Yellow Cabs and Weather

Emily Padvorac[1]

[1]NYU Center for Urban Science & Progress

December 15, 2017

Yellow Cabs and Weather <Emily Padvorac, ep2247, ep2247>

Abstract:

This project examined the relationship between yellow taxi cabs and weather data. The purpose is to see if shorter taxi trips were taken when adverse weather conditions happen. Previous work has examined the relationship between taxi ridership and weather conditions, mostly with the conclusion that during adverse weather, taxi drivers make more money. This research will further investigate the relationship between taxi ridership and weather to see if any further results can be determined.

Introduction:

This research seeks to explore the relationship between taxi ridership and weather in New York City. The goal is to understand if during events of adverse weather, more people are prone to take taxi trips for a short distance. This analysis could prove helpful to taxi companies and city planners as it can shed light on when a higher volume of taxi traffic and ridership would be expected.

Previous work has sought to look at relationships between taxi ridership and weather events. "During adverse weather, taxi drivers tend to make more money" (Kamga et al., 2013). Kamga et al. also found that during days when adverse weather took place, there was a higher demand for taxis in Manhattan compared to the other boroughs, and that more people were prone to take shorter taxi trips.

To do this, data will be used from the hottest and coldest month on average during the year of 2016. The hypothesis is that people are more prone to take shorter taxi trips on adverse weather days, compared to days when there is no adverse weather taking place. In this research, adverse weather is defined as days were the temperature was below 35 degrees, or above 90 degrees, or days when rainfall was present.

Data:

Weather Data

The months used in this study were determined by New York City's climate. Climate data was viewed from the National Weather Service, New York City office online to obtain the hottest and coldest monthly average temperatures, as measured by the Central Park weather station for the year of 2016.

National Climatic Data Center archives climate and historical weather data. These data are archived from FAA operated weather stations located in various areas around the country. KNYC (the Central Park Station) was selected for this study the since the only other two stations available were in outlying New York City Boroughs (LaGuardia and JFK). Since central park was located in the city, it was determined to be the most representative of the weather conditions in the immediate New York City area. The weather station from which the data was fetched had data available in hourly, daily and monthly time scales. The daily observations were chosen since hourly data was too precise for the scope of this study and the taxi

data was formatted into a daily time scale. For this research, daily summary reports were obtained for the months of study. In this dataset, the daily maximum dry bulb temperature (daily high temperature), daily precipitation amount, and daily snowfall amount were the variables looked at.

Taxi Data

New York City Taxi & Limousine Commission(TLC) issues annual taxi ridership data. Taxi data was available for every month for multiple years, but 2016 was the most recently available data and was chosen for this reason. For this analysis, yellow taxi cabs were considered. Data was obtained for January 2016, and August 2016, the coldest and hottest months of the year respectively. This data included a variety of variables, pick up and drop off times, passenger count, etc. For the purpose of this research, trip distance was the main variable looked at from this dataset, since this directly relates to the stated hypothesis. The data was filtered into "short" and "long" distances, with 2 miles being the delimiter between the two.

Taxi Zone Shapefile

New York City TLC also issues a shapefile that contains all of the taxi zones in all five boroughs. This shapefile was used to plot the pickup locations for August 2016.

Data weaknesses

Weaknesses and limitations were found in the taxi data. The January 2016 data, did not have any information for the pickup location ID, and the drop off location ID's. Due to this, the research does not allow for the use of geopandas in order to plot the pickup locations for taxi data for the month of January. The August 2016 data, did not have any latitude or longitude information available on both the pickup and drop off locations.

Data Wrangling

Multiple merges were done in order to get the data in its proper form. Both taxi datasets, were filtered for trip distances less then 2 miles. The first merge came from merging the August taxi data with the taxi zone shapefile on the pickup locations, and location id. After this geopandas was used to plot the trip distances, as seen in figure 1. For the month of August, most of the trips that were less then 1 mile took place in Manhattan. Most of the trips that were between 1 and 2 miles took place in Queens. Most of the pickup locations were located in Manhattan, as previous research discussed. Geopandas was also used to plot the trip distances in January, as seen in figure 2. Like August, most of the trip distances less then 1 mile took place in August.

Next, data was merged on dates between the taxi datasets and the weather datasets to create new data frame. After the multiple line and bar plots were plotted to see the variation in temperature, precipitation, and snow (for the month of January) for both months (figures 3,4,5,6). In this analysis, snow was accounted for in precipitation totals for the day. These plots determined that our definition for adequate weather in this study is temperatures that are greater than 90 degrees, or less then 35 degrees, and precipitation was defined as anything greater than 0.1 inches. After defining the threshold, datasets were merged into 4 new dataset's to include the days of study.

Data was also filtered for trip distances that were greater than 2 miles. The data for the long taxi trip distances was also filtered to the adequate weather datasets.

Methodology

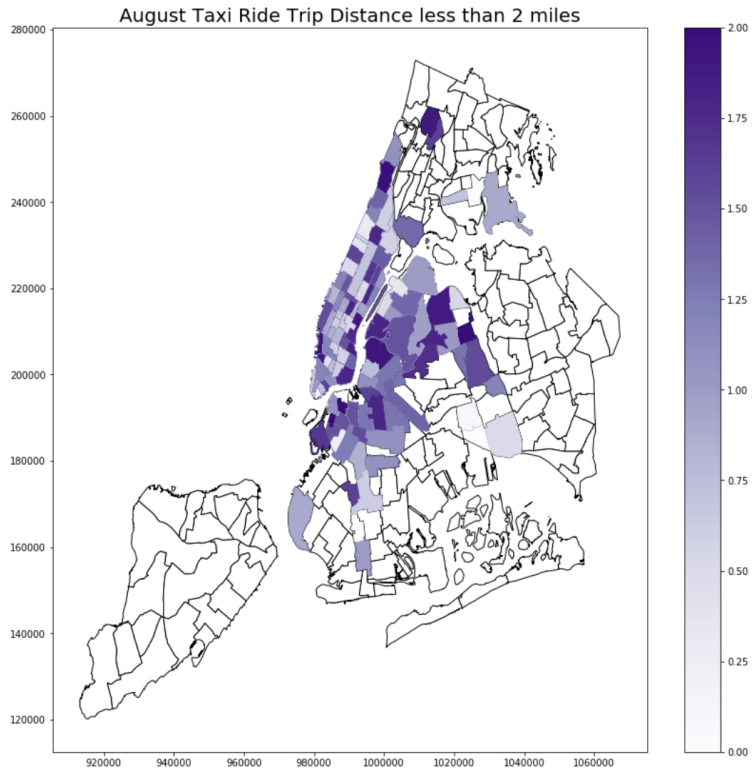

Figure 1: Short Trip Distances plotted for the month of August.

Figure 1: Short taxi trip distances less than 2 miles plotted for the month of August 2016

In this analysis, multiple linear regressions and multivariate regressions were performed on all of the variables for the two considered months. An alpha of 0.05 was set to define our significant threshold. Other methods that should have been considered for this analysis are logistic regression, and clustering.

The first set of linear regressions was run between adverse weather for the two months for short trips only. The first regression was between August temperature and August precipitation, with temperature being the dependent variable and precipitation being the regressor. The R-squared value for this linear regression was 0.305. The next set of linear regression was run between the adverse weather days for January temperature and January precipitation. In this linear regression, temperature was the dependent variable, and precipitation was the regressor. The R-squared value was 0.203, this was slightly less than the R-squared value for the month of August.

The next set of linear regressions were run between long trips vs. short trips on the adverse weather days for the two months. Running the linear regression between short distances (less than 2 miles) and long distances (greater than 2 miles) for both January and August, we get an R-squared value of 0. There is no correlation between the two variables for either of the months being studied.

Linear regressions were also performed for both long and short distances seeing if either temperature or precipitation had any effect on who takes taxis. For the month of August, running the regression between long distance and temperature, an R-square value is 0.001, in January the R-square was 0. Regression for short distance and temperature in both August and January, the R-square is 0.

Multivariate regression for both months was performed with the dependent variable being short distance and the regressors being temperature and precipitation. In January a R-square value of 0 was found, and

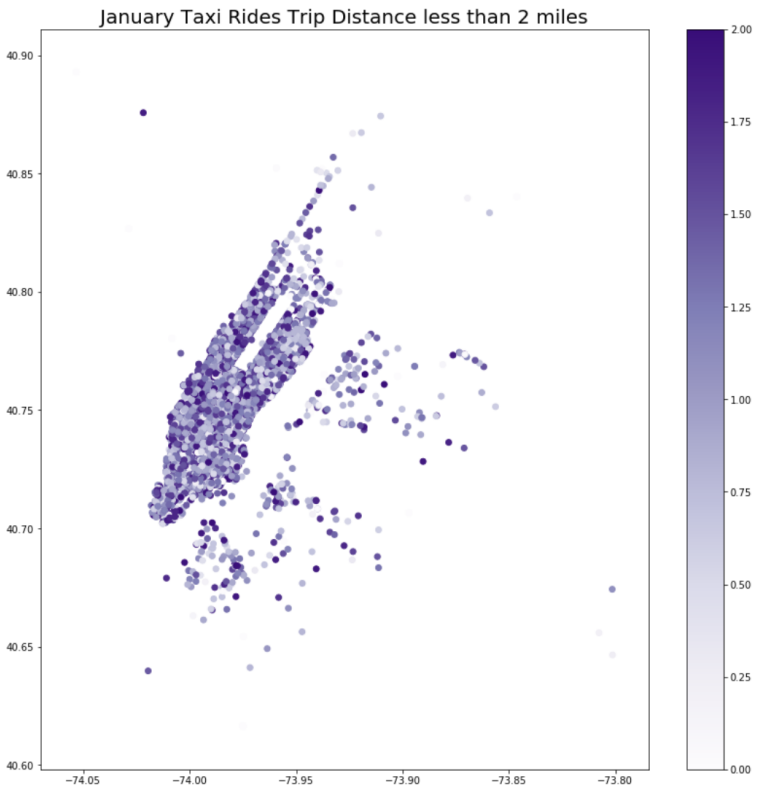

Figure 2: Short taxi trip distances less than 2 miles plotted for the month of January 2016

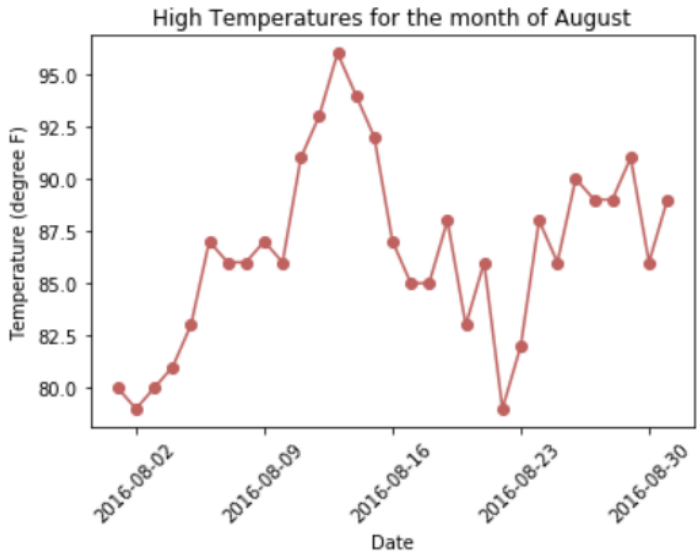

Figure 3: Daily high temperatures (degree F.) for the month of August 2016

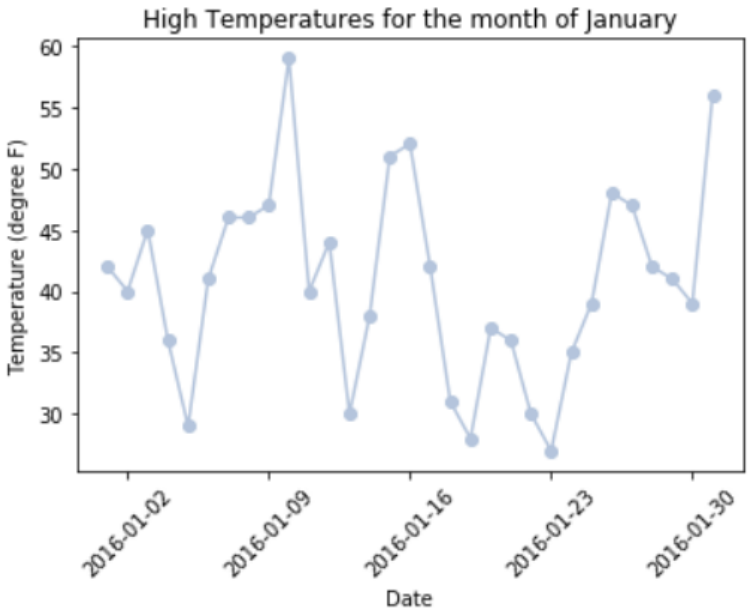

Figure 4: Daily high temperatures (degree F.) for the month of January 2016

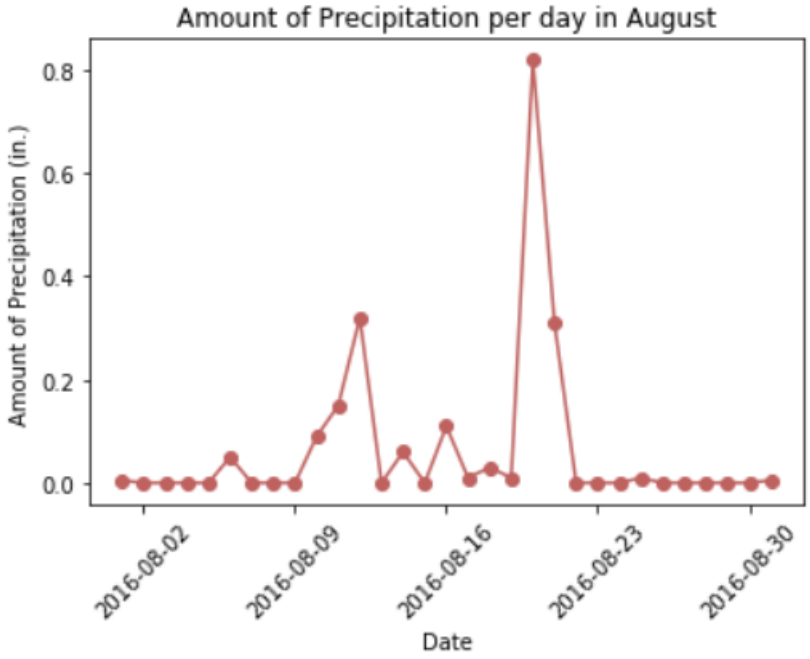

Figure 5: Daily precipitation amounts (in inches) for the month of August 2016

in August a R-square value of 0.002 was found.

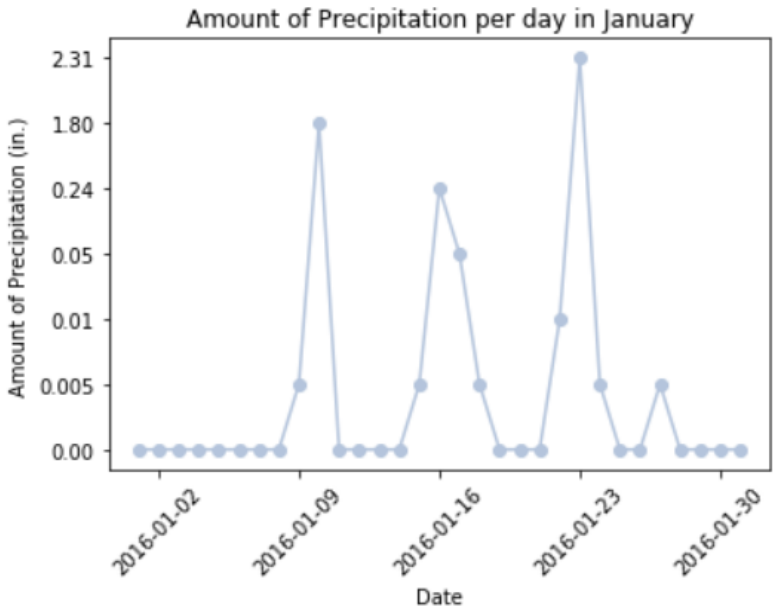

Figure 6: Daily precipitation amounts (in inches) for the month of January 2016

Conclusion:

For this analysis, models showed no indication that there is any significance or correlation between short taxi distances and temperature and precipitation during adverse weather days. Starting this project, the hypothesis assumed that there would be some relationship on people taking taxi's on adverse weather days, but there ended up not being any.

Future work:

For future studies, we can consider more variables in the datasets to improve our analysis. One weather variable to consider adding to the adverse weather condition would be wind speed. For the taxi data, it would be helpful to explore in the future if maybe there is an increase in passenger ridership during adverse weather conditions, or if people pay more money to take shorter trips. For short distances, more people in Manhattan were prone to take taxis so it would be helpful to study only Manhattan. It also might be useful to look at Uber and Lyft data to see if more people are prone to take an Uber or Lyft during adverse weather conditions. It would also be good to consider income data, however due to the fact that people don't stay in their homes during the day due to work, school, other activities it might not be best to use that data.

Links:

https://github.com/ep2247/PUI2017_ep2247/tree/master/Extra%20Credit%20Project

References:

Kamga, C., Yazici, M., Singhai, A. 2013. Hailing in the Rain: Temporal and Weather-Related Variations in Taxi Ridership and Taxi Demand- Supply Equilibrium. ResearchGate

https://www.researchgate.net/publication/255982467

Links: links to code and data (in the spirit of reproducibility i should be able to reproduce identically all plots you include using your code and your data) must be uploaded within authorea as data and notebooks embedded in the appropriate figures like this https://www.authorea.com/users/9932/arti how-to-insert-jupyter-and-ipython-notebooks-in-authorea-articles

Bibliography