Wrangle report

Introduction:

This project focuses on gathering, assessing, cleaning, and analyzing data from the WeRateDogs Twitter account. The main dataset includes tweet data about dogs, which we enhance using additional sources like image predictions and tweet metadata. The goal is to practice real-world data wrangling techniques using Python libraries like Pandas, NumPy, and Matplotlib, leading to meaningful insights and visualizations.

1. Gathering Data

In this project, I need to gather data from several sources and different of formats.

1. The WeRateDogs Twitter archive. The file was provided by the project and can be downloaded directly from Udacity website.
2. The tweet image predictions. The file is hosted on Udacity's servers. I downloaded this file programmatically by using the *Requests* library in Python.
3. Get retweets count and favorite count information missing from the Twitter archive from another file. I chose to download the tweet JSON file programmatically by using the Requests library since I don't have a Twitter account.

2. Assessing data

After gathering the data, I assessed the data both visually and programmatically to identify any data quality and tidiness issues. Quality relates to content while tidiness relates to data structure. Tidy data requirements: each variable forms a column, each observation forms a row, each type of observational unit forms a table. Since the datasets were not large, I was able to open them in EXCEL and scrolled through the data and spot any obvious issues. I also used code in Jupyter Notebook to view specific portions and summaries of the data, for example, pandas' info, head, sample, value_counts, duplicated, query, and describe methods. I've made

notes of the observations while assessing the data so that I can fix the issues later in the cleaning step.

Quality Issues

twitter_archive table:

- datatypes for in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id are float, should all be int
- only need original ratings with pictures, retweets and replies should be removed, related columns should be removed too. The picture part will be fixed later.
- timestamp is str, should be datetime, remove +0000 in timestamp

- abnormal values in rating_denominator, e.g., 170, 150, 130, etc. The rating_denominator is almost always 10
- abnormal values in rating_numerator, e.g., 1776, 960, 666, 204, 165, etc. make no sense.
  - source info redundant, not easy to read .

image_prediction table:

- inconsistent capitalization in p1, p2 and p3 columns
- jpg url duplicates
- many entries are not dogs, e.g., jaguar, mailbox, peacock, cloak, etc.

tweet_json table:

- missing data probably due to retweets in twitter_archive

Tidiness Issues

- twitter_archive: doggo, floofer, pupper, puppo are all stages of dog, should be in one column
- The three tables should be combined into one since they're all related to the same type of observational unit according to tidy data requirements.

Cleaning data

I worked on cleaning the dataset based on the issues I found earlier. I didn’t clean every single problem in the data — that would take too long and some of them aren’t important for what I want to analyze. So I just focused on the ones that actually affect my analysis.

For each issue, I followed the basic process: define it, write the code to fix it, then test it to make sure it worked. I also made copies of the original data before starting any changes, just in case.

Some issues were related, so when I cleaned one, others disappeared. For example, I had columns like retweeted_status_id and in_reply_to_status_id that had wrong data types, but since I removed all retweets and replies (because the project only needs original tweets), those columns were no longer needed, so I dropped them and didn’t have to worry about their types.

Another thing was the rating columns. Some tweets had weird values for rating_numerator or rating_denominator. I fixed some manually by checking the tweet text, but in some cases, removing non-dog images helped get rid of the bad ratings automatically.

I also changed some incorrect names like “a” or “the” to None and converted the timestamp column to datetime format. For tidiness, Most of the problems were solved also during solving the quality issues, but the second tidiness I removed all the null values in twitter_archive_clean.jpg_url

After fixing all the issues, reassess the dataset and iterate if necessary. Then store clean data in a csv file.