**模块B：数据分析（40分）**

**竞赛题目：广告数据分析**

**背景简介**

随着互联网人口红利的消失，各大互联网产品都争先恐后开始做用户增长，希望以最低的成本拉新促活更多用户。用户增长的策略和手段有很多，其中互联网效果广告投放是非常重要的一种手段。

在互联网广告投放过程中，负责投放的运营同学为了保证最优的投放收益率，会从广告的各个角度去做投放优化，如：文案、图片、排版、媒体位置以及定向人群等等。而数据分析能从量化的角度指导运营如何去投放广告，如何优化投放组合，从而降低用户获取的成本。

**数据介绍**

（1）ad1.csv

| **字段** | **字段描述** | **数据含义** |
| --- | --- | --- |
| userid | 用户ID |  |
| time\_stamp | 时间戳 |  |
| adgroup\_id | 广告ID |  |
| pid | pid资源位 |  |
| noclk | 是否未点击 | 1：是，0：否 |
| clk | 是否点击 | 1：是，0：否 |
| cate\_id | 商品ID |  |
| campaign\_id | 广告计划ID |  |
| customer | 用户ID |  |
| brand | 品牌ID |  |
| price | 价格 |  |
| final\_gender\_code | 性别特征 | 1：男，2：女 |
| age\_level | 年龄层次 |  |
| pvalue\_level | 消费等级 | 1：低，2：中，3：高 |
| shopping\_level | 购物深度 | 1：浅，2：中，3：深 |
| occupation | 是否学生 | 1：是，0：否 |
| new\_user\_class\_level | 城市等级 |  |

（2）ad2.csv

| **字段** | **字段描述** |
| --- | --- |
| ad\_id | 广告ID |
| xyz\_campaign\_id | 广告系列 |
| age | 年龄 |
| gender | 性别 |
| interest | 兴趣 |
| Impressions | 展示次数 |
| Clicks | 点击次数 |
| Spent | 广告花费 |
| Total\_Conversion | 咨询次数 |
| Approved\_Conversion | 购买次数 |

**环境介绍**

在本竞赛模块环境中，已自动启动了 jupyter notebook 环境。环境中已安装 Python 中多种数据分析库，包括：

* Numpy
* Pandas
* Scipy
* Matplotlib
* Datetime
* Sklearn
* 其它

如果选手需要用到其他分析库，可根据实际情况自行安装。

**任务列表**

1. 数据探索与处理
2. 广告渠道分析
3. 投放时间分析
4. 用户画像分析
5. 广告成本分析

**任务启动**

**任务1：数据探索与处理**

**1.1 数据读取**

读取data/ad1.csv文件，指定第一行为列名，存入变量df中。

**1.2 处理缺失值**

对df变量进行操作，移除空值率超过60%的列字段，保存到变量df\_notnull中，并输出结果到results/1\_2.txt文件里。

**1.3 重复值处理**

对df\_notnull变量进行操作，删除多余行重复值，保存到变量df\_drop\_duplicates中，并输出结果到results/1\_3.txt文件里。

**1.4 异常值处理**

对df\_drop\_duplicates变量进行操作，按以下指定要求处理，最终结果保存到变量df\_clean中，并输出结果到results/1\_4.txt文件里。

* 只保留商品价格小于100000的记录

**1.5 数据格式处理**

对df\_clean变量进行操作，按以下指定要求处理:

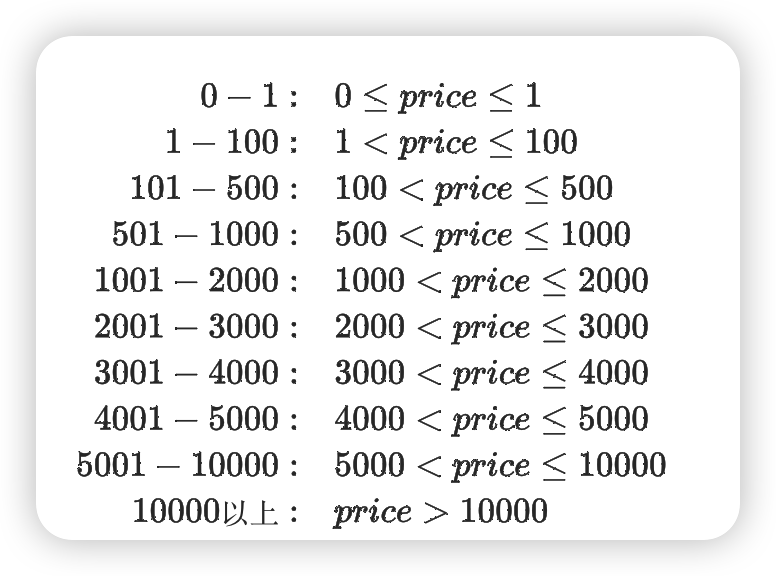
涉及时间计算时，一律采用东八区时间

* 将time\_stamp字段转换为时间类型，保存在day新字段（格式为：yyyy-mm-dd HH:mm:ss）,并输出结果到results/1\_5\_day.txt文件里。（0.5分）
* 提取day字段中的日期信息，保存到date新字段（格式为：yyyy-mm-dd）,并输出结果到results/1\_5\_date.txt文件里。（0.5分）
* 提取day字段中的周信息，保存到week新字段（返回本周的第几天，注意星期天为第7天）,并输出结果到results/1\_5\_week.txt文件里。（0.5分）
* 提取day字段中的小时信息，保存到hour新字段（24小时计算，只返回小时数）,并输出结果到results/1\_5\_hour.txt文件里。（0.5分）

**任务2：广告渠道分析**

**2.1 探查商品价格对点击量的影响**

对df\_clean变量进行操作，按以下指定要求处理:

* 根据以下要求划分price字段，并计算各价格区间点击量，保存到price\_level新字段
* 查找出点击量最高的价格区间，保存到变量result\_2\_1中，并输出结果到results/2\_1.txt文件里。

**2.2 高点击量价格区间分析**

对df\_clean变量进行操作，按以下指定要求处理:

* 在点击量最高的价格区间数据范围中处理，计算深度用户的占比，结果保存到变量result\_2\_2中，并输出结果到results/2\_2.txt文件里。（保留四位小数，无需转换为百分比）

**2.3 商品与点击率分析**

对df\_clean变量进行操作，按以下指定要求处理:

* 查找出展示量前十的商品，然后计算不同商品的点击率，查询出点击率最低的cate\_id（商品id)，结果保存到变量result\_2\_3中，并输出结果到results/2\_3.txt文件里。（四舍五入，保留三位小数）

**2.4 低点击率分析**

**2.4.1 性别差异**

对df\_clean变量进行操作，按以下指定要求处理:

* 用2.3题查询出的cate\_id对数据进行筛选。
* 在筛选后的数据中，计算男女不同性别点击率，查询出点击率最低的final\_gender\_code（性别），结果保存到变量result\_2\_4\_1中，并输出查询结果到results/2\_4\_1.txt文件里。

**2.4.2 年龄层差异**

对df\_clean变量进行操作，按以下指定要求处理:

* 用2.3题查询出的cate\_id对数据进行筛选。
* 在筛选后的数据中，计算不同年龄层的点击率，查询出点击率最低的age\_level（年龄层），结果保存到变量result\_2\_4\_2中，并输出结果到results/2\_4\_2.txt文件里。

**2.5 高点击商品与用户性别关系**

对df\_clean变量进行操作，按以下指定要求处理:

* 查找出展示量前十的商品所有数据，然后计算其中不同商品不同性别（final\_gender\_code）的点击率，查询出男性点击率的最高值，结果保存到变量result\_2\_5中，并输出结果到results/2\_5.txt文件里。（四舍五入，保留三位小数）

**2.6 高点击商品与资源位点击量分析**

对df\_clean变量进行操作，按以下指定要求处理:

* 查找出展示量前十商品的所有数据，然后查询出其中最高点击量的pid（资源位），结果保存到变量result\_2\_6中，并输出结果到results/2\_6.txt文件里。

**任务3：投放时间分析**

涉及时间计算时，一律采用东八区时间

**3.1 各时间段点击率分析**

对df\_clean变量进行操作，按以下指定要求处理:

* 计算每个时段（hour）平均点击率
* 在上一步结果中，查询出点击率最高的时间段（hour），结果保存到变量result\_3\_1中，并输出结果到results/3\_1.txt文件里。（保留三位小数）

**3.2 各星期点击率分析**

对df\_clean变量进行操作，按以下指定要求处理:

* 以week字段分组，计算点击率，查询出点击率最高的week，并输出结果到results/3\_2.txt文件里。（保留三位小数）

**3.3 周末商品各时段点击量**

对df\_clean变量进行操作，按以下指定要求处理:

* 计算星期六每个时间段（hour）的平均点击率
* 在上一步结果中，查询出点击率最高的时间段（hour），结果保存到变量result\_3\_3中，并输出结果到results/3\_3.txt文件里。（保留三位小数）

**任务4：用户画像分析**

**4.1 年龄层与消费档次**

对df\_clean变量进行操作，按以下指定要求处理:

* 计算不同年龄层次（age\_level）与不同消费档次（pvalue\_level）之间的用户人数。
* 在上一步结果中，查询出在用户最多的年龄层次（age\_level）中，不同消费档次（pvalue\_level）最高有多少用户，结果保存到变量result\_4\_1中，并输出结果到results/4\_1.txt文件里。（取整数）

**4.2 购物深度与性别关系**

对df\_clean变量进行操作，按以下指定要求处理:

* 计算不同购物深度（shopping\_level）与不同性别（final\_gender\_code）之间的用户人数。
* 在上一步结果中，查询出在深度用户中，男性占比为多少，结果保存到变量result\_4\_2中，并输出结果到results/4\_2.txt文件里。（保留四位小数）

**4.3 不同年龄层次、不同性别的用户点击率差异性**

对df\_clean变量进行操作，按以下指定要求处理:

* 计算不同年龄层次（age\_level）与不同性别（final\_gender\_code）之间的点击率。
* 在上一步结果中，判断是否所有年龄层次点击率都是女性高于男性，并输出是或否到变量result\_4\_3中，并输出结果到results/4\_3.txt文件里。

**4.4 用户价值分析**

对df\_clean变量进行操作，按以下指定要求处理:

* 选用以下三个指标作为判断用户价值的标准，运用 rfm 模型对用户划分成五大类客户。
  + **用户平均购物层次:** 体现用户在平台上的持续购物深度
  + **用户平均广告点击率:** 体现用户对广告的接受程度
  + **浏览广告的商品平均价格:** 体现用户通过广告推广能带来的潜在收益
* 五大类客户类型分别对应为：重要保持用户、重要发展用户、重要挽留用户、一般用户以及低价值用户。
  + **重要保持用户:** 这类用户有着高点击率，有更大的可能将用户的点击率转换为实际的购买行为.
  + **重要发展用户:** 这类用户点击率不高、对广告商品的价格也不高，但购物深度深。
  + **重要挽留用户:** 这类用户点击率偏低，但是其浏览的广告商品的价格很高.
  + **一般用户:** 这类用户点击率低，浏览广告的商品价格适中，购物深度也适中。
  + **低价值用户:** 这类用户点击率低，浏览广告商品价格低，购物深度也低.
* 在上一步结果中，查询并判断userid为794330、823441、1135599的用户是哪种类型客户，结果保存到变量result\_4\_4中，并输出结果到results/4\_4.txt文件里。（保存格式：'客户类型1、客户类型2、客户类型3'）

**任务5：广告成本分析**

**5.1 数据读取**

读取data/ad2.csv文件，指定第一行为列名，存入变量data中。

**5.2 计算每类广告系列平均花费**

对data变量进行操作，按以下指定要求处理:

* 计算不同系列广告每万条展示量的平均价格。（保留两位小数）
* 在上一步计算结果中，保存xyz\_campaign\_id为916的广告每万条展示量的平均价格到price\_916变量中，并输出结果到results/5\_2\_916.txt文件里。
* 在上一步计算结果中，保存xyz\_campaign\_id为936的广告每万条展示量的平均价格到price\_936变量中，并输出结果到results/5\_2\_936.txt文件里。
* 在上一步计算结果中，保存xyz\_campaign\_id为1178的广告每万条展示量的平均价格到price\_1178变量中，并输出结果到results/5\_2\_1178.txt文件里。

**5.3 计算每类广告系列平均点击率**

对data变量进行操作，按以下指定要求处理:

* 计算不同系列广告平均点击率。（保留六位小数）
* 在上一步计算结果中，保存xyz\_campaign\_id为916的广告平均点击率到clk\_rate\_916变量中，并输出结果到results/5\_3\_916.txt文件里。
* 在上一步计算结果中，保存xyz\_campaign\_id为936的广告平均点击率到clk\_rate\_936变量中，并输出结果到results/5\_3\_936.txt文件里。
* 在上一步计算结果中，保存xyz\_campaign\_id为1178的广告平均点击率到clk\_rate\_1178变量中，并输出结果到results/5\_3\_1178.txt文件里。

**5.4 计算每类广告系列平均咨询率**

对data变量进行操作，按以下指定要求处理:

* 排除Clicks（点击量）为0的数据。
* 916、936、1178这三种系列广告分别对应线上转化率为10%、40%、100%。
* 计算不同系列广告平均线上咨询率。（保留四位小数）
* 在上一步计算结果中，保存xyz\_campaign\_id为916的广告平均线上咨询率到clk\_rate\_916变量中，并输出结果到results/5\_4\_916.txt文件里。
* 在上一步计算结果中，保存xyz\_campaign\_id为936的广告平均线上咨询率到clk\_rate\_936变量中，并输出结果到results/5\_4\_936.txt文件里。
* 在上一步计算结果中，保存xyz\_campaign\_id为1178的广告平均线上咨询率到clk\_rate\_1178变量中，并输出结果到results/5\_4\_1178.txt文件里。

**5.5 计算每类广告系列平均付费率**

对data变量进行操作，按以下指定要求处理:

* 计算不同系列广告平均付费率。（保留四位小数）
* 在上一步计算结果中，保存xyz\_campaign\_id为916的广告平均付费率到pay\_916变量中，并输出结果到results/5\_5\_916.txt文件里。
* 在上一步计算结果中，保存xyz\_campaign\_id为936的广告平均付费率到pay\_936变量中，并输出结果到results/5\_5\_936.txt文件里。
* 在上一步计算结果中，保存xyz\_campaign\_id为1178的广告平均付费率到pay\_1178变量中，并输出结果到results/5\_5\_1178.txt文件里。