Anexo II

Julio Ladron de Guevara Jimenez

2023-08-21

```
options(repos = c(CRAN = "https://cloud.r-project.org"))
```

```
library(latexpdf)
library(ggplot2)

library(psych)

library(tidyverse)
library(kableExtra)

library(nortest)

library(dplyr)
library(tidyr)

library(WRS2)

library(waffle)
library(vcd)
library(rcompanion)
```

Estudio de significancia de variables y creacion del modelo predictivo

Lo primero que vamos a hacer es cargar el dataset.

Nuestro estudio se basa en realizar un modelo lineal para predecir el riesgo de que una persona sufra un "stroke" dadas unas variables de partida, llamadas predictoras. Stroke es la variable objetivo

```
datos <- read.csv("/home/guincho/Desktop/stroke_corregido.csv")

head(datos)
```

```
##   gender age hypertension heart_disease ever_married     work_type
## 1   Male  67            0             1            1       Private
## 2 Female  61            0             0            1 Self-employed
## 3   Male  80            0             1            1       Private
## 4 Female  49            0             0            1       Private
## 5 Female  79            1             0            1 Self-employed
## 6   Male  81            0             0            1       Private
##   residence_type avg_glucose_level  bmi  smoking_status stroke        imc_str
## 1          Urban            228.69 36.6 formerly smoked      1     Obeso g II
## 2          Rural            202.21   NA    never smoked      1
## 3          Rural            105.92 32.5    never smoked      1      Obeso g I
```

```
## 4          Urban            171.23 34.4          smokes      1      Obeso g I
## 5          Rural            174.12 24.0    never smoked      1 Peso saludable
## 6          Urban            186.21 29.0 formerly smoked      1      Sobrepeso
##   grupo_edad
## 1        65+
## 2      50-64
## 3        65+
## 4      35-49
## 5        65+
## 6        65+
```

```
print(colnames(datos))
```

```
##  [1] "gender"            "age"               "hypertension"
##  [4] "heart_disease"     "ever_married"      "work_type"
##  [7] "residence_type"    "avg_glucose_level" "bmi"
## [10] "smoking_status"    "stroke"            "imc_str"
## [13] "grupo_edad"
```

Las variables son las siguientes: “gender” “age” “hypertension” “heart_disease” “ever_married” “work_type” “residence_type” “avg_glucose_level” “bmi” “smoking_status” “stroke” “imc_str”“grupo.edad”

Nos tenemos que preguntar, cuáles de ellas tienen son estadísticamente significativas para sufrir un accidente cerebrovascular

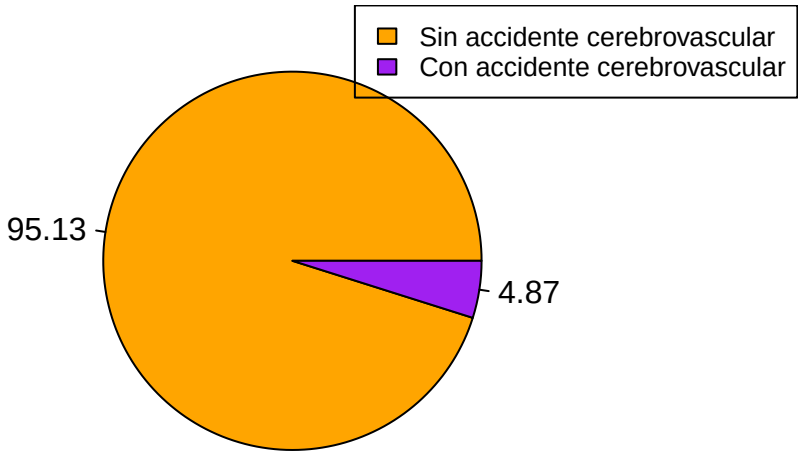
- 1) Variable Objetivo: Stroke
- 2) Variables predictoras categóricas: gender, work_type, Residence_type,smoking_status, hypertension, heart_disease, ever_married, stroke
- 3) Variables numericas continuas: age, avg_glucose_level, bmi

Estudio de significancia de variables

Lo primero que queremos ver es qué porcentaje de la poblacion ha sufrido un accidente cerebrovascular

```
x <- c(sum(datos$stroke == 0), sum(datos$stroke == 1))

etiquetas = c("Sin accidente cerebrovascular", "Con accidente cerebrovascular")

piepercent <- round(100*x/sum(x), 2)

colores = c("orange", "purple")


pie(x, piepercent, main = "Porcentaje de personas que sufren un stroke", col = colores)
legend("topright", etiquetas, cex = 0.8, fill = colores)
```

Porcentaje de personas que sufren un stroke



Un 4.87% de la población ha sufrido un accidente cerebrovascular.

A continuación mostramos la tabla con estadísticos descriptivos de Stroke

A) Estudio de variables continuas en relacion con Accidentes cerebrovasculares

A.1) Edad y accidente cerebrovascular

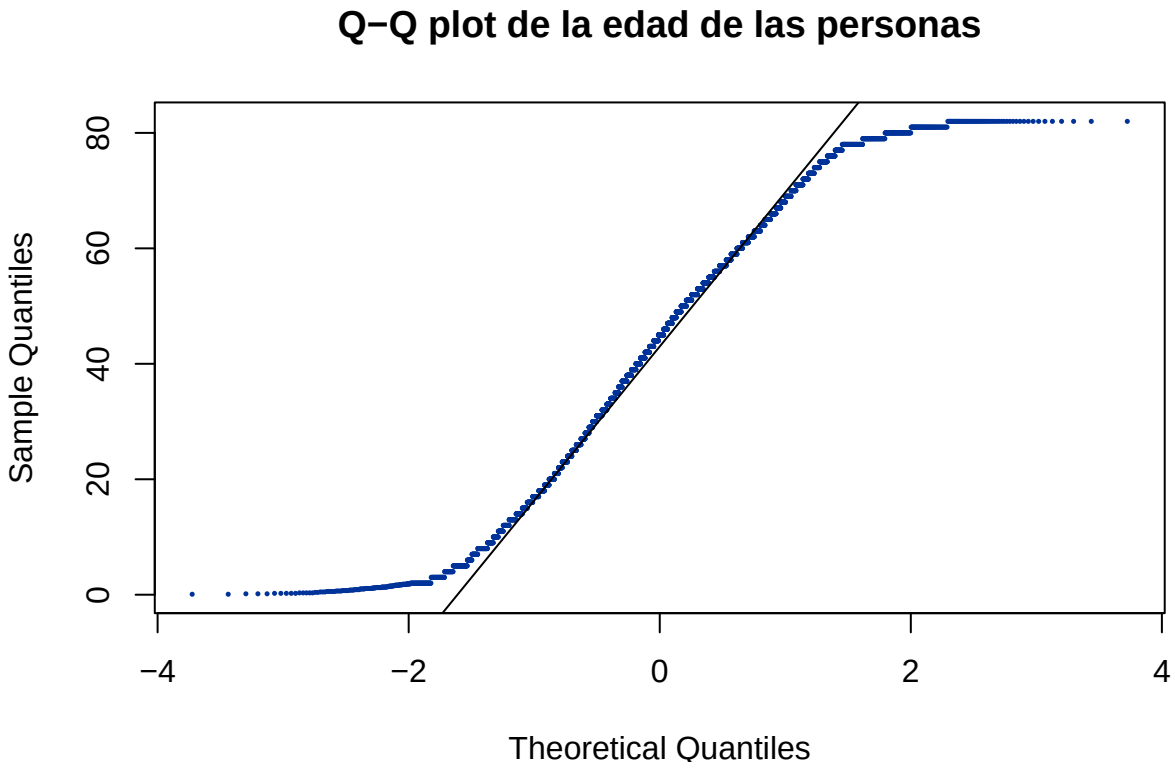
La edad viene representada en nuestro dataset como age.

Vamos a ver si la edad sigue una distribucion normal. Haremos un Q-Q plot como método gráfico y un test de Anderson-Darling como método numérico.

La prueba de Anderson-Darling es una prueba estadística que se utiliza para determinar si un conjunto de datos dado sigue o no una distribución normal.

$$\begin{cases} H_0 : \textit{La muestra sigue una distribución normal} \\ H_1 : \textit{La muestra no sigue una distribución normal} \end{cases}$$

```
stats::qqnorm(datos$age,
              main = "Q-Q plot de la edad de las personas",
              pch = 16, cex = 0.35,
              col = rgb(0, 0.2, 0.6))

stats::qqline(datos$age)
```




```
nortest::ad.test(datos$age )
```

```
##
##  Anderson-Darling normality test
##
## data:  datos$age
## A = 33.856, p-value < 2.2e-16
```

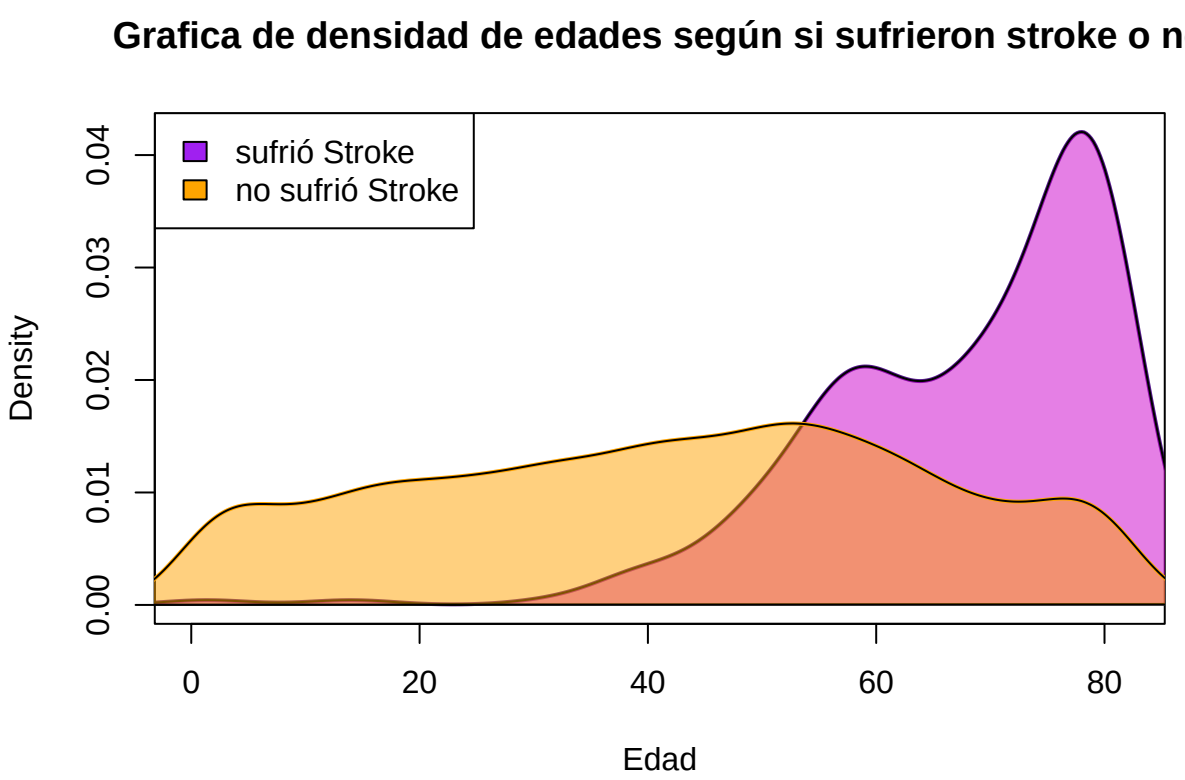
p-value < 2.2e-16

Dado que el valor p es mucho más pequeño que el nivel de significancia utilizado de 0.05, los datos no siguen una distribución normal.

Crearemos una gráfica que muestre ambas funciones de distribución

```
datos_stroke_1 <- subset(datos, stroke == 1)
datos_stroke_0 <- subset(datos, stroke == 0)

# Calculamos la funcion de desnidad de la edad para ambos grupos
densidad_stroke_1 <- density(datos_stroke_1$age)
densidad_stroke_0 <- density(datos_stroke_0$age)

# Creamos la ventana donde irá la grafica

plot(densidad_stroke_1, lwd = 2,
     main = "Grafica de densidad de edades según si sufrieron stroke o no",
     xlab = "Edad", col = "purple4", xlim = c(min(datos$age), max(datos$age)),
     ylim = c(0, max(densidad_stroke_1$y, densidad_stroke_0$y)))

# Densidad para stroke = 1

polygon(densidad_stroke_1, col = rgb(0.8, 0, 0.8, alpha = 0.5))
```

```r
# Densidad para stroke = 0

lines(densidad_stroke_0, lwd = 2, col = "orange1")
polygon(densidad_stroke_0, col = rgb(1, 0.64, 0, alpha = 0.5))

# Leyenda

legend("topleft", legend = c("sufrió Stroke", "no sufrió Stroke"),
       fill = c("purple", "orange"))
```

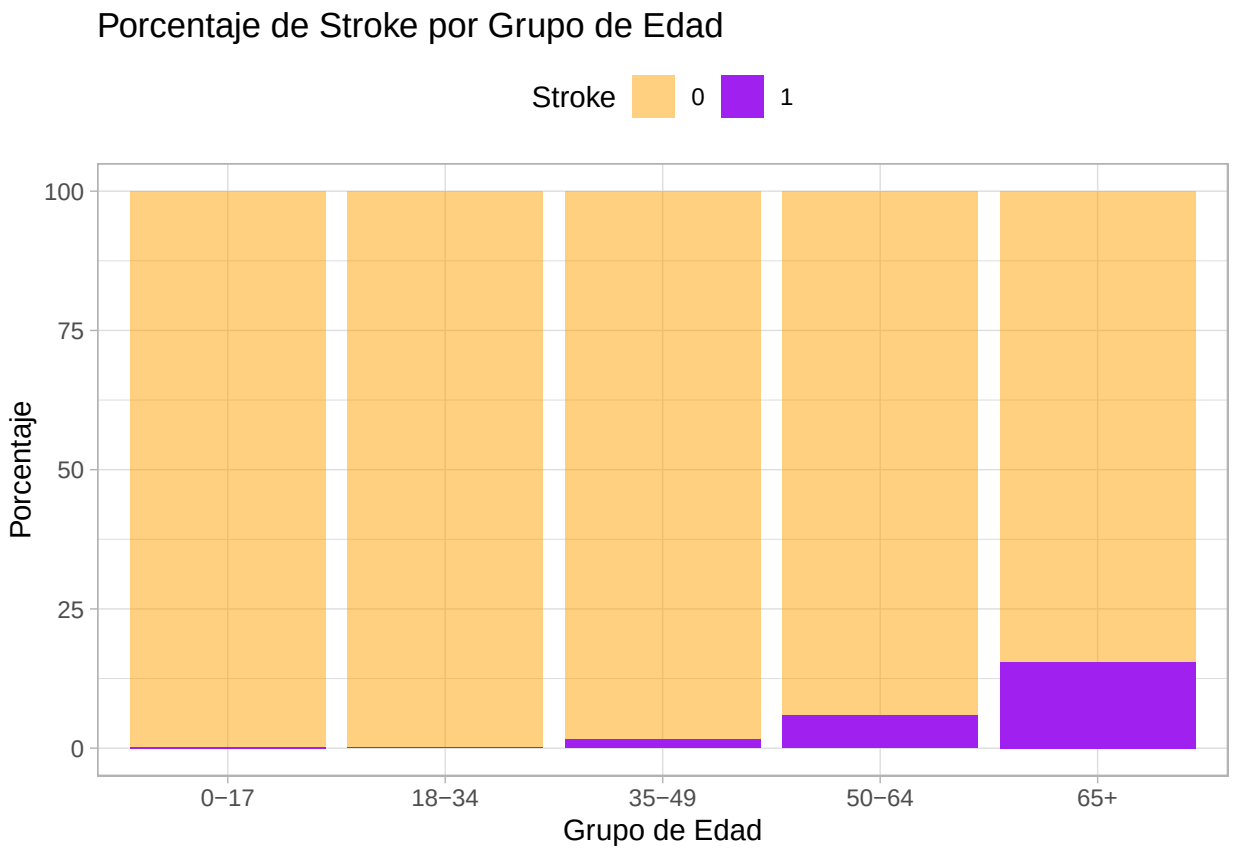


A simple vista, la edad parece tener importancia significativa en la posibilidad de sufrir un stroke

Vamos a ver que cantidad de pesonas sufre un stroke según si pertenece a los diferentes grupos de edad creados en la variable grupo_edad

```r
colores = c(rgb(1, 0.64, 0, alpha = 0.5), "purple")

# Calcular las frecuencias de stroke por grupo de edad
frecuencias <- datos %>%
  group_by(grupo_edad, stroke) %>%
  summarise(count = n()) %>%
  group_by(grupo_edad) %>%
  mutate(percentage = (count / sum(count)) * 100)

# Crear el gráfico de barras
ggplot(frecuencias, aes(x = grupo_edad, y = percentage, fill = factor(stroke))) +
  geom_bar(stat = "identity", position = "stack") +
  labs(x = "Grupo de Edad", y = "Porcentaje", fill = "Stroke") +
  scale_fill_manual(values = colores, name = "Stroke", labels = c("0", "1")) +
  theme_light() +
```

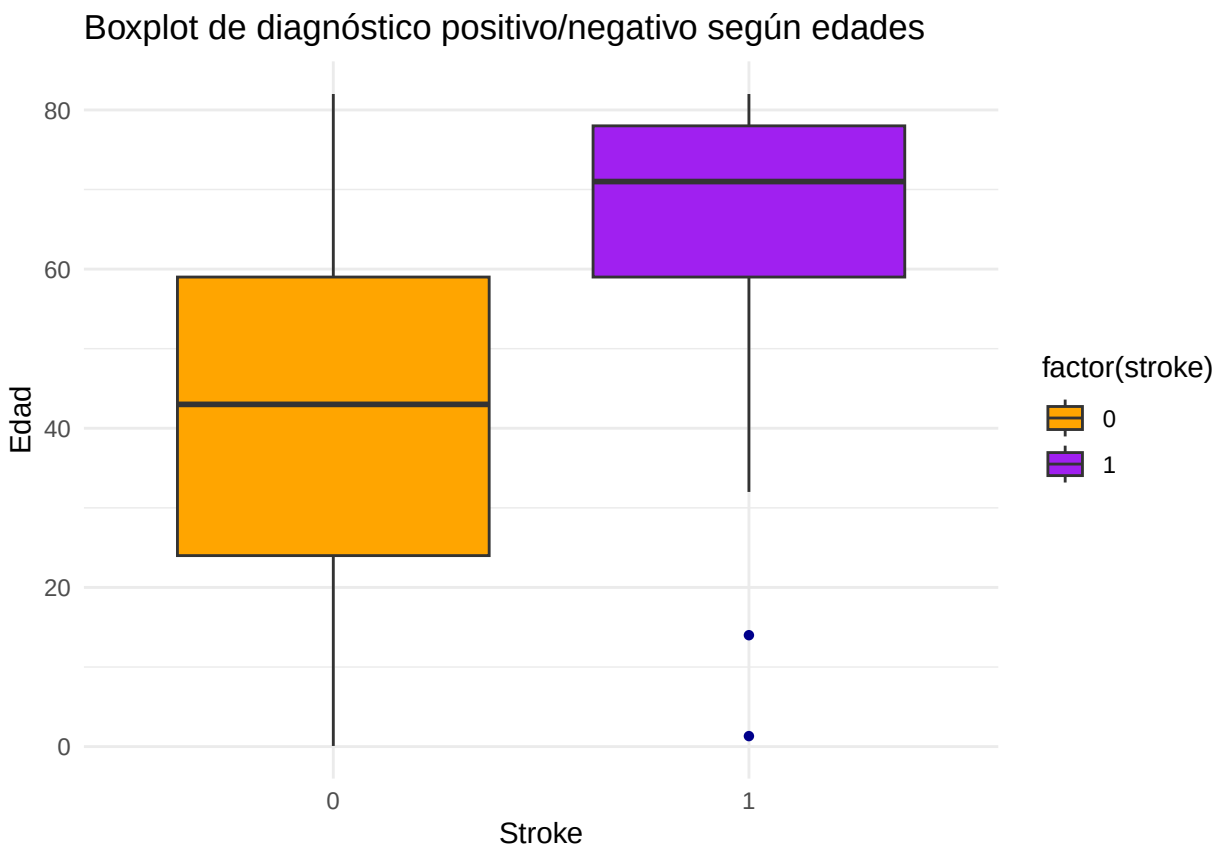
```
  theme(legend.position = "top") +
  ggtitle("Porcentaje de Stroke por Grupo de Edad")
```



A medida que la edad aumenta, aumenta claramente la predisposición a sufrir un stroke

Vamos a estudiar si tiene valores atípicos con el boxplot

```
# Definir paleta de colores
colores <- c("orange", "purple")

# Crear un gráfico de boxplot con colores personalizados y resaltado de outliers
ggplot(datos, aes(x = factor(stroke), y = age, fill = factor(stroke))) +
  labs(title = "Boxplot de diagnóstico positivo/negativo según edades ")+
  geom_boxplot(outlier.color = "blue4", outlier.shape = 16) +
  scale_fill_manual(values = colores) +
  labs(y = "Edad", x = "Stroke") +
  theme_minimal()
```



No hay una cantidad significativa de outliers, a simple vista vemos que hay gran diferencia entre las medianas

Para mostrar si es significativa la diferencia utilizaremos una prueba estadística no paramétrica llamada U de Mann-Whitney (tambien conocida como Wilcoxon-Mann-Whitney) para comparar las medianas de edad de el grupo Stroke = 0 y el grupo Stroke = 1 (pagina 348)

$$\begin{cases} H_0 : No\ hay\ diferencia\ entre\ las\ medianas \\ H_1 : Hay\ diferencia\ significativa\ entre\ las\ medianas \end{cases}$$

```
wilcox.test(age ~ stroke,
            data = datos)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  age by stroke
## W = 200264, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

p-value $<$ 2.2e-16

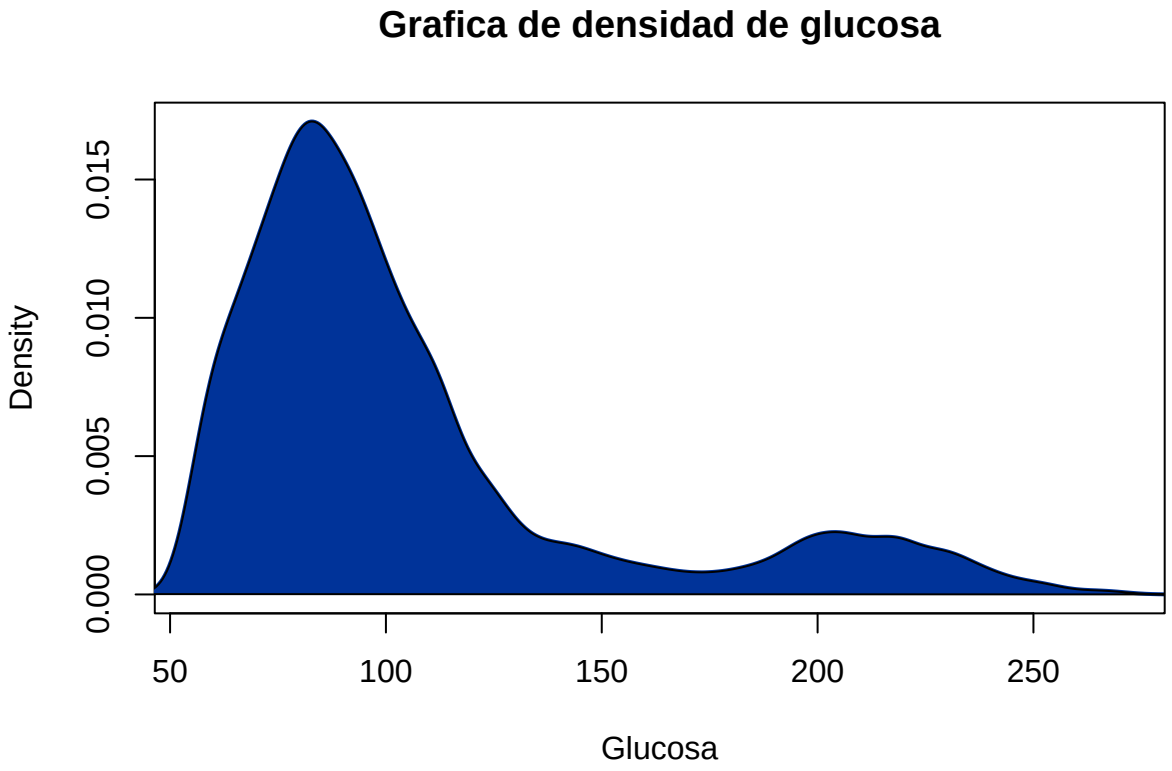
El p-value es lo suficicentemente pequeño para rechazar la hipótesis nula

Nos ha dado un p-valor mucho más pequeño que 0.05, lo que significa que la edad es significativamente diferente entre las personas que sufren un accidente cerebrovascular

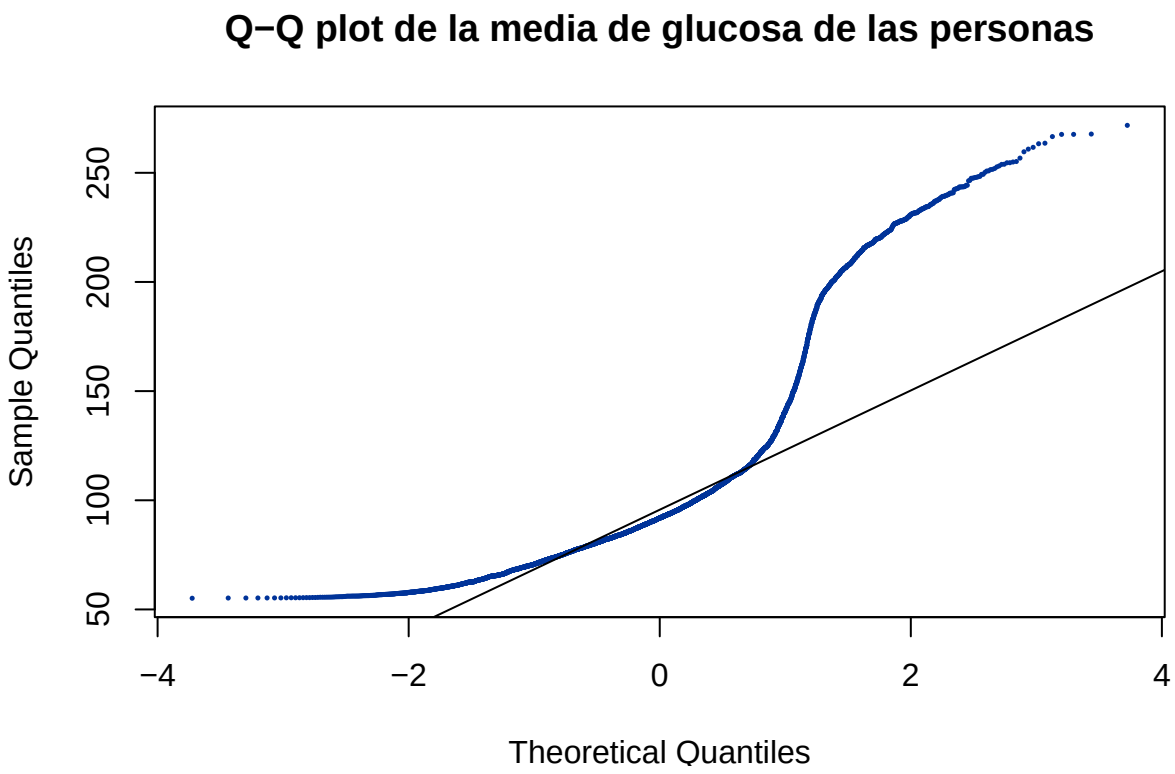
A.2) Glucosa y accidente cerebrovascular

La glucosa viene representada en nuestro dataset como avg_glucose_level.

```
# Calculamos la funcion de densidad
densidad <- density(datos$avg_glucose_level)

# Creamos la ventana donde irá la grafica

plot(densidad, lwd = 2,
     main = "Grafica de densidad de glucosa",
     xlab = "Glucosa", col = rgb(0, 0.2, 0.6),
     xlim = c(min(datos$avg_glucose_level), max(datos$avg_glucose_level)),
     ylim = c(0, max(densidad$y, densidad$y)))

# Densidad

polygon(densidad, col = rgb(0, 0.2, 0.6,))
```




```
stats::qqnorm(datos$avg_glucose_level,
              main = "Q-Q plot de la media de glucosa de las personas",
              pch = 16, cex = 0.35,
              col = rgb(0, 0.2, 0.6))
stats::qqline(datos$avg_glucose_level)
```



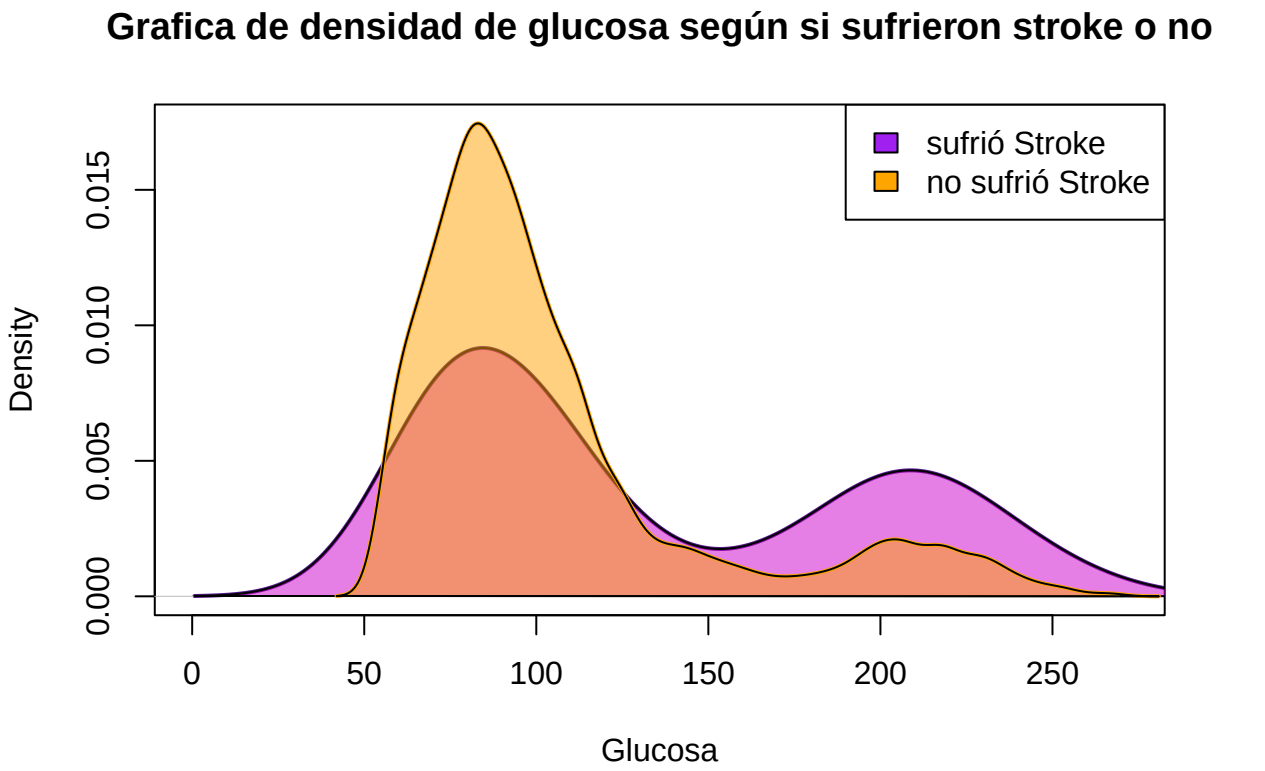
Podemos observar una gráfica de densidad con dos máximos relativos. Sin embargo como la primera campana es mucho más alta que la segunda, podemos concluir que la mayoría de la población tiende a tener un nivel de glucosa alrededor de 80. Es una distribucion asimetrica a la derecha

Además observamos claramente en la gráfica Q-Q plot que no es distribucion normal

Veamos ahora la grafica de densidad separando los grupos stroke = 1 y stroke = 0

```
# Calculamos la funcion de densidad de la glucosa para ambos grupos
densidad_stroke_1 <- density(datos_stroke_1$avg_glucose_level)
densidad_stroke_0 <- density(datos_stroke_0$avg_glucose_level)

# Creamos la ventana donde irá la grafica

plot(densidad_stroke_1, lwd = 2,
     main = "Grafica de densidad de glucosa según si sufrieron stroke o no",
     xlab = "Glucosa", col = "purple4", xlim = c(0, max(datos$avg_glucose_level)),
     ylim = c(0, max(densidad_stroke_1$y, densidad_stroke_0$y)))

# Densidad para stroke = 1

polygon(densidad_stroke_1, col = rgb(0.8, 0, 0.8, alpha = 0.5))

# Densidad para stroke = 0

lines(densidad_stroke_0, lwd = 2, col = "orange1")
polygon(densidad_stroke_0, col = rgb(1, 0.64, 0, alpha = 0.5))


# Leyenda
```

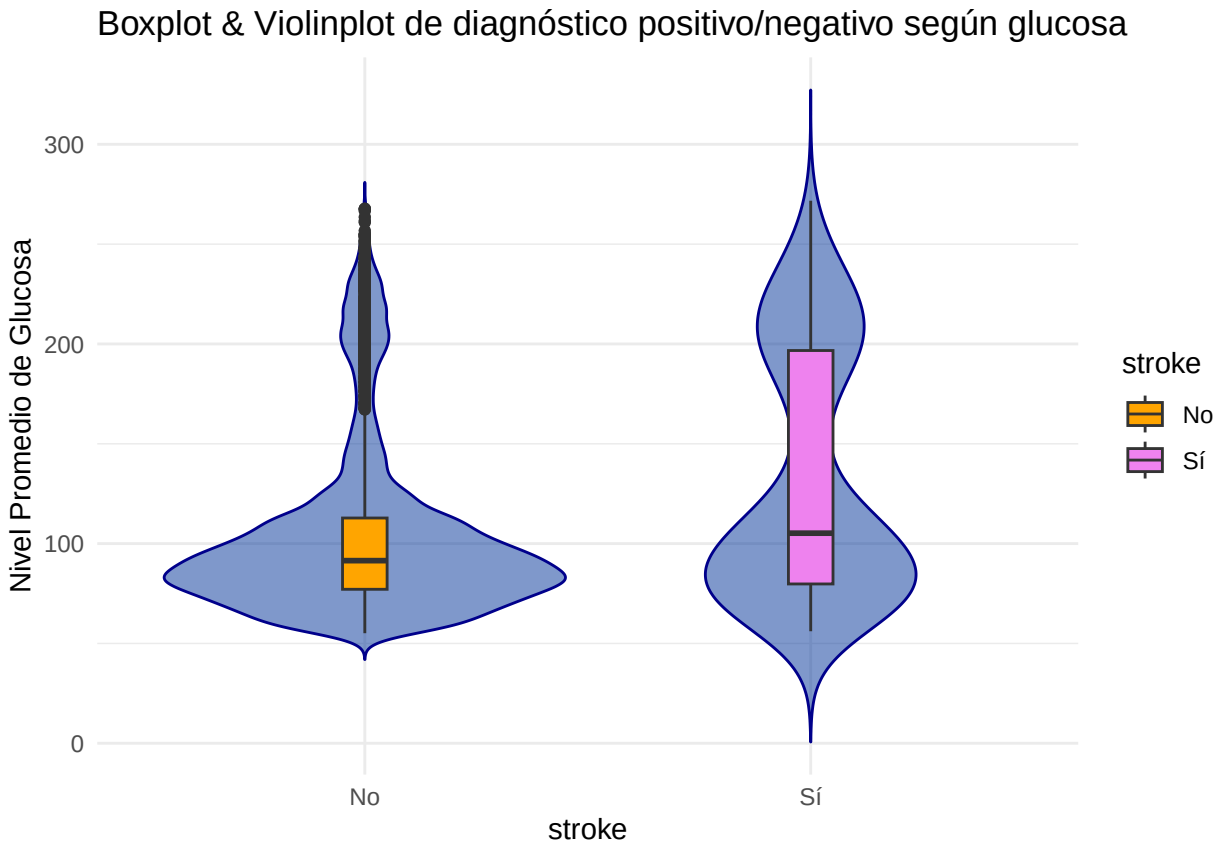
```r
legend("topright", legend = c("sufrió Stroke", "no sufrió Stroke"),
       fill = c("purple", "orange"))
```



La gráfica de las personas que sufrieron stroke muestra una diferencia con la de las personas que no sufrieron stroke. El segundo máximo relativo es mucho más alto que el las personas que no lo sufrieron. Esto puede indicar una correlacion positiva a sufrir un stroke con la media de glucosa en sangre.

Para obtener una comparativa más clara, vamos a crear un gráfico de caja (boxplot) y un gráfico de violín (violinplot). El violinplot se caracteriza por mostrar la densidad de la distribución de los datos en diferentes valores de la variable

```r
# Crear una copia temporal del conjunto de datos llamado "datos"
temp <- datos

# Cambiamos valores numéricos a categoricos
temp$stroke <- ifelse(temp$stroke == 0, "No", 'Sí')

# Definir una paleta de colores para las cajas
colores_cajas <- c("orange", "violet")  # Puedes ajustar los colores aquí

# Gráfico de Violines y Cajas de la Edad en Pacientes con y sin Accidentes Cerebrovasculares
ggplot(temp, aes(x=stroke, y=avg_glucose_level, fill = stroke)) +
  labs(title = "Boxplot & Violinplot de diagnóstico positivo/negativo según glucosa ")+
  geom_violin(trim=FALSE, fill=rgb(0, 0.2, 0.6, alpha = 0.5), color="blue4") +
  geom_boxplot(width=0.1) +
  scale_fill_manual(values = colores_cajas) +  # Asignar colores a las cajas
  labs(y = "Nivel Promedio de Glucosa")  +
  theme_minimal()
```



Gracias al gráfico anterior podemos ver que las medianas no están a la misma altura. Además vemos que hay una cantidad enorme de valores atípicos. Son los puntos que hay en el boxplot de aquellas personas que no sufrieron accidente cerebrovascular

Analizaremos la significancia estadística mediante la prueba pb2gen del paquete WRS2.

pb2gen es una herramienta valiosa para realizar pruebas de comparacion de poblaciones en situaciones donde los supuestos paramétricos no se cumplen y hay preocupaciones sobre la presencia de outliers. Permite realizar pruebas estadísticas robustas y confiables que se adaptan a diversos tipos de datos y distribuciones.

$$\begin{cases} H_0 : No\ hay\ diferencia\ entre\ las\ medianas \\ H_1 : Hay\ diferencia\ significativa\ entre\ las\ medianas \end{cases}$$

```
pb2gen(avg_glucose_level ~ stroke,
       data = datos)
```

```
## Call:
## pb2gen(formula = avg_glucose_level ~ stroke, data = datos)
##
## Test statistic: -27.6995, p-value = 0.00668
## 95% confidence interval:
## -49.5458    -4.9741
```

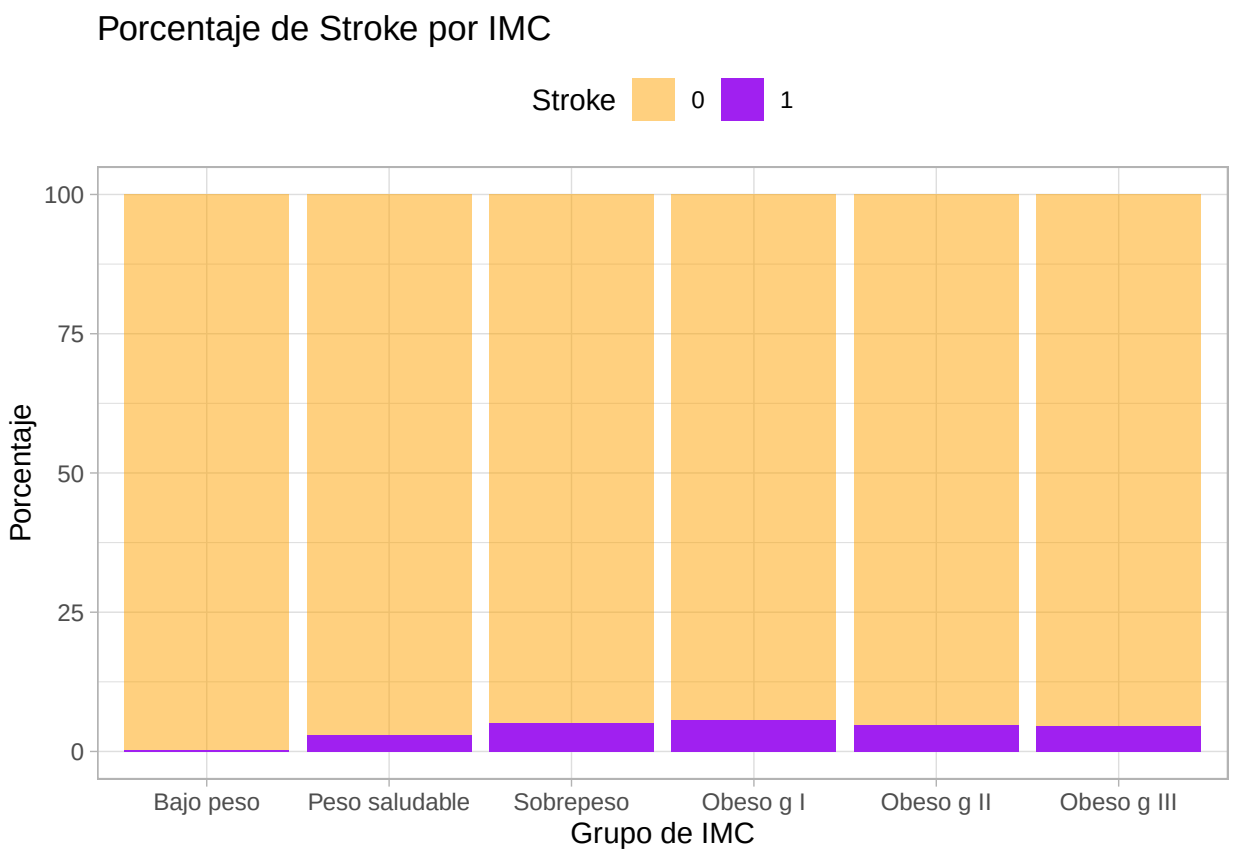
El valor p es extremadamente pequeño muy cercano a 0, lo que sugiere que hay evidencia estadística sólida para rechazar la hipótesis nula.

El intervalo de confianza del 95% no incluye el valor cero, lo que refuerza la conclusión de que hay una diferencia significativa en las medianas de nivel promedio de glucosa entre los grupos “Stroke = 0” y “Stroke = 1”

A.3) Indice de masa corporal y accidente cerebrovascular

Vamos a crear un gráfico que muestre la diferencia de porcentaje de strokes entre las personas de los distintos rangos del IMC

```
temp = na.omit(datos)


colores = c(rgb(1, 0.64, 0, alpha = 0.5), "purple")

# Calcular las frecuencias de stroke por grupo de IMC

frecuencias <- temp %>%
  group_by(imc_str, stroke) %>%
  summarise(count = n()) %>%
  group_by(imc_str) %>%
  mutate(percentage = (count / sum(count)) * 100)

# Crear el gráfico de barras
ggplot(frecuencias, aes(x = imc_str, y = percentage, fill = factor(stroke))) +
  geom_bar(stat = "identity", position = "stack") +
  labs(x = "Grupo de IMC", y = "Porcentaje", fill = "Stroke") +
  scale_x_discrete(limits = c("Bajo peso", "Peso saludable", "Sobrepeso",
                              "Obeso g I", "Obeso g II", "Obeso g III")) +
  scale_fill_manual(values = colores, name = "Stroke", labels = c("0", "1")) +
  theme_light() +
  theme(legend.position = "top") +
  ggtitle("Porcentaje de Stroke por IMC")
```

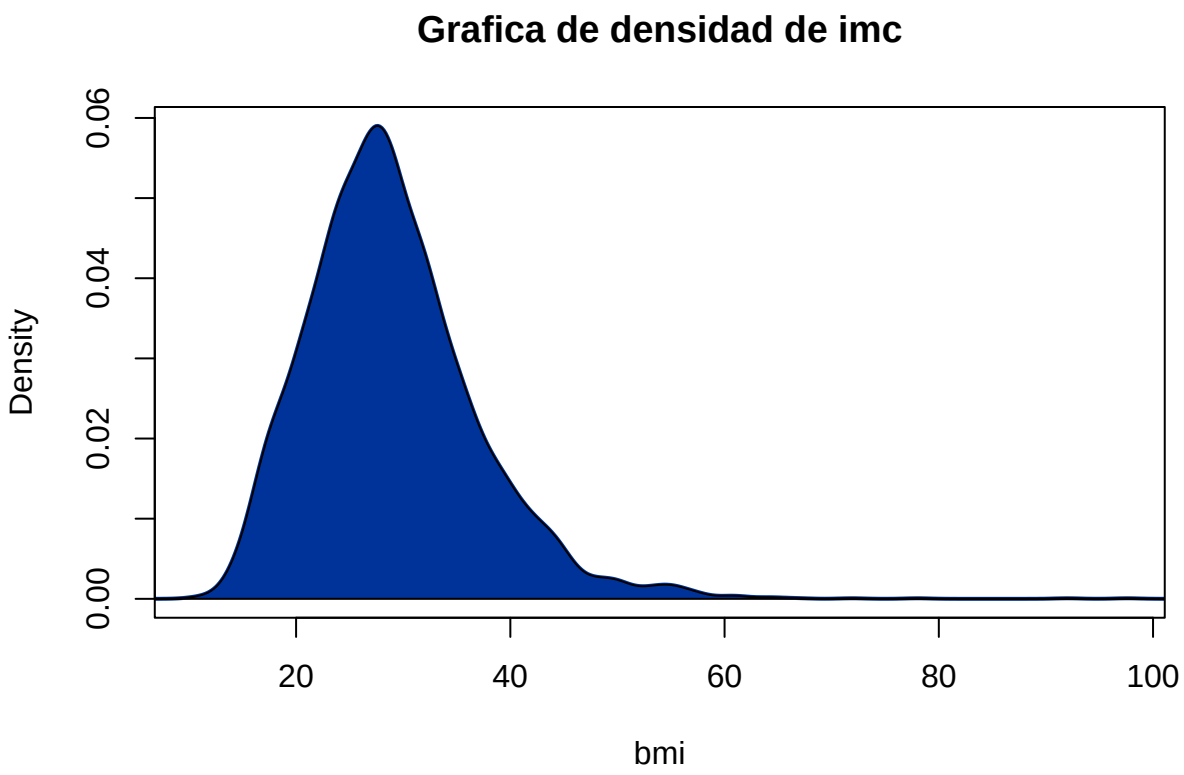


A simple vista, parece que hay una tendencia entre las personas con más imc a sufrir un stroke. Mientras que las personas con menos imc no lo sufren tanto

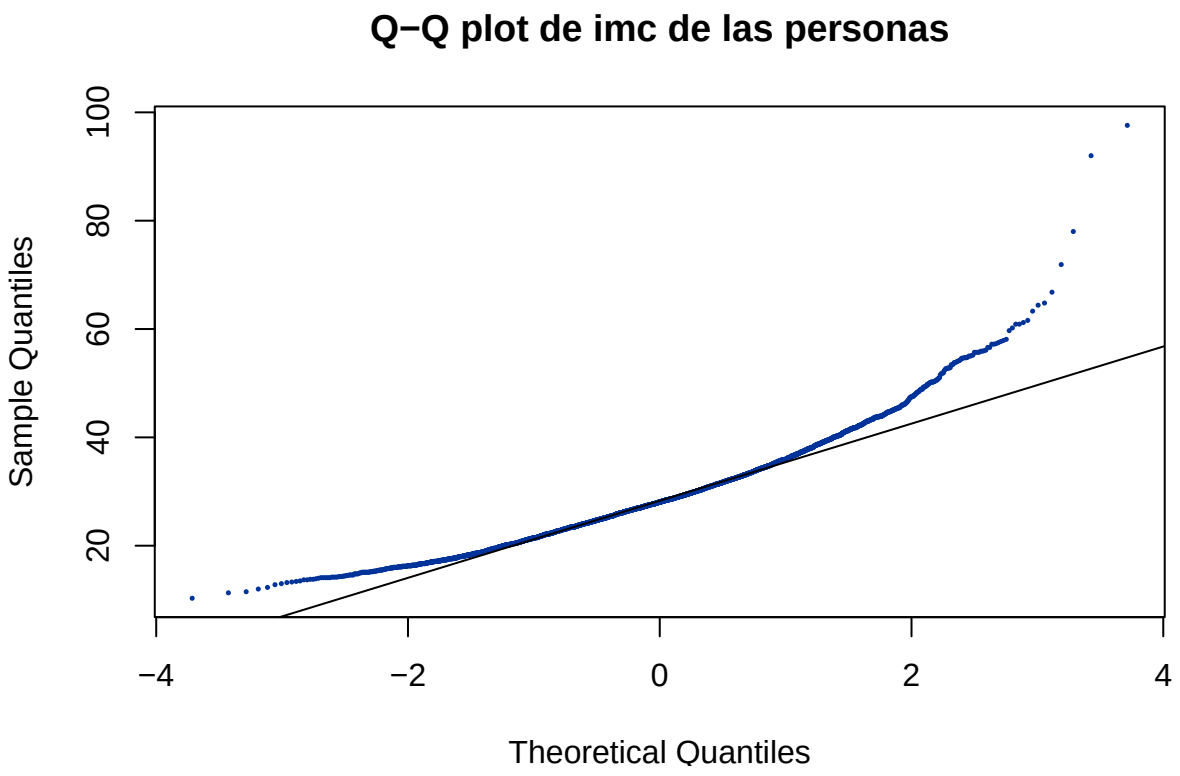
Vamos a realizar el análisis exploratorio de esta variable sin tener en cuenta los datos NA

Dibujaremos la gráfica de densidad de bmi, su Q-Q plot y también un test de Anderson-Darling

```
# Calculamos la funcion de densidad
densidad <- density(temp$bmi)


# Creamos la ventana donde irá la grafica

plot(densidad, lwd = 2,
     main = "Grafica de densidad de imc",
     xlab = "bmi", col = rgb(0, 0.2, 0.6),
     xlim = c(min(temp$bmi), max(temp$bmi)),
     ylim = c(0, max(densidad$y, densidad$y)))

# Densidad para stroke

polygon(densidad, col = rgb(0, 0.2, 0.6,))
```




```
stats::qqnorm(temp$bmi,
              main = "Q-Q plot de imc de las personas",
              pch = 16, cex = 0.35,
              col = rgb(0, 0.2, 0.6))

stats::qqline(temp$bmi)
```




```
nortest::ad.test(temp$bmi )
```

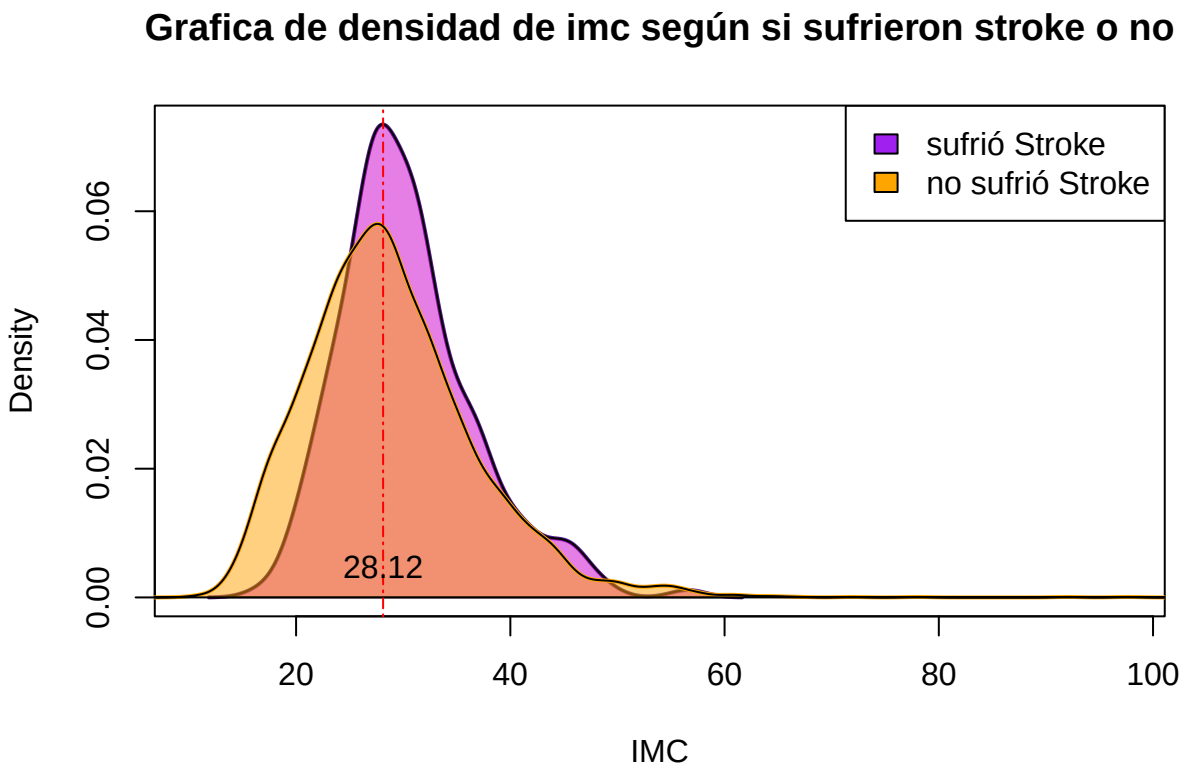
```
##
##  Anderson-Darling normality test
##
## data:  temp$bmi
## A = 32.57, p-value < 2.2e-16
```

Es una función asimétrica a la derecha. No sigue distribución normal.

Ahora toca dibujar las funciones de distribucion con y sin accidente cerebrovascular

```
datos_stroke_1 <- subset(temp, stroke == 1)
datos_stroke_0 <- subset(temp, stroke == 0)

# Calculamos la funcion de desnidad de la edad para ambos grupos
densidad_stroke_1 <- density(datos_stroke_1$bmi)
densidad_stroke_0 <- density(datos_stroke_0$bmi)

maximo = max(densidad_stroke_1$y)
x_maximo = densidad_stroke_1$x[which.max(densidad_stroke_1$y)]

# Creamos la ventana donde irá la grafica

plot(densidad_stroke_1, lwd = 2,
     main = "Grafica de densidad de imc según si sufrieron stroke o no",
     xlab = "IMC", col = "purple4", xlim = c(min(temp$bmi), max(temp$bmi)),
     ylim = c(0, max(densidad_stroke_1$y, densidad_stroke_0$y)))

# Densidad para stroke = 1
```

```r
polygon(densidad_stroke_1, col = rgb(0.8, 0, 0.8, alpha = 0.5))

# Densidad para stroke = 0

lines(densidad_stroke_0, lwd = 2, col = "orange1")
polygon(densidad_stroke_0, col = rgb(1, 0.64, 0, alpha = 0.5))

# Valor maximo

abline(v = x_maximo, col = "red", lty=6)

text(x_maximo, 0, labels=round(x_maximo, 2), pos = 3, col = "black")


# Leyenda

legend("topright", legend = c("sufrió Stroke", "no sufrió Stroke"),
       fill = c("purple", "orange"))
```

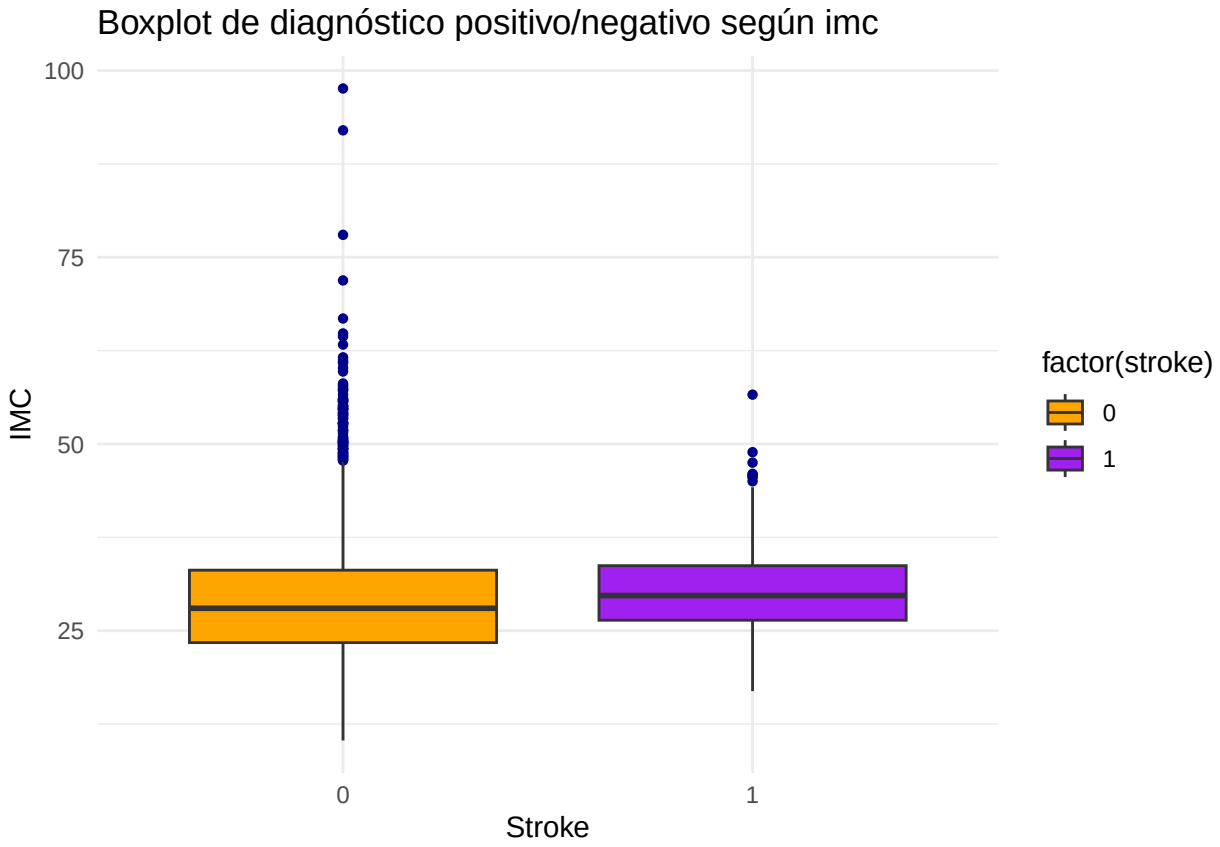


Parece que cuando bmi está alrededor de 28.12 o por encima, hay mas posibilidad de sufrir stroke. Esto es sobrepeso

Primero haremos un estudio de sus valores atípicos

```r
# Definir paleta de colores
colores <- c("orange", "purple")

# Crear un gráfico de boxplot con colores personalizados y resaltado de outliers
ggplot(datos, aes(x = factor(stroke), y = bmi, fill = factor(stroke))) +
  labs(title = "Boxplot de diagnóstico positivo/negativo según imc")+
```

```
  geom_boxplot(outlier.color = "blue4", outlier.shape = 16) +
  scale_fill_manual(values = colores) +
  labs(y = "IMC", x = "Stroke") +
  theme_minimal()
```



Volveremos a usar la función pb2gen, por todos los valores atípicos que aparecen de azul en el boxplot

$$\begin{cases} H_0 : No\ hay\ diferencia\ entre\ las\ medianas \\ H_1 : Hay\ diferencia\ significativa\ entre\ las\ medianas \end{cases}$$

```
pb2gen(bmi ~ stroke,
       data = datos)
```

```
## Call:
## pb2gen(formula = bmi ~ stroke, data = datos)
##
## Test statistic: -1.4274, p-value = 0
## 95% confidence interval:
## -2.6632    -0.5404
```

El valor p es muy cercano a 0, lo que significa que la diferencia en las medias de "bmi" entre los grupos "stroke = 0" y "stroke = 1" es extremadamente pequeña y estadísticamente significativa.

El intervalo de confianza del 95% no incluye el valor cero, lo que respalda la conclusión de que hay una diferencia significativa entre las medias.

A.4) Estadísticos de las variables continuas

A continuación mostramos la tabla con estadísticos descriptivos de las variables continuas

```
# Creamos un subconjunto con las filas donde stroke es igual a 1
sub_continuas_1 <- datos %>%
  filter(stroke == 1) %>%
  select(age, bmi, avg_glucose_level) %>%
  rename(
    "Edad & diagnostico positivo" = age,
    "IMC & diagnostico positivo" = bmi,
    "Nivel medio de glu & diagnostico positivo" = avg_glucose_level)

# Creamos un subconjunto con las filas donde stroke es igual a 0
sub_continuas_0 <- datos %>%
  filter(stroke == 0) %>%
  select(age, bmi, avg_glucose_level) %>%
  rename(
    "Edad & diagnostico negativo" = age,
    "IMC & diagnostico negativo" = bmi,
    "Nivel medio de glu & diagnostico negativo" = avg_glucose_level)



# Creamos la funcion para calcular estadísticos

calcular_estadisticos <- function(dataset) {
  estadisticos <- data.frame(
    Mediana = sapply(dataset, median, na.rm = TRUE),
    PrimerCuartil = sapply(dataset, quantile, probs = 0.25, na.rm = TRUE),
    TercerCuartil = sapply(dataset, quantile, probs = 0.75, na.rm = TRUE),
    IQR = sapply(dataset, function(x) IQR(x, na.rm = TRUE))
  )
  return(estadisticos)
}
# Ahora calculamos los estadisticos y los unimos

estadisticos_0 = calcular_estadisticos(sub_continuas_0)

estadisticos_1 = calcular_estadisticos(sub_continuas_1)

# Unimos los dataframes utilizando rbind

estadisticos <- rbind(estadisticos_0, estadisticos_1)

# Ordenamos rownames alfabeticamente para intercalarlos

estadisticos_ordenado <- estadisticos[order(rownames(estadisticos)), ]


estadisticos_ordenado %>%
  kbl(caption = "Tabla de estadísticos de variable continuas") %>%
  pack_rows("Edad p-value = 0", 1, 2) %>%
  pack_rows("IMC p-value = 0", 3, 4) %>%
  pack_rows("Glucosa p-value = 0.003", 5, 6) %>%
```

```
kable_styling(full_width = FALSE, position = "center") %>%
row_spec(seq(1, nrow(estadisticos_ordenado), by = 2),
         background = rgb(1.0, 0.8, 0.6)) %>%
row_spec(seq(2, nrow(estadisticos_ordenado), by = 2),
         background = rgb(0.8, 0.6, 1.0)) %>%
column_spec(1, background = "white") %>%
kable_styling(latex_options = c("hold_position"))
```

Table 1: Tabla de estadísticos de variable continuas

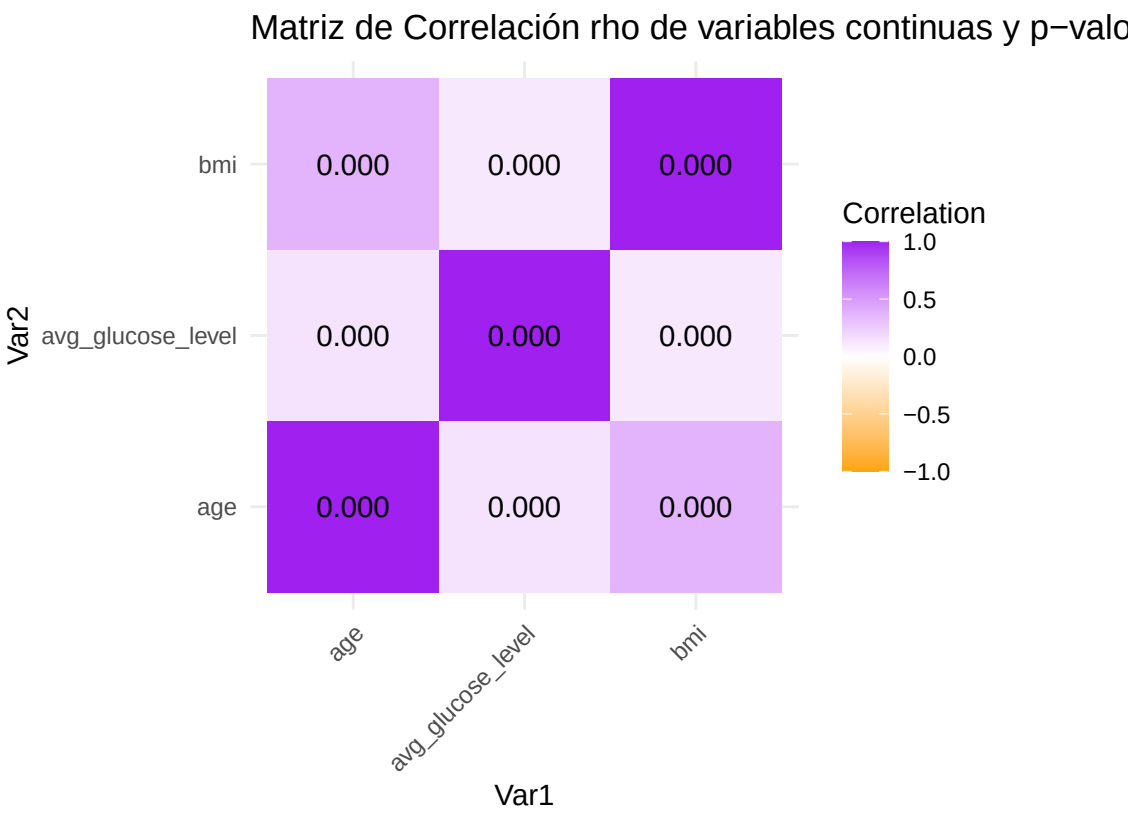
	Mediana	PrimerCuartil	TercerCuartil	IQR
Edad p-value = 0				
Edad & diagnostico negativo	43.00	24.00	59.00	35.00
Edad & diagnostico positivo	71.00	59.00	78.00	19.00
IMC p-value = 0				
IMC & diagnostico negativo	28.00	23.40	33.10	9.70
IMC & diagnostico positivo	29.70	26.40	33.70	7.30
Glucosa p-value = 0.003				
Nivel medio de glu & diagnostico negativo	91.47	77.12	112.83	35.71
Nivel medio de glu & diagnostico positivo	105.22	79.79	196.71	116.92

A.5) Matriz de correlaciones de variables continuas

La correlación de Spearman es un método estadístico no paramétrico para medir la relación monotónica entre dos variables. Se basa en los rangos de los valores en lugar de las relaciones lineales. El coeficiente de correlación de Spearman, denotado como ρ (rho), varía entre -1 y 1.

ρ cerca de +1 indica una correlación positiva fuerte, donde los valores altos de una variable se asocian con valores altos de la otra. ρ cerca de -1 indica una correlación negativa fuerte, donde los valores altos de una variable se asocian con valores bajos de la otra. ρ cerca de 0 sugiere una correlación débil o nula, sin una relación monotónica clara.

```
datos_continuos <- datos[, c("age", "avg_glucose_level", "bmi")]

matriz_cor <- corr.test(datos_continuos, method = "spearman", use = "complete.obs")$r
matriz_pval <- abs(corr.test(datos_continuos, method = "spearman", use = "complete.obs")$p)

ggplot(data = as.data.frame(as.table(matriz_cor))) +
  geom_tile(aes(Var1, Var2, fill = Freq)) +
  geom_text(aes(Var1, Var2, label = sprintf("%.3f", matriz_pval)), color = "black") +
  scale_fill_gradient2(low = "orange", high = "purple", mid = "white",
                       midpoint = 0, limit = c(-1, 1), space = "Lab",
                       name = "Correlation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_fixed() +
  labs(title = "Matriz de Correlación rho de variables continuas y p-valores",
       caption = "Como todas las p-value cumplen |p-value| < 0.05 => son significativas")
```



Como todas las p−value cumplen |p−value| < 0.05 => son significativas

B) Estudio de variables categoricas

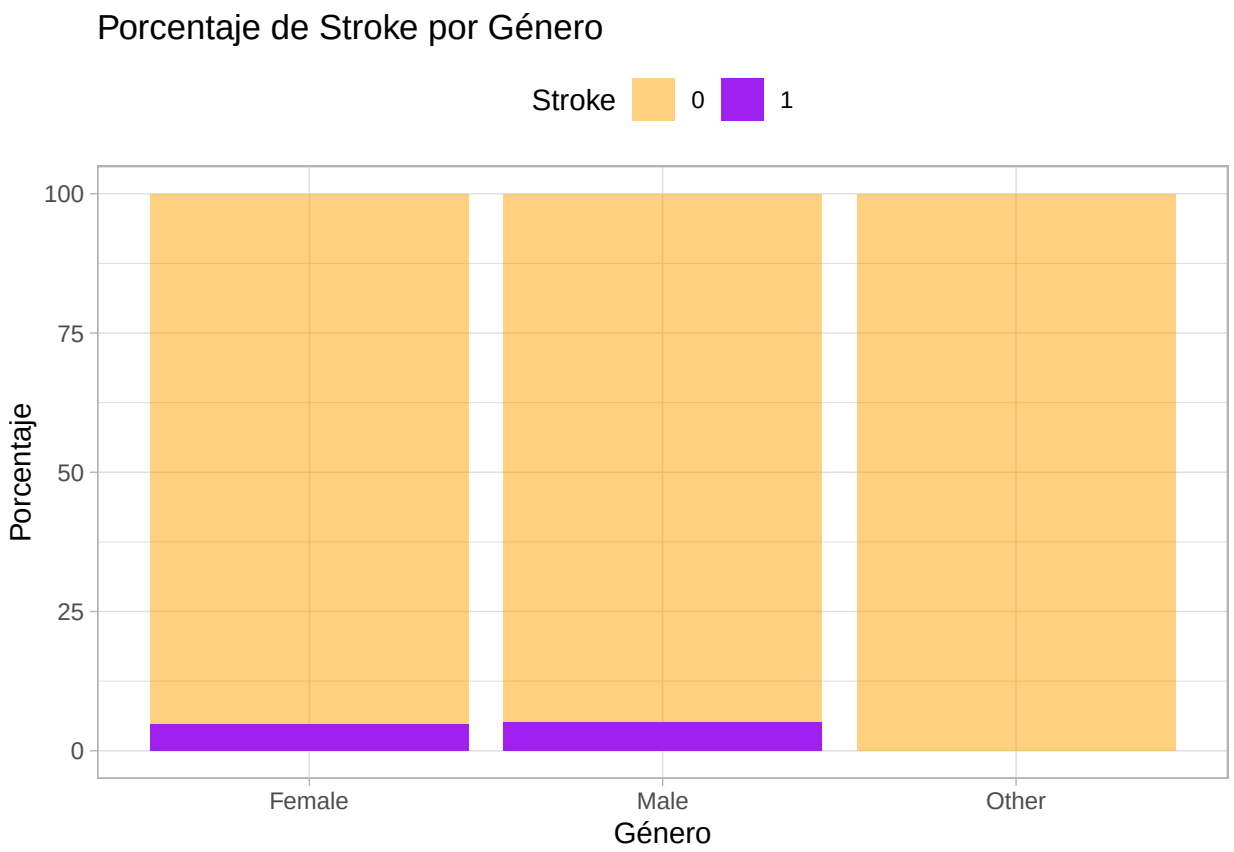
B.1) Genero y stroke

Estudiemos la variable gender de nuestro dataset

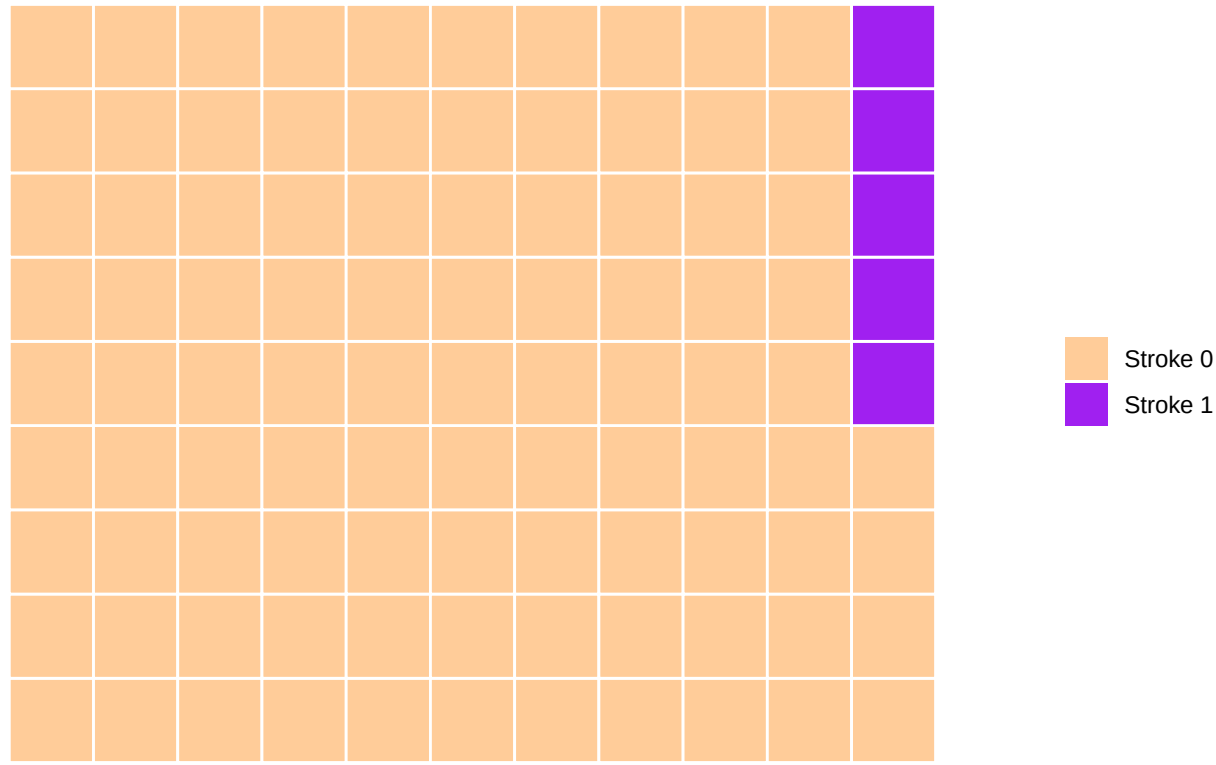
```
table(datos$gender)
```

```
##
## Female   Male  Other
##   2994   2115      1
```

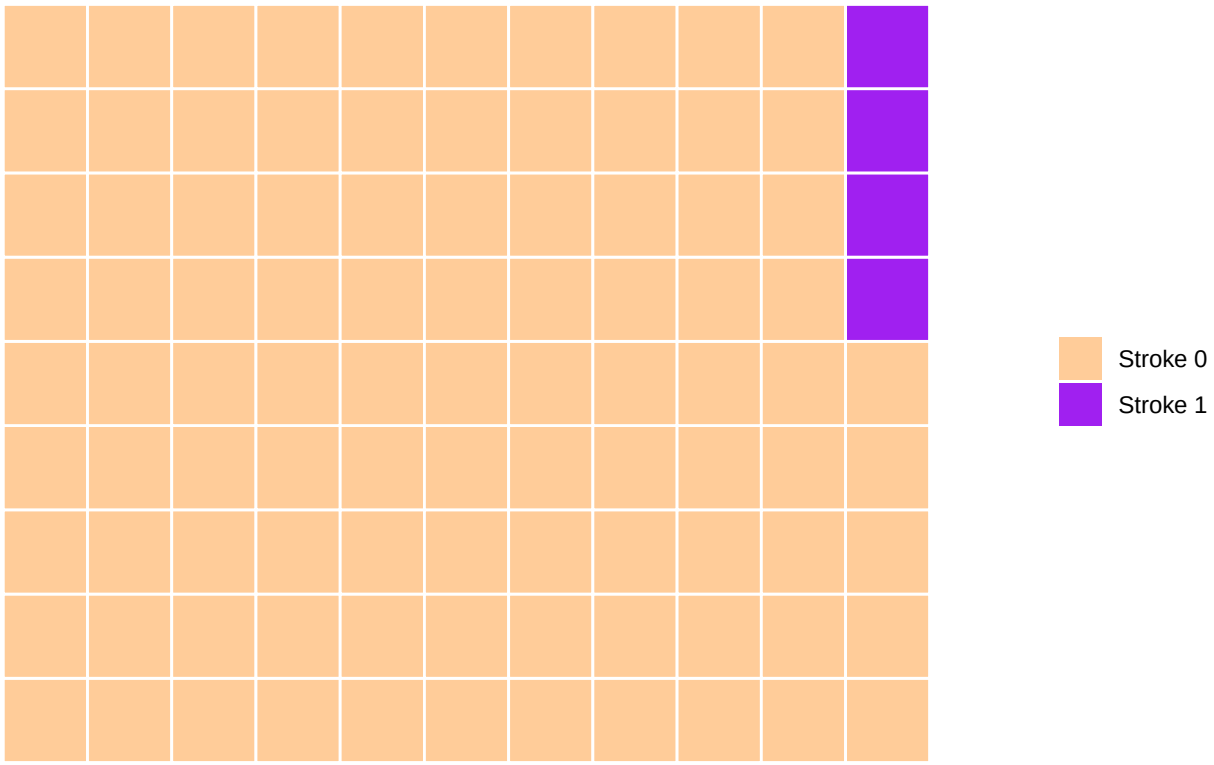
```
colores = c(rgb(1, 0.64, 0, alpha = 0.5), "purple")

# Calcular las frecuencias de stroke por grupo de género

frecuencias <- datos %>%
  group_by(gender, stroke) %>%
  summarise(count = n()) %>%
  group_by(gender) %>%
  mutate(percentage = (count / sum(count)) * 100)

# Creamos el gráfico de barras
ggplot(frecuencias, aes(x = gender, y = percentage, fill = factor(stroke))) +
  geom_bar(stat = "identity", position = "stack") +
  labs(x = "Género", y = "Porcentaje", fill = "Stroke") +
  scale_fill_manual(values = colores, name = "Stroke", labels = c("0", "1")) +
  theme_light() +
  theme(legend.position = "top") +
  ggtitle("Porcentaje de Stroke por Género")
```




```
p_male_s1 = frecuencias$percentage[frecuencias$gender == "Male" & frecuencias$stroke == 1]
p_male_s0 = frecuencias$percentage[frecuencias$gender == "Male" & frecuencias$stroke == 0]


p_female_s1 = frecuencias$percentage[frecuencias$gender == "Female" & frecuencias$stroke == 1]
p_female_s0 = frecuencias$percentage[frecuencias$gender == "Female" & frecuencias$stroke == 0]


datos_hombres = c("Stroke 0" = p_male_s0, "Stroke 1" = p_male_s1)

datos_mujeres = c("Stroke 0" = p_female_s0, "Stroke 1" = p_female_s1)


waffle(datos_hombres, rows = 9, size = 0.5, colors = c(rgb(1.0, 0.8, 0.6), "purple"),
       title = "Distribución de Stroke para Hombres")
```




```
# Crear gráfico waffle para mujeres
waffle(datos_mujeres, rows = 9, size = 0.5, colors = c(rgb(1.0, 0.8, 0.6), "purple"),
       title = "Distribución de Stroke para Mujeres")
```



Hay 3 variables. El estudio está realizado sobre 2995 mujeres, 2115 hombres y un valor "otros"

No podremos obtener ningún resultado significativo del valor otros porque su tamaño muestral no es lo suficientemente grande

A simple vista no hay una diferencia porcentual elevada entre los hombres que sufren stroke y las mujeres, si acaso levemente superior en los hombres.

Para ver si hay alguna diferencia significativa entre la población de hombres y la de mujeres usaremos un Test de Chi Cuadrado, ya que estamos comparando poblaciones de variables categóricas

Se pretende validar o rechazar lo siguiente:

H_0: Las poblaciones female y male sufren accidentes cerebrovasculares con la misma frecuencia. H_1: Las poblaciones female y male no sufren accidentes cerebrovasculares con la misma frecuencia.

El test de Chi Cuadrado en R se hace con chisq.test()

```
chisq.test(datos$gender, datos$stroke)
```

```
##
##  Pearson's Chi-squared test
##
## data:  datos$gender and datos$stroke
## X-squared = 0.47259, df = 2, p-value = 0.7895
```

p-value = 0.7895, que es mayor a 0.05. No hay suficiente evidencia para rechazar la hipótesis nula. Entonces el género no afecta de manera significativa a la frecuencia de sufrir un accidente cerebrovascular

Vamos a hacer una tabla de contingencia con genero

```
tabla <- table(datos$gender, datos$stroke)
tabla_contingencia <- addmargins(tabla)
```

```r
num_columnas <- ncol(tabla_contingencia)

colores <- c(rgb(0.8, 0.6, 1.0), rgb(1.0, 0.8, 0.6))

tabla_formateada <- tabla_contingencia %>%
  kable(row.names = TRUE) %>%
  kable_styling(full_width = FALSE, position = "center") %>%
  column_spec(seq(1, num_columnas, by = 2), background = colores[1]) %>%
  column_spec(seq(2, num_columnas, by = 2), background = colores[2]) %>%
  column_spec(1,background = "white") %>%
  row_spec(4, background = "white")  %>%
  kable_styling(latex_options = c("hold_position"))

tabla_formateada
```

	0	1	Sum
Female	2853	141	2994
Male	2007	108	2115
Other	1	0	1
Sum	4861	249	5110

B.2) Tipo de trabajo y accidente cerebrovascular

Miraremos la influencia que tiene el tipo de trabajo en la incidencia de stroke

```r
table(datos$work_type)
```

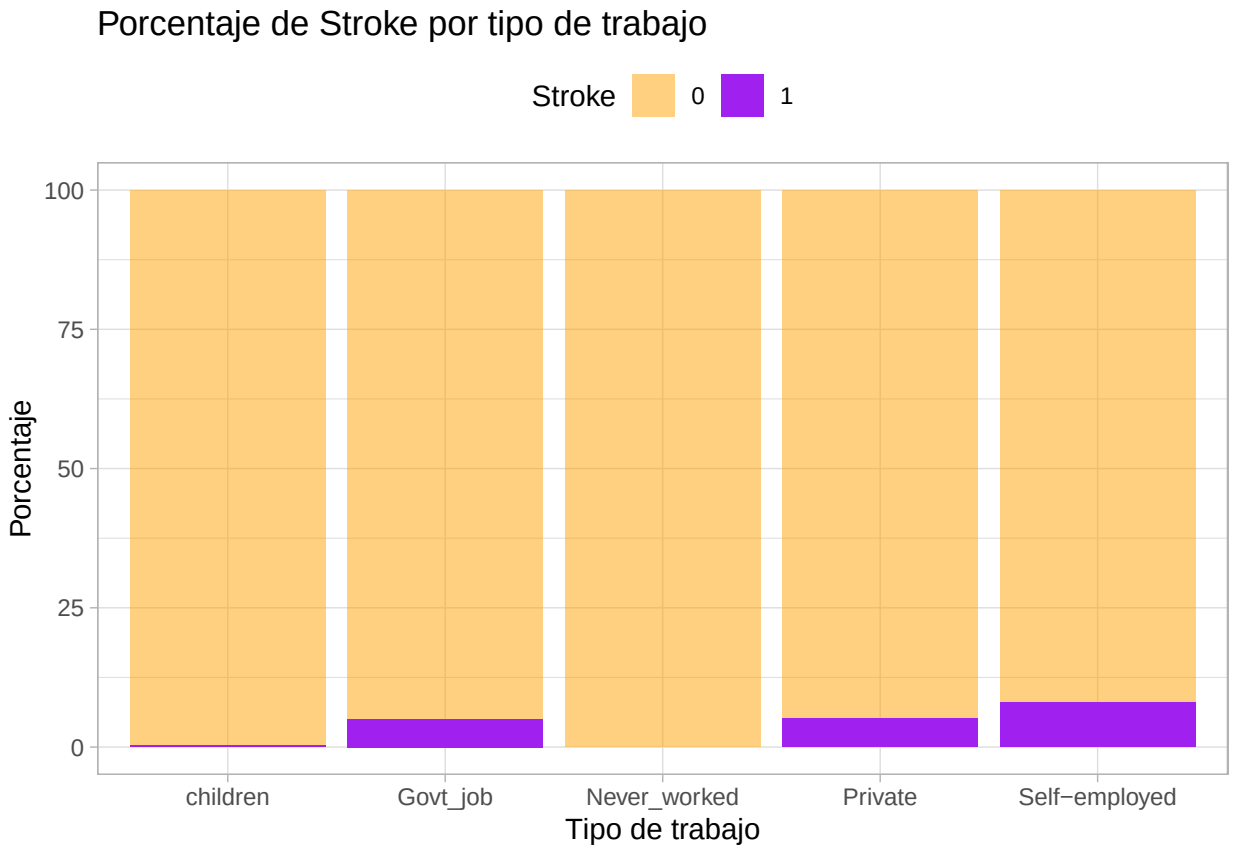
```
##
##      children      Govt_job  Never_worked       Private Self-employed
##           737           654            10          2896           813
```

```r
n_w_observations <- filter(datos, work_type == "Never_worked")

max_age_never_worked <- max(n_w_observations$age)

max_age_never_worked
```

```
## [1] 23
```

```r
colores = c(rgb(1, 0.64, 0, alpha = 0.5), "purple")

# Calcular las frecuencias de stroke por grupo de género

frecuencias <- datos %>%
  group_by(work_type, stroke) %>%
  summarise(count = n()) %>%
  group_by(work_type) %>%
  mutate(percentage = (count / sum(count)) * 100)

# Creamos el gráfico de barras
ggplot(frecuencias, aes(x = work_type, y = percentage, fill = factor(stroke))) +
  geom_bar(stat = "identity", position = "stack") +
  labs(x = "Tipo de trabajo", y = "Porcentaje", fill = "Stroke") +
  scale_fill_manual(values = colores, name = "Stroke", labels = c("0", "1")) +
```

```
  theme_light() +
  theme(legend.position = "top") +
  ggtitle("Porcentaje de Stroke por tipo de trabajo")
```



La persona maś mayor que nunca ha trabajado tiene 23 años. Ademas este grupo tiene pocos representantes, porque aquellos menores de 14 que nunca habían trabajado los cambiamos a children. Este grupo no tiene suficiente tamaño muestral para obtener conclusiones significativas

A simple vista parece que los niños y aquellas personas que nunca han trabajado son las personas que tienen menos predisposicion a sufrir accidentes cerebrovasculares. Puede ser que se deba a que hay una relación directa de estos "tipos de empleo" con la edad.

Para ver si hay alguna diferencia significativa entre las personas con diferentes tipos de trabajo usaremos un Test de Chi Cuadrado, ya que estamos comparando poblaciones de variables categóricas

```
chisq.test(datos$work_type, datos$stroke)
```

```
##
##  Pearson's Chi-squared test
##
## data:  datos$work_type and datos$stroke
## X-squared = 51.78, df = 4, p-value = 1.534e-10
```

El p-value = 1.534e-10, es mucho menor que 0.05, asi que es significativo

Vamos a hacer una tabla de contingencia de los tipos de trabajo con stroke

```
tabla <- table(datos$work_type, datos$stroke)
tabla_contingencia <- addmargins(tabla)
```

```r
num_columnas <- ncol(tabla_contingencia)

colores <- c(rgb(0.8, 0.6, 1.0), rgb(1.0, 0.8, 0.6))

tabla_formateada <- tabla_contingencia %>%
  kable(row.names = TRUE) %>%
  kable_styling(full_width = FALSE, position = "center") %>%
  column_spec(seq(1, num_columnas, by = 2), background = colores[1]) %>%
  column_spec(seq(2, num_columnas, by = 2), background = colores[2]) %>%
  column_spec(1,background = "white") %>%
  row_spec(6, background = "white")  %>%
  kable_styling(latex_options = c("hold_position"))

tabla_formateada
```

	0	1	Sum
children	735	2	737
Govt_job	621	33	654
Never_worked	10	0	10
Private	2747	149	2896
Self-employed	748	65	813
Sum	4861	249	5110

B.3) Tipo de residencia y accidente cerebrovascular

Vamos a estudiar el tipo de residencia

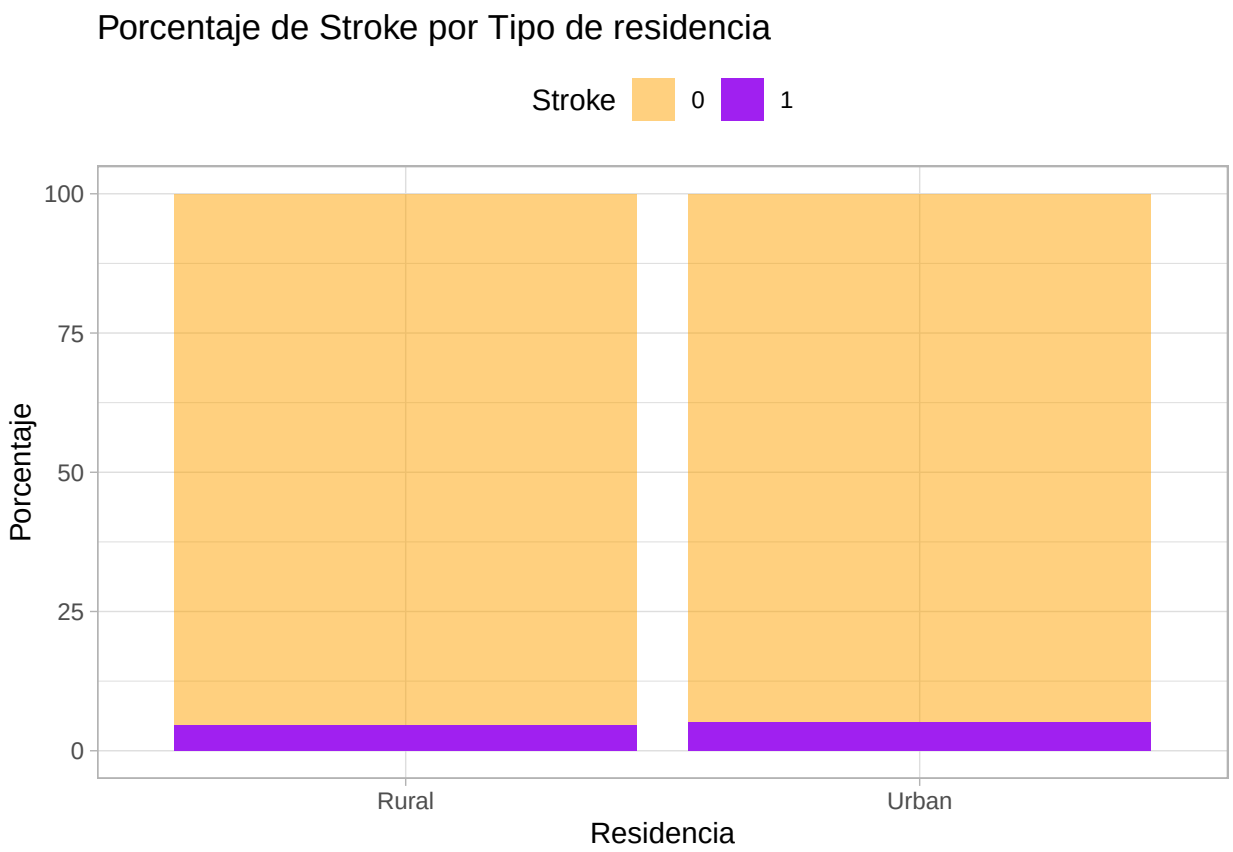
Vemos que hay dos tipos. Rural y urbano

Tambien imprimimos las gráficas de barras con el porcentaje

```
table(datos$residence_type)
```

```
##
## Rural Urban
##  2514  2596
```

```
colores = c(rgb(1, 0.64, 0, alpha = 0.5), "purple")

# Calcular las frecuencias de stroke por grupo de género

frecuencias <- datos %>%
  group_by(residence_type, stroke) %>%
  summarise(count = n()) %>%
  group_by(residence_type) %>%
  mutate(percentage = (count / sum(count)) * 100)

# Creamos el gráfico de barras
ggplot(frecuencias, aes(x = residence_type, y = percentage, fill = factor(stroke))) +
  geom_bar(stat = "identity", position = "stack") +
  labs(x = "Residencia", y = "Porcentaje", fill = "Stroke") +
  scale_fill_manual(values = colores, name = "Stroke", labels = c("0", "1")) +
  theme_light() +
  theme(legend.position = "top") +
  ggtitle("Porcentaje de Stroke por Tipo de residencia")
```



Para ver si hay alguna diferencia significativa entre las personas con diferentes tipos de residencia usaremos un Test de Chi Cuadrado, ya que estamos comparando poblaciones de variables categóricas

```
chisq.test(datos$residence_type, datos$stroke)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  datos$residence_type and datos$stroke
## X-squared = 1.0816, df = 1, p-value = 0.2983
```

p-value = 0.2983

El p-value no es lo suficientemente pequeño para rechazar la hipótesis nula además es > 0.1. Así que no rechazamos la hipótesis. Es por ello que la residencia no es significativa en sufrir un accidente cerebrovascular

Vamos a hacer una tabla de contingencia con tipos de residencia

```
tabla = table(datos$residence_type, datos$stroke)
tabla_contingencia <- addmargins(tabla)

num_columnas <- ncol(tabla_contingencia)

colores <- c(rgb(0.8, 0.6, 1.0), rgb(1.0, 0.8, 0.6))

tabla_formateada <- tabla_contingencia %>%
  kable(row.names = TRUE) %>%
  kable_styling(full_width = FALSE, position = "center") %>%
  column_spec(seq(1, num_columnas, by = 2), background = colores[1]) %>%
  column_spec(seq(2, num_columnas, by = 2), background = colores[2]) %>%
```

```
  column_spec(1,background = "white") %>%
  row_spec(3, background = "white")  %>%
  kable_styling(latex_options = c("hold_position"))

tabla_formateada
```

	0	1	Sum
Rural	2400	114	2514
Urban	2461	135	2596
Sum	4861	249	5110

B.4) Tipo de fumador y accidente cerebrovascular

Vamos a estudiar los tipos de fumadores

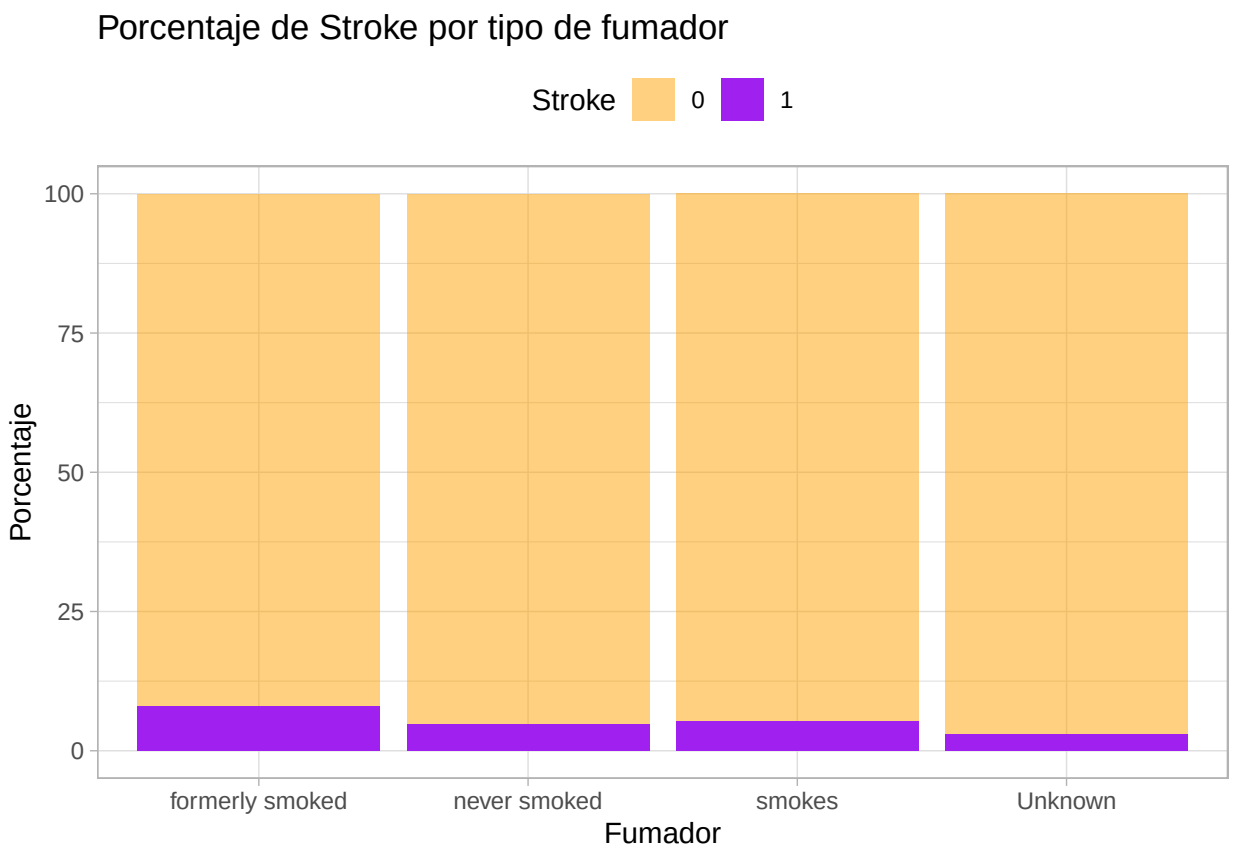
Vemos que hay 4 tipos

Tambien imprimimos las gráficas de barras con el porcentaje

```
table(datos$smoking_status)
```

```
##
## formerly smoked    never smoked          smokes         Unknown
##             885            1892             789            1544
```

```
colores = c(rgb(1, 0.64, 0, alpha = 0.5), "purple")

# Calcular las frecuencias de stroke por grupo de género

frecuencias <- datos %>%
  group_by(smoking_status, stroke) %>%
  summarise(count = n()) %>%
  group_by(smoking_status) %>%
  mutate(percentage = (count / sum(count)) * 100)

# Creamos el gráfico de barras
ggplot(frecuencias, aes(x = smoking_status, y = percentage, fill = factor(stroke))) +
  geom_bar(stat = "identity", position = "stack") +
  labs(x = "Fumador", y = "Porcentaje", fill = "Stroke") +
  scale_fill_manual(values = colores, name = "Stroke", labels = c("0", "1")) +
  theme_light() +
  theme(legend.position = "top") +
  ggtitle("Porcentaje de Stroke por tipo de fumador")
```



H_0: Las poblaciones segun tipo de fumador sufren accidentes cerebrovasculares con la misma frecuencia. H_1: Las poblaciones segun tipo de fumador no sufren accidentes cerebrovasculares con la misma frecuencia.

```
chisq.test(datos$smoking_status, datos$stroke)
```

```
##
##  Pearson's Chi-squared test
##
## data:  datos$smoking_status and datos$stroke
## X-squared = 29.147, df = 3, p-value = 2.085e-06
```

p-value = 2.085e-06

Las poblaciones de distintos tipos de fumador no sufren accidentes cerebrovasculares con la misma frecuencia. Rechazamos la H. nula

Vamos a comparar los fumadores con los exfumadores

```
fumadores_ex <- datos[datos$smoking_status == "smokes" |
                      datos$smoking_status == "formerly smoked", ]


chisq.test(fumadores_ex$smoking_status, fumadores_ex$stroke)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  fumadores_ex$smoking_status and fumadores_ex$stroke
## X-squared = 4.0649, df = 1, p-value = 0.04378
```

p-value = 0.04378

Las poblaciones de fumadores y ex fumadores tienen un p-valor suficientemente pequeño para rechazar la hipótesis nula. Aunque como no es demasiado pequeño existe la posibilidad de cometer un error tipo I al rechazar la hipótesis nula. Así que no extraeremos conclusiones de la segunda comparación

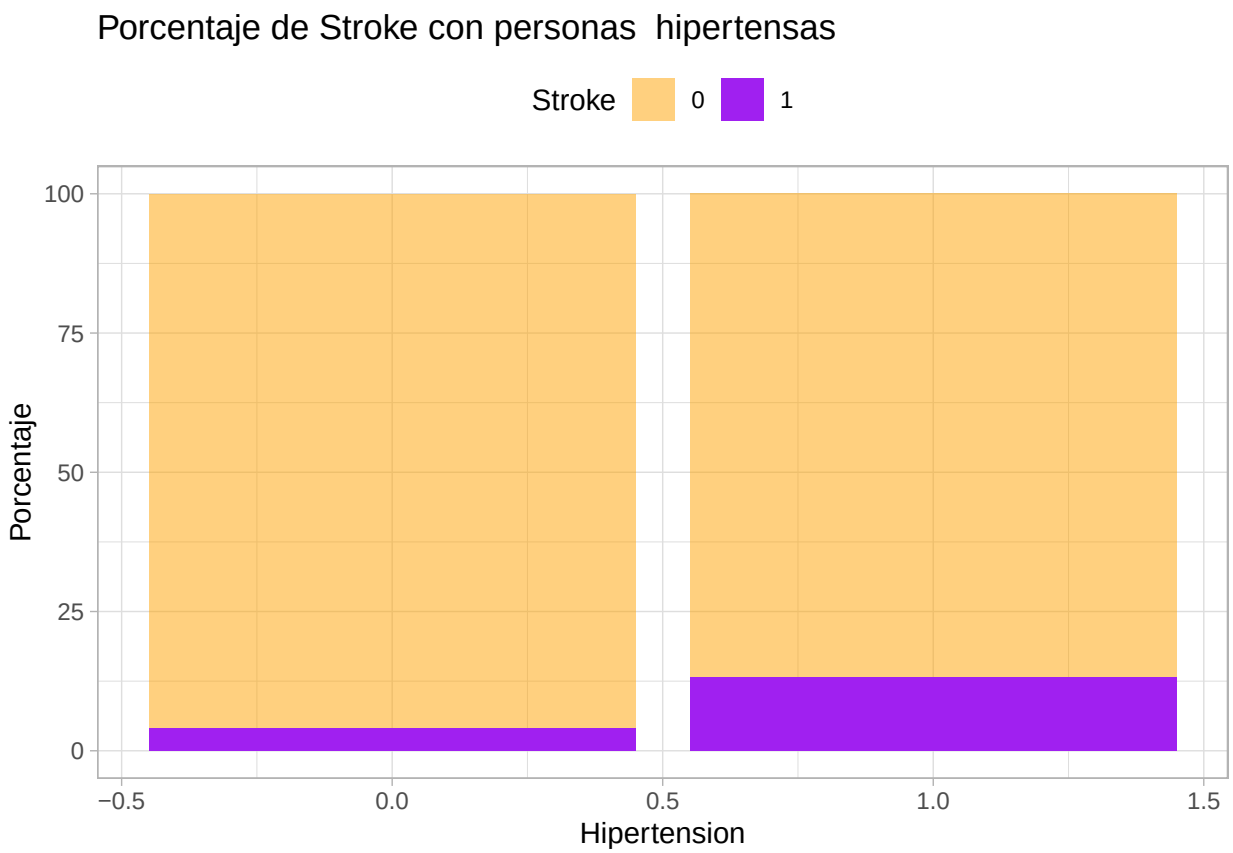
Tabla de contingencia

```
tabla = table(datos$smoking_status, datos$stroke)

tabla_contingencia <- addmargins(tabla)

num_columnas <- ncol(tabla_contingencia)

colores <- c(rgb(0.8, 0.6, 1.0), rgb(1.0, 0.8, 0.6))

tabla_formateada <- tabla_contingencia %>%
  kable(row.names = TRUE) %>%
  kable_styling(full_width = FALSE, position = "center") %>%
  column_spec(seq(1, num_columnas, by = 2), background = colores[1]) %>%
  column_spec(seq(2, num_columnas, by = 2), background = colores[2]) %>%
  column_spec(1,background = "white") %>%
  row_spec(5, background = "white")  %>%
  kable_styling(latex_options = c("hold_position"))

tabla_formateada
```

	0	1	Sum
formerly smoked	815	70	885
never smoked	1802	90	1892
smokes	747	42	789
Unknown	1497	47	1544
Sum	4861	249	5110

B.5) Relacion de la hipertension con accidentes cerebrovasculares

```
colores = c(rgb(1, 0.64, 0, alpha = 0.5), "purple")

# Calcular las frecuencias de stroke por grupo de hipertension

frecuencias <- datos %>%
  group_by(hypertension, stroke) %>%
  summarise(count = n()) %>%
  group_by(hypertension) %>%
  mutate(percentage = (count / sum(count)) * 100)

# Creamos el gráfico de barras
ggplot(frecuencias, aes(x = hypertension, y = percentage, fill = factor(stroke))) +
  geom_bar(stat = "identity", position = "stack") +
  labs(x = "Hipertension", y = "Porcentaje", fill = "Stroke") +
  scale_fill_manual(values = colores, name = "Stroke", labels = c("0", "1")) +
  theme_light() +
  theme(legend.position = "top") +
  ggtitle("Porcentaje de Stroke con personas  hipertensas")
```



A simple vista, parece que las personas que tienen hipertension, tienen una mayor posibilidad de sufrir un accidente cerebrovascular

Creamos su tabla de contingencia

```
tabla = table(datos$hypertension, datos$stroke)

tabla_contingencia <- addmargins(tabla)

num_columnas <- ncol(tabla_contingencia)

colores <- c(rgb(0.8, 0.6, 1.0), rgb(1.0, 0.8, 0.6))

# Nuevos nombres de filas
nombres_filas <- c("No hipertenso", "Hipertenso", "Sum")

# Asignar nuevos nombres de filas
rownames(tabla_contingencia) <- nombres_filas

tabla_formateada <- tabla_contingencia %>%
  kable(row.names = TRUE) %>%
  kable_styling(full_width = FALSE, position = "center") %>%
  column_spec(seq(1, num_columnas, by = 2), background = colores[1]) %>%
  column_spec(seq(2, num_columnas, by = 2), background = colores[2]) %>%
  column_spec(1,background = "white") %>%
  row_spec(3, background = "white")  %>%
  kable_styling(latex_options = c("hold_position"))
```

```
tabla_formateada
```

	0	1	Sum
No hipertenso	4429	183	4612
Hipertenso	432	66	498
Sum	4861	249	5110

Hagamos un chisq.test

$$\begin{cases} H_0 : No\ hay\ diferencia\ entre\ la\ frecuencia\ de\ sufrir\ stroke \\ H_1 : Hay\ diferencia\ significativa\ entre\ la\ frecuencia\ de\ sufrir\ stroke \end{cases}$$

```
chisq.test(datos$hypertension, datos$stroke)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  datos$hypertension and datos$stroke
## X-squared = 81.605, df = 1, p-value < 2.2e-16
```

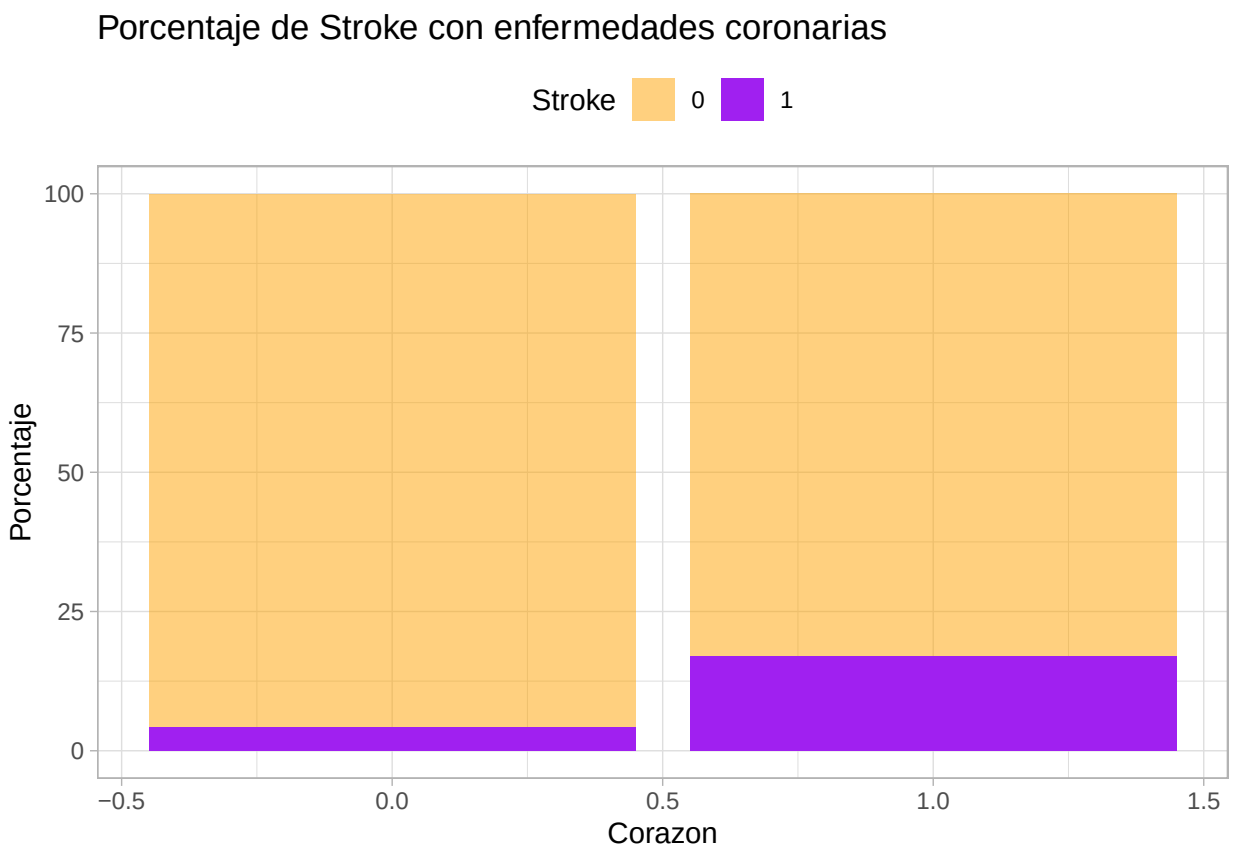
p-value $<$ 2.2e-16

Rechazamos la hipótesis nula

Su p-value es muy pequeño, por lo tanto el hecho de sufrir hipertension es bastante significativo en la posibilidad de sufrir un accidente cerebrovascular

B.6) Relacion de las enfermedades de corazon con accidentes cerebrovasculares

```
colores = c(rgb(1, 0.64, 0, alpha = 0.5), "purple")

# Calcular las frecuencias de stroke por enfermedades de corazon

frecuencias <- datos %>%
  group_by(heart_disease, stroke) %>%
  summarise(count = n()) %>%
  group_by(heart_disease) %>%
  mutate(percentage = (count / sum(count)) * 100)

# Creamos el gráfico de barras
ggplot(frecuencias, aes(x = heart_disease, y = percentage, fill = factor(stroke))) +
  geom_bar(stat = "identity", position = "stack") +
  labs(x = "Corazon", y = "Porcentaje", fill = "Stroke") +
  scale_fill_manual(values = colores, name = "Stroke", labels = c("0", "1")) +
  theme_light() +
  theme(legend.position = "top") +
  ggtitle("Porcentaje de Stroke con enfermedades coronarias")
```



A simple vista, parece que las personas que tienen enfermedades coronarias, tienen una mayor posibilidad de sufrir un accidente cerebrovascular

Creamos su tabla de contingencia

```
tabla = table(datos$heart_disease, datos$stroke)

tabla_contingencia <- addmargins(tabla)

num_columnas <- ncol(tabla_contingencia)

colores <- c(rgb(0.8, 0.6, 1.0), rgb(1.0, 0.8, 0.6))


# Nuevos nombres de filas
nombres_filas <- c("Sin enf. coronaria", "Con enf. coronaria", "Sum")

# Asignar nuevos nombres de filas
rownames(tabla_contingencia) <- nombres_filas

tabla_formateada <- tabla_contingencia %>%
  kable(row.names = TRUE) %>%
  kable_styling(full_width = FALSE, position = "center") %>%
  column_spec(seq(1, num_columnas, by = 2), background = colores[1]) %>%
  column_spec(seq(2, num_columnas, by = 2), background = colores[2]) %>%
  column_spec(1,background = "white") %>%
  row_spec(3, background = "white")  %>%
  kable_styling(latex_options = c("hold_position"))
```

```
tabla_formateada
```

	0	1	Sum
Sin enf. coronaria	4632	202	4834
Con enf. coronaria	229	47	276
Sum	4861	249	5110

$$\begin{cases} H_0 : No\ hay\ diferencia\ entre\ la\ frecuencia\ de\ sufrir\ stroke \\ H_1 : Hay\ diferencia\ significativa\ entre\ la\ frecuencia\ de\ sufrir\ stroke \end{cases}$$

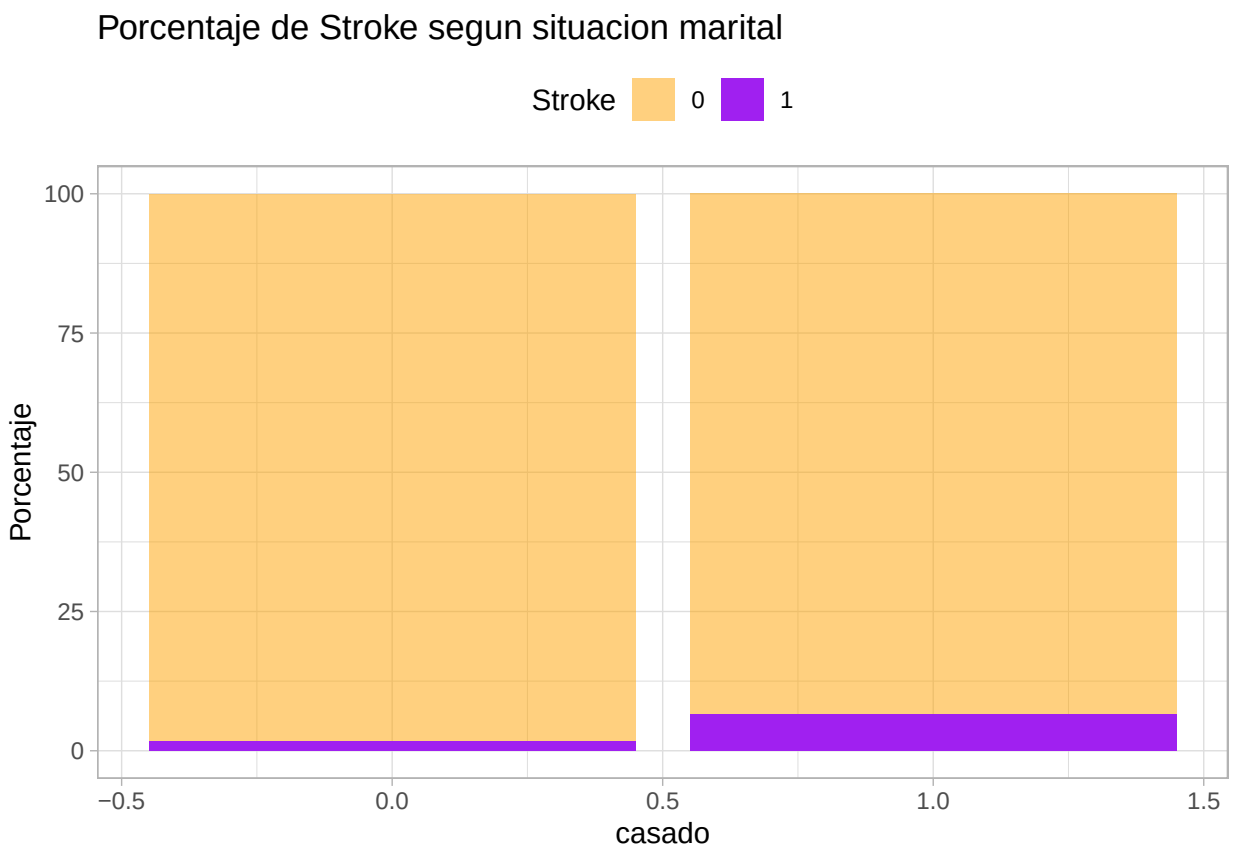
```
chisq.test(datos$heart_disease, datos$stroke)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  datos$heart_disease and datos$stroke
## X-squared = 90.26, df = 1, p-value < 2.2e-16
```

p-value es muy pequeño, por tanto la relacion entre stroke y enfermadades coronarias es significativa

B.7) Relacion de situacion marital con accidentes cerebrovasculares

```
colores = c(rgb(1, 0.64, 0, alpha = 0.5), "purple")

# Calcular las frecuencias de stroke por situacion marital

frecuencias <- datos %>%
  group_by(ever_married, stroke) %>%
  summarise(count = n()) %>%
  group_by(ever_married) %>%
  mutate(percentage = (count / sum(count)) * 100)

# Creamos el gráfico de barras
ggplot(frecuencias, aes(x = ever_married, y = percentage, fill = factor(stroke))) +
  geom_bar(stat = "identity", position = "stack") +
  labs(x = "casado", y = "Porcentaje", fill = "Stroke") +
  scale_fill_manual(values = colores, name = "Stroke", labels = c("0", "1")) +
  theme_light() +
  theme(legend.position = "top") +
  ggtitle("Porcentaje de Stroke segun situacion marital")
```



Creamos su tabla de contingencia

```
tabla = table(datos$ever_married, datos$stroke)

tabla_contingencia <- addmargins(tabla)

num_columnas <- ncol(tabla_contingencia)

colores <- c(rgb(0.8, 0.6, 1.0), rgb(1.0, 0.8, 0.6))


# Nuevos nombres de filas
nombres_filas <- c("No casado", "Casado", "Sum")

# Asignar nuevos nombres de filas
rownames(tabla_contingencia) <- nombres_filas

tabla_formateada <- tabla_contingencia %>%
  kable(row.names = TRUE) %>%
  kable_styling(full_width = FALSE, position = "center") %>%
  column_spec(seq(1, num_columnas, by = 2), background = colores[1]) %>%
  column_spec(seq(2, num_columnas, by = 2), background = colores[2]) %>%
  column_spec(1,background = "white") %>%
  row_spec(3, background = "white")  %>%
  kable_styling(latex_options = c("hold_position"))

tabla_formateada
```

	0	1	Sum
No casado	1728	29	1757
Casado	3133	220	3353
Sum	4861	249	5110

$$\begin{cases} H_0 : No\ hay\ diferencia\ entre\ la\ frecuencia\ de\ sufrir\ stroke \\ H_1 : Hay\ diferencia\ significativa\ entre\ la\ frecuencia\ de\ sufrir\ stroke \end{cases}$$

```
chisq.test(datos$ever_married, datos$stroke)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  datos$ever_married and datos$stroke
## X-squared = 58.924, df = 1, p-value = 1.639e-14
```

p-value es muy pequeño, por tanto la relacion entre stroke y estado marital es significativa

B.8) Matriz de correlaciones

Para calcular la matriz de correlación entre uvarias variables categóricas utilizamos el coeficiente de correlación Cramer's V. Este coeficiente es especialmente adecuado para medir la asociación entre variables categóricas.

El coeficiente de Cramer varía de 0 a 1 donde:

0 indica ninguna asociación entre las dos variables. 1 indica una fuerte asociación entre las dos variables.

Se calcula de la siguiente manera:

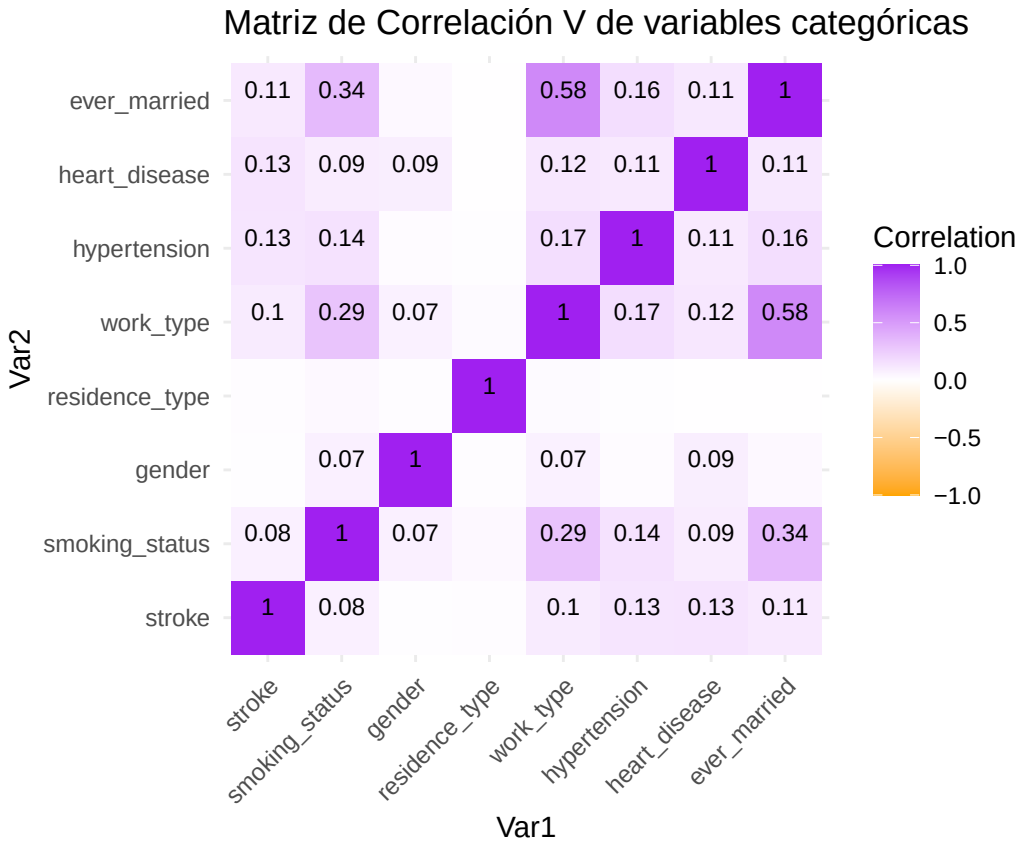
$$V \text{ de Cramer} = \frac{\sqrt{X^2/n}}{\min(c-1, r-1)}$$

X^2: La estadística de Chi-cuadrado n: Tamaño total de la muestra r: Número de filas c: Número de columnas

En R, la función cramerV() del paquete "rcompanion" calcula V utilizando la función chisq.test() del paquete "stats".

```
subset_data <- datos[, c("stroke", "smoking_status", "gender", "residence_type",
                         "work_type", "hypertension", "heart_disease", "ever_married" )]

# Creamos una matriz vacía para almacenar los coeficientes de Cramer
num_vars <- ncol(subset_data)
cramer_matrix <- matrix(0, nrow = num_vars, ncol = num_vars)
colnames(cramer_matrix) <- rownames(cramer_matrix) <- colnames(subset_data)

# Llenamos la matriz con los coeficientes de Cramer
for (i in 1:num_vars) {
  for (j in 1:num_vars) {
    cramer_matrix[i, j] <- cramerV(subset_data[, i], subset_data[, j])
  }
}


threshold <- 0.05
```

```r
filtered_cramer_data <- as.data.frame(as.table(cramer_matrix))
filtered_cramer_data <- filtered_cramer_data[filtered_cramer_data$Freq > threshold, ]

p <- ggplot(data = as.data.frame(as.table(cramer_matrix))) +
  geom_tile(aes(Var1, Var2, fill = Freq)) +
  geom_text(data = filtered_cramer_data, aes(Var1, Var2, label = round(Freq, 2)),
            color = "black", size = 3, nudge_y = 0.15) +
  scale_fill_gradient2(low = "orange", high = "purple", mid = "white",
                       midpoint = 0, limit = c(-1, 1), space = "Lab",
                       name = "Correlation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_fixed() +
  labs(title = "Matriz de Correlación V de variables categóricas")

print(p)
```



B.9) Tabla de variables categóricas

```r
subset_data <- datos[, c( "smoking_status", "gender", "residence_type",
                          "work_type", "hypertension", "heart_disease",
                          "ever_married","stroke")]



# Calculamos las estadísticas por categoría para una columna dada
```

```r
estadisticos_categoricos <- function(data, column) {
  data %>%
    group_by(.data[[column]]) %>%
    summarise(
      total_count = n(),
      count_s0 = sum(stroke == 0),
      count_s1 = sum(stroke == 1),
      percentage_s0 = round(count_s0 / total_count * 100, 2),
      percentage_s1 = round(count_s1 / total_count * 100, 2)
    )
}



summary_list <- list()

# Iterar a través de las columnas del subconjunto de datos
for (col in names(subset_data)) {
  if (col != "stroke") {
    summary <- estadisticos_categoricos(subset_data, col)
    colnames(summary)[1] <- "Variable & p-value"
    summary_list[[col]] <- summary
  }
}

# Combina los dataframes en dataframe_list usando rbind
combinadas <- do.call(rbind, summary_list)

colnames(combinadas) <- c("Variable & p-value","Número total", "Diagnostico negativo",
                          "Diagnostico positivo", "Porcentaje diagnostico negativo",
                          "Porcentaje diagnostico positivo")


colores = c(rgb(1.0, 0.8, 0.6), rgb(0.8, 0.6, 1.0))

combinadas %>%
  kbl(caption = "Tabla de estadísticos de variable categóricas") %>%
  pack_rows("Tipo de fumador; p-value = 2.085e-06", 1, 4) %>%
  pack_rows("Género; p-value =  0.7895, NO SIGNIFICATIVO", 5, 7) %>%
  pack_rows("Tipo de vivienda; p-value = 0.2983, NO SIGNIFICATIVO", 8, 9) %>%
  pack_rows("Tipo de trabajo; p-value = 1.534e-10", 10, 14) %>%
  pack_rows("Hipertension; p-value =2.2e-16", 15, 16) %>%
  pack_rows("Enfermedad coronaria; p-value = 2.2e-16", 17, 18) %>%
  pack_rows("Casado alguna vez; p-value =  1.639e-14", 19, 20) %>%
  kable_styling(full_width = FALSE, position = "center") %>%
  column_spec(c(3, 5), background = colores[1]) %>%
  column_spec(c(4, 6), background = colores[2]) %>%
  kable_styling(latex_options = c("hold_position")) %>%
  kable_styling(latex_options = "scale_down")
```

Como hemos visto, hay 2 variables No significativas. gender y residence_type. Además hay 2 variables que hemos creado nosotros que son imc_str y grupo_edad, a partir de otras para explicar el dataset. Vamos a crear un nuevo archivo csv para realizar los modelos predictivos sobre el mismo. Que no tenga estas columnas

Table 2: Tabla de estadísticos de variable categóricas

Variable & p-value	Número total	Diagnostico negativo	Diagnostico positivo	Porcentaje diagnostico negativo	Porcentaje diagnostico positivo
Tipo de fumador; p-value = 2.085e-06					
Unknown	1544	1497	47	96.96	3.04
formerly smoked	885	815	70	92.09	7.91
never smoked	1892	1802	90	95.24	4.76
smokes	789	747	42	94.68	5.32
Género; p-value = 0.7895, NO SIGNIFICATIVO					
Female	2994	2853	141	95.29	4.71
Male	2115	2007	108	94.89	5.11
Other	1	1	0	100.00	0.00
Tipo de vivienda; p-value = 0.2983, NO SIGNIFICATIVO					
Rural	2514	2400	114	95.47	4.53
Urban	2596	2461	135	94.80	5.20
Tipo de trabajo; p-value = 1.534e-10					
Govt_job	654	621	33	94.95	5.05
Never_worked	10	10	0	100.00	0.00
Private	2896	2747	149	94.85	5.15
Self-employed	813	748	65	92.00	8.00
children	737	735	2	99.73	0.27
Hipertension; p-value =2.2e-16					
0	4612	4429	183	96.03	3.97
1	498	432	66	86.75	13.25
Enfermedad coronaria; p-value = 2.2e-16					
0	4834	4632	202	95.82	4.18
1	276	229	47	82.97	17.03
Casado alguna vez; p-value = 1.639e-14					
0	1757	1728	29	98.35	1.65
1	3353	3133	220	93.44	6.56

Este nuevo archivo tendrá los valores NA de bmi de la mediana. Es decir, imputaremos por la mediana, poque hay muchas observaciones anómalas y la mediana es mas robusta que la media

```
mediana_bmi <- median(datos$bmi, na.rm = TRUE)

datos$bmi[is.na(datos$bmi)] <- mediana_bmi


# Eliminamos las columnas
columnas_a_eliminar <- c("gender", "grupo_edad", "imc_str", "residence_type")

datos_finales <- datos[, !(names(datos) %in% columnas_a_eliminar)]

head(datos_finales)
```

```
##   age hypertension heart_disease ever_married     work_type avg_glucose_level
## 1  67            0             1            1       Private            228.69
## 2  61            0             0            1 Self-employed            202.21
## 3  80            0             1            1       Private            105.92
## 4  49            0             0            1       Private            171.23
## 5  79            1             0            1 Self-employed            174.12
## 6  81            0             0            1       Private            186.21
##    bmi  smoking_status stroke
## 1 36.6 formerly smoked      1
## 2 28.1    never smoked      1
## 3 32.5    never smoked      1
## 4 34.4          smokes      1
## 5 24.0    never smoked      1
## 6 29.0 formerly smoked      1
```

```r
write.csv(datos_finales, file = "/home/guincho/Desktop/stroke_final.csv", row.names = FALSE)
```