Data Wrangling Report

Introduction

This project is a data wrangling project, which mainly focus on fixing the data quality and tidiness issues using python 3. We have gathered dog rating from 3 different resources each one represented in data frame.

1- Gathering from SCV file.
2- Import tsv file from HTML link.
3- Connecting with twitter API .

Gathering:

1- Gathering from CSV file that's given from Udacity course. I used pd.read_csv to import data to work space, which is stored in twitter_archive data frame.

2- Image prediction , what breed of dog (or other objects, animal, etc.) is present in each tweet according to a neural network. Data must imported from HTML link which is hosted on Udacity server and downloaded programmatically using the requests library and the provided url. (**https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv**), which is stored in image_pred data frame.

3- Tweets data which is using twitter API twhic is o import data. Unfortunately, my request has been rejected , so, I used json file instead, which is given from Udacity. This data stored in tweets data frame.

Data Assessment:

Tidiness

1- In table twitter_archive we have 4 coulmns (doggo, floofer, pupper, and puppo)have to be merged in one column called "type"

2- Merging image_pred with twitter_archive.

3- In tweets table we have 2 colmuns (favorite_count and retweet_count) have to merge with twitter_archive table.

Quality

1- Timestamp in twitter_archive datatype is incorrect.

2- Thers is some missing values in twiiter_archive and unnecessary columns have to be dropped

3- The standred rating_denominator is 10 and it includes some valuse less and more than 10 .

4- There is some rating_numerator less than 10.

5- Some of dog names is inccorrect

6 - The columns name in image_pred p1,p2 and p3 isn't clear

7 - Some of dogs name have lower case have to started with capital letters

8- Some of dog' type has None values.¶

Data Cleaning:

Tidiness

1- Merging 4 types of dogs in one column to make analysis easy.
2- Merging three data frames in one data frame. I encountered an issues in this step. After merging the rows double 4 times. So to solve this problem I dropped duplicated before merging.

Quality:

1- Convert some of columns datatypes.
2- Drop columns that unnecessary
3- Rapelace values that not equal to 10 or its multiplise by 10
4- reating_numerator less than 10 will be dropped because its mean its not a dog.
5- The dataset has dog's name such a, the, and an have to repalce it by None.
6- Some columns not understood, i will replacet by prediction and confidence.
7- Change lower case by upper case.
8- Replace None values by dog's name.