Wrangle Report

Introduction:

The WeRateDogs Twitter archive contains basic tweet data for I 2356+ of their tweets.

WeRateDogs is a twitter account that rates people's dogs with some comments about the dog. This project works through the gathering of data (which consists of three files) , assessing, cleaning of the data and finally visualization and observation from the analysis provided. In this project I will describe the processes I've made to gathe, assess and clean the Twitter archive.

Gather data:

In order to do analyze the archive, three files were gathered:

- The WeRateDogs Twitter archive: (*twitter_archive_enhanced.csv)* was downloaded manually and consists of more than 2300 tweets from WeRateDogs Account.
- The tweet image predictions: (*image_predictions.tsv*) i.e what breed of dog is present in each tweet according to a neural network .It was downloaded programmatically.
- Additional Data via the Twitter API (*.json file*) contains data of how many retweets and favorites each tweet has.

Asses data:

After gathering those files, they were loaded into separate Pandas dataframes to be assessed.
Each file was evaluated both visually and programmatically, and the following issues were observed during the process:

Quality issues:

1. Timestamp should be in datetime format (The WeRateDogs Twitter archive)

2. Drop columns with too many missing values (The WeRateDogs Twitter archive)

3. Change tweet_id to type str in order to merge with the other 2 tables (The WeRateDogs Twitter archive)

4. Keep original ratings (no retweets) that have images (The WeRateDogs Twitter archive)

5. Errors in name column (a, an,...) in (The WeRateDogs Twitter archive)

6. Keep original tweets (.json file)

7. drop columns (tweet image predictions)

8. Drop jpg_url duplicated (tweet image predictions)

Tidiness issues:

1. Melt the doggo, floofer, pupper and puppo columns (The WeRateDogs Twitter archive)

2. merge all the data frames in one Dataset

Clean data:

Quality and tidiness issues were cleaned using techniques such as:

- Change date format with .to_datetime
- Dropping columns with a lot of missing values
- Change data types of some columns
- Keeping only the original ratings that have images
- Deleting the names of dogs that were incorrect
- keep original tweets only
- Drop columns that were not useful for the analysis
- Delete rows with duplicated images
- create additional columns with all the values of the four columns and then create a function to clean the values
- create additional columns with all the values of the four columns and then create a function to clean the values
- merging the three dataframes in one Dataset