

# wrangle\_report

February 10, 2021

## 0.0.1 WeRateDogs - Twitter Data

**1. Gather Data** I looked at the instructions given by the Udacity team on how to gather data for this data wrangling analysis.

- I downloaded the data twitter-archive-enhanced.csv from Udacity.
- Next I downloaded the file image predictions file which is in the tsv format.
- Then I tried creating my twitter developer account but didn't get approval from tweeter so I downloaded the tweeter\_api provided by udacity.

Once I had all the above three files, I created them into 3 different dataframes which are shown below.

- archive - this is a dataset "twitter-archive-enhanced.csv" which was converted into a dataframe and gives information on basic tweet data.
- image\_predictions - This dataset will contain information about predictions about the image
- tweets\_api - This dataset will contain information like tweet\_id, no of retweets and no of favorites etc.,

## 0.0.2 2 Assessing

- In the assessing step, I tried to gather some quality and tidiness issue from the data I gathered. I assessed all files concurrently for similar quality or tidiness issues

Below are the quality and tidiness issues I accessed ##### QUALITY

### Archive

- timestamp and retweeted\_timestamp in object form
- Correct denominators other than 10.
- Source columns have HTML tags
- tweet\_id is in int type in both archive and image\_prediction
- Rating\_numerator with decimal values incorrectly extracted
- incorrect dog names contained(a, an, just,infugured....etc) ##### Image\_prediction
- dog breeds inconsistent, contains underscores, and have different case formatting
- Rename Columns(p1,p2,p3,p1\_dog,p2\_dog,p3\_dog,p1\_conf,p2\_conf,p3\_conf)to improv clarity ##### Tidiness ##### archive\_df table
- the columns doggo, floofer, pupper and puppo should be variables in a column dog\_stage ##### All tables
- All three tables share the column tweet\_id and should be merged to archive\_df

### 0.0.3 3 Cleaning

For cleaning all the 3 dataframes, Below are the steps I took to clean the dataframes from observations made in the accessing phase

- I made a copy of all dataframes
- I joined the 3 dataframes based on their tweet\_id
- I converted the datatype of "tweet\_id" into string
- I created a column called dog\_stage for the ( puppo, pupper, doggo and the floofer stages)
- I dropped all duplicates including the ones as a result of creating the dog\_stage column
- I converted timestamp which was in the string format to the datetime format
- I removed the underscores from the dog breeds and removed the inconsistency in the name format which was a mixture of both lower and upper case
- I removed the "><" from the source column so that the information would be properly extracted
- I made all rows in the rating\_denominator 10 to remove row information that contained >10 or <10 since 10 is the only rating denominator
- I converted the ratings\_denominators that are in decimal to float datatype
- Rename Columns(p1,p2,p3,p1\_dog,p2\_dog,p3\_dog,p1\_conf,p2\_conf,p3\_conf)to improve clarity
- I changed incorrect dog names (a, an, just, atually,all....etc)to none
- I Rename Columns(p1,p2,p3,p1\_dog,p2\_dog,p3\_dog,p1\_conf,p2\_conf,p3\_conf)to improve clarity

### 0.0.4 Storing Data

- I stored the final dataframe into csv file with name twitter\_archive\_master.csv with final data of 2060 rows and 27 columns

In [ ]: