基于全球温度变化的数据挖掘报告

计算2001孙铭俊

学号：2007010119

指导老师：张学辉

汇报提纲

一、数据描述与预处理

二、数据分析及可视化展示

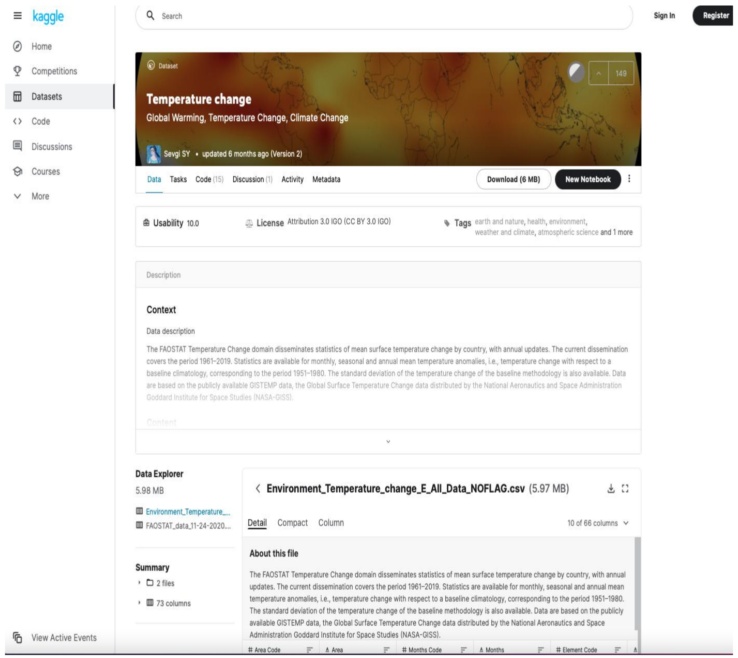
三、总结与反思

一、数据描述与预处理

数据来源

➢本项目数据集取自 https://www.kaggle.com/sevgisarac/temperature-change ，由世界粮农组织发布，且基于公开获得的GISTEMP，即美国国家航空航天局戈达德空间研究所（NASA-GISS）分发的全球表面温度变化数据，其内容涵盖了1961-2019年间各个国家的地表温度变化。

➢Environment_Temperature_change_E_All_Data_NOFLAG.csv 为温度变化数据表，数据量大，本报告将围绕其进行详细分析。



Temperature change

Global Warming, Temperature Change, Climate Change

Sevgi SY • updated 6 months ago (Version 2)

Usability 10.0

License Attribution 3.0 IGO (CC BY 3.0 IGO)

Tags earth and nature, health, environment, weather and climate, atmospheric science and 1 more

Description

Context

Data description

The FAOSTAT Temperature Change domain disseminates statistics of mean surface temperature change by country, with annual updates. The current dissemination covers the period 1961–2019. Statistics are available for monthly, seasonal and annual mean temperature anomalies, i.e., temperature change with respect to a baseline climatology, corresponding to the period 1951–1980. The standard deviation of the temperature change of the baseline methodology is also available. Data are based on the publicly available GISTEMP data, the Global Surface Temperature Change data distributed by the National Aeronautics and Space Administration Goddard Institute for Space Studies (NASA-GISS).

Content

Data Explorer

5.98 MB

Environment_Temperature_...

FAOSTAT_data_11-24-2020....

Summary

2 files

73 columns

Environment_Temperature_change_E_All_Data_NOFLAG.csv (5.97 MB)

Detail Compact Column 10 of 66 columns

About this file

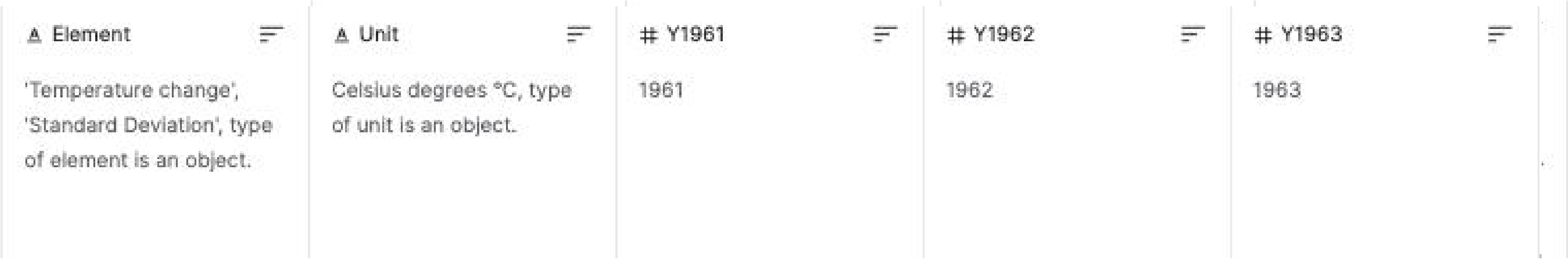
The FAOSTAT Temperature Change domain disseminates statistics of mean surface temperature change by country, with annual updates. The current dissemination covers the period 1961–2019. Statistics are available for monthly, seasonal and annual mean temperature anomalies, i.e., temperature change with respect to a baseline climatology, corresponding to the period 1951–1980. The standard deviation of the temperature change of the baseline methodology is also available. Data are based on the publicly available GISTEMP data, the Global Surface Temperature Change data distributed by the National Aeronautics and Space Administration Goddard Institute for Space Studies (NASA-GISS).

# Area Code	A Area	# Months Code	A Months	# Element Code

一、数据描述与预处理

主要字段

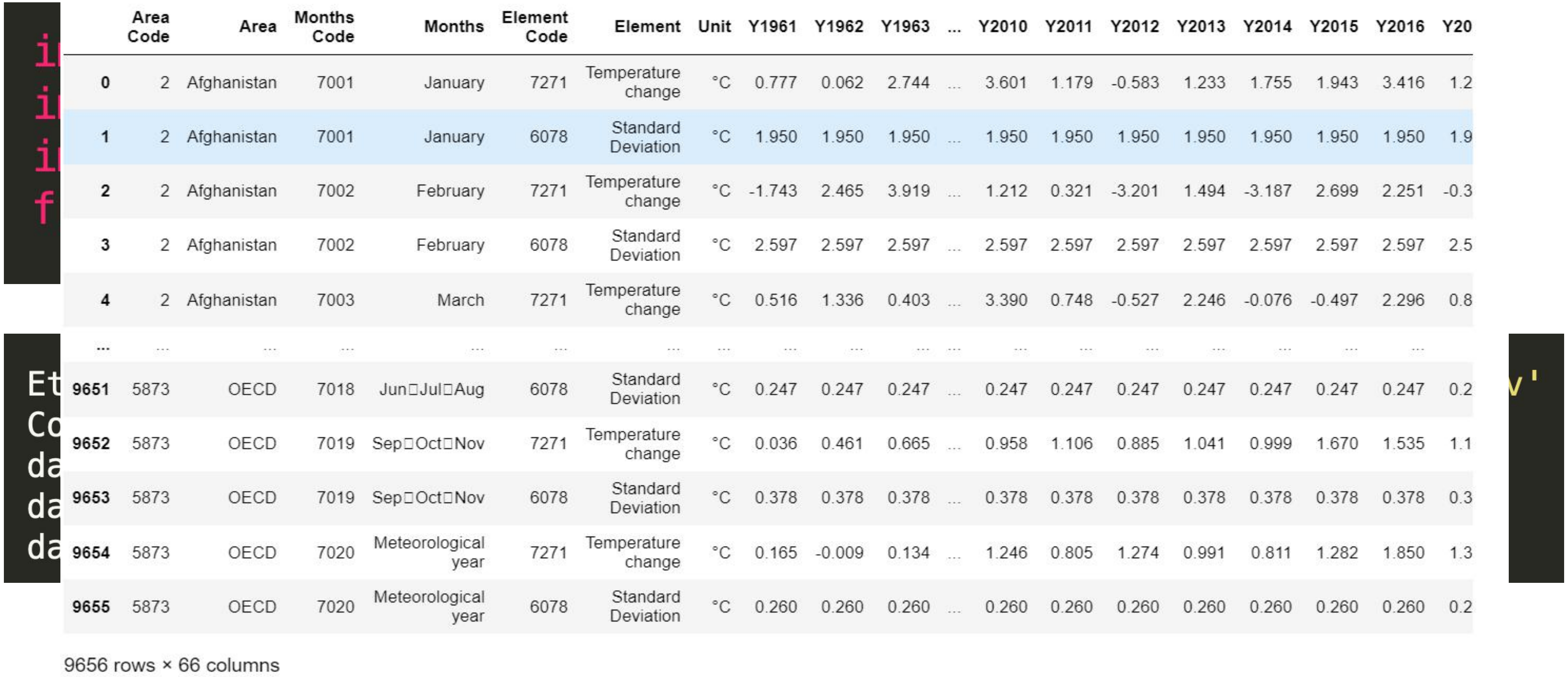
- 涵盖的国家/地区(Area字段):190个国家和37个其他领土实体。
- 月份(Months):1到12月，4个季度，气象年
- 测量单位(Unit):摄氏度℃
- 统计元素(Element):
- (1).温度变化(Temperature change)
- (2).标准差(Standard Deviation)



Element	Unit	# Y1961	# Y1962	# Y1963
'Temperature change', 'Standard Deviation', type of element is an object.	Celsius degrees °C, type of unit is an object.	1961	1962	1963

一、数据描述与预处理

读取数据



```
i
i
i
f

Et                                                    v'
Co
da
da
da
```

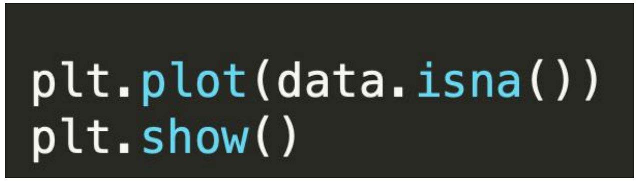
	Area Code	Area	Months Code	Months	Element Code	Element	Unit	Y1961	Y1962	Y1963	...	Y2010	Y2011	Y2012	Y2013	Y2014	Y2015	Y2016	Y20
0	2	Afghanistan	7001	January	7271	Temperature change	°C	0.777	0.062	2.744	...	3.601	1.179	-0.583	1.233	1.755	1.943	3.416	1.2
1	2	Afghanistan	7001	January	6078	Standard Deviation	°C	1.950	1.950	1.950	...	1.950	1.950	1.950	1.950	1.950	1.950	1.950	1.9
2	2	Afghanistan	7002	February	7271	Temperature change	°C	-1.743	2.465	3.919	...	1.212	0.321	-3.201	1.494	-3.187	2.699	2.251	-0.3
3	2	Afghanistan	7002	February	6078	Standard Deviation	°C	2.597	2.597	2.597	...	2.597	2.597	2.597	2.597	2.597	2.597	2.597	2.5
4	2	Afghanistan	7003	March	7271	Temperature change	°C	0.516	1.336	0.403	...	3.390	0.748	-0.527	2.246	-0.076	-0.497	2.296	0.8
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
9651	5873	OECD	7018	Jun□Jul□Aug	6078	Standard Deviation	°C	0.247	0.247	0.247	...	0.247	0.247	0.247	0.247	0.247	0.247	0.247	0.2
9652	5873	OECD	7019	Sep□Oct□Nov	7271	Temperature change	°C	0.036	0.461	0.665	...	0.958	1.106	0.885	1.041	0.999	1.670	1.535	1.1
9653	5873	OECD	7019	Sep□Oct□Nov	6078	Standard Deviation	°C	0.378	0.378	0.378	...	0.378	0.378	0.378	0.378	0.378	0.378	0.378	0.3
9654	5873	OECD	7020	Meteorological year	7271	Temperature change	°C	0.165	-0.009	0.134	...	1.246	0.805	1.274	0.991	0.811	1.282	1.850	1.3
9655	5873	OECD	7020	Meteorological year	6078	Standard Deviation	°C	0.260	0.260	0.260	...	0.260	0.260	0.260	0.260	0.260	0.260	0.260	0.2

9656 rows × 66 columns

一、数据描述与预处理

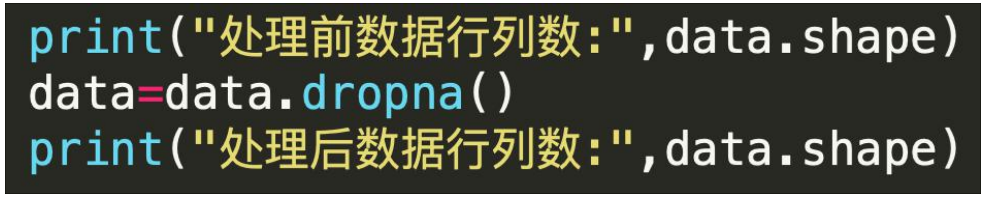
删除数据缺失值

- ➢ plot作图分析数据集

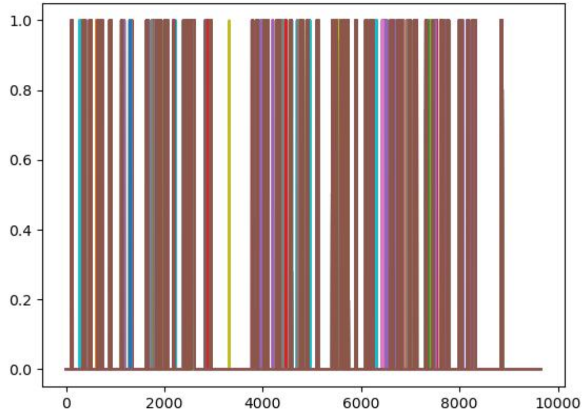


```
plt.plot(data.isna())
plt.show()
```

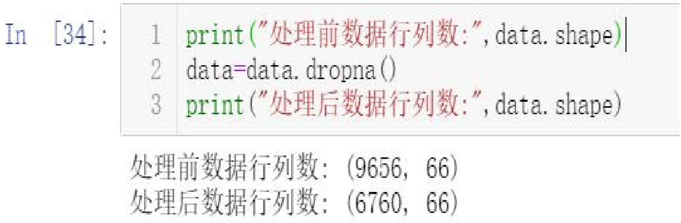
- ➢ 处理缺失数据



```
print("处理前数据行列数:",data.shape)
data=data.dropna()
print("处理后数据行列数:",data.shape)
```



plot画图结果



```
In  [34]:
print("处理前数据行列数:",data.shape)
data=data.dropna()
print("处理后数据行列数:",data.shape)

处理前数据行列数: (9656, 66)
处理后数据行列数: (6760, 66)
```

数据处理结果

汇报提纲

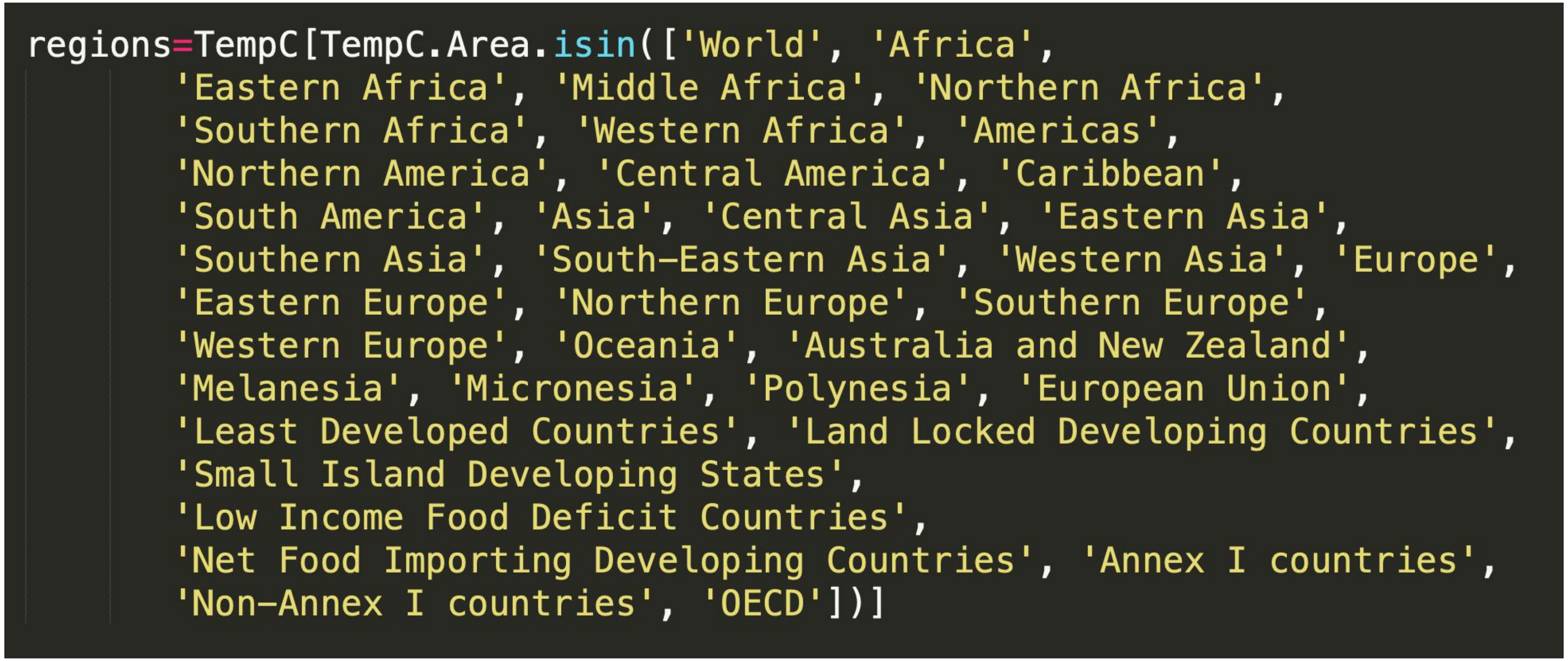
一、数据描述与预处理

二、数据分析及可视化展示

三、总结与反思

二、数据分析及可视化展示

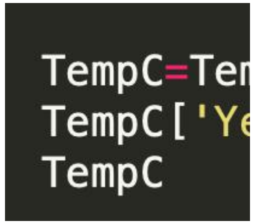
数据分析——总体分析世界温度变化



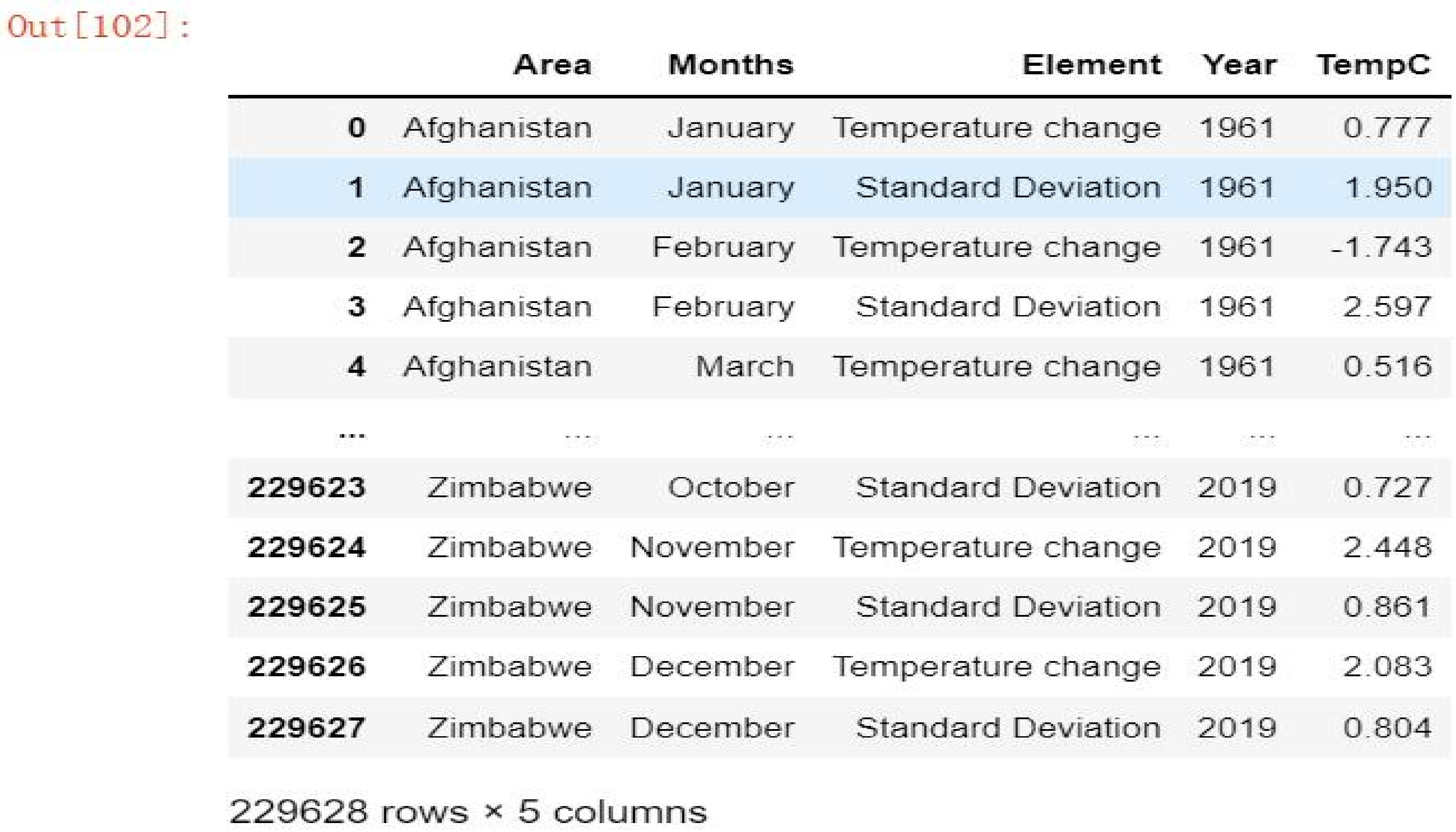
```
regions=TempC[TempC.Area.isin(['World', 'Africa',
       'Eastern Africa', 'Middle Africa', 'Northern Africa',
       'Southern Africa', 'Western Africa', 'Americas',
       'Northern America', 'Central America', 'Caribbean',
       'South America', 'Asia', 'Central Asia', 'Eastern Asia',
       'Southern Asia', 'South-Eastern Asia', 'Western Asia', 'Europe',
       'Eastern Europe', 'Northern Europe', 'Southern Europe',
       'Western Europe', 'Oceania', 'Australia and New Zealand',
       'Melanesia', 'Micronesia', 'Polynesia', 'European Union',
       'Least Developed Countries', 'Land Locked Developing Countries',
       'Small Island Developing States',
       'Low Income Food Deficit Countries',
       'Net Food Importing Developing Countries', 'Annex I countries',
       'Non-Annex I countries', 'OECD'])]
```

二、数据分析及可视化展示

数据分析——总体分析世界温度变化



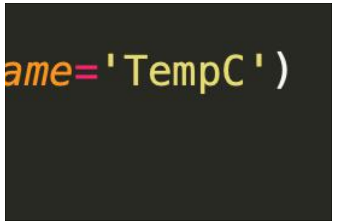
```
TempC=Te
TempC['Y
TempC
```



Out[102]:

	Area	Months	Element	Year	TempC
0	Afghanistan	January	Temperature change	1961	0.777
1	Afghanistan	January	Standard Deviation	1961	1.950
2	Afghanistan	February	Temperature change	1961	-1.743
3	Afghanistan	February	Standard Deviation	1961	2.597
4	Afghanistan	March	Temperature change	1961	0.516
...	...	...	...	...	...
229623	Zimbabwe	October	Standard Deviation	2019	0.727
229624	Zimbabwe	November	Temperature change	2019	2.448
229625	Zimbabwe	November	Standard Deviation	2019	0.861
229626	Zimbabwe	December	Temperature change	2019	2.083
229627	Zimbabwe	December	Standard Deviation	2019	0.804

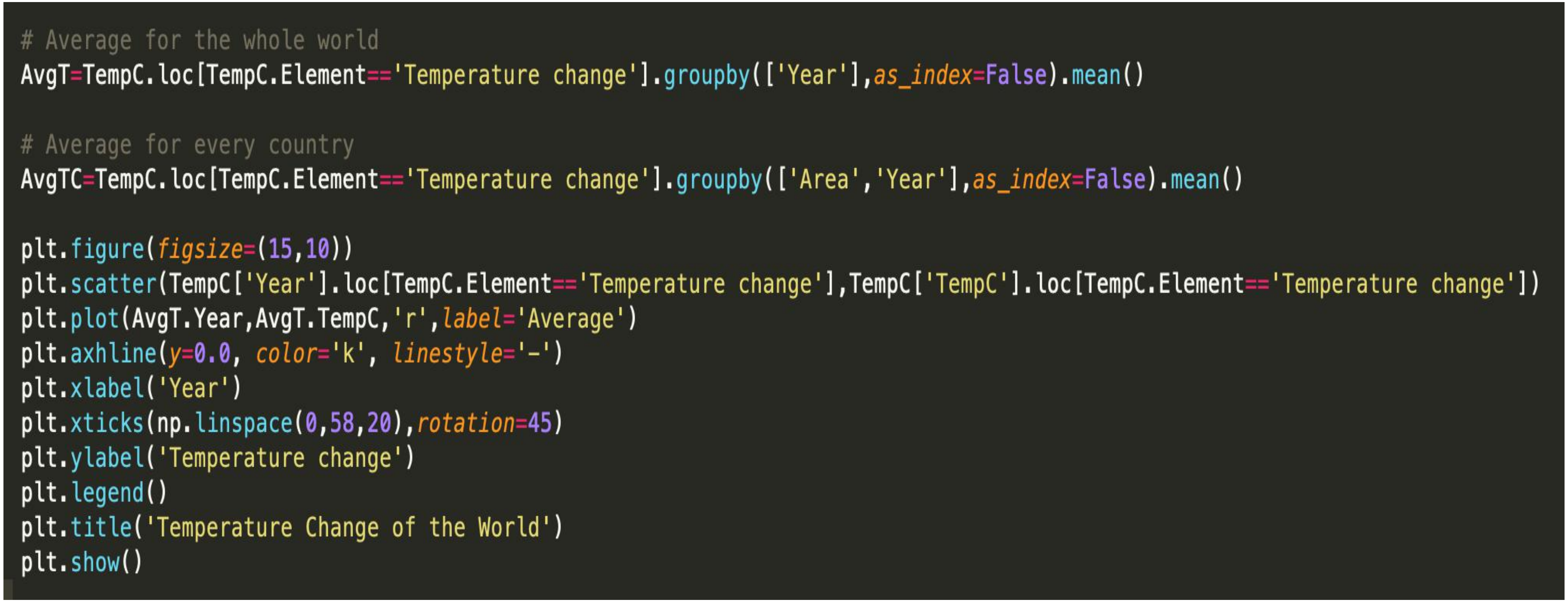
229628 rows × 5 columns



```
ame='TempC')
```

二、数据分析及可视化展示

绘制图像——世界温度散点图



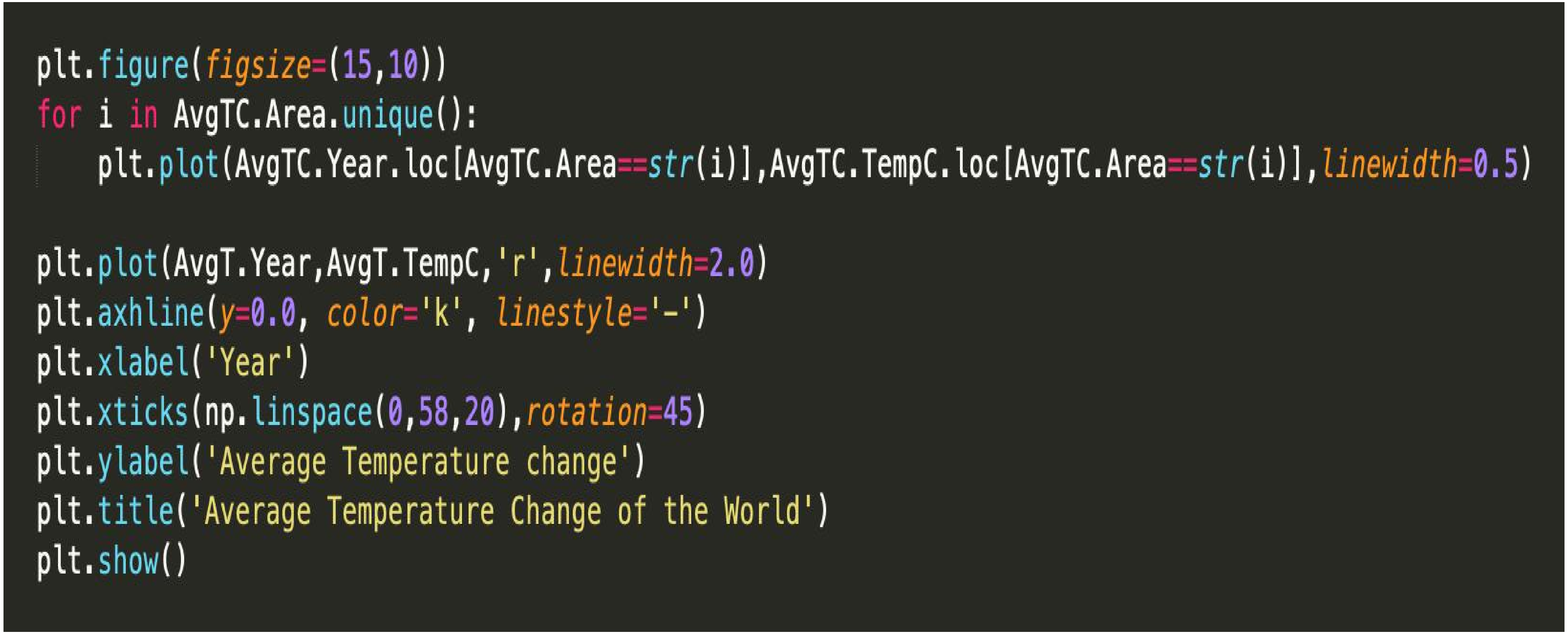
```
# Average for the whole world
AvgT=TempC.loc[TempC.Element=='Temperature change'].groupby(['Year'],as_index=False).mean()

# Average for every country
AvgTC=TempC.loc[TempC.Element=='Temperature change'].groupby(['Area','Year'],as_index=False).mean()

plt.figure(figsize=(15,10))
plt.scatter(TempC['Year'].loc[TempC.Element=='Temperature change'],TempC['TempC'].loc[TempC.Element=='Temperature change'])
plt.plot(AvgT.Year,AvgT.TempC,'r',label='Average')
plt.axhline(y=0.0, color='k', linestyle='-')
plt.xlabel('Year')
plt.xticks(np.linspace(0,58,20),rotation=45)
plt.ylabel('Temperature change')
plt.legend()
plt.title('Temperature Change of the World')
plt.show()
```

二、数据分析及可视化展示

绘制图像——每个国家年份均温折线图

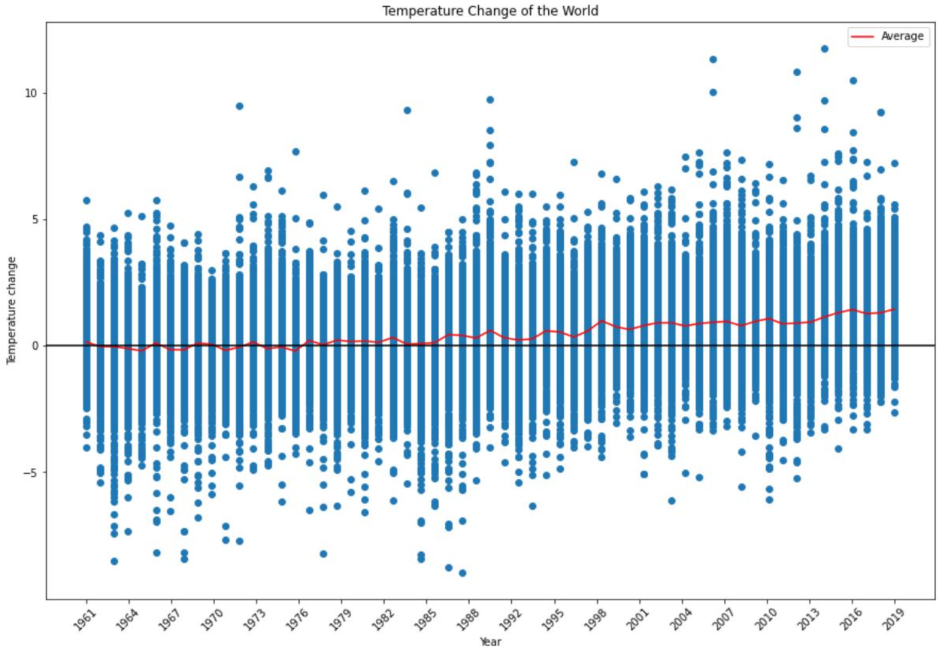


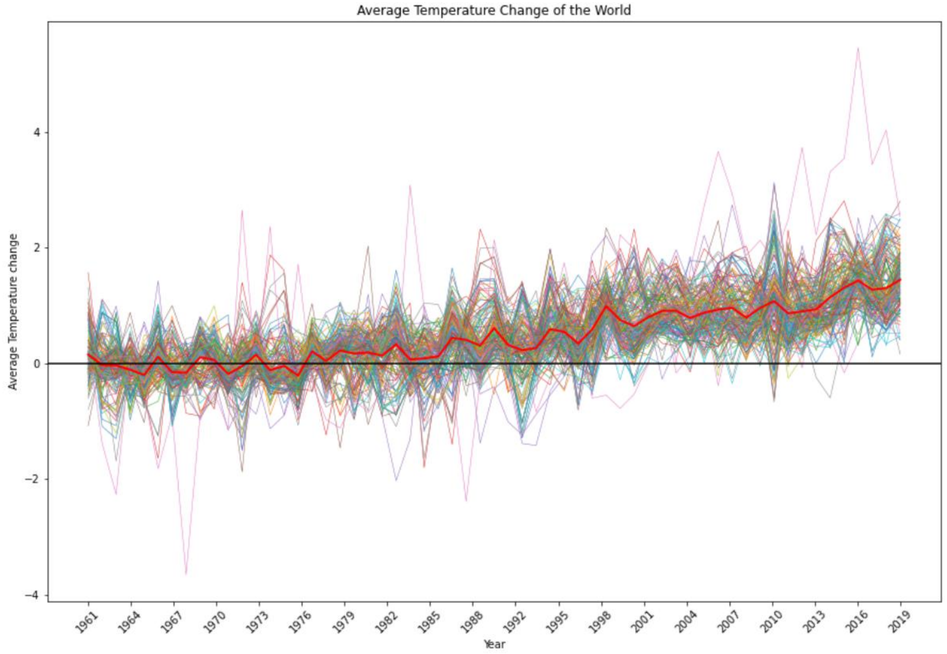
```
plt.figure(figsize=(15,10))
for i in AvgTC.Area.unique():
    plt.plot(AvgTC.Year.loc[AvgTC.Area==str(i)],AvgTC.TempC.loc[AvgTC.Area==str(i)],linewidth=0.5)

plt.plot(AvgT.Year,AvgT.TempC,'r',linewidth=2.0)
plt.axhline(y=0.0, color='k', linestyle='-')
plt.xlabel('Year')
plt.xticks(np.linspace(0,58,20),rotation=45)
plt.ylabel('Average Temperature change')
plt.title('Average Temperature Change of the World')
plt.show()
```

二、数据分析及可视化展示

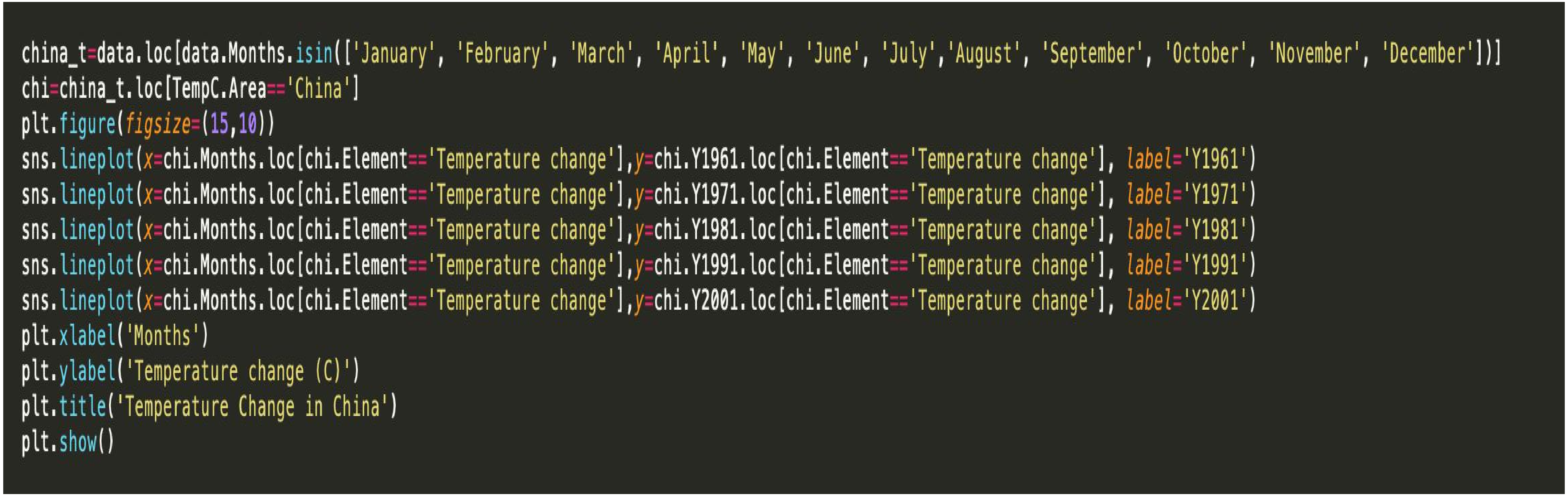
分析图像





二、数据分析及可视化展示

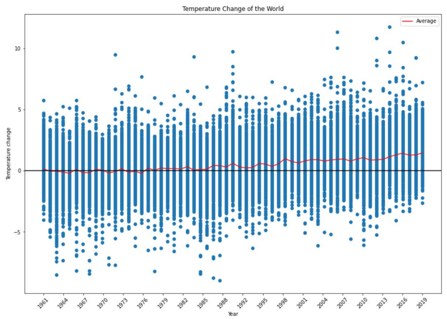
绘制图像——中国温度变化趋势折线图及中国温度变化标准差趋势折线图

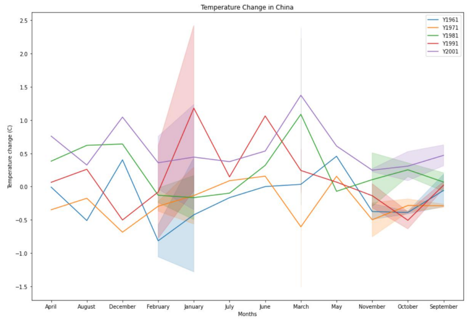


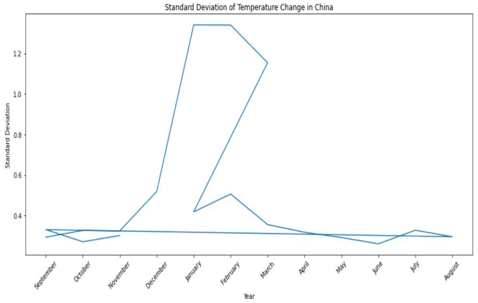
```
china_t=data.loc[data.Months.isin(['January', 'February', 'March', 'April', 'May', 'June', 'July','August', 'September', 'October', 'November', 'December'])]
chi=china_t.loc[TempC.Area=='China']
plt.figure(figsize=(15,10))
sns.lineplot(x=chi.Months.loc[chi.Element=='Temperature change'],y=chi.Y1961.loc[chi.Element=='Temperature change'], label='Y1961')
sns.lineplot(x=chi.Months.loc[chi.Element=='Temperature change'],y=chi.Y1971.loc[chi.Element=='Temperature change'], label='Y1971')
sns.lineplot(x=chi.Months.loc[chi.Element=='Temperature change'],y=chi.Y1981.loc[chi.Element=='Temperature change'], label='Y1981')
sns.lineplot(x=chi.Months.loc[chi.Element=='Temperature change'],y=chi.Y1991.loc[chi.Element=='Temperature change'], label='Y1991')
sns.lineplot(x=chi.Months.loc[chi.Element=='Temperature change'],y=chi.Y2001.loc[chi.Element=='Temperature change'], label='Y2001')
plt.xlabel('Months')
plt.ylabel('Temperature change (C)')
plt.title('Temperature Change in China')
plt.show()
```

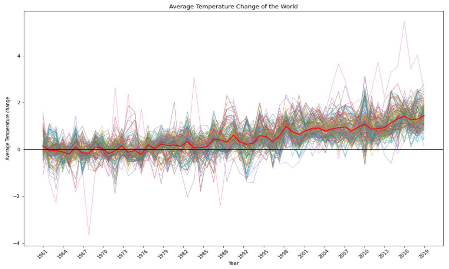
二、数据分析及可视化展示

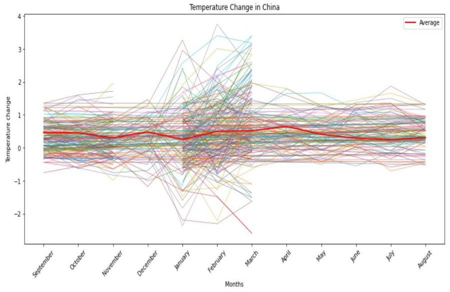
得出结论











本次分析可以获得中国在1961-2019年间冬季温度增长速度变快；本次数据记录的是地表温度，因此可以猜测是城市化的加快或能源结构的调整导致了冬季温度变化的差异。

汇报提纲

一、数据描述与预处理

二、数据分析及可视化展示

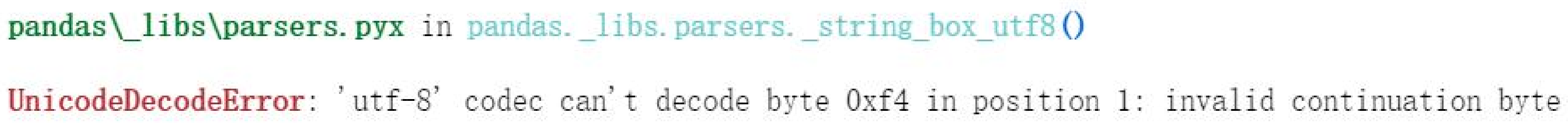
三、总结与反思

三、总结与反思

问题及其解决方案

1.在读入csv文件中编码出现问题

在运行data = pd.read_csv(Etemp_path)时出现报错



```
pandas\_libs\parsers.pyx in pandas._libs.parsers._string_box_utf8()

UnicodeDecodeError: 'utf-8' codec can't decode byte 0xf4 in position 1: invalid continuation byte
```

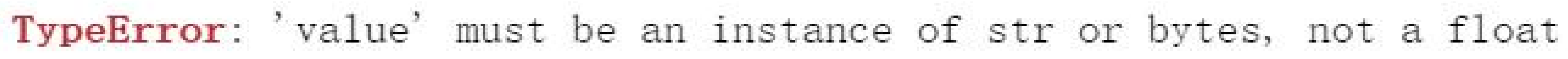
在查阅相关资料后才发现，Unit单元地摄氏度是Latin字符，

故改为data = pd.read_csv(Etemp_path，encoding='latin-1')

三、总结与反思

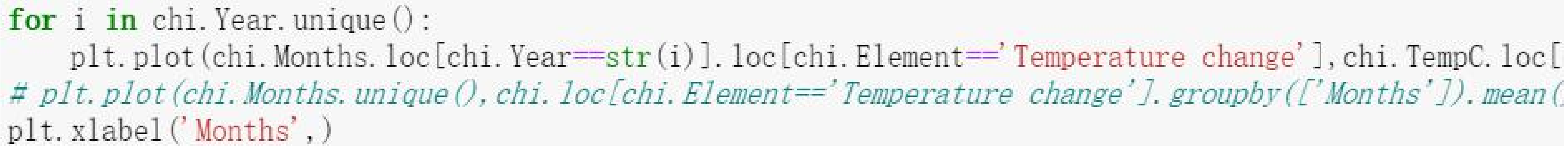
问题及其解决方案

2.在可视化过程中类型出现问题



```
TypeError: 'value' must be an instance of str or bytes, not a float
```

发现是months,years存在重复值，重复分类在groupby()中会报错，进而引起value的值出现问题，因此在year,months后添加unique()函数即可解决问题。



```
for i in chi.Year.unique():
    plt.plot(chi.Months.loc[chi.Year==str(i)].loc[chi.Element=='Temperature change'],chi.TempC.loc[
# plt.plot(chi.Months.unique(),chi.loc[chi.Element=='Temperature change'].groupby(['Months']).mean(
plt.xlabel('Months',)
```

三、总结与反思

反思

本次项目的实践存在几点不足：

1、数据集具体特征的选取，在没有对数据进行清洗以及后续对部分数据的丢弃之前，我们运行出来的结果往往会有问题抑或是运行速度很慢，在对数据进行处理之后，效率才得到了提升；

2、在进行数据分析时，问题最多的就是类型的报错以及索引的使用不当，所有应当更加熟悉pandas库，要对dataframe,series有一个充分的了解；

因此我认为在项目实践中对应模块的掌握程度很大程度上决定了对数据分析的处理效率，对于数据进行科学的分析则能够更好地为结果进行服务。

中國石油大學（华东）
CHINA UNIVERSITY OF PETROLEUM

谢谢大家！