ISTEA

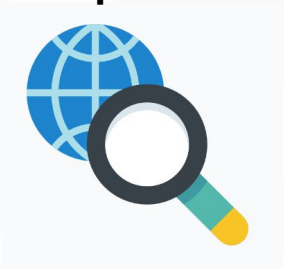
Análisis Exploratorio de Datos

(EDA)

Objetivo general

Explorar, analizar y visualizar datos de manera efectiva **antes de aplicar técnicas de minería de datos.**

Explorar



Examinar con detenimiento

Analizar

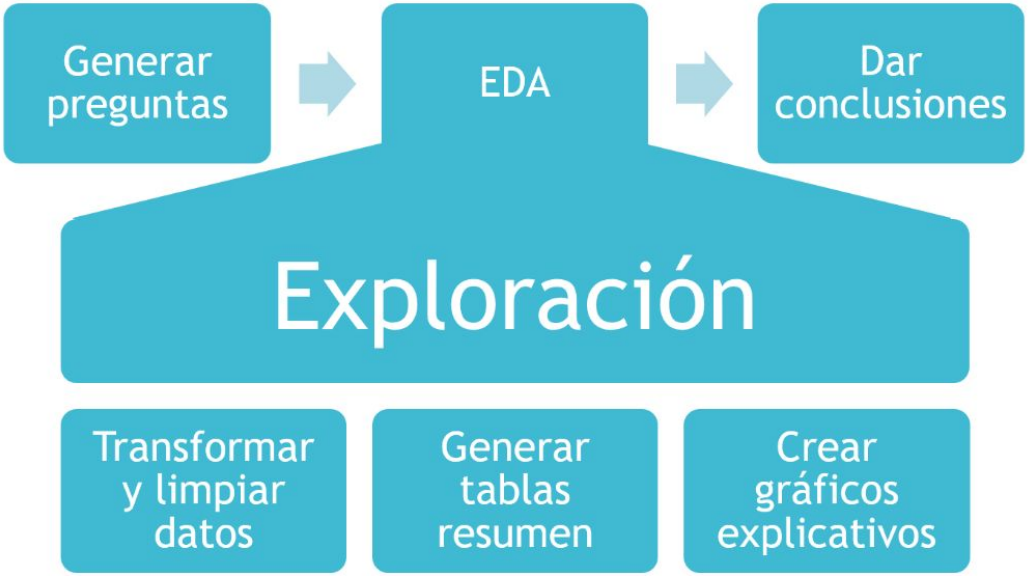
Descomponer en partes para comprender mejor

Visualizar

Representar algo por medio de imágenes

Análisis exploratorio de datos

Abreviado como EDA (Exploratory Data Analysis), es el proceso de investigar y comprender un conjunto de datos utilizando **técnicas estadísticas** y **visuales**.



El EDA y el proceso de la minería de datos

Transformar y limpiar datos

…pero con profundidad estadística.

Objetivos del EDA

- **Comprensión de los datos**: Permite una comprensión profunda de la naturaleza de los datos, lo que es esencial para tomar decisiones informadas en cualquier análisis posterior.
- **Detección de problemas**: Ayuda a identificar problemas en los datos, como valores faltantes, errores de entrada o valores atípicos, que deben ser tratados antes de realizar análisis más avanzados.
- **Selección de características**: Puede ayudar en la selección de las características más relevantes para un problema, lo que puede mejorar el rendimiento de los modelos.
- **Generación de hipótesis**: La EDA a menudo conduce a la formulación de hipótesis que pueden ser probadas más adelante utilizando técnicas de modelado.

Transformar y limpiar datos

Se siguen los mismos pasos que en las etapas de limpieza y transformación del proceso general de la minería de datos (clase anterior):

- Manejo de valores faltantes
- Tratamiento de valores atípicos
- Codificación de variables categóricas
- Normalización y estandarización

Generar tablas resumen

Una vez que los datos han sido transformados y limpiados, es útil generar tablas resumen que proporcionen una visión general de las estadísticas y distribuciones de las variables:

- Estadísticas descriptivas
- Tablas de contingencia
- Matrices de correlación

Generar tablas resumen

Estadísticas descriptivas

- Calcular estadísticas descriptivas básicas para las variables numéricas, como la media, la mediana, la desviación estándar, el mínimo y el máximo.
- Estas estadísticas proporcionan una **visión general** de la tendencia central y la dispersión de los datos.

Generar tablas resumen

Tablas de contingencia

- Para variables categóricas, se pueden crear tablas de contingencia que muestren la relación entre diferentes categorías.
- Estas tablas son útiles para comprender las relaciones y las distribuciones de las variables categóricas.

Generar tablas resumen

Matrices de correlación

- Calcular matrices de correlación para identificar relaciones lineales entre pares de variables numéricas.
- Las matrices de correlación muestran coeficientes de correlación que van desde -1 (correlación negativa perfecta) hasta 1 (correlación positiva perfecta).

Ejemplo

Rendimiento de estudiantes:

- Crear un dataframe con los siguientes datos:

Librerías Python a utilizar:

- Pandas (documentación)

Student_ID	Name	Gender	Age	Math_Score	Science_Score	History_Score
1	Juan	M	18	85	90	75
2	María	F	19	92	88	80
3	Carlos	M	20	78	85	72
4	Luisa	F	18	88	92	88
5	Ana	F	20	76	78	65

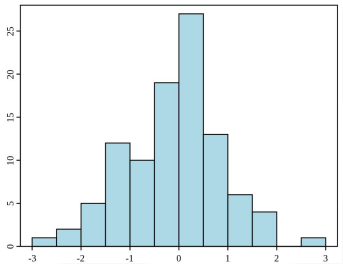
Ejemplo

Rendimiento de estudiantes:

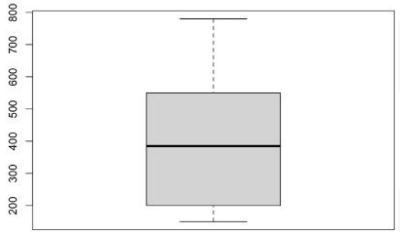
- Obtener e interpretar las estadísticas descriptivas
- Obtener e interpretar tablas de contingencias
- Obtener e interpretar matriz de correlaciones

Generar gráficos explicativos

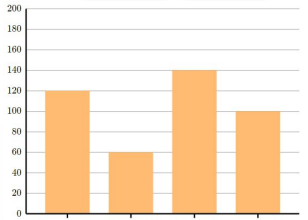
La visualización de datos desempeña un papel fundamental en la exploración de datos. Los gráficos ayudan a comprender mejor los patrones y las relaciones en los datos.



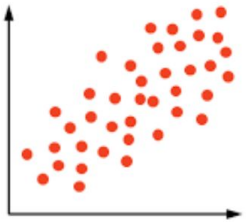
Histograma

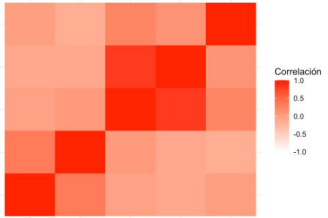


Caja



Barras


Dispersión



Calor

Generar gráficos explicativos

Histograma

Es un gráfico en el cual se representa una determinada distribución de frecuencias de la variable (discreta o continua):

- Frecuencia absoluta (f_i): Número de veces que se repite el elemento.
- Frecuencia relativa (f_r): f_i / N, donde N es el tamaño de la muestra
- Frecuencia acumulada (F): Suma de la frecuencia inmediata superior o el mismo.
- Frecuencia relativa acumulada (F_r): F / N

Generar gráficos explicativos

Histograma (ejemplo)

Variable discreta

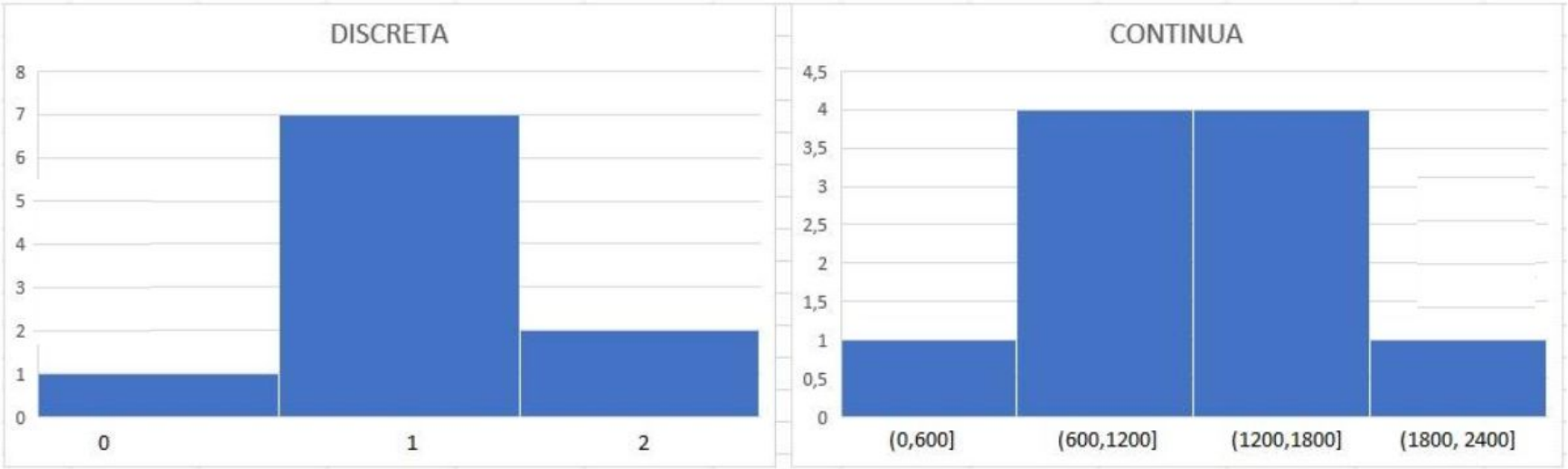
Encuesta a 10 personas preguntándoles por el número de smartphones que poseen y los resultados que obtenemos son:
1, 2, 1, 1, 1, 2, 0, 1, 1, 1.

Variable continua

Encuesta a 10 personas y sus salarios:
1005,38; 2100,52; 567,98; 848,82; 654,36; 1653,22; 1308,54; 1789,12; 762,95; 1234,33

Generar gráficos explicativos

Histograma



Generar gráficos explicativos

Barras

Un diagrama de barras es un gráfico usado para mostrar de forma resumida un grupo de datos que puede incluir variables cualitativas y cuantitativas.

- Se compone de columnas o barras de diferentes alturas, estas pueden ser horizontales o verticales.
- Tiene un eje horizontal o eje x, donde se ubica una variable, por lo general, cualitativa.
- Tiene un eje vertical o eje y, donde se ponen los valores que determinan la altura de las barras. A estos números se les conoce como **frecuencia**.
- El ancho de las barras y el espacio entre cada una debe ser el mismo.
- Las barras también sirven para comparar valores.

Generar gráficos explicativos

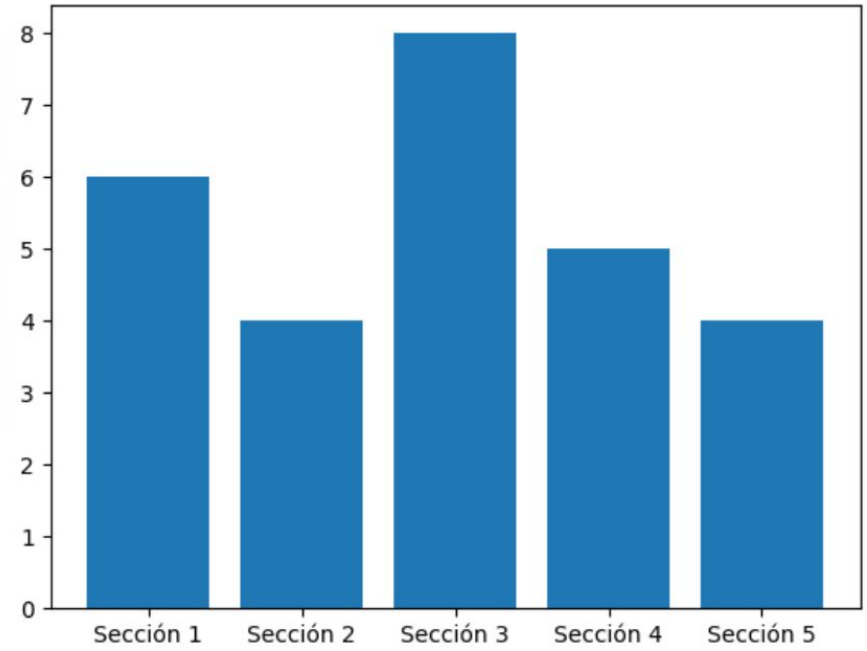
Barras (ejemplo #1)

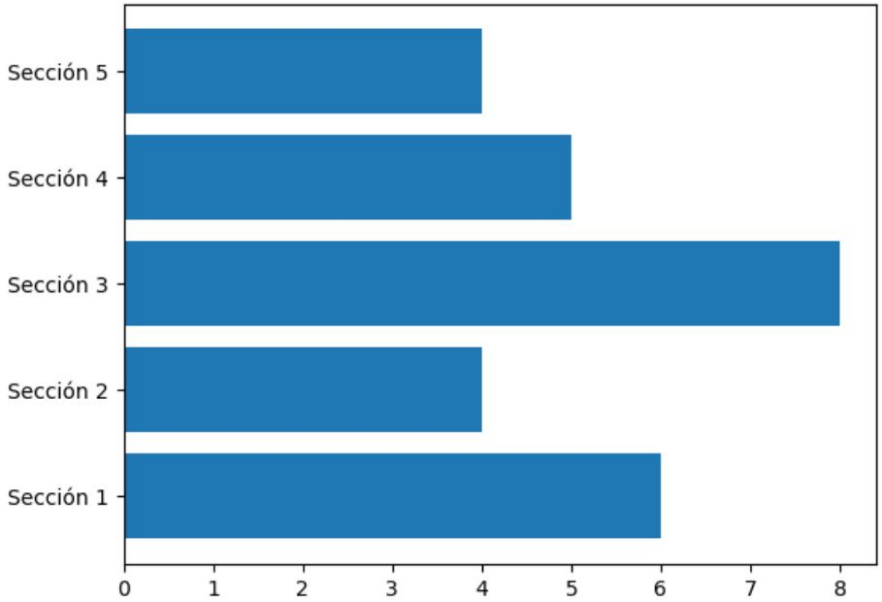
En un colegio se desea comparar las puntuaciones promedio de matemáticas en las diferentes secciones de 5to grado (notas en escala 1-10). Existen 5 secciones y los promedios; el desvío estándar y el error estándar se muestran en la siguiente tabla:

Sección	n	Prom	σ	σ / √n
Sección 1	25	6	1	1/5
Sección 2	25	4	2	2/5
Sección 3	25	8	3	3/5
Sección 4	25	5	1	1/5
Sección 5	25	4	2	2/5

Generar gráficos explicativos

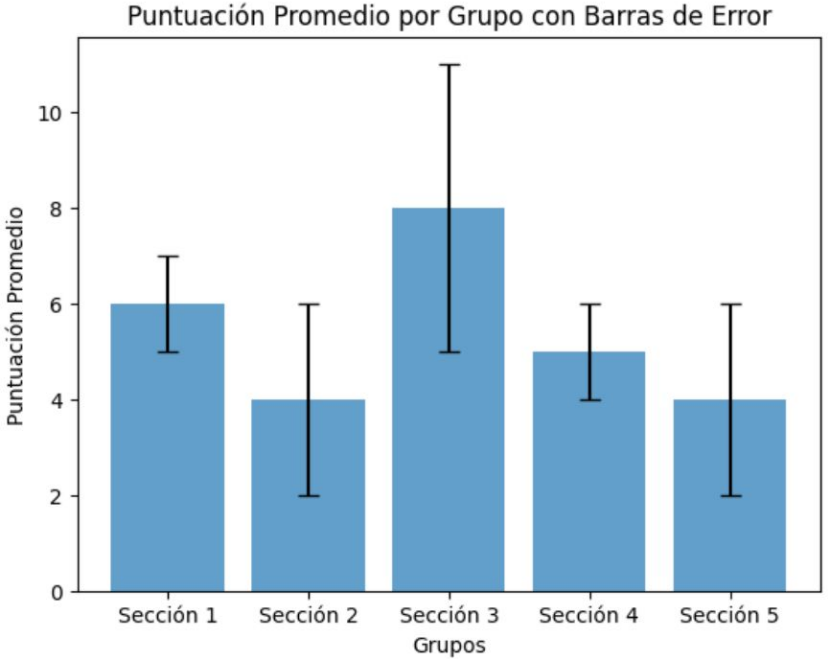
Barras





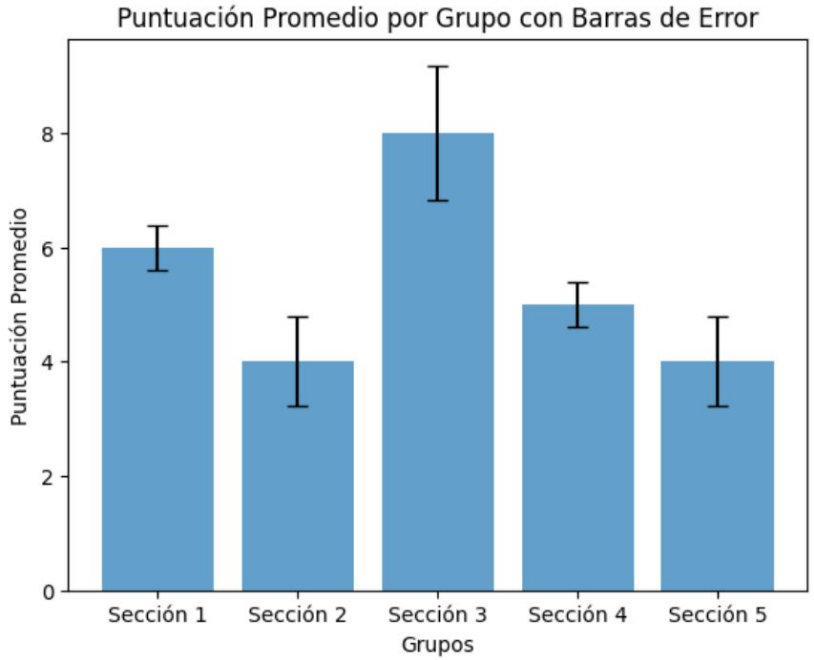
Generar gráficos explicativos

Barras con error desvío estándar



Generar gráficos explicativos

Barras con error intervalo de confianza al 95% (Z = 1.96)



Generar gráficos explicativos

Barras (ejemplo #2)

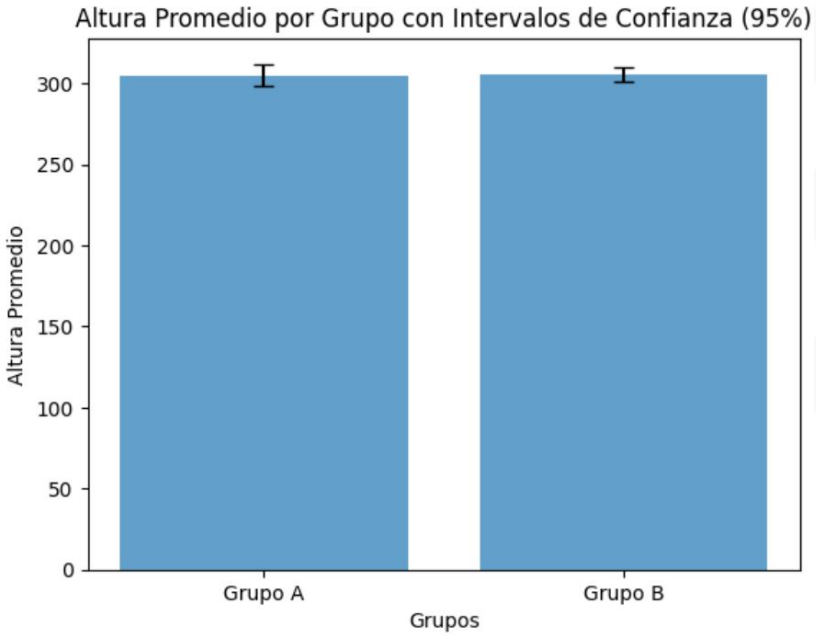
Suponga dos grupos A y B con medias de 305 y 305.4, respectivamente y un Error Estándar de la Media (SEM) de 3.54 y 2.44, respectivamente.

Determine en el gráfico por qué se puede concluir que, con un nivel de confianza del 95%, se puede decir que no existe diferencia estadística significativa entre las medias de ambos grupos.

La prueba t dio un valor p = 0.9281

Generar gráficos explicativos

Barras con error intervalo de confianza al 95% (Z = 1.96)



Generar gráficos explicativos

Dispersión

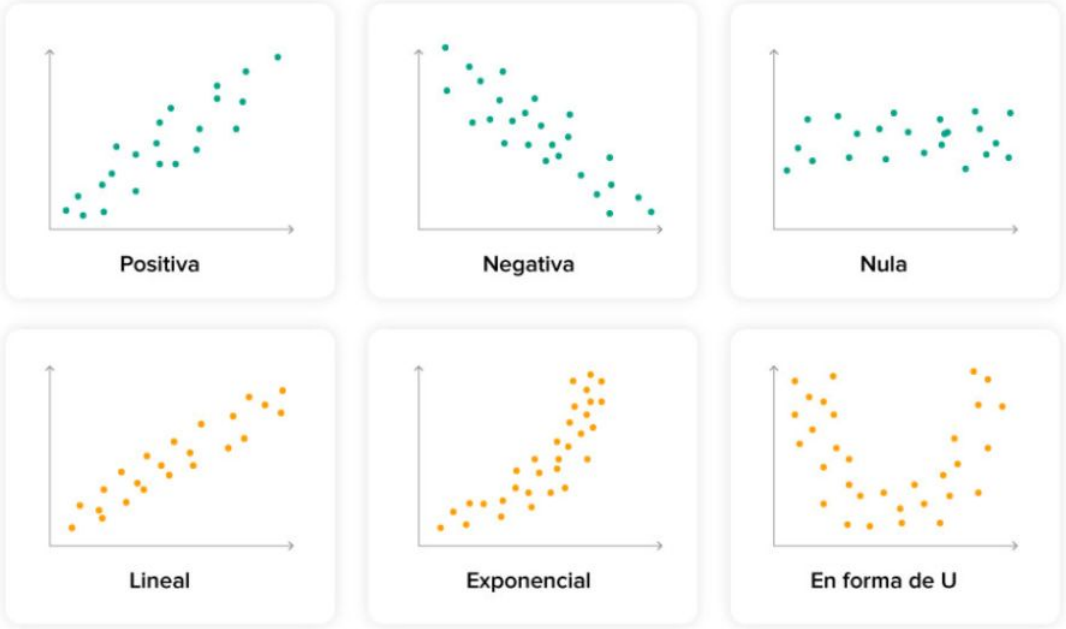
El diagrama o gráfica de dispersión es una herramienta de control y apoyo para verificar la existencia de una correlación o relación entre variables cuantitativas.

- Muestra la relación entre dos variables.
- Es el mejor método para mostrar un patrón no lineal.
- Se puede determinar el rango de flujo de datos, es decir, el valor máximo y mínimo.
- La observación y la lectura son sencillas.
- Trazar el diagrama de dispersión es fácil.

Generar gráficos explicativos

Dispersión

Tipos de correlación



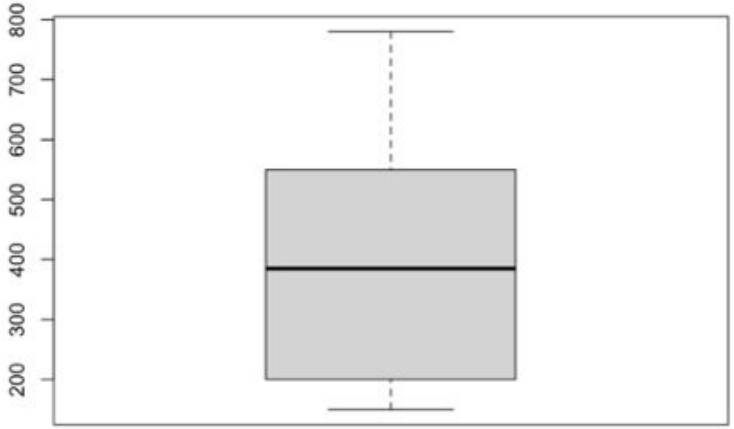
Generar gráficos explicativos

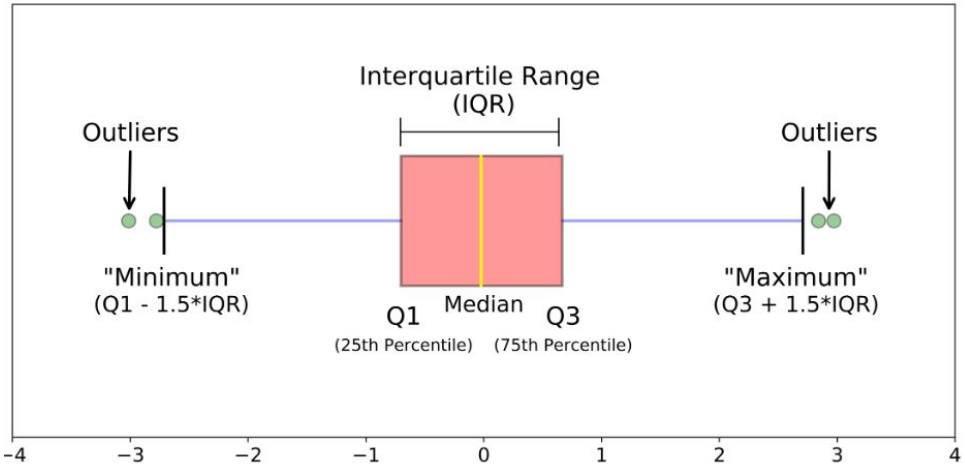
Dispersión

Generar gráficos explicativos

Caja

Un diagrama de caja (boxplot), es una representación de una variable cuantitativa o categórica con el propósito de identificar rápidamente los cuartiles del conjunto de datos.





Generar gráficos explicativos

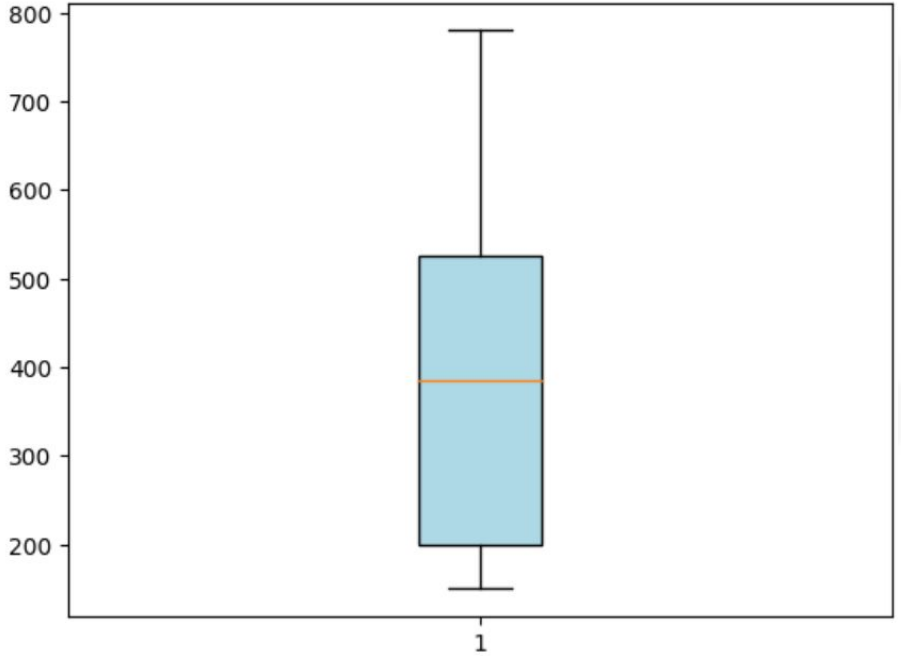
Caja (ejemplo)

Suponemos que queremos representar el número de ciclistas que pasan por delante de nuestra casa a lo largo de un año. Primero, contamos los ciclistas y recogemos la información en una tabla.la:

Mes	Ciclistas
Enero	200
Febrero	150
Marzo	200
Abril	300
Mayo	370
Junio	400
Julio	600
Agosto	700
Septiembre	780
Octubre	500
Noviembre	400
Diciembre	200

Generar gráficos explicativos

Caja (ejemplo)



Generar gráficos explicativos

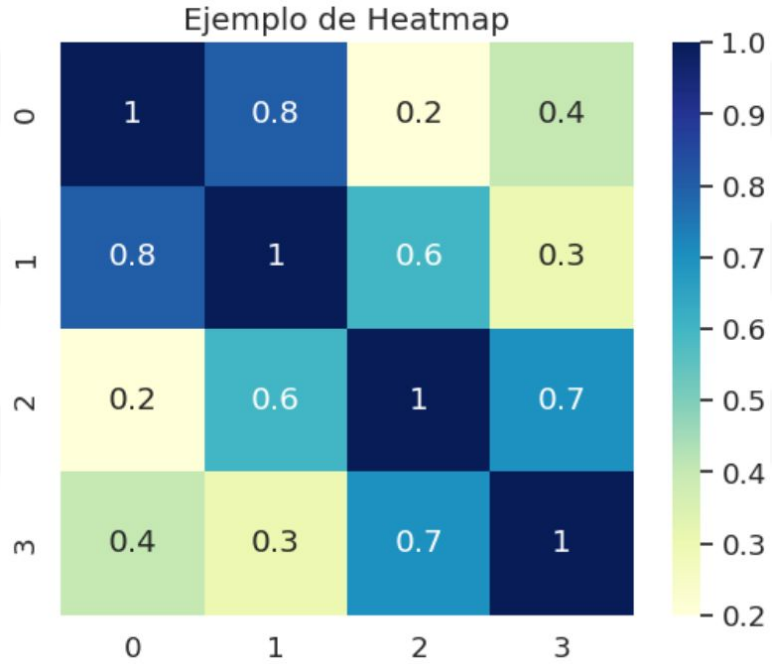
Calor

Un gráfico de calor, también conocido como mapa de calor, es una representación visual que se utiliza para mostrar la relación numérica entre dos variables categóricas o para representar la magnitud y cambios de una tercera variable en un gráfico bidimensional.

- Utiliza colores para resaltar valores en una cuadrícula o matriz, donde cada valor se asigna a un color específico.
- Los colores intensos suelen indicar valores más altos o relaciones más fuertes, mientras que los colores suaves representan valores más bajos o relaciones más débiles.

Generar gráficos explicativos

Calor (ejemplo)



Ejemplo

Explorar el siguiente dataset:

- IBM HR Analytics Employee Attrition

Librerías Python a utilizar:

- Pandas (documentación)

Resumen

- Visualización de Datos: La visualización de datos implica representar información numérica o categórica de manera gráfica para facilitar su comprensión. Algunas de las herramientas comunes incluyen gráficos de barras, gráficos de dispersión, histogramas y diagramas de calor.
- EDA (Análisis Exploratorio de Datos): EDA es un proceso crítico en la exploración de datos que implica resumir, visualizar y entender los datos antes de aplicar técnicas de modelado o inferencia. Se compone de pasos como limpieza de datos, identificación de valores atípicos, estadísticas descriptivas y visualización.
- Importancia de la Visualización y EDA: La visualización de datos y el EDA son cruciales para comprender la estructura y las tendencias en los datos, detectar valores atípicos, identificar patrones y relaciones, y tomar decisiones informadas en análisis de datos y toma de decisiones.

Preguntas…