

火锅团购数据分析

数据介绍：HotPoT.csv，案例数据提供了某线上团购网站的西安市区内在 2017 年 01 月 05 日这一时间节点的共 2688 个火锅团购产品的销售数据，数据说明表如下：

变量类型		变量名		详细说明	取值范围	备注
因变量	季均销量	quarterly.sales.volume	季均销量	累计销量/团购在线天数 X90	0.0~220.8	单位：个
自变量	基本信息	字段 1	团购名称	定性变量	——	后续进行了拆解
		字段 2	广告栏关键词	定性变量	——	即团购广告栏上展示的内容
		online.intervel	在线天数	连续变量		
	价格信息	price	套餐售价	连续变量	1~2599	单位：元
		discount	团购折扣力度	团购套餐售价/团购套餐原价	0.01~0.99	——
		privilege.for.new	是否钱包新客优惠	定性变量共 2 个水平	——	除价格折扣外的其他优惠
		privilege.for.sale	是否立减优惠	定性变量共 2 个水平	——	除价格折扣外的其他优惠
	地理信息	district	所属行政区	定性变量共 10 个水平	碑林、雁塔等	西安市所辖城区
	服务信息	if.nonworkdays.available	是否节假日通用	定性变量共 2 个水平	节假日通用、不通用	——
		time.duration	营业时长	连续变量	0~24	单位：小时
		if.appointment.needed	是否需预约	定性变量	需预约、无需预约	——
		level.of.participants.limited	每张团购券使用人数限制	定性变量共 4 个水平	不限、限 1 人、限 2-4 人，限 5-12 人	进行了水平合并
		level.of.tickets.limited	每次消费团购张数限制	定性变量共 2 个水平	不限张数、限张数	——
		WiFi.available	是否提供 WIFI	定性变量共 2 水平	提供 WiFi、未提及	——
		parking	停车位	定性变量共 4 个水平	免费、收费、无、未提及	——
		rooms.available	包间	定性变量共 3 个水平	仅限大厅、均可使用、未提及	——
	口碑信息	评分		连续变量	1.0~5.0	满分为 5 分
		评论数		连续变量	0~8032	——
	店名信息	Place	是否包含地名	定性变量共 2 个水平	是、否	

		Food	是否包含食材	定性变量 共 2 个水平	是、否	
		Taste	是否包含口味	定性变量 共 2 个水平	是、否	
		Cook	是否包含烹饪方式	定性变量 共 2 个水平	是、否	
		People	是否包含人物	定性变量 共 2 个水平	是、否	
		Style	菜系	定性变量 共 4 个水平	北系、川系、粤 系、其他	
		HighFreq	使用 TOP20 高频词 的个数	连续变量	0~6	

分析任务：

1. 请完成数据读入与简单清洗；
2. 将团购的季均销量转换为对数，对季均销量、团购上线天数、折扣力度绘制分布直方图，简单陈述你观察到的现象；
3. 将火锅团购商家店名高频词统计出来，并绘制出词云，简单陈述你观察到的现象；提示：在统计火锅团购商家店名高频词时，可以将一些无意义的词组剔除，如“区”、“路”、“火锅”、“小区”、“分店”、“店”等；同时还需要先剔除非中文字符（数字和字母）。
4. 用箱线图表示销量与团购价的关系，及销量与节假日通用的关系；尝试解读这两个箱线图呈现的现象；
5. 用箱线图表示店名中包含的各类词汇与团购销量的关系，并尝试解读箱线图的含义；
8. 建立线性回归模型，探究影响团购销量的主要因素，并对模型结果进行适当解读。