Introduction

The goal of this project was to wrangle the WeRataDogs Twitter data to create interesting and trustworthy analysis and visualizations. Before this could be done, the data has to be gathered from different sources and cleaned.

Gathering Data

Data for this project was gathered from three different sources.

- File in hand: this is the WeRateDogs Twitter archive. This file was manually downloaded and imported into Jupyter Notebook. It was assigned the name: twitter_enhanced.
- Udacity's servers (internet file): this file was downloaded programmatically from the internet using Python's requests library. This file contain dog images from the Twitter archive and a prediction of dog breeds. It was assigned the name: image_predictions.
- Twitter's Api: data were gathered by querying json files from Twitter's API. This required authentication – that was eventually granted. The data gathered were: retweet_count, favorite_count, and friends_count. It was assigned the name: tweet_api.

Assessing Data

The three files were assessed visually and programmatically for issues. The focus of the assessment was not to identify all errors in the data but to concentrate on issues that will influence analysis. At least eight quality and two tidiness issues were to be identified.

Visual Assessment

Some issues were identified by visually assessing the dataframes.

- The number of rows were not the same – twitter_enhanced had 2356 rows, image_predictions had 2075 rows, and tweet_api 2331 rows.
- The dog stages in the twitter_enhanced table did not follow the rules of tidy data. They needed to be merged to a single column.
- The image_predictions table needed to have just one column for image predictions and one for the confidence level instead of three columns for each.
- The ratings columns in the twitter_enhanced table were visually assessed with the text column to spot errors in recording figures. It was noticed that some rows had two ratings.

Programmatic Assessment

The following issues were spotted:

- The twitter_enhanced table contain retweets and replies
- The ratings columns in the twitter_enhanced table had very high and low values – some were apparently errors while others was as a result of having multiple dogs in a row.
- The friends_count column in the tweet_api table had a single number (17) throughout. It was not very useful.
- Some dog names in the twitter_enhanced table were incorrect. The incorrect names began with lowercase letters.

- There were 66 duplicated images in the image_predictions table.
- Some datatypes in the twitter_enhanced and image_predictions were incorrect.

Cleaning Data

In this section, the identified issues were cleaned. The cleaning stage was subdivided into three according to the nature of cleaning task to perform. For each cleaning task, the issue was defined, code written to resolve the issue(s) and testing performed on the data to verify if the issue was resolved.

- **Quality issues – Completeness:**
  - Rows containing retweets were removed.
  - Rows containing replies were removed.
- **Tidiness issues:**
  - The four columns containing dog stage were melted into one column of dog stage and the original dog stage columns dropped. The resulting duplicated rows were also dropped.
  - The breed prediction and confidence level columns were melted into single columns each and original columns dropped.
  - The three tables were merged into a single table.
  - Irrelevant columns were removed from the resulting dataframe.
- **Quality issues:**
  - Wrong datatypes were adjusted accordingly.
  - Erroneously recorded rating values were corrected using data from the text column and incorrect ratings removed from the data.
  - Incorrect dog names were removed.

Conclusion

After the data cleaning process, the resulting dataframe was stored in a csv file with the name: twitter_archive_master.csv. This master data was then used for analysis.