

数据分析师的体检报告

作为一个数据分析师从业者，数据分析占据了我大部分时间。但是当拥有两年的从业经验后，我逐渐有些迷茫。数据分析师是会用 EXCEL、Python、SQL 处理数据？还是负责每月甚至每周各式各样的报表？其实这些问题都可以归结为数据分析师日常工作是怎样？既然有了这个疑问，作为一个数据分析从业者，我就当一次数据医生，替数据分析师这份职业做一份体检报告。

1. 数据收集

有了明确问题，那就必须围绕这个问题构想需要什么样的数据。数据分析师日常工作资料可以来自 1. 招聘网站 2. 数据分析从业者 3. 职业论坛。但后两者的数据显然是来源于供给方的，资料相对零散，不容易进行统计。故我从招聘网站下手。

招聘网站好处：1. 数据集中 2. 数据来自需求方，具有客观性。知道了数据的来源，那就必须为数据定下指标。

指标：重点词的出现占比，这个指标能表明数据分析师日常工作是什么？需求方对哪方面职责的需求比较大，通过指标可以回答数据分析师日常工作是做什么。

操作：通过 Python 从 BOSS 直聘上爬取了 296 条数据分析师的数据。（如图 1、图 2、图 3）因为职位信息与岗位职责分别处于不同的网页。故爬取流程

为：职位信息→获取岗位职责网址→爬取岗位职责。

area	title	company	money	exp	company	company	web
重庆 渝北	【数据分析】	马上消费金融	20-35K	13-5	10年本科	互联网金融	员工旅游, https://www.zhipin.com/job_detail/0509178d18ece2c01n162Nm5GVtR.html
北京 大兴	【数据分析】	京东世纪	15-25K	15-1	3年本科	互联网	年终奖, https://www.zhipin.com/job_detail/8bf1027116a1ebee1n1409y1GF8Q.html
北京 朝阳	【数据分析】	BOSS直聘	15-30K	16-1	1年以内	人力资源	加班补助, https://www.zhipin.com/job_detail/cf37d9f3fbb99d841XV_3Nu0GFJQ.html
北京 海淀	【数据分析】	滴滴	20-40K	13-3	5年本科	移动互联网	五险一金, https://www.zhipin.com/job_detail/75834cb09b5745901n1629u9ElpR.html
北京 朝阳	【数据分析】	中信百信	15-30K	14-3	5年本科	银行	交通补助, https://www.zhipin.com/job_detail/cb26b94688c131610H1_2t24FY-.html
上海 闵行	【数据分析】	森马	13-25K	13-3	5年本科	生活服务	员工旅游, https://www.zhipin.com/job_detail/a21165a5080764c51nxy2tW6GVZW.htm
北京 朝阳	【数据分析】	美团	20-40K	16-3	5年本科	互联网	员工旅游, https://www.zhipin.com/job_detail/7087fe121ea7527b1nV82N-5EVZU.html
北京 海淀	【数据分析】	快手	20-40K	16-1	3年本科	社交网络	定期体检, https://www.zhipin.com/job_detail/b5fc6f8c3752b58a1nZ73NS5FV8U.html
北京 朝阳	【数据分析】	狮桥集团	15-30K	14-3	5年本科	投资/融资	定期体检, https://www.zhipin.com/job_detail/11a2f0f030ba48581nx42tq_E1pR.html
上海 杨浦	【数据分析】	OYO酒店	25-35K	3-5	5年本科	互联网	带薪年假, https://www.zhipin.com/job_detail/b9a63a24a678258703d_2NW_FFM-.htm
广州 海珠	【数据分析】	大参林	15-25K	13-5	10年本科	医疗健康	带薪年假, https://www.zhipin.com/job_detail/b39eb2a6935942511nx_2ti_E1FS.html
深圳 福田	【数据分析】	vivo	35-50K	5-10	10年本科	移动互联网	股票期权, https://www.zhipin.com/job_detail/4f6ab6c1fd13ee4e1n1y2dW7FJQJ.html
佛山 顺德	【数据分析】	美的	13-24K	15-3	5年本科	其他行业	五险一金, https://www.zhipin.com/job_detail/049d3a743544bdec1XR-3tS_FtQ.html
上海 长宁	【数据分析】	拼多多	30-60K	18-3	经验不限	移动互联网	五险一金, https://www.zhipin.com/job_detail/4e3585a674feb6d1XV439m-FIBW.html
北京 大兴	【数据分析】	京东集团	14-18K	16-1	在校/应届	电子商务	带薪年假, https://www.zhipin.com/job_detail/af19b20f149de5801n1909q5EVRr.html
杭州 拱墅	【数据分析】	海天塑机	12-18K	13-3	经验不限	机械设备	自助午餐, https://www.zhipin.com/job_detail/96af89b9ee6134061n153968EJS.html
北京 海淀	【数据分析】	北京字节	150-200元	4天/周	3个月	移动互联网	住房补贴, https://www.zhipin.com/job_detail/91469e099ec2f1891n1_29S8E1pS.html
福州 马尾	【数据分析】	国通星驿	6-10K	1-3	3年本科	互联网金融	五险一金, https://www.zhipin.com/job_detail/c9fb58a80a5867e13nR_2d--GFI-.html
大连 中山	【数据分析】	大连品管	3-5K	1-3	3年本科	在线教育	员工旅游, https://www.zhipin.com/job_detail/ec196dd591fa5de11nF8392_FIRO.html

图 1（职位信息）

job_description	company_name	full_company_name	web
岗位职责：1) 负责经营指标数据的	马上消费金融	马上消费金融股份有限	https://www.zhipin.com/job_detail/0509178d18ece2c01n162Nm5GVtR.html
【职位描述】1、对接项目数据提取	京东世纪贸易有限公司	北京京东世纪贸易有限	https://www.zhipin.com/job_detail/8bf1027116a1ebee1n1409y1GF8Q.html
工作职责1、用数据分析、数据挖掘	BOSS直聘	北京华品博睿网络技术	https://www.zhipin.com/job_detail/cf37d9f3fbb99d841XV_3Nu0GFJQ.html
职位描述1、负责用户相关运营和营	滴滴	北京滴滴无限科技发展	https://www.zhipin.com/job_detail/75834cb09b5745901n1629u9ElpR.html
职位职责：1、负责部门相关报表开	中信百信银行	中信百信银行股份有限	https://www.zhipin.com/job_detail/cb26b94688c131610H1_2t24FY-.html
岗位职责：1、基于市场环境、竞争	森马	浙江森马服饰股份有限	https://www.zhipin.com/job_detail/a21165a5080764c51nxy2tW6GVZW.html
岗位职责：1.深入了解美团外卖业务	美团	北京三快在线科技有限	https://www.zhipin.com/job_detail/7087fe121ea7527b1nV82N-5EVZU.html
职位描述1、基于深度思考和数据整	快手	北京快手科技有限公司	https://www.zhipin.com/job_detail/b5fc6f8c3752b58a1nZ73NS5FV8U.html
工作职责1、负责不同业务线的指标	狮桥集团	狮桥融资租赁（中国）	https://www.zhipin.com/job_detail/11a2f0f030ba48581nx42tq_E1pR.html
职位职责：1、负责商业化指标体系	OYO酒店	遨游酒店信息技术（深	https://www.zhipin.com/job_detail/b9a63a24a678258703d_2NW_FFM-.html
职位描述：1.负责APP整体数据运营	大参林	大参林医药集团股份有	https://www.zhipin.com/job_detail/b39eb2a6935942511nx_2ti_E1FS.html
岗位职责：1.搭建OS数据体系，能	vivo	维沃移动通信有限公司	https://www.zhipin.com/job_detail/4f6ab6c1fd13ee4e1n1y2dW7FJQJ.html
工作职责：构建运营业务的分析体系	美的	广东美的制冷设备有限	https://www.zhipin.com/job_detail/049d3a743544bdec1XR-3tS_FtQ.html
岗位职责【岗位职责】1.分析业务	拼多多	上海寻梦信息技术有限	https://www.zhipin.com/job_detail/4e3585a674feb6d1XV439m-FIBW.html
【职位描述】1、负责重要产品线的数据提取及	京东集团	北京京东世纪贸易有限	https://www.zhipin.com/job_detail/af19b20f149de5801n1909q5EVRr.html
工作内容：数据分析、算法开发、	海天塑机	海天塑机集团有限公司	https://www.zhipin.com/job_detail/96af89b9ee6134061n153968EJS.html
质量检测			

图 2（岗位职责）

```
正在爬取第296页的数据内容
职位描述
1、深入了解电商业务，能通过指导分析业务逻辑的风险与问题，为电商业务的风控分析提供高质量的数据支持；
2、通过数据分析和挖掘，洞察业务安全风险和隐患，协助制定高效的安全策略，找到作弊者的行为特点，快速形成有效可落地的检测和打击策略模型，持续迭代优化电商业务场景的风控效果。
任职要求
1、本科及以上学历，计算机、数学、统计学等相关专业优先，有数据分析类实习经验优先；
2、熟悉MySQL/Hive/Hadoop/Spark，能对海量数据进行高效分析和处理。
3、思维开阔敏捷，自我驱动，结果导向，有强烈责任感，良好的执行力和团队合作精神。【滴滴】北京滴滴无限科技发展有限公司

Process finished with exit code 0
```

图 3（Python 爬取过程）

2. 数据清洗

数据清理操作如下:

1. 岗位名称：BOSS 直聘搜索数据分析师，它会推荐数据分析师相关的工作。我们这次分析的方向主要为商业数据分析为主，所以大数据工程师，数据爬虫等技术类岗位进行去除。

2. 岗位职责：由于 BOSS 直聘岗位职责与任职要求文本没有明确分开，所以我通过 EXCEL 文本函数 find&mid 做了文本处理，分开了岗位职责与任职要求的

信息。

3. 数据建模

岗位职责内容这么多，如何才能找到重点标签？最好的办法是做一个词云，找出重点的词作为我们分析的标签。（如图 4）



图 4

我用 Python 做了岗位职责词云并且圈选了一部分重点的词，同时为对应的词做了分类。

分类情况

工作对象：客户、运营、产品、用户

数据建模：开发/设计、建模、指标、

数据可视化：报告/报表

数据落地：监控、方案、项目、优化、策略

【注：】由于开发/设计、报告/报表代表的内容比较相同，所以只要岗位职责出现其中 1 个词，都计算 1 次出现次数。

4. 数据可视化

数据可视化操作：

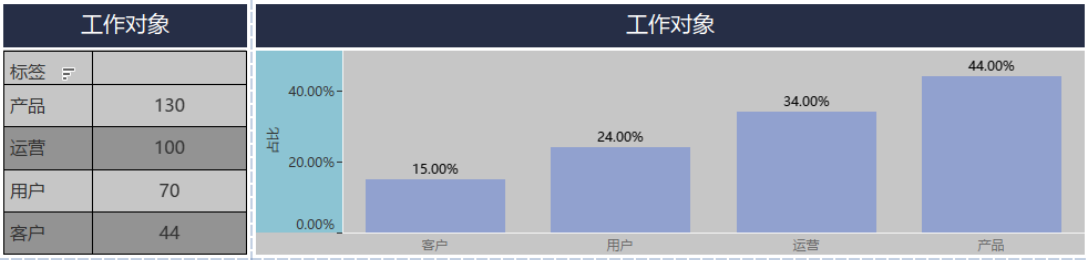
1. 数据处理：需要对各个重点词统计出现次数与出现占比，通过运用 Python 循环+SQL 语句的查询功能，可以避免在 EXCEL 中做多次筛选获取数据的尴尬。（如图 5）

2. 数据可视化：输出重点词出现次数与出现占比数据传输到 sql，运用 Tableau 进行数据可视化。（如图 6、图 7）

```
for word in f:
    word = word.strip()
    if re.search('[.!',word):
        my_text1 = pd.read_sql_query(f"SELECT "
                                     count(*) as '{word}数里',
                                     round(count(job_demand)/(SELECT count(*) FROM full_data),2) as '{word}占比'
                                     FROM
                                     full_data
                                     WHERE
                                     job_demand
                                     REGEXP
                                     '{word}',engine)
```

图 5

主要分析对象-产品



主要建模内容-开发指标

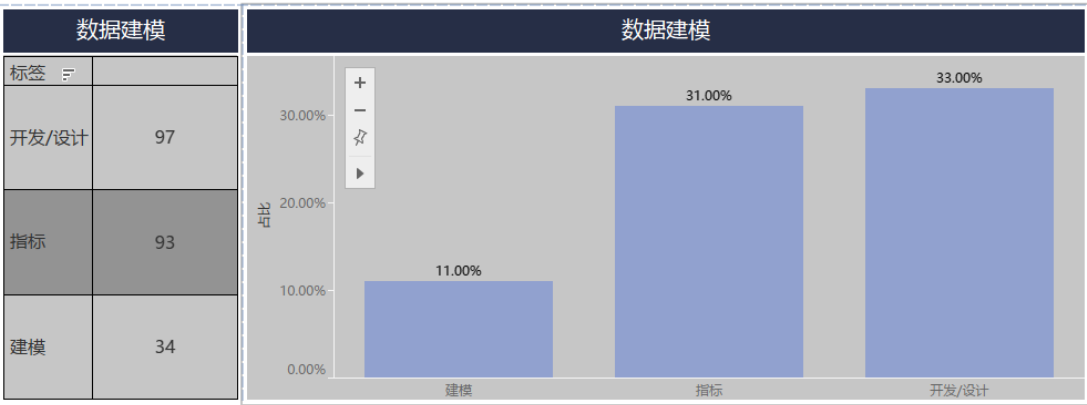
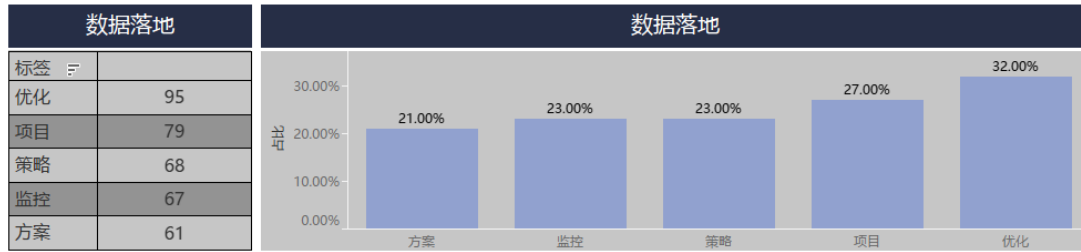


图 6

主要工作任务-优化



呈现报表占比接近一半



图 7

5. 数据报告

作为一个医生，对应的数据分析 CT 图出来了，就需要断症。

1. 数据主要分析对象-产品

可以看到产品出现占比为 44%，高出第二名运营的 10%。可以表明企业对数据分析师的需求侧重在产品端。

2. 主要建模内容-开发指标

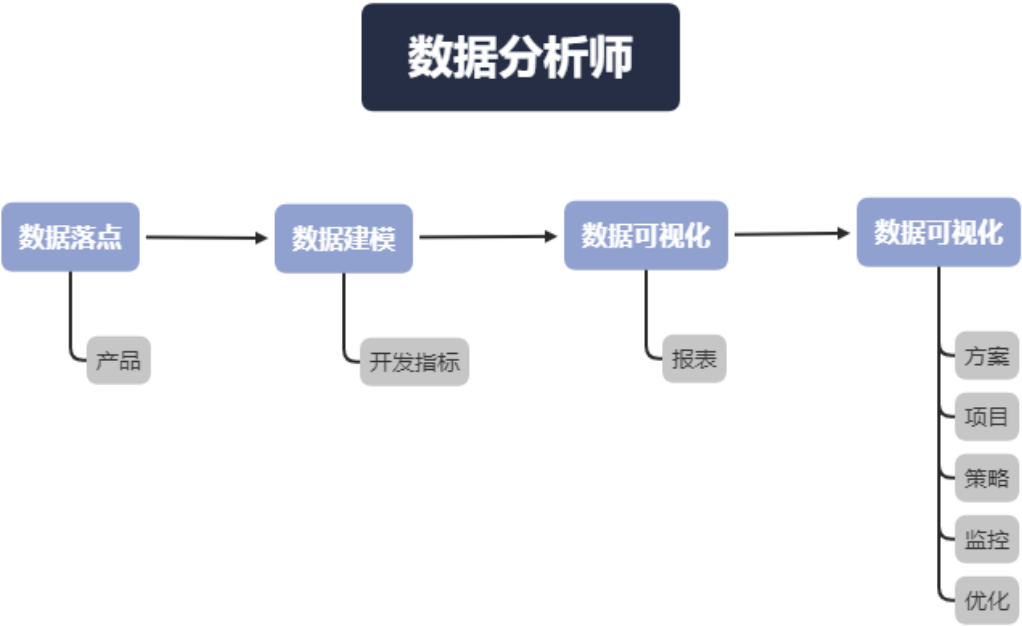
可以看到指标（重点词）出现占比 31%，开发/设计出现占比 33%。表明数据分析师建模当中最多的工作内容是开发与设计不同的模型，并根据实际项目设置相应的数据指标。

3. 主要工作任务-优化

可以看到优化占比 32%，但优化与其他重点词出现占比差距并不大。虽然数据分析师主要工作任务为优化，但方案、监控、策略、项目也是数据分析师的工作任务之一。

4. 呈现报表占比接近一半

报表/报告占比接近一半。表明报表设计与呈现能力是数据分析师必不可少的能力。



6. 数据优化

优化是数据分析师最主要的任务，做为一个好的数据分析师数据报告呈现后必须根据数据报告的内容做优化分析。

1. 数据报告维度单一。这次分析的数据维度为 1 维，只能得出一个数据分析师日常做什么？但是如果这份报告需要用到求职场景，求职肯定会涉及地点、薪资、行业等因素。故把岗位职责+地点、岗位职责+薪资再进行二维的分析是必须的。

2. 数据报告结论广泛。以结论 1 为例子，互联网行业中产品分为实体产品和服务型产品，这里指的是什么产品？实体产品也有具体的分工链条：生产→库存→销售→财务，报告是指产品的哪项分工，需要根据上述步骤重新做一次有关产品的词云，才能下准确结论。

3. 缺少时间维度。通过定期收集数据，做一个关键词出现次数折线图。以 7 天为一个周期，观察每个时间段关键词的趋势情况。通过观察关键词出现次数增长或衰减，分析每个时间段数据分析师岗位需求。

后续会根据数据优化方向出数据分析师体检 2.0 报告。