Wrangle_Project

Introduction:

In this project, I will gather, assess, and clean data for famous account on Twitter which called: WeRateDogs , in this account people provide images, humorous comment and rates of dogs then shared these details on the account. Trying extract these tweets of this account, assess then clean them (wrangling process) then act on it through analysis, visualization and/or modeling.

Data wrangling, which consists of:

1) Gathering data
2) Assessing data
3) Cleaning data

Gathering Data:

Here I am gathering three datasets.

1) twitter_archive: import the data on Jupyter Notebook then read it by panda library.

2) image_predictions: downloaded programmatically by using the Requests library then read it by panda library.

3) tweet_json.txt: it is a Jason file, converting JSON data and read it line by line to create dataframe then read it by panda library.

Assessing data:

After gathering the data, assess them visually and programmatically for quality and tidiness issues for cleaning the data later.

The four main data quality dimensions are: Completeness, Validity, Accuracy, Consistency. For data which considered tidy, each variable must form a column, each observation must form a row, and each type of observational unit must form a table. First, I am trying to define the info(), head() and describe() function .

After assessing the data visually and programmatically, I found some quality and tidiness issues.

Quality Issues:

1) 'twitter_archive' table

-tweet_id is integer not string

-'text' column which contain value &amp' instead of '&'

-'source' column has long value.

-'timestamp' column string not datetime.

-'name' column contain lower case and incorrect name.

-remove retweets column.

-Empty values in: in_reply_to_status_id,in_reply_to_user_id.

-Empty value in: retweeted_status_id,retweeted_status_user_id.

-Tweets which were without images.

-Tweets without rating

2)'image_predictions' table

-Remove (-) from p1,p2,p3 columns to be more readable.

Tidiness Issue:

1) 'twitter_archive' table

-'doggo','floofer','pupper','puppo' columns combine them in one column.

2) 'tweet_info' tabl

-Rename 'id' coloumn to 'tweet_id'.

3) All tables

join all three tables through tweet_id.

Cleaning Data:

In this step, I define some issue then write code for it to satisfy quality and tidiness by using panda's library.