Udacity Data Analytics Nanodegree

Project 04 – Data Wrangling

Wrangle Report

1 – Data Wrangling

The data wrangling processes consists of three steps that for the first time there's a sequence, but it's totally iterable. It consists in gathering the data from the text file, then assessing programmatically and visually with pandas functions, furthermore classifying data in quality issue and tidiness issue. Then, cleaning the dataset with several pandas functions, and saving in a dataframe.

1.1 – Gathering Data

The gathering data is the very first step of data wrangling. In here, we must have in mind what kind of data we want to gather. That's why it's so important to understand the plataform you are gathering from. In this case, I used tweets-json.txt file provided by Udacity. This file was basically all tweets information in the json format.
To extract all data, it was necessary to make a few lines of code to extract everything needed and then create the dataframe.

1.2 – Assessing Data

In this step, it's very important to look at your dataset carefully. You can do this by two ways: manually or programmatically. Both ways are equally important, and cannot be left behind.

First, for the Twitter dataset, I glanced at the dataframe and check some tadiness and quality issues. Then, to make sure the issues would repeat themselves, I assessed them programmatically. And with this coding structure, it was possible to classify the dataframe's issue.

Take a quick look of how I separated them:

a) **Quality Issues**
   - Date is string, not datetime
   - Remove RTs, they're not the targeted data.
   - Date has inconsistent numbers (000)
   - Change text range from list (string) to int
   - `tweets_id`should be obj not int
   - Outliners numbers in `numerator` (45)
   - Change all `denominator` to 10.
   - `numerator`change to float.
   - Dogs names are incorrect, for example `a carrot`, `a might`.

b) **Tadiness Issues**

   - One column with 3 variables (full text, rate and url)
   - Put together the dataframes
   - Create a column with dog stages

1.3 – Cleaning Data

Cleaning data is very challenging part, because you must think what steps you are going to take, and avoid getting you data messy. For that, I first cleaned all tidiness issues and then all quality issues. Below are some pandas functions that I used in this very important step.

- pd.str.extract: to extract regex codes, for dates, names and urls.
- pd.merge: to merge two dataframes.
- pd.to_numeric: to convert string to float.
- pd.str.contains: to check a few sentences in the dog's name
- pd.str.islower: to get all dog's name that started with lowercase and delete it.

1.4 – Storing Data

After cleaning everything, it's fundamental to store your data. For that, I used pd.to_csv and created a new file called "tweets_cleaned.csv". With this cleaned dataframe, it'll be possible to make further analysis and understand the numbers behind the WeRateDogs' profile.