



Programa de formación MACHINE LEARNING AND DATA SCIENCE MLDS

Facultad de
INGENIERÍA



Módulo 1

Análisis y visualización de datos con Python

Unidad 2

Análisis de datos con Pandas
Clase sincrónica

Felipe Restrepo Calle, PhD.

Facultad de
INGENIERÍA

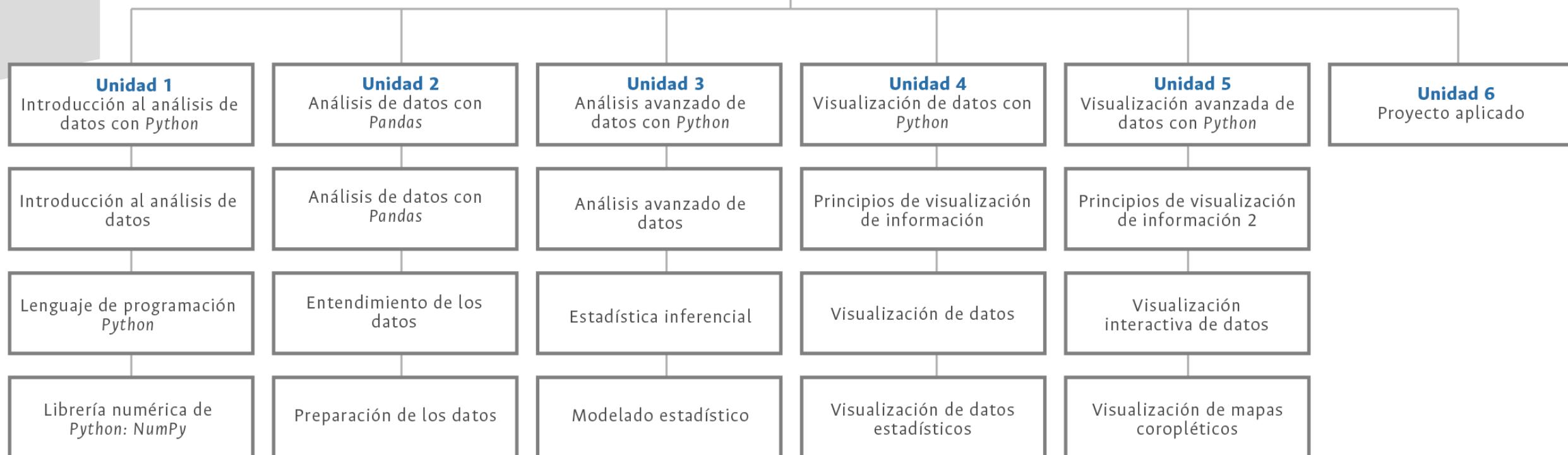




Mapa de contenidos

Módulo 1

Análisis y visualización de datos con Python





Análisis de datos con CRISP-DM





Agenda

1

Entendimiento de los datos

- Análisis exploratorio de datos
- Estadística descriptiva
- Visualización de datos

2

Preparación de los datos

- Limpieza de datos
- Selección de características
- Preprocesamiento y transformación de datos

1

Entendimiento de los datos



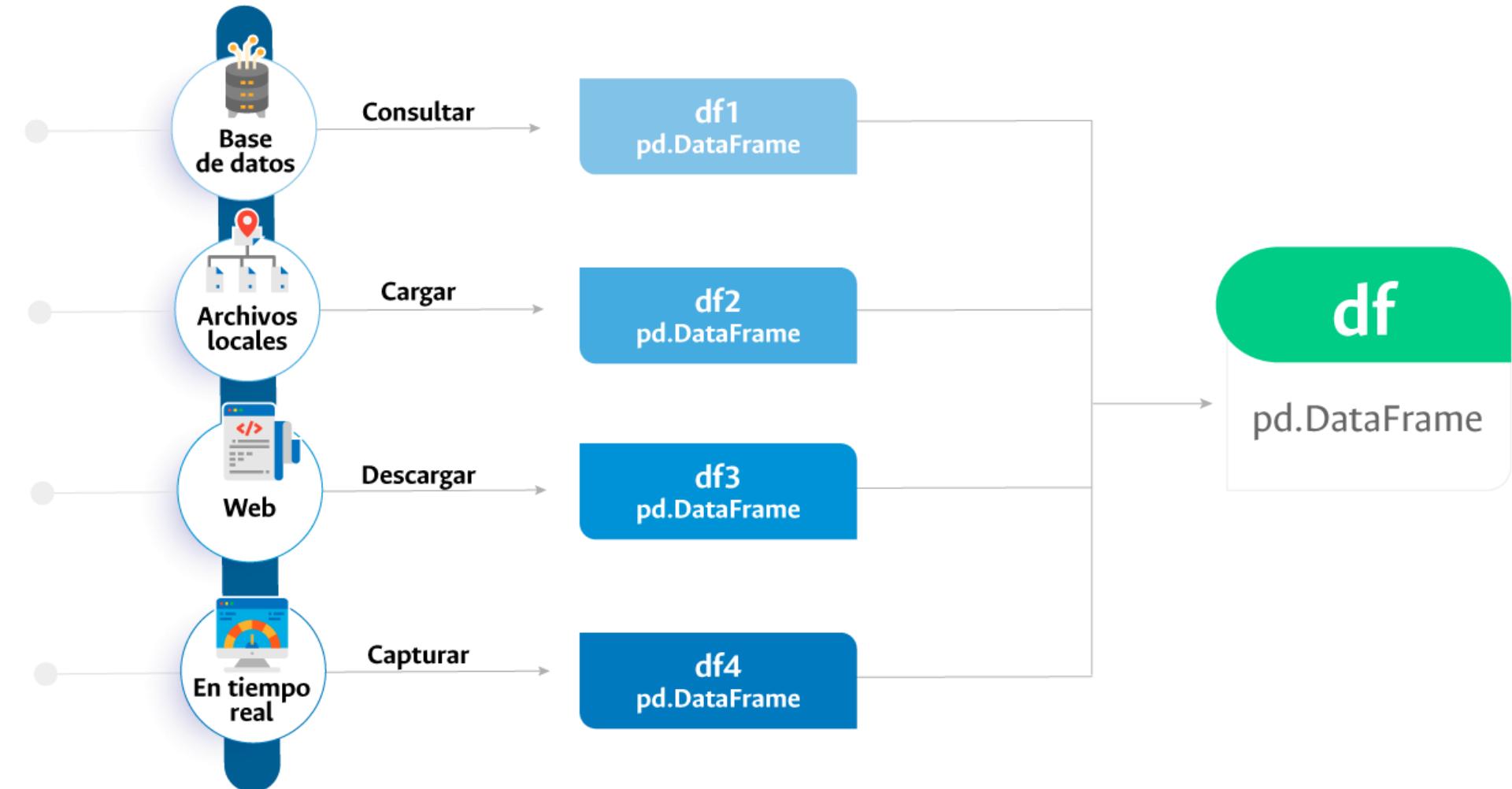
Entendimiento de los datos

Adquisición de datos

Identificar fuentes de datos

Recolectar los datos

Integrar los datos



Entendimiento de los datos

 Análisis exploratorio de datos

- Se realiza cuando se entra en contacto con los datos.
- Permite hacer una aproximación inicial a las características generales que se pueden identificar en los datos.
- En ocasiones, el análisis exploratorio detallado permite identificar soluciones a problemas sin necesidad de recurrir a modelos especializados.

Entendimiento de los datos

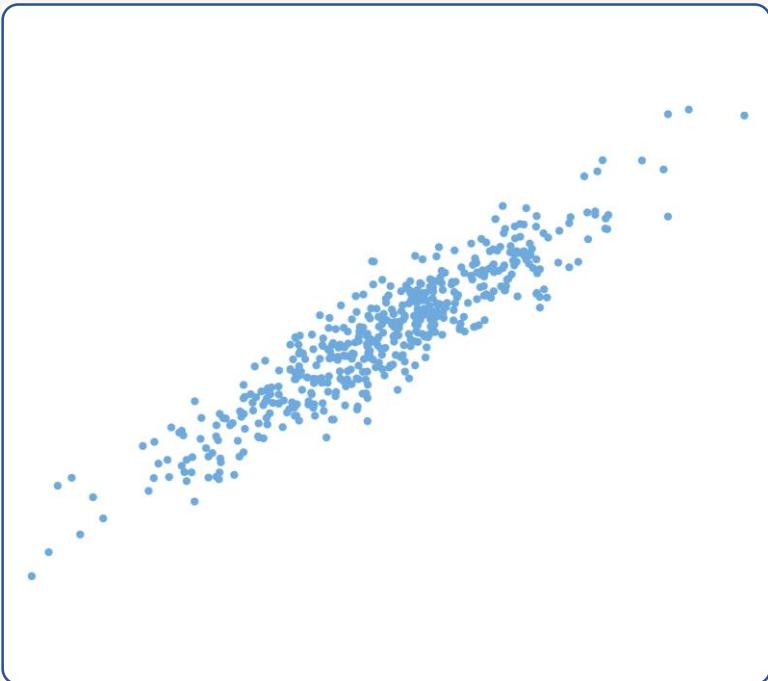


Análisis exploratorio de datos

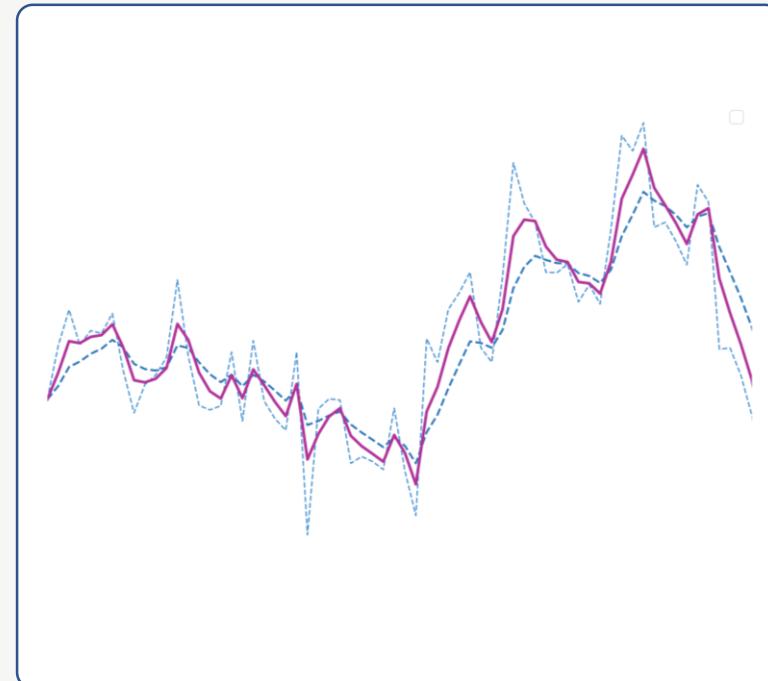
¿Qué se busca?



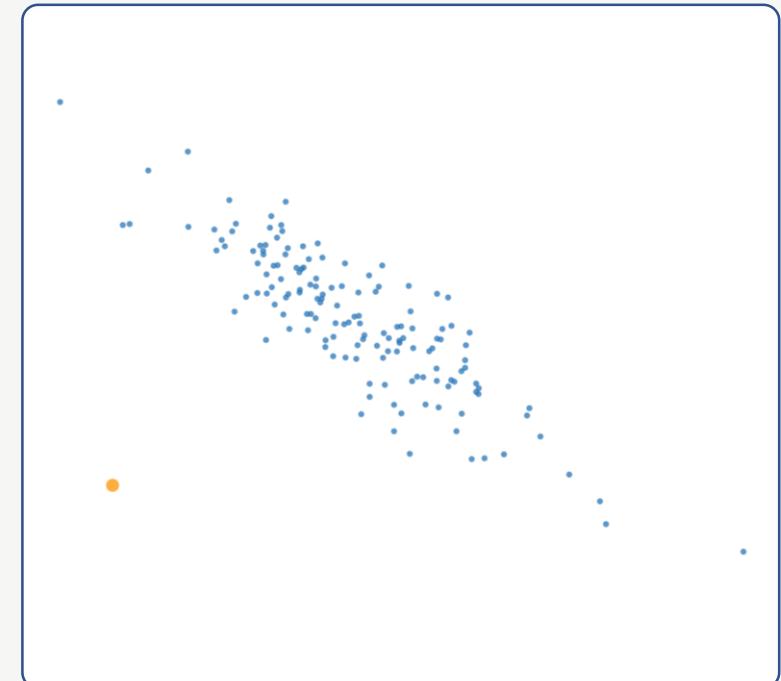
Correlaciones



Tendencias



Valores atípicos (outliers)



Entendimiento de los datos

 Análisis exploratorio de datos

¿Cómo se lleva a cabo?



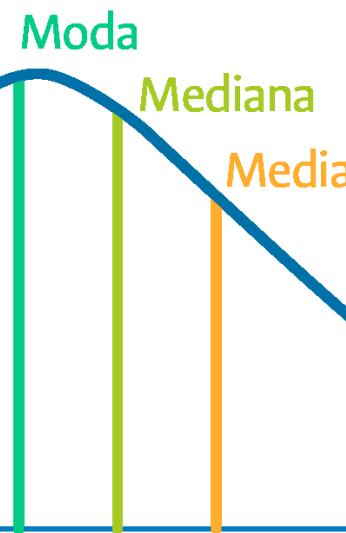
- **Estadística descriptiva:** Se encarga de organizar, caracterizar y presentar descripciones o resúmenes de los datos mediante:
 - **Medidas de posición**
 - **Medidas de forma**
 - **Medidas de dispersión**
 - **Medidas multivariadas**
- **Visualización de información:** campo de la ciencia de datos, que busca representar de manera gráfica la naturaleza de los datos a los que referencia.



Medidas de posición

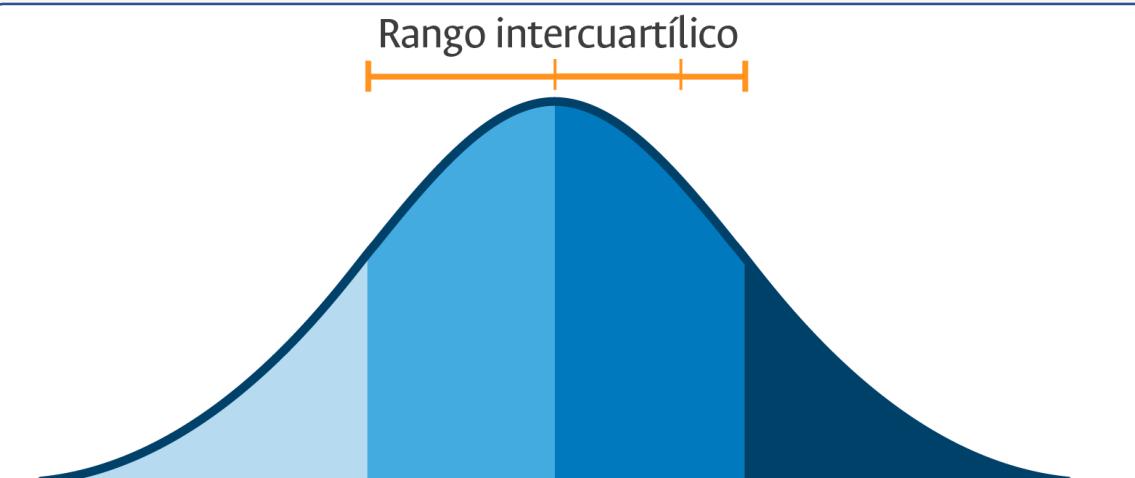


Medidas de tendencia central



Cuantiles

Permiten ilustrar la división de una distribución entre un número de grupos equidistantes de muestras. Los cuartiles, los deciles y los percentiles son los más comunes.

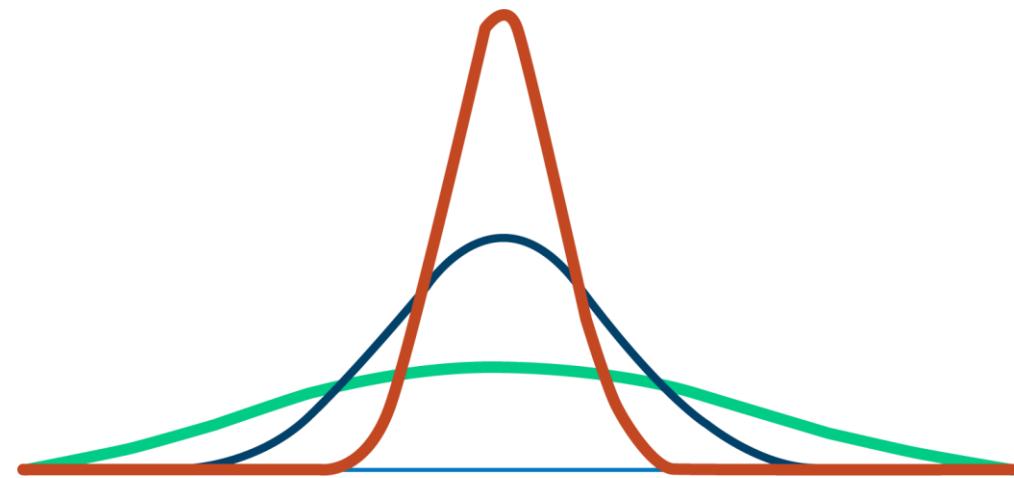




Medidas de forma

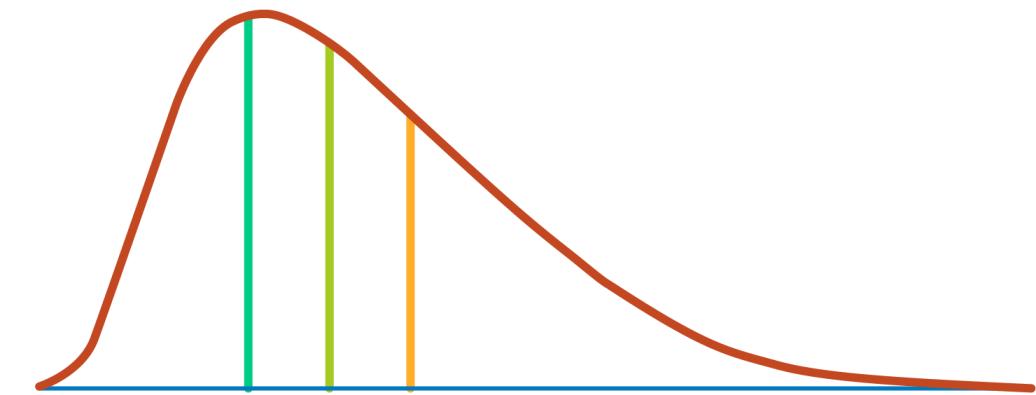
Curtosis

Representa el apuntamiento de una distribución y el grosor de sus picos, que refleja la concentración de los valores cerca de la media. **Positivo:** rojo; **negativo:** verde.



Asimetría o skewness

Representa la concentración o sesgo de los datos centrales de una distribución en una dirección particular. Se relaciona con el orden de las medidas de tendencia central. **Positivo:** izquierda; **negativo:** derecha.

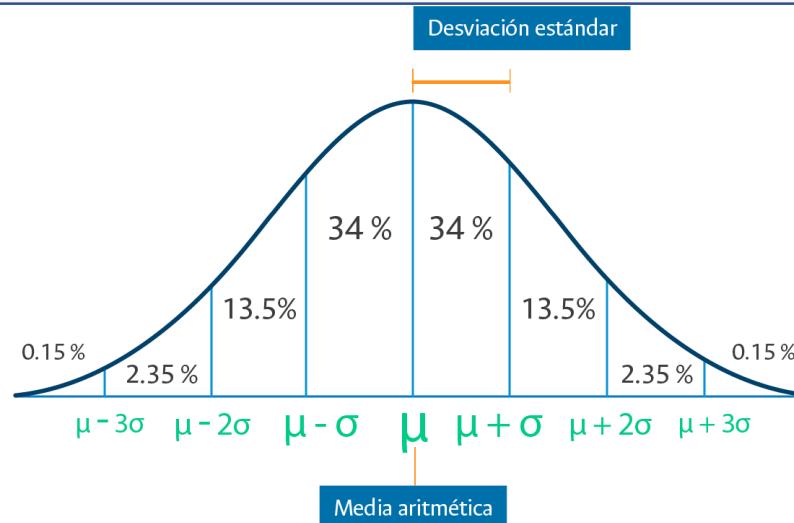




Medidas de dispersión

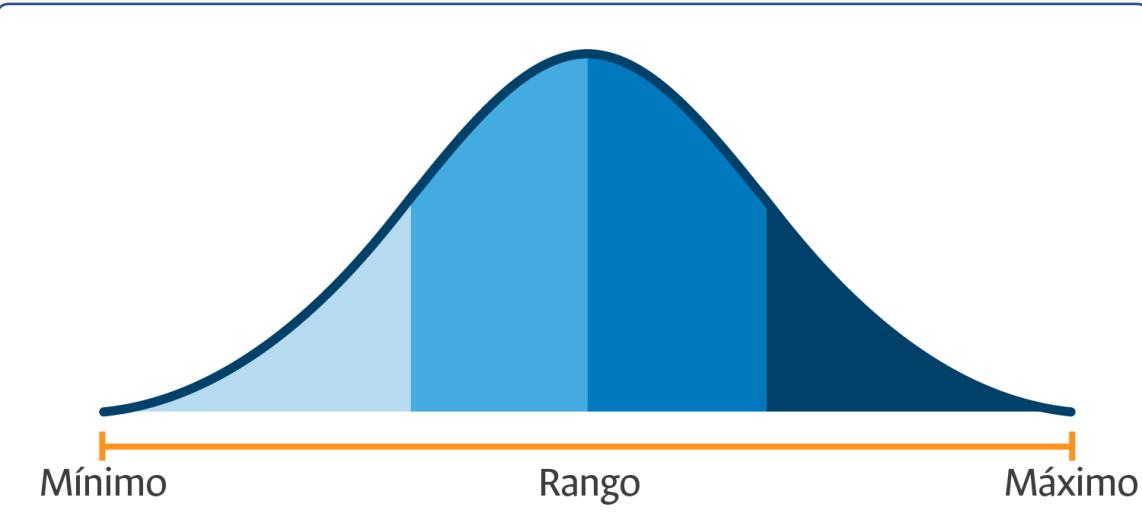
Desviación estándar y varianza

Representan la variación o distancia de los datos de una distribución con respecto a su media. Necesarias para entender la dispersión de las medidas, que no puede inferirse a partir de medidas centrales como la media.



Rango, mínimo y máximo

El rango es la diferencia entre el valor máximo y el valor mínimo de una distribución. Permiten caracterizar los extremos de una distribución y son útiles para encontrar valores atípicos.



Entendimiento de los datos



Estadística descriptiva

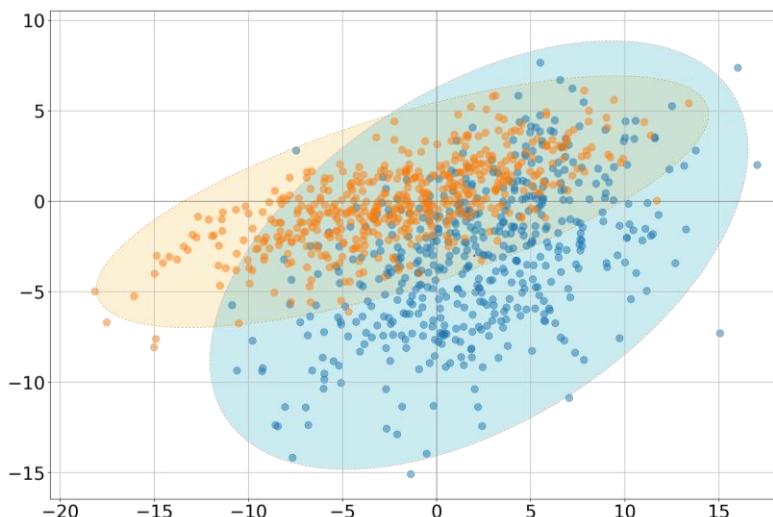




Medidas multivariadas

Correlaciones y covarianza

Medidas que representan la dependencia y variación conjunta de dos variables numéricas continuas. Son importantes para identificar relaciones entre características de los datos de interés.



Tablas de contingencia

Método tradicional para representar la relación entre dos variables cualitativas o categóricas. En esta tabla se caracterizan los conjuntos que comparten propiedades con cálculos como frecuencia o la media aritmética.

	Categoría X	Categoría Y	Categoría Z	Todos
Categoría A	25	12	26	63
Categoría B	18	20	16	54
Categoría C	9	12	16	37
Categoría D	15	23	16	54
Categoría E	16	17	11	44
Categoría F	18	21	8	47
Categoría G	14	24	14	52
Categoría H	14	17	16	47
Categoría I	16	11	20	47
Categoría J	18	21	16	55
Todos	163	178	159	500

Entendimiento de los datos

 **Visualización de datos****Tipos de visualizaciones comunes:**

- Histogramas
- Diagramas de cajas
- Gráfica de barras
- Gráfica circular
- Gráfica de líneas
- Gráfica de áreas
- Gráfica de dispersión
- Gráfica hexagonal

Entendimiento de los datos

 Visualización de datos

Histogramas

Representación de la distribución de variables numéricas en intervalos discretos. Útil para identificar valores atípicos, tendencias y representar la forma de la distribución de los datos.

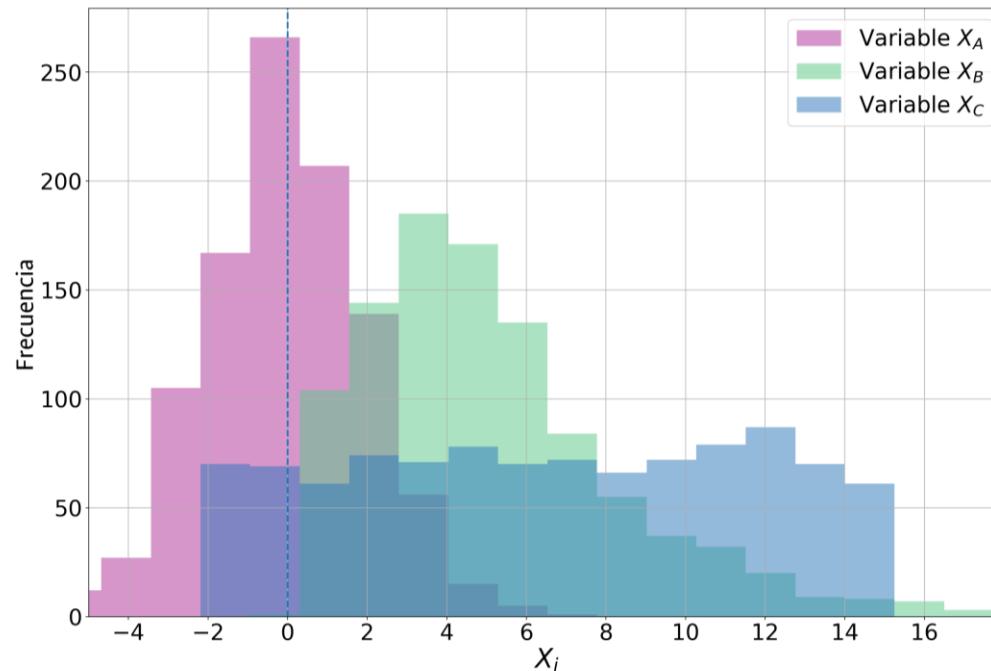
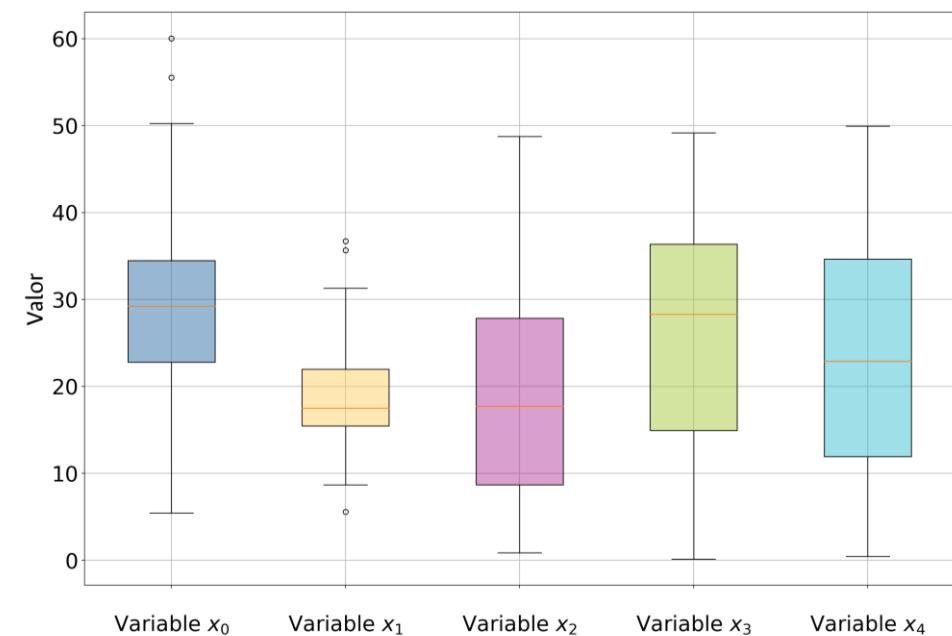


Diagrama de cajas

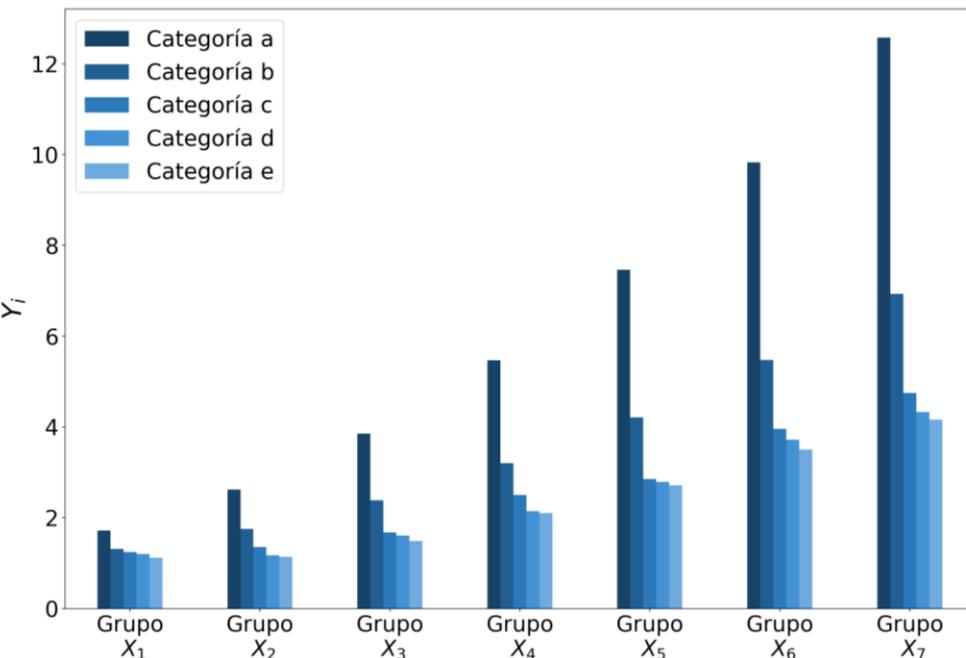
Representación que muestra a simple vista la mediana, los cuartiles, el mínimo, el máximo y el rango. También puede representar los valores atípicos.



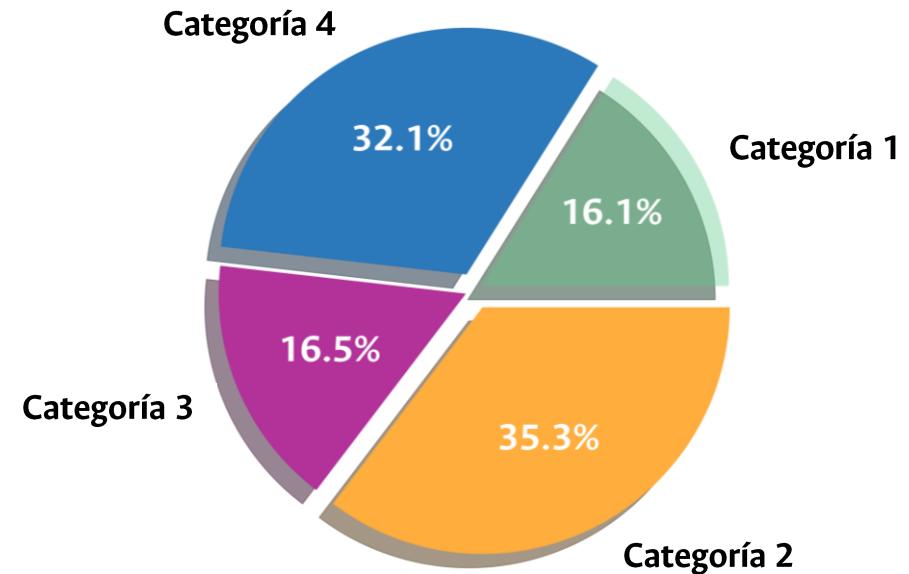
Entendimiento de los datos

 Visualización de datos Gráfica de barras

Representación de datos categóricos mediante barras con alturas proporcionales a los valores que representan. Se usa para hacer comparaciones de magnitud entre variables de cada categoría.

 Gráfica circular (torta)

Representación de las proporciones o porcentajes de los valores en una variable categórica. Permite visualizar de forma sencilla la proporción numérica de los posibles valores de una variable.



Entendimiento de los datos



Visualización de datos



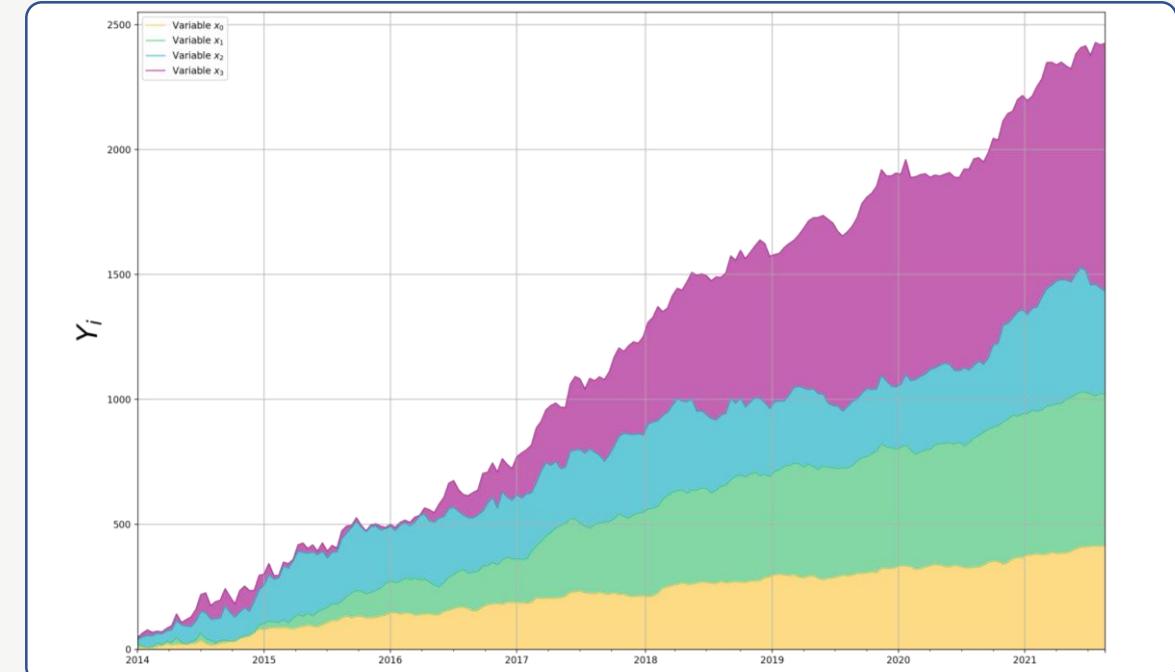
Gráfica de líneas

Representa la evolución de una variable numérica en relación a otra variable ordenada (e.g., tiempo). Permite identificar patrones o tendencias en los datos y comparar la evolución entre las variables.



Gráfica de áreas

Representa la proporción o magnitud de varias variables en relación a una variable ordenada, como el tiempo. Es usada para comparar la evolución y magnitud de distintas variables.



Entendimiento de los datos

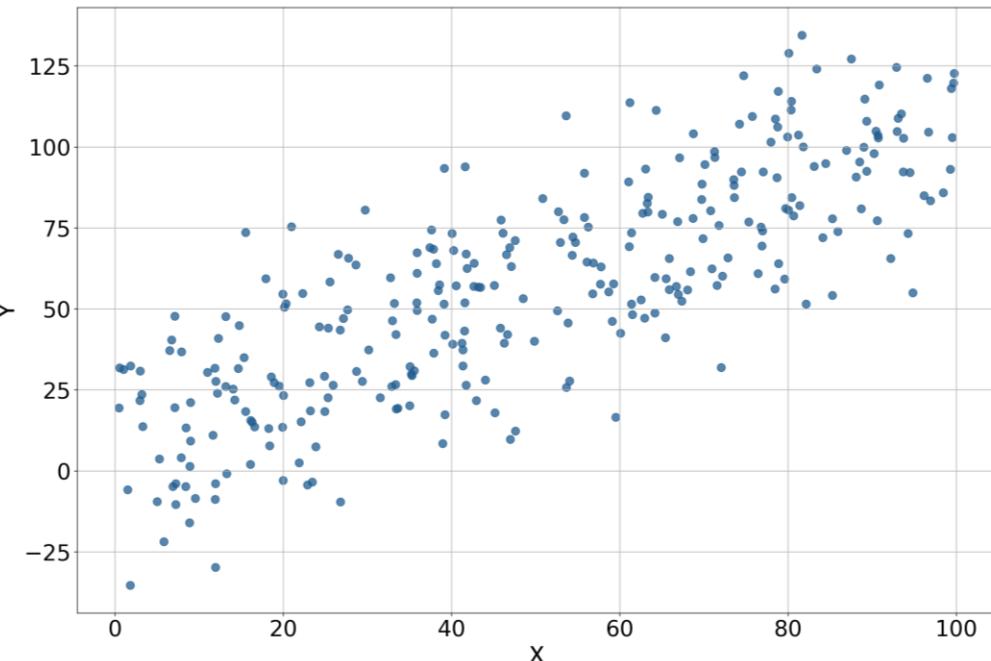


Visualización de datos



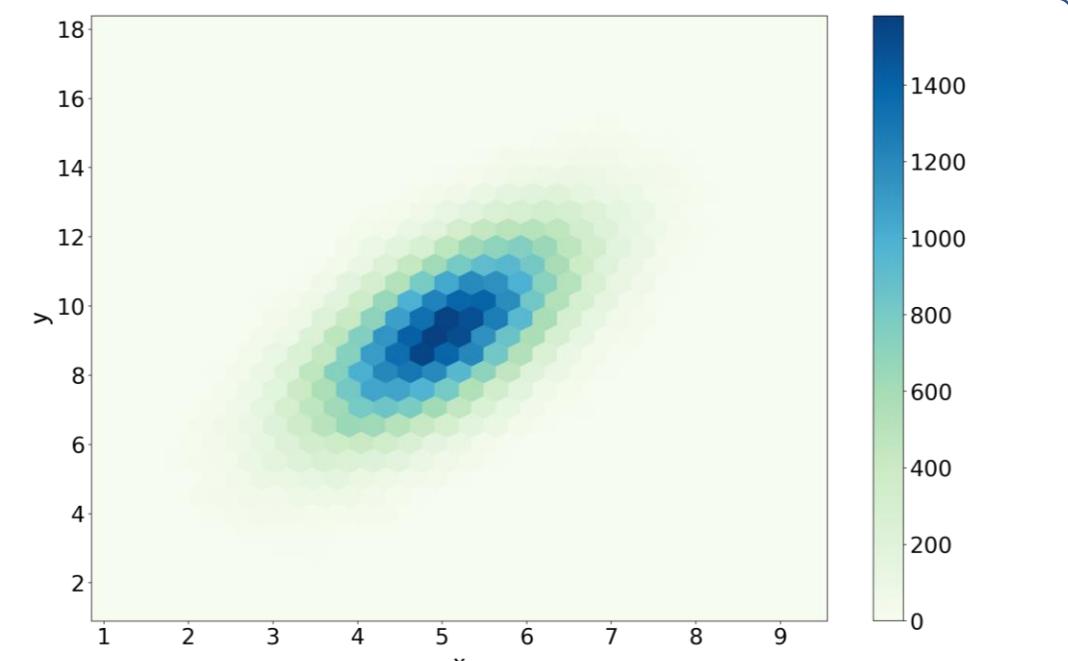
Gráfica de dispersión (scatter)

Representación espacial de la relación entre dos variables cuantitativas. Es usada principalmente para identificar e ilustrar relaciones entre variables.



Gráfica hexagonal

Representación de la frecuencia de valores dentro de intervalos en dos variables distintas. Se considera como un histograma de dos dimensiones. Permite identificar concentraciones de valores asociados en dos variables. Útil para datos con alta densidad.



2

Preparación de los datos



Preparación de los datos



- Antes de analizar o modelar los datos, es importante realizar tareas relacionadas con la **preparación de los datos**:
 - Limpieza de datos
 - Selección de características
 - Preprocesamiento y transformación

Preparación de los datos



Limpieza de datos



Valores faltantes

Existen variables de carácter no obligatorio, que es posible que no se registren en una observación. De acuerdo con las **necesidades del problema**, puede ser necesario **eliminar** toda la fila o toda la columna. También, es posible **imputar** datos para completar los valores faltantes.

	Nombre	Saldo
1001	Andrea	2000000
1002	José	-
1003	Alberto	-
1004	María	540000



Datos duplicados

Es posible encontrar datos duplicados en un conjunto de datos. En este caso puede ser necesario **eliminar** alguno de los registros o **integrarlos** para no perder datos valiosos.

	Nombre	Apellido
2001	José	Peñalosa
2002	Ana	Rodríguez
2003	Antonio	Becerra
2004	antonio	becerra

Preparación de los datos



Limpieza de datos



Valores atípicos (outliers)

No siempre se requiere limpiar datos atípicos. Si es necesario, se pueden reemplazar o eliminar de manera que no se afecte el análisis. Es necesario intentar comprender la razón de sus valores inesperados. Por ejemplo, un carácter faltante o adicional debido a un error de digitación.

	Nombre	Estatura
3001	Andrea	175
3002	José	16
3003	María	1680
3004	Antonio	180



Datos inconsistentes

Al integrar datos, es posible encontrar **inconsistencias** en el formato o en la unidad de medida utilizada. Para esto, es importante **entender** la naturaleza de los datos para **modificar** los registros y obtener uniformidad y consistencia.

	Nombre	Fecha de nacimiento
4001	Andrea	1980-02-23
4002	José Díaz	1975-15-08
4003	María	12/03/1998
4004	Antonio	27-mar-1989

Preparación de los datos



Limpieza de datos

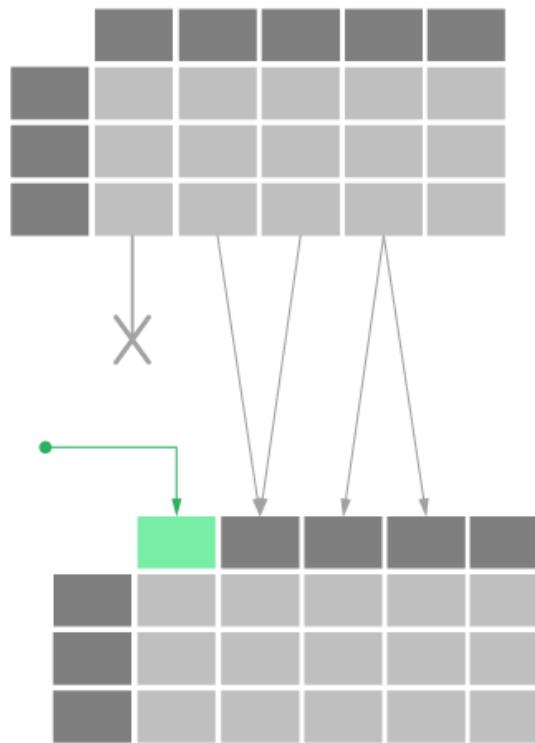
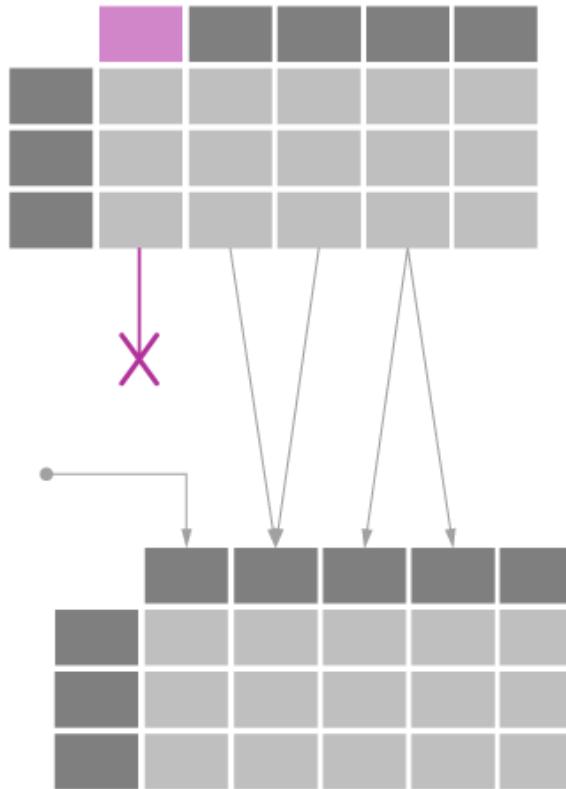
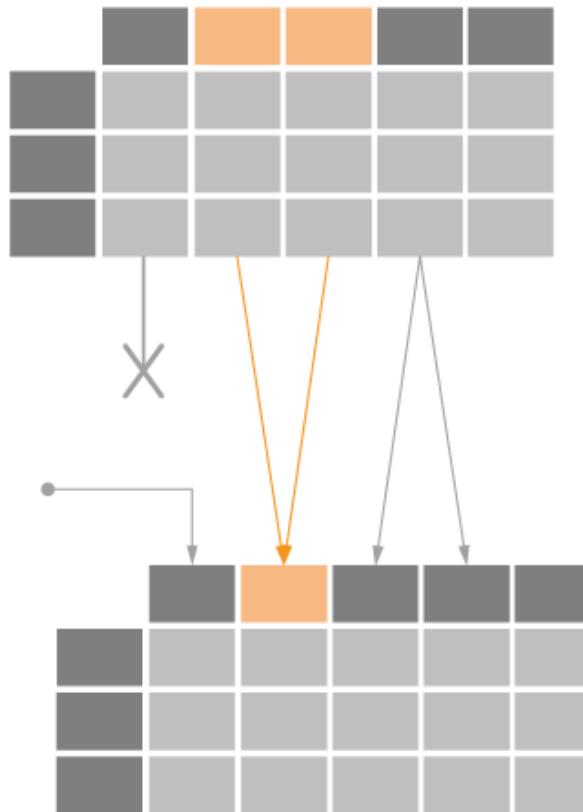
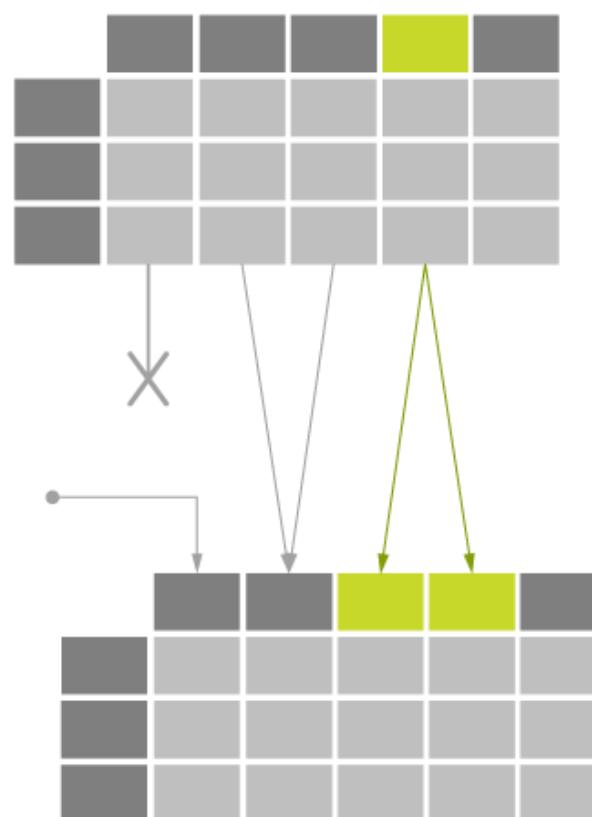


Datos corruptos o ruido

Es posible encontrar datos corruptos o ruido en los datos; si se encuentran, hay que intentar **recuperar** los datos, entendiendo las posibles causas del ruido como el formato de origen, la codificación o componentes físicos de almacenamiento en mal estado.

		Nombre	Apellido
5001		Andrea	Rodríguez
5002	'José'		Moreno
5003	'María'		Varón
5004	'Antonio\n'		Quiroga

Preparación de los datos

 Selección de características Añadir Eliminar Combinar Dividir

Preparación de los datos

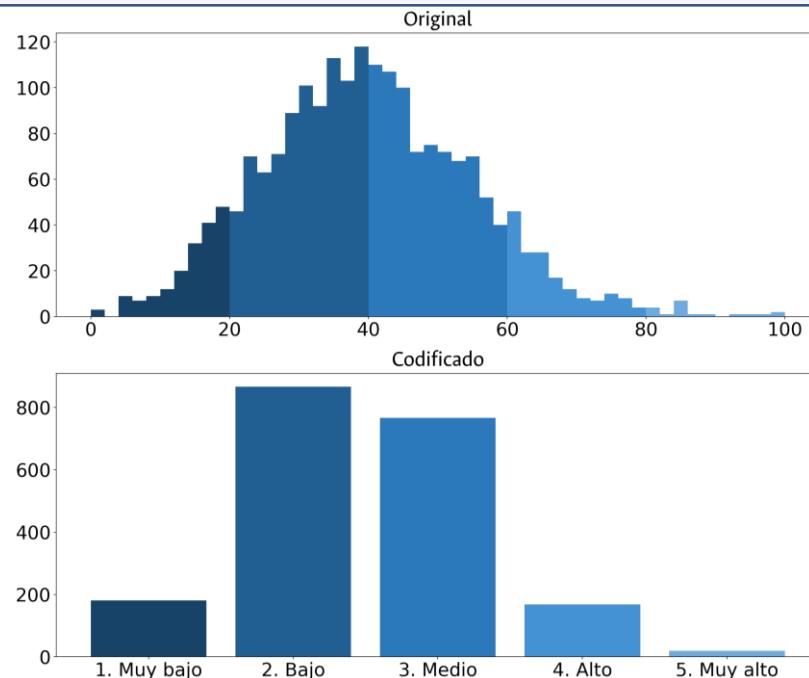


Preprocesamiento de datos



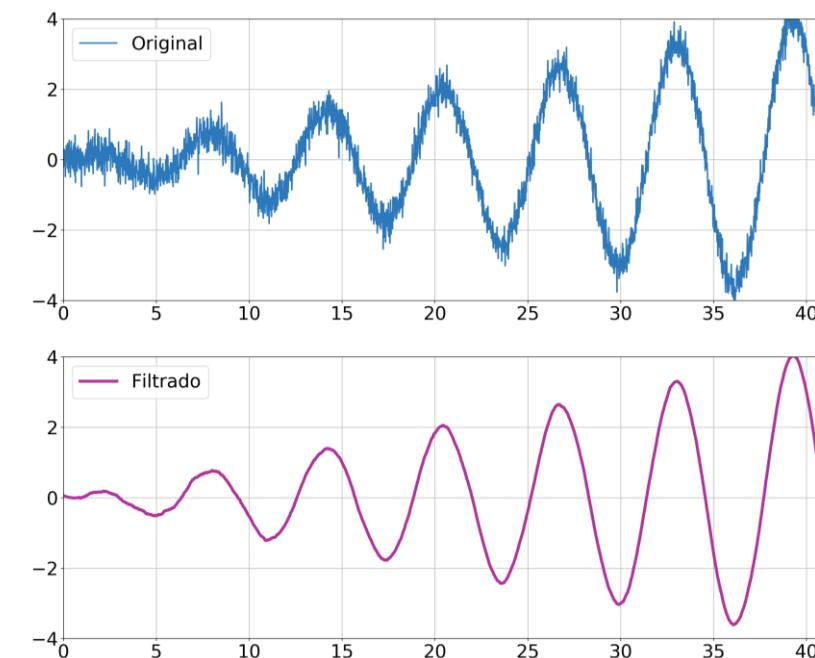
Codificar

Puede que sea necesario cambiar la codificación de algunas variables. Este es el caso de la **discretización** o **recodificación** de variables numéricas a categóricas como en los modelos de clasificación.



Filtrar

Es común encontrar componentes de **ruido** que pueden ser filtrados con métodos de limpieza de señales. Esto permite mejorar la precisión de los resultados.

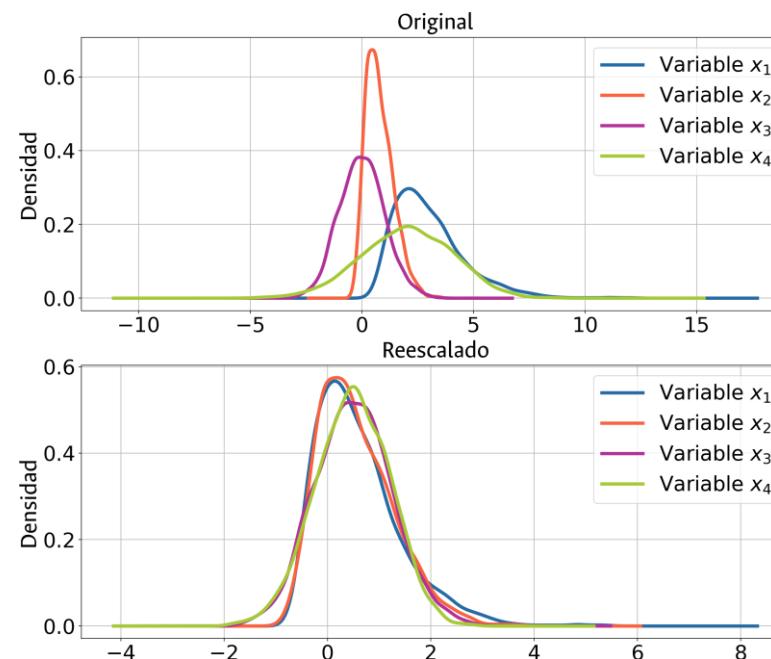


Preparación de los datos

 Preprocesamiento de datos

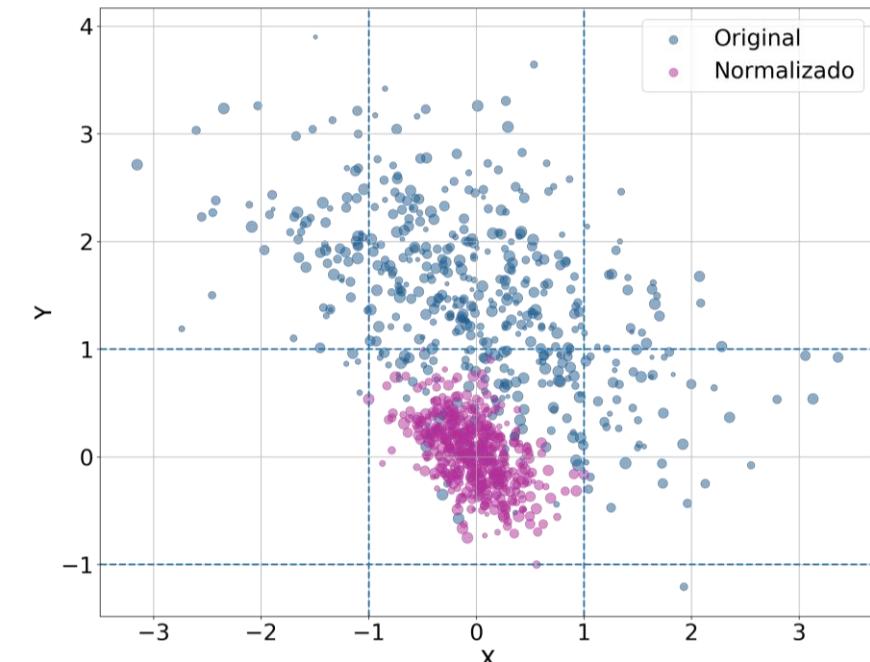
Escalar

Un método importante es el de re-escalado de datos, que consiste en alterar la dimensión de los datos, de modo que el rango de una o varias variables se limiten a un intervalo definido.



Normalizar

Esta es una tarea de redimensionado en la que se transforman los datos para que correspondan a una distribución con media aritmética centrada en 0 y desviación estándar de 1.





Análisis de datos con CRISP-DM



Logística

**Actividades académicas Unidad 2**

1

Objeto Virtual de Aprendizaje (OVA)

- Análisis de datos con Pandas



4

Documentos (pdf)

- Sitios para encontrar datasets
- Guía de referencia rápida de Pandas



2

Pandas - Entendimiento de los datos:

- Taller guiado (notebook)
- Quiz 3 (notebook)



5

Tarea 2 – Análisis de datos con Pandas

- Notebook



3

Pandas – Preparación de los datos:

- Taller guiado (notebook)
- Quiz 4 (notebook)



6

Campuswire:

- Dudas en los foros.
- Colaboración entre estudiantes.

Logística

 Proyecto aplicado

Análisis y visualización de datos con Python

- Unidad 0: Obligatorio
- Unidad 1: Introducción al análisis de datos con Python
- Unidad 2: Análisis de datos con Pandas
- Unidad 3: Análisis avanzado de datos con Python
- Unidad 4: Visualización de datos con Python
- Unidad 5: Visualización avanzada de datos con Python
- **Proyecto Aplicado**
- Evaluación



Referencias

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0. <ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/>

Gunderson, B., West, B.T. & Shedden, K. (s.f.). Understanding and Visualizing Data with Python [Comprender y visualizar datos con Python]. Universidad de Michigan. <https://www.coursera.org/specializations/statistics-with-python>

Nguyen, M. & Altintas, I. (s.f.). Machine Learning with Big Data [Aprendizaje automático con Big Data]. UC San Diego. <https://www.coursera.org/learn/big-data-machine-learning>

Diez, D., Çetinkaya-Rundel, M., Barr, C.D. (2019). OpenIntro Statistics. (4.ª ed.). OpenIntro. <https://www.openintro.org/book/os/>

Kapil, A. R. (2018) Data exploration and preparation. DataVedas. <https://www.datavedas.com/data-exploration-and-preparation/>

Kunin, D., Guo, J., Devlin, T.D., Xiang, D. (2019). Seeing Theory [Software] <https://seeing-theory.brown.edu/>

Recurso computacional en línea de la UCLA. (2014). Probabilidad y estadística Ebook. UCLA. <http://wiki.stat.ucla.edu/socr/index.php/EBook>



Créditos

Facultad de
INGENIERÍA

Autores

Felipe Restrepo Calle, PhD

Asistente docente

Alberto Nicolai Romero Martínez

Diseño instruccional

Claudia Patricia Rodríguez Sánchez

Diseño gráfico

Clara Valeria Suárez Caballero
Milton R. Pachón Pinzón

Diagramadora PPT

Daniela Duque

2022

