Examen_II_G1

RandalPicadoBermudezC36024

2025-11-30

Table of contents

Grupo 1: Luis Alberto Juárez Potoy

Parte1

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```r
library(scales)
```

```
Adjuntando el paquete: 'scales'

The following object is masked from 'package:purrr':

    discard

The following object is masked from 'package:readr':

    col_factor
```

```r
library(readxl)
moras <- read_excel("moras.xlsx")
```

Parte2

```r
resumen <- moras %>%
  summarise(
    media.edad = mean(EDAD, na.rm = TRUE),
    minimo.edad = min(EDAD, na.rm = TRUE),
    maximo.edad = max(EDAD, na.rm = TRUE),
    q1.edad = quantile(EDAD, 0.25, na.rm = TRUE),
    q3.edad = quantile(EDAD, 0.75, na.rm = TRUE),

    media.salario = mean(SALARIO, na.rm = TRUE),
    minimo.salario = min(SALARIO, na.rm = TRUE),
```

```
    maximo.salario = max(SALARIO, na.rm = TRUE),
    q1.salario = quantile(SALARIO, 0.25, na.rm = TRUE),
    q3.salario = quantile(SALARIO, 0.75, na.rm = TRUE),
      )

resumen
```

```
# A tibble: 1 x 10
  media.edad minimo.edad maximo.edad q1.edad q3.edad media.salario
       <dbl>       <dbl>       <dbl>   <dbl>   <dbl>         <dbl>
1       41.9          14          99      31      52       657534.
# i 4 more variables: minimo.salario <dbl>, maximo.salario <dbl>,
#   q1.salario <dbl>, q3.salario <dbl>
```

Parte3

```
mean.salario <- mean(moras$SALARIO, na.rm = TRUE)
mean.edad <- mean(moras$EDAD, na.rm = TRUE)
desviacion.salario <- sd(moras$SALARIO, na.rm = TRUE)
desviacion.edad <- sd(moras$EDAD, na.rm = TRUE)

moras <- moras %>%
  mutate(
    Z.SALARIO = (SALARIO - mean.salario) / desviacion.salario,
    Z.EDAD = (EDAD - mean.edad) / desviacion.edad
    ) %>%
  mutate(
    ATIPICO.SALARIO = if_else(abs(Z.SALARIO) > 1.96, "ATIPICO", "NO ATIPICO"),
    ATIPICO.EDAD = if_else(abs(Z.EDAD) > 1.96, "ATIPICO", "NO ATIPICO")
  )
```

Parte4

```
porcentajes.na <- data.frame(
  dato = names(moras),
  faltante = colSums(is.na(moras))
) %>%
```

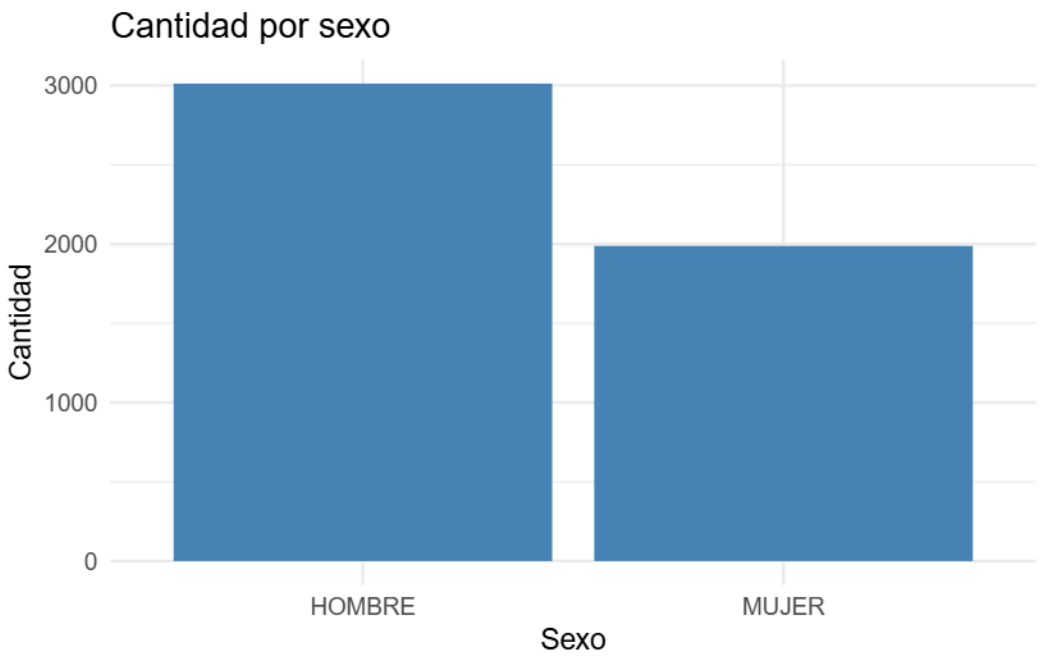
```
  mutate(
    porcentajes.na = (faltante/5000)* 100
  )
```

Parte5

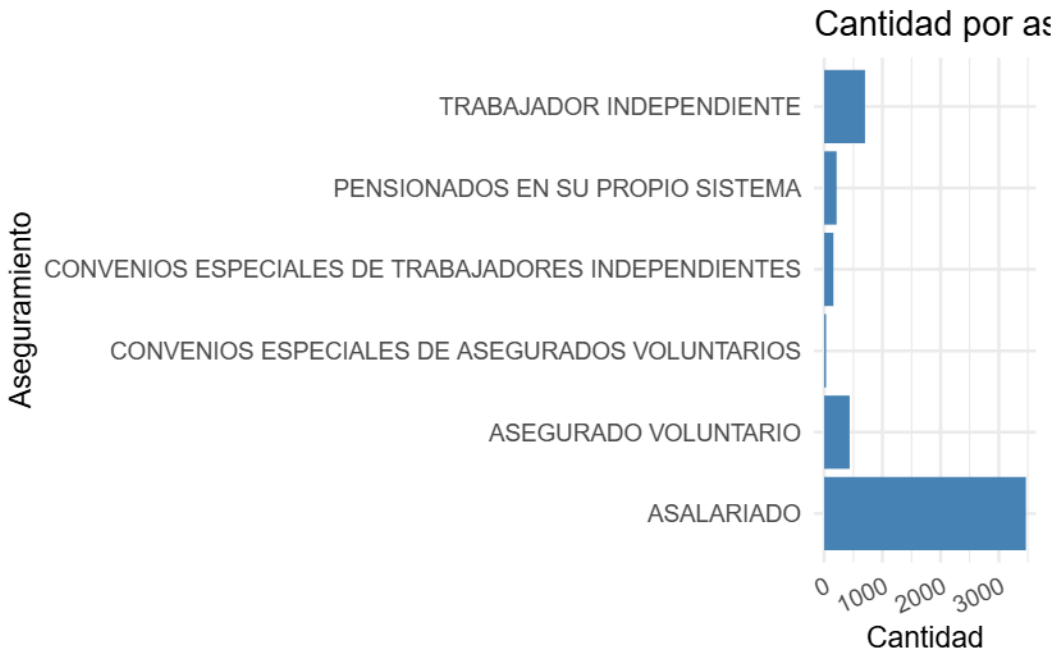
```
#Los unicos con valores atipicos son Salario, tipo Aseguramiento y edad,
#con 3,2 y 3 valores respectivamente, lo cual no es significativo dentro del total
#(5000), por lo tanto por eficiencia se va a imputar con media y con moda.

moda.aseguramiento <- moras %>%
  group_by(TIPO_ASEGURAMIENTO) %>%
  summarise( contador = n()) %>%
  arrange(desc(contador)) %>%
  head(1)

moda.aseg <- moda.aseguramiento$TIPO_ASEGURAMIENTO

moras <- moras %>%
  mutate(
    EDAD = if_else(is.na(EDAD), mean.edad, EDAD),
    SALARIO = if_else(is.na(SALARIO), mean.salario, SALARIO),
    TIPO_ASEGURAMIENTO = if_else(is.na(TIPO_ASEGURAMIENTO), moda.aseg, TIPO_ASEGURAMIENTO)
  )
```

Parte6

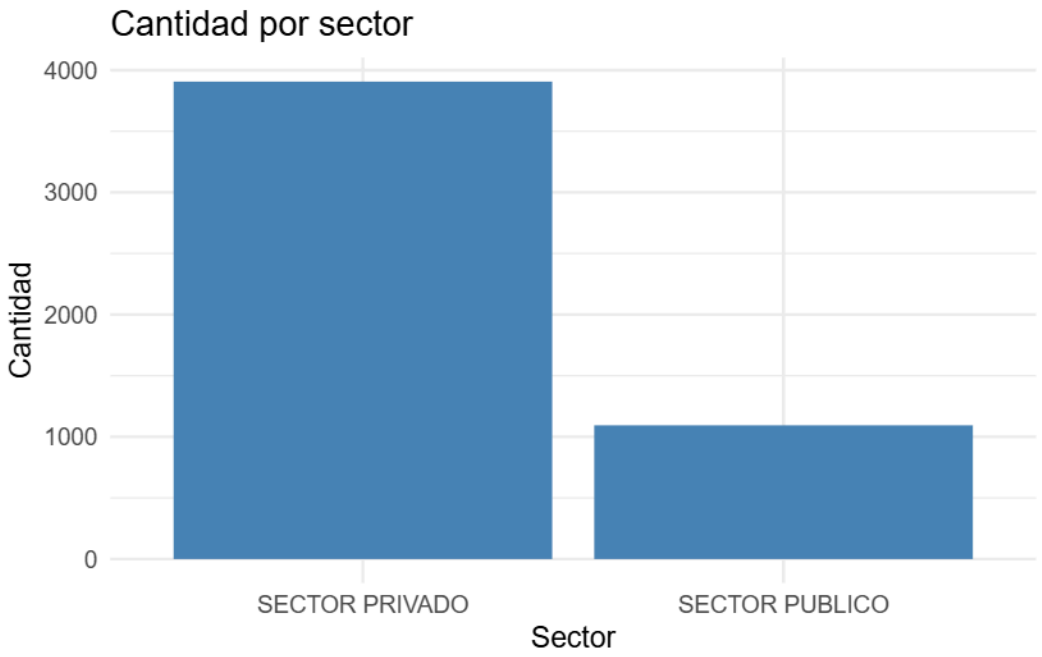
```
graf.cantidadporsexo <- moras %>%
  group_by(SEXO) %>%
  summarise(contador = n()) %>%
  ggplot(aes(x = SEXO, y = contador)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Cantidad por sexo", x = "Sexo", y = "Cantidad")+
  theme_minimal()

graf.cantidadporsexo
```



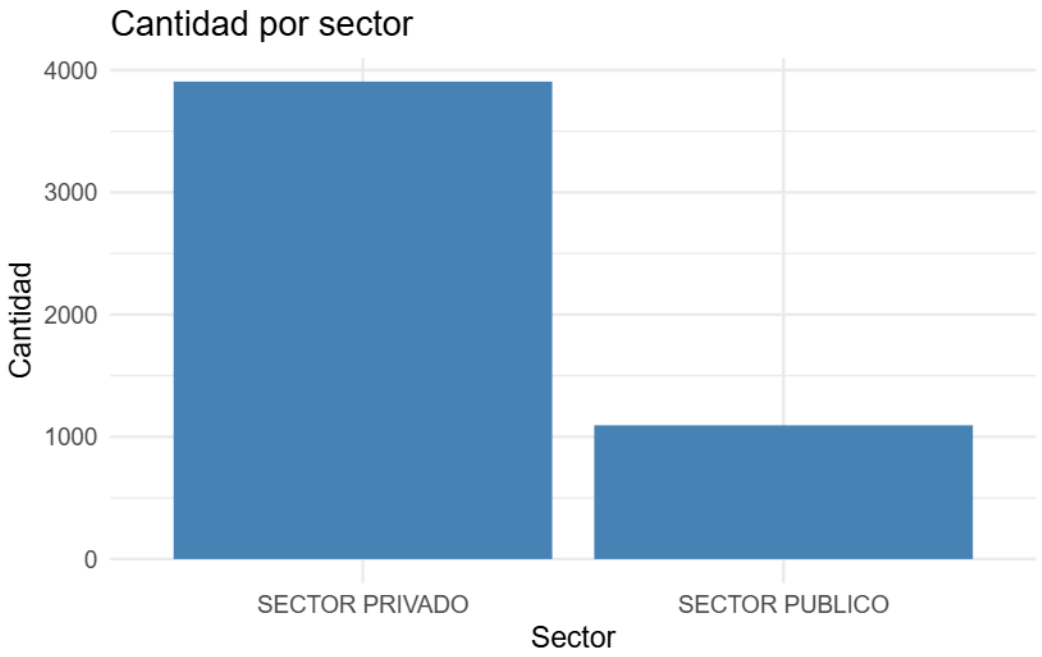

```
graf.cantidadporaseguramiento <- moras %>%
  group_by(TIPO_ASEGURAMIENTO) %>%
  summarise(contador = n()) %>%
  ggplot(aes(x = TIPO_ASEGURAMIENTO, y = contador)) +
  coord_flip()+
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Cantidad por aseguramiento", x = "Aseguramiento", y = "Cantidad")+
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 25, hjust = 1))

graf.cantidadporaseguramiento
```




```r
graf.cantidadporsector <- moras %>%
  group_by(SECTOR) %>%
  summarise(contador = n()) %>%
  ggplot(aes(x = SECTOR, y = contador)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Cantidad por sector", x = "Sector", y = "Cantidad")+
  theme_minimal()

graf.cantidadporsector
```

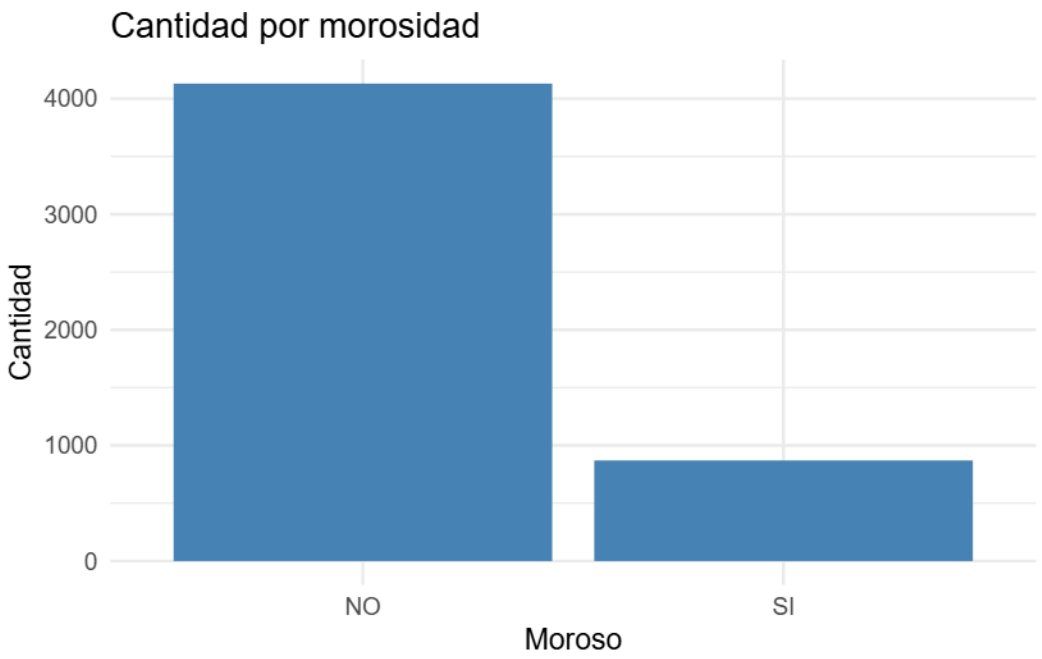



```
graf.cantidadporextranjeros<- moras %>%
  group_by(INDICADOR.EXTRANJERO) %>%
  summarise(contador = n()) %>%
  ggplot(aes(x = INDICADOR.EXTRANJERO, y = contador)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Cantidad por extranjeros", x = "Extranjero", y = "Cantidad")+
  theme_minimal()

graf.cantidadporsector
```




```
graf.cantidadporactividad <- moras %>%
  group_by(INDICADOR_ACTIVO) %>%
  summarise(contador = n()) %>%
  ggplot(aes(x = INDICADOR_ACTIVO, y = contador)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Cantidad por actividad", x = "Activo", y = "Cantidad")+
  theme_minimal()

graf.cantidadporactividad
```


```
graf.cantidadpormorosidad <- moras %>%
  group_by(INDICADOR_MOROSO) %>%
  summarise(contador = n()) %>%
  ggplot(aes(x = INDICADOR_MOROSO, y = contador)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Cantidad por morosidad", x = "Moroso", y = "Cantidad")+
  theme_minimal()

graf.cantidadpormorosidad
```

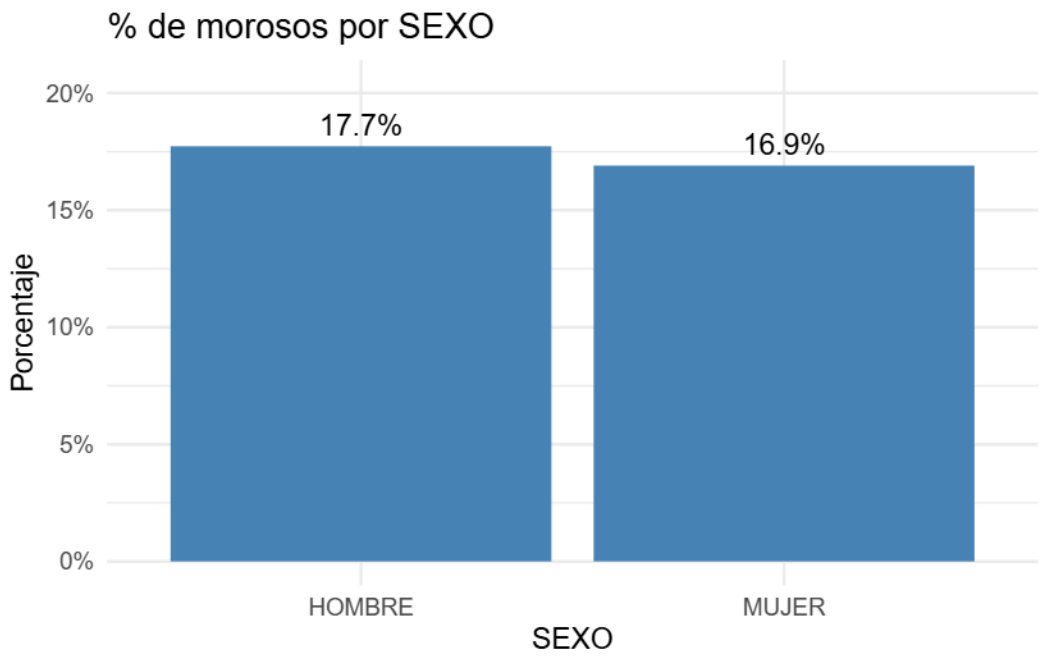



```
#La interpretacion de estas graficas es analoga al de excel, pues son los mismos graficos.
```

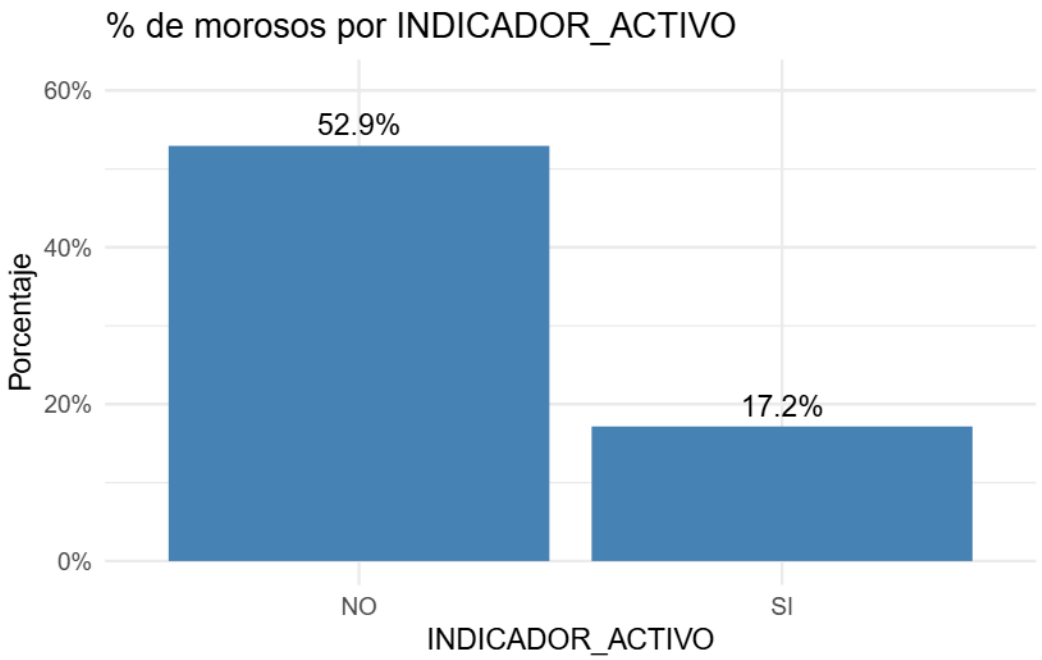
parte7

```
graf.porcentajemorosidad <- function(data, categoria){

  datos <- data %>%
    group_by(.data[[categoria]]) %>%
    summarise(
      cantidad   = n(),
      morosos    = sum(INDICADOR_MOROSO == "SI"),
      porcentaje = morosos / cantidad
    )

  ggplot(datos, aes(x = .data[[categoria]], y = porcentaje)) +
    geom_col(fill = "steelblue") +
    geom_text(aes(label = percent(porcentaje, accuracy = 0.1)),
              vjust = -0.5, size = 4) +
    scale_y_continuous(labels = percent_format(accuracy = 1),
                       limits = c(0, max(datos$porcentaje) * 1.15)) +
```

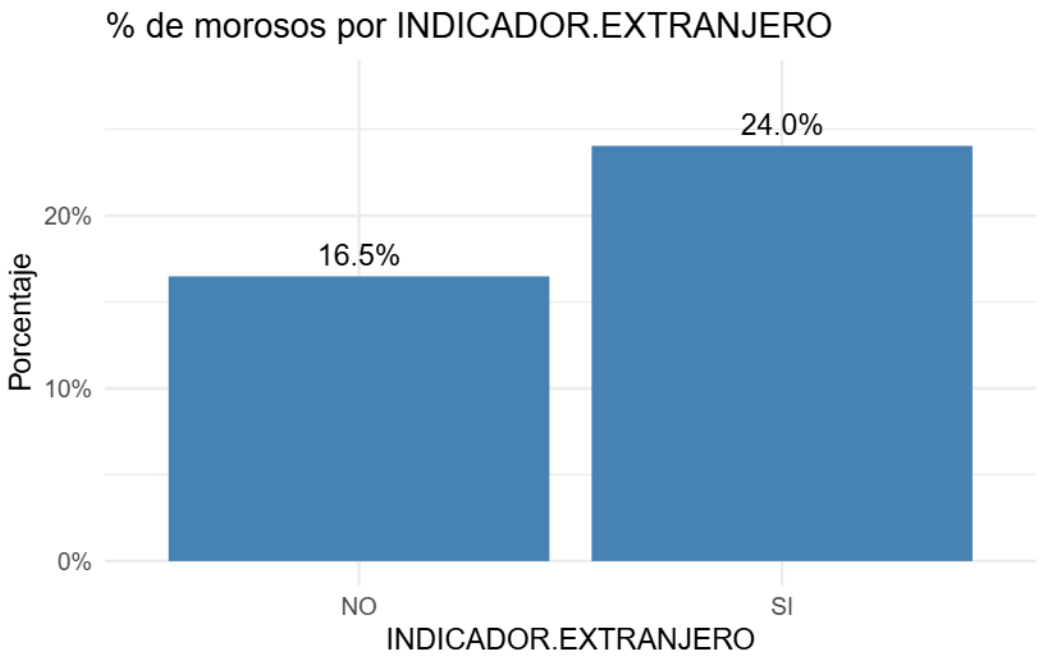
```
    labs( title = paste("% de morosos por", categoria), x = categoria,y = "Porcentaje"
    ) +
    theme_minimal()
}

graf.porcentajemorosidad(moras, "SEXO")
```



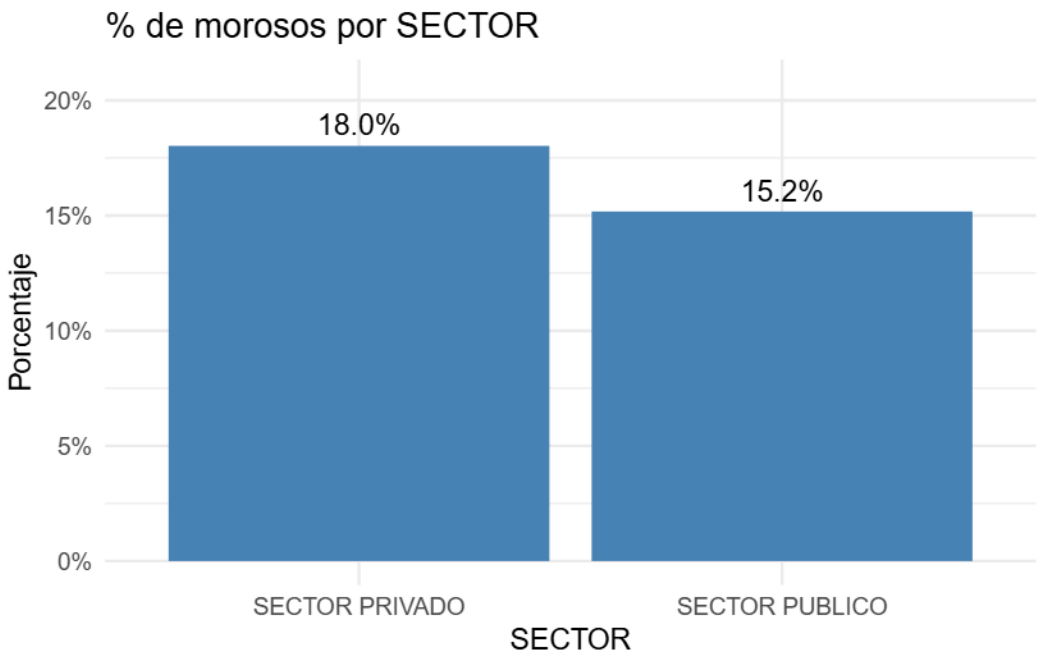

```
#Note que aunque hay diferencia es minima, no es ni un 1% la diferencia entre
#el porcentaje de hombres y mujeres morosas, lo indica que esta categoria no da
#mucha informacion para decidir si una persona va a ser morosa por su SEXO

graf.porcentajemorosidad(moras, "INDICADOR_ACTIVO")
```



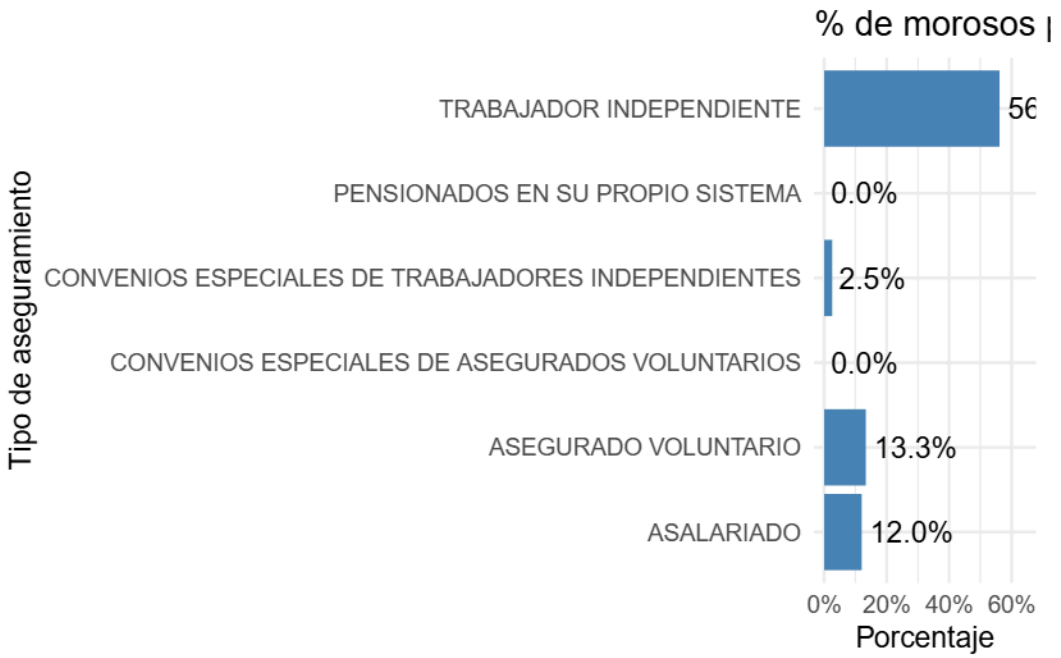

```
#Note que aquí la diferencia es bastante, las personas no activan son mucho mas
#morosas que las activas, lo cual tiene un poco de sentido, pues es de esperar
#que las personas activas tiendan a cumplir mas con los cumplimientos

graf.porcentajemorosidad(moras, "INDICADOR.EXTRANJERO")
```




```
#Este grafico nos muestra como las personas extranjeras tienden a ser mas morosas
#que las no extranjeras, las razones de esto pueden ser varias, como la estabilidad
#laboral, pues en muchas ocaciones los extranjeros no tienen salario o empleo fijo
#(en los locales tambien se da pero menos)

graf.porcentajemorosidad(moras, "SECTOR")
```




```
#El sector privado tiene una mayor cantidad de morosos (18%), lo cual se puede
#deber a que las personas en el sector publico suelen ser tener mejores salarios
#y mejor estabilidad lo cual lo llevan a cumplir mas con sus pagos.


distribucion.aseg <- moras %>%
  group_by(TIPO_ASEGURAMIENTO) %>%
  summarise(
    cantidad = n(),
    morosos = sum(INDICADOR_MOROSO == "SI"),
    porcentaje = morosos / cantidad
  )

ggplot(distribucion.aseg, aes(x = porcentaje, y = TIPO_ASEGURAMIENTO)) +
  geom_col(fill = "steelblue") +
  geom_text(aes(label = percent(porcentaje, accuracy = 0.1)),
            hjust = -0.1, size = 4) +
  scale_x_continuous(labels = percent_format(accuracy = 1),
                     limits = c(0, max(distribucion.aseg$porcentaje) * 1.15)) +
  labs(title = "% de morosos por tipo de aseguramiento",x = "Porcentaje",y = "Tipo de asegura
  ) +
  theme_minimal()
```

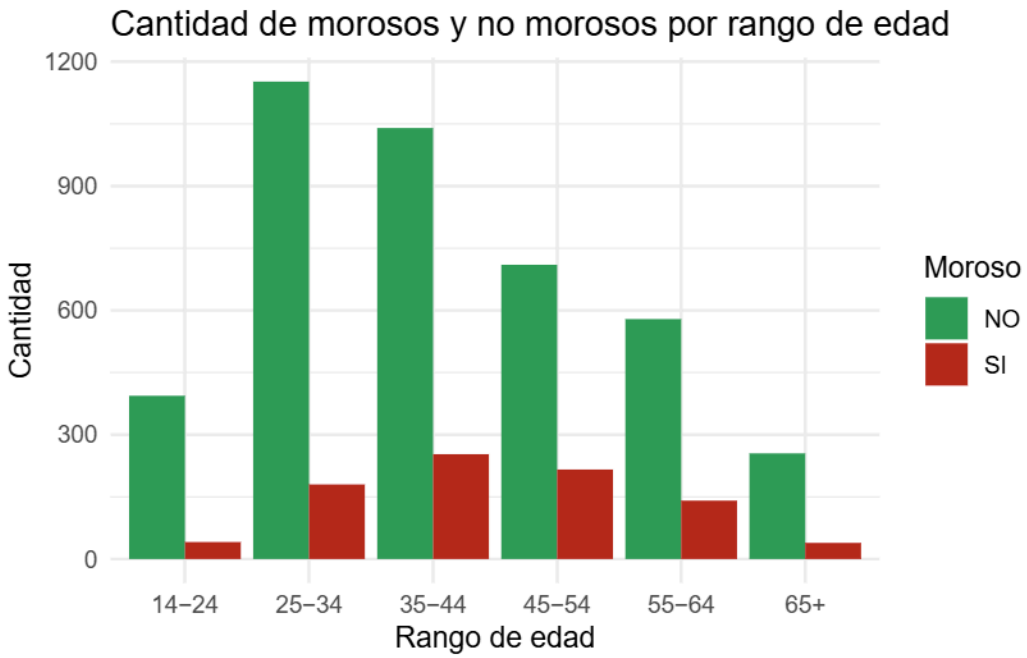



```
#Note que el trabajadores independientes es extremadamente alto (56%) lo cual se
#puede deber a varias razones, como a que esta población el salario que perciben
#suele variar mucho dependiendo las fechas del año y diversos factores, lo cual
#puede llevar a que muchas veces no puedan realizar los pagos.
```

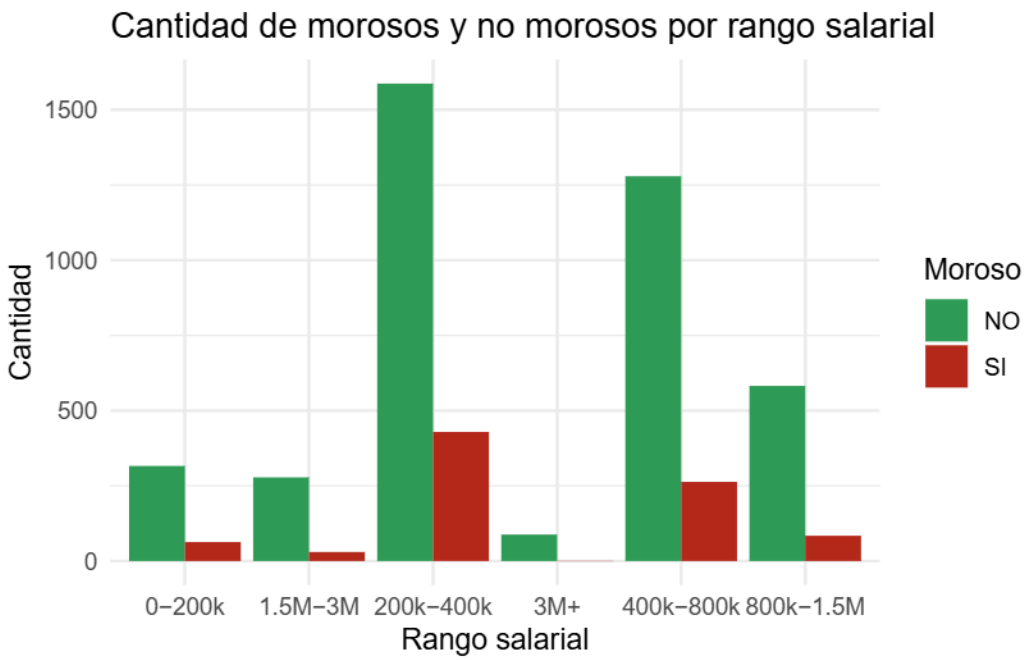
Parte8

```
#Se opta por crear categorias para poder graficar y analizar de mejor manera
#este dato, pues es un dato continuo y en un grafico de barras no se podria
#hacer bien. En salario se usa k para hacer referencia a miles.

moras <- moras %>%
  mutate(
    RANGO.EDAD = case_when(
      EDAD >= 14 & EDAD <= 24 ~ "14-24",
      EDAD >= 25 & EDAD <= 34 ~ "25-34",
      EDAD >= 35 & EDAD <= 44 ~ "35-44",
      EDAD >= 45 & EDAD <= 54 ~ "45-54",
      EDAD >= 55 & EDAD <= 64 ~ "55-64",
      EDAD >= 65              ~ "65+",
```

```
      TRUE ~ NA_character_
    )
  )

moras <- moras %>%
   mutate(
    RANGO_SALARIO = case_when(
      SALARIO >= 0       & SALARIO < 200000  ~ "0-200k",
      SALARIO >= 200000  & SALARIO < 400000  ~ "200k-400k",
      SALARIO >= 400000  & SALARIO < 800000  ~ "400k-800k",
      SALARIO >= 800000  & SALARIO < 1500000 ~ "800k-1.5M",
      SALARIO >= 1500000 & SALARIO < 3000000 ~ "1.5M-3M",
      SALARIO >= 3000000                     ~ "3M+",
      TRUE ~ NA_character_
    )
   )

edad.totales <- moras %>%
  group_by(RANGO.EDAD, INDICADOR_MOROSO) %>%
  summarise(cantidad = n(), .groups = "drop")


ggplot(edad.totales, aes(x = RANGO.EDAD, y = cantidad, fill = INDICADOR_MOROSO)) +
  geom_col(position = "dodge") +
  labs(
    title = "Cantidad de morosos y no morosos por rango de edad", x = "Rango de edad", y = "
    fill = "Moroso"
  ) +
  scale_fill_manual(values = c("SI" = "#B8291D", "NO" = "#309C58")) +
  theme_minimal()
```




```
#De aqui podemos ver que la mayor cantidad de no morosos son las personas entre
#los 25-34 y de morosos se encuentran en el rango de edad de los 35 a los 44 años,
#y que apartir de esta edad empiezan a decrecer, pero lo hacen tanto los morosos
#como los no morosos, por lo cual se podria concluir que es que hay menos personas
#en esos rangos de edad,

salarios.totales <- moras %>%
  group_by(RANGO_SALARIO, INDICADOR_MOROSO) %>%
  summarise(cantidad = n(), .groups = "drop")


ggplot(salarios.totales, aes(x = RANGO_SALARIO, y = cantidad, fill = INDICADOR_MOROSO)) +
  geom_col(position = "dodge") +
  labs(
    title = "Cantidad de morosos y no morosos por rango salarial", x = "Rango salarial", y =
    fill = "Moroso"
  ) +
  scale_fill_manual(values = c("SI" = "#B8291D", "NO" = "#309C58")) +
  theme_minimal()
```




```
#De este grafico lo que podemos ver es que en las personas con salario mayor a
#los 3 millones no hay morosos (salario tan alto se espera un cumplimiento con
#las obligaciones), y en general exxiste esta relacion entre salario y morosos,
#entre menor salario mas cantidad de personas morosas, sin embargo igual entre
#mas alto el salario menos personas totales, la unica excepción de esto esta en
#el grupo de 0 a 200k.
```

Parte9

```
#Es esto que se hace en R
```

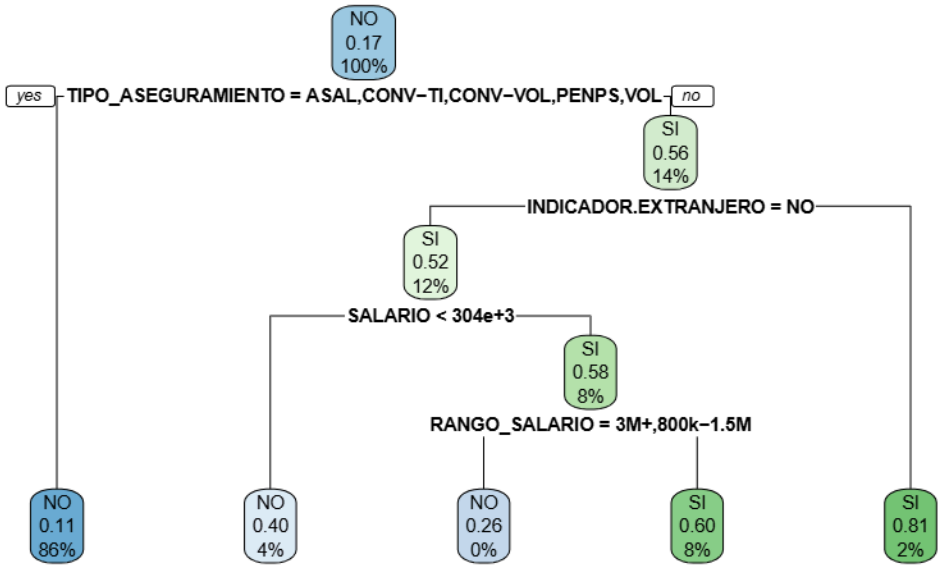
Parte 10

```
#Los nombres de los asegurados no se logran apreciar, por lo que se pondra una abreviatura

moras.arbol <- moras %>%
  mutate(
```

```
    TIPO_ASEGURAMIENTO = case_when(
      TIPO_ASEGURAMIENTO == "ASALARIADO" ~ "ASAL",
      TIPO_ASEGURAMIENTO == "PENSIONADOS EN SU PROPIO SISTEMA" ~ "PENPS",
      TIPO_ASEGURAMIENTO == "ASEGURADO VOLUNTARIO" ~ "VOL",
      TIPO_ASEGURAMIENTO == "TRABAJADOR INDEPENDIENTE" ~ "INDEP",
      TIPO_ASEGURAMIENTO == "CONVENIOS ESPECIALES DE TRABAJADORES INDEPENDIENTES" ~ "CONV-TI
      TIPO_ASEGURAMIENTO == "CONVENIOS ESPECIALES DE ASEGURADOS VOLUNTARIOS" ~ "CONV-VOL",
      TRUE ~ "OTRO"
    )
  )


#| eval: false
library(rpart)
library(rpart.plot)
```

```
Warning: package 'rpart.plot' was built under R version 4.5.2
```

```
tree <- rpart(INDICADOR_MOROSO ~ ., data = moras.arbol)

rpart.plot(
  tree)
```



```
#El árbol primero divide a las personas según su tipo de aseguramiento.
#En la mayoría de tipos de aseguramiento, casi todos aparecen como no morosos.
#Después, para los casos que quedan, el árbol revisa si la persona es extranjera.
#Dentro de ese grupo, las personas extranjeras tienen más probabilidad de ser morosas
#que las que no lo son.Luego, en otro nivel, el árbol utiliza el salario para
#separar aún más. En ese subgrupo, los salarios más altos se relacionan con más
#morosidad.En general, según este árbol, las variables que más influyen en la
#morosidad en este conjunto de datos son: el tipo de aseguramiento, si la persona
#es extranjera o no, y el salario.
```