# INTRO A PYTHON PARA CIENCIA DE DATOS

LevelUP

# Clase 6:

01/06/2024

LevelUP

- EDA
- Hipótesis.
- Visualizaciones. Estadísticas
- Querys al Dataset

# EDA

# (EXPLORATORY DATA ANALYSIS)

LevelUP
Tech Academy

# EDA

Un **EDA (Exploratory Data Analysis, o Análisis Exploratorio de Datos)** es una etapa crucial en el proceso de análisis de datos.

- Conocer nuestro dataset
- Realizar limpiezas y transformaciones de datos
- Resumirlo utilizando técnicas estadísticas y gráficas
- Entrender estructura y relaciones entre los datos

# OBJETIVOS DEL EDA

- **Comprensión de los Datos:** Obtener una visión general de las variables y sus características.
- **Identificación de Patrones y Relaciones:** Descubrir correlaciones y patrones entre variables.
- **Detección de Anomalías:** Identificar valores atípicos o datos erróneos.
- **Verificación de Suposiciones:** Comprobar las suposiciones estadísticas que podrían guiar el análisis posterior.
- **Generación de Hipótesis:** Formular hipótesis iniciales basadas en observaciones preliminares.

# TÉCNICAS COMUNES EN EDA

- **Estadísticas Descriptivas:** Cálculo de medidas como la **media, mediana, moda, desviación estándar, mínimo, máximo, percentiles**, etc (conocidas como medidas de Tendencia Central).

- **Visualizaciones Gráficas**

- **Análisis de Valores Nulos:** Detectar y cuantificar la presencia de datos faltantes.

- **Análisis de Distribuciones:** Comparar las distribuciones de las variables con distribuciones teóricas.

- **Transformaciones de Datos:** Aplicar transformaciones para mejorar la interpretación y análisis de los datos.

# HIPÓTESIS

LevelUP
Tech Academy

# HIPÓTESIS

En el contexto del Análisis Exploratorio de Datos (EDA), las hipótesis juegan un papel crucial en guiar la exploración y el análisis.

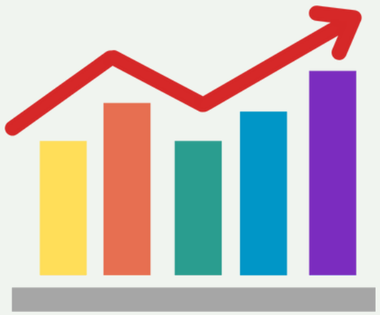
Una hipótesis ***es una suposición o propuesta preliminar*** sobre la relación entre variables o sobre las características de los datos que se pretende investigar y validar.

Durante un **EDA**, las hipótesis pueden surgir de ***observaciones preliminares***, del conocimiento previo sobre el dominio del problema, o de la literatura existente

# VISUALIZACIONES.

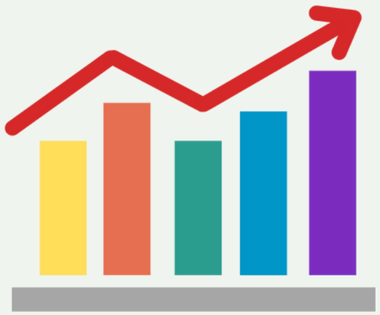
LevelUP
Tech Academy

# VISUALIZACIONES



Las visualizaciones son herramientas fundamentales en el ***Análisis Exploratorio de Datos (EDA)*** porque permiten interpretar y comunicar información compleja de manera **clara y eficiente**.

# TIPOS DE DATOS



Cualitativos: Propiedades **categóricas** de nuestro dataset

- Nominal: “Soltero” / “Casado” / “Divorciado”
- Ordinal: “Regular” / “Bien” / “Muy Bien”

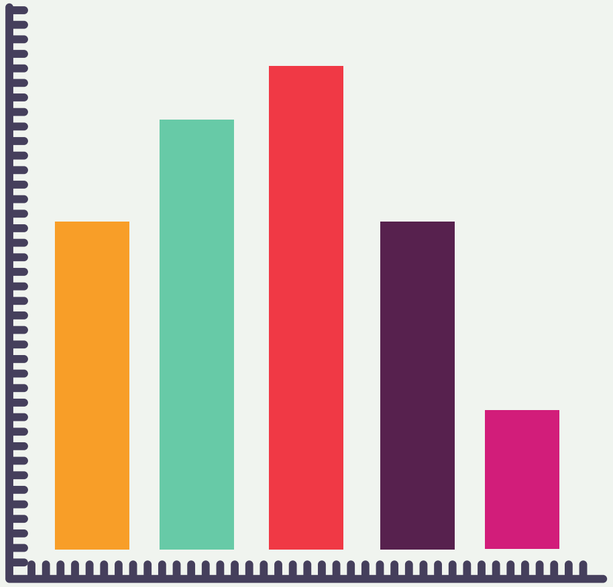
Cuantitativos: Propiedades **numércias** de nuestro dataset

- Discreto: Toma valores **determinados** Reales
- Contunio: Tova valores **cualesquiera dentro de un intervalo** de Nros Reales

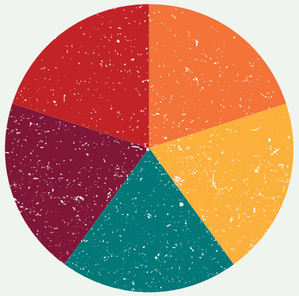
# GRÁFICOS DE BARRAS (BAR PLOTS)

Los gráficos de barras **representan la frecuencia o el valor de categorías específicas** mediante barras de longitud proporcional.

**Cuándo usar:** Para comparar cantidades entre diferentes categorías.



# GRAFICOS DE TORTA



Los gráficos de torta ( o pie charts) son una herramienta visual común utilizada para mostrar la **proporción de diferentes categorías dentro de un conjunto de datos**. Cada sector del gráfico representa una categoría y su tamaño es proporcional a su porcentaje del total.

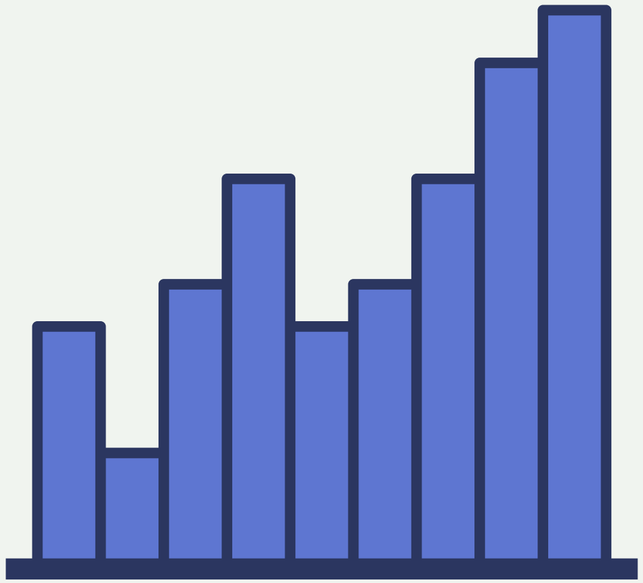
Cuándo Usar Gráficos de Torta

- **Proporciones**: Para mostrar cómo se divide un total en partes proporcionales.
- **Comparaciones de Partes**: Para comparar la proporción de diferentes categorías en un conjunto de datos pequeño.
- **Visualización Simple**: Cuando hay *pocas* categorías (generalmente menos de 6) para mantener la claridad.

# HISTOGRAMAS

Los histogramas muestran la **distribución de una variable continua dividiendo el rango de valores en intervalos (`bins`)** y contando la frecuencia de los valores en cada intervalo.

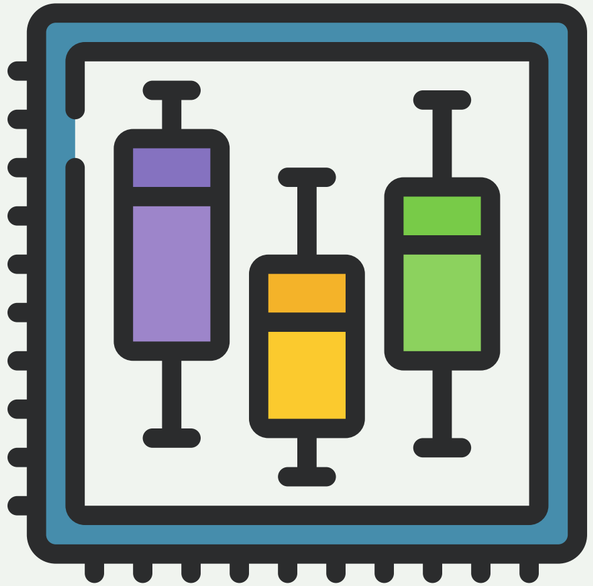
**Cuándo usar:** Para entender la distribución de una variable continua y detectar sesgos, asimetrías, o presencia de valores atípicos.



# DIAGRAMAS DE CAJA (BOXPLOTS)

Los boxplots muestran la ***distribución*** de una **variable continua** a través de sus **cuartiles**, **destacando la mediana y los valores atípicos**.

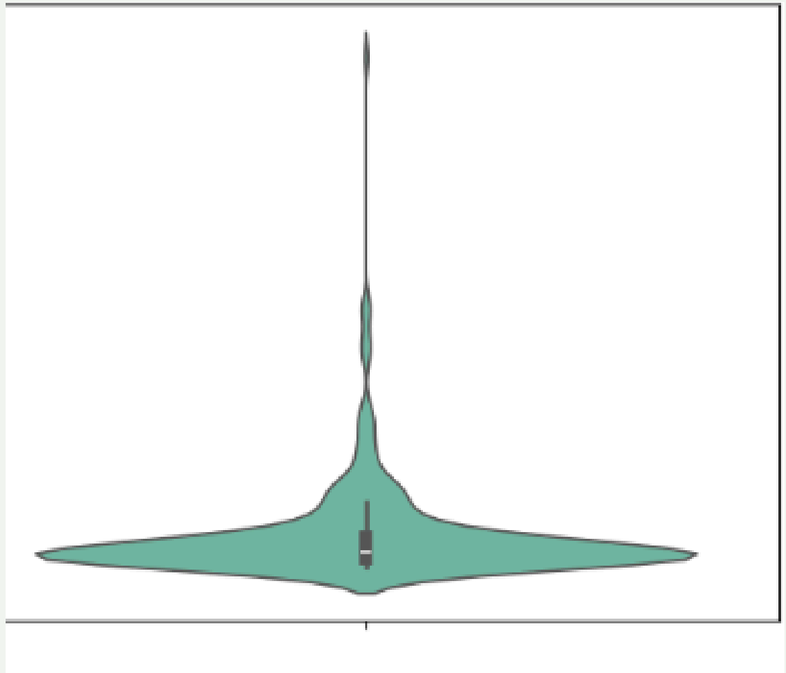
**Cuándo usar:** Para comparar distribuciones entre diferentes grupos o detectar valores atípicos.



# DIAGRAMAS DE VIOLÍN

Los violin plots combinan características de los boxplots y los KDE plots (Kernel Distribution Estimator) para mostrar la **distribución de una variable y sus densidades en diferentes niveles**.

**Cuándo usar:** Para comparar distribuciones entre grupos y entender la forma de las distribuciones.



# CONSULTAS SQL


LevelUP
Tech Academy

# QUE ES SQL

SQL (Structured Query Language) es un lenguaje de programación específico para gestionar y manipular **bases de datos relacionales**.

# FUNCIONALIDADES PRINCIPALES DE SQL

## Consultas de Datos:

- **SELECT:** Para recuperar datos de una base de datos.

```
SELECT columna1, columna2 FROM tabla WHERE condición;
```

Si bien en el curso no profundizaremos en sintaxis SQL, dejamos las más comunes:

## Manipulación de Datos:

**INSERT:** Para insertar nuevos registros en una tabla

```
INSERT INTO tabla (columna1, columna2) VALUES (valor1, valor2);
```

**UPDATE:** Para modificar registros existentes

```
UPDATE tabla SET columna1 = valor1 WHERE condición;
```

**DELETE:** Para eliminar registros de una tabla

```
DELETE FROM tabla WHERE condición;
```

# Definición de Datos:

**CREATE TABLE:** Para crear una nueva tabla

```
CREATE TABLE tabla (
  columna1 tipo_dato,
  columna2 tipo_dato,
  ...
);
```

**ALTER TABLE:** Para modificar la estructura de una tabla existente

```
ALTER TABLE tabla ADD columna tipo_dato;
```

**DROP TABLE:** Para eliminar una tabla

```
DROP TABLE tabla;
```

# VENTAJAS DE USAR SQL

- **Estandarización:** Es un lenguaje estándar, lo que significa que se **utiliza ampliamente** y de manera consistente en muchos sistemas de bases de datos.

- **Facilidad de Uso:** Es intuitivo y **fácil de aprender** para aquellos que tienen conocimientos básicos de programación y manejo de datos.

- **Flexibilidad:** Permite **realizar operaciones complejas de manera eficiente**, desde simples consultas hasta manipulaciones avanzadas de datos.

- **Interoperabilidad:** Puede **interactuar** con muchos otros lenguajes de programación y aplicaciones.

# PANDASQL

Pandasql es una biblioteca de Python que permite **utilizar SQL** para consultar datos en **DataFrames** de pandas.

Esto puede ser especialmente útil para quienes están más familiarizados con SQL que con las operaciones de pandas o para combinar consultas SQL con las capacidades de pandas. pandasql proporciona una manera conveniente de escribir consultas SQL directamente en Python y ejecutarlas sobre DataFrames.

```
# Definir una consulta SQL
query3 = """
SELECT sum(Fare)
FROM df
"""

# Ejecutar la consulta SQL usando pandasql
result3 = psql.sqldf(query3,)

# Mostrar el resultado
print("Suma total de todas las tarifas pagadas \n",result3)
```

# GRACIAS

Team Level Up

LevelUP
Tech Academy