WRANGLE REPORT

Project: We Rate Dogs - Data Wrangling and Visualization

This report summarizes my Wrangling and Cleaning efforts made for the Project “We Rate Dogs – Data Wrangling and Visualization”.

Following are the major 3 steps that have I followed to perform Data Wrangling:

1. Data Gathering.
2. Data Assessment and
3. Data Cleaning.

DATA GATHERING:

For this project, I have gathered three datasets using three different techniques:

i. Gathering Twitter Archive Dataset:

Twitter Archive was just downloaded and is in the form of csv file. So, it is imported in the project workspace using Pandas read_csv() method. It is the simplest and easiest of the three methods.

ii. Gathering Image Predictions Dataset:

Gathering this data set is little trickier than the previous one. The dataset is stored in one of the Udacity’s server. So, it is downloaded from the server using “Requests” library. Also, this file was in the tsv format, so it is imported in the workspace using Pandas read_csv() method with the “\t” de-limiter.

iii. Gathering Retweet and Favorite Counts:

This method is the most challenging and hardest in all the three techniques. The retweet and favorite count for each tweet in the Twitter Archive Dataset is fetch from the Twitter API using Python Twitter Library called “Tweepy” and are stored in a text file in the form of json data using json Library. This json text file is then read line by line to create the DataFrame.

DATA ASSESSMENT:

The second step in Data Wrangling. Here, I’m inspecting my gathered datasets for Quality and Tidiness issues. I have used both Visual and Programming techniques to perform the assessment. Some of the functions I used for Programmatic assessment are info(), describe(), head(), tail(), value_counts(), query(), sample(), etc. and some looping methods.

Below are my Assessment findings in terms of Quality and Tidiness:

Quality

- twitter_archive dataset:

- Missing expanded_urls.
- Retweet information's included in the dataset.
- Erroneous Data types (tweet_id, in_reply_to_status_id, in_reply_to_user_id, timestamp, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, doggo, floofer, pupper, puppo and tweet_id).
- Links are included in the text.
- Incorrect rating_numerator and rating_denominator. (e.g. for tweet_id '835246439529840640')
- 'None' written in place of null (NaN) in name column.
- Default names 'a','an','the','such','all', etc. in the name column data.
- There are names in the text which are not recorded in the name column data.
- Number of dog types stated by the text are not matching with the total counts on the dog types columns.

- prediction_data dataset:

- Erroneous datatype for tweet_id
- Different case structures for the values at p1, p2 and p3. (e.g. some are in lowercase and some are in uppercase).

- api_data dataset:

- Erroneous datatype for tweet_id

Tidiness

- twitter_archive dataset:

- Multiple columns with the same type of data in different columns - (doggo, floofer, pupper, puppo).
- **Data for same information i.e. tweets are stored in three different datasets.**

DATA CLEANING:

This is the Final Step in my Data Wrangling process and is for fixing the Quality and the Tidiness issues identified in the Data Assessment Step. Before everything, a copy of each dataset is made using copy() method and the Cleaning tasks are performed on these copied datasets. First, I worked on fixing the missing data's and then move on to fixing the messy data's - Tidiness and Quality issues.

For each issue identified, three steps are performed:

i. Define: Defining what and how to do to resolve the issue.

ii. Code: Actual coding and resolving the issue.

iii. Test: Testing if the issue has been successfully resolved.

Even though it is said as the final step, this step has been iterated itself many times as more issues came up while fixing and working other issues.