Project WeRateDogs Data Wrangle Report

This is Cassi of breed doggo rated 14/10

Introduction

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

Project Steps Overview

The main task for this project was to:

1. Gather the data
2. Assess the data
3. Clean the data
4. Store the data either as a .csv or SQLite database
5. Analyzing, and visualizing data
6. Reporting

The focus for this document will be from step 1 to 4, I am going to take you through the gathering process, assessing process, cleaning process and storing the data.

Gathering Data

In this step, data was gathered from three different sources and in different file formats.

Data 1 - Enhanced Twitter Archive: This contains basic tweet data for all 5000+ of their tweets, but not everything. Additional data is still needed to create interesting and trustworthy analyses and visualizations. This file came in a .csv file format (twitter-archive-enhanced.csv). The data was downloaded manually, uploaded into the project root directory and then read into a pandas Dataframe using the .read_csv function.

Data 2 - The tweet image predictions: The second step of gathering the data needed for this project was to programmatically download the tweet image prediction which is in .tsv file format (image_predictions.tsv) using the Requests library and then read it into a Pandas Dataframe.

Data 3 - The Additional data from the Twitter API: The third step of gathering the data was to use the Tweepy library to query the retweet count and the fovarite count from twitter, query using the tweet id from twitter-archive-enhanced.csv so as to get the data for those tweet ids only, write each data into a .txt file format (tweet_json.txt) and then read the data line by line into a pandas Dataframe.

Assessing Data

All the three gathered data were assessed in two formats, visual and programmatic assessment, the purpose of this step was to spot and highlight at least eight (8) quality issues and two (2) tidiness issues.

- **Visually Assessment:** In this step, all the three data were displayed in a pandas dataframe by calling the name of each dataframe, also, other tools like spreadsheet, text editor etc. were also used to access the data visually. Missing values were easily seen and some inconsistent input in some columns
- **Programmatic Assessment:** Here some pandas methods/functions were used, such as .info(), .value_counts(), .duplicated() etc. the data were assessed one after the other and on each data, quality and tidiness issues were documented.

At the end of assessing the data, the below quality and tidiness issues were gathered.

Quality issues

- **df_archive DataFrame**

1. Change tweet_id from int to string
2. rating_numerator should be correctly extracted and should be of datatype float
3. Remove 181 retweets and 78 replies which are not needed
4. Drop in_reply_to_status_id and in_reply_to_user_id column, it will not be needed for analysis
5. timestamp should be datetime datatype
6. Drop retweeted_status _id, _user_id and _timestamp since we are not interested in retweets only real tweets
7. Drop expanded_urls
8. Clean rows in name columns that are likely not a dog name e.g his, a, an, the ect.
9. Replace None in doggo, floofer, pupper and puppo column as null
10. Replace None in name column as null
11. Drop all rows that have no dog name in the name column (all NaN)
12. Rename name column to more descriptive name - dog_name

- **df_image_prediction DataFrame**

13. tweet_id should be string

14. Drop duplicated jpg_url cells
15. Drop some predictions that are not of breed dog which returns false
16. Fix inconsistency upper and lower case letters in p1, p2, and p3 columns

Tidiness issues

1. Combine doggo, floofer, pupper and puppo into one column called dog_stage (in df_archive DataFrame)
2. Combine all 3 Dataset into one DataFrame

Cleaning Data

The third step in the wrangle process is to clean the data; in this section, all the quality and tidiness issues found during the assessing process were cleaned using pandas methods/function. First, copies of the original data were made before cleaning, also, the rules of tidy data were put into consideration. Finally, the result of the cleaned high-quality and tidy data was stored in a master pandas DataFrame (twitter_archive_master.csv).

Storing Data

The data was stored in a .csv file format called **twitter_archive_master.csv** using the .to_csv method in pandas.