那些年我们去过的日本

——日本旅游价格分析

第四组
林耿 王众 张雨晴 陆静颐

一、背景介绍

随着我国经济的不断发展，我国人民的生活水平有了一个很大的提升，我国人民在追求物质生活的基础上，对于精神生活也有了更高的要求。而旅游项目就是一个非常好的选择，因为在旅途过程中我们可以感受更多，学到很多。在国内外众多旅游景点中，日本的旅游热度一直居高不下，并且近年来呈现出不断上升的趋势。

赴日游客为何连年增长？他们一般都去哪儿玩，怎么玩，体验怎么样？据统计，日本作为中国公民出国旅游的第二大目的地国家，仅 2018 上半年中国大陆赴日游客就已经突破 400 万人次。数据显示，2018 年 1~9 月前往日本旅游的中国大陆游客人次达 644.5 万，同比去年增长近 16%，中国继续蝉联赴日旅游第一大客源国。

在收入增长和旅游消费升级推动，以及签证、航班等便利因素影响下，赴日旅游的人数持续增长。今年以来，赴日航班越来越密集，包括从二三线城市出发到日本京都大阪地区、一线城市出发到日本名古屋、旭川、静冈、四国等小众目的地的航班，以及中日联航，包括春秋、乐桃，捷星等航司的入场让中日航班变得越来越便捷、便宜。近年来，日本对中国游客推出了多次旅游签证，并放宽了多次签证的条件。

在线旅游平台为赴日游客带来更大的便捷。消费者可以通过网络实现自由行、跟团游、定制旅行、目的地参团、度假酒店等产品购买，当地一日游和当地门票玩乐产品预订，获取食住行游购娱导一站式服务。

中国游客赴日旅游，哪些目的地最火？中国哪些城市赴日旅游积极性最高？我们根据 1949 条日本旅游产品的数据，试图分析影响各大平台上众多日本游产品价格的因素。

二、数据处理

1. 原始变量

变量类型	变量名	详细说明	取值范围
因变量	价格	连续变量	[9, 75599]
自变量	名称	文本数据	例：日本东京-大阪 6晚7日 自由行
	出游类型	定性变量(17水平)	自由行，跟团，定制游，度假酒店等
	优惠方式	定性变量(34水平)	立减优惠，餐厅优惠券，交通卡/券等
	有无住宿	定性变量(2水平)	有/无
	住宿条件	定性变量(88水平)	wifi，住近市区，住近购物中心等
	出发地	定性变量(89水平)	西安，北京，成都等
	供应商	定性变量(146水平)	途牛国旅，斑马旅游，牛人专线等
	满意度	连续变量	[0, 100%]
	出游人数	连续变量	[0, 7798]
	点评数	连续变量	[0,1992]

2. 衍生变量

从原始变量到衍生变量的处理：

变量类型		变量名	详细说明	取值范围
因变量		价格	连续变量	[9, 75599]
自变量	内部因素	天数(day)	定性变量(12水平)	0~10之间的整数，0.5
		夜晚(night)	定性变量(12水平)	0~9之间的整数，3~4，4~5
		城市	定性变量(11水平)	东京，大阪，京都，奈良等
		出游类型	定性变量(7水平)	自由行，跟团，定制游，半自助等
		有无住宿	定性变量(2水平)	有/无
		住宿条件	定性变量(9水平)	餐厅，湖光山色，免费WIFI等
	外部因素	优惠方式其它	定性变量(2水平)	是/否
		出发地	定性变量(3水平)	国内出发，日本出发，其它
		供应商	定性变量(5水平)	途牛国旅，无二之旅，中国国旅苏州旗舰店，玩转当地，其它
		满意度	连续变量	[0, 100%]
		出游人数	连续变量	[0, 7798]
		点评数	连续变量	[0,1992]

① 从名称中提取日、夜、城市信息，三个均为定性变量。

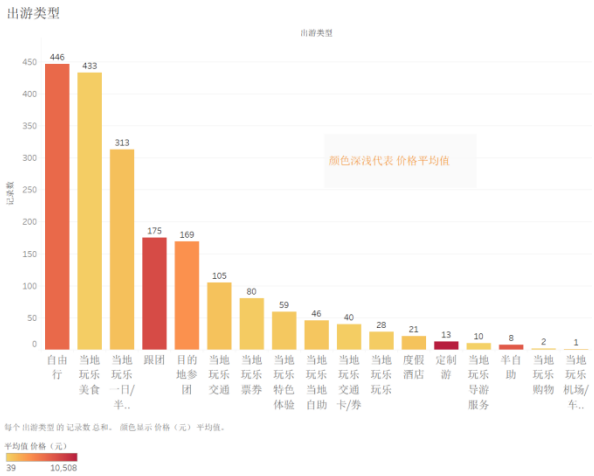
② 出游类型、住宿条件、供应商分别选取频率最高的 7、9、5 项。

③ 将出发地信息划分为国内、日本与其它三大类。

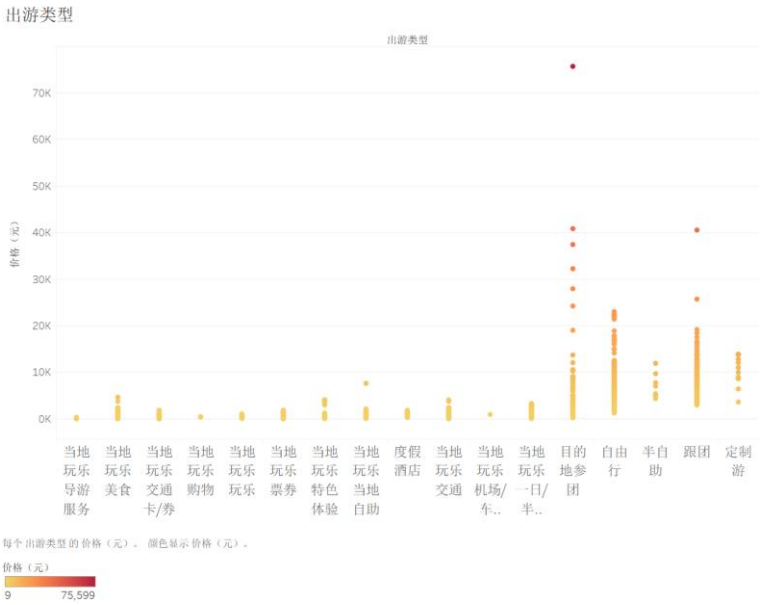
由于部分旅游产品为当地玩乐、门票、餐厅预约等项目，而我们重点对完整旅游项目进行分析，因此在进行分析时，我们暂时删去了当地玩乐、门票、餐厅预约等项目，而留下"自由行","目的地参团","跟团","定制游","半自助"等项目，对该筛选后数据集进行分析。

三、描述分析

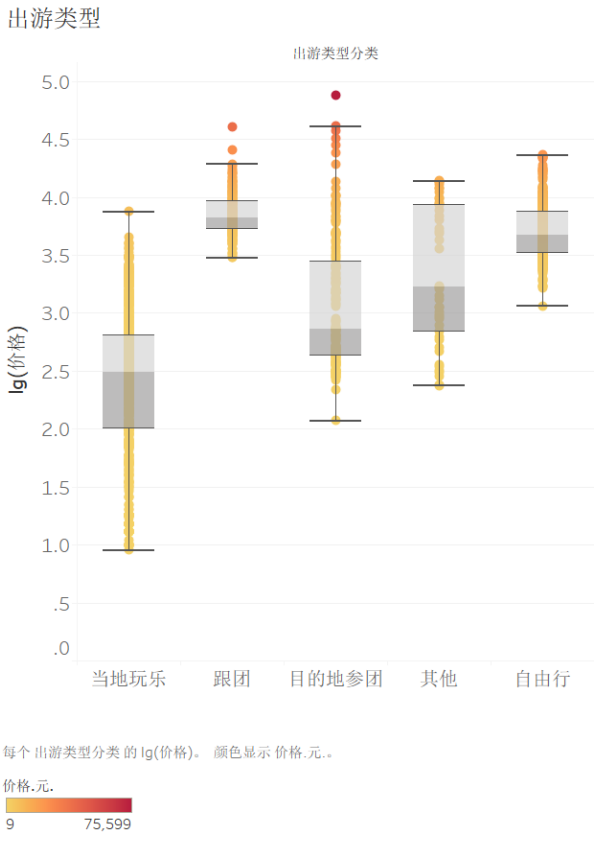
1. 出游类型



从上图中的柱状图高度我们可以看出，选择自由行的人数比例最高，跟团与目的地参团的人数较少。从柱状图颜色反映出的价格可看出：定制游相比之下是最贵的，跟团其次，自由行相对而言较便宜。



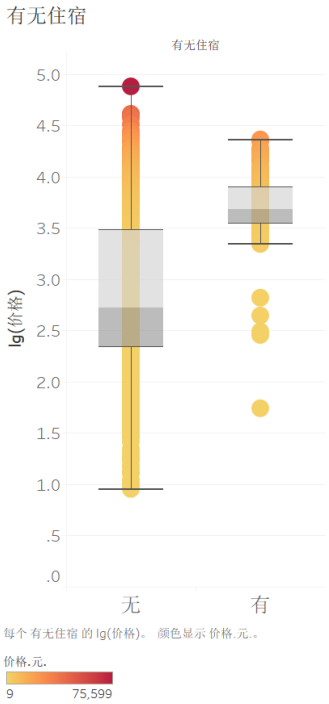
从出行方式——价格的关系图中我们可以看出，当地玩乐系列的价格和其他出游类型相比低很多。因此我们把当地玩乐美食、购物等等变量合在一起，并画出 出游类型——lg(价格）的箱线图。



2. 优惠方式

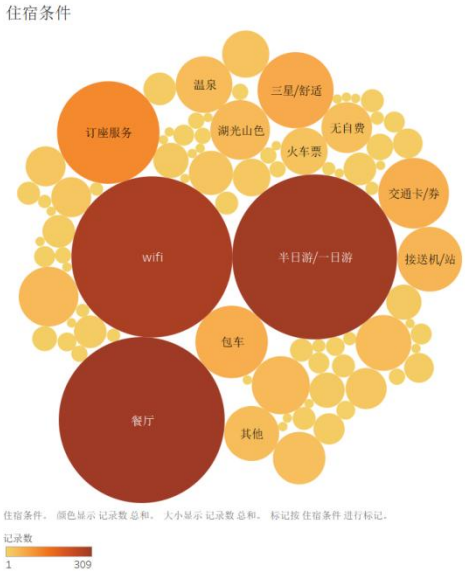
因为绝大多数优惠方式都为“立减优惠”，我们把优惠方式 分为“立减优惠”及“其他优惠方式”。

3. 有无住宿

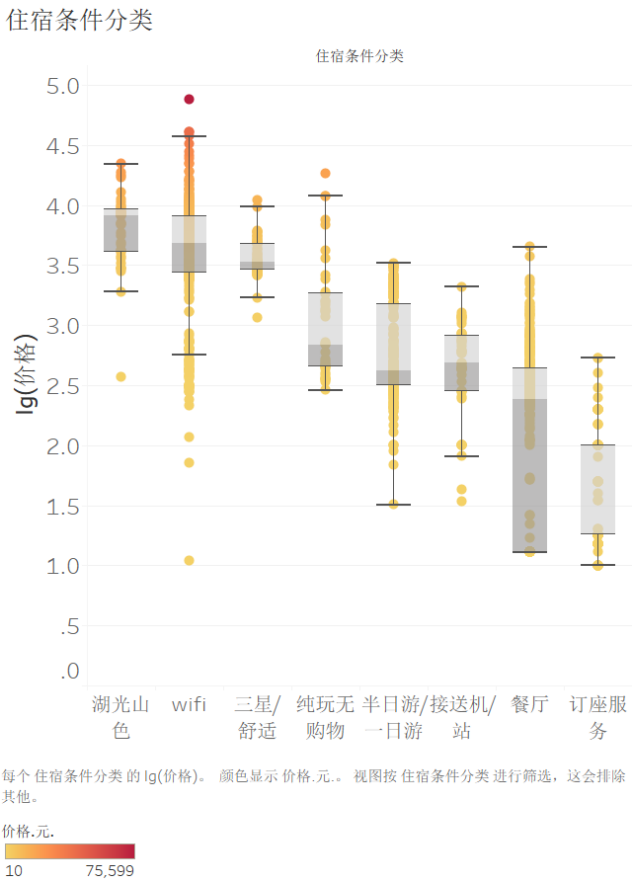


画出“有无住宿”——“lg（价格）”的箱线图，可以看出有住宿的价格平均会比无住宿的高，且价格较为集中，这也符合我们的常规认识。

4. 住宿条件



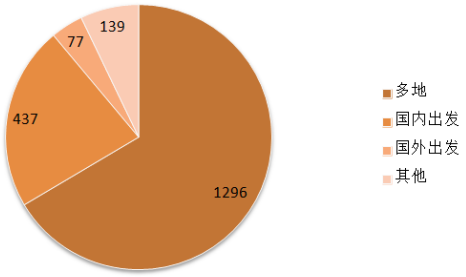
住宿条件中，有“wifi””、“餐厅”、“可以订座服务”占了很高的比例。

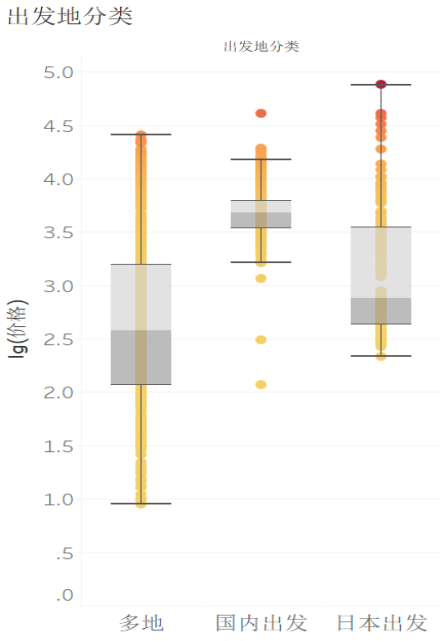


我们提取了一些数目较多的变量，画出住宿条件与“lg（价格）”的箱线图。可以看到可以欣赏“湖光山色”这一住宿条件对应的价格较高。

5. 出发地

出发地主要分为“国内出发”、“日本出发”、“多地出发”三种。

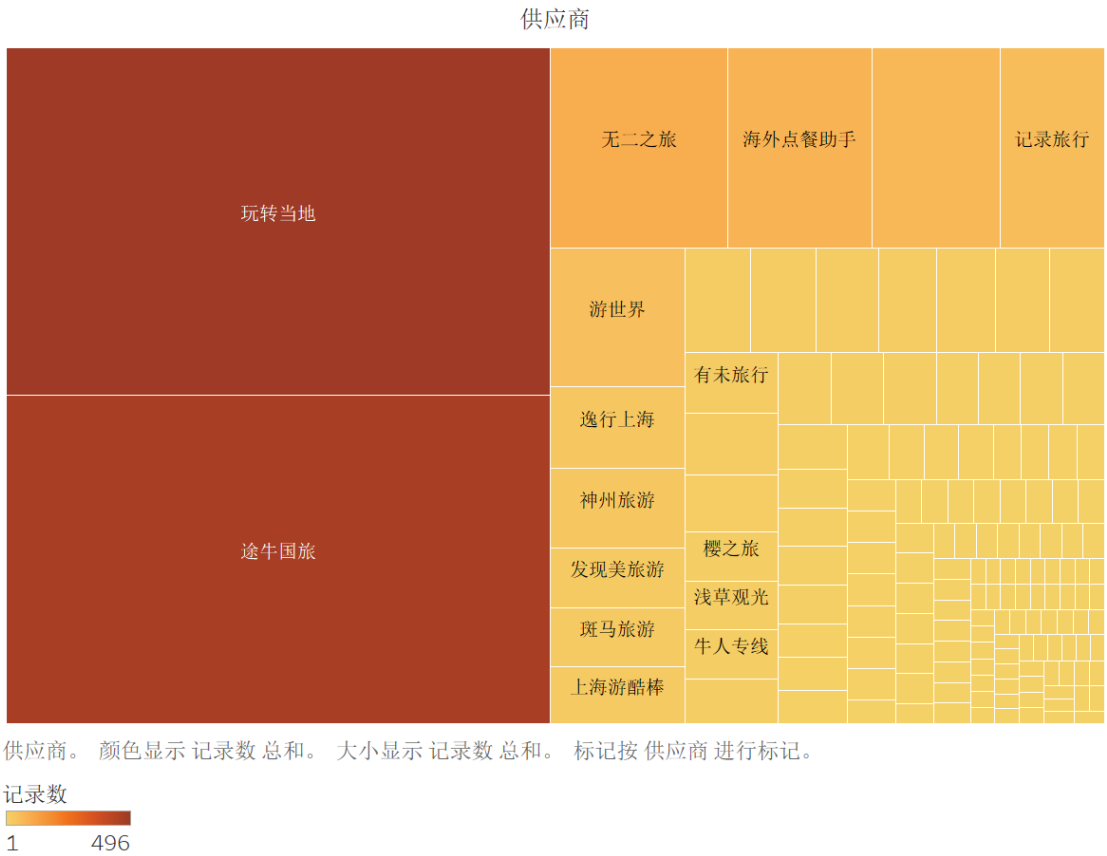




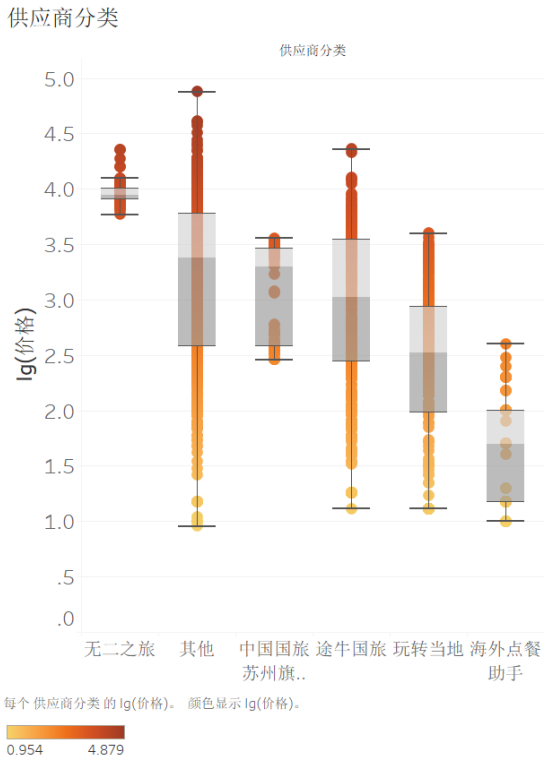
从中国地图可看出，在我们的数据中，从西安、北京、上海出发的人数较多。

箱线图可以看出从国内出发会比从日本出发的价格更高，这也是因为距离更远。

6. 供应商

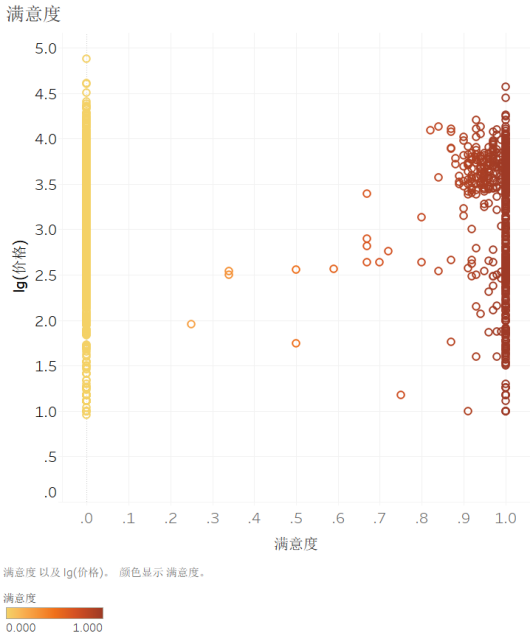


从图中可看出，选择“玩转当地”和“途牛国旅”的人几乎占了全部人数的半壁江山。



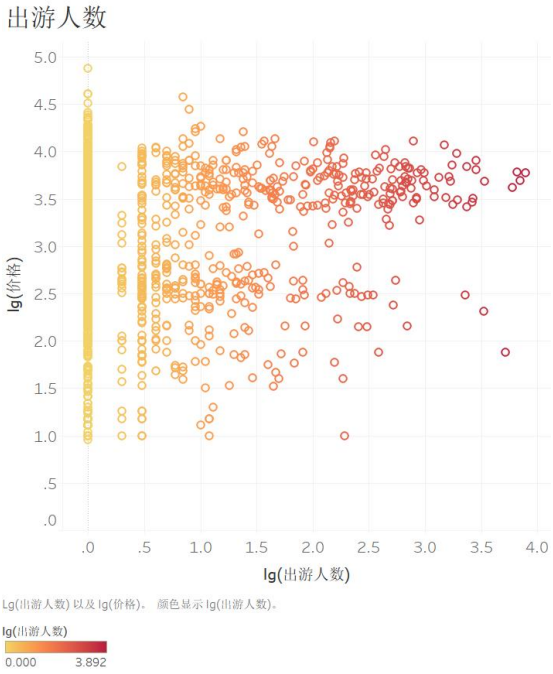
我们保留变量里出现次数较多的供应商，画出箱线图。我们可以看出“无二之旅”的平均价格相对高很多，而“途牛国旅”会比“玩转当地”平均价格高。

7. 满意度



我们看到满意度主要呈现两极分化的状态。通过“满意度”-“lg（价格）”的散点图，在满意度不为 0 时，我们还是可以看出“满意度越高，价格越高”的趋势。

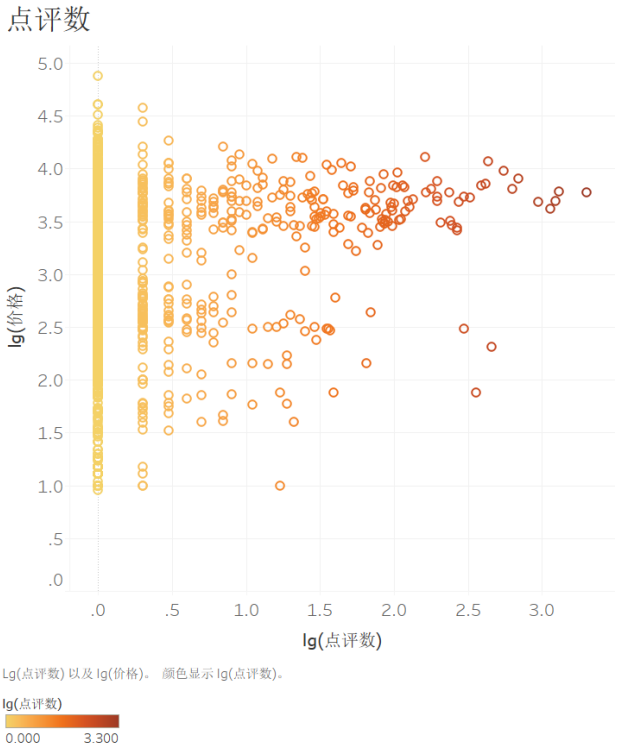
8. 出游人数



通过“lg(出游人数)”-“lg（价格）”的散点图，我们看到大部分的项目出行人数很少，集中于 0-1000。

总体可看出出游人数越多，价格越高的趋势。

9. 点评数



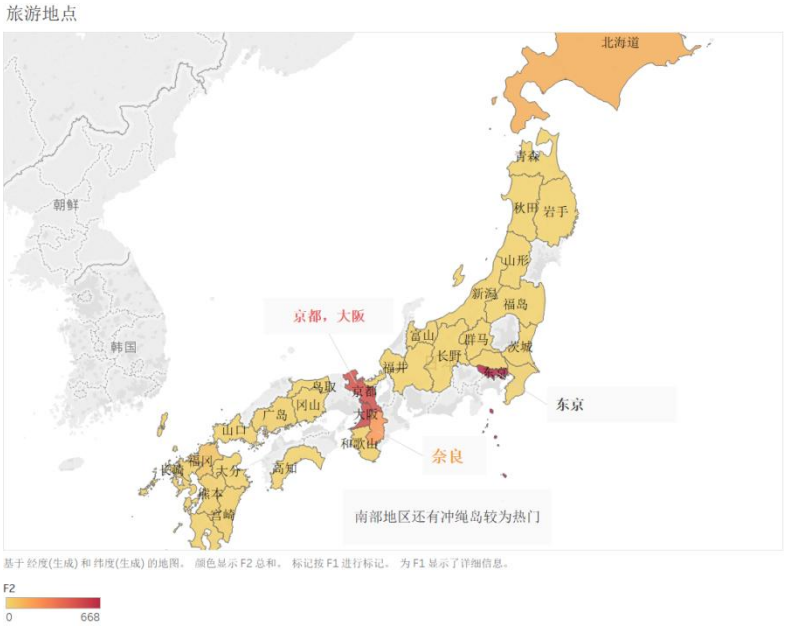
通过“lg(点评数)”-“lg（价格)”的散点图，我们看到大部分的项目点评数很少，集中于 0-1000。

总体可看出点评数越多，价格越高的趋势。

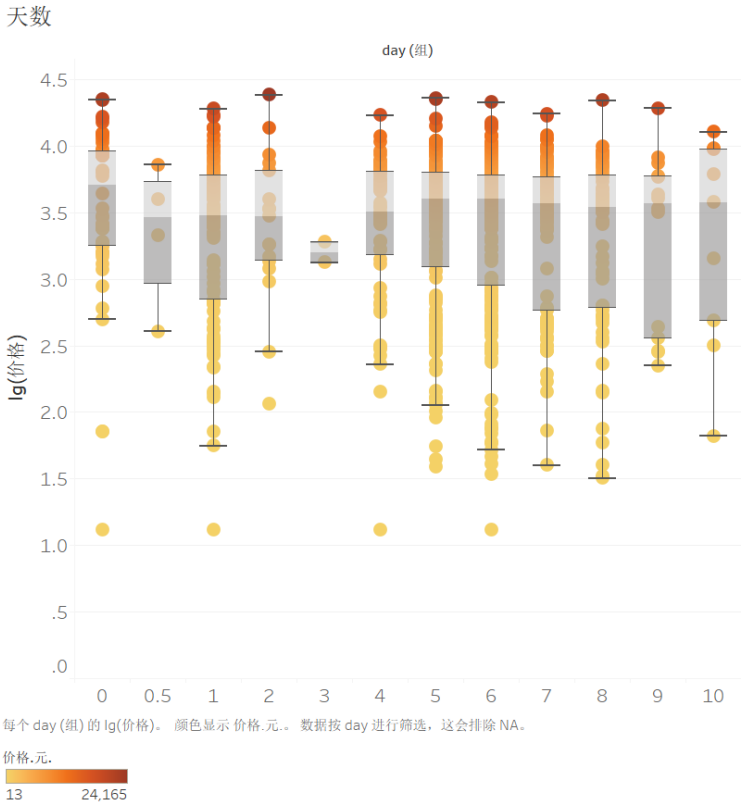
10. 旅游地点

通过提取“名称”中的日本地名，我们画出地图。

可以看出：旅游地点可分为东南西北四部分。其中，关东的东京与关西的大阪、京都最为热门。南北的冲绒与北海道也是很多人的选择。推测热门地点的旅游价格也较高。



11. 天数



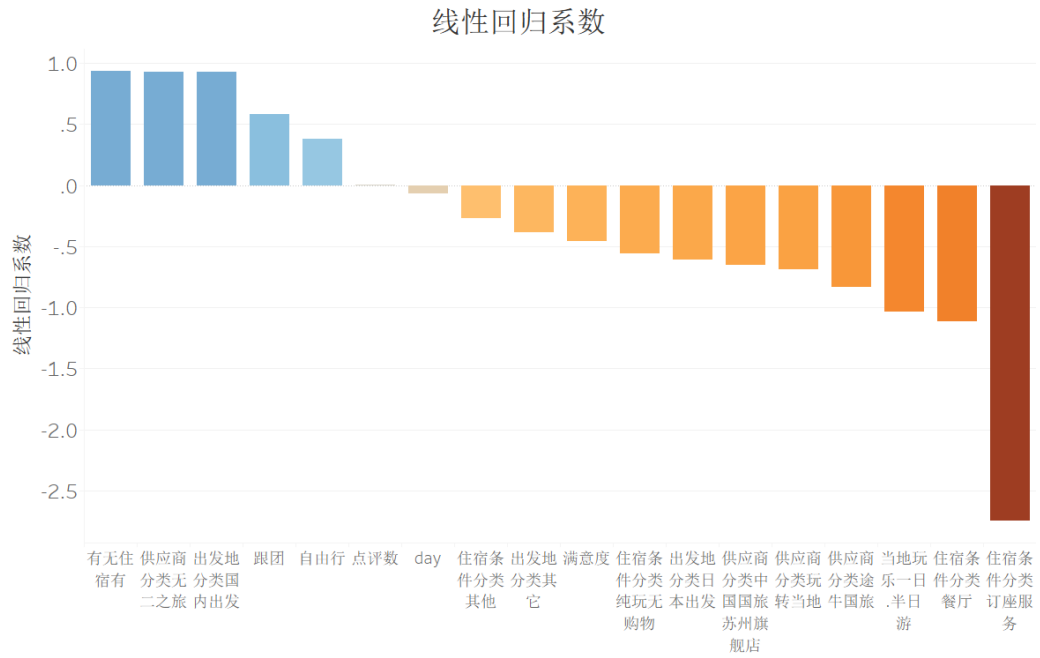
通过提取“名称”中的“几天几夜”，我们画出天数与价格的箱线图。但是图表中并没有反映出天数与价格之间的正向关系，关联性不大。

四、模型构建

1. 线性回归模型

在线性回归模型中，我们使用价格作为因变量，其余变量作为自变量，得到了如下所示的线性回归模型。

变量名		系数
出游人数		-4.31E-04
点评数		2.32E-03
天		-6.22E-02
有无住宿有		9.07E-01
住宿条件	餐厅	-1.08E+00
	纯玩无购物	-4.90E-01
	订座服务	-2.63E+00
	其他	-2.53E-01
出发地	国内出发	1.03E+00
	日本出发	-5.24E-01
供应商	途牛国旅	-9.23E-01
	玩转当地	-5.79E-01
	无二之旅	1.05E+00
	中国国旅苏州旗舰店	-6.43E-01



① 在经过线性回归变量筛选后，我们可以发现，在与价格相关的因素中，点评数、有住宿对价格有正影响，满意度、旅游天数对价格有负影响。不难理解，点评数高的旅游产品往往关注度和曝光度都很高，因此定价也会相应高；而有住宿的旅游产品包含酒店的价格，会相对无住宿的旅游产品更昂贵。

② 令人意外的是，满意度越高、旅游天数越多的旅游产品价格竟然会相对更低，因此我们可以得出结论，人们现在更关注旅游产品的性价比，即人们在评判某旅游产品是否符合心理预期时会更倾向于考虑其价格，因此在这些旅游产品中会存在许多通过压低价格来提高口碑及销量的部分；另外，旅游天数也成为了供应商的一大噱头，由于人们在选择时更倾向于选择价格更低、时间更久的旅游项目，旅行社也通过人们的这一心理推出更多旅游时间很长但是服务项目数量较少、质量较低的项目。

③ 从回归结果中我们发现，出发地为国内出发的旅游产品价格相比日本出发以及其他形式出发的旅游产品会更高；无二之旅出品的旅游产品价格相比途牛国旅、玩转当地、中国国旅苏州旗舰店价格更高，而其中途牛国旅出品的旅游产品价格最低；类型为跟团的旅游产品的价格倾向于最高，紧接着依次为自由行和当地玩乐一日/半日游。

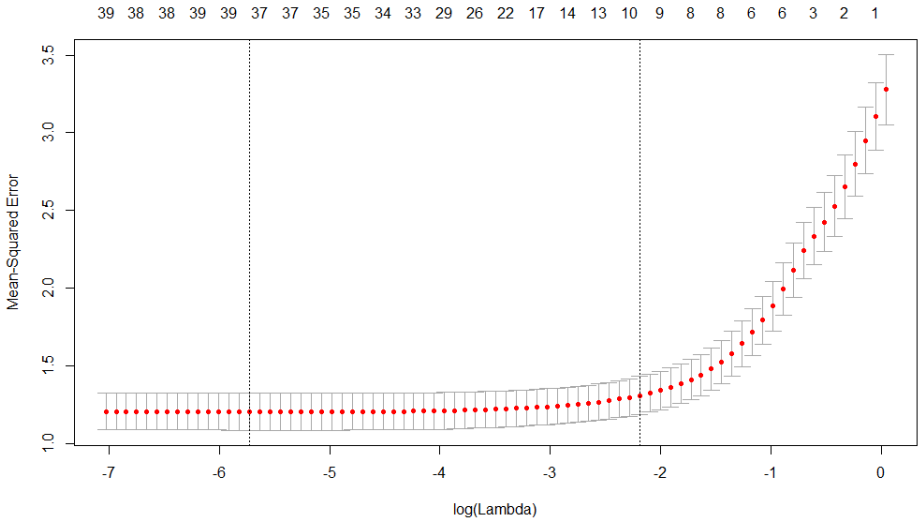
④ 另外，我们发现，模型中所有城市变量都在筛选变量的过程中被删除了，我们通过观察各个旅游产品的具体描述发现，大部分日本旅游产品不局限于一个城市，而涉及到很多城市，因此在模型构建时很难通过产品覆盖城市判断价格的高低。

2. Lasso 模型

不同于线性回归模型，lasso 模型要最小化的式子加入了惩罚项，如下所示。Lasso 模型可以解决原先与最小二乘法等价的线性回归中可能面临的过拟合问题，lasso 模型的特殊惩罚项可以帮助筛选变量。

$$\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij})^2+\lambda\sum_{j=1}^{p}|\beta_j|=\mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|$$

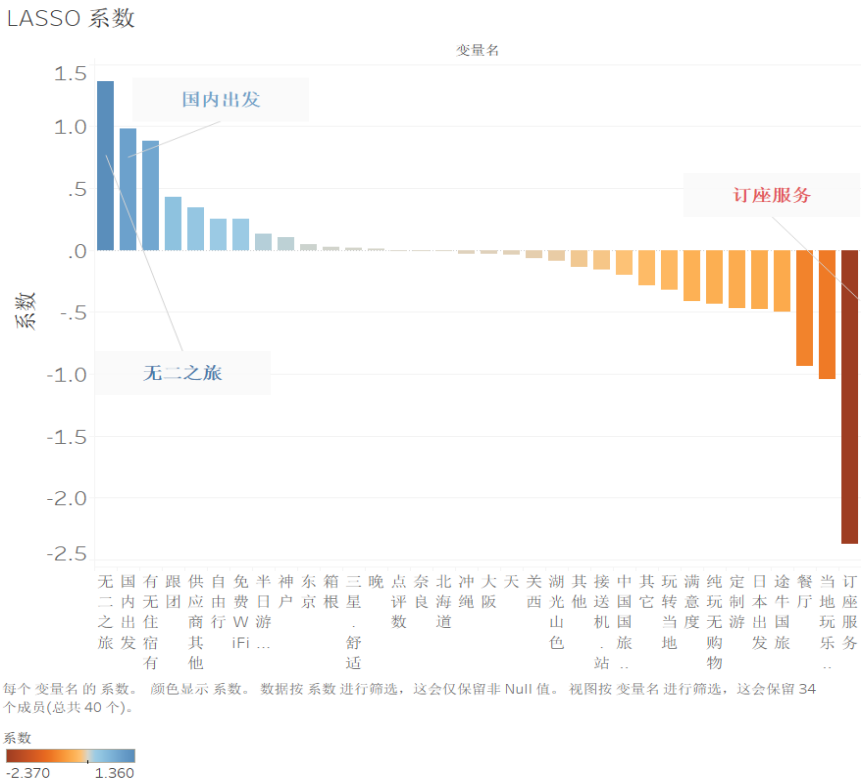
在使用不同的 lambda 系数的情况下，取价格的对数作为因变量，得到的模型对测试数据进行预测的 MSE 如下所示：



通选选择 MSE 较小的模型，最终采用的 lambda 值约为 0.0036，得到模型系数如下：

满意度		-4.12E-01	**供应商**	**其他**	3.44E-01
出游人数.人.		剔除		**途牛国旅**	-4.98E-01
点评数		3.51E-04		**玩转当地**	-3.20E-01
天		-4.03E-02		**无二之旅**	1.36E+00
晚		1.11E-02		**中国国旅苏州旗舰店**	-2.03E-01
优惠方式分类其他	**其他**	剔除	**路线包含城市**	**东京**	4.75E-02
有无住宿有	**有**	8.79E-01		**大阪**	-3.39E-02
住宿条件	**半日游.一日游**	1.32E-01		**京都**	剔除
	餐厅	-9.37E-01		**奈良**	-8.37E-03
	纯玩无购物	-4.32E-01		**富士山**	剔除
	订座服务	-2.37E+00		**冲绳**	-2.92E-02
	湖光山色	-8.73E-02		**北海道**	-8.46E-03
	接送机.站	-1.61E-01		**关西**	-6.55E-02
	免费 WiFi	2.50E-01		**箱根**	2.87E-02
	其他	-1.36E-01		**神户**	1.02E-01
	三星.舒适	1.82E-02	**出游类型**	**自由行**	2.55E-01
出发地	**国内出发**	9.80E-01		**跟团**	4.31E-01
	其它	-2.86E-01		**目的地参团**	剔除
	日本出发	-4.76E-01		**当地玩乐一日.半日**	-1.04E+00

				游	
				定制游	-4.73E-01
				其他	剔除
			特色	**温泉**	剔除



模型分析：

① 从被剔除的变量中可以发现，在 lasso 模型的假设下，出游方式、优惠方式、路线中是否包含富士山或京都、是否有温泉服务与价格并没有比较强的关系。天数和晚数与价格的关系与一般的认识有出入，考虑共线性后也没有得到正相关结果，在分析数据集后推测可能这部分的缺失值较多，影响了回归结果。在住宿方面，有住宿的情况下价格要高一些，而从系数之间的差值上可以看出，具有免费 wifi、半日游一日游、湖光山色等特质的住宿条件的项目要贵一些。

② 从出发地可以看出，国内出发的价格相对于日本出发和其他要更高，比如相对于日本出发，国内出发的价格平均要高出 exp(0.98+0.476) 元。

③ 从供应商部分的结果可以看出，无二之旅、其他供应商提供的旅游项目价格水平要比途牛国际、运转当地、中国国旅苏州旗舰店要高。

④ 而从我们比较看重的出游类型来看，自由行、跟团的旅游项目的平均价格显然要比定制游、一日半日游的要高。

⑤ 对于加入的“温泉”的特色来看，该变量与价格没有足够的相关性，说明温泉服务不是对价格产生很大影响的因素。

该模型最终在数据集上的 MSLE 为 1.12。

3. Random Forrest 回归模型

随机森林是一个高度灵活的机器学习方法，拥有广泛的应用前景，从市场营销到医疗保健保险。 既可以用来做市场营销模拟的建模，统计客户来源，保留和流失，也可

用来预测疾病的风险和病患者的易感性。

随机森林是一个可做回归和分类的模型，具备处理大数据的特性。

为对比线性回归模型和 lasso 模型，我们另外用 Random Forrest 模型进行拟合，以探索更好的预测模型。

最终得到 Random Forrest 回归模型的 MS LE 为 1.01。

4. K-means 聚类

为了探究各种日本旅行项目之间的异同，我们对这个数据进行了聚类分析，删掉了区别较小的变量，得到了以下的结果。

类别	价格	满意度	出游人数	天	出发地分类国内出发	供应商分类途牛国旅	大阪	京都	自由行
1	4346.46	0.97	150.68	6.18	0.71	0.65	0.84	0.76	0.08
2	4089.75	0.97	5262.63	5.88	0.63	0.38	0.75	0.50	0.50
3	3732.25	0.97	121.50	5.53	0.45	0.43	0.32	0.16	0.88
4	3594.84	0.96	1.82	2.01	0.39	0.18	0.20	0.29	0.01
5	9193.99	0.96	14.00	4.83	0.06	0.05	0.48	0.21	0.95
6	713.13	0.94	1.35	0.92	0.01	0.15	0.13	0.19	0.00

针对聚类结果，我们把各种旅游项目项目分为了六种类型：大阪京都偏爱型、网红爆款型、经济旅游型、冷门短期出游型、土豪自由放飞型、超短期体验型。

大阪京都偏爱型：这部分的项目偏好大阪和京都，但是并非自由行，大部分由途牛国际安排，由国内出发居多。

网红爆款型：这部分的项目有非常多人购买，满意度也不错。

经济旅游型：这部分项目较便宜，也有不少游客购买。

冷门双休出游型：这部分的项目天数的均值为 2，价格也比较便宜。

土豪自由放飞型：这部分项目的价格比较高，并且绝大部分为自由行项目。

超短期体验型：这部分的平均天数约为 1，价格非常便宜。

五、结论

1. 旅行者在选择旅游产品的时候，应该多多甄别产品的质量，而不应该被旅游天数迷惑，因为会存在部分旅游产品故意通过延长旅游天数，降低产品质量来吸引眼球；也会存在部分旅游产品的天数与实际不符，如，部分旅游产品的第一天和最后一天是在飞机上度过。
2. 在选择产品时，会存在有些旅行团提供廉价但是质量高的产品，也会有如从多地出发而不是同城出发的低价产品，旅游者可以多多注意这些产品。
3. 如果更灵活一点，在产品的组合上也可以做些选择，如我们发现存在一些低价的自由行产品以及低价的当地游产品，通过两者的组合，旅行者很有可能找到总价比直接跟团旅游更便宜的选项。
4. 如果用户怕选到不靠谱的旅游项目，可以选择已出行人数多，旅游时长 5～7 天的旅游产品，通过前面分析，我们发现这些产品一般价格中规中矩，而且用户满意度都很不错，这样就避免出现一些如踩坑的情况。