WeRateDogs

Introduction:

This project is to collect, clean and analysis twitter's data from WeRateDogs account. There are steps to be fallowed during this process (Gathering, Assessing, Cleaning and Analyzing).

Gathering Data:

There were three sources of data (Twitter-archive-enhanced.csv, Image-predictions.tsv and Twitter API/tweet-json.txt). The Twitter API will require twitter approval, so tweet-jason.txt was used insisted. Image-predictions.tsv was on Udacity server, so using the request url method was the choice.

Assessing Data:

During this step the data has been viewed visually. There are some qualities issues that have been addressed along with the tidiness issues. These issues as fallow:

- **Twitter_archive_df:**
  Qualities issues
  - timestamp column has +0000. This should be removed and the time should be extracted out.
  - timestamp column has dtpye as object. It should be changed to datetime dtype
  - source column shows the (href ..) link. The source should be exracted out of the link
  - name column has some rows with strange names. Name's examples (an, None, 0, a, Al, my, this, all, old, infuriating, the). It should be repalced so it will be replaced NaN.
  - Original tweets only. Rows with retweets and replay should be removed
  - Removing (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp) related to replays and retweets, because only the original tweets to be kept.
  - Some dogs appeare in more than one stage e.x. (doggo and floofer). Therefore, firstly getting rid of None with in the columns (doggo,floofer,pupper,puppo)
  - Represent dog stage one column dog_stage
  - There are empty values in dog_stage column replace with NaN
  - some of the values dog_stage column have two value in it e.x(doggopuppo) will have (,) to separate it.
  - There some decimal dog rating values which displayed in text column

  Tidiness issues
  - doggo, floofer, pupper and puppo should be represented in one column named dog_stage.
  - Joining tables with twitter_archive_df_clean the main dataframe.Main dataframe will be created, which will contain the 3 dataframes.

- **image_predictions_df**
  Qualities issues
  - columns(p1,p2,p3): showing first letter issue (capital and small letter) and underscore (_) need to be replaced with space
  - Convert the result of predictions into one column breed_dog. Relaying on (p1_dog,P2_dog,p3_dog) if it's true it will take the name from the name related to this prediction, which can be found in (p1,p2,p3). Breed_dog will take the name from the first true prediction.
- **tweet_df**
  Qualities issues:
  - Rename id column to be tweet_id
  - (id, retweet_count and favorite_count) these the only columns that should be kept, other columns will be removed.

Cleaning Data:

- Removing +0000 from timestamp column
- Changing the dtype for timestamp column
- Extracting the source from source column:
  twitter_archive_df_clean.source.str.extract('>(.*)<')
- Replacing the strange names with np.nan
- Dropping retweet and replays rows from the dataset (keeping the original)
- Removing (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp)
- Some dogs appeare in more than one stage e.x. (doggo and floofer). Therefore, firstly getting rid of None with in the columns (doggo,floofer,pupper,puppo)
- Represent dog stage one column dog_stage
- There are empty values in dog_stage column replace with NaN
- some of the values dog_stage column have two value in it e.x(doggopuppo) will have (,) to separate it.
- Changing rating_numthe decimal values in the text field
- Capitalize (p1,p2,p3) values and replace the underscore between names with space
- Create a function to get the name of the first true value from our prediction's formula e.x. (if p1_dog is true it will get the value from p1), and then store these values in new column breed_dog
- Dropping columns from tweet_df and keeping (id, retweet_count and favorite_count)
- Rename tweet_df['id'] to tweet_df ['tweet_id']
- Joining all the dataset together relaying on twitter_archive_df as the main dataset and tweet_id as the relational column