**皇包车数据挖掘项目**

**背景**

皇包车（HI GUIDES）是一个为中国出境游用户提供全球中文包车游服务的平台。拥有境外10万名华人司机兼导游（司导），覆盖全球90多个国家，1600多个城市，300多个国际机场。截止2017年6月，已累计服务400万中国出境游用户。为更好服务用户同时也扩大盈利，皇包车推出了精品旅游服务，区别于普通旅游服务，精品旅游服务为用户提供高质量的旅游包车服务，该服务也成为平台最希望用户购买的服务产品。

要为用户提供更好的服务，首先就要更深刻理解他们的行为，为此，皇包车打算从用户个人信息、用户订单数据、APP行为数据以及用户评论数据进行分析，从中发现有价值的线索。例如对现有5万多名用户在旅游app中的浏览行为记录分析，能否发现哪些用户在浏览之后完成了订单，且享受了精品旅游服务，而哪些用户却在中途退出了，没有下单呢？皇包车最希望用户能够预定精品旅游服务，那么该如何分析用户的个人信息和浏览行为，从而预测用户是否会在短期内购买精品旅游服务呢？。

* **用户个人信息：**用户id、性别、省份、年龄段。
* **用户历史订单数据**：用户id，订单id，订单时间，订单类型，旅游城市，国家，大陆。其中订单类型中1表示购买了精品旅游服务，0表示普通旅游服务。
* **用户行为信息**：用户id、行为类型、发生时间。行为类型一共有9个，其中1是唤醒app；2~4是浏览产品，无先后关系；5~9则是有先后关系的，从填写表单到提交订单再到最后支付。
* **用户评论数据**：用户id，订单id，评分，标签，评论内容。

如有乱码显示，注意编码设为“UTF-8”

user\_profile**用户个人信息**

|  |  |
| --- | --- |
| **变量名称** | **解释** |
| Userid | 用户id |
| Gender | 性别 |
| province | 省份 |
| Age | 年龄段 |

orderHistory **用户历史订单数据**

|  |  |
| --- | --- |
| **变量名称** | **解释** |
| userid | 用户id |
| orderid | 订单id |
| orderTime | 订单时间 |
| orderType | 订单类型，1表示购买了精品旅游服务，0表示普通旅游服务 |
| city | 旅游城市 |
| country | 旅游国家 |
| continent | 所在洲 |

user\_action **用户行为信息**

|  |  |
| --- | --- |
| **变量名称** | **解释** |
| userid | 用户id |
| actionTime | 行为时间 |
| actionType | 行为类型，行为类型一共有9个，其中1是唤醒app；2~4是浏览产品，无先后关系；5~9则是有先后关系的，从填写表单到提交订单再到最后支付 |

user\_comment **用户评论数据**

|  |  |
| --- | --- |
| **变量名称** | **解释** |
| userid | 用户id |
| orderid | 订单id |
| rating | 评分 |
| tags | 标签 |
| commentsKeyWords | 评论内容 |

**要求：**

1. 以小组为单位，对数据进行初步的查看和分析，基于以上数据信息，讨论可能存在哪些业务问题值得深入分析？针对这些业务问题，你们计划进行哪些数据分析？
2. 基于1中的业务问题和数据分析计划，利用Python、SPSS Modeler等工具进行实现，同时完成数据分析报告，分析报告需要包括几部分内容（业务问题、数据分析基本过程、分析结果解读、对皇包车平台的建议），数据流文件需要满足几个要求（可运行、有一定注释、呈现方式尽量整齐、数据处理逻辑合理），数据挖掘过程至少需要采用本课程里涉及的关联、聚类、分类中的**两类方法**。
3. 以小组为单位提交modeler数据流文件和word数据分析报告文件（07.10）。

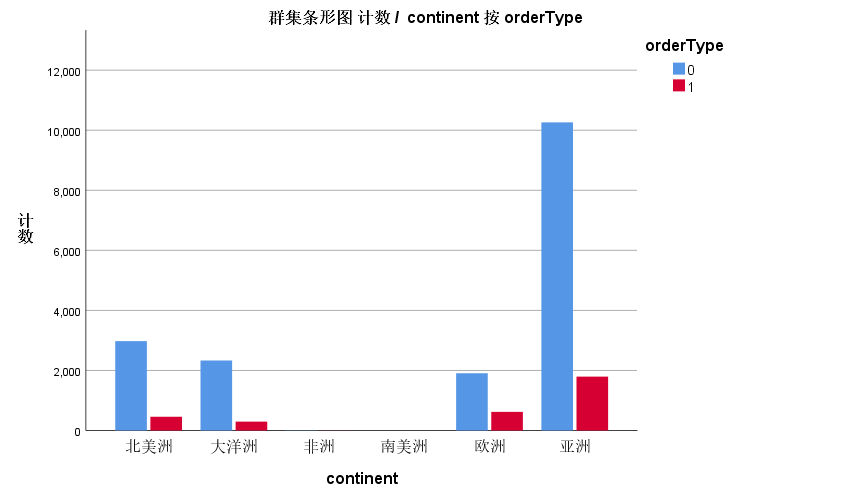
orderHistory **用户历史订单数据**

缺失值：

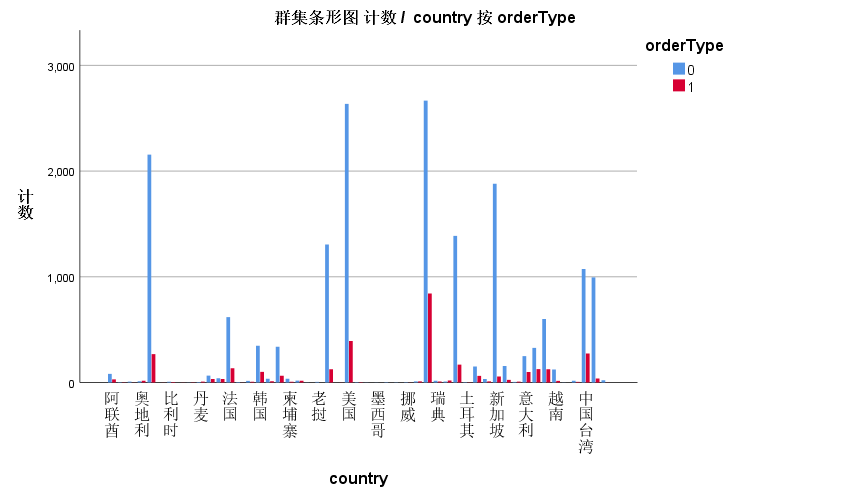
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | userid | orderid | orderTime | orderType | city | country | continent |
| 个案数 | 有效 | 20653 | 20653 | 20653 | 20653 | 20653 | 20653 | 20653 |
| 缺失 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

因此该数据表不存在缺失值，由于其内容为字符型，故也不存在离群值和极端值等问题需要处理。

分析：通过分析用户的订单类型和所在地区，得到了如下的结果：



1. 关于订单类型和大洲：可以看出大多数顾客更愿意选择亚洲，少部分愿意选择北美洲、大洋洲和欧洲，而在每个洲中，选择普通订单和精品订单的顾客比例差不多。



1. 关于订单类型和国家：可以看出，对于大部分国家，选择普通订单的顾客人数明显更多；而对于一部分欧洲国家，愿意选择精品订单的人数明显更多。

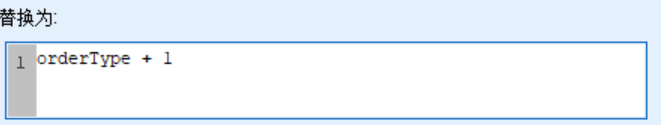
想法：

* + - 1. RFM

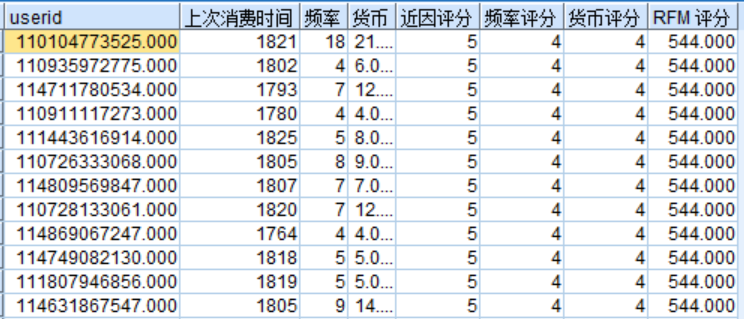
针对聚类这一方面，我们决定从RFM分析和用户类型聚类入手。使用RFM分析可以帮助该平台更好地分析哪些用户忠诚度更高，哪些用户更值得花更大的营销成本。

* 1. 数据预处理

为了区分普通服务的顾客和未购买服务的顾客，我们将所有的orderType字段的值加1。



随后，我们依据userid、orderTime和orderType字段来信息RFM分析，其中orderType字段为“值”，因为其相对地显示出顾客的购买金额（0,1,2）。得到结果如下：



（仅截取部分）

RFM算法能够有效地衡量客户价值和客户创利能力，该模型通过一个客户近期购买行为、购买的总体频率以及花了多少钱三项指标来描述该客户的价值状况，强调以客户的行为来区分客户。通过寻找RFM排名靠前的顾客id，可以确定哪些顾客更具有保留的价值，值得做专人营销。

* + - 1. 寻找哪种类型的顾客需要我们额外关注？