Tipología y ciclo de vida de los datos: Practica 2

Autor: Borja López Gómez y Sergio Beltrán Nuez

Diciembre 2021

Contents

Introducción

Presentación

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

Competencias

En esta práctica se desarrollan las siguientes competencias del Máster 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

Objetivos

Los objetivos concretos de esta práctica son: * 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.Tipología y ciclo de vida de los datos Práctica 2 pág 1

- 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 les permitan 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

Descripción de la PEC a realizar

La prueba está estructurada en 1 ejercicio teórico/práctico y 1 ejercicio práctico que pide que se desarrolle la fase de preparación en un juego de datos.
Deben responderse todos los ejercicios para poder superar la PEC.

Recursos

Los siguientes recursos son de utilidad para la realización de la práctica: * Calvo M., Subirats L., Pérez D. (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.

- Megan Squire (2015). Clean Data. Packt Publishing Ltd.
- Jiawei Han, Micheine Kamber, Jian Pei (2012). Data mining: concepts and techniques. Morgan Kaufmann.
- Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.
- Peter Dalgaard (2008). Introductory statistics with R. Springer Science & Business Media.
- Wes McKinney (2012). Python for Data Analysis. O'Reilley Media, Inc.
- Tutorial de Github https://guides.github.com/activities/hello-world.

1.Descripción del dataset

##Origen de datos

Para la realización de esta practica se propone la utilización del dataset Cervical Cancer Risk Classification, un dataset que recoge diferentes marcadores relacionados con la aparición del cancer de cuello uterino. El dataset puede encontrarse tanto en el enlace de Kaggle https://www.kaggle.com/loveall/cervical-cancer-risk-classification como en el repositorio UCI (https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29) Se trata de un dataset de origen público publicado por el Hospital Universitario de Caracas, Venezuela.

##Motivación

Pese a que se trata de unos de los canceres más prevenibles y el número de nuevos casos ha disminuido constantenmente en los últimos años, aproximadamente 4000 muejeres en Estado Unidos y 300000 en todo el mundo son diagnosticadas cada año. La mortalidad provocada por este cancer se ha reducido de manera notable gracias a las pruebas de cribado. No obstante, numerosos estudios confirman que el nivel de probreza y otros factores socioeconómicos están relacionados con bajas tasas de cribado. Por todo esto consideramos que se trata de un tema de actualidad que requiere de mucha colaboración ya que entender bien la información de que disponemos puede facilitar el estudio de futuros trabajos.

##Carga incial

Una vez obtenido el dataset el primer paso sería cargar los datos, entender de que tipo de información disponemos, analizar las variables y realizar un primer análisis de la información para ver la calidad de los datos y obtener las primeras conclusiones.

Para poder ilustrar el ejemplo, cargamos el fichero de datos

```
cervical_data <- read.csv('kag_risk_factors_cervical_cancer.csv',stringsAsFactors = FALSE)
rows=dim(cervical_data)[1]
```

Instalamos y cargamos las librerías ggplot2 y dplry

```
# https://cran.r-project.org/web/packages/ggplot2/index.html
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
# https://cran.r-project.org/web/packages/dplyr/index.html
if (!require('dplyr')) install.packages('dplyr'); library('dplyr')
```

2.Integración y selección de los datos de interés

##Análisis de variables

En primer lugar comprobamos la estructura del fichero:

```r
str(cervical_data)
```

```
## 'data.frame':    858 obs. of  36 variables:
##  $ Age                                 : int  18 15 34 52 46 42 51 26 45 44 ...
##  $ Number.of.sexual.partners           : chr  "4.0" "1.0" "1.0" "5.0" ...
##  $ First.sexual.intercourse            : chr  "15.0" "14.0" "?" "16.0" ...
##  $ Num.of.pregnancies                  : chr  "1.0" "1.0" "1.0" "4.0" ...
##  $ Smokes                              : chr  "0.0" "0.0" "0.0" "1.0" ...
##  $ Smokes..years.                      : chr  "0.0" "0.0" "0.0" "37.0" ...
##  $ Smokes..packs.year.                 : chr  "0.0" "0.0" "0.0" "37.0" ...
##  $ Hormonal.Contraceptives             : chr  "0.0" "0.0" "0.0" "1.0" ...
##  $ Hormonal.Contraceptives..years.     : chr  "0.0" "0.0" "0.0" "3.0" ...
##  $ IUD                                 : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ IUD..years.                         : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs                                : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs..number.                       : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs.condylomatosis                 : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs.cervical.condylomatosis        : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs.vaginal.condylomatosis         : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs.vulvo.perineal.condylomatosis: chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs.syphilis                       : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs.pelvic.inflammatory.disease    : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs.genital.herpes                 : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs.molluscum.contagiosum          : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs.AIDS                           : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs.HIV                            : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs.Hepatitis.B                    : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs.HPV                            : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs..Number.of.diagnosis           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs..Time.since.first.diagnosis    : chr  "?" "?" "?" "?" ...
##  $ STDs..Time.since.last.diagnosis     : chr  "?" "?" "?" "?" ...
##  $ Dx.Cancer                           : int  0 0 0 1 0 0 0 0 1 0 ...
##  $ Dx.CIN                              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Dx.HPV                              : int  0 0 0 1 0 0 0 0 1 0 ...
##  $ Dx                                  : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ Hinselmann                          : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ Schiller                            : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ Citology                            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Biopsy                              : int  0 0 0 0 0 0 1 0 0 0 ...
```

Se puede observar como existen 858 observación y un total de 36 variables que corresponden a los diferentes indicadores a tener en cuenta en la identificación del cancer de cuello de útero. La descripción de las variables es la siguiente:

Age Edad del paciente

Number of sexual partners Numero de personas con las que ha mantenido relaciones

First sexual intercourse Edad a la que mantuvo su primera relación

Num of pregnancies Número de embarazos

Smokes Indicador de si es fumador (1/0)

Smokes years Número de años siendo fumadora

Smokes pack years Valor que cuantifica el consumo de tabaco

Hormonal contraceptives Utilización de anticonceptivos hormonales (1/0)

Hormonal Contraceptives years Utilización de anticonceptivos hormonales en años

IUD Dispositivo intrauterino (1/0)

IUD years Dispositivo intrauterino en años

STDs Enfermedades de transmisión sexual (ETS)

STDs number Número de enfermedades (ETS)

STDs condylomatosis ETS condilomatosis (1/0)

STDs cervical.condylomatosis ETS condilomatosis cervical (1/0)

STDs vaginal.condylomatosis ETS condilomatosis vaginal (1/0)

STDs vulvo.perineal.condylomatosis ETS condilomatosis vulvo perineal (1/0)

STDs syphilis ETS sifilis (1/0)

STDs pelvic.inflammatory.disease ETS enfermedad pelvica inflamatoria (1/0)

STDs genital.herpes ETS herpes (1/0)

STDs molluscum.contagiosum ETS molusco contagioso (1/0)

STDs AIDS ETS SIDA (1/0)

STDs HIV ETS VIH (1/0)

STDs Hepatitis.B ETS Hepatitis B (1/0)

STDs HPV ETS Virus del papiloma humano (1/0)

STDs..Number.of.diagnosis Número de ETS diagnosticadas

STDs..Time.since.first.diagnosis Tiempo desde la primera vez diagnosticada

STDs..Time.since.last.diagnosis Tiempo desde la última vez diagnosticada

Dx.Cancer Diagnostico Cancer (1/0)

Dx.CIN Diagnostico lesiones precancerosas(1/0)

Dx.HPV Diagnostico VPH (Virus del papiloma humano)

Dx Existe diagnostico

Hinselmann Prueba de Hinselmann (1/0)

Schiller Prueba de Schiller(1/0)

Citology Prueba Citologia (1/0)

Biopsy Prueba Biopsia (1/0)

Selección de variables

Una vez identificadas las variables, el siguiente paso sería realizar un proceso de limpieza de los datos. Descartarlos datos con mala calidad o que, a priori, pensemos no aportan mucho valor en los resultados esperados. En este caso concreto podríamos eliminar las columnas relacionadas a los diferentes tipos de ETS dejando únicamente el indicador general que informa si ha padecido alguna enfermedad de este tipo y guardar los datos en un subset adicional por si fuera necesario lanzar análisis más detallados en el futuro.

```
cervical_data_subset<-cervical_data[c(1:13,26:36)]
```

3.Limpieza de datos

##Transformación de variables

Para poder obtener resultados estadísticos de los datos es necesario transformar algunos datos que viene como string a tipo entero.

```
cervical_data_subset$Number.of.sexual.partners<-as.integer(cervical_data_subset$Number.of.sexual.partne
cervical_data_subset$First.sexual.intercourse<-as.integer(cervical_data_subset$First.sexual.intercourse
cervical_data_subset$Smokes<-as.integer(cervical_data_subset$Smokes)
cervical_data_subset$Smokes..years.<-as.integer(cervical_data_subset$Smokes..years.)
cervical_data_subset$Num.of.pregnancies<-as.integer(cervical_data_subset$Num.of.pregnancies)
```

Otro paso dentro de la limpieza de datos podría ser eliminar los datos redundantes o duplicados, en este caso concreto, al no tener ningún cambo que identifique al paciente como único aceptaremos la premisa de que diferentes pacientes pueden peteir diferentes valores y por lo tanto no eliminaremos los registros duplicados.

##Eliminación de valores nulos o no informados

Siguiendo con el proceso de limpieza, el siguiente paso sería identificar los valores nulos, no informados, o informados con caracteres extraños.

Estadísticas de valores vacíos

```
colSums(is.na(cervical_data_subset))
```

```
##                                 Age        Number.of.sexual.partners
##                                   0                               26
##         First.sexual.intercourse                 Num.of.pregnancies
##                                   7                               56
##                              Smokes                    Smokes..years.
##                                  13                               13
##               Smokes..packs.year.           Hormonal.Contraceptives
##                                   0                                0
##  Hormonal.Contraceptives..years.                              IUD
##                                   0                                0
##                       IUD..years.                             STDs
##                                   0                                0
##                     STDs..number.        STDs..Number.of.diagnosis
##                                   0                                0
## STDs..Time.since.first.diagnosis  STDs..Time.since.last.diagnosis
##                                   0                                0
##                           Dx.Cancer                           Dx.CIN
##                                   0                                0
##                              Dx.HPV                               Dx
##                                   0                                0
```

```
##                         Hinselmann                            Schiller
##                                  0                                   0
##                           Citology                              Biopsy
##                                  0                                   0
```

Estadísticas de valores nulos

```
colSums(cervical_data_subset=="")
```

```
##                                 Age          Number.of.sexual.partners
##                                   0                                  NA
##           First.sexual.intercourse                  Num.of.pregnancies
##                                  NA                                  NA
##                              Smokes                      Smokes..years.
##                                  NA                                  NA
##                Smokes..packs.year.            Hormonal.Contraceptives
##                                   0                                   0
##  Hormonal.Contraceptives..years.                                   IUD
##                                   0                                   0
##                        IUD..years.                                STDs
##                                   0                                   0
##                      STDs..number.          STDs..Number.of.diagnosis
##                                   0                                   0
## STDs..Time.since.first.diagnosis  STDs..Time.since.last.diagnosis
##                                   0                                   0
##                            Dx.Cancer                              Dx.CIN
##                                   0                                   0
##                              Dx.HPV                                  Dx
##                                   0                                   0
##                         Hinselmann                            Schiller
##                                   0                                   0
##                           Citology                              Biopsy
##                                   0                                   0
```

Estadísticas de valores informados con '?'

```
colSums(cervical_data_subset=="?")
```

```
##                                 Age          Number.of.sexual.partners
##                                   0                                  NA
##           First.sexual.intercourse                  Num.of.pregnancies
##                                  NA                                  NA
##                              Smokes                      Smokes..years.
##                                  NA                                  NA
##                Smokes..packs.year.            Hormonal.Contraceptives
##                                  13                                 108
##  Hormonal.Contraceptives..years.                                   IUD
##                                 108                                 117
##                        IUD..years.                                STDs
##                                 117                                 105
##                      STDs..number.          STDs..Number.of.diagnosis
##                                 105                                   0
```

```
## STDs..Time.since.first.diagnosis  STDs..Time.since.last.diagnosis
##                              787                             787
##                       Dx.Cancer                          Dx.CIN
##                               0                               0
##                           Dx.HPV                              Dx
##                               0                               0
##                       Hinselmann                        Schiller
##                               0                               0
##                         Citology                          Biopsy
##                               0                               0
```

Una vez identificados, se procede a aplicar el tratamiento de falta de datos que más se adapte a cada tipo de dato teniendo en cuenta que un alto número de valores sin informar puede ser un indicador de incosistencia.Con ciertas variables podemos tomar el valor medio de los datos o el valor más repetido, en otros casos podemos tomar valores condicionales en base a la edad, por ejemplo si es fumador o no, fijando el valor NO para menores de 18 años. En otros casos es posible que nos convenga mantener los valores nulos y trabajar únicamente con el subset de datos informados como puede ser el indicador de si una persona estaba embarazada o no.

Antes de entrar a realizar tareas de transformación es conveniente categorizar las variables y comprobar que están informadas de un modo coherente. Por ejemplo, comprobar que los campos de edad no tiene valores muy elevados o caracteres en lugar de números, eliminar los espacios de las variables de tipo string, comprobar que los indicadores se informar únicamente con valores de 0 y 1...

Transformación de las variables Number of sexual partners y first sexual intercourse por la media:

```
cervical_data_subset$Number.of.sexual.partners[is.na(cervical_data_subset$Number.of.sexual.partners)] <-
cervical_data_subset$First.sexual.intercourse[is.na(cervical_data_subset$First.sexual.intercourse)] <-me
```

Filtramos todos los valores con el registro IUD =‘?’ y les aplicamos el valor NA, se trata de una variable muy importante en nuestro estudio y no es fácilmente interpolable por lo que al ser un número elevado de muestras puede afectar al análisis de otras variables.

```
cervical_data_subset$IUD[cervical_data_subset$IUD=="?"] <- "NA"
```

##Identificación y tratamiento de valores extremos

A continuación se procede a normalizar el campo edad para poder realizar representaciónes gráficas de un modo más sencillo

```
summary(cervical_data_subset[,"Age"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.00   20.00   25.00   26.82   32.00   84.00
```

```
cervical_data_subset["Age_segment"] <- cut(cervical_data_subset$Age, breaks = c(0,10,20,30,40,50,60,70,
```

---

4. Análisis de los datos

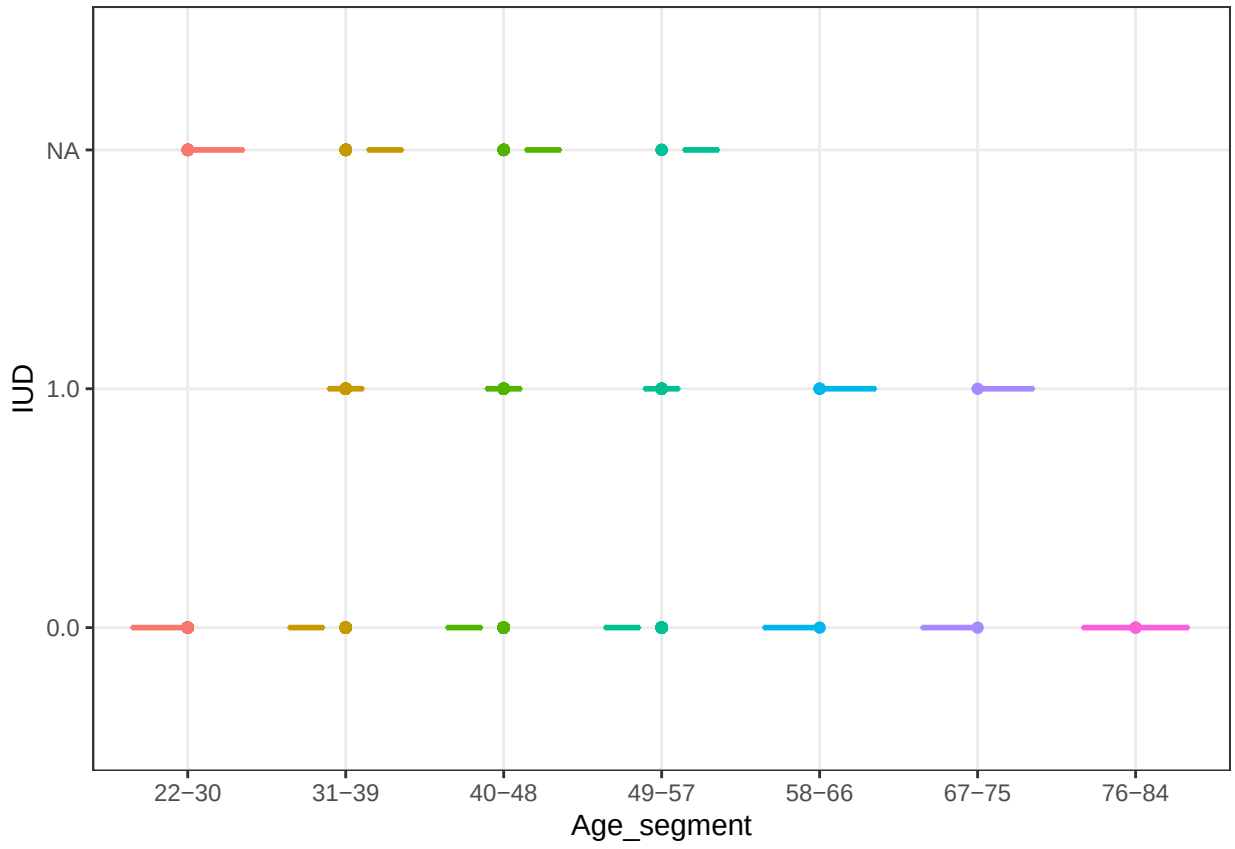
##Selección de los datos

Precisamos un primer set de datos para continuar con los análisis escogemos las variables más significativas como edad segmentada, variables relacionadas con el tabaco, los metodos hormonales y dispositivos intrauterinos Y

```
analisis_subset<-cervical_data_subset[c(25,5:13)]
str(analisis_subset)
```

```
## 'data.frame':    858 obs. of  10 variables:
##  $ Age_segment                     : Factor w/ 8 levels "13-21","22-30",..: 2 2 4 6 5 5 6 3 5 5 ...
##  $ Smokes                          : int  0 0 0 1 0 0 1 0 0 1 ...
##  $ Smokes..years.                  : int  0 0 0 37 0 0 34 0 0 1 ...
##  $ Smokes..packs.year.             : chr  "0.0" "0.0" "0.0" "37.0" ...
##  $ Hormonal.Contraceptives         : chr  "0.0" "0.0" "0.0" "1.0" ...
##  $ Hormonal.Contraceptives..years.: chr  "0.0" "0.0" "0.0" "3.0" ...
##  $ IUD                             : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ IUD..years.                     : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs                            : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs..number.                   : chr  "0.0" "0.0" "0.0" "0.0" ...
```

```
ggplot(data = analisis_subset, aes(x = Age_segment, y = IUD, colour = Age_segment)) +
  geom_boxplot() +
  geom_point() +
  theme_bw() +
  theme(legend.position = "none")
```

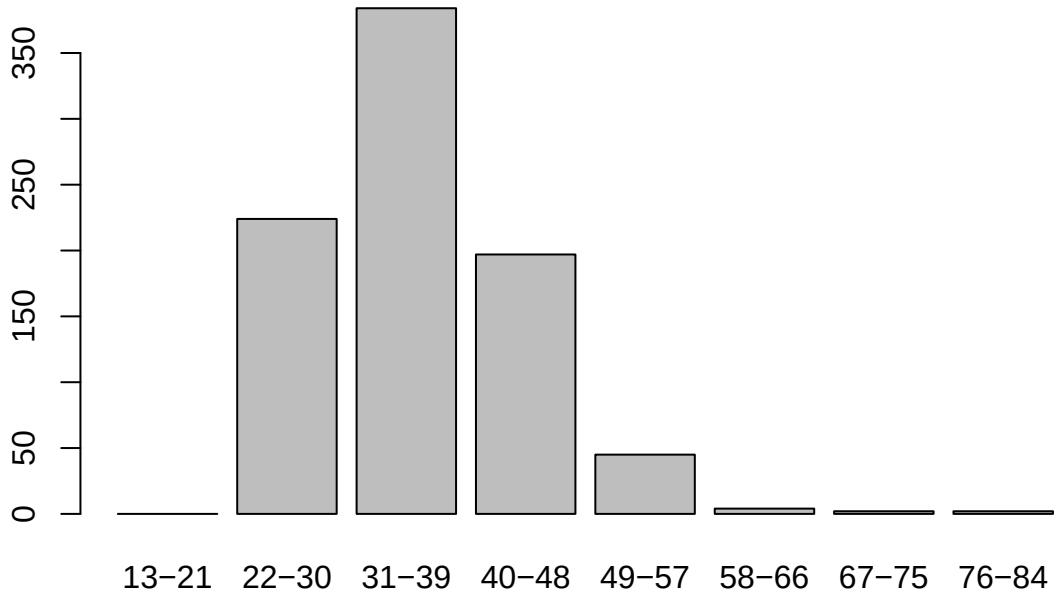


5. Representación de los resultados

##Representación gráfica Para entender los datos más en profundidad procedemos a realizar varias representaciones gráficas para obtener una visión más general de como se distribuye la información, de este modo nos será más sencillo identificar la naturaleza de los datos.

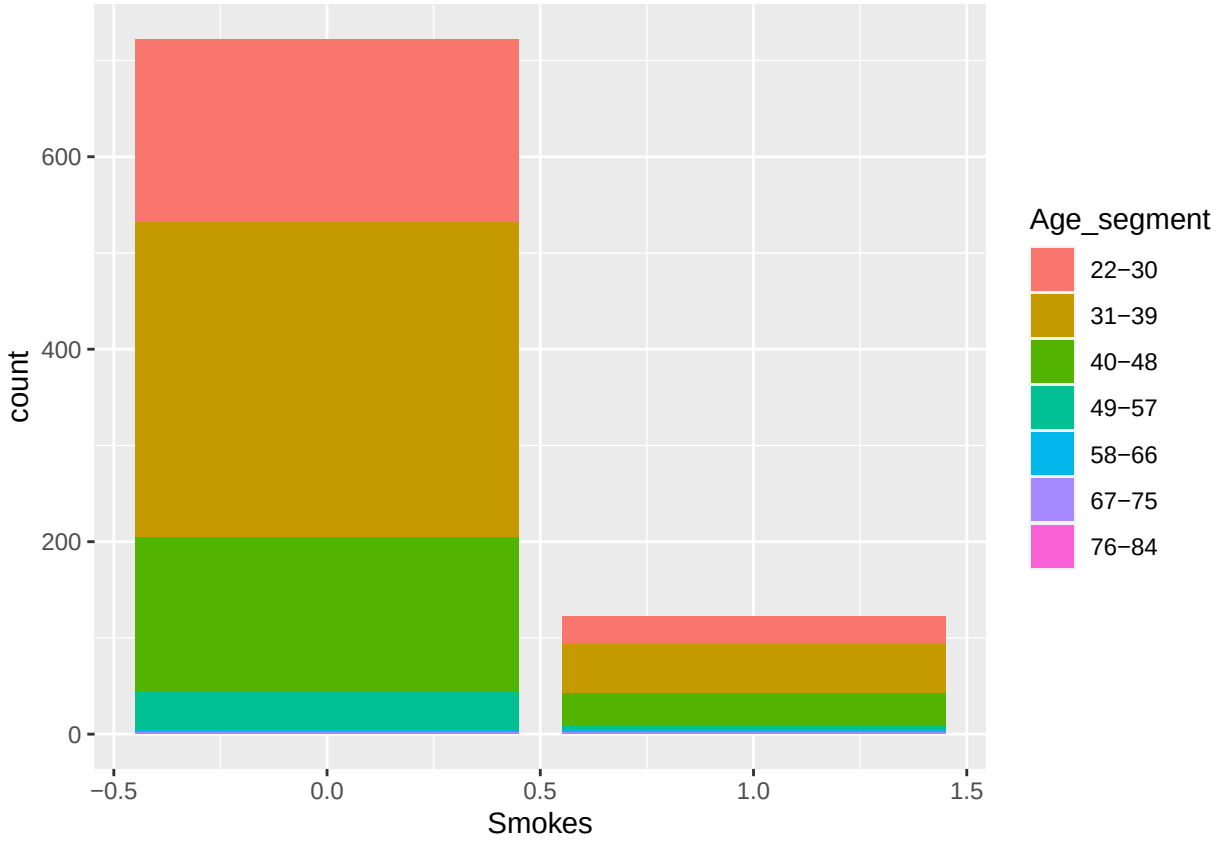
Distribución de los grupos de edad

```
plot(cervical_data_subset$Age_segment)
```



Representación de los grupos de edad en base a si son fumadores o no:

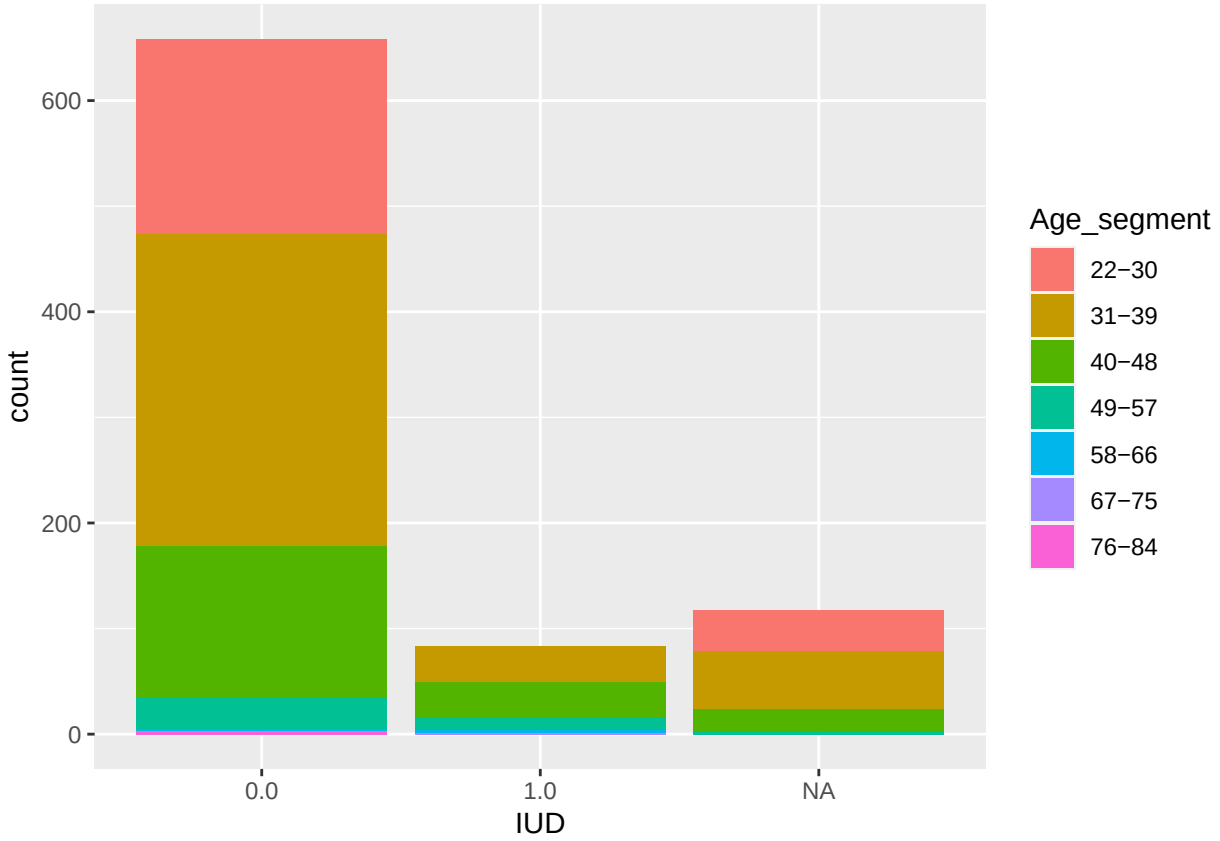
```
ggplot(data=cervical_data_subset[1:rows,],aes(x=Smokes ,fill=Age_segment))+geom_bar()
```



En la grafica vemos como muchos más casos corresponden a personas no fumadoras independiente del rango de edad ya que este se distribuye de acuerdo a la distribución anterior por grupos de edad, en todos los grupos encontramos personas fumadoras excepto en el grupo 0 de 13 a 21 años. En este caso quizás podría darnos pie a pensar que el tabaco no es un factor de riesgo en este tipo de cancer pero habría que confirmarlo con análisis más avanzados.

Representación de los grupos de edad en base a si son fumadores o no:

```
ggplot(data=cervical_data_subset[1:rows,],aes(x=IUD ,fill=Age_segment))+geom_bar()
```



6. Resolución de problemas

Una vez analizado el dataset en profundidad se abre un amplio número de posibilidades.Por ejemplo, seríamos capaces de identificar diferentes agrupaciones en base a las variables descritas y lanzar las campañas de concienciación personalizadas a los grupos más numerosos. Una vez establecida la clasificación sería necesario **analizar la calidad del modelo** para identificar que efectivamente los sujetos dentro de cada grupo tienen similitudes, existen diferencias entre los diferentes grupos y la distancia de cada muestra con el centro de su cluster es apropiada para definir un buen nivel de calidad. Por último, una vez establecido el modelo y teniendo en cuenta que todas las mujeres que se encuentran en el dataset han sufrido cáncer. Si hemos sido capaces de identificar grupos con características comunes con un aceptable nivel de calidad, seremos capaces de **lanzar campañas de prevención personalizadas** lo más efectivas posible. Un posible resultado podría ser por un lado, proponer a las mujeres de entre 30 y 40 que no utilicen dispositivo intrauterino; por otro lado, proponer la realización de pruebas diagnosticas gratuitas a las mujeres de entre 20 y 30 que nunca se han realizado pruebas con anterioridad y una tercera campaña podría ir dirigida a las personas mayores de 40 años que han fumado durante más de 5 años para que dejen de fumar y concienciarles de los riesgos de seguir fumando.