Wrangle report

The main objective of this project is to improve data wrangling skills through the utilization of real-world data. The data wrangling process entails three key steps: gathering, assessing, and cleaning. In this particular case, the dataset utilized for the project is the tweet archive from the Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs in a humorous manner, providing amusing comments about each dog. Initially, I will collect data from various sources in different formats. Subsequently, I will assess the data both visually and programmatically to identify any issues related to data quality and tidiness. Once identified, I will proceed with programmatic cleaning to resolve these issues. Finally, I will analyze the cleaned dataset and present the results through visualizations.

1. Gathering Data

To complete the project, I need to collect data from different sources and formats.

- Download and upload file twitter-archive-enhanced.csv and read it into a Pandas DataFrame
- Download file image-predictions.tsv and read it into a Pandas DataFrame
- I have queried each tweet's retweet count and favorite ("like") count using the Tweepy library and stored the data in tweet_json.txt
- I have read the tweet_json.txt line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count.

2. Assessing data

After collecting the data, I evaluated it comprehensively using visual and programmatic methods to identify any issues related to data quality and tidiness. Data quality refers to the content, while tidiness refers to the structure of the data. Tidy data should have variables as columns, observations as rows, and each observational unit corresponding to a table. Additionally, I utilized Jupyter Notebook with code (e.g., info, head, sample, value_counts, duplicated, query, and describe methods) to examine specific sections and summaries of the data. Throughout the assessment, I took notes on my observations to address these concerns during the subsequent cleaning phase.

Quality Issues

df_twitter_archive_enhanced dataframe:

1. The columns in_reply_to_status_id, in_reply_to_user_id, in_reply_to_user_id and retweeted_status_user_id of data type is float64, but should be convert to int64 same column tweet_id.
2. Column timestamp is object should be convert to datetime64 and remote '+0000'.
3. Only original ratings accompanied by pictures are required, so entries related to retweets and replies should be removed.
4. Column rating_denominator has most of the values 10, the other values appear only 1,2 or 3 times. These values appear to be abnormal.
5. Abnormal values(count value few occurrences) also exist in the rating_numerator column, such as 88, 84, 960 ...
6. The doggo, floofer, pupper, and puppo columns represent different stages of dogs and should be combined into a single column.
7. The source information is redundant and not easily readable.
8. Delete columns that won't be used for analysis.

df_image_prediction dataframe:

1. Column jpg_url have duplicate record. Should remove duplicate
2. Numerous entries do not correspond to dogs; they include unrelated items such as "orange", "banana", "ox" ...
3. Inconsistencies in capitalization are observed in the p1, p2, and p3 columns.
4. Delete columns that won't be used for analysis

df_tweet_ dataframe:

1. It is likely that missing data in the " df_twitter_archive_enhanced " can be attributed to retweets and should be considered as part of the table.

Tidiness Issues

1. In df_twitter_archive_enhanced dataframe, all stages of a dog, including doggo, floofer, pupper, and puppo, should be listed in a single column
2. The According to the principles of tidy data, the three dataframe, which pertain to the same type of observational unit, need to be merged into a single dataframe

3. Cleaning data

I addressed documented assessment issues, but due to the numerous problems in the dataset, it was time-consuming to clean all of them. Therefore, I focused on issues relevant to my analysis. The programmatic data cleaning involved three steps: defining, coding, and testing.

For example, abnormal rating numerators and denominators seen earlier disappeared after removing non-dog images from the image_prediction dataset. This automatically resolved part of the abnormal ratings. While most cleaning was done programmatically, some required manual intervention, like rectifying abnormal ratings by carefully reviewing the text.

After resolving all problems, the dataset was reevaluated, iterating if necessary. Finally, the cleaned data was stored in a CSV file named 'twitter_archive_master.csv'.