Introduction

- The dataset I have wrangled is the tweet archive of Twitter account called @dog_rates (https://twitter.com/dog_rates), also known as WeRateDogs (https://twitter.com/dog_rates). WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.
- In this report I will briefly describe my Wrangling efforts done on the data associated with WeRateDogs Twitter account

Data Wrangling Process

- The process is divided in three steps:
  - Gathering Data
  - Assessing Data
  - Cleaning Data

1- Gathering Data

In this project I gathered three datasets of different sources as described below in a ipynb Jupyter Notebook file titled as `wrangle_act.ipynb`:

1) The WeRateDogs Twitter archive. This was a file on hand. I download this file manually by clicking on the link provided on the project details page. The file was named as `twitter_archive_enhanced.csv`.

2) The tweet image predictions. That is associated with predictions regarding what breed of dog according to a model of neural network. This file `image_predictions.tsv` was hosted on Udacity's servers and I downloaded it programmatically using the Requests library and this link (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv)

3) The third dataset was supposed to be gathered using Tweepy to query Twitter's API for additional data beyond the data included in the WeRateDogs Twitter archive. This additional dataset was required to include retweet count and favorite count as minmum . Gathering data using the Twitter API was requiring me to get a Twitter Developer Account. However, I applied for a Twitter Developer Account , but unfortunately my applications was **NOT APPROVED**

So, I collected the Twitter data without actually creating a Twitter Developer account By Downloading the two files that were included on Twitter API Page in Udacity's Project Section:

1. twitter_api.py: I copied the code that was included in this python document and pasted it on a Code Cell on my Jupyter Notebook `wrangle_act.ipynb`. I also commented the whole cell not to cause any break in the kernell when running the code.
2. tweet-json.zip: This zip folder contains a txt file that supposed to be the resulting data from successfully running the code included in the twitter_api.py. I extracted this zip file using the zipfile library

Then read it line by line into a pandas DataFrame to get include retweet count and favorite count data.

Finally, I was successfully gathered the three datsets with total number of **6785** records combined.

2- Assessing Data.

2- Assessing Data I assessed the gathered data both visually and programmatically for quality and tidiness issues.

- The Programmatic Assessment required using a range of Pandas functions and methods such as : .info(), .describe(), .head(), .sample(), .value_counts() and .unique().
- Here are the issues I've detected:

Qulity issues:

- `twitter_archive`
  - **Missing Values** :
    - in_reply_to_status_id, in_reply_to_user_id : 78 instead of 2356
    - retweeted_status_id,retweeted_status_user_id,retweeted_status_timestamp 181 instead of 2356
    - expanded_urls : 2297 instead of 2356
  - We are interested in the tweet ONLY not the retweet
  - We are interested in the tweet ONLY not the reply to the original tweet
  - **tweet_id** is saved as int datatype instead of/ "better to be" string (object)
  - **timestamp , retweeted_status_timestamp** are saved as object datatype (str) instead of date/timestamp
  - **source** column is writen in html containg `<a>` tags
  - column **name** :
    - some values are not titled `untitled_unlowers` ('BeBe','DonDon','CeCe',, 'JD', 'DayZ')
    - some are inacuarte values : `lowers ['such', 'a', 'quite', 'not', 'one', 'incredibly', 'mad','an', 'very', 'just', 'my', 'his', 'actually', 'getting','this', 'unacceptable', 'all', 'old', 'infuriating', 'the','by', 'officially', 'life', 'light', 'space']`
  - **rating_numerator & rating_denominator**:
    - datatype for rating_numerator should be float instead of int
    - fix:

        - @45 13.5/10 instead of 5/10
        - @ 313 13/10 instead of 960/0
        - @ 2335 : 9/10 instead of 1/2
        - @ 1068 : 14/10 instead of 9/11
        - @1165: 13/10 instead of 4/20
        - @ 1202 : 11/10 instead of 50/50
        - @ 1662 10/10 instead of 7/11
        - @ 695 : 9.75/10 instead of 75/10
        - @763 : 11.27/10 instead of 27/10
        - @1712 :11.26/10 instead of 26/10
      - drop:
        - @ 516 no rating
        - @342 inaccurate (account start date)
      - invistigate(outliers):
        - @ 315 https://t.co/YbEJPkg4Ag (https://t.co/YbEJPkg4Ag) 0/10
        - @979 1776/10
        - @ 1634 : https://t.co/kRK51Y5ac3 (https://t.co/kRK51Y5ac3) 143/130
        - @2074 :420/10
        - @1274 names
    - columns **doggo,floofer,pupper, puppo** has None values instead of Null.
    - We are interested in dogs , **text** column reveals the truth about that some tweets are not related to dogs
    - expanded_urls is too bulky we are interested in tweet link only.
- `image_pred`
    - some images are not for dogs
    - tweet_id is saved as *int* datatype instead of *object* datatype
    - replace the underscore in breeds values with space and title all breeds values (p1 &p2& p3)
- `twitter_json`
    - column **id** is saved as *int* datatype instead of *object* datatype & rename as tweet_id
- `All_datasets`
    - we have completeness issue not all the datasets have the same number of observation

Tidiness issues:

- `twitter_archive`
    - text column has two variables text and short urls,create short_urls column, drop expanded_urls
    - The values of four columns (doggo,floofer,pupper,puppo) in `twitter_archive` dataset should be in one column dog_stage with a category datatype.
    - rating_numerator and rating_denominator columns in `twitter_archive` dataset should form one column dog_rating normalized out of 10.
    - make new columns for `day_name` and `month` from the timstamp column
- `image_pred`
    - Columns p1, p1_dog, p1_conf , p2, p2_dog, p2_conf , p3, p3_dog, p3_conf could be condenced to two columns `dog_breed` and `confidence`
- `All datasets`
    - tweet_id is present in two datasets and after renaming it will appear in all datasets
    - `tweet_json` and `image_pred` datasets should be part of our main dataset `twitter_archive`.

3- Cleaning Data

My main approach in the cleaning process is **First things first** So, I started with the:

- Missing Data ,
- Tidiness
- Other Quality Issues.

- It is important to state here that I focus more on Tideness issues and do what ever quality cleaning leads to that point.

Resources

The following are some of the Resources I used to Clean and analyze the Data:

- Stackoverflow : delete list of rows in pandas (https://stackoverflow.com/questions/14661701/how-to-drop-a-list-of-rows-from-pandas-dataframe)
- DOCS: Get the name of the day in pandas (http://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.dt.day_name.html)
- Conditional replace (https://stackoverflow.com/questions/21608228/conditional-replace-pandas)
- Outliers :Independence Day (https://en.wikipedia.org/wiki/Independence_Day_(United_States))
- Outliers: Snoop dogg (https://en.wikipedia.org/wiki/Snoop_Dogg)
- Holidays (https://www.edarabia.com/school-holidays-united-states/)

> **Important Note** : I assessed and cleaned (if necessary) the data upon which my analyses and visualizations are based.

My Final product from this Data Wrangling Process was a high quality and tidy master pandas titled: