Data Wrangling Report prepared by Caleb Omariba

Udacity Data wrangling of datasets from “WeRateDogs” tweets.

INTRODUCTION

The WeRateDogs which is the source of our data is a Twitter account that rates people’s dogs with humorous comments about the dog. The account was started in 2015 by a college student Matt Nelson. This account has attracted 8.8 million+ followers ever since.

Data wrangling steps in this project:

Step 1: Gathering data
Step 2: Assessing data
Step 3: Cleaning data

Gathering data.

The project required me to gather data from three different sources.

a) I downloaded one dataset given by Udacity I.e (twitter-archive-enhanced.csv)which I successfully loaded. Which i named (twits_arc_df)
b) Programmatically downloaded (`image_predictions.tsv`) using the Requests library. Which i named (predictions_df)
c) I used Tweepy, a Python library, to query Twitter’s API for WeRateDogs Twitter data whereby i create a table called (api_tweets_df)

Assessing data

In the assessment process, I found some issues and documented it in two parts, the **quality issues** and the **tidiness issues**

Quality issues observed.

twits_arc_df

- **in_reply_to_status_id**, **in_reply_to_user_id**, **retweeted_status_id**, **retweeted_status_user_id**, **retweeted_status_timestamp** variables have a lot of missing data and, moreover, we do not need them for the analysis.
- The **timestamp** column datatype is wrongly string data type.
- The standard **rating_numerators** range from 0 to a maximum of 15 the rest of the values are considered as outliers.
- 10 is the standard **rating_denominator** that is used in the WeRateDogs handle. The rest of the values are incorrect.
- We can observe that the **majority of Dog names start with a capital letter** so those that do not are wrong.

predictions_df

- The **column names in the predictions_df are not meaningful** to be understood clearly.
- we have **66 duplicates in the jpg_url column**
- The **dog breeds' names have some names beginning with uppercase letters while others begin with lowercase letters** which is not consistent.

Tidiness Issues.

twits_arc_df

- we have **4 columns for the dog stages which is wrong.**

api_tweets_df

- *The api_tweets_df table should be combined with the twits_arc_df table.*

Cleaning data

1. Dropped all columns with a lot of missing data which will not be helpful in the analysis process. I used the drop() method to remove the specified redundant values.

2. Changed the timestamp to datetime type data type. I used the pd.to_datetime() function to the strings to datetime data type.

3. Drop rating_numerator that are >15. Since the majority values of the rating_numeratir were less than 15 the ones more than 15 were considered as outliers.

4. Dropped the rating_denominator column since it will not be useful because it is a constant that is all the denominator values are to be 10 though some are not 10.

5. Converted the first letter of the dog names to capital for all the dogs names to make them consistent.

6.Ranamed the column titles of the predictions_df to meaningful titles.

7.Dropped the duplicate values from the jpg_url column in the predictions_df.

8.Converted all dog names to begin with uppercase letters in the predictions_df.

9. When handling the tidiness issues I used the melt function Melting the 4 dog stages into one column.

10.Merged twits_arc_df and api_tweets_df tables together using the merge function and finally I joined the two with the predictions_df to form the final master dataset table that I used for analysis and visualizations.