wrangle_report

February 9, 2023

0.1 WeRateDogs Data Analysis Project (wrangle_report)

This project focuses on the data wrangling part of data analysis, emphasizing its three aspects namely: data gathering, data assessing and data cleaning. Analyzing the dog ratings of the Twitter account "WeRateDogs", three data sets were used. These data were gathered using three methods, downloading the "twitter_archive_enhanced.csv" data set manually from the project resources, downloading the "Image-prediction" programmatically and querying Twitter API to get additional data for the retweet and favorite counts.

In gathering the "Image-prediction" dataset, the request library was used to get the file content from the given URL and it was saved as a TSV file. The third file was gotten by querying Twitter API tweepy, obtaining authentication using customer key and secret and access token and access secret. The retrieved information was saved as a txt file and read into a python data frame using the pd.read_json function.

For the data assessment phase, these data sets were assessed both visually by looking through the data sample and programmatically using pandas functions and methods. The ".info()" method was used to observe the columns and their data types, ".duplicated()" method to check for duplicates, ".query()" for filtering, ".isnull()" to check for null values, ".value_counts()" to return counts of unique values and other methods for assessment. The observation from the assessment was then documented as quality and tidiness issues. 10 quality issues and 2 tidiness issues were identified. These issues are listed below.

0.1.1 Quality Issues

- Twitter archive table

1. data type for time stamp is an object instead of date-time format
2. Some rating denominators are not 10
3. Some rating numerators exceed normal ratings (e.g. 420, 170, 50, 1776, 27, 60)
4. Null values in 'expanded_urls' column hence, some tweets do not have images and some values are duplicated
5. Retweet column indicates tweets that are not original
6. 'name' column contains some wrong names (e.g. a, an, such)
7. Redundant source, retweet, in_reply and rating_denominator columns

- Image prediction table

8. Some data samples are not dogs (hence column P1 are not dog names and P1_dog returns false)
9. Selecting needed columns from the table

- Tweet table

10. Selecting needed columns

0.1.2 Tidiness Issues

1. Redundant columns [doggo floofer Tits Puppo] should form a column 'stage'
2. Column names need to be renamed to describe the observation better and make it understandable

For the third part of the data wrangling process, cleaning, the first step was to create a quality copy of the data frames. The Define, Code, and Test technique was then used to fix each issue, defining the issue to be fixed, writing a block/blocks of python code to fix it, and then testing if the problem has been corrected. After the cleaning was done, the resulting clean data frames were then merged into a single master data frame, "twitter_archive_master.csv", which was then used for the analysis and visualiation of the project.

For the analysis and the visualization phase of this report, a few insights were highlighted. These are:

1. Most popular dog breeds
2. Most popular dog stages
3. Dog breeds with the highest number of retweet count and favorite count
4. Dog stages with the highest number of retweet count and favorite count
5. Top rated dog breeds
6. Highest rated dog breeds
7. Common dog names

Finally, these questions were answered and two of the insights (Most popular dog stages and Most loved dog stages) were visualized.