1. Introduction

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog according to Wikipedia. The Twitter archive for this accounts is going to be analyzed using Python so the data of this archive has to be cleaned, in order to do so we have to do the Data Wrangling Procedure according to the following steps:

1. Data Gathering
2. Data Assessing
3. Data Cleaning

In this report I will breakdown each step and what we have done to reach our final result in this project.

2. Data Gathering

We had to gather 3 Dataframes from different sources and these df's are the following:

1. twitter_archive_enhanced.csv
2. image_predictions.tsv
3. tweet_json.txt

2.1. twitter_archive_enhanced.csv

The first data file we had in our hands was the Archive in csv file called "twitter_archive_enhanced.csv", this data file contained every detail that was related to the raw content of the tweet, and its contains the following columns: (tweet_id, in_reply_to_status_id, in_reply_to_user_id, Timestamp, Source, Text, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls, rating_numerator, rating_denominator, Name, Doggo, Floofer, Pupper, Puppo)

2.2. image_predictions.tsv

The second file we had in our hands also was a tsv file called "image_predictions.tsv", this data file was created by using neural network that can classify dogs based on their photos that was tweeted, this file contained the following columns: (tweet_id, jpg_url, img_num, p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog)

2.3. tweet_jason.txt

This file was created using Twitter API in order to extract number of like and number of retweets for each tweet in the first file "twitter_archive_enhanced.csv" and it contains 3 columns: (tweet_id, faviorite_id, retweet_id)

3. Data Assessing

An assessment was done to the previous gathered data, the assessment was divided into two main group:

3.1. Tidiness Issues

The assessment found the following issued related to the tidiness of data:

1. Dog stages should be in one column.
2. All tables should be one table.

3.2. Quality Issues

The assessment found the following issued related to the quality of data:

1. 'tweet_id's should be object instead of int in all df's also 'favorite_count' and 'retweet_count' should be int type instead of float.
2. Retweets included in the table.
3. 'timestamp' column should be timestamp dtype.
4. Zero numerator, denominator rows.
5. Denominator != 10
6. image_pred df has 2075 entries while twitter_arch has 2356 entries.
7. All algorithm columns (p1, p2 and p3) inconsistent with the first letter either capital or small.
8. Undescriptive column names.
9. 'source' column has only 4 values that are not readable.
10. 'a', 'an' and 'the' values in 'name' column.

4. Data Cleaning

The cleaning of the data was done according to issues found in the assessment as following:

4.1. Tidiness Issues

1. last four columns of twitter_arch_clean df will be milted together in one column called 'dog_stage' through replacing "None" values with "" then adding all strings in four columns together in dog_stage column then the four columns in question will be dropped out of the df.
2. json_df and twitter_arch_clean and image_predc will be merged together through merge statement

4.2. Quality Issues:

1. Change the tweet_id column type to their proper types dtype.
2. Drop all retweeted data from our table then drop 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id','retweeted_status_user_id', 'retweeted_status_timestamp' columns.
3. Change time stamp from object type to timestamp type.
4. After merging both dataframes the zero denomenator disappeared and only 2 rows left with zero numerator so I will check them manually and change them upon this process. After checking the manually tweet both rows will be dropped from both dataframes because the first one is a joke clearly and the second one is non dog related.
5. drop all rows with denominator != 10
6. All tweets without image will be dropped.
7. Change the first letter of each value to capital.
8. Change the columns p1,p1_conf,p1_dog,..... to descriptive column names.
9. Extract the human readable part of the source type
10. As the assessments showed that any name starts with small letter is not an actual name, so any name value starts with small letter will be replaced by 'None', then we will try to find if there is

actual names in these tweets because in many of these tweets the actual name is written after word 'named'.