

Tarea 1

Estudiantes

**John Daniel hoyos Arias
Ivan Santiago Rojas Martinez
Genaro Alfonso Aristizabal Echeverri**

Docente

Juan Carlos Salazar Uribe

Asignatura

Analitica de datos



Sede Medellín
17 de septiembre del 2022

Índice

1. Ejercicio1	4
2. Ejercicio2	6
3. Ejercicio3	6
3.1. K-nearest neighbors (KNN)	6
3.2. a) Distancia a cada observación	7
3.3. b) Predicción para $K = 1$	8
3.4. c) Predicción para $K = 3$	8
3.5. d) Frontera de decisión de Bayes	8
4. Ejercicio4	9
4.1. a) Use the <code>read.csv()</code> function to read the data into R. Call the loaded data <code>college</code> . Make sure that you have the directory set to the correct location for the data	9
4.2. b) Look at the data using the <code>fix()</code> function. You should notice that the first column is just the name of each university. We don't really want R to treat this as data. However, it may be handy to have these names for later. Try the following commands:	9
4.3. c) Use the <code>summary()</code> function to produce a numerical summary of the variables in the data set.	10
4.4. II) Use the <code>pairs()</code> function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using <code>A[,1:10]</code>	10
4.5. III) Use the <code>plot()</code> function to produce side-by-side boxplots of <code>Outstate</code> versus <code>Private</code>	11
4.6. V) Use the <code>hist()</code> function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command <code>par(mfrow=c(2,2))</code> useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.	13
4.7. VI) Continue exploring the data, and provide a brief summary of what you discover.	14

Índice de figuras

1. Ejercicio 1

Se tiene que la tasa de error de prueba está dada por:

$$\text{Average}(I(y_0 \neq \hat{y}_0)) = E[I(y_0 \neq \hat{y}_0)]$$

Se supone que se está trabajando con dos clases (Aunque también se puede generalizar para n clases). Donde se suponen los siguientes datos de prueba:

$$y_{0i} = \begin{cases} 0, & i = 1, 2, \dots, n; \text{ Siendo } x_{0i} \text{ los datos de prueba} \\ 1 \end{cases}$$

Y su respectivo clasificador de Bayes está dado por

$$P(y_{0i} = j | X_{0i} = x_{0i}), \text{ con } j = 0, 1$$

Donde

$$\underbrace{P(y_{0i} = 0 | x_{0i})}_{P_0} > \text{ ó } < \underbrace{P(y_{0i} = 1 | x_{0i})}_{P_1}$$

Se toma la máxima probabilidad de las dos probabilidades condicionales anteriores, es decir $\max\{P_0, P_1\}$

El error se minimiza cuando $\max\{P_0, P_1\}$ lleva a aciertos, el cual se reduce al minimizar la tasa de error promedio cuando se usa una indicadora (Clasificador):

La siguiente igualdad es un resultado de probabilidad (Solo se cumple para la indicadora):

$$\underbrace{E[I(y_{0i} \neq \hat{y}_{0i})] | x_{0i}}_A = P(y_{0i} \neq \hat{y}_{0i} | x_{0i}) = 0 * P(I_A = 0) + 1 * (I_A = 1)$$

Se desea minimizar

El menor valor que puede obtener es:

$$E[I_A | x_{0i}] = 0 \text{ que sucede cuando } y_{0i} = \hat{y}_{0i} \implies (y_{0i} - \hat{y}_{0i}) = 0 \implies (y_{0i} - \hat{y}_{0i})^2 = 0$$

Seguidamente, se intentará probar que $y_{0i} = \hat{y}_{0i}$ produce el menor error Cuando se usa el método de Bayes.

Sea $f(y) = (y - \hat{y})^2$. Donde $E[f(y)] = E[(y - \hat{y})^2]$ es el MSE , que es mínimo justamente cuando y_i es el promedio de los errores.

Posteriormente, se utiliza una variable binaria que indique cuando se comete o no un error:

$I\{y_{0i} \neq \hat{y}_{0i}\}$ el cual toma valores de $\begin{cases} 0 \\ 1 \end{cases}$ al igual que y_{0i} y el clasificador \hat{y}_{0i} .

Entonces $(y_{0i} - \hat{y}_{0i})^2 = I\{y_{0i} \neq \hat{y}_{0i}\}$, esta igualdad se prueba a continuación:
sea:

$$y_{0i} = I\{y_{0i} = j\} \quad \hat{y}_{0i} = I\{\hat{y}_{0i} = j\} \longrightarrow \text{ambas son indicadoras.}$$

Utilizando la función de perdida

$$\delta(x) = (y_{0i} - \hat{y}_{0i})^2 = [I\{y_{0i} = j\} - I\{\hat{y}_{0i} = j\}]^2$$

Se prueba que:

$[I\{y_{0i} = j\} - I\{\hat{y}_{0i} = j\}]^2$	$I\{y_{0i} \neq \hat{y}_{0i}\}$
$(1 - 1)^2 = 0$	0
$(1 - 0)^2 = 1$	1
$(0 - 1)^2 = 1$	1
$(0 - 0)^2 = 0$	0

Así que $(y_{0i} - \hat{y}_{0i})^2$ es equivalente a $I\{y_{0i} \neq \hat{y}_{0i}\}$

Este proceso de clasificación esta basado en la regla de clasificación de Bayes por lo tanto minimiza $(y_{0i} - \hat{y}_{0i})^2$ y minimiza la tasa de error de prueba.

Esta expresión anterior se demostrará a continuación:

Se define la función de pérdida posterior como:

$$L(y_{0i}, \delta(x)) = (y_{0i} - \hat{y}_{0i})^2 = [I\{y_{0i} = j\} - I\{\hat{y}_{0i} = j\}]^2$$

Consideremos $y = y_{0i}$, $x = x_{0i}$ para facilitar el proceso algebraico a continuación:

$$L(y, \delta(x)) = (y - \hat{y})^2 = [I\{y = j\} - I\{\hat{y} = j\}]^2$$

Sea la función de pérdida posterior esperada:

$$\begin{aligned}
 \gamma(y, \delta(x)) &= E[L(y, \delta(x))|X = x] \\
 &= E[(y - \hat{y})^2|X = x] \\
 &= E[(y - E[y|x] + E[y|x] - \hat{y})^2|X = x] \\
 &= E[(y - E[y|x])^2 + (E[y|x] - \hat{y})^2 + 2(y - E[y|x])(E[y|x] - \hat{y})|X = x] \\
 &= E[(y - E[y|x])^2|X = x] + (E[y|x] - \hat{y})^2 + \underbrace{2 E[(y - E[y|x])(E[y|x] - \hat{y})|X = x]}_{\blacksquare}
 \end{aligned}$$

$$\begin{aligned}
 \blacksquare E[(y - E[y|x])(E[y|x] - \hat{y})|X = x] &= E[y \cdot E[y|x] - y \cdot \hat{y} - E^2[y|x] + \hat{y} \cdot E[y|x]|X = x] \\
 &= \color{blue}{E[y|x]E[y|x]} - \color{red}{\hat{y}E[y|x]} - \color{blue}{E^2[y|x]} + \color{red}{\hat{y}E[y|x]} \\
 &= 0
 \end{aligned}$$

Por lo tanto

$$\gamma(y, \delta(x)) = E[(y - E[y|x])^2 | X = x] + (E[y|x] - \hat{y})^2 \geq \underbrace{E[(y - E[y|x])^2 | X = x]}_{\text{Cota inferior: m\'ınimo}}$$

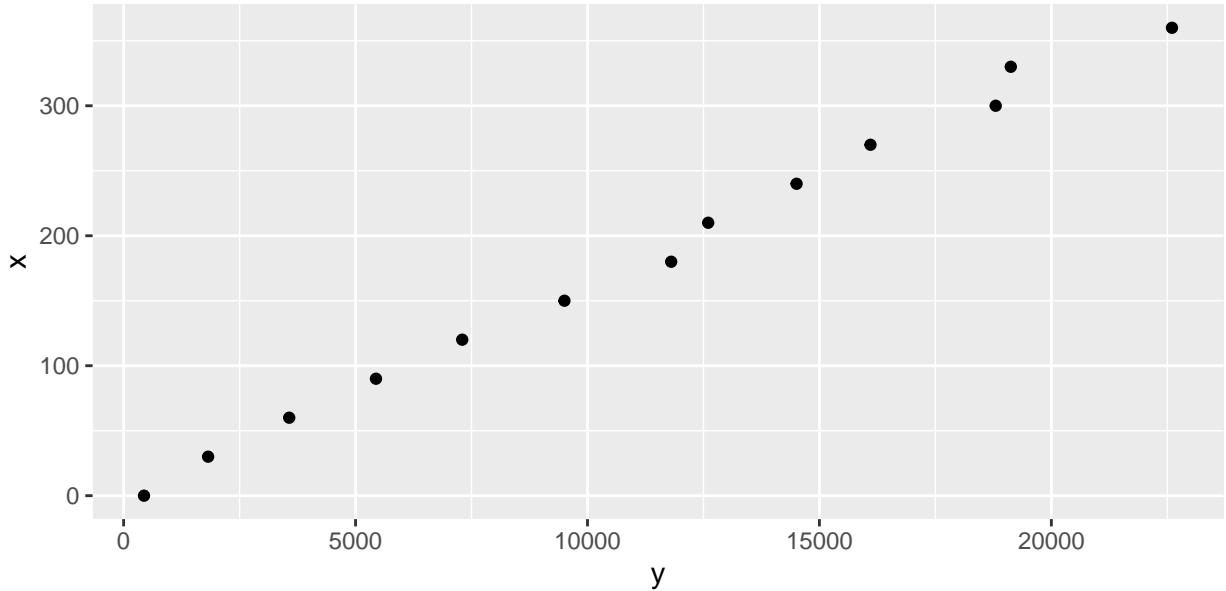
Por lo tanto el m\'ınimo valor que toma $\gamma(y, \delta(x))$ es:

$$E[(y - E[y|x])^2 | X = x] \text{ que es cuando } (E[y|x] - \hat{y})^2 = 0$$

Dado que $E[(y - E[y|x])^2 | X = x]$ es la cota m\'ınima, se prueba que esta utiliza el estimador de Bayes para estimar \hat{y} y lo esta utilizando mediante $E[y|x]$: ($E[(y - E[y|x])^2 | X = x]$) donde:

$$\hat{y} = E[y|x] = \underbrace{P(y|x)}_{\text{Estimador de Bayes}}$$

2. Ejercicio2



3. Ejercicio3

3.1. K-nearest neighbors (KNN)

$$Pr(Y = J | X = x_0) \approx \frac{1}{K} \sum_{i \in N_0} I(y_i = j)$$

Cuadro 1: Base de datos

X1	X2	X3	Y
0	3	0	Red
2	0	0	Red
0	1	3	Red
0	1	2	Green
-1	0	1	Green
1	1	1	Red

3.2. a) Distancia a cada observación

Usando la distancia euclíadiana entre dos punto u y v definida como:

$$d(u, v) = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + (u_3 - v_3)^2}$$

Calculamos la distancia entre cada observación y el punto P con características $X_1 = X_2 = X_3 = 0$

- $d(p, x_1) = \sqrt{(p_1 - x_{1,1})^2 + (p_2 - x_{1,2})^2 + (p_3 - x_{1,3})^2} = \sqrt{(0 - 0)^2 + (0 - 3)^2 + (0 - 0)^2} = \sqrt{0 + 9 + 0} = \sqrt{9} = 3$
- $d(p, x_2) = \sqrt{(p_1 - x_{2,1})^2 + (p_2 - x_{2,2})^2 + (p_3 - x_{2,3})^2} = \sqrt{(0 - 2)^2 + (0 - 0)^2 + (0 - 0)^2} = \sqrt{4 + 0 + 0} = \sqrt{4} = 2$
- $d(p, x_3) = \sqrt{(p_1 - x_{3,1})^2 + (p_2 - x_{3,2})^2 + (p_3 - x_{3,3})^2} = \sqrt{(0 - 0)^2 + (0 - 1)^2 + (0 - 3)^2} = \sqrt{0 + 1 + 9} = \sqrt{10} = 3.162278$
- $d(p, x_4) = \sqrt{(p_1 - x_{4,1})^2 + (p_2 - x_{4,2})^2 + (p_3 - x_{4,3})^2} = \sqrt{(0 - 0)^2 + (0 - 1)^2 + (0 - 2)^2} = \sqrt{0 + 1 + 4} = \sqrt{5} = 2.236068$
- $d(p, x_5) = \sqrt{(p_1 - x_{5,1})^2 + (p_2 - x_{5,2})^2 + (p_3 - x_{5,3})^2} = \sqrt{(0 + 1)^2 + (0 - 0)^2 + (0 - 1)^2} = \sqrt{1 + 0 + 1} = \sqrt{2} = 1.414214$
- $d(p, x_6) = \sqrt{(p_1 - x_{6,1})^2 + (p_2 - x_{6,2})^2 + (p_3 - x_{6,3})^2} = \sqrt{(0 - 1)^2 + (0 - 1)^2 + (0 - 1)^3} = \sqrt{1 + 1 + 1} = \sqrt{3} = 1.732051$

Ahora procedemos a calcularla con R:

```
point <- c(0, 0, 0)
dist_eucl <- function(x){
```

```

ans <- c()
for (i in 1:nrow(x)){
  xi <- as.numeric(t(as.vector(x[i, ])))
  result <- sqrt(sum((xi-point)^2))
  ans <- append(ans, result)
}
return (ans)
}

db <- mutate(db, dist = dist_eucl(db[1:3]))

```

Cuadro 2: Distancia a cada observación desde el punto $X_1 = X_2 = X_3 = 0$

Observación	Grupo	Distancia Euclidiana
1	Red	3.000000
2	Red	2.000000
3	Red	3.162278
4	Green	2.236068
5	Green	1.414214
6	Red	1.732051

3.3. b) Predicción para $K = 1$

Con una selección de $K = 1$. Knn identifica la observación más cercana al punto con características $X_1 = X_2 = X_3 = 0$ y en este caso la observación mas cercana es la **numero 5** con una distancia de **1.414214**. Dando así Knn una estimación de 1/1 de pertenecer al grupo **Green**. Por ende la estimación es pertenecer a la clase **Green**.

3.4. c) Predicción para $K = 3$

Con una selección de $K = 3$. Knn identifica las 3 observaciones más cercanas al punto con características $X_1 = X_2 = X_3 = 0$ y en este caso las observaciones mas cercana son la **numero 5**, la **numero 6** y la **numero 2** que consisten en 2 observaciones de la clase **Red** y una observación de la clase **Green**, dando como resultado una estimación de 2/3 de pertenecer a la clase **Red** y 1/3 de pertenecer a la clase **Green**. Por consiguiente se estima pertenecer a la clase **Red**.

3.5. d) Frontera de decisión de Bayes

Si la frontera de decisión de Bayes en este problema es altamente no lineal, ¿esperaríamos que el mejor valor de K fuera grande o pequeño? ¿Por qué?

A medida que K crece, el método se vuelve menos flexible y produce un límite de decisión cercano a lineal. Teniendo en cuenta esto, es de esperarse que el mejor valor para K sea pequeño.

4. Ejercicio4

- 4.1. a) Use the `read.csv()` function to read the data into R. Call the loaded data `college`. Make sure that you have the directory set to the correct location for the data**

Inicialmente cargamos los datos de la siguiente manera:

```
library(ISLR)
College=ISLR::College
kable(head(College))
```

	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Unc
Abilene Christian University	Yes	1660	1232	721	23	52	
Adelphi University	Yes	2186	1924	512	16	29	
Adrian College	Yes	1428	1097	336	22	50	
Agnes Scott College	Yes	417	349	137	60	89	
Alaska Pacific University	Yes	193	146	55	16	44	
Albertson College	Yes	587	479	158	38	62	

- 4.2. b) Look at the data using the `fix()` function. You should notice that the first column is just the name of each university. We don't really want R to treat this as data. However, it may be handy to have these names for later. Try the following commands:**

Para esta sección, mostramos la estructura de los datos, la cual, a groso modo cuenta con datos de 777 universidades

```
fix(College)
rownames(College)=College[,1]
College=College [,-1]
fix(College)
```

4.3. c) Use the summary() function to produce a numerical summary of the variables in the data set.

Luego, calculamos el resumen estadístico general para todas las variables de la base de datos, #este resumen es muy importante, dado que nos da una mejor visión sobre la estructura y el comportamiento de nuestra información.

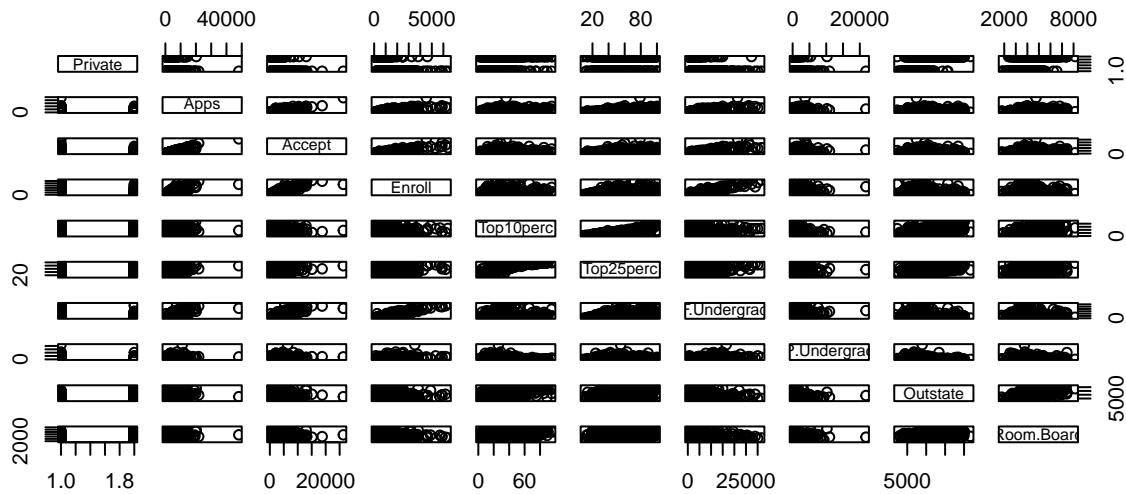
Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.U.
No :212	Min. : 81	Min. : 72	Min. : 35	Min. : 1.00	Min. : 9.0	Min.
Yes:565	1st Qu.: 776	1st Qu.: 604	1st Qu.: 242	1st Qu.:15.00	1st Qu.: 41.0	1st
NA	Median : 1558	Median : 1110	Median : 434	Median :23.00	Median : 54.0	Med
NA	Mean : 3002	Mean : 2019	Mean : 780	Mean :27.56	Mean : 55.8	Mea
NA	3rd Qu.: 3624	3rd Qu.: 2424	3rd Qu.: 902	3rd Qu.:35.00	3rd Qu.: 69.0	3rd
NA	Max. :48094	Max. :26330	Max. :6392	Max. :96.00	Max. :100.0	Max

#por ejemplo, en promedio el numero de estudiantes matriculados son 780 por universidad, el costo promedio de los libros #es aproximadamente 549.4 dólares, la cantidad promedio de empleados para cada universidad es de 13441 personas, otros datos interesantes como #La razón promedio de graduación que es de 65.46, indica que de cada 100 estudiantes aproximadamente 66 se gradúan.

4.4. II) Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10].

En esta sección realizamos un diagrama de dispersión, con el fin de observar el grado de asociación lineal entre las primeras 10 variables cuantitativas, el resultado fue el siguiente:

```
pairs(College[,1:10])
```

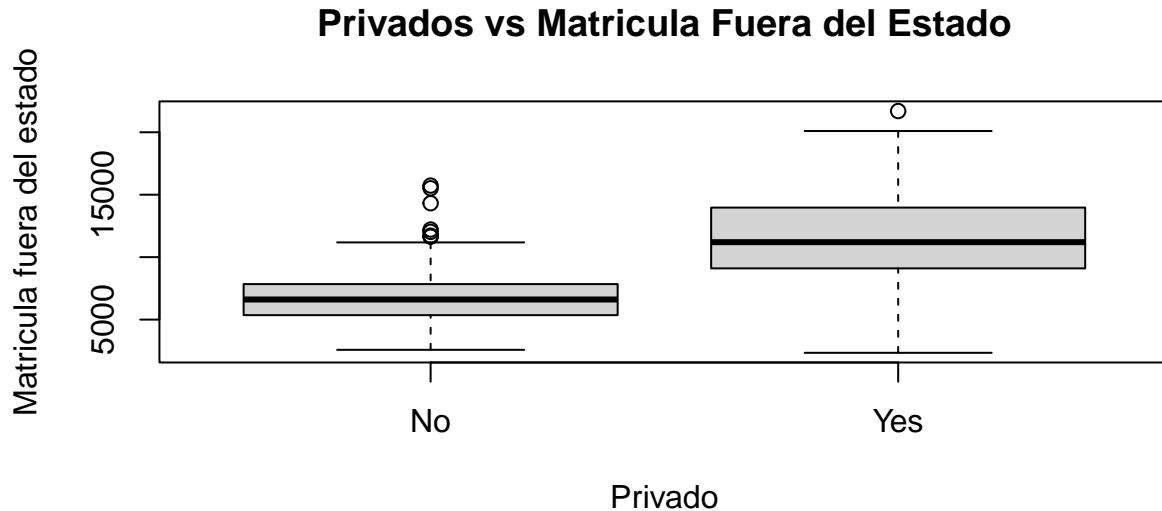


#initialmente observamos una alta asociación entre las variables número de aplicaciones recibidas(apps) y numero de aplicaciones #aceptadas(accept), esto indica que en general, que a medida que aumenta la recepción de aplicaciones aumenta también su #aceptacion, lo cual tiene mucho sentido. #de la misma manera existe una alta asociación entre aplicaciones aceptadas(accept) y #número de estudiantes matriculados(enroll) #otra relación bastante fuerte es entre los estudiantes nuevos que hacen parte del 10 % y 25 % superior de secundaria, #inidica que hacer parte de estos porcentajes puede incrementar la probabilidad e asistir a una universidad.

4.5. III) Use the plot() function to produce side-by-side boxplots of Outstate versus Private.

#luego, agregamos un grafico entre universidad publica o privada y el numero de matriculas fuera del estado

```
plot(College$Outstate~as.factor(College$Private), main="Privados vs Matricula Fuera del E",
     ylab='Matricula fuera del estado')
```



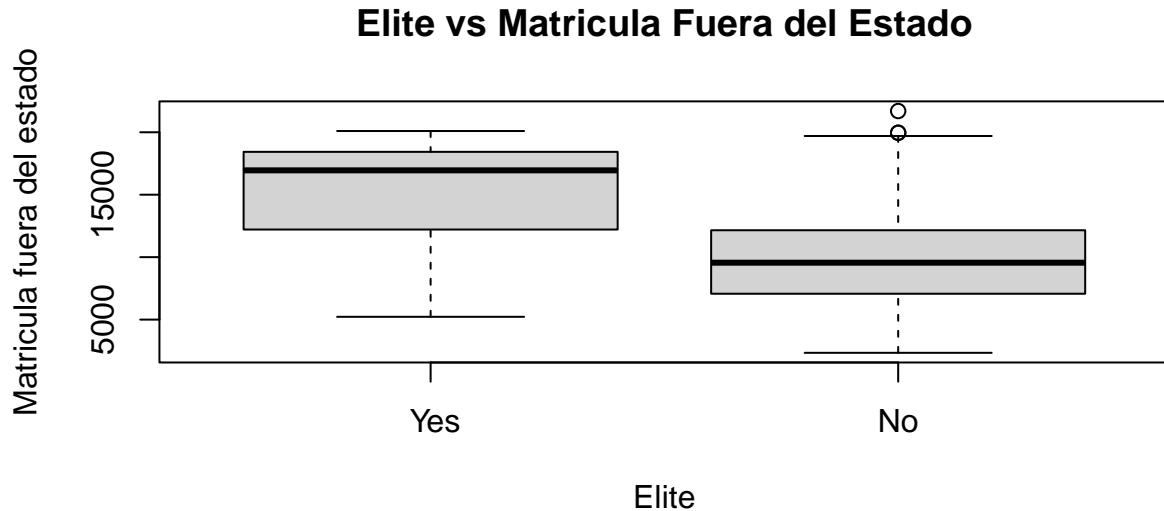
#el grafico anterior nos muestra que aparentemente no existe una diferencia significativa sobre el comportamiento #medio para el numero de matrículas fuera del estado para universidades públicas o privadas, aunque se podría pensar que # es un poco mayor para las privadas, pero el traslape de las cajas no es muy evidente.

\subsection{IV) Create a new qualitative variable, called Elite, by binning the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10 \% of their high school classes exceeds 50 \%.

#En esta sección realizamos una agrupación en una nueva variable categórica llamada Elite, donde los estudiantes con #un rendimiento superior del 10 \% en secundaria, se agrupan en si, siempre y cuando la universidad contenga 50 o mas de ellos, #en caso contrario se agrupan en no, el resultado fue el siguiente:

	Cantidad
Yes	78
No	699

```
plot(College$Outstate~as.factor(College$Elite), main="Elite vs Matricula Fuera del Estado",
      ylab='Matricula fuera del estado')
```

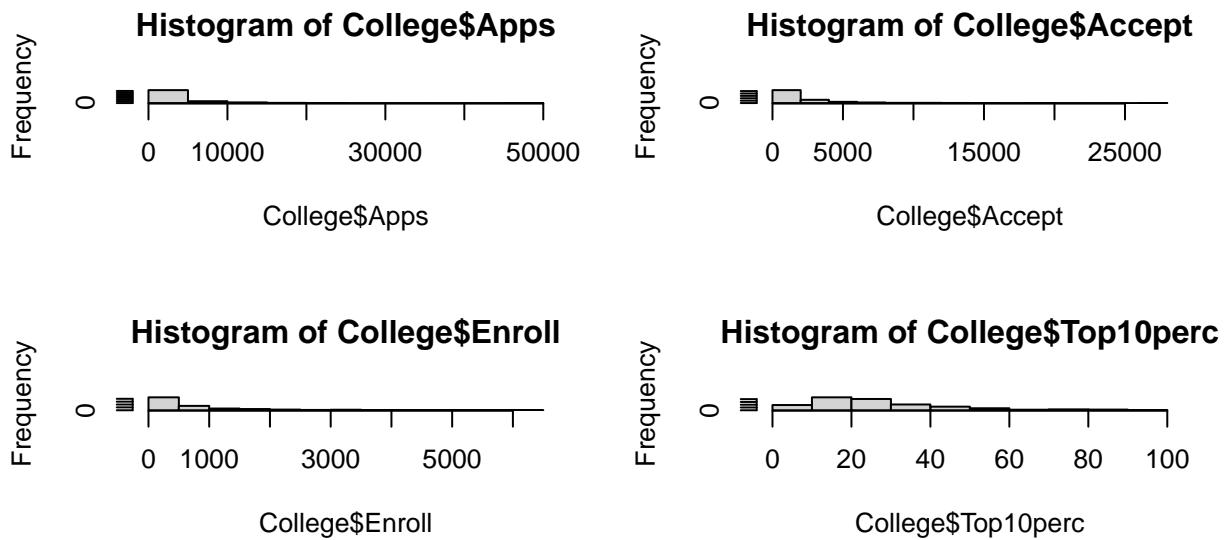


#graficamente parece haber mayor cantidad de matrículas fuera del estado para los estudiantes elite, pero esta diferencia #no es muy clara, pero da indicios muy fuertes.

- 4.6. V)** Use the `hist()` function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command `par(mfrow=c(2,2))` useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.

#los siguientes histogramas muestran la distribución de algunas de las variables, sin embargo, no es posible concluir #acerca de una posible distribución.

```
par(mfrow=c(2,2))
hist(College$Apps)
hist(College$Accept)
hist(College$Enroll)
hist(College$Top10perc)
```



4.7. VI) Continue exploring the data, and provide a brief summary of what you discover.

Continúe explorando los datos.

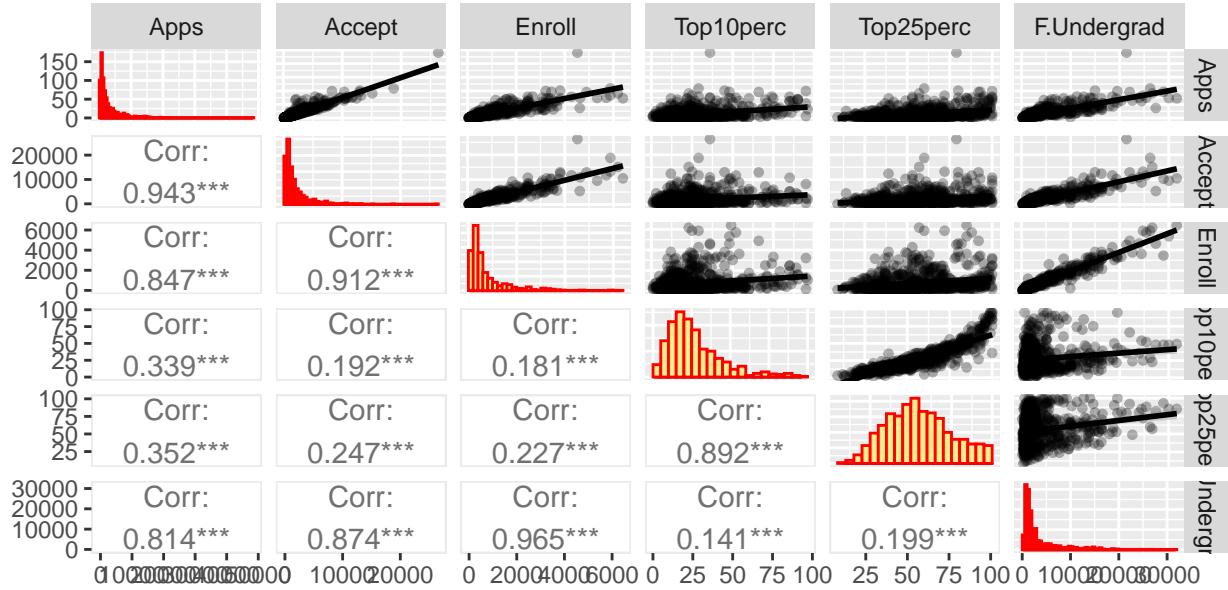
```
base<-College
base<-base[,2:7]

library(ggplot2)
library(GGally)

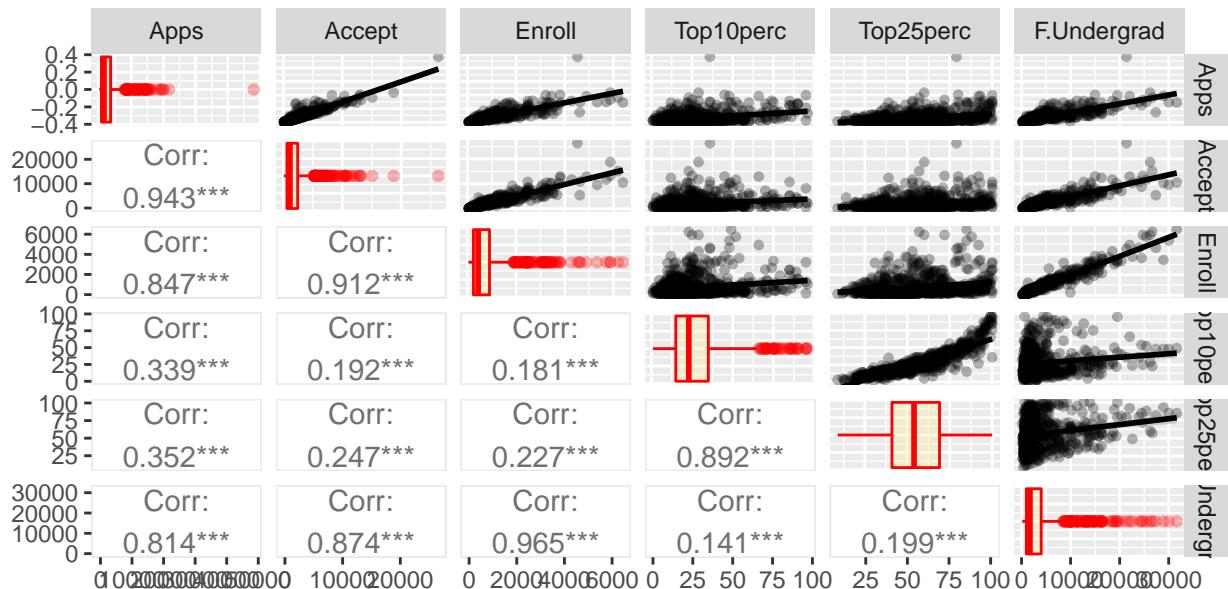
gg2<-ggpairs(base,upper=list(continuous = wrap("smooth",alpha = 0.3, size=1.2,
                                              method = "lm")),lower=list(continuous = "white"))

for(i in 1:ncol(base)){
  gg2[i,i]<-gg2[i,i] +
    geom_histogram(breaks=hist(base[,i],breaks = "FD",plot=F)$breaks,
                  colour = "red",fill="lightgoldenrod1")
}

gg2
```



```
ggpairs(base, diag=list(continuous=wrap("box_no_facet", color="red",
                                         fill="lightgoldenrod1", alpha=0.3)),
        upper=list(continuous = wrap("smooth", alpha = 0.3, size=1.2, method = "lm")),
        lower=list(continuous ="cor"))
```



#el siguiente grafico representa el comportamiento de las primeras 8 variables cuantitativas, la siguiente imagen, reúne
un conjunto de gráficos muy importantes a la hora de hacer un análisis descriptivo, por su parte también nos enseña #la relación que existe entre unas u otras y del grado de asociación numérico dictado por el coeficiente de correlación. #si existe o no una posible distribución asociada a cada variable.