Limpieza y Análisis de Datos

Diciembre 2020

Contents

1 - DESCRIPCIÓN ACTIVIDAD

El objetivo de esta actividad es el tratamiento de un dataset, que puede ser el creado en la práctica 1 o bien estar disponible en Kaggle. En nuestro caso se trata de un dataset disponible en ***https://data.world/ https://data.world/pablosdt/domestic-violence-in-spain***, y trata sobre la violencia de género en España.

1.1 - OBJETIVOS

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que permita continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

1.2 - COMPETENCIAS

En esta práctica se desarrollan las siguientes competencias del Master de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

2 - RESOLUCIÓN

2.1 - DESCRIPCIÓN DEL DATASET / IMPORTANCIA

El dataset elegido contiene información sobre casos de mujeres asesinadas en España, por sus parejas o ex-parejas, entre los años 2003 y 2017. Los datos provienen de la web de estadística de la violencia de género en España http://estadisticasviolenciagenero.igualdad.mpr.gob.es/

El dataset está formado por 10 características (columnas) que presentan 900 sucesos (filas o registros):

- Year (integer)

- Month (Text)
- Autonomous Comunity (string)
- Province (string)
- Victim's age group (string)
- Agressor's age group (string)
- Partner / ex partner (string)
- Cohabitation (string)
- Existence of previous police report of gender violence (string)
- Number of victims (integer)

La información contenida en el dataset es importante, ya que proporciona datos de contexto sobre los casos de asesinatos de mujeres por parte de de sus parejas o ex-parejas.

A partir de este conjunto de datos, se plantea la problemática de determinar qué variables influyen más sobre el hecho de que ya se hubiera producido algún abuso previo, mediante el análisis de las correlaciones entre el hecho de reportes de abusos previos y otras características que definen el suceso, la realización de contrastes de hipótesis que nos proporcionen relaciones interesantes inferidas de los datos de la población, como por ejemplo la cohabitación, las edades, etc. Entendemos que el hecho de que ya se hubiera producido algún abuso previo, es un factor determinante en el hecho de que dicho abuso se convierta en agresión y que esta pueda llevar al asesinato.

Este análisis puede ser de gran relevancia, ya que podria permitir a la policía determinar a partir de informes previos, que comportamiento pueden tener casos con características y/o relaciones entre ellas similares, que todavía no han acabado en fatalidad, como por ejemplo incumplimiento de órdenes de alejamiento en el caso de que el agresor y la posible víctima no convivan, agresiones previas denunciadas o recogidas por la policía, servicios sociales etc...

2.2 - INTEGRACIÓN Y SELECCIÓN DE DATOS

Una vez definido el objetivo, creemos que las características más relevantes a considerar son:

Year

De cara a poder obtener información de progresión del número de casos en el tiempo, con el objetivo de evaluar si las políticas de prevención que se estén aplicando están dando resultado o no. No es el caso, pero esta característica podría ser determinante en un futuro, cuando se disponga de información de estos sucesos durante los periodos de confinamiento provocados por el COVID.

Month / Autonomous Comunity

Es sabido que la época del año y la localización, determinan factores ambientales, como el excesivo calor, que facilitan las reacciones violentas en determinadas personas. Otros factores desencadenantes de agresiones previas pueden ser el nivel de desempleo en la zona, desahucios, etc...

Victim's age group / Agressor's age group / Partner - Ex Partner / Cohabitation

Este grupo de características nos ayudarán también a establecer las correlaciones y pruebas de hipótesis en relación con la característica principal del estudio. Por ejemplo, en que grado determina la edad del agresor, el hecho de que haya habido algún informe de violencia previo. Los comportamientos machistas, que pueden acabar en agresiones extremas, se extienden cada vez más a poblaciones más jóvenes.

Previous abuse report

Característica principal del estudio

Por lo que las columnas del fichero que utilizaremos para nuestro estudio son:

Month, Year, Autonomous Community, Relation, Victim Age, Agressor Age, Previous Abuse Report, Living Together

Descartando la provincia y el número de víctimas, ya que entendemos que no son necesarias para el estudio.

2.3 - LIMPIEZA DE LOS DATOS

Se realiza una inspección prelminar del archivo mediante Excel, donde, de entrada, no se observan valores vacíos, ni otro tipo de información que pueda ser problemática. El archivo csv viene separado por comas.

Hacemos la carga de las librerías necesarias:

```
# Lectura de los datos

ViolenciaGenero <- read.csv("GenderViolenceSpain.csv", sep =",", header = TRUE)

head(ViolenciaGenero)
```

```
##     Month Year Autonomous.Community               Province   Relation
## 1 January 2003            Andalucía                Almería    Partner
## 2 January 2003            Andalucía                Granada    Partner
## 3 January 2003            Andalucía                 Málaga    Partner
## 4 January 2003             Canarias Santa Cruz de Tenerife    Partner
## 5 January 2003             Cataluña              Barcelona Ex-partner
## 6 January 2003             Cataluña              Barcelona    Partner
##    Victim.Age Agressor.Age Previous.Abuse.Report Living.Together Victims
## 1 41-50 years  51-64 years               Unknown             Yes       1
## 2 75-84 years  75-84 years               Unknown             Yes       1
## 3 21-30 years      Unknown               Unknown             Yes       1
## 4 31-40 years      Unknown               Unknown             Yes       1
## 5 31-40 years      Unknown               Unknown              No       1
## 6 31-40 years  51-64 years               Unknown             Yes       1
```

```
# Tipos de datos asignados a cada campo

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

```
##                 Month                  Year  Autonomous.Community
##           "character"             "integer"           "character"
##              Province              Relation            Victim.Age
##           "character"           "character"           "character"
##          Agressor.Age Previous.Abuse.Report       Living.Together
##           "character"           "character"           "character"
##               Victims
##             "integer"
```

Comprobamos que los tipos proporcinados para cada columna coinciden con los del dataset.

2.3.1 - Selección de los datos de interes

Siguiendo el criterio establecido en el apartado 2.2, vamos a eliminar del datset las columnas ***Victims*** y ***Province***:

2.3.2 - Ceros y elementos vacíos

Vamos a comprobar si tenemos ceros y/o elementos vacíos

```
# Comprobamos valores NA y nulos


sapply(ViolenciaGenero, function(x) sum(is.na(x)))
```

```
##                  Month                   Year  Autonomous.Community
##                      0                      0                      0
##               Province               Relation             Victim.Age
##                      0                      0                      0
##          Agressor.Age Previous.Abuse.Report       Living.Together
##                      0                      0                      0
##                Victims
##                      0
```

```
sapply(ViolenciaGenero, function(x) sum(is.null(x)))
```

```
##                  Month                   Year  Autonomous.Community
##                      0                      0                      0
##               Province               Relation             Victim.Age
##                      0                      0                      0
##          Agressor.Age Previous.Abuse.Report       Living.Together
##                      0                      0                      0
##                Victims
##                      0
```

2.3.3 - Identificación y tratamiento de outliers

Un outlier es una observación anormal y extrema en una muestra estadística o serie temporal de datos, que puede afectar potencialmente a la estimación de los parámetros del mismo. En nuestro caso, la única variable numérica es el año, por lo que no aplica la identificación de outliers.

Pero hemos comprobado que sólo hay un registro donde la variable Previous.Abuse.Report = 'Ex-officio'. Esto puede llevarnos a valores cero dentro de las tablas de contingencia y tests que realizaremos más tarde, por lo que podemos considerarlo como un "outlier", por lo que decidimos que es oportuno eliminar dicho registro.

2.3.4 - Exportación de los datos preprocesados

Exportamos los datos preprocesados a un fichero .csv

```
# Exportación de los datos preprocesados a un fichero .csv

write.csv(ViolenciaGenero,"GenderViolenceSpain_data_clean.csv")
```

2.4 - ANÁLISIS DE LOS DATOS

2.4.1 - Factorización y niveles de las variables cuantitativas

De cara a poder estudiar la homogeneidad y la normalidad de la varianza, vamos a factorizar y convertir a valores númericos las variables cualitativas a cuantitativas categóricas.

Convertimos en factores y vemos sus niveles

```
# Convertimos en factores y vemos sus niveles


levels(factor(ViolenciaGenero$Previous.Abuse.Report))
```

```
## [1] "No"      "Unknown" "Yes"
```

```
levels(factor(ViolenciaGenero$Autonomous.Community))
```

```
##  [1] "Andalucía"                 "Aragón"
##  [3] "Canarias"                  "Cantabria"
##  [5] "Castilla - La Mancha"      "Castilla y León"
##  [7] "Cataluña"                  "Ceuta"
##  [9] "Comunidad de Madrid"       "Comunidad Foral de Navarra"
## [11] "Comunidad Valenciana"      "Comunitat Valenciana"
## [13] "Extremadura"               "Galicia"
## [15] "Illes Balears"             "Islas Baleares"
## [17] "La Rioja"                  "Melilla"
## [19] "País Vasco"                "Principado de Asturias"
## [21] "Región de Murcia"
```

```
levels(factor(ViolenciaGenero$Month))
```

```
##  [1] "April"     "August"    "December"  "February"  "January"   "July"
##  [7] "June"      "March"     "May"       "November"  "October"   "September"
```

```
levels(factor(ViolenciaGenero$Relation))
```

```
## [1] "Ex-partner"            "In separation process" "Partner"
```

```
levels(factor(ViolenciaGenero$Victim.Age))
```

```
##  [1] "<16 years"   ">85 years"   "16-17 years" "18-20 years" "21-30 years"
##  [6] "31-40 years" "41-50 years" "51-64 years" "65-74 years" "75-84 years"
## [11] "Unknown"
```

```
levels(factor(ViolenciaGenero$Agressor.Age))
```

```
##  [1] ">85 years"   "16-17 years" "18-20 years" "21-30 years" "31-40 years"
##  [6] "41-50 years" "51-64 years" "65-74 years" "75-84 years" "Unknown"
```

```r
levels(factor(ViolenciaGenero$Living.Together))
```

```
## [1] "No"      "Unknown" "Yes"
```

```r
levels(factor(ViolenciaGenero$Province))
```

```
##  [1] "A Coruña"              "Álava"                 "Albacete"
##  [4] "Alicante"              "Alicante/Alacant"      "Almería"
##  [7] "Asturias"              "Ávila"                 "Badajoz"
## [10] "Barcelona"             "Bizkaia"               "Burgos"
## [13] "Cáceres"               "Cádiz"                 "Cantabria"
## [16] "Castellón"             "Castellón/Castelló"    "Ceuta"
## [19] "Ciudad Real"           "Córdoba"               "Cuenca"
## [22] "Girona"                "Granada"               "Guadalajara"
## [25] "Guipúzcoa"             "Huelva"                "Huesca"
## [28] "Illes Balears"         "Islas Baleares"        "Jaén"
## [31] "La Rioja"              "Las Palmas"            "León"
## [34] "Lleida"                "Lugo"                  "Madrid"
## [37] "Málaga"                "Melilla"               "Murcia"
## [40] "Navarra"               "Ourense"               "Palencia"
## [43] "Pontevedra"            "Salamanca"             "Santa Cruz de Tenerife"
## [46] "Segovia"               "Sevilla"               "Soria"
## [49] "Tarragona"             "Teruel"                "Toledo"
## [52] "Valencia"              "Valencia/València"     "Valladolid"
## [55] "Vizcaya"               "Zamora"                "Zaragoza"
```

Factorizamos los valores para cada columna

```r
ViolenciaGenero$Previous.Abuse.Report <- as.numeric(fct_rev(factor(ViolenciaGenero$Previous.Abuse.Repor

ViolenciaGenero$Autonomous.Community <- as.numeric(factor(ViolenciaGenero$Autonomous.Community))

ViolenciaGenero$Month <- as.numeric(factor(ViolenciaGenero$Month))

ViolenciaGenero$Relation <- as.numeric(factor(ViolenciaGenero$Relation))

ViolenciaGenero$Victim.Age <- as.numeric(factor(ViolenciaGenero$Victim.Age))

ViolenciaGenero$Agressor.Age <- as.numeric(factor(ViolenciaGenero$Agressor.Age))

ViolenciaGenero$Living.Together <- as.numeric(factor(ViolenciaGenero$Living.Together))

ViolenciaGenero$Province <- as.numeric(factor(ViolenciaGenero$Province))
```

```r
# Exportación de los datos procesados a un fichero .csv

write.csv(ViolenciaGenero,"GenderViolenceSpain_TFM.csv")
```

2.4.2 - Selección de grupos de datos

Seleccionamos un conjunto inicial de grupos de datos que nos pueden resultar interesantes de analizar y/o comparar.

Los valores de la variable Previous.Abuse.Report han quedado así:

1: Yes 2: Unknown 3: No

Agrupación por Comunidad Autónoma con temperaturas muy elevadas en verano

Comunidades: Andalucía (1), Aragón (2), Región de Murcia (21), Castilla - La Mancha (5)

Meses: Junio (7), Julio (6), Agosto (2)

```
head(ViolenciaGenero.Televada.Verano)
```

```
##   Month Year Autonomous.Community Province Relation Victim.Age Agressor.Age
## 1     7 2003                    2       57        3          6            10
## 2     7 2003                   21       39        3          6            10
## 3     6 2003                    5       51        3          5            10
## 4     7 2004                    1       23        3          7            10
## 5     7 2004                    1       23        3          8             8
## 6     6 2004                    1        6        1          6             6
##   Previous.Abuse.Report Living.Together Victims
## 1                     2               3       1
## 2                     2               3       1
## 3                     2               3       1
## 4                     2               3       1
## 5                     2               3       1
## 6                     2               1       1
```

Agrupación por agresores jóvenes (16-17 years (2), 18-20 years(3))

```
head(ViolenciaGenero.Agresores.Jovenes)
```

```
##   Month Year Autonomous.Community Province Relation Victim.Age Agressor.Age
## 1     9 2004                    7       49        3          1             3
## 2     6 2004                   21       39        3          1             3
## 3    10 2004                   19       55        1         11             2
## 4     4 2006                    7       10        1          5             3
## 5     2 2006                    1       30        1          4             3
## 6     3 2006                    1       30        3          5             3
##   Previous.Abuse.Report Living.Together Victims
## 1                     2               1       1
## 2                     2               1       1
## 3                     2               1       1
## 4                     3               1       1
## 5                     3               1       1
## 6                     3               3       1
```

Agrupación por agresores adultos (21-30 years (4), 31-40 years (5), 41-50 years (6), 51-64 years (7))

```
head(ViolenciaGenero.Agresores.Adultos)
```

```
##   Month Year Autonomous.Community Province Relation Victim.Age Agressor.Age
## 1     5 2003                    1        6        3          7             7
## 2     5 2003                    7       10        3          6             7
```

```
## 3     5 2003                    7       10        3          11             5
## 4     5 2003                   17       31        3           6             5
## 5     4 2003                   17       31        1           5             5
## 6     8 2003                    1       14        3           6             7
##   Previous.Abuse.Report Living.Together Victims
## 1                     2               3       1
## 2                     2               3       1
## 3                     2               3       1
## 4                     2               3       1
## 5                     2               1       1
## 6                     2               3       1
```

Agrupación por agresores mayores (65-74 (8), 74-85 (9), > 85 years (1))

```
head(ViolenciaGenero.Agresores.Mayores)
```

```
##   Month Year Autonomous.Community Province Relation Victim.Age Agressor.Age
## 1     5 2003                    1       23        3         10            9
## 2     4 2003                    1       37        3         10            8
## 3     4 2003                   14        1        3          9            8
## 4     8 2003                    9       36        3          9            8
## 5     7 2003                    7       34        3          9            8
## 6    12 2003                    1       30        3          2            1
##   Previous.Abuse.Report Living.Together Victims
## 1                     2               3       1
## 2                     2               3       1
## 3                     2               3       1
## 4                     2               3       1
## 5                     2               3       1
## 6                     2               3       1
```

Agrupación por Convivientes = Yes (3)

```
head(ViolenciaGenero.Convivientes.Yes)
```

```
##   Month Year Autonomous.Community Province Relation Victim.Age Agressor.Age
## 1     5 2003                    1        6        3          7            7
## 2     5 2003                    1       23        3         10            9
## 3     5 2003                    1       37        3          5           10
## 4     5 2003                    3       45        3          6           10
## 5     5 2003                    7       10        3          6            7
## 6     5 2003                    7       10        3         11            5
##   Previous.Abuse.Report Living.Together Victims
## 1                     2               3       1
## 2                     2               3       1
## 3                     2               3       1
## 4                     2               3       1
## 5                     2               3       1
## 6                     2               3       1
```

Agrupación por Relation = Yes (3)

```r
head(ViolenciaGenero.Relation.Expartner)
```

```
##   Month Year Autonomous.Community Province Relation Victim.Age Agressor.Age
## 1     5 2003                    7       10        1          6           10
## 2     4 2003                   17       31        1          5            5
## 3     8 2003                    1       30        1          5           10
## 4     8 2003                    1       30        1          6           10
## 5     9 2003                    7       49        1          7           10
## 6     7 2003                    6       44        1          7           10
##   Previous.Abuse.Report Living.Together Victims
## 1                     2               1       1
## 2                     2               1       1
## 3                     2               1       1
## 4                     2               1       1
## 5                     2               1       1
## 6                     2               1       1
```

2.4.3 - Comprobación de homogeneidad y normalidad de la varianza

Para la comprobación de que los valores que toman nuestras variables cuantitativas provienen de una población distribuida normalmente, utilizaremos la prueba de normalidad deShapiro. Se comprueba si el p-valor es superior al nivel de significación prefijado alfa = 0.05. Si esto se cumple, entonces se considera que variable en cuestión sigue una distribución normal.

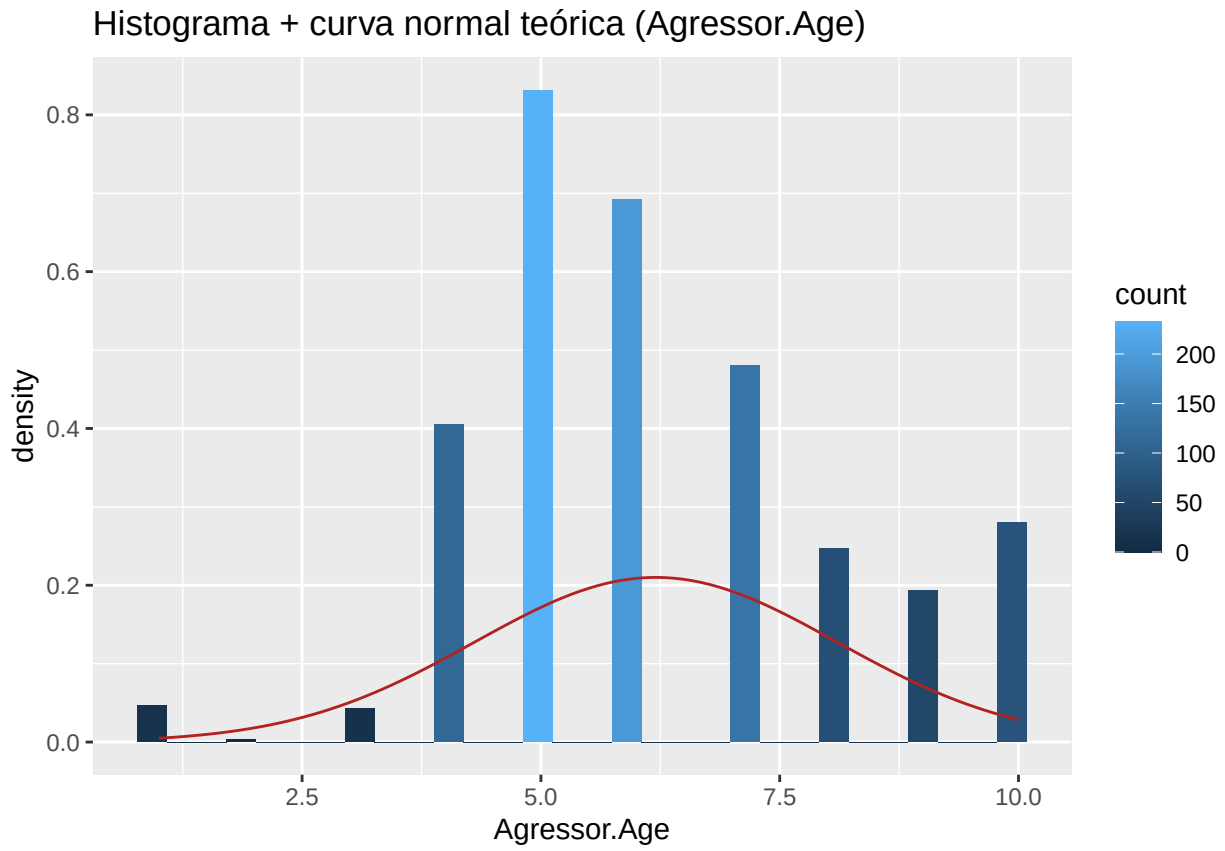
```r
alpha = 0.05

col.names = colnames(ViolenciaGenero)

for (i in 1:ncol(ViolenciaGenero))
  {
    if (i == 1) cat("Variables que no siguen una distribución normal:\n")
      if (is.integer(ViolenciaGenero[,i]) | is.numeric(ViolenciaGenero[,i]))
        {
          p_val = shapiro.test(ViolenciaGenero[,i])$p.value
          if (p_val < alpha) {
                                cat(col.names[i])
                                if (i < ncol(ViolenciaGenero) - 1) cat(", ")
                                if (i %% 2 == 0) cat("\n")
                              }
        }
  }
```

```
## Variables que no siguen una distribución normal:
## Month, Year,
## Autonomous.Community, Province,
## Relation, Victim.Age,
## Agressor.Age, Previous.Abuse.Report,
## Living.TogetherVictims
```

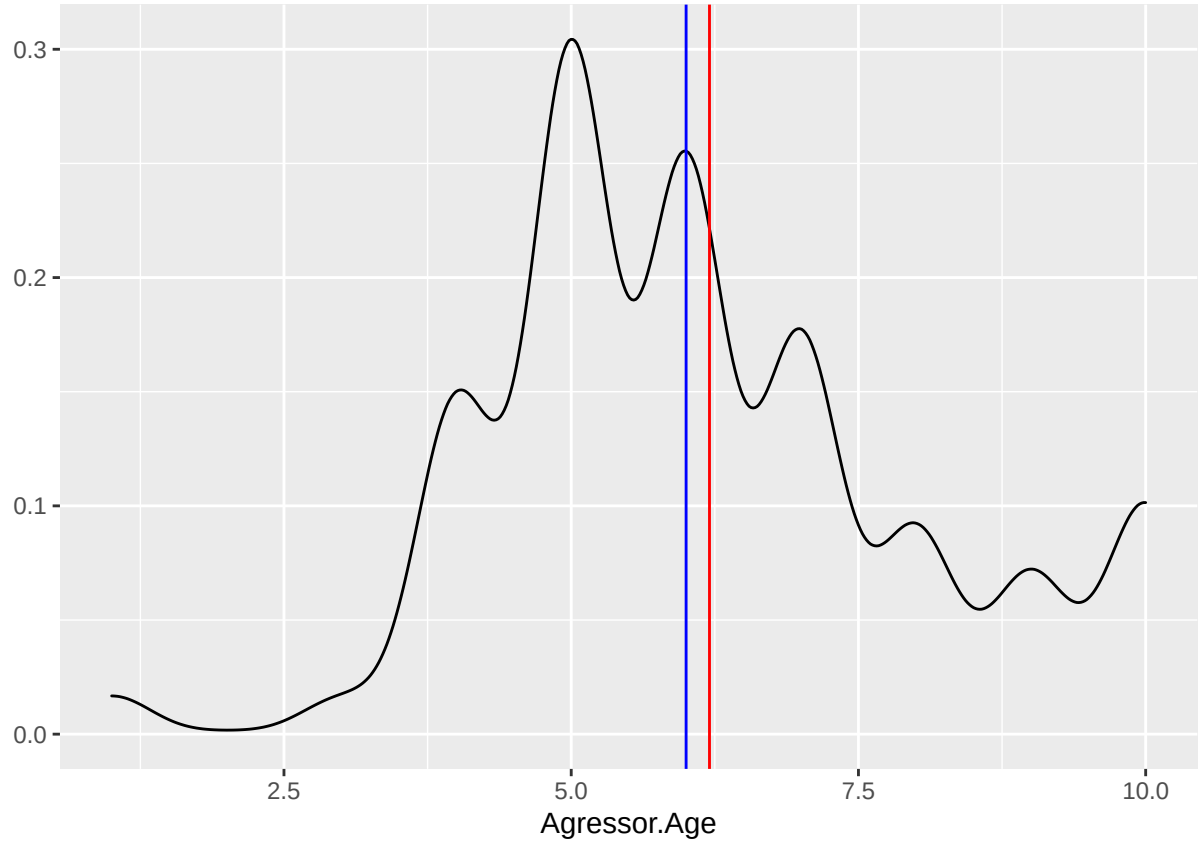
Podemos comprobarlo gráficamente por ejemplo con la variable Agessor.Age.

```
ggplot(data = ViolenciaGenero, aes(x = Agressor.Age)) +
  geom_histogram(aes(y = ..density..,fill = ..count..)) +
  stat_function(fun = dnorm, colour = "firebrick",
                args = list(mean = mean(ViolenciaGenero$Agressor.Age),
                            sd = sd(ViolenciaGenero$Agressor.Age))) +
  ggtitle("Histograma + curva normal teórica (Agressor.Age)")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
qplot(Agressor.Age, data = ViolenciaGenero,geom="density")+ geom_vline(xintercept =
mean(ViolenciaGenero$Agressor.Age),color="red")+ geom_vline(xintercept =
median(ViolenciaGenero$Agressor.Age), color="blue")
```



Vemos que ninguna variable es Normal

Seguidamente, pasamos a estudiar la homogeneidad de varianzas.

Como ninguna de las variables es normal aplicaremos el test de Levene entre Victims y las variables que vamos a utilizar. Este test prueba la hipótesis nula de que las varianzas poblacionales son iguales.

```
leveneTest(Victims ~ factor(ViolenciaGenero$Month), ViolenciaGenero, center=median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group  11   1.029 0.4181
##       887
```

```
leveneTest(Victims ~ factor(ViolenciaGenero$Year), ViolenciaGenero , center=median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group  14  1.0172 0.4332
##       884
```

```
leveneTest(Victims ~ factor(ViolenciaGenero$Autonomous.Community), ViolenciaGenero, center=median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group  20  0.3649 0.9954
##       878
```

```r
leveneTest(Victims ~ factor(ViolenciaGenero$Relation), ViolenciaGenero, center=median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2  0.1716 0.8424
##       896
```

```r
leveneTest(Victims ~ factor(ViolenciaGenero$Victim.Age), ViolenciaGenero, center=median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group  10  0.2589 0.9894
##       888
```

```r
leveneTest(Victims ~ factor(ViolenciaGenero$Agressor.Age), ViolenciaGenero, center=median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   9  0.3173 0.9695
##       889
```

```r
leveneTest(Victims ~ factor(ViolenciaGenero$Previous.Abuse.Report), ViolenciaGenero, center=median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2   0.374 0.6881
##       896
```

```r
leveneTest(Victims ~ factor(ViolenciaGenero$Living.Together), ViolenciaGenero, center=median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2  0.2719  0.762
##       896
```

```r
leveneTest(Victims ~ factor(ViolenciaGenero$Province), ViolenciaGenero, center=median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group  56  0.3098      1
##       842
```

Si el P-valor resultante de la prueba de Levene es inferior a un cierto nivel de significación (típicamente 0.05), es poco probable que las diferencias obtenidas en las variaciones de la muestra se hayan producido sobre la base de un muestreo aleatorio de una población con varianzas iguales. Por lo tanto, la hipótesis nula de igualdad de varianzas se rechaza y se concluye que hay una diferencia entre las variaciones en la población.

Los resultados indican que **no hay diferencias significativas** entre las varianzas de los grupos, es decir existe homogeneidad de varianza u homocedasticidad.

2.4.4 - Tablas de Contingencia

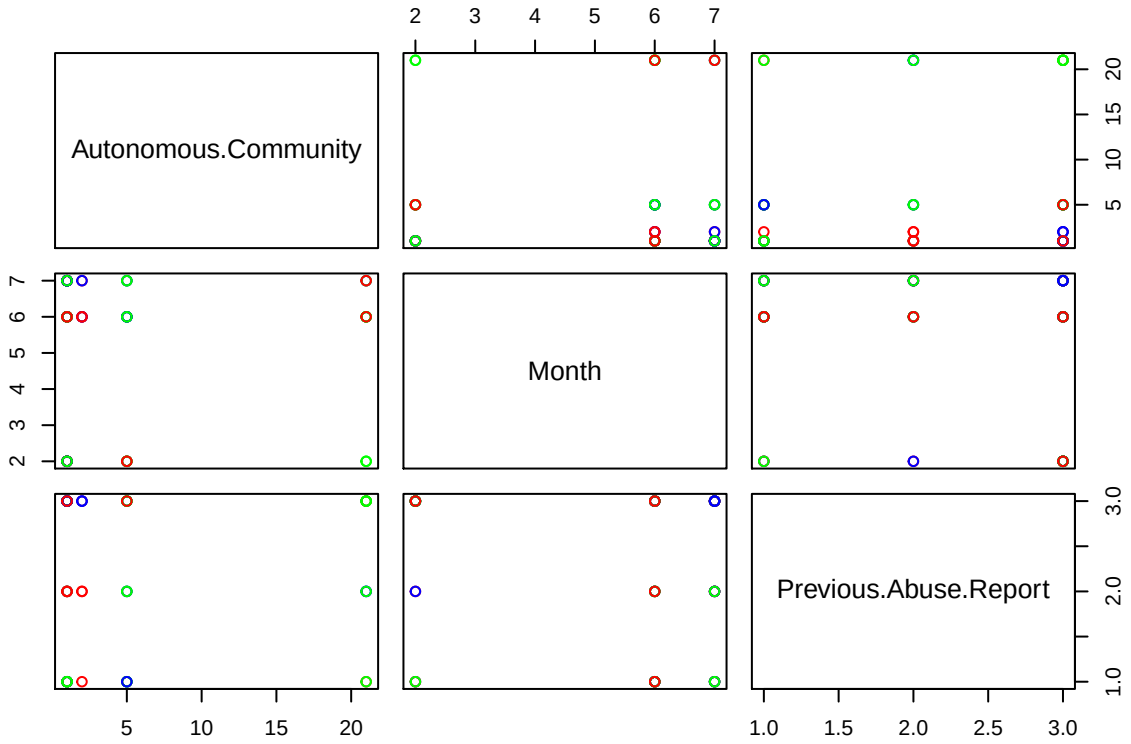
En nuestro caso, todas las variables que vamos a utilizar son categóricas, por lo que el análisis de sus relaciones se ha de obtener mediante Tablas de contingencia y pruebas Chi-Cuadrado

Tablas de Contingencia

```
TablaCAMonth.PAR <- ftable(ViolenciaGenero.Televada.Verano[, c("Autonomous.Community", "Month", "Previou

TablaCAMonth.PAR
```

```
##                             Previous.Abuse.Report 1 2 3
## Autonomous.Community Month
## 1                    2                            2 1 7
##                      6                            4 2 7
##                      7                            3 2 9
## 2                    2                            0 0 0
##                      6                            1 1 2
##                      7                            0 1 1
## 5                    2                            1 1 3
##                      6                            3 1 2
##                      7                            1 1 1
## 21                   2                            0 0 2
##                      6                            1 2 2
##                      7                            1 1 1
```

```
 #plot(TablaCAMonth.PAR, col = c("red", "blue", "green"), main = "Autonomous Community vs. Previous abu

plot(ViolenciaGenero.Televada.Verano[, c("Autonomous.Community", "Month", "Previous.Abuse.Report")],col
```

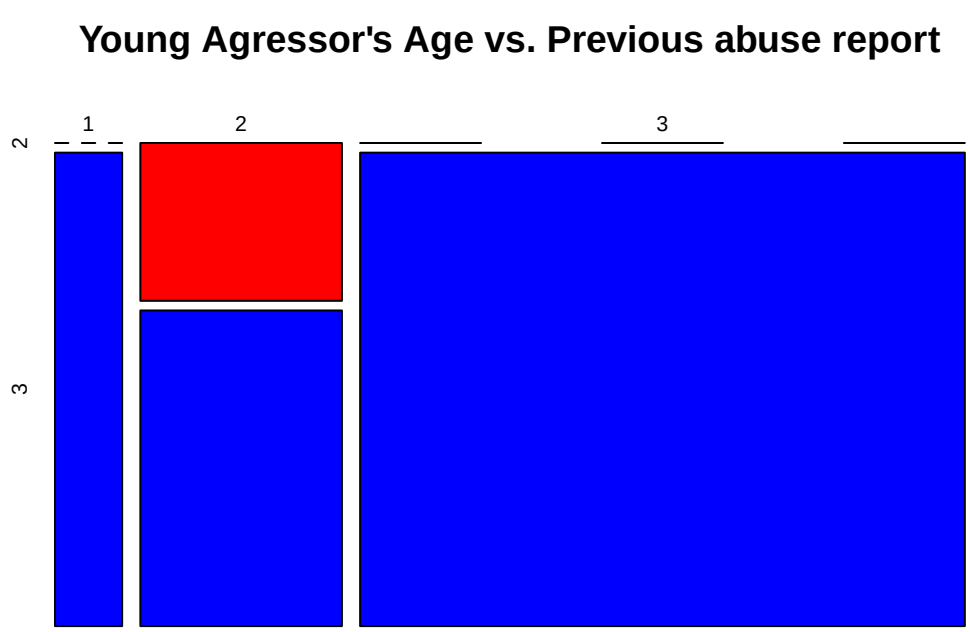


Podemos comprobar que donde más asesinatos se producen, es en la Comunidad Andaluza (1) en los meses de Junio (7) y Julio (6), sin que haya constancia de abusos previos, seguida por Castilla La Mancha (5) en el més de Julio (6), también sin constancia de abusos previos.

```
TablaJovenes.PAR <- table(ViolenciaGenero.Agresores.Jovenes$Previous.Abuse.Report,ViolenciaGenero.Agres

#TablaJovenes.PAR <- ViolenciaGenero[, c("Previous.Abuse.Report","Agressor.Age")]
#TablaJovenes.PAR <- df(ViolenciaGenero.Agresores.Jovenes$Previous.Abuse.Report,ViolenciaGenero.Agresor

TablaJovenes.PAR
```

```
##
##     2 3
##   1 0 1
##   2 1 2
##   3 0 9
```

```
#plot(TablaJovenes.PAR, col = c("red", "blue","green"), main = "Young Agressor's Age vs. Previous abuse
#axis(1, at=c(1,2,3), labels = c("Yes","Unknown","No"))

#hist(TablaJovenes.PAR)

plot(TablaJovenes.PAR, col = c("red", "blue", "green", "cyan"), main = "Young Agressor's Age vs. Previou
```

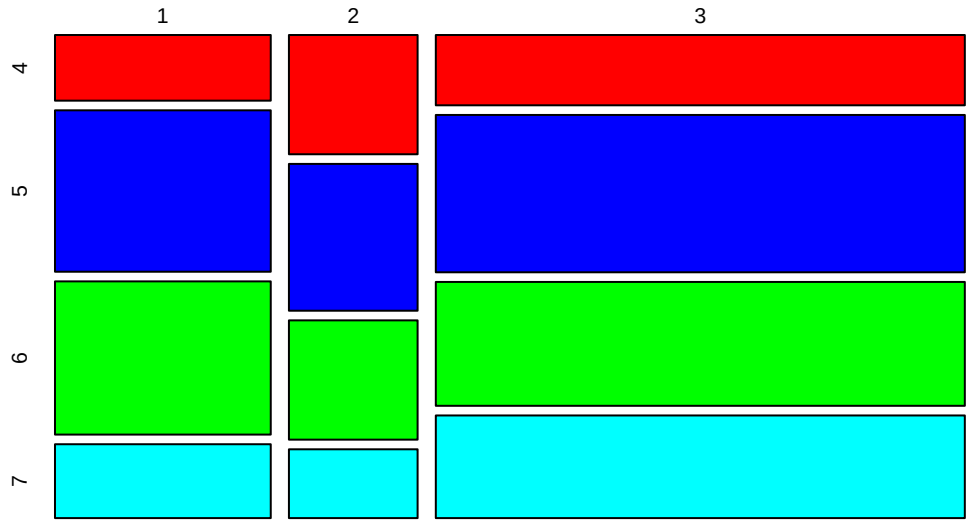


Podemos comprobar que el mayor número de asesinatos en jóvenes se producen en la franja de 18-20 years(3) y sin constancia de abusos previos

```
TablaAdultos.PAR <- table(ViolenciaGenero.Agresores.Adultos$Previous.Abuse.Report,ViolenciaGenero.Agres

TablaAdultos.PAR
```

```
##
##       4   5   6   7
##   1  24  59  56  27
##   2  26  32  26  15
##   3  63 141 111  92
```

```
plot(TablaAdultos.PAR, col = c("red", "blue", "green", "cyan"), main = "Adult Agressor's Age vs. Previou
```

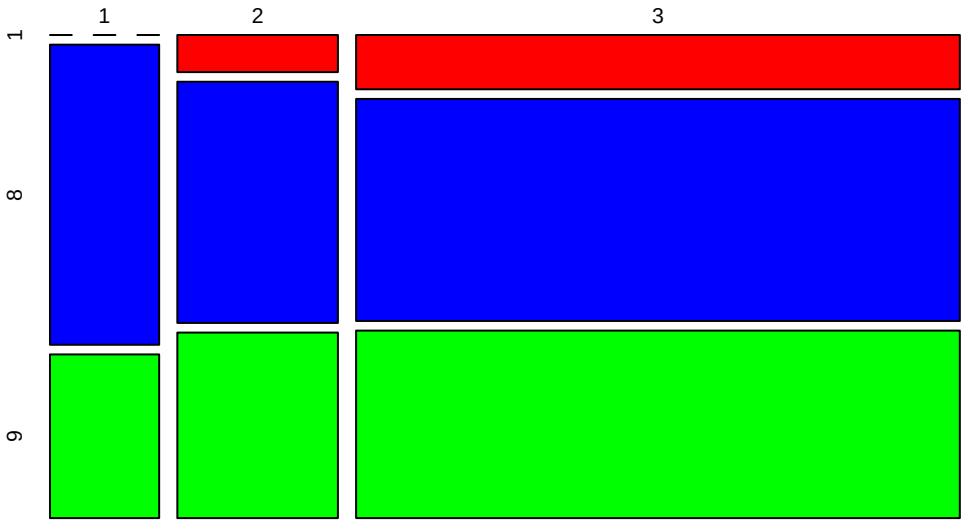


Podemos comprobar que el mayor número de asesinatos en adultos se producen en las franjas 31-40 years (5), 41-50 years (6)

```
TablaMayores.PAR <- table(ViolenciaGenero.Agresores.Mayores$Previous.Abuse.Report,ViolenciaGenero.Agres

TablaMayores.PAR
```

```
##
##      1  8  9
##   1  0 11  6
##   2  2 13 10
##   3 11 45 38
```

```
plot(TablaMayores.PAR, col = c("red", "blue","green"), main = "Old Agressor's Age vs. Previous abuse re
```

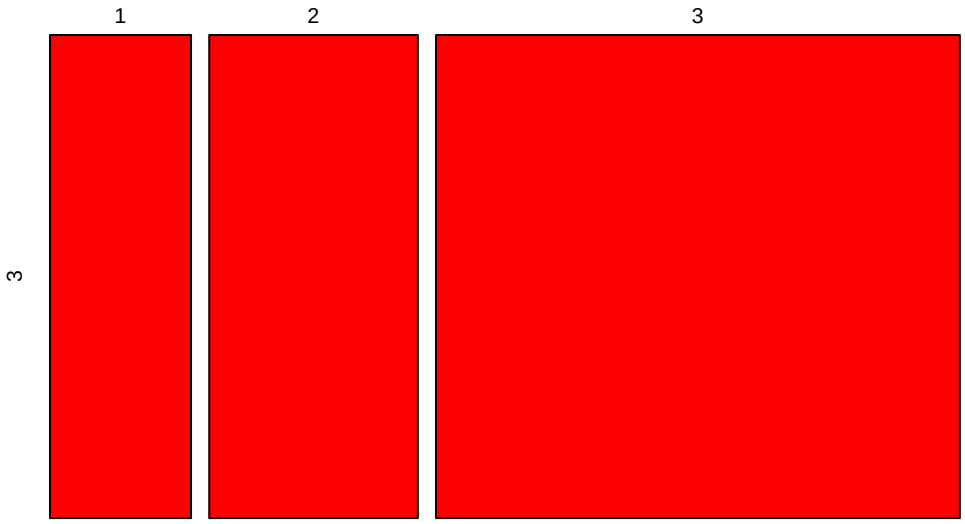


Podemos comprobar que el mayor número de asesinatos se producen en todas las franjas de edades, especialmente en las franjas 65-74 (8) y 74-85 (9) , y sin constancia de abusos previos

```
TablaConvivientes.PAR <- table(ViolenciaGenero.Convivientes.Yes$Previous.Abuse.Report,ViolenciaGenero.C

TablaConvivientes.PAR
```

```
##
##       3
##   1  94
##   2 139
##   3 349
```

```
plot(TablaConvivientes.PAR, col = c("red"), main = "Living.Together.Yes vs. Previous abuse report")
```

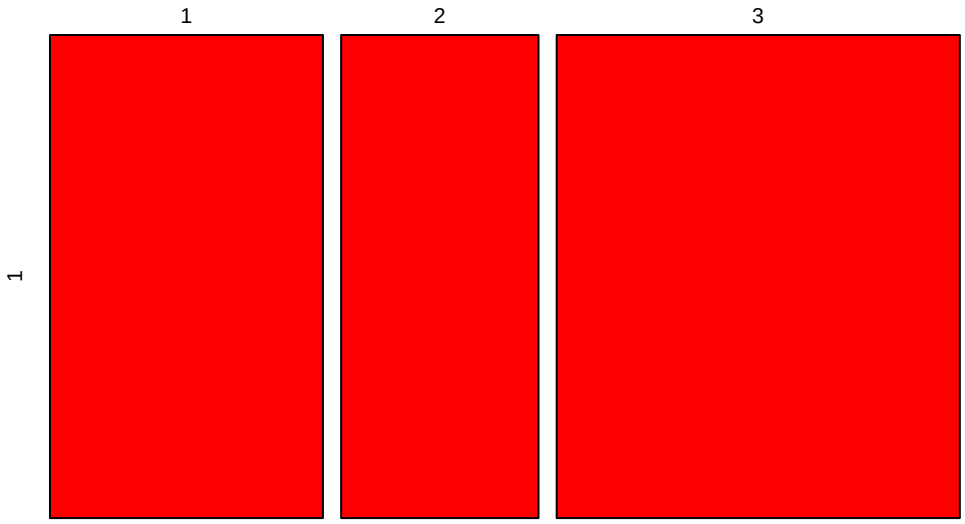


Podemos comprobar que el mayor número de asesinatos entre convivientes Yes (3), se producen sin constancia de abusos previos

```
TablaRelation.PAR <- table(ViolenciaGenero.Relation.Expartner$Previous.Abuse.Report,ViolenciaGenero.Rela

TablaRelation.PAR
```

```
##
##      1
##   1 65
##   2 47
##   3 96
```

```
 plot(TablaRelation.PAR, col = c("red"), main = "Relation.ExtPartner vs. Previous abuse report")
```



2.4.5 - Aplicación de pruebas estadísticas

2.4.5.1 - Estudio de la Correlación / Tests Chi-Squared

Estamos tratando variables cuantitativas politómicas y nominales, por lo que el test Chi-squared resulta adecuado en algunos casos, y el test exacto de Fisher en otros para valorar la independencia.

La Edad de los agresores la consideramos nominal, ya que no las vamos a utilizar estableciendo relaciones de tipo mayor/menor, ni vamos a evaluar distancias entre los diferentes rangos de edades.

```
#chisq.test(TablaCAMonth.PAR)

#Aplicamos Fisher, ya que al utilizar Chi-Square daba el error: "Chi-squared approximation may be incor
#El problema es que la aproximación de Chi-cuadrado a la distribución del estadístico de prueba se basa

fisher.test(TablaCAMonth.PAR, hybrid = TRUE,conf.level = 0.95, simulate.p.value = TRUE)
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
##
## data:  TablaCAMonth.PAR
## p-value = 0.9855
## alternative hypothesis: two.sided
```

Como el p-value es > 0.05 no podemos rechazar la hipotésis nula que indica independencia entre las variables (casi total). Por lo tanto no existe correlación entre ellas.

```
fisher.test(TablaJovenes.PAR, conf.level = 0.95, simulate.p.value = FALSE)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  TablaJovenes.PAR
## p-value = 0.3077
## alternative hypothesis: two.sided
```

Como el p-value es > 0.05 no podemos rechazar la hipotésis nula que indica independencia entre ambas variables. Por lo tanto no existe correlación entre ellas.

```
chisq.test(TablaAdultos.PAR)
```

```
##
##  Pearson's Chi-squared test
##
## data:  TablaAdultos.PAR
## X-squared = 12.096, df = 6, p-value = 0.05986
```

Como el p-value es > 0.05 no podemos rechazar la hipotésis nula que indica independencia entre ambas variables. Por lo tanto no existe correlación entre ellas.

```
fisher.test(TablaMayores.PAR, conf.level = 0.95, simulate.p.value = FALSE)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  TablaMayores.PAR
## p-value = 0.6504
## alternative hypothesis: two.sided
```

Como el p-value es > 0.05 no podemos rechazar la hipotésis nula que indica independencia entre ambas variables. Por lo tanto no existe correlación entre ellas.

```
chisq.test(TablaConvivientes.PAR)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  TablaConvivientes.PAR
## X-squared = 190.98, df = 2, p-value < 2.2e-16
```

Como el p-value es $<= 0.05$ podemos rechazar al hipotésis nula que indica independencia entre ambas variables. Por lo tanto existe correlación entre ellas

```r
chisq.test(TablaRelation.PAR)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  TablaRelation.PAR
## X-squared = 17.721, df = 2, p-value = 0.0001419
```

Como el p-value es <= 0.05 podemos rechazar al hipotésis nula que indica independencia entre ambas variables. Por lo tanto existe correlación entre ellas

2.4.5.3 - Regresión Logística (Multinomial)

En este caso tenemos una variable dependiente Previous.Abuse.Report de caracter politómica y nominal.

Vamos a plantear modelos de regresión logística multinomial que nos permitan trabajar con los grupos de datos construídos anteriormente donde hayamos encontrado correlación entre las variables que los componen.

Calcularemos Odds Ratios e Intervalos de Confianza:

```r
# Calculamos un modelo relativo a los datos  Violencia.Convivientes.Yes


model.vgenero.convivientes.yes = multinom(Previous.Abuse.Report ~ Agressor.Age, data = ViolenciaGenero.
```

```
## # weights:  9 (4 variable)
## initial  value 639.392352
## iter  10 value 518.018643
## iter  10 value 518.018643
## final  value 518.018643
## converged
```

```r
# Obtenemos el summary

summary(model.vgenero.convivientes.yes)
```

```
## Call:
## multinom(formula = Previous.Abuse.Report ~ Agressor.Age, data = ViolenciaGenero.Convivientes.Yes)
##
## Coefficients:
##   (Intercept) Agressor.Age
## 2  -2.8302764    0.4812557
## 3   0.8063664    0.0840625
##
## Std. Errors:
##   (Intercept) Agressor.Age
## 2   0.5246014   0.07701679
## 3   0.3954953   0.06371025
##
## Residual Deviance: 1036.037
## AIC: 1044.037
```

La línea de coeficientes que comienza con 2 hace referencia al modelo comparando la probabilidad de que no sepamos nada sobre el informe previo, respecto a que si lo haya La línea de coeficientes que comienza con 3 hace referencia al modelo comparando la probabilidad de que no haya informe previo, respecto a que si lo haya

Vamos a evaluar ahora los ***odds ratio***. Los ***odds*** es la razón de la probabilidad de ocurrencia de un suceso entre la probabilidad de su no ocurrencia.Vamos a ver como transformamos los coeficientes en odds ratios. En este primer modelo vamos a tratar de ser algo didáctivos y vamos a explicar en detalle su cálculo.

En esta expresión, el modelo está expresado en términos del ***log-odds*** para el modelo (2):

$$ln(\frac{P(Y=1/X)}{1-P(Y=1/X)}) = -2.83 + 0.481 * Agressor.Age$$

Si se escribe en términos de odds, se tiene:

$$\frac{P(Y=1/X)}{1-P(Y=1/X)} = \frac{e^{b_0} + \sum_{i=1}^{n}(b_i x_i)}{1 + e^{b_0} + \sum_{i=1}^{n}(b_i x_i)}$$

Se calculan los distintos valores de las probabilidades para las cuatro combinaciones entre la variable dependiente Y con la independiente X:

$$\frac{P(Y=1/X=1)}{1-P(Y=1/X=1)} = \frac{e^{b_0+b_1}}{1+e^{b_0+b_1}}$$

$$\frac{P(Y=1/X=0)}{1-P(Y=1/X=0)} = \frac{e^{b_0}}{1+e^{b_0}}$$

$$\frac{P(Y=0/X=1)}{1-P(Y=0/X=1)} = \frac{1}{1+e^{b_0+b_1}}$$

$$\frac{P(Y=0/X=0)}{1-P(Y=0/X=0)} = \frac{1}{1+e^{b_0}}$$

Los ***odds-ratio (OR)*** se calculan como la razón entre los ***odds***, donde la variable respuesta Y está presente entre los individuos, es decir, toma el valor Y = 1, y la variable independiente X puede estar presente o no, es decir, tomar los valores X = 1 y X = 0.

$$OR = \frac{\frac{P(Y=1/X=1)}{1-P(Y=1/X=1)}}{\frac{P(Y=1/X=0)}{1-P(Y=1/X=0)}} = e^{b_1}$$

• Un OR = 1 implica que no existe asociación entre la variable respuesta y la covariable. • Un OR inferior a la unidad se interpreta como un factor de protección, es decir, el suceso es menos probable en presencia de dicha covariable. • Un OR mayor a la unidad se interpreta como un factor de riesgo, es decir,el suceso es más probable en presencia de dicha covariable.

$$ln(\frac{P(Y=1/X)}{1-P(Y=1/X)}) = 0.806 + 0.084 * Agressor.Age$$

```
# Coefcientes Modelo

coefmodel.vgen.conv.yes <- coef(model.vgenero.convivientes.yes)

coefmodel.vgen.conv.yes
```

```
##   (Intercept) Agressor.Age
## 2  -2.8302764    0.4812557
## 3   0.8063664    0.0840625
```

```
# Odds Ratios Modelo



exp(coefmodel.vgen.conv.yes)
```

```
##   (Intercept) Agressor.Age
## 2  0.05899654     1.618105
## 3  2.23975489     1.087697
```

```
# Intervalos de confianza odds ratio


Modelo.vgenero.conv.yes.IC <- confint(model.vgenero.convivientes.yes)

Modelo.vgenero.conv.yes.IC
```

```
## , , 2
##
##                   2.5 %     97.5 %
## (Intercept)  -3.8584763 -1.8020766
## Agressor.Age  0.3303056  0.6322058
##
## , , 3
##
##                    2.5 %    97.5 %
## (Intercept)   0.03120986 1.5815230
## Agressor.Age -0.04080731 0.2089323
```

```
# Calculamos un modelo relativo a los datos  Violencia.Relation.Expartner

model.vgenero.relation.expartner = multinom(Previous.Abuse.Report ~ Relation, data = ViolenciaGenero.Rel
```

```
## # weights:  9 (4 variable)
## initial  value 228.511356
## final  value 219.738384
## converged
```

```
# Obtenemos el summary

summary(model.vgenero.relation.expartner)
```

```
## Call:
## multinom(formula = Previous.Abuse.Report ~ Relation, data = ViolenciaGenero.Relation.Expartner)
##
## Coefficients:
##   (Intercept)   Relation
```

```
## 2  -0.1621210 -0.1621210
## 3   0.1949799  0.1949799
##
## Std. Errors:
##   (Intercept)   Relation
## 2  0.09573560 0.09573560
## 3  0.08031386 0.08031386
##
## Residual Deviance: 439.4768
## AIC: 443.4768
```

En este caso tenemos las siguientes ecuaciones:

$$ln(\frac{P(Y=1/X)}{1-P(Y=1/X)}) = -0.162 - 0.162 * Relation$$

$$ln(\frac{P(Y=1/X)}{1-P(Y=1/X)}) = 0.194 - 0.194 * Relation$$

```
# Coefcientes Modelo

coefmodel.vgen.rel.expartner <- coef(model.vgenero.relation.expartner)

coefmodel.vgen.rel.expartner
```

```
##   (Intercept)   Relation
## 2  -0.1621210 -0.1621210
## 3   0.1949799  0.1949799
```

```
# Odds Ratios Modelo


exp(coefmodel.vgen.rel.expartner )
```

```
##   (Intercept)  Relation
## 2   0.8503383 0.8503383
## 3   1.2152865 1.2152865
```

```
# Intervalos de confianza odds ratio


Modelo.vgen.rel.expartner.IC <- confint(model.vgenero.relation.expartner)

Modelo.vgen.rel.expartner.IC
```

```
## , , 2
##
##                   2.5 %     97.5 %
## (Intercept) -0.3497593 0.02551732
## Relation    -0.3497593 0.02551732
##
## , , 3
```

```
##
##                   2.5 %    97.5 %
## (Intercept) 0.03756759 0.3523921
## Relation    0.03756759 0.3523921
```

Vamos a considerar también el modelo relativo a agresores adultos, ya que su p-value estaba al límite de descartar la hipótesis nula

```
# Calculamos un modelo relativo a los datos  Violencia.Agresores.Adultos


model.vgenero.agresores.adultos = multinom(Previous.Abuse.Report ~ Agressor.Age, data = ViolenciaGenero
```

```
## # weights:  9 (4 variable)
## initial  value 738.267458
## final  value 622.894505
## converged
```

```
# Obtenemos el summary

summary(model.vgenero.agresores.adultos)
```

```
## Call:
## multinom(formula = Previous.Abuse.Report ~ Agressor.Age, data = ViolenciaGenero.Agresores.Adultos)
##
## Coefficients:
##   (Intercept) Agressor.Age
## 2   0.6853508  -0.22222470
## 3   0.6033503   0.05293524
##
## Std. Errors:
##   (Intercept) Agressor.Age
## 2   0.7108247   0.12993230
## 3   0.5227703   0.09300185
##
## Residual Deviance: 1245.789
## AIC: 1253.789
```

En este caso tenemos:

$$ln(\frac{P(Y=1/X)}{1-P(Y=1/X)}) = 0.685 - 0.222 * Agressor.Age$$

$$ln(\frac{P(Y=1/X)}{1-P(Y=1/X)}) = 0.603 + 0.053 * Agressor.Age$$

```
# Coefcientes Modelo

coefmodel.vgen.agr.adultos <- coef(model.vgenero.agresores.adultos)

coefmodel.vgen.agr.adultos
```

```
##   (Intercept) Agressor.Age
## 2   0.6853508  -0.22222470
## 3   0.6033503   0.05293524
```

```
# Odds Ratios Modelo


exp(coefmodel.vgen.agr.adultos )
```

```
##   (Intercept) Agressor.Age
## 2    1.984468    0.8007354
## 3    1.828234    1.0543614
```

```
# Intervalos de confianza odds ratio


Modelo.vgenero.agr.adultos.IC <- confint(model.vgenero.agresores.adultos)

Modelo.vgenero.agr.adultos.IC
```

```
## , , 2
##
##                    2.5 %     97.5 %
## (Intercept)  -0.7078401 2.07854160
## Agressor.Age -0.4768873 0.03243794
##
## , , 3
##
##                    2.5 %    97.5 %
## (Intercept)  -0.4212607 1.6279613
## Agressor.Age -0.1293450 0.2352155
```

2.5 - REPRESENTACIÓN DE RESULTADOS

Interpretación de Modelos

Tabla resumen

model.vgenero.convivientes.yes (Abuso.Previo)	Variables Independientes	B,(EE)	OR	IC95% OR	p-valor
No	Intercept	-2.830,(0.525)	0.058	(-3.858 ;-1.802)	< 0.001
No	Agressor.Age	0.481,(0.077)	1.618	(0.33; 0.632)	< 0.001
Unknown	Intercept	0.806,(0.395)	2.239	(0.031;1.581)	< 0.001
Unknown	Agressor.Age	0.084, (0.064)	1.087	(-0.04; 0.208)	< 0.001
model.vgenero.relation.expartner (Abuso.Previo)	Variables Independientes	B (EE)	OR	IC95% OR	p-valor
No	Intercept	-0.162,(0.958)	0.85	(-0.35;0.22)	< 0.001
No	Relation	-0.162,(0.096)	0.85	(-0.35;0.22)	< 0.001
Unknown	Intercept	0.195,(0.08)	1.22	(0.38;0.352)	< 0.001
Unknown	Relation	0.195, (0.08)	1.22	(0.38;0.352)	< 0.001
model.vgenero.agresores.adultos(Abuso.Previo)	Variables Independientes	B (EE)	OR	IC95% OR	p-valor
No	Intercept	0.685, (0.71)	1.984	(-0.708;2.079)	0.05986
No	Agressor.Age	-0.222, (0.129)	0.801	(-0.477;0.032)	0.05986
Unknown	Intercept	0.603, (0.523)	1.829	(-0.421;1.628)	0.05986
Unknown	Agressor.Age	0.052,(0.093)	1.054	(-0.129;0.235)	0.05986

Figure 1: Tabla resumen modelos

Cuando interpretamos las odds ratios de cada variable, se asume que el resto de variables independientes se mantienen fijas. Interpretaremos cada una de las variables independientes entre los distintos tipos de Abuso.Previo tomando como referencia Abuso.Previo = Yes (1).

2.6 - RESOLUCIÓN DEL PROBLEMA

REFERENCIAS

https://bookdown.org/content/2274/modelos-con-variables-cualitativas.html

https://www.usj.es/sites/default/files/tarjetas/aprendizaje/EstadisticaConceptosClave.pdf

https://biocosas.github.io/R/060_analisis_datos_categoricos.html

https://www.ccg.unam.mx/~vinuesa/R4biosciences/docs/Tema8_correlacion.html

https://www.cienciadedatos.net/documentos/22.2_test_exacto_de_fisher_chi-cuadrado_de_pearson_mcnemar_qcochran#(chi%5E2)_de_Pearson_(test_de_independencia)

https://rpubs.com/rslbliss/r_logistic_ws

https://en.wikipedia.org/wiki/Multinomial_distribution

https://rpubs.com/Joaquin_AR/220567

https://masteres.ugr.es/moea/pages/tfm1011/modelosderespuestamultinomialconraplicacionparaelestudiodeladepresionenpa

http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/Categor/Tema5Cate.pdf (Modelos Logit para respuestas nominales)