DATA WRANGLING DOCUMENTATION

BY: TOLULOPE OLUWADARE

Introduction

The Data Wrangling for this project required a cursory look at the data available and the use of multiple Pandas functions to make the data suitable for analysis. It must be noted, however, that there are still many other data wrangling techniques that could've been applied to the datasets provided. This report states how I wrangled the data to have a dataset suitable for analysis. All the data provided was for the WeRateDogs Twitter account.

Data Gathering

There were three different datasets in three different formats and from different sources. They are:

archive.csv

This dataset contained 17 columns and 2356 rows of data and was downloaded manually. This data contained a lot of information on the tweets from the WeRateDogs Twitter account.

image-predictions.tsv

This was programmatically downloaded from Udacity and contained data related to dog breed predictions made on the dogs in the tweets made by WeRateDogs with 2075 rows and 12 columns.

tweet-json.txt

This dataset contains json file data but is saved as a txt. It has 2354 rows and 31 columns. It contained data similar to the archive.csv file and data it did not have as well.

Data Assessment

The three different datasets had a varying number of rows which meant possible missing, duplicate or incorrect data across all three datasets. Each dataset was assessed individually for data quality issues. Below is the assessment summary:

Archive data

- Remove retweets. i.e., tweets where rt_status_id is not null

- Drop irrelevant columns

- Change the datatype of timestamp and tweet_id columns

Prediction data

- Change the datatype of the tweet_id column

- Adjust inconsistencies in values in p1_dog, p2_dog, p3_dog columns

- Change column names to make them easy to understand

- Remove duplicates in the jpg_url column

- Drop irrelevant columns

Tweet data

- Change the datatype of the id column

- Change the name of the id column to tweet_id

- Change HTML values in the source column

Data Cleaning & Storage

Cleaning the data involved resolving all the issues identified in the data assessment including dropping unnecessary columns and merging all three datasets using the tweet_id column as the primary key. The merged data was checked for possible duplicates and then stored as a csv file. Data Analysis was performed on the merged data created.