Report

1.Gathering data

In this first part we gathered the data that we would be working on from three different sources. The data is the @WeRateDogs Twitter data, where @WeRateDogs is a popular Twitter hash tag, as the name tells, people rate dogs with a denominator of 10 and the numerator is usually higher than 10 to show how lovely the dog is.

Type of methode that we used to gather this data :

1. We downloaded and uploaded the (twitter_archive_enhanced.csv) into a dataframe
2. We used the Request library to fetch and download data from a link where the data was represented in a tsv file
3. We gathered data from twitter archive using twiter API (tweepy library)

2.Assessing data

in this part we used ou knowledge to assess our data visually and programmaticly where we found some quality issues that were :

1. existing of retweets in **text** column
2. in the lines 2352, 2353, 2354 the **name** columns contains ('a') is an error, not a dog name + the existance of irregular names such as: ['a', 'the', 'an', 'very', 'just', 'quite', 'one', 'getting', 'actually','mad', 'not', 'old', 'life', 'officially', 'light', 'by', 'infuriating','such', 'all', 'unacceptable', 'this', 'his', 'my', 'incredibly','space']
3. **timestamp** and **retweeted_status_timestamp** need to be datetime
4. **tweet_id** is integer (should be a string type)
5. some of the **rating_denominator** values are different than 10
6. some values in the **p1,p2,p3** columns are lowerCase where others are UpperCase
7. some of the **rating_numerator** values are the composition of multiple ratings, some of them are not even ratings
8. **Missing data** in the twitter_archive dataset in column (expanded_urls)
9. **tweet_id** is integer type, it needs to be a String

And also we found a couple of tidiness issues :

1. Merge the tweet_json table and twitter_archive table and and image_predictions
2. the columns doggo, pupper, puppo, floofer need to be in one column

3.Cleaning the data

In this part, Numerous techniques including pandas join, combining multiple columns, pandas subsetting, eliminating missing values, and others, were used to clean up the data quality and tidiness issues listed above.

4.Saving the data

At the end, we save our data (the clean version) into a csv file using the .to_csv methode.