

**本科生毕业论文(设计)**

题目 **招聘岗位分析**

院 (系) **信息科学与技术学院**

专 业  软件工程

班 级  **f200152**

学 号  **f20015227**

学生姓名  **严文振**

指导教师及职称 **张青**

定稿日期： 2023 年 5 月 16 日

目 录

摘要

第一章 数据采集 ····························································5

1.1 数据获取·······························································5

1.2 导入数据集 ····························································5

1.3 本章小结 ······························································5

第二章 数据清洗·····························································6

2.1 数据清洗概述 ··························································6

2.2 去除重复数据 ··························································6

2.3 去除带有空值的数据 ····················································6

2.4 确认清洗完的数据的数据 ················································7

2.5 本章小结 ······························································7

第三章 数据建模·····························································8

3.1 数据建模概述 ··························································8

3.2 计算薪资范围的最低值和最高值 ··········································8

3.3 计算每月发薪资次数的平均值·············································8

3.3 计算平均薪资···························································9

3.4 本章小结·······························································9

第四章 数据可视化··························································10

4.1 数据可视化概述·························································11

4.2 岗位类型与薪资·························································11

4.2.1 查看岗位类型······················································12

4.2.2 岗位类型与薪资柱状图··············································12

4.2.3 岗位类型与薪资柱状图分析··········································12

4.2.4 岗位类型与薪资箱线图··············································12

4.2.5 岗位类型与薪资箱线图分析··········································13

4.3 岗位数与薪资··························································13

4.3.1 岗位数与薪资直方图················································14

4.3.2 岗位数与薪资直方图分析·········································14 4.3.3 岗位数与薪资回归预测···············································14

4.3.4 岗位数与薪资聚类分析···············································14

4.3.5 岗位数与薪资决策树·················································15

4.4 学历与薪资·····························································15

4.4.1 查看学历类型·······················································15

4.4.2 学历与薪资柱状图···················································15

4.4.3 学历与薪资柱状图分析···············································16

4.5 各学历占比·····························································16

4.5.1 学历饼图···························································16

4.5.1 学历饼图分析·······················································16

4.6 工资经验与薪资······················································17

4.6.1 查看工资经验类型··················································17

4. 6. 2 工资经验与薪资柱状图··············································17

4.6.3 工资经验与薪资柱状图分析··········································18

4.7 各工作经验占比························································18

4.7.1 工资经验饼图······················································19

4.7.2 工资经验饼图分析··················································19

4.8 城市与薪资····························································20

4.8.1 查看城市类型······················································20

4.8.2 城市与薪资柱状图··················································21

4.8.3 城市与薪资柱状图分析··············································21

4.8.4 城市与薪资箱线图···················································22

4.8.5 城市与薪资箱线图分析···············································23

第五章 总结································································24

摘要

随着就业压力的不断增大，毕业就能找到工作变成了许多人的奢望。找工作 + 高薪资是许多人一生奋斗的目标，而工作中最为关注的薪资往往与所在城市、学历、工作经验以及工作类型密切相关。所以这个招聘岗位分析将选择当下待遇相对比较好的几个工作类型从城市、学历、工作经验不同角度分别分析与薪资的关系。

关键字： 薪资、城市、学历、工作经验、工作类型

第一章 数据采集

1.1 数据获取

数据获取来源于Boss直聘、前程无忧、智联招聘等各大招聘网站拼凑而成。

1.2 导入数据集

采用Pandas库中的read\_excel()函数导入存放数据的excel表格。

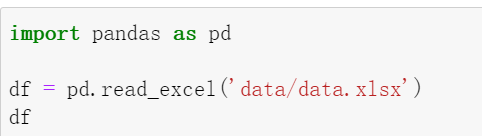


图1-1 导入数据集代码

根据图1-2可知，数据的属性一共有type，name，company，salary，city，experience，edu组成，分别对应着招聘岗位的工作类型，工作名称，所在公司，薪资，城市，工作经验和学历。



图1-2 导入后的数据表

1.3 本章小结

招聘岗位分析的第一步是读取存放数据的excel表格，并且查看数据集的各种属性，事先计划好后面的分析方向。属性一共有工作类型，工作名称，所在公司，薪资，城市，工作经验和学历。

其中可以分别进行对工作类型、城市、工作经验和学历对薪资的影响进行分析，以此来确认工作类型、城市、工作经验和学历哪个因素对薪资影响的权重最大。

第二章 数据清洗

2.1 数据清洗概述

导入的数据并不一定是理想的数据，可能带有重复的数据或者是缺失的数据，所以需要进行数据清洗，去除重复的数据和带有空值的数据，防止干扰到之后的分析。

2.2 去除重复数据

采用drop\_duplicates()函数去除重复数据。



图2-2 去除重复数据

2.3 去除带有空值的数据

采用dropna含糊去除任意属性带有至少一个空值的数据。

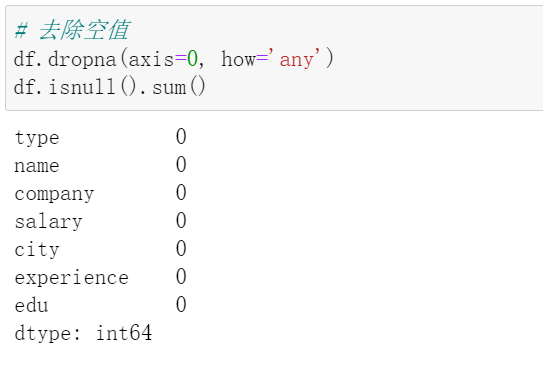


图2-3 去除带有空值的数据

2.4 确认清洗完的数据的数据

清洗完数据需要使用info()函数确认清洗完的数据是否符合要求。 根据图2-3可知全部的数据都为非空数据，符合分析的要求。

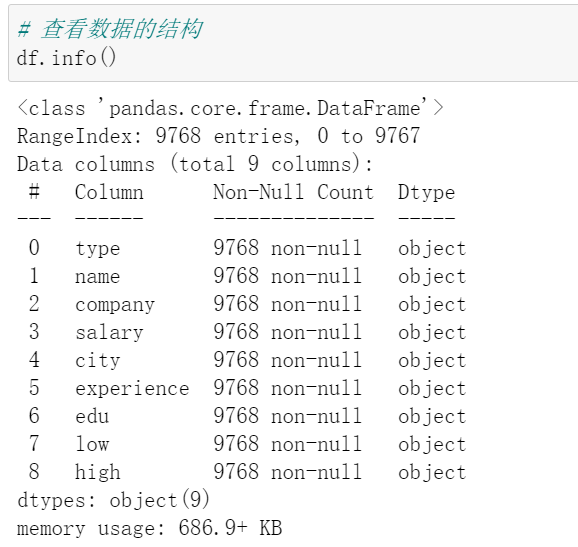


图2-3 清洗完的数据

2.5 本章小结

从网上获取的数据很多情况下并不是非常理想的数据，往往可能或有多个数据缺少必要的数值或者是相同的数值，如果不进行数据清洗工作，在数据分析的过程中代码可能会频繁报错，或者造成最终结果可能不正确。所以数据清洗必不可少。

确认数据清洗完成，具备进行分析的条件后，才可以进行数据分析的工作。

第三章 数据建模

3.1 数据建模概述

因为数据中的薪资表示的是一个范围，所以为了方便分析，需要计算出薪资的平均值。

3.2 计算薪资范围的最低值和最高值

这里采用pandas中的str.extract()函数截取需要的数字。

正则表的式str.extract('(\d+)')中的()表示需要截取的内容的位置，\d+表示截取多个数字。(\d+)表示的是截取从一开始到后面非数字的全部数字，例如数据20-40k，此处取出的就是20。-(\d+)表示的是取出范围符号‘ - ’后的全部数字，例如数据20-40k，此处取出的就是40。

如果薪资为固定的数字，例如20，则-(\d+)则取不到任何值，因为没有范围符号‘ - ’，所以直接用低值填充，这样将来计算平均值的时候还是原来本身。

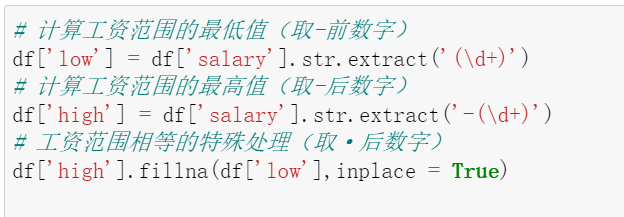


图3-1 计算薪资范围的最低值和最高值代码

3.3 计算每月发薪资次数的平均值

并不是所有薪资都是每个月发一次的，例如20-40K·17薪表示一年发17次工资，所以需要计算平均每个月发k次工资，其中k为x薪与12月的比值。

计算步骤是先利用函数str.extract()取出每年发x次工资，之后全部除12计算出每个月发薪资次数的平均值。

str.extract(·(\d{2}))表示取·后的全部数字，例如20-40K·17薪，取出的为结果17。如果数据是20-40K则默认每个月发一次工资， 但是缺少符号·将导致最终计算的结果为空值，所以全部用12填充。由于取出的数据全部是字符串，并且是先计算一年发工资的次数，所以最终需要转换为数字类型再除12计算出每个月发薪资的次数的平均值。

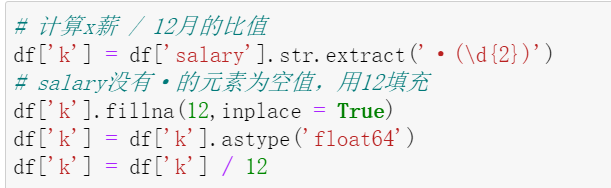


图3-2 计算每月发薪资次数的平均值代码

3.3 计算平均薪资

计算公式： 平均薪资 = （最低值 + 最高值）/ 2 \* （x薪 / 12）

avg =（low + high）/ 2 \* k

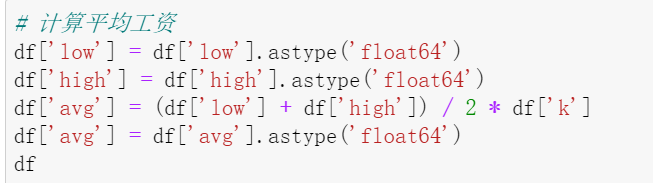


图3-3 计算平均薪资代码

输入代码可计算出平均薪资：



图3-4 计算平均薪资结果

3.4 本章小结

逐步计算出范围的最低薪资和最高薪资后，再求出每月的平均发工资次数，就可以去计算薪资的平均值。薪资的平均值计算完成后，就可以进行数据的可视化分析。

第四章 数据可视化

4.1 数据可视化概述

数据的可视化能够更加直观的分析大量数据的相互之间的关系，能够帮助我们更快速和准确的得出我们想要的答案。

导入的数据集一共有工作类型，工作名称，所在公司，薪资，城市，工作经验和学历这些属性，而分析的目的最终是围绕着薪资展开的，所以接下来将逐步分析工作类型、工作经验、学历以及城市分别和薪资的关系。

4.2 岗位类型与薪资

4.2.1 查看岗位类型

数据一共包括4种不同类型的岗位，分别为产品经理、java、数据分析和python。

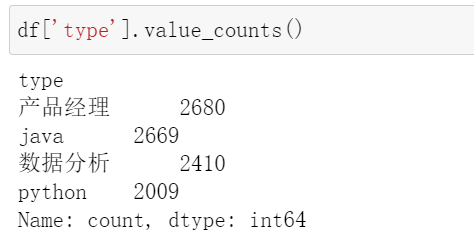


图4-1 查看岗位类型

4.2.2 岗位类型与薪资柱状图

岗位与薪资柱状图代码如下：



图4-2 岗位与薪资柱状图代码

岗位与薪资柱状图如下：

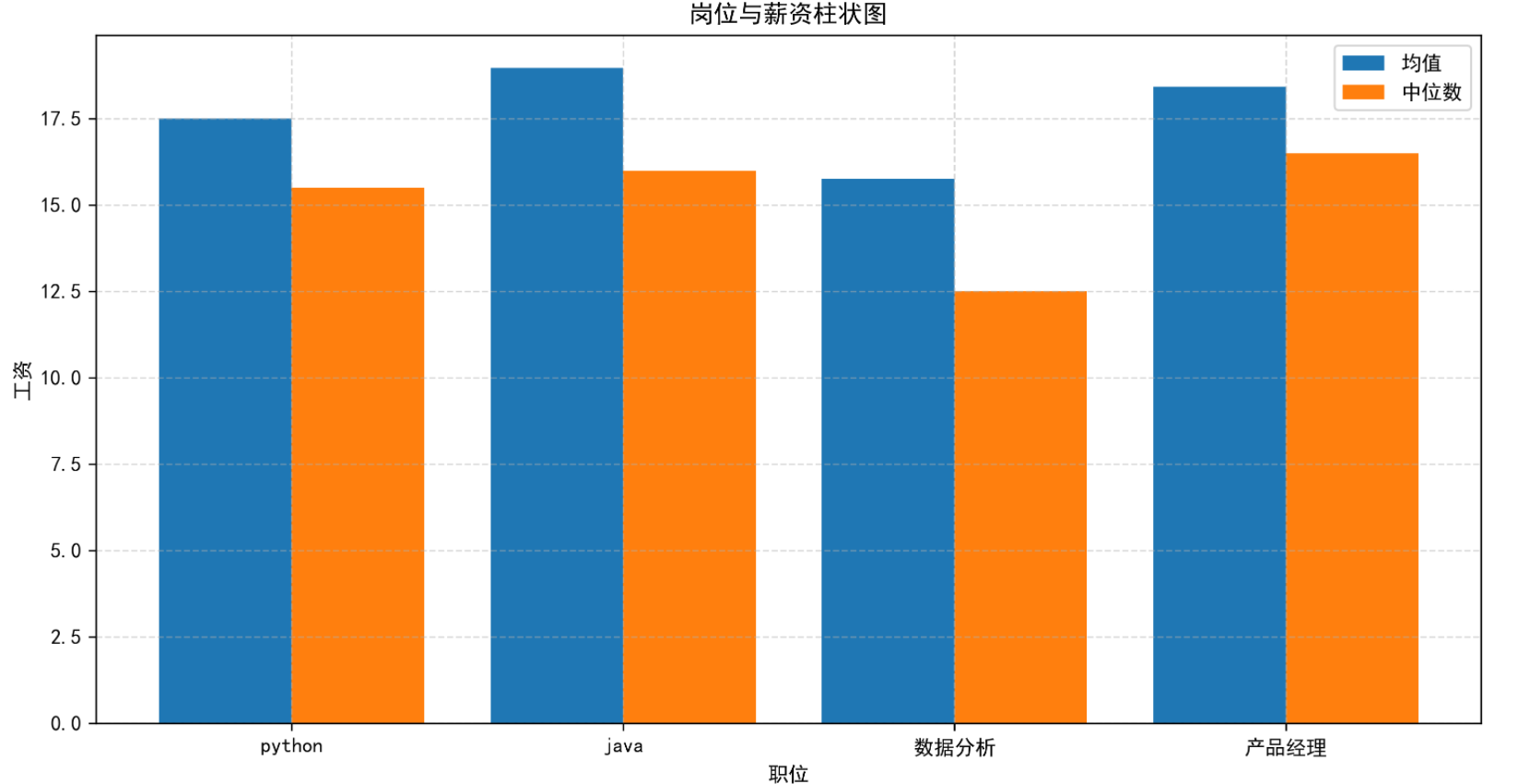


图4-3 岗位与薪资柱状图

4.2.3 岗位类型与薪资柱状图分析

四个岗位的薪资均值都在15k-19k，最高的是java，约为18.5k左右，最低的是数据分析，约为15.5k左右。Python、java、产品经理的薪资中位数大致相同，约为1.6k，但是数据分析的薪资前三个略低2.5k左右。数据分析总体要不如Python、java和产品经理。

4.2.4 岗位类型与薪资箱线图

岗位于薪资箱线图代码如下：



图4-4 岗位与薪资箱线图代码

岗位与薪资箱线图如下：

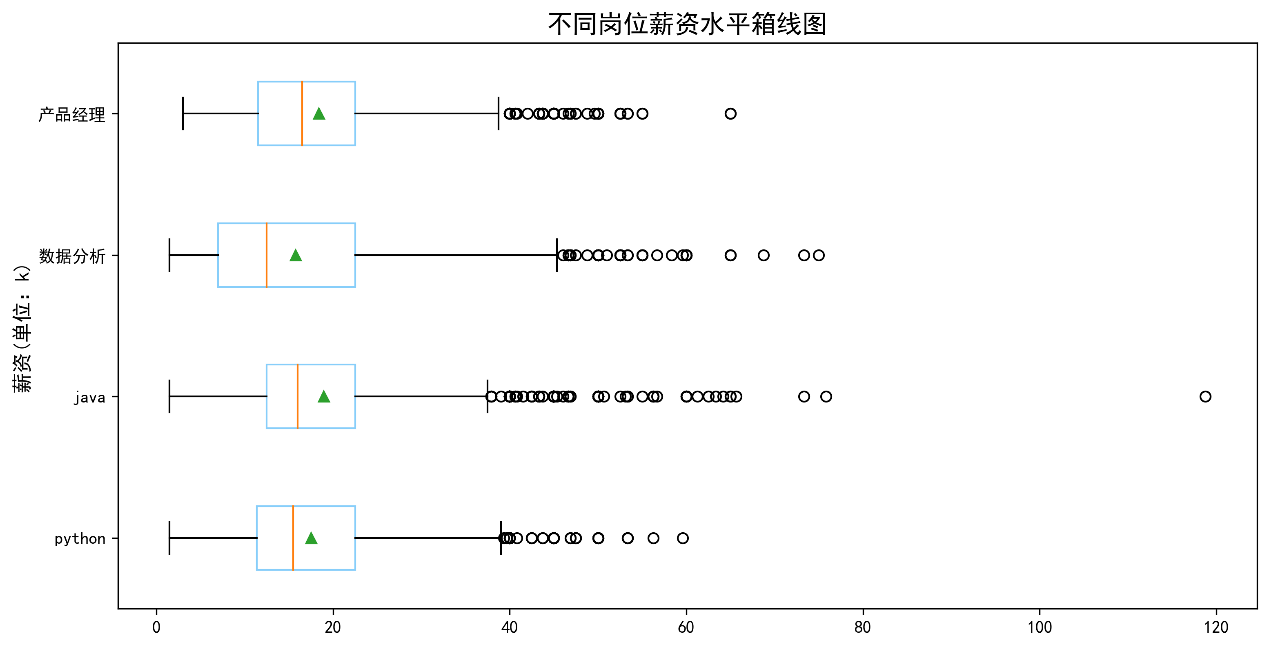


图4-5 岗位与薪资箱线图代码

4.2.5 岗位类型与薪资箱线图分析

箱线图的最左边的线表示的是下限，最右边的是上限，箱子的左宽是25%，右宽是75%，箱子中间的线是中位数，三角形是均值。箱子的长度表示波动程度，也就是左宽到右宽的长。

根据图4-5可知，产品经理、java、python的薪资的波动程度大致相同，但是数据分析的波动程度相对较大，中位数和平均值都要比产品经理、java、python低，但是上限都要比他们高，说明数据分析是一个非常看重个体能力的职业。同时四种职位的异常点都比较多，上限要比75%的人高出非常多，说明只有少部分人可以拿非常高的薪资。40k大致为绝大多数人的薪资上限。

4.3 岗位数与薪资

4.3.1 岗位数与薪资直方图

岗位类型与薪资直方图代码：



图4-6 岗位数与薪资直方图代码

岗位类型与薪资直方图：

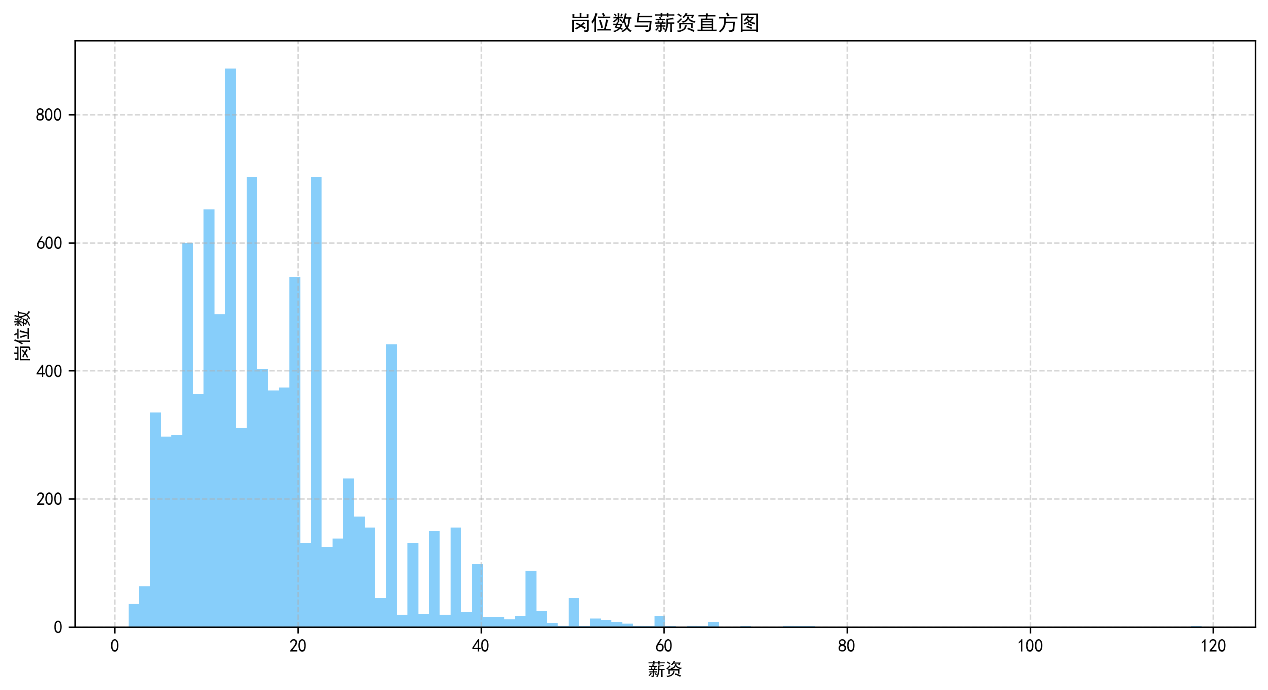


图4-7 岗位数与薪资直方图

4.3.2 岗位数与薪资直方图分析

从图中可以得知，绝大多数的岗位薪资都集中在7k-20k之间，再往上就会出现断崖式下降，随着薪资越来越高，其提供的岗位数则越来越少，从几百个到个位数。说明只有极少部分人能拿到非常高的薪水。

4.3.3 岗位数与薪资回归预测

岗位数与薪资回归预测代码：



图4-8 岗位数与薪资回归预测代码

岗位数与薪资回归预测图：

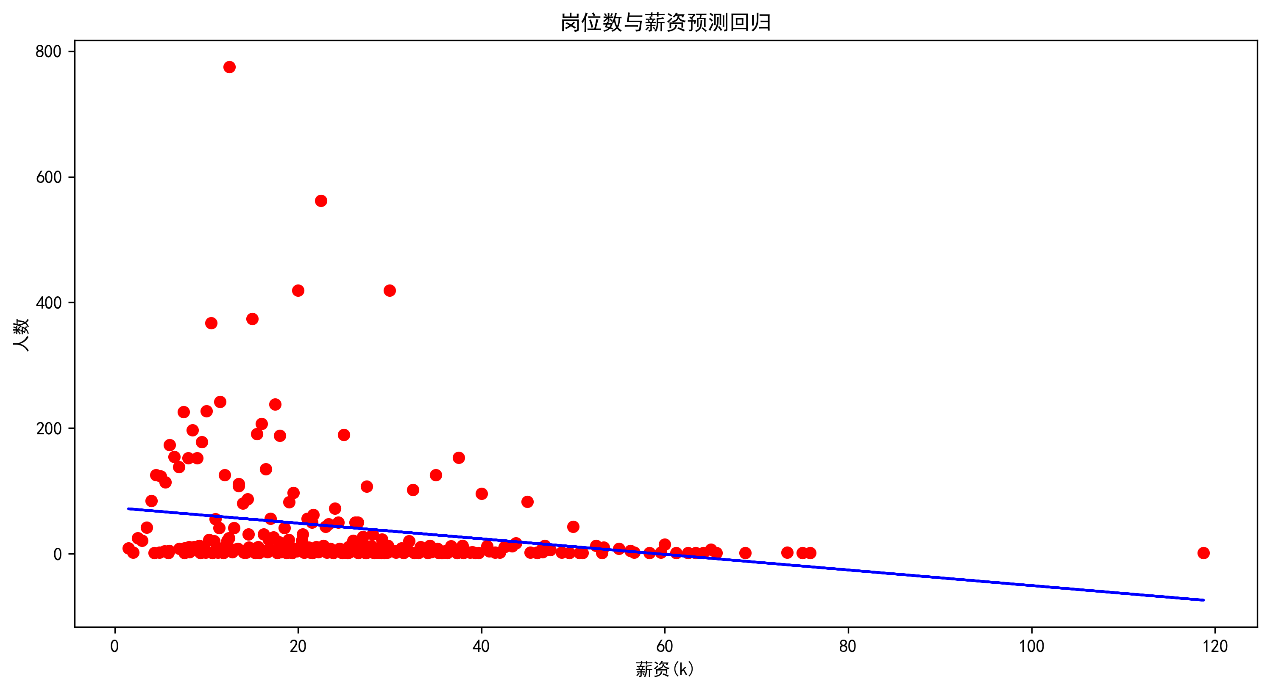


图4-9 岗位数与薪资回归预测图

4.3.4 岗位数与薪资回归预测分析

回归预测就是把预测的相关性作为基础，把影响预测目标的各因素找出来，然后找出这些因素和预测目标之间的函数关系的近似表达，并且用数学的方法找出来。在利用样本数据对其模型估计参数，并且对模型进行误差检验。如果模型确定，就可以用模型对因素的的变化值进行预测。根据图4-9可知，函数的斜率k呈现下降的趋势，说明薪资越高的人，人数会越少。

4.3.5 岗位数与薪资聚类

岗位数与薪资聚类分析代码：



图4-10 岗位数与薪资聚类分析代码

岗位数与薪资聚类分析图：

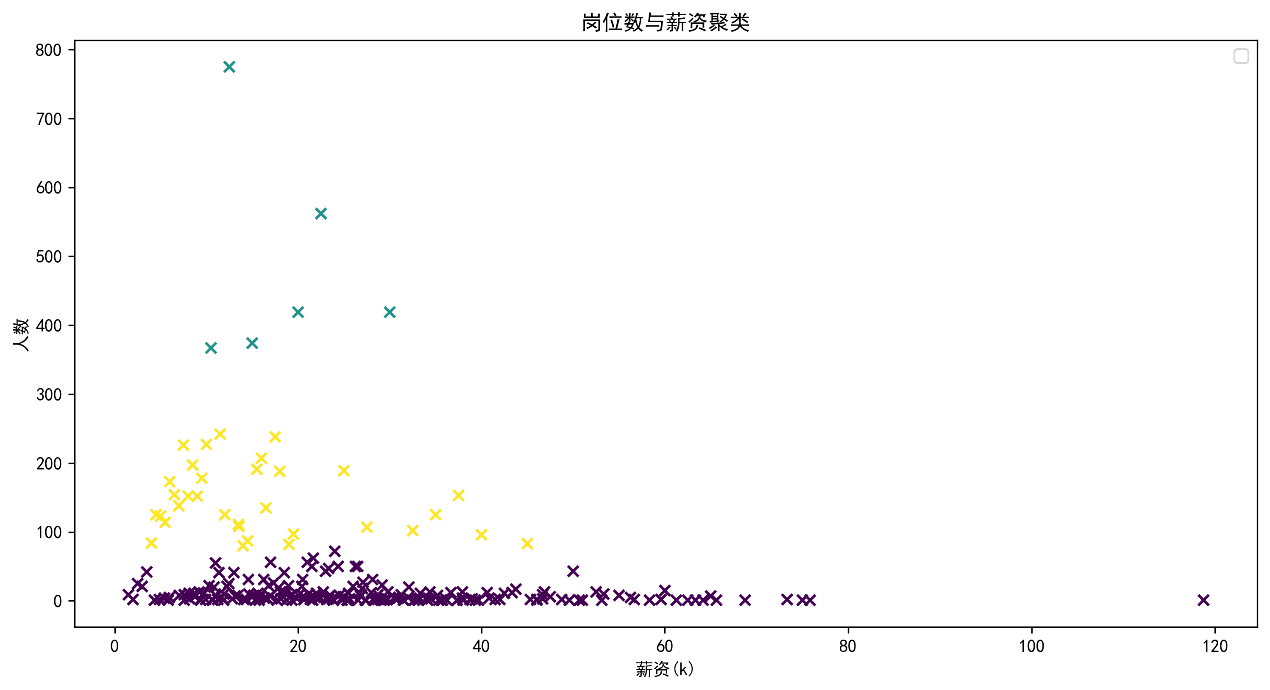


图4-11 岗位数与薪资聚类分析图

4.3.6 岗位数与薪资聚类分析

k均值聚类算法是一种迭代求解的聚类分析算法，其步骤是，预将数据分为K组，则随机选取K个对象作为初始的聚类中心，然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个聚类。每分配一个样本，聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是没有（或最小数目）对象被重新分配给不同的聚类，没有（或最小数目）聚类中心再发生变化，误差平方和局部最小。

根据图4-11可知，人数比较多的那一档的薪资集中在10k-30k，然后是3k-4.5k，最后是全部，说明绝大多数的人的薪资分布是比较均匀的，都集中在中间范围，而薪资越高或越低的人会越来越少，呈现正态分布。

4.3.7 岗位数与薪资决策树

岗位数与薪资决策树代码：



图4-12 岗位数与薪资决策树代码

岗位数与薪资决策树图：

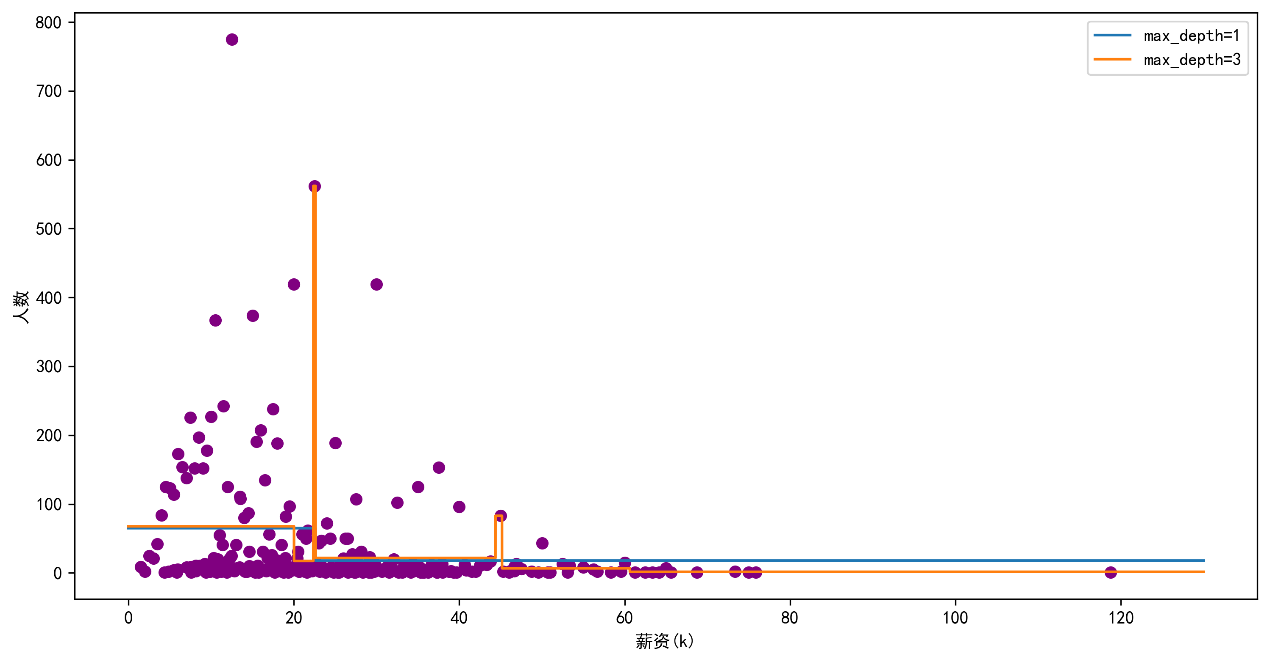


图4-13 岗位数与薪资决策树图

4.3.8 岗位数与薪资决策树分析

决策树是在已知各种情况发生概率的基础上，通过构成决策树来求取净现值的期望值大于等于零的概率，评价项目风险，判断其可行性的决策分析方法，是直观运用概率分析的一种图解法。由于这种决策分支画成图形很像一棵树的枝干，故称决策树。在机器学习中，决策树是一个预测模型，他代表的是对象属性与对象值之间的一种映射关系。决策树是一种树形结构，其中每个内部节点表示一个属性上的测试，每个分支代表一个测试输出，每个叶节点代表一种类别。这是一种十分常用的分类方法。它是一种监督学习，所谓监督学习就是给定一堆样本，每个样本都有一组属性和一个类别，这些类别是事先确定的，那么通过学习得到一个分类器，这个分类器能够对新出现的对象给出正确的分类。这样的机器学习就被称之为监督学习。根据图4-13可知20k对人数的预测值比较高，说明20k是多数人的选择，别的情况如120k则显得非常低，说明拿各种薪资的人参差不齐。

4.4 学历与薪资

4.4.1 查看学历类型

查看学历类型代码：

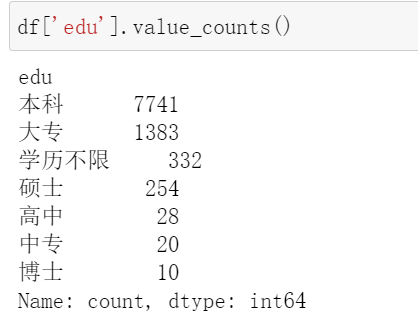


图4-14 查看学历类型

4.4.2 学历与薪资柱状图

学历与薪资柱状图代码：

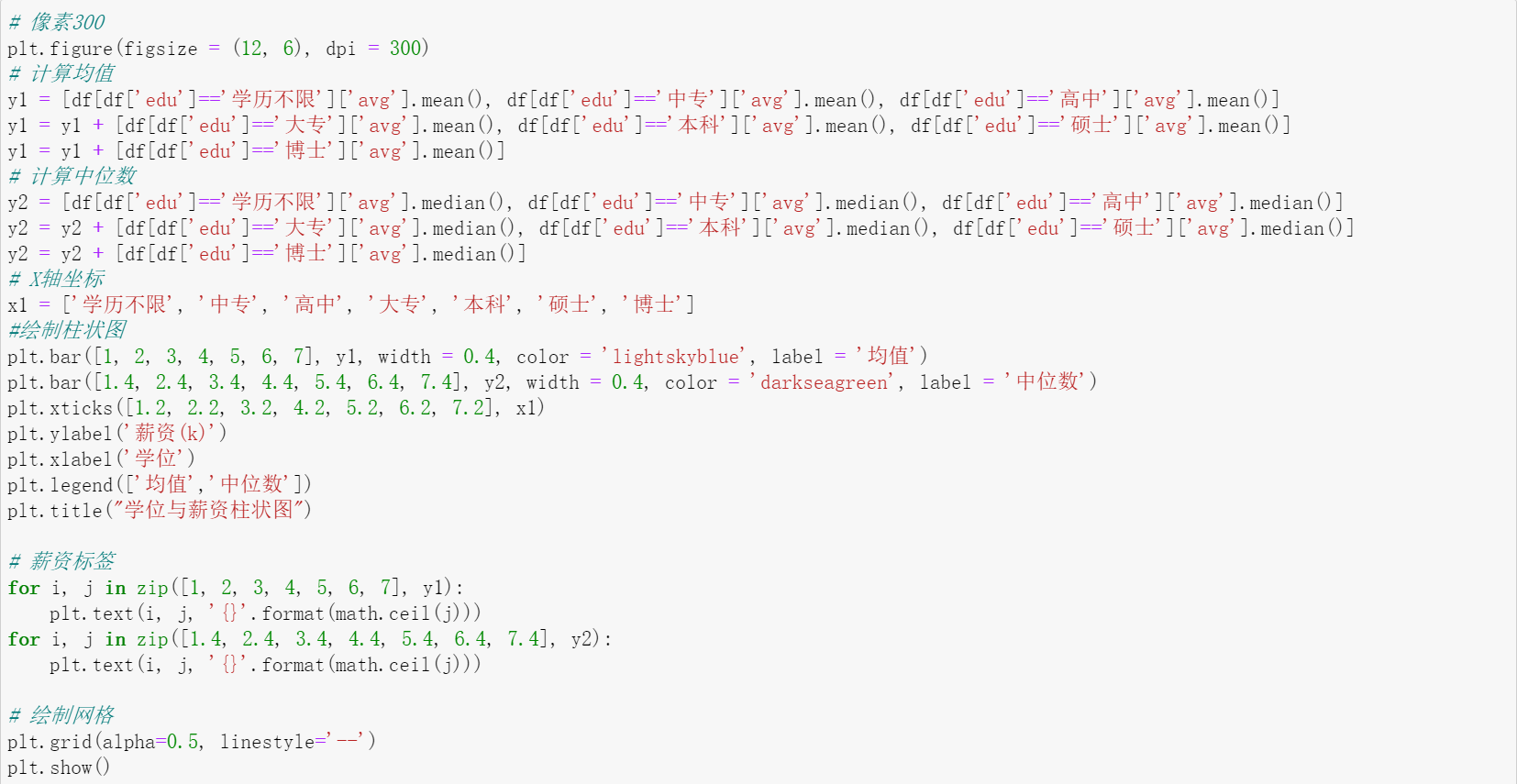


图4-15 学历与薪资柱状图代码

学历与薪资柱状图：

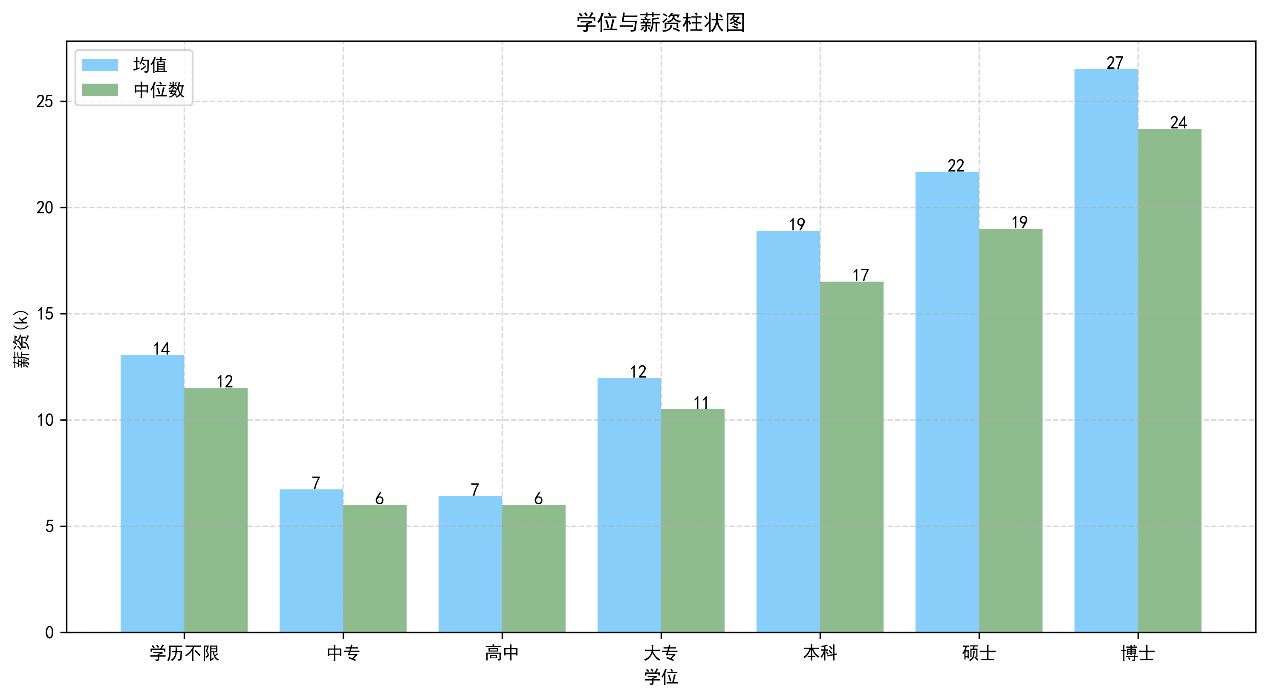


图4-16 学历与薪资柱状图代码

4.4.3 学历与薪资柱状图分析

根据图4-16可知，薪资随着学历的增加而增加，中专最低，博士最高。学历不限并不是最低的，虽然学历不限代表任何学历都可以入职，但是略大于大专而低于本科。所以去除学历不限来看，薪资和学历高低呈正相关的关系。学历不限是全部学历的集合，可能是采取了一种均值的数据。因此略大于大专。

4.5 各学历占比

4.5.1 学历饼图

学历饼图代码：

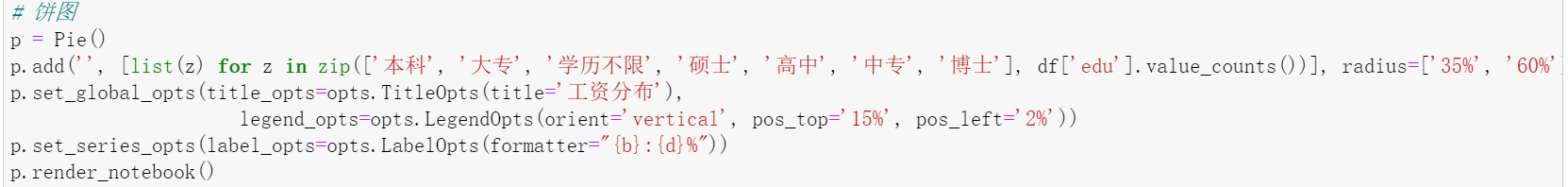


图4-17 学历饼图代码

学历饼图：

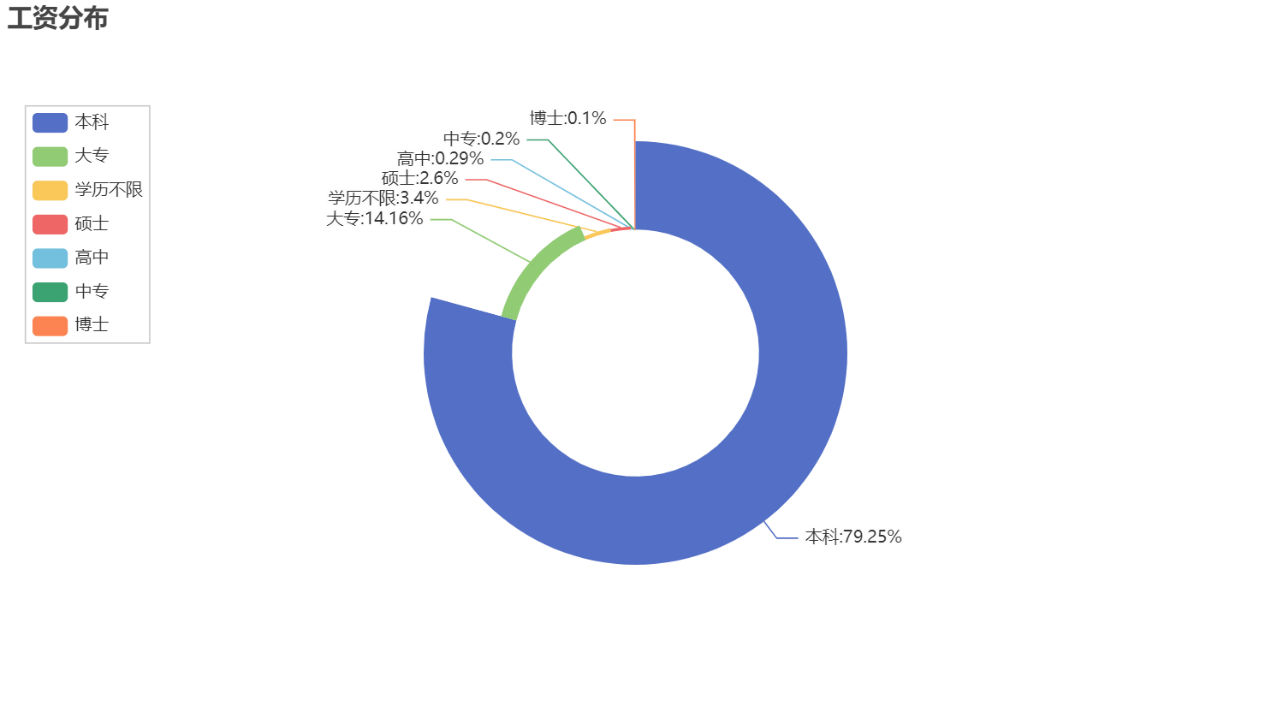


图4-18 学历饼图

4.5.2 学历饼图分析

本科占所有岗位的绝大多数，其次是大专、学历不限、硕士、高中、中专、博士。本科79.25%，大专14.16%，学历不限3.4%，硕士2.6%，高中0.29%，中专0.2%，博士占比仅仅只有0.1%。所以学历取得本科是大多数岗位要求的前提条件，本科的占比达到8成！

4.6 工资经验与薪资

4.6.1 查看工资经验类型

查看工作经验类型代码：

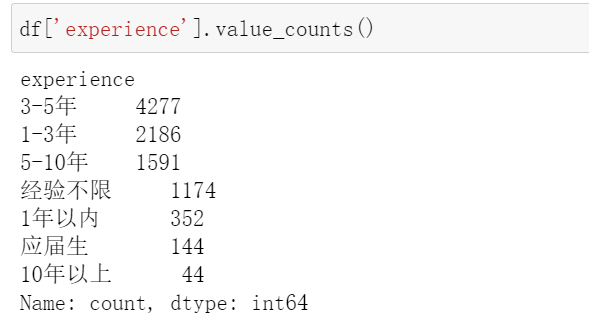


图4-19 查看工作经验类型

4.6.2工资经验与薪资柱状图

工作经验与薪资柱状图代码：



图4-20 工作经验与薪资代码图

工作经验与薪资柱状图：

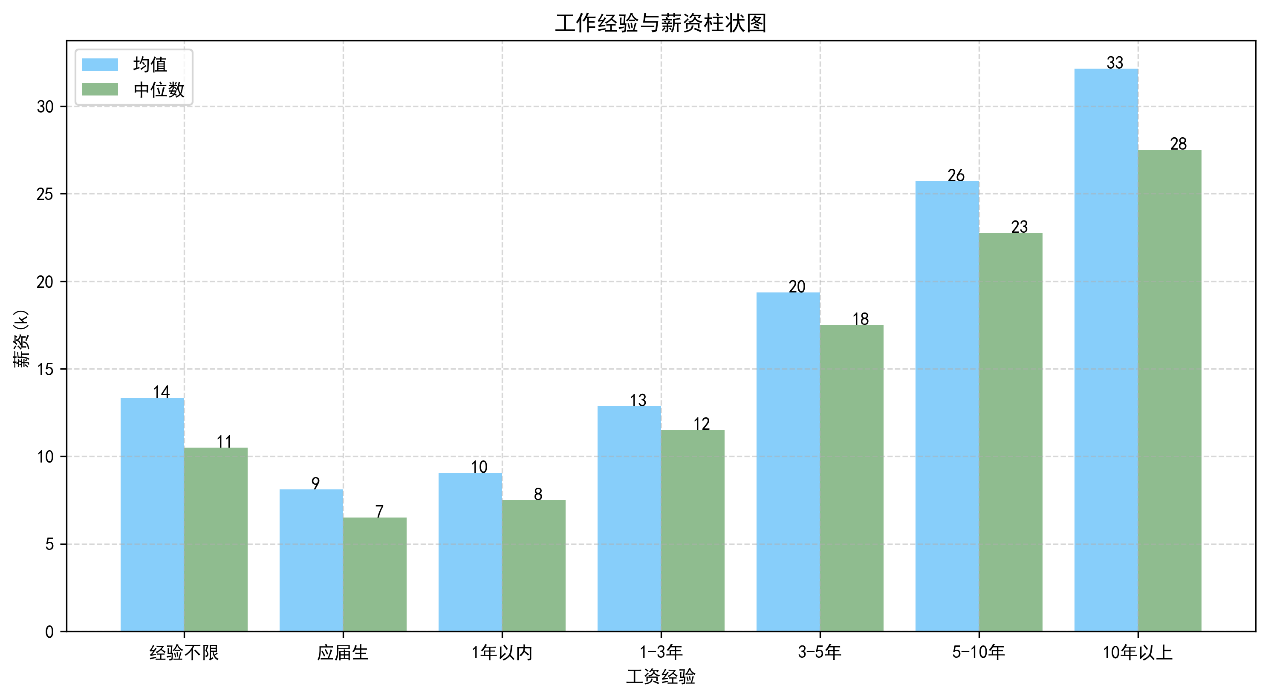


图4-21 工作经验与薪资柱状图

4.6.3工资经验与薪资柱状图分析

根据图4-16可知，薪资随着工作经验的增加而增加，应届生最低，10年以上最高。经验不限并不是最低的，虽然经验不限代表任何工作经验的人都可以入职，但是略大于1-3年而低于3-5年。所以去除经验不限来看，工作经验和薪资呈正相关的关系。经验不限是全部学历的集合，可能是采取了一种均值的数据。因此略大于1-3年。

4.7 各工作经验占比

4.7.1工资经验饼图

工作经验饼图代码：

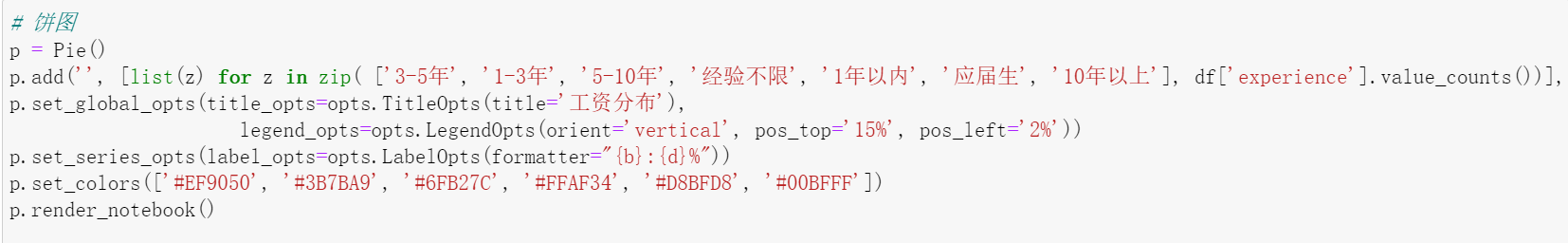


图4-22 工作经验饼图代码

工作经验饼图：

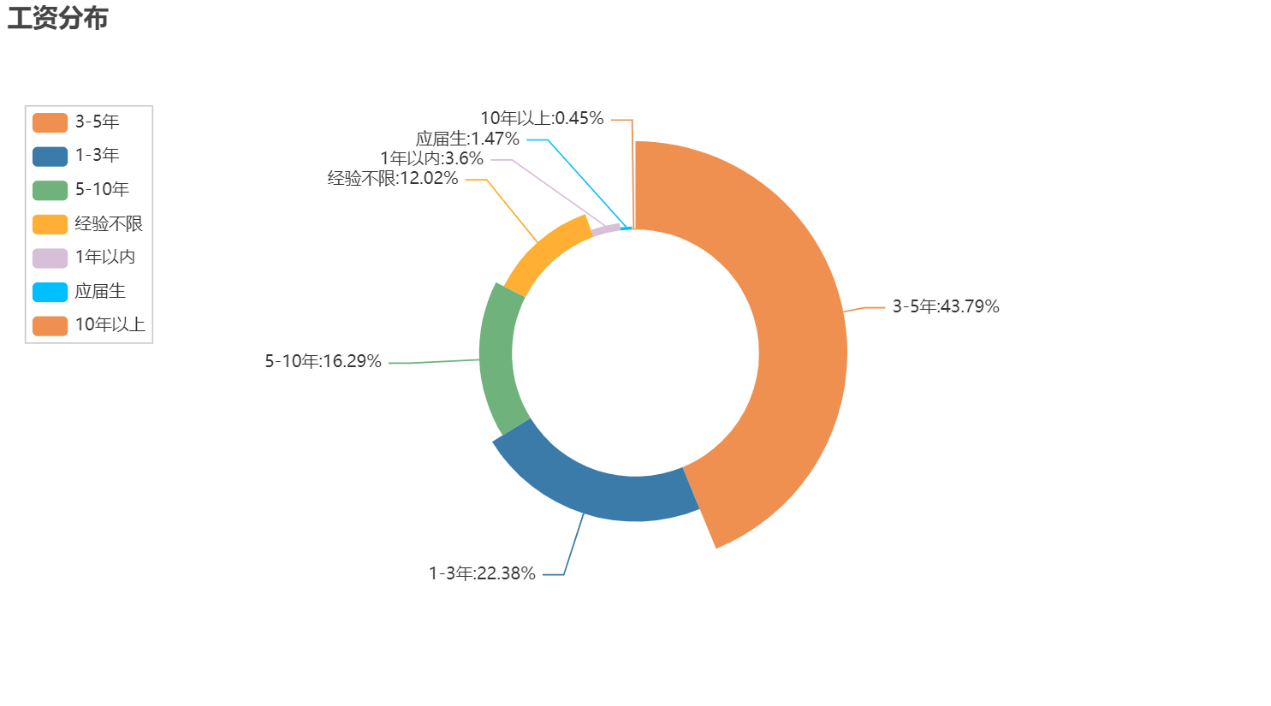


图4-23 工作经验饼图

4.7.2工资经验饼图分析

3-5年工资经验提供的岗位最多，其次是1-3年的工作经验。应届生有且仅有1.47%，经验不限占比12.02%。所以多数岗位对工作经验的要求是3-5年，只有3-5年的工作经验才有资格胜任大部分岗位。

4.8 城市与薪资

4.8.1查看城市类型

查看城市类型代码：

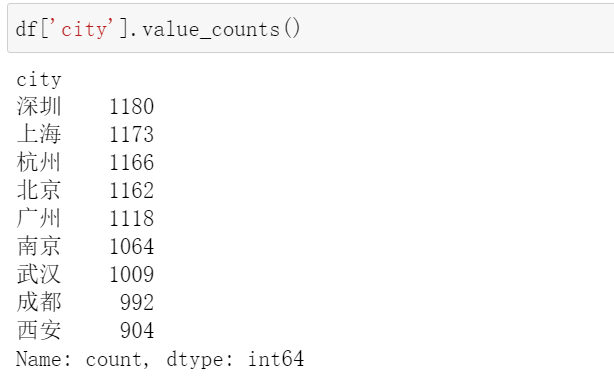


图4-24 城市类型

4.8.2 城市与薪资柱状图

城市与薪资柱状图代码：



图4-25 城市与薪资代码

城市与薪资柱状图：

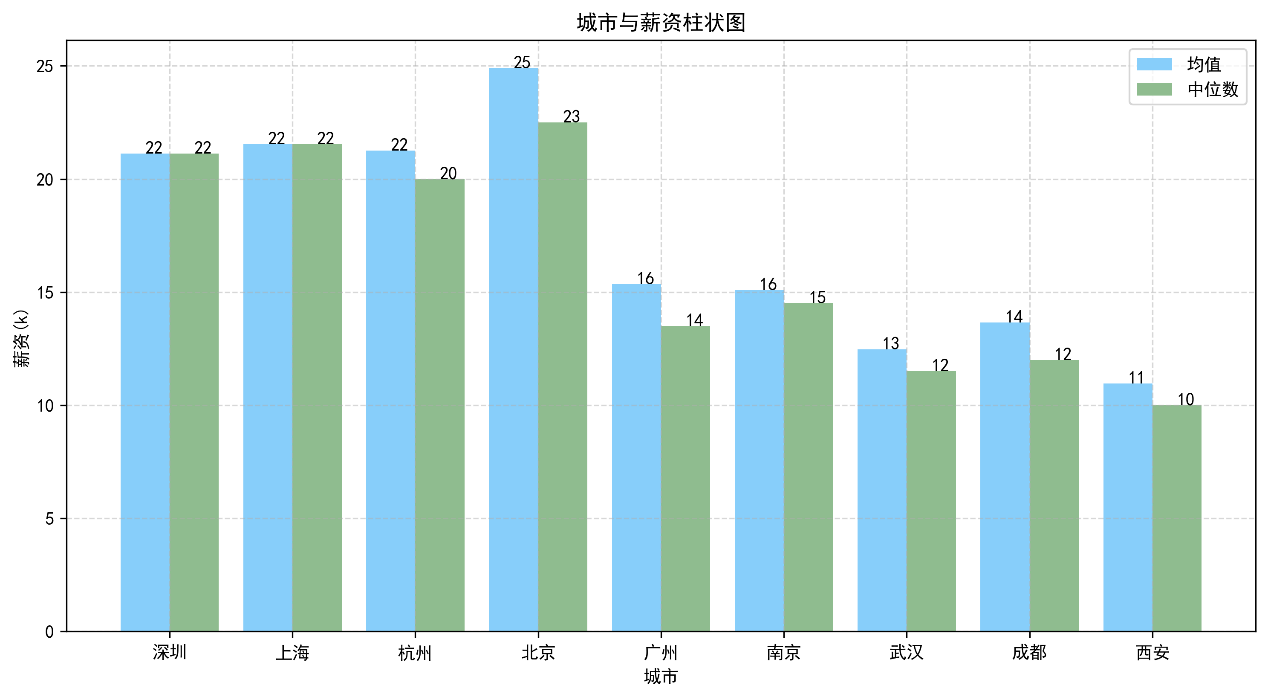


图4-26 城市与薪资柱状图

4.8.3 城市与薪资柱状图分析

深圳、上海、杭州、北京属于一档，广州、南京、武汉、程度、西安属于另一档。深圳、上海、杭州、北京要明显高于广州、南京、武汉、程度、西安，其中北京最高。说明IT行业的薪资与城市的科技水平相关。城市的科技水平越高，平均薪资就越高。

4.8.4 城市与薪资箱线图

城市与薪资箱线图代码：



图4-27 城市与箱线图代码

城市与薪资箱线图：

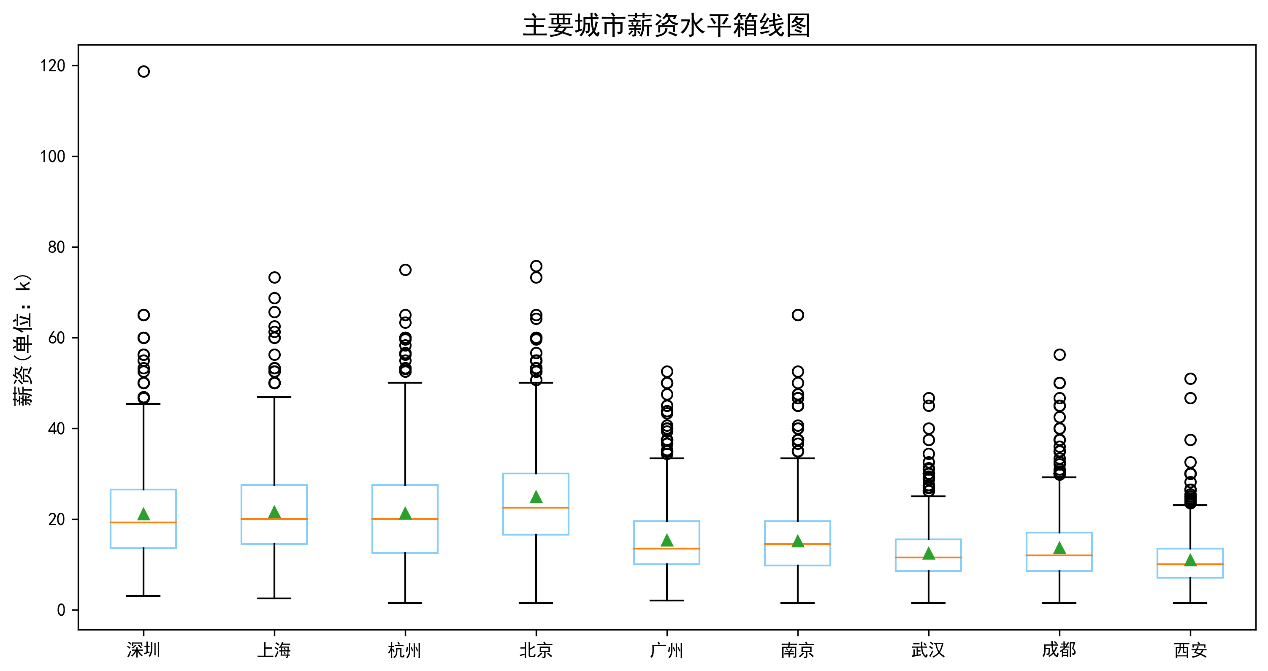


图4-28 城市与薪资箱线图

4.8.5 城市与薪资箱线图分析

深圳、上海、杭州、北京的综合薪资要明显高于广州、南京、武汉、成都、西安。但是广州、南京、武汉、成都、西安仍然有人能远超深圳、上海、杭州、北京的高水平，说明个人能力并不局限于城市的科技水平。城市的科技水平只能反映整体的情况。能力强的人同样可以取得非常高的薪资。

第五章 总结

在python、java、数据分析、产品经理这几个主流的IT行业中，薪资和行业类型的关系不是特别的大，四个行业的薪资平均水平都没有明显高于其他各个行业，但是薪资和学历、工作经验以及城市科技水平有着比较大的关系。大多数人拿的薪资都在10k-20k之内，只有少数人才能拿比较高的薪水。

一个人的学历越高、工作经验越丰富，薪资也就越高。同时城市的科技水平越发达，整体的薪资水平就越高。一个人的工作经验和学历是这个人能力的直接体现，但是科技相对落后的城市又有人能达到甚至超过科技发达城市的高薪资水平。

这证明了薪资很大程度上和一个人的能力是呈正相关的关系。

所以总的来说城市的科技水平决定了薪资的下限，反映了绝大多数人的真实薪资状况；而薪资的上限由一个人的能力决定，能力越强的人拿的薪资就越高，并不会受城市的发达水平约束，在科技相对落后的城市中，能力最强的人的薪资超过发达城市中能力较强的人的情况也是真实存在的。