wrangle_report

June 24, 2022

1 Reporting: Data_Wrangling_Project_report

1.1 Introduction

We are performing Data Wrangling on 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 comments about the dog.

This data have some quality and tidiness issues that needs to be identified and clean by following the data wrangling process for the data to be fit for analysis.

At the end of the project: 1. Identify at least 8 quality issues and 3 tidiness issues 2. Clean the identified issues 3. Perform some analysis and bring out at least 3 insights and 1 visualization from the cleaned data.

1.1.1 Gathering Data

The first step of data wrangling is to gather the required data to be cleaned. This involves collating the data from different sources either manually or programmatically.

For this project, we used 3 separate data, each of which was gathered separately.

1. **Twitter_archive_enhanced data**: This data was downloaded manually, upload to the project workspace, and read into pandas DataFrame.
2. **The tweet image_predictions**: This data was gathered programmatically by using the Request library and the given URL. It is a .tsv file. It was downloaded programmatically into the workspace and read into pandas DataFrame with tabular separation ()͡
3. **Tweet_json data**: This required querying Twitter API data using the Tweepy library and stored in tweet_json.txt. > I used the provided `twitter_api.py` code and `tweet_json.txt` due to my inability to secure access to the Twitter API data after waiting for a day. So, I used the provided one and I opened and read the JSON text file to extract the needed column (tweet_id, favorite_count, and retweet count) to be stored in a pandas dataframe called 'tweet_json_df

1.1.2 Assessing Data

I assessed the 3 data visually and programmatically using `(.info(), .describe(), .duplicated(), .isna(), .value_counts())` methods.

After assessing the data, the following quality and tidiness issues were identified:

Quality issues twitter_archive_enhanced

1. The null value on the doggo, floofer, pupper, and puppo column are recorded as none
2. Name of some dogs were not properly captured or not captured at all as some dogs were given names such as a, an, such, quite. Other unavailable names are recorded as None
3. There are 181 retweets data in the datasets to be removed
4. Some columns are not useful for analysis, most of it even contains large missing data (source, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded url)
5. The rating denominator has some values not equal to 10 as 10 is the standard denominator
6. The rating numerator has some extreme values which seems unrealistic, like (420, 666, 182, 1776)
7. Other rating numerator were incorrectly captured due to having multiples of the numerators in tens or the decimal part of a decimal ratings
8. Timestamp was recorded as object instead of datetime datatype and tweet_id is int instead of string

image_predictions

9. tweet_id is int instead of string
10. jpg_url has 66 duplicated values
11. The predicted dog breed is not well defined, someone can struggle to know what p1, p2, or p3 is
12. The dog breed should have a consistence case format so that repeated breed can be capture without issue

tweet_json_df

13. tweet_id is int instead of string

Tidiness issues

1. **twitter_archive_enhanced** - The classes of the dogs (doggo, floofer, pupper, puppo) was recorded in separate column instead of one column for the class of dogs and some dogs have more than one classification
2. **image_predictions** - Prediction 1 has the highest confidence level, then we can drop other predictions with their correspondence.
3. The three datasets can be combined into one data.

1.1.3 Cleaning Data

I performed cleaning following the established way of cleaning identified issues by defining, coding, and testing. Each of the issues was properly addressed with defined cleaning methods and properly tested.

After cleaning all the quality and tidiness issues, the 3 datasets was merged and stored in `twittwe_archive_master` in a CSV file. This master data was later used for analysis and visualization.