**商业香水排行前200名的研究**

1710751104 韦利仁 2018/11/22

**目录**

1. 抓取数据
2. 数据整理
3. 数据分析

（一）、分析前200名商业香水的品牌分布情况。

（二）、分析前200名商业香水的属性分布情况。

（三）、香水诞生年份分析

（四）、香水口碑分析

1. 参考

**前言**

对于大多数人来说，想要让自己在人群中脱颖而出，除了靠衣着之外，另一个吸睛的方法便是依靠香水。香水是极其私人的东西，甲之少女，乙之妇女，更或者，不男不女。私以为，看再多的香评，都不如你的鼻子诚实，你喜欢的，便是最好的。在香水榜单方面，国内由于文化等诸多方面的影响，没有十分突出的榜单，选择了知名度比较高的香水时代网站的排行榜，一定程度上反映了国内用香人的品味。本次分析实验所用的数据排行是基于大多数的人评价，是相对的，并不绝对，你喜欢的就是第一。因此本次分析实验则是选择了评价人数较多的商业香水，而非沙龙香水。分析实验用R语言数据可视化分析了商业香水前200名中的品牌、属性及香调等的特性，得出大多数人对于香水的感觉，但也有一些误差，比如是否存在女性使用香水的人数大于男性人数的问题，这可能会导致香水排行的不准确等等。本次实验数据来源香水时代<https://www.nosetime.com>。

1. **抓取数据**

# 抓取香水名称

# 抓取香水评分

# 抓取香水详情网址

# 抓取香水具体细节

get\_perfum <- function(url){

page <- read\_html(url)

`商业香水top200` <- tibble(

row1 = page %>%

html\_nodes("div.trade-article div.trade-content-article h2 a") %>%

html\_text(),

row2 = page%>%

html\_nodes("div.score span.tiaoz") %>%

html\_text(),

row3 = page %>%

html\_nodes("div.trade-article div.trade-content-article h2 a") %>%

html\_attr("href") %>%,

paste0("https://www.nosetime.com", .),

row4 = page %>%

html\_nodes("div.intro") %>%

html\_text() %>%

str\_replace\_all("\n\t\t\t\t\t\t\t", "") %>%

str\_replace\_all("\n\t\t\t\t\t\t", "") %>%

sub("\\s+", "", .)

)

}

urls1 <- str\_c("https://www.nosetime.com/top200.php?type=trade&page=",1:10,"#list")

perfums <- urls1 %>%

map\_df(get\_perfum)

perfums

View(perfums)

1. **整理数据**

首先分析名称的特点，例如：爱马仕 大地Hermes Terre d'Hermes, 2006，可以得出名称的第一个空格前就是品牌名以及逗号之后是香水诞生的年份的特点，利用此特点可以整理出香水所属品牌以及她们的诞生年份。此外有一些特殊的例如：”爱马仕之旅Hermes Voyage d'Hermes, 2010”和”香奈儿 19号Chanel No 19 Eau de Parfum”这两个名称没有之前我们总结的特点，就需要我们先对其进行处理。

perfums$row1[11] = paste0(perfums$row1[11], ", 1970")

perfums$row1[19] = paste0(perfums$row1[19], ", 2007")

perfums$row1[171] = paste0(perfums$row1[171], ", 2006")

perfums$row1[195] = paste0(perfums$row1[195], ", 1960")

perfums$row1[46] = paste0("爱马仕 ", perfums$row1[46])

perfums$row1[53] = paste0("爱马仕 ", perfums$row1[53])

`品牌` = perfums$row1 %>%

regexpr(" ", .) %>%

map(1) %>%

substr(map(perfums$row1, 1), 1, .)

`年份` = perfums$row1 %>%

regexpr(",", .) %>%

map(1) %>%

substring(map(perfums$row1, 1), .) %>%

str\_replace\_all(", ", "") %>%

map(1) %>%

as.numeric()

其次，我们要从香水的具体内容里将属性和香调整理出来。其中有一些香水的香调不能确定，将对他们不做要求。

`属性` = perfums$row4 %>%

regexpr("属性：", .) %>%

map(1) %>%

map(~. + 3) %>%

substr(map(perfums$row4, 1), ., map(., ~. + 2))

`香调` = perfums$row4 %>%

regexpr(" 香调：", .) %>%

map(1) %>%

map(~. + 4)%>%

substring(map(perfums$row4, 1), ., last = 100000L)

之后将其形成一个整理好的表格（见附1），有利于我们后面进行分析。

**三、数据分析**

（一）、分析前200名商业香水的品牌分布情况。

香水的品牌有时候决定了香水的品质，香水的价格，这使得我们在购买香水时，多了几分思考。

先将品牌这一列提取出来，并用table()函数进行数量统计，将统计好的数据形成一个数据框并进行排名，升序。之后并利用ggplot2画出品牌数量的条形图（见附2）。

mytable <- table(aa$`品牌`)

df <- as.data.frame(mytable) %>%

arrange(Freq)

df %>%

ggplot(aes(x = Var1, y = Freq, fill = Var1)) +

geom\_bar(stat = "identity", width = 2) +

geom\_text(aes(label = Freq, hjust = -0.5, color = Var1), show.legend = TRUE)+

coord\_flip()

其中品牌数量的前五排名如下：

1 爱马仕 19

2 娇兰 19

3 迪奥 15

4 香奈儿 12

5 宝格丽 11

不难看出，前200 名一共有62个品牌入围，前四名被法国的品牌囊括，其中爱马仕和娇兰以19个香水入围数量并排第一， 迪奥有15款香水入围，香奈儿则有12款香水入围，意大利的宝格丽以11款紧随其后。

（二）、分析前200名商业香水的属性分布情况。

香水的属性分别是，男香，女香，中性香。不同属性代表了不同的格调，这也成为我们分析的原因。

先对属性进行筛选，并统计他们的数量，直接将统计好的数量化成饼状图（见附3）。

y = c(

nrow(filter(aa, `属性` == "男香 ")),

nrow(filter(aa, `属性` == "女香 ")),

nrow(filter(aa, `属性` == "中性香"))

)

labels = c("男香", "女香", "中性香")

pie(y, labels, col=rainbow(3), radius = 0.9)

从附3可以看出，女香在前两百中占有一半的数量，初步分析，这与女性使用香水的数量较多，而使得调香师们调出的女香较多会有一定的关系。但更多原因需要进一步的分析。而男香与中性香所占的比例依次排下。

（三）、香水诞生年份分析

商业香水top200中，最早的一款香水是1792年来自摩勒沃兹的4711原始古龙水，最近的一款是2018年来自迪奥的典藏系列-琥珀幽夜。榜单中各个年份的香水数量各有不同，按该年份入围香水数量多少降序排列。

mytable1 <- table(aa$年份) %>%

as.data.frame() %>%

arrange(-Freq)

前五名分别是：

1 2006 18

2 2007 18

3 2008 14

4 2010 11

5 2003 10

可以看出，入围榜单的大部分香水是2000年以后的香水，这可能是与人们的接受能力有关，较以前的香水，最近的香水可能接受度较高。

（四）、香水口碑分析

评分机制只能一定程度上体现香水的质量，但是很多时候不能体现香水的全面信息。如同电影一般，一部电影的口碑一致性越高，说明这是一部好电影的可能性越高，香水为私人物品，与电影一样，都是以人的感受为主，但又不同的是，电影会传达某种价值观，但香水没有对不对，只有喜不喜欢。所以这里的分析只是抱着尝试的态度进行分析。

一般来说，如果是口碑一致的好评香水，其评分的形状大致是 P 形，即从五星至一星逐渐递减。而对于口碑分化严重的评分形状大多是C形，即五星与一星都较少，中间星级所占的比例则较多。

为了量化我们上面的观察结果，我们需要计算两个指标：

1. 平均星级：用各个星级所占的比例对星级进行加权平均，求出加权平均星级，作为衡量电影质量的指标。
2. 星级标准差：在平均星级的基础上，以各个星级所占比例为权重，我们可以计算出星级评价的标准差，这个指标可以量化口碑的不一致程度，标准差越大，则口碑越不一致。

首先需要从香水的详情页上抓取星级评分的比例，之后再进行求加权平均值和标准差的计算。

getdetail <- function(url){

page <- read\_html(url)

tibble(

`五星` = page %>%

html\_nodes("ul.item\_score div.nows:nth-child(1)") %>%

html\_text()

)

}

getdetail\_safely <- safely(getdetail)

jobdetail <- perfums$row3 %>%

map(getdetail\_safely)

b = jobdetail %>%

map\_df("result")

e = b$五星 %>%

regexpr("%", .) %>%

map(., ~. - 1) %>%

substr(b$五星, 1, .) %>%

as.numeric() %>%

map(., ~. \* 0.01)

mys <- as.numeric()

mystar\_std <- as.numeric()

for(i in 1:200){

s = e[(5 \* i - 4):(5 \* i)] %>%

as.numeric()

mys[[i]] <- s[1]\*10 + s[2]\*8 + s[3]\*6 + s[4]\*4 + s[5]\*2

mystar\_std[[i]] <- sum(c(

(10 - mys[[i]])^2 \* s[1]),

((8 - mys[[i]])^2 \* s[2]),

((6 - mys[[i]])^2 \* s[3]),

((4 - mys[[i]])^2 \* s[4]),

((2 - mys[[i]])^2 \* s[5])) %>%

sqrt() %>%

print()

}

arrange(mystar\_std)

将计算好的标准差进行排序之后，我们可以看看榜单中，口碑比较一致的前五名是：

名称 评分 mys mystar\_std

<chr> <dbl> <dbl> <dbl>

1 汤姆·福特 白色广藿香Tom Ford White Patchouli, 2008 8.7 7.7 1.51

2 娇兰 蓝调时光Guerlain L'Heure Bleue, 1912 9.3 8.22 1.60

3 娇兰 香根草（伟之华）Guerlain Vetiver, 2000 9.1 8.06 1.64

4 D二次方 落基山木缘DSQUARED² He Wood Rocky Mountain Wood, 2009 8.9 7.84 1.64

5 迪奥 沙丘男士Dior Dune, 1997 8.6 7.82 1.64

口碑最不一致的五款香水是：

名称 评分 mys mystar\_std

<chr> <dbl> <dbl> <dbl>

1 蒂埃里·穆勒 天使Mugler Angel, 1992 8.4 7.26 2.46

2 汤姆·福特 同名男士Tom Ford for Men, 2007 8.2 7.12 2.39

3 密使 同名香水Agent Provocateur, 2000 8.3 7.1 2.30

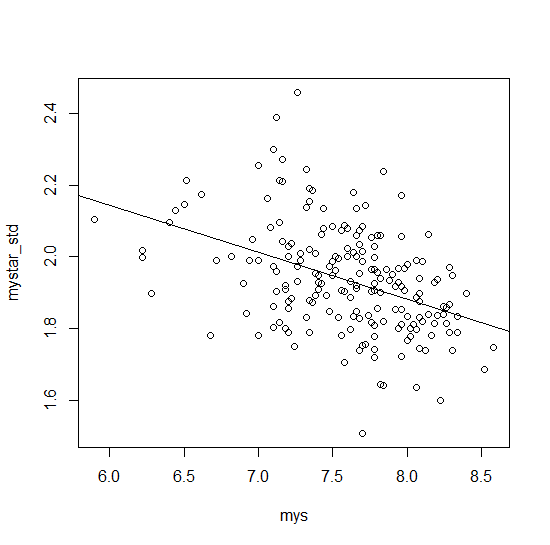
4 水莲娜 粉红棉花糖Aquolina Pink Sugar, 2004 8.1 7.16 2.27

5 娇兰 理想男士（偷心）Guerlain L'Homme Ideal, 2014 8 7 2.26

这五款香水的评分还是很好的，但是由于他们独特的香味，会让一部分人为之沉迷，一部分人逃离，使得口碑差异化较为严重。

(一) 电影评分与其口碑一致性的关系

为了更直观的理解电影的评分和他的口碑一致度的关系，我们可以以香水的平均星级为横轴，香水的星级标准差为纵轴绘制散点图如下：



我们可以看到，平均值与标准差之间关系不大明显，为了我们更好判断，我们需要做一个回归的模型。

fit <- lm(ss$mystar\_std ~ ss$mys, data = ss)

summary(fit)

Call会的值如下：

Call:

lm(formula = ss$mystar\_std ~ ss$mys, data = ss)

Residuals:

Min 1Q Median 3Q Max

-0.41516 -0.09140 -0.00715 0.07825 0.48053

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.93377 0.15249 19.238 < 2e-16 \*\*\*

ss$mys -0.13142 0.02005 -6.555 4.73e-10 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1339 on 198 degrees of freedom

Multiple R-squared: 0.1783, Adjusted R-squared: 0.1741

F-statistic: 42.96 on 1 and 198 DF, p-value: 4.728e-10

可以看到回归系数（-0.1314）显著不为0（P值小于0.001），但是由于R方很小，我们有必要做个相关性分析。

> cor(sss, method = "pearson")

mys mystar\_std

mys 1.0000000 -0.4222541

mystar\_std -0.4222541 1.0000000

可以看到相关系数的绝对值是0.42，说明他们两者中等相关。根据这个方法我们发现，在口碑与评分之间并没有存在很明显的相关性，这也很好的诠释了我们一直在说的香水是十分私密的物品，每个人对香水的喜爱都非常不一样的。

四、参考

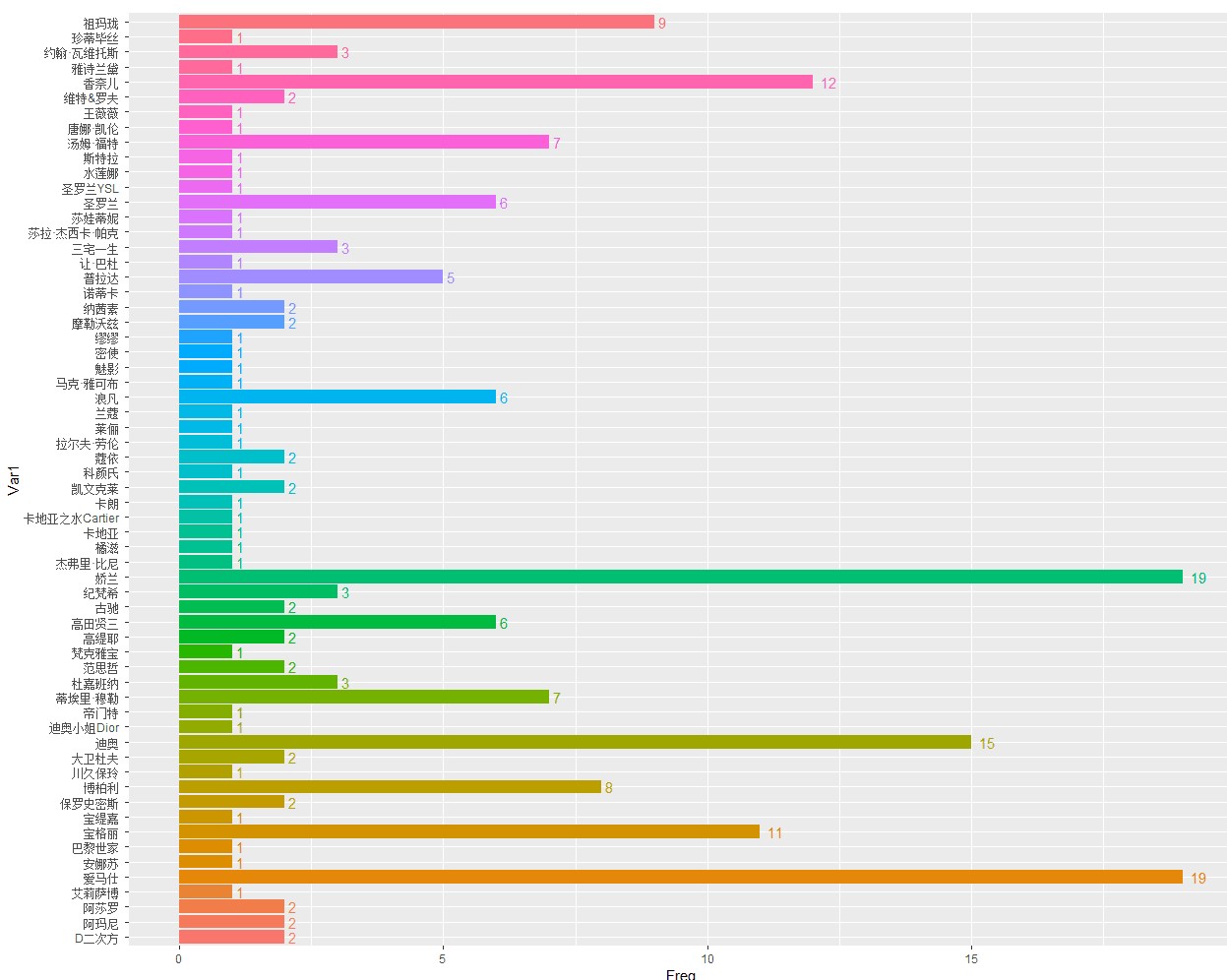
1.数据源地址<https://www.nosetime.com/top200.php?type=trade>

2.项目代码地址https://github.com/MeetLiren/R

附1（由于纸张大小原因只列出前25名）：



附2：



附3：

