Data Wrangling Report

Introduction

This is a data wrangle report that demonstrates the data wrangling process for the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. 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.

Gathering Data

The data that was gathered was downloaded from three sources and stored in three separate files:
The first dataset which is the twitter_archive dataset was downloaded manually.

The second dataset was downloaded programmatically through the url that was given,we made use of the python request library to extract it..

Additional data from Twitter API were downloaded from querying the Twitter API for each tweets Json data using python tweepy library and then each tweets entire set of Json data was stored in a file tweet_json_txt file. Each tweet json data was written to its own line. Then read the json txt file line by line into a pandas DataFrame with tweet_id,retweet_count and favorite_count

Assessing Data

THe three files we obtained from the gathering stage were loaded into pandas dataframe individually for assessment.

Each of the DataFrame was assessed both visually and programmatically.

Visual assessment involves visually assessing it by getting the first five rows of the dataset and scrolling through.

Programmatically involved us using code to get a more efficient assessment of the dataset
.
The following quality and Tidiness issues were identified below

Quality Issues:

1.Wrong datatype of the timestamp column

2. Wrong datatype of tweet_id in both the the twitter_archive and image_predictions dataset

3. The rating_denominator columns should be 10.0 or greater than

4. We remove retweets as we want original ratings that have images, and drop the columns that are irrelevant to the analysis

5. Text column rename to tweets

6. The name column has different values such as 'a' and 'None' and other lowercase names

7. The name column is irregular where both lowercase and uppercase are present.

8. The column names such as p1,p2,p1_conf,p2_conf,p3_conf,p1_dog,p2_dog,p3_dog are not descriptive enough and drop the unwanted column 'img_num' in the image prediction dataset

9. Replace the None values with NaN

TIDINESS ISSUES:

2.have a new column 'stage' that shows all types of dog stage(doggo,floofer,pupper,puppo) and drop the initial four columns of doggo,floofer,pupper,puppo.

3. Merge the three datasets into one big dataset.

Clean Data

Firstly copies of the three datasets were made

The datasets were cleaned individually using programmatically techniques

We made use of Define ,Code and Test

- Dropping the columns that are not necessary for the analysis
- Converting the wrong datatypes to the correct datatype format
- Renaming the columns that are not well descriptive
- Standardize the rating numerator to 10.0 or greater than
- Creating new columns for year month and day
- Replacing the None values with NaN

-
- Finally combining the three datasets into a single DataFrame

Storing the cleaned data

The dataset is now cleaned and ready for the analysis and saved as master_cleaned csv file