FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

Para estudiantes de Educación Básica y Media.
UNIVERSIDAD DE CHILE

Clase IV: Análisis estadístico y visualización de datos

FM849 Proyecto de Ciencia de Datos: Inteligencia Artificial (IA) y sus aplicaciones

Máximo Flores Valenzuela (mflores@dcc.uchile.cl)

Universidad de Chile • 16 de agosto de 2025

Roadmap de la clase

1. Medidas de tendencia central, dispersión, y otros valores estadísticos.
2. Tablas de contingencia para el análisis de variables categóricas.
3. Visualización de información.

Medidas de tendencia central, dispersión, y otros valores estadísticos

Estadísticas de resumen

Corresponden a valores que describen propiedades de los datos, p. ej., frecuencias, medidas de dispersión y de tendencia central.

- Las medidas de dispersión miden la variabilidad en los datos. Acá tenemos la desviación estándar (σ), el rango ($R = \max(X) - \min(X)$), etc.

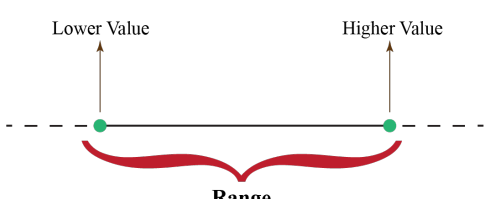

Figura 3: Visualización del rango.

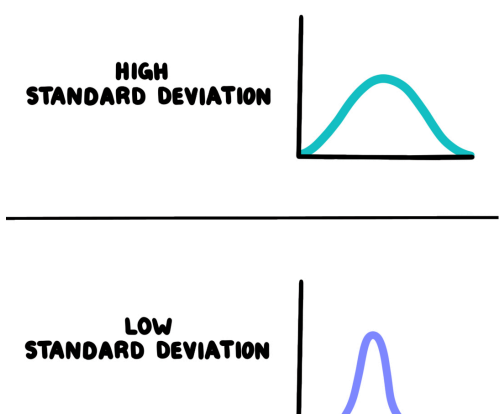

Figura 4: Visualización de la desviación estándar σ.

Estadísticas de resumen

- Por otro lado, las medidas de tendencia central corresponden a un «resumen» de los datos en un sólo concepto. Acá tenemos la media (promedio), mediana y moda.

$$\overline{X} = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i$$

Media ($\overline{X}$): todos los datos ponderan igual.

Figura 5: Visualización de la mediana.

Mediana: es el dato del medio después de ordenar.

Moda: es el dato que más se repite, es decir, el que tiene mayor frecuencia.

Frecuencia y moda

- La frecuencia de un valor (f) es el conteo de cuántas veces aparece.
  - La frecuencia porcentual (f_p) de un valor es la relación entre cuántas veces aparece y el total de datos: $f_p = f/N$, donde N es el total de datos.
  - Usando la librería `pandas`, podemos usar `df['col'].value_counts()` para saber las frecuencias de los valores en la columna `col`.
- La moda M es el valor que aparece más veces:

$$M = \arg\max_c f(c), \quad \text{donde } f(c) \text{ es el conteo de veces que aparece } c$$

  - En `pandas`, basta usar `df.mode()`.

! **Importante:** Estos estadísticos se suelen usar para variables de naturaleza categórica.

Tendencia central: Media o promedio $\overline{x}$

- Tiene un problema muy común: es muy sensible a valores atípicos (*outliers*).
  - Imaginemos el caso de los sueldos en Chile: sueldos muy altos desplazan el promedio hacia arriba, haciendo que parezca que los chilenos ganamos mucho más de lo que es la realidad.

Renta promedio por región

Región	Promedio noviembre 2022	Promedio noviembre 2021	Variación 2022 vs 2021
Tarapacá	$983.975	$759.134	29,62%
Atacama	$943.098	$827.365	13,99%
Antofagasta	$874.496	$912.560	-4,17%
Metropolitana	$795.268	$795.110	0,02%
La Araucanía	$719.906	$653.388	10,18%
Valparaíso	$687.164	$672.638	2,16%
Los Lagos	$687.094	$786.557	-12,65%
Biobío	$686.458	$744.051	-7,74%
Maule	$680.762	$725.564	-6,17%
O´Higgins	$675.769	$708.772	-4,66%
Coquimbo	$662.301	$669.643	-1,10%
Arica y Parinacota	$656.806	$885.027	-25,79%
Magallanes	$599.830	$735.870	-18,49%
Ñuble	$594.088	$677.077	-12,26%
Los Ríos	$554.316	$766.954	-27,73%
Aysén	$482.151	$789.319	-38,92%
País	**$766.848**	**$784.786**	**-2,29%**

Fuente: Trabajando.com. Rubros con mejores sueldos son construcción, minería y actividades

Figura 6: Tabla de renta promedio por región en Chile.

Tendencia central: Media truncada

La media truncada soluciona el problema anterior definiendo un conjunto de datos que excluye a los que están en un porcentaje inferior y superior al ordenar los datos de menor a mayor (o al revés).

- En el caso anterior, podemos eliminar el 5% de los que menos ganan, y el 5% de los que más ganan, eliminando un 10% de los datos de la muestra.

Problema: a veces los *outliers* sí cuentan una historia. No podemos eliminarlos porque sí, hay que realizar un análisis previo.

- Usando la librería `scipy`, se puede implementar de la siguiente forma para el 10% de valores extremos:

```
> from scipy import stats
> media_truncada_datos = stats.trim_mean(datos, 0.1)
```

Obviamente, el valor 0.1 puede cambiar según cuántos datos queremos eliminar.

Tendencia central: Mediana

- Como vimos, separa a la mitad inferior y superior de la muestra luego de que los datos están ordenados.
- No nos cuestionamos: ¿qué pasa si el número de datos es par? ¿Cuál es el del medio?

Figura 7: Cantidad par de datos. ¿Qué pasa con el cálculo de la mediana?

La solución es sencilla: tomamos a los dos del medio (personas 2 y 3 de izq. a der. en el ejemplo de arriba), y promediamos sus valores asociados.

- Es más robusta que la media cuando hay valores atípicos (*outliers*).
- Con la librería `numpy`, se puede calcular a través de la instrucción `np.median(datos)`.

Tendencia central: Resumen

Con lo que hemos discutido, podemos afirmar que estas medidas no necesariamente son iguales. Cada una tiene sus propiedades. Cuándo usar cada una dependerá del contexto.

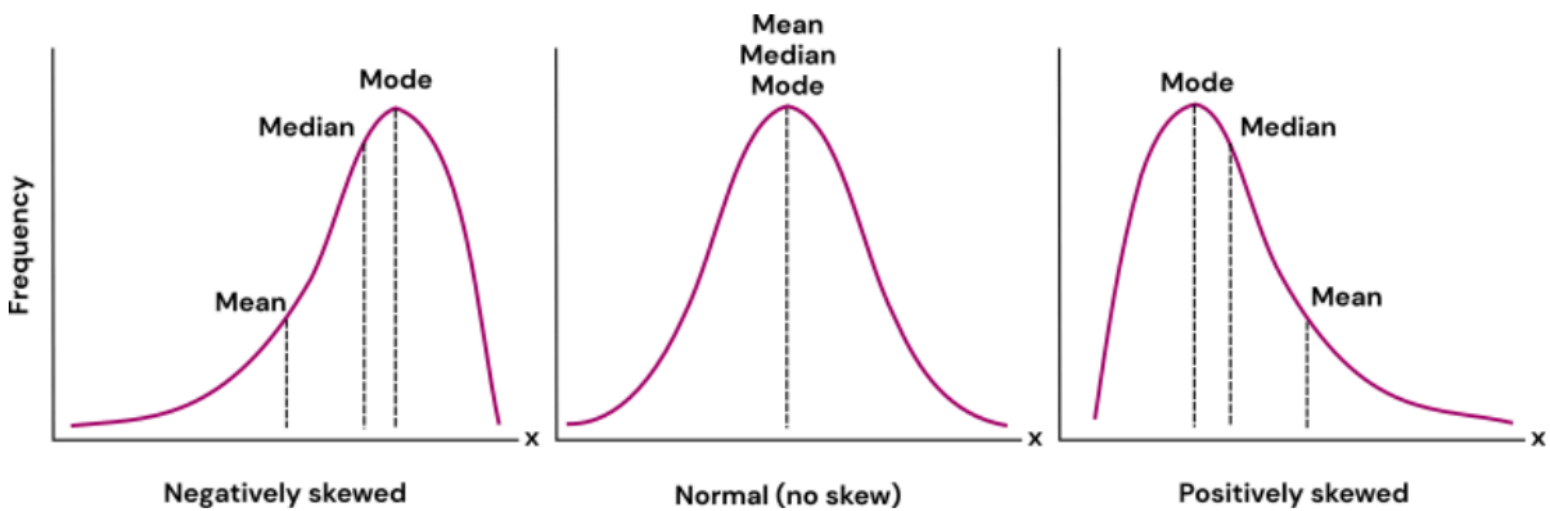

Figura 8: Diferencia entre la media, mediana y moda.

Extensión de la mediana: Percentiles

- El percentil p es el valor que separa los datos en su $p\%$ inferior y su $(100 - p)\%$ superior cuando están ordenados. Con $p = 50$ recuperamos la definición de la mediana.
- Una definición equivalente son los **cuantiles**, que usan la versión percentual. Por ejemplo, usando los valores $0.1, 0.2, \ldots, 0.9$ para el 10%, 20%, ..., 90% respectivamente.
- Con la librería `pandas`, se calculan con la función `quantile`:

```
> cuantiles = df['col'].quantile([i / 100 for i in range(101)])
```

Cuartiles

Los cuartiles son un caso particular de los percentiles. Dividen la muestra en 4 partes:

- El primer cuartil, q_1, equivale al percentil 25.
- El segundo cuartil, q_2, equivale al percentil 50, es decir, la mediana.
- El tercer y último cuartil, q_3, equivale al percentil 75.

En código, con `pandas`, incluyendo el mínimo ($p = 0$) y el máximo ($p = 100$):

```
> cuartiles = df['col'].quantile([0, 0.25,
0.5, 0.75, 1])
```

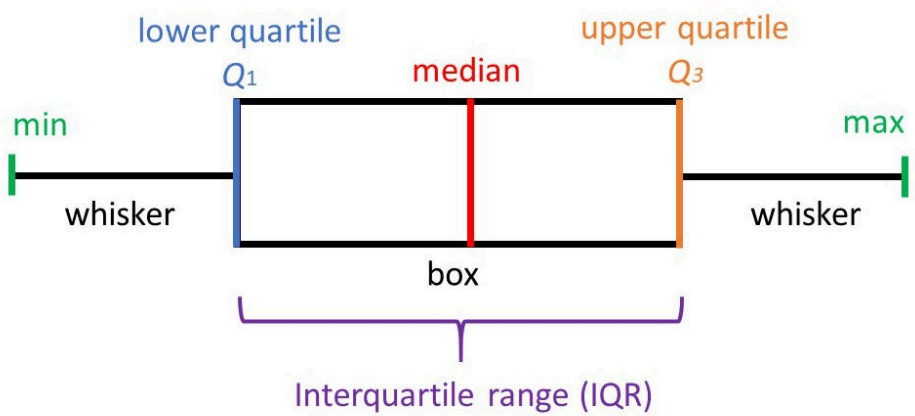

Figura 9: Visualización de los cuartiles.

Resumen de un *Data Frame*

Para resumir las estadísticas de un *Data Frame* con `pandas`, podemos usar la instrucción `describe()`. En el caso de variables numéricas, nos indica el mínimo, los cuartiles, la media, y el máximo.

```
> df.describe()
```

	Company Names	Cars Names	Engines	CC/Battery Capacity	HorsePower	Total Speed	Performance(0 - 100)KM/H	Cars Prices	Fuel Types	Seats	Torque
count	1218	1218	1218	1215	1218	1218	1212	1218	1218	1218	1217
unique	37	1201	356	311	456	114	180	535	23	19	263
top	Nissan	Mistral	I4	1984 cc	355 hp	250 km/h	6.5 sec	$35,000	Petrol	5	400 Nm
freq	149	2	64	31	23	145	45	36	871	692	72

`count` son los valores no nulos, `unique` es la cantidad de valores únicos (excluyendo nulos), `top` es la moda, y `freq` cuántas veces aparece la moda.

Este *Data Frame* lo vimos en el Tutorial II. No aparecen las estadísticas de variables numéricas porque las transformaciones de tipos de datos se hicieron posteriormente.

Medidas de dispersión

Las medidas de dispersión responden a la pregunta: ¿qué tan dispersos son mis datos?

- Podemos hablar de qué tan separados están entre sí, o qué tantos valores abarcan.
- Generalmente, el valor fijado para establecer el nivel de separación es una medida de tendencia central, como el promedio.

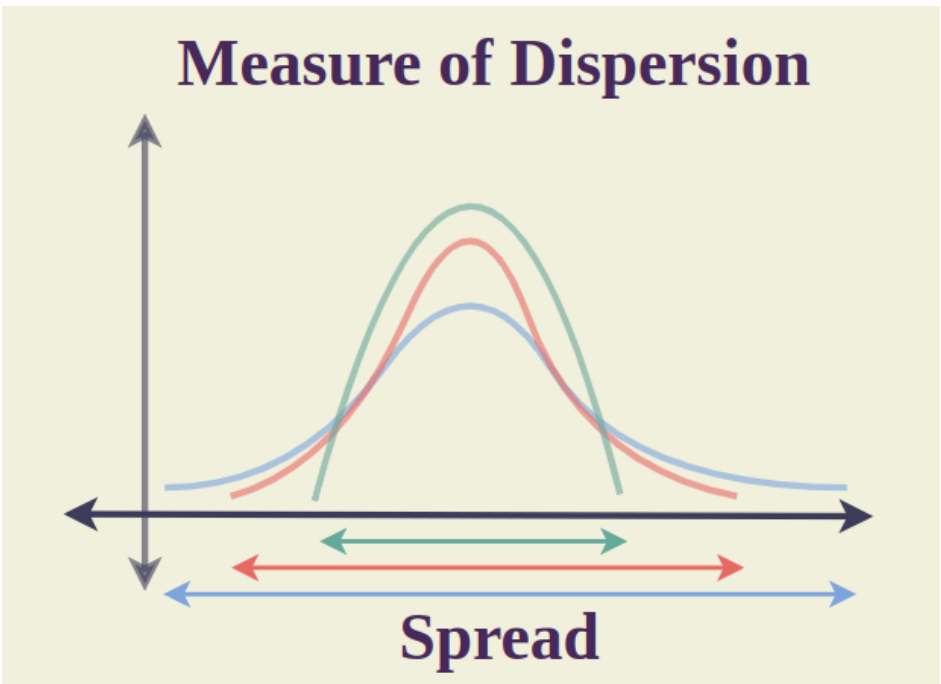

Figura 10: Dispersión de distintas distribuciones de datos.

Dispersión: Rango

El rango es la distancia de Manhattan entre los valores extremos:

$$\text{rango}(X) = |\max(X) - \min(X)| = \max(X) - \min(X)$$

En Python, usando `pandas`, podemos usar las instrucciones `max()` y `min()`, y restar los valores resultantes:

```
> df["col"].max() - df["col"].min()
```

Dispersión: Varianza σ^2 y desviación estándar σ

La desviación estándar corresponde a la raíz cuadrada de la varianza. La varianza mide el «promedio» de las diferencias cuadráticas de las observaciones con respecto a la media:

$$\mathbb{V}\text{ar}(X) = \frac{1}{N-1} \cdot \sum_{i=1}^{N} (x_i - \bar{x})^2; \quad \sigma(X) = \sqrt{\mathbb{V}\text{ar}(X)}$$

La razón principal por la cual no se ocupa solamente la varianza es porque al sacar la raíz cuadrada quedan resultados con unidades interpretables (p. ej., pasando de cm^2 a cm).

En Python, usando `pandas`, se ocupan las instrucciones `var()` y `std()` para su cálculo:

```
> varianza = df["col"].var()
> desv_estandar = df["col"].std()
```

Dispersión: Varianza σ^2 y desviación estándar σ

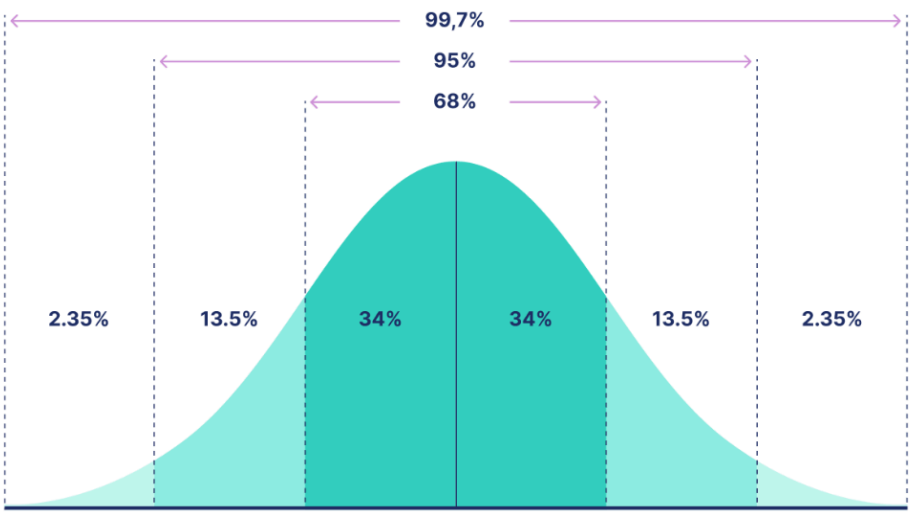

Figura 11: Efecto de la desviación estándar σ en una distribución normal.

Recién, dijimos que la varianza mide el «promedio» entre comillas. La razón de esto es porque el promedio real resulta de dividir la suma por N y no por $N - 1$.

El motivo de ocupar $N - 1$ es porque la media está fija en la fórmula, entonces se quita un grado de libertad. Esta modificación se llama corrección de Bessel y genera un estimador insesgado para σ^2.

Dispersión: Alternativa a la varianza

Una alternativa a la varianza ocupa cualquier medida de tendencia central $m(X)$ de $X = (x_1, ..., x_N)$ (usualmente la mediana) y se llama **desviación absoluta promedio** (AAD):

$$\text{AAD}(X) = \frac{1}{N} \cdot \sum_{i=1}^{N} |x_i - m(X)|$$

Observación: En Python, el valor absoluto se calcula con la instrucción `abs()`:

```
> abs(7 - 15)
>>> 8
```

Dispersión: Desviación mediana absoluta

Una medida de desviación que está basada directamente en la mediana se llama **desviación mediana absoluta** (MAD), que ocupa una constante de escala C (usualmente 1.4826 para que sea consistente con la desviación estándar):

$$\text{MAD}(X) = C \cdot \text{median}(|x_i - \text{median}(X)|)$$

El término $\text{median}(|x_i - \text{median}(X)|)$ se calcula generando una nueva muestra $X' = (x'_1, ..., x'_N)$ donde cada x'_i es el valor absoluto de la resta del dato original x_i con la mediana original de X. Posteriormente, se calcula la mediana de la nueva muestra X'.

En Python, con la librería `scipy`, se puede calcular de la siguiente forma:

```
> from scipy import stats
> stats.median_abs_deviation(df["col"])
```

Dispersión: Rango intercuartil

El rango intercuartil se calcula usando la misma lógica del rango normal, salvo que ahora se excluye el 25% menor y el 25% mayor de los datos.

$$\text{IQR}(X) = |q_3 - q_1| = q_3 - q_1$$

En Python, también se puede usar `pandas` para calcularlo:

```
> df["col"].quantile(0.75) - df["col"].quantile(0.25)
```

Estadística multivariada: Covarianza

En ocasiones, queremos ver cómo varía una variable con respecto a otra. Para medir el grado de variación lineal conjunta de un par de variables se usa la covarianza:

$$\mathbb{C}\text{ov}(X, Y) = \frac{1}{N-1} \cdot \sum_{i=1}^{N} (x_i - \bar{x}) \cdot (y_i - \bar{y}) \in \mathbb{R}$$

Notar que $\mathbb{C}\text{ov}(X, X) = \mathbb{V}\text{ar}(X) = \sigma^2$.

En Python, usando `pandas`, se calcula con la instrucción `X.cov(Y)`, donde `X` e `Y` son características:

```
> df["col_1"].cov(df["col_2"])
```

- Si dos variables son independientes, entonces su covarianza es cero. **Al revés** no necesariamente es cierto (siempre indica que no hay relación lineal).
- Si la covarianza entre dos variables es menor que 0, entonces tienden a moverse linealmente en direcciones opuestas. Si es mayor que 0, se mueven linealmente en la misma dirección.

Estadística multivariada: Matriz de covarianzas

Cuando tenemos un *dataset* con muchas columnas numéricas, podemos calcular una matriz de covarianzas (denotada Σ).

- Cada celda Σ_{ij} de esta matriz contiene la covarianza entre los atributos i y j.

$$\Sigma = \begin{pmatrix} \mathbb{C}\text{ov}(X_1, X_1) & \mathbb{C}\text{ov}(X_1, X_2) & \dots & \mathbb{C}\text{ov}(X_1, X_N) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}\text{ov}(X_N, X_1) & \mathbb{C}\text{ov}(X_N, X_2) & \dots & \mathbb{C}\text{ov}(X_N, X_N) \end{pmatrix}$$

- La matriz Σ es simétrica, es decir, lo que está por sobre la diagonal es lo mismo que está por debajo. Esto es porque $\mathbb{C}\text{ov}(X, Y) = \mathbb{C}\text{ov}(Y, X)$.

En Python, usando `pandas`, esta matriz se puede calcular aplicando la instrucción `cov()` a un *Data Frame* de variables numéricas:

```
> df.cov(numeric_only=True)
```

Estadística multivariada: Correlación de Pearson

La covarianza sufre de un problema de interpretabilidad: cuando las escalas entre variables son muy distintas, los coeficientes de la matriz Σ también son dispares.

Para solventar esto, se introduce el coeficiente de correlación de Pearson (ρ) como:

$$\rho(X, Y) = \frac{\mathbb{C}\text{ov}(X, Y)}{\sigma(X) \cdot \sigma(Y)} \in [-1, 1]$$

- Un valor cercano a 1 indica que crecen en el mismo sentido linealmente.
- Un valor cercano a -1 indica que crecen en sentidos contrarios linealmente.
- Si $\rho(X, Y) = 0$, entonces $\mathbb{C}\text{ov}(X, Y) = 0$, por lo tanto, las variables no tienen una relación lineal.

En Python, usando `pandas`, la matriz de coeficientes de Pearson se puede calcular con `corr()`:

```
> df.corr(method="pearson", numeric_only=True)
```

Tablas de contingencia para el análisis de variables categóricas

Tablas de contingencia

Las variables de tipo numérico no siempre van a ser las únicas a las cuales nos enfrentemos. También existen las variables categóricas.

Supongamos que tenemos los siguientes datos, con satisfacción baja/media/alta y grupos A/B:

Grupo	Satisfacción
A	Alta
B	Baja
B	Media
A	Media
A	Baja
B	Baja

Tabla 1: Datos categóricos sobre nivel de satisfacción según grupo.

Tablas de contingencia

En la Tabla 1, se puede ver que no hay nada de tipo numérico. Son sólo categorías. Una tabla de contingencia permite resumir estas características con datos de frecuencias de aparición:

Grupo / Satisfacción	Baja	Media	Alta
A	1	1	1
B	2	1	0

Tabla 2: Tabla de contingencia para datos categóricos sobre nivel de satisfacción según grupo.

En Python, usando `pandas`, se puede calcular usando la instrucción `crosstab()`:

```
> import pandas as pd
> ...
> df['grupo'] = pd.Categorical(df['grupo'])
> df['satisfaccion'] = pd.Categorical(df['satisfaccion'])
> tabla_contingencia = pd.crosstab(df['grupo'], df['satisfaccion'])
```

Visualización de información

Tipos de *datasets*: Tablas

En general, existen varios tipos de *datasets*: tablas, redes, árboles, datos geométricos, etc. En este curso, por alcance, sólo nos enfocaremos en las **tablas**.

- Una tabla es una colección de elementos que puede ser pensada como una matriz.
- Cada columna es una variable, que puede ser numérica o categórica, y cada fila es una observación. Una celda puntual es un valor asociado a la característica de una observación.
- Las operaciones fundamentales que se pueden hacer son: filtrado, agrupación y agregación.

¿Por qué visualizar información?

Perfectamente podríamos graficar porque sí, pero el sentido de las visualizaciones radica en un motivo profundo: **contar una historia**.

Antes de pensar en generar una visualización, tenemos que pensar en las tareas que tiene como trasfondo. Una tarea se define como una acción seguida de un objetivo.

En este curso, nos centraremos en la tarea de analizar. En específico, se busca presentar datos.

Recomendado: `@chartosaur` *(TikTok).*

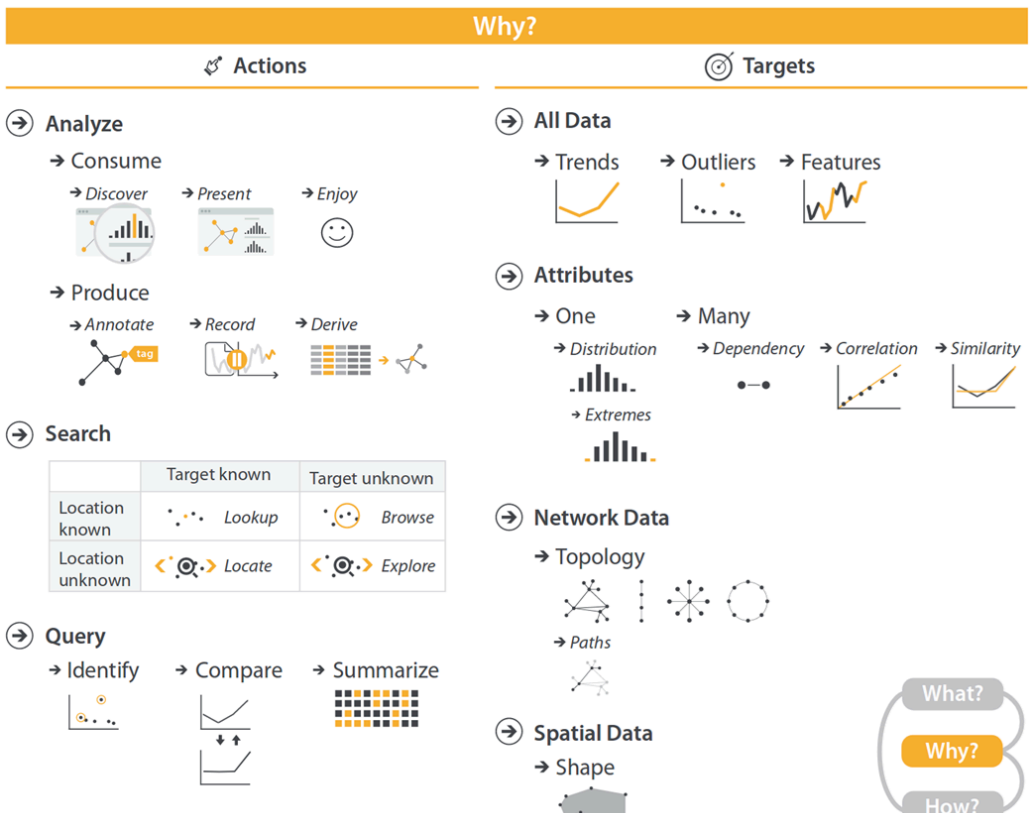

Gráfico de dispersión (*scatter plot*)

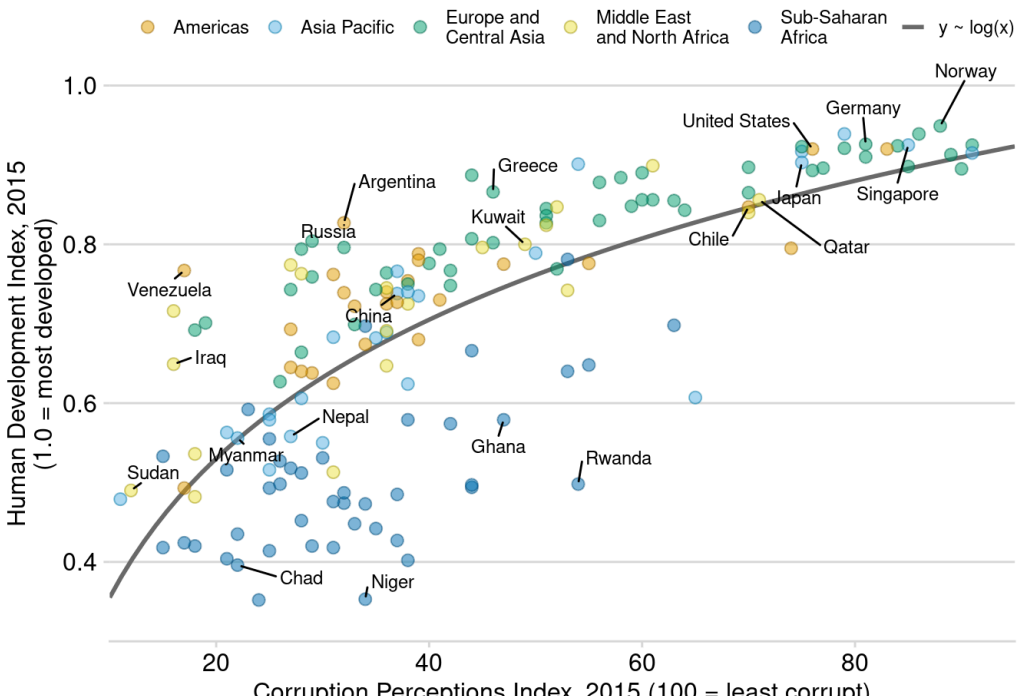

Figura 13: Gráfico de dispersión del índice de percepción de corrupción vs. el índice de desarrollo humano.

- Expresa valores cuantitativos.
- Requiere 2 variables de esta naturaleza.
- La marca son los puntos.
- Los canales son las posiciones: horizontal y vertical.

Se puede usar para encontrar patrones en los datos (qué relación siguen, si hay *outliers*, hay *clusters*, etc.)

Gráfico de líneas (*line chart*)

- Requiere llaves y valores, al igual que los diccionarios que vimos las clases pasadas.
- Las marcas son puntos que se conectan por líneas.
- La posición vertical expresa un valor cuantitativo, mientras la posición horizontal contiene las llaves ordenadas.

Se puede usar para encontrar tendencias y patrones. Las llaves deben estar ordenadas para que el gráfico tenga sentido.

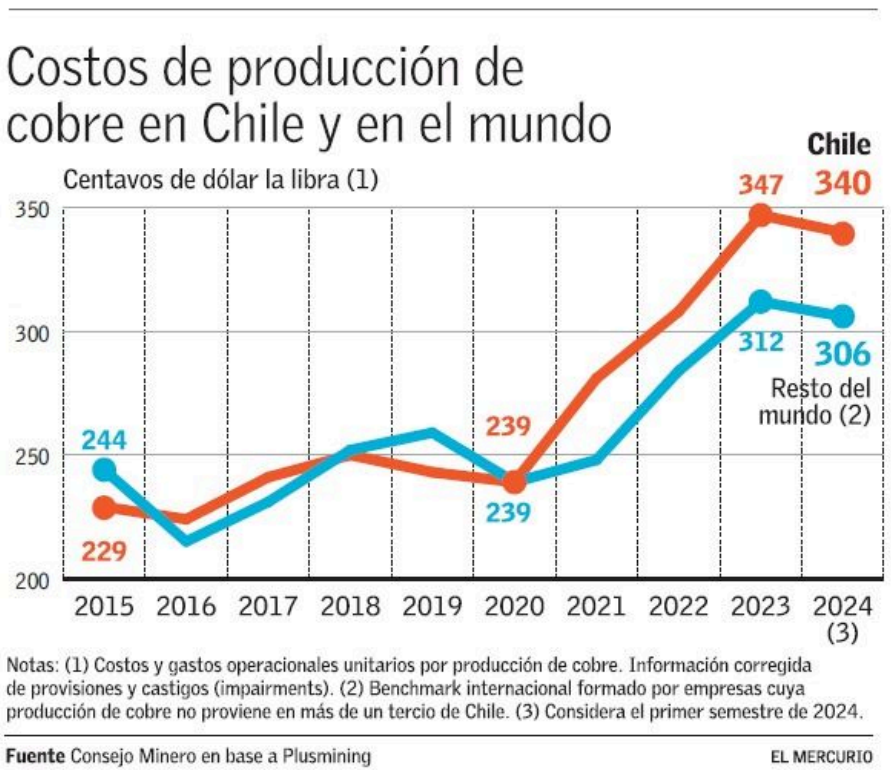

Figura 14: Gráfico de líneas de los costos de producción de cobre en Chile vs. el mundo.

Gráfico de barras (*bar chart*)

- Requiere 1 atributo categórico y 1 cuantitativo.
- Las marcas son barras.
- Los canales que codifican información incluyen: el largo de la barra para expresar un valor cuantitativo y una separación en el espacio para representar otra categoría.

Se puede usar para comparar categorías y encontrar casos extremos.

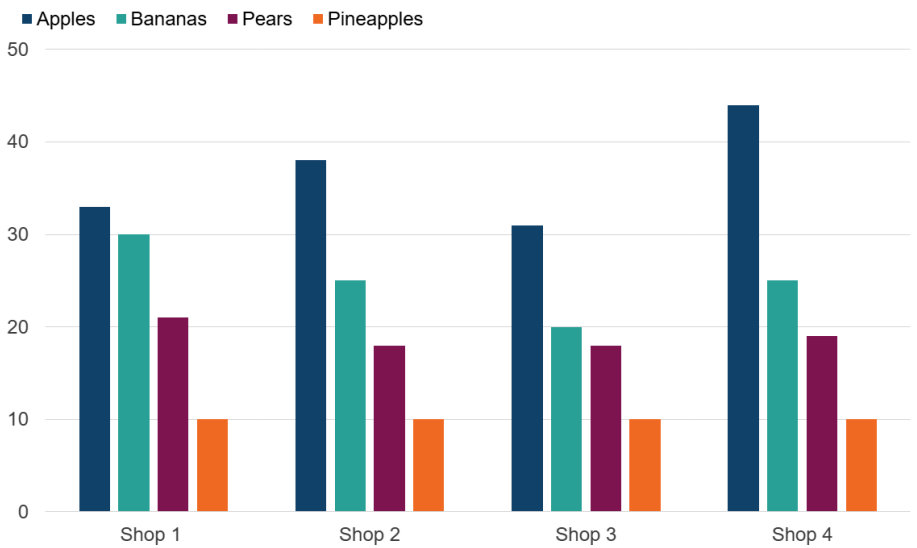

Figura 15: Gráfico de barras para los artículos vendidos en diferentes tiendas.

Gráfico de barras apilado (*stacked bar chart*)

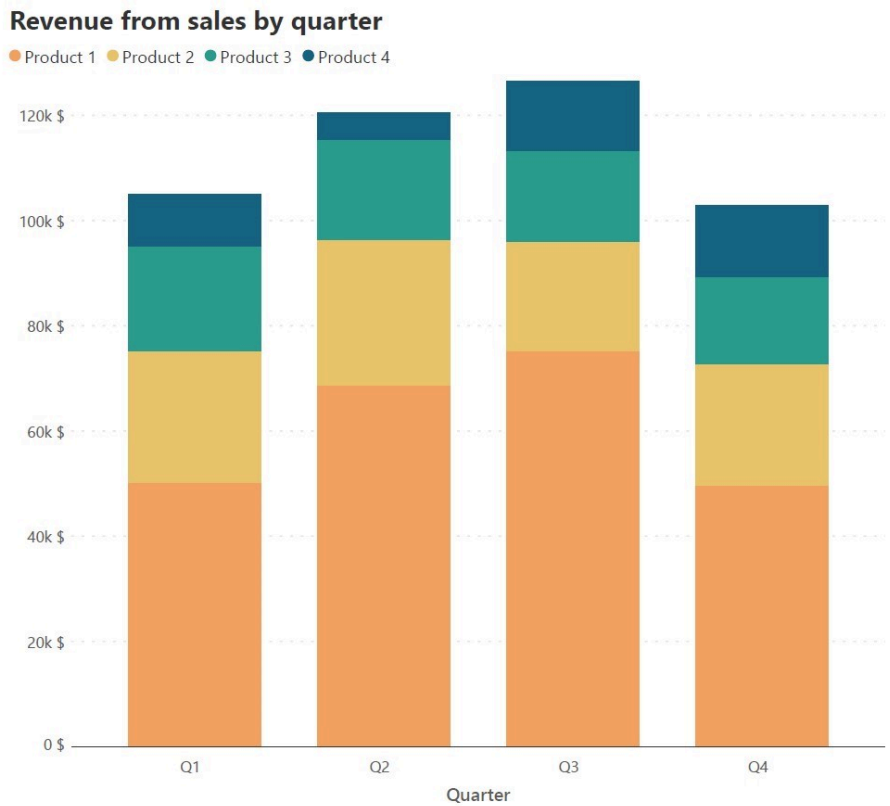

Figura 16: Gráfico de barras apilado sobre la ganancia en ventas por trimestre según producto.

- Requiere 2 atributos categóricos (llave + categorías de apilación) y 1 atributo cuantitativo.
- La marca es una pila vertical de múltiples barras.
- Los canales de codificación son el largo de cada barra, el tono de color, y las regiones espaciales que separan llaves.
- Sólo la barra inferior se alinea, las demás no.

Se puede usar para comparar categorías según su dominancia, dada una llave.

Normalización de barras apiladas (*normalized stacked bar chart*)

Una variante del gráfico anterior se obtiene normalizando las barras apiladas.

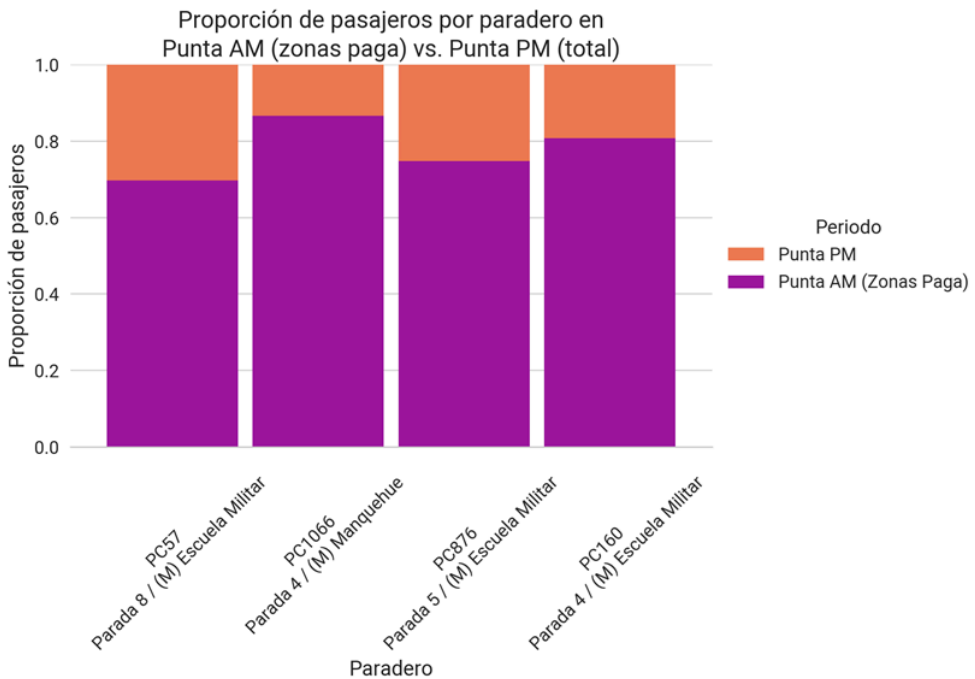

Figura 17: Proporción de pasajeros por paradero más frecuentado en el periodo Punta AM (zonas paga) y Punta PM (total) en el transporte de buses Red en la Región Metropolitana.

Histograma

Un histograma corresponde a un gráfico de barras con atributos categóricos ordinales, donde cada categoría abarca **un rango de datos**.

El rango puede mostrarse o no de manera explícita. Si no se hace, el inicio y el fin de la barra deben tener claramente los límites del rango.

Como aproximan la distribución de los datos, se recomienda que no haya separación entre las barras.

Figura 18: Gráfico de barras de estudiantes por nivel educacional e histograma de puntajes de una prueba.

Diagrama de caja [y bigotes] (*box plot*)

- Requiere 1 atributo categórico y 1 atributo cuantitativo.
- La marca es la caja, y opcionalmente puntos que pueden representar mejor la distribución.
- El principal canal de codificación es el largo de la caja (rango intercuartil) y de los bigotes (asociados a *outliers*).
- La línea sólida es la mediana, y la línea discontinua, la media.

Se usa para comparar las distribuciones de diversas categorías en función de sus cuartiles y *outliers*.

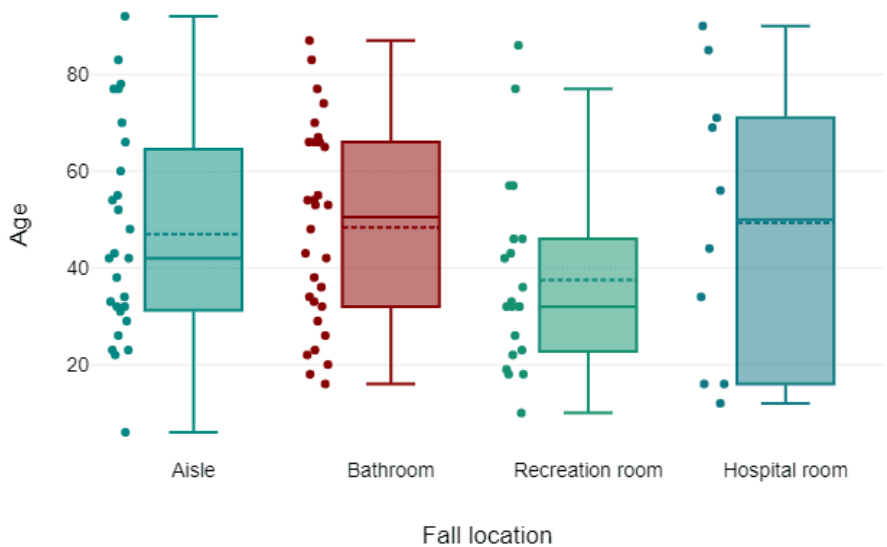

Figura 19: Diagrama de caja y bigotes de lugares de caída y edades.

Mapa de calor (*Heatmap*)

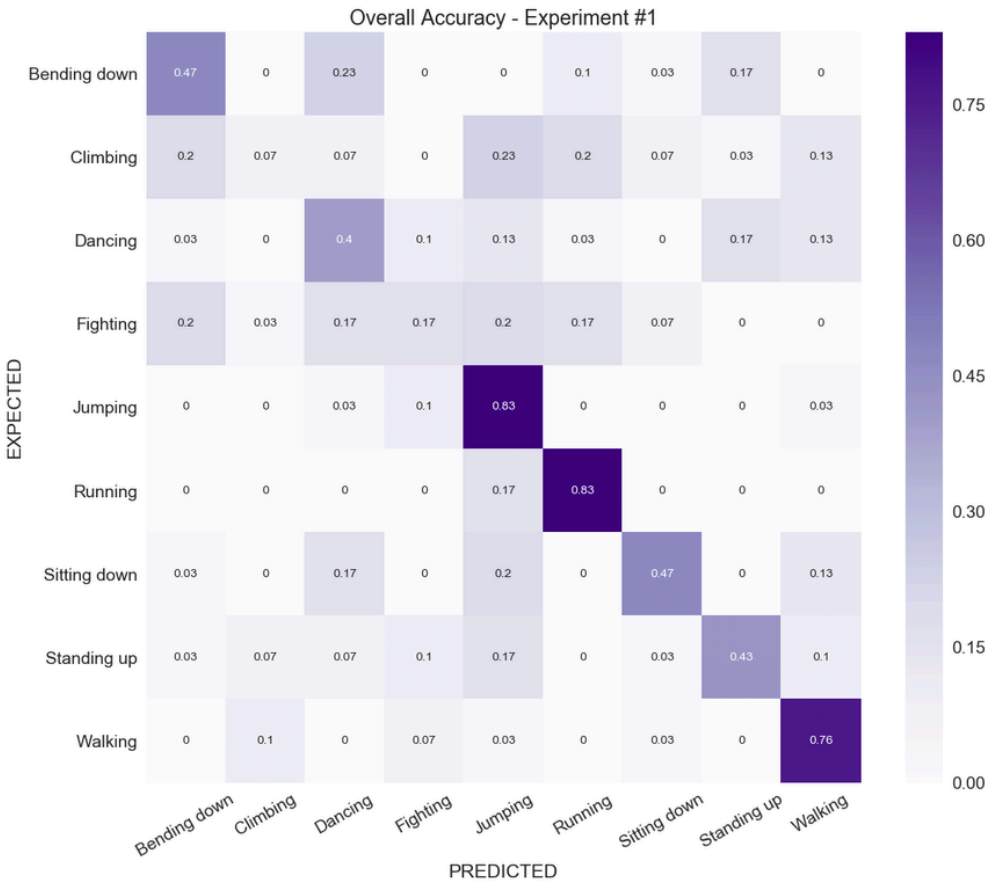

Figura 20: Matriz de confusión.

- Requiere dos llaves (atributos categóricos) y un valor (cuantitativo).
- El canal es el color determinado por el atributo cuantitativo.
- La marcas son áreas alineadas en una matriz de dos dimensiones.
- Una celda de la matriz corresponde al área que representa la intersección de dos llaves.

Se pueden usar para encontrar *clusters* o valores atípicos. También, suelen ser usados para construir las **matrices de confusión**, que evalúa los errores y aciertos de un modelo.

Pie chart y polar area chart

Son gráficos que codifican áreas con ángulos. Los ángulos y áreas son menos precisos para el ojo humano que el largo de una línea o barra, por la razón de cambio. Requieren 1 llave categórica, y 1 atributo cuantitativo. La diferencia entre ellos es que *polar area chart* codifica también con largo.

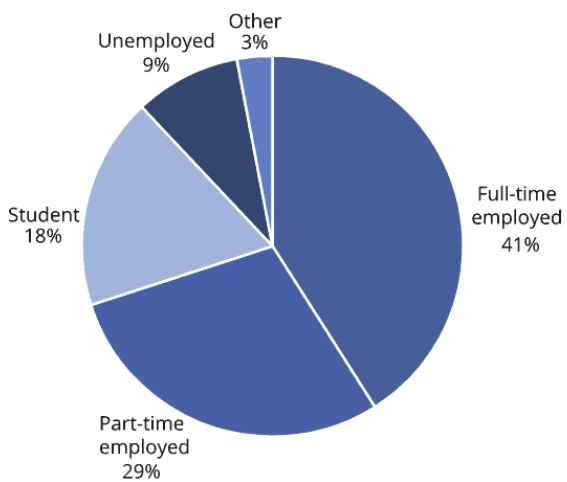

Figura 21: *Pie chart* sobre ocupación de personas.

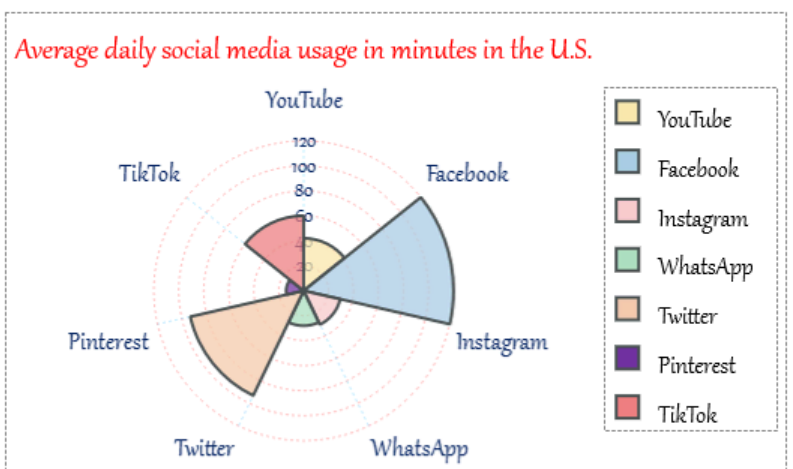

Figura 22: *Polar area chart* sobre el uso de RR. SS. en EE. UU.

Sirven para entender las proporciones y dominancias de distintas categorías con respecto a un todo.

¿Preguntas?