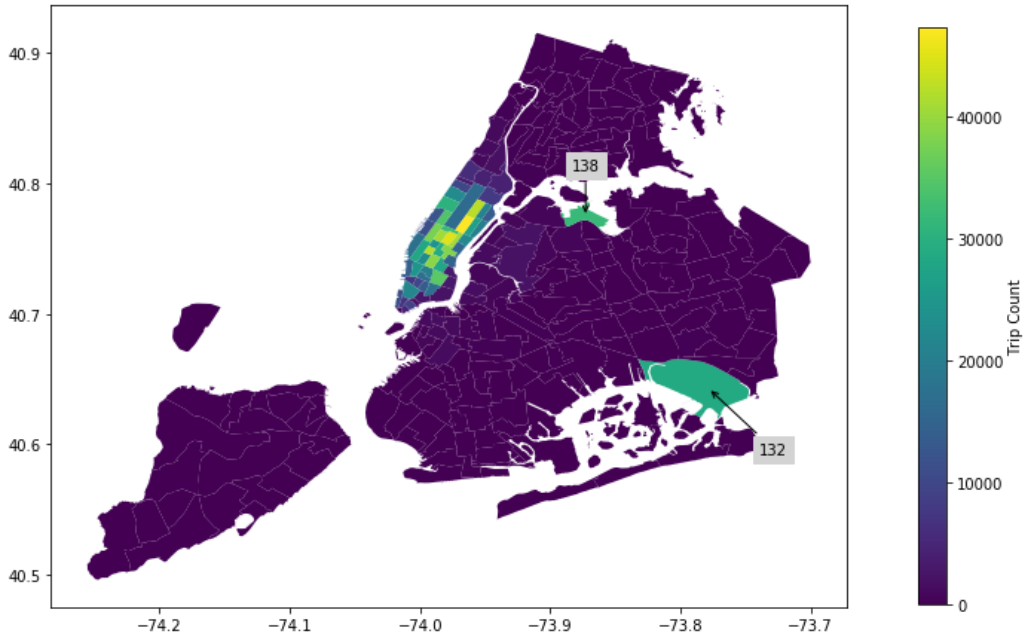

Figure 1: Pickups in 2017 Taxi Zone

Majority of the areas in Staten Island, Brooklyn, Queens and Bronx are deep purple which represent zones with trip count less than 10,000. Significant number of trips took place at central and lower western Manhattan, with trip

count in most of these zones more than 20,000. Also, Zone 132 and 138 in Queens are also distinguishable. We now zoom into these regions separately and have a closer look.

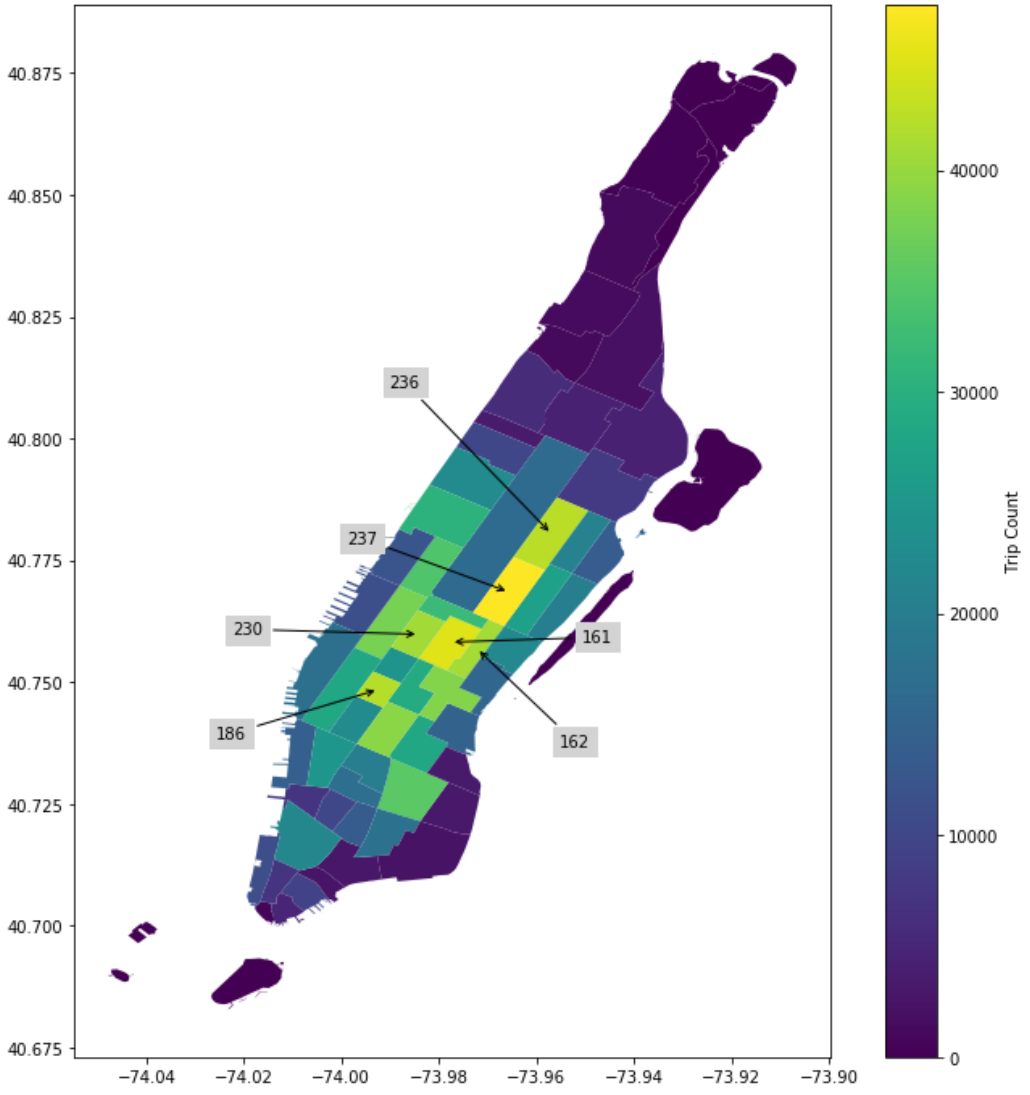

Figure 2: Pickups in Manhattan

As we can see, Most pickups took place at the center and lower west of manhattan, especially along Midtown Manhattan and Upper East Side. The most busy zones are 161 (Midtown Center), 236(Upper East Side North) and

237 (Upper East Side South) and 186 (Penn Station/Madison Sq West) with pick-up count above 42,000. These areas are some of New York City's largest and busiest business districts and hence we should not be surprised that large number of rides took place in these areas. Now we move on to Queens.

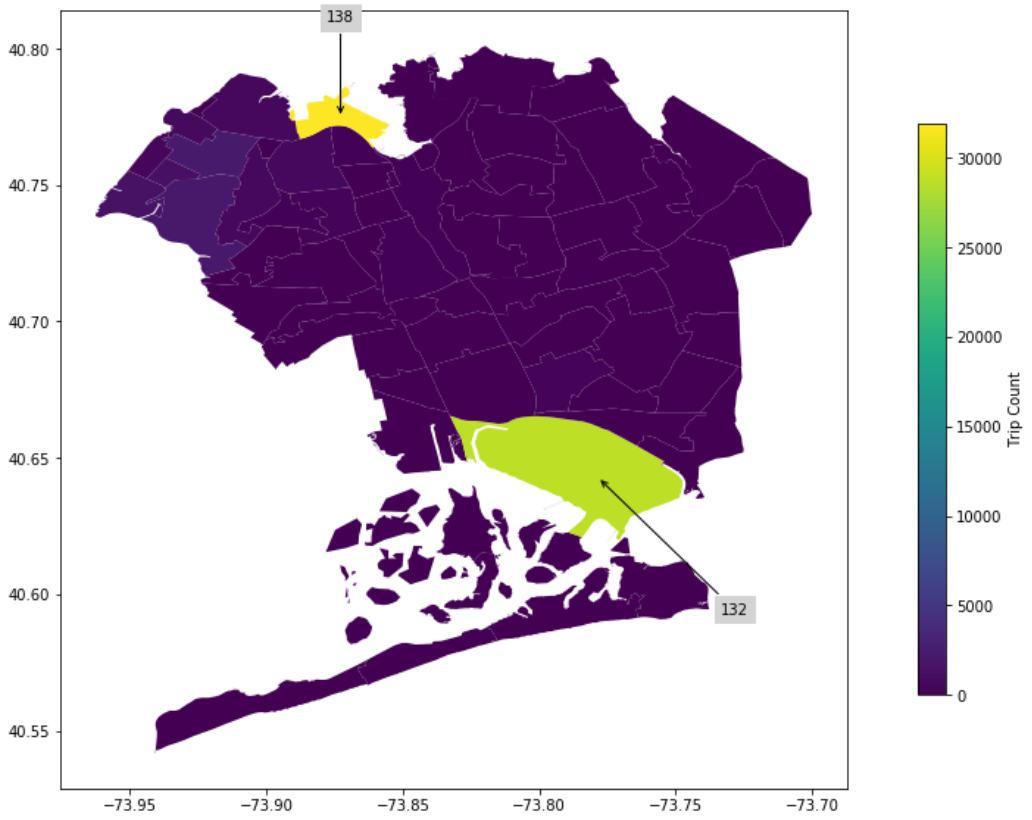

Figure 3: Pickups in Queens

Zone 138 and 132 have seen large amount of pickups (32,100 for 138 and 29,184 for 132). This is most likely because 138 lies in laguardia airport and 132 lies in John F. Kennedy International Airport. Therefore, demand for taxi is high here as people who get off planes take taxi to their destinations.

3.2 Weekdays vs Weekends

In this section, I compared pickup counts between weekdays and weekends by taxi zones. Again, I use 100k monthly sampled data from last section

but split the dataset into weekdays and weekends. To ensure there is equal number of days selected from weekdays and weekends, I randomly selected 100k trips from each period.

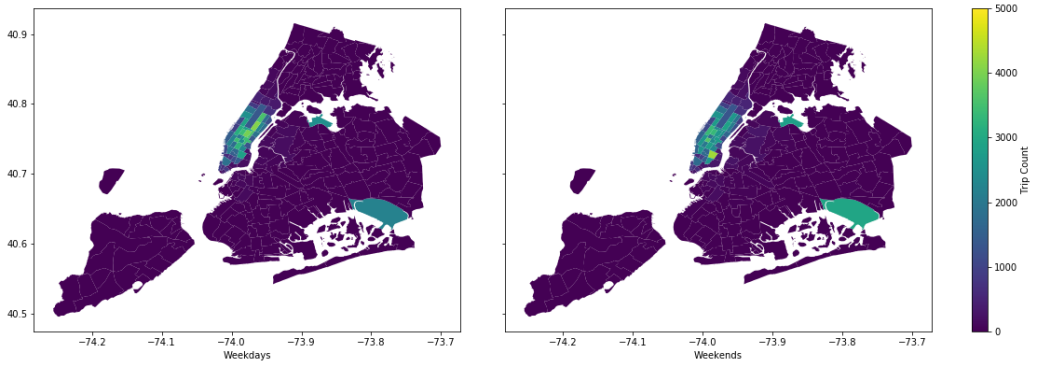

Figure 4: Pick-up Count Weekdays vs Weekends

The figure shows that, again, most trips took place at center and lower west portion of Manhattan and two airports in queens throughout the week. However, there are also differences between the two period. At the airports, it appears that more trips were taken during weekends than weekdays. This could be because more people perfer to travel on the weekends. Also, Manhattan is apparently more busy during weekdays. I now zoom in for more details.

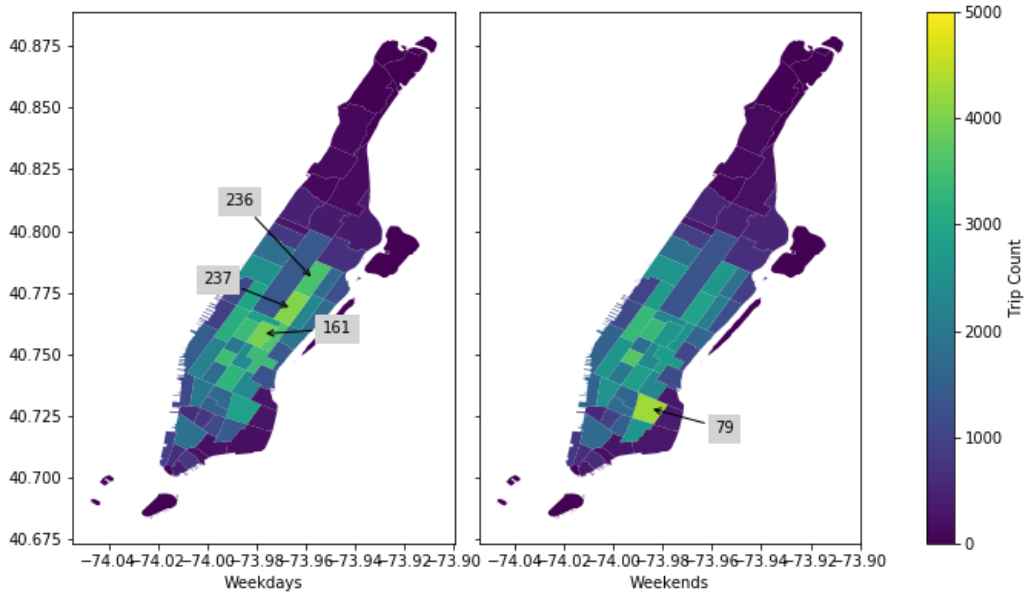

Figure 5: Pick-up Count Weekdays vs Weekends in Manhattan

As we can see, more pick-ups took place at zone 236 (Upper East Side North), 237(Upper East Side South), 161(Midtown Center), central and lower Manhattan during weekdays as more people travel and work during weekdays.

However, zone 79 (East Village) has more pickups during weekends. This is most likely because East village is known for for its happening nightlife, cool bars, music venues and performance spaces [3]. It is therefore no surprise that more people come here for a good time during the weekends.

3.3 Rate of tip per minute by Taxi Zone

Next, I investigated rate of tip per minute (RTM) by taxi zones according to pick-up locations. If both trip A and B incur a 5 dollar tip but trip A takes 5 minutes and trip B takes an hour, it is certainly more profitable for drivers to choose trip A. Therefore, RTM helps taxi driver to decide if certain trips is more profitable than others by choosing the right zones to drive in.

I again used the 1.2M monthly data from 2017 but only keep trips that are paid by credit (since cash payments do not incur tips). Then I calcualte

RTM for each zone as such:

$$RTM = \frac{\text{mean tip amount (dollar)}}{\text{mean trip duration (minute)}} \tag{1}$$

After calculating the RTM for each zone, I noticed that there are a few zones with RTM being abnormally high (greater than 3.0). Close inspection of trips taken place in these zones revealed that most of these trip have unusually short trip durations as compared to tip amount (One trip have a duration of 1 second and tip amount of 5 dollars). These could be results of human error/system malfunctions and therefore are omitted from visualization. Also, I omitted zones with RTM = 0 because trips paid with credit card in these zones have 0 tip amount. These trips would not give any additional information about relationship between tip and duration. As a result, some zones may be missing from the plot.

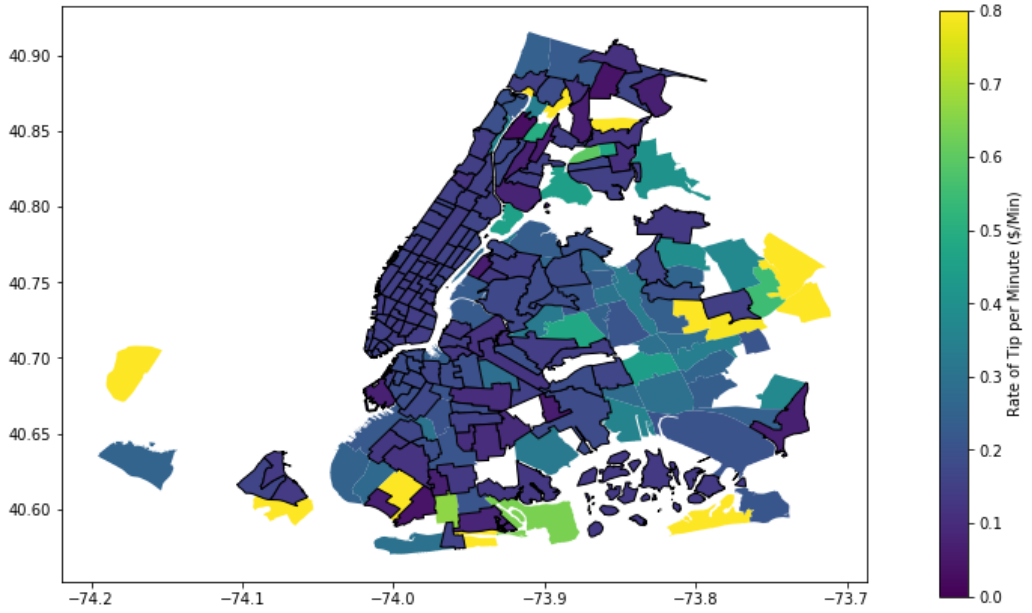

Figure 6: Rate of Tip Per Minute (RTM) by Taxi Zones

The figure shows that entirety of Manhattan, large portions of Brooklyn and Bronx have lower RTM. These zones are in deep purple with black outline and have RTM less than 0.17 dollar/min. On the other hand, Majority of

Queens and Staten islands have high RTM. They are mostly in light blue or yellow color and have RTM as high as 0.8 dollar/min. With this knowledge in mind, taxi drivers can choose to drive in areas with higher RTM for potentially earning more tips. However, we also need to understand the reason behind difference in RTM so that we can make prediction for regions with unknown RTM.

If a trip has high RTM, considering tip amount being fixed, this trip must have a longer duration. For a trip to have a longer duration, it could be due to longer trip distance or stuck in slow traffic. In another word, average trip distance may be shorter and traffic must be slower in Manhattan, Brooklyn and Bronx. I tried to justify my theory by visualizing average trip distance and average taxi speed.

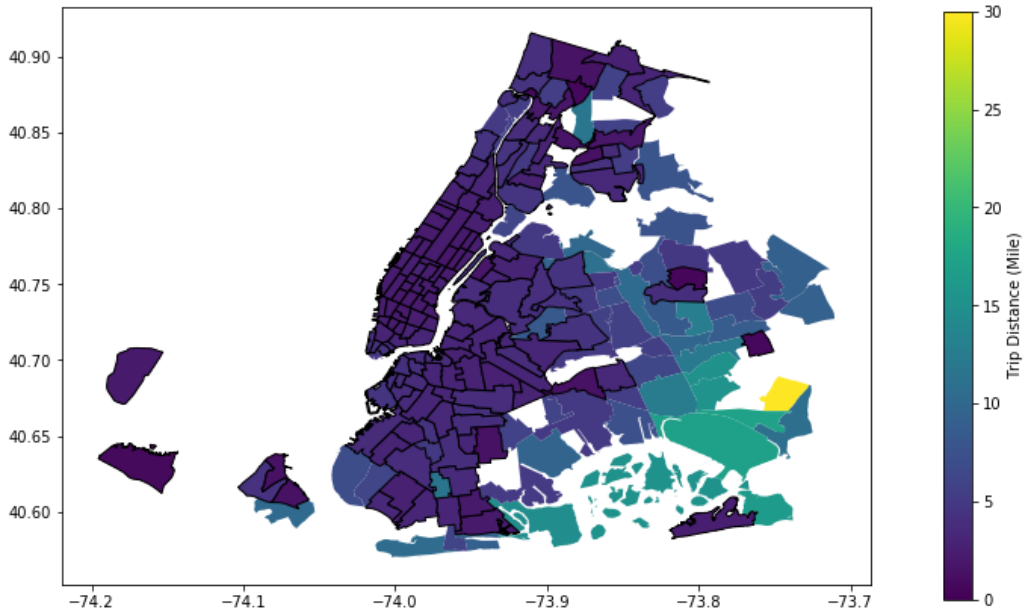

Figure 7: Trip Distance By Taxi Zones

As we can see, the distribution of trip distance is similar to the distribution of RTM. Trips took place in majority of Manhattan, Brooklyn and Bronx are shorter than the other regions. These areas are outlined in black and have average trip distance less than 4.6 miles.

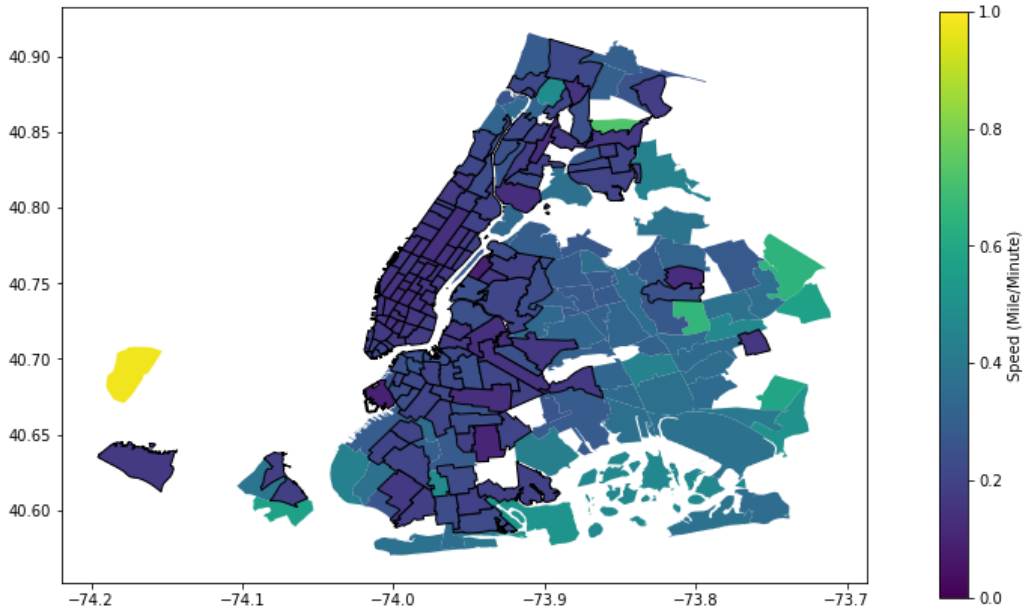

Figure 8: Average Speed By Taxi Zones

As we can see, the the distribution of average speed is also similar to RTM. Trips took place in majority of Manhattan, Brooklyn and Bronx have average speed lower than the other regions. These areas are outlined in black and have average speed less than 0.25 miles/minute = 15.0 miles/hour.

Next time, if a driver wishes to earn more tip from driving for the same amount of time, he/she may choose to drive in an area where he/she can expect to drive for longer distance in faster traffic.

3.4 The Effect of Rain on Ridership

In this final section, I investigated the effect of rain on trip count and RTM by comparing both of the attributes on rainy and dry days of Manhattan. I randomly selected 2 groups of 9 days from 2017. The rain day group consists of days when there was rain level ranging from light rain to heavy rain during the day. The dry day group consists of days when there was no rain.

I also ensure that the chosen days come from the same day of the week in the same month to reduce variance from individual days. Also, I only

looked at Manhattan for this investigation as weather data was recorded at Central Park[1].

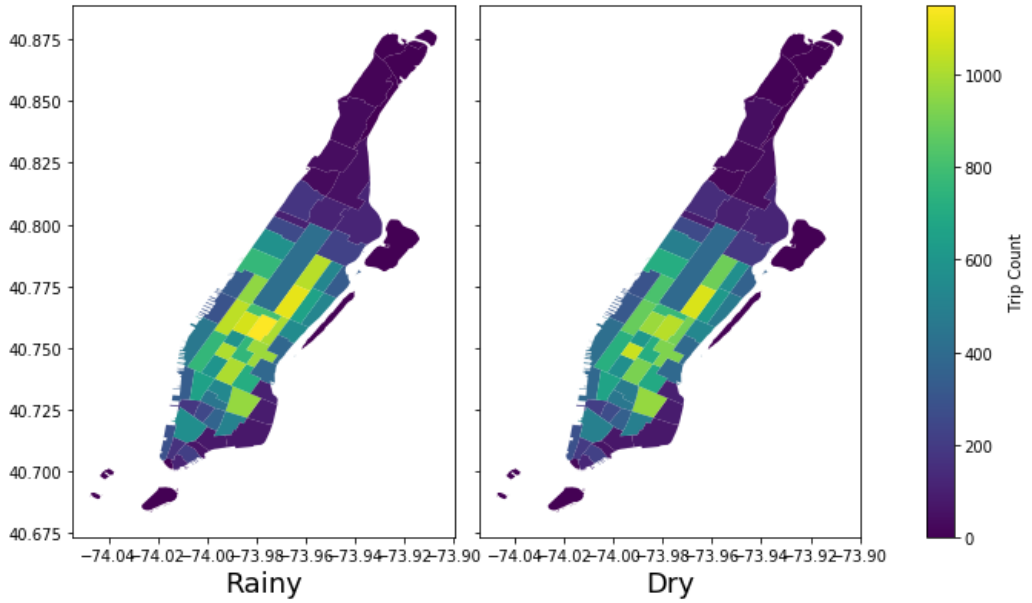

Figure 9: Pick-up Count Rainy Days vs Dry days

The figure shows that pick-up count generally shows the same distribution on both groups of days but the difference are not easy to see. In order to see the different more clearly, I constructed a contrast plot which only shows the difference in trip counts.

[1] `https://www.timeanddate.com/weather/usa/new-york/historic?month=10&year=2017`

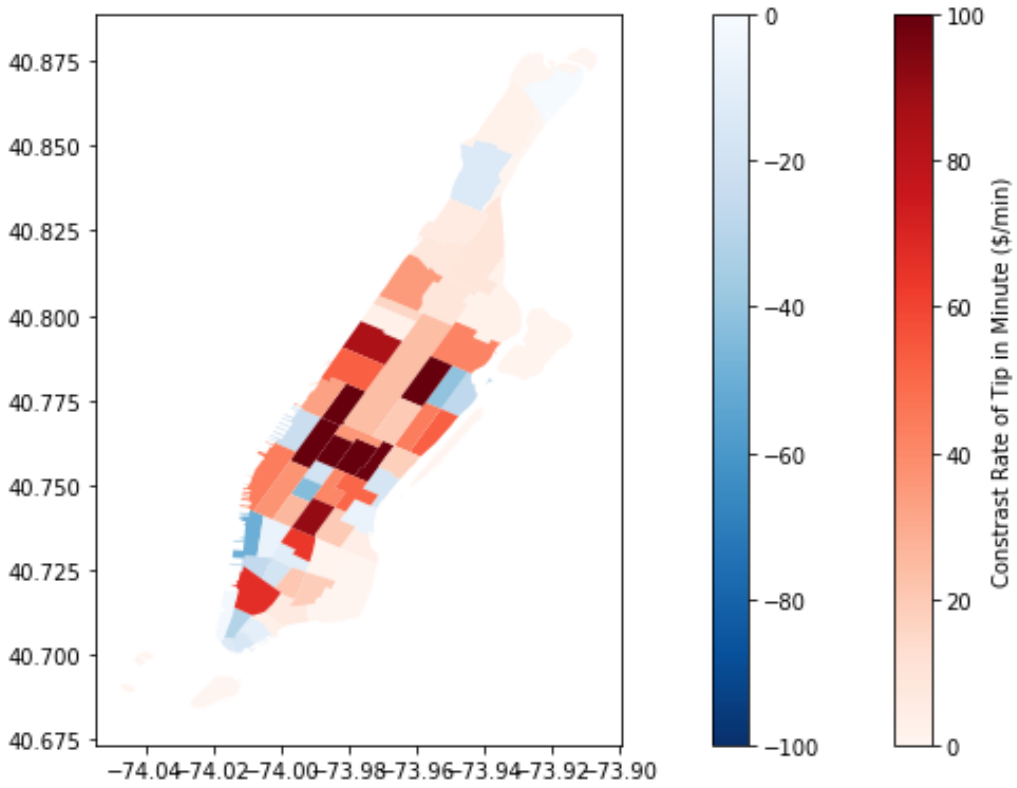

Figure 10: Pick-up Count Difference

In this contrast plot, red zones are zones with trip count higher in rainy days and blue zones are the ones with trip count higher in dry days. The darker the color, the larger the contrast. We can immediately see that majority of the zones have higher trip count in rainy days (41 out of 69), especially along Central Park and Midtown Manhattan. This could be due to higher demand for rides on rainy days. Therefore, it seems that increasing taxi supply on rainy days can be a profitable strategy.

However, although it seems that more trips are taken on rainy days, is it actually more profitable for the drivers? To answer this questioin, I look at Rate of Tip Per Minute (RTM) for the two groups. I will only show contrast plot so that the difference in RTM can be seen more easily.

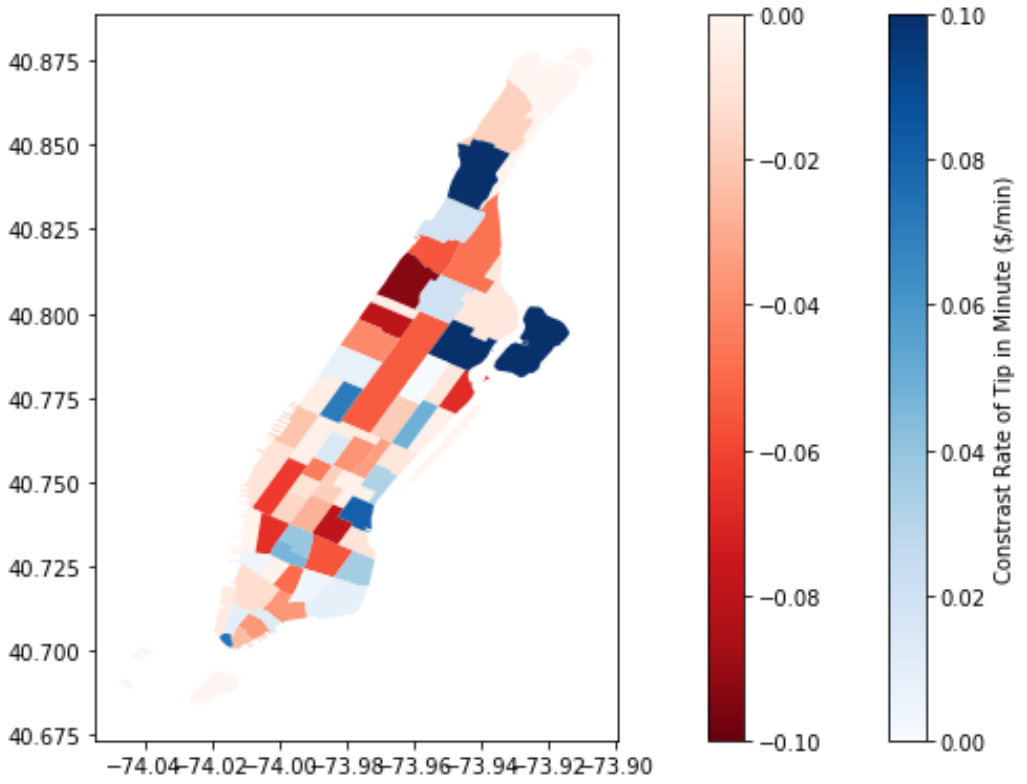

Figure 11: Contrast Plot for RTM Difference by Taxi Zones

The plot shows the numeric difference in RTM between rainy and dry days. Red zones are zones with lower RTM in rainy days and blue zones are the ones with higher RTM on rainy days. As we can see, majority of the zones (40 out of 69) are red hence have lower RTM in rainy days. This could be due to slow traffic as rain could slow down traffic. (Assume there is little difference in trip distance between rainy and dry days)

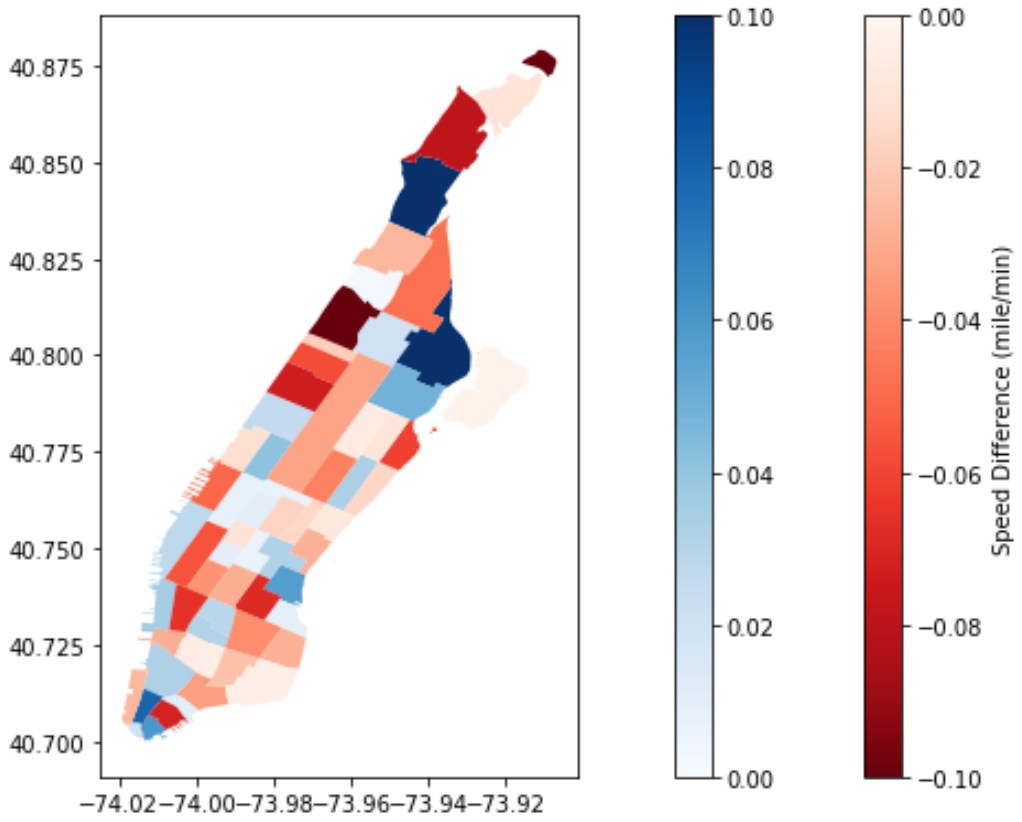

Figure 12: Contrast Plot for Average Speed Difference by Taxi Zones

As expected, this plot follows approximately same distribution as the contrast plot for RTM. Average speed in majority of zones is lower during rainy days (as indicated by red region),

Next time, a taxi driver may prefer to drive in rainy days because of high demand. But he/she must also be aware that it is may not be entirely profitable since tip earned per minute has actually gone down due to slower traffic.

4 Conclusion

In this project, I explored trip count and rate of tip based on taxi zones, compared pick-up counts between weekdays and weekends and analyse trip count and rate of tip per minute by area due to rain. I discovered that trips

are more frequently taken at center of Manhatan and airports in general, and are more often near CBD during weekdays whereas more often close to entertaining area such as East Village during weekends. I also learnt that rate of tip per minute can be affected by trip distance traffic condition. Also, rain is an important factor that drives demand for taxi. But drivers also need to be aware of lower rate of tip per minute on rainy days because of slow traffic. As for future work, I may compare data from other years and explore impact of other factors such as major events on ridership.

References

[1] A. Sraders, "Uber vs. taxi: What's the difference?," 2019.

[2] "Is uber better than taxis?," Feb 2016.

[3] L. Boulter, "'the east village is one of the few places in manhattan hanging on to some character'," Jan 2017.