- I started by checking twitter-archive-enhanced.csv .I checked head(),tail(),describe .. and so on.
- I noticed timestamp is not as datetime format and tweet_id is considerd int (which is not correct as it is supposed to be read as string because we cant do operation like mean ,add, median to it) .
- I also noticed that max **rating_numerator** =1776 and max **rating_denominator** =170 which may be considered a typo. Moreover min = 0 at numerator and denominator(here it can make errors if division i guess).
- Next I noticed that 'doggo, floofer, puppo, pupper' columns are better to be looked as values in a united column "**stage of dog".**
- **Name** column has a lot of **none** names + some names are not actually names like (a,the…).I will replace them with none later.
- I checked if there was any duplicate tweet_id but there was not any.
- Next step was checking 'image-predictions.tsv'. I used same methods of programmatic assessing .
- I noticed all entries are are present and tweet_id is int so it also need to be converted to string. Furthermore, there was no duplicates in tweet_id.
- Then I contact Twitter and I made a developer account.I used the keys given to make file with data needed.
- I loaded the Jason.txt and made df with the loaded data,followed by programmatic assessment.
- The visualization of correlation between retweets and favorites. Positive correlation with r=0.80 was found.
- 25 **tweet_id** was missing**.(**which I guess I can not make anything about because I got it directly from twitters API)
- Next step I started cleaning.
- I started melting (doggo, floofer, puppo, pupper) columns. Then I remove the duplicate from this melting process .
- Next I converted tweet_id to str. Then I merged all 3 columns toghter.
- Then I continued fixing datatypes (timestamp, retweeted_status_timestamp) to datetime
- And (retweeted_status_id,retweeted_status_user_id) to str.
- After that I changed some strange names to NONE.
- Then I began visualization process .