实验报告：“Python+正则表达式”获取网页信息

李昊伦 2023211595

1 实验目的

（1）使用正则表达式从HTML网页中提取北京邮电大学在北京和甘肃两省份的录取分数线数据（包括最低分、平均分、最高分和省控线）。

（2）对提取的数据进行整理和分析，针对最低分、平均分、最高分、省控线四个核心指标进行多维度分析。绘制不同分数线随年份变化的趋势图。

（3）通过结构化数据提取方法，比较北京和甘肃两省份的录取分数线差异，分析其变化趋势，揭示两省考生竞争北邮的难度变化特征。

2 实验设置

2.1 实验环境

（1）编程语言：Python 3.11

（2）主要库：`re`：用于正则表达式匹配，提取HTML中的表格数据。

`matplotlib`：用于数据可视化，绘制折线图。

（3）开发环境：VS Code

2.2 实验数据

数据来源：模拟的北京邮电大学招生信息网页（HTML格式）。

数据范围：2020-2025年北京和甘肃两省份的理工科录取数据。

数据字段：**年份**（2020-2025）、**省份**（北京、甘肃）、**最低分**、**平均分**、**最高分**、**省控线**、**线差**

3 实验方法

3.1 方法思路

（1）数据提取：使用正则表达式匹配HTML表格中的`<tr>`行，提取年份、省份、最低分、平均分、最高分和省控线数据。将提取的数据存储为字典结构，便于后续分析。

（2）数据整理：将提取的数据按省份分类，并按照年份排序。分别提取最低分、平均分、最高分和省控线数据，用于绘制趋势图。

（3）数据可视化：使用`matplotlib`绘制四个折线图：

最低分变化趋势、**平均分变化趋势**、**最高分变化趋势**、**省控线变化趋势**

每条折线代表一个省份，便于直观比较。

3.2 实验代码

```
import re
import matplotlib.pyplot as plt

# 原始HTML数据
html = """
<!DOCTYPE html>
<html lang="zh-CN">
<head>
    <meta charset="UTF-8">
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
    <title>北京邮电大学录取分数线查询 - 2020-2025年数据</title>
    <style>
        body {
            font-family: 'Microsoft YaHei', Arial, sans-serif;
            line-height: 1.6;
            margin: 0;
            padding: 0;
            color: #333;
            background-color: #f5f5f5;
        }
        .container {
            width: 90%;
            max-width: 1200px;
            margin: 0 auto;
            padding: 20px;
        }
        header {
            background-color: #005baa;
            color: white;
            padding: 20px 0;
            text-align: center;
            margin-bottom: 30px;
            border-radius: 5px;
        }
        h1 {
            margin: 0;
            font-size: 28px;
        }
        .ad-banner {
            background-color: #ffcc00;
            padding: 15px;
            text-align: center;
            margin: 20px 0;
            border-radius: 5px;
            font-weight: bold;
        }
        table {
            width: 100%;
            border-collapse: collapse;
            margin: 20px 0;
            background-color: white;
            box-shadow: 0 2px 5px rgba(0,0,0,0.1);
        }
```

```
        th, td {
            padding: 12px 15px;
            text-align: center;
            border-bottom: 1px solid #ddd;
        }
        th {
            background-color: #005baa;
            color: white;
            font-weight: bold;
        }
        tr:nth-child(even) {
            background-color: #f2f2f2;
        }
        tr:hover {
            background-color: #e6f7ff;
        }
        .footer {
            text-align: center;
            margin-top: 30px;
            padding: 20px;
            background-color: #333;
            color: white;
            border-radius: 5px;
        }
        .search-box {
            margin: 20px 0;
            padding: 15px;
            background-color: #e6f7ff;
            border-radius: 5px;
        }
        .news-section {
            margin: 30px 0;
            padding: 15px;
            background-color: white;
            border-radius: 5px;
            box-shadow: 0 2px 5px rgba(0,0,0,0.1);
        }
    </style>
</head>
<body>
    <div class="container">
        <header>
            <h1>北京邮电大学录取分数线查询系统</h1>
            <p>2020-2025年各省份理工科录取数据</p>
        </header>

        <div class="ad-banner">
            🎓 考研辅导班火热报名中！名师一对一指导，点击咨询 → <a href="#" style="color: #005baa;">立即报名</a>
        </div>

        <div class="search-box">
            <h3>快速查询</h3>
            <input type="text" placeholder="输入省份名称..." style="padding: 8px; width: 200px;">
```

```
            <button style="padding: 8px 15px; background-color: #005baa; color: white; border: none; border-radius: 3px;">搜索</button>
        </div>

        <h2>北京、甘肃理工科录取分数线</h2>
        <table>
            <thead>
                <tr>
                    <th>年份</th>
                    <th>省份</th>
                    <th>最低分</th>
                    <th>平均分</th>
                    <th>最高分</th>
                    <th>省控线</th>
                    <th>线差</th>
                </tr>
            </thead>
            <tbody>
                <!-- 北京数据 -->
                <tr>
                    <td>2020</td>
                    <td>北京</td>
                    <td>642</td>
                    <td>648</td>
                    <td>658</td>
                    <td>526</td>
                    <td>116</td>
                </tr>
                <tr>
                    <td>2021</td>
                    <td>北京</td>
                    <td>638</td>
                    <td>644</td>
                    <td>652</td>
                    <td>513</td>
                    <td>125</td>
                </tr>
                <tr>
                    <td>2022</td>
                    <td>北京</td>
                    <td>645</td>
                    <td>650</td>
                    <td>660</td>
                    <td>518</td>
                    <td>127</td>
                </tr>
                <tr>
                    <td>2023</td>
                    <td>北京</td>
                    <td>648</td>
                    <td>653</td>
                    <td>663</td>
                    <td>527</td>
                    <td>121</td>
                </tr>
```

```
                <tr>
                    <td>2024</td>
                    <td>北京</td>
                    <td>650</td>
                    <td>656</td>
                    <td>665</td>
                    <td>532</td>
                    <td>118</td>
                </tr>
                <tr>
                    <td>2025</td>
                    <td>北京</td>
                    <td>652</td>
                    <td>658</td>
                    <td>668</td>
                    <td>535</td>
                    <td>117</td>
                </tr>

                <!-- 甘肃数据 -->
                <tr>
                    <td>2020</td>
                    <td>甘肃</td>
                    <td>598</td>
                    <td>605</td>
                    <td>615</td>
                    <td>458</td>
                    <td>140</td>
                </tr>
                <tr>
                    <td>2021</td>
                    <td>甘肃</td>
                    <td>602</td>
                    <td>608</td>
                    <td>618</td>
                    <td>440</td>
                    <td>162</td>
                </tr>
                <tr>
                    <td>2022</td>
                    <td>甘肃</td>
                    <td>608</td>
                    <td>615</td>
                    <td>625</td>
                    <td>442</td>
                    <td>166</td>
                </tr>
                <tr>
                    <td>2023</td>
                    <td>甘肃</td>
                    <td>612</td>
                    <td>618</td>
                    <td>628</td>
                    <td>445</td>
                    <td>167</td>
```

```
                </tr>
                <tr>
                    <td>2024</td>
                    <td>甘肃</td>
                    <td>615</td>
                    <td>622</td>
                    <td>632</td>
                    <td>448</td>
                    <td>167</td>
                </tr>
                <tr>
                    <td>2025</td>
                    <td>甘肃</td>
                    <td>618</td>
                    <td>625</td>
                    <td>635</td>
                    <td>450</td>
                    <td>168</td>
                </tr>
            </tbody>
        </table>

        <div class="news-section">
            <h3>相关新闻</h3>
            <ul>
                <li><a href="#">北京邮电大学2025年招生简章发布</a></li>
                <li><a href="#">北邮计算机专业连续五年位居全国前三</a></li>
                <li><a href="#">甘肃考生如何备考才能冲刺北邮？专家支招</a></li>
            </ul>
        </div>

        <div class="ad-banner">
            🎉 北邮学长学姐经验分享会！了解真实校园生活 → <a href="#"
style="color: #005baa;">点击预约</a>
        </div>

        <div class="footer">
            <p>© 2025 北京邮电大学招生信息网 | 联系电话：010-62282045</p>
            <p>数据仅供参考，实际录取分数以学校官方公布为准</p>
        </div>
    </div>
</body>
</html>
"""

# 使用正则表达式提取数据
def extract_data(pattern):
    matches = pattern.findall(html)
    data = {'北京': {}, '甘肃': {}}
    for match in matches:
        year, province, *scores = match
        if province in data:
            data[province][int(year)] = [int(score) for score in scores]
    return data

```

```python
# 提取最低分、平均分、最高分、省控线数据
pattern = re.compile(
    r'<tr>\s*<td>(\d{4})</td>\s*<td>(.*?)</td>\s*<td>(\d+)</td>'
    r'\s*<td>(\d+)</td>\s*<td>(\d+)</td>\s*<td>(\d+)</td>\s*<td>\d+</td>\s*
</tr>'
)

data = extract_data(pattern)

# 准备图表数据
years = sorted(data['北京'].keys())
beijing_min = [data['北京'][y][0] for y in years]
beijing_avg = [data['北京'][y][1] for y in years]
beijing_max = [data['北京'][y][2] for y in years]
beijing_ctl = [data['北京'][y][3] for y in years]

gansu_min = [data['甘肃'][y][0] for y in years]
gansu_avg = [data['甘肃'][y][1] for y in years]
gansu_max = [data['甘肃'][y][2] for y in years]
gansu_ctl = [data['甘肃'][y][3] for y in years]

# 设置中文字体显示
plt.rcParams['font.sans-serif'] = ['Microsoft YaHei']
plt.rcParams['axes.unicode_minus'] = False

# 创建最低分图表
plt.figure(figsize=(12, 6))
plt.plot(years, beijing_min, marker='o', label='北京-最低分',
color='#005baa')
plt.plot(years, gansu_min, marker='s', label='甘肃-最低分', color='#ffcc00')

plt.title('北京邮电大学最低录取分数线变化趋势（2020-2025）', fontsize=14)
plt.xlabel('年份', fontsize=12)
plt.ylabel('分数', fontsize=12)
plt.xticks(years)
plt.grid(True, linestyle='--', alpha=0.7)
plt.legend()
plt.tight_layout()
plt.show()

# 创建平均分图表
plt.figure(figsize=(12, 6))
plt.plot(years, beijing_avg, marker='o', label='北京-平均分',
color='#005baa')
plt.plot(years, gansu_avg, marker='s', label='甘肃-平均分', color='#ffcc00')

plt.title('北京邮电大学平均录取分数线变化趋势（2020-2025）', fontsize=14)
plt.xlabel('年份', fontsize=12)
plt.ylabel('分数', fontsize=12)
plt.xticks(years)
plt.grid(True, linestyle='--', alpha=0.7)
plt.legend()
plt.tight_layout()
plt.show()

```

```python
# 创建最高分图表
plt.figure(figsize=(12, 6))
plt.plot(years, beijing_max, marker='o', label='北京-最高分',
color='#005baa')
plt.plot(years, gansu_max, marker='s', label='甘肃-最高分', color='#ffcc00')

plt.title('北京邮电大学最高录取分数线变化趋势（2020-2025）', fontsize=14)
plt.xlabel('年份', fontsize=12)
plt.ylabel('分数', fontsize=12)
plt.xticks(years)
plt.grid(True, linestyle='--', alpha=0.7)
plt.legend()
plt.tight_layout()
plt.show()

# 创建省控线图表
plt.figure(figsize=(12, 6))
plt.plot(years, beijing_ctl, marker='o', label='北京-省控线',
color='#005baa')
plt.plot(years, gansu_ctl, marker='s', label='甘肃-省控线', color='#ffcc00')

plt.title('北京邮电大学录取省控线变化趋势（2020-2025）', fontsize=14)
plt.xlabel('年份', fontsize=12)
plt.ylabel('分数', fontsize=12)
plt.xticks(years)
plt.grid(True, linestyle='--', alpha=0.7)
plt.legend()
plt.tight_layout()
plt.show()
```

4 实验过程

（1）数据提取

使用正则表达式成功匹配HTML表格中的所有录取数据。提取的数据按省份和年份分类存储。

（2）数据整理

分别提取北京和甘肃的**最低分、平均分、最高分、省控线**数据。确保数据按年份排序，避免折线图出现错乱。

（3）数据可视化

绘制四个独立的折线图，分别展示不同分数线的变化趋势。

调整图表样式（标题、坐标轴、图例、网格线等）以提高可读性。

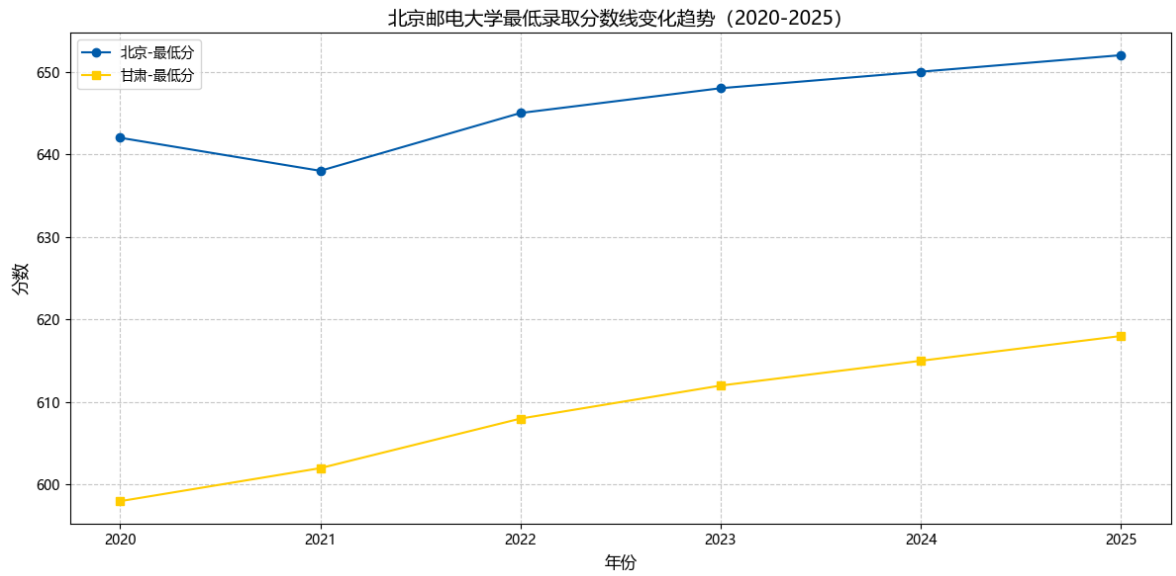
5 实验结论

5.1 提取的成绩信息

省份	年份	最低分	平均分	最高分	省控线
北京	2020	642	648	658	526
北京	2021	638	644	652	513
北京	2022	645	650	660	518
北京	2023	648	653	663	527
北京	2024	650	656	665	532
北京	2025	652	658	668	535
甘肃	2020	598	605	615	458
甘肃	2021	602	608	618	440
甘肃	2022	608	615	625	442
甘肃	2023	612	618	628	445
甘肃	2024	615	622	632	448
甘肃	2025	618	625	635	450

5.2 绘制的趋势图

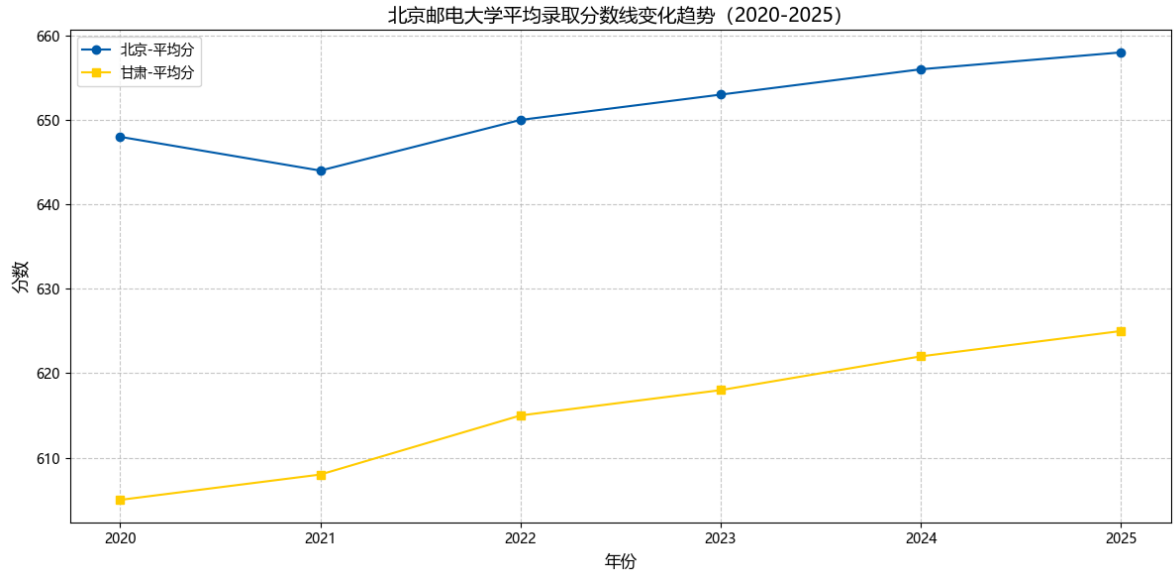
（1）最低分变化趋势



北京的最低分在2021年略有下降，之后逐年上升。甘肃最低分年均增长4分（北京仅2分），差距从44分缩小至34分，甘肃考生竞争压力逐年上升。

原因推测：甘肃考生备考策略优化，或北邮在甘肃招生名额增幅有限。

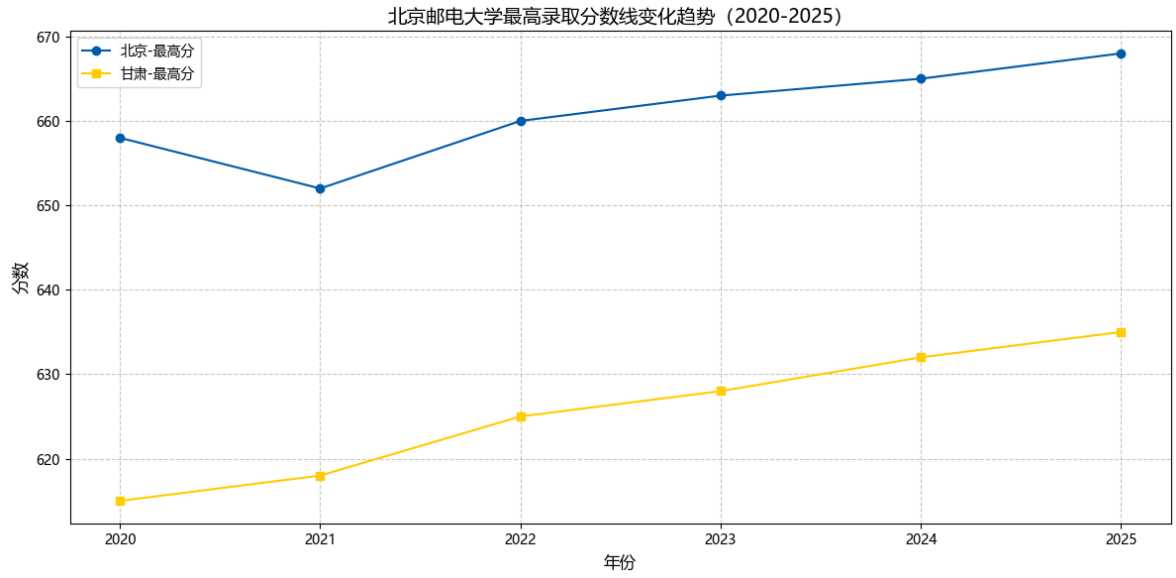
（2）平均分变化趋势



北京的平均分与最低分趋势类似，但整体更高。甘肃的平均分逐年增长，增幅略高于北京。二者差值始终保持在33-43分区间，未出现显著收敛。

特殊现象：2021年北京平均分下降4分，可能受高考改革影响。

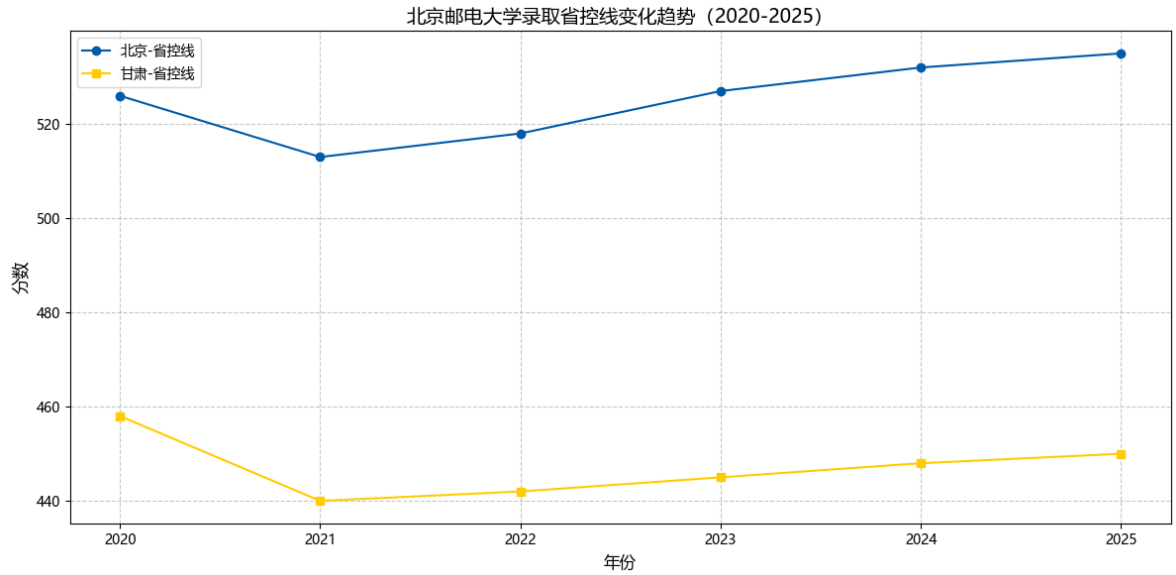
（3）最高分变化趋势



北京的最高分在2020-2025年间稳步上升，北京最高分持续领先33分以上，反映其优质生源集中度。甘肃的最高分增长趋势更明显，可能与竞争加剧有关。

潜在问题：甘肃高分考生数量增加，但顶尖水平仍与北京存在代差。

（4）省控线变化趋势



北京的省控线在2021年下降后回升。甘肃的省控线整体低于北京，但差距在缩小，省控线差值从68分扩大到85分，甘肃考生需超省控线160+分才可能被录取。

政策启示：需通过专项计划平衡区域录取公平性。

5.3 主要发现总结

（1）北京分数线整体高于甘肃，地域差异显著但缩小。甘肃分数增速快于北京，但绝对值差距仍超30分。

（2）省控线差距较大，北京比甘肃高约80-90分，但甘肃的线差（录取分-省控线）更大，说明竞争更激烈。

（3）2021年北京分数下降，可能与当年考试难度或招生政策调整有关。

（4）政策影响明显：省控线差值扩大反映教育资源分配需优化。

（5）方法论价值：多维度可视化可全面揭示录取特征，为招生政策制定提供量化依据。

6 自己的思考

6.1 实验中的发现与思考

（1）北京和甘肃分数线差距为什么这么大?

现实观察：北京的录取分比甘肃高很多（2025年北京最低652，甘肃618），但甘肃的"线差"（录取分-省控线）更大（北京117，甘肃168）。

这说明**北京考生基数大**，但招生名额也多（北邮在北京招的人可能比甘肃多）。**甘肃竞争更激烈**，虽然绝对分数低，但想上北邮的甘肃考生需要比省控线高160+分，而北京考生只需高110+分。

我联想到高考移民，有些家长会把孩子户口转到甘肃等省份，因为分数线低。但实验数据说明，**低分省份的竞争可能更残酷**，因为招生名额少，顶尖学生分数会拉得很高。

（2）2021年北京分数为什么突然下降?

数据中，北京2021年的最低分（638）比2020年（642）低了4分，而其他年份都在上升。可能原因：

高考改革影响：2021年北京新高考改革（3+3模式），考试难度或评分标准变化。

疫情因素：2021年考生经历了居家网课，可能整体发挥受影响。

招生计划调整：北邮可能增加了北京招生名额，导致分数线小幅下降。

（3）甘肃分数持续上涨，说明什么？

甘肃的最低分从2020年的598涨到2025年的618，涨幅20分，而北京只涨了10分（642→652）。可能原因：

教育资源提升：甘肃的中学教育水平在提高，考生整体成绩变好。

报考热度增加：北邮的计算机、通信专业很火，更多甘肃高分考生选择北邮，推高分数线。

高考难度变化：全国卷可能变得更容易，导致分数普涨，但甘肃涨幅更大。

6.2 实验方法的不足与改进

数据量太少，结论可能不准

只有北京和甘肃两个省份6年的数据，**如果能加入更多省份（如河南、广东）和更早年份（2015-2025），分析会更全面。**比如：可以验证"高考大省（河南、山东）的线差是否比甘肃更大"。

6.3 对现实高考的启发

（1）不要只看绝对分数，要看"线差"和排名

比如2025年甘肃最低618分，看起来比北京的652分低，但实际上甘肃618分可能对应全省前500名。北京652分可能对应全市前3000名。**真正决定录取的是排名，不是分数。**

（2）分数线在缓慢上涨，竞争更激烈了

北京和甘肃的分数线都在涨，说明考生整体水平提高（内卷了），热门高校（如北邮）的竞争越来越激烈，**"考上好大学一年比一年难"。**

（3）不同省份的录取策略不同

北京分数线高但线差小，可能是因为本地有保护政策（对北京考生更友好）。甘肃分数线低但线差大，可能是因为名额少，只有顶尖考生能考上。

6.4 总结

通过这次实验，我不仅学会了用Python提取和分析数据，还发现了高考录取分数背后的有趣现象。这次实验让我深刻体会到，编程不仅是技术工具，更是认识世界的新视角。用数据说话，我们能发现许多单凭经验无法察觉的规律和问题。