模块一 - 北美旅游产品选择攻略

张笑竹 / 201618070114

2019年7月5日

背景介绍

伴随互联网的普及，越来越多热爱旅游的驴友们开始选择在网上购买旅游产品，但是网上的旅游产品名目众多，面对一个相对陌生的旅行目的地，驴友们往往很难选择出最“物美价廉”的那一个。本案例通过对网上各种北美旅游产品的相关信息进行统计分析，帮助驴友们花更少的钱，购买最优质的北美旅游产品。

案例数据提供了2926条北美旅游数据观测，共25个变量，其中包括产品名称、旅游方式、供应商、等级、景点个数、交通情况、用餐情况、是否有自由活动、客户评分、出游人数、评价人数、报价信息、旅游线路等。

1 任务一

在任务一中，我们首先读入数据，筛选出产品名称、旅游方式、供应商、等级、景点个数、交通、用餐、自由活动、总评价、出游人数、出发地、每日报价、旅游线路变量，并重命名为英文变量名。

```
setwd('C:\\Users\\小竹子\\Desktop\\统计仿真实验\\1\\1. 模块1--探索性数据分析\\案例难度: ____\\2. 北美旅游产品选择攻略')
library(openxlsx)
travel_dat <- read.xlsx('travel.xlsx', sheet = 1, startRow = 1, colNames = T)
travel_dat <- travel_dat[,c('产品名称','旅游方式','供应商', '等级', '景点个数',
                            '交通', '用餐', '自由活动', '总评价', '出游人数', '出发地',
                            '周日报价', '周一报价', '周二报价', '周三报价', '周四报价',
                            '周五报价','周六报价','旅游线路')]
colnames(travel_dat) <- c('Product', 'TravelMethod', 'Agency', 'Star',
                          'Place', 'Traffic', 'Meal', 'FreeActivitie', 'Evaluate',
                          'Sale', 'Depart', 'SunPrice', 'MonPrice', 'Tuesprice', 'WedPrice',
                          'ThusPrice', 'Friprice', 'Satprice', 'Routine')
```

然后，去除产品名称中带有“健康医疗”以及从国内出发（出发地为上海或北京）的样本。将数据集命名为travel_dat，查看数据的前几行。

```
# 去除产品名称中带有"健康医疗"以及从国内出发（出发地为上海或北京）的样本。
library(stringr)
r1 <- as.integer(rownames(travel_dat[str_detect(travel_dat$Product,'健康医疗'),]))
travel_dat <- travel_dat[-r1,]
r2 <- as.integer(rownames(travel_dat[str_detect(travel_dat$Depart, '北京|上海'),]))
travel_dat <- travel_dat[-r2,]
head(travel_dat)
```

```
##
Product
## 1                                                                                              美国洛杉
矶+旧金山6日5晚跟团游(3钻)·七大主题项目任选二【限时特惠】含必付
## 2                                                    美国西海岸+洛杉矶+拉斯维加斯+旧金山+黄石国家
公园10日9晚跟团游(3钻)·大峡谷+羚羊彩穴+主题项目选一 含必付
## 3 美国黄石国家公园+洛杉矶+拉斯维加斯+盐湖城5日半自助游·大提顿国家公园、布莱斯国家公园·(5天)美西黄石经济游：绝美黄石
公园、秀丽大提顿国家公园、布莱斯国家公园奇景不断
## 4                                                                                             美国洛杉
矶+旧金山+硅谷5日4晚跟团游·17里湾优胜美地二选一【含主题乐园门票】
## 5                                                                                                 美国
洛杉矶+拉斯维加斯+旧金山14日13晚半自助游·1号公路+6大国家公园自驾
## 6                                                                                                 美国
拉斯维加斯+洛杉矶7日6晚跟团游(3钻)·七大主题公园任选三+携程大礼包
##   TravelMethod                    Agency          Star
## 1       跟团游           供应商：熊大国旅         5晚3钻
## 2       跟团游           供应商：熊大国旅         9晚3钻
## 3     半自助游 供应商：途风（熊大旗下） 暂无酒店信息
## 4       跟团游 供应商：Namei Group Inc.        4晚3钻
## 5     半自助游           供应商：熊大国旅 暂无酒店信息
## 6       跟团游           供应商：熊大国旅         6晚3钻
##                            Place          Traffic         Meal
## 1   共7个景点，包含7个经典景点：     暂无交通信息 暂无用餐信息
## 2 共14个景点，包含12个经典景点：  行车时长38小时 暂无用餐信息
## 3  共12个景点，包含5个经典景点：     暂无交通信息 暂无用餐信息
## 4  共11个景点，包含9个经典景点：    行车时长6小时 暂无用餐信息
## 5 共54个景点，包含15个经典景点：  行车时长33小时       42次自理
## 6   共6个景点，包含6个经典景点：     暂无交通信息 暂无用餐信息
##      FreeActivitie      Evaluate      Sale Depart SunPrice MonPrice
## 1        自由活动1次 4.3<U+00A0>分  39人出游 洛杉矶 ￥4796起 ￥4796起
## 2        自由活动1次 3.8<U+00A0>分 136人出游 洛杉矶 ￥6885起 ￥6885起
## 3 暂无自由活动信息 3.0<U+00A0>分  42人出游 洛杉矶 实时计价
## 4        自由活动1次 3.3<U+00A0>分  27人出游 洛杉矶 ￥3058起 ￥3058起
## 5 暂无自由活动信息 5.0<U+00A0>分  13人出游 洛杉矶 ￥8898起 ￥8898起
## 6        自由活动1次 5.0<U+00A0>分  27人出游 洛杉矶 ￥5047起 ￥5047起
##   Tuesprice WedPrice ThusPrice Friprice Satprice Routine
## 1  ￥4796起 ￥4496起  ￥4496起 ￥4496起 ￥4496起  西海岸
## 2  ￥6885起 ￥6585起                     ￥6585起  西海岸
## 3           ￥2970起  ￥2970起           ￥2970起  西海岸
## 4  ￥3058起 ￥3058起  ￥3058起 ￥3058起 ￥3058起  西海岸
## 5  ￥8898起 ￥8898起  ￥8898起 ￥8898起 ￥8898起  西海岸
## 6  ￥5047起 ￥4747起  ￥4747起 ￥4747起 ￥4747起  西海岸
```

至此，基本的数据导入和再加工（data munging）就完成了。

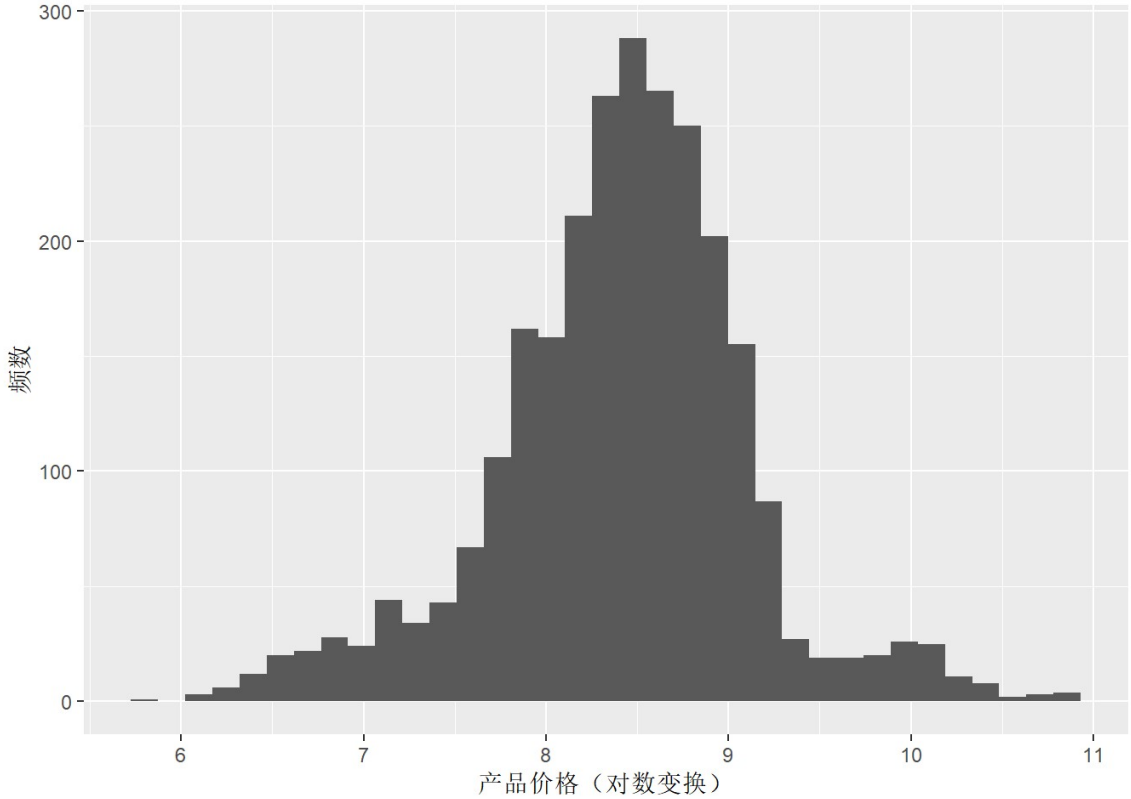
2 任务二

在任务二中，我们提取周一到周日报价中的数值部分，计算一周报价的均值（若一周7天均无报价则缺失），并以新变量“Price”存入数据集travel_dat中；然后，剔除平均价格缺失的样本。

```
l_Sun <- str_extract_all(travel_dat$SunPrice, '[0-9]+[0-9]')
l_Mon <- str_extract_all(travel_dat$MonPrice, '[0-9]+[0-9]')
l_Tue <- str_extract_all(travel_dat$Tuesprice, '[0-9]+[0-9]')
l_Wed <- str_extract_all(travel_dat$WedPrice, '[0-9]+[0-9]')
l_Thu <- str_extract_all(travel_dat$ThusPrice, '[0-9]+[0-9]')
l_Fri <- str_extract_all(travel_dat$Friprice, '[0-9]+[0-9]')
l_Sat <- str_extract_all(travel_dat$Satprice, '[0-9]+[0-9]')
l_Sun <- str_extract_all(travel_dat$SunPrice, '[0-9]+[0-9]')


l2c <- function(list){
  v <- vector('integer', length(list))
  for(i in 1:length(list)){
    if(length(list[[i]])==0){
      v[i] <- NA
    }else{
      v[i] <- list[[i]]
    }
  }
  return(as.integer(v))
}

v_Sun <- l2c(l_Sun)
v_Mon <- l2c(l_Mon)
v_Tue <- l2c(l_Tue)
v_Wed <- l2c(l_Wed)
v_Thu <- l2c(l_Thu)
v_Fri <- l2c(l_Fri)
v_Sat <- l2c(l_Sat)

dat1 <- data.frame(v_Sun, v_Mon, v_Tue, v_Wed, v_Thu, v_Fri, v_Sat)
logic <- vector('logical',nrow(dat1))
for(i in 1:nrow(dat1)){
  logic[i] <- sum(is.na(dat1[i,]))
}

dat1 <- dat1[logic != 7,]
index <- as.integer(rownames(dat1))

# 求平均值
price <- apply(dat1, 1, function(x) mean(x,na.rm = T))
travel_dat <- travel_dat[index,]
travel_dat <- cbind(travel_dat, price)

# 画出直方图
library(ggplot2)
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
ggplot(data = travel_dat, aes(log(1+price)) ) +
  geom_histogram(bins=35) +
  scale_fill_manual(values = c('yellow')) +
  xlab('产品价格（对数变换）') +
  ylab('频数')
```



观察该对数分布直方图，不难看出，价格的对数几乎呈现正态分布，即价格本身应该是对数正态分布。大部分旅游产品的价格位于5000到8000元之间，且更多的报价位于靠近5000元一侧。总体而言，北美旅游市场的产品报价较为合理，可以进行进一步探究。

3 任务三

在任务三中，从“Place”变量中提取出全部景点数和经典景点数，并分别以“AllPlace”和“ClassicPlace”变量存入数据集travel_dat，变量类型为数值型。

```
P <- str_extract_all(travel_dat$Place, '\\d+')

AllPlace <- vector('integer', length(P))
ClassicPlace <- vector('integer', length(P))

for(i in 1:length(P)){
  if(length(P[[i]])==0){
    AllPlace[i] <- NA
    ClassicPlace[i] <- NA
  }else{
    AllPlace[i] <- P[[i]][1]
    ClassicPlace[i] <- P[[i]][2]
  }
}
AllPlace <- as.numeric(AllPlace)
ClassicPlace <- as.numeric(ClassicPlace)
```

然后，将全部景点数按由少到多分成4组，分别为“9个及以下”，“10-16个”，“17-25个”，“25个以上”，以变量“AllPlacesGroup”变量保存在数据集travel_dat中，计算每一组内产品的平均价格并使用dplyr包的summarise()函数进行展示。

```
travel_dat <- cbind(travel_dat, AllPlace, ClassicPlace)

library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
AllPlacesGroup <- vector('character',length(P))
AllPlacesGroup[AllPlace <= 9] <- '(0,9]'
AllPlacesGroup[AllPlace > 9 & AllPlace <= 16] <- '(9,16]'
AllPlacesGroup[AllPlace > 16 & AllPlace <= 25] <- '(16,25]'
AllPlacesGroup[AllPlace > 25 & AllPlace <= 77] <- '(25,77]'
AllPlacesGroup[AllPlace > 77] <- NA
AllPlacesGroup <- factor(AllPlacesGroup, levels = c('(0,9]','(9,16]','(16,25]','(25,77]',NA),
                         ordered = T)

travel_dat <- cbind(travel_dat, AllPlacesGroup)
summarise(group_by(travel_dat,AllPlacesGroup),mean(price), na.rm = T)
```

```
## # A tibble: 5 x 3
##   AllPlacesGroup `mean(price)` na.rm
##   <ord>                  <dbl> <lgl>
## 1 (0,9]                  4203. TRUE
## 2 (9,16]                 5300. TRUE
## 3 (16,25]                6564. TRUE
## 4 (25,77]                6981. TRUE
## 5 <NA>                   4596. TRUE
```

根据输出的结果，可以观察一个基本的规律，即景点数量越多，产品本身价格越高。当然，由于边际收益递减的原理，随着景点数目增加，增幅在逐渐减小。

4 任务四

在任务四中，利用周一到周日是否有报价，计算每周出团日期，若当天有报价则视为出团，无报价则为未出团。提取出“仅工作日”，“仅周末”，“工作日和周末”三类出团情况并以“Date”变量存入数据集travel_dat（不属于上述任何一种情况的以缺失值存储），计算每一类出团日期的平均价格并展示。

```
datef <- function(x){
  a <- sum(is.na(x[1]) + is.na(x[7]))
  b <- sum(is.na(x[2]) + is.na(x[3]) + is.na(x[4]) +
           is.na(x[5]) + is.na(x[6]))
  return(list(a=a,b=b))
}

Date <- vector('character', nrow(dat1))
for(i in 1:nrow(dat1)){
  l = datef(dat1[i,])
  if(l[[1]]==2){
    Date[i] <- '仅工作日'
  }
  if(l[[2]]==5){
    Date[i] <- '仅周末'
  }
  if(l[[1]]!=2 & l[[2]]!=5){
    Date[i] <- '工作日和周末'
  }
}

travel_dat <- cbind(travel_dat, Date)
summarise(group_by(travel_dat,Date),mean(price))
```

```
## # A tibble: 3 x 2
##   Date          `mean(price)`
##   <fct>                 <dbl>
## 1 工作日和周末          5787.
## 2 仅工作日              5275.
## 3 仅周末                6929.
```

根据输出的结果，我们得到了一个3*2的矩阵，其中“仅周末出团”的旅游产品平均价格高于“工作日和周末出团”的旅游产品平均价格，而其价格又高于“仅工作日出团”的价格。事实上，这样的现象并不难理解，周末（假期）的工资往往要高于工作日， 而北美的工会力量异常强大，导致人工费用（尤其是假期人工费用）极高；因此，“周末”出团往往要在无形中提高出游的成本。

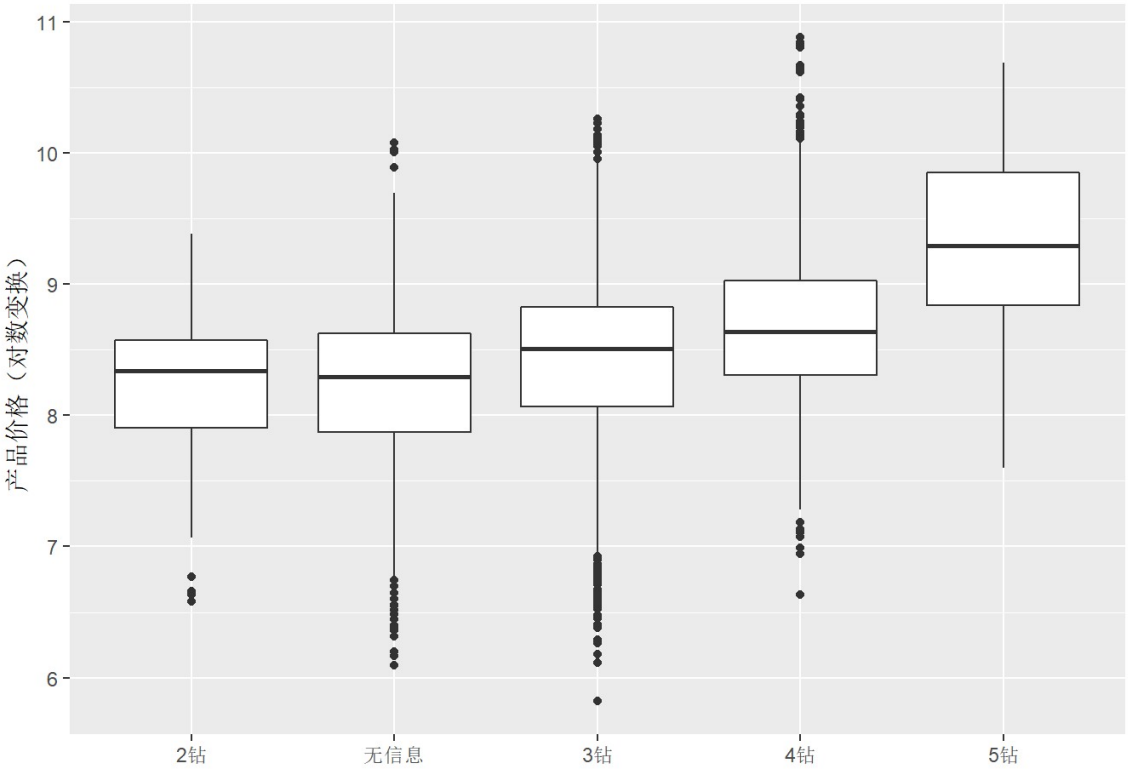
5 任务五

最后，在任务五中，我们提取“Star”变量中的“钻”字符来表示产品等级；当一个产品包含多个钻级时取最大钻级，并将产品钻级以新变量“Star2”存入数据集travel_dat中，变量类型为因子型。

```
loc <- str_locate_all(travel_dat$Star,'钻')

Star2 <- vector('character',nrow(travel_dat))
for(i in 1:length(loc)){
  if(length(loc[[i]][,1])==0){
    Star2[i] <- '无信息'
  }else{
    m <- max(as.integer(str_sub(travel_dat$Star[i],loc[[i]][,1]-1,loc[[i]][,2]-1)))
    Star2[i] <- paste(m, '钻', sep='')
  }
}

Star2 <- factor(Star2, levels = c('2钻','无信息','3钻','4钻','5钻'), ordered = T)

travel_dat <- cbind(travel_dat, Star2)
```

下面画出价格的对数对产品等级的分组箱线图，并按每一等级的平均价格由低到高进行排列。

```
ggplot(data = travel_dat, aes(x = Star2,y = log(price))) + geom_boxplot() +
  xlab('') + ylab('产品价格（对数变换）')
```



箱线图所揭示的信息非常直观：“钻”级越高，产品价格也相应越高。“钻”级往往是根据线路规划、服务态度、服务水平等一系列标准进行评定的；“钻”级越高，服务质量越好，旅行社的运营成本自然也越高。值得注意的是，“无钻级信息”的产品尽管平均价格高于“2钻”产品，但是其价格中位数却是最低的，因此不可以草率地认为“无钻级信息”的产品就好于“2钻”产品。此外，“5钻”产品的价格箱体几乎不与其他等级重叠，其价格远远高于其他等级产品；这一方面是因为“5钻”服务质量较高，另一方面或许也与旅行社的营销策略有关：相比于其他等级，“5钻”往往更易营造一种“高水准”的印象，而更高的价格则与“钻级”相互佐证，从而吸引支付能力更强的一批顾客。

6 结论

通过上述的分析，我们对数据集有了一个较为全面的了解。尽管没有使用任何无监督学习算法或监督学习算法，探索性数据分析本身也是非常关键的。通过对原始数据提取和清理的特征工程，我们可以获得更多可以使用的变量；而描述统计本身则提供了最为直观的解释。

通过上述分析，我们可以针对不同情况的驴友提供不同的建议。对于重视旅游体验，预算充足的旅游，可以直接选择钻级较高的旅游产品，该类产品价格高、方差小，有十分安全的保障。对于预算适中、力求物美价廉的旅游，不妨选择钻级在“3钻”或者“4钻”的产品，其服务质量有保障，价格相差也不大。在景点数目的选择上，如果想小而精，9到16个为宜，因为其价格比16到25个低了不少，景点数量的降幅却不大；而对于想要求大求全的旅游，完全可以选择25到77个景点的区间，毕竟价格只比上一档贵了400元左右，景点数目的增幅却很大。

最后，需要提醒各位驴友，如非必要，完全可以避开周末开团的旅行产品，毕竟对于相差不多的项目和服务，仅仅因为时间因素，就要多花费几百元甚至上千元，并不是非常值得。