**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；

2. 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；

3. 除去中频次最高的5个行业类别，绘制剩余行业的词云图；

4. 以对数平均薪资为对象，进行简单的描述分析；

5. 使用箱线图，探究学历对岗位薪资的影响，对结果进行解读；

6. 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；

7. 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，探求这些比变量对于薪资的影响

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 = ）