Data Wrangling

The first step was to acquire the data from Kaggle. After retrieving and loading the data, it was essential to look at some of the content of the data such as the shape, column names and summary statistics. The original data consisted of 1458644 entries and 11 columns. The datatypes included objects, floats and integers. The data consisted of some useful columns, such as trip dates and times, pickup and drop-off coordinates, vendor-id, trip duration and passenger count. The summary statistic showed high variance for trip durations which it was analyzed in EDA.

After doing a quick inspection of the data, it was time to figure out the missing values per column since this is necessary to do before starting EDA. This time around the data did not have any missing values and there was no need for any imputation or dropping entries or complete columns. Since the goal is to solve a regression problem, the next step was to create dummy variables for the useful object datatypes. Due to most of the datatypes being floats, everything was converted to float to make data wrangling easier.

The data had two columns with dates so it was easier just to convert it to date time datatype and from the date columns split it into multiple columns with month, week day and hour of each trip. From here, the data was ready to do EDA and figure out any possible outliers or errors. So it was decided to continue data wrangling after understanding what the data was trying to tell.

EDA helped to understand the data and possible errors in the data were found. In data cleaning, the aim was to remove trips that were unlikely to happen, such as low and high trip duration times and short and long distance trips. To start cleaning the data, the first step was to tackle coordinate precision. Coordinate precision is given by the number of decimals so in this case, the decision was to use 5 decimal precision, which is worth up to 1.1 meter. All trips with a lower precision were dropped and all trips with a higher precision were rounded to five decimals.

After making sure all of the coordinates had the same precision, it was time to remove all the unlikely trips. For this part, there was a need to engineer some features that would give us a better idea of what was happening, such as direct distance (minimum distance from A to B) and speed of the trip. After doing feature engineering, everything was ready to start removing possible outliers and errors in the data.

The data had short trips that had a direct distance of 0 and trip duration of more than 60 seconds. Also, some of the trips showed unusual records showing trip durations of less than 5 minutes and really large distances with speeds of more than 100 mph.

Long trips was also a concern. Some of the entries recorded trip durations of more than 8 hours which is highly unlikely for a taxi ride, but also, the direct distance of the

trip was less than 120 miles. It is true that New York City can sometimes have traffic issues, but seems unrealistic to take 8 hours to get out of the city.

Finally, after dropping any suspicious data, the columns not useful for modeling were dropped, such as the complete date times, id and speed. This helped our modeling focus on the relevant features and perform better.