



中国石油大学 (华东)  
CHINA UNIVERSITY OF PETROLEUM

## 2021—2022 学年第 2 学期 《程序设计(Python)》课程报告

专业班级	
姓 名	
学 号	
成绩	
评阅教师	张学辉

2022 年 6 月 1 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  $GrowthRate = (\text{年末人口数} - \text{年初人口数}) / \text{年平均人口} \times 1000\%$

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。可以视之作为一种改良（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:

---

```
D:\pyy\new\Scripts\python.exe D:/pythonProject5/main.py
```

	Year	Prediction
0	2021	1.375952e+09
1	2022	1.391362e+09
2	2023	1.406771e+09
3	2024	1.422181e+09
4	2025	1.437590e+09

```
进程已结束, 退出代码为 0
```

## 2. 数据处理

将数据按照年份排序

```
data = data.sort_values('Year').set_index('Year')
print(data)
```

效果展示:

```
D:\pyy\new\Scripts\python.exe D:/pythonProject5/main.py
```

```
      Population  GrowthRate
Year
1950    376325200         0.00
1951    382376948         1.61
1952    388799073         1.68
1953    395544369         1.73
1954    402578596         1.78
...      ...      ...
2017    1338676785         1.07
2018    1352642280         1.04
2019    1366417754         1.02
2020    1380004385         0.99
2021    1393409038         0.97
```

```
[72 rows x 2 columns]
```

```
进程已结束，退出代码为 0
```

### 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：

```
D:\pyy\new\Scripts\python.exe D:/pythonProject5/main.py
```

```
      Population  GrowthRate
Year
1950    376325200         0.00
1960    450547679         1.98
1970    555189792         2.23
1980    698952844         2.34
1990    873277798         2.10
2000   1056575549         1.78
2010   1234281170         1.36
2020   1380004385         0.99
```

```
进程已结束，退出代码为 0
```

#### 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:

D:\pyy\new\Scripts\python.exe D:/pythonProject5/main.py

	Population	GrowthRate
Year		
1961	459642165	2.02
1962	469077190	2.05
1963	478825608	2.08
1964	488848135	2.09
1965	499123324	2.10
1966	509631500	2.11
1967	520400576	2.11
1968	531513824	2.14
1969	543084336	2.18
1970	555189792	2.23
1971	567868018	2.28
1972	581087256	2.33
1973	594770134	2.35
1974	608802600	2.36
1975	623102897	2.35
1976	637630087	2.33
1977	652408776	2.32
1978	667499806	2.31
1979	682995354	2.32
1980	698952844	2.34
1981	715384993	2.35
1982	732239504	2.36
1983	749428958	2.35
1984	766833410	2.32
1985	784360008	2.29

1986	801975244	2.25
1987	819682102	2.21
1988	837468930	2.17
1989	855334678	2.13
1990	873277798	2.10
1991	891273209	2.06
1992	909307016	2.02

进程已结束，退出代码为 0

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。

### 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：

D:\pyy\new\Scripts\python.exe D:/pythonProject5/main.py

	Year	Prediction
0	2021	1.375952e+09
1	2022	1.391362e+09
2	2023	1.406771e+09
3	2024	1.422181e+09
4	2025	1.437590e+09
5	2026	1.452999e+09
6	2027	1.468409e+09
7	2028	1.483818e+09
8	2029	1.499228e+09
9	2030	1.514637e+09
10	2031	1.530047e+09
11	2032	1.545456e+09
12	2033	1.560865e+09
13	2034	1.576275e+09
14	2035	1.591684e+09
15	2036	1.607094e+09
16	2037	1.622503e+09
17	2038	1.637913e+09
18	2039	1.653322e+09
19	2040	1.668731e+09
20	2041	1.684141e+09
21	2042	1.699550e+09
22	2043	1.714960e+09
23	2044	1.730369e+09
24	2045	1.745779e+09
25	2046	1.761188e+09

```
26 2047 1.776597e+09
27 2048 1.792007e+09
28 2049 1.807416e+09
29 2050 1.822826e+09
30 2051 1.838235e+09
31 2052 1.853645e+09
32 2053 1.869054e+09
33 2054 1.884463e+09
34 2055 1.899873e+09
35 2056 1.915282e+09
36 2057 1.930692e+09
37 2058 1.946101e+09
38 2059 1.961511e+09
39 2060 1.976920e+09
40 2061 1.992330e+09
41 2062 2.007739e+09
42 2063 2.023148e+09
43 2064 2.038558e+09
44 2065 2.053967e+09
45 2066 2.069377e+09
46 2067 2.084786e+09
47 2068 2.100196e+09
48 2069 2.115605e+09
```

进程已结束，退出代码为 0

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

##### 1 人口数量柱状图绘制

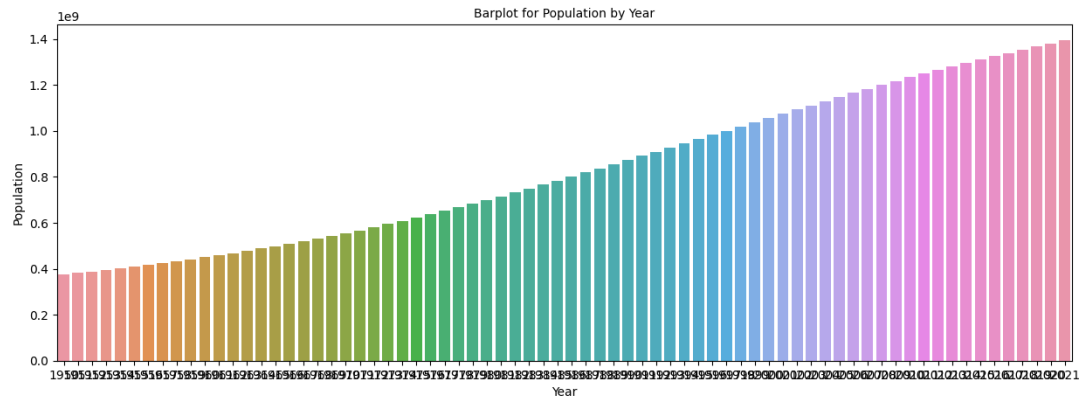
```
plt.figure(figsize=(15,5))

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

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

plt.show()
```

效果展示：



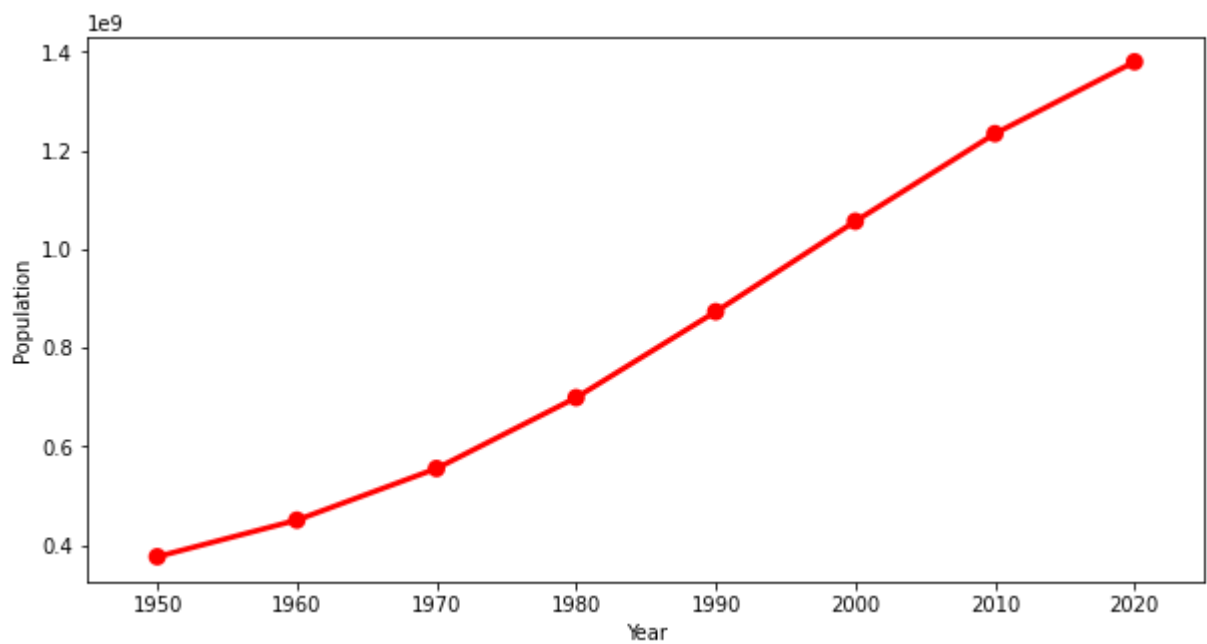
## 2 缩小区间折线图

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

sns.pointplot(year_data.index, year_data.Population, color='red')

plt.show()
```

效果展示:



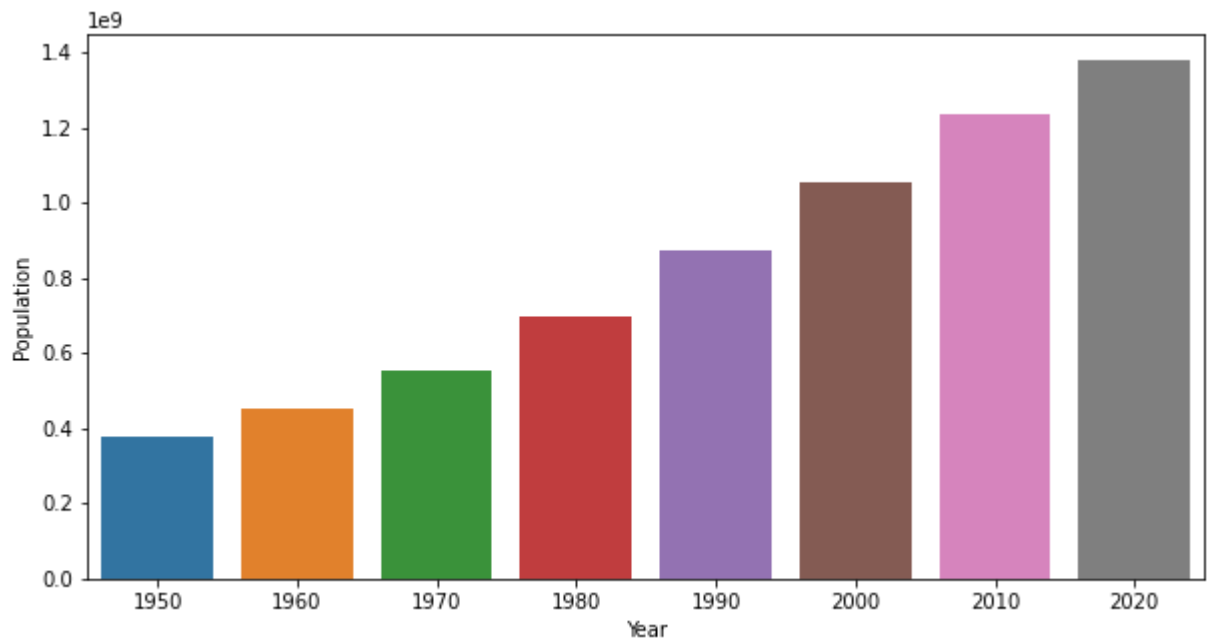
## 3 缩小区间柱状图

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

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

plt.show()
```

效果展示：



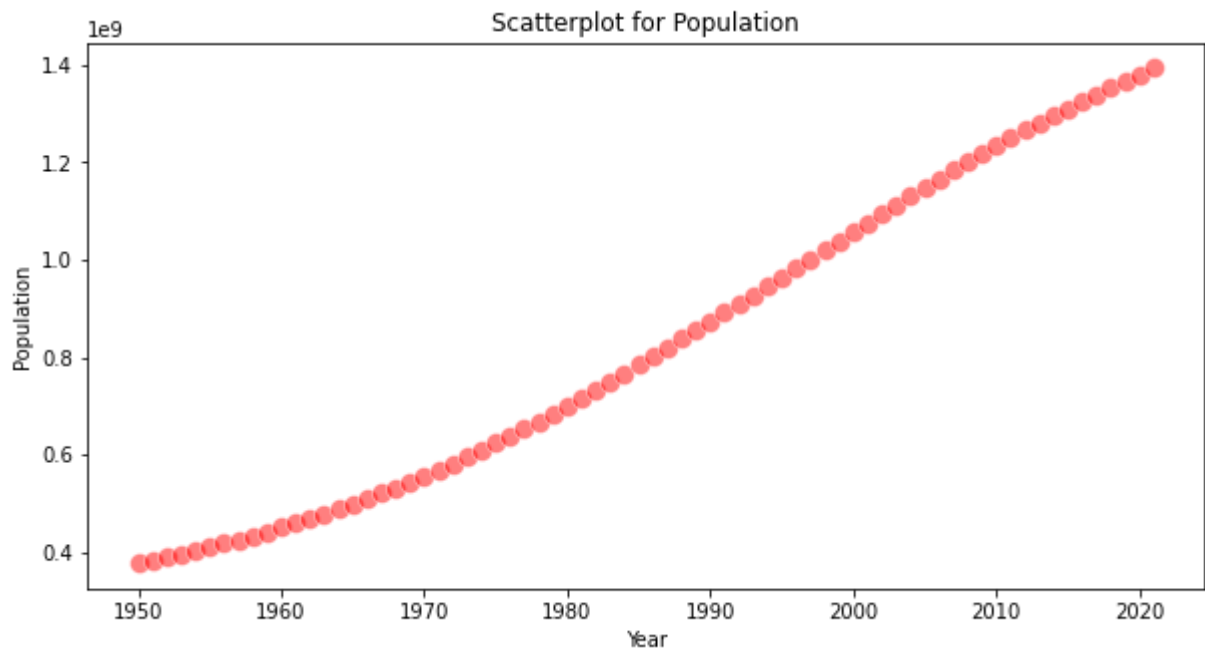
#### 4 人口数量点状图绘制

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

plt.title('Scatterplot for Population')

sns.scatterplot(data.index, data.Population,
                s=100, alpha=0.5, color='red')
```

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。

#### 5 增长率点状图

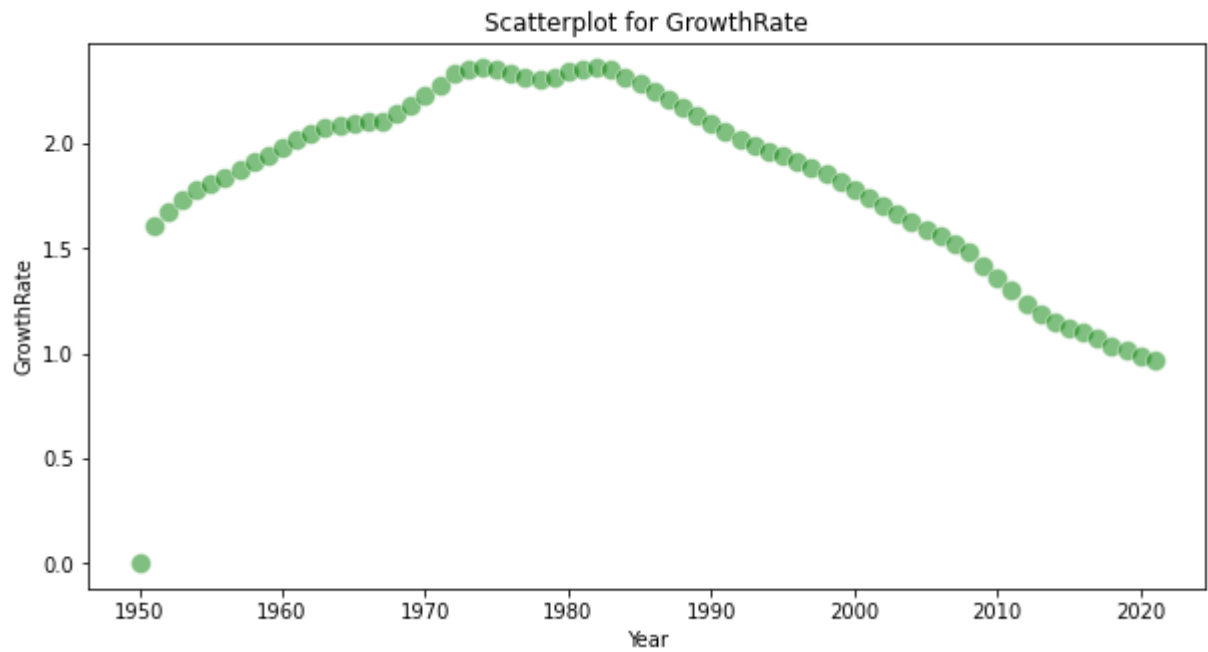
```
plt.figure(figsize=(10,5))

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。

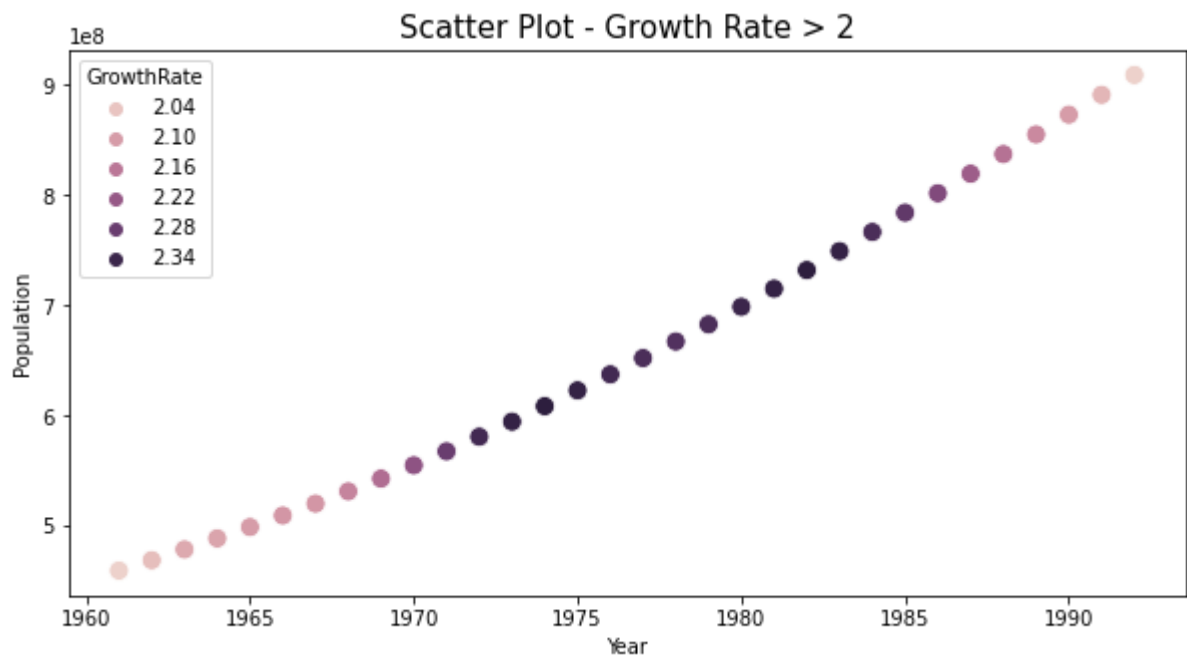
6 增长率大于 2 点状图

```
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)
```

效果展示:



7 增长率大于 2 柱状图

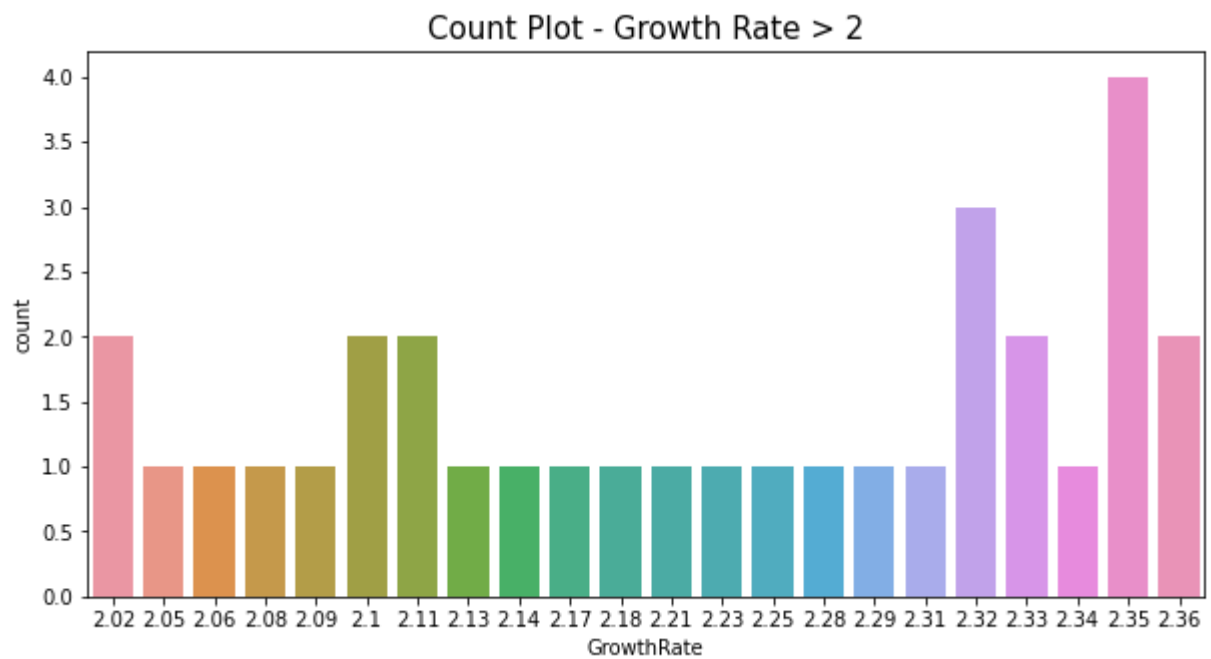
```
plt.figure(figsize=(10,5))

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

sns.countplot(gr_year.GrowthRate)

plt.show()
```

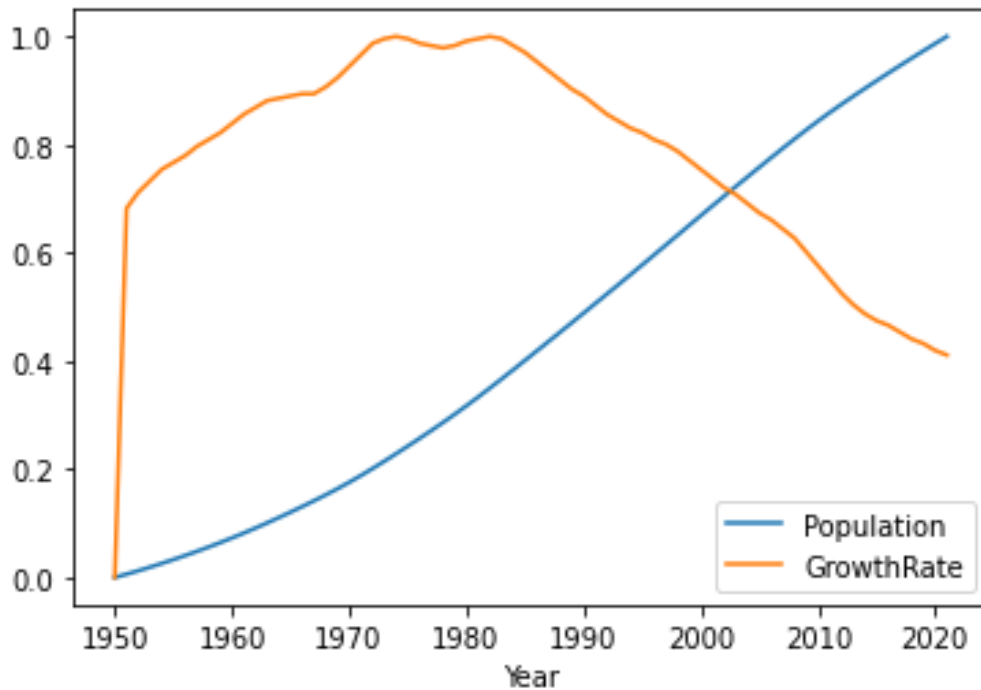
效果展示:



8 曲线图

```
norm_data.plot()  
  
plt.show()
```

效果展示:



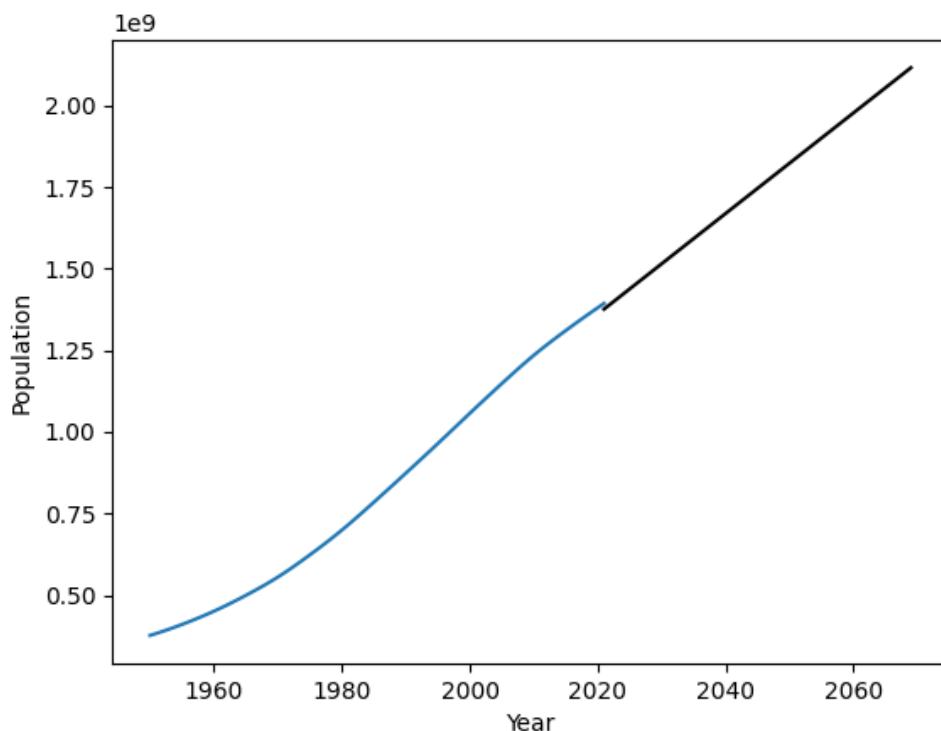
由图我们将两条曲线结合分析可以得出人口数量的曲线虽然一直上升，但是斜率却发生了变化，总体的人口数量曲线由下凸变为上凸，与人口增长率的曲线对应。

#### 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：

```
sns.lineplot(x=data['Year'],y=data['Population'])  
  
sns.lineplot(x=Pred_X['Year'],y=Pred_X['Prediction'],color='black')
```

效果展示:



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

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：经过排除文件名拼写错误后，得出问题在于路径。将 csv 文件复制到项目文件夹中，使用相对路径使问题得到解决。

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！

评分参考	评价				
	优秀	良好	中等	及格	不及格
数据描述：各字段描述清楚					
数据预处理：是否进行了预处理，为什么要进行相关处理，原因阐述是否清晰					
数据分析：多角度数据分析，描述是否清楚准确					
可视化展示：多种图表展示，类型选择是否恰当，图表是否完整美观，能否根据图表得出相应结论					
问题及解决方法：遇到的问题是否描述清楚，解决方法是否清晰					
文档结构完整、格式排版美观、描述清楚准确					
答辩：视频录制完整，讲解清晰，重点突出，作品效果					