Capstone Project 1: Data Wrangling

Weather Data Cleaning

The 2016 NYC weather data was originally acquired from the Wunderkind API, and provided here.
https://www.kaggle.com/meinertsen/new-york-city-taxi-trip-hourly-weather-data

First we dropped all the columns that we do not think that we'll use.

As we check the data, we notice that it's fairly well organized. The rain and snow columns are are separate boolean columns. The data would be easier to read if it was a single column with a categorical values, so we combine the two columns into one precip_type column with values 'rain', 'snow', or an empty string for clear weather.

From examining the weather data, we notice something is off with the temperatures for certain days. Plotting the daily average temperatures for every month, we quickly notice that the for the days before the 12th of the month, the average temperatures rise and fall as you would expect throughout the year, and after the 12th of the month the average temperatures spread out and differ similar to how you expect them to differ between the various months. From this we can conclude that for the days that are before the 12, the month and day have been accidentally reversed in the data. After correcting for this, we plot the average temperatures again and we see that it now somewhat behaves how we would expect.

Merging Weather and Taxi Data

Now that the dates in the weather data are correct, we can merge the weather and the taxi data on their pickup times using merge_asof, matching each taxi ride pickup time with its nearest weather recording.

Sampling

Since our data set is fairly large with over 60 million rows, due to hardware limitations we would need to take a sample to work with. With the goal of analyzing if the weather affects the tip amounts that taxi riders give, we should take a stratified sample of taxi rides in the rain, in the snow, and in clear weather. Using 10% of the original data would still give us 6 million samples, which seems reasonable.

Data Cleaning

With a merged data set, now we can start to clean it up. The first thing we should do is calculate the time of each taxi trip using the pick up and drop off datetimes. Once we do this we

no longer need the drop off pickup time columns, so we can drop that. Now with the trip time info, we can calculate the average speed of each trip using the time and the distance. Using the average speed, we can also eliminate speeds that are either very low or very high and that do not make sense. Lastly we should drop the rest of the columns that don't appear useful. The thunder, hail, and tornado columns have only 0 values so we should drop those. Now our data is ready to be explored.