Udacity Data Analyst Nanodegree

Wrangle and Analyze Data Project

Project Details

The tasks in this project are as follows:

- Data wrangling, which consists of:
  - Gathering data
  - Assessing data
  - Cleaning data
- Storing, analyzing, and visualizing the wrangled data

Gathering Data:

The data for this project consist on three different dataset that were obtained as following:

- **The Twitter archive file:**
  The (twitter_archive_enhanced.csv) file was provided by Udacity, I downloaded it manually and stored it as a DataFrame.

- **The tweet image predictions file:**
  This file (image_predictions.tsv) is hosted on Udacity's servers; I downloaded it programmatically using the Requests library and stored it as a DataFrame.

- **Data from Twitter:**
  The (tweet_json.txt) text File with jSON structure, provided by Udacity, I read the tweet's JSON data from this file line by line into a list of dictionaries then create a DataFrame from this list.

Assessing data:

- **Visually:**
  By checking the Twitter archive CSV file in Excel.

- **Programmatically**:
  By using different methods:
  - info()
  - duplicated()
  - value_counts()
  - describe()
  - sort_values
  - head()

I found the following issues:

<table>
<tr><td colspan="2">• <b>Quality issues:</b></td></tr>
<tr><td>(twitter_archive) table:
- tweet_id is an int not string.
- Timestamp is an object not data time.
- Retweets columns are not necessary, as this project concerns wrangling and exploring original rating.
- By manually checking the top 10 rating_numerator with Rating_denominator =10, I found that:
- 1 not dog image.
- 3 without image.
- 6 extraction issues, ex:the actual rating is "11.27/10" while the rating in the set is "27/10</td><td>- Some columns will not be used for analysis.
- 'source' column contain HTML tags.
- Rating_denominator not equal to 10 in 23 rows:
- 1 row with denominator =0.
- 2 rows < 10.
- 20 > 10, the images in most of them are contains more than 1 dog.</td></tr>
<tr><td>(tweet_json) table:
- tweet_id is an int not string.</td><td>(image_predictions) table:
- tweet_id is an int not string.</td></tr>
<tr><td colspan="2">• <b>Tidiness issues:</b></td></tr>
<tr><td colspan="2">(twitter_archive) table:
- Dog Stages represented in four columns (doggo, floof, pupper, and puppo).</td></tr>
<tr><td colspan="2">(image_predictions) table:
- The table should be part of the master table.</td></tr>
<tr><td colspan="2">(image_predictions) table:
- The table must be merged with the other tables.</td></tr>
</table>

Cleaning data:

- First of all I create a copy of the three DataFrames to keep the originals.
- Example for cleaning:
    - Remove the HTML tags by replacing them with readable sources.
    - Drop the row with rating_denominator values not equal to 10.
    - Drop the rows with rating_numerator > 17.
    - Create column for dog Stages and drop columns :(doggo,floofer,pupper,puppo).
    - Merge the three tables.
- For each issue described in the assessing section I followed the approach of Define, Code and Test.
- I used different methods
    - astype()
    - type()
    - to_datetime()
    - isnull()
    - value_counts()
    - list()
    - drop()
    - replace()
    - sum()
    - sort_values()
    - extract()
    - head()
    - merge()