MDPI

Article

Proyecto de AED - Inserción laboral de titulados universitarios

Joan Merlos Cremades[1,*], Héctor Torres Muñoz[2]

[1] Universitat de València; joanmer@alumni.uv.es
[2] Universitat de València; hecto2@alumni.uv.es
* Correspondence: joanmer@alumni.uv.es

Simple Summary: Proyecto de la asignatura Ánalisis exploratorio de datos del Máster en Ciencia de datos de la Universitat de València.

Abstract: En las siguientes páginas, se encuentra un estudio exploratorio de los datos correspondientes a la inserción laboral de los estudiantes universitarios que finalizaron su grado durante el curso 2013-2014, dichos datos fueron recogidos por el Instutito Nacional de Estadística (INE) y se pueden descargar de forma libre accediendo al siguiente enlace https://www.ine.es/ftp/microdatos/eilu/datos_2019.zip. Dicho análisis consistirá en transformar los datos a sus correspondientes tipos, renombrar los factores para que tengan nombres representativos, estudiar las posibles relaciones entre las diferentes variables e imputar variables de interés.

Keywords: Análisis exploratorio de datos, INE, Inserción laboral.

Citation: . Proyecto de AED - Inserción laboral de titulados universitarios. *Journal Not Specified* **2024**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

1. Introducción

En la carpeta de datos original, se pueden observar dos archivos de datos CSV diferentes. El archivo que se va a proceder a analizar es el de la gente que se graduó en una universidad durante el curso 2013-2014 (**EILU_GRAD_2019.csv**). El otro archivo corresponde a la gente que obtuvo un máster en el curso 2013-2014, por lo que no será objeto de nuestro estudio. Habiendo aclarado ya el conjunto de datos a utilizar podemos proceder a importar las librerías que nos permitirán analizarlo, limpiarlo y estudiarlo.

```
pacman::p_load(readr,dplyr,ggplot2,vcd,patchwork,knitr)
```

2. Importación y limpieza de datos

En este apartado vamos a proceder a importar el archivo de datos con el que vamos a trabajar. Para ello primero debemos de adivinar su codificación mediante el uso de la función *guess_encoding*.

```
guess_encoding("./data/EILU_GRAD_2019.csv")
```

```
## # A tibble: 1 x 2
##   encoding confidence
##   <chr>         <dbl>
## 1 ASCII             1
```

Una vez que sabemos que la codificación de nuestros datos es ASCII, podemos proceder a importarlos teniendo en cuenta que estan en formato CSV y que los diferentes valores se encuentran separados por tabulador.

```
df_raw <- read_delim("./data/EILU_GRAD_2019.csv",
    delim = "\t", escape_double = FALSE,
    trim_ws = TRUE, locale = locale(encoding = "ASCII"))
```

El dataframe importado tiene una longitud de 299 variables, así que para falicitar nuestro estudio nos quedaremos únicamente con las variables principales que afectan a los estudios, la situación laboral y el sueldo de los diferentes individuos. Para conocer que representa cada una de las distintas variables debemos acudir al *codebook* de los metadatos, este viene dado por el archivo xlsx denominado **dr_EILU_GRAD_2019.xlsx**. Cabe destacar que en el proyecto de R donde se encuentra este documento pdf también hemos añadido dicho archivo para que cualquier persona pueda identificar las variables sin necesidad de abrir el archivo original de metadatos.

```
var <- c('IDENT','SEXO','RAMA','EST_B17_M1','EST_B17_M2','EST_B17_M3',
         'EST_B21','IDIOMAS','TRBPRN1','TR_SUELDO','HATR_E11')

df <- df_raw[var]
```

Una vez que ya tenemos seleccionadas las variables con las que vamos a realizar nuestro estudio debemos pasar cada una de ellas a su respectivo tipo. Dicha tarea se vuelve un poco tediosa debido a que las diferentes variables de tipo factor han sido codificadas mediante índices numéricos. Para saber que simboliza cada uno de estos números debemos, otra vez, acudir al *codebook* de los metadatos. Una vez que podemos comprender que significa cada número debemos convertir las diferentes columnas en tipo factor mediante la función *factor* introduciendo en el parámetro *levels* los valores codificados correspondientes a cada nivel del factor y en la parámetro *labels* el nuevo nombre que queremos asociar a cada nivel del factor, el cual debe ser lo suficientemente indicativo para poder interpretarlo más adelante en las distintas gráficas exploratorias que se realizaran. Por otro lado, para modelizar los datos faltantes hemos utilizado el tipo NA. Cabe destacar que todo este proceso se encuentra en el documento Rmd asociado, ya que es algo extenso y consideramos redundante introducirlo en este informe.

Antes de pasar al siguiente apartado, vamos a intentar convertir todas las variables que nos indican la rama de los estudios realizados por cada individuo en una sola. La variable que nos aporta la rama del primer estudio universitario realizado por el individuo se denomina *RAMA*. No osbtante, después tenemos más variables que nos indican la rama de los másteres realizados por el individuo o la rama del doctorado más importante. Para reducir todas estas múltiples variables a una sola vamos a crear la variable *URAMA*, la cual nos va a indicar la rama del estudio más elevado realizado por el individuo, en la mayoría de casos dichos estudios coincidiran con los últimos realizados por el individuo. Cabe destacar que vamos a ignorar la variable dedicada a los estudios del del tipo MIR, ya que todo aquel que se encuentre realizándolo o ya lo haya realizado necesariamente debe haber realizado una carrera o máster de la rama de ciencias de la salud. Por tanto, nos es redundante conocer esa información para crear la columna con la rama del último estudio realizado.

Para crear la variable *URAMA* simplemente debemos tener en cuenta que cuando el individuo no ha dado información sobre el estudio o directamente no ha realizado dicho estudio la columna aparecera como NA, así que vamos estudiando que columnas continen datos no faltantes empezando por los doctorados, siguiendo por los másteres y terminando en los grados o licenciaturas.

```
df <- df %>%
  mutate(URAMA = ifelse(!is.na(EST_B21), EST_B21,
    ifelse(!is.na(EST_B17_M3), EST_B17_M3,
        ifelse(!is.na(EST_B17_M2), EST_B17_M2,
          ifelse(!is.na(EST_B17_M1), EST_B17_M1,
                 RAMA)))))

codi_urama <- c(1, 2, 3, 4, 5)
desc_urama <- c("Artes y humanidades", "Ciencias",
```

```
                "Ciencias sociales y jurídicas", "Ingeniería y arquitectura",
                "Ciencias de la salud")

df$URAMA <- factor(df$URAMA, levels=codi_urama, labels=desc_urama)
```

Una vez se ha obtenido la columna *URAMA* podemos eliminar del *data frame* las variables con información sobre las ramas de los másteres o la columna que contiene la rama del doctorado más importante.

```
df <- df %>% select(-c('EST_B17_M1','EST_B17_M2','EST_B17_M3','EST_B21'))
```

3. Inspección de datos faltantes

En este apartado, vamos a proceder a estudiar los datos faltantes en cada columna del *data frame* con información, es decir, vamos a obviar la columna con el identificador.

Table 1. Cantidad de NA's en función de la variable

Variable	Número de datos faltantes
SEXO	0
RAMA	0
IDIOMAS	271
TRBPRN1	0
TR_SUELDO	5266
HATR_E11	914
URAMA	0

Como podemos observar las únicas columnas con datos faltantes son *IDIOMAS*, *TR_SUELDO* y *HATR_E11*. La variable con más datos faltantes es TR_SUELDO con un `16.64` % de datos con NA. No obstante, no todos ellos son datos faltantes, ya que puede darse el caso de que la persona no esté actualmente trabajando y, por tanto, haya evitado rellenar a drede la casilla de *TR_Sueldo*. Para observar cuales son en realidad nuestros datos faltantes procedemos a observar el número de NA's de la variable *TR_SUELDO* una vez hemos eliminado aquellas personas que no se encuentran trabajando.

Table 2. Cantidad de Na's en la variable TR_SUELDO cuando se elimina la gente que no está activa laboralmente

Variable	Cálculo	Número de datos faltantes
TR_SUELDO	Con todos los datos	5266
TR_SUELDO	Con la gente activa laboralmente	739

Como podemos observar hay aproximadamente 4500 valores NA que pertenecen a la población no activa laboralmente, la cual es lógico que no tenga un salario. No obstante, los otros 700 datos con NA no parecen seguir ninguna lógica. Por ello intentaremos relacionar la variable *TR_SUELDO* con alguna otra variable del conjunto que no contenga datos faltantes.

4. Imputación de datos faltantes

Queremos imputar los valores de la variable con el sueldo para aquellas personas que se encuentran activas, para ello lo que vamos a hacer va a ser primero aplicar el test del χ^2 a la combinación de la variable *TR_SUELDO* con el resto de variables, evitando las variables IDENT porque no tiene sentido relacionarla con el identificador y la variable *TRBPRN1* porque, si eliminamos los NA's de la variable *TR_SUELDO*, aquellas personas que no estan

laboralmente activas se eliminan y no tiene sentido realizar la tabla de contingencia cuando dos columnas valen siempre 0.

Table 3. p-valor y V de Cramer para el sueldo

Variable 1	Variable 2	p-valor	V de Cramer
TR_SUELDO	SEXO	1.94e-192	1.87e-01
TR_SUELDO	IDIOMAS	1.37e-34	4.30e-02
TR_SUELDO	HATR_E11	6.99e-195	8.17e-02
TR_SUELDO	URAMA	1.33e-322	1.24e-01

Cuando el valor de p obtenido en dicho test es menor que 0.05 podemos concluir que hay una relación entre las dos variables categóricas sobre las que se ha aplicado el test, por tanto podemos decir que todas las variables se encuentran relacionadas en mayor o menor medida. La significancia estadística de dicha relación nos la de el valor de la V de Cramer, dicho estadístico consiste en una normalización del parámetro χ^2 de modo que un mayor valor de la V de Cramer quiere decir que hay una mayor significancia estadística entre las variables. Ante dicho descubrimiento podríamos pensar que utilizar la variable *SEXO* es lo mejor. No obstante, esto no tiene ningún sentido, ya que la cantidad del salario de una persona no debería depender únicamente de su respectivo sexo. Ante este hecho vamos a proceder a utilizar la variable *URAMA* para realizar la imputación de los datos, ya que es la que tiene mayor significancia estadística después de la variable *SEXO*.

A continuación se realiza la imputación condicional de la variable *TR_SUELDO* empleando la moda, el objetivo es completar los valores faltantes en la variable, la cual corresponde con el rango salarial de las personas encuestadas. Para garantizar la coherencia estadística se realiza la imputación solamente en individuos que constan como "Trabajando", además se emplea la variable *RAMA* de manera condicional, usando como criterio de sustitución la moda salarial de cada rama.

Para poder imputar utilizando la rama deberíamos garantizar que los datos son MAR (*Missing at random*), es decir, que se han perdido de forma completamente aleatoria. Para poder observar si se cumple dicho propósito vamos a realizar una tabla con la frecuencia de datos faltantes en cada tipo del factor *URAMA*.

Table 4. Porcentaje de valores faltantes de la gente activa laboralmente en TR_SUELDO en función de la rama de estudio

	No faltantes	Faltantes	Total casos	% Faltantes
Artes y humanidades	2352	81	2433	3.33
Ciencias	2368	70	2438	2.87
Ciencias sociales y jurídicas	11951	301	12252	2.46
Ingeniería y arquitectura	5310	157	5467	2.87
Ciencias de la salud	4404	130	4534	2.87

Como podemos observar todas las columnas presentan una proporción de aproximadamente un 3 % de datos faltantes, esto nos garantiza que los datos faltantes están esparcidos de forma aleatoria dentro de nuestro conjunto. Por tanto, podemos aplicar una imputación del tipo moda sin preocuparnos por afectar en exceso a la estructura de los datos.

En primer lugar convertimos las variables *RAMA* y *TR_SUELDO* en factores ordenados. Esto es obligatorio ya que ambas variables son categóricas y presentan un número finito de niveles. Además, destacar la conversión de *TR_SUELDO* a factor ordinal ya que los rangos salariales deben tener un orden lógico, en este caso de menor a mayor. Esto permitirá ordenar correctamente las categorías y realizar comparaciones válidas entre niveles.

A continuación, se calcula la moda salarial, la cual corresponde con el nivel de sueldo más frecuente dentro de cada rama de estudio. Para calcularla se han considerado únicamente observaciones de los individuos que se encuentran trabajando y cuyo valor de *TR_SUELDO* no es NA.

El cálculo de la moda se lleva a cabo usando las funciones *group_by* que agrupa los registros por rama, *count* que contabiliza la frecuencia de cada nivel salarial dentro del grupo, y *slice_max* que selecciona el valor más frecuente (la moda) en cada rama, todas ellas pertenecen a la librería *dplyr*. El resultado es una tabla auxiliar que asocia a cada rama su correspondiente moda salarial.

Cuando la moda salarial de cada rama está calculada, se puede comenzar la imputación. Empleando la función *left_join()*, se une la tabla de modas al conjunto de datos original a través de la variable *URAMA*. Luego, usando *mutate()* se reemplazan los valores NA con las condiciones *ifelse(TRBPRN1 == "Trabajando" & is.na(TR_SUELDO)*, es decir, que esté trabajando y que el valor sea NA, que es lo que se debe imputar. Tras la imputación, *TR_SUELDO* se reconvierte nuevamente en factor ordenado, asegurando que la variable conserva la estructura ordinal original y también las etiquetas.

Con el propósito de examinar la distribución *TR_SUELDO* antes y después de la imputación para ello recurrimos a la siguiente tabla.

Table 5. Número de datos en los diferentes niveles salariales de la variable TR_SUELDO

Nivel salarial	Datos sin imputación	Datos con imputación
Menos de 700 euros	1756	1756
De 700 a 999 euros	2524	2524
De 1.000 a 1.499 euros	8415	8997
De 1.500 a 1.999 euros	7887	8044
De 2.000 a 2.499 euros	3442	3442
De 2.500 a 2.999 euros	1212	1212
De 3.000 euros en adelante	1149	1149
NA's	5266	4527

Se oberva una imputación satisfactoria con reducción de aproximadamente 700 casos de NA en trabajador. No obstante, es necesario destacar todavía la presencia de 4527 valores faltantes, los cuales corresponden, como ya se ha comentado anteriormente, con NA de personas que no trabajan.

5. Gráficas exploratorias

En este apartado, vamos a proceder a realizar diferentes gráficas de barras, ya que estamos tratando con datos categóricos, para observar como se separan los datos en función de las variables que utilicemos. Cabe destacar que en todos los casos vamos a proceder a utilizar el color para expresar otra variable dentro de un mismo gráfico, realizando así un análisis bivariante.

Para empezar, vamos a hacer dos gráficos de barras con la variable RAMA y URAMA para observar si hay mucha discrepancia entre ambas variables.

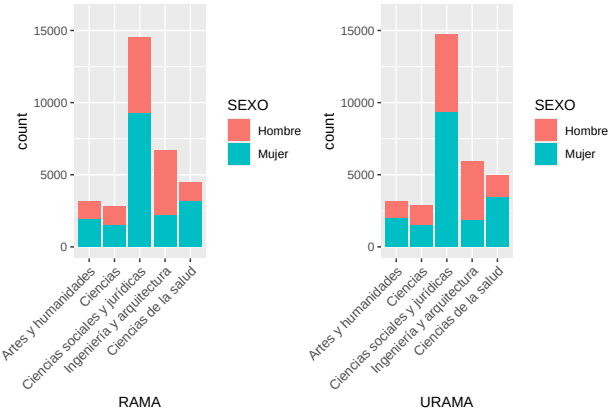

Figure 1. Gráfico de barras de RAMA/URAMA coloreado con SEXO

Como podemos observar los dos gráficos son muy similares, lo cual tiene sentido, ya que la gente suele realizar estudios que tengan relación con sus estudios ya completados.

Ahora graficamos el tiempo transcurrido hasta encontrar trabajo en función del tipo de RAMA del estudio más importante realizado.

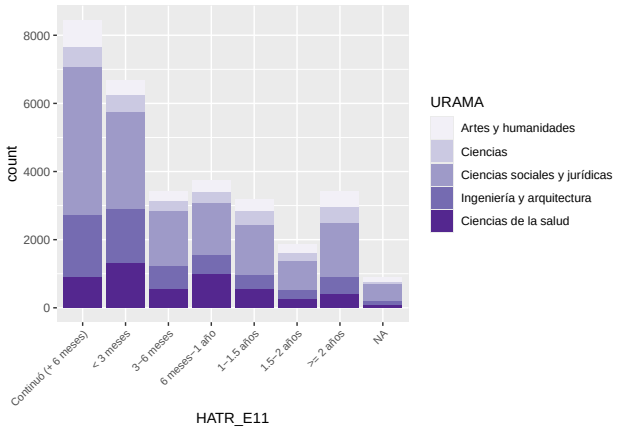

Figure 2. Gráfico de barras del tiempo transcurrido hasta que empezo a trabajar coloreado con el tipo de RAMA del estudio más importante realizado

La anterior gráfica nos permite apreciar como la mayor parte de la gente tiende a continuar en el mismo trabajo que tenía después de finalizar la universidad. Podemos destacar que la gráfica presenta una tendencia descendiente, de modo que cuanto más tiempo pasa es menos probable que la gente se encuentre sin trabajo. No osbtante, observamos un aumento anómalo en la última columna con datos no vacíos, lo cual podría estar relacionado con la realización de estudios de post-grado como másteres o doctorados donde uno no está trabajando directamente pero ya ha finalizado su grado o licenciatura. Por otro lado, no parece haber una relación entre el tipo de la rama y el tiempo que se tarda en encontrar trabajo, ya que todas las columnas presentan proporciones similares de todos las ramas.

Para proseguir vamos a graficar la rama del estudio más elevado en función de la situación laboral.

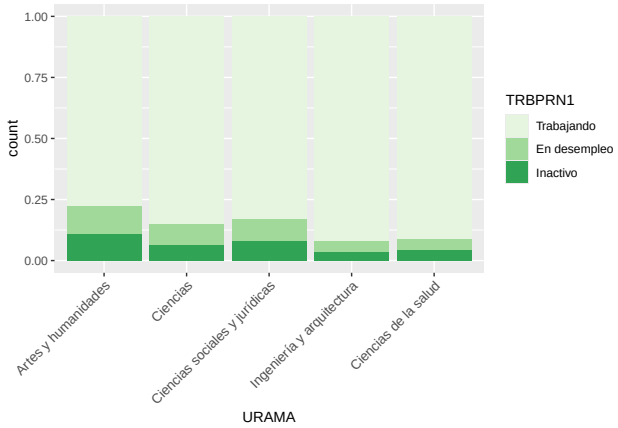

Figure 3. Gráfico de barras normalizado de la rama del estudio más elevado coloreado con la situación laboral

Como podemos observar en la anterior gráfica, hay una cantidad mayor de gente que no se encuentra trabajando en las ramas de Artes y humidades y Ciencias sociales y ciencias jurídicas.

A continuación, vamos a proceder a realizar diferentes gráficos de barras del número de idiomas en función de la situación laboral, los cuales además estarán coloreados por el sexo.

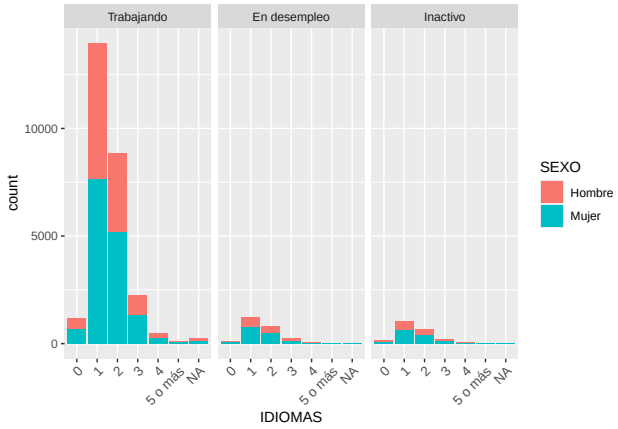

Figure 4. Gráficos de barras del número de idiomas en función de la situación laboral y coloreado por el sexo

Observamos como en todos los casos la columna con mayor peso es 1, lo cual nos indica que la gente suele enseñarse un único idioma diferente a los idiomas maternos. Dicha relación tiene sentido, ya que actualmente la gente suele centrarse en estudiar el inglés y dejar el resto de idiomas más de lado. También se puede notar que la segunda columna con más cantidad de datos es la de dos idiomas, dentro de la cual predomina el sexo femenino. Por otro lado, no parece haber una relación directa entre la cantidad de idiomas diferentes del materno que conoce el individuo con la situación laboral que este ostenta.

En primer lugar vamos a gráficar la variación del sueldo según la rama del estudio más elevado realizado.

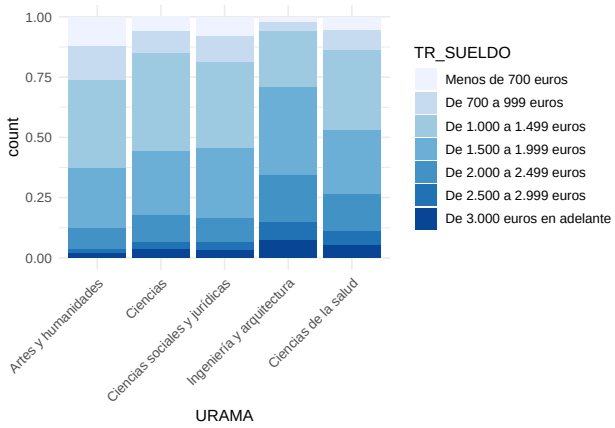

Figure 5. Variación del sueldo según la rama de estudio

El propósito de esta gráfica es analizar la distribución de los niveles salariales, permitiendo identificar diferencias claras en la distribución entre ramas. Las ramas de Ingeniería y arquitectura y Ciencias de la salud presentan una mayor proporción de niveles salariales altos, lo que sugiere una mejor inserción laboral o remuneración media más elevada. En cambio, Artes y humanidades y Ciencias concentran una mayor proporción de sueldos en los tramos bajos o medios, reflejando una estructura salarial menos favorable. La rama de Ciencias sociales y jurídicas se sitúa en una posición intermedia, mostrando una distribución equilibrada entre niveles salariales medios y altos.

Seguidamente, podemos representar la rama del último estudio realizado coloreada con el sexo de los individuos.

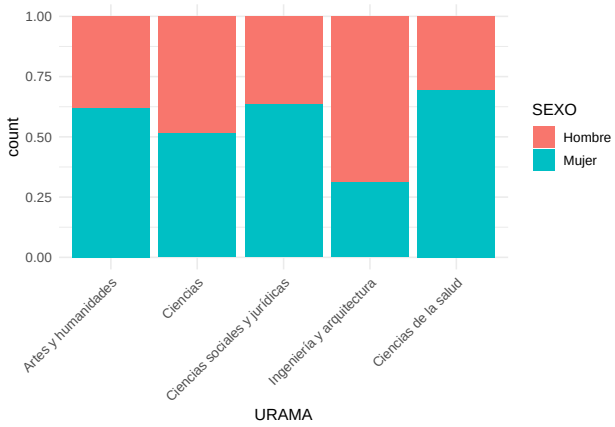

Figure 6. Distribución de hombres y mujeres en cada rama de estudio

El objetivo de esta gráfica es mostrar la distribución proporcional del alumnado por sexo dentro de cada rama de estudio. De esta forma, se puede observar la representación relativa de hombres y mujeres en las diferentes áreas académicas y detectar posibles desequilibrios de género en la elección de estudios universitarios. De manera general, las ramas de Ciencias de la salud, Artes y humanidades y Ciencias sociles y jurídicas, muestran una mayor proporción de mujeres, mientras que Ingeniería y arquitectura presenta una clara predominancia masculina. Por otro lado, la rama de Ciencias es la más equilibrada.

A continuación, podemos hacer una comparación de la distribución salarial según el sexo.

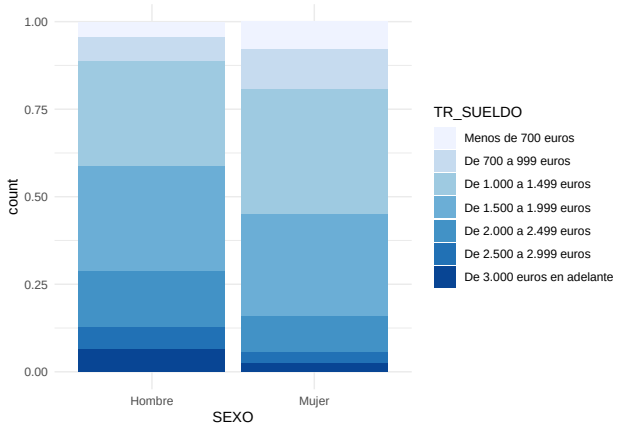

Figure 7. Distribución salarial según el sexo

El objetivo de esta gráfica es identificar si existen diferencias significativas en la estructura de los ingresos entre hombres y mujeres dentro de la población laboral activa analizada, evidenciando las disparidades de distribución entre hombres y mujeres. Aunque las proporciones pueden variar según el contexto específico del estudio, la tendencia reflejada sugiere una mayor concentración de mujeres en los tramos salariales bajos o intermedios, mientras que los hombres presentan una mayor proporción en los rangos salariales altos.

Finalmente, procedemos a representar la relación entre número de idiomas y rango salarial.

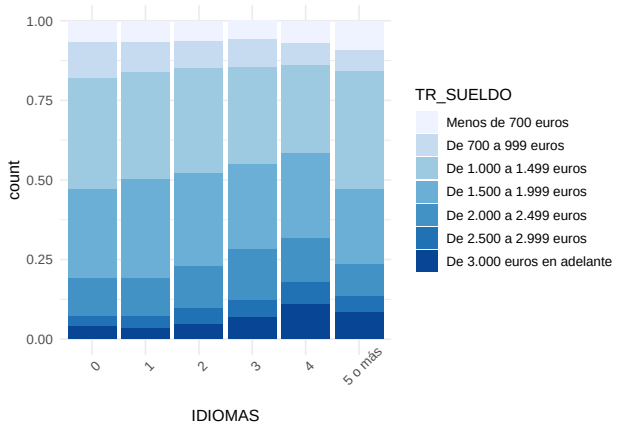

Figure 8. Distribución salarial según el número de idiomas aprendidos

El objetivo de esta gráfica es analizar la relación entre el número de idiomas conocidos; *IDIOMAS*, y el nivel salarial, *TR_SUELDO*, entre las personas que se encuentran trabajando. A través de esta representación se pretende identificar si existe una tendencia ascendente en la remuneración a medida que aumenta el número de idiomas. De manera general, se observa una tendencia ascendente; no obstante, el gráfico también muestra un retroceso al alcanzar cinco idiomas o más.

6. Conclusión

A lo largo de este proyecto hemos aprendido diferentes cosas tanto del conjunto de datos con el que estábamos tratando como del análisis exploratorio de datos en general. Nuestros objetivos desde un principio fueron utilizar el conjunto de datos proporcionado por el INE para estudiar la inserción laboral en función de las diferentes ramas del conocimiento e intentar encontrar alguna relación del sueldo con la rama de estudio asociada.

Lo primero que aprendimos fue como tratar con un conjunto de datos real, ya que tuvimos que reducir las variables de nuestro *data frame* y quedarnos solo con aquellas que coincidían con nuestros objetivos. Esto nos permitió pasar de un conjunto de datos con 299 a uno con 12. Cabe destacar que de estas 12 variables iniciales conseguimos combinar aquellas que se relacionaban con la rama de los diferentes estudios en una sola, lo que nos redujo el conjunto a 8 variables.

Además de aprender como tratar con las diferentes variables también tuvimos que realizar una reconversión del tipo de nuestros datos, ya que nuestras variables que eran del tipo factor se encontraban codificadas y en formato cadena. Para poder empezar a trabajar con dichas variables tuvimos que utilizar el *codebook* proporcionado por el INE para decodificar las variables y transformarlas en un factor cuyos niveles fueran fácilmente interpretables.

Seguidamente, nos dispusimos a estudiar la presencia de datos faltantes a lo largo de nuestras variables. En dicho proceso observamos que la variable que contenia el sueldo de los individuos poseía una gran cantidad de valores faltantes. A pesar de que gran parte de estos provenían de las personas que no se encontraban activas laboralmente, todavía seguía quedando una pequeña parte que no conseguíamos explicar. Dichos datos fueron tratados como MAR después de comprobar que se esparcían de forma similar entre las diferentes ramas del conocimiento, lo que nos permitió imputarlos utilizando la moda de cada nivel de *URAMA*. Para realizar dicho proceso antes comprobamos con el test χ^2 y la V de Cramer que las variables *TR_SUELDO* y *URAMA* tenían una relación que era estadísticamente significativa.

Para finalizar con nuestro proyecto realizamos diferentes gráficos exploratorios en los que pudimos sacar diferentes conclusiones. Los resultados confirman que existen diferencias significativas en la empleabilidad y el salario según la rama de estudio, siendo Ingeniería y Arquitectura y Ciencias de la Salud las áreas con mejores indicadores de inserción y con mayor concentración de sueldos elevados. En cambio, las ramas de Artes y Humanidades y Ciencias Sociales y Jurídicas presentan una estructura salarial menos favorable y mayores tasas de precariedad laboral. Asimismo, se observa una brecha salarial por género, donde los hombres se concentran en los tramos de mayor remuneración y las mujeres en niveles medios o bajos, incluso en ramas con predominio femenino. También se constata una cierta influencia del número de idiomas, que se asocia positivamente con salarios más altos, aunque de forma no lineal.

7. Bibliografía

[1] Pérez, C. *Técnicas de análisis multivariante de datos: Aplicaciones con SPSS*. Pearson Educación, 2004.

[2] Provost, F.; Fawcett, T. *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, 2013.