Wrangle Report

By Geo Niju Shanth G

In this data wrangling project I have worked with three different source of data. It was very challenging but finally able to achieve what I wanted. We have one file which is provided by udacity as csv file other source one is from a http link and the other through twitter API. Twitter archive data is provided in the form of csv files. The image predictions data is programmatically downloaded using requests python library file as a tsv file. Each tweet's JSON data is queried from Twitter API using Python's Tweepy library. Tweepy was really interesting where in we can analyse the data on a real time basis. In the case of this project we have pulled the data for corresponding tweet ids from which we can get the retweet and favourite counts.

Once the data from the three different sources are available the challenge was to clean the data for analysis. I first assessed the data for quality and tidiness issues. Once the issues were found and document the next step was to do the cleaning process. Datatype issues which were found were fixed by converting to appropriate datatypes. Wrong names were found in the dog name column which might have happened while extracting the name from the tweet. Those issues were fixed by replacing names which start with lowercase letters as 'None'. Records which were retweets were removed by removing records which have not null retweet status id column values. Underscore '_' between names were removed using replace function and names were captitalized using title function. Dog stage columns were combined into a single column. The final step was to combine all three datasets into a single file using merge function and doing inner join.

This project was very challenging especially extracting data from twitter API and the cleaning process and I have acquired enough skills to do data wrangling using python.