Caso PIMA

- Los Pima son un grupo de nativos de Arizona. En los ultimos anios se observo un aumento en la prevalencia de diabetes tipo 2, asociado a cambios en la dieta y a una disminucion de la actividad fisica, por lo que han sido objeto de muchos estudios.
- La base de datos pertenece al Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales. 768 mujeres adultas
- El objetivo del conjunto de datos es predecir si una paciente tiene diabetes o no, basandose en ciertas mediciones de diagnostico incluidas en el conjunto de datos.

Describa la base de datos en forma univariada mediante graficos y estadisticos para las variables cuantitativas y tablas de frecuencias para las cualitativas

Para las cuanti:

1. Cual es la tendencia central de las variables y su rango?

```
library(faraway)

db <- faraway::pima
pima <- db

#Cambio los ceros por NA

pima$glucose[pima$glucose == 0] <- NA
pima$diastolic[pima$diastolic == 0] <- NA
pima$triceps[pima$triceps == 0] <- NA
pima$insulin[pima$insulin == 0] <- NA
pima$bmi[pima$bmi == 0] <- NA

#Clasifique a las pacientes segC:n su peso en normales, con sobrepeso o con
#obesidad
pima$cat_peso <- ifelse(pima$bm < 25, "Normal",
                        ifelse(pima$bm >= 25 & pima$bm < 30, "Con sobrepeso",
                               "Obesidad"))
#Genere la variable diabetes a partir de test (si = 1, no=0)

pima$result <- ifelse(pima$test == 1, "si","no")

colnames(pima)[7]<-"diabetes_gen"

sapply(pima, class)
```

```
##     pregnant      glucose    diastolic      triceps      insulin          bmi
##    "integer"    "integer"    "integer"    "integer"    "integer"    "numeric"
## diabetes_gen          age         test     cat_peso       result
##    "numeric"    "integer"    "integer"  "character"  "character"
```

```r
pima$pregnant = as.numeric(pima$pregnant)
pima$cat_peso = as.factor(pima$cat_peso)
pima$result = as.factor(pima$result)

summary(pima)
```

```
##     pregnant         glucose        diastolic         triceps
##  Min.   : 0.000   Min.   : 44.0   Min.   : 24.00   Min.   : 7.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 64.00   1st Qu.:22.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :29.00
##  Mean   : 3.845   Mean   :121.7   Mean   : 72.41   Mean   :29.15
##  3rd Qu.: 6.000   3rd Qu.:141.0   3rd Qu.: 80.00   3rd Qu.:36.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##                   NA's   :5       NA's   :35       NA's   :227
##     insulin            bmi         diabetes_gen         age
##  Min.   : 14.00   Min.   :18.20   Min.   :0.0780   Min.   :21.00
##  1st Qu.: 76.25   1st Qu.:27.50   1st Qu.:0.2437   1st Qu.:24.00
##  Median :125.00   Median :32.30   Median :0.3725   Median :29.00
##  Mean   :155.55   Mean   :32.46   Mean   :0.4719   Mean   :33.24
##  3rd Qu.:190.00   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
##  Max.   :846.00   Max.   :67.10   Max.   :2.4200   Max.   :81.00
##  NA's   :374      NA's   :11
##       test               cat_peso   result
##  Min.   :0.000   Con sobrepeso:179   no:500
##  1st Qu.:0.000   Normal       :106   si:268
##  Median :0.000   Obesidad     :472
##  Mean   :0.349   NA's         : 11
##  3rd Qu.:1.000
##  Max.   :1.000
##
```

2. Que variables presentan mayor variabilidad?

```r
library(pastecs)

stat.desc(pima)
```

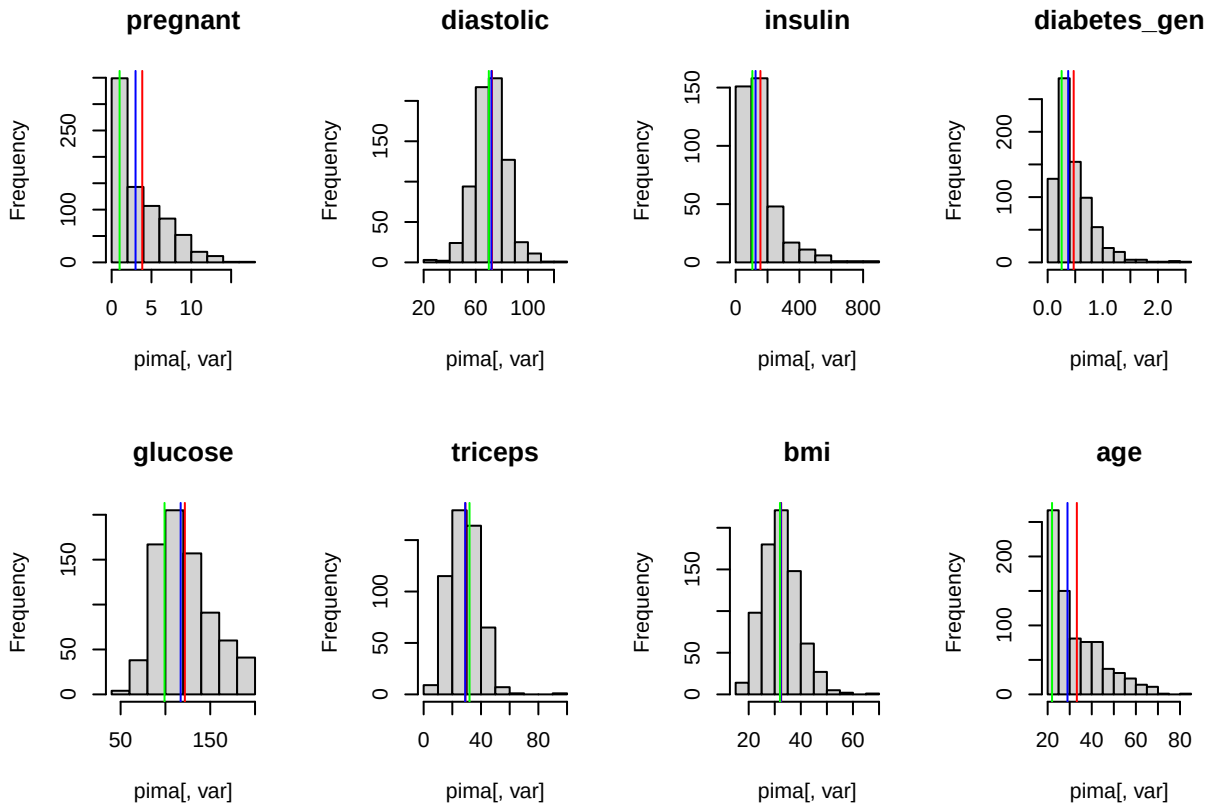
```
##                    pregnant      glucose    diastolic      triceps      insulin
## nbr.val       768.0000000 7.630000e+02 7.330000e+02 5.410000e+02 3.940000e+02
## nbr.null      111.0000000 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
## nbr.na          0.0000000 5.000000e+00 3.500000e+01 2.270000e+02 3.740000e+02
## min             0.0000000 4.400000e+01 2.400000e+01 7.000000e+00 1.400000e+01
## max            17.0000000 1.990000e+02 1.220000e+02 9.900000e+01 8.460000e+02
## range          17.0000000 1.550000e+02 9.800000e+01 9.200000e+01 8.320000e+02
## sum          2953.0000000 9.284700e+04 5.307300e+04 1.577200e+04 6.128600e+04
## median          3.0000000 1.170000e+02 7.200000e+01 2.900000e+01 1.250000e+02
## mean            3.8450521 1.216868e+02 7.240518e+01 2.915342e+01 1.555482e+02
## SE.mean         0.1215892 1.105464e+00 4.573454e-01 4.504407e-01 5.983841e+00
## CI.mean.0.95    0.2386871 2.170117e+00 8.978652e-01 8.848307e-01 1.176434e+01
## var            11.3540563 9.324254e+02 1.533178e+02 1.097672e+02 1.410770e+04
## std.dev         3.3695781 3.053564e+01 1.238216e+01 1.047698e+01 1.187759e+02
## coef.var        0.8763413 2.509364e-01 1.710120e-01 3.593740e-01 7.635951e-01
```

```
##                      bmi diabetes_gen          age         test cat_peso
## nbr.val      7.570000e+02 768.00000000 7.680000e+02 768.00000000       NA
## nbr.null     0.000000e+00   0.00000000 0.000000e+00 500.00000000       NA
## nbr.na       1.100000e+01   0.00000000 0.000000e+00   0.00000000       NA
## min          1.820000e+01   0.07800000 2.100000e+01   0.00000000       NA
## max          6.710000e+01   2.42000000 8.100000e+01   1.00000000       NA
## range        4.890000e+01   2.34200000 6.000000e+01   1.00000000       NA
## sum          2.457030e+04 362.40100000 2.552900e+04 268.00000000       NA
## median       3.230000e+01   0.37250000 2.900000e+01   0.00000000       NA
## mean         3.245746e+01   0.47187630 3.324089e+01   0.34895833       NA
## SE.mean      2.516930e-01   0.01195579 4.243608e-01   0.01721050       NA
## CI.mean.0.95 4.941002e-01   0.02346996 8.330464e-01   0.03378527       NA
## var          4.795546e+01   0.10977864 1.383030e+02   0.22748262       NA
## std.dev      6.924988e+00   0.33132860 1.176023e+01   0.47695138       NA
## coef.var     2.133558e-01   0.70215138 3.537882e-01   1.36678604       NA
##              result
## nbr.val          NA
## nbr.null         NA
## nbr.na           NA
## min              NA
## max              NA
## range            NA
## sum              NA
## median           NA
## mean             NA
## SE.mean          NA
## CI.mean.0.95     NA
## var              NA
## std.dev          NA
## coef.var         NA
```

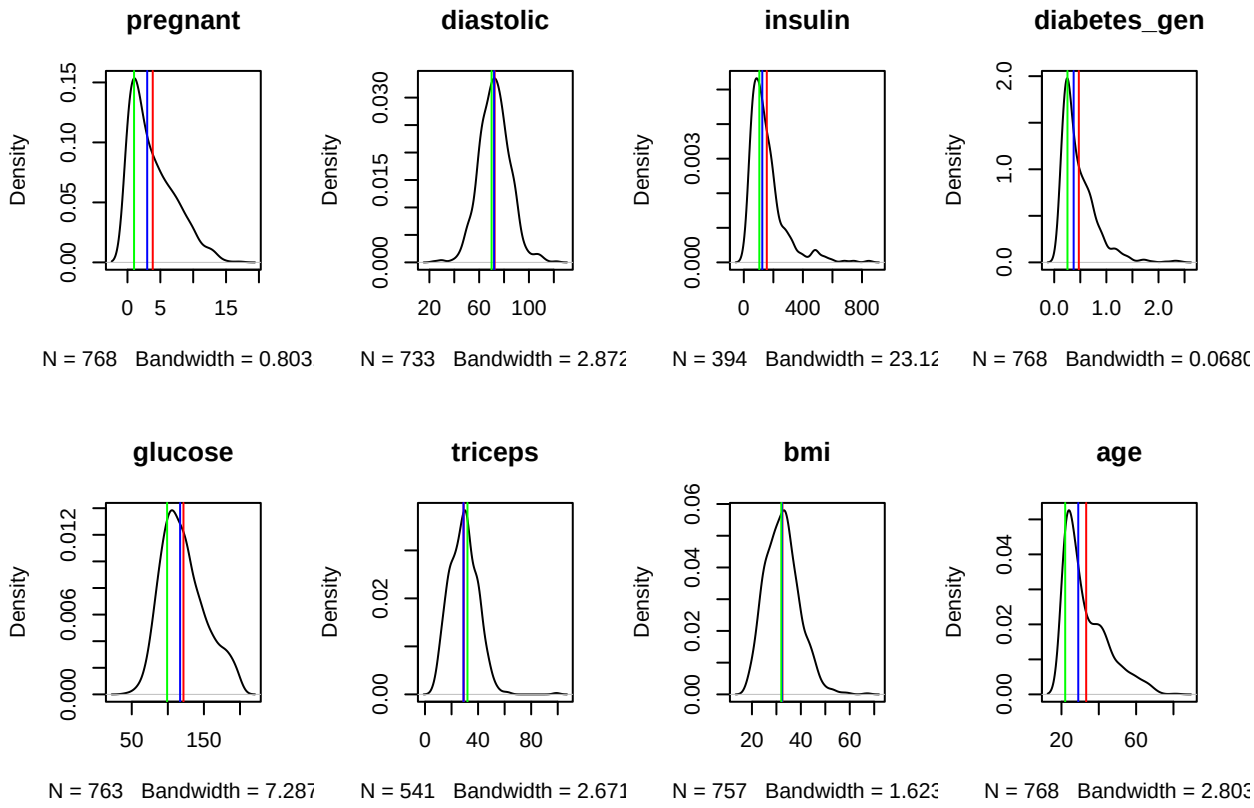
3. Que tipo de asimetria presentan las variables?

```
getmode<-function(x){
  return(as.numeric(rownames(data.frame(which.max(table(x))))))
}
layout(matrix(nrow = 2,ncol = 4,data = c(1:8)))
for (var in colnames(pima)[c(1:8)]){
    hist(pima[,var],main = var)
    #mean in red, median in blue, mode in green
    abline(v=mean(na.omit(pima[,var])),col="red")
    abline(v=median(na.omit(pima[,var])),col="blue")
    abline(v=getmode(na.omit(pima[,var])),col="green")

}
```

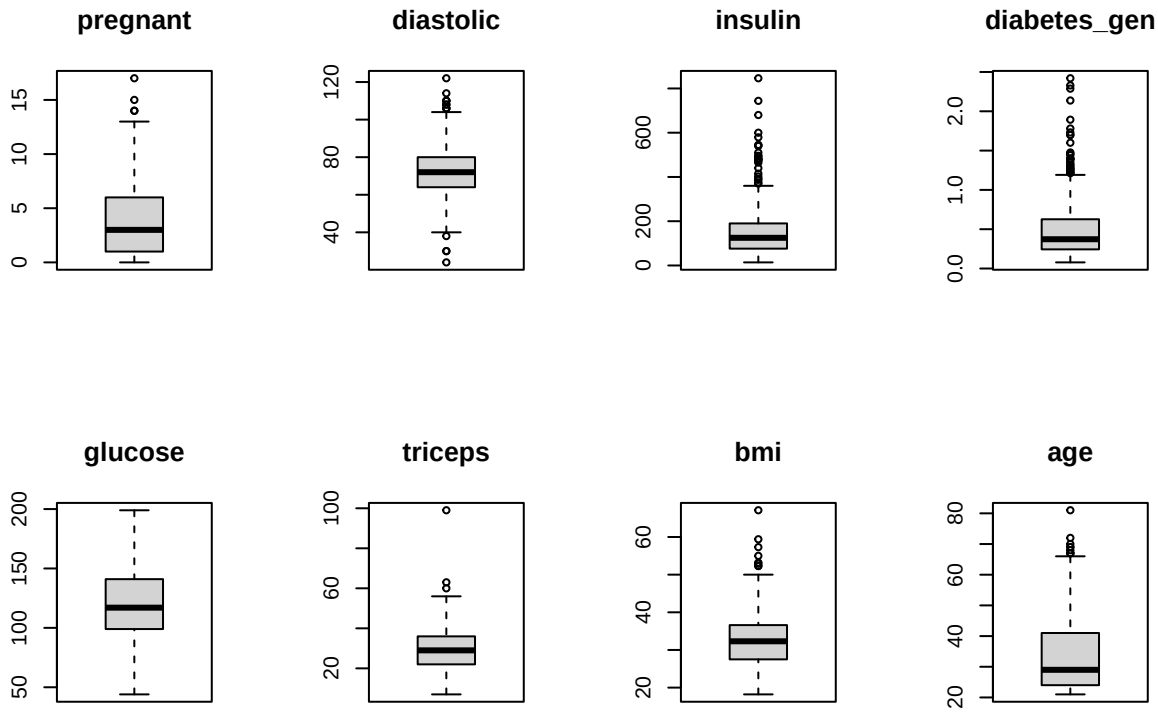



```
getmode<-function(x){
  return(as.numeric(rownames(data.frame(which.max(table(x))))))
}
layout(matrix(nrow = 2,ncol = 4,data = c(1:8)))
for (var in colnames(pima)[c(1:8)]){
  plot(density(na.omit(pima[,var])),main = var)
  #mean in red, median in blue, mode in green
  abline(v=mean(na.omit(pima[,var])),col="red")
  abline(v=median(na.omit(pima[,var])),col="blue")
  abline(v=getmode(na.omit(pima[,var])),col="green")
}
```



4. Hay valores atipicos?

```
layout(matrix(nrow = 2,ncol = 4,data = c(1:8)))
for (var in colnames(pima)[c(1:8)]){
  boxplot(pima[,var],main=var)
}
```



Para las cuali: 1. Cuales son las categoricas mas frecuentes? En que porcentaje?

```
prop.table(table(pima$cat_peso))
```

```
##
## Con sobrepeso        Normal      Obesidad
##     0.2364597     0.1400264     0.6235139
```

```
prop.table(table(pima$result))
```

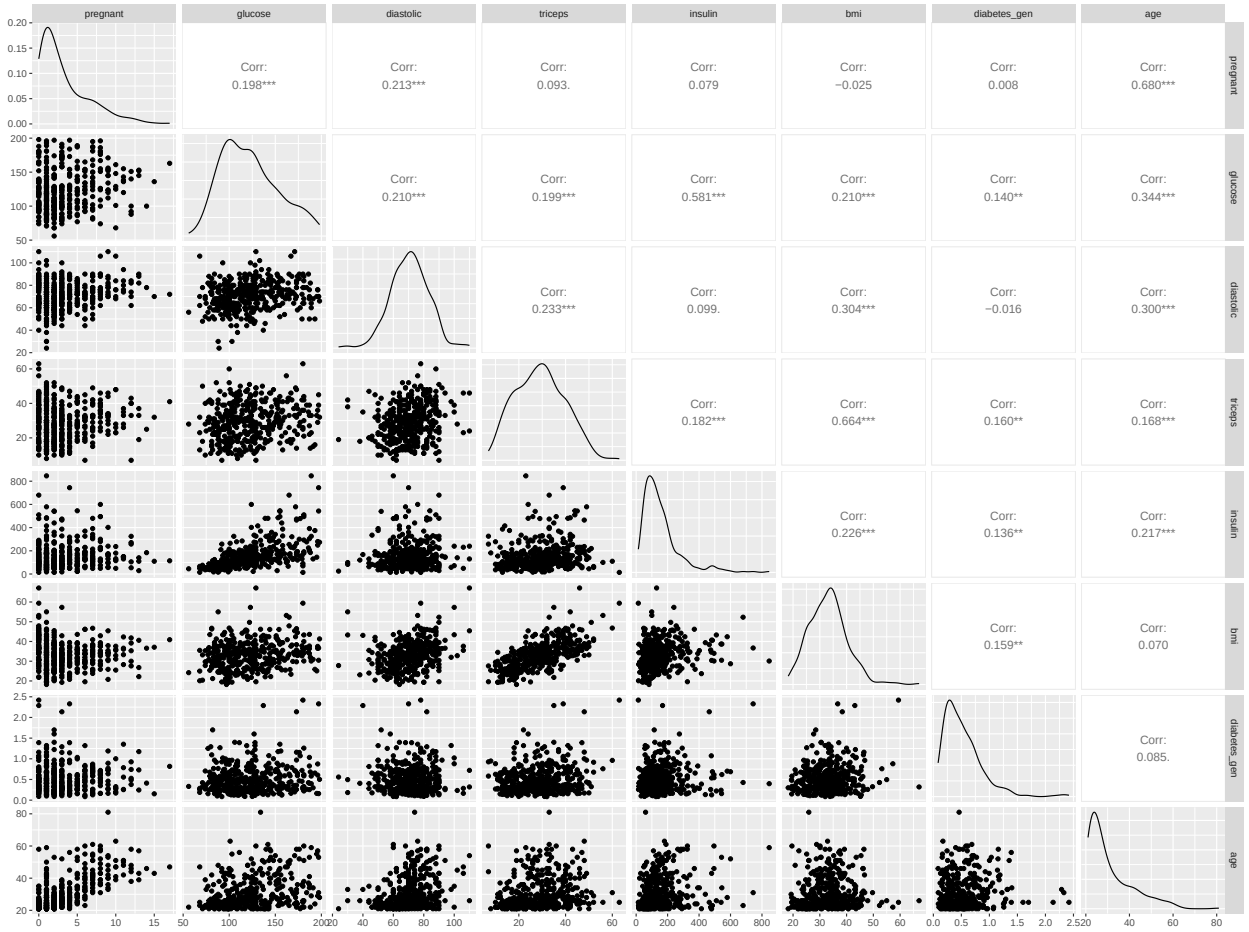
```
##
##        no        si
## 0.6510417 0.3489583
```

Lo mas comun es obesidad. El 62% de las mujeres de la muestra son obesas El 86% presenta sobrepeso u obesidad

El 35% presenta diabetes

BIVARIADA 1. Describa la base de datos en forma bivariada mediante graficos y estadisticos

```
library(GGally)
ggpairs(na.omit(pima),c(1:8))
```

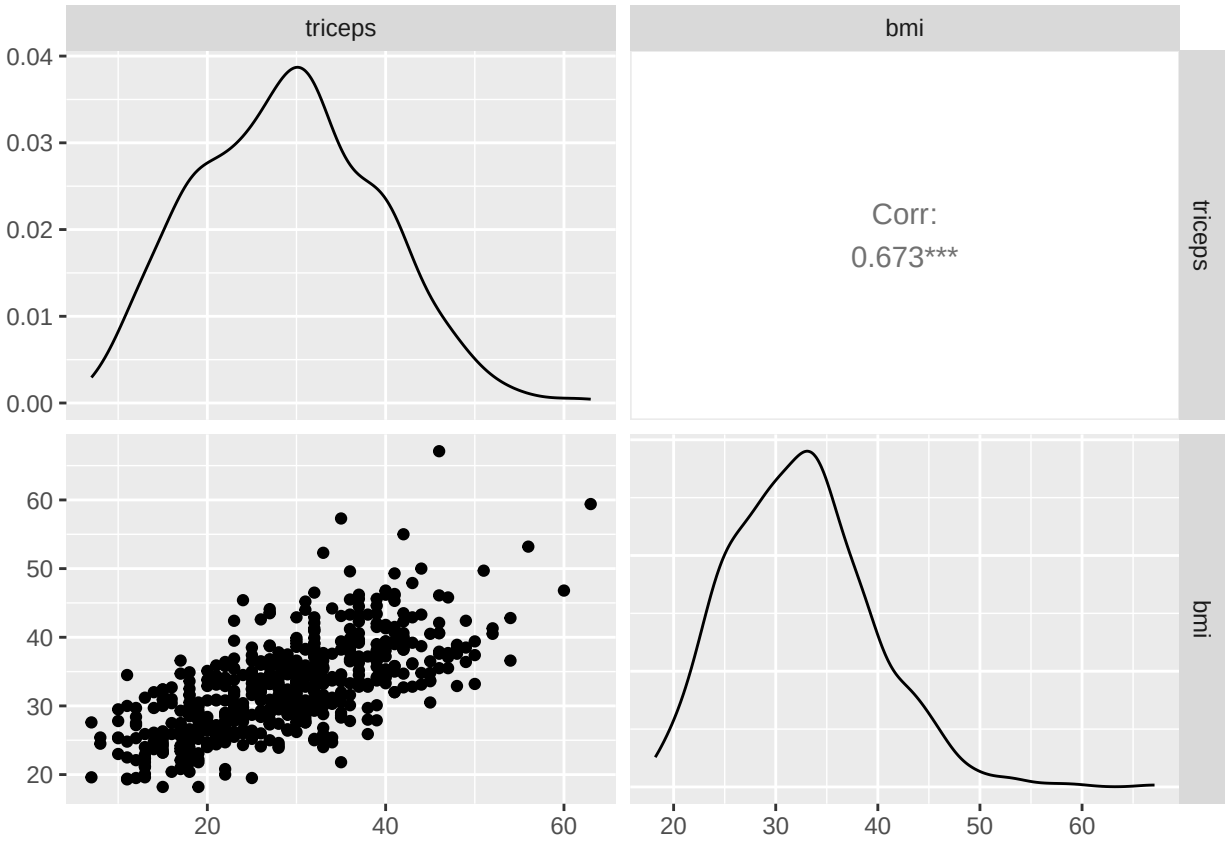


Se decide descartar la observacion de triceps mas alta porq consideramos que es un error de carga

```
pima$triceps[pima$triceps == 99] <- NA
```

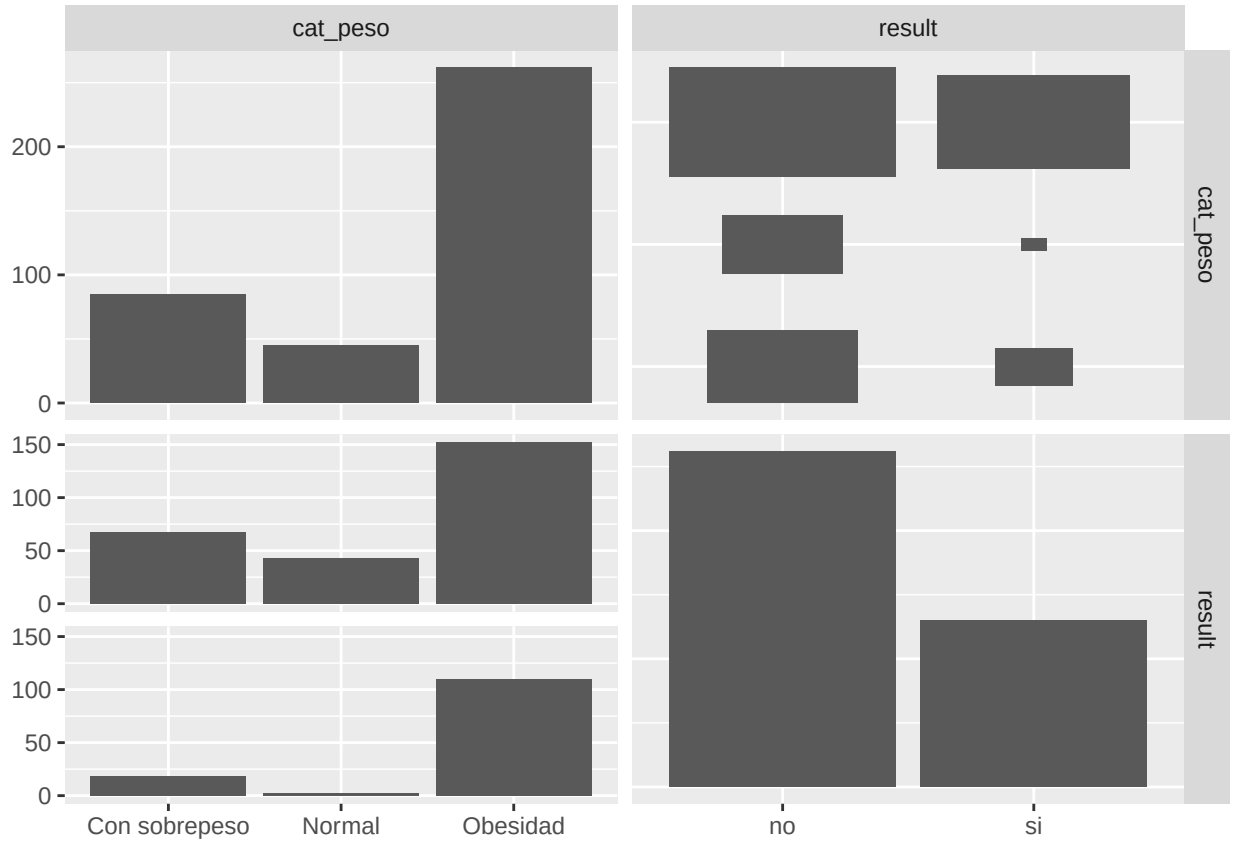
Las asociaciones mas fuertes son: bmi y triceps, nro embarazos y edad, glucosa e insulina. Todas asociaciones moderadas (r>0.33) y positivas

```
ggpairs(pima,columns = c(4,6))
```



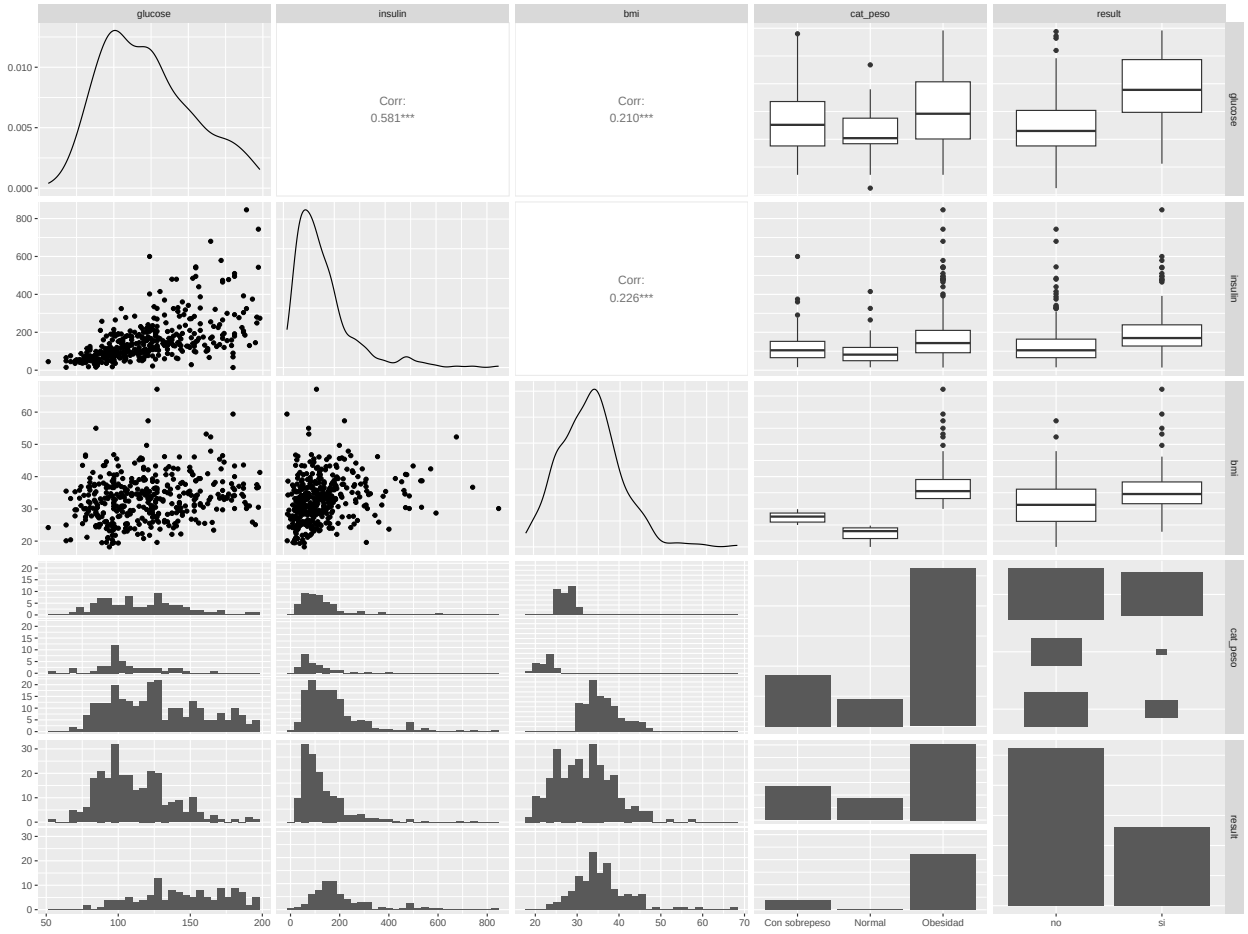
Cualitativas sin NA

```
library(GGally)
ggpairs(na.omit(pima),columns = c(10,11))
```



Cualitativa~Cuantitativa

```
ggpairs(na.omit(pima),columns=c(2,5,6,10,11))
```



2. Cual es el porcentaje de mujeres con sobrepeso y obesidad entre las que tienen o no diabetes?

```
table(pima$cat_peso, pima$result)
```

```
##
##                  no  si
##   Con sobrepeso 139  40
##   Normal         99   7
##   Obesidad      253 219
```

```
prop.table(table(pima$cat_peso, pima$result)) ##100% esta en toda la tabla
```

```
##
##                        no          si
##   Con sobrepeso 0.183619551 0.052840159
##   Normal        0.130779392 0.009247028
##   Obesidad      0.334214003 0.289299868
```

El 13% de todas las mujeres es normal y no diab El 33% de todas las mujeres es obesa y no diab El 18% de todas las mujeres es normal y diab ...

```r
prop.table(table(pima$cat_peso, pima$result),1) ##100% esta por fila
```

```
##
##                          no         si
##   Con sobrepeso 0.77653631 0.22346369
##   Normal        0.93396226 0.06603774
##   Obesidad      0.53601695 0.46398305
```

El 93% de las normales no son diabeticas

```r
prop.table(table(pima$cat_peso, pima$result),2) ##100% esta por columnas
```

```
##
##                          no         si
##   Con sobrepeso 0.28309572 0.15037594
##   Normal        0.20162933 0.02631579
##   Obesidad      0.51527495 0.82330827
```

El 20% de las no diab tienen peso normal El 3% de las diab tienen peso normal

```r
prop.table(table( pima[which(pima$cat_peso!="normales"), 10:11 ] ))
```

```
##               result
## cat_peso                   no          si
##   Con sobrepeso 0.183619551 0.052840159
##   Normal        0.130779392 0.009247028
##   Obesidad      0.334214003 0.289299868
```

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:pastecs':
##
##     first, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 4.2.3
```

```r
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

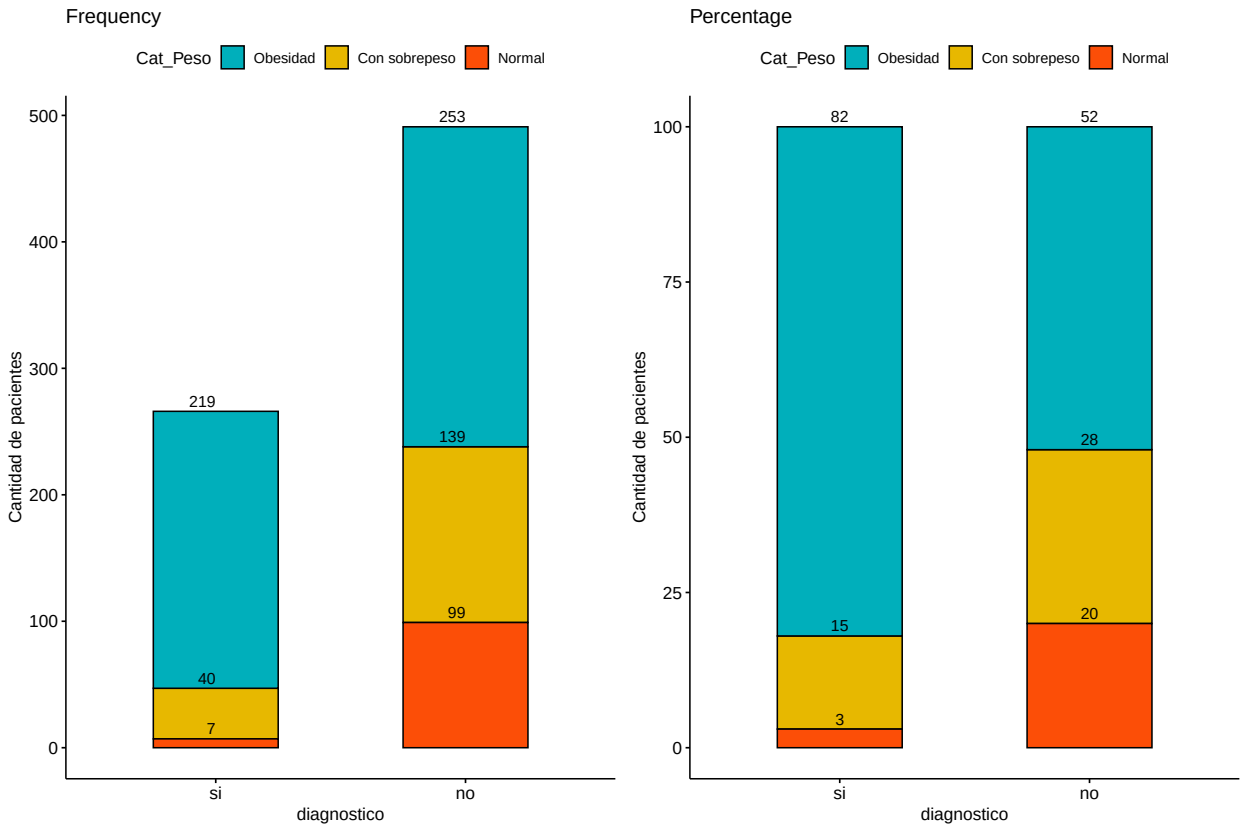
```r
datos1 <- data.frame(table(pima$cat_peso, pima$result))
colnames(datos1) <- c("Cat_Peso", "Result", "Freq")
datos1$Cat_Peso <- factor(datos1$Cat_Peso, levels = rev(c("Normal", "Con sobrepeso", "Obesidad")))
datos1$Result <- factor(datos1$Result, levels = rev(c("no", "si")))

# Use dplyr filter to calculate Porc
datos1 <- datos1 %>%
  group_by(Result) %>%
  mutate(Porc = round(Freq / sum(Freq) * 100, 0))

plot1 <- ggbarplot(datos1,x="Result",y="Freq",fill="Cat_Peso",
                   palette =c("#00AFBB", "#E7B800", "#FC4E07"),
                   xlab = "diagnostico",ylab="Cantidad de pacientes",label = T,  width=0.5,
                   title = "Frequency",lab.hjust = T)

plot2 <- ggbarplot(datos1,x="Result",y="Porc",fill="Cat_Peso",
                   palette =c("#00AFBB", "#E7B800", "#FC4E07"),
                   xlab = "diagnostico",ylab="Cantidad de pacientes",label = T,  width=0.5,
                   title = "Percentage")

grid.arrange(plot1, plot2, ncol = 2 )
```



3. Compare las variables cuantitativas entre pacientes con o sin diabetes segun tendencia central y variabilidad

```
library(arsenal)
```

```
## Warning: package 'arsenal' was built under R version 4.2.3
```

```
tab1 <- tableby(result~glucose + insulin +pregnant+triceps+age+bmi+diabetes_gen, data =pima)
summary(tab1)
```

```
##
##
## |                           |    no (N=500)     |    si (N=268)     |   Total (N=768)   | p value|
## |:--------------------------|:-----------------:|:-----------------:|:-----------------:|-------:|
## |**glucose**                |                   |                   |                   | < 0.001|
## |   N-Miss   |         3         |         2         |         5         |        |
## |   Mean (SD)| 110.644 (24.777)  | 142.320 (29.599)  | 121.687 (30.536)  |        |
## |   Range    | 44.000 - 197.000  | 78.000 - 199.000  | 44.000 - 199.000  |        |
## |**insulin**                |                   |                   |                   | < 0.001|
## |   N-Miss   |        236        |        138        |        374        |        |
## |   Mean (SD)| 130.288 (102.482) | 206.846 (132.700) | 155.548 (118.776) |        |
## |   Range    | 15.000 - 744.000  | 14.000 - 846.000  | 14.000 - 846.000  |        |
## |**pregnant**               |                   |                   |                   | < 0.001|
## |   Mean (SD)|   3.298 (3.017)   |   4.866 (3.741)   |   3.845 (3.370)   |        |
## |   Range    |  0.000 - 13.000   |  0.000 - 17.000   |  0.000 - 17.000   |        |
```

```
## |**triceps**                |                  |                  |                  | < 0.001|
## |   N-Miss    |       139        |       89         |       228        |        |
## |   Mean (SD) | 27.235 (10.026)  | 32.631 (9.091)   | 29.024 (10.045)  |        |
## |   Range     | 7.000 - 60.000   | 7.000 - 63.000   | 7.000 - 63.000   |        |
## |**age**                    |                  |                  |                  | < 0.001|
## |   Mean (SD) | 31.190 (11.668)  | 37.067 (10.968)  | 33.241 (11.760)  |        |
## |   Range     | 21.000 - 81.000  | 21.000 - 70.000  | 21.000 - 81.000  |        |
## |**bmi**                    |                  |                  |                  | < 0.001|
## |   N-Miss    |        9         |        2         |       11         |        |
## |   Mean (SD) | 30.860 (6.561)   | 35.407 (6.615)   | 32.457 (6.925)   |        |
## |   Range     | 18.200 - 57.300  | 22.900 - 67.100  | 18.200 - 67.100  |        |
## |**diabetes_gen**           |                  |                  |                  | < 0.001|
## |   Mean (SD) |  0.430 (0.299)   |  0.550 (0.372)   |  0.472 (0.331)   |        |
## |   Range     |  0.078 - 2.329   |  0.088 - 2.420   |  0.078 - 2.420   |        |
```

En las pacientes diabeticas la glucosa, la insulina, edad, bmi, ???, son en promedio significativamt mayores que en las no diabeticas