|  |  |
| --- | --- |
| 学号 | 10214518330 |



课 程 设 计

课程名称 python程序设计课程设计

|  |  |
| --- | --- |
| 题 目 | 实习岗位状况的数据分析 |
| 专 业 | \_\_\_\_\_\_软件工程\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 班 级 | \_\_\_\_\_\_软件1182\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 姓 名 | \_\_\_\_\_\_\_李欣\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 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 设计题目 1](#_Toc10783)

[2 开发环境 1](#_Toc17306)

[3 分析思想 1](#_Toc6725)

[4 主要分析目标 1](#_Toc10081)

[5 分析步骤 2](#_Toc7528)

[5.1 csv文件信息读取 2](#_Toc12593)

[5.2 数据清洗 3](#_Toc27238)

[6 数据分析 5](#_Toc7814)

[6.1 学历与岗位数据分析 5](#_Toc27291)

[6.2 城市与岗位数据分析 7](#_Toc30189)

[6.3 工资数据分析 9](#_Toc19065)

[6.4 公司规模数据分析 12](#_Toc22975)

[7 设计体会 15](#_Toc29657)

[参 考 文 献 16](#_Toc16823)

# 1 设计题目

实习岗位状况的数据分析

# 2 项目工具

项目托管平台：GitHub

编码工具：jupyter

编码语言：python

# 3 需求分析

选题背景：1.大学毕业前夕面临找工作和实习就业的选择，通过本次python课设可以对实习岗位进行大概的了解和分析，更好的为后面实习就业做准备。2.通过数据的分析和处理可以明确的凸显实习岗位受影响的条件和因素，让我们尽量做出正确的选择。

通过对实习岗位状况的数据分析，分析自身所处行业的用人要求与岗位的需求量，不同城市的岗位需求量，不同城市能够提供的薪资水平，大中小型企业对数据分析岗位的需求情况。

分析目标：1.学历与岗位需求量分析

2.城市与岗位需求量分析

3.薪资水平分析

4.公司规模分析

# 4分析思想

1. 将已有的实习岗位信息数据进行数据读取并去重。

2. 处理数据，筛选出后面分析需要的数据。

3. 对处理好的数据进行统计分析

4. 实现数据的可视化展示

# 5 分析步骤

## 5.1 csv文件信息读取

读取文件信息并去重，并保存成DataFrame格式;读取有效字段的数据；具体代码如下：

In [22]: import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import matplotlib as mpl

%matplotlib inline

import seaborn as sns

from pyecharts.charts import Bar

from pyecharts import options as opts

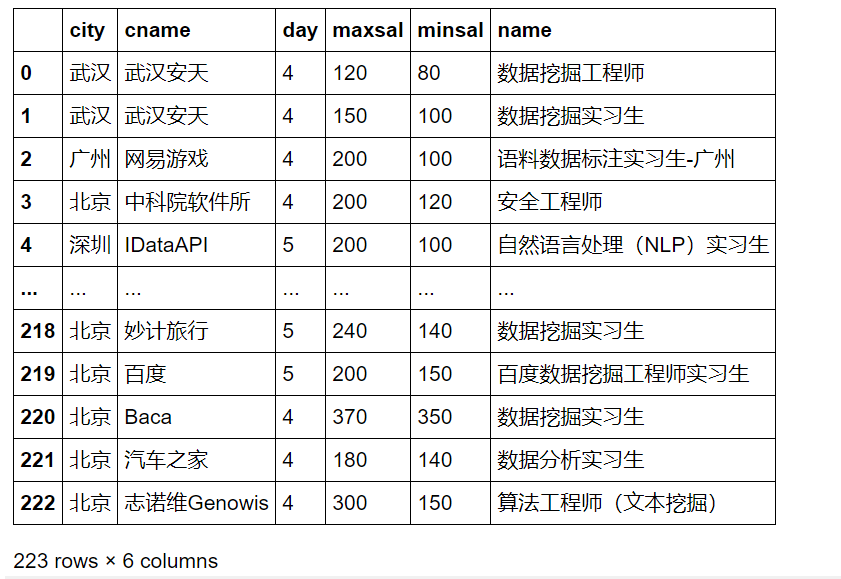
#设置中文标签的显示

plt.rcParams['font.sans-serif']=['SimHei']#设置显示中文字体

In [23]: #读取job\_list.csv中的有效字段的数据

job\_list\_data = pd.read\_csv('job\_list.csv')[['city','cname','day','maxsal','minsal','name']]

job\_list\_data



In [24]: #读取job\_detailed.csv中有效字段的数据

job\_detailed\_data=pd.read\_csv('job\_detailed.csv')[['city','cname','day','degree','iname','industry','maxsal','minsal','month','attraction']]

job\_detailed\_data



## 5.2 数据清洗

对数据进行去重，并生成新的索引删除原索引；提取需要的字段并去空；查看数据规模以及是否有缺失项；最后将清洗之后的数据转存为.csv文件，导入数据库进行处理。具体代码如下：

In [25]: print("原始行数:",len(job\_detailed\_data))

原始行数: 223

In [26]: #去重，并生成新索引删除原索引

job\_detailed\_data=job\_detailed\_data.drop\_duplicates().reset\_index().drop('index',axis=1)

print('现在数据行数:',len(job\_detailed\_data))

job\_detailed\_data



In [27]: #提取需要的字段并去空

job\_detailed\_data=job\_detailed\_data[['city','cname','day','degree','iname','industry','maxsal','minsal','month','attraction']]

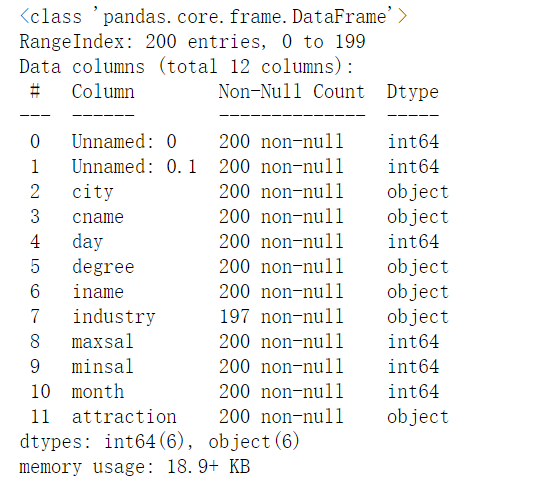
job\_detailed\_data.head()

print('现在数据行数:',len(job\_detailed\_data))

现在数据行数: 200

In [18]: #看数据规模以及是否有缺失项

job\_detailed\_data.info()

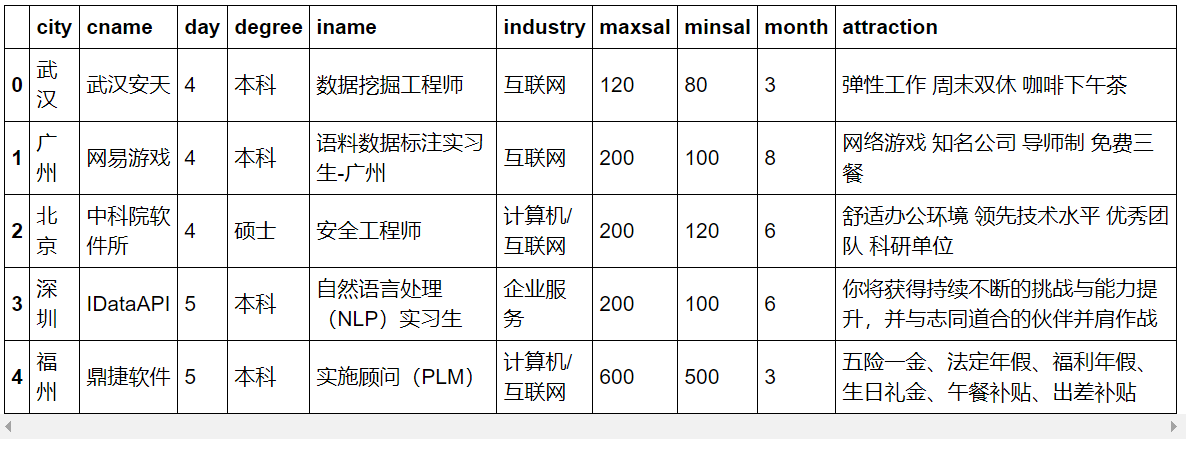


把清洗之后的数据转存为.csv文件,导入Mysql数据库对数据进行处理之后，再读取修改之后的.csv文件

In [33]: #job\_detailed\_data.to\_csv('E:\\testcsv.csv',encoding='utf-8')

job\_detailed\_data=pd.read\_csv('E:\\mod\_testcsv.csv')[['city','cname','day','degree','iname','industry','maxsal','minsal','month','attraction']]

job\_detailed\_data.head()



# 6 数据分析

## 6.1 学历与岗位数据分析

探究岗位与相关学历是否有要求或联系

In [34]: degree=job\_detailed\_data['degree'].value\_counts()

degree

Out[34]:

本科 124

硕士 36

不限 30

大专 6

Name: degree, dtype: int64

In [35]: 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')

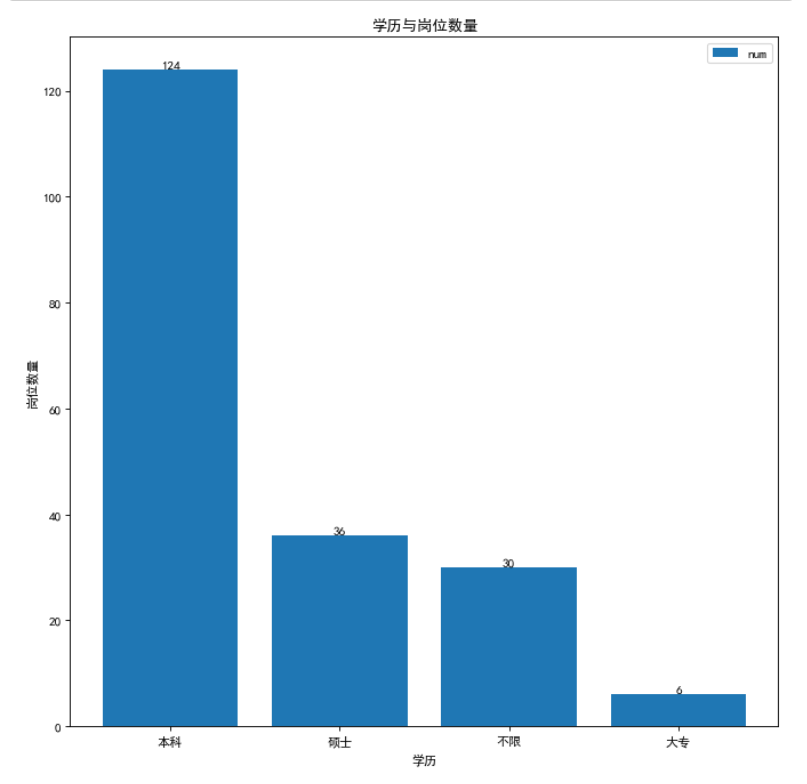
plt.xlabel('学历')

plt.ylabel('岗位数量')

plt.title('学历与岗位数量')

plt.legend(loc="upper right")

plt.show()



In [61]: plt.figure(figsize=(10, 10), dpi=80)

x = degree[:15].index

y = degree[:15].values

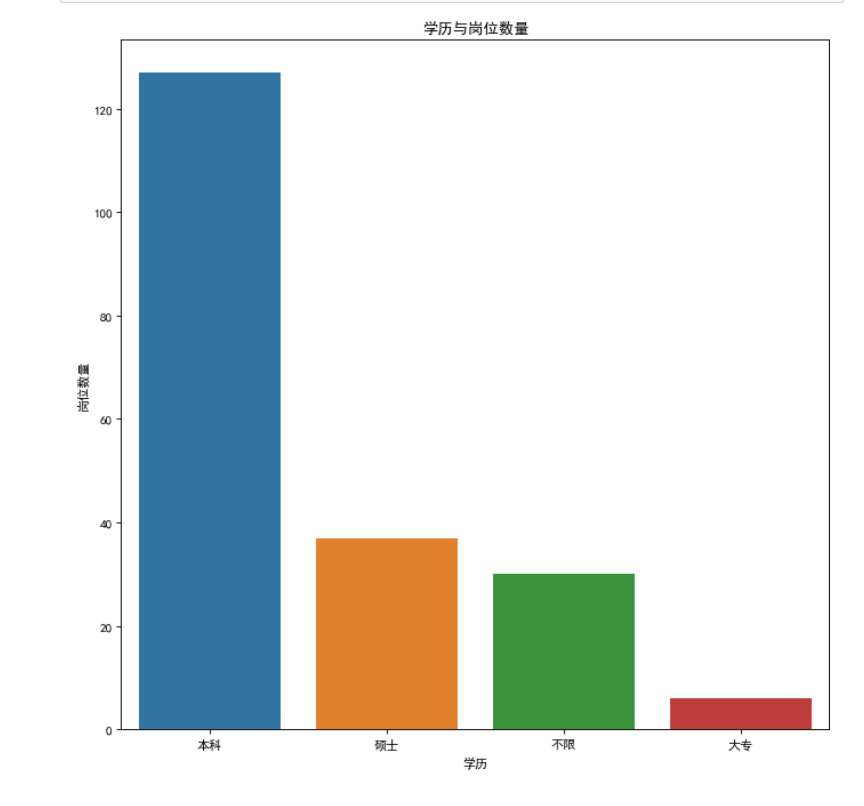
sns.barplot(x=x, y=y, data=job\_detailed\_data,capsize=.05)

plt.xlabel('学历')

plt.ylabel('岗位数量')

plt.title('学历与岗位数量')

plt.show()



结论

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。

## 6.2 城市与岗位数据分析

In [36]: city=job\_detailed\_data['city'].value\_counts()

city

Out[36]: 北京 96

上海 42

广州 14

深圳 9

武汉 7

合肥 4

杭州 4

成都 3

南京 3

全国 2

青岛 2

西安 2

宁波 2

福州 1

天津 1

重庆 1

珠海 1

厦门 1

苏州 1

Name: city, dtype: int64

In [37]: 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')

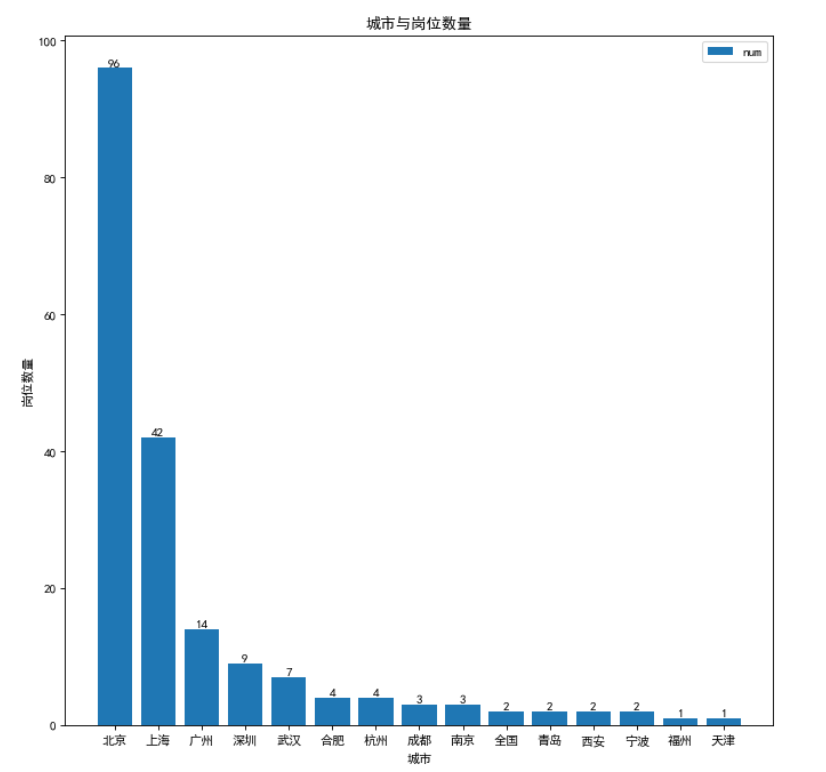
plt.xlabel('城市')

plt.ylabel('岗位数量')

plt.title('城市与岗位数量')

plt.legend(loc="upper right")

plt.show()



In [65]: plt.figure(figsize=(10, 10), dpi=80)

x = city[:15].index

y = city[:15].values

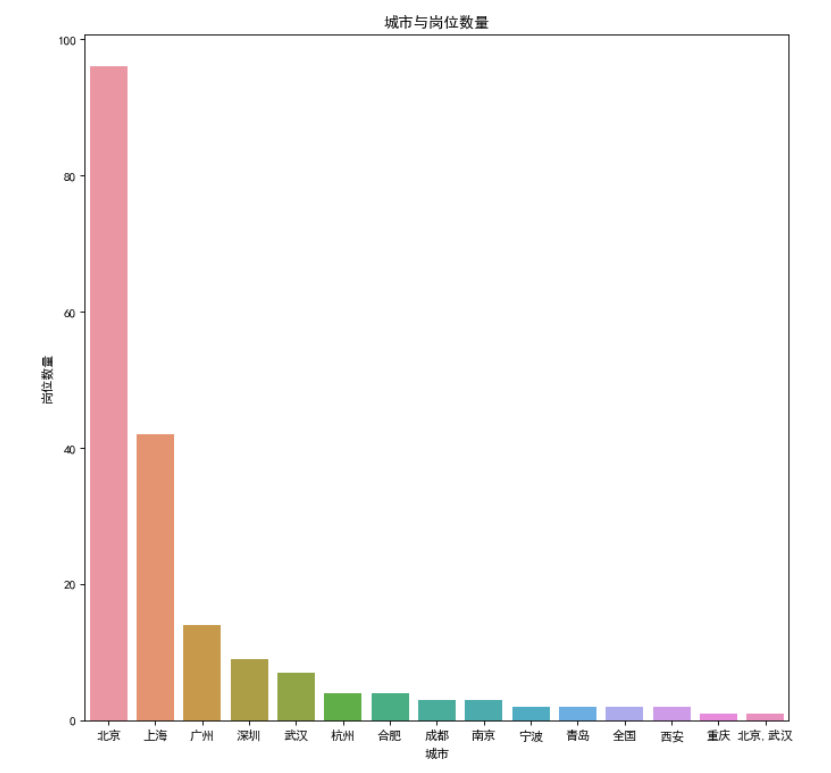
sns.barplot(x=x, y=y, data=job\_detailed\_data,capsize=.05)

plt.xlabel('城市')

plt.ylabel('岗位数量')

plt.title('城市与岗位数量')

plt.show()

  
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。

## 6.3 工资数据分析

In [45]: #maxsal=job\_detailed\_data['maxsal'].value\_counts()

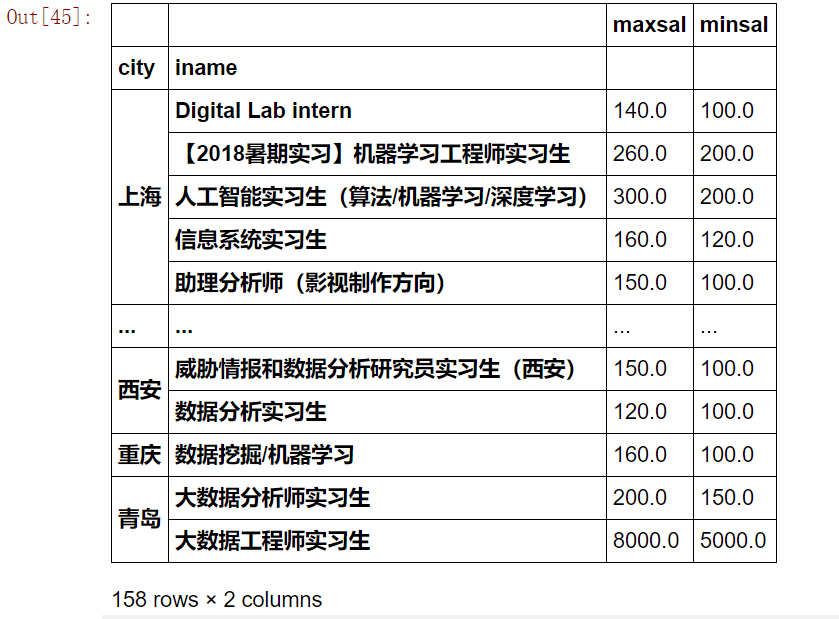
#maxsal

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

salary\_by\_city

<ipython-input-45-c168b863a684>:3: FutureWarning: Indexing with multiple keys (implicitly converted to a tuple of keys) will be deprecated, use a list instead.

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

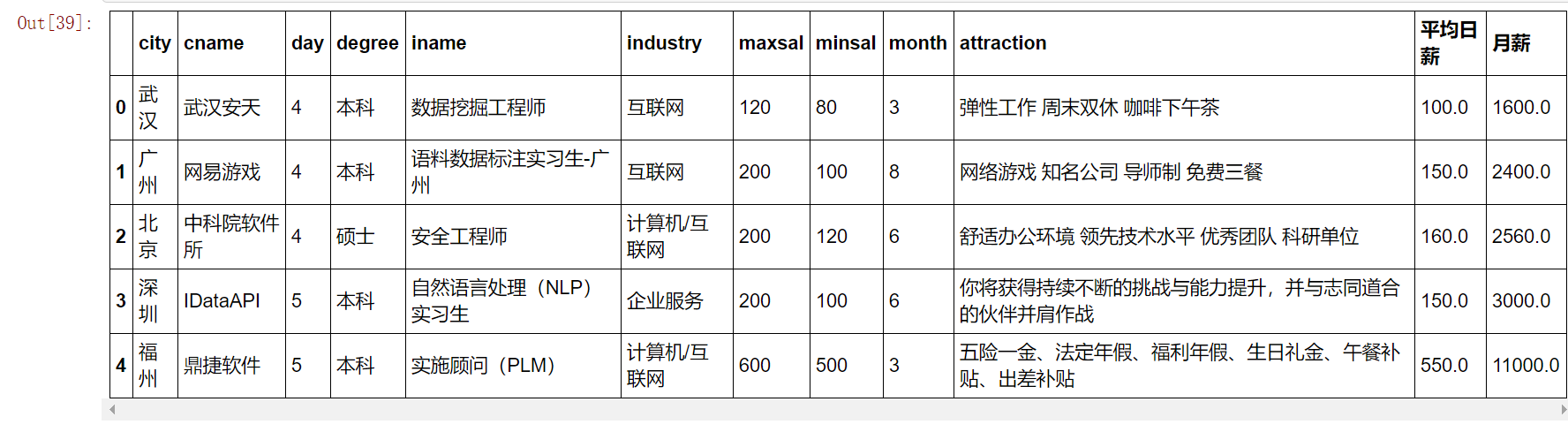


平均日薪=(maxsal+minsal)/2

In [38]: df1=job\_detailed\_data

df1['平均日薪']=(df1['maxsal']+df1['minsal'])/2

df1.head()



### 全国的平均月薪

由于每周工作的时长不同 我们看月薪 每周工作时长X平均日薪X4

In [39]: df2=df1

df2['月薪']=df2['day']\*df2['平均日薪']\*4

df2.head()



In [40]: print('全国的平均月薪:',df2['月薪'].mean())

全国的平均月薪: 3969.3061224489797

月薪与城市

In [41]: salary\_by\_city = df2.groupby('city')['月薪'].mean()

salary\_by\_city

Out[41]: city

上海 2918.761905

全国 3030.000000

北京 3675.291667

南京 2000.000000

厦门 3490.000000

合肥 2575.000000

天津 1440.000000

宁波 3000.000000

广州 3218.857143

成都 1780.000000

杭州 3432.500000

武汉 2685.714286

深圳 3616.000000

珠海 1760.000000

福州 11000.000000

苏州 1560.000000

西安 1690.000000

重庆 2600.000000

青岛 66750.000000

Name: 月薪, dtype: float64

In [42]: city=job\_detailed\_data['city'].value\_counts()

city\_pct = (city/city.sum()).map(lambda x: '{:,.2%}'.format(x))

city\_pct

Out[42]: 北京 48.98%

上海 21.43%

广州 7.14%

深圳 4.59%

武汉 3.57%

合肥 2.04%

杭州 2.04%

成都 1.53%

南京 1.53%

全国 1.02%

青岛 1.02%

西安 1.02%

宁波 1.02%

福州 0.51%

天津 0.51%

重庆 0.51%

珠海 0.51%

厦门 0.51%

苏州 0.51%

Name: city, dtype: object

In [43]:s alary\_by\_city\_sort = salary\_by\_city[city\_pct.index].sort\_values(ascending=False)

salary\_by\_city\_sort

Out[43]: 青岛 66750.000000

福州 11000.000000

北京 3675.291667

深圳 3616.000000

厦门 3490.000000

杭州 3432.500000

广州 3218.857143

全国 3030.000000

宁波 3000.000000

上海 2918.761905

武汉 2685.714286

重庆 2600.000000

合肥 2575.000000

南京 2000.000000

成都 1780.000000

珠海 1760.000000

西安 1690.000000

苏州 1560.000000

天津 1440.000000

Name: 月薪, dtype: float64

In [50]: plt.figure(figsize=(20, 10), dpi=80)

x = salary\_by\_city\_sort.index

y = salary\_by\_city\_sort.values

p = plt.bar(x, y, label="num", color="#87CEFA")

for x,y in zip(x,y):

plt.text(x,y, '%.0f'%y, ha='center',va='bottom')

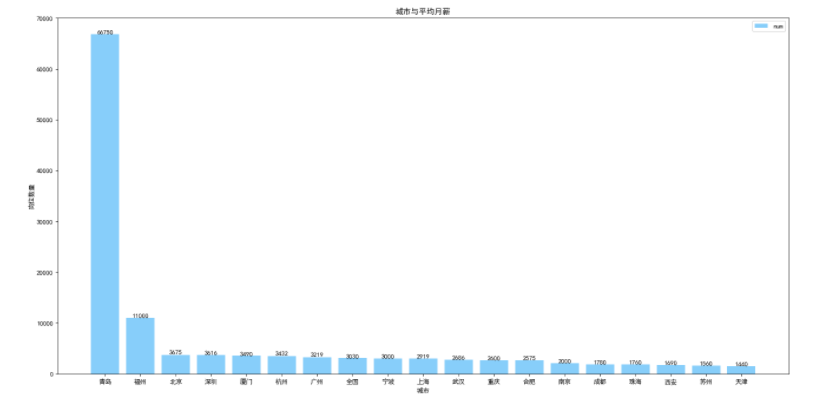
plt.xlabel('城市')

plt.ylabel('平均月薪')

plt.title('城市与平均月薪')

plt.legend(loc="upper right")

plt.show()



不同城市、岗位，每日工资最高、最低情况

In [23]: #maxsal=job\_detailed\_data['maxsal'].value\_counts()

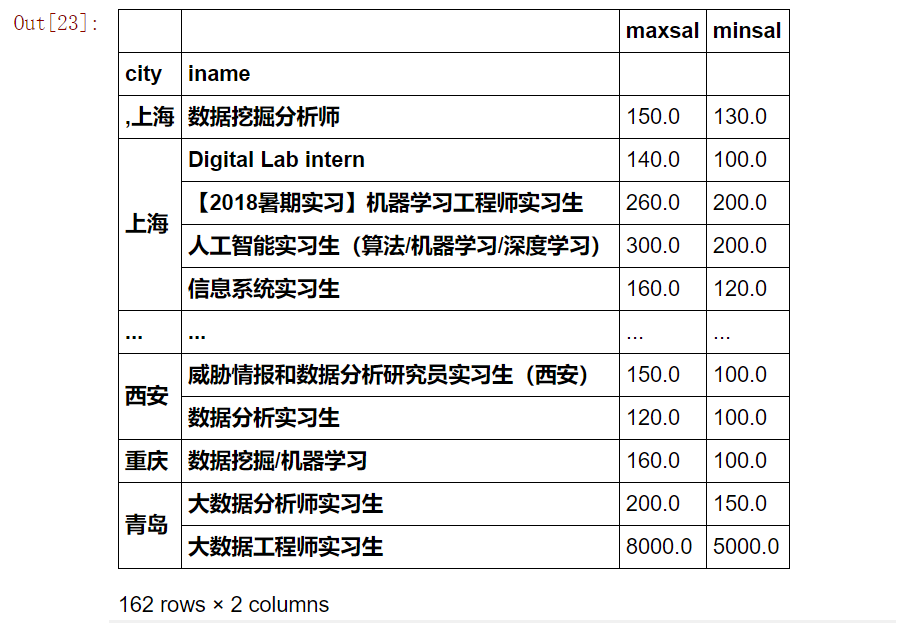
#maxsal

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

salary\_by\_city

<ipython-input-23-c168b863a684>:3: FutureWarning: Indexing with multiple keys (implicitly converted to a tuple of keys) will be deprecated, use a list instead.

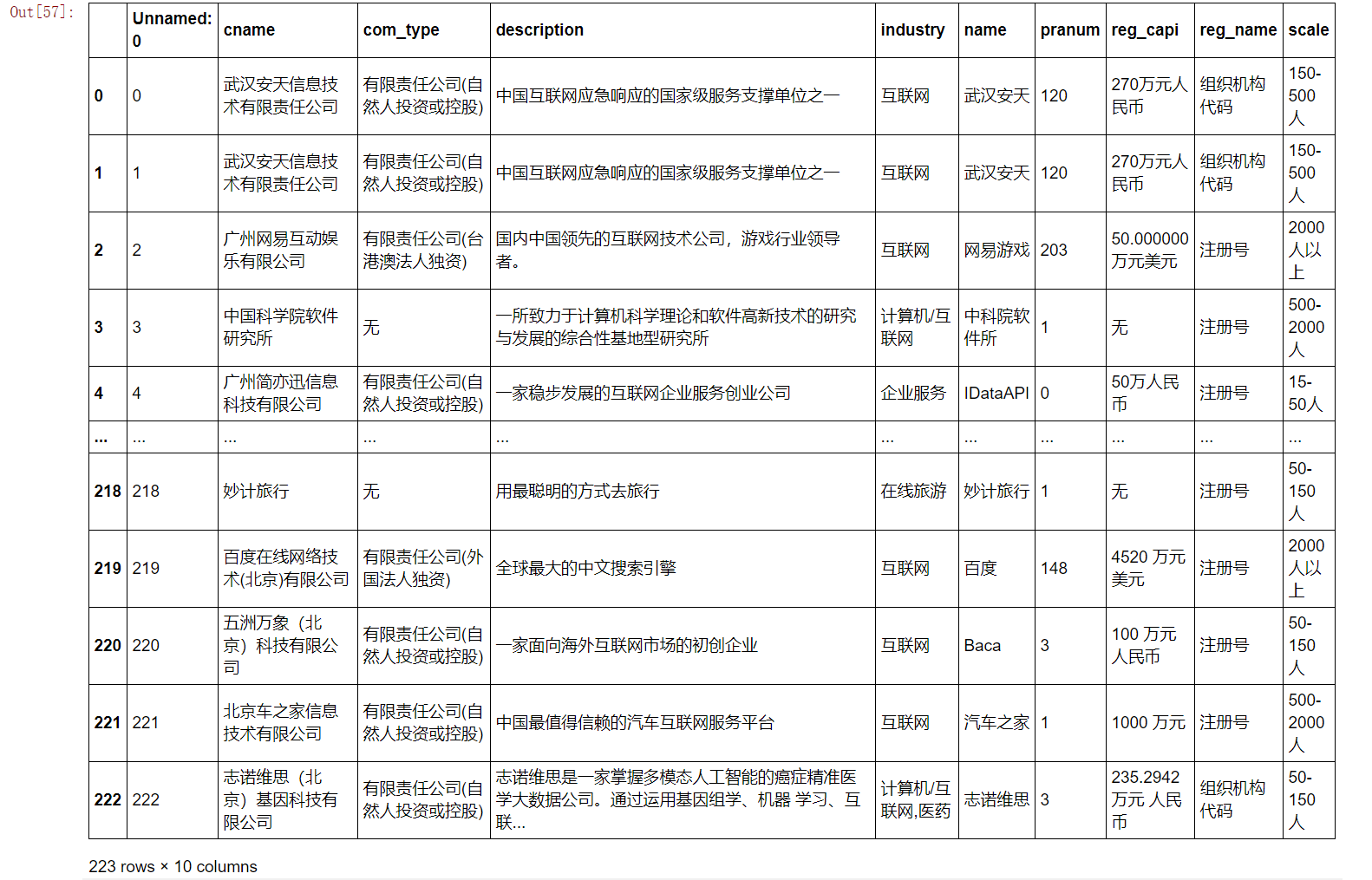
salary\_by\_city = job\_detailed\_data.groupby(['city','iname'])['maxsal','minsal'].mean()



## 6.4 公司规模数据分析

In [57]: df3 = pd.read\_csv('E:\\test\_com.csv', delimiter=',')

df3



In [ ]: print('原始数据行')

In [58]: df3.loc[:,'scale'].value\_counts()

Out[58]: 2000人以上 63

500-2000人 44

50-150人 42

15-50人 34

150-500人 33

少于15人 6

500-1000人 1

Name: scale, dtype: int64

大型企业：63 中型企业：45 小型企业：115

In [74]: 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)

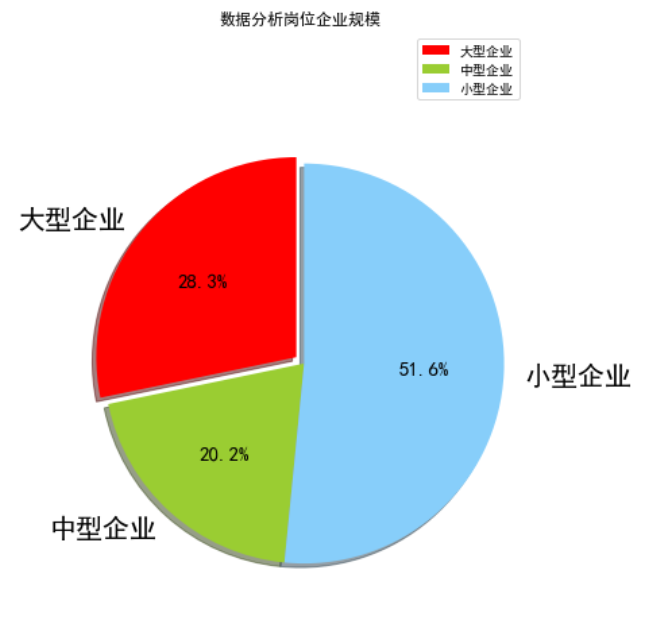
# 设置x，y轴刻度一致，这样饼图才能是圆的

plt.axis('equal')

plt.title('数据分析岗位企业规模')

plt.legend()

plt.show()



In [75]: 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 = [63,45,115]

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:

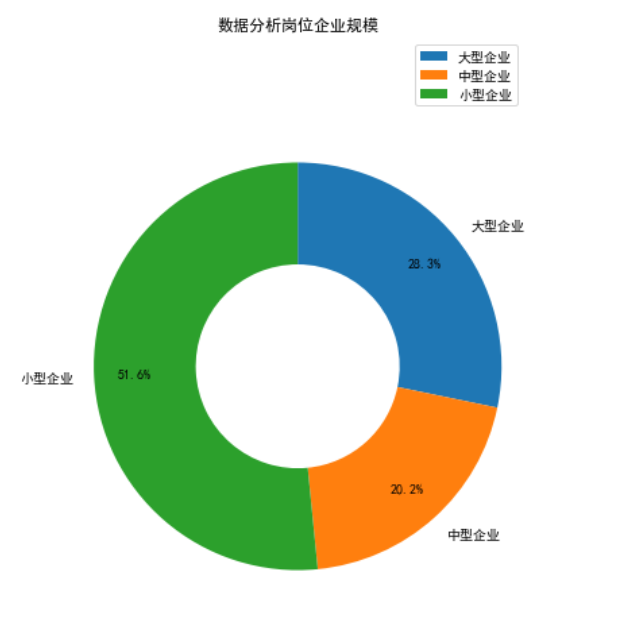
t.set\_size(15)

plt.axis('equal')

plt.title('数据分析岗位企业规模')

plt.legend()

plt.show()



# 7 设计体会

选择这个课题的时候，经过本学期的python学习，我从最初的Python小白到现在能简单运用python语言实现一些简单案例和进行数据分析，其中的过程虽然艰辛但也很欣慰，艰辛的是刚接触的时候觉得好难，在加上之前了解不够深入，以为python是一门很难的语言，所以从刚开始学的时候就十分在意，生怕掉队，欣慰的是经过一段时间的学习，我发现对于这门课程的学习，我有了新的认识。

python最大的优点就是有很多已经存在的第三方库，这是任何一个编程软件都无法代替的，正是第三方库的存在使得编程更简单，让我有了信心在这个实习岗位数据分析的过程中，我发现python的实用性巨大，且潜力无限，利用简单的库就能进行很完善的数据分析，相对于其他语言，python优势非常明显，短短的代码可以蕴藏无限的奥义，让人不得不对python的未来充满期待！

最后感谢老师的辛勤授课及陪伴，也希望日后自己在学习python的路途上能学有所成，在数据分析这方面能有所成就。

# 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：  1.如何对数据进行处理 答：读取并筛选有效的数据，然后进行数据清洗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。 2.数据分析采用了哪些库 答：采用了numpy、pandas、seaborn、pyecharts库  3.在做的过程中遇到哪些问题，是如何解决的？  答：在数据清洗的时候遇到了数据不完整、数据值重复、数据无用这些问题。数据不完整进行了人工手动填入，或者对无法补全的数据进行剔除；数据值重复使用sql去除重复记录；对无用的数据进行剔除。 | | | |
| **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 | | | |