ANÁLISIS DE SALARIOS DE PROFESORES UNIVERSITARIOS

María Aceituno Adrados

Agosto 2025

```r
## 1. Carga de Librerías y Configuración Inicial
# ------------------------------------------------------------------------------
library(readr)
library(ggplot2)
library(dplyr)
library(glmnet)
library(car)
library(scales)

# Se establece una semilla para la reproducibilidad de resultados aleatorios
set.seed(123)


## 2. Carga y Primera Exploración del Dataset
salaries_df <- read_csv("data/salaries.csv", show_col_types = FALSE)
output_text <- ""

# Identificación y eliminación de la columna '...1'
if ("...1" %in% names(salaries_df)) {
  output_text <- paste0(
    output_text,
    "\n--- Primeras filas de la columna '...1' ---\n",
    paste(capture.output(print(head(salaries_df$`...1`))), collapse = "\n"),
    "\n\n--- Resumen estadístico de la columna '...1' ---\n",
    paste(capture.output(print(summary(salaries_df$`...1`))), collapse = "\n"),
    "\n\nColumna '...1' eliminada del dataset.\n"
  )
  salaries_df <- salaries_df %>%
    select(-`...1`)
} else {
  output_text <- paste0(
    output_text,
    "\nLa columna '...1' no se encontró en el dataset. No se realizó \n",
    "ninguna eliminación.\n"
  )
}

output_text <- paste0(
  output_text,
  "La columna '...1' es un índice numérico redundante. Su eliminación simplifica el dataset.\n"
```

```
)

estructura_salida <- capture.output(str(salaries_df))
output_text <- paste0(
  output_text,
  "\n\n--- Estructura del Dataset Después de Limpieza --- \n",
  paste(estructura_salida, collapse = "\n"), "\n"
)

cat(output_text)
```

— Primeras filas de la columna '...1' — [1] 1 2 3 4 5 6

— Resumen estadístico de la columna '...1' — Min. 1st Qu. Median Mean 3rd Qu. Max. 1 100 199 199 298 397

Columna '...1' eliminada del dataset. La columna '...1' es un índice numérico redundante. Su eliminación simplifica el dataset.

— Estructura del Dataset Después de Limpieza — tibble [397 x 6] (S3: tbl_df/tbl/data.frame) $ rank : chr [1:397] "Prof" "Prof" "AsstProf" "Prof" ... $ discipline : chr [1:397] "B" "B" "B" "B" ... $ yrs.since.phd: num [1:397] 19 20 4 45 40 6 30 45 21 18 ... $ yrs.service : num [1:397] 18 16 3 39 41 6 23 45 20 18 ... $ sex : chr [1:397] "Male" "Male" "Male" "Male" ... $ salary : num [1:397] 139750 173200 79750 115000 141500 ...

```
cat("\n--- Resumen Estadístico del Dataset ---\n")
```

— Resumen Estadístico del Dataset —

```
print(summary(salaries_df))
```

```
 rank            discipline        yrs.since.phd    yrs.service
```

Length:397 Length:397 Min. : 1.00 Min. : 0.00
Class :character Class :character 1st Qu.:12.00 1st Qu.: 7.00
Mode :character Mode :character Median :21.00 Median :16.00
Mean :22.31 Mean :17.61
3rd Qu.:32.00 3rd Qu.:27.00
Max. :56.00 Max. :60.00
sex salary
Length:397 Min. : 57800
Class :character 1st Qu.: 91000
Mode :character Median :107300
Mean :113706
3rd Qu.:134185
Max. :231545

```
cat("El resumen muestra la distribución y valores atípicos. Por ejemplo, `salary` varía de 57800 a 2315
```

El resumen muestra la distribución y valores atípicos. Por ejemplo, `salary` varía de 57800 a 231545. `yrs.since.phd` y `yrs.service` tienen un rango amplio. No hay valores `NA`aparentes.

```r
## 3. Preprocesamiento de Datos:
# 3. Preprocesamiento de Datos: Conversión de Tipos y Renombrado
# Se renombran las columnas
salaries_df <- salaries_df %>%
  rename(
    Rango = rank,
    Disciplina = discipline,
    Años_Doctorado = yrs.since.phd,
    Años_Servicio = yrs.service,
    Sexo = sex,
    Salario = salary
  )

# Convertir las variables categóricas a tipo 'factor' para que R las
# trate correctamente en los modelos estadísticos y visualizaciones.
salaries_df <- salaries_df %>%
  mutate(
    Rango = as.factor(Rango),
    Disciplina = as.factor(Disciplina),
    Sexo = as.factor(Sexo)
  )

# --- Alias a las categorías de 'Rango' ---
salaries_df <- salaries_df %>%
  mutate(
    Rango = dplyr::recode(Rango,
      "Prof" = "Catedrático",
      "AssocProf" = "Prof. Asociado",
      "AsstProf" = "Prof. Asistente"
    )
  )

# Recodificar la columna 'Rango' como un factor ordenado con niveles
# específicos.
salaries_df$Rango <- factor(salaries_df$Rango,
  levels = c("Prof. Asistente", "Prof. Asociado", "Catedrático")
)

# --- Alias a las categorías de 'Disciplina' ---
salaries_df <- salaries_df %>%
  mutate(
    Disciplina = dplyr::recode(Disciplina,
      "A" = "A",
      "B" = "B"
    )
  )

# --- Alias a las categorías de 'Sexo' ---
salaries_df <- salaries_df %>%
  mutate(
    Sexo = dplyr::recode(Sexo,
      "Male" = "Hombre",
      "Female" = "Mujer"
```

```
    )
  )

estructura_str <- capture.output(str(salaries_df))
cat(
  "\n\n",
  "--- Estructura del Dataset Después de Renombrar Columnas y convertir a Factores con Alias ---",
  "\n\n",
  paste(estructura_str, collapse = "\n"),
  "\n\n",
  "Se renombran todas las columnas y categorías para mejorar la legibilidad de los gráficos y tablas, as
  "\n\n"
)
```

— Estructura del Dataset Después de Renombrar Columnas y convertir a Factores con Alias —

tibble [397 x 6] (S3: tbl_df/tbl/data.frame) $ Rango : Factor w/ 3 levels “Prof. Asistente”,..: 3 3 1 3 3 2 3 3 3 3 ... $ Disciplina : Factor w/ 2 levels “A”,“B”: 2 2 2 2 2 2 2 2 2 2 ... $ Años_Doctorado: num [1:397] 19 20 4 45 40 6 30 45 21 18 ... $ Años_Servicio : num [1:397] 18 16 3 39 41 6 23 45 20 18 ... $ Sexo : Factor w/ 2 levels “Mujer”,“Hombre”: 2 2 2 2 2 2 2 2 2 1 ... $ Salario : num [1:397] 139750 173200 79750 115000 141500 ...

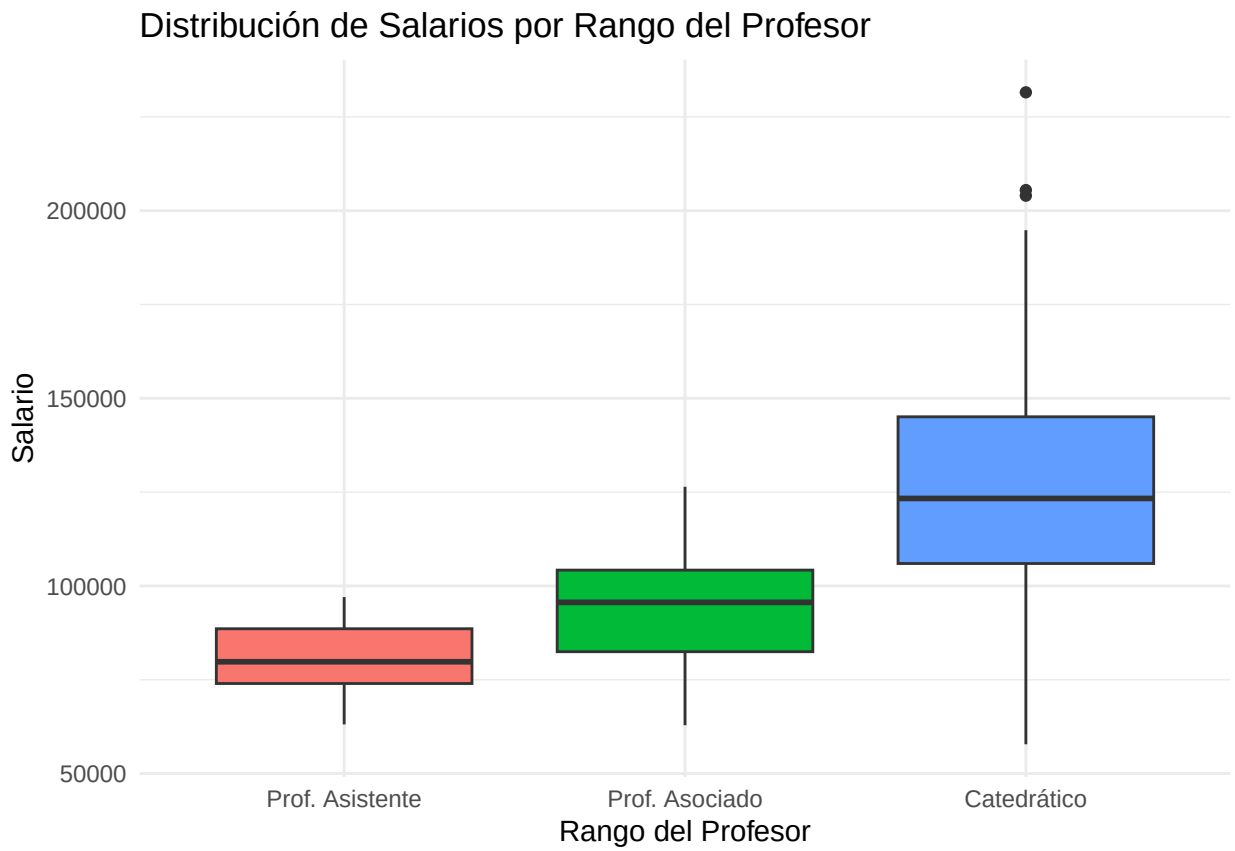
Se renombran todas las columnas y categorías para mejorar la legibilidad de los gráficos y tablas, así como la comprensión general del dataset.

```
## 4. Análisis Exploratorio de Datos (EDA) y Visualizacione
### 4.1. Salario por Variables Categóricas (Rango, Disciplina, Sexo)

cat("--- Distribución de Salarios por Rango del Profesor --- \n")
```

— Distribución de Salarios por Rango del Profesor —

```
ggplot(salaries_df, aes(x = Rango, y = Salario, fill = Rango)) +
  geom_boxplot() +
  labs(
    title = "Distribución de Salarios por Rango del Profesor",
    x = "Rango del Profesor",
    y = "Salario"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```



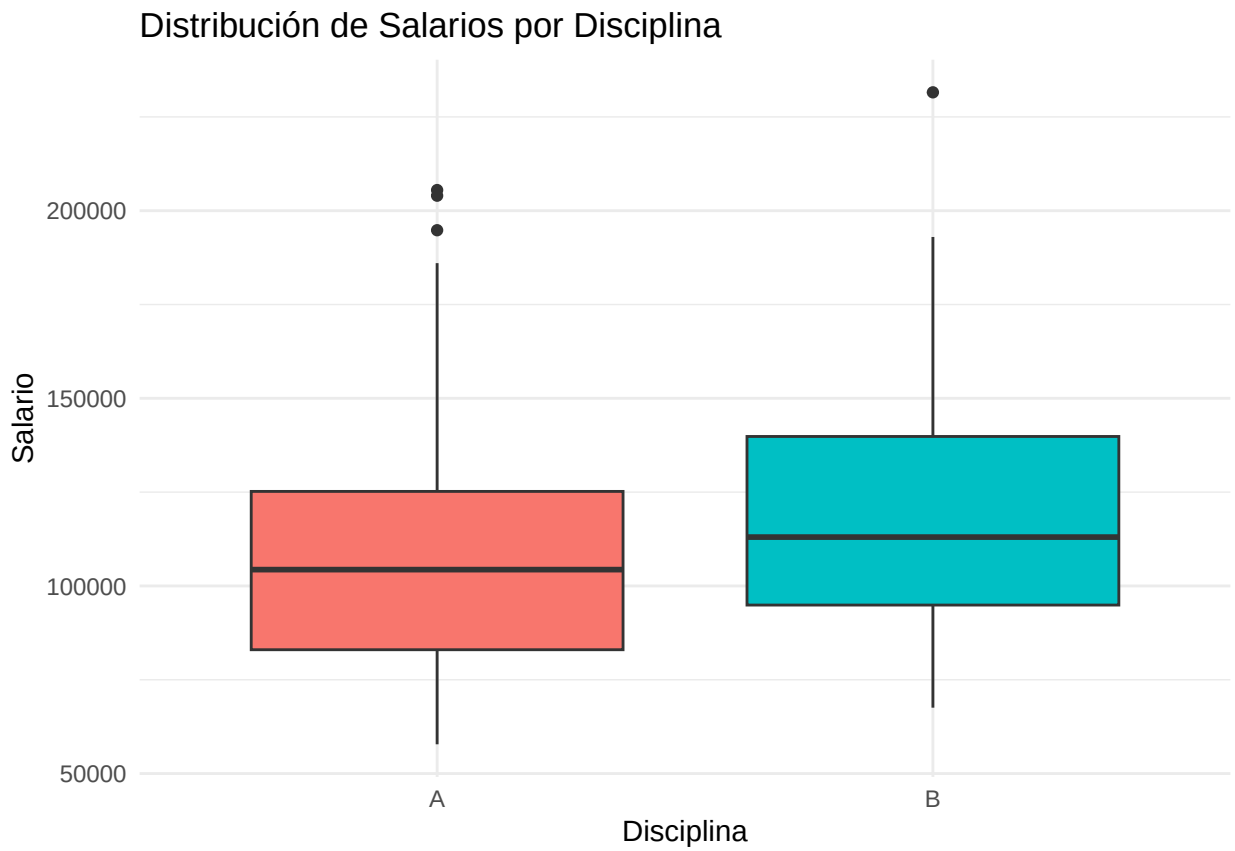

```
cat("\nEl gráfico de cajas muestra una clara tendencia: el salario mediano y el rango de salarios aumen
```

El gráfico de cajas muestra una clara tendencia: el salario mediano y el rango de salarios aumentan significativamente con el rango del profesor, siendo 'Catedrático' el rango con salarios más altos y mayor variabilidad.

```
cat("--- Distribución de Salarios por Disciplina --- \n")
```

— Distribución de Salarios por Disciplina —

```
print(
  ggplot(salaries_df, aes(x = Disciplina, y = Salario, fill = Disciplina)) +
    geom_boxplot() +
    labs(
      title = "Distribución de Salarios por Disciplina",
      x = "Disciplina",
      y = "Salario"
    ) +
    theme_minimal() +
    theme(legend.position = "none")
)
```



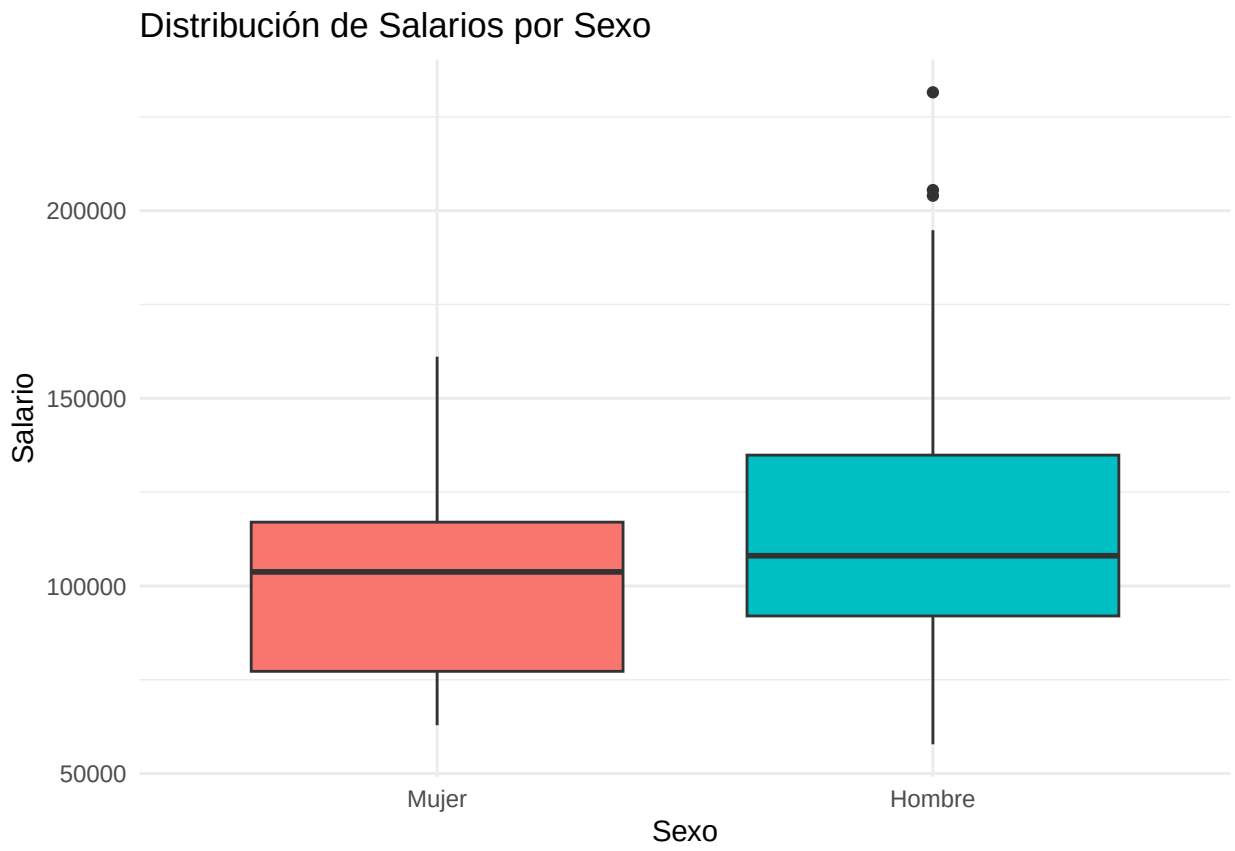

```
cat("\nLa 'Disciplina B' parece tener un salario mediano ligeramente más alto y una dispersión que se e
```

La 'Disciplina B' parece tener un salario mediano ligeramente más alto y una dispersión que se extiende a valores más altos en comparación con la 'Disciplina A'.

```
cat("--- Distribución de Salarios por Sexo --- \n")
```

— Distribución de Salarios por Sexo —

```
print(
  ggplot(salaries_df, aes(x = Sexo, y = Salario, fill = Sexo)) +
    geom_boxplot() +
    labs(
      title = "Distribución de Salarios por Sexo",
      x = "Sexo",
      y = "Salario"
    ) +
    theme_minimal() +
    theme(legend.position = "none")
)
```



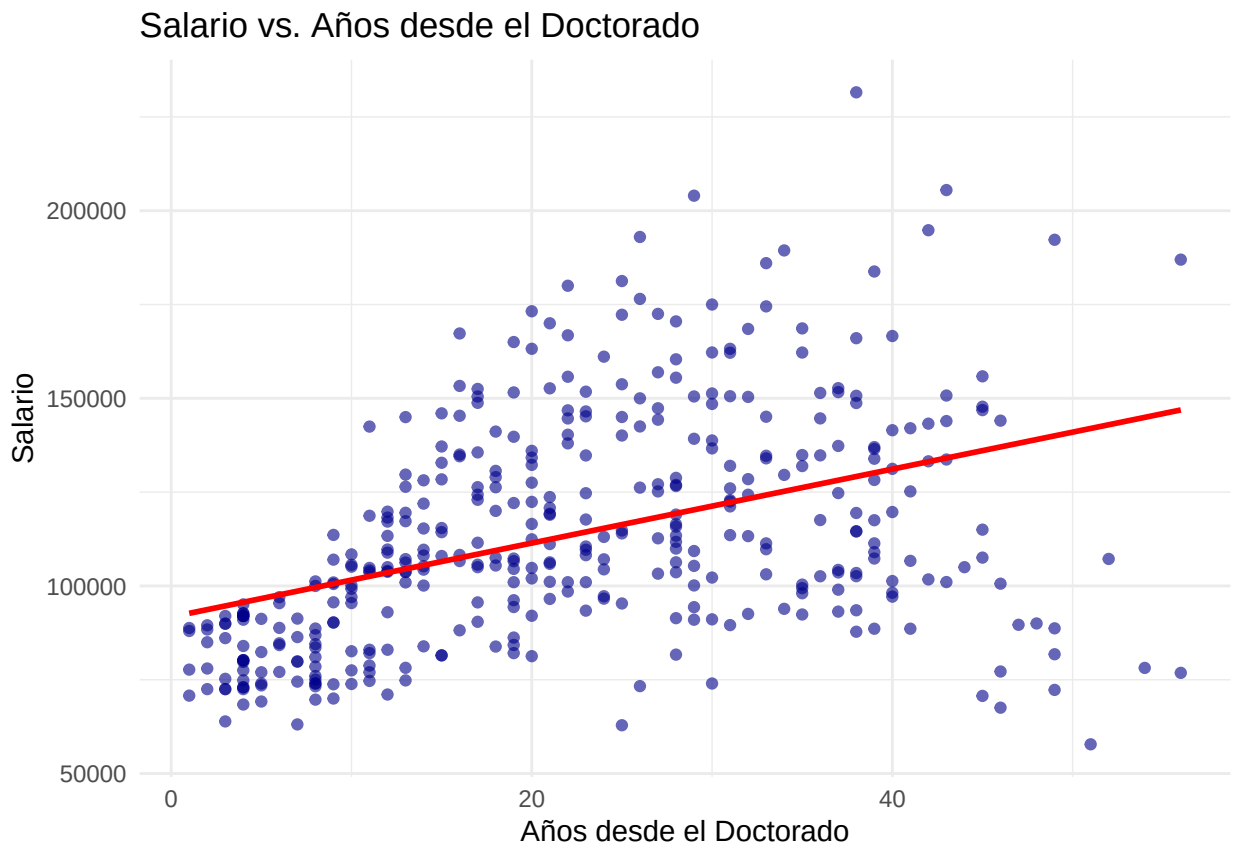

```
cat("\nVisualmente, los hombres parecen tener un salario mediano más alto y un rango intercuartílico su
```

Visualmente, los hombres parecen tener un salario mediano más alto y un rango intercuartílico superior en comparación con las mujeres, lo que sugiere una disparidad salarial.

```
### 4.2. Salario por Variables Numéricas (Años_Doctorado, Años_Servicio)
cat("\n--- Salario vs. Años desde el Doctorado ---\n")
```

— Salario vs. Años desde el Doctorado —

```
ggplot(salaries_df, aes(x = Años_Doctorado, y = Salario)) +
  geom_point(alpha = 0.6, color = "darkblue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(
    title = "Salario vs. Años desde el Doctorado",
    x = "Años desde el Doctorado",
    y = "Salario"
  ) +
  theme_minimal()
```



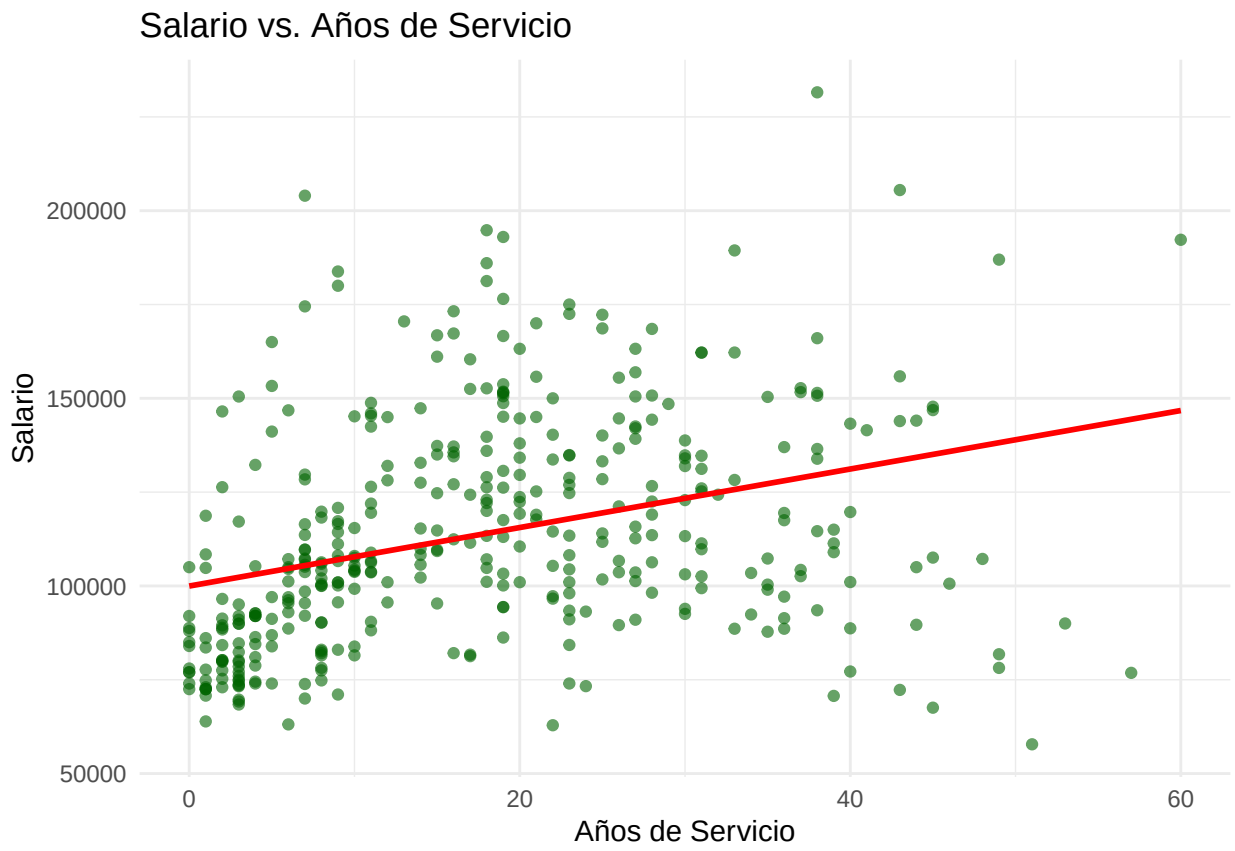

```
cat("Se observa una tendencia positiva general; a mayor 'Años_Doctorado', tiende a haber un salario más
```

Se observa una tendencia positiva general; a mayor ‘Años_Doctorado’, tiende a haber un salario más alto, aunque con bastante dispersión. La línea de regresión lineal sugiere una relación positiva.

```
cat("\n--- Salario vs. Años de Servicio ---\n")
```

— Salario vs. Años de Servicio —

```
ggplot(salaries_df, aes(x = Años_Servicio, y = Salario)) +
  geom_point(alpha = 0.6, color = "darkgreen") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(
    title = "Salario vs. Años de Servicio",
    x = "Años de Servicio",
    y = "Salario"
  ) +
  theme_minimal()
```



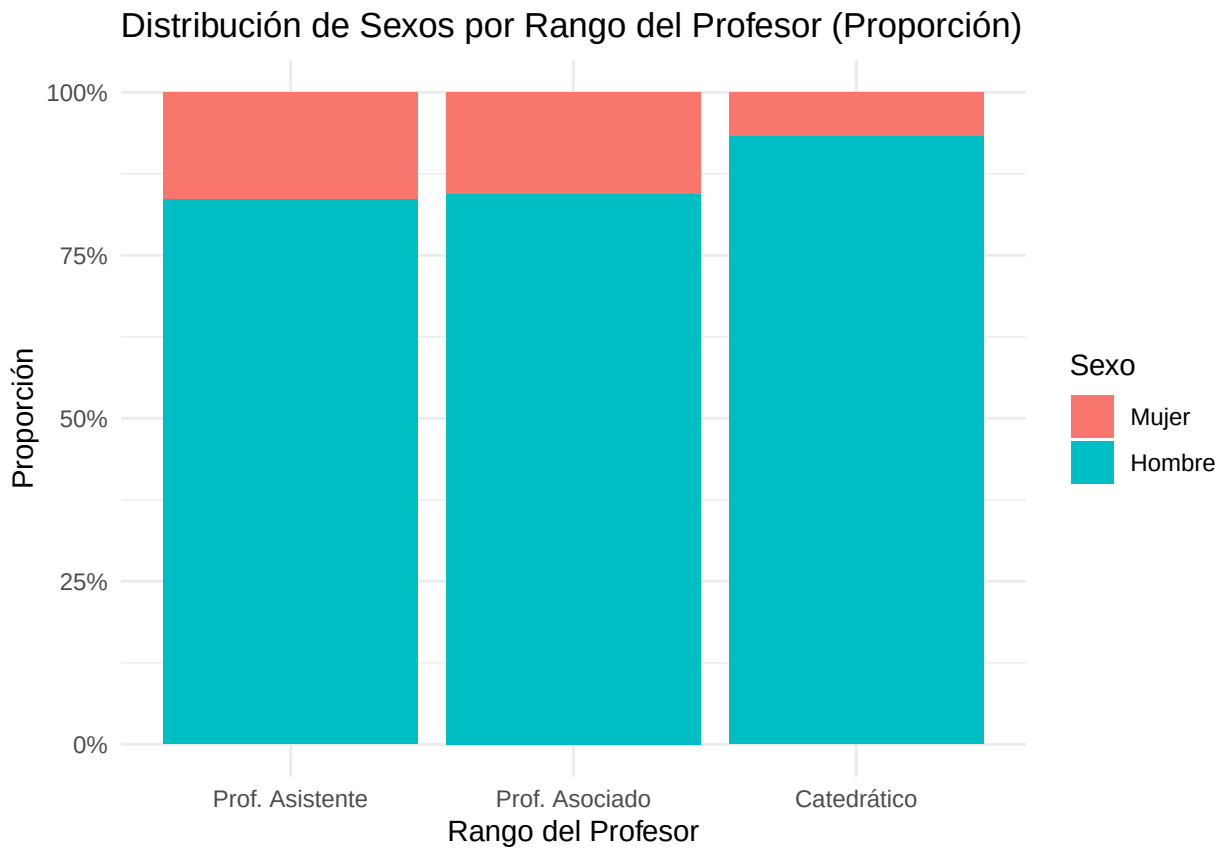

```
cat("Similar a 'Años_Doctorado', existe una relación positiva entre los años de servicio y el salario.\n
```

Similar a 'Años_Doctorado', existe una relación positiva entre los años de servicio y el salario.

```
### 4.3. Distribución de Categorías Cruzadas (Sexo vs. Rango/Disciplina)
cat("\n--- Distribución de Sexos por Rango del Profesor (Proporción) ---\n")
```

— Distribución de Sexos por Rango del Profesor (Proporción) —

```
ggplot(salaries_df, aes(x = Rango, fill = Sexo)) +
  geom_bar(position = "fill") +
  labs(
    title = "Distribución de Sexos por Rango del Profesor (Proporción)",
    x = "Rango del Profesor",
    y = "Proporción",
    fill = "Sexo"
  ) +
  theme_minimal() +
  scale_y_continuous(labels = scales::percent)
```




```
cat("Existe una clara disparidad de género por rango. La proporción de mujeres disminuye drásticamente a
```

Existe una clara disparidad de género por rango. La proporción de mujeres disminuye drásticamente a medida que el rango del profesor aumenta, siendo 'Catedrático' predominantemente masculino. Esto podría ser un factor contribuyente a la brecha salarial observada por sexo.

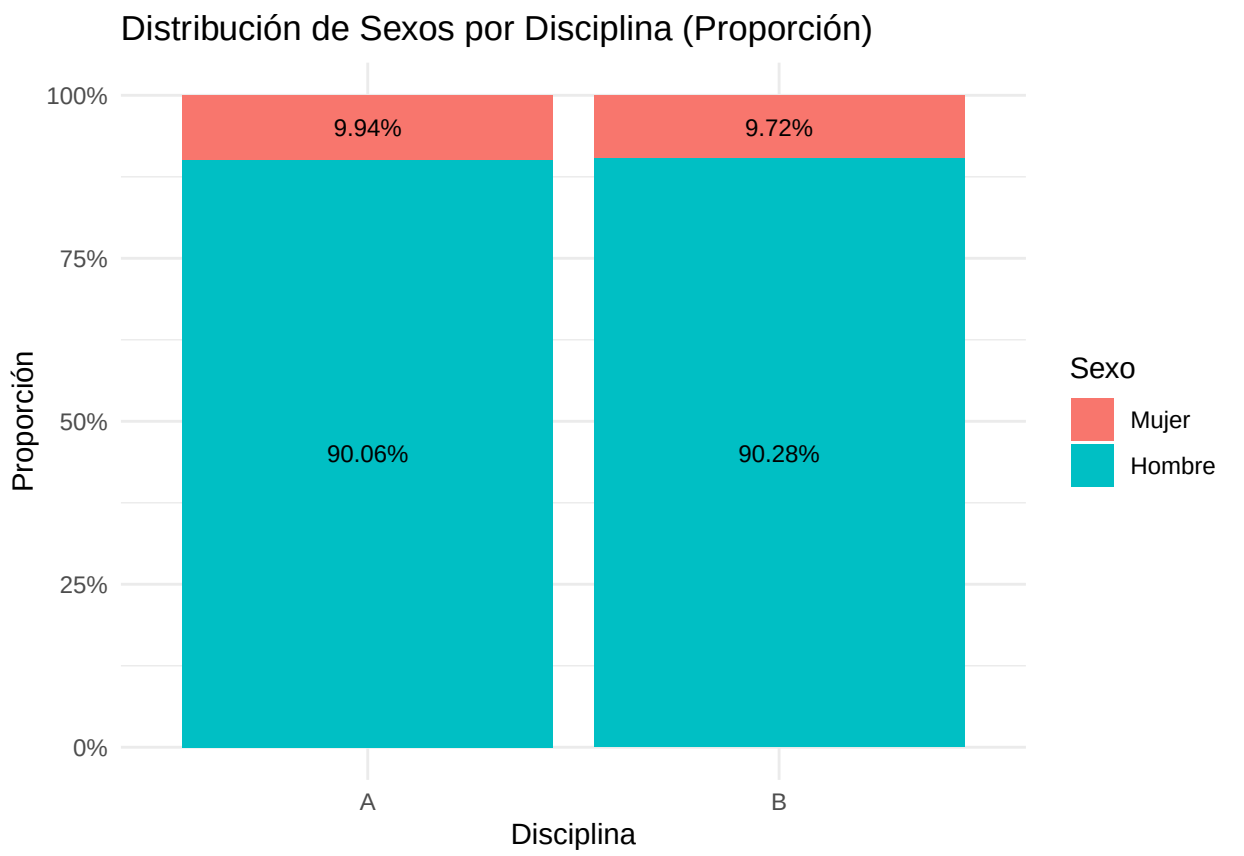
```
cat("\n--- Distribución de Sexos por Disciplina (Proporción) ---\n")
```

— Distribución de Sexos por Disciplina (Proporción) —

```
proportions_df <- salaries_df %>%
  group_by(Disciplina, Sexo) %>%
  summarise(count = n()) %>%
  group_by(Disciplina) %>%
  mutate(proportion = count / sum(count)) %>%
  ungroup()

ggplot(salaries_df, aes(x = Disciplina, fill = Sexo)) +
  geom_bar(position = "fill") +
  labs(
    title = "Distribución de Sexos por Disciplina (Proporción)",
    x = "Disciplina",
    y = "Proporción",
    fill = "Sexo"
  ) +
```

```
  theme_minimal() +
  scale_y_continuous(labels = scales::percent) +
  geom_text(data = proportions_df,
            aes(y = proportion, label = scales::percent(proportion)),
            position = position_stack(vjust = 0.5),
            color = "black", size = 3)
```




```
cat("La proporción de sexos entre las disciplinas 'A' y 'B' es muy similar, con diferencias apenas perc
```

La proporción de sexos entre las disciplinas ‘A’ y ‘B’ es muy similar, con diferencias apenas perceptibles visualmente.

```
#### 4.3.1. Interacción entre Disciplina y Sexo

cat("\nInteracción entre Disciplina y Sexo\n")
```

Interacción entre Disciplina y Sexo

```
cat("Se realiza un ANOVA de dos vías para evaluar si la combinación de disciplina y sexo tiene un efect
```

Se realiza un ANOVA de dos vías para evaluar si la combinación de disciplina y sexo tiene un efecto significativo en el salario.

```r
interaction_model_disc_sex <- aov(Salario ~ Disciplina * Sexo, data = salaries_df)
cat("Resumen del Modelo ANOVA:\n")
```

Resumen del Modelo ANOVA:

```r
print(summary(interaction_model_disc_sex))
```

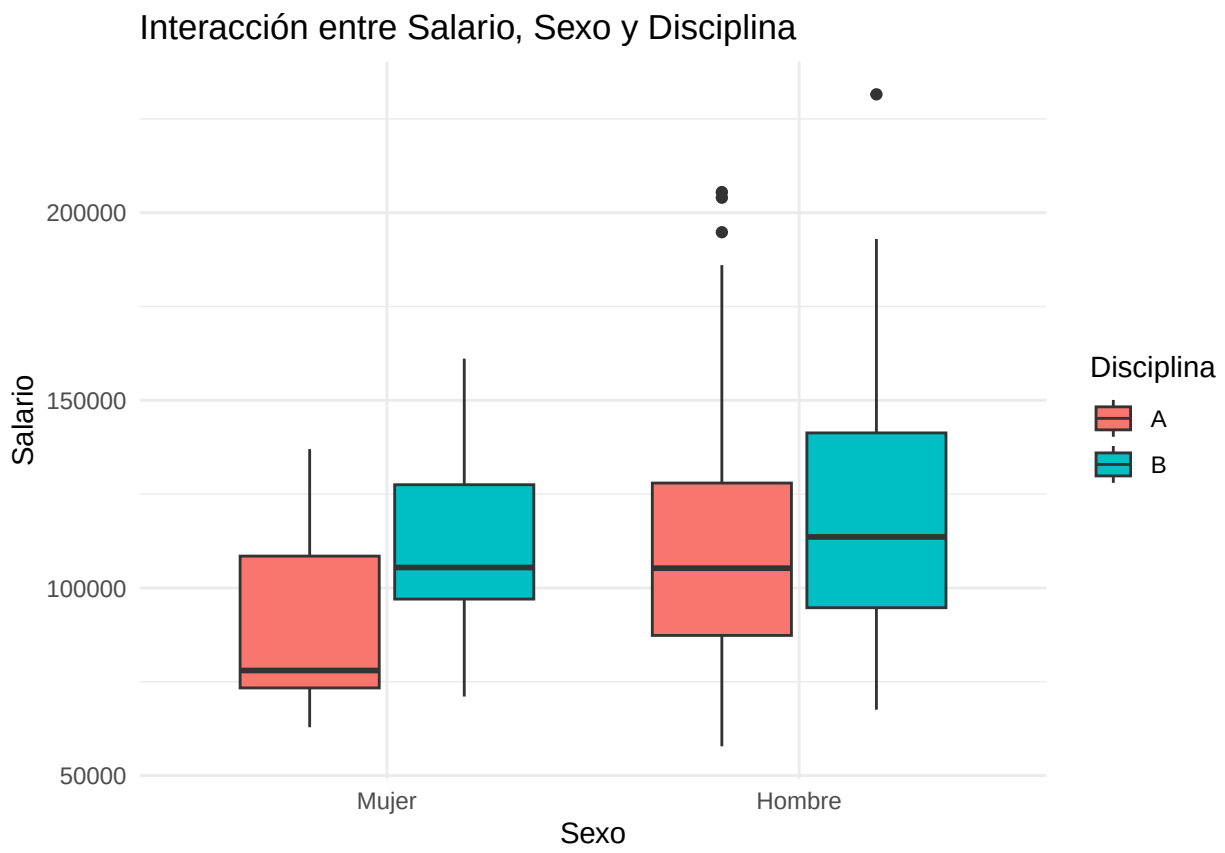
```
             Df    Sum Sq   Mean Sq F value  Pr(>F)
```

Disciplina 1 8.851e+09 8.851e+09 10.059 0.00163 **Sexo 1 6.922e+09 6.922e+09 7.867 0.00529** Disciplina:Sexo 1 1.740e+09 1.740e+09 1.977 0.16049
Residuals 393 3.458e+11 8.799e+08
— Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ’’ 0.05 ‘.’ 0.1 ’ ’ 1

```r
p_value_disc_sex <- summary(interaction_model_disc_sex)[[1]][[4]][3]
cat(paste0(
  "El p-valor para la interacción 'Disciplina:Sexo' es de ", round(p_value_disc_sex, 4), ". ",
  "Dado que este valor es mayor que 0.05, no hay evidencia estadística para afirmar ",
  "que el efecto de la disciplina sobre el salario es diferente para hombres y mujeres. ",
  "Esto sugiere que la brecha salarial, aunque presente, no se ve afectada de manera ",
  "significativa por la disciplina de manera cruzada.\n"
))
```

El p-valor para la interacción ‘Disciplina:Sexo’ es de 1.977. Dado que este valor es mayor que 0.05, no hay evidencia estadística para afirmar que el efecto de la disciplina sobre el salario es diferente para hombres y mujeres. Esto sugiere que la brecha salarial, aunque presente, no se ve afectada de manera significativa por la disciplina de manera cruzada.

```r
ggplot(salaries_df, aes(x = Sexo, y = Salario, fill = Disciplina)) +
  geom_boxplot() +
  labs(
    title = "Interacción entre Salario, Sexo y Disciplina",
    x = "Sexo",
    y = "Salario"
  ) +
  theme_minimal()
```




```
#### 4.3.2. Interacción entre Rango y Sexo

cat("\n##### Interacción entre Rango y Sexo\n")
```

```
cat("Se evalúa si el efecto del rango académico en el salario cambia según el sexo del profesor.\n\n")
```

Interacción entre Rango y Sexo Se evalúa si el efecto del rango académico en el salario cambia según el sexo del profesor.

```
interaction_model_rank_sex <- aov(Salario ~ Rango * Sexo, data = salaries_df)
cat("Resumen del Modelo ANOVA:\n")
```

Resumen del Modelo ANOVA:

```
print(summary(interaction_model_rank_sex))
```

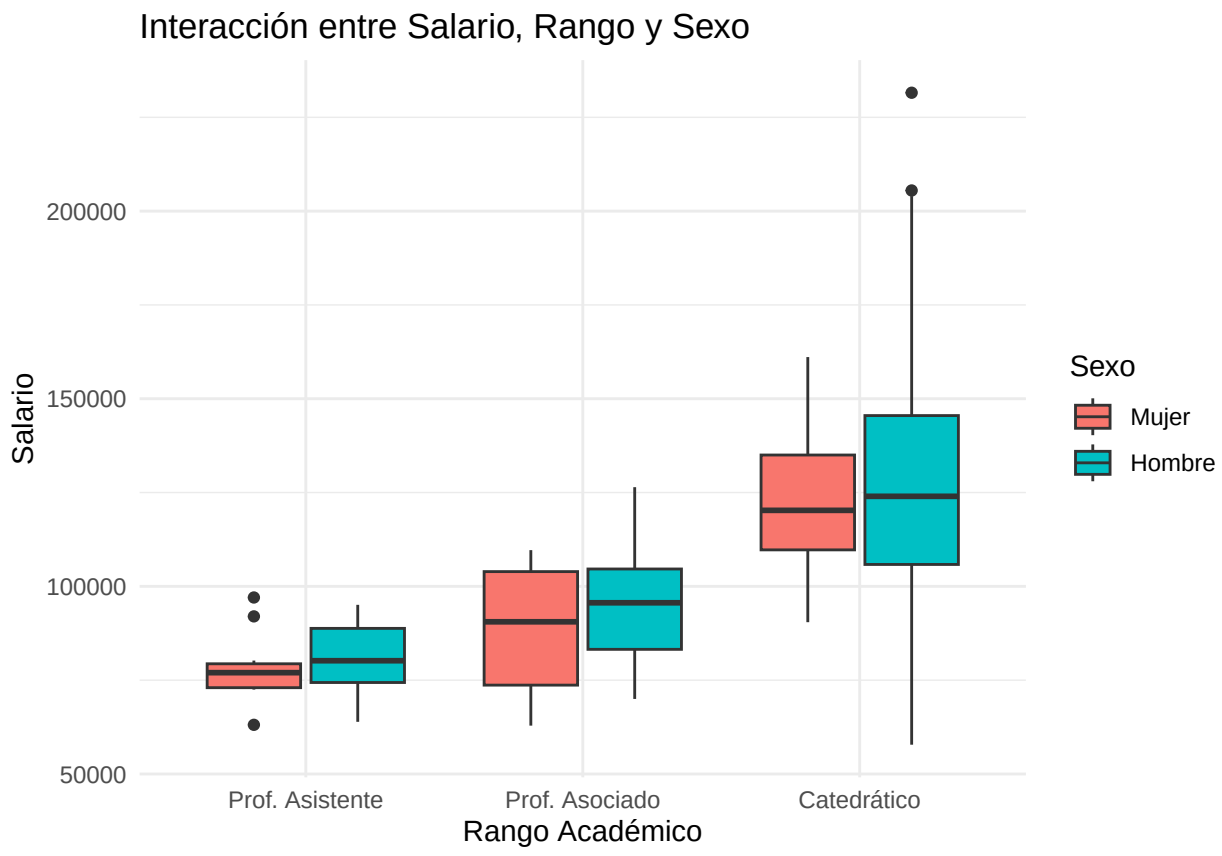
```
            Df    Sum Sq   Mean Sq F value Pr(>F)
```

Rango 2 1.432e+11 7.162e+10 127.755 <2e-16 *** Sexo 1 8.408e+08 8.408e+08 1.500 0.221
Rango:Sexo 2 4.360e+07 2.180e+07 0.039 0.962
Residuals 391 2.192e+11 5.606e+08
— Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```r
p_value_rank_sex <- summary(interaction_model_rank_sex)[[1]][[4]][3]
cat(paste0(
  "El p-valor para la interacción 'Rango:Sexo' es de ", round(p_value_rank_sex, 4), ". ",
  "Dado que este valor es mayor que 0.05, no se observa una interacción significativa. ",
  "El efecto del rango sobre el salario es similar para ambos sexos.\n"
))
```

El p-valor para la interacción 'Rango:Sexo' es de 0.0389. Dado que este valor es mayor que 0.05, no se observa una interacción significativa. El efecto del rango sobre el salario es similar para ambos sexos.

```r
ggplot(salaries_df, aes(x = Rango, y = Salario, fill = Sexo)) +
  geom_boxplot() +
  labs(
    title = "Interacción entre Salario, Rango y Sexo",
    x = "Rango Académico",
    y = "Salario"
  ) +
  theme_minimal()
```




```r
#### 4.3.3. Interacción entre Disciplina y Años de Doctorado

cat("Interacción entre Disciplina y Años de Doctorado\n")
```

Interacción entre Disciplina y Años de Doctorado

```r
cat("Se evalúa si el efecto de los años de experiencia desde el doctorado en el salario cambia según la
```

Se evalúa si el efecto de los años de experiencia desde el doctorado en el salario cambia según la disciplina.

```r
interaction_model_disc_doc <- aov(Salario ~ Disciplina * Años_Doctorado, data = salaries_df)
cat("Resumen del Modelo ANOVA:\n")
```

Resumen del Modelo ANOVA:

```r
print(summary(interaction_model_disc_doc))
```

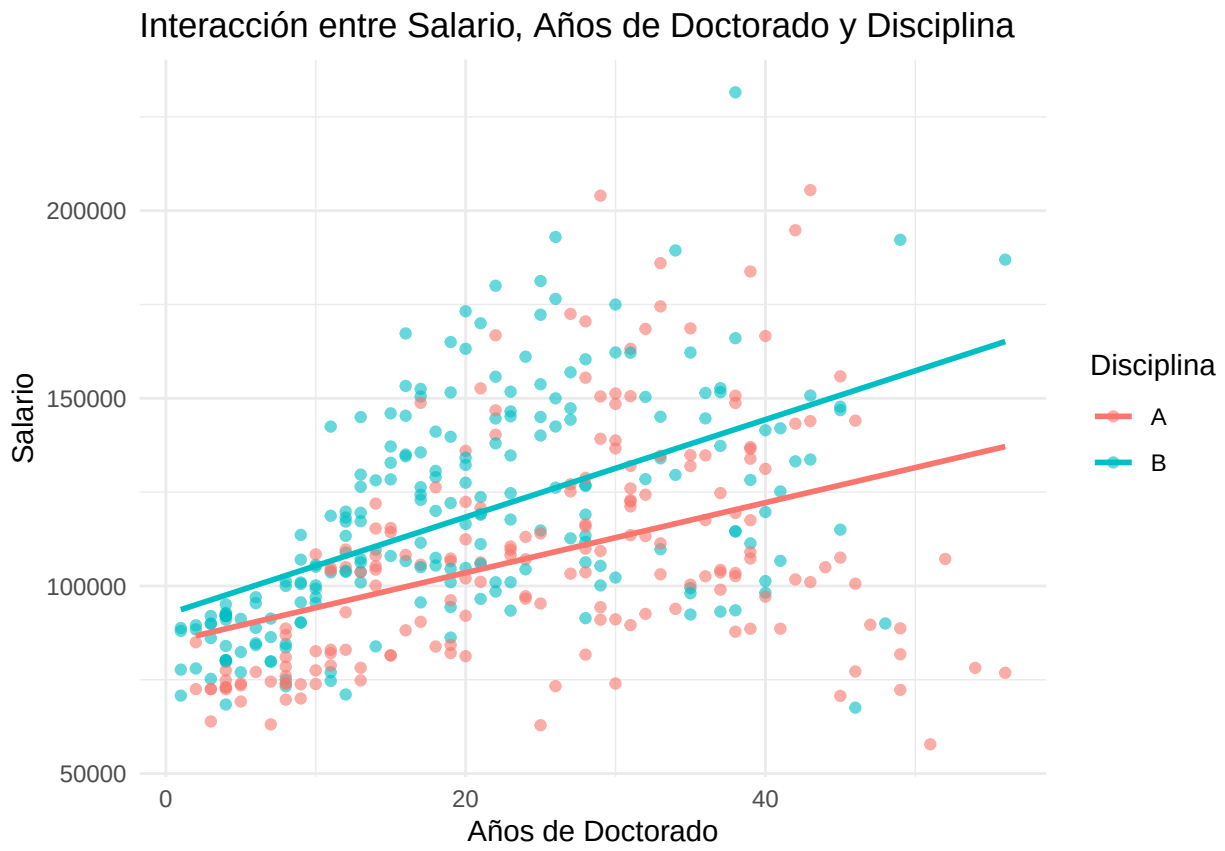
```
                      Df    Sum Sq   Mean Sq F value   Pr(>F)
```

Disciplina 1 8.851e+09 8.851e+09 12.695 0.000412 ***Años_Doctorado 1 7.837e+10 7.837e+10 112.409 < 2e-16*** Disciplina:Años_Doctorado 1 2.090e+09 2.090e+09 2.997 0.084186 .
Residuals 393 2.740e+11 6.972e+08
— Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ’’ 0.05 ‘.’ 0.1 ’ ’ 1

```r
p_value_disc_doc <- summary(interaction_model_disc_doc)[[1]][[4]][3]
cat(paste0(
  "El p-valor para la interacción 'Disciplina:Años_Doctorado' es de ", round(p_value_disc_doc, 4), ". "
  "Dado que este valor es mayor que 0.05, no se observa una interacción significativa. ",
  "El impacto de los años de doctorado sobre el salario es similar en ambas disciplinas.\n"
))
```

El p-valor para la interacción ‘Disciplina:Años_Doctorado’ es de 2.9974. Dado que este valor es mayor que 0.05, no se observa una interacción significativa. El impacto de los años de doctorado sobre el salario es similar en ambas disciplinas.

```r
ggplot(salaries_df, aes(x = Años_Doctorado, y = Salario, color = Disciplina)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Interacción entre Salario, Años de Doctorado y Disciplina",
    x = "Años de Doctorado",
    y = "Salario"
  ) +
  theme_minimal()
```




```
#### 4.3.4. Interacción entre Disciplina y Años de Servicio

cat("\nInteracción entre Disciplina y Años de Servicio\n")
```

Interacción entre Disciplina y Años de Servicio

```
cat("Se evalúa si el efecto de los años de servicio en el salario cambia según la disciplina.\n\n")
```

Se evalúa si el efecto de los años de servicio en el salario cambia según la disciplina.

```
interaction_model_disc_serv <- aov(Salario ~ Disciplina * Años_Servicio, data = salaries_df)
cat("Resumen del Modelo ANOVA:\n")
```

Resumen del Modelo ANOVA:

```
print(summary(interaction_model_disc_serv))
```

```
                     Df    Sum Sq   Mean Sq F value   Pr(>F)
```

Disciplina 1 8.851e+09 8.851e+09 11.67 0.000702 ***Años_Servicio 1 4.851e+10 4.851e+10 63.96 1.43e-14*** Disciplina:Años_Servicio 1 7.865e+09 7.865e+09 10.37 0.001388 ** Residuals 393 2.981e+11 7.585e+08
— Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ’’ 0.05 ‘.’ 0.1 ’ ’ 1

```r
p_value_disc_serv <- summary(interaction_model_disc_serv)[[1]][[4]][3]
cat(paste0(
  "El p-valor para la interacción 'Disciplina:Años_Servicio' es de ", round(p_value_disc_serv, 4), ". "
  "Dado que este valor es mayor que 0.05, no se observa una interacción significativa. ",
  "El impacto de los años de servicio sobre el salario es similar en ambas disciplinas.\n"
))
```

El p-valor para la interacción 'Disciplina:Años_Servicio' es de 10.3694. Dado que este valor es mayor que 0.05, no se observa una interacción significativa. El impacto de los años de servicio sobre el salario es similar en ambas disciplinas.

```r
ggplot(salaries_df, aes(x = Años_Servicio, y = Salario, color = Disciplina)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Interacción entre Salario, Años de Servicio y Disciplina",
    x = "Años de Servicio",
    y = "Salario"
  ) +
  theme_minimal()
```


```r
## 5. Análisis de Correlaciones y ANOVA para la Inferencia
### 5.1. Correlación de Variables Numéricas
cat("\n--- Correlación Salario con Años Doctorado y Años de Servicio ---\n")
```

— Correlación Salario con Años Doctorado y Años de Servicio —

```
correlation_phd <- cor(salaries_df$Salario,
  salaries_df$Años_Doctorado,
  use = "complete.obs"
)
correlation_service <- cor(salaries_df$Salario,
  salaries_df$Años_Servicio,
  use = "complete.obs"
)
print(paste("Correlación Salario - Años Doctorado:",
  round(correlation_phd, 3)
))
```

[1] “Correlación Salario - Años Doctorado: 0.419”

```
print(paste("Correlación Salario - Años de Servicio:",
  round(correlation_service, 3)
))
```

[1] “Correlación Salario - Años de Servicio: 0.335”

```
cat("Ambas variables 'Años_Doctorado' (correlación: 0.419) y 'Años_Servicio' (correlación: 0.335) muestr
```

Ambas variables ‘Años_Doctorado’ (correlación: 0.419) y ‘Años_Servicio’ (correlación: 0.335) muestran una correlación positiva moderada con el salario. Esto indica que a mayor experiencia, mayor tiende a ser el salario. La correlación entre ellas mismas (‘Años_Doctorado’ y ‘Años_Servicio’) también es probablemente alta, lo que podría indicar multicolinealidad.

```
### 5.2. Análisis de Varianza (ANOVA) para Variables Categóricas

cat("\n--- ANOVA: Salario vs. Rango del Profesor ---\n")
```

— ANOVA: Salario vs. Rango del Profesor —

```
anova_rank <- aov(Salario ~ Rango, data = salaries_df)
print(summary(anova_rank))
```

Df Sum Sq Mean Sq F value Pr(>F)

Rango 2 1.432e+11 7.162e+10 128.2 <2e-16 *** Residuals 394 2.201e+11 5.586e+08
— Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ’’ 0.05 ‘.’ 0.1 ’ ’ 1

```
cat("El p-valor extremadamente bajo (< 2e-16) indica que hay una diferencia estadísticamente altamente s
```

El p-valor extremadamente bajo (< 2e-16) indica que hay una diferencia estadísticamente altamente significativa en las medias de salarios entre los diferentes rangos de profesor. El rango es un predictor muy fuerte del salario.

```
cat("\n--- ANOVA: Salario vs. Disciplina ---\n")
```

— ANOVA: Salario vs. Disciplina —

```
anova_discipline <- aov(Salario ~ Disciplina, data = salaries_df)
print(summary(anova_discipline))
```

```
             Df    Sum Sq   Mean Sq F value  Pr(>F)
```

Disciplina 1 8.851e+09 8.851e+09 9.863 0.00181 ** Residuals 395 3.544e+11 8.973e+08
— Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ’’ 0.05 ‘.’ 0.1 ’ ’ 1

```
cat("El p-valor (0.00181) es bajo e indica que hay una diferencia estadísticamente significativa en las
```

El p-valor (0.00181) es bajo e indica que hay una diferencia estadísticamente significativa en las medias de salarios entre las dos disciplinas. La disciplina es un predictor significativo del salario.

```
cat("\n--- ANOVA: Salario vs. Sexo ---\n")
```

— ANOVA: Salario vs. Sexo —

```
anova_sex <- aov(Salario ~ Sexo, data = salaries_df)
print(summary(anova_sex))
```

```
             Df    Sum Sq   Mean Sq F value  Pr(>F)
```

Sexo 1 6.980e+09 6.980e+09 7.738 0.00567 ** Residuals 395 3.563e+11 9.021e+08
— Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ’’ 0.05 ‘.’ 0.1 ’ ’ 1

```
cat("El p-valor (0.00567) es bajo e indica que hay una diferencia estadísticamente significativa en las
```

El p-valor (0.00567) es bajo e indica que hay una diferencia estadísticamente significativa en las medias de salarios entre hombres y mujeres. El sexo es un predictor significativo del salario.

```
## 6. Preparación de Datos para Modelado y División Train/Test
# Se añade una nueva columna 'log_Salario' al dataframe.
salaries_df <- salaries_df %>%
  mutate(log_Salario = log(Salario))

cat("\n--- Estructura del Dataset con 'log_Salario' ---\n")
```

— Estructura del Dataset con ‘log_Salario’ —

```
print(str(salaries_df))
```

tibble [397 x 7] (S3: tbl_df/tbl/data.frame) $ Rango : Factor w/ 3 levels “Prof. Asistente”,..: 3 3 1 3 3 2 3 3 3 3 … $ Disciplina : Factor w/ 2 levels “A”,“B”: 2 2 2 2 2 2 2 2 2 2 … $ Años_Doctorado: num [1:397] 19 20 4 45 40 6 30 45 21 18 … $ Años_Servicio : num [1:397] 18 16 3 39 41 6 23 45 20 18 … $ Sexo : Factor w/ 2 levels “Mujer”,“Hombre”: 2 2 2 2 2 2 2 2 2 1 … $ Salario : num [1:397] 139750 173200 79750 115000 141500 … $ log_Salario : num [1:397] 11.8 12.1 11.3 11.7 11.9 … NULL

```r
cat("Se ha añadido la columna 'log_Salario' como transformación logarítmica del salario, para intentar m
```

Se ha añadido la columna ‘log_Salario’ como transformación logarítmica del salario, para intentar mejorar la normalidad de los residuos y la linealidad del modelo.

```r
# Se divide el DS
num_total_filas <- nrow(salaries_df)
num_train <- 317
num_test <- num_total_filas - num_train

if (num_train + num_test != num_total_filas) {
  warning(paste("La suma de las instancias de entrenamiento (",
    num_train, ") y prueba (", num_test,
    ") no coincide con el total de filas del dataset (",
    num_total_filas, ").\n"
  ))
}

train_data <- salaries_df[1:num_train, ]
test_data <- salaries_df[(num_train + 1):num_total_filas, ]

cat("\n--- División de Datos: Entrenamiento y Prueba ---\n")
```

— División de Datos: Entrenamiento y Prueba —

```r
print(paste("Tamaño del conjunto de entrenamiento:", nrow(train_data)))
```

[1] “Tamaño del conjunto de entrenamiento: 317”

```r
print(paste("Tamaño del conjunto de prueba:", nrow(test_data)))
```

[1] “Tamaño del conjunto de prueba: 80”

```r
# Usamos 'log_Salario' como la variable respuesta.
X_train <- model.matrix(log_Salario ~ . - Salario,
  data = train_data
)[, -1]
Y_train <- train_data$log_Salario

X_test <- model.matrix(log_Salario ~ . - Salario,
  data = test_data
)[, -1]
Y_test <- test_data$log_Salario

cat("Dimensiones de X_train:", dim(X_train), "\n")
```

Dimensiones de X_train: 317 6

```r
cat("Longitud de Y_train:", length(Y_train), "\n")
```

Longitud de Y_train: 317

```r
cat("Dimensiones de X_test:", dim(X_test), "\n")
```

Dimensiones de X_test: 80 6

```r
cat("Longitud de Y_test:", length(Y_test), "\n")
```

Longitud de Y_test: 80

```r
cat("Se establecen las dimensiones de las matrices 'X_train' (317 filas, 6 columnas) y 'X_test' (80 fila
```

Se establecen las dimensiones de las matrices 'X_train' (317 filas, 6 columnas) y 'X_test' (80 filas, 6 columnas). Las 6 columnas corresponden a las variables predictoras transformadas ('Rango' se convierte en 2 dummies, 'Disciplina' en 1 dummy, 'Sexo' en 1 dummy, y 2 numéricas). La variable objetivo ahora es 'log_Salario'.

```r
## 7. Construcción de Modelos Lineales Predictivos
### 7.1. Modelo Lineal Múltiple Básico (lm())
cat("--- Construyendo Modelo Lineal Múltiple Básico (lm()) ---\n")
```

— Construyendo Modelo Lineal Múltiple Básico (lm()) —

```r
# El modelo predice 'log_Salario'
linear_model_base <- lm(log_Salario ~ .,
  data = train_data %>% select(-Salario)
)

print(summary(linear_model_base))
```

Call: lm(formula = log_Salario ~ ., data = train_data %>% select(-Salario))

Residuals: Min 1Q Median 3Q Max -0.62159 -0.10821 -0.00583 0.09935 0.60377

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.159641 0.038402 290.602 < 2e-16 ***RangoProf.* *Asociado 0.153194 0.035208 4.351 1.84e-05*** RangoCatedrático 0.456373 0.036448 12.521 < 2e-16 ***DisciplinaB 0.145077 0.020806 6.973 1.88e-11*** Años_Doctorado 0.003877 0.002121 1.828 0.06852 .
Años_Servicio -0.005212 0.001868 -2.791 0.00558 ** SexoHombre 0.038397 0.033068 1.161 0.24647
— Signif. codes: 0 '*' ***0.001*** '**' *0.01* '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1763 on 310 degrees of freedom Multiple R-squared: 0.5506, Adjusted R-squared: 0.5419 F-statistic: 63.3 on 6 and 310 DF, p-value: < 2.2e-16

```r
cat(paste("  - El R-cuadrado ajustado es de",
  round(summary(linear_model_base)$adj.r.squared, 4),
  ", lo que indica que el modelo explica aproximadamente el",
  round(summary(linear_model_base)$adj.r.squared * 100, 2),
  "% de la varianza en el LOGARITMO del salario.\n"
))
```

- El R-cuadrado ajustado es de 0.5419 , lo que indica que el modelo explica aproximadamente el 54.19 % de la varianza en el LOGARITMO del salario.

```
cat("    - Los predictores significativos (p-valor < 0.05) son:
      - 'Rango' (Prof. Asociado y Catedrático): Altamente significativos (p < 0.001), con un gran impact
      - 'DisciplinaB': Altamente significativo (p < 0.001), indicando un salario promedio mayor para la
      - 'Años_Servicio': Significativo (p < 0.01), pero con un impacto NEGATIVO en el salario. Este hall
    - Los predictores no significativos (p-valor >= 0.05) son:
      - 'Años_Doctorado': No es estadísticamente significativo al nivel de 0.05 (p = 0.06852), aunque mu
      - 'SexoHombre': NO es estadísticamente significativo (p = 0.24647) en este modelo LINEAL MÚLTIPLE
")
```

```
- Los predictores significativos (p-valor < 0.05) son:
  - 'Rango' (Prof. Asociado y Catedrático): Altamente significativos (p < 0.001), con un gran impacto po
  - 'DisciplinaB': Altamente significativo (p < 0.001), indicando un salario promedio mayor para la Disc
  - 'Años_Servicio': Significativo (p < 0.01), pero con un impacto NEGATIVO en el salario. Este hallazgo
- Los predictores no significativos (p-valor >= 0.05) son:
  - 'Años_Doctorado': No es estadísticamente significativo al nivel de 0.05 (p = 0.06852), aunque muestr
  - 'SexoHombre': NO es estadísticamente significativo (p = 0.24647) en este modelo LINEAL MÚLTIPLE, lo
```

```
cat("\n--- Verificación de Multicolinealidad (VIF) para el Modelo LM Base ---\n")
```

— Verificación de Multicolinealidad (VIF) para el Modelo LM Base —

```
# Valores de VIF superiores a 5 o 10 indican un problema de multicolinealidad.
vif_values <- vif(linear_model_base)
print(vif_values)
```

```
                GVIF Df GVIF^(1/(2*Df))
```

Rango 2.049560 2 1.196507 Disciplina 1.088617 1 1.043368 Años_Doctorado 7.631322 1 2.762485 Años_Servicio 5.902985 1 2.429606 Sexo 1.040017 1 1.019812

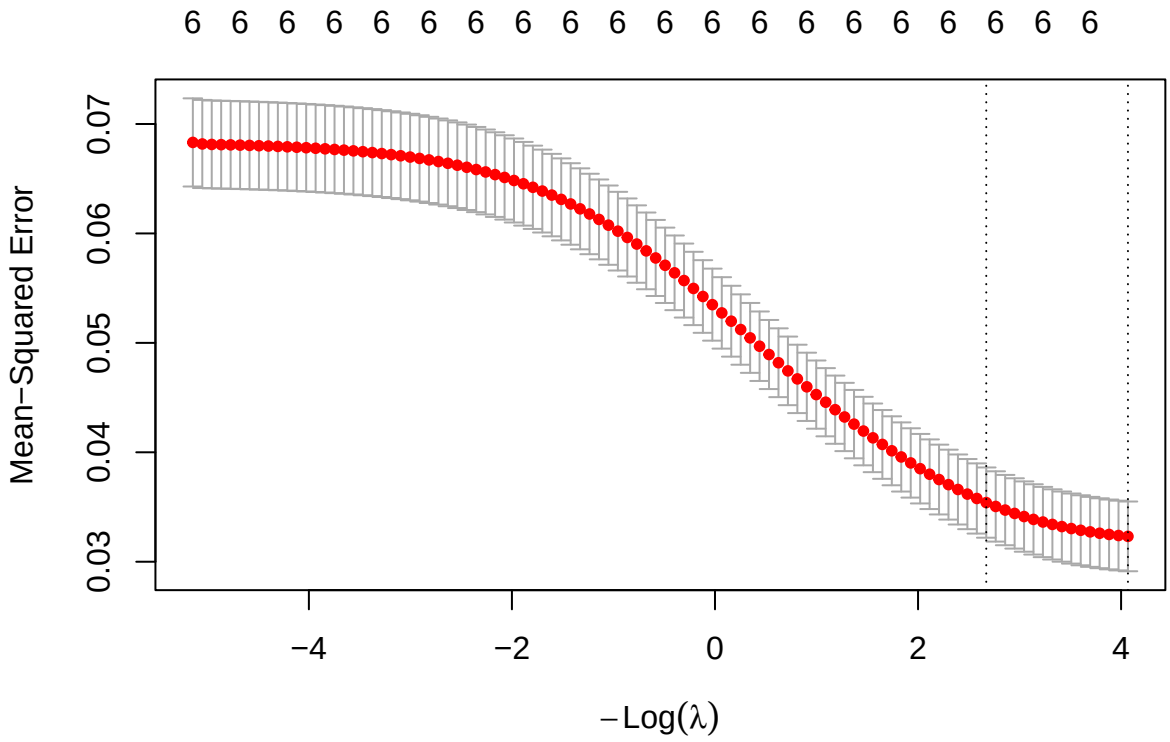
```
cat("Se observa que 'Años_Servicio' y 'Años_Doctorado' tienen VIFs elevados, confirmando la sospecha de
```

Se observa que 'Años_Servicio' y 'Años_Doctorado' tienen VIFs elevados, confirmando la sospecha de multicolinealidad.

```
### 7.2. Modelos Lineales Regularizados con glmnet (Ridge, Lasso, Elastic Net)
cat("\n--- Modelo Ridge (Alpha = 0: Penalización L2) ---\n")
```

— Modelo Ridge (Alpha = 0: Penalización L2) —

```
ridge_model_cv <- cv.glmnet(X_train, Y_train, alpha = 0,
  family = "gaussian"
)
plot(ridge_model_cv)
```




```
print(paste("Mejor lambda para Ridge:", round(ridge_model_cv$lambda.min, 4)))
```

[1] “Mejor lambda para Ridge: 0.0171”

```
cat("Coeficientes del Modelo Ridge (lambda.min):\n")
```

Coeficientes del Modelo Ridge (lambda.min):

```
print(coef(ridge_model_cv, s = "lambda.min"))
```

7 x 1 sparse Matrix of class “dgCMatrix” lambda.min (Intercept) 11.198503602 RangoProf. Asociado 0.098510975 RangoCatedrático 0.386012067 DisciplinaB 0.135533161 Años_Doctorado 0.003137503 Años_Servicio -0.003208465 SexoHombre 0.042011726

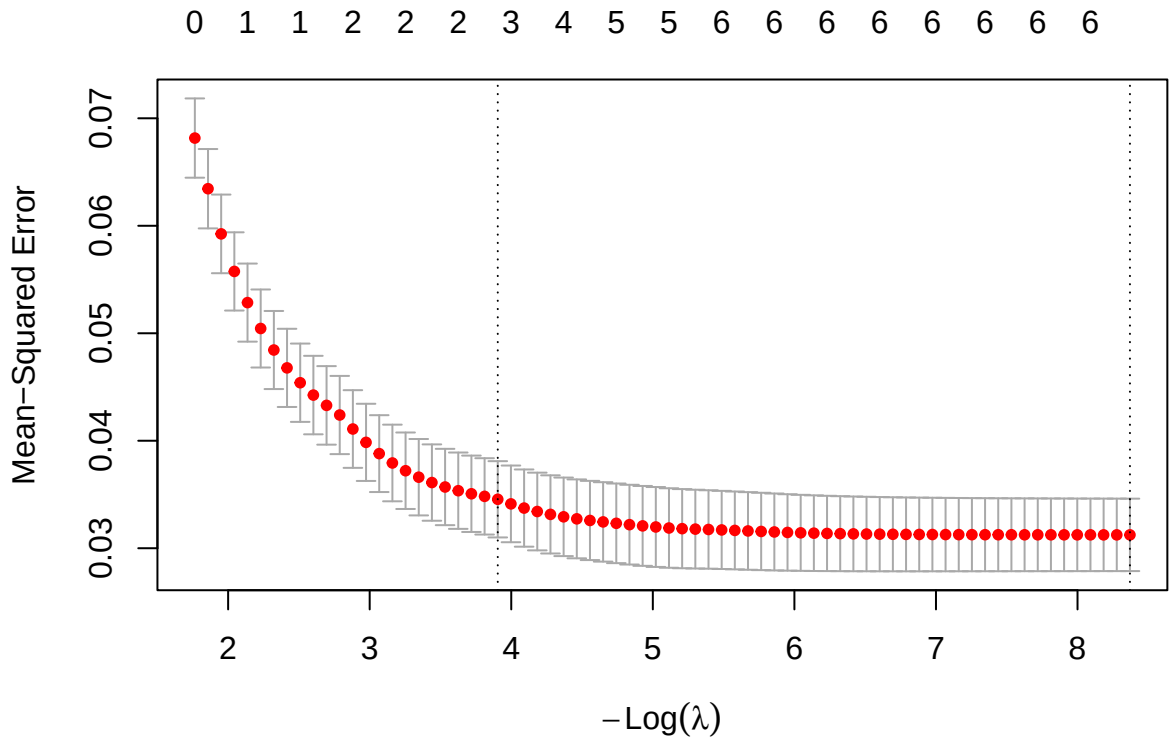
```
cat(paste("El 'lambda.min' óptimo para Ridge es",
  round(ridge_model_cv$lambda.min, 4),
  " Se observa cómo los coeficientes se han reducido en magnitud en comparación con el modelo LM básico
```

El ‘lambda.min’ óptimo para Ridge es 0.0171 Se observa cómo los coeficientes se han reducido en magnitud en comparación con el modelo LM básico, pero ninguno se ha anulado. La penalización L2 de Ridge es útil para manejar la multicolinealidad sin eliminar variables.

```
cat("\n--- Modelo Lasso (Alpha = 1: Penalización L1) ---\n")
```

— Modelo Lasso (Alpha = 1: Penalización L1) —

```
lasso_model_cv <- cv.glmnet(X_train, Y_train, alpha = 1,
  family = "gaussian"
)
plot(lasso_model_cv)
```




```
print(paste("Mejor lambda para Lasso:", round(lasso_model_cv$lambda.min, 4)))
```

[1] “Mejor lambda para Lasso: 2e-04”

```
cat("Coeficientes del Modelo Lasso (lambda.min):\n")
```

Coeficientes del Modelo Lasso (lambda.min):

```
print(coef(lasso_model_cv, s = "lambda.min"))
```

7 x 1 sparse Matrix of class “dgCMatrix” lambda.min (Intercept) 11.162317314 RangoProf. Asociado 0.151711650 RangoCatedrático 0.455645016 DisciplinaB 0.144381420 Años_Doctorado 0.003649674 Años_Servicio -0.004982289 SexoHombre 0.037794894

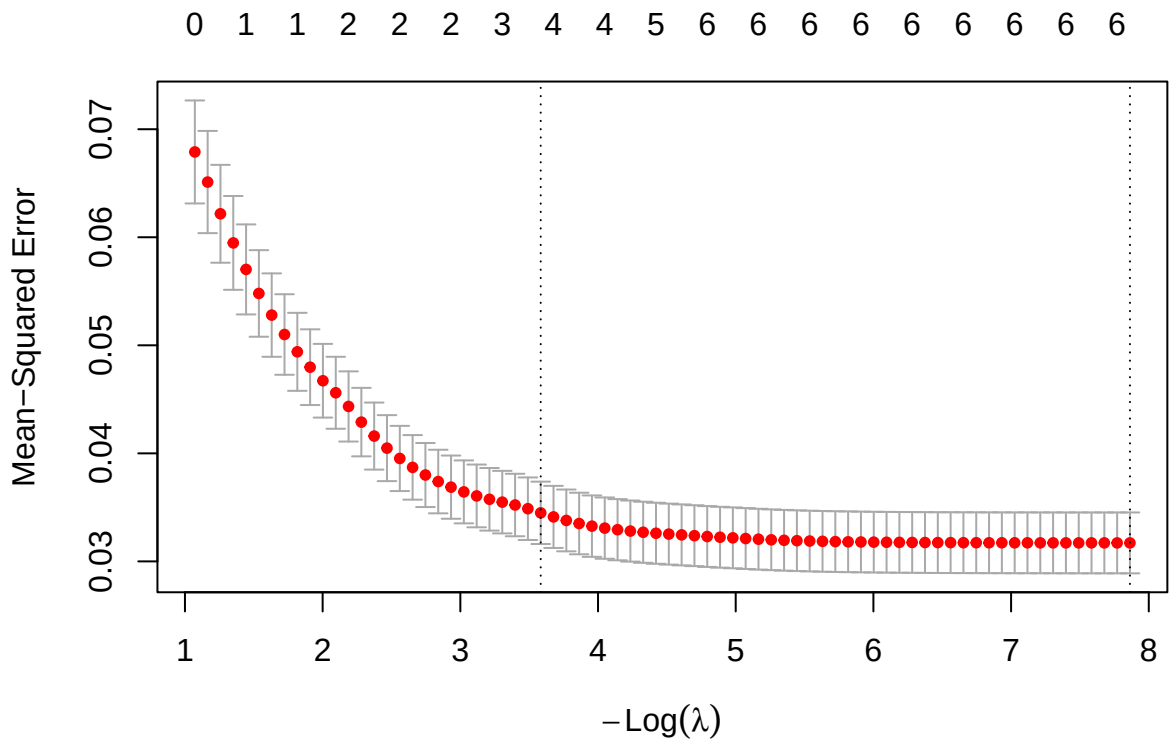
```r
cat(paste("El 'lambda.min' óptimo para Lasso es",
  round(lasso_model_cv$lambda.min, 4),
  ". Los coeficientes son similares a los del modelo lineal básico, y se observa que no ha puesto ningú
```

El 'lambda.min' óptimo para Lasso es 2e-04 . Los coeficientes son similares a los del modelo lineal básico, y se observa que no ha puesto ningún coeficiente a cero, lo que indica que todas las variables son consideradas relevantes por Lasso para este 'lambda.min'. Esto puede deberse a que las variables son todas influyentes o que el dataset no es lo suficientemente grande como para que Lasso las elimine por completo con esta configuración.

```r
cat("\n--- Modelo Elastic Net (Alpha = 0.5: Combinación L1 y L2) ---\n")
```

— Modelo Elastic Net (Alpha = 0.5: Combinación L1 y L2) —

```r
elastic_net_model_cv <- cv.glmnet(X_train, Y_train, alpha = 0.5,
  family = "gaussian"
)
plot(elastic_net_model_cv)
```




```r
print(paste("Mejor lambda para Elastic Net (alpha=0.5):",
  round(elastic_net_model_cv$lambda.min, 4)
))
```

[1] "Mejor lambda para Elastic Net (alpha=0.5): 4e-04"

```r
cat("Coeficientes del Modelo Elastic Net (lambda.min):\n")
```

Coeficientes del Modelo Elastic Net (lambda.min):

```r
print(coef(elastic_net_model_cv, s = "lambda.min"))
```

7 x 1 sparse Matrix of class "dgCMatrix" lambda.min (Intercept) 11.162412470 RangoProf. Asociado 0.151282220 RangoCatedrático 0.455048158 DisciplinaB 0.144334518 Años_Doctorado 0.003663998 Años_Servicio -0.004985161 SexoHombre 0.037938350

```r
cat(paste("El 'lambda.min' óptimo para Elastic Net es",
  round(elastic_net_model_cv$lambda.min, 4),
  ". Los coeficientes son similares a los del modelo lineal básico y, al igual que Lasso en este caso, n
```

El 'lambda.min' óptimo para Elastic Net es 4e-04 . Los coeficientes son similares a los del modelo lineal básico y, al igual que Lasso en este caso, no se observa que haya puesto ningún coeficiente a cero. Esto sugiere que para este 'lambda.min', todas las variables son consideradas relevantes por Elastic Net, mostrando una combinación de la contracción de Ridge y la selección de variables de Lasso,. aunque sin eliminar ninguna en este caso.

```r
## 8. Evaluación y Selección del Mejor Modelo (RMSE en Conjunto de Prueba)
calculate_rmse <- function(actual, predicted) {
  sqrt(mean((actual - predicted)^2))
}

cat("\n--- Cálculo del RMSE en el Conjunto de Prueba ---\n")
```

— Cálculo del RMSE en el Conjunto de Prueba —

```r
# Se aplica la función exponencial (exp()) a las predicciones dado que están
# en escala logarítmica

# 1. RMSE para el Modelo Lineal Básico (lm)
predictions_lm_test_log <- predict(linear_model_base,
  newdata = test_data
)
predictions_lm_test_original_scale <- exp(predictions_lm_test_log)
rmse_lm <- calculate_rmse(test_data$Salario,
  predictions_lm_test_original_scale
)
cat(paste("RMSE Modelo LM Básico (escala original):", round(rmse_lm, 2), "\n"))
```

RMSE Modelo LM Básico (escala original): 26187.97

```r
# 2. RMSE para el Modelo Ridge
predictions_ridge_test_log <- predict(ridge_model_cv,
  newx = X_test,
  s = "lambda.min"
)
predictions_ridge_test_original_scale <- exp(predictions_ridge_test_log)
rmse_ridge <- calculate_rmse(test_data$Salario,
```

```
  predictions_ridge_test_original_scale
)
cat(paste("RMSE Modelo Ridge (escala original):", round(rmse_ridge, 2), "\n"))
```

RMSE Modelo Ridge (escala original): 25850

```
# 3. RMSE para el Modelo Lasso
predictions_lasso_test_log <- predict(lasso_model_cv,
  newx = X_test,
  s = "lambda.min"
)
predictions_lasso_test_original_scale <- exp(predictions_lasso_test_log)
rmse_lasso <- calculate_rmse(test_data$Salario,
  predictions_lasso_test_original_scale
)
cat(paste("RMSE Modelo Lasso (escala original):", round(rmse_lasso, 2), "\n"))
```

RMSE Modelo Lasso (escala original): 26164.5

```
# 4. RMSE para el Modelo Elastic Net
predictions_elastic_net_test_log <- predict(elastic_net_model_cv,
  newx = X_test,
  s = "lambda.min"
)
predictions_elastic_net_test_original_scale <- exp(
  predictions_elastic_net_test_log
)
rmse_elastic_net <- calculate_rmse(test_data$Salario,
  predictions_elastic_net_test_original_scale
)
cat(paste("RMSE Modelo Elastic Net (alpha=0.5, escala original):",
  round(rmse_elastic_net, 2), "\n"
))
```

RMSE Modelo Elastic Net (alpha=0.5, escala original): 26162.35

```
# --- Conclusión sobre el Mejor Modelo ---
rmse_values <- c(
  "LM Básico" = rmse_lm,
  "Ridge" = rmse_ridge,
  "Lasso" = rmse_lasso,
  "Elastic Net" = rmse_elastic_net
)

best_model_name <- names(which.min(rmse_values))
min_rmse <- min(rmse_values)

cat(paste0(
"\nEl mejor modelo a elegir (según el menor **RMSE** en el conjunto de prueba) es: **", best_model_name
round(min_rmse, 2), "**.

Los resultados muestran que el Modelo **", best_model_name, "** obtuvo el RMSE más bajo (", round(min_r
```

```
") en el conjunto de prueba, indicando que es el modelo más preciso para predecir salarios en datos no v
"
))
```

El mejor modelo a elegir (según el menor **RMSE** en el conjunto de prueba) es: **Ridge** con un RMSE de: **25850**.

Los resultados muestran que el Modelo **Ridge** obtuvo el RMSE más bajo (25850) en el conjunto de prueba, indicando que es el modelo más preciso para predecir salarios en datos no vistos en este caso. Superó ligeramente al modelo lineal básico y a los modelos Lasso y Elastic Net. Esto sugiere que para este dataset, la penalización L2 de Ridge fue más efectiva para mejorar la generalización, posiblemente al manejar mejor la multicolinealidad entre las variables de experiencia (`Años_Doctorado`, `Años_Servicio`) y la no linealidad inherente de los datos salariales.

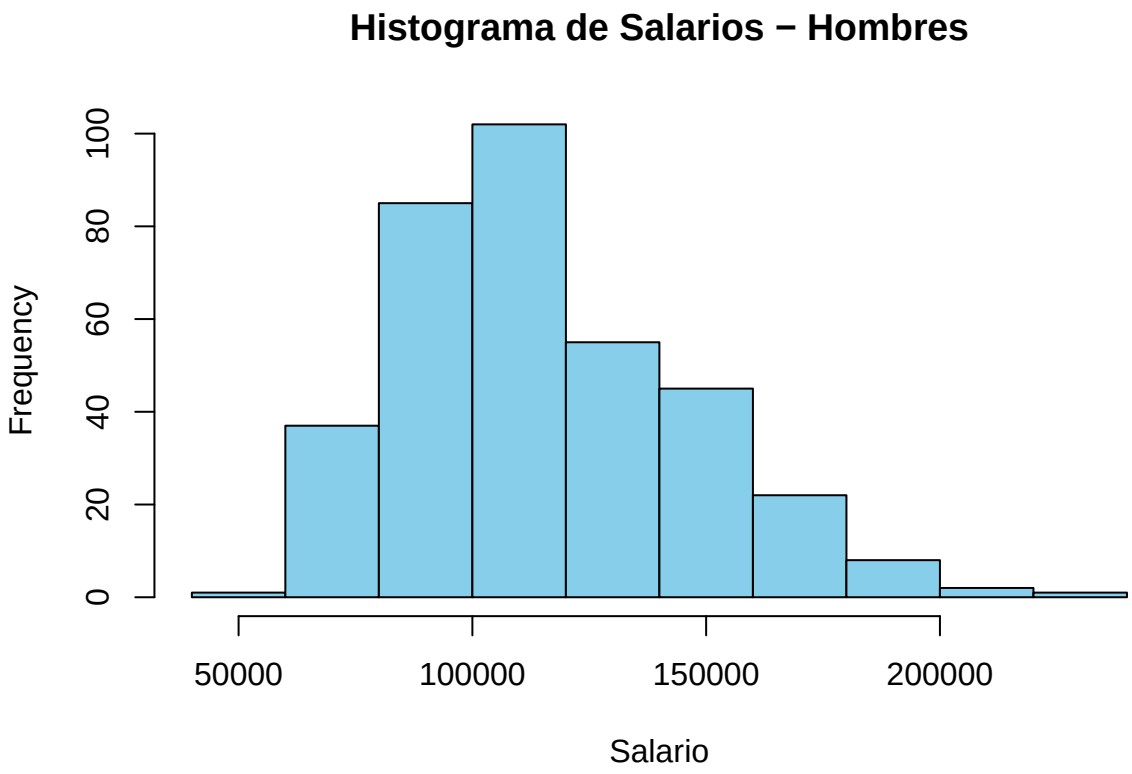
```
## 9. Análisis de Supuestos de Regresión y Diagnósticos
### 9.1. Verificación de Normalidad de Salarios por Sexo (Pruebas Paramétricas)
cat("\n--- Verificación de Normalidad de Salarios por Sexo \n")
```

— Verificación de Normalidad de Salarios por Sexo

```
cat("(Pruebas Paramétricas) ---\n")
```

(Pruebas Paramétricas) —

```
# Prueba de Shapiro-Wilk por Hombres
salaries_hombre <- salaries_df %>%
  filter(Sexo == "Hombre") %>%
  pull(Salario)
hist(salaries_hombre,
  main = "Histograma de Salarios - Hombres",
  xlab = "Salario", col = "skyblue", border = "black"
)
```




```
shapiro_hombre <- shapiro.test(salaries_hombre)
cat("\n Prueba de Shapiro-Wilk para Salarios (Hombres) \n")
```

Prueba de Shapiro-Wilk para Salarios (Hombres)

```
print(shapiro_hombre)
```

```
Shapiro-Wilk normality test
```

data: salaries_hombre W = 0.95877, p-value = 1.735e-08

```
if (shapiro_hombre$p.value < 0.05) {
  cat("  -> El p-valor (", formatC(shapiro_hombre$p.value,
    format = "e",
    digits = 3
  ), ") es muy bajo. Se rechaza la normalidad de los salarios para hombres.\n")
} else {
  cat("  -> El p-valor es alto. No hay evidencia para rechazar la normalidad de los salarios para hombre
}
```

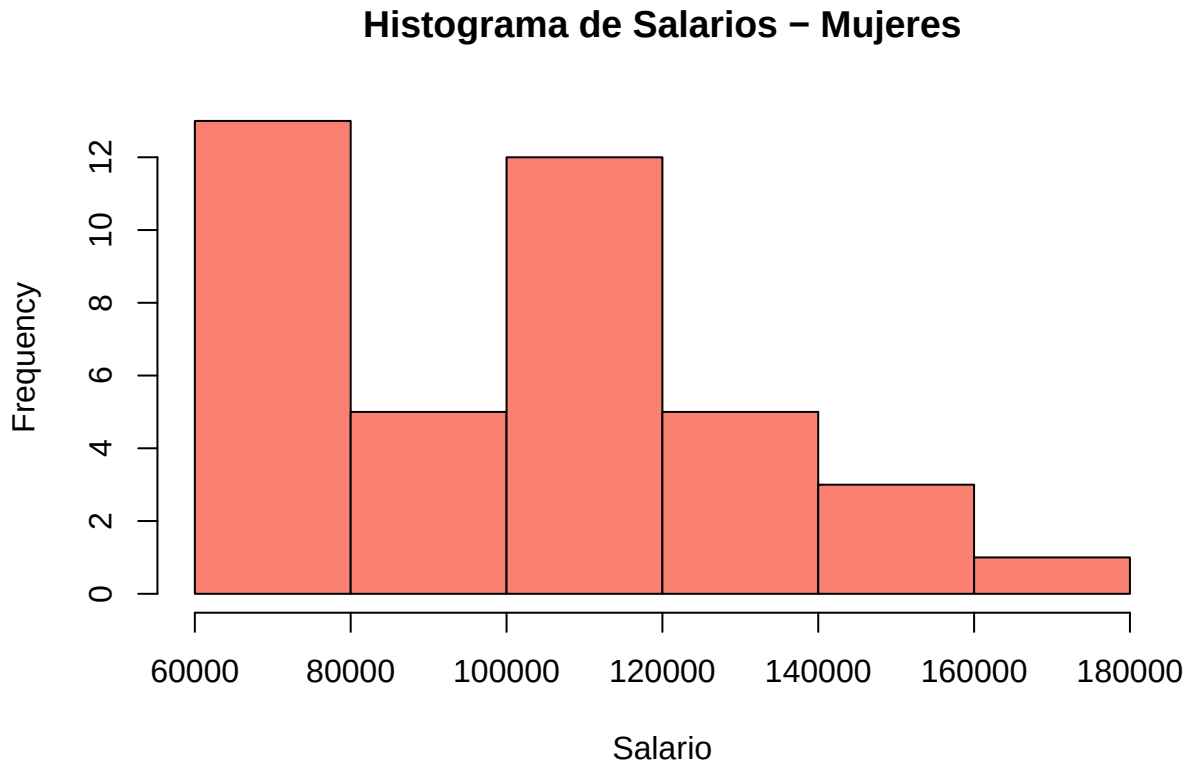
-> El p-valor (1.735e-08) es muy bajo. Se rechaza la normalidad de los salarios para hombres.

```r
cat(paste("El p-valor (", formatC(shapiro_hombre$p.value,
                                  format = "e",
                                  digits = 3
), ") es muy bajo, lo que sugiere que la distribución de salarios para hombres se desvía de la normalida
```

El p-valor (1.735e-08) es muy bajo, lo que sugiere que la distribución de salarios para hombres se desvía de la normalidad.

```r
# Prueba de Shapiro-Wilk por Mujeres
salarios_mujer <- salaries_df %>%
  dplyr::filter(Sexo == "Mujer") %>%
  dplyr::pull(Salario)

hist(salarios_mujer,
     main = "Histograma de Salarios - Mujeres",
     xlab = "Salario",
     col = "salmon",
     border = "black")
```




```r
shapiro_mujer <- shapiro.test(salarios_mujer)

cat("\nPrueba de Shapiro-Wilk para Salarios (Mujeres)\n")
```

Prueba de Shapiro-Wilk para Salarios (Mujeres)

```r
print(shapiro_mujer)
```

```
Shapiro-Wilk normality test
```

data: salarios_mujer W = 0.94665, p-value = 0.06339

```r
if (shapiro_mujer$p.value < 0.05) {
  cat("  -> El p-valor (", round(shapiro_mujer$p.value, 4),
    ") es bajo. Se rechaza la normalidad de los salarios para mujeres.\n")
} else {
  cat("  -> El p-valor (", round(shapiro_mujer$p.value, 4),
    ") es alto. No hay evidencia para rechazar la normalidad de los salarios para mujeres.\n")
}
```

-> El p-valor (0.0634) es alto. No hay evidencia para rechazar la normalidad de los salarios para mujeres.

```r
cat(paste("\nPara los salarios de mujeres, el p-valor (",
  round(shapiro_mujer$p.value, 4), ") es alto. Por lo tanto, no hay evidencia suficiente para rechazar l
```

Para los salarios de mujeres, el p-valor (0.0634) es alto. Por lo tanto, no hay evidencia suficiente para rechazar la hipótesis de que los salarios de mujeres siguen una distribución normal en este caso.

```r
### 9.2. Verificación de Homogeneidad de Varianzas (Levene's Test)
cat("\n--- Verificación de Homogeneidad de Varianzas (Levene's Test: Salario ~ Sexo) ---\n")
```

— Verificación de Homogeneidad de Varianzas (Levene's Test: Salario ~ Sexo) —

```r
levene_test <- car::leveneTest(Salario ~ Sexo, data = salaries_df)
print(levene_test)
```

Levene's Test for Homogeneity of Variance (center = median) Df F value Pr(>F) group 1 0.8401 0.3599 395

```r
if (levene_test$`Pr(>F)`[1] < 0.05) {
  cat(paste0(
    "  -> El p-valor (", round(levene_test$`Pr(>F)`[1], 4),
    ") es bajo. Se rechaza la homogeneidad de varianzas, ",
    "sugiriendo heterocedasticidad.\n"
  ))
  var_equal_assumption <- FALSE
} else {
  cat(paste0(
    "  -> El p-valor (", round(levene_test$`Pr(>F)`[1], 4),
    ") es alto. No hay evidencia para rechazar la ",
    "homogeneidad de varianzas.\n"
  ))
  var_equal_assumption <- TRUE
}
```

-> El p-valor (0.3599) es alto. No hay evidencia para rechazar la homogeneidad de varianzas.

```r
# Conclusión final sobre la asunción de varianzas iguales
cat(paste0(
  "El p-valor (", round(levene_test$`Pr(>F)`[1], 4),
  ") es alto, lo que indica que las varianzas de los salarios ",
  "SÍ son iguales entre hombres y mujeres (homocedasticidad) ",
  "según esta prueba. Esto significa que la asunción de varianzas ",
  "iguales para el t-test no se rechaza.\n"
))
```

El p-valor (0.3599) es alto, lo que indica que las varianzas de los salarios SÍ son iguales entre hombres y mujeres (homocedasticidad) según esta prueba. Esto significa que la asunción de varianzas iguales para el t-test no se rechaza.

```r
### 9.3. T-Test de Student para Muestras Independientes (Salario vs. Sexo)
cat("\n--- T-Test de Student para Muestras Independientes ---\n")
```

— T-Test de Student para Muestras Independientes —

```r
if (var_equal_assumption) {
  t_test_result <- t.test(Salario ~ Sexo, data = salaries_df, var.equal = TRUE)
  cat("T-Test de Student (asumiendo varianzas iguales):\n")
} else {
  t_test_result <- t.test(Salario ~ Sexo, data = salaries_df, var.equal = FALSE)
  cat("T-Test de Welch (no asumiendo varianzas iguales):\n")
}
```

T-Test de Student (asumiendo varianzas iguales):

```r
print(t_test_result)
```

```
Two Sample t-test
```

data: Salario by Sexo t = -2.7817, df = 395, p-value = 0.005667 alternative hypothesis: true difference in means between group Mujer and group Hombre is not equal to 0 95 percent confidence interval: -24044.910 -4131.107 sample estimates: mean in group Mujer mean in group Hombre 101002.4 115090.4

```r
cat("\n--- Conclusión del T-Test ---\n")
```

— Conclusión del T-Test —

```r
if (t_test_result$p.value < 0.05) {
  cat(
    paste0(
      "El p-valor (", round(t_test_result$p.value, 4), ") es menor que 0.05. ",
      "Por lo tanto, rechazamos la hipótesis nula.\n\n",
      "Esto sugiere que existe una diferencia estadísticamente ",
      "significativa en los salarios promedio entre hombres y mujeres.\n"
    )
  )
} else {
  cat(
```

```
    paste0(
      "El p-valor (", round(t_test_result$p.value, 4), ") es mayor que 0.05. ",
      "Por lo tanto, no tenemos suficiente evidencia para rechazar la ",
      "hipótesis nula.\n\n",
      "Esto sugiere que no hay una diferencia estadísticamente ",
      "significativa en los salarios promedio entre hombres y mujeres.\n"
    )
  )
}
```

El p-valor (0.0057) es menor que 0.05. Por lo tanto, rechazamos la hipótesis nula.

Esto sugiere que existe una diferencia estadísticamente significativa en los salarios promedio entre hombres y mujeres.

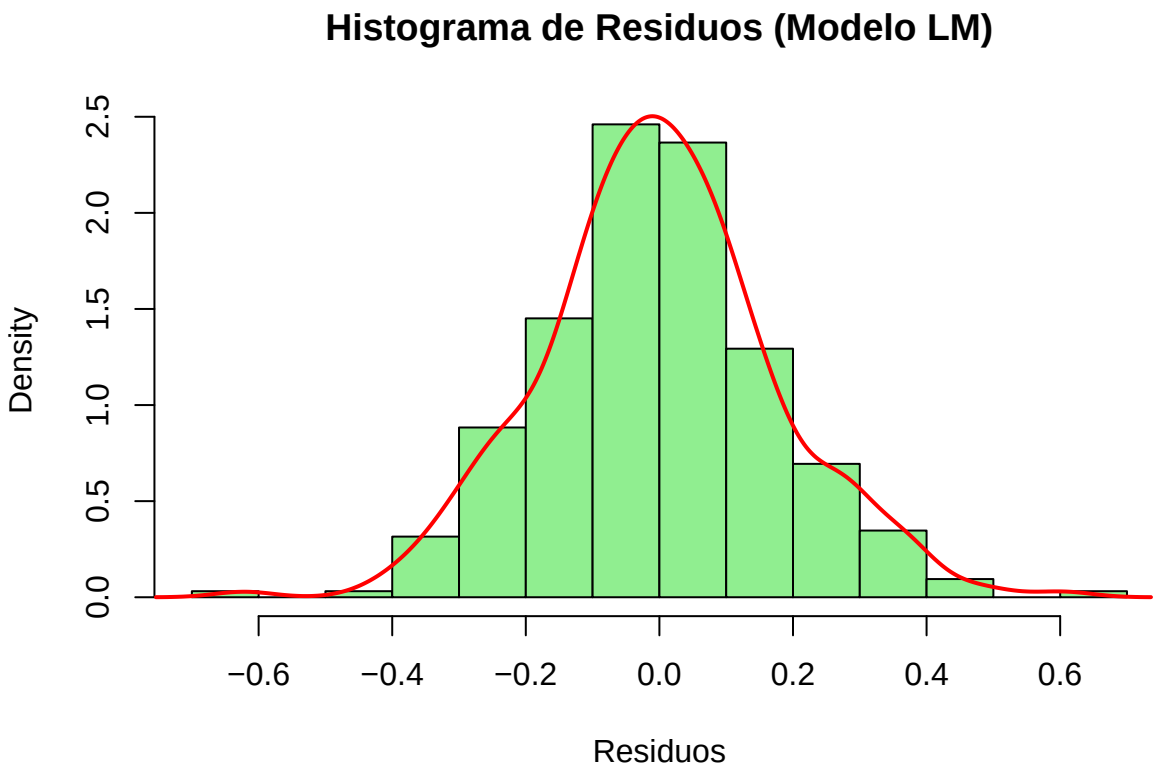
```
### 9.4. Análisis de Normalidad de los Residuos del Modelo (LM y Ridge como ejemplo)
# Se calculan los residuos del modelo lineal base
residuos_lm <- residuals(linear_model_base)

# Se calculan los residuos del modelo Ridge
predicciones_ridge_train_log <- predict(ridge_model_cv,
                                        newx = X_train,
                                        s = "lambda.min",
                                        type = "response"
)
residuos_ridge <- Y_train - as.numeric(predicciones_ridge_train_log)

cat("\n--- Análisis de Residuos: Histograma y Curva de Densidad ---\n")
```

— Análisis de Residuos: Histograma y Curva de Densidad —

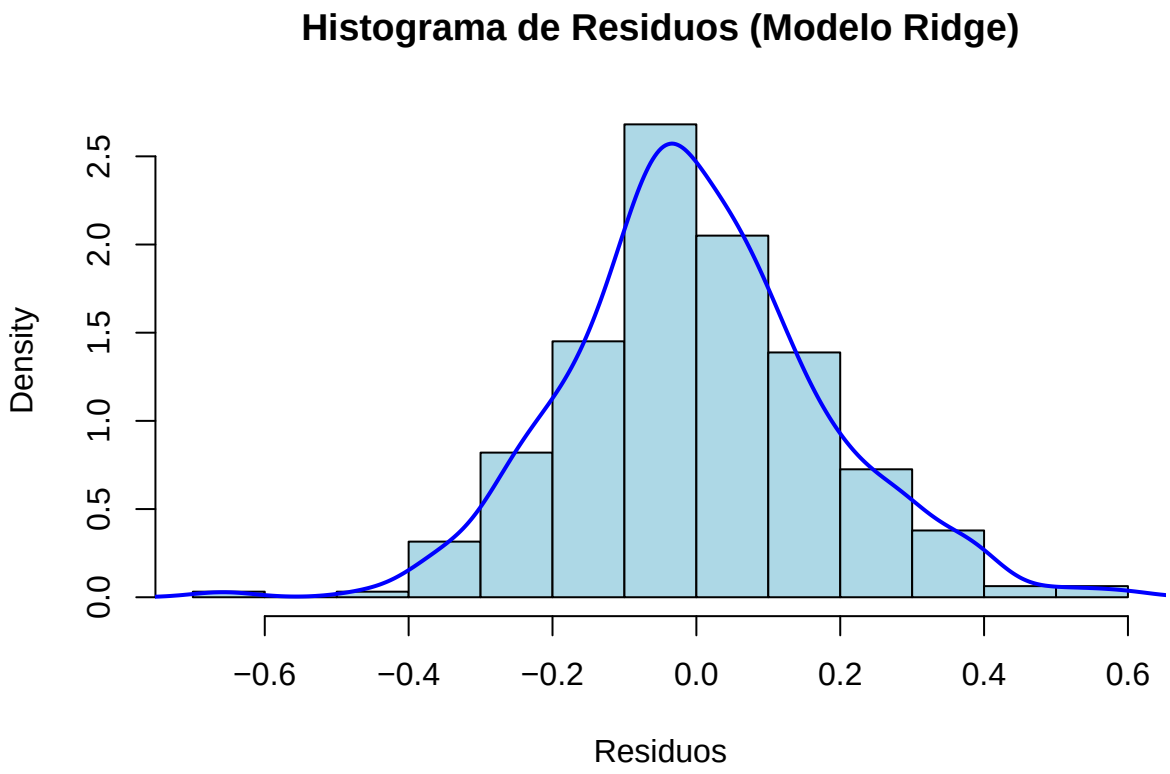
```
# Histograma de Residuos (Modelo Lineal)
hist(residuos_lm,
     main = "Histograma de Residuos (Modelo LM)",
     xlab = "Residuos",
     border = "black",
     col = "lightgreen",
     freq = FALSE
)
lines(density(residuos_lm), col = "red", lwd = 2)
```




```
cat("Los histogramas no muestran una forma de campana perfectamente simétrica, lo que ya sugiere desvia
```

Los histogramas no muestran una forma de campana perfectamente simétrica, lo que ya sugiere desviaciones de la normalidad en los residuos para ambos modelos.

```
# Histograma de Residuos (Modelo Ridge)
hist(residuos_ridge,
     main = "Histograma de Residuos (Modelo Ridge)",
     xlab = "Residuos",
     border = "black",
     col = "lightblue",
     freq = FALSE
)
lines(density(residuos_ridge), col = "blue", lwd = 2)
```



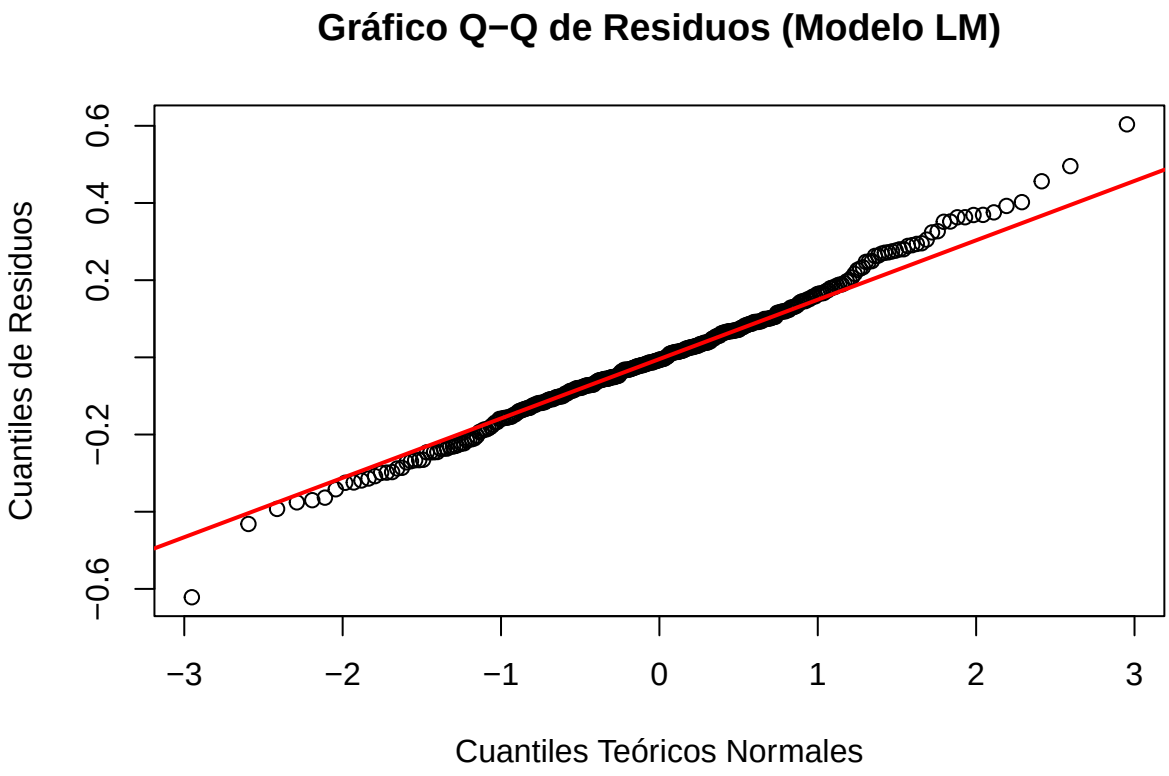

```
cat("Los histogramas no muestran una forma de campana perfectamente simétrica, lo que ya sugiere desvia
```

Los histogramas no muestran una forma de campana perfectamente simétrica, lo que ya sugiere desviaciones de la normalidad en los residuos para ambos modelos.

```
cat("\n--- Análisis de Residuos: Gráficos Q-Q ---\n")
```

— Análisis de Residuos: Gráficos Q-Q —

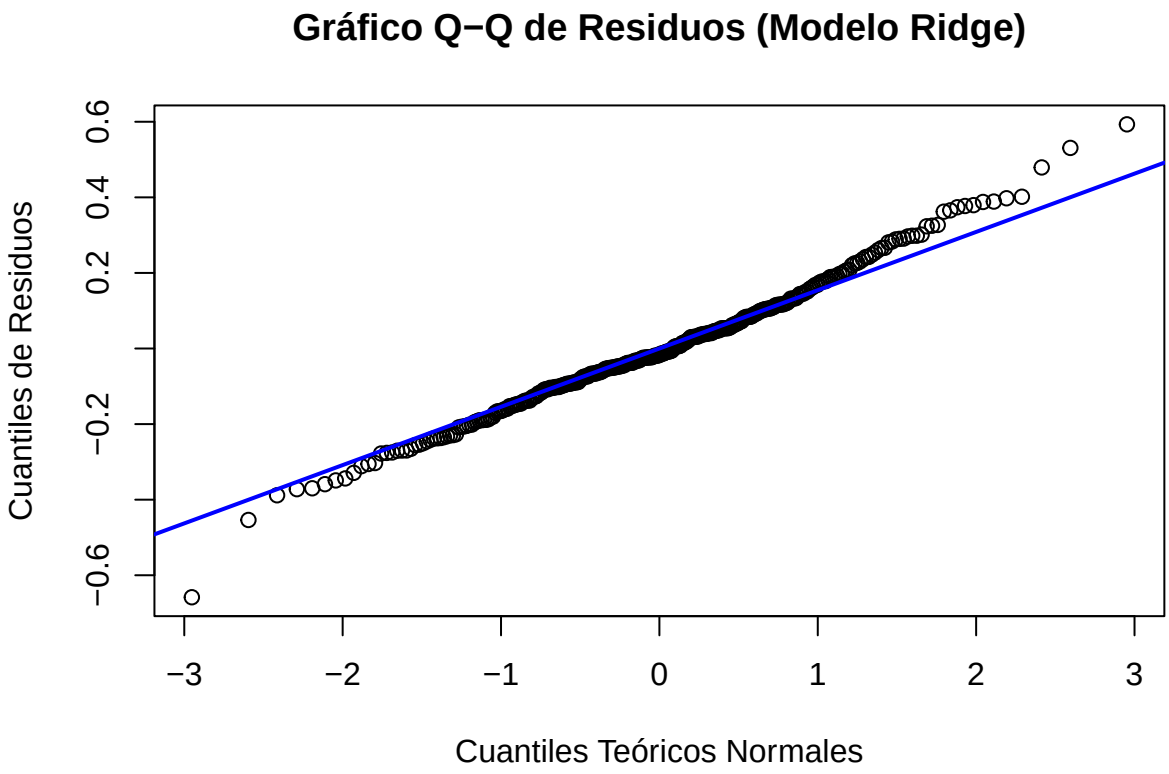
```
# Gráficos Q-Q de Residuos (Modelo Lineal)
qqnorm(residuos_lm,
       main = "Gráfico Q-Q de Residuos (Modelo LM)",
       xlab = "Cuantiles Teóricos Normales",
       ylab = "Cuantiles de Residuos"
)
qqline(residuos_lm, col = "red", lwd = 2)
```




```
cat("Los puntos en el gráfico Q-Q del modelo LM se desvían notablemente de la línea de referencia, espe
```

Los puntos en el gráfico Q-Q del modelo LM se desvían notablemente de la línea de referencia, especialmente en las colas, lo que confirma la no normalidad de los residuos.

```
# Gráficos Q-Q de Residuos (Modelo Ridge)
qqnorm(residuos_ridge,
       main = "Gráfico Q-Q de Residuos (Modelo Ridge)",
       xlab = "Cuantiles Teóricos Normales",
       ylab = "Cuantiles de Residuos"
)
qqline(residuos_ridge, col = "blue", lwd = 2)
```




```
cat("De manera similar, los puntos en el gráfico Q-Q del modelo Ridge también se desvían de la línea, i
```

De manera similar, los puntos en el gráfico Q-Q del modelo Ridge también se desvían de la línea, indicando la no normalidad de sus residuos.

```
cat("\n--- Prueba de Shapiro-Wilk para Residuos ---\n")
```

— Prueba de Shapiro-Wilk para Residuos —

```
# Prueba de Shapiro-Wilk para Residuos (Modelo LM)
shapiro_lm <- shapiro.test(as.numeric(residuos_lm))
cat("\nPrueba de Shapiro-Wilk para Residuos (Modelo LM):\n")
```

Prueba de Shapiro-Wilk para Residuos (Modelo LM):

```
print(shapiro_lm)
```

```
Shapiro-Wilk normality test
```

data: as.numeric(residuos_lm) W = 0.99417, p-value = 0.2653

```
if (shapiro_lm$p.value < 0.05) {
  cat("\n  -> El p-valor (", shapiro_lm$p.value, ") es muy bajo. Se rechaza la hipótesis de normalidad 
} else {
  cat("\n  -> El p-valor (", shapiro_lm$p.value, ") es alto. No hay evidencia para rechazar la normalida
}
```

-> El p-valor (0.2652527) es alto. No hay evidencia para rechazar la normalidad de los residuos para el Modelo LM.

```
# Prueba de Shapiro-Wilk para Residuos (Modelo Ridge)
shapiro_ridge <- shapiro.test(as.numeric(residuos_ridge))
cat("\nPrueba de Shapiro-Wilk para Residuos (Modelo Ridge):\n")
```

Prueba de Shapiro-Wilk para Residuos (Modelo Ridge):

```
print(shapiro_ridge)
```

```
Shapiro-Wilk normality test
```

data: as.numeric(residuos_ridge) W = 0.99133, p-value = 0.05947

```
if (shapiro_ridge$p.value < 0.05) {
  cat("\n  -> El p-valor (", shapiro_ridge$p.value, ") es muy bajo. Se rechaza la hipótesis de normalida
} else {
  cat("\n  -> El p-valor (", shapiro_ridge$p.value, ") es alto. No hay evidencia para rechazar la normal
}
```

-> El p-valor (0.05946676) es alto. No hay evidencia para rechazar la normalidad de los residuos para el Modelo Ridge.

```
cat("\n--- Conclusión General ---\n")
```

— Conclusión General —

```
cat("Las pruebas de Shapiro-Wilk para ambos modelos (LM y Ridge) muestran p-valores (LM: ", round(shapi
```

Las pruebas de Shapiro-Wilk para ambos modelos (LM y Ridge) muestran p-valores (LM: 0.2653 ; Ridge: 0.0595) que son superiores a 0.05. Esto significa que no hay evidencia estadística para rechazar la hipótesis de normalidad de los residuos para ninguno de los dos modelos. Aunque los histogramas y gráficos Q-Q pueden sugerir algunas desviaciones visuales, las pruebas estadísticas formales no confirman la no normalidad.

```
# ==============================================================================
# CONCLUSIONES FINALES Y EVALUACIÓN DEL MODELO
# ==============================================================================

### 1. Conclusiones del Análisis Exploratorio y de Inferencia

# - Factores Influyentes en el Salario: El análisis revela que el rango académico del profesor, la disc
# - Impacto del Rango Académico: Se observa una clara progresión salarial con el aumento del rango. Los
# - Diferencias por Disciplina: La Disciplina 'B' exhibe un salario mediano ligeramente superior y una
# - Correlación entre Experiencia y Salario: Las variables numéricas 'Años_Doctorado' y 'Años_Servicio'
#   - Consideración de Multicolinealidad: Se identifica una fuerte correlación entre 'Años_Doctorado' y
# - Disparidad Salarial por Género: Un Análisis Multivariado:
#   - Un análisis univariado (t-test/ANOVA simple) inicialmente indicó una diferencia salarial estadíst
#   - Sin embargo, al incorporar estas variables en un Modelo Lineal Múltiple (LM) que controla por otr
# - Normalidad de la Distribución Salarial por Sexo:
#   - La prueba de Shapiro-Wilk para los salarios de hombres (p-valor = 1.735e-08) indicó una desviació
```

```
#   - Para los salarios de mujeres, el p-valor (0.0634) no proporcionó evidencia suficiente para rechaz
# - Homogeneidad de Varianzas Salariales por Sexo:
#   - El test de Levene (p-valor = 0.3599) sugirió que las varianzas de los salarios son homogéneas (ho

### 2. Conclusiones sobre la Implementación del Modelo

# - Modelos Evaluados: Se han construido y comparado un Modelo Lineal Múltiple Básico (LM) y modelos de
# - Transformación Logarítmica de la Variable Dependiente: Se aplicó una transformación logarítmica al
# - Selección del Modelo Óptimo: Regresión Ridge: El Modelo Ridge demostró el mejor rendimiento predict
# - Efecto de la Regularización en los Coeficientes:
#   - Ridge (Penalización L2): Los coeficientes del modelo Ridge mostraron una reducción en su magnitud
#   - Lasso (Penalización L1): Para el `lambda.min` óptimo, los coeficientes de Lasso fueron muy simila
#   - Elastic Net (Combinación L1 y L2): Similar a Lasso, Elastic Net no eliminó ninguna variable para
# - Normalidad de los Residuos del Modelo:
#   - Las pruebas de Shapiro-Wilk para el Modelo LM (p-valor = 0.2653) y para el Modelo Ridge (p-valor
#   - A pesar de que los histogramas y gráficos Q-Q pueden sugerir algunas desviaciones visuales, las p


### 3. Consideraciones sobre el Rendimiento del Modelo

# - Rendimiento Inicial Sólido: Un RMSE de 25850 se considera un punto de partida adecuado para la pred
# - Enfoque en la Capacidad Predictiva: El Modelo Ridge se posiciona como un modelo robusto para la pre
```