REPORT

Data Wrangling: Gather, Assess and Clean

By: Nacer KROUDIR

2023

Data Wrangling:

The purpose of this project is to put into practice the art of data manipulation using real-world data. The process of data manipulation is divided into three parts:

- Data Gathering.
- Data Assessing.
- Data Cleaning.

The data used for this project is the tweet history of the Twitter user *@dog_rates*, also known as **WeRateDogs**. **WeRateDogs** is a Twitter account that evaluates dogs and adds a humorous comment. The first step is to gather data from various sources in various forms. Then, we will evaluate the data through both visual and technical methods to pinpoint any issues with the data's quality or organization. Once issues have been identified, we will then use programming to clean and refine the data. Finally, we will analyze the refined data and present our findings through visualization.

Data Gathering:

In this project, data was gathered from multiple sources:

- The WeRateDogs Twitter archive is provided as *twitter-archive-enhanced.csv*.
- The predictions of images in the tweets can be obtained programmatically by downloading the *image-predictions.tsv* file using the Requests library from this link.
- Twitter API and Python's Tweepy library to gather each tweet's retweet count and favorite ("like") count at minimum, and any additional data I find interesting.

Data Assessing:

The problems in the data are grouped in two categories:

Quality:

Accuracy:

- *tweet_id* should be an object instead of integer.
- *Timestamp* and *retweeted_status_timestamp* should be of type datetime instead of object.
- *Created_at* should be of type datetime instead of object.
- Column names not clear.

Validity:

- p1, p2 and p3 columns may contain invalid data (not a dog breed).
- There are 181 retweets by looking at *retweeted_status_id* and 78 replies by looking at *in_reply_to_user_id*.

Consistency:

- *Source* column is not clean.
- Some *rating_denominator* are not on the same scale.
- *id* column in df3 should be renamed to *tweet_id* and converted to object.
- p1, p2 and p3 columns are inconsistent in capitalization.

Completeness:

- Missing values in multiple columns.

Tidiness:

- The last 4 columns (*doggo, floofer, pupper, puppo*) represent one variable that could be named *type.*
- The three tables should be merged in one as they represent different data attributes for the same entity (tweets).

Data Cleaning:

After the assessment, cleaning the data took place through the three steps of Define, Code and Test.

- Change the datatypes of the columns that needs another datatype like *timestamp, tweet_id.*
- Rename columns to clearer names such as *p1* becomes *prediction_1*
- Create a new column *breed* that summarizes *p1, p2* and *p3*.
- Clean breed values and make them consistent.
- Clean the *source* column.
- Fix the wrong values of *rating_denominator* and scale them to be 10. Scale rating_numerator accordingly.
- Drop columns with missing values as they are not needed in the analyses.
- Create a new column *type* to summarize (*doggo, floofer, pupper, puppo*).
- Drop unused data columns.
- Merge the three dataframes.