Data wrangling Report

Introduction:

In this project we will do the process of data wrangling witch is gathering data, assessing data and cleaning the data.

A. Data gathering

In this project, we will gather three data set:

1- df_twitter_archive
importing it using pandas to read csv file into a data frame

columns: tweet_id, in_reply_to_status_id, in_reply_to_user_id, timestamp, source, text, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls, rating_numerator, rating_denominator', name, doggo, floofer, pupper, puppo

2- df_image_predictions
I downloaded this file from Udacity project section importing it using pandas

columns: tweet_id, jpg_url, img_num, p1, p1_conf, p1_dog, p2,p2_conf, p2_dog, p3, p3_conf, p3_dog

3- df_json
I downloaded this file (tweet_json.txt)from the project section and I reading it using pandas since I could not complete sitting the API because it take long time to make a developer account in twitter.

Columns: created_at, id, id_str, full_text, truncated, display_text_range, entities, extended_entities, source, in_reply_to_status_id, in_reply_to_status_id_str, in_reply_to_user_id, in_reply_to_user_id_str, in_reply_to_screen_name, user, geo, coordinates, place, contributors, is_quote_status, retweet_count, favorite_count'favorited, retweeted, possibly_sensitive, possibly_sensitive_appealable, lang, retweeted_status, quoted_status_id, quoted_status_id_str, quoted_status

B. Data assessing

Assessing data is the second step in data wrangling. When assessing inspecting dataset for two things: data quality issues (i.e. content issues) and lack of tidiness (i.e. structural issues).

Observations

Quality:

1- Remove all columns which are have retweets

2- Delete unnessary columns
3- There are some dogs have multiple stages, need to add 6 more columns, then add it to dog stages
4- Some of (doggo , floofer, pupper, puppo) clumns have None value insted of NaN
5- In the (rating_numerator)column data type is not decimal
6- In visual assessment I realize that in some rows there is a defferent numbers in (rating_denominator) which is an incorrect values, it has to be 10,There are 23 incorrect value in (rating_denominator) column
7- The columns (timestamp) and (retweeted_status_timestamp) the data type is not datetime variable
8- p1, p2, and p3 contain underscores instead of spaces
9- There are 66 jpg_url duplicated in (df_image_predictions) shoud be dropped
10- Change tweet_id from an integer to a string in all dataframes
11- Rename id in df_json to tweet_id

Tidyness:

1- No need to have 4 columns describe the stages of dogs (doggo , floofer, pupper, puppo)

2- No need to have 2 columns (rating_numerator) and (rating_denominator), we can combine it in one column

3- merge all 3 tabels in one dataset, the common column is (tweet_id)

C. Cleaning data

First, I created copies of all 3 data frames before cleaning data. I cleaned the data using thr define , code, test steps

1- Remove all columns which are have retweets
2- Delete unnessary columns
3- There are some dogs have multiple stages, need to add 6 more columns, then add it to dog stages
4- Change data type in (rating_numerator) to decimal
5- Change all values in (rating_denominator) to 10
6- Change the data type in (timestamp) and (retweeted_status_timestamp) columns to datetime
7- Replace underscore and dashes with spaces in p1, p2, and p3
8- Dropped all duplicated 66 jpg_url in (df_image_predictions)
9- Change tweet_id from an integer to a string in all dataframes
10- Rename id in df_json to tweet_id

11- Combine 4 columns describe the stages of dogs (doggo , floofer, pupper, puppo) in one column called (dog_stages)

12- Cobine 2 columns (rating_numerator) and (rating_denominator), in one column called (rating).

13- - merge all 3 tabels in one dataset, the common column is (tweet_id)

Conclusion:

Data wrangling is a very important skill that anyone handles data should be familiar with.

The most interesting thing I learned from this project is there are a several ways to clean a data and how to merge several data frames in one dataset