Wrangle Report

In this report i will explain what i did to initialize the datasets to be used in creating analyzes and visualizations for Twitter user @dog_rates .

I have done 3 main steps ; i will explain each step in details .

| FIRST STEP : Gathering Data .

- I loaded all libraries and packages i needed .

- I readed the 3 dataset needed : "twitter-archive-enhanced.csv ", "image_predictions.tsv" , "df_tweet" ; "df_tweet" is the file i got keys for the Consumer API keys, and the Access Token and Access Token Secret that you i needed from my developer account on Twitter .

| SECOND STEP : Assessing Data.

to assess any issues twitter_archive ,image_predictions ,and df_tweet ; so I found some issues need to proccess :

Quality Issues : I found 11 issues .

1- Find any (NaN)or Null Values and fill it in all dataframes .
2- Search on the text column of invalid names and replaced it by the text what i found about it and drop null values from it .
3- Drop retweets columns .
4- Convert the type of tweet_id in (twitter_archive_copy) from (int64) to (string).
5- Convert the type of tweet_id in (image_predictions_copy) from (int64) to (string).
6- Convert the type of tweet_id in (df_tweet_copy) from (int64) to (string)
7- Convert the type of rating_numerator from (int64)to (float) .
8- Convert the type of rating_denominator from (int64) to (float).
9- Fix outliers on Rating columns in (twitter_archive_copy) .
10- Merge rating_numerator column with rating_denominator in one column called 'Rating' .
11- Drop column no need it .

* Tidiness Issues : I found 3 issues .

1- Combine (doggo, floofer, pupper, and puppo) in one column named (dog_stage) & Drop (doggo, floofer, pupper, and puppo) columns .
2- Create list of prediction confidence as (p_con) and list of prediction_images as (pred).
3- Merge all the tables .

| THIRD STEP : Cleaning Data .

-Before any clean proccess , i copied all the 3 datatstes .

-I cleaned the issues what i found it on (Asses) step .

- The Quality Issues :

1- I find any null values then i filled it by [np.nan].

2- I search on the text column of invalid names(None. The , a) and replaced it by the text what i found about it and drop null values from it .

3-I Removed the retweets columns :

['in_reply_to_status_id' , 'in_reply_to_user_id' , 'retweeted_status_id' , 'retweeted_status_user_id' , 'retweeted_status_timestamp']

4- I Converted the type of tweet_id in (twitter_archive_copy) to (str)

5- I Converted the type of tweet_id in (image_predictions_copy) to (str)

6- I Converted the type of tweet_id in df_tweet_copy to (str)

7- I Converted the type of rating_numerator from (int64)to (float) .

8- I Converted the type of rating_denominator from (int64) to (float) .

9-I Fixed the outliers on Rating columns (rating_numerator column & rating_denominator) by do three steps :

1. create boxplot to show outliers value
2. calculate the (IQR) to used it to find The outliers :

```
Q1= twitter_archive_copy[f].quantile(0.25
Q3 = twitter_archive_copy[f].quantile(0.75)
IQR = Q3 - Q1
upper_limit = Q3 + 1.5 * IQR
lower_limit = Q1 - 1.5 * IQR
```

upper, lower = outliers(twitter_archive_copy, "rating_numerator")
and
upper, lower = outliers(twitter_archive_copy, "rating_ denominator ")

3. I unite rating_denominator on = 10 ; because it's the basement of the rating numbers, we couldn't have several numbers in denominator

4- fix the outliers & drop the outliers have invalid values in rating_numerator by –

-Checked the text of it if it have correct value in rating_numerator , then replaced it by the correct number of rate what I found it

-Filled each an other rating_numerator it didn't t have any number of rate in the tweets by null values ,then removed the null values in rating_numerator column

10-I merged the rating_numerator column with thr rating_denominator in one column called 'Rating' as float type by :

1. I divide the rating_numerator column with thr rating_denominator in new column named it rating
2. I changed the type of rating column to float
3. I removed the numbers have outliers value
4. I changed the outlier values (1776.0 , 75.00,50.0,27.00,1.0) in rating_numerator to null values then removed it .

11-I removed the column no need it ['timestamp', 'source' ,'rating_numerator' ,'rating_denominator']

- The Tidiness Issues :

1- I Combined (doggo, floofer, pupper, and puppo) in one column named (dog_stage)

and filled any tweet have none value in (dog_stage) column :

a- I checked how many (None) values in (puppo , pupper ,floofer , dooger)

b- I merged (doggo, floofer, pupper and puppo columns) to new column (dog_stage)

c- I removed doggo, floofer, pupper, and puppo columns

2- I created list of prediction confidence as (p_con) and list of prediction_images as (pred)

a- I used the function to check the value if True will collected
b- I created new columns : pred , p_conf
c- I removed rows that have None
d- I fixed the index after delete rows
e- I drop columns no need :

['img_num','p1','p1_dog','p1_conf','p2', 'p2_conf', 'p2_dog', 'p3', 'p2_conf', 'p3_dog', 'p3_conf']

3- I merged all datasets :

a - I merged image_predictions_copy with df_tweet_copy

b – I merged masterdf with twitter_archive_copy

| FOURTH STEP : Storing the data .

- I make a copy of masterdf :

```
cleandf=masterdf.copy()
```

- I store the cleandf :

```
cleandf.to_csv('twitter_archive_master.csv',encoding = 'utf-8', index = False)
```

| FIVETH STEP : Analyzing and Visualizing Data .

In Act Report file .
