Wrangle Report

Introduction

The dataset wrangled in the project is from the Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates other users' submitted dog pictures and also writes a humorous comment about the dogs.

The project goals are as follows:

- Data wrangling, which consists of:
    - Gathering data
    - Assessing data
    - Cleaning data
- Storing, analyzing, and visualizing the wrangled data
- Reporting on the data wrangling efforts and the data analyses and visualizations

Gathering Data

The project required three different pieces of data to be gathered. They had to be different file formats and from different sources.

- The first data was the WeRateDogs Twitter archive. This file was provided and downloaded manually. This file included basic information such as: tweet id, timestamp, etc.
- The second data is Tweet image predictions. The file is hosted on Udacity's servers. The file was programmatically downloaded using the Requests library.
- The third data is to be queried from Twitter's API. The query will get each tweet's JSON data, and it will then be stored in a file called tweet_json.text. The file is then read into a pandas DataFrame.

Assessing Data

After gathering the data, the data was assessed for any quality and tidiness issues.

Quality issues are issues with the data's content. Is there any data missing? Does the data conform to realistic values? (Ex. A person cannot have negative height.) Is the data accurate? Is the data in a standard format?

In the twitter archive, I found 6 quality issues.

- There was missing data in some of the columns.

- Some columns were the wrong data type.
- Some of the values in rating_numerator were below 10 or were truncated because of decimal points.
- Some of the values in rating_denominator were not equal to 10.
- There were invalid names such as: a, an, by, etc.
- The data contained retweets, which we do not want in our data set.

In the image predictions file, I found 3 quality issues.

- There was missing data.
- I found inconsistent capitalization in the p1, p2, and p3 columns.
- There were duplicate values in jpg_url.

In the query data, I found 2 quality issues.

- There was missing data.
- There are also retweets in the data.

Tidiness issues are issues with the data's structure. There are 3 guidelines tidy data follows. Each variable forms a column. Each observation forms a row. Each type of observational unit forms a table.

In the Twitter archive, I found that the columns doggo, floofer, pupper, and puppo violate the variable forming a column guideline. These four columns can be combined into one single dog_stage column.

I found that the image predictions data should be joined with the Twitter archive. It violates the observational unit forming a table.

I also found that the query data should be joined with the Twitter archive. It violates the observational unit forming a table.

Cleaning Data

1. Merged the clean versions of the twitter_archive and image_preds.
2. Merged the clean versions of twitter_archive and query_data.
3. Deleted the retweets in the merged dataset.
4. Removed the columns related to retweets.
5. Removed columns that weren't necessary for analysis.
6. Converted the timestamp column from object type to datetime type.
7. For the rating_numerator that have decimal values, they were manually fixed.
8. There were also some rating_numerators that were wrong. In the tweet text, there are multiple instances of what could be considered ratings, so those were programmatically fixed.

9. Remove rows with no ratings.
10. Fix values in the name column.
11. Fix inconsistent capitalization in the p1, p2, and p3 columns.
12. Combine the doggo, floofer, pupper, and puppo columns into one column.