

# wrangle\_report

May 29, 2022

## 0.0.1 Introduction

This project is all about data wrangling or better put data pre-processing. Data pre-processing or wrangling involves three key stages which were exploited in this project. The stages of data pre-processing are:

- Gathering data
- Assessing Data for Quality and Tidiness Issues
- Cleaning data to enable meaningful usage ranging from data storytelling to making predictions with them.

**Gathering Data** This stage of the project is all about collection of datasets used in the WeRateDogs Twitter data. This stage explored three different types of data gathering. These are:

- Downloading data from a website and loading the downloaded data thereafter to be assessed for further usage. Precisely, archive data was downloaded from <https://learn.udacity.com/nanodegrees/nd002-ent-masterschool/parts/aa944879-6dc1-4280-a760-44f8ea68716d/lessons/ls2232/concepts/06e57daa-cd07-4a1f-936e-c426f7353cc9>
- Using python **Requests** library to get data from the following URL: [https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad\\_image-predictions/image-predictions.tsv](https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv). The twitter image prediction data was gathered using the get method of the requests library.
- Finally, data was gathered using an Application Programming Interface (API). In particular, twitter API was used to collect the retweet count and favorite counts for the WeRateDogs twitter data.

**Assessing Data** The above collected datasets were assessed both visually and programmatically for both data quality and tidiness issues. The visual assessment involves displaying the datasets individually and checking their qualities. The programmatic assessment comprises the use of the following pandas methods:

- head
- describe
- info
- value\_counts
- tail

The following quality issues were observed in the three datasets:

- missing values in some features such as `in_reply_to_status_id`, `in_reply_to_user_id`, `retweeted_status_timestamp`, `retweeted_status_id`
- `rating_denominator` has some values less than 10
- none values in `doggo`, `flooffer`, `pupper`, `puppo`
- there are inconsistencies in `p1`, `p2`, `p3` columns; some of them are capitalized while others are only in lowercases
- `id_str` is of type string rather than integer and should be renamed to `tweet_id`

A number of tidiness issues were detected:

- there are 66 duplicated '`jpg_urls`' and 17 duplicated '`tweet_id`' in image prediction dataset
- in `df_twitter` dataframe, `doggo`, `floofer`, `pupper` and `puppo` are all identifiers of dog breeds which can be merged.

### 0.0.2 Data Cleaning:

The goal of identifying data quality and tidiness issues is to enable seamless cleaning of the datasets. The identified issues were cleaned methodically in four stages, namely:

- stating the issue(s) to be addressed as identified
- defining the steps to be taken in addressing the issues followed by codes for the cleaning.
- Code: implementations of the steps itemized in the previous stage
- Test for checking if the issues have been addressed.

Missing values were generally filled with two pandas fillna methods of `ffill` and `bfill` to recursively fill the missing values with the values after and before them. All the `rating_denominator` values were changed to 10 in conformity with the standard of the **WeRateDogs**. All the duplicates in `df_image_pred` dataframe's '`jpg_url`' and '`tweet_id`' were dropped while keeping the first of the duplicates. The '`None`' values are replaced with '`_`' and the `doggo`, `floofer`, `pupper` and `puppo` were collapsed into one column followed by other processes shown in the other notebook. Finally, the `id_str` was converted to integer type and the column `id_str` renamed to `tweet_id`

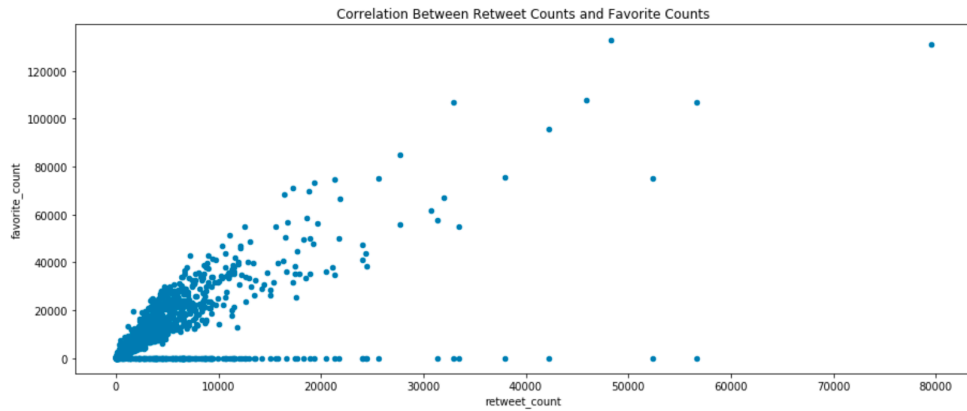
### 0.0.3 Storage

The cleaned datasets were stored for future use.

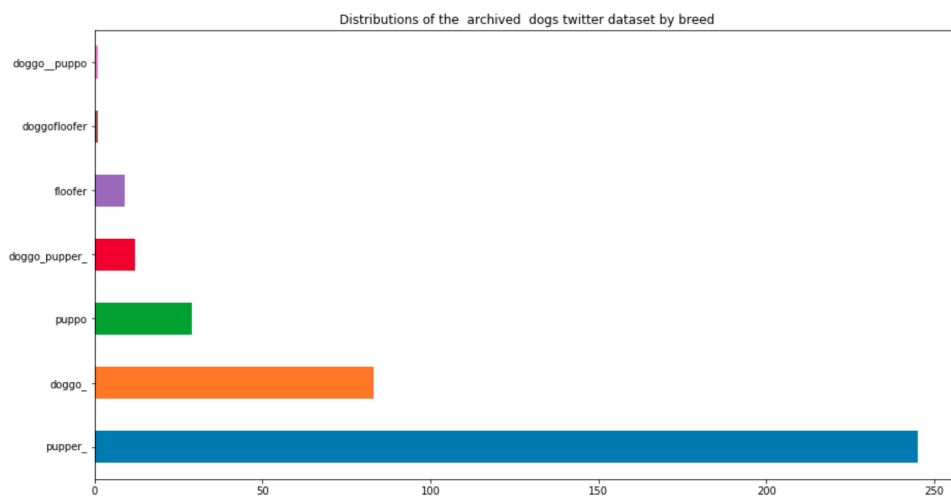
### 0.0.4 Results/Insights

The three gathered, assessed and cleansed datasets are all about the data from **WeRateDogs**. They are the archived dataset, image prediction and retweet counts and favorite datasets. In the archived datasets, the following insights were obtained:

1. There are 380 dogs with known breed names where the most popular breed is `pupper` (245 dogs). 1976 dogs do not have their breeds known or predicted.
2. The source of most dogs data is from **iphone**". 2221 of them are from this source which represents 94 % of all the archived dogs datasets.



'Correlation'



'dog\_breeds'

3. The contents of each tweet as shown in the 'text' column is unique. This implies that tweets about the dogs were completely personalized.
4. The most popular dog p1 value (name) is Golden Retriever (139) followed by Labrador Retriever as given in the image prediction datasets.

The scatter plot below shows the correlation of the retweet\_count and favorite\_count from the twitt dataset. It clearly shows that there is strong positive correlation between the two variables with over 94 % of the retweet\_counts being less than 10000 and most favorite counts are less than 30000.

The chart below shows the distribution of dog breeds in the dataset.