|  |  |
| --- | --- |
| 学号 | 10212818128 |



课 程 设 计

课程名称 python程序设计课程设计

|  |  |
| --- | --- |
| 题 目 | 节假日长假景点人满为患的数据分析 |
| 专 业 | \_\_\_\_\_\_\_ 软件工程\_\_\_\_\_\_\_\_\_\_ |
| 班 级 | \_\_\_\_\_\_\_\_\_\_软件1181\_\_\_\_\_\_\_\_\_\_ |
| 姓 名 | \_\_\_\_\_\_\_\_\_\_\_薛文\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 成 绩 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 指 导 老 师 | \_\_\_\_\_\_\_\_\_\_\_徐勇\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

2021 年 1 月 11 日至 2021 年 1 月 22 日

武汉华夏理工学院信息工程学院

**课 程 设 计 任 务 书**

课程名称：python程序设计课程设计 指导教师：徐勇

班级名称： 软件1181-1182 开课院、系：计算机与网络工程系

**一、课程设计目的与任务**

“python程序设计课程设计”是一个综合性的学习实践型实验教学环节，将在“python程序设计”课程的授课基础上，对python的基础语法、python 对文件的操作、python 对数据的操作、数据处理基础、数据可视化等若干个知识点进行综合运用。

python是一种具有天然开源基因的编程语言，了解开源社区和广泛的使用开源工具，也是Python学习实践的重要环节。因此在本次课程设计中，要求掌握jupyter notebook、Git等常用工具，以及github等重要开源社区的使用。

**二、课程设计的内容与基本要求**

数据分析的基本技术和方法在“python程序设计”课程中已经进行完整的讲授，由于数据分析本身是基于业务场景的，因此本次课程设计环节更加偏重于实际的业务场景的实践。通过对近期互联网热点的调查，准备了15个具体的业务场景，用于本次课程设计的具体任务场景。具体包括如下：

1. 北上广深租房状况分析；
2. 蔡某坤粉丝数及转发数据真假状况分析；
3. 地震的数据分析；
4. 英文名字的数据分析；
5. 外籍英文老师收入虚高情况数据分析；
6. 我国城市空气污染和烟花燃放的关系分析；
7. 针对996工作，程序员群体的看法的分析；
8. 吴某凡微博热点的分析；
9. 节假日长假景点人满为患的数据分析；
10. 针对荔枝的品种、销售地等维度，进行价格数据分析；
11. 分析芒果TV《我是大侦探》的观众评论数据；
12. 针对当前儿科医生的缺乏，对相关数据进行分析；
13. 著名网游《绝地求生》的数据分析；
14. 实习岗位状况的数据分析；
15. 电影《流浪地球》的观众评价的数据分析。

每三个学生组建一个课程设计小组，最后的任务输出包括代码每小组一份、课程设计报告每人一份、答辩ppt每小组一份，并进行课程设计成果答辩。小组成员均参与前述工作，但是每个人的侧重点不同。

每个课程设计小组可以从上述15个场景中选取一个作为课程设计的选题，选题中提供了待分析的数据，和现有的分析方法。各小组，通过学习和实践现有的分析方法，理解实战分析的思维过程并锻炼实际动手能力，再此基础上可以扩展更多维度的分析和数据展现形式。

每个班每个选题最多只能被两个小组选中，先选先得。同时，如果各小组发现更有意思的场景，并能够获取到相关待分析的数据，也可以申请作为选题方向。

本次课程设计的目标是培养学生的团队协作能力、对python知识点的综合运用、对实际场景的理解和适应能力、针对答辩的表达能力等。注重过程，期待成果，但不强求结果的尽善尽美。

**三、学时分配进度安排**

|  |  |  |
| --- | --- | --- |
| 序号 | 设计内容 | 所用时间 |
| 1 | 下发任务书，学生查阅资料 | 1天 |
| 2 | 组建团队，并进行选题和团队匹配 | 1天 |
| 3 | 各小组对自己的选题开始进行研究 | 2天 |
| 4 | 各小组开始准备数据并形成初步处理意见 | 1天 |
| 5 | 代码和实现分析的动手实践 | 2天 |
| 6 | 形成初步的报告书和ppt | 1天 |
| 7 | 答辩并完成报告书 | 2天 |
| 合 计 | | 2周 |

**四、课程设计考核及评分标准**

**1.设计报告要求**

课程设计报告要求逻辑清晰、层次分明、书写整洁。课程设计报告为每人一份，同一个小组的各成员的整体报告内容结构一致，但个人侧重点不同，个人着重撰写自己的工作内容，其他人的内容只要体现文档结构的完整性即可。

课程设计考核将综合考虑学生考勤和参与度、团队协作能力，过程管理能力、成果达成情况等。

**2.过程要求**

整个过程要求通过github来进行过程化发布，即阶段性地在github上提交过程结果。

**3.评分标准**

|  |  |
| --- | --- |
| **评分依据** | **评分成绩** |
| 1．团队协作能力 | 25分 |
| 2．python综合运用能力 | 25分 |
| 3．态度认真、刻苦钻研、创新能力 | 10分 |
| 4．过程完成、对工具的使用、对github的运用 | 20分 |
| 5．课程设计答辩逻辑清晰，内容正确 | 10分 |
| 6. 课程设计期间的课堂考勤、遵守纪律 | 10分 |
| 总分 | 100分 |

注：按上述六项分别记分后求和，根据小组成员贡献率综合评定，记载个人最后成绩。

成绩等级：优（90分—100分）、良（80分—89分）、中（70分—79分）、及格（60分—69分）、60分以下为不及格。

**五、指导地点与时间**

本课程设计将安排在第20-21周，采用腾讯会议和QQ群的方式，以在线形式进行。具体安排如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 周次 | 星期一 | 星期二 | 星期三 | 星期四 | 星期五 |
| 第20周 | 第5-8节 | 第5-8节 |  | 第5-8节 |  |
| 第21周 |  | 第5-8节 |  | 第5-8节 |  |

执笔:徐 勇 日期：2021-1-8

审阅:钱小红 日期：2021-1-8

目录

[1 绪论 1](#_Toc62247358)

[1.1 背景 1](#_Toc62247359)

[1.2 结构 1](#_Toc62247360)

[1.3 小组分工 1](#_Toc62247361)

[2 开发技术 1](#_Toc62247362)

[2.1 GitHub简介 1](#_Toc62247363)

[2.2 Jupyter notebook简介 2](#_Toc62247364)

[2.3 Numpy简介 2](#_Toc62247365)

[2.3 Pandas简介 2](#_Toc62247366)

[2.4 Python语言简介 2](#_Toc62247367)

[3 数据的收集与整理 3](#_Toc62247368)

[3.1 数据的来源 3](#_Toc62247369)

[3.2数据的读取 3](#_Toc62247370)

[3.3 数据清理 3](#_Toc62247371)

[3.3.1 数据去重 4](#_Toc62247372)

[3.3.2 各字段数据清洗 4](#_Toc62247373)

[4 问题与问题分析 6](#_Toc62247374)

[4.1 问题 6](#_Toc62247375)

[4.2 问题分析 6](#_Toc62247376)

[4.2.1 哪些城市/省份旅游选择最多 6](#_Toc62247377)

[4.2.2 热门城市/景点情况统计 7](#_Toc62247378)

[4.2.3 各省份好评热门景点 8](#_Toc62247379)

[4.2.4 各省份热门但评分不高景点 9](#_Toc62247380)

[4.2.5 优惠力度最大景点 9](#_Toc62247381)

[4.2.6 各级别景区门票价格 10](#_Toc62247382)

[5.总结 10](#_Toc62247383)

[6.参考文献 11](#_Toc62247384)

### 1 绪论

### 1.1 背景

近几年来，随着社会的发展和进步，人们的生活也得到了很大的改善，闲暇之余就会有很多人会选择外出放松一下。不过平日里大家都在为了生计而忙碌不已，只有在节假日的时候才能有时间和朋友一起出去玩，所以每到节假日的时候我国的众多景点就会出现这样的一幕，没有放假的时候景区几乎杳无人烟，但是一到假期的时候无人问津的地方都挤满了人。

中国游客在过去10年如何发生变化——从旗子跟随者到自由自在的行者中国的出境游过去10年发生巨变，2010年中央王国出境游仅5700万人次，今年将达1.8亿。10年增加的数字比日本总人口还多。但正如中国游客当初很快发现的那样，一个庞大的游客群并未受到目的地热烈欢迎——尤其当他们齐刷刷从一长串客车上下来时。

### 1.2 结构

本课程设计是对2019年各个城市/省份的五一节日长假人满为患的数据进行分析、统计分析热门景点、打折力度最大、哪些城市最受青睐、各级别景区的门票价格，观察这一年其最受青睐的城市，整理出来的数据，运用matplolib进行编程绘图更直观的描述出其变化，并对这数据进行重洗、清洁，删除异常点，进行绘图，并用matplolib进行分析。总的来说可分为四大部分：

首先，对问题进行重选，根据问题的要求对问题进行分析，找出相应的解决方法。

其次，根据对问题的分析，收集相关的数据并进行处理。

再次，建立相关的魔性，对模型进行分析，模型的准备，建立模型并求解模型。

最后，对模型进行总结。

### 1.3 小组分工

胡佳（组长）：主要负责代码中出现的问题分析以及GitHub软件的使用上传。

李文娟：主要负责后期数据分析的结果进行汇总出来，撰写PPT以及答辩。

薛文：主要负责代码前期的数据清洗，PPT前期制作以及报告。

总的来说，在整个代码、PPT三个人都一起参与其中，有出现的问题一起讨论解决。

### 2 开发技术

### 2.1 GitHub简介

GitHub是一个面向[开源](https://baike.baidu.com/item/%E5%BC%80%E6%BA%90/20720669" \t "_blank)及私有[软件](https://baike.baidu.com/item/%E8%BD%AF%E4%BB%B6/12053)项目的托管平台，因为只支持Git作为唯一的版本库格式进行托管，故名GitHub。

GitHub于2008年4月10日正式上线，除了[Git](https://baike.baidu.com/item/Git/12647237" \t "_blank)代码仓库托管及基本的Web管理界面以外，还提供了订阅、讨论组、文本渲染、在线文件编辑器、协作图谱（报表）、代码片段分享（Gist）等功能。目前，其注册用户已经超过350万，托管版本数量也是非常之多，其中不乏知名开源项目[Ruby](https://baike.baidu.com/item/Ruby/11419" \t "_blank) on Rails、[jQuery](https://baike.baidu.com/item/jQuery/5385065" \t "_blank)、[python](https://baike.baidu.com/item/python/407313" \t "_blank)等。

2018年6月4日，微软宣布，通过75亿美元的股票交易收购代码托管平台GitHub。

### 2.2 Jupyter notebook简介

Jupyter Notebook（此前被称为 IPython notebook）是一个交互式笔记本，支持运行 40 多种编程语言。

Jupyter Notebook 的本质是一个 Web[应用程序](https://baike.baidu.com/item/%20%E5%BA%94%E7%94%A8%E7%A8%8B%E5%BA%8F/5985445)，便于创建和共享文学化程序文档，支持实时代码，数学方程，可视化和 [markdown](https://baike.baidu.com/item/markdown/3245829)。 用途包括：数据清理和转换，数值模拟，统计建模，机器学习等等 。

### 2.3 Numpy简介

Numpy（Numerical Python）是Python的一种开源的数值计算扩展。这种工具可用来存储和处理大型矩阵，比Python自身的嵌套列表（nested list structure)结构要高效的多（该结构也可以用来表示矩阵（matrix）），支持大量的维度数组与矩阵运算，此外也针对数组运算提供大量的数学函数库 。

### 2.3 Pandas简介

pandas 是基于[NumPy](https://baike.baidu.com/item/NumPy/5678437" \t "_blank) 的一种[工具](https://baike.baidu.com/item/%E5%B7%A5%E5%85%B7/81891" \t "_blank)，该工具是为解决数据分析任务而创建的。Pandas 纳入了大量库和一些标准的[数据模型](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E6%A8%A1%E5%9E%8B/1305623" \t "_blank)，提供了高效地操作大型数据集所需的工具。pandas提供了大量能使我们快速便捷地处理数据的函数和方法。你很快就会发现，它是使Python成为强大而高效的数据分析环境的重要因素之一。

### 2.4 Python语言简介

Python由荷兰数学和计算机科学研究学会的Guido van Rossum 于1990 年代初设计，作为一门叫做[ABC语言](https://baike.baidu.com/item/ABC%E8%AF%AD%E8%A8%80/334996" \t "_blank)的替代品。Python提供了高效的高级[数据结构](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E7%BB%93%E6%9E%84/1450" \t "_blank)，还能简单有效地[面向对象](https://baike.baidu.com/item/%E9%9D%A2%E5%90%91%E5%AF%B9%E8%B1%A1/2262089)编程。Python语法和动态类型，以及[解释型语言](https://baike.baidu.com/item/%E8%A7%A3%E9%87%8A%E5%9E%8B%E8%AF%AD%E8%A8%80/8888952" \t "_blank)的本质，使它成为多数平台上写脚本和快速开发应用的编程语言，随着版本的不断更新和语言新功能的添加，逐渐被用于独立的、大型项目的开发。

Python[解释器](https://baike.baidu.com/item/%E8%A7%A3%E9%87%8A%E5%99%A8/10418965)易于扩展，可以使用C或[C++](https://baike.baidu.com/item/C%2B%2B/99272" \t "_blank)（或者其他可以通过C调用的语言）扩展新的功能和[数据类型](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E7%B1%BB%E5%9E%8B/10997964" \t "_blank)。 Python 也可用于可定制化软件中的扩展程序语言。Python丰富的标准库，提供了适用于各个主要系统平台的源码或[机器码](https://baike.baidu.com/item/%E6%9C%BA%E5%99%A8%E7%A0%81/86125)。

### 3 数据的收集与整理

### 3.1 数据的来源

主要按全国各个城市抓取飞猪“景点门票”栏的景点门票销售数据，并且分析五一哪些景点会人挤人，哪些景点值得一去。

主要的文件为：

1.city\_data.csv: 全国城市及所属省份列表。

2.get\_data.py: 数据爬取代码。

3.laborday.ipynb:Jupyter notebook代码，对景点门票数据进行分析。

飞猪景点门票销售数据5万条下载链接：

<https://pan.baidu.com/s/11ydiAP9tibcsEJvoVpEAyw>

### 3.2数据的读取

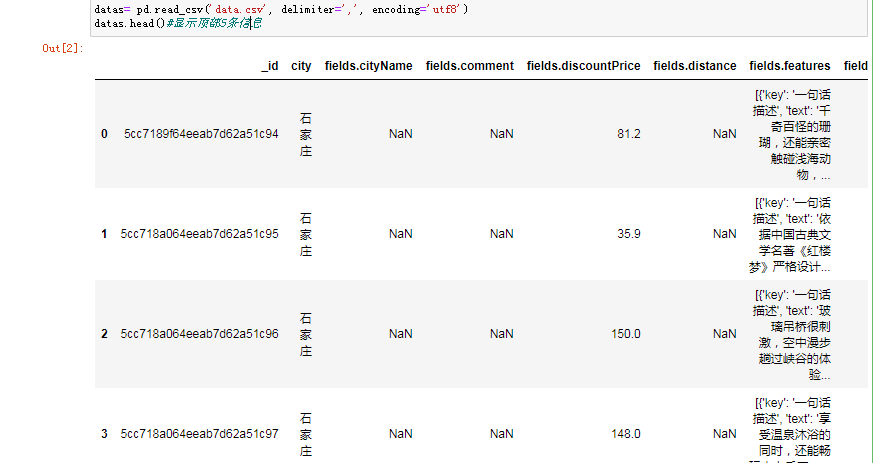
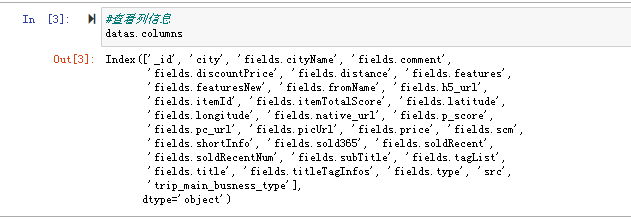
读取data.csv文件中的数据，以逗号作为分隔符来添加字段名，如下图所示：

图 3.2 读取数据图

### 3.3 数据清理

提取信息并进行去重、规范化处理，如下图所示：



### 3.3.1 数据去重

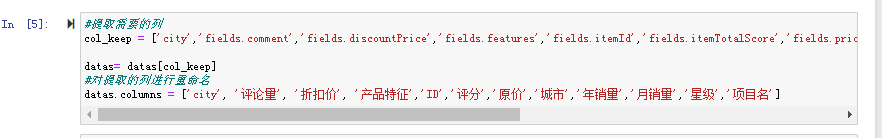
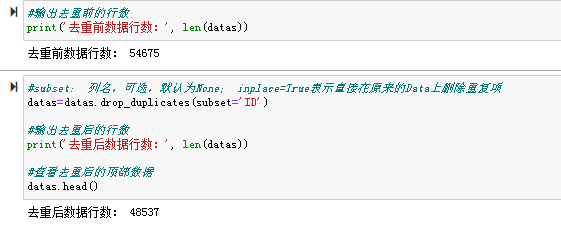


图 3.3.1 数据去重图

图 3.3 提取所需字段图

### 3.3.2 各字段数据清洗

以“产品特征”为关键字的数据清洗，如下图所示：

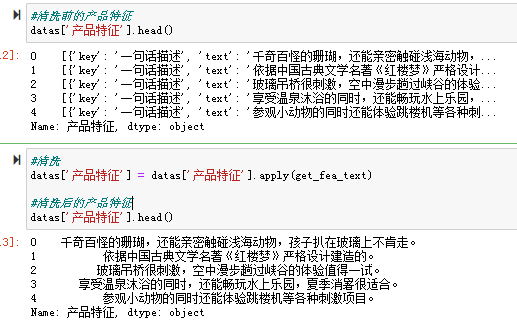


图 3.3.2 “产品特征”为关键字数据清洗图

以“星级”为关键字的数据清洗，如下图所示：

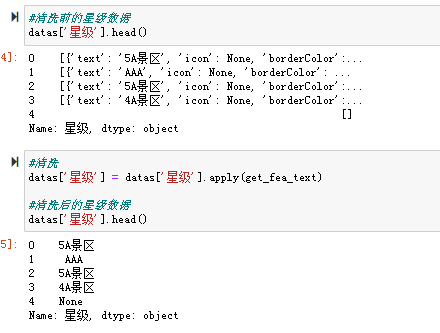


图 3.3.2 “星级”为关键字数据清洗图

以“年销量”为关键字的数据清洗，如下图所示：



图 3.3.2 “年销量”为关键字数据清洗图

### 4 问题与问题分析

### 4.1 问题

1. 哪些城市/省份的旅游选择最多？
2. 热门城市/景点的情况统计
3. 各省份的好评热门景点
4. 很热门但是评分不好的景点
5. 优惠力度最大的景点
6. 各级别景区的门票价格的数据分析

### 4.2 问题分析

### 4.2.1 哪些城市/省份旅游选择最多

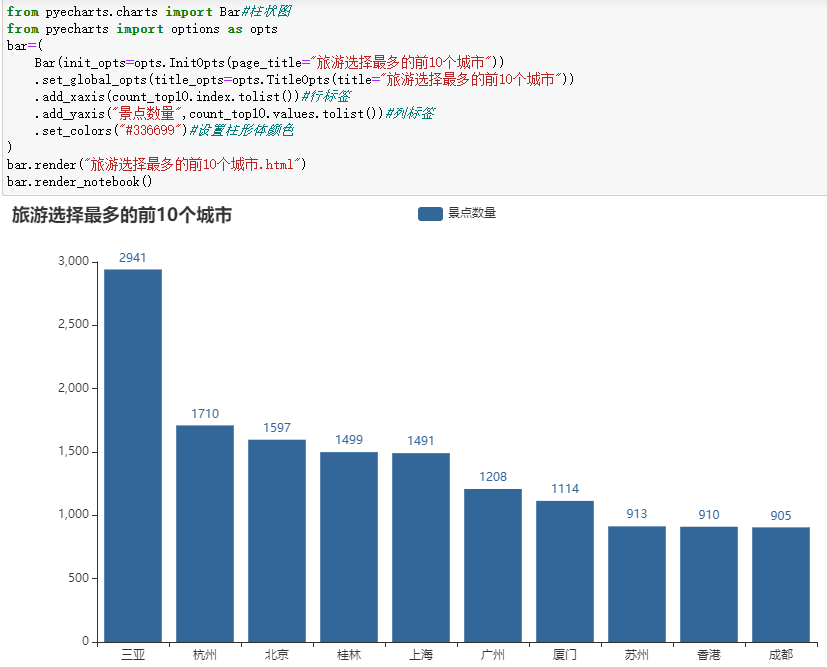
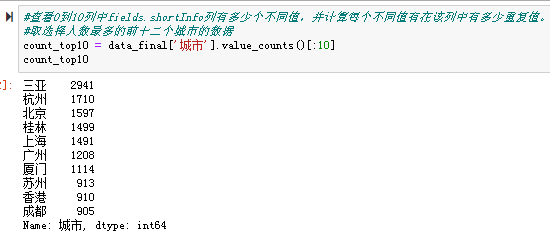


图 4.2.1 对城市可选择景点统计图

分析：数据分析显示，三亚的景点数量是最多的，远超其后的杭州，北京等。通过浏览资料，我们可以发现三亚在地理位置，气候风光等方面有着得天独厚的优势，其景区整体销售量也处于较高的位置。



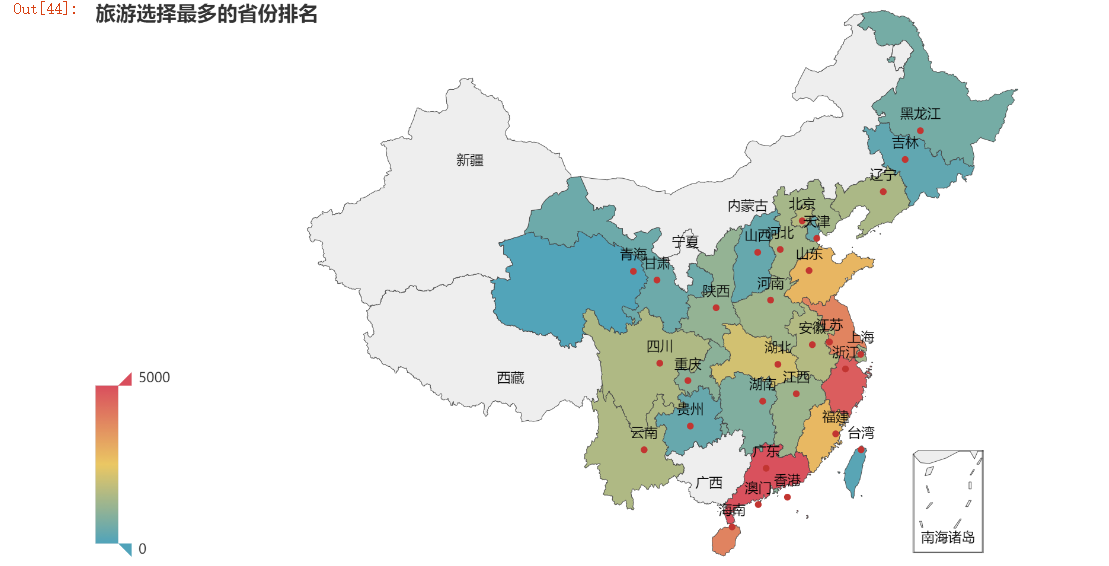


图 4.2.1 对省份可选择景点统计图

分析：根据数据分析结果，我们可以发现广东省的景点数量最多，随着其后的是浙江上海等，最少的是青海省。浏览相关资料，可以发现其地理位置与香港，澳门接壤，与海南隔海相望。其独特的地理位置为其旅游行业的发展提供了良好的条件。

### 4.2.2 热门城市/景点情况统计

图 4.2.2 对热门城市统计图

分析：根据数据结果，我们可以看出杭州是月售最多的城市，其次是上海北京等地。分析原因，我们不难发现，图中所示的热门城市皆为经济发展较为良好的地区，可见其城市的经济发展情况与其旅游业的发展也是密切相关的。



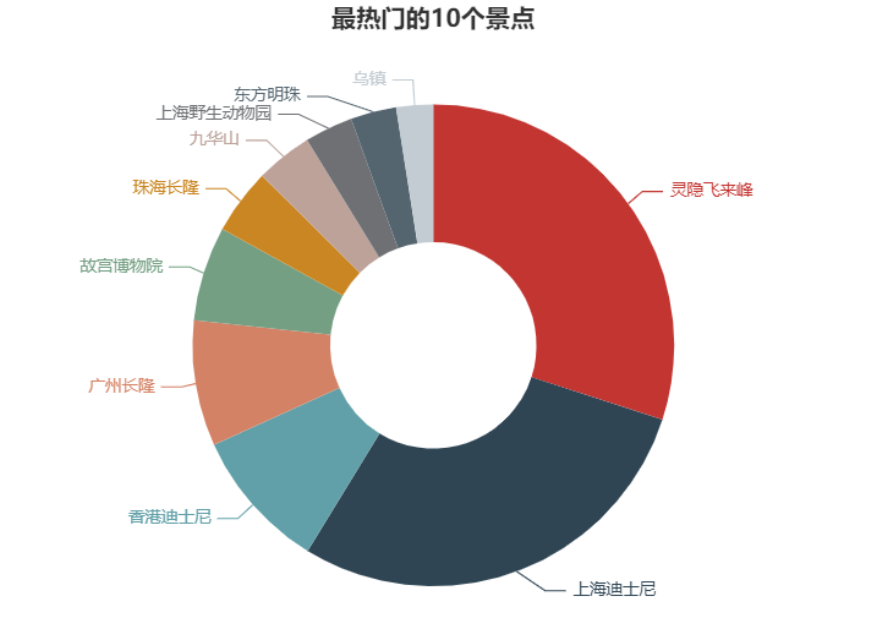


图 4.2.2 对热门景点统计图

分析：根据数据分析结果，我们不难看出灵隐飞来峰是最多人心仪的目标景点，其次为上海迪士尼度假区与故宫博物院。浏览相关资料，我们可以发现在热门景点中，普遍存在多样性的游乐项目及较多的互动性强的设施。查阅资料，我们发现尤其上海迪士尼等（具有较多的全国首发游乐项目）景点，由此可以得出结论：景点较强的创新性也是吸引游客的重要因素之一。

### 4.2.3 各省份好评热门景点

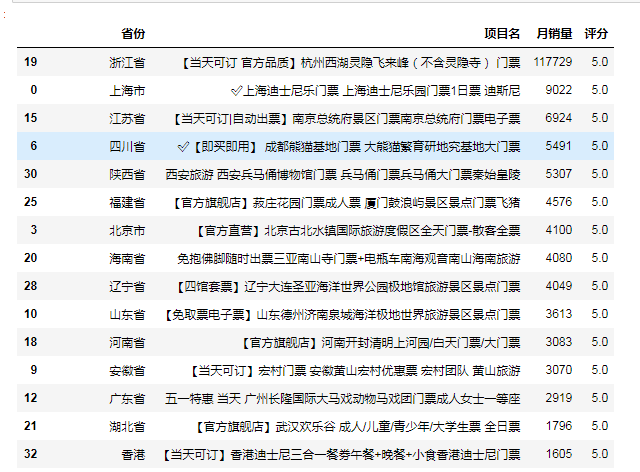


图 4.2.3 各省份好评热门景点图

分析：根据分析结果所示，在各省份好评热门景点的数据中，其好评数最多的是杭州西湖灵隐飞来峰，其次是上海迪士尼乐园门票。

其结果显示，在一定程度上怀旧也对游客具有较高的吸引力。也印证了上述游客对创新性较高的景点较为感兴趣这一结论。

### 4.2.4 各省份热门但评分不高景点

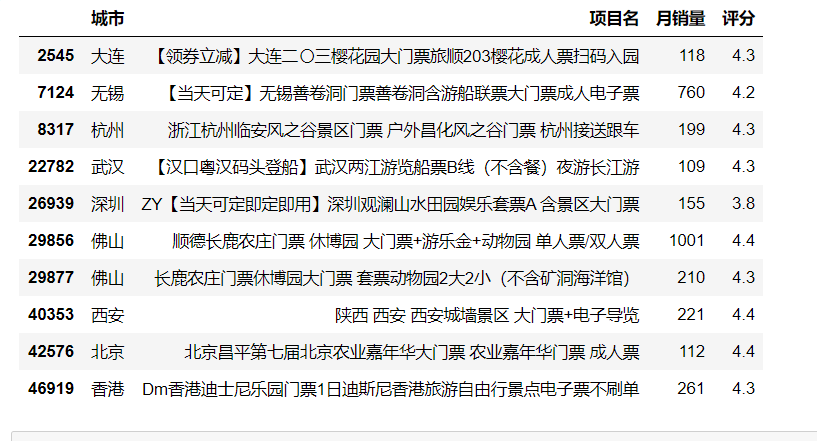


图 4.2.4 热门但评分不高的景点图

分析：结果显示，在总数据中，存在左侧十条数据月销量在100以上，景区评分小于4.5分。结合网上资料分析数据，此类景点存在部分单一性，如其中观赏类景点互动性较差，园馆类景点存在时间限制，地理位置等原因，致使其评分略低。若部分游客对热门景点较为好奇，建议参考第三条好评热门景点或仔细浏览各景点相关信息，避免踩雷。

### 4.2.5 优惠力度最大景点

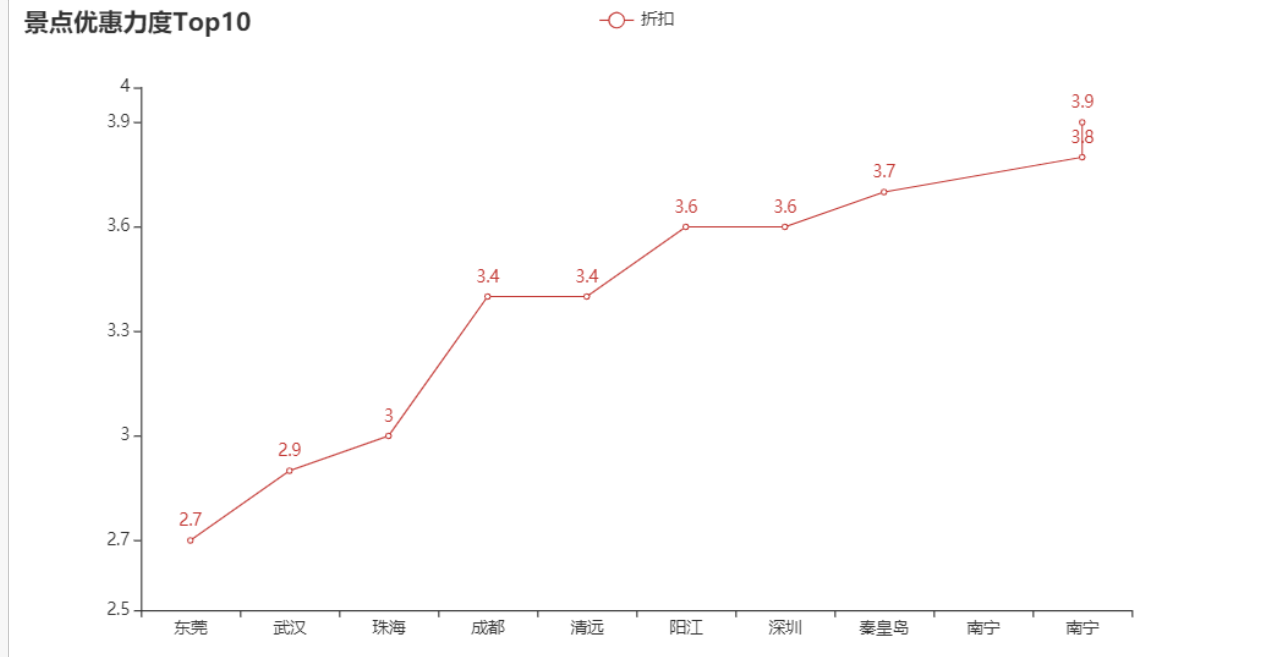


图 4.2.5 优惠力度最大景点图

分析：在总数据中，存在上图所示等部分景点存在较大的优惠力度。结合相关资料分析数据可以发现，部分景点由于天气或旅游淡季或一些突发状况等现象，对不同时段的售票价格及优惠力度采取不同形式。此处建议游客提前浏览目标景点及其城市的相关信息，在能确保拥有良好的旅行体验的情况下，选择目标城市。

### 4.2.6 各级别景区门票价格

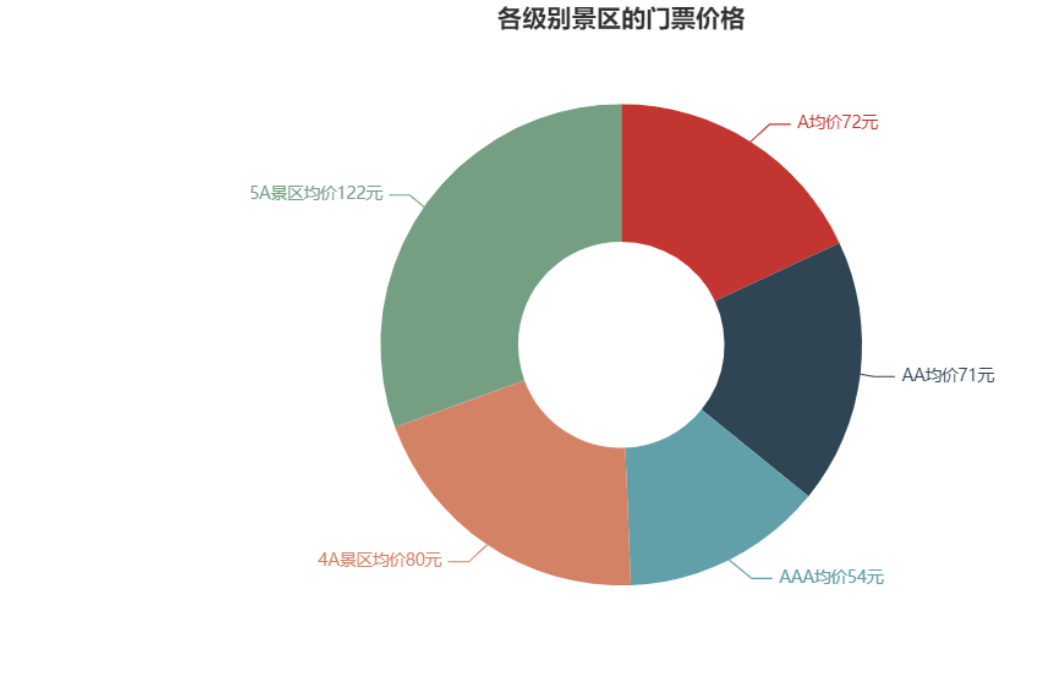
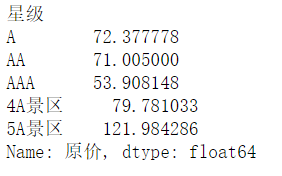


图 4.2.6 各级别景区门票价格图

分析：在总数据中，各级别景区的平均售价分别如下所示。结合相关资料分析数据可以发现，由于历史，人文情怀，景区地理位置，互动性等方面，5A景区确实可能拥有较好的旅行体验。但是部分宣传力度较弱的4A，3A景区确实存在其性价比高于5A景区的现象。此处建议游客在选择目标景点时，可参考同等力度的景区性价比，不一定非要执着于5星级。

### 5.总结

通过选择比较，部分旅客希望能在较短的时间内游玩到更多的景点。对于各城市及省份的景区选择数量，我们分别进行了统计。根据上面三亚为选择最多的城市，杭州其次。在省份中，澳门，香港，广东等选择最多。此处建议游客选择目标景区时，可考虑上述城市或省份。注意，疫情期间，减少出国，关注疫情，避免前往高风险地区。

通过热点比较，武汉，澳门，海口等城市较为热门，灵隐飞来峰，上海迪士尼，香港迪士尼等景区较为热门。游客可根据其景区整体销售量，参考评论及其景区互动性等特点决定具体目标景点。

通过价格比较，各星级平均票价与城市优惠力度如上图所示，总数据中3A景区平均售价最为划算，东莞，武汉等地区景点优惠力度最大。确定目标景区时，游客可参考具体景点及整体售价特点决定目标景点。前往景区或人流量较大的场所时，请佩戴做好防护，注意消毒。

通过这次的课程设计，我们深刻的了解到GitHub软件功能的使用以及jupyter notebook的使用，并对书本再次的回顾学习，也认识到自己学到的东西还是非常浅显的离不开课本，我们要多思考、多应用。我们也有收获，增强了分析问题与小组合作交流的能力，到了下午我们小组经常开会议一起讨论做分析，不得不说团结就是力量这句话没有错，一起完成了这次课设。

### 6.参考文献

1. 刘宇宙.python3.5从零开始学【M】. 北京：清华大学出版社，2017.

2. 董付国.python可以这样学【M】. 北京：清华大学出版社，2017.

3. 刘凌霞，吴海涛.21天学通python【M】.北京：电子工业出版社，2016.

4． https://zhuanlan.zhihu.com/p/196535136

|  |  |  |  |
| --- | --- | --- | --- |
| 课程设计题目 | 数据分析的学习与实践 | | |
| 课程设计学生答辩或质疑记录：  1.数据清洗的目的是什么？  答：数据清洗的目的有两个，第一是通过清洗让数据可用。第二是让数据变的更适合进行后续的分析工作。换句话说就是有”脏”数据要洗，干净的数据也要洗。  2.怎么进行数据清洗的？  答：数据录入过程、数据整合过程都可能会产生重复数据，直接删除是重复数据处理的主要办法。Pandas提供查看、处理重复数据的方法drop\_duplicates。  3.怎么读取数据信息？  答：在Python中pandas模块是专门用来数据分析的工具，通过原始文件（utf-8格式）用read\_csv函数读写，将原始数据另存为utf-8格式的数据，重新读入到TXT数据中，CSV文本文件是非常常用的一种数据存储格式。 | | | |
| **评 分 依 据** | | **分 值** | **评分成绩** |
| 1．团队协作能力 | | 25分 |  |
| 2．python综合运用能力 | | 25分 |  |
| 3．态度认真、刻苦钻研、创新能力 | | 10分 |  |
| 4．过程完成、对工具的使用、对github的运用 | | 20分 |  |
| 5．课程设计答辩逻辑清晰，内容正确 | | 10分 |  |
| 6. 课程设计期间的课堂考勤、遵守纪律 | | 10分 |  |
| 总 分 | | 100分 |  |
| 最终评定等级为：  指导老师签字：  2021 年 1 月 15日 | | | |

**课程设计成绩评定表**