Gathering data

1- First file imported with pandas library using -> pd.read_csv(file path)

2- Second file downloaded using request library then saved it into a dataFrame

Assessing data

-Checked duplicated values

-I found that there are many tweets are duplicated with different ids , then I realized it's a retweet

- rating numerator must be 10 or more , and there are values that wasn't normal like 900

- rating denominator must be 10

issues in quality and tidiness

Quality

1- in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id,retweeted_status_user_id and retweeted_status_timestamp have too many missing values
2- tweet_id or any id in any DF must be string ,, img_num in img_prediction_df must be string
3- timestamp have ('0000') must be removed
4- rating_numerator must be greater than 10
5- rating_denominator must be equal 10
6- timestamp must be datetime type
7- there are alot of none in name column in arch_df and some with lower and uppercase
8- remove retweets

Tidiness

1-doggo floofer ,pupper, pupper and puppo must be in one column

2- tweet data fram is related to arch_df must be merged

Cleaning data

1- i will drop this columns with too many missing values with -> df.drop('a', inplace=True, axis=1) (done)
2- id,img_num will be converted to string -> .astype(str)
3- slice timestamp to extract (" +0000")
4- convert timestamp to datetime

5- **all recored in rating_numerator will be > 10 and will be 14 for big numbers**
6- **rating_denominator must be equal 10**
7- **make all names with lowercase**
8- **remove retweets from arch_df by removing records with retweeted_status_id that's not null then drop this columns**
9- **merge doggo, floofer, pupper and puppo and put them all in stage column , then drop them all**
10- **tweet_info related to arch_df so I merged them In one data frame called tweet_data**
11- **stage data type changed to category**
12- **I cleaned source column from un needed words to get a clear source**