FUNDACIÓN YPF

Módulo III | Clase 5

Análisis exploratorio de datos: Introducción al AED

FUNDACIÓN YPF

¿Ponemos a grabar el taller?

FUNDACIÓN YPF

Primera pre-entrega

¡En esta clase, deberán entregar los desafíos resueltos en la notebook 5 y 7. Los mismos serán subidos como repositorio a su cuenta de Github.

Presentarán lo trabajado entregando el link al de Github en el foro del aula virtual

FUNDACIÓN YPF

¿Qué vamos a ver hoy?

- ¿Qué es el análisis exploratorio de datos?
- Importancia del análisis exploratorio de datos
- ¿Qué debo buscar en un AED?

- Exploración de datos parte 1:

  Descripción de datos

  Visualización de datos

FUNDACIÓN YPF

¿Qué es el análisis exploratorio?

FUNDACIÓN YPF

Análisis exploratorio de datos

El análisis exploratorio de datos tiene por objetivo **identificar** las principales **características** de un conjunto de datos mediante un número reducido de gráficos y/o números

FUNDACIÓN YPF

¿Por qué es importante?

Nos ayuda a **organizar la información** que nos provee los datos a fin de **detectar patrones** presentes así como también datos que se apartan de manera sobresaliente del modelo subyacente.

Debe ser el **primer paso** en cualquier modelado

FUNDACIÓN YPF

¿Qué buscamos en el análisis exploratorio?

FUNDACIÓN YPF

Buscamos…

Análisis

Individual de cada variable

De las relaciones subyacente entre ellas

Mediante

Métodos gráficos

Medidas resumen

FUNDACIÓN YPF

¿Qué buscamos?

- **¿Cuantos features tengo (Columnas)? ¿Nombre?**

  Feature o atributo: Propiedad o característica individual y medible de un fenómeno observado.

- **¿Cuántas observaciones tengo (filas)?**

- **¿Cada fila contiene sólo atributos de una observación?**

FUNDACIÓN YPF

¿Qué buscamos?

¿Qué tipos de variables tengo?

Categóricas o cualitativas: ¿Están representadas por números o strings?

Cuantitativas discretas

Cuantitativas continuas

El tipo de datos determina el método de análisis apropiado y específico

FUNDACIÓN YPF

¿Qué buscamos?

¿Qué tipos de valores faltantes?

Faltantes random

Faltantes NO random

¿Hay valores erróneos?

FUNDACIÓN YPF

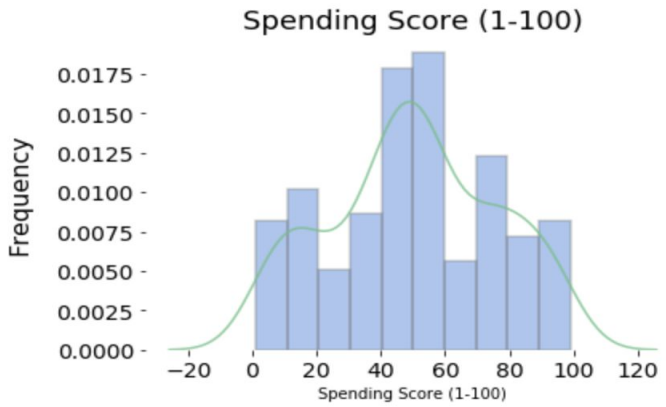
Distribución de las variables continuas

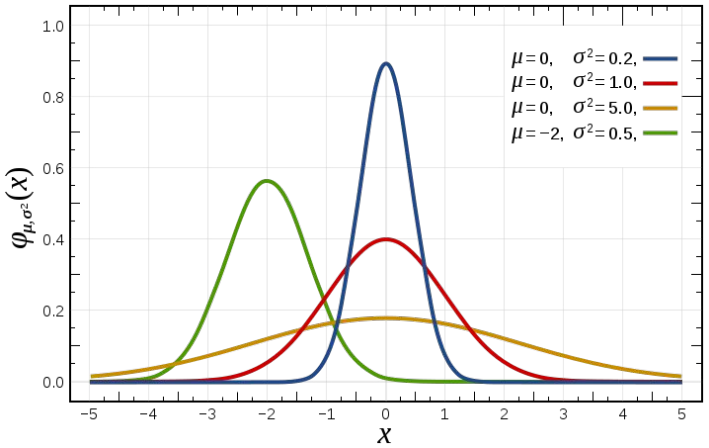
A

Gráficos: A través de histogramas, distplots o boxplots.

B

Numéricos: Media aritmética, Mediana, rango y quartiles, medidas de dispersión

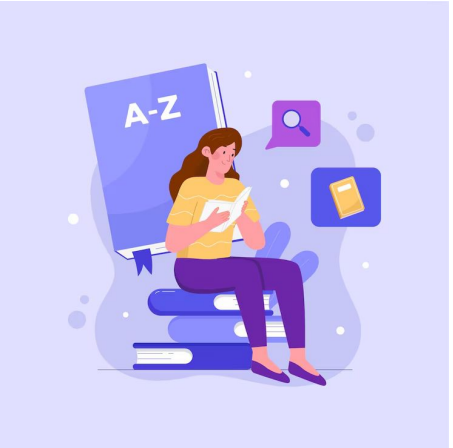




FUNDACIÓN YPF

Distribución de las variables continuas

- Permite ver si la distribución de mi variable es normal, logarítmica, sesgada a la izquierda o a la derecha, etc.
- Algunos algoritmos asumen la distribución de nuestra variable target es normal
- Para poder filtrar outliers debo saber la distribución de una variable



FUNDACIÓN YPF

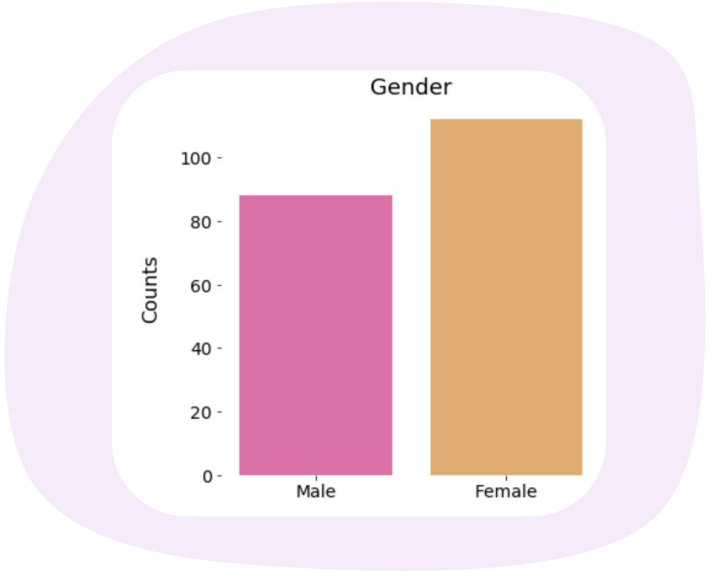
Frecuencia de cada categoría

A

Gráficos: Countplots o barras

B

Tablas de frecuencias



FUNDACIÓN YPF

Frecuencia de cada categoría

- Permite ver si cada nivel en una variable categórica está representada o si hay sobrerrepresentación de un nivel
- Los algoritmos de clasificación requieren que el dataset no tenga desbalance de clases.

FUNDACIÓN YPF

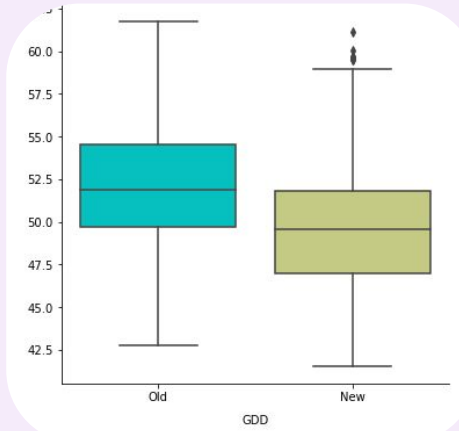
Comparar varios conjuntos de datos

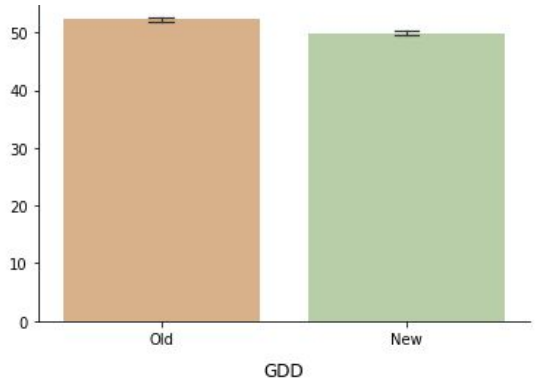
A

Gráficos: A través de boxplots, gráfico de barras.

B

Numéricos: Comparar la media aritmética, mediana, rango y quartiles, medidas de dispersión





FUNDACIÓN YPF

Evaluar relaciones: Correlación

A

Gráficos: Scatterplots, pairplots.

B

Numéricos: Análisis de correlación

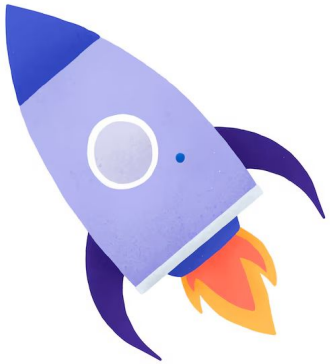
```
correlacion = df.corr(method='pearson')
```

	nombre	saldo	categoria
nombre	1.000000	0.084695	-0.028462
saldo	0.084695	1.000000	-0.023454
categoria	-0.028462	-0.023454	1.000000

FUNDACIÓN YPF

Evaluar relaciones: Correlación

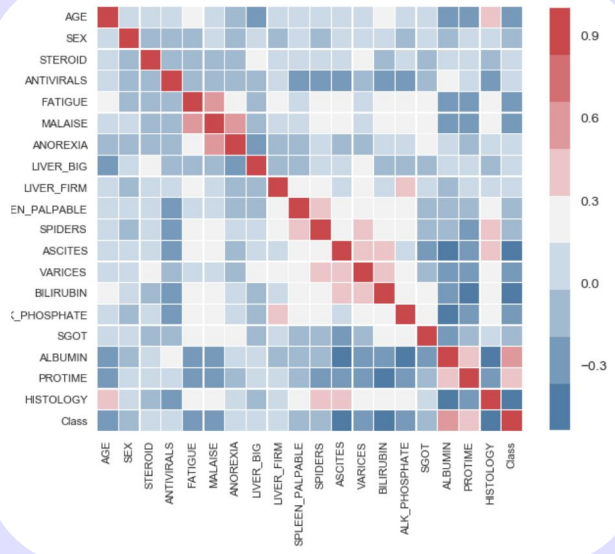
- Permite identificar la relación entre las variables independientes y la variable target.
- Permite establecer correlaciones entre las variables independientes
- Evitar COLINEALIDAD



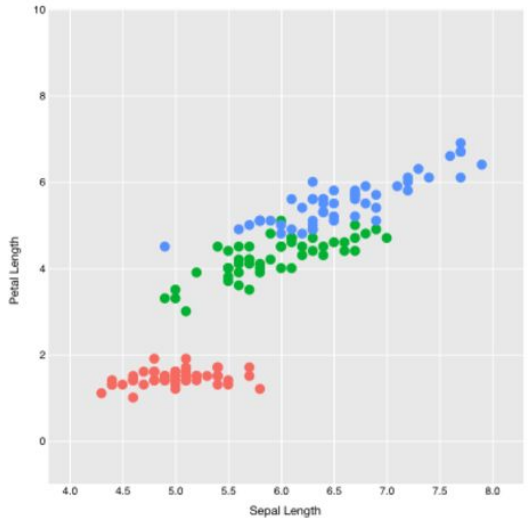
FUNDACIÓN YPF

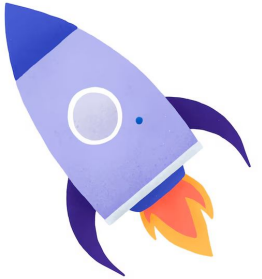
Evaluar relaciones: Correlación

Heatmap



Scatterplot





FUNDACIÓN YPF

Descanso

Nos vemos en 10 minutos

FUNDACIÓN YPF

Repasamos en Kahoot

FUNDACIÓN YPF

Juguemos un rato

Respondemos algunas preguntas en Kahoot

FUNDACIÓN YPF

¿Alguna consulta?

FUNDACIÓN YPF

DESAFÍO 5

Para la siguiente clase:

- Elegir un tema y un Dataset para el trabajo final. Publicar el mismo en el foro del aula virtual
- Además, describir cuál podría ser el objetivo de su proyecto

Pueden reconfigurar los grupos de trabajo o volver a armarlos. En tal caso, cuando entreguen el desafío, deberán colocar el nombre de las participantes e informar a la tutora

FUNDACIÓN YPF

Repasamos

FUNDACIÓN YPF

Análisis exploratorio de datos

El análisis exploratorio de datos tiene por objetivo **identificar** las principales **características** de un conjunto de datos mediante un número reducido de gráficos y/o números

FUNDACIÓN YPF

Buscamos...

Análisis

¿Cuantos features tengo (Columnas)? ¿Nombre? ¿Cuántas observaciones tengo (filas)? ¿Cada fila contiene sólo atributos de una observación?

¿Qué tipo de variables tengo?

¿Hay valores faltantes? ¿Hay valores erróneos?

FUNDACIÓN YPF

Sección práctica:

Aprendemos cómo describir datos con la primera parte de la Notebook 8

FUNDACIÓN YPF

En la sala general

Describiendo datos

Trabajamos con la primera parte de la Notebook 8

Demostraremos cómo describir los datos con Pandas.

FUNDACIÓN YPF

Descanso

Nos vemos en 10 minutos

FUNDACIÓN YPF

Herramientas para visualización de Datos

FUNDACIÓN YPF

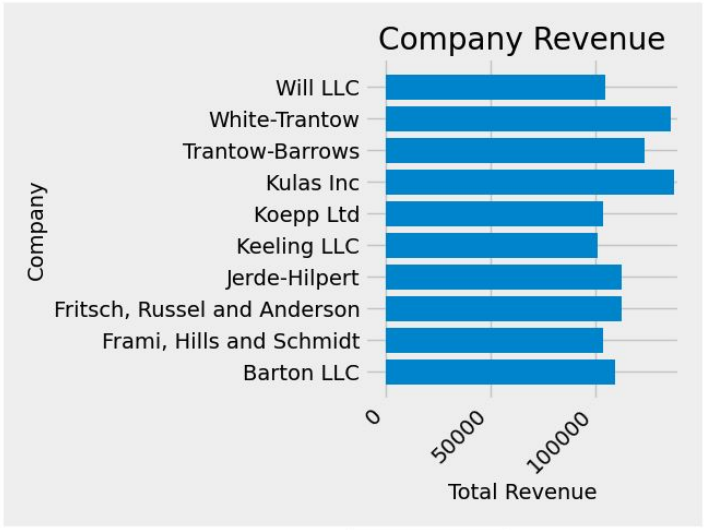
Matplotlib

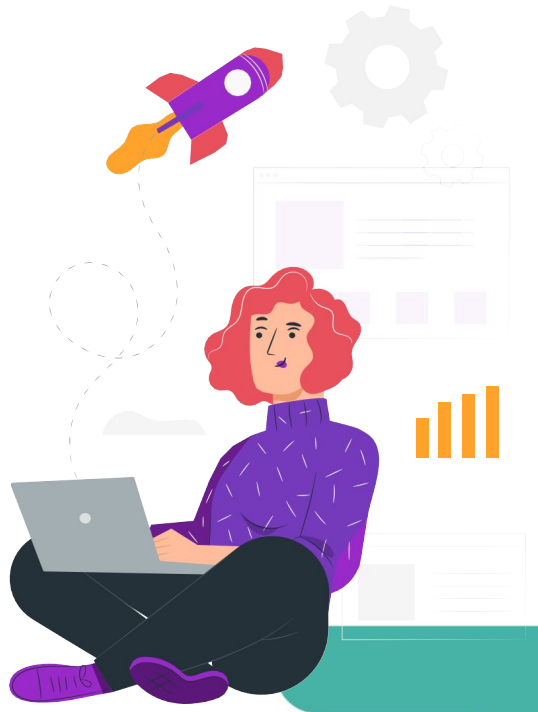
Matplotlib es una librería *open source* desarrollada por John Hunter en **2002** con el fin de imitar las funcionalidades de creación gráfica de Matlab en Python.

FUNDACIÓN YPF

¿Por qué Matplotlib?

Matplotlib ofrece una **amplia variedad de gráficos** básicos y avanzados. Brinda métodos para poder modificar y añadir elementos a los gráficos de manera muy sencilla. Ofrece tutoriales para aprender como hacer los distintos plots.





FUNDACIÓN YPF

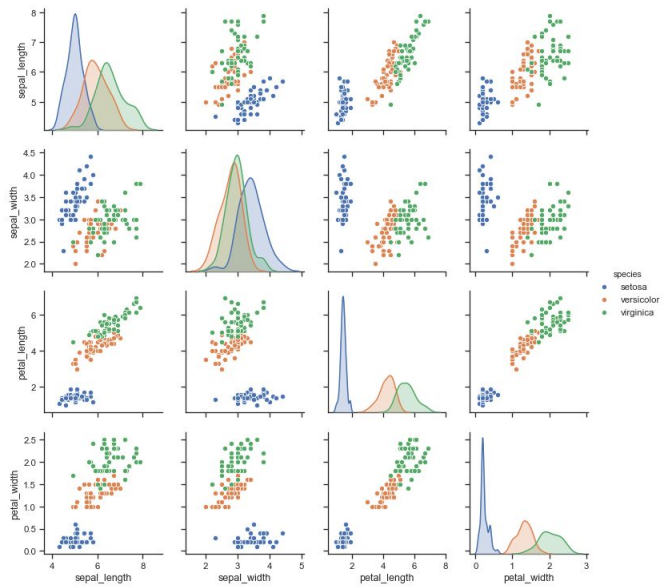
Seaborn

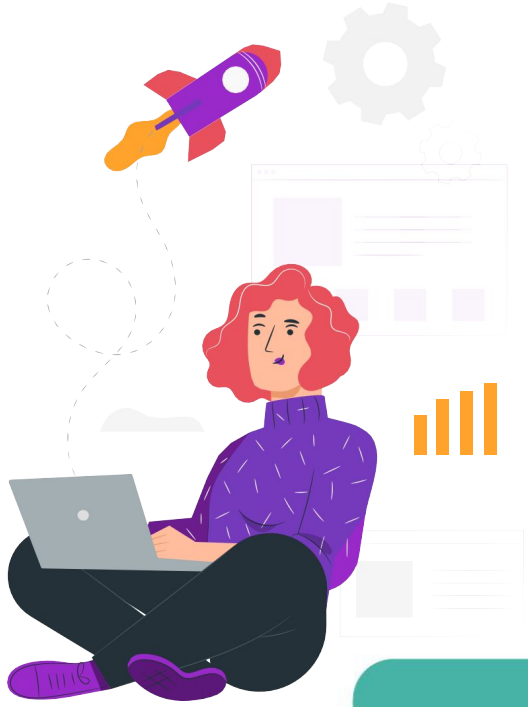
Seaborn es una librería para representación estadística desarrollada sobre Matplolib, pero que está diseñada para facilitar tanto el trabajo con los datos como la **personalización de los gráficos.**

FUNDACIÓN YPF

¿Por qué Seaborn?

Seaborn representa fácilmente distribuciones de datos o agregaciones sin tener que darle formatos complejos al DataFrame. Su galería de gráficos es más amplia que Matplotlib y está más focalizada en la estética, por lo que permite una personalización de gráficos es mayor.





FUNDACIÓN YPF

Sección práctica:

Aprendemos cómo visualizar datos con la segunda parte de la Notebook 8

FUNDACIÓN YPF

En la sala general

Visualización de datos

Trabajamos con la segunda parte de la Notebook 8

Demostraremos cómo visualizar los datos con Pandas.

FUNDACIÓN YPF

DESAFÍO 6

Para la siguiente clase:

Practicar cómo describir y visualizar datos en Pandas. Se sugiere ir avanzando con sus propios Dataset elegidos en grupo, de modo de llevar al proyecto propio lo visto en clase.

FUNDACIÓN YPF

¿Alguna consulta?

FUNDACIÓN YPF

¡Muchas gracias!