Proyecto II Análisis exploratorio

Valentina Yañez

2025-06-22

Objetivo General

Aplicar técnicas de análisis exploratorio de datos utilizando la librería ggplot2 en R para explorar visualmente las características de la base de datos Wisconsin Breast Cancer Dataset, identificar patrones, relaciones y posibles agrupaciones dentro de los datos.

Carga y revisión inicial de los datos

- Leer la base de datos en R
- Revisar estructura(str()), dimesiones(dim()), y valores faltantes (summary(), anyNA()).

```
# cargar librerias necesarias

# cargar los datos (ubicada en ./data/wdbc.data)
data <- read.csv("./data/wdbc.data", header = FALSE)
```

```
# generar encabezado de los datos
# Atributos base
features <- c("radius", "texture", "perimeter", "area", "smoothness",
              "compactness", "concavity", "concave_points", "symmetry", "fractal_dimension")

# Generar nombres por bloques
mean_names <- paste(features, "mean", sep = "_")
se_names <- paste(features, "se", sep = "_")
worst_names <- paste(features, "worst", sep = "_")

# Todos los nombres completos
column_names <- c("id", "diagnosis", mean_names, se_names, worst_names)

# Asignar nombres a las columnas
colnames(data) <- column_names

# Revisar la estructura de los datos
str(data)
```

```
## 'data.frame':    569 obs. of  32 variables:
##  $ id                     : int  842302 842517 84300903 84348301 84358402 843786 844359 84458202 844
##  $ diagnosis              : chr  "M" "M" "M" "M" ...
##  $ radius_mean            : num  18 20.6 19.7 11.4 20.3 ...
##  $ texture_mean           : num  10.4 17.8 21.2 20.4 14.3 ...
##  $ perimeter_mean         : num  122.8 132.9 130 77.6 135.1 ...
##  $ area_mean              : num  1001 1326 1203 386 1297 ...
##  $ smoothness_mean        : num  0.1184 0.0847 0.1096 0.1425 0.1003 ...
```

```
##  $ compactness_mean       : num  0.2776 0.0786 0.1599 0.2839 0.1328 ...
##  $ concavity_mean         : num  0.3001 0.0869 0.1974 0.2414 0.198 ...
##  $ concave_points_mean    : num  0.1471 0.0702 0.1279 0.1052 0.1043 ...
##  $ symmetry_mean          : num  0.242 0.181 0.207 0.26 0.181 ...
##  $ fractal_dimension_mean : num  0.0787 0.0567 0.06 0.0974 0.0588 ...
##  $ radius_se              : num  1.095 0.543 0.746 0.496 0.757 ...
##  $ texture_se             : num  0.905 0.734 0.787 1.156 0.781 ...
##  $ perimeter_se           : num  8.59 3.4 4.58 3.44 5.44 ...
##  $ area_se                : num  153.4 74.1 94 27.2 94.4 ...
##  $ smoothness_se          : num  0.0064 0.00522 0.00615 0.00911 0.01149 ...
##  $ compactness_se         : num  0.049 0.0131 0.0401 0.0746 0.0246 ...
##  $ concavity_se           : num  0.0537 0.0186 0.0383 0.0566 0.0569 ...
##  $ concave_points_se      : num  0.0159 0.0134 0.0206 0.0187 0.0188 ...
##  $ symmetry_se            : num  0.03 0.0139 0.0225 0.0596 0.0176 ...
##  $ fractal_dimension_se   : num  0.00619 0.00353 0.00457 0.00921 0.00511 ...
##  $ radius_worst           : num  25.4 25 23.6 14.9 22.5 ...
##  $ texture_worst          : num  17.3 23.4 25.5 26.5 16.7 ...
##  $ perimeter_worst        : num  184.6 158.8 152.5 98.9 152.2 ...
##  $ area_worst             : num  2019 1956 1709 568 1575 ...
##  $ smoothness_worst       : num  0.162 0.124 0.144 0.21 0.137 ...
##  $ compactness_worst      : num  0.666 0.187 0.424 0.866 0.205 ...
##  $ concavity_worst        : num  0.712 0.242 0.45 0.687 0.4 ...
##  $ concave_points_worst   : num  0.265 0.186 0.243 0.258 0.163 ...
##  $ symmetry_worst         : num  0.46 0.275 0.361 0.664 0.236 ...
##  $ fractal_dimension_worst: num  0.1189 0.089 0.0876 0.173 0.0768 ...
```

```
# revisar dimensiones
dim(data)
```

```
## [1] 569  32
```

```
# revisar valores faltantes
summary(data)
```

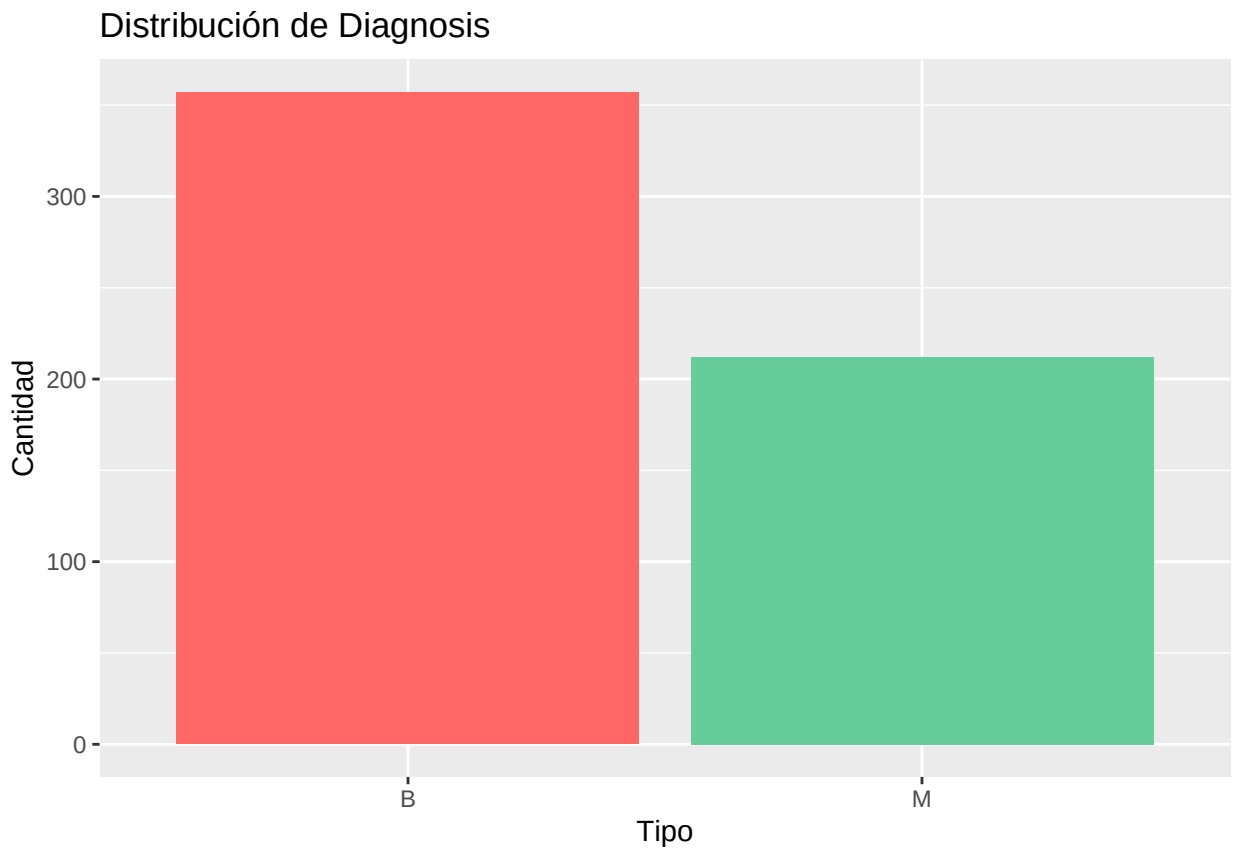
```
##        id              diagnosis          radius_mean      texture_mean
##  Min.   :     8670   Length:569         Min.   : 6.981   Min.   : 9.71
##  1st Qu.:   869218   Class :character   1st Qu.:11.700   1st Qu.:16.17
##  Median :   906024   Mode  :character   Median :13.370   Median :18.84
##  Mean   : 30371831                      Mean   :14.127   Mean   :19.29
##  3rd Qu.:  8813129                      3rd Qu.:15.780   3rd Qu.:21.80
##  Max.   :911320502                      Max.   :28.110   Max.   :39.28
##  perimeter_mean     area_mean      smoothness_mean   compactness_mean
##  Min.   : 43.79   Min.   : 143.5   Min.   :0.05263   Min.   :0.01938
##  1st Qu.: 75.17   1st Qu.: 420.3   1st Qu.:0.08637   1st Qu.:0.06492
##  Median : 86.24   Median : 551.1   Median :0.09587   Median :0.09263
##  Mean   : 91.97   Mean   : 654.9   Mean   :0.09636   Mean   :0.10434
##  3rd Qu.:104.10   3rd Qu.: 782.7   3rd Qu.:0.10530   3rd Qu.:0.13040
##  Max.   :188.50   Max.   :2501.0   Max.   :0.16340   Max.   :0.34540
##  concavity_mean    concave_points_mean symmetry_mean    fractal_dimension_mean
##  Min.   :0.00000   Min.   :0.00000     Min.   :0.1060   Min.   :0.04996
##  1st Qu.:0.02956   1st Qu.:0.02031     1st Qu.:0.1619   1st Qu.:0.05770
##  Median :0.06154   Median :0.03350     Median :0.1792   Median :0.06154
##  Mean   :0.08880   Mean   :0.04892     Mean   :0.1812   Mean   :0.06280
##  3rd Qu.:0.13070   3rd Qu.:0.07400     3rd Qu.:0.1957   3rd Qu.:0.06612
##  Max.   :0.42680   Max.   :0.20120     Max.   :0.3040   Max.   :0.09744
##    radius_se        texture_se      perimeter_se       area_se
```

```
##  Min.   :0.1115   Min.   :0.3602   Min.   : 0.757   Min.   :  6.802
##  1st Qu.:0.2324   1st Qu.:0.8339   1st Qu.: 1.606   1st Qu.: 17.850
##  Median :0.3242   Median :1.1080   Median : 2.287   Median : 24.530
##  Mean   :0.4052   Mean   :1.2169   Mean   : 2.866   Mean   : 40.337
##  3rd Qu.:0.4789   3rd Qu.:1.4740   3rd Qu.: 3.357   3rd Qu.: 45.190
##  Max.   :2.8730   Max.   :4.8850   Max.   :21.980   Max.   :542.200
##  smoothness_se      compactness_se      concavity_se     concave_points_se
##  Min.   :0.001713   Min.   :0.002252   Min.   :0.00000   Min.   :0.000000
##  1st Qu.:0.005169   1st Qu.:0.013080   1st Qu.:0.01509   1st Qu.:0.007638
##  Median :0.006380   Median :0.020450   Median :0.02589   Median :0.010930
##  Mean   :0.007041   Mean   :0.025478   Mean   :0.03189   Mean   :0.011796
##  3rd Qu.:0.008146   3rd Qu.:0.032450   3rd Qu.:0.04205   3rd Qu.:0.014710
##  Max.   :0.031130   Max.   :0.135400   Max.   :0.39600   Max.   :0.052790
##   symmetry_se       fractal_dimension_se  radius_worst   texture_worst
##  Min.   :0.007882   Min.   :0.0008948    Min.   : 7.93   Min.   :12.02
##  1st Qu.:0.015160   1st Qu.:0.0022480    1st Qu.:13.01   1st Qu.:21.08
##  Median :0.018730   Median :0.0031870    Median :14.97   Median :25.41
##  Mean   :0.020542   Mean   :0.0037949    Mean   :16.27   Mean   :25.68
##  3rd Qu.:0.023480   3rd Qu.:0.0045580    3rd Qu.:18.79   3rd Qu.:29.72
##  Max.   :0.078950   Max.   :0.0298400    Max.   :36.04   Max.   :49.54
##  perimeter_worst    area_worst     smoothness_worst  compactness_worst
##  Min.   : 50.41   Min.   : 185.2   Min.   :0.07117   Min.   :0.02729
##  1st Qu.: 84.11   1st Qu.: 515.3   1st Qu.:0.11660   1st Qu.:0.14720
##  Median : 97.66   Median : 686.5   Median :0.13130   Median :0.21190
##  Mean   :107.26   Mean   : 880.6   Mean   :0.13237   Mean   :0.25427
##  3rd Qu.:125.40   3rd Qu.:1084.0   3rd Qu.:0.14600   3rd Qu.:0.33910
##  Max.   :251.20   Max.   :4254.0   Max.   :0.22260   Max.   :1.05800
##  concavity_worst  concave_points_worst symmetry_worst   fractal_dimension_worst
##  Min.   :0.0000   Min.   :0.00000      Min.   :0.1565   Min.   :0.05504
##  1st Qu.:0.1145   1st Qu.:0.06493      1st Qu.:0.2504   1st Qu.:0.07146
##  Median :0.2267   Median :0.09993      Median :0.2822   Median :0.08004
##  Mean   :0.2722   Mean   :0.11461      Mean   :0.2901   Mean   :0.08395
##  3rd Qu.:0.3829   3rd Qu.:0.16140      3rd Qu.:0.3179   3rd Qu.:0.09208
##  Max.   :1.2520   Max.   :0.29100      Max.   :0.6638   Max.   :0.20750
```

Análisis univariado con ggplot2

- Generar histogramas y boxplots de al menos 4 variables numéricas.
- Generar gráfico de barras para la variable categórica diagnosis.
- Comentar la distribución de las variables.

```r
#install.packages("ggplot2")

# graficar diagnosis
library(ggplot2)
ggplot(data, aes(x = diagnosis)) +
  geom_bar(fill = c("#FF6666", "#66CC99")) +
  labs(title = "Distribución de Diagnosis", x = "Tipo", y = "Cantidad")
```




```
#install.packages(c("ggplot2", "FactoMineR", "factoextra", "dplyr"))
library(ggplot2)
library(FactoMineR)
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(dplyr)
```

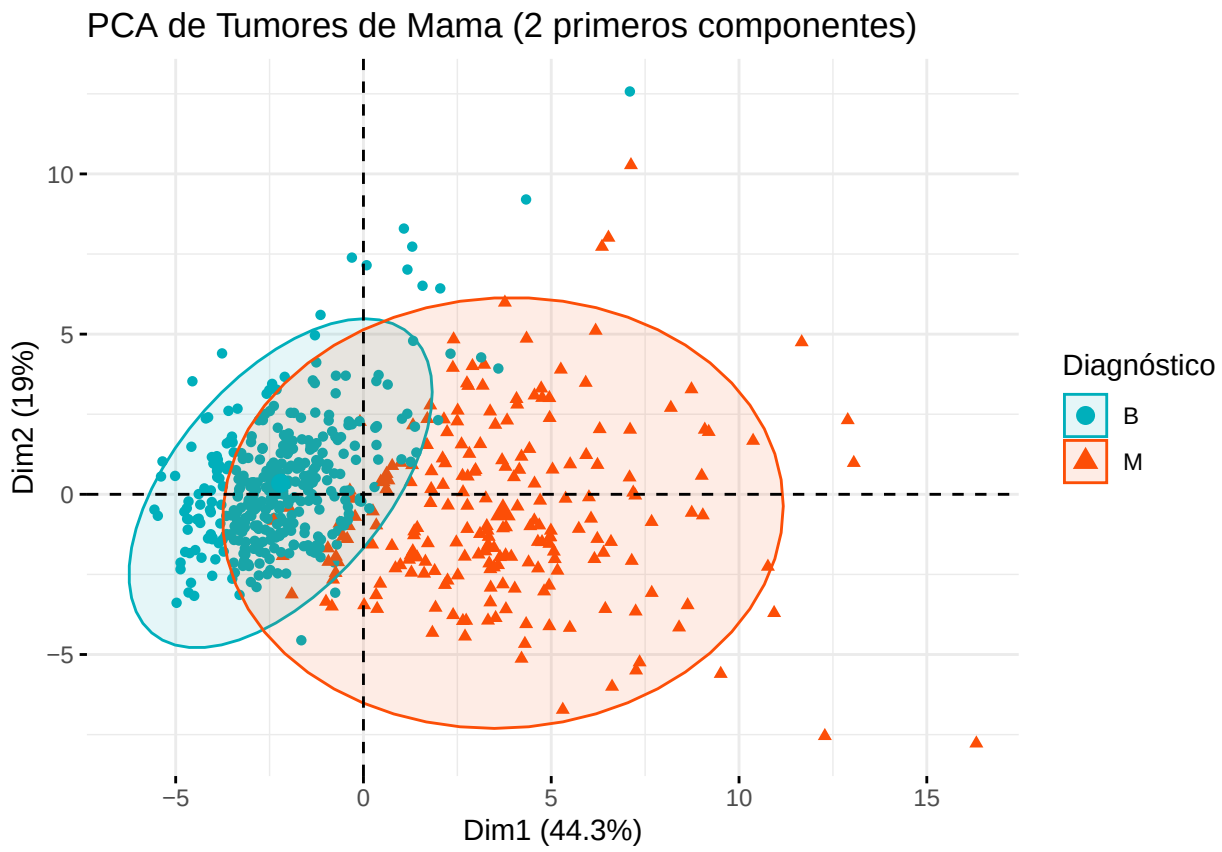
```
##
## Adjuntando el paquete: 'dplyr'
```

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

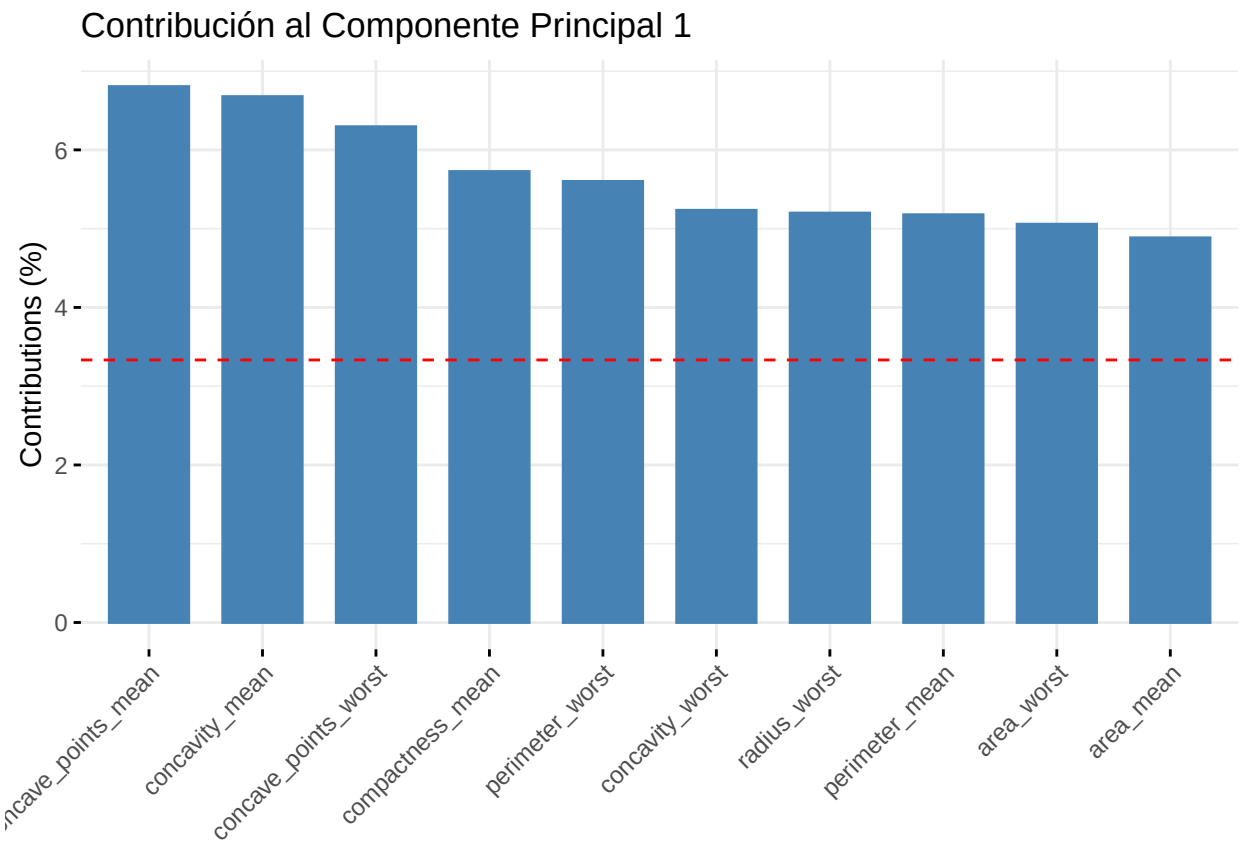
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Convertir diagnosis a factor
data$diagnosis <- as.factor(data$diagnosis)

# Seleccionar solo variables numéricas
data_numeric <- data[, 3:ncol(data)]  # Quita id y diagnosis

# Ejecutar PCA
pca_result <- PCA(data_numeric, graph = FALSE)
```

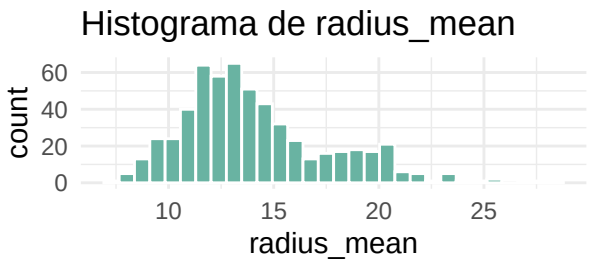
```r
# Visualización con color por diagnóstico
fviz_pca_ind(pca_result,
             geom.ind = "point",
             col.ind = data$diagnosis,
             palette = c("#00AFBB", "#FC4E07"),
             addEllipses = TRUE,
             legend.title = "Diagnóstico") +
  ggtitle("PCA de Tumores de Mama (2 primeros componentes)")
```

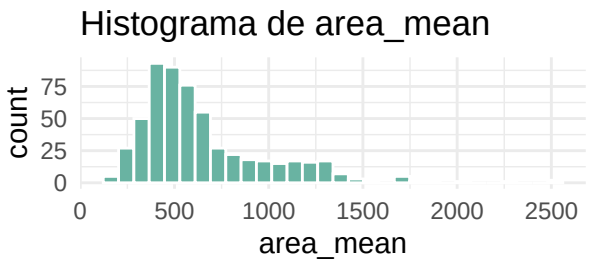


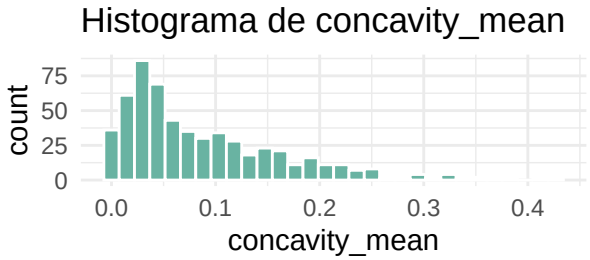

```r
# Ver cómo contribuye cada variable al primer componente
fviz_contrib(pca_result, choice = "var", axes = 1, top = 10) +
  ggtitle("Contribución al Componente Principal 1")
```

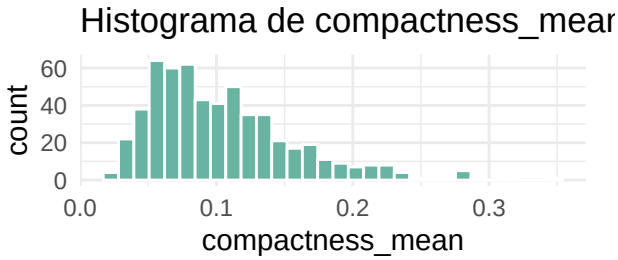


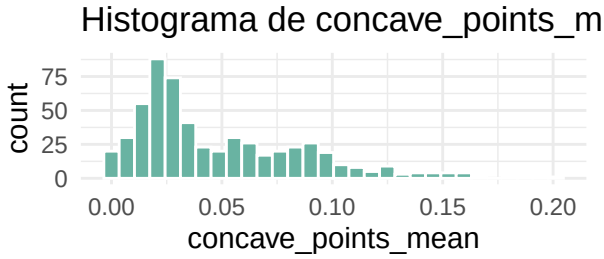

```r
# Histogramas de variables numéricas

#install.packages("patchwork")
library(ggplot2)
library(patchwork)

features <- c("radius_mean", "area_mean", "concavity_mean", "compactness_mean", "concave_points_mean",

plots <- list()

for (f in features) {
  plots[[f]] <- ggplot(data, aes(x = .data[[f]])) +
    geom_histogram(bins = 30, fill = "#69b3a2", color = "white") +
    labs(title = paste("Histograma de", f)) +
    theme_minimal()
}

# Combinar todos los gráficos
wrap_plots(plots, ncol = 2)  # 2 columnas
```

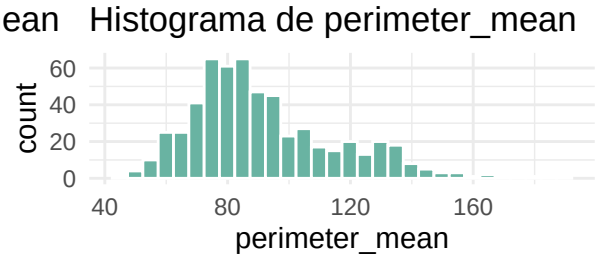




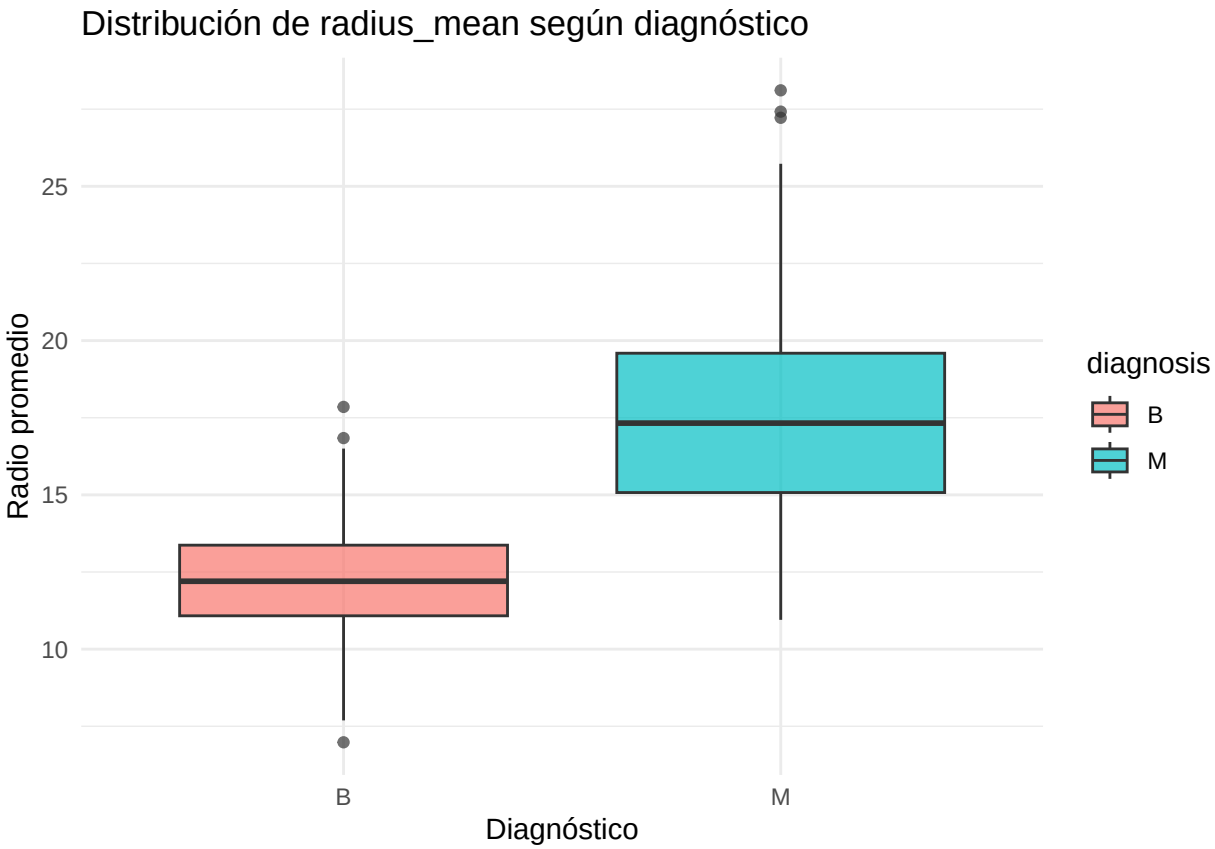




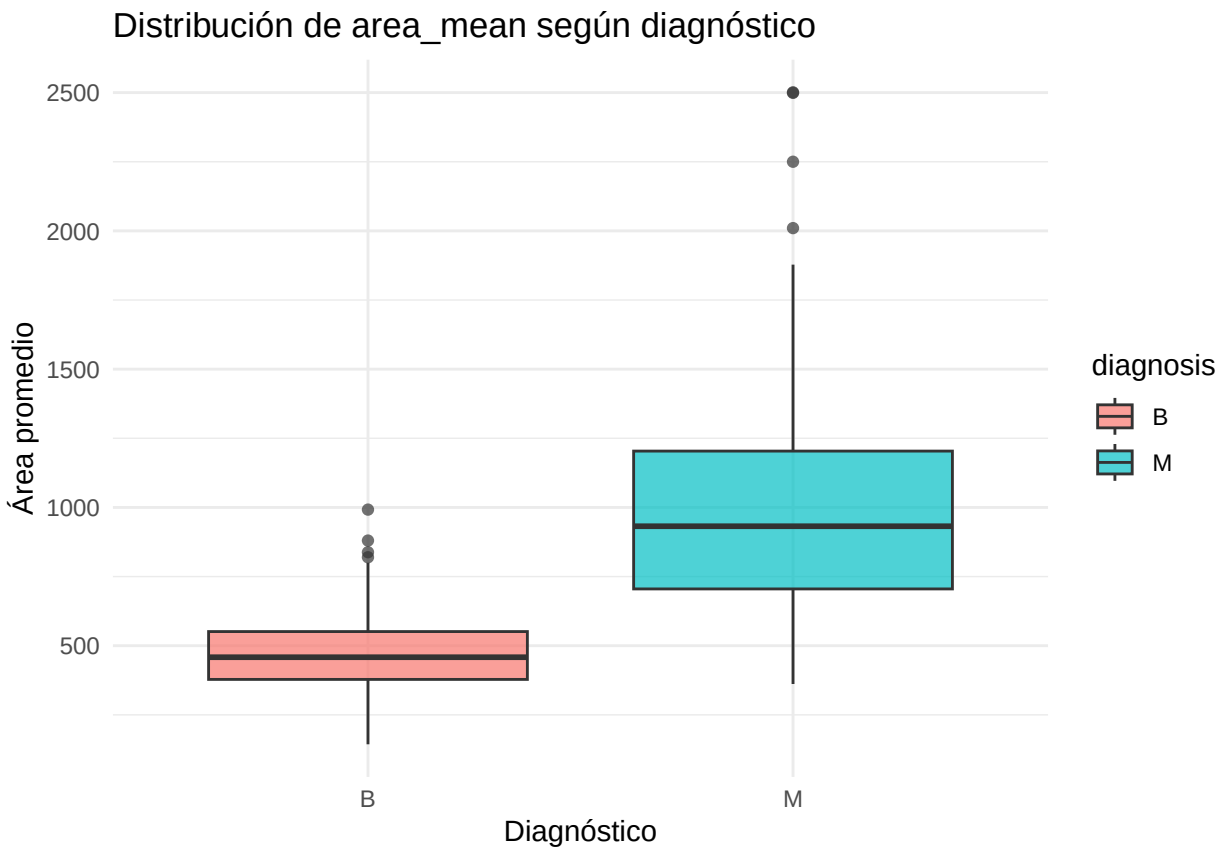




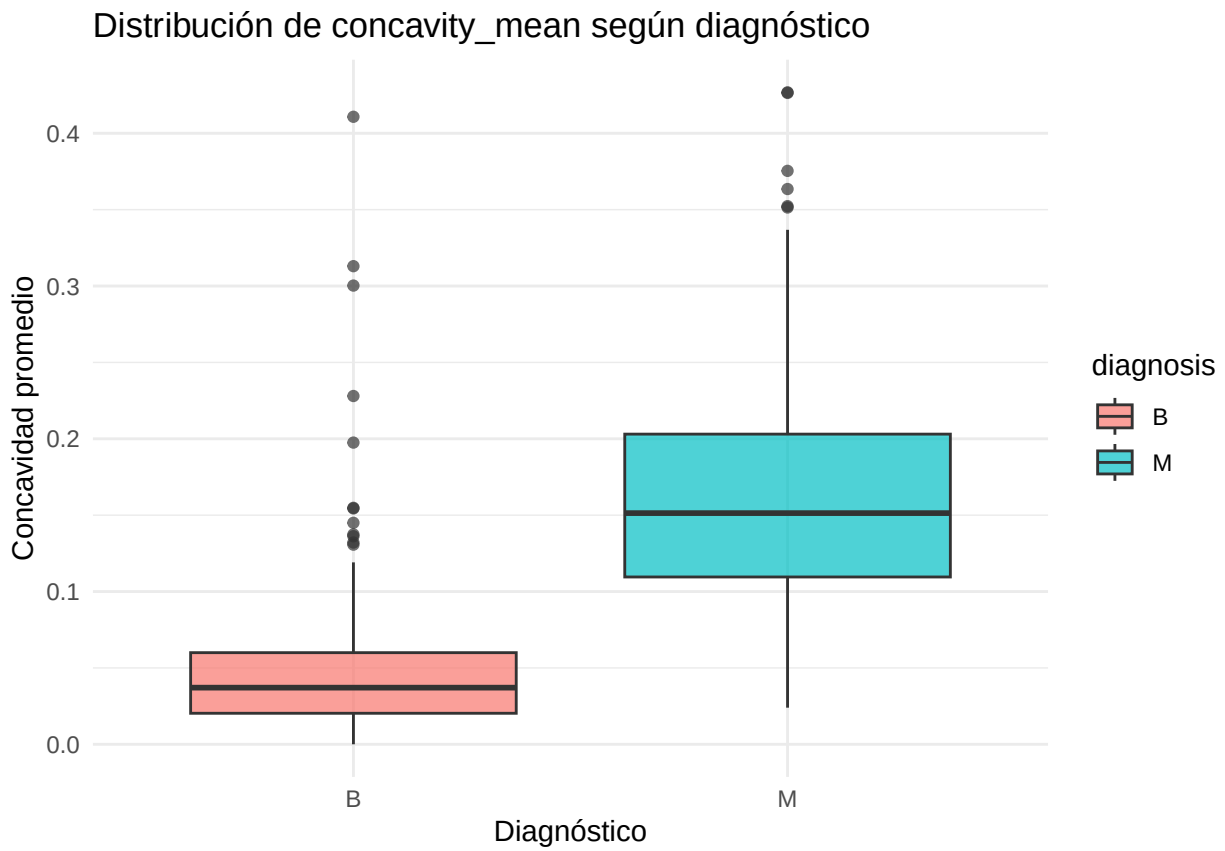

```r
# Boxplots de variables numéricas por diagnóstico
ggplot(data, aes(x = diagnosis, y = radius_mean, fill = diagnosis)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "Distribución de radius_mean según diagnóstico",
       x = "Diagnóstico", y = "Radio promedio") +
  theme_minimal()
```



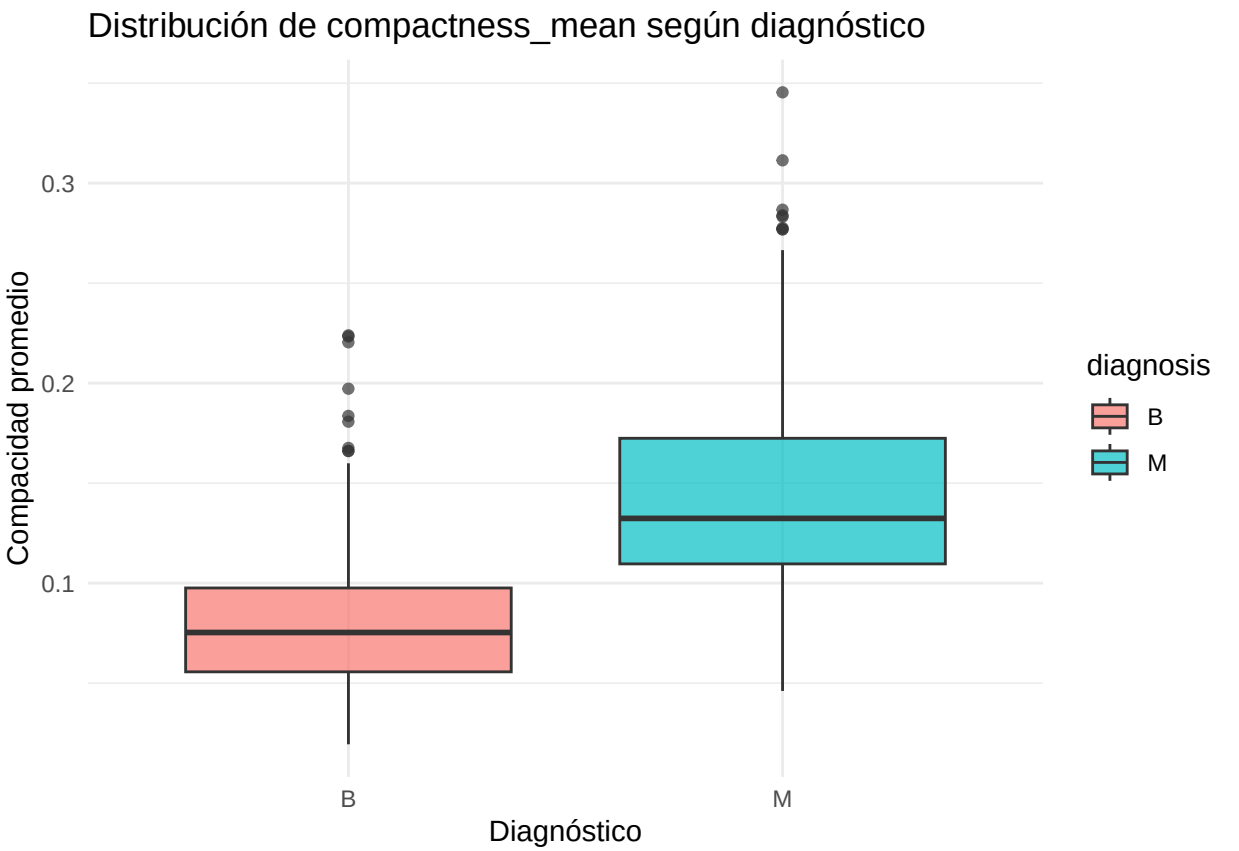

```
ggplot(data, aes(x = diagnosis, y = area_mean, fill = diagnosis)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "Distribución de area_mean según diagnóstico",
       x = "Diagnóstico", y = "Área promedio") +
  theme_minimal()
```




```
ggplot(data, aes(x = diagnosis, y = concavity_mean, fill = diagnosis)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "Distribución de concavity_mean según diagnóstico",
       x = "Diagnóstico", y = "Concavidad promedio") +
  theme_minimal()
```




```r
ggplot(data, aes(x = diagnosis, y = compactness_mean, fill = diagnosis)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "Distribución de compactness_mean según diagnóstico",
       x = "Diagnóstico", y = "Compacidad promedio") +
  theme_minimal()
```



Análisis de histogramas

La variable radius_mean, que representa el radio promedio del tumor, mostró una distribución asimétrica a la derecha. La mayoría de los valores se concentran entre 10 y 15 unidades, lo que sugiere que la mayor parte de los tumores analizados tienen un tamaño moderado. Sin embargo, se observan algunos valores más altos, cercanos a 25, que podrían estar asociados a tumores de mayor tamaño y, posiblemente, de carácter maligno.

area_mean mide el área promedio del tumor, presentó una distribución similar, pero con una gran concentración de valores por debajo de las 1000 unidades mas o menos, tiene amplia dispersión, lo que refuerza su importancia como atributo diferenciador, aunque también sugiere la necesidad de normalización.

concavity_mean describe el grado de concavidad de los bordes del tumor, tiene alta concentración de valores próximos a cero. Común en tumores benignos al parecer. Un subconjunto más reducido presenta valores más altos, lo que reflejaría la morfología más irregular asociada a tumores malignos.

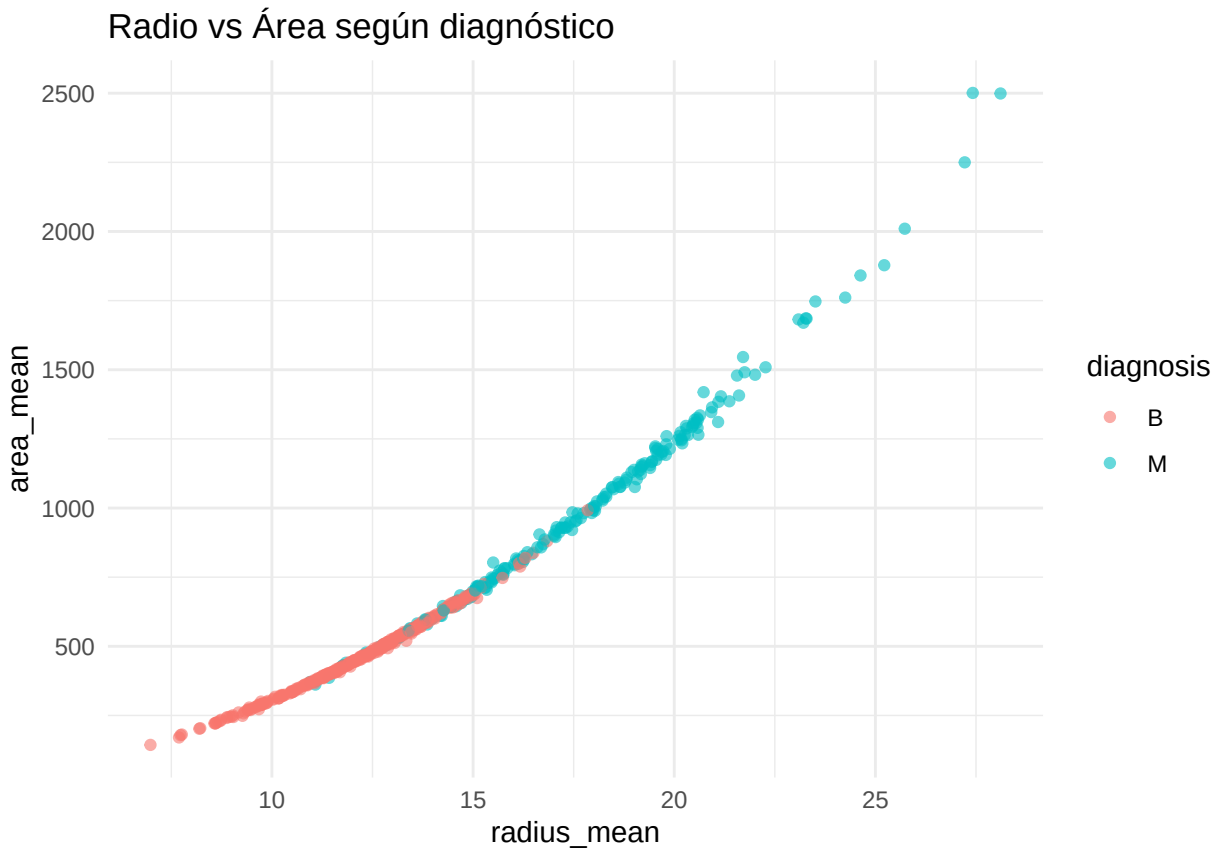
Y en final compactness_mean también tiene una distribución sesgada de forma positiva, con la mayoría de los datos situados entre 0.05 y 0.15. Este comportamiento sugiere que la mayoría de los tumores son relativamente compactos en su forma, aunque también existen algunos casos más dispersos o irregulares.

Análisis bivariado

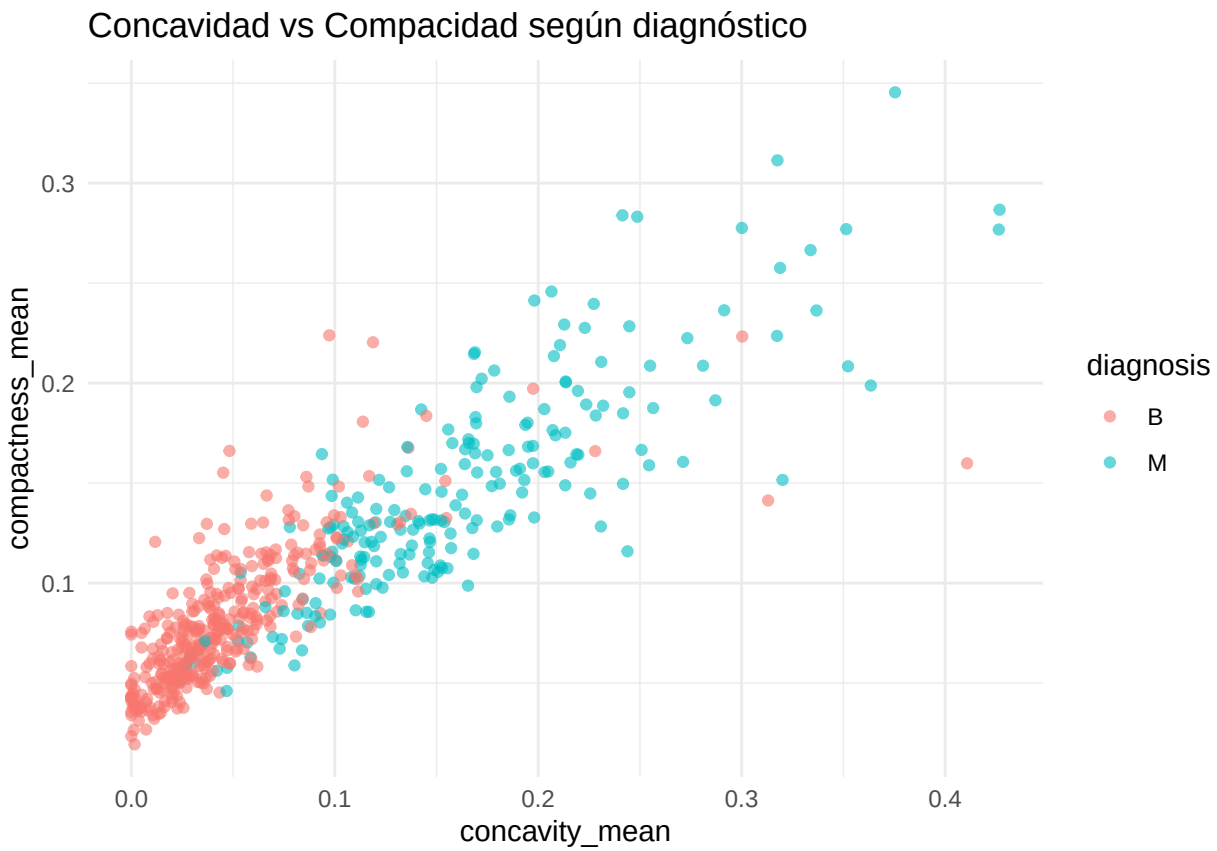
- Crear diagramas de dispersión (scatter plots) entre pares de variables, diferenciando por diagnóstico (usando color = diagnosis).
- Evaluar la posible correlación entre algunas variables.

```
library(ggplot2)

ggplot(data, aes(x = radius_mean, y = area_mean, color = diagnosis)) +
```

```
  geom_point(alpha = 0.6) +
  labs(title = "Radio vs Área según diagnóstico") +
  theme_minimal()
```




```
ggplot(data, aes(x = concavity_mean, y = compactness_mean, color = diagnosis)) +
  geom_point(alpha = 0.6) +
  labs(title = "Concavidad vs Compacidad según diagnóstico") +
  theme_minimal()
```



Matrices de correlación

- Calcular la matriz de correlaci´on.
- Visualizarla con corrplot o ggcorrplot.
- Identificar las variables m´as correlacionadas.

```
cor(data$radius_mean, data$area_mean)
```

```
## [1] 0.9873572
```

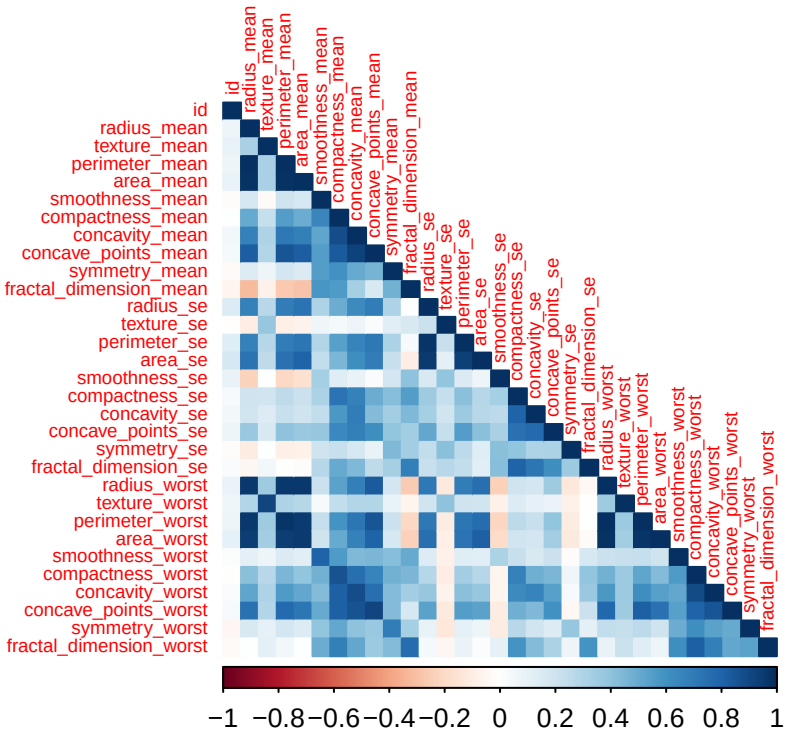
```
cor(data$concavity_mean, data$compactness_mean)
```

```
## [1] 0.8831207
```

```
#install.packages("corrplot")
library(corrplot)
```

```
## corrplot 0.95 loaded
```

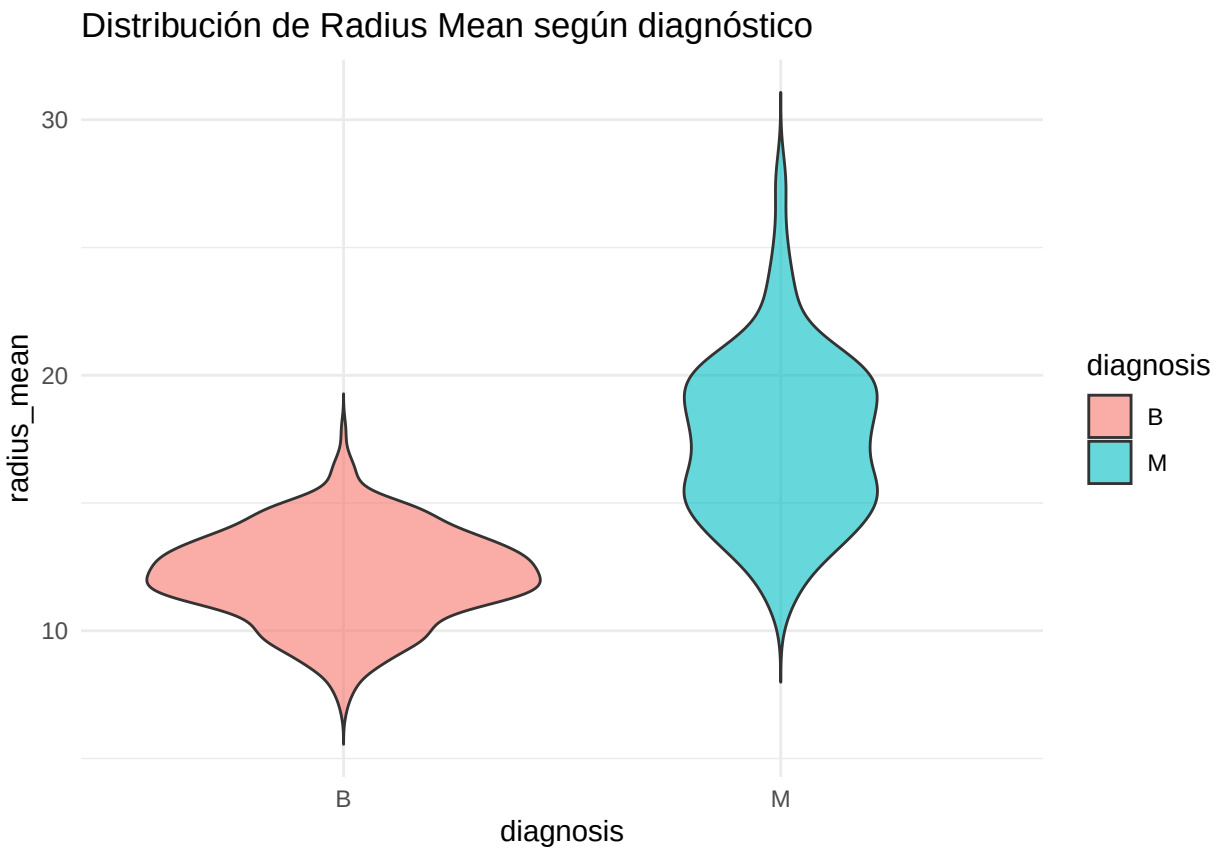
```
# Eliminar columnas no numéricas
numeric_data <- data[sapply(data, is.numeric)]
corr_matrix <- cor(numeric_data)

corrplot(corr_matrix, method = "color", type = "lower", tl.cex = 0.6)
```



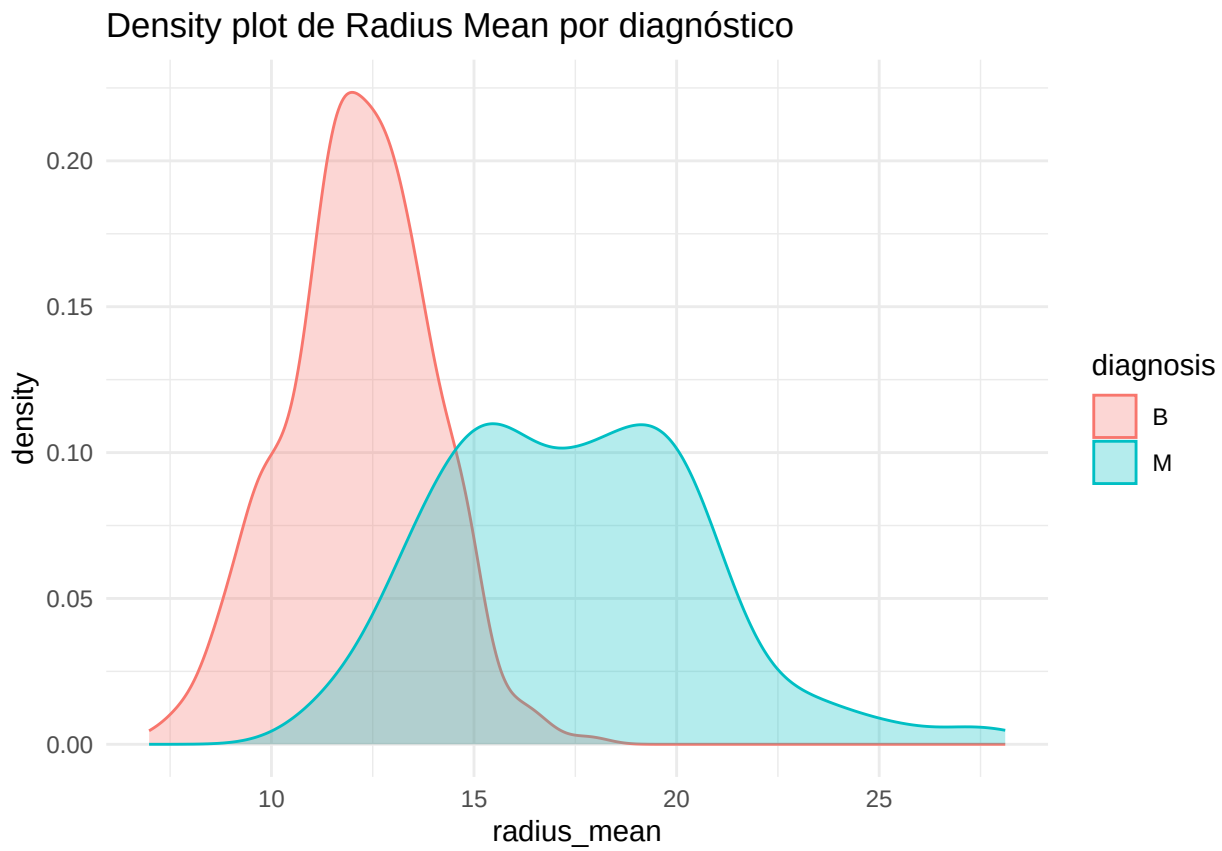
Gráficos avanzado

- Crear violin plots o density plots para comparar las distribuciones según diagnóstico.
- Utilizar gráficos facetados para explorar más de una variable simultáneamente.

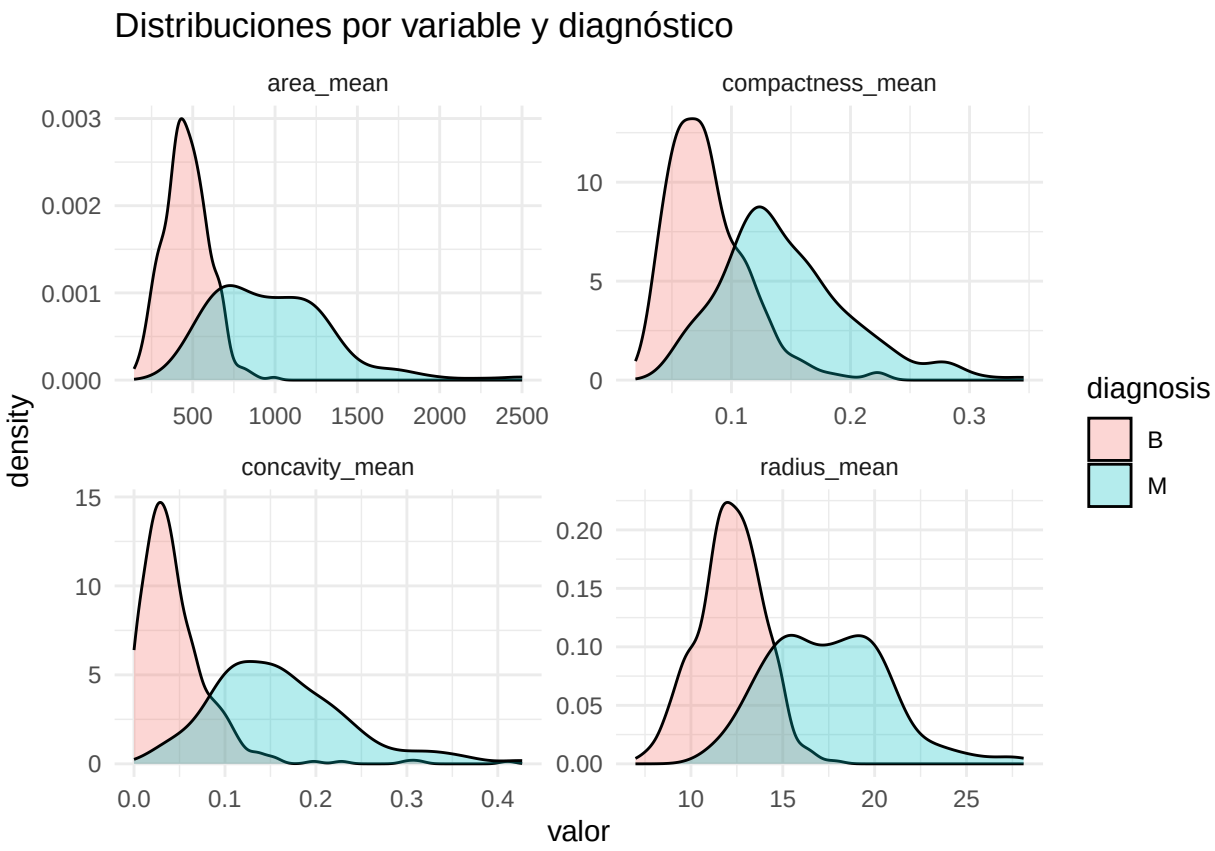
```
library(ggplot2)

ggplot(data, aes(x = diagnosis, y = radius_mean, fill = diagnosis)) +
  geom_violin(trim = FALSE, alpha = 0.6) +
  labs(title = "Distribución de Radius Mean según diagnóstico") +
  theme_minimal()
```




```
ggplot(data, aes(x = radius_mean, color = diagnosis, fill = diagnosis)) +
  geom_density(alpha = 0.3) +
  labs(title = "Density plot de Radius Mean por diagnóstico") +
  theme_minimal()
```




```r
# Variables de interés
vars <- c("radius_mean", "area_mean", "concavity_mean", "compactness_mean")

# Reorganizar el dataframe a formato largo (long format)
#install.packages("tidyr")
library(tidyr)
long_data <- pivot_longer(data, cols = all_of(vars), names_to = "variable", values_to = "valor")

# Density plot facetado
ggplot(long_data, aes(x = valor, fill = diagnosis)) +
  geom_density(alpha = 0.3) +
  facet_wrap(~variable, scales = "free") +
  labs(title = "Distribuciones por variable y diagnóstico") +
  theme_minimal()
```



Conclusiones