We Rate Dogs:- Wrangle Report

Introduction

The process of handling real-world data is often complex and requires the use of Python and its libraries. Through this project, we have gathered data from various sources and in different formats. Our objective has been to evaluate the quality and cleanliness of the data, followed by its cleaning process, which is commonly known as data wrangling.

This report aims to provide a summary of the work undertaken in this project. The dataset under examination pertains to the tweet archive of a well-known Twitter user, @dog_rates, also referred to as WeRateDogs. WeRateDogs is a popular Twitter account that assigns ratings to dogs posted by users, accompanied by clever and humorous comments. Notably, these ratings consistently employ a denominator of 10, adhering to a standardized scale. However, the numerators associated with the ratings often exceed 10. For instance, it is common to see ratings such as 11/10, 12/10, 13/10, and so on. The account humorously justifies these inflated ratings with the catchphrase "they're good dogs Brent." WeRateDogs has amassed a substantial following of over 4 million users and has gained considerable attention from media outlets worldwide. The analysis conducted on the Twitter archive, encompassing all tweets until August 1, 2017, has yielded several findings, which are presented in this report.

Steps followed:

1. Gathering Data
2. Assesing the data
3. Cleaning the data

1. Gathering Data

In this project, we have collected three different sets of data:

1. The WeRateDogs Twitter archive: This dataset was provided by Udacity and can be accessed through a downloadable link. It contains the Twitter archive of WeRateDogs, which includes various information about the tweets, such as text, timestamp, and rating.

2. Tweet image predictions: This dataset, named "image_predictions.tsv," is hosted on Udacity's servers. To obtain this file, we programmatically download it using the Requests library. The file contains predictions of what breed of dog or other objects are present in each tweet image. These predictions were generated by a neural network.

3. Retweet and favorite counts: To gather the retweet count and favorite (like) count for each tweet in the WeRateDogs Twitter archive, we utilize the Twitter API. By querying the API using Python's Tweepy library and the tweet IDs from the archive, we can retrieve the JSON data for each tweet. We store this data in a file called "tweet_json.txt," with each tweet's JSON data written on its own line. Finally, we read this file line by line into a pandas DataFrame, extracting the tweet ID, retweet count, and favorite count.

It is important to note that the WeRateDogs Twitter archive, tweet image predictions, and retweet/favorite counts are separate datasets that are combined and analyzed in this project to gain insights about the tweets and their associated information.

2. Assesing Data

After gathering the data, the next step involves evaluating the collected data for quality and tidiness issues. We conducted both visual and programmatic assessments to identify and document any problems that may affect the data's content and structure. By examining the data visually and programmatically, we aimed to ensure a comprehensive review, minimizing the chances of overlooking any issues and facilitating easy access to the identified problems. The assessment focused on two main types of issues:

1. Quality Issues: These pertain to problems with the data content, commonly referred to as "dirty data." We aimed to identify and document any instances of low-quality data that could impact the accuracy and reliability of the analysis.

2. Tidiness Issues: These relate to structural problems that hinder straightforward analysis, often referred to as "messy data." Our objective was to evaluate the data's structure and verify if it met the criteria for tidy data, which include having each variable represented as a column, each observation as a row, and each type of observational unit as a separate table.

We conducted a visual assessment by scrolling through the data in our Jupyter Notebook, examining its overall appearance and identifying any visible irregularities. Additionally, we performed a programmatic assessment using code to obtain specific subsets and summaries of the data, enabling a more systematic and detailed analysis. Based on these assessments, we documented all the quality and tidiness issues that were observed.

2.1. Quality of Data

2.1.1. Twitter_archive_enhanced table

- The columns "name," "doggo," "floofer," "pupper," and "puppo" contain entries labeled as "None," indicating missing or unspecified data.
- The values in the "source" column are encoded in a format that is not easily interpretable by humans, requiring further processing or decoding.
- The "name" column includes invalid or unconventional entries such as "such," "quite," "a," and "an," which may need to be addressed for data accuracy.
- In certain instances, the "rating_denominator" column deviates from the assumed standard value of 10, suggesting the presence of non-standard or inconsistent data.
- The "rating_numerator" column contains unusually high values, resulting in exceptionally high ratings, which could potentially indicate inaccuracies or anomalies.
- Six instances of the "rating_numerator" column contain incorrectly entered values, warranting further investigation or cleaning.
- Some tweets in the dataset are retweets of posts from the Twitter account @dog_rates, associated with the We Rate Dogs platform.

2.1.2. Image Predictions table

- Inconsistencies are observed in the usage of underscores and capitalization in the values of the p1, p2, and p3 columns.
- The column names are not adequately labeled, which hinders the comprehension of their intended significance and function.
- Duplicated image predictions are found, indicated by distinct tweet IDs but identical jpg_urls. The remaining data in these instances remains unchanged.

2.1.3. Tweets json table

- The column containing tweet identifiers is labeled as "tweet_id" in this particular scenario, whereas it might be referred to as "id" in other situations. It is important to highlight that there are cases where the values of retweet_count and favorite_count seem to be replicated in specific rows, which raises concerns about the reliability or uniqueness of these values.

2.2. Tidiness of Data

- The dataset consists of a variable named "dog_comb_stage" that spans four separate columns: doggo, floofer, pupper, and puppo. These columns serve to categorize different stages in a dog's life, and the distribution of the "dog_comb_stage" variable is divided among these four columns.
- The goal of this task was to merge two tables with the purpose of combining the information and data they hold. By performing this merging operation, a unified view can be attained by consolidating the pertinent data from both sources. This consolidation process facilitates a more comprehensive analysis and comprehension of the data.

3. Cleaning Data

To ensure optimal efficiency and prevent the need for redundant code in the future, a systematic approach will be taken in this section. The approach will encompass the following key aspects:

1. Handling Missing Data: Identifying any instances of missing data and implementing suitable strategies to effectively address them.

2. Resolving Tidiness Concerns: Rectifying any concerns related to the organization and structure of the data, ensuring it is restructured in a coherent and standardized manner.

3. Addressing Additional Quality Matters: Thoroughly examining and resolving various quality issues, such as inconsistencies, inaccuracies, or anomalies, to uphold the overall integrity and reliability of the data.

By methodically addressing each of these areas, we can effectively tackle the identified issues and enhance the overall quality and usability of the data without encountering any repetitive code in the future.

To ensure a streamlined and efficient approach in this section, a systematic plan will be implemented, encompassing the following key aspects:

1. Handling Missing Data: A thorough analysis will be conducted to identify any missing data present in the dataset. Suitable strategies and techniques will then be employed to effectively address and manage these missing values.

2. Resolving Tidiness Issues: The dataset will be carefully examined to identify any tidiness issues that may hinder its organization and coherence. By reorganizing and restructuring the data in a standardized and logical format, these tidiness issues will be resolved.

3. Addressing other Data Quality Concerns: A comprehensive evaluation will be undertaken to identify and address various data quality concerns. Inconsistencies, inaccuracies, and anomalies within the dataset will be thoroughly examined and rectified to ensure the integrity and reliability of the data.

By following this systematic plan, the identified issues within the dataset can be systematically addressed and resolved, thereby enhancing the overall quality and usability of the data for subsequent analyses and insights.

3.1. Handling missing Data

3.1.1. Several columns in the dataset contain missing values, including in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, and expanded_urls.

Definition: To optimize the dataset, specific columns such as in_reply_to_status_id, in_reply_to_user_id, retweeted_status_timestamp, and retweeted_status_user_id are omitted from the analysis. In the tweet_data_clean table, the missing values in the expanded_urls column are resolved by extracting the tweet_id, which corresponds to the final part of the tweet URL after "status/". This ensures that the expanded_urls column contains comprehensive and pertinent information that can be utilized for subsequent data processing.

3.2. Resolving Tidiness Issues

3.2.1. In the dataset, there is a single variable named "dog_stage" that is represented across four different columns: "doggo", "floofer", "pupper", and "puppo".

Definition: Merge the `doggo` , `floofer` , `pupper` , and `puppo` columns into a single column named `dog_status` . Once the `dog_status` column is created, we can remove the unnecessary individual columns associated with the different dog stages. This consolidation will streamline the dataset and enhance its clarity and organization.

3.2.2. There is a discrepancy in the title of the tweet ID. In some instances, it is referred to as "id," while in others, it is denoted as "tweet_id."

Definition: The renaming process involves changing the name of the `tweet_id` column in the `tweet_count_clean` dataset to `tweet_ids` .

3.2.3. Combining the tweet_cleaned and tweet_cnt_cleaned tables.

Definition: The merging process involves combining the `tweet_cleaned` and `tweet_cnt_cleaned` tables based on the matching values in the `tweet_id` column. By merging these tables, we create a single table that consolidates the information from both sources, allowing us to analyze and work with the combined data effectively.

3.3. Addressing other Data Quality Concerns

3.3.1. tweet_cleaned_comb_cnt table:

3.3.1.1. In the dataset, there are certain columns with erroneous data types that require attention. The `timestamp` and `retweeted_status_timestamp` columns need to be split into separate date and time components for better analysis. Additionally, the `dog_stage` column should be categorized correctly as it represents different stages of dogs. Furthermore, the columns `tweet_id` , `in_reply_to_status_id` , `in_reply_to_user_id` , `retweeted_status_id` , and `retweeted_status_user_id` should be of string data type to maintain consistency and facilitate appropriate data handling.

3.3.1.2. It has been observed that some tweets in the dataset are retweets of posts made by `@dog_rates` , which is the Twitter handle for the popular account called `We Rate Dogs` . These retweets can be identified and distinguished from the original tweets, indicating the engagement and influence of `We Rate Dogs` in the dataset.

Definition: Convert the `timestamp` column to the datetime data type and filter out the rows where the `tweet_id` matches the `retweeted_status_id` .

3.3.1.3. In some cases, the `rating_denominator` deviates from the assumed standard value of 10, which indicates the presence of inaccurate data.

Definition: The objective is to rectify and clean the inaccurate observations that have incorrect denominator ratings. The plan is to standardize the denominator rating to 10 and adjust the corresponding numerator rating to align with the revised denominator rating.

3.3.1.4. The `name` column contains invalid data entries such as "such", "quite", "a", and "an".

Definition: Validate and replace any occurrences of the "name" column containing invalid data with NaN.

3.3.1.5.. The columns 'name', 'doggo', 'floofer', 'pupper', and 'puppo' contain instances where the value is 'None'.

Definition: The columns 'doggo', 'floofer', 'pupper', and 'puppo' were combined into a single column called 'dog_stage'. However, the 'dog_stage' column contains several incorrectly entered data entries, including instances with 'None' as the value. Therefore, it is necessary to correct and reorganize the entire column to ensure accurate and consistent data.

3.3.1.6. The values in the source column are encoded and not easily understandable by humans.

Definition: Converting the values in the `source` column into easily understandable text format.

3.3.2. img_cleaned table:

3.3.2.1. The columns p1, p2, and p3 exhibit inconsistencies in their values. Instead of using spaces, underscores (_) are used to separate words. Additionally, the capitalization of the values varies, with some being in uppercase and others in lowercase.

Definition: To modify the variables `p1` , `p2` , and `p3` , we will replace any occurrences of `_` with a space character and convert the values to uppercase.

3.3.2.2.The column names lack clear descriptive labels.

Definition: Assign more meaningful and descriptive names to the columns.

3.3.2.3. Instances where the same image, identified by its jpg_url, appears to have multiple predictions associated with it, but with different tweet ids. The remaining data attributes remain unchanged across these duplicate predictions.

Definition: Remove the instances where the `image_url` is duplicated in the dataset.

References

1. Ignoring warnings
2. Merging Dataframes
3. Piechart
4. Null value imputations