DMC | INSTITUTE

Dip. Data Sciencie

Curso:

Estadística Descriptiva

Sesión 02

Docente: Nilton Yanac

Enero, 2025

REGLAS

Se requiere **puntualidad** para un mejor desarrollo del curso.

Para una mayor concentración **mantener silenciado el micrófono** durante la sesión.

Las preguntas se realizarán **a través del cha**t y en caso de que lo requieran **podrán activar el micrófono**.

Realizar las actividades y/o tareas encomendadas en **los plazos determinados**.

Identificarse en la sala Zoom con el primer nombre y primer apellido.

ITINERARIO

07:00 PM – 07:30 PM	***Soporte técnico DMC***
07:30 PM – 08:50 PM	***Agenda***
08:50 PM – 09:00 PM	***Pausa Activa***
09:00 PM – 10:30 PM	***Agenda***

Horario de Atención Área Académica y Soporte

Lunes a Viernes 09:00 am a 10:30 pm / Sábado 09:00 am a 02:00pm

SILABO

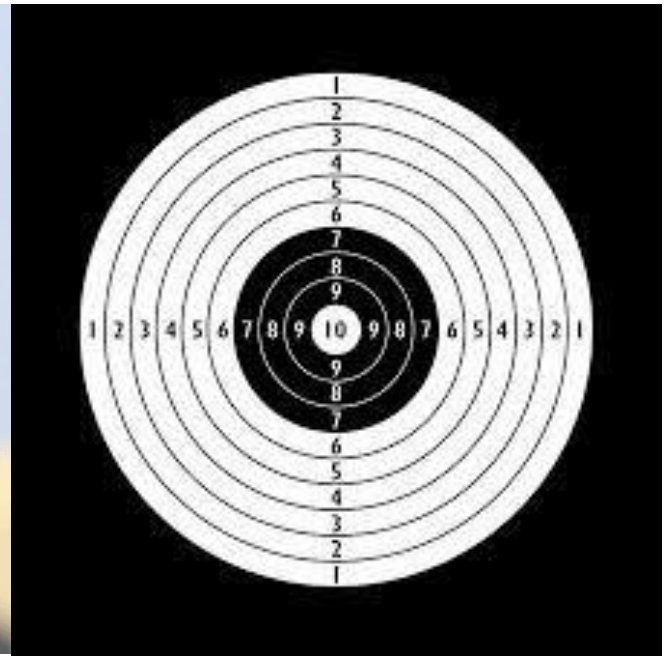
Objetivo del curso:

Conocer las principales metodologías de análisis de datos para la toma de decisiones en los negocios

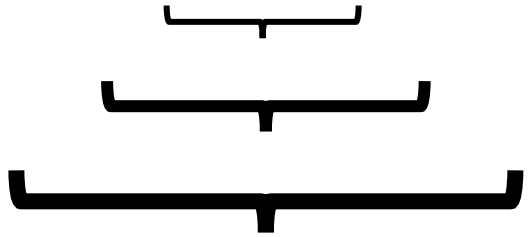
Agenda de la sesión 02:

- *Tema 01: REPASO DE LA SESIÓN 01*
- *Taller 01: ANÁLISIS DESCRIPTIVO*
- *Tema 02: ANÁLISIS EXPLORATORIO DE DATOS*
- *Taller 02:: ANÁLISIS EXPLORATORIO CON COLAB*

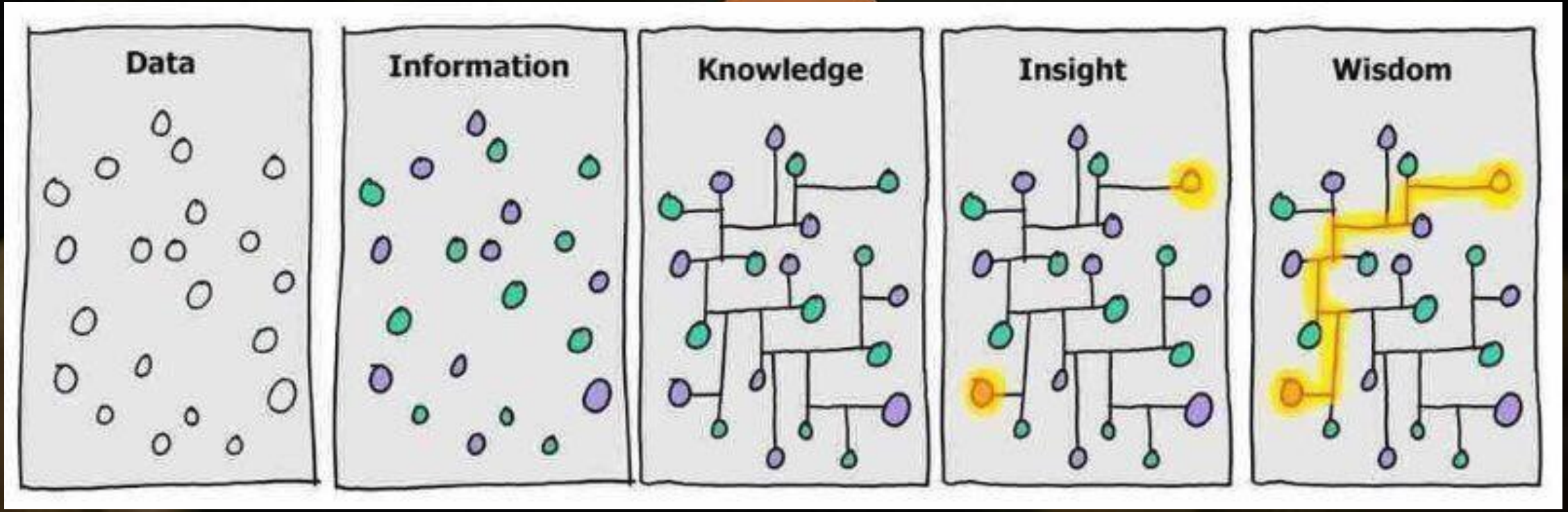
¿Qué sabe hacer mejor el que usa estadística?



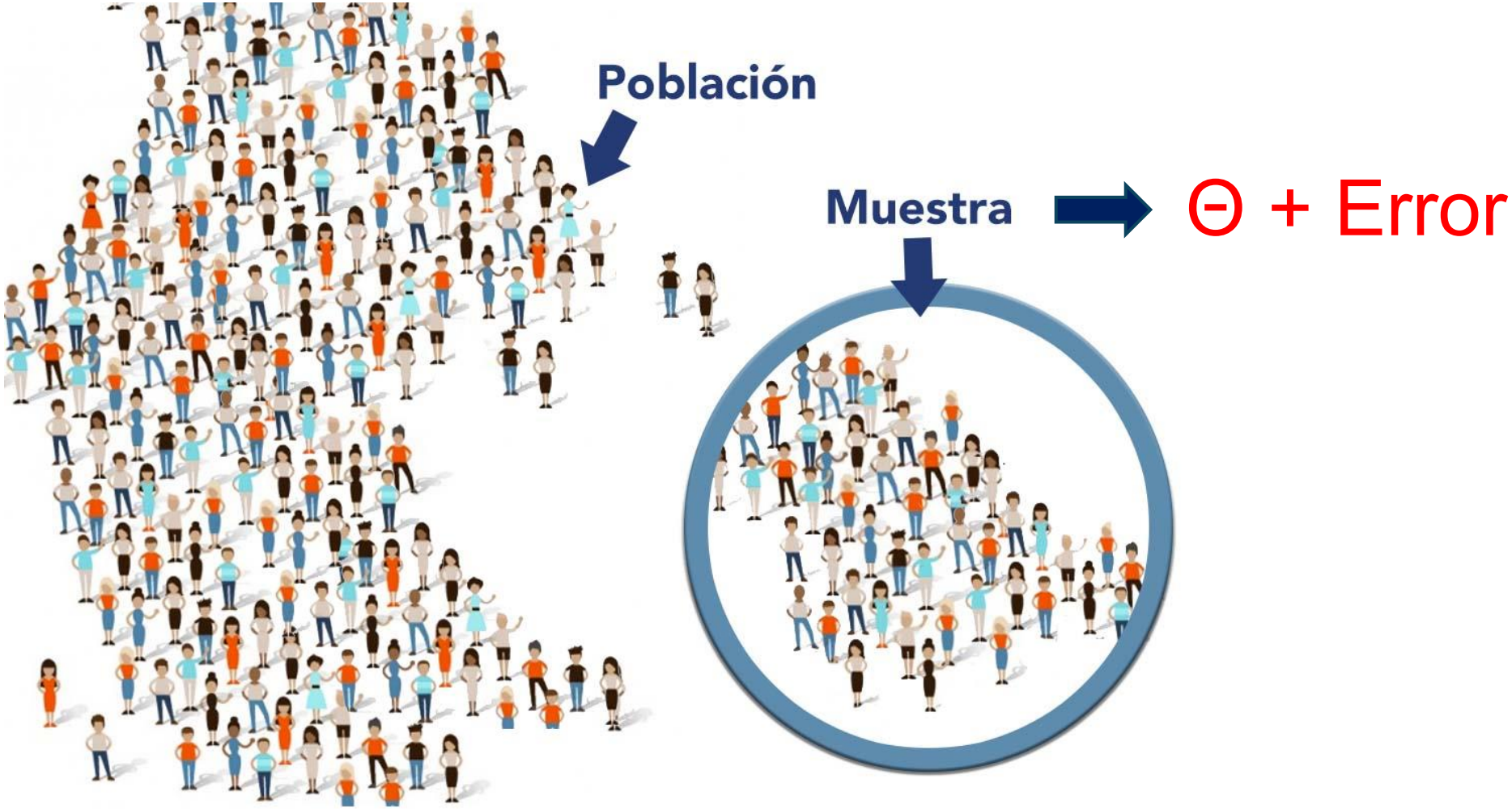
Θ + Error



Del Dato al Insight...



¿Por qué existe el error en la estadística?



Análisis Descriptivo de Datos

Medidas de centralidad. Medidas para entender entorno a qué valores se distribuye la variable

Media

- Suma de todos los valores, partido por el número de ellos.
- Sensible a valores extremos.

Mediana

- Ordenados los valores de menor a mayor, punto por debajo del que se encuentran la mitad de valores.
- Es un valor del conjunto. Si son pares, el inmediato inferior.
- Poco sensible a valores extremos, más robusta.

Moda

Valor más repetido.

- En el caso continuo se trata del rango con más casos.
- En el caso discreto, el más repetido.

Análisis Descriptivo de Datos

Varianza

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

- En general, promedio de la diferencia de cada valor respecto a la media del conjunto, al cuadrado. (unidades cuadradas)

Desviación Típica

$$\hat{\sigma} = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$$

- Raíz de la varianza, misma unidad.

Análisis Descriptivo de Datos

Coeficiente de Variación

Es una medida de dispersión relativa, no posee unidades.

Para una muestra	Para una Población
$cv = \frac{S}{\bar{x}} x100$	$CV = \frac{\sigma}{\mu} x100$

Se recomienda su uso para comparar la variabilidad de dos o más conjuntos de datos que tienen diferentes unidades de medida o cuando los promedios de los conjuntos de datos a comparar son diferentes.

TALLER 01

El área de RRHH de la empresa MANUFACTURAS SA ha seleccionado aleatoriamente a 30 personas del área de producción (se encuentran personal operario, personal técnico, jefe de área, y supervisores) Dicha área está buscando identificar la relación que existe entre el salario actual de su personal y otras características.

Con los hallazgos, la gerencia general busca desarrollar una política de sueldos para sus nuevos colaboradores.

La cantidad de trabajadores del área de producción es de 280 personas, incluyendo a los operarios, supervisores, jefes y sub gerentes. La empresa cuenta con 320 empleados en total.

Luego de analizar la muestra inicial de 30 trabajadores del área de producción, se le ha contratado a usted para que confirme los resultados obtenidos o sugiera alguna mejora antes de llevar dichos números a la junta de directorio con los inversionistas, donde se tomará la decisión de abrir una nueva planta de producción con una inversión de USD 200 millones, de los cuales el 35% representa el gasto en planilla para los primeros 24 meses de operación. Use los datos que se le ha brindado y el link a continuación:

https://colab.research.google.com/drive/1Sq0fBEPr0MrmL9bedxAX8tGTnONzje6W?usp=sharing

TEMA 02:
ANÁLISIS EXPLORATORIO DE DATOS (EDA)

¿Qué es el análisis exploratorio de datos (EDA)?

Conjunto de técnicas estadísticas dirigidas a explorar, describir y resumir la información que contienen los datos, maximizando su comprensión.

1. Realizar un análisis descriptivo
2. Revelar la presencia de datos atípicos
3. Identificar posibles errores
4. Comprobar la relación entre variables

Es esencial para garantizar que los resultados de cualquier análisis sean consistentes y veraces

¿Qué pasos hay que seguir en el análisis exploratorio de datos (EDA)?

1. El análisis descriptivo, permite conocer los tipos de datos, descubrir patrones y preparar los datos para futuros análisis.

2. Una mala codificación de las variables puede influir negativamente en la agrupación de los datos o los resultados de los análisis.

3. Los datos ausentes pueden generar problemas a la hora de aplicar técnicas de machine learning, elaborar modelos predictivos, realizar análisis estadísticos o generar representaciones gráficas.

4. Pueden modificar los resultados y restar potencia a los análisis estadísticos o técnicas de machine learning aplicadas.

5. Entre otras razones, para descartar posibles variables que aporten información redundante en el conjunto de datos, ocasionando ruido en los análisis.

Proceso del análisis exploratorio de datos

BOXPLOT:

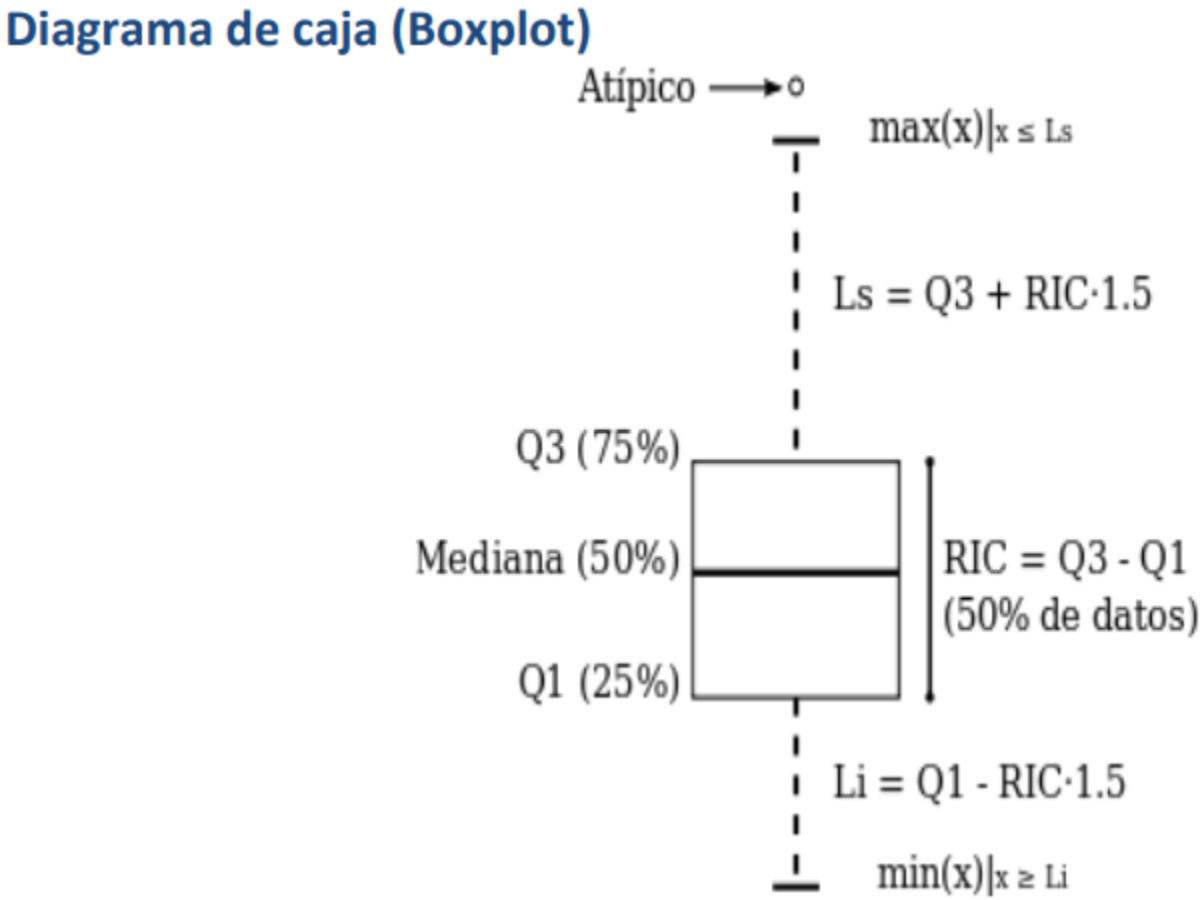
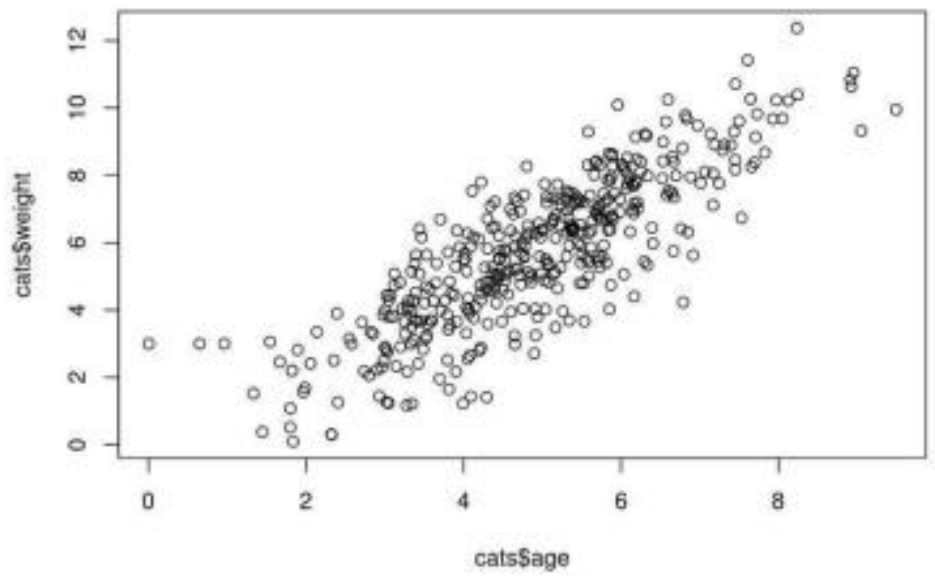


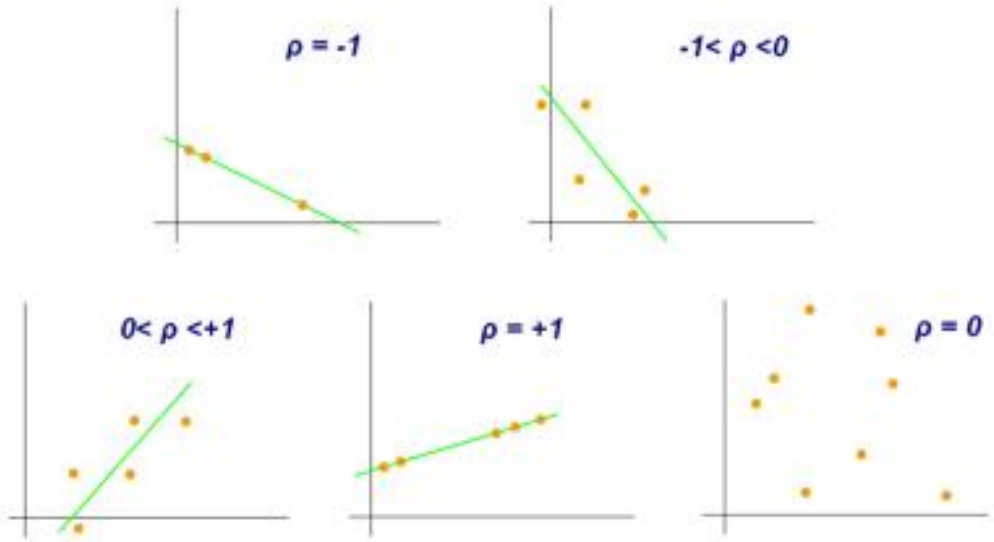
Diagrama de caja (Boxplot)
Atípico
max(x)|x ≤ Ls
Ls = Q3 + RIC·1.5
Q3 (75%)
Mediana (50%)
Q1 (25%)
RIC = Q3 - Q1
(50% de datos)
Li = Q1 - RIC·1.5
min(x)|x ≥ Li

GRÁFICOS DE DISPERSIÓN

Gráfico de dispersión



Análisis de dispersión y distribución



OUTLIERS Y TRANFORMACIÓN DE VARIABLES

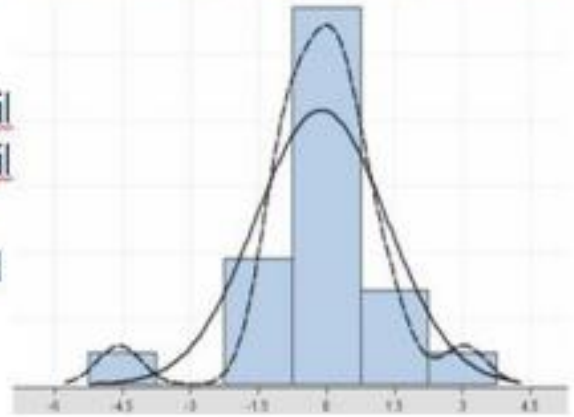
Valores missing y extremos

- Usar el rango intercuartil: [Q1 – 1.5 * IRQ , Q3 + 1.5 * IRQ]
- Donde:

  Q1: 1er cuartil

  Q3: 3er cuartil

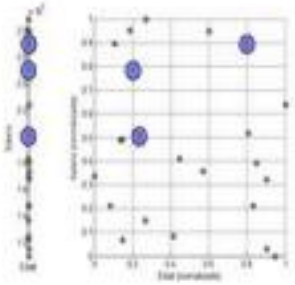
  IRQ: Q3 – Q1



Transformación de variables

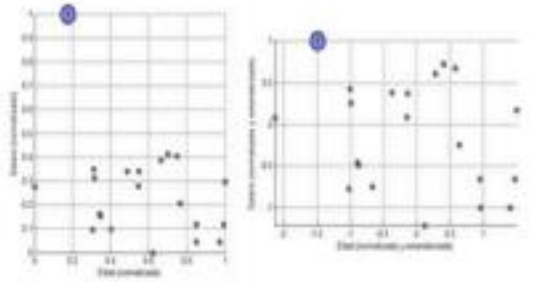
- Normalización min-max

$$x_{new} = (x_i - \min(X)) / (\max(X) - \min(X))$$



- Estandarización z-score

$$x_{new} = (x_i - \text{mean}(X)) / \text{sd}(X)$$



TALLER 02:
ANÁLISIS EXPLORATORIO EN COLAB – IMPUTACIÓN DE DATOS

https://colab.research.google.com/drive/1IV0hgFXWKja2CAAbnA6xulJaQWeglyf_?usp=sharing

Imputación: usando la media o promedio

```
import numpy as np
import pandas as pd
from sklearn.impute import SimpleImputer

# Crear un dataframe con valores faltantes
df = pd.DataFrame({'A': [1, 2, np.nan, 4],
                   'B': [5, np.nan, np.nan, 8],
                   'C': [9, 10, 11, 12]})

# Verificar valores faltantes antes de la imputación
print("Valores faltantes antes de la imputación:\n", df.isnull().sum())

# Crear un imputador con estrategia de media
imputer = SimpleImputer(strategy='mean')

# Imputar los valores faltantes
df_imputed = pd.DataFrame(imputer.fit_transform(df), columns=df.columns)

# Mostrar el DataFrame después de la imputación
print("\nDataFrame después de la imputación:\n", df_imputed)
```

Imputación: usando la moda

```
import numpy as np
import pandas as pd
from sklearn.impute import SimpleImputer

# Crear un DataFrame con valores faltantes
df = pd.DataFrame({'A': [1, 2, np.nan, 4, 2],
                   'B': [5, np.nan, np.nan, 8, 5],
                   'C': [9, 10, 11, 12, 10]})

# Verificar valores faltantes antes de la imputación
print("Valores faltantes antes de la imputación:\n", df.isnull().sum())

# Crear un imputador con estrategia de moda (most_frequent)
imputer = SimpleImputer(strategy='most_frequent')

# Imputar los valores faltantes
df_imputed = pd.DataFrame(imputer.fit_transform(df), columns=df.columns)

# Verificar valores faltantes después de la imputación
print("\nValores faltantes después de la imputación:\n", df_imputed.isnull().sum())

# Mostrar el DataFrame después de la imputación
print("\nDataFrame después de la imputación:\n", df_imputed)
```

Imputación: usando la mediana

```
import numpy as np
import pandas as pd
from sklearn.impute import SimpleImputer

# Crear un DataFrame con valores faltantes
df = pd.DataFrame({'A': [1, 2, np.nan, 4],
                   'B': [5, np.nan, np.nan, 8],
                   'C': [9, 10, 11, 12]})

# Verificar valores faltantes antes de la imputación
print("Valores faltantes antes de la imputación:\n", df.isnull().sum())

# Crear un imputador con estrategia de mediana
imputer = SimpleImputer(strategy='median')

# Imputar los valores faltantes
df_imputed = pd.DataFrame(imputer.fit_transform(df), columns=df.columns)

# Mostrar el DataFrame después de la imputación
print("\nDataFrame después de la imputación:\n", df_imputed)
```

Imputación: usando el método de interpolación

```
import pandas as pd
import numpy as np

# Crear un DataFrame con valores faltantes
df = pd.DataFrame({'A': [1, 2, np.nan, 4],
                   'B': [5, np.nan, np.nan, 8],
                   'C': [9, 10, 11, 12]})

# Verificar valores faltantes antes de la imputación
print("Valores faltantes antes de la imputación:\n", df.isnull().sum())

# Imputar los valores faltantes usando interpolación lineal
df_imputed = df.interpolate(method='linear', axis=0,
limit_direction='forward')

# Verificar si quedan valores faltantes después de la interpolación
print("\nValores faltantes después de la imputación:\n",
df_imputed.isnull().sum())

# Mostrar el DataFrame después de la imputación
print("\nDataFrame después de la imputación:\n", df_imputed)
```

Imputación: usando el método de extrapolación por regresión

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression

# Crear un DataFrame con datos de ejemplo
data = {
   'X': [1, 2, 3, 4, 5, 6],
   'Y': [5, 7, np.nan, np.nan, 15, 17],  # np.nan en lugar de
None
   'Z': [10, 12, 14, 16, 18, 20]
}

df = pd.DataFrame(data)

# Paso 1: Separar los datos conocidos y desconocidos
df_conocidos = df.dropna(subset=['Y'])
df_desconocidos = df[df['Y'].isnull()]
```

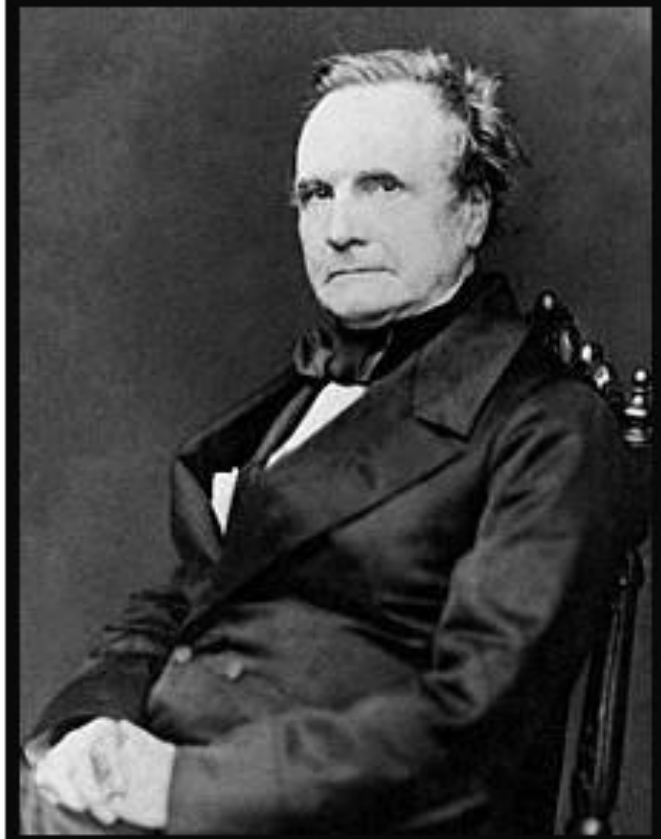
```
# Verificar si hay valores faltantes en 'Y' antes de entrenar el modelo
if not df_desconocidos.empty:
   # Paso 2: Preparar los datos para el modelo de regresión
   X_train = df_conocidos[['X', 'Z']]
   y_train = df_conocidos['Y']

   # Paso 3: Entrenar el modelo de regresión
   regression_model = LinearRegression()
   regression_model.fit(X_train, y_train)

   # Paso 4: Predecir los valores faltantes
   X_test = df_desconocidos[['X', 'Z']]
   predicted_values = regression_model.predict(X_test)

   # Paso 5: Reemplazar los valores faltantes en el DataFrame original
   df.loc[df['Y'].isnull(), 'Y'] = predicted_values

# Mostrar el DataFrame después de la imputación
print("\nDataFrame después de la imputación por regresión:\n", df)
```



Los errores por usar datos inadecuados son muchos menos que por no usar ningún dato en absoluto.

(Charles Babbage)

akifrases.com

DMC | INSTITUTE

¡Gracias… Totales!

Docente: Nilton Yanac

Junio, 2025