**数据说明：**

1. 本案例所采用数据为“Mydata.xlsx”文件为数据分析岗位招聘数据，数据集包含了7493条岗位招聘信息，涵盖了较有代表性的六个省市地区:数据分析岗位密集的北京市、上海市和深圳市，以及相对稀少的陕西省、山西省、河北省。本数据集的每一列分别对应：招聘职位、招聘公司名称、公司所在地区、公司类别、公司规模、行业类别、招聘职位年限要求、学历要求、招聘人数以及职位描述、最高薪资和最低薪资。详细的变量含义见下表。

|  |  |  |  |
| --- | --- | --- | --- |
| 变量类型 | 变量名称 | 详细说明 | 取值范围 |
| 因变量 | 平均薪资 | (最低薪资+最高薪资)/2 | 1500-400000元/月 |
| 自变量 | 招聘职位 | 字符型变量 | 例：数据分析员，战略分析经理 |
| 公司名称 | 字符型变量 | 例：北京中航讯科技股份有限公司 |
| 公司所在地 | 定性变量  共6个水平 | 分别为陕西、山西、河北、北京、上海、深圳。 |
| 公司类别 | 定性变量  共8个水平 | 分别为民营公司、上市公司、外资、国企、非营利机构、合资、事业单位、创业公司 |
| 公司规模 | 定性变量  共7个水平 | 分别为少于50人、50-150人、150-500人、500-1000人、1000-5000人、5000-10000人，10000人以上 |
| 公司行业类别 | 定性变量  共443个水平 | 例：计算机服务，互联网/电子商务 |
| 经验要求 | 数值型变量 | 0-10年 |
| 学历要求 | 定性变量  共7个水平 | 分别为无、高中、中专、大专、本科、硕士、博士。 |
| 招聘人数 | 数值型变量 | 0-9人 |
| 最低薪资 | 数值型变量 | 1500-200000元/月 |
| 最高薪资 | 数值型变量 | 1500-599999元/月 |
| 岗位描述 | 字符型变量 | 包括职位描述，任职条件等。例：针对海量数据进行分析建模，挖掘数据潜在价值。掌握一定数据统计及分析工具，熟练运用SPSS或SAS或R等数据挖掘工具的优先等。 |

**分析任务：**

1. 读入数据并完成简单的数据清洗；

使用Rstudio提供的交互界面从excel中导入数据，然后根据各变量含义，对数据进行类型转换，然后求出各个岗位的平均薪资。使用的R代码如下：

1. 提取出频次最高的5个行业类别，并绘制该五个行业的频数直方图，对结果进行解读；

提取最高频次的5个行业类别的R代码以及绘制频数直方图如下所示：

频数最高的五个行业类别按照频数由高到低分别是： 互联网/电子商务，金融/投资/证券，计算机软件，快速消费品(食品、饮料、化妆品)，服装/纺织/皮革。其中，互联网/电子商务的频数远高于其他四个行业的频数，这可能与近年来互联网相关的产业遇到窗口期告诉发展有关。此外，金融/投资/证券，与计算机软件频数相差不大。快速消费品(食品、饮料、化妆品)，与服装/纺织/皮革的频数相差也不是很大。

1. 除去中频次最高的5个行业类别，绘制剩余行业的词云图；

绘制的除去频数最高的5个行业的剩余行业词云图如下：

本题所使用的R代码如下：

1. 以对数平均薪资为对象，进行简单的描述分析；

对数平均薪资的平均数是8.960，中位数是8.854，数据分布基本无偏，最小值是7.313，最大值是12.899，四分卫数分别为8.566和9.393。最小值出现在用友优普信息技术有限公司的“商务、行政实习生”职位上。最大值出现在汇鼎财富（北京）投资有限公司的“金融衍生品数据分析/交易员”职位上。

下面，绘制出对数平均薪资的直方图如下：

可以看出，在取对数之后，对数平均薪资的分布接近于正态分布，说明原的分布呈现比较严重的右偏，也即大多数岗位的薪资是比较低的。

本题所使用的R代码如下：

1. 使用箱线图，探究学历对岗位薪资的影响，对结果进行解读；

从第4题可知，工资的分布存在比较严重的右偏，因此在本题中，均对工资取对数进行讨论。分别做出对数平均薪资、对数最高薪资、对数最低薪资的箱线图如下：

可以看出，薪资的高低和学历的高低整体上呈现正相关的关系，即学历越高，薪资越高。博士生的平均薪资、最高薪资和最低薪资的中位数都是最高的；相反的，中专生和高中生的薪资最低。但仅仅从箱线图上看，还不能说明学历对薪资有统计意义上的显著影响。

此外，可以发现大专和本科的薪资有较多的离群值，可能这部分的薪资统计有误。硕士生和博士生的箱宽较大，说明硕士和博士毕业生的薪资分布的离散趋势比较大，可能是因为硕士和博士所从事的职位比较多元化。

本题所使用的R代码如下

1. 从岗位描述中提取对软件的要求，并使用箱线图展示不同软件要求对薪资的影响；提示：这些职位要求的软件不外乎R、SPSS、Excel、Python、MATLAB、Java、SQL、SAS、Stata、EViews、Spark、Hadoop这12种；

使用R语言中的grepl函数可以检索岗位描述中是否出现了所需要的软件。下面分别展示了是否要求12种软件对薪资的影响的箱线图：

从箱线图可以看出，是否要求Excel对于薪资基本没有影响，这可能是因为掌握Excel属于最基本的职业能力。类似的，是否掌握Stata对薪资的影响也比较小，这可能是因为Stata属于较容易入门的软件，入职后学习成本并不高。而像是否掌握Java，Spark，Hadoop等软件，则对薪资影响较高，掌握这些软件的求职者更可能获得较高的薪资，这可能与这些软件入门成本较高有关。

1. 以对数平均薪资为因变量，以地区、公司类别、公司规模、学历、经验要求以及12种软件需求为自变量，建立线性回归模型，探求这些变量对于薪资的影响

用R语言进行线性回归的R代码如下：

回归结果整理如下：

对回归结果的解读如下。控制其他因素不变时：

地区：河北的薪资最低，深圳和上海的薪资最高；河北的薪资平均比北京低41%，深圳和上海的薪资平均比北京高1.6%

公司类别：事业单位的薪资最低，创业公司的薪资最高，创业公司的薪资平均比事业单位高24%

公司规模：少于50人的公司的薪资最低，150-500人规模的公司薪资最高，少于50人的公司的薪资平均比1000-5000人规模的公司薪资低4.3%，150-500人规模的公司薪资平均比1000-5000人规模的公司薪资高1.8%。

学历：高中生的薪资最低，博士生的薪资最高。高中生的薪资比无学历要求的薪资怕平均低25.2%，博士生的薪资比无学历要求的薪资平均高94.9%。

经验：经验每增加一年，薪资平均增加9.8%。

软件掌握情况：掌握Excel和Matlab的薪资比无任何软件要求的薪资低，根据常识，这属于异常现象，可能是因为无任何软件要求的数据分析师的招聘信息太少，数据量不足，导致回归结果异常。掌握Hadoop的薪资最高，比无任何软件要求的薪资平均高21.4%。

线性模型的回归诊断示意图如下：

可以发现，残差图基本不存在非线性问题，说明对平均薪资取对数后去除了非线性问题。但是，残差图上仍然可以看出存在异方差的问题，说明对数线性回归模型仍有改善空间。在QQ图上，残差基本服从正态分布，但在右侧尾部向上偏离直线，可能说明数据中还有一些信息可供进一步挖掘。此外，在Cook’s distance中，没有超过1的点，说明在该模型下，没有离群的数据点。

8. 使用第7问得到的模型，进行模型预测准确度评估。需要满足以下要求：

1) 进行50次重复实验；

2）每次实验中，对数据做五折交叉验证（请设定每次划分样本的随机数种子，便于结果验证），得到五个预测误差结果；

3) 计算所有预测误差的均值（提示：对250个预测误差求平均）

（提示：对于样本量为N的样本，记预测值为 ，真实值为 ，则预测误差 = ）

完成本题的R代码如下：

在设定的随机数下，平均的预测误差为0.18969。