Actividad de Modelos Lineales Generalizados (GLM)

Daniel Arteta Salazar

2025-11-08

Contents

Resumen

1. Datos y variables

1.1. Importar librerías

```
library(readxl)
library(tidyverse)
library(ggplot2)
```

1.2. Preparación de los datos

```
# Importar base de datos
base <- read_excel("../datos/base_seguro_15000.xlsx")

# Adecuar el nombre de las variables
base <- base %>% janitor::clean_names()

# Tamaño de la tabla
dim(base)
```

```
## [1] 15000     7
```

```
# Resumen de las variables
summary(base)
```

```
##       edad           sexo               tipo              region
##  Min.   :18.00   Length:15000       Length:15000       Length:15000
##  1st Qu.:33.00   Class :character   Class :character   Class :character
##  Median :49.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :49.07
##  3rd Qu.:65.00
##  Max.   :80.00
##     reclamos       costo_esperado       prima_pura
##  Min.   :0.00000   Min.   :   26.18   Min.   :    0.0
##  1st Qu.:0.00000   1st Qu.: 1929.99   1st Qu.:    0.0
##  Median :0.00000   Median : 3660.01   Median :    0.0
##  Mean   :0.03433   Mean   : 4940.43   Mean   :  203.6
##  3rd Qu.:0.00000   3rd Qu.: 6375.56   3rd Qu.:    0.0
##  Max.   :2.00000   Max.   :46043.90   Max.   :38720.5
```

```
# Registros duplicados
sum(duplicated(base))
```

```
## [1] 0
```

```
# Número de valores nulos por variable
colSums(is.na(base))
```

```
##           edad           sexo           tipo         region       reclamos
##              0              0              0              0              0
## costo_esperado     prima_pura
##              0              0
```

```
# Agregamos una variable categórica por edad
base$grupo_etario <- cut(base$edad,
                     breaks = c(0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100),
                     labels = c("1-10", "11-20", "21-30", "31-40", "41-50",
                                "51-60", "61-70", "71-80", "81-90", "91-100"),
```

```
                        right = TRUE)
```

1.3. Resumen de la sección

No se encuentran valores anormales en las variables, registros duplicados ni valores nulos, por lo que no se realiza ninguna modificación a los datos.

2. Análisis descriptivo

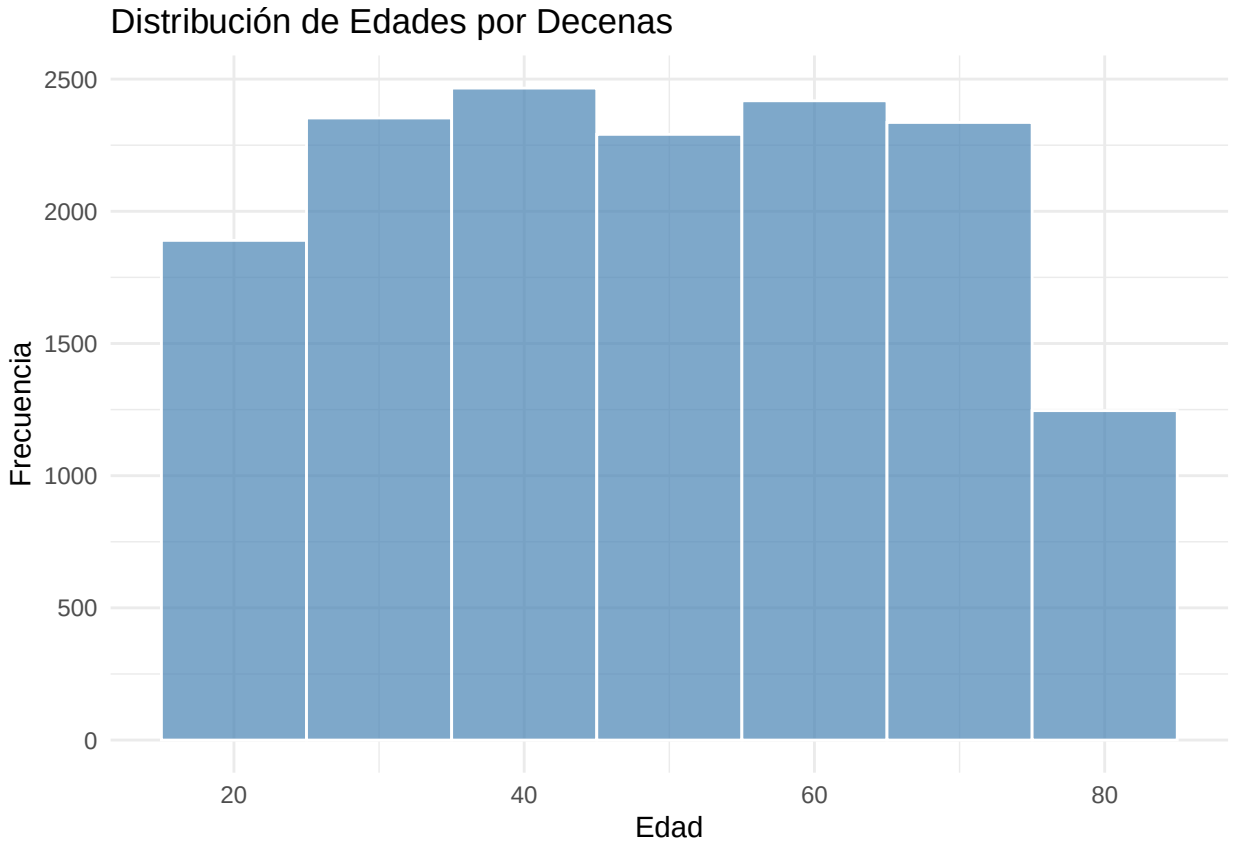
2.1. Análisis univariado

2.1.1. Edad

```
summary(base$edad)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   33.00   49.00   49.07   65.00   80.00
```

```
ggplot(base, aes(x = edad)) +
  geom_histogram(binwidth = 10, fill = "steelblue", color = "white", alpha = 0.7) +
  labs(title = "Distribución de Edades por Decenas",
       x = "Edad",
       y = "Frecuencia") +
  theme_minimal()
```

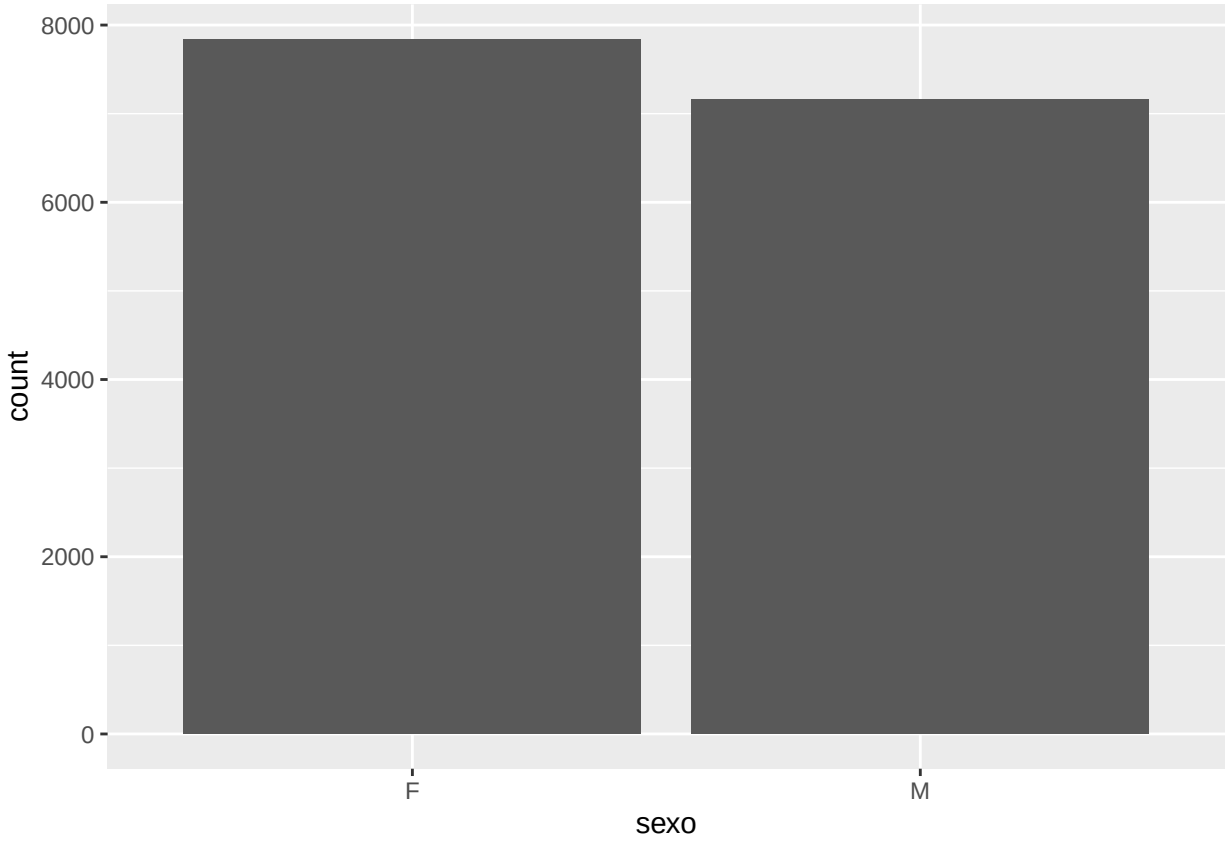


2.1.1. Sexo

```
# Tabla de frecuencia
base %>%
  count(sexo) %>%
  mutate(
    proporcion = n / sum(n),
    porcentaje = proporcion * 100
  ) %>%
  rename(frecuencia = n)
```

```
## # A tibble: 2 x 4
##   sexo  frecuencia proporcion porcentaje
##   <chr>      <int>      <dbl>      <dbl>
## 1 F           7841      0.523       52.3
## 2 M           7159      0.477       47.7
```

```
# Gráfico de barras
ggplot(base, aes(x = sexo)) +
  geom_bar()
```



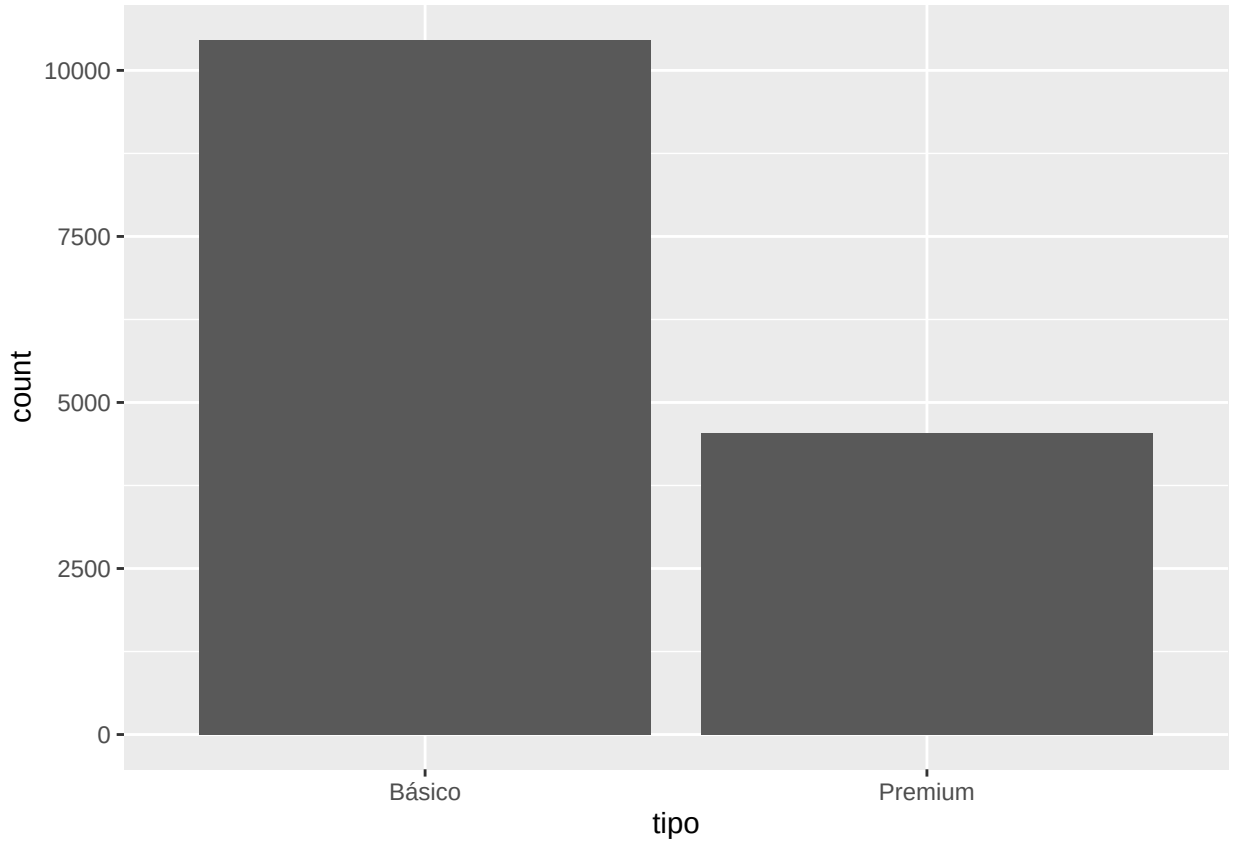
2.1.2. Tipo

```
# Tabla de frecuencia
base %>%
  count(tipo) %>%
  mutate(
    proporcion = n / sum(n),
```

```
    porcentaje = proporcion * 100
  ) %>%
  rename(frecuencia = n)
```

```
## # A tibble: 2 x 4
##   tipo    frecuencia proporcion porcentaje
##   <chr>        <int>      <dbl>      <dbl>
## 1 Básico       10462      0.697       69.7
## 2 Premium       4538      0.303       30.3
```

```
# Gráfico de barras
ggplot(base, aes(x = tipo)) +
  geom_bar()
```



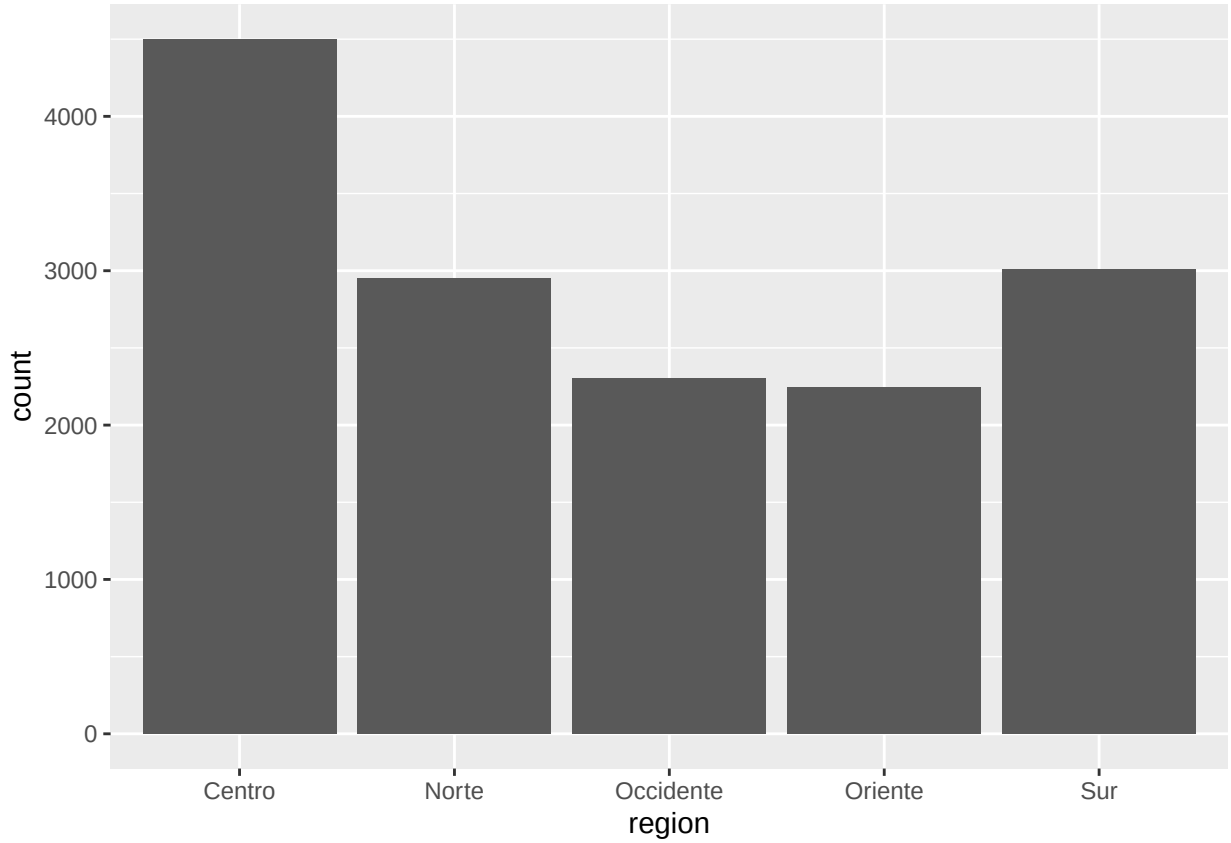
2.1.3. Región

```
# Tabla de frecuencia
base %>%
  count(region) %>%
  mutate(
    proporcion = n / sum(n),
    porcentaje = proporcion * 100
  ) %>%
  rename(frecuencia = n)
```

```
## # A tibble: 5 x 4
##   region    frecuencia proporcion porcentaje
##   <chr>          <int>      <dbl>      <dbl>
```

```
## 1 Centro          4501      0.300       30.0
## 2 Norte           2948      0.197       19.7
## 3 Occidente       2302      0.153       15.3
## 4 Oriente         2242      0.149       14.9
## 5 Sur             3007      0.200       20.0
```

```r
# Gráfico de barras
ggplot(base, aes(x = region)) +
  geom_bar()
```

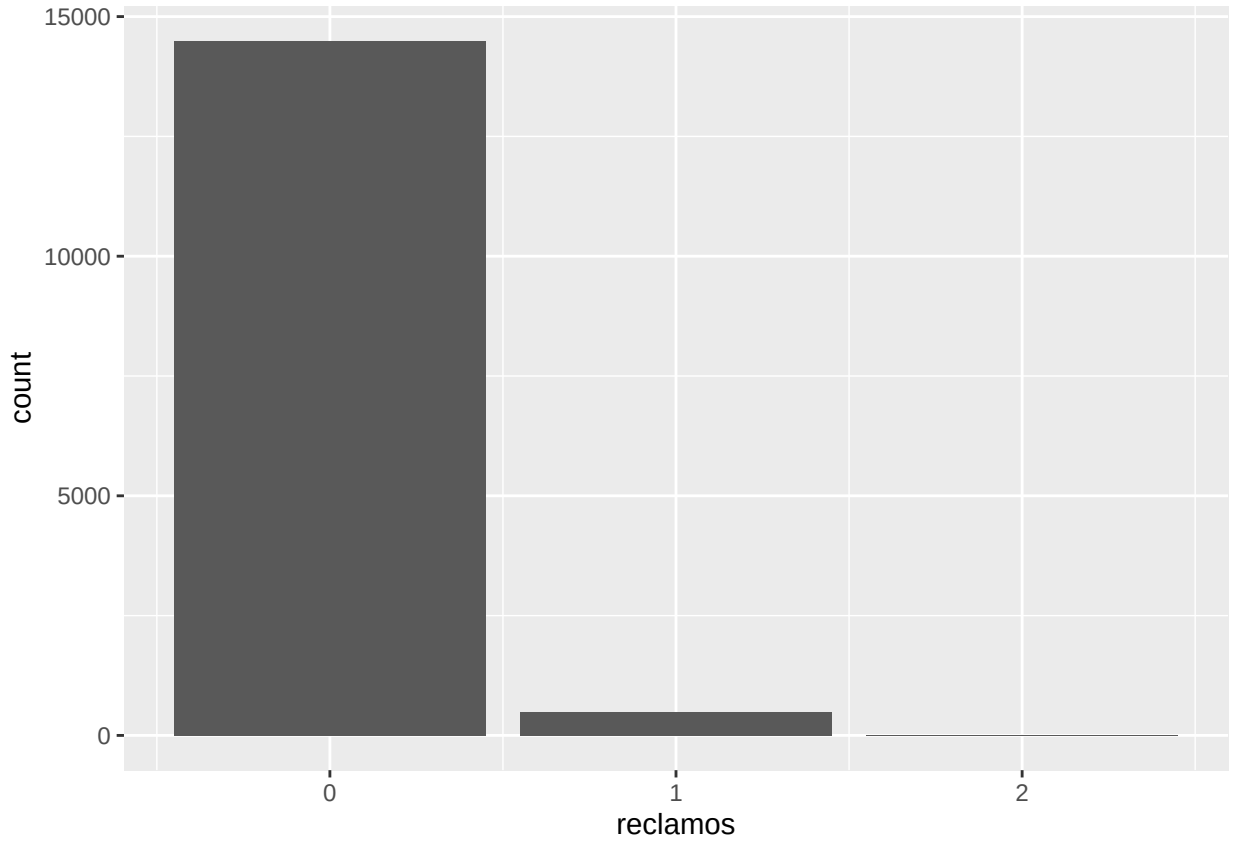


2.1.3. Reclamo

```r
# Tabla de frecuencia
base %>%
  count(reclamos) %>%
  mutate(
    proporcion = n / sum(n),
    porcentaje = proporcion * 100
  ) %>%
  rename(frecuencia = n)
```

```
## # A tibble: 3 x 4
##   reclamos frecuencia proporcion porcentaje
##      <dbl>      <int>      <dbl>      <dbl>
## 1        0      14499   0.967       96.7
## 2        1        487   0.0325       3.25
## 3        2         14   0.000933     0.0933
```

```r
# Gráfico de barras
ggplot(base, aes(x = reclamos)) +
  geom_bar()
```

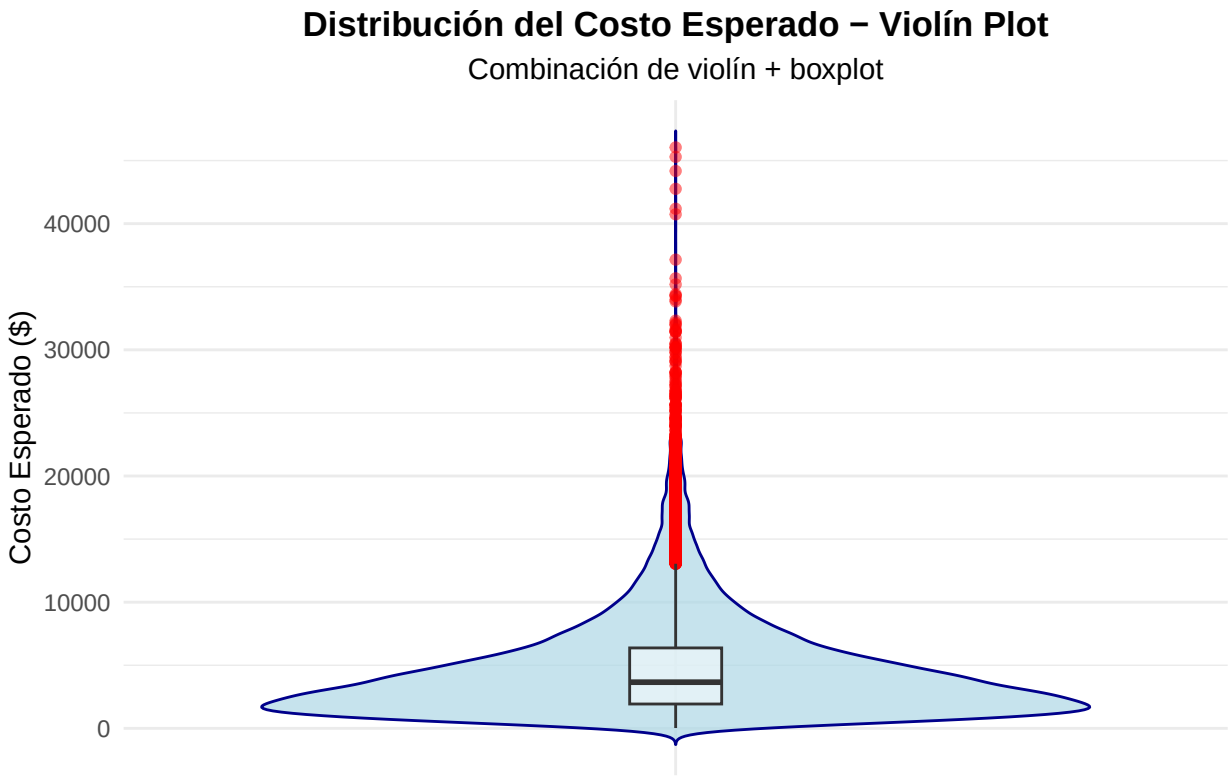


2.1.4. Costo esperado

```r
summary(base$costo_esperado)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    26.18  1929.99  3660.01  4940.43  6375.56 46043.90
```

```r
ggplot(base, aes(x = "", y = costo_esperado)) +
  geom_violin(fill = "lightblue", color = "darkblue", alpha = 0.7, trim = FALSE) +
  geom_boxplot(width = 0.1, fill = "white", alpha = 0.5, outlier.color = "red") +
  labs(title = "Distribución del Costo Esperado - Violín Plot",
       subtitle = "Combinación de violín + boxplot",
       y = "Costo Esperado ($)",
       x = "") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"),
        plot.subtitle = element_text(hjust = 0.5))
```



2.2. Análisis bivariado

2.2.1. Costo esperado

2.2.2. Costo esperado

2.2.3. Costo esperado

2.2.4. Costo esperado

3. Correlaciones y relaciones clave

4. Modelado GLM (frecuencia y severidad)

5. Prima pura por segmentos

6. Recomendaciones tarifarias

7. Limitaciones del análisis