Dataset Salaries

Datos Salaries

En el conjunto de datos Salaries del paquete carData tenemos información del salario de un período académico de nueve meses (2008-2009) de profesores asistentes, profesores asociados y profesores titulares de una institución de enseñanza superior en los Estados Unidos.

Los datos fueron recogidos como parte de un esfuerzo de la administración para monitorear diferencias salariales entre hombres y mujeres.

La información se encuentra recogida en las variables:

- rank: El rango del profesor
- discipline: departamento del profesor; A - Teóricas, B - Aplicadas
- yrs.since.phd: años desde la consecución del Ph.D. del profesor
- yrs.service: años en servicio del profesor
- sex: sexo
- salary: salario (correspondiente a nueve meses del período académico)

```
library(carData)
data("Salaries")
head(Salaries)
```

```
##        rank discipline yrs.since.phd yrs.service  sex salary
## 1      Prof          B            19          18 Male 139750
## 2      Prof          B            20          16 Male 173200
## 3  AsstProf          B             4           3 Male  79750
## 4      Prof          B            45          39 Male 115000
## 5      Prof          B            40          41 Male 141500
## 6 AssocProf          B             6           6 Male  97000
```

La administración quiere saber:

1. ¿Cuál es el número de varones y de mujeres en la muestra? ¿Cuántos profesores asistentes, asociados y titulares hay? En total y por sexo.
2. ¿Cómo se distribuye el tiempo de servicio de los profesores y de las profesoras por rango? ¿Quiénes presentan más variabilidad? ¿Existen outliers?
3. ¿Cómo se distribuye el salario de los profesores y de las profesoras? ¿Quiénes presentan más variabilidad? ¿Existen outliers?
4. Evaluar gráficamente qué relación existe entre el salario y los años de servicio en los varones y en las mujeres.
5. Finalmente da una interpretación global sobre la diferencia salarial entre varones y mujeres en esta institución. ¿Qué les recomendarías?

1. ¿Cuál es el número de varones y de mujeres en la muestra? ¿Cuántos profesores asistentes, asociados y titulares hay? En total y por sexo.

Analizamos el número de participantes en función de su sexo y rango.

```r
library(summarytools)

# ¿Cuál es el número de varones y de mujeres en la muestra?
# utiliza una función que te permita obtener las frecuencias
summary(Salaries$sex)
```

```
## Female   Male
##     39    358
```

```r
#  ¿Cuántos profesores asistentes, asociados y titulares hay?
#  utiliza la misma función que antes para obtener las frecuencias
summary(Salaries$rank)
```

```
##  AsstProf AssocProf      Prof
##        67        64       266
```

```r
# combina las dos variables para obtener sus frecuencias
# utiliza una nueva función para obtener las frecuencias para cada celda (combinación)
# en una tabla de contingencia

with(Salaries,ctable(sex,rank))
```

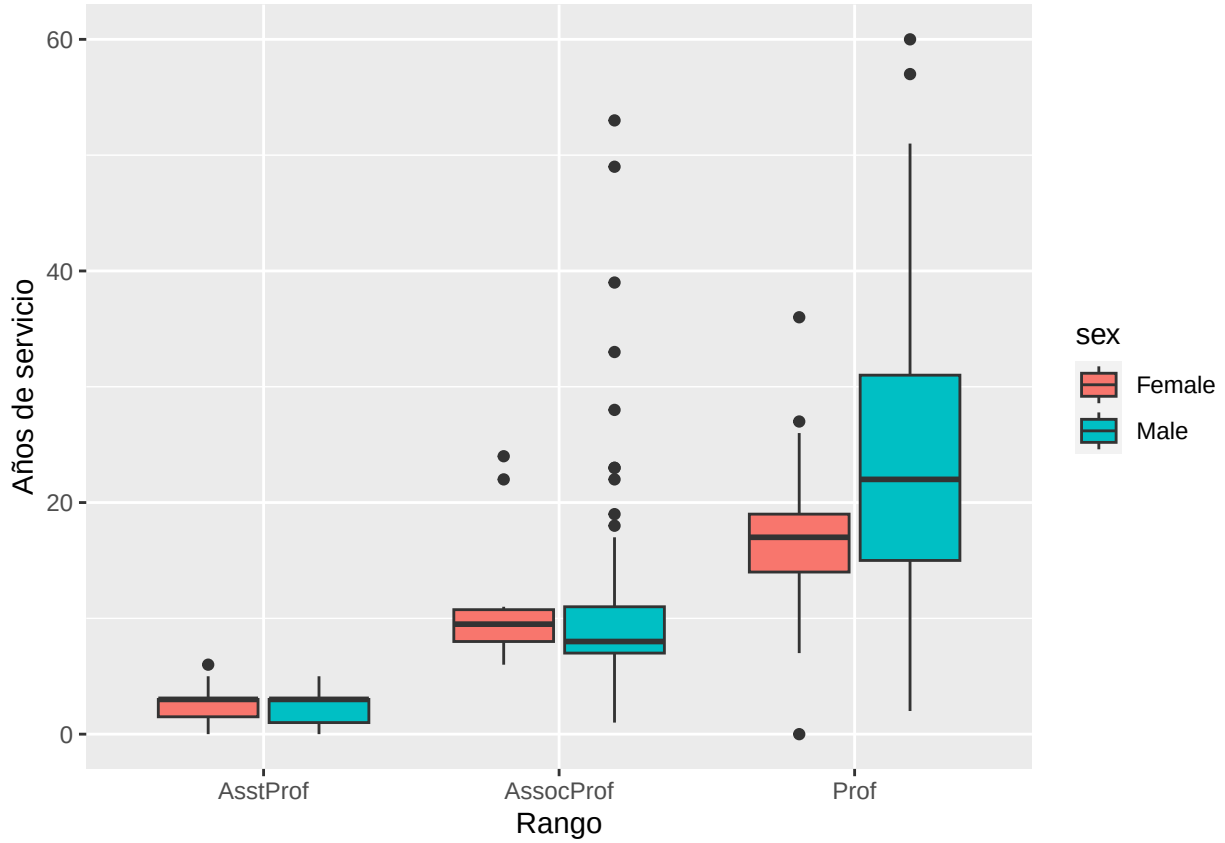
```
## Cross-Tabulation, Row Proportions
## sex * rank
## Data Frame: Salaries
##
## -------- ------ ------------ ------------ ------------- --------------
##             rank     AsstProf    AssocProf          Prof          Total
##      sex
##   Female          11 (28.2%)   10 (25.6%)    18 (46.2%)    39 (100.0%)
##     Male          56 (15.6%)   54 (15.1%)   248 (69.3%)   358 (100.0%)
##    Total          67 (16.9%)   64 (16.1%)   266 (67.0%)   397 (100.0%)
## -------- ------ ------------ ------------ ------------- --------------
```

2. ¿Cómo se distribuye el tiempo de servicio de los profesores y de las profesoras por rango? ¿Quiénes presentan más variabilidad? ¿Existen outliers?

Investigamos el tiempo de servicio por sexo y rango de los profesores.

```r
# ¿Cómo se distribuye el tiempo de servicio de los profesores y de las profesoras por rango?

library(ggplot2)
ggplot(Salaries, aes(x=rank, y=yrs.service, fill=sex))+
  geom_boxplot()+labs(y="Años de servicio",x="Rango") #crea el gráfico para evaluar la distribución
```



Interpretación de los resultados

Hay una mayor variabilidad en los hombres que están en el rango *Profesor*, ya que los "bigotes" que se extienden desde las cajas son muy largos. Además, el rango intercuartílico de esta variable en hombres es muy grande, lo que indica que la distribución del 50% de los valores es muy amplia. La mediana en ese rango entre hombres y mujeres es muy diferente, siendo mucho menor en mujeres que en hombres. El percentil 75 en hombres es mucho mayor que en mujeres. En mujeres el percentil 75 se encuentra por debajo de los 20 años de servicio, mientras que en hombres el valor central está por encima de los 20 años, y los valores por encima de esa medida central están entre los 20 y 30 años de servicio.

Además, hay muchos **outliers** o valores atípicos en el rango de *Profesor asociado*, categoría masculina, e indican que aunque los profesores asociados llevan años de servicio con una mediana en torno a 10 (tanto hombres como mujeres), existen valores atípicos de hombres con ese rango que llevan años de servicio entre 20 y 60 años y que no han pasado a un rango superior. La distribución en mujeres en este rango es simétrica, encontrándose el 50% de los casos por encima de la mediana, y el otro 50 por debajo. El percentil 75 en este rango es parecido entre hombres y mujeres.

En el caso del rango de *Profesor asistente*, tanto hombres como mujeres representan una distribución parecida, con una mediana en torno a 3 años. La distribución es asimétrica positiva en ambos casos, estando todos los valores por debajo de la mediana.

```
# ¿Quiénes presentan más variabilidad?

library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
```

```
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.1     v tidyr     1.3.0
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x tibble::view()  masks summarytools::view()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
Salaries %>%
  group_by(sex, rank) %>%     # agrupación
  summarize(media = mean(yrs.service),      # resumen
            mediana = median(yrs.service),
            varianza = var(yrs.service),
            desviacion = sd(yrs.service),
            CV = (sd(yrs.service) / mean(yrs.service) * 100) )
```

```
## `summarise()` has grouped output by 'sex'. You can override using the `.groups`
## argument.
```

```
## # A tibble: 6 x 7
## # Groups:   sex [2]
##   sex    rank      media mediana varianza desviacion    CV
##   <fct>  <fct>     <dbl>   <dbl>    <dbl>      <dbl> <dbl>
## 1 Female AsstProf   2.55     3       3.47       1.86  73.2
## 2 Female AssocProf 11.5      9.5    39.2        6.26  54.4
## 3 Female Prof      17.1     17      65.6        8.10  47.3
## 4 Male   AsstProf   2.34     3       2.05       1.43  61.2
## 5 Male   AssocProf 12.0      8     115.        10.7   88.9
## 6 Male   Prof      23.2     22     137.        11.7   50.4
```

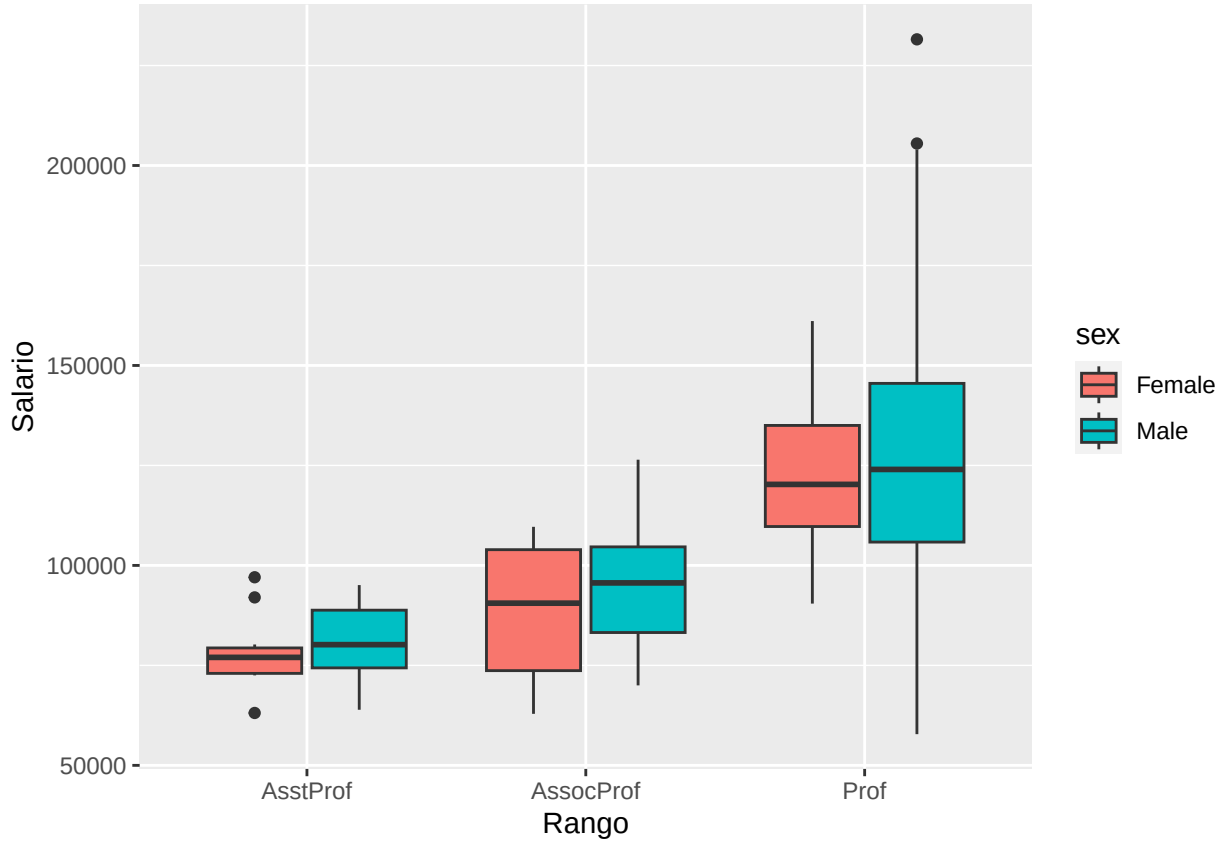
Interpretación de los resultados

En los estadísticos descriptivos, se puede observar que en cuanto **años de servicio**, la media es más alta en mujeres en el rango de *Profesor asistente*, pero en los demás rangos es mayor en hombres. La varianza aumenta a medida que aumentamos el rango, tanto en hombres como en mujeres, siendo mayor en hombres excepto en el rango más bajo. Esto quiere decir que hay una mayor dispersión de los datos de años de servicio a medida que aumentamos el rango.

3. ¿Cómo se distribuye el salario de los profesores y de las profesoras por rango? ¿Quiénes presentan más variabilidad? ¿Existen outliers?

Ahora investigamos el salario por sexo de manera similar a la anterior.

```
# Gráfico
ggplot(Salaries, aes(x=rank, y=salary, fill=sex))+
  geom_boxplot()+labs(y="Salario",x="Rango") #crea el gráfico para evaluar la distribución
```



Interpretación de los resultados

La mayor variabilidad la representan los hombres en la categoría de *Profesor*, ya que el gráfico de cajas presenta un mayor rango intercuartílico y bigotes más largos. También presentan outliers de salarios elevados. Además de presentar más variabilidad, son los que más salario cobran, ya que el percentil 75 es mayor que las demás condiciones.

En el rango de *profesor asociado* tienen un mayor rango intercuartílico las mujeres, siendo el percentil 25 más bajo que en el caso de los hombres. Los hombres en este rango tienen mayor variabilidad, ya que sus bigotes son más largos. El percentil 75 en ambos caso es muy parecido, lo que quiere decir que el 50% de los datos por encima de la mediana en ambos casos llega al mismo salario tanto en hombres como mujeres.

En el rango más bajo, el de *profesor asistente*, se puede observar que el gráfico de cajas asociado al salario de las mujeres tiene un rango intercuartílico muy pequeño, con algunos outliers por encima y por debajo de la mediana, y su percentil 75 está al mismo nivel que la mediana de los datos de los hombres. En el caso de los hombres, el gráfico de cajas presenta simetría y el rango intercuartñilico, así como el percentil 75 y la mediana, son mayores que las mujeres en este rango.

```
#Descriptivos
Salaries %>%
  group_by(sex, rank) %>%    # agrupación
  summarize(media = mean(salary),        # resumen
            mediana = median(salary),
            varianza = var(salary),
            desviacion = sd(salary),
            CV = (sd(salary) / mean(salary) * 100) )
```

```
## `summarise()` has grouped output by 'sex'. You can override using the `.groups`
```

```
## argument.
```

```
## # A tibble: 6 x 7
## # Groups:   sex [2]
##   sex    rank        media mediana   varianza desviacion    CV
##   <fct>  <fct>       <dbl>   <dbl>      <dbl>      <dbl> <dbl>
## 1 Female AsstProf   78050.  77000   87834311.      9372. 12.0
## 2 Female AssocProf  88513.  90556. 322751501.     17965. 20.3
## 3 Female Prof      121968. 120258. 384928020.     19620. 16.1
## 4 Male   AsstProf   81311.  80182   62431217.      7901.  9.72
## 5 Male   AssocProf  94870.  95626. 166173174.     12891. 13.6
## 6 Male   Prof      127121. 123996  796018943.     28214. 22.2
```

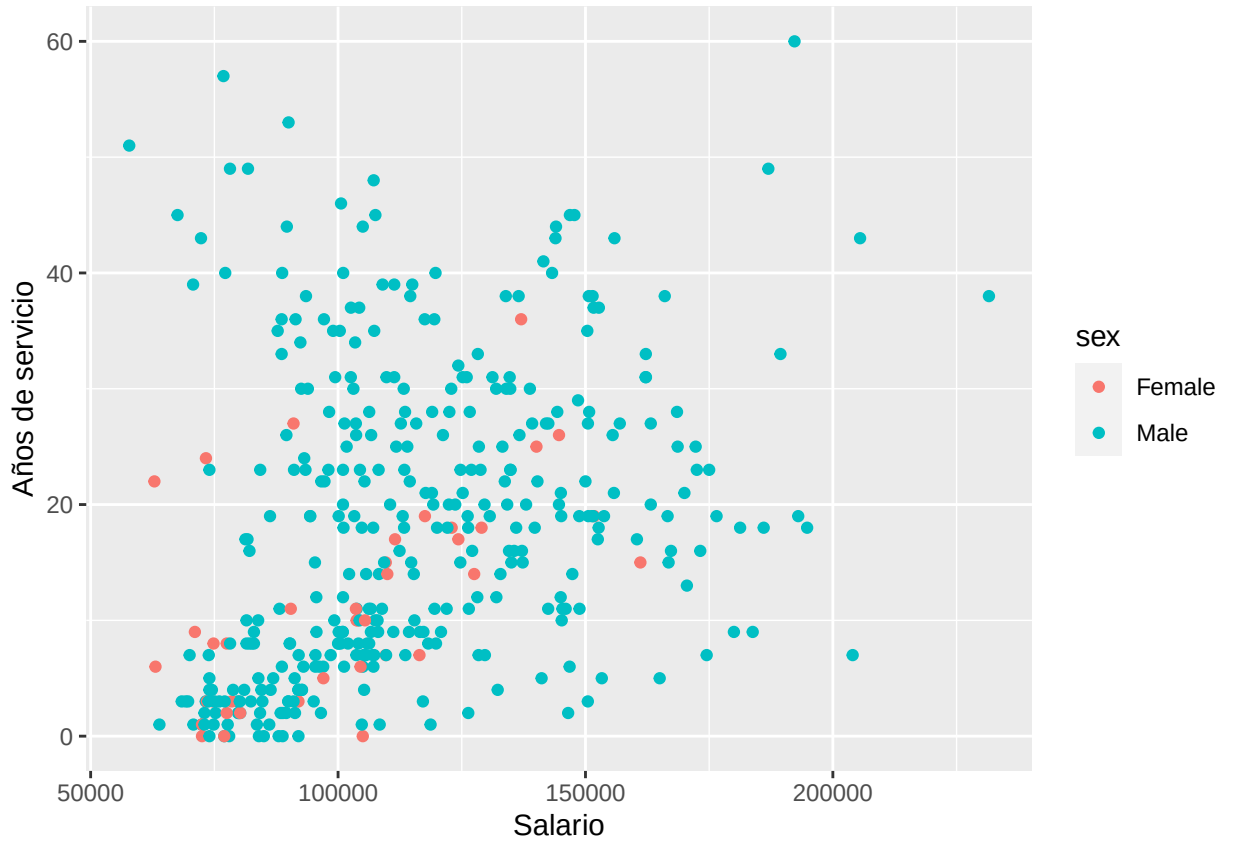
Interpretación de los resultados

En los estadísticos descriptivos, se puede observar como la media y la mediana de los datos de salario son menores en mujeres que en hombres en todos los rangos. En el caso de la varianza, esta es mayor en mujeres que en hombres para cada rango, lo que indica una mayor dispersión de los datos alrededor de la media en el caso del salario de las mujeres.

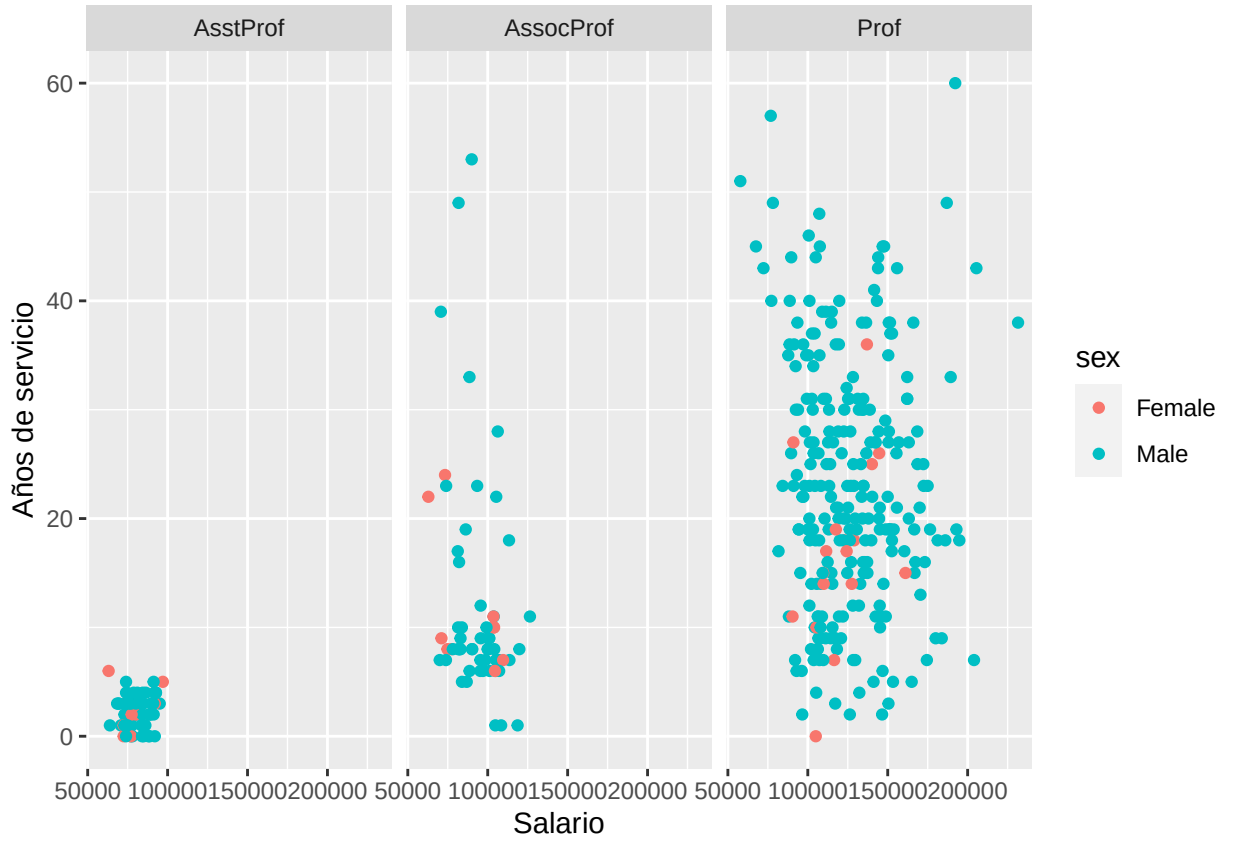
4. Evaluar gráficamente qué relación existe entre el salario y los años de servicio en los varones y en las mujeres.

Queremos evaluar la relación entre 2 variables numéricas (el salario y los años de servicio) diferenciando/coloreando según el sexo del encuestado.

```
# Gráfico
ggplot(Salaries, aes(x = salary, y = yrs.service, col=sex)) +
  geom_point() + labs(y="Años de servicio", x="Salario")  #crea el gráfico de dispersión
```




```
# si queremos dibujarlo por tipo de profesor (rank)
# Gráfico
ggplot(Salaries, aes(x = salary, y = yrs.service, col=sex)) +
  geom_point() +  #crea el gráfico de dispersión
  facet_wrap(~rank) + # divide el gráfico en subpantallas
    labs(y="Años de servicio", x="Salario")
```



5. Finalmente da una interpretación global sobre la diferencia salarial entre varones y mujeres en esta institución. ¿Qué les recomendarías?

Según el análisis de estos datos, está claro que hay diferencias en el salario de hombres y mujeres en todos los rangos. Creo que debería haber más conciencia sobre la importancia de la equidad salarial y la igualdad de oportunidades dentro de esta institución. También, habría que evaluar las políticas en cuanto a promoción del personal, y ver si hay sesgos de género o no, para que, en su caso, ajustarlas de manera justa, ya que, además de cobrar menos, las mujeres se quedan en los puestos más bajos, siendo muchas menos que hombres a medida que se avanza de rango.