CSCI 6350 Data Science Assignment 4: Examining the Effects of Weather on Taxi Service in New York City

Stephen Notley, Diana Edwards, Ian Gross, Jianhui Lu, & Animesh Tripathy

12/1/2017

1. The team chose to examine the relationship between weather and taxi usage statistics in New York City in order to better understand how weather effects taxi service demand and performance. The specific questions posed will be elaborated on in greater detail in Part 2 below.

   (a) The team chose the 2016 Taxi data because of its convenient and detailed information about pickup and dropoff times and locations (in longitude and latitude), number of passengers, and trip duration in seconds, which will allow for many interesting relationships to be explored. This data was sourced from a Kaggle competition two months prior to this paper with the challenge for the data science community to propose a better way to predict ride times.

   The weather data was sourced from Weather Underground (page found here), which has historical weather data available for New York City, measured at JFK International Airport, which contains daily metrics of temperature, wind speed, visibility, and conditions, among other stats. This was obtained by design and use of a scraping script to gather the data from the monthly pages throughout 2016.

   The size of both of these datasets were not so large as to prevent team members from storing them on local machines, and so the data was managed that way, with one clean, ground-truth version of the datasets available on a shared Google Drive.

   (b) The taxi data came in a zipped csv file, with each row representing one trip and each column representing various statistics about each trip. The metadata here was simply a plain text pairing system of column title and an explanation of what it meant in plain English with units. The weather data was available in a web page, from which a script was used to scrape the JavaScript into a csv file, which was labeled with appropriate column names as metadata. The only additional piece of metadata collected on the weather data was the location of its collection: JFK International Airport. The team also decided to aggregate the unwieldy individual taxi trip data into a statistics csv file, which consisted of a datapoint for each day, giving the total number of taxi passengers that day, the average number of passengers per ride, the time per unit distance of the trip, and average trip duration. The team thought that these aggregated statistics could be used to see meaningful correlations during analysis.

   It should also be noted that the "unit distance" here was calculated from the difference in latitude and difference in longitude for each taxi pickup and dropoff. The absolute values of these differences were summed in order to simulate the traversal of city blocks. It should be noted that this may not entirely accurately depict actual distance traveled, but given the grid-system nature of New York City, the team judged that this would provide a sufficient approximation for the purposes of this study.

2. (a) The two questions that the group hoped to answer using these datasets were:
      i. How do various weather conditions affect taxi trip durations in New York City?
      ii. How do various weather conditions affect the amount of people using taxis in New York City?

   (b)

   (c)

3. (a)

(b)

(c)

4.