This project contains the three steps of Data wrangling: Gathering, Assessing, and Cleaning the data.

As for gathering:

Data was collected in three ways:

1-Gather the data from a website and save the data in a folder and Data Frame. (Images)

2-Read a file using read_csv.

3-Read a json file.

Assessing Data:

It is done by checking the duplicated values, nulls, and info of the data frames. Then, writing down the issues for the next step which is cleaning.

There are two types of issues Quality and Tideness.

Quality issues:

- Tweet_df:

1-drop the columns that we don't need (will keep [id_str,created_at, favorite_count,full_text, retweet_count].

2- drop columns with null values

3-id_str should be an object as it is the id of the tweet and should be changed to tweet_id to match the other dataframes.

- Twitter_archive:

1-drop columns that we don't need (will keep [tweet_id,timestamp,rating_numerator,rating_denominator,name, doggo,floofer, pupper,puppo].

2- drop columns with null values

3-tweet_id is an integer and it should be an object.

- Image_predection_df:

1- tweet_id should be an object.

2- change p1 to prediction1.

- Master_df:

1-Remove unneeded columns.

Tideness:

1-Stages of dogs in Twitter_archive should be in one column.

2-Collect all the data frames into one master data frame.

Cleaning the data frames:

It started by taking a copy of each of the data frames to ensure the safety of the data.

Then dropping the columns unneeded.

Drop columns with null values.

Changing columns names or type to match the rest of the data frames.

For the tindeness part:

In the twitter archive data frame the dog stages where in 4 columns which isn’t convenient. For that the combining was needed into one column which is stages_of_dog.