Wrangle Report

Data wrangling is a crucial phase in any data analysis, and in this context, it involves several key steps. The aim of this document is to provide an in-depth overview of the data wrangling efforts undertaken in this project. Below, we outline the main steps involved in this process:

Gathering Data

The first step in data wrangling is gathering data from various sources. In our project, we have identified three primary sources:

1. The WeRateDogs Twitter Archive: This dataset contains essential tweet data related to the WeRateDogs Twitter account. To obtain this data, a manual download is required. It serves as the foundation for our analysis.
2. Tweet Image Predictions: This dataset contains predictions about the breed of dogs in the images associated with each tweet. To acquire this data, we utilize a programmatic approach for downloading and storing it. This dataset enhances our analysis by providing information about the dog breeds featured in the tweets.
3. Additional Data via the Twitter API: To enrich our dataset, we utilize the tweet IDs from the WeRateDogs Twitter archive to query the Twitter API. This allows us to gather supplementary information such as retweet count and favorite count, which are valuable metrics for assessing tweet popularity.

Assessing Data

With the data gathered, the next step involves assessing it for quality and tidiness issues. This assessment is conducted through a combination of visual inspection in a Jupyter Notebook and in excel sheets, and programmatic evaluation using pandas functions and methods. The primary objectives are to identify and document any discrepancies or anomalies in the data. During this process, we aim to pinpoint:

Quality issues

1. df1: extra data for the retweets, columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp.

2. df1: name column has the name 'a' repeated for 55 entries.
3. df1: name column has inconsistent capitalization.
4. df1: timestamp column type is object and should be timestamp.
5. df1: rating_denominator is larger than 10 for 19 tweets.
6. df1: expanded_urls column has 59 null values.
7. df1: some rating_denominator values are zero.
8. df2: p1, p2, p3 have inconsistent capitalization.

Tidiness issues

1. df1: doggo, floofer, pupper, puppo better to be in one column.
2. df3 better be merged with df2 and df1 to be as one master table.

Cleaning Data

This phase involves defining, coding, and thoroughly testing cleaning procedures to ensure data integrity and accuracy. It is crucial to create copies of the original data before making any changes to preserve the source information. All identified quality and tidiness issues have beeen resolved to prepare the data for analysis.

Storing, Analyzing, and Visualizing Data

Following the data cleaning process, the clean data is stored in a CSV format. Analysis and visualization are carried out using pandas and matplotlib within a Jupyter Notebook.

Insights:

1. Top 10 dogs favorite count
2. Top 10 dogs with highest rate
3. Top 10 breads favorite count

By following these structured steps, we ensure that our data wrangling efforts are systematic, transparent, and in line with project requirements and guidelines. This approach is essential for producing reliable and actionable insights for subsequent stages of the project.