

Hola

Tratamiento de datos

ENCUESTA PERMANENTE DE HOGARES



Encuesta Permanente de Hogares

Diseño de registros y estructura para las bases preliminares Hogar y Personas

Buenos Aires, agosto de 2025

Antes de empezar, una aclaración:

Antes de empezar, una aclaración:

Todas las conclusiones que se puedan llegar a desprender de este análisis deben interpretarse con cautela. Se trata únicamente de un ejercicio pedagógico cuyo propósito es ilustrar un posible flujo de trabajo con datos reales.

Ahora sí, empecemos

Tratamiento de datos

LIBRERÍAS

```
# para hacer todo lo que aprendieron la clase pasada  
library(tidyverse)
```

Tratamiento de datos LIBRERÍAS

```
# para hacer todo lo que aprendieron la clase pasada  
library(tidyverse)  
  
# para hacer todo lo que aprendieron la clase pasada,  
# aún con objetos de R nativo  
library(broom)
```

Tratamiento de datos

LECTURA DE DATOS

```
1 df <- read_delim(  
2   file = "usu_individual_T125.txt",  
3   delim = ";",  
4 )
```

CODUSU	ANO4	TRIMESTRE	NRO_HOGAR	COMPONENTE	H15	REGION
TQRMNOVQVHJOLOCDEFKID00875778	2025	1	1	2	1	43
TQRMNOVQVHJOLOCDEFKID00875778	2025	1	1	3	0	43
TQRMNOVQVHJOLOCDEFKID00875778	2025	1	1	4	0	43
TQRMNOVQVHJOLOCDEFKID00875778	2025	1	1	5	0	43
TQRMNOUPQHLOLOCDEFKID00851757	2025	1	1	1	1	43
TQRMNOUPQHLOLOCDEFKID00851757	2025	1	1	2	1	43

```
1 nrow(df)  
[1] 45.425
```

Tratamiento de datos

VARIABLES DE INTERÉS

```
1 df <- df %>%
2   rename(
3     edad = CH06 ,
4     salario = P21
5   #
6 )
```

Características de los miembros del hogar		
Campo	Tipo (longitud)	Cuestionario Hogar
		Descripción
CH03	N (2)	Relación de parentesco 1 = Jefe/a 2 = Cónyuge/pareja 3 = Hijo/a/hijastro/a 4 = Yerno/nuera 5 = Nieto/a 6 = Madre/padre 7 = Suegro/a 8 = Hermano/a 9 = Otros familiares 10 = No familiares
CH04	N (1)	Sexo 1 = Varón 2 = Mujer
CH05	date	Fecha de nacimiento (día, mes y año)
CH06	N (2)	¿Cuántos años cumplidos tiene?

Tratamiento de datos

F I L T R O S

Nos quedamos con la gente que tiene algún trabajo (formal o informal)

```
1 df_filt <- df %>%
 2   filter(
 3     ESTADO == 1,
 4     salario > 0,
 5   )
```

Podemos ver cuánto filtramos comparando

```
1 nrow(df)
[1] 45425

1 nrow(df_filt)
[1] 15802
```

Tenemos los datos, ahora queremos graficarlos

Graficos

GGPLOT

```
1 ggplot(df_filt,                      # dataframe con los datos  
2   aes(x = edad, y = salario))      # qué vars del df quiero mapear
```

Graficos

GGPLOT

```
1 ggplot(df_filt,                      # dataframe con los datos
2   aes(x = edad, y = salario)) + # qué vars del df quiero mapear
3   geom_point(alpha = 0.2)       # representa la vars con puntos
```

Graficos

GGPLOT

```
1 ggplot(df_filt,                      # dataframe con los datos
2   aes(x = edad, y = salario)) + # qué vars del df quiero mapear
3   geom_point(alpha = 0.2) +      # representa la vars con puntos
4   labs(                          # modificá los labels
5     x = "Edad",
6     y = "Salario",
7   )
```

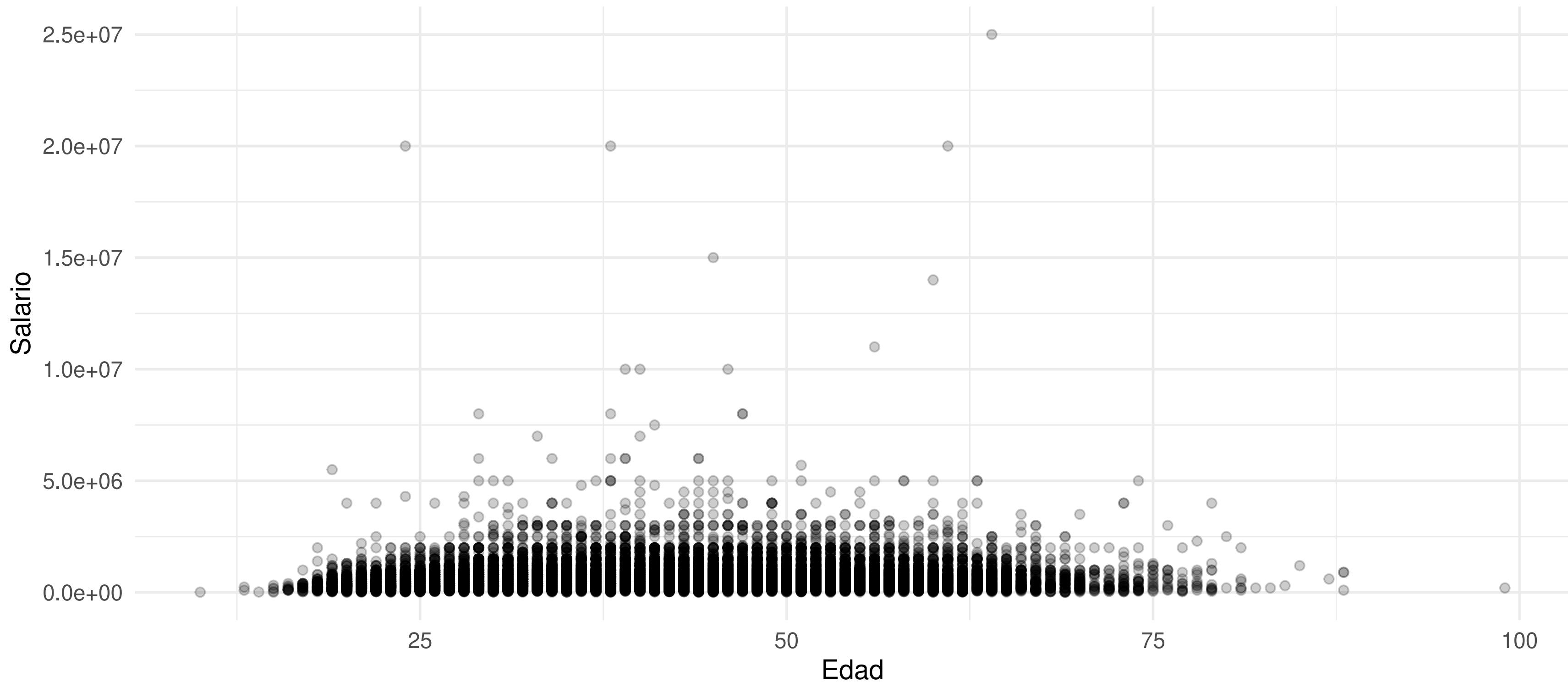
Graficos con GGPlot

INTRODUCCIÓN

```
1 ggplot(df_filt,                      # dataframe con los datos
2   aes(x = edad, y = salario)) + # qué vars del df quiero mapear
3   geom_point(alpha = 0.2) +      # representa la vars con puntos
4   labs(                          # modificá los labels
5     x = "Edad",
6     y = "Salario",
7   ) +
8   theme_minimal()               # que sea más limpio
```

Graficos con GGPlot

EDAD VS SALARIO



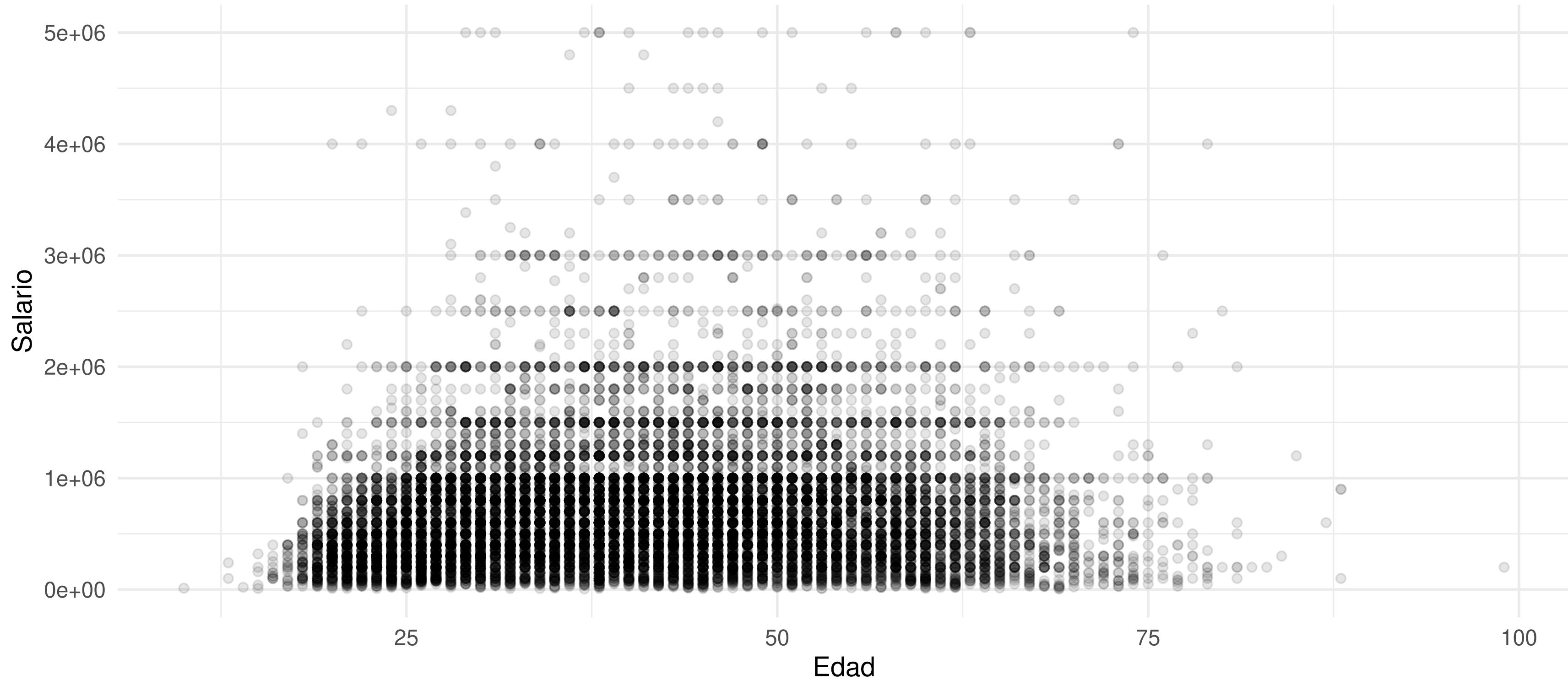
Graficos con GGPlot

EXPLORACIÓN GRÁFICA

```
1 ggplot(df_filt,                      # dataframe con los datos
2   aes(x = edad, y = salario)) + # qué vars del df quiero mapear
3   geom_point(alpha = 0.2) +      # representa la vars con puntos
4   labs(                          # modificá los labels
5     x = "Edad",
6     y = "Salario",
7   ) +
8   ylim(0, 5000000) +            # solo salarios menores a 5M
9   theme_minimal()              # que sea más limpio
```

Graficos con GGPlot

EXPLORACIÓN GRÁFICA



En busca de la correlación

EDAD VS SALARIO PROMEDIO

```
1 edad_salario <- df_filt %>%
```

En busca de la correlación

EDAD VS SALARIO PROMEDIO

```
1 edad_salario <- df_filt %>%
2   # Agrupo todas las filas por edades
3   group_by(edad) %>%
```

En busca de la correlación

EDAD VS SALARIO PROMEDIO

```
1 edad_salario <- df_filt %>%
2   # Agrupo todas las filas por edades
3   group_by(edad) %>%
4
5   # Calculo el salario promedio de cada grupo
6   summarise(salario_prom = mean(salario)) %>%
```

En busca de la correlación

EDAD VS SALARIO PROMEDIO

```
1 edad_salario <- df_filt %>%
2   # Agrupo todas las filas por edades
3   group_by(edad) %>%
4
5   # Calculo el salario promedio de cada grupo
6   summarise(salario_prom = mean(salario)) %>%
7
8   # Desarmo el grupo
9   ungroup() %>%
```

En busca de la correlación

EDAD VS SALARIO PROMEDIO

```
1 edad_salario <- df_filt %>%
2   # Agrupo todas las filas por edades
3   group_by(edad) %>%
4
5   # Calculo el salario promedio de cada grupo
6   summarise(salario_prom = mean(salario)) %>%
7
8   # Desarmo el grupo
9   ungroup() %>%
10
11  # Ordeno el df por edad de forma creciente
12  arrange(edad)
```

En busca de la correlación

EDAD VS SALARIO PROMEDIO

```
1 # Calculo el salario promedio por edad
2 edad_salario <- df_filt %>%
3   group_by(edad) %>%
4   summarise(salario_prom = mean(salario)) %>%
5   ungroup() %>%
6   arrange(edad)
```

En busca de la correlación

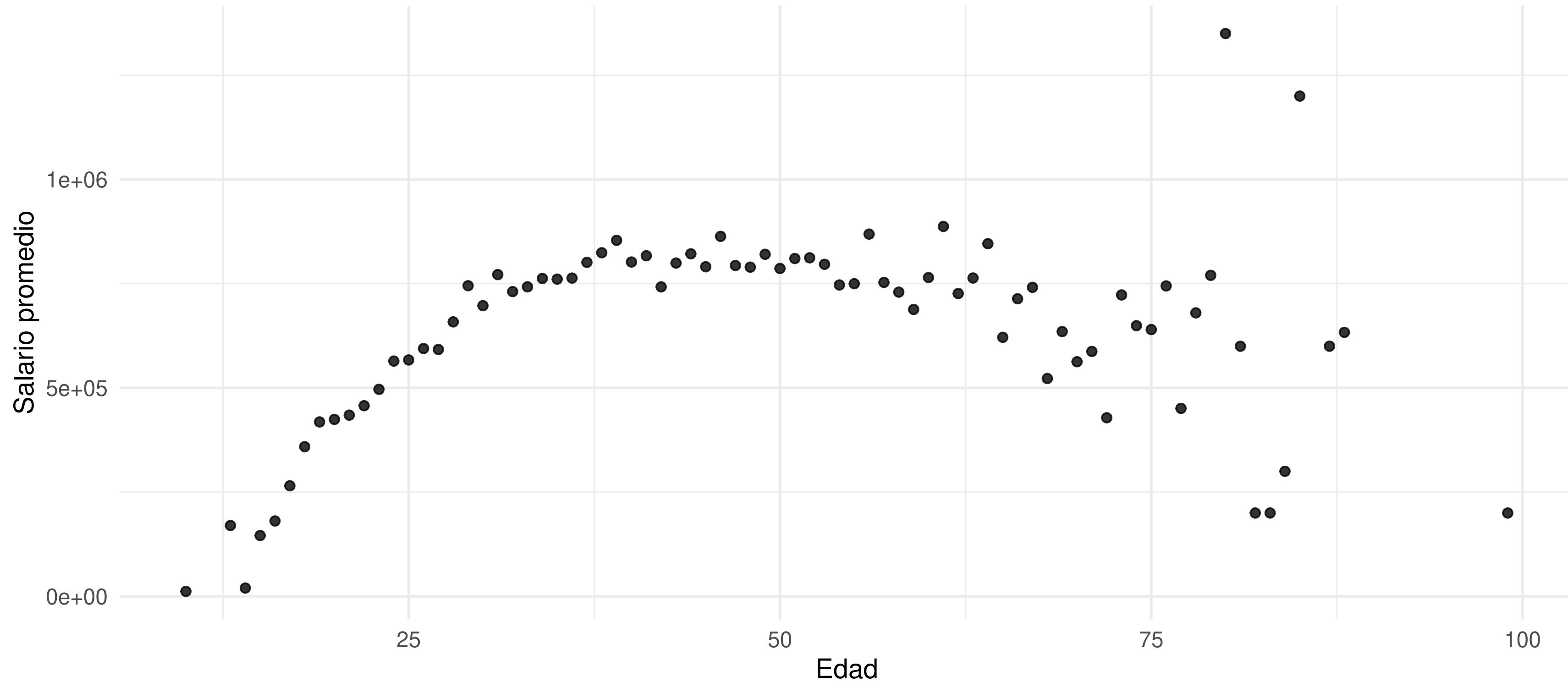
EDAD VS SALARIO PROMEDIO

Graficamos igual que antes:

```
1 ggplot(edad_salario, aes(x = edad, y = salario_prom)) +  
2   geom_point(alpha = 0.8) +  
3   labs(  
4     x = "Edad",  
5     y = "Salario promedio",  
6   ) +  
7   theme_minimal()
```

En busca de la correlación

EDAD VS SALARIO PROMEDIO



En busca de la correlación

EDAD VS SALARIO PROMEDIO

Filtrando por edad nos queda...

```
1 # Me quedo las personas entre 18 y 40 años que sean asalariadas
2 df_filt_18_40 <- df %>%
3   filter(
4     salario > 0,
5     ESTADO == 1,
6     between(edad, 18, 40)
7   )
8 # Calculo el promedio de su salario por edad
9 edad_salario_18_40 <- df_filt_18_40 %>%
10   group_by(edad) %>%
11   summarise(salario_prom = mean(salario)) %>%
12   ungroup() %>%
13   arrange(edad)
```

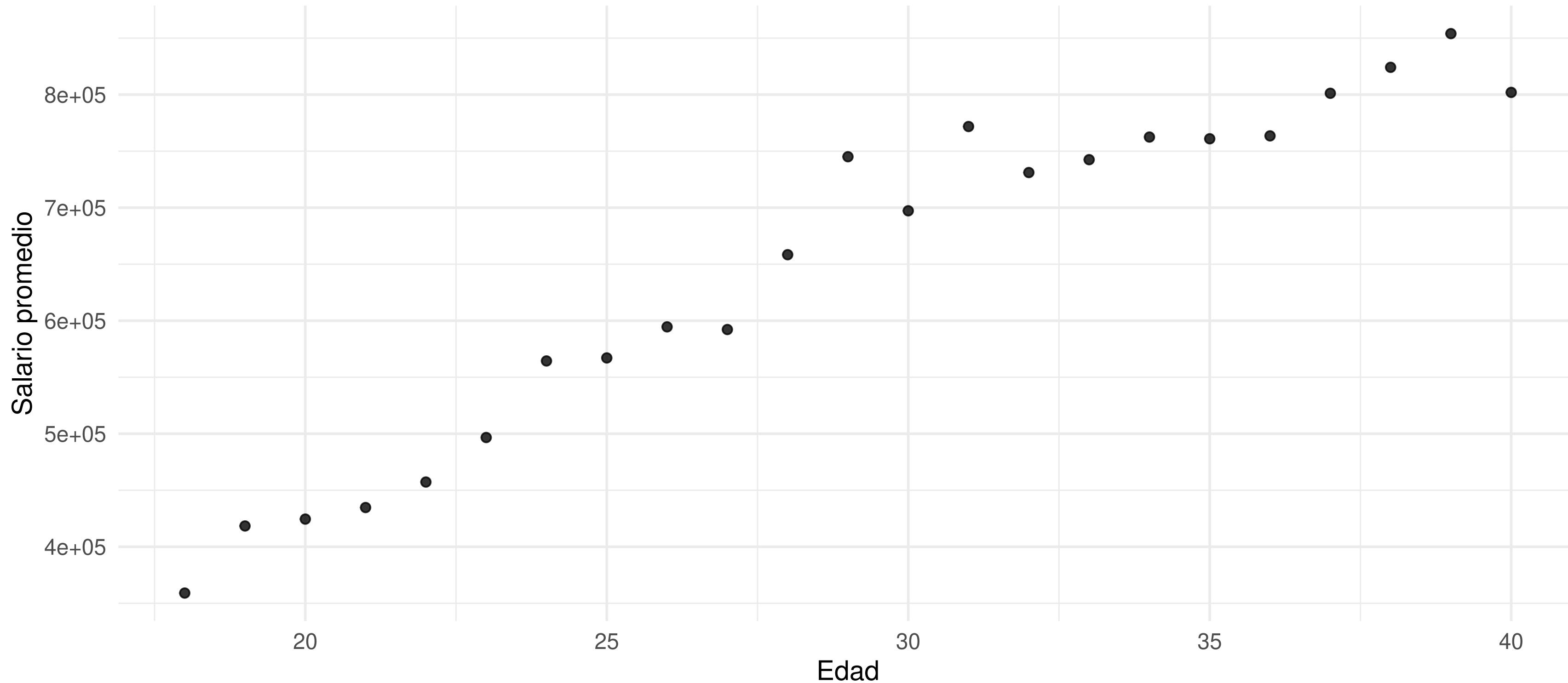
En busca de la correlación

EDAD VS SALARIO PROMEDIO

```
1 # Grafico el salario promedio vs la edad
2 ggplot(edad_salario_18_40, aes(x = edad, y = salario_prom)) +
  geom_point(alpha = 0.8) +
  labs(
    x = "Edad",
    y = "Salario promedio",
  ) +
  theme_minimal()
```

En busca de la correlación

EDAD VS SALARIO PROMEDIO



Modelo lineal simple

INTRODUCCIÓN AL MODELO LINEAL SIMPLE

Viendo los datos resulta razonable proponer un modelo lineal:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

donde β_i son los parámetros de nuestro modelo y ϵ el error.

O bien Y es una var. aleatoria tal que su esperanza condicionada a X está dada por

$$\mathbb{E}(Y|X) = \beta_0 + \beta_1 X$$

En cualquier caso, dados pares de datos (x_i, y_i) , los mejores estimadores para β_0 y β_1 que podemos construir son

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Modelo lineal simple

MODELO LINEAL A MANO

En nuestro código esta cuenta

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}$$

se escribe

```
1 x_i <- edad_salario_18_40$edad           # [18, 19, ..., 40]
2 y_i <- edad_salario_18_40$salario_prom # [400K, 300K, ..., 800K]
3 # ...
```

Modelo lineal simple

MODELO LINEAL A MANO

En nuestro código esta cuenta

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}$$

se escribe

```
1 x_i <- edad_salario_18_40$edad # [18, 19, ..., 40]
2 y_i <- edad_salario_18_40$salario_prom # [400K, 300K, ..., 800K]
3 x_prom <- mean(x_i)
4 y_prom <- mean(y_i)
5 # ...
```

Modelo lineal simple

MODELO LINEAL A MANO

En nuestro código esta cuenta

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

se escribe

```
1 x_i <- edad_salario_18_40$edad # [18, 19, ..., 40]
2 y_i <- edad_salario_18_40$salario_prom # [400K, 300K, ..., 800K]
3 x_prom <- mean(x_i)
4 y_prom <- mean(y_i)
5 b1 <- sum((x_i - x_prom) * (y_i - y_prom))/sum((x_i - x_prom)^2)
6 b0 <- y_prom - b1 * x_prom
```

Modelo lineal simple

MODELO LINEAL A MANO

En nuestro código esta cuenta

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

se escribe

```
1 x_i <- edad_salario_18_40$edad # [18, 19, ..., 40]
2 y_i <- edad_salario_18_40$salario_prom # [400K, 300K, ..., 800K]
3 x_prom <- mean(x_i)
4 y_prom <- mean(y_i)
5 b1 <- sum((x_i - x_prom) * (y_i - y_prom))/sum((x_i - x_prom)^2)
6 b0 <- y_prom - b1 * x_prom
7 cat(b0, b1)

>>> 14446 21726
```

Modelo lineal simple

MODELO LINEAL EN R

Podemos pedirle R que haga la cuenta por nosotros

Modelo lineal simple

MODELO LINEAL EN R

Podemos pedirle R *que haga la cuenta por nosotros*

```
1 mod <- lm(  
2   formula = salario_prom ~ edad,  
3   data = edad_salario_18_40  
4 )  
5 mod
```

Modelo lineal simple

MODELO LINEAL EN R

Podemos pedirle R que haga la cuenta por nosotros

```
1 mod <- lm(  
2   formula = salario_prom ~ edad,  
3   data = edad_salario_18_40  
4 )  
5 mod
```

Coefficients:

(Intercept)	edad
14446	21726

Modelo lineal simple

MODELO LINEAL EN R

Podemos pedirle R que haga la cuenta por nosotros

```
1 mod <- lm(  
2   formula = salario_prom ~ edad,  
3   data = edad_salario_18_40  
4 )  
5 mod
```

Coefficients:

(Intercept)	edad
14446	21726

```
1 b0 = mod$coefficients[1]  
2 b1 = mod$coefficients[2]  
3 cat(b0, b1)  
  
>>> 14446 21726
```

Modelo lineal simple

DATOS + LM

```
1 ggplot(edad_salario_18_40, aes(x = edad, y = salario_prom)) +
2   geom_point(alpha = 0.8) +
3   labs(
4     x = "Edad",
5     y = "Salario promedio",
6   ) +
7   geom_function(fun = \((x)\) b0 + b1*x, # Agregamos una recta
8     linewidth = 1.5, col = "#C95E61" # b0 + b1*edad
9   ) +
10  theme_minimal()
```

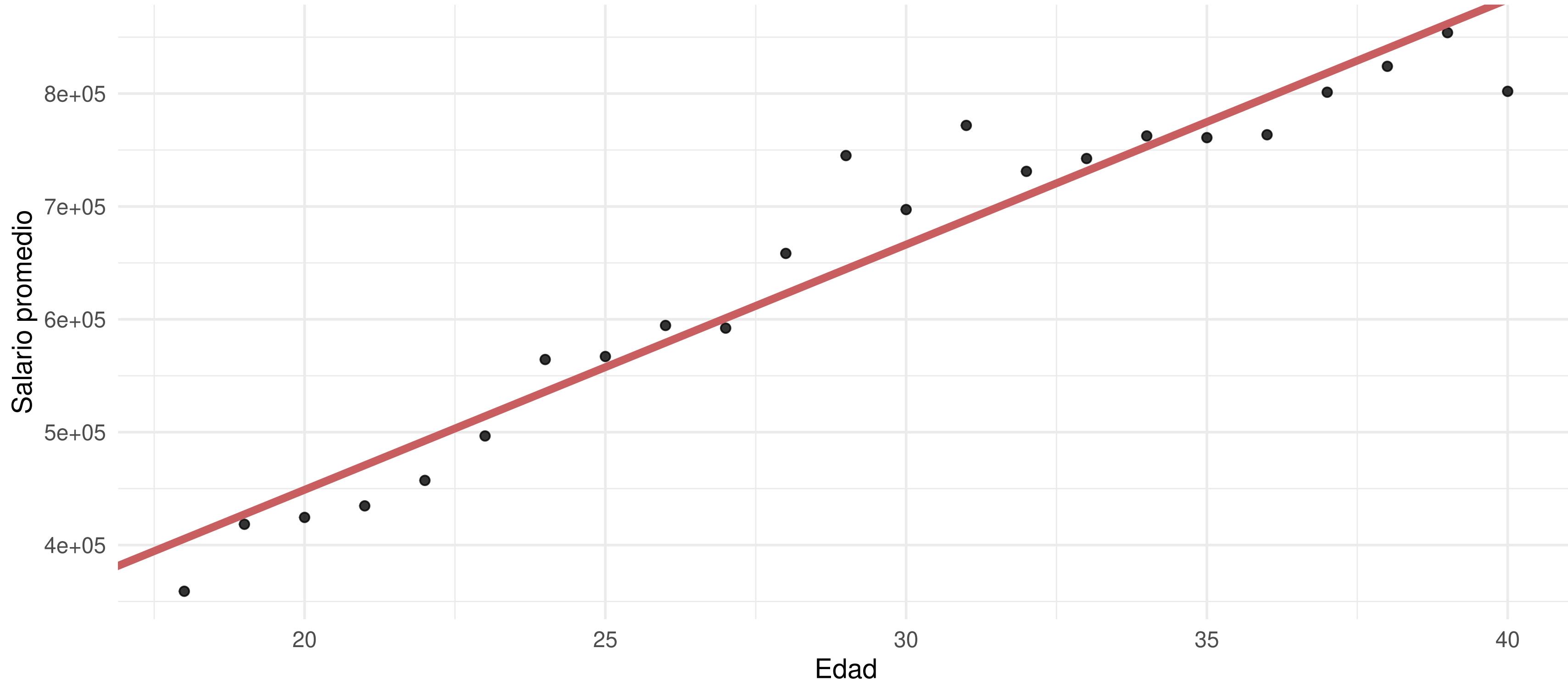
Modelo lineal simple

DATOS + LM

```
1 ggplot(edad_salario_18_40, aes(x = edad, y = salario_prom)) +
2   geom_point(alpha = 0.8) +
3   labs(
4     x = "Edad",
5     y = "Salario promedio",
6   ) +
7   geom_abline(intercept = b0, slope = b1, # También se puede
8               linewidth=1.5, col = "#C95E61"           # hacer así
9   ) +
10  theme_minimal()
```

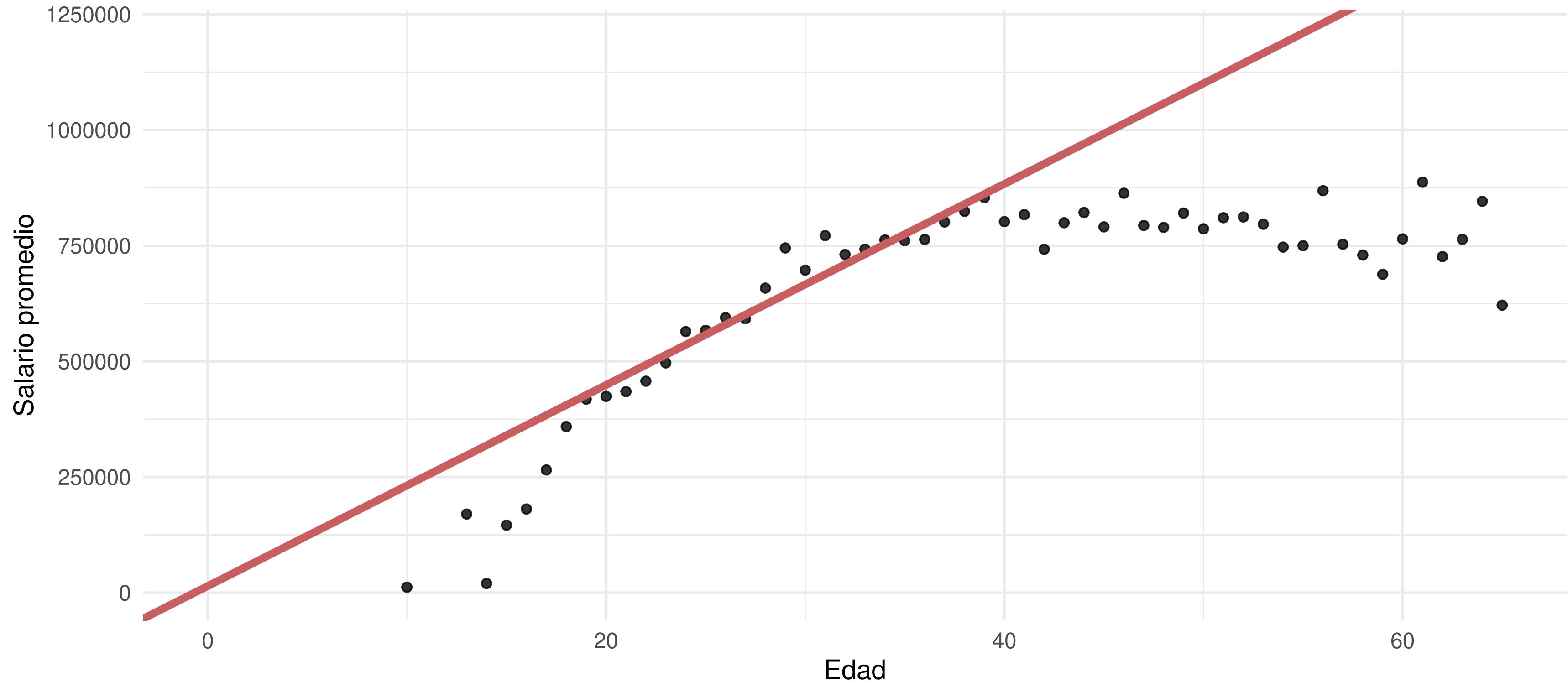
Modelo lineal simple

DATOS + LM



Modelo lineal simple

DATOS + LM



Modelo lineal simple

PREDICCIONES DEL MODELO

Recordemos que

$$E(Y|X) = \beta_0 + \beta_1 X$$

Modelo lineal simple

PREDICCIONES DEL MODELO

Recordemos que

$$E(Y|X) = \beta_0 + \beta_1 X$$

Lo que quiere decir que el modelo me puede decir cuánto predice que es el sueldo promedio (Y) para una persona de determinada edad (X):

$$E(\text{ sueldo} | \text{edad}) = 14446 + 21726 \text{ edad}$$

Modelo lineal simple

PREDICCIONES DEL MODELO

Recordemos que

$$E(Y|X) = \beta_0 + \beta_1 X$$

Lo que quiere decir que el modelo me puede decir cuánto predice que es el sueldo promedio (Y) para una persona de determinada edad (X):

$$E(\text{ sueldo} | \text{edad}) = 14446 + 21726 \text{ edad}$$

ej: $E(\text{ sueldo} | 35) = 14446 + 21726 \cdot 35 = \774844

Modelo lineal simple

PREDICCIONES (Y LIMITACIONES) DEL MODELO

```
1 library(glue)
2
3 glue("Alguien con 35 años gana en promedio ${round(b0 + 35 * b1)}")
4 glue("Alguien recién nacido gana en promedio ${round(b0 + 0 * b1)}")
5 glue("Alguien con 90 años gana en promedio ${round(b0 + 90 * b1)}")

>>> Alguien con 35 años gana en promedio $774844
```

Modelo lineal simple

PREDICCIONES (Y LIMITACIONES) DEL MODELO

```
1 library(glue)
2
3 glue("Alguien con 35 años gana en promedio ${round(b0 + 35 * b1)}")
4 glue("Alguien recién nacido gana en promedio ${round(b0 + 0 * b1)}")
5 glue("Alguien con 90 años gana en promedio ${round(b0 + 90 * b1)}")

>>> Alguien con 35 años gana en promedio $774844
>>> Alguien recién nacido gana en promedio $14446
>>> Alguien con 90 años gana en promedio $1969754
```

Esto es todo.