Introduction

The goal of the project is conduct data analysis and visualisation by wrangling data from Twitter user ’WeRateDogs’. With Twitter archive in hand, the objective is to gather comprehensive data to supplement the archive, which contains basic tweet information, to begin and complete data analysis on the data. The data will be assessed on its tidiness and quality and cleaned effectively before developing a visual.

Gather

The task began with obtaining further data by querying the Twitter API, which was then imported to Jupyter notebook for assessment. The data in hand contains basic information and to get a complete picture of data from WeRateDogs, one could be sourced out Twitter API based on the Tweet IDs which are already present in the archive. The data is in the csv format.

The idea is to gain access to the API by registering a developer account with Twitter and subsequently obtained their access code. This will create an API object that you can use to gather Twitter data. The data obtained here was saved in JSON format, of which extraction was done via Tweepy.

For this project, there is a neural network from Udacity showcasing tweet image predictions and data therein is on its server, which was downloaded programmatically using Requests.

All data obtained here was amalgamated for an easy review.

Assess

The assessment of the data revealed several tidiness and data quality issues. Not all issues were rectified, however. These include:

Quality Issues:

- To remove rows that contained retweets, by selecting rows that had the string "RT" and utilising the ‘drop’ method.
- Once these rows with retweets were deleted, they had no value. Hence, were deleted.
- Rows in columns ['*in_reply_to_status_id*', '*in_reply_to_user_id*'] contained missing values. They were deleted by the ‘drop’ method.
- Different datatypes in ‘timestamp’, ‘favorite_count’ and ‘retweet_count’. To change timestamp to time and date using to_datetime. To change the other two from float to int using astype
- Column 'name' contains peculiar names like 'a' and 'none'.
- Column 'expanded url' contains inconsistent url data. Some rows have more than one url.
- Rating numerator have values above 100 and more, above the usual range.
- Rows in ‘favorite_count’ and ‘retweet_count’ had NaN values. To fillna the rows.

Tidiness:

- Column 'name' contained names that are not consistently capitalised. Deployed the series.title.capitalize function.
- Columns 'img_num', 'p1' to 'p3', 'p1_conf' to 'p3_conf' and 'p1_dog' to 'p3_dog' can be grouped/collapsed as/into one big column.
- Dog stages can be grouped as one column and converting the None to NaN values.

Cleaning

The identified issues were cleaned programmatically. As seen in the assessment, the issues and parameters were initially defined. Then individual issue was cleaned using codes which were eventually tested to make sure the operations worked. The issue with missing data and data that was not complete was tackled first. Then quality and tidiness issues, in that order. For this exercise, only 8 quality issues and 2 tidiness issues were rectified.

Summary / Conclusion

The project began with data gathering which required opening a developer account with Twitter in order to access the API. Once that was done, the next stage began easier. The assessment of the combined data was tedious. There are many issues in the data as described above. However, for this project only 8 quality issues and 2 tidiness issues were resolved. The final product has 25 columns down from 32 columns from all three tables. The data now is ready for visualisation. A total of seven plots were made based on the findings.