wrangle_report

March 31, 2021

0.1 Project 4 : Wrangeling the data

Through this project our main focus was to wrangle the data which consisted of many steps Gathering,Assessing, and cleaning.

0.1.1 Gathering

which is the most important step due that we here has collect the required data to extract the desired insights , we gathered the data in multiple ways

> Downloading the TSV file ('image-predictions.tsv') programmatically then load it into data frame Uploading the twitter-archive-enhanced.csv file manually through the notebook page Loading the tweet_json.text to a data frame

0.1.2 Assessing

this step comes after gathering all the needed data and its done by two ways visually and programmatically which we focus on seeing what needs to be proccessd before analyzing and extracting insights.after this I concluded the below issues that need to be taken care of:

Quality issues 1.change the type of timestamp to be datetime, through this step we need to assgin the correct type for this column which will ease any steps that can be done through this column

2.remove unwanted columns and the ones that does have alot of null values,in this step it is better to delete which will make querying the columns more faster also for the columns that has alot of null values these are useless so it is better to get ride of them

3.convert a and an to none instead in name column as we can see from our inspection there is around 70 entries that needs to be converted

4.change columns name in df_images> >'p1':'prediction_of_golden_retriever', >'p1_conf':'prediction_confident_1', >'p1_dog':'result_for_prediction_1', >'p2':'prediction_of_Labrador_retriever', >'p2_conf':'prediction_confident_2'

5.seperate date and time into two columns which will help us in aggregating date or time alone when required to analyze the data

Tideness 1.dog calssifications must be one column with three variables doggo, floofer, pupper, puppo columns into one column

2.Combine three different dataframes into one master data set

0.1.3 Cleaning

for the cleaning part I used numpy and panda libraries for shaping the data