I. Data Gathering

The data was collected of the three (3) different sources on the Project Details page.

1) Gathering csv.
2) Gathering url.
3) Gathering API.

II. Data Assessment and Cleaning

The assessment and cleaning work was in csv archive. I found seven task of Quality aspect. The issues are:

1) CSV df_twitter table
ii.> In expanded url column, there are null values.
iv.> Some rows have different rating_denominator than 10.
vi.> There are some names of dog with transcription error (for example: a, an, all, by, my).
vii.> In column text, some values have a special character wich "".
viii.> Some column can be removed such as reply_to_status_id, retweeted_status_user_id, retweeted_status_id, retweeted_status_timestamp.
ix.> The column timestamp should be a datatime type for make time series plots.

2) URL df_twitter table
a.) Text of columns p1, p2 and p3 have mayusc and minusc.
b.) Text of columns p1, p2 and p3 have special character (for example: "_")

I found three task of Tidiness aspect. The issues are:

1.> rating_numerator and rating_denominator could be a only one column.
2.> The variables doggo, floofer, pupper and puppo could be a one column that describe the type of image or something about this.
3.> The text column, contain url that could be other column.

The clean was focused in:

- Create a column rating (This is rating numerator / rating denominator, fixed some outliers).
- Transform timestamp to datatime type. For this I used the function to_datatime.
- Create a new column call meme. This column describe the type of image. If the image is doggo, floofer, pupper and puppo
- Also was replaced some name with transcription error (for example: a, an all, my, the). For fix that, I used the function replace.
- I drop many column was I not used. For this I used the function drop.

I clean the dataset was saved and stored in a csv (df_tweeter_clean.csv)