Data Wrangling Report

Firstly, I want to thank my mentor for his support. I couldn't complete this project without his support. My work in this project is divided into five sections. The first three for data wrangling for each data source, the fourth one for merging all the data together and the last one was for visualizations and insights.

The first data source was "twitter_archive_master.csv" file. I began my work assessing its data visually and programmatically. I realized that programmatic assessment is more focused than visual assessment. After assessment I have detected some problems like missing data in some columns like "expanded_url", "in_replay_to_status_id", "retweet_status_id" etc. I discovered latter after reading some tutorials online about Twitter api that some of these columns belongs to replays and retweets on Twitter so I dropped all these columns. Also, I removed rows that doesn't have NaN values for these columns. I removed records that have empty value at "expanded_url". Some of "expanded_url" values contained multiple links with comma separated there was duplication in this links. So, I removed duplicated links. There were some values at name column that aren't names so I handled that by replacing all values that don't have uppercase letter with "None" using regular expression. Columns (doggo, floofer,pupper ,puppo) represented a value rather than a category. So, I removed all and converted them to values for a new dog_stage column. There were some denominators that doesn't have value of 10 so I forced all values to 10. "timestamp" wasn't datetime data type. So, I changed it to datetime. Also, there is no need for tweet_id column to be an integer. So, I changed it to string.

After being satisfied with my work on the first dataset, I downloaded "image predictions dataset" programmatically. After assessment, data quality and tidiness issues were not apparent for this data set. Then I decided to use the third data source by using Twitter api to get "favorite_count" and "retweet_count". There was a lot of data that was not useful to us so I used "dataframe.filter" function to keep only these three columns "id" , " favorite_count" and "retweet_count".

After gathering and cleaning all the data individually. I should merge our data in a single data frame. So, I used "merge" function with "how=inner" parameter to drop all tweets that don't exist in all of our dataframes and to merge our data based on matching tweet ids between them. After that there was two columns have tweet_id values "id" and "tweet_id". I removed "id" column. Finally, it was necessary to save the resulting dataframe to a new csv file.