**实验报告**

|  |  |  |
| --- | --- | --- |
| **课程名称**： **应用回归分析实验** | **学期： 2023年秋季学期** | **成绩**： |
| **指导教师**： **陈丹** | **学生姓名**： **枫叶** | **学生学号**： |
| **实验名称**： **酒店数据分析** | | |
| **实验编号**：1 | **实验日期**： | **实验学时**： |
| **学院： 数学与统计学院** | **专业： 统计学** | **年级**： **2021级** |

1. **实验目的**

应用描述统计方法对酒店数据进行分析

**二、使用环境**

R语言

**三、实验内容**

（1）背景介绍

据国家旅游局发布的数据显示，2015年中国国内旅游人次、出境旅游人次、国内旅游消费及境外旅游消费位居世界第一，庞大的旅游需求带来了庞大的酒店需求。作为消费者的我们最关心的自然是酒店的价格，一方面是对开销的考量，另一方面是直觉上来说价格更贵的酒店会拥有更好的服务和更好的设施，但是酒店的定价也可能仅是因为具有服务于特殊群体的特色而更高，这就需要消费者谨慎挑选，价高者就一定好吗？价格会因为哪些因素而高呢？本文试图基于统计分析方法探究酒店价格的影响因素，为消费者提供一定参考。

（2）数据说明

本文数据爬取自携程旅行网北京地区酒店信息，采样时间为2023年10月9日，所爬取信息为若当天入住的情况，共采集样本261条，经数据预处理后得到了酒店名称、价格、距市中心直线距离、酒店评分、标签等，下表提供了一个变量概览。

表 1

|  |  |  |  |
| --- | --- | --- | --- |
| 变量类型 | 变量名 | 详细说明 | 取值范围 |
| 因变量 | 酒店价格 | 起步价（元） | 94-1633 |
| 自变量 | 酒店名称 | 无 | 无 |
| 评分 | 所有客户评分的平均值  （5分满分） | 4-5 |
| 距离 | 距市中心直线距离  （公里） | 1.1-830 |
| 点评数 | 留下评论的人数 | 12-6499 |
| 标签 | 12个二分类变量 | 0或1 |

（3）数据分析

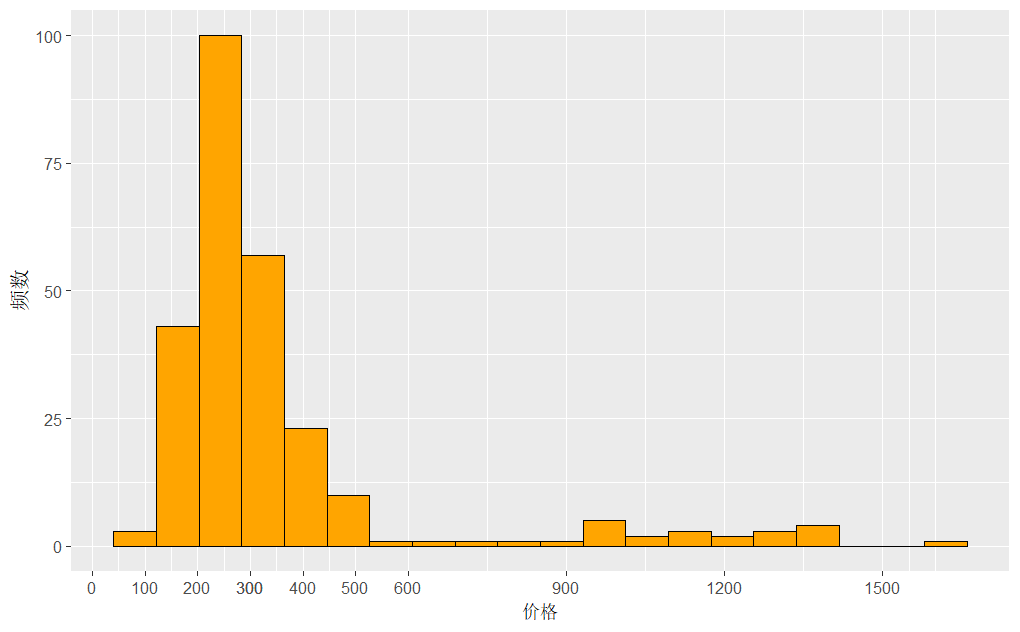


图 1

从上图可以发现大部分酒店的定价都集中在100-500元之间，100-300元又占据其中的绝大部分。



图 2

从评分和价格的散点图（图2）来看，价格有略微的随评分上升而上升的关系，但并不明显，考虑到二者数量级差距，对价格作对数处理，线性关系更明显了一些，但设定评分前20%为高分组，后80%为低分组，作出箱线图后注意到低分组价格平均来说甚至比高分组更高，进一步调整4.6分为高分标准，作出箱线图后得到相同结论。这或许说明价格贵不一定好，物美价廉才是普遍现象。箱线图可以分析不同组别之间是否存在显著差异，在这一点上比散点图更为直观。

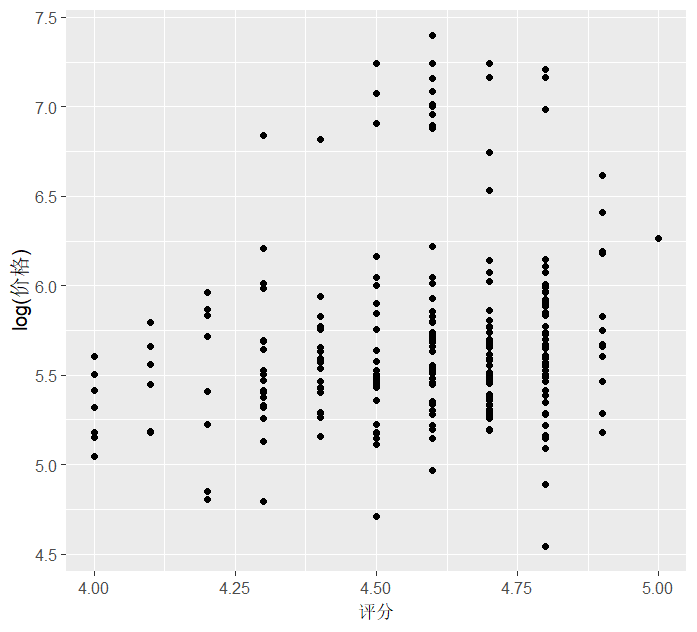


图 3

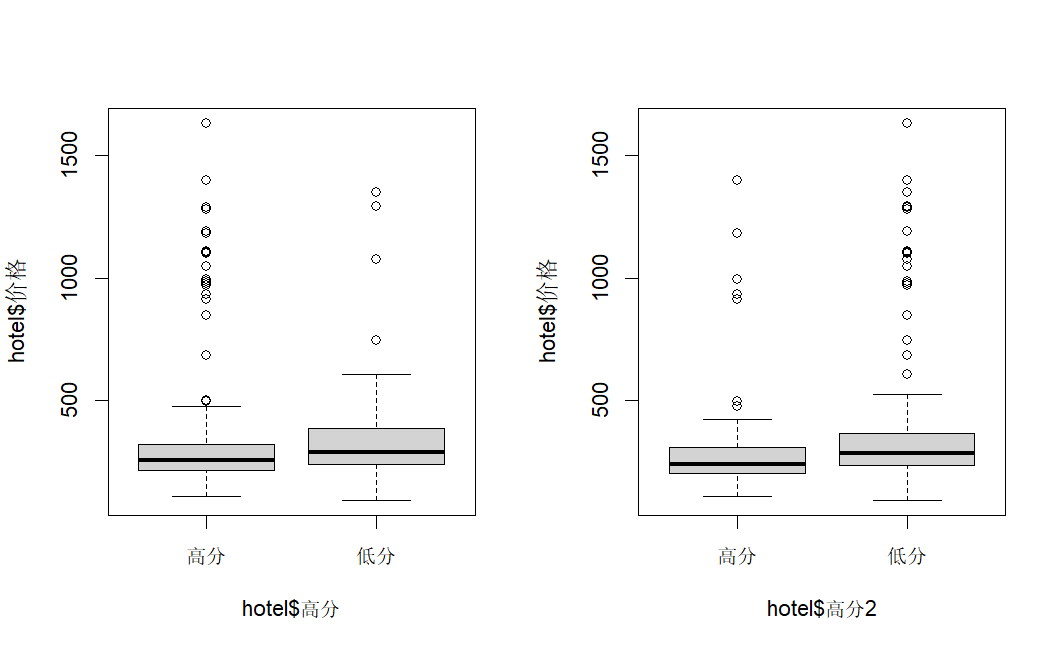


图 4

对价格和距离作出散点图（图5），发现大量酒店距市中心距离都很小，对二者均取对数处理，结果如图6所示，二者间存在一定的线性关系，且系数应该为负

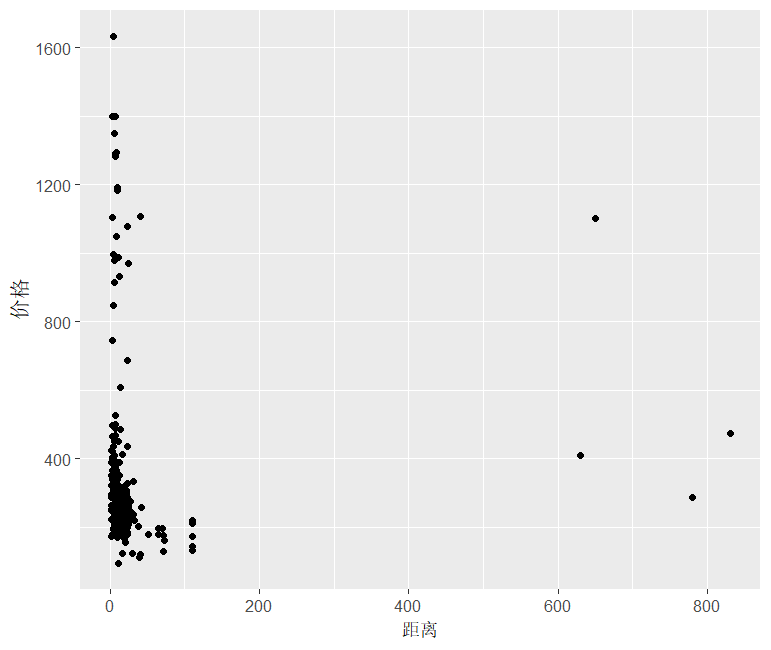


图 5

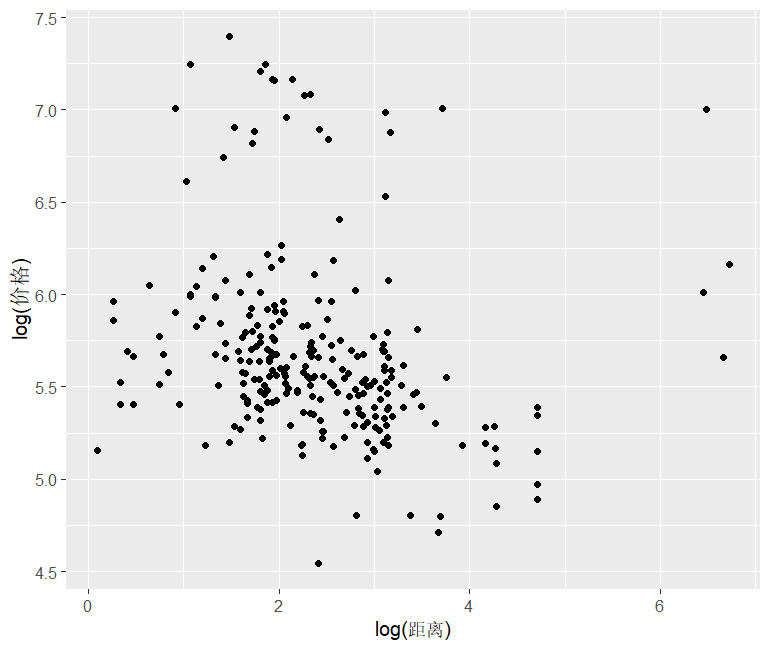


图 6

由于爬虫获取到的价格并不对应房型，对于房间类型和价格的统计图仅在理论上作出回答，我认为小提琴图或箱线图可以分析房间类型对价格的影响。作为补充，我选取酒店所列出的标签进行分析，一个酒店可以含多个标签，不同酒店之间可以有相同的标签，考虑到北京的酒店可能多为公务出差者居住，故选取了叫醒服务、茶室、投影仪等相关标签，此外前往北京旅游者也是一大酒店使用群体，故选取了亲子主题房、咖啡厅、家庭房、情侣房、茶室等相关标签，一些可以提升住宿体验的标签如花园、咖啡厅等也被选中。通过简单的回归模型，并基于AIC准则进行变量筛选，所有变量我们得出如表2所示结果

表 2

|  |  |
| --- | --- |
| 变量/模型 | log(价格) |
| 截距项 | 4.236\*\*\* |
| 评分 | 0.448\*\*\* |
| log（距离） | -0.118\*\*\* |
| 花园 | 0.165\* |
| 洗衣房 | -0.114\*\* |
| 家庭房 | -0.168\*\*\* |
| 叫醒服务 | -0.324\*\*\* |
| 情侣房 | -0.243\*\* |
| 修正 | 0.276 |

注：\*\*\*,\*\*,\*\*分别表示在 0.01,0.05,0.1显著性水平下显著

和前文结果一致，评分具有显著的正向作用，对数处理后的距离变量具有显著负向作用，此外，带花园的酒店价格更高，而带有洗衣房、家庭房、叫醒服务、情侣房的酒店价格更低，推测带花园的酒店多属于高端酒店，而带有洗衣房等标签的酒店定位于中低端消费者，价格低廉的同时更适合家庭或情侣出游。

（4）总结

大部分酒店价格集中在100-300元；距市中心的直线距离在一定程度上衡量了酒店的偏僻程度，越靠近市中心价格越贵；评分高的酒店并不一定价格更贵，也可能更加便宜实惠；回归结果事实上可以归结为带花园的高端酒店价格更贵，带有家庭房等标签的酒店价格更低，对于旅游出行来说更加友好

1. **算法介绍及结果**

|  |
| --- |
| #实验1  library(readxl)  library(ggplot2)  library(tidyverse)  library(stringi)  hotel <- read\_excel("D:/预删除文件夹/大三上/应用回归分析/酒店数据(携程).xlsx")  #数据预处理  hotel$评分 <- as.numeric(hotel$评分)  hotel$价格 <- hotel$价格 %>%  gsub("¥","",.) %>%  gsub(",","",.) %>%  as.numeric()  hotel$距离 <- hotel$酒店位置 %>%  gsub("\*.\*直线","",.) %>%  gsub("公里","",.) %>%  gsub("米","",.) %>%  as.numeric()  hotel$点评数 <- hotel$点评数 %>%  gsub("条点评","",.) %>%  gsub(",","",.) %>%  as.numeric()  hotel <- hotel %>%  rowwise() %>%  unite("标签",4:13)  hotel <- hotel %>%  mutate(across(starts\_with("标签"),~str\_split(.x, "\_") |>  map(as.vector)))  hotel <- hotel %>%  rowwise() %>%  mutate(花园 = ifelse("花园" %in% 标签,1,0)) %>%  mutate(洗衣房 = ifelse("洗衣房" %in% 标签,1,0)) %>%  mutate(亲子主题房 = ifelse("亲子主题房" %in% 标签,1,0)) %>%  mutate(咖啡厅 = ifelse("咖啡厅" %in% 标签,1,0)) %>%  mutate(无烟楼层 = ifelse("无烟楼层" %in% 标签,1,0)) %>%  mutate(家庭房 = ifelse("家庭房" %in% 标签,1,0)) %>%  mutate(叫醒服务 = ifelse("叫醒服务" %in% 标签,1,0)) %>%  mutate(免费停车 = ifelse("免费停车" %in% 标签,1,0)) %>%  mutate(情侣房 = ifelse("情侣房" %in% 标签,1,0)) %>%  mutate(套房 = ifelse("套房" %in% 标签,1,0)) %>%  mutate(茶室 = ifelse("茶室" %in% 标签,1,0)) %>%  mutate(投影仪 = ifelse("投影仪" %in% 标签,1,0))  #描述  summary(hotel)  #画图  hotel$n <- 1  ggplot(data = hotel,aes(x=价格)) +  geom\_histogram(bins=20,color="black",fill="orange") +  scale\_x\_continuous(breaks=c(100,200,300,400,500,seq(0,1800,300))) +  ylab("频数")  ggplot(data = hotel,aes(x=log(距离),y=log(价格))) +  geom\_point()  ggplot(data = hotel,aes(x=距离,y=价格)) +  geom\_point()  ggplot(data = hotel,aes(x=评分,y=价格)) +  geom\_point()  ggplot(data = hotel,aes(x=评分,y=log(价格))) +  geom\_point()  high\_line1 <- quantile(hotel$评分,0.8)  high\_line2 <- 4.6  hotel <- hotel %>%  rowwise() %>%  mutate(高分 = ifelse(评分 >= high\_line,1,0))  hotel <- hotel %>%  rowwise() %>%  mutate(高分2 = ifelse(评分 >= high\_line2,1,0))  hotel$高分 <- factor(hotel$高分,labels = c("高分","低分"))  hotel$高分2 <- factor(hotel$高分2,labels = c("高分","低分"))  par(mfrow = c(1,2))  boxplot(hotel$价格~hotel$高分)  boxplot(hotel$价格~hotel$高分2)  #回归  model <- lm(data=hotel,log(价格)~评分+log(距离)+花园+洗衣房+亲子主题房+咖啡厅+无烟楼层+家庭房+叫醒服务+免费停车+情侣房+套房+茶室+投影仪)  summary(model)  model2 <- step(model)  summary(model2) |

**六、参考文献**

**七、教师评语**