Wrangle Report

Purpose of the report

The target of the report is to show the steps done in wrangling the data related to the “WeRateDogs” tweets. The wrangling process consists of three stages. Data Gathering, Assessing and Cleaning

Data Gathering

Three sources of data will be used in the project as identified below

Enhanced Twitter Archive

The WeRateDogs Twitter archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 2356 of their tweets. This data is given as part of the project resources

Image Predictions File

The tweet image predictions file contains predictions for what breed of dog is present in each tweet image according to neural network algorithm. The data of this file exists online in a url hosted in udacity and will be downloaded programmatically using request api

Twitter API

Additional data can be retrieved from Twitter other than those exist in the enhanced twitter archive file, like retweet and favorite counts. I will retrieve this data from the Twitter's API.

Output Data Gathering

The 3 data frames are:-

- twitter_archive_df - contains data read from provided csv file
- image_predictions_df - contains data read (by using requests) from tsv file hosted on server
- api_df - contains data obtained from twitter handle by using tweepy library

Data Assessment

The target of assessing the data is to identify the issue that exists on the data which will prevent from doing the analysis in later stage or will lead to wrong analysis. In this stage I will identify a set of quality issues and tidiness issues I found in the gathered datasets

Issue No	Description	Place	Issue Type
Q1	As per project constraint We only need original rating no retweet that have images should be included – The variable "expanded_urls" has few missing values ,which means some tweets have no images, any rating without images shouldn't be taken	Twitter Archive	Quality
Q1	As per project constraint We only need original rating no retweet or reply that have images should be included – The are many retweet and reply tweets that exist in the 3 datasets	All Data Frames	Quality
Q2	Unneeded fields in the analysis in_reply_to_status,id,in_reply_to_user_id retweeted_status_id,retweeted_status_user_id'	Twitter Archive	Quality
Q3	Errorenous DataTypes in timestamp field	Twitter Archive	Quality
Q4	Missing data in column "name" are showed as non-null values "None"	Twitter Archive	Quality
Q5	Incorrect values in column “name” ,like a,an	Twitter Archive	Quality
Q6	Dog develepment stages (puppo,floofer,doggo,pupper) have inconsistent presentation of null values as 'None' when Nan should be used	Twitter Archive	Quality
T1	Column headers (Doggo,pupper,floofer,puppo) are values for a variable dog stage	Twitter Archive	Tidiness
Q7	The name of variables p1,p2,p3,p1_conf,p2_conf,p2_dog is not indicative	Image Prediction	Quality
Q8	Source column need to be simplified	Twitter Archive	Quality

T2	Combine 3 data frame into one master data frame		Tidiness

Data Cleaning

All identified issues in the assessment are cleaned as follows

1. **As per project constraint We only need original rating no retweet that have images should be included**

    - Delete retweets by filtering the NaN of retweeted_status_user_id in twitter archive data frame
    - Delete retweets by filtering the NaN of in_reply_to_status_id
    - Delete Tweet with no images by filtering the not is NAN of expanded urls
    - Filtering tweets in archive data frame based on tweet with images in image prediction dataframe
    - Delete from image prediction data frame all tweets that are retweet or reply
    - Delete from twitter api all tweets that are retweet or reply
2. **Removing fields from tweet archive data frame that will not be needed in the analysis**
3. **Change the data type of the timestamp field in the archive data frame**
4. **In case the “name value” is “None” Change the “name” value to the name in the text field or to null instead of “None”. The dog name is found in the "text" variable using regular expression. Searching for the pattern "named is ,"named"**
5. **In case the “name value” is lower than 3 characters and starts with lower character Change the name value to null or to the name of the dog extracted from the text field. The dog name is found in the "text" variable using reqular expression.Searching for the pattern "named is ,"named"**

6. **Create a new variable “dog_stage” to capture the values in the columns headers (Doggo, pupper , puppo,floofer) and drop the 4 columns after that (Q6,T1 issues)**
    a. Replace "None" values in columns (puppo,floofer,doggo,pupper) by empty string
    b. Create new column dog_stage that concatenate the values in columns (puppo,floofer,doggo,pupper)-**T1 issue**
    c. drop old columns (puppo,floofer,doggo,pupper)
    d. Remove inconsisteny in dog_breed presentation of null values as empty string- **Q6 Issue**
7. Change Column names (p1,p2,p3) in Image Predictions Data frame to more clear names
8. Change the source of tweet to more readable names
9. Merging the 3 data frames to one master data frame based on twitter id –**T2 issue**

Output of Data Cleaning

- Twitter_df: contains the merging of the 3 data sets

Data Storing

Store the master data frame “twitter_df” into one file “twitter_archive_master.csv”

Data Analysis & Visualizing

Use the master data file created in the Data storing, “twitter_archive_master.csv ” in the analysis