wrangle_report

August 23, 2020

1 Project: Wrangle and Analyze "WeRateDogs"

1.1 I. Introduction

In my report I will explain the wrangling process involved by using the dataset provided by Twitter account, "WeRateDogs."

The Data wraning process will include the following;

1. Gathering Data from source.
2. Asessing and evaluating data.
3. Cleaning the data by checkng for any quality issues.

1.2 II. Gathering Data

I begin my wrangling process by gathering the three different datasets, provided to me, downloaded and extracted all contents of each zipped file.

The first file I manually downloaded, *twitter-archive-enhanced.csv* was not very valuable to use because, upon searching for a tweet_id after using tweepy.api, most of the id's in the list returned no existing page, or status. I believe we only successfully found around 8-10 pages of existing accounts. For the second file, *image-predictions.tsv* I saw the file came in as tab delimited. I seperated each column with a tab and loaded the dataset into a pandas dataframe. I saved the ids into a list then called Twitter's API to ge the status of each, individual id and then I wrote it back into a file, which was dynamically named *tweet_json.txt*. I created a dataframe and used the .merge() function, to merge the initial dataframe(image_predictioons) into the new dataframe(tweet_status_info_filtered). Within df_merged I used a callable function, *www.mathguide.de API* to know the meaning of the language code.

1.3 III. Data Assessment

One of the assessments I did was look up the information of the dataframe. This gave me information like the count of all rows, check if the rows had any null-types or not, and we aslo looked for thed dataytpe. After viewing this information I begin using the .unique() function on each column to analyze the values in each field. I also checked for NaN values on each field and checked the scale for each numerical field.

1.4 IV. Data Cleaning and Quality Check

1.4.1 Data Quality issue 1.

I extracted the dog rating from the text using a regular expression because the rating values was hidden within the full text of the tweet to perform some analysis. I then, created a column to hold the rating from *WeRateDogs*. I moved on to then create new columns to hold the rating as an integer which was used for analysis later on...

1.4.2 Data Quality issue 2

I renamed the field holding the language of each twitter to something meaningful and readable. My plan was to get the full meaning of each langauge code to make the data more readable.

1.4.3 Data Quality issue 3

I noticed that **favorite_count** and **retweet_count** data type were float, so I perceived in converting the datatypes to integers.

1.4.4 Data Quality issue 4

I extracted date and time from **created_at**, which helped me to be able to analyze data and group them by year.

1.4.5 Data Quality issue 5

I removed hyper links from the text beause I noticed that most of the links where bad and also would not add any value to my analysis.

1.4.6 Data Quality issue 6

I noticed that some rating were higher than the rating scale of 100 percent; this presented some anomalies on the dataset. I had to reset the rating higher than 100, back to 100%, to avoid outliers.

1.4.7 Data Quality issues 7.

I looked for duplicate values on **tweet_id** and **jpg_url** and dropped the records tha had duplicate values.

1.4.8 Data Quality issues 8.

Lastly, I dropped all records with pictures that had a probability of not being a dog.

```
In [ ]:
```