Influencia del Red Bull en la frecuencia cardíaca

María Vázquez Jiménez

2025-05-22

Índice

Introducción

En este proyecto se va a analizar cómo afecta el **consumo de Red Bull** a la **frecuencia cardíaca** de los individuos, cuyos datos se han obtenido gracias a la plataforma Kaggle. Para analizar dichos datos se han empleado diversas **herramientas de R y del entorno Tidyverse**, utilizando tanto **modelos estadísticos** como **gráficos exploratorios**, que ayudan a comprender cuáles son las variables que afectarán al hecho de consumir esta bebida energética.

Entorno

Se ha utilizado el lenguaje de programación R debido a que se emplea bastante a la hora del análisis estadístico y de datos por su flexibilidad, potencia y gran ecosistema de paquetes. En relación a esto último, se han empleado diversas librerías que pueden agruparse en distintos entornos funcionales:

Entorno Tidyverse:

dplyr: manipular y transformar eficientemente los datos.

tidyr: reorganizar y limpiar estructuras de datos.

readr: ayuda a la lectura rápida y estructurada de archivos CSV.

ggplot2: crear gráficas claras y personalizables.

Manejo de rutas y estructuras del proyecto:

here: facilitar el trabajo con rutas relativas dentro del proyecto.

Modelado estadístico y machine learning:

nnet: realizar regresiones logísticas multinomiales.

caret: dividir los datos en conjuntos de entrenamiento y prueba.

rpart: construir árboles de decisión.

rpart.plot: visualizar de forma clara y directa dichos árboles.

splines: incluir funciones spline en modelos estadísticos para capturar relaciones no lineales.

Datos

Tabla de datos

La tabla de datos, extraída de la página web de Kaggle, tiene un total de 5 filas y 120 columnas.

Resumen estadístico

Volunteer_ID	sex	agegrp	bp_before	bp_after
Min. : 1.00	Length:120	Length:120	Min. :138.0	Min. :145.0
1st Qu.: 30.75	Class :character	Class :character	1st Qu.:147.0	1st Qu.:160.0
Median : 60.50	Mode :character	Mode :character	Median :154.5	Median :169.0
Mean : 60.50	NA	NA	Mean :156.4	Mean :170.6
3rd Qu.: 90.25	NA	NA	3rd Qu.:164.0	3rd Qu.:178.2
Max. :120.00	NA	NA	Max. :185.0	Max. :204.0

El resumen estadístico muestra una visión general de las características principales de las variables incluidas en el conjunto de datos, estas son:

Volunteer_ID: Es una variable numérica que sirve como identificador de cada individuo. Aunque no tiene un significado estadístico, observamos se corresponden a los 120 individuos que se analizaron para este conjunto de datos.

bp_before y bp_after: Representan la frecuencia cardíaca de los individuos antes y después de tomar Red Bull. Comparando la media y la mediana de ambas variables vemos que los valores de frecuencia cardíaca tienden a aumentar tras el consumo de la bebida. Por otro lado, tanto los cuartiles como los máximos muestran un desplazamiento hacia valores más altos después del consumo, lo que puede significar que dicho aumento se deba un efecto fisiológico.

sex y agegrp: Son variables categóricas las cuales nos muestran a que rango de edad pertenecen los individuos muestreados (este va de 30-45, 46-59 y 60 o más) y el sexo de los mismos.

Análisis sobre aquellos individuos a los que les afecta más tomar Red Bull

Volunteer_ID	sex	agegrp	bp_before	bp_after	diferencia
17	Male	30-45	141	162	21
97	Female	46-59	144	169	25
114	Female	60+	151	177	26
73	Female	30-45	141	168	27
57	Male	60+	147	176	29
28	Male	46-59	142	183	41

Se puede apreciar que al individuo al cual le afecta más tomar Red Bull es a un hombre de entre 46-59 años, pues presenta una diferencia de 42 lpm entre la frecuencia cardíaca antes y después de tomar Red Bull.

Por consiguiente, nos vamos a centrar en dicho rango de edad.

Individuos de entre 46-59 años a los que les afecta más el tomar Red Bull

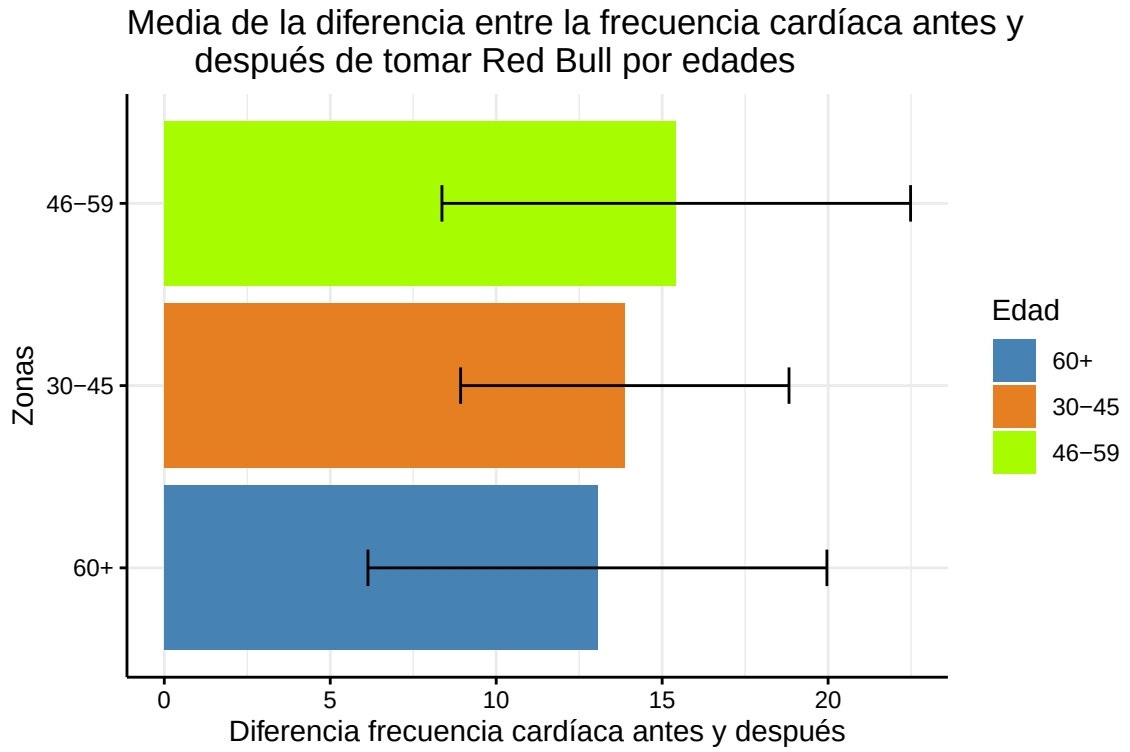
Volunteer_ID	sex	agegrp	bp_before	bp_after
39	Male	46-59	185	204
32	Male	46-59	184	202
31	Male	46-59	175	193
35	Male	46-59	170	187
82	Female	46-59	170	187

Observamos que los individuos más afectados por el consumo de Red Bull son hombres. Posteriormente veremos si el sexo influye en la frecuencia cardíaca a la hora de tomar dicha bebida energética.

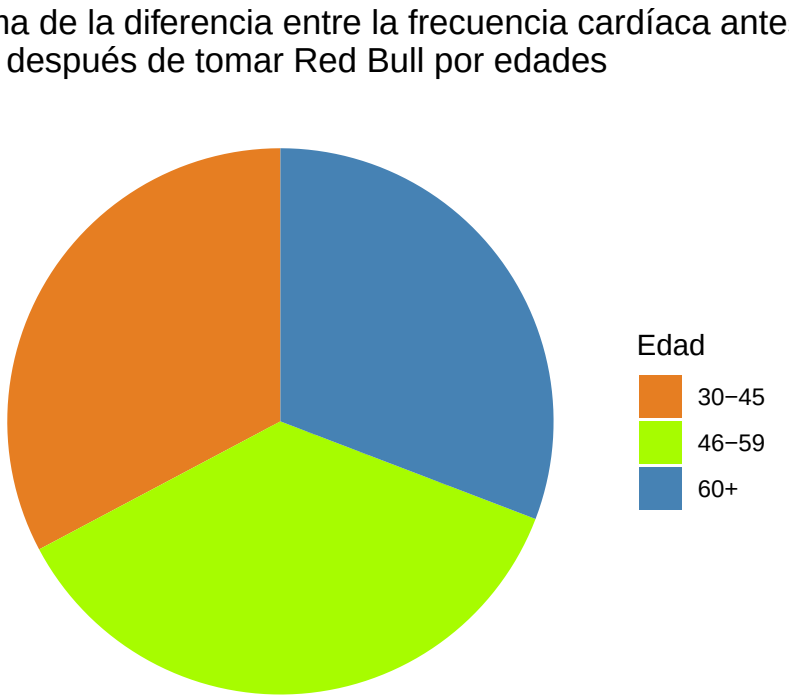
Visualización de los datos

Estas son algunas gráficas que facilitarán la comprensión de nuestro conjunto de datos, observando si hay alguna relación entre las distintas variables del experimento en cuestión.

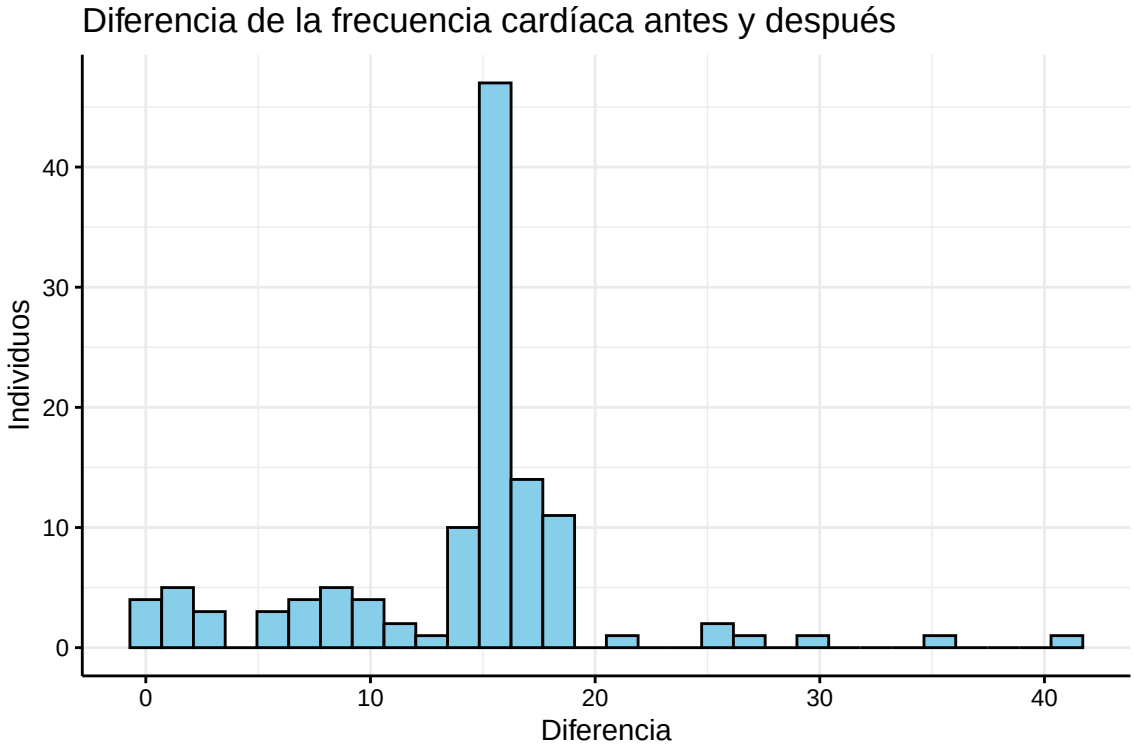
Gráfico de barras

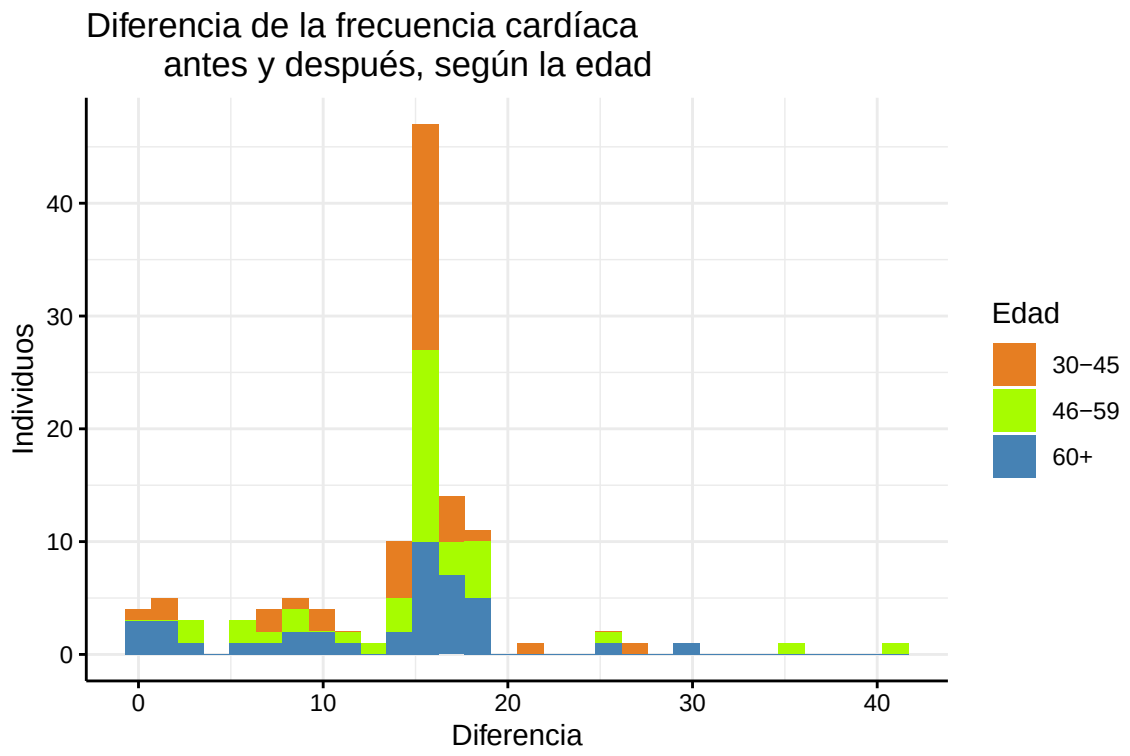


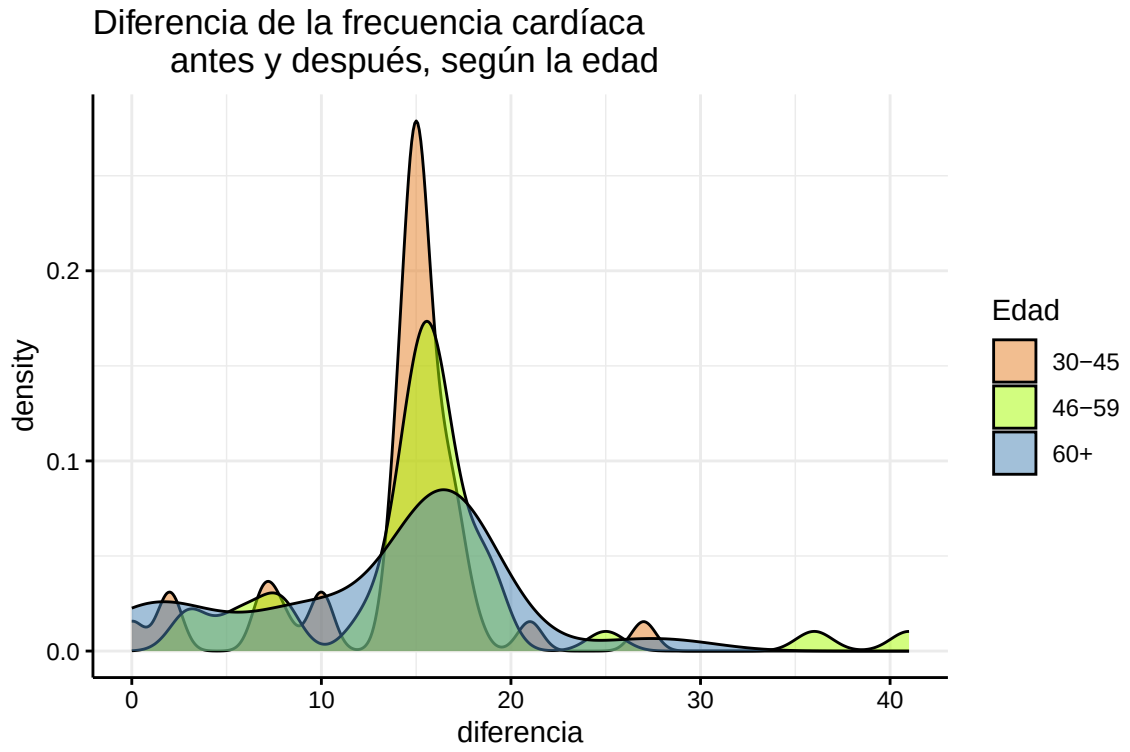
Gráfica de sectores



Histogramas

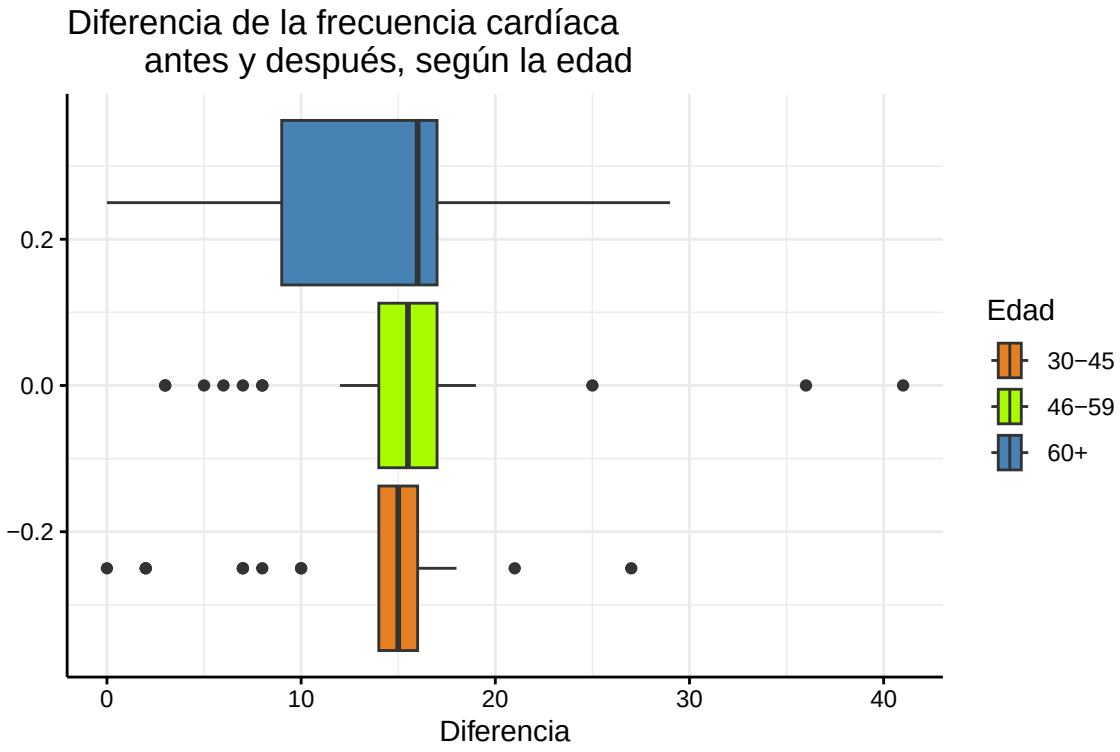




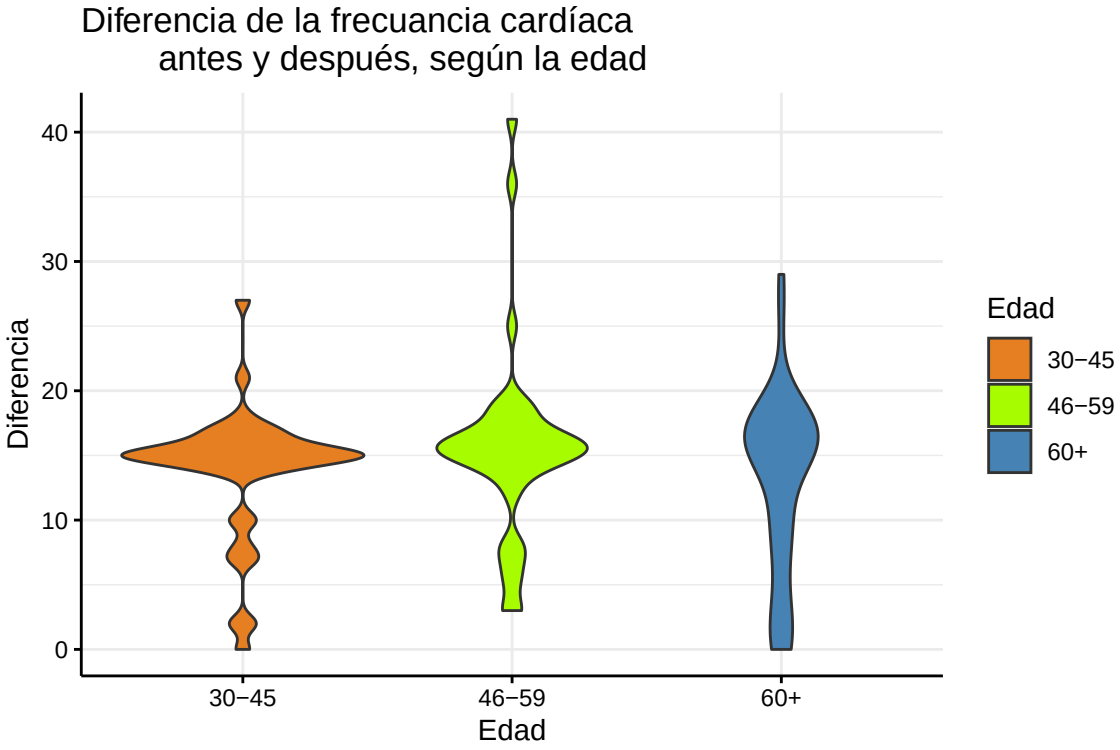


Podemos ver que la mayoría de los individuos de entre 30 y 35 años presentan una diferencia de 15 lpm en la frecuencia cardíaca. Por otro lado, los individuos a quienes más les ha afectado el consumo de Red Bull son unos pocos que pertenecen al grupo de edad de entre 46 y 59 años, representando incluso diferencias cercanas a 40 lpm.

Boxplot



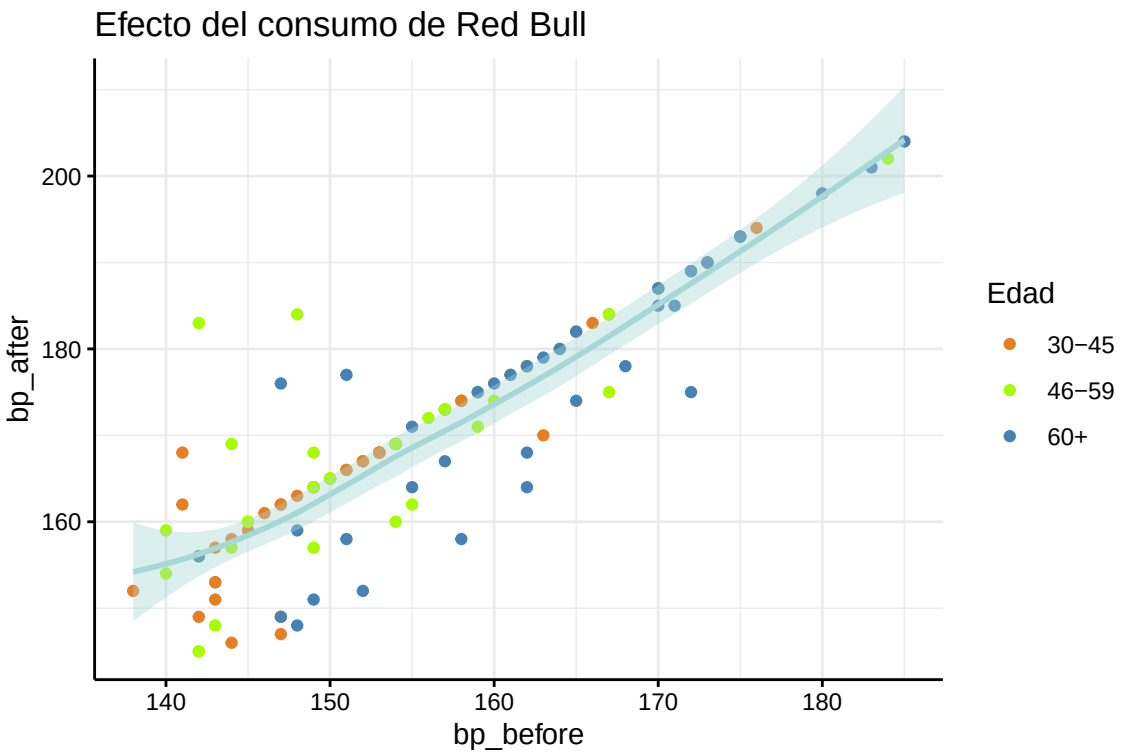
Gráfica de violines



Gráficas clave

Veamos ahora las gráficas que más han llamado la atención.

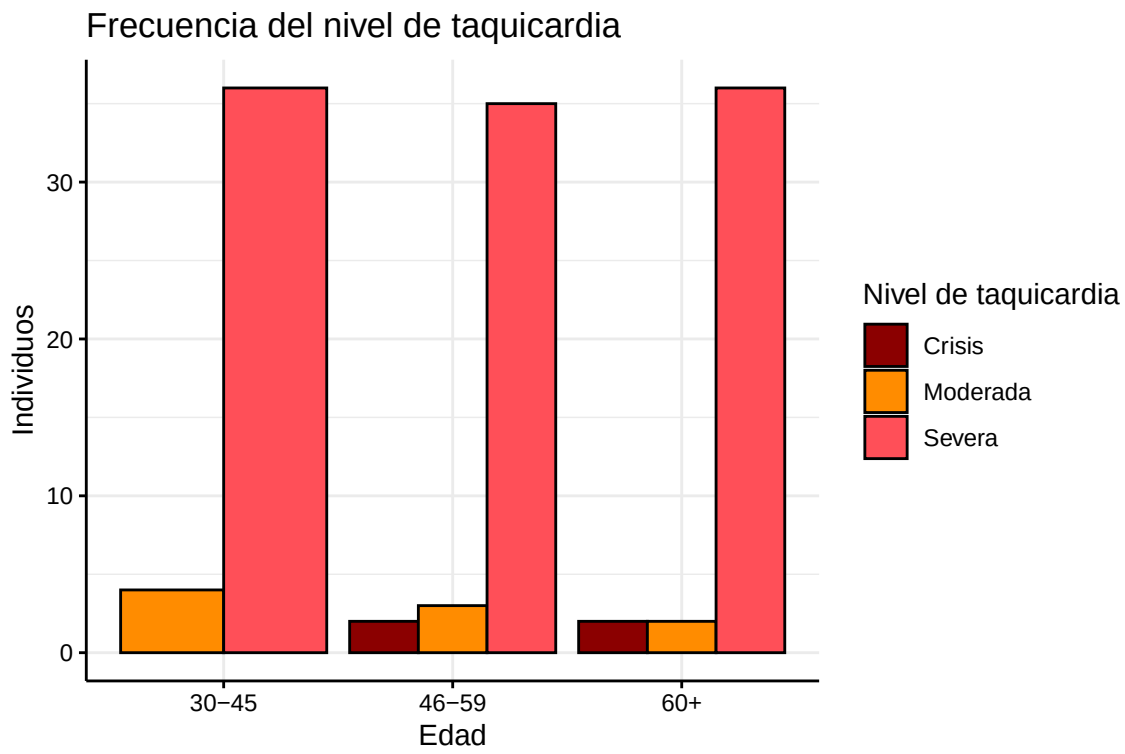
Gráfica de dispersión con una curva de tendencia añadida.



La curva se ha generado utilizando geom_smooth con el método LOESS (Locally Estimated Scatterplot Smoothing), el cual es un método de suavizado no paramétrico que se usa para representar la tendencia general de los datos sin asumir una forma relación específica entre las distintas variables. La relación entre la frecuencia cardíaca antes y después de tomar Red Bull es directa, lo que significa que ambas son directamente proporcionales.

Gráfica de barras agrupadas.

Esta gráfica nos muestra los distintos niveles de taquicardia que puede presentar un individuo después de tomar Red Bull, dependiendo de su edad y sexo, siendo estos niveles: Leve = 100-120, Moderada = 121-150, Severa = 151-200, Crisis = +200



La mayoría de los individuos presenta taquicardia severa tras consumir Red Bull. En cambio, los individuos de entre 46 y 59 años, así como los mayores de 60, presentan taquicardia crítica con la misma frecuencia. Por otro lado, ningún individuo desarrolla taquicardia leve después de consumir Red Bull. La taquicardia moderada, aunque poco común en los tres grupos de edad, se observa con mayor frecuencia en el grupo de 30 a 45 años y con menor frecuencia en el de mayores de 60 años.

Modelos

Sabiendo la frecuencia cardíaca y la edad de un individuo, buscamos deducir cuánta frecuencia cardíaca tendrá después de tomar Red Bull.

Modelo lineal predictivo

A continuación se muestra el modelo lineal predictivo que se ha realizado con el 50% de los datos de los individuos que pertenecen a cada sexo y a cada rango de edad. Hemos cogido el 50% de los datos para reservar el resto con el fin de comprobar si el modelo se ajusta correctamente con los datos que no se han observado.

Resumen del modelo

```
##
## Call:
## lm(formula = bp_after ~ bp_before + agegrp + sex, data = Tabla)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.218  -1.066   1.514   2.352  16.431
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.56867   11.62230  -0.909    0.367
## bp_before     1.14989    0.07717  14.901   <2e-16 ***
## agegrp46-59   1.46307    1.95920   0.747    0.458
## agegrp60+     0.36351    2.04545   0.178    0.860
## sexMale       0.74063    1.62644   0.455    0.651
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.062 on 55 degrees of freedom
## Multiple R-squared:  0.835,  Adjusted R-squared:  0.823
## F-statistic: 69.59 on 4 and 55 DF,  p-value: < 2.2e-16
```

A la vista de los coeficientes anteriores, podemos ver que la frecuencia cardíaca después de tomar Red Bull aumenta en 1.15 lpm por cada lpm que aumenta la frecuencia cardíaca antes de tomar Red Bull. Además, la frecuencia cardíaca después de tomar Red Bull aumenta en 1.46 lpm si el individuo pertenece al grupo de edad de 46-59 años, y aumenta 0.36 lpm si el individuo pertenece al grupo de edad de 60 años o más, todo esto respecto del caso de los individuos de entre 30-45 años. Por último, si el individuo es un hombre, su frecuencia cardíaca aumentará 0.74 con respecto a una mujer.

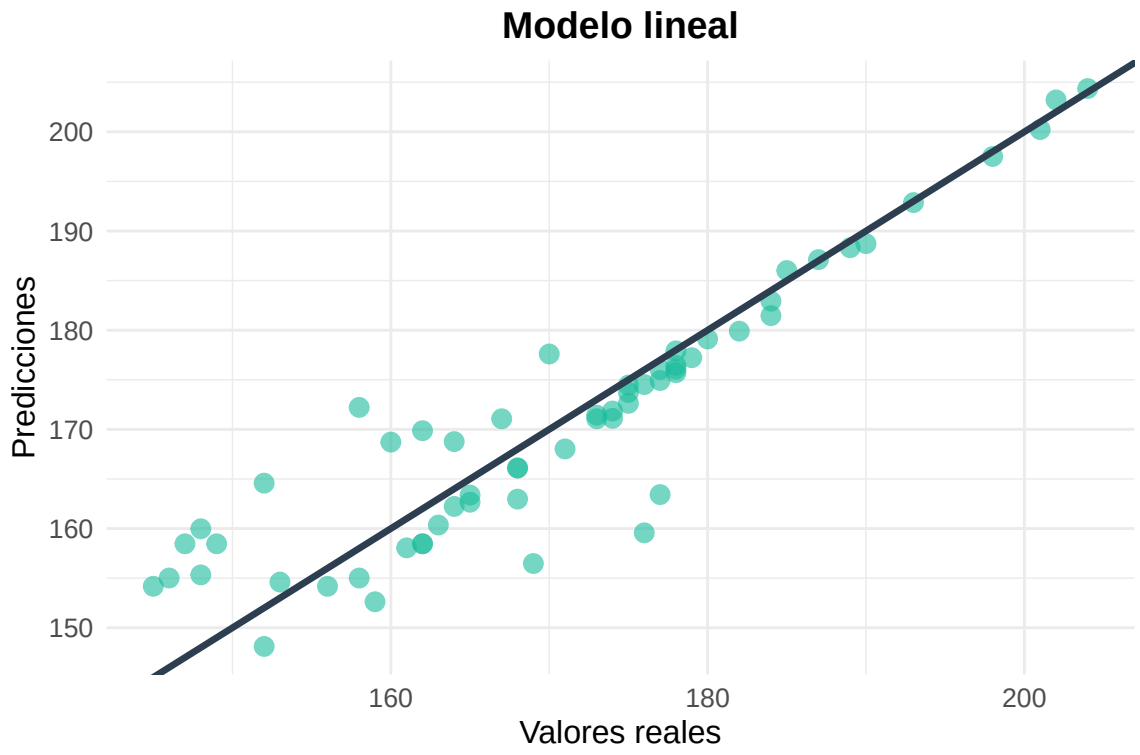
El modelo lineal tendría la siguiente forma:

$$y = alfa * bp_before + beta * agegrp46 - 59 + gamma * agegrp60 + mu * sexMale$$

$$y = -10.56867 + 1.14989 * bp_before + 1.46307 * agegrp46 - 59 + 0.36351 * agegrp60 + 0.74063 * sexMale$$

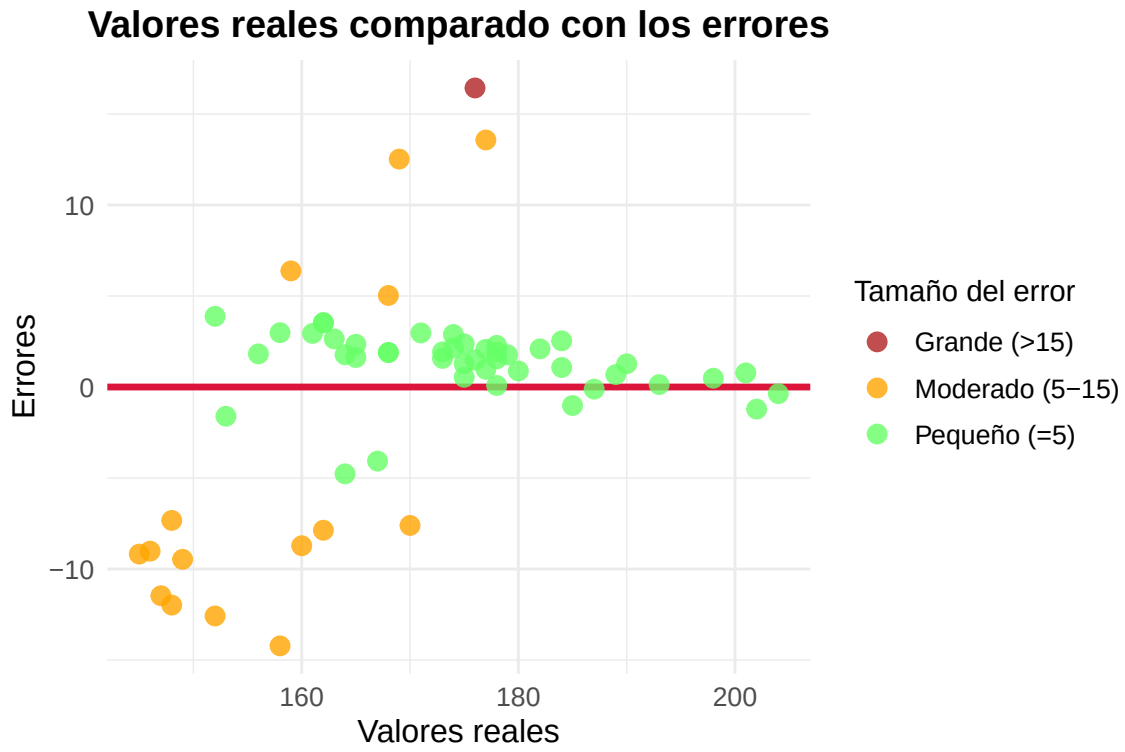
Al tener un R cuadrado de **0.823** significa que el modelo explica el 82.3% de la variabilidad en la presión arterial posterior, es decir, que el modelo es bastante bueno a la hora de predecir la presión arterial después de tomar Red Bull. Por otro lado, en este modelo, solo la variable `bp_before` muestra una relación estadísticamente significativa con la presión arterial posterior, ya que el p-valor es menor que 0.05, mientras que las variables de edad y sexo no presentan evidencia suficiente para respaldar que influyen de manera significativa en la variable dependiente (bp_after), puesto que el p-valor es mayor que 0.05.

Gráfica



En la gráfica anterior se puede observar que el modelo lineal predictivo se ajusta bastante bien a los datos, ya que la mayoría de los puntos están cerca de la recta y=x. Sin embargo, los puntos de aquellas frecuencias cardíacas comprendidas entre 150 y 177 lpm se desvían bastante de la línea de regresión, lo que indica que el modelo no es perfecto y que hay otros factores que pueden estar influyendo en la frecuencia cardíaca después de tomar Red Bull y que no los estamos teniendo en cuenta, por lo tanto, más adelante se desarrollará un modelo que compare el hecho de añadir más variables.

Gráfica: valores reales comparados con los errores



Podemos observar que la mayoria de datos se han predicho correctamente, pero esto no impide que existan varios puntos que se desvían significativamente de la línea, sobre todo uno que varía más de 15 lpm de la frecuencia cardíaca real de dicho individuo, por tanto, como hemos dicho anteriormente, posteriormente se aplicará un modelo que compare el hecho de añadir más variables.

Modelos con splines (interpolación)

Vamos a analizar el hecho de añadir más variables a un modelo distinto, en este caso será un modelo con splines el que nos diga si realmente es necesario añadir más variables para tener un gráfico más representativo sobre cómo afecta el Red Bull en las personas de distinto sexo y rango de edad. Para ello también hemos cogido el 50% de los datos de los individuos que pertenecen a cada sexo y a cada rango de edad, pues el resto nos servirá para comprobar si el modelo se ha ajustado bien al resto de los datos.

Resumen del modelo al que simplemente se ha comparando la frecuencia cardíaca antes y después.

```
##
## Call:
## lm(formula = bp_after ~ bs(bp_before, df = 4), data = Tabla)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.7297  -0.7328   1.2563   2.3876  16.9987
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)             151.415      4.545  33.314  < 2e-16 ***
## bs(bp_before, df = 4)1    3.987      7.543   0.529 0.599222
## bs(bp_before, df = 4)2   21.980      6.058   3.628 0.000626 ***
## bs(bp_before, df = 4)3   43.289      8.668   4.994 6.32e-06 ***
## bs(bp_before, df = 4)4   51.810      5.783   8.960 2.45e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.073 on 55 degrees of freedom
## Multiple R-squared:  0.8344, Adjusted R-squared:  0.8223
## F-statistic: 69.27 on 4 and 55 DF,  p-value: < 2.2e-16
```

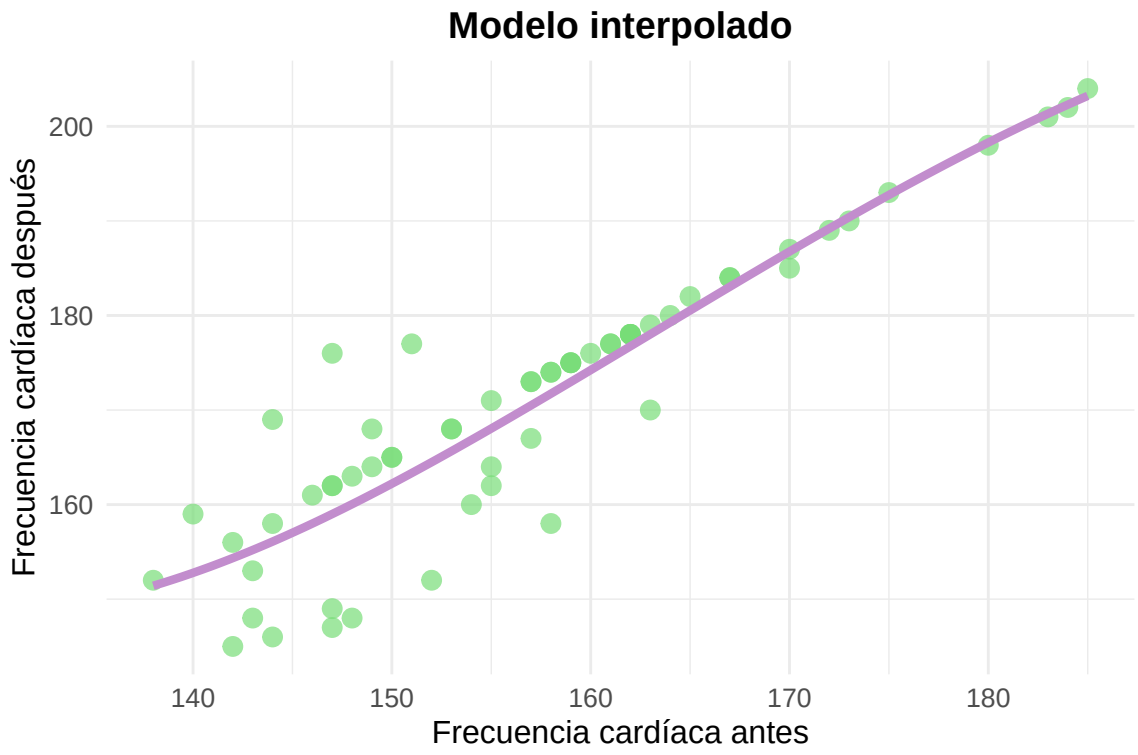
El modelo, al tener un R cuadrado de **0.8223**, presenta un buen ajuste para predecir cómo varía la frecuencia cardíaca tras el consumo de Red Bull. Además, al tener un p-valor menor a 0.05, podemos decir que los resultados son significativos, lo que significa que el modelo realmente tiene una relación importante con la variable que estamos estudiando.

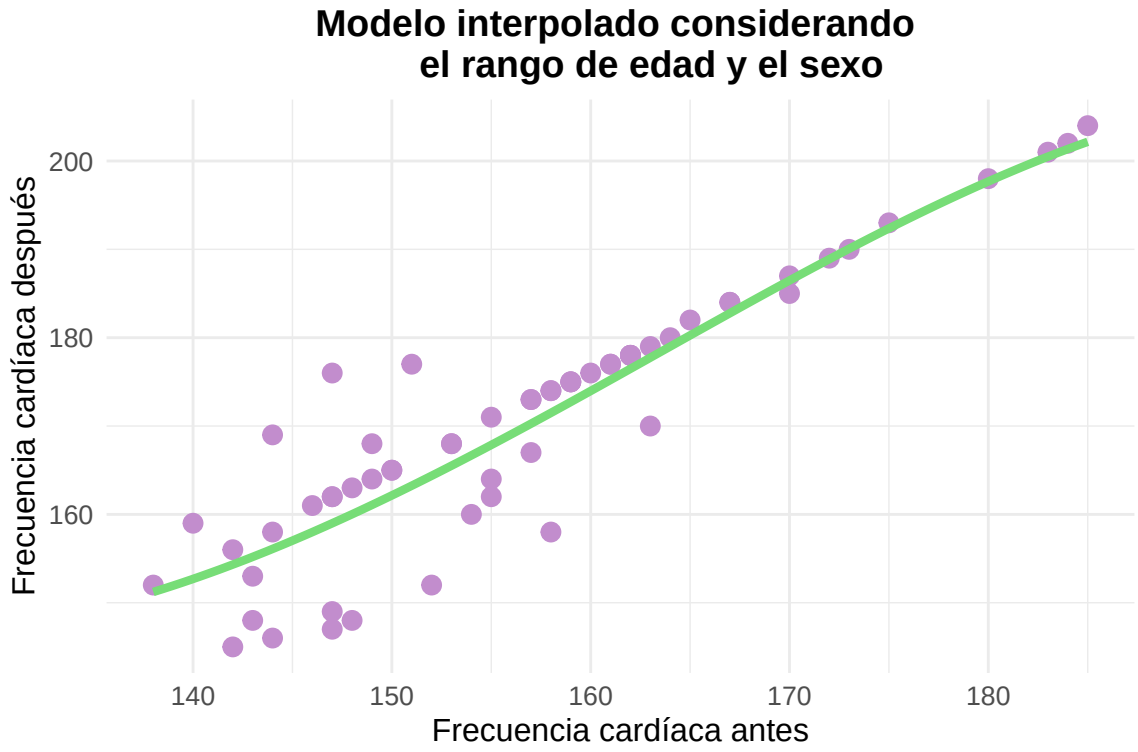
Resumen del modelo al que se le han añadido las variables sexo y rango de edad.

```
##
## Call:
## lm(formula = bp_after ~ bs(bp_before, df = 4) + agegrp + sex,
##     data = Tabla)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.957  -1.161   1.541   2.493  16.542
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             150.3746     4.8120  31.250  < 2e-16 ***
## bs(bp_before, df = 4)1   4.4034     7.7520   0.568  0.57246
## bs(bp_before, df = 4)2  21.5381     6.3985   3.366  0.00144 **
## bs(bp_before, df = 4)3  43.5042     8.9801   4.845 1.18e-05 ***
## bs(bp_before, df = 4)4  50.9540     6.1374   8.302 4.25e-11 ***
## agegrp46-59              1.4229     2.0230   0.703  0.48498
## agegrp60+                0.4669     2.1120   0.221  0.82591
## sexMale                  0.8445     1.6726   0.505  0.61576
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.201 on 52 degrees of freedom
## Multiple R-squared:  0.8367, Adjusted R-squared:  0.8148
## F-statistic: 38.07 on 7 and 52 DF,  p-value: < 2.2e-16
```

Al tener un R cuadrado de **0.8148** podemos decir que el modelo indica un buen poder predictivo. Sin embargo, al comparar este valor con el del modelo anterior ($R^2 = 0.8223$), se observa una ligera disminución, lo que sugiere que la inclusión de las variables “rango de edad” y “sexo” no ha mejorado el ajuste del modelo, sino que lo ha empeorado ligeramente. Por otro lado, el hecho de que el modelo tenga un p-valor menor a 0.05 indica que es estadísticamente significativo.

Gráficas realizadas





Comparando ambas gráficas vemos que varía muy poco, pero ese poco puede ser determinante, por lo que es importante tener en cuenta todos los factores posibles.

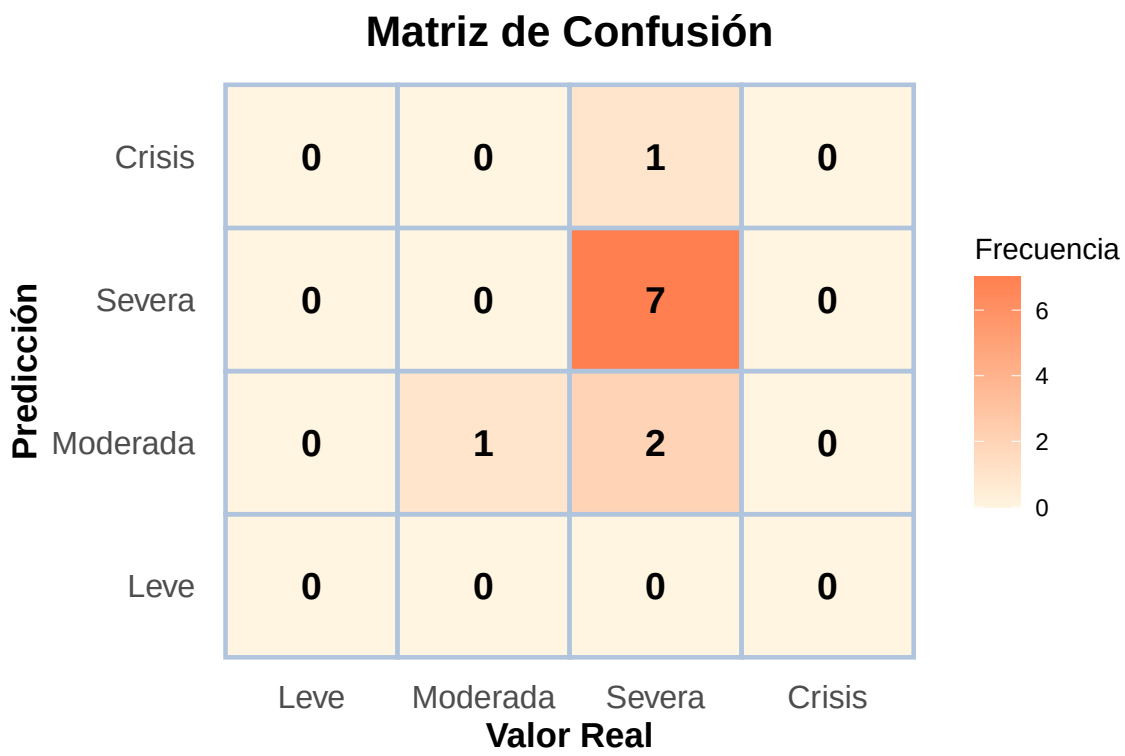
Árbol de decisión

Este es un tipo de modelo predictivo el cual utiliza una estructura de árbol para representar distintas decisiones y sus posibles consecuencias. A continuación, lo utilizaremos para clasificar a los individuos según su frecuencia cardíaca después de consumir Red Bull, teniendo en cuenta las distintas variables del conjunto de datos.

Este modelo se ha realizado teniendo un 80% de los datos se separó para el entrenamiento de dicho modelo.

Podemos observar como aquellos que tienen un número mayor o igual a 179 lpm, antes de tomar Red Bull, presentan una taquicardia crítica. Por otro lado, aquellos que tenían un número menor a 179 lpm y mayor o igual a 149 presentan una taquicardia severa. Por último, aquellos que tenían un número menor a 149 lpm taquicardia moderada si son mujeres, mientras que aquellos que pertenecen al sexo masculino presentan una taquicardia severa.

Matriz de confusión



La matriz de confusión indica que el modelo clasificó correctamente al único individuo que presentó taquicardia moderada. Sin embargo, clasificó incorrectamente a tres de los diez individuos que presentaron taquicardia severa después de consumir la bebida, a dos de ellos los clasificó como casos de taquicardia moderada y a uno como taquicardia crítica, cuando en realidad los tres deberían haber sido clasificados como taquicardia severa.

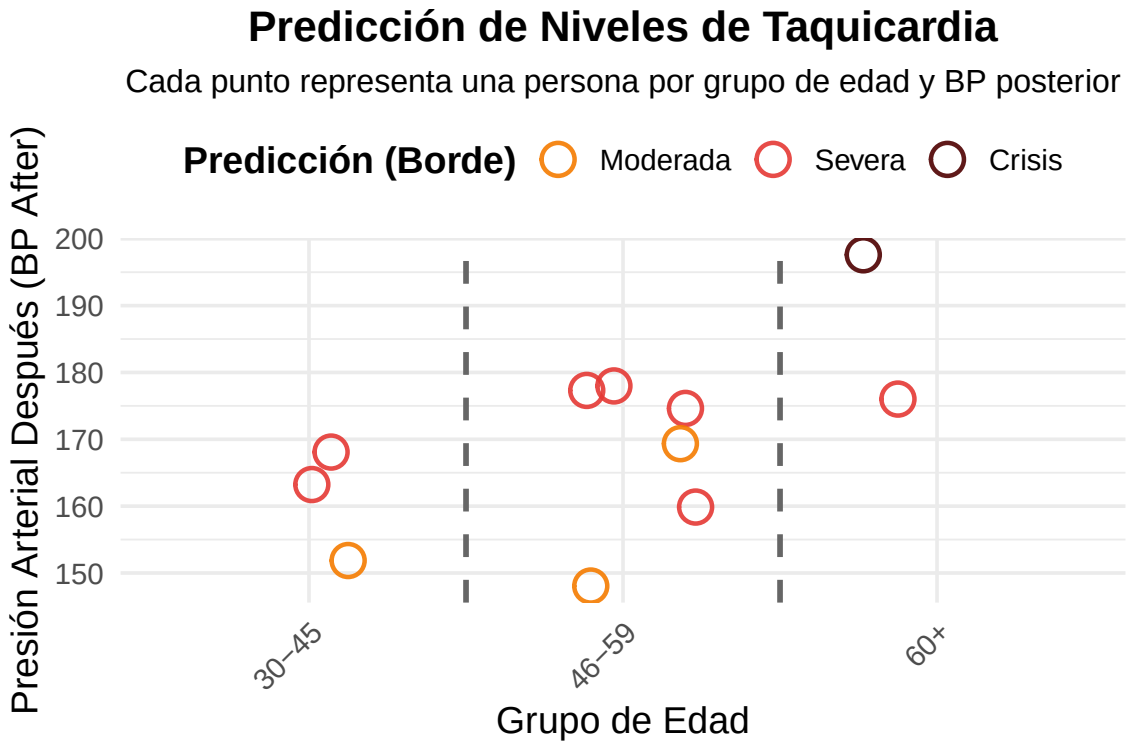
Precisión del modelo

```
## Precisión del modelo:  0.7272727
```

La precisión del modelo es de 0.73 aproximadamente, lo que significa que el modelo es capaz de clasificar correctamente el 73% de los casos. Esto indica que el modelo tiene un buen rendimiento en la clasificación de los casos de taquicardia después de consumir Red Bull, pero que aún hay margen de mejora.

Gráfica de la matriz de confusión

Esta gráfica es una aproximación propia realizada con ggplot2 y que permite observar no solo los aciertos y errores del modelo, sino también cómo se distribuyen las predicciones según la edad y la presión arterial tras el consumo de la bebida.



Inferencia estadística del arbol de decisión

Regresión logística multinomial Este modelo se ha utilizado para predecir la categoría de taquicardia (Leve, Moderada, Severa o Crítica) en función de variables como la frecuencia cardíaca antes de consumir Red Bull, el rango de edad y el sexo. La regresión logística multinomial estima la probabilidad de que una observación pertenezca a cada una de estas categorías, comparándolas con una categoría de referencia, que en este caso es la taquicardia leve.

```
## Call:
## multinom(formula = taquicardia ~ bp_before + agegrp + sex, data = Tabla)
##
## Coefficients:
##        (Intercept) bp_before agegrp46-59  agegrp60+  sexMale
## Severa   -34.37619 0.2394752   0.5020138 -0.5623896 141.2056
## Crisis  -391.77797 2.1793036  24.9317187 10.1429706 126.4745
##
## Std. Errors:
##         (Intercept) bp_before agegrp46-59  agegrp60+    sexMale
## Severa 19.682542699 0.1352097   1.2100586 1.66473866 0.06249529
## Crisis  0.004119922 0.2042455   0.0212453 0.02519449 0.06249529
##
## Residual Deviance: 21.61346
## AIC: 41.61346
```

Se puede observar que a medida que la frecuencia cardíaca aumenta, la probabilidad de tener taquicardia severa o una taquicardia crítica también aumenta. Las personas de 46-59 años tiene una mayor probabilidad de tener una taquicardia severa, mientras que este grupo y el de mayores de 60 años tienen una mayor probabilidad de tener una taquicardia crítica en comparación con el grupo base (30-45 años). El coeficiente es extremadamente alto para hombres, lo que indica que los hombres presentan una probabilidad significativamente mayor de tener taquicardia severa o una taquicardia crítica en comparación con las mujeres (grupo base). Este valor elevado podría indicar un sesgo en los datos o un problema de escalado de las variables.

P-valores

```
##        (Intercept)  bp_before agegrp46-59 agegrp60+ sexMale
## Severa  0.08071852 0.07653758    0.678239 0.7354954       0
## Crisis  0.00000000 0.00000000    0.000000 0.0000000       0
```

Los resultados sugieren que el modelo es sensible a algunas variables, especialmente el sexo y la presión arterial antes de consumir Red Bull, para predecir la taquicardia crítica. Sin embargo, la edad y el sexo no parecen tener un impacto tan relevante en la taquicardia severa.

Conclusiones

En este proyecto se ha analizado cómo afecta el consumo de Red Bull a la frecuencia cardíaca de los individuos, utilizando diversas herramientas de R y del entorno Tidyverse. A continuación se presentan las conclusiones más relevantes:

Aprendizaje

Los **hombres de entre 46 y 59 años son los más afectados** por el consumo de Red Bull, mostrando un incremento significativo en la frecuencia cardíaca, con diferencias de hasta 42 lpm entre el antes y el después del consumo.

Visualización

Aunque la mayoría de los individuos presentan una taquicardia severa después de tomar Red Bull, los **grupos de entre 46-59 años y mayores de 60 años** son los únicos que **llegan a presentar taquicardia crítica**.

El **modelo lineal y el modelo con splines muestran un ajuste similar**, con un R^2 de 0.823 y 0.8223 respectivamente, siendo estos los **más precisos** para explicar cómo influye el Red Bull en la frecuencia cardíaca.

Modelos

Al **añadir las variables de edad y sexo** al modelo con splines, el R^2 disminuyó ligeramente ($R^2 = 0.8148$), lo que sugiere que la inclusión de estas variables no mejora el ajuste del modelo, sino que lo **empeora ligeramente**.

El **árbol de decisión** tiene **dificultades para predecir la taquicardia severa**, lo que indica que pueden faltar factores importantes en el modelo. Aunque la frecuencia cardíaca y el sexo parecen ser variables clave, la **edad no aporta mucho** y podría estar complicando el análisis.

Trabajo Futuro

El árbol intenta disminuir el error dividiendo los datos paso a paso, eligiendo en cada momento la variable que mejor separa los casos. Sin embargo, al no considerar todas las combinaciones posibles, puede perder relaciones más complejas entre las variables. Por eso, **sería útil incluir nuevas variables fisiológicas o ciertos hábitos de consumo para mejorar la precisión en futuros estudios**.