Wrangle and Analyze Data

(Adel Abu Hashim)

Wrangling data contains three steps

- Gathering Data
- Assessing Data
- Clean Data

Actually we don't stop after cleaning but we move up again for the previous step and repeat until we get best results.

Gathering Data

Data was gathered from 3 different sources:

- First data set (The WeRateDogs Twitter archive) - Downloaded manually, We convert it automatically into data frame using pandas library; through read_csv method
- Second data set (The tweet image predictions) -Downloaded programmatically using python requests library, the we store it in our device using os library the it is easy to convert it into data frame using pandas again.
- Third data set (Tweets JSON data) - I downloaded data manually but in JSON format which must be converted into a format in which we could get its data frame.

Assessing Data

We doing assessing on data in order to find quality and tidiness issues;

- Quality:

1. Wrong data types is some columns like (timestamp,retweeted_status_timestamp)
2. Null values in columns of (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls)
3. Some records have dominarator more than 10(outliers), Some records have denominators not equal to 10.
4. There are no columns for ratings.
5. 745 records have 'None' names.
6. Doggo, Puppo, Pupper and Floofer columns have so many 'None' values.
7. Wrong names like 'a' and ‘the’ and ‘by’ for example.
8. . All retweets and replies should be deleted.
9. P1, P2, P3 has no real dog names sometimes
10. The dataset doesn't have records for all 2354 ids. it only have 2075 ids

- Tidiness:

. All three datasets should be joined together.

. Doggo, Puppo, Pupper and Floofer columns as dog_stages(I knew by search LOL!).

Cleaning Data

This process came after conclusions on assessing data trying to fix quality and tidiness issues like;

- Correct wrong data types.
- Delete Extraneous columns on data.
- Fix Null problem.

This problem contains three steps:

- Define
- Code
- Test

Storing Data

After cleaning data we save it in order to use it any time.

Analysis Data and Visualization

These steps are not part of the data wrangling process. However, it cannot reflect correct and accurate insights without performing data wrangling first. Visualizations and insights are provided in ‘act_report.pdf’.