**Report 4**

**Part 1: Course exercises**

1.1:NAN

1.2:堆叠合并

1.3:哑变量

1.4:异常值

1.5:pivot（）函数

2.1:TRUE

2.2:TRUE

2.3:FALSE

2.4:FALSE

2.5:FALSE

3.1:B

3.2:A

3.3:C

3.4:D

3.5:A

4.1: 异常值是指数据集中的个别值明显偏离它所属数据集的其余值，这些数值是不合理的或错误的。

4.2:堆叠合并、主键合并、根据索引合并、合并重叠数据，总共四种。

5.1:

*import* pandas *as* pd

*import* numpy *as* np

group\_a = pd.DataFrame({'A': [2,3,5,2,3],

                             'B': ['5',np.nan,'2','3','6'],

                             'C': [8,7,50,8,2],

                             'key': [3,4,5,2,5]})

group\_b = pd.DataFrame({'A': [3,3,3],

                        'B': [4,4,4],

                        'C': [5,5,5]})

print(group\_a)

print(group\_b)

5.2:

group\_a  = group\_a.combine\_first(group\_b)

group\_a

5.3:

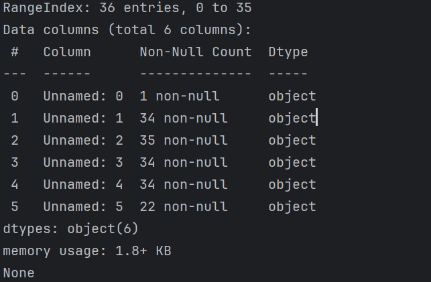
group\_a.rename(*columns*={'key':'D'})

**Part 2:**

**Download the files from the Moodle and complete the following Pandas practices and screenshot your solutions and results.**

1. 读取2004年的数据，并展示查看数据的格式.

import numpy as np  
import pandas as pd  
df = pd.read\_csv('./by\_year/2014.csv')  
print(df.info())

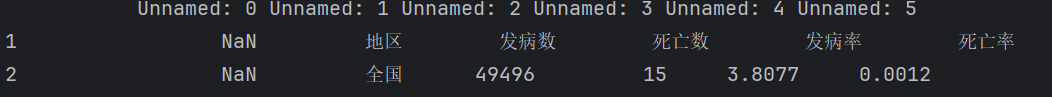


2. 对第一年的数据进行预处理：删除第一行，选取真实列名所在的第一行，更改列名，删除多余的行.

df = df.drop(0)

# 更改前的结果 ：   


# 更改后的结果：



df.columns = df.iloc[0]

#更改 df 列名称



df = df.drop(df.columns[0],axis=1) # 删除多余的列



3. 第二步删除行后，index被打乱：重新设置被打乱的index，添加年份变量.

df = df.reset\_index(drop=True)

# 重新设置索引

# 更改列名称 添加年份变量

df['年份'] = 2023



4. 批量读取连接数据：自定义一个函数对后面所有年份进行批量预处理（删除第0列，删除前朗行后后一行，重塑年份变量）.

def modify\_file(path,year) :  
 df = pd.read\_csv(path)  
 df = df.drop(0) # 删除 列名  
 df.columns = df.iloc[0]  
 df = df.drop(1) # 删除第一行  
  
 df = df.drop(df.columns[0], axis=1) # 删除多余的列  
 df = df.reset\_index(drop=True)  
 df['年份'] = year

5. 定义函数用于批量读取及凭借数据，调用第四步中定义的函数进行批量处理：建立空列表用于存放数据，通过循环遍历读取文件，读取数据，一次累加年份，进行预处理及其重塑变量等.

list\_path = []# 用于存放文件的地址

year = 2004  
for i in list\_path :  
 modify\_file(i,year)  
 year += 1

6. 重塑列名，链接数据，并填充缺失值.

other\_data.columns = col\_name.append(pd.Series("年份")) # 重塑数据的列名

flu\_data = pd.concat([dat0, other\_data]) # 连接数据，命名为flu\_data

flu\_data.fillna(0, inplace = True) # 使用0填充缺失值

7. 检查数据：对数据进行计数，替换文字中的空格，删除多余类型数据，替换有歧义数据，再次检查.

flu\_data["地区"].value\_counts() # 对数据进行计数

flu\_data["地区"] = flu\_data["地区"].apply(lambda x: x.replace(" ", "")) # 替换文字中的空格

flu\_data = flu\_data.loc[flu\_data["地区"] != "建设兵团"] # 删除地区为建设

flu\_data.loc[flu\_data['地区'] =='黑龙江','地区']='黑龙江省' # 将黑龙江替换为黑龙江省

flu\_data["地区"].value\_counts()

8. 人口数据的清洗与重塑：修改列名，删除多余的行，统一格式.

people = pd.read\_csv("people\_2.csv",encoding="gbk") # 读取人口数据，命名为people

people.columns = people.iloc[2] # 用第2行作为列名

people.drop([0,1,2,len(people)-1,len(people)-2], axis = 0, inplace = True) # 删除多余的行

people.reset\_index(inplace=True, drop=True) # 删除多余的行

## 统一地区名的格式

people.loc[people['地区'] == '内蒙古自治区','地区']='内蒙古'

people.loc[people['地区'] == '广西壮族自治区','地区']='广西'

people.loc[people['地区'] == '西藏自治区','地区']='西藏'

people.loc[people['地区'] == '宁夏回族自治区','地区']='宁夏'

people.loc[people['地区'] == '新疆维吾尔自治区','地区']='新疆'

peo\_name = list(people.columns) # 获取people的变量名

peo\_name.remove("地区") # 去除地区变量，得到年份数据

change\_people = pd.melt(people, id\_vars=["地区"], value\_vars=peo\_name, \

var\_name="年份", value\_name="总人口数") # 通过melt重塑数据

change\_people["年份"] = change\_people["年份"].apply(lambda x: re.findall("\d+", x)[0]) # 去除年份的“年”字

change\_people["年份"] = change\_people["年份"].astype(np.int)

9. 拼接数据：填充数据，转换数据类型，最终查看数据并保存输出.

result = pd.merge(flu\_data, change\_people, on=['年份', '地区'])

change\_list = ['发病率','死亡率','总人口数','发病数','死亡数']

result[change\_list] = result[change\_list].apply(pd.to\_numeric)

result.head()

*Ref：Pandas数据预处理与数据重塑案例（*[*https://zhuanlan.zhihu.com/p/44677396*](https://zhuanlan.zhihu.com/p/44677396)*）*

问题：所给出的数据文件CSV 2004.csv、2005.csv、2006.cs、2008.csv、2016.csv文件的编码方式是GB2312 但是其中2009.csv、2010.csv、2011.csv、2013.csv、2014.csv文件的编码方式未知。查找到了是GBK文件