Temperature Report

Ran Lyu
7810ICT Software Technologies
September 22, 2017

Abstract

With the development of data science, the importance of data mining, data analysis, and visualization is increasing for industries and academic. This assignment is a practice for data cleansing, extracting value from large datasets to do further analysis.
Data visualization is an important tool to find interesting evidence or abnormal value, which should be cleaned in the data cleansing step. After that, the data is more valuable for further analysis.
In this report, after analysing the yearly average temperature of different states in Australia, and major cities in China, we may safely draw a conclusion that usually the cities' yearly average temperature are higher if they close to equator.

Introduction

As mentioned in the abstract, data analysis and visualization is increasingly necessary and helpful to have insight for some problems. The purpose of this report is to extract information and summary from large volume of raw data, and using python data visualization library to make the data in a more readable way, such as table and plots, for readers.

There raw data is structured data from three excel spreadsheets:

(A) temperature by state,

(B) temperature by country,

(C) temperature by major city.

Two libraries are used to analyse data:

1. Data analysis from excel: python Openpyxl library,
2. Data visualization: python Matplotlib library.

Database Structure

Schema of Sqlite database temperature.db (table names, attribute types)

Table Name 1: GlobalTemperatureByCountry

Attributes	**Data type**
date	date
averageTemperature	real
averageTemperatureUncertainty	real
country	text

Table Name 2: GlobalTemperatureByState

Attributes	**Data type**
date	date
averageTemperature	real
averageTemperatureUncertainty	real
state	text
country	text

Table Name 3: GlobalTemperatureByMajorCity

Attributes	Data type
date	date
averageTemperature	real
averageTemperatureUncertainty	real
city	text
country	text
latitude	text
longitude	text

Table Name 4: Southern-cities

Attributes	Data type
city	text
country	text
latitude	text
longitude	text

Table Name 5: ChineseCityYearlyAvgTemperature

Attributes	Data type
date	date
temp	real
city	text

- ## Chinese City Temperature Data

Methods

The method to calculate the yearly mean temperature of each city in China is to create table 'ChineseCityYearlyAvgTemperature' to store the values that need to be further used and analysed from table 'GlobalTemperatureByMajorCity'. Because table 'GlobalTemperatureByMajorCity' size is large and what we want to analyse is just several columns and partial rows, so we select date, temperature, and city from 'GlobalTemperatureByMajorCity' table, and insert into the 'ChineseCityYearlyAvgTemperature' table.
Then sum of yearly total temperature, sum of months have records, city, and year are selected (group by city and then year). The yearly average temperature is total temperature/sum of months.

Results

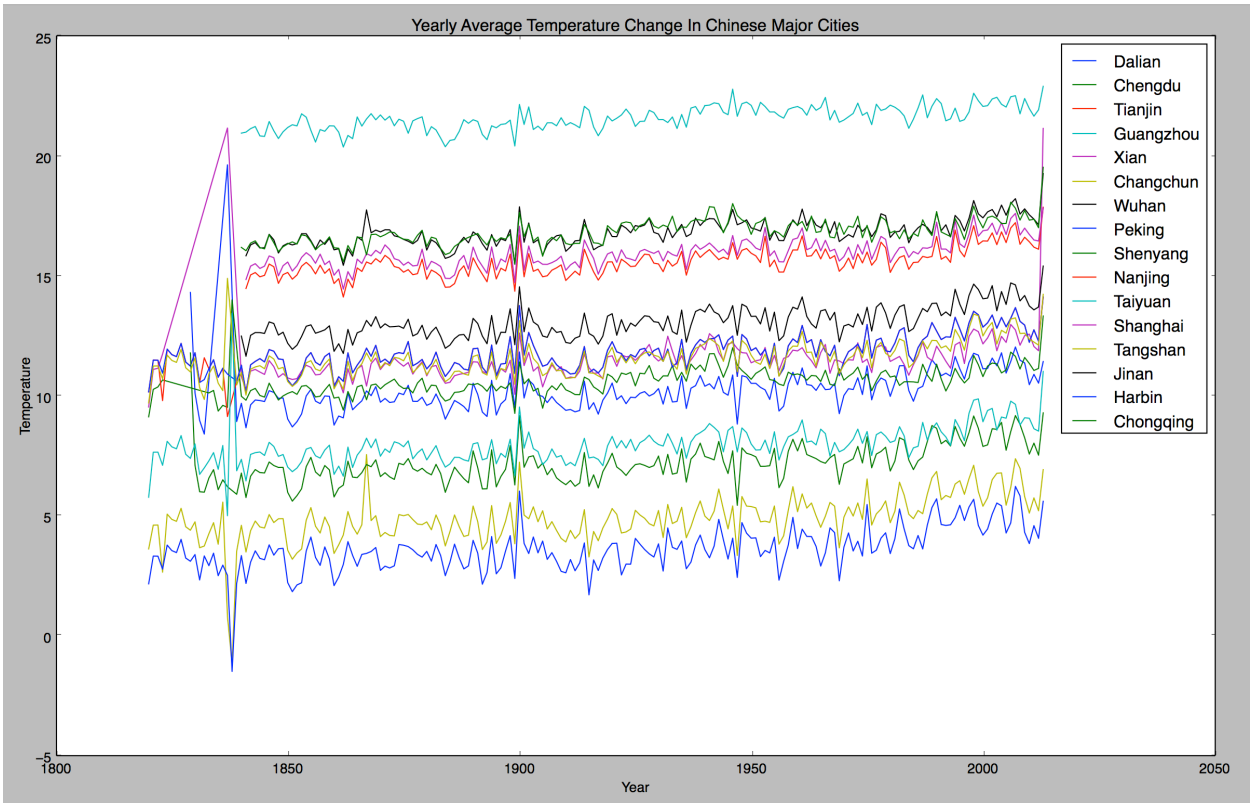

Figure 1

As we can see in Figure 1, there are some abnormal value evidence, such as the sharp increase and decrease between 1820 and 1840, especially Xian (purple), Dalian (Navy blue). After analyzing, I found that 'ChineseCityYearlyAvgTemperature' table has many temperature value 0. This is because there are some missing values in the original table 'GlobalTemperatureByMajorCity'. However, in the program the Null or None values are handled and considered as 0.

For handling missing data, two conventional methods are used (Marina, 2013).

First, listwise deletion method is used when the around 60% - 80% of value are missing in one year, such as Figure 2, 3, 4, &5.

236889	1823-12-01			Xian	China	34.56N	108.97E
236890	1824-01-01			Xian	China	34.56N	108.97E
236891	1824-02-01			Xian	China	34.56N	108.97E
236892	1824-03-01			Xian	China	34.56N	108.97E
236893	1824-04-01			Xian	China	34.56N	108.97E
236894	1824-05-01			Xian	China	34.56N	108.97E
236895	1824-06-01			Xian	China	34.56N	108.97E
237078	1839-09-01			Xian	China	34.56N	108.97E
237079	1839-10-01			Xian	China	34.56N	108.97E
237080	1839-11-01			Xian	China	34.56N	108.97E
237081	1839-12-01			Xian	China	34.56N	108.97E

Figure 2

176457	1837-12-01	-2.359	2.158	Peking	China	39.38N	116.53E
176458	1838-01-01			Peking	China	39.38N	116.53E
176459	1838-02-01			Peking	China	39.38N	116.53E
176460	1838-03-01			Peking	China	39.38N	116.53E
176461	1838-04-01			Peking	China	39.38N	116.53E
176462	1838-05-01			Peking	China	39.38N	116.53E
176463	1838-06-01			Peking	China	39.38N	116.53E
176464	1838-07-01			Peking	China	39.38N	116.53E
176465	1838-08-01			Peking	China	39.38N	116.53E
176466	1838-09-01			Peking	China	39.38N	116.53E
176467	1838-10-01			Peking	China	39.38N	116.53E
176468	1838-11-01			Peking	China	39.38N	116.53E
176469	1838-12-01			Peking	China	39.38N	116.53E
176470	1839-01-01	-5.424	1.986	Peking	China	39.38N	116.53E

Figure 3

1832-02-01			Dalian	China	39.38N	120.69E
1832-03-01			Dalian	China	39.38N	120.69E
1832-04-01			Dalian	China	39.38N	120.69E
1832-05-01			Dalian	China	39.38N	120.69E
1832-06-01			Dalian	China	39.38N	120.69E
1832-07-01	22.793	1.873	Dalian	China	39.38N	120.69E
1832-08-01			Dalian	China	39.38N	120.69E

Figure 4

1838-07-01			Dalian	China	39.38N	120.69E
1838-08-01			Dalian	China	39.38N	120.69E
1838-09-01			Dalian	China	39.38N	120.69E
1838-10-01			Dalian	China	39.38N	120.69E
1838-11-01			Dalian	China	39.38N	120.69E
1838-12-01			Dalian	China	39.38N	120.69E
1839-01-01	-5.587	2.159	Dalian	China	39.38N	120.69E

Figure 5

Second, mean imputation method is used when a small number of value are missing, such as one or two months in a year are missing (Figure6, 7).

237051	1837-06-01			Xian	China	34.56N	108.97E
237052	1837-07-01	24.505	1.733	Xian	China	34.56N	108.97E
237053	1837-08-01	22.41	1.706	Xian	China	34.56N	108.97E
237054	1837-09-01	16.647	1.586	Xian	China	34.56N	108.97E
237055	1837-10-01			Xian	China	34.56N	108.97E

Figure 6

1837-04-01			Dalian	China	39.38N	120.69E
1837-05-01	14.346	1.831	Dalian	China	39.38N	120.69E
1837-06-01	19.067	1.732	Dalian	China	39.38N	120.69E
1837-07-01	23.164	1.661	Dalian	China	39.38N	120.69E
1837-08-01	23.116	1.566	Dalian	China	39.38N	120.69E
1837-09-01	18.562	1.545	Dalian	China	39.38N	120.69E
1837-10-01			Dalian	China	39.38N	120.69E

Figure 7

After implementing these two methods to the missing data, the plot changed to Figure. According to Figure, there was an increase for all of the Chinese cities around 1900. Guangzhou, a city in South China, has the highest average temperature above 20, whereas Harbin, Dalian, Shenyang, cities in North China, have the lowest average temperature below 5. Other cities in the middle of China are sharing the same pattern and range from 5 to 20 degree.

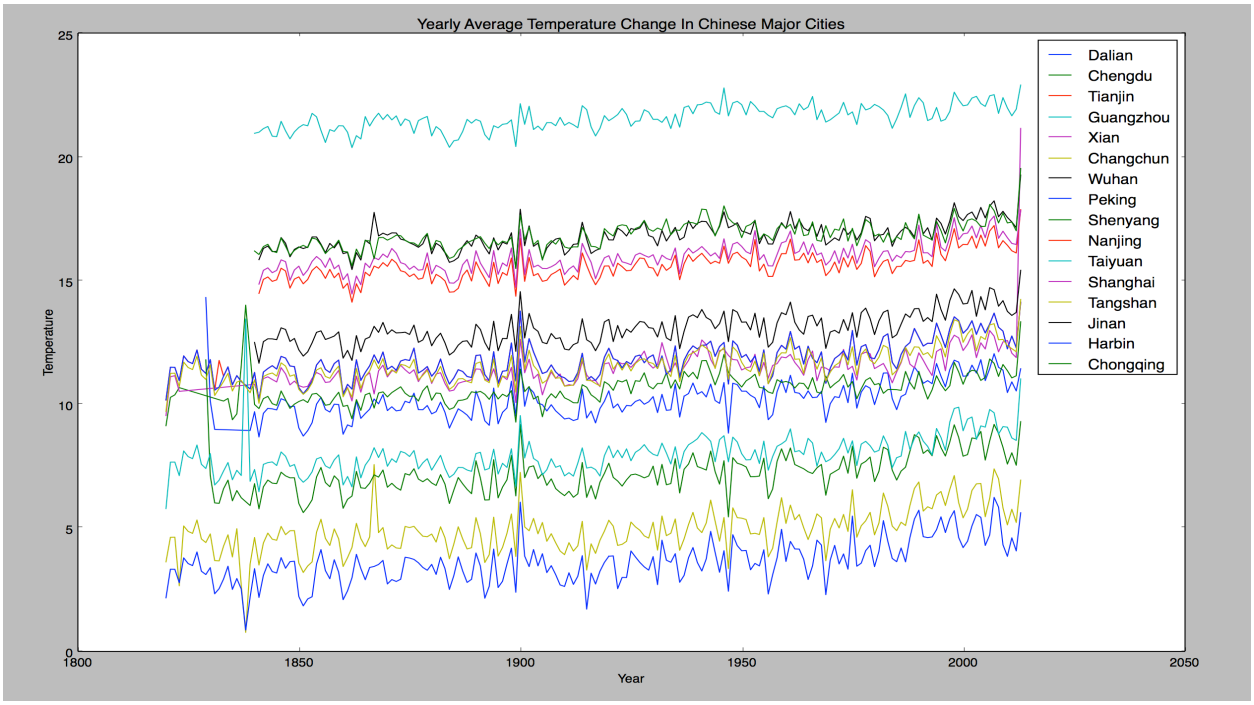

Figure 8

- ## Australian State and National Temperature Data

Methods

The method to get Australia state and national temperature is similar like the task3. Instead, rather than only calculate each Australia state's average temperature, the national yearly average temperature need to be append to the results for comparison purpose. Figure 9 is the initial rough plot for task 4. Apparently, from figure 10, Western Australia and North territory have some abnormal value. Check file 'World temperature.xlsx', we can see some value are 0 due to the missing data in the raw data file.

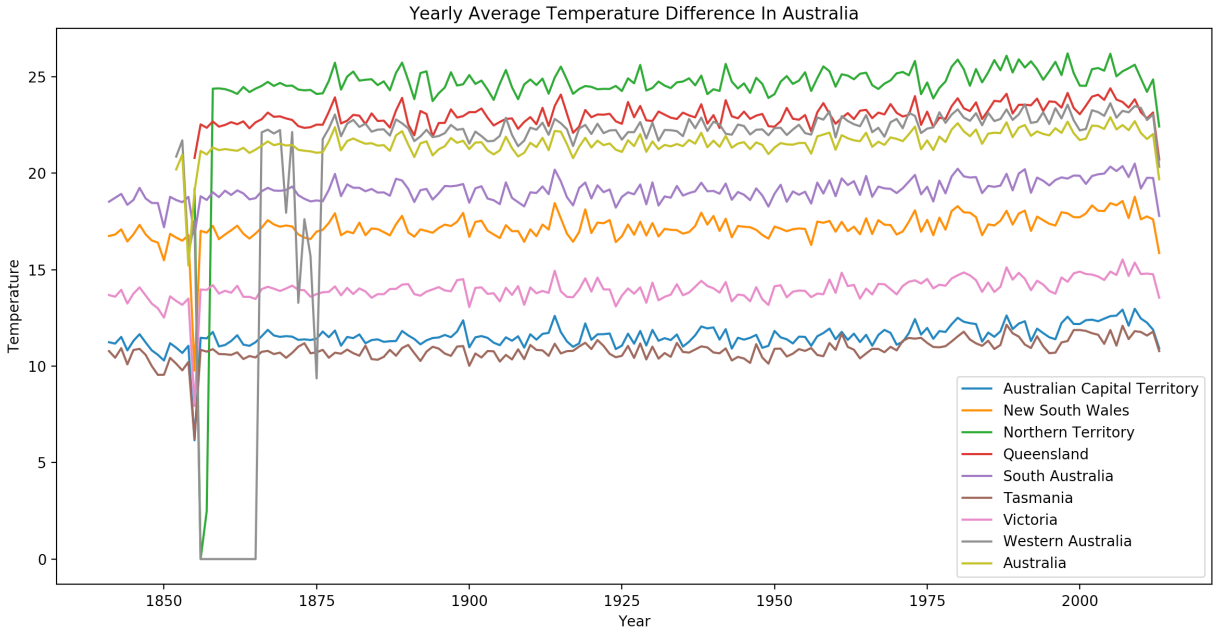

Figure 9

1188	17.52575	1855	Western Australia
1189	0	1856	Western Australia
1190	0	1857	Western Australia
1191	0	1858	Western Australia
1192	0	1859	Western Australia
1193	0	1860	Western Australia
1194	0	1861	Western Australia
1195	0	1862	Western Australia
1196	0	1863	Western Australia
1197	0	1864	Western Australia
1198	0	1865	Western Australia
1199	22.11117	1866	Western Australia

Figure 10

Data Cleansing

Two same methods as used in task 3 to handle the missing value are used in task4 as well. Figure11, &12 is to guess the missing value by calculating the neighbour two monthes average temperature.

1850-08-01	5.403	1.741	Australian Capital Territory	Australia
1850-09-01			Australian Capital Territory	Australia
1850-10-01	10.715	1.671	Australian Capital Territory	Australia

Figure 11

1850-08-01	5.403	1.741	Australian Capital Territory	Australia
1850-09-01	8.059	1	Australian Capital Territory	Australia
1850-10-01	10.715	1.671	Australian Capital Territory	Australia

Figure 12

Instead of supplementing the whole year's missing data, I deleted the rows with empty temperature. For example, in Figure 12, 24 rows in year 1856 and 1857 are deleted. After Cleansing the data, the results plot is like Figure 14.

1855-11-01	28.33	1.798	Northern Territory	Australia
1855-12-01			Northern Territory	Australia
1856-01-01			Northern Territory	Australia
1856-02-01			Northern Territory	Australia
1856-03-01			Northern Territory	Australia
1856-04-01			Northern Territory	Australia
1856-05-01			Northern Territory	Australia
1856-06-01			Northern Territory	Australia
1856-07-01			Northern Territory	Australia
1856-08-01			Northern Territory	Australia
1856-09-01			Northern Territory	Australia
1856-10-01			Northern Territory	Australia
1856-11-01			Northern Territory	Australia
1856-12-01			Northern Territory	Australia
1857-01-01			Northern Territory	Australia
1857-02-01			Northern Territory	Australia
1857-03-01			Northern Territory	Australia
1857-04-01			Northern Territory	Australia
1857-05-01			Northern Territory	Australia
1857-06-01			Northern Territory	Australia
1857-07-01			Northern Territory	Australia
1857-08-01			Northern Territory	Australia
1857-09-01			Northern Territory	Australia
1857-10-01			Northern Territory	Australia
1857-11-01			Northern Territory	Australia
1857-12-01	29.876	1.784	Northern Territory	Australia

Figure 13

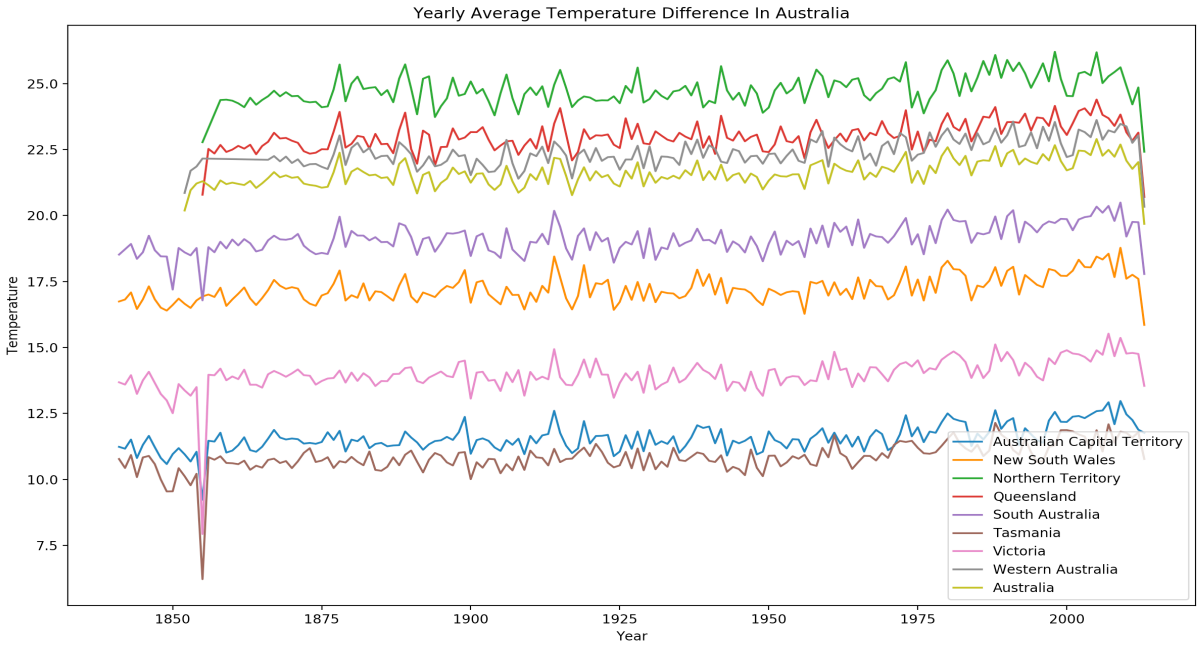

Figure 14

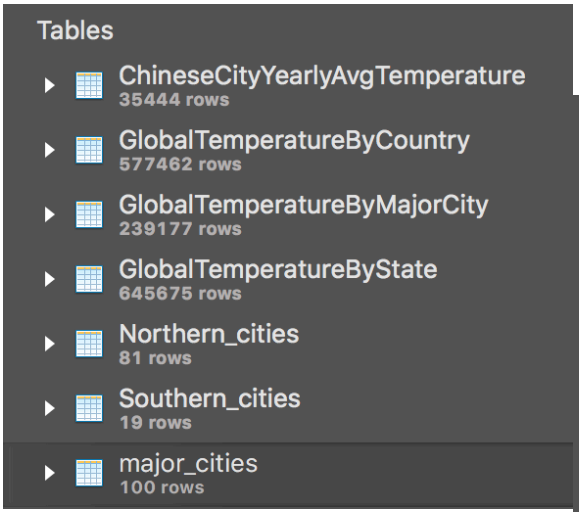

Figure 15

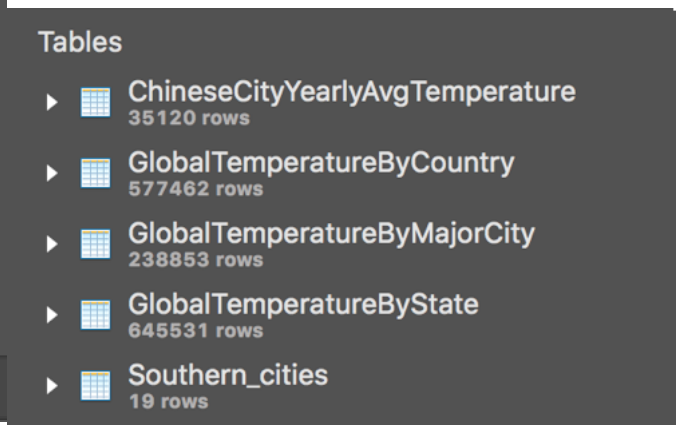

Figure 16

Discussion

There is no specific requirement for how to deal with missing data, which results in the inconsistency of the plot. However, we can easily find out here are some strange conditions when data is visualized. Figure 15 is the tables created before clean the missing value, Figure 16 is after, which has less rows. As for to ignore, delete, or supplement the missing value is not a requirement for this assignment.

Reference list

Marina, 2013 Dealing with missing data: Key assumptions and methods for applied analysis