

INTRODUCCIÓN AL ANÁLISIS DE DATOS

M.Sc. Henry López



"LOS DATOS SON EL NUEVO PETRÓLEO. ES VALIOSO, PERO SI NO ESTÁ REFINADO, REALMENTE NO SE PUEDE USAR. TIENE QUE CAMBIARSE A GAS, PLÁSTICO, PRODUCTOS QUÍMICOS, ETC. PARA CREAR UNA ENTIDAD VALIOSA QUE IMPULSE LA ACTIVIDAD RENTABLE; ENTONCES LOS DATOS DEBEN DESGLOSARSE, ANALIZARSE PARA QUE TENGAN VALOR".

— Clave humby 2006 & Michael Palmer

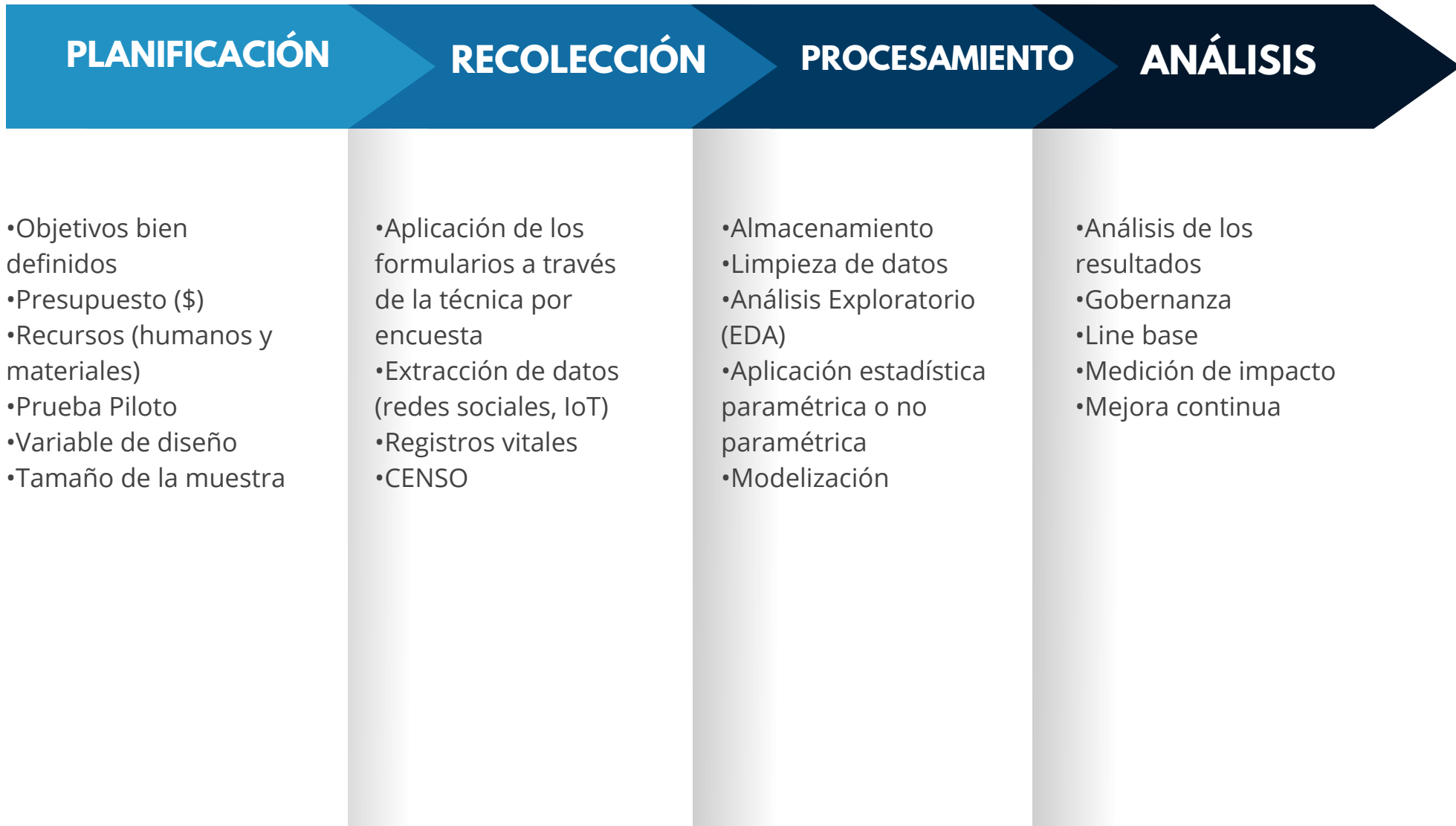
"SEGÚN TUKEY, EL ANÁLISIS EXPLORATORIO DE DATOS (EDA) ES UN PROCESO DE EXAMINAR Y COMPRENDER LOS DATOS UTILIZANDO ESTADÍSTICA, GRÁFICOS Y OTRAS TÉCNICAS PARA EXPLORAR PATRONES, DESCUBRIR RELACIONES Y SACAR CONCLUSIONES ÚTILES PARA LA INCORPORACIÓN POSTERIOR EN LA MODELIZACIÓN Y LA TOMA DE DECISIONES. ".

— Tukey (1977)

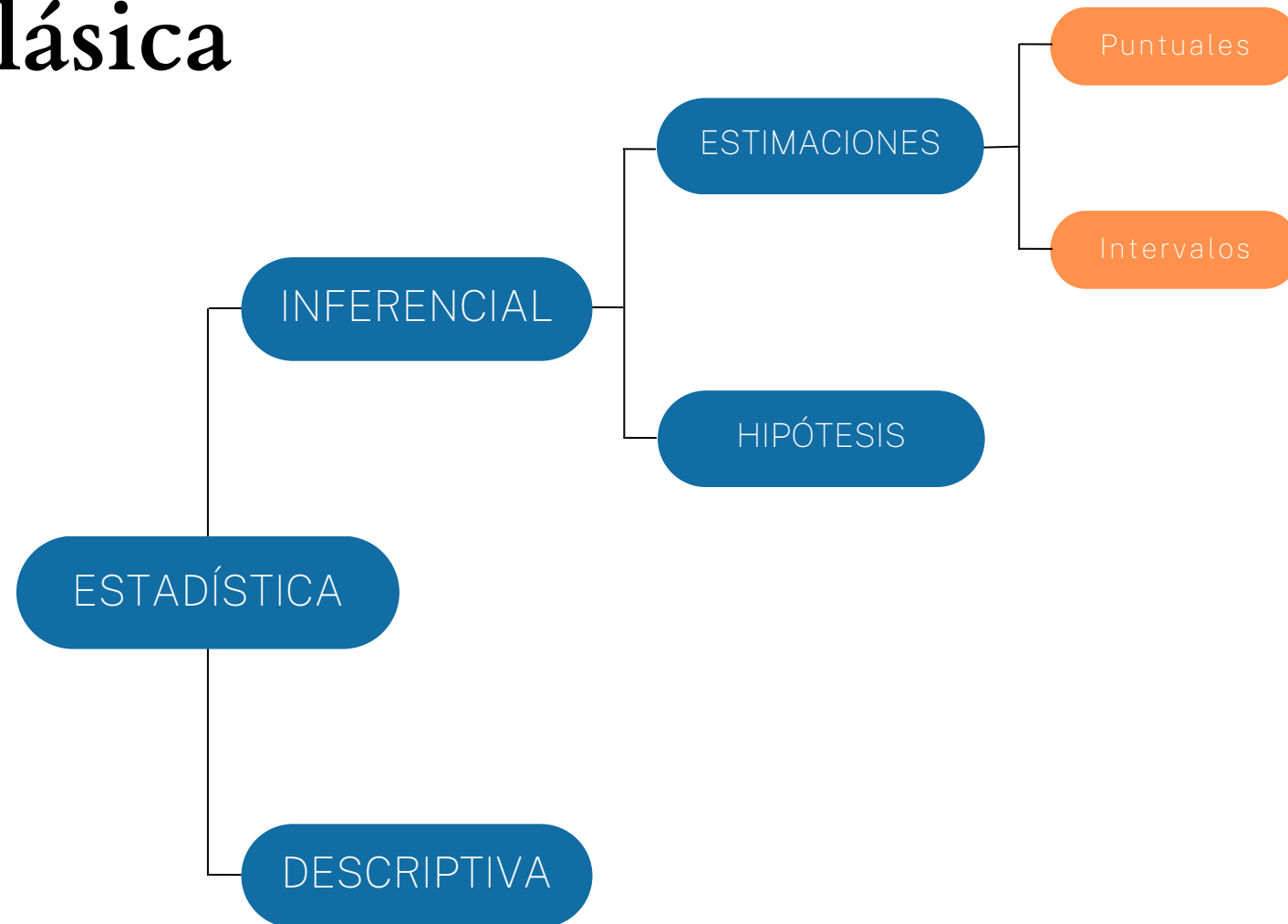
Competencia

Aplica la estadística descriptiva para la toma de decisión, utilizando de manera responsable las herramientas tecnológicas para el análisis de datos.

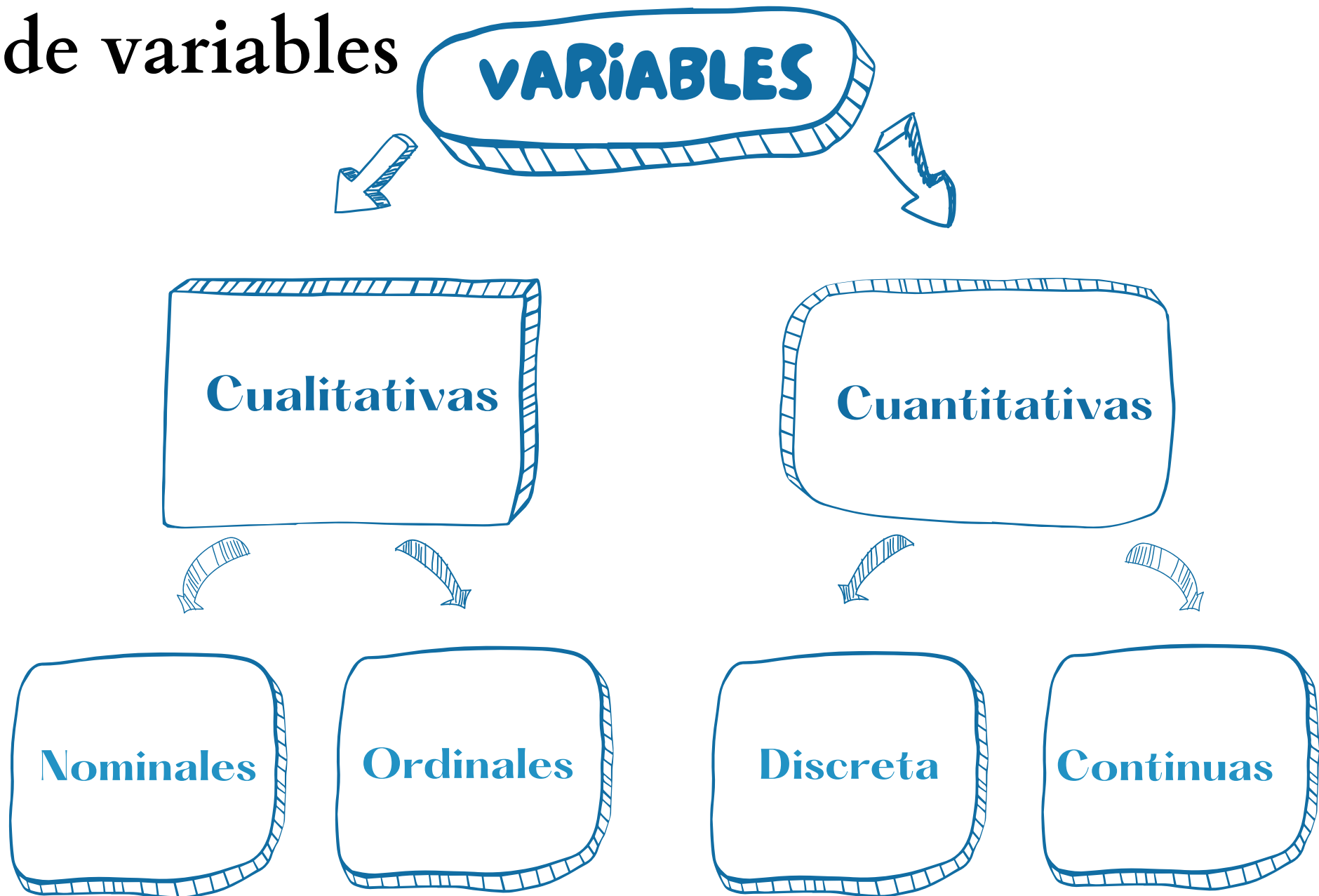
Estadística



División clásica



Tipos de variables



Tipos de variables

CONTINUOUS

measured data, can have ∞ values within possible range.



I AM 3.1" TALL
I WEIGH 34.16 grams

NOMINAL

UNORDERED DESCRIPTIONS



i'm a
TURTLE!



i'm a
Snail!—



i'm a
butterfly!

BINARY

ONLY 2 MUTUALLY
EXCLUSIVE OUTCOMES

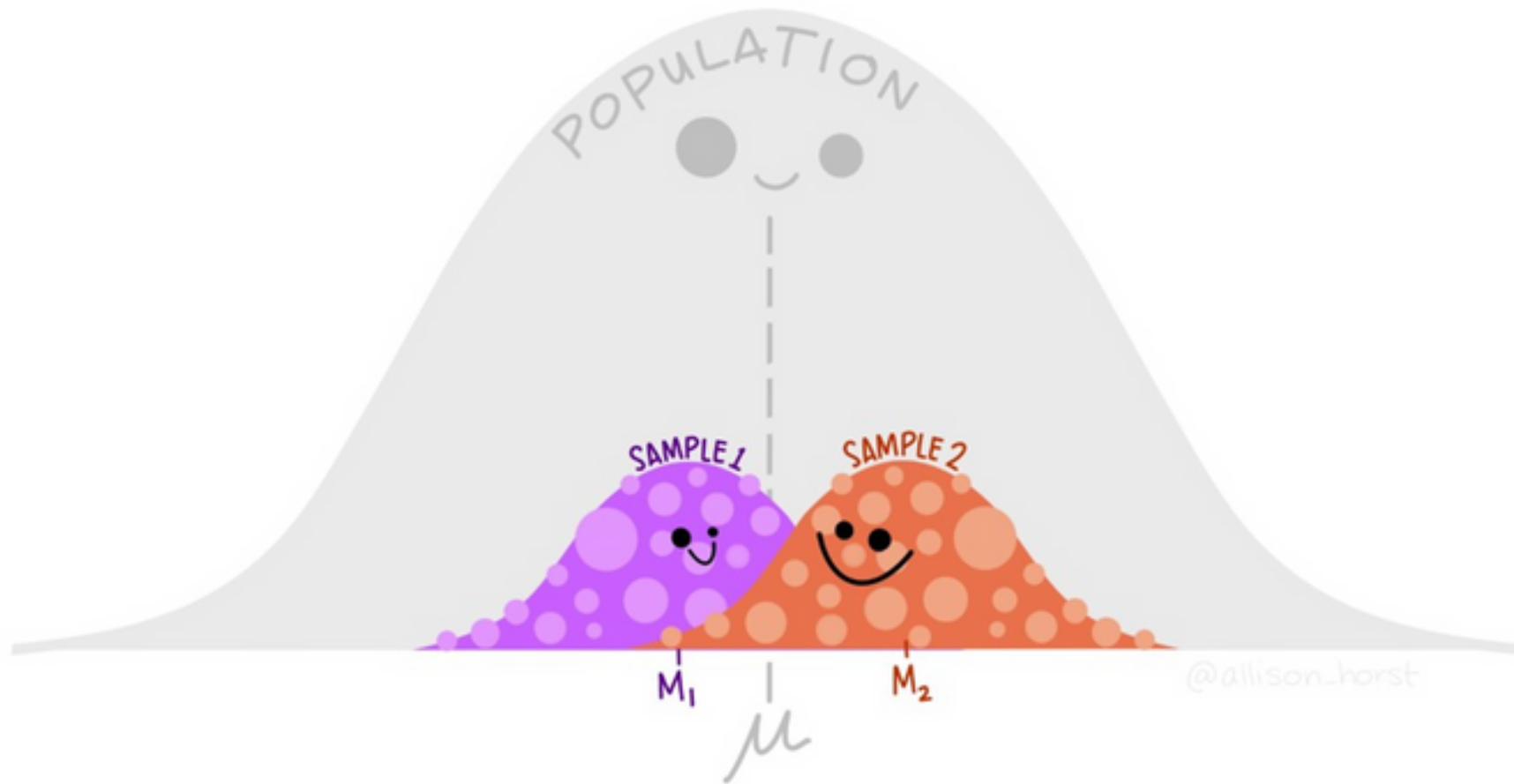


I AM
EXTINCT!—



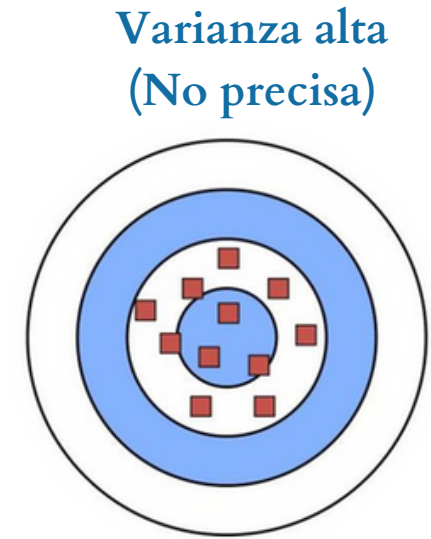
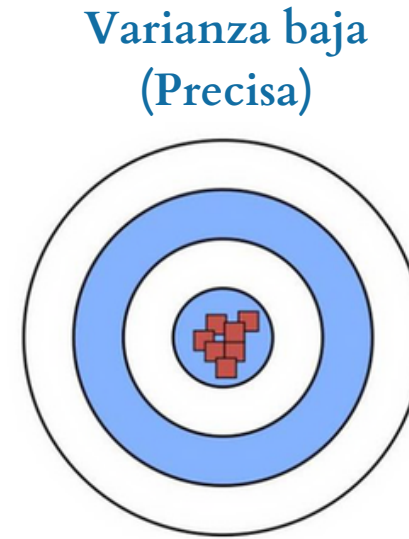
HA.

Población y muestra

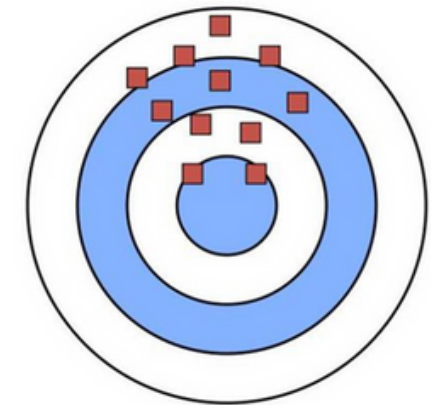
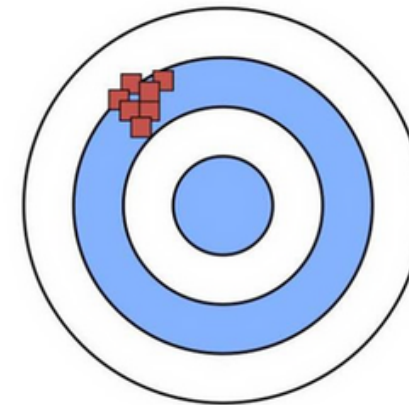


Descripción de datos

Sesgo bajo
(Precisa)

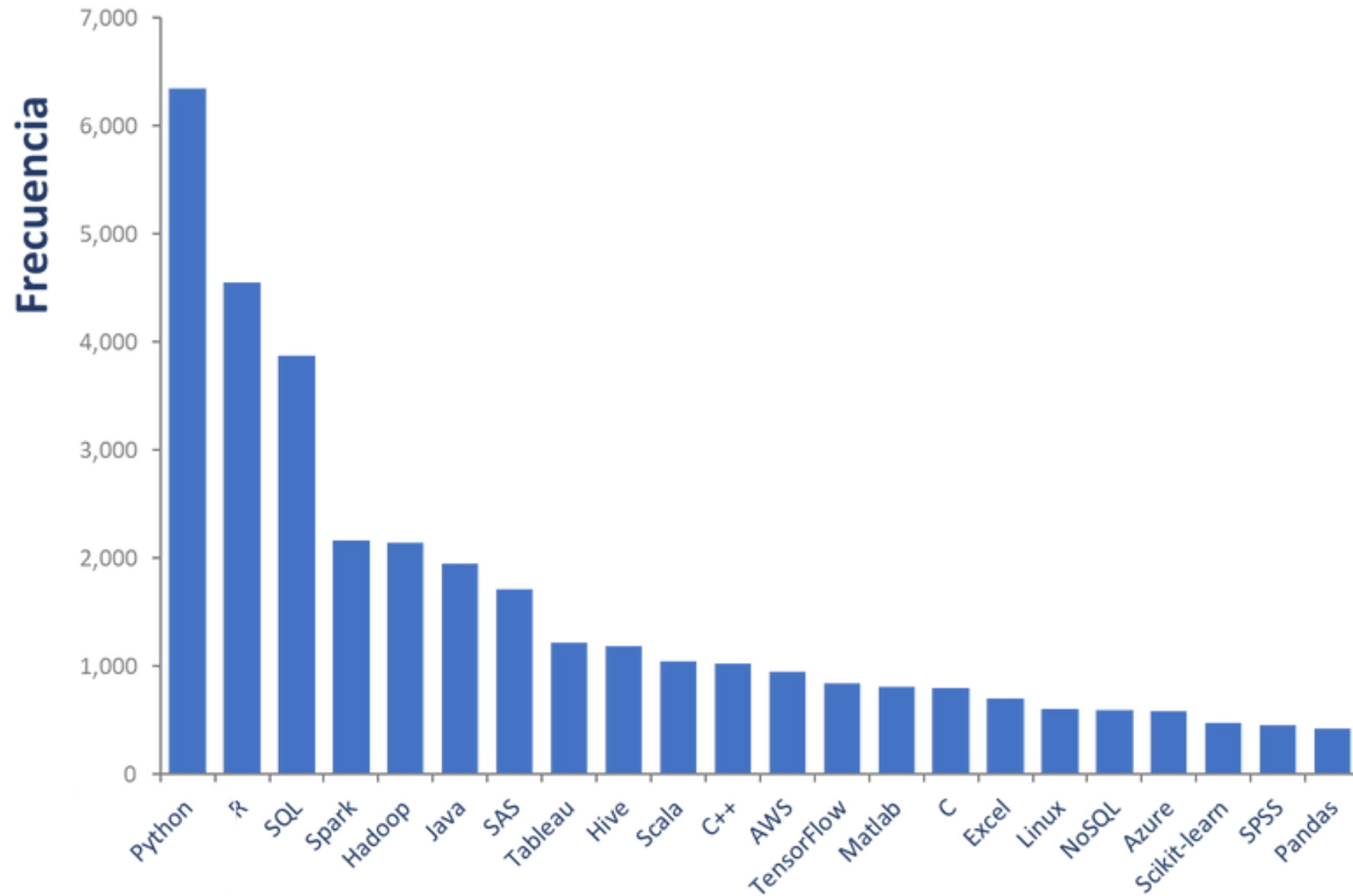


Sesgo alto
(No Precisa)



This work by Sebastian Raschka is licensed under a
Creative Commons Attribution 4.0 International License.

Software más usados



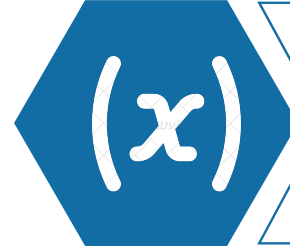
Métodos estadísticos



Tabulación



Métodos gráficos



Descripción de datos

Tabulación de datos categóricos

Tabla 1

Unidades agrícolas por niveles de bienestar

Pobreza	Año	Tamaño de la unidad Agrícola (manzanas)					Total
		<2	2 to 5	5 to 20	20 to 50	>50	
Pobre	2001	16.7	22.1	36.4	24.7	0	100
	2011	43.3	32.7	21.2	2.9	0	100
No pobre	2001	0	0	0	0	100	100
	2011	0	0	30.1	30.4	39.5	100
Total	2001	12.3	16.3	26.9	18.2	26.2	100
	2011	25.4	19.2	24.9	14.3	16.3	100

Source: Castro-Leal and Laguna (2015) using CENAGRO 2001 and 2011

Tabulación de datos categóricos

Tabla 2

Distribución de la población según nivel de alfabetismo por macro región

Macro Región	EMNV 2009			Total	EMNV 2014			Total
	Lee y escribe	Solo sabe leer	No sabe ni leer ni escribir		Lee y escribe	Solo sabe leer	No sabe ni leer ni escribir	
- Managua	57.8	0.5	4.1	62.4	36.9	0.5	2.4	39.7
- Pacífico	14.1	0.2	1.9	16.3	20.0	0.3	1.8	22.1
- Central	8.9	0.1	2.3	11.4	20.0	0.3	2.6	22.9
- Atlántico	7.2	0.3	2.5	10.0	12.7	0.4	2.2	15.3
Total	88.0	1.1	10.9	100.0	89.6	1.4	9.1	100.0

Nota: Se refleja el porcentaje de la población por nivel de alfabetismo según macro región.

Datos abiertos. (INIDE- EMNV's 2011,2016).

Visualización

→
Analizar

Estadísticos descriptivos

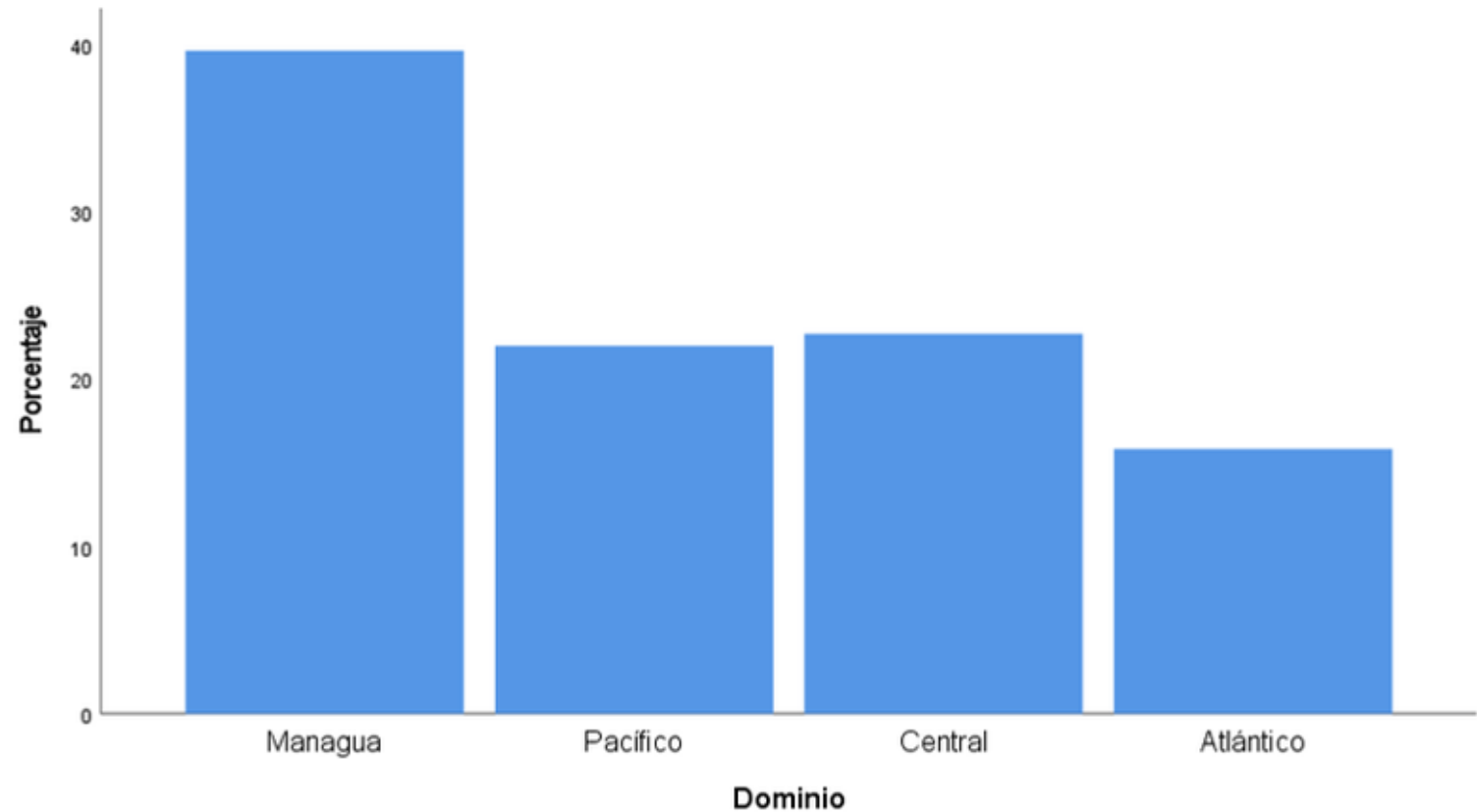
Variables

Dominio

Gráficos

Gráficos de barras

Figura 1. Dominio de estudio



Visualización

Figura 2. Área de residencia

→
Analizar

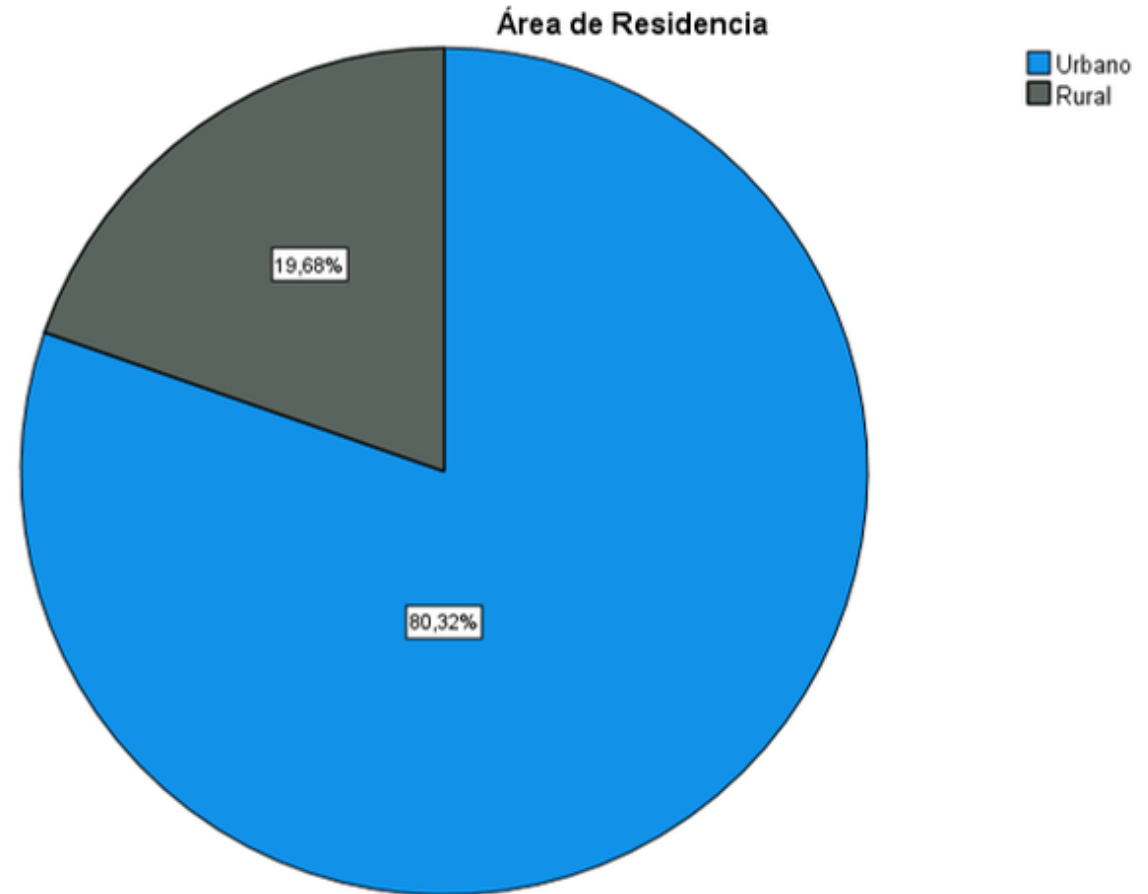
Estadísticos descriptivos

Variables

Área de residencia

Gráficos

Gráficos circulares



Visualización

Figura 3. Área de residencia según sexo

Gráficos

Cuadro de diálogos antiguos

Barras

Agrupadas

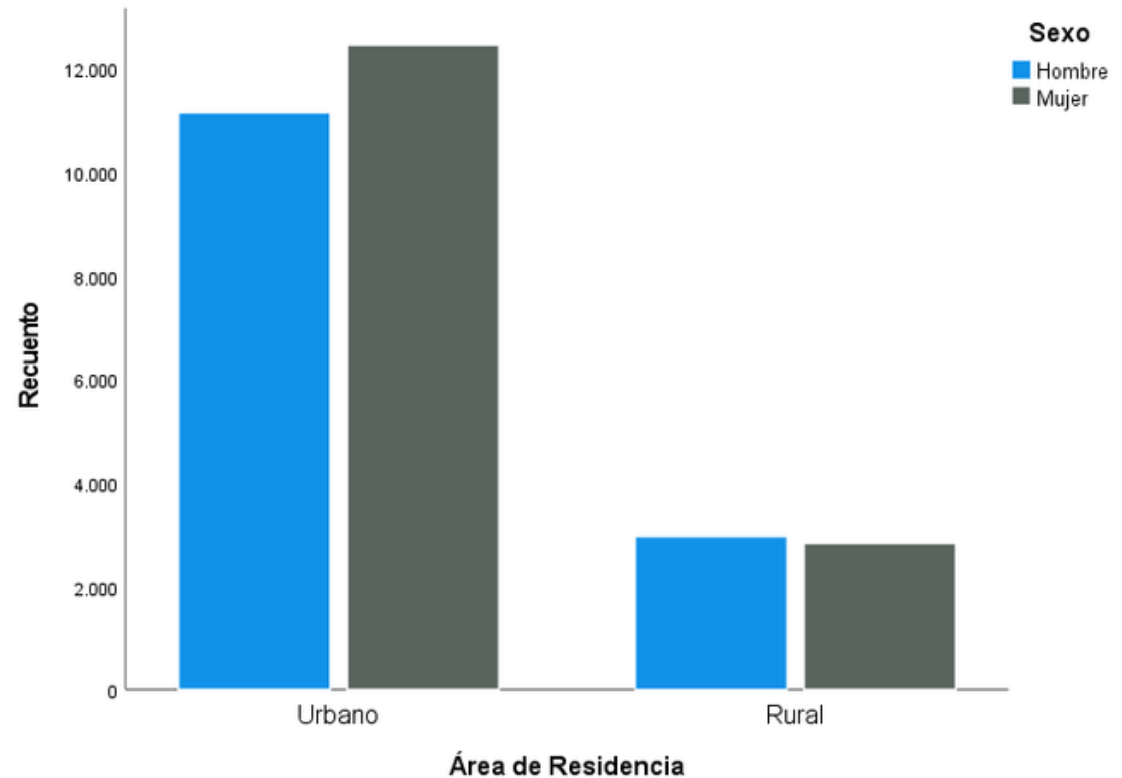
Resúmenes para grupos de caso

Eje de categorías

Área de residencia

Definir grupos por

Cuál es el sexo de



Visualización

Figura 4. Distribución de la población por sexo y edades simple

Gráficos

Cuadro de diálogos antiguos

Barras

Agrupadas

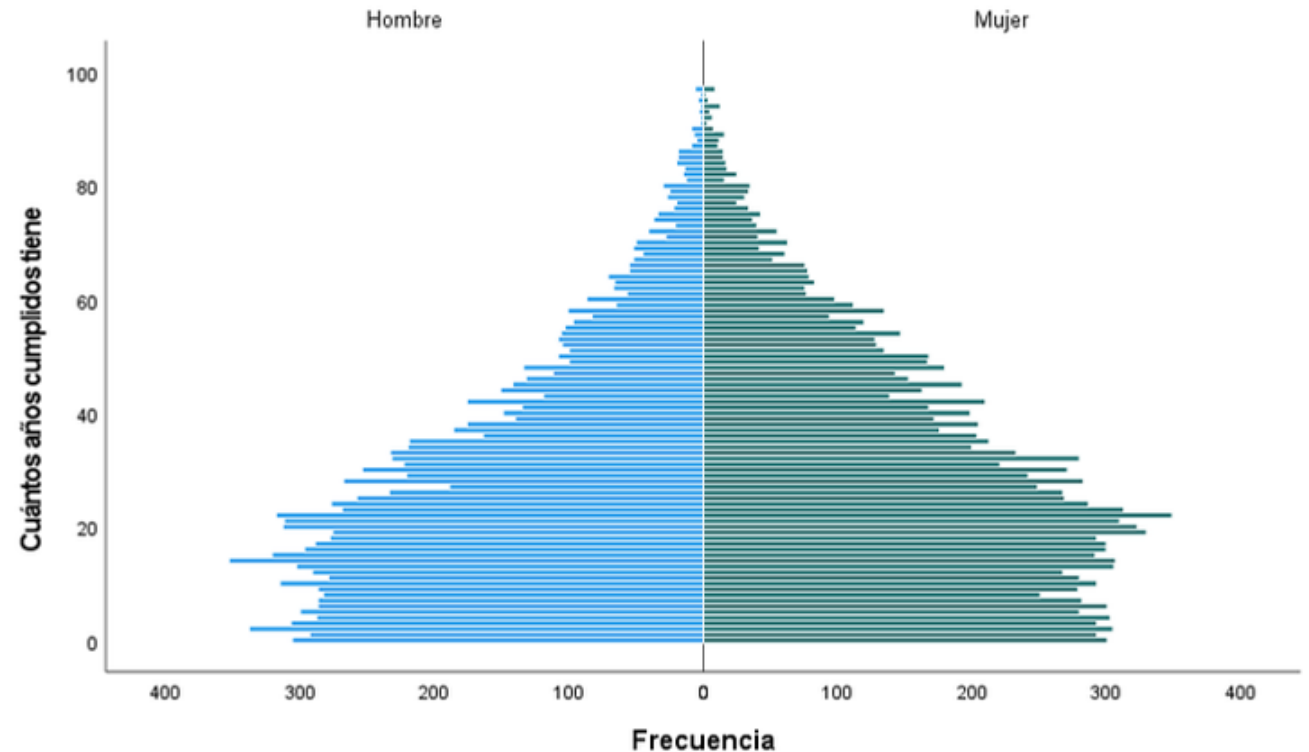
Resúmenes para grupos de caso

Eje de categorías

Área de reasidencia

Definir grupos por

Cuál es el sexo de



Visualización

Figura 5. Edades simple según dominio de procedencia

Gráficos

Cuadro de diálogos antiguos

Diagrama de cajas

Simples

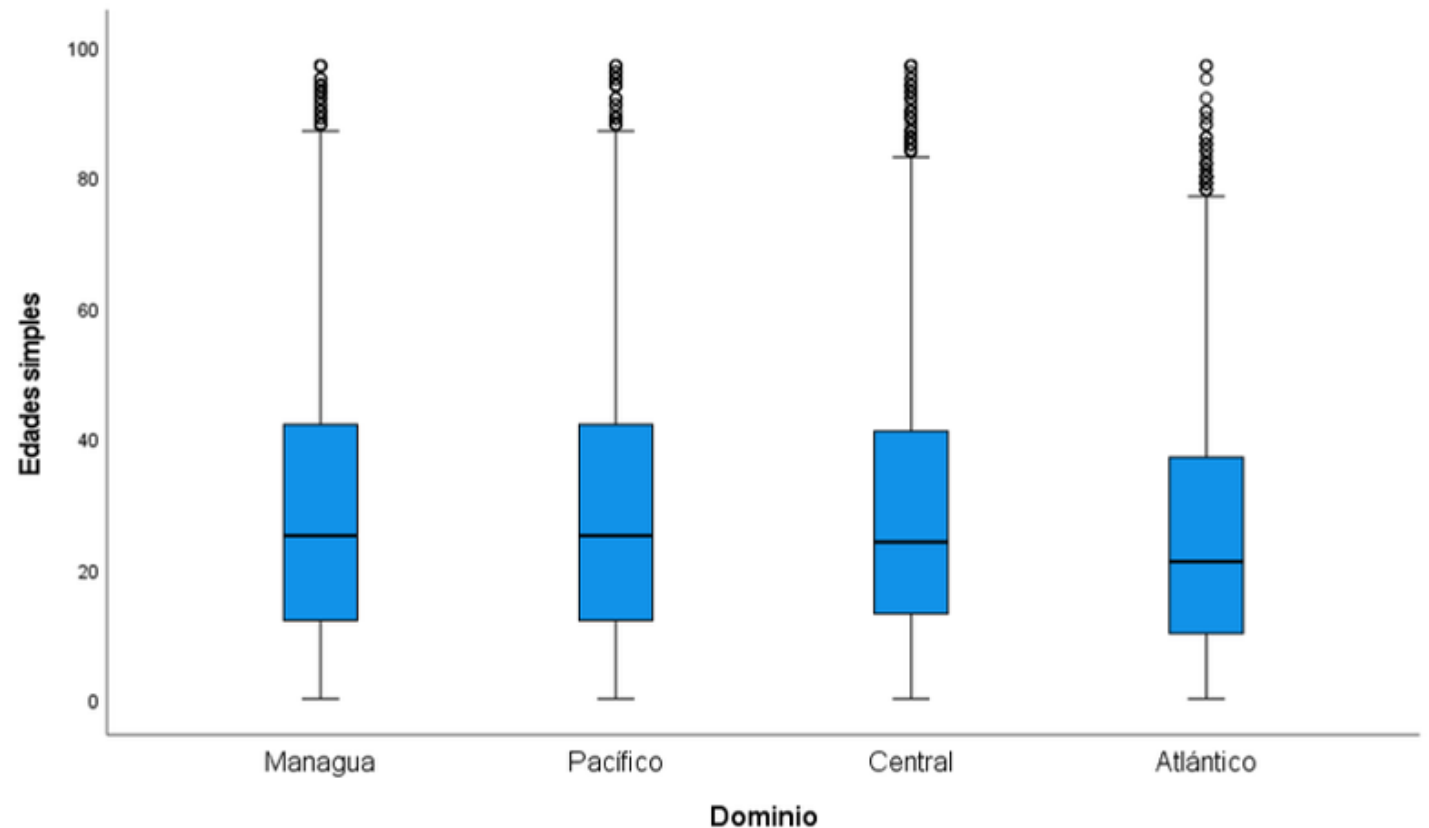
Resúmenes para grupos de casos

Variables

Cuántos años cumplido tiene

Eje de categoría

Dominio



Visualización

Figura 6. Precio según año de los vehículos

Gráficos

Cuadro de diálogos antiguos

Dispersión/puntos..

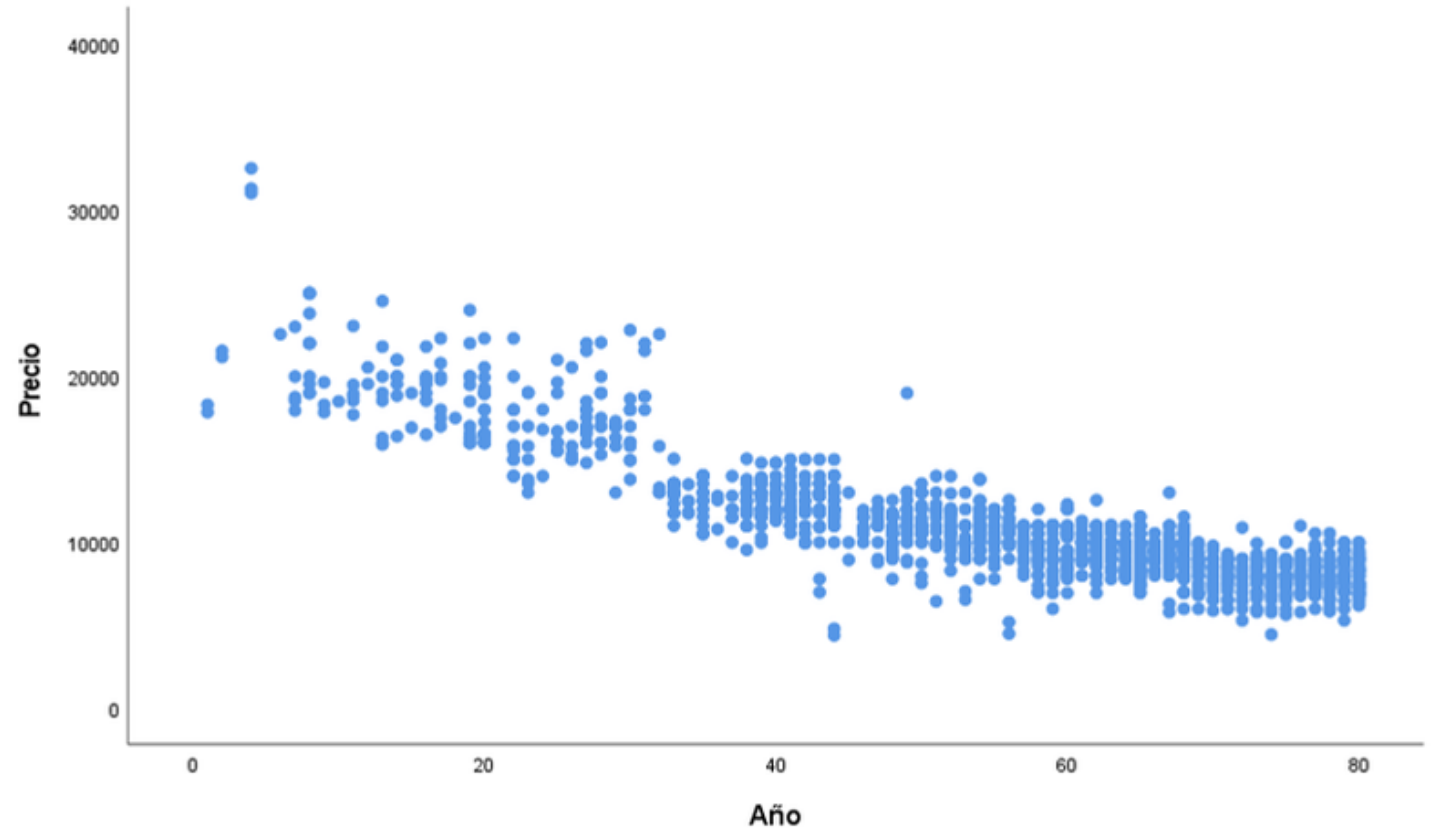
Dispersión simple.

Eje y

Prece

Eje x

Age



Descripción de datos



Analizar

Estadísticos descriptivos

Frecuencia

Variables

Price

Tendencia central

Media

Mediana

Tabla 3.

Medidas de tendencia central

Estadísticos

		Price	KM
N	Válido	1436	1436
	Perdidos	0	0
Media		10730,82	68533,26
Mediana		9900,00	63389,50

Descripción de datos



Analizar

Estadísticos descriptivos

Frecuencia

Variables

Price

Tendencia dispersión

Desv. Desviación

Varianza

Mínimo

Máximo

Tabla 3.
Medidas de dispersión

Estadísticos		Price	KM
N	Válido	1436	1436
	Perdidos	0	0
Desv. Desviación		3626,965	37506,449
Varianza		13154872,100	1406733707.0
Mínimo		4350	1
Máximo		32500	243000

Referencias

1. Binek, R. (2015). Kosaciec szczecinkowaty Iris setosa [Image]. Retrieved from https://commons.wikimedia.org/wiki/File:Kosaciec_szczecinkowaty_Iris_setosa.jpg#/media/File:Kosaciec_szczecinkowaty_Iris_setosa.jpg
2. Chihara, L. M., & Hesterberg, T. C. (2018). *Mathematical Statistics with Resampling and R* (2nd ed.). Wiley.
3. Kloeke, J., & McKean, J. W. (2014). *Nonparametric Statistical Methods Using R (Chapman & Hall/CRC The R Series Book 25) (English Edition)* (1.^a ed.). Chapman and Hall/CRC.
4. González, G. C., Liste, V. A., & Felpeto, B. A. (2011). *Tratamiento de datos con R, Statistica y SPSS* (1.^a ed.). Ediciones Diaz de Santos.
5. Rasch, D., Pilz, J., Verdooren, L. R., & Gebhardt, A. (2011). *Optimal Experimental Design with R (English Edition)* (1.^a ed.). Chapman and Hall/CRC.
6. Husson, F., Le, S., & Pagès, J. (2017). *Exploratory Multivariate Analysis by Example Using R* (2nd ed.). CRC Press.