Introduction

In this project, I was working in Wrangling Data Project which is Analyzing "WE RATE DOGS Account tweets", this project was awesome and very powerful for practice what I learn during this course, without more talking let dig into the details of it.

For any wrangling project we work in 3 main steps to get the cleaned data and in step no. 4 we analyze it, the three steps are:

1- **Gathering Data.**
2- **Assessing Data.**
3- **Cleaning Data.**

Let's talk about each step in detail.

Gathering Data:

We collect data from 3 files, with different extensions: ".csv", "requesting and .tsv", "from Twitter API .txt".

The twitter developer account take some days to be approved which I should take in actual projects in the future.

The gathering process wasn't hard in this project because of your awesome guidance and saving the Data while working is also great.

Assessing Data:

The assessing project work in two steps, visually and programmatically and this get me a lot of notes like:

[Issues in enhanced twitter archive file]

1- There are 181 retweets (retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp)

2- There are 78 replies (in_reply_to_status_id, in_reply_to_user_id)

3- The timestamp field is in string format

4- There are 2297 tweets with expanded_urls which mean that 59 tweets with missing data

5- There are 4 columns for dog stages (doggo, floofer, pupper, puppo)

6- The columns related to retweets are not applicable for original tweets

7- The columns related to replies are not applicable for original tweets

8- Some of the rows from the tail() output below have invalid strings in the name column, e.g. "a", "an", "in".

9- The important cols are the end of the table

Issues in Image Predictions file

1- The p1, p2 ... names are difficult to understand

2- Drop unneeded cols

Issues in Image Predictions file

No issues here as I only get what I needed here

Cleaning Data:

Cleaning data process needs a lot of effort and I started by merging all the data frames by the tweet id and do all the work on it by removing retweets & replies, removing tweets having no images, putting all dog stages in one column, editing the dog names,.....

All these take a lot of time to apply, test for each step but it was amazing to research for solutions for problems which faced me.

Analyzing Data:

I apply the .describe() function on the cleaned data frame and get the average values and the 5 values which get a clear overview of what in the df.

After that some plots which will be presented in the *act_report.pdf.*

Conclusion:

This project was awesome and help me in the simple app I build with my graduation project to plot difficult files and continue now developing it.

So thanks a lot.