wrangle_report

December 11, 2018

0.1 Wrangle and Analyze Data

This project comes under the curriculum of the **Data Analyst Nanodegree by Udacity**. In this project, I perform Data Wrangling and then Data Analysis.

The 1st phase of this project is **Data Wrangling**.

In the Data Wrangling process, we first gather the data from various sources like Databases, Web Scrapping etc. Then we assess the gathered data both manually and programmatically for quality and tidiness of the data. After assessing the data, we clean it to get a well structured and organised data.

0.1.1 Gather

For this project, I gathered the data from three different sources. >01. The WeRateDogs Twitter archive from a csv file downloaded manually from a link provided and loaded into the dataframe twitter_archive. >02. The tweet image predictions data from a link programatically using the python library requests and loaded the data into the dataframe image_predictions. >03. Lastly, I used Tweepy python access library to access the tweet data for the Tweet IDs. I queried the tweepy library, then I got JSON data. I wrote the JSON data to a text file. Then I extracted tweet ID, retweet count, and favorite count from the text file line by line and loaded the data into dataframe tweet_json.

0.1.2 Assess

I first assessed the data visually and then programatically using python's pandas library. I used >01. The info() function to assess the datatypes, number of records in each column. >02. The isnull()funtion to look for the null values in the dataframes. >03. The describe() function to get the statistic summary of the dataframes. >04. The value_counts() function to get the number of the different values in a column.

Through my assessments both manually and programatically, I got some quality and tidiness issues in the dataframes. The issues are the following: #### Quality >01. Some columns like **in_reply_to_status_id** (2278) **in_reply_to_user_id** (2278) **retweeted_status_id** (2175) **retweeted_status_user_id** (2175) and **retweeted_status_timestamp** (2175) have very large number of null values in them. It is better to drop them. >02. Column **timestamp**, **rating_denominator** and **rating_numerator** have incorrect datatypes. >03. Sources are not mentioned properly in the **source** column. >04. A lot of dog names are like "None" and "a". These aren't the names of dogs in real life, usually. >05. The **rating_denominator** and **rating_numerator** columns have some invalid values. >06. We can extract the gender of the dog from the **text** column. >07. Replace underscores from the dog breeds given in the columns **p1**, **p2**, **p3** in image_predictions table.

>08. **tweet_id** in all the three dataframes is in int data type. It is better that it should be in string datatype. >09. tweet_json dataframe has 20 duplicate rows.

Tidiness

1. Keep only the true prediction column about the dogs instead of every prediction column.
2. Combine each dog stage column into one column.
3. The **rating_denominator** and **rating_numerator** columns can be converted into a single **rating** column.
4. All the three dataframes can be merged into a single dataframe **df_clean**.

0.1.3 Cleaning

The process of Cleaning contains 3 steps- >01. **Define**: In this step we define the issue related to the data and how we are going to deal with it using the code. >02. **Code**: In this step we enter the code to resolve the issue defined earlier. >03. **Test**: In this step we test whether the code we entered has resolved the issue or not.

For the Cleaning purpose I ,generally, used some assessment functions like info(), value_counts() and some Pandas Library functions drop(), merge(), apply() etc.

The 2st phase of this project is **Data Analysis**.

After Cleaning part of Data Wrangling, I analyzed the clean dataframe. Through my analysis I tried to find the answer of the following questions: >01. Which tweet ID has the maximum retweet count? >02. Which tweet ID has the maximum favorite count or likes? >03. Which is the most frequent source for tweeting? >04. Which is the popular gender of dogs? >05. Tweet ID associated with highest ratings. >06. Which is the most frequently adopted breed of dogs? >07. Which names do most people like to have for their dogs?

To understand the analysis in a better way, I plotted the graphs using Python's Seaborn library, Matplotlib library and plot() function.