Data Wrangling report for WeRateDogs assignment

- Mukul Pathak.

Gathering data

The data were made available with three different .csv files. These three files were :-

1. **twitter_archive_enhanced.csv** : This file was provided from the beginning.
2. i**mage_predictions.tsv :** This file I downloaded with the help of program. This tsv file contains the neural network program to predict dog breeds.
3. **tweet_json.txt :** This file was again gathered programatically with the help of twitter API and tweepy package. This file contained the tweet id, favorite count and retweet count.

Assessing data

To assess the data properly and to find as many issues with the data possible, I relaid upon :

1. .head()
2. .describe()
3. .info()
4. .sample()
5. .value_counts()
6. .isnull()
7. .duplicated()
8. .rating_numerator()

After going through the outputs of the above mentioned, I took few decisions and the resultant issues found out were the following,

Data Tidiness

1. The columns doggo, pupper, puppo, and floofer are variations of an entity so they all can be merged.
2. For a better result, all the three dataframes should become a single dataframe.

Data Quality

1. In many dataframes, there's issue of wrong datatype. In tweet_archive, we have issues with tweet_id, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and timestamp.
   In doggo_df the issue is with tweet_id. In doggo_data the isssue is with tweet_id.
2. There are issues with dog's name
3. Missing image data of breed predictions
4. We don't need reweets.

Cleaning the data

When data assess was successful and thorough, data cleaning became quite easy, All I had to do was solve the above mentioned issues related to data tidiness and data quality. With merging all the three DataFrames into one, things became quite simpler and I did not had to repeat same tasks on different DataFrames and make a fun journey mundane.

With various methods and techniques all the issues were resolved and data became better for the next stage ie, visualisation.

Analysing and Visualisation

Now comes the part which makes the whole process look beautiful. With different plots and graphs, the cleaned data was very well represented. Which later helped me in analysing and reaching to a conclusion. Overall, the journey of data wrangling was quite challenging and amusing at the same time.