Práctica 2: Limpieza y análisis de datos

Dataset: Income of Adults

04 de enero de 2022

Contenido

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

El conjunto de datos objeto de análisis se ha obtenido a partir del dataset de kaggle [1]. Este dataset llamado ***"Income of Adults"*** recoge diferentes métricas para ayudar a predecir los ingresos de las personas (si su ingreso es mayor o menor de 50k) según su grupo demográfico. Está constituido por 15 características (columnas) que presentan 32.562 encuestados (filas).

Los campos de este conjunto de datos son los detallados en la siguiente tabla:

Tabla 1. Análisis descriptivo de las columnas del dataset a tratar.

Campo	Tipo	Descripción	Ejemplo
age	Numérico	Edad de la persona	39
workclass	Texto	Clase de empleado. Campo categorizado.	State-gov
fnlwgt	Numérico	Peso final dado por la oficina que ha recogido los datos y que da el número de unidades en la población objetico en base a una fórmula no facilitada.	77516
education	Texto	Nivel de estudios. Campo categorizado.	Bachelors
education-num	Texto	Años dedicados a los estudios.	13
marital-status	Texto	Situación sentimental. Campo categorizado.	Never-married
occupation	Texto	Trabajo actual.	Adm-clerical
relationship	Texto	Relación familiar actual. Campo categorizado.	Not-in-family
race	Texto	Raza. Campo categorizado.	White
sex	Texto	Sexo. Campo categorizado.	Male
capital-gain	Numérico	Ganancia de capital.	2174
capital-loss	Numérico	Pérdida de capital.	0
hours-per-week	Numérico	Horas de trabajo por semana.	40
native-country	Texto	Nacionalidad. Campo categorizado.	United-States
salary	Lógico	Salario.	<=50K

El dataset original se comparte bajo la licencia CC0: Public Domain. La fuente de los datos es la Oficina del Censo de EE. UU (The U.S. Census Bureau), último censo disponible, año 2020.

Fue creado en mayo de 2021, y se espera que se actualice anualmente, la versión actual es la 1ª.

Atendiendo a la documentación subida al repositorio, Kaggle le ha asignado una usabilidad de 9.1 sobre 10.

¿Por qué es importante y qué pregunta/problema pretende responder?

A partir del análisis de este conjunto de datos, se pretende dar respuesta a una serie de preguntas relacionadas con la vida laboral y en particular de los ingresos de la población de Estados Unidos.

Inicialmente se escoge este dataset para ver si se puede estudiar qué salario puede alcanzar una persona en base a su pertenencia a un grupo demográfico, con las características descritas en el conjunto de datos.

Las personas que no alcanzan un determinado umbral de sueldo tienen una peor calidad de vida, se encuentran con más dificultades para hacer frente a cualquier eventualidad negativa y esto provoca otras consecuencias (problemas de adicción, riesgos de problemas de salud, tanto físicos como mentales, etcétera). Determinados grupos demográficos pueden convertirse en colectivos más vulnerables de la sociedad. Factores como la raza, el sexo o el nivel de estudios, influyen en el acceso a la vida laboral y pueden ser determinantes en el momento de conseguir una mejor situación laboral.

Por eso, estudiar este tipo de datasets ayuda a determinar y remarcar los posibles sesgos sociales en la población.

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

Se inicia la práctica con la lectura del fichero de datos. Para ello, cargaremos los paquetes necesarios, cuyo número se ha ido incrementando conforme se avanzó en el desarrollo de la práctica.

Paquetes utilizados

- dplyr (manejo de dataframes)
- rmarkdown (informe dinámico)
- nortest (test lillie)
- ggplot2 (gráficas)
- gridExtra (formato gráficas)
- mlbench (regresión logística)
- Corrplot (correlation matrix)
- Tidyr (ordenación de datos)
- Caret (métodos de entrenamiento y clasificación)
- gbm (regresión logística)

Lectura del fichero

Se realiza la lectura con la función read.csv, indicando la coma como carácter separador y diciéndole que contamos con una primera fila de cabecera de columnas.

Dimensión del conjunto de datos

Comprobamos mediante la función dim() sí las dimensiones coinciden con las observadas en el estudio previo: 32.561 filas y 15 columnas.

Características de las variables leídas

Inspeccionamos la estructura del dataset y la de sus diferentes variables con la función str().

```
'data.frame':   32561 obs. of  15 variables:
 $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
 $ workclass     : chr  " State-gov" " Self-emp-not-inc" " Private" " Private" ...
 $ fnlwgt        : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...
 $ education     : chr  " Bachelors" " Bachelors" " HS-grad" " 11th" ...
 $ education.num : int  13 13 9 7 13 14 5 9 14 13 ...
 $ marital.status: chr  " Never-married" " Married-civ-spouse" " Divorced" " Married-civ-spouse" ...
 $ occupation    : chr  " Adm-clerical" " Exec-managerial" " Handlers-cleaners" " Handlers-cleaners" ...
 $ relationship  : chr  " Not-in-family" " Husband" " Not-in-family" " Husband" ...
 $ race          : chr  " White" " White" " White" " Black" ...
 $ sex           : chr  " Male" " Male" " Male" " Male" ...
 $ capital.gain  : int  2174 0 0 0 0 0 0 0 14084 5178 ...
 $ capital.loss  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ hours.per.week: int  40 13 40 40 40 40 16 45 50 40 ...
 $ native.country: chr  " United-States" " United-States" " United-States" " United-States" ...
 $ salary        : chr  " <=50K" " <=50K" " <=50K" " <=50K" ...
```

Para seguir con el análisis preliminar, podemos, por ejemplo, hacer una inspección de 10 observaciones

elegidas de manera aleatoria como ejemplo para ilustrar los datos con los que contamos. Aunque dichos valores ya se muestran en el estudio de la estructura, visualizar el dataset en columnas siempre es más claro.

	age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	native.country	salary
	<int>	<chr>	<int>	<chr>	<int>	<chr>	<chr>	<chr>	<chr>	<chr>	<int>	<int>	<int>	<chr>	<chr>
17401	51	Private	441637	HS-grad	9	Married-civ-spouse	Tech-support	Husband	White	Male	0	0	40	United-States	<=50K
24388	38	Private	115289	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-States	<=50K
4775	30	Private	53158	Assoc-acdm	12	Never-married	Tech-support	Not-in-family	White	Female	0	0	40	United-States	<=50K
26753	38	Private	172538	Bachelors	13	Married-civ-spouse	Tech-support	Husband	White	Male	0	1977	40	United-States	>50K
13218	21	Private	200973	Some-college	10	Never-married	Sales	Own-child	White	Female	0	0	20	United-States	<=50K
26109	34	Private	80058	Prof-school	15	Never-married	Exec-managerial	Own-child	White	Male	0	0	50	United-States	<=50K
29143	33	Private	90668	HS-grad	9	Never-married	Adm-clerical	Not-in-family	White	Female	0	0	40	United-States	<=50K
10539	66	Private	185336	HS-grad	9	Widowed	Sales	Other-relative	White	Female	0	0	35	United-States	<=50K
8462	38	Private	108140	Bachelors	13	Never-married	Other-service	Not-in-family	White	Male	0	0	20	United-States	<=50K
4050	28	Private	25955	Assoc-voc	11	Divorced	Craft-repair	Not-in-family	Amer-Indian-Eskimo	Male	0	0	40	United-States	<=50K

10 rows

Dado el estudio que se pretende hacer, y tras el estudio de las variables, se desestiman para la práctica la ganancia y pérdida de capital.

```r
datos <- dplyr::select (datos,
                        -capital.gain,
                        -capital.loss)

str (datos)
```

Seguimos con la inclusión de una variable secuencial como identificador único de cada fila. Además, realizamos una primera reordenación de las columnas.

```r
filas <- nrow (datos)

datos <- cbind (datos, 'id' = 1:filas)

names (datos)

datos <- subset (datos,
                 select = c(14,1,2,3,4,5,6,7,8,9,10,11,12,13))

names (datos)
```

```
 [1] "age"            "workclass"      "fnlwgt"
 [4] "education"      "education.num"  "marital.status"
 [7] "occupation"     "relationship"   "race"
[10] "sex"            "hours.per.week" "native.country"
[13] "salary"         "id"
 [1] "id"             "age"            "workclass"
 [4] "fnlwgt"         "education"      "education.num"
 [7] "marital.status" "occupation"     "relationship"
[10] "race"           "sex"            "hours.per.week"
[13] "native.country" "salary"
```

La segmentación por rangos de edad permite establecer nichos más concretos y a profundizar en el estudio. Este procedimiento, conocido como discretización (transformar variables numéricas en variables categóricas) ayuda a hacer el dataset más agradable, y lo hace más fácil de analizar y extraer conclusiones. Se establecen categorías de edad para el estudio de empleabilidad vs salario en las siguientes escalas:

- De 0 a 16 años → cat_0
- De 17 a 20 años → cat_1
- De 21 a 34 años → cat_2
- De 35 a 45 años → cat_3
- De 46 a 65 años → cat_4
- Mayores de 65 años → cat_5

```r
datos <- cbind (datos,
                age_cat = cut (datos$age,
                               breaks = c(0,16,20,34,45,65,150),
                               labels = c("cat_0",
                                          "cat_1",
                                          "cat_2",
                                          "cat_3",
                                          "cat_4",
                                          "cat_5"),
                               right = TRUE))

names (datos)
```

```
 [1] "id"             "age"            "workclass"      "fnlwgt"
 [5] "education"      "education.num"  "marital.status" "occupation"
 [9] "relationship"   "race"           "sex"            "hours.per.week"
[13] "native.country" "salary"         "age_cat"
```

Se ha observado que algunas variables tienen un espacio en blanco al inicio. Por lo que procedemos a eliminarlo con una función propia.

```r
ltrim <- function (x) {trimws(x,which = c("left"))}

datos$workclas <- ltrim (datos$workclass)
datos$education <- ltrim (datos$education)
datos$marital.status <- ltrim (datos$marital.status)
datos$occupation <- ltrim (datos$occupation)
datos$relationship <- ltrim (datos$relationship)
datos$race <- ltrim (datos$race)
datos$sex <- ltrim (datos$sex)
datos$salary <- ltrim (datos$salary)
```

Preprocesado – variables categóricas

Podemos combinar algunos valores categóricos para reducir la dispersión y facilitar los análisis. Especialmente útil en las variables en las que el reparto de las cuales no ayude o no sea relevante en los análisis, como en el caso del estado civil, dónde la diferencia entre separado y divorciado no aporta ninguna información relevante.

La información de la combinación se encuentra en la siguiente tabla.

Variable	Valor original	Valor después de combinación
marital.status	Married-AF-spouse	Married
	Married-civ-spouse	
	Married-spouse-absent	
	Separated	Not-Married
	Divorced	
	Never-Married	
workclass	Local-gov	Other-Govt
	State-gov	
	Self-emp-inc	Self-Employed
	Self-emp-not-inc	
	Without-pay	Not-Working
	Never-worked	
occupation	Farming-fishing	Blue-Collar
	Handlers-cleaners	
	Machine-op-inspct	
	Transport-moving	
	Priv-house-serv	Service
	Other-service	

3. Limpieza de los datos.

3.1 ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Tras un análisis visual, y utilizando rstudio, se localizan datos perdidos. A continuación, pasamos a la revisión de las variables para su análisis y limpieza.

Tras el análisis de las distintas variables hemos encontrado valores vacíos, representados por '?', en las siguientes variables

- workclass
- occupation
- country

Procedemos a sustituirlos.

```
```{r}
datos [datos == "?"] <- NA
```
```

Hay un total de 3.679 valores vacíos. Procedemos a su identificación por columnas.

```
```{r}
apply (is.na (datos),
       2,
       sum)
```
```

```
            id            age      workclass         fnlwgt      education 
             0              0           1836              0              0 
 education.num marital.status     occupation   relationship           race 
             0              0           1843              0              0 
           sex hours.per.week native.country         salary        age_cat 
             0              0              0              0              0 
```

Su porcentaje es el siguiente.

```
```{r}
apply (is.na (datos),
       2,
       mean)
```
```

```
            id            age      workclass         fnlwgt      education  education.num 
    0.00000000     0.00000000     0.05638647     0.00000000     0.00000000     0.00000000 
 marital.status     occupation   relationship           race            sex hours.per.week 
    0.00000000     0.05660146     0.00000000     0.00000000     0.00000000     0.00000000 
 native.country         salary        age_cat 
    0.00000000     0.00000000     0.00000000 
```

Estos valores representan una proporción reducida de valores dentro del total en el dataset, un 13.06% valores son valores perdidos. Fraccionado por columnas, tendríamos que la columna workclass tiene 5.6% de valores perdidos, la columna occupation tendría 5.66% y native.country un 1.8%.

Los campos workclass y occupation incluyen categorías para todo tipo de casos en los que se pueda encontrar un individuo, por ejemplo, workclass incluye una categoría de otros. Por tanto, consideramos que se trata, o bien de errores en la medición o de campos donde el individuo se ha negado a cumplimentarlo. Cualquiera de estas situaciones supone una dificultad añadida para los análisis ya que son individuos que no se pueden clasificar en ninguna categoría.

Por esto, y viendo que estos valores representan una fracción muy reducida en el volumen total de datos, decidimos eliminarlos.

En el caso del país de origen, no vemos que haya una categoría de otros, y debido a que desconocemos los motivos por los cuales este campo se ha dejado vació (puede que el individuo tema a posibles

represalias o simplemente no lo sepa), procedemos a crear una categoría llamada “Unknown” dónde se sitúen aquellos individuos sin país de origen definido.

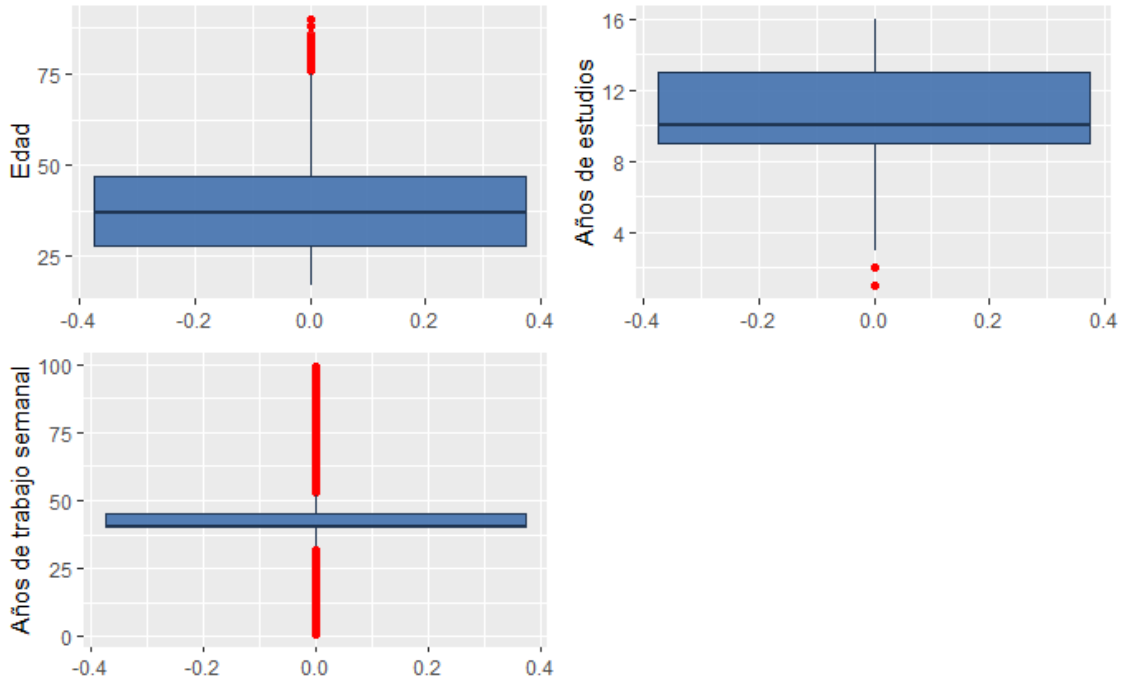
```{r}
datos$native.country [is.na (datos$native.country)] <- "Unknown"
```

Una vez, eliminados la dimensión del dataset ha cambiado. Ahora tenemos, 30.718 filas y 15 columnas.

3.2 Identificación y tratamiento de valores extremos.

Los valores extremos o outliers son aquellos que parecen no ser congruentes sin los comparamos con el resto de los datos. Para identificarlos, podemos hacer uso diferentes vías, nosotros nos decantamos por representar un diagrama de caja por cada variable y ver qué valores distan mucho del rango intercuartílico (la caja).

Tratamos las variables numéricas.



Podemos identificar los valores extremos por variables, y observamos que:

- En la variable “edad” los outliers adquieren valores de entre 77 a 90 años. Se corresponden con edades en las que se podría asumir que la persona ya no está en trabajo activo.
- En la variable “años de educación” los outliers adquieren valores de entre 1 a 2 años de estudios. Asumimos que son individuos que no tienen estudios.
- En la variable horas de trabajo semanal es la más dispar y los outliers van desde 0 a 47 horas y de 50 a las 100 horas. Estos valores, aunque alejados de la media, y considerados extremos, entran dentro del rango posible de horas de trabajo semanal, y, por tanto, son datos relevantes para el proyecto analítico.

Concluimos que no podemos considerar estas medidas como no validas ya que, no parece que provengan de datos erróneos. Aunque sean poco probables, son valores extremos que pueden pertenecer a la población muestreada.

4. Análisis de los datos.

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

El estudio que interesa es la relación de las distintas variables con el umbral de salario especificado.

Así, a priori, se debería realizar análisis de las variables:

- age ~ salary. Si la capacidad de ganar más aumenta con la edad, o hay algún tipo de predisposición hacia los grupos de más experiencia que pueda representar una barrera para los más jóvenes.
- sex ~ salary. Si hay censo social que favorece a los hombres frente a las mujeres.
- race~salary. Si hay censo social que discrimen a los individuos no considerados de raza blanca.

También, se pretende analizar las otras variables, como workclass, occupation, education o hours.per.week para determinar qué peso tienen a la hora de determinar la probabilidad del individuo de ganar más de 50k.

Para analizar dichas cuestiones, se realizan sobre las variables cuantitativas (education, age, hours.per.week) un análisis de correlación. Sobre las variables sex and race, se realizará un contraste de hipótesis. Además, estos análisis estadísticos se verán acompañados de modelos de regresión lineal y modelos de regresión logística para sospesar el peso en la predicción de la variable salary.

4.2 Comprobación de la normalidad y homogeneidad de la varianza.

4.2.1. Normalidad

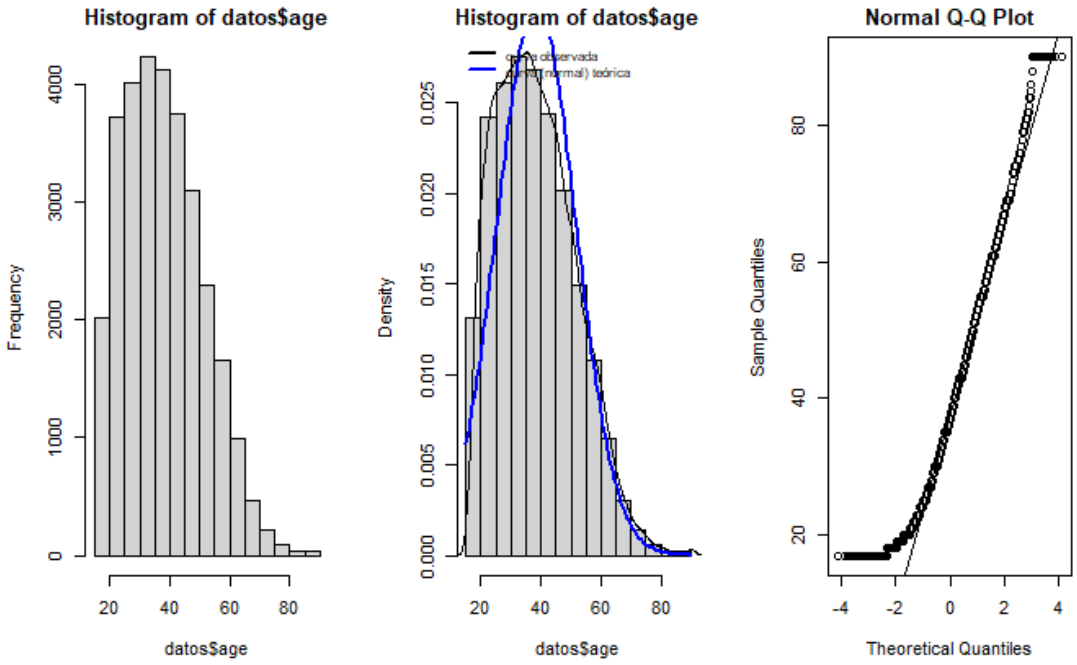
Dado el elevado número de datos, podría aplicarse el teorema del límite central. También puede usarse uno de los tests disponibles en r. Dado el elevado tamaño de la muestra se usará el test Lilliefors (Kolmogorov-Smirnov). Si el valor de probabilidad (p-value) que obtenemos por la prueba es menor a 0.05 se rechaza la hipótesis nula y los datos no siguen una distribución normal. Si el valor de probabilidad es mayor a 0.05, no se puede rechazar la hipótesis nula y los datos seguirían una distribución normal.

- Para la variable Age.

El valor p-value del test de Lilliefors es mucho menor al 0.05 por tanto se rechaza la hipótesis nula y se concluye que la variable edad no sigue una distribución normal.

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  datos$age
D = 0.060447, p-value < 2.2e-16
```

La representación visual de dicha distribución sería la siguiente:

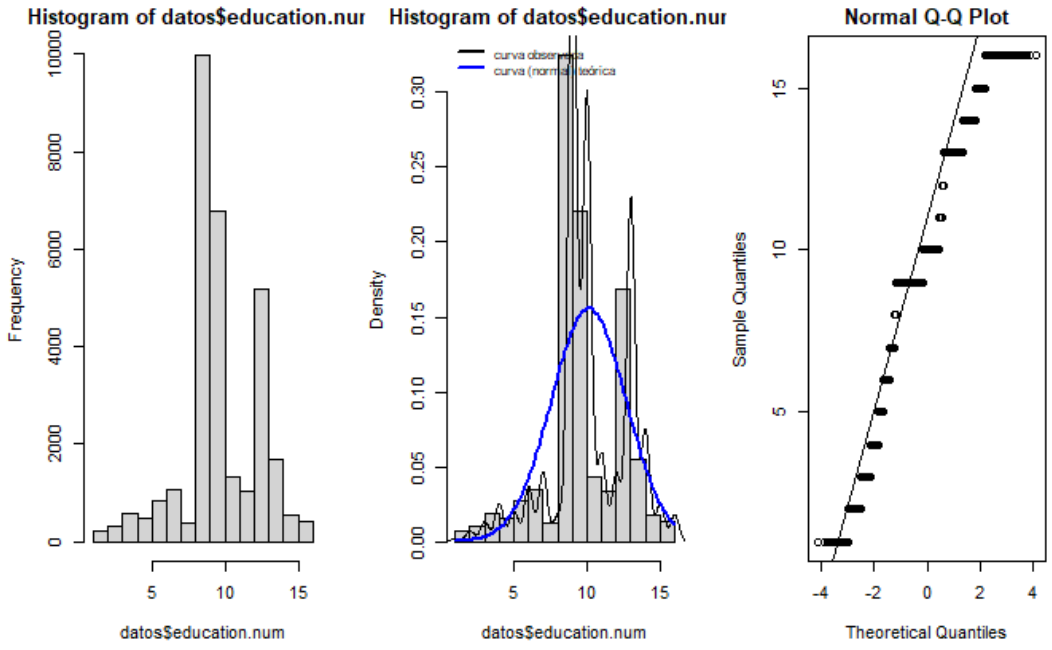


- Para la variable Education.

El valor p-value del test de Lilliefors es mucho menor al 0.05 por tanto se rechaza la hipótesis nula y se concluye que la variable número de años de estudios no sigue una distribución normal.

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  datos$education.num
D = 0.20518, p-value < 2.2e-16
```

Su representación visual es la siguiente:

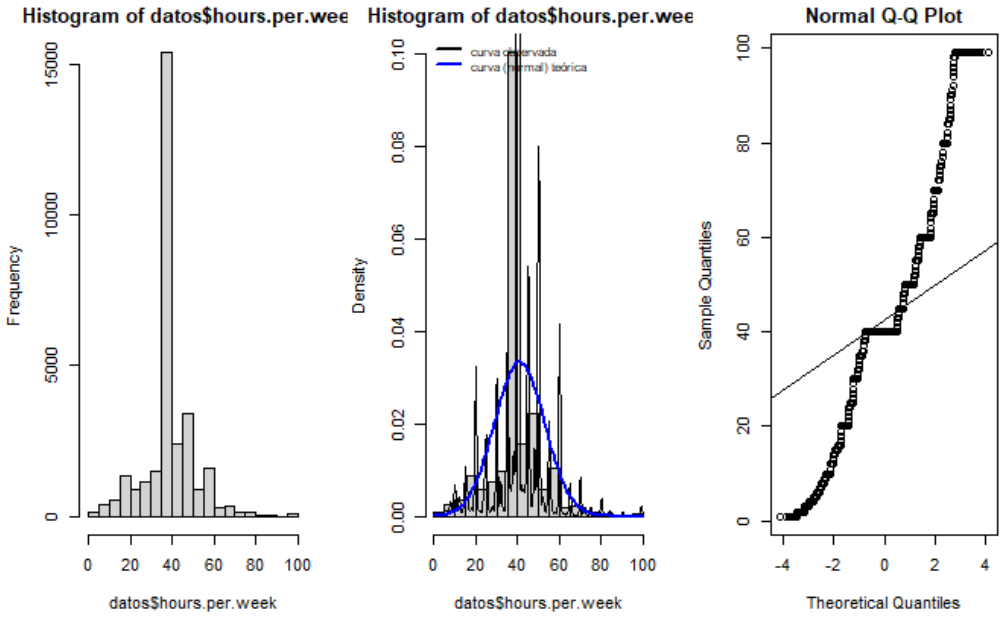


- Para la variable Hours Per Week

El valor p-value del test de Lilliefors es mucho menor al 0.05 por tanto se rechaza la hipótesis nula y se concluye que la variable horas de trabajo a la semana no sigue una distribución normal. Visualmente quedaría

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  datos$hours.per.week
D = 0.24645, p-value < 2.2e-16
```

Visualmente la distribución queda de la siguiente forma:



4.2.2. Homogeneidad

Implementamos el test de Fligner-Killeen, recordemos que se trata de la alternativa no paramétrica, utilizada cuando los datos no cumplen con la condición de normalidad.

La hipótesis nula asume igualdad de varianzas en los diferentes grupos de datos, por lo que p-valores inferiores al nivel de significancia indican heterocedasticidad.

En este caso, estudiaremos las diferencias en la varianza en los grupos de edad, horas semanales y estudios con la probabilidad de tener un salario superior a los 50k.

- Para la variable age.

```
        Fligner-Killeen test of homogeneity of variances

data:  age by salary
Fligner-Killeen:med chi-squared = 597.28, df = 1, p-value <
2.2e-16
```

- Para la variable hours.per.week

```
        Fligner-Killeen test of homogeneity of variances

data:  hours.per.week by salary
Fligner-Killeen:med chi-squared = 34.465, df = 1, p-value =
4.341e-09
```

- Para la variable education.num

```
        Fligner-Killeen test of homogeneity of variances

data:  education.num by salary
Fligner-Killeen:med chi-squared = 119.31, df = 1, p-value <
2.2e-16
```

Al obtener un valor p menor que el valor de significancia 0,05, rechazamos la hipótesis nula que presuponía homocedasticidad y concluimos que la variable age, hours.per.week y education.num presentan varianzas estadísticamente diferentes en las personas con un sueldo superior a los 50k y las personas con un sueldo inferior a los 50k.

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

4.3.1 ¿Qué variables cuantitativas influyen más en el salario?

En primer lugar, procedemos a realizar un análisis de correlación entre las diferentes variables cuantitativas para determinar cuáles de ellas ejercen una mayor influencia en el momento de determinar la probabilidad de tener un salario superior o inferior a los 50K.

Puesto que hemos determinado que las variables no siguen una distribución normal, utilizaremos el coeficiente de correlación de Spearman. Esta es una alternativa no paramétrica que mide el grado de dependencia entre dos variables. La suposición que deben cumplir las variables es que deben de medirse en una escala ordinal.

Realizado el test, con los resultados de la siguiente figura, nos fijamos en el p-value. En todos ellos su valor es menor que 0,05 lo que implica que las correlaciones son estadísticamente significantes.

Ahora nos fijamos en el rango estimado de Spearman, identificamos cuáles son las variables más correlacionadas con el precio en función de su proximidad con los valores -1 y +1. En todas ellas su rango es positivo lo que implica que la correlación entre cada uno de ellos es positiva. Por ejemplo, a mayor nivel de educación, mayor probabilidad de obtener un saldo mayor de 50K. Entre las diferentes correlaciones analizadas, vemos que la mayor se obtiene en la variable education. Por tanto, concluimos que esta variable tiene mayor peso que las otras dos en determinar el salario.

```
        Spearman's rank correlation rho

data:  datos$salary and datos$age
S = 3.4916e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2772296


        Spearman's rank correlation rho

data:  datos$salary and datos$education.num
S = 3.2363e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.3300795


        Spearman's rank correlation rho

data:  datos$salary and datos$hours.per.week
S = 3.5448e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2662159
```

4.3.2 ¿La probabilidad de tener un salario superior a 50K aumenta si el individuo es un hombre de raza blanca?

La segunda prueba estadística que se aplicará consistirá en un contraste de hipótesis sobre dos muestras para determinar si el sexo y la raza influyen en el salario del individuo. ¿Hay algún tipo de discriminación racial o brecha salarial entre hombre y mujeres?

Para ello, obtendremos diferentes muestras del dataset, una para cada sexo y para cada raza listado en el censo.

Como se ha determinado que la normalidad y la homocedasticidad no se cumplen (recordamos que p-valores eran menores al nivel de significancia) aplicaremos pruebas no paramétricas, como la prueba de Mann-Whitney.

```
        Wilcoxon rank sum test with continuity correction

data:  datos.female.salary and datos.male.salary
W = 82539763, p-value < 2.2e-16
alternative hypothesis: true location shift is less than 0
```

El contraste de hipótesis de dos muestras sobre la diferencia de medias es unilateral atendiendo a la formulación de la hipótesis alternativa. Se ha planteado si la media de la primera muestra (salario mujeres) es menor (hipótesis alternativa) que la media de la segunda muestra (salario hombres).

Puesto que obtenemos un p-valor menor que el valor de significación, rechazamos la hipótesis nula, y concluimos que, efectivamente, la probabilidad de tener un salario superior a 50K será mayor si el individuo es hombre.

Ahora vamos a determinar la posible discriminación salarial por raza. Utilizamos el mimo test, ya que las condiciones para aplicarlo se conservan.

```
        Wilcoxon rank sum test with continuity correction

data:  datos.black.salary and datos.white.salary
W = 33125646, p-value < 2.2e-16
alternative hypothesis: true location shift is less than 0


        Wilcoxon rank sum test with continuity correction

data:  datos.asian.salary and datos.white.salary
W = 12990645, p-value = 0.8383
alternative hypothesis: true location shift is less than 0


        Wilcoxon rank sum test with continuity correction

data:  datos.indian.salary and datos.white.salary
W = 3215168, p-value = 1.439e-08
alternative hypothesis: true location shift is less than 0
```

Comparamos la población identificada como raza “Black” con los de raza “White”, los de raza “Asian-Pac-Islander" con los de raza “White”, y lo de raza “Amer-Indian-Eskimo" con los de raza “White”.

Solamente en el caso de la población de raza Asiática, el p-valor es superior a 0,05. Por tanto, para la población asiática no podríamos descartar la hipótesis nula. En cambio, para los otros grupos sí y concluiríamos que existen diferencias salariares según la raza del individuo.

4.3.3. Modelo de regresión lineal

La tercera prueba estadística es la creación de un modelo de regresión lineal, el cual nos permitirá poder realizar predicciones sobre el salario de un individuo por sus características intrínsecas del individuo y logros. Así, se realizará un modelo de regresión lineal utilizando regresores tanto cuantitativos como cualitativos con el que poder realizar las predicciones del salario.

Para obtener un modelo de regresión lineal considerablemente eficiente, lo que haremos será obtener varios modelos de regresión utilizando las variables que estén más correlacionadas con respecto al salario, según los resultados obtenidos anteriormente. Así, de entre todos los modelos que tengamos, escogeremos el mejor utilizando como criterio aquel que presente un mayor coeficiente de determinación (R^2).

```r
```{r}
# Regresores cuantitativos
educationNum = datos$education.num
ageIndividual = datos$age
hoursPerWeek = datos$hours.per.week
# Regresores cualitativos
sexIndividual = datos$sex
raceIndividual = datos$race
workclassInd = datos$workclass
# Variable a predecir
salary50k = datos$salary
```
```

```
# Generación de varios modelos
# No age, sex or race
modelo1 <- lm(salary50k ~ educationNum +
                hoursPerWeek +
                workclassInd,
              data = datos)
# No sex or race
modelo2 <- lm(salary50k ~ educationNum +
                hoursPerWeek +
                workclassInd +
                ageIndividual,
              data = datos)
# No educationNum
modelo3 <- lm(salary50k ~ ageIndividual +
                sexIndividual +
                raceIndividual +
                workclassInd +
                hoursPerWeek,
              data = datos)
# No educationNum or workclass
modelo4 <- lm(salary50k ~ sexIndividual +
                raceIndividual +
                hoursPerWeek,
              data = datos)
# No education or age
modelo5 <- lm(salary50k ~ sexIndividual +
                raceIndividual +
                workclassInd +
                hoursPerWeek,
              data = datos)
# Only education, hours and workclass
modelo6 <- lm(salary50k ~ educationNum +
                hoursPerWeek +
                workclassInd,
              data = datos)
# Only sex, race and education
modelo7 <- lm(salary50k ~ educationNum +
                sexIndividual +
                raceIndividual,
              data = datos)
# Only education
modelo8 <- lm(salary50k ~ educationNum,
              data = datos)
```

Para los anteriores modelos de regresión lineal múltiple obtenidos, podemos utilizar el coeficiente de determinación para medir la bondad de los ajustes y quedarnos con aquel modelo que mejor coeficiente presente.

```
     Modelo        R^2
[1,]      1 0.15506756
[2,]      2 0.19598784
[3,]      3 0.14085148
[4,]      4 0.08592147
[5,]      5 0.10277441
[6,]      6 0.15506756
[7,]      7 0.16021941
[8,]      8 0.11198415
```

En este caso, tenemos que el segundo modelo sería el que tiene un mayor coeficiente de determinación. Pero a pesar de ser el mejor entre los modelos utilizados, el coeficiente de determinación no sería lo suficientemente grande para hacer predicciones, pero nos permitirá extraer algunas conclusiones que comentaremos en el apartado del mismo nombre.

4.3.3. Modelo de regresión logística

Otra prueba estadística que podemos aplicar es la regresión logística. Esta es un tipo de análisis de regresión utilizada para predecir el resultado de una variable dicotómica dependiente, en función de una serie de variables independientes o predictoras. Utiliza una escala transformada basada en una función logística.

Realizamos la validación cruzada 10-fold, para dividir la muestra original, con el fin de minimizar el posible overfitting.

```{r}
set.seed (1000)

trainCtrl = trainControl (method = "cv",
                          number = 10)

regresionModelo = train (salary ~ age +
                           workclass +
                           education.num +
                           marital.status +
                           occupation +
                           relationship +
                           race +
                           sex +
                           hours.per.week,
                         trControl = trainCtrl,
                         method = "gbm",
                         data = train,
                         verbose = FALSE)
```

El resultado de aplicarlo sobre nuestros datos es el siguiente:

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 21347  1721
         1  3243  4407

               Accuracy : 0.8384
                 95% CI : (0.8342, 0.8425)
    No Information Rate : 0.8005
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5372

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8681
            Specificity : 0.7192
         Pos Pred Value : 0.9254
         Neg Pred Value : 0.5761
             Prevalence : 0.8005
         Detection Rate : 0.6949
   Detection Prevalence : 0.7510
      Balanced Accuracy : 0.7936

       'Positive' Class : 0
```

En este caso, la bondad del modelo se evalúa mediante la medida AIC (criterio de información de Akaike). Cuando menor es el AIC mejor será la bondad de ajuste del modelo.

En este caso, solo hemos elaborado un modelo con las características que hemos ido viendo que han sido más significativas.

La matriz de confusión muestra una precisión general en la muestra de ~ 83%, sensibilidad de ~ 86% y especificidad de ~ 71%. Esto implica que el 83% de las veces, el modelo ha clasificado correctamente el nivel de ingresos, el 86% de las veces, siendo el nivel de ingresos menor o igual a USD 50000 en clasificar

correctamente y el 71% de las veces, siendo el nivel de ingresos mayor que 50000 USD está clasificado correctamente.

Realizamos otro modelo de regresión logística, pero ahora no incorporamos las variables sex o race, ya que partimos de la premisa que ni el sexo ni la raza del individuo debería de condicionar el salario, y por tanto el modelo no se vería demasiado influenciado por la falta de dichas variables.

Realizamos su matriz de confusión.

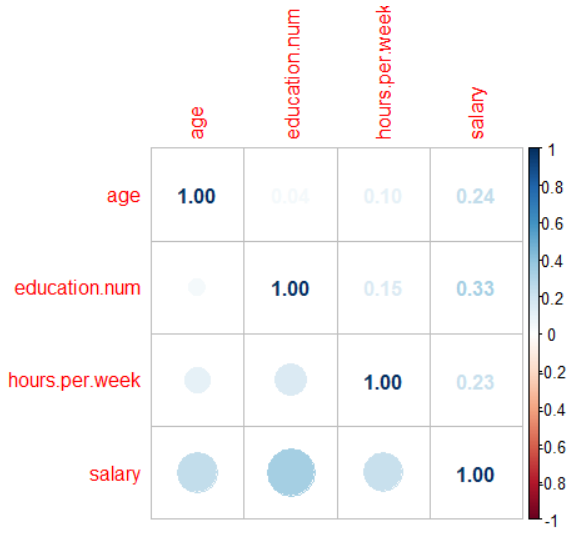
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 21370  1698
         1  3277  4373

               Accuracy : 0.838
                 95% CI : (0.8339, 0.8421)
    No Information Rate : 0.8024
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5349

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8670
            Specificity : 0.7203
         Pos Pred Value : 0.9264
         Neg Pred Value : 0.5716
             Prevalence : 0.8024
         Detection Rate : 0.6957
   Detection Prevalence : 0.7510
      Balanced Accuracy : 0.7937

       'Positive' Class : 0
```

Se obtiene un nuevo modelo con una precisión de 83.8%, una sensibilidad de 86.70% y una especificidad de 72.03%.

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

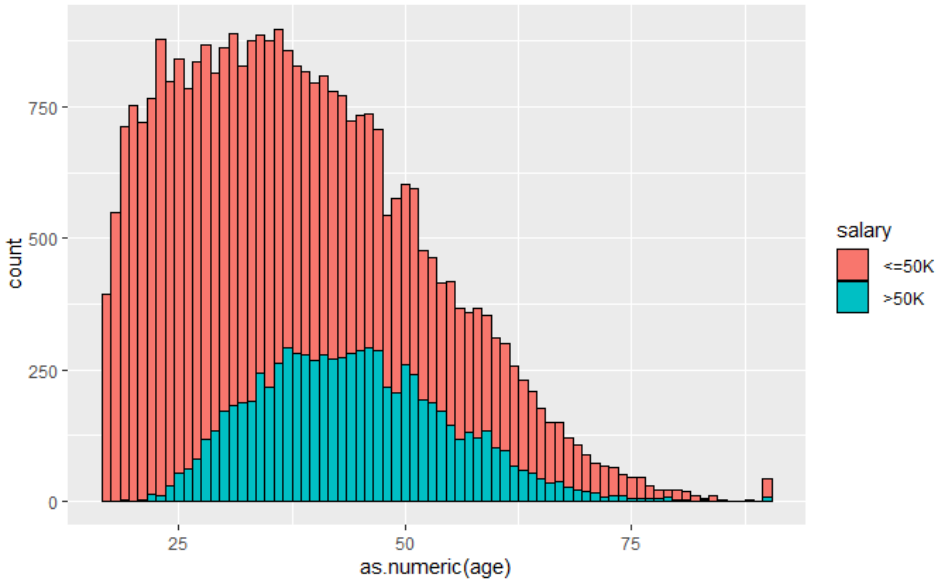
Representamos gráficamente el estudio de correlaciones entre atributos, realizado antes. Este estudio solamente puede hacerse con atributos numéricos, por lo que debemos asegurarnos antes de tener solamente atributos numéricos.

En la figura se describe las correlaciones con el color rojo indicando una correlación negativa y el color azul una positiva. Cuando mayor sea la correlación entre atributos, mayor será su índice (sobre 1) y mayor será la intensidad del color.

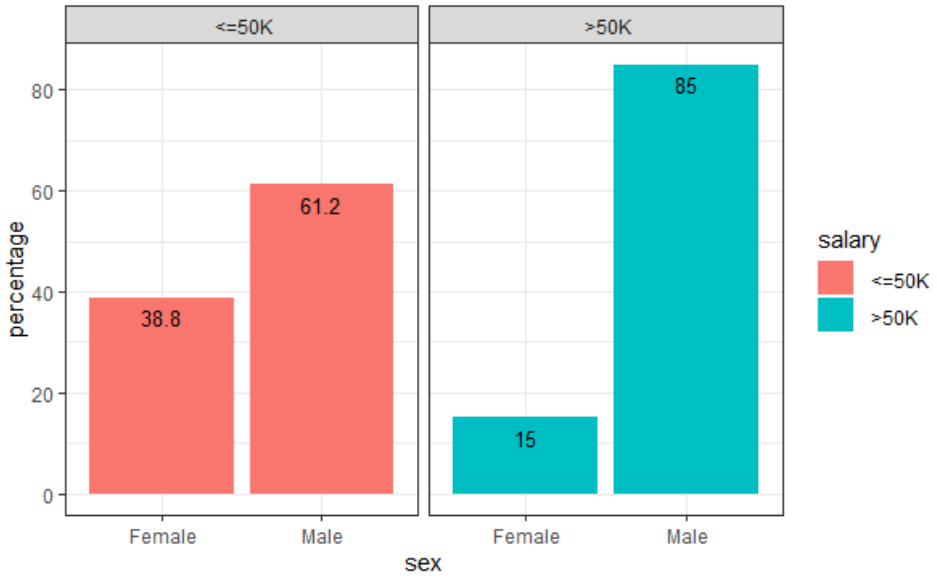


La correlación que se nos presenta es escasa, y positiva. Este gráfico nos muestra visualmente las conclusiones llegadas en el apartado anterior. La variable education es la más influyente (dentro de las variables numéricas) en la determinación del salario.

Ahora pasaremos a realizar el análisis de las variables respecto a la clase (salary) a la que pertenecen.

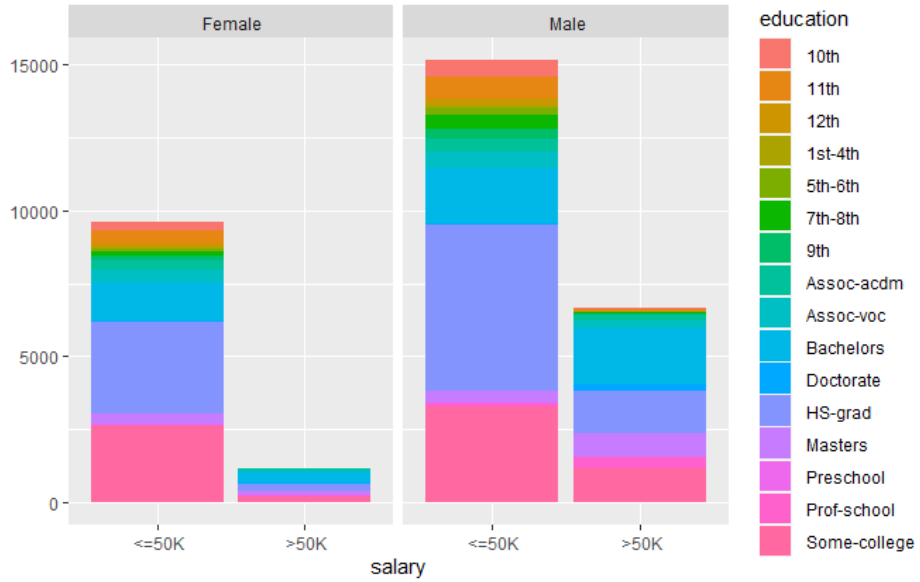


Los censados se agrupan dentro del grupo correspondiente a aquellos que ganan menos de 50000$ (en adelante 50K) y apreciamos que cuando la edad se aproxima a la edad de jubilación esa tendencia va disminuyendo hasta casi hacerse cero, dado que los individuos pasan a un estado inactivo laboral.



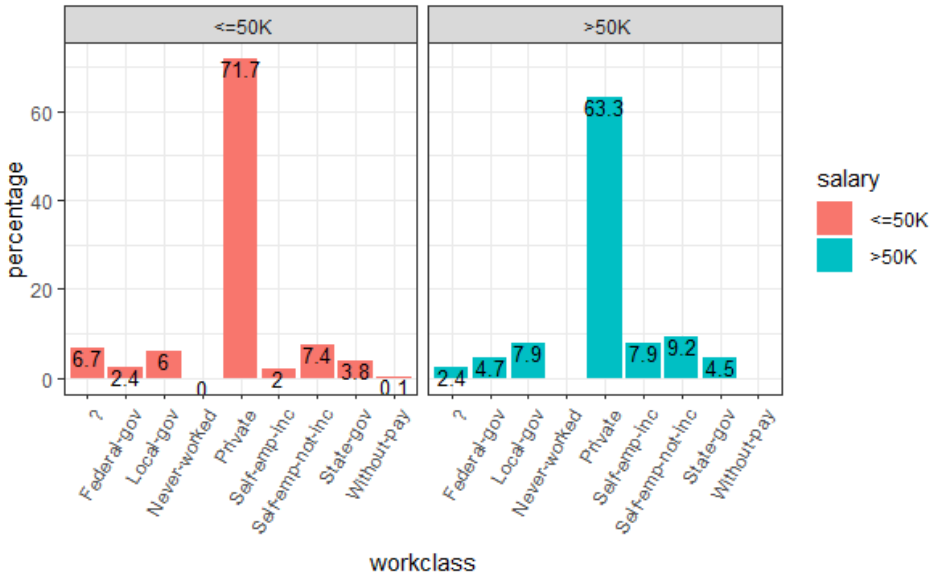
Podemos observar claramente que los hombres son más propensos a ganar más de 50K que las mujeres. Del mismo modo que también son ellos quienes ganan, en menos de 50K, aunque en este caso la proporción es más baja, por lo que podríamos decir que los puestos que ganan más de 50K están más ocupados por hombres.

Lo confirmaremos con el grado de estudios que tienen los censados.

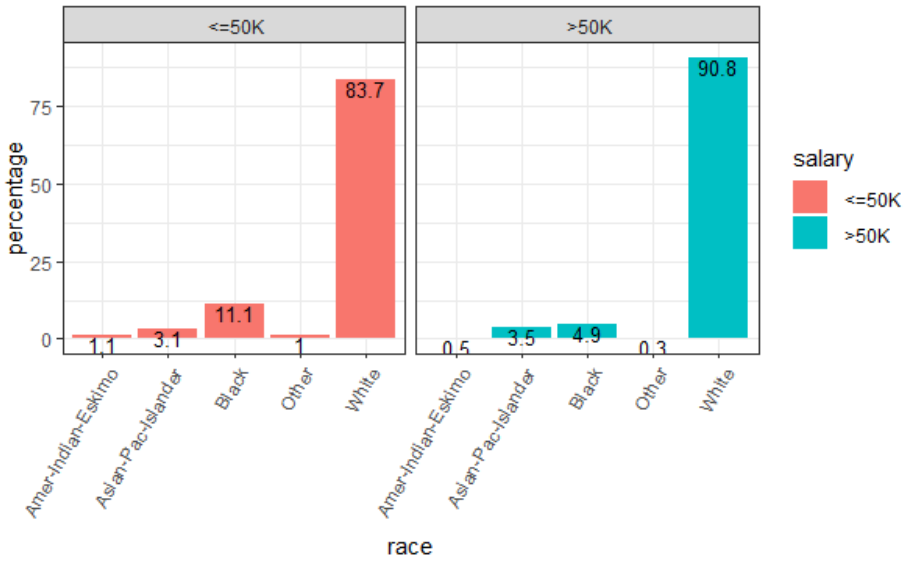


Comprobamos que la mayoría de ellos tienen estudios universitarios. Aquellos que únicamente tienen estudios básicos se encuentran en la clase de menos de 50K pues es obvio que para los puestos de responsabilidad se requieren algún tipo de estudio post universitario. También, vemos que hay más número de doctores en la categoría de más de 50K que en la de menos lo cual corrobora la idea anterior.

Si analizamos el puesto de trabajo con los ingresos podemos ver que no hay diferencia respecto al género del censado y que la gran mayoría de ellos (independientemente de la variable salary) pertenecen al sector privado.



La relación respecto a la raza tampoco aporta mucha información al ser el grupo de raza blanca el predominante en ambas categorías, confirmando que es muy difícil encontrar a una persona de otra raza diferente de la blanca en altos puestos de dirección (y menos si es mujer, independientemente de la raza).



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

Por la distribución de la variable salario, es evidente que existe un sesgo o discriminación hacia un tipo determinado de individuos.

Entre los 30 y 50 años, son cuando se alcanza la mayor probabilidad de obtener un mayor salario. Los hombres, estadísticamente, son más propensos a ganar más de 50K. También, son más propensos a ganar menos de 50K, pero esto solamente evidenciaría que hay más población de hombres trabajadores que de mujeres. La educación, aunque es la variable más influyente, mayores niveles de educación no parece aumentar la probabilidad de ganar más. El grupo más grande estaría conformado por aquellos con “Bachelors” para ambos sexos. Dentro del sector laboral, el sector privado es donde la probabilidad

de ganar más es mayor. La población de raza blanca tiene mayor acceso a la vida laboral, y de igual forma, su salario es potencialmente mayor. También se observa una discriminación evidente en el grupo de raza "negra" y "india", los cuales son los que tienen menor probabilidad de ganar más de 50K.

Si atendemos a los diferentes tests estadísticos, extraemos las siguientes conclusiones:

Entre las diferentes variables cuantitativas analizadas (education, age, hours.per.week), vemos que la más determinante es la variable education. Por tanto, concluimos que esta variable tiene mayor peso que las otras dos en determinar el salario. Aunque las diferencias son mínimas, la escala de influencia sería la siguiente: education >> age > hours.per.week

En el contraste de hipótesis, hemos determinado sesgos. Los hombres tienen más probabilidad de ganar más que las mujeres. Los individuos de raza blanca son más propensos a ganar más que los de otras razas (a excepción de los de raza asiática).

Los modelos de regresión lineal no nos servirían para predecir el salario, pero nos dan una idea de la influencia de cada variable. La variable education sigue siendo la más influyente, las variables race o sex aunque influyen, no condicionan en gran medida el modelo.

En el modelo de regresión logística, hemos podido observar un patrón similar. Aunque las variables race o sex, se han determinado como influyentes y se ha concluido la existencia de un sesgo, no incluir estas variables en el modelo apenas ha condicionado la precisión, sensibilidad o especificidad de este. No podríamos concluir que dichas variables no influyan, porque estaríamos cayendo en error, sino que deberíamos concluir que los hombres de raza blanca conforman la mayoría de los encuestados y por tanto el modelo es preciso para este colectivo, pero de ser aplicado a los grupos discriminados, obtendríamos un modelo totalmente erróneo.

7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

Adjuntamos los siguientes vínculos del repositorio Github:

- Fichero de datos originales: https://github.com/UOCPgarcia/Factores_vs_Salario_Cleaning/blob/main/data/adult_data.csv
- Fichero de código: Factores_vs_Salario_Cleaning/cleaning.rmd at main · UOCPgarcia/Factores_vs_Salario_Cleaning (github.com)
- Fichero de datos procesados: https://github.com/UOCPgarcia/Factores_vs_Salario_Cleaning/blob/main/data/adult_data_processed.csv
- Salida del informe dinámico formato pdf: https://github.com/UOCPgarcia/Factores_vs_Salario_Cleaning/blob/main/code/cleaning.pdf
- Salida del informe dinámico formato html: https://github.com/UOCPgarcia/Factores_vs_Salario_Cleaning/blob/main/code/cleaning.html

8. Referencias.

[1]. Debasish Dutta. (2021). *Income of Adults* [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/2241218

Nota: Formato APA.

9. Tabla de contribuciones al trabajo.

Contribuciones	Firma
Investigación previa	M.P.G.R, A.C.A
Redacción de las respuestas	M.P.G.R, A.C.A
Desarrollo código	M.P.G.R, A.C.A