|  |  |
| --- | --- |
| 学号 | 10214518327 |



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-实习岗位状况的数据分析 |
| 专 业 | 软件工程 |
| 班 级 | 软件1182 |
| 姓 名 | 高维程 |
| 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**

[摘要 6](#_Toc62151892)

[1. 分析目标 7](#_Toc62151893)

[2.分析流程 7](#_Toc62151894)

[3.分析实施 7](#_Toc62151895)

[3.1数据获取 7](#_Toc62151896)

[3.2数据预处理 9](#_Toc62151897)

[3.3 数据标准化 11](#_Toc62151898)

[4.数据分析 11](#_Toc62151899)

[4.1 学历与岗位数量 11](#_Toc62151900)

[4.2 城市与岗位数量 14](#_Toc62151901)

[4.3 工资分析 16](#_Toc62151902)

[4.3.1不同城市、岗位，每日工资最高、最低情况 16](#_Toc62151903)

[4.3.2 平均日薪 17](#_Toc62151904)

[4.3.3 全国的平均月薪 17](#_Toc62151905)

[4.3.4 月薪与城市 18](#_Toc62151906)

[4.3.5 不同城市岗位的占比 18](#_Toc62151907)

[4.3.6 月薪情况(降序排列) 19](#_Toc62151908)

[4.4 公司规模分析 20](#_Toc62151909)

[4.思考 24](#_Toc62151910)

[5.总结 24](#_Toc62151911)

[参考文献 26](#_Toc62151912)

[课程设计成绩评定表 27](#_Toc62151913)

# 摘要

在互联网技术不断升级的今天，数据获取变的越来越容易，大数据行业越来越火，基于海量数据进行分析的数据分析人才逐渐成为各大企业追逐的宠儿。大数据推动起数据的深度分析与挖掘，使得研究结果更加客观，并且多维。

实习是每个大学生必须拥有的一段经历，它使我们在实践中了解社会，让我们学到了很多在课堂上根本就学不到的知识，受益匪浅，也打开了视野，增长了见识，为我们以后进一步走向社会打下坚实的基础。受去年疫情的影响，各行各业就业形势不容乐观，企业实习岗位情况也受到了相应冲击。本文针对所提供的相关课题的数据进行实习岗位的状况的分析,

主要从学历与实习岗位需求量、城市与岗位需求量、薪资水平、公司规模这四个方面进行分析。

# 1.分析目标

学历与岗位需求量

城市与岗位需求量

薪资水平分析

公司规模分析

# 2.分析流程

数据获取:找出需要用作数据分析的数据并进行获取。

数据预处理:对获取数据进行去重、缺失值处理等一系列数据清洗操作。

数据可视化:文字不如数据，数据不如图片，可视化是对于数据最好的展现方式，让人一目了然。

深入分析:对于数据深入分析，寻找不同因素之间与薪资之间的关系。

# 3.分析实施

## 3.1数据获取

由于数据老师已提供，这里仅放爬虫的代码，代码如下:

#!/usr/bin/python

# -\*- coding:utf-8 -\*-

import requests

import pandas as pd

def sxs\_crawl(pages=30, kw='数据挖掘', c='全国'):

list\_urls = ["https://iosapi.shixiseng.com/app/interns/search?c={}&d=&ft=&i=&k={}"

"&m=&page={}&s=-0&st=&t=zj&x=&z=".format(c, kw, page) for page in range(pages)]

job\_list\_data = []

for url in list\_urls:

response = requests.get(url)

if response.json()['msg']:

job\_list\_data.extend(response.json()['msg'])

else:

break

job\_list = pd.DataFrame(job\_list\_data)

job\_list.to\_csv('/Users/apple/Desktop/job\_list.csv', index=False)

# 职位详情ID爬取

uuids = list(job\_list['uuid'])

job\_detailed\_url = ['https://iosapi.shixiseng.com/app/intern/info?uuid={}'.format(uuid) for uuid in uuids]

job\_detailed\_data = []

for url in job\_detailed\_url:

response = requests.get(url)

job\_detailed\_data.append(response.json()['msg'])

job\_detailed = pd.DataFrame(job\_detailed\_data)

job\_detailed.to\_csv('/Users/apple/Desktop/job\_detailed.csv', index=False)

# 公司信息爬取

cuuids = list(job\_detailed['cuuid'])

com\_detailed\_url = ['https://iosapi.shixiseng.com/app/company/info?uuid={}'.format(cuuid) for cuuid in cuuids]

com\_detailed\_data = []

for url in com\_detailed\_url:

response = requests.get(url)

com\_detailed\_data.append(response.json()['msg'])

com\_detailed = pd.DataFrame(com\_detailed\_data)

com\_detailed.to\_csv('/Users/apple/Desktop/com\_detailed.csv', index=False)

print('Successfully crawled {} jobs.'.format(job\_list.shape[0]))

if \_\_name\_\_ == '\_\_main\_\_':

sxs\_crawl(pages=30, kw='数据挖掘', c='全国')

## 3.2数据预处理

对抓取完的数据，首先进行的是数据清洗。然后进行重复项删除，最后只留下唯一的数据。对job\_detailed.csv文件的数据清洗如图3.1所示，对com\_detailed.csv文件的数据清洗如图3.2所示。

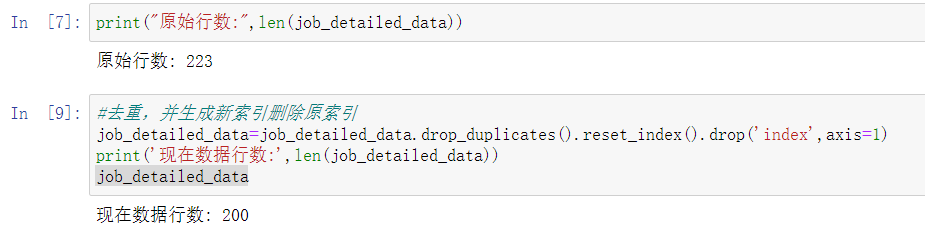


图3.1

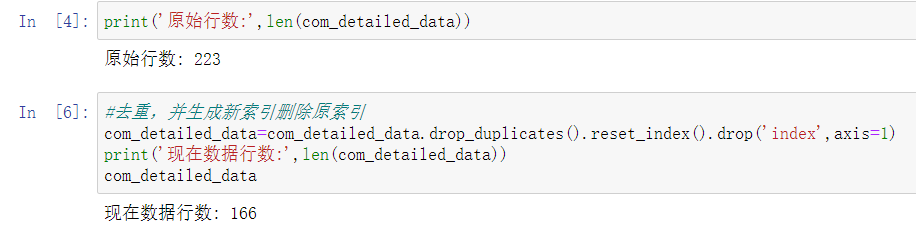


图3.2

针对前三个问题(学历与岗位需求量、城市与岗位需求量、薪资水平分析)的分析，因部分数据内容与分析岗位状况无任何关系，故将不需要的数据不做筛选。我的判定方式是将testjob.csv文件中包含"city"、"cname"、"day"、"degree"、"iname"、"industry"、"maxsal"、"minsal"、"month"、"attraction"中的的字段保留,然后导入Mysql。处理之后数据如图3.3所示。



图3.3

由于"city"这个字段的有些数据内容存在多个城市（图3.2），在之后的数据可视化图形展示中会存在歧义，故将这些数据直接剔除。

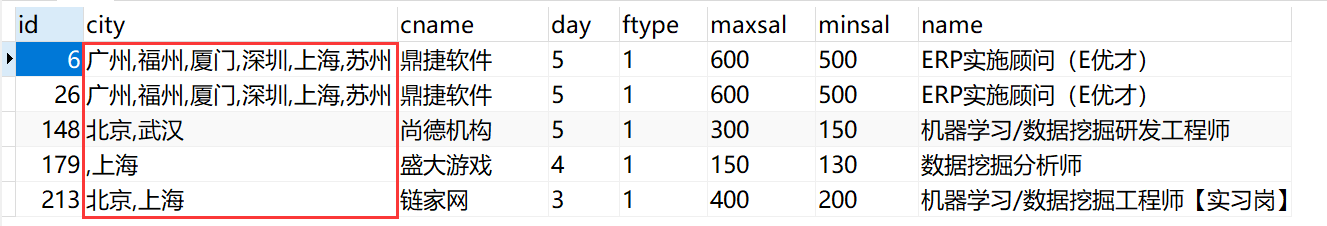


图3.2

针对最后一个问题(公司规模)的分析,因部分数据内容与分析岗位状况无任何关系，故将不需要的数据不做筛选。我的判定方式是将com\_detailed.csv文件中包含"cname"、"com\_type”、"description"、”industry”、"name"、" pranum"、" reg\_capi"、" reg\_name"、" scale"中的的字段保留。处理之后数据如图3.1所示。

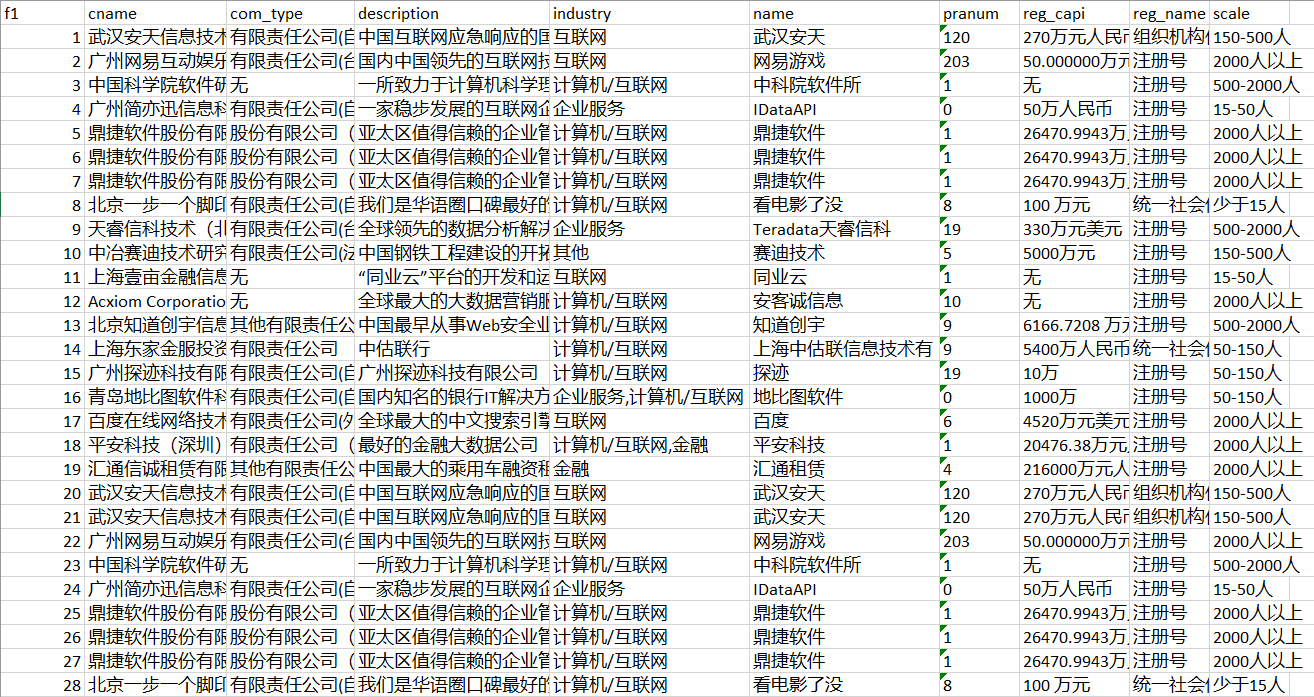


图3.1

## 3.3 数据标准化

标准化主要是对薪资和所属行业进行规范。这次统计，取薪资上的最高值与最低值的平均数为计算数据。

同时需要标准化的还有所属行业。所属行业抓取的结果是用“/”隔开的，所以先对其分列，在分别定义维度。

# 4.数据分析

## 4.1 学历与岗位数量

通过Python代码进行统计学历与岗位数量的关系，如图4.1

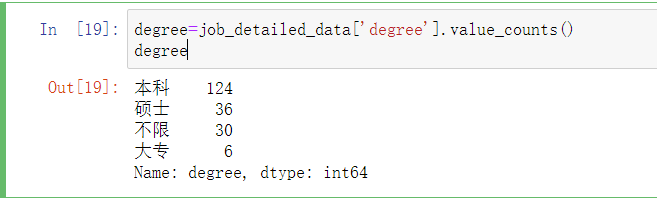


图4.1

通过sql语句进行统计学历与岗位数量的关系，如图4.2

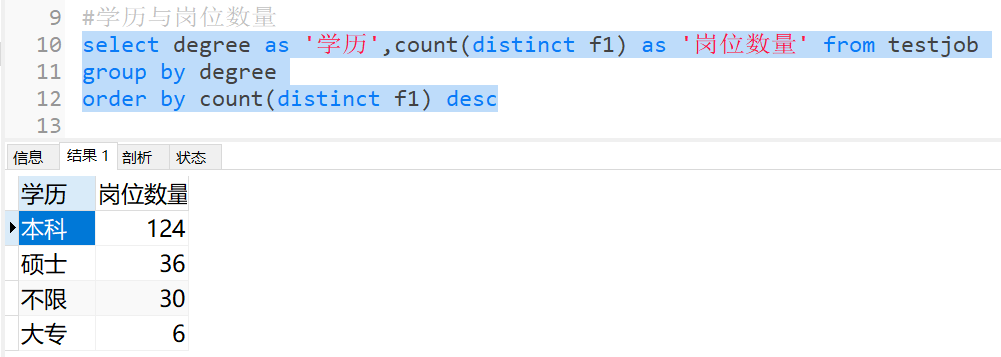


图4.2

通过Python代码实现学历与岗位数量的数据可视化，代码如下,数据可视化之后如图4.3所示。

|  |
| --- |
| plt.figure(figsize=(10, 10), dpi=80)  x = degree[:15].index  y = degree[:15].values  p = plt.bar(x, y, label="num")  for x,y in zip(x,y):  plt.text(x,y, '%.0f'%y, ha='center',va='bottom')  plt.xlabel('学历')  plt.ylabel('岗位数量')  plt.title('学历与岗位数量')  plt.legend(loc="upper right")  plt.show() |

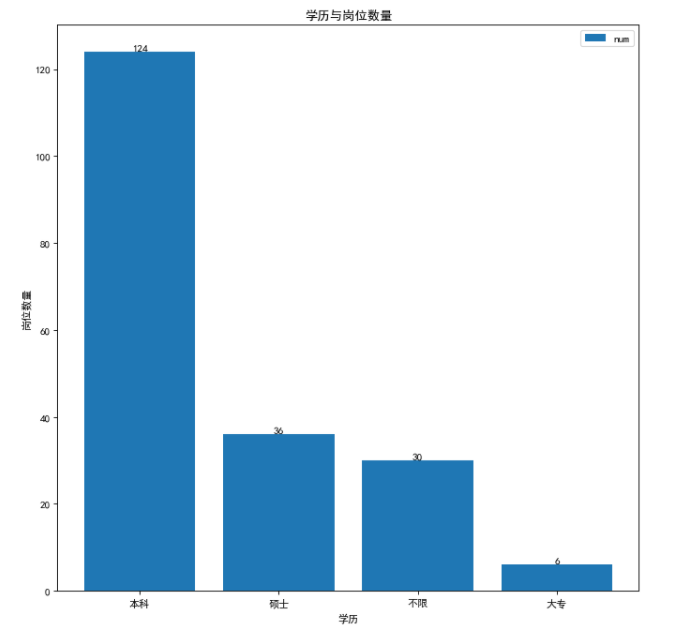


图4.3

仔细观察就会发现图4.3每个柱状都是用同一颜色填充的，为了更清晰、直观的观察图形，对每一个柱状进行用不同颜色填充处理，代码如下，如下图4.4所示。

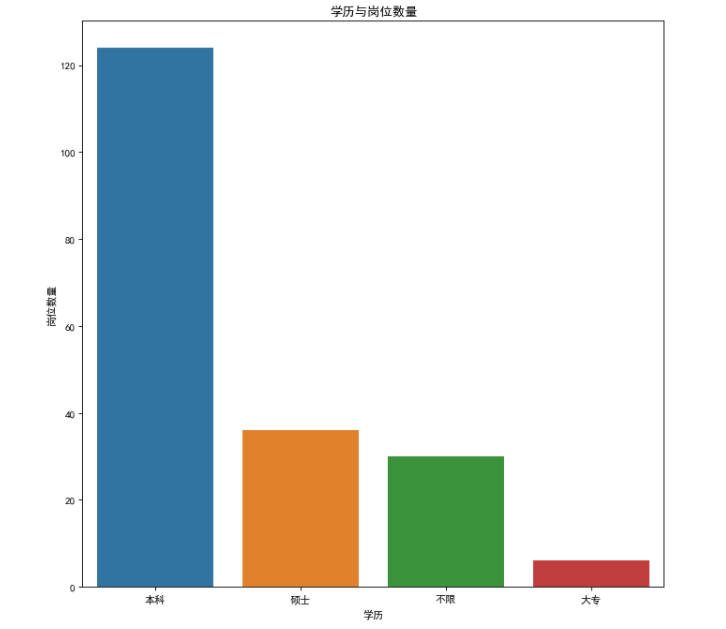


图4.4

结论: 数据不一定非常准确，但是，从图中还是可看，绝大多数岗位对实习生的学历要求是本科及以上，对学历要求是硕士及以上的岗位数量和对学历不做要求差不多，极少数岗位对学历的最低要求是大专。

## 4.2 城市与岗位数量

通过Python代码进行统计城市与岗位数量的关系，如图4.5所示。

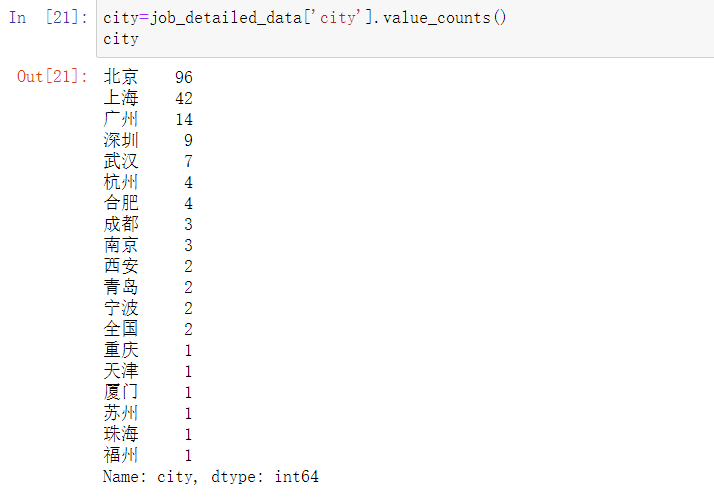
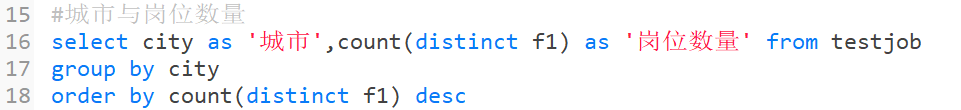


图4.5

通过Mysql数据库进行统计城市与岗位数量的关系，如图4.6所示。



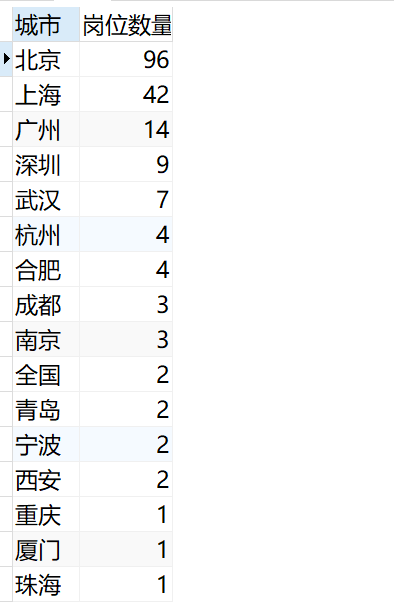


图4.6

通过Python代码实现城市与岗位数量的数据可视化，代码如下,数据可视化之后如图4.7所示。

|  |
| --- |
| plt.figure(figsize=(10, 10), dpi=80)  x = city[:15].index  y = city[:15].values  p = plt.bar(x, y, label="num")  for x,y in zip(x,y):  plt.text(x,y, '%.0f'%y, ha='center',va='bottom')  plt.xlabel('城市')  plt.ylabel('岗位数量')  plt.title('城市与岗位数量')  plt.legend(loc="upper right")  plt.show() |

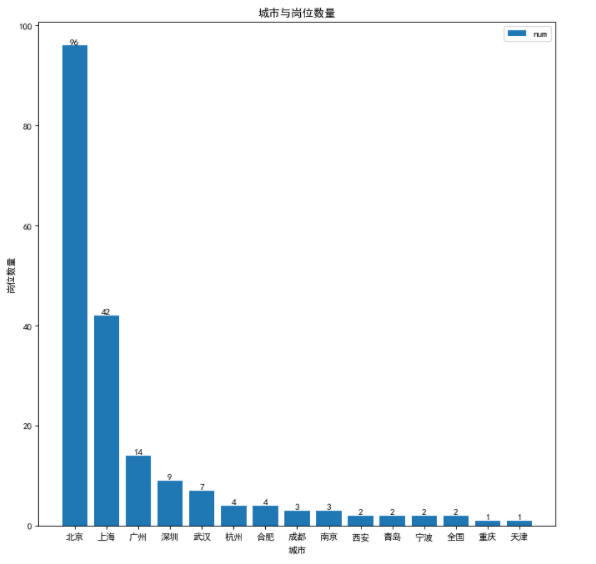


图4.7

仔细观察就会发现图4.7每个柱状都是用同一颜色填充的，为了更清晰、直观的观察图形，对每一个柱状进行用不同颜色填充处理，代码如下，如下图4.8所示。

|  |
| --- |
| plt.figure(figsize=(10, 10), dpi=80)  x = city[:15].index  y = city[:15].values  sns.barplot(x=x, y=y, data=job\_detailed\_data,capsize=.05)  plt.xlabel('城市')  plt.ylabel('岗位数量')  plt.title('城市与岗位数量')  plt.show() |

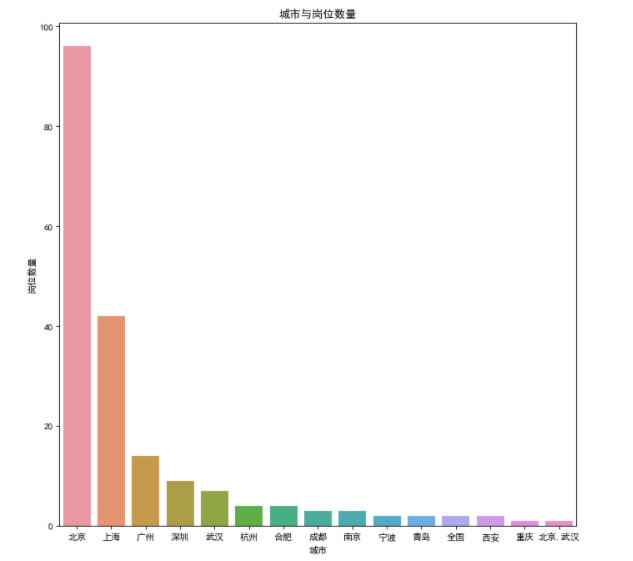


图4.8

结论: 一线城市(北京、上海、广州、深圳)特别是北京上海提供的岗位数量相比其他城市多一些,而且北京提供的岗位数量几乎占到这些城市提供的岗位数总和的一半。

## 4.3 工资分析

4.3.1不同城市、岗位，每日工资最高、最低情况

代码如下,结果如图4.9所示。

|  |
| --- |
| salary\_by\_city = job\_detailed\_data.groupby(['city','iname'])['maxsal','minsal'].mean()  salary\_by\_city |

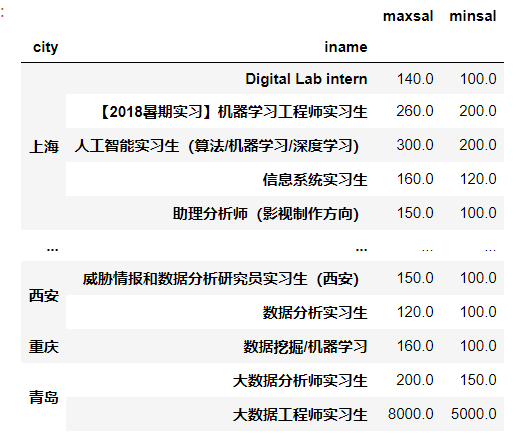


图4.9

### 4.3.2 平均日薪

由于数据中只给出每日的最高薪资和最低薪资，后面统计月薪需要平均日薪，这里需要计算平均日薪，其公式为平均日薪=(maxsal+minsal)/2，代码如下，结果如图4.10。

|  |
| --- |
| df1=job\_detailed\_data  df1['平均日薪']=(df1['maxsal']+df1['minsal'])/2  df1.head() |



图4.10

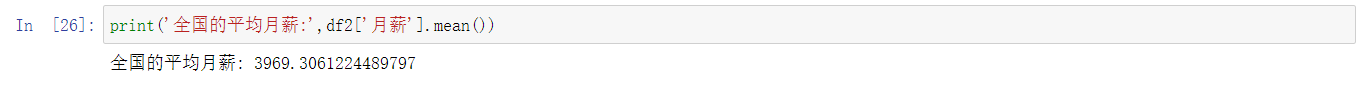
### 4.3.3 全国的平均月薪

由于每周工作时长是以天为单位，我们看月薪，其计算公式为月薪=每周工作时长X平均日薪X4，代码如下，结果如图4.11所示。

|  |
| --- |
| df2=df1  df2['月薪']=df2['day']\*df2['平均日薪']\*4  df2.head() |



图4.11



### 4.3.4 月薪与城市

不同城市，月薪情况是怎样的，代码如下，结果如图4.12。

|  |
| --- |
| salary\_by\_city = df2.groupby('city')['月薪'].mean()  salary\_by\_city |

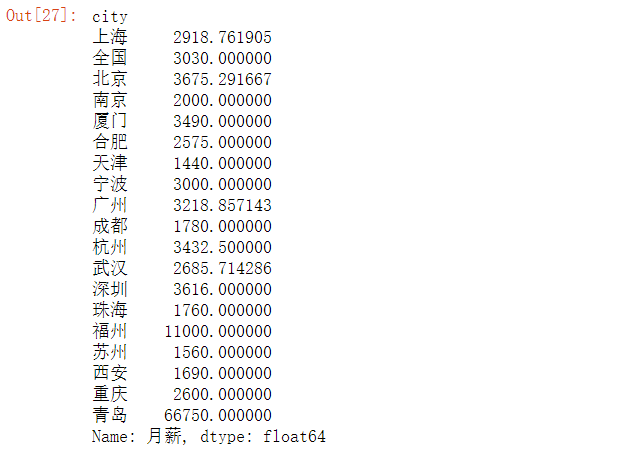


图4.12

### 4.3.5 不同城市岗位的占比

不同城市，岗位占比情况是怎样的(按降序排列)，代码如下，结果如图4.13。

|  |
| --- |
| city=job\_detailed\_data['city'].value\_counts()  city\_pct = (city/city.sum()).map(lambda x: '{:,.2%}'.format(x))  city\_pct |

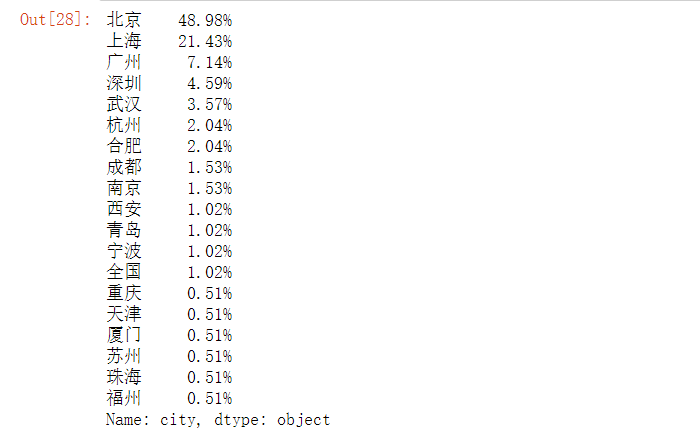


图4.13

### 4.3.6 月薪情况(降序排列)

全国各城市，月薪情况是怎样的，代码如下,结果如图4.14所示。

|  |
| --- |
| salary\_by\_city\_sort = salary\_by\_city[city\_pct.index].sort\_values(ascending=False)  salary\_by\_city\_sort |

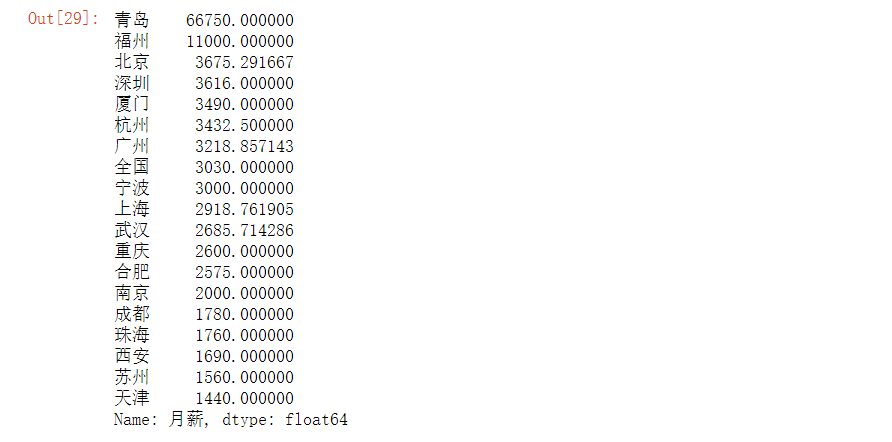


图4.14

统计完之后，在对全国各城市的平均月薪进行数据可视化分析，代码如下，结果如图4.15所示。

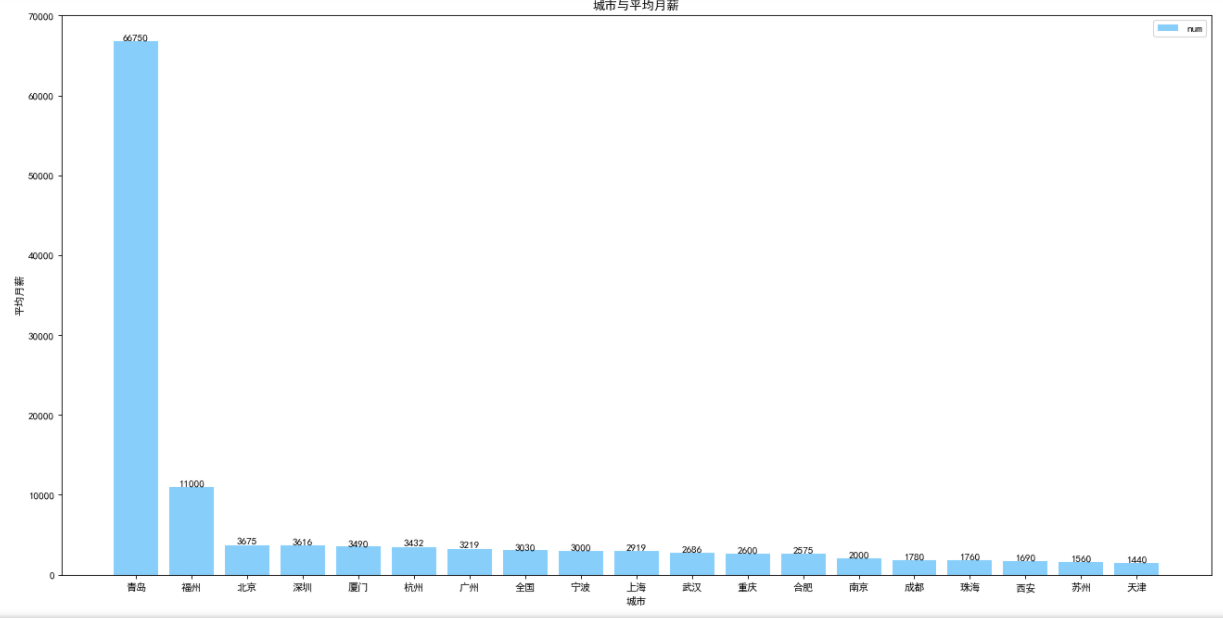


图4.15

## 4.4 公司规模分析

读取数据清洗、处理之后的数据文件test\_com.csv,代码如下，结果如图4.16所示。

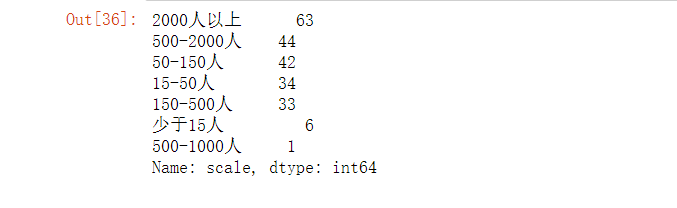
|  |
| --- |
| df3 = pd.read\_csv('E:\\test\_com.csv', delimiter=',')  df3 |



图4.16

对公司的规模进行统计，代码如下，结果如图4.17所示。

|  |
| --- |
| df3.loc[:,'scale'].value\_counts() |



对统计的结果进行数据可视化(饼图)，代码如下，结果如图4.18所示。

|  |
| --- |
| plt.figure(figsize=(6,9))  #定义饼状图的标签，标签是列表  labels = [u'大型企业',u'中型企业',u'小型企业']  #每个标签占多大，会自动去算百分比  sizes = [63,45,115]  colors = ['red','yellowgreen','lightskyblue']  #将某部分爆炸出来， 使用括号，将第一块分割出来，数值的大小是分割出来的与其他两块的间隙  explode = (0.05,0,0)    patches,l\_text,p\_text = plt.pie(sizes,explode=explode,labels=labels,colors=colors,  labeldistance = 1.1,autopct = '%3.1f%%',shadow = True,  startangle = 90,pctdistance = 0.6)  # 参数：  #labeldistance，文本的位置离远点有多远，1.1指1.1倍半径的位置  #autopct，圆里面的文本格式，%3.1f%%表示小数有三位，整数有一位的浮点数  #shadow，饼是否有阴影  #startangle，起始角度，0，表示从0开始逆时针转，为第一块。一般选择从90度开始比较好看  #pctdistance，百分比的text离圆心的距离  #patches, l\_texts, p\_texts，为了得到饼图的返回值，p\_texts饼图内部文本的，l\_texts饼图外label的文本    #改变文本的大小  #方法是把每一个text遍历。调用set\_size方法设置它的属性  for t in l\_text:  t.set\_size(20)  for t in p\_text:  t.set\_size(15)  # 设置x，y轴刻度一致，这样饼图才能是圆的  plt.axis('equal')  plt.title('数据分析岗位企业规模')  plt.legend()  plt.show() |

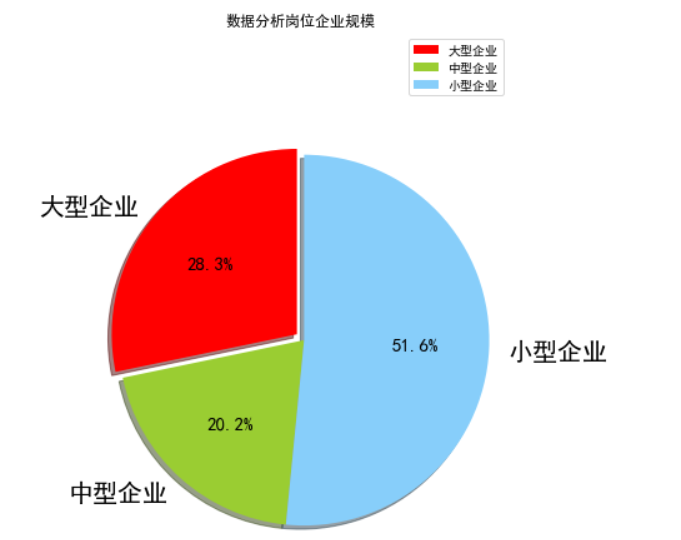
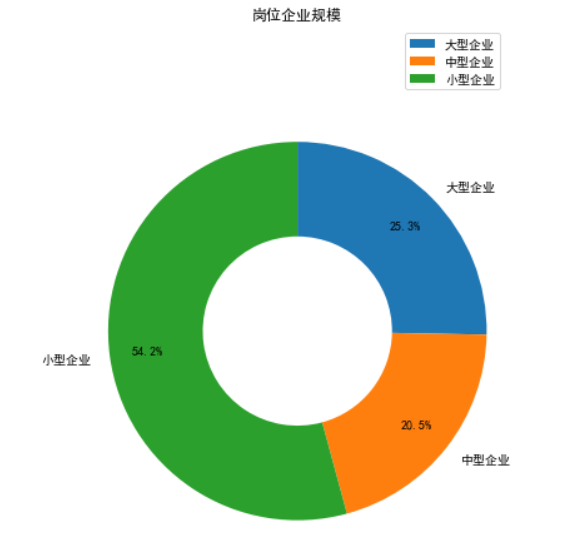


图4.18

对统计的结果进行数据可视化(扇形图)，代码如下，结果如图4.19所示。

|  |
| --- |
| plt.figure(figsize=(6,9))  labels = [u'大型企业',u'中型企业',u'小型企业']  #每个标签占多大，会自动去算百分比  sizes = [42,34,90]  plt.pie(sizes, labels=labels, startangle = 90,autopct = '%3.1f%%',pctdistance = 0.8,  counterclock = False, wedgeprops = {'width' : 0.5});  for t in l\_text:  t.set\_size(20)  for t in p\_text:  t.set\_size(15)  plt.axis('equal')  plt.title('岗位企业规模')  plt.legend()  plt.show() |



# 4.思考

此次课程设计主要是对之前学习的SQL及可视化图表的实践，真实的数据分析远不止这些，还需要扎实的教学、统计学基础，良好的数据敏感度，开拓但严谨的思维等，是一个终生学习的过程。

# 5.总结

进行本次Python课设的过程中，查阅了许多资料，在徐勇老师和同学的提点下，完成了这个实习岗位状况的数据分析。在这个过程中，我学到了很多，如怎样去全面的看待并分析一个问题等，同时也在总体上巩固了自己的Python知识，发现自己之前从来没有像现在一样将零散的知识点结合在一起，完成了这个实习岗位状况的数据分析后，让我在整体上又认识了一次Python。

在这次的实习岗位状况的数据分析中，有了平时学习的基础，完成本次Python课设的效率也因此大大提高，在设计过程中虽然遇到了一些问题，但经过多次的思考及反复的实践，检查出问题可能出现的原因，从而也暴露出了我在这方面的知识的欠缺和经验的不足，这使我意识到，自己学习还是只停留在表面，所以，我将努力在今后的时间里改正，并将所学的知识吃透、弄懂。

此外在这次课设报告的撰写过程中，我得到了许多人的帮助。

首先我要感谢徐勇老师给予我们的指导、提供给我们的支持和帮助，导师渊博的知识，严谨的治学态度，一丝不苟的工作作风，平易近人的性格都是我们学习的楷模。徐勇平时讲学很仔细，这学期对我们指导性建设要求提出了很多意见，促进了当代教育的发展。在课设及修正期间，导师给了我们很大的支持和鼓励，才使得课设得以顺利的完成，在此谨向导师表示忠心的感谢和崇高的敬意，这是我们能顺利完成这次课设的主要原因，在此期间，我们不仅学到了许多新的知识，而且也开阔了视野，提高了自己的设计能力。

其次，我要感谢帮助过我的同学，他们也为我解决了不少我不太明白的设计上的难题和一些我不太理解的语句。

Python课设的不足望老师谅解，今后我会更加努力地巩固所学的知识，希望自己能够学以致用，不辜负老师的教导。

# 参考文献

[1]刘宇宙.Python 3.5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，郝宁波，吴海涛，21天学通Python[M].北京:电子工业出版社，2016.

[4]基于Python的数据可视化[J].罗博炜.信息记录材料.2019(12)

[5]基于Python的招聘网站信息爬取与数据分析[J].王芳.信息技术与网络安全.2019(08)

# 课程设计成绩评定表

|  |  |  |  |
| --- | --- | --- | --- |
| 课程设计题目 | 实习岗位状况的数据分析 | | |
| 课程设计学生答辩或质疑记录：  如何对数据进行处理？  首先读取并筛选有效的数据，然后进行数据清洗(包括检查数据的一致性、有无缺失项、有无重复)，再把清洗好的数据另存为.csv文件，再将其导入Mysql数据库，观察数据，通过sql语句，对有问题的数据进行处理。  在做的过程中，应该注意什么？  在做的过程中，应尽可能多地观察数据，发现有问题的数据及时进行处理，否则  之后进行分析实现数据可视化的图形很容易产生歧义。  在做的过程中遇到了哪些问题，如何解决的？  在做的过程中对数据进行清洗之后，发现.csv文件仍然没有导入到Mysql数据库里，之后意识到没有把清洗之后的数据转存为.csv文件，导致.csv文件的数据并没有改变。 | | | |
| **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 | | | |