Reporting: wrangle_report

The dataset is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. Now we have 3 pieces of information on hand:

The steps I used to complete this project are:

Wrangling the twitter data through the following processes:

- Gathering data
- Assessing data
- Cleaning data

- Storing data

- Analyzing, and Vsualizing the wrangled data

- Reporting on the data wrangling steps, and data visualizations

Data Gathering

In this step I gathered three pieces of data for this project and then loaded them in the notebook, these three are from three different sourses:

- twitter-archive-enhanced.csv: opened directly with pandas method ‘read_csv’ and then stored it in dataframe called ‘twArchive’.
- image-predictions.tsv: used Request and get method, then stored it in dataframe named ‘imagePred’.
- twitter api and json: I created a loop to get the tweet id and created dataframe 'json' and save it in a file and called it ‘twJson’.

twArchive got 2356 tweets, imagePred got 2075 tweets, twJson got 2354.

Assessing Data

In this step my goal is to detect and document at least eight quality issues and two tidiness issue, I tried to use both:

- **Visual assessment**: in the visual assessment I used .head() to be able to catch some tidiness and quality issues easily by looking.
- **Programmatic assessment**: in the programmatic assessment I used info() to be able to identify duplicates and datatypes and do some observations like doing value count of some columns to identify some issues.

Quality issues:

1. `twArchive` Remove any retweets and replies, just want tweets of dog_rates account
2. `twArchive` Delete unnecessary columns that we won't use it.
3. `twArchive` *timestamp* datatype should be datetime.
4. `twArchive` Some null values are 'None' *string* instead of NaN
5. `twArchive` some rating_denominator values are not 10
6. `imagePred` There are some images that are duplicated (66 images)
7. `twArchive` Some rows don't have expanded_urls so the tweet is without picture
8. `imagePred` in p1, p2 and p3: stabilize capitalization and remove underscores

Tidiness issues

1. `twArchive` *doggo, floofer, pupper* and *puppo* variables are unique columns and should be combined into one column.
2. Merge these three dataframes: `twArchive`, `imagePred`, and `twJson` into one dataframe based on 'tweet_id'.

Cleaning Data

This is the step to fix each of the quality and tidiness issues, I used three steps to each issue, first is **Define** which is defining the issue and how I am planning to fix it, second is **Code** which is the code that will clean the issue, finally is **Test** which is a code test that prove that the problem is fixed.

Before any cleaning, I assure that each piece of data have its copy.

Storing Data

In this step I saved this gathered, assessed, and cleaned master dataset to a CSV file and named it 'twitter_archive_master.csv. by using <.to_csv>

Analyzing and Visualizing Data

The final step is Analyzing and Visualizing Data, which I provide a separate report that focuses on this final step and provided it in 'act_report.pdf'