Prueba 1 - ANOVA

Rubén Sierra Serrano y Alfredo Robledano Abasolo

2023-03-20

Creación y descripción de la base de datos.

El primer paso para trabajar sobre una base de datos es, conocerla, es decir, conocer el contenido de la base de datos y, además, ser capaces de clasificar las diferentes variables en función de su clase o, lo que es lo mismo, saber distinguir las variables continuas de las discretas (categóricas).

Otro punto interesante es el de crear nuevas variables dependiendo de los resultados que estamos buscando según nuestros intereses.

Leemos a continuación la base de datos:

```
 id gender age hypertension heart_disease ever_married     work_type
```

1 9046 Male 67 0 1 Yes Private 2 51676 Female 61 0 0 Yes Self-employed 3 31112 Male 80 0 1 Yes Private 4 60182 Female 49 0 0 Yes Private 5 1665 Female 79 1 0 Yes Self-employed 6 56669 Male 81 0 0 Yes Private Residence_type avg_glucose_level bmi smoking_status stroke 1 Urban 228.69 36.6 formerly smoked 1 2 Rural 202.21 NA never smoked 1 3 Rural 105.92 32.5 never smoked 1 4 Urban 171.23 34.4 smokes 1 5 Rural 174.12 24.0 never smoked 1 6 Urban 186.21 29.0 formerly smoked 1

Encontramos una base de datos que consta de 5110 observaciones y de 12 variables.

Algunas de estas variables sirven para identificar a cada una de las personas que se ha hecho seguimiento, como son la edad o el género. Otras, sin embargo, hacen referencia a temas de salud, como el nivel de glucosa, el índice de masa corporal o si tienen o han tenido problemas cardiovasculares.

Entendemos que es interesante, para este estudio, comprobar qué factores pueden influir sobre el índice de masa corporal (BMI en inglés), tomamos esta como nuestra variable objetivo.

Por último, encontramos variables que hacen referencia a hábitos de personas, como pueden ser el tipo de trabajo y con que regularidad fuman. • age: Edad de cada individuo. • avg_glucose_level: Nivel de glugosa promedio de cada individuo • bmi: Índice de masa corporal de cada individuo • gender: Sexo de los pacientes. • hypertension: Determina si los pacientes tienen hipertensión. • heart_disease: Determina si los pacientes han tenido problemas cardiacos. • work_type: Tipo de trabajo que ejerce el paciente. • ever_married: Determina si los pacientes se han casado. • Residence_type: Lugar de residencia de los pacientes. • smoking_status: Nivel de tabaquismo de los pacientes. • stroke: Determina si el paciente ha tenido previamente infartos. Vamos a eliminar la variable id ya que se trata de un identificador empleado para diferenciar a los pacientes:

Limpieza de la base de datos - estadística descriptiva

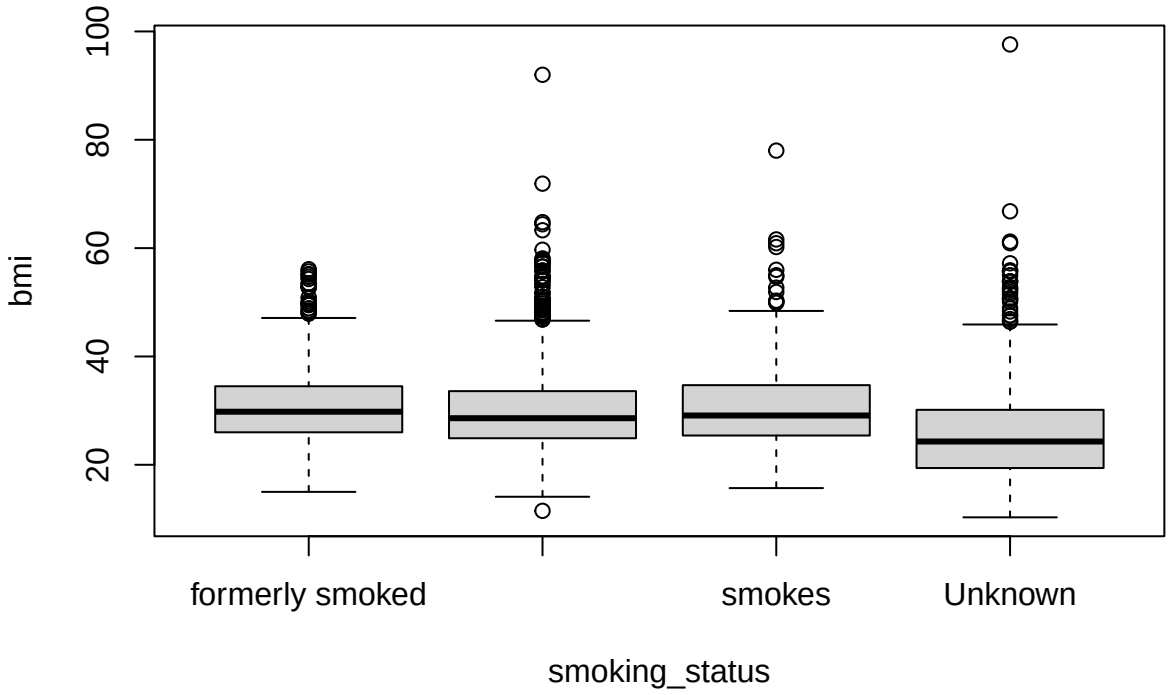
Una vez hemos definido las variables y sabemos lo que significan, tenemos que depurar la base de datos para poder trabajar con ella.

En la variable gender, observamos que hay una única obsevación (no significativa) que adopta el valor Other, la convertimos en NA:

```
df$gender <- ifelse(df$gender == "Other", NA, df$gender)
```

De las variables que hemos elegido el índice de masa corportal (bmi) tiene 3.9% de variables que son NA, por tanto, eliminaremos esas observaciones. Por otro lado existe un número significativo de observaciones que tienen variable smoking_status de valor Unknown, realizamos un breve análisis de su relación con el bmi.

```
boxplot(bmi~smoking_status, data=df)
```

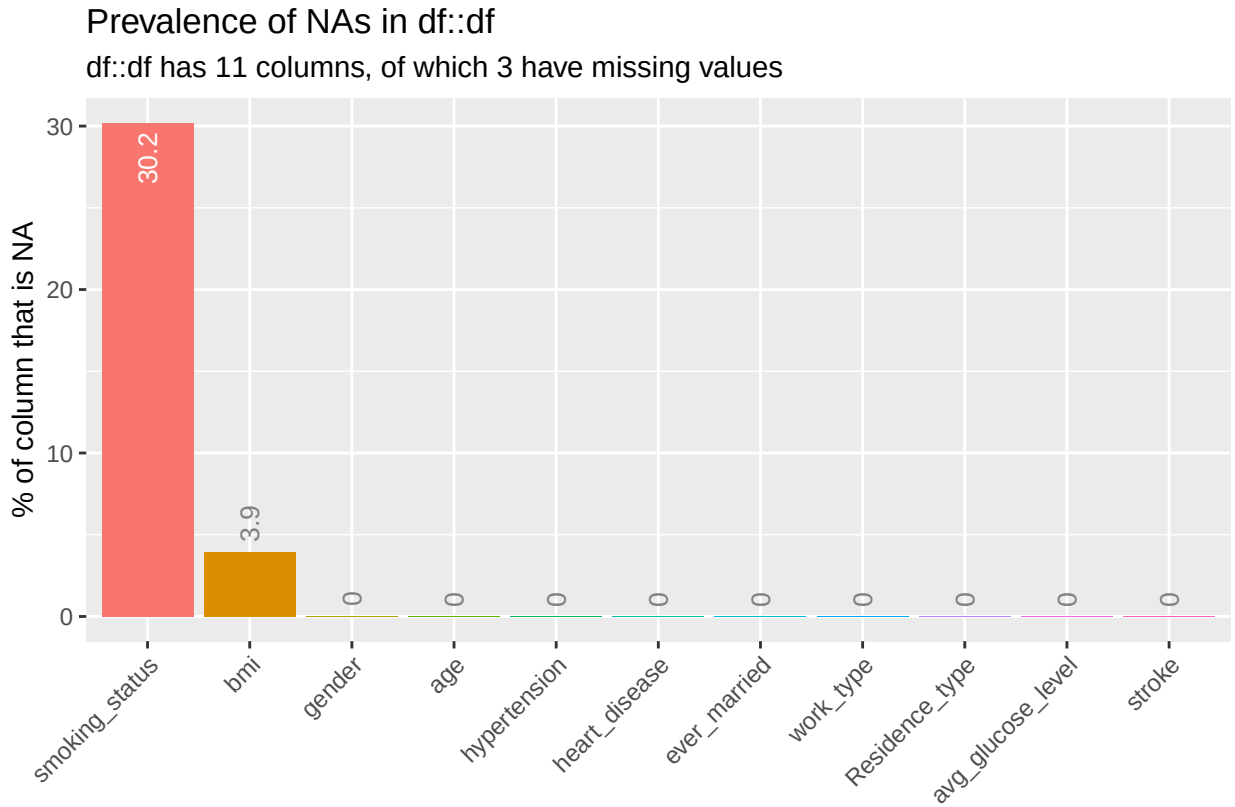


Si realizamos un diagrama de cajas que relacione las variables bmi y smoking_status observamos que entre las variables que no son unknown no hay mucha diferencia, pero esta si difiere y por tanto al realizar anova saldría como valor de variable significativo unknown pero realmente no sabríamos porque al no especificarse un valor concreto. Descartamos esas observaciones al no darnos información útil.

```
df$smoking_status <- ifelse(df$smoking_status == "Unknown", NA, df$smoking_status)
```

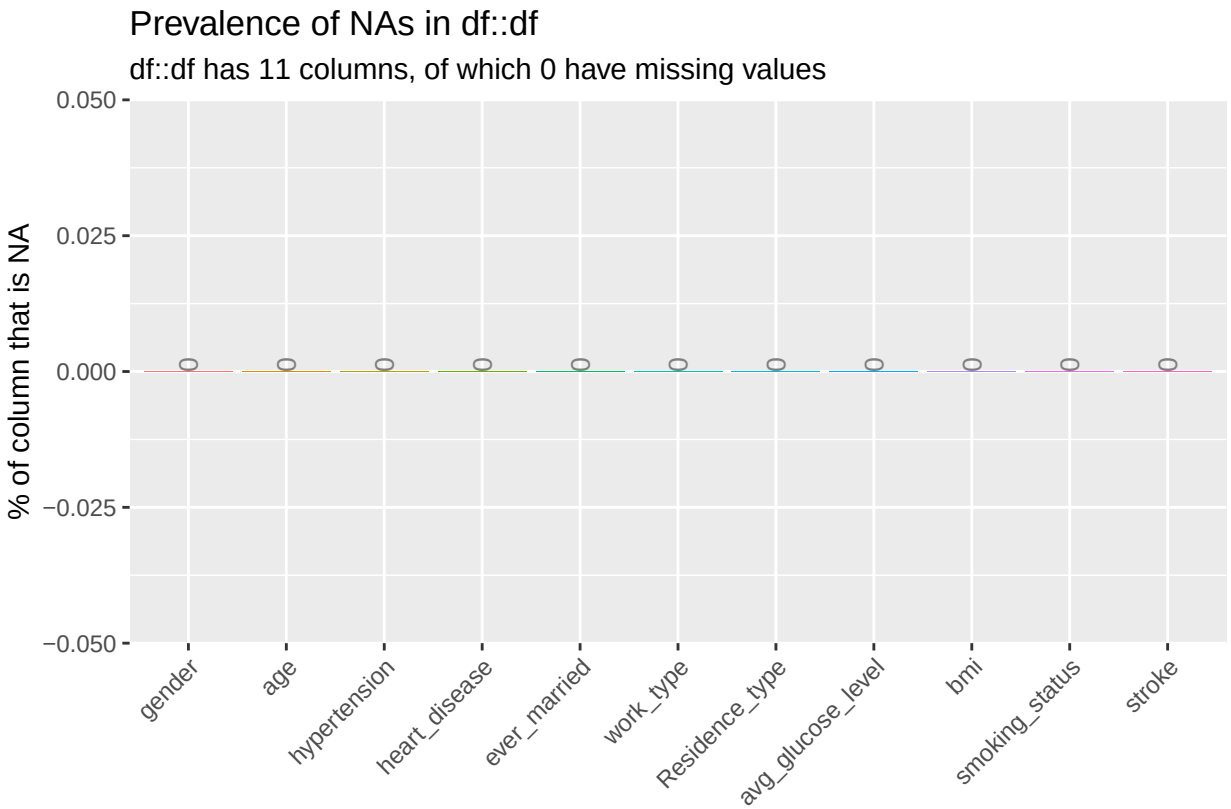
Hemos convertido en NA todos aquellos valores de variables de observaciones no útiles.

```
library(inspectdf);
show_plot(inspect_na(df))
```



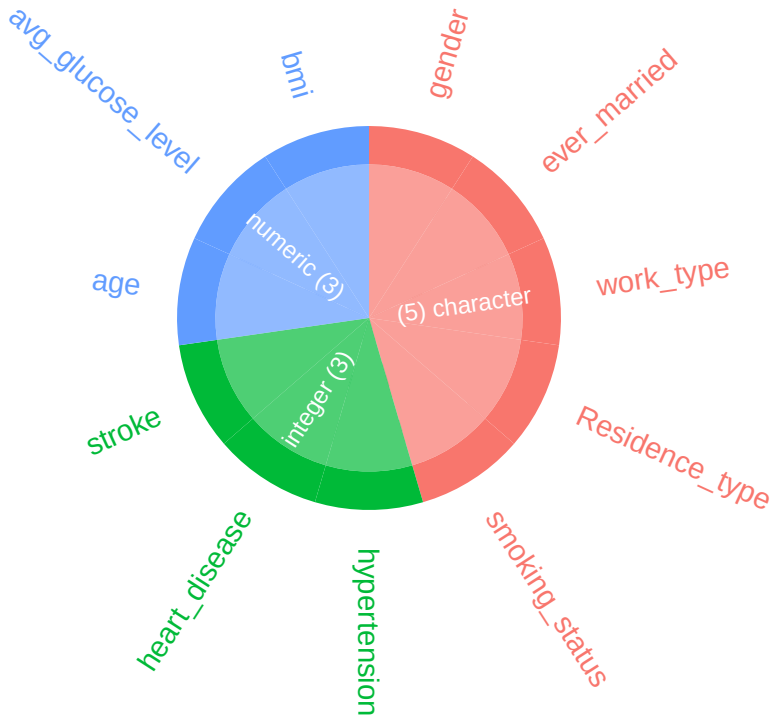
Eliminamos todas aquellas observaciones no útiles.

```
df <- na.omit(df)
show_plot(inspect_na(df))
```



Lo siguiente que debemos realizar es un estudio de la naturaleza de las variables, es decir, asignar el tipo factor a aquellas que consideremos categóricas y comprobar que las numéricas estén bien identificadas:

```
show_plot(inspect_types(df))
```



Tenemos 5 variables de tipo char que pasaremos a factor. Además, debemos pasar a factor las variables que representan categorías aún siendo variables numéricas, en nuestro caso las variable hypertension, heart_disease y stroke.

```
df[,c(1, 3, 4, 5, 6, 7, 10, 11)] <- lapply(df[,c(1, 3, 4, 5, 6, 7, 10, 11)], factor)
summary(df)
```

```
##     gender          age        hypertension heart_disease ever_married
##  Female:2086   Min.   :10.00   0:3017       0:3219        No : 826
##  Male  :1339   1st Qu.:34.00   1: 408       1: 206        Yes:2599
##                Median :50.00
##                Mean   :48.65
##                3rd Qu.:63.00
##                Max.   :82.00
##          work_type    Residence_type avg_glucose_level      bmi
##  children     :  68   Rural:1680     Min.   : 55.12    Min.   :11.50
##  Govt_job     : 514   Urban:1745     1st Qu.: 77.23    1st Qu.:25.30
##  Never_worked :  14                  Median : 92.35    Median :29.10
##  Private      :2200                  Mean   :108.31    Mean   :30.29
##  Self-employed: 629                  3rd Qu.:116.20    3rd Qu.:34.10
##                                      Max.   :271.74    Max.   :92.00
##          smoking_status stroke
##  formerly smoked: 836   0:3245
##  never smoked   :1852   1: 180
##  smokes         : 737
##
```

```
##
##
```

Observamos que hay 3 variables continuas, de las que posteriormente estudiaremos las distribuciones y se tomarán las decisiones oportunas: • age: Edad de cada individuo. • avg_glucose_level: Nivel de glugosa promedio de cada individuo • bmi: Índice de masa corporal de cada individuo En contraposición, hay 5 variables discretas: • gender: Sexo de los pacientes. • hypertension: Determina si los pacientes tienen hipertensión. • heart_disease: Determina si los pacientes han tenido problemas cardiacos. • work_type: Tipo de trabajo que ejerce el paciente. • ever_married: Determina si los pacientes se han casado. • Residence_type: Lugar de residencia de los pacientes. • smoking_status: Nivel de tabaquismo de los pacientes. • stroke: Determina si el paciente ha tenido previamente infartos.

Nos podemos percatar que la variable edad puede tomar valores no enteros, por simplicidad convertiremos la edad a valores enteros y trabajaremos en meses porque parte de la muestra son bebés.

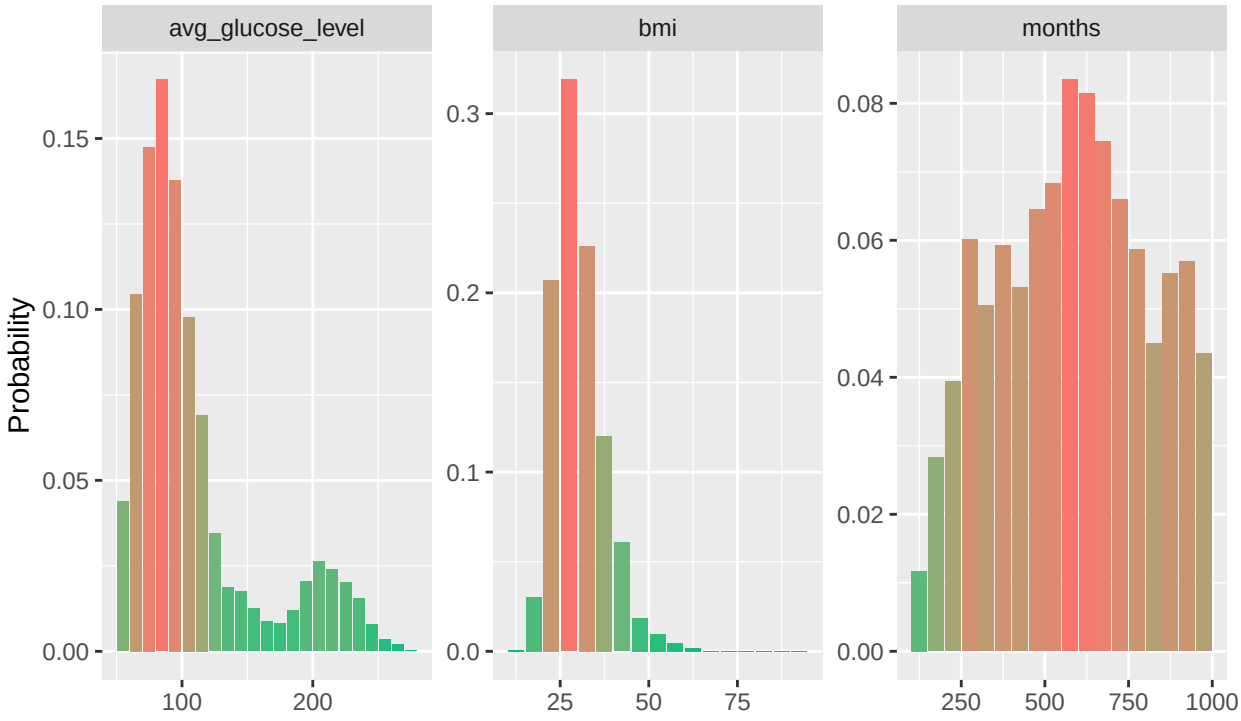
```
df[2]<-lapply(df[2],function(x) round(x*12))
colnames(df)[2]  <- "months"
summary(df)
```

```
##     gender         months      hypertension heart_disease ever_married
##  Female:2086   Min.   :120.0   0:3017       0:3219        No : 826
##  Male  :1339   1st Qu.:408.0   1: 408       1: 206        Yes:2599
##                Median :600.0
##                Mean   :583.8
##                3rd Qu.:756.0
##                Max.   :984.0
##          work_type    Residence_type avg_glucose_level      bmi
##  children     :  68   Rural:1680     Min.   : 55.12    Min.   :11.50
##  Govt_job     : 514   Urban:1745     1st Qu.: 77.23    1st Qu.:25.30
##  Never_worked :  14                  Median : 92.35    Median :29.10
##  Private      :2200                  Mean   :108.31    Mean   :30.29
##  Self-employed: 629                  3rd Qu.:116.20    3rd Qu.:34.10
##                                      Max.   :271.74    Max.   :92.00
##          smoking_status stroke
##  formerly smoked: 836   0:3245
##  never smoked   :1852   1: 180
##  smokes         : 737
##
##
##
```

Analizando el summary, la media y el tercer cuartil del IMC y de los niveles de glucosa distan mucho de sus valores máximos, por tanto, podemos concluir que tiene valores outliners.

Lo primero que tenemos que hacer para eliminar los outliners es graficar las funciones numéricas para observa cuánto se aleja la cola del resto de la función:

```
library(inspectdf);
show_plot(inspect_num(df))
```



En el IMC, los outliners son los valores que se escapan por las colas, de tal forma que una persona es muy extraño que tenga un IMC ínfimo o superior al 50%. Por tanto, nuestros outliners serán los valores menores a un 15% y mayores a un 50%:

```
df <- subset(df, bmi > 15 & bmi < 50)
summary(df)
```

```
##     gender         months      hypertension heart_disease ever_married
##  Female:2042   Min.   :120.0   0:2972       0:3159        No : 814
##  Male  :1322   1st Qu.:405.0   1: 392       1: 205        Yes:2550
##                Median :600.0
##                Mean   :584.3
##                3rd Qu.:756.0
##                Max.   :984.0
##          work_type    Residence_type avg_glucose_level      bmi
##  children     :  68   Rural:1647     Min.   : 55.12    Min.   :15.30
##  Govt_job     : 505   Urban:1717     1st Qu.: 77.22    1st Qu.:25.20
##  Never_worked :  14                  Median : 92.30    Median :28.95
##  Private      :2156                  Mean   :108.04    Mean   :29.86
##  Self-employed: 621                  3rd Qu.:116.00    3rd Qu.:33.70
##                                      Max.   :271.74    Max.   :49.90
##          smoking_status stroke
##  formerly smoked: 822   0:3185
##  never smoked   :1817   1: 179
##  smokes         : 725
##
```

```
##
##
```

En los niveles de glucosa, los outliners son los valores que se escapan por las colas, de tal forma que una persona es muy extraño que tenga un nivel de glucosa superior a los 250 mg/dL:

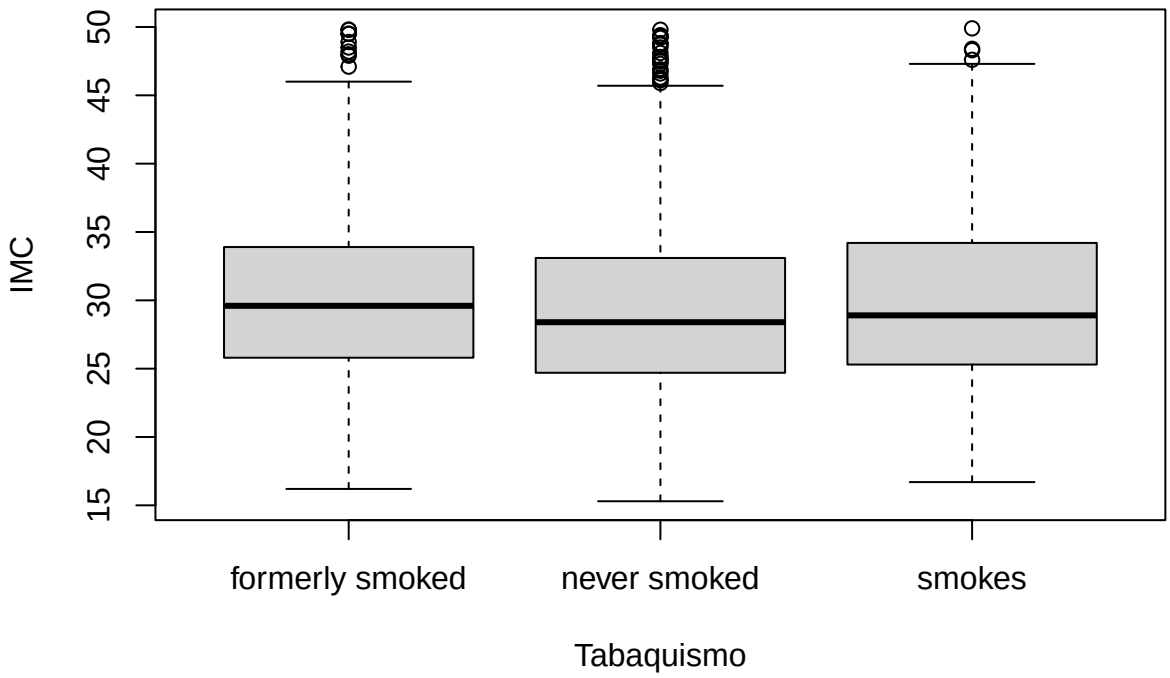
```
df <- subset(df, avg_glucose_level< 225)
summary(df)
```

```
##     gender         months      hypertension heart_disease ever_married
##  Female:1976   Min.   :120.0   0:2874       0:3049        No : 804
##  Male  :1257   1st Qu.:396.0   1: 359       1: 184        Yes:2429
##                Median :588.0
##                Mean   :577.2
##                3rd Qu.:744.0
##                Max.   :984.0
##          work_type    Residence_type avg_glucose_level      bmi
##  children     :  68   Rural:1579     Min.   : 55.12    Min.   :15.30
##  Govt_job     : 488   Urban:1654     1st Qu.: 76.46    1st Qu.:25.10
##  Never_worked :  14                  Median : 91.02    Median :28.80
##  Private      :2077                  Mean   :102.76    Mean   :29.74
##  Self-employed: 586                  3rd Qu.:111.96    3rd Qu.:33.50
##                                      Max.   :224.71    Max.   :49.90
##          smoking_status stroke
##  formerly smoked: 782   0:3075
##  never smoked   :1752   1: 158
##  smokes         : 699
##
##
##
```

Análisis a priori - influencia de las variables.

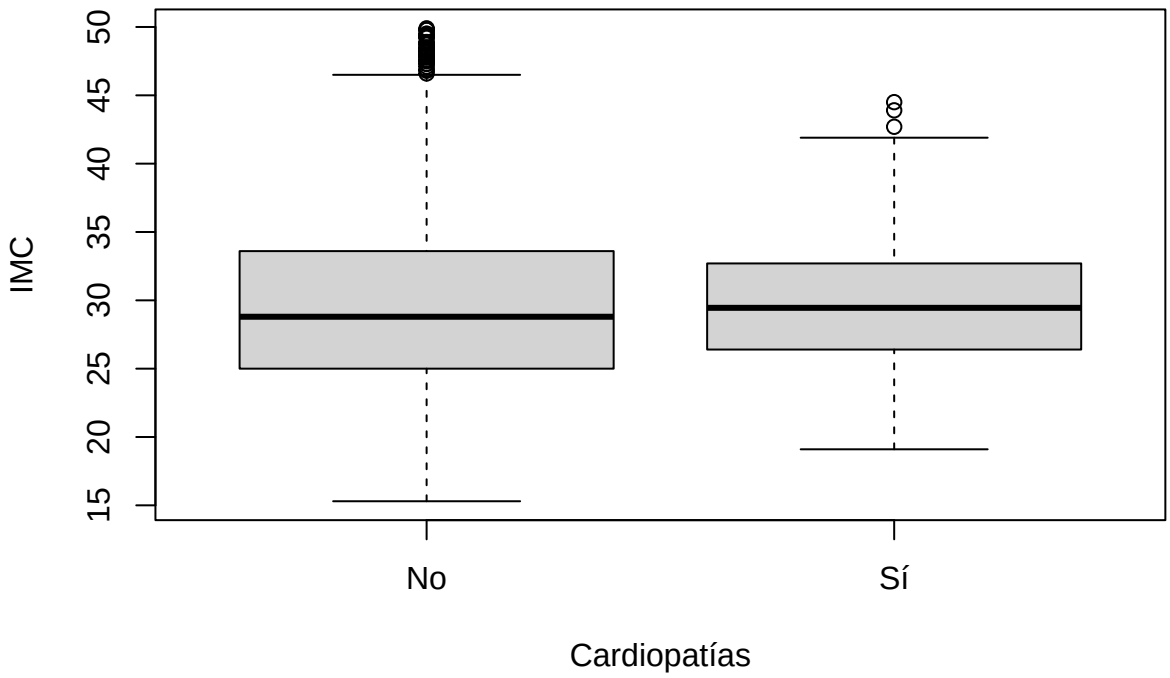
Ahora que conocemos la base de datos y la hemos depurado, vamos a empezar a sacar algunas conclusiones. Por ejemplo, nos podría interesar comprobar si el índice de masa corporal viene influido por el tabaquismo:

```
boxplot(bmi ~ smoking_status, data = df, ylab = "IMC", xlab = "Tabaquismo")
```

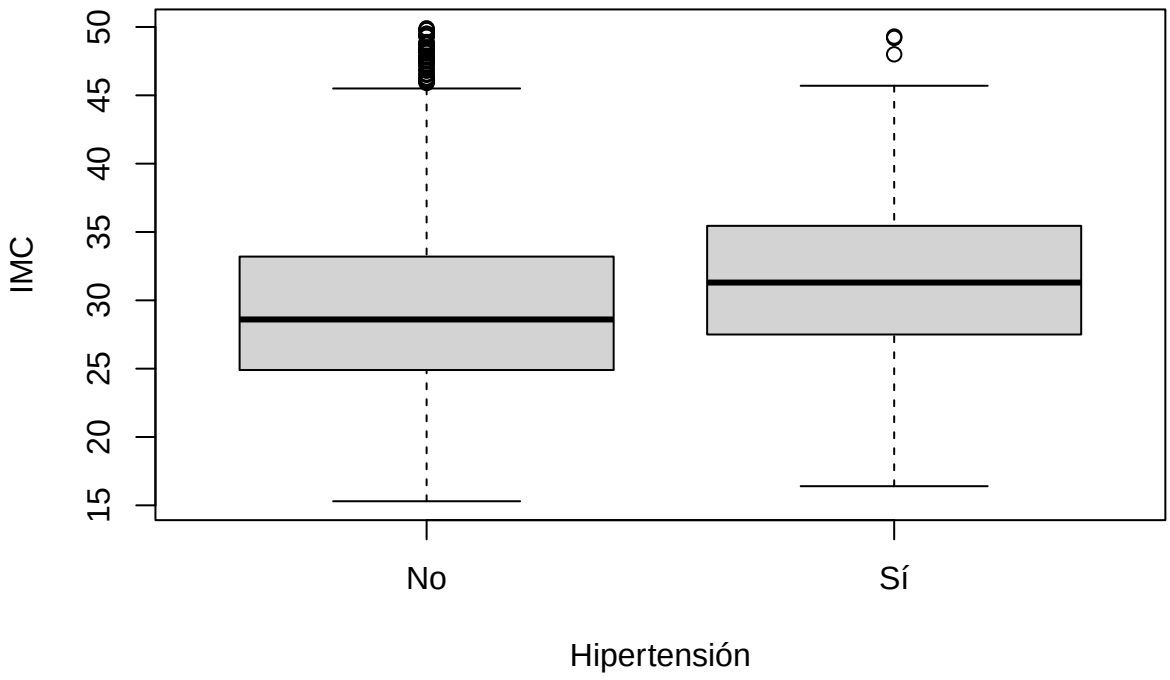


También podría ser interesante mirar la influencia entre los problemas cardiacos, tener hipertensión, el sexo o si viven en el campo y el IMC:

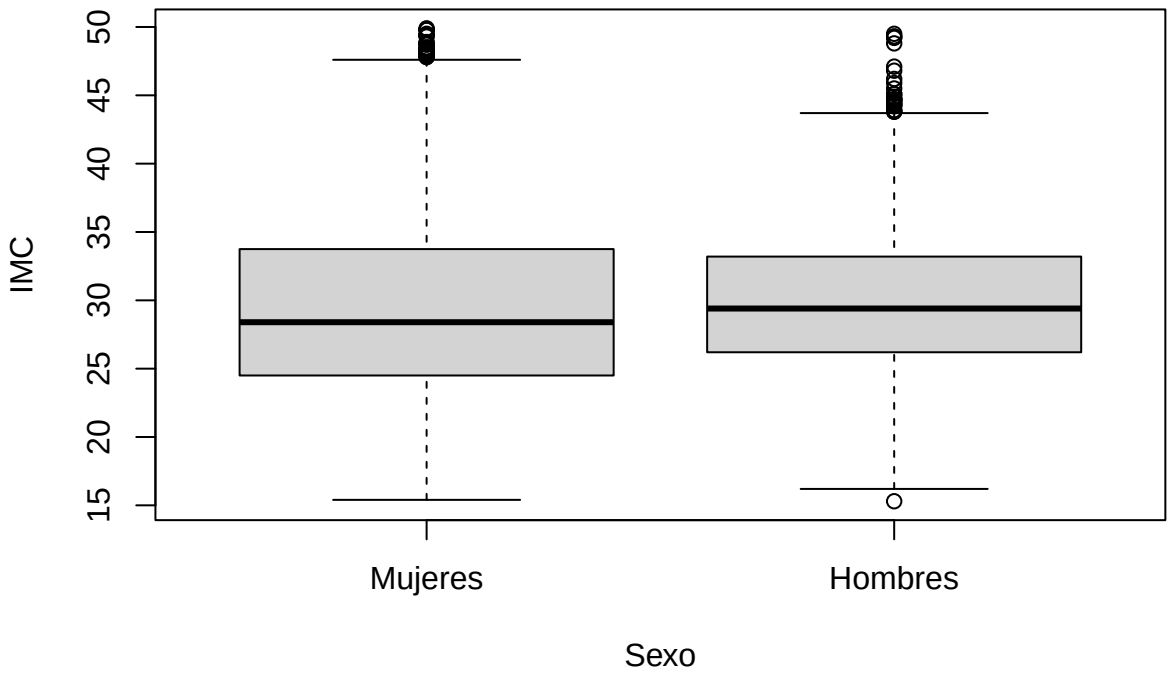
```
boxplot(bmi ~ heart_disease, data = df, ylab = "IMC", xlab = "Cardiopatías", names = c("No", "Sí"))
```



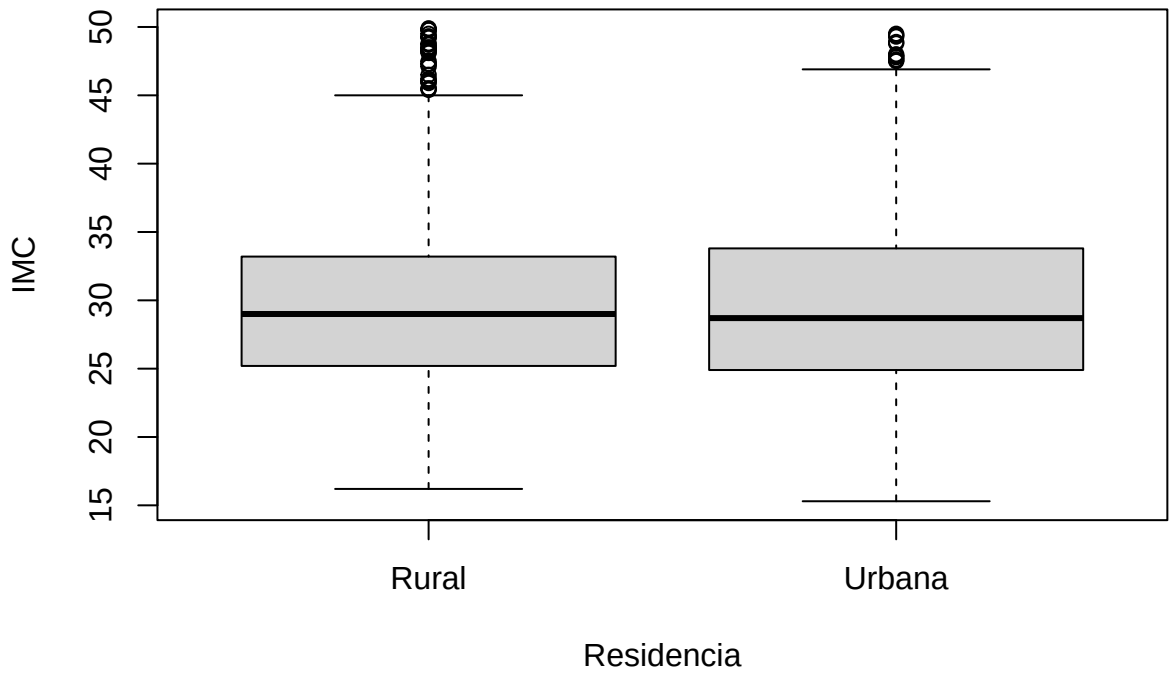

```
boxplot(bmi ~ hypertension, data = df, ylab = "IMC", xlab = "Hipertensión", names = c("No", "Sí"))
```




```
boxplot(bmi  ~  gender, data = df, ylab = "IMC", xlab = "Sexo", names = c("Mujeres", "Hombres") )
```



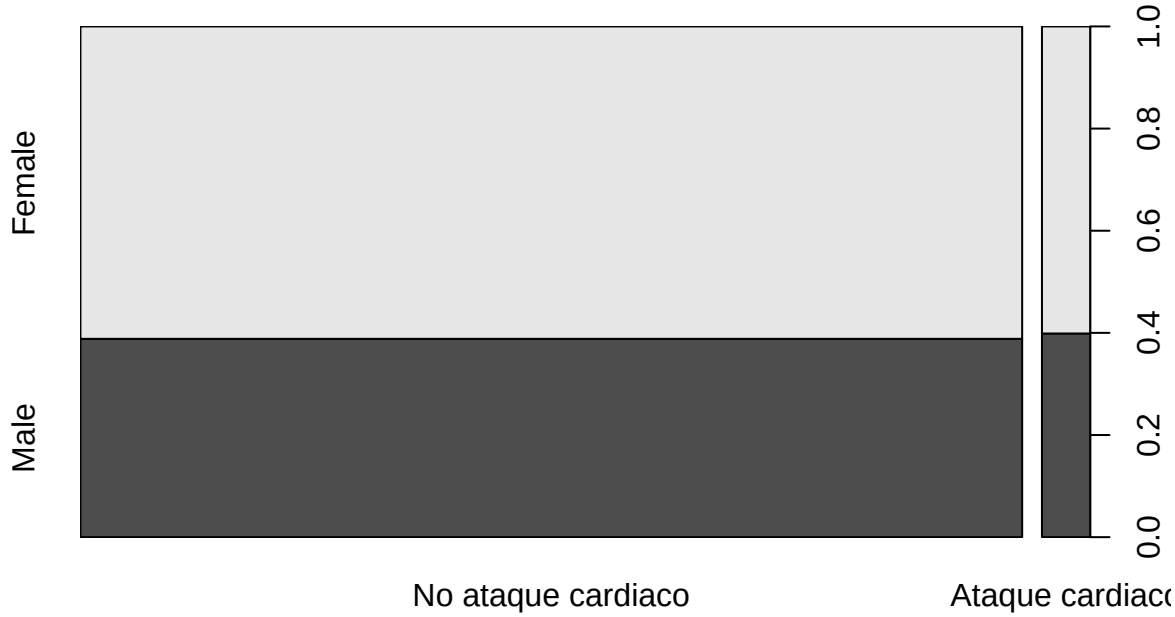

```
boxplot(bmi  ~  Residence_type, data = df, ylab = "IMC", xlab = "Residencia", names = c("Rural", "Urbana
```



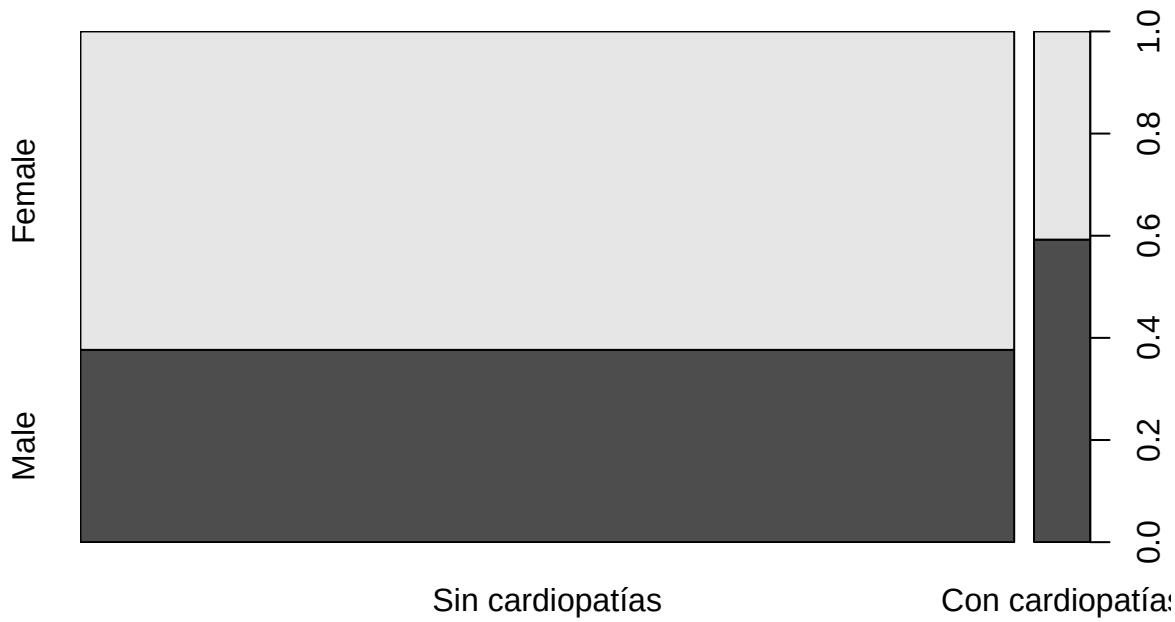
Vamos a comprobar, de forma gráfica, si el sexo puede afectar a la hipertensión, los ataques cardiacos y las cardiopatías. De esta forma, en el futuro, podremos realizar más comprobaciones de forma analítica.

```
spineplot(table(df$hypertension,df$gender), xaxlabels = c("Sin hipertensión", "Con hipertensión"))
```


```
spineplot(table(df$stroke,df$gender), xaxlabels = c("No ataque cardiaco", "Ataque cardiaco"))
```




```
spineplot(table(df$heart_disease,df$gender), xaxlabels = c("Sin cardiopatías", "Con cardiopatías"))
```



Intervalo de confianza En este apartado crearemos un intervalo de confianza para el IMC. Hemos visto anteriormente el estimador puntual para este valor, aproximadamente el 29.7%, pero vamos a crear un intervalo de confianza para este valor y poder estimar cuánto valdría este valor en la población.

Queremos hacer el intervalo de confianza para el IMC **media**.

```
# Intervalo para la media

conf <- t.test(x =  df$bmi,                    # Muestra 1
               y = NULL,                       # Muestra 2
               alternative = c("two.sided"),   # Tipo de intervalo
               paired = FALSE,                 # Variables dependientes
               var.equal = FALSE,              # Varianzas iguales
               conf.level = 0.95)              # Nivel de confianza (1-nivel significación)

# Muestro el resultado del intervalo

conf$conf.int
```

```
## [1] 29.51491 29.95574
## attr(,"conf.level")
## [1] 0.95
```

Contraste de hipótesis

Hemos visto de forma gráfica, que la residencia no afectaba al IMC pero tener hipertensión sí podía influir.

Crearemos un primer contraste de forma analítica para comprobar que la media de IMC entre los que viven en el campo y en la ciudad es igual. Es decir, queremos plantear el siguiente contraste de hipótesis:

$$\begin{cases} H_0: & \mu_{rural} = \mu_{urbano} \\ H_1: & \mu_{rural} \neq \mu_{urbano} \end{cases}$$

```
# Contraste para la diferencia de medias
rural <- df[df$Residence_type == 'Rural', 'bmi']
urbano <- df[df$Residence_type == 'Urban', 'bmi']

t.test(x = rural,                    # Muestra 1
       y = urbano,                    # Muestra 2
       alternative = c("two.sided"),  # Signo hipótesis alternativa
       paired = FALSE,                # Variables dependientes
       var.equal = FALSE,             # Varianzas iguales
       conf.level = 0.95)             # Nivel de confianza (1-nivel significación)
```

```
##
##  Welch Two Sample t-test
##
## data:  rural and urbano
## t = -0.19098, df = 3230.3, p-value = 0.8486
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4836161  0.3977663
## sample estimates:
## mean of x mean of y
##  29.71336  29.75629
```

Como el p-valor es mayor que el nivel de significación, podemos concluir que no hay evidencias suficientes para rechazar H_0 y por tanto no existen diferencias entre el IMC entre las personas que tienen residencia rural y residencia urbana.

Ahora vamos a comprobar qué pasa con las personas con hipertensión. Para ello, vamos a plantear la siguiente hipótesis, pues gráficamente es la que hemos visto que se cumple:

$$\begin{cases} H_0: & \mu_{hipertensión} \leq \mu_{nohipertensión} \\ H_1: & \mu_{hipertensión} > \mu_{nohipertensión} \end{cases}$$

```
# Contraste para la diferencia de medias

yes <- df[df$hypertension == 1, 'bmi']
no <- df[df$hypertension == 0, 'bmi']

t.test(x = yes,                       # Muestra 1
       y = no,                        # Muestra 2
       alternative = c("greater"),    # Signo hipótesis alternativa
       paired = FALSE,                # Variables dependientes
       var.equal = FALSE,             # Varianzas iguales
       conf.level = 0.99)             # Nivel de confianza (1-nivel significación)
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  yes and no
## t = 6.5064, df = 463.38, p-value = 1e-10
## alternative hypothesis: true difference in means is greater than 0
## 99 percent confidence interval:
##  1.42336     Inf
## sample estimates:
## mean of x mean of y
##  31.70864  29.48883
```

El p-valor es pequeño, lo que quiere decir que nuestro estadístico de diferencia de medias está en la región crítica, por lo que tenemos evidencias suficientes para rechazar la hipótesis nula y podemos concluir que hay diferencia entre el IMC de personas con hipertensión y las que no tienen hipertensión, teniendo las personas con hipertensión un mayor IMC. #ANOVA Por último, vamos a comprobar si el tabaquismo puede influir en el IMC. Gráficamente hemos visto que las personas que nunca han fumado suelen tener un IMC menor.

Definimos el siguiente contraste:

$$\begin{cases} H_0: & \mu_{nofumador} = \mu_{exfumador} = \mu_{fumador} \\ H_1: & Diferentes \end{cases}$$

```r
# Aplicamos ANOVA
anova_str <- aov(bmi ~ smoking_status, data = df)
summary(anova_str)
```

```
##                   Df Sum Sq Mean Sq F value  Pr(>F)
## smoking_status    2    494  247.14   6.068 0.00234 **
## Residuals      3230 131557   40.73
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como resultado, obtenemos la tabla de ANOVA vista en clase, con la diferencia de encontrar el p-valor en lugar del valor teórico para la distribución de la F.

El criterio de rechazo es el mismo que tenemos en los contrastes de hipótesis, es decir, si el p-valor es mayor que el nivel de significación, rechazamos H_0.

En este caso tenemos un p-valor muy pequeño, bajo un nivel de significación 0.05, rechazamos H_0, es decir, no hay evidencias suficientes para aceptar que las medias de IMC en los diferentes grupos según si la persona nunca ha fumado, es exfumador o fuma actualmente sean iguales. Podemos concluir que estas medias son diferentes y por tanto, el tabaquismo sí es un factor influyente en el IMC.