First, I will define the issues as quality and tidiness :

Quality issues:

- Dealing with null values.
- In the name of the dog in 'df' have some issues some times is written a, an ,the .
- The data should be non-retweets.
- Change 'None' in 'name',doggo','floofer','pupper','puppo' colums to null.
- Drop the replayes cause i only need the original tweets
- Change the rating for tweet_id = '666287406224695296' to 9/10.
- Drop the tweet that has tweet_id = '810984652412424192' because there is'not rating in the tweet.
- Change the 'timestamp' from String to datetime.

Tidness issues:

- Solve the 'doggo' 'floofer' 'pupper' 'puppo' colums and make it a one colume named type of dog.
- Change the names in 'image_predictions' to make it easier to understand.
- merge the three tables.

And in the cleaning phase I cleaned those issues and combined the 3 data frames into 1 data frame.

So after that I started the analyzing and visualization phase :

I will explain every one with the code and every thing.

First insight : I want to find the dog who have the largest rating and print the picture.

I used this code :

```
# i want to find the dog who have the largest rating and print the picture.

im =df_finale[df_finale['rating_numerator'] == df_finale.rating_numerator.max()]

img_url = im.jpg_url

img_url=img_url.values[0]

from PIL import Image

import requests

img = Image.open(requests.get(img_url, stream=True).raw)

img
```

to print this picture :

The second insight: I wanted to see which prediction phase have identified dog are the most ?

Code :

```
df_finale.p1_isdog.value_counts()

df_finale.p2_isdog.value_counts()

df_finale.p3_isdog.value_counts()
```

so from this we knows that phase 2 have identified 1479 dog and it's the best phase.

The third insight : I have been curios about which are most popular name for dogs.

Code:

```
df_finale.name.value_counts()
```

from that i notice the name Charlie is the popular name for a dog.

Visualization:

I had a question about Which dog_stage in the dogs is the most?

So I used a bar chart to see which one :

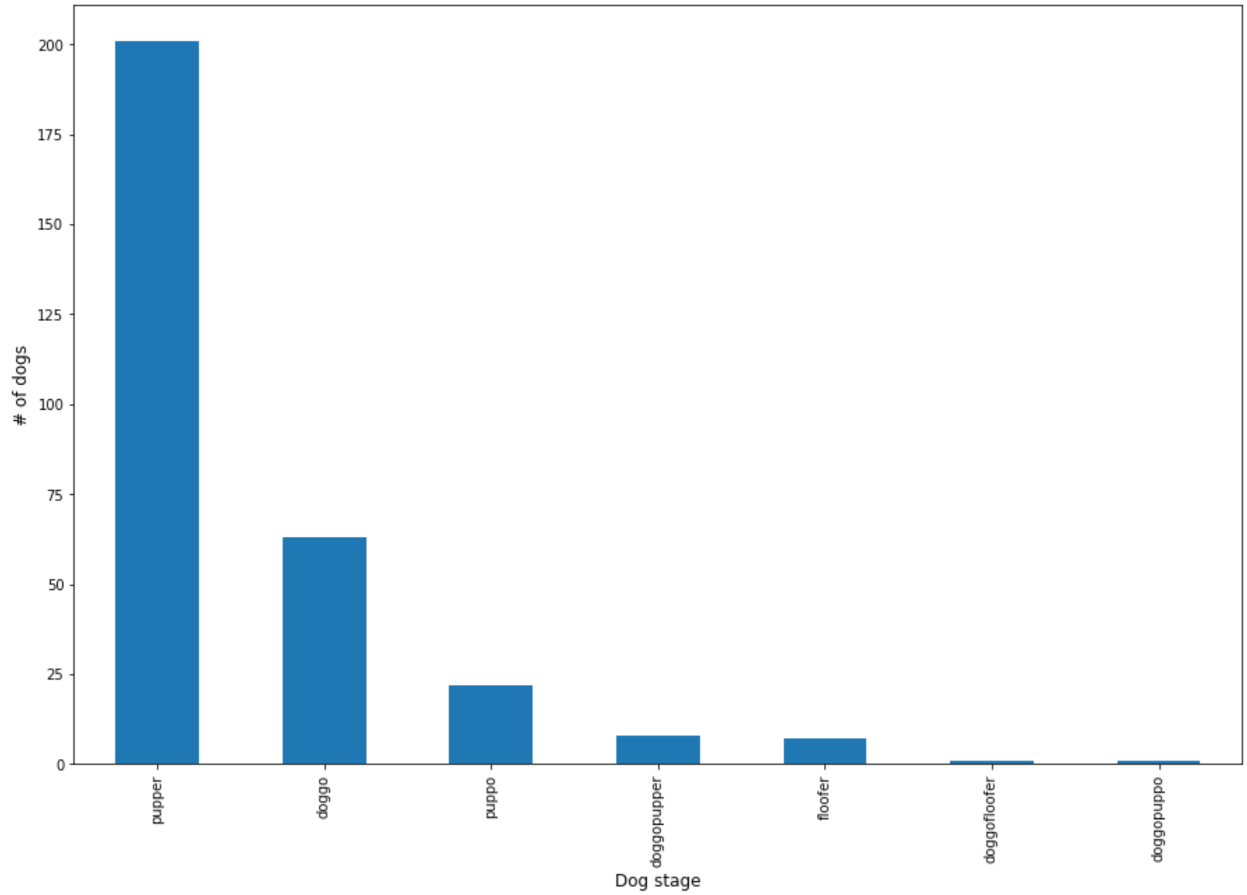
Code :

```
axx=df_finale.dog_stage.value_counts().plot(kind='bar',figsize=(15,10))

axx.set_xlabel("Dog stage",fontsize=12);

axx.set_ylabel("# of dogs",fontsize=12);
```

output :



and that’s was my project.