Data Wrangling

Introduction

In this project, we will practice the process of data wrangling that we have learned. we will gather data, asses them by defining the quality and tidiness issues, then clean them.

Gather Data:

In this process I have gathered data from different files and import it to my notebook to start working on it.

1- **Twitter_archive**

   This data was provided in the project, I have download it and import it to my notebook using panda's function to read CSV files.

2- **Tweet_df**
   After downloading the json file (tweet_json.txt) from the project resources, I parse it using panda and extracted the data of interest.

3- **Image_prediction**
   I have downloaded this file from the project resources and import it to my notebook it using panda's function to read CSV. It contains predictions of what is the dog breed from the images.

Assess Data:

In this step, I will go through the files I have imported before and try to find quality and tidiness in these data files.

Quality of the data:

Going through the file and analyzing them I have found in the twitter_archive these quality issues:

- tweet_id,in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, and retweeted_status_user_id should be strings

- timestamp should be in datetime type.
- nulls in dogs name represented as none
- nulls is dogs staged represented as none
- some dogs does not have dog stage(doggo,floofer,ect.)
- null data in (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp)
- wrong rating_numerator value it should be greater than 10
- wrong rating_denominator value it should be equal to 10
- expanded urls are from different websites

going through the tweet_df I found this quality issue:

- retweet and favorate count should be integar.

Going through image_prediction I have found these quality issues:

- tweet id should be string
- null values

Checking the tidiness of the data these are the tidiness issue I have found:

- merge the three dataframes in one data frame.
- merge dog stages(doggo,floofer,ect.) in one column in twitter_archive.
- merge (p1,p2,p3) in one columns in image_prediction.
- timestamp have extra digits that is not related to time.
- dogs breed have are in three different columns
- drop columns which we are not interested in.

Clean Data:

Using what I have learn since the start of this course, I have fixed all the quality and tidiness issues. The codes that I have used to clean data is included in the zip file.