# Devolución general del Ej 2

Eduardo Fernández

2023

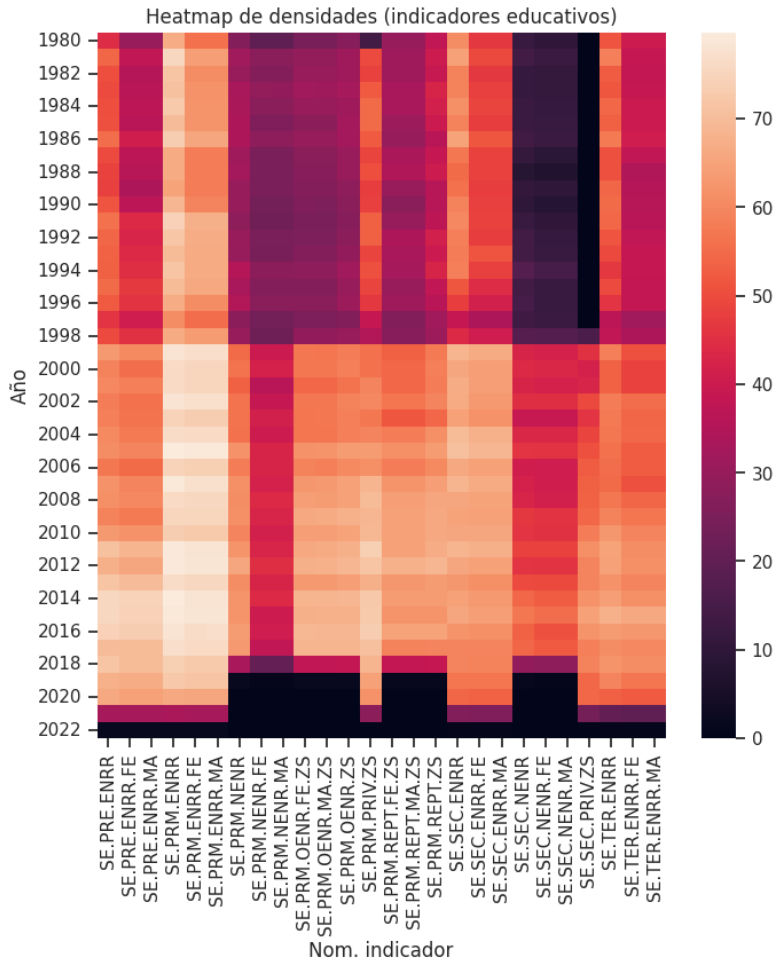
# Etapas

1. Averiguar la estructura de los datos.
2. Realizar sanity check.
   - Realizar preguntas iniciales que sirvan para explorar los datos.
3. Realizar más preguntas.

# Datos

- World-Development-Indicators https://databank.worldbank.org/reports.aspx?source=world-development-indicators Indicadores de países según el Banco Mundial.
- https://datasf.org Datos de crímenes en San Francisco.
- https://courses.cs.washington.edu/courses/cse512/21sp/data/weather.csv.gz estaciones meteorológicas en EEUU.
- https://catalogodatos.gub.uy/dataset/intendencia-montevideo-viajes-realizados-en-los-omnibus-del-stm líneas de bus en Montevideo.
- https://capitalbikeshare.com/system-data alquiler de bicicletas en Washington DC.
- https://www.kaggle.com/datasets/shivamb/netflix-shows Netflix

# Mapa de calor para buscar densidad de ausencias de datos



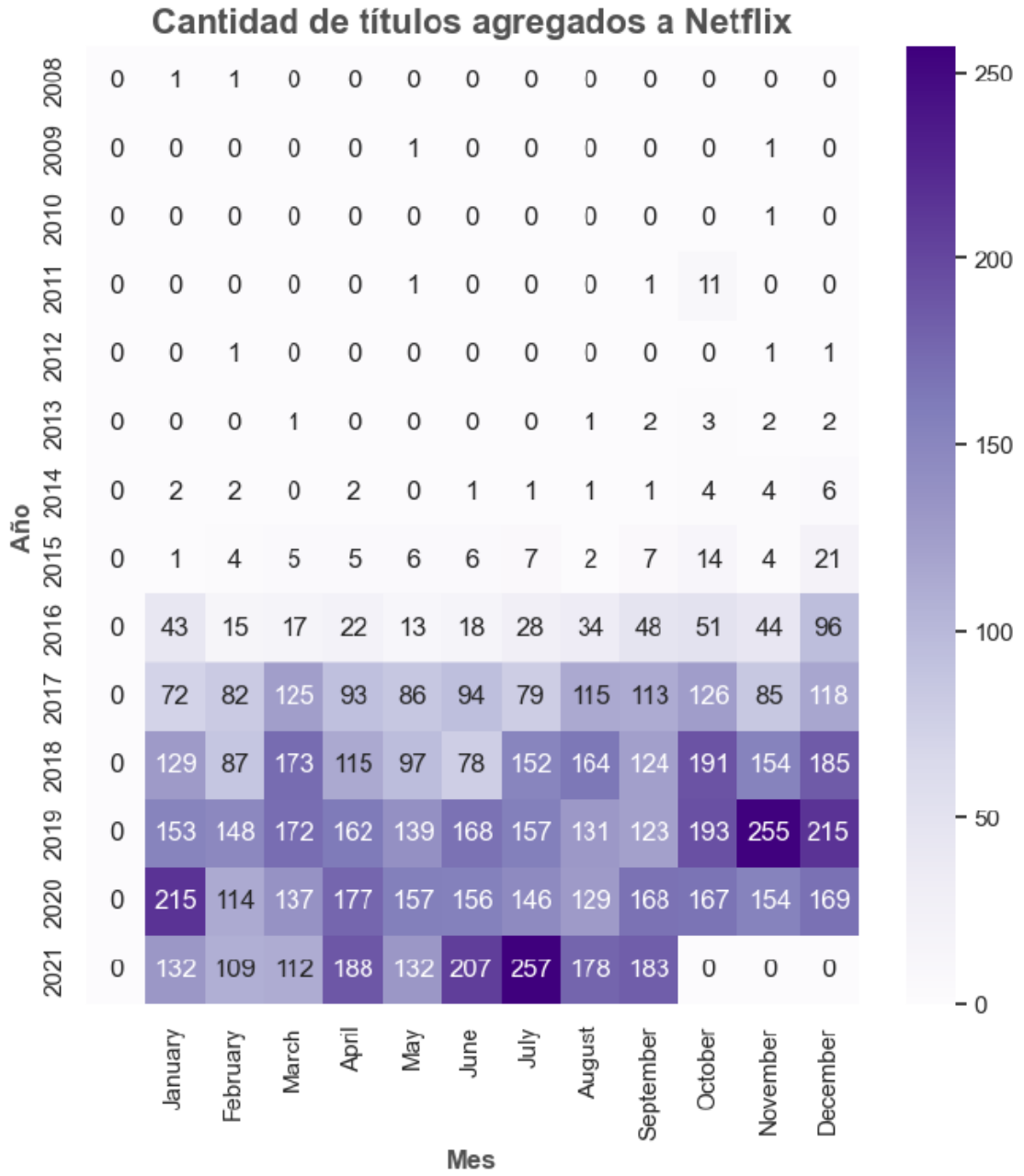

# Matriz de correlación de indicadores


**Podría aportar el uso de dendogramas para ordenar indicadores por similaridad.**

# Mapa de Calor

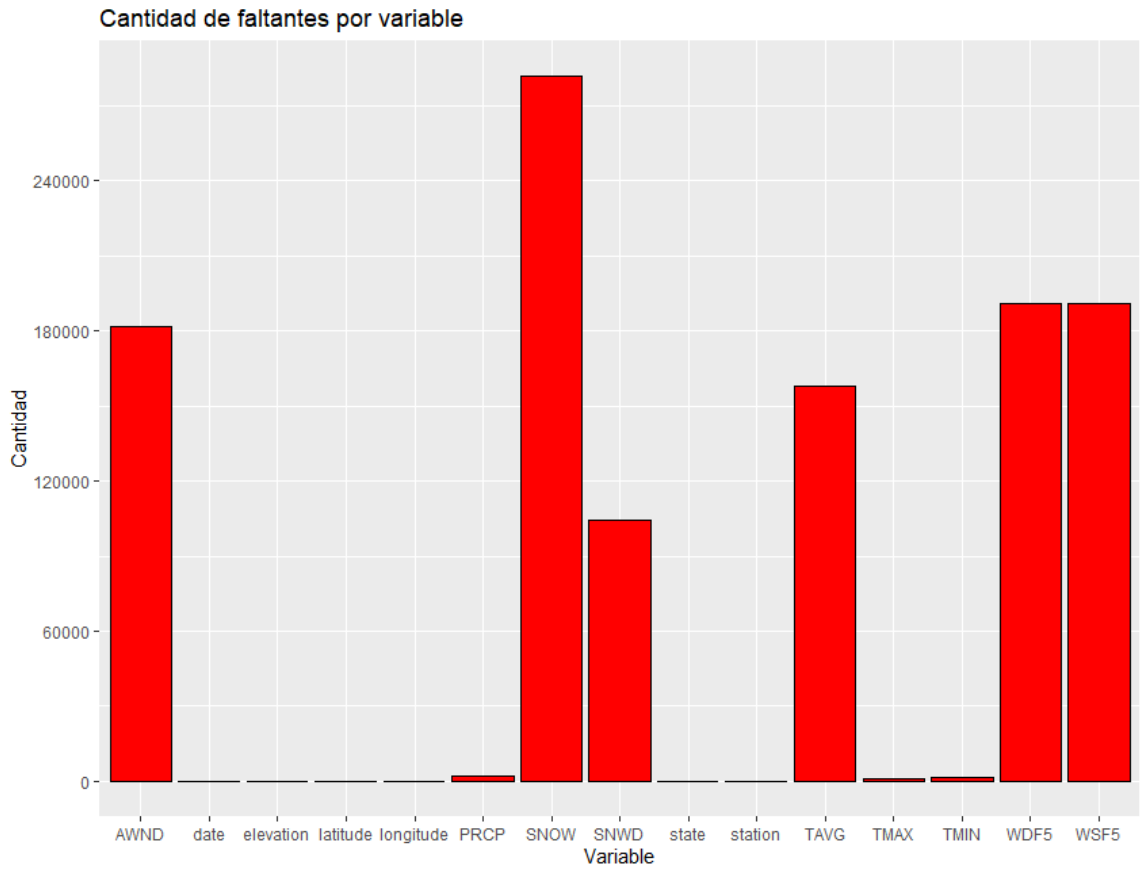
**. Colores complementados con números.**

- El color da la impresión general.
- La precisión la da el número.

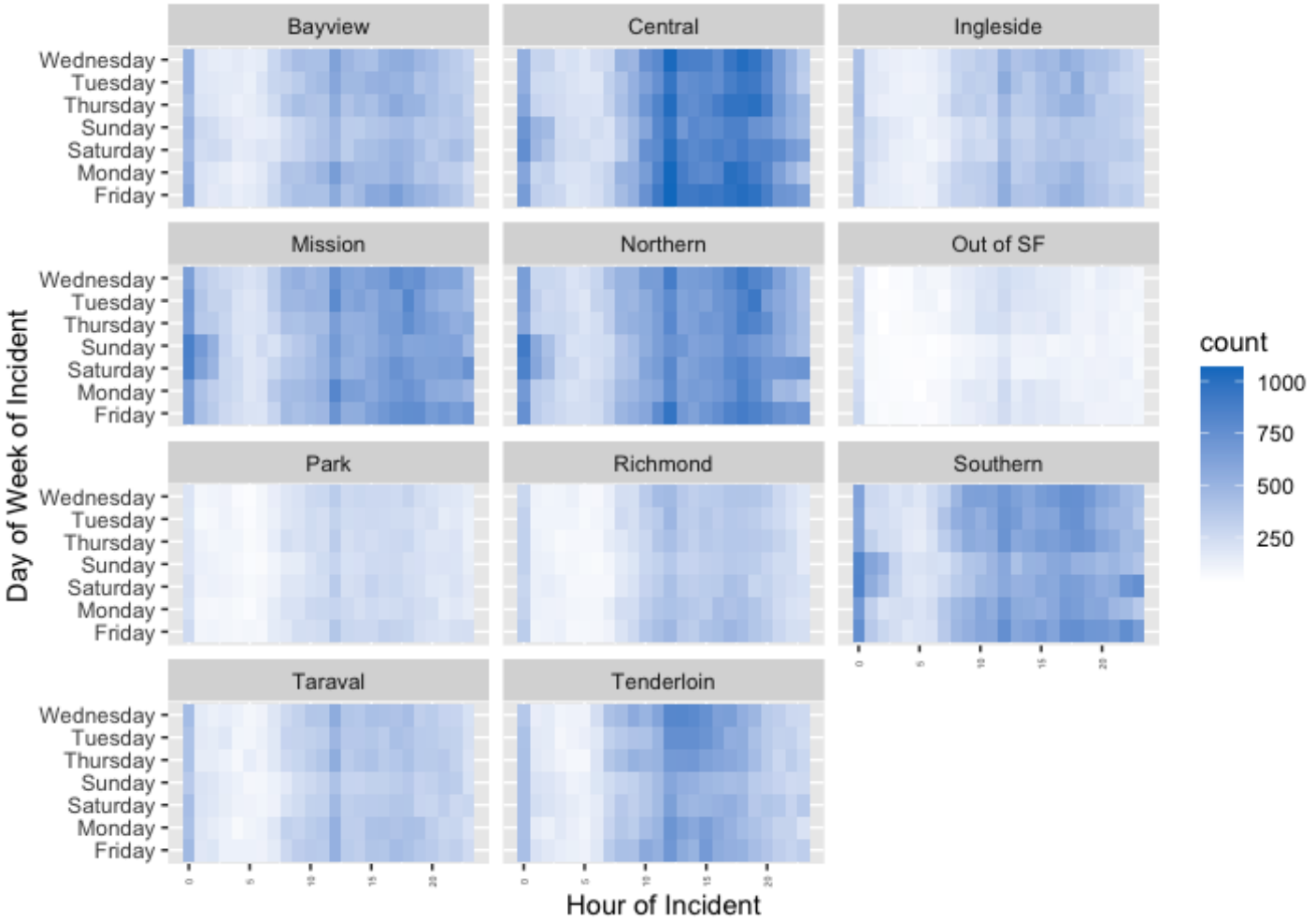


# Bar plots

- **Quizá que falten datos es razonable, y no invalida la tupla o fila de datos.**
- En Hawaii nadie pone la nieve que NO hay.
- En estaciones precarias o automáticas quizá no se calcula el viento/temperatura promedio del día.



# Múltiples mapas de calor para comparar datos

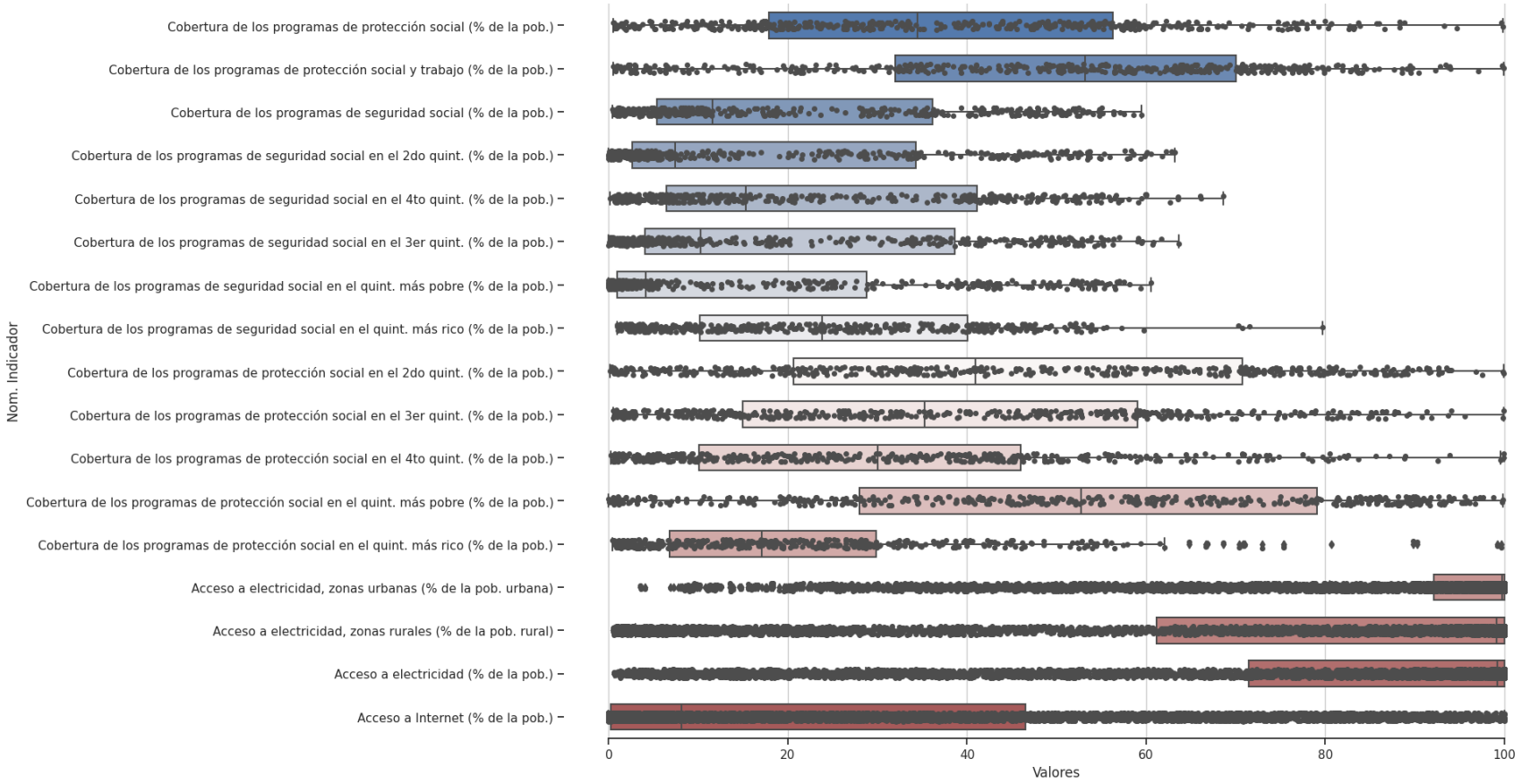


# Box-plot para analizar distribuciones.

**¿Los colores aportan?**

**A veces la densidad de puntos ocultan la forma del box plot.**

**¿Alternativas?: Violín, histogramas, Ridgeline**

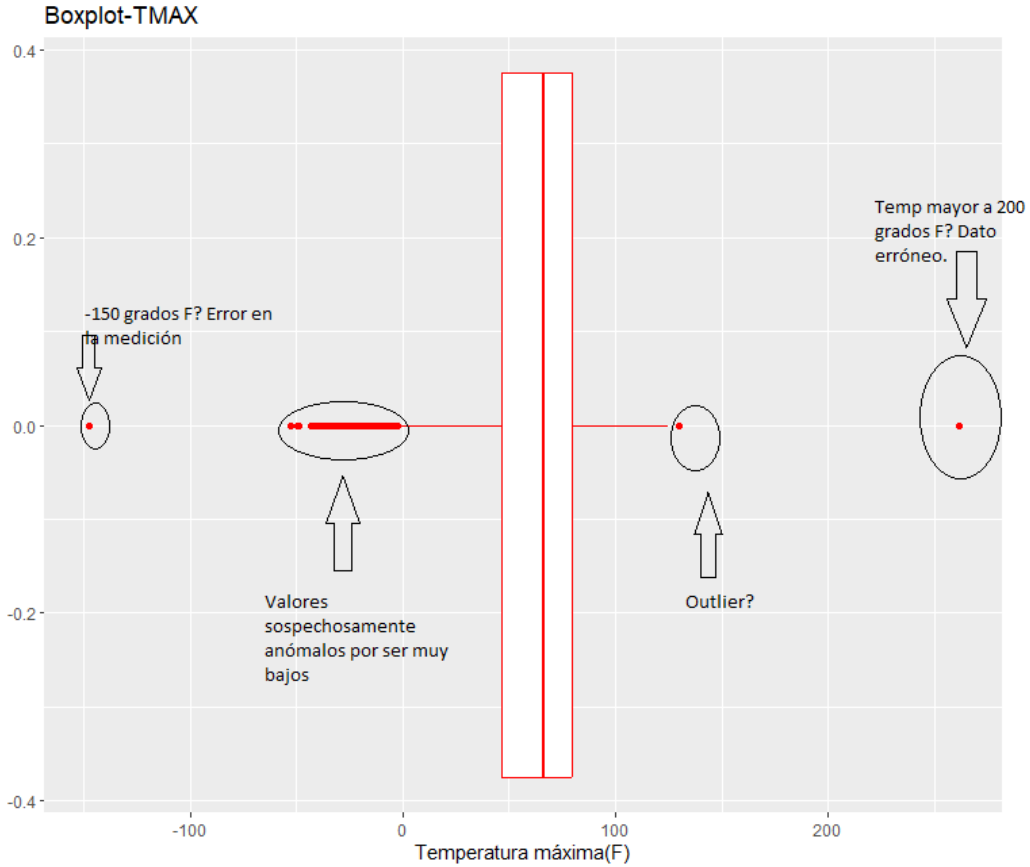


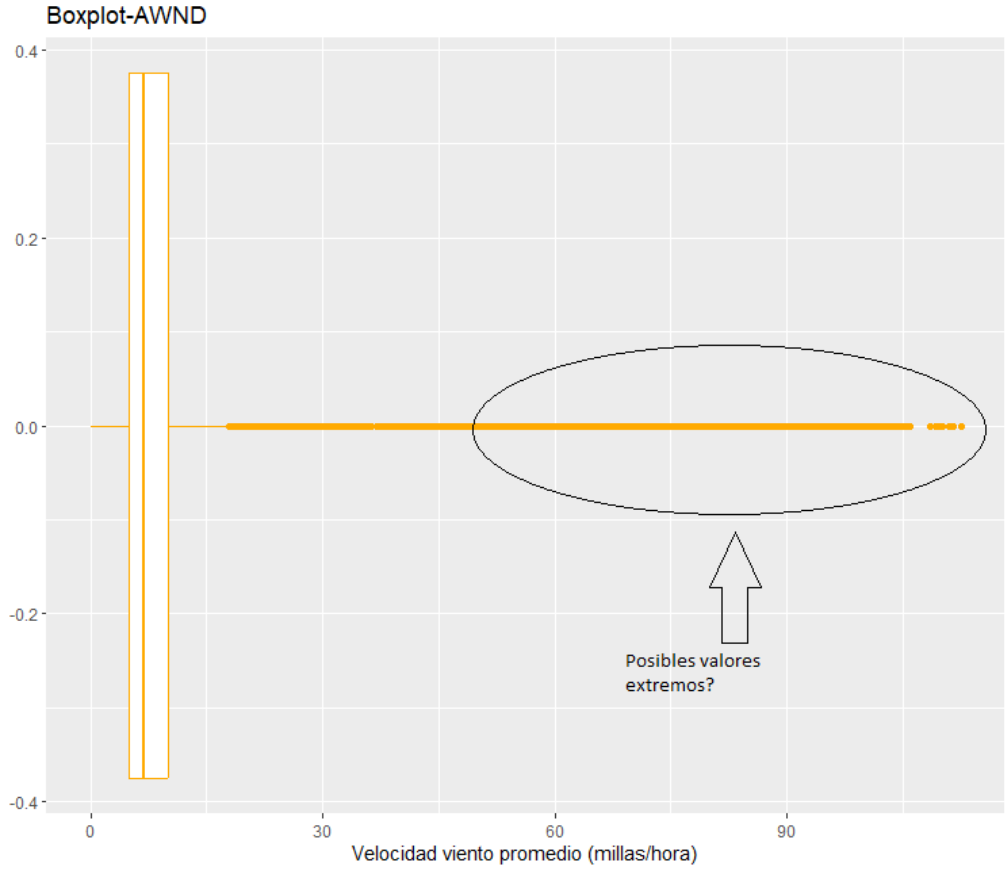
# Box-plot para analizar distribuciones.

**Outlier no significa que esté mal el dato.**

Puede estar correcto, o no.

También puede haber datos que no sean outliers pero estén mal.

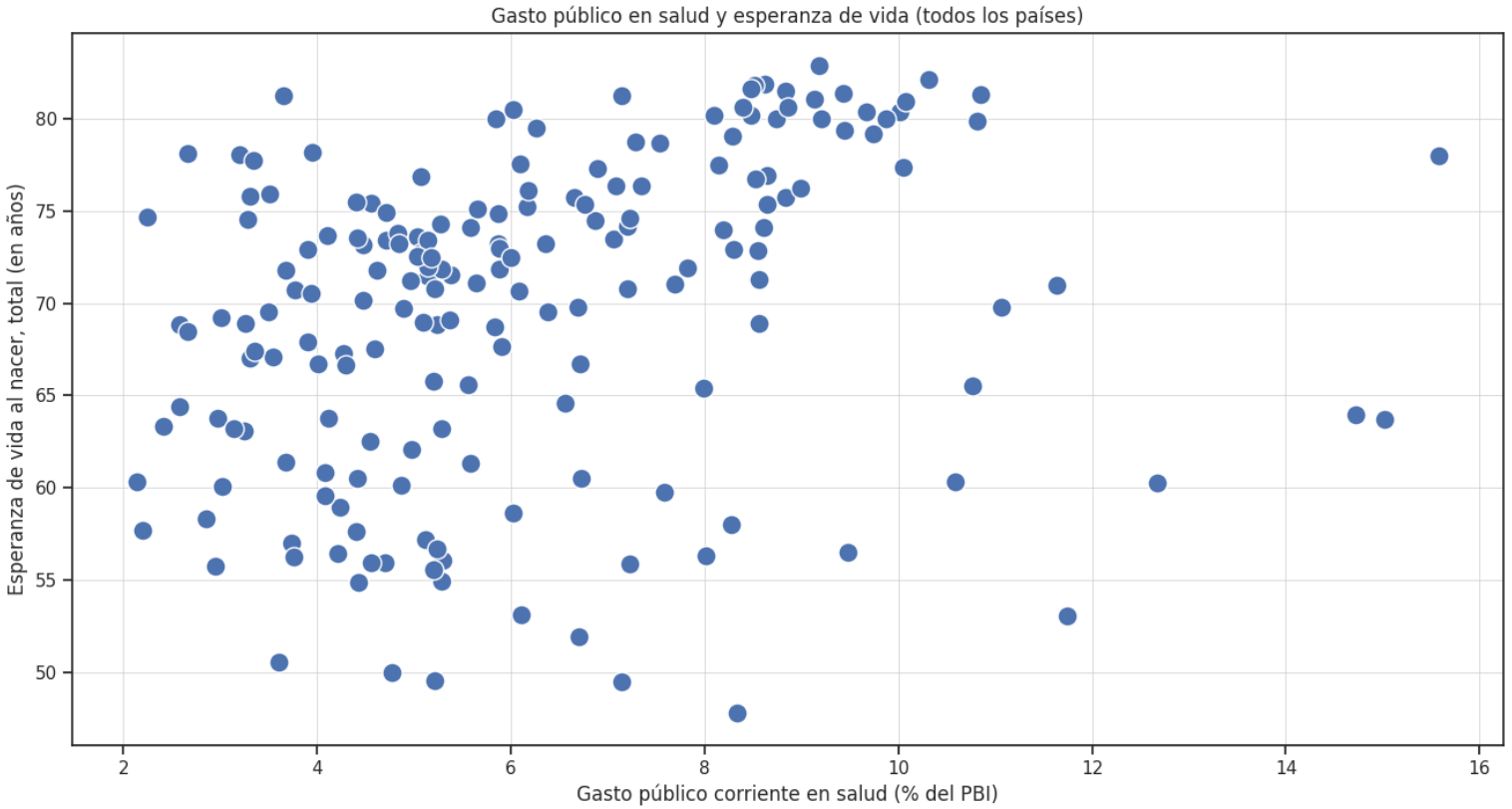


# Scatter plots

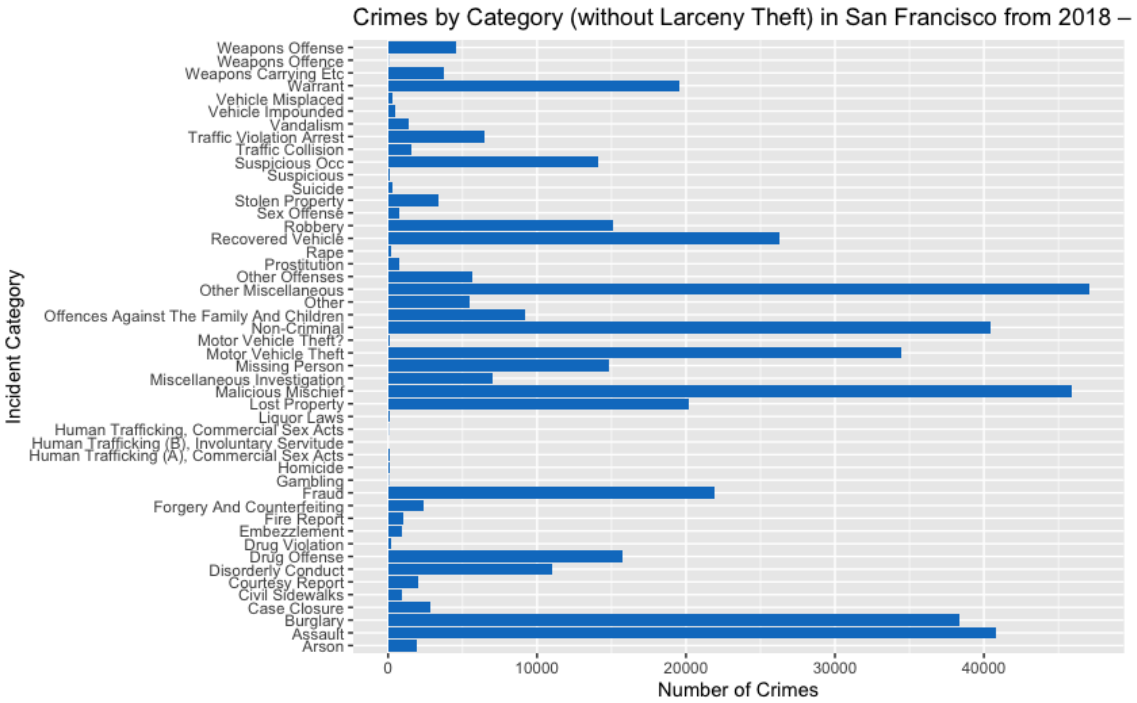
**Correlaciones no del todo claras.**

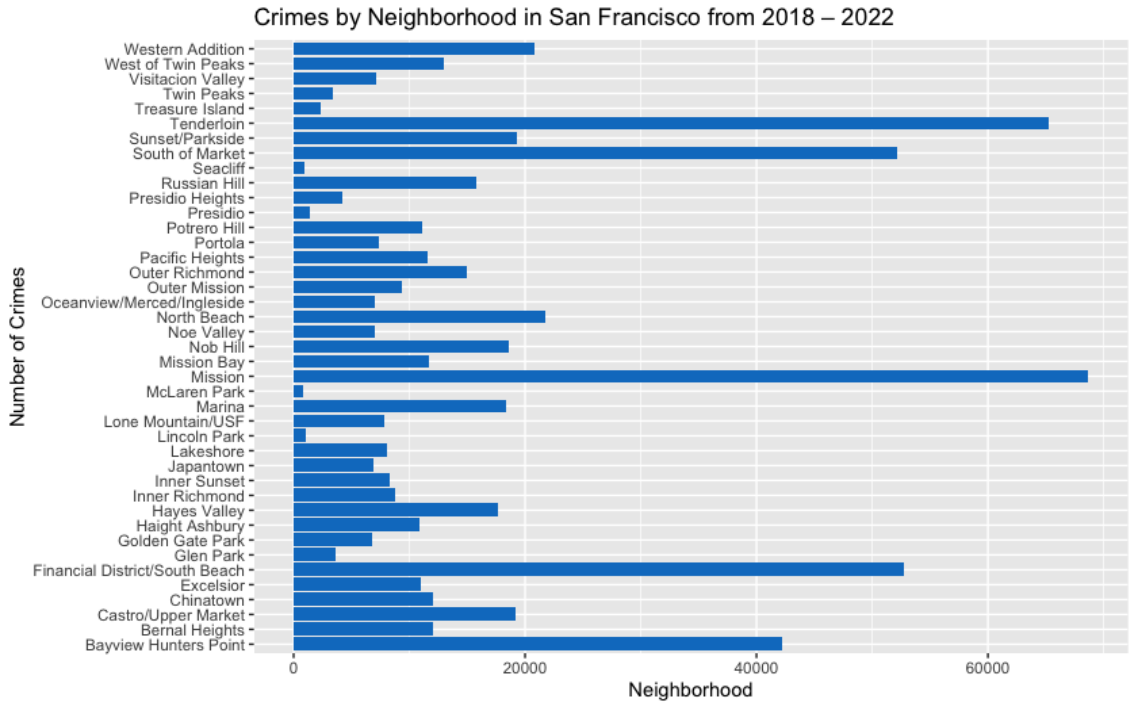
**Quizá conviene acompañarlas con una línea que señale la correlación.**



# Bar plots

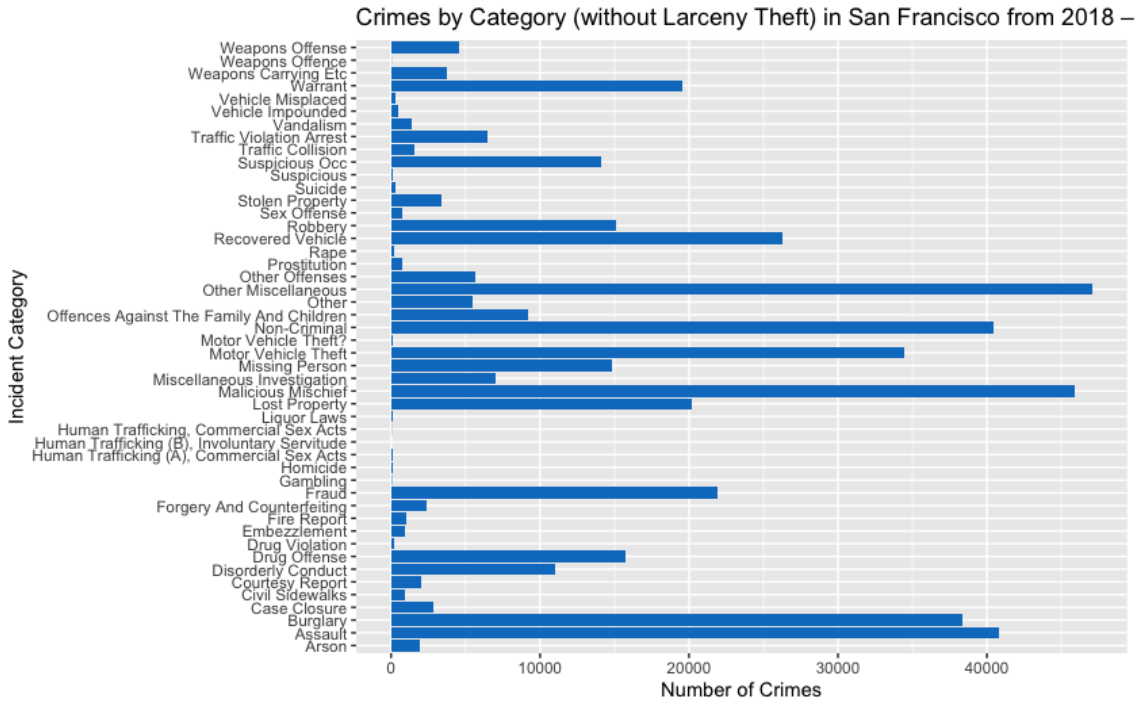
- **No todos los crímenes se realizan por igual en todos los vecindarios.**
- **Agrupar los crímenes por categorías según su peligrosidad y condena, y luego comparar vecindarios, normalizando valores.**
- **Los números en ratios de la población total.**

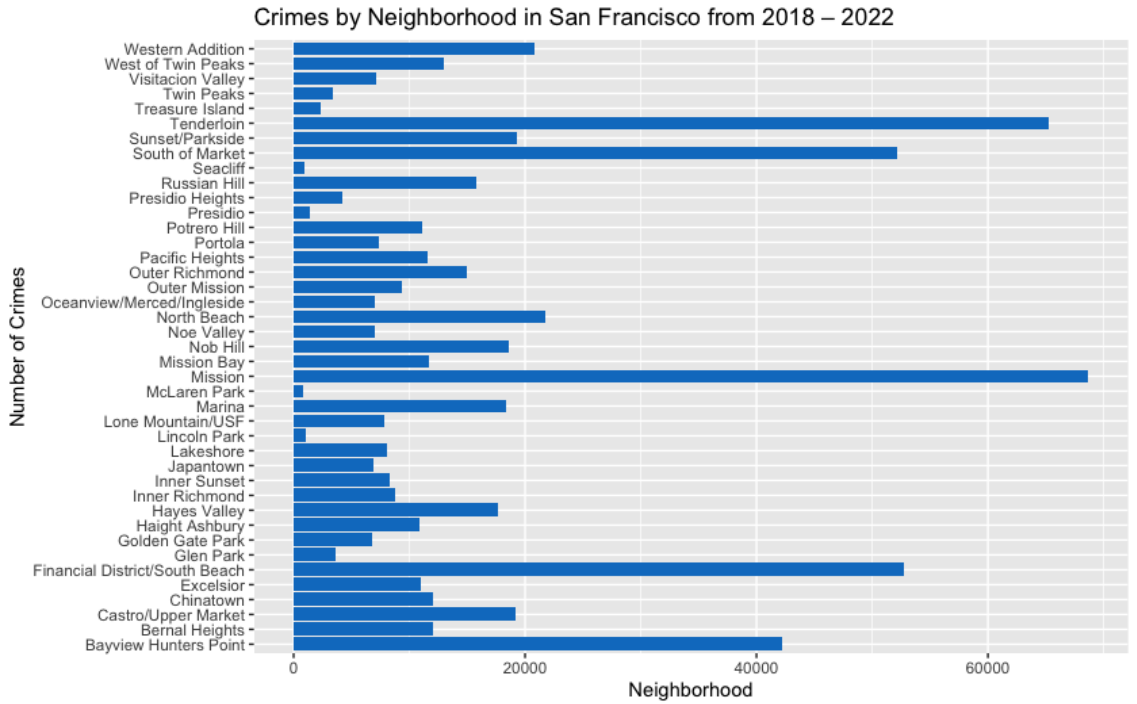


# Bar plots

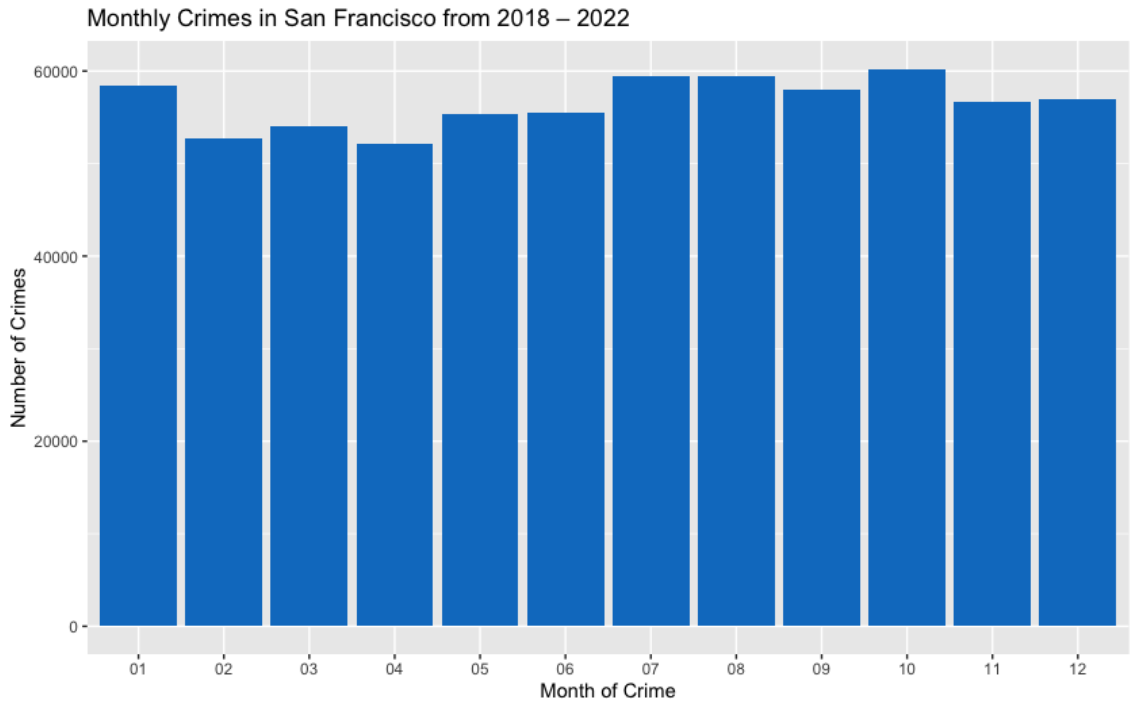
- **No todos los crímenes se realizan por igual en todos los vecindarios.**
- **Agrupar los crímenes por categorías según su peligrosidad y condena, y luego comparar vecindarios, normalizando valores.**
- **Los números en ratios de la población total.**



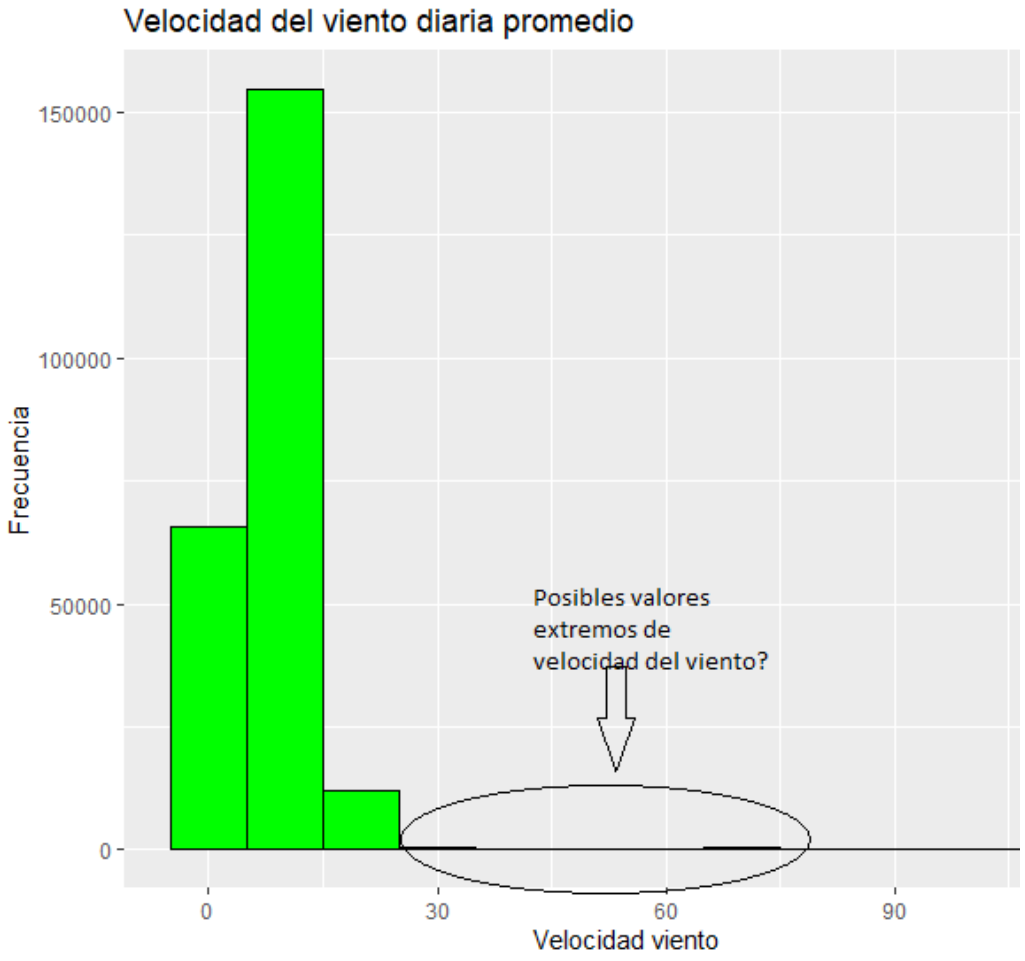
# Bar plots

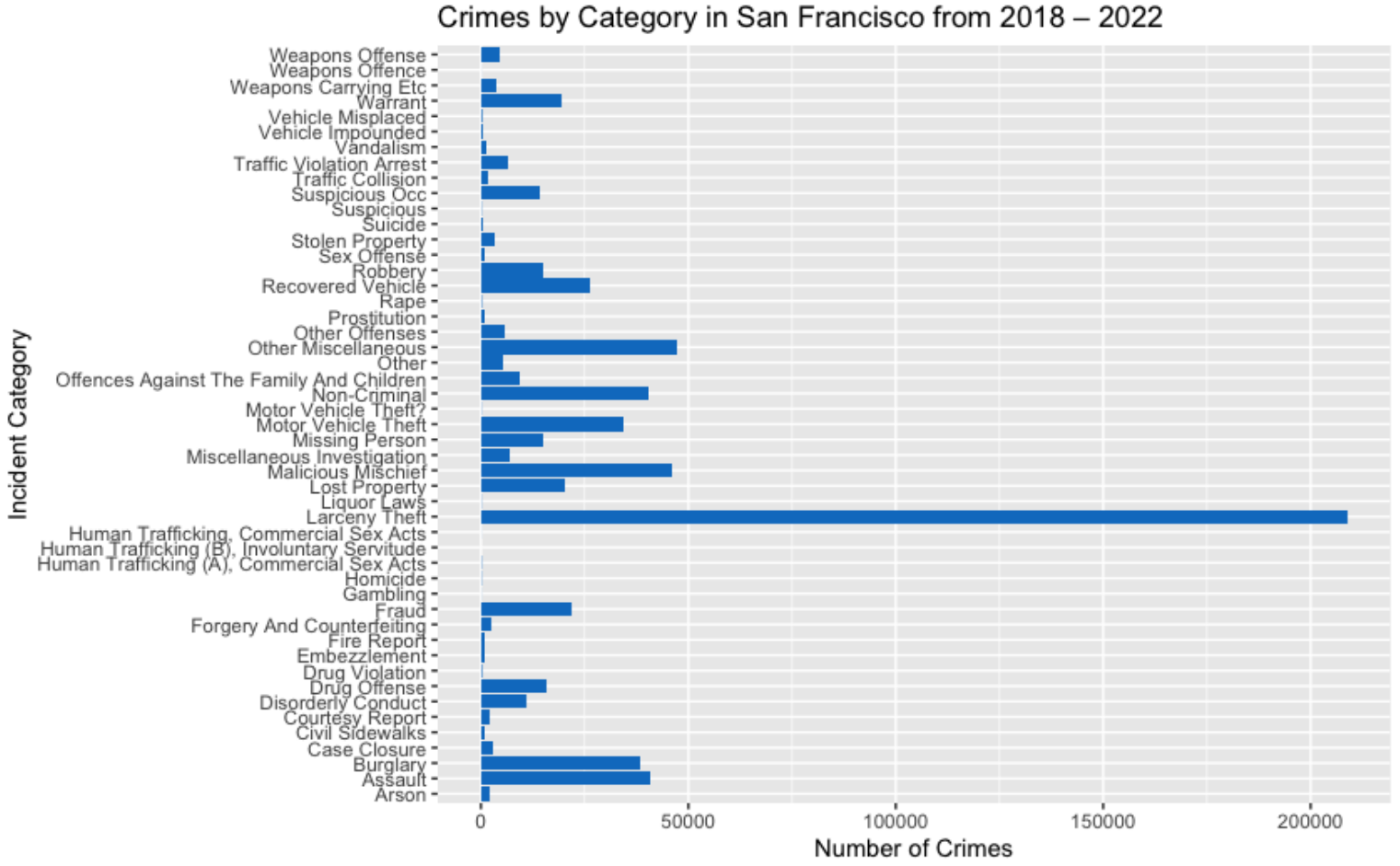
- **No todos los crímenes se realizan por igual en todos los meses.**
- **Agrupar los crímenes por categorías según su peligrosidad y condena, y luego comparar meses, normalizando valores.**



# Bar plots

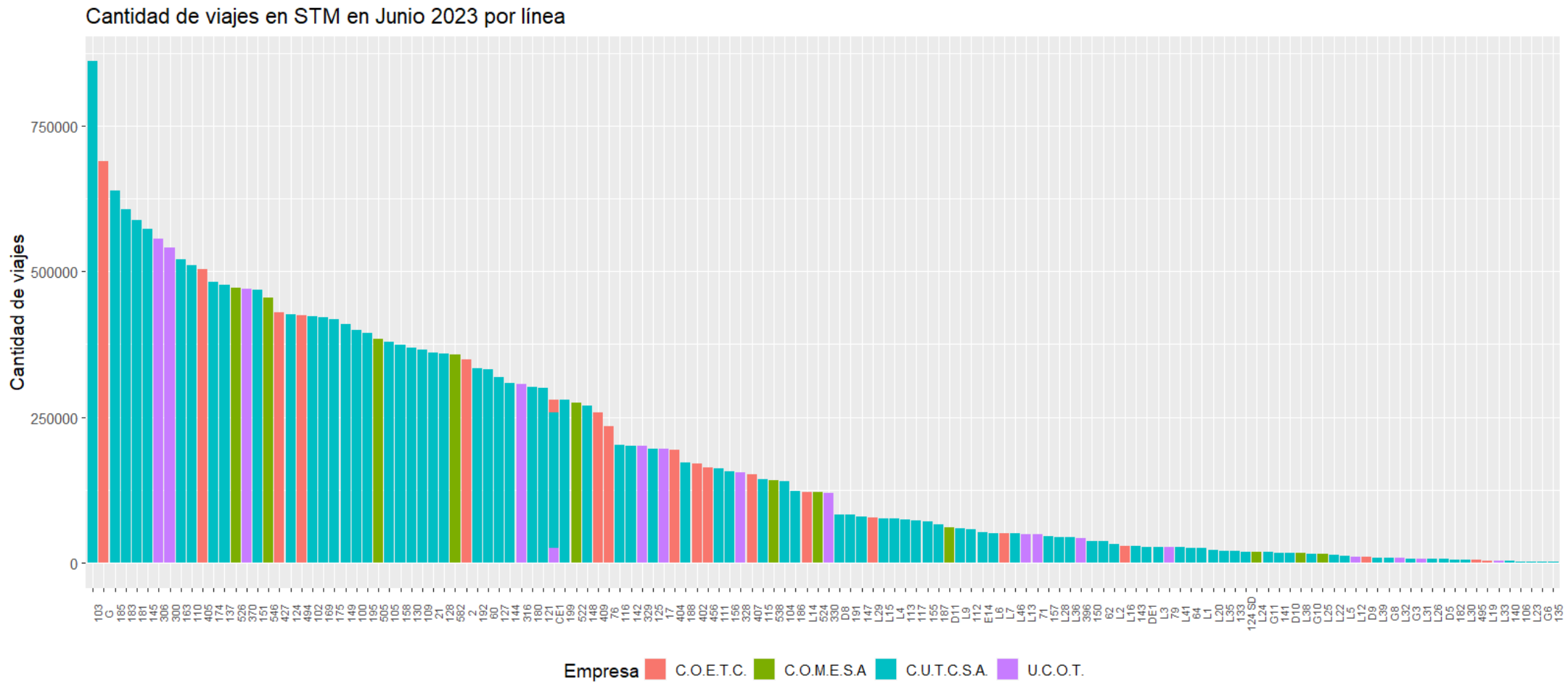
- **Escala logarítmica en el eje Y?**
- ¿Separarlos por su frecuencia y analizarlos aparte?



# Bar plots

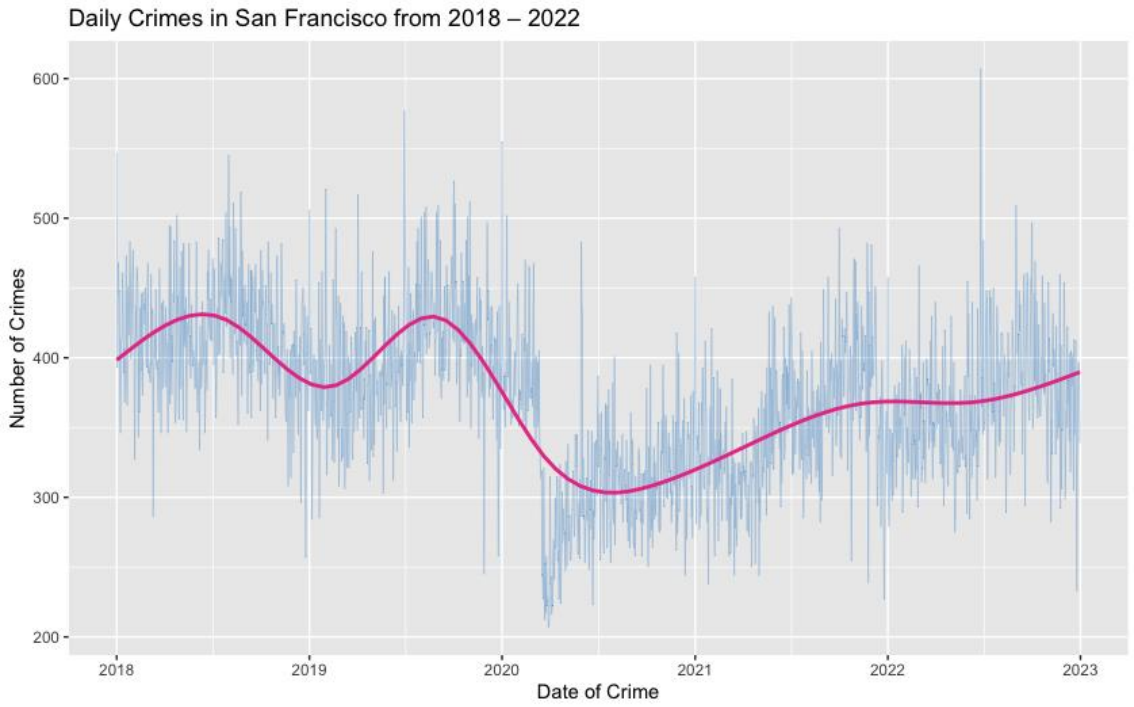
- **¿Escala logarítmica en el eje Y?**
- ¿Separarlos por su frecuencia y analizarlos aparte?

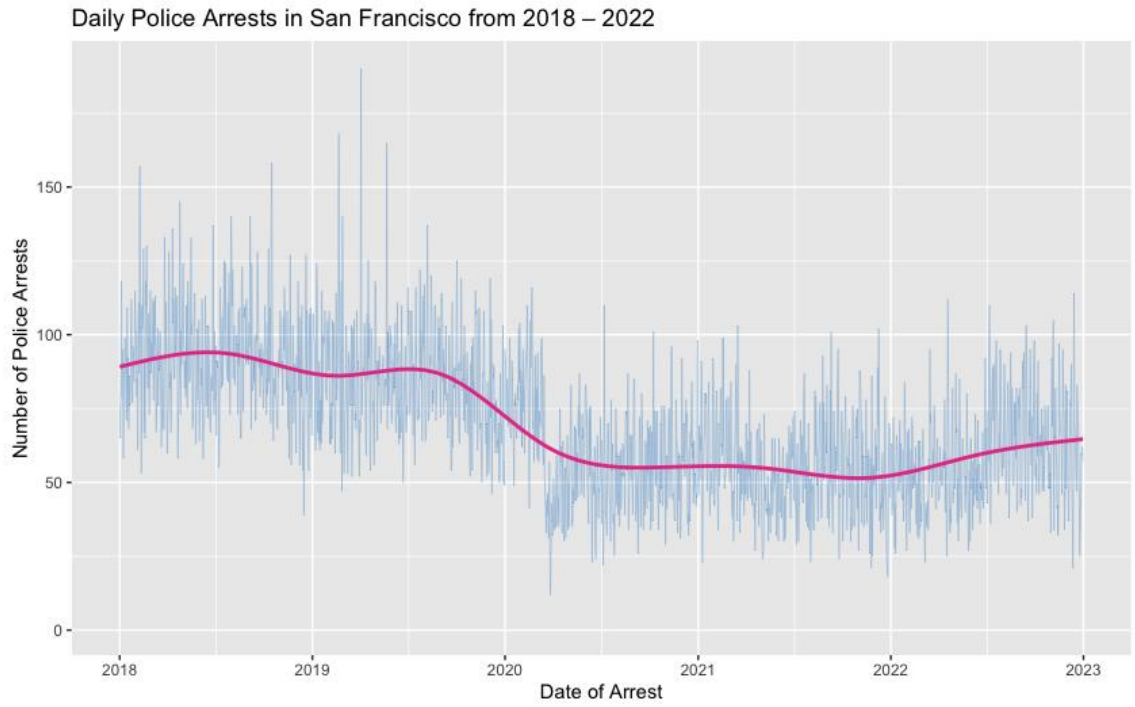


# Interpolar

- **Mostrar tendencias.**
  - Los crímenes vuelven a los números pre-pandemia, pero la policía aun no vuelve al número de arrestos.

- **Ojo, la interpolación no deja de ser una interpretación.**
  - Hay que hacer análisis más finos.
  - Quizá hubo un cambio en el tipo de crimen, o cambiaron los procedimientos policiales. En consecuencia, varían los arrestos.

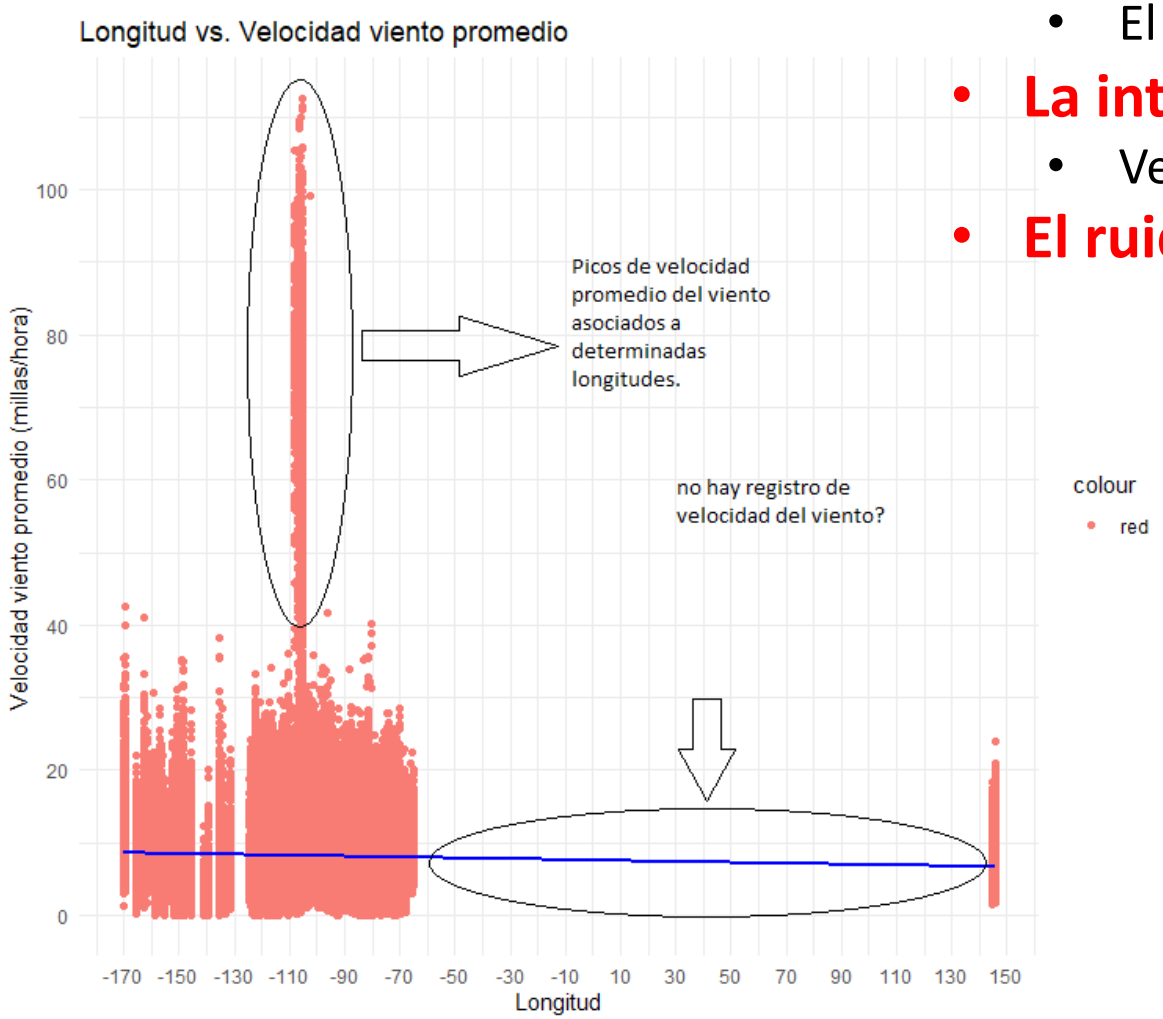


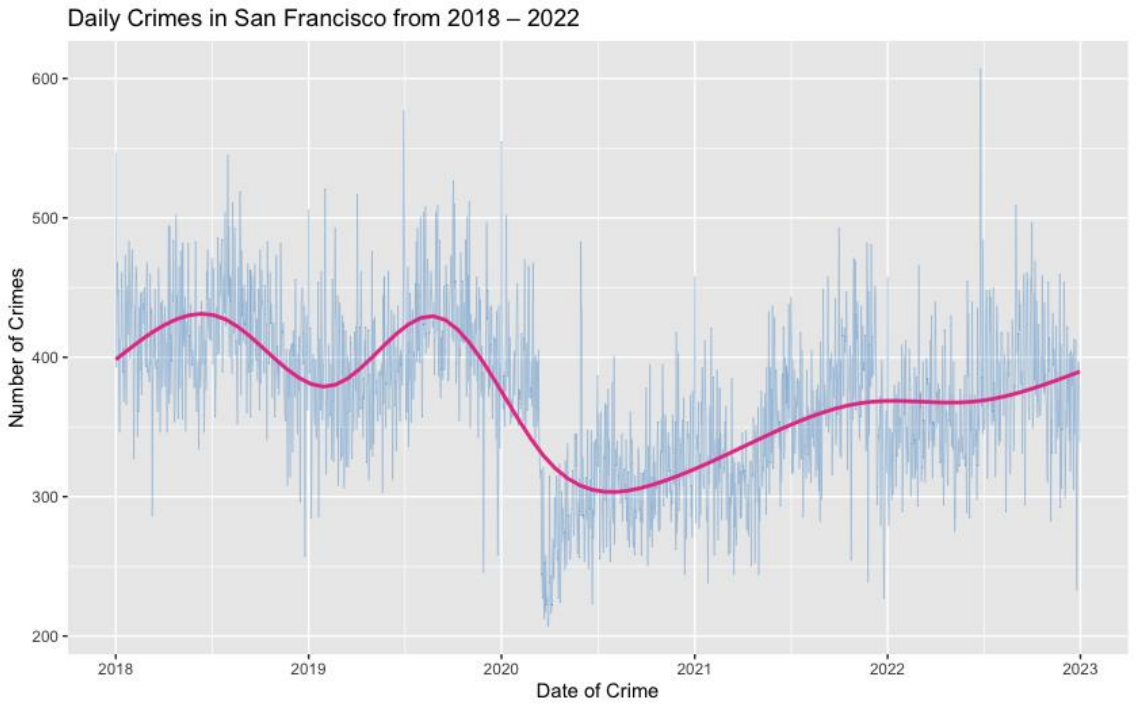



# Interpolar

## La interpolación no reemplaza a los datos.

- **Las interpolaciones pueden obviar los picos.**
  - Los picos hay que analizarlos aparte.
- **Puede dar continuidad a la discontinuidad.**
  - El cambio de régimen requiere de análisis particulares.
- **La interpolación puede inventar valores donde no hay.**
  - Ver donde es pertinente aplicarla.
- **El ruido (la varianza) también es dato importante.**

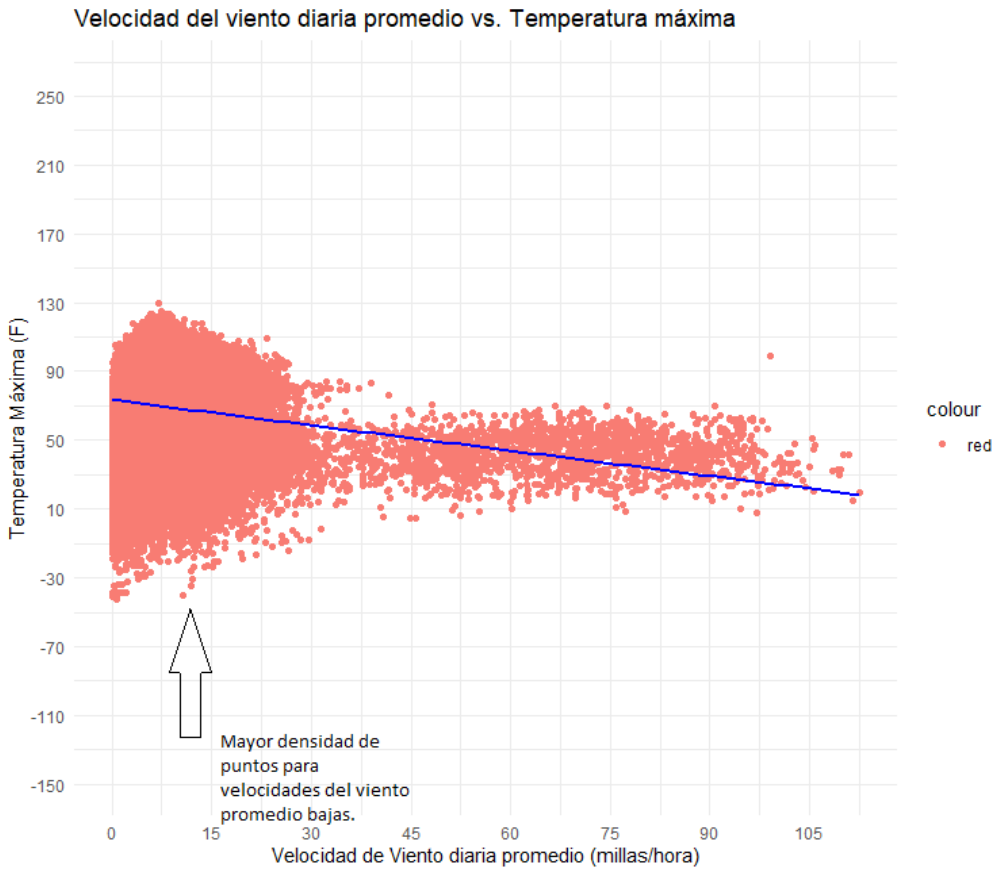


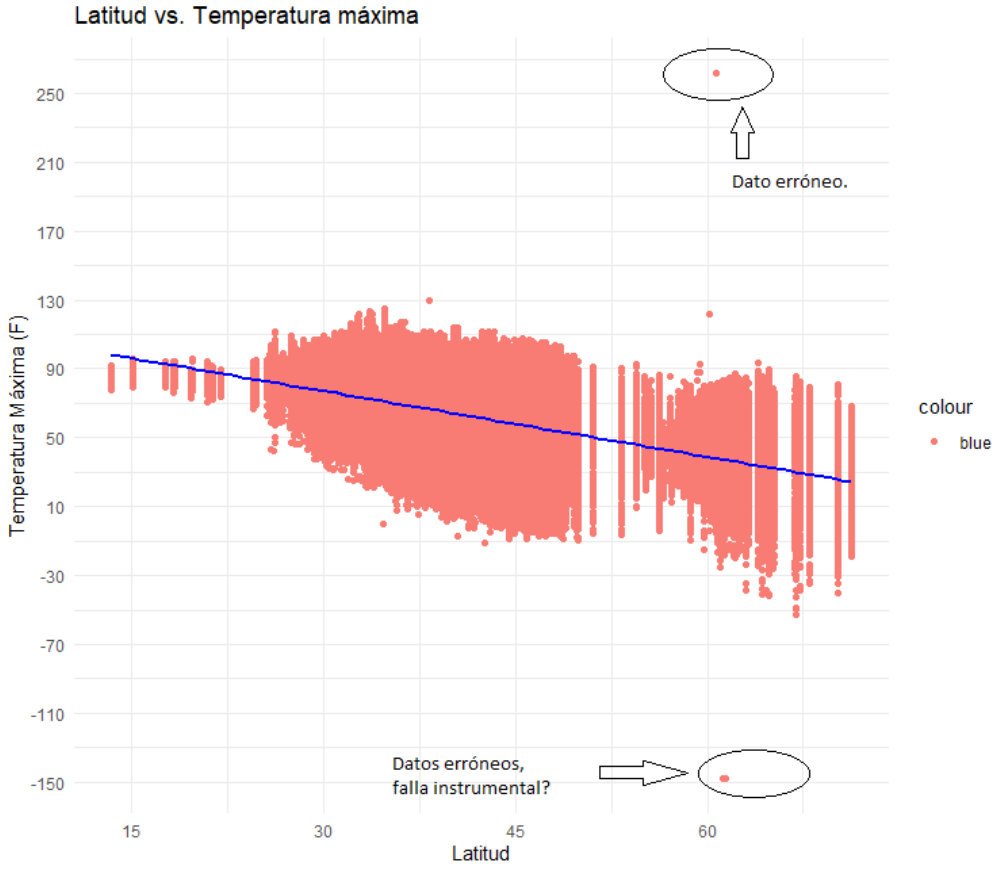



# Dot plots e interpolación

**Acotar las escalas.**

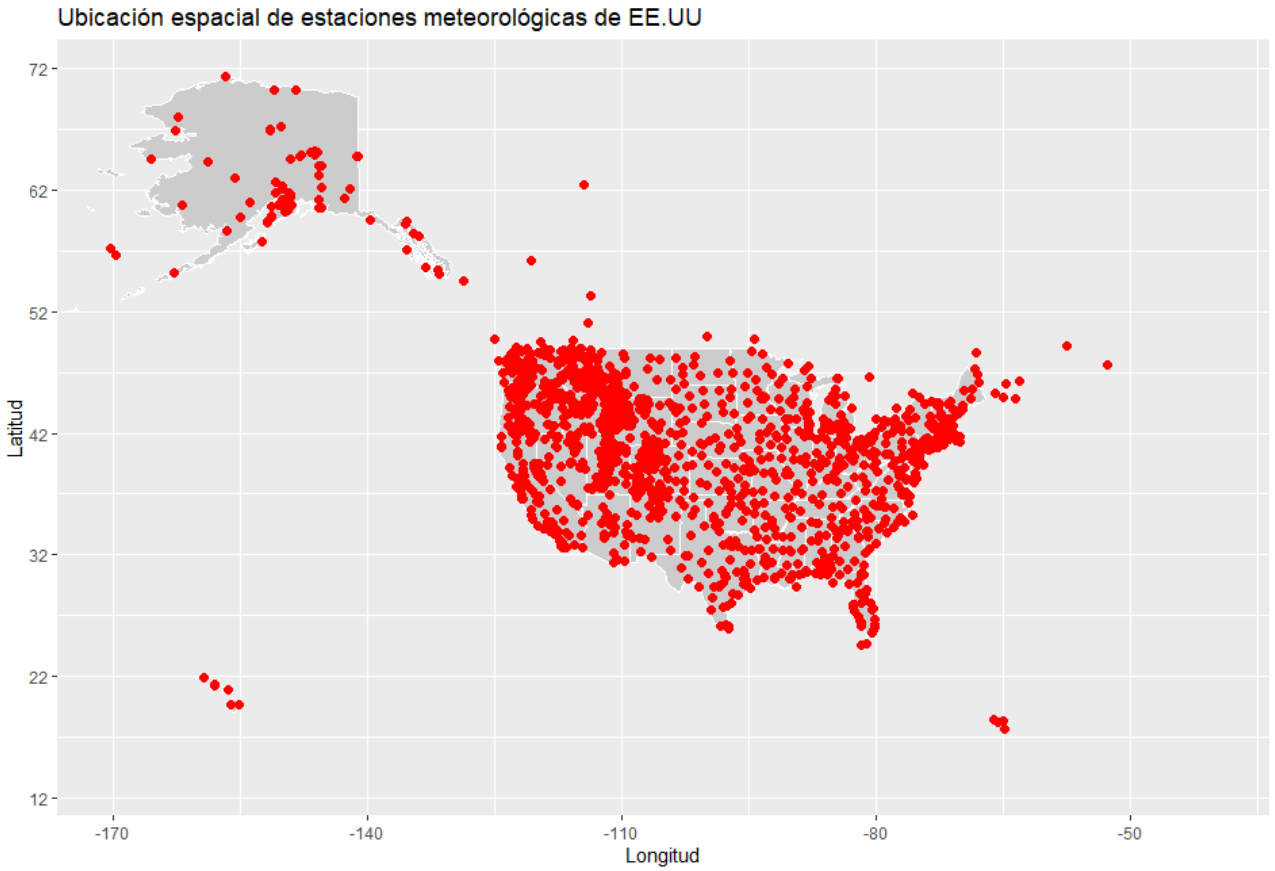
- Intervalo  en Y es el doble de lo necesario ([-50 140]).
- Quizá haya que quitar outliers para ver mejor las tendencias.
- Las tendencias quedan más horizontales.



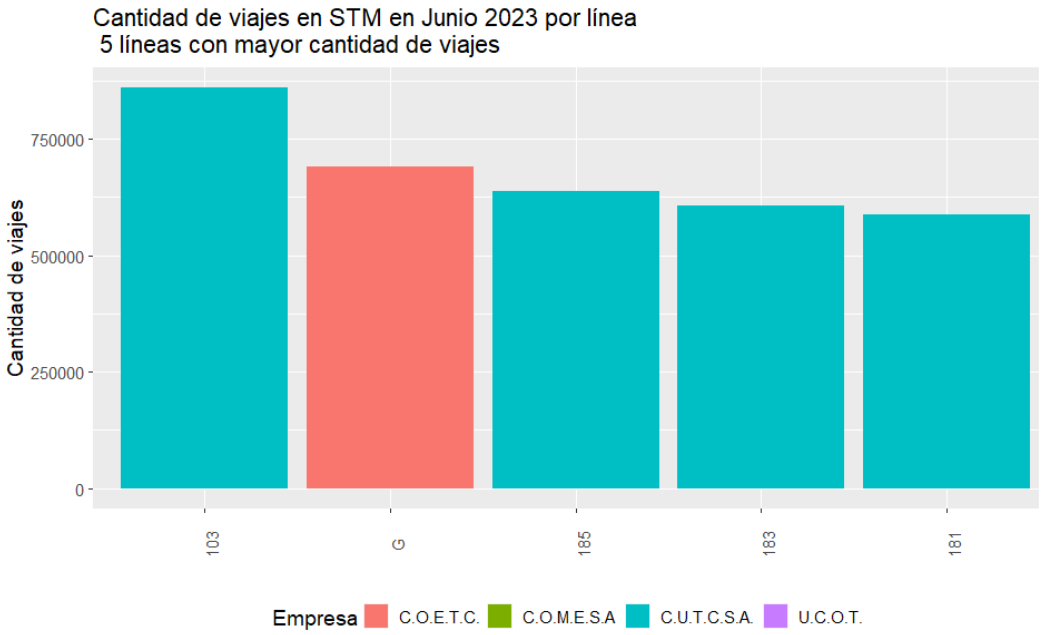
# Dot plot para mostrar densidad

- **Cuando se superponen muchos puntos, hay áreas con color uniforme, pero eso no significa que la densidad sea uniforme.**
- Quizá poner transparencia a los puntos (Alpha=0.1), entonces cuando los puntos se superponen, se oscurece la zona.



# Colores

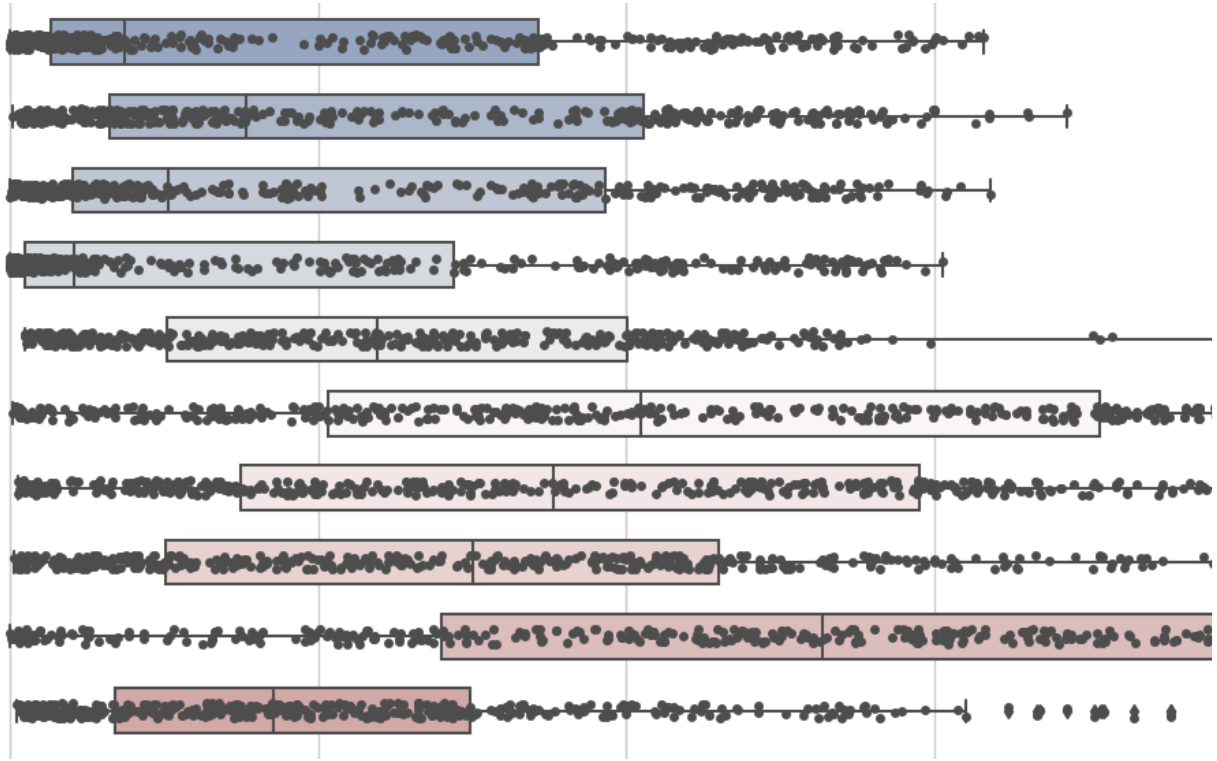
- **Colores que no se utilizan.**



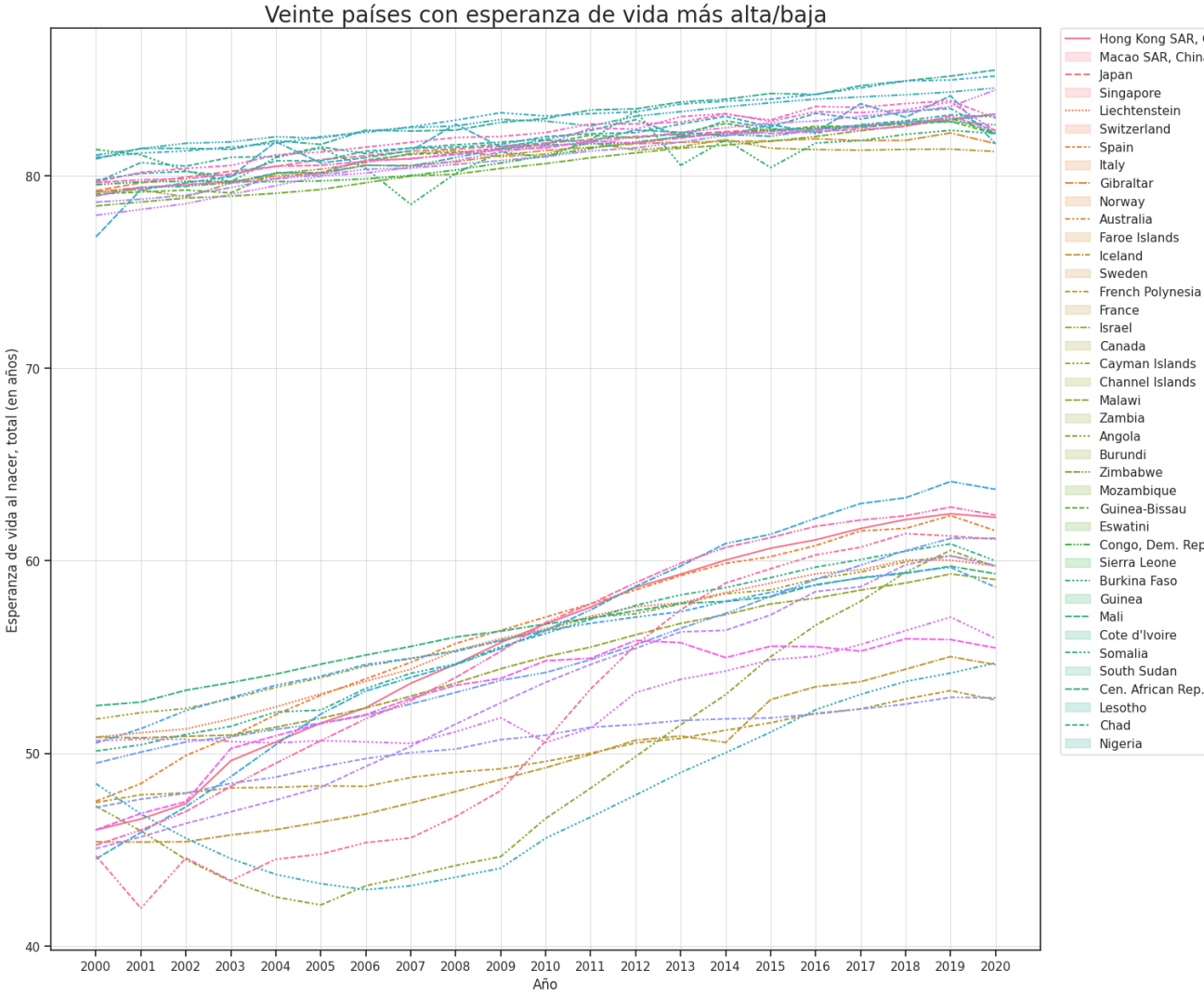

- **Los colores significan diferentes cosas según el gráfico.**

# Colores

- **Gama de colores secuencial sin significado.**
  - Quizá poner paleta cualitativa como set  1



- **Apreciar los límites de la percepción.**
  - Muchos datos dificultan la percepción.



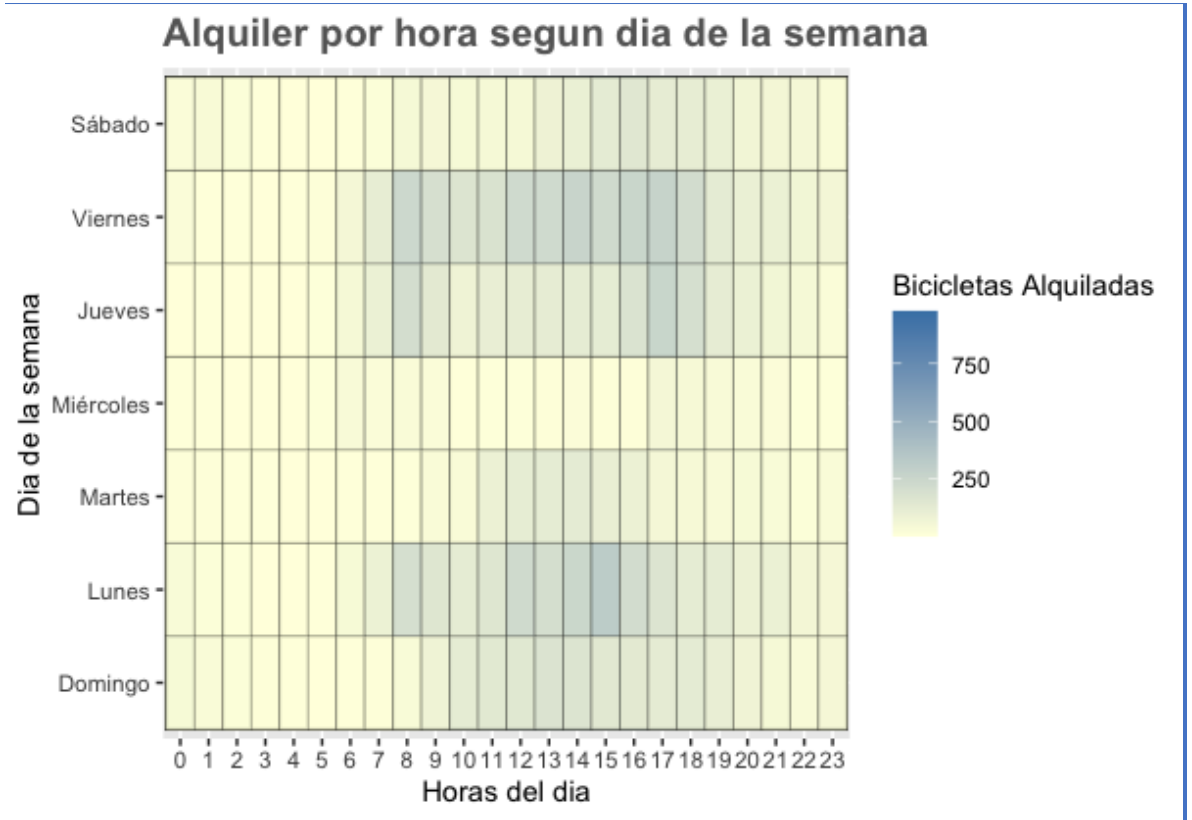

Nota: Los nombres de países en la leyenda fueron ordenados por valor de esperanza de vida decreciente

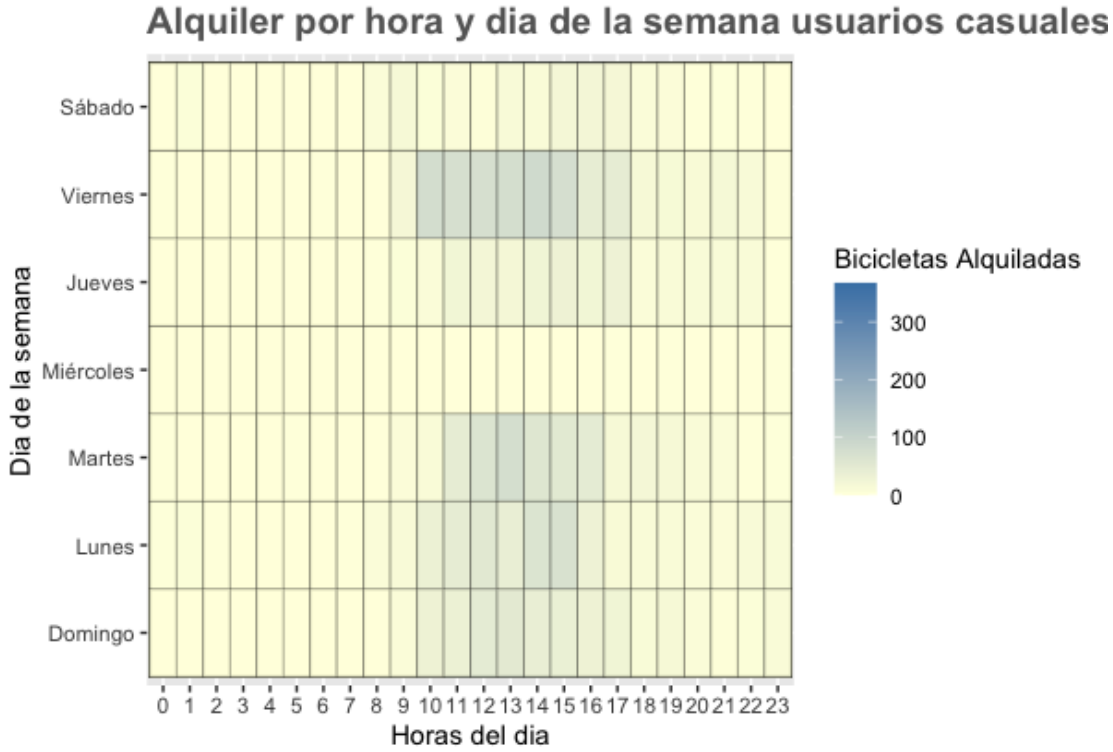
# Colores

- **Color maps con escalas diferentes.**
  - Si queremos comparar gráficas conviene uniformizar.
  - Si las escalas en las gráficas son muy diferentes conviene manejarlas de forma independiente.

**. Paleta de colores corrida.**

- Nunca se llega al color del extremo superior.

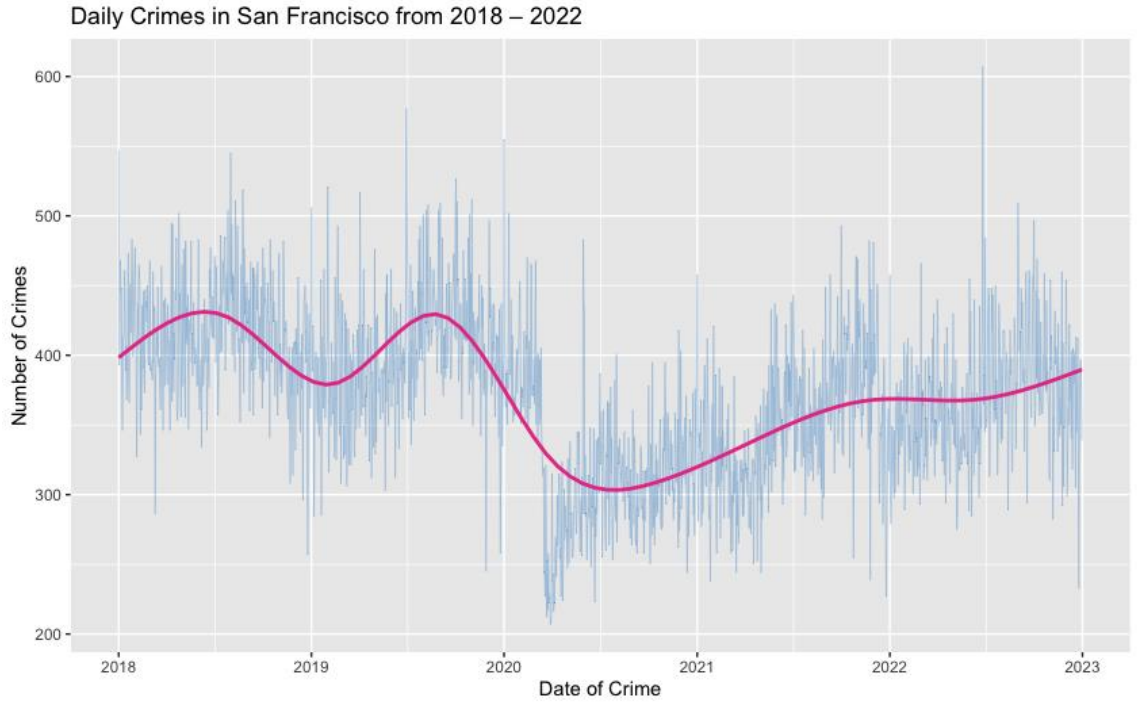


# Colores

**. El color jerarquiza los datos.**

- Importa la interpolación, pero también los datos.
- Colores fuertes para las línea interpolada y pastel para los datos/líneas individuales, donde importa más la distribución global y no cada valor individual.




**. Colores complementados con números.**

- El color da la impresión general.
- La precisión la da cada número.