Data Wrangling Report by Michal Ezeh

The dataset wrangled, analyzed, and visualized 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 comments about the dog.

Project Objectives

The objectives of the project included:

- Performing data wrangling (gathering, assessing, and cleaning) on the different data sources provided.
- Storing, analyzing, and visualizing the wrangled data.
- Reporting on: a. My data wrangling efforts b. My data analyses and visualizations.

Project Overview

Step 1: Gathering Data

In this step, three different datasets were gathered and read into dataframes. The following datasets were:

- **The WeRateDogs Twitter archive**

The file had to be downloaded manually by clicking the following link: twitter_archive_enhanced.csv. Once it is downloaded, it was then uploaded and read into a pandas DataFrame.

- **The Tweet Image predictions**

This file (image_predictions.tsv) had to be downloaded programmatically using the requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv.

- **Additional data from the Twitter API**

Using the tweet IDs in the WeRateDogs Twitter archive, I had to query the Twitter API for each tweet's JSON data (each tweet ID, retweet count, and favorite at the minimum). Using python's Tweepy library then to store each tweet's entire set of JSON data in a file called tweet_json.txt.

Step 2: Assessing Data

The datasets were assessed both visually and programmatically and while working, a number of issues were observed. The issues with the datasets were separated into quality and tidiness issues.

Quality Issues

From the WeRateDog Twitter archive dataset(archive_enhanced table)

- Missing values in the in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp.
- Source Colum has HTML tags present.
- The expanded_url column contains both twitter and non-twitter links.
- Erroneous datatype for timestamp and retweeted_timetamp.
- Erroneous datatype for tweet_id.
- Inconsistent data entry as the name column has entries both in upper case and lower case.

image_predictions table (image predictions dataset)

- Inconsistent data entry format as p1, p2, and p3 have entries both in uppercase and lowercase.
- The missing number of records between the image_predictions and the archive_enhanced table.
- Erroneous datatype for tweet_id.

Tweet_json table (Additional Twitter API dataset)

- Erroneous datatype for tweet_id (should be string).
- Missing number of records (number of retrieved tweets is less than the number of tweet ids in archive_enhanced).

Tidiness issues

Archive_enhanced table (Twitter archive dataset)

- doggo, floofer, pupper, puppo variables all represent values of a single column, dog_stage.

Tweet_json table (Additional Twitter API dataset)

- tweet_json table should be joined to the archive_enhanced table.

Step 3: Cleaning Data

After assessing the data and finding out the issues with the datasets, I set out to find solutions in cleaning the data. The table below shows the issues assessed from the data with the actions carried out to clean it.

QUALITY ISSUES

Dataset	Issues Observed	Solution
archive_enhanced table	Missing values in the in_reply_to_status_id,in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp.	Dropping the above-listed columns as they columns that have missing data (number of entries lesser than 2356).
	Source Colum has HTML tags present.	Due to improper formatting, there is the presence of HTML tags in the entries of the source column. This will be cleaned by using string slicing.
	The expanded_url column contains both twitter and non-twitter links.	Replace the non-Twitter URLs with the Twitter version by joining the https://twitter.com/dog_rates/status/ to the tweet_id
	Erroneous datatype for timestamp and retweeted_timetamp.	The datatype for timestamp and retweeted_timestamp should be datetime and not integer.
	Erroneous datatype for tweet_id.	The datatype for tweet_id should be string and not integer.
	Inconsistent data entry as the name column has entries both in upper and lower case.	Allow entries to the dog name column to all be in lowercase for consistency.

image_predictions table	Inconsistent data entry format as p1, p2, and p3 have entries both in uppercase and lowercase.	For consistency, all data entries in the p1,p2, and p3 columns will be changed to lowercase.
	Erroneous datatype for tweet_id.	The datatype for tweet_id should be string and not integer

TIDINESS ISSUES

Dataset	Issues Observed	Solution
Archive_enhanced table	doggo, floofer, pupper, puppo variables all represent values of a single column, dog_stage.	Melt the rows into one, making sure rows where all values are none are represented as None
Tweet_json table	tweet_json table should be joined to the archive_enhanced table.	Merge the tweet_json_clean dataframe with archive_enhanced_clean using pandas merge function.