REPORT

DATA WRANGLING OF THE WeRateDogs TWEET DATASET

The project was carried out in different stages as given below:

1. Data Gathering.
2. Data Accessment.
3. Data Cleaning.
4. Data Storage.
5. Data Analysis and Visualization.

The data used in this project was gotten from three diferent sources. The first (twitter-arc) was downloaded directly as a csv file. The second was downloaded using the request library(tweet_pred) and the third involved using the tweepy API to query twitter for the data(twitter_api). The datasets were downloaded, read into different dataframes which allowed for assessment.
The data was assessed using pandas and its methods, excel was also used.

Looking Into Data With Excel and Pandas

Using excel, inconsistent rating scores were being used for both the denominator and numerator, with the prescribe rating being over 10. Peculiar dog names were also observed e.g; 'None', 'a', 'an' etc.

Using pandas, we observed incomplete data collected and inappropriate data types given to certain columns amongst others.

Issues Observed

Given below are the observed issues we observed while assesing the data from the three datasets;

Quality (The Validity,Completeness, Consistency, Accuracy Issues)

1. "twitter_arc" has 2356 rows while the "twitter_pred" has only 2075 rows, which can be attributed to retweets and missing photos.
2. Not needed retweet and in_reply columns(user_id, retweeted_status_id, in_reply_to_user_id).
3. Unusual dog names, such as; 'a', 'an','all', etc., we also had missing names.

4. Wrong numerator and denominator numbers/values.
5. The source data is not clear as we have html tags surrounding the needed info.
6. Timestamp is stored in a wrong data types and has additional info that is not needed.
7. tweet_id is stored in a wrong data type.
8. Incomplete data in other rows, with a bulk of some having 'None' as data.

Tidiness (The structural issues)

1. All three dataframes, should be together to form a complete dataframe.
2. The doggo, floffer, pupper, and puppo columns in the "twitter_arc" should be together in one column.

The data was cleaned using python in a jupyter notebook. Major cleaning operations included merging of the three(3) datasets to get a comprehensive single data sheet(arc_api). Notable cleaning operations include;
- Removing of redundant columns.
- Getting equal amount of data for the remaining columns. This involved filling empty rows.
- Merging the four different columns that signified growth stage of the dogs into a single column.
- Source information was also cleaned to get the needed info.
- Columns with wrong data types were also worked on.

Upon the completion of the cleaning operation, the data was stored with the prescribed name and format(twitter_archive_master.csv).
Further operations like visualization and analysis were also carried out using the cleaned data.

In [ ]: