Ejercicio 4 smith

Jhonatan Smith Garcia

14/1/2022

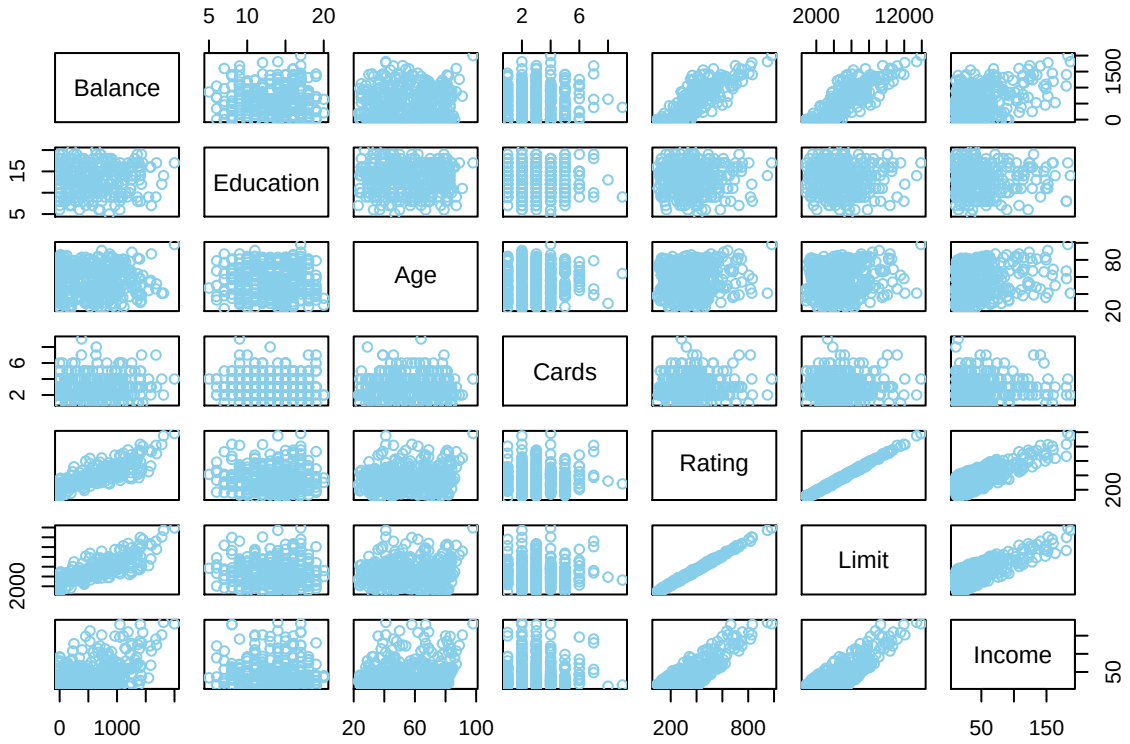
3)

Considere la base de datos "Credit" del paquete ISLR. Se procede a analizar un conjunto de datos simulados de 10.000 clientes. Suponga a "Balance" como variable de interes.

A continuacion, se realiza una breve descripcion de las variables de la base de datos:

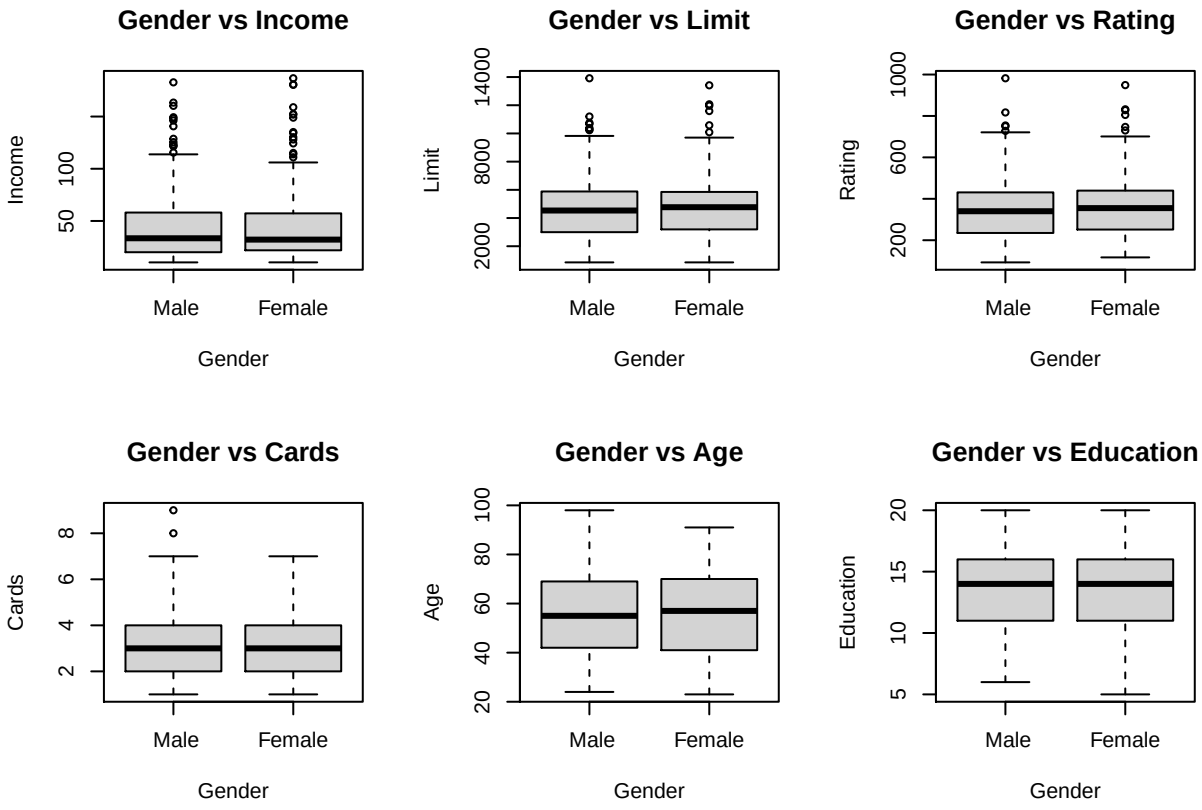
ID: Identificacion *Income:* Ingresos en $10.000 *Limit:* Límite de crédito *Rating:* Calificacion crediticia *Age:* edad *Education:* Numeros de años de estudio *Cards:* Numero de tarjetas de credito *Gender:* Genero *Student:* Un factor con niveles "Si" y "no" que indica si el individuo es o no estudiante *Married:* Factor con niveles "Si" y "No" Que indica si el individuo está casado *Ethnicity:* Un factor de niveles "Afrodescendiente", "Asiatico", "caucasico" que indica la etnia del individuo. *Balance:* saldo promedio de la tarjeta de credito en $

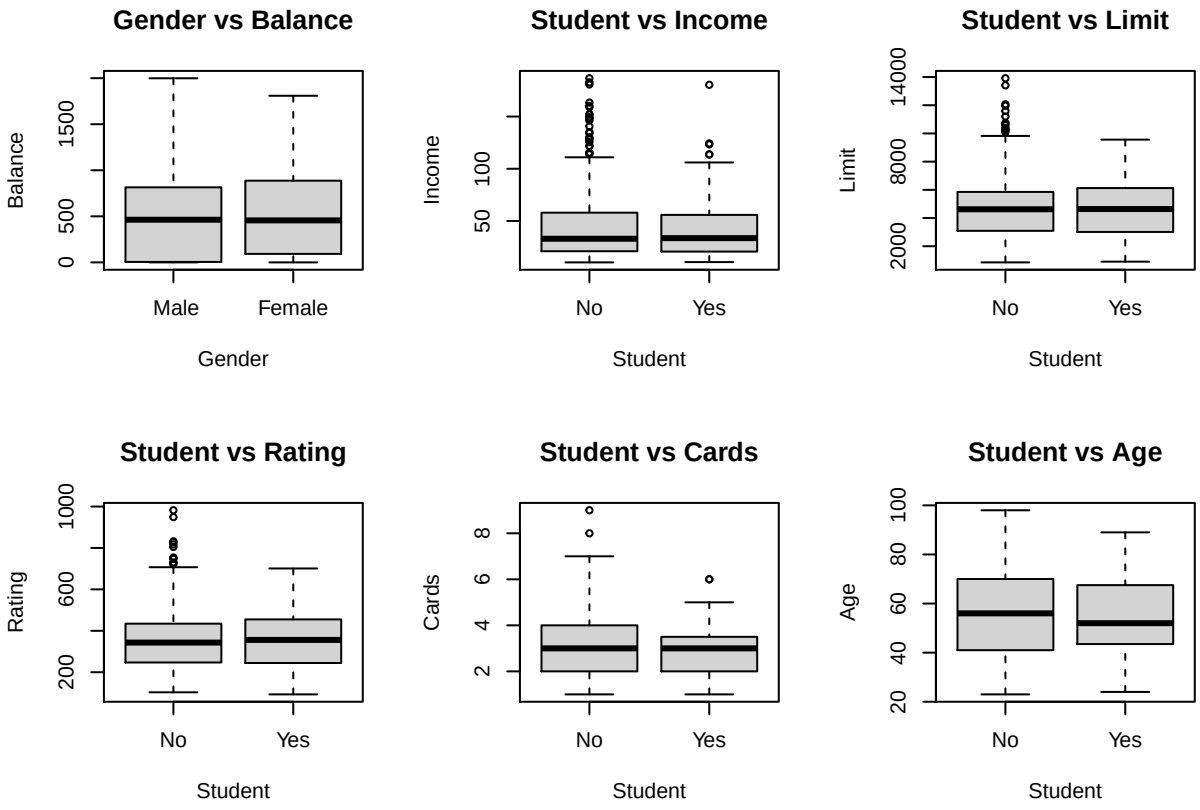
Un primer acercamiento a los datos se realiza con un analisis descriptivo de los mismos. Esto, con la intencion de en un principio, entender y ver el comportamiento de los datos dado el problema planteado (predecir acorde a la variable de interes Balance).

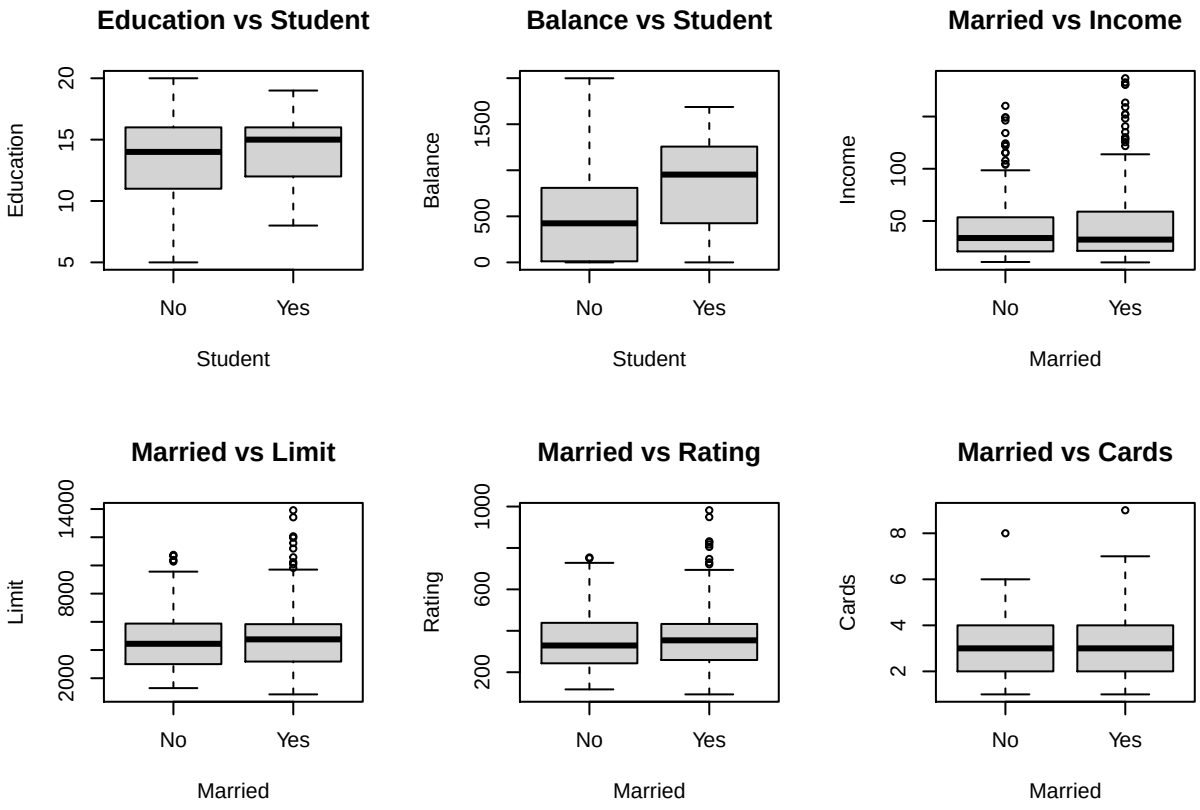


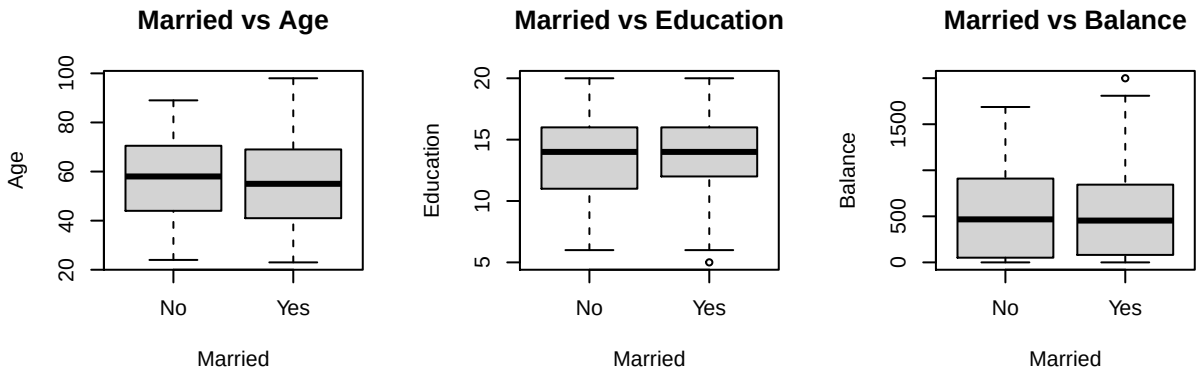
Este grafico se utiliza para ver tendencia entre los datos de tipo numerico. Se observa que existe relaciones entre ciertas variables. Por ejemplo, Limit y Rating tienen una tendencia positiva, asi como Limit e Income.

La variable respuesta Balance se relaciona de manera positiva con limit y rating.


Gender vs Income
Income
50
100
Male
Female
Gender
Gender vs Limit
Limit
2000
8000
14000
Male
Female
Gender
Gender vs Rating
Rating
200
600
1000
Male
Female
Gender
Gender vs Cards
Cards
2
4
6
8
Male
Female
Gender
Gender vs Age
Age
20
40
60
80
100
Male
Female
Gender
Gender vs Education
Education
5
10
15
20
Male
Female
Gender

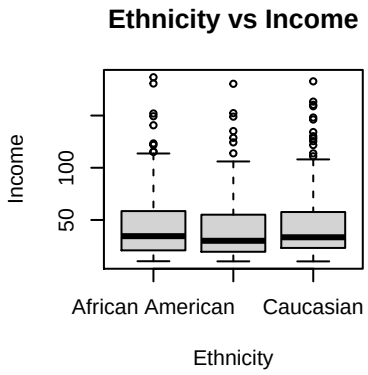

Gender vs Balance
Balance
0
500
1500
Male
Female
Gender
Student vs Income
Income
50
100
No
Yes
Student
Student vs Limit
Limit
2000
8000
14000
No
Yes
Student
Student vs Rating
Rating
200
600
1000
No
Yes
Student
Student vs Cards
Cards
2
4
6
8
No
Yes
Student
Student vs Age
Age
20
40
60
80
100
No
Yes
Student

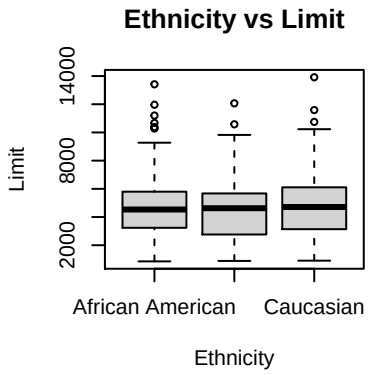

Education vs Student
Education
5
10
15
20
No
Yes
Student
Balance vs Student
Balance
0
500
1500
No
Yes
Student
Married vs Income
Income
50
100
No
Yes
Married
Married vs Limit
Limit
2000
8000
14000
No
Yes
Married
Married vs Rating
Rating
200
600
1000
No
Yes
Married
Married vs Cards
Cards
2
4
6
8
No
Yes
Married

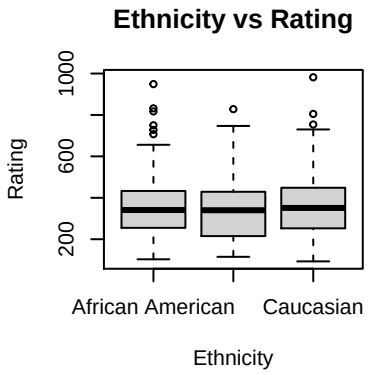

Married vs Age
Age
20
40
60
80
100
No
Yes
Married

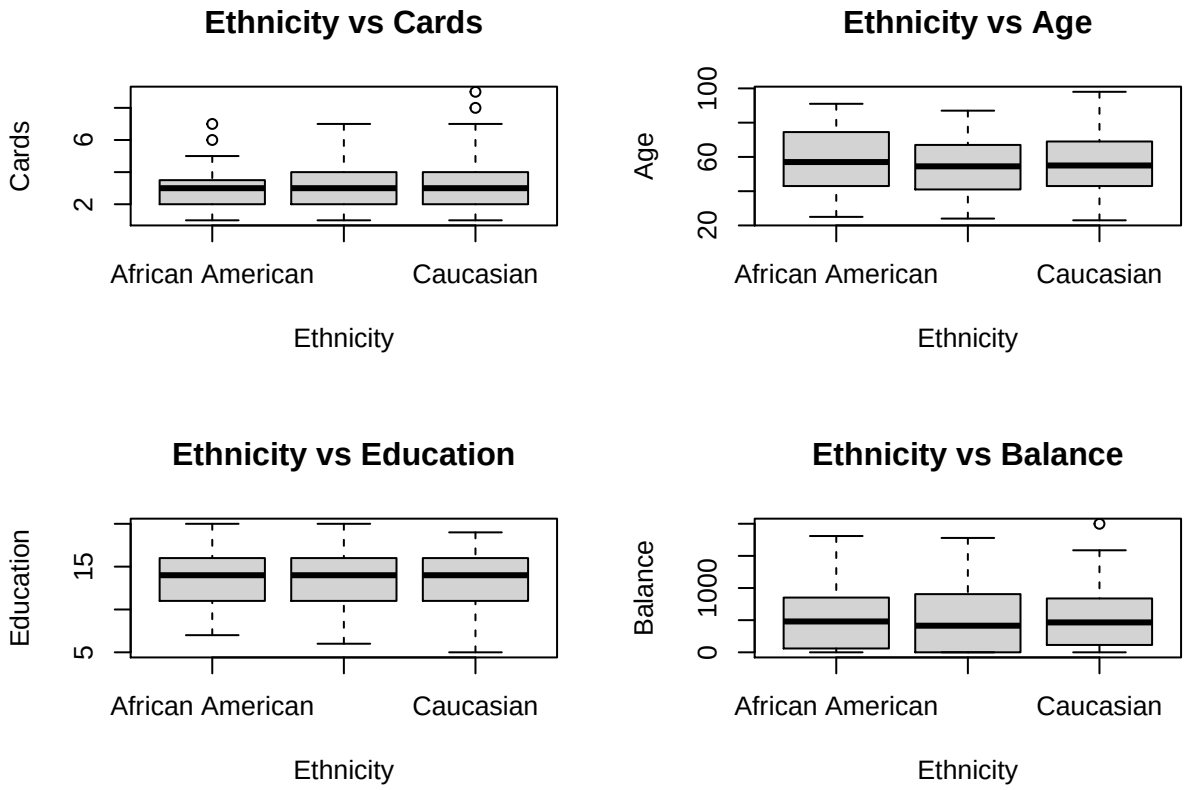
Married vs Education
Education
5
10
15
20
No
Yes
Married

Married vs Balance
Balance
0
500
1500
No
Yes
Married


Ethnicity vs Income
Income
50
100
African American
Caucasian
Ethnicity


Ethnicity vs Limit
Limit
2000
8000
14000
African American
Caucasian
Ethnicity


Ethnicity vs Rating
Rating
200
600
1000
African American
Caucasian
Ethnicity



El anterior es un analisis descriptivo de las variables categoricas vs continuas. En general no se observan diferencias por categorias exceptuando *Student vs Balance* donde las personas que fueron estudiantes en promedio, tenian un saldo de tarjetas mas altos.

Dado en analisis anterior se decide tomar el conjunto de variables Income, Limit, Student y Rating para modelar a Balance. Se ajustan los modelos GAM anidados de la siguiente manera.

```
require(gam)

modelo.gam.1 <- gam(Balance~Limit+Rating, data = Credit)
modelo.gam.2 <- gam(Balance~Limit+Rating+Student, data = Credit)
modelo.gam.3 <- gam(Balance~Limit+Rating+Student+Income, data = Credit)
modelo.gam.4 <- gam(Balance~Limit+Rating+Student+s(Income, df = 3), data = Credit)

anova(modelo.gam.1,modelo.gam.2,modelo.gam.3,modelo.gam.4, test = "F")
```

```
## Analysis of Deviance Table
##
## Model 1: Balance ~ Limit + Rating
## Model 2: Balance ~ Limit + Rating + Student
## Model 3: Balance ~ Limit + Rating + Student + Income
## Model 4: Balance ~ Limit + Rating + Student + s(Income, df = 3)
##   Resid. Df Resid. Dev Df Deviance        F  Pr(>F)
## 1       397   21427162
## 2       396   15655583  1  5771579  571.624 < 2e-16 ***
## 3       395    4032502  1 11623082 1151.165 < 2e-16 ***
## 4       393    3968044  2    64457    3.192 0.04216 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se observa que el p-value de los modelos 2 y 3 es pequeño. Esto se traduce en que las variables Income y Student son de importancia para explicar el comportamiento de Balance. Sin embargo, esto complica la variable Income con dos nodos ya que no hay un aporte significativo para el modelo. En general, aplicar un modelo con n nodos (n>0) complica el modelo y no hay una ganancia significativa.

La seleccion del modelo para explicar Balance (Saldo en Tarjeta de Credito), se procede a realizar validacion cruzada (CV) con la base de datos, dividiendo un 70% para entrenamiento y un 30% para prueba.

```
set.seed(1998)

prop = 0.7
t1 = sample(length(Credit$Balance), size = (length(Credit$Balance)*prop))

train = Credit[t1,]
test <- Credit[-t1,]
y = train$Balance
y1 <- test$Balance
```

Esta es la forma en que se seleccionan las bases de datos para ajustar cada uno de los modelos anterior mente seleccionados. Se procede a analizar el MSE de los modelos y dado este criterio se decide cual modelo es el más optimo a trabajar.

```
modelo.gam.1.t <- gam(Balance~Limit+Rating, data = train)
modelo.gam.2.t <- gam(Balance~Limit+Rating+Student, data = train)
modelo.gam.3.t <- gam(Balance~Limit+Rating+Student+Income, data = train)
modelo.gam.4.t <- gam(Balance~Limit+Rating+Student+s(Income, df = 3), data = train)
```

Note algo. Se parte del modelo mas sencillo, hasta un modelo mas complejo. Se procede a realizar un analisis de varianza con la tabla ANOVA de los datos de entrenamiento.

```
anova(modelo.gam.1.t, modelo.gam.2.t, modelo.gam.3.t,modelo.gam.4.t, test = "F")
```

```
## Analysis of Deviance Table
##
## Model 1: Balance ~ Limit + Rating
## Model 2: Balance ~ Limit + Rating + Student
## Model 3: Balance ~ Limit + Rating + Student + Income
## Model 4: Balance ~ Limit + Rating + Student + s(Income, df = 3)
##   Resid. Df Resid. Dev     Df Deviance        F Pr(>F)
## 1       277   15861831
## 2       276   11656437 1.0000  4205394 411.3955 <2e-16 ***
## 3       275    2831532 1.0000  8824904 863.3022 <2e-16 ***
## 4       273    2790682 1.9997    40850   1.9984 0.1375
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note que el modelo 4, que utiliza un spline (complicando el proceso) no tuvo un efecto significativo. Por este motivo, el modelo 4 es descartado.

Ahora, el modelo 2 y 3 son significativos, lo que indica que el agregar las variables (partiendo del modelo 1) y llegar hasta el modelo 3, es el modelo que mejor explica la proporcion de varianza. Surje una pregunta. ¿Que variable tiene papel preponderante para el modelo?

```
## Anova for Parametric Effects
##            Df   Sum Sq  Mean Sq  F value    Pr(>F)
## Limit       1 42018376 42018376 4080.848 < 2.2e-16 ***
```

```
## Rating      1   272874   272874   26.502 5.021e-07 ***
## Student     1  4205394  4205394  408.430 < 2.2e-16 ***
## Income      1  8824904  8824904  857.080 < 2.2e-16 ***
## Residuals 275  2831532    10296
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En general, la variable que mayor proporcion de varianza explica de Balance es Limit. Esto se cumple para todos los otros modelos. Esto indica que Limit es una variable que si o si, debe de ir en cualquier modelo que sea seleccionado. Todas las variables son significativas.

Se procede a analizar el AIC de cada modelo.

```
## Loading required package: magrittr
```

```
## Loading required package: kableExtra
```

```
gam1<-modelo.gam.1.t$aic
gam2<-modelo.gam.2.t$aic
gam3<-modelo.gam.3.t$aic
gam4<-modelo.gam.4.t$aic

tabla = cbind(c(gam1,gam1, gam3, gam4))
colnames(tabla) = c("AIC")
rownames(tabla )=c("modelo 1", "modelo 2", "modelo 3", "modelo 4")
tabla
```

```
##               AIC
## modelo 1 3867.104
## modelo 2 3867.104
## modelo 3 3388.636
## modelo 4 3388.567
```

Por criterio de AIC, se escoje el modelo 3 nuevamente. Si bien, no es el que tiene el valor minimo, su diferencia es infima respecto a la ganancia obtenida vs la complejidad del mejor (modelo 4) por tal motivo, se escoje nuevamente el modelo 3. Ahora, se procede a comparar los MSE de todos los modelos entre si con los datos de prueba.

```
mse1 <- mean((y1- predict(modelo.gam.1.t, test))^2)
mse2 <- mean((y1- predict(modelo.gam.2.t, test))^2)
mse3 <- mean((y1- predict(modelo.gam.3.t, test))^2)
mse4 <- mean((y1- predict(modelo.gam.4.t, test))^2)

tabla1 = cbind(c(mse1, mse2,mse3,mse4))
colnames(tabla1) = c("MSE")
rownames(tabla1)=c("modelo 1 ", "modelo 2 ", "modelo 3 ", "modelo 4 ")
tabla1
```

```
##                MSE
## modelo 1  46590.26
## modelo 2  33623.41
## modelo 3  10128.27
## modelo 4  10005.29
```

El modelo con un menor MSE es el modelo 4. Sin embargo, la ganancia no es mucha respecto al modelo 3 dado que utiliza splines en el proceso. Finalmente, el modelo mas optimo dado el principio de parsimonia y, dado el criterio del MSE (No necesariamente el menor) y AIC, el modelo a seleccionar es el modelo que predice Balance con las variables Limit, Rating, Student e Income.