Fundamentos de ciencia de datos con R - Módulo 4

Clase 4: Gráficos básicos - dispersón

CEPAL - Unidad de Estadísticas Sociales

2025-11-04

Introducción

Los gráficos de dispersión muestran la relación entre dos variables numéricas ubicando puntos en el plano (x, y). Son ideales para explorar tendencias, patrones, no linealidades, outliers y heterocedasticidad.

Para esta clase utilizaremos la librería tidyverse, que agrupa varias herramientas fundamentales para el análisis de datos en R. Dentro de ella se encuentra ggplot2, el paquete que usaremos para crear nuestros gráficos de barras.

```
library(tidyverse)
```

Introducción

Nota

"Antes de ajustar un modelo, mira el diagrama de dispersión." — Regla de oro exploratoria

¿Cuándo usar un gráfico de dispersión?

Usa dispersión cuando:

- Ambas variables son numéricas continuas (p. ej., ingreso_hh vs gasto_hh).
- Quieres evaluar asociación (positiva/negativa), forma (lineal/no lineal) y dispersión.
- Necesitas comparar subgrupos con color, forma o facetas.

Evítalo cuando:

- Una o ambas variables son categóricas (usa barras o boxplots).
- Hay millones de puntos sin transparencia/agrupamiento (prefiere hexbin o agregados).

Carga base de datos ejemplo

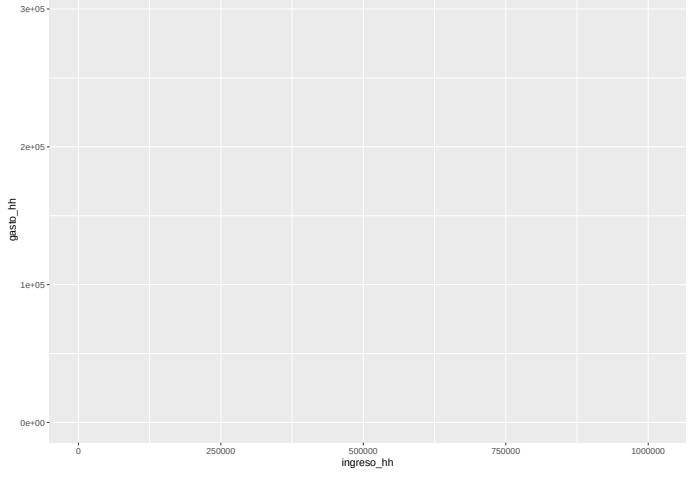
```
datos <- readRDS("../Data/base_personas_gasto.rds")
head(datos[,2:9], 5)
```

id_pers	upm	estrato	area	fep	pobreza	ingreso_hh	gasto_hh
1	1100100006	11001	1	19	3	10783.05	10783.05
2	1100100006	11001	1	19	3	10783.05	10783.05
1	1100100006	11001	1	16	3	21259.72	21259.72
2	1100100006	11001	1	16	3	21259.72	21259.72
3	1100100006	11001	1	16	3	21259.72	21259.72

Paso a paso: construyendo un gráfico de dispersión (gasto_hh vs ingreso_hh)

Paso 1: capa base (mapear x e y)

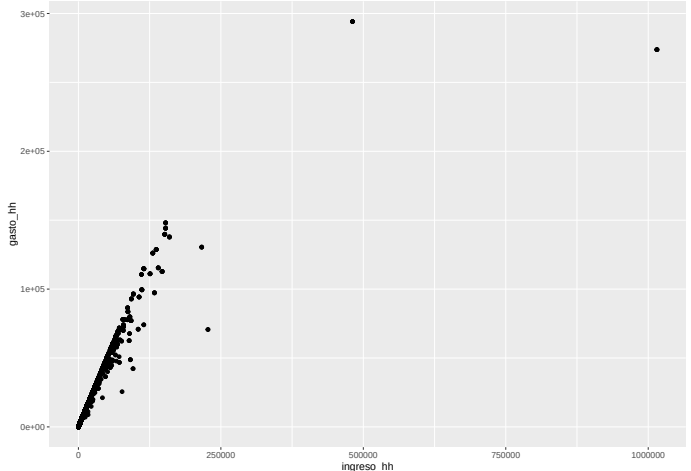
```
ggplot(datos, aes(x = ingreso_hh, y = gasto_hh))
```



Paso a paso: construyendo un gráfico de dispersión (gasto_hh vs ingreso_hh)

Paso 2: añadir geometría de puntos

```
ggplot(datos, aes(x = ingreso_hh, y = gasto_hh)) +
geom_point()
```



Paso a paso: construyendo un gráfico de dispersión (gasto_hh vs ingreso_hh)

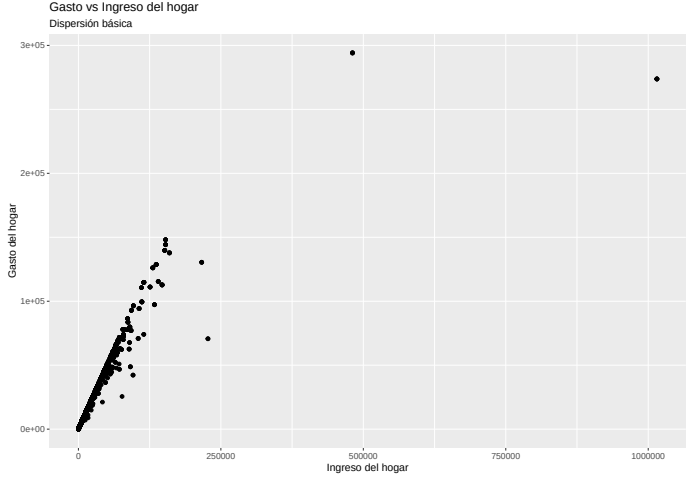
Paso 3: títulos y ejes

```
disp3 <- ggplot(datos, aes(x = ingreso_hh, y = gasto_hh)) +
geom_point() +
labs(
title = "Gasto vs Ingreso del hogar",
subtitle = "Dispersión básica",
x = "Ingreso del hogar", y = "Gasto del hogar"
)
```

Paso a paso: construyendo un gráfico de dispersión (gasto_hh vs ingreso_hh)

Paso 3: títulos y ejes

```
disp3
```



Paso a paso: construyendo un gráfico de dispersión (gasto_hh vs ingreso_hh)

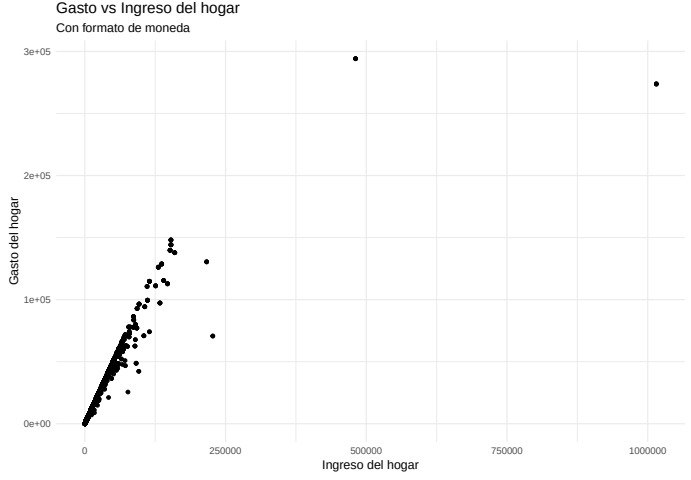
Paso 4: tema

```
disp4 <- ggplot(datos, aes(x = ingreso_hh, y = gasto_hh)) +
geom_point() +
labs(
title = "Gasto vs Ingreso del hogar",
subtitle = "Con formato de moneda",
x = "Ingreso del hogar", y = "Gasto del hogar"
) +
theme_minimal(base_size = 13)
```

Paso a paso: construyendo un gráfico de dispersión (gasto_hh vs ingreso_hh)

Paso 4: tema y)

```
disp4
```



Paso a paso: construyendo un gráfico de dispersión (gasto_hh vs ingreso_hh)

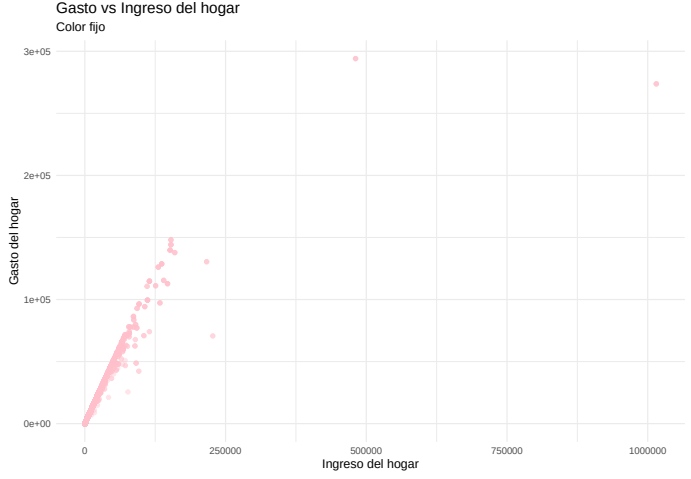
Paso 5: color, tamaño y transparencia

```
disp5 <- ggplot(datos, aes(x = ingreso_hh, y = gasto_hh)) +
geom_point(color = "pink", alpha = 0.4,
           size = 1.8) +  # color fijo + transparencia
labs(
title = "Gasto vs Ingreso del hogar",
subtitle = "Color fijo",
x = "Ingreso del hogar", y = "Gasto del hogar"
) +
theme_minimal(base_size = 13)
```

Paso a paso: construyendo un gráfico de dispersión (gasto_hh vs ingreso_hh)

Paso 5: color, tamaño y transparencia

```
disp5
```



Paso a paso: construyendo un gráfico de dispersión (gasto_hh vs ingreso_hh)

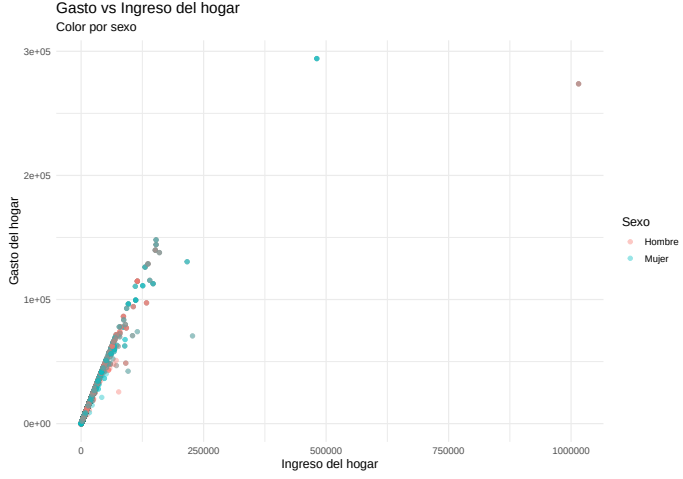
Paso 5: color, tamaño y transparencia

```
disp5b <- ggplot(datos, aes(x = ingreso_hh, y = gasto_hh,
                            color = sexo)) +
geom_point(alpha = 0.4, size = 1.8) +  # color mapeado a la variable
labs(
title = "Gasto vs Ingreso del hogar",
subtitle = "Color por sexo",
x = "Ingreso del hogar", y = "Gasto del hogar",
color = "Sexo"
) +
theme_minimal(base_size = 13)
```

Paso a paso: construyendo un gráfico de dispersión (gasto_hh vs ingreso_hh)

Paso 5: color, tamaño y transparencia

```
disp5b
```



Paso a paso: construyendo un gráfico de dispersión (gasto_hh vs ingreso_hh)

Paso 5: color, tamaño y transparencia

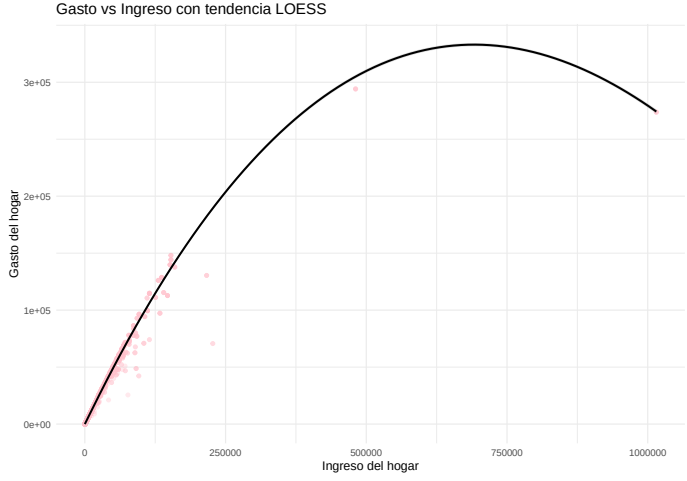
> **Nota**
>
> En `geom_point()`,
>
> - `color = "pink"` → aplica un color fijo a todos los puntos.
> - `aes(color = sexo)` → asigna un color distinto según la categoría de la variable `sexo`. Ambos pueden combinarse con `alpha` (transparencia) y `size` (tamaño del punto).

Ejemplo 1: Línea de tendencia

```
ej1 <- ggplot(datos, aes(x = ingreso_hh, y = gasto_hh)) +
geom_point(alpha = 0.35, size = 1.6, color = "pink") +
geom_smooth(method = "loess", se = TRUE, color = "black") +
labs(
title = "Gasto vs Ingreso con tendencia LOESS",
x = "Ingreso del hogar", y = "Gasto del hogar"
) +
theme_minimal(base_size = 13)
```

Ejemplo 1: Línea de tendencia

```
ej1
```



Ejemplo 2: Modelo lineal

```
ej2 <- ggplot(datos, aes(x = ingreso_hh, y = gasto_hh)) +
geom_point(alpha = 0.35, size = 1.6, color = "pink") +
geom_smooth(method = "lm", se = TRUE, color = "black") +
labs(
title = "Relación lineal: gasto ~ ingreso",
x = "Ingreso del hogar", y = "Gasto del hogar"
) +
theme_minimal(base_size = 13)
```

Ejemplo 2: Modelo lineal

```
ej2
```