wrangle_report

August 2, 2022

0.1 Reporting: wrangle_report

In this project,I gathered, assessed,cleaned and stored the cleaned data.

The gathering part was divided into three sections as the methods of getting the data differed.

1)Twitter-archive_data.I downloaded it manually by clicking a link that was given, twitter_archive_enhanced.csv. Once downloaded, I uploaded it and read data into a pandas DataFrame.

2)Image-predictions.I downloaded this data programmatically using the Requests library using a url that was provided.

3)Tweet_json data. I used the one provided by Udacity.

The three datasets are df_arc,df_img and df respectively.

After gathering the data,i went to the next part which is assessment. The assessment requried me to look for quality and tidiness issues using visual assessment and programmatic assessement. In visual assessment i looked at the excel files and realised things like the rating numerator and rating denominator having values that are not consistent,in the columns with the different dog stages none was used instead of nan. In programmatic assessment for example,I saw timestamp dtype not in datetime dtype.

After assessing i went aheadto clean the data and came up with the final cleaned dataset,df_master which i stored in csv format as twitter_archive_master.csv