****



**《Python语言与数据科学》**

**总结报告**

**Project Report**

**姓 名： 陆淼 学 号： 22120122**

**院 系：仪器科学与工程学院 课程编号：**

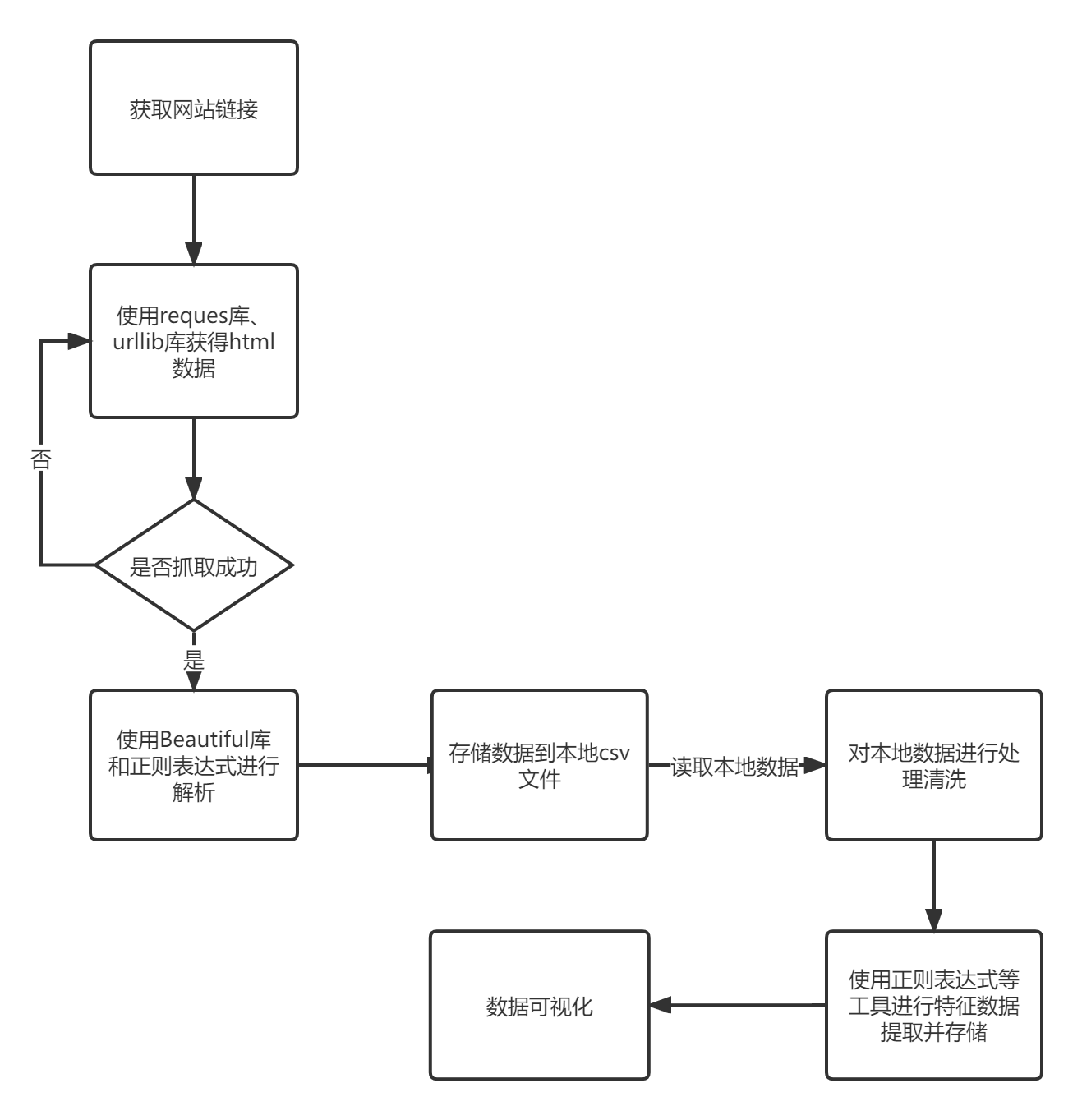
**上课地点： 完成时间： 2022.6.1**

**上课教师：**

**一、项目综述**

本次大作业主要以影响力较大的电影榜单豆瓣TOP250作为数据抓取对象，通过Python网络爬虫爬取相关数据并进行数据清洗整理，存储到本地文件中。后续通过读取文件、数据处理和提取整理出了一些有价值的统计信息，最后用可视化方法直观展示数据。

大致思路如下：



主程序为douban.py。运行后会自动生成抓取数据分析处理生成原数据集douban.py、特征数据集douban\_info.py、可视化结果county.png、score.png、type.png、year.png。

**二、爬取部分简介**

本次爬取的目标网站为<https://movie.douban.com/top250?start=>，网站共分为十页每页二十五条数据共250条。采取循环递增改变目标URL的方法获取全部html信息。爬取后的部分代码如下：

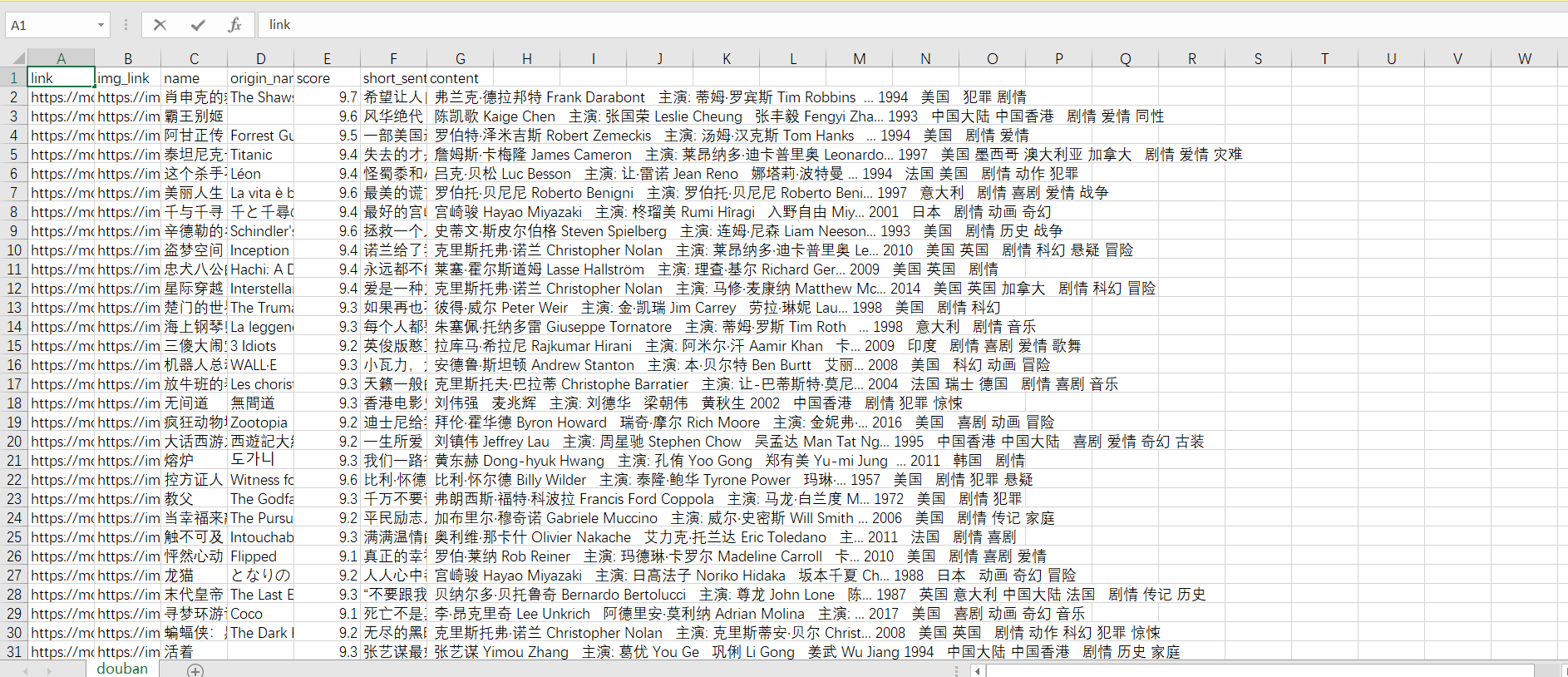
1. **<div** class="item"**>**
2. **<div** class="pic"**>**
3. **<em** class=""**>**1**</em>**
4. **<a** href="https://movie.douban.com/subject/1292052/"**>**
5. **<img** width="100" alt="肖申克的救赎" src="https://img2.doubanio.com/view/photo/s\_ratio\_poster/public/p480747492.webp" class=""**>**
6. **</a>**
7. **</div>**
8. **<div** class="info"**>**
9. **<div** class="hd"**>**
10. **<a** href="https://movie.douban.com/subject/1292052/" class=""**>**
11. **<span** class="title"**>**肖申克的救赎**</span>**
12. **<span** class="title"**>** / The Shawshank Redemption**</span>**
13. **<span** class="other"**>** / 月黑高飞(港)  /  刺激1995(台)**</span>**
14. **</a>**

17. **<span** class="playable"**>**[可播放]**</span>**
18. **</div>**
19. **<div** class="bd"**>**
20. **<p** class=""**>**
21. 导演: 弗兰克·德拉邦特 Frank Darabont   主演: 蒂姆·罗宾斯 Tim Robbins /...**<br>**
22. 1994 / 美国 / 犯罪 剧情
23. **</p>**

26. **<div** class="star"**>**
27. **<span** class="rating5-t"**></span>**
28. **<span** class="rating\_num" property="v:average"**>**9.7**</span>**
29. **<span** property="v:best" content="10.0"**></span>**
30. **<span>**2627118人评价**</span>**
31. **</div>**
33. **<p** class="quote"**>**
34. **<span** class="inq"**>**希望让人自由。**</span>**
35. **</p>**
36. **</div>**
37. **</div>**
38. **</div>**

通过分析html源码的组成方式写出相应正则表达式，进行七项特征的数据解析：电影播放链接、封面链接、电影名、电影本土名、电影评分、电影简评和电影有关信息（包含导演、主要演员、年份和制作有关国家等）。封装成函数，返回值是已经填充好的数据。

**三、爬取后的数据存储和清洗：**

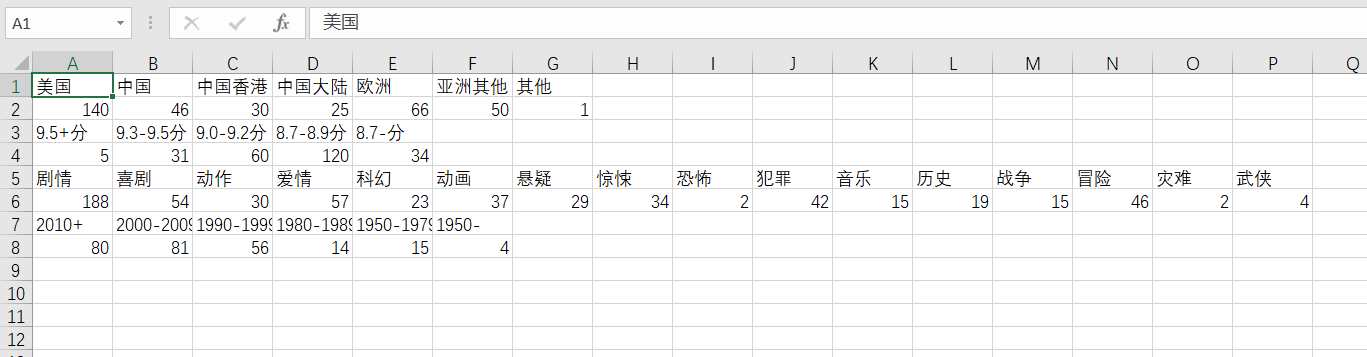
利用存储功能将提取出的数据存储为’douban.csv’。基本效果图如下：

通过正则表达式和一些基本操作去除数据中的一些不必要空格和无关信息等，为接下来数去特征的提取做准备。

**四、数据特征提取：**

提取四项有一定价值的数据：电影拍摄主要国家、评分、题材和制作年份。对于评分可以直接进行遍历统计，其他三项包含在有关信息的字符串中可以通过正则表达式简洁地提取并统计。

统计的结果存储为’douban\_info.csv’，结果如下：

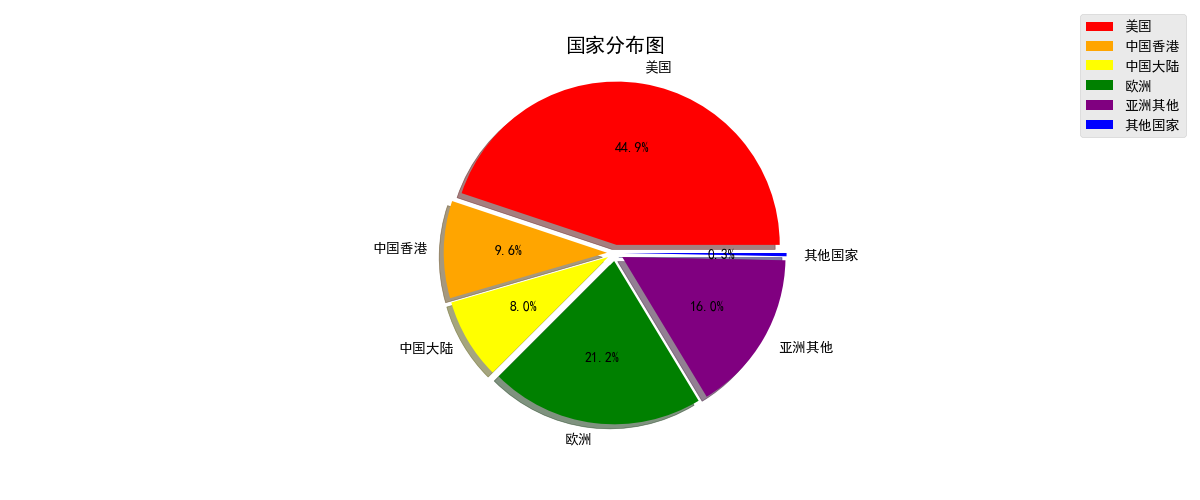


作为后续可视化地数据来源。

**五、数据可视化：**

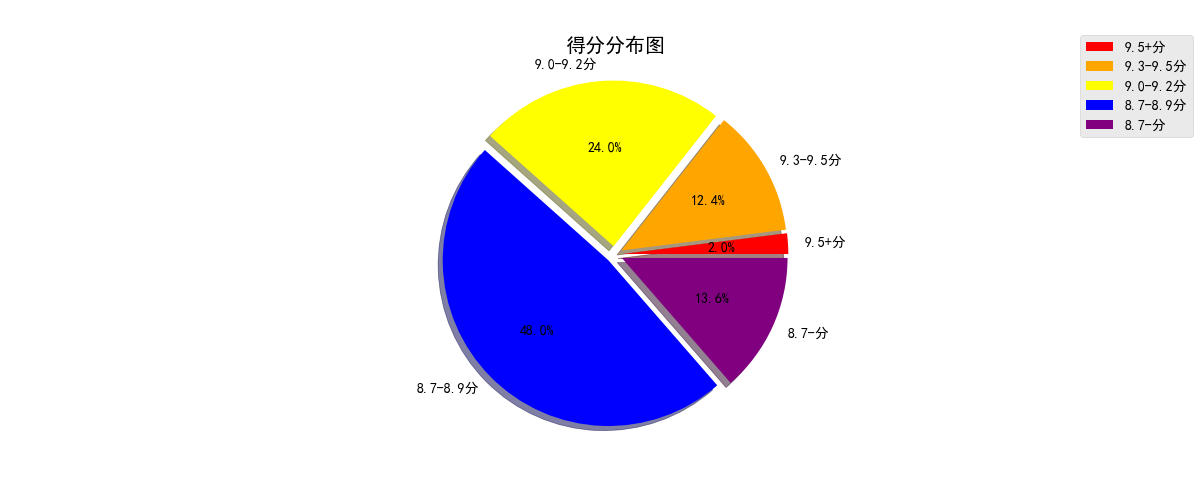
根据爬取、清洗和提取后的特征数据的特点，针对性地使用合适的可视化方案。

1.主要国家：(country.png)



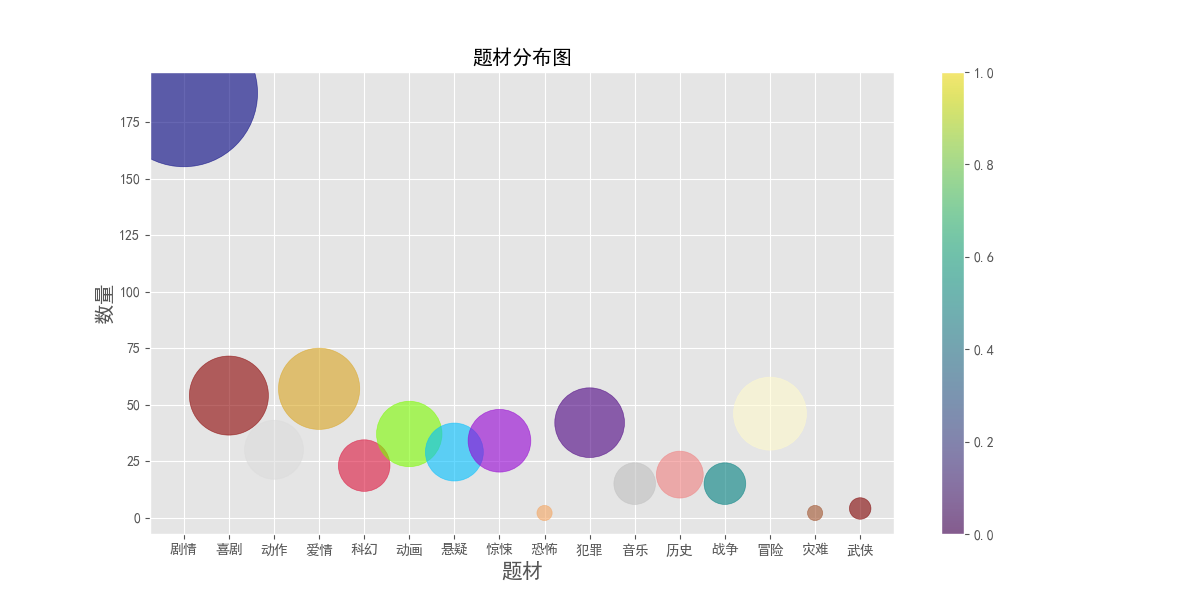
可以看到豆瓣TOP250中美国制作的电影很多占到了几乎半数，中国、欧洲和日本韩国等亚洲其他国家分布比较均匀。

2.评分:(score.png)



可以看到大部分分布在8.7-8.9分，9.5+的高分和8.7-的低分较少。

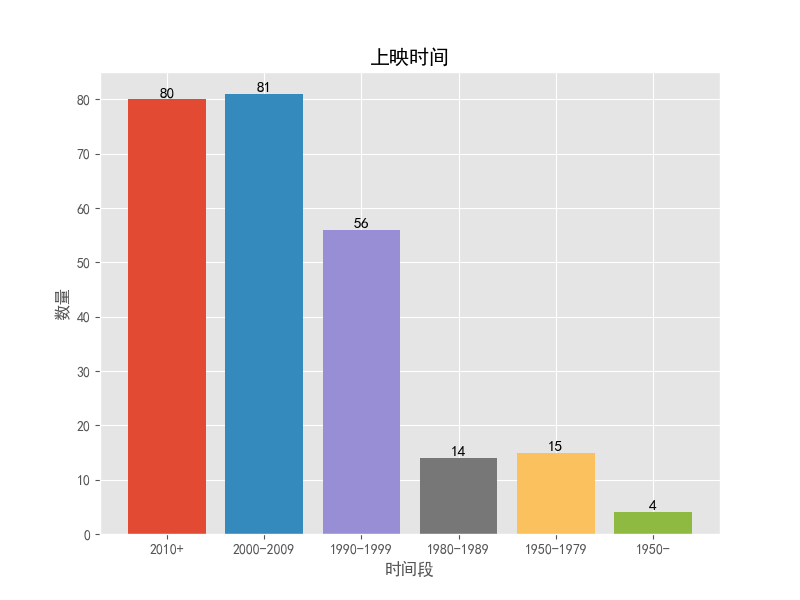
3.题材:(type.png)



注意数据点的圆尺寸正比于其数据量。

大部分的电影都有剧情的标签，反映出高分电影更注重剧情性。同时电影的题材丰富多彩，分布较为均匀。

4.年份:(year.png)



可以看出影片年份集中在1990年及以后，更老的电影占比较少。

**六、总结：**

通过这次通选课我基本了解了数据从获取到清洗处理存储再到可视化的基本流程，掌握了很多有用的知识。我发现最难的部分应该是爬取，源代码的解析即使使用解析库XPATH和BeautifulSoup也较为困难，编写正则表达式也容易出错。因此这次爬取的目标网站较为简单，网站源代码排版十分整齐便于找出标签。

数据存储的时候采用了本地存储的方式。原本也尝试了数据库的方式不过存储数据量不大且csv文件也有较多的接口函数因此没有使用。

特征数据提取的部分较为简单，使用了Python自带的字符串检索语法与Re正则表达式相结合，轻松提取出了四项有价值的数据。

可视化部分则应用了三种绘图方式进行展示。由于此次爬取并提取的数据以分类为主而不是变化的数据，因此采用了柱状图、饼图等清晰的展示。

总的来说通过这次大作业实践受益匪浅，希望以后能有更多的机会应用学习到的知识。