Wrangle & Analyze Data

Introduction

In this Project we are working on wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

Gather Data

Below are the files Gathered as part of this Project:

- **twitter_archive_enhanced.csv** – This file needs to be downloaded manually and upload it to Jupyter notebook. There are total 2356 records present in this file.
- **image_predictions.tsv** – This file is tweet image predictions file which basically tells which breed of dog is present in each tweet according to neural network. This file is hosted on Udacity server where we need to download using programmatically and place it into the Jupyter notebook.
- **tweet_json.txt** – This file needs to be created using tweet ID's in the WeRateDogs Twitter Archive, query the Twitter API for each tweets JSON data using Pythons Tweepy library and store each tweets entire set of JSON data.

Assessing Data

After gathering the above three files, we will be using Pandas to read those three files and create three dataframes for assessment. Now each dataframe needs to be assess Visually and Programmatically to check the Quality and Tidiness issues.

Quality Issues:

Visual Assessment

- In twit_arc_enh, Drop all 181 rows containing retweets, where these columns will be non-null: retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp.
- In twit_arc_enh, Drop all 78 rows containing replies, where those that have non-null values in these columns: in_reply_to_status_id and in_reply_to_user_id.
- In twit_arc_enh, timestamp and retweeted_status_timestamp are of object type. It should be timestamp type.
- In img_pred, Greatest of all probablity with TRUE value will be taken into one column

- In img_pred, Prediction names are having _ hence remove it with space

Programmatic Assessment

- In twit_arc_enh, numerator values as 1776, 960, 666, 420, 143
- In twit_arc_enh, denominator values as 11, 170, 50 are moved to 10
- In twit_arc_enh, extract the string in between the html tags for Source column
- In retwt_DataFrame, remove 163 retweets

Tidiness Issues

- Dog Type columns should be re-structured into one columns to show the value in one.
- Merge all the Dataframes into one to visualise

Cleaning Data

The quality and tidiness issues which are addressed in the above step are cleaned here.

- Using Programmatic techniques amended all the data wherever it is necessary.
- Created new Data Frame to visualise the data in better way.

- Unwanted columns in twit_arc_enh needs to be dropped.
- Unwanted columns in img_pred need to be dropped after we finalise Prediction and Confidence for each dog
- Unwanted columns in retwt_DataFrame needs to be dropped