Data Analysis Nanodegree

Data Wrangling Documentation

By Reem Alhathbi

INTRODUCTION:

Real-world data rarely come clean. Using Python and its libraries, we will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling. We will document our wrangling efforts, plus showcase them through analyses and visualizations using Python (and its libraries) and/or SQL.

The dataset that we will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

The purpose and goal of this project are to create a trustworthy and interesting analysis and Visualization based on the wrangled data.

WHAT IS DATA WRANGLING?

The data wrangling process begins with data collection from various sources, followed by an assessment of data quality, and then by cleaning the data to create a dataset that can be used for exploratory data analysis.

PROJECT DETAILS:

This part of the project is divided into three steps, which are as follows:

- Data Gathering.
- Data Assessment.
- Data Cleaning

DATA GATHERING:

Three datasets are required for this project, which are as follows:

- **Twitter archive**: Import CSV for WeRateDogs Twitter archive using the provided Twitter-archive-enhanced.csv file.

- **Tweet image predictions**: -Programmatically download the image predictions.tsv file through the Requests library.
- **Twitter JSON:** Download data JSON file called tweet_json.txt.

DATA ASSESSING

Assessing is the second step in the data wrangling process: After gathering each of the above pieces of data, assess them visually and programmatically for quality and tidiness issues assessing. Visually, by printing the datasets in the Jupyter Notebook and viewing them via Excel. In addition, Pandas functions can be used programmatically. Finally, I sorted the problems into quality problems (which are issues in validity, completeness, accuracy and consistency, i.e., issues in the content), and tidiness problems (which are issues in the structure) and outlined them.

1-twitter_archive dataset:

- **Quality:**

1-Null values recorded as None and NaN (missing values)
2- tweet_id type will change tweet_id data type to string
3-convert timestamp to be Date Time and rename the column into tweet_date
4-We have some columns that contain unnecessary data delete unneeded column

- **Tidiness:**

1- The 4 different columns doggo, floofer, pupper and puppo, combine in one column represent stages_of_dogs"

2- image_predictions dataset:

- **Quality:**

1-tweet_id should be string type
2-The types of dogs in columns p1, p2, and p3 had some uppercase \ lowercase letters.
3- Change the underscores to spaces in the columns (p1, p2 and p3).
4- Delete The columns that will not be used in the analyses.

3- data_tweet:

- **Quality:**

1- id column should be named 'tweet_id' as the other data have
2- 'tweet_id' dtype should be string, 'sourse' change data type to category
3- Drop some columns that contain unnecessary data.
4- Source mixed html tag Rewrite the tweet source, from iphone.etc.

- **Tidiness:**

Merge twitter_archive_copy, data_tweet_copy and image_predictions_copy to df_merge dataframe

CLEANING DATA

After I have assessed the gathered data, I will fix the quality and tidiness issues. This is the third and final step in the data wrangling process. The following three steps are applied to each assessment point so that it can be cleaned. There are three stages:

1- Define: which states what I am going to do.
2- Code: The issue is fixed programmatically here.
3- Test: This is where I ensure that we have cleaned the point properly.

DATA STORING:

The twitter archive master dataset is stored in a CSV file. At this point, the dataset was successfully wrangled and ready for analysis.

ANALYZING AND VISUALIZING DATA

Questions:

1- Are there any relationships between retweets and likes?
To compare favorites and likes, I use scatter.

2- What is the most popular source?
To view the most popular tweets, I use the count plot

3- What is the most stage for dogs?
To view the most stage for dogs, I use a pie chart

4- What is the average monthly number of tweets?
I use line plot to view the average number of tweets per month