Wrangle Report

This project report will represent the three critical steps for wrangling the data for analysis and visualizations which starts with gathering the data, assessing the data with two kinds of visual assessment and programmatically assessment, and with two outputs: quality issues, and tidiness issues. Finally, cleaning the data with three sub-steps, define the issues, run the codes, and test the process.

Gathering Data

In the beginning, there were three different but related data sets, each data was imported programmatically in an appropriate way. The Twitter archive was directly imported using 'pd.read_csv' code. Image predictions data was imported using request library through URL to get the response, then read the extracted file using 'pd.read_csv' . The tweets json data was imported by first unzipping the file, and then read the file using 'pd.read_json' code.

Assessing Data

This step was important after importing the data sets and it helped to understand each data set's issues to solve later in the cleaning step. Through visual assessments, I got a glance at understanding the structure of each data set and the content of each data set. Throughout the programmatic assessment, I got a chance to understand each data set's issues which are:

The following quality issues in twitter archive data:

1- Date type for timestamp and retweeted_status_timestamp columns instead of the object type
2- String type for tweet_id instead of integer

3- Remove retweets and replies from df_twitter_archive_enhanced data frame

4- Fix incorrect numerator values

5- Fix incorrect denominator values

6- Remove un-needed columns like the retweets and replies information

The following quality issue in image predictions data:

7- String type for tweet_id instead of integer

The following quality issue in twitter json data:

8- String type for id instead of integer

9- Different id column name in comparison other 'tweet_id' column name

The following tidiness issues in both the twitter archive and tweet json:

1- Dogs stage column instead of various columns for doggo, floofer, pupper, and puppo

2- Retweets and favorites information can be joined to the other data frames

Cleaning Data

In this step, every reported issue will be solved for each data set through three sub-steps as mentioned before: define, code, test. In the defined sub-step, I have re-written the reported issue in detail with the action to take and also wrote the appropriate method to use. In the code sub-step, I have written the code and run it. Finally, in the test sub-step, I have tested the success of the code in solving the reported issue.

The following process was taken in order to solve the issues:

As for the data type issues, the ‘astype’ code, and ‘pd.to_datetime’ were used to solve these issues. After that, removing the unwanted rows that are related to WeRateDogs retweets and replies was done successfully by using drop code, and later the related columns were removed by using drop code as well. As for inaccurate numerator and denominator values, these issues were solved by re-extracting the values from the text by using extract code. Through the assessment

process, I've realized that the tweet json table has different id names than the rest tables and this could cause an error in combining the tables into one table. Hence, this issue was solved by using the rename method. On the tidiness issues side, the different dog stages name that appeared in different variables was joined into one variable column using both combine and 'if statements' methods. After solving all these issues, I was able to join the three tables into one table successfully and then store it for visualizing it.