“日暮腾云起，东瀛万里行。”

——日本旅游价格水平研究报告

复旦第三区交通委小组

李泽君、赵雅滢、谢炳辉、王维实、钟诚

2019.6.1

目录

一、 背景介绍

“日暮腾云起,东瀛万里行。”

中日作为一衣带水的邻邦，自古以来就常有经贸互通和人员往来。上溯唐朝，就有日本僧人来华学习的记录，从 2013 年开始，随着日本签证的进一步放宽，国际航线扩充及邮轮停靠数量增加，赴日中国游客逐年增加。2017 年，中国大陆访日游客比上年增加 15.4%，达 735.58 万人次。根据携程发布的《2018 年中国游客赴日旅游报告》，日本作为中国公民出国旅游的第二大目的地国家，今年上半年中国大陆赴日旅客突破 400 万人次。无论是游客规模还是消费额，中国大陆游客均已经成为了日本旅游市场的主力军。携程的客户调查也发现，每年新增大量第一次赴日本的游客，同时多次赴日旅行的回头客很多。

随着中国赴日旅游市场不断扩大，赴日旅游的选择也越来越多。北海道温泉、富士山朝雾、秋叶原祭典、朝圣寿司之神，旅行社也推出了多样的旅行套餐供消费者选择，为帮助消费者选择性价比最高的旅游线路，本课题着眼于旅行社赴日旅游的 1949 条相关数据，探究影响赴日旅游价格的因素，力求为旅人的出行提供最优的方案。

二、 数据预处理

2.1 概述

数据由以下几栏组成：价格、名称、出游类型、优惠方式，有无住宿、住宿条件、出发地、供应商、满意度、出游人数、点评数，其中价格为连续变量，为需要研究的因变量；其余为自变量。

变量类型		变量名	取值水平	备注
因变量		价格	[7,75599]	
自变量	文本数据	名称		旅游项目的简单介绍
	分类数据	出游类型	自由行、跟团、定制游等	
		优惠方式	立减、餐厅优惠券、交通卡等	
		住宿条件	Wifi，住近市区等	
		出发地	西安、北京、成都等	
		供应商	途牛国旅、斑马旅游等	
	连续变量	满意度	[0,100%]	
		出游人数	[0,7798]	
		点评数	[0,1992]	

表 2.1　日本旅游项目价格相关特征概述

2.2　错误值处理

（1）数据中“有无数据列”一栏，部分标记为“无住宿”的样本在其他列中有住宿酒店的信息，逻辑矛盾，故综合两者信息重新构建有无住宿列。

2.3 衍生变量提取

（1）在住宿条件一栏中，提取关于酒店星级、住所是否近市区、自然美景、购物中心，有无 WIFI 的信息，构建新的数据列。

（2）初步处理出发地数据后发现从中国出发的项目价格比日本高，因而将中国和日本出发的项目作为两个类别，同时对多地出发的项目另作一个类别进行归类。

（3）根据各个水平的样本数，平均对数价格的描述，把本地玩乐停车，交通和交通卡都合并为“当地交通”水平；把当地特色服务，购物玩乐，自助和导游服务合并为“本地服务”水平，把定制游，度假酒店和半自助这几个样本量很少的水平合并成“其他”。

（4）对于“优惠方式”一列，很多水平都和“优惠方式”几乎无关，比如团建拓展，文娱体验等，应为项目特点，所以把这些值都设为“未提及”。其余的大多是“立减优惠”，而将其他数量较小的优惠方式设为“其他”。为了不损失信息，把这些项目特点存放到另外一列属性列里。

（5）供应商变量中，选择将样本数少于 30 的类别合并为“其他”。

（6）提取项目名称中的地点，如东京、大阪、北海道、神户、温泉等

（7）提取名称中关于旅游天数的描述作为新的连续变量，如果不是旅行或者没有提及，则置为 0。

（8）提取满意人数 = 满意率*出游人数，点评占比 = 点评数/出游人数作为协同变量。

三、 变量分析

3.1 因变量

首先对因变量水平进行分析并绘制直方图。

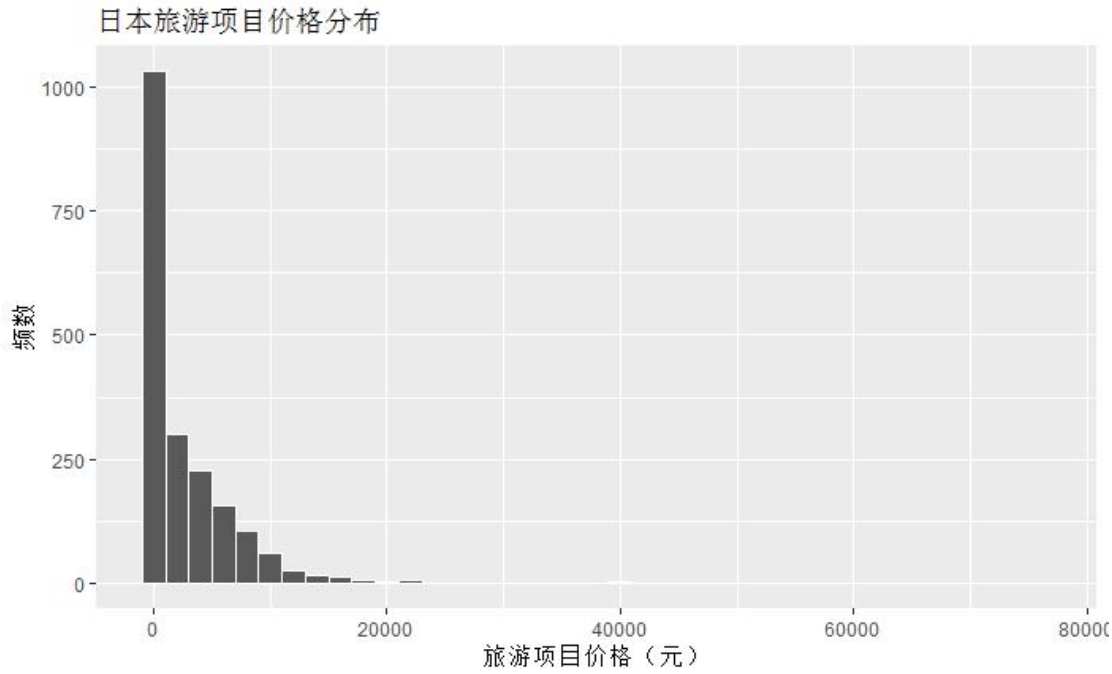

图 3.1.1　日本旅游项目价格分布直方图

通过观察价格分布直方图，可以发现样本中旅游项目的价格整体右偏，均值大于中位数。价高的旅游项目的价格要远高于大部分的旅游项目，例如价格最高的项目是“癌研有明医院体检 C 套餐 + 和服体验 5 天 4 晚游”，价格因为医疗服务高达 7.5 万。

而从图中我们可以看到绝大多数的样本的价格不超过 1 万元，导致绝大多数数据集中在最左侧而且难以观察数据的分布情况。且对数函数的特性在于当数值小的时候函数值变化大，数值很大的时候函数值变化小，故我们考虑使用价格的对数来作为因变量。

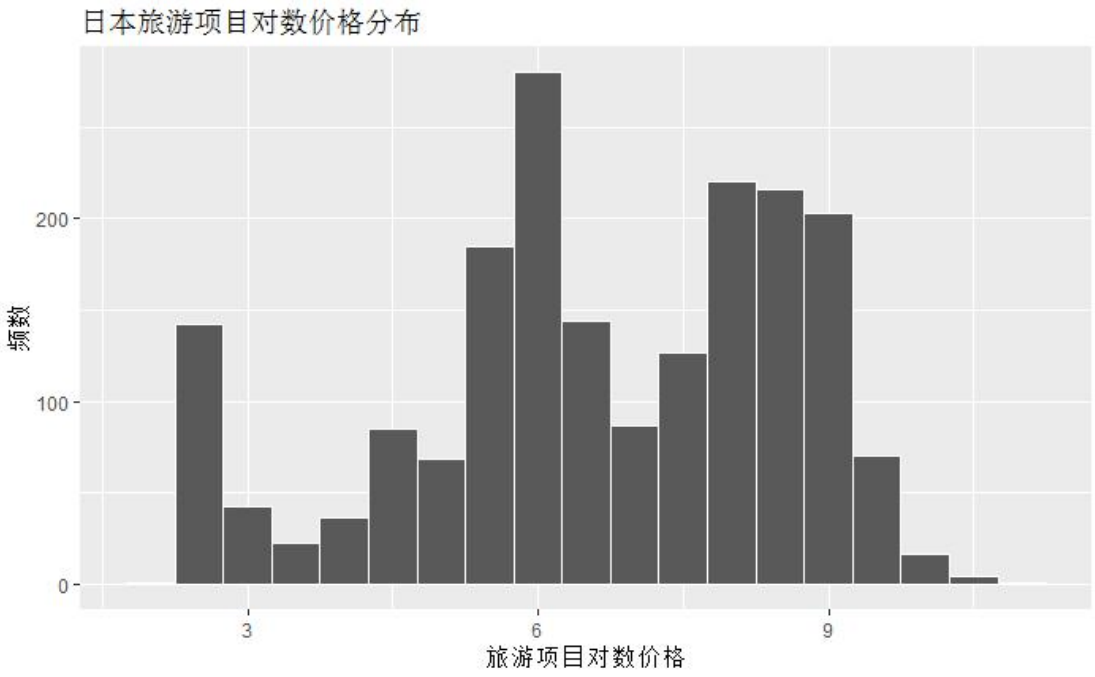

图 3.1.2　日本旅游项目对数价格分布直方图

观察对数价格的直方图，我们可以看出对数价格的分布相对集中，样本之间的差距也被缩小，同时大部分样本的对数价格位于 5.5 至 9 之间，即 300 元至 9000 元区间，其中价格偏低的是小型的旅游活动如半日游，当地的导游服务。偏高的是完整的旅行团的活动。此外，低于 300 元的多是一些优惠券，优惠门票，美食项目。

3.2 出游类型

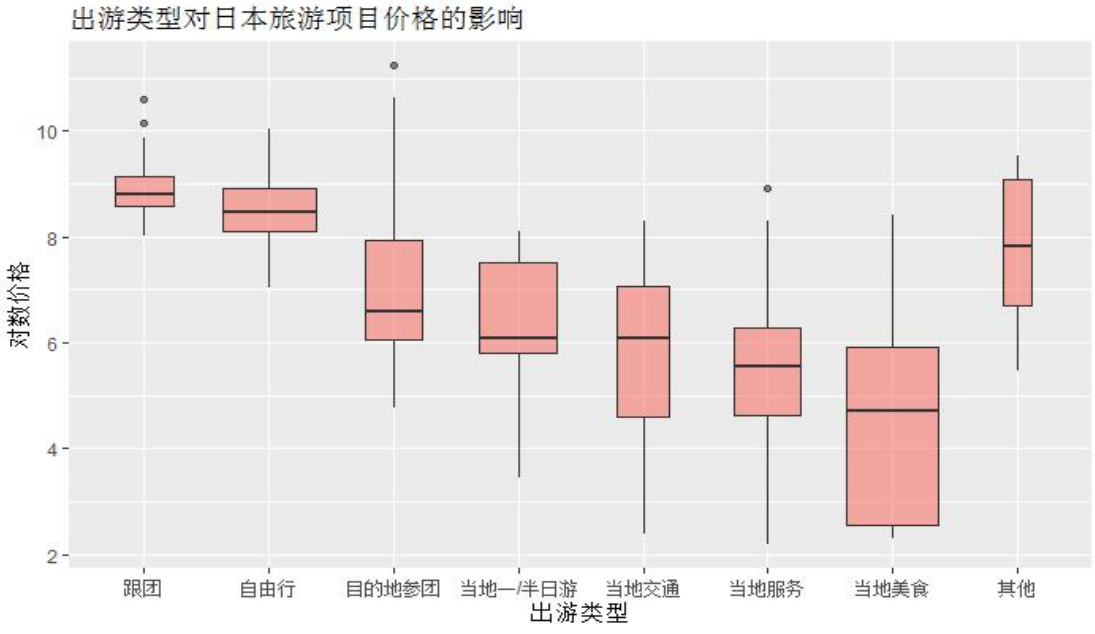

图 3.2　出游类型对日本旅游项目对数价格影响箱线图

通过观察数据，“当地”类旅游项目指当地餐厅订座，当地交通卡和门票的优惠项目，也即在旅游地当地享受的服务项目，而其他的项目则是有完整计划的旅游团。

从箱线图来看，对于各个箱子（类别的）总体高度，可以发现当地的旅游项目比其他的项目总体上价格更低，因为当地的旅游服务项目不需要承担交通费用和人力管理费用，而且当地的项目多数规模较小，所以价格更低。而观察箱子的宽度（代表样本量），我们可以发现数据中当地的服务项目数量更多，说明旅游平台不仅仅针对于旅行团和外地旅行，也有较多的本地项目。所以接下来对当地旅行和长途旅行来分别分析。

在当地的旅游项目中，美食是样本数最多的，而交通类服务项目最少，这种大小关系说明了本地旅游项目的市场倾向，美食市场要大于其他的需求，而交通服务项目较少。从高度（平均价格水平）和箱体长度（方差）来说，一日游的平均水平最高，其次是交通，美食最低，这样的结果也符合我们的价格预期，旅行成本最高，而吃喝相对比较便宜。美食项目的方差比其他类别的都要更大，说明服务类项目和一日游的价格比较接近，但是对于美食来说，餐厅的消费水平差距较大。

在长途旅行中，自由行最多，显示市场倾向于选择自由行而非跟团。总体价格上，跟团最高，其次是自由行，然后是目的地参团，与交通、人力上的花费相印证，“其他”水平中主要是定制游和度假酒店项目，自然价格也偏高。样本方差上，跟团和自由行的方差较小，这说明相对完整的旅游项目的价格比较均衡，目的地参团可能因为选择更多导致方差很大，其他类别同理方差较大。

3.3 优惠方式

从箱线图可以看出，立减优惠是最常见的优惠方式，其旅游项目的价格也总体偏高，因为立减优惠常用于刺激顾客进行高消费，故项目价格总体较高。而未提及样本稍低，“其他”水平优惠方式则是最低的，说明使用其他的优惠方式的项目总体价格偏低，方差较大，推测是采用其他优惠方式的项目属于比较便宜普通的项目如服务类并且种类较多。

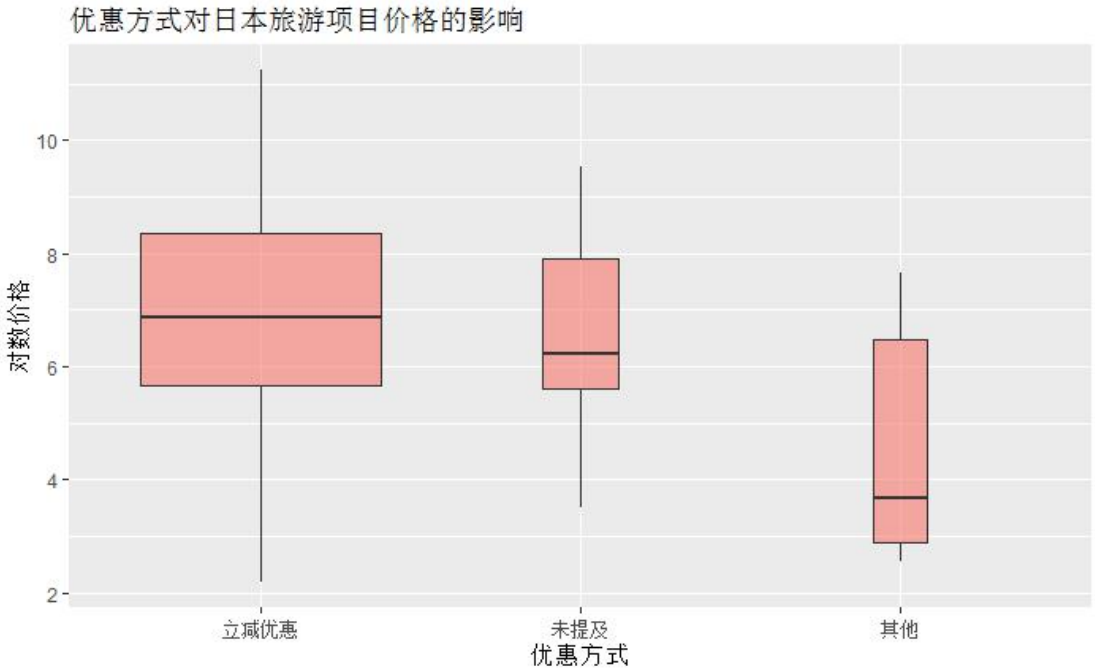

图 3.3　优惠方式对日本旅游项目对数价格影响箱线图

3.4 供应商

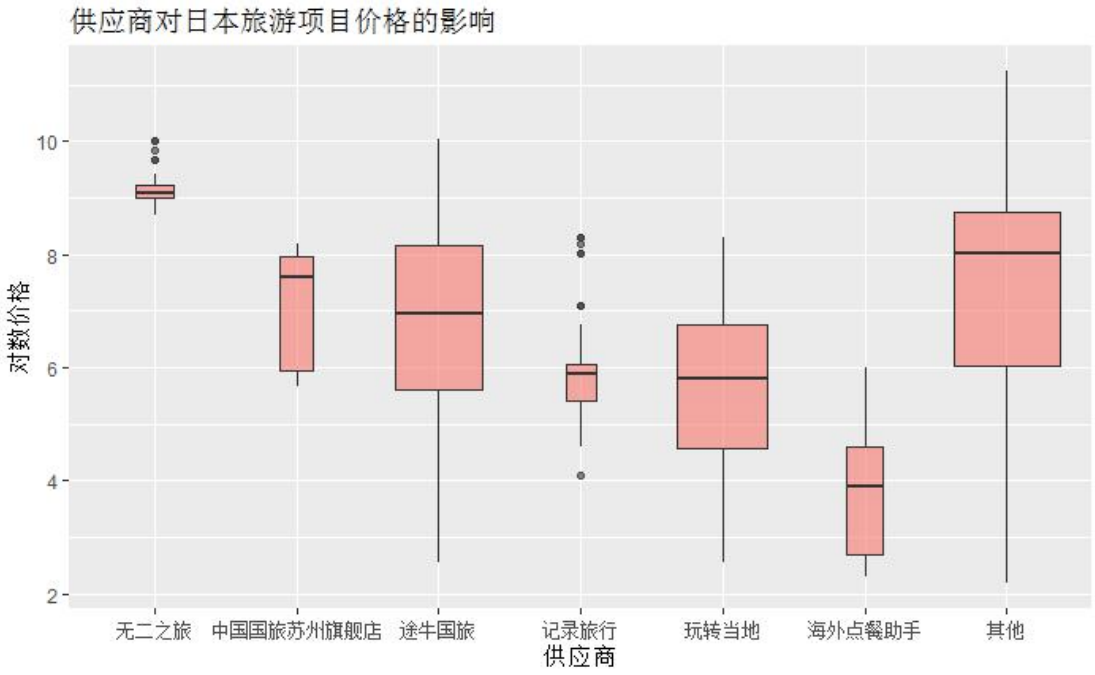

图 3.4　供应商对日本旅游项目对数价格影响箱线图

从箱线图的高度上来说，可以看出两个大供应商“途牛国旅”和“玩转当地”的总体价格水平处于中间位置，和“苏州国旅”也比较接近，这三家旅行社，两家是主要的大平台，一家是国旅，所以总体价格处于中游；而总体价格水平最高的是“无二之旅”的旅游项目，

说明其针对高端客户，而“记录旅行”则是针对普通消费者；“海外点餐助手”主要经营本地餐饮项目，所以价格较低。其他水平主要为小旅行社，针对高消费的小众群体，比如医疗旅游项目，所以价格稍高。国旅和大平台内部包含多种旅行项目，所以方差较大，小旅行社针对性强，所以方差较小。

3.5 有无住宿

不同的租赁平台面向不同层次的顾客，自然面向高端顾客的平台租金会比中低端顾客的高。在我们选择租赁平台的时候，也会根据房屋的质量选择不同的租赁平台，我们对不同平台上租金价格水平进行研究，把样本数较高的链家、若航寓与小租乐等大平台和其他样本数较低的租赁平台进行比较。

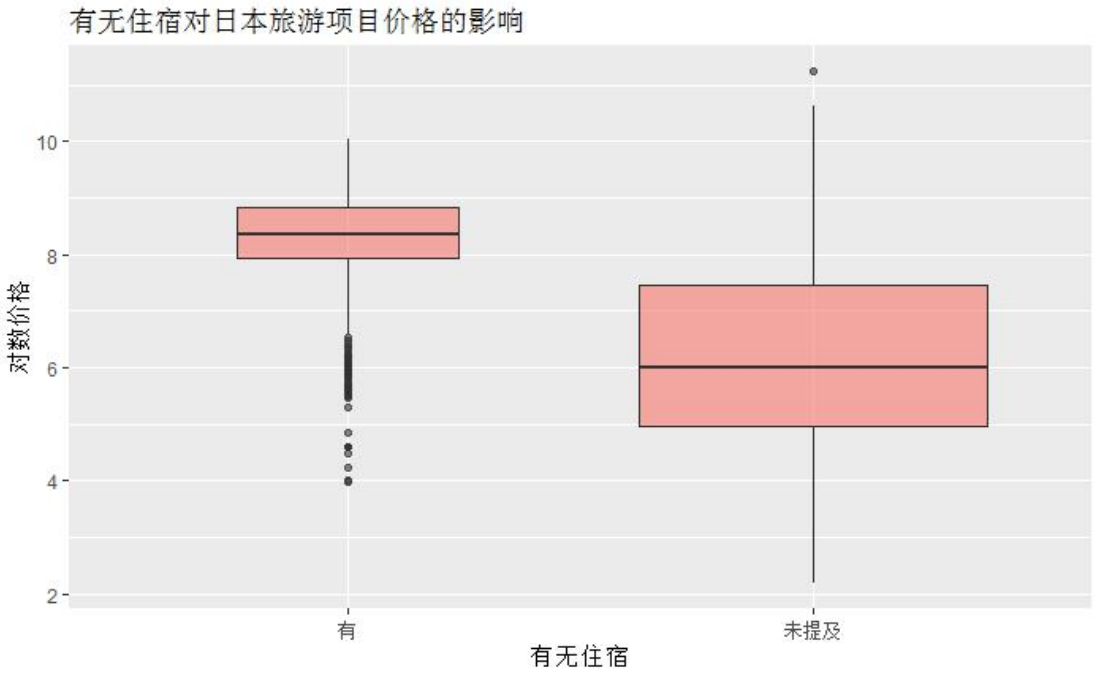

图 3.5　有无住宿对日本旅游项目对数价格影响箱线图

比较两箱线图，可以看到有住宿的项目的价格更高，方差更小。因为包含住宿费的项目价格起点较高，且聚集在高价位。而没有提及住宿的项目包含时间短且价格不高的主题娱乐公园门票、一日游项目等，也包含多天自由行的项目，因而方差较大；同时由于没有包括住宿的价格，整体价格相对较低。

3.6 住宿条件

在有住宿的样本中，我们进一步分析住宿条件对于日本旅游项目价格的影响。我们研究的住宿条件，包括住处是否靠近海滨、山、湖等自然美景，以及酒店星级。

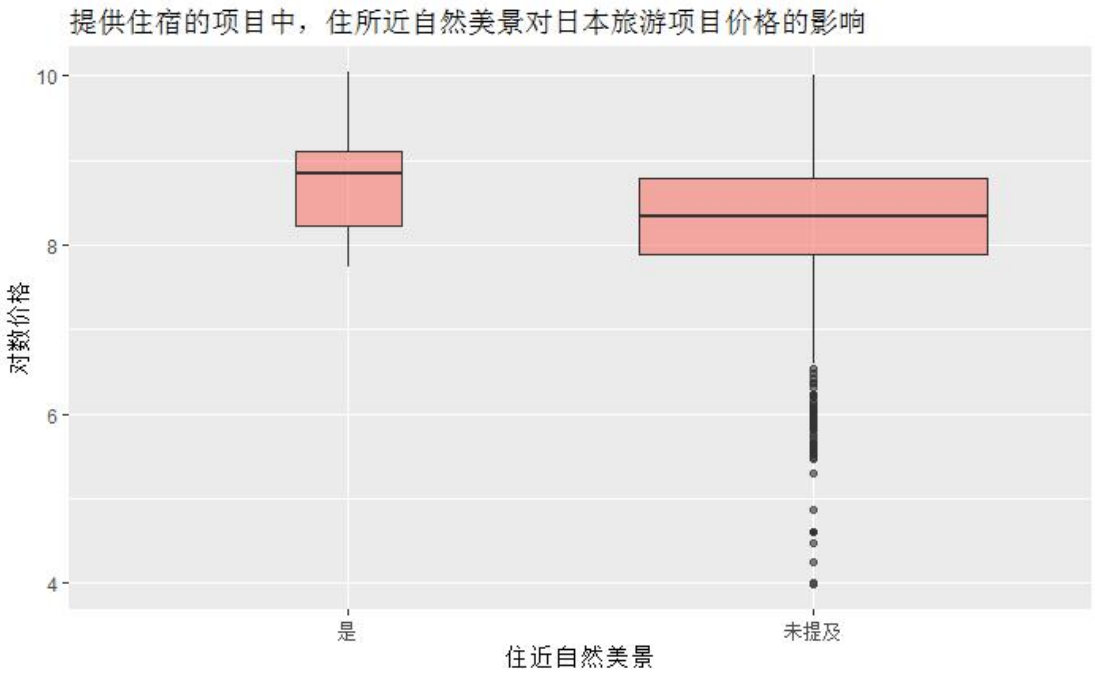

图 3.6.1　住宿条件对日本旅游项目对数价格影响箱线图

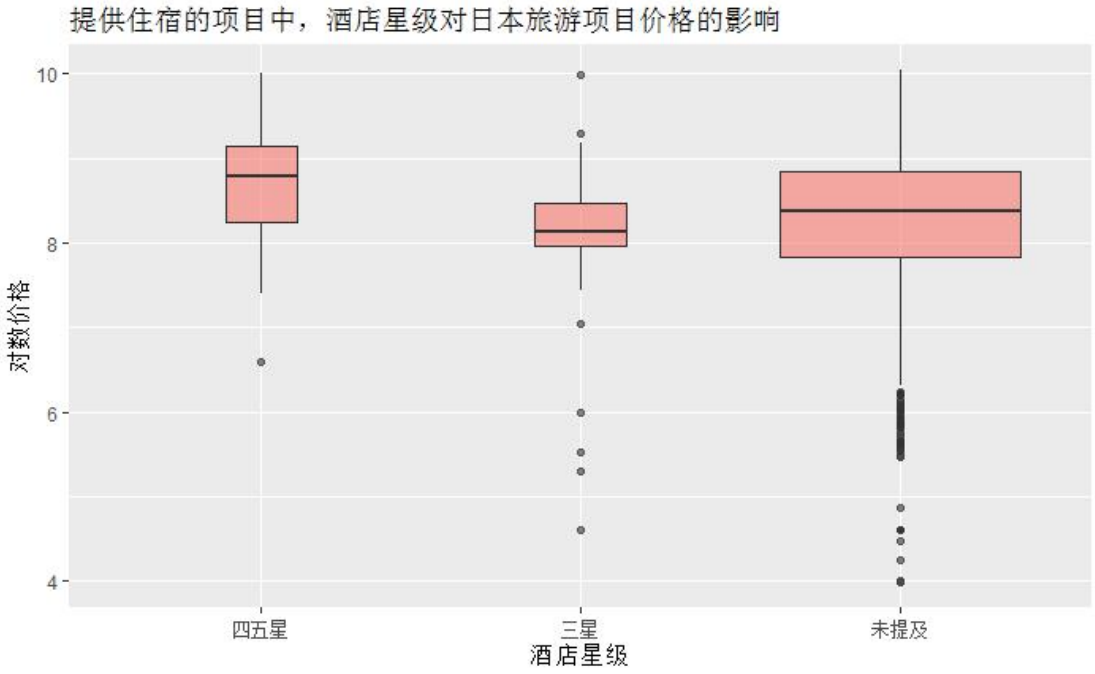

图 3.6.2　酒店星级对日本旅游项目对数价格影响箱线图

箱线图显示，靠近自然美景的酒店价格一般会贵一些，而对住处的风景有所要求的旅客一般会对旅程整体质量有较高要求，因而这种项目价格偏高。

同时可见住四五星酒店的项目价格较高，而住所是三星的项目价格低于未提及酒店星级的项目。推断为四五星酒店住宿的项目面向高端客户，因而项目整体服务价格偏高，而住三星酒店的客户会偏向于经济实惠的旅游项目。

3.7 出发地

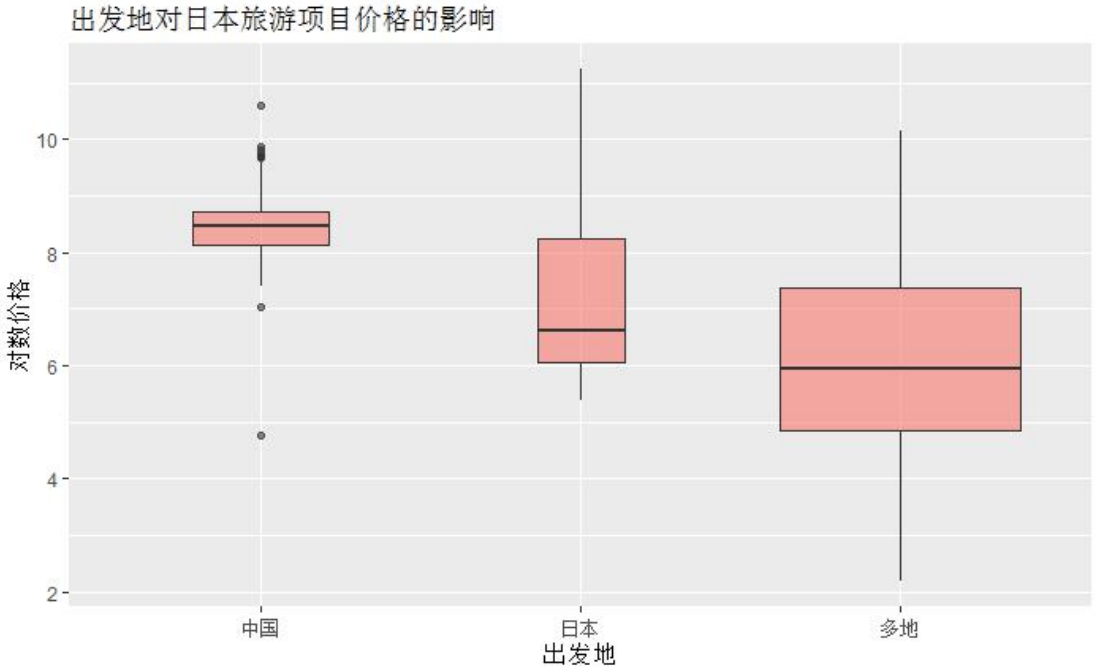

图 3.7　出发地对日本旅游项目对数价格影响箱线图

从中国出发的项目价格较高且方差小，因为从中国出发的项目包含了机票价格，且从中国出发项目长期偏多，所以从中国出发的项目价格明显要高。从日本出发的项目价格较低，因为国内飞机票会比跨国价格低，同时日本出发的项目方差较大，包含了很多 1 日游的项目，从日本出发的项目价格主要取决于项目本身的价格，方差大。“多地”水平的项目包括景点门票、娱乐公园门票、机票等，大多是小项目，总体价格低，且方差较大。

3.8 近购物中心

可见行程中有购物，或者住处接近购物中心的项目的价格偏高，而明显提及无购物的项目的价格偏低，方差大。推断想要选择购物的客户是对本次旅程有一定量的购物预算，因而项目普遍比较贵，而无购物的项目可能出现两种情况，1）不想被旅行团带去商场消费，尽量节省不必要的开支，这种项目的性价比比较高。2）项目的花费来源于购物外的体验，倾向于花更多的钱在旅程中，这种项目价格高。因而无购物的项目区别较大。

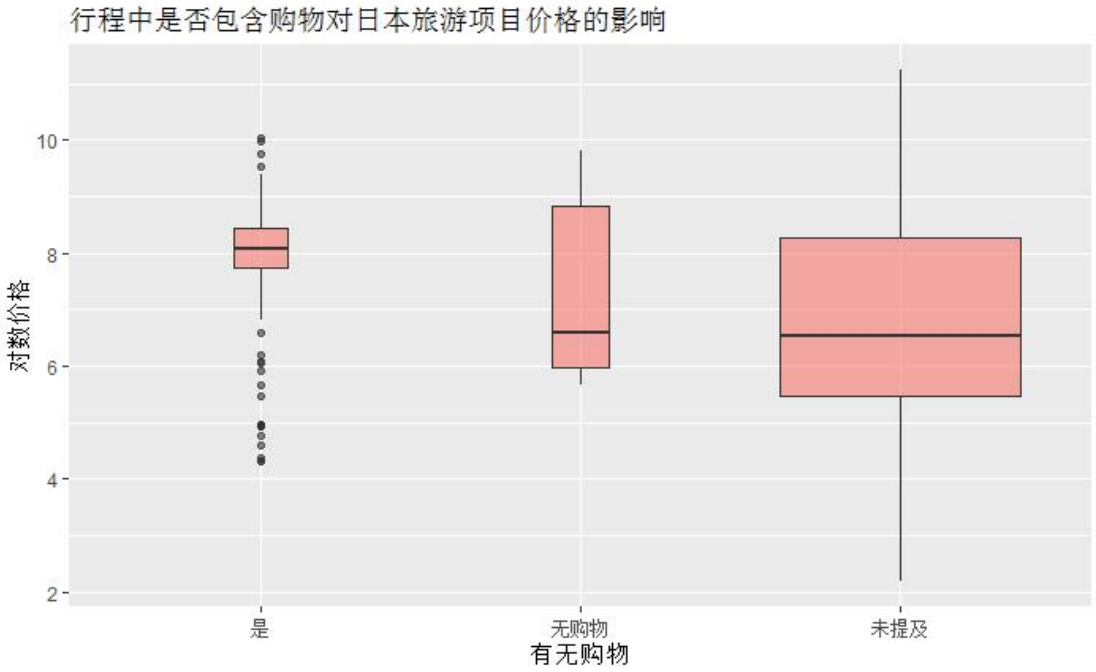

图 3.8　购物项目对日本旅游项目对数价格影响箱线图

3.9 有无 wifi

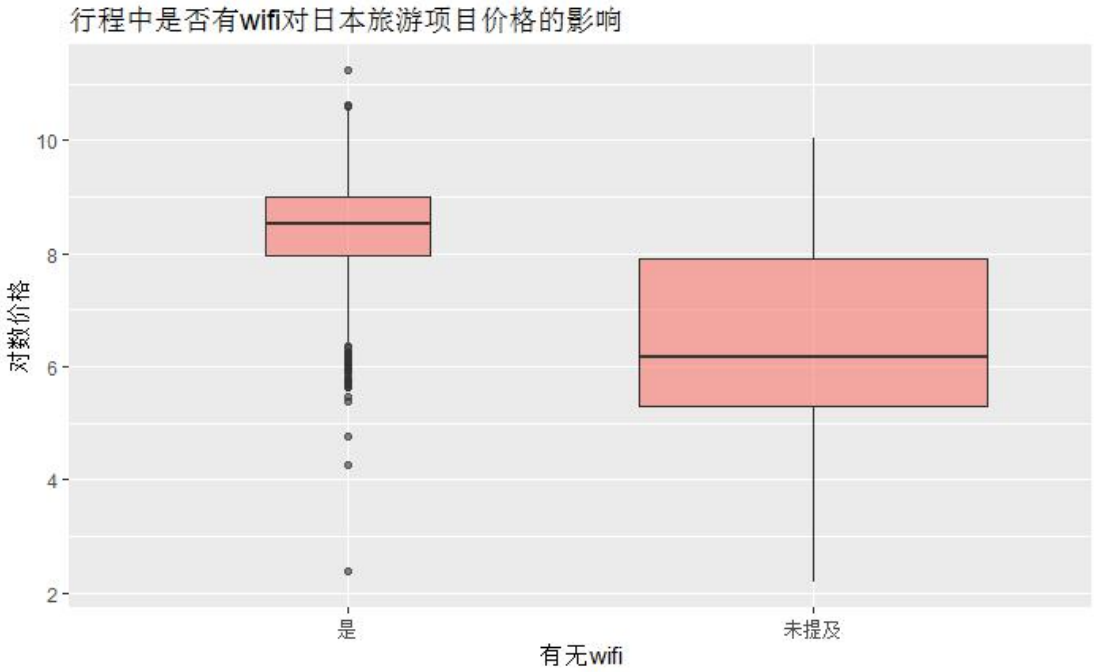

图 3.9　WIFI 对日本旅游项目对数价格影响箱线图

可见旅程中提供 wifi 的项目价格明显较高。在现代社会，无网寸步难行，旅客们对于 wifi 的需求较高，提供 wifi 的项目除需要包含 wifi 租赁费，也说明了这个项目是一个服务导

向的优质项目；这类项目也自然不包含娱乐园区的门票、景点门票等小项目，这些都使得提供 wifi 的项目价格偏高。

3.10 目的地

接下来根据分析各个旅游项目所经过的地点城市（包括温泉这个日本旅游特色）对于旅游项目价格的影响。

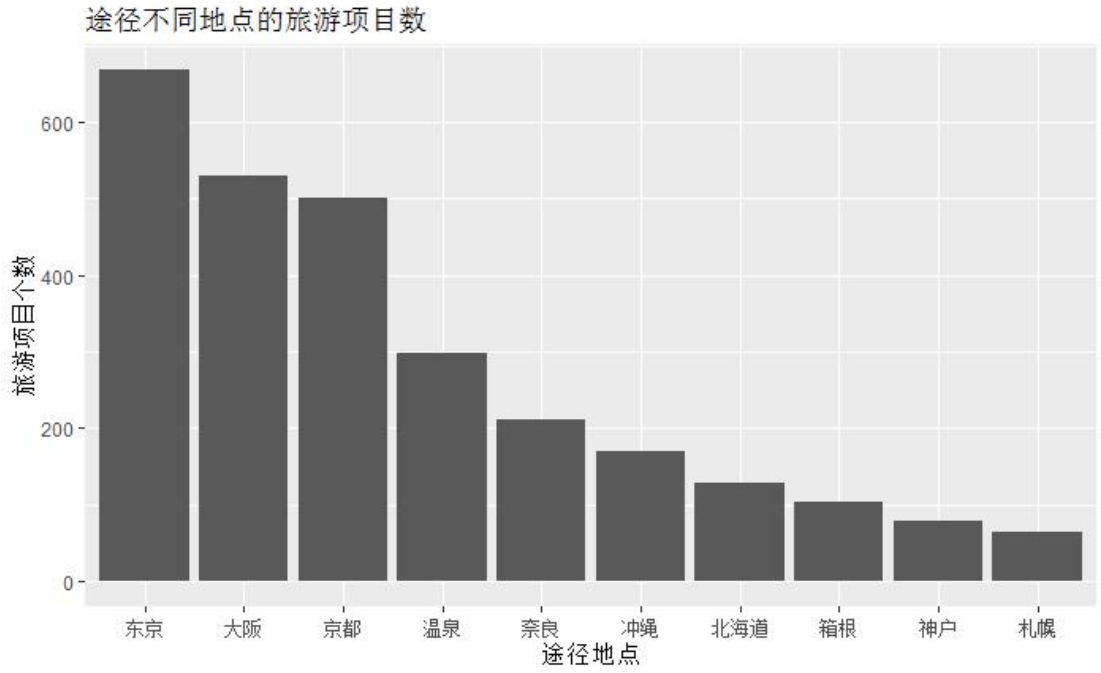

图 3.10.1——途径不同地点的旅游项目数柱状图

从柱状图可以看出，日本旅游项目里，最热门的三大地点是东京、京都和大阪，作为三个最具代表性的日本地区，经过这三个地方的旅游项目也多于其他地点。通过考虑“温泉”这个日本的特色旅游项目，可以发现温泉也是很受欢迎的旅游项目，而类似于北海道，札幌，冲绳这样的特点鲜明的地方旅游项目较少，一方面是因为知名度较低，另一方面是因为去这些地方旅游目的性更强，市场更小。

接下来我们分析各个目的地对于价格的影响：

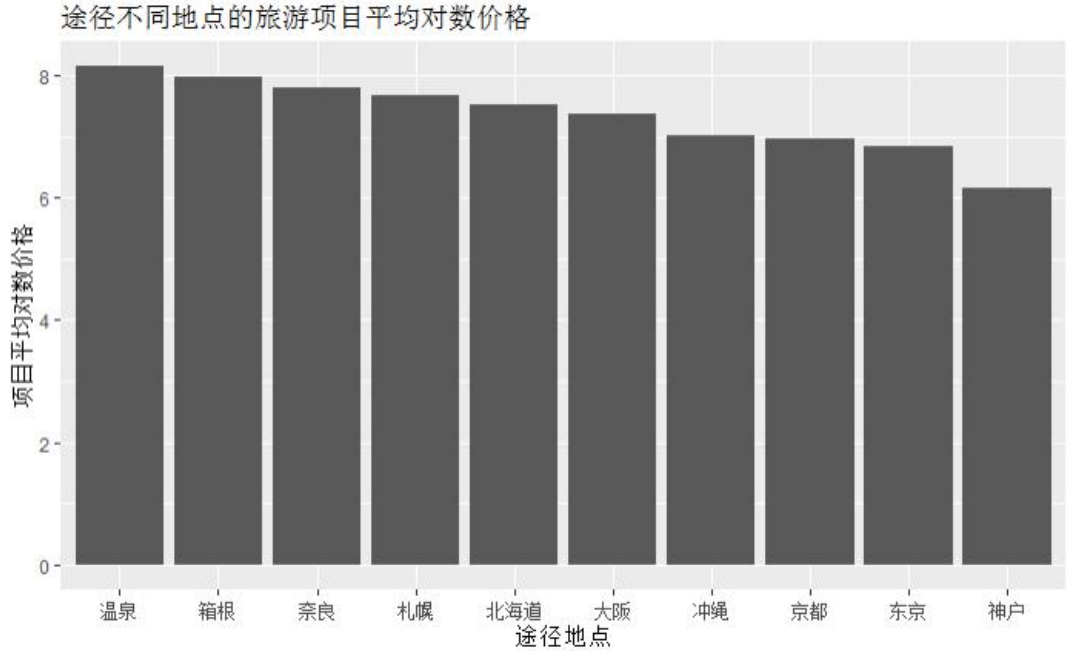

图 3.10.2——途径不同地点的旅游项目平均对数价格柱状图

需要注意的是，这里的目的地并不是一个分类变量，一个旅游的路线可能途径多个地点。观察柱状图我们可以发现，前往不同地方的旅游项目间价格差距不大，因为一个项目可能去到多个地方，比如东京和京都通常在同一线路上，所以价格也很接近，神户是这些地点里总体价格最低的，可能是因为神户因为牛肉出名，前往旅游的项目价格不高，主要为饮食消费。相对来说总体价格比较高的是箱根，可能是因为箱根主要的旅游活动的花费成本较高。而温泉作为一个可以附加在各条路线里的项目，可以看出温泉游的价格较高，甚至高于地点起到的作用，也可能是因为各个地点的均值受到一些“本地玩乐”的项目的影响，拉低了均值。

3.11 天数

一般而言，旅行时间越长相应的价格也越高，这里我们通过描述性分析的方法来探究这个特点：

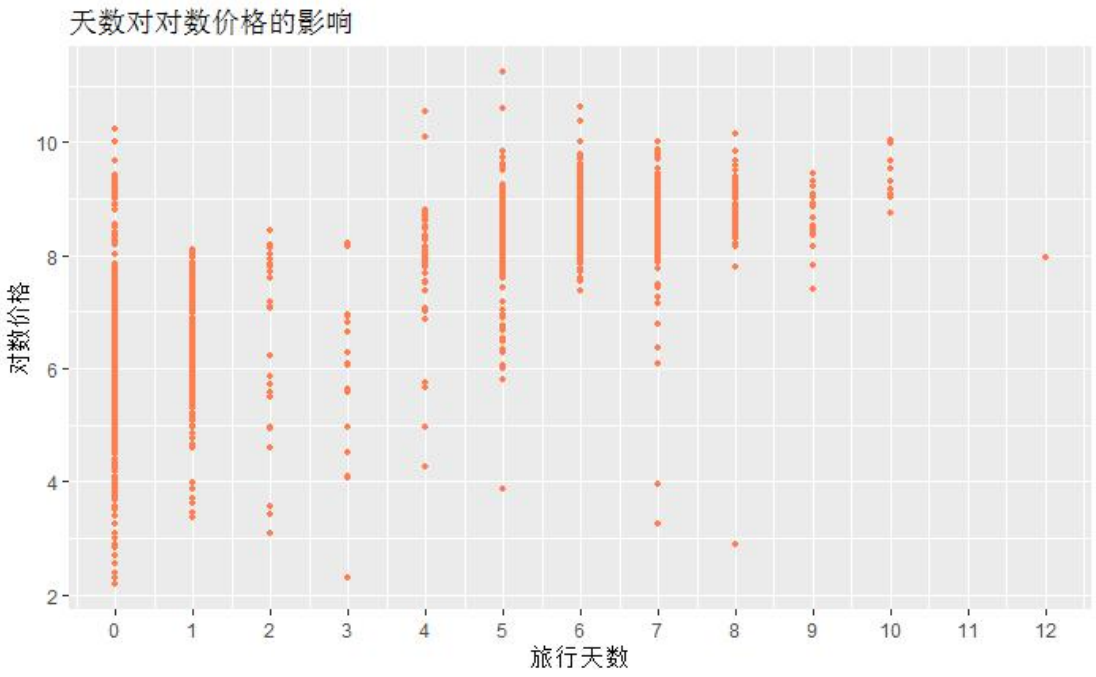

图 3.11——旅行天数对日本旅游对数价格散点图

需要说明的一点是，在这里旅行天数是 0 的旅游项目代表的是一些类似于“订座”，“门票”这样的项目。观察这些旅游项目，我们可以看出这些项目的价格更加分散，也符合我们的样本的分布特征，因为这些旅游项目种类繁多，有东京最高档的餐厅和一般的街边小吃与门票订座，所以差异很大；

观察其他样本，我们可以大致勾勒出对数价格随旅行天数增加而增加的趋势，符合我们的预期，也就是时间长的旅行项目成本（交通规划等等）增加，且同样天数的旅游项目内部差异，随天数增加而减小（分散度减小），说明越是长期的旅游项目其市场中的项目价格差异会越小，可能因为更模式化，弹性减小而具有相似的成本。同时我们发现，主要的旅游项目样本天数都分布在一周左右，说明一周是大家觉得比较合适的一个旅行时长，两三天和 9 天以上样本都比较少，因为 2，3 天在日本可能难以规划线路，而 9 天以上的旅行受到签证和假期的影响，市场较小。

3.12 点评数量

点评数能在一定程度上反应这个旅游项目给游客带来的印象是否深刻，感触很深的游客会更有意愿去评论，这里对点评数占出游人数比和价格的关系进行探究。

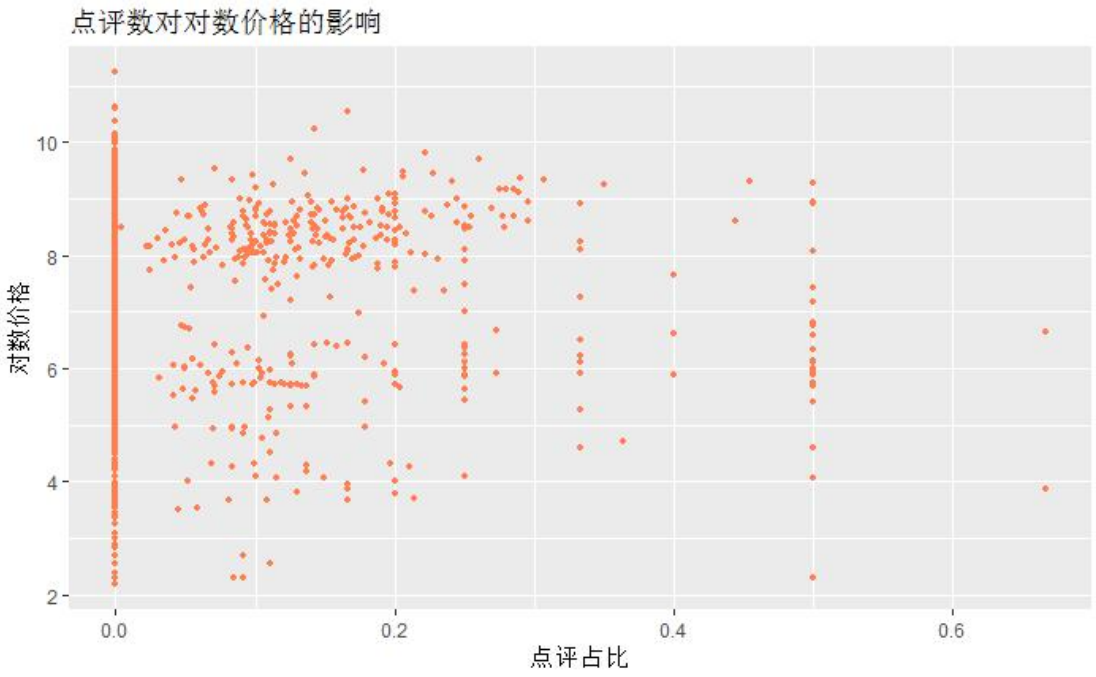

图 3.12——点评占比对日本旅游对数价格散点图

从散点图上可以看出，不少的旅游项目无人问津的（没有评论），大多数的旅游项目评论数占出游人数比都低于 0.3，说明大概不到三分之一的人愿意去为旅游项目写点评，说明这个平台的社区不够活跃。同时价格和点评占比并没有显著的关系。

3.13 人数

点评数能在一定程度上反应这个旅游项目给游客带来的印象是否深刻，感触很深的游客会更有意愿去评论，这里对点评数占出游人数比和价格的关系进行探究。

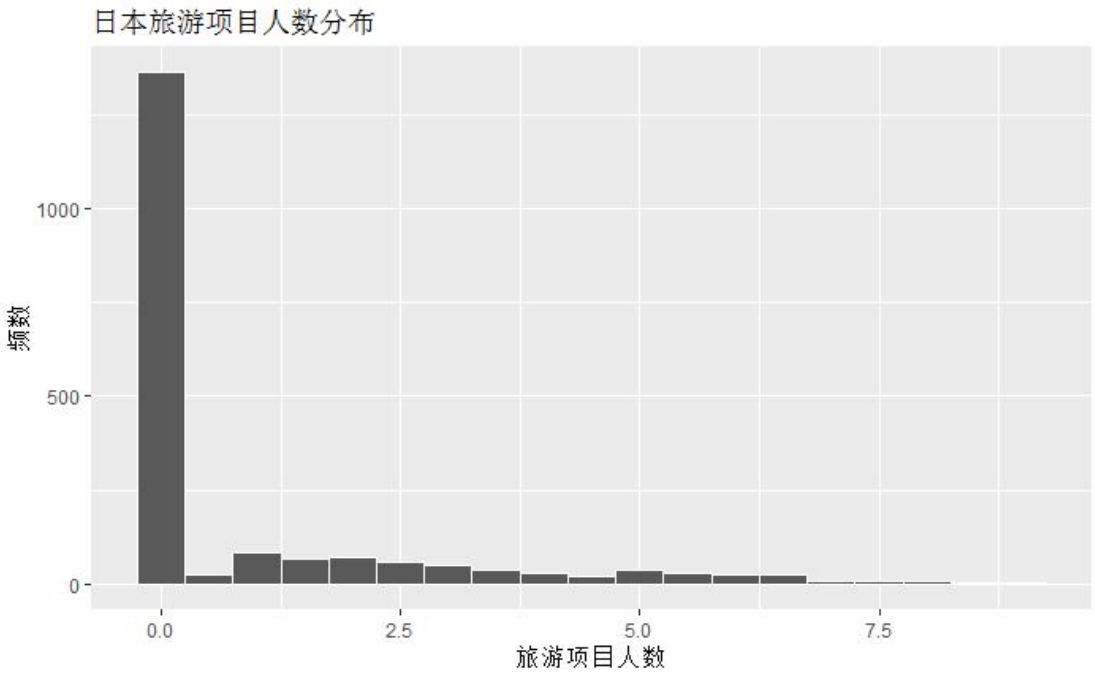

图 3.13.1——旅游项目人数散点图

百分之 70 的旅游项目出游人数是 0，剩下的有出游人数的旅游项目的人数分布非常的零散。出游人数最高的项目是从中国出发的日本东京-富士山-京都-大阪 6 日跟团游，出游人数 7798 人，满意度 0.97，价格 6004，在样本中算稍贵的项目。因而我们使用 log(出游人数+1)作为变量进行考察。

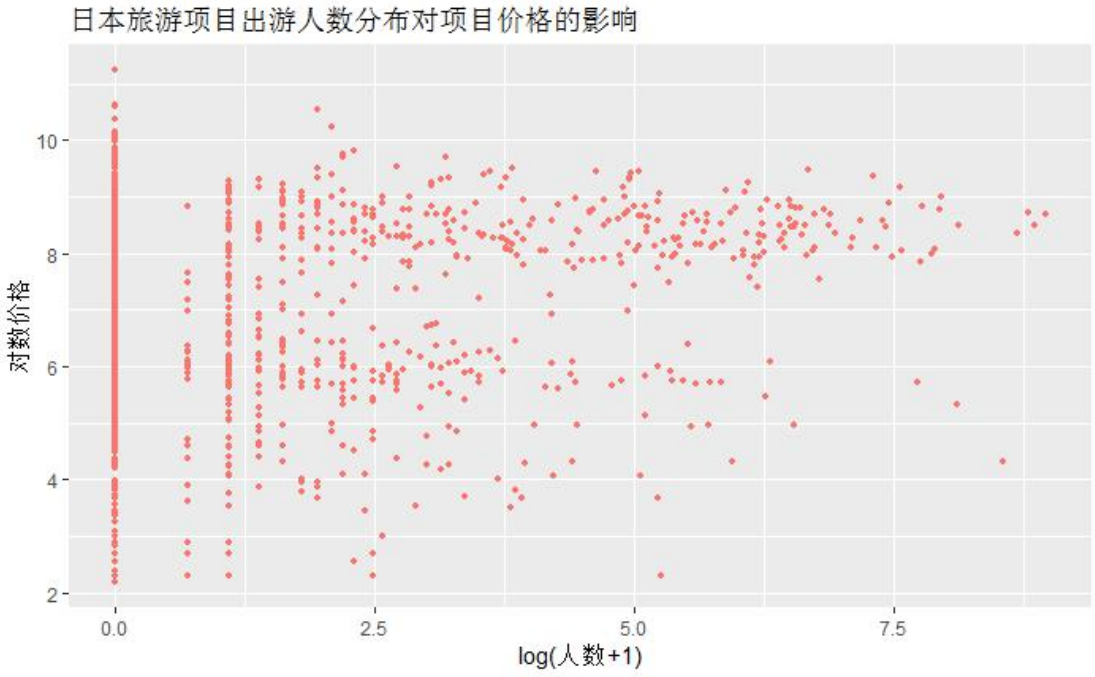

图 3.13.2——log(人数+1)对日本旅游对数价格散点图

可见人数为 0 的项目的价格是很零散无规律的，当 log(人数+1)大于 5，即出游人数大于 150 时，项目价格都是比较高的。可见吸引更多顾客的项目的价格集中于 3000 以上。但是总体的分布还是比较零散，说明出游人数和项目价格直接的影响并不大。

3.14 满意度

满意度是一个[0,1]取值的连续变量，表征客户对于旅游项目的满意程度。

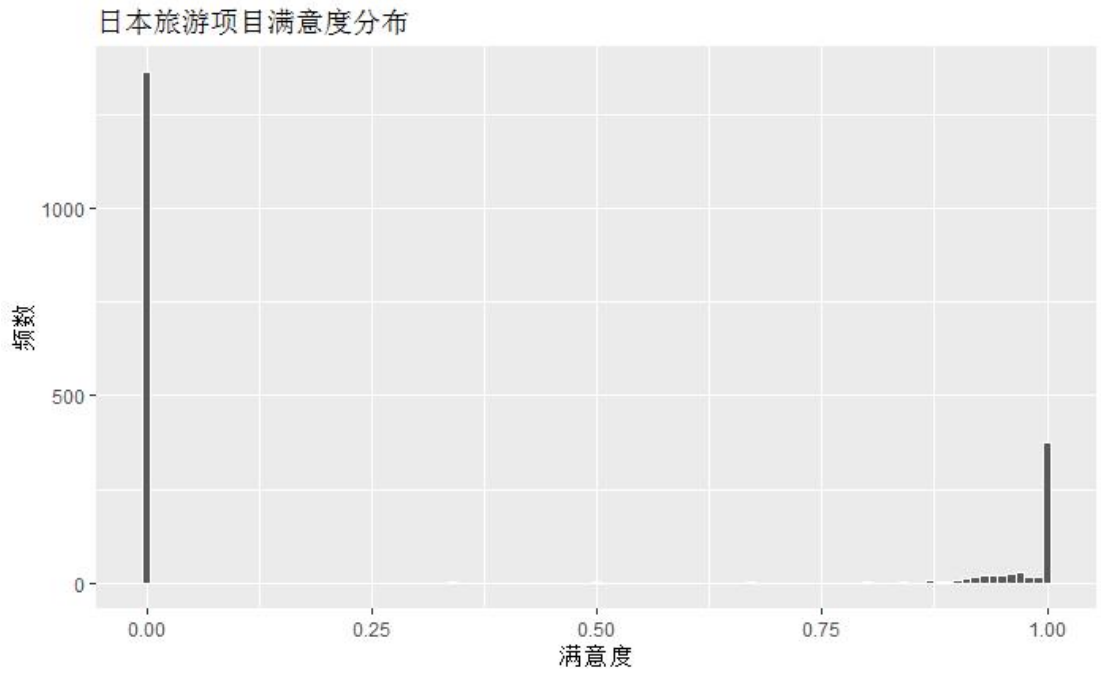

图 3.14.1——满意度频数分布直方图

我们发现 70%的项目的满意度均为 0，这个比例和出游人数为 0 的比例是一样的。认为无人问津的旅游项目的满意度默认为 0。在(0.00,0.85)区间几乎没有数据，在(0.85,1.00)区间有少量的满意度分布，还有很多项目满意度为 1，可能是很多旅客随手给了满分好评。

接下来我们考察满意度对项目价格的影响。

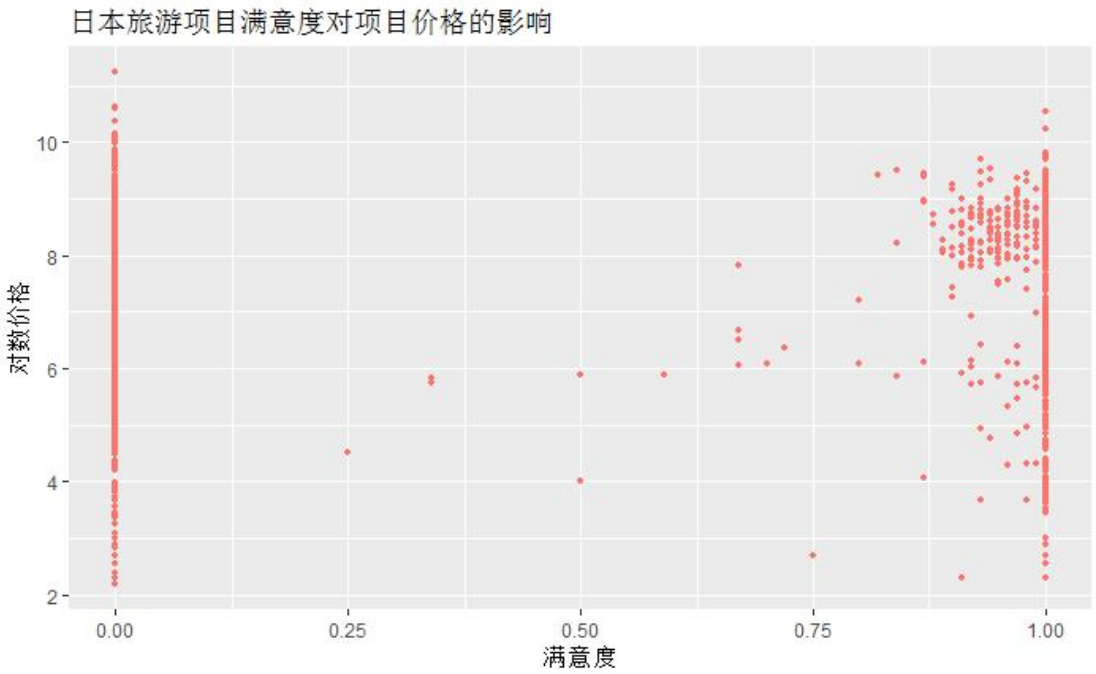

图 3.14.2——满意度频数分布直方图

我们发现满意度为 0 或者 1 的项目的价格分布均零散而无规律。首先，有人给出满意度评分说明这个项目的出游人数不为 0，其次满意度较高的项目的价格偏高。具体来说，满意度在(0.85, 1.00) 区间的项目价格都偏高，集中在 3000 元以上。这些大多数是服务优良、客户反馈好的优质旅游项目。

四、 模型拟合

在对数据集中各特征进行描述分析，并根据数据集对其进行分类后，我们可以对其使用不同的模型进行拟合，比较其优劣之后取得更好的拟合效果。

各模型拟合结果如下表：

模型	Adjusted R-square
线性回归	0.72
GBDT	0.25（MSE）
k-NN	0.52
CART	0.52
SVR	0.56
随机森林	0.66（MSE）

表 4--各模型拟合效果

4.1 线性回归

我们在划分训练集、测试集、去除异常点并逐步优化拟合方程后，得到的拟合结果如下：

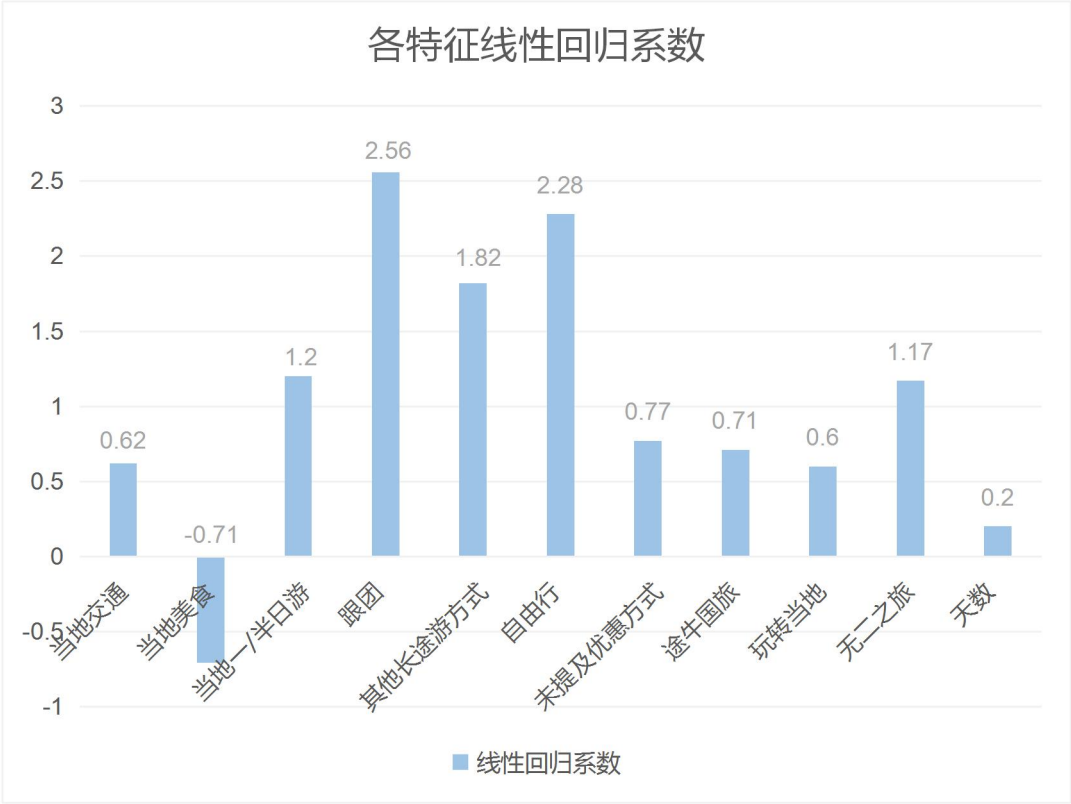

图 4.1—各特征线性回归系数（所示特征 p 值均小于 0.001）

从优化后的线性回归模型，可以看到出游类型、旅游项目、旅行社、天数对于价格的影响较大。其中在出游类型方面，交通项目与短期游对价格有提升作用，美食项目反之。旅游项目上，长途跟团对旅行价格提升最高，自由行次之，其他方式如目的地参团等最低，另一方面，大旅行社的旅游项目价格也相对较高，天数增加也对价格起到推动作用。而优惠方式、供应商、出发地等因素影响较小，而且满意度并未与价格呈正相关。同时该模型 MSE 的值为 1.144，Adjusted R-square 值为 0.72。

4.2 GBDT 模型

使用 GBDT 模型对数据集进行建模。得到各特征的相关性如下

特征	相关性
出游类型	68.01
天数	23.76

供应商	3.58

表 4.2 GBDT 模型各特征相关系数

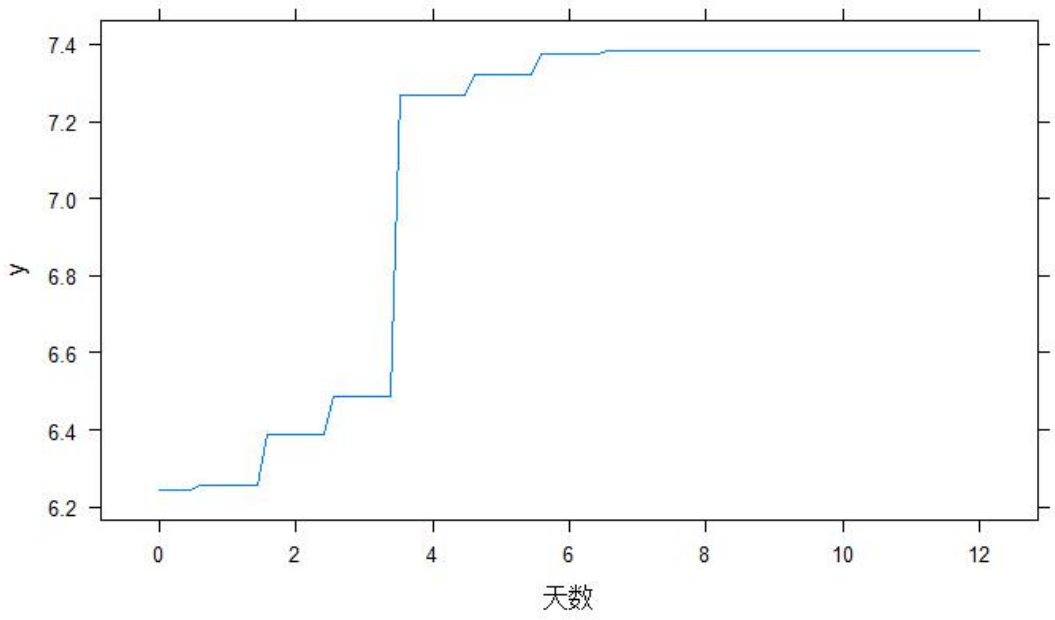

图 4.2 GBDT 模型天数边际效用图

从模型得到的重要性可以看出，对价格影响较大的是出游类型和天数，其中出游类型影响尤为明显，与线性回归模型得到的结果相互印证，其他因素较不重要。而天数中主要起到比较大边际效用的范围为 2-8。模型的测试集 MSE 为 0.252，这个结果表明 GBDT 模型在这个数据集的表现非常好，预测的数据值得信赖。

4.3 k-NN 模型

在使用 k-NN 模型对数据集进行拟合时，我们发现当 k 选择 9，距离选择曼哈顿距离时效果最好，但最好情况下 Adjusted R-square 值也仅能达到 0.52，建模结果并不理想。

4.4 CART 模型

我们使用 CART 模型对数据集进行拟合，同时计算数据集中各特征重要性，得到重要性图如下：

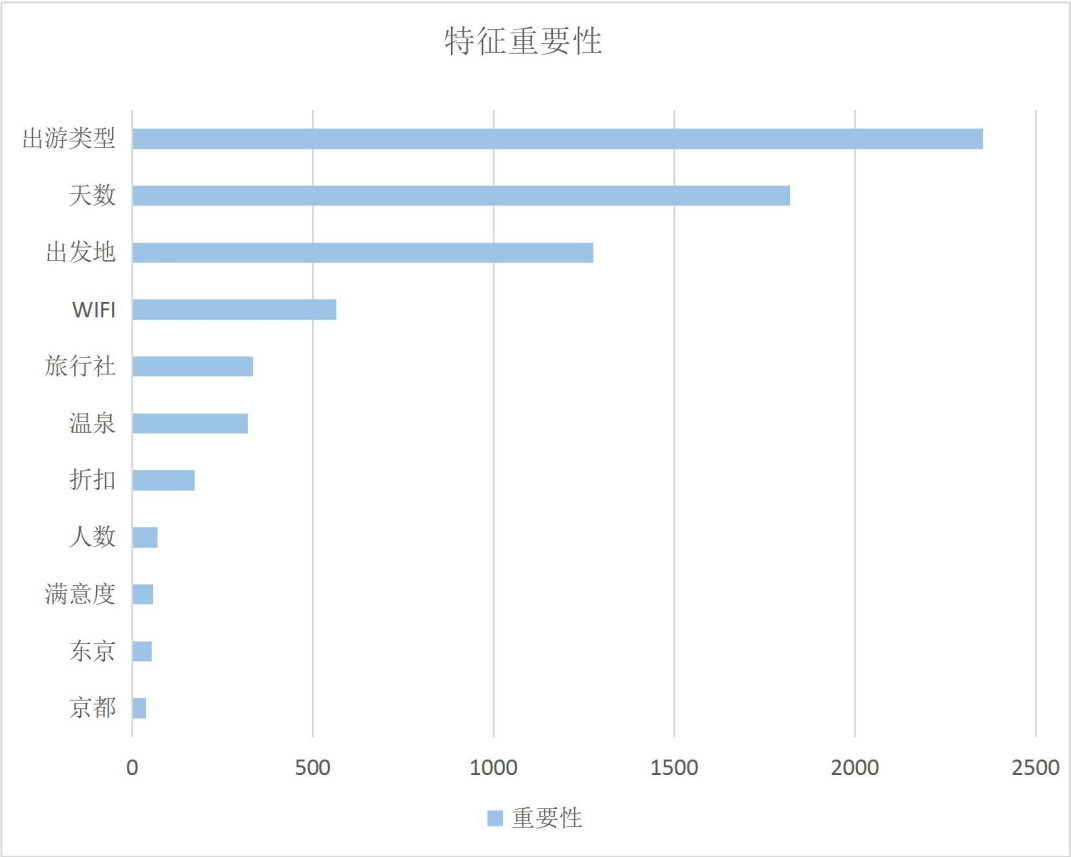

图 4.4 CART 模型中各特征重要性

从各特征重要性排序中我们可以得出与前述模型相似的结论，对旅行价格影响最大的事旅游的类型，如餐饮、医疗、娱乐之类，其次是旅行的长度，当然，出发地和 WIFI 也是重要的考量因素，旅行社的选择也比较重要。CART 模型建模得到的 Adjusted R-square 值为 0.52，建模效果一般。

4.5 SVR 模型

我们使用 SVM 模型衍生出的 SVR 模型对数据集进行拟合，针对残差平方和最小进行调参，最后得到的 Adjusted R-square 值为 0.56，建模效果一般。

4.6 随机森林

最后，我们使用随机森林进行建模，得到模型中各变量重要程度如下图：

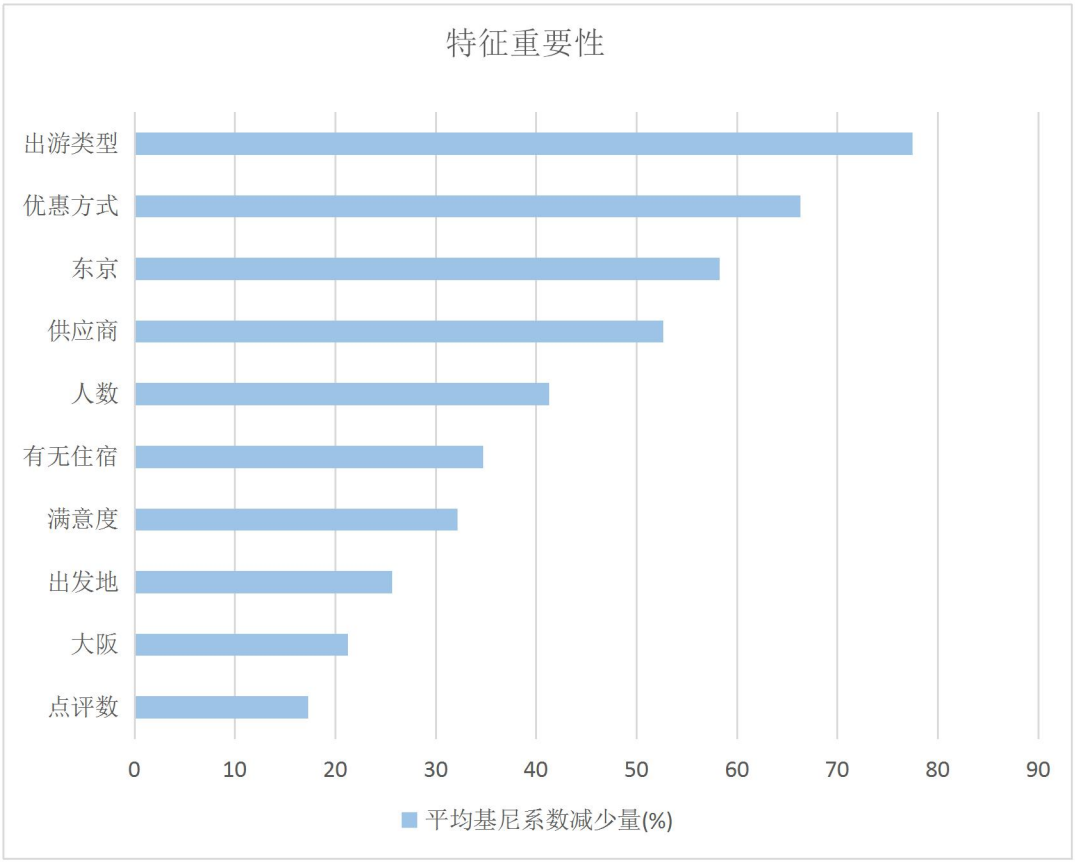

图 4.6 随机森林模型中各特征重要性

可以看到，随机森林中较为重要的变量为出游类型、优惠方式和目的地为东京的旅行团。与上文所示模型基本印证，测试集的 MSE 相较于 GDBT 模型稍微差一点，为 0.6647，也是相当不错的表现。

4.7 重要变量

综合几种分类模型中提取出的重要变量，如下表：

模型名称	重要变量 1	重要变量 2	重要变量 3
GBDT	出游类型	天数	供应商
CART	出游类型	天数	出发地
随机森林	出游类型	优惠方式	目的地东京

表 4.7 各模型中重要变量一览

五、 模型的改进方向

在对日本旅游数据集进行建模时，出现大量的评分数、出游人数、满意度均为 0 的数据，而在满意度有值得数据中，大量样本集中在 90%以上，体现了用户“随手评分”或者是不愿意留下评价的特点，使得我们构造的“满意人数”，“点评占比”两个新变量效果不佳。

六、 结论

从建模结果上看，简单的线性回归模型和 GBDT、随机森林模型取得了较好的拟合效果。在进行特征分裂时，出游类型、出发地、出游天数、优惠方式、供应商以及一些知名目的地能对日本旅游的价格产生较为显著的影响。

基于上述信息，我们对准备近日赴日旅游的游客朋友们提出以下建议：

1、明确自己赴日旅游的目的和目的地，比如赴日泡温泉、饮食、医疗等，选择相应的旅游计划，避免无谓的消费。合理安排日程，争取在尽量短的时间内完成旅行计划，劳逸结合、合理出行。

2、如果具有相应的语言能力和计划能力，推荐目的地参团和自驾游而非跟团游的模式，既可以省钱还可以拥有一个自由的日程。

3、尽管大旅行社的旅行团价格会比较高，但基于跨国旅行和安全性的考虑，还是推荐选择品牌好，品质佳的老牌旅行社。

“日暮腾云起, 东瀛万里行。”希望每一位旅客都能拥有美好的日本旅行体验！

七、 小组分工

李泽君、赵雅滢：数据预处理、描述分析

谢炳辉、王维实：建模，建模作图

钟诚：撰写报告

复旦大学第三交通委小组

2019.6.3