|  |  |
| --- | --- |
| 学号 | 10212818129 |



课 程 设 计

课程名称 python程序设计课程设计

|  |  |
| --- | --- |
| 题 目 | 数据分析的学习与实践-节假日旅游分析 |
| 专 业 | 软件工程 |
| 班 级 | 软件1181 |
| 姓 名 | 曾逸雯 |
| 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 研究背景 5](#_Toc62240197)

[2 开发环境与相关技术 5](#_Toc62240198)

[2.1项目托管平台github 5](#_Toc62240199)

[2.2开发平台jupyter notebook 6](#_Toc62240200)

[2.3 编程语言python语言 6](#_Toc62240201)

[3.程序设计 7](#_Toc62240202)

[3.1 数据分析库导入 7](#_Toc62240203)

[3.2 数据清洗 7](#_Toc62240204)

[3.2.1数据读取 7](#_Toc62240205)

[3.2.2字段保留 8](#_Toc62240206)

[3.2.3数据去重 8](#_Toc62240207)

[3.2.4类型转换 8](#_Toc62240208)

[3.2.5字段清洗 9](#_Toc62240209)

[3.2.6数据合并 9](#_Toc62240210)

[4.分析方向与问题分析 10](#_Toc62240211)

[4.1 分析方向 10](#_Toc62240212)

[4.2 问题分析 10](#_Toc62240213)

[4.2.1选择最多的城市与省份 10](#_Toc62240214)

[4.2.2最受游客青睐的城市 11](#_Toc62240215)

[4.2.3最热门的景点 12](#_Toc62240216)

[4.2.4 各省好评热门景点 13](#_Toc62240217)

[4.2.5热门但评分不好的景点 14](#_Toc62240218)

[4.2.6折扣大的景点 15](#_Toc62240219)

[4.2.7不同级别景区票价 15](#_Toc62240220)

[4.3 结论 16](#_Toc62240221)

[5.总结 16](#_Toc62240222)

[参考文献 17](#_Toc62240223)

# 1 研究背景

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。同时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。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但是旅游地及景点的多样化使人们多了很多选择，所以决定对旅游景点的分析作为选题为游客提供数据参考。

# 2 开发环境与相关技术

## 2.1项目托管平台github

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。

GitHub项目本身自然而然的也在GitHub上进行托管，只不过在一个私有的，公共视图不可见的库中。开源项目可以免费托管，但私有库则并不如此。ChrisWanstrath, GitHub的开发者之一，肯定了通过付费的私有库来在财务上支持免费库的托管这一计划。在GitHub，用户可以十分轻易地找到海量的开源代码。

## 2.2开发平台jupyter notebook

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。Jupyter Notebook 的本质是一个 Web 应用程序，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。 用途包括：数据清理和转换，数值模拟，统计建模，机器学习等等用户可以通过电子邮件，Dropbox，GitHub 和 Jupyter Notebook Viewer，将 Jupyter Notebook 分享给其他人。在Jupyter Notebook 中，代码可以实时的生成图像，视频，LaTeX和JavaScript。

## 2.3 编程语言python语言

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语言的替代品。Python提供了高效的高级数据结构，还能简单有效地面向对象编程。Python语法和动态类型，以及解释型语言的本质，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解释器易于扩展，可以使用C或C++等其他可以通过C调用的语言扩展新的功能和数据类型，也可用于可定制化软件中的扩展程序语言。Python中丰富的标准库，提供了适用于各个主要系统平台的源码或机器码。

Python在设计上坚持了清晰划一的风格，这使得Python成为一门易读、易维护，并且被大量用户所欢迎的、用途广泛的语言。在设计Python语言时，如果面临多种选择，Python开发者一般会拒绝花俏的语法，而选择明确的没有或者很少有歧义的语法。由于这种设计观念的差异，Python源代码通常被认为比Perl具备更好的可读性，并且能够支撑大规模的软件开发。Python的作者有意的设计限制性很强的语法，使得不好的编程习惯(例如if语句的下一行不向右缩进)都不能通过编译。其中很重要的一项就是Python的缩进规则。一个和其他大多数语言的区别就是，一个模块的界限，完全是由每行的首字符在这一行的位置来决定的(而C语言 是用一对花括号{}来明确的定出模块的边界的，与字符的位置毫无关系)。这一点曾经引起过争议。因为自从C这类的语言诞生后，语言的语法含义与字符的排列方式分离开来,曾经被认为是一种程序语言的进步。不过不可否认的是，通过强制程序员们缩进(包括if, for和函数定义等所有需要使用模块的地方)， Python确实使得程序更加清晰和美观。

# 3.程序设计

## 3.1 数据分析库导入

通过导入pandas 、Numpy、Matplotlib、Seaborn为代码所涉及的维度数组数学函数库及分析数据的图形化显示。

使用%matplotlib命令可以将matplotlib的图表直接嵌入到Notebook之中，或者使用指定的界面库显示图表，它有一个参数指定matplotlib图表的显示方式。inline表示将图表嵌入到Notebook中。如下图3.1所示。

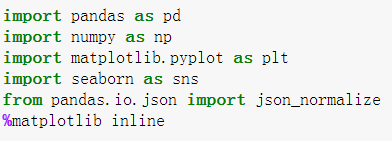


图3.1数据分析库导入

## 3.2 数据清洗

### 3.2.1数据读取

直接从本地磁盘读取从飞猪获得的旅游信息，如图3.2所示。



图3.2数据读取

### 3.2.2字段保留

因为数据中总字段多达31个所以通过选择将数据分析中所需要的字段存放到col\_keep中保留起来，如图3.3保留字段所示。



图3.3保留字段

### 3.2.3数据去重

通过drop\_duplicates()函数将fields.itemId中重复的数据删除，如图3.4去重所示。



图3.4去重

### 3.2.4类型转换

将fields.comment、fields.discountPrice、fields. itemId、fields. price、fields. soldRecentNum等字段的数据类型转换成合理的类型，如图3.5数据类型转换所示。



图3.5数据类型转换

### 3.2.5字段清洗

通过定义get\_fea\_text()函数来判断原数据中fields.features传过来的数据，如果没有则返回空，如果有则转换为字典类型的数组再将text里的值返回。如图3.6产品特点清理所示。



图3.6产品特点清理

首先对产品的年销量通过判断fields.sold365是否为空，若为空则返回np.nan，在判断数据中是否包含”万”来返回销量是数量还是金额，若是”万”则将”售”与”万”之间的数据截取出来乘以10000，如果不是”万”则将”售”与”笔”之间的数据截取出来返回。如图3.7产品年销售清理。



图3.7产品年销售清理

### 3.2.6数据合并

将city\_data.csv与清洗后的数据通过city字段合并到一起，完成数据的清洗。如图3.8数据合并所示。

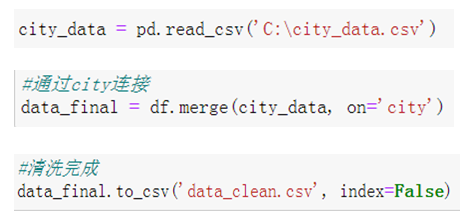


图3.8数据合并

# 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选择最多的城市与省份

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。再其次就是广州、厦门、苏州、香港、成都、重庆、清远。通过查看地图发现大多数游客所选择旅游的省份都在长江中下游，秦岭淮河以南的城市。如图4.1、4.2所示。

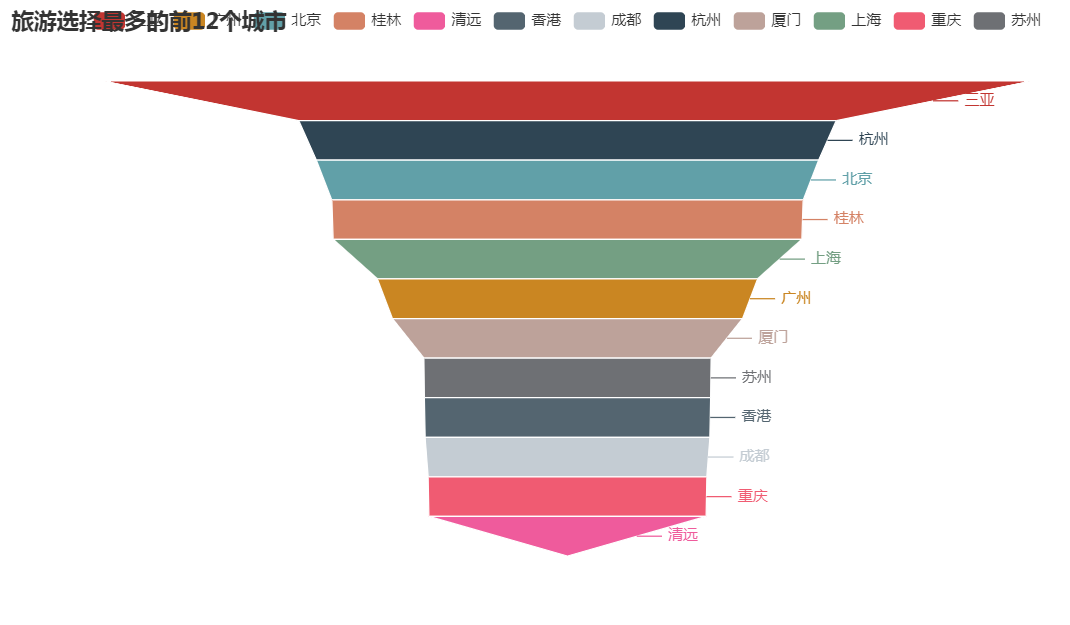


图4.1旅游选择最多的12个城市

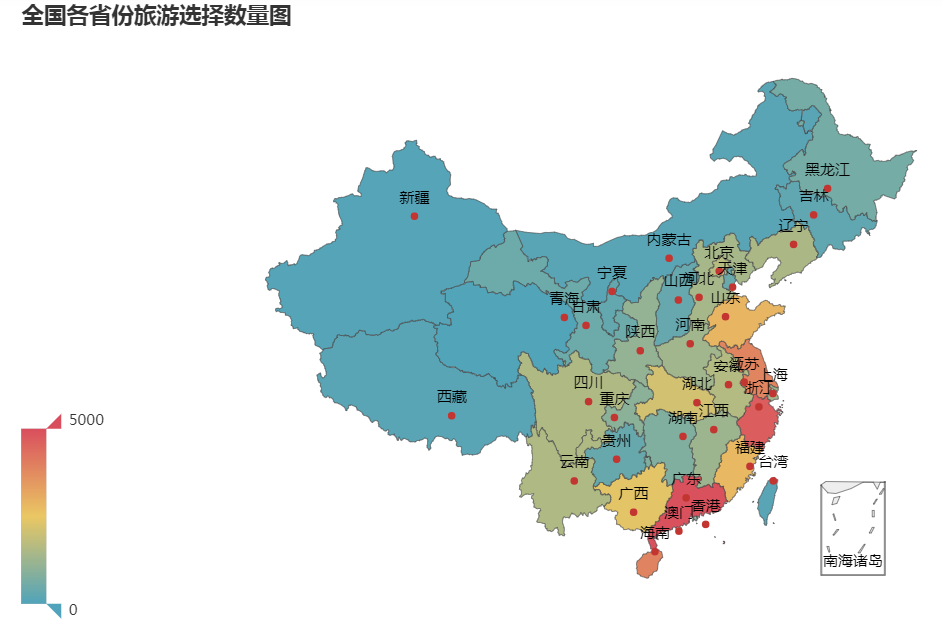


图4.2全国各省份旅游选择数量图

### 4.2.2最受游客青睐的城市

通过查看最近一个月售出的门票发现游客们对武汉、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4.3所示



图4.3近一个月售出门票top20

### 4.2.3最热门的景点

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4.4所示

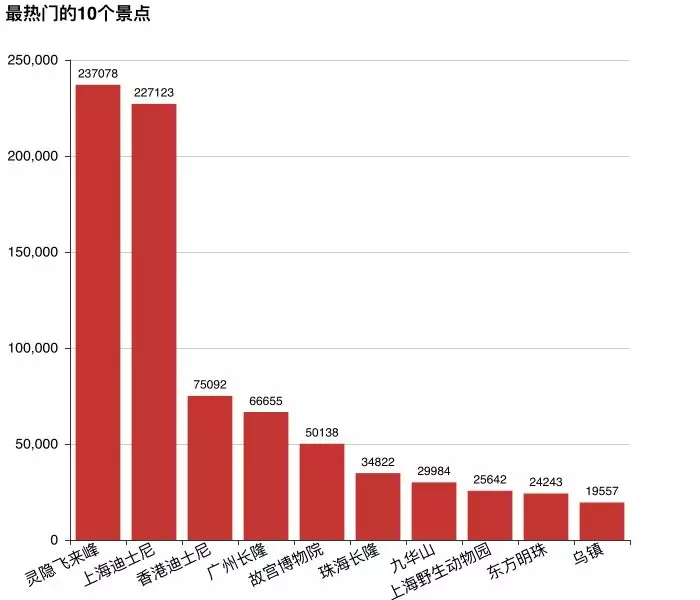


图4.4近一个月售出门票top20

### 4.2.4 各省好评热门景点

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，其评分都为5.0。如图4.5、图4.6所示



图4.5各省受好评的热门景点(一)



图4.6各省受好评的热门景点(二)

### 4.2.5热门但评分不好的景点

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。



图4.7评分不高的热门景点

### 4.2.6折扣大的景点

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，还是有一部分景点的折扣在4折左右。如图4.8所示。



图4.8折扣大的景点

### 4.2.7不同级别景区票价

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。

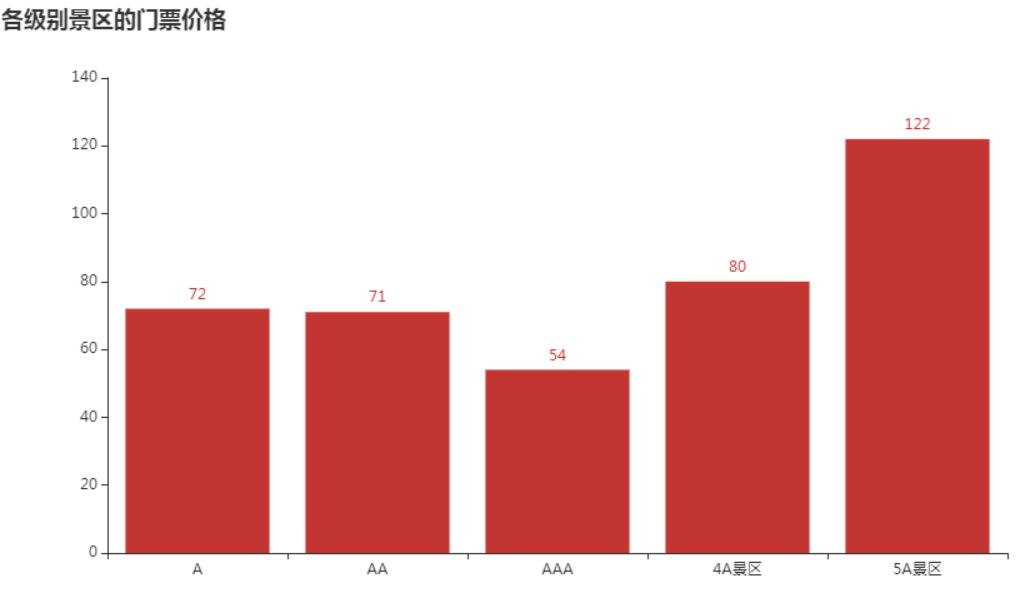


图4.9不同级别的景点售价

## 4.3 结论

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。大多数游客都比较偏向于去那种名气比较大的旅游景点，比如说杭州的西湖、上海的迪斯尼、广州的长隆、北京的故宫等。南方沿海城市是大多数游客的选择但是那些热门景点的折扣相对来说很小。

# 5.总结

课程设计是培养学生综合运用所学知识，发现，提出，分析和解决实际问题，锻炼实践能力的重要环节，是对学生实际工作能力的具体训练和考察的过程，这次的课程设计完成的是节假日旅游景点人满为患的数据分析，使用github和jupyter notebook进行开发设计。

虽然完成的过程并不是一帆风顺，遇到了一些问题，例如运行出错、图表不显示等…但是这些犯的错误中使我明白，做任何事都要懂得坚持，任何时候都不能自满自足，学海无涯，还有很多东西自己还不会需要去学习，所以不断请教他人非常有必要。回顾起此课程设计，至今我仍感慨颇多，从理论到实践，在这段日子里，可以说得:是苦多于甜，但是可以学到很多很多的东西，同时不仅可以巩固了以前所学过的知识,，而且学到了很多在书本上所没有学到过的知识。通过这次课程设计使我懂得了理论与实际相结合是很重要的，只有理论知识是远远不够的，只有把所学的理论知识与实践相结合起来，从理论中得出结论，才能真正为社会服务，从而提高自己的实际动手能力和独立思考的能力。在设计的过程中遇到问题，可以说得是困难重重，但可喜的是最终都得到了解决。最有收获的是提供了一个向老师学习的好机会。我从老师身上学到了很多东西，老师教会了我们一个从理论走向实际运用的方法，也教会了我们同学间宝贵的合作精神。老师认真负责的工作态度，严谨的治学精神和深厚的理论水平都使我受益匪浅。无论在理论上还是在实践中，都给予我很大的帮助，使我得到不少的提高。

# 参考文献

[1] 丁辉.Python基础与大数据应用.人民邮电出版社

[2] 刘宇宙.python3.5从零开始学[M]. 北京：清华大学出版社，2017.

[3] 埃里克.Python编程：从入门到实践.人民邮电出版社

**课程设计成绩评定表**

|  |  |  |  |
| --- | --- | --- | --- |
| 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：  问题一：怎么获取数据的？  答：通过搜集数据将其存放在csv文件中，直接通过访问磁盘获取数据。  问题二：图表不显示问题怎么解决的？  答： 是因为pyecharts的版本问题导致的，将pyecharts版本更改的后就能成功打印图表了。  问题三：简述Seaborn的功能。  答：Seaborn其实是在matplotlib的基础上进行了更高级的API封装，使用seaborn就能做出很具有吸引力的图，而使用matplotlib就能制作具有更多特色的图。 | | | |
| **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 | | | |