

摘要

本研究旨在设计并实现一个基于机器学习和统计分析的招聘广告大数据可视化系统，以应对就业市场快速变化的需求。该系统采用数据挖掘、聚类分析等技术，处理数十万条招聘信息，实现了职位差异度量化评估、多角度职位画像展示、薪酬待遇模式分析、地域招聘活动特征挖掘，以及行业发展动态和新兴职位识别等功能。研究发现，职位差异可通过多角度多维属性有效量化，薪酬待遇与职位、行业、地域呈现显著相关性，地域招聘活动展现出明显的产业集群特征。同时，基于薪酬水平、学历和经验要求能发现急需人才的新兴职位。本研究创新性地构建了一个集数据处理、多维分析、可视化展示于一体的平台，为求职者提供决策支持，并为企业招聘活动提供数据依据。未来研究将致力于优化预测算法和拓展多源异构数据分析，以提高系统的分析精度和预测能力。

关键词：招聘大数据数据可视化，机器学习与统计分析，多维职位画像

1 假设

表 1-1 研究假设

假设名称	假设内容
数据代表性	本研究所使用的数据集是对真实就业市场的有效抽样，能够反映实际招聘环境的主要特征和规律。
教育-经验-薪酬关联	在整体就业市场中，教育水平和工作经验与薪酬水平呈正相关关系。
薪酬区间有效性	招聘信息中提供的薪酬范围能够有效反映该职位的实际薪酬水平，并包含了合理的薪酬波动空间。
截面数据动态性	本研究使用的数据集虽为特定时间点的截面数据，但恰好捕捉了部分行业正在经历显著变革的动态特征。
技能需求真实性	招聘信息中的技能要求准确反映了当前就业市场对这些技能的实际需求。
地理经济差异	不同地区的经济发展水平差异直接导致了各地区间的薪酬水平差异。
招聘频率-需求关系	特定职位或行业的招聘信息发布频率与其人才需求程度呈正相关关系。
行业薪酬异质性	不同行业间存在显著的薪酬水平和结构差异，这反映了各行业的特性和市场需求状况。
薪酬面议含义	标注为“薪酬面议”的职位往往代表了更高级别或更专业化的岗位。
市场信息不对称	就业市场中存在信息不对称现象，求职者和雇主可能无法获得完全一致的市场信息。

2 数据介绍与处理

2.1 原始数据概览

本作品所使用的赛事官方招聘数据集包含 430,664 条招聘通知，涉及 169,540 个独立岗位。数据集覆盖 371 个城市，267,296 家企业，和 158 个不同的行业类别，充分反映了就业市场的多样性和复杂性。数据集还包含了 10 种不同的工作经验要求和 12 种学历要求类别，为分析职位要求提供了丰富的维度。

表 2-2-1 原始数据概览

数据特征	数量
总记录数	430,664
独立岗位数	169,540
涉及城市数	371
企业数量	267,296
行业类别数	158
经验要求类别	10
学历要求类别	12
薪酬表达形式	—

注：薪酬表达形式包括年薪制（标明薪资结构如 12 薪、13 薪等）、月薪制、周薪制、日薪制、时薪制、单结制（按单位工作量计酬）以及面议。

2.2 数据预处理与清洗

本作品依据招聘通知数量对原始数据中各职位进行排序。排序结果显示，有 148201 个职位仅具有一条招聘广告，数据具有极强的长尾效应。由于难以从招聘通知数量过少的数据中分析出有意义的信息，最终选取在招聘通知数量排序中前 1% 的职位作为分析对象。

2.2.1 数据划分

1. 薪资数据划分

原始数据包含年薪、月薪、周薪等其中不同薪酬表达形式。为了使分析更加直观，本作品依据《中华人民共和国劳动合同法》将薪酬数据划分为固定用工和灵活用工两类。

为统一薪资的表示模式，本作品通过正则表达式匹配抽取每条数据的薪资上界和下界，同时依据劳动和社会保障部颁布的《关于职工全年月平均工作时间和工资折算问题的通知》，将不同薪酬模式换算为年薪。对于单结及面议的薪酬信息，由于其不确定性高且数据量少，故按同职位其他薪酬的平均数处理。

2. 地域数据划分

基于原始数据，本作品依据行政区划首字母将地域信息划分为一级行政区划和二级行政区划。即认为属性“city”中所有首字母相同的二级行政区划属于同一类一级行政区划。本作品认为各省份内部经济状况相似，不同地区间的经济差异直接影响薪酬水平。

2.2.2 数据清洗

考虑到灵活用工和固定用工的内在差异，本作品基于统一的薪酬模式，对两类形式的招聘信息分别进行数据清洗。首先基于一级行政区划与职位将招聘信息分为若干小组，对于分组内的招聘信息，本作品采取四分位距（IQR）方法检测异常值并排除。公式如下：

$$L = Q_1 - 1.5 \times IQR$$

$$U = Q_3 + 1.5 \times IQR$$

其中， Q_1 为第一四分位数（第 25%分位数）， Q_3 第三四分位数（第 75%分位数）， L 为合理数值下界， U 为合理数值上界。

2.2.3 数据语义识别

数据清洗后，本作品统计各经验要求与学历要求在数据中的分布情况，并与从招聘网站上爬取的 600 万条真实招聘数据的经验要求、学历要求进行对比。结合中位数、平均数等进行学历、经验要求虚拟代码的语义分析，得出虚拟代码与实际内涵的语义对应关系。

同时，依据我国对中、高等教育的划分，将高等教育及以上学历要求定义为高学历要求，中等教育以上学历要求定义为中学历要求，中等教育以下定义为低学历要求。依据经验要求数据，本作品以五年为分界线，将经验要求在五年及以上的定义为高经验要求，将经验要求在五年以下的定义为低经验要求。对应如下：

表 2-2 学历要求虚拟代码与语义对应

学历	Gh	Go	GP	GI	Gy	Gx	GZ	Gz
语义对应	博士	硕士	本科	大专	高中	中专	初中	不限学历
学历划分	高学历要求				中学历要求		低学历要求	

表 2-3 经验要求虚拟代码与语义对应

经验	ESu	Eby	EzN	EaZ	EdD	Eqh	Eas	EKk
语义对应	10 年及以上	8-9 年	5-7 年	5 年及以上	3-4 年	2 年	1 年	无需经验
经验划分	高经验要求				低经验要求			

其中，Gw、Gi、GJ 三类学历要求未能在真实数据集中找到合理对应，但由于具有这三种学历的招聘信息数量极少，去除这三类学历要求对整体分析无较大影响，故本作品对上述三类学历要求不予以考虑。

2.3 最终数据构建与特征

经过数据的预处理和清洗，本作品得到了两类最终数据：**概览数据**和**属性视图数据**。概览数据包含 166,118 条招聘记录，涉及 1,713 个独立岗位。数据覆盖 26 个一级行政区划、370 个二级行政区划，136,149 家企业，158 个不同的行业类别，包含 8 种不同的工作经验要求和 8 种学历要求类别。

表 2-4 最终数据构建形式

数据名称	数据意义	数量
data	招聘信息	166,118
_id	招聘信息涉及的岗位，每种岗位分配唯一的虚拟代码	1,713
province	招聘信息涉及的一级行政区划	26
city	招聘信息涉及的二级行政区划	370
salary	招聘信息的薪酬数据，一个包含上下界信息的字符串	-
salary_type	招聘信息薪酬模式，每种薪酬模式分配唯一虚拟代码	7
salary_base	固定用工招聘信息的薪酬期数	-
salary_upper_bound	统一为年薪后招聘信息的薪酬数据的上界	-
salary_lower_bound	统一为年薪后招聘信息的薪酬数据的下界	-
company	招聘信息涉及的公司，每个公司分配唯一的虚拟代码	136,149
companyType	招聘信息涉及的行业，每个行业分配唯一的虚拟代码	158
education	招聘信息的学历要求	11
experience	招聘信息的经验要求	8
category	招聘信息的技能要求，由学历要求和经验要求决定	4

经过数据清洗与预处理，我们最终得到两类概览数据：**固定用工类概览数据**和**灵活用工类概览数据**。两类数据均全面且深入地反映了职位的学历要求、经验要求、薪酬模式、薪资上下限、地域、行业、等多维度内容。

为探索各维度信息间的关系，本作品基于不同视角，构建了一系列的属性视图数据，主要概括如下：

2.3.1 职位属性视图数据

职位属性视图数据集中关注不同职位在就业市场中的需求情况，是本作品对就业市场进行分析的重要基础。该数据包含每个职位的薪酬范围数据、每个职位的学历要求与经验要求分布区间以及各职位学历要求与经验要求的众数。

2.3.2 地域属性视图数据

地域属性视图数据反映各地域内部招聘信息情况，同时为跨地域招聘信息比较研究提供了依据。数据涵盖一级行政区划所包含的二级行政区划数量数据、各地域（包括一级和二级行政区划，下同）所包含的职位数量和招聘信息数量数据、各地域薪资统计数据。除一级行政区划包含二级行政区划数量数据外，上述所有数据均具备灵活用工、固定用工两个版本，以便分别分析。而各地域薪资统计数据则进一步地包含薪资上下限、招聘信息中最高薪资和最低薪资的平均值、中位数、25 和 75 分位数等。

2.3.3 行业属性视图数据

与职位属性视图数据类似，行业属性视图数据用于展现不同行业在就业市场中的样貌，相关数据主要包括每个行业的薪酬分布数据、学历/经验要求分布数据，以及不同行业的平均薪资等薪酬数据。

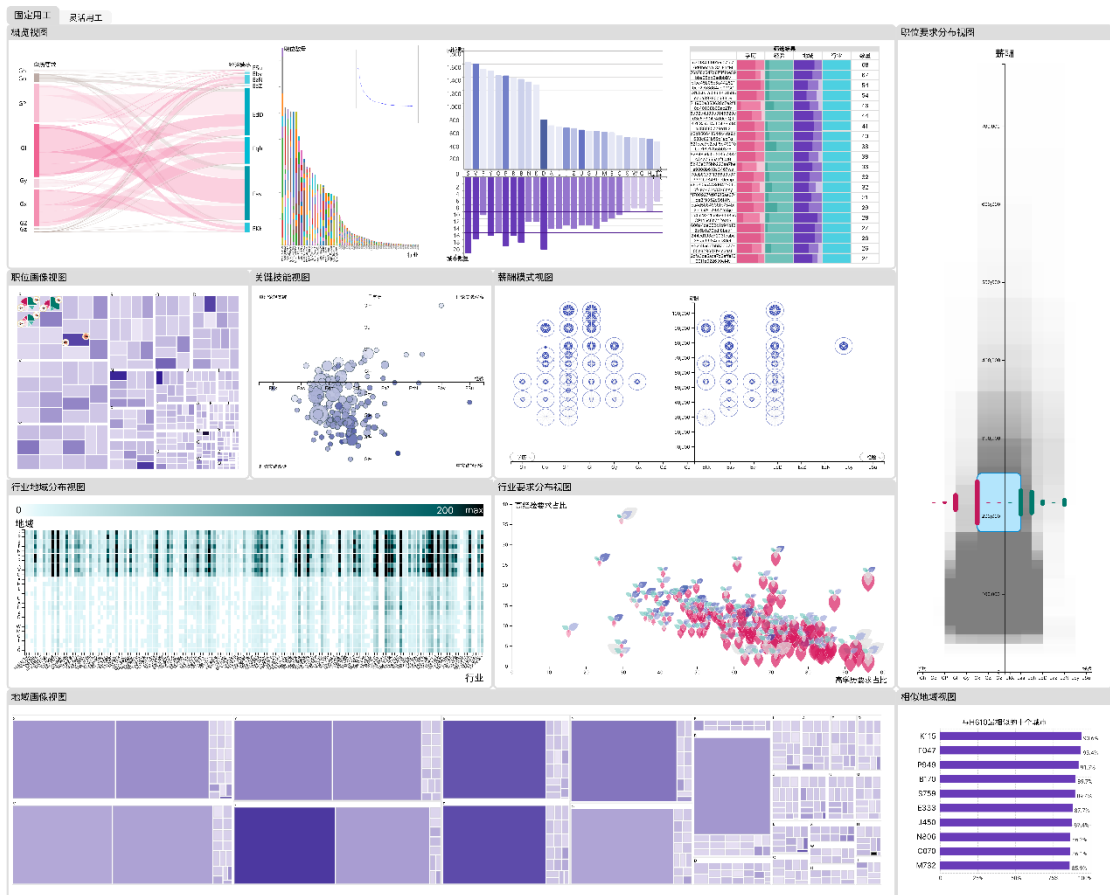
2.3.4 行业-地域视图数据

本作品认为不同的地域的招聘市场因经济等原因具备差异性，因此有必要基于不同的地域对具体的行业情况进行研究。行业-地域视图数据主要包括行业在不同地域中的分布情况数据、各行业在不同地域中的招聘数量数据等。

3 可视化系统设计

3.1 系统概览

本作品设计并实现了一个市场招聘信息可视化分析系统。该系统包含多种视图，采用合理的视觉编码，确保信息的可读性和易理解性。系统支持多种交互功能，使用者能够深入且个性化地探索招聘信息的多维度细节。通过这些功能，使用者可以全面直观地感知和理解市场招聘动态，为企业招聘与求职者决策提供可行建议。



图表 3-1 市场招聘信息可视化分析系统总览图

市场招聘信息可视化分析系统最上方设有 1 个职位类型切换选项，为系统的一级筛选条件。用户在选定“固定用工”或“灵活用工”后，视图中展现结果均来自与该类型下的招聘数据。

可视化系统分为 9 个视图，共包含 5 个部分。

A 为概览视图，包括学历要求-经验要求桑基图、地域-职位城市数量双向条形图、行业类别与职位数量堆积条形图和职位比较列表，主要用于招聘市场整体规律展现。此外，概览视图中各个维度均可由使用者点击筛选，起到展示条件筛选作用。

B、C 为职位画像视图和关键技能视图，用于展现选定职位的关键特征分析。

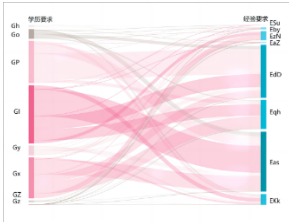
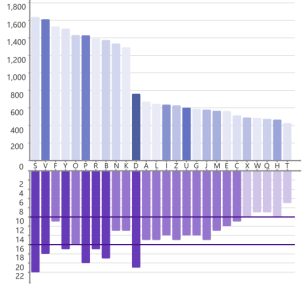
D 为薪酬视图，用于展示选定职位在选定维度下的薪酬水平变化、薪酬模式，以及其他同筛选条件下职位的薪酬分布对比。

G、H、分别为地域画像视图和相似地域视图，其中 G 可通过交互分别展现一级地域区划视图和二级地域区划视图，用于展现地域招聘活动细节与具有相似招聘活动的地域。

E、F、I 分别为行业地域分布视图、行业要求分布视图和职位要求分布视图。其中，E、F 用于分析行业发展动态，I 用于识别与佐证可能的急需人才的新兴职位。

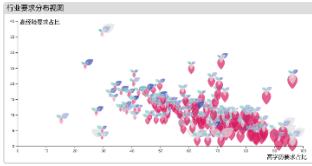
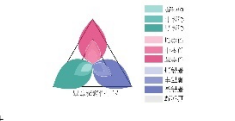
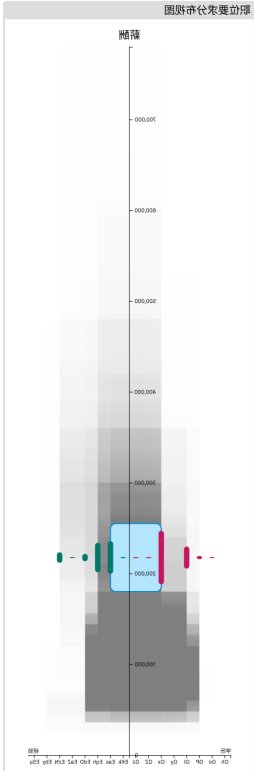
通过上述视图的配合使用，整个系统能够为使用者提供全方位、多粒度的市场招聘分析结果，从而为招聘决策和求职决策提供支持。

3.2 可视化词典

任务	视图		总览	细节拆分	设计内容	交互模式
T1	A	A1	<div></div> <p>学历要求-经验要求桑基图 展示各职位学历要求和经验要求间的变化关系。</p>	—	A1 左右侧分别以粉红色和蓝绿色色相表示学历要求和经验要求，颜色明度代表了学历要求及经验要求的不同类别，流线的粗细表示了数据量的多少，直观展现学历要求及经验要求两个维度间的关系。	点击 A1 两侧的不同标签，选中相应的学历要求和经验要求层次。
		A2	<div></div> <p>地域-职位城市数量双向条形图 展示地域维度上的职位数量和城市数量分布。</p>	—	A2 横轴上下侧分别用橙色和紫色色相区别职位数量分布和城市数量分布，各个地域的职位数量沿横轴正向方向依次递减，职位数量条形图的高度代表不同地域的职位数量，颜色明度用以表示不同行业薪资水平的相对高低。	点击坐标横轴上的地域标签，选中相应的地域。
		A3	<p>行业类别与职位数量堆积条形图 呈现行业维度上的薪资分布情况。</p>	—	A3 中各个地域的职位数量沿横轴正向方向依次递减，条形图的高度表示不同行业的职位数量，不同颜色明度用以区别同行业中的不同职位。	点击坐标横轴上的行业标签，选中相应的行业。

					显示具体的薪酬。不同色相分别代表日薪、周薪、时薪。	
	C	<div></div> <p>关键技能视图（C）展示选定职位所需要的实践技能和理论技能的分布，以及技能类型与薪酬的关系。</p>	—	<p>C 中横轴代表经验要求，纵轴代表学历要求，两坐标轴将视图划分为四个象限，分别对应四种不同的关键技能类型。</p> <p>图中每一个散点代表选定职位所在的一个行业类型，其所在位置又所有该行业招聘信息所在位置计算余弦相似度而得。散点大小代表岗位数量，颜色深度表示该行业平均薪酬。</p>		
T3	D	<div></div> <p>薪酬模式散点图（D）将薪酬模式区分为固定用工薪酬模式和灵活用工薪酬模式，分别展示薪酬水平与学历、经验、行业、地域的关系，以及不同筛选组合中薪酬随选中维度变化的规律。</p>	<div></div>	<p>D 中横轴正负半轴分别可选择一个维度加以展示。每个点代表筛选出的一条招聘信息。展示固定用工时，每个散点的中心圆的大小表示月薪高低，外环上圆的数量表示发放奖金的月份数量。展示灵活用工时，每个散点的不同颜色用以区分周薪/日薪/时薪，外圈的弧表示该条招聘信息的薪酬在同类型的所有数据中所处的相对位置。</p>	<p>在 D 横轴处下拉框点击选择横轴展示维度，坐标轴中展示筛选职位在相应维度下的散点分布。</p>	
T4	G	G1	<div></div> <p>总体地域画像视图（G1）通过矩形树图的形式展示所有地域间的层级关系。</p>	<div></div>	<p>G1 树图中，一个灰色框中包括的为个省，其中的矩形代表各个二级行政区划。矩形的大小体现各地域及二级行政区划间招聘数量的差异，面积越大，招聘数量越多。色彩明度表示各地域及二级行政区划间平均薪酬的差异，明</p>	<p>点击选中一个灰色框中的省，H 中展示与该省相似度最高的十个省的排序及其相似度。</p>

					度从高到低代表平均薪酬由高到低。	
		G2	 <p>筛选地域画像视图（G2）以矩形树图的形式展示所有地域间的层级关系，同时在矩形内嵌入三层环形图，展示对应二级行政区划的行业偏好与职位偏好，以及平均薪酬的高低。</p>	—	<p>G2 中树图的信息表现方式与 G1 一致。G2 内嵌的三层环形图中，外环和中间环展示招聘数量最高的 5 个职位和行业各自的占比。外环展示职位偏好（蓝色环），中间环展示行业偏好（靛蓝色环），颜色明度高低和弧长长短代表偏好程度，明度越低、弧长越长，代表偏好程度更高。环形图的内环展示平均薪酬高低，内环色彩明度越低，代表平均薪酬水平越高，反之薪酬水平越低。</p>	<p>在 G1 中点击选择一个省以后，G1 切换至 G2，显示该省所包含的二级行政区划，并显示三层环图。</p> <p>点击矩形选中二级行政区划，高该二级行政区划，同时 H 展示与该二级行政区划相似度最高的十个二级行政区划的排序及其相似度。</p>
		H	 <p>相似地域视图（H）以条形图形式在省级推荐相似省，在二级行政区划级推荐相似二级行政区划。</p>	—	<p>H 中显示与 G 中点击选择的省（市）相似度最高的十个省（市）。条形图长度按百分比数值降序排列，筛选出相似度更高的省市地域。</p>	<p>点击 H 中的条形，可以选中对应的省（市），G 中展示出对应省（市）的地域画像。</p>
T5		E	 <p>行业地域分布视图展示行业与地域的分布关系，显示不同行业的地域偏好，不同地域的行业偏好。</p>	—	<p>E 中横轴展现 158 个行业，从左到右按平均薪酬越来越低排序；纵轴 26 个省，从上到下按平均薪酬越来越低排序。</p> <p>每个格子颜色深度表示招聘数量。</p>	

	F	 <p>行业要求分布视图</p>	 <p>图例</p>	<p>F 图结合散点图与改编后的雷达图。每个花型散点代表一个行业，红色花瓣代表学历，绿色花瓣代表经验，蓝色花瓣代表薪资，灰色代表缺失值。横纵坐标轴分别为高学历/经验要求的招聘信息在该点对应行业中的占比。</p>	
	I	 <p>职位要求分布视图</p>		<p>I 图由箱线图改编而成。每个箱型色块表示一个职位。每个色块高度的跨度表示该职位主要涉及的学历范围和范围，宽度的跨度表示主要设计的薪酬范围。</p> <p>“线”依据离散型数据特性被改编为条带，每个条带对应横轴上的一种学历或经验，条带长度代表该学历/经验要求的占比。</p>	

4 问题解答

4.1 问题一

分析职位招聘信息，从行业类别、薪资待遇、经验要求等多维度属性量化评估职位差异度

4.1.1 分析与聚类

本作品以职位招聘信息为基础，从学历要求、经验要求、地域分布、行业类别、薪酬水平等多维度评估职位间的差异度。

考虑到综合各维度评估时，各维度权重难以确定与合理解释，且从实际情况考虑，横跨学历经验、行业、地域等多个维度共同比较职位间的差异度也较为少见，不具有实际意义。同时，本作品认为各维度上会存在多个集中的类别，每个类别的差异较小，故本作品对每个维度也进行了分组（表 2-2，表 2-3）。因此，本作品将维度分组聚合，从以下三个角度评估职位差异度：1. 学历、经验要求与薪资；2. 地域分布与招聘信息数；3. 行业分布。

针对角度一，本研究基于学历要求、经验要求和平均薪资水平对全体职位进行 K-means 聚类。聚类结果如下：

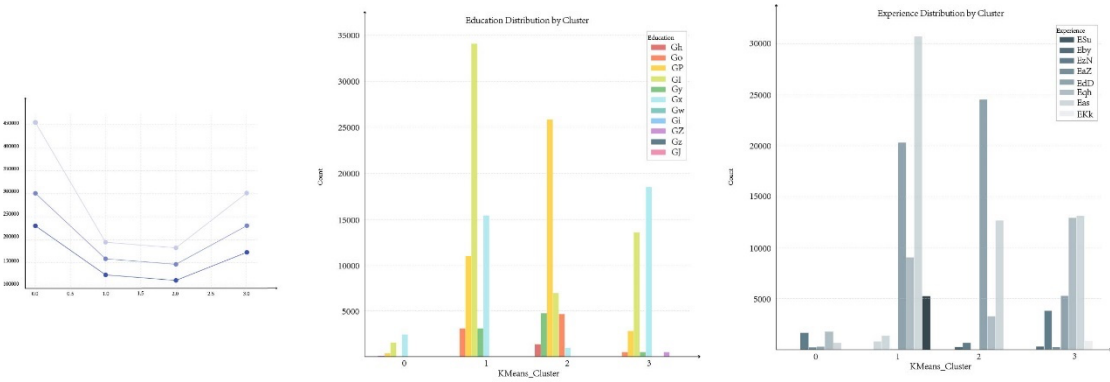


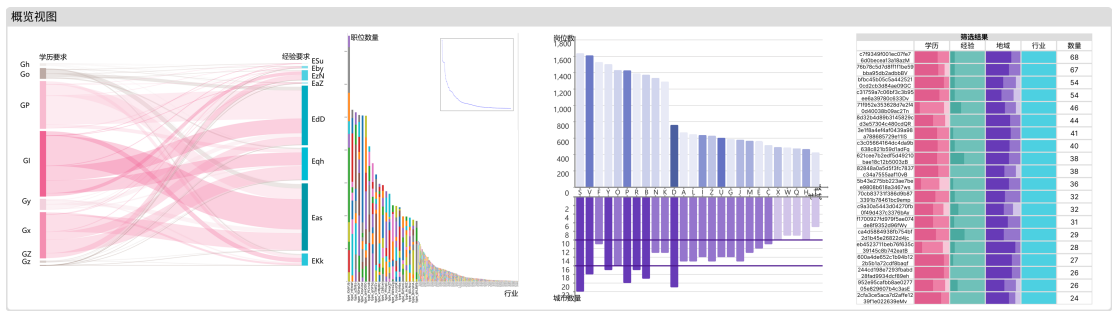
图 4-1 聚类结果薪资分布（左）、学历要求分布（中）、经验要求分布（右）

可以发现，综合考量上述三个维度后，职位被划分为具有差异的四个类型，薪资水平由高到低依次为：低学历高经验型、低学历低经验型、中等学历低经验型、高学历高经验型。这反映了四种类型职位在学历经验要求上的偏好差异与相应的薪酬差异。

角度二、角度三仅针对单个维度展开，在此不做聚类处理，分析结果将与可视化内容一同展现。

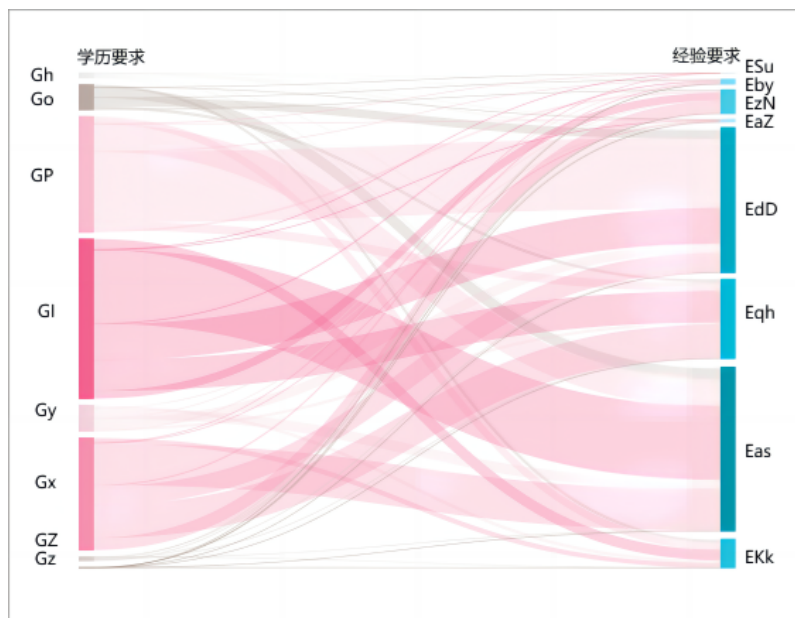
4.1.1 设计与案例呈现

为了更直观地展现职位间的差异度，本作品设计概览视图（A），如下所示：



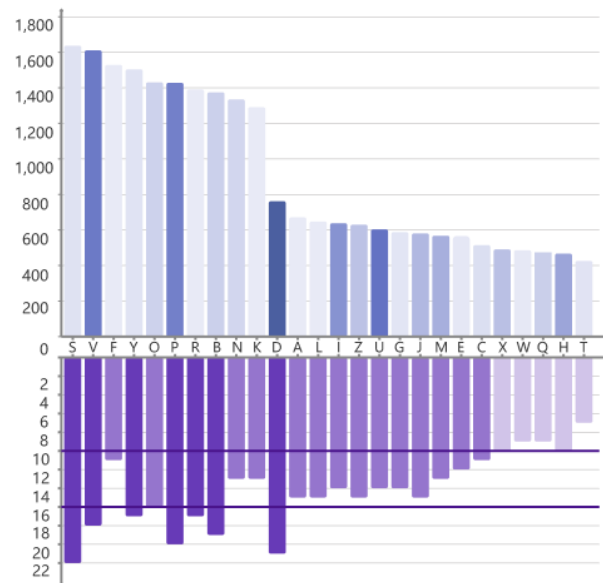
图表 4-2 视图 A 总览

从学历-经验桑基图（A1）中可分析得，大多数职位在招聘时不要求求职者具有较多经验，经验要求基本集中于五年以下（EdD 及以下）。但在低经验中，高学历人才（如 GP，本科）被要求有相对较高的经验（EdD，3-4 年），中低学历人才（如 Gx，中专）所受到的经验要求不高（Eas，1 年）。



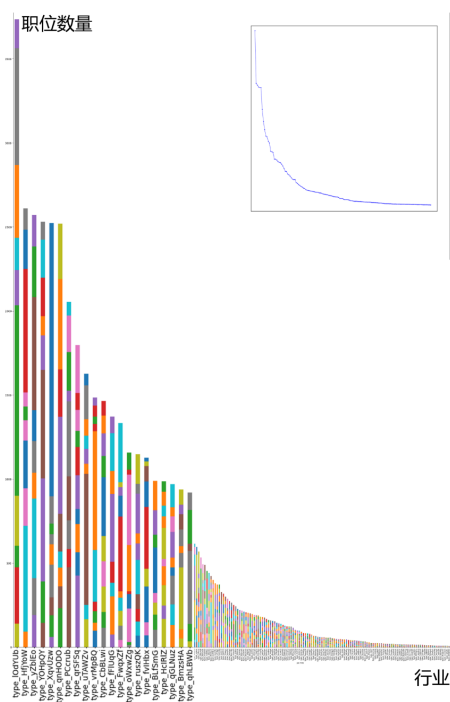
图表 4-3 学历-经验桑基图（A1）

由地域-职位城市数量双向条形图（A2）可得，职位在一级行政区划上大致分为具有差异的两类：在 S 至 K 行政区划上进行人数量大的招聘，和在 D 至 T 行政区划上进行人数量少的招聘。其中二级行政区划数少但岗位数多的地域（F，K 等）被推测为发展态势较好，反之二级行政区划数多但岗位数少的地域（D）被推测为发展态势较差。



图表 4-4 地域-职位城市数量双向条形图（A2）

由行业-职位数堆叠图（A3）可得，行业的招聘数量呈现极为明显的长尾现象。其中，type_10dYUb 这一行业的招聘数量明显高于其他行业，type_ZcPYrP、type_IWPTYt 等行业位于尾部于头部中间，为较为平庸的行业。



图表 4-5 行业-职位数堆叠图（A3）

在 A3 中选择 type_ZcPYrP 行业后后，其下职位在 A4 中展现如下：

	筛选结果				数量
	学历	经验	地域	行业	
c7f9349f001ec07fe7					68
6d0becea13a18azM					67
76b78c5d7d8ff1f1be59					67
bba95db2adbbBV					54
bfbcb45b05c5a442521					54
0cd2cb3d84ae09GC					54
c31759a7c06bf3c3b95					54
ee6a39780c633Dv					46
71f952e353628d7e2f4					46
0d40038b09ac2Tn					44
8d32b4d89b3145829c					44
d3e57304c480cdQR					41
3e1f8a4ef4af0439a98					41
a788685729e11IS					40
c3c05664164dc4da9b					40
638c821b59d1adFq					38
621cee7b2edf5d49210					38
bae18c12b5003zB					38
82848a0a5d5f3c7837					38
c34a7555aaf10vB					36
5b43e275bb223ae7be					36
e9808b618a3467ws					36
70cb83731f386d9b87					32
3391b78461bc9emp					32
c9a30a5443d04270fb					32
0f49d437c3376bAy					31
f1700927fd979f5ae074					31
de8f9352d96fWv					29
ca4d5884938fb754bf					29
2d1b45e26822d4jc					28
eb4523711beb76f635c					28
39145c8b742eatB					27
600a4de652c1b94b12					27
2b5b1a72cdf8baqf					26
244cd198e7293fbabd					26
28fad9934dcf89eh					26
952e95cafbb8ae0277					26
05e829607b4c3asE					24
2cfa3ce5aca7d2affe12					24
39f1e022639eMv					24

图表 4-6 职位筛选结果

可见该行业偏好高学历低经验型职位，且这些职位较多分布于经济发展水平较高的地域。

4.1.2 小结与建议

1. 小结

依据职位在上述三个角度上的差异性，本作品分别将职位划分为以下类型：

角度一：低学历高经验型、低学历低经验型、中等学历低经验型、高学历高经验型

角度二：大规模招聘型、小规模招聘型

角度三：大体量行业型、小体量行业型。

2. 建议

对于企业而言，**企业可根据职位性质合理设置学历和经验要求**：对实操性强的岗位，可以降低学历要求，提高经验要求，并当提高薪酬以吸引和留住优秀人才；对于创新性和研究性岗位则更注重学历背景，提供更多的职业发展和培训机会。同时，**企业也可依据招聘规模选择相应发展形势的地区**。

对于求职者而言，**低学历求职者应注重积累实际工作经验**，**高学历求职者应综合考虑地区经济发展水平**，以得到满意的薪资。

4.2 问题二

结合职位的关键特征挖掘，从多角度设计并展示职位画像，如关键技能、偏好

城市、薪酬待遇等

4.2.1 分析与设计

1. 关键技能分析

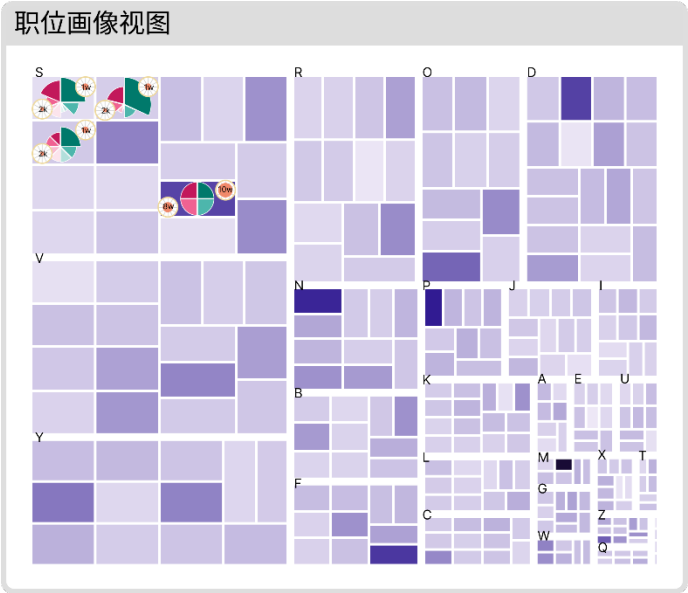
本作品将关键技能定义为通过理论学习获得的相关技能和通过实际操作和经验积累获得的相关技能。由于数据均以虚拟代码代称，关键技能无法通过数据内容准确分析而得，本作品借助对经验要求和学历要求的语义分析结果（表 2-2，表 2-3），将关键技能划分为四个类型：**理论实践双高型技能、重理论轻实践型技能、重实践轻理论型技能和理论实践双低型技能**。基于上述技能类型划分，本作品从职位所在行业维度、学历要求和经验要求三个维度对关键技能进行间接呈现。

2. 视图设计

基于上述定义，本作品从直接呈现和间接呈现两个角度设计并展示职位画像，分别为职位画像视图（B）和关键技能视图（C）。其中，B 包含职位偏好城市、薪酬待遇、经验要求、学历要求等与数据内容相对应的维度，C 用于展现和分析职位的关键技能。

4.2.2 案例呈现

使用者在 A1 视图中分别选定 GP 和 EdD 为学历要求和经验要求的筛选条件后，接着选择在 A4 中排位为第一的职位“bfbc45b05c5a4425210cd2cb3d84ae09GC”作为 B 和 C 的展示对象。B 中视图如下所示：

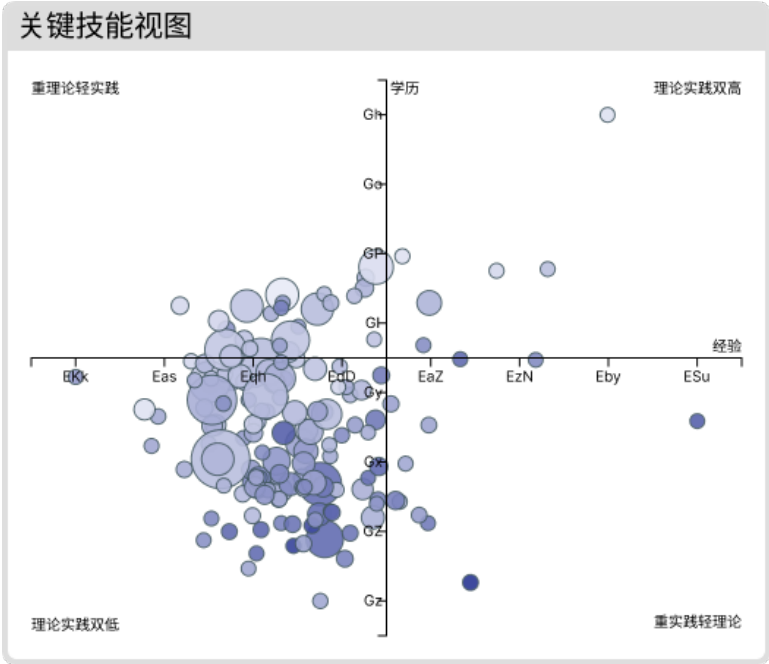


图表 4-7 职位画像视图 B

可见该职位偏好 S、V、Y 等一级行政区划以及其下城市，其在这类一级行政区划中招聘人数最多。鼠标悬浮于各矩形可显示相应二级行政区划名称，使用者发现，在 B 这一

级行政区划中，该职位在 S759 这一二级行政区划中具有较高薪酬水平。以 S759 为例，该职位在该城市偏好学历要求为 GI（大专）和 Gy（中专），可推断得该职位在该地域偏好技术型人才；该职位在该城市偏好经验要求为 Ekk（无需经验）和 Eas（1 年），可见该职位得技术门槛相对较低。

C 视图结果如下所示：



图表 4-8 关键技能视图 C

使用者发现，该职位所在行业对技能的需求主要分布于“理论事件双低”类型，少数分布于“重理论轻实践”类型。可见该职位对于求职者关键技能的要求并不高。点击最大的几个代表行业的圆形可得，该职位更偏向于与 type_10dYUb、type_qrSFSq、type_yZblE o 等行业相关的关键技能。

4.2.3 小结与建议

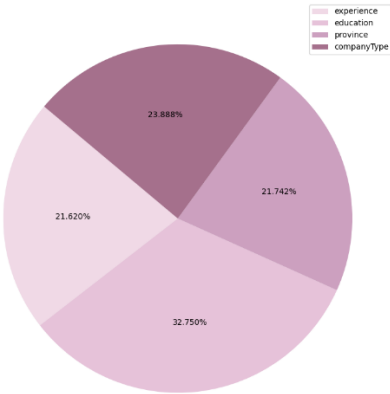
4.3 问题三

不同的职位具有不同的薪酬待遇模式。请对薪酬待遇与职位、行业、地域等之间的潜在关系进行建模，识别薪酬模式和潜在的薪酬差异，利用图表的形式呈现结果并简要分析。

4.3.1 分析与设计

1. 潜在关系建模与分析

本作品从行业、一级行政区划、学历要求和经验要求四个维度分析职位的薪酬待遇。为了确定上述维度对于薪酬待遇的影响程度，本作品使用决策树模型进行分析。分析结果表明，上述维度对薪酬待遇影响程度由高到低依次为：**学历要求，行业，经验要求，一级行政区划**。



图表 4-9 决策树结果

2. 相关定义

本作品将薪酬模式定义为企业发放薪酬的形式，包括月薪制、日薪制、时薪制等，以及月薪中的不同绩效形式（例如，除固定薪酬外还可能发放奖金）。前者（如日薪制和时薪制）在灵活用工中较为常见，而后者（如月薪加绩效奖金）在固定用工中较为普遍。

这种分类反映了职位间潜在的薪酬差异。例如，有些职位尽管月薪较低，但因额外绩效奖金而整体薪酬较高；而另一些职位尽管月薪较高，但因缺乏额外绩效奖金，其整体薪酬可能并不具有竞争力。

3. 视图设计

由图表 4-9 可得，四个维度对于薪酬待遇的影响程度十分相近，仅从单个维度分析薪酬待遇较为单一，无法全面地反映薪酬待遇与其他指标间的潜在关系。故本作品设计双向坐标轴图表，结合散点图与折线图，直观地同时展示薪酬待遇在多个维度上的分布情况。

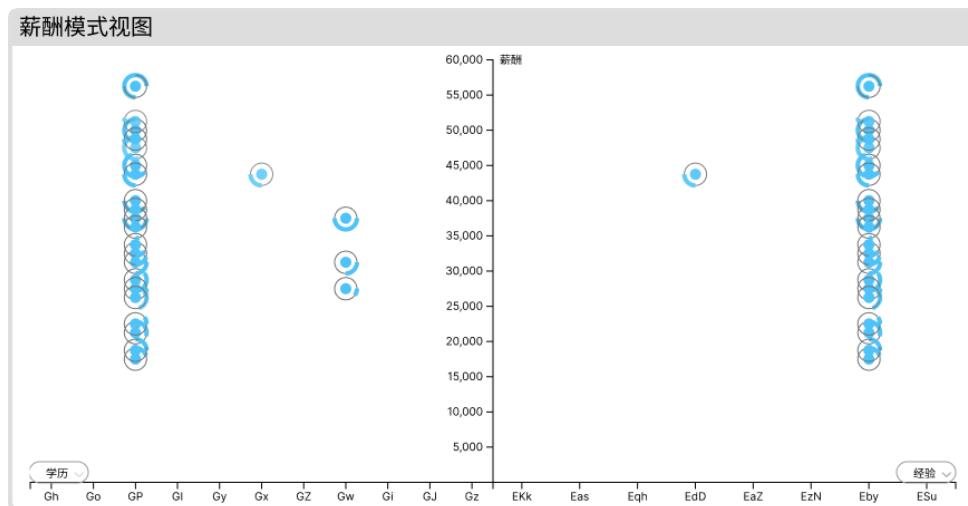
同时，为了直观呈现薪酬模式与潜在薪酬差异，本作品对散点图进行改进，将薪酬模式、薪酬水平、绩效奖金等内容直观地呈现在各散点上。

4.3.2 案例呈现

1. 灵活用工

使用者选择“灵活用工”模式后，在 A 中不进行任何筛选。A4 中展现所有灵活用工的职位对比，使用者选择其中排位靠前的“15d3753085d386f193e30f663a2137a2iY”作为 D 的展现对象。

使用者首先在横轴上勾选“学历”和“经验”两维度。



图表 4-10 灵活用工薪酬模式

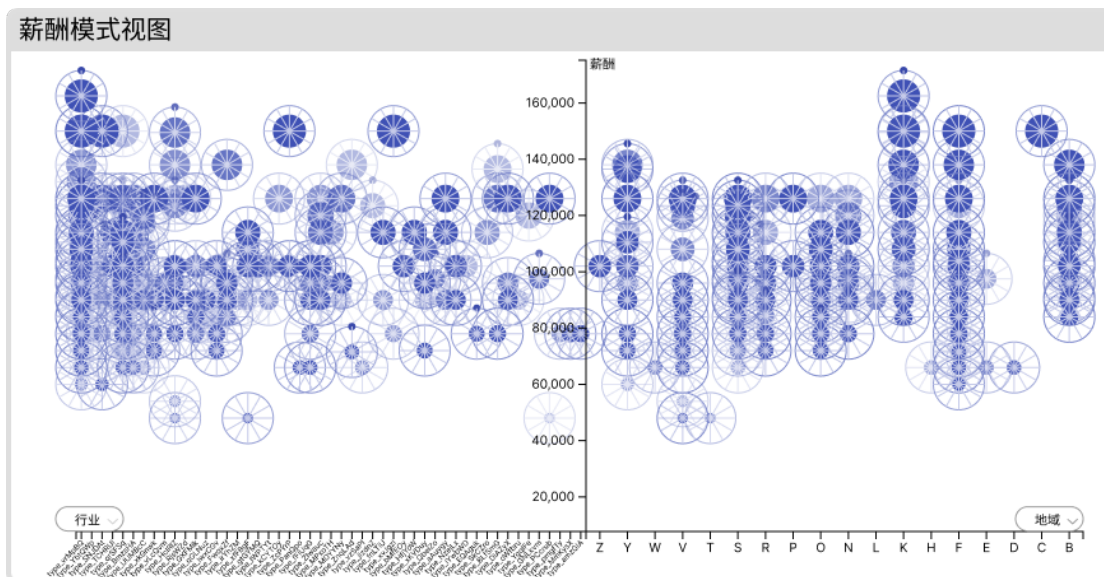
从散点细节上看，该职位**薪酬模式全部为日结模式**，且薪资水平浮动极大，无明显规律。从宏观上看，该职位在具有 8-9 年工作经验的本科人才上展现了极大的偏好。

使用者在 A4 视图选择若干其他灵活用工职位，发现了同样的规律：灵活用工招聘信息中，有极大部分招聘信息面向**日结、高学历、工作经验高**的人才，推断此类招聘信息可能面向**摄影师、制片等影视类工作者**。

2. 固定用工

使用者选择“固定用工”模式后，首先在 A 视图中选定 Eas（一年工作经验）和 Gx（中专）为职位筛选条件，并选择在 A4 视图中招聘数最多的“6be068c4e74275ca8ed9975805e2f763Em”作为展现对象。

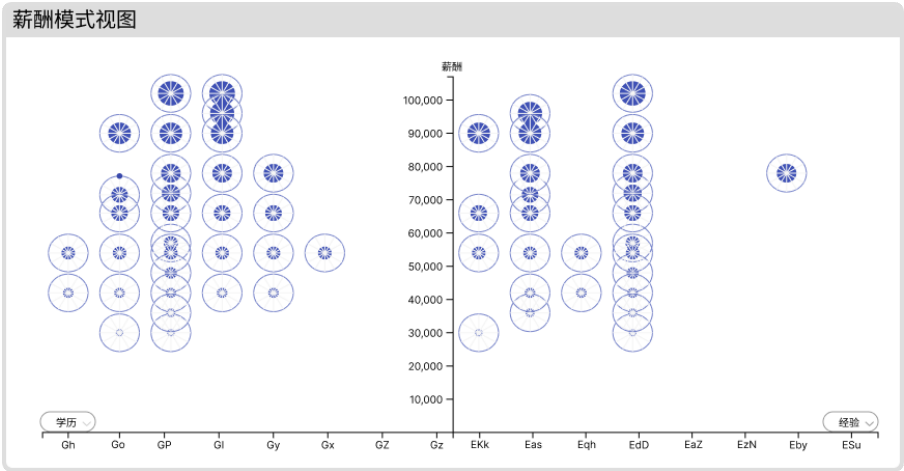
使用者勾选“行业”和“地域”两横轴维度，D 视图展现如下：



图表 4-11 固定用工薪酬模式-1

从散点细节上看，该职位薪酬模式多数为不具奖金的月薪模式，少量招聘信息为求职者给予一个月的奖金；从宏观上看，该职位偏好 type_qrSFSq、type_vrMpBQ 两行业，在地域上无显著偏好，但在 K 和 F 两个一级行政区划上具有较高薪酬水平。职位整体薪资水平浮动较大，为每年 5-16 万。

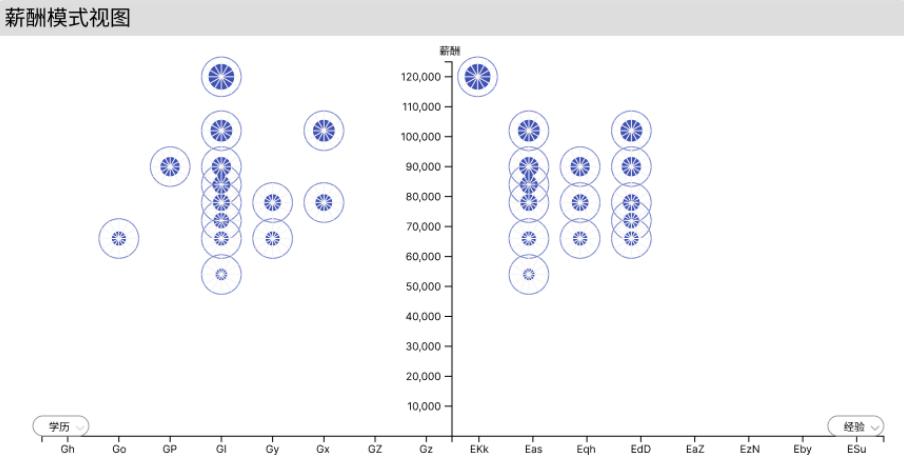
随后使用者在 A 视图中选择“D”行政区划作为筛选条件，选择 A4 视图中招聘数第一的职位“bfbc45b05c5a4425210cd2cb3d84ae09GC”进行展现。在横轴上勾选“学历要求”和“经验要求”后，得到 D 视图如下所示：



图表 4-12 固定用工薪酬模式-2

从散点细节上看，该职位薪酬模式大多数月薪制，仅有条招聘信息的薪酬具有一个月奖金；从宏观上看，该职位偏好有 3-4 年工作经验（EdD）的本科学历（GP）人员。职位整体薪资水平为每年 3-10 万。

使用者将分析职位切换至 A4 中招聘数第十的职位“534952ab1cf024773caddc874099da11qn”，横轴展示维度不做变化，D 视图如下所示：



图表 4-13 固定用工薪酬模式-3

从细节上看，该职位**薪酬模式均为月薪**，不存在奖金模式；从宏观上看，该职位**偏好具有 1 年工作经验（Eas）的大专学历（GI）人员**。职位整体薪资水平为每年 5-12 万。

使用者发现，该职位的整体薪资明显高于上一个职位，且大专学历（GI）和中专学历（Gx）工作者的薪资上限高于本科学历（GP）工作者。

4.3.3 小结与建议

灵活用工与固定用工呈现出薪酬模式上的差异：灵活用工主要以日薪制为主，薪资水平浮动较大，常见于高学历、高经验要求的职位，如摄影师、制片等影视类工作者。固定用工以月薪制为主，部分职位提供绩效奖金。薪资水平相对稳定，但奖金模式较少见，更偏好低学历低经验和高学历高经验的人才。

职业型人才可能比理论型人才获得更多薪酬，大专学历、中专学历等人才的薪资上下限均高于本科学历人才。但招聘时对职业型教育人才的经验要求低于理论型教育人才，也许意味着**理论型人才具有更多跳槽的机会，有更大的晋升空间**。

4.4 问题四

挖掘和总结地域招聘活动画像，从多角度设计并展示招聘活动的地域特征，如偏好职位和行业类别等，并识别具有相似招聘特征的地域。

4.4.1 分析与设计

1. 相似地域定义与识别

本作品将具有相似招聘信息的地域定义为在地域经济发展水平、平均薪酬水平、偏好职位和偏好行业四个维度上相似的地域。具有相似招聘特征的地域通过向量化计算来评估和识别。在对数据进行标准化处理后，各地域在四个维度上的特征值被转换为向量，并采用余弦相似度计算地域之间的相似指数。

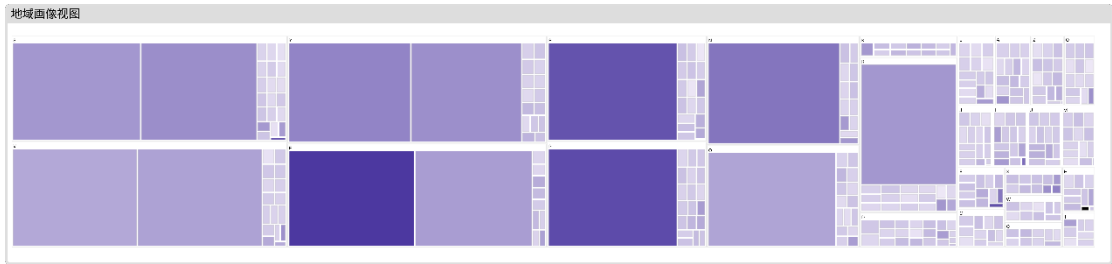
2. 视图设计

在地域画像视图（G）上，依据对一、二级行政区划的划分，本作品将地域招聘活动分为两级展现：一级视图展现各一级行政区划的招聘数量和平均薪酬水平，交互后的二级视图展现各二级行政区划招聘活动的偏好职位、偏好行业、最高与最低薪资水平。此外，相似地域视图（H）中将展现与所选中的一级或二级行政区划最为相似的一级或二级行政区划。

4.4.2 案例呈现

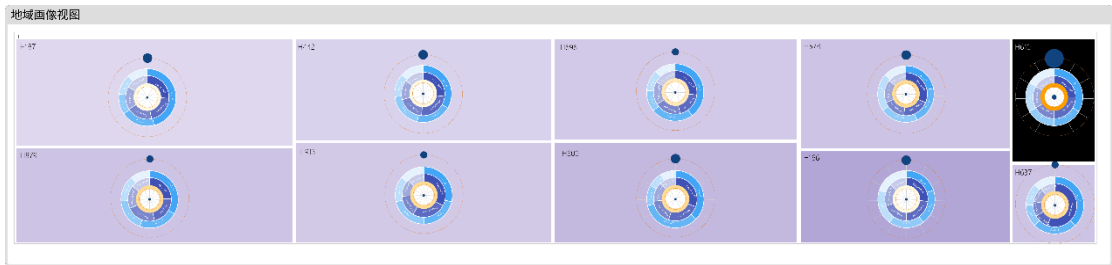
使用者首先浏览 G 视图的一级界面，注意到 S、P、Y、F、B、V、N、O 等一级行政区划具有较多的招聘信息数，这主要是因为这些地域下有 1-2 个招聘数量极高的二级行政区

划。使用者进一步分析发现，F、P、B 等地区招聘信息数最多的二级行政区划具有较好的薪酬待遇。



图表 4-14 地域画像一级视图

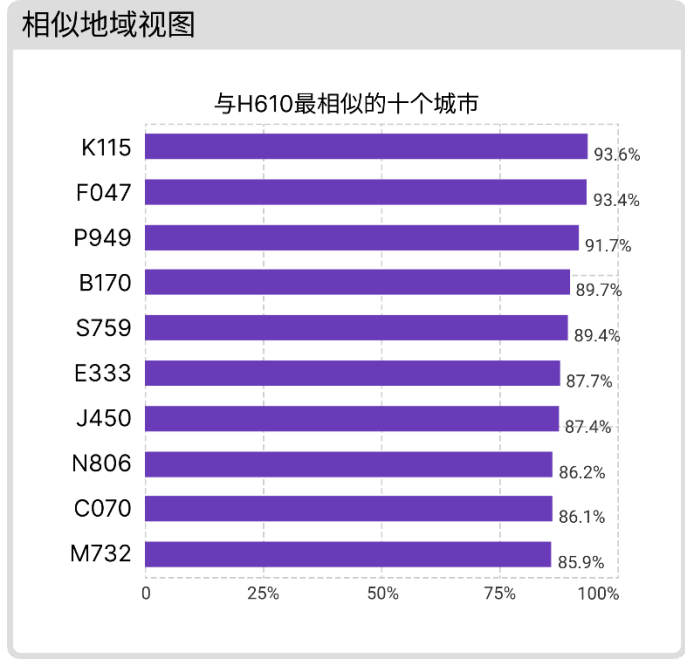
随后，使用者发现 H 这一一级行政区划中有一个二级行政区划的薪酬水平极高。使用者通过点击交互进入 H 的二级界面，如下图所示：



图表 4-15 地域画像二级视图

使用者发现 H610 这一二级行政区划的薪酬水平极高，其薪酬下限和薪酬上限均显著高于其他地区的水平。此外，使用者还发现“H”一级行政区划**偏好 type_10dYUb、type_qrSFSq 两行业和 bfb45b05c5a4425210cd2cb3d84ae09GC、d5483194d58df66a5fc42c069ee4fe2axN 两职位。**

使用者点击 H610 后，右侧的 H 视图呈现出与 H610 最相似的 10 个二级行政区划及其相似程度，如下图所示：



图表 4-16 相似地域视图

4.4.3 小结与建议

总体而言，通过一级界面的浏览，使用者能够快速识别出招聘信息集中的一级行政区划。这些地域的就业机会较多，还伴随着较好的薪酬待遇，招聘活动较为活跃，就业机会较多。对于需要高薪人才的企业，可优先考虑在这些地域开展招聘活动；求职者也可以优先考虑在这些地域求职，以增加求职成功率。

在对二级界面的深入分析中，使用者能够对特定地域的职位偏好、行业偏好及薪资上下限进行细致分析，从而挖掘出两类重要的地域：1. 招聘信息数量较少但薪资水平较高的**潜力发展地域**；2. 招聘信息数量较多但薪资水平较低**的高需求地域**。

对于企业而言，在潜力发展型地域，可根据当地的职位和行业偏好，精准定位招聘需求，有针对性地进行招聘宣传和岗位设置；在高需求地域，可以通过优化薪酬结构和提供额外福利来提升吸引力。对于希望获得高薪酬和快速职位发展的求职者而言，可优先考虑潜力发展地域，尤其是具有高学历、多经验的求职者；对于希望尽快找到工作的求职者，则可优先考虑招聘信息数较多的高需求地域。

4.5 问题五

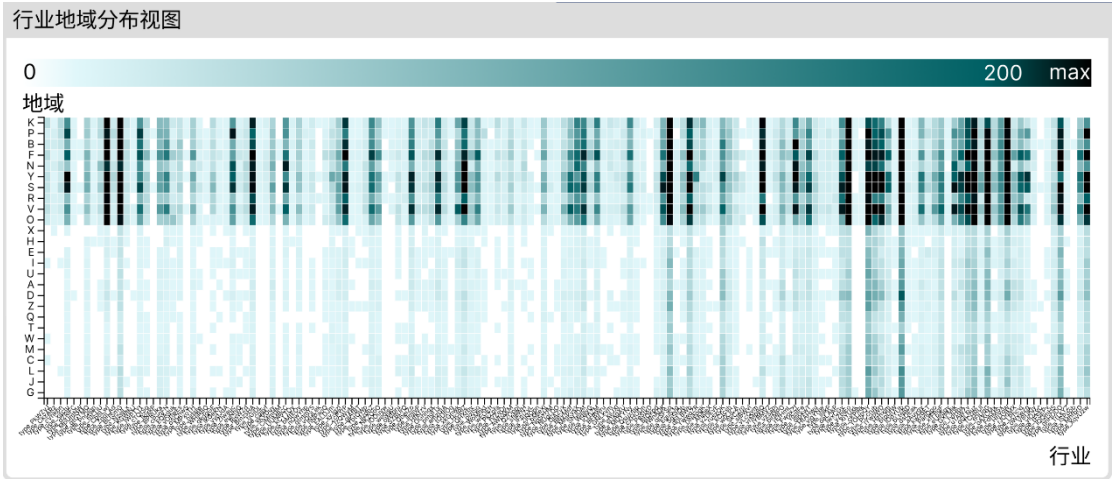
结合上述分析结果，是否可以总结行业发展动态，并发现急需人才的新兴职位，并简要说明理由。

4.5.1 定义分析

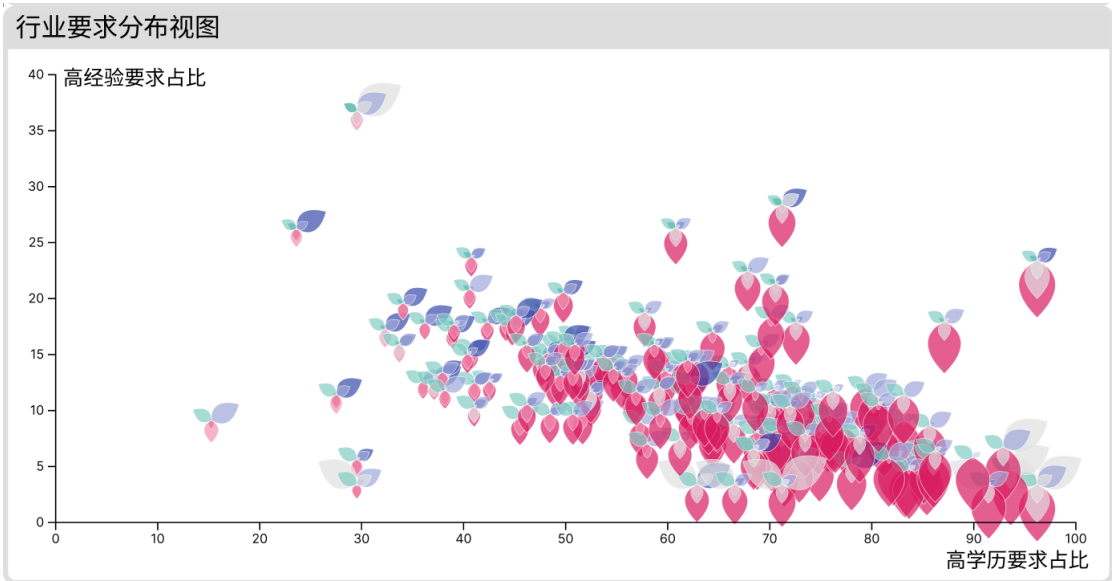
本作品将行业发展动态分为**增长型**与**衰退型**。其中，增长型行业为在经济发展状况良好的行政区划内，高学历和高经验要求的职位占比较大，且高薪职位比例较高的行业，具有向上的发展趋势与较好的前景；衰退型行业则与之相反。

此外，本作品将急需人才的新兴职位的特征定义为：**对学历和经验要求较低，但薪酬水平处于较高层次**，且显著高于具有相同学历和经验要求的职位。这类职位的出现往往反映了行业内特定技能或岗位的紧缺，表现出强烈的市场需求和较高的竞争力。其较低的入职门槛和相对优渥的薪酬待遇，吸引大量求职者，并促进相关行业的快速发展。

4.5.2 总结行业发展动态



图表 4-17 行业地域分布视图



图表 4-18 行业要求分布视图

依据行业地域分布视图（E），本作品依据定义认为**在靠上区域颜色更深的列对应增长型行业**，在靠下区域颜色深的为衰退型行业。依据行业要求分布视图（F），本作品依据定义认为**靠右上区域且粉色花瓣更大更深的图形对应增长型行业**，靠左下区域且粉色花瓣小而浅的图形对应行业为衰退型行业。但由于经验要求多集中于低经验类别内，故只考虑相对位置。

综合考虑上述内容后，本作品认为**能够总结行业发展动态**：type_l0dyUb、type_qrSFSq、type_BLfSmG 等为增长型行业，type_Vl0ZvC、type_XiYUxk 等为衰退型行业。

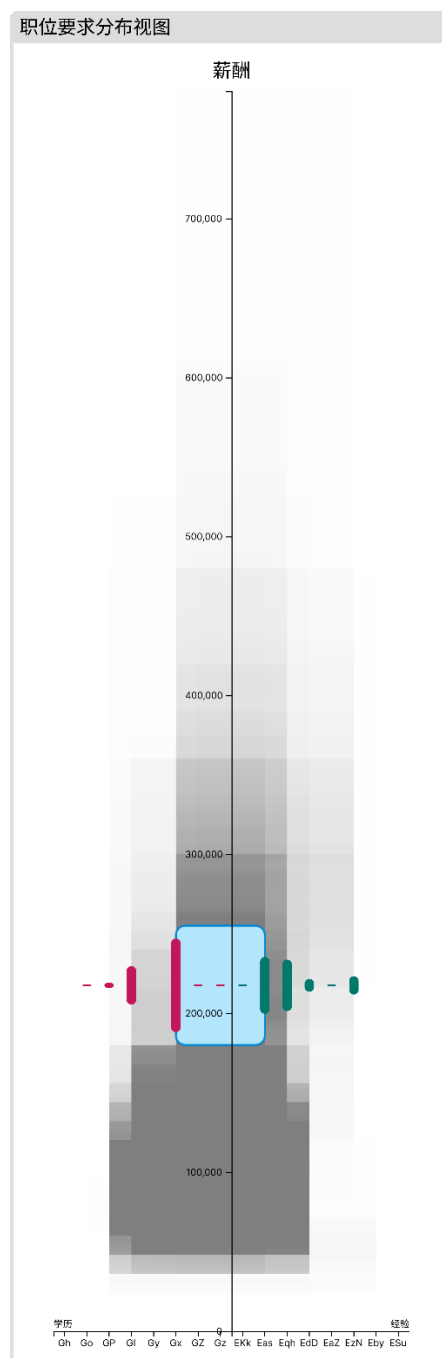
此外，本作品还总结得以下可能的行业发展规律：

1. 技术型人才整体薪资水平优于理论型人才
2. 固定用工招聘时倾向于低经验要求
3. 灵活用工中经验要求对薪资的影响比学历要求大
4. 中专、大专型人才容易受到高经验要求，本科及以上学历型人才更容易受到低经验要求

4.5.3 急需人才的新兴职位

结合 G 视图中“H”这一发展态势较好的一级行政区划，本研究在 A 视图中进行类别为“学历要求为 Gx（中专）+经验要求为 Eas（1 年）+地域为 H”的筛选，并在 A4 视图中锁定一个可能的急需人才的新兴职位：d5483194d58df66a5fc42c069ee4fe2axN。

选中这一职位后，在职位要求分布视图（I）中观察该行业的分布位置，如下图所示：



图表 4-19 职位要求分布视图

可以发现，该职位薪资水平处于整体薪资水平的前列，但经验要求与学历要求大多分布于中等学历和低经验类别。该职位符合本作品对于新兴职位的定义，可以判断它有较大可能为急需人才的新兴职位。

5 讨论与总结

5.1 讨论

5.1.1 经验教训

本研究的一个关键经验是认识到在招聘数据可视化分析中，过度依赖宏观统计结果可能会掩盖数据中蕴含的重要细节信息。我们发现，深入探究数据的分布特征、异常值和细分模式往往能够揭示更有价值的洞察。例如，在分析薪资数据时，不仅要关注平均值和中位数，还应该可视化整个分布曲线，识别多峰分布或长尾现象。这种细致的观察使我们能够发现不同职位类别或地区间的微妙差异，从而得出更加准确和有针对性的结论。

这一经验启示我们，在设计可视化系统时，应该**为用户提供多尺度的数据探索能力**，既能够概览整体趋势，又能深入查看局部细节。同时，我们意识到有必要开发更加灵活的可视化工具，使用户能够轻松地**在不同粒度级别之间切换**，从而全面把握数据的复杂性。这种方法不仅能够提高分析的准确性，还能帮助识别潜在的市场机会或风险，为招聘决策提供更加可靠的依据。

5.1.2 不足

本研究在多维数据可视化方面存在一定局限性。现有的可视化方案在同时展示和对比多个岗位的学历、经验、薪资和行业特征等多维属性时效果欠佳，难以充分呈现各维度间的复杂关联。这些局限性可能影响用户对招聘市场复杂性的全面理解，从而制约了数据分析的深度和广度。

5.2 总结

本研究旨在设计和实现一个基于机器学习和统计分析的招聘广告大数据可视化系统，以深入分析就业市场的多维度特征和动态趋势。

通过处理和分析包含 430,664 条招聘记录的大规模数据集，本研究在多个方面取得了显著成果。首先，我们成功构建了一个集数据处理、多维分析和可视化展示于一体的分析平台，为就业市场参与者提供了全面的决策支持工具。其次，通过创新的可视化技术，我们揭示了职位差异、薪酬结构、地域特征等多个维度的复杂关系，为理解就业市场动态提供了新的视角。特别是，本研究强调了深入探究数据细节和分布特征的重要性，发现这种细致观察能够揭示传统宏观统计方法可能忽略的重要洞察。

此外，研究中开发的交互式可视化方法增强了用户探索复杂数据关系的能力，为数据驱动的招聘决策提供了有力支持。总的来说，本研究不仅为招聘市场分析提供了创新的方法和工具，也为大数据可视化在人力资源管理领域的应用做出了重要贡献，为政策制定者、企业和求职者提供了宝贵的决策参考。