wrangle_report

March 4, 2019

1 Wrangel Report

In order to complete this data wrangling project, I repeatedly watched the teacher's video and carefully recorded how the teacher did each step. I really like the working framework that the teacher designed in the video to do data wrangling. Every step is very clear.

According to the courses instructions, my data wrangling project divided into three steps.

1.1 Step 1: Gathering Data

In this project, we need to gather three parts of data. 1. Enhanced Twitter Archive. Udacity offered this part of the data as a CSV file. So I just need to load the data using pandas read_csv method. 2. Additional Data via the Twitter API. I registered the Twitter account and applied for the consumer_key etc. So I could download this part of data and save it into a CSV file to use later. 3. Image Predictions File. Import `requests` library and use `get` method to get the data from Http link and save the data into CSV file to use later.

1.2 Step 2: Accessing Data

There are two ways of accessing data. Visulization and programmatically.

Firstly, I opened the three CSV files using excel. Find two quality issues in the `twitter-archive-enhanced` table according to the 4 aspects of Data Quality Dimensions.

Then, I used pandas offered methods like head, sample, info, describe etc to programmatically found data type problem, unreasonable values, mixed format data value etc quality problems.

Finally, I found 2 tidiness issus according to the defination of Tidy Data.

1.3 Step 3: Cleaning Data

The structure in the courses for cleaning data is very clear, so I use `Define`, `Code` and `Test` three step to fix the problems one by one.