Report

The Data Wrangling Process

The data wrangling process of the project is performed in three (3) different stages. These include gathering the data, assessing it, and cleaning it. These activities are important to ensure that insight generated from the data are not impacted negatively by inaccurate, misleading, or missing an important information.

- The Data Gathering Stage
  Data for the project was gathered from twitter, specifically from the WeRateDogs handle (archive) and kept in a csv file. These gave access to three different data which together formed the data set for the project and were downloaded from the website. The data provides tweets about dogs ratings. The twitter-archive-enhanced.csv contains tweet details while the tweet_json.txt file is used to keep important details like favourite and retweet counts of each tweet id. Meanwhile, the image-prediction.tsv file is downloaded from here.
  These three files were read into jupyter notebook using pandas as dogs_rating_df, additional_rating_df and image_predict_df.

- Assessing the Data
  The data read was assessed for quality and tidiness. It was very useful to understand the quality of the data and how its structure fit for our analysis.
  The following issues were unveiled.

  Quality issues
  1. Missing data in columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, expanded_urls, names).
  2. The rating_denominator column contains values other than 10.
  3. Replies and retweets exist in the data. Only original tweets with ratings are required.
  4. Wrong data types (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, timestamp).
  5. The name column contains inaccurate values, eg. a, an.
  6. Null values represented as none in columns (doggo, floofer, pupper, puppo).
  7. There are two rows less in this column.
  8. Two column names and values combined into one column in the image_predict_df table.
  9. The values in the text column comprise of the text and the link to the tweets.

  Tidiness issues
  1. The rating column should be a constant value. The column need not to be created.
  2. `dogs_rating_df table` doggo, pupper, floofer, puppo should be values and not columns.
  3. `image_predict_df table` two column names and values combined into one column
  4. The three dataset can be merged with the necessary columns.

- Cleaning the Data
  The issues identified were carefully analysed and cleaned. The cleaning process involves dropping columns that are not required for the analysis. These columns are thus removed: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, expanded_urls, retweeted_status_timestamp, and the rating_denominator. Retweets and replies were removed. The timestamp for the tweets were also converted to datetime.
  The text column includes the url to the tweets, numbers, punctuations. These were removed hence the text will be taken through sentiment analysis. All none values were also converted to null and the dog stages were melted to make them values under a new column called dog_stages.
  A few rating values in the rating_numerator column were extreme given that we want to analyse those scores. These values are thus replaced with the mean rating scores.
  The three tables were then combined (combined_dogs_tweet_df).