Wrangle Report

By Michelle Bleyl

Introduction:

The Wrangle data and Analysis Project is the tweet archive of WeRateDogs(@dog_rates) that is a Twitter account that rates people's dogs with a humorous comment and rating about the dog.

- The wrangling process includes:
  1. Gathering data
  2. Assessing data
  3. Cleaning data
- Storing, analyzing, and visualizing our wrangled data
- Report our wrangle process, analysis, and visualization.

Gathering Data

1. **WeRateDogs Twitter Archive** is already given by Udacity page. I download manually and got "twitter_archived_enhanced.csv".
2. **The tweet image prediction**, i.e., what breed of is present in each tweet according to a neural network (image_predictions.tsv). This file is provided by Udacity server and downloaded programmatically using requests library.
3. **Twitter API & JSON**. Using the tweet IDs in the WeRateDogs twitter archive. I quired the Twitter API for each tweet's JSON data using Python's tweepy library. Gather each retweet count and favorite count as minimum, and added followers count too for analysis.

Assessing Data

Now we have gathered the data, check these data and find at least 8 quality issues and 2 tidiness issue.

Quality Issues:

- Completeness
  1. Remove Retweets. If "retweeted_status_id", "retweeted_status_user_id", "retweeted_status_timestamp" exists that means the tweet is a retweet.
  2. Missing data in the following columns: "in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_id", "retweeted_status_user_id", "retweeted_status_timestamp", and "expanded_urls". (Twitter Archive)
  3. Tweet_id data type is int. (Twitter Archive and Image Prediction)
- Validity
  4. Some dog's names are "None", "a", "an", "very" etc. (Tweet Archive)
- Accuracy
  5. Timestamp is an object. (Twitter Archive)
  6. Some rating_numerator is inaccurate. For example, rating numerator goes up to 1776.
- Consistency
  7. Some row in P1, p2, p3 are in lower case. (Image Predictions)

8. rating_denominator is a standard 10, but there are values other than 10. For example, there are ratings like 88/80.
9. We have duplicated tweets based on their jpg_url meaning that they are using the same photo. (Image Predictions)

Tidiness Issues:

1. doggo, floofer, pupper, puppo are related to each other but they are separated by columns.(Twitter Archive)
2. p1, p2, p3 are all about the dog's breed. They are all related. (Image Predictions)
3. Twitter Archive, Image Predictions, and Twitter API & JSON are all related.

Cleaning Data

After assessing data, now clean their quality issues and tidiness issues by the following process. Use define, code, and test method.

1. Remove retweets: Only store rows that are null in "retweeted_status_id", "retweeted_status_user_id", "retweeted_status_timestamp".
2. Drop columns that have too much missing data: "in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_id", "retweeted_status_user_id", "retweeted_status_timestamp", and "expanded_urls" in Twitter Archive are all removed.
3. Tweet_id data type is int: Change data type from integer to string for both Twitter Archive and Image Predictions.
4. Some dog's names are invalid: Change these invalid names all into "None".
5. Timestamp is object: Change their data type to timestamp.
6. Some rating_numerator is inaccurate: Delete these inaccurate ones.
7. Some row in P1, p2, p3 are in lower case: Capitalized the first letter.
8. Some rating_denominator is other than 10: change rating into rating_numerator divided by rating_denominator. For example, if rating_numerator is 88 and rating_denominator is 80 then we change this to 88/80=11/10=1.1. After dividing all ratings, delete rating_numerator and rating_denominator.
9. We have duplicated tweets based on their jpg_url: remove these duplicated jpg_urls.
10. doggo, floofer, pupper, puppo are related to each other: Combine these 4 columns into 1 column that is a categorical variable.
11. p1, p2, p3 are all about the dog's breed: Combine these columns into one column called "Breed".
12. Twitter Archive, Image Predictions, and Twitter API & JSON are all related: Merge all these dataframe into one dataframe. Name this "twitter_df".

Storing Data

Store "twitter_df" dataframe into csv file using