口红市场数据挖掘及营销建议

《计算新闻传播学》课程期末作业

一、 背景调研

1. 口红整体市场份额

（1） 全球市场：在全球市场，唇妆是全球彩妆市场增长最快的品类，2012 到 2014 年间增占率达到 9%。2017 年全球唇部护理产业总价值已达 19 亿美元。（约合人民币 115.8 亿元人民币）

（2） 中国市场：截止 2016 年中国美妆市场交易额已达 1618.3 亿元，预计 2018 年将达到 2600 亿元；而彩妆的交易额达到了 44%，并且口红最多。

（3） 整体增速：口红在 2016 年销量增速已达到美妆整体增速的 5 倍。

2. 口红目标消费市场

（1） 整体概况：

目前中国女性经济市场规模近 2.5 万亿元，到 2019 年市场规模有望达到 4.5 万亿元。

（2） 目标对象特点：

- 愿意接受新产品，拥有新体验；消费集中度低，均衡多样。
- 收入多少和所消费品牌的关系在美妆品类上不再阶层分明，中等收入消费者对于高端品牌的需求已经日渐凸显。
- 年轻潮流女性扛起了高端唇膏消费的大旗。
- 爆款产品的消费者跟风现象明显。

（3） 地区喜好差异：

根据艾瑞咨询于 2016 年发布的年度口红消费报告数据，不同地区的消费者对口红喜好也有所不同。其中，东三省最爱姨妈红，河南、海南、江西最爱豆沙色，内蒙、宁夏、青海最爱正红，安徽、河南、江西最爱南瓜红，甘肃、西藏、宁夏最爱枚红和裸色。

（4） 当前口红十大品牌：其他热销的还有安娜苏，阿玛尼，资生堂，倩碧，植村秀，NARS，兰芝等。

3. 品牌活动现状：

（1） 研发新产品：

除了 MAC 广为人知的 128 支“子弹头”外，阿玛尼红管哑光唇釉系列和黑管唇釉分别有 24 和 18 个色号。

同样热门的还有 Tom Ford 的 50 支 Lips & Boys 系列、YSL 圆管和方管分别有 24 和 50 多个色号、Christian Louboutin 的“萝卜丁”口红则有 38 个色号、Chanel 的 Rouge Coco Shine 系列多达 24 个色号、资生堂新品“臻红”系列唇膏根据颜色深浅饱和度的不同开发了 16 个色号。

（2） 拓展营销手段：

- 男星代言美妆品牌，如杨洋代言娇兰，陈伟霆代言美宝莲；
- 植入电视节目，如三生三世十里桃花中的“杨幂色”；
- 利用社交网络内容营销，如 YSL 星辰的火热；
- 红人明星直播，如网红口红试色

二、 数据分析

1. 数据总体描述

笔者爬取了淘宝网站上 1696 条店铺口红销售数据，每条代表一个口红商品的信息，共包含变量 17 个。（数据表格见文件夹内“口红-数据”）

（1） 评分最高的店铺

描述分		价格分		质量分		服务分	
亚玛迪美妆店	4.7	小粥粥美妆国	4.7	敏恩玩美妆	4.7	MOUENE慕漾官方店	4.85
美谁妹妹的店	4.7	MOUENE慕漾官方店	4.7	仟佰媚美妆	4.7	小粥粥美妆国	4.8
小粥粥美妆国	4.7	大凡美颜公社	4.68	小粥粥美妆国	4.7	猫猫家美妆	4.76
敏恩玩美妆	4.7	猫猫家美妆	4.68	MOUENE慕漾官方店	4.7	大凡美颜公社	4.74
MOUENE慕漾官方店	4.7			大凡美颜公社	4.7		
大凡美颜公社	4.7			猫猫家美妆	4.7		
猫猫家美妆	4.7						
仟佰媚美妆	4.7						
露娜美妆	4.7						

（2） 评价数最多的店铺：时尚韩国美妆

- 评价关键词：口红唇膏防脱色保湿持久不掉色防水滋润口红液韩国咬唇妆不脱色
- 评价数量：45852

（3） 销量最高的店铺：韩熙贞官方旗舰店

- 销量冠军单品:【第二支 1 元】韩熙贞持久保湿不掉色口红 咬唇妆滋润唇膏
- 销售量：32182

2. 口红销量影响因素

笔者通过 R 软件对于原始数据进行分析，总结口红销量影响因素。具体过程见下：

（1） 数据说明：

变量类型	变量名	详细说明	取值范围
因变量	总销量	单位：支	0~32192
自变量	颜色丰富度	单位：种	1~82
	价格	单位：元	9~999
	总评价数	单位：个	0~45852
	综合评分	对描述、价格、服务、质量分的平均	4.16~4.74
	是否防晒	定类变量，共2个水平	是(1)/否(0)
	国家	定类变量，共15个水平	爱尔兰、比利时、德国、法国、韩国、加拿大、美国、日本、意大利、英国、泰国、中国台湾、中国内地、其他

（2） 分词处理：

对口红颜色进行分词处理，形成“颜色丰富度”的可操作指标。

- 注：#strsplit()函数对 excel 中挤在一个单元格内的各种口红颜色进行基于“换行”（\\s）为分隔符的字符串分割；#lapply()与 length()函数对字符串得到具体店铺与具体标题下的口红共有几种颜色，形成“颜色丰富度”的可操作指标

```
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
setwd("/Users/cecilia/Desktop")
getwd()

color <- read.table(pipe("pbpaste"),header = T,fill=TRUE)
View(color)
color1 <- as.character(color$颜色)
sen1 <- strsplit(color1,"\\s")
Lsen <- lapply(sen1,length)
Lsen <- unlist(Lsen)
```
```

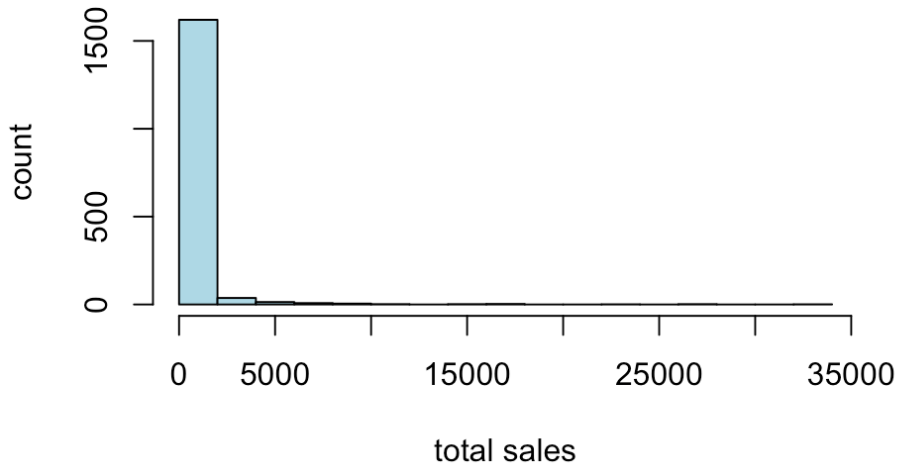
（3） 对因变量的详细说明：

- 淘宝店铺口红总销量（N=1695）

```
sales <- read.table(pipe("pbpaste"),header = TRUE,fill = TRUE)
summary(sales$总销量)
hist(sales$总销量,xlab="total
sales",ylab="count",main="",col="lightblue")
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0     5.0    26.0   422.2   140.5 32192.0
```

- 各购买链接销量分布的均值、极值、销量分布及其不均



（4） 多元回归分析：

- 采用 pipe（“pbpaste”）的剪切板函数将各类相关数据读取进 R 中
- 为了方便回归分析，将字符串性质转为 integer 并将区间值取中位值
- 与之前构建的概念“颜色丰富度”一起，将总评价数、综合评分、价格、是否防晒、国家等作为自变量，进行多元回归分析，发现除“国家”以外均显著相关

```
price <- read.table(pipe("pbpaste"),header = TRUE,fill = TRUE)
comment <- read.table(pipe("pbpaste"),header = TRUE,fill = TRUE)
suntan <- read.table(pipe("pbpaste"),header = TRUE,fill = TRUE)
country <- read.table(pipe("pbpaste"),header = TRUE,fill = TRUE)
grade <- read.table(pipe("pbpaste"),header = TRUE,fill = TRUE)
class(price$价格)
price$价格 <- as.integer(price$价格)
lm1=lm(sales$总销量~color$Freq + comment$总评价数 + suntan$防晒 +
price$价格+ grade$综合评分 +country$国家)
summary(lm1)
par(mfrow=c(2,2))
plot(lm1,which=c(1:4))

lm2 <- lm(log(sales$总销量 + 1)~color$Freq + comment$总评价数 +
suntan$防晒 + price$价格 +grade$综合评分 )
summary(lm2)
par(mfrow=c(2,2))
plot(lm2,which=c(1:4))
```

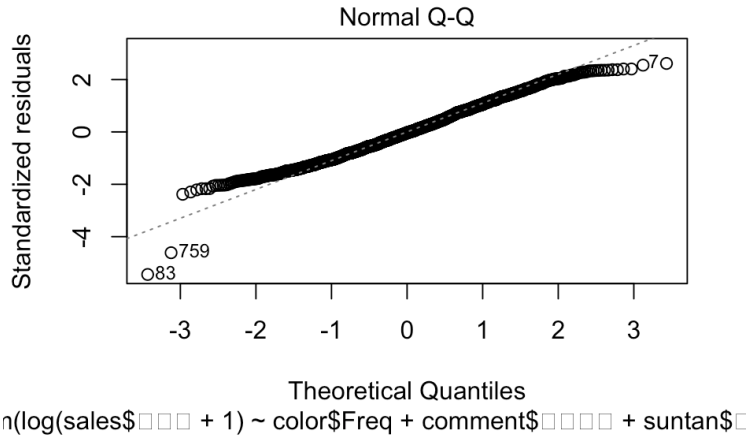
- 在回归诊断中出现异方差情况，故对因变量进行了 log（y+1）的变换，（考虑到 y 可能为 0 故+1）
- 变换后的回归模型符合正态性诊断，诊断图见下

```
Call:
lm(formula = log(sales$总销量 + 1) ~ color$Freq + comment$总评价数 +
    suntan$防晒 + price$价格 + grade$综合评分)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4871 -1.4783 -0.0669  1.4834  5.2043

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.347e+01  2.304e+00   5.849 5.95e-09 ***
color$Freq        2.218e-02  8.011e-03   2.768   0.0057 **
comment$总评价数  3.101e-04  2.182e-05  14.213  < 2e-16 ***
suntan$防晒      -4.141e-01  1.899e-01  -2.181   0.0293 *
price$价格       -2.533e-03  3.487e-04  -7.265 5.70e-13 ***
grade$综合评分   -2.082e+00  5.097e-01  -4.085 4.61e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.001 on 1665 degrees of freedom
  (24 observations deleted due to missingness)
Multiple R-squared:  0.1657,    Adjusted R-squared:  0.1631
F-statistic: 66.12 on 5 and 1665 DF,  p-value: < 2.2e-16
```



n(log(sales$□□□ + 1) ~ color$Freq + comment$□□□□ + suntan$□

（5） 对数线性模型结果解读

变量	回归系数（$\times 10^{-1}$）	P值
截距项	1.347	＜0.01
颜色丰富度	4.9195	＜0.01
总评价数	92.4713	＜0.01
是否防晒	-4.141	＜0.01
价格	-16.252	＜0.01
综合评分	-0.48	＜0.01
F检验	P值＜2.2^{-16}	调整后R方0.1631

- 整个模型解释程度不是很高，但在各类变换中效果最佳
- 在控制其他因素不变时，
  颜色每多一种，销量可增加 49%；
  总评价数每多一条，销量可增加 924%；
  防晒的口红比不防晒的口红销量少 41%；
  价格每贵一元，销量下降 162%；
  综合评分每多一分，销量下降 4.8% 。

3. 评分对口红销量的影响

- 经过多次变换后得到对数线性模型的拟合程度最好，除“质量分”一项不显著外均显著
- “描述分”呈负相关，推测存在“销量较差的淘宝店铺为提高销量而自行刷分行为”
- “价格分”与“服务分”分别与销量呈正相关，“服务分”系数更大，证明良好的服务反馈比合理的价格评论对销量促进更多

```
gradems <- read.table(pipe("pbpaste"),header = TRUE,fill = TRUE)
gradejg <- read.table(pipe("pbpaste"),header = TRUE,fill = TRUE)
gradezl <- read.table(pipe("pbpaste"),header = TRUE,fill = TRUE)
gradefw <- read.table(pipe("pbpaste"),header = TRUE,fill = TRUE)
lm3 <- lm(sales$总销量 ~ gradems$描述分 + gradejg$价格分 +
gradezl$质量分 + gradefw$服务分 )
lm4 <- lm(sqrt(sales$总销量) ~ gradems$描述分 + gradejg$价格分 +
gradezl$质量分 + gradefw$服务分 )
lm5 <- lm(log(sales$总销量 + 1) ~ gradems$描述分 + gradejg$价格分 +
gradezl$质量分 + gradefw$服务分 )
summary(lm5)
```

```
Call:
lm(formula = log(sales$总销量 + 1) ~ gradems$描述分 + gradejg$价格分
+
    gradezl$质量分 + gradefw$服务分)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6276 -1.5452 -0.1378  1.3396  6.9731

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     15.729      3.562   4.415 1.07e-05 ***
gradems$描述分  -24.981      3.900  -6.405 1.94e-10 ***
gradejg$价格分    9.075      2.552   3.556 0.000387 ***
gradezl$质量分   -3.526      3.327  -1.060 0.289314
gradefw$服务分   16.798      2.838   5.919 3.92e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.107 on 1690 degrees of freedom
Multiple R-squared:  0.07487,   Adjusted R-squared:  0.07268
F-statistic: 34.19 on 4 and 1690 DF,  p-value: < 2.2e-16
```

4. 产品名称与功效词云：

（1） 商家最青睐的标题词

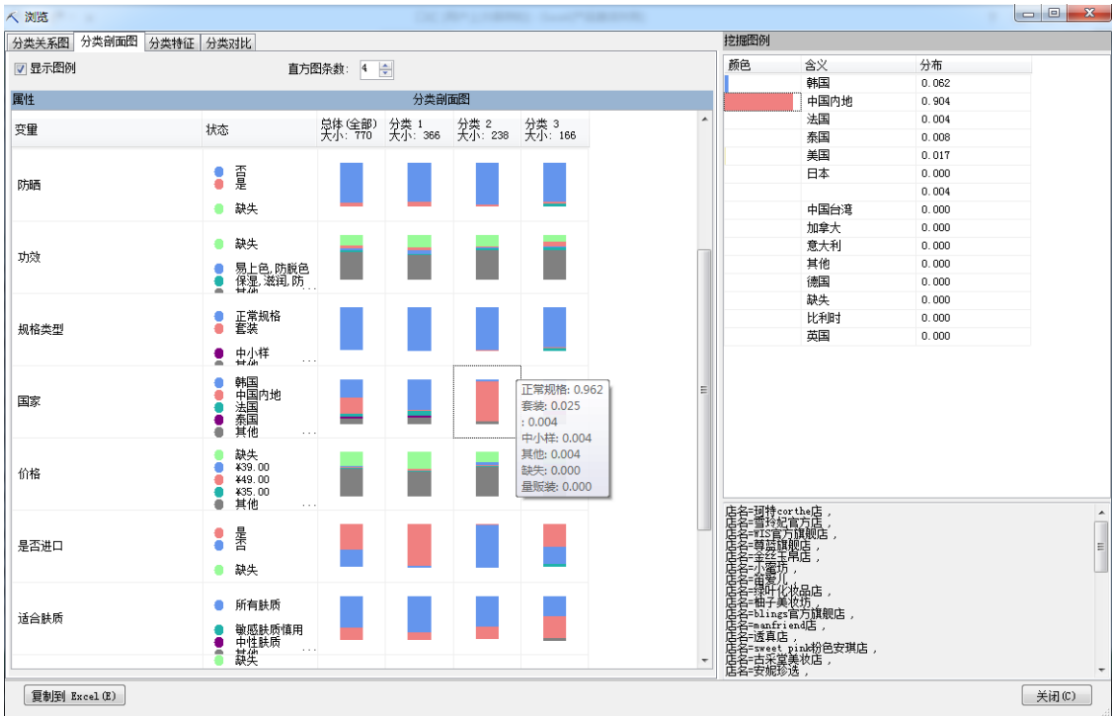
（2） 何种功效的口红最多

（3） 何种颜色的口红最多

（4） 口红产地

三、 市场预测及营销建议

1. 现有口红市场分类特征：



分类一	分类二	分类三
韩国	内地	
49	39	
易上色	保湿、滋润	易上色、防脱、保湿、滋润、持久
防晒	不防晒	不防晒
正常规格	套装	
3年	3年	3年
蜜糖、敏恩、碧黛	丽子美妆 茉莉美妆	丽子美妆 美谁妹妹美妆 果粒美妆
所有肤质	敏感肤质慎用	

2. 营销建议

（1） 提高淘宝店铺内服务质量，增加良好的服务反馈；

（2） 可以将统一分类的产品打包销售，如第一类：韩国产、易上色、防晒适合所有肤质的口红，定价为 49 左右；第二类：内地产、保湿滋润口红，定价为 39 左右；

（3） 根据数据分析结果，并非价格越便宜销量越高，因此店铺不应一味打价格战；

（4） 增加店内商品可选择性、丰富色号，根据地域喜好进行一定的消费推荐或智能 push 推荐。