|  |  |
| --- | --- |
| 学号 | 10212818107 |



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-地震的数据分析 |
| 专 业 | \_\_\_\_\_\_\_\_\_\_\_软件工程\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 班 级 | \_\_\_\_\_\_\_\_\_\_\_软件1181\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 姓 名 | \_\_\_\_\_\_ \_\_\_\_\_龚吉\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_ |
| 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

[前言 1](#_Toc61701281)

[1编程环境与相关技术 2](#_Toc61701282)

[1.1硬件平台 2](#_Toc61701283)

[1.2项目托管平台-GitHub 2](#_Toc61701284)

[1.3开发平台-jupyter 3](#_Toc61701285)

[1.4编程语言-python 4](#_Toc61701286)

[2程序设计（仅展示关键代码） 6](#_Toc61701287)

[2.1数据分析库 6](#_Toc61701288)

[2.1.1数据分析库导入见下图2.1-1 6](#_Toc61701289)

[2.1.2数据分析库注解 6](#_Toc61701290)

[2.2数据表 7](#_Toc61701291)

[2.2.1csv文件的读取2.2-1 7](#_Toc61701292)

[2.3数据清洗 8](#_Toc61701294)

[2.3.1数据清洗代码见下图2.1-1 8](#_Toc61701295)

[2.4数据分析图 9](#_Toc61701296)

[2.4.1数据提取代码见下图2.4-1 9](#_Toc61701297)

[2.4.2生成分析图代码见下图2.4-2 9](#_Toc61701298)

[2.5 统计全国招聘岗位工作天数 10](#_Toc61701299)

[2.5.1统计图生成代码见下图2.5-1 10](#_Toc61701300)

[2.6 全国岗位招聘要求统计图 11](#_Toc61701301)

[2.6.1统计图生成代码见下图2.6-1 11](#_Toc61701302)

[2.7 各个城市岗位的平均最高和最低工资统计图 12](#_Toc61701303)

[2.7.1统计图生成代码见下图2.7-1和2.7-2 12](#_Toc61701304)

[3数据分析 16](#_Toc61701310)

[3.1问题引入 16](#_Toc61701311)

[3.2数图分析 16](#_Toc61701312)

[4个人体会 16](#_Toc61701310)

# 

# 前言

古人有云：“纸上得来终觉浅，绝知此事要躬行”。  
  
　　自从走进了大学，距离工作就不远了，学校为了拓展我们学生自身的知识面，扩大与社会的接触面，增加个人在社会竞争中的经验，锻炼和提高我们的能力，以便在以后顶岗后能真正走入社会，能够适应国内外的经济形势的变化，在学习市场营销专业知识两年之后，组织我们进行认识实习。认识实习是教学与生产实际相结合的重要实践性教学环节。它不仅让我们学到了很多在课堂上根本就学不到的知识，还使我们开阔了视野，增长了见识，了解一个企业是怎样进行营销工作，怎样进行生产的。为我们以后更好把所学的知识运用到实际工作中打下坚实的基础。在认识实习过程中，学校也以培养学生观察问题、解决问题和向生产实际学习的能力和方法为目标。

# 1编程环境与相关技术

## 1.1硬件平台

CPU：Intel(R) Core(TM) i5-8300H CPU @ 2.30GHz (8 CPUs), ~2.3GHz

最低要求：64位7代i3处理器及以上

内存：16G

最低要求：4G

硬盘空间：SSD-120G

最低要求：机械硬盘-20G

## 1.2项目托管平台-GitHub

github作为开源代码库以及版本控制系统，Github拥有超过900万开发者用户。随着越来越多的应用程序转移到了云上，Github已经成为了管理软件开发以及发现已有代码的首选方法。

如前所述，作为一个分布式的版本控制系统，在Git中并不存在主库这样的概念，每一份[复制](https://baike.baidu.com/item/%E5%A4%8D%E5%88%B6" \t "_blank)出的库都可以独立使用，任何两个库之间的不一致之处都可以进行合并。

GitHub可以托管各种git库，并提供一个web界面，但它与外国的[SourceForge](https://baike.baidu.com/item/SourceForge/6562141" \t "_blank)、[Google Code](https://baike.baidu.com/item/Google%20Code" \t "_blank)或中国的[coding](https://baike.baidu.com/item/coding/8921246" \t "_blank)的服务不同，GitHub的独特卖点在于从另外一个项目进行分支的简易性。为一个项目贡献代码非常简单：首先点击项目站点的“fork”的按钮，然后将代码检出并将修改加入到刚才分出的代码库中，最后通过内建的“pull request”机制向项目负责人申请代码合并。已经有人将GitHub称为代码玩家的MySpace。

在GitHub进行分支就像在[Myspace](https://baike.baidu.com/item/Myspace" \t "_blank)（或[Facebook](https://baike.baidu.com/item/Facebook" \t "_blank)…）进行交友一样，在社会关系图的节点中不断的连线。

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](https://baike.baidu.com/item/%E5%BC%80%E6%BA%90%E9%A1%B9%E7%9B%AE" \t "_blank)可以免费托管，但私有库则并不如此。Chris Wanstrath，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。

通过与客户的接洽，开发FamSpam，甚至是开发GitHub本身，GitHub的私有库已经被证明了物有所值。任何希望节省时间并希望和团队其它成员一样远离页面频繁转换之苦的人士都会从GitHub中获得他们真正想要的价值。

在GitHub，用户可以十分轻易地找到海量的[开源](https://baike.baidu.com/item/%E5%BC%80%E6%BA%90/20720669" \t "_blank)代码。

## 1.3开发平台-jupyter

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

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。

## 1.4编程语言-python

Python的设计哲学是“优雅”、“明确”、“简单”。因此，[Perl语言](https://baike.baidu.com/item/Perl%E8%AF%AD%E8%A8%80/1346108" \t "_blank)中“总是有多种方法来做同一件事”的理念在Python开发者中通常是难以忍受的。Python开发者的哲学是“用一种方法，最好是只有一种方法来做一件事”。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。这些准则被称为Python格言。在Python[解释器](https://baike.baidu.com/item/%E8%A7%A3%E9%87%8A%E5%99%A8)内运行import this可以获得完整的列表。

Python开发人员尽量避开不成熟或者不重要的优化。一些针对非重要部位的加快运行速度的补丁通常不会被合并到Python内。所以很多人认为Python很慢。不过，根据二八定律，大多数程序对速度要求不高。在某些对运行速度要求很高的情况，Python设计师倾向于使用[JIT](https://baike.baidu.com/item/JIT" \t "_blank)技术，或者用使用C/C++语言改写这部分程序。可用的JIT技术是[PyPy](https://baike.baidu.com/item/PyPy" \t "_blank)。

Python是完全[面向对象](https://baike.baidu.com/item/%E9%9D%A2%E5%90%91%E5%AF%B9%E8%B1%A1/2262089" \t "_blank)的语言。[函数](https://baike.baidu.com/item/%E5%87%BD%E6%95%B0/18686609)、模块、数字、[字符串](https://baike.baidu.com/item/%E5%AD%97%E7%AC%A6%E4%B8%B2/1017763)都是对象。并且完全支持继承、重载、派生、多继承，有益于增强源代码的复用性。Python支持重载运算符和动态类型。相对于[Lisp](https://baike.baidu.com/item/Lisp/22083" \t "_blank)这种传统的函数式编程语言，Python对函数式设计只提供了有限的支持。有两个标准库(functools, itertools)提供了[Haskell](https://baike.baidu.com/item/Haskell" \t "_blank)和Standard ML中久经考验的函数式程序设计工具。

虽然Python可能被粗略地分类为“[脚本语言](https://baike.baidu.com/item/%E8%84%9A%E6%9C%AC%E8%AF%AD%E8%A8%80)”（script language），但实际上一些大规模软件开发计划例如[Zope](https://baike.baidu.com/item/Zope" \t "_blank)、[Mnet](https://baike.baidu.com/item/Mnet" \t "_blank)及[BitTorrent](https://baike.baidu.com/item/BitTorrent/142795" \t "_blank)，[Google](https://baike.baidu.com/item/Google/86964" \t "_blank)也广泛地使用它。Python的支持者较喜欢称它为一种高级动态编程语言，原因是“脚本语言”泛指仅作简单程序设计任务的语言，如shellscript、[VBScript](https://baike.baidu.com/item/VBScript" \t "_blank)等只能处理简单任务的编程语言，并不能与Python相提并论。

Python本身被设计为可扩充的。并非所有的特性和功能都集成到语言核心。Python提供了丰富的[API](https://baike.baidu.com/item/API/10154" \t "_blank)和工具，以便程序员能够轻松地使用[C语言](https://baike.baidu.com/item/C%E8%AF%AD%E8%A8%80/105958" \t "_blank)、[C++](https://baike.baidu.com/item/C%2B%2B/99272)、Cython来编写扩充模块。Python编译器本身也可以被集成到其它需要脚本语言的程序内。因此，很多人还把Python作为一种“胶水语言”（glue language）使用。使用Python将其他语言编写的程序进行集成和封装。在Google内部的很多项目，例如Google Engine使用C++编写性能要求极高的部分，然后用Python或Java/Go调用相应的模块。《[Python技术手册](https://baike.baidu.com/item/Python%E6%8A%80%E6%9C%AF%E6%89%8B%E5%86%8C" \t "_blank)》的作者[马特利](https://baike.baidu.com/item/%E9%A9%AC%E7%89%B9%E5%88%A9)（Alex Martelli）说：“这很难讲，不过，2004 年，Python 已在[Google](https://baike.baidu.com/item/Google" \t "_blank) 内部使用，Google 召募许多 Python 高手，但在这之前就已决定使用Python，他们的目的是 Python where we can, C++ where we must，在操控硬件的场合使用[C++](https://baike.baidu.com/item/C%2B%2B/99272" \t "_blank)，在快速开发时候使用 Python。”

# 2程序设计（仅展示关键代码）

## 2.1数据分析库

### 2.1.1数据分析库导入见下图2.1-1

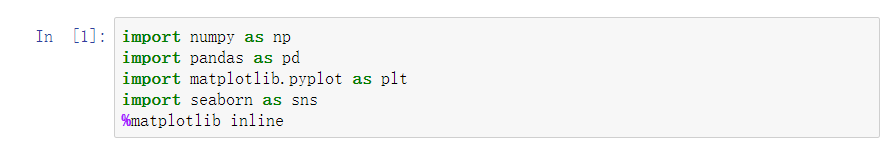


图2.1-1

### 2.1.2数据分析库注解

pandas 是 python 的数据分析处理库 NumPy是Python语言的一个扩充程序库。支持高级大量的维度数组与bai矩阵运算，此外也针对数组运算提供大量的数学函数库。

Numpy内部解除了CPython的GIL（全局解释器锁），运行效率极好，是大量机器学习框架的基础库！

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。

%matplotlib inline Seaborn是一种基于matplotlib的图形可视化python libraty。它提供了一种高度交互式界面，便于用户能够做出各种有吸引力的统计图表。

Seaborn其实是在matplotlib的基础上进行了更高级的API封装，从而使得作图更加容易，在大多数情况下使用seaborn就能做出很具有吸引力的图，而使matplotlib就能制作具有更多特色的图。

Matplotlib是一个Python的2D绘图库，它以各种硬拷贝格式和跨平台的交互式环境生成出版质量级别的图形。通过Matplotlib，开发者可以仅需要几行代码，便可以生成绘图。一般可绘制折线图、散点图、柱状图、饼图、直方图、子图等等。Matplot使用Numpy进行数组运算，并调用一系列其他的Python库来实现硬件交互。

## 2.2数据表

### 2.2.1csv文件的读取2.2-1

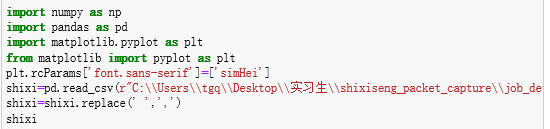


图2.2-1

## 2.3数据清洗

### 2.3.1数据清洗代码见下图2.1-1



图2.3-1

## 2.4数据分析

### 2.4.1数据提取代码见下图2.4-1

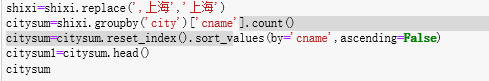


图2.4-1

### 2.4.2生成分析图见下图2.4-2

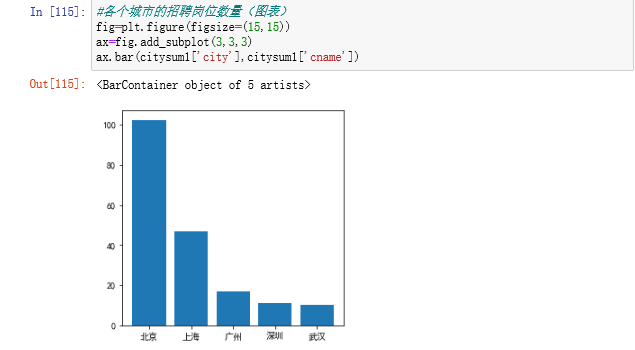


图2.4-2

## 2.5 统计全国招聘岗位工作天数

### 2.5.1统计图生成代码见下图2.5-1

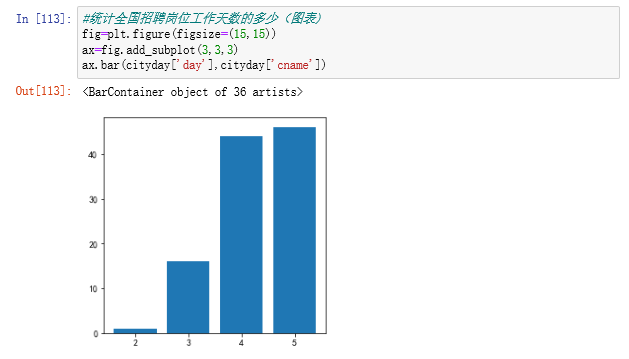


图2.5-1

## 2.6 全国岗位招聘要求统计图

### 2.6.1统计图生成代码见下图2.6-1

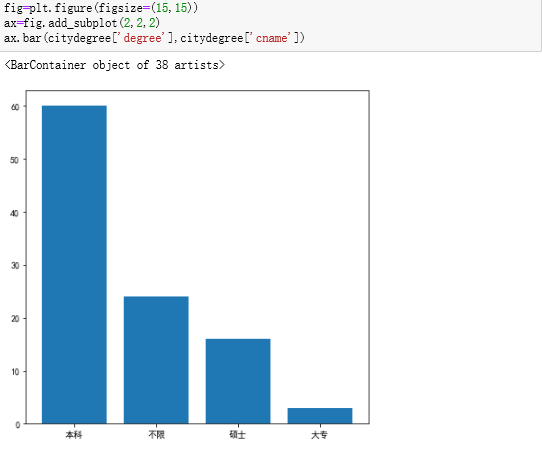


图2.6-1

## 2.7 各个城市岗位的平均最高和最低工资统计图

### 2.7.1统计图生成代码见下图2.7-1

IMG_256

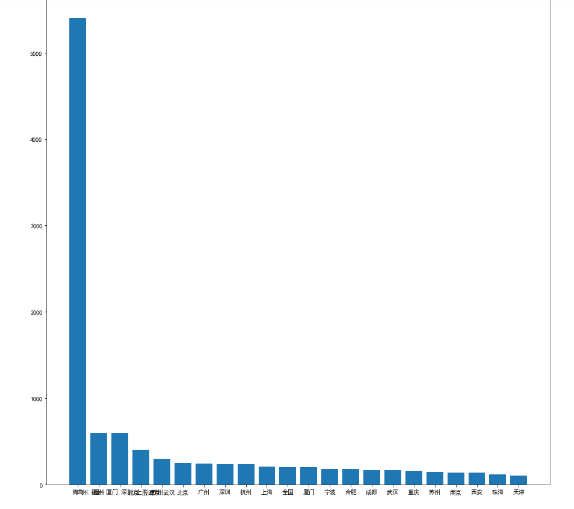


图2.7-1

IMG_256

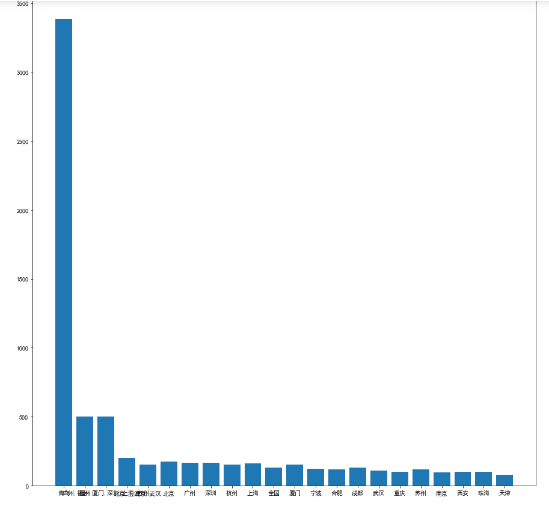


图2.7-2

# 3数据分析

## 3.1引入问题

对于非北上广的普通院校应届生申请实习，相关单位会提供机会吗?相关网络言论见图3.1-1

这是网络上的回答，第一眼看的不是能力，而是做事要脚踏实地。

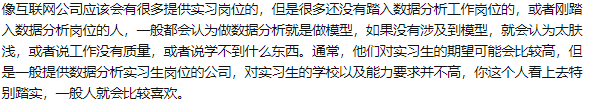


图3.1-1

## 3.2数图分析

我们先看各地方的岗位实习职位数量，见图3.2-1

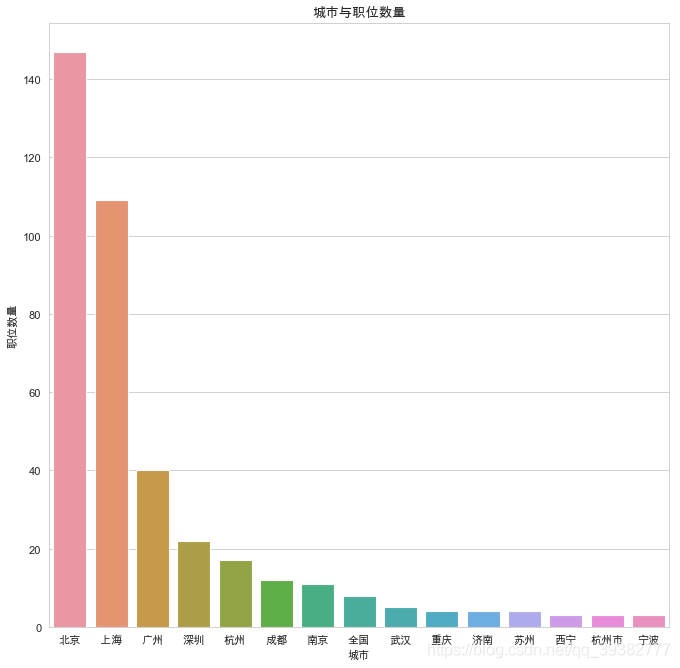
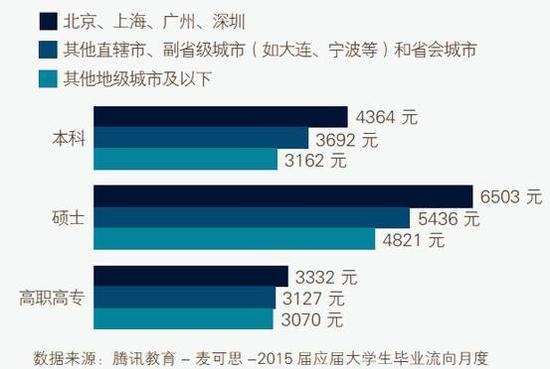


图3.2-1

可以看到，北上广深等一线城市职位数量遥遥领先，而后面的城市职位数量就非常少，从图也可发现去大城市发展的人很多，竞争也很激烈。

以下是北上广深学历和工资平均对比图。



可见， 一线城市的职位平均工资要高一些，但是个城市生活水平质量也不同，每个人的选择也不同，但是大多数还是会选择一线城市。

各年级大学生关于实习的态度

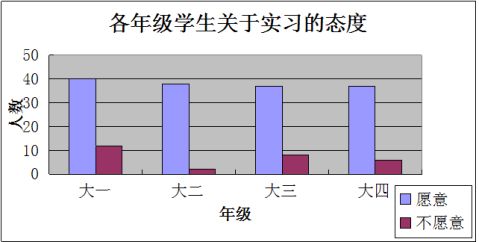


图3.2-3

绝大部分在校大学生希望参加实习。从调查数据中可以看出，对实习持赞成态度的学生占到了总数的84%，说明大部分本科生还是比较向往参加实习工作的。但是，各个年级的比例却有一些相对的差距。大一学生可能由于刚刚步入大学，对周边环境还不了解，赞成率仅为76.9%。大二学生对环境已逐渐熟悉，所以赞成率达到95.0%。而大三学生由于考研压力，反对率达到了18.8%。而大四学生的未来前景选择各异，赞成率为86.1%。综合上述，首先可以看出大部分学生倾向于参加实习工作，向往体验职场生活；其次可以看出，周边环境对学生态度有所影响；最后，对实习的态度产生影响的因素较多，因人而异。

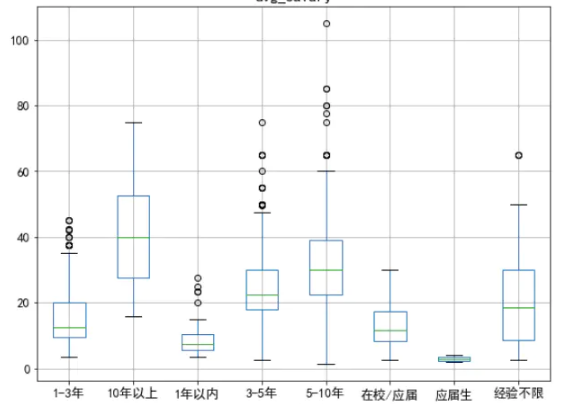


图3.2-4

工作经验上看，薪资的差距拉开的很大，应届生和有工作经验的

薪资对比，不在一个维度上，但这恰恰说明数据类岗位是可以通过个人能力的提高来获取更大的收益的，它是可成长的。你对技术、业务越熟悉，你的机会就越多，薪资就越高。

# 4个人体会

这次做论文使我受益良多。一方面，通过这次课设的完成，我对Python有了更深刻的理解，能够更加熟练的运用这门语言，更加理解到了编程的魅力。而另一方面，使我明白了认真做事的重要性。使我体会到完成一件事的喜悦。以后面对困难的时候，都能有直面困难并克服它的决心。本次课设能够顺利完成，也感谢老师的辅导和同学们互相帮助学习，使我能够很好的掌握和运用专业知识，并在设计中得以体现。在以后的学习生活中我一定会能有更好的发挥，克服自己的缺点，不断改进，不断成长学习。

|  |  |  |  |
| --- | --- | --- | --- |
| 课程设计题目 | 数据分析的学习与实践-实习岗位状况 | | |
| 程序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。   1. 数据表中的重复值怎么处理。   答：pandas中有两个函数是专门用来处理重复值的，第一个是duplicated函数。Duplicated函数用来查找并显示数据表中的重复值。第二个是drop\_duplicates函数用来删除数据表中的重复值，判断标准和逻辑与duplicated函数一样。  3.数据的空值怎么处理。  答：在python中空值被显示为NaN。在处理空值之前我们先来检查下数据表中的空值数量。对于一个小的数据表，我们可以人工查找，但对于较为庞大的数据表，就需要寻找一个更为方便快捷的方法了。Pandas中查找数据表中空值的函数有两个，一个是函数isnull，如果是空值就显示True。另一个函数notnull正好相反，如果是空值就显示False。 | | | |
| **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 | | | |

**课程设计成绩评定表**