Wrangle Report

Introduction

In this report I will briefly discuss my wrangling effort during the course of the project. From gathering the data, accessing it visually and programmatically to cleaning the different data used in this project.

Gathering

The first dataset is a comma separated file twitter-archive-enhanced.csv(comma separated file) provided by the project supervisor. The second file was downloaded programmatically from a URL provided using the request library as image_prediction.tsv (Tab Separated File). The third file was gotten from a user's tweet on Twitter by querying Twitter API using my provided Authentication details.

Accessing

After the Datasets were gathered, I read them accordingly and accessed them both visually by using the head and tail methods to observe any physical issue I could easily see and programmatically using info, describe, shape, value_counts etc. methods, to fix out the quality and tidiness issue whatsoever. After I encountered bunches of these issues like wrong column data types, inconsistencies and null values and a host of others, I made sure they were all detailedly documented for cleaning sake to serve as a guide when cleaning the Datasets. Below are the issues derived

Quality Issues

1. tweet_id column in integer Datatype, also on image_pred Dataset and df-tweets Dataset
2. timestamp column has wrong datatype, also time_created column in df_tweets Dataset
3. Inconsistency in rating denominator in df_dataset
4. Irrelevance and composition of to many null values in the underlisted columns of df_dataset
    a. text
    b. in_reply_to_status_id
    c. in_reply_to_user_id
    d. retweeted_status_id
    e. retweeted_status_user_id
    f. retweeted_status_timestamp
    g. expanded_urls
    h. Timestamp and source columns are present in both df and df_tweet datasets

5. Some names are wrong in df dataset
6. Some dogs have None stage in df dataset
7. Multiple Dog stages
8. dog breeds and level of confidence in different columns in image_pred dataset

Tidiness Issues

1. Dog stage in df has 4 columns
2. 3 separate files

Cleaning

After accessing the datasets, I followed the documented issues, and devised programmatically ways to get rid of these issues. I used the Define, Code and Test methods to solve each issue. For example, I created new columns like dog_stage by melting the four stages columns by also adding **(multiple stages)** because some dogs are assigned more than one stage, extracting day, month and year into separate columns. I also created new Tables where need arises. I set all dogs with no assigned name as **'None'** which are disregarded when exploring.

Storing

After all the datasets are well cleaned and tidied, I merged the three datasets by using the inner join method on tweet ids. Finally, I stored the merged files as **Twitter Master Dataset** (twitter_master_dataset.csv); a comma separated file.