Gene Hong and Chris Au

7/26/20

CS 301

Weather Impact on NYC Taxi Fare

For our code we started by importing all the files we would need such as pandas, numpy, seaborn and matplotlib.pyplot. We used the centralparkweather.csv data from the professor's link provided, and combined it with the train.csv data to make a combined data csv to show the fare price given the weather. We used two functions; one to computer distance in km by using the haversine function and multiplying it by Earth's radius, another to show the time of day. We adjusted the data to fit our needs by dropping non existing data, minimum and maximum fares, and observations that don't make sense. We divided up the longitude and latitude of new york by a grid of 100 by 100 rectangles and organized them into dropoff and pickup places in the grid. After making a histogram to visualize the densities of pickup and dropoff points, we made a subsection of the points to zoom in on more data. After we found the 100 most densely populated pairs in the grid by sorting through the occurrences and converting the values into a list. We found the observations that fit into these pairs and used them for analysis. There were some discrepancies for heavy snowfall and rainfall, so if there was snowfall accumulating to 6" or more in 24 hours or less, that would be heavy snowfall. Heavy rain would be more than 0.3 inches of rain per hour and since it is over a 24 hour period, we multiplied that value by 24 for it to suffice as heavy rain. Since distance is a factor, we segmented the observations by distance and tried to find the average fare price for each distance. As a result there are 2 different lists for

each geographic zone and thus we created a dataframe where the columns are the percent differences due to rain and snow and the rows are the 100 dominant zones.