有个直播间

——基于 RFM 聚类

姓名：于沁涵 学号：1910227

一、背景介绍

“你看过直播吗？”相信很多人的答案都是肯定的。随着互联网的飞速发展，现如今直播行业的发展如火如荼，直播间在各大软件里几乎无处不在。直播，是一种实时性、互动性显著的互联网传播内容的形式。不同于传统的文字、图片、视频等传播形式，直播紧密的将用户与直播内容交互在一起，用户本身也是内容生产的一份子。

在之前国内互联网刚刚起步的时候就已经出现了第一批网络主播，在网上喊麦活跃气氛直播，之后随着智能手机的出现，直播行业就貌似被打了一针肾上腺素一样，出现了飞跃性的发展，如今很多的直播平台已经相对成熟，比较出名的有熊猫 TV，触手 TV，龙珠，网易 CC，战旗，YY，虎牙，斗鱼，企鹅电竞等等。在 2020 年，方正证券预测直播市场将达到 600 亿元，华创证券更是给出了 1060 亿的乐观预测。这也更加反映出了人们对直播平台用户数快速增长的良好预期。

然而，在以流量为支撑的直播间中许多问题也应运而生。网络直播通过通过粉丝打赏，融资和广告收入，为直播平台实现盈利，但这一切都需要丰富的流量。例如抖音平台，我们时常可以看到有些拥有几千万粉丝的主播拥有可观的流量，但也不乏只有几个人的直播间，如果没有流量，那么一切都是徒劳，就像赌博一样，赌赢了就是暴富，赌输了只能一败涂地。因此，很多人即使在面对如此诱人的“赚大钱”方式面前，仍然选择了犹豫或退缩。

那么流量从何而来呢？粉丝是其中最直接的一条途径，直播的对象是看直播的人，那么拥有的粉丝数越多也说明看直播的人就会越多，流量自然也就更好，因此，掌握好粉丝的喜好和心理是每一个主播都需要做的一门功课。

为了帮助主播了解粉丝的喜好和行为，我们考虑利用已有的相关数据对直播间粉丝留下的“行为”进行分析。利用直方图、折线图、RFM 聚类对数据进行分析，并得到结论。

二、数据来源和说明

本文数据爬取了某直播平台两个 TOP 房间的粉丝操作记录，分别记为王者荣耀和颜值，数据分别有 210977 和 235159 个。时间跨度从 2017 年 4 月 27 日至 2017 年 6 月 27 日，其中每个数据包含 13 个变量，分别为房间 ID、房间类型、时间、操作类型、用户名、操作内容、星期、年、月、日、时、分、秒，具体变量说明如表 1:

表 1 数据变量说明

变量名称	详细说明	取值范围	备注
房间 ID	固定值	863、1038864	两个直播房间的 ID 号
房间类型	分类变量	王者荣耀、颜值	直播房间的名称
时间戳	连续变量	1493469201～1498491814	POSIXct 类型，距指定时间的秒数
操作类型	分类变量	登录、弹幕、送礼物	弹幕占比 47.7%
用户名	字符型变量	49～39002	网络中用户给自己起的昵称
操作内容	连续变量	21556～189660	弹幕发送的内容以及刷礼物的内容
星期	分类变量	0、1、2、3、4、5、6	0 代表周日、其余分别代表相应数字的星期
年	固定值	2017	2017 年
月	分类变量	4、5、6	4～6 月
日	连续变量	1～30	一个月中的每一天
时	连续变量	0～23	一天中的每小时
分	连续变量	0～59	每小时中的每分钟
秒	连续变量	0～59	每分钟的每一秒

三、描述性分析

（一）两个直播间的弹幕词云图

“词云”是文本数据的一种可视化展现方式，它一般是由文本数据中提取的词汇组成某些彩色图形。词云图的核心价值在于以高频关键词的可视化表达来传达大量文本数据背后的有价值的信息。因此，我们可以利用弹幕中的词云图来直观的看一看直播间最常出现的词有哪些，不同的直播间常出现的词有什么不同，利用 R 语言和万页软件，得到了结果图图 1 和图 2。

图 1 王者荣耀直播间弹幕词云图

图 2 颜值直播间弹幕词云图

由图可以看出，王者荣耀直播间的最常见关键词为“主播”、“洪斌”、“加油”、“666”等，这说明在游戏类直播间，主播和粉丝之间的互动是占主导地位的。在游戏直播间看直播的人倾向于选择自己喜欢的主播或者游戏类型，因此在发送弹幕的时候更喜欢发送一些与主播相关的弹幕，或者评价主播对游戏操作的言论，例如“加油”、“666”等。此外，与游戏内容相关的弹幕也脱颖而出，游戏英雄的名字、游戏使用的技能、段位.....这也说明观看直播的粉丝也喜欢讨论游戏内容的话题，因此，从王者荣耀直播间中我们可以看出粉丝发送弹幕的喜好主要来自于两方面：与主播的互动、游内容的讨论。

而颜值直播间中的最常见关键词包括“emot”、“仙宝”、“主播”、“dy115”等，这说明在颜值直播间，粉丝的聚焦点大多数在和主播的互动和倾诉上。其中在词云图中出现更多的还有“漂亮”、“好看”、“蓝瘦”等形容词，这说明粉丝留在直播

间也与主播的颜值有关，而直播间讨论的内容大多数在情感、生活、颜值方面，这与游戏直播间中讨论游戏内容形成比较鲜明的对比。

总体来看，两个直播间共有的特点是直播间中的粉丝与主播有很大的关系，而王者荣耀直播间的气氛比较诙谐和热闹，颜值直播间的气氛比较温馨和贴近生活。

（二）两个直播间的周内日均操作量分布图

为了探究直播间的火热程度，首先我们分析粉丝周内日均操作量分布，如图 3:

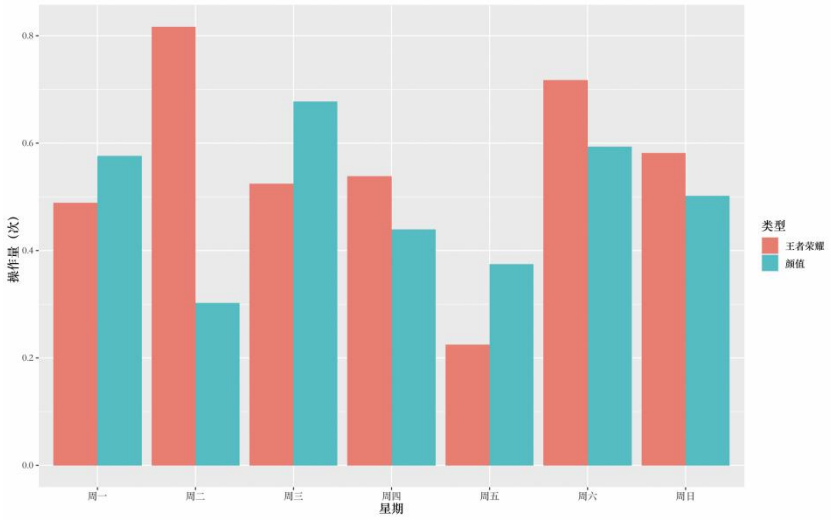

图 3 粉丝周内日均操作量分布

由图中看出王者荣耀的粉丝操作在周二、周四、周六和周日均高于颜值直播间，其中周二的操作量最高，而颜值直播间的粉丝操作在周三最多。这从侧面反映了直播娱乐呈现轻量化，即不再局限于周末或特定时间，追求随时随地的娱乐观赏。

（三）两个直播间的粉丝日操作量分布图

随后，我们观察了在 2017 年 4 月 27 日至 2017 年 6 月 27 日中，两个直播间粉丝的日操作量折线图，如图 4。

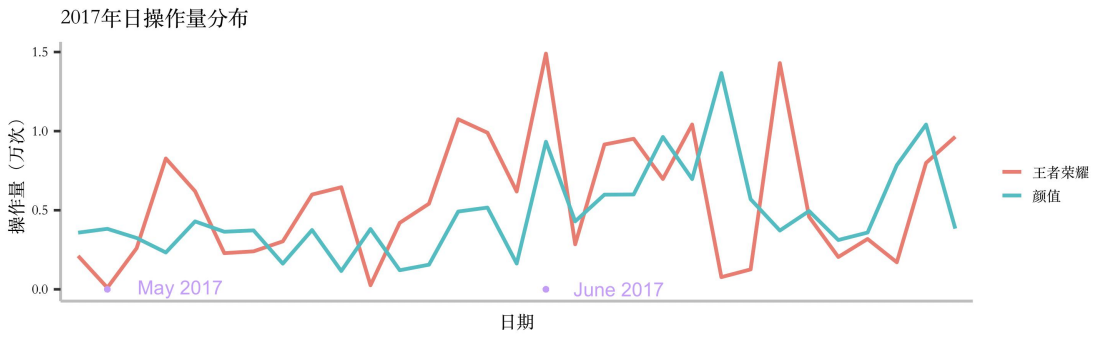

图 4 数据时间内粉丝日均操作分布

在图 2 中，操作量为 0 的日子，是主播当天没有直播。可以看到，王者荣

耀房间，两天左右直播一次，颜值房间，一天左右直播一次，这说明颜值主播直播的频率更多一些，这是因为颜值主播需要一种更高频率的持久性来获取稳定的流量，否则流量来的快去得也快。到 6 月中，颜值房间较王者荣耀房间，增长更快些。大有日操作量越前者之势。这说明每天直播，更容易积累粉丝。

（四）两个直播间的 24 小时操作量分布图

下一步我们继续将范围缩小，来研究 24 小时操作量的分布图，如图 5。

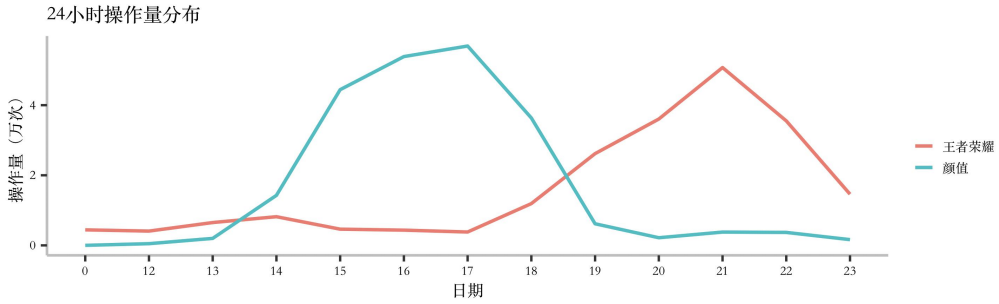

图 5 数据时间内粉丝 24 小时内操作分布

图中可以看到王者荣耀直播间的粉丝活跃时间是晚上。颜值直播间粉丝是活跃时间在下午。在颜值房间，晚上 7 点之后，粉丝就不再活跃了，而在下午 3、4、5 点左右，粉丝非常活跃，这说明人们喜欢利用下午的时间来放松自己，和自己喜欢的颜值主播聊天。而王者荣耀直播间的活跃时间主要是 21 点，这个时间也是很多游戏迷的娱乐时间，因此选择去看他们喜欢的游戏直播也是一个热门的方式，这也正符合了现在年轻人喜欢熬夜打游戏的心理。

（五）两个直播间不同类型操作的直方图

直播间的操作类型有哪些？我们可以主要将其分为三类：弹幕、登陆、礼物。其中弹幕是最常见的方式，在此不详细进行解释。登陆表示每次进入直播间的行为，这说明操作者对正要进入的直播间感兴趣。礼物指的是在直播间以金币形式购买礼物并送给主播的一种方式，这一般是粉丝对主播表达喜欢的一种方式，有的粉丝会花几千甚至上万的钱为主播刷礼物，这也是提高直播间人气的方式之一。为了探究两个直播间在此方面的差异，我们绘制了直方图，如图 6。

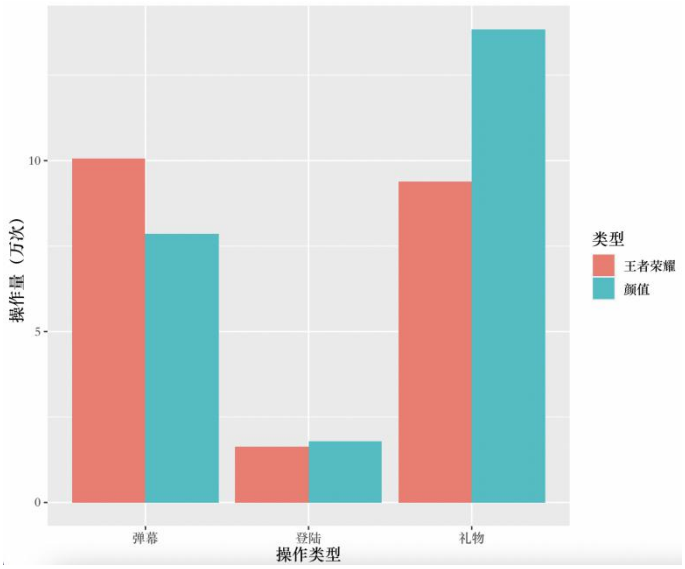

图 6 数据时间不同操作类型的分组直方图

王者荣耀直播间粉丝相比送礼物，更喜欢发弹幕与主播互动，很多人选择当一个聊天者，而不是选择作为金主。而颜值直播间粉丝则更喜欢送礼物，两个直播间的登陆操作基本持平。这是因为颜值直播更以刷礼物为主流，粉丝中更多是以人为吸引点去观看直播，因此会选择刷礼物的方式让主播记住自己，与主播进行更多互动，或者通过刷礼物让主播唱歌、表演等，而王者荣耀直播间则更娱乐化，更多的人是在观看主播的操作或者游戏效果，因此在主播玩的很好的时候，才会有更多的人送出礼物，而大部分时间可能都在观看或者聊天。因此，颜值主播一般会考虑利用直播刷礼物的方式进行“赚大钱”的行为，而王者荣耀主播则有更多的盈利方式，礼物不作为主要盈利来源。

四、RFM 模型

（一）指标的构建

RFM 模型是衡量客户价值和客户创造利益能力的重要工具和手段。在众多的客户关系管理的分析模式中，RFM 模型是被广泛提到的。该机械模型通过一个客户的近期购买行为、购买的总体频率以及花了多少钱 3 项指标来描述该客户的价值状况。因此，在本案例中，找到合适的指标是我们首先要考虑的事情。考虑本文数据，我们设定基于 RFM 模型的指标维度是：

- 近度 R（Recency）：最近一次操作距今小时数，标准化，R 值越小，粉丝的价值越高；
- 频度 F（Frequency）：平均操作时间间隔，标准化，F 值越高越小，粉丝的价值越高；
- 强度 M（Monetary）：操作总数，标准化，M 值越高，粉丝的价值越高。

（二）模型的建立

我们们选取两个房间的 TOP20%粉丝的操作记录，将 RFM 值分别按均值离散化，并初步将粉丝命名，后利用 k-means 聚类，选用上述指标，将粉丝聚类，得到两个直播间的聚类结果如下：-

]

.

七、结论与展望

本文基于国家卫健委发布的疫情数据，试图对未来一日的新增确诊人数增长率进行预测。基于增长率指标，我们建立线性回归模型，最终将新增确诊人数增长率、累计确诊人数增长率以及累计死亡人数增长率作为主要因素建立模型，得到了比较好的结果。模型对人们了解疫情走势有着重要的作用，可以帮助人们进行每日生活计划的制定，帮助国家进行疫情防控措施的制定，为新冠疫情预测局面提供良好的结果与解决方案。

此外，模型还可以在本文的基础上增加更多因素。例如，加入密切接触者、无症状感染者以及病毒的类型因素等，也可以增加预测的准确性，更好的进行建模。

附录 1　R 语言程序代码

```
##批量加载包
library(ggplot2)
library(reshape2)
library(cowplot)
library(tidyr)
library(gcookbook)
library(gridExtra)
library(rfm)
library(plyr)
library(data.table)
library(stringr)
library(lubridate)
dt1 <- read.csv("/Users/yuqinhan1229/Desktop/863.csv",header=TRUE)
dt2 <- read.csv("/Users/yuqinhan1229/Desktop/1038864.csv")
##输出所有弹幕
a <- dt1[dt1$act_type=="danmu",]$content
write.table(a,"/Users/yuqinhan1229/Desktop/863.txt",row.names=F,col.names    = F)
b <- dt2[dt2$act_type=="danmu",]$content
write.table(b,"/Users/yuqinhan1229/Desktop/1038864.txt",row.names=F,col.names
= F)

dt <- rbind(dt1,dt2)
##查看粉丝数量
##一周内从周一到周日操作量分布

weekday <- c('周日','周一','周二','周三','周四','周五','周六','周日','周一','周二','周三
','周四','周五','周六')
act <-
c(34910/6,39093/8,32684/4,10494/2,32339/6,11218/5,50239/7,30122/6,40374/7,2116
5/7,40638/6,35110/8,26218/7,41532/7)/10000
type <- c('王者荣耀','王者荣耀','王者荣耀','王者荣耀','王者荣耀','王者荣耀','王者
荣耀','颜值','颜值','颜值','颜值','颜值','颜值','颜值')
mydata <- data.frame(type,weekday,act)
week <- factor(mydata$week,levels = c('周一','周二','周三','周四','周五','周六','周日
'))
```

```
ggplot(mydata,aes(x=week,y
=act,fill=type))+geom_bar(stat="identity",position="dodge")+
   xlab("星期") + ylab("操作量（万次）") + labs(fill="类型")+
   theme(text=element_text(family="Songti SC",size=12,face = "bold"))

##日操作量分布
###将日期命名
dt1['date'] <- tidyr::unite(dt1[,9:10],"date",m,d,sep=".",remove=FALSE)
dt2['date'] <- tidyr::unite(dt2[,9:10],"date",m,d,sep=".",remove=FALSE)
##计算直播的星期天数
data <- dt2[!duplicated(dt2$date),]
table(data$weekday)
count1 <- as.data.frame(table(dt1$date))
count1$Var1 <- as.vector(count1$Var1)
count1$Freq <- as.vector(count1$Freq/10000)
count2 <- as.data.frame(table(dt2$date))
count2$Var1 <- as.vector(count2$Var1)
count2$Freq <- as.vector(count2$Freq/10000)
count <- merge(count1,count2,by='Var1')
mydata <- melt(count,id="Var1")
mydata$variable <- c(rep(x = '王者荣耀', times = 31), rep(x = '颜值', times = 31))
g1 <- ggplot(mydata,aes(factor(Var1),value,
                        color=variable,group=variable))+
   geom_line(cex=1.3)+
   theme_test(base_size = 20)+
   theme( panel.border = element_rect(size=2,fill = 'transparent'),
          axis.text = element_text(color='black'))+
    xlab("日期")+ylab("操作量（万次）")+scale_colour_hue(",labels=c('王者荣耀','
颜值'))+
   theme(text=element_text(size=13, family="Songti SC"))+
   theme(axis.ticks.x = element_blank())+
   theme(axis.text.x = element_blank())+
   theme(panel.border = element_blank())+     ## 删去外层边框
   theme(axis.line = element_line(size=1, colour = "grey")) +

geom_point(x=2,y=0,color="#CC99FF")+annotate('text',x=4.5,y=0,label=expression('
May 2017'),size=5,color = '#CC99FF')+
```

```
geom_point(x=17,y=0,color="#CC99FF")+annotate('text',x=19.5,y=0,label=expressio
n('June 2017'),size=5,color = '#CC99FF')+
   ggtitle("2017 年日操作量分布")

##24 小时操作量分布
count1 <- as.data.frame(table(dt1$H))
count1$Var1 <- as.vector(count1$Var1)
count1$Freq <- as.vector(count1$Freq/10000)
count2 <- as.data.frame(table(dt2$H))
count2$Var1 <- as.vector(count2$Var1)
count2$Freq <- as.vector(count2$Freq/10000)
count <- merge(count1,count2,by='Var1')
mydata <- melt(count,id="Var1")
mydata$variable <- c(rep(x = '王者荣耀', times = 13), rep(x = '颜值', times = 13))

g2 <- ggplot(mydata,aes(factor(Var1),value,
                        color=variable,group=variable))+
   geom_line(cex=1.3)+
   theme_test(base_size = 20)+
   theme( panel.border = element_rect(size=2,fill = 'transparent'),
          axis.text = element_text(color='black'))+
   xlab("日期")+ylab("操作量（万次）")+scale_colour_hue('',labels=c('王者荣耀','
颜值'))+
   theme(text=element_text(size=13, family="Songti SC"))+
   theme(panel.border = element_blank())+      ## 删去外层边框
   theme(axis.line = element_line(size=1, colour = "grey")) +
   ggtitle("24 小时操作量分布")
grid.arrange(g1,g2,nrow =2)
##不同类型的操作数量
count1 <- as.data.frame(table(dt1$act_type))
count1$Var1 <- as.vector(count1$Var1)
count1$Freq <- as.vector(count1$Freq/10000)
count2 <- as.data.frame(table(dt2$act_type))
count2$Var1 <- as.vector(count2$Var1)
count2$Freq <- as.vector(count2$Freq/10000)
count <- merge(count1,count2,by='Var1')
```

```
mydata <- melt(count,id="Var1")
mydata$Var1 <- rep(x = c('弹幕','登陆','礼物') ,times = 2)
mydata$variable <- c(rep(x = '王者荣耀', times = 3), rep(x = '颜值', times = 3))
ggplot(mydata,aes(x=Var1,y=value,fill=variable))+geom_bar(stat="identity",position
="dodge")+
   xlab("操作类型") + ylab("操作量（次）") + labs(fill="类型")+
   theme(text=element_text(family="Songti SC",size=12,face = "bold"))

##rfm
#数据处理
#R（Recency）：最近一次操作距今小时数，R 值越小，粉丝的价值越高；
#用户名降序，操作时间降序排列
order<-dt1[order(dt1$user_name,dt1$time,decreasing=T),]
#按用户名列去重，会保留最上面的那行
order_new<- order[!duplicated(order$user_name),]
new_name<- order_new$user_name          #新表格的用户名列
options(digits.secs=3)
order$time <- .POSIXct(order$time, tz="UTC")
order_new$time <-.POSIXct(order_new$time, tz="UTC")
dt_all = interval(order_new$time,ymd_hms("2017-07-08 00:00:00"))#区间
R<-time_length(dt_all,'hour') #时间间隔，新表格的 R 值(未标准化)

#频度 F（Frequency）：平均操作时间间隔，F 值越高越小，粉丝的价值越高；
#强度 M（Monetary）：操作总数，M 值越高，粉丝的价值越高。
#最早与最晚的时间间隔
name0<-ddply(order,.(user_name),function(order){data.frame(name0=time_length(int
erval(min(order$time),max(order$time))))})
name0<-name0[!is.na(name0$user_name),]  #删除用户名是 NA 的行
#每个用户的操作次数，也就是强度 M(未标准化)
name1<-aggregate(order$time, by=list(type=order$user_name),length)
#计算 F，未标准化，加到 x3_3 的最后一列
name1$F<-name0$name0/name1$x
```