VI Jornadas Argentinas de Educación Estadística

Enseñanza del Análisis Exploratorio de Datos mediante planillas de cálculo

Mg. LUIS ARENAS – Mg. GUILLERMO SABINO

NOVIEMBRE 2024

Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016

¿Qué?

1. Enseñar pensamiento estadístico.
2. Enfocar en la comprensión conceptual.

¿Cómo?

3. Integrar datos reales con un contexto y un propósito.
4. Fomentar el aprendizaje activo.
5. Utilizar la tecnología para explorar conceptos y analizar datos.
6. Utilizar las evaluaciones para mejorar y evaluar el aprendizaje de los estudiantes.

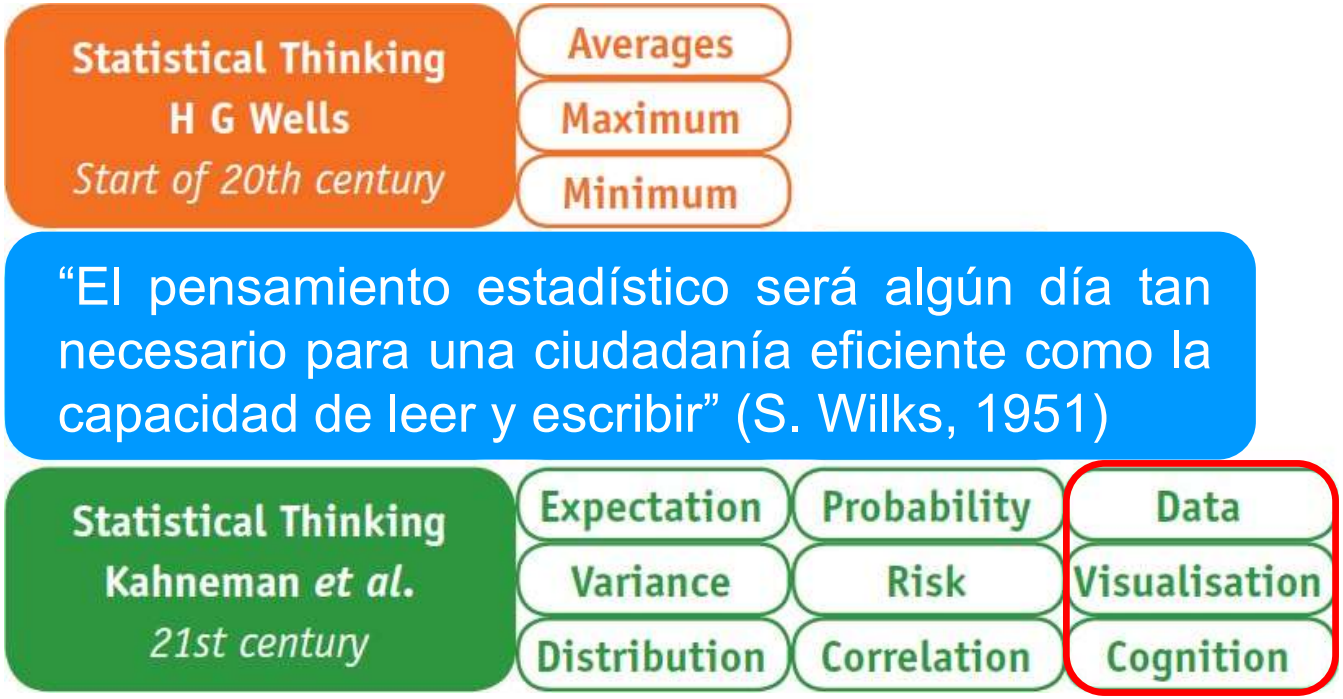
Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016

1. **Enseñar pensamiento estadístico.**
2. **Enfocar en la comprensión conceptual.**

¿Qué?

3. Integrar datos reales con un contexto y un propósito.
4. Fomentar el aprendizaje activo.
5. Utilizar la tecnología para explorar conceptos y analizar datos.
6. Utilizar las evaluaciones para mejorar y evaluar el aprendizaje de los estudiantes.

"El pensamiento estadístico es una forma de entender un mundo complejo describiéndolo en términos relativamente sencillos que, sin embargo, captan aspectos esenciales de su estructura o función, y que también nos dan una idea de lo incierto que estamos sobre ese conocimiento" (R. A. Poldrack, 2018)



(N. Marriott, 2014)

Wild y Pfannkuch (1999) caracterizan el pensamiento estadístico a partir de cuatro dimensiones:

1. Ciclo investigativo

2. Tipos de pensamiento

3. Ciclo interrogativo

4. Disposiciones

Wild y Pfannkuch (1999) caracterizan el pensamiento estadístico a partir de cuatro dimensiones:

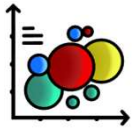
1. Ciclo investigativo
2. Tipos de pensamiento
   a. Reconocer la necesidad de los datos
   b. Transnumeración
   c. Consideración de la variación
   d. Razonamiento con modelos estadísticos
   e. Integración de la Estadística y el contexto
3. Ciclo interrogativo
4. Disposiciones

Wild y Pfannkuch (1999) caracterizan el pensamiento estadístico a partir de cuatro dimensiones:

1. Ciclo investigativo
2. Tipos de pensamiento
   a. Reconocer la necesidad de los datos
   b. Transnumeración
   c. Consideración de la variación
   d. Razonamiento con modelos estadísticos
   e. Integración de la Estadística y el contexto
3. Ciclo interrogativo
4. Disposiciones
   a. Escepticismo
   b. Imaginación
   c. Curiosidad y conciencia
   d. Apertura a ideas que cambien preconcepciones
   e. Propensión a buscar significados profundos
   f. Perseverancia y compromiso

Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016

1. Enseñar pensamiento estadístico.
2. Enfocar en la comprensión conceptual.
3. **Integrar datos reales con un contexto y un propósito.**
4. **Fomentar el aprendizaje activo.**
5. **Utilizar la tecnología para explorar conceptos y analizar datos.**
6. Utilizar las evaluaciones para mejorar y evaluar el aprendizaje de los estudiantes.



¿Qué es Análisis Exploratorio de Datos?

Teniendo en cuenta que los datos están constituidos por dos partes: la “regularidad” y las “desviaciones”.

- La regularidad indica la estructura simplificada de un conjunto de observaciones.
- Las diferencias de los datos con respecto a esta estructura representan las desviaciones o residuos de los datos.

El AED es básicamente el desglose de los datos en esas dos partes. En lugar de imponer, en hipótesis, un modelo a las observaciones, se genera dicho modelo desde las mismas.

¿Qué es Análisis Exploratorio de Datos?

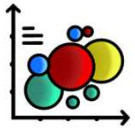
Siguiendo a Batanero et al., (1991):

Es una nueva filosofía en la aplicación de los métodos de análisis de datos.

- Consiste en el estudio de los datos desde todas las perspectivas, y con todas las herramientas posibles.
- El propósito es extraer cuánta información sea posible, generar hipótesis nuevas, en el sentido de conjeturar sobre las observaciones de las que disponemos.

Siguiendo a Pearson et al., (2018):

- Es el proceso de investigar datos de una manera organizada y cuidadosa en un esfuerzo por entender la estructura contenida en ellos.
- En ese proceso las visualizaciones gráficas juegan un rol central.



¿Cómo se piensa a los datos en el AED?

- Los datos se conciben como la materia prima fundamental para el proceso de investigación y descubrimiento y pueden incluir tanto información útil, como ruido o errores.
- Los datos brutos suelen necesitar limpieza por estar incompletos o con inconsistencias, por ello en el AED se trabaja para identificar y manejar problemas de calidad de datos, como valores faltantes, duplicados y anomalías en los datos.
- Son multivariados y obtenidos mayormente de manera no estructurada, es decir, pueden o no responder a un muestreo formal.

¿Cómo se hace el AED?

Pearson et al., (2018) indica la siguiente estrategia general:

1. Evaluar las características generales del conjunto de datos, por ejemplo:
   a. ¿cuántos registros tiene? ¿cuántas variables?
   b. ¿qué indican los nombres de cada una de las variables? ¿tienen algún significado?.
   c. ¿qué tipo de variables hay?
   d. ¿cuántos valores únicos tiene cada variable?
   e. ¿hay datos duplicados?
   f. ¿qué valor ocurre más frecuentemente y cuán a menudo ocurre?
   g. ¿hay observaciones o datos faltantes ? en ese caso, ¿cuán frecuentemente eso ocurre?

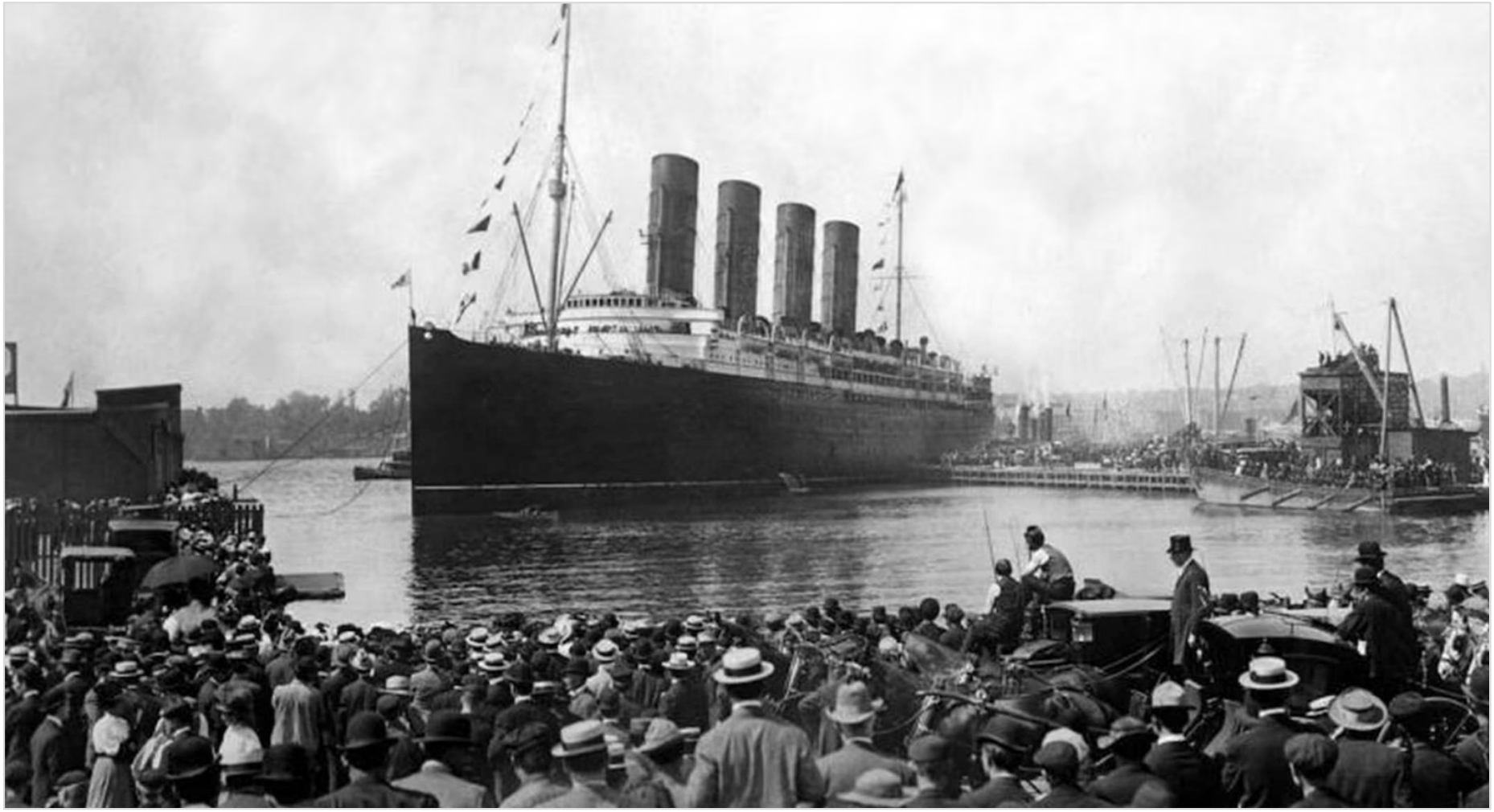
¿Cómo se hace el AED?

2. Examinar las estadísticas descriptivas de cada variable.
3. Examinar gráficas exploratorias de todas las variables de interés particular.
4. Aplicar procedimientos para buscar anomalías de los datos.
5. Buscar relaciones entre las variables clave.
6. Resumir todos los resultados anteriores en un diccionario de datos. Este documento debe incluir:
   a. Un resumen de toda la información desde el paso 1 en adelante.
   b. Metadatos, es decir, información acerca del conjunto de datos y su contenido publicado por otras fuentes.

¿Por qué AED con Microsoft Excel?

- Interfaz amigable e intuitiva: es fácil de usar lo que permite trabajar rápidamente con datos y realizar análisis básicos sin una curva de aprendizaje pronunciada.
- Capacidad de resumen y visualización rápida: ofrece gráficos y tablas dinámicas que permiten explorar tendencias, patrones y relaciones de manera visual sin tener que realizar cálculos complejos.
- Análisis estadístico básico: incluye funciones estadísticas y de manejo de datos, así como la herramienta Análisis de Datos.
- Herramientas de limpieza y manipulación de datos: permite realizar operaciones de limpieza y transformación de datos, filtrado, búsqueda y reemplazo, etc..

TITANIC



Titanic: metadatos

- Este conjunto es utilizado frecuentemente para aprender conceptos de modelado de datos, predicciones y para practicar el análisis de datos con algoritmos de clasificación, como el de predicción de supervivencia de los pasajeros.
- Fue creado recopilando información pública y registros históricos sobre los pasajeros que estaban a bordo del RMS Titanic, el transatlántico británico que se hundió tras chocar con un iceberg en 1912. Esta información incluye detalles que se documentaron después del accidente.
- La mayoría de los datos provienen de fuentes históricas, incluyendo archivos de la White Star Line, informes de la investigación británica y estadounidense del accidente, y otros registros públicos.

Análisis

1. Univariado ¡DATOS!

a) Variable Cuantitativa Discreta

b) Variable Cuantitativa Continua

2. Visualización de Datos

a) ¿Por qué graficamos los datos?

b) Gráficos erróneos

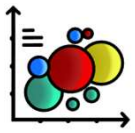
c) Percepción y visualización de datos

d) Tareas visuales y decodificación de datos

e) Dashbord básicos

   i. Variables Cualitativas (Univariado)

   ii. Bivariado



Análisis

1. Univariado
   a) Variable Cuantitativa Discreta
   b) Variable Cuantitativa Continua

a) ¿Cuánto tiempo le dedican?
b) ¿Qué tipos de gráficos ven?
c) ¿Con qué profundidad?

2. **Visualización de Datos**
   a) ¿Por qué graficamos los datos?
   b) Gráficos erróneos
   c) Percepción y visualización de datos
   d) Tareas visuales y decodificación de datos
   e) Dashbord básicos
      i. Variables Cualitativas (Univariado)
      ii. Bivariado

Análisis

1. Univariado
a) Variable Cuantitativa Discreta
b) Variable Cuantitativa Continua

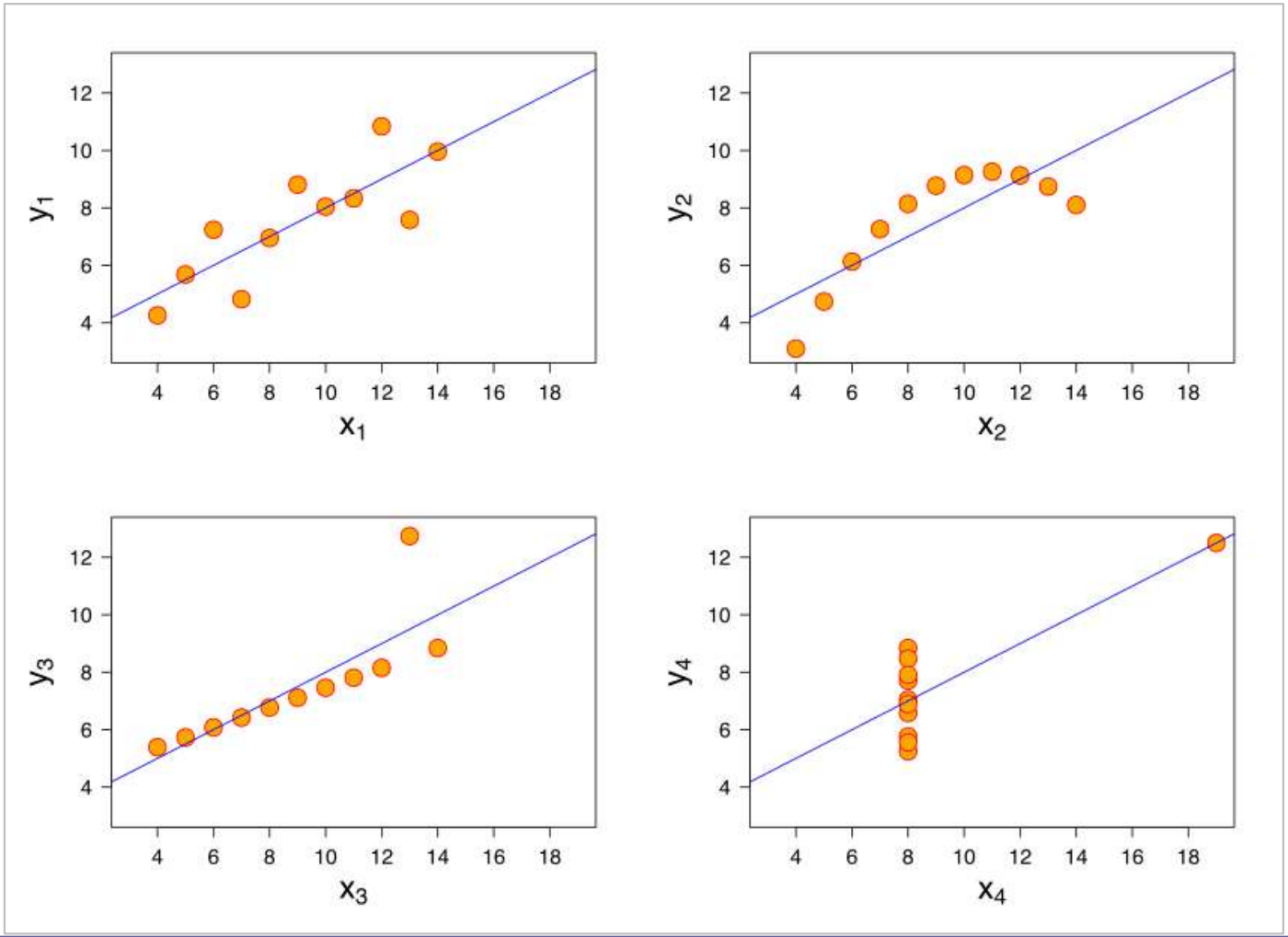
2. Visualización de Datos
a) ¿Por qué graficamos los datos?
b) Gráficos erróneos
c) Percepción y visualización de datos
d) Tareas visuales y decodificación de datos
e) Dashbord básicos
   i. Variables Cualitativas (Univariado)
   ii. Bivariado

Encontrando patrones

Se puede apreciar como las medidas de resumen estadísticas coinciden, sin embargo, sus representaciones visuales marcan claramente las diferencias entre cada conjunto de datos.

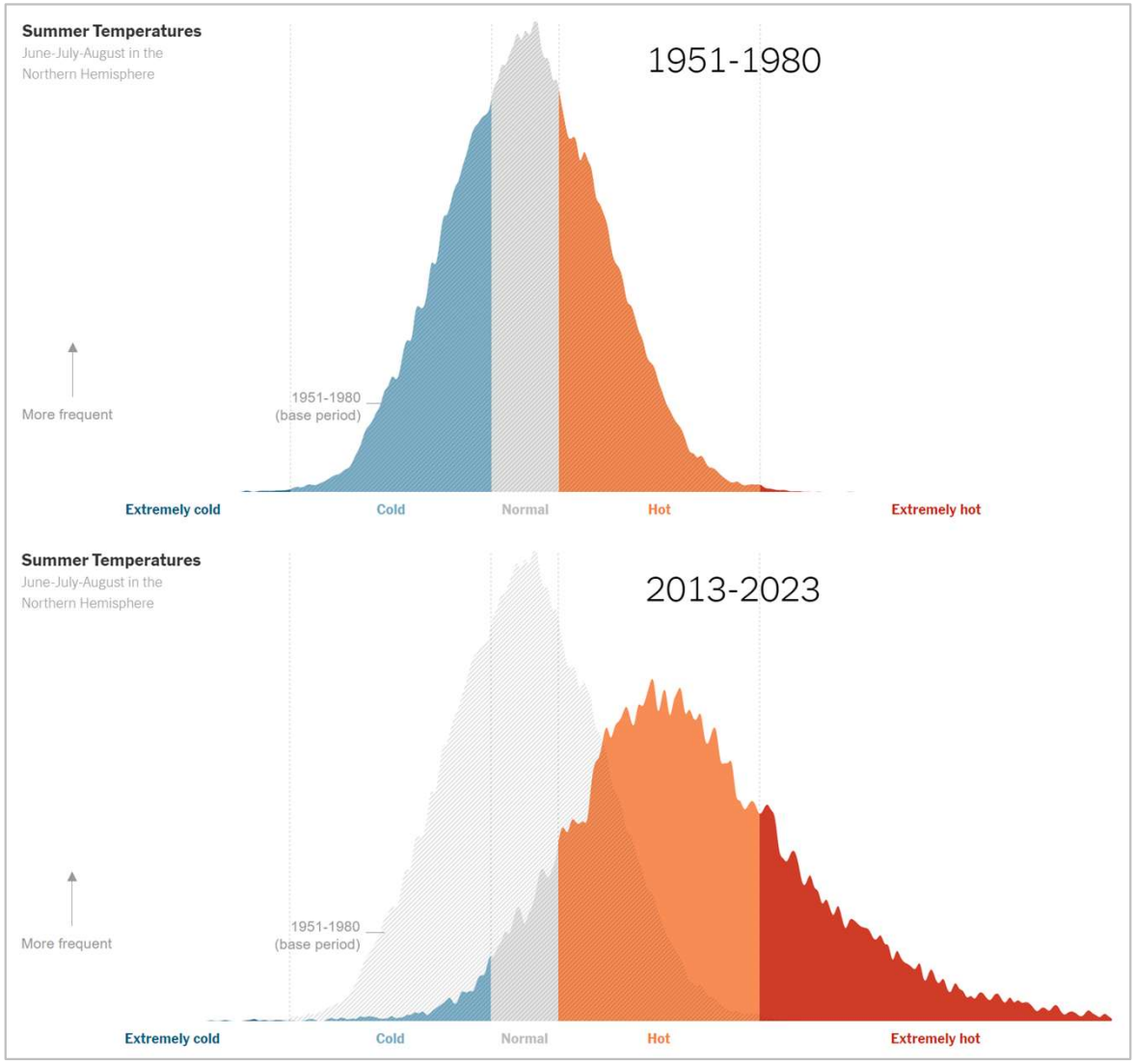
Dataset I		Dataset II		Dataset III		Dataset IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Estadísticas	Valor
media de x	9
Varianza de x	11
media de y	7,5
varianza de y	4,125
correlacion	0,816
regresion lineal	y = 3.00 + 0.500x
R2	0,67



Transmitir información

La representación gráfica nos permite de un solo vistazo condensar información que de otra manera llevaría muchísimo tiempo y un desarrollo extenso.



Análisis

1. Univariado

a) Variable Cuantitativa Discreta

b) Variable Cuantitativa Continua

2. Visualización de Datos

a) ¿Por qué graficamos los datos?

b) Gráficos erróneos

c) Percepción y visualización de datos

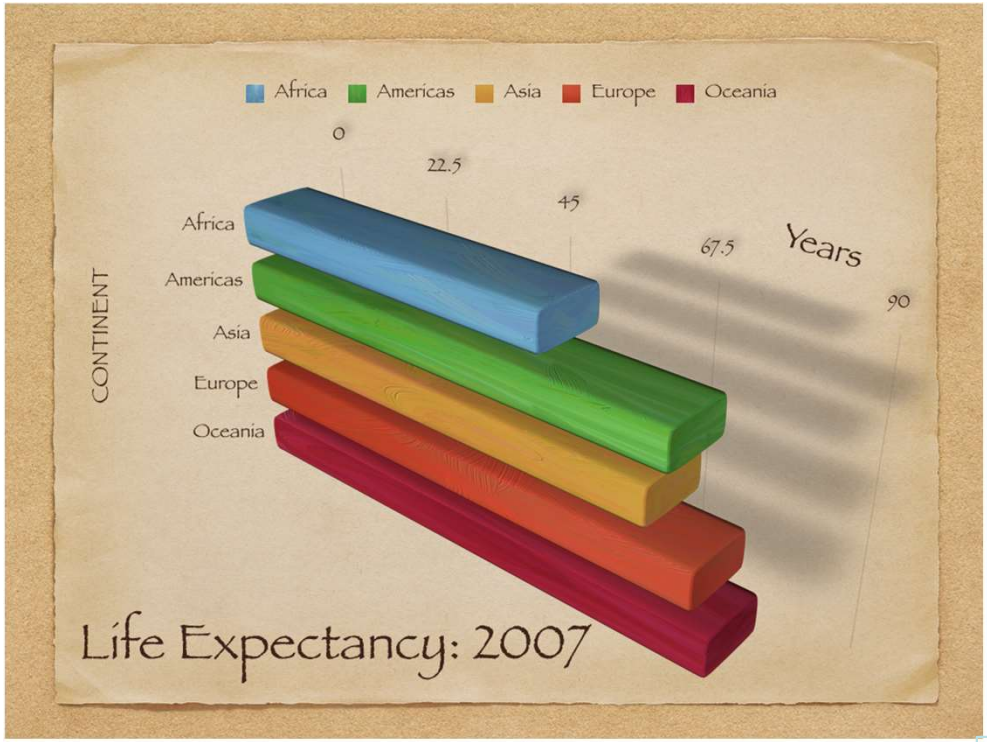
d) Tareas visuales y decodificación de datos

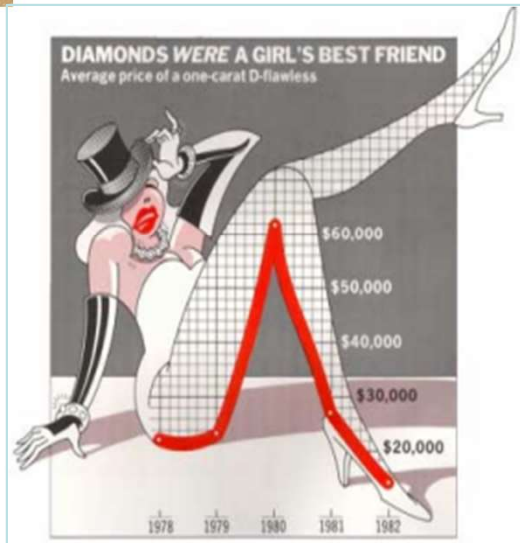
e) Dashbord básicos

   i. Variables Cualitativas (Univariado)

   ii. Bivariado

1. Mal gusto

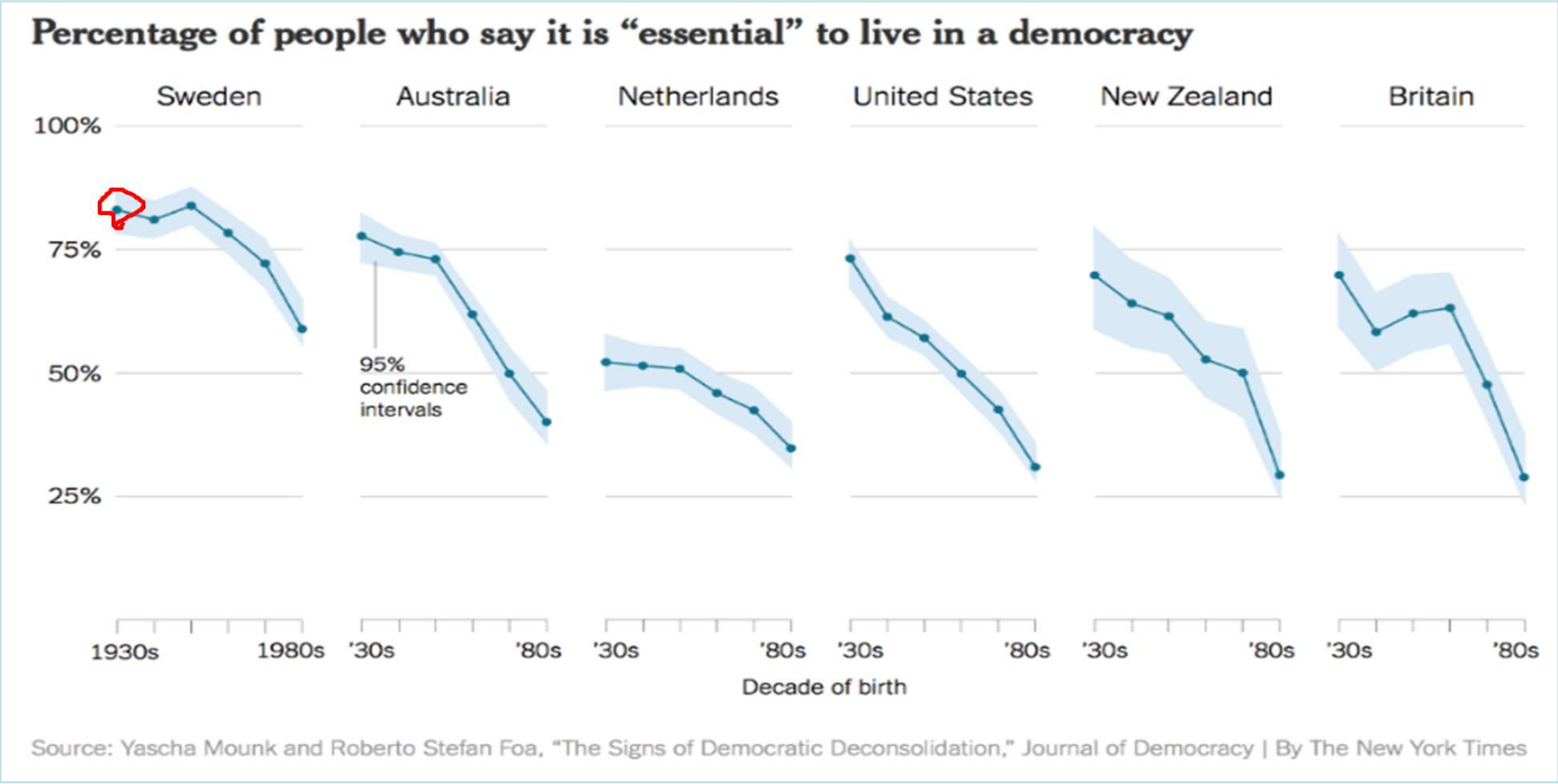

Africa
Americas
Asia
Europe
Oceania
0
22.5
45
67.5
90
Years
CONTINENT
Life Expectancy: 2007


DIAMONDS WERE A GIRL'S BEST FRIEND
Average price of a one-carat D-flawless
$60,000
$50,000
$40,000
$30,000
$20,000
1978
1979
1980
1981
1982


DIAMONDS WERE A GIRL'S BEST FRIEND
Average price of a one-carat D-flawless
$60,000
$50,000
$40,000
$30,000
$20,000
1978
1979
1980
1981
1982

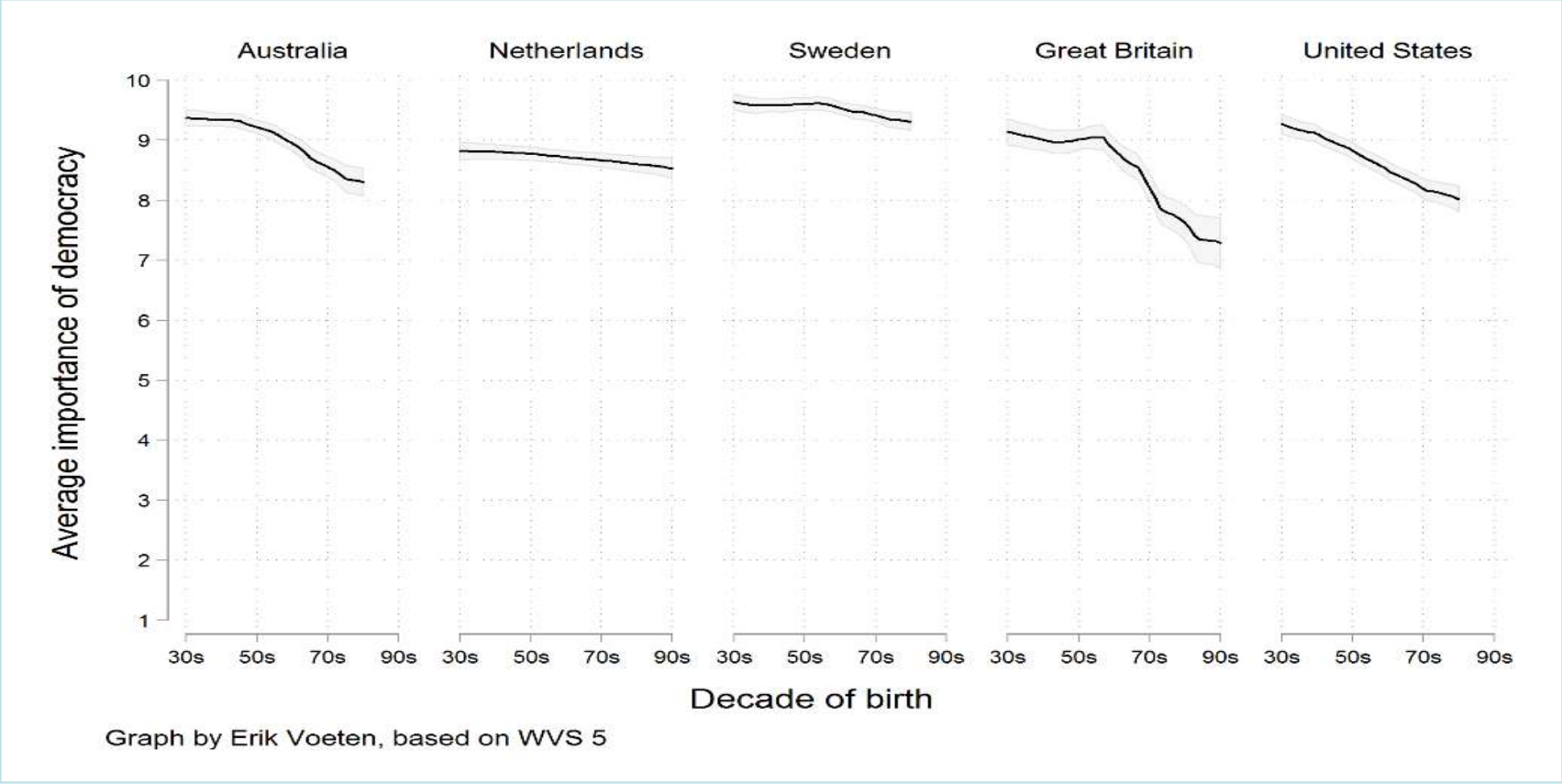
2. Malos datos

En noviembre de 2016, el diario *The New York Times* (TNYT) informó sobre una investigación sobre la confianza de la gente en las instituciones de la democracia. Había sido publicado en una revista académica por el politólogo Yasha Mounk. La pregunta que se les hizo a todos los entrevistados del estudio en un momento dado fue que calificaran la importancia de vivir en una democracia en una escala de diez puntos, donde 1 era "Nada importante" y 10 era "Absolutamente importante".

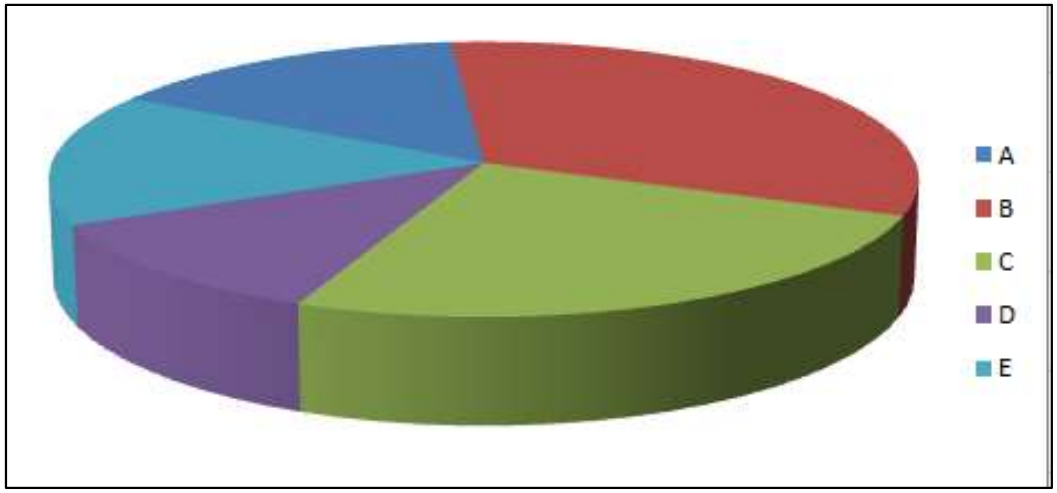


2. Malos datos

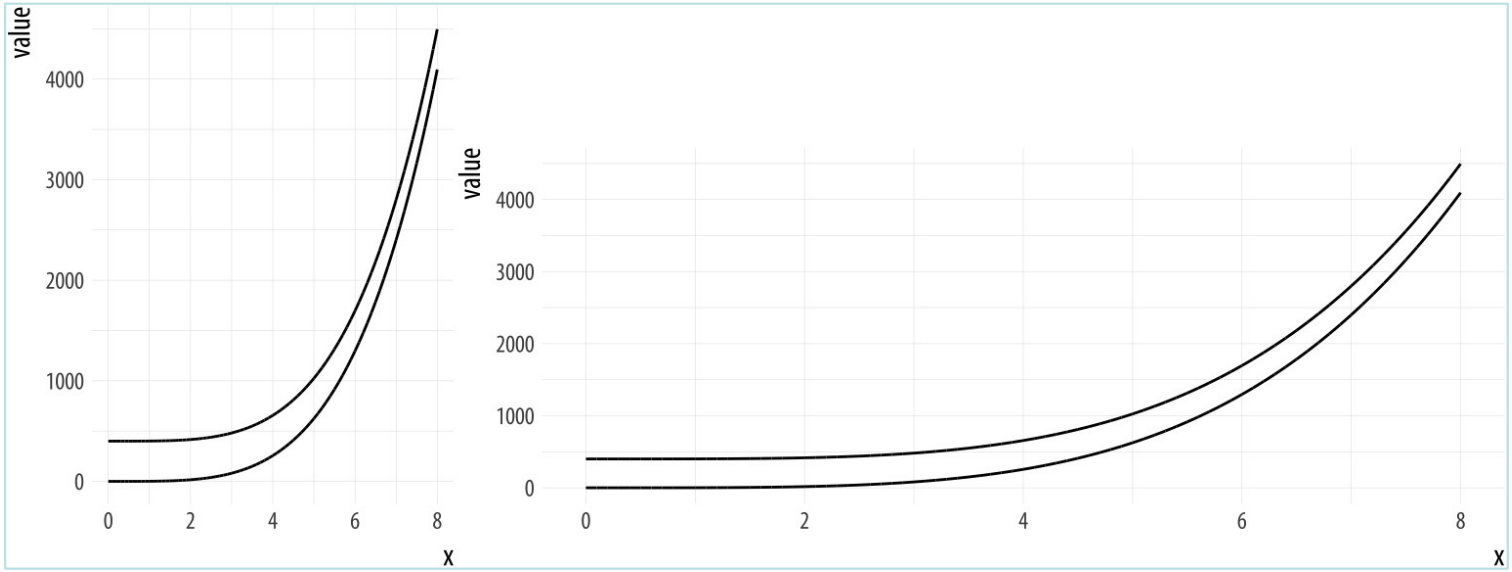
En noviembre de 2016, el diario *The New York Times* (TNYT) informó sobre una investigación sobre la confianza de la gente en las instituciones de la democracia. Había sido publicado en una revista académica por el politólogo Yasha Mounk. La pregunta que se les hizo a todos los entrevistados del estudio en un momento dado fue que calificaran la importancia de vivir en una democracia en una escala de diez puntos, donde 1 era "Nada importante" y 10 era "Absolutamente importante".



3. Mala percepción



Banco	Intereses
A	15,3%
B	32,5%
C	25,1%
D	11,6%
E	15,5%

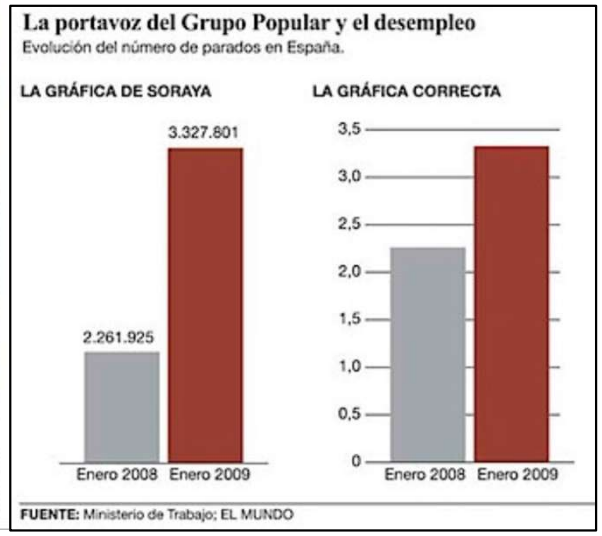


3. Mala percepción

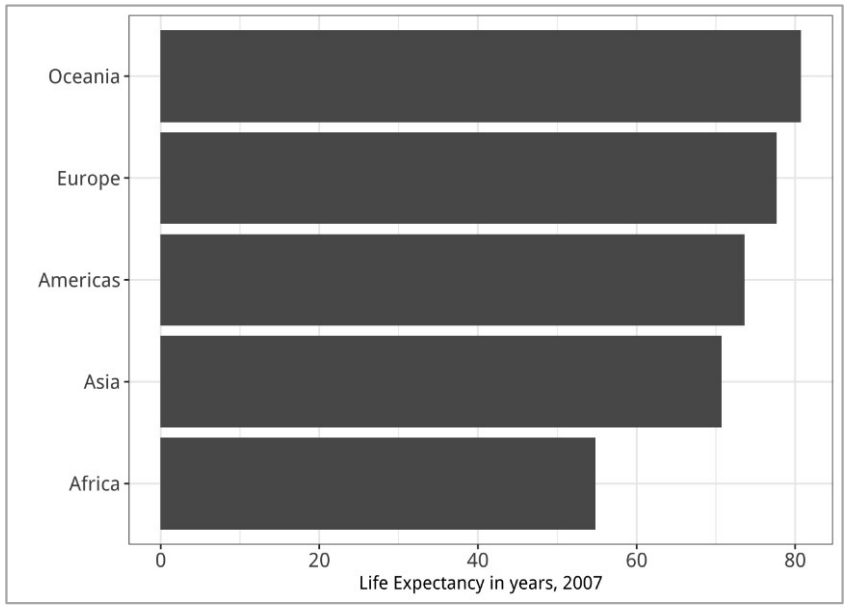
EMILIO NARANJO/EFE

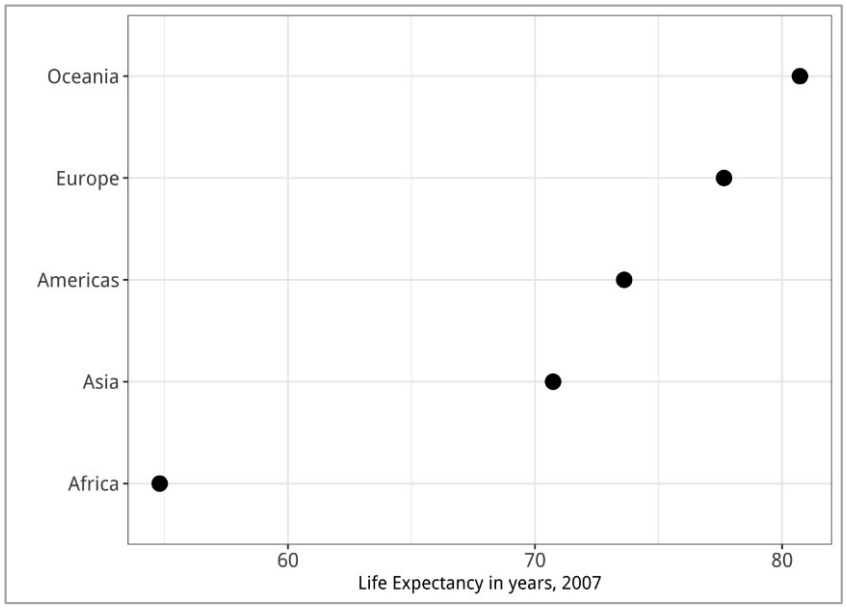
La foto de Soraya que no querría ver el Gobierno

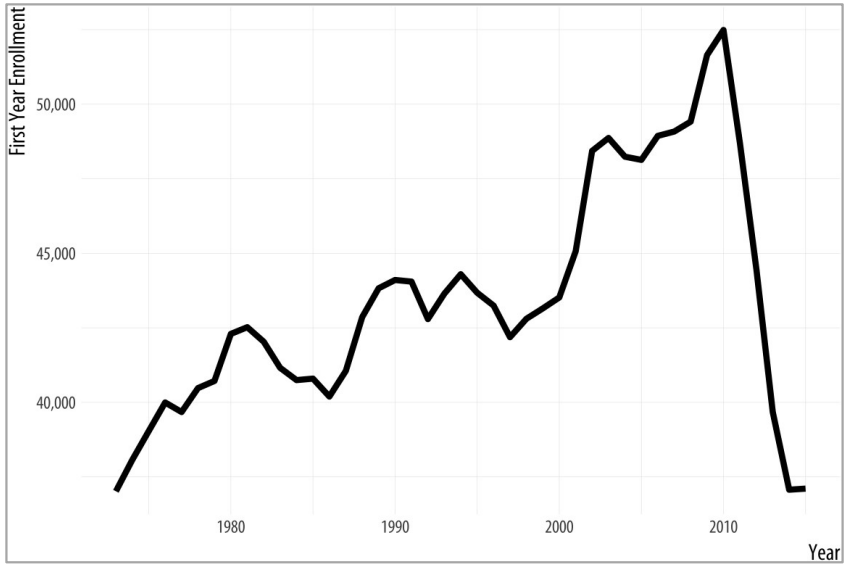
La portavoz del Grupo Popular, Soraya Sáenz de Santamaría, vuelve a ser foto de portada. Pero esta vez su imagen –gráfico en mano–, lejos de cualquier pose, representa un duro reproche al presidente del Gobierno: los 1.065.876 parados registrados en el último año, «6.600 diarios» durante el mes de enero. A su juicio, este incremento constituye «la radiografía de su promesa de pleno empleo». La dirigente *popular* reclamó con firmeza a Zapatero que «se deje de palabrerías» y «solo» que el paro entre sus prioridades». Página 6

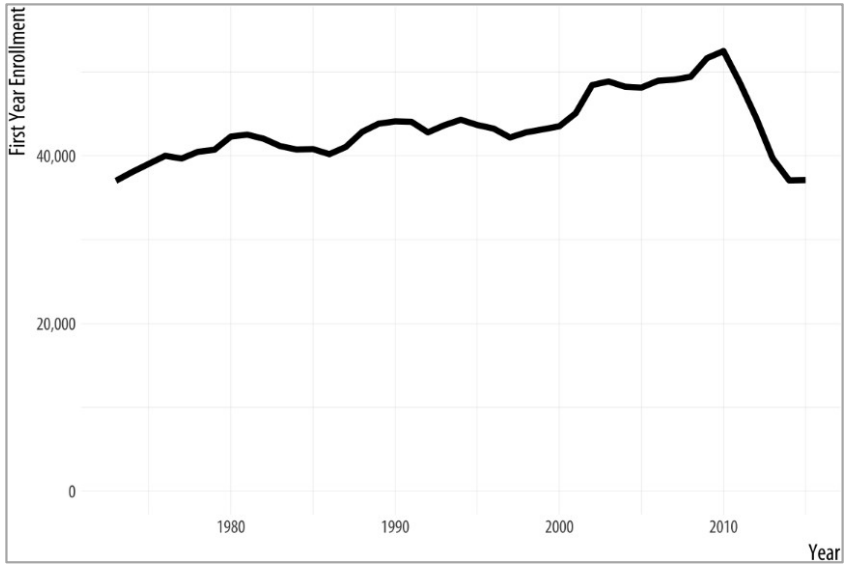


4. Problemas de honestidad y buen juicio









Análisis

1. Univariado

a) Variable Cuantitativa Discreta

b) Variable Cuantitativa Continua

2. Visualización de Datos

a) ¿Por qué graficamos los datos?

b) Gráficos erróneos

c) Percepción y visualización de datos

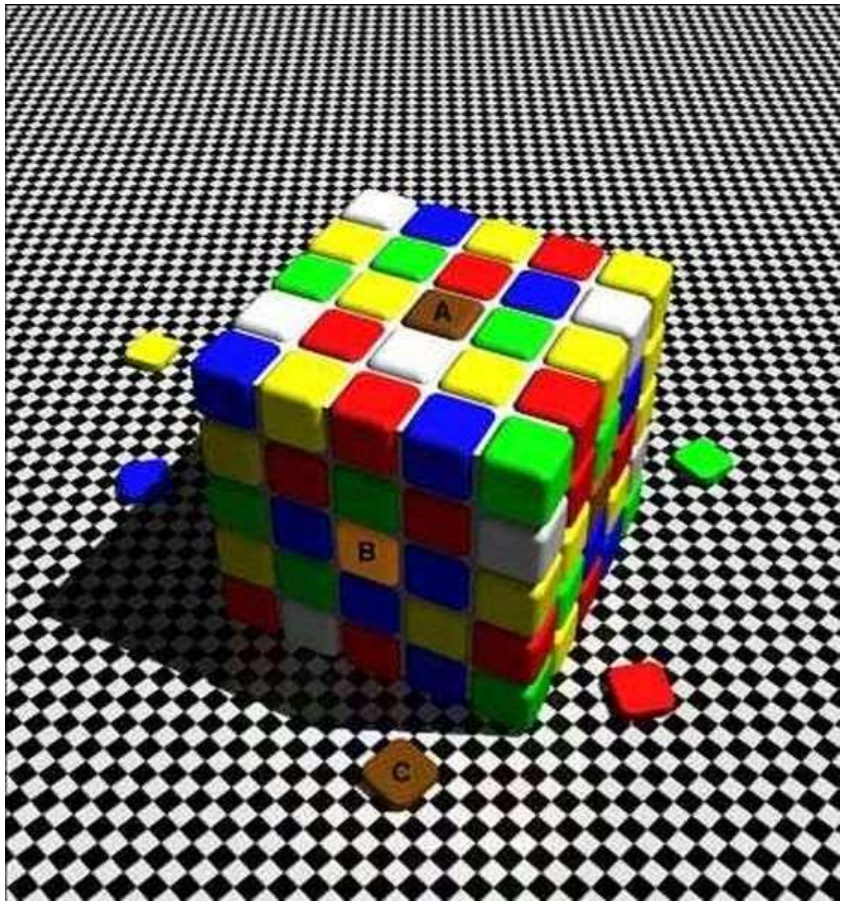
d) Tareas visuales y decodificación de datos

e) Dashbord básicos

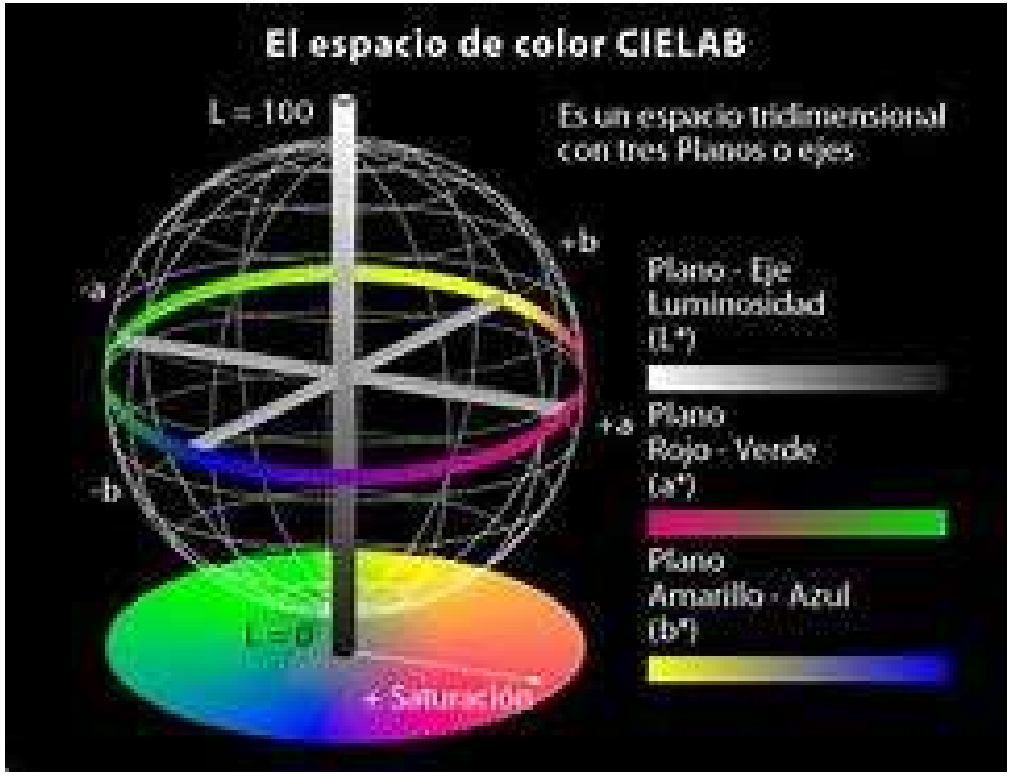
   i. Variables Cualitativas (Univariado)

   ii. Bivariado

1. Contrastes y Contextos



2. Colores

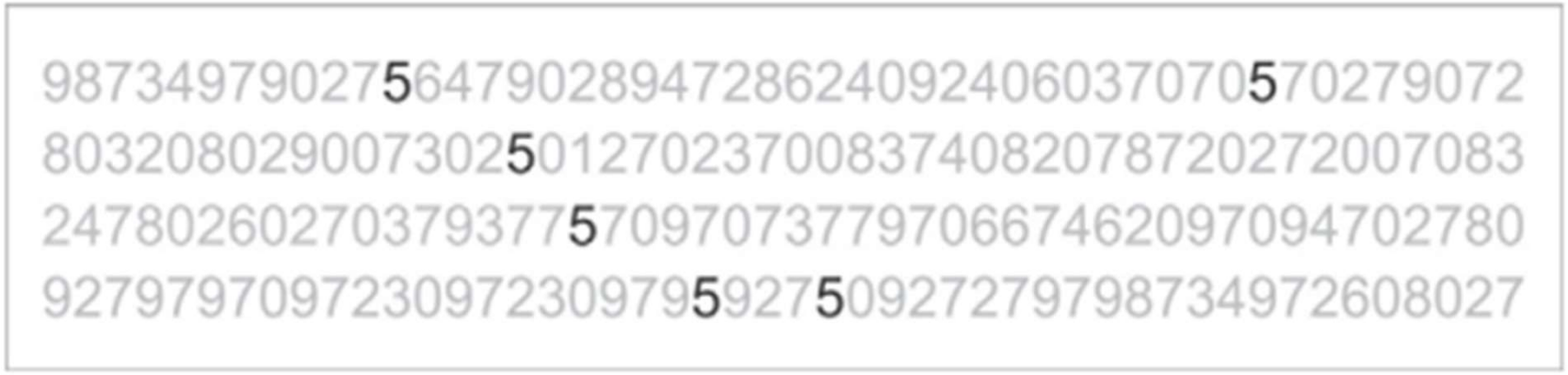

El espacio de color CIELAB
L = 100
Es un espacio tridimensional con tres Planos o ejes
+b
-a
+a
-b
Plano - Eje Luminosidad (L*)
Plano Roja - Verde (a*)
Plano Amarillo - Azul (b*)
L=0
+ Saturación


Sequential Grayscale
Sequential Blue to Gray
Sequential Terrain
Diverging
Unordered Hues

3. Preatencional y lo que destaca

Tarea: Contar cuantos números “5” hay

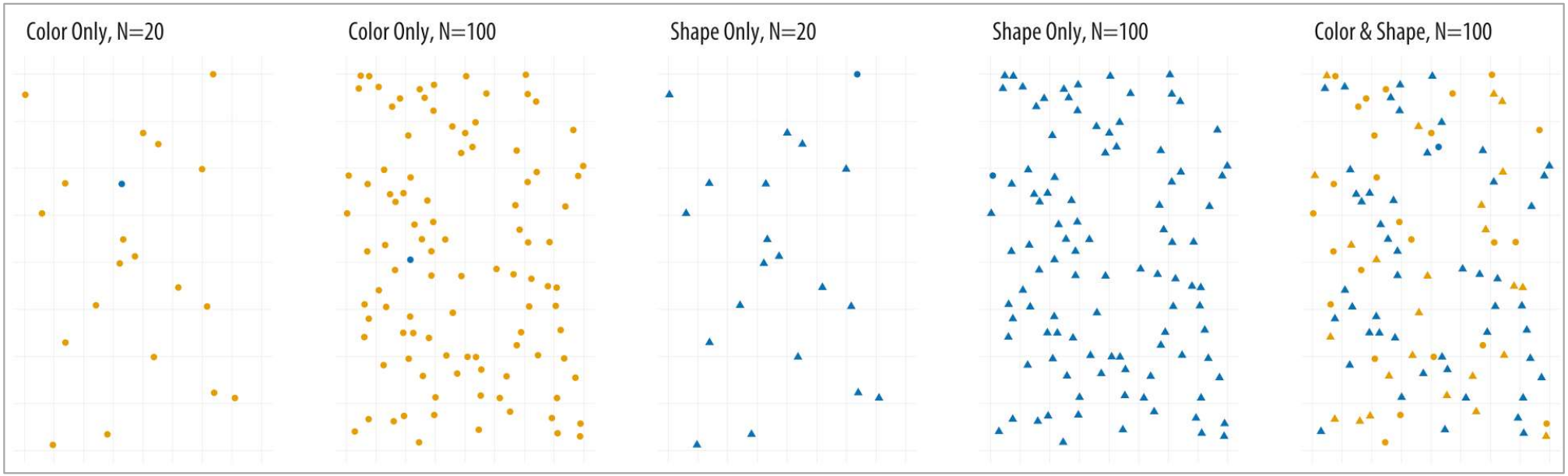
Procesamiento ATENTO



Procesamiento PREATENCIONAL

3. Preatencional y lo que destaca

Tarea: Encontrar el único círculo azul

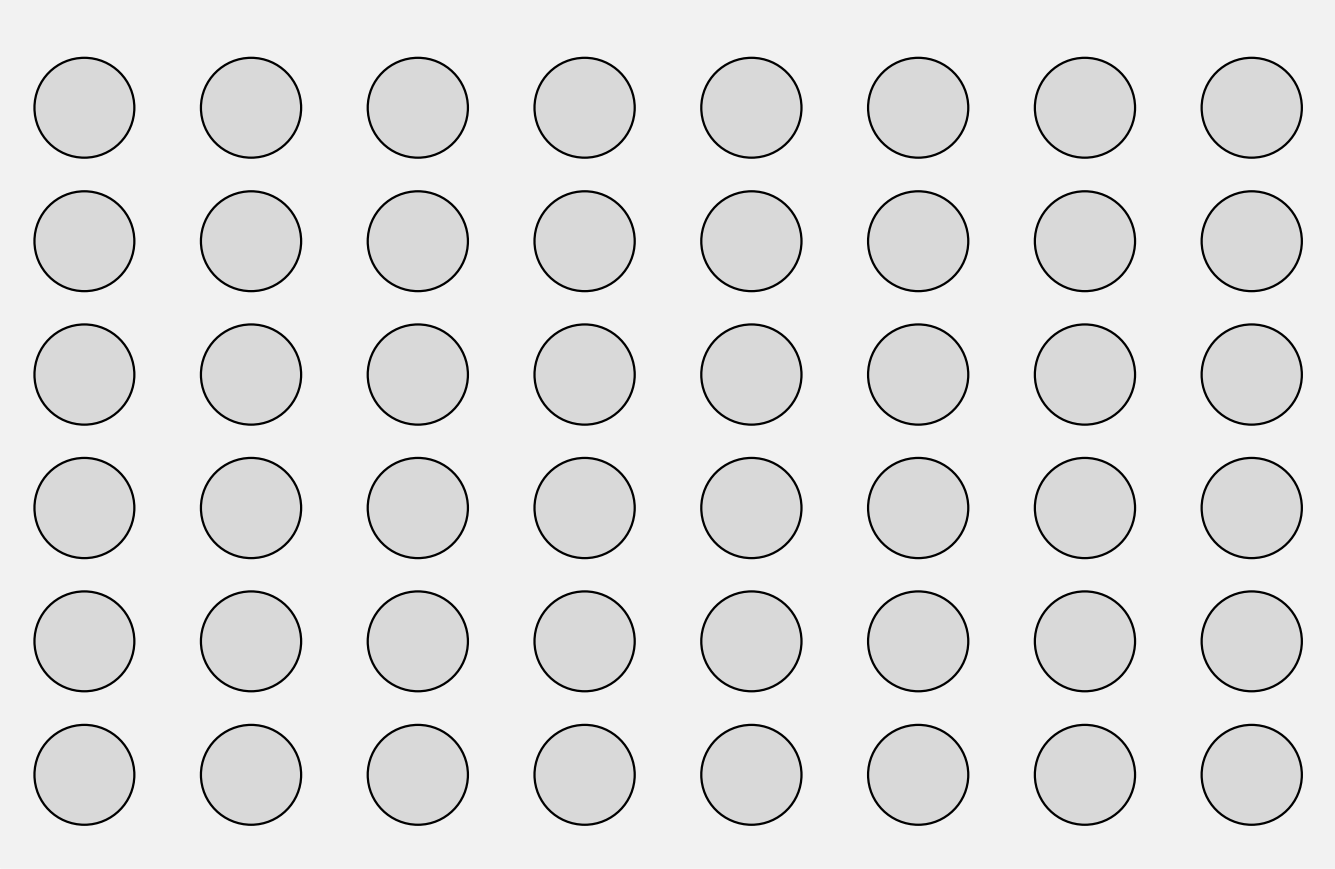


Más fácil

Más difícil

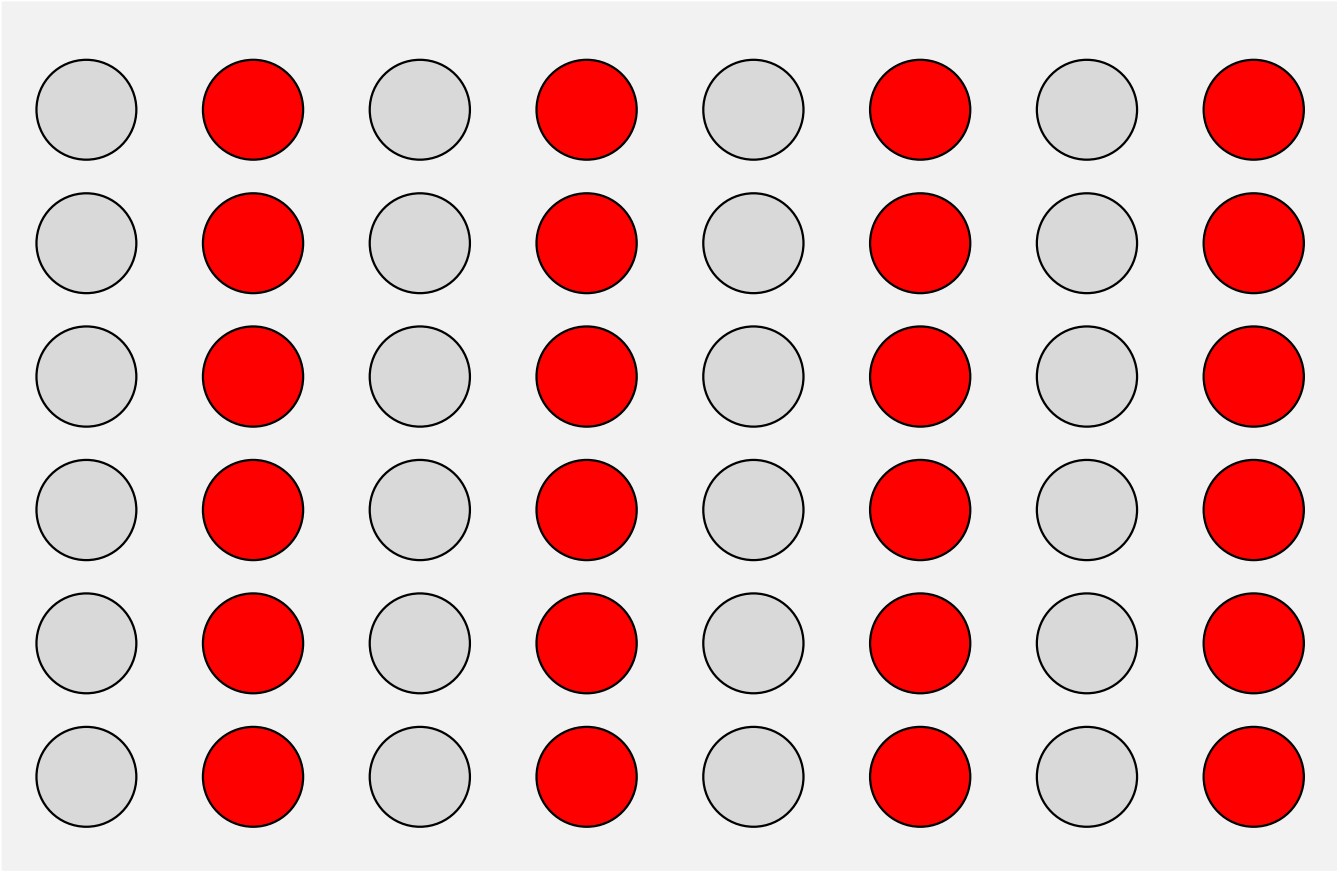
4. Reglas de la Gestalt

¿Cuántos grupos hay?



4. Reglas de la Gestalt

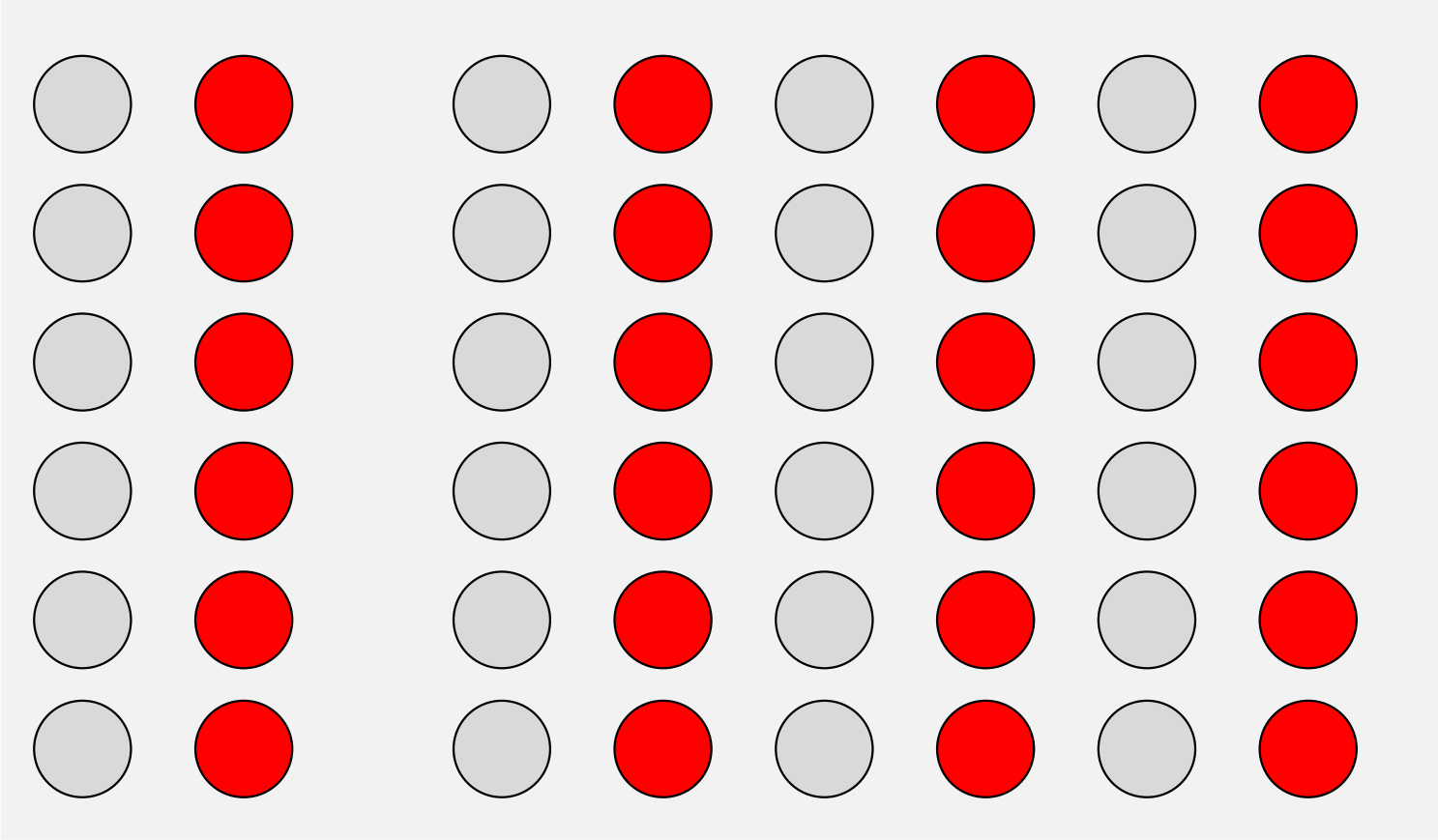
¿Cuántos grupos hay?



SIMILARIDAD

4. Reglas de la Gestalt

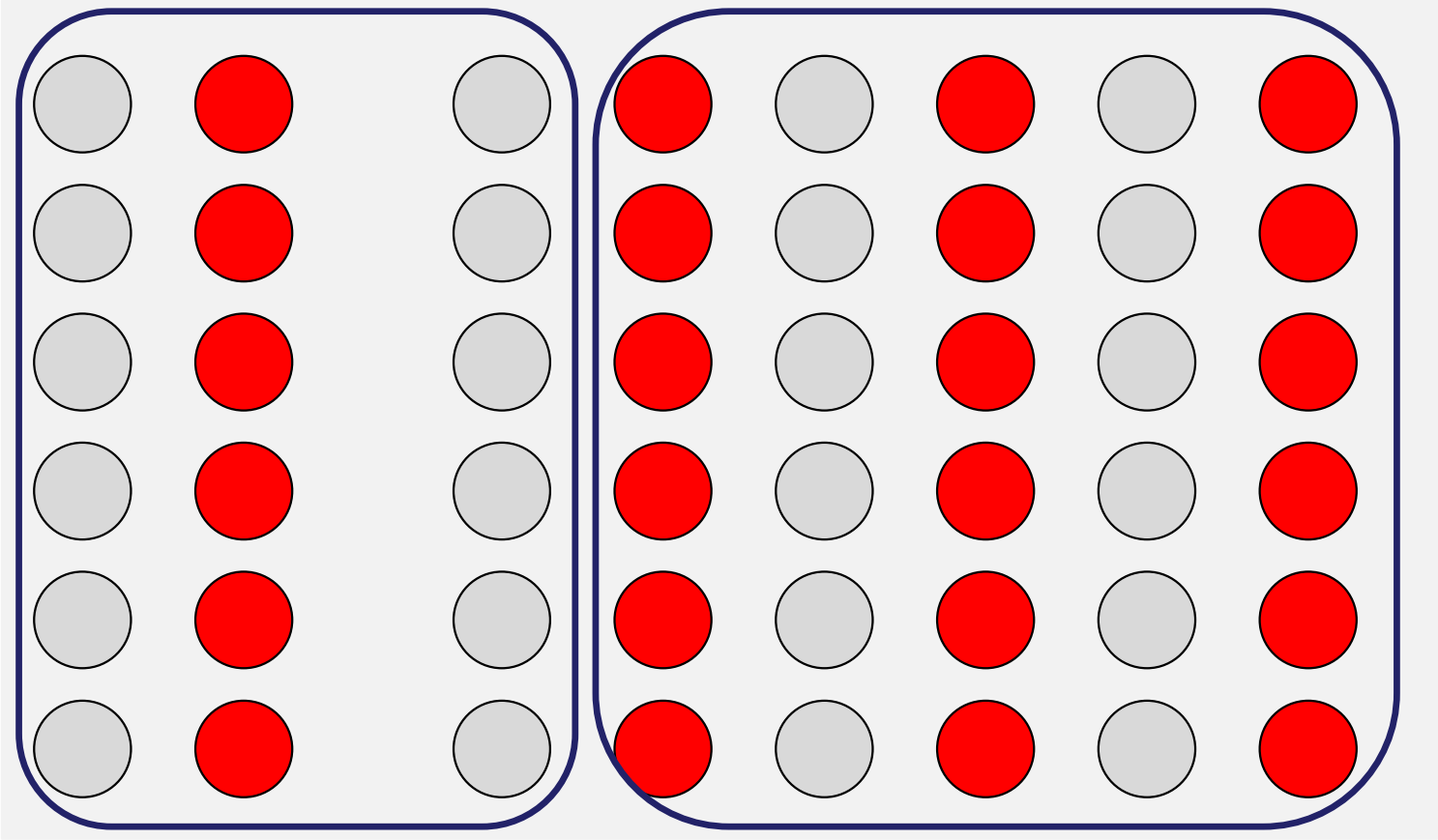
¿Cuántos grupos hay?



PROXIMIDAD

4. Reglas de la Gestalt

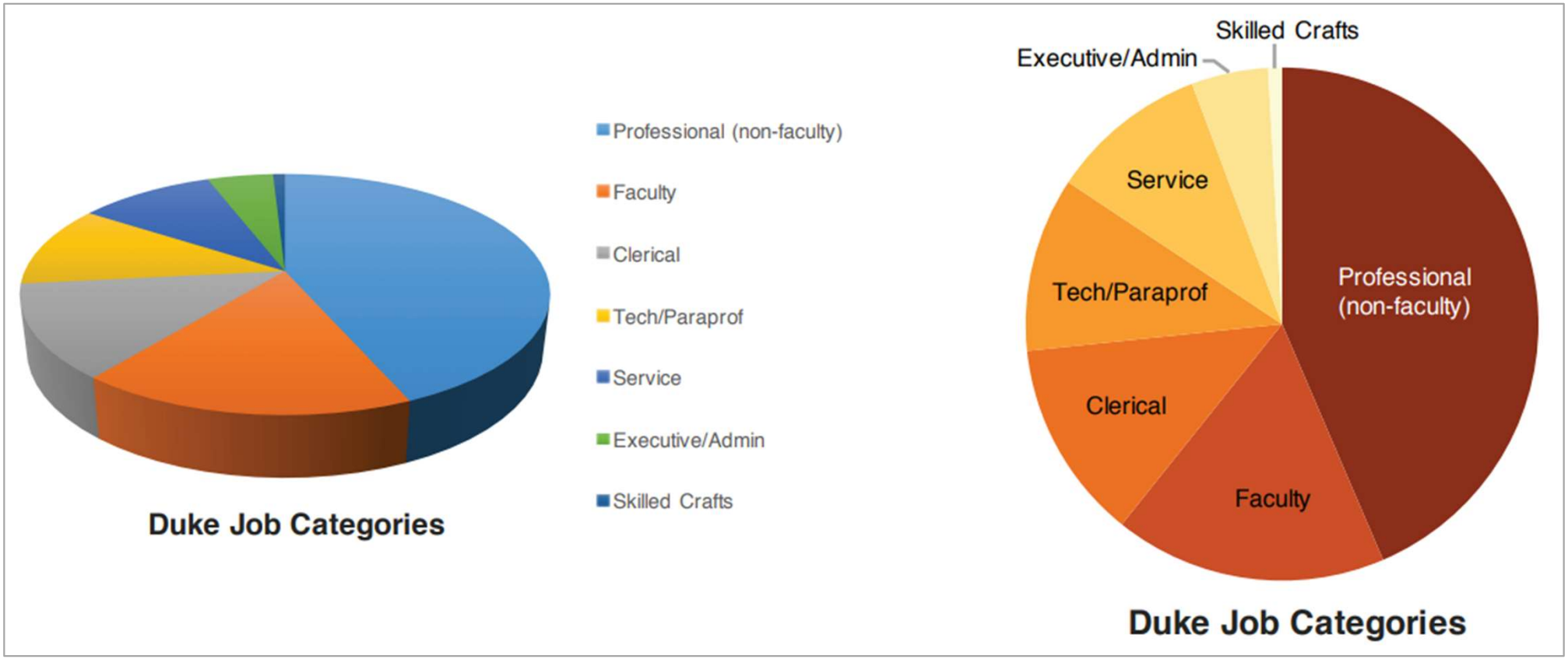
¿Cuántos grupos hay?



ENCIERRO

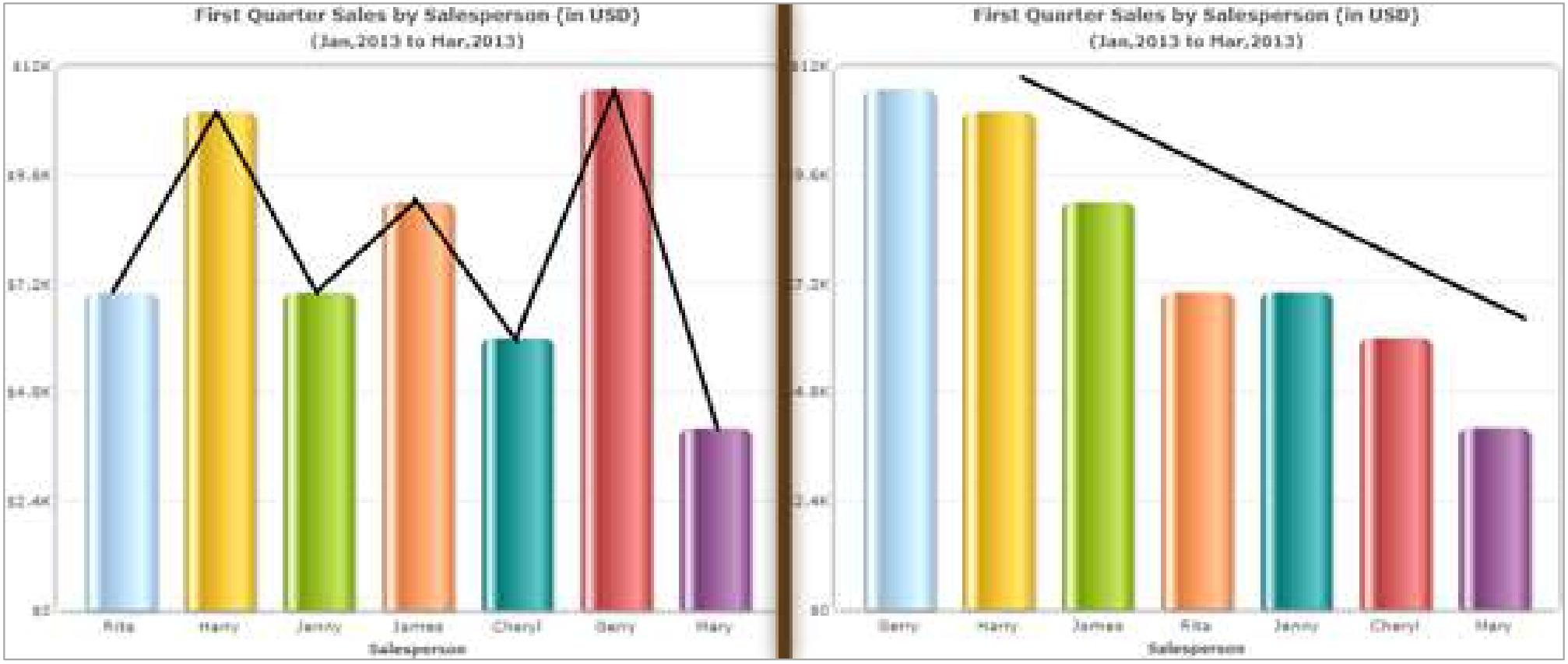
4. Reglas de la Gestalt

Ley de Prägnanz: Hacelo simple. Organizar los datos de forma lógica siempre que sea posible.



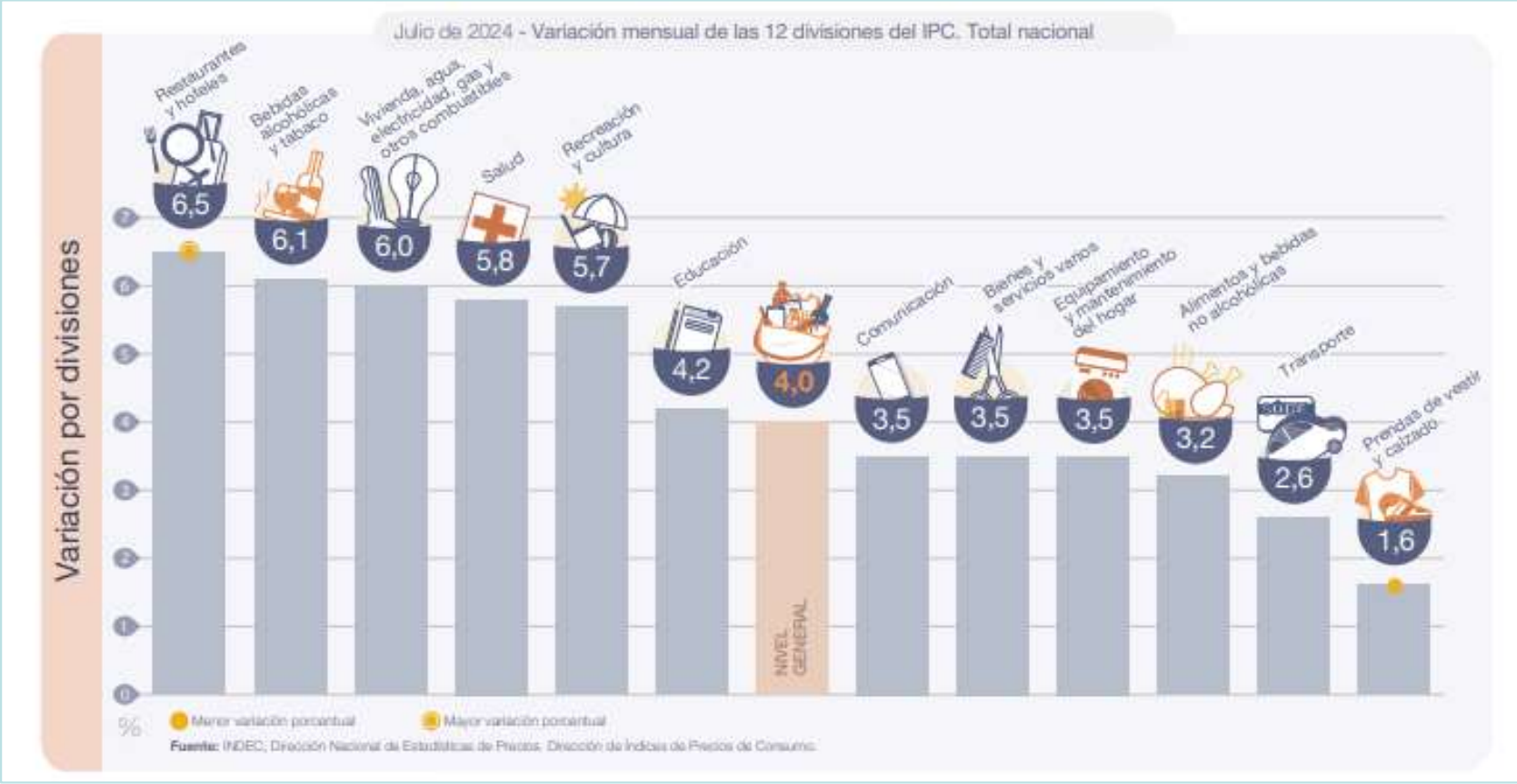
4. Reglas de la Gestalt

Ley de Continuidad: Organizar los objetos en una línea para facilitar agrupación y comparación.



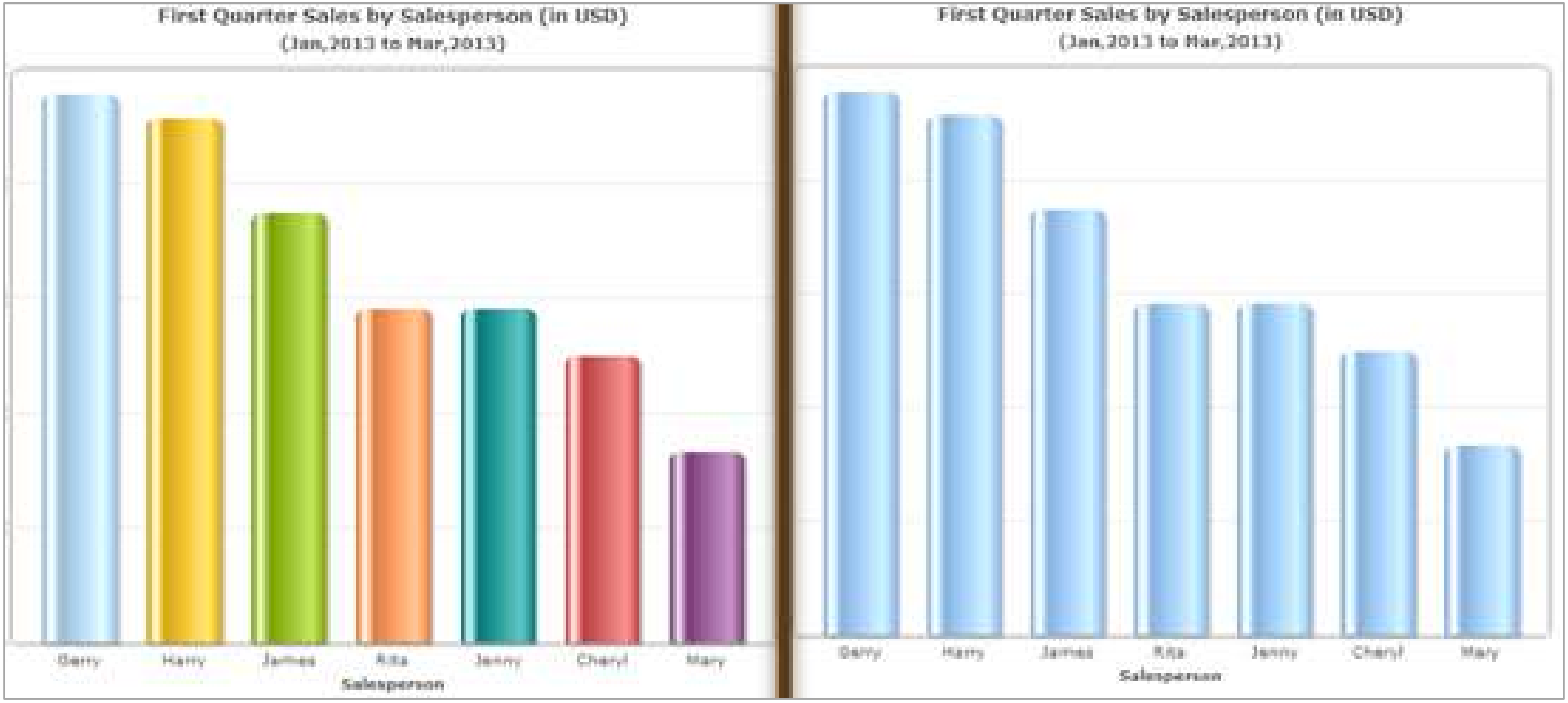
4. Reglas de la Gestalt

Ley de Continuidad: Organizar los objetos en una línea para facilitar agrupación y comparación.



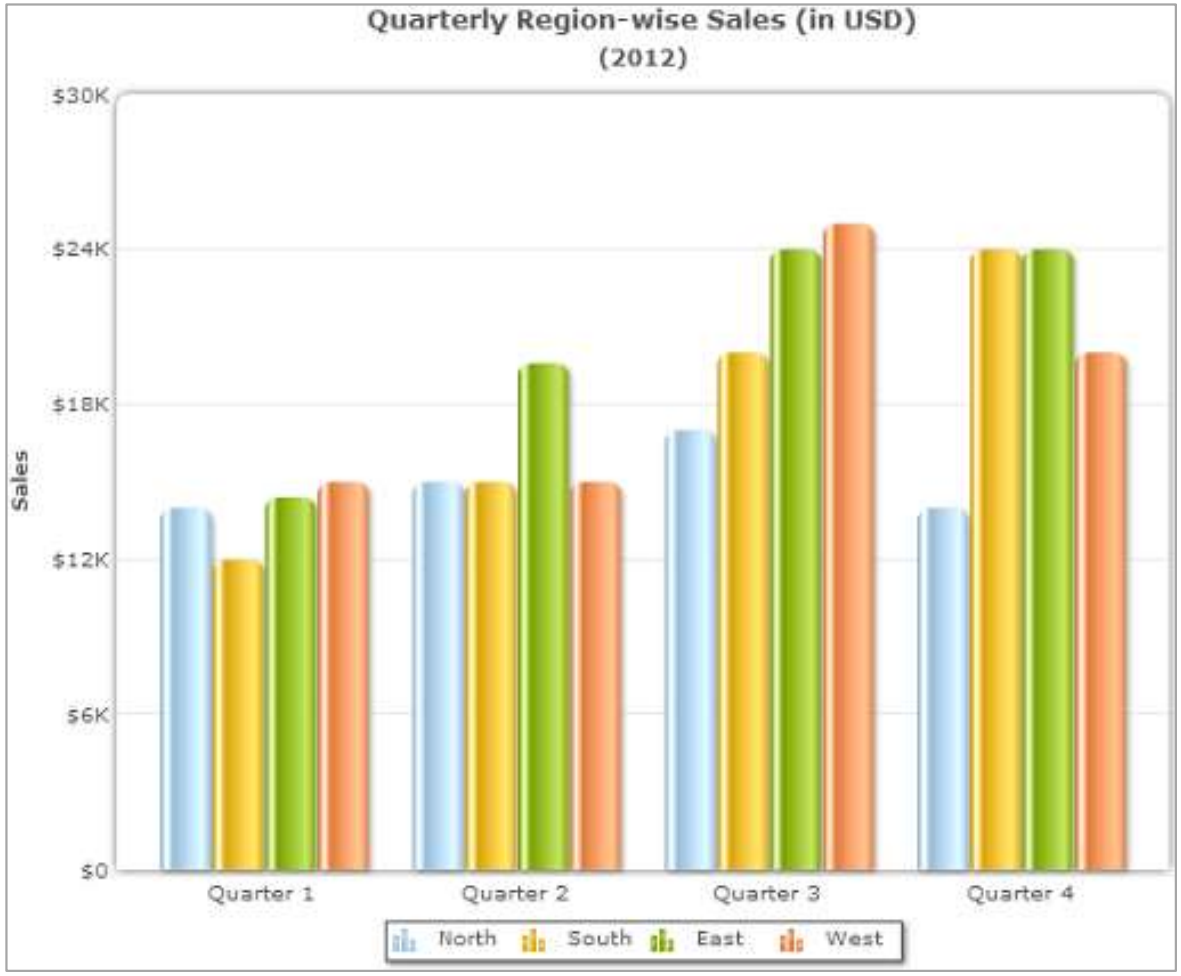
4. Reglas de la Gestalt

Ley de Similaridad: Utilizar características similares para establecer relaciones.



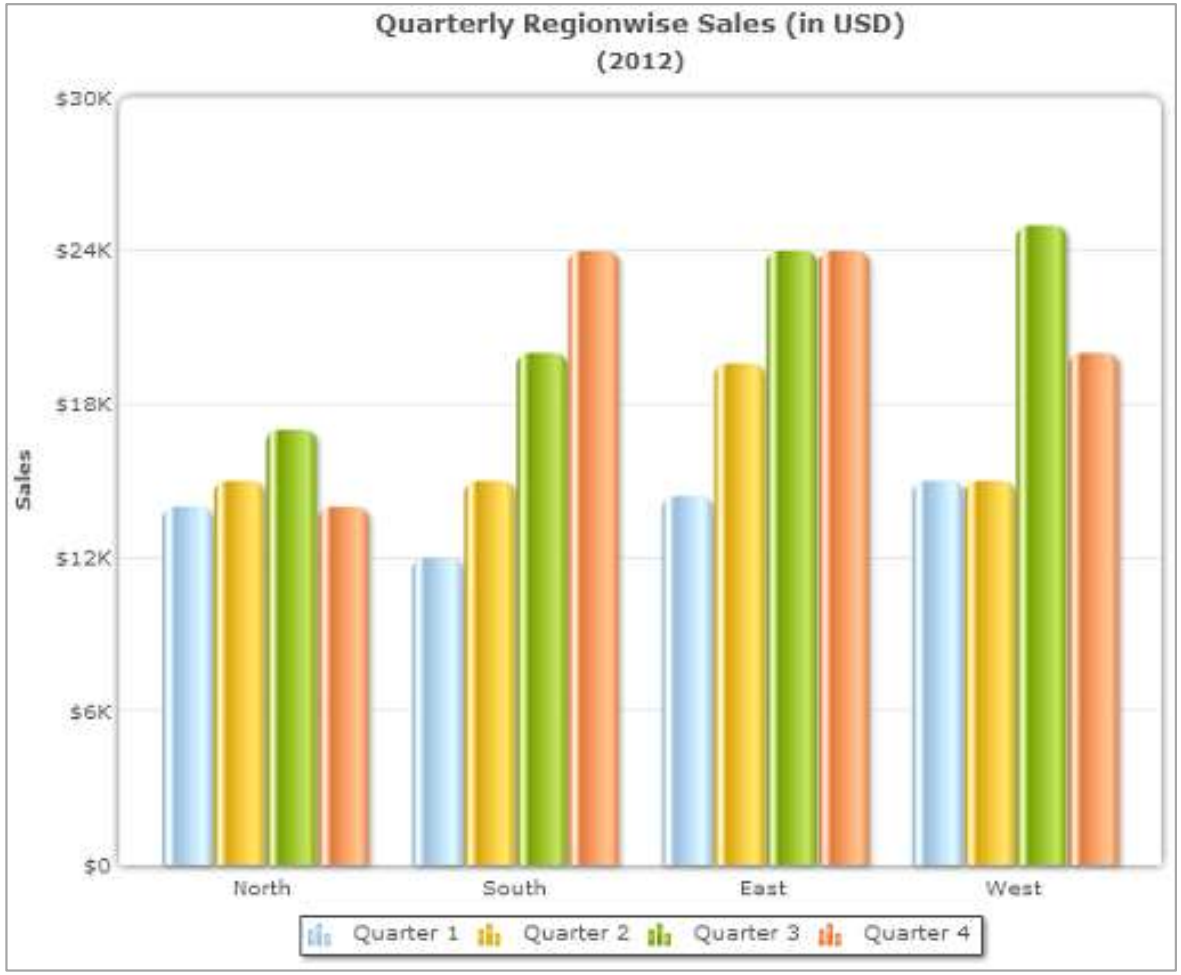
4. Reglas de la Gestalt

Ley de proximidad: Crear agrupaciones por proximidad para respaldar el mensaje.



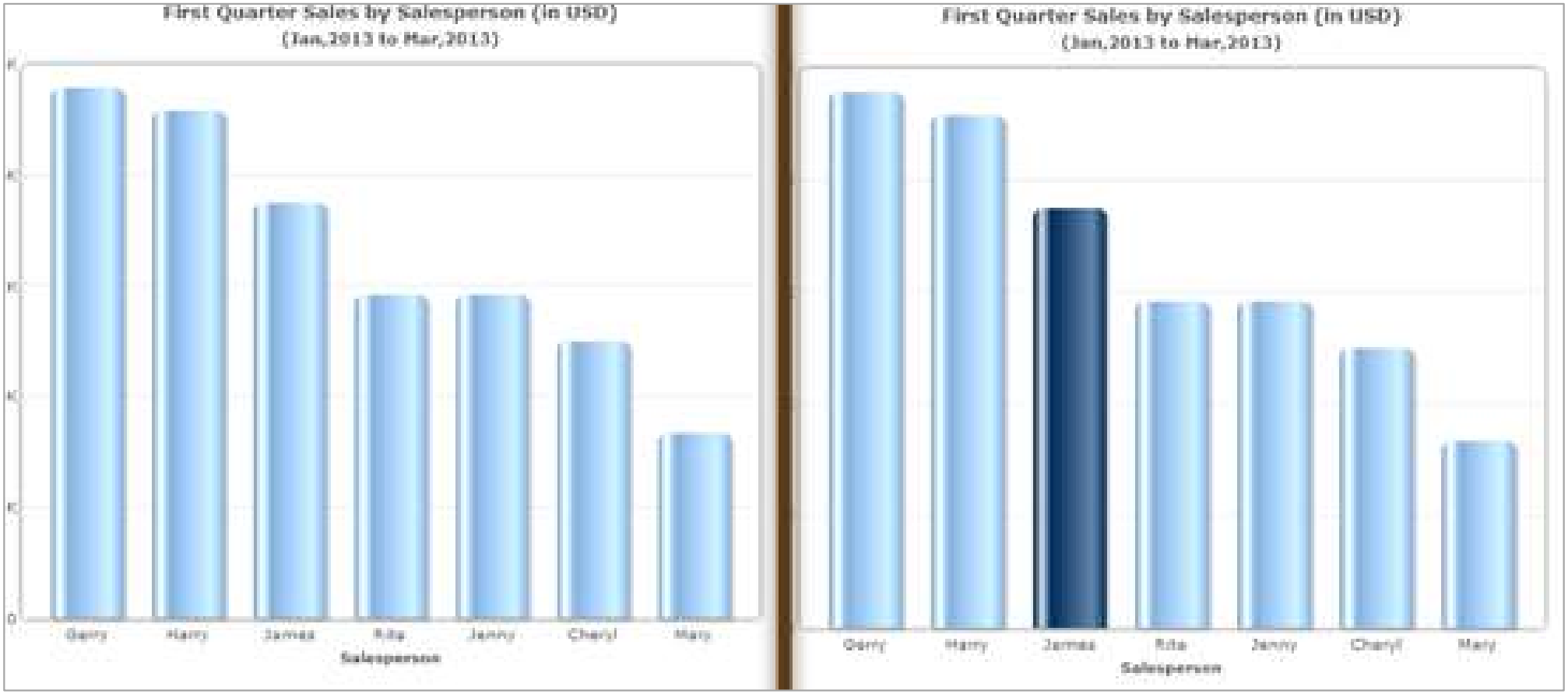
4. Reglas de la Gestalt

Ley de proximidad: Crear agrupaciones por proximidad para respaldar el mensaje.



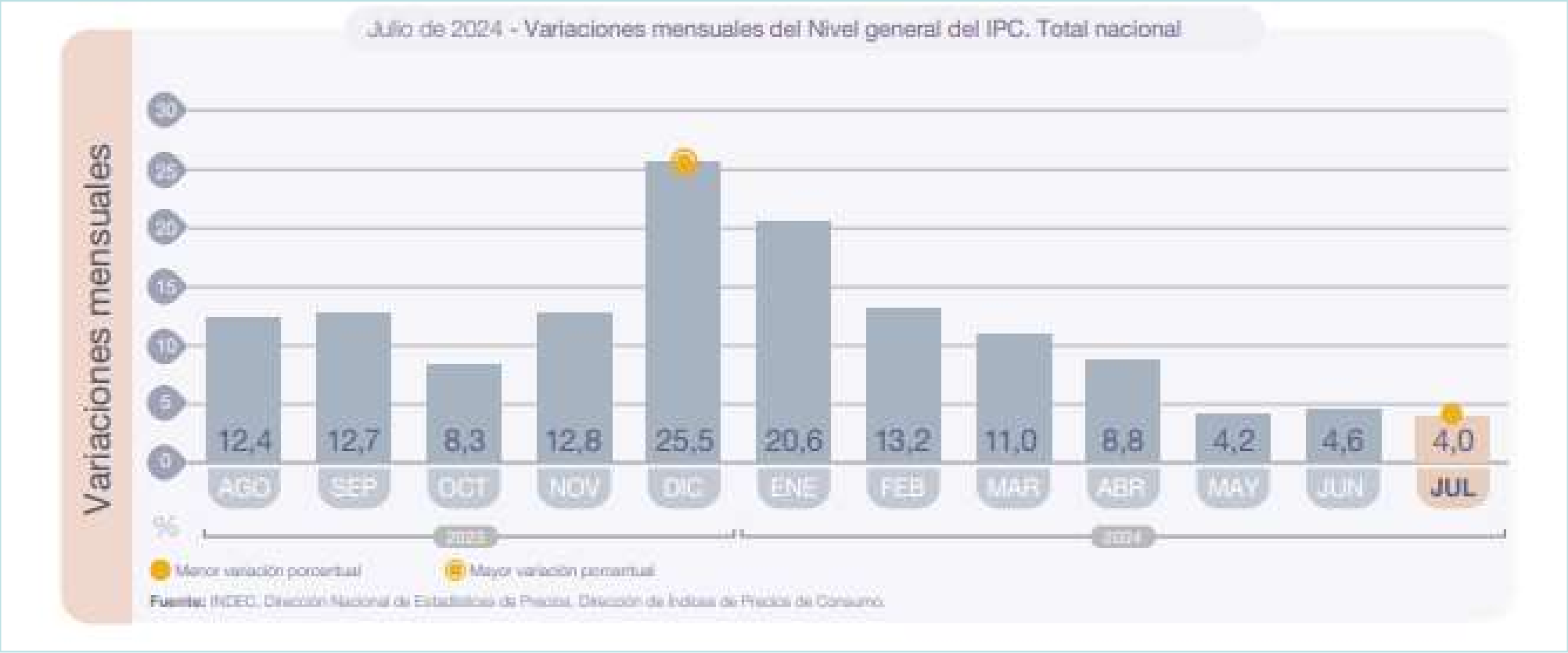
4. Reglas de la Gestalt

Ley del Punto Focal: Usa características para resaltar y crear puntos focales.



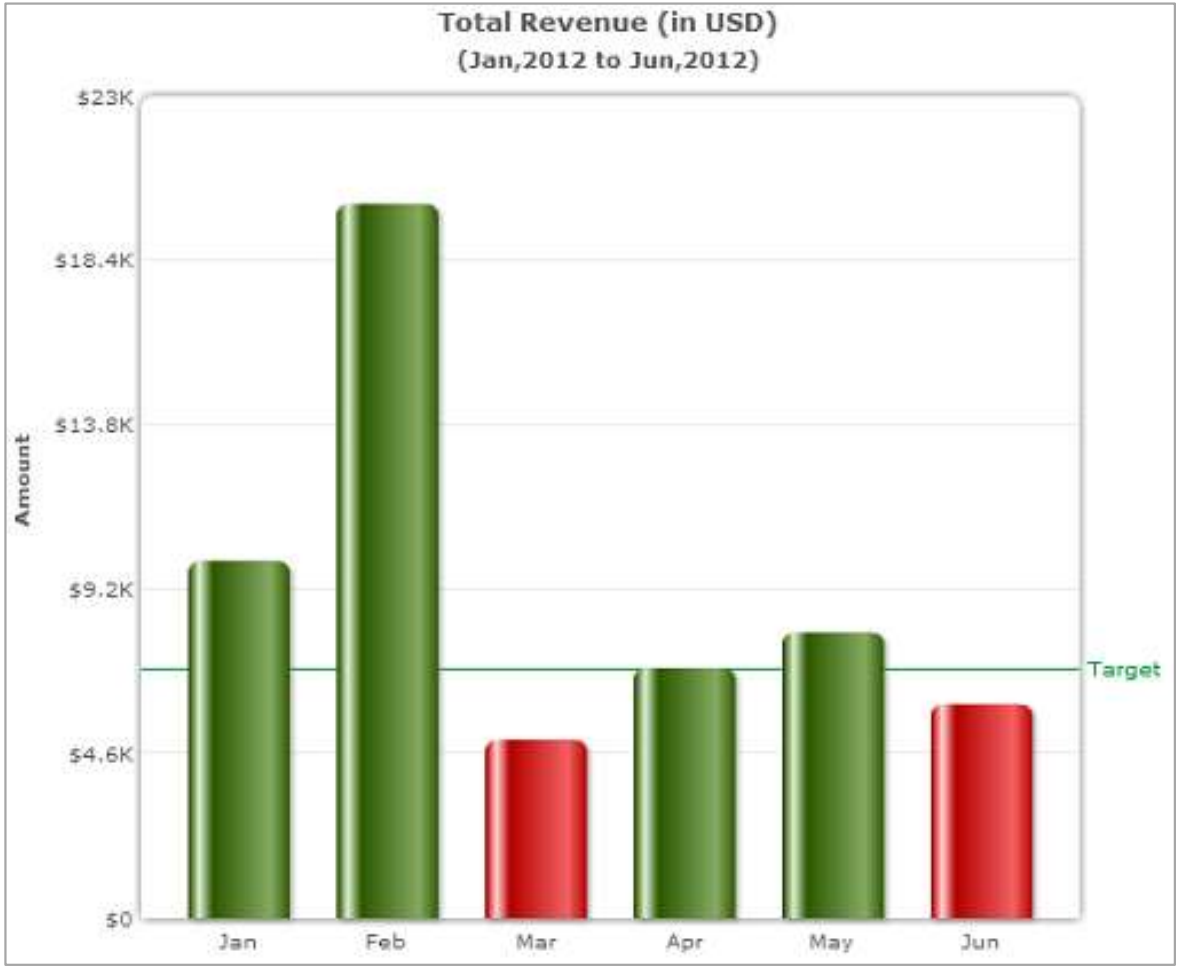
4. Reglas de la Gestalt

Ley del Punto Focal: Usa características para resaltar y crear puntos focales.



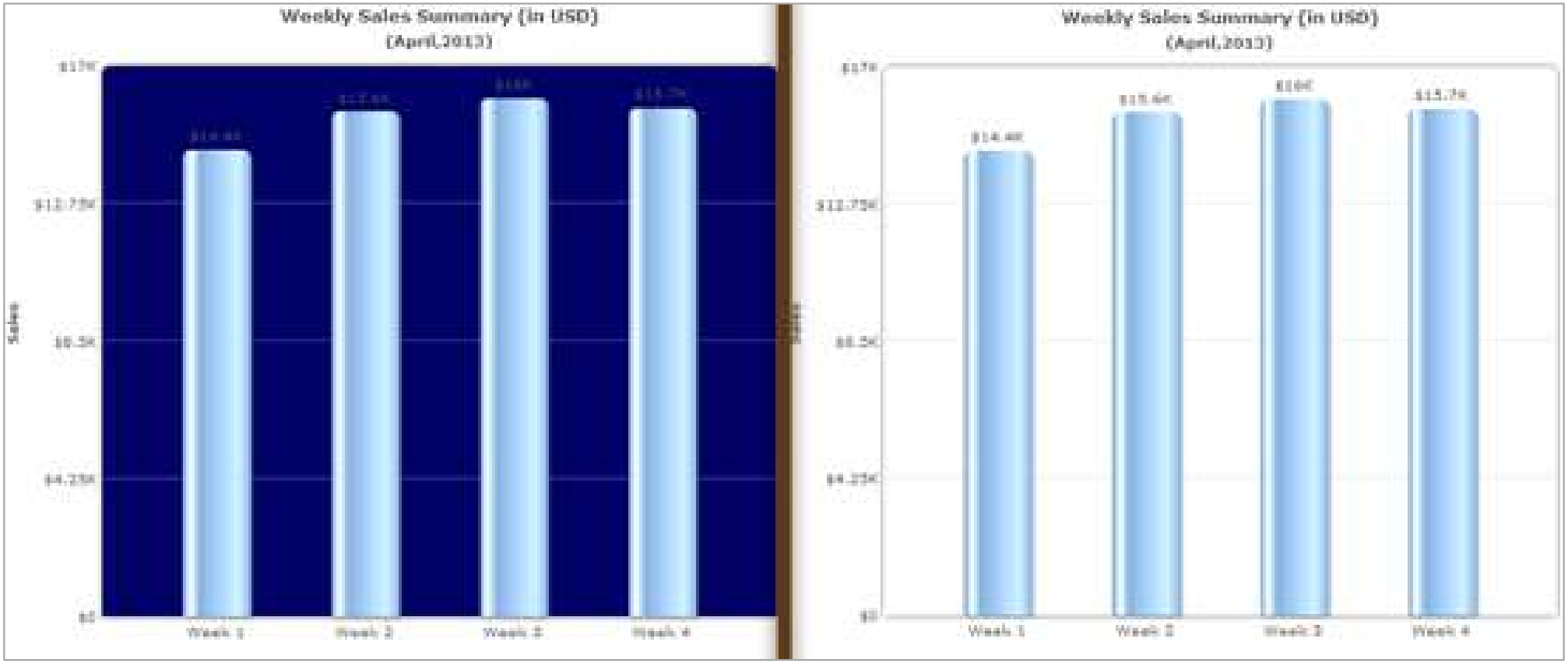
4. Reglas de la Gestalt

Ley de la Correspondencia Isomórfica: Convenciones establecidas (verde +; rojo -)



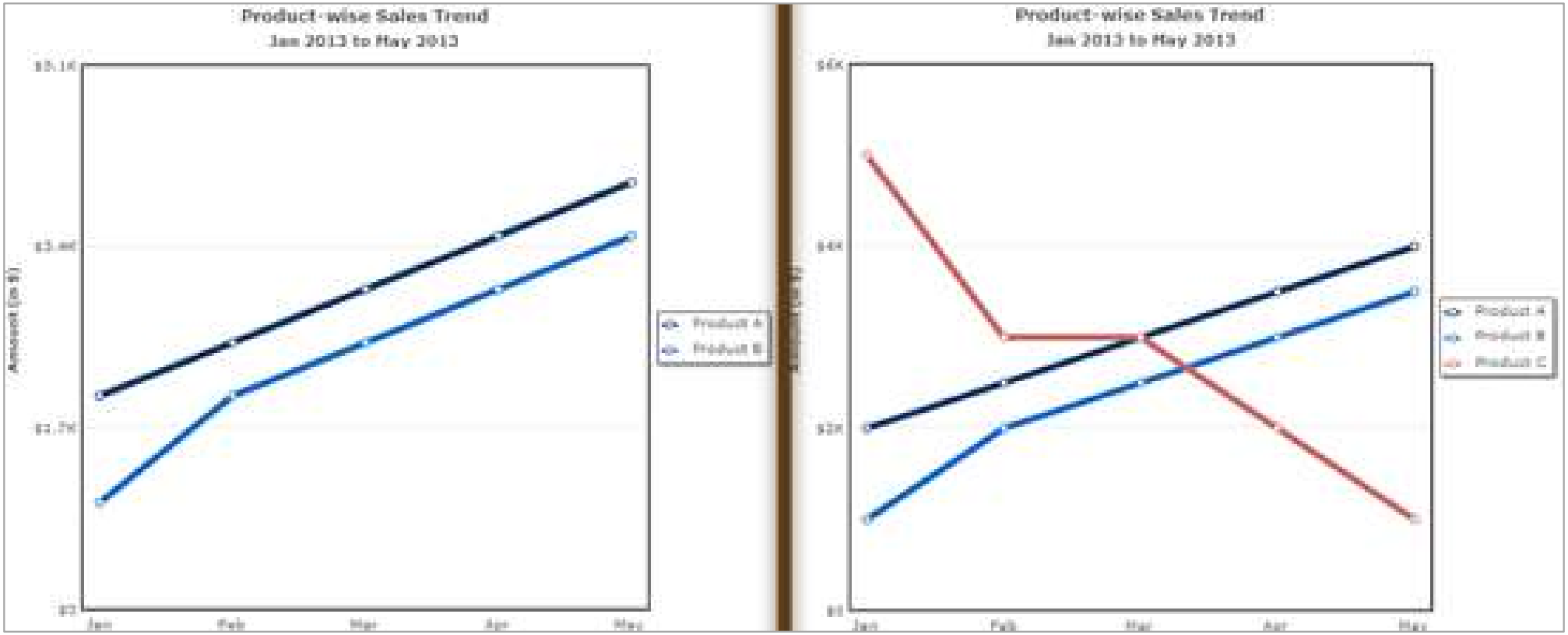
4. Reglas de la Gestalt

Ley de la Figura/Fondo: Suficiente contraste entre el primer plano y el fondo.



4. Reglas de la Gestalt

Ley del Destino Común: Utilizar dirección y/o el movimiento para establecer o negar relaciones.



Análisis

1. Univariado

a) Variable Cuantitativa Discreta

b) Variable Cuantitativa Continua

2. Visualización de Datos

a) ¿Por qué graficamos los datos?

b) Gráficos erróneos

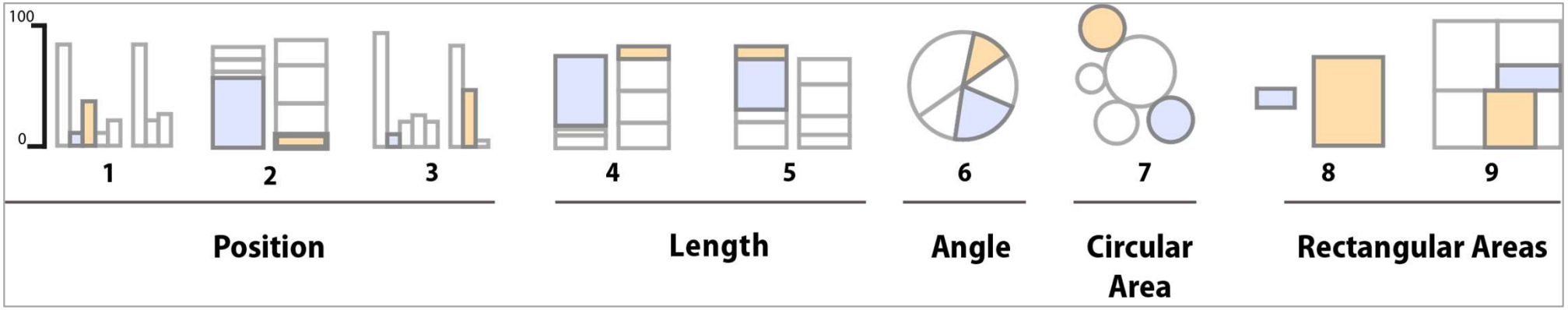
c) Percepción y visualización de datos

d) Tareas visuales y decodificación de datos

e) Dashbord básicos

i. Variables Cualitativas (Univariado)

ii. Bivariado



Menos error

Más error

Análisis

1. Univariado

a) Variable Cuantitativa Discreta

b) Variable Cuantitativa Continua

2. Visualización de Datos

a) ¿Por qué graficamos los datos?

b) Gráficos erróneos

c) Percepción y visualización de datos

d) Tareas visuales y decodificación de datos

e) Dashbord básicos

i. Variables Cualitativas (Univariado)

ii. Bivariado

Análisis

1. Univariado

a) Variable Cuantitativa Discreta

b) Variable Cuantitativa Continua

2. Visualización de Datos

a) ¿Por qué graficamos los datos?

b) Gráficos erróneos

c) Percepción y visualización de datos

d) Tareas visuales y decodificación de datos

e) Dashbord básicos

i. Variables Cualitativas (Univariado)

ii. Bivariado

¿De qué depende sobrevivir o no?

- Mujeres y Niños… ¿primero?
- Clase social (¿?)
- Número de familiares a bordo (¿?)
- Puerto de embarque (¿?)

Referencias

[1] Carver, R., Everson, M., Gabrosek, J., Rowell, G. H., Norton, N., Lock, R., & Wood, B. (2016, February). Draft: Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report [online]. http://www.amstat.org/education/gaise/collegeupdate/GAISE2016_DRAFT.pdf

[2] Wild, C., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. International Statistical Review, 67(3), 223–248. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x

[3] R. A. Poldrack, (2018). Statistical Thinking for the 21st Century. https://statsthinking21.org

[4] Marriot, N. (2014). The future of Statistical Thinking. Significance, Volume: 11, Issue: 5, Pages: 78-80. https://doi.org/10.1111/j.1740-9713.2014.00787.x

[5] Camm, Jeffrey D., Cochran, James J., Fry, Michael J., Ohlmann, Jeffrey W., (2021) Data Visualization Exploring and Explaining with Data. Cengage South-Western, p. 448.

[6] Pearson, Ronald K. (2018). Exploratory Data Analysis Using R. Chapman & Hall/CRC data mining and knowledge discovery, No. 45, CRC Press Taylor & Francis Group: Boca Raton, FL.

[7] Batanero C., Estepa A. y Godino J. D. (1991). Análisis Exploratorio de Datos: sus posibilidades en la enseñanza secundaria. *Suma* , No. 9, p. 25-31. https://www.ugr.es/~batanero/pages/ARTICULOS/anaexplora.pdf

Referencias

[8] Healy, K (2018). Data Visualization. A practical introduction. Princeton University Press

[9] Tufte, E. R (1983). The Visual Display of Quantitative Information. Graphics Press.

[10] Tukey, J. W (1977). Exploratory Data Analysis. Addison-Wesley.

[11] Cleveland, W. S., & McGill, R. (1984). Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. Journal of the American Statistical Association

[12] Cleveland, W. S. (1985). The Elements of Graphing Data. Wadsworth.

[13] Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10). Association for Computing Machinery, New York, NY, USA, 203–212. https://doi.org/10.1145/1753326.1753357