Data Wrangling Report

By Ahmed Abdelhamed Elgayar

The Udacity Data Professional Track Project 2 : this is a report of the main steps for data wrangling of "WeRateDogs".

A- Data Gathering:

1- Twitter_archive_enhanced.csv file, this file downloaded manually and uploaded to project workspace then imported to working environment using Pandas function "pd.read_csv".

2- Image_prediction.tsv file , this file downloaded manually and uploaded to project workspace then imported to working environment using Pandas function "pd.read_csv".

3- Tweet-json.text this file should be extracted from twitter via the Tweepy library by quering the API but i don't get the (consumer & access) so i used anther method that i downloaded manually and uploaded to project workspace then imported to working environment.

B- Data Assessment & Cleaning:

Here i investigated the imported datasets both visually and programmatically for quality and tidiness issues.

1- The visual assessment done on "Twitter_archive_enhanced" spreadsheet and then the programmatic assessment in Jupiter notebook for the 3 datasets.

2- In the jupiter notebook i mentioned all the data quality and tidiness issues

3- Cleaning data : for each issue in the project I defined the issues then the code to solve it then I tested my code.

Issue	Solution
Quality Issues	
all tweets_id are integers they should be string	Convert tweets_id to string type
Data types(consistency issues); All timestamps are object type	Convert time stamp to datetime data type
Inconsistent representation of null values as "None" strings in the (name, doggo, floofer, pupper, puppo) columns	Clean "None" strings in the (name, doggo, floofer, pupper, puppo) columns
Erroneous pet names like the letter "a" and "an"	Extract the correct names for ('a','an') if they existed
There were retweets and replies in the dataset that we don't use them	Remove retweets and replies in the dataset that we don't use them
Some tweets may not include any image so we should delete them	Delete tweets without images
Non-descriptive columns's names in the image_df dataset such as the predicted breed(p1,p2,p3)	Remane the columns in image_df suach as the prediction breed(p1,p2,p3)
Incorrect values of the rating_numerator and the rating_denominator and should be as a float type not int	Extract the right values for rating_numerator and rating_denominator as a float type
Tidiness Issues	
(doggo, floofer, pupper, puppo) columns in archive_df data frame should be in 4 columns instead of 1 column	Gathering all dog breed in one column
(p1,p2,p3) (p1_conf,p2_conf,p3_conf),and (p1_dog ,p2_dog ,p3_dog) coulums in image_df are in 3 columns instead of one	Reshape(p1,p2,p3) (p1_conf,p2_conf,p3_conf),and (p1_dog ,p2_dog ,p3_dog) to be in one column