Capstone 1 – Data Wrangling

For my analysis, I have chosen 3 datasets to work with, each of which will require some degree of data wrangling.

1. **NYC Taxi Trips 2016** – A time-series dataset from Kaggle containing 1.5 million rows detailing yellow cab transactions that occurred in 2016. Columns include date/time of pickup and drop-off, pickup/drop-off coordinates, passenger count and trip duration.
2. **NYC Weather 2016** – A time-series dataset from the National Oceanic and Atmospheric Administration containing daily weather observations from 7 NYC weather stations. Columns include station number, station name, coordinates, elevation, date, temperature, wind speed, precipitation, snow and other weather events.
3. **New York Neighborhood Map** – A geo-spatial dataset from Zillow outlining the boundaries of neighborhoods in all five boroughs of NYC as well as other municipalities in New York State.

In order to begin Exploratory Data Analysis, the first two datasets need to be cleaned and merged. The steps taken in this process are outlined below:

1. NYC Taxi Trips
   1. Imported as trip\_df. Since the dataset is large, I looped over the csv file in chunks of 10,000 and then concatenated the list of dataframes.
   2. Columns called “pickup\_datetime” and “dropoff\_datetime” were actually stored as strings. I needed to parse dates in order to merge with weather dataframe in a later step, so the pickup datetime strings were sliced to only include MM/DD/YYYY characters, then converted to datetimes and stored in a new column called ‘date’.
   3. The full pickup datetime and dropoff datetime strings were then converted to datetimes and stored as columns ‘pickup\_datetime’ and ‘dropoff\_datetime’.
   4. Trip duration was stored as an Int64, so I subtracted pickup\_datetime from dropoff\_datetime to get a TimeDelta and stored it as ‘trip\_duration’.
2. Weather
   1. Imported as weather.
   2. The date column was stored as strings, so these were converted to datetime.
   3. The data was reviewed for each weather station. Since some weather stations were missing metrics altogether (e.g. Central Park didn’t have average temperatures), I selected LaGuardia Airport as the weather station that would be used in conjunction with the taxi trip data. Additionally, this is the most centrally located station for all 5 boroughs.
   4. Weather data was merged with NYC Taxi data using a left join (taxi data was the left table) and the date as the key. This gives every taxi trip and associated weather pattern.
3. Merged dataset
   1. There are seven columns indicating the presence of different weather patterns using binary logic (1 for yes, NA for no). I iterated over these columns filling NA’s with zero’s and then converted them to Booleans using a dictionary mapping 1 to True and 0 to False.
   2. I listed columns that could be discarded (as they were either duplicates or no longer needed for analysis) and iterated over the list with a delete command.
   3. I renamed all remaining columns to make them more readily understandable. The columns for weather events were named by referencing the NOAA’s documentation on daily weather observations.
   4. Wrote the clean dataset to a csv called nyc\_taxi.csv

NOTE: The geo-spatial dataset was left separate and may need to be merged later, depending on the steps taken to visualize and classify pickup points. However, this is not known at the time of data wrangling.