project

Khan

August 15, 2018

R Markdown

I have used this data for my project for ST511. The reason for me to choose this data is to get the idea of the people's choice about the mobile application. This data reperesents the information of the most used mobile application in the mobile from the app store. This .csv file has 17 columns and 7197 rows. Following code shows thw head of my .csv file. So my interesting questions for this data is about the popularity of application type and money invested on the application. Since I have more than two groups so I would be doing ANOVA test to check whether mean amount spends on the games, education, health-fitness, and social-networking are same or not? I will be doing other test as well but let's start with the Anova first.

```
data<-read.csv(file="AppleStore.csv",header=TRUE,sep=",")
n=nrow(data)
n
```

```
## [1] 7197
```

```
levels(data$prime_genre)
```

```
##  [1] "Book"              "Business"          "Catalogs"
##  [4] "Education"         "Entertainment"     "Finance"
##  [7] "Food & Drink"      "Games"             "Health & Fitness"
## [10] "Lifestyle"         "Medical"           "Music"
## [13] "Navigation"        "News"              "Photo & Video"
## [16] "Productivity"      "Reference"         "Shopping"
## [19] "Social Networking" "Sports"            "Travel"
## [22] "Utilities"         "Weather"
```

```
GAM<-data[data$prime_genre=="Games",]
EDU<-data[data$prime_genre=="Education",]
NWS<-data[data$prime_genre=="News",]
SN<-data[data$prime_genre=="Social Networking",]
HF<-data[data$prime_genre=="Health & Fitness",]
total<-sum(nrow(GAM),nrow(EDU),nrow(NWS),nrow(SN),nrow(HF))
total
```

```
## [1] 4737
```

```
maindata<-rbind(GAM ,EDU , NWS , SN ,  HF)
```

I am interested to find the money spends on the game application (Games) , education application(Education), Social Networking, and Health & Fitness.As stated above my H_0 null hypothesis for the ANOVA test is avarage money spends on the Mobile application such as games, health-fitness, education and social-networking is same. In notational form μ_g=μ_s=μ_h=μ_e. And alternative hypothesis is H_A is $\mu_g \neq \mu_s \neq \mu_h \neq \mu_e$

```
# Create a vector of IDs which list which rows correspond to each diet.
IDG <- which(data$prime_genre == "Games")
IDE <- which(data$prime_genre == "Education")
IDH <- which(data$prime_genre == "Health & Fitness")
IDS <- which(data$prime_genre == "Social Networking")
IDN <- which(data$prime_genre == "News")



# Create a vector of lifetime means by group
groupmeans <- c(mean(data$price[IDG]), mean(data$price[IDE]),
mean(data$price[IDH]), mean(data$price[IDS]), mean(data$price[IDN]))
groupmeans

## [1] 1.4329234 4.0282340 1.9164444 0.3398802 0.5177333

SSW <- sum((data$price[IDG]-groupmeans[1])^2) +
  sum((data$price[IDE]-groupmeans[2])^2) +
  sum((data$price[IDH]-groupmeans[3])^2) +
  sum((data$price[IDS]-groupmeans[4])^2) +
  sum((mean(data$price[IDN]-groupmeans[5])^2))
SSW

## [1] 183342.8

gm<-mean(maindata$price)
gm

## [1] 1.646462

SST<-sum((maindata$price-gm)^2)
SST

## [1] 186576.6
```

Let’s calculate SSB as follows

```
SSB <- SST - SSW
SSB

## [1] 3233.783
```

Degrees of freedom

We know how many total observations there are and how many groups there are, so we can find the degrees of freedom.

```
n<-nrow(maindata)
n

## [1] 4737

I <- 5
dfW <- n - I
dfT <- n - 1
dfB <- I - 1
```

Note that the degrees of freedom BETWEEN is also the *Extra degrees of freedom*.

Mean squares

The mean square is the sum of squares divided by the corresponding degrees of freedom.

```
MSW <- SSW / dfW
MSW

## [1] 38.74531

MSB <- SSB / dfB
MSB

## [1] 808.4459
```

F-statistic

The F-statistic is the mean square BETWEEN divided by the mean square WITHIN:

```
F <- MSB / MSW
F

## [1] 20.86565
```

We also saw that we could think of it as the (extra sum of squares/extra degrees of freedom), divided by the pooled variance, where we know the pooled variance is equal to the mean square WITHIN. You can calculate this to show it's the same.

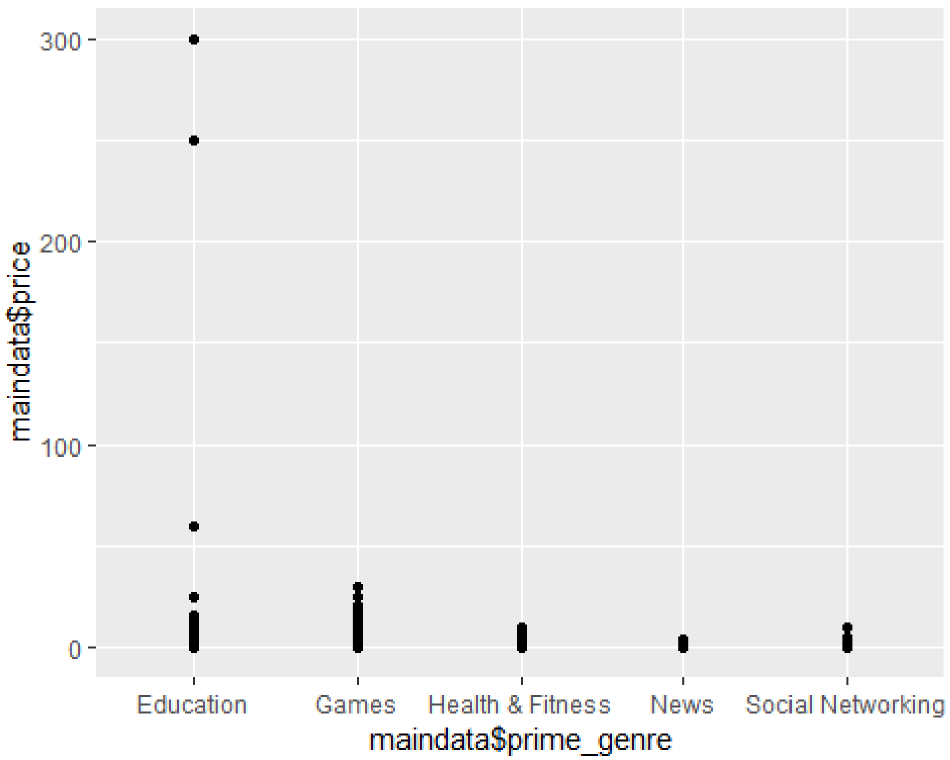
p-value

We want to look up the area in the F distribution with dfW and dfB degrees of freedom that is above our observed F-statistic.

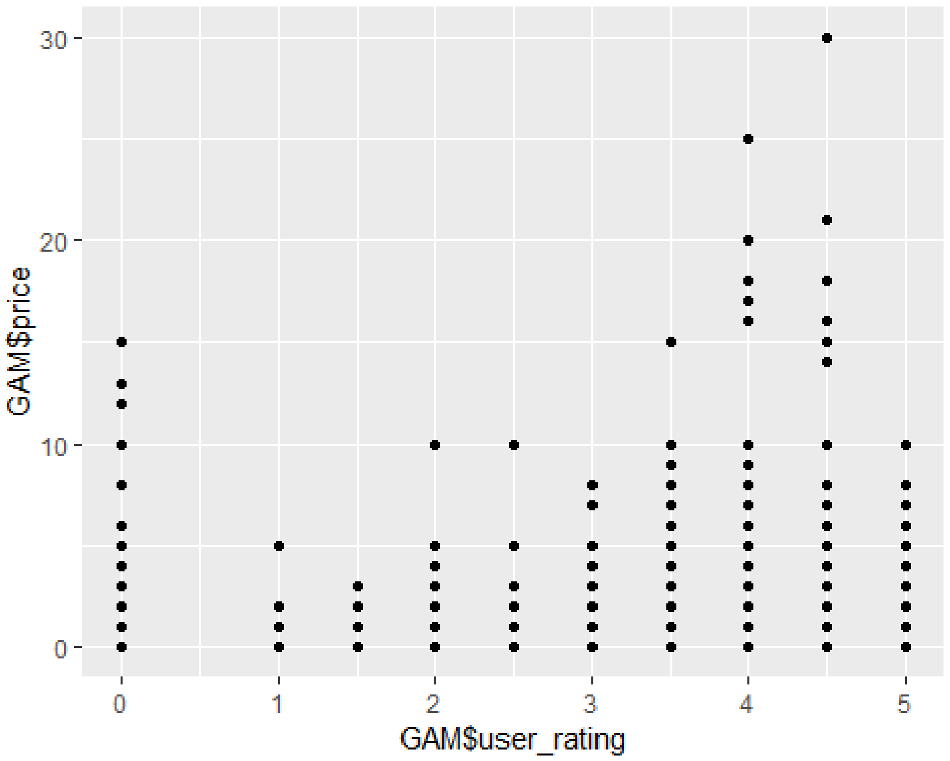
```
1-pf(F, df1=dfB, df2=dfW)

## [1] 0
```

This p-value is very, very small – so close to zero that 0 gets printed out due to rounding. We would reject the null hypothesis that the average money spend on the all the type of applications is same. We say that money spend on atleast one type of application is different than other. following code represents the anova test by using the inbuild function.

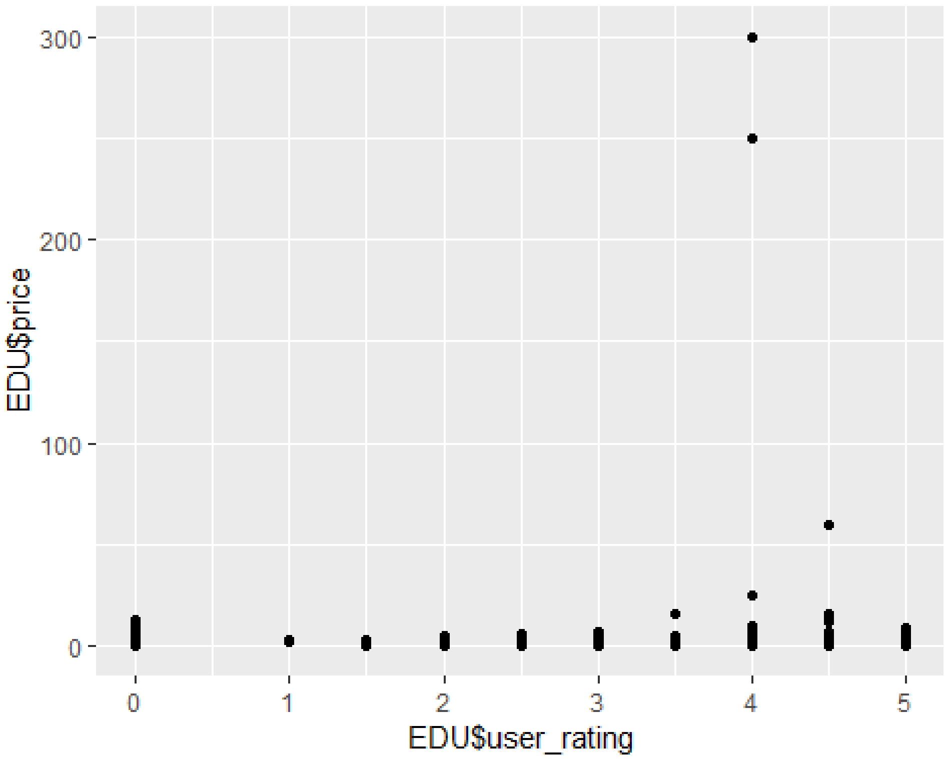
```
anova(lm(maindata$price ~ maindata$prime_genre, data=maindata))

## Analysis of Variance Table
##
## Response: maindata$price
##                        Df Sum Sq Mean Sq F value    Pr(>F)
## maindata$prime_genre    4   3140  784.92  20.248 < 2.2e-16 ***
## Residuals            4732 183437   38.77
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(aov(maindata$price ~ maindata$prime_genre, data=maindata))

##                        Df Sum Sq Mean Sq F value Pr(>F)
## maindata$prime_genre    4   3140   784.9   20.25 <2e-16 ***
## Residuals            4732 183437    38.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

kruskal.test(maindata$price ~ maindata$prime_genre, data=maindata)

##
##  Kruskal-Wallis rank sum test
##
## data:  maindata$price by maindata$prime_genre
## Kruskal-Wallis chi-squared = 257.96, df = 4, p-value < 2.2e-16

ggplot(maindata, aes(x=maindata$prime_genre, y=maindata$price)) +
geom_point()
```



```
ggplot(data = GAM, aes(GAM$user_rating, GAM$price)) +
  geom_point()
```



```
ggplot(data = EDU, aes(EDU$user_rating, EDU$price)) +
  geom_point()
```



This point graph shows that the variation of the data within in not much and hence F factor will be more since we have formula for F=MSB/MSW.