Wrangle Report

Data Gathering

gathering ‘twitter archive’ and ‘image_predictions’ was straight forward ,but gathering tweets info from tweet API was pretty challenging. It takes many steps to achieve it. Connecting to tweeter API then asking for tweets info though listed tweet IDs then saving it to local file with txt format then using JSON library to parse the file then make DataFrame from it.

Data Assessing

Visual assessing through Jupyter notebook was challenging because the full text doesn’t appear. One way I used to assess the data was thought ‘LibreOffice Calc’ this way doesn’t show all issue ,but it shows the pretty obvious ones.

Data Cleaning

first thing I made sure of that taking copy of each DataFrame to prevent losing the data.
Rating_numerator and rating_denominator columns has been extracted with a lot of mistakes like taking only the fraction instead of the whole value and extracting a date instead of the rating. I needed to do the extraction again instead of fixing each value separately. It took two days to learn regular expression.
dealing with new type of issues was overwhelming but looking back they seem easy to do.

After all, the data-set contains a lot of missing values like ‘ dog stage’ it has %84 missing values. I tried to find an API that takes the dog breed found by the image-predictions and return the dog-stage ,but I didn’t find any. Another issue I found that image-predictions has a lot of low confidence. I assumed that %70 is the deal breaker. It was challenging and enjoyable journey. It took me two weeks to do it.