Análisis exploratorio

Roberto Muñoz

Astrónomo y Data Scientist

Director científico MetricArts

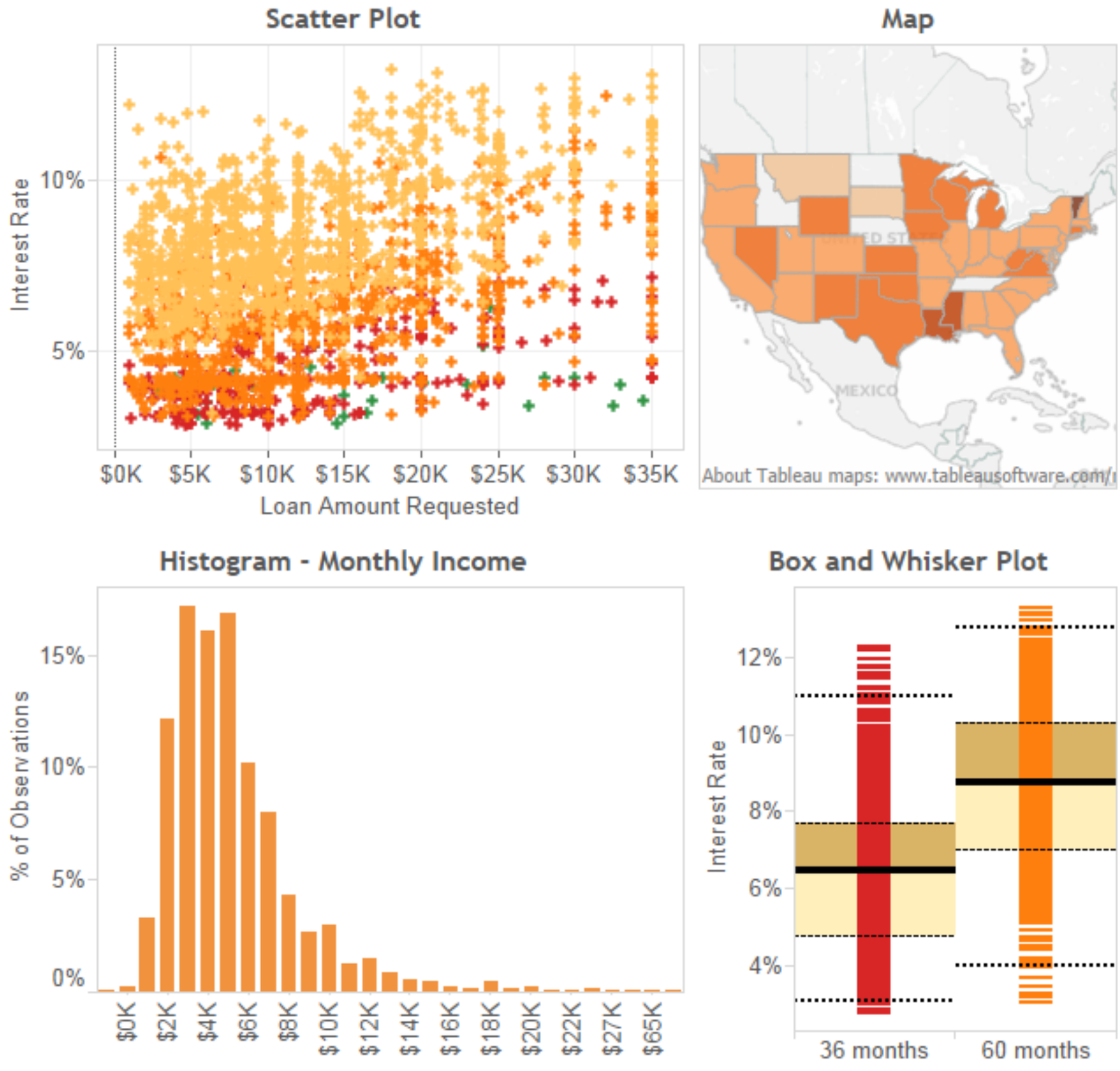
github.com/rpmunoz

@RobertoKPax

¿Qué es el análisis exploratorio?

- El análisis exploratorio (exploratory data analysis o EDA) es una metodología diseñada para entender las características principales de un conjunto de datos (dataset).
- **Objetivo**: Explorar los dataset para lograr un entendimiento básico de sus datos y de las relaciones existentes entre las variables analizadas.

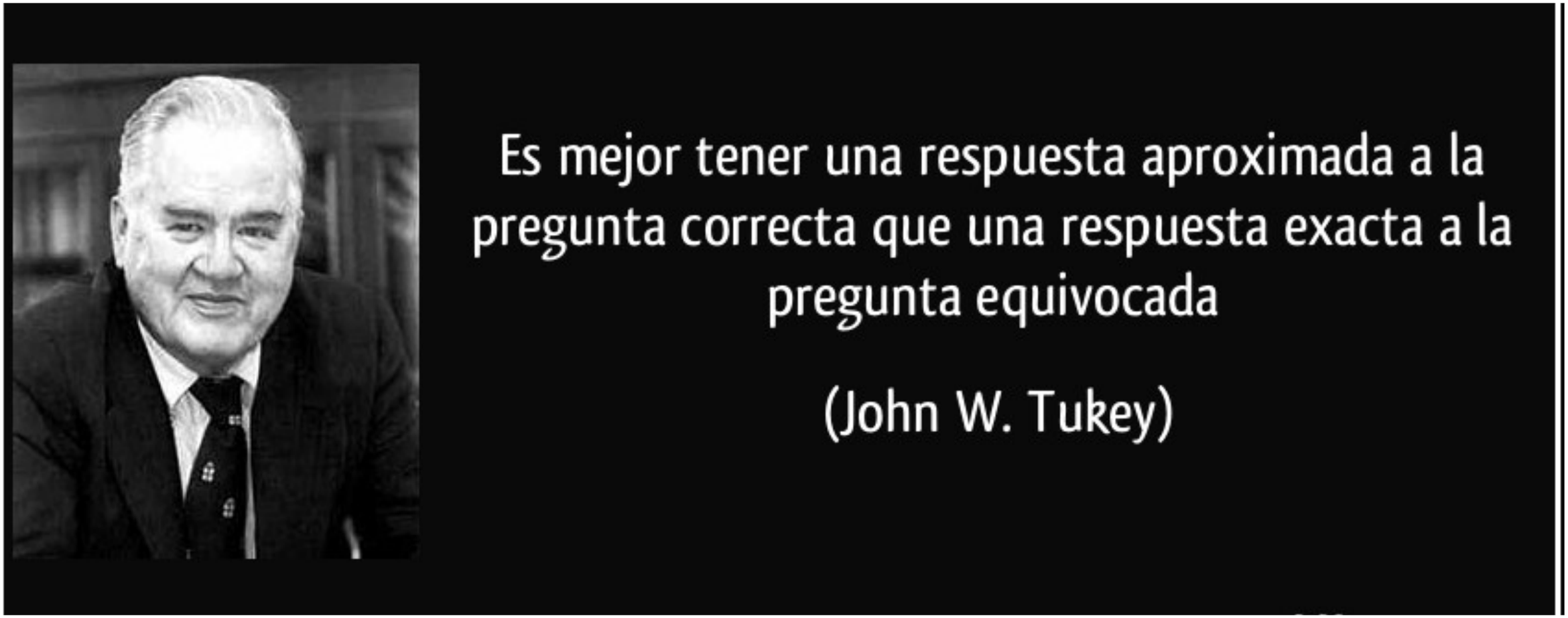
Visualizaciones



Historia del análisis exploratorio

- El análisis exploratorio fue propuesto por el estadístico norteamericano John Tukey el año 1977.
- Tukey hizo notar que la Estadística de aquel entonces le daba mucho énfasis a los test de hipótesis estadísticos.
- Sugirió usar los datos para generar nuevos test de hipótesis y generar nuevos experimentos.

John Tukey


Es mejor tener una respuesta aproximada a la pregunta correcta que una respuesta exacta a la pregunta equivocada
(John W. Tukey)

Tareas del EDA

- Principales tareas del análisis exploratorio

  1. Entendimiento de los datos
  2. Limpieza de los datos
  3. Transformación y manipulación de los datos
  4. Preparación de los datos para su posterior análisis estadístico

- El examen previo de los datos es un paso necesario

Etapas del EDA

1. Preparar los datos para hacerlos accesibles
2. Realizar un análisis descriptivo y examen gráfico de la naturaleza de las **variables individuales**
3. Realizar un análisis descriptivo y examen gráfico de las **relaciones entre las variables** analizadas
4. Evaluar algunos **supuestos básicos** como normalidad
5. Identificar **casos atípicos** (outliers) y evaluar impacto potencial
6. Evaluar el impacto potencial de **datos ausentes** (missing)

1. Preparación de los datos

- Definir método de entrada (teclado, archivo o web) y hacer datos accesibles a cualquier análisis estadístico.
- Aplicar operaciones sobre los datos
  - Combinar datasets
  - Seleccionar subconjunto
  - Transformar variables
  - Ordenar casos

2. Análisis de variables individuales

- Hacer análisis estadístico gráfico y numérico de las variables del problema
- Armar un mapa de la información contenida en el dataset
- Dependiendo de escala de medida o tipo de datos, se sugieren ciertas visualizaciones y resúmenes descriptivos.

Análisis de variables individuales

Escala de medida	Visualización	Medidas de localización	Medidas de dispersión
Nominal (cualitativa)	Diagrama de barras Diagrama de líneas Gráfico de torta	Moda	
Ordinal (cuantitativa)	Diagrama de cajas o Boxplot	Mediana	Rango intercuartílico
Intervalo	Histogramas Polígono de frecuencias	Media	Desviación estándar
Razón		Media geométrica	Coeficiente de variación

Variables cualitativas

- Encuesta acerca del estado civil de clientes de supermercado

Tabla 2
Tabla de frecuencias del Estado Civil

Estado Civil

	Frecuencia	Porcentaje
Soltero	77	19.2
Casado	305	75.9
Viudo	16	4.0
Separado	4	1.0
Total	402	100.0

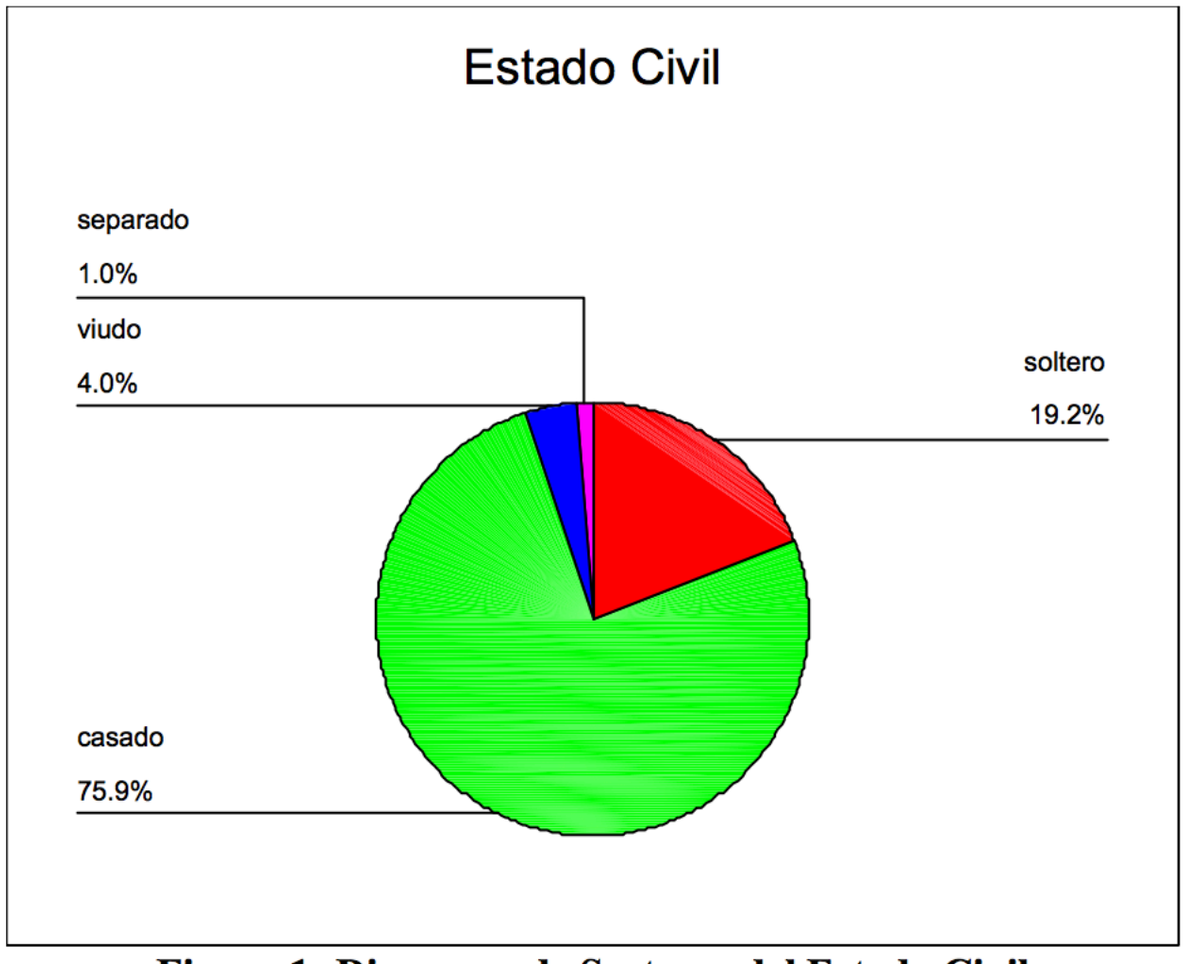


Figura 1: Diagrama de Sectores del Estado Civil

Variables cuantitativas

- Encuesta acerca del número de miembros del grupo familiar de clientes de supermercado

Tabla 4
Tabla de frecuencias del Número de Miembros que viven en casa

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	0	1	.2	.3	.3
	1	30	7.5	7.5	7.8
	2	91	22.6	22.8	30.5
	3	87	21.6	21.8	52.3
	4	129	32.1	32.3	84.5
	5	43	10.7	10.8	95.3
	6	12	3.0	3.0	98.3
	7	7	1.7	1.8	100.0
	Total	400	99.5	100.0	
Perdidos	Sistema	2	.5		
Total		402	100.0		

Tabla 5
Estadísticos descriptivos de la variable Número de Miembros que viven en casa

Estadísticos

miembros que viven en casa

N	Válidos	400
	Perdidos	2
Media		3.31
Mediana		3.00
Moda		4
Desv. típ.		1.33
Asimetría		.234
Error típ. de asimetría		.122
Curtosis		-.107
Error típ. de curtosis		.243
Mínimo		0
Máximo		7
Percentiles	25	2.00
	50	3.00
	75	4.00

Variables cuantitativas

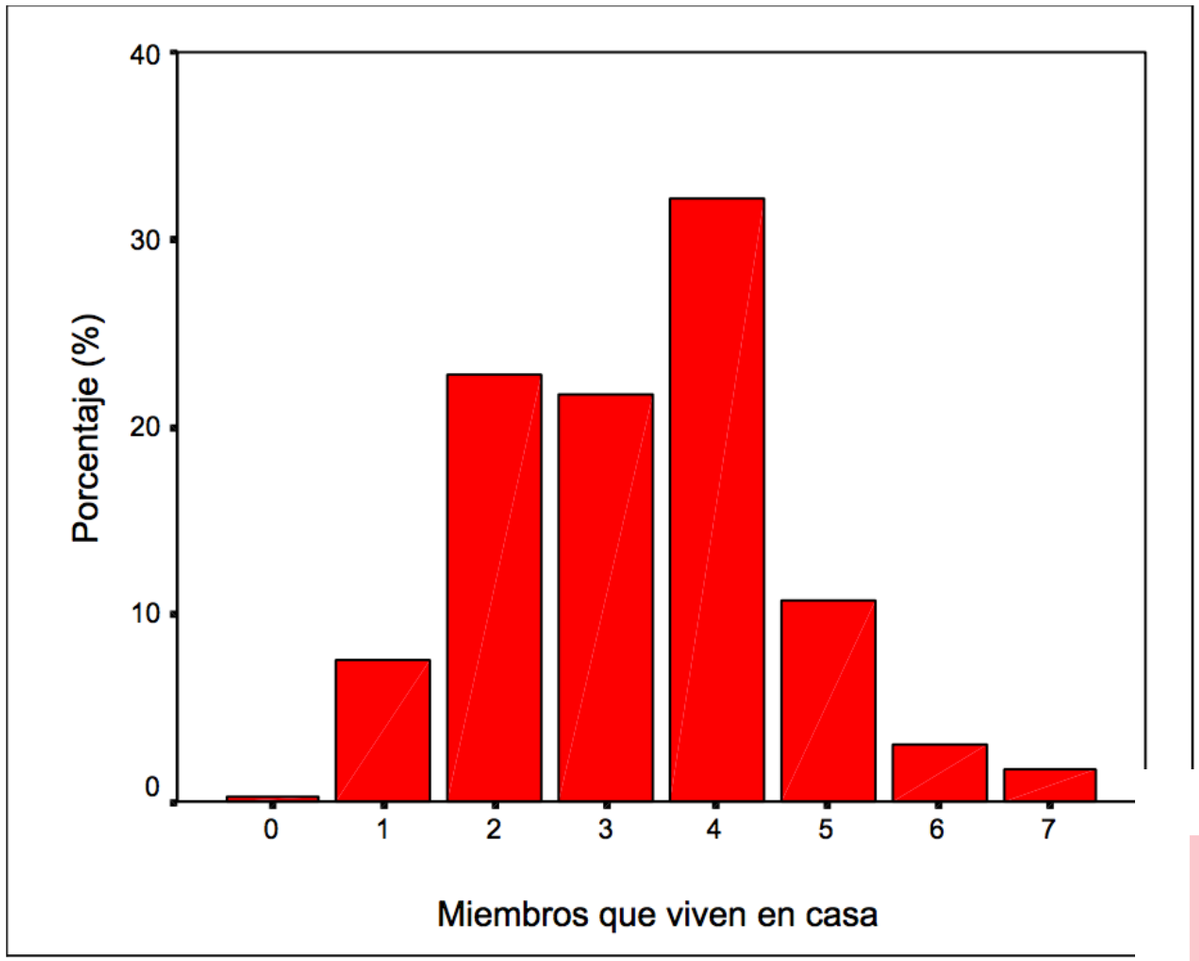


Figura 3: Diagrama de Barras del Número de Miembros que viven en casa

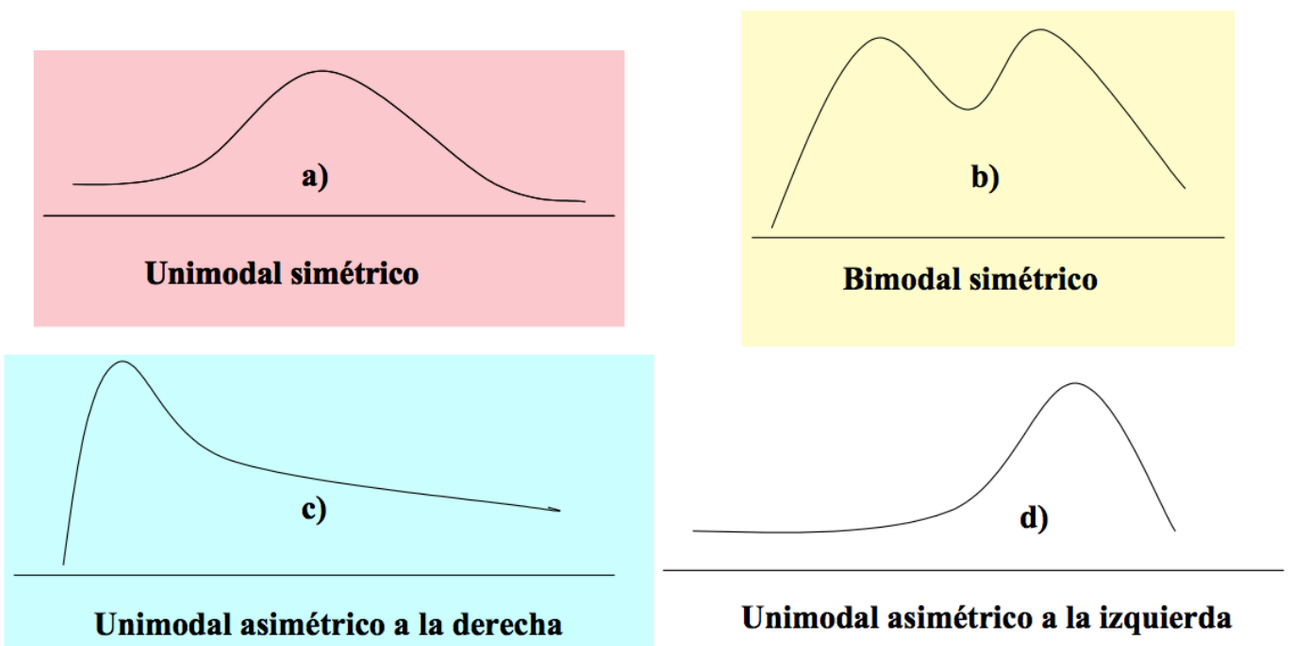


Figura 4:
Tipología de las distribuciones de frecuencias agrupadas

3. Análisis de relaciones entre variables

- Analizar la existencia de posibles relaciones entre las variables del dataset.
- En general se aplica análisis bidimensional (dos variables). Los casos típicos son,
  - Ambas variables son cualitativas
  - Ambas variables son cuantitativas
  - Una variable es cuantitativa y la otra cualitativa

Análisis de dos variables cuantitativas

- La distribución conjunta de dos variables puede expresarse gráficamente mediante un diagrama de dispersión queproporciona una buena descripción de la relación entre las dos variables.
- La relación entre las variables también puede expresarse de forma numérica. Una medida de la relación entre dos variables que resuma la información del gráfico de dispersión y que no dependa de las unidades de medida es el coeficiente de correlación lineal.

Casos de heterogeneidad

- A) Hay un dato atípico o discordante con el resto, que modifica el signo de la correlación. Puede comprobarse que si el punto A no existiese, el coeficiente de correlación sería positivo, mientras que su presencia hace la correlación negativa.

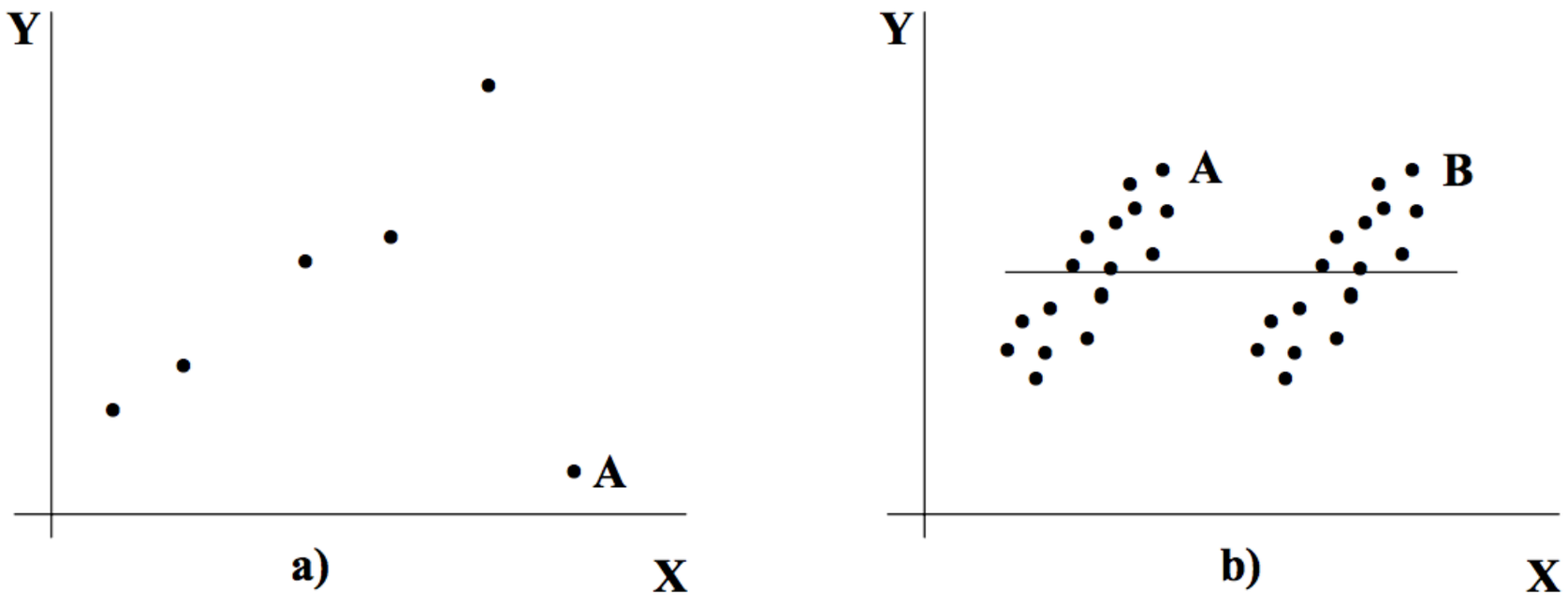


Figura 22: Dos casos frecuentes de heterogeneidad

Casos de heterogeneidad

- B) En este caso el gráfico indica que la relación entre las variables es distinta para los individuos del grupo A que para los del B y si calculamos un coeficiente de correlación para todos los datos obtendremos un valor muy pequeño.

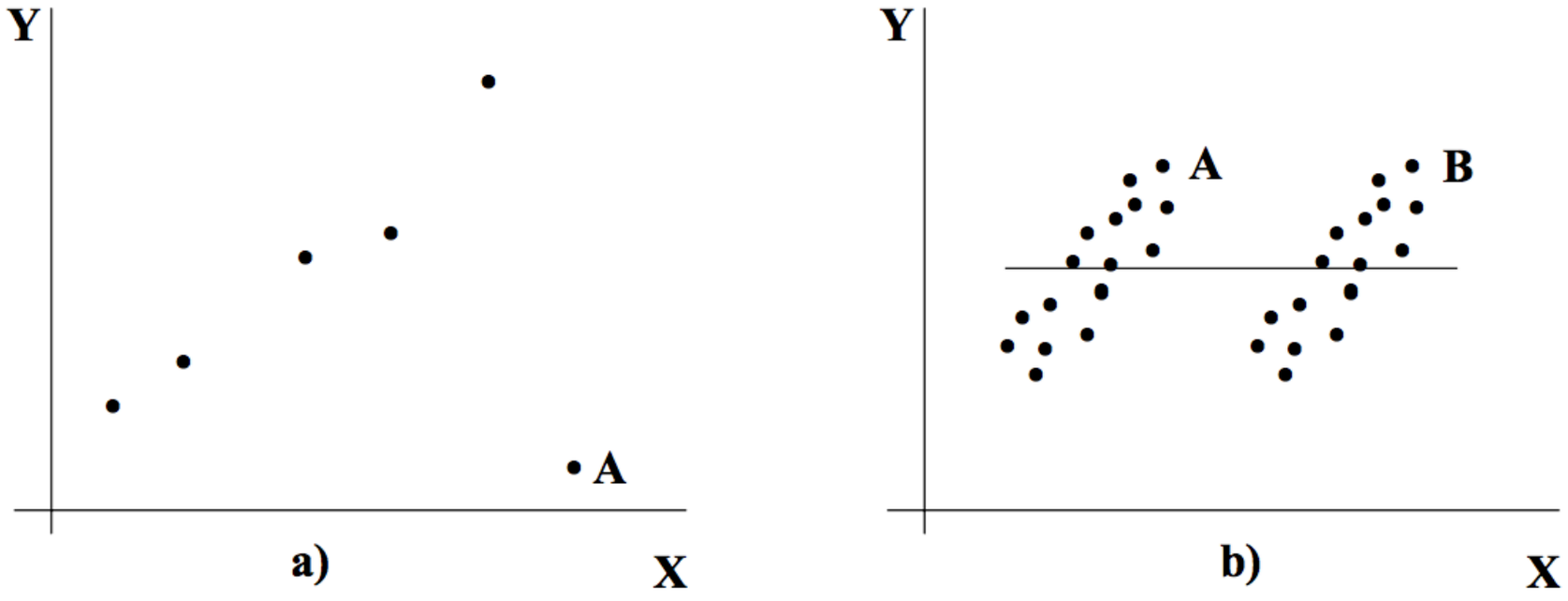


Figura 22: Dos casos frecuentes de heterogeneidad