《大数据开发技术（Hadoop）》

实训报告

主题：基于分布式 Hadoop 的智联招聘网站数据分析与可视化

组长：黄骏锋

成员：杨熠铭，范文镜

完成日期：2024年1月5日

目录

1.项目实现目的 1

1.1项目目的 1

1.2 项目实训环境 1

1.3 编程语言 2

2.项目实现思路介绍 2

2.1整体实现思路 2

2.2实现步骤介绍 3

3.项目实现流程介绍 6

3.1数据收集部分 6

3.2数据存储部分 7

3.3数据预处理部分 8

3.4数据分析与可视化部分 11

4.项目分析结果说明 15

4.1数据分析的结论 15

# 1.1项目实训目的

* 掌握完全分布式 Hadoop 集群和Spark的搭建方法和原理，并学会基础的配置环境操作。
* 熟练运用网络爬虫技术采集网站数据。
* 学会使用 Hadoop 和Spark进行数据存储、数据预处理和分析。
* 掌握数据可视化的方法和工具，能够将分析结果直观地展示出来。

## 1.2项目实训环境

### **实训环境**

* ****操作系统****：CentOS 7
* **Hadoop 版本**：Hadoop 2.7.3
* **Spark 版本**：Spark2.4.8
* ****编程语言****：Python 3.6
* ****数据可视化工具****：Spark、Matplotlib

## 1.3 编程语言

### **1.Java**

### **2.Python**

### **3.SQL**

### **4.XShell 脚本语言**

# 2.项目实现思路与步骤介绍

## 2.1整体实现思路

#### **完全分布式 Hadoop 集群搭建**

#### **完全分布式Spark集群搭建**

#### **3. 数据采集**

#### **4. 数据存储**

#### **5. 数据预处理**

#### **6. 数据分析**

#### **7. 数据可视化**

## 2.2实现步骤介绍

#### 1. 完全分布式 Hadoop 集群搭建

* ****安装前准备****：在每台节点服务器上安装 CentOS 7 操作系统，配置好网络，确保各节点之间能够相互通信，并安装 Java Development Kit（JDK）。
* ****配置集群节点****：修改每台节点的主机名，配置 hosts 文件，将集群中各节点的主机名和 IP 地址进行映射。
* ****安装 Hadoop****：在主节点上下载并解压 Hadoop 安装包，然后配置 Hadoop 的核心文件，如 hadoop-env.sh、core-site.xml、hdfs-site.xml、mapred-site.xml 和 yarn-site.xml 等，设置文件存储路径、节点角色等相关参数。
* ****启动集群****：在主节点上使用 start-dfs.sh 和 start-yarn.sh 命令启动 Hadoop 集群，通过 jps 命令检查各节点上的进程是否正常启动，包括 NameNode、DataNode、ResourceManager 和 NodeManager 等。

2.Spark集群搭建

* **系统准备**
* **安装 CentOS 系统**：在服务器或虚拟机上安装 CentOS 操作系统，配置分区、网络等参数，设置静态 IP 地址。
* **更新系统**：使用sudo yum update -y命令更新系统软件包。
* **配置主机名和 hosts 文件**：为每个节点设置不同主机名，编辑/etc/hosts文件添加节点 IP 地址和主机名映射关系。
* **安装和配置 Java 环境**
* **安装 JavaJDK**：可从 Oracle 官网下载或使用 yum 安装 OpenJDK，如sudo yum install -y java - 11 - openjdk - devel。
* **配置 Java 环境变量**：编辑/etc/profile文件，添加JAVA\_HOME和PATH环境变量，使配置生效。
* **安装和配置 SSH**
* **安装 SSH 服务**：使用 yum 安装 SSH 服务，启动并设置开机自启。
* **配置 SSH 免密码登录**：在 Master 节点生成 SSH 密钥对，将公钥复制到 Worker 节点，实现免密码登录。
* **安装和配置 Spark**
* **下载 Spark**：从 Spark 官方网站下载适合版本，如spark - 2.4.8 - bin - hadoop2.7.tgz，使用wget命令下载到指定目录，解压并重命名。
* **配置 Spark 环境变量**：编辑/etc/profile文件，添加SPARK\_HOME和PATH环境变量，使其生效。
* **配置 Spark 集群文件**
* 配置spark - env.sh文件：复制模板文件，编辑添加 Java Home、Spark Master 地址和端口等信息。
* 配置slaves文件：复制模板文件，编辑添加从节点名称。
* **配置 HistoryServer（日志服务）**
* 复制spark - defaults.conf文件，编辑添加日志相关配置，指定 Spark 将日志输入到 HDFS 中。
* 在spark - env.sh文件末尾添加 HistoryServer 启动参数，使其读取 HDFS 中的 Spark 日志。
* 创建 HDFS 中的日志目录，使用hdfs dfs - mkdir - p /spark\_log命令。
* **分发和运行**
* **分发 Spark 安装包**：使用scp - r命令将 Spark 安装包分发到其他节点。
* **启动 Spark 集群**：切换到 Spark 目录，使用sbin/start - all.sh启动 Master 和 Slaves，使用sbin/start - history - server.sh启动 HistoryServer。若启动报错，需检查并修改spark - env.sh文件中的错误（如将Export改为export），然后重新分发并启动。
* **测试**：打开浏览器，输入hadoop01:8080查看 Spark 集群管理页，确认集群搭建成功。

#### 3. 数据采集

* ****选择网站和确定采集目标****：选择智联招聘网站作为数据采集来源，目标是采集公司名称、工作类型、学历限制、薪酬、地区等信息。
* 对目标网站进行数据爬取：使用八爪鱼数据采集软件对智联招聘网站进行数据爬取。

#### 4. 数据存储

* ****创建 HDFS 目录****：在 Hadoop 集群的 HDFS 中创建用于存储招聘网站数据的目录。
* ****传数据到 HDFS****：使用 Hadoop 的命令行工具，将采集到的数据集上传到 HDFS 的目录下。

#### 5. 数据预处理

* 数据清洗：检查数据集中有无缺失值，并且剔除不需要使用到的数据，再根据数据分析的需求对数据进行转换和裁切

#### 6. 数据分析

* 对各个城市的工作岗位招聘数量进行分析
* 对与计算机相关的工作类型及占比进行分析
* 对智联招聘网站中北上广深四大发达城市的岗位招聘学历限制条件进行分析
* 对计算机相关领域中工作经验与薪资之间的关系进行分析

#### 7. 数据可视化

* 使用maplotlib和pandas库进行数据的柱状图、饼图、峰状图、折线图绘制，并对图分析讨论，结合实际情况得出最终分析结果

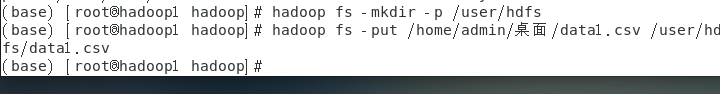
# 项目实现流程介绍

## 3.1数据收集阶段

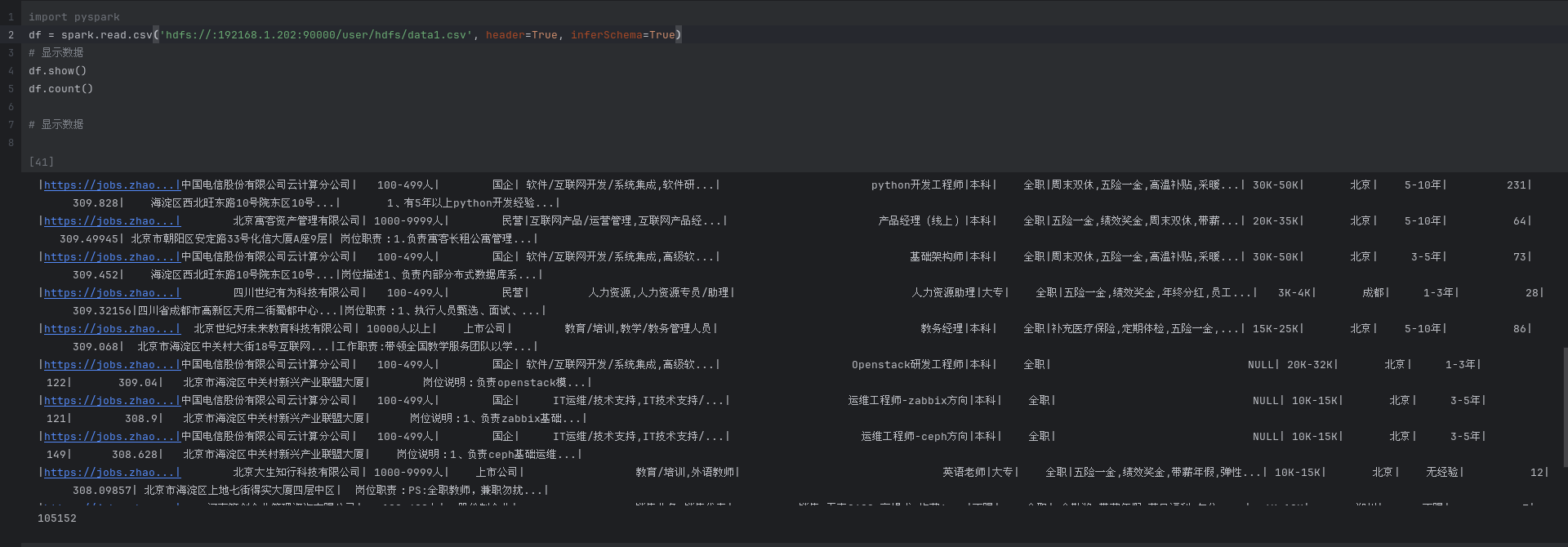
* 在八爪鱼软件上设置好需要提取的信息后，启动采集并导出数据为csv格式如图所示（下图为数据集中的部分信息示例，数据来源于智联招聘网站）



## 3.2数据存储阶段



* 将采集到的数据data1.csv上传到hdfs中



* 使用pyspark从hdfs中获取数据进行分析

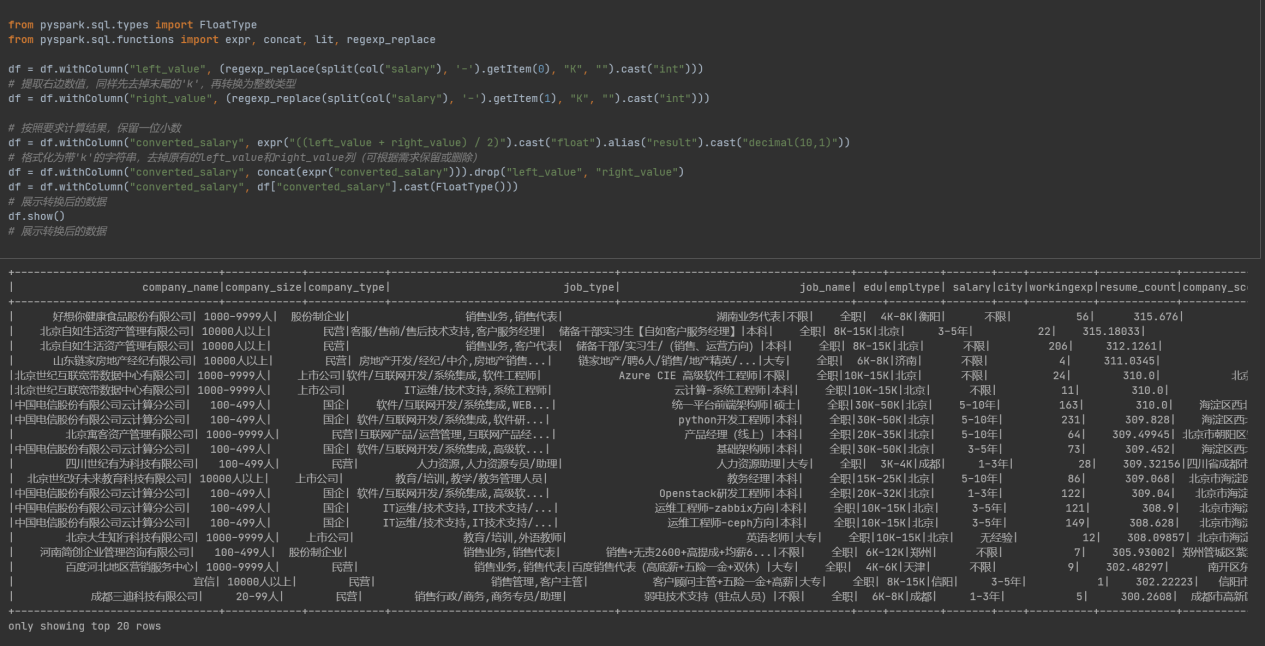
## 3.3数据预处理阶段



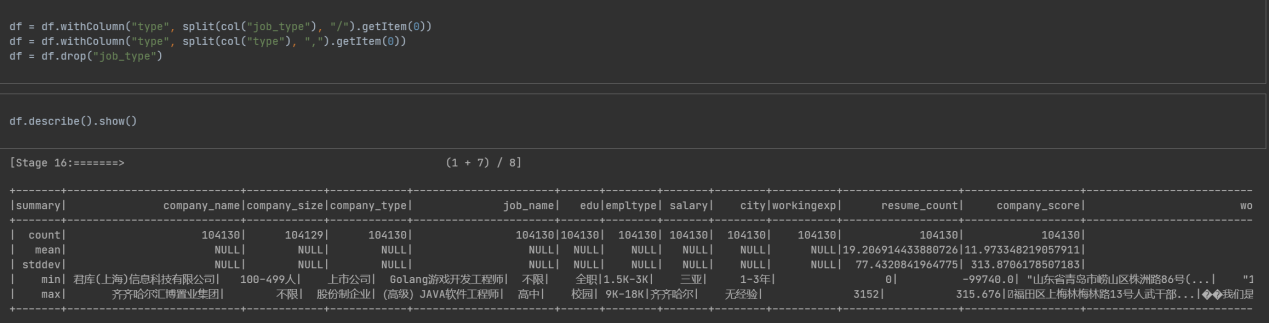
* 如图所示代码主要实现了使用 PySpark 从一个 CSV 文件中读取数据，创建一个 DataFrame，然后展示数据内容并统计数据的行数



* 如图所示代码主要实现了从 PySpark 的 DataFrame 中删除指定的列 tag 和 url，删除不需要的列，以便后续的数据分析和处理工作，然后展示删除列后的 DataFrame 内容。

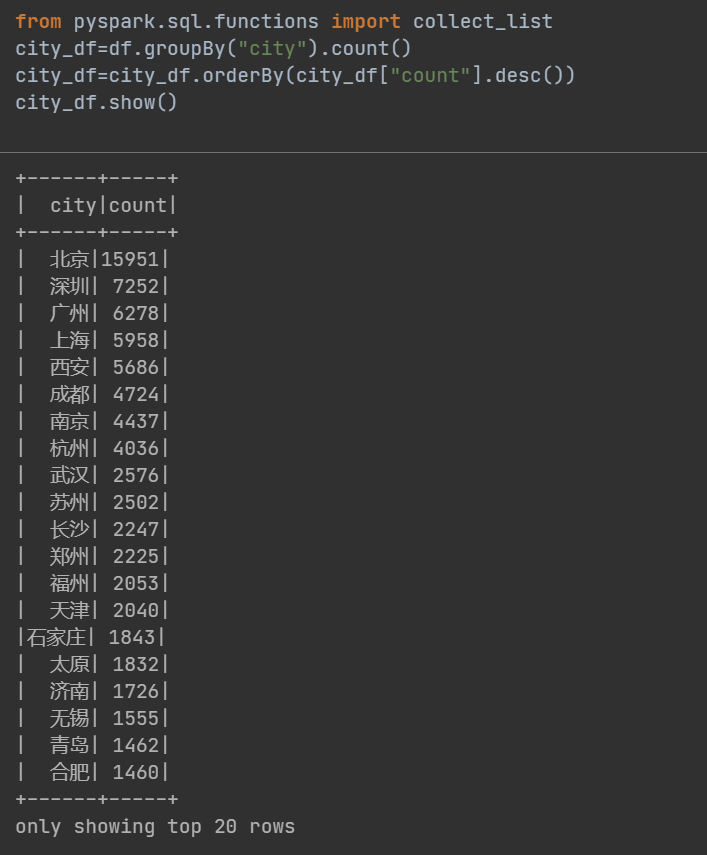


* 如图所示代码主要是对一个包含薪资范围数据的 PySpark DataFrame 进 行数据处理和转换，具体操作包括：**提取并转换左边数值、提取并转换右边数值、计算平均值并格式化格、式化为带 'k' 的字符串并删除多余列、转换数据类型并展示数据**



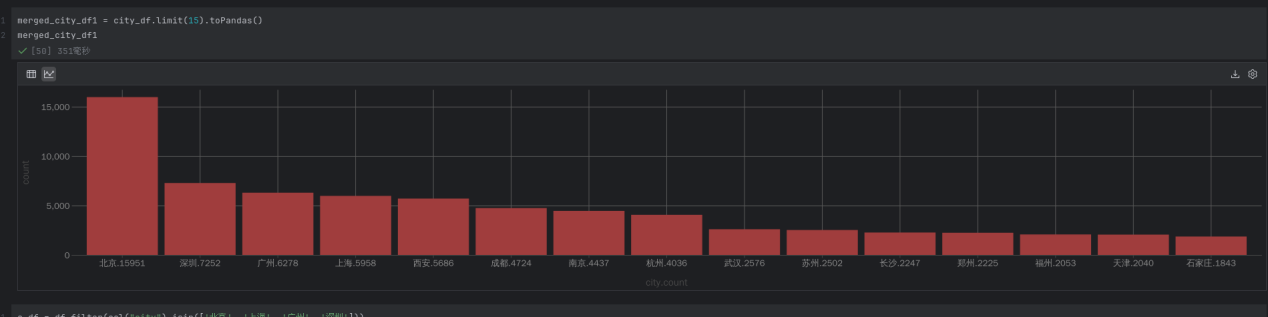
* 这段代码主要是在 PySpark 中对 DataFrame 进行列的提取、转换和删除操

作，最后展示数据的描述性统计信息。

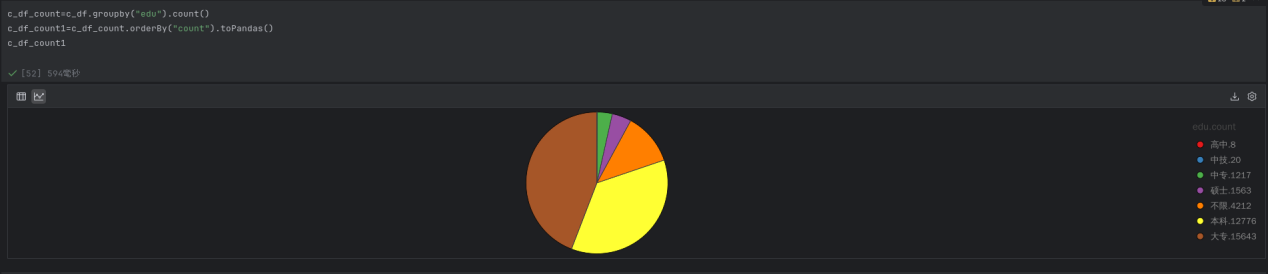


* 这段代码的主要功能是从 PySpark 的 DataFrame 中按照 "city" 列进行分组，计算每个城市的记录数量，然后按照记录数量降序排列并展示结果。

## 3.4数据分析与可视化阶段



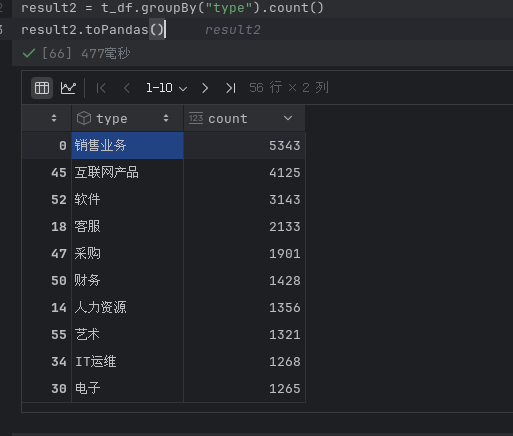
* 如图所示，这是对数据集中的工作岗位分布进行降序排序后得到的柱状图，由图可知各个城市的工作岗位招聘数量。由图北京、上海、广州、深圳等一线城市，在智联招聘上的职位数量通常较为丰富，涵盖了各个行易知业和领域。



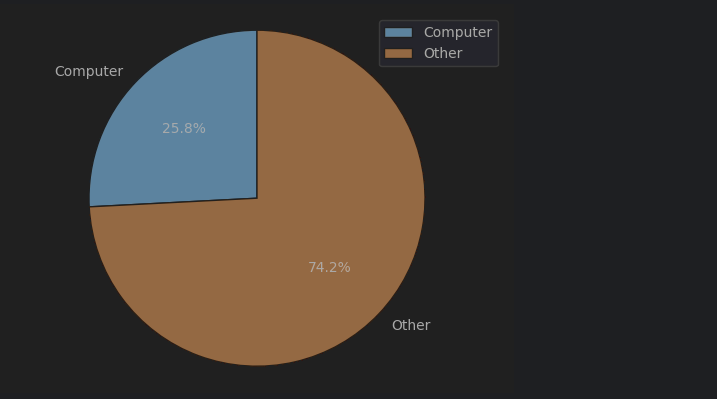
* 上图是对智联招聘网站中所有城市的岗位招聘学历限制条件进行数据可视化，由图可知，大部分工作的招聘条件还是大专及本科以上，大专生、本科生仍是就业市场的绝对主体



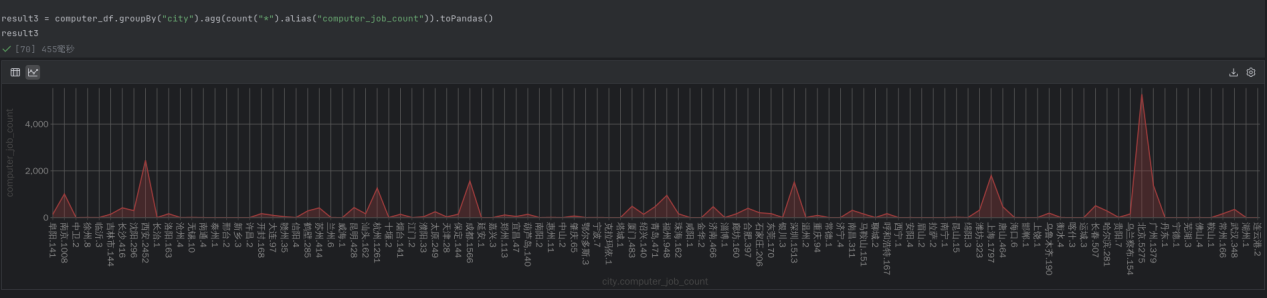
* 以上根据受教育程度：学历来观察不同学历的平均薪资之间的差异，由图显而易见随着学历增高平均薪资也在逐渐增高。



* 上图代码是对与计算机相关的工作类型进行降序排列，在北上广深四个城市中，与计算机相关的工作占比突出，并且由图可知各行各业的运转都与计算机息息相关，计算机技能与是大部分工作都是接轨的。



* 上图是对招聘网站中的所有招聘岗位进行数据统计制作的分布饼图，在全部的工作类型中，与计算机相关的工作占比25%，由此可知，计算机仍是当前就业环境的主流



* 这张图片展示了一个峰状图，用于分析不同城市的计算机相关岗位数量。
* **岗位数量较多的城市**：
  + 北京的计算机岗位数量最多，远远超过其他城市，柱状图高度突出，接近 5000 个岗位。
  + 上海、深圳、广州等一线城市的计算机岗位数量也相对较多，分别有一定数量的岗位，仅次于北京。
  + 一些新一线城市或省会城市，如杭州、成都、武汉等，也有较多的计算机岗位，在图中表现为相对较高的柱状图。
* **岗位数量较少的城市**：
  + 大部分其他城市的计算机岗位数量相对较少，柱状图高度较低，很多城市的岗位数量在几百甚至几十的水平。例如，阜阳、临沂等城市的岗位数量较少。



* 这张图片展示了一个折线图，用于分析计算机相关领域（基于智联招聘等平台的数据）中工作经验与薪资之间的关系，同时还区分了不同学历的情况。（蓝绿色折线是研究生，红色折线是本科生，黄色折线是大专，蓝色折线是中专，绿色折线是高中）
* ****工作经验与薪资的总体趋势****：
* 对于大多数学历层次，薪资随着工作经验的增加而呈现上升趋势。例如，在 5 - 10 年工作经验区间，各学历层次的薪资普遍达到较高水平，尤其是蓝绿色折线（代表研究生）在该区间薪资达到峰值接近 40k。
* 工作经验在 10 年以上时，蓝色折线（代表中专）的薪资有明显的大幅上升，超过了其他学历层次在该经验区间的薪资水平，可能反映了该学历层次在特定工作经验阶段的薪资优势或行业需求特点。
* ****不同学历层次的薪资差异****：
* 在工作经验不限的情况下，各学历层次的薪资起点较为接近，都在 10k 左右。
* 随着工作经验的增加，不同学历层次的薪资差异逐渐显现。例如，在 5 - 10 年工作经验区间，蓝绿色折线（代表研究生）的薪资高于其他学历层次，而红色折线（代表本科生）在某些区间也表现出较高的薪资水平。
* 在无经验和 1 年以下工作经验区间，各学历层次的薪资水平较为接近且较低，说明刚进入职场时，学历对薪资的影响相对较小，工作经验的积累更为重要。

# 4.项目分析结果说明

## 4.1数据分析的结论

### **一、行业需求与趋势**

* ****热门行业****
  1. 通过分析智联招聘上各行业的职位发布数量，可以发现互联网、金融、医疗健康、新能源等行业的职位数量相对较多，反映出这些行业目前处于快速发展阶段，对人才的需求较为旺盛。例如，互联网行业的技术岗位（如软件工程师、数据分析师等）招聘需求持续增长，这与数字化转型和线上业务的蓬勃发展密切相关。
  2. 而一些传统行业，如制造业、销售等，职位发布数量相对稳定或略有下降，可能受到产业升级和市场竞争的影响，对人才的需求结构也在发生变化，更加注重招纳与计算机相关的人才。
* ****新兴行业与岗位****
  1. 随着科技的发展，一些新兴行业如人工智能、区块链、大数据等领域的职位逐渐增多。这些行业的岗位往往对专业技能要求较高，薪资待遇也相对优厚，吸引了大量求职者的关注。例如，人工智能领域的算法工程师、机器学习工程师等岗位，虽然门槛较高，但市场需求增长迅速，成为就业市场的新亮点。

### **二、地域差异**

* ****一线城市就业机会与竞争****
  1. 像北京、上海、广州、深圳等一线城市，在智联招聘上的职位数量通常较为丰富，涵盖了各个行业和领域。这些城市经济发达，企业众多，尤其是大型企业和跨国公司的总部或分支机构集中，提供了多样化的就业机会。
  2. 然而，一线城市的就业竞争也非常激烈。一方面，大量的求职者涌入这些城市，包括本地高校毕业生和外地求职者；另一方面，这些城市的生活成本较高，对求职者的综合素质和能力要求也更高。求职者需要具备较高的学历的同时还要有较强的专业技能、丰富的实践经验和良好的综合素质才能脱颖而出。

### **三、学历与技能要求**

* ****学历门槛****
  1. 从智联招聘的数据来看，同一城市的不同岗位对学历的要求存在差异。一些高端技术岗位、管理岗位和研发岗位通常要求本科及以上学历，甚至硕士、博士学历在某些领域更具竞争力。例如，金融行业的投资分析师、科研机构的研发人员等岗位，高学历往往是进入的门槛之一。
  2. 然而，也有一些行业和岗位对学历的要求相对较低，更加注重实际技能和工作经验。例如，汽车销售、制造业的一线操作岗位等，可能更看重求职者的技能水平和工作态度。
* ****技能需求****

1.除了学历，企业越来越注重求职者的技能水平。同时，随着技术的发展，与行业相关的专业技能需求也在不断变化。例如，在互联网行业，掌握编程语言（如 Python、Java 等）、数据分析工具、云计算等技能的求职者更受欢迎；在新媒体行业，具备内容创作、视频剪辑、社交媒体运营等技能的人才需求较大。此外，一些软技能如创新能力、适应能力、领导力等也逐渐成为企业选拔人才的重要考量因素。

### **四、薪资水平与福利待遇**

* ****薪资分布****
  1. 智联招聘的数据显示，不同行业、地域和岗位的薪资水平存在较大差异。一般来说，一线城市的薪资水平相对较高，尤其是一些热门行业和高端岗位。例如，互联网行业的技术岗位、金融行业的核心岗位等，薪资待遇较为优厚，年薪可能达到数十万元甚至更高。
  2. 而一些传统行业和二三线城市的薪资水平相对较低，但也在逐步提高。此外，不同学历和工作经验的求职者薪资也有所不同，通常学历越高、工作经验越丰富，薪资水平也越高。例如，硕士毕业生的起薪往往高于本科毕业生，具有 5 年以上工作经验的求职者薪资也会比应届毕业生高出不少。
* ****福利待遇****
  1. 除了薪资，企业提供的福利待遇也成为求职者关注的重点。一些企业提供完善的五险一金、带薪年假、定期体检、培训机会、员工旅游等福利，这些福利不仅能够提高员工的生活质量和工作满意度，也有助于企业吸引和留住人才。例如，一些大型企业和外资企业在福利待遇方面相对较为优厚，而一些中小企业可能在福利方面相对简单，但可能会提供更具挑战性的工作机会和发展空间。

### **五、就业市场的挑战与机遇**

* ****挑战****
  1. 就业市场竞争激烈，尤其是在热门行业和一线城市，求职者面临着较大的压力。如何在众多求职者中脱颖而出，是一个需要思考的问题。
  2. 行业变化快速，一些传统行业面临转型和升级，可能导致部分岗位的需求减少或消失，求职者需要不断学习和更新知识技能，以适应市场的变化。例如，随着自动化技术的发展，一些制造业的流水线工人可能面临失业风险，需要学习新的技能以转型到其他岗位。
  3. 毕业生数量逐年增加，就业压力较大。应届毕业生往往缺乏工作经验，在求职过程中可能会遇到一些困难，需要通过实习、培训等方式提升自己的竞争力。
* ****机遇****
  1. 新兴行业和领域的发展为求职者提供了更多的选择和机会。例如，人工智能、大数据等领域的兴起，创造了大量新的就业岗位，对于具备相关技能和知识的求职者来说，是一个很好的发展机遇。
  2. 远程办公和灵活就业模式的兴起，为求职者提供了更多的工作方式选择。一些求职者可以通过远程办公的方式，为不同地区的企业工作，拓宽了就业范围和机会。同时，一些自由职业者和创业者也可以利用互联网平台，开展自己的业务，实现自我价值。
  3. 政府和企业对人才培养和发展的重视，提供了更多的培训和学习机会。例如，一些地方政府推出了职业技能培训补贴政策，鼓励求职者提升自己的技能水平；企业也越来越注重内部培训和人才发展，为员工提供晋升和发展的通道。

综上所述，目前的就业环境既有挑战也有机遇。求职者需要根据自身的情况和市场需求，不断提升自己的学历和技能水平，关注行业发展趋势，选择适合自己的职业发展道路。同时，政府和企业也需要共同努力，创造更好的就业环境和发展机会，促进就业市场的健康发展。