

Exploración de los datos y Preprocesamiento

Roberto Saborit Roig

22/3/2021

Contents

Exploración de los datos	1
Tipos de variables	1
Datos y valores ausentes	2
Distribución de la variable respuesta	3
Distribución de las variables predictoras	5
Distribución de las variables cualitativas	9

Exploración de los datos

Antes de comenzar a generar el modelo, incluso antes del preprocesamiento, vamos a realizar una exploración de los datos con los siguientes objetivos:

- Ver si existen valores ausentes en el conjunto de datos y ver su distribución entre las distintas variables.
- Explorar los tipos de variable y ver si necesitamos cambiar el tipo de alguna variable.
- Ver la distribución de las variables, tanto de la respuesta como de las variables descriptivas.

Tipos de variables

La primera comprobación que haremos será ver los tipos de variables que hay y si todas tienen el tipo de valor que le corresponde:

```
oasis_cross_sectional <- read.csv("oasis_cross-sectional.csv")
oasis_longitudinal <- readxl::read_excel("oasis_longitudinal_demographics.xlsx")

str(oasis_cross_sectional)
```

```
## 'data.frame':   436 obs. of  12 variables:
## $ ID   : chr   "OAS1_0001_MR1" "OAS1_0002_MR1" "OAS1_0003_MR1" "OAS1_0004_MR1" ...
## $ M.F  : chr   "F" "F" "F" "M" ...
## $ Hand : chr   "R" "R" "R" "R" ...
## $ Age  : int   74 55 73 28 18 24 21 20 74 52 ...
## $ Educ : int    2 4 4 NA NA NA NA NA 5 3 ...
## $ SES  : int    3 1 3 NA NA NA NA NA 2 2 ...
```

```
## $ MMSE : int 29 29 27 NA NA NA NA NA 30 30 ...
## $ CDR : num 0 0 0.5 NA NA NA NA NA 0 0 ...
## $ eTIV : int 1344 1147 1454 1588 1737 1131 1516 1505 1636 1321 ...
## $ nWBV : num 0.743 0.81 0.708 0.803 0.848 0.862 0.83 0.843 0.689 0.827 ...
## $ ASF : num 1.31 1.53 1.21 1.1 1.01 ...
## $ Delay: chr "N/A" "N/A" "N/A" "N/A" ...
```

```
str(oasis_longitudinal)
```

```
## tibble [373 x 15] (S3: tbl_df/tbl/data.frame)
## $ Subject ID: chr [1:373] "OAS2_0001" "OAS2_0001" "OAS2_0002" "OAS2_0002" ...
## $ MRI ID : chr [1:373] "OAS2_0001_MR1" "OAS2_0001_MR2" "OAS2_0002_MR1" "OAS2_0002_MR2" ...
## $ Group : chr [1:373] "Nondemented" "Nondemented" "Demented" "Demented" ...
## $ Visit : chr [1:373] "1" "2" "1" "2" ...
## $ MR Delay : num [1:373] 0 457 0 560 1895 ...
## $ M/F : chr [1:373] "M" "M" "M" "M" ...
## $ Hand : chr [1:373] "R" "R" "R" "R" ...
## $ Age : num [1:373] 87 88 75 76 80 88 90 80 83 85 ...
## $ EDUC : num [1:373] 14 14 12 12 12 18 18 12 12 12 ...
## $ SES : num [1:373] 2 2 NA NA NA 3 3 4 4 4 ...
## $ MMSE : num [1:373] 27 30 23 28 22 28 27 28 29 30 ...
## $ CDR : num [1:373] 0 0 0.5 0.5 0.5 0 0 0 0.5 0 ...
## $ eTIV : num [1:373] 1987 2004 1678 1738 1698 ...
## $ nWBV : num [1:373] 0.696 0.681 0.736 0.713 0.701 ...
## $ ASF : num [1:373] 0.883 0.876 1.046 1.01 1.034 ...
```

Datos y valores ausentes

Vamos a comprobar ahora el número de datos que tenemos y la cantidad de valores ausentes que hay:

```
#El número total de filas nos indica la cantidad de medidas
```

```
nrow(oasis_cross_sectional)
```

```
## [1] 436
```

```
nrow(oasis_longitudinal)
```

```
## [1] 373
```

```
any(is.na(oasis_cross_sectional)); any(is.na(oasis_longitudinal))
```

```
## [1] TRUE
```

```
## [1] TRUE
```

```
#Tenemos datos ausentes en ambos conjuntos
```

```
#Podemos comprobar que variables tienen mayor porcentaje de NA y la cantidad total
```

```
apply(is.na(oasis_cross_sectional), 2, mean); apply(is.na(oasis_cross_sectional), 2, sum)
```

```
##      ID      M.F      Hand      Age      Educ      SES      MMSE      CDR
## 0.0000000 0.0000000 0.0000000 0.0000000 0.4610092 0.5045872 0.4610092 0.4610092
##      eTIV      nWBV      ASF      Delay
## 0.0000000 0.0000000 0.0000000 0.0000000
```

```
##      ID      M.F      Hand      Age      Educ      SES      MMSE      CDR      eTIV      nWBV      ASF      Delay
##      0      0      0      0      201      220      201      201      0      0      0      0
```

```
apply(is.na(oasis_longitudinal), 2, mean); apply(is.na(oasis_longitudinal), 2, sum)
```

```
## Subject ID      MRI ID      Group      Visit      MR Delay      M/F      Hand
## 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
##      Age      EDUC      SES      MMSE      CDR      eTIV      nWBV
## 0.00000000 0.00000000 0.05093834 0.00536193 0.00000000 0.00000000 0.00000000
##      ASF
## 0.00000000
```

```
## Subject ID      MRI ID      Group      Visit      MR Delay      M/F      Hand
##      0      0      0      0      0      0      0
##      Age      EDUC      SES      MMSE      CDR      eTIV      nWBV
##      0      0      19      2      0      0      0
##      ASF
##      0
```

En el caso del estudio seccional tenemos una gran cantidad de datos ausentes en las variables Educ, SES, MMSE, CDR, que suponen casi un 50% de los datos, en esas variables, esto será importante para tenerlo en cuenta al dividir los datos, y que los datos de entrenamiento o de test, no tengan un gran número de datos ausentes, ya que esto puede afectar al modelo.

En cambio en el conjunto de datos longitudinal no tenemos apenas datos ausentes, solo unos pocos en la variable SES, que es el estatus socioeconómico.

Distribución de la variable respuesta

```
library(ggplot2)

par(mfrow=c(1,2))

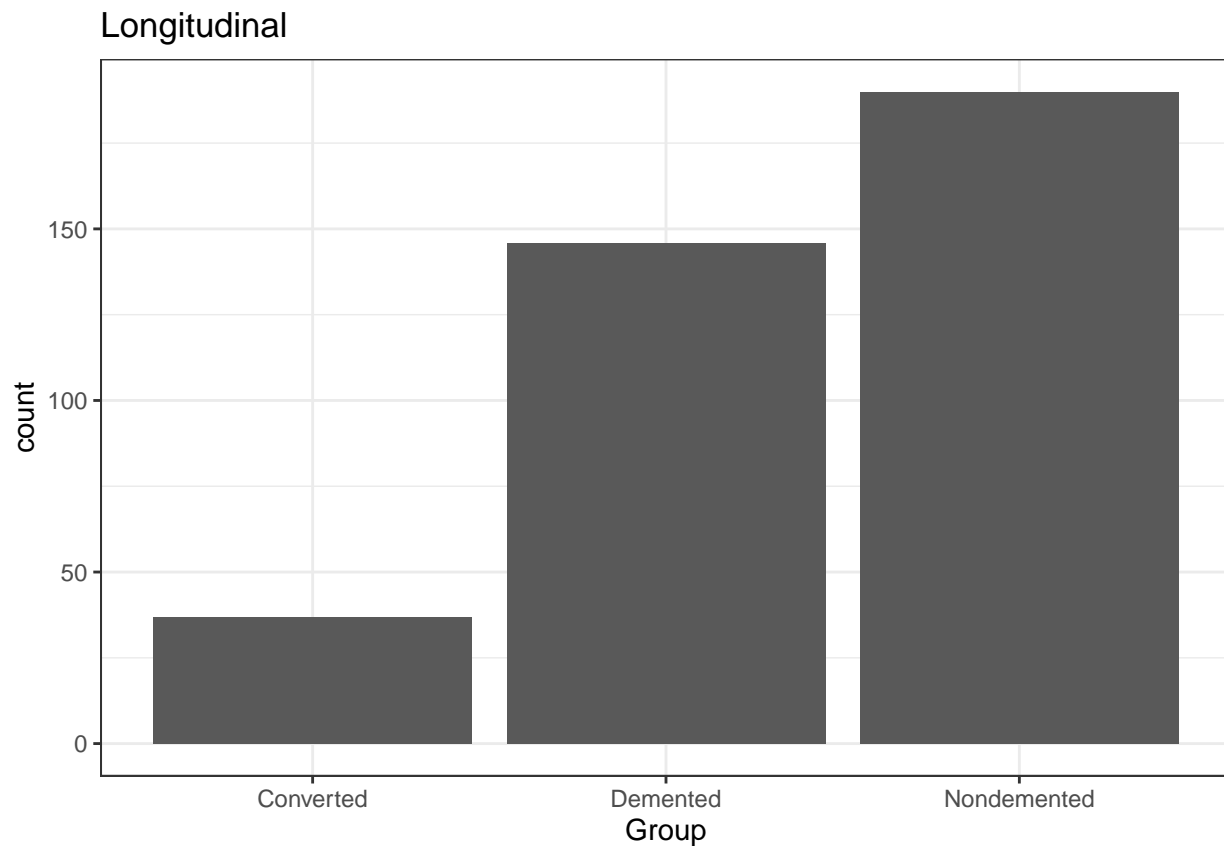
table(oasis_longitudinal$Group)
```

```
##
##      Converted      Demented      Nondemented
##      37      146      190
```

```
round(prop.table(table(oasis_longitudinal$Group)), 2)
```

```
##
##      Converted      Demented      Nondemented
##      0.10      0.39      0.51
```

```
ggplot(data = oasis_longitudinal, aes(x = Group, y = ..count.., )) +
  geom_bar() +
  labs(title = "Longitudinal") +
  theme_bw() +
  theme(legend.position = "bottom")
```

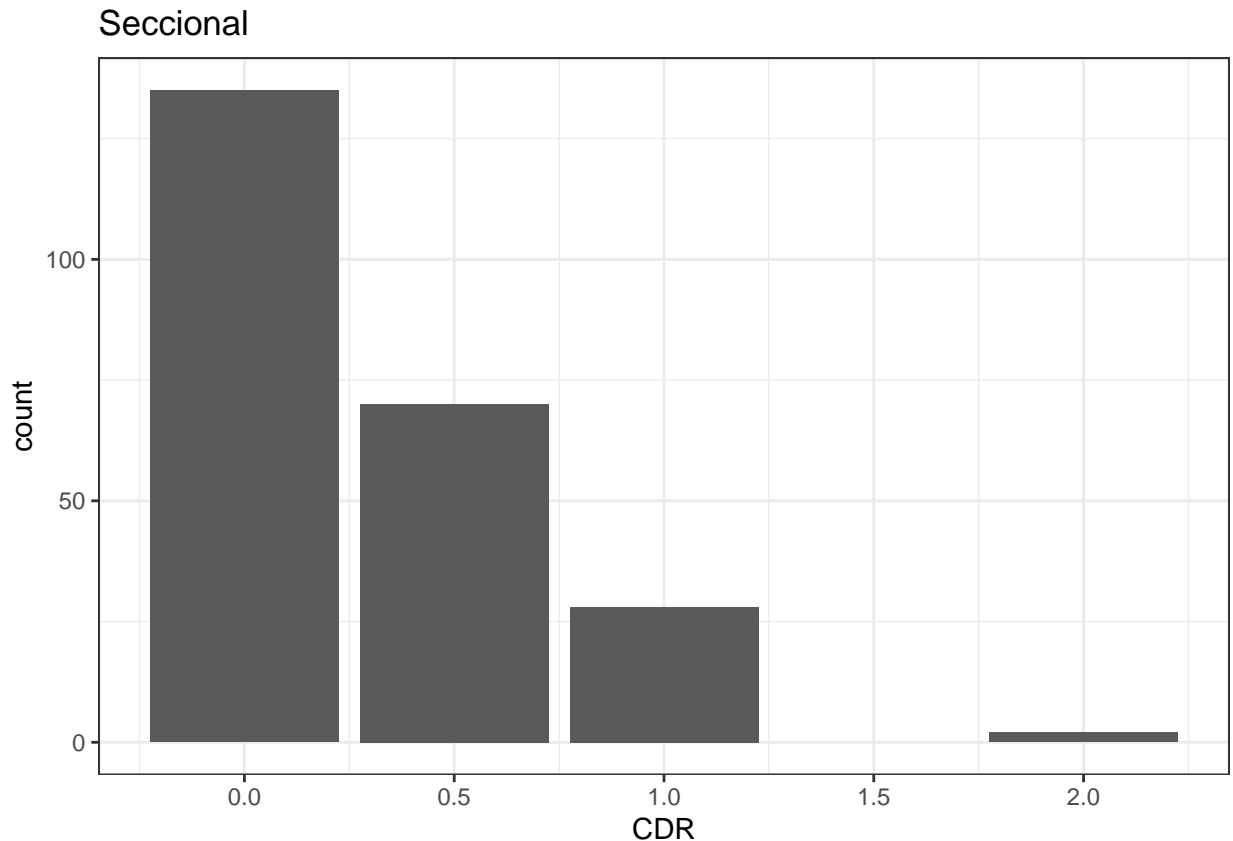


```
table(oasis_cross_sectional$CDR)
```

```
##
##  0 0.5  1  2
## 135 70 28  2
```

```
ggplot(data = oasis_cross_sectional, aes(x = CDR, y = ..count.., )) +
  geom_bar() +
  labs(title = "Seccional") +
  theme_bw() +
  theme(legend.position = "bottom")
```

```
## Warning: Removed 201 rows containing non-finite values (stat_count).
```



Como vemos el porcentaje de convertidos que son aquellos que al principio del experimento no tenían demencia y en las sucesivas medidas la desarrollaron, es del 10% de los datos, esto es importante conocerlo, si queremos crear un modelo efectivo es importante que acierte más del 10% de convertidos, que podría acertarse si simplemente clasificamos todos los sujetos como convertidos.

En el caso del estudio seccional lo hemos dividido en grupos según la variable CDR que muestra si se no se tiene demencia (0), si se tiene ver-mild-dementia (0.5), mild-dementia (1) o demencia (2).

Distribución de las variables predictoras

Variables continuas

```
library(ggpubr)
```

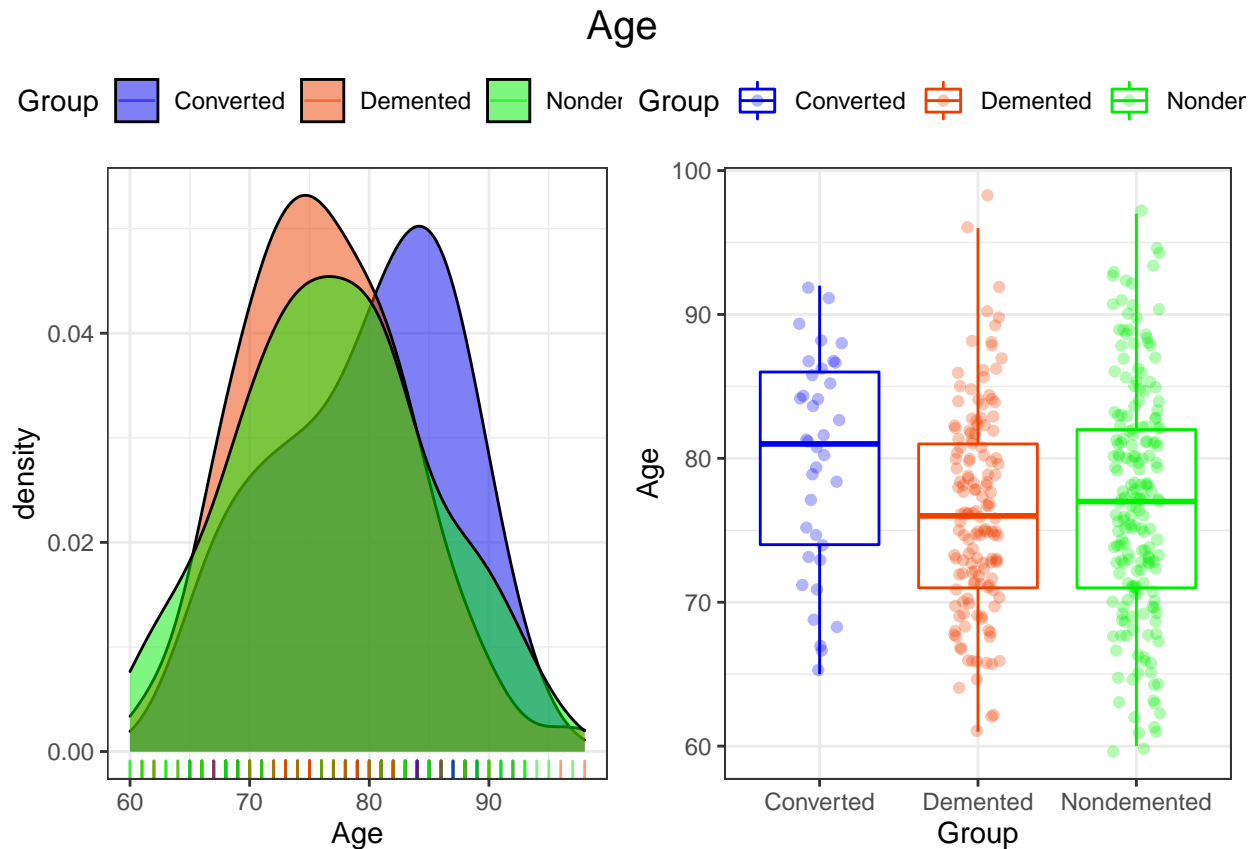
```
## Warning: package 'ggpubr' was built under R version 4.0.3
```

```
p1 <- ggplot(data = oasis_longitudinal, aes(x = Age, fill = Group)) +
  geom_density(alpha = 0.5) +
  scale_fill_manual(values = c("blue2", "orangered2", "green2")) +
  geom_rug(aes(color = Group), alpha = 0.5) +
  scale_color_manual(values = c("blue2", "orangered2", "green2")) +
  theme_bw()
p2 <- ggplot(data = oasis_longitudinal, aes(x = Group, y = Age, color = Group)) +
  geom_boxplot(outlier.shape = NA) +
```

```

geom_jitter(alpha = 0.3, width = 0.15) +
scale_color_manual(values = c("blue2", "orangered2", "green2")) +
theme_bw()
final_plot <- ggarrange(p1, p2, legend = "top")
final_plot <- annotate_figure(final_plot, top = text_grob("Age", size = 15))
final_plot

```



```

tapply(oasis_longitudinal$Age, oasis_longitudinal$Group, mean)

```

```

##   Converted   Demented Nondemented
##   79.75676   76.26027   77.05789

```

Como se ve el grupo converted tiene una edad media significativamente más baja que las otras dos variables.

```

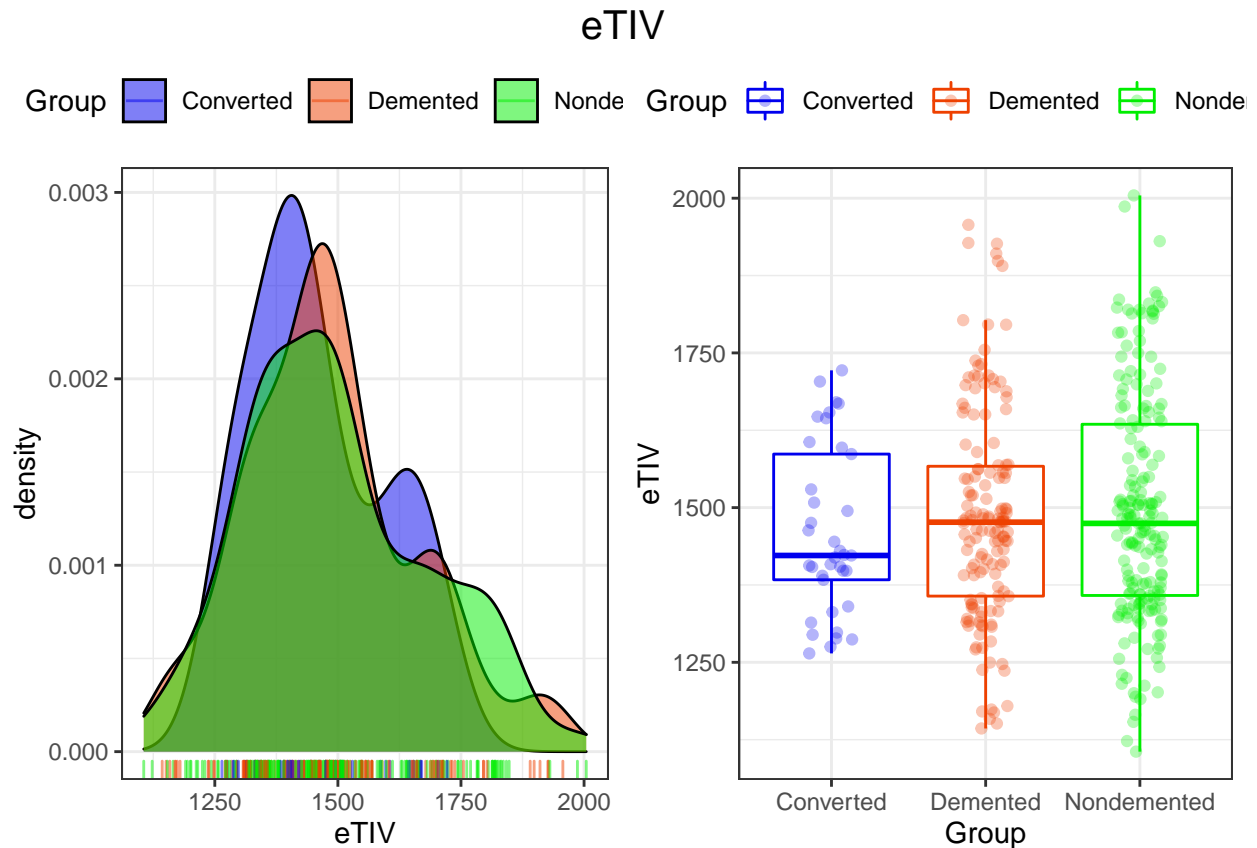
p1 <- ggplot(data = oasis_longitudinal, aes(x = eTIV, fill = Group)) +
  geom_density(alpha = 0.5) +
  scale_fill_manual(values = c("blue2", "orangered2", "green2")) +
  geom_rug(aes(color = Group), alpha = 0.5) +
  scale_color_manual(values = c("blue2", "orangered2", "green2")) +
  theme_bw()
p2 <- ggplot(data = oasis_longitudinal, aes(x = Group, y = eTIV, color = Group)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(alpha = 0.3, width = 0.15) +
  scale_color_manual(values = c("blue2", "orangered2", "green2")) +

```

```

theme_bw()
final_plot <- ggarrange(p1, p2, legend = "top")
final_plot <- annotate_figure(final_plot, top = text_grob("eTIV", size = 15))
final_plot

```



```

tapply(oasis_longitudinal$eTIV, oasis_longitudinal$Group, mean)

```

```

##   Converted   Demented Nondemented
##   1459.347   1485.848   1495.472







```

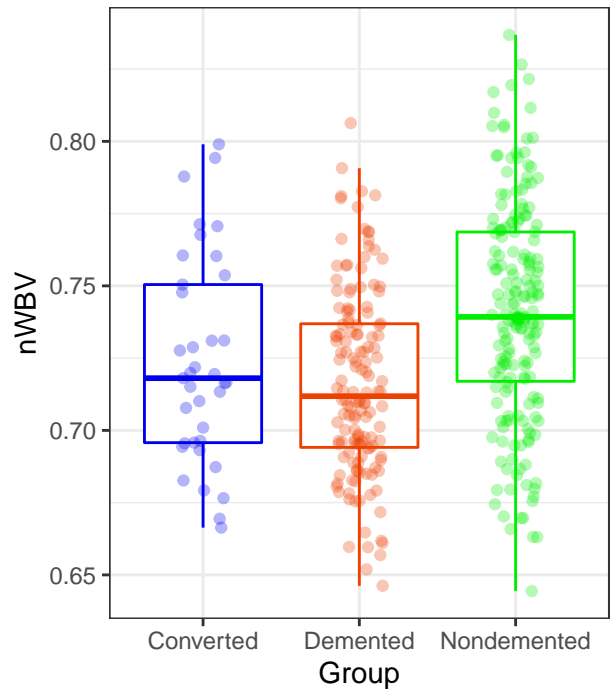
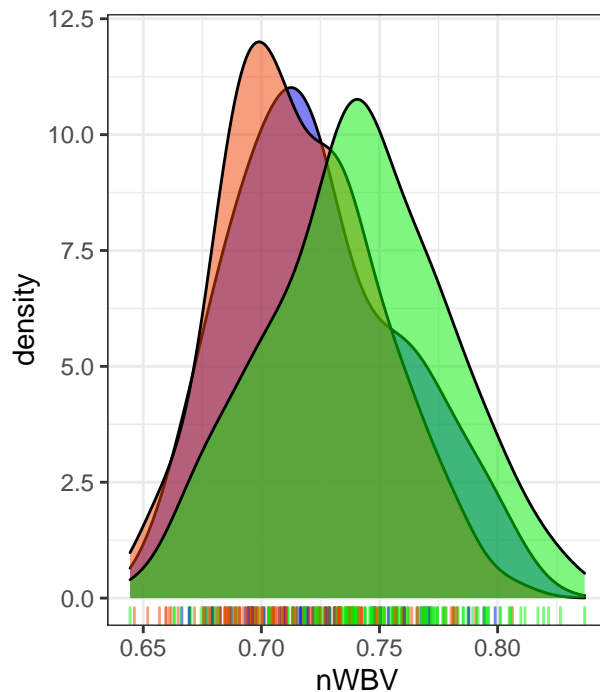
```

p1 <- ggplot(data = oasis_longitudinal, aes(x = nWBV, fill = Group)) +
  geom_density(alpha = 0.5) +
  scale_fill_manual(values = c("blue2", "orangered2", "green2")) +
  geom_rug(aes(color = Group), alpha = 0.5) +
  scale_color_manual(values = c("blue2", "orangered2", "green2")) +
  theme_bw()
p2 <- ggplot(data = oasis_longitudinal, aes(x = Group, y = nWBV, color = Group)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(alpha = 0.3, width = 0.15) +
  scale_color_manual(values = c("blue2", "orangered2", "green2")) +
  theme_bw()
final_plot <- ggarrange(p1, p2, legend = "top")
final_plot <- annotate_figure(final_plot, top = text_grob("nWBV", size = 15))
final_plot

```

nWBV

Group  Converted  Demented  Nondemented Group  Converted  Demented  Nondemented



```
tapply(oasis_longitudinal$nWBV, oasis_longitudinal$Group, mean)
```

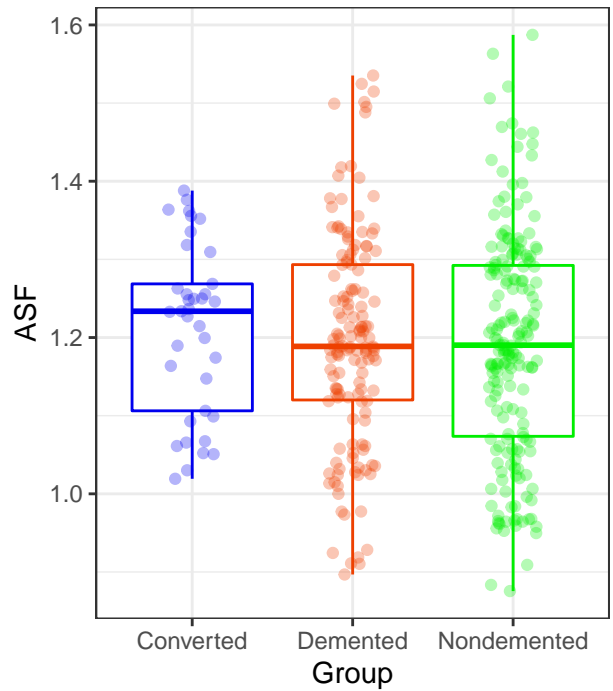
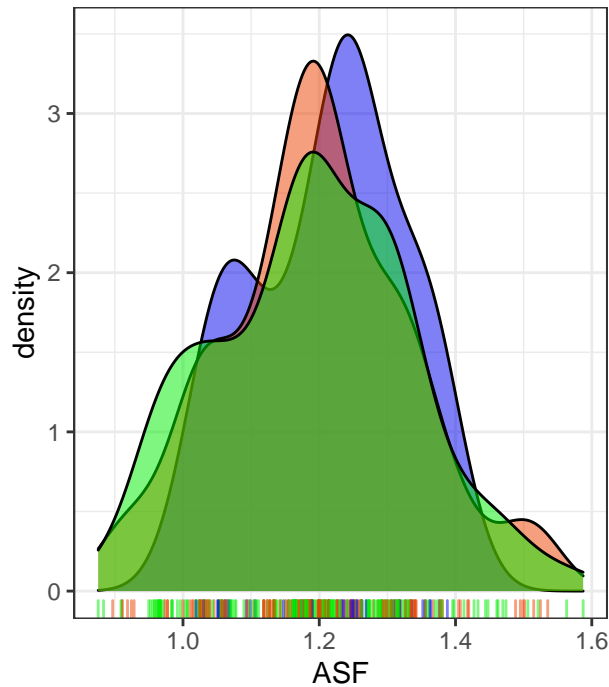
```
##   Converted   Demented Nondemented
##   0.7237336   0.7163034   0.7408726
```

En esta variable sí se ven diferencias entre los grupos, los no dementes tienen claramente valores más altos, que los dementes y los convertidos.

```
p1 <- ggplot(data = oasis_longitudinal, aes(x = ASF, fill = Group)) +
  geom_density(alpha = 0.5) +
  scale_fill_manual(values = c("blue2", "orangered2", "green2")) +
  geom_rug(aes(color = Group), alpha = 0.5) +
  scale_color_manual(values = c("blue2", "orangered2", "green2")) +
  theme_bw()
p2 <- ggplot(data = oasis_longitudinal, aes(x = Group, y = ASF, color = Group)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(alpha = 0.3, width = 0.15) +
  scale_color_manual(values = c("blue2", "orangered2", "green2")) +
  theme_bw()
final_plot <- ggarrange(p1, p2, legend = "top")
final_plot <- annotate_figure(final_plot, top = text_grob("ASF", size = 15))
final_plot
```


ASF

Group ■ Converted ■ Demented ■ Nondem Group ■ Converted ■ Demented ■ Nonden



```
tapply(oasis_longitudinal$ASF, oasis_longitudinal$Group, mean)
```

```
##   Converted   Demented Nondemented
##   1.212422   1.196880   1.191066
```

En este caso no parece haber difeencias en la distribución

Distribución de las variables cualitativas