Sanmi Ayeni

Introduction

The tweet archive of the Twitter user @dog rates, commonly known as WeRateDogs, is the subject of this dataset. This account rate people's dogs and comments accordingly. To start the data wrangling and analysis process, the required datasets will be gathered using the request library and via the Twitter API, then assessed visually and programmatically. The datasets will then be cleaned to remove unclean and untidy data. To produce the necessary results, an exploratory data analysis (EDA) will be conducted using matplotlib and seaborn.

Part One

1.0 Gathering Data

Three datasets are needed for this project and were gathered by different means. The *twitter_archived_enhanced.csv* file was downloaded manually followed by the *image_prediction.tsv* file. The CSV file was gotten using the request function in python to download the file programmatically. Lastly, the *tweet_ json.txt* file was gotten using Twitter API, this process can be regarded as web scraping. All files were loaded on jupyter lab and saved to a local drive for assessment.

2.0 Assessing Data

The assessment of the three datasets was done to highlight quality and tidiness issues in the dataset. The following issues were noticed.

Data Quality issues identified while assessing data

- The timestamp in df_twitter has the wrong data type and should be DateTime
- 181 retweets, indicating duplicated rows
- Missing values in the name column appearing as 'None'.
- Invalid names like a, an, by, getting, his, incredibly, infuriating, officially, quite, space, such, the this, unacceptable and very
- The names wrongly spelt like (billl, carll, Jennifur, klevin, paull, samsom, shawwn, traviss zooey)
- 55 users with invalid name 'a'
- Inconsistency in rating_numerator and rating denominator
- The three data frames do not have the same shape
- Image prediction data does not have descriptive columns names

Data Tidiness Issues identified while assessing the data

- The doggo, floofer, pupper and puppo should be categorical data
- The archived data and Twitter API data should be merged

3.0 Cleaning Data

A copy of each dataset was created to avoid tampering with the original data. furthermore, the data cleaning process involved three stages namely:

Define: establishes how the data will be cleaned

Code: Converts the defined steps to code

Test: Test the result of the code to confirm if it was properly implemented.

Issues with data quality and tidiness, including inaccurate/invalid names, missing data, improper data types, etc., were fixed. After the cleaning process, the three datasets were merged into one for easier analysis and visualisation.