|  |  |
| --- | --- |
| 学号 | 10212818105 |



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 | 龚成龙 |
| 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、遵守纪律 | 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、创新能力 | 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](#_Toc62226850)

[1.1 背景 1](#_Toc62226851)

[1.2 小组成员及任务划分 1](#_Toc62226852)

[2 相关技术简介 1](#_Toc62226853)

[2.1 MySQL简介 1](#_Toc62226854)

[2.1.1 MySQL优缺点 1](#_Toc62226855)

[2.1.2MySQL的优化 2](#_Toc62226856)

[2.2 jupyter notebook简介 2](#_Toc62226857)

[2.2.1 jupyter notebook的主要特点 2](#_Toc62226858)

[2.3 pandas简介 2](#_Toc62226859)

[2.3.1 pandas数据结构 3](#_Toc62226860)

[2.4 numpy简介 3](#_Toc62226861)

[2.4.1 numpy数据类型 3](#_Toc62226862)

[2.5 GitHub简介 4](#_Toc62226863)

[3 数据处理 5](#_Toc62226864)

[3.1把需要的字段挑出来 5](#_Toc62226865)

[3.2去重 5](#_Toc62226866)

[3.3数据类型的转换 5](#_Toc62226867)

[3.4各字段清洗 5](#_Toc62226868)

[3.4.1 fields.features字段的清洗 5](#_Toc62226869)

[3.4.2 字段fields.sold365的清洗 6](#_Toc62226870)

[4 分析方向与问题分析 7](#_Toc62226871)

[4.1 分析方向 7](#_Toc62226872)

[4.2 问题分析 7](#_Toc62226873)

[4.2.1哪些城市的旅游选择最多？ 7](#_Toc62226874)

[4.2.2哪些省份的旅游选择最多？ 7](#_Toc62226875)

[4.2.3哪些城市最受游客青睐？ 8](#_Toc62226876)

[4.2.4最热门的景点？ 9](#_Toc62226877)

[4.2.5各省份的好评热门景点？ 10](#_Toc62226878)

[4.2.6很热门但是评分不好景点？ 12](#_Toc62226879)

[4.2.7打折力度最大的景点？ 13](#_Toc62226880)

[4.2.8各级别景区的门票？ 14](#_Toc62226881)

[5.总结 14](#_Toc62226882)

[6.参考文献 15](#_Toc62226883)

# 1 绪论

## 1.1 背景

* 随着国内经济的发展，人民在追求物质生活的基础上增加了对精神文化生活的追求。
* 科学技术的进步使移动端用户在线订票和选择旅游地变得更加方便，国内旅游市场稳步上升。
* 在当今社会下旅游成为大多数人们缓解工作、生活的压力的一种很普遍的方式。
* 旅游地及景点的多样化使人们多了很多选择。

## 1.2 小组成员及任务划分

曾逸雯（负责数据清洗、PPT制作）

龚成龙（GitHub管理、代码整合）

伍泽蒙（数据收集、报告模板制作）

代码由3人共同参与分析完成

# 2 相关技术简介

## 2.1 MySQL简介

MySQL 是一款安全、跨平台、高效的，并与 PHP、Java 等主流编程语言紧密结合的数据库系统。该数据库系统是由瑞典的 MySQL AB 公司开发、发布并支持，由 MySQL 的初始开发人员 David Axmark 和 Michael Monty Widenius 于 1995 年建立的。MySQL 的象征符号是一只名为 Sakila 的海豚，代表着 MySQL 数据库的速度、能力、精确和优秀本质。

### 2.1.1 MySQL优缺点

优点：MySQL数据库是一个复杂程度较低且高性能的数据库系统，它具有丰富的数据类型，可以存储多种数据类型的数据，按照对应的数据类型存储数据，从而可以节约数据的存储空间。MySQL的核心是完全多线程，同时支持多处理器，在很大程度上提高了MySQL数据库处理数据的速度和效率，没有内存漏洞。MySQL数据库可以控制哪些人可以看到自己的数据哪些人不能看到具有很高的安全性。

缺点：MySQL数据库不支持系统在运作正常状态下备份数据，容易让数据丢失。MySQL没有存储过程的一种语言，这在很大程度上了限制了企业级数据库操作员的操作。Linux系统上的MySQL数据库若是由第三方安装的，那么第三方必须支付许可费方能给用户安装。

### 2.1.2MySQL的优化

MySQL数据库的优化可以通过以下八个方面来进行操作，分别是创建索引，复合索引，使用较短的索引，索引不会包含有空值的列，like语句的操作，不在列上进行运算，排序的索引，不使用NOT IN操作。创建索引的时候可以让我们快速找到所满足条件的数据，但并不是所有的情况都需要创建索引，避免过度索引的情况，因此，创建索引的时候需要考虑添加的索引是否有效，以免出现过度索引的状况。为了提高数据库查询数据的速度，可以利用短索引优化数据库。

## 2.2 jupyter notebook简介

Jupyter Notebook是基于网页的用于交互计算的应用程序。其可被应用于全过程计算：开发、文档编写、运行代码和展示结果。简而言之，Jupyter Notebook是以网页的形式打开，可以在网页页面中直接编写代码和运行代码，代码的运行结果也会直接在代码块下显示。如在编程过程中需要编写说明文档，可在同一个页面中直接编写，便于作及时的说明和解释。

### 2.2.1 jupyter notebook的主要特点

 编程时具有**语法高亮**、*缩进*、*tab补全*的功能。

 可直接通过浏览器运行代码，同时在代码块下方展示运行结果。

 以富媒体格式展示计算结果。富媒体格式包括：HTML，LaTeX，PNG，SVG等。

 对代码编写说明文档或语句时，支持Markdown语法。

 支持使用LaTeX编写数学性说明。

## 2.3 pandas简介

Pandas是python的一个数据分析包，最初由AQR Capital Management于2008年4月开发，并于2009年底开源出来，目前由专注于Python数据包开发的PyData开发team继续开发和维护，属于PyData项目的一部分。Pandas最初被作为金融数据分析工具而开发出来，因此，pandas为时间序列分析提供了很好的支持。 Pandas的名称来自于面板数据（panel data）和python数据分析（data analysis）。panel data是经济学中关于多维数据集的一个术语，在Pandas中也提供了panel的数据类型。

### 2.3.1 pandas数据结构

Series：一维数组，与Numpy中的一维array类似。二者与Python基本的数据结构List也很相近。Series如今能保存不同种数据类型，字符串、boolean值、数字等都能保存在Series中。

Time- Series：以时间为索引的Series。

DataFrame：二维的表格型数据结构。很多功能与R中的data.frame类似。可以将DataFrame理解为Series的容器。

Panel ：三维的数组，可以理解为DataFrame的容器。

Panel4D：是像Panel一样的4维数据容器。

PanelND：拥有factory集合，可以创建像Panel4D一样N维命名容器的模块。

## 2.4 numpy简介

一个用python实现的科学计算，包括：1、一个强大的N维数组对象Array；2、比较成熟的（广播）函数库；3、用于整合C/C++和Fortran代码的工具包；4、实用的线性代数、傅里叶变换和随机数生成函数。numpy和稀疏矩阵运算包scipy配合使用更加方便。

NumPy（Numeric Python）提供了许多高级的数值编程工具，如：矩阵数据类型、矢量处理，以及精密的运算库。专为进行严格的数字处理而产生。多为很多大型金融公司使用，以及核心的科学计算组织如：Lawrence Livermore，NASA用其处理一些本来使用C++，Fortran或Matlab等所做的任务。

NumPy 的前身为 Numeric ，最早由 Jim Hugunin 与其它协作者共同开发，2005 年，Travis Oliphant 在 Numeric 中结合了另一个同性质的程序库 Numarray 的特色，并加入了其它扩展而开发了 NumPy。NumPy 为开放源代码并且由许多协作者共同维护开发。

### 2.4.1 numpy数据类型

numpy 支持的数据类型比 Python 内置的类型要多很多，基本上可以和 C 语言的数据类型对应上，其中部分类型对应为 Python 内置的类型。如图列举了常用 NumPy 基本类型。



## 2.5 GitHub简介

GitHub是一个面向开源及私有软件项目的托管平台，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 on Rails、jQuery、python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。

# 3 数据处理

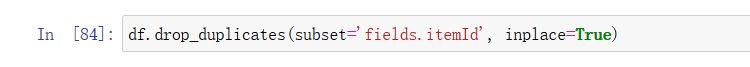
## 3.1把需要的字段挑出来

如图先定义一个列表存储自己需要的字段，然后直接筛选出需要的字段。



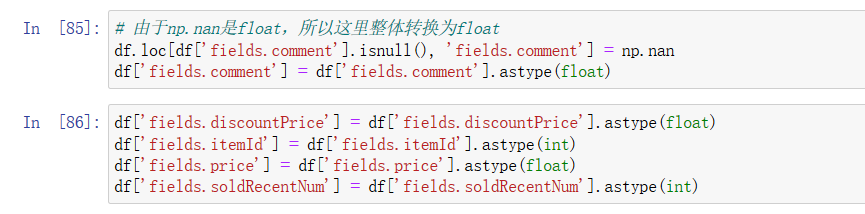
## 3.2去重

如图这里通过pandas.DataFrame.drop\_duplicates函数实现将字段”fields.itemId”重复的去除。



## 3.3数据类型的转换

如图先筛选出字段”fields.comment”为np.nan的数据，然后将此字段转换为float类型，然后将字段”fields.discountPrice”和字段”fields.price”转换为float类型”，将“fields.itemId”和”fields.soldRecentNum”转换为int类型



## 3.4各字段清洗

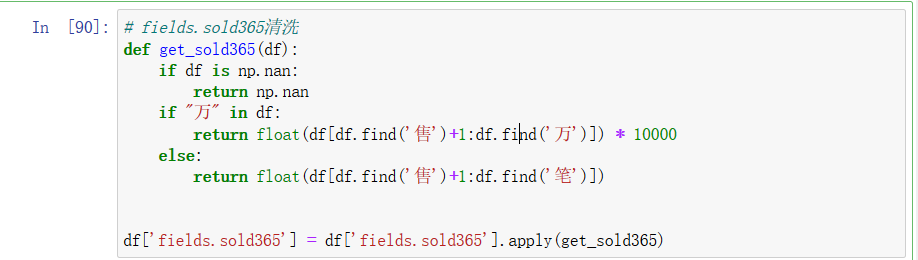
### 3.4.1 fields.features字段的清洗

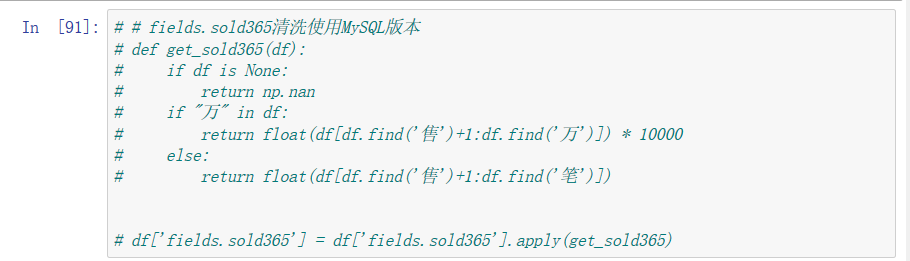
如图定义函数get\_fea\_text（）函数来实现字段fields.features的清洗，函数主要将传过来的列表数据先判断是否为空，为空直接返回None，否则转换为字典类型然后得到一个字典类型的数组，然后如果字典数组有值我们就将数组第一个数据的key为text的value值返回，否则返回None。



### 3.4.2 字段fields.sold365的清洗

如图，首先这个地方采用了两种方案，主要区别在于使用MySQL后数据为空存储的为None，而pandas中的空为np.nan，原理都是先判断传入数据是否为空，如果为空则返回np.nan，否则先判断数据是否包含字符”万”，如果包含则将字符”售”和字符”万”之间的数据截取出来然后乘以10000返回，如果不包含字符”万”的话将字符”售”和字符”笔”之间的数据截取出来返回即可。





# 4 分析方向与问题分析

## 4.1 分析方向

1. 哪些城市/省份的旅游选择最多？
2. 哪些城市最受游客青睐？
3. 最热门的景点
4. 各省份的好评热门景点
5. 很热门但是评分不好的景点
6. 打折力度最大的景点
7. 各级别景区的门票价格

## 4.2 问题分析

### 4.2.1哪些城市的旅游选择最多？

通过数据分析筛选出了旅游选择最多的前12个城市，其中三亚遥遥领先于其他11个城市，其次是杭州、北京、桂林、上海等自然风景，人文风景著名的地点。在其次就是广州、厦门、苏州、香港、成都、重庆、清远这类南方城市。如图2.1所示

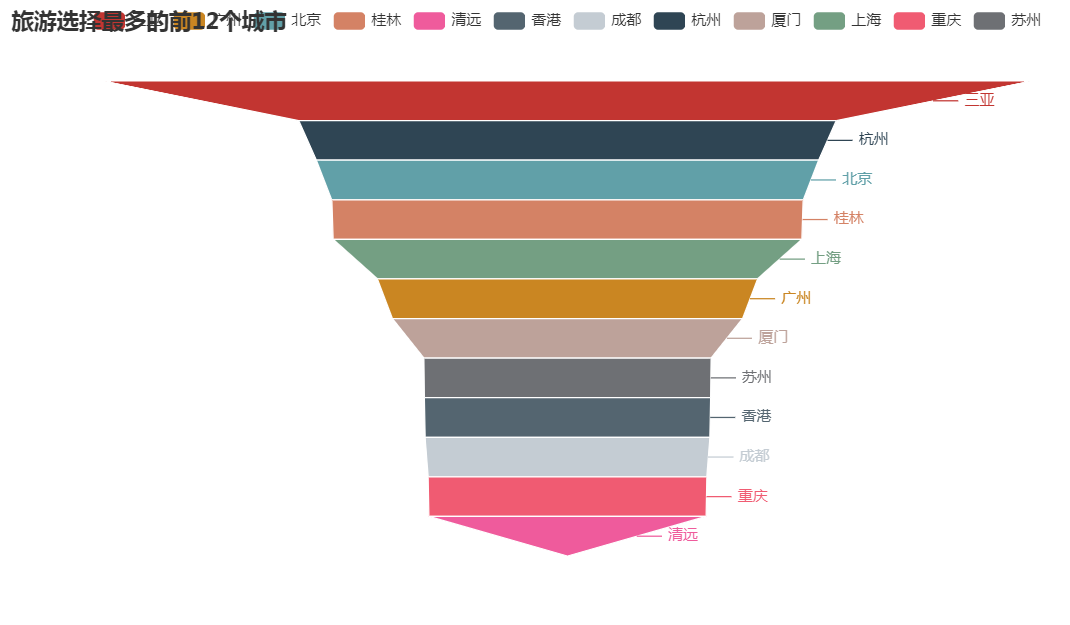


图2.1 旅游选择最多的前12个城市

### 4.2.2哪些省份的旅游选择最多？

以省份为单位，计算了每个省份在飞猪上能搜索到的的景点门票项数，发现旅游选择最多的省份主要集中在沿海，包括广东、香港、澳门、海南、浙江、江苏等。内陆地区的省份旅游选择较少。西北地区的旅游选择则更少。如图2.2所示

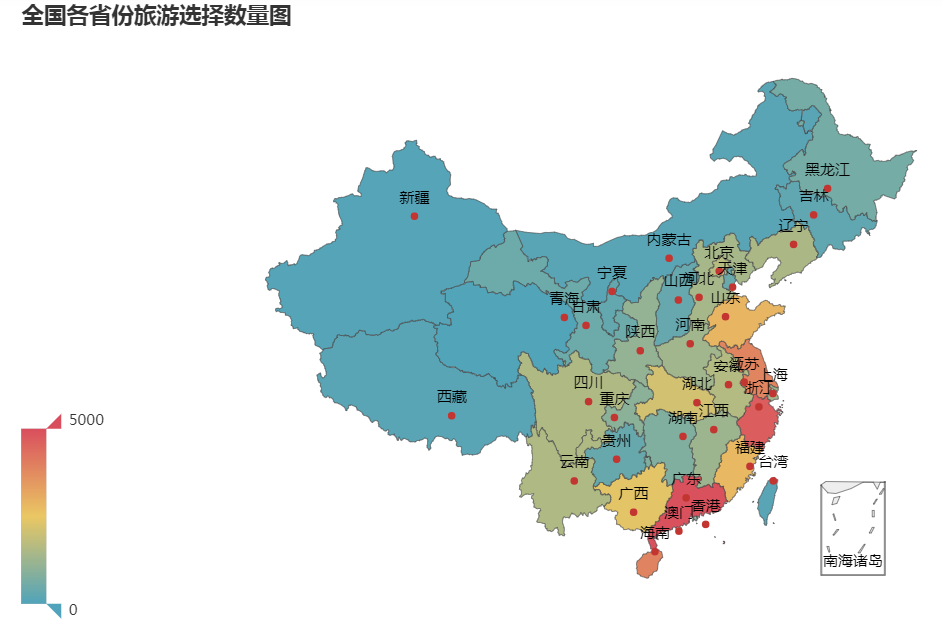


图2.2 全国各省份旅游选择数量图

### 4.2.3哪些城市最受游客青睐？

游客们对武汉、澳门、海口、杭州、常州、苏州等城市旅游景点的评价很高。通过结合”选择旅游最多的城市”及”评分最高的城市”发现杭州、澳门、苏州等地在游客的选择和评价来说都是较高的。如图2.3所示



图2.3 最近一个月售出门票Top10城市云图

### 4.2.4最热门的景点？

通过城市最近一个月售出门票数量和城市景点的平均分这两个指标来确定可见杭州和上海遥遥领先。从而推测出游客可能更加偏爱杭州和上海的风景和旅游项目。如图2.4所示。

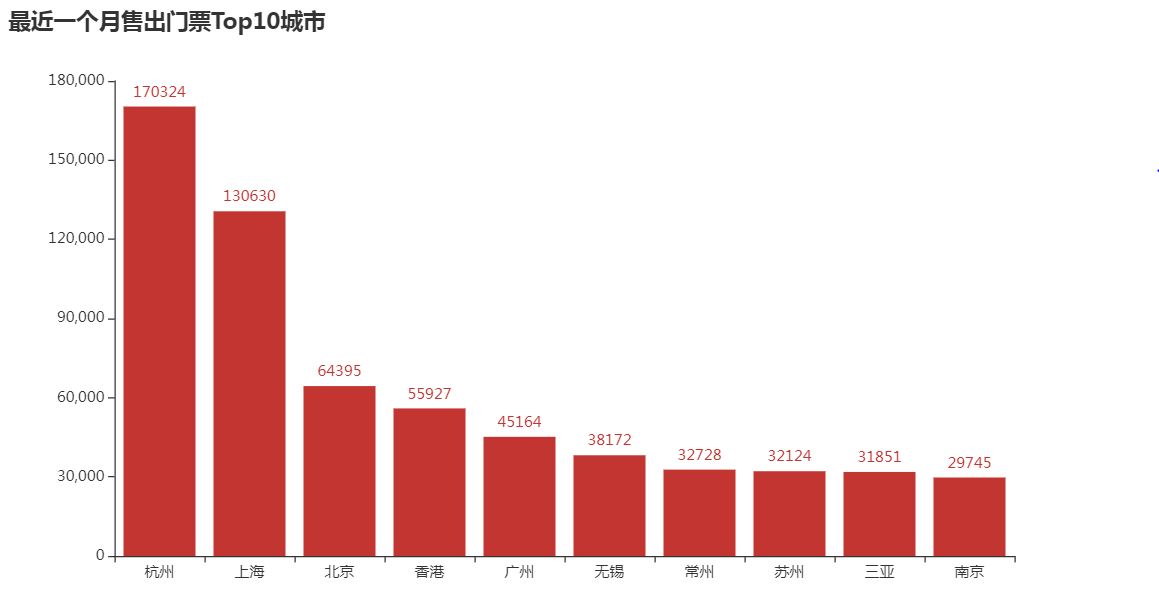


图2.4 最近一个月售出门票Top10城市

从卖出的门票数据可以看到，游客大多选择来杭州、上海旅游的原因是杭州的灵隐飞来峰和上海迪士尼。另外还有香港迪士尼、广州长隆、故宫博物院、珠海长隆、九华山、上海野生动物园、东方明珠、乌镇等都备受游客青睐。如图2.5所示

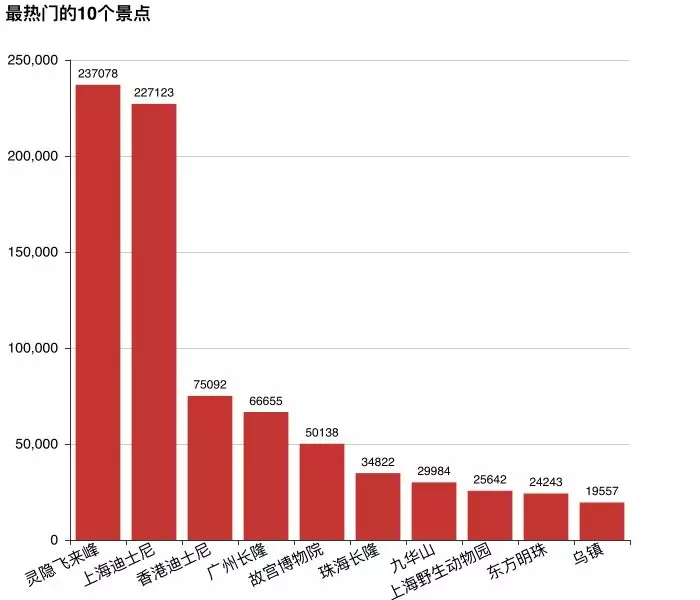


图2.5 最热门的10个景点

### 4.2.5各省份的好评热门景点？

把这些景点门票数据以省级行政区为单位，筛选出了每个省级行政区最受好评且销量最高的景点。如图2.6，2.7所示：



图2.6 最热门的景点（省份为单位）



图2.7 最热门的景点（省份为单位）

### 4.2.6很热门但是评分不好景点？

在很多景点中虽然热门但是去过的游客对这些地点的评价却不是很高，我们筛选出了销量大于100且评分小于4.5的地点，其中深圳的”观澜山水田园娱乐套票”评分最低只有3.8分。希望可以给出行旅游的人提个醒。如图2.8所示。



图2.8 很热门但是评分不好的景点

### 4.2.7打折力度最大的景点？

通过景点门票打折状况分布图可以看出大多数景点的折扣都在8折到全价之间，也有少部分景点门票在3折左右。如图2.9所示。

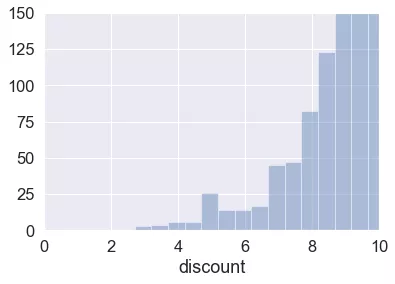


图2.9 打折力度最大的景点

门票折扣价最低的是2.7折，是位于广东 东莞的“南社明清古村落”，折后仅需13.5元就够购买门票。优惠力度最大的是珠海的“海泉湾海洋温泉”，打折前为1980.0元，折后为589.0元，推荐有意向去的旅客可以考虑购买。如图2.10所示。



图2.10景点详细打折情况

### 4.2.8各级别景区的门票？

级别越高，景区门票的价格就越高。不过，A级和AA级的平均价格差不多，都在71元左右。AAA级景区的门票平均价格最低，4A景区门票均价在80元，5A景区门票均价在122元，比4A的价格高出了42元。最高的5A级景区门票比最低的3A级景区门票高出了整整68元。如图2.11所示

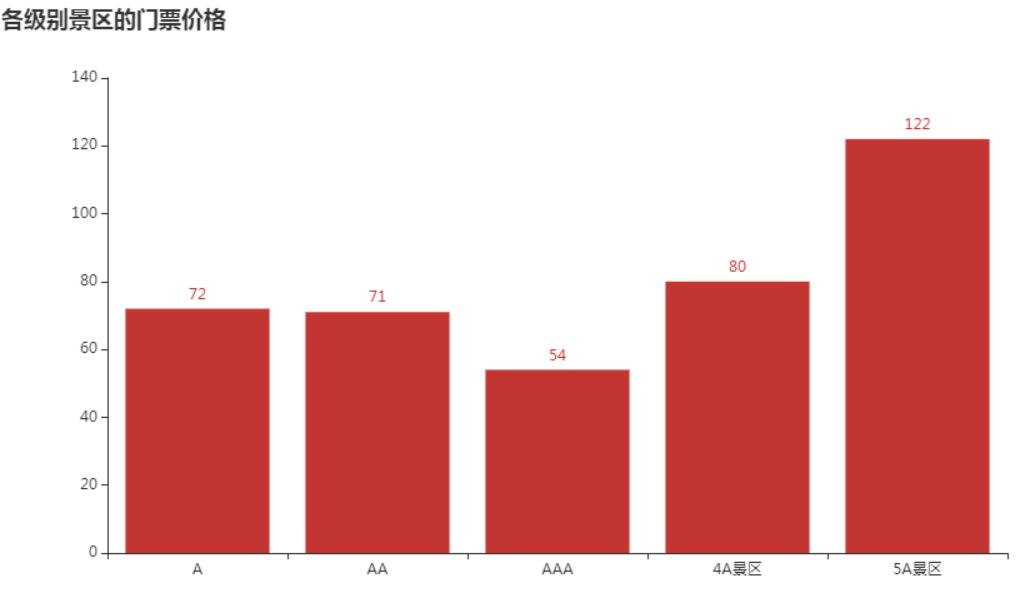


图2.11各级别景区的门票价格

# 5.总结

人文风光大多数在南方地区，主要是古镇、游乐园、动物园、古迹等，自然风光大多在北方，例如壁画、草原、戈壁、峡谷等。

大多数游客的旅游地都在南方沿海一带，去的旅游景点都是十分著名的热门景点，例如杭州的西湖、上海的迪斯尼、广州的长隆、北京的故宫等，当然旅游体验也很好评分都很高但是却没什么折扣。

总而言之，旅游地的选择和景点的选择大多要依靠游客自己的主观感受，以上数据仅供参考。

# 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.

**课程设计成绩评定表**

|  |  |  |  |
| --- | --- | --- | --- |
| 课程设计题目 | 节假日长假景点人满为患的数据分析 | | |
| 程设计学生答辩或质疑记录：  1.什么是pyecharts？  pyecharts 是一个用于生成 Echarts 图表的类库。 echarts 是百度开源的一个数据可视化 JS 库，主要用于数据可视化。pyecharts 是一个用于生成 Echarts 图表的类库。实际上就是 Echarts 与 Python 的对接。使用 pyecharts 可以生成独立的网页，也可以在 flask , Django 中集成使用。  2.简述pandas中的apply函数？  apply()函数的基本形式为：DataFrame.apply(func, axis=0, broadcast=False, raw=False, reduce=None, args=(), \*\*kwds)其中，func是你想要这个DataFrame应用的函数，可以自己编写，也可以是已经存在的，相当于C/C++的函数指针。如果这个参数是自己实现的话，函数的传入参数根据axis来定，比如axis = 1，就会把一行数据作为Series的数据结构传入给自己实现的函数中，我们在函数中实现对Series不同属性之间的计算，返回一个结果，则apply函数会自动遍历每一行DataFrame的数据，最后将所有结果组合成一个Series数据结构并返回。 | | | |
| **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、遵守纪律 | | 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、创新能力 | | 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 | | | |