Segundo Ejerccio P2P

Francisco Ramirez

15 de noviembre, 2016

Segunda tarea P2P

Del url http://world.openfoodfacts.org/data se puede descargar la tabla de datos FoodFacts.csv que contiene información sobre más de 65,000 productos alimentarios: en concreto, 156 variables, de las que 87 corresponden a contenidos nutricionales. La tabla completa ocupa más de 160 Mb (y además va aumentando casi cada día), así que para este ejercicio (en MiriadaX no nos dejan subir tablas de más de 5 Mb), a partir de esta tabla hemos construido una tabla con sólo algunos productos y las siguientes variables:

. product_name: nombre del producto

. country: país donde se adquirió (cuando un producto ha sido adquirido en varios países, aparece una fila para cada país)

. continent: continente del país

. nutrition_grade_fr: escala nutricional francesa

. additives_n: número de aditivos

. main_category: categoría

. energy: calorías por 100 g del producto

. fat: grasa (en g) por 100 g del producto

. sugars: azúcares (en g) por 100 g del producto

. fiber: fibra (en g) por 100 g del producto

. proteins: proteínas (en g) por 100 g del producto

. sodium: sal (en g) por 100 g del producto

. alcohol: alcohol (en g) por 100 g del producto

. vitamin_b6 : vitamina B6 (en g) por 100 g del producto

El resultado final es la tabla FoodFactsMooc.csv que encontraréis en el repositorio del curso (url https:// miriadax.net/documents/28098821/74010125/FoodFactsMooc.csv/c1b38463-6006-4a3b-b94a-a72d31d54831) A partir de esta tabla, vamos a estudiar algunos índices nutricionales por productos, por países y por continentes.

Pregunta - 1 (A)

Cargad la tabla de datos en un data frame “global” llamado DF_G; a continuación, cread un data frame para cada continente, y llamadlos DF_Europa, DF_Africa, DF_Asia, DF_AmericaN, DF_AmericaS, DF_Oceanía, respectivamente

En algunos apartados necesitaremos eliminar repeticiones de productos de estos data frames. Un mismo producto puede aparecer varias veces en la tabla de partida, porque haya sido adquirido en diferentes países; si varias filas corresponden exactamente al mismo producto, sólo se diferencian en el país y, si corresponde, en el continente. Así que también tenéis que construir un data frame “global y sin repeticiones” DFU_G, que no contenga las variables correspondientes al país y el continente y donde cada producto aparezca una sola vez. ¿Cuántas repeticiones había en la tabla de datos original?

Asimismo, para cada continente, cread un data frame sin repeticiones (DFU_Europa, DFU_Africa, etc., que no contenga la variable correspondiente al país (pero dejad el continente, os va a ser ser útil dentro de un rato), donde cada producto aparezca una sola vez.

Finalmente, concatenad por filas los 6 data frames anteriores de la forma DFU_Continente en un único data frame DFU_Continentes. Este data frame contendrá productos repetidos, pero nunca dentro de un mismo continente.

(Indicación: Si aplicáis la función unique a un dataframe, construís un nuevo dataframe eliminando las repeticiones de filas.)

Respuesta:

Para no extender demasiado la respuesta, al comprobrar cada data frame se han listado solo las 2 primeras filas y 5 columnas.

```
# Cargamos   la tabla de datos "FoodFactsMooc.csv"  en el data frame DF_G
DF_G <- read.csv("C:/Users/Andres/Desktop/PAPA/R/Curso Introduccio a R/FoodFactsMooc.csv")

## Comprobamos el data frame general DF_G
    DF_G[c(1:2),c(1:5)]
```

	product_name	country	continent	nutrition_grade_fr	additives_n
10	Luxury Christmas Pudding	United Kingdom	Europe	c	5
11	Luxury Christmas Pudding	United Kingdom	Europe	c	5

```
   ##  Creamos los data frame de los continentes
    DF_Europa = subset(DF_G, DF_G$continent== "Europe")
    DF_Africa = subset(DF_G, DF_G$continent== "Africa")
    DF_Asia = subset(DF_G, DF_G$continent== "Asia")
    DF_AmericaN = subset(DF_G, DF_G$continent== "North America")
    DF_AmericaS = subset(DF_G, DF_G$continent== "South America")
    DF_Oceania = subset(DF_G, DF_G$continent== "Oceania")
   ## Comprobamos los  data frame
    DF_Europa[c(1:2),c(1:5)]
```

	product_name	country	continent	nutrition_grade_fr	additives_n
10	Luxury Christmas Pudding	United Kingdom	Europe	c	5
11	Luxury Christmas Pudding	United Kingdom	Europe	c	5

```
    DF_Africa[c(1:2),c(1:5)]
```

	product_name	country	continent	nutrition_grade_fr	additives_n
1222	Solid Pack Pumpkin	United Arab Emirates	Africa	a	0
3625	Foul Medammas	Lebanon	Africa	a	1

```
DF_Asia[c(1:2),c(1:5)]
```

	product_name	country	continent	nutrition_grade_fr	additives_n
2045	Beatrix Potter Butter Cookies	China	Asia		0
2125	Lee Kum Kee Premium Gold Soy Sauce	Hong Kong	Asia	b	NA

```
DF_AmericaN[c(1:2),c(1:5)]
```

	product_name	country	continent	nutrition_grade_fr	additives_n
53	Pizza Parlanno	United States	North America	d	0
60	Mac 'n Cheese	United States	North America	c	3

```
DF_AmericaS[c(1:2),c(1:5)]
```

	product_name	country	continent	nutrition_grade_fr	additives_n
3342	Great Value Regular Potato Crisps	Brazil	South America		0
7439	Boulgour	French Guiana	South America		0

```
DF_Oceania[c(1:2),c(1:5)]
```

	product_name	country	continent	nutrition_grade_fr	additives_n
199	Honey Barbecue Sauce	Australia	Oceania	e	3
200	Sweet Baby Rays Steakhouse Marinade & Sauce	Australia	Oceania	e	1

```
##  Creamos el data frame general sin repeticiones DFU_G
##  Sin las variables pais y continente
DFU_G=unique(subset(DF_G, select = c(1, 4:14)))

##  Comprobamos el data frame
DFU_G[c(1:2),c(1:8)]
```

	product_name	nutrition_grade_fr	additives_n	main_category	energy	fat	sugars	fiber
10	Luxury Christmas Pudding	c	5	Desserts	1284	7	44	4.7
11	Luxury Christmas Pudding	c	5	Sugary snacks	1284	7	44	4.7

```
## creamos los data frame  de continentes sin repeticiones
## Sin la variable pais
DFU_Europa=unique(subset(DF_Europa, select = c(1, 3:14)))
DFU_Africa=unique(subset(DF_Africa, select = c(1, 3:14)))
DFU_Asia=unique(subset(DF_Asia, select = c(1, 3:14)))
DFU_AmericaN=unique(subset(DF_AmericaN, select = c(1, 3:14)))
DFU_AmericaS=unique(subset(DF_AmericaS, select = c(1, 3:14)))
```

```
DFU_Oceania=unique(subset(DF_Oceania, select = c(1, 3:14)))

## Comprobamos los  data frame
DFU_Europa[c(1:2),c(1:5)]
```

	product_name	continent	nutrition_grade_fr	additives_n	main_category
10	Luxury Christmas Pudding	Europe	c	5	Desserts
11	Luxury Christmas Pudding	Europe	c	5	Sugary snacks

```
DFU_Africa[c(1:2),c(1:5)]
```

	product_name	continent	nutrition_grade_fr	additives_n	main_category
1222	Solid Pack Pumpkin	Africa	a	0	Aliments-d-origine-vegetale
3625	Foul Medammas	Africa	a	1	Plant-based foods and beverages

```
DFU_Asia[c(1:2),c(1:5)]
```

	product_name	continent	nutrition_grade_fr	additives_n	main_category
2045	Beatrix Potter Butter Cookies	Asia		0	Sugary snacks
2125	Lee Kum Kee Premium Gold Soy Sauce	Asia	b	NA	

```
DFU_AmericaN[c(1:2),c(1:5)]
```

	product_name	continent	nutrition_grade_fr	additives_n	main_category
53	Pizza Parlanno	North America	d	0	Meals
60	Mac 'n Cheese	North America	c	3	Meals

```
DFU_AmericaS[c(1:2),c(1:5)]
```

	product_name	continent	nutrition_grade_fr	additives_n	main_category
3342	Great Value Regular Potato Crisps	South America		0	pt:Salgadinho-de-batata
7439	Boulgour	South America		0	Plant-based foods and b

```
DFU_Oceania[c(1:2),c(1:5)]
```

	product_name	continent	nutrition_grade_fr	additives_n	main_category
199	Honey Barbecue Sauce	Oceania	e	3	Groceries
200	Sweet Baby Rays Steakhouse Marinade & Sauce	Oceania	e	1	Barbeque-sauce

```
    ##Concatenamos las 6 continentes
    DFU_Continentes = rbind(DFU_Europa,DFU_Africa,DFU_Asia,
                  DFU_AmericaN,DFU_AmericaS,DFU_Oceania )

    ## Comprobamos el data frame

    DFU_Continentes[c(1:2),c(1:5)]
```

	product_name	continent	nutrition_grade_fr	additives_n	main_category
10	Luxury Christmas Pudding	Europe	c	5	Desserts
11	Luxury Christmas Pudding	Europe	c	5	Sugary snacks

El número de repeticiones es la diferencia de filas entre el data frame DF_G y D_FUG, que en este caso nos da:

```
## Calculamos el numero de repeticiones en la tabla general
    dim(DF_G)[1]-dim(DFU_G)[1]  ## en este caso es de 1529
```

```
## [1] 1529
```

Pregunta - (B) En la rúbrica no se menciona esta pregunta

¿Qué porcentajes de valores NA hay en las diferentes variables del data frame DFU_G? Sería conveniente que presentaseis esta información en forma de tabla. ¿Qué variables presentan un mayor porcentaje de valores NA? ¿Tenéis alguna explicación para ello?

Respuesta:

Primero con la funcion summary, aplicada a las variables de los valores numéricos (columnas 3 y de la 5 a la 12) vemos en la fila 7 el número de NA's por variable.

```
    Res_NA= summary(DFU_G[,c(3,5:12)])
    Res_NA
```

additives_n	energy	fat	sugars	fiber	proteins	sodium
Min. : 0.000	Min. : 0	Min. : 0.00	Min. : -0.50	Min. : 0.000	Min. : 0.000	Min. : 0.000
1st Qu.: 0.000	1st Qu.: 367	1st Qu.: 1.00	1st Qu.: 1.10	1st Qu.: 0.500	1st Qu.: 1.700	1st Qu.: 0.031
Median : 1.000	Median : 974	Median : 5.80	Median : 4.10	Median : 1.800	Median : 5.800	Median : 0.201
Mean : 1.888	Mean :1071	Mean : 12.93	Mean : 12.79	Mean : 2.816	Mean : 7.432	Mean : 0.438
3rd Qu.: 3.000	3rd Qu.:1623	3rd Qu.: 20.80	3rd Qu.: 15.80	3rd Qu.: 3.570	3rd Qu.:10.100	3rd Qu.: 0.472
Max. :21.000	Max. :4134	Max. :101.00	Max. :105.00	Max. :94.800	Max. :86.000	Max. :42.000
NA's :2614	NA's :2119	NA's :2128	NA's :5746	NA's :15487	NA's :2546	NA's :5493

Extraemos esta fila 7 y creamos una tabla con los valores númericos de NA's por variable.

```r
#creamos un vector,con solo los numeros  NA por variable
Res_NA = Res_NA[7,]
Num_NA=as.numeric(gsub("\\D", "",Res_NA))

#Creamos un vector con los  nombres de las variables
Variable=names(Res_NA)

#Creamos un data frame con estos dos vectores
DF_V_Res = data.frame(Variable,Num_NA )

## Extraemos el numero de filas de DFU_G  en este caso es 37415
m = dim(DFU_G)[1]

## Añadimos una columna con el % de NA's por variable
DF_V_Res$Porcentage_NA = round((DF_V_Res$Num_NA/m)*100, 2)
DF_V_Res
```

Variable	Num_NA	Porcentage_NA
additives_n	2614	6.99
energy	2119	5.66
fat	2128	5.69
sugars	5746	15.36
fiber	15487	41.39
proteins	2546	6.80
sodium	5493	14.68
alcohol	35178	94.02
vitamin_b6	36736	98.19

Vemos que alcohol y vitaminas B6 son las que tiene un mayor porcentage de NA's con mucha diferencia con el resto,
el alcohol 35178 (94.02%) y vitamin_B6 36736 (98.19%), de un total de 37415 productos.

La explicación a este resultado es que el resto de varibles corresponden a elementos comunes a todos o la mayoria de alimentos y sin embargo el alcohol y suplemento de vitamina_b6 son exclusivos de este tipo de productos.

Pregunta - 2 (C)

Calculad las correlaciones entre las variables numéricas del dataframe DFU_G (sus últimas 8 variables). Escoged la opción use que produzca menos NA. Opcionalmente, representad los valores absolutos de estas correlaciones mediante un diagrama de calor. A continuación:

1. Si lo habéis hecho bien, aparecerá una sola pareja de variables con correlación NA. ¿A qué se debe este valor?
2. Determinad el par de variables numéricas diferentes con mayor correlación en valor absoluto, y dibujad su diagrama de dispersión incluyendo su recta de regresión.

Respuesta:

Calculamos las correlaciones de las últimas 8 vatiables y se comprueba que la opcion "pairwise.complete.obs" es la que menos NA's produce

```r
# Creamos un data frame  con solos estas variable
DFU_G_Var  = DFU_G[,5:12]

#Calculamos las correlaciones con las dos opciones
##use para comprobrar con que opcion hay menos NA's
    round(cor(DFU_G_Var, use = "pairwise.complete.obs"),2)
```

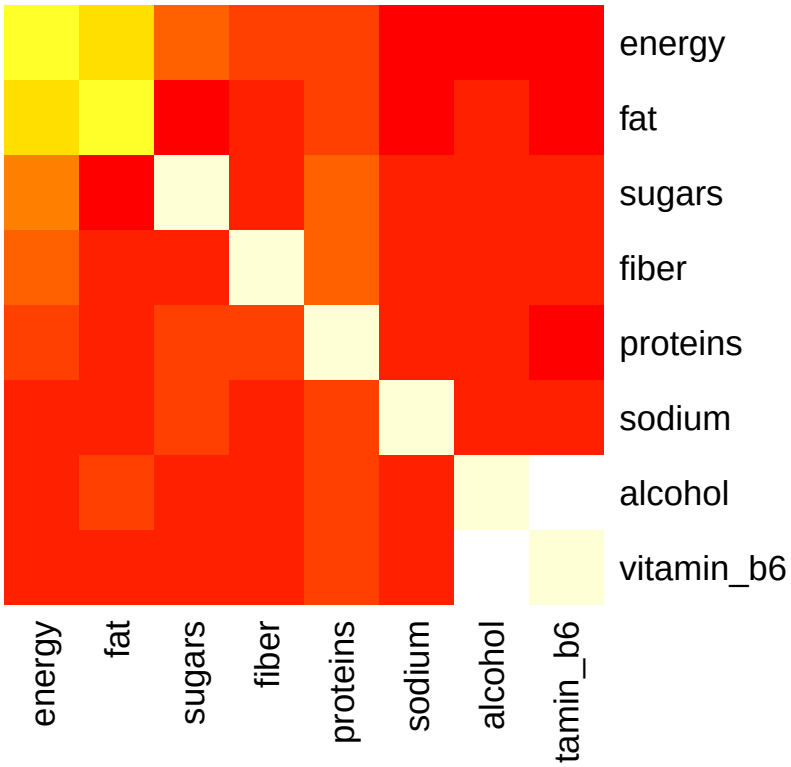
	energy	fat	sugars	fiber	proteins	sodium	alcohol	vitamin_b6
energy	1.00	0.79	0.35	0.28	0.23	-0.02	-0.05	-0.01
fat	0.79	1.00	0.02	0.09	0.15	0.01	-0.09	-0.03
sugars	0.35	0.02	1.00	0.06	-0.23	-0.09	-0.03	-0.05
fiber	0.28	0.09	0.06	1.00	0.23	-0.02	-0.07	-0.03
proteins	0.23	0.15	-0.23	0.23	1.00	0.08	-0.12	0.06
sodium	-0.02	0.01	-0.09	-0.02	0.08	1.00	0.03	-0.02
alcohol	-0.05	-0.09	-0.03	-0.07	-0.12	0.03	1.00	NA
vitamin_b6	-0.01	-0.03	-0.05	-0.03	0.06	-0.02	NA	1.00

```r
    round(cor(DFU_G_Var, use = "complete.obs"),2)
```

	energy	fat	sugars	fiber	proteins	sodium	alcohol	vitamin_b6
energy	1.00	0.38	0.58	0.87	0.73	0.77	NA	0.26
fat	0.38	1.00	-0.16	0.41	0.15	0.36	NA	-0.05
sugars	0.58	-0.16	1.00	0.29	0.42	0.32	NA	0.35
fiber	0.87	0.41	0.29	1.00	0.70	0.57	NA	0.01
proteins	0.73	0.15	0.42	0.70	1.00	0.57	NA	0.13
sodium	0.77	0.36	0.32	0.57	0.57	1.00	NA	0.32
alcohol	NA	NA	NA	NA	NA	NA	1	NA
vitamin_b6	0.26	-0.05	0.35	0.01	0.13	0.32	NA	1.00

Creamos el gráfico de color.

```r
   heatmap(abs(cor(DFU_G_Var, use = "pairwise.complete.obs")), Rowv = NA,
           Colv= NA, revC = TRUE)
```



La pareja que tiene variables de correlación NA's es "alcohol" con "vitamin_b6", esto debido a que su covarianza Sij es NA ya que los productos que tienen la variable vitamina_b6 tienen NA o 0 en alcohol.

Buscamos el valor más alto de correlaciones a parte de 1.

```
round(sort(as.numeric(cor(DFU_G_Var, use = "pairwise.complete.obs"))),2)
```

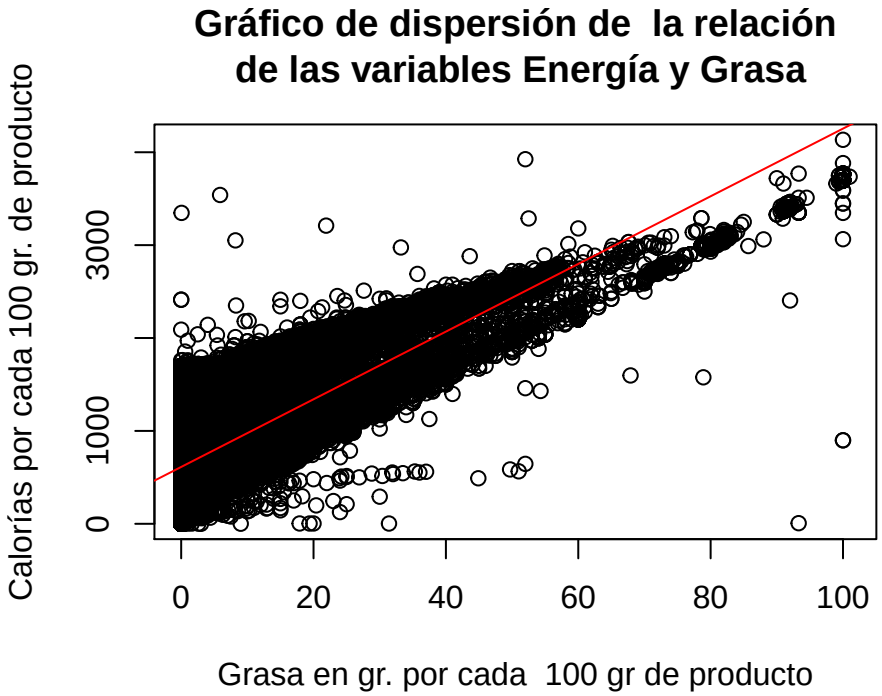
```
## Warning in cor(DFU_G_Var, use = "pairwise.complete.obs"): the standard
## deviation is zero
```

```
##  [1] -0.23 -0.23 -0.12 -0.12 -0.09 -0.09 -0.09 -0.09 -0.07 -0.07 -0.05
## [12] -0.05 -0.05 -0.05 -0.03 -0.03 -0.03 -0.03 -0.03 -0.03 -0.02 -0.02
## [23] -0.02 -0.02 -0.02 -0.02 -0.01 -0.01  0.01  0.01  0.02  0.02  0.03
## [34]  0.03  0.06  0.06  0.06  0.06  0.08  0.08  0.09  0.09  0.15  0.15
## [45]  0.23  0.23  0.23  0.23  0.28  0.28  0.35  0.35  0.79  0.79  1.00
## [56]  1.00  1.00  1.00  1.00  1.00  1.00  1.00
```

Tenemos que es 0.79 y que corresponde a la pareja de variables energy–fat según tabla anterior.

Dibujamos el diagrama de dispersion y le añadimos la recta de regresión

```
plot(DFU_G_Var$fat,DFU_G_Var$energy,
 main="Gráfico de dispersión de  la relación\n de las variables Energía y Grasa",
 xlab="Grasa en gr. por cada  100 gr de producto",
 ylab= "Calorías por cada 100 gr. de producto")
abline (lm(energy~fat,DFU_G_Var), col="red")
```



Pregunta - 3 (D)

El Actimel de Danone "ayuda al funcionamiento normal del sistema inmunitario" porque contiene 0.21 mg de vitamina B6 por cada 100 g. ¿Qué porcentaje de los productos de los que en la tabla DFU_G se indica su cantidad de vitamina B6, tienen como mínimo tanta vitamina B6 por cada 100 g como el Actimel? ¿Qué 5 productos de esta tabla tienen la mayor cantidad de vitamina B6 por cada 100 g?

Respuesta:

Primero extraemos de DFU_G los productos que tiene vitamina B6

```
DFU_V_B6  =droplevels(subset(DFU_G, vitamin_b6 >= 0,  select=(c(1,12))))
   DFU_V_B6$product_name = as.character(DFU_V_B6$product_name)
   head(DFU_V_B6)
```

	product_name	vitamin_b6
163	Strawberry Lemonade	0.000139
218	Bluberry Protein Shake	0.000800
219	mango protein flavored soy protein shake	0.000789
306	Cheerios	0.001790
307	Cheerios	0.001790
308	Raisin Nut Bran	0.001020

Ordenamos la tabla en orden decreciente por la variable vitamina. Primero sacamos un vector con los valores de las vitaminas ordenados de mayor a menor

```
Vi_ord = sort(DFU_V_B6$vitamin_b6,  decreasing = TRUE)
    head(Vi_ord)
```

```
## [1] 0.21500 0.20000 0.02080 0.01400 0.00690 0.00645
```

Calculamos las frecuencias relativa absolutas y acumuladas de este vector.

```
    Fre_rel = prop.table(table(Vi_ord))
    Fre_rel_acu = cumsum(prop.table(table(Vi_ord)))
```

Sacamos la Frecuencia relativa acumulada hasta los 0,00021 gr, pero lo que buscamos es desde 0,00021 hasta el máximo. El calculo será restando de 1 la frecuencia relativa acumulada hasta los 0,00021 y sumandole la frecuencia relativa absoluta de 0,00021, pasandolo despues a %, en este caso será:

```
 round((1-Fre_rel_acu["0.00021"]+ Fre_rel["0.00021"]) *100,2)
```

```
## 0.00021
##   90.43
```

Para sacar los 5 productos con mayor cantidad de vitamina B6, ordenamos la tabla por orden decreciente de la variable "vitamin_b6" y visualizamos los 5 primeros

```
Sort.DFU_V_B6=DFU_V_B6[order(-DFU_V_B6$vitamin_b6),]
Sort.DFU_V_B6[1:5,]
```

	product_name	vitamin_b6
26883	Levure de biÃ¨re VitalitÃ© BeautÃ©	0.2150
55658	Zanahorias "VegaTajo"	0.2000
3299	Pop Tarts Hot Fudge Sundae	0.0208
47915	bcaa caps	0.0140
24641	Levure de biÃ¨re en paillettes	0.0069

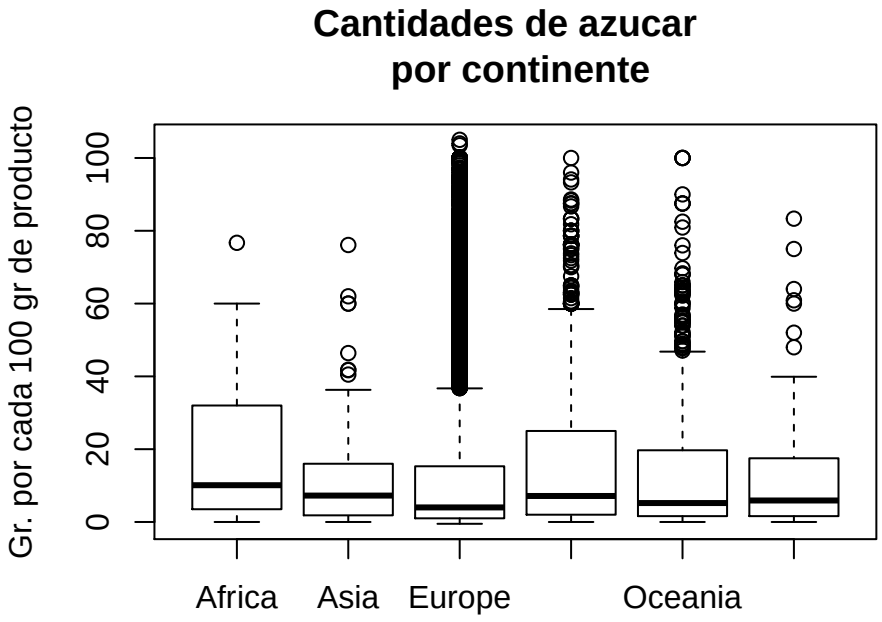
Pregunta - 4 (E)

Producid un gráfico que muestre, para cada continente, los diagramas de caja de las cantidades de azúcar en 100g en los productos adquiridos en el continente; y lo mismo para las cantidades de sal, de grasa y el número de aditivos. Usad el dataframe sin repeticiones dentro de continentes DFU_Continentes. ¿Se observan diferencias entre los consumos en los continentes? Comentadlas.

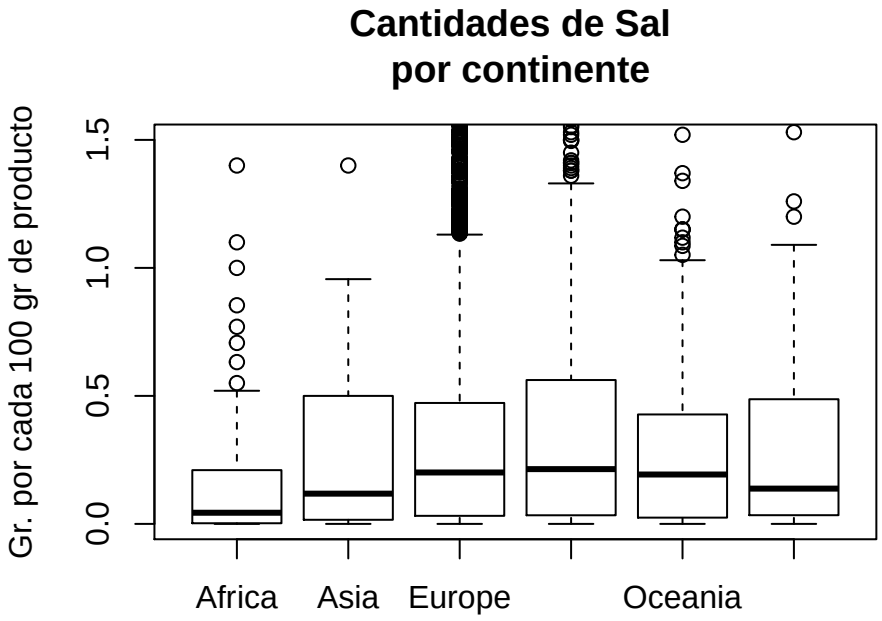
Respuesta:

Creamos los 4 diagramas de cajas, en el caso de la sal se ha cambiado los limites del eje de las ordenadas (y) para poder ver el diagrama correctamente.

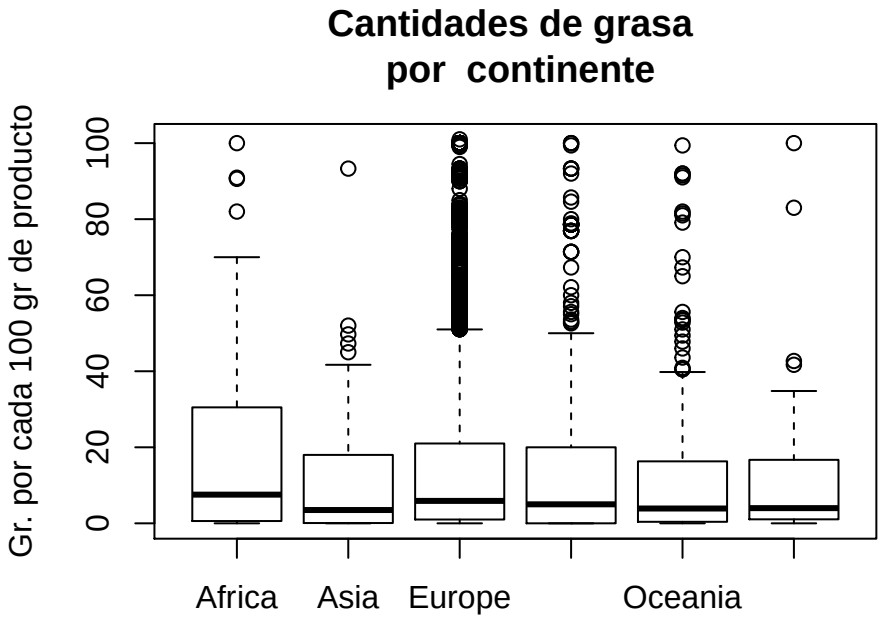
```
boxplot(sugars~continent, data = DFU_Continentes,
 main= "Cantidades de azucar  \n por continente",
 ylab="Gr. por cada 100 gr de producto")
```



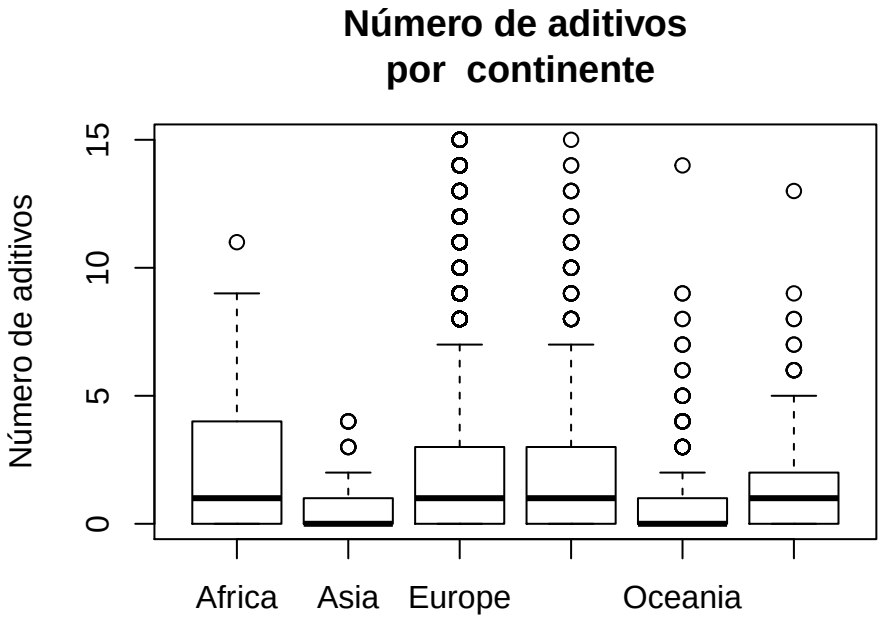

```
boxplot(sodium~continent, data = DFU_Continentes,
 main= "Cantidades de Sal \n por continente",
 ylab="Gr. por cada 100 gr de producto",ylim=c(0,1.5))
```



```r
boxplot(fat~continent, data = DFU_Continentes,
 main= "Cantidades de grasa \n por  continente",
 ylab="Gr. por cada 100 gr de producto")
```




```r
boxplot(additives_n~continent, data = DFU_Continentes,
 main =" Número de aditivos \n por  continente",
 ylab="Número de aditivos",ylim=c(0,15))
```



Se observa que las medianas son muy limilares y que la dispersión es mayor en los continentes mas desarrollados que deben de ser tipos en el procesado de los alimentos de estos continentes.

Pregunta - 5 (F)

Vamos a fijarnos ahora en las bebidas alcohólicas (aquellas que tengan valor "Beverage" en la variable main_category y contenido de alcohol mayor que 0). Usando los dataframes sin repeticiones DFU_G y DFU_Continentes, dibujad histogramas del contenido de alcohol en estas bebidas: uno global, y uno para cada continente. Procurad que los grupos sean los mismos en cada histograma (aunque algunos queden vacíos), para poder compararlos mejor. Poned nombres adecuados a los histogramas. ¿Se observan diferencias entre los continentes? ¿Qué continente presenta una distribución del contenido de alcohol en sus bebidas alcohólicas más parecido al global? ¿Se os ocurre por qué?

Respuesta:

Primero extraemos el valor de la variable alcohol de los registros que contienen alcohol en las tablas DFU_G y DFU_Continentes y los convierto a vectores.

```
#Extraemos los datos
Alc_G = subset(DFU_G,  main_category == "Beverages" &  alcohol > 0, select = alcohol  )
Alc_Europa =subset(DFU_Continentes,  main_category== "Beverages" &  alcohol> 0 &
                   continent == "Europe", select = alcohol)
Alc_Africa = subset(DFU_Continentes,  main_category== "Beverages" &  alcohol> 0 &
                        continent == "Africa", select = alcohol)
Alc_Asia = subset(DFU_Continentes,  main_category== "Beverages" &  alcohol> 0 &
                        continent == "Asia", select = alcohol)
Alc_AmericaN = subset(DFU_Continentes,  main_category== "Beverages" &  alcohol> 0 &
```

```
                        continent == "North America", select = alcohol)
Alc_AmericaS = subset(DFU_Continentes,  main_category== "Beverages" &  alcohol> 0 &
                        continent == "South America", select = alcohol)
Alc_Oceania = subset(DFU_Continentes,  main_category== "Beverages" &  alcohol> 0 &
                       continent == "Oceania", select = alcohol)

## Los convierto a vectores numéricos
Alc_G = as.numeric(Alc_G$alcohol)
Alc_Europa = as.numeric(Alc_Europa$alcohol)
Alc_Africa = as.numeric(Alc_Africa$alcohol)
Alc_Asia = as.numeric(Alc_Asia$alcohol)
Alc_AmericaN = as.numeric(Alc_AmericaN$alcohol)
Alc_AmericaS = as.numeric(Alc_AmericaS$alcohol)
Alc_Oceania = as.numeric(Alc_Oceania$alcohol)
```

Calculamos el número de clases con la regla de Sturges.

```
nclass.Sturges(Alc_G)   ##  Nos da un valor de   k de 12
```

```
## [1] 12
```

Calculamos los rangos con un número de 12 clases teniendo en cuenta que la precisión es de una decima

```
    diff(range(Alc_G))/12
```

```
## [1] 5.741667
```

Redondeamos a 5.8 y calculamos los valores del vector L de las clases, teniendo en cuenta que la precisión es de una decima.

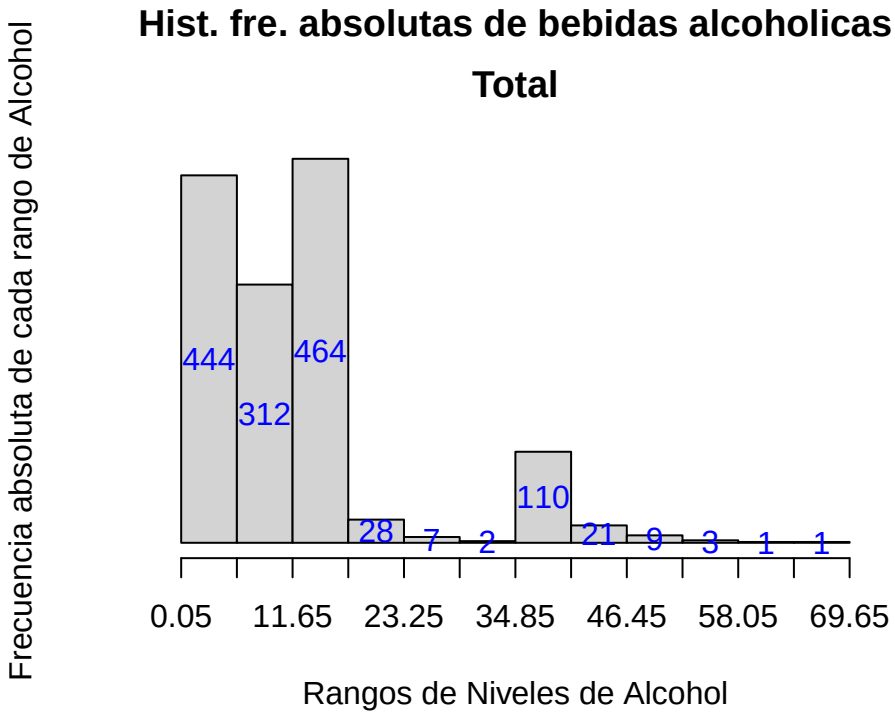
```
    L= (min(Alc_G)-(0.1/2)) + 5.8*(0:12)   ## nos  da un valor de 0.05
```

Hacemos los histogramas Utilizando la función de la lección 11 para hacer los histogramas

```
     hist_abs = function (x,L, nombre= " "){
    Titulo = c("Hist. fre. absolutas de bebidas alcoholicas\n", nombre)
      h= hist(x, breaks =L, right =FALSE , freq =FALSE ,
              axes =FALSE , col =" lightgray ",
              main =Titulo ,
              xlab ="Rangos de Niveles de Alcohol",
              ylab ="Frecuencia absoluta de cada rango de Alcohol")
      axis (1, at=L)
      text (h$mids , h$density /2, labels =h$counts , col="blue")
    }
```

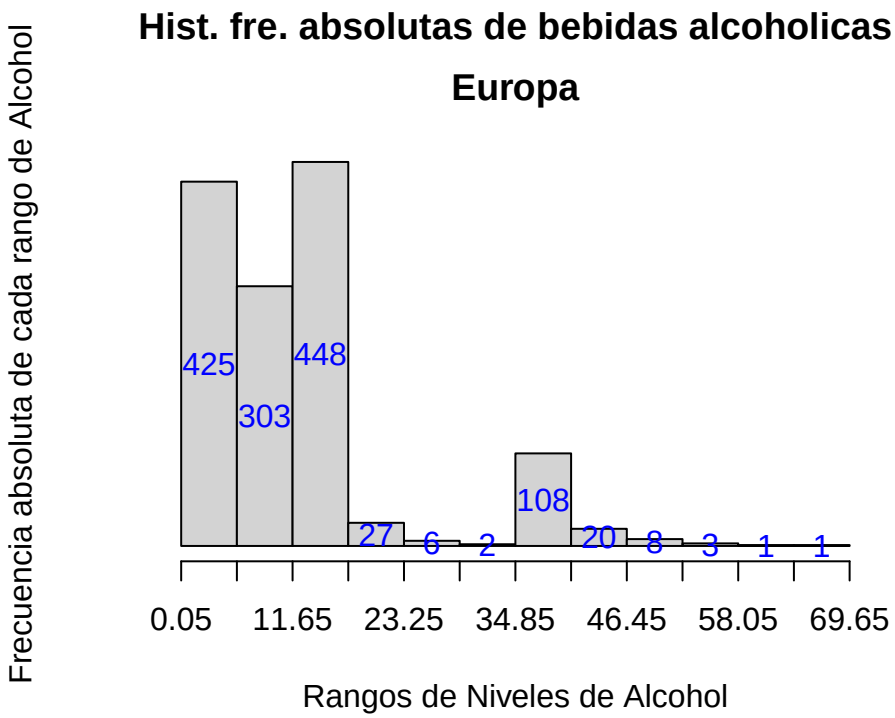
Realizamos el histograma general.

```
    hist_abs(Alc_G, L, "Total")
```

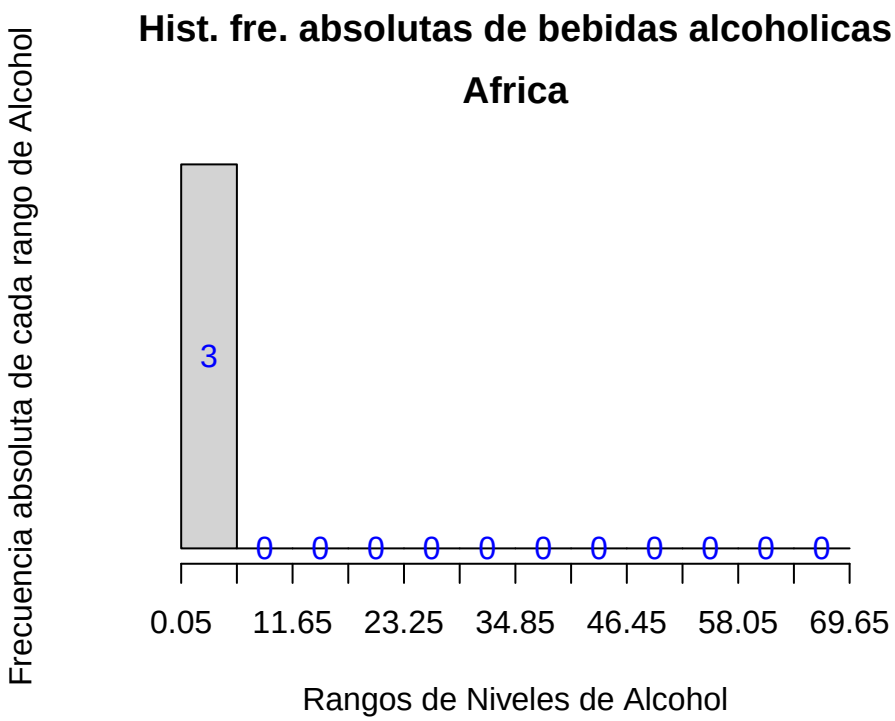


Ceamos los histogramas de los continentes con la misma función

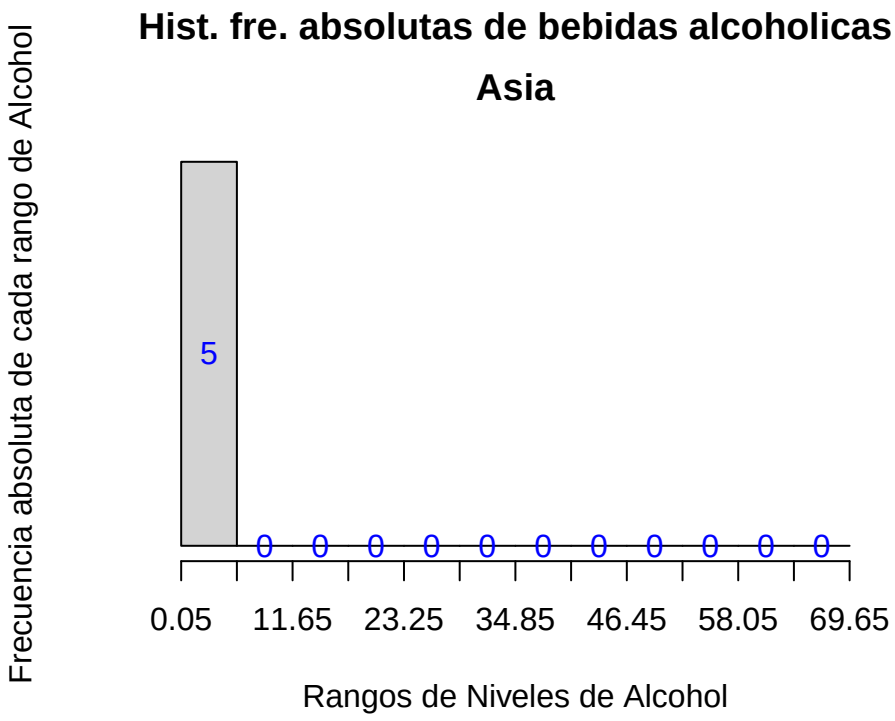
```
    hist_abs(Alc_Europa, L, "Europa")
```



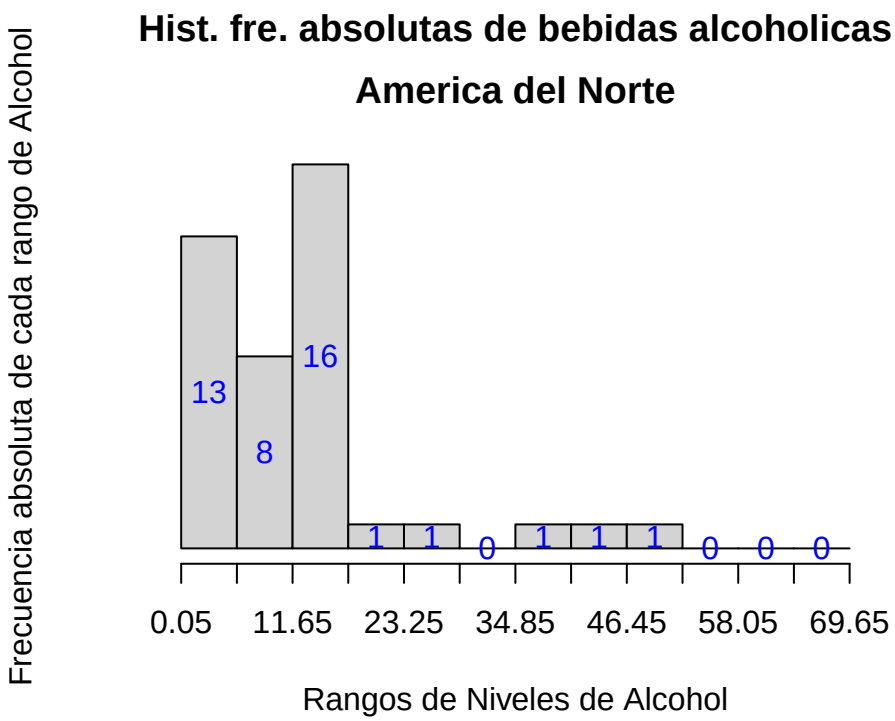
```
hist_abs(Alc_Africa, L, "Africa")
```



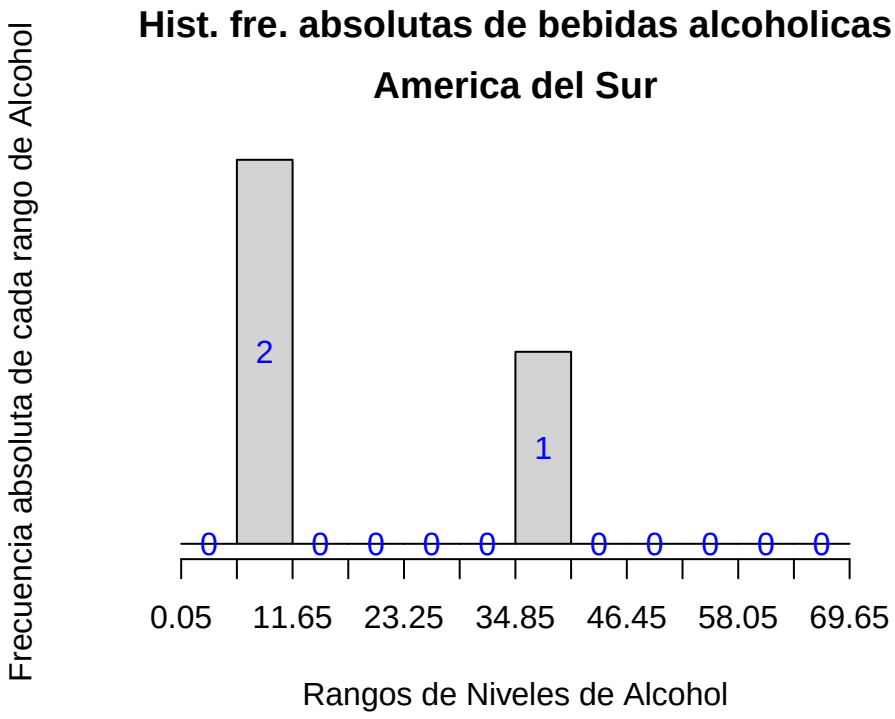

```
hist_abs(Alc_Asia, L, "Asia")
```



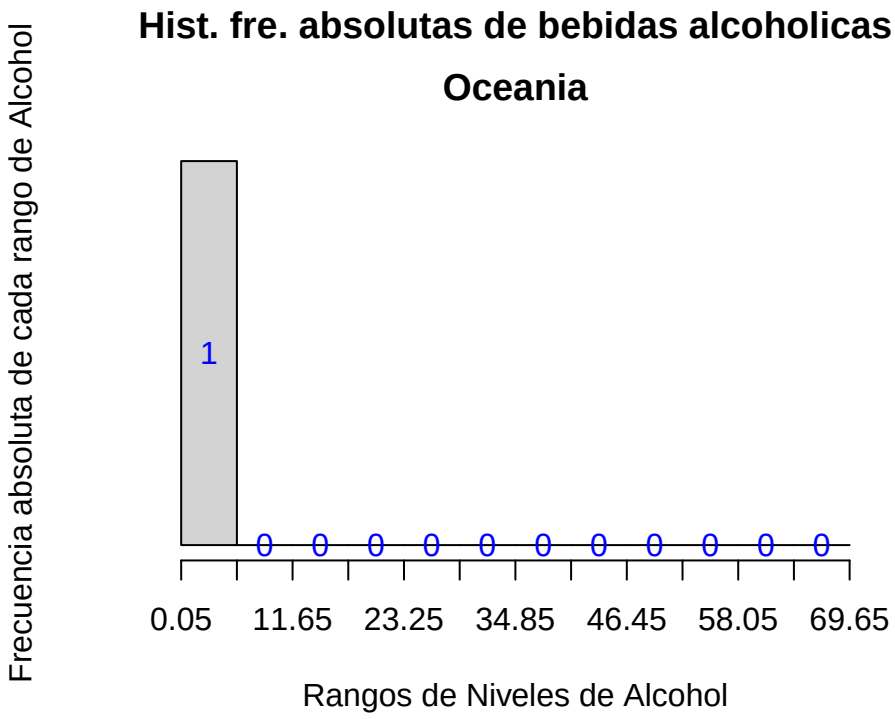
```
hist_abs(Alc_AmericaN, L, "America del Norte")
```




```
hist_abs(Alc_AmericaS, L, "America del Sur")
```



```
hist_abs(Alc_Oceania, L, "Oceania")
```



Vemos que el número de muestras son casi todas de Europa y de Norte América, para hacer una comparación mejor de este dato creamos una tabla con los porcentage de muestras de alcohol por continentes con respecto al total de las muestras que tienen alcohol.

```
Nom_Continentes = c("Europa","Africa","Asia","Norte America", "Sudamerica", "Oceania")
Valor_Continentes =round(c(length(Alc_Europa),
                           length(Alc_Africa),
                           length(Alc_Asia),
                           length(Alc_AmericaN),
                           length(Alc_AmericaS),
                           length(Alc_Oceania))/length(Alc_G)*100,2)

  data.frame(Nom_Continentes, Valor_Continentes)
```

Nom_Continentes	Valor_Continentes
Europa	96.43
Africa	0.21
Asia	0.36
Norte America	3.00
Sudamerica	0.21
Oceania	0.07

Vemos que la mayoría de muestras son de Europa con 96.43% luego América del Norte con solo un 3% y el resto con un porcentage mínimo, por lo que los gráficos de estos últimos no son representativos por tener un

muestra tan pequeña. Solo podemos comparar ambos continentes, Europa con América del Norte y aunque la de América del Norte es también una muestra pequeña se aprecia una distribución similar a la de Europa.

Sin embargpo si comparamos Europa con el total la similitud es prácticamente idéntica, como es lógico ya que el 96.43% del total son muestras de Europa.

Pregunta - 6 (G)

La variable nutrition_grade_fr es un factor con niveles a,b,c,d,e. Hay otro nivel, vacío, para los productos sin categoría (correspondería al NA). Usando el data frame DFU_Continente, dibujad un diagrama de barras que muestre, para cada continente, el porcentaje de productos en cada categoría. No incluyáis los productos sin categoría asignada.

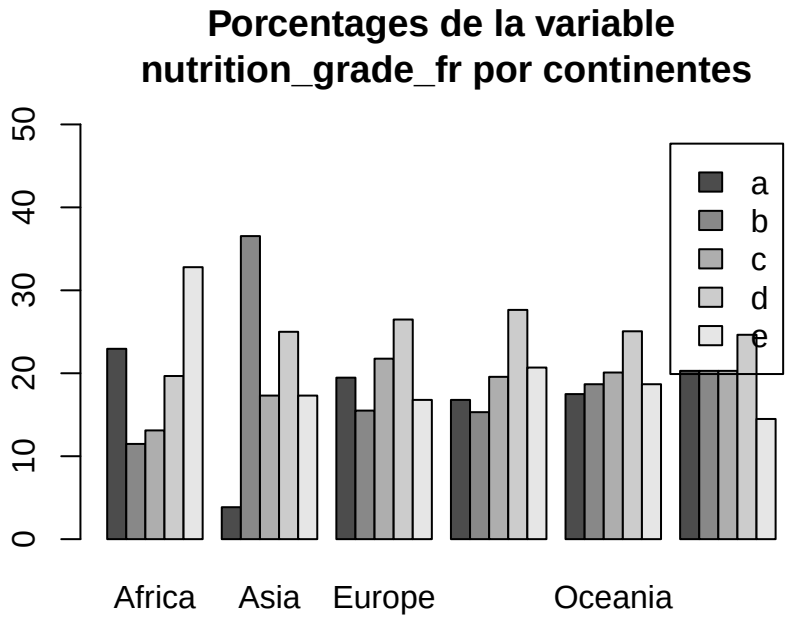
Respuesta:

Primero eliminanos las filas con los datos vacios de la variable nutritian_grade_fr del data frame DFU_Continentes, creamos una tabla con los porcentages en cada categoría por continentes y dibujamos el gráfico.

```
## par( ps=10, fin =c(7,5))
    ##Eliminamos las filas vacias
   DF_Tabl_nut =droplevels(subset(DFU_Continentes, nutrition_grade_fr != "",
                select = c(2,3)))

#Creamos la tabla con los porcentages por categoria
Tabl_nutricion = round(prop.table(table(DF_Tabl_nut$nutrition_grade_fr,
            DF_Tabl_nut$continent),margin = 2)*100,2)

    ##Dibujamos el diagrama de barras
    barplot(Tabl_nutricion, beside = TRUE, legend.text = TRUE,
            main = "Porcentages de la variable\n nutrition_grade_fr por continentes",
            ylim=c(0,50))
```



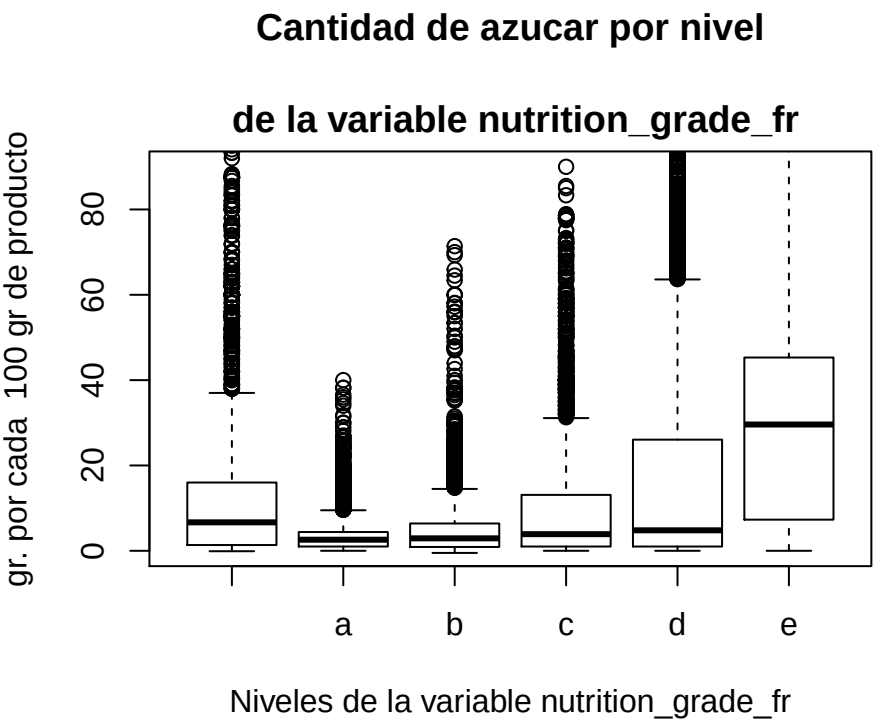
Pregunta - 7 (H)

A partir del dataframe DFU_G, dibujad un gráfico que contenga los diagramas de caja de los contenidos de azúcar por 100 g de los alimentos de cada nivel de la variable nutrition_grade_fr. Repetid este gráfico para los contenidos de sal, grasa, fibra y proteínas. A partir de estos gráficos, ¿podéis interpretar cómo clasifica esta variable los alimentos? Si queréis, podéis confirmar si vuestra interpretación va en la dirección correcta consultando el informe oficial (en francés) sobre la clasificación nutricional francesa.

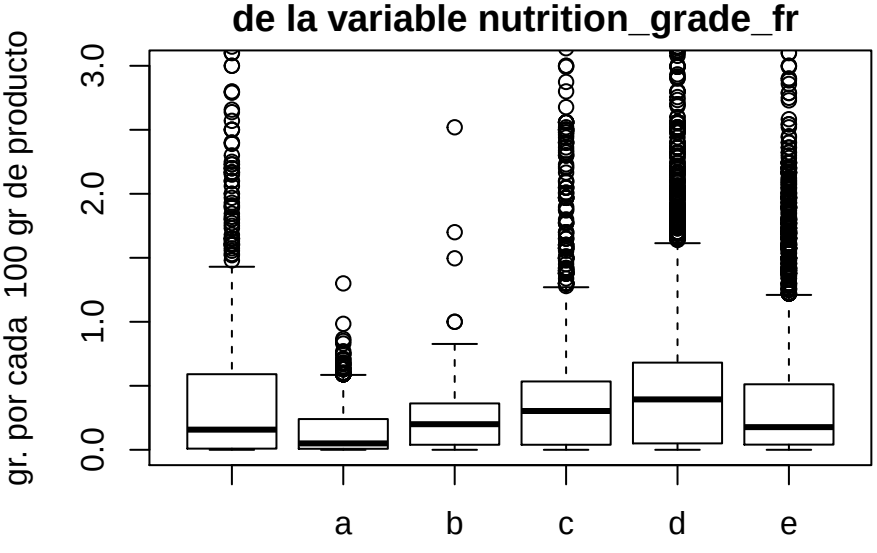
Respuesta:

Dibujamos los diagramas:

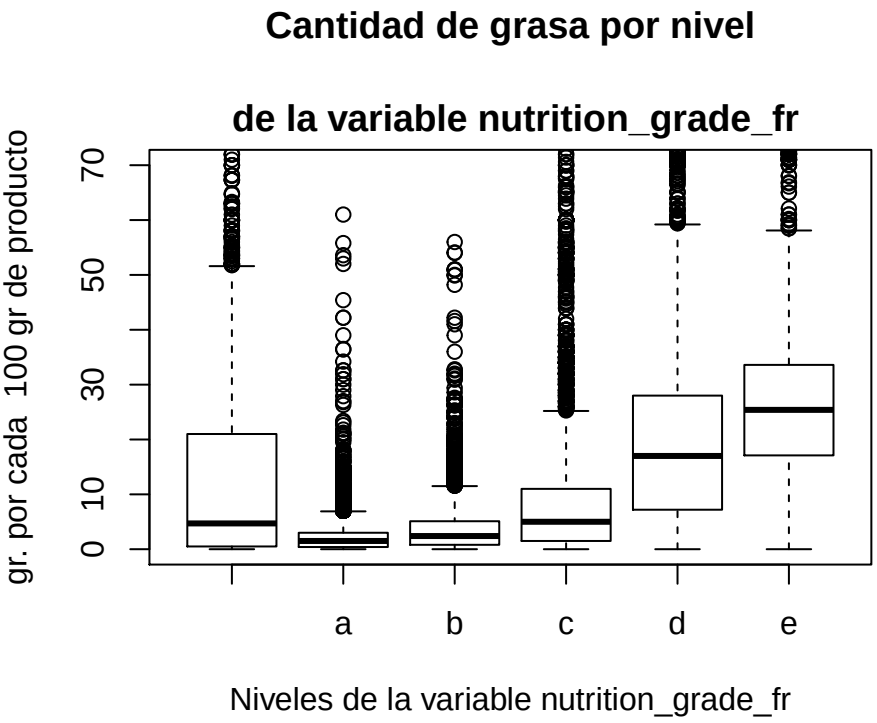
```
boxplot(sugars~nutrition_grade_fr, data = DFU_G, main= "Cantidad de azucar por nivel
        \n de la variable nutrition_grade_fr", ylim=c(0,90),
        xlab="Niveles de la variable nutrition_grade_fr",
        ylab= "gr. por cada  100 gr de producto")
```



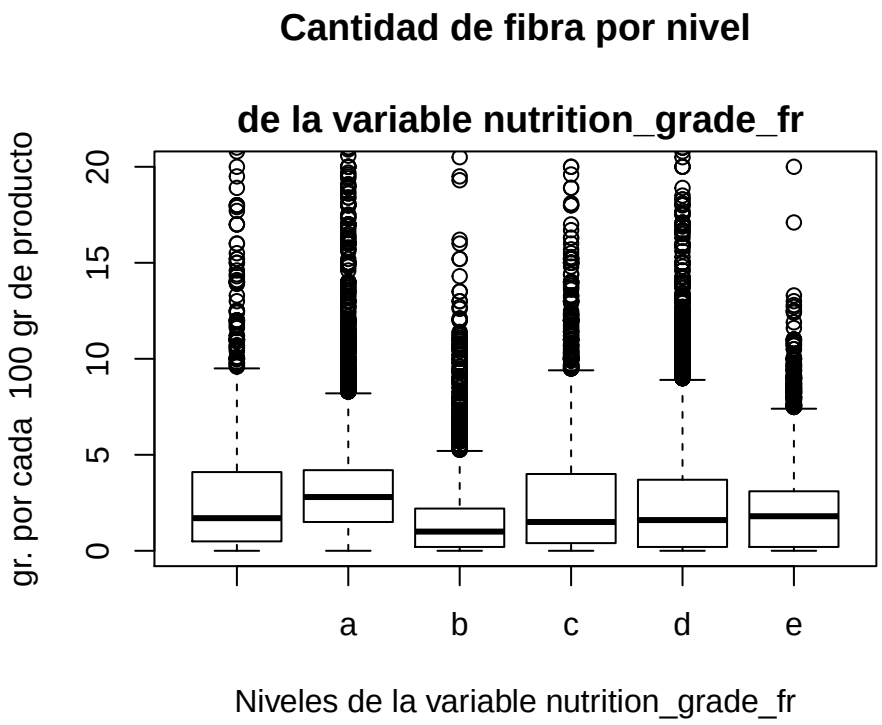

```
boxplot(sodium~nutrition_grade_fr, data = DFU_G, main= "Cantidad de sal por nivel
        \n de la variable nutrition_grade_fr", ylim=c(0,3),
        xlab="Niveles de la variable nutrition_grade_fr",
        ylab= "gr. por cada  100 gr de producto")
```



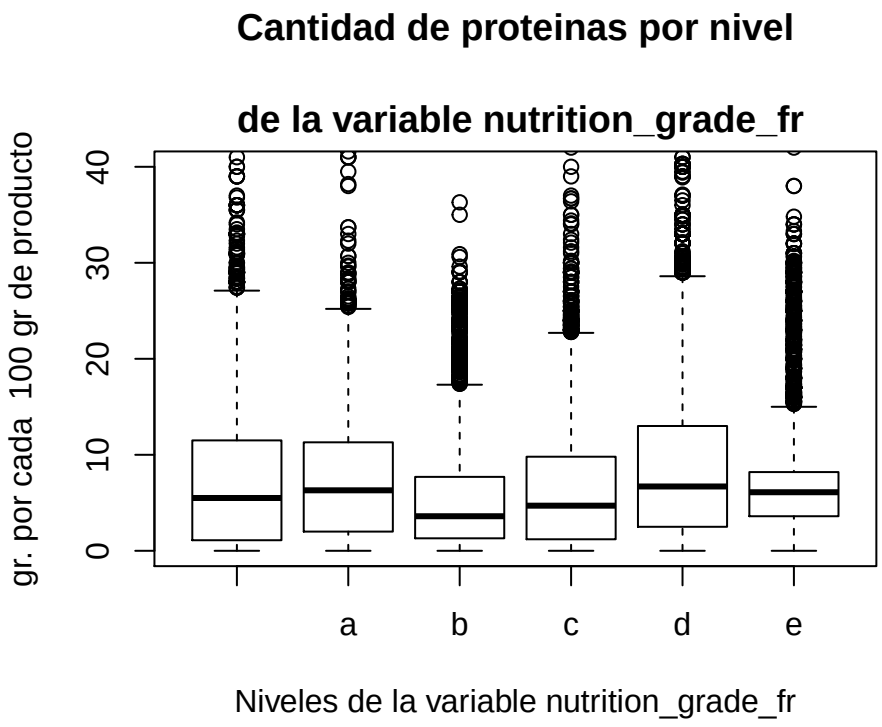
```
boxplot(fat~nutrition_grade_fr, data = DFU_G, main= "Cantidad de grasa por nivel
        \n de la variable nutrition_grade_fr", ylim=c(0,70),
        xlab="Niveles de la variable nutrition_grade_fr",
        ylab= "gr. por cada  100 gr de producto")
```




```
boxplot(fiber~nutrition_grade_fr, data = DFU_G, main= "Cantidad de fibra por nivel
        \n de la variable nutrition_grade_fr", ylim=c(0,20),
        xlab="Niveles de la variable nutrition_grade_fr",
        ylab= "gr. por cada  100 gr de producto")
```




```
boxplot(proteins~nutrition_grade_fr, data = DFU_G,
   main= "Cantidad de proteinas por nivel
        \n de la variable nutrition_grade_fr", ylim=c(0,40),
        xlab="Niveles de la variable nutrition_grade_fr",
        ylab= "gr. por cada  100 gr de producto")
```



Analizando los diagramas de caja se ve que los clasifica según los niveles de azucar y niveles de grasa, de menor (a) a mayor(e), lo que equivaldría a decir que la clasificación es de menor a mayor nivel enérgetico.

Si vemos el informe de la HSCP, (en la página 7) viene a decir que para la clasificación de los alimentos a adoptado un sistema de 5 clases (5-C, a<–>e) que es la representación gráfica de un sistema de información nutricional de un marcador continuo de calidad nutricional calculados a partir de la declaración nutricional (marcador FSA)

También se indica que este marcador FSA y el sistema 5-C estan muy relacionados con la densidad energética.

Esto nos confirma lo analizado en el párrafo anterior sobre la distribución de los diagramas de caja de las variables azucar y grasa.