UDACITY

Data Analyst Nanodegree

Wrangle and Analyze Data

Wrangle Report

Introduction

In the Data Analyst Nanodegree by Udacity, the project “Wrangle and Analyze Data” consists in wrangling data, one part of the data analysis process, which is divided in three phases: 1. Gathering data, which means collect data from different sources and different file formats (depends on the project), 2. Assessing data, which means check and evaluate visually and programmatically the quality and tidiness issues of the data, and finally, 3. Cleaning data, which means fix the found quality and tidiness issues.

Project tasks are as follow:

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

Wrangling Data

Wrangle Report

Gathering Data

The Gathering Data process consisted in collect each of the three piece of data as described below:

- The WeRateDogs Twitter archive is a csv file, was provided by the Udacity instructor, I downloaded it manually, loaded in the Udacity workspace and reading as a dataframe with the pandas read_csv function.
- The tweet image predictions is a tsv file, hosted on Udacity's servers, it was dowload programmatically using the Requests library.
- Each tweet's JSON data was consulted using the Twitter API and the Python's Tweepy library. Then, a txt file was created with the entire set of JSON data and was read line by line into a pandas DataFrame.

Assessing Data

The Assessing Data process consisted in check for quality and tidiness issues as follow:

- Visually assessment: Each piece of data collected I checked in the Jupyter Notebook within the Udacity's workspace in the files section.
- Programmatically assessment: I checked programmatically each piece of data collected using pandas' functions and/or methods like info, head, sample, query, describe, value_counts, sort_values, dtypes, shape, duplicated and isnull.

Findings in the assessment process

Quality Issues

- Missing values in multiple columns.
- Incorrect data type in multiple columns.
- Unstandardized rating records.
- Retweets and replies stored as tweets.
- Tweets without image.
- Not readable records on several columns (like source).

Tidiness Issues

- Unnecessary columns.
- Dog stages' variable stored in multiple columns.
- Individual piece of data that must to be merged in one DataFrame.

Cleaning Data

Before the cleaning process, I made copies of the original pieces of data. The cleaning process consist in three sub-phases:

- Define: I defined the respective process to fix each issue.
- Code: I ran scripts of pure Python, pandas, re, datatime and numpy libraries to fix the issues identified in the assess phase. For example, I used functions and/or methods like drop, melt, drop_duplicates, merge, to_datetime, dropna, apply, among others.
- Test: I tested each scripts that I ran for ensured that the issues was fixed. For this sub-phase I used functions and/or methods like print, info, query, dtypes, len, among others.

Conclusion

- The wrangling data is essential in the data analysis process, and it is not always a linear and straightforward task.
- Wrangling data is challenging and worthy at the same time.
- Python give us several tools to gather, assess and clean data.