Wrangling Report

Udacity DAND: Wrangle and Analyze Data

By: Sahil Yadav

This project is the third project of the term 2 of the Data Analyst Nanodegree Program. My goal was to wrangle and analyse the data given to me based on the twitter handle @dogs_rates which is also known as WeRateDogs. Here I would like to report my wrangling efforts which basically consisted of three parts:

1) Gathering the Data.

2) Assessing the Data.

3) Tidying the Data.

1) Gathering:

The data was obtained from a single source which was the Udacity data server. I wasn't really interested in creating a twitter account because that would bring me to another social media platform and I'm on just too many of them right now.

I downloaded the twitter_archive_enhanced file, images file and twitter_json file manually. The images file used machine leaning techniques to determine which breed the dog was. The twitter_json file was consisting of more information like favourite_count and retweet_count.

The main motive of this was to collect data from different sources to perform analysis as most data has to be researched and downloaded according the analysis goal.

2) Assessing the Data:

In this part of the wrangling I went through the information about the dataset to find out more about my dataset, look into the abruptness, removing unnecessary data variables and highlighting the ones needed.

For this I mainly used the following methods :

- head()
- sample()
- info()
- value_counts()

Next I highlighted the issues I encountered while the processing of assessing the data which are as follows :

Quality :

1) tweet_id is an integer

2) retweet_id doesn't consist of image in most cases. They must be removed.

3) The name column has many entries which do not look like names. The most frequent entry in name column is "a", which is not a name. There are also entries like "an" and "they" which do not make any sense.

4) The rating_numerator and rating_denominator columns have unusual values.

5) timestamp and retweeted_status_timestamp are currently of type 'object' which need to be a datetime object.

6) The number of rows in images dataframe do not match the number of rows in the archive dataframe. There are 2075 rows in the images dataframe and 2356 rows in the archive dataframe. There must be a few missing pictures or they may be deleted

7) p1, p2, and p3 contain underscores instead of spaces in the labels

8) doggo, floofer, pupper, and puppo have values that are the string "None" instead of NaN

Tidiness:

1) We can create one single variable that can consist of doggo, floofer, pupper, puppo. We'll name it Dog Stage

2) All the dataframes need to be merged.

3) Removing unnecessary variables that make the dataframe difficult to view

3) Cleaning the Data:

Using the appropriate methods I solved almost all the issues that I noticed and made sure that my data was ready to undergo analysis. Especially, the rating one was very tricky as there is no real way in which you can identify the rating system. But I have given it a try, I hope it works out well enough.