TASK5：公开广告数据的处理和市场基本竞争情况

截至：11月20日  22:00

在TASK4中，各位学到的是一套以公开的广告数据为基础的市场和竞品研究方法。接下的几个任务会要求大家自己来逐一实现这个方法中的各个模块，在练习中加深对解决方案设计思路的理解。TASK5要求各位完成的主要内容包括：清理数据、描绘广告主所属行业的SEM投放格局、为广告主圈定主要的竞争对手。

按照TASK4要求所取得的原始的公开广告数据，需要经过清理才能使用。**清理数据通常包含两个主要任务**：评估数据质量；对数据进行加工。其中，**评估数据质量**有一些**通用的**项目，也有在解决某个特定问题时**因事项特殊要求而设立**的评估标准。

以这份广告数据为例，首先它是一份结构化数据。一般的，对于结构化的数据，我们要评价各个变量取值的**缺失**情况、评价每个变量的**取值范围**和分布是否有异常、是否有重复观测或其他**无效数据**等。

而这次分析的目标——了解广告主所在行业的典型竞争者的投放策略——要求我们所使用的数据能够**覆盖行业中典型广告主**在所有**重要营销角度**方面的投放；也就是说，这份数据要包含足够多的广告主、足够丰富的关键词、各个典型的时段(如一天24小时、工作日与休息日等)、广告主业务可以辐射到的各个地域。当然，后两个标准是针对真实的生产环境而言的，在这个项目中我们重点是走通流程、理解方法，在这两个方面可以适当放宽。

通过质量评估的数据，通常仍然不适合直接进行分析，需要按照分析目的进行进一步的加工。**对数据进行加工**是对通过质量评估的数据进行筛选、打通、抽取、转换、计算等操作，使它们可以被用于分析和建模。**筛选**是要取得符合分析目的的数据。以我们这个竞品研究的例子来说，假定你是一个售卖高端礼品的电商，产品中有一些高端床品，那么当你搜索这些高端床品相关的词时很可能会发现出现一些制售床品的厂家的广告；如果你把自己定义为一家电商公司，那么在开始分析工作之前就可以把这些非电商类的生产厂家先过滤掉。除了与分析目标无关的数据之外，筛选工作经常还涉及处理质量不合格的数据、在预研阶段从全量数据中抽取样本等情况。**打通**数据的例子，大家可以回顾TASK2中通过关键词ID关联搜索引擎报表和网站访问日志的操作。**抽取**是指把原始数据加工成适合分析的格式。比如TASK4中要求大家记录的“飘红”这个变量，其实就可以看成是从原始创意文本中抽取出来的一部分。典型的**转换**动作比如，把文本替换为编码；这样可以减少冗余，节省存储空间。还有时候，我们需要对明细数据进行汇总，或者计算衍生指标(比如根据收益和成本计算ROI)，这都是一些为了方便后续的分析而进行的计算操作。

除了数据清理的工作，在TASK5中各位还会开始进行画图的工作。图形是发现和传递数据价值的重要工具，尤其是在初步探索数据特征的阶段，一张好图的表现力之强可称得上是“一图胜千言”。在后半段的几个任务中，画图都会是任务重点之一。在TASK5中，我们首先来建立评价一张好图的基本标准。

一张好图，首先要做到“**选型正确**”，成为“正确的图”是成为“好图”的必要条件。什么是“正确的图”？从工作步骤上来讲，画图是工具，不是目的。决定目的的工作，要在画图之前。在有了一个目的之后，再根据这个目的和数据的特征来选择恰当的图形。举例来说：

1. 你想要看哪些因素会影响用户购买金额之间的关系，可能的影响因素有性别等分类变量，也有注册时长等连续型变量。对于分类变量和客单价的关系，可以画分类的盒图；对于连续型变量，可以画散点图、气泡图等

2. 你要探索或展现一个时间序列的变化，柱状图和折线图其实都可以胜任。但是稍微细致一点的话，观测期数比较少(比如只有三五个数据点)的时候，柱状图会更美观一点；而观测期数比较多的时候(比如超过十个)的时候，一般折线图会更适合一点

3. 柱状图和条形图看起来很像，好像只是陈列的方式分别是水平和垂直排列而已。大部分情况下它们之间的区别确实不大，不过如果你的分类标签比较长、类别数又不太少的话(比如每个分类标签都有十个左右汉字、一共有七八个类别)，那么条形图看起来会更整齐清爽、布局合理

上面的这些例子中给出的方案，不是唯一的，甚至算不上是最好的，但它们至少不是错的——选型正确，一张图形至少能够得到一个及格分。相反，再精妙新颖的图形，如果和你的分析目标以及要表现的数据特征不搭配，也只具备陈列和炫耀的意义，不会成为一张实践场景中的“好图”。

这次任务中画图相关的工作，在选型上难度不大。但是本着“要事先行”的原则，我们还是要先从这个基础的标准说起。如果评价标准不正确，画图的时候很可能会舍本逐末，浪费了这个好工具。希望大家在完成下面几个任务时、甚至是在后续的学习和职场生涯中，都能够逐渐成长为绘制数据图形的高手。

TASK5需要以各位在TASK4中所抓取到的公开广告数据为素材开展工作。如果有些同学实在没有能力自己获取一份合格的数据，或者暂时还没有积累好足量的数据，可以先用一份历史数据来完成TASK5的要求。不过如果选择使用历史数据，报告的得分上限会被压缩；并且历史数据中各自都有一些小问题，会影响到个别模块的产出，或者需要你补充一些数据。所以，仍然鼓励各位尽量自己抓取到一份质量和数量都足够可靠的数据。

【任务内容】

操作与问答(文字描述合计不超过1500字)：

1. 评估数据质量：对TASK4中抓取到的公开广告数据进行**系统的**质量评估。列出你进行的所有评估项目，每个评估项目列明评估方法、评估标准和结论。如：

**评估项目**：是否覆盖了足够多的关键词

**评估方法**：统计每个搜索词对应的广告数量

**评估标准**：

(1) 对全部数据：搜索词数量/关键词样本中关键词数量>XX%

(2) 按关键词样本中的分类分组统计每一组：搜索词数量/关键词样本中关键词数量>XX%

**评估结果**：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 是否覆盖足量关键词 | 抓取数据中的搜索词数  (a) | 关键词样本中的关键词数  (b) | 覆盖率(a/b) | 评估结论 |
| 全体 | …… | …… | …… | 通过 |
| 分类1 | …… | …… | …… | 数据不足，补充抓取 |
| 分类2 | …… | …… | …… | 通过 |
| …… | …… | …… | …… | …… |

2. 数据加工：将原始数据清理为可以按照TASK4方法实施分析的数据。内容包括：

(1) 将URL转换为广告主名称

(2) 删除质量评估中不符合要求的数据；增补质量评估中要求增补的数据

(3) 补充分析所需的其他数据(如需)

(4) 其他你认为需要完成的数据加工工作：逐项列出

(5) 经过加工的数据，共计取得了多少条广告？覆盖了关键词样本的比例是多少？

以Excel表格形式提供加工好的数据。(4)~(5)以问答题格式提交在PDF文档中。

3. 描述市场基本格局：用你认为最恰当的方式展示广告主所在市场在如下几个方面的特征

(1) 市场整体竞争格局

(2) 广告主在市场中的竞争地位

(3) 广告主应当主要关注哪些竞争对手

(4) 其他你认为重要的市场整体特征

**按照数据分析报告的格式提交**。

【提交要求】

提交**Excel文件**(加工好的数据)和**Word文档**(问答题答案)。文档内容包含姓名、题目、回答。宋体，小四号字，1.5倍行距，每道问答题后列明字数。**存在以下任何一种情况存在，本次任务计0分：**

1、 迟交

2、 雷同

3、 抄袭

【参考资料】

1、 关于基于SEM广告的行业和竞品分析方法

a) TASK4参考文档《一种使用SEM广告展现数据优化投放策略的方法》

b) <https://mp.weixin.qq.com/s/RGNbO9_Y7XFZiJf7tv0aXw>

c) <http://teach.xiong99.com.cn/mooc/courseware/detail/139507670481637376>

注意：以上参考文档中的图形有部分是帮助理解思路的示意图，并不是数据分析报告中规范的统计图表。规范的统计图表和数据分析报告撰写方式，请查阅第2、3项资料。

2、 关于撰写数据分析报告：

a) 《数据思维实践》系列课程：<http://teach.xiong99.com.cn/mooc/user/home/16/1>

3、 关于画图：

a) 丑图百讲：<https://mp.weixin.qq.com/s/MHLA0twJf27GcOQGzz-1Aw>

b) 好图范例：

i. 谢益辉《现代统计图形》，第一章

ii. (选读)<https://www.ted.com/speakers/hans_rosling>

c) 统计图选型的基本思路：<https://extremepresentation.typepad.com/blog/2006/09/choosing_a_good.html>

可供使用的数据集：

链接: <https://pan.baidu.com/s/1F7d52ZC5cpiT7Att-LRyrQ>  提取码: p91v

链接: <https://pan.baidu.com/s/1Q48_8cFohpsQD0bUttK5uw>  提取码: ny2z

链接: <https://pan.baidu.com/s/1g-IVNa8mXLFb_0nITr7CBg>  提取码: ymze