数据分析与数据挖掘岗位职责与任职要求等差异化分析

1. 背景：

随着互联网及移动互联网的高速发展，我们在享受其便利的同时，产生了大量的用户数据。这些数据被收集在企业的数据库中，能否利用这些数据提升公司利润是每个公司心心念念的事情。数据分析师以及数据挖掘工程师就是应此需求产生的岗位。但其实，这里也说的并不准确。数据分析师已经存在很久了，著名的啤酒与尿布的分析结论便是来源于传统的零售商－沃尔玛。但是，当时的数据并没有和用户产生如此紧密的联系。现在的数据分析师和数据挖掘工程师更贴近于用户，毕竟用户才是企业的金主。数据能做的事情简直太多了，比如最近Facebook数据泄漏以致影响美国选民在竞选中的投票。数据如此重要，那懂数据的人一定很值钱。为什么会对比这两个工作的差异，因为目前有很多应届毕业生和想转行的小伙伴们对两者的工作内容很模糊，希望从数据出发，希望帮助大家对两种岗位有更清晰的认识。

1. 数据爬取：

此次数据来源于互联网招聘平台拉勾网。共爬取数据分析师岗位信息450条，数据挖掘工程师岗位信息445条。数据爬取分两部分完成：一是通过访问<https://www.lagou.com/jobs/positionAjax.json>文件得到每个招聘需求的基本信息, 具体包括：城市，公司名，公司规模，所在城区，融资阶段，岗位名称，学历要求，工作年限要求，薪资等；二是根据岗位的唯一标识positionID，访问<https://www.lagou.com/jobs/positionID.html> 文件得到每个招聘需求的详细信息，具体包括：岗位职责和任职要求。

1. 数据存储：

将上述爬取的数据存储为csv文件。

1. 数据读取：

将csv文件用Pandas读入。

1. 数据评估与清洗：

* 由于在爬取数据挖掘相关岗位信息时，出现了爬取中断，所以一次完整的爬取过程被强制拆分成了多次，导致产生了部分重复信息，需要利用drop\_duplicates函数将重复数据删除。
* 由于岗位名称的定义不一致，添加kd列来统一岗位名称。
* 将工作描述提取为两部分，岗位职责和任职要求。
* 由于薪资给出的格式包含上下限，如[10k-20k]，则新建三列，分别存储下限，均值和上限。

1. 数据分析

该部分具体分为两部分，一是用Tableau做基本的可视化分析；二是对岗位职责和任职要求绘制词云。

* Tableau做基本的可视化分析

1. 薪资差异对比

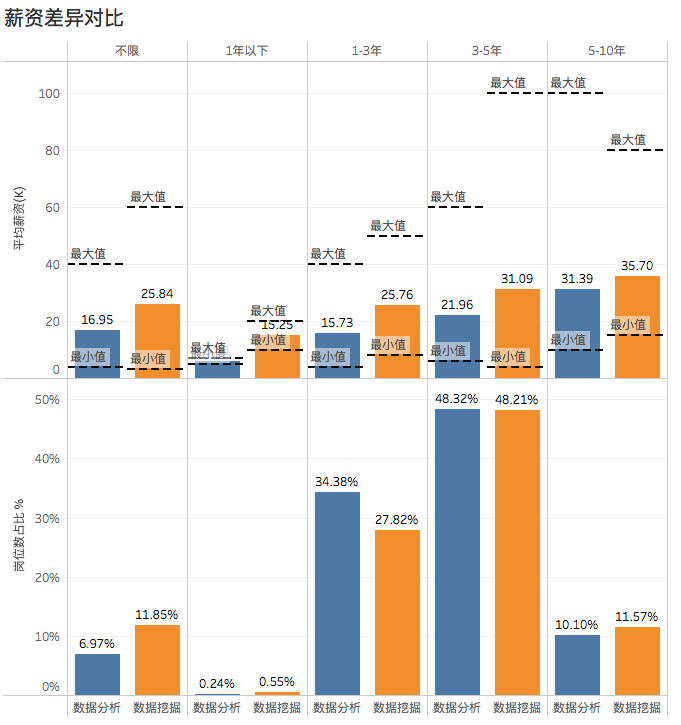


图 1

从图1中可看出：

* 数据挖掘的平均薪资高于数据分析。
* 工作年限越高，平均薪资越高。
* 工作年限在5-10年的数据分析师的薪资上限高于数据挖掘工程师。同时平均薪资的差距降低。由于样本有限，只能说明经验丰富的数据分析师的待遇非常丰富。
* 工作年限在3-5年以及1-3年的岗位需求最多。
* 工作经验不限的平均薪资并不低，这个大概率是招聘需求填写不明确。

1. 学位要求差异：

从图2中可以看出：

* 大本岗位需求明确要求本科以上
* 数据分析师少部分学位需求降低到大专
* 数据挖掘工程师部分学位需求提高到博士
* 数据挖掘工程师硕士学位要求占比较多

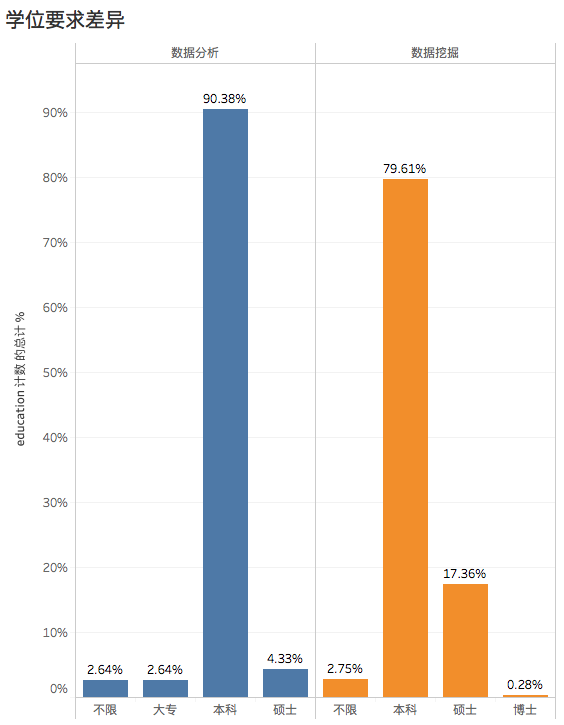


图 2

1. 公司处于各个阶段的岗位需求差异：

从图3中可以看出：

* 较大规模的公司对数据挖掘工程师的需求相对较多。这较合理，毕竟数据量较大才需要挖掘。

1. 城区分布差异：

从图4中可看出：

* 大部分的招聘公司集中在海淀和朝阳。这较合理，海淀和朝阳集中了大部分科技公司。

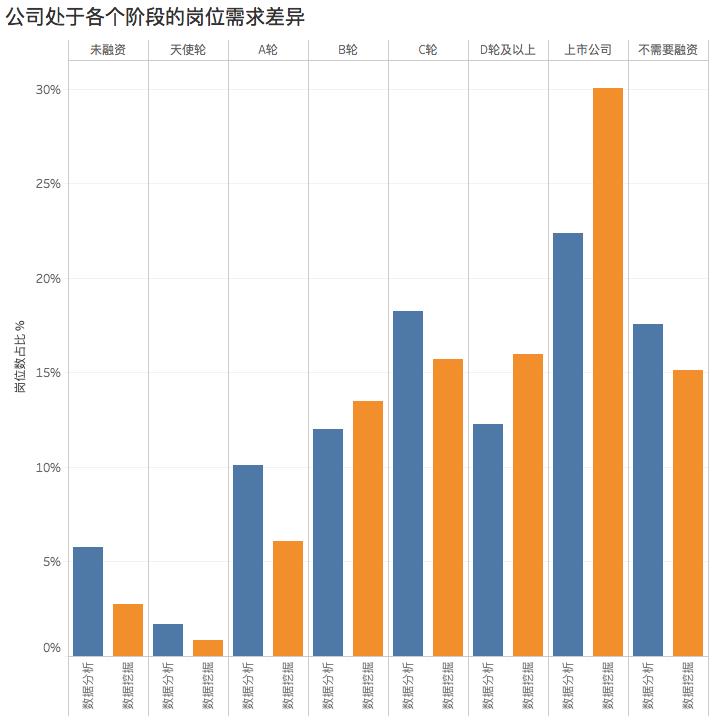


图 3

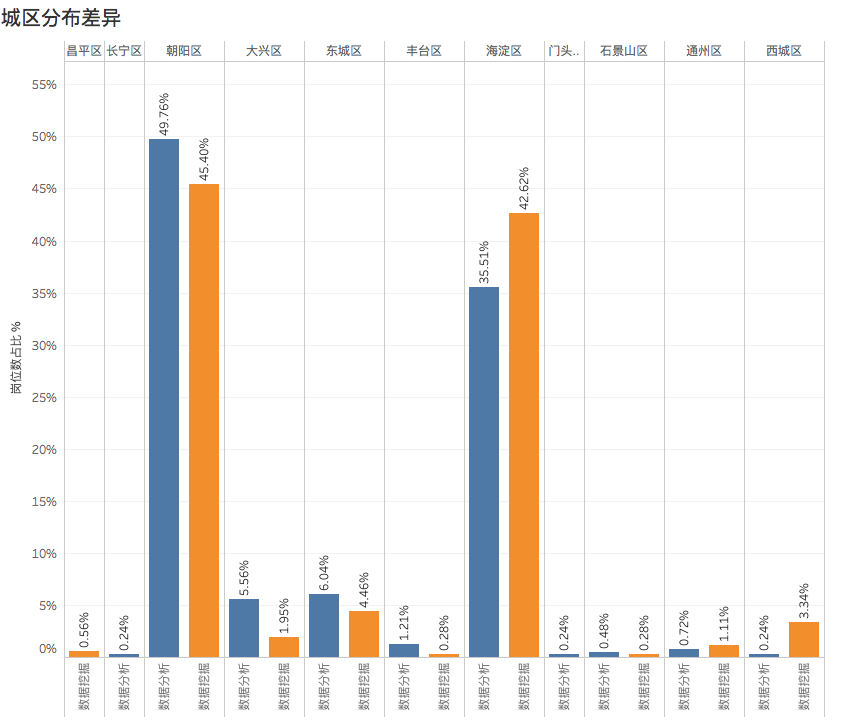


图 4

* 岗位职责和任职要求绘制词云

1. 数据分析师与数据挖掘工程师词云

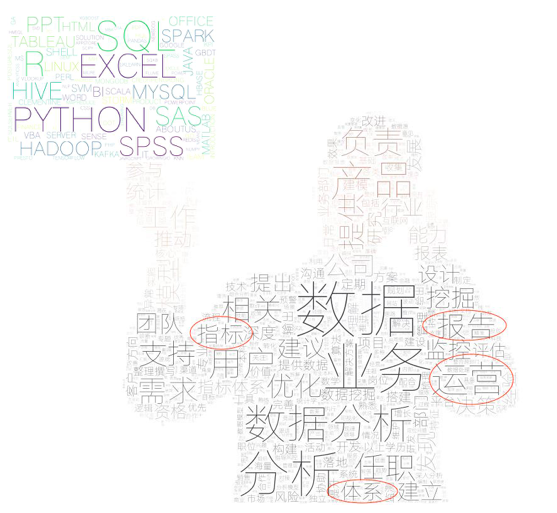


图 5

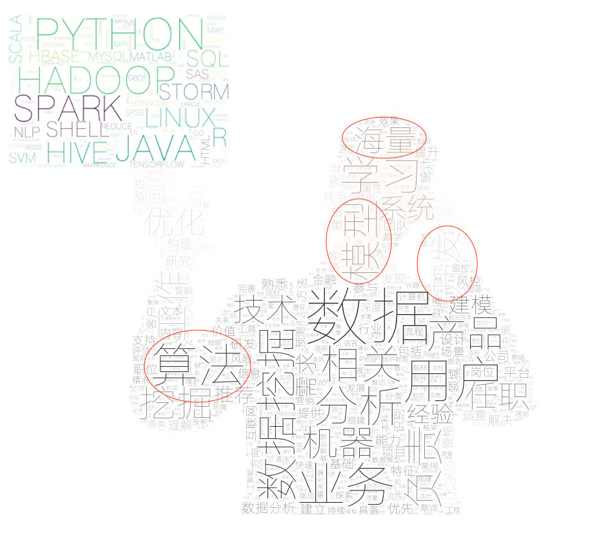


图 6

* 从图5，6中可以看出，目前的招聘需求中，数据分析师偏向于运营和产品；数据挖掘工程师偏向于开发。我用红圈标出了一些有趣的差异词，我们可以看到，数据分析师会参与一些指标的建立并输出一些报告（比如目前的用户增长类数据分析师，通过对数据的分析，提高用户转化率等）；而数据挖掘工程师更偏重于从海量数据中利用算法建立模型进行挖掘（比如目前的推荐系统工程师）
* 数据分析师的技能牌中SQL，Excel，Python和R出现频次最高，尤其是SQL。数据挖掘工程师的技能牌中Python，Java，Hadoop，Spark出现频次较高，尤其是Python。从技能牌情况可看出，数据挖掘工程师对开发能力要求较高。
* Python是数据分析以及数据挖掘工程师共有的核心技能。