**数据说明：**

1. 本案例所采用数据为“data.csv”文件为北京外卖商家销量数据，数据集包含了32043条商家信息，为2017年8月从北京地区某外卖平台爬取的数据。每一条观测数据为一个商家的信息，变量包括月销量、平均每单价格、配送方式、配送费等。自变量共分为五个维度，分别为外卖商户基本信息、评分信息、位置信息、优惠信息和衍生字段信息，具体信息如下表所示。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 变量类型 | | 变量名称 | 详细说明 | 备注 |
| 因变量 | | 月销量 | 单位：单  取值范围：0-13571 |  |
| 自变量 | 基本信息 | 商家配送方式 | 定性变量  商家配送、蜂鸟专送 | 商家配送占比60.04% |
| 配送费 | 单位：元  取值范围0-230 |  |
| 起送价 | 单位：元  取值范围0-1000 |  |
| 是否新店 | 定性变量  共2个水平 | 新店占比8.5% |
| 是否品牌店 | 定性变量  共2个水平 | 品牌店占比12.07% |
| 配送时间 | 单位：分钟  取值范围0-102 |  |
| 平均每单费用 | 数值型变量 |  |
| 是否支持外卖保 | 定性变量  共2个水平 |  |
| 是否可开发票 | 定性变量  共2个水平 |  |
| 每天营业时长 | 单位：小时  取值范围0-24 |  |
| 8点是否营业 | 定性变量  共2个水平 |  |
| 12点是否营业 | 定性变量  共2个水平 |  |
| 16点是否营业 | 定性变量  共2个水平 |  |
| 23点是否营业 | 定性变量  共2个水平 |  |
| 商家分类 | 定性变量  共6个水平 | 包含快餐简餐、特色菜系、异国料理、小吃夜宵、甜品饮品、果蔬生鲜六类 |
| 评分信息 | 比较评分 | 单位：分  （百分比数据） | 指和周围的商家对比所得的得分，满分1分，即100% |
| 食物评分 | 单位：分 | 满分5分 |
| 服务评分 | 单位：分 | 满分5分 |
| 星级 | 单位：星 | 最高5星 |
| 位置信息 | 行政区划1 | 定性变量  共16个水平 | 包含北京全部16个行政区 |
| 行政区划2 | 定性变量  共7个水平 | 将北京城六区意外的其他区合并为非城区 |
| 优惠信息 | 满减折扣率 | 取值范围0-96% |  |
| 新用户减免 | 单位：元  取值范围0-45 |  |
| 其他活动数量 | 单位：个  取值范围0-30 |  |
| 衍生字段信息 | | 对数月销量 |  |  |
| 对数平均每单费用 |  |  |
| 对数配送费 |  |  |

字段对照：

1．基本信息：

delivery\_mode(商家配送方式)；float\_delivery\_fee(配送费)；float\_minimum\_order\_amount(起送价)；is\_new(是否新店)；is\_premium(是否品牌店)；order\_lead\_time(配送时间)；recent\_order\_num(月销量)；average\_cost(平均每单费用)；保(是否支持外卖保)；票(是否可开发票)；open\_time(每天营业时长，单位：小时)；open08(8点是否营业)；open12(12点是否营业)；open18(18点是否营业)；open23(23点是否营业)；shop\_type(商家分类)

2．评分信息：

compare\_rating(比较评分，通过和周边商家对比得到的得分)；food\_score(食物评分)；service\_score(服务评分)；star\_level(星级)；

3．位置信息

district(行政区划1，包含北京全部16个行政区)；district0(行政区划2，其中将北京城六区外的其他区合并为非城区)；

4．优惠信息

manjian\_discount(满减折扣率)；new\_user(新用户减免)；other\_activities(其他活动数量，除了满减和新用户减免外的其他活动)；

5．衍生字段信息

log\_recent\_order\_num(对数月销量：月销量加1再取对数)；log\_average\_cost(对数平均每单费用：平均每单费用加1再取对数)；log\_float\_delivery\_fee(对数配送费：配送费加1再取对数)；

**分析任务：**

1. 读入数据，查看数据情况；

2. 绘制对数月销量的直方图，观察其特征，并作出解释；

3. 绘制行政区划与月销量对比箱线图，并作出解释（要求箱线图的箱子按中位数大小排列）；

4. 绘制四个用餐高峰期是否营业（open08、open12、open18、open23）与对数月销量对比箱线图（四个子图）。要求：在一张大图中放入四个子图（提示：使用基础绘图函数时可用par参数，使用ggplot2时使用facet\_wrap分面或者ggpubr包中的ggarrange函数）；

5. 建立回归模型，使用BIC准则进行逐步回归得到最终模型，查看最终模型中的系数大小及其显著性，并解读系数。要求：

1. 因变量：对数月销量
2. 自变量：对数平均每单费用、配送模式、对数配送费、对数起送费、起送时间、所在行政区（北京市16区）、是否有外卖保、是否开发票、满减折扣率、其他优惠活动数、新用户减免、每天营业时长、8点是否营业、12点是否营业、16点是否营业、23点是否营业、商家分类、是否新店、是否品牌店
3. 对于行政区划变量，使用昌平区作为基准水平；
4. 对于商家分类变量，使用果蔬生鲜作为基准水平；

6. 从模型中提取出行政区变量的系数与显著性，分别绘制系数柱状图，要求：

1. 回归系数按数值大小排列；
2. 显著与不显著使用不同颜色标注；

7. 假设你要开一家外卖店，并有如下打算，请预测其未来的（已经不是新店）月销量。

1. 在长宁区经营快餐简餐（非品牌店），目标客户是学生党，平均每单费用约为20元（可参考北京与长宁区定位相似的行政区）；
2. 使用蜂鸟专送，配送费6元，配送时间约为45分钟；
3. 不可开发票；
4. 综合评分(new\_score)约为0.65分；
5. 满减折扣率约为35%、其他活动数5个、没有新用户减免；
6. 每天营业11小时，且在18点和23点都营业。