基于大众点评网平台的火锅店数据分析

姓　　名：李慧敏

学　　号：14210720074

专　　业：微电子学院

时　　间：2015-7-8

目录

1 实验要求

请到 http://www.dianping.com/search/category/1/10/g110 抓取 40 家火锅店的店名、位置、点评条数、人均消费额、口味评分（均分）、环境评分（均分）、服务评分（均分）、点评的均分、有无团购、有无送外卖 、有无订座等数据（还可以采集标签等数据，这部分数据可选，但有可能影响分析效果），采用数据挖掘中的分类、聚类、关联或回归等中的至少 3 种方法（算法），对火锅店进行分类、找出每类火锅店的特征，并分析点评均分与其他变量之间的关系。可以使用 IBM SPSS Statistics、IBM SPSS Modeler 等工具。

2 实验目的

近年来互联网行业迅猛发展，快速渗透到餐饮、娱乐、购物等各行各业，与我们的生活息息相关。例如被我们所熟知的“大众点评网”就是一个比较成功的例子。线上竞争对于餐饮业的业绩影响越来越大。

如何获得更高的评分？如何赢取更多的人气？这都成为商家十分关注的问题。这次实验，通过从大众点评网上抓取多家火锅店的相关数据，来探索影响其线上业绩的因素。具体来说，有以下几个目的：

（1） 找到影响点评分数的关键因素；

（2） 找到影响点评条数的关键因素，因为点评条数一定程度上反映了通过该平台去该餐厅就餐的人数；

（3） 分析不同的影响因素之间的关联关系；

（4） 对火锅店进行聚类，分析不同类别的特点，并联系实际给出有利于市场的建议。

3 实验工具

本次实验我主要使用了 IBM SPSS Modeler，为了生成一些可视化图表，我结合使用了 SAP Lumira 作为辅助。

4 实验过程

4.1 数据预处理

首先我对数据进行了预处理。观察发现，原始数据中不存在空值或明显异常

的无效数据。可以看到，先除去标签数据不考虑，可能的影响因素有位置、点评条数、口味、服务、环境、点评分数、人均消费额、有无团购、有无外送、有无订座，而其中只有位置、有无团购、有无外送、有无订座是离散的数据。点评条数、口味、服务、环境、点评分数、人均消费额，这几项都是连续数值，因此要先把这几项做离散化处理。

为了科学地进行离散化，我先使用了 Lumira 中的箱线图对这几项数据的分布概况做一个简单的统计。箱线图是一种能够比较直观反映数据分布的图，里面包含了中值、上四分位数、下四分位数、上限、下限，还能看出是否有异常值存在。但是这里的异常值我们不能简单地直接删掉不做考虑，因为这是通过实际收集到的真实数据。通过箱线图，制定合理的离散化方案。

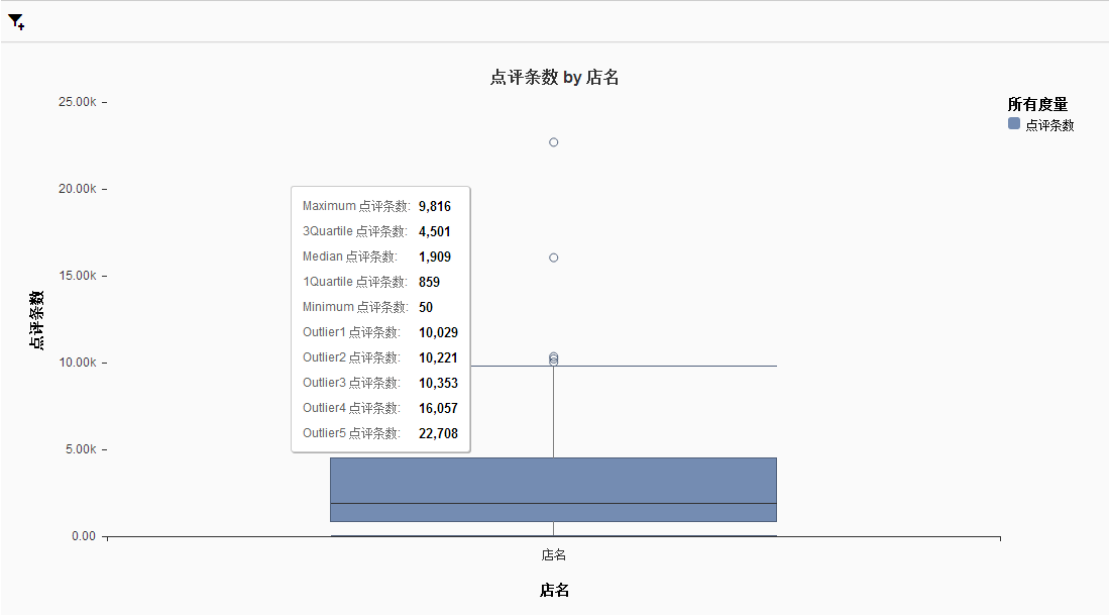


图 1 点评条数箱线图

根据图 1 中的数据，这里将点评条数离散化为四个分类，分别以 500、2000、5000 作为三个分界点，得到的四类是“点评条数少”、“点评条数较少”、“点评条数较多”和“点评条数多”。

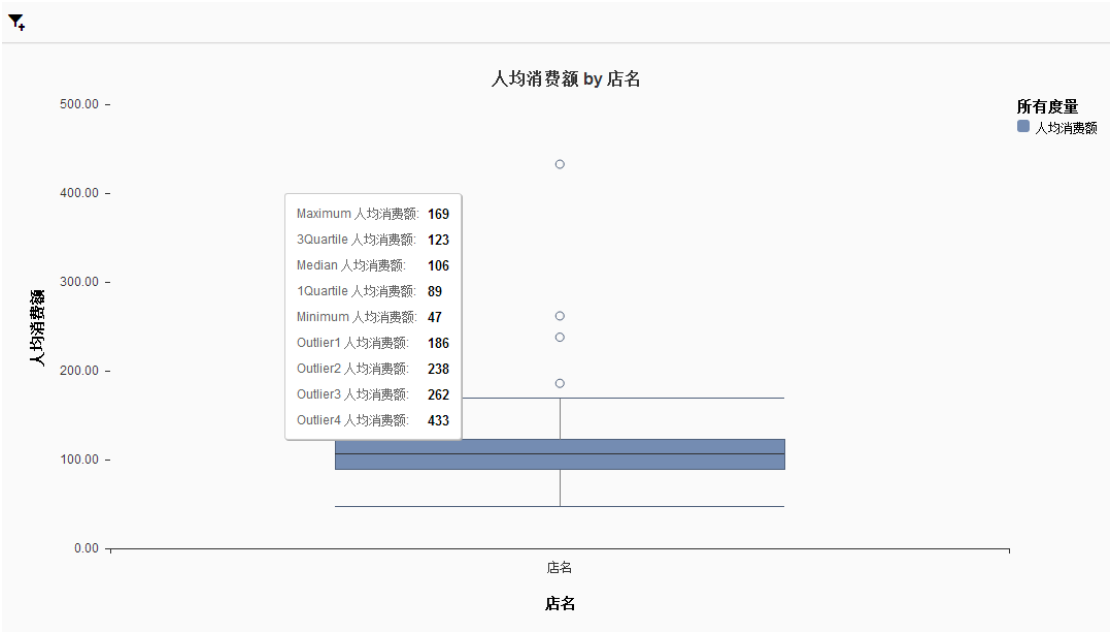


图 2 人均消费额箱线图

离散化的时候分为三类，是因为考虑到人均消费额的差距不是特别大，如果分类太多可能导致分类效果不明显，所以分为三类。分别以 80、150 为分界线，三个分类为“人均消费低”、“人均消费一般”和“人均消费高”。

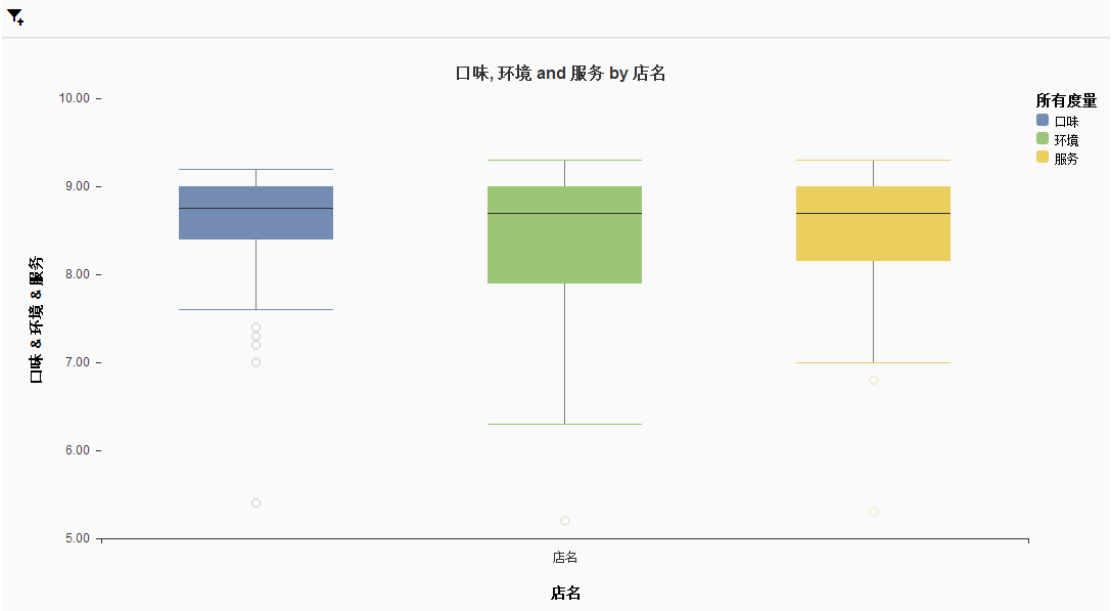


图 3 口味、环境和服务评分箱线图

可以看到口味、环境和服务三个的分布差别不大，可以以同样标准进行分类，分为三类，两个分类节点为 7.6 和 8.8。三个分类是“评分较低”、“评分一般”和“评分较高”。

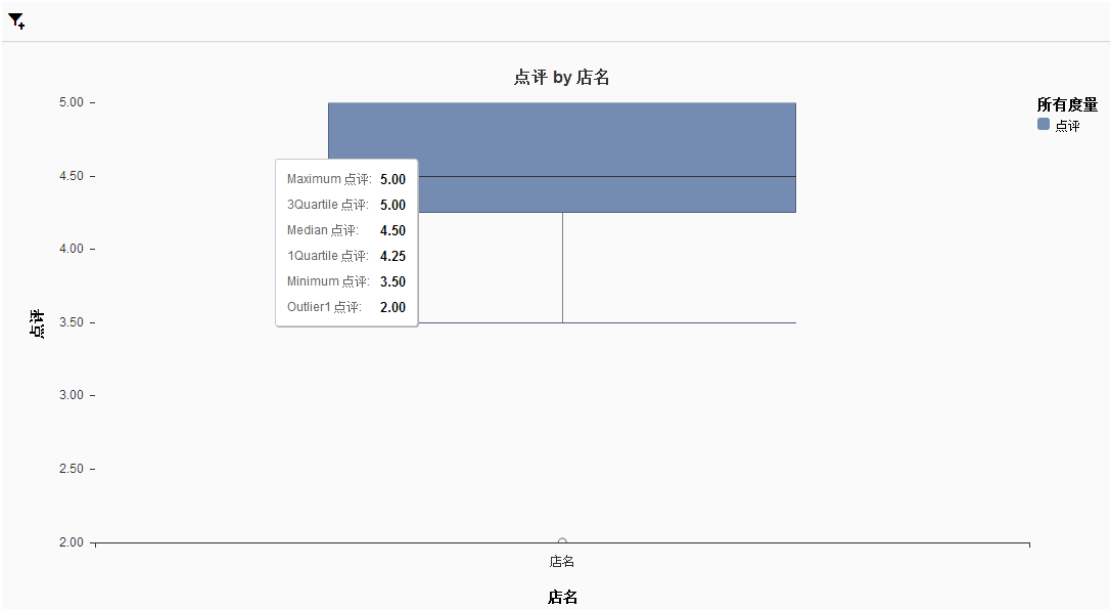


图 4 点评箱线图

将其分为三类，小于等于 4 为“点评较低”，大于 4 小于等于 4.5 为“点评一般”，大于 4.5 为“点评较高”。

离散的时候使用“if-else”公式，如图 5 所示。离散化后的表格如图 6 所示。

图 5 对点评的离散化处理

	店名	位置	...	...	...	...	...	...	有无团购	有无外送	有无订座	...	...	...	...	...	...	...	...	...	点评条数new	人均消费new	口味new	环境new	服务new	点评new
1	Mo-Mo牧场1	张江	...	...	...	...	...	...	无	无	有	...	...	...	...	...	...	...	...	...	点评条数较少	人均消费一般	口味较好	环境较好	服务较好	点评较高
2	Mo-Mo牧场2	淮海路	...	...	...	...	...	...	无	无	有	...	...	...	...	...	...	...	...	...	点评条数多	人均消费高	口味一般	环境一般	服务较好	点评一般
3	Season三层堂和牛主题餐厅	淮海路	...	...	...	...	...	...	有	无	无	...	...	...	...	...	...	...	...	...	点评条数较少	人均消费高	口味较好	环境一般	服务一般	点评较高
4	煲宫1	人民广场	...	...	...	...	...	...	有	无	无	...	...	...	...	...	...	...	...	...	点评条数多	人均消费一般	口味较好	环境一般	服务较好	点评较高
5	煲宫2	徐家汇	...	...	...	...	...	...	有	无	有	...	...	...	...	...	...	...	...	...	点评条数较少	人均消费一般	口味较好	环境较好	服务较好	点评较高
6	打望重庆火锅	淮海路	...	...	...	...	...	...	无	无	无	...	...	...	...	...	...	...	...	...	点评条数较少	人均消费一般	口味一般	环境一般	服务一般	点评较低
7	大队长主题火锅	徐家汇	...	...	...	...	...	...	有	无	有	...	...	...	...	...	...	...	...	...	点评条数多	人均消费一般	口味较好	环境较好	服务较好	点评较高
8	东来福热气涮羊肉	陆家嘴	...	...	...	...	...	...	无	无	无	...	...	...	...	...	...	...	...	...	点评条数少	人均消费低	口味较差	环境较差	服务较差	点评较低
9	洞子张老匠火锅	五角场	...	...	...	...	...	...	有	无	无	...	...	...	...	...	...	...	...	...	点评条数较多	人均消费一般	口味一般	环境一般	服务一般	点评一般
10	豆捞坊1	人民广场	...	...	...	...	...	...	有	无	无	...	...	...	...	...	...	...	...	...	点评条数多	人均消费一般	口味一般	环境一般	服务一般	点评较高
11	豆捞坊2	莘庄	...	...	...	...	...	...	有	无	无	...	...	...	...	...	...	...	...	...	点评条数多	人均消费一般	口味一般	环境较好	服务一般	点评一般
12	豆捞坊3	静安寺	...	...	...	...	...	...	有	无	无	...	...	...	...	...	...	...	...	...	点评条数较多	人均消费一般	口味较好	环境较好	服务较好	点评较高
13	凡之鱼	虹桥	...	...	...	...	...	...	无	无	无	...	...	...	...	...	...	...	...	...	点评条数较少	人均消费低	口味一般	环境一般	服务较好	点评一般
14	呷哺呷哺	莘庄	...	...	...	...	...	...	无	无	无	...	...	...	...	...	...	...	...	...	点评条数较多	人均消费低	口味较差	环境较差	服务较差	点评较低
15	锅德城市火锅	静安寺	...	...	...	...	...	...	有	无	有	...	...	...	...	...	...	...	...	...	点评条数较少	人均消费高	口味较好	环境较好	服务较好	点评较高
16	海底捞火锅1	五角场	...	...	...	...	...	...	无	有	无	...	...	...	...	...	...	...	...	...	点评条数较多	人均消费一般	口味较好	环境较好	服务较好	点评较高
17	海底捞火锅2	徐家汇	...	...	...	...	...	...	无	无	无	...	...	...	...	...	...	...	...	...	点评条数较多	人均消费一般	口味较好	环境较好	服务较好	点评较高
18	红辣椒	五角场	...	...	...	...	...	...	有	无	无	...	...	...	...	...	...	...	...	...	点评条数多	人均消费一般	口味较好	环境较好	服务较好	点评较高
19	皇城根	淮海路	...	...	...	...	...	...	无	无	有	...	...	...	...	...	...	...	...	...	点评条数较多	人均消费一般	口味一般	环境较好	服务一般	点评一般
20	黄记煌三汁焖锅	莘庄	...	...	...	...	...	...	无	无	无	...	...	...	...	...	...	...	...	...	点评条数较少	人均消费一般	口味一般	环境一般	服务一般	点评一般
21	季悦火锅	淮海路	...	...	...	...	...	...	有	有	有	...	...	...	...	...	...	...	...	...	点评条数较少	人均消费高	口味较好	环境较好	服务较好	点评较高
22	辣府	人民广场	...	...	...	...	...	...	有	无	无	...	...	...	...	...	...	...	...	...	点评条数多	人均消费一般	口味较好	环境一般	服务一般	点评一般
23	老工厂重庆火锅	徐家汇	...	...	...	...	...	...	有	无	有	...	...	...	...	...	...	...	...	...	点评条数较多	人均消费一般	口味较好	环境一般	服务较好	点评较高
24	老李羊蝎子涮锅	人民广场	...	...	...	...	...	...	有	无	无	...	...	...	...	...	...	...	...	...	点评条数较少	人均消费一般	口味一般	环境一般	服务一般	点评较低
25	老渝城正宗重庆老火锅1	五角场	...	...	...	...	...	...	有	无	有	...	...	...	...	...	...	...	...	...	点评条数较少	人均消费一般	口味一般	环境一般	服务一般	点评一般
26	老渝城正宗重庆老火锅2	五角场	...	...	...	...	...	...	有	无	有	...	...	...	...	...	...	...	...	...	点评条数少	人均消费一般	口味一般	环境一般	服务一般	点评一般
27	龙门鱼府	张江	...	...	...	...	...	...	有	无	有	...	...	...	...	...	...	...	...	...	点评条数少	人均消费低	口味较差	环境较差	服务较差	点评较低
28	卢记麻辣鱼火锅	虹桥	...	...	...	...	...	...	无	无	有	...	...	...	...	...	...	...	...	...	点评条数少	人均消费一般	口味一般	环境一般	服务一般	点评较低
29	梅亭	静安寺	...	...	...	...	...	...	有	有	有	...	...	...	...	...	...	...	...	...	点评条数少	人均消费一般	口味一般	环境一般	服务一般	点评一般
30	陌生人火锅餐厅	五角场	...	...	...	...	...	...	有	无	有	...	...	...	...	...	...	...	...	...	点评条数较少	人均消费一般	口味较好	环境较好	服务较好	点评较高
31	牧羊传奇火锅	张江	...	...	...	...	...	...	有	无	有	...	...	...	...	...	...	...	...	...	点评条数少	人均消费一般	口味较好	环境一般	服务一般	点评一般
32	漂亮石头火锅	陆家嘴	...	...	...	...	...	...	无	无	无	...	...	...	...	...	...	...	...	...	点评条数少	人均消费低	口味较差	环境较差	服务较差	点评较低
33	齐民市集	虹桥	...	...	...	...	...	...	有	无	无	...	...	...	...	...	...	...	...	...	点评条数较少	人均消费一般	口味一般	环境较好	服务较好	点评一般
34	热辣一号	淮海路	...	...	...	...	...	...	有	无	有	...	...	...	...	...	...	...	...	...	点评条数较多	人均消费一般	口味一般	环境一般	服务一般	点评一般
35	三人行得头王	徐家汇	...	...	...	...	...	...	有	有	有	...	...	...	...	...	...	...	...	...	点评条数多	人均消费低	口味一般	环境一般	服务一般	点评一般
36	三人行得头王火锅1	人民广场	...	...	...	...	...	...	有	有	有	...	...	...	...	...	...	...	...	...	点评条数多	人均消费一般	口味一般	环境一般	服务一般	点评一般
37	三人行得头王火锅2	虹桥	...	...	...	...	...	...	有	有	有	...	...	...	...	...	...	...	...	...	点评条数多	人均消费低	口味一般	环境一般	服务较好	点评一般

图 6 离散化后的表格

4.2 C5.0 决策树模型探索影响因素

4.2.1 分析影响点评分数的因素

对以上的几项离散化生成几个新的维，之后就可以使用 C5.0 决策树进行分类从而分析找出关键影响因素了。

这里我把影响因素分为三类：（1）有无团购、外送、有无订座；（2）口味、服务、环境、点评条数；（3）位置、人均消费额。

（1） 平台额外促销便利因素

输入：有无团购、有无外送、有无订座

目标：点评

决策树模型：生成了一个一层的决策树，可以看到根据“有无团购”分成了两类。有团购时点评主要分布在点评一般和点评较高，两者在有团购的店家中占到了 85.715%的比重，无团购的店家点评主要分布在点评较低这一类，占了 58.824%。因此也可以得出结论，有无团购是影响点评评分的一个因素。

结论：这一结论不难理解，当有团购时，价格会优惠很多。我们会认为性价比相对高一些，购买的体验也会提升，因此我们会倾向于给更高的评分。而针对有无外送，我觉得可能是由于火锅这一特定类型基本都不支持外送，所以本身的数据也很少，也就无法成为关键影响因素。

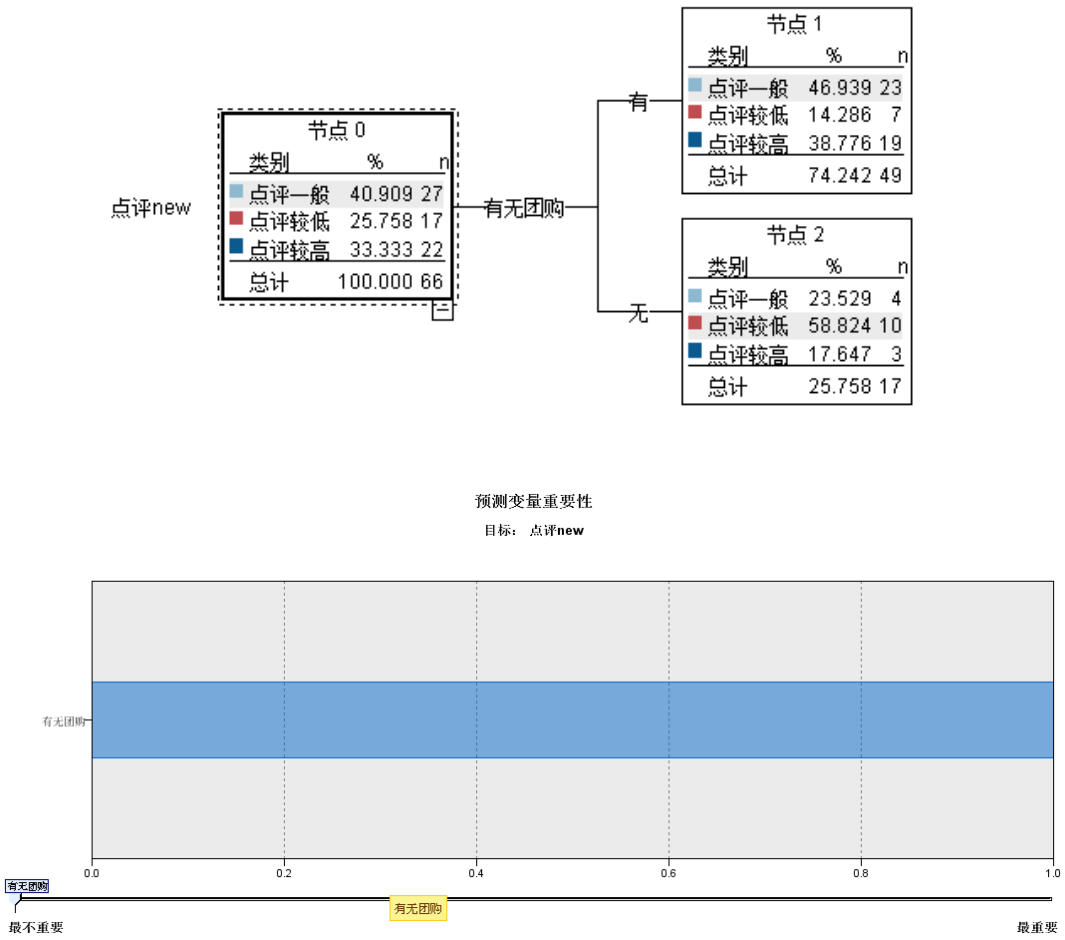


图 7 团购、外送、订座作为影响因素的 C5.0 模型

（2） 其他顾客提供的信息

输入：口味、服务、环境、点评条数

目标：点评

决策树模型：可以看到这里生成了一个两层决策树模型。首要的影响因素是服务，可以看到服务一般的店家里面有 71.429%是点评一般，服务较好的店家里 74.074%是点评较高，服务较差店家里面 100%都是点评较低。而服务较好的一类里面，口味又成为影响因素。口味一般的里面 100%是点评一般，而口味较好的点里面 86.957%是点评较高的。

结论：大家在给出点评分数的时候，对服务的质量还是很注重的。在很大程度上，服务的质量直接影响到了点评的分数。而口味也在较小的程度上影响点评分数。由此看来，火锅店家除了把味道做好之外一定要在服务上下点功夫，争取留住顾客的心啊。

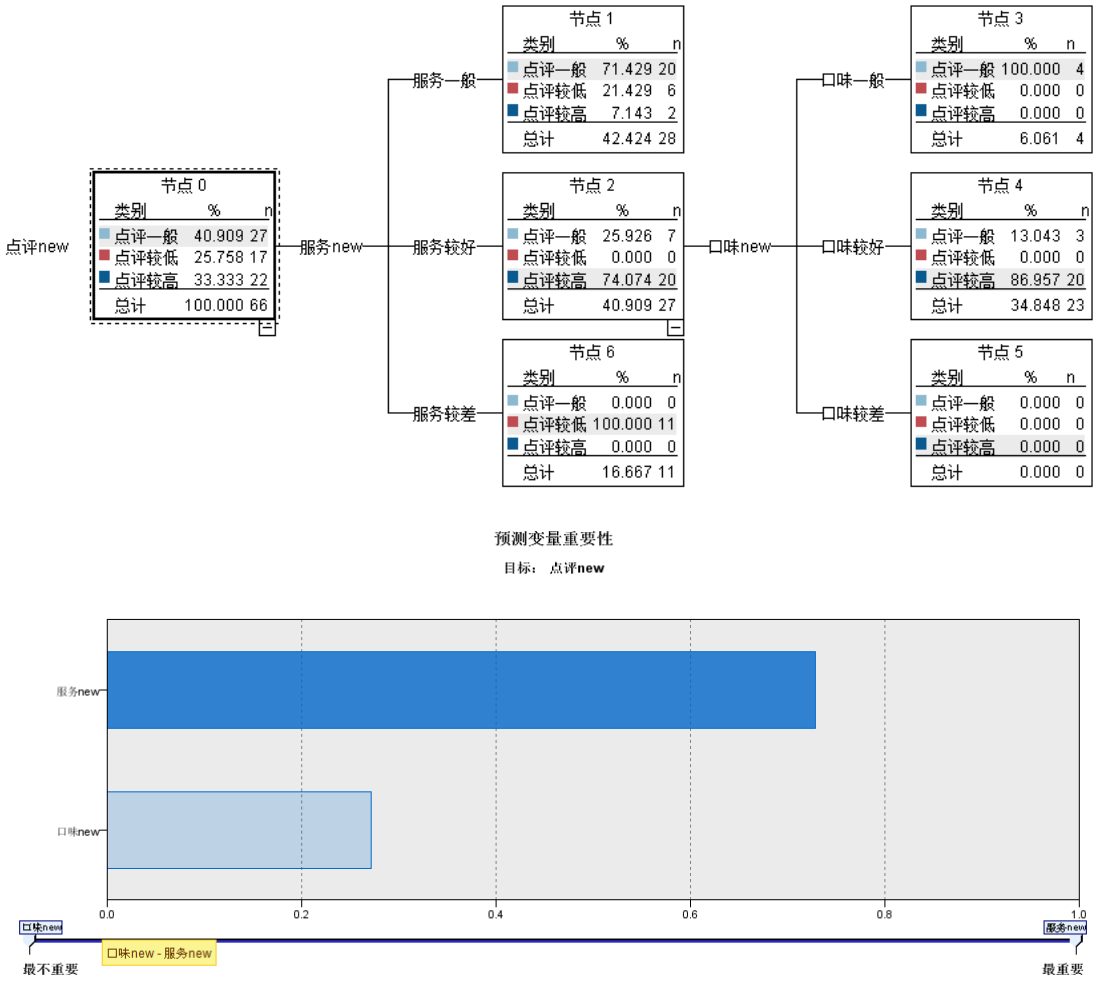


图 8 口味、服务、环境、点评条数作为影响因素的 C5.0 模型

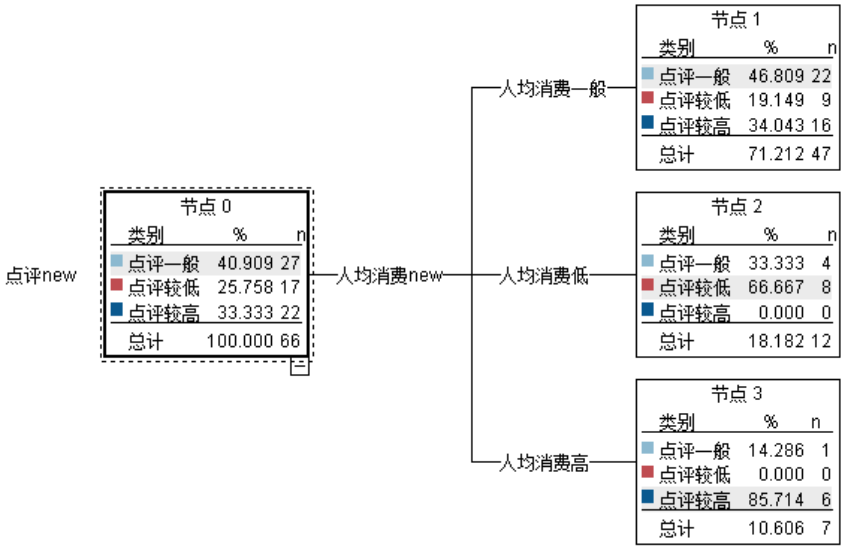
（3） 店家本身固定属性

输入：位置、人均消费额

目标：点评

决策树模型：生成了一个一层的决策树模型，可以看到根据人均消费额分成了三类，人均消费的高、低、一般分别对应点评的高、低、一般。因此人均消费是一个影响因素。

结论：一分价钱一分货是自古以来的说法，在这里得到了体现。可以看到，人均价格高的绝大部分都是点评较高的，而人均价格低的绝大部分都是点评较低的。看来，消费者想要更好的体验，还是要多下点血本啊。



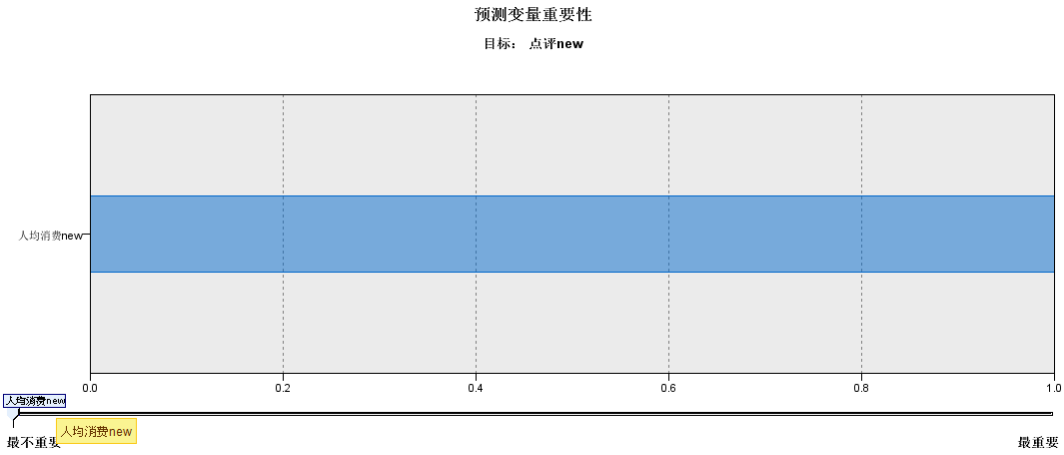


图 9 位置、人均消费额作为影响因素的 C5.0 模型

（4） 整体分析

输入：有无团购、服务、人均消费额

目标：点评

决策树模型：生成的决策树有一层，与（2）的结果一模一样，服务的影响还是最大的。

结论：这次进一步说明了店家要想让自己在大众点评网上评分高，就要提高服务质量的道理。

此时建立的模型结构图如图 10 所示。

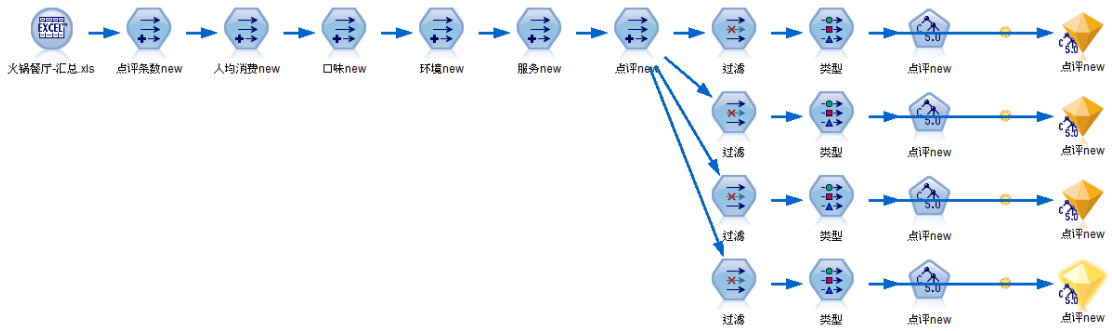


图 10 以点评为目标的决策树分析模型结构图

4.2.2 分析影响点评条数的因素

由于点评条数在一定程度上反映了顾客通过“大众点评网”这个平台在某家火锅店消费的情况，因此这里也将点评条数作为目标，分析其影响因素。以制定更好的策略拉动更多的顾客。

类似于 4.2.1 中，这里我也是把影响因素分为三类：（1）有无团购、外送、有无订座；（2）口味、服务、环境、点评；（3）位置、人均消费额。

（1） 平台额外促销便利因素

输入：有无团购、有无外送、有无订座

目标：点评条数

决策树模型：可以看到生成了一个一层的决策树。有无外送是一个影响因素，在有外送的店家里，55.555%的店家点评数较多；在无外送的店家里，38.596%的点评数较少。但是也可以注意到，无外送的店家里，28.07%的点评数较多，位居第二。说明虽然有无外送是一个影响因素，但是影响不是很大。而且可以看到，只有 9 家火锅店有外送，数据量较少，不是很能反应出真实的情况。这一点在前面也有提到过。

结论：有无外送能在一定程度上影响点评条数，毕竟有外送的话会带来一些便利条件，相应的购买人数就多了。但是，毕竟支持火锅外送的店家还是极少数的。而且这次的数据量较少，所以该结论的真实度还有待考证。

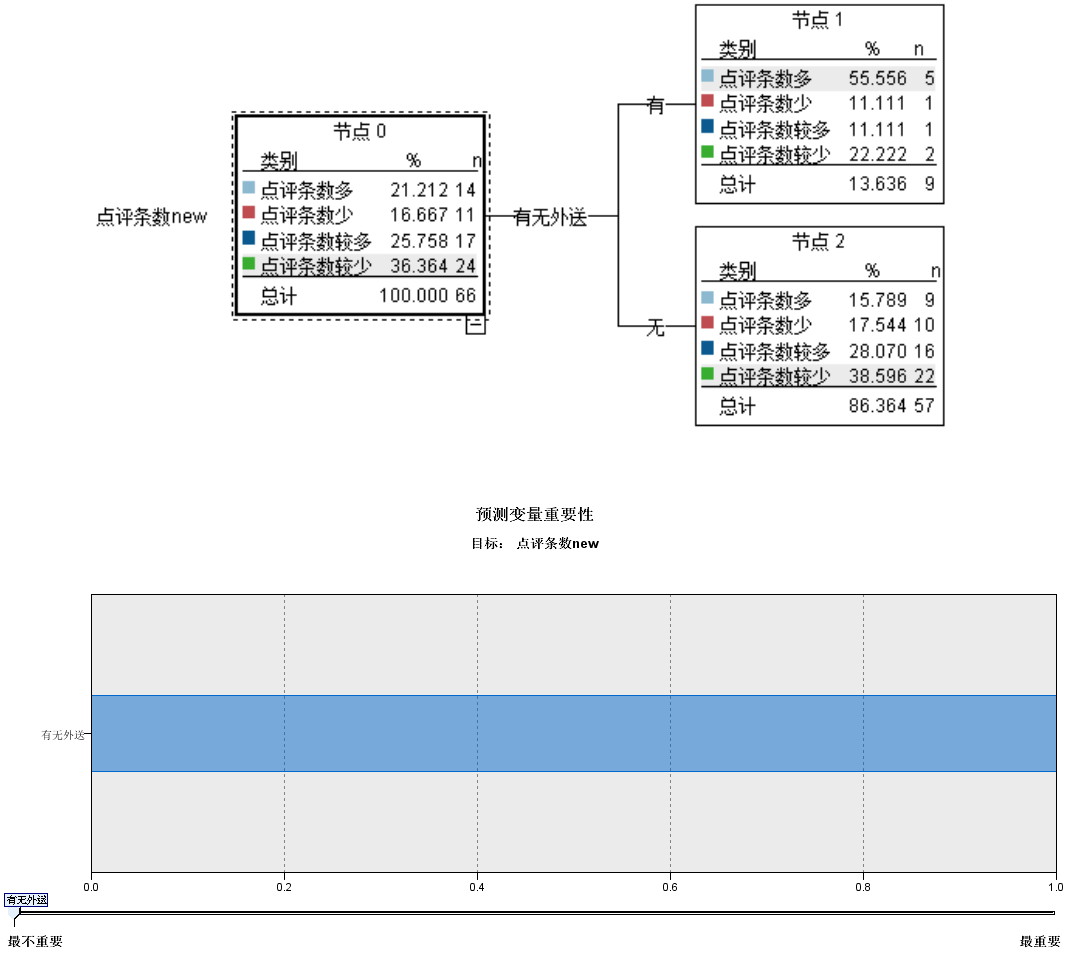


图 11 团购、外送、订座作为影响因素的 C5.0 模型

（2） 其他顾客提供的信息

输入：口味、服务、环境、点评

目标：点评条数

决策树模型：生成了一层决策树。可以看出，点评条数多的都在口味一般和口味较好类里面，说明一定程度上口味还是会影响点评条数的。但是由于口味较差里面只有 6 家店，略少，因此该决策树存在一定的问题。

结论：口味在一定程度上会影响点评条数，毕竟人们会更加倾向于选择口味好的火锅店。

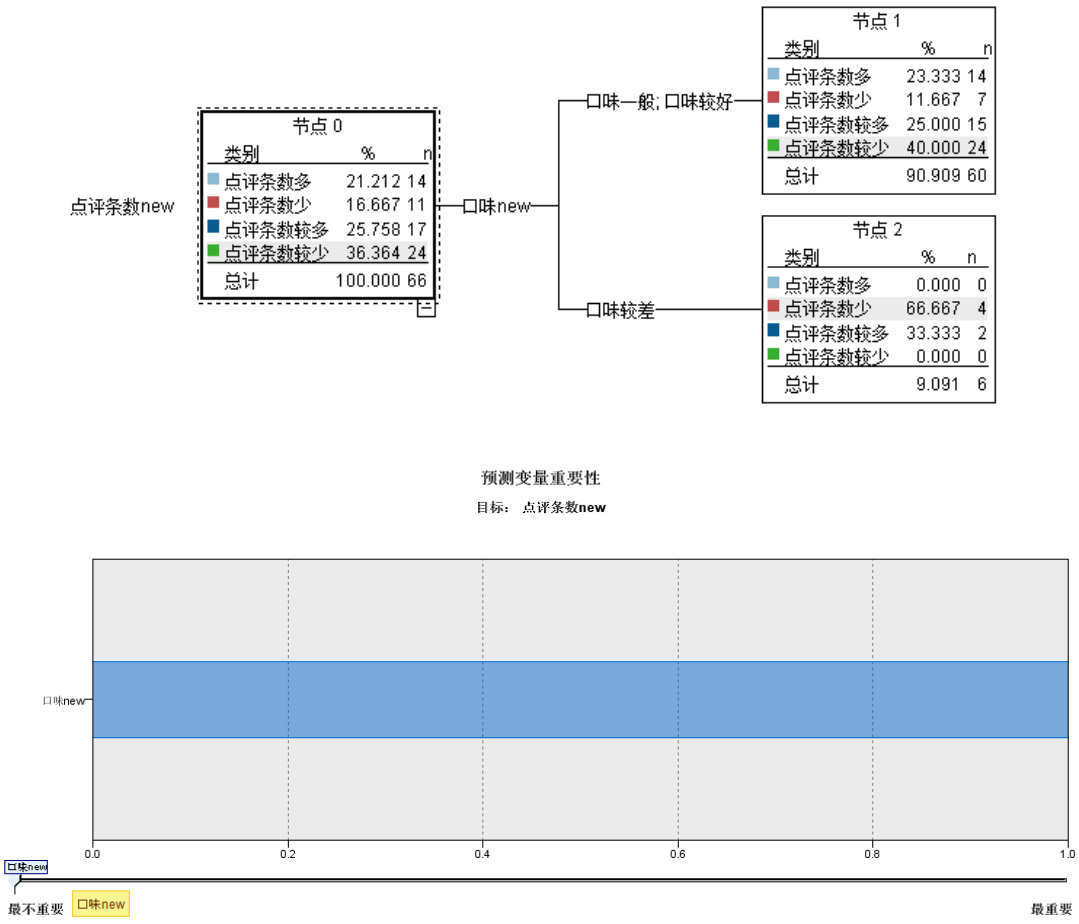


图 12 口味、服务、环境、点评作为影响因素的 C5.0 模型

（3） 店家本身固定属性

输入：位置、人均消费额

目标：点评条数

决策树模型：生成了一维决策树。影响因素是位置。可以看到人民广场和徐家汇的点评条数多；五角场、莘庄、静安寺的点评条数较多，淮海路和虹桥的点评数较少，张江和陆家嘴的点评条数最少。这一结果可以结合不同商圈的特性来解释。例如人民广场、徐家汇都是非常繁华、人流密集的地方，而且是休闲娱乐、逛街放松的地带，因此吃火锅的人相对多一些。而张江人口相对稀少，陆家嘴又是金融白领上班一族聚集的地方，工作很忙，可能没有太多时间吃火锅。虹桥地处上海西南角，稍有一些偏，去的人很多是去机场或火车站，所以吃火锅的相对少，淮海路是上海著名的小资地界，有很多酒吧咖啡厅，可能在那里火锅并不是大家的宠爱。而五角场、莘庄和静安寺也都是人较密集的地方，但是又没有人民广场那么拥挤，所以点评数较多。

结论：不同的商圈适合开不同类型的餐饮店，可以看到像人民广场、徐家汇这样的人口密集、娱乐休闲的商圈比较适合开火锅店。

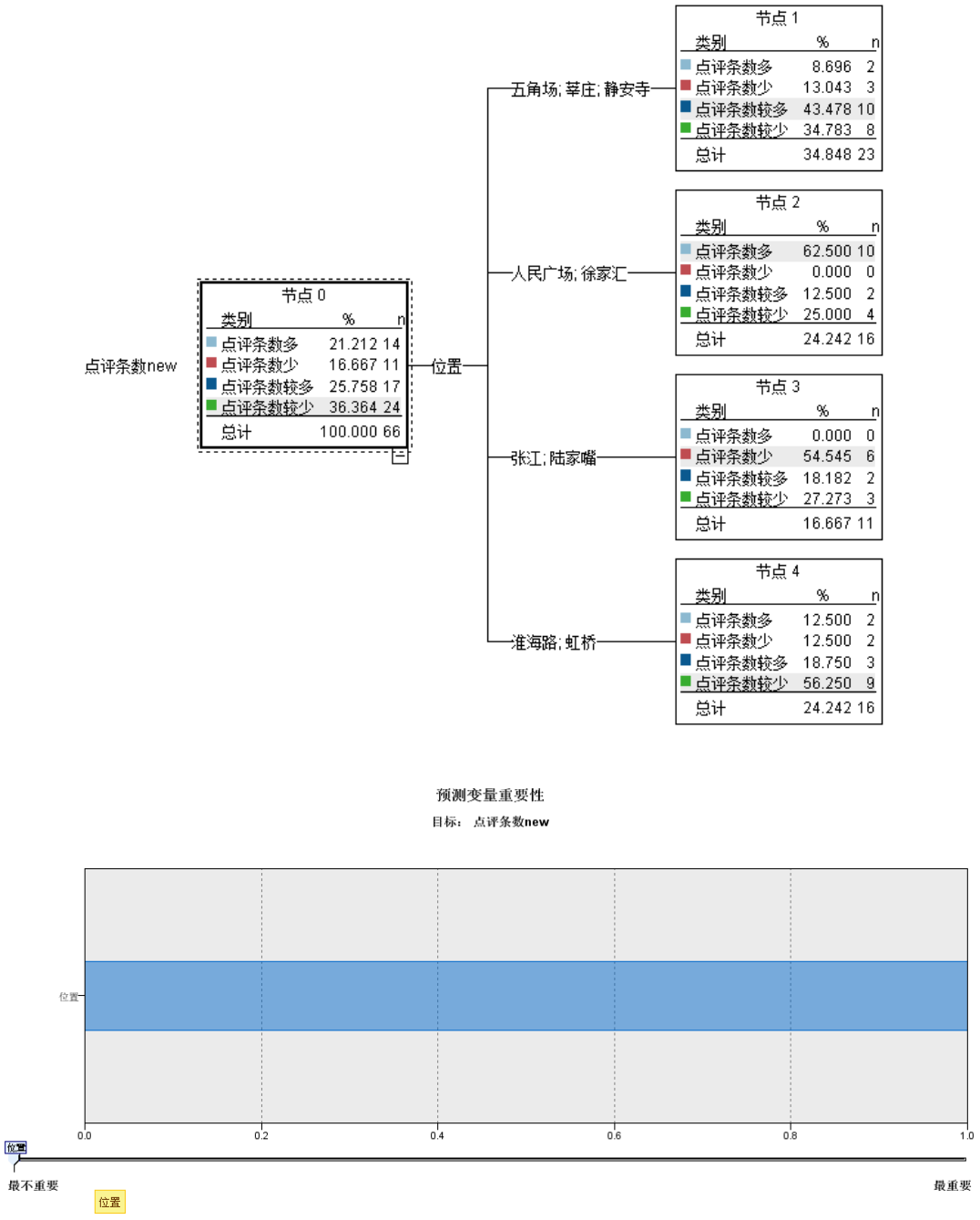


图 13 位置、人均消费额作为影响因素的 C5.0 模型

（4） 整体分析

输入：有无外送、口味、位置

目标：点评条数

决策树模型：生成的结果和（3）中一致，位置是关键影响因素。而且从（1）和（2）中我们也分析了有无外送和口味对点评条数虽然有影响，但是影响并不是特别大。

结论：开火锅店需要评估商圈的位置和定位，这样才能准确把握并迎

合消费者的心理，从而拥有更多的客户。

在以点评条数为目标的决策树分析中，生成的模型如图 14 所示。

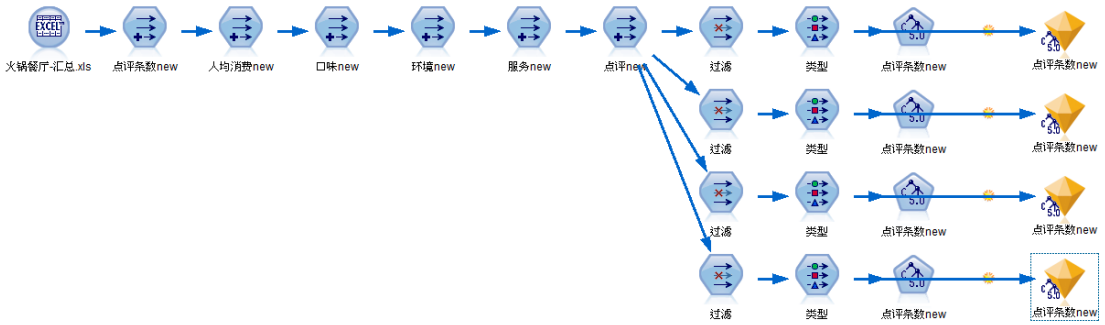


图 14 以点评条数为目标的决策树分析模型结构

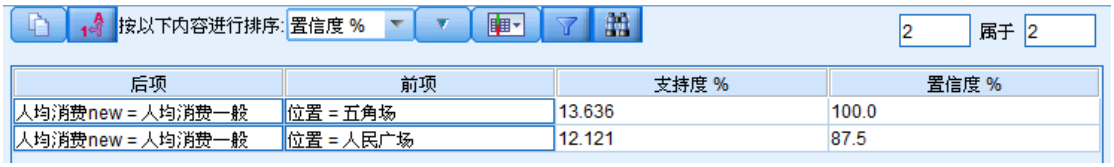
4.3 Apriori 算法进行关联分析

在这一部分里，我对影响因素进行了关联分析，使用了 Apriori 算法。首先还是按照 4.2 中的三个分类进行关联分析，之后又进行了一些交叉分析。所做的关联分析分别有以下几组：（1）位置与人均消费额；（2）有无团购、外送、订座；（3）口味、服务、环境；（4）口味、服务、环境、人均消费额；（5）位置和有无团购、外送、订座。这里，我设置最低条件支持度 10%，最小置信度 80%，最大前项数 5 项。

（1） 店家本身固定属性

分析因素：位置、人均消费额

分析：可以看到五角场的几家火锅店全部都是人均消费水平一般，人民广场的大部分也是人均消费一般。由我推测可能是由于五角场主要是大学生较多，而人民广场这种休闲娱乐场所也要面向广大层次的用户，因此火锅店的消费水平是中等的。



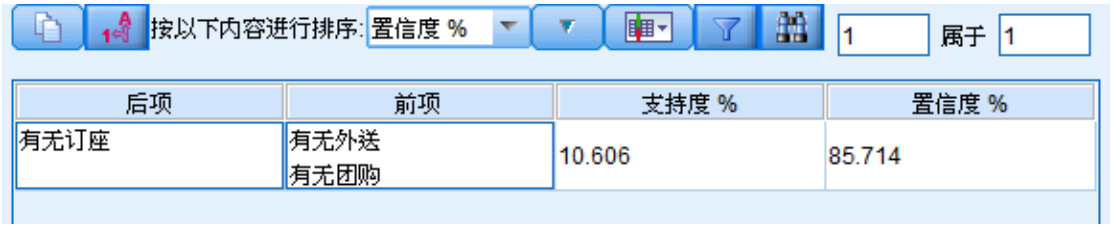
后项	前项	支持度 %	置信度 %
人均消费new = 人均消费一般	位置 = 五角场	13.636	100.0
人均消费new = 人均消费一般	位置 = 人民广场	12.121	87.5

图 15 店家本身固定属性因素关联分析

（2） 平台额外促销便利因素

分析因素：有无团购、有无外送、有无订座

分析：可以看到当该商家有团购和外送服务时，大部分都有订座服务。



后项	前项	支持度 %	置信度 %
有无订座	有无外送 有无团购	10.606	85.714

图 16 平台额外促销便利因素关联分析

（3） 其他顾客提供的信息

分析因素：口味、服务、环境

分析：从下图中可以得到的结论是，往往口味、环境和服务是要好都好，要差都差。因此可知往往好的店家在各个的方面都比较注重，而一些差的火锅店则各个方面都不尽如人意。

后项	前项	支持度 %	置信度 %
服务new = 服务较差	环境new = 环境较差	16.667	100.0
环境new = 环境较差	服务new = 服务较差	16.667	100.0
口味new = 口味较好	环境new = 环境较好 服务new = 服务较好	31.818	95.238
服务new = 服务较好	环境new = 环境较好 口味new = 口味较好	31.818	95.238
环境new = 环境较好	服务new = 服务较好 口味new = 口味较好	34.848	86.957
环境new = 环境一般	服务new = 服务一般 口味new = 口味一般	34.848	86.957
服务new = 服务一般	环境new = 环境一般 口味new = 口味一般	34.848	86.957
环境new = 环境一般	服务new = 服务一般	42.424	85.714
口味new = 口味较好	服务new = 服务较好	40.909	85.185
服务new = 服务较好	环境new = 环境较好	37.879	84.0
口味new = 口味较好	环境new = 环境较好	37.879	84.0
口味new = 口味一般	服务new = 服务一般 环境new = 环境一般	36.364	83.333
服务new = 服务较好	口味new = 口味较好	42.424	82.143
口味new = 口味一般	服务new = 服务一般	42.424	82.143
服务new = 服务一般	环境new = 环境一般	45.455	80.0

图 17 其他顾客提供的信息关联分析

（4） 不同方面因素的交叉分析

分析因素：口味、服务、环境、人均消费

分析：除了得到（3）中的结论之外，我们还可以看到，当人均消费水平低的时候，往往服务和环境也较差，而当人均消费水平高的时候，服务和环境一般会比较好。但是也可以注意到，人均消费水平一般的时候，也有很多餐厅的口味、服务、环境不错的。

后项	前项	支持度 %	置信度 %
服务new = 服务较差	环境new = 环境较差	16.667	100.0
环境new = 环境较差	服务new = 服务较差	16.667	100.0
服务new = 服务较差	环境new = 环境较差 人均消费new = 人均消费低	10.606	100.0
环境new = 环境较差	服务new = 服务较差 人均消费new = 人均消费低	10.606	100.0
服务new = 服务一般	环境new = 环境一般 口味new = 口味一般 人均消费new = 人均消费一般	25.758	100.0
口味new = 口味较好	环境new = 环境较好 服务new = 服务较好	31.818	95.238
服务new = 服务较好	环境new = 环境较好 口味new = 口味较好	31.818	95.238
口味new = 口味较好	服务new = 服务较好 人均消费new = 人均消费一般	28.788	94.737
口味new = 口味较好	环境new = 环境较好 服务new = 服务较好 人均消费new = 人均消费一般	24.242	93.75
服务new = 服务较好	环境new = 环境较好 口味new = 口味较好 人均消费new = 人均消费一般	24.242	93.75
环境new = 环境较好	服务new = 服务较好 口味new = 口味较好	34.848	86.957
环境new = 环境一般	服务new = 服务一般 口味new = 口味一般	34.848	86.957
服务new = 服务一般	环境new = 环境一般 口味new = 口味一般	34.848	86.957
服务new = 服务一般	环境new = 环境一般 人均消费new = 人均消费一般	34.848	86.957
人均消费new = 人均消费一般	服务new = 服务一般 口味new = 口味一般	34.848	86.957
服务new = 服务较好	人均消费new = 人均消费高	10.606	85.714
口味new = 口味较好	人均消费new = 人均消费高	10.606	85.714
环境new = 环境一般	服务new = 服务一般	42.424	85.714

图 18 口味、服务、环境、人均消费关联分析

（5） 不同方面因素的交叉分析

分析因素：位置与团购、外送、订座

分析：除了得到（2）中的结论外，还可以看到地处人民广场和五角场两个地方的火锅店往往都有团购。这可能也是由于商圈的特点造成的。五角场是大学区，而人民广场又是年轻人休闲逛街约会爱去的地方，客流量又较大，因此有团购是不难理解的。

后项	前项	支持度 %	置信度 %
有无团购	位置 = 人民广场	12.121	100.0
有无团购	位置 = 五角场	13.636	88.889
有无订座	有无外送 有无团购	10.606	85.714

图 19 位置与团购、外送、订座关联分析

模型结构流程图如下图 20 所示。

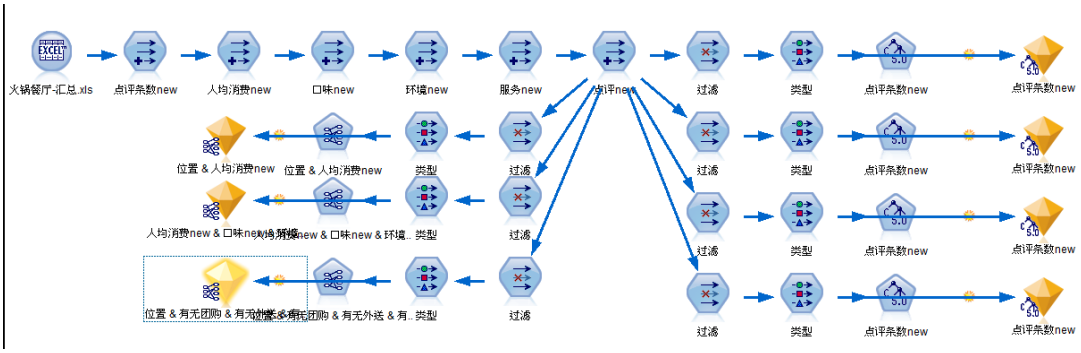


图 20 做完关联分析的模型结构流程图

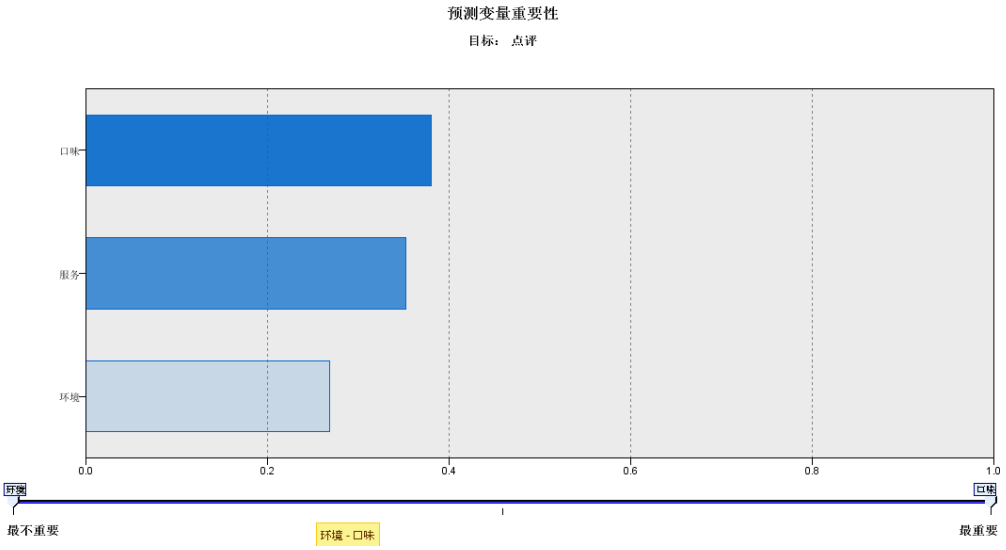
4.4 线性回归

这里之所以要做线性回归分析，是因为在之前的分析中，我观察到点评与口味、服务、环境三者的关系十分密切，基本是要高都高，要低都低的状态，所以通过线性回归，可以得到其中更加具体的关系，从而更好地了解点评分数的影响因素。

输入：口味、服务、环境

目标：点评

分析：如图 21 所示，在“Model Summary”里面，可以看到调整后的判定系数 R 方为 0.883，说明拟合度还是较高的，不被解释的变量较少。而从“ANOVA”表格中看到回归方程显著性检验的概率为 0，小于显著性水平 0.05。说明被解释变量与解释变量全体的线性关系是显著的，可建立线性方程。但是在系数表格里的环境一栏的显著性检验结果不小于 0.05，说明有问题，于是我又将环境这一因素排除，重新做了一次线性回归。



Variables Entered/Removed(a)

Model	Variables Entered	Variables Removed	Method
1	服务, 口味, 环境(b)	.	Enter
a. Dependent Variable: 点评			
b. All requested variables entered.			

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.943(a)	.888	.883	.183474
a. Predictors: (Constant), 服务, 口味, 环境				

ANOVA(a)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	16.629	3	5.543	164.662	.000(b)
	Residual	2.087	62	.034		
	Total	18.716	65			
a. Dependent Variable: 点评						
b. Predictors: (Constant), 服务, 口味, 环境						

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.872	.316		-5.922	.000
	口味	.456	.103	.552	4.428	.000
	环境	-.141	.100	-.219	-1.413	.163
	服务	.428	.141	.612	3.039	.003
a. Dependent Variable: 点评						

图 21 点评与口味、服务、环境的线性回归

输入：口味、服务

目标：点评

分析：可以在图 22 中看到，这次每个系数的显著性检验结果都小于 0.05 了。调整后的 R 方也有轻微的提高。说明这次是一个合理的线性回归。

Variables Entered/Removed(a)

Model	Variables Entered	Variables Removed	Method
1	服务, 口味(b)	.	Enter
a. Dependent Variable: 点评			
b. All requested variables entered.			

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.941(a)	.885	.881	.184920
a. Predictors: (Constant), 服务, 口味				

ANOVA(a)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	16.562	2	8.281	242.161	.000(b)
	Residual	2.154	63	.034		
	Total	18.716	65			
a. Dependent Variable: 点评						
b. Predictors: (Constant), 服务, 口味						

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.830	.317		-5.770	.000
	口味	.468	.103	.567	4.523	.000
	服务	.271	.088	.388	3.099	.003
a. Dependent Variable: 点评						

图 22 点评与口味、服务的线性回归

4.5 K-means 聚类分析

前面我们使用 C5.0 的方法来进行分类，但是那种是通过不同的影响因素的信息熵一层一层向下细分，最后可能会分出很多类。这里我们采用 K-means 的聚类方法，相当于是在多维空间中根据不同点分布的情况来分类，这样的分类还是很有必要的，可以从一种比较宏观的角度来分类。

但是这里聚类的类别数是需要自己指定的，最开始我们也不知道要分为几类合适，这就需要经过多次尝试。在这里我分别试了 2 类、3 类、4 类。我认为**分为 4 类**时的结果比较有实际意义，所以最后采用聚类为 4 类。而输入我采用了 5 维数据，没有采用“点评”这一维，这是由于在 4.4 中我们看到“点评”与“口味、服务”两个维度的数据有着很强的线性关系，所以就没有再列为输入因素之一。

输入：口味、服务、环境、点评条数、人均价

分析：可以看到聚类的结果还是比较好的，这四类中分别有 50、13、2、1 个样本。

从图 24 中可以看到在聚类-1 中，聚类中心的服务、口味、环境都非常高、人均消费额也是比较高的，点评条数相对也较高。因此，我们可以把聚类-1 定位为**高档火锅餐厅**。聚类-4 的服务、口味、环境的评分都比较低，人均消费额也很低，点评条数属于中等偏低的，因此这一类可以定位为**中低档火锅餐厅**。聚类-3 的服务、口味、环境评分都还不错，处于中等水平，但是口味还是相对较为突出的，人均消费额中等偏低，但是可以注意到其点评条数特别高，这一类可以定位为少数那些比较**经济实惠而且口味也还不错的特别受欢迎的火锅店**。聚类-4 的服务、口味、环境都特别低，人均消费也是特别低，点评条数也很少，这一类可以定位为那些**少数的特别差的不受欢迎的火锅餐厅**。

可以看到，属于聚类-1 的火锅餐厅最多，有 50 家，其次是有 13 家的聚类-4，说明上海的高档火锅店还是很多的，而且很受欢迎，其次中低档也较多，而且受欢迎程度也还可以。聚类-3 说明有少数口味好的中低档火锅餐厅也可以赢得极大的人气，这一点应该引起关注，看来中低档火锅如果注重口味的话还是能有很大市场的。而聚类-2 这种低档次的火锅店在上海是不太受欢迎的，虽然消费水平低，但是服务、口味、环境也很低造成人气低下，说明上海的顾客在选择火锅店时还是比较注重体验和品质的，宁可多花一些钱来获得好的体验。

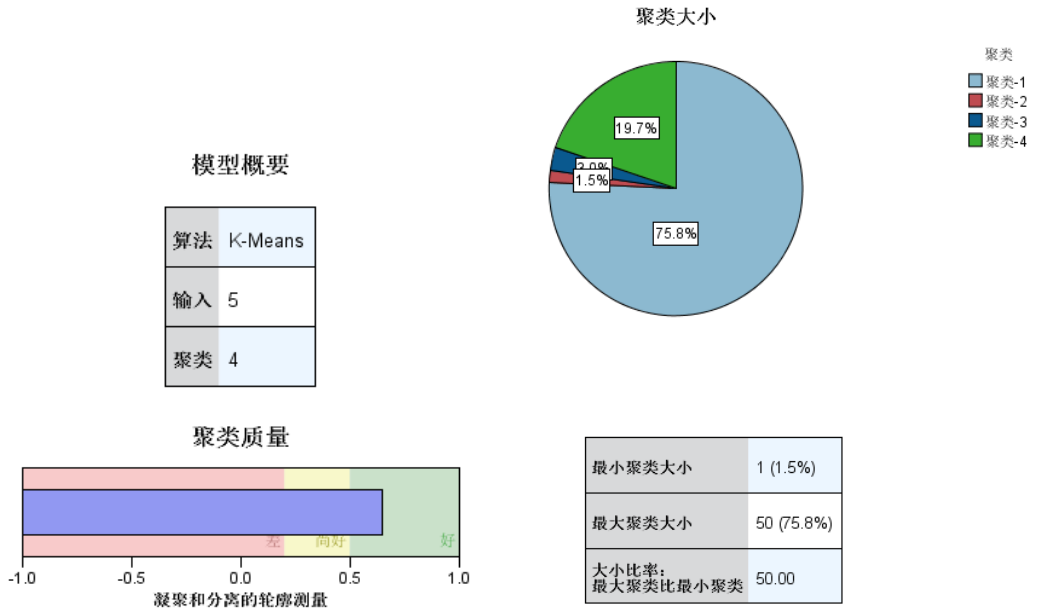


图 23 聚类模型总体概况

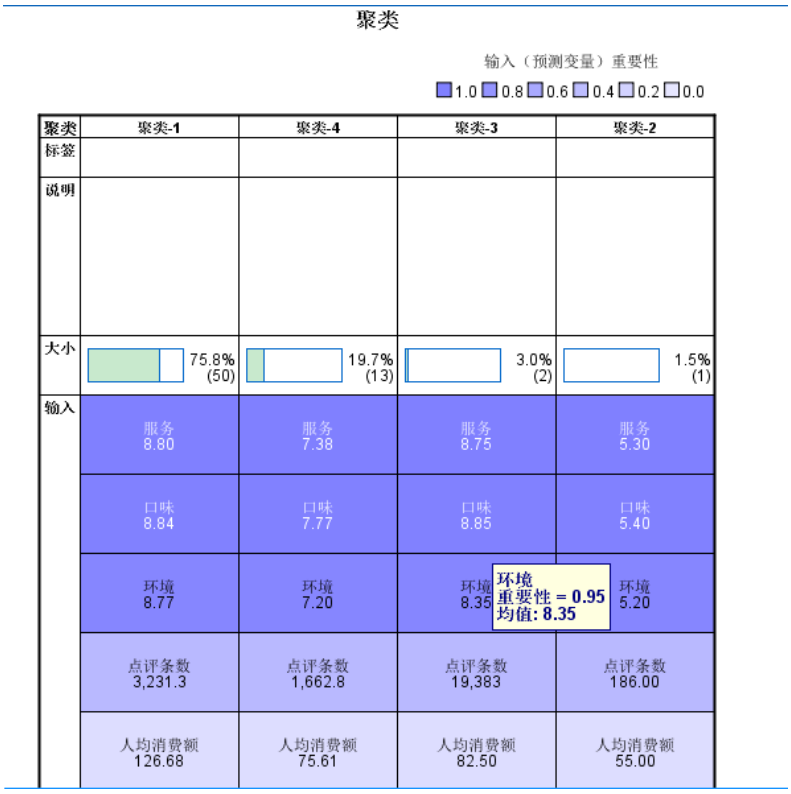


图 24 聚类中心数值分布

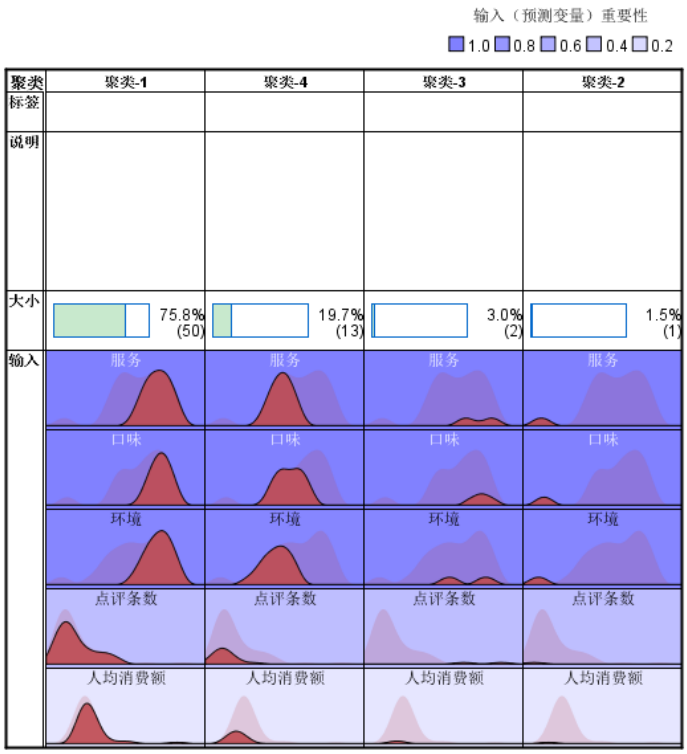


图 25 聚类模型中不同类别的分布情况

这里为了更直观地观察聚类后的分布，我借助 Lumira 工具中的散点图来展

示不同店家的分布情况。

首先我选取了聚类结果中影响最高的两个维度—服务和口味作为横纵坐标，结果如图 26 所示，其中蓝色是聚类-1，绿色是聚类-2，黄色是聚类-3，橙色是聚类-4。

可以看出，除了聚类-1 与聚类-3 分界线不明确之外，其它几类的聚类情况还是很不错的。而聚类-1 与聚类-3 的分界线可能在其它维度有所体现。之后我又采用环境和点评条数作为横纵坐标，分布情况如图 27 所示，这印证了我的想法，这里可以看到聚类-1 与聚类-3 分界线就很明确了。之后我还使用了人均消费额和环境作为横纵坐标来进一步展示。

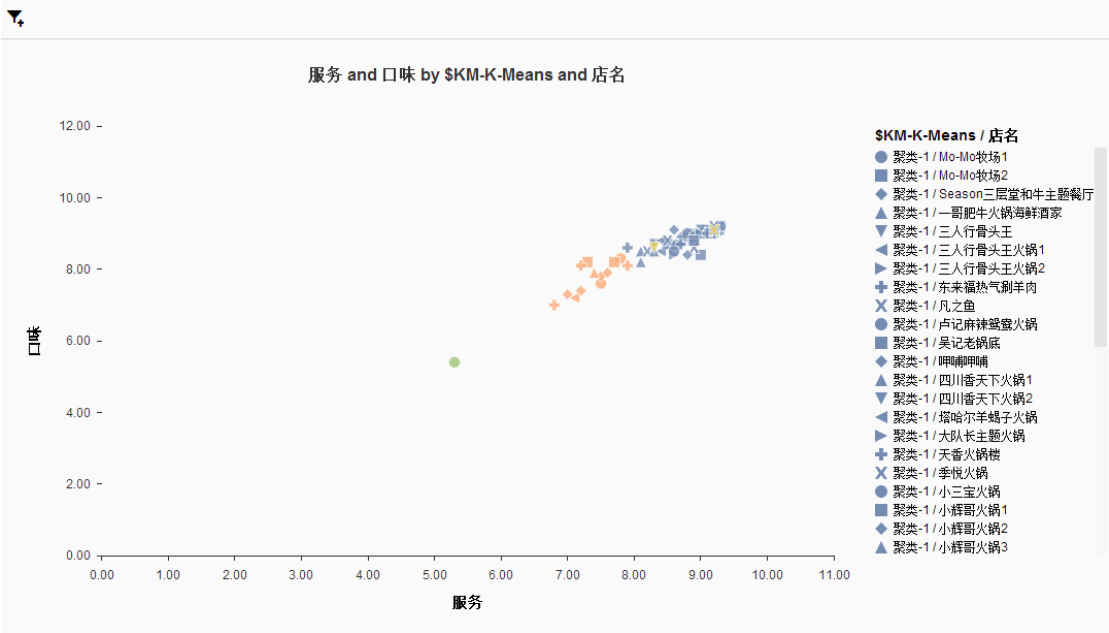


图 26 服务和口味作为横纵坐标的散点分布

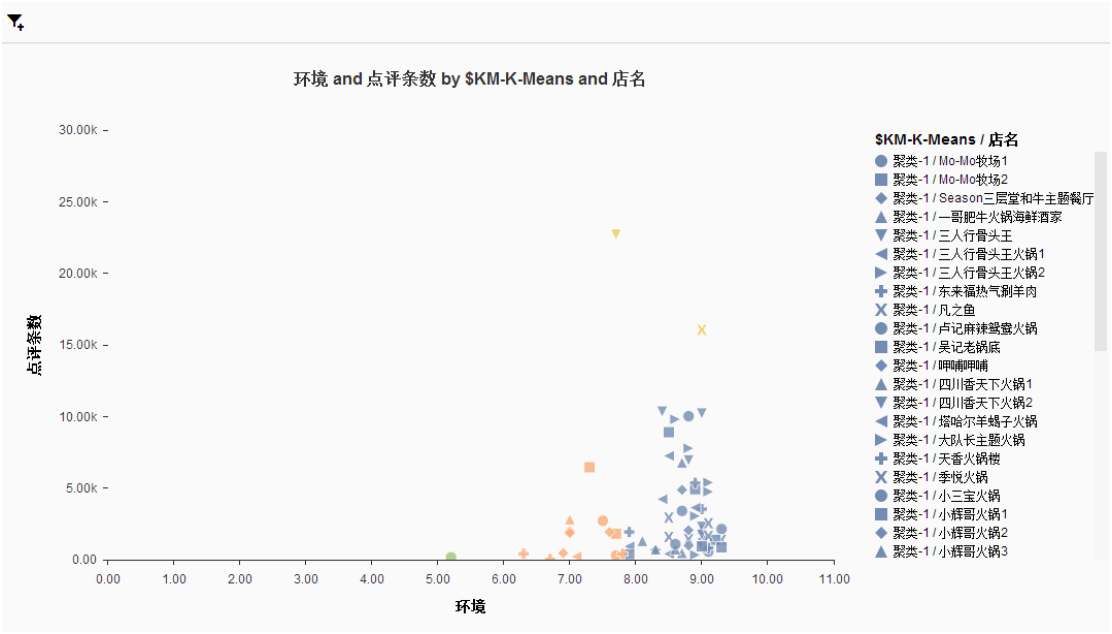


图 27 环境和点评条数作为横纵坐标的散点分布

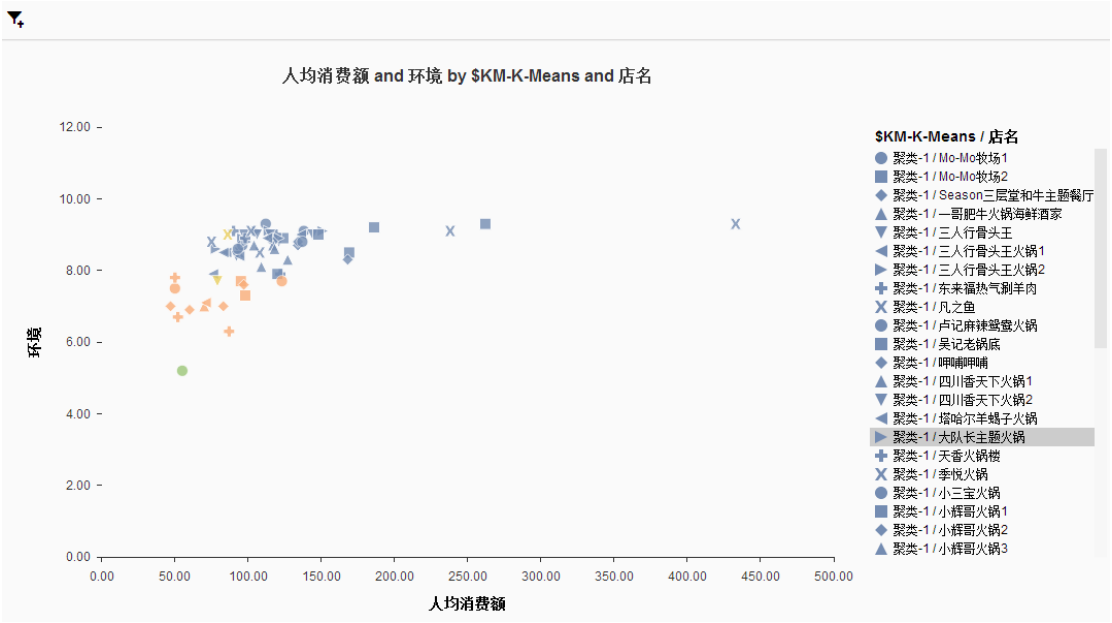


图 28 人均消费额和环境作为横纵坐标的散点分布

4.6 对标签的分析

在收集数据的时候，我们把标签一起采集进来了，但是排除了一些信息量过少的标签，比如那些只有个别店家才有的标签。标签是通过从用户的评论中提取出来的，因此一定程度上反映了用户是出于什么原因选择这家火锅店，或者说用户最看重这家火锅店的什么特点。所以我认为对标签的分析还是有一定意义的。

对于标签的分析，我没有像处理其他维度那样先进行离散化分为几类，再作为可能的影响因素来分析，其中的一个因素是考虑到标签数据中有很大一部分是0。

对于标签的分析，我使用了 Lumira 工具。

4.6.1 数据预处理

由于标签是从评论中提取出来的，我认为直接拿收集来的标签个数并不能反映出真实的信息，因为有的店评论量很多，自然该标签的数值就会大，而我更想关注的是来这家火锅店的顾客中有多大比例的人是注重“无线上网”的，多大比例是注重了“免费停车”的条件，多大比例是注重“可以刷卡”这一因素等等。所以我在预处理中，都将标签数值除以点评条数，以得到相对值。这里为了防止信息损失过大，我又乘上 100 将比值扩大了 100 倍，如图 29 所示。

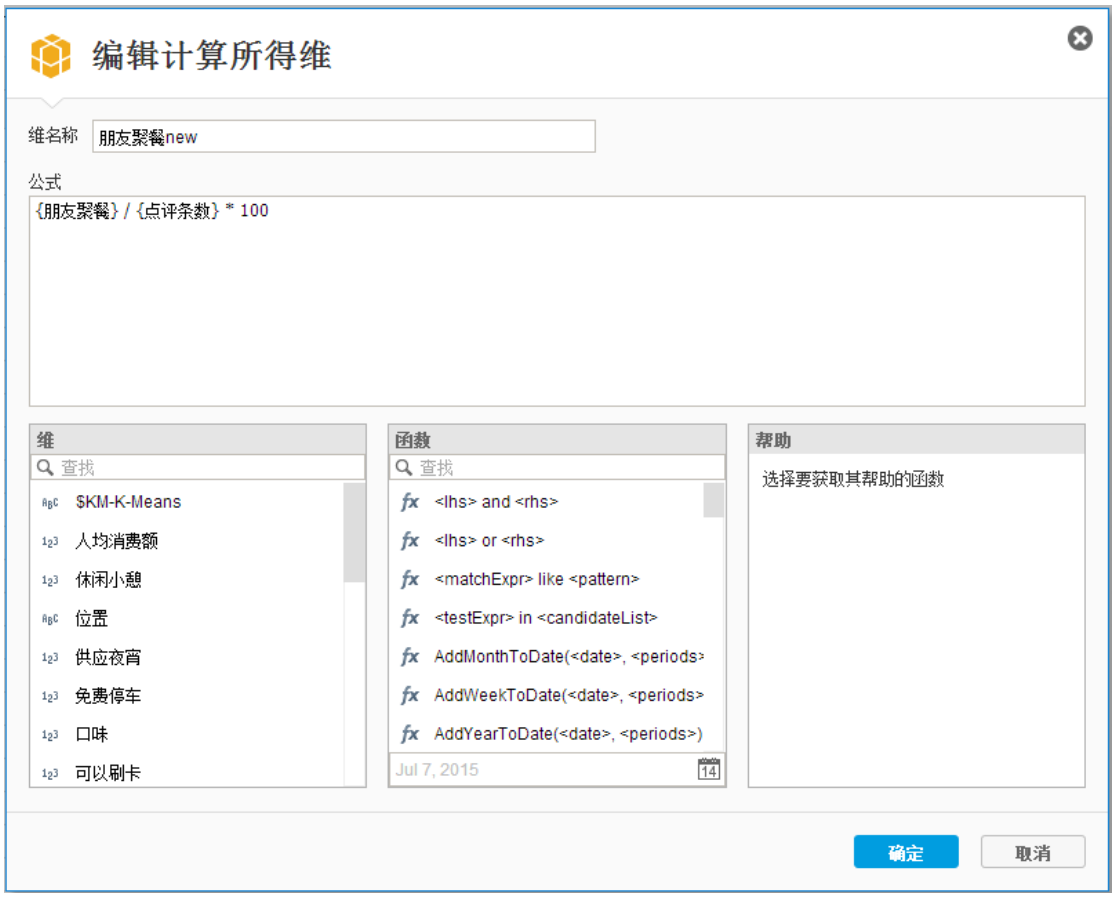


图 29 标签的数据预处理

4.6.2 标签分析

首先，对所有标签进行一个总体的统计，如下图所示。可以看到人们比较在意的几点是“可以刷卡”，其次“无线上网”也有一定的影响。而来吃火锅的动机最主要的是“朋友聚餐”，其次是“家庭聚会”和“情侣约会”。这一点也符合火锅的传

统定位。

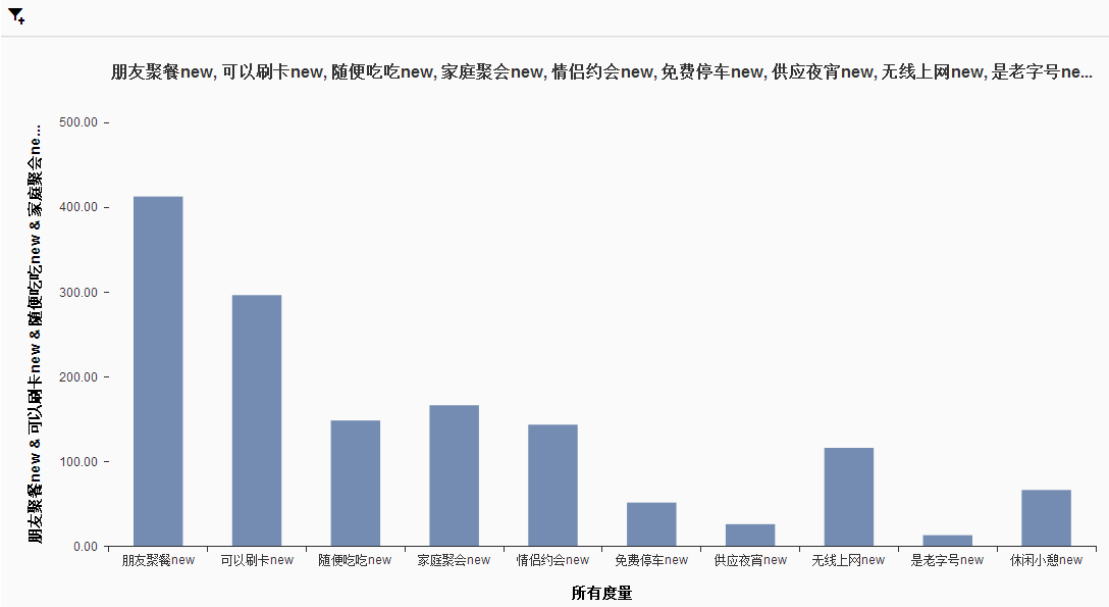


图 30 标签整体统计图

之后，我采用四个不同的聚类作为四个维度来对标签进行统计。结果如图所示。需要**注意**的一点是不同聚类之间的标签统计值大小不具有可比性，因为不同聚类分类中的样本个数不同。我们知道，不同的四个聚类代表了四种不同类型的火锅店。通过下图，我想分析的是在四种类型的火锅店中人们分别主要关注该店哪些方面。

可以看到，在高档餐厅中人们会非常在意“可以刷卡”，而是否可以刷卡在其它几类中的体现就不如在聚类-1 中突出了。同时，“无线上网”也是只有在高档餐厅中比较被注重。“朋友聚餐”在各个聚类中都被重视，这是由于火锅传统定位的原因，因此不做过多分析。而在中低档火锅店中，“随便吃吃”、“休闲小憩”的比重较高，看来大家对中低档火锅店还是没有足够的钟爱。

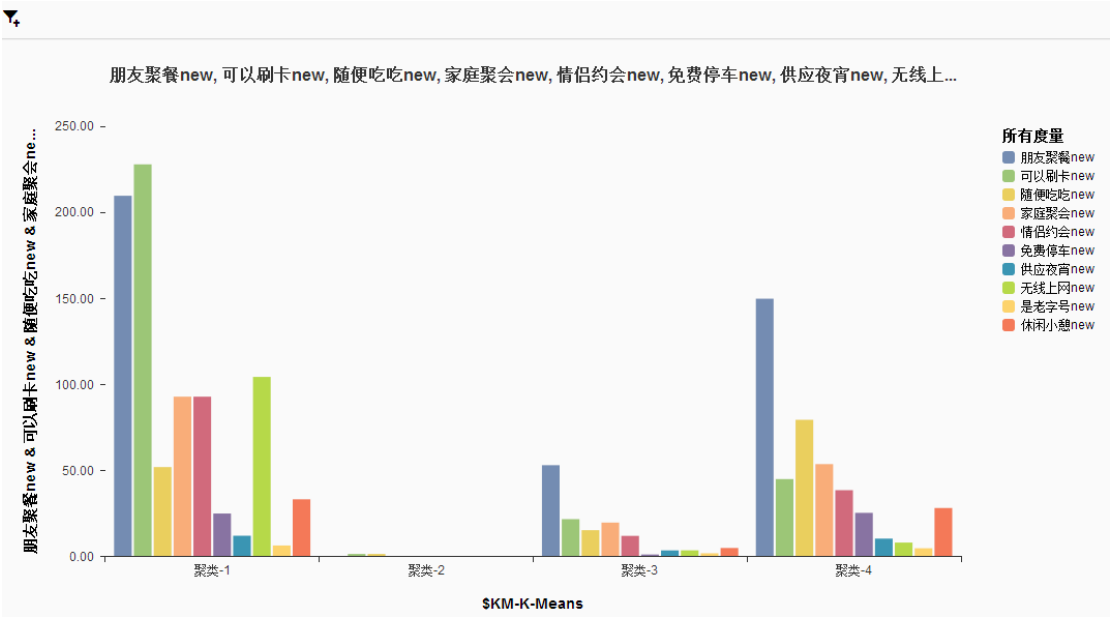


图 31 基于四个聚类维度对标签的柱状统计图

之后，基于位置维度，我又对这些标签进行了统计。也可以发现很多有趣的现象，如图 32 所示。

（1）“朋友聚餐”唯一在陆家嘴商圈的关注度非常低，而在陆家嘴商圈最受关注的是“可以刷卡”和“无线上网”以及“随便吃吃”，同时“家庭聚会”、“休闲小憩”和“情侣约会”在这里也很低，可见在快节奏的金融白领工作区，大家常常是一边上网一边自己随便吃吃。针对此情况，可以在陆家嘴开一些类似于快餐的单人小火锅店，同时要提供 WIFI 和刷卡功能。

（2）“是老字号”这一因素只有在徐家汇和淮海路两个地方关注度较高，其次在虹桥也还可以，而这两个地方都是上海曾经的繁华地带，现在也是一些较为高档、有格调的消费区，因此可以推测这里的人比较关注是否为老字号。所以一些品质格调较低的火锅店还是考虑一下重新找个合适的地盘吧。

（3）“免费停车”在张江的受关注度十分突出，在其他几个区都不太明显。因此位于张江的火锅店可以考虑增加这一便利条件来吸引顾客。

（4）可以看到“情侣约会”主要是集中在淮海路、人民广场等地，因此，这里的火锅店也可以考虑增加一些浪漫氛围，或者一些情侣主题火锅店等，应该还是会有一定的市场。

但是需要注意的是，这里由于使用的样本较少，得出的一些结论可能并不具有广泛的代表性，反映的信息也不够全面。此处只是根据已有的这一些信息进行了一个简单的分析。

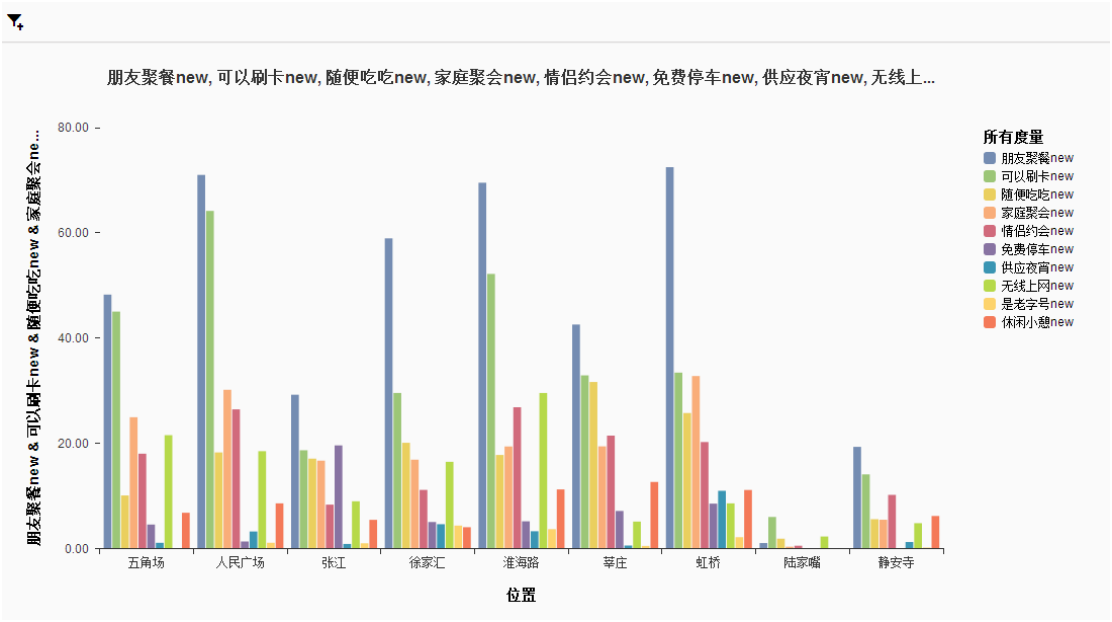


图 32 基于位置维度对标签的柱状统计图

5 总结

本次实验，通过对大众点评网上的实际数据进行分析处理的实际例子，让我对电子商务中的数据分析有了一个初步的感受。

在实验中，我分别使用了 C5.0 决策树、Apriori 关联算法、线性回归、K-means 聚类算法对数据的不同方面进行了分析。达到了实验目的中的四条。分析了影响点评条数和点评分数的关键因素，并分析了不同影响因素之间的关系。之后使用聚类算法对火锅店进行分类，并对分类结果进行分析总结。最后，还对火锅店的标签进行了简单的统计和分析。