Data Wrangling Report

Done by

Nahayo Gilbert

July 8, 2022

Wrangle Report

This project is about analyzing the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. This Twitter account rates people's dogs with a humorous comment about the dog and it has over 4 million followers and has received international media coverage. This report will cover all the wrangling steps carried out include gathering, assessing, cleaning, and storing the combined three datasets into single master data frame.

Gathering Data

In this project, three different sources were used to gather data which include:

1. WeRateDogs Twitter archive which was given to us but did not have all the required data to complete the project.

2. Tweet image predictions file which was downloaded programmatically from url given to us.

3. Twitter json file which has missing data from the twitter archive, the file was also given to us due to complexity of getting Twiter's developer account and fetching data directly using API. Codes to download Data using the API is included in the notebook but commented out.

Assessing Data

The project focused mainly on at least eight quality and two tidiness issues.

Quality Issue

Below are quality issues found in all three data sets.

1. Columns in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp,and expanded_urls have empty columns,
2. Source column in archive_df contain irrelevent values
3. rating_numerator and rating_denominator should be float not int format
4. Timestamp should should be date and time format
5. name from archive_df values are not irrelevent and wrong capitalization
6. Columns p1, p2, and p3 from prediction_df dataset are wrong capitalization
7. Remove Rt from the text variable
8. tweet id in all three data set should be object

Tidiness Issue:

Below are tidiness issues found in the twitter archive and image predictions data sets.

9. In prediction_df data frame, columns doggo, pupper, flooter, and puppo should be one single column, similarly p1, p2, and p3 shoulb also be one column containing all those informatiom.
10. A single observational unit is stored in multiple tables

Cleaning Data

After identifying above issues, Different approaches and techniques were used to clean them and few of techniques include:

1. Removing Retweets from columns.
2. Changing different datatpyes for better visualization.
3. Merging datasets into a single table.

Storing Data

Finally, all cleaned data sets were merged into a single master file and was saved to be used for visualization.