Wrangle and Analysis Data

Wrangle prosces

Introduction

In this report I will talk about the wrangle process briefly to explain the what was the problem with the data sets and how we get over it .

Bs:This report made for the udacity project review .

Wrangle Process

Firstly: I start with data gathering process by collection twitter_archive _enhanced Download this file manually then image_prediction file by downloading it programmatically by using request func and there was missing data related to tweets favourite and retweet count so I use api to collect the json file to extract the missing data

Secondly:I start to assess the data data visually and programmatically and notice the following observation:

quality Issue like

In tweet archive dataframe

- tag is include in the url "remove the tag"

- expanded_urls multi img links for same rating post in df_img

- tweet_ID should be a string in every dataframe

- in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id and

  retweeted_status_user_id is to be removed

- wrong data type for timestamp

- last five columns in archive data describing the dog stage and name is supposed to be nulls not none

- Errors in the naming of the dogs some such names like 'me,not,one,very ,the ,this,etc',officially which might have been a problem with the csv itself due to wrong extraction

image prediction

- rename columns for the img production as they are ambiguous names

- when joining tables we would not join retweets

- drop production model 2,3 as their coffecent are less accurate than 1

Tidiness issues:

- Information about one type of observational unit (tweets) is spread across three different dataframes. Therefore, these three dataframes should be merged as they are part of the same observational unit.
- dog "stage" is untidy"Column headers are values, not variable names."there are 4 columns describing the dog stage which it should be clearly only one colun
- expanded_urls there is more than one link leading to the same website kepping one link and removing the rest should be ideal
- Lastly the rate was separated in different column 'rating_numerator','rating_denominator'

Third:Clean process

The I start to clean the data programmatically based on the above observation