Análisis de datos

HEART ATTACK

María Nieto

índice

Introducción

Los ataques cardíacos representan una amenaza significativa para la salud cardiovascular global, siendo una de las principales causas de morbimortalidad en todo el mundo. El análisis de datos relacionados con la predicción y el estudio de ataques cardíacos se ha convertido en una herramienta esencial para comprender los factores de riesgo, identificar patrones subyacentes y desarrollar estrategias preventivas efectivas. En este contexto, el presente trabajo se centra en un exhaustivo análisis de datos que abarca variables clínicas, de estilo de vida y biomarcadores relevantes, con el objetivo de discernir patrones y relaciones que puedan contribuir a la predicción y comprensión de eventos cardíacos adversos.

La relevancia de este análisis radica en la necesidad de avanzar en la capacidad predictiva de los modelos existentes, así como en la identificación de nuevas variables clave que puedan mejorar la precisión en la evaluación del riesgo cardiovascular. Al abordar este conjunto de datos, buscamos no solo predecir eventos futuros, sino también profundizar en la comprensión de las complejas interacciones entre diversos factores que contribuyen al desarrollo de ataques cardíacos.

A través de técnicas avanzadas de análisis estadístico y modelado predictivo, pretendemos identificar patrones de riesgo, evaluar la relevancia de variables específicas y proporcionar una base sólida para estrategias de intervención personalizadas. La aplicación de enfoques analíticos avanzados no solo permite identificar relaciones lineales entre variables, sino también descubrir patrones no lineales y complejas interconexiones que pueden ser cruciales para una evaluación precisa del riesgo cardiovascular.

En resumen, este trabajo se propone contribuir al avance de la investigación en la predicción y análisis de ataques cardíacos, utilizando un enfoque integral que aprovecha las capacidades de análisis de datos avanzadas para mejorar la comprensión de los factores de riesgo y, en última instancia, promover la salud cardiovascular y la prevención de enfermedades.

Sobre los Datos

Estas son las categorías que hay en el conjunto de datos:

- Age : Age of the patient
- Sex : Sex of the patient
- cp : Chest Pain type:
  - Value 0: typical angina
  - Value 1: atypical angina
  - Value 2: non-anginal pain
  - Value 3: asymptomatic
- trtbps : resting blood pressure (in mm Hg)
- chol: cholesterol in mg/dl fetched via BMI sensor
- fbs: fasting blood sugar > 120 mg/dl:
  - 1 = true
  - 0 = false
- rest_ecg: resting electrocardiographic results:
  - Value 0: normal
  - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
  - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- thalach: maximum heart rate achieved
- exang: exercise induced angina:
  - 1 = yes
  - 0 = no
- old peak: ST depression induced by exercise relative to rest
- slp: the slope of the peak exercise ST segment:

  - 0 = unsloping
  - 1 = flat
  - 2 = downsloping
- caa: number of major vessels (0-3)
- thall : thalassemia:
  - 0 = null
  - 1 = fixed defect
  - 2 = normal
  - 3 = reversable defect
- output: diagnosis of heart disease (angiographic disease status):
  - 0: < 50% diameter narrowing. less chance of heart disease
  - 1: > 50% diameter narrowing. more chance of heart disease

La traducción de estos, sería:

- Edad: Edad del paciente.
- Sexo: Sexo del paciente.
- cp: tipo de dolor en el pecho:
  - Valor 0: angina típica
  - Valor 1: angina atípica
  - Valor 2: dolor no anginoso
  - Valor 3: asintomático
- trtbps: presión arterial en reposo (en mm Hg)
- chol: colesterol en mg/dl obtenido mediante el sensor de IMC
- fbs: azúcar en sangre en ayunas > 120 mg/dl:
  - 1 = verdadero

- 0 = falso

- rest_ecg: resultados electrocardiográficos en reposo:
  - Valor 0: normal
  - Valor 1: tener anomalía de la onda ST-T (inversiones de la onda T y/o elevación o depresión del ST > 0,05 mV)
  - Valor 2: muestra hipertrofia ventricular izquierda probable o definitiva según los criterios de Estes
- thalach: frecuencia cardíaca máxima alcanzada
- exang: angina inducida por el ejercicio:
  - 1 = si
  - 0 = no
- oldpeak: depresión del ST inducida por el ejercicio en relación con el reposo.
- slp: la pendiente del segmento ST del ejercicio máximo:
  - 0 = sin pendiente
  - 1 = plano
  - 2 = descendente
- caa: número de buques principales (0-3)
- thall: talasemia:
  - 0 = nulo
  - 1 = defecto arreglado
  - 2 = normales
  - 3 = defecto reversible

- Resultado: diagnóstico de enfermedad cardíaca (estado de enfermedad angiográfica):

- 0: < 50% de estrechamiento del diámetro. menos posibilidades de enfermedad cardíaca
- 1: > 50% de estrechamiento del diámetro. más posibilidades de sufrir enfermedades cardíacas

Análisis Descriptivo de Datos

La base de datos es la siguiente, y con la función nrow, nos indica el número de datos que contiene.

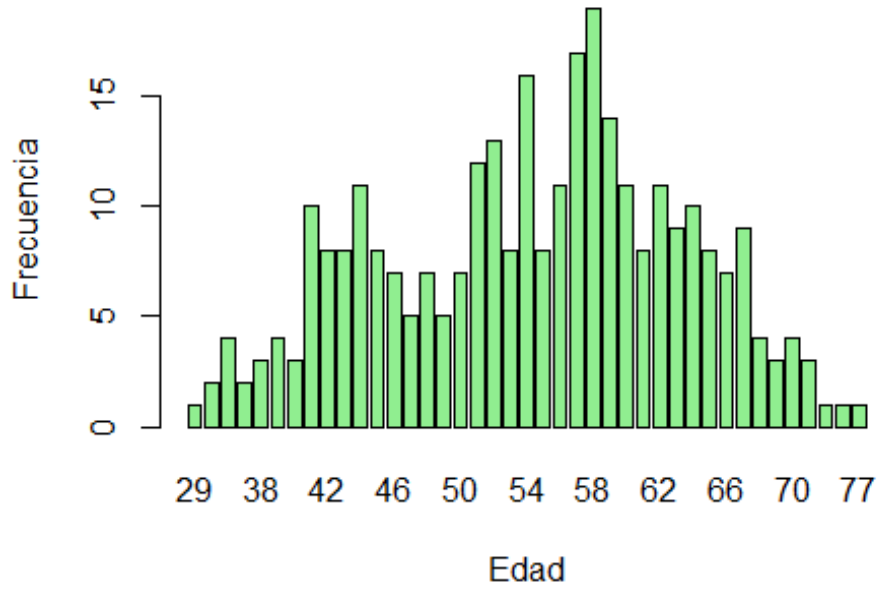
```
heart_heart_csv=read.csv("heart - heart.csv.csv",head=T,sep=",")
head(heart_heart_csv)
```

```
##    age      sex  cp  trtbps  chol  fbs  restecg  thalachh  exng  oldpeak  slp  caa
thall
## 1   63    Male   3     145   233    1        0       150     0      2.3    0   0
1
## 2   37    Male   2     130   250    0        1       187     0      3.5    0   0
2
## 3   41  Female   1     130   204    0        0       172     0      1.4    2   0
2
## 4   56    Male   1     120   236    0        1       178     0      0.8    2   0
2
## 5   57  Female   0     120   354    0        1       163     1      0.6    2   0
2
## 6   57    Male   0     140   192    0        1       148     0      0.4    1   0
1
##   output
## 1      1
## 2      1
## 3      1
## 4      1
## 5      1
## 6      1
```

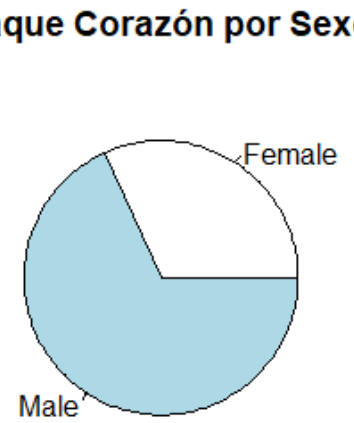
```
## [1] 303
```

nrow nos indica el número de datos que contiene la base de datos de los ataques al corazón (heart).

```
##
## 29 34 35 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58
59
##  1  2  4  2  3  4  3 10  8  8 11  8  7  5  7  5  7 12 13  8 16  8 11 17 19
14
## 60 61 62 63 64 65 66 67 68 69 70 71 74 76 77
## 11  8 11  9 10  8  7  9  4  3  4  3  1  1  1
```




```
##
## Female   Male
##     96    207
```



A la vista de todos estos datos y gráficos relativos a las variables cualitativas de la base de datos, podemos decir que el mayor números de casos de ataques cardíacos ocurren entre los 56 y los 60 años. Si lo desglosamos por gérero se comprueba que los hombres tienen más tendencias de sufrir estos infartos agudos de miocardio. No resulta completamente claro por qué los hombres enfrentan un mayor riesgo de desarrollar enfermedades cardíacas, pero en promedio, el riesgo cardiovascular de una mujer es equivalente al de un hombre que es 20 años mayor. Uno de los factores cruciales en esta disparidad se atribuye a las hormonas. Se presume que los estrógenos producidos por los ovarios tienen un efecto protector. De hecho, es a partir de la menopausia cuando el riesgo de enfermedades cardiovasculares en las mujeres aumenta significativamente.

Resumen de Datos Numéricos

En este apartado, se puede observar la información relativa de cada una de las variables que componen la base de datos.

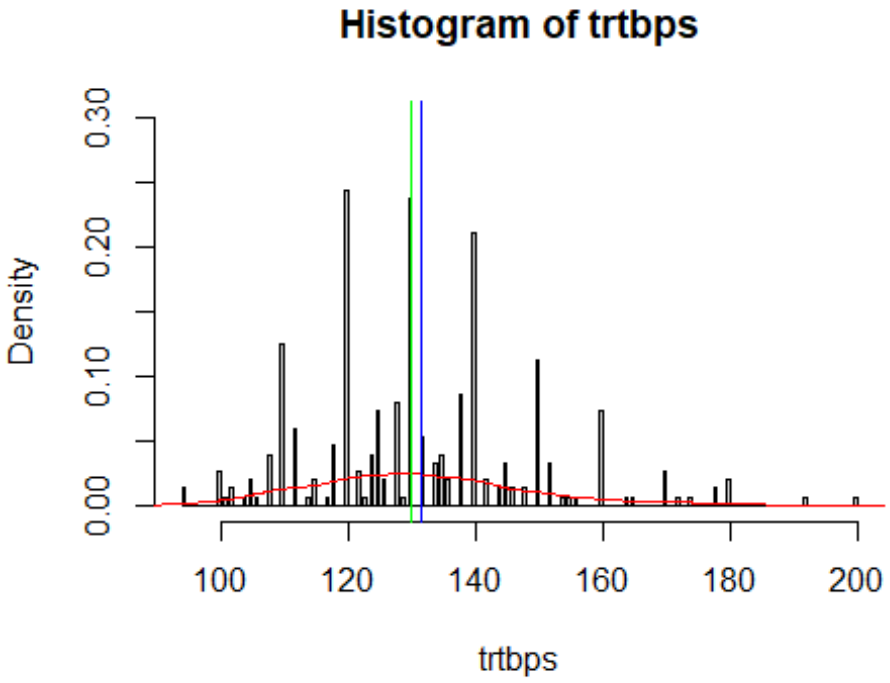
```
##       age            sex                 cp            trtbps     
##  Min.   :29.00   Length:303         Min.   :0.000   Min.   : 94.0  
##  1st Qu.:47.50   Class :character   1st Qu.:0.000   1st Qu.:120.0
```

```
##  Median :55.00   Mode  :character   Median :1.000   Median :130.0
##  Mean   :54.37                      Mean   :0.967   Mean   :131.6
##  3rd Qu.:61.00                      3rd Qu.:2.000   3rd Qu.:140.0
##  Max.   :77.00                      Max.   :3.000   Max.   :200.0
##       chol            fbs             restecg          thalachh
##  Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0
##  1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5
##  Median :240.0   Median :0.0000   Median :1.0000   Median :153.0
##  Mean   :246.3   Mean   :0.1485   Mean   :0.5281   Mean   :149.6
##  3rd Qu.:274.5   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:166.0
##  Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0
##       exng           oldpeak          slp              caa
##  Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.80   Median :1.000   Median :0.0000
##  Mean   :0.3267   Mean   :1.04   Mean   :1.399   Mean   :0.7294
##  3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :6.20   Max.   :2.000   Max.   :4.0000
##      thall           output
##  Min.   :0.000   Min.   :0.0000
##  1st Qu.:2.000   1st Qu.:0.0000
##  Median :2.000   Median :1.0000
##  Mean   :2.314   Mean   :0.5446
##  3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :3.000   Max.   :1.0000
```

A la vista de los resultados se puede apreciar como la media de edad se encuentra en 55 años. Además, el tipo de dolor de pecho que sufren los enfermos de media es de tipo 1, una angína típica. Sin embargo, exiten otros dos tipos. El tipo 2, que es una angina atípica; y, el tipo 3, un dolor no anginoso. Por otro lado, la presión arterial media en reposo (medido en mmHg al ingresar al hospital) es de 130.

Análisis de la Distribución de los Datos “trtbps”

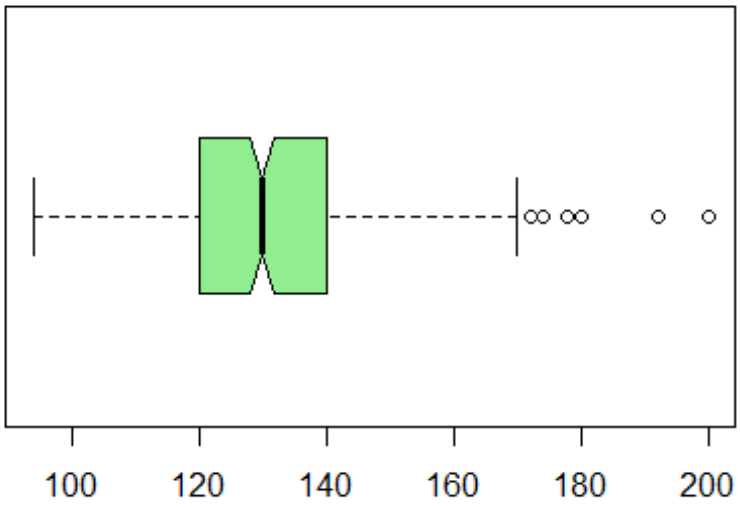
```
## Warning: package 'moments' was built under R version 4.1.3
```




```
## [1] 0.7102301
```

En este punto, analizamos la variable cuantitativa de la presión arterial en reposo y, de la que podemos decir, que estamos ante una función de densidad multimodal porque se observan varios picos locales, estando comprendidos entre 120 y 140, que nos indican donde se encuentra el centro de la distribución. Al calcular el coeficiente de asimetría, obtenemos un resultado de 0.7102301 lo que nos indica que la distribución tiene una asimetría hacia la derecha o positiva, puesto que este coeficiente es positivo o mayor que 0. También, se puede ver comparando la media y la mediana, ya que si la media toma un valor muy superior a la mediana, estamos ante un caso de asimetría positiva. En este caso, no existe una gran diferencia pues el valor de la media es de 131.6, y de el de la mediana es de 130.

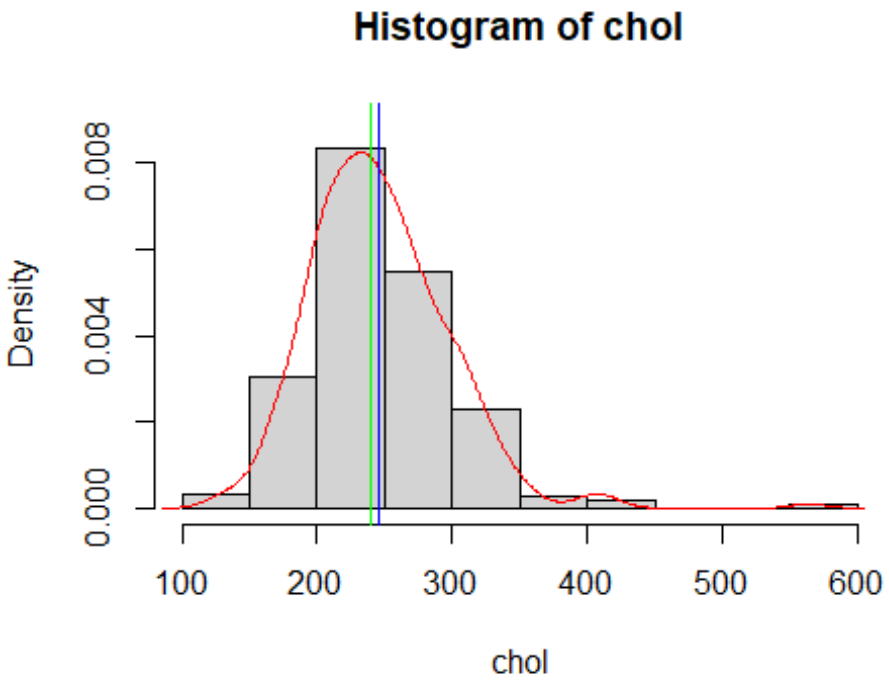
```
## The following objects are masked from heart_heart_csv (pos = 4):
##
##     age, caa, chol, cp, exng, fbs, oldpeak, output, restecg, sex, slp,
##     thalachh, thall, trtbps
```



Por otra parte, el diagrama de cajas o boxplot es similar al histograma pero no idéntico. Mientras que el histograma con su densidad es bueno para visualizar el centro, la dispersión, las colas y la forma de la distribución, no resulta útil a la hora de comparar distribuciones. Un gráfico que permite observar lo anterior y también comparar distribuciones en una misma figura es el boxplot.

Al analizar el boxplot podemos observar una gran concentración de valores en torno al rango entre 120 y 140. Esto confirma, como se mencionó anteriormente, que en la muestra hay una proporción significativa de pacientes con presión arterial en ese rango. Gracias a la disposición de estos datos, podemos identificar los límites de la caja, determinados por el rango intercuartílico. Asimismo, se observan varios valores atípicos, aquellos que se encuentran fuera de dicha caja.

Análisis de la Distribución de los Datos “chol”



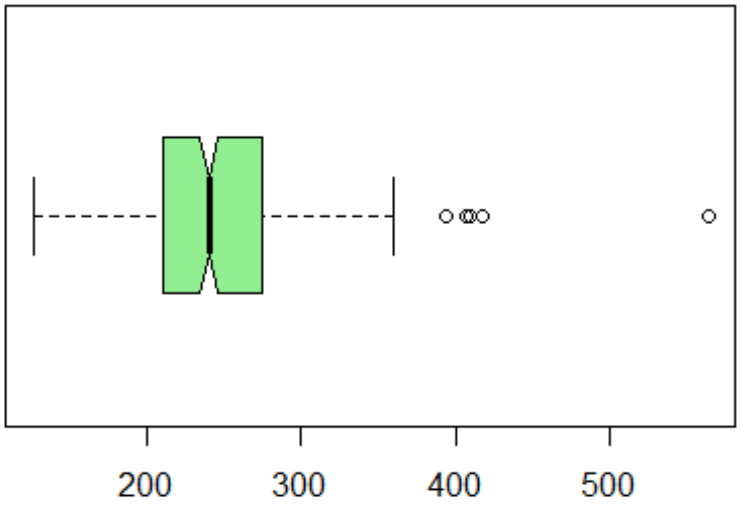
Se atribuye la condición de densidad multimodal a la presente distribución debido a la presencia de múltiples puntos locales en su representación, indicando la existencia de diversas concentraciones significativas.

```
## [1] 1.137733
```

Al calcular el coeficiente de asimetría, obtenemos un resultado de 1.137733, lo cual indica que la distribución exhibe una asimetría positiva o sesgo hacia la derecha, dado que dicho coeficiente es positivo o mayor que 0. Esta conclusión también se corrobora al comparar la media y la mediana; cuando la media supera significativamente a la mediana, se sugiere una asimetría positiva. En este caso particular, aunque la media es de 246.3 y la mediana es de 240, la disparidad entre ambas no es considerable.

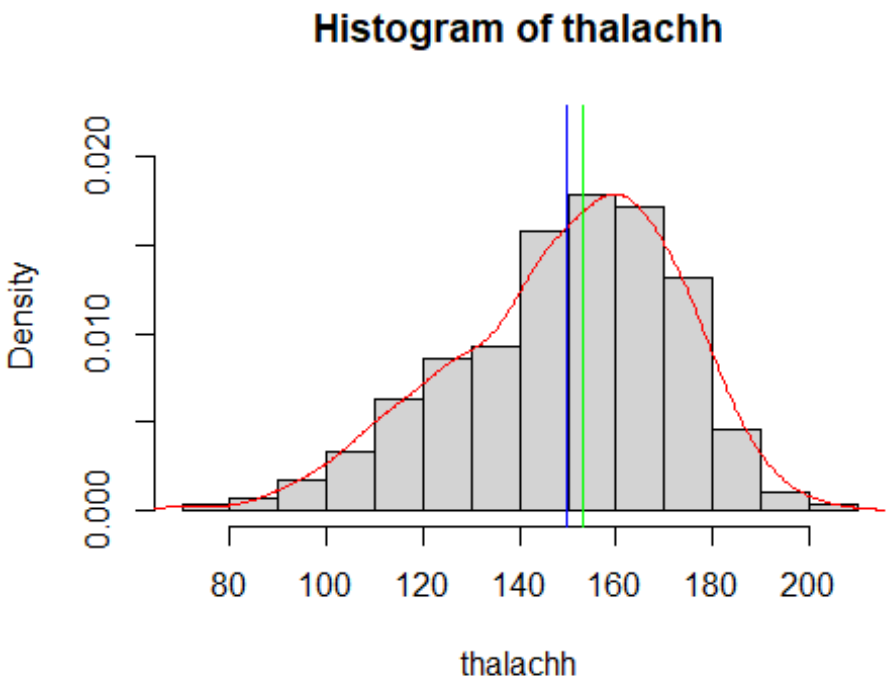
```
## The following objects are masked from heart_heart_csv (pos = 3):
##
```

```
##     age, caa, chol, cp, exng, fbs, oldpeak, output, restecg, sex, slp,
##     thalachh, thall, trtbps

## The following objects are masked from heart_heart_csv (pos = 5):
##
##     age, caa, chol, cp, exng, fbs, oldpeak, output, restecg, sex, slp,
##     thalachh, thall, trtbps
```



Al examinar el diagrama de cajas, se observa una notable concentración de valores en el intervalo comprendido entre 200 y 300. Además, se puede confirmar la presencia de valores extremos, ya que estos se sitúan fuera del rango intercuartílico.

Análisis de la Distribución de los Datos “thalachh”



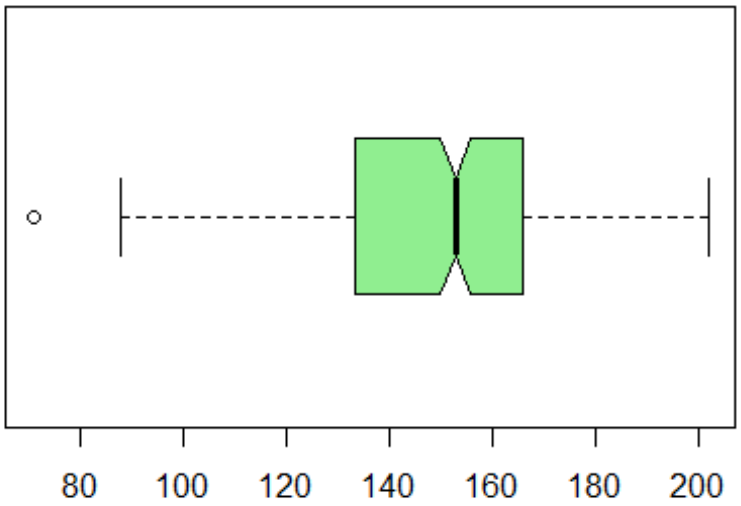
La caracterización de esta distribución como multimodal se fundamenta en la presencia de múltiples puntos locales en su representación, lo que señala la existencia de diversas concentraciones significativas.

```
## [1] -0.5347455
```

Estamos ante una asimetría negativa, con un valor de -0.5347, esto indica una distribución unilateral que se extiende hacia valores más negativo.

```
## The following objects are masked from heart_heart_csv (pos = 3):
##
##     age, caa, chol, cp, exng, fbs, oldpeak, output, restecg, sex, slp,
##     thalachh, thall, trtbps

## The following objects are masked from heart_heart_csv (pos = 4):
##
##     age, caa, chol, cp, exng, fbs, oldpeak, output, restecg, sex, slp,
##     thalachh, thall, trtbps
```

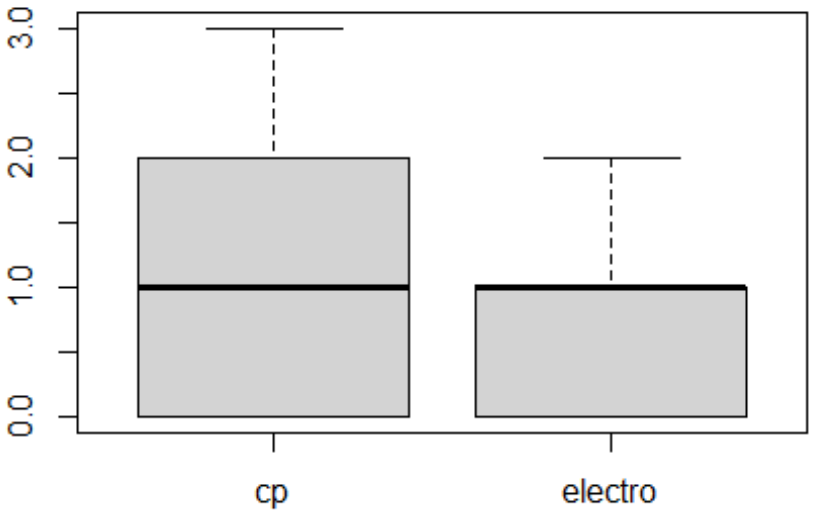
```
## The following objects are masked from heart_heart_csv (pos = 6):
##
##     age, caa, chol, cp, exng, fbs, oldpeak, output, restecg, sex, slp,
##     thalachh, thall, trtbps
```



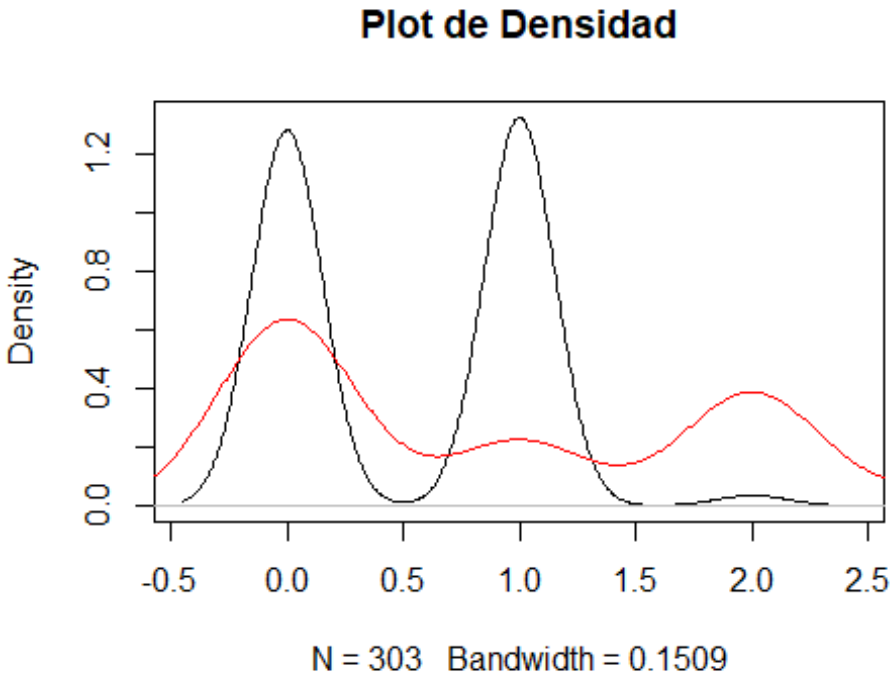
Al analizar el diagrama de cajas, se nota una destacada concentración de valores en el intervalo de 130 a 170. Asimismo, se confirma la existencia de valores extremos, evidenciados por su ubicación fuera del rango intercuartílico.

Comparación de Variables Numéricas

En este caso, vamos a comparar las variables “número de shares”cp, que mide el dolor en el pecho, y “número de sads”restecg, que son los resultados electrográficos en reposo, con el objetivo de poder relacionar ambas magnitudes.



En el boxplot proporcionado, se comparan las distribuciones de las variables “cp” (tipo de dolor de pecho) y “restecg” (resultados electrocardiográficos en reposo). Ambos boxplots presentan rangos intercuartílicos similares, ya que las cajas abarcan desde el primer cuartil hasta el tercer cuartil. Es interesante observar que ambos boxplots se extienden hasta sus valores máximos.



El gráfico de densidad muestra la distribución de datos en un intervalo de tiempo continuo, este gráfico es una alteración de un Histograma. Los picos de este gráfico de densidad ayudan a mostrar en dónde están ubicados los valores que se concentran en el intervalo. Las áreas de mayor densidad en el gráfico pueden indicar regiones donde la combinación de ciertos niveles de dolor de pecho y resultados electrocardiográficos en reposo es más común. La forma y dirección de las curvas de densidad conjunta pueden revelar patrones en la asociación entre ambas variables. Las áreas de baja densidad pueden señalar combinaciones menos comunes de valores, y la presencia de valores atípicos podría destacarse en regiones de baja densidad.

Datos Multivariantes

En esta sección, nos enfocaremos en el análisis y descripción de un conjunto de datos multivariante. En su mayoría, las variables serán de naturaleza numérica o cuantitativa, por lo que nuestro interés se centrará en comprender las relaciones entre estas. Con el propósito de cuantificar la variación conjunta de dos variables, emplearemos la covarianza y calcularemos la matriz de varianzas y covarianzas entre dichas variables. Este enfoque

nos permitirá obtener una comprensión más profunda de cómo las distintas variables numéricas interactúan entre sí, proporcionando una herramienta fundamental para explorar la complejidad y las interrelaciones dentro del conjunto de datos.

```
##                   cp      trtbps         chol          fbs      restecg
## cp         1.06513234   0.8617140   -4.1137740  0.034719035  0.024107709
## trtbps     0.86171399 307.5864533  111.9672153  1.109042030 -1.052324438
## chol      -4.11377396 111.9672153 2686.4267480  0.245426529 -4.116702730
## fbs        0.03471903   1.1090420    0.2454265  0.126876926 -0.015769458
## restecg    0.02410771  -1.0523244   -4.1167027 -0.015769458  0.276528315
## thalachh   6.99161804 -18.7591305  -11.8004940 -0.069897056  0.531462418
## exng      -0.19116779   0.5571110    1.6319913  0.004294800 -0.017474264
## oldpeak   -0.17882106   3.9344863    3.2467937  0.002376893 -0.035882893
## slp        0.07613708  -1.3128319   -0.1289642 -0.013146679  0.030151028
## caa       -0.19108037   1.8183726    3.7372522  0.050258999 -0.038740629
## thall     -0.10220095   0.6680218    3.1354884 -0.006983149 -0.003857671
## output     0.22332962  -1.2679496   -2.2038555 -0.004983280  0.035997639
##               thalachh        exng      oldpeak         slp         caa
## cp          6.99161804 -0.19116779 -0.178821061  0.07613708 -0.19108037
## trtbps    -18.75913055  0.55711101  3.934486263 -1.31283195  1.81837257
## chol      -11.80049396  1.63199134  3.246793653 -0.12896422  3.73725220
## fbs        -0.06989706  0.00429480  0.002376893 -0.01314668  0.05025900
## restecg     0.53146242 -0.01747426 -0.035882893  0.03015103 -0.03874063
## thalachh  524.64640570 -4.07629008 -9.153517802  5.45936878 -4.99323542
## exng       -4.07629008  0.22070684  0.157215920 -0.07461806  0.05560291
## oldpeak    -9.15351780  0.15721592  1.348095207 -0.41321881  0.26439578
## slp         5.45936878 -0.07461806 -0.413218805  0.37973466 -0.05051035
## caa        -4.99323542  0.05560291  0.264395777 -0.05051035  1.04572378
## thall      -1.35249055  0.05947151  0.149462330 -0.03952746  0.09506480
## output      4.81876598 -0.10235394 -0.249452495  0.10632090 -0.19982296
##                 thall      output
## cp       -0.102200949  0.22332962
## trtbps    0.668021769 -1.26794964
## chol      3.135488383 -2.20385548
```

```
## fbs      -0.006983149 -0.00498328
## restecg  -0.003857671  0.03599764
## thalachh -1.352490547  4.81876598
## exng      0.059471510 -0.10235394
## oldpeak   0.149462330 -0.24945249
## slp      -0.039527463  0.10632090
## caa       0.095064804 -0.19982296
## thall     0.374882521 -0.10507508
## output   -0.105075077  0.24883614
```

Con la finalidad de evaluar la interdependencia entre variables, se procede a su estudio mediante el cálculo de la matriz de correlaciones.

```
##                   cp      trtbps         chol          fbs     restecg
## cp         1.00000000  0.04760776 -0.076904391  0.094444035  0.04442059
## trtbps     0.04760776  1.00000000  0.123174207  0.177530542 -0.11410279
## chol      -0.07690439  0.12317421  1.000000000  0.013293602 -0.15104008
## fbs        0.09444403  0.17753054  0.013293602  1.000000000 -0.08418905
## restecg    0.04442059 -0.11410279 -0.151040078 -0.084189054  1.00000000
## thalachh   0.29576212 -0.04669773 -0.009939839 -0.008567107  0.04412344
## exng      -0.39428027  0.06761612  0.067022783  0.025665147 -0.07073286
## oldpeak   -0.14923016  0.19321647  0.053951920  0.005747223 -0.05877023
## slp        0.11971659 -0.12147458 -0.004037770 -0.059894178  0.09304482
## caa       -0.18105303  0.10138899  0.070510925  0.137979327 -0.07204243
## thall     -0.16173557  0.06220989  0.098802993 -0.032019339 -0.01198140
## output     0.43379826 -0.14493113 -0.085239105 -0.028045760  0.13722950
##              thalachh        exng      oldpeak         slp         caa
## cp         0.295762125 -0.39428027 -0.149230158  0.11971659 -0.18105303
## trtbps    -0.046697728  0.06761612  0.193216472 -0.12147458  0.10138899
## chol      -0.009939839  0.06702278  0.053951920 -0.00403777  0.07051093
## fbs       -0.008567107  0.02566515  0.005747223 -0.05989418  0.13797933
## restecg    0.044123444 -0.07073286 -0.058770226  0.09304482 -0.07204243
## thalachh   1.000000000 -0.37881209 -0.344186948  0.38678441 -0.21317693
## exng      -0.378812094  1.00000000  0.288222808 -0.25774837  0.11573938
```

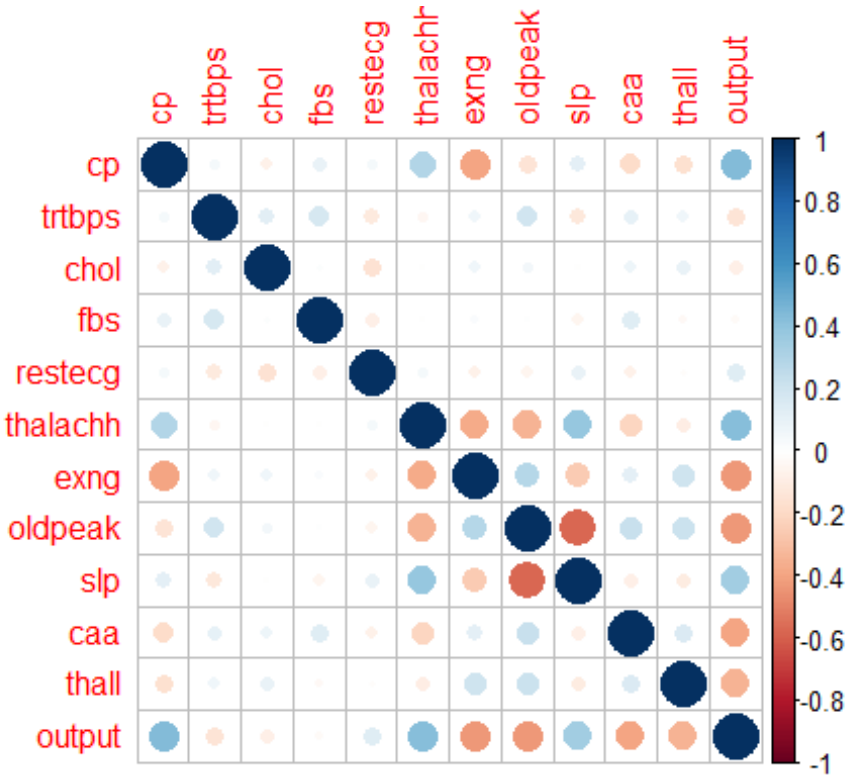
```
## oldpeak  -0.344186948  0.28822281  1.000000000 -0.57753682  0.22268232
## slp       0.386784410 -0.25774837 -0.577536817  1.00000000 -0.08015521
## caa      -0.213176928  0.11573938  0.222682322 -0.08015521  1.00000000
## thall    -0.096439132  0.20675379  0.210244126 -0.10476379  0.15183213
## output    0.421740934 -0.43675708 -0.430696002  0.34587708 -0.39172399
##                thall      output
## cp       -0.16173557  0.43379826
## trtbps    0.06220989 -0.14493113
## chol      0.09880299 -0.08523911
## fbs      -0.03201934 -0.02804576
## restecg  -0.01198140  0.13722950
## thalachh -0.09643913  0.42174093
## exng      0.20675379 -0.43675708
## oldpeak   0.21024413 -0.43069600
## slp      -0.10476379  0.34587708
## caa       0.15183213 -0.39172399
## thall     1.00000000 -0.34402927
## output   -0.34402927  1.00000000

## [1] 0.1287076
```

Dado que el determinante de la matriz de correlación se encuentra próximo a 0, esto sugiere un elevado grado de dependencia entre las variables.

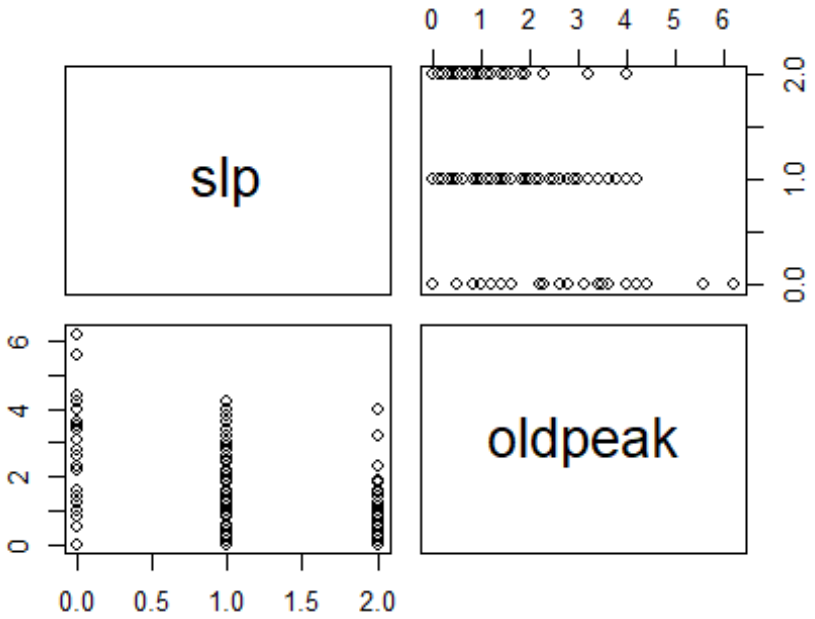
Dependencia Lineal entre las Variables

Además, en este punto, es relevante crear gráficos de dispersión para visualizar las relaciones entre las variables en pares. Estos diagramas son fundamentales para identificar posibles relaciones no lineales, ya que en tales casos la matriz de varianzas-covarianzas podría no ser suficiente para resumir la dependencia entre las variables.

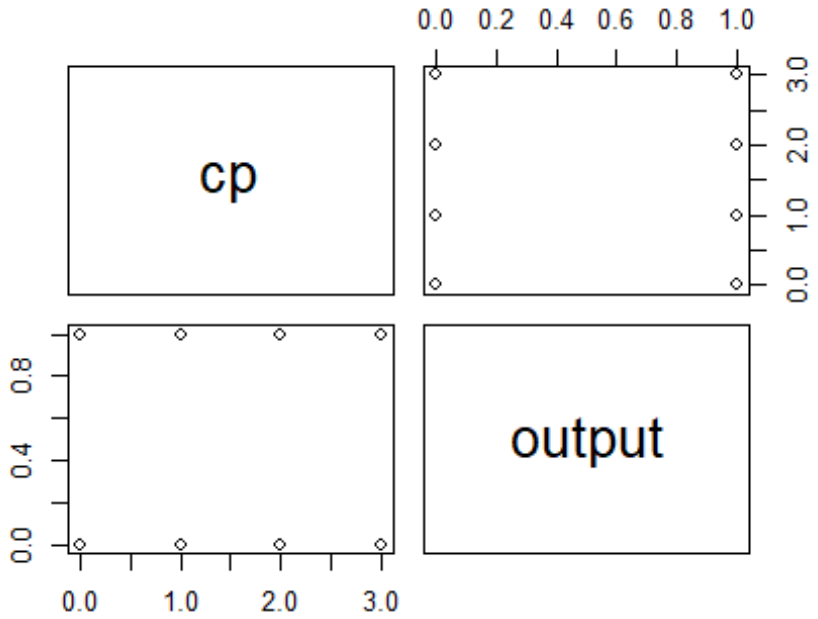
```
## Warning: package 'corrplot' was built under R version 4.1.3

## corrplot 0.92 loaded
```



Como se evidencia en el diagrama de dispersión, son escasas las relaciones lineales fuertes entre las variables. Por lo tanto, a continuación, nos enfocaremos en analizar la relación más fuerte identificada y aquella que presenta una conexión mínima.



Podemos concluir que la relación entre ambas variables es prácticamente nula. Esto se debe a que la variable “slp”, que mide la pendiente del segmento ST durante el ejercicio máximo, y la variable “oldpeak”, que evalúa la depresión del segmento ST inducida por el ejercicio en comparación con el reposo, están vinculadas a la actividad física y a cambios en el electrocardiograma, pero miden conceptos distintos.



Si “output” asigna probabilidades a la posibilidad de tener un ataque al corazón, es razonable suponer que podría haber una conexión entre “cp” y las predicciones de riesgo de ataque cardíaco. En este sentido, sería plausible anticipar que niveles más elevados de dolor en el pecho se correlacionen con una mayor probabilidad de que el modelo predice un riesgo más alto de ataque cardíaco. Por consiguiente, es factible que exista una relación lineal entre ambas variables.

Selección de los Componentes Principales

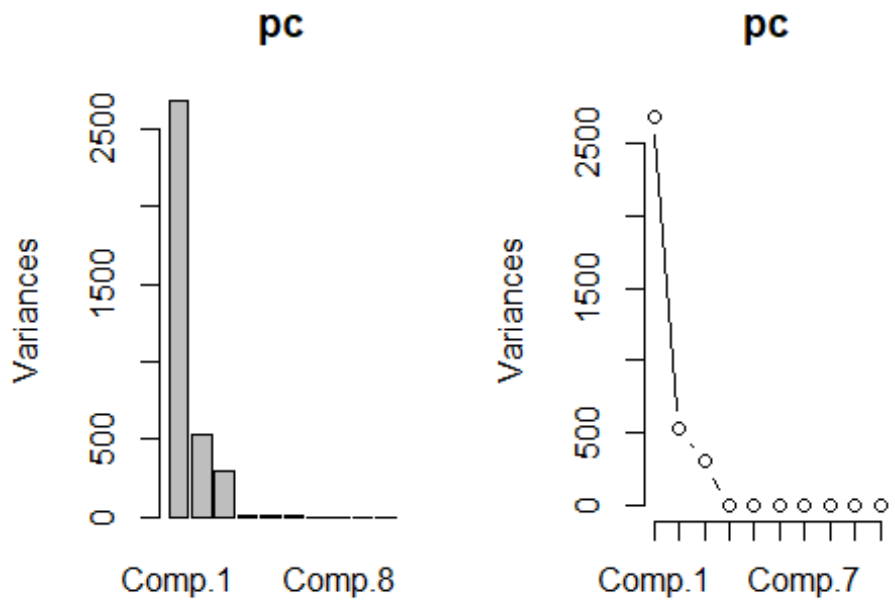
A continuación, procederemos a calcular los componentes principales. Para llevar a cabo este cálculo, es necesario seleccionar la matriz apropiada para la determinación de los autovalores y los autovectores.

```
##
## Loadings:
##            Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
Comp.10
##  cp                                          0.353   0.663   0.605   0.165
0.173
## trtbps                  0.996
## chol     -0.999
## fbs                                                          0.104  0.112
0.141
## restecg                                                     -0.917  0.365
0.106
## thalachh         0.996
## exng                              -0.112        -0.130  0.112  0.105  0.105
0.866
## oldpeak                        -0.737  0.558 -0.113 -0.116 -0.146 -0.306
## slp                                0.217 -0.226               -0.295 -0.858
0.211
## caa                            -0.445 -0.428  0.767
##  thall                                    -0.165                      0.945
-0.219
## output                                 0.214                   -0.176 -0.154
-0.311
##          Comp.11 Comp.12
## cp
## trtbps
## chol
## fbs      -0.832   0.509
```

```
## restecg
## thalachh
## exng             -0.416
## oldpeak
## slp               0.126
## caa              -0.127
## thall    -0.136  -0.101
## output   -0.516  -0.719
##
##                   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
Comp.9
## SS loadings      1.000   1.000   1.000   1.000   1.000   1.000   1.000   1.000
1.000
## Proportion Var   0.083   0.083   0.083   0.083   0.083   0.083   0.083   0.083
0.083
## Cumulative Var   0.083   0.167   0.250   0.333   0.417   0.500   0.583   0.667
0.750
##                Comp.10 Comp.11 Comp.12
## SS loadings      1.000   1.000   1.000
## Proportion Var   0.083   0.083   0.083
## Cumulative Var   0.833   0.917   1.000
```

Para tomar la decisión sobre qué componentes retener, disponemos de tres enfoques de estudio para orientar nuestra elección.

Gráficos de Sedimentación



En el gráfico izquierdo, se aprecia que los tres primeros componentes explican la mayor parte de la variabilidad de los datos. En el gráfico derecho, se observa que a partir del segundo componente se puede aproximar a una línea recta. Este enfoque visual sugiere la elección de los dos o tres primeros componentes como candidatos más relevantes para retener.

Proporción de la Varianza

```
## Importance of components:
##                             Comp.1     Comp.2      Comp.3       Comp.4
## Standard deviation     51.7967661 22.9076588 17.31797723 1.1991291037
## Proportion of Variance  0.7638997  0.1494144  0.08539362 0.0004094143
## Cumulative Proportion   0.7638997  0.9133142  0.99870777 0.9991171859
##                               Comp.5       Comp.6       Comp.7       Comp.8
## Standard deviation     1.0104314122 0.9253485447 5.918983e-01 5.255744e-01
```

```
## Proportion of Variance 0.0002906998 0.0002438045 9.975285e-05 7.865013e-05
## Cumulative Proportion  0.9994078858 0.9996516903 9.997514e-01 9.998301e-01
##                            Comp.9      Comp.10      Comp.11      Comp.12
## Standard deviation     4.582512e-01 4.015061e-01 3.412882e-01 0.3302322094
## Proportion of Variance 5.979134e-05 4.590028e-05 3.316451e-05 0.0000310506
## Cumulative Proportion  9.998899e-01 9.999358e-01 9.999689e-01 1.0000000000
```

De la operación anterior, la primera fila nos suministra las desviaciones estándar asociadas a cada componente. Además, los autovalores reflejan la varianza relacionada con cada componente; por lo tanto, el cuadrado de estos valores (la varianza) equivale a los autovalores. La segunda fila indica la proporción de varianza explicada por cada componente, mientras que la tercera fila exhibe la proporción de varianza acumulada al agregar la varianza explicada por la componente actual a las anteriores. Al centrarnos en la última fila, y al igual que se evidenció en el gráfico de sedimentación, se observa que el primer componente abarca por sí solo el 66.3% de la variabilidad. Para alcanzar un nivel del 90%, sería necesario retener los dos primeros componentes principales. Sin embargo, si el objetivo es alcanzar el 99%, se requeriría conservar los tres primeros componentes.

Cálculo de Autovalores

```
##          Comp.1          Comp.2         Comp.3         Comp.4         Comp.5
Comp.6
##  2682.9049737    524.7608331    299.9123354       1.4379106       1.0209716
0.8562699
##          Comp.7          Comp.8          Comp.9        Comp.10        Comp.11
Comp.12
##       0.3503437       0.2762284       0.2099941       0.1612071       0.1164776
0.1090533

## [1] 292.6764
```

El valor de referencia en esta situación es 292.6764. Al observar detenidamente, notamos que los tres primeros componentes poseen un autovalor asociado mayor que dicho valor.

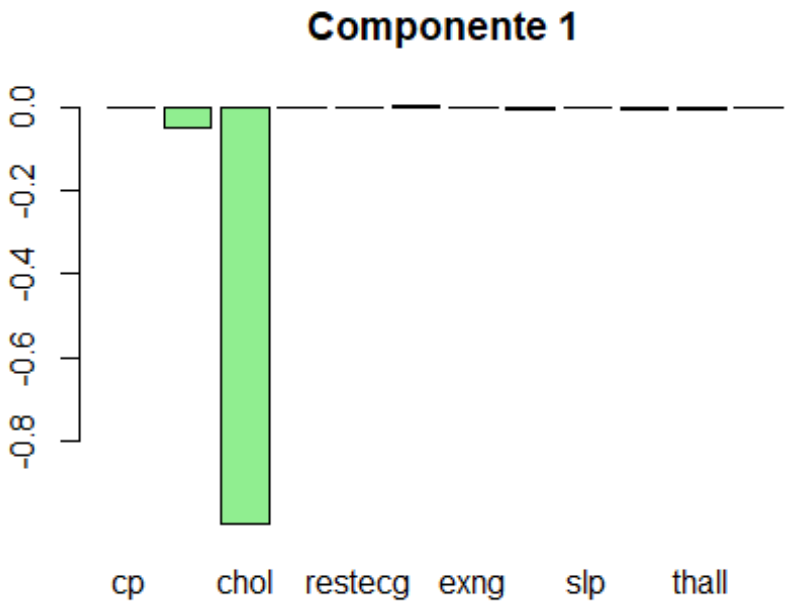
Por lo tanto, este tercer método también respalda la decisión de retener los tres primeros componentes.

Tras realizar estos cálculos mediante los tres métodos, podemos afirmar que optamos por conservar los tres primeros componentes principales.

```
##                  Comp.1        Comp.2        Comp.3
## cp         1.527711e-03  0.0130660971  0.0053855273
## trtbps    -4.695977e-02 -0.0808990493  0.9955018118
## chol      -9.988741e-01  0.0097270963 -0.0463424267
## fbs       -1.105958e-04 -0.0002987658  0.0036155697
## restecg    1.547379e-03  0.0010959148 -0.0027081078
## thalachh   5.863633e-03  0.9962467354  0.0815078358
## exng      -6.245369e-04 -0.0077866839  0.0004926436
## oldpeak   -1.294510e-03 -0.0179360342  0.0100901359
## slp        8.301833e-05  0.0105573834 -0.0028630768
## caa       -1.430361e-03 -0.0096975728  0.0041110936
## thall     -1.178518e-03 -0.0026186910  0.0013683051
## output     8.509534e-04  0.0092985993 -0.0025613835
```

El primer componente se puede interpretar de la siguiente forma:

R1 = 1.527711e-03X1 + (-4.695977e-02X2) + (-9.988741e-01X3) + (-1.105958e-04X4) + 1.547379e-03X5 + 5.863633e-03X6 + (-6.245369e-04X7) + (-1.294510e-03X8) + 8.301833e-05X9 + (-1.430361e-03X10) + (-1.178518e-03X11) + 8.509534e-04X12



Como se evidencia, los coeficientes presentan tanto valores positivos como negativos. En consecuencia, procederemos a clasificar las variables en dos grupos según el signo de los coeficientes que las acompañan:

```
R1    =    (1.527711e-03*datos_num[,1]    +    1.547379e-03*datos_num[,5]    +
5.863633e-03*datos_num[,6]        +        8.301833e-05*datos_num[,9]        +
8.509534e-04*datos_num[,12])       -       (4.695977e-02*datos_num[,2]       +
9.988741e-01*datos_num[,3]        +        1.105958e-04*datos_num[,4]        +
6.245369e-04*datos_num[,7]        +        1.294510e-03*datos_num[,8]        +
1.430361e-03*datos_num[,10] + 1.178518e-03*datos_num[,11])
R1

##      [1]  -238.6661  -254.7282  -208.8682  -240.3250  -358.2820  -197.4896
-299.3478
##      [8]  -267.3241  -205.9018  -173.8334  -244.3684  -279.9771  -270.8016
-215.0834
##     [15]  -288.7736  -223.4610  -344.2406  -232.1198  -252.7649  -244.4203
-239.1338
```

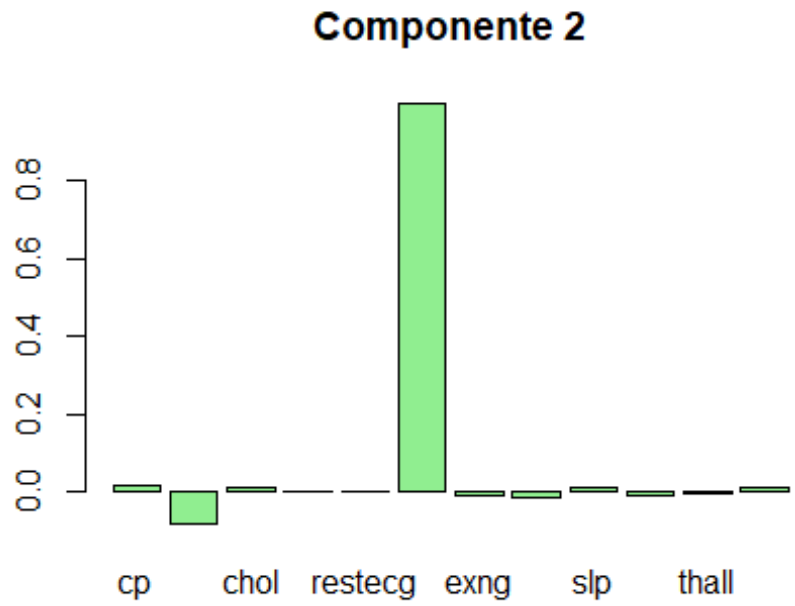
```
##    [22]  -237.7907  -231.2760  -248.9659  -204.3054  -308.2253  -217.8836
-179.2448
##    [29]  -422.1851  -201.9918  -201.7225  -181.6165  -223.7556  -277.6719
-217.8978
##    [36]  -182.5317  -308.9973  -237.8168  -275.1059  -366.2222  -313.3966
-249.7736
##    [43]  -211.7877  -268.9710  -326.1441  -329.2592  -240.2526  -262.2746
-221.0893
##    [50]  -239.2801  -260.9418  -306.4117  -235.9932  -144.8844  -257.0456
-206.1406
##    [57]  -226.3898  -264.0242  -186.3128  -307.7402  -269.1048  -312.8085
-190.2134
##    [64]  -208.3342  -216.3677  -188.2088  -225.6065  -238.8158  -224.3889
-213.6317
##    [71]  -262.4828  -230.2556  -208.6904  -266.1918  -217.5189  -255.1158
-249.6223
##    [78]  -226.3628  -209.6995  -243.7585  -253.9252  -312.6671  -321.4918
-303.7568
##    [85]  -268.7784  -567.8288  -281.3446  -200.6060  -217.9971  -251.7041
-259.5095
##    [92]  -211.9821  -228.2409  -292.9441  -164.2685  -231.7661  -399.2131
-236.9529
##    [99]  -319.8008  -250.8161  -249.6323  -277.2082  -200.3064  -244.2266
-200.8781
##   [106]  -215.7236  -240.4805  -241.3258  -248.4102  -257.9486  -332.1843
-131.8874
##   [113]  -318.4403  -214.9849  -266.8992  -219.3930  -218.8797  -197.4684
-207.6908
##   [120]  -248.3176  -308.0536  -276.1095  -271.9481  -270.7901  -202.1373
-214.1782
##   [127]  -208.1912  -282.8156  -201.1734  -273.6249  -207.3297  -276.0359
-299.3514
##   [134]  -239.0021  -310.6149  -273.8459  -182.8687  -212.9557  -205.2017
```

```
-268.1024
##  [141]  -299.3815  -306.9993  -213.3823  -226.8971  -202.6691  -251.2112
-246.3937
##  [148]  -245.7674  -230.3865  -185.0193  -234.4508  -153.3607  -233.8175
-283.6517
##  [155]  -225.3383  -202.1155  -257.7671  -196.4909  -224.7837  -225.9012
-244.3724
##  [162]  -346.8401  -161.3895  -180.2715  -180.2715  -292.5674  -233.6312
-273.3443
##  [169]  -258.9635  -208.4452  -260.9849  -232.9244  -288.3803  -228.9394
-211.1089
##  [176]  -171.3158  -234.3044  -340.2683  -181.7396  -282.0804  -358.0328
-231.1312
##  [183]  -334.7445  -234.0381  -249.0266  -294.0395  -257.9812  -270.8927
-238.3573
##  [190]  -176.0488  -309.9328  -221.0107  -192.7661  -287.6696  -190.4603
-332.8038
##  [197]  -236.9243  -258.6334  -271.7618  -251.9645  -200.9095  -262.7618
-276.0941
##  [204]  -281.2679  -170.4945  -259.7836  -243.0709  -263.8426  -192.6142
-182.4291
##  [211]  -233.8758  -264.5302  -223.4772  -312.6128  -253.7513  -346.0254
-268.2417
##  [218]  -334.9699  -259.3174  -260.9440  -412.6956  -222.6890  -287.1437
-296.3001
##  [225]  -243.1653  -179.8863  -285.7200  -202.6546  -294.7257  -313.7557
-246.9044
##  [232]  -295.7050  -295.3445  -250.8026  -327.1128  -304.2278  -304.5358
-298.2555
##  [239]  -308.5849  -286.6887  -275.5586  -256.0535  -217.8030  -280.3189
-189.3817
##  [246]  -278.5453  -413.9620  -252.5359  -290.5575  -259.4396  -303.5357
-252.0921
```

```
##  [253]  -299.5355  -302.6333  -279.4708  -314.4668  -263.9678  -205.8035
-249.8694
##  [260]  -235.3108  -235.1414  -234.0644  -286.9095  -272.7844  -210.3047
-216.2512
##  [267]  -334.4027  -153.6391  -290.7368  -288.1885  -253.5149  -239.1813
-241.9540
##  [274]  -237.5212  -279.1697  -216.6522  -224.0009  -265.7031  -324.1397
-171.5696
##  [281]  -320.3014  -208.8668  -222.8866  -228.8273  -212.5400  -316.5283
-209.1130
##  [288]  -238.0111  -338.9577  -210.0228  -208.7807  -322.1825  -231.8821
-218.0211
##  [295]  -173.6044  -192.5315  -201.8051  -182.9809  -246.5846  -268.0932
-198.7274
## [302] -136.2885 -240.8210
```

Por lo tanto, se puede interpretar como la suma de dos medias ponderadas. Donde se aprecia un alto peso negativo en la variable chol, que mide el colesterol en mg/dl obtenido mediante el sensor de IMC; este valor negativo separa a los afectados por infartos en los que poseen un nivel de colesterol superior y los que no.

El segundo componente se puede interpretar de la siguiente forma: R2 = 0.0130660971X1 + (-0.0808990493X2) + 0.0097270963X + (-0.0002987658X4) + 0.0010959148X5 + 0.9962467354X6 + (-0.0077866839X7) + (-0.0179360342X8) + 0.0105573834X9 + (-0.0096975728X10) + (-0.0026186910X11) + 0.0092985993X12




```
R2    =    (0.0130660971*datos_num[,1]    +    0.0097270963*datos_num[,3]    +
0.0010959148*datos_num[,5]        +        0.9962467354*datos_num[,6]        +
0.0105573834*datos_num[,9]        +       0.0092985993*datos_num[,12])       -
(0.0808990493*datos_num[,2]        +        0.0002987658*datos_num[,4]        +
0.0077866839*datos_num[,7]        +        0.0179360342*datos_num[,8]        +
0.0096975728*datos_num[,10] + 0.0026186910*datos_num[,11])
R2
```

```
##      [1]  139.97739  178.18155  162.83502  169.94462  156.13145  137.99741
143.96402
##      [8]  165.23774  149.45355  162.86993  150.40314  130.68519  162.45730
136.62680
##     [15]  152.05602  149.84248  165.00617  103.63332  160.62529  141.44596
151.75462
##     [22]  170.11515  168.23065  126.73045  167.97170  151.39847  146.36143
115.38334
##     [29]  149.16803  142.83724  160.83066  131.50483  178.94599  143.98415
116.51211
```

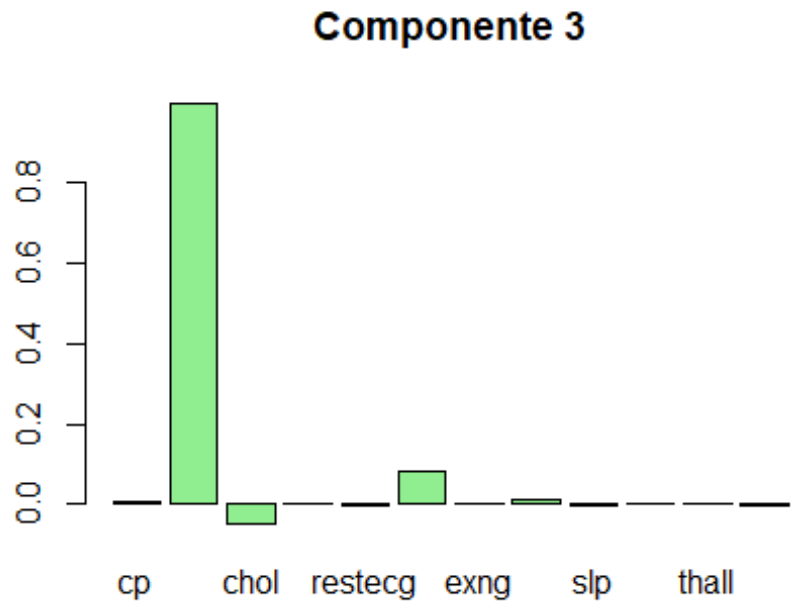
```
##    [36]  149.63080  161.44972  154.52253  137.55981  141.02812  133.15182
171.21477
##    [43]  141.00728  134.52180  173.16475  164.84361  170.33572  146.80159
106.37089
##    [50]  150.53673  140.45636  143.67040  137.15996  167.00869  162.93560
148.53695
##    [57]  177.61680  177.55648  165.63555  151.01094  123.23217  149.71979
181.60647
##    [64]  122.58920  155.15827  171.93406  136.54533  166.11937  161.83336
154.41597
##    [71]  139.28093  148.05784  192.74753  176.53219  156.62116  151.90870
157.64512
##    [78]  154.23984  174.98741  147.28168  171.75165  162.04105  154.28370
167.96441
##    [85]  115.87194  155.59160  143.60373  149.19635  150.60284  115.86120
166.82482
##    [92]  158.72015  159.38439  150.54642  130.00647  101.30882  148.91044
139.97005
##    [99]  153.94776  164.24859  167.76290  132.64737  168.91903  185.91251
153.91116
##   [106]  106.92675  119.88264  142.56427  154.07711  152.00019  142.03996
161.47764
##   [113]  124.26573  153.57290  146.48920  161.79779  158.93905  153.57853
164.88370
##   [120]  142.63595  113.93298  172.81406  164.94413  160.28455  172.71175
183.80276
##   [127]  135.41140  161.79490  159.30890  113.47173  151.44222  153.21632
154.59292
##   [134]  145.85206  155.21077  154.51420   87.71533  131.18064  118.56698
96.80108
##   [141]  149.61289  173.95877  164.71761  135.06238  106.17821  132.26441
141.27540
##   [148]  160.60717  160.90854  140.72342  126.74252  116.90650  142.91389
```

```
142.35343
##  [155]  142.44724  121.91264  170.32465  165.38419  135.46731  154.05665
161.01054
##  [162]  158.04277  173.17551  162.90247  162.90247   97.37429  120.96433
150.59127
##  [169]  138.36997  144.99543  133.44582  160.68643  152.44019  163.81348
122.92395
##  [176]  106.25670  152.13328  149.38280  111.51349  102.10915  124.22435
103.58154
##  [183]  161.07464  157.53155  117.70448  146.19209  135.37537  101.09258
153.33823
##  [190]  150.19441  133.89141  122.18066  104.66568  132.40501  144.86953
128.81611
##  [197]  136.52835  154.72704   91.46201  150.90875  169.35919  132.80300
101.06597
##  [204]  137.53246  132.95902  152.51793  134.85661  146.72214  130.56407
151.78467
##  [211]  141.32140  132.21745  132.04096  136.68486  135.73521  128.06863
88.67599
##  [218]  124.14143  118.01543  141.39605  145.14793  101.25331  164.93532
119.01533
##  [225]  118.88906  114.43181   95.61980  121.69747  147.49011  123.39147
145.09920
##  [232]  112.95276  134.29393   88.26249  101.13932  163.91109  163.15780
160.84790
##  [239]  154.21563  147.96972  101.21281  130.80752  121.78902   78.00307
95.66981
##  [246]  158.00443  142.51636  108.98385  181.50489  136.56071  113.00650
134.14123
##  [253]   97.24080  119.31133  114.29757  137.93210  121.59803  115.80182
143.63407
##  [260]  173.82303  152.15684  152.58329   87.37664  162.19214  100.68737
124.50262
```

```
##  [267]  105.11854  117.42891  108.39763   94.80467  136.17260  135.87103
63.31943
##  [274]  149.60359  111.30307  159.29624   94.87338  133.00012  143.53943
114.90474
##  [281]  116.56157  147.02937  125.41000  170.20750  128.13139  111.19320
152.55825
##  [288]  153.20194  136.75547  121.10259  150.40192  133.23683  133.81802
139.21698
##  [295]  135.33599  133.86696  127.37294   78.07704  113.54914  125.19510
130.62178
## [302] 105.29054 165.13434
```

Este segundo componente principal, lo podemos interpretar como la contraposición entre dos variables, siendo la negativa la presión arterial en reposo (trtbps), y la positiva, la frecuencia cardíaca máxima alcanzada (thalach). Esto se debe a que, en general, cuando la presión arterial disminuye, el cuerpo puede responder aumentando la frecuencia cardíaca para compensar la disminución del flujo sanguíneo y mantener un suministro adecuado de oxígeno a los órganos y tejidos.

El tercer componente se puede interpretar de la siguiente forma: R3 = 0.0053855273X1 + 0.9955018118X2 + (-0.0463424267X3) + 0.0036155697X4 + (-0.0027081078X5) + 0.0815078358X6 + 0.0004926436X7 + 0.0100901359X8 + (-0.0028630768X9) + 0.0041110936X10 + 0.0013683051X11 + (-0.0025613835X12)




```
R3    =    (0.0053855273*datos_num[,1]    +    0.9955018118*datos_num[,2]    +
0.0036155697*datos_num[,4]           +      0.0815078358*datos_num[,6]      +
0.0004926436*datos_num[,7]        +        0.0100901359*datos_num[,8]        +
0.0041110936*datos_num[,10]     +     0.0013683051*datos_num[,11]     )     -
(0.0463424267*datos_num[,3]        +        0.0027081078*datos_num[,5]        +
0.0028630768*datos_num[,9]  + 0.0025613835*datos_num[,12] )
R3
```

```
##      [1]  145.81794  133.11515  133.99469  123.03700  116.33906  142.53294
138.23209
##      [8]  121.37151  175.22098  155.74076  141.33952  128.00519  131.02917
111.49620
##     [15]  149.43895  122.21098  117.72565  148.18363  151.82341  140.63638
136.67240
##     [22]  133.21439  143.39700  149.25021  144.68039  158.49853  152.31968
111.42930
##     [29]  132.86321  132.70164  109.04644  122.66585  134.58955  124.19554
124.78460
```

```
##    [36]  146.22547  134.16711  152.04535  153.91041  154.91799  136.69537
132.73747
##    [43]  105.98420  128.83780  139.33398  118.41742  143.15641  138.18969
126.79015
##    [50]  139.57082  129.70651  117.77384  130.64752  115.25521  136.73902
136.96996
##    [57]  126.31811  117.50708  123.22786  126.34078  107.83342  105.90810
124.34811
##    [64]  135.74289  143.04963  143.73888  100.93576  132.84373  123.11834
127.03418
##    [71]  119.49901   95.61813  136.42580  142.43028  125.03647  135.94672
126.65005
##    [78]  142.49348  132.92222  105.97371  114.50301  127.00693   99.85217
152.04248
##    [85]   99.20776  101.41243  116.95813  104.13356  112.48768   98.00862
125.90313
##    [92]  135.50028  140.83867  131.02742  115.32949  139.93155  133.91749
108.70813
##    [99]  128.04743  132.13702  150.56201  176.56558  144.92873  124.17185
132.62491
##   [106]  119.07858  159.13589  138.83145  121.36516  110.68842  176.67348
157.60061
##   [113]  135.71151  112.84282  129.90436  123.35544  133.21954  123.75441
109.09031
##   [120]  138.50504  125.34044  139.64933  113.10149  108.75780   98.94744
123.39100
##   [127]  113.69072  152.50539  140.08905  116.86301  163.25786  134.04270
118.99060
##   [134]  111.08255  124.53535  130.22664  119.04215  129.19955  110.46922
123.79709
##   [141]  118.59721  115.20058  123.88088  106.76593  139.70947  155.59984
118.41153
##   [148]  152.15791  122.76417  133.30229  159.97859  114.79041  171.37018
```

```
144.86146
##  [155]  139.57848  130.96396  132.28302  126.73297  126.00136  132.46054
122.11571
##  [162]  129.09666  127.01601  143.38914  143.38914  154.85704  119.39850
140.03902
##  [169]  129.64539  142.63595  129.14924  112.60297  119.36364  135.17605
130.66190
##  [176]  111.07982  118.87833  136.72885  121.06551  145.67280  125.82164
148.21378
##  [183]  127.89407  114.32759  148.52454  110.52872  129.44213  120.02754
141.87772
##  [190]  114.41092  126.86604  128.12806  119.97500  142.89152  143.47153
165.57768
##  [197]  150.64608  125.96492  115.18148  110.90043  114.80375  124.00808
145.86713
##  [204]  178.75094  163.57777  128.72995  110.02143  150.20135  122.11925
144.37219
##  [211]  129.05815  118.86175  118.74196  142.03261  124.65579  126.72416
125.15896
##  [218]  124.91425  133.00687  129.78846  143.07005  138.41971  138.53093
196.65108
##  [225]  108.73074  146.50078  114.85817  120.89561  168.86785  120.82340
108.64725
##  [232]  160.99909  157.71988  115.91427  123.41375  139.62703  124.47944
139.66983
##  [239]  123.56373  125.07876  155.98622  173.33133  145.30911  145.80639
131.46185
##  [246]  124.28099  126.69849  157.67500  193.92334  139.54395  135.55790
131.63808
##  [253]  132.42669   95.89994  156.83045  139.03717  126.05779  144.36458
150.58171
##  [260]  123.64307  180.10300  113.87711  117.14883  108.84695  108.76598
112.43327
```

```
##  [267]  173.60203  120.86237  117.69307  124.71999  119.66453  134.41953
114.27137
##  [274]  101.42207  106.39352  128.32043  143.82176  122.84354  133.00844
139.91569
##  [281]  130.99331  130.69423  126.28912  155.73050  141.04761  134.76411
137.17441
##  [288]  155.92963  105.66951  128.54066  151.04930  110.21409  170.74749
153.73794
##  [295]  123.39288  142.48880  125.39549  162.46217  138.22824  108.05663
145.94546
## [302] 132.73302 132.67016
```

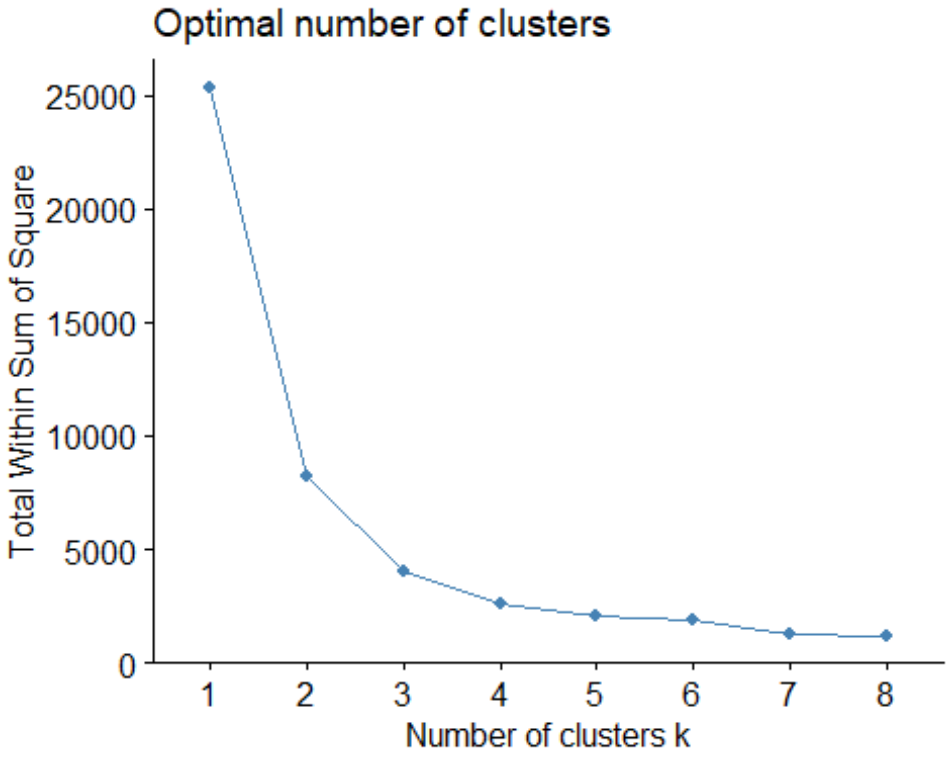
En este último componente principal, observamos que la variable "trtbps", que mide la presión arterial en reposo, es la que más peso positivo tiene. Por contraposición, el colesterol es la variable más pesada en los negativos.

Análisis de Conglomerados

En este apartado nos vamos a centrar en el análisis de conglomerados, cuyo objetivo es agrupar las observaciones de nuestra base de datos en grupos homogéneos que tengan características comunes entre ellos.

Vamos a comenzar realizando la agrupación de variables cuantitativas que no midan una escala de dolor como la variable “cp”.

```
## Warning in age$oldpeak = subset(heart_heart_csv, select = c("age",
"oldpeak")):
## Realizando coercion de LHD a una lista

## Warning: package 'factoextra' was built under R version 4.1.3

## Loading required package: ggplot2

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



Al analizar la gráfica generada por esta función, es crucial identificar el punto en el que se forma un “codo”, ya que a partir de ese punto la reducción en la variabilidad disminuye. En este escenario, podemos concluir que 3 grupos podrían ser un número apropiado para representar la estructura subyacente de este conjunto de datos. Después de determinar el número apropiado de grupos, implementamos el algoritmo para agrupar los datos en cada uno de esos grupos mediante la función kmeans() de R. Esta función lleva a cabo una asignación aleatoria de los centros iniciales, y la cantidad de asignaciones aleatorias se especifica mediante el parámetro nstart() de la función

```
## K-means clustering with 3 clusters of sizes 130, 90, 83
##
## Cluster means:
##        age  oldpeak
## 1 54.80000 1.050769
## 2 64.67778 1.392222
## 3 42.50602 0.639759
##
## Clustering vector:
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
19  20
##   2   3   3   1   1   1   1   3   1   1   1   3   1   2   1   1   1   2
3   2
##  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38
39  40
##   1   3   3   2   3   2   1   1   2   1   3   2   3   1   1   3   1   1
2   2
##  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58
59  60
##   1   3   3   1   3   1   3   3   1   1   1   2   2   3   2   1   3   3
3   1
##  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78
79  80
##   2   1   1   3   1   3   1   3   3   2   1   1   3   1   3   1   1   1
```

```
1   1
##  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98
99 100
##   3   3   2   1   3   2   2   3   1   1   3   1   1   1   3   1   2   1
3   1
## 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118
119 120
##   3   1   2   3   1   2   2   3   1   1   2   1   2   3   1   3   3   1
3   3
## 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138
139 140
##   2   1   3   1   3   3   3   2   1   2   1   1   3   3   3   1   2   2
1   2
## 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158
159 160
##   1   3   3   2   2   2   3   2   3   3   2   2   2   2   3   1   3   3
1   1
## 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178
179 180
##   1   1   3   3   3   2   2   2   2   1   1   3   1   1   2   3   2   2
3   1
## 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198
199 200
##   1   2   2   1   1   3   2   1   1   3   1   1   1   2   2   1   3   2
2   2
## 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218
219 220
##   3   2   1   2   2   1   1   2   1   1   1   2   3   2   1   3   2   2
2   3
## 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238
239 240
##   2   1   2   1   1   2   2   3   1   2   3   1   1   2   2   1   1   2
2   3
```

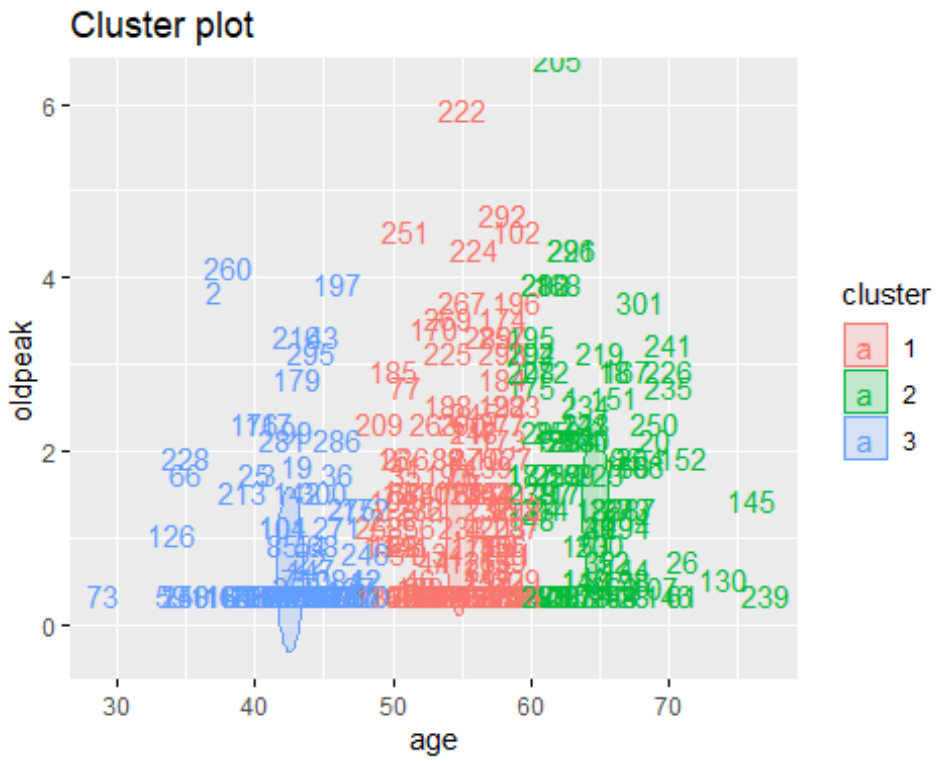
```
## 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258
259 260
##   2   1   2   1   1   3   1   2   1   2   1   3   2   2   1   3   1   1
2   3
## 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278
279 280
##   2   1   1   2   1   2   1   1   1   1   3   2   2   1   3   1   1   1
1   2
## 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298
299 300
##   3   1   1   3   2   3   1   1   1   1   2   1   1   2   3   2   2   1
1   3
## 301 302 303
##   2   1   1
##
## Within cluster sum of squares by cluster:
## [1] 1353.325 1350.780 1319.746
##  (between_SS / total_SS =  84.1 %)
##
## Available components:
##
##  [1]  "cluster"          "centers"          "totss"             "withinss"
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

Después de haber realizado la agrupación de datos, puede ser relevante, para investigaciones futuras, añadir una columna en el conjunto de datos que identifique a qué clúster ha sido asignada cada unidad experimental.

```
## Warning: package 'dplyr' was built under R version 4.1.3

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

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

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

##   age oldpeak cluster
## 1  63     2.3       2
## 2  37     3.5       3
## 3  41     1.4       3
## 4  56     0.8       1
## 5  57     0.6       1
## 6  57     0.4       1
```

A continuación, representaremos los grupos en un gráfico para visualizar la asignación de cada país a su respectivo grupo.



Lo que hemos hecho ha sido agrupar a las personas por edad y la depresión del ST inducida por el ejercicio en relación con el reposo. En la zona azul del cluster plot se encuentran las personas con menos edad, en la zona roja se encuentran las personas con una edad intermedia, entre 50 y 60 años. Por último, la zona verde que representa los datos de las personas ancianas.

Comparación de Medias de Dos poblaciones

En este apartado, vamos a estudiar la relación entre una variable cuantitativa y una cualitativa con el fin de analizar la base de datos más profundamente.

¿Existen Diferencias Significativas entre el Sexo de una persona y la probabilidad de enfermedad cardíaca?

Nuestra variable objeto de estudio será la probabilidad de una enfermedad cardíaca (output), mientras que, el sexo será el factor con dos niveles de estudio: Femenino y Masculino. Para esto. es necesario comprobar que la variable destinada a factor sea de clase "factor".

```
## [1] "character"
```

Al comprobar que la clase no es factor, debemos cambiarla.

```
## [1] "factor"
```

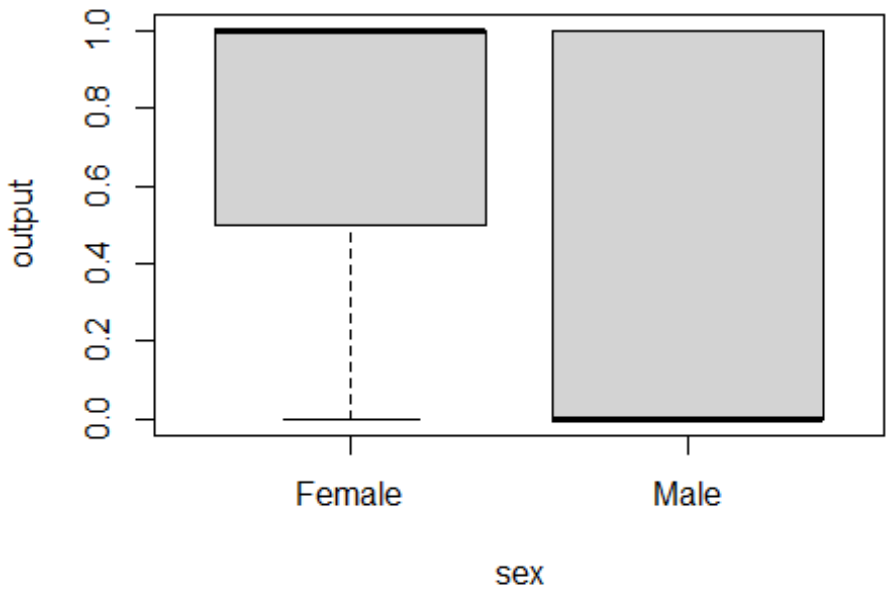
A continuación, analizaremos los datos con el comando summary para estudiar el tamaño muestral de la variable en cada población:

```
##      sex          output
##  Female: 96   Min.   :0.0000
##  Male  :207   1st Qu.:0.0000
##               Median :1.0000
##               Mean   :0.5446
##               3rd Qu.:1.0000
##               Max.   :1.0000
```

Como podemos observar, el tamaño muestral de los hombres es de 207, mientras que el de las mujeres es de 96. En ambos casos el tamaño muestral es muy grande, por lo que no es necesario analizar la normalidad.

Antes de abordar la resolución de los tests de hipótesis, podemos llevar a cabo un análisis descriptivo examinando las medias muestrales de la variable en estudio para cada nivel del factor. Además, podemos visualizar estos datos utilizando un diagrama de cajas por niveles.

```
## Warning: package 'car' was built under R version 4.1.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.1.3

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

##    Female      Male
## 0.7500000 0.4492754
```



A la vista del diagrama de cajas y de los valores numéricos de las medias parece que existe una diferencia notable en la probabilidad de tener una enfermedad cardíaca entre diferentes sexos. Para confirmar esto debemos resolver el test t-Student previo análisis de igualdad de varianzas.

Homocedasticidad o Igualdad de Varianzas

Para estudiar si existe igualdad de varianzas en ambas poblaciones aplicamos el test de Levene con la intención de resolver el siguiente test:

$\{H_0: \sigma_1 = \sigma_2 \; H_1: \sigma_1 \neq \sigma_2$

```
## Levene's Test for Homogeneity of Variance (center = "mean")
##        Df F value   Pr(>F)
## group   1  56.409 6.74e-13 ***
##       301
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Una vez realizado el test de Levene y tras obtener un p-valor menor que 2.2e-16 siendo menor que 0,05 rechazamos H0 y, por tanto, no existe igualdad de varianzas en ambas poblaciones. Ahora realizaremos el test de la t-Student para resolver el contraste de igualdad de medias.

$\{H_0: \mu_1 = \mu_2 \; H_1: \mu_1 \neq \mu_2$

```
##
##  Welch Two Sample t-test
##
## data:  output by sex
## t = 5.3372, df = 209.95, p-value = 2.44e-07
## alternative hypothesis: true difference in means between group Female and
group Male is not equal to 0
## 95 percent confidence interval:
##  0.1896497 0.4117996
```

```
## sample estimates:
## mean in group Female   mean in group Male
##            0.7500000            0.4492754
```

Tras realizar el t-test obtenemos un p-valor menor que 2.2e-16 siendo menor que 0,05 por lo que rechazamos H0 y, también la igualdad de medias, afirmando que existen diferencias significativas entre la probabilidad de tener enfermedades cardíacas en diferentes sexos.

Conclusión

Una vez concluido el estudio de los datos de Ataques al corazón (Heart Attack), podemos decir que las personas con mayor edad tienen mayor riesgo de desarrollar enfermedades cardiovasculares. Por otro lado, hemos visto que existe una diferencia significativa entre hombres y mujeres a la hora de sufrir un infarto. Esto puede deberse a que las mujeres pueden experimentar síntomas de ataque cardíaco diferentes a los de los hombres, y estos síntomas pueden ser menos reconocidos o atribuidos erróneamente a otras condiciones de salud. Las mujeres pueden presentar síntomas más sutiles o atípicos, como fatiga, falta de aliento, náuseas o dolor en la espalda, en lugar del típico dolor en el pecho. Estas diferencias en los síntomas y en la percepción de riesgo a veces pueden llevar a un subdiagnóstico y, en consecuencia, a un tratamiento más tardío en mujeres.

Por último, en los componentes principales, especialmente en el segundo de ellos, se observa que, en términos generales, cuando la presión arterial disminuye, el cuerpo responde incrementando la frecuencia cardíaca para contrarrestar la reducción del flujo sanguíneo.