Data Wrangling Project

By: Nashwa Shafik Shokry

WeRateDogs Data Analysis:

Gathering:

- Requested a developer account from Twitter to be able to collect the number of favourites and retweets.
- Programmatically downloaded the predictions file and saved it.
- I created a config.py file that contains my keys as a dictionary, which I imported in my notebook.
- Used tweepy APIs to get the tweet status by tweet_id (some of the tweets are no longer existing, may be they were deleted) but I decided to keep them in my analysis! If I want to ignore them later, I could filter them out.
- Looked for a solution in Jupyter Notebooks that can skip running the cells that get the tweets, after creating the tweet_json.txt file, but for simplicity, decided to check for the file existence.

Assessment:

- Opened the available csv file (twitter-archive-enhanced.csv) in Excel for a fast visual assessment.
- Performed programmatic assessment for archive_df and documented the findings.
- Opened the predictions ".tsv" file in word pad to check the gathered data.
- Performed a visual and programmatic assessment of the predictions file and documented the findings.
- I picked some of the tweets with no dog predictions and checked them in twitter: some of them contain dogs others don't.
- After getting the tweets from twitter, I performed a quick assessment of the data. As the most important data are the favorite_count and retweet_count, I didn't pay much attention to other columns. Just checked for possible useful data that can give more insights.
- The ID and ID_STR were confusing, as they are different, but after some search, found that I should use the ID.

Cleaning:

- First copied the 3 data frames into cleaning data frames, to keep the original ones as a reference such that I can re-run the cells below when needed.
- Started by removing unwanted data from the main archive data frame, so, removed retweets and replies as well as rows having no images.
- Then I performed some tidying. Although there are not many tweets with dog categories, I found it might be interesting to compare ratings of different categories of dog.

- While I was performing the tidying, I had to re-assess the data frame, and discovered some errors in categorizing dogs, I resolved them.
- To simplify the analysis, I added the highest predicted dog type prediction to the archive data frame.
- When testing, I discovered very low confidence values, but after picking a sample and checked the tweets in twitter, I decided to accept them.
- Resolved quality issues.

Storing, Analyzing, and Visualizing Data for this Project

- I stored the dataframe to a csv named "twitter archive master.csv" file and checked its contents using Excel.
- And finally, the most exciting part is the analysis and visualization. During it I added some columns that were not saved to the file. I added a column to plot against the year and month. I ignored August 2017 from time analysis as it only has one day data. I added a column to sum the interactions (favorites and retweets).
- I got some insights that are summarized in the insights report (act_report.pdf).

Regression test:

- I re-run all the notebook and fixed errors that were due to changing the order of cells (referencing variables that were not defined).
- And then made a copy of config.py to ban the keys.