WeRateDogs Data Wrangling and Analysis

In this short report, we will cover the steps of wrangling and analyzing data from WeRateDogs. WeRateDogs is an account on Twitter that allows users to rate dogs. We will cover the following steps:
Step 1: Gathering data
Step 2: Assessing data
Step 3: Cleaning data
Step 4: Storing data
Step 5: Analyzing, and visualizing data

Gathering Data

The data gathering occurred using 3 sources:

1. Directly downloaded Twitter archive data from file twitter_archive_enhanced.csv
2. Tab separated file (image_predictions.tsv) is downloaded using the Requests library. The file contains tweet image prediction
3. Json data gathered through Twitter API using the Tweepy library (tweet_json.txt)

Assessing Data

The data assessment was done by:

1. Visual Assessment
2. Programmatically using Pandas, Matplotlib, and Seaborn

The following issues were found in the three files:
Twitter_archive_enhanced:
The data were loaded into the Pandas' **twitter_archive_enhanced_data** dataframe.

A. Quality Issues

1. columns that contain null: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls.

2. doggo, floofer, pupper and puppo columns contain None which is defined as an object data type instead of being counted as null type.

3. rating_denominator contains 0
4. tweet_id is integer
5. name column: for the unique sorted values in this column, all names start with capital values, the last 25 values are not names.

6. Source column contains data as an html tag

B. Tidiness Issues

Structural issue: the categories doggo, floofer, pupper and puppo in separate columns although they represent the same feature.

Image_predictions:

The data was loaded into thePandas' **image_predictions_data** dataframe

A. Quality Issues

1. 'p1', 'p2', 'p3' columns have inconsistency formatting issues in breeds name some are capitalized, some are hyphenated and others are not.

B. Tidiness Issues

1. display_text_range contains a list that has two values

2. image_predictions_data should be in the same table as twitter_archive_enhanced_data instead of a separate table

Tweet_json:

The data were loaded into the Pandas' **api_data** dataframe.

A. Quality Issues

1. 'contributors', 'coordinates', 'geo', 'in_reply_to_screen_name', 'in_reply_to_user_id', 'in_reply_to_user_id_str', 'is_quote_status' contain null values

2. created_at is not a time stamp and condensed in a single column

B. Tidiness Issues

1. columns that contain dictionaries which could be separated into their own columns: entities, extended_entities, user

---

Cleaning Data

To clean the data each dataframe was copied into a new one.

Twitter_archive_enhanced_data → archive_enhanced_copy
Image_predictions_data → img_pred_copy
Api_data → api_copy

archive_enhanced_copy:

A. Quality Issues Solution

1. Drop columns that contain null: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls.

2. Using melt function condense values of doggo, floofer, pupper and puppo columns into a single column stage.
3. Drop rows where rating_denominator contains 0
4. Convert tweet_id into string
5. Drop rows where the name column doesn’t represent dog names.
6. Extract the source from the anchor tag into the source column

Tidiness Issues Solution
Using the melt function for the doggo, floofer, pupper, and puppo into the stage column.

img_pred_copy:

A. Quality Issues Solution

1. Applying lowercase and removing - using the replace method to 'p1', 'p2', 'p3' columns.

B. Tidiness Issues Solution

1. Extracting the length of tweet from the display_text_range.

2. Merging the img_pred_copy with archive_enhanced_copy into df1

api_copy:

A. Quality Issues Solution

1. Dropping 'contributors', 'coordinates', 'geo', 'in_reply_to_screen_name', 'in_reply_to_user_id', 'in_reply_to_user_id_str', 'is_quote_status' columns.

2. Split the created_at column into 6 columns.

B. Tidiness Issues Solution

1. Dropping entities, extended_entities, user columns.

Final Dataframe
Merging df1 with api_copy and dropping all columns that will not be used in analysis or contain redundant data including: 'timestamp', 'id', 'id_str', and 'in_reply_to_status_id'.

Storing Data

Cleaned data is stored into the file "twitter_archive_master.csv

---

Analyzing, and visualizing data

The insights were found using several pandas methods including groupby(), sort_values(), and value_counts().
Visualizations included: Line plots, scatter plot, bar plot, and Seaborn’s count plot.
