WRANGLE REPORT: WeRateDogs

Table of Contents

1. Introduction

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dogs. WeRateDogs has provided the dataset to be wrangled and analyzed.

2. Gathering

I gathered three datasets i.e.

a) Enhanced Twitter archive data
The data set was provided by WeRatesDogs to be downloaded directly. I downloaded the CSV file end credits into a data frame using pandas.

b) Image predictions data
I downloaded the file programmatically. I made a directory to store the dataset. I used the requests library to read data from the URL here and wrote it into a file using the os library. I uploaded and read the file into a data frame using pandas

c) Additional data from the Twitter API
I used Tweepy to query tweet as API for data in the WeRateDogs Twitter archive. To do this, I required consumer and access tokens Which I got from Twitter after applying for a developer account. I queried each tweet ID as in the enhanced Twitter archive data. I wrote the JSON data to a tweet_json.txt file. I read the data from the tweet_json.txt file into a Python list using the loads method in the JSON library. I then read the data into a pandas data frame and converted it into a CSV file.

3. Assessing

I assessed the gathered data both programmatically and manually. I accessed for both quality and tidiness issues.

Quality issues our issues with the content of the data i.e. validity, completeness and consistency of the data. Tidiness issues are issues with the structure of the data.

Quality issues

1. Enhanced Twitter archive data

- There are retweets in the dataset.
- Incorrect data type i.e. timestamp is an object.
- Denominator ratings are more than 10 in some rows.
- Incorrect dog_name values such as a, an

2. Image predictions data
    - Duplicate jpg_url values.
    - Inconsistent first letters in p1, p2 and p3 columns

3. Columns that are not needed for analysis.

Tidiness issues

1. Two variables in one column i.e. timestamp.
2. Dog types are in different columns.
3. All the three datasets are to be part of one table.

4. Cleaning

After assessing the datasets, I created a copy of each and cleaned them.

Quality issues

1. I deleted retweets and kept only original tweets.
2. I converted the timestamp data type from an object to add date time.
3. I dropped the rows with rating_denominator above 10.
4. I replaced the incorrect dog_name values with NaN.
5. I deleted the duplicate jpg_url values.
6. I capitalized the first letters in p1, p2 and p3 columns.
7. I dropped the columns that were not needed for analysis.

Tidiness issues

1. I created new columns (date and time) and extracted date and time values from the time stamp column into their respective new columns.
2. I created a new column dog_type and extracted the dog type values from their columns.
3. I merged the three datasets on the column tweet ID.

5. Storing

I saved the Clean Master data frame to a CSV file.