Gathering, Cleaning and Assessing

We are dealing with three files here:

- Twitter-archive-enhanced.csv – This has all the information that was pre-extracted from twitter
- Image-predictions.tsv – This has all the data about the image that was uploaded with the tweet. The image was run through a neural network. It shows the dog type in the image.
- Tweet_json.txt – This is data that is extracted from twitter using an API after which it was saved locally and then loaded into a pandas dataframe.

For Gathering –

- Having never used Twitter API, it was a good exposure for me to use the Twitter API and access the data programmatically. Tweepy was the library I have used for this exercise.
- The csv file was downloaded manually and stored in a pandas dataframe.
- The tsv file was downloaded using the request library programmatically and then stored in a pandas DataFrame
- All of the above mentioned packages have been used in combination with pandas and numpy library.

For Assessing and Cleaning –

- Have used numpy and pandas inbuilt functions extensively for assessing.
- Majority of the issues seen here were around having incorrect data types.
- There were a lot of junk data as well including the tweet urls and image urls which does not help any kind of analysis.

For Visualization –

- Only used seaborn given the ease of use as well how well the default format is.
- For this exercise, I have primarily made the use of horizontal bar plots since they reflect the story around the data correctly.