Udacity Nanodegree - Project 4

Wrangle and Analyze Data | WeRateDogs Twitter

Project Requirements

Step 1: Gathering Data

In this step, there were data points from 3 different sources that i had to gather data from. They included gathering data from the following sources:

- The WeRateDogs Twitter archive. The twitter_archive_enhanced.csv file was provided to Udacity students. Mainly downloaded this .csv file, loaded it up onto github "https://raw.githubusercontent.com/bakedbry/udacity/main/twitter-archive-enhanced.csv"
- The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This .tsv file was downloaded from online cloudfront.net.
- Twitter API and Python's Tweepy library to gather each tweet's retweet count and favorite ("like") count at minimum. Applied Twitter API to get this data.

Step 2: Accessing Data

Once the data was gathered, I began to assess the data on both quality and tidiness issues. Here were the issues listed out (completeness & tidiness:

Quality Issues

The four main data quality dimensions are:

- Completeness: missing data?
- Validity: does the data make sense?
- Accuracy: inaccurate data? (wrong data can still show up as valid)
- Consistency: standardization?

Twitter archive table (twitter_archive)

* Keep original ratings (no retweets) that have images
* Drop columns not needed for our analysis
* Erroneous datatypes in these columns (tweet_id, rating_denominator,rating_numerator, in_reply_to_status_id, in_reply_to_user_id, timestamp, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, doggo, floofer, pupper, and puppo)
* Missing values in 'name' and dog stages represented as 'None'
* Error in dog names (e.g a,an,actually) are not a dog's name.
* Clean numerator as some line items have very high numerator ratings.

Image prediction table (image_prediction)

* Erroneous datatype (tweet_id)
* Missing images for some rows of data

* Create new column dog_breed

Twitter API Data (twitter_data)

* Missing data

Tidiness Issues

Three requirements for tidiness:

- Each variable forms a column
- Each observation forms a row
- Each type of observational unit forms a table

Twitter archive table (twitter_archive)

* The last four columns all relate to the same variable (dogoo, floofer, pupper, puppo)

Image prediction table (image_prediction)

* Image predictions table should be added to twitter archive table.
* Creating a new dog_breed column using the image prediction data

Twitter API Data (twitter_data)

* twitter api table columns(retweet_count, favorite_count, followers_count) should be added to twitter archive table.

Step 3: Cleaning Data

With that, i began to start the cleaning process of issues that i've accessed in step 2. Also, made a copy of all the tables before proceeding to do cleaning by doing the following:

1. Keep original ratings (no retweets) that have images" by deleting retweets by filtering the NaN of retweeted_status_user_id
2. Drop columns not needed for our analysis
3. Fix erroneous data types. Ie Convert tweet_id to str from twitter_archive, image_prediction, twitter_data tables.
   twitter_archive_clean.tweet_id = twitter_archive_clean.tweet_id.astype(str)
4. Fix Incorrect dog names
5. Created a function to check and clean dog stages column
6. Created a new dog_breed column using the image prediction data
7. Dropped tweets with no images
8. Clean Numerator. As some line items have very high numerator ratings.

Udacity Nanodegree - Project 4

Wrangle and Analyze Data | WeRateDogs Twitter

9. Moved twitter api table and image prediction table to twitter archive table. Merged tables.

Step 4: Storing Data

After cleaning the data, i then stored it as “twitter_archive_master.csv” file for analysis.

Links to colab notebook:
https://colab.research.google.com/drive/1jEHMNRY00lJp70aaM-CCAhgBgXi5L-Xt?usp=sharing