Data Wrangling Report

Abdullah Alhussain

18/11/2022

Introduction:

This project aims to analyze data for a Twitter account called “WeRateDogs @dog_rates”. In this account, dogs are rated with humorous comments, and although the rating denominator is 10, the rating numerator is usually higher than 10, e.g., 11 and 12. These funny facts about this account are some of the reasons behind its popularity.

In this report, the wrangling effort of the data collected from this account is discussed. The wrangling process is mainly three steps: data gathering, data assessment and data cleaning.

Data Gathering:

Originally, we have the data in three files, which are:

- 'twitter-archive-enhanced.csv'
- 'image_predictions.tsv'
- ‘tweet-json.txt’

The three data files were opened directly using Pandas after getting them.

Data Assessment:

After assessing the quality of the data, many issues were found. In this report, only eight quality issues and two tidiness issues are discussed.

Tidiness Issues:

1- Dataset is separated into three datasets.
2- Dataset contains four columns for dog stages.

Quality Issues:

1- Dataset contains retweets.
2- Dataset contains useless columns.
3- The column ‘source’ is not readable.
4- The column 'tweet_id' has inappropriate datatype.
5- The column 'timestamp' has inappropriate datatype and unnecessary details.
6- There are errors in data extraction and improper values in the data.
7- The dataset contains tweets missing images.
8- Wrong datatype for categorical columns.

Data Cleaning:

In this part, the quality and tidiness issues found in the data assessment steps are cleaned. For the code implementation, kindly refer to the Jupyter Notebook ‘wrangle_act.ipynb’.

Issue #1: Dataset is separated into three datasets (Tidiness Issue)

As described in the data gathering part, the dataset comes from three different data files (sources), and this fact is a tidiness issue that needs to be cleaned because one data frame is enough, no need for three data frames.

Action:

Merge the three data frames into a single data frame based on the ‘tweet_id’ column.

Issue #2: Dataset contains four columns for dog stages (Tidiness Issue)

The dataset contains four columns for dog stages which are 'doggo', 'floofer', 'pupper', and 'puppo'. The dog stage is a categorical variable, and all stages are different possible values for a single variable. Therefore, one column is enough, no need for four columns.

Action:

Combine the four columns 'doggo', 'floofer', 'pupper', and 'puppo' into a single column.

Issue #3: Dataset contains retweets (Quality Issue)

The dataset contains tweets that are retweets (duplicated tweets). These tweets need to be removed from the data to improve the quality of the data and analysis.

Action:

Drop rows containing retweets.

Issue #4: Dataset contains useless columns (Quality Issue)

There are columns that are useless because they contain too many missing values, or because they will not be used in our analysis. These columns are:

- 'in_reply_to_status_id'
- 'in_reply_to_user_id'
- 'retweeted_status_id’
- 'retweeted_status_user_id'
- 'retweeted_status_timestamp'
- 'p1'
- 'p1_conf'
- 'p1_dog'
- 'p2'
- 'p2_conf'
- 'p2_dog'
- 'p3'
- 'p3_conf'

- 'p3_dog'

Action:

Drop useless columns that contain too many missing values, and columns that will not be used in our analysis.

Issue #5: The column ‘source’ is not readable (Quality Issue)

The ‘source’ column has HTML structure that has an impact on the readability of the column.

Action:

Remove HTML structure from 'source' column to make it readable.

Issue #6: The column 'tweet_id' has inappropriate datatype (Quality Issue)

The column ‘tweet_id’ has an integer datatype. This datatype is not appropriate because ‘tweet_id’ is nominal data. The ‘tweet_id’ is like phone number, national ID number, badge number, student ID number, etc. Usually, we are not interested in performing mathematical operations on them.

Action:

Change datatype of 'tweet_id' column from integer to string.

Issue #7: The column 'timestamp' has inappropriate datatype and unnecessary details (Quality Issue)

The ‘timestamp’ column has an object datatype. Transforming this column to ‘datetime’ datatype will make dealing with this it easier. Furthermore, hour of the day, minutes and seconds are not necessary for our analysis. Therefore, removing these details will make the column much cleaner.

Action:

Change datatype of 'timestamp' column from object to datetime. Also, extract date information only, i.e., year, month, and day, from 'timestamp' and filter other details, i.e., hour, minute and second.

Issue #8: Errors in data extraction and improper values in the data (Quality Issue)

Some numerators for the ratings were not extracted successfully. Furthermore, some denominators have improper values, i.e., zeroes.

Action:

Extract numerators and denominators from text and drop rows where denominator is equal to zero.

Issue #9: The dataset contains tweets missing images (Quality Issue)

Some tweets do not have images. So, this issue shall be cleaned.

Action:

Remove tweets that do not have images from the dataset.

Issue #10: Wrong datatype for categorical columns (Quality Issue)

Two categorical columns ‘dog_stage’ and ‘source’ have wrong datatype, which is object.

Action:

Convert the datatype of the columns 'dog_stage' and 'source' from object to category.

Conclusion:

In this report, we have discussed the data wrangling process for the data collected from the Twitter account “WeRateDogs @dog_rates”. The process has three main steps which are: Data Gathering, Data Assessment, and Data Cleaning. First, in the data gathering step, the data was gathered from three sources. After data is gathered, eight quality issues and two tidiness issues were found in the data in the data assessment step. Finally, these quality and tidiness issues were solved in the data cleaning step.