咖啡厅经营成功的秘密

——基于大众点评网站数据的北京市咖啡厅客流量影响因素分析

董心诣 金融科技21 2021310447

目录

一、背景介绍与研究问题

1.1 背景介绍：线下餐饮的线上营销（020）模式

O2O（Online To Offline），即将线下商务的机会与互联网结合在一起，让互联网成为线下交易的前台。这样线下服务就可以用线上来揽客，消费者可以用线上来筛选服务，还有成交可以在线结算。该模式最重要的特点是：推广效果可查，每笔交易可跟踪。

“餐饮O2O”是指利用互联网的方式把线上的用户引流到线下为餐饮商家带去客源，增加收入;或把线下的顾客引流到线上进行维护或客户关系管理，以延长顾客的消费周期，提高顾客的消费次数。受到之前新冠疫情带来的消费习惯改变的影响，目前我国餐饮 O2O 市场处于高速增长的阶段。2020 年我国餐饮 O2O 市场规模接近 1.87 亿元。占整个 O2O 市场规模的 71.2%。餐饮 O2O 在整个餐饮行业市场规模的渗透率加速上涨至 47.37%。

在餐饮O2O行业中，大众点评是最早进入并开展相关业务的互联网平台，取得了一定的领先优势。在**消费端层面**，随着生活节奏的加快，消费者外出就餐的需求也随之增长，大众点评已经成为越来越多人吃饭找餐厅的必备软件。**商家端层面，**餐饮商户愿意通过入驻大众点评平台的方式进行营销和推广，提高餐厅的知名度，将顾客引流至线下餐厅进行消费。

1.2 研究问题：咖啡厅客流量的影响因素分析

对于咖啡厅商家的角度来说，影响一家咖啡厅客流量的因素有很多，比如品牌知名度、地理位置、价格等，因此可以通过分析大众点

评排行榜热门咖啡厅的数据信息，找到哪些因素会为咖啡厅带来更多的客流量，对咖啡厅的经营产生影响。因此，本报告收集了大众点评网站上北京市热门咖啡厅排行榜 top75 的咖啡厅相关数据，对影响咖啡厅客流量的因素进行研究。

二、数据来源和相关说明

2.1 数据获取

本报告使用的是大众点评网站上北京市咖啡厅排行榜 top75 的咖啡厅相关数据，数据获取方式为网络爬虫，数据采集时间范围为 2023 年 4 月 1 日至 5 月 1 日。数据类型可分为咖啡厅店铺信息和评论信息。

咖啡厅店铺信息包括北京市咖啡厅排行榜前 75 家的咖啡厅的总评论数量、人均消费、类别标签、位置标签、团购信息和评分。数据来源如下图所示：

图 2-1：大众点评商家数据信息

评论信息指的是北京市咖啡厅排行榜前 75 家的咖啡厅在 2023 年 4 月 1 日至 5 月 1 日一个月内新增的评论数，即月评论数。数据来自大众点评网的评论界面。本报告探究的因变量是客流量，希望

获取的是商家历史订单量的时间序列数据，但历史销量数据无法获取。考虑到美团点评平台对日常到店餐饮场景的高渗透率，假定单位时间内到点评评论的人数/到店消费人数的比率保持稳定，因此可以用按时间加总的评论量来近似到店餐饮消费的人数。同时，而带有时间标签的评论数据可以在大众点评网上直接浏览到，可以通过对网络源代码爬虫获取。

图 2-2：咖啡厅的评论信息界面

2.2 相关说明

数据共包括 7 个变量，其中 4 个为定量变量，3 个为定性变量。因变量为月评论数，用来近似咖啡厅的月客流量，其他变量为自变量。

表 2-1:数据变量说明表

变量类型	变量名	说明	取值范围	备注
因变量	月评论数	单位：条	0～330	
自变量	总评论数	单位：条	428～10974	
	人均消费	单位：元	26～218	
	评分		4.0～5.0	
	类别	定性变量：		建模时处

	共 5 个水平		理为是否为综合型咖啡厅建模时处理为是否位于热门商区
位置	定性变量：共 37 个水平		
是否提供团购优惠	定性变量：共 2 个水平	1 代表提供团购优惠，0 代表不提供	38.67%的咖啡厅提供团购优惠

三、描述性分析

3.1 因变量：咖啡厅的月评论数

本报告选取了 75 家咖啡厅店铺数据以及评论数据，下表展示了因变量即咖啡厅的月评论数的数值特征。

表 3-1：咖啡厅月评论数的数值特征

变量名	平均数	中位数	最小值	最大值	标准差	偏度	峰度
月评论数	45.826	29	0	330	55.207	3.124	11.841

在中心趋势方面，月评论数的平均数为 45.826，中位数为 29，

中位数小于平均数，说明样本中 50%的咖啡厅的月评论数小于平均值。

在离散程度方面，最小值为 0，即本月没有新增评论，所对应的咖啡厅是“糖房咖啡(东四店)”，最大值为 330，所对应的咖啡厅是“TIGERS 老虎西餐(亦庄店)”，极差为 330，标准差为 55.207，说明月评论数比较分散。

在分布形态方面，偏度为 3.124 大于 0，说明呈现右偏分布，说明样本中存在少数“网红”咖啡厅，吸引许多网友前来打卡，使其月评论数远超过大部分咖啡厅。峰度为 11.841，相比正态分布，月评论数呈现尖峰厚尾的特征。总体来说，咖啡厅的月评论数存在较大差距，反映大众点评榜单上 top75 咖啡厅存在客流量方面存在较大差距。

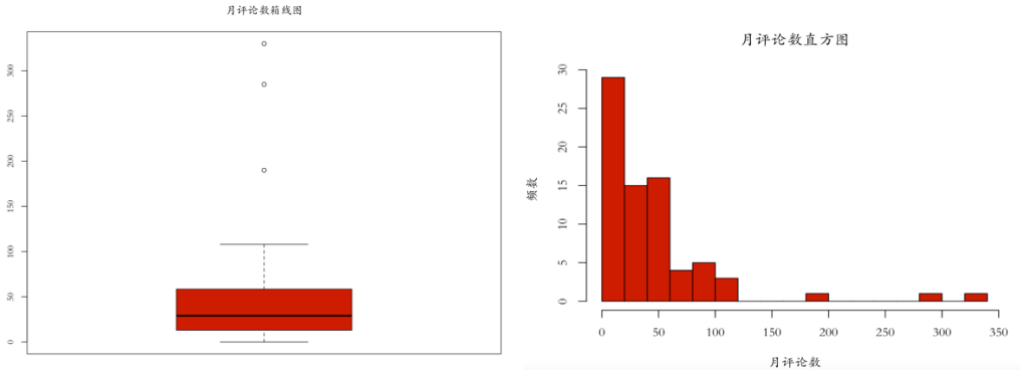

图 3-1：咖啡厅月评论数据的分布情况

3.2 自变量：

3.2.1 定量自变量

本次报告中选择的定量自变量有咖啡厅的总评论数、人均消费以及评分，下表展示了这些定量自变量的数值特征。

表 3-2：咖啡厅的总评论数、人均消费以及评分的数值特征

变量名	平均数	中位数	最小值	最大值	标准差	偏度	峰度
总评论数	2562.187	1885	428	10974	1853.736	1.812	4.489
人均消费	66.520	58	26	218	30.009	2.422	8.034
评分	4.433	4.5	4	5	0.300	0.052	-0.373

接下来，考察月评论数据与每个自变量的相关关系，计算因变量与自变量之间的相关系数并绘制散点图。其中，咖啡厅的总评论数与月评论数的相关系数为 0.571，人均消费与月评论数的相关系数为 0.485，评分与月评论数的相关系数为 0.284。说明总评论数、人均消费与月评论数的线性相关程度较高，评分与月评论数的线性相关程度较低。

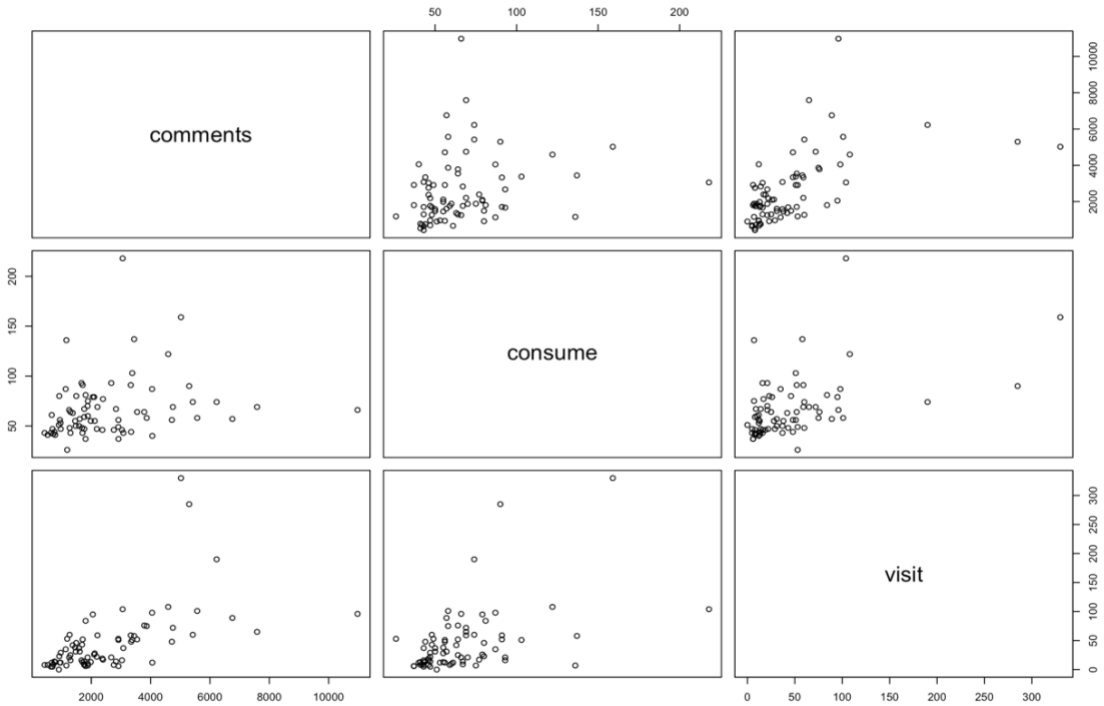

图 3-2：咖啡厅总评论数、人均消费与月评论数的散点图矩阵

3.2.2 定性自变量

本次报告中选择的定性自变量有：咖啡厅种类、是否为综合型咖啡厅、咖啡厅位置、是否位于热门商区、是否提供团购优惠。下图展示了这些定性自变量的分布情况。

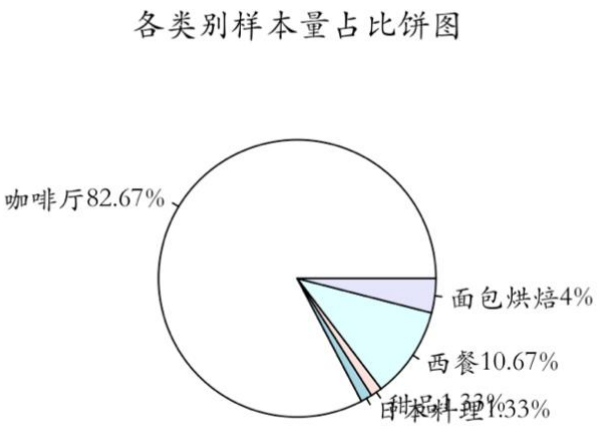

图 3-3：咖啡厅各类别样本量占比饼图

（1）对于咖啡厅种类，由于大众点评将一些提供咖啡产品的西餐、面包店等也加入了咖啡厅榜单当中，但是这些餐厅的在网站显示的种类标签并不是“咖啡厅”，而是“西餐”、“面包烘焙”等，从饼图中可以看出这些非咖啡厅标签的样本量较少，因此可以加总归入“综合型咖啡厅”，该定性变量转化为“是否为综合型咖啡厅”的虚拟变量进行研究。

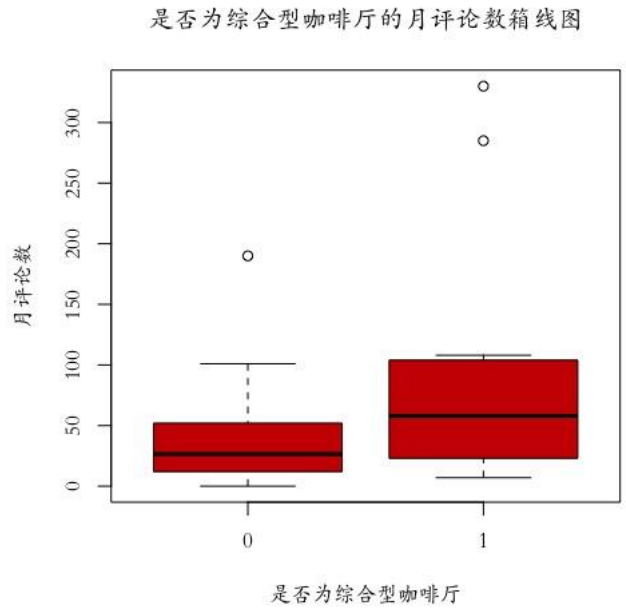

图 3-4：是否为综合型咖啡厅的月评论数分组箱线图

（2）对于是否为综合型咖啡厅，从分组箱线图中可以看出是否为综合型咖啡厅对月评论数的影响，即综合型咖啡厅的平均月评论数比单一的咖啡厅的平均月评论数高。可能的原因在于综合型咖啡厅所提供的美食和饮品种类更多，可以吸引更多消费者前来消费，这个初步结论符合大众的预期。

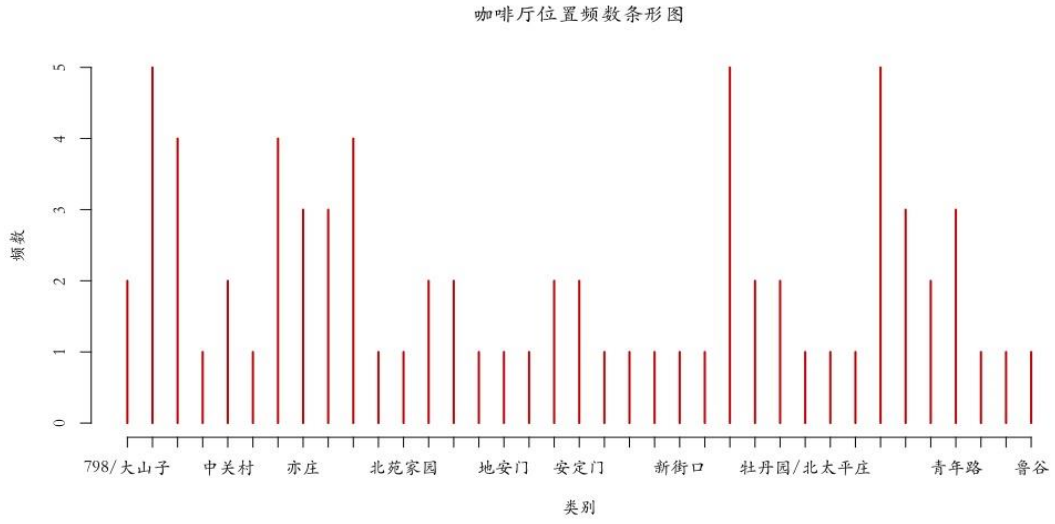

图 3-5：咖啡厅位置频数条形图

（3）对于咖啡厅位置，从上图可以看出 75 个咖啡厅样本一共涵盖了北京 37 个不同地区的咖啡厅。同时，大众点评网站还对北京市的热门商区进行了标记（见图 3.6），因此我们可以根据咖啡厅的位置判断其是否属于北京的热门商区，因此该定性变量转化为“是否位于热门商区”的虚拟变量进行研究。

图 3-6：大众点评网站显示的北京热门商区

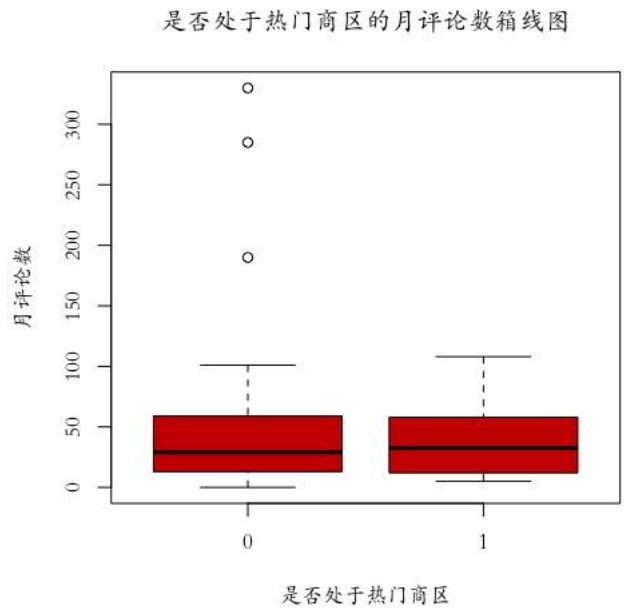

图 3-7：是否位于热门商区的月评论数分组箱线图

（4）对于是否位于热门商区，从上图的分组箱线图来看，热门商区对于平均月评论数的影响似乎不大，即平均数相差不大。不过想要得到确切的结论，需要进行下一步的建模分析。

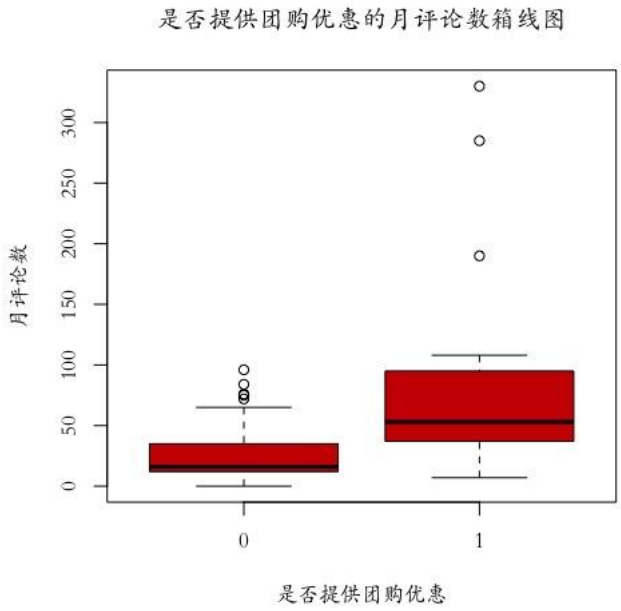

图 3-8：咖啡厅是否提供团购优惠的月评论数分组箱线图

（5）对于是否提供团购优惠，从上图的分组箱线图来看，团购优惠对于平均月评论数有较大影响，即提供团购优惠的咖啡厅的平均月评论数高于不提供团购优惠的咖啡厅。可能的原因在于提供团购优惠能吸引更多消费者前来消费，这个初步结论符合大众的预期。

四、统计建模

4.1 多元线性回归模型：咖啡厅客流量影响因素分析

4.1.1 模型

为分析咖啡厅月客流量的影响因素，本报告选取了多元线性回归模型进行建模，并建立回归模型如下：

$$\mathbf{visit} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 comments} + \boldsymbol{\beta_2 consume} + \boldsymbol{\beta_3 star} + \boldsymbol{\beta_4 D_{shop}} + \boldsymbol{\beta_5 D_{class}} + \boldsymbol{\beta_6 D_{promo}} + \boldsymbol{\mu}$$

使用最小二乘法进行参数估计，结果如下表所示：

表 4-1：咖啡厅月评论数的多元线性回归模型估计结果

变量	系数估计	t 检验 p 值	备注
截距项	-190.144	0.005*	
累计人气（总评论数）	0.014	<0.001*	
人均消费	0.558	0.005*	
评分	35.827	0.017*	
是否位于热门商区	-19.979	0.046*	基准组：不位于热门商区
是否为综合型咖啡厅	1.307	0.934	基准组：非综合型咖啡厅
是否提供团购优惠	25.495	0.013*	基准组：不提供团购优惠
F 统计量	15.890	<0.001	
R^2:0.584	调整后的R^2：0.547		

代入回归系数可得估计的回归方程如下：

$$\widehat{visit} = -190.144 + 0.014 \times comments + 0.558 \times consume + 35.827 \times star + 19.979 \times D_{shop} + 1.307 \times D_{class} + 25.495 \times D_{promo}$$

4.1.2 模型结果分析

针对模型结果的拟合优度、F 检验、t 检验和结构参数β解读如下：

（1）R^2为 0.584，说明回归方程能解释数据总变异的的 58.4%，拟合程度尚可。

（2）F 检验的 p 值小于 0.05，说明回归方程整体显著，至少有一个自变量对因变量咖啡厅的月评论数有显著影响。

（3）对于由总评论数反映的累计人气（comments），t 检验的 p 值小于 0.001，说明回归系数显著不为 0，累计人气对月评论数存在显著影响。控制其他因素不变时，总评论数每提升 1 个单位，月评论数平均增长 0.014 条。

（4）对于人均消费（consume），t 检验的 p 值为 0.005，小于 0.05,说明回归系数显著不为0,人均消费对月评论数存在显著影响。控制其他因素不变时,人均消费每提升 1 元,月评论数平均增长 0.558 条。

（5）对于评分（star），t 检验的 p 值为 0.017，小于 0.05，说明回归系数显著不为 0，评分对月评论数存在显著影响。控制其他因素不变时，评分每提升 1 个单位，月评论数平均增长 35.827 条。

（5）对于是否位于热门商区（shop），t 检验的 p 值为 0.046，小于 0.05，说明回归系数显著不为 0，是否位于热门商区对月评论数存在显著影响。控制其他因素不变时，平均而言，位于热门商区比不位于热门商区的咖啡厅的月评论数少 19.979 条。

（6）对于是否提供团购优惠（promo），t 检验的 p 值为 0.013，

小于 0.05，说明回归系数显著不为 0，是否提供团购优惠对月评论数存在显著影响。控制其他因素不变时，平均而言，提供团购优惠比不提供团购优惠的咖啡厅的月评论数多 25.495 条。

4.1.3 回归模型诊断

如下表所示，计算 6 个变量的 VIF 值均未超过 5，说明不存在多重共线性问题。

表 4-2：自变量的 VIF 值

变量名	VIF 值
总评论数	1.099
人均消费	1.823
团购	1.277
评分	1.033
热门商区	1.091
综合型	1.929

4.2 月新增评论数和月营业额预测

通过本报告的回归方程，代入一家咖啡厅的总评论数、评分和人均水平等参数，可以估计出店铺的大众点评月评论数增量。以下图所示的咖啡厅为例进行预测，代入$comments = 362$、$consume = 41$、$star = 4.5$、$D_{shop} = 0$、$D_{class} = 1$、$D_{promo} = 1$，可得预测的平均新增月评论数为 3.381。

图 4-1：大众点评咖啡厅信息页面

除此之外，已知所有到店消费的顾客大约有 15%的订单会留下评论，则可以通过以下公式计算咖啡厅营业额的提升：

$$\text{月营业额} = \text{月新增评论数} \div 15\% \times \frac{\text{总到店消费人数}}{\text{总订单数}} \times \text{人均消费}$$

4.3 文本分析：咖啡厅的热门“推荐菜”词频统计

在浏览排行榜单上咖啡厅的推荐菜标签时，发现许多咖啡厅的推荐菜有重复。猜测通过分析哪些咖啡品类在咖啡厅推荐菜上出现频率较高，以此推测这些咖啡品类更加获得消费者的喜爱。

图 4-2：大众点评网站上“推荐菜”标签

因此，本报告爬取了大众点评网站北京地区的咖啡厅排行榜前 75 家咖啡厅的推荐菜，每 1 家咖啡厅有 3 个推荐菜，一共 225 个标签。进行词频统计并绘制了词云图，统计结果如下：

表 4-3：咖啡厅“推荐菜”词频统计结果

单词	词频
Dirty	15
拿铁	11
澳白	10
手冲咖啡	9
美式咖啡	9
拿铁咖啡	6
提拉米苏	5
咖啡拿铁	3
污咖啡	3
美式	3
黑森林拿铁	2
青酱鸡肉三明治	2
椰子冰拿铁	2
富士山	2
老北京污 dirty	2
脏脏包	2
夜来香	2
精品手冲咖啡	2
杏仁可颂	2
苹果薄塔	2

图 4-3：咖啡厅“推荐菜”词云图

根据词频统计结果，可以发现北京地区的热门咖啡厅的主打咖啡为：Dirty、拿铁、澳白、手冲、美式；主打美食主要为提拉米苏、三明治、脏脏包和可颂等。

五、结论与商业建议

5.1 结论

本案例对 2023 年 4 月大众点评北京市咖啡厅榜单上 top75 的咖啡厅店铺信息和评论数据进行统计分析，得到如下结论：影响咖啡厅客流量的主要因素有：咖啡厅累计人气、人均消费、评分、咖啡厅位置是否位于热门商区和是否提供团购优惠。而咖啡厅的经营定位，即以出售咖啡饮品为主还是综合型餐厅，并不会对咖啡厅客流量产生显著性的影响。

5.2 咖啡厅的经营策略

基于以上结论，本报告对咖啡厅的经营策略给出建议如下：

（1）首先，对于累计人气和评分，历史累计人气和评分高的咖啡厅说明在当地积累一定的名声并且在大众点评等点评网站获得更高的曝光率，会带来吸引更多的消费者前来消费，引发“滚雪球效应”。

（2）其次，对于人均消费来看，回归结果显示控制其他因素不变，平均而言，人均消费越高会带来更多的客流量。经济学原理告诉我们，对于同一种商品价格越高，需求量越低，但是此处的人均消费对应的并不是同一种商品的价格，因此没有违背经济学原理。可能的原因在于人均消费越高，反映的是这家咖啡厅的咖啡品质越好，或者提供的菜品种类更丰富，让消费者更愿意消费。因此对于咖啡厅经营者来说，牺牲质量的低价策略，即“价格战”并不能吸引更多的客流量，应该优先提高咖啡口味和质量。

（3）同时，对于咖啡厅是否位于热门商区，和之前预测的初步结论相反，控制其他因素不变，位于热门商区比不位于热门商区的咖啡厅的平均客流量要少。可能的原因在于热门商区的咖啡厅虽然有机会接触到更多的顾客，但由于热门商区内部的咖啡厅更加密集，所以往往会面对更加激烈咖啡厅之间的竞争，不利于吸引客流量。同时考虑到热门商区的租金更高，因此对于咖啡厅经营者来说不宜把咖啡厅位置选在竞争激烈的热门商区。

（4）最后，控制其他因素不变，提供团购优惠可以带来更多的客流量。因此咖啡厅经营者通过在大众点评网站上提供团购优惠吸引

新顾客前来消费。

5.3 不足与展望

由于在通过 python 爬虫获取数据时遭到了大众点评的网站限制，本报告未能实现全部预想，还有诸多可以改进的地方。比如，样本量应该进一步扩大至 500 家咖啡厅以上。

同时，在未来的研究中可以考虑加入更多的因素，比如对每家咖啡厅的网友的评论进行爬虫并文本分析，以获取更多影响客流量的因素，也可以将模型推广到其他城市，进一步考虑不同城市对咖啡厅经营策略的影响。