Subway and Taxi Usage in New York City and Chicago

Transportation-Transformation

Stephanie Carrero (scarrero), Holly Zheng (yzheng40), Becky Mathew (rmathew7)

Hypothesis

We investigate the use of public transportation vs. private ridership. New York City and Chicago are our two cities of focus, and we examined subway and taxi usage within each city as well as across the two cities. The factors that we thought would have a significant influence on the use of taxi are:

- number of subway stations in the zip code -- this potentially reflects the availability of one type of public transportation
- number of yearly subway rides total in the zip code -- this potentially reflects a general habit of utilizing public transportation among people of the area
- population and income of the zip code -- demographic information

We hypothesize that more taxi rides in a zip code is the result of fewer subway stations, lower subway rides, higher population, and higher income of the zip code.

Data

We collected two datasets with information on yearly taxi rides, yearly subway rides, population, income, and number of subway stations for 88 zip codes in New York City and 31 zip codes in Chicago. The data also includes whether the taxi record is a pickup or dropoff. All sources come from government sites.

Main efforts in preparing the data include matching a location's latitude/longitude to nearest zip code based on Euclidean distance, resolving mismatched names of subway stations (eg. "Times Square" vs. "Times Sq"), and scaling data to the same units (eg. millions vs thousands). From exploratory histograms, we concluded that our data for both cities is not skewed.

Findings

1. Response to our initial hypothesis

An increase in the number of subway rides in a zip code corresponds to an increase in taxi pickups and dropoffs in that zip code. In New York City, an increase in income of a zip code corresponds to an increase in taxi pickups and dropoffs in that zip code.

The subway rides coefficient is positive and statistically significant in the multiple regressions for the combined data ($p=0.000$), for the full Chicago pickup data ($p = 0.008$), for the full Chicago dropoff data ($p = 0.010$) for the scaled NYC pickup data ($p=0.023$), and for the scaled NYC dropoff data ($p=0.023$).

In NYC alone, the income coefficient is positive and statistically significant for the pickup data and the dropoff data (p=0.000 for both regressions).

We investigated 5 Chicago zip codes that have significantly higher taxi ride numbers than other zip codes. These outliers, however, do not affect our claim from the above paragraph -- they do not change the statistical significance or direction of effect of the independent variables on taxi ride usage in Chicago.

After we found that one of the Chicago outlier zip codes is the region where the O'Hare airport is, we incorporated AARP livability index of transportation for each zip code as an indication of the residential condition of each zip code area. These indices were not statistically significant to the regression (p=0.565), which means for future work we need more fine-grained descriptions of zip codes, such as availability of tourist attractions.

2. Follow-up on comparison between NYC and Chicago

The city that a zip code belongs to does not influence the relations described in Claim #1. The relationship between taxi and subway usage behaves similarly across New York City and Chicago.

We combined the datasets for Chicago and New York pickups and dropoffs. We added a binary variable that indicated if a particular row was for a pickup and a binary variable that indicated if a particular row was from Chicago data. The p-value for the city binary variable (is_chicago) is p=0.715. Therefore, we fail to accept the hypothesis that the city a zip code belongs to (Chicago or New York) has a statistically significant relationship with the number of taxi rides in a zip code.

Interaction terms on which city it is and its subway rides, population, and income are also not significant in affecting the regression results. In multiple regression including these interaction terms, the p-values on these variables are all higher than 0.05.

Conclusion

Our hypothesis that higher income corresponds to greater taxi usage is supported by our investigation. However, contrary to our expectations, greater subway usage and lower population correspond to greater taxi usage across both cities.