WeRateDogs Project report
Omar Samir Khalil Mohamed Soliman
19 March 2021

Resources (For 3 Data frames)

- Twitter-archive-enhanced.csv
- Tweet_json.txt
- https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

Wrangling process

- Gathering

  I have gathered my three data frames from 3 resources mentioned above from udacity resources

- Assessing

  I have assessed the three data frames both visually and programmatically using python methods info(), describe() samples(#Num) and head(), also I have needed to list some values to detect invalid values as I made to the name column to get invalid names.

- Ceaning

  There are issues I have found (All fixed and cleaned):

  Data types validity

  1. Twitter_archive_df
     a. tweet_id is int => String
     b. timestamp is str => datetime64
     c. in_reply_to_status_id is float => String
     d. in_reply_to_user_id is float => String
     e. retweeted_status_id is float => String
     f. retweeted_status_user_id is float => String
     g. retweeted_status_timestamp is str => datetime64
     h. numerator and denominator are int => float
     i. Clean invalid dog names like ‘a’, ‘such’, ‘getting’, 'quite', 'not', 'actually' … etc
     j. Refine HTML code from source column
     k. You can remove not needed retweet-related columns

l. Merge doggo, floofer, pupper, puppo columns into one column (unified_dog_stage)
m. Convert unified_dog_stage data type to category.
n. replace a none value in unified_dog_stage column with np.nan to get a real stats from the available data.
o. Remove duplicated rows with duplicated tweet_id and different unified_dog_stage values and set unified_dog_stage value equal **multiple_stages.**

p. Rename column headers to be more descriptive as img_num, p1, p1_conf, p1_dog as follows
    1. 'p1' = > 'prediction_1'
    2. 'p2': 'prediction_2'
    3. 'p3': 'prediction_3'
    4. 'p1_conf': 'prediction_1_confidence'
    5. 'p2_conf': 'prediction_2_confidence'
    6. 'p3_conf': 'prediction_3_confidence'
    7. 'P1_dog':'is_prediction_1_dog_breed'
    8. 'P2_dog':'is_prediction_2_dog_breed'
    9. 'P3_dog':'is_prediction_3_dog_breed'
    10. 'img_num':'dog_image_number'

2. Img_predictions_df
    a. tweet_id is int => String

3. Tweets_df
    a. favourites is float => int
    b. retweets is float => int

Tidiness

1. dog stages/breeds in multiple columns => All merged in unified_dog_stage column.
2. the three tables should be merged into one table => three data frames merged together in one data frame and saved to twitter_archive_master.csv

Analysis

1. I used a distribution plot to describe the distribution in retweets and favorites.
2. I used a pie chart to show the volume of dog stages/breeds
3. I used a regplot to show correlation between favorites and retweets and how can we predict future engagement for upcoming tweets.
4. Finally, I used a wordcloud to represent the most used dog names which show us that the majority of tweets records does not include names for dogs.

   * all there analytical figures mentioned in details in `act_report.pdf`

- # Stored results
  - I have stored all cleaned data merged together for 3 data frames in file twitter_archive_master.csv
  - Also there is another document for analytics `act_report.pdf`