Wrangle Report

Introduction

The purpose of this project to analyze the database given by the **WeRateDogs** account which is a Twitter account which rates the dogs , These ratings almost always have a denominator of 10. The numerators, always greater than 10. 11/10, 12/10, 13/10, etc. Because "they're good dogs Brent."
WeRateDogs has over 4 million followers and has received international media coverage.

The Main goal of the project:

Wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. Additional gathering, then assessing and cleaning is required for "*Wow!*"-worthy analyses and visualizations.

Data wrangling, consists of:

- Gathering data :
  This stage we will depends on a given file which is “twitter_archive_enhanced.csv” and download file “image_predictions.tsv” programmatically using the Requests library and the URL

  There is another file which is “tweet_json.txt” which can be downloaded using using Python's Tweepy library

  Gathering data stage goal is to gather the required data and read it in the notebook file in order to assess these data.

- Assessing data
  This stage target to detect the quality and the tidiness of data in order to achieve convenient analysis, by assessing the data, I find many issues regarding the quality and tidiness of data which is:

twitter_archive

- There are many unwanted columns with high counts of missing values("in_reply_to_status_id","in_reply_to_user_id,retweeted_status_id", "retweeted_status_user_id", "retweeted_status_timestamp")
- datatype issue in coloumn'timestamp' need to convert to timedate and convert it to (year,month,day) columns
- invalid data in rating_numerator column for instance(1776,960)
- invalid data in rating_denominator column which are greater than or less than 10
- Convert 'None' values to 'NaN' to union all none values
- Make a new column for net rate of dogs
- change datatype for 'tweet_id' & 'dogs_stage'

image_predictions

- There are 66 'jpg_url' duplicated
- Remove unwanted columns to make a convenient analysis

Tidiness

twitter_archive

- 1 variable (dog stage) is present in 4 different columns (doggo, floofer, pupper, and puppo)
- twitter_archive & Twitter_json dataframes can be merged for a convenient analysis.
- As we can always iterate on the database , dataframes (twitter_archive_clean,image_predictions_clean) can be merged in one dataframe to provide more tidy database for analysis

- Cleaning data:
  In this stage you clean the data using python equations and pandas library by defining the issues and how to solve it then test the codes.
  The following screens shots shows an example of cleaning process

Quality

1-There are many unwanted columns with high counts of missing values("in_reply_to_status_id","in_reply_to_user_id,retweeted_status_id", "retweeted_status_user_id", "retweeted_status_timestamp")

Define

drop all the unwanted columns

Code

```
twitter_archive_clean.drop(['in_reply_to_status_id','in_reply_to_user_id','retweeted_status_id','retweeted_status_user_id',
                            'retweeted_status_timestamp','expanded_urls','source'], axis=1, inplace=True)
```

Test

```
twitter_archive_clean.head(2)
```

	tweet_id	timestamp	text	rating_numerator	rating_denominator	name	doggo	floofer	pupper	puppo
0	892420643555336193	2017-08-01 16:23:56 +0000	This is Phineas. He's a mystical boy. Only eve...	13	10	Phineas	None	None	None	None
1	892177421306343426	2017-08-01 00:17:27 +0000	This is Tilly. She's just checking pup on you....	13	10	Tilly	None	None	None	None

The most common notice about assessing data is the iteration process (you assess then you discover another issues while you working with cleaning the back to assessment then clean and so on

Cleaning process considered to be the heavy task on the project in order to achieve copy of a clean data which make the following task (Analysis & Visualization more easier)