Data Wrangling Report

As assignment for the Udacity Data analysis Nanodegree, This report will show the main steps done for Twitter account “WeRateDogs”

Data Gathering:

This is the first step done, there were three main sources for the data to be gathered:

1. Twitter-archive-enhanced.csv file, it was in the classroom and downloaded manually to our working directory and then imported to our working environment notebook.
2. Image-predictions.tsv, this file was hosted on a web-page and it was requested by the its url by requests library in python and saved in our working directory.
3. The final dataset was gathered from Twitter REST API via Tweepy library. Twitter developer account must be created to make an API and tokens to get the data, then querying on it and get extra information: retweets count and favorite count.

Data Assessment:

This step to define the issues in two ways, visually and programmatically.

1. Visual assessment: done on spreadsheet like excel or text editor.
2. Programmatically assessment: done in jupyter notebook using pandas functions.

The main two issues are quality and tidiness issues:

1. Quality issues: the issues with content:
    1. Rows in Twitter-archive-enhanced that is retweets or replies.
    2. Missing values in column ‘expended_urls’ in Twitter-archive-enhanced dataset.
    3. Columns (doggo, floofer, pupper, puppo) have ‘None’ instead of np.nan for missing values.
    4. Column ‘name’ has inconsistent values (None, a, an,…).
    5. Column ‘timestamp’ has wrong datatype.
    6. Columns ‘rating_numerator’ and ‘rating_dominator’ have inaccurate values.
    7. Columns (p1, p2, p3, p1_conf, p2_conf, p3_conf, p1_dog, p2_dog, p3_dog) in Image-predictions.tsv dataset are nondescriptive.
    8. There is difference between number of tweets in Twitter-archive-enhanced dataset, api dataset and Image-predictions dataset.
2. Tidiness issues: the issues with structure that prevent easy analysis:
    1. Columns (doggo, floofer, pupper, puppo) are values not variables.

2. The three datasets should be joined.

Data cleaning:

This step to make solution for issues that defined in the data assessment step.
The solutions:
For quality issues:

1. Drop these rows from dataframe.
2. Drop these rows from dataframe.
3. Replace the ‘None’ value with np.nan using pd.replace function.
4. Extract the right name from the tweet text and replace (‘None’, ‘a’, ‘an’, …) with its right name, and if there None values remain replace it with np.nan.
5. Convert datatype to datetime using pd.to_datetime.
6. Check the rating dominator with value more than 10 and get the dogs count and then divide rating numerator over dogs count to get the right numerator.
7. Replace these names with descriptive names.
8. Delete rows from Twitter-archive-enhanced dataset that haven’t image.

For Tidiness issues:

1. Compress these columns to one column that has the assigned category.
2. Merge the datasets using pd.merge function.

Data storing:

Store the results in csv file using dataframe.to_csv function.