Data Wrangling Process

Project Overview

The Twitter account WeRateDogs took the world by storm with the popular meme "they're good dogs Brent" back in 2016. The account is known for it's hilariously adorable dogs featuring they're unique rating system of how good of boys and girls they are.

The purpose of this project is to derive insights from the data by cleaning various data sources and communicating the findings. The data used for this analysis includes dog breeds, dog classifications (ranging from "puppo" to "fluffer", how adorable), Twitter engagement data, and the WeRateDogs rating system.

Gathering Data

To derive interesting and trustworthy analyses and visualizations from the WeRateDogs account, three data sources were used:

- **Twitter archive file**

  An archive file of 2,000+ tweets from the WeRateDogs Twitter account after the year 2017. This data includes the tweet text, time the tweet was posted, the name of the dog featured, the rating given, and the classification of the dog (i.e. "doggo", "pupper", "fluffer", "puppo", how adorable). This file was given as a csv file and read in as a DataFrame using pandas' `read_csv()` function.

- **Image predictions file**

  A dataset compiled through use of a neural network that classify breeds of dogs based on tweet images was used in this analysis. The dataset includes the URL of the image and the top three breed predictions along with confidence levels for each prediction. The dataset was downloaded programmatically from a given URL using the `requests` package and read in as a tsv file.

- **Twitter API data**

  Numeric data of the tweets given in the archive file via the Twitter API. This includes the number of favorites the tweet received and the number of retweets. This data was gathered using the `tweepy` package and downloaded programmatically as a json into a txt file. The data was then extracted from the txt file and read into a DataFrame using pandas' `read_json` function.

Data Assessment

Once the data was successfully read into three separate DataFrames I assessed the data using the following methods:

- **Visual assessment**

  First I visually assessed the data by printing the contents of each pandas DataFrame, both printing in their entirety and using the `.head()` method.

- **Programmatic assessment**

I continued my assessment of the data by using a few different methods:

  - `.head()`
  - `.info()`
  - `.describe()`
  - `.value_counts()`
  - `.count()`
  - `.duplicated()`

By visually and programmatically assessing the data, I identified 9 data quality issues and 2 data tidiness issues, ranging from incorrect data types to null or incorrect values in categorical columns.

Cleaning the Data

After assessing the data, I cleaned the data via three steps: defining, coding, and then testing. I carried this out for each of the 11 points identified during the assessment phase individually. I originally found 7 quality issues during the assessment phase, however through cleaning I was able to iterate and discover 2 additional data quality issues.

To clean the data, I first began with addressing missing data and changing data types. This included converting the timestamp to a `datetime` object to allow for time series analysis, and removing tweets that were retweets rather than originals. Then I addressed data correctness issues, which predominantly included the string "None" rather than NaN.

Lastly, I addressed data tidiness in the original archive file regarding dog classifications. I used the pandas `melt()` function to collapse the columns into a single column to enable easier categorical analysis of each row.

Conclusion

After iterating through the cleaning and assessment phases a few times, I was able to manipulate the three DataFrames into datasets optimal for analysis. At the end of the cleaning process, I merged the three DataFrames using the `tweet_id` of each column. The result is a singular DataFrame containing the dog classifications, dog ratings, image predictions, and the engagement performance of each tweet. The final dataset is what was used for analysis.