****

**2024—2025学年第2学期**

**《数据分析(Python)》课程报告**

|  |  |
| --- | --- |
| **专业班级** |  |
| **姓 名** |  |
| **学 号** |  |
| **序 号** |  |
| **评阅教师** | **张学辉** |

**2025 年 5 月 28 日**

**基于印度人口数据挖掘报告**

**一、数据描述**

1数据来源:

数据是关于2020年世界各国家人口的数据，是关于1950年至2021年印度人口数量和人口增长率的数据。

2数据获取：

1950年至2021年印度人口数量和人口增长率等并将获这些信息暂存于Excel文件中。

3数据种类：

年份Year

印度人口数量Population

人口增长率GrowthRate=（年末人口数-年初人口数）/年平均人口×1000‰

人口增长率是反映人口发展速度和制定人口计划的重要指标，也是计划生育统计中的一个重要指标，它表明人口自然增长的程度和趋势。人口的力量正在把世界分成两部分：人口增长缓慢的地区，生活条件正在得到改善；人口增长迅速的地区，生活条件正在恶化。人口的迅速增长抵销了农业和经济方面的部分差距。

4使用工具:

Python是一种广泛使用的解释型、高级编程、通用型编程语言，由吉多·范罗苏姆创造，第一版发布于1991年。可以视之为一种改良（加入一些其他编程语言的优点，如面向对象）的LISP。

Python的设计哲学强调代码的可读性和简洁的语法（尤其是使用空格缩进划分代码块，而非使用大括号或者关键词）。相比于C++或Java，Python让开发者能够用更少的代码表达想法。不管是小型还是大型程序，该语言都试图让程序的结构清晰明了。

5数据主要内容

通过这两个方面的数据分析，我们可以分析出近几十年来印度人口的数量变化趋势，来进一步分析印度经济发展和人口的关系。

此份数据主要是针对印度从1950年到2021年每年的人口数量以及人口增长率及相应变化，通过人口数量以及人口增长率可以分析印度人口的具体情况，明确印度人口数量的多少；通过人口数量和人口增长率可以分析影响各印度人口变化的主要因素，并且通过各因素的情况进一步研究深入推理出印度相应政策以及经济状况的影响。

**二、数据预处理**

数据中具体的人口数量和增长率，并没有通过已有数据对未来的人口数量进行预测，不便于解决如印度人口预计在哪一年达到20亿人口等问题。我们可以通过线性回归分析，线性回归是利用数理统计中回归分析，来确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法，运用十分广泛。其表达形式为y = w'x+e，e为误差服从均值为0的正态分布。从而得出从2021年到2069年的预测人口数量，便于绘图进行分析印度过去和未来的人口变化趋势。

1.首先导入代码所需库

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import warnings

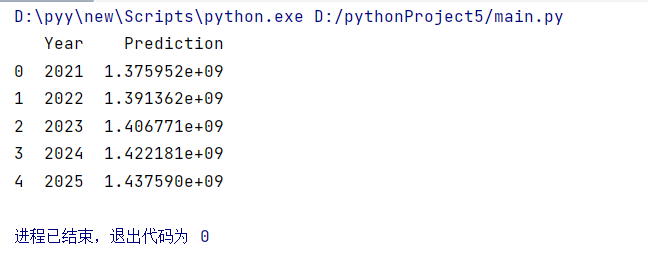
warnings.filterwarnings('ignore')

2.数据读入

data = pd.read\_csv('IndiaPopulation\_2021.csv')

print(data.head())

效果展示：



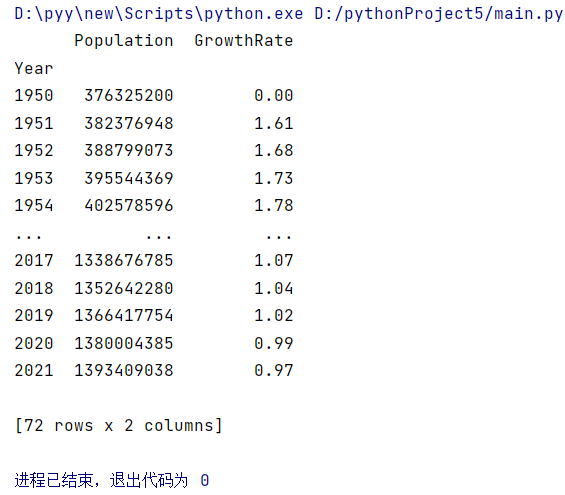
2.数据处理（**是否有缺失值，异常值和重复值**）

将数据按照年份排序

data = data.sort\_values('Year').set\_index('Year')

print(data)

效果展示：



**三、数据分析**

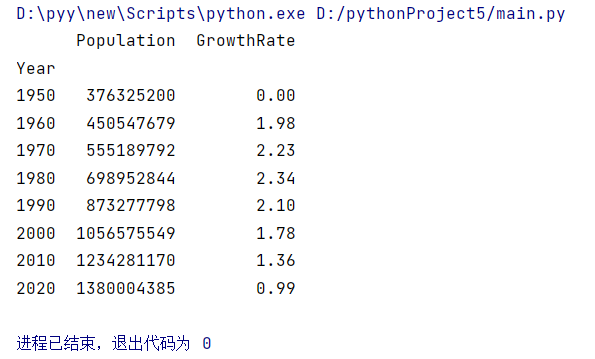
1缩小区间

由于数据复杂，我们需要进行区间的缩小增大数据的可视化。

year\_data = data.iloc[0:len(data):10]

print(year\_data)

效果展示：

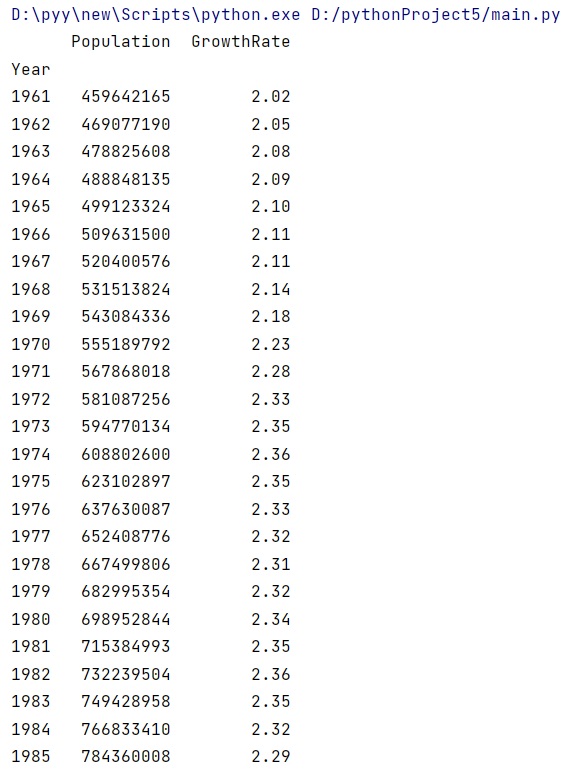


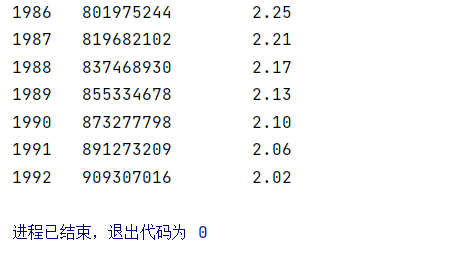
2筛选出增长率大于2的年份

gr\_year = data[data['GrowthRate']>2]

Print(gr\_year)

效果展示：





**由此我们得出，大量年份人口的增长率大于2，是印度人口增长迅速的直接原因之一。（能否查找其他资料，找出人口增长迅速的主要原因？？？？）**

3人口预测

from sklearn.linear\_model import LinearRegression

lr = LinearRegression()

X = data[['Year']]

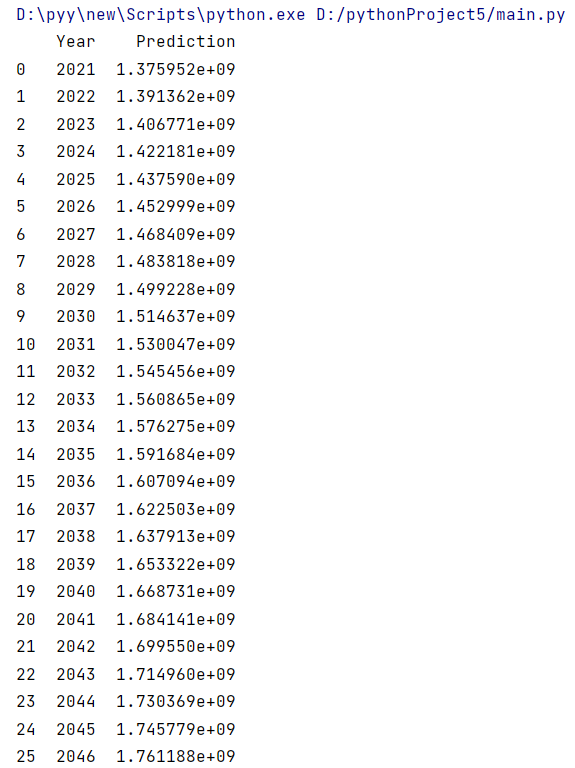
Y = data[['Population']]

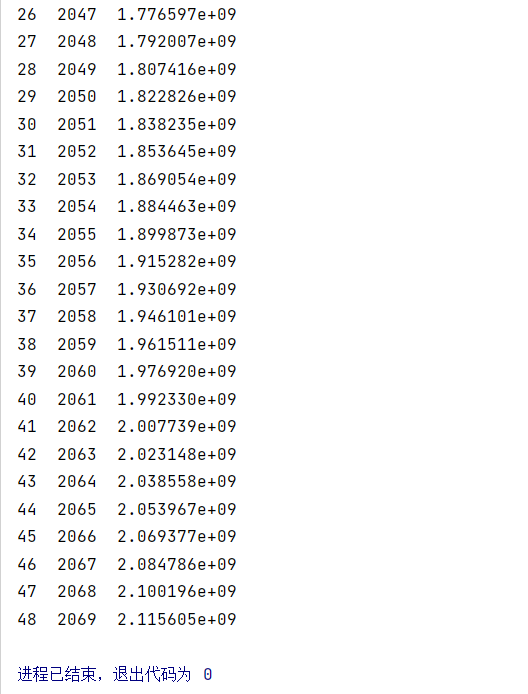
model = lr.fit(X,Y)

Pred\_X = pd.DataFrame(list(range(2021,2070)),columns=['Year'])

Pred\_X['Prediction'] = model.predict(Pred\_X[['Year']])

效果展示：





**由线性回归分析我们得到印度人口将在2062年达到20亿人。**

**四、数据可视化展示**

**能否说明每个图表的意义是什么？**

1人口数量柱状图绘制

plt.figure(figsize=(15,5))

plt.title('Barplot for Population by Year', fontsize=15)

sns.barplot(data.index, data.Population)

plt.show()

plt.figure(figsize=(10,5))

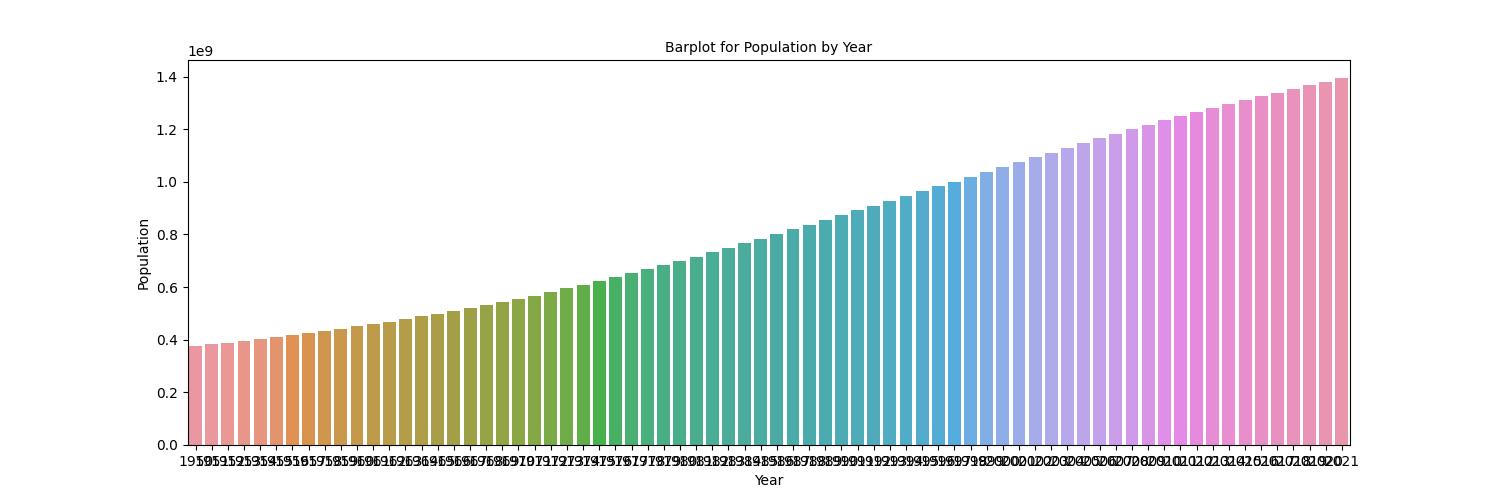
plt.title('Scatterplot for Population')

sns.scatterplot(data.index, data.Population,

s=100, alpha=0.5, color='red')

plt.show()

效果展示：



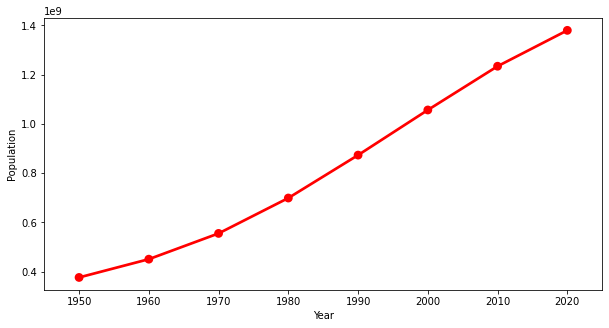
2缩小区间折线图

plt.figure(figsize=(10,5))

sns.pointplot(year\_data.index, year\_data.Population, color='red')

plt.show()

效果展示：



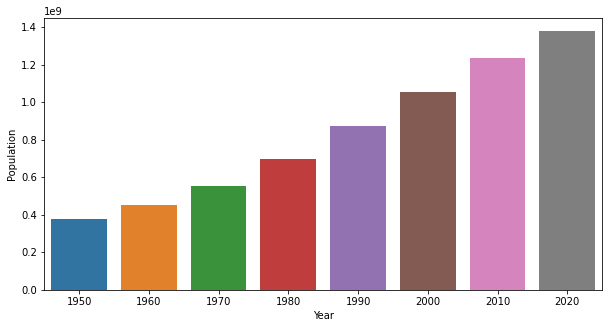
3缩小区间柱状图

plt.figure(figsize=(10,5))

sns.barplot(year\_data.index, year\_data.Population)

plt.show()

效果展示：



4人口数量点状图绘制

plt.figure(figsize=(10,5))

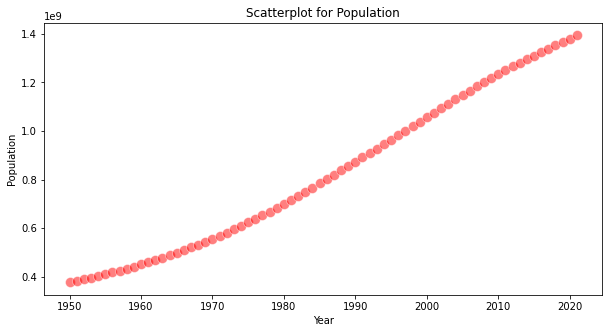
plt.title('Scatterplot for Population')

sns.scatterplot(data.index, data.Population,

s=100, alpha=0.5, color='red')

plt.show()

效果展示：



**由此我们可以很清楚的看出印度人口数量几乎以不变的速率增长，并且增长速度极快，我们需要结合人口的增长率进一步进行分析。（能否查阅相关资料，进一步分析？？？？）**

5增长率点状图

plt.figure(figsize=(10,5))

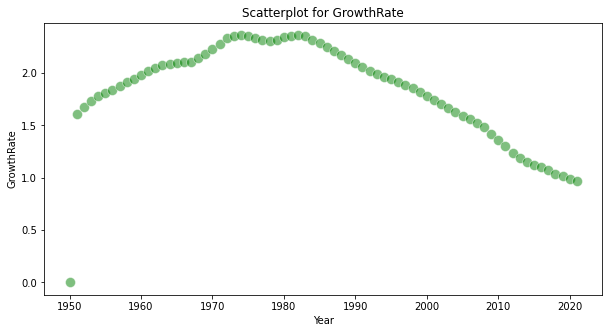
plt.title('Scatterplot for GrowthRate')

sns.scatterplot(data.index, data.GrowthRate,

s=100, alpha=0.5, color='green')

plt.show()

效果展示：



**由此我们可以看出增长率呈现一个先增长后减少的趋势，并且在增加的过程中有一段时间增长率几乎不变，可能与当时印度采取了一定的措施有关。（那么到底采取了什么措施？？）**

6增长率大于2点状图

plt.figure(figsize=(10,5))

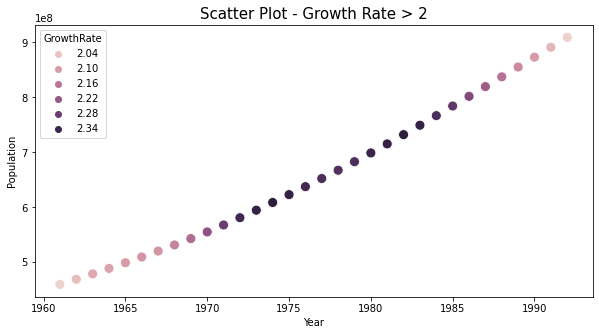
plt.title('Scatter Plot - Growth Rate > 2', fontsize=15)

sns.scatterplot(gr\_year.index, gr\_year.Population,

hue=gr\_year.GrowthRate,s=100)

plt.show()

效果展示：



7增长率大于2柱状图

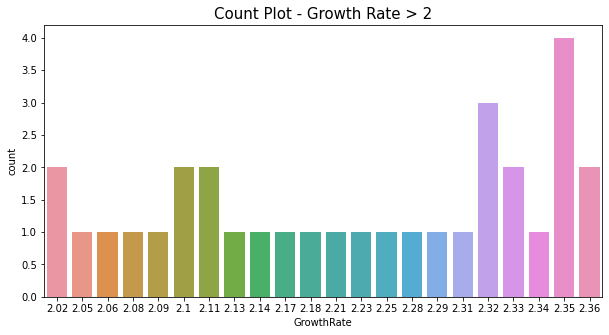
plt.figure(figsize=(10,5))

plt.title('Count Plot - Growth Rate > 2', fontsize=15)

sns.countplot(gr\_year.GrowthRate)

plt.show()

效果展示：

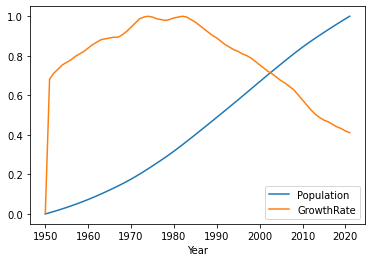


8曲线图

norm\_data.plot()

plt.show()

效果展示：



由图我们将两条曲线结合分析可以得出人口数量的曲线虽然一直上升，但是斜率却发生了变化，总体的人口数量曲线由下凸变为上凸，与人口增长率的曲线对应。

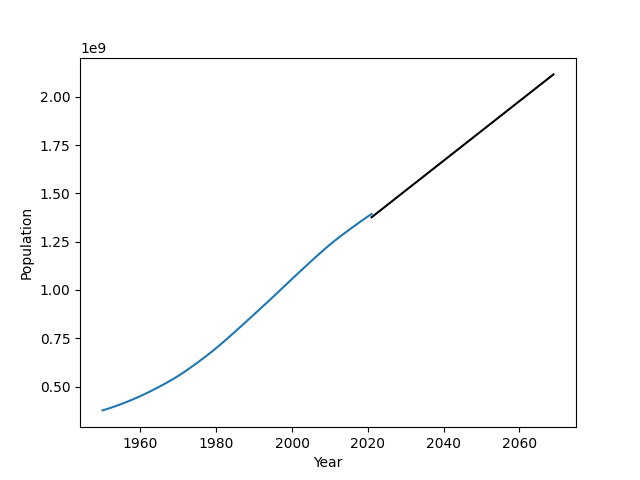
9线性回归分析

将利用线性回归分析得出的印度2021年到2069年的人口数量绘制在人口数量折线图中，如下：

效果展示：

sns.lineplot(x=data['Year'],y=data['Population'])

sns.lineplot(x=Pred\_X['Year'],y=Pred\_X['Prediction'],color='black')



通过对数据的可视化分析，得出以下三点结论：

1. 印度人口增长较快。虽然印度政府采取措施控制人口增长，但因为人口基数大，每年新增人口1500多万。
2. 根据线性回归分析，印度将在2062年达到20亿人
3. 印度人口众多，人口增长快，造成人均资源、人均粮食不足，延缓国民经济的发展。

**通过查阅资料，得出近50年来印度人口增长率下降的原因是：50年来除了1977-1979年计划生育工作一度陷入停顿以外，计划生育工作一直在不断加强。中央政府也很重视从经济上支持计划生育，并不断增加这一事业的财政拨款。印度政府对医疗卫生和计划生育的投资额从“三五”计划时期的25.08亿卢比增加到“六五”计划时的341.33亿卢比，再到“七五”时的680.94亿卢比，及“八五”时的1596.56亿卢比。大量的预算投入使得印度的人口数量得到了有效的控制。**

**启示人口大国要采取积极正确的政策来解决人口问题，中国实行计划生育效果很好，人口得到了非常有效的控制。今年的5月31号政府实行了三孩政策，为了防止人口老龄化。人口问题是每个国家都必须重视的问题，只有将人口问题解决好，才能让每一个人都拥有幸福的生活。**

**五、遇到的问题与解决方法**

列举3-5个典型问题及解决方法

1、问题：采用相对路径读入csv文件时，文件没有放入项目文件夹中，导致报错。

解决：**列出解决问题的具体方法。**

2.问题：保存新的csv文件时，准备修改资料文件夹中的同名csv，但是始终没有效果。

解决：打开项目文件夹发现程序自动创建了一个csv文件。

**六、学习总结与反思**

本次的python数据挖掘作业，通过将数据进行可视化展示，通过这次的大作业我收获颇多，主要分为如下两点：

一、将数据表格进行可视化展示：我学习到了如何用python读取数据表格，如何再利用读取到的数据绘制可视化的图形。在绘制图形时，我选取了matplotlib库，通过网上学习，我掌握了matplotlib库绘制柱状图和折线图的方法，极大的锻炼了我的代码能力和自学能力。

二、加强了数据分析能力：拿到数据后，首先大致对印度人口的数量有一个大致的了解，之后选择要研究的问题：印度近几十年来的人口数量变化趋势，印度人口将在哪一年达到20亿人。印度人口数量不仅仅是一些数字，也从侧面反映出印度经济的发展和政策的实施，人口数量。

在这次的大作业中我也有许多的思考，在制作PPT时，一定要选用简洁大方的背景图片，不能与文本文字的颜色相冲突，并且要对数据的变化进行必要的资料查阅。此外，我在代码能力上还有所欠缺，对python的操作不是很熟悉，需要多加练习。我在今后一定会加强练习，不断提高自己的编程能力，学好计算机语言。

经过半学期的学期，加上这一次的数据分析大作业，我已经基本能够用python解决一些简单的、小规模的问题了。经过一次大作业，我的的确确收获很多。在这之前虽然写过一些实用代码，但是都是碎片化的、随写随用的一点小脚本，十几行，难度都不大，这次大作业可以说是我第一次解决一个有一定规模的实际问题，让我对编程的学习有了更深的了解……谢谢老师提供了这次实践的机会，也希望在以后的学习生活中，我能掌握更多的技能，学到更多的知识！