Udacity Project

Wrangle and Analyze Data

By Hanan Alshehri

"Wrangle Report"

What is Data Wrangling?

Data wrangling is a core skill that everyone who works with data should be familiar with since so much of the world's data isn't clean.

We need to wrangle our data for good outcomes, otherwise there could be consequences. If we analyze, visualize, or model our data before we wrangle it, our consequences could be making mistakes, missing out on cool insights, and wasting time. So best practices say wrangle. Always.

Three Steps of Data Wrangling:

- Gathering

- Assessing

- Cleaning

In this Project, I will identify each step of the data wrangling process.

1- Gather

Gathering data is the first step in data wrangling. Before gathering, we have no data, and after it, we do.

So, I will gather each of the three pieces of data: the WeRateDogs archive, the tweet image predictions, and tweet's JSON data.

- **The WeRateDogs Twitter archive**

Which contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which I used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced." Of the 5000+ tweets, I have filtered for tweets with ratings only (there are 2356).

- **The tweet image predictions**

i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network.

- **The twitter API data**

Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

2- Assess

Assessing the data is the second step in data wrangling. When assessing, I'm like a detective at work, inspecting your dataset for two things: data quality issues (i.e. content issues) and lack of tidiness (i.e. structural issues).

So, I will detect and document at least eight (8) quality issues and two (2) tidiness issues.

The Quality Issues in this project:¶

1- The WeRateDogs Twitter archive:

- Erroneous datatypes: "in_reply_to_status_id, in_reply_to_status_id, timestamp and tweet_id".
- Unnecessary URL in "source" column.
- Contains retweets record.
- Incorrect dog names: "None", "a", "the", "an".

2- The tweet image predictions:

- Erroneous datatypes: "tweet_id".
- Rename p1, p2 and p3 columns.
- Unnecessary underscore in p1, p2 and p3 columns.
- Contains duplicated jpg_url.

3- The twitter API data:

- Erroneous datatypes: "tweet_id".
- Rename "id" to "tweet_id" to match other tables.

The Tidiness Issues in this project:¶

- Cause each variable forms a column: merge "doggo, floofer, pupper, puppo" columns to one column.
- Merge all these dataframes into one by using tweet_id.

3- Clean

Cleaning the data is the third step in data wrangling. It is where you fix the quality and tidiness issues that I identified in the assess step.

First, I will take a copy of the data to clean them, then I will do the manipulation on it

The process is divided to define, code and test. programmatically, I will solve either quality or tidiness problems. When I finished the cleaning process, I merge the datasets together to store the clean DataFrame(s) in a CSV file with the main one named twitter_archive_master.csv.