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 hourly metrics of temperature, wind speed, visibility, and conditions, among other stats. This was obtained by design and use of a scraping application to gather the data from the daily pages throughout 2016.

   The weather data was sourced from a 2016 NYC weather data file from Kaggle, which was created for the same challenge as the Taxi data, (but is a separate dataset, simply over the same time-range) and which can be found here. This dataset contains dates with temperature ranges and average, precipitation and snow-fall/snow-depth stats, which the team thought would influence commuting conditions and could potentially lead to interesting results.

   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 dataset 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.

2. (a)

   (b)

   (c)

3. (a)

   (b)

   (c)

4.