

DATA WRANGLING REPORT

Data wrangling had five steps as follows:

Step 1: Gathering data.

In this step I started with “twitter-archive-enhanced.csv” data which is provided by Udacity. After that, I used the given URL to request image predictions to be downloaded programmatically in “image_predictions.tsv”. Final data is “tweets_info.txt” which is provided by Udacity, then I read its content and converted it from dictionary to data frame.

Step 2: Assessing data.

To assess these data, I used different methods, such as “.info()” to know general info about the data, “.describe()” to know the data statistics.

The observations that were found are as follows:

A. Quality

1. "tweet_id", "in_reply_to_status_id", "in_reply_to_user_id" have int/float data type.
2. "timestamp" has an object data type.
3. "retweeted_status_id", "retweeted_status_user_id" and "retweeted_status_timestamp" are unnecessary for this project.
4. Some of these names are incorrect, and we can notice they start with small letters.
5. "id" could be more clear.
6. "id" is not the first column.

B. Tidiness

1. Three datasets can be merged in one table.
2. Dog stages need to be combined into one column

Step 3: Cleaning data.

This step is important to make the data more readable and clear. In this step, I clean the data programmatically to rename some columns, drop others, change some types and merge data frames.

Step 4: Storing data.

After the cleaning process, I store the data into a new file.

Step 5: Analysis and Visualization

This is the final step in this project. I analyzed and visualized the cleaned data programmatically to find insights. The largest and smallest rated dog was found, the retweets counts and favorite counts were found by visualization.