软件综合实训设计报告

题目： 大数据职业前景分析

班级： 计本163

姓名：张清, 钟鹏荣,王泰龙,吴小凡

学号： 2016021214,2016021193

2016021076,2016021199

指导教师： 姜翠霞

2019年 6月

**数据采集课程设计成绩单**

**一．组内自评成绩表**

|  |  |  |
| --- | --- | --- |
| 项目名称 | 大数据职业前景分析 | |
| 自评排序 | \* 请按照项目中各成员的工作任务、工作态度、完成质量等自行排序。 | |
| 序号 | 姓名 | 负责的工作任务 |
| 1 | 张清 | 数据分析，总结报告 |
| 2 | 钟鹏荣 | 数据爬虫 |
| 3 | 王泰龙 | 数据爬虫 |
| 4 | 吴小凡 | 数据可视化 |

**二．总成绩表（**表2由指导教师填写）

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 姓名 | 答辩成绩 | **总评成绩** |
| 1 | 张清 |  |  |
| 2 | 钟鹏荣 |  |  |
| 3 | 王泰龙 |  |  |
| 4 | 吴小凡 |  |  |

题目： 大数据职业前景分析

摘要：

本次课程设计通过使用scrapy + selenium + xpath + MongoDb+csv 对数据进行获取,存储。通过我们学习到的python的切片，正则表达式，numpys和padas库的内容对数据进行分析，通过使用matplotlib库对数据进行可视化。通过本次设计我们可以清楚的看到大数据的需求。

关键字：爬虫爬虫，数据分析，数据可视化。

1. 引言:

大数据的行业覆盖面积已经达到了百分之九十九，纵观未来三十年都属于紧俏行业，目前我国对于大数据的就业人员需求很大。

伴随着大数据技术的成熟，大数据应用的普及和发展才刚 刚开始，我们预计未来二十年，甚至更长一段时间都是大数 据黄金发展阶段，相关的行业将引来巨大的发展机遇。大部 分行业都需要，市场、营销、运营相关的需求很多。大数据 不是职位，学完大数据认证后你可以从事大数据挖掘专家， 高级行业分析师，大数据业务架构师，大数据架构师，大数 据算法工程师，大数据开发工程师，大数据运维工程师。不 管是国内还是国外，大数据相关的人才都是供不应求的局面。 目前市场急需运用大数据分析结果的大数据相关管理人才。

大数据分析在业务中使用的流程分为数据获取和预处理、 数据存储管理、数据分析建模、数据可视化。

1. 需求分析

伴随着大数据技术的越来越火爆，越来越多的人，在考虑大数据方向，比较重要的是什么，经验，学历，还是其他的，我们通过这个数据分析对我们的大数据的职业前景做一个数据分析，通过对职业需求的生成词云，来展示现在职场最需要的大数据的东西，通过研究学历和薪水的关系，经验和薪水的关系得出简单结论。

1. 爬虫分析
   1. 目标网站：<https://www.zhipin.com/>
   2. 要爬取的网页连接：

<https://www.zhipin.com/c101010100-p100599/?ka=search_100599>

<https://www.zhipin.com/c101010100-p100506/?ka=search_100506>

<https://www.zhipin.com/c101010100-p100507/?ka=search_100507>

<https://www.zhipin.com/c101010100-p100508/?ka=search_100508>

<https://www.zhipin.com/c101010100-p100509/?ka=search_100509>

<https://www.zhipin.com/c101010100-p100511/?ka=search_100511>

<https://www.zhipin.com/c101010100-p100512/?ka=search_100512>

<https://www.zhipin.com/c101010100-p100513/?ka=search_100513>

* 1. 要获取的数据：

职位名，薪水，工作地点，经验，学历，职位福利，职位要求。

* 1. 对页面进行分析：

每一个页面都可以找到其他链接的网址，每一个目标页面的链接都是有固定格式的。故通过CrawlSpider进行爬虫。



图3-1

* 1. 代码编写
     1. 编写Item

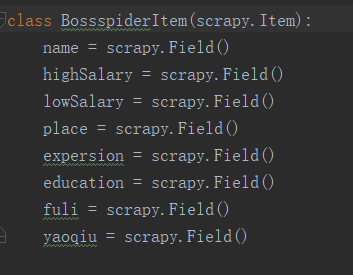


图3-2

* + 1. 编写Spider

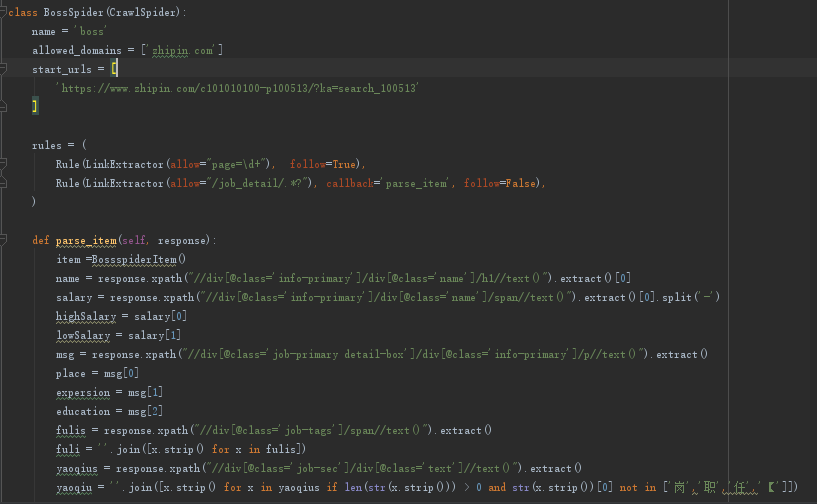


图3-3

* + 1. 编写middlewares下载器中间件，设置随机请求头和使用selenium做简单的反爬虫。

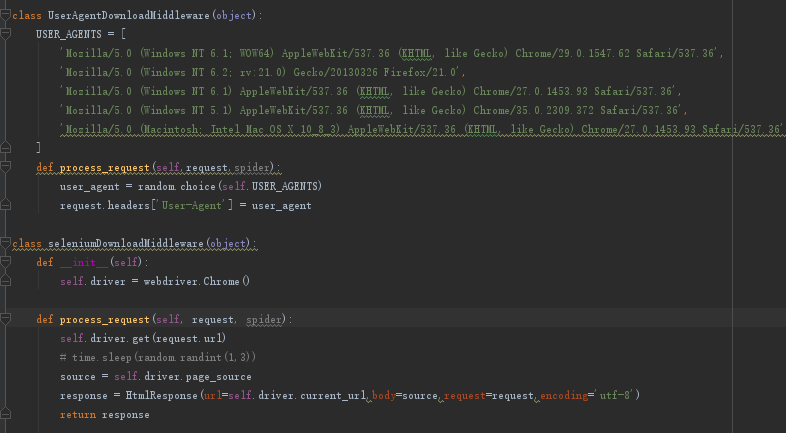


图3-4

* + 1. 编写pipeline管道

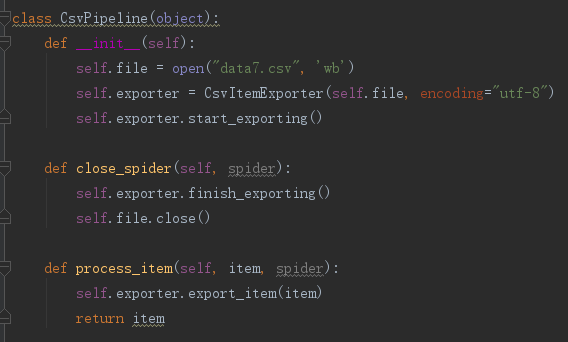


图3-5

* + 1. 编写settings文件

修改robits协议，设置下载时延，开启下载器中间件，开启管道。

* + 1. 程序运行

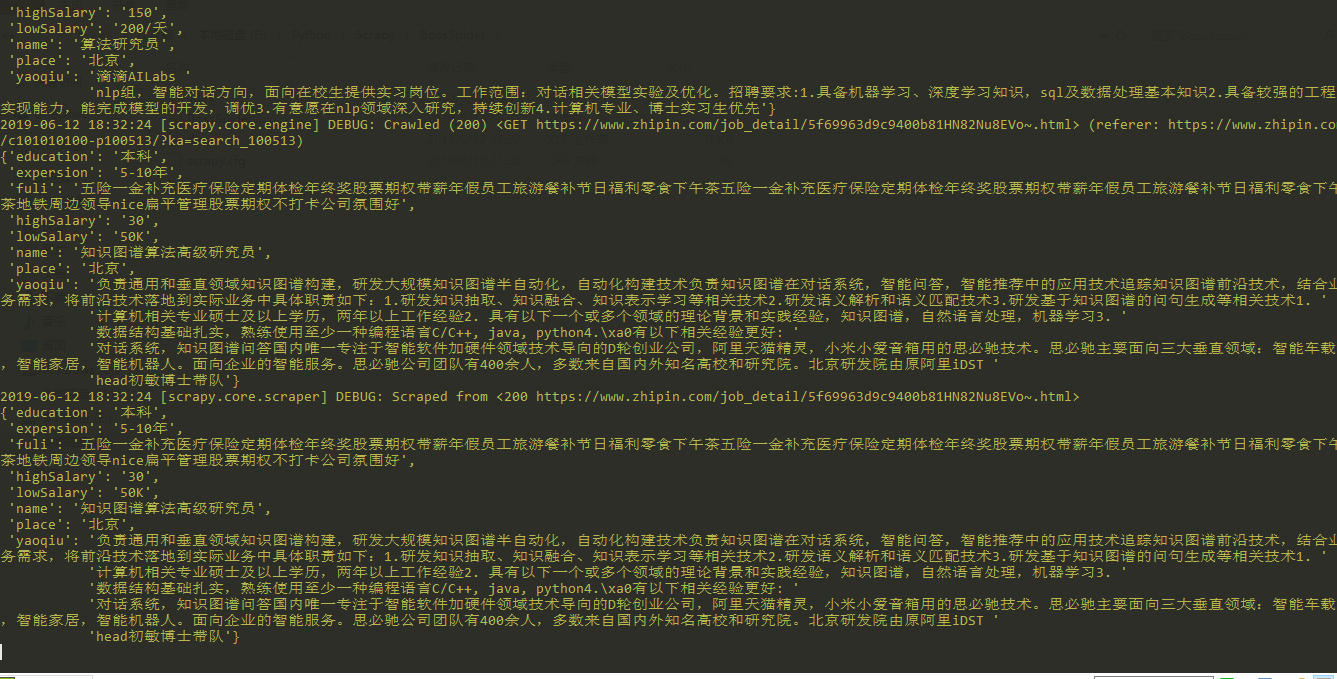


图3-6

* + 1. 运行结果



图3-7

1. 数据解析：
   1. 通过切片去掉一些多余的数据
   2. 通过padas统计所有出现的数据的次数，去掉一些异常的数据和可以忽略的数据

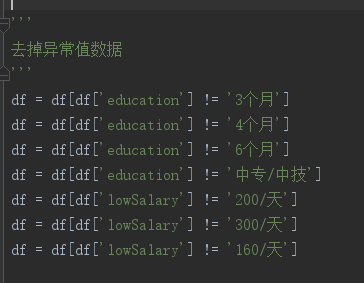


图4-1

* 1. 通过对padas数据的添加一列数据
  2. 通过padas的pivot\_table方法求出平均数，中位数，方差，标准差。
  3. 对数据进行分组研究

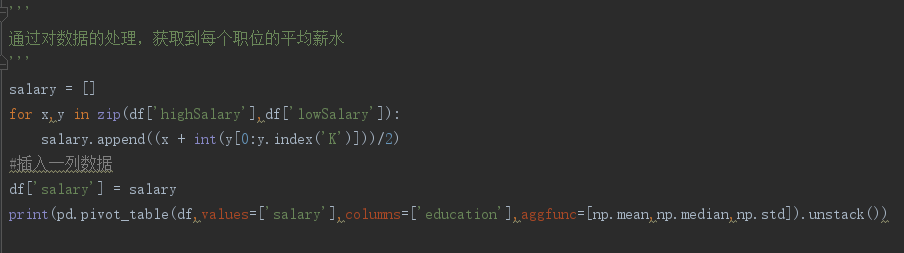


图4-2

1. 数据可视化
   1. 对大数据整体薪水的可视化

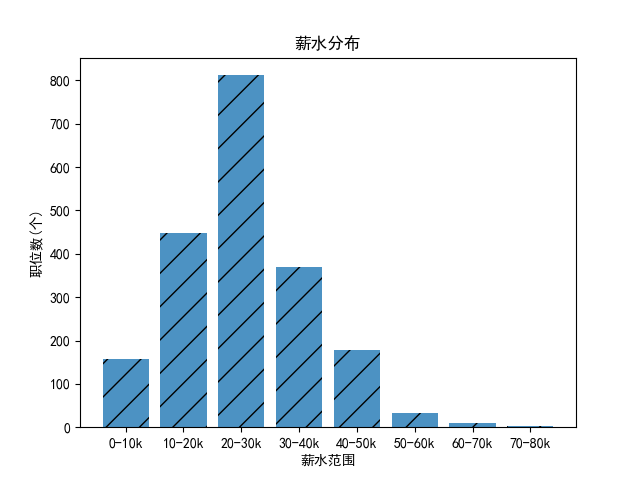


图5-1

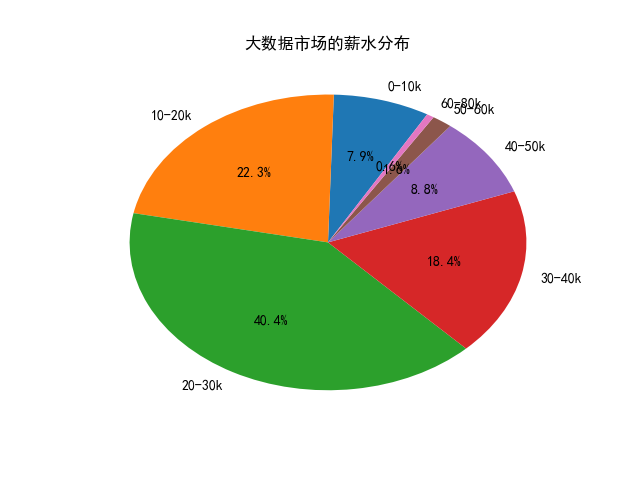


图5-2

结论:

薪酬主要集中在10k-40k，其中在10k-20k，20K-30K，中居多，起薪还是挺高的，薪酬的提升幅度也是令人期待的。从总体薪酬可以看出，数据分析师的收入还是挺可观的。

* 1. 对学历和薪水的关系进行研究

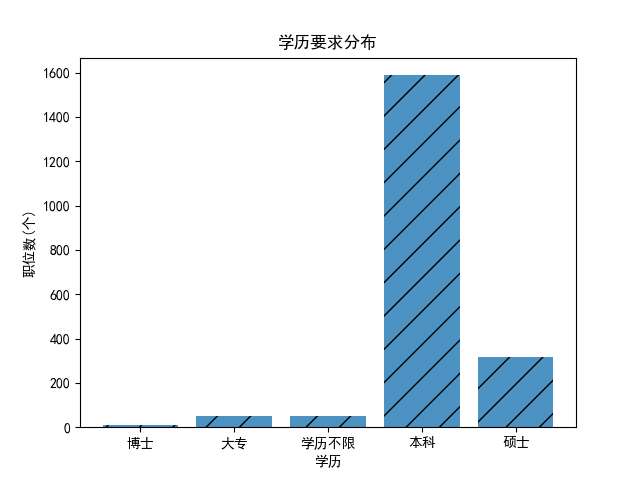


图5-3

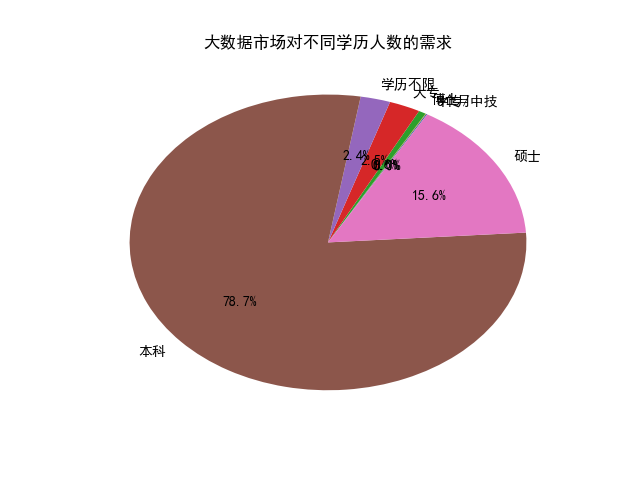


图5-4

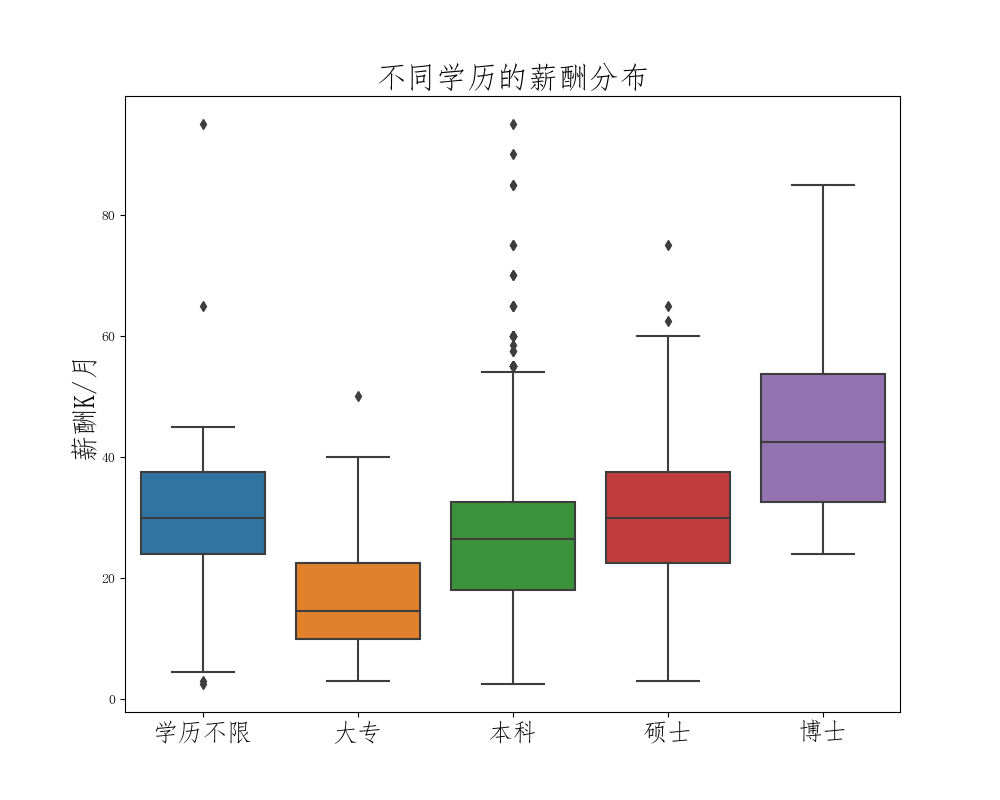


图5-5

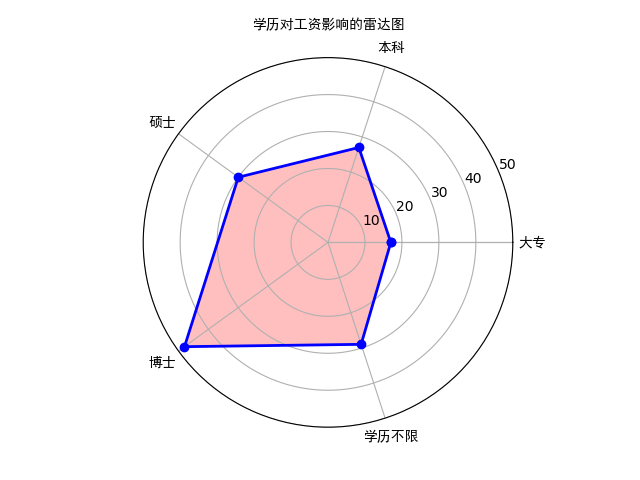


图5-6

结论:

通过对数据的分析可以得出，大数据方向对本科生的需求最大，其次硕士，不限学历和大专学历需求很少，对于博士需求基本没有。同时发现学历不限和硕士学历拿到的薪水差不多，所有在查看学历不限的时候要注意。同时由于在设计的时候，没有考虑到一些高端职位对学历的需求，在以后的工作中一定做出调整。

* 1. 对经验和薪水的关系进行研究

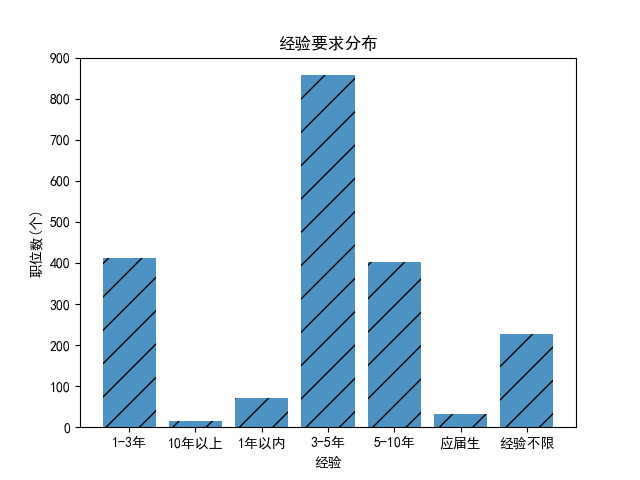


图5-7

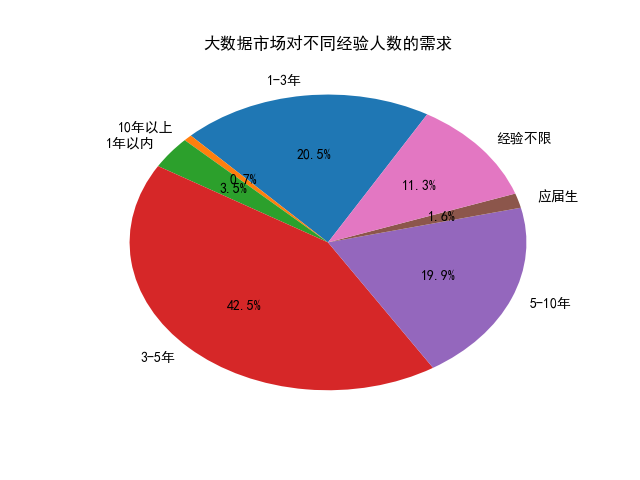


图5-8

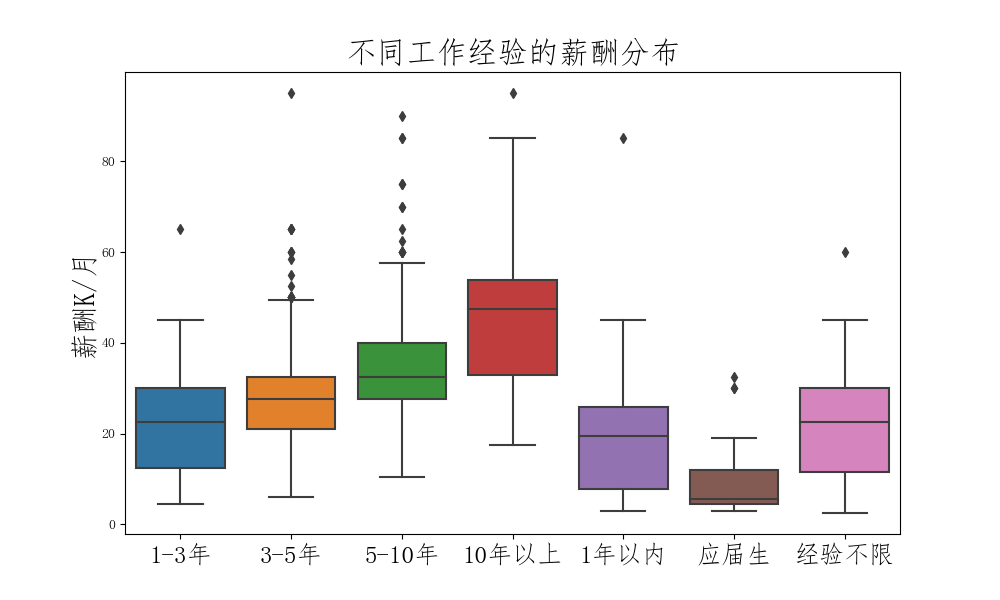


图5-9

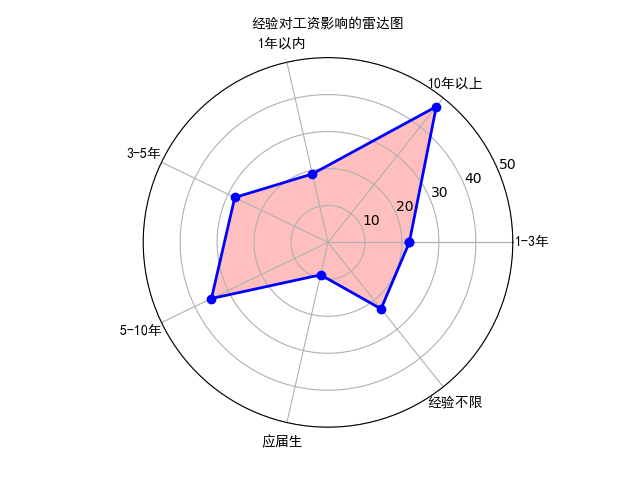


图5-10

结论:

随着工作经验的增加，薪酬也会增加。同样工作经验的薪酬差距也随着年限的上升而增大，差距变大，逐渐出现分水岭。5-10年的薪酬分布已经变得比较大，进一步说明5年是一个分水岭。另外，不限经验明显会比1年以下的薪酬高，比较接近1-3年经验的薪酬。所以，不能太乐观认为不限工作经验的就是不需要经验的。

1-3年经验,3-5年经验和5-10年经验的需求比较大，这要集中在这两个年限。从这个情况看，虽然数据分析需求还是挺大的，但是对经验要求比较大，可能会造成招不到人的火爆现象，实际对没有经验需求还是比较少的，转行就要被破冷水了。从长远看，10年以上经验的需求还是会比较少的，可以看出5年来说对一个从事大数据工作的人来说是比较关键的。

* 1. 对工作福利生成词云

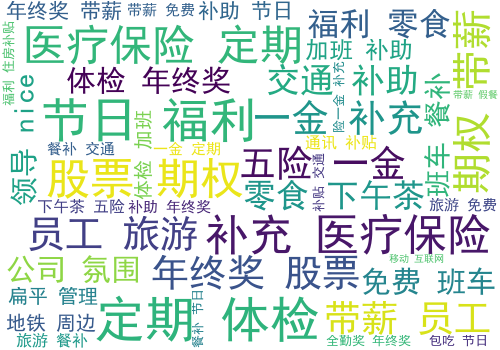


图3-9

结论：

薪酬是应聘者考虑的一个重要因素，但也不是唯一，除了薪酬，五险一金，双休，带薪年假，发展空间，每年公司组织旅游等等，这些也是要考虑的。

* 1. 对工作要求生成词云



图3-10

结论:

偏向大数据方向的会比较高薪，Hadoop，java，python是能让你提高薪酬的技能。大数据还需要一定的经验和沟通能李和自我的学习能力。在IT这个行业不断学习是非常重要的。

1. 验证影响因素对薪水是否有显著性影响。
   1. 验证学历对薪水是否有显著性影响。
      1. 代码

d2 = read.csv("F:\\Python\\BossSpider\\BossSpider\\data\\data\_deal.csv")

fm1 = lm(salary~education,data=d2)

fm1

anova(fm1)

summary(fm1)

* + 1. 结果

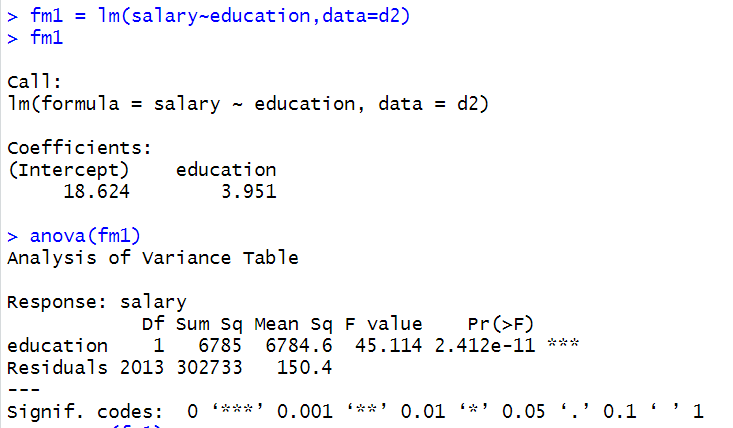


图6-1

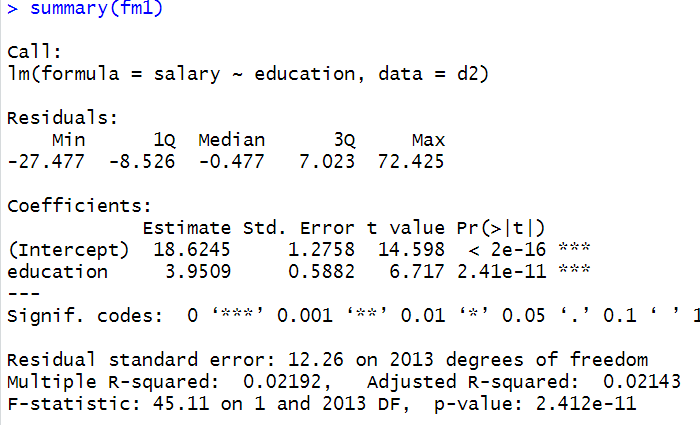


图6-2

* + 1. 结论

通过对学历和薪水的关系做模型方差得出的值为2.412e-11 < 0.05 所有具有显著性影响。本回归系数具有统计学意义。

通过对学历和薪水的关系做t检验的出各个学历的P值均小于0.05所有可以得出学历和薪水具有显著性关系.本回归系数具有统计学意义。

* 1. 验证经验对薪水是否有显著性影响。
     1. 代码

fm2 = lm(salary~expe,data=d2)

fm2

anova(fm2)

summary(fm2)

* + 1. 结果

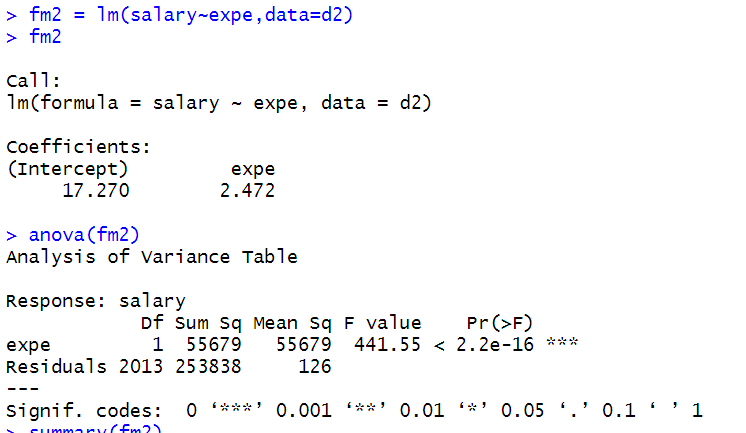


图6-3

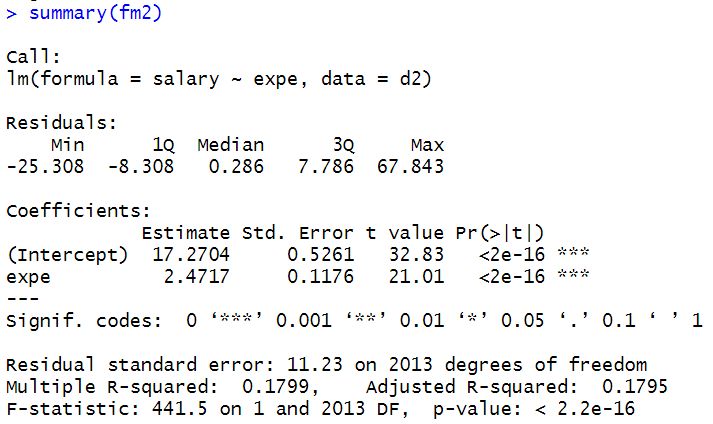


图6-6

* + 1. 结论

通过对经验和薪水的关系做模型方差得出的值为2.2e-16 < 0.05 所有具有显著性影响。本回归系数具有统计学意义。

通过对经验和薪水的关系做t检验的出各个学历的P值均小于0.05所有可以得出学历和薪水具有显著性关系.本回归系数具有统计学意义。

* 1. 验证学历和经验两个因素是否对薪水有显著性影响，哪个变量对数据的影响性更大。
     1. 代码

fm3 = lm(salary~education+expe,data=d2)

fm3

anova(fm3)

summary(fm3)

#标准化回归系数，求出哪个因素对结果影响比较大

library(mvstats)

coef.sd(fm3)

图6-7

* + 1. 结果

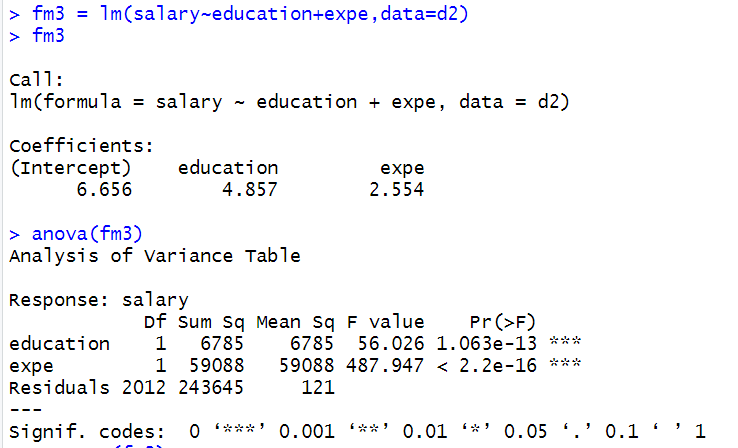


图6-5

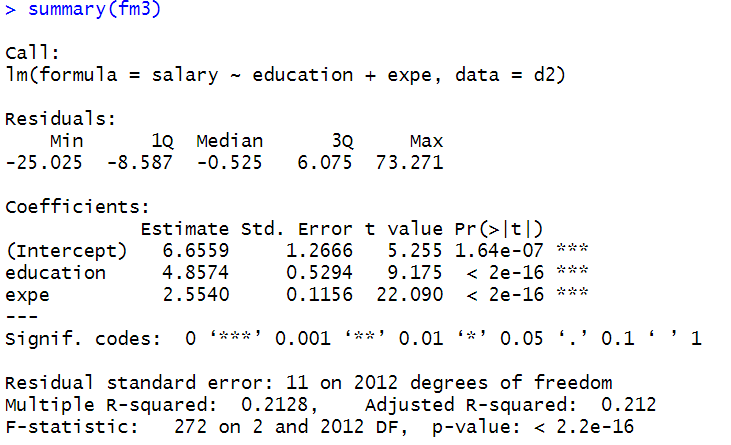


图6-6

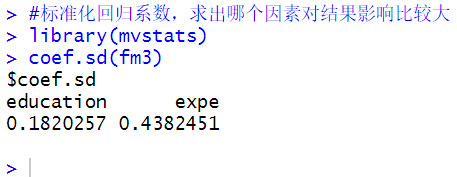


图6-7

* + 1. 结论

通过对经验和学历对薪水的影响做t检验，得出p值为2.2e-16,得出经验和学历对于薪水具有显著性影响，本回归系数具有统计学意义。

通过对数据做多元线性回归模型可得表达式为:

Y = 6.656 + 4.857 \* x1 + 2.554 \* x2

通过对数据做标准化回归系数可得出，经验对薪水的影响性更大一些。

1. 结束语

这个分析还是比较简单的，没有深入探究更多，有待进一步分析。通过本次课程设计，我们充分应用了我们所学到的知识点，同时在学的基础上通过查阅资料，网上搜索，对知识点有了进一步的理解和体会。

参考文献

1. 王仁武，python与数据科学.华东师范大学出版社
2. Jacqueline Kazil,Katharine Jarmul,Python数据处理 人民邮电出版社
3. 刘大成,Python数据可视化之matplotlib实践,电子工业出版社
4. 王斌会,多元统计分析及R语言建模 暨南大学出版社