

# Práctica 2 - Tipología y ciclo de vida de los datos

Daniel Padilla Ortega y Fernando Álamo

Junio-2021

## Contents

0.Carga del dataset . . . . .	2
1.Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder? . . . . .	2
2. Integración y selección de los datos de interés a analizar . . . . .	3
Selección de los atributos necesarios para el estudio. . . . .	5
Tratamiento de valores ausentes en las variables categóricas . . . . .	7
Tratamiento de valores ausentes en las variables numéricas. . . . .	10
Aproximación de valores ausentes con modelos de regresión lineal* . . . . .	11
Imputación valores ausentes de IMC en aquellos que tenemos altura y peso . . . . .	15
Imputación missing values N_CIGARRILLOS . . . . .	15
Imputación missing values GANANCIA_PESO_MADRE . . . . .	16
Imputación missing values PESO_NACER_GM . . . . .	16
Identificación y tratamiento de Outliers . . . . .	17
Identificación para las variables de la madre. . . . .	17
Discretización de la variable N_CIGARRILLOS . . . . .	19
Tratamiento Outliers EDAD_MADRE . . . . .	19
Tratamiento Outliers IMC . . . . .	19
Tratamiento Outliers PESO_MADRE . . . . .	21
Tratamiento Outliers ALTURA_MADRE . . . . .	21
Tratamiento Outliers GANANCIA_PESO_MADRE . . . . .	22
Identificación para la variable PESO_NACER_GM del niño . . . . .	23
4. Análisis de los datos . . . . .	26
4.1 Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar). . . . .	26
4.2 Comprobación de la normalidad y homogeneidad de la varianza . . . . .	27
4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes. . . . .	31

¿Las madres de raza caucasiana tienen hijos en media con mayor edad que las de raza afroamericana ? . . . . .	31
¿Las madres de raza caucasiana tienen su primer hijo en media con mayor edad que las de raza afroamericana? . . . . .	32
¿Las madres de raza asiática tienen su primer hijo en media con mayor edad que las de raza no asiática? . . . . .	33
¿Las madres con estudios universitarios tienen en media su primer hijo con mayor edad que las que no tienen estudios superiores? . . . . .	35
¿Qué factores sociológicos ,fisiológicos y culturales favorecen que una mujer tenga su primer embarazo? . . . . .	35
Análisis de predicción del peso del recién nacido en función de las características fisiobiológicas de la madre . . . . .	42
Análisis del peso de los bebés en función de la raza de las madres. . . . .	45
5. Representación de los resultados a partir de tablas y gráficas. . . . .	47
6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema? . . . . .	51

## 0.Carga del dataset

```

filename <-("US_births(2018).csv")
data <- read.csv(filename)
dim(data)

## [1] 3801534      55

names(data)

##  [1] "ATTEND"      "BFACIL"       "BMI"          "CIG_O"        "DBWT"
##  [6] "DLMP_MM"     "DLMP_YY"      "DMAR"         "DOB_MM"       "DOB_TT"
## [11] "DOB_WK"       "DOB_YY"       "DWgt_R"       "FAGECOMB"    "FEDUC"
## [16] "FHISPX"       "FRACE15"      "FRACE31"      "FRACE6"       "ILLB_R"
## [21] "ILOP_R"       "ILP_R"        "IMP_SEX"      "IP_GON"       "LD_INDL"
## [26] "MAGER"        "MAGE_IMPFLG" "MAR_IMP"      "MBSTATE_REC" "MEDUC"
## [31] "MHISPX"       "MM_AICU"      "MRACE15"      "MRACE31"     "MRACEIMP"
## [36] "MRAVE6"        "MTRAN"        "M_Ht_In"     "NO_INFEC"    "NO_MMORB"
## [41] "NO_RISKS"      "PAY"          "PAY_REC"      "PRECARE"     "PREVIS"
## [46] "PRIORDEAD"    "PRIORLIVE"   "PRIORTERM"   "PWgt_R"      "RDMETH_REC"
## [51] "RESTATUS"      "RF_CESAR"     "RF_CESARN"   "SEX"          "WTGAIN"

```

## 1.Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

El conjunto de datos seleccionado contiene información de todos los nacimientos registrados en Estados Unidos en el año 2018 (un total de más de 3.8 millones de registros). Se ha obtenido de la página de internet kaggle, en concreto de link siguiente: <https://www.kaggle.com/des137/us-births-2018>

Se trata de una subselección de 55 atributos de un dataset con más de 150 atributos, que contiene los datos publicados por el CDC (Centers for Disease Control and Prevention) perteneciente a la administración pública estadounidense.

El año 2018 puede ser un año representativo de la década 2010-2019, por lo que este conjunto de datos consideramos que una muestra insesgada de los nacimientos de niños durante toda la década. El dataset original permite hacer múltiples estudios en profundidad con respecto al peso de los bebés al nacer, pudiendo determinar si las características fisiológicas de la madre (edad, peso, altura, raza, etc) tienen influencia en el peso de los bebés.

Adicionalmente, se pueden estudiar otros fenómenos sociológicos como la influencia del nivel de estudios, estado civil, raza, etc., en la edad en la que las madres americanas tienen su primer hijo, y si los neonatos tienen características homogéneas por razas.

Por último, los resultados se podrían usar para generar un modelo predictivo que usar para, una vez ajustado por semana de embarazo, tener indicios si el feto se está desarrollando de forma adecuada.

En esta práctica vamos a analizar los datos desde un punto de vista sociológico con respecto a la edad de las madres, y desde un punto de vista biológico con respecto al peso de los bebés al nacer, tratando de responder a las siguientes preguntas: \* ¿Es la raza un factor que influye en la edad de las madres? ¿Hay alguna preponderancia sobre qué raza tiene a sus hijos más mayores? \* ¿Influye el nivel de estudios en la edad en que las madres norteamericanas tienen sus hijos? ¿Y en su primer hijo? \* ¿Es el peso de la madre, la altura, o el IMC factores explicativos del peso del bebé? \* ¿Se puede aproximar con un 95% de confianza el peso del bebe a través de los factores sociológicos y fisiológicos de la madre?

---

## 2. Integración y selección de los datos de interés a analizar

**Descripción de los campos del dataset** (en negrita los campos que van a objeto de limpieza y análisis en la práctica)

- 1-ATTEND: Tipo de asistente en el parto (1)Doctor en Medicina, (2)Doctor en Osteopatía, (3)Matrona, (4)Otro tipo de enfermera, (5)Other
- 2-BFACIL: Instalación para el nacimiento (1)Hospital, (2)Centro de maternidad independiente, (3)Hogar(Previsto), (4)Hogar(No Previsto), (5)Hogar(Desconocido), (6)Clínica, (7)Otro
- 3-BMI: Índice de Masa Corporal de la madre.
- 4-CIG\_0: Cigarros diarios antes del embarazo
- 5-DBWT: Peso del bebe al nacer en gramos
- 6-DLMP\_MM: Último mes de menstruación
- 7-DLMP YY: Año de la última menstruación
- 8-DMAR: Estado Civil de la madre (1)Casada, (2)Soltera
- 9-DOB\_MM: Mes de nacimiento
- 10-DOB\_TT: Hora de nacimiento
- 11-DOB\_WK: Día de la semana de nacimiento
- 12-DOB YY: Año de nacimiento
- 13-DWgt\_R: Recodificación del peso de en el parto

- 14-FAGECOMB: Años combinados de los padres
- 15-FEDUC: Grado de estudios del padre. (1)8th grado o menos, (2)De 9th a 12th grado sin diploma, (3)Graduado de la escuela secundaria o GED completado, (4)Algunos créditos universitarios, pero sin título, (5)Grado asociado (AA, AS), (6)Licenciatura (BA, AB, BS), (7)Maestría (MA, MS, MEng, MEd, MSW, MBA), (8)Doctorado (PhD, EdD) o Título Profesional (MD, DDS,DVM, LLB, JD), (9)Desconocido
- 16-FHISPX: Origen hispánico del padre. (0)No hispánico, (1)Mexicano, (2)Puerto Riqueño, (3)Cubano, (4)Centro americano o sudamericano, (5)Dominicano, (6)Otro, (9)Desconocido
- 17-FRACE15: Clasificación de la raza del padre en 15 diferentes tipologías.
- 18-FRACE31: Clasificación de la raza del padre en 31 diferentes tipologías.
- 19-FRACE6: Clasificación de la raza del padre: (1)Blanca, (2)Negra, (3)Indoamericana o nativa de Alaska, (4)Asiática, (5) Hawaiana o de otra isla del pacifico, (6)Más de una raza, (9)Desconocido.
- 20-ILLB\_R: Recodificación:Intervalo desde el último de nacimiento de un hermano vivo (000-003) Parto múltiple, (004-300) Meses desde el último nacimiento.
- 21-ILOP\_R: Registro del intervalo de nacimiento (000-003) Parto múltiple, (004-300) Meses desde el último nacimiento, 888 No aplicable/Primer nacimiento, (999)Desconocido.
- 22-ILP\_R: Intervalo desde el último registro de nacimiento, (000-003) Plural delivery, (004-300) Meses desde el último nacimiento, 888 No aplicable/Primer nacimiento, (999)Desconocido.
- 23-IMP\_SEX: Sexo imputado (Registro en blanco)Sin imputar, (1)Imputado.
- 24-IP\_GON: Presencia de infección del tipo Gonorrhea, (Y)Si, (N)No, (U)Desconocido
- 25-LD\_INDL: Parto inducido (Y)Si, (N)No, (U)Desconocido
- 26-MAGER: Edad de la madre
- 27-MAGE\_IMPFLG: Campo de edad de la madre imputado (Registro en blanco)Sin imputar, (1)Imputado.
- 28-MAR\_IMP: Campo de estado civil de la madre imputado (Registro en blanco)Sin imputar, (1)Imputado.
- 29-MBSTATE\_REC: Procedencia de la madre (1)Nacida en EEUU, (2)Nacida fuera de EEUU, (3)Desconocido
- 30-MEDUC: Grado de estudios de la madre, (1)8th grado o menos, (2)De 9th a 12th grado sin diploma, (3)Graduado de la escuela secundaria o GED completado, (4)Algunos créditos universitarios, pero sin título, (5)Grado asociado (AA, AS), (6)Licenciatura (BA, AB, BS), (7)Maestría (MA, MS, MEng, MEd, MSW, MBA), (8)Doctorado (PhD, EdD) o Título Profesional (MD, DDS,DVM, LLB, JD), (9)Desconocido
- 31-MHISPX: Origen hispánico de la madre. (0)No hispánico, (1)Mexicano, (2)Puerto Riqueño, (3)Cubano, (4)Centro americano o sudamericano, (5)Dominicano, (6)Otro, (9)Desconocido
- 32-MM\_AICU: Ingreso en cuidados intensivos, (Y)Si, (N)No, (U)Desconocido
- 33-MRACE15: \*\*\*\*
- 34-MRACE31: \*\*\*\*
- 35-MRACEIMP: Campo de clasificación de la raza de la madre imputada (Registro en blanco)Sin imputar, (1)Raza desconocida imputada, (2)Raza Imputada

- 36-MRAVE6: Clasificación de la raza de la madre. (1)Blanca, (2)Negra, (3)Indoamericana o nativa alaseña, (4)Asiatica, (5) Hawaiana o de otra isla del pacifico, (6)Más de una raza, (9)Desconocido.
- 37-MTRAN: Madre transferida, (Y)Si, (N)No, (U)Desconocido
- 38-M\_Ht\_In: Altura de la madre en pulgadas
- 39-NO\_INFEC: Sin infecciones reportadas, (Y)Si, (N)No, (U)Desconocido
- 40-NO\_MMORB: Sin morbilidad materna reportada, (Y)Si, (N)No, (U)Desconocido
- 41-NO\_RISKS: Sin factores de riesgo reportados, (Y)Si, (N)No, (U)Desconocido
- 42-PAY: Procedencia del pago del parto, (1)Seguro público (Medicaid), (2)Seguro privado
- 43-PAY\_REC: Registro del pago, (1)Medicaid, (2)Seguro privado, (3)Pago sin seguro, (4)Otro, (9)Desconocido
- 44-PRECARE: Mes en que comenzó la atención prenatal (00)Sin atención prenatal, (01-10)Mes en el que comenzó la atención prenatal, (99)Desconocido
- 45-PREVIS: Número de visitas prenatales
- 46-PRIORDEAD: Número de niños muertos de nacimientos de niños previos.
- 47-PRIORLIVE: Número de niños vivos de nacimientos de niños previos.
- 48-PRIORTERM: Otras terminaciones anteriores.
- 49-PWgt\_R: Recodificación de peso antes del embarazo, peso en libras.
- 50-RDMETH\_REC: Registro del método de parto, (1)Vaginal, (2)Vaginal después de una cesárea anterior, (3)Cesárea primaria, (4)Ceárea repetida, (5)Vaginal (desconocido si hubo cesárea anterior), (6)Sección C (desconocido si hubo sección c anterior), (9)Sin establecer.
- 51-RESTATUS: Estatus residencial, (1)Residente, (2) EL estado de ocurrencia y la residencia es la misma pero el condado es diferente, (3) El estado de ocurrencia y la residencia es diferente pero ambos son uno de los 50 estados de EE. UU o del distrito de Columbia (4)Residente extranjero
- 52-RF\_CESAR: Cesárea previa, (Y)Si, (N)No, (U)Desconocido
- 53-RF\_CESARN: Numero de cesáreas previas
- 54-SEX: Sexo del recien nacido, (M)Masculino, (F)Femenino
- 55-WTGAIN: Incremento de peso de la madre en libras

#### Selección de los atributos necesarios para el estudio.

```
data2 <- data[,c(3,4,5,8,26,30,36,38,41,46,47,48,49,54,55)]
dim(data2)
```

```
## [1] 3801534      15
```

```
names(data2)
```

```
## [1] "BMI"        "CIG_0"       "DBWT"        "DMAR"        "MAGER"       "MEDUC"
## [7] "MRAVE6"     "M_Ht_In"    "NO_RISKS"    "PRIORDEAD"   "PRIORLIVE"   "PRIORTERM"
## [13] "PWgt_R"      "SEX"         "WTGAIN"
```

Tipos de datos:

```
sapply(data2, function(x) class(x))

##      BMI      CIG_0      DBWT      DMAR      MAGER      MEDUC
##  "numeric"  "integer"  "integer"  "integer"  "integer"  "integer"
##  MRAVE6    M_Ht_In   NO_RISKS PRIORDEAD PRIORLIVE PRIORTERM
##  "integer"  "integer"  "integer"  "integer"  "integer"  "integer"
##  PWgt_R     SEX      WTGAIN
##  "integer"  "character" "integer"
```

En el dataset no tenemos datos suficientes para el tratamiento de enfermedades durante el embarazo, como preclampsia, diabetes gestacional, etc. En todo caso, estamos interesados en analizar los nacimientos en los que se han identificado factores de riesgo en el embarazo. Dado que puede haber muchos otros riesgos no explicitados en el dataset, ni tampoco hay información relevante sobre los mismos, filtramos el dataset borrando las madres con factores de riesgo identificados.

```
data2 <- data2[data2$NO_RISKS == 1,-9]
dim(data2)
```

```
## [1] 2608956      14
```

```
names(data2)
```

```
## [1] "BMI"      "CIG_0"    "DBWT"    "DMAR"    "MAGER"    "MEDUC"
## [7] "MRAVE6"   "M_Ht_In"  "PRIORDEAD" "PRIORLIVE" "PRIORTERM" "PWgt_R"
## [13] "SEX"       "WTGAIN"
```

Uno de los análisis que queremos realizar es sobre la influencia del nivel de estudios de la madre en la edad del primer nacimiento. Para ello, vamos a generar una variable dicotómica obteniendo si es el primer nacimiento mediante los atributos PRIOTERM, PRIORLIVE y PRIORDEAD.

El primer paso es tratar los registros con valores ausentes en alguna de las variables PRIOTERM, PRIORLIVE y PRIORDEAD. En este caso, 99 es el valor centinela de dato no válido.

```
nrow(data2[data2$PRIOTERM == 99 | data2$PRIORLIVE == 99 | data2$PRIORDEAD == 99 ,])/nrow(data2)*100

## [1] 0.4282556
```

El porcentaje de valores ausentes es del 0.43% de los registros por lo que podemos eliminarlos sin perder información relevante en nuestros datos, ni producir sesgos aparentes.

```
data2 <- data2[!(data2$PRIOTERM == 99 | data2$PRIORLIVE == 99 | data2$PRIORDEAD == 99) ,]
```

Una vez borrados, generamos el atributo PRIMER\_EMBARAZO, como variable dicotómica. 0 si no es el primero, 1 si lo es.

```
data2$PRIMER_EMBARAZO = 0
data2$PRIMER_EMBARAZO[data2$PRIOTERM == 0 & data2$PRIORLIVE == 0 & data2$PRIORDEAD == 0 ] = 1
data2 <- data2[,-(9:11)]
names(data2)
```

```

## [1] "BMI"           "CIG_O"          "DBWT"           "DMAR"
## [5] "MAGER"         "MEDUC"          "MRAVE6"         "M_Ht_In"
## [9] "PWgt_R"        "SEX"            "WTGAIN"         "PRIMER_EMBARAZO"

```

Para facilidad de uso vamos a cambiar el nombre de las variables y reordenar los campos por afinidad.

```

names(data2) <- c("IMC", "N_CIGARRILLOS", "PESO_NACER_GM", "ESTADO_CIVIL", "EDAD_MADRE", "EDUC_MADRE",
                  "RAZA_MADRE", "ALTURA_MADRE", "PESO_MADRE_LB", "SEXO_HIJO", "GANANCIA_PESO_MADRE",
                  "PRIMER_EMBARAZO")

```

```

data2 <- data2[, c(5, 9, 8, 1, 4, 6, 7, 2, 12, 11, 10, 3)]
names(data2)

```

```

## [1] "EDAD_MADRE"      "PESO_MADRE_LB"    "ALTURA_MADRE"
## [4] "IMC"              "ESTADO_CIVIL"     "EDUC_MADRE"
## [7] "RAZA_MADRE"       "N_CIGARRILLOS"   "PRIMER_EMBARAZO"
## [10] "GANANCIA_PESO_MADRE" "SEXO_HIJO"       "PESO_NACER_GM"

```

Después de estos tratamientos básicos contamos con 5 variables categóricas y 8 variables numéricas.

### Tratamiento de valores ausentes en las variables categóricas

```

numvars <- c(1:4, 8, 10, 12)
catvars <- c(5:7, 9, 11)

```

```
names(data2[,catvars])
```

```

## [1] "ESTADO_CIVIL"    "EDUC_MADRE"      "RAZA_MADRE"      "PRIMER_EMBARAZO"
## [5] "SEXO_HIJO"

```

```
summary(data2[,catvars])
```

```

## ESTADO_CIVIL EDUC_MADRE RAZA_MADRE PRIMER_EMBARAZO
## Min. :1.0 Min. :1.000 Min. :1.000 Min. :0.0000
## 1st Qu.:1.0 1st Qu.:3.000 1st Qu.:1.000 1st Qu.:0.0000
## Median :1.0 Median :4.000 Median :1.000 Median :0.0000
## Mean :1.4 Mean :4.409 Mean :1.513 Mean :0.3697
## 3rd Qu.:2.0 3rd Qu.:6.000 3rd Qu.:2.000 3rd Qu.:1.0000
## Max. :2.0 Max. :9.000 Max. :6.000 Max. :1.0000
## NA's :327987
## SEXO_HIJO
## Length:2597783
## Class :character
## Mode :character
##
## 
## 
## 
```

La variable de ESTADO\_CIVIL tiene 327.987 valores ausentes, y no tenemos una buena forma de inferir los datos correctos. La opción, dado el gran volumen de registros que tenemos, sería descartar los valores ausentes.

Como precaución previa, vamos a ver como están distribuidas en el resto de variables categóricas.

Para obtener las proporciones correctamente, lo primero que tenemos que poner es un valor a los NA, y a continuación observar las proporciones.

```
data2$ESTADO_CIVIL[is.na(data2$ESTADO_CIVIL)] = 0
proportions(table(data2$ESTADO_CIVIL, data2$SEXO_HIJO), margin = 2)

##
##          F            M
## 0 0.1266121 0.1259153
## 1 0.5130440 0.5141785
## 2 0.3603440 0.3599062

proportions(table(data2$ESTADO_CIVIL, data2$RAZA_MADRE), margin = 2)

##
##          1            2            3            4            5            6
## 0 0.12607588 0.04338809 0.06531935 0.30688383 0.16057599 0.17661312
## 1 0.56378878 0.27580491 0.29167733 0.60099141 0.36229429 0.34822332
## 2 0.31013534 0.68080699 0.64300333 0.09212475 0.47712972 0.47516356

proportions(table(data2$ESTADO_CIVIL, data2$EDUC_MADRE), margin = 2)

##
##          1            2            3            4            5            6
## 0 0.13496108 0.10834899 0.11889235 0.12572229 0.10311150 0.12559291
## 1 0.46000502 0.23543503 0.33043689 0.43454284 0.60377532 0.75842508
## 2 0.40503389 0.65621597 0.55067075 0.43973487 0.29311319 0.11598201
##
##          7            8            9
## 0 0.12640230 0.14314415 0.52185565
## 1 0.81140801 0.81444277 0.23663863
## 2 0.06218969 0.04241308 0.24150571
```

Desde el punto de vista del sexo del bebe, el borrado no debería generar ningún riesgo de sesgo, ya que vemos que las proporciones se mantienen prácticamente iguales para los dos sexos.

Sin embargo, tanto respecto a la raza de la madre, como a su nivel de estudios, nos encontramos que las proporciones de los datos a borrar no son despreciables en el total de la muestra de cada categoría, con proporciones mayores a 10% en la mayoría de los casos.

Llegado a este punto, y después de haber probado a imputar estos datos ausentes con métodos de imputación basados en la similitud tipo kNN, decidimos borrar estos registros con los datos ausentes. Nos parece que es más coherente que tratar el dato como una categoría adicional (“Soltera”, “Casada”, “Desconocido”). Esta decisión conlleva la asunción de que no hay razones para que las personas que no aporten su estado civil estén sesgadas a estar “Casadas” o “Solteras”. Es probable que en el pasado esta asunción fuera más débil por cierto rechazo social que entendemos no está presente en nuestros días.

Borramos por tanto los registros con estado civil desconocido.

```
data2 <- data2[data2$ESTADO_CIVIL != 0,]  
nrow(data2)
```

```
## [1] 2269796
```

La educación va a ser un eje de análisis. Para ellos vamos a categorizar la educación en tres grupos, hasta primaria, hasta secundaria, universitarios y postuniversitarios.

```
data2$EDUC_MADRE <- cut(data2$EDUC_MADRE, c(-Inf, 2, 4, 6, 8, +Inf),  
                           labels = c("Primaria", "Secundaria", "Universitarios",  
                                     "Postuniversitarios", "Desconocido"),  
                           ordered = FALSE)
```

El número de registros sigue siendo considerablemente alto, más de 2.2 millones de registros.

Transformación del resto de las variables categóricas a factor.

```
data2$ESTADO_CIVIL <- factor(data2$ESTADO_CIVIL,  
                               levels = c(1:2),  
                               labels = c("Casada", "Soltera"))
```

```
data2$RAZA_MADRE <- factor(data2$RAZA_MADRE,  
                            levels = c(1:7),  
                            labels = c("Caucasiana",  
                                      "Afroamericana",  
                                      "Indoamericana",  
                                      "Asiatica",  
                                      "Hawaiiana",  
                                      "Más de una raza",  
                                      "Desconocido"))
```

```
data2$PRIMER_EMBARAZO <- factor(data2$PRIMER_EMBARAZO,  
                                    levels = c(0, 1),  
                                    labels = c("Si", "No"))
```

```
data2$SEXO_HIJO <- factor(data2$SEXO_HIJO)
```

```
summary(data2[, catvars])
```

```
##   ESTADO_CIVIL          EDUC_MADRE          RAZA_MADRE  
##   Casada :1334281    Primaria      : 290001  Caucasanina :1688854  
##   Soltera: 935515    Secundaria     :1032808  Afroamericana : 376863  
##               Universitarios : 657148  Indoamericana : 21922  
##               Postuniversitarios: 274317  Asiatica     : 117452  
##               Desconocido     : 15522   Hawaiian     :  6937  
##                                         Más de una raza: 57768  
##                                         Desconocido     :       0  
##  
##   PRIMER_EMBARAZO SEXO_HIJO  
##   Si:1442681      F:1110987  
##   No: 827115      M:1158809  
##  
##
```

```
##  
##  
##
```

Tratamiento de valores ausentes en las variables numéricas.

```
names(data2[,numvars])
```

```
## [1] "EDAD_MADRE"          "PESO_MADRE_LB"        "ALTURA_MADRE"  
## [4] "IMC"                 "N_CIGARRILLOS"       "GANANCIA_PESO_MADRE"  
## [7] "PESO_NACER_GM"
```

```
summary(data2[,numvars])
```

```
##   EDAD_MADRE    PESO_MADRE_LB    ALTURA_MADRE      IMC  
##  Min.   :12.00  Min.   :75.0   Min.   :30.00  Min.   :13.00  
##  1st Qu.:24.00  1st Qu.:128.0  1st Qu.:62.00  1st Qu.:21.90  
##  Median :28.00  Median :147.0  Median :64.00  Median :25.00  
##  Mean   :28.13  Mean   :171.2  Mean   :64.41  Mean   :27.92  
##  3rd Qu.:32.00  3rd Qu.:175.0  3rd Qu.:66.00  3rd Qu.:29.90  
##  Max.   :50.00  Max.   :999.0  Max.   :99.00  Max.   :99.90  
##   N_CIGARRILLOS  GANANCIA_PESO_MADRE  PESO_NACER_GM  
##  Min.   : 0.000  Min.   : 0.00   Min.   : 227  
##  1st Qu.: 0.000  1st Qu.:21.00   1st Qu.:3000  
##  Median : 0.000  Median :30.00   Median :3317  
##  Mean   : 1.594  Mean   :32.17   Mean   :3295  
##  3rd Qu.: 0.000  3rd Qu.:40.00   3rd Qu.:3635  
##  Max.   :99.000  Max.   :99.00   Max.   :9999
```

Observamos la presencia de valores centinela que nos muestran la presencia de valores ausentes en todas las variables excepto EDAD\_MADRE, vamos a proceder con su imputación.

Cambiamos los valores centinela a NaN para que no afecten a la calidad del modelo de regresión que vamos a intentar usar para llenar valores ausentes.

```
data2$PESO_MADRE_LB[data2$PESO_MADRE_LB == 999.0] = NaN  
data2$ALTURA_MADRE[data2$ALTURA_MADRE == 99.0] = NaN  
data2$IMC[data2$IMC == 99.90] = NaN  
data2$N_CIGARRILLOS[data2$N_CIGARRILLOS == 99.0] = NaN  
data2$GANANCIA_PESO_MADRE[data2$GANANCIA_PESO_MADRE == 99.0] = NaN  
data2$PESO_NACER_GM[data2$PESO_NACER_GM == 9999] = NaN  
summary(data2[,numvars])
```

```
##   EDAD_MADRE    PESO_MADRE_LB    ALTURA_MADRE      IMC  
##  Min.   :12.00  Min.   :75.0   Min.   :30.00  Min.   :13.00  
##  1st Qu.:24.00  1st Qu.:127.0  1st Qu.:62.00  1st Qu.:21.80  
##  Median :28.00  Median :145.0  Median :64.00  Median :24.90  
##  Mean   :28.13  Mean   :154.3  Mean   :64.24  Mean   :26.26  
##  3rd Qu.:32.00  3rd Qu.:174.0  3rd Qu.:66.00  3rd Qu.:29.30  
##  Max.   :50.00  Max.   :375.0  Max.   :78.00  Max.   :69.80  
##                NA's   :45361   NA's   :10964   NA's   :51167
```

```

##  N_CIGARRILLOS      GANANCIA_PESO_MADRE PESO_NACER_GM
##  Min.   : 0.000      Min.   : 0.00          Min.   : 227
##  1st Qu.: 0.000      1st Qu.:20.00        1st Qu.:3000
##  Median : 0.000      Median :30.00        Median :3317
##  Mean   : 1.185      Mean   :30.06        Mean   :3290
##  3rd Qu.: 0.000      3rd Qu.:39.00        3rd Qu.:3634
##  Max.   :98.000      Max.   :98.00        Max.   :8165
##  NA's    :9493       NA's    :69509        NA's    :1630

```

Cambiamos a medidas europeas para mayor facilidad de comprensión, es decir a kg y cm

```

data2$PESO_MADRE <- data2$PESO_MADRE_LB * 0.453592
data2$ALTURA_MADRE <- data2$ALTURA_MADRE * 2.54
data2$GANANCIA_PESO_MADRE <- data2$GANANCIA_PESO_MADRE * 0.453592

data2<- select(data2, -PESO_MADRE_LB)

```

#### Aproximación de valores ausentes con modelos de regresión lineal\*

```

modelo_altura1 <- lm(ALTURA_MADRE ~ PESO_MADRE, data = data2)
summary(modelo_altura1)

```

```

##
## Call:
## lm(formula = ALTURA_MADRE ~ PESO_MADRE, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q      Max
## -92.553  -4.580  -0.199   4.495  38.097
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.542e+02  1.888e-02 8167.1  <2e-16 ***
## PESO_MADRE  1.283e-01  2.617e-04   490.2  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.797 on 2218856 degrees of freedom
## (50938 observations deleted due to missingness)
## Multiple R-squared:  0.09772,    Adjusted R-squared:  0.09772
## F-statistic: 2.403e+05 on 1 and 2218856 DF,  p-value: < 2.2e-16

```

```

modelo_altura2 <- lm(ALTURA_MADRE ~ PESO_MADRE + EDAD_MADRE, data = data2)
summary(modelo_altura2)

```

```

##
## Call:
## lm(formula = ALTURA_MADRE ~ PESO_MADRE + EDAD_MADRE, data = data2)
##
## Residuals:
##
```

```

##      Min     1Q Median     3Q    Max
## -92.703 -4.537 -0.192  4.483 38.738
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.507e+02  2.857e-02 5274.3   <2e-16 ***
## PESO_MADRE 1.261e-01  2.605e-04   484.0   <2e-16 ***
## EDAD_MADRE 1.309e-01  7.994e-04   163.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.757 on 2218855 degrees of freedom
## (50938 observations deleted due to missingness)
## Multiple R-squared:  0.1085, Adjusted R-squared:  0.1085
## F-statistic: 1.35e+05 on 2 and 2218855 DF, p-value: < 2.2e-16

modelo_altura3 <- lm(ALTURA_MADRE ~ PESO_MADRE + EDAD_MADRE + RAZA_MADRE, data = data2)
summary(modelo_altura3)

```

```

##
## Call:
## lm(formula = ALTURA_MADRE ~ PESO_MADRE + EDAD_MADRE + RAZA_MADRE,
##      data = data2)
##
## Residuals:
##      Min     1Q Median     3Q    Max
## -92.581 -4.525 -0.174  4.464 41.075
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.509e+02  2.864e-02 5267.503   <2e-16 ***
## PESO_MADRE 1.214e-01  2.636e-04   460.578   <2e-16 ***
## EDAD_MADRE 1.418e-01  8.065e-04   175.869   <2e-16 ***
## RAZA_MADREAfroamericana -1.152e-01  1.242e-02   -9.274   <2e-16 ***
## RAZA_MADREIndoamericana -9.749e-01  4.644e-02  -20.992   <2e-16 ***
## RAZA_MADREAsiatica -2.818e+00  2.083e-02  -135.314   <2e-16 ***
## RAZA_MADREHawaiiana -3.327e+00  8.338e-02  -39.903   <2e-16 ***
## RAZA_MADREMás de una raza 2.731e-01  2.879e-02    9.484   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.726 on 2218850 degrees of freedom
## (50938 observations deleted due to missingness)
## Multiple R-squared:  0.1166, Adjusted R-squared:  0.1166
## F-statistic: 4.183e+04 on 7 and 2218850 DF, p-value: < 2.2e-16

```

A diferencia de lo que se podría pensar en una primera instancia, la altura de la madre no se puede aproximar con modelos de regresión lineal simple o múltiple con los datos que tenemos. En todos ellos tenemos un  $R^2$  ajustado por debajo del 12%. Tenemos que desestimar los modelos de regresión.

```

modelo_peso1 <- lm(PESO_MADRE ~ ALTURA_MADRE, data = data2)
summary(modelo_peso1)

```

```

## 
## Call:
## lm(formula = PESO_MADRE ~ ALTURA_MADRE, data = data2)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -51.246 -11.823  -3.752   8.255 112.181 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -54.281761  0.253772 -213.9   <2e-16 ***
## ALTURA_MADRE  0.761592  0.001554   490.2   <2e-16 ***  
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 16.56 on 2218856 degrees of freedom
## (50938 observations deleted due to missingness)
## Multiple R-squared:  0.09772,    Adjusted R-squared:  0.09772 
## F-statistic: 2.403e+05 on 1 and 2218856 DF,  p-value: < 2.2e-16 

modelo_peso2 <- lm(PESO_MADRE ~ ALTURA_MADRE + EDAD_MADRE, data = data2)
summary(modelo_peso2)

```

```

## 
## Call:
## lm(formula = PESO_MADRE ~ ALTURA_MADRE + EDAD_MADRE, data = data2)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -50.872 -11.778  -3.739   8.225 112.385 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -54.839443  0.254955 -215.09   <2e-16 *** 
## ALTURA_MADRE  0.757380  0.001565  484.03   <2e-16 ***  
## EDAD_MADRE    0.044272  0.001971   22.47   <2e-16 ***  
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 16.56 on 2218855 degrees of freedom
## (50938 observations deleted due to missingness)
## Multiple R-squared:  0.09792,    Adjusted R-squared:  0.09792 
## F-statistic: 1.204e+05 on 2 and 2218855 DF,  p-value: < 2.2e-16 

modelo_peso3 <- lm(PESO_MADRE ~ ALTURA_MADRE + EDAD_MADRE + RAZA_MADRE, data = data2)
summary(modelo_peso3)

```

```

## 
## Call:
## lm(formula = PESO_MADRE ~ ALTURA_MADRE + EDAD_MADRE + RAZA_MADRE,
##      data = data2)
## 
## Residuals:

```

```

##      Min     1Q Median     3Q    Max
## -52.965 -11.540 -3.539 8.092 108.128
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -51.026933   0.253809 -201.04 <2e-16 ***
## ALTURA_MADRE          0.718758   0.001561  460.58 <2e-16 ***
## EDAD_MADRE            0.116525   0.001974   59.02 <2e-16 ***
## RAZA_MADREAfroamericana 4.544722   0.030068 151.15 <2e-16 ***
## RAZA_MADREIndoamericana 4.422761   0.112973  39.15 <2e-16 ***
## RAZA_MADREAsiatica     -7.655427   0.050627 -151.21 <2e-16 ***
## RAZA_MADREHawaiiana    6.306843   0.202903  31.08 <2e-16 ***
## RAZA_MADREMás de una raza 1.752171   0.070052  25.01 <2e-16 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.37 on 2218850 degrees of freedom
## (50938 observations deleted due to missingness)
## Multiple R-squared:  0.1189, Adjusted R-squared:  0.1189
## F-statistic: 4.278e+04 on 7 and 2218850 DF, p-value: < 2.2e-16

```

Como era de esperar después del resultado anterior, tampoco se puede estimar el peso en función de la altura y las otras características físicas de la madre.

Esta aproximación si la hemos visto funcionar en otros casos, como puede ser conjuntos de deportistas. En esos casos la altura marca muy certeramente el peso, dado que para tener una buena condición física el IMC es relativamente parecido en todos y por tanto se puede ajustar bastante bien altura y peso en caso de datos ausentes.

Lamentablemente, este es el caso en nuestro dataset. Lo único que se puede hacer es borrar los registros con datos ausentes.

```

data2 <- data2[!(is.na(data2$PESO_MADRE) | is.na(data2$ALTURA_MADRE)),]
summary(data2[,numvars])

```

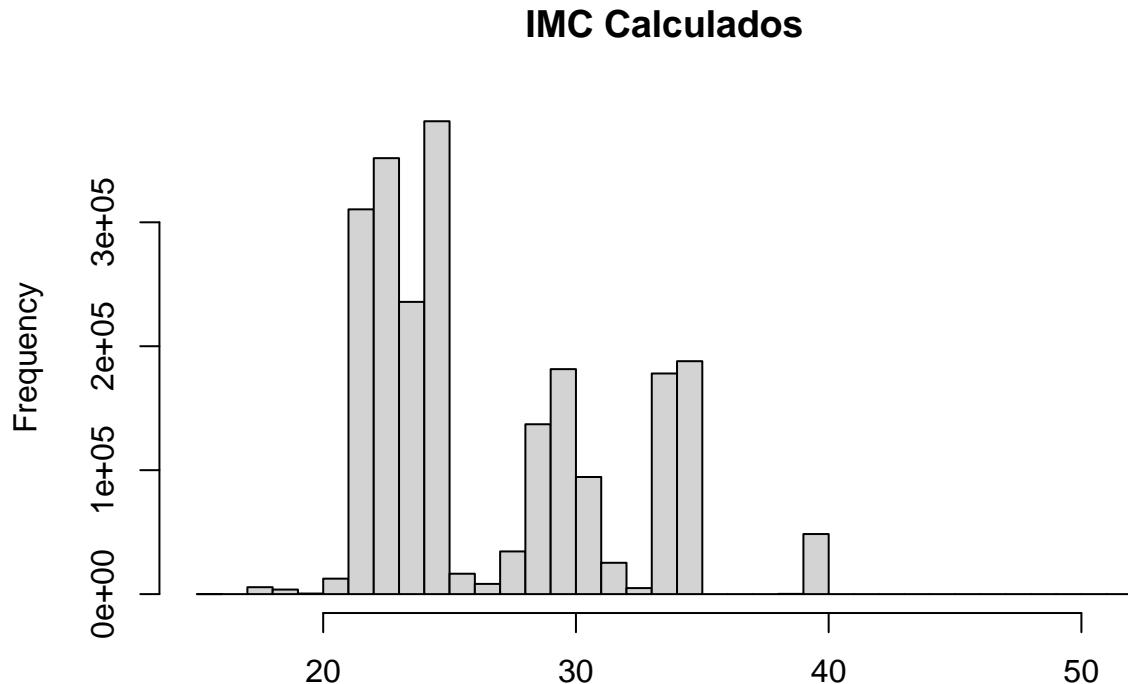
```

##      EDAD_MADRE      ALTURA_MADRE       IMC      ESTADO_CIVIL
##      Min.    :12.00    Min.    : 76.2    Min.    :13.00  Casada :1307579
##  1st Qu.:24.00    1st Qu.:157.5    1st Qu.:21.80  Soltera: 911279
##  Median :28.00    Median :162.6    Median :24.90
##  Mean   :28.12    Mean   :163.2    Mean   :26.26
##  3rd Qu.:32.00    3rd Qu.:167.6    3rd Qu.:29.30
##  Max.   :50.00    Max.   :198.1    Max.   :69.80
##                  NA's    :229
##      PRIMER_EMBARAZO SEXO_HIJO      PESO_MADRE
##      Si:1409253      F:1086273    Min.    : 34.02
##      No: 809605      M:1132585    1st Qu.: 57.61
##                          Median : 65.77
##                          Mean   : 70.00
##                          3rd Qu.: 78.93
##                          Max.   :170.10
##
```

## Imputación valores ausentes de IMC en aquellos que tenemos altura y peso

Después de este borrado la variable IMC tiene 229 datos ausentes, pero este atributo tiene una fórmula directa para su cálculo =>  $\text{peso}(kg)/\text{estatura}^2(m^2)$

```
data2$IMC <- data2$PESO_MADRE/(data2$ALTURA_MADRE/100)^2  
  
hist(data2$IMC[data2$IMC], breaks = 50, main = "IMC Calculados", xlab="")
```



El cálculo de estos IMC revela que los datos de altura/peso o no son correctos, o tenemos mujeres en categoría de riesgo bien por delgadez extrema (IMC<16) u obesidad mórbida (IMC>40). Estos datos los vamos a filtrar posteriormente en el tratamiento de outliers.

## Imputación missing values N\_CIGARRILLOS

Imputamos los valores ausentes del número de cigarrillos por la mediana.

```
sum(is.na(data2$N_CIGARRILLOS))  
  
## [1] 7798  
  
i_cig <- is.na(data2$N_CIGARRILLOS)  
data2$N_CIGARRILLOS[i_cig] <- median(data2$N_CIGARRILLOS[!i_cig])  
  
sum(is.na(data2$N_CIGARRILLOS))
```

```
## [1] 0  
  
median(data2$N_CIGARRILLOS)  
  
## [1] 0
```

#### Imputación missing values GANANCIA\_PESO\_MADRE

```
sum(is.na(data2$GANANCIA_PESO_MADRE))  
  
## [1] 23524  
  
i_gpes <- is.na(data2$GANANCIA_PESO_MADRE)  
data2$GANANCIA_PESO_MADRE[i_gpes] <- median(data2$GANANCIA_PESO_MADRE[!i_gpes])  
  
sum(is.na(data2$GANANCIA_PESO_MADRE))
```

```
## [1] 0
```

```
median(data2$GANANCIA_PESO_MADRE)
```

```
## [1] 13.60776
```

#### Imputación missing values PESO\_NACER\_GM

```
sum(is.na(data2$PESO_NACER_GM))  
  
## [1] 1258  
  
i_gpesn <- is.na(data2$PESO_NACER_GM)  
data2$PESO_NACER_GM[i_gpesn] <- median(data2$PESO_NACER_GM[!i_gpesn])  
  
sum(is.na(data2$PESO_NACER_GM))
```

```
## [1] 0
```

```
median(data2$PESO_NACER_GM)
```

```
## [1] 3320
```

```
summary(data2[,numvars])
```

```

##      EDAD_MADRE      ALTURA_MADRE       IMC      ESTADO_CIVIL
##  Min.   :12.00   Min.   : 76.2   Min.   : 10.46  Casada :1307579
##  1st Qu.:24.00  1st Qu.:157.5  1st Qu.: 21.80  Soltera: 911279
##  Median :28.00  Median :162.6  Median : 24.89
##  Mean   :28.12  Mean   :163.2  Mean   : 26.26
##  3rd Qu.:32.00  3rd Qu.:167.6  3rd Qu.: 29.29
##  Max.   :50.00  Max.   :198.1  Max.   :195.30
## PRIMER_EMBARAZO SEXO_HIJO      PESO_MADRE
##  Si:1409253    F:1086273   Min.   :34.02
##  No: 809605    M:1132585   1st Qu.:57.61
##                                         Median :65.77
##                                         Mean   :70.00
##                                         3rd Qu.:78.93
##                                         Max.   :170.10

```

## Identificación y tratamiento de Outliers

Los valores extremos pueden alterar las propiedades de las herramientas analítico estadísticas dando como resultado conclusiones erróneas en nuestros análisis. Vamos a tratar los valores atípicos de cada una de las variables.

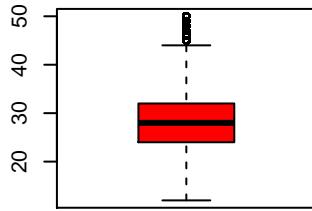
**Identificación para las variables de la madre.** Una forma de tener una idea simple de como son las distribuciones de la variables, así como la identificación de outliers, es dibujando los boxplot de las variables.

```

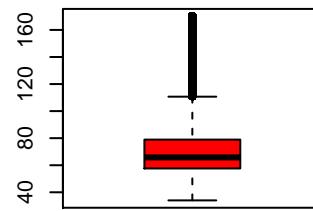
conf2x3 = matrix(c(1:6), nrow=2, byrow=TRUE)
layout(conf2x3)

boxplot(data2$EDAD_MADRE, xlab = "EDAD_MADRE", col="red")
boxplot(data2$PESO_MADRE, xlab = "PESO_MADRE", col="red")
boxplot(data2$ALTURA_MADRE, xlab = "ALTURA_MADRE", col="red")
boxplot(data2$IMC, xlab = "IMC", col="red")
boxplot(data2$N_CIGARRILLOS, xlab = "N_CIGARRILLOS", col="red")
boxplot(data2$GANANCIA_PESO_MADRE, xlab = "GANANCIA_PESO_MADRE", col="red")

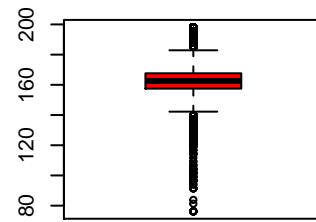
```



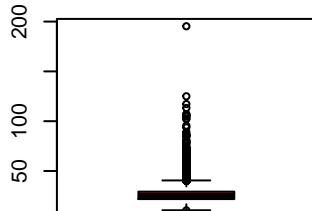
EDAD\_MADRE



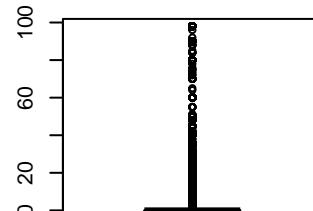
PESO\_MADRE



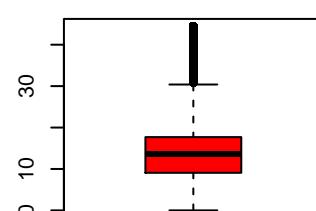
ALTURA\_MADRE



IMC



N\_CIGARRILLOS



GANANCIA\_PESO\_MADRE

A simple vista se aprecia la gran cantidad de outliers que tenemos en la muestra. IMC arroja una estructura bastante desconcertante que puede conllevar que hay datos erróneos. Es casi imposible encontrar personas con IMC de más de 50, y en esta muestra se tienen de casi 200. La variable N\_CIGARRILLOS tiene más el aspecto de una categórica que de una variable numérica. En el peso de la madre, hay valores relativamente altos para lo que estamos acostumbrados en España, pero bien pueden ser casos reales. Por el contrario, en la altura de la madre hay valores sorprendentemente bajo, que podrían ser casos extremos (pero válidos) de embarazos en niñas y personas con enanismo u osteocondrodisplasia.

Porcentaje de outliers en las variables

```
n_data2 <- nrow(data2)

sprintf("%% Outliers edad de la madre %0.2f%%", length(boxplot.stats(data2$EDAD_MADRE)$out)/n_data2*100)

## [1] "% Outliers edad de la madre 0.12%"

sprintf("%% Outliers IMC de la madre %0.2f%%", length(boxplot.stats(data2$IMC)$out)/n_data2*100)

## [1] "% Outliers IMC de la madre 3.25%"

sprintf("%% Outliers peso de la madre %0.2f%%", length(boxplot.stats(data2$PESO_MADRE_LB)$out)/n_data2*100)

## [1] "% Outliers peso de la madre 0.00%"
```

```

sprintf("%% Outliers altura de la madre %0.2f%%", length(boxplot.stats(data2$ALTURA_MADRE)$out)/n_data2)

## [1] "% Outliers altura de la madre 0.46%"

sprintf("%% Outliers cigarrillos fumados %0.2f%%", length(boxplot.stats(data2$N_CIGARRILLOS)$out)/n_data2)

## [1] "% Outliers cigarrillos fumados 8.98%"

sprintf("%% Outliers ganancia de peso de la madre %0.2f%%", length(boxplot.stats(data2$GANANCIA_PESO_MADRE)$out)/n_data2*100)

## [1] "% Outliers ganancia de peso de la madre 1.44%"

```

### Discretización de la variable N\_CIGARRILLOS

La variable N\_CIGARRILLOS presenta un porcentaje importante de outliers y además el rango de la variable es demasiado elevado por lo que se va a proceder a discretizar la variable y así mantener sus propiedades.

```

data2$N_CIGARRILLOS <- cut(data2$N_CIGARRILLOS, c(-Inf, 0, 10, 20, 40, Inf),
                                labels = c("No fumadora", "menos de 10", "entre 10 y 20", "entre 20 y 40", "40 o más"),
                                ordered = TRUE)

summary(data2$N_CIGARRILLOS)

##      No fumadora    menos de 10    entre 10 y 20    entre 20 y 40      40 o más
##          2019650           119115            67810           10223             2060

```

### Tratamiento Outliers EDAD\_MADRE

```

sprintf("%% Outliers edad de la madre %0.2f%%", length(boxplot.stats(data2$EDAD_MADRE)$out)/n_data2*100)

## [1] "% Outliers edad de la madre 0.12%"

unique(boxplot.stats(data2$EDAD_MADRE)$out)

## [1] 47 45 50 46 48 49

```

Los valores extremos para la variable EDAD\_MADRE suponen el 0.12% de la muestra, observamos que estas edades son perfectamente factibles en nuestros días, y no provienen de ningún error en los datos por lo que no será necesario realizar ningún tratamiento.

### Tratamiento Outliers IMC

```

sprintf("%% Outliers IMC de la madre %.2f%%", length(boxplot.stats(data2$IMC)$out)/n_data2*100)

## [1] "% Outliers IMC de la madre 3.25%"

head(sort(unique(boxplot.stats(data2$IMC)$out)),50)

## [1] 10.46026 40.54560 40.54608 40.54723 40.55124 40.56167 40.56469 40.56508
## [9] 40.57155 40.57892 40.58581 40.59663 40.59676 40.60328 40.60388 40.60582
## [17] 40.60995 40.62176 40.62344 40.63525 40.67341 40.67341 40.68051 40.68236
## [25] 40.68686 40.70245 40.72131 40.72528 40.74222 40.74675 40.74881 40.74931
## [33] 40.75459 40.75762 40.76718 40.76968 40.77703 40.77800 40.78678 40.79860
## [41] 40.80311 40.81239 40.81706 40.82249 40.83482 40.85215 40.86475 40.87793
## [49] 40.89030 40.89279

tail(sort(unique(boxplot.stats(data2$IMC)$out)),50)

## [1] 68.83018 69.33619 69.35034 70.08473 70.09906 70.18484 70.30690
## [8] 70.64130 70.66590 70.85484 70.97773 71.47729 71.47868 72.15137
## [15] 72.17202 72.25987 72.49250 72.51489 72.54299 72.98294 73.20585
## [22] 73.23635 73.61103 75.58805 77.47317 78.37422 79.06232 79.33938
## [29] 80.14987 81.34683 84.71884 84.93074 85.69152 85.74757 86.43912
## [36] 86.71184 89.05541 93.74253 95.52289 102.24688 102.53090 104.67916
## [43] 105.46035 105.46035 106.77861 107.10817 113.27223 117.17817 124.99005
## [50] 195.29695

length(boxplot.stats(data2$IMC)$out)

## [1] 72038

sum(boxplot.stats(data2$IMC)$out<15)

## [1] 1

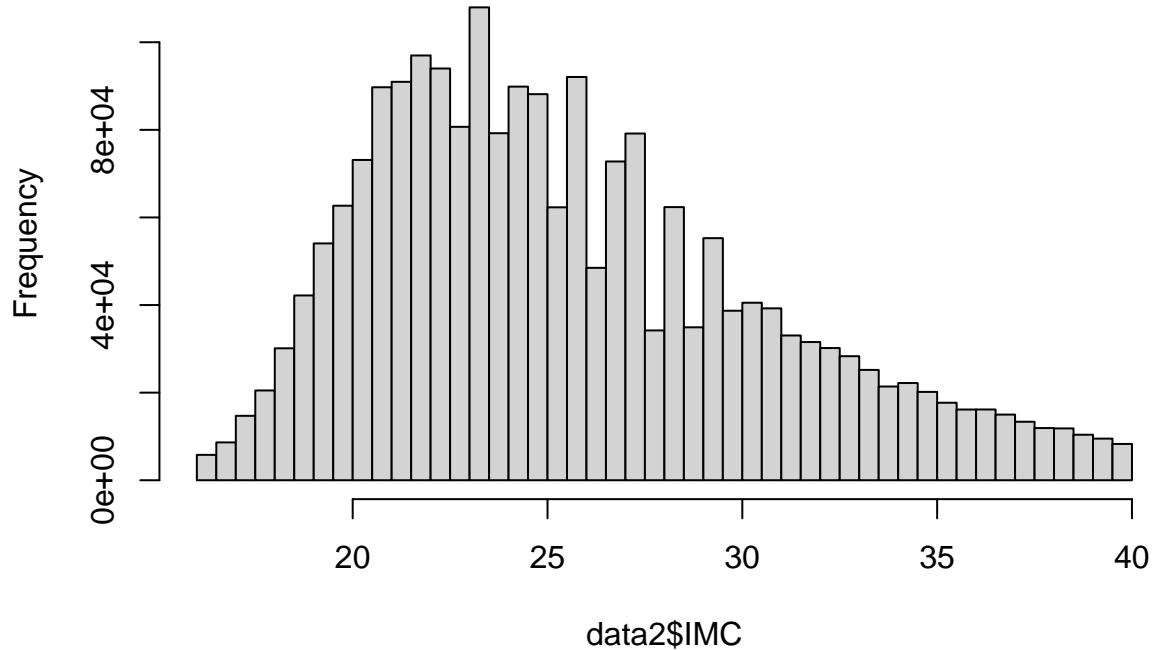
sum(boxplot.stats(data2$IMC)$out>40)

## [1] 72037

i_imc <- which((data2$IMC >= 16) & (data2$IMC <= 40))
data2 <- data2[i_imc,]
n_data2 <- nrow(data2)
hist(data2$IMC, breaks = 50, main = "IMC Calculados")

```

## IMC Calculados



### Tratamiento Outliers PESO\_MADRE

```
sprintf("%% Outliers peso de la madre %0.2f%%", length(boxplot.stats(data2$PESO_MADRE_LB)$out)/n_data2*100)

## [1] "% Outliers peso de la madre 0.00%"

head(sort(unique(boxplot.stats(data2$PESO_MADRE_LB)$out)))

## NULL

tail(sort(unique(boxplot.stats(data2$PESO_MADRE_LB)$out)))

## NULL
```

Después del tratamiento del IMC, hemos excluido también los outliers de peso.

### Tratamiento Outliers ALTURA\_MADRE

```
sprintf("%% Outliers altura de la madre %0.2f%%", length(boxplot.stats(data2$ALTURA_MADRE)$out)/n_data2*100)
```

```

## [1] "% Outliers altura de la madre 0.42%"

length(boxplot.stats(data2$ALTURA_MADRE)$out)

## [1] 8863

head(sort(unique(boxplot.stats(data2$ALTURA_MADRE)$out)))

## [1] 101.60 114.30 116.84 119.38 121.92 124.46

tail(sort(unique(boxplot.stats(data2$ALTURA_MADRE)$out)))

## [1] 185.42 187.96 190.50 193.04 195.58 198.12

```

Como podemos observar los valores extremos provienen de mujeres con una estatura muy baja o muy alta pero desde luego posible por lo que no apreciamos la presencia de errores de imputación en los datos.

### Tratamiento Outliers GANANCIA\_PESO\_MADRE

```

sprintf("%% Outliers ganancia de peso de la madre %0.2f%%", length(boxplot.stats(data2$GANANCIA_PESO_MADRE)$out) / n_data2 * 100)

## [1] "% Outliers ganancia de peso de la madre 1.74%"

head(sort(unique(boxplot.stats(data2$GANANCIA_PESO_MADRE)$out)), 10)

## [1] 29.93707 30.39066 30.84426 31.29785 31.75144 32.20503 32.65862 33.11222
## [9] 33.56581 34.01940

tail(sort(unique(boxplot.stats(data2$GANANCIA_PESO_MADRE)$out)), 10)

## [1] 40.36969 40.82328 41.27687 41.73046 42.18406 42.63765 43.09124 43.54483
## [9] 43.99842 44.45202

```

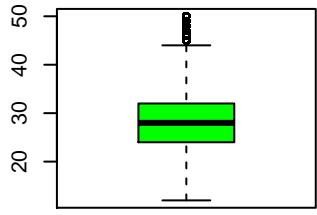
Consideramos que la ganancia de peso durante el embarazo es asimismo factible (embarazos múltiples, etc), por lo que aunque se puedan considerar outliers, van a aportar información al dataset.

```

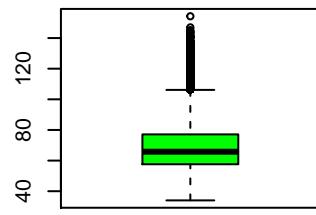
conf2x3 = matrix(c(1:6), nrow=2, byrow=TRUE)
layout(conf2x3)

boxplot(data2$EDAD_MADRE, xlab = "EDAD_MADRE", col="green")
boxplot(data2$PESO_MADRE, xlab = "PESO_MADRE", col="green")
boxplot(data2$ALTURA_MADRE, xlab = "ALTURA_MADRE", col="green")
boxplot(data2$IMC, xlab = "IMC", col="green")
boxplot(data2$N_CIGARRILLOS, xlab = "N_CIGARRILLOS", col="green")
boxplot(data2$GANANCIA_PESO_MADRE, xlab = "GANANCIA_PESO_MADRE", col="green")

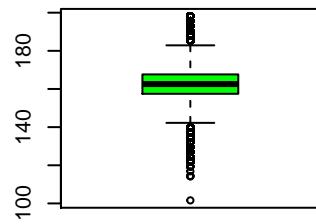
```



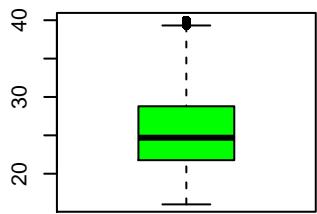
EDAD\_MADRE



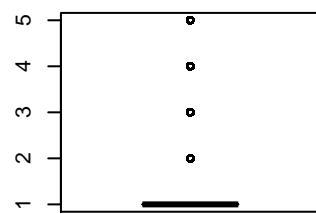
PESO\_MADRE



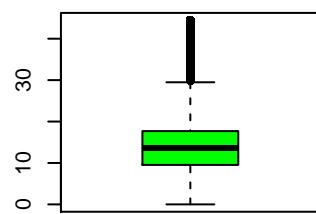
ALTURA\_MADRE



IMC



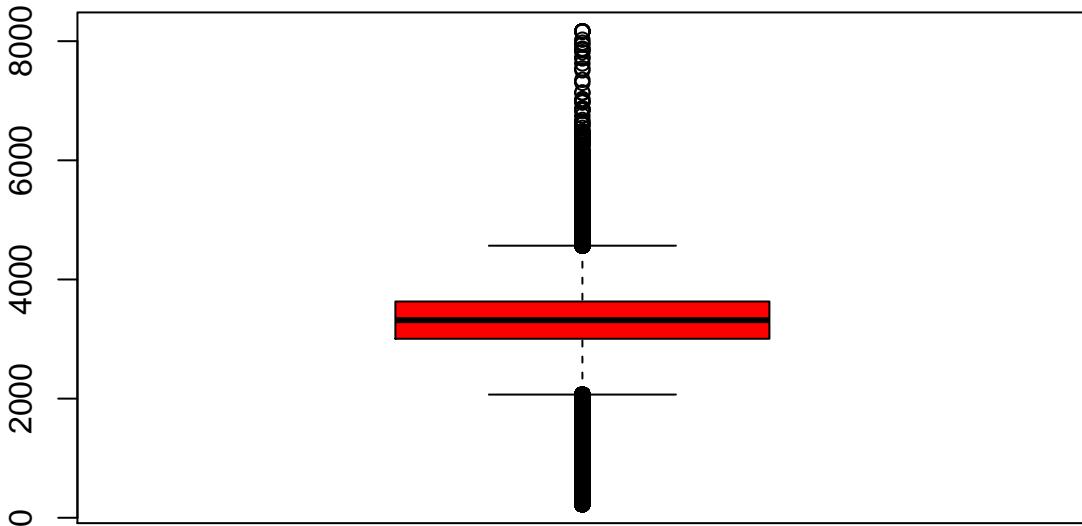
N\_CIGARRILLOS



GANANCIA\_PESO\_MADRE

Identificación para las variable PESO\_NACER\_GM del niño

```
boxplot(data2$PESO_NACER_GM, xlab = "PESO_NACER_GM", col="red")
```



PESO\_NACER\_GM

```
sprintf("%% Outliers del peso del bebe al nacer %0.2f%%",
       length(boxplot.stats(data2$PESO_NACER_GM)$out)/n_data2*100)
```

```
## [1] "% Outliers del peso del bebe al nacer 3.01%"
```

```
head(sort(unique(boxplot.stats(data2$PESO_NACER_GM)$out)),10)
```

```
## [1] 227 228 229 230 231 232 233 234 235 236
```

```
tail(sort(unique(boxplot.stats(data2$PESO_NACER_GM)$out)),10)
```

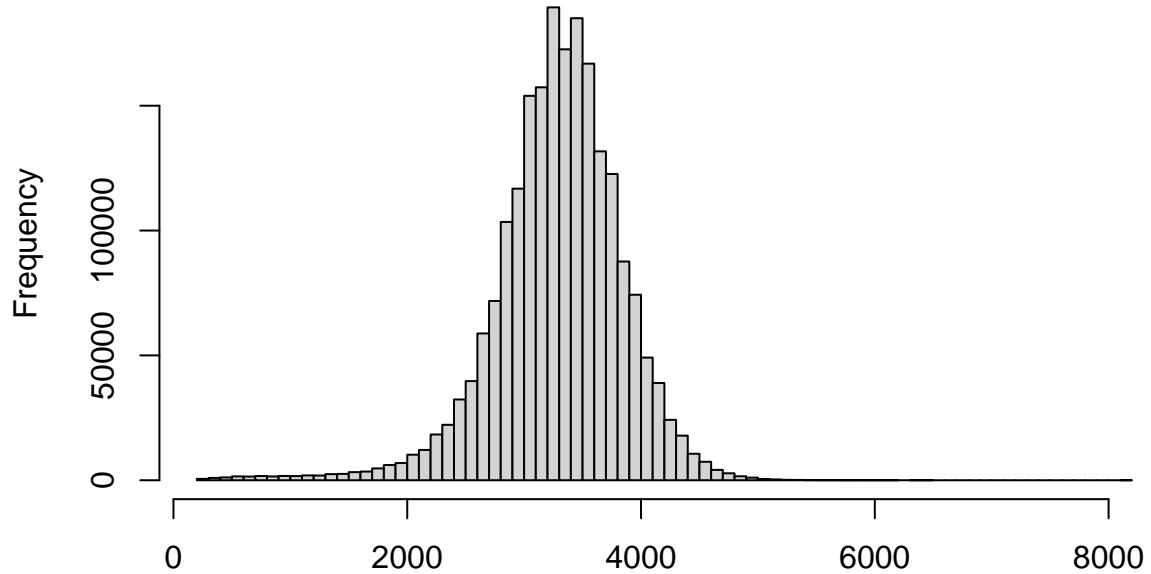
```
## [1] 7815 7853 7860 7875 7940 7975 8025 8160 8161 8165
```

El peso de los recien nacidos es complicado determinar un umbral de dato erróneo, ya que por un lado podemos tener bebes prematuros, con pesos muy, muy bajos (en el año 2018 se dio el record de peso mínimo de bebe prematuro con 245 gramos, Saybie, en San Diego). Además el dataset origen no especifica si los bebes fueron viables (sobrevivieron) o no.

Observando los valores superiores, tampoco hay un salto en un valor que podamos identificar como un error de input.

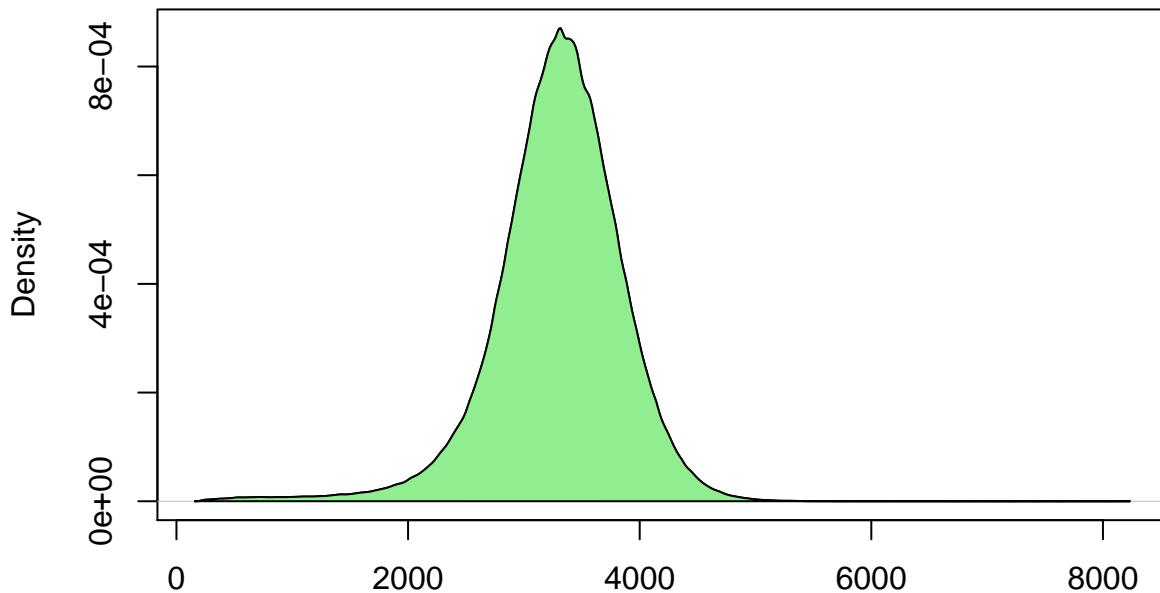
```
hist(data2$PESO_NACER_GM,breaks = 100, main = "Peso del bebé", xlab="")
```

## Peso del bebé



```
plot(density(data2$PESO_NACER_GM), main = "Densidad de peso de los recien nacidos", xlab="")
polygon(density(data2$PESO_NACER_GM) , col="light green", border="black")
```

## Densidad de peso de los recien nacidos



De hecho al graficar con un histograma, o un diagrama de densidad, podemos decir que son valores outliers pero no errores de imputación, por tanto no vamos a hacer ningún tratamiento adicional.

Con este tratamiento consideramos que el dataset está preparado para los análisis que nos hemos planteado en este práctica. Vamos a salvar este dataset ya limpio en el fichero csv US2018\_births\_clean.csv

```
write.csv(data2, "US2018_births_clean.csv", sep = ",", col.names = TRUE, fileEncoding = "Latin1" )

## Warning in write.csv(data2, "US2018_births_clean.csv", sep = ",", col.names =
## TRUE, : attempt to set 'col.names' ignored

## Warning in write.csv(data2, "US2018_births_clean.csv", sep = ",", col.names =
## TRUE, : attempt to set 'sep' ignored
```

## 4. Análisis de los datos

### 4.1 Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Nuestros análisis quieren responder preguntas tanto de índole sociológica (edad, estudios) como de índole biológica (edad, peso, raza), por tanto, vamos a hacer los siguientes grupos de datos a analizar.

```
#RAZA
madres_rc <- data2[data2$RAZA_MADRE == "Caucasiana",]
madres_raf <- data2[data2$RAZA_MADRE == "Afroamericana",]
```

```

madres_rasi <- data2[data2$RAZA_MADRE == "Asiatica",]
madres_r_no_asi <- data2[data2$RAZA_MADRE != "Asiatica",]

madres_rindo <- data2[data2$RAZA_MADRE == "Indoamericana",]
madres_rhaw <- data2[data2$RAZA_MADRE == "Hawaiiana",]
madres_rrresto <- data2[data2$RAZA_MADRE == "Más de una raza",]

madres_rc_primer <- madres_rc[madres_rc$PRIMER_EMBARAZO == "Si",]
madres Raf_primer <- madres Raf [madres Raf $PRIMER_EMBARAZO == "Si",]
madres_rasi_primer <- madres_rasi [madres_rasi $PRIMER_EMBARAZO == "Si",]
madres_r_no_asi_primer <- madres_r_no_asi [madres_r_no_asi $PRIMER_EMBARAZO == "Si",]

#EDUCACION
madres_uni <- data2[data2$EDUC_MADRE == "Universitarios" | data2$EDUC_MADRE == "Postuniversitarios",]
madres_no_uni <- data2[data2$EDUC_MADRE == "Primaria" | data2$EDUC_MADRE == "Secundaria",]

```

Estos son los grupos principales de análisis, sin embargo, como vamos a trabajar con diferentes condicionalidades, sobre el primer embarazo, o pesos de los bebés, decidimos para no cargar de variables el entorno y disminuir la dificultad de lectura, no generar todos los grupos de estudio.

## 4.2 Comprobación de la normalidad y homogeneidad de la varianza

En primer lugar, nuestras cuestiones se basan en la edad de la madre y al peso del hijo por tanto debemos contrastar si estas siguen una distribución normal y así realizar contrastes de hipótesis paramétricos o no paramétricos.

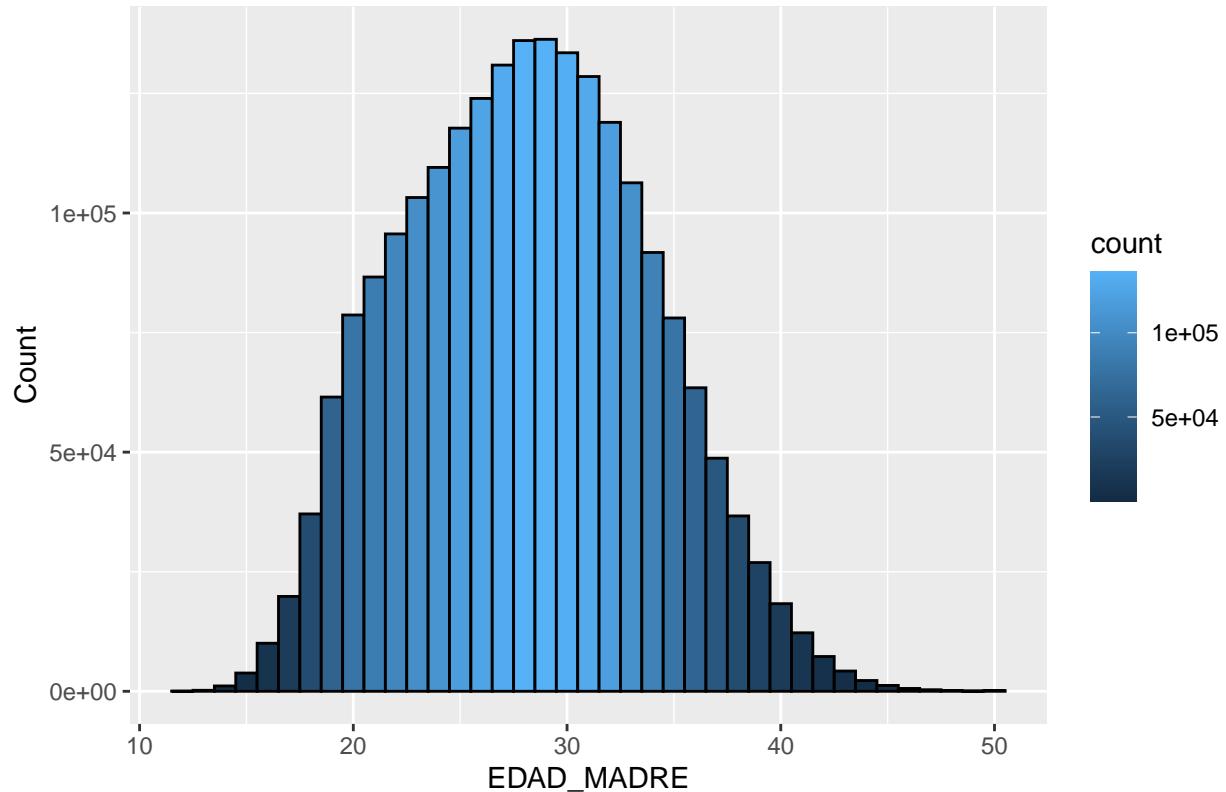
### *Normalidad de la variable EDAD*

```

pl <- ggplot(data2, aes(x=EDAD_MADRE))
pl2 = pl + geom_histogram(binwidth=1, aes(fill=..count..), col='black')
pl2 + xlab('EDAD_MADRE') + ylab('Count') + ggtitle('Histograma EDAD_MADRE')

```

## Histograma EDAD\_MADRE



Vamos a aplicar el test de normalidad de Anderson Darling.

```
ad.test(data2$EDAD_MADRE)
```

```
##  
##  Anderson-Darling normality test  
##  
##  data:  data2$EDAD_MADRE  
##  A = 5565.1, p-value < 2.2e-16
```

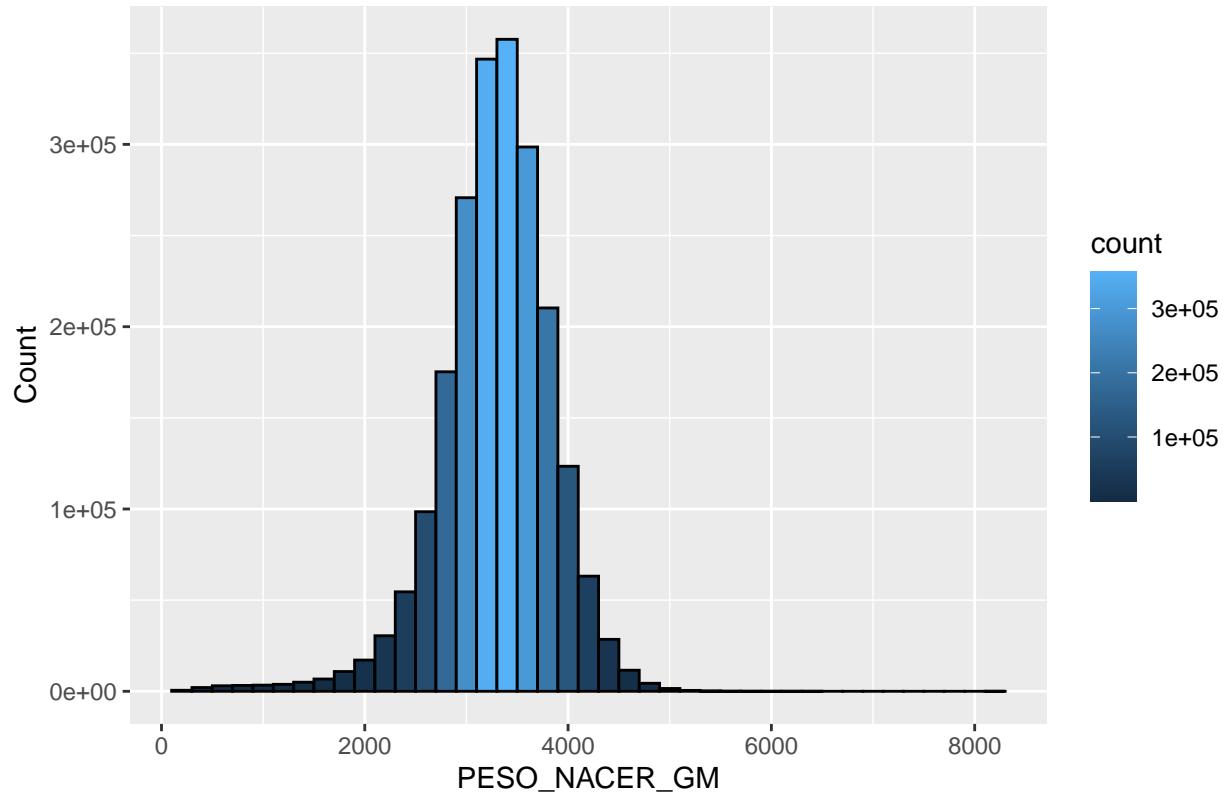
Como podemos ver el p\_value es muy inferior a 0.05 y podemos asegurar que la variable no sigue una distribución normal.

Dado que las preguntas que nos vamos a intentar responder son relativas a hipótesis sobre las medias, por el teorema del límite central, podemos decir que las medias si van a distribuirse como una normal.

*Normalidad de la variable PESO\_NACER\_GM*

```
pl <- ggplot(data2, aes(x=PESO_NACER_GM))  
pl2 = pl + geom_histogram(binwidth=200, aes(fill=..count..), col='black')  
pl2 + xlab('PESO_NACER_GM') + ylab('Count') + ggtitle('Histograma PESO_NACER_GM')
```

## Histograma PESO\_NACER\_GM



Aplicando el test de normalidad de Anderson Darling.

```
ad.test(data2$PESO_NACER_GM)
```

```
##  
##  Anderson-Darling normality test  
##  
##  data:  data2$PESO_NACER_GM  
##  A = 11557, p-value < 2.2e-16
```

Como podemos ver el p\_value es muy inferior a 0.05 y podemos asegurar que la variable no sigue una distribución normal.

Dado que las preguntas que nos vamos a intentar responder son relativas a hipótesis sobre las medias, por el teorema del límite central, podemos decir que las medias si van a distribuirse como una normal.

*¿Son los grupos de pares homocedásticos? Test de Fligner-Killeen*

Por el límite central ya sabemos que las medias de edad de las mujeres van a seguir una distribución normal, pero para determinar el test a realizar, y el estadístico correspondiente, vamos a comprobar si la varianza de las muestras es igual o diferente.

## EDAD MADRE CAUCASIANA VS EDAD MADRE AFROAMERICANA

Test sobre la varianza:

```

var.test(madres_rc$EDAD_MADRE, madres_raf$EDAD_MADRE)

##
## F test to compare two variances
##
## data: madres_rc$EDAD_MADRE and madres_raf$EDAD_MADRE
## F = 0.92021, num df = 1596805, denom df = 341530, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.9158537 0.9245827
## sample estimates:
## ratio of variances
## 0.9202131

```

El resultado nos indica que la varianza no es la misma en ambos grupos de muestras, dado que p-value es casi cero, muy por debajo del nivel de significancia 0.05.

### **EDAD MADRE CAUCASIANA PRIMER HIJO VS EDAD MADRE AFROAMERICANA PRIMER HIJO**

Test sobre la varianza:

```

var.test(madres_rc_primer$EDAD_MADRE, madres_raf_primer$EDAD_MADRE)

```

```

##
## F test to compare two variances
##
## data: madres_rc_primer$EDAD_MADRE and madres_raf_primer$EDAD_MADRE
## F = 0.9229, num df = 1013747, denom df = 223486, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.9174996 0.9283220
## sample estimates:
## ratio of variances
## 0.922903

```

### **EDAD MADRE ASIATICA PRIMER HIJO VS EDAD MADRE NO ASIATICA PRIMER HIJO**

Test sobre la varianza:

```

var.test(madres_rasi_primer$EDAD_MADRE, madres_r_no_asi_primer$EDAD_MADRE)

```

```

##
## F test to compare two variances
##
## data: madres_rasi_primer$EDAD_MADRE and madres_r_no_asi_primer$EDAD_MADRE
## F = 0.77292, num df = 63646, denom df = 1288751, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.7645006 0.7814857
## sample estimates:
## ratio of variances
## 0.7729229

```

## EDAD MADRE CON ESTUDIOS UNIVERSITARIOS SIENDO SU PRIMER HIJO VS EDAD MADRE SIN ESTUDIOS UNIVERSITARIOS SIENDO SU PRIMER HIJO

Test sobre la varianza:

```
var.test(madres_uni$EDAD_MADRE[madres_uni$PRIMER_EMBARAZO == "Si"], madres_no_uni$EDAD_MADRE[madres_no_1

## 
## F test to compare two variances
##
## data: madres_uni$EDAD_MADRE[madres_uni$PRIMER_EMBARAZO == "Si"] and madres_no_uni$EDAD_MADRE[madres_no_
## F = 0.62916, num df = 538579, denom df = 805672, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.6267905 0.6315431
## sample estimates:
## ratio of variances
## 0.6291601
```

Como en el caso anterior, la varianza es diferente ya que tenemos un p\_value muy inferior al nivel de significancia, 0.05. Por tanto, usamos el mismo test que en el caso anterior

Todos los pares de grupos presentan heterocedasticidad por lo que a la hora de aplicar los contrastes de hipótesis se realizarán de la manera pertinente.

**4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.**

**¿Las madres de raza caucasiana tienen hijos en media con mayor edad que las de raza afroamericana ?**

La hipótesis se formula de la siguiente forma:

$$H_0 : \mu_C = \mu_A \quad H_1 : \mu_C > \mu_A$$

donde  $\mu_C$  es la media de edad de las mujeres de raza caucasiana, y  $\mu_A$  es la media de edad de las mujeres de raza afroamericana.

Como las varianzas son diferentes, tal y como hemos comprobado en el test del punto 4.2, para resolver nuestra pregunta debemos hacer el test sobre dos muestras con distribuciones normales y varianzas desconocidas diferentes  $N(\mu_C, \sigma_C), N(\mu_A, \sigma_A)$ , con lo que el estadístico de contraste es:  $t = \frac{\bar{X}_C - \bar{X}_A}{\sqrt{(\frac{s_C^2}{n_C} + \frac{s_A^2}{n_A})}}$ , que se comporta como una t de Student de  $\nu$  grados de libertad, donde  $\nu$  se calcula como  $\nu = \frac{(\frac{s_C^2}{n_C} + \frac{s_A^2}{n_A})^2}{\frac{(s_C^2/n_C)^2}{n_C-1} + \frac{(s_A^2/n_A)^2}{n_A-1}}$

Usando la función en r t.test con el parámetro var.equal = FALSE :

```
t.test(madres_rc$EDAD_MADRE, madres Raf$EDAD_MADRE, alternative = "greater", var.equal = FALSE)

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

```

## data: madres_rc$EDAD_MADRE and madres_raf$EDAD_MADRE
## t = 136.23, df = 485179, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.475576      Inf
## sample estimates:
## mean of x mean of y
## 28.32864 26.83503

```

Con el p-value muy pequeño se puede afirmar que en media, las mujeres de raza caucasiana que dan a luz en Estados Unidos son tienen mayor edad que las de raza afroamericana.

### **¿Las madres de raza caucasiana tienen su primer hijo en media con mayor edad que las de raza afroamericana?**

La hipótesis se formula de la siguiente forma:

$$H_0 : \mu_C = \mu_A \quad H_0 : \mu_C > \mu_A$$

donde  $\mu_C$  es la media de edad de las mujeres de raza caucasiana, y  $\mu_A$  es la media de edad de las mujeres de raza afroamericana

Por el límite central ya sabemos que la medias de edad de las mujeres van a seguir una distribución normal, pero para determinar el test a realizar, y el estadístico correspondiente, vamos a comprobar si la varianza es igual o diferente.

Igual que en el caso anterior, las varianzas de los grupos son diferentes. Por tanto para resolver nuestra pregunta debemos hacer el test sobre dos muestras con distribuciones normales y varianzas desconocidas diferentes  $N(\mu_C, \sigma_C)$ ,  $N(\mu_A, \sigma_A)$ , con lo que el estadístico de contraste es:  $t = \frac{\bar{X}_C - \bar{X}_A}{\sqrt{(\frac{s_C^2}{n_C} + \frac{s_A^2}{n_A})}}$ , que se comporta como una t de Student de  $\nu$  muestras, donde  $\nu$  se calcula como  $\nu = \frac{(\frac{s_C^2}{n_C} + \frac{s_A^2}{n_A})^2}{\frac{(s_C^2/n_C)^2}{n_C-1} + \frac{(s_A^2/n_A)^2}{n_A-1}}$

Usando la función en r t.test con los parámetros adecuados nos lo resuelve:

```
t.test(madres_rc_primer$EDAD_MADRE, madres_raf_primer$EDAD_MADRE, alternative = "greater", var.equal =
```

```

##
## Welch Two Sample t-test
##
## data: madres_rc_primer$EDAD_MADRE and madres_raf_primer$EDAD_MADRE
## t = 98.204, df = 320751, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.231928      Inf
## sample estimates:
## mean of x mean of y
## 29.55012 28.29721

```

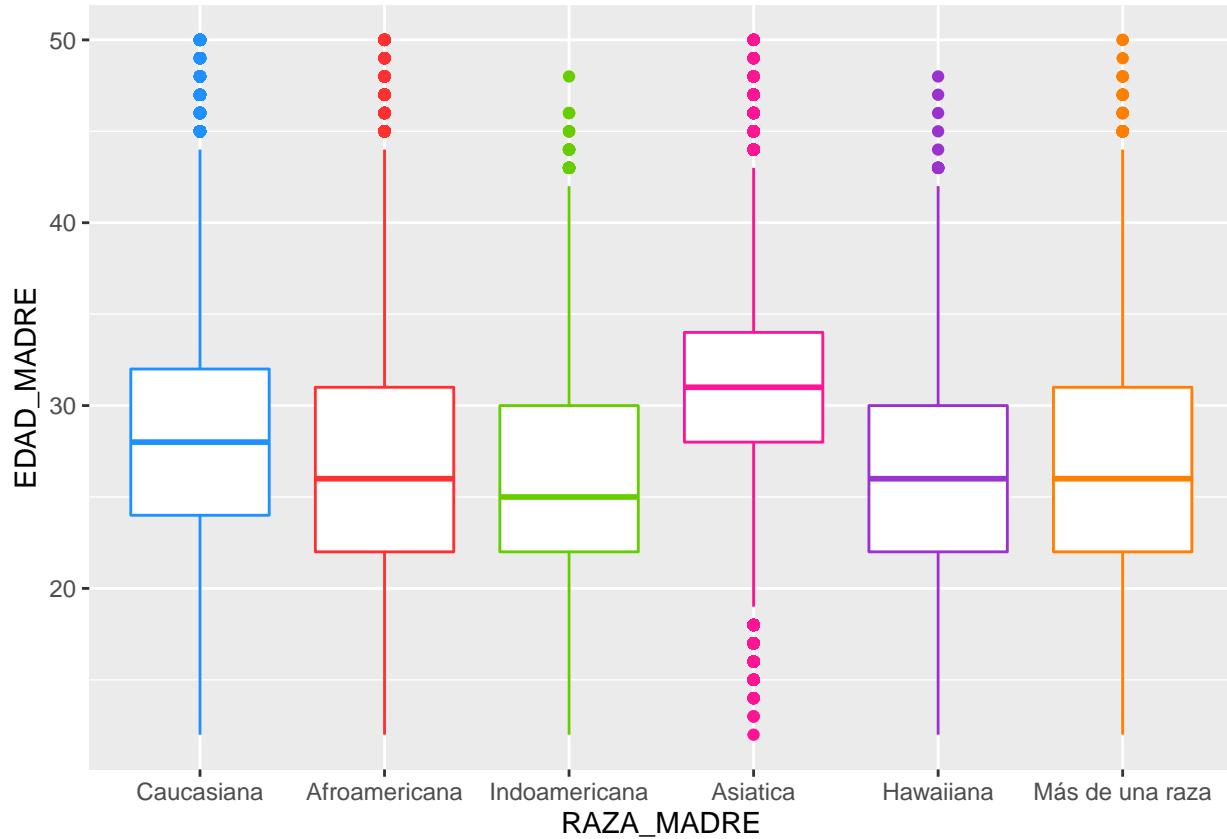
La diferencia aún es mayor.

Llegado a este punto, podemos intentar hacer un test ANOVA para comprobar de una forma más sistemática si estas diferencias las encontramos en general entre todas las razas, tal y como podría sugerir el siguiente gráfico boxplot

```

col <- c("dodgerblue", "firebrick1", "chartreuse3",
       "deeppink", "darkorchid3", "darkorange1")
ggplot(data2, aes(x = RAZA_MADRE, y = EDAD_MADRE)) + geom_boxplot(color = col)

```



A simple vista parece que las mujeres que tienen los hijos en media a mayor edad son las asiáticas.

**¿Las madres de raza asiática tienen su primer hijo en media con mayor edad que las de raza no asiática?**

Vamos a hacer una comprobación también. La hipótesis se formula de la siguiente forma:

$$H_0 : \mu_A = \mu_{noA} \quad H_0 : \mu_A > \mu_{noA}$$

donde  $\mu_A$  es la media de edad de las mujeres de raza caucasiana, y  $\mu_{noA}$  es la media de edad de las mujeres de raza afroamericana

Por el límite central ya sabemos que las medias de edad de las mujeres van a seguir una distribución normal, pero para determinar el test a realizar, y el estadístico correspondiente, vamos a comprobar si la varianza es igual o diferente.

Igual que en el caso anterior, las varianzas de los grupos son diferentes. Por tanto para resolver nuestra pregunta debemos hacer el test sobre dos muestras con distribuciones normales y varianzas desconocidas diferentes  $N(\mu_C, \sigma_C), N(\mu_A, \sigma_A)$ , con lo que el estadístico de contraste es:  $t = \frac{\bar{X}_A - \bar{X}_{noA}}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_{noA}^2}{n_{noA}}}}$ , que se comporta

como una t de Student de  $\nu$  muestras, donde  $\nu$  se calcula como  $\nu = \frac{\left(\frac{s_A^2}{n_A} + \frac{s_{noA}^2}{n_{noA}}\right)^2}{\frac{(s_A^2/n_A)^2}{n_A-1} + \frac{(s_{noA}^2/n_{noA})^2}{n_{noA}-1}}$

Usando la función en r t.test con los parámetros adecuados nos lo resuelve:

```
t.test(madres_rasi_primer$EDAD_MADRE, madres_r_no_así_primer$EDAD_MADRE, alternative = "greater", var.e
```

```
##  
## Welch Two Sample t-test  
##  
## data: madres_rasi_primer$EDAD_MADRE and madres_r_no_así_primer$EDAD_MADRE  
## t = 131.01, df = 72025, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## 2.490087      Inf  
## sample estimates:  
## mean of x mean of y  
## 31.78415 29.26241
```

Efectivamente, las mujeres asiáticas tienen su primer hijo siendo en media más mayores que la media del resto de las razas.

Llegado a este punto vamos a ver como de bueno podría ser la raza como un elemento diferenciador de la media de la edad de las madres por raza mediante un ANOVA.

```
m_anova_raza <- aov(EDAD_MADRE ~ RAZA_MADRE, data = data2[data2$PRIMER_EMBARAZO=="Si",])  
summary(m_anova_raza)
```

```
##          Df  Sum Sq Mean Sq F value Pr(>F)  
## RAZA_MADRE     5  792020 158404    5643 <2e-16 ***  
## Residuals 1352393 37962458       28  
## ---  
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Efectivamente, el test de ANOVA nos dice con un p-value mucho menor que 0.05, que al menos hay dos grupos que son diferentes de cara al primer embarazo, como ya habíamos visto.

Revisamos varianzas por grupo.

```
bartlett.test(EDAD_MADRE ~ RAZA_MADRE, data2)
```

```
##  
## Bartlett test of homogeneity of variances  
##  
## data: EDAD_MADRE by RAZA_MADRE  
## Bartlett's K-squared = 5887.2, df = 5, p-value < 2.2e-16
```

Este test nos dice que no hay homogeneidad ya que el p-value es menor que el umbral de significancia, con lo que tendríamos que ir por pares para ver si par a par las medias son significativamente diferentes.

*Vamos ahora a analizar desde el punto de vista de nivel de estudios. Para ello vamos a hacernos la pregunta*

**¿Las madres con estudios universitarios tienen en media su primer hijo con mayor edad que las que no tienen estudios superiores?**

La hipótesis se formula de la siguiente forma:

$$H_0 : \mu_{uni} = \mu_{nouni} \quad H_1 : \mu_{uni} > \mu_{nouni}$$

donde  $\mu_{uni}$  es la media de edad de las mujeres con estudios superiores, y  $\mu_{nouni}$  es la media de edad de las mujeres que no tienen estudios superiores

Por el límite central ya sabemos que las medias de edad de las mujeres van a seguir una distribución normal y además como comprobamos anteriormente la varianza es diferente entre los grupos.

```
t.test(madres_uni$EDAD_MADRE[madres_uni$PRIMER_EMBARAZO == "Si"], madres_no_uni$EDAD_MADRE[madres_no_uni$PRIMER_EMBARAZO == "No"],  
        alternative = "greater", var.equal = FALSE)
```

```
##  
##  Welch Two Sample t-test  
##  
## data:  madres_uni$EDAD_MADRE [madres_uni$PRIMER_EMBARAZO == "Si"] and madres_no_uni$EDAD_MADRE [madres_no_uni$PRIMER_EMBARAZO == "No"]  
## t = 514.26, df = 1305707, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
##  4.248094      Inf  
## sample estimates:  
## mean of x mean of y  
##  31.93185   27.67012
```

*¿Podríamos decir que la diferencia es mayor de 4 años con un nivel de confianza del 99%?*

```
t.test(madres_uni$EDAD_MADRE[madres_uni$PRIMER_EMBARAZO == "Si"], madres_no_uni$EDAD_MADRE[madres_no_uni$PRIMER_EMBARAZO == "No"],  
        alternative = "greater", mu=4, conf.level=0.99, var.equal = FALSE)
```

```
##  
##  Welch Two Sample t-test  
##  
## data:  madres_uni$EDAD_MADRE [madres_uni$PRIMER_EMBARAZO == "Si"] and madres_no_uni$EDAD_MADRE [madres_no_uni$PRIMER_EMBARAZO == "No"]  
## t = 31.582, df = 1305707, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is greater than 4  
## 99 percent confidence interval:  
##  4.242446      Inf  
## sample estimates:  
## mean of x mean of y  
##  31.93185   27.67012
```

El resultado es concluyente dado el p\_value, efectivamente el nivel de estudios influye en la edad del primer hijo, retrasando en media más de cuatro años con un nivel de confianza del 99%.

**¿Qué factores sociológicos ,fisiológicos y culturales favorecen que una mujer tenga su primer embarazo?**

### Modelo logistico

Variables

```

PRIMER_EMBARAZO <- data2$PRIMER_EMBARAZO
EDAD_MADRE <- data2$EDAD_MADRE
ESTADO_CIVIL <- data2$ESTADO_CIVIL
RAZA_MADRE <- data2$RAZA_MADRE
EDUC_MADRE <- data2$EDUC_MADRE

```

Vamos a evaluar diferentes modelos incluyendo nuevas variables acerca de la madre de manera que podamos obtener un modelo logístico que se ajuste de manera adecuada a los datos.

### Modelo1

```

model.logist1=glm(formula=PRIMER_EMBARAZO~EDAD_MADRE ,family=binomial(link=logit))
summary(model.logist1)

```

```

##
## Call:
## glm(formula = PRIMER_EMBARAZO ~ EDAD_MADRE, family = binomial(link = logit))
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -1.6735   -0.9392   -0.7026    1.1382    2.5054
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 2.5969687  0.0077975 333.0   <2e-16 ***
## EDAD_MADRE -0.1138222  0.0002813 -404.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2799465  on 2131842  degrees of freedom
## Residual deviance: 2616118  on 2131841  degrees of freedom
## AIC: 2616122
##
## Number of Fisher Scoring iterations: 4

```

### Modelo2

```

model.logist2=glm(formula=PRIMER_EMBARAZO~EDAD_MADRE + ESTADO_CIVIL ,family=binomial(link=logit))
summary(model.logist2)

```

```

##
## Call:
## glm(formula = PRIMER_EMBARAZO ~ EDAD_MADRE + ESTADO_CIVIL, family = binomial(link = logit))
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -1.6728   -0.9347   -0.7046    1.1668    2.6656
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)

```

```

## (Intercept)      3.0343686  0.0094030  322.70  <2e-16 ***
## EDAD_MADRE     -0.1253883  0.0003155 -397.39  <2e-16 ***
## ESTADO_CIVILSoltera -0.2885796  0.0033589  -85.91  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2799465  on 2131842  degrees of freedom
## Residual deviance: 2608630  on 2131840  degrees of freedom
## AIC: 2608636
##
## Number of Fisher Scoring iterations: 4

```

### Modelo3

```

model.logist3=glm(formula=PRIMER_EMBARAZO~EDAD_MADRE + ESTADO_CIVIL + RAZA_MADRE ,family=binomial(link=logit))
summary(model.logist3)

```

```

##
## Call:
## glm(formula = PRIMER_EMBARAZO ~ EDAD_MADRE + ESTADO_CIVIL + RAZA_MADRE,
##       family = binomial(link = logit))
##
## Deviance Residuals:
##    Min      1Q   Median      3Q      Max
## -1.9290 -0.9426 -0.6933  1.1522  2.7468
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)            3.0876593  0.0094707 326.02  <2e-16 ***
## EDAD_MADRE           -0.1281453  0.0003185 -402.37  <2e-16 ***
## ESTADO_CIVILSoltera -0.2088259  0.0034937  -59.77  <2e-16 ***
## RAZA_MADREAfroamericana -0.2206733  0.0043143  -51.15  <2e-16 ***
## RAZA_MADREIndoamericana -0.4838331  0.0159026  -30.43  <2e-16 ***
## RAZA_MADREAsiatica    0.6066863  0.0065442   92.71  <2e-16 ***
## RAZA_MADREHawaiiana   -0.4157200  0.0287755  -14.45  <2e-16 ***
## RAZA_MADREMás de una raza -0.1459217  0.0095183  -15.33  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2799465  on 2131842  degrees of freedom
## Residual deviance: 2595631  on 2131835  degrees of freedom
## AIC: 2595647
##
## Number of Fisher Scoring iterations: 4

```

### Modelo4

```

model.logist4=glm(formula=PRIMER_EMBARAZO~EDAD_MADRE + ESTADO_CIVIL + RAZA_MADRE + EDUC_MADRE ,family=binomial(link=logit))
summary(model.logist4)

##
## Call:
## glm(formula = PRIMER_EMBARAZO ~ EDAD_MADRE + ESTADO_CIVIL + RAZA_MADRE +
##       EDUC_MADRE, family = binomial(link = logit))
##
## Deviance Residuals:
##      Min        1Q     Median        3Q        Max
## -2.1090   -0.9439   -0.5945    1.0911    3.3198
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 3.4011265  0.0106087 320.597 < 2e-16 ***
## EDAD_MADRE                  -0.1776700  0.0003797 -467.898 < 2e-16 ***
## ESTADO_CIVILSoltera          0.1544363  0.0038029  40.610 < 2e-16 ***
## RAZA_MADREAfroamericana    -0.2016737  0.0044488 -45.332 < 2e-16 ***
## RAZA_MADREIndoamericana    -0.3613723  0.0163344 -22.123 < 2e-16 ***
## RAZA_MADREAsiatica          0.4873470  0.0067872  71.804 < 2e-16 ***
## RAZA_MADREHawaiiana         -0.1928076  0.0298063 -6.469 9.89e-11 ***
## RAZA_MADREMás de una raza   -0.1409949  0.0097153 -14.513 < 2e-16 ***
## EDUC_MADRESecundaria        0.4631045  0.0051977  89.097 < 2e-16 ***
## EDUC_MADREUniversitarios    1.4247052  0.0061070 233.292 < 2e-16 ***
## EDUC_MADREPostuniversitarios 1.9521025  0.0073072 267.146 < 2e-16 ***
## EDUC_MADREDesconocido        0.7483914  0.0219825  34.045 < 2e-16 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2799465 on 2131842 degrees of freedom
## Residual deviance: 2486584 on 2131831 degrees of freedom
## AIC: 2486608
##
## Number of Fisher Scoring iterations: 4

```

## Modelos

Modelo 1: PRIMER\_EMBARAZO ~ EDAD\_MADRE  
 Modelo 2: PRIMER\_EMBARAZO ~ EDAD\_MADRE + ESTADO\_CIVIL  
 Modelo 3: PRIMER\_EMBARAZO ~ EDAD\_MADRE + ESTADO\_CIVIL + RAZA\_MADRE  
 Modelo 4: PRIMER\_EMBARAZO ~ EDAD\_MADRE + ESTADO\_CIVIL + RAZA\_MADRE + EDUC\_MADRE

Matriz de confusión de los modelos y cálculo de la sensibilidad y especificidad

```

# Predicciones del modelo
p1=predict(model.logist1, data2, type="response")
p2=predict(model.logist2, data2, type="response")
p3=predict(model.logist3, data2, type="response")
p4=predict(model.logist4, data2, type="response")

```

```

mc1 <- table(p1>0.5, data2$PRIMER_EMBARAZO)
mc2 <- table(p2>0.5, data2$PRIMER_EMBARAZO)
mc3 <- table(p3>0.5, data2$PRIMER_EMBARAZO)
mc4 <- table(p4>0.5, data2$PRIMER_EMBARAZO)

s1 <- mc1[2,1]/(mc1[1,1]+mc1[2,1])
e1 <- mc1[1,2]/(mc1[2,2]+mc1[1,2])
s2 <- mc2[2,1]/(mc2[1,1]+mc2[2,1])
e2 <- mc2[1,2]/(mc2[2,2]+mc2[1,2])
s3 <- mc3[2,1]/(mc3[1,1]+mc3[2,1])
e3 <- mc3[1,2]/(mc3[2,2]+mc3[1,2])
s4 <- mc4[2,1]/(mc4[1,1]+mc4[2,1])
e4 <- mc4[1,2]/(mc4[2,2]+mc4[1,2])

modelo <- c("Modelo1","Modelo2","Modelo3","Modelo4")
sens <- c(s1,s2,s3,s4)
esp <- c(e1,e2,e3,e4)

confs <- data.frame(modelo = modelo, sensibilidad = sens, especificidad = esp)
confs

##     modelo sensibilidad especificidad
## 1 Modelo1      0.1083593    0.6818322
## 2 Modelo2      0.1213998    0.6698711
## 3 Modelo3      0.1332743    0.6460990
## 4 Modelo4      0.1530702    0.5434271

```

Tras realizar los distintos modelos obtenemos los siguientes resultados:

- Todas las variables incluidas son significativas a un nivel de confianza del 99%.
- Analizamos la devianza y en todos los modelos va mejorando ya que reduce la devianza, si bien no es una reducción drástica.
- Los AIC de los modelos son:
  - AIC Modelo 1: 2616122
  - AIC Modelo 2: 2608636
  - AIC Modelo 3: 2595647
  - AIC Modelo 4: 2486608

Por lo que el último modelo es el que presenta un mejor ajuste.

Calculamos la calidad de los modelos y obtenemos según su capacidad predictiva para un umbral de discriminación del 50% y obtenemos que el modelo con unos mejores parámetros de sensibilidad es el modelo 4 y el modelo con mayor especificidad es el 1.

Para analizar el comportamiento de los modelos en diferentes umbrales de probabilidad vamos a realizar el análisis ROC.

### **CURVA ROC DE LOS MODELOS**

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```

##  

## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':  

##  

##     cov, smooth, var

# Análisis ROC
r1 = roc(data2$PRIMER_EMBARAZO, p1, data=data, auc=T, ci=T)

## Setting levels: control = Si, case = No

## Setting direction: controls < cases

r2 = roc(data2$PRIMER_EMBARAZO, p2, data=data, auc=T, ci=T)

## Setting levels: control = Si, case = No
## Setting direction: controls < cases

r3 = roc(data2$PRIMER_EMBARAZO, p3, data=data, auc=T, ci=T)

## Setting levels: control = Si, case = No
## Setting direction: controls < cases

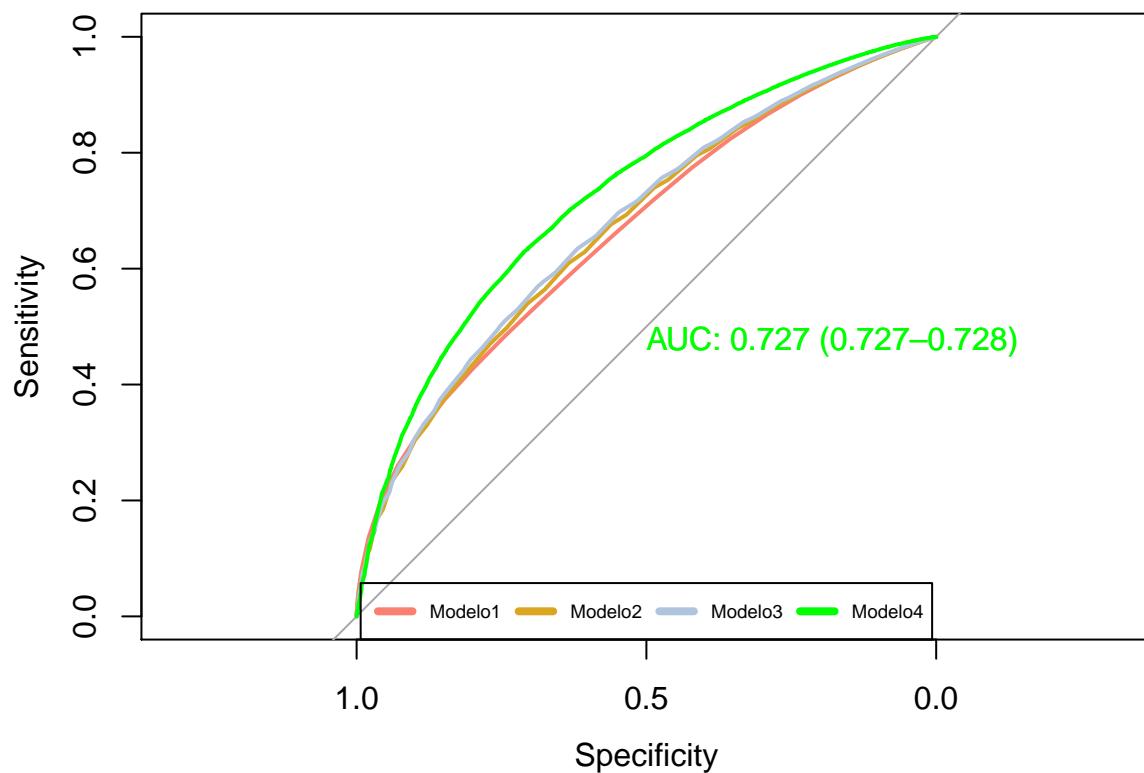
r4 = roc(data2$PRIMER_EMBARAZO, p4, data=data, auc=T, ci=T)

## Setting levels: control = Si, case = No
## Setting direction: controls < cases

plot.roc(r1, plot=TRUE, legacy.axes=FALSE, col="salmon", lwd=2)
plot.roc(r2, col="goldenrod", lwd=2, add=TRUE)
plot.roc(r3, col="lightsteelblue", lwd=2, add=TRUE)
plot.roc(r4, col="green", lwd=2, print.auc=TRUE, add=TRUE)

legend("bottom",
       legend=c("Modelo1", "Modelo2", "Modelo3", "Modelo4"),
       col=c("salmon", "goldenrod", "lightsteelblue", "green"),
       lwd=4, cex = 0.6, xpd = TRUE, horiz = TRUE)

```



El AUC proporciona una medición agregada del rendimiento en todos los umbrales de probabilidad posibles. Observamos que el modelo 4 es el modelo que obtiene un mayor AUC con diferencia y por tanto un mayor rendimiento ante diferentes umbrales.

#### Interpretación del modelo

Vamos a realizar una interpretación de los coeficientes del modelo.

En el modelo esta construido teniendo como referencia a madres casadas de raza blanca con educación primaria. Analizando los signos de los coeficientes podemos observar que la probabilidad de que este sea el primer hijo de la madre disminuye conforme aumentan la edad de la madre.

A nivel de estado civil observamos que la probabilidad de que sea el primer hijo de la madre aumenta en el caso de que esta sea soltera con respecto a que este casada.

A nivel de raza la probabilidad de que sea el primer hijo de las madres afroamericanas, indoamericanas, hawaianas y más de una raza disminuye con respecto a las madres blancas. En el caso de las madres asiáticas aumenta considerablemente.

A nivel de educación observamos que la probabilidad de que sea el primer hijo aumenta para todos los niveles superiores a primaria y considerablemente en el caso de que la madre tenga estudios universitarios o postuniversitarios.

#### Predictión con el modelo 4

¿Qué probabilidad existe de que una mujer embarazada de 19 años de edad, soltera, de raza caucasiana y con estudios de primaria de a luz a su primer hijo?

```
p <- predict(model.logist4, newdata = data.frame( EDAD_MADRE = 19,
ESTADO_CIVIL = "Soltera",
```

```

        RAZA_MADRE = "Caucasiana",
        EDUC_MADRE = "Primaria"))

sprintf("La probabilidad de que una mujer embarazada de 19 años de edad, soltera, de raza caucasiana y c

```

```
## [1] "La probabilidad de que una mujer embarazada de 19 años de edad, soltera, de raza caucasiana y c
```

¿Qué probabilidad existe de que una madre embarazada de 25 años de edad, casada, de raza asiática y con estudios de universitarios de a luz a su primer hijo?

```

p <- predict(model.logist4, newdata = data.frame( EDAD_MADRE = 25,
                                                ESTADO_CIVIL = "Casada",
                                                RAZA_MADRE = "Caucasiana",
                                                EDUC_MADRE = "Postuniversitarios"))

sprintf("La probabilidad de que una madre embarazada con 25 años de edad, soltera, de raza caucasiana y c

```

```
## [1] "La probabilidad de que una madre embarazada con 25 años de edad, soltera, de raza caucasiana y c
```

¿Y con las mismas condiciones pero con raza Afroamericana?

```

p <- predict(model.logist4, newdata = data.frame( EDAD_MADRE = 25,
                                                ESTADO_CIVIL = "Casada",
                                                RAZA_MADRE = "Afroamericana",
                                                EDUC_MADRE = "Postuniversitarios"))

sprintf("La probabilidad de que una madre embarazada con 19 años de edad, soltera, de raza Afroamericana

```

```
## [1] "La probabilidad de que una madre embarazada con 19 años de edad, soltera, de raza Afroamericana
```

## Análisis de predicción del peso del recién nacido en función de las características físico-biológicas de la madre

El primer paso sería hacer un análisis de correlación de las variables numéricas

```

numvars <- c(1:3,9,11,12)
mat.cor <- cor(data2[,numvars], method = "pearson")

mat.cor

##                                     EDAD_MADRE ALTURA_MADRE          IMC GANANCIA_PESO_MADRE
## EDAD_MADRE             1.00000000  0.12272721  0.01762468      -0.01434936
## ALTURA_MADRE           0.12272721  1.00000000 -0.04602352       0.12471226
## IMC                   0.01762468  -0.04602352  1.00000000      -0.21045551
## GANANCIA_PESO_MADRE   -0.01434936  0.12471226 -0.21045551       1.00000000
## PESO_NACER_GM          0.07939566  0.16435121  0.08371561       0.19778856
## PESO_MADRE              0.06371045  0.36585170  0.90926647      -0.14529211
##                                     PESO_NACER_GM    PESO_MADRE
## EDAD_MADRE               0.07939566  0.06371045
## ALTURA_MADRE             0.16435121  0.36585170

```

```

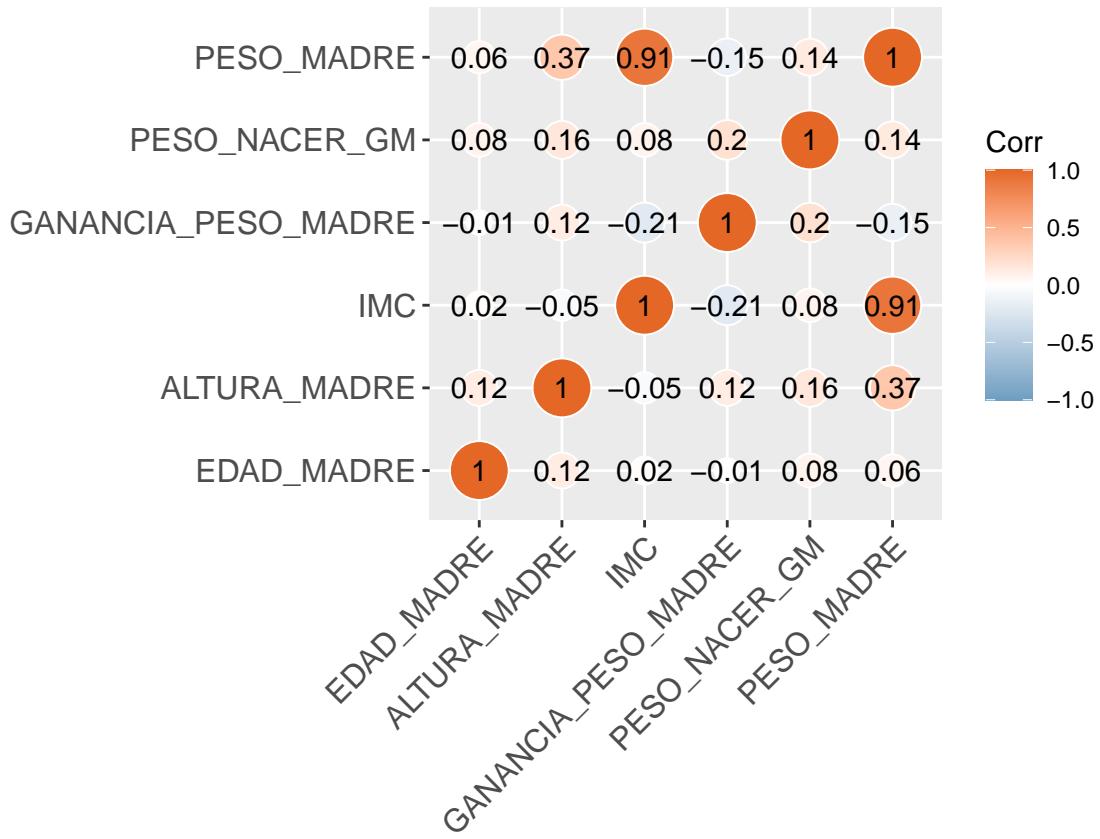
## IMC          0.08371561  0.90926647
## GANANCIA_PESO_MADRE 0.19778856 -0.14529211
## PESO_NACER_GM    1.00000000  0.14370823
## PESO_MADRE      0.14370823  1.00000000

```

```

ggcorrplot(mat.cor,
           method = "circle",
           lab = TRUE,
           outline.color = "white",
           ggtheme = ggplot2::theme_gray,
           colors = c("#6D9EC1", "white", "#E46726"))

```



Los resultados son a priori sorprendentes, dado que se tiende a pensar que las mujeres grandes suelen tener niños más grandes. Sin embargo, desde un punto de vista de correlación no parece que haya ninguna relación suficientemente fuerte. En todo caso, vamos a ver si esta no relación se mantiene por razas.

```

m_peso_bb1 <- lm(PESO_NACER_GM ~ PESO_MADRE, data = data2)
summary(m_peso_bb1)

```

```

##
## Call:
## lm(formula = PESO_NACER_GM ~ PESO_MADRE, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0000 -0.5000 -0.2500  0.2500  1.0000
## 
```

```

## -3381.6 -284.8 32.3 340.3 5011.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.923e+03 1.782e+00   1640 <2e-16 ***
## PESO_MADRE 5.411e+00 2.552e-02    212 <2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 545.9 on 2131841 degrees of freedom
## Multiple R-squared: 0.02065, Adjusted R-squared: 0.02065
## F-statistic: 4.496e+04 on 1 and 2131841 DF, p-value: < 2.2e-16

```

Tal y como se podría predecir de la matriz de correlación, nos encontramos con  $R^2$  muy, muy bajo, del 0.02, muy lejos de ser una variable explicativa.

Incluyendo además otras variables cualitativas

```
m_peso_bb2 <- lm(PESO_NACER_GM ~ PESO_MADRE + RAZA_MADRE, data = data2)
summary(m_peso_bb2)
```

```

##
## Call:
## lm(formula = PESO_NACER_GM ~ PESO_MADRE + RAZA_MADRE, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q      Max
## -3423.8 -280.1    31.6   334.4  5220.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2945.6419    1.7892 1646.314 < 2e-16 ***
## PESO_MADRE           5.7730    0.0255  226.377 < 2e-16 ***
## RAZA_MADREAfroamericana -246.9341   1.0182 -242.529 < 2e-16 ***
## RAZA_MADREIndoamericana -24.0576   3.8095  -6.315 2.7e-10 ***
## RAZA_MADREAsiatica -103.5596   1.6694 -62.033 < 2e-16 ***
## RAZA_MADREHawaiiana -99.7836   6.9217 -14.416 < 2e-16 ***
## RAZA_MADREMás de una raza -89.0536   2.3574 -37.777 < 2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 538.3 on 2131836 degrees of freedom
## Multiple R-squared: 0.04779, Adjusted R-squared: 0.04778
## F-statistic: 1.783e+04 on 6 and 2131836 DF, p-value: < 2.2e-16

```

```
m_peso_bb3 <- lm(PESO_NACER_GM ~ RAZA_MADRE, data = data2)
summary(m_peso_bb3)
```

```

##
## Call:
## lm(formula = PESO_NACER_GM ~ RAZA_MADRE, data = data2)
##
## Residuals:
```

```

##      Min     1Q   Median     3Q    Max
## -3112.0 -285.6    33.0  341.0 5054.4
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3339.0357    0.4311 7745.546 <2e-16 ***
## RAZA_MADREAfroamericana -228.4159    1.0270 -222.412 <2e-16 ***
## RAZA_MADREIndoamericana   -7.1355    3.8543  -1.851  0.0641 .
## RAZA_MADREAsiatica       -152.9678    1.6749 - 91.331 <2e-16 ***
## RAZA_MADREHawaiiana        -86.2300    7.0041 -12.311 <2e-16 ***
## RAZA_MADREMás de una raza -83.0743    2.3854 -34.826 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 544.7 on 2131837 degrees of freedom
## Multiple R-squared:  0.0249, Adjusted R-squared:  0.02489
## F-statistic: 1.089e+04 on 5 and 2131837 DF, p-value: < 2.2e-16

```

```

data3 <- data2[data2$RAZA_MADRE == "Asiatica" | data2$RAZA_MADRE == "Afroamericana",]
m_peso_bb4 <- lm(PESO_NACER_GM ~ IMC + RAZA_MADRE, data = data3)
summary(m_peso_bb4)

```

```

##
## Call:
## lm(formula = PESO_NACER_GM ~ IMC + RAZA_MADRE, data = data3)
##
## Residuals:
##      Min     1Q   Median     3Q    Max
## -3035.3 -275.8    41.6  347.0 5155.1
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2854.927    4.500 634.39 <2e-16 ***
## IMC                  9.604     0.165  58.20 <2e-16 ***
## RAZA_MADREAsiatica 106.621     2.022  52.72 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 568.8 on 454820 degrees of freedom
## Multiple R-squared:  0.01062, Adjusted R-squared:  0.01062
## F-statistic: 2442 on 2 and 454820 DF, p-value: < 2.2e-16

```

Con estos resultados tenemos que descartar construir un modelo de regresión lineal que pueda explicar el peso del bebe en función de estas variables.

### Análisis del peso de los bebes en función de la raza de las madres.

Sin embargo, vamos a hacer una comparativa entre pesos de las madres y los bebes por razas que al principio no lo teníamos en el alcance inicial del trabajo.

```
tapply(data2$PESO_MADRE, data2$RAZA_MADRE, mean)
```

```

##      Caucasiana    Afroamericana    Indoamericana    Asiatica    Hawaiiana
##      68.14353       71.35124       71.07477       59.58507       70.49127
## Más de una raza     Desconocido          NA
##      69.17926

```

```
tapply(data2$PESO_NACER_GM, data2$RAZA_MADRE, mean)
```

```

##      Caucasiana    Afroamericana    Indoamericana    Asiatica    Hawaiiana
##      3339.036       3110.620       3331.900       3186.068       3252.806
## Más de una raza     Desconocido          NA
##      3255.961

```

Llama la atención que las mujeres asiáticas tengan hijos con mayor peso que las afroamericanas, teniendo éstas un peso medio considerablemente mayor que las asiáticas.

Hacemos un test sobre la varianza para determinar el estadístico a estudiar.

```

var.test(data2$PESO_NACER_GM[data2$RAZA_MADRE=="Asiatica"], data3$PESO_NACER_GM[data2$RAZA_MADRE=="Afroamericana"])

##
##  F test to compare two variances
##
## data: data2$PESO_NACER_GM[data2$RAZA_MADRE == "Asiatica"] and data3$PESO_NACER_GM[data2$RAZA_MADRE == "Afroamericana"]
## F = 0.77485, num df = 113291, denom df = 90853, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7653376 0.7844679
## sample estimates:
## ratio of variances
##                 0.774847

```

Una vez más, la varianza es diferente por lo que tenemos que repetir el mismo test que en los puntos anteriores.

```

t.test(data2$PESO_NACER_GM[data2$RAZA_MADRE=="Asiatica"], data3$PESO_NACER_GM[data2$RAZA_MADRE=="Afroamericana"],
       alternative = "greater", var.equal = FALSE)

```

```

##
## Welch Two Sample t-test
##
## data: data2$PESO_NACER_GM[data2$RAZA_MADRE == "Asiatica"] and data3$PESO_NACER_GM[data2$RAZA_MADRE == "Afroamericana"]
## t = 22.393, df = 182372, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  49.95185      Inf
## sample estimates:
## mean of x mean of y
## 3186.068   3132.156

```

Tal y como se aprecia, efectivamente es así, las mujeres asiáticas tienen hijos en media de un peso superior a las afroamericanas.

Por último, vamos a hacer un test de hipótesis para ver si los hijos de las mujeres caucasianas tienen hijos que en media pesan más que los de las otras razas.

De nuevo, test sobre las varianzas:

```
var.test(data2$PESO_NACER_GM[data2$RAZA_MADRE=="Caucasiana"], data2$PESO_NACER_GM[data2$RAZA_MADRE!="Caucasiana"])
```

```
##  
## F test to compare two variances  
##  
## data: data2$PESO_NACER_GM[data2$RAZA_MADRE == "Caucasiana"] and data2$PESO_NACER_GM[data2$RAZA_MADRE != "Caucasiana"]  
## F = 0.87943, num df = 1596805, denom df = 535036, p-value < 2.2e-16  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.8760969 0.8827619  
## sample estimates:  
## ratio of variances  
## 0.8794263
```

Que nos vuelve a dar diferentes

```
t.test(data2$PESO_NACER_GM[data2$RAZA_MADRE=="Caucasiana"], data2$PESO_NACER_GM[data2$RAZA_MADRE!="Caucasiana"],  
       alternative = "greater", var.equal = FALSE)
```

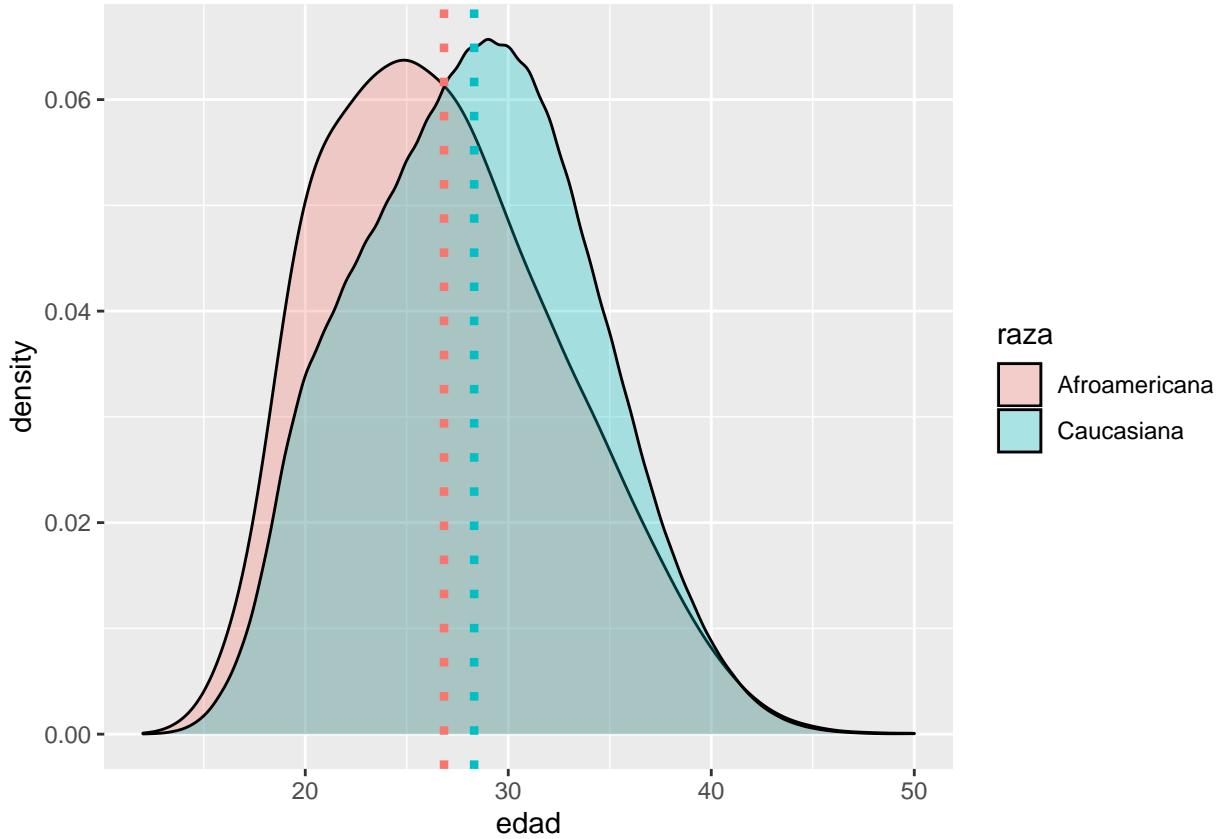
```
##  
## Welch Two Sample t-test  
##  
## data: data2$PESO_NACER_GM[data2$RAZA_MADRE == "Caucasiana"] and data2$PESO_NACER_GM[data2$RAZA_MADRE != "Caucasiana"]  
## t = 211.06, df = 871453, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## 186.3511 Inf  
## sample estimates:  
## mean of x mean of y  
## 3339.036 3151.221
```

Efectivamente, se comprueba que en media los hijos de las mujeres caucasianas pesan más que las del resto de razas.

## 5. Representación de los resultados a partir de tablas y gráficas.

*¿Las madres de raza caucasiana tienen hijos en media con mayor edad que las de raza afroamericana ?*

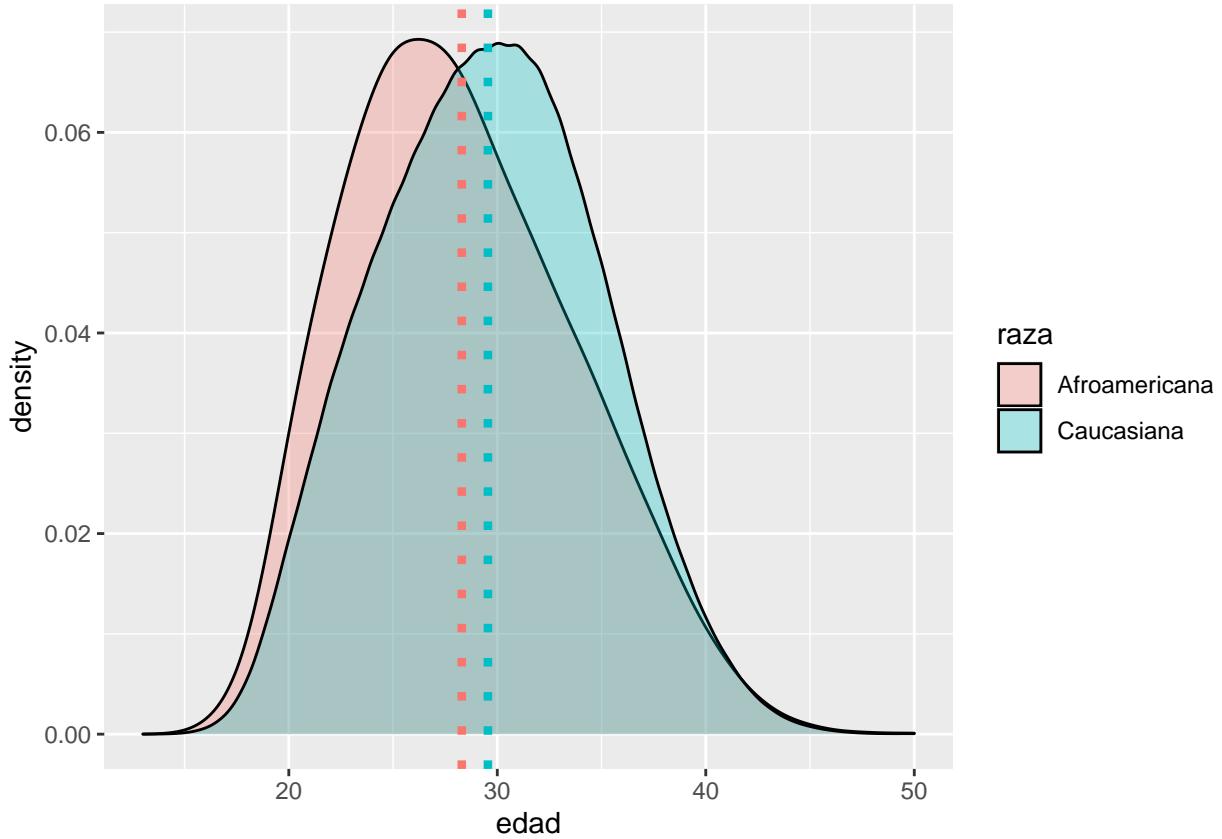
```
tmp <- rbind(data.frame(raza = "Caucasiana", edad = madres_rc$EDAD_MADRE),  
              data.frame(raza = "Afroamericana", edad = madres_raf$EDAD_MADRE))  
  
ggplot(tmp, aes(x = edad, fill = raza)) + geom_density(adjust = 2, alpha=0.3) +  
  geom_vline(xintercept = 28.32, color="#00BFC4", linetype = "dotted", size=1.5) +  
  geom_vline(xintercept = 26.835, color="#F8766D", linetype = "dotted", size=1.5)
```



*¿Las madres de raza caucasiana tienen su primer hijo en media con mayor edad que las de raza afroamericana?*

```
tmp <- rbind(data.frame(raza = "Caucasiana", edad = madres_rc_primer$EDAD_MADRE),
               data.frame(raza = "Afroamericana", edad = madres_raf_primer$EDAD_MADRE))

ggplot(tmp, aes(x = edad, fill = raza)) + geom_density(adjust = 2, alpha=0.3) +
  geom_vline(xintercept = 29.55, color="#00BFC4", linetype ="dotted", size=1.5) +
  geom_vline(xintercept = 28.297,color="#F8766D", linetype ="dotted", size=1.5)
```

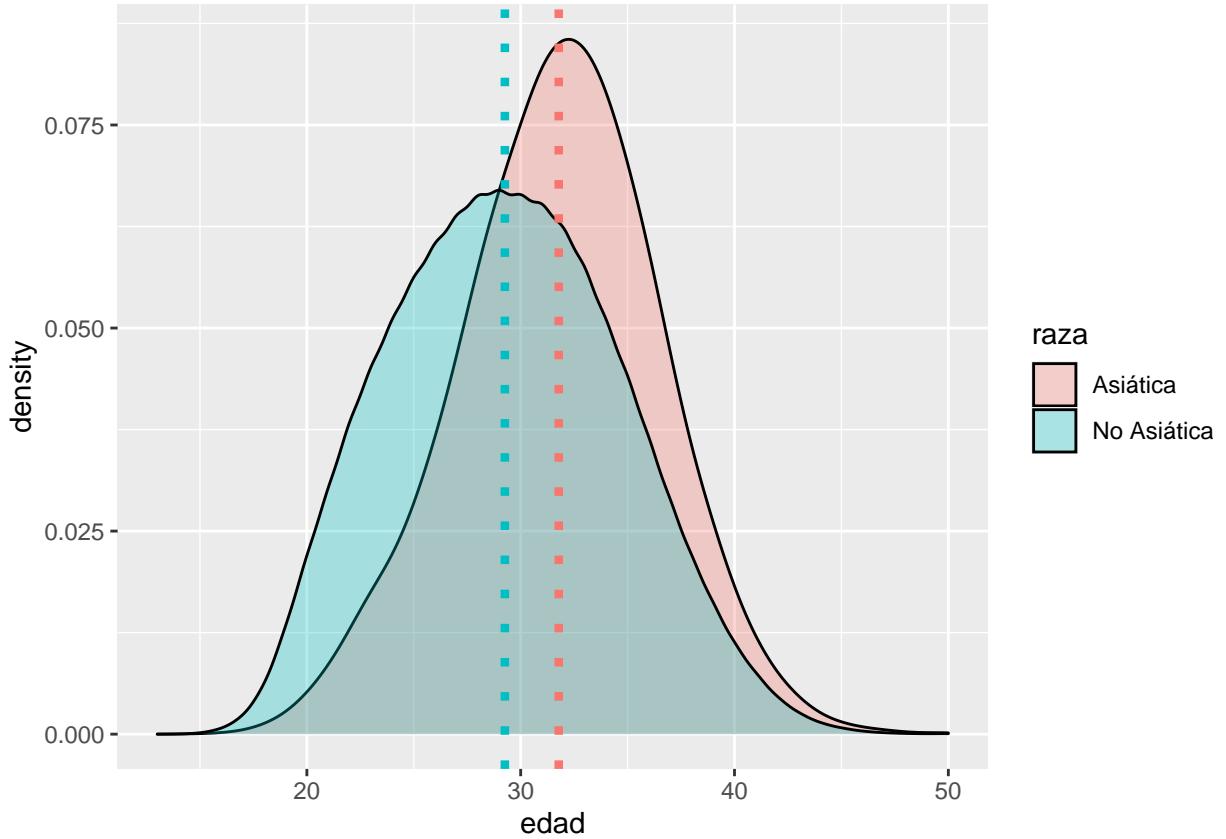


En ambos casos se puede apreciar fácilmente en los gráficos que en media las mujeres caucasianas esperan más a tener los hijos.

*¿Las madres de raza asiática tienen su primer hijo en media con mayor edad que las de raza no asiática?*

```
tmp <- rbind(data.frame(raza = "Asiática", edad = madres_rasi_primer$EDAD_MADRE),
               data.frame(raza = "No Asiática", edad = madres_r_no_asi_primer$EDAD_MADRE))

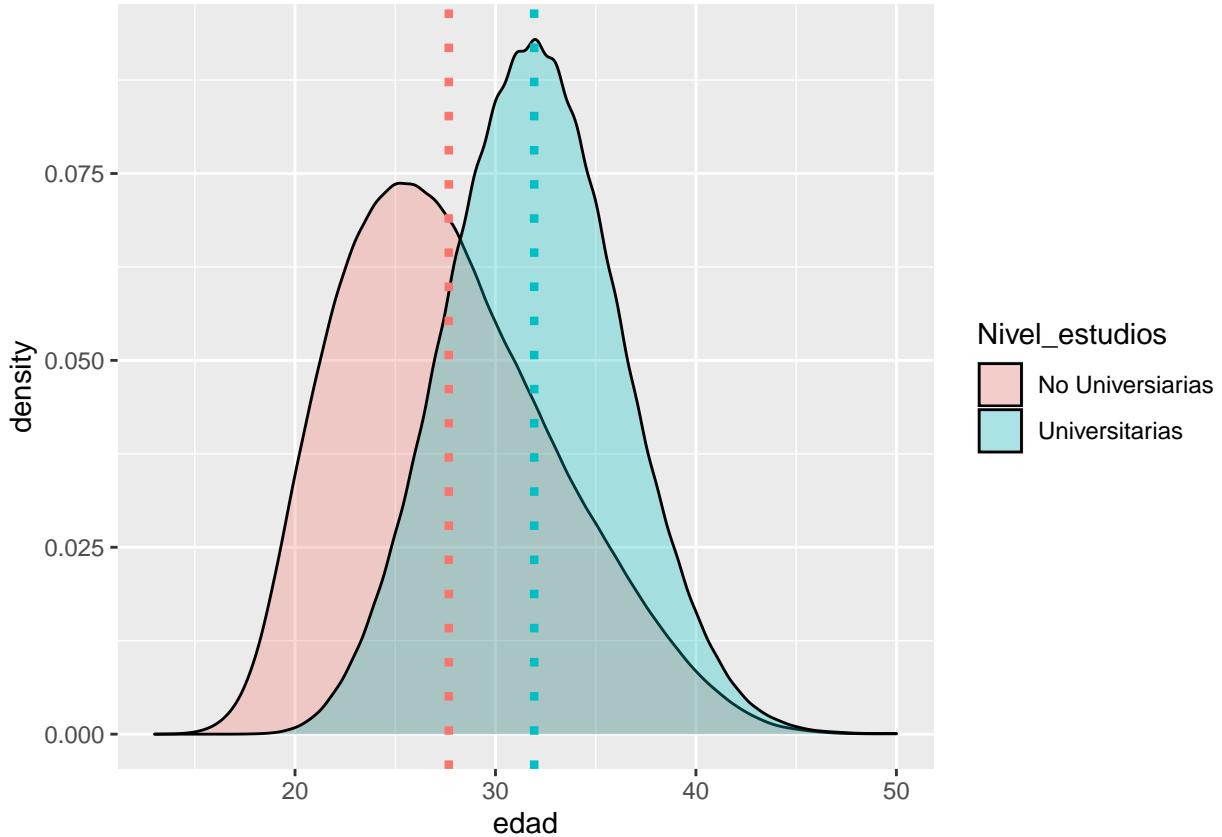
ggplot(tmp, aes(x = edad, fill = raza)) + geom_density(adjust = 2, alpha=0.3) +
  geom_vline(xintercept = 29.26, color="#00BFC4", linetype ="dotted", size=1.5) +
  geom_vline(xintercept = 31.78, color="#F8766D", linetype ="dotted", size=1.5)
```



*¿Las madres con estudios universitarios tienen en media su primer hijo con mayor edad que las que no tienen estudios superiores?* t.test(madres\_uni\$EDAD\_MADRE[madres\_uni\$PRIMER\_EMBARAZO == "Si"], madres\_no\_uni\$EDAD\_MADRE[madres\_no\_uni\$PRIMER\_EMBARAZO == "Si"], alternative = "greater", var.equal = FALSE) ‘

```
tmp <- rbind(data.frame(Nivel_estudios = "Universitarias", edad = madres_uni$EDAD_MADRE[madres_uni$PRIMER_EMBARAZO == "Si"]),
               data.frame(Nivel_estudios = "No Universitarias", edad = madres_no_uni$EDAD_MADRE[madres_no_uni$PRIMER_EMBARAZO == "Si"]))

ggplot(tmp, aes(x = edad, fill = Nivel_estudios)) + geom_density(adjust = 2, alpha=0.3) +
  geom_vline(xintercept = 31.93, color="#00BFC4", linetype ="dotted", size=1.5) +
  geom_vline(xintercept = 27.67,color="#F8766D", linetype ="dotted", size=1.5)
```



## 6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

El dataset ha sido mucho más complejo de lo esperado. Las variables no siguen una distribución normal y los grupos que hemos analizado son heterocedasticos respecto a estas variables.

Adicionalmente, nos hemos encontrado una seria dificultad al respecto de generar modelos de predicción del peso en función de las características biológicas y las circunstancias socioculturales de la madre. Realmente, más allá de una dificultad es una imposibilidad, ya que creemos que ha quedado claramente demostrado que las características biológicas no determinan/influyen en el peso de los bebés al nacer.

Como resultados finales, si hemos encontrado que la raza, en su contexto sociocultural, influye definitivamente en la edad en la que las mujeres están embarazadas y fundamentalmente cuando tienen su primer hijo, siendo la raza caucásica la que más tarda en tenerlo.

También hemos confirmado estadísticamente que las mujeres con estudios universitarios tienen su primer hijo con mayor edad que las mujeres no universitarias, siendo 4 años mayores con un 99% de confianza.

Adicionalmente y de forma sorpresiva, las mujeres caucásicas son las que en media sus hijos pesan más de todas las razas, siendo muy llamativo el ejemplo de que las mujeres asiáticas tienen en media, bebés más pesados que las mujeres afroamericanas, siendo éstas más altas y más pesadas en media que las asiáticas.

Por último, y también como factor anecdotico, hemos encontrado que las mujeres asiáticas tienen su primer hijo a una edad más avanzada que el resto de las razas.

En resumen, hemos encontrado ciertos factores a priori insospechados que influyen en las edades de las mujeres al tener hijos y hemos confirmado que los factores biológicos de las mujeres no determinan el peso de los bebés al nacer.

Contribuyentes	FIRMA
-----	-----
Investigación Previa	DP, FA
Redacción de las respuestas	DP, FA
Desarrollo de código	DP, FA
-----	-----