Wrangle report

by Michael Bong

Overview

In this project we wrangled 'WeRateDogs' Twitter data by putting it a combination of 3 data sets through a process of gathering, assessing and cleaning. The 3 data sets are:

- Data set 1 (df_archive) - 'twitter_archive_enhanced.csv': Historical tweet data
- Data set 2 (df_image) - 'image_predictions.tsv': Predicted dog type data
- Data set 3 (df_tweet) - 'tweet_json.txt': Live tweet data

Once the data sets have been gathered, each of them is assessed individually to highlight any Quality or Tidiness issues that needed to be addressed before we can combine the datasets for use in further analysis. Both types of issues are addressed in turn.

Quality issues

Quality issues cover instances where a data set contains data that should not be there. Alternatively, it also covers instances where the data set is missing data that should be present.

We will briefly explain the quality issues that have been highlighted for each data set below. Explanation will focus on reason the cleaning needs to be done and how the cleaning was done.

Data set 1: df_archive

- (A) Retweets need to be removed
    - As we are only interested in original, initial tweets, we need to remove all tweets that are retweets. Here, retweets are not supposed to be in the data set. This is done by identifying retweets using the retweeted status ID. Once identified, these retweets were removed from the data set.
- (C) Remove unrequired columns such as: 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id' 'retweeted_status_user_id', 'retweeted_status_timestamp'
    - The initial data set has many columns that will not be used in further analysis, therefore for neatness, these columns need to be removed. The unrequired columns are dropped from the data set.
- (D) 'source' has a lot of unnecessary characters and can be cleaned to show only relevant information
    - The 'source' column contains a lot of unnecessary content such as characters for address references. The surplus content is removed and only key pieces of information

    - such as ('Twitter for iPhone', 'Vine - Make a Scene', 'Twitter Web Client', 'TweetDeck') remain.
- (E) All identifier columns such as 'tweet_id' and 'id' should be made to a string object, as we do not need to do any calculations on them
    - Identifier columns initially are a mix of string and integer objects. To ensure consistency and as these columns will not be needed for calculations, their data types have been changed to string object.
- (L) Object is data type of 'timestamp', when it should be datetime
    - The 'timestamp' column has an object data type. As these are date and times, the data type of this column has been changed to datetime.
- (M) Remove tweets without images (expanded_urls)
    - Tweets without images need to be removed. These anomalous tweets can be identified as tweets without image urls. These tweets are removed from the data set as without an image, there cannot be an associated rating.
- (Z2) Remove 'doggo', 'pupper', 'puppo', 'floofer' columns after a consolidated 'dog_type' column has been created
    - Instead of having a separate column for each dog type, it is cleaner to have a 'dog_type' column. An if-else function is created to convert the contents of the 4 initial columns into a single 'dog_type' column.
    - Once this consolidated column is created, the initial 4 separate dog type columns can be dropped.

DATA SET 2: DF_IMAGE

- (H) All identifier columns such as 'tweet_id' and 'id' should be made to a string object, as we do not need to do any calculations on them
    - Identifier columns initially are a mix of string and integer objects. To ensure consistency and as these columns will not be needed for calculations, their data types have been changed to string object.
- (N) The predictions in 'p1', 'p2', 'p3' have underscores, which should be replaced by spaces
    - This is a simple step to replace the underscores in the content of the columns 'p1', 'p2', 'p3' with blank spaces.
- (O) Only keep dog predictions that are dogs, with the highest level of confidence
    - In this project, we are only interested in predictions that are dogs. Non-dog predictions do not add any value and therefore needs to be removed.
    - Additionally, an if-else function is created to ensure we only keep predictions that are dogs and of the highest level of confidence. The output of this function for each tweet is stored in a new column called 'predicted_dog_breed'.
- (P) Remove columns that are no longer needed: 'img_num', 'p1', 'p1_conf', 'p1_dog', 'p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog'
    - After the new 'predicted_dog_breed' column has been created, the original columns are no longer required for analysis and can be dropped from the data set.

DATA SET 3: DF_TWEET

- (I) Retweets need to be removed
    - As we are only interested in original, initial tweets, we need to remove all tweets that are retweets. Here, retweets are not supposed to be in the data set. This is done by identifying retweets using the retweeted status ID. Once identified, these retweets were removed from the data set.
- (J) All identifier columns such as 'tweet_id' and 'id' should be made to a string object, as we do not need to do any calculations on them
    - Identifier columns initially are a mix of string and integer objects. To ensure consistency and as these columns will not be needed for calculations, their data types have been changed to string object.
- (K) Only keep the columns that are needed ['id', 'retweet_count', 'favorite_count']
    - We only need to keep the columns needed for analysis. All unused columns can be dropped.

DATA SET 4: DF_MASTER

- (W) Add in actual_rating, which is 'rating_numerator' / 'rating_denominator'. Remove outliers.
    - 'Df_master' is our combined data set that will be used for analysis. In order to perform meaningful analysis, a new calculated column needs to be created called actual_rating'. This will enable tweets to be compared from a ratings perspective.
    - After the 'actual_rating' column is created, outliers need to be removed, otherwise they may skew the analysis.

TIDINESS ISSUES

Tidiness issues cover mostly: data should be there, but not in the correct forms, structure

DATA SET 1: DF_ARCHIVE

- (Z1) The classifications of dog stages from the Dogtionary ('doggo', 'pupper', 'puppo', 'floofer') should be consolidated into a column
    - Instead of having a separate column for each dog type, it is cleaner to have a 'dog_type' column. An if-else function is created to convert the contents of the 4 initial columns into a single 'dog_type' column.

- (X) Join df_archive with df_image and df_tweet
    - A 'Df_master'data set is created and it is the combination of all the 3 data sets that have been cleaned. This combined data set will be used for analysis.

DATA SET 3: DF_TWEET

- (Y) 'id' column should be renamed to 'tweet_id' to make is consistent with other data frames
  - This simply involves renaming the ‘id’ column into ‘tweet_id’ to prevent confusion and make it consistent with the other data sets.

CONCLUSION

Having completed the cleaning of the data sets, we now have a combined and cleaned ‘Df_master’ data set that can be used to perform further analysis and gain new insights.