In This Nanodegree of Data Analyst, and as a data scientist student.I present the visual side of my coding, analysis, and all I have learned during this section of the nano degree course so far. So, starting from the data gathering point. Since we got many files to collect our datasets. In many different sources. Such as csv, tsv, and Json.
To Assessing data. Which contains Quality and Tidiness issues. But the cleaning section comes first. So, I took a copy before doing the cleaning set test in case of any damage. Then I start fixing all the issues I had written. to get store all the work in a csv file called “twitter_archive_master.csv”.
Finally, with the last part of the project is analyzing and visualizing the data I got.

Step 1: Gathering the data

1. I’ve downloaded the WeRateDogs Twitter achieve data
2. I used the file ('image-predictions.tsv') to be able to use the Requests library.
3. We have stored the (minimum tweet ID, retweet count, and favorite count) in a Jason file I named it ('tweet-json.txt')

Step 2: Step 2: Assessing Data
First let’s take a quick look at our data for the ("twitter-archive-enhanced.csv") and ('image-predictions.tsv') files. To know what are we dealing with.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 17 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   tweet_id                    2356 non-null   int64  
 1   in_reply_to_status_id       78 non-null     float64
 2   in_reply_to_user_id         78 non-null     float64
 3   timestamp                   2356 non-null   object 
 4   source                      2356 non-null   object 
 5   text                        2356 non-null   object 
 6   retweeted_status_id         181 non-null    float64
 7   retweeted_status_user_id    181 non-null    float64
 8   retweeted_status_timestamp  181 non-null    object 
 9   expanded_urls               2297 non-null   object 
 10  rating_numerator            2356 non-null   int64  
 11  rating_denominator          2356 non-null   int64  
 12  name                        2356 non-null   object 
 13  doggo                       2356 non-null   object 
 14  floofer                     2356 non-null   object 
 15  pupper                      2356 non-null   object 
 16  puppo                       2356 non-null   object 
dtypes: float64(4), int64(3), object(10)
memory usage: 313.0+ KB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 12 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   tweet_id  2075 non-null   int64  
 1   jpg_url   2075 non-null   object 
 2   img_num   2075 non-null   int64  
 3   p1        2075 non-null   object 
 4   p1_conf   2075 non-null   float64
 5   p1_dog    2075 non-null   bool   
 6   p2        2075 non-null   object 
 7   p2_conf   2075 non-null   float64
 8   p2_dog    2075 non-null   bool   
 9   p3        2075 non-null   object 
 10  p3_conf   2075 non-null   float64
 11  p3_dog    2075 non-null   bool   
dtypes: bool(3), float64(3), int64(2), object(4)
memory usage: 152.1+ KB
```

There are many missing values and probably some duplicates as we can see very clearly. It appears that there are no missing values in the ('image-predictions.tsv') file, but there may be some duplicates. In addition, there are some issues with the type of data

Quality issues
As I mentioned above, Before I start cleaning anything I took a copy of all the 3 files.

1. Dropping the useless columns: Dropped 4 columns: *doggo, floofer, pupper, and puppo*. They aren't useless but I did not wasn't going to do anything about them in the first place. So, I prefer to drop them. Also, I dropped *in_reply_to_status_id, in_reply_to_user_id, 'retweeted_status_user_id, retweeted_status_timestamp.* Because of the many null issues.

2. Converting the string column of timestamp to datetime: Changing the *timestamp* section and splitting it into 3 subsections.
Year, month, and *day* to be more specific during the analysis phase.

3. Remove time from data timestamp column. (keeping date only): Since the tweet time isn't useful to me. (also looks bad). Removing it from the timestamp column was a good decision.

4. Rename timestamp column: Renaming the timestamp column into *day_of_tweet* because it is clearer now than the previous one.

5. Clean texts: Removing symbols such as *&amp to "&",* and remove *\n.*
I guess this will not be clear or useful for others to know. it is just noisy to look at.

6. Delete retweets: Deleting another column. *retweeted_status_id.*
I figure it out a little bit late. That does not add any good value to me.

7. check duplication: Luckily, did not have any duplicates in any of the rest of the columns.

8. Convert to category datatype: Changing the *Source* column datatype into Category instead of Object. Because it is small.

9. Retweet Sources: Changing the form of HTML to a readable and easier form. So, everyone can get to know the source of the Twitter users.

10. Incorrect Dog Names: As mentioned in Project Motivation, the dog stages need to be cleaned.
So, it will show only the rows with multiple dog stages. Even the dog names are not actually names, like a, the, or such. Plus, some of the names are lowercase and some of them are in uppercase.

Step 3 Tidiness issues observations:

1- dog_type: the 4 dogs types were melted into this single column.

2- df1_copy: all data frames were stored in this one.

Step 4: **Storing Data**

Last but not least, this section is about storing all the changes we make in the gathered, assessing, and cleaning sections into a csv file called *twitter_archive_master.csv.*

Step 5: Analyzing and Visualizing data

Finally, this is the last part of the project. So, we report it in the other report file called *act_report.pdf.*