Wrangle Report

This project deals with data about rating of images of dogs posted by people on twitter by WeRateDogs.

The wrangling process included 3 steps.

1. Gathering Data

2. Assessing Data

3. Cleaning Data

After wrangling data was analysed and visualised to draw conclusions.

Data Wrangling Process:

1. The data was collected from 3 sources.

    - Data was read by a file, twitter_archive_enhanced.csv, that was provided. It contained data archieve of tweets.

    - Data was read by a file that was programmatically downloaded by a URL. It contained data about breed of the dog.

    - Data was then collected by the Twitter API, which had all the information about the tweets.

2. The collected data then underwent assessing. In this part the data was checked for major two issues - Quality Issues and Tidiness Issues.

    -**Quality Issues**: The content related issues of the data. It included four major dimensions - Completeness,Validity,Accuracy,Consistency.

    The major Quality issues found are:

    1.in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id,retweeted_status_user_id read as float instead of int

    2. retweeted_status_timestamp read as string instead of datatime object

    3. Invalid values of numerator and denomenator

    4. Unusual dog names like 'a','an','None'

    5. Null values represented as 'None' which will be read as string

    6. There are retweets in twitter_archieve

    7. There are tweets that don't contain any images

    8. id and id_str both read as integer and are different in some observations.

9. It takes first fraction in the text to be the rating_numerator and rating_denominator without looking further

10. Id number 810984652412424192 has no rating and rating_nnumerator and rating_denominator are mistaken to be the first fraction encountered.

11. More descriptive column names can be given

12. Columns with no values at all can be dropped as they only missing data

13. Duplicates

-**Tidiness Issues**: The structural related issues of the data. The guildlines to find to find this are

Each variable forms a column.

Each observation forms a row.

Each type of observational unit forms a table.

The major tidiness issues found are:

1. Different stages of dog in 4 variables (doggo,floofer,pupper,puppo) instead of one

2. Irrelevant and similar data spread across 3 tables. We can merge them in single table containing table of our interest

Hence, we found various issues with the dataset.

3. The assessed data with various issues now has to be cleaned. The error has to be defined, coded to remove the issue and then tested to see that the issue is removed.

1. First of all tidiness issues were removed.

- The 4 variables were made to single column.
- The 3 tables were merged to a single one.
- Irrelevant columns were removed.

2. We see that due to the above steps, some quality issues were solved automatically.

- id and id_str columns were removed.
- Somes ids were removed. Hence no need to change type.
- Columns with no entries were removed

3. Next, Quality issues were cleaned.

- Retweets were deleted. Along with them duplicate entries and entries with no image were also removed.

- Multiple predictions were merged to single column. (A tidiness issue but seen during quality issues. No problem as this process is iterative.)

- Null values with 'None' replaced with NaNs and 0s.

- Standardising numerator and denominator by keeping decimal values rather than a fraction. Invalid values of numerator and denominator hence removed.

- Converting each data type to its correct form.

- Renaming with more descriptive names.

Hence the data was cleaned and stored in twitter_archive_master.csv. The data is now ready for analysis and visualisation.