MDPI

Article

Mini Proyecto de Análisis Exploratorio de Datos

Nicolás Camañes Antolín, Rubén Castillo Carrasco, Erik González Soler, Jesús Martínez Leal

* Correspondence:

Simple Summary: En este proyecto de Análisis Exploratorio de Datos se explora un conjunto de datos real con el fin de obtener valiosas perspectivas y conclusiones.

Abstract: Nowadays, data has become a very powerful tool. By collecting, processing and utilizing data, companies gain multiple benefits, both economic and strategic. Such is te potencial of data that in recent decades, a large number of companies have emerged and grown, focusing their activities solely on data management. Through the implementation of this project, the aim is to put into practice the concepts learned in the Exploratory Data Analysis subject of the Master's degree in Data Science. The project consists of various stages. Firstly, the search for a suitable dataset. Next, the initial preprocessing of the dataset after loading it to be able to work in the programing language in which the project has been carried out (R, in our case). Thirdly, the generation of different graphs and tables to examine the characteristics of the study data, as well as the analysis of anomalies in the dataset. As a result of the analysis, information has been obtained regarding the different magnitudes and relationships existing between the variables studied.

Keywords: Análisis exploratorio de datos; Patrones en datos; Encuesta de población; Limpieza de datos; Exploratory Data Analysis

1. Introducción.

El conjunto de datos que se utiliza para el análisis pertenece al Instituto Nacional de Estadística (INE). En concreto, pertenece a una encuesta realizada a la población en el año 2021 que tiene como propósito el proporcionar información detallada sobre personas, viviendas y edificios que no puede obtenerse a través de registros administrativos. En nuestro caso, hemos decidido hacer la exploración en su mayoría de la parte relacionada a cuestiones que se les hizo a adultos (personas de 16 años o más). Para agregar más variables de interés se escogieron algunas de otro conjunto de datos, también de esta encuesta, hecho a todos los integrantes de la vivienda y no solo a adultos. El enlace a los ficheros de microdatos se encuentra disponible aquí.

Nota importante: Algunas gráficas se verán con relativamente mala calidad en este archivo *.PDF* ya que, debido a que tenemos un gran número de puntos en algunos gráficos, no fue posible la incorporación de los gráficos a nivel vectorial (ocupan demasiada memoria RAM en la visualización en PDF). Es por ello que se tuvo que hacer un `ggsave()` y guardar a *.png* y luego incrustarlo con un `includegraphics()`. Para una mejor visualización de dichos gráficos es mejor acudir a la carpeta *./figure* del directorio, donde habrá una versión en *HD* para ellas.

1.1. Carga de librerías y datos necesarios para el análisis

Primeramente, se cargan todas las librerías necesarias en las diferentes fases del proyecto. Esto se hace de manera más elegante utilizando el paquete `pacman` de nuestro lenguaje de programación, R.

A continuación, se realiza la carga del conjunto de datos, presente tanto en *.txt* como *.csv* en la carpeta *./data* incluida en el repositorio del proyecto. Asimismo, se carga el

Citation: Mini Proyecto de Análisis Exploratorio de Datos. *Journal Not Specified* **2023**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

fichero *.xlsx* que nos muestra el diseño de registro y valores válidos de las variables. Este fichero será especialmente útil, puesto que permitirá una automatización para convertir el conjunto de datos, que actualmente se encuentra en estado *Raw data* a *Technically correct data*, teniendo cada variable su tipo correspondiente.

A pesar de que resultó de gran utilidad el fichero que ofrecía INE para la transformación de las variables a un tipo correspondiente, surgieron algunos fallos que se debieron solventar de manera manual. Determinadas variables no fueron adecuadamente codificadas por parte del INE, de tal forma que fue necesario modificar ligeramente el fichero *.xlsx*. Concretamente, algunos diccionarios de variables en los que no terminaba de estar correctamente determinado el código (lo que sería el level en un factor). El problema presentado (en algunas variables) consistía, básicamente, en que aparecía "01" en lugar de "1", valor recogido en las observaciones del dataset.

Una vez hecha la conversión del dataset que contaba con un mayor número de variables se decidió que era de especial interés hacer una unión con el otro dataset mencionado al inicio, con el fin de obtener variables interesantes como `SEXO` o `EDAD`.

Las variables `EDAD`, `SEXO`, `NACIM` y `PNACIM` de este nuevo dataset deben ser convertidas a sus tipos correspondientes, por lo que se hace uso de la información que proporciona el INE sobre dichas variables, de manera similar a como se hizo anteriormente.

Para lograr la unión de ambos datasets, se emplea la función `left_join` del paquete `dplyr` utilizando las variables de identificación `IDEN`, `NPV` y `FACTOR`, de tal manera que podamos establecer una relación unívoca entre los adultos presentes en el primer y segundo dataset.

1.2. Características generales de los datos

Es posible hacerse una idea rápida de cuáles son los datos pertenecientes al dataframe, *df_merged*, haciendo uso de la función *glimpse*, perteneciente a la librería `dplyr`.

En resumen, nuestro dataset cuenta con 361934 observaciones y con 52 variables. Muchas de estas variables no nos serán de utilidad para las preguntas que deseamos responder, por lo que convendrá deshacernos de ellas por una cuestión de comodidad y optimización en los cálculos. Las variables están explicadas en *dr_ECEPOVadultos_2021.xlsx* y
dr_ECEPOVhogar_2021.xlsx. Tras un análisis detallado, las variables que nos convienen rescatar son las mostradas en la Tabla 1.

Table 1. Variables de interés para el estudio.

Variable	Descripción
IDEN	Identificador de la vivienda
IDQ_PV	Código de la provincia de residencia.
TAM_MUNI	Tamaño del municipio.
EC	Estado civil legal.
EDADEC	Edad de adquisición del estado civil legal.
ESTUDIOS	Nivel de estudios alcanzado.
ANOESTUD	Año que alcanzó su mayor nivel de estudios.
EDADESTUD	Edad a la que alcanzó su mayor nivel de estudios.
CAMPO	Campo de los estudios.
SITLAB	Situación laboral durante la última semana.
FLEXI	¿Puede flexibilizar, adaptar o acomodar su jornada laboral?
LUGTRAB	Lugar de trabajo
SATISTIEMP	Grado de satisfacción en el tiempo de desplazamiento al trabajo/estudio.
COMPRAINT	Realización de compras por internet en el último mes.
HIJOS	Tenencia de hijos.
NHIJOS	Número total de hijos.
TDOMEST	Grado de participación en las tareas domésticas del hogar.
SEXO	Sexo de la persona.
EDAD	Edad de la persona.
NACIM	Lugar de nacimiento de la persona (España u otro sitio).
PNACIM	País de nacimiento exacto de la persona.

En primer lugar, ahora que tenemos un conjunto más reducido, podemos cerciorarnos que algunas reglas sencillas se cumplen, como que las variables de `EDAD`, `EDADEC`, `SATISTIEMP`, `ANOESTUD` estén entre intervalos numéricos coherentes. Para ello, utilizamos el paquete `editrules`, para posteriormente aplicarlo sobre nuestro dataframe reducido, **df**.

Tras aplicar estas reglas a nuestro conjunto de datos, obtenemos que no hay ninguna violación de estas (obtenemos un NULL tras usar la función `violatedEdits` sobre nuestro dataframe), por lo que no será necesario hacer una imputación o eliminado de dichas observaciones.

1.3. Análisis de missing data en nuestro conjunto de interés

Una representación precisa de cómo se distribuyen los *missing values* por nuestro conjunto de datos lo otorga la librería `VIM` con la función `aggr()`. Se han representado solamente aquellas variables que cuentan con algún NA para facilitar la visualización. En la Figura 1 puede observarse tanto la proporción total de valores faltantes en dichas variables individualmente (izquierda), como una serie de combinaciones posibles entre ellas de valores faltantes.

Así, encontramos por ejemplo que la combinación de NA en (`SATISTIEMP`, `FLEXI`, `ANOESTUD`, `EDADESTUD`, `CAMPO`, `LUGTRAB`) cuenta con una proporción del 25 % de los datos totales.

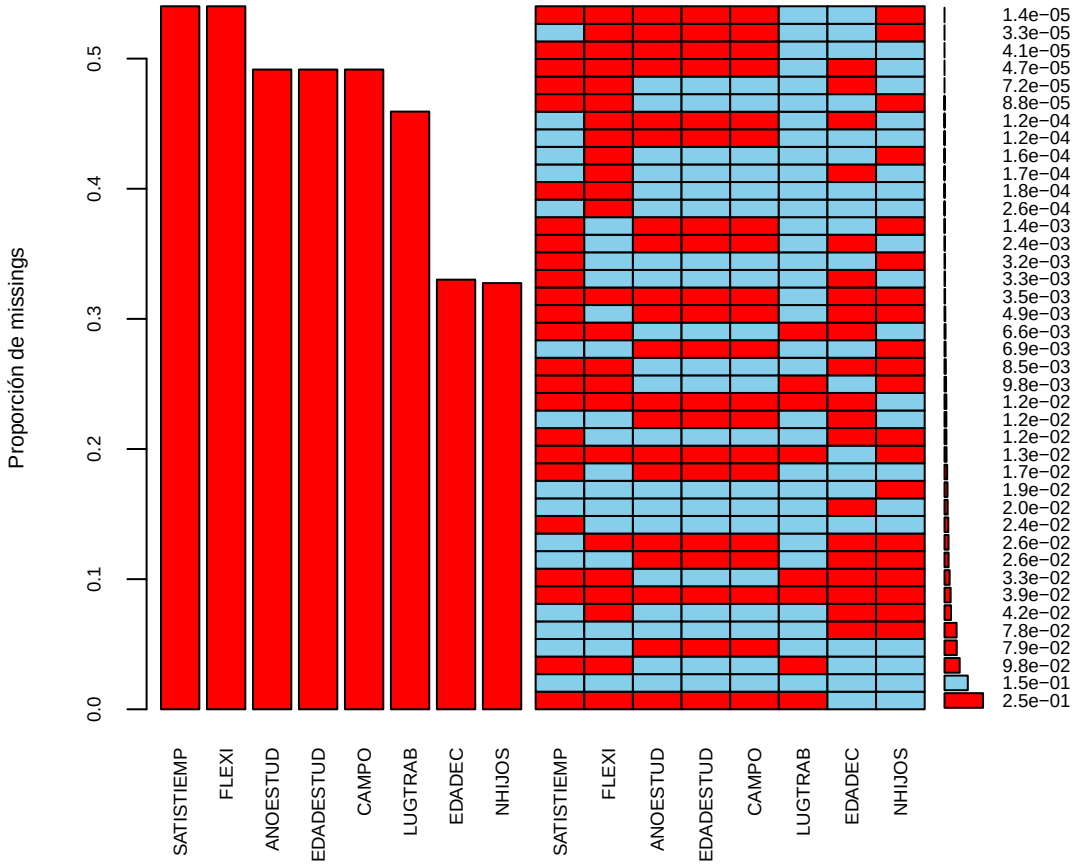

Figure 1. Esquema de valores missing en nuestro dataset

Podemos observar en el gráfico de barras de la izquierda que hay proporciones de missings que parecen repetirse. Esto nos da pie a plantearnos si acaso existe una razón tras ellos más allá de la pura aleatoriedad (*missing completely at random*). Quizás existen variables observadas en nuestro conjunto de datos que según tengan un valor u otro inducen un *NA* en otras variables…

Esta cuestión puede hacerse frente utilizando la información adicional que INE nos proporciona para la base de datos. En este caso, haremos uso del cuestionario dado. Se llega a las siguientes conclusiones:

- Si la variable `EC` (*Estado civil*) es "Soltero/a" ⇒ la variable `EDADEC` debe contener un *NA*.
- Si la variable `ESTUDIOS` es una de estas: "No sabe leer o escribir", "Sabe leer y escribir pero fue menos de 5 años a la escuela", "Educación primaria completa o fue a la escuela al menos 5 años", "Primera etapa de educación secundaria y similar (EGB, Bachiller elemental, ESO, certificado de Estudios Primarios, certificado de Escolaridad o certificado de Profesionalidad niveles 1 o 2)" ⇒ las variables `ANOESTUD`, `EDADESTUD` y `CAMPO` deben contener un *NA*.
- Si la variable `SITLAB` (*Situación laboral actual*) es **distinta** de "Ocupado/a" (a tiempo completo o a tiempo parcial) ⇒ la variable `FLEXI` (*Flexibilidad en el horario de trabajo*) se pone como *NA*, automáticamente.
- Por otra parte, de manera relacionada a la anterior, la variable `SATISTIEMP` puede aparecer con un valor distinto de *NA* solo si la categoría de `SITLAB` es "Ocupado/a" (a

tiempo completo o a tiempo parcial) o "Estudiante" **Y** además la variable `LUGTRAB` es distinta de "En el propio domicilio" o "En varios municipios (soy comercial, repartidor, taxista...)".

- La variable `NHIJOS` (*número de hijos*) se pone como *NA* si y solo si la variable booleana `HIJOS` (¿ha tenido hijos?) está en "No".

Nota: es importante recalcar que en el propio cuestionario se pedía que a las personas encuestadas que estuvieran *teletrabajando* debido al COVID-19 respondieran teniendo en cuenta su situación anterior.

Esta serie de conclusiones de pertenencia de NA o no se comprueban en el código presente en el *.Rmd* del directorio.

```
[1] "Todas las condiciones se satisfacen, tal y como menciona el INE."
```

Habiéndose cumplido lo que prometía el INE podemos asegurar que en nuestro conjunto de datos (al menos en nuestras variables seleccionadas de interés) no hay ningún ningún valor que se haya perdido estrictamente. Todos los NA responden a que determinadas variables (presentes en el dataset) toman un valor u otro (*missing at random*).

En nuestro análisis, hemos decidido que los únicos valores NA que pueden cobrar sentido y ser imputados son los de la variable `NHIJOS`, ya que consideramos que con 16 años (edad mínima en nuestro conjunto de datos) este valor no tendría por qué ser siempre 0. Esto permitirá que el mínimo de `NHIJOS` sea 0, y no 1.

Tras hacer este cambio, tenemos datos consistentes pudiendo así empezar a extraer conocimiento sobre nuestro conjunto de datos.

En primer lugar, mediante la función *summary*, obtenemos información acerca de las características de cada variable del dataset. Para las categóricas (factor) encontramos las frecuencias de los distintos niveles, mientras que para las numéricas se indica el mínimo, máximo y mediana, entre otras cosas.

Ya que la función *summary* devuelve demasiada información como para introducirse en este .PDF, hemos realizado una modificación de esta, incluyéndose en la Tabla 2.

Table 2. Información relevante acerca de las variables de estudio.

variable	type	levels	topLevel	topCount	topFrac	missFrac
IDEN	character	172444	060772	17	0.000	0.000
IDQ_PV	factor	52	Madrid	39559	0.109	0.000
TAM_MUNI	factor	4	De 50.000 habitantes o menos	143289	0.396	0.000
SEXO	factor	2	Mujer	187860	0.519	0.000
EDAD	numeric	94	55	7609	0.021	0.000
NACIM	factor	2	España	323737	0.894	0.000
PNACIM	factor	161	España	323737	0.894	0.000
EC	factor	6	Casado/a en primeras nupcias	181494	0.501	0.000
EDADEC	numeric	90	25	16159	0.067	0.330
ESTUDIOS	factor	12	Primera etapa de educación secundaria y simi...	110093	0.304	0.000
ANOESTUD	numeric	88	2021	9047	0.049	0.492
EDADESTUD	numeric	65	18	26174	0.142	0.492
CAMPO	factor	11	Arquitectura, Construcción, Formación Técnic...	26759	0.145	0.492
SITLAB	factor	9	Ocupado/a - A tiempo completo	142613	0.394	0.000
FLEXI	factor	3	No	94653	0.569	0.540
LUGTRAB	factor	7	En el municipio en el que resido	86875	0.444	0.459
SATISTIEMP	numeric	12	10	41244	0.248	0.540
COMPRAINT	factor	2	No	214970	0.594	0.000
NHIJOS	numeric	19	2	122321	0.338	0.000
HIJOS	factor	2	Sí	243404	0.673	0.000
TDOMEST	factor	4	Me encargo de una parte importante de las ta...	122130	0.337	0.000

2. Exploración / visualización

Una vez hemos asegurado que nuestros datos estén en la estructura de data.frame, tengan los valores correctamente etiquetados y estén almacenados con el tipo correcto, así como de conocer el origen de los NA, podemos empezar a buscar posibles patrones en las instancias o entre las características. Una serie de preguntas surgen a nosotros en lo que es un *brainstorming* tras observar qué datos nos eran interesantes.

- ¿Cómo estarán distribuidas las edades de nuestros encuestados? ¿Se seguirá el comportamiento de distribución normal?
- ¿En qué situación laboral estarán nuestros adultos (más de 16 años) encuestados? ¿Serán la mayoría trabajadores a jornada completa?
- ¿De qué manera estará el hecho de la edad que posee un encuestado con su número de hijos? ¿Habrá también una relación entre la edad y su satisfacción respecto al desplazamiento al trabajo/estudio?
- ¿Existirá una diferencia relevante entre la satisfacción de desplazamiento y la provincia de la que es el encuestado? ¿Serán los lugareños de Madrid y Barcelona los más desquiciados en este aspecto?
- ¿Será posible ver una distinción entre campos de estudio según el sexo del encuestado? ¿Será cierto que los hombres tienden a ir a ramas más relacionadas con la Arquitectura y Construcción, mientras que las mujeres más a aspectos sociosanitarios?
- ¿Habrá una cierta tendencia a colaborar en las tareas domésticas o no según la situación laboral del encuestado? ¿Dependerá de haber tenido hijos?
- ¿Qué variables numéricas de nuestro conjunto de datos tendrán cierta correlación, ya sea lineal o no lineal?
- ¿Qué concordancia existirán entre las variables categóricas de nuestro estudio? ¿Existirá entre el estado civil y tener o no hijos?
- ¿Podremos ir más allá y tratar de implementar un algoritmo de clustering con el objetivo de detectar instancias atípicas en nuestro conjunto de datos?

Todas estas preguntas se intentarán responder en el presente documento, mostrando representaciones gráficas o por métodos más estadísticos.

2.1. Análisis univariante.

Distinguiremos el análisis entre las variables de tipo numérico y las de tipo categórica (factor) presentes, ya que ciertos estadísticos descriptivos (como por ejemplo la media) carecen de sentido en las pertenecientes al último tipo.

2.1.1. Variables de tipo numérico

Se ha creado una función para poder obtener diferentes estadísticos de las variables numéricas. Dichos estadisticos se representan en la Tabla 3. Nótese que emplear el argumento `na.rm = TRUE` tiene sentido, tal y como se ha demostrado anteriormente.

Table 3. Estadísticos de variables numéricas

Variable	Media	Mediana	DesvEst	Asimetria	Curtosis
EDAD	51.42	52	18.68	0.01	-0.68
EDADEC	33.56	29	14.17	1.75	2.77
ANOESTUD	1998.05	1998	15.91	-0.34	-0.61
EDADESTUD	23.10	22	6.22	2.30	7.41
SATISTIEMP	7.38	8	2.44	-1.05	0.83
NHIJOS	1.41	1	1.30	1.04	3.03

Es destacable ver que la variable `EDAD` posee una asimetría cercana a 0, lo que indica que la distribución tiene una forma que es aproximadamente igual a ambos lados del valor central.

Continuando el análisis univariante de las variables numéricas, se ha generado el siguiente gráfico. Este se corresponde con el histograma de la variable `EDAD` y con la función de distribución normal asociada (hace uso de media y desviación típica de los datos).

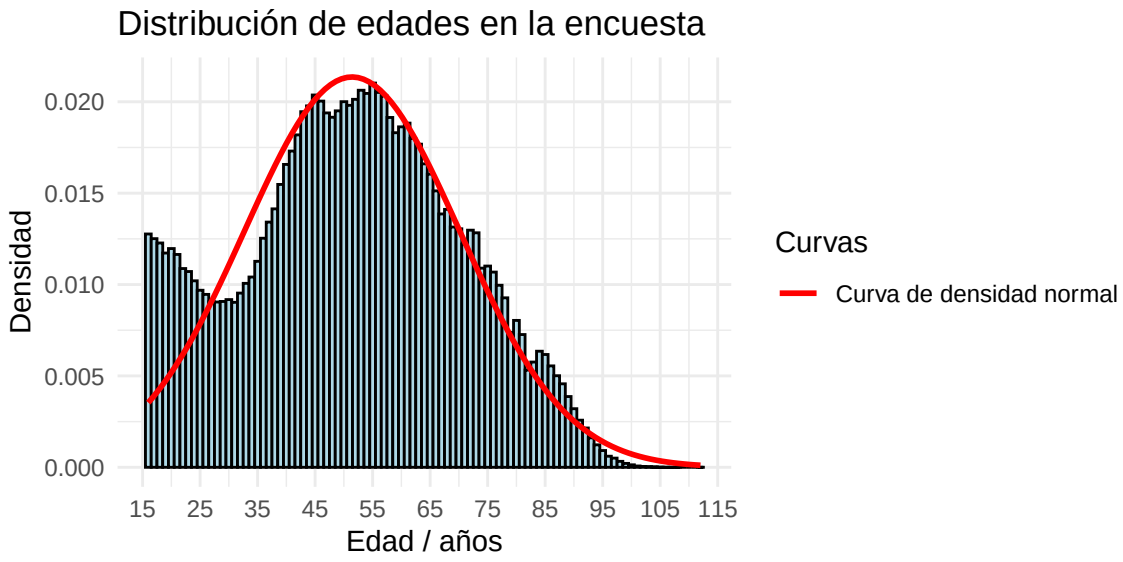

Figure 2. Distribución de edades para la encuesta con la curva gaussiana

Mediante el gráfico anterior, se aprecia que las edades más fracuentes en el dataset van de 40 a 60 años. Apreciamos además que en cierta parte la concordancia con una gaussiana es clara, yéndose esta similitud hacia las colas. Esto se observa también en la Figura 3, donde se hace un *quantile-quantile plot*.

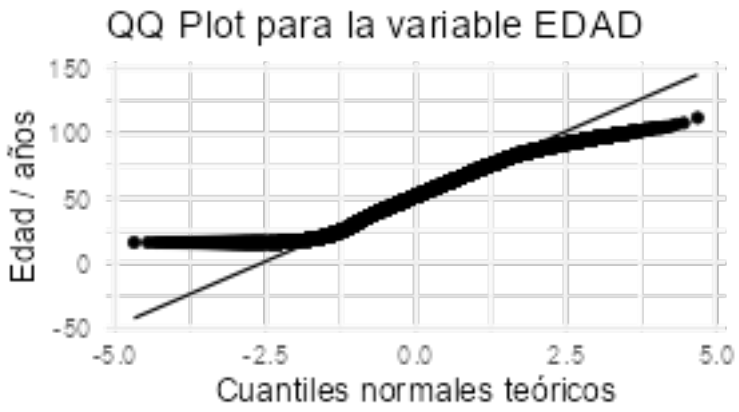

Figure 3. Q-Q plot para la variable EDAD.

La linea diagonal indica el comportamiento gaussiano ideal.

2.1.2. Variables de tipo categórico

A diferencia de las variables numérica, en las variables categóricas no es posible obtener muchos estadísticos. El más común es la moda, que se corresponde con la columna *topLevel* de la Tabla 2.

Otra información que se puede analizar en el análisis univariante de una variable categórica es la frecuencia de aparición de cada una de las categorías. Se representa esto

en la Tabla 4. Es posible también representar esta información visualmente, a través de un gráfico de barras como en la Figura 4.

Table 4. Frecuencias de cada grupo según su situación laboral.

SITLAB	Frecuencia
Ocupado/a - A tiempo completo	142613
Ocupado/a - A tiempo parcial	23806
Estudiante	29287
Parado/a - Ha trabajado anteriormente	36129
Parado/a - No ha trabajado anteriormente	5241
Jubilado/a, prejubilado/a	80865
Incapacitado/a permanentemente para trabajar	9436
Dedicado/a las tareas del hogar	27462
Otro tipo de inactividad	7095

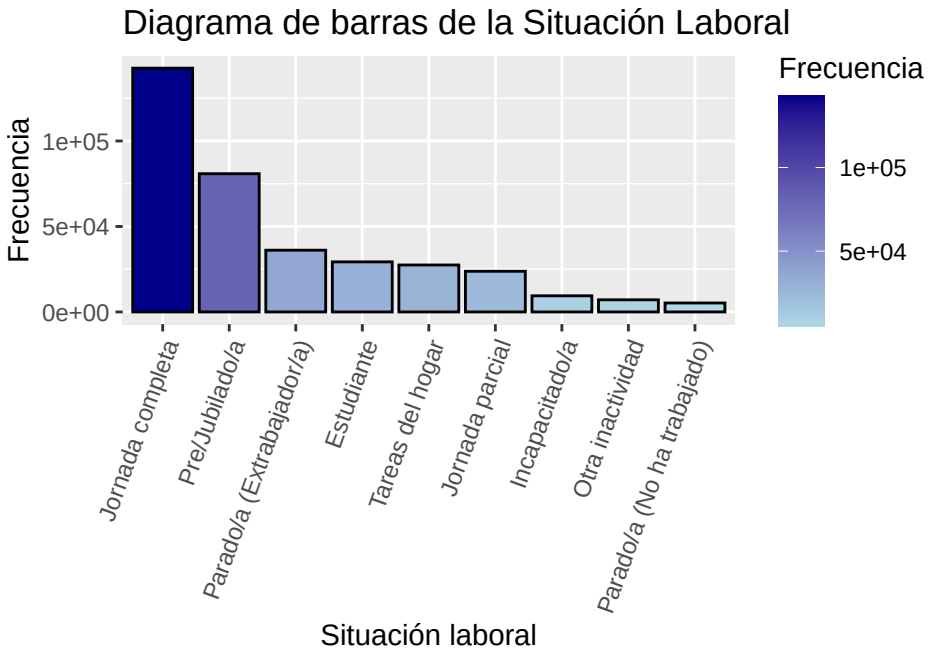

Figure 4. Diagrama de barras para la variable SITLAB.

Tal y como se aprecia en el gráfico, en el momento en que se realizó la encuesta, predominaban claramente las personas ocupadas a tiempo completo, seguidas de las jubiladas o prejubiladas. Por otra parte, las personas paradas que no hubieran nunca trabajado representaban el grupo más pequeño.

2.1.3. Análisis de outliers univariante

Para la detección de outliers, observaciones que parecen "diverger" del patrón de comportamiento del resto de datos, resulta fundamental establecer un criterio para su identificación. En nuestro caso, hemos utilizado 4 métodos diferentes:

- Regla 3 σ: asume distribución gaussiana seguida por los datos.
- Identificador Hampel: no presupone distribución gaussiana y utiliza la mediana.
- Regla boxplot: a partir del boxplot se identifican como lo que queda fuera de los bigotes.
- Regla $P_5 - P_{95}$ de percentiles: lo que queda por debajo del 5% o por encima del 95% es tratado como outlier.

Una aplicación de los 4 métodos a la variable `EDAD` lleva a la obtención de la Tabla 5.

Table 5. Identificación de Outliers de la variable EDAD.

method	n	nMiss	nOut	lowLim	upLim	minNom	maxNom
tresSigma	361934	0	2	-4.615272	107.4627	16	107
Hampel	361934	0	1	-5.821400	109.8214	16	108
ReglaBoxplot	361934	0	4	-2.500000	105.5000	16	105
p5-p95	361934	0	34188	20.000000	83.0000	21	82

En esta tabla apreciamos el método empleado (*Method*), el número de observaciones totales de la variable *n*, el número de datos perdidos *nMiss*, el número de outliers total detectado *nOut*, el extremo inferior y superior del cálculo (*lowLim* y *upLim*), así como el mínimo y máximo de la variable que han sido recogidos sin entrar en la identificación como outlier (*minNom* y *maxNom*).

Más allá de esto, es posible identificar visualmente los outliers en una representación gráfica, tal y como se muestra en la Figura 5 para la variable EDAD. Una generalización a todas las variables numéricas se ha realizado en el código de R para la identificación de los Outliers, pero no se mostrará por cuestiones de espacio.

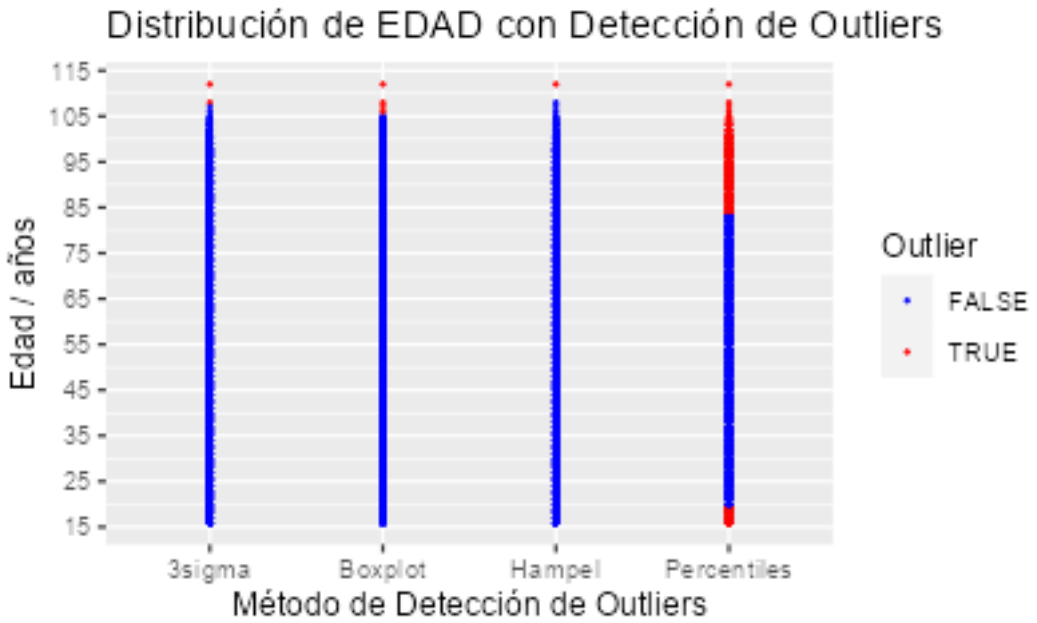

Figure 5. Outliers para la variable EDAD por diferentes métodos.

En la Figura 5 podemos apreciar que los 3 primeros métodos (3σ, Boxplot, Hampel) han resultado muy permisivos con los outliers, mientras que el de percentiles ha eliminado bastantes valores, fruto de que el umbral superior de aceptación de no ser outlier era 83 años, como se observa en la Tabla 5.

Analizando nuestro conjunto de datos, hemos llegado a la conclusión de que los Outliers que aparecen no se deben en ningún caso a datos erróneos. Esto es así porque el INE ha debido de hacer un preprocesado de posibles valores incorrectos con anterioridad. Así pues, consideramos que carece de sentido imputarlos e incluso eliminarlos, ya que pueden resultar de interés en nuestro estudio.

2.2. *Análisis multivariante*

Como su nombre indica, la fase de análisis multivariante, consiste en la búsqueda de relación entre dos variables o más. Para ello existen diferentes formas de representación, en función del tipo de variables que se esté tratando tratando.

2.2.1. Gráficos

Análisis entre variables numéricas

Mediante la realización del siguiente gráfico se ha tratado de representar la relación existente entre las variables: edad (`EDAD`), número de hijos (`NHIJOS`) y grado de satisfacción con el tiempo de desplazamiento al trabajo/estudio (`SATISTIEMP`).

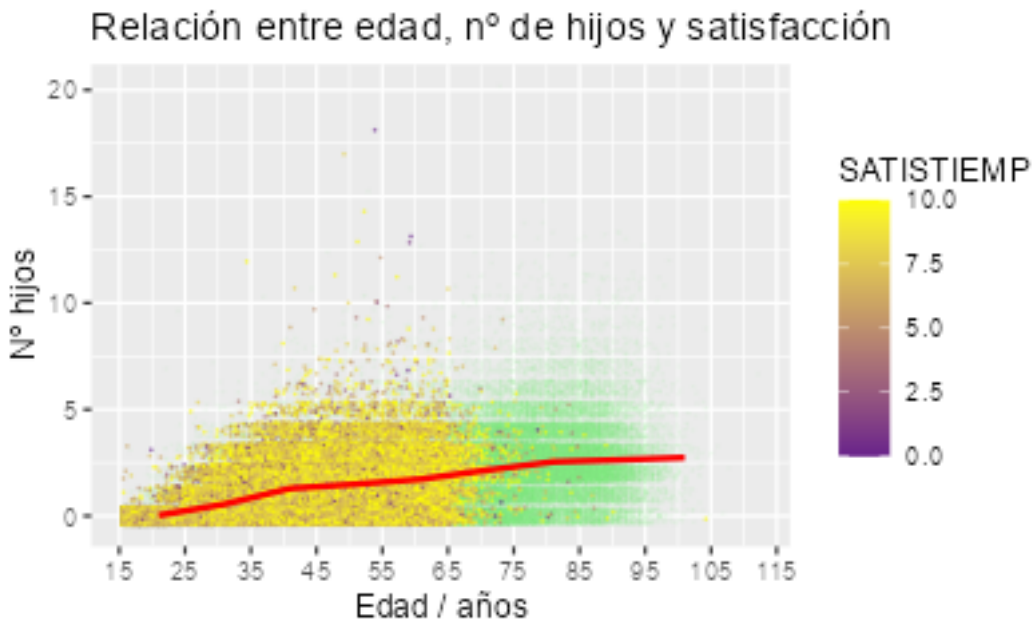

Figure 6. Diagrama de dispersión para representar la relación entre EDAD, NHIJOS y SATISTIEMP

Este gráfico corresponde a uno de dispersión con puntos "jittered" o dispersos, lo que resulta útil dada la enorme superposición de puntos presentes. La línea roja se ha hecho a partir de las medias en el número de hijos para los intervalos ("16-26", "27-36", "37-46", "47-56", "57-66", "67-76", "77-86", "87-96", "97-106", "107-116"), cogiendo el punto medio en cada caso. Es evidente notar que hay una tendencia creciente en que el número de hijos sea mayor conforme la edad es mayor. Esto es lo que uno esperaría, ya que antiguamente en España se solían tener mayor cantidad de hijos. Vemos además que los NA para la variable `SATISTIEMP` [color verde] está muy concentrada en edades avanzadas, resultado de que la mayoría de gente está jubilada. Los resultados en formato tabla no se muestran en el documento en aras de la brevedad, pero están disponibles en el *.Rmd*, al igual que muchas otras cosas.

Los datos exactos se recogen en la Tabla 6.

Table 6. Media de hijos de y de satisfacción de desplazamiento según el grupo de edades indicado.

Grupo edades	Media de hijos	Media de satisfacción
16-26	0.03	7.02
27-36	0.55	7.23
37-46	1.29	7.37
47-56	1.49	7.47
57-66	1.73	7.68
67-76	2.15	8.03
77-86	2.54	7.19
87-96	2.63	8.50
97-106	2.74	10.00

Es notable apreciar que la media de hijos se incrementa según avanzamos en el grupo de edades, como cabría esperar. La media de satisfacción por otra parte parece seguir una ascensión conforme se incrementa la edad. Sin embargo, en el intervalo de "77-86" se ve una caída.

Análisis entre variables numéricas y categóricas

A partir de la Figura 7 podemos hacer una representación de este tipo. En este caso, relacionaremos la variable `SATISTIEMP` (Satisfacción en el tiempo al trabajo / estudio) junto con la variable categórica de provincia (`IDQ_PV`). Para poder tener algo de interacción se recomienda correr el código en .Rmd.

Figure 7. Valor medio de la variable SATISTIEMP en función de la provincia.

A la vista del mapa, es claro apreciar que la media de satisfacción desplazamiento en la provincia de Madrid es algo inferior al resto de sitios, seguida por la de Barcelona.

Analisis entre variables categóricas.

Es también es posible representar la relación entre varias variables categóricas. Para ello, lo más común es emplear el gráficos de barras o mosaicos. En la siguiente representación (Figura 8) se ha expresado la relación entre la el campo de los estudios (`CAMPO`) con el sexo de las personas (`SEXO`).

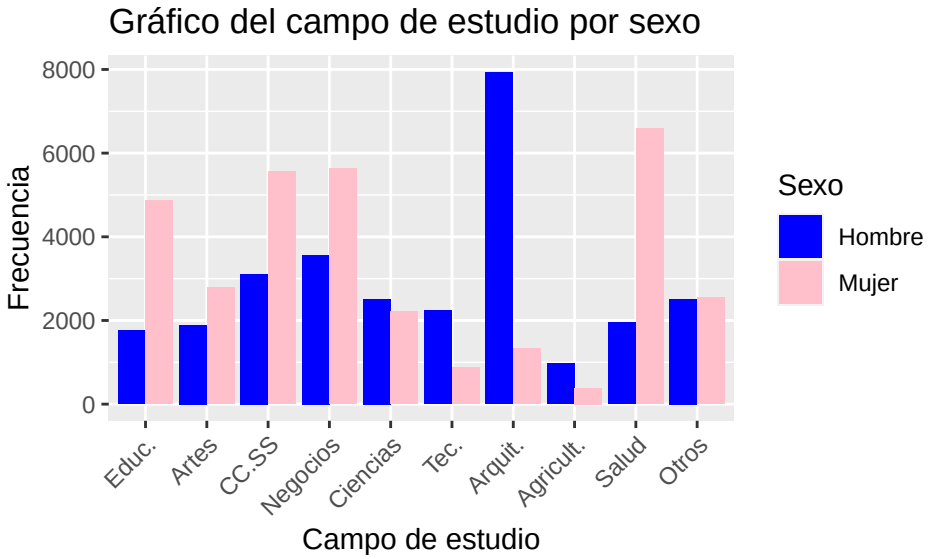

Figure 8. Gráfico de barras del campo de estudio en función del sexo.

Como se puede apreciar en el gráfico de barras anterior, hay campos de estudio en los que predominan los hombres frente a las mujeres, como es en el caso de las Arquitecturas. En otros campos pasa lo contrario, como en el caso de la salud o la educación.

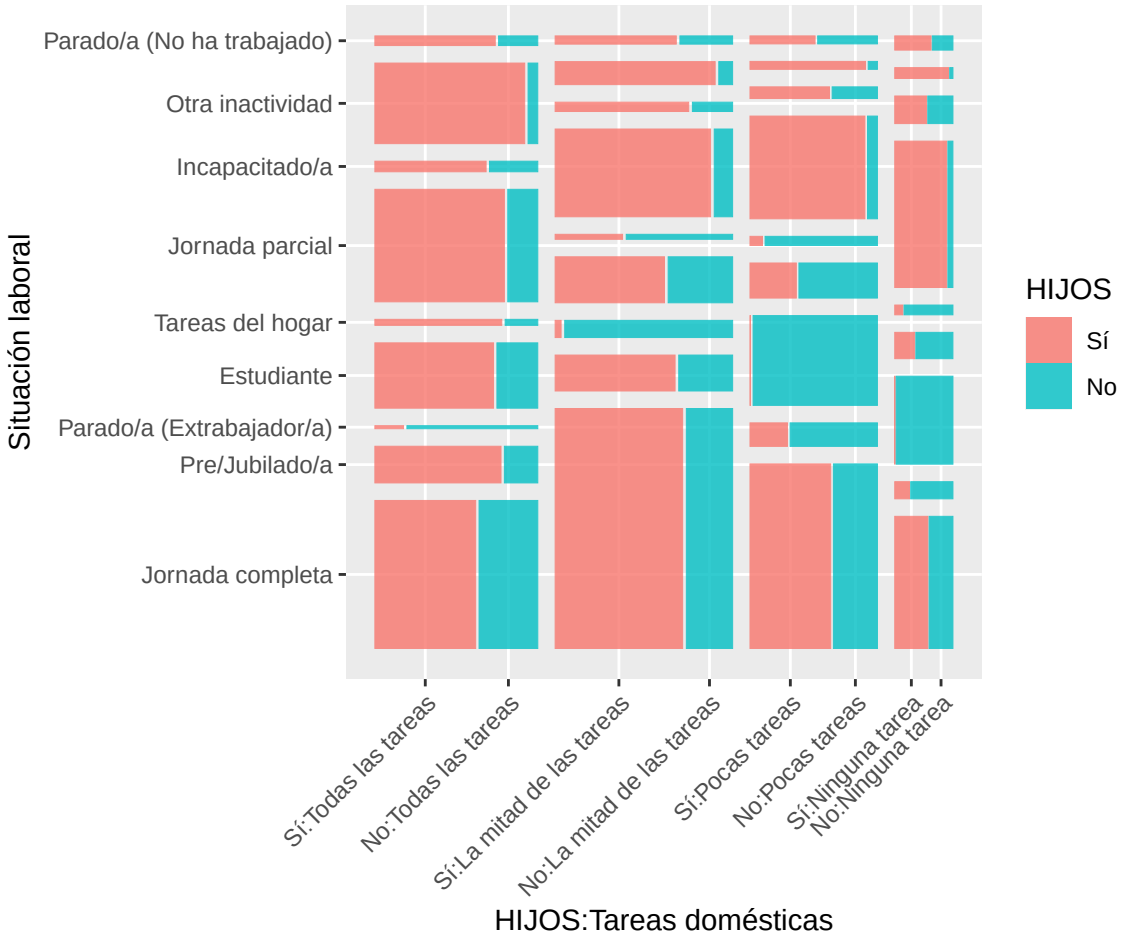

Figure 9. Gráfico de barras del campo de estuido en función del sexo

Un gráfico de mosaico es una herramienta de visualización muy poderosa. Podemos visualizar 3 variables de manera simultánea en la Figura 9. Puede observarse en primer lugar la proporción relativa visualmente según la situación laboral, ya que actúa como una especie de gráfico de barras, pudiendo asociarse con la altura de los rectángulos (ver Figura 4). Por otra parte, es posible apreciar dentro de cada grupo según la situación laboral y su función en realizar las tareas domésticas, cuál es la tendencia a tener hijos. Así, puede observarse una clara tendencia a que los "Estudiantes", según incrementan su contribución a las tareas domésticas, son más propensos a haber tenido hijos.

2.2.2. Caracterización

Otra forma de analizar la relación existente entre variables es mediante el cálculo de estadísticos, como es el caso de las correlaciones o covarianzas para las variables numéricas.

Variables numéricas

Para el caso de las variables numéricas, una técnica común es la creación de un mapa de correlaciones con todas las variables del dataset. En primer lugar, se ha generado un mapa para poder representar las correlaciones de Pearson. Que se corresponde con el mostrado en la Figura 10.

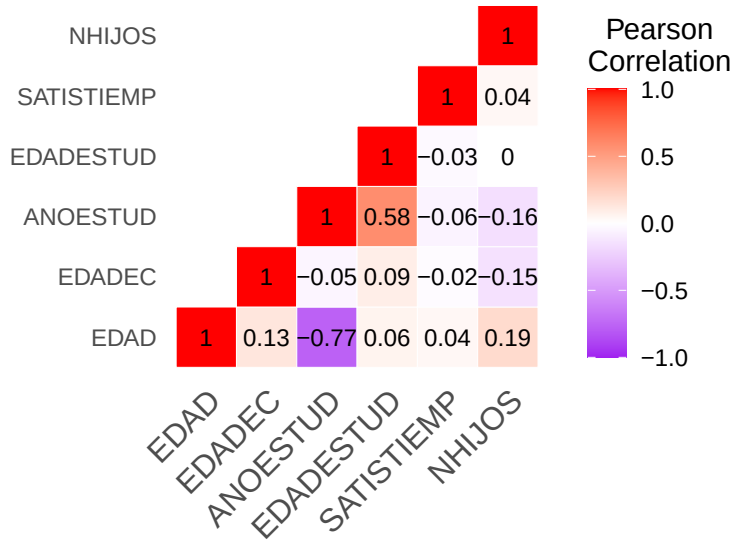

Figure 10. Mapa de correlaciones de Pearson para las variable numéricas

La correlación de Pearson es de gran utilidad para poder determinar la dependencia lineal entre diferentes variables. Analizando los resultados obtenidos, sólo se puede apreciar una dependencia algo significativa entre los pares de variables (`EDAD`, `ANOESTUD`). Esto de cierta manera tiene sentido que estén correlacionadas y además con una dependencia negativa, ya que lógicamente un incremento en tu `EDAD` suele llevar a que acabaras tus estudios máximos años anteriores, siendo además lineal por cuestiones de linealidad temporal. Esta relación puede apreciarse en la Figura 11. Entre el resto de variables numéricas no existe una dependencia lineal significativa.

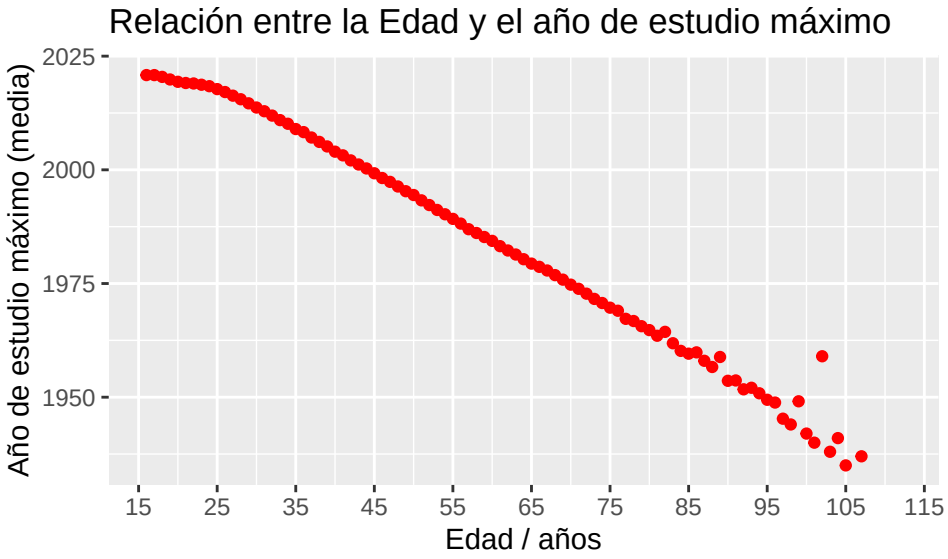

Figure 11. Relación entre la variable EDAD y la media de ANOESTUD.

Otra manera de ver una correlación entre variables numéricas lo da el índice de correlación de Spearman. En este caso, este es un estadístico que nos da la relación monótona entre pares de variables, no midiendo únicamente una relación estrictamente lineal.

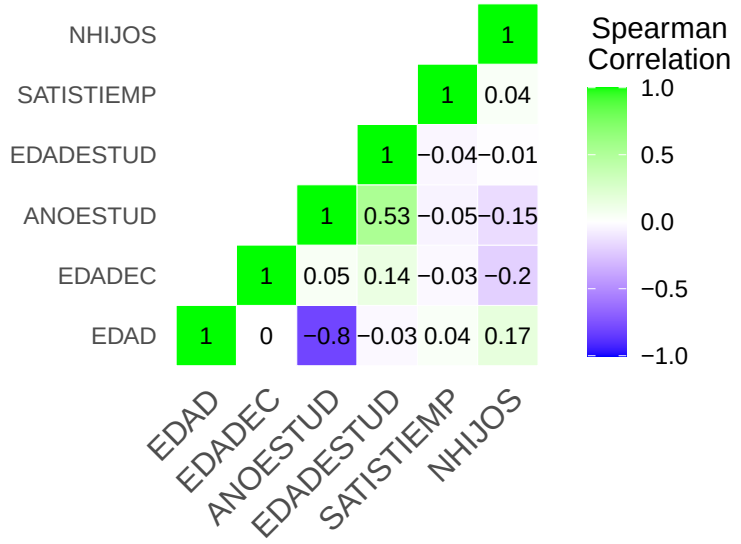

Figure 12. Mapa de correlaciones de Spearman para las variable numéricas

Como se puede observar, en el gráfico de la Figura 12, no parecen existir relaciones entre variables numéricas desconocidas hasta el momento. sin embargo, era capital comprobar las correlaciones de Spearman, ya que puede darse el caso de que variables que no presentan una relación lineal entre sí, sí presentasen una relación no lineal.

Variables categóricas

En el caso de las variables categóricas, hay diferentes formas de analizar la relación entre las variables. Se pueden estudiar variables a pares mediante la realización del test de chi-cuadrado. En otros casos, también es común el empleo de la V de Cramer. Como alternativa a estos estadísticos, se puede emplear la tau de Goodman y Kruskal.

Calculamos con el test de chi-cuadrado entre las variables `ESTUDIOS` y `LUGTRAB`, permitiéndonos evaluar la independencia entre estas dos variables categóricas.

```
Se rechaza la hipótesis nula de independencia entre las variables

con un nivel de significancia de 0.05. Hay cierta asociación entre ellas.
```

En el caso de la tau de Goodman y Kruskal [1], el valor de tau es asimétrico (hecho que no sucede con otros estadísticos como es la correlación), debiéndose esto a que dicho estadístico se basa en la fracción de variabilidad de la categoría y que puede ser explicada por la variable categórica x. Teniendo en cuenta lo anterior, se ha tratado de representar un mapa con este estadístico para las variables categóricas.

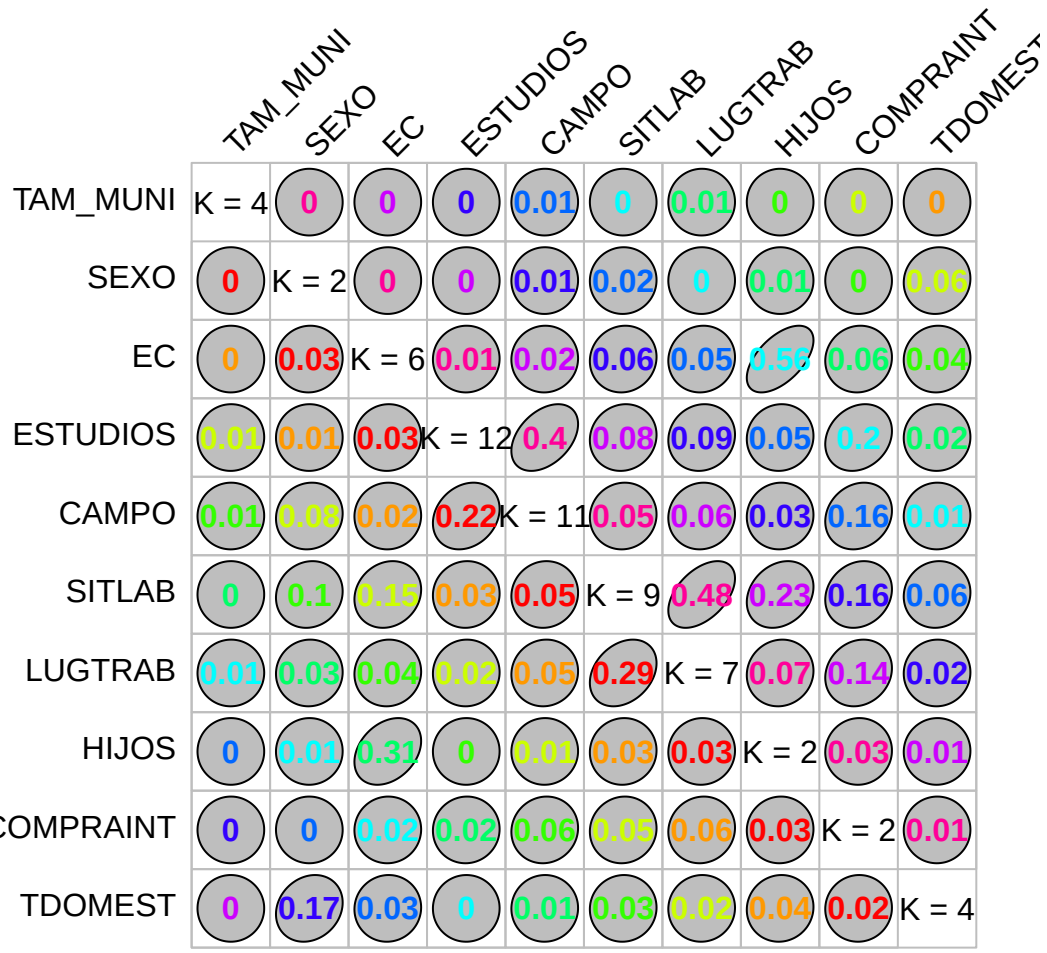

Figure 13. Mapa de GoodManKruskal para las variables categóricas

Al ser un mapa asimétrico, la forma de interpretar los resultados de la Figura 13 es un poco peculiar. Del resultado obenido podemos sacar las siguientes conclusiones:

- El nivel de estudios de una persona es algo influyente en el campo de trabajo (0.4). En cambio, el campo de trabajo no es tan influyente en los estudios de una persona (0.22).
- El estado civil de una persona es ciertamente influyente en si tendrá o no hijos (0.56).
- La situación laboral de una persona es algo influyente en su lugar de trabajo (0.48). En cambio, el lugar de trabajo de una persona, no es tan influyente en su situación laboral.

La diagonal del mapa anterior, es decir los valores de K, aportan información acerca del número de valores distintos de cada factor.

2.3. Búsqueda de outliers multivariante

Se hace uso de un algoritmo de clustering para detectar instancias atípicas dentro del conjunto de datos. Este es un algoritmo no supervisado que genera grupos entre las instancias, obteniendo una serie de centroides, pudiendo identificar a una instancia como outlier bien por estar a una distancia elevada de su centroide o bien por el hecho de que conforme un grupo pequeño y aislado del resto.

Hay que tener en cuenta los siguientes puntos en la aplicación de este algoritmo de ML:

- Es sensible a la inicialización: lo que implica que hemos de fijar la semilla generadora de números aleatorios para garantizar la reproducibilidad del cuaderno.
- Se ha de parametrizar el número de núcleos, del que dependen directamente los resultados obtenidos.

En nuestro problema, realizamos dos aproximaciones, una primera donde solo se capturan subconjuntos de variables numéricas, seguida de otra donde se capturan un subconjunto mayor de variables, tomando variables tanto numéricas como categóricas.

2.3.1. Clustering con atributos numéricos

Para la primera aproximación, los resultados se muestran en la Figura 14. Seleccionaremos las variables numéricas `EDAD`, `SATISTIEMP` y `ANOESTUD`.

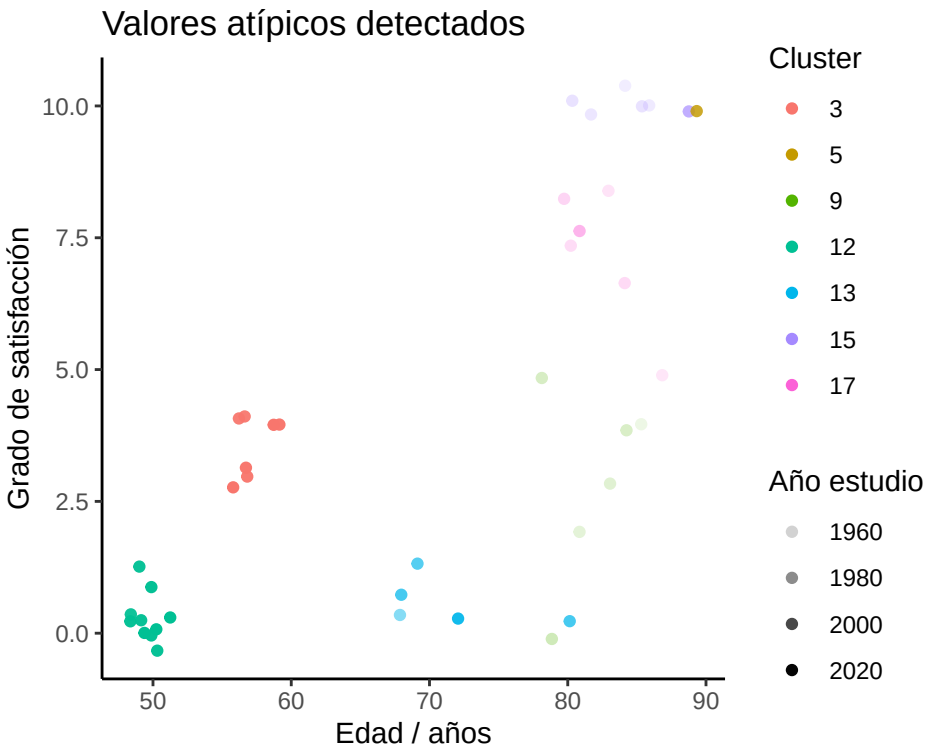

Figure 14. Clustering en variables numéricas de EDAD, SATISTIEMP y ANOESTUD.

Como se observa en el gráfico, en este conjunto de datos aparecen valores bastante peculiares, como el hecho de que existan personas de casi 80 años con estudios recientes. Asimismo, tampoco hay evidencia de que estos datos sean incorrectos, como sucedería si se encontraran con jóvenes con estudios anteriores a sus nacimientos.

2.3.2. Clustering con variables categóricas y numéricas

En cuanto al tratamiento con variables categóricas, este es similar. En esta ocasión es necesario aplicar One-Hot Encoding a las variables para transformarlas en numéricas, al ser requerido por el algoritmo de clustering.

Destaca que el algoritmo es capaz de encontrar instancias claramente atípicas, como señores de más de 80 años de edad que puntúen de forma negativa sus desplazamientos al trabajo (lo que implica que siguen trabajando), que realicen compras en internet o que tengan estudios recientes terminados.

Table 7. Outliers encontrados en el conjunto de datos

	COMPRAINTSí	COMPRAINTNo	HIJOSNo	EDAD	SATISTIEMP	ANOESTUD	clust
75142	0	1	0	85	4	1954	8
60579	0	1	1	87	5	1958	8
102962	1	0	1	89	10	2009	18
20617	0	1	1	81	2	1958	8
13490	0	1	0	86	10	1960	8
60326	0	1	0	83	3	1960	8

3. Conclusión

A modo de conclusión, nos gustaría expresar lo que ha supuesto realizar este trabajo. Haber respondido a las preguntas particulares que teníamos sobre este dataset planteadas al inicio, ya sea el cómo se distribuye una variable numérica como la `EDAD`, una categórica como la situación laboral (`SITLAB`), buscar relaciones entre distintas variables identificando por su tipo (numérico-numérico, categórico-numérico, categórico-categórico) es ciertamente importante. Más aún es el haber podido acceder a métodos más poderosos como el clustering para nuestro análisis.

Todo ello nos ha aportado un dominio del conocimiento del problema en cuestión en poco tiempo sin necesidad de consagrarnos como expertos en área de estudio. Además, hemos aprendido cómo problemáticas como la falta de datos o los errores de los mismos pueden suponer verdaderos dolores de cabeza en el análisis y exploración de estos.

Sin embargo, consideramos que lo más fructífero de este asunto es el habernos enfrentado por primera vez a una parte de un problema real en Ciencia de Datos, algo que va más allá de lo que meramente se puede llegar a impartir en clase con ejemplos sencillos. Además, se han debido tomar ciertas decisiones en colectivo, lo que nos ha hecho ganar además competencias relacionadas con el trabajo en equipo, algo que con total seguridad nos será muy útil en el futuro sea donde sea que vayamos a parar.

Acknowledgments: Es necesario mostrar nuestro agradecimiento a ValgrAI por el apoyo en el pago de la matrícula para el Máster en Ciencia de Datos (UV). Además, no nos podemos olvidar del profesorado que nos permitirá formarnos en este ámbito.

Abbreviations

The following abbreviations are used in this manuscript:

AED	Análisis Exploratorio de Datos

References

1. Pearson, R. The GoodmanKruskal package: Measuring association between categorical variables. https://cran.r-project.org/web/packages/GoodmanKruskal/vignettes/GoodmanKruskal.html, 2020. Accessed: 2023-11-14.