Módulo: Estadística Descriptiva con Python

Agenda

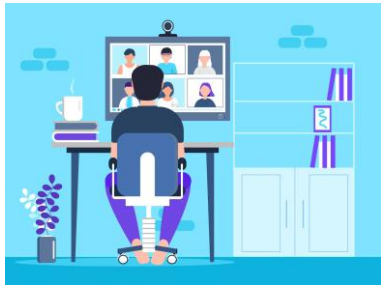
0. Reglas del Juego
1. Repaso: Introducción a la Estadística y Fases de un Proyecto
2. ¿Qué es el análisis exploratorio?
3. Análisis Univariado
4. Análisis Bivariado
5. Taller de Estadística con Python

Reglas del Juego

DATA ANALÍTICA

Mantener el micrófono apagado en caso de que no vayamos a hablar.

Nos encantaría verte.
Ten tu cámara encendida y conozcámonos virtualmente.



Preguntar en caso que tengan dudas.

Disfruta de este espacio.
Desconecta del resto y participa.

Por cada clase tendremos 10 min o 15 min de receso.


BIEN CLARAS SON LAS
REGLAS DEL JUEGO
memegenerator.es

Modo de Evaluación

20%

Evaluación continua:

Notebooks de ejercicios, formularios de ejercicios o tareas (challenges).

20%

Pruebas de Entrada:

Exámenes de 5 a 10 preguntas que serán tomados después del receso de la 3ra y 5ta sesión.

60%

Proyecto Final:

Caso aplicativo haciendo uso de las herramientas y aprendizajes obtenidos a lo largo del curso. Presentación y exposición final en la 5ta sesión.

PROYECTO FINAL

Proyecto Final

Objetivo:
- Consolidar lo aprendido durante el curso.

Contenido:
- Contexto y casos de uso (Importancia)
- Datos
- Flujo de desarrollo (Limpieza de Datos, Análisis)
- Conclusiones

Tiempo de Exposición: 10 min por grupo

Grupos: De 4 o 5 personas

Entregable:
- Diapositivas
- Notebooks (ordenado y replicable)

Fecha de presentación: Viernes 21 de abril (sesión 5)

Repositorios de Datos:

Kaggle:

https://www.kaggle.com/datasets?topic=trendingDataset

UCI ML Repository:

https://archive.ics.uci.edu/ml/datasets.php

https://archive-beta.ics.uci.edu/datasets

Repaso

INTRODUCCIÓN A LA ESTADÍSTICA

¿Cómo podemos definir la estadística?

Ciencia aplicada que nos proporciona un conjunto de técnicas o procedimientos para recopilar, organizar, analizar y presentar datos con el fin de describirlos o de realizar generalizaciones válidas

"Es el arte de analizar los datos utilizando técnicas matemáticas para resolver problemas del mundo real"

Detección de correo spam	Pronóstico del tipo de cambio	Estimación de ingresos
Intención de voto en elecciones	Pronósticos de temperatura	Detección de objetos
Detectar clientes propensos a mora	Pronósticos del ganador de un partido	Reducción de cantidad de variables

Nota: Las ciencias aplicada utilizan el conocimiento científico para resolver problemas prácticos

¿Cómo se clasifica o divide la estadística?

La estadística se clasifica en 2 grandes ramas:

Estadística Descriptiva

Técnicas para resumir y describir los datos utilizando tablas, gráficos o algunos cálculos

Estadística Inferencial

Técnicas para inferir o generalizar el conocimiento sobre una población utilizando una muestra

Estadística Paramétrica

Estadística No Paramétrica

Reportería

Muestreo / Modelos predictivos

Nota: Un modelo es una simplificación de la realidad mediante una representación matemática

¿Cuáles son los elementos de la estadística?

- **Población**: Conjunto de elementos (personas, objetos, etc) sobre el cual se quiere obtener información observando o midiendo una o más características.

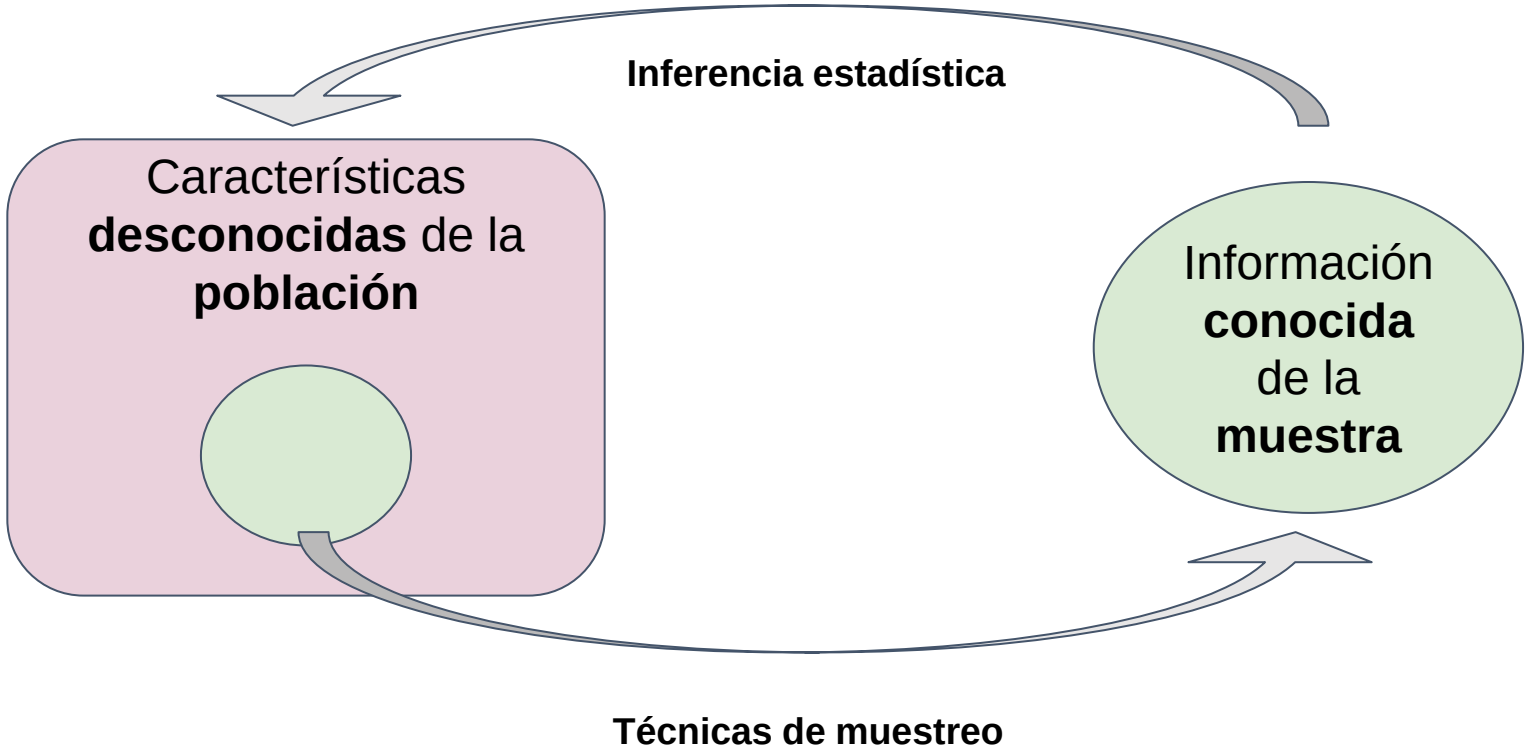
- **Muestra**: Es una parte o subconjunto de la población, el cual debe ser **representativa** y seleccionarse con **técnicas de muestreo**.

- **Parámetro**: Es una medida descriptiva que resume una característica de la **población**.

- **Estadístico**: Es una medida descriptiva que resume una característica de la **muestra**.

Para recordarlo siempre…

Ciclo básico de un análisis estadístico



Nota: Para realizar inferencia estadística se debe conocer sobre probabilidad

Tipos de Variables

CUANTITATIVAS

Discreta: Sus valores son números enteros, por ejemplo: número de alumnos, cantidad de hijos, cantidad de empleados.

Continua: Si entre dos valores, existen infinitos valores posibles, por ejemplo: altura, edad, peso.

CUALITATIVAS

Nominal: Sus valores no se pueden ordenar, por ejemplo: Estado civil, lugar de nacimiento, Fuma (Si/No), Color preferido.

Ordinal: Sus valores se pueden ordenar, por ejemplo: Nivel de satisfacción, intensidad de calor, nivel de estudios.

Repaso

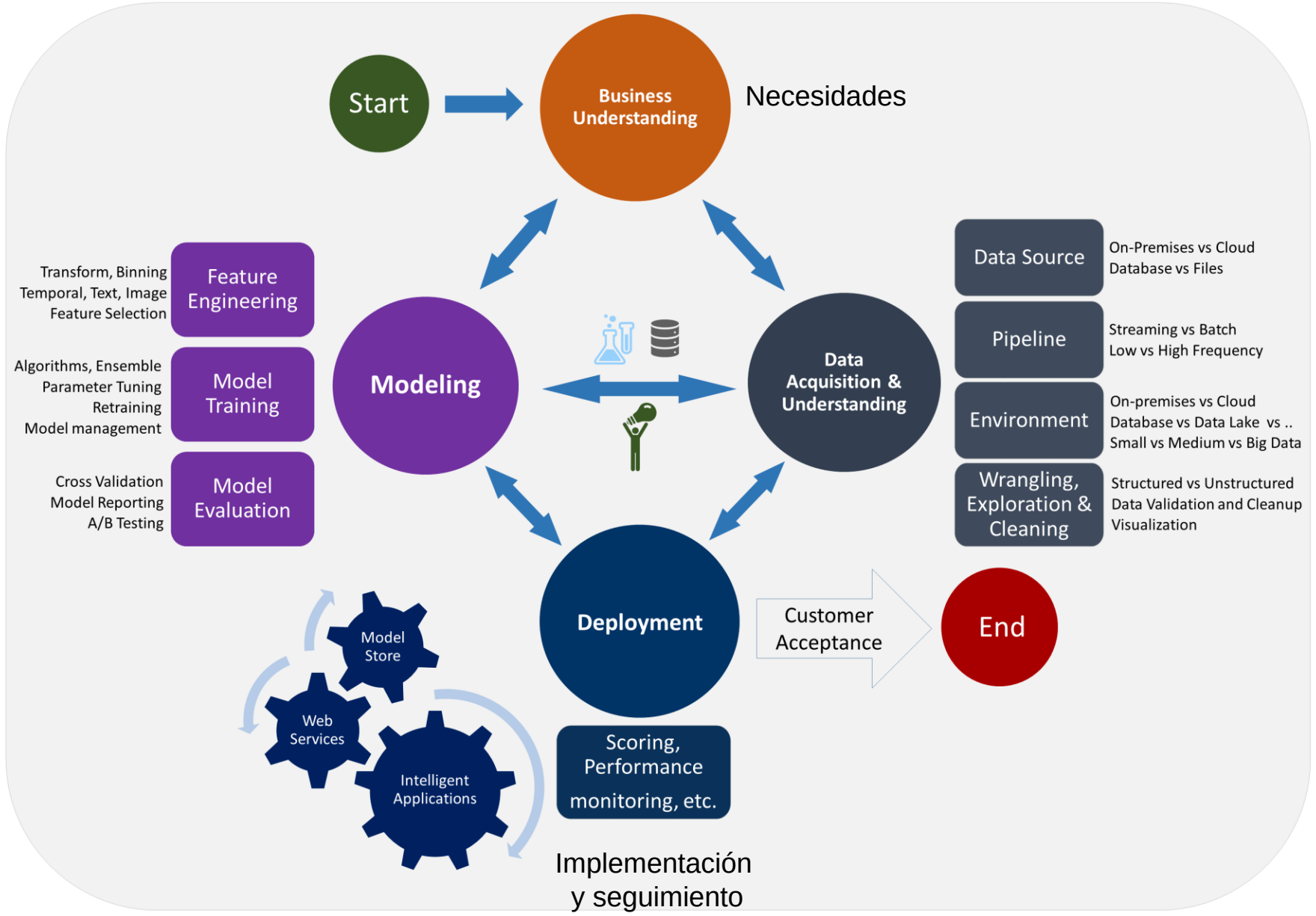
FASES DE UN PROYECTO

Creación de variables

Entrenamiento

Evaluación

Data Science Lifecycle



Fuentes de datos

Flujo de datos

Infraestructura tecnológica

Limpieza y exploración

The Team Data Science Process lifecycle https://learn.microsoft.com/en-us/azure/architecture/data-science-process/lifecycle

Proceso de análisis de datos

Identificación del problema

¿Qué problema está tratando de resolver la empresa? ¿Qué necesita medir y cómo lo medirá?

Recopilación de datos

Puede provenir de fuentes internas o de fuentes secundarias.

Limpieza de datos

Depuración de datos duplicados y anómalos, la resolución de incoherencias, la estandarización de la estructura y el formato de los datos y el tratamiento de los espacios en blanco y otros errores de sintaxis.

What Is Data Analysis? (With Examples) https://www.coursera.org/articles/what-is-data-analysis-with-examples

Proceso de análisis de datos

Análisis de datos

Hacemos uso de diferentes técnicas y herramientas de análisis de datos

Interpretación de resultados

¿Qué recomendaciones puede hacer con base en los datos? ¿Cuáles son las limitaciones de sus conclusiones?

What Is Data Analysis? (With Examples) https://www.coursera.org/articles/what-is-data-analysis-with-examples

NIVELES DE ANÁLISIS

Niveles de análisis que necesita para una mejor toma de decisiones

Análisis descriptivo

¿Que sucedió?

Análisis de diagnóstico

¿Por qué sucedió?

Análisis predictivo

¿Qué podría pasar en el futuro?

Análisis prescriptivo

¿Qué acción tomar?

ANÁLISIS EXPLORATORIO

¿Qué es el análisis exploratorio (EDA)?

- Metodología de análisis de datos que implica el **uso de estadísticas y visualizaciones** para entender los datos, descubrir patrones, tendencias y relaciones entre variables.
- Su objetivo principal es obtener una comprensión profunda de los datos antes de aplicar algún algoritmo de machine learning.
- Ayuda a identificar posibles problemas con los datos, como valores atípicos, errores de medición, distribuciones sesgadas y otros.

Análisis exploratorio de datos – IBM : https://www.ibm.com/es-es/cloud/learn/exploratory-data-analysis
¿Qué es el análisis exploratorio de datos o EDA? – Unipython: https://unipython.com/que-es-el-analisis-exploratorio-de-datos/

Fases del Análisis Exploratorio

- Identificación de variables
- Análisis Univariado
- Análisis Bivariado
- Tratamiento de valores perdidos
- Tratamiento de outliers
- Transformación de variables
- Creación de variables

QUIZ TIME

Medidas de Tendencia Central

Medidas de Tendencia Central

Imaginemos que tenemos un grupo de 7 estudiantes:

Edades: 7, 8, 10, 8, 6, 15, 12

Medidas de Tendencia Central

Edades: 7, 8, 10, 8, 6, 15, 12

MEDIA: Para nuestro ejemplo, la media será:

(Suma de las edades) / (Cant. de estudiantes) = (7 + 8 + 10 + 8 + 6 + 15 + 12) / 7 = 9.43

MEDIANA:

Para su cálculo tenemos que ordenar los valores de menor a mayor:

6, 7, 8, 8, 10, 12, 15

Dado que el número de estudiantes que conocemos es impar (7 estudiantes) entonces la Mediana es el valor de centro. En este caso la edad de la 4ta posición, 8 años.

Medidas de Tendencia Central

Edades: 7, 8, 10, 8, 6, 15, 12

MODA:

¿Cuál es el valor que más se repite?.

Para nuestro caso la edad que más se repite es 8 años. Dos estudiantes tienen dicha edad.

¿Cómo calculo la mediana si tengo una cantidad par de números?

Imaginemos que tenemos un grupo de 8 estudiantes

Edades: 7, 8, 10, 8, 6, 15, 12, 18

Para su cálculo tenemos que ordenar los valores de menor a mayor:
6, 7, 8, 8, 10, 12, 15, 8

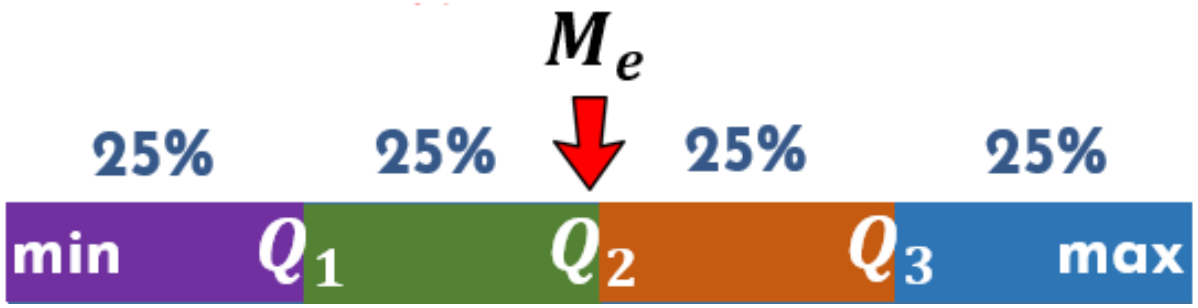
Dado que el **número de estudiantes que conocemos es par** (8 estudiantes) entonces la **Mediana es el valor promedio entre los número centrales (4ta y 5ta posición).**
En este caso la edad de la 4ta posición, 8 años y la edad de la 5ta posición es 10 años.

Por lo que el valor de la mediana es:
(8+10)/2 = 9

Medidas de Posición

Medidas de Posición

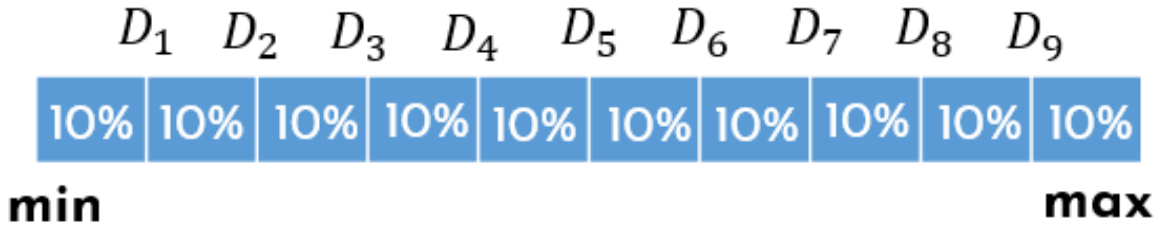
Cuartiles: Es cada uno de los tres valores que pueden dividir un grupo de números, ordenados de menor a mayor, **en cuatro partes iguales**. En otras palabras, cada cuartil determina la separación entre uno y otro subgrupo.



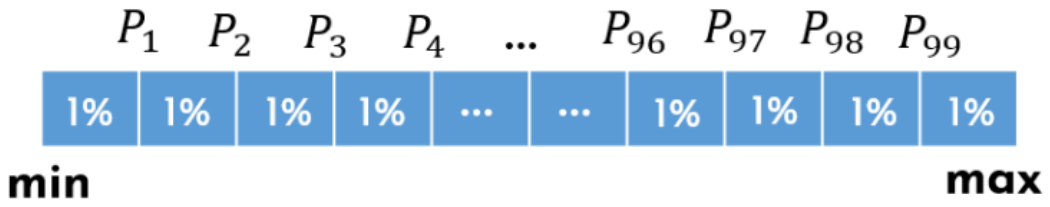
Lasmatesfaciles - Cuartiles deciles y percentiles para datos agrupados https://lasmatesfaciles.com/2021/06/21/cuartiles-deciles-y-percentiles-para-datos-agrupados/

Medidas de Posición

Deciles: Estamos ante un cuantil que divide los datos en diez partes iguales. Existen nueve deciles, de D1 a D9. El D5 se corresponde con la mediana. Por su lado, los valores superiores e inferiores (equivalentes a los diferentes cuartiles) se sitúan en puntos intermedios entre estos.



El percentil: Por último, este cuantil divide la distribución en cien partes. Hay 99 percentiles. Tiene, a su vez, una equivalencia con los deciles y cuartiles.



Economipedia - Medidas de posición https://economipedia.com/definiciones/medidas-de-posicion.html
Lasmatesfaciles - Cuartiles deciles y percentiles para datos agrupados https://lasmatesfaciles.com/2021/06/21/cuartiles-deciles-y-percentiles-para-datos-agrupados/

Medidas de Dispersión

Medidas de Dispersión

Varianza: Representa la variabilidad de los datos respecto a su media.

$$\sigma_x^2 = \frac{\sum(x_i - \overline{x})^2}{N} \qquad S_x^2 = \frac{\sum(x_i - \overline{x})^2}{n-1}$$

Desviación estandar: Es la raíz cuadrada positiva de la varianza.

Medidas de Dispersión

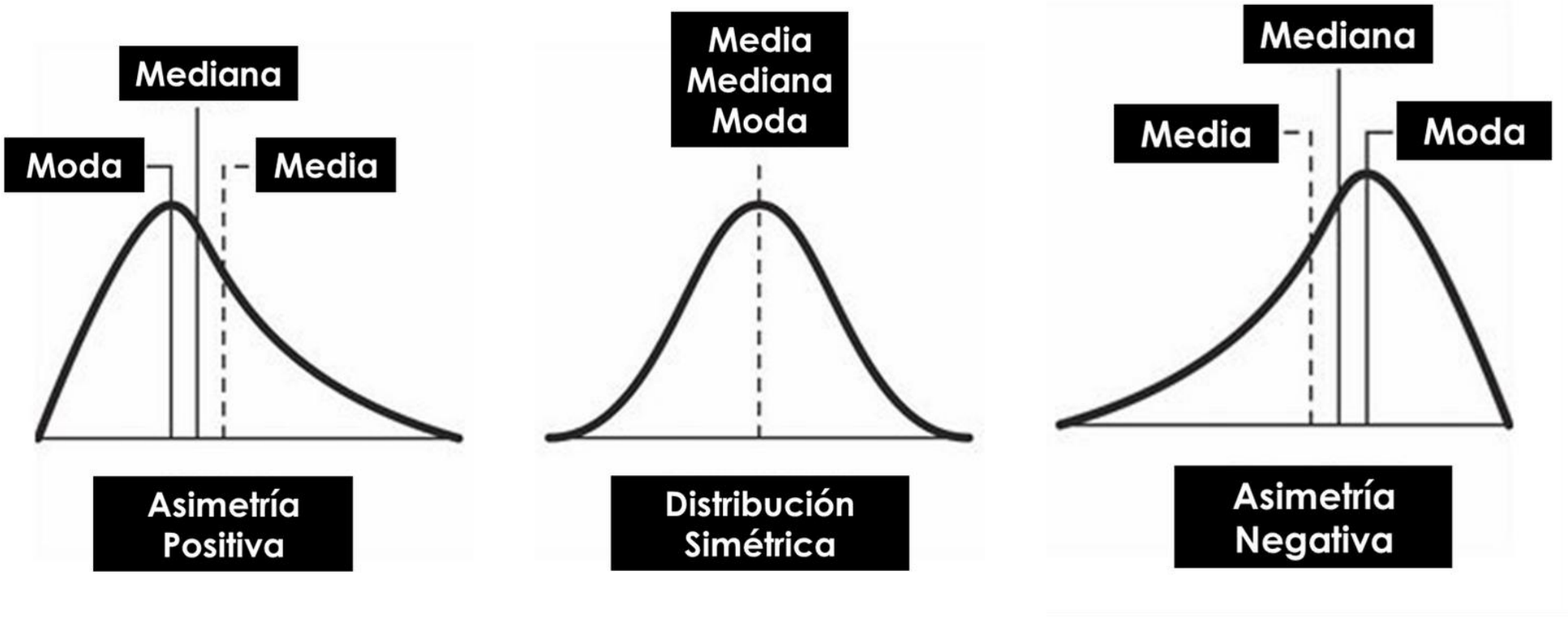
Rango: Max - Min

Rango Intercuartílico: Q3 – Q1

Medidas de Forma

Medidas de Forma

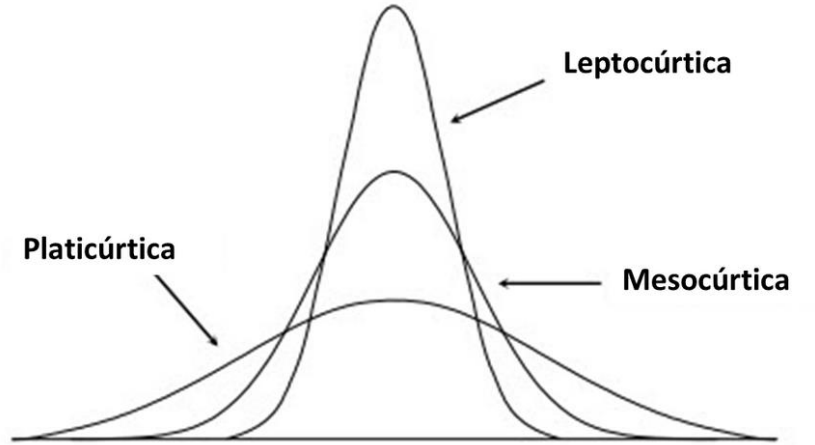
Asimetría:



Statistics for Data Science: What is Skewness and Why is it Important? https://www.analyticsvidhya.com/blog/2020/07/what-is-skewness-statistics/

Medidas de Forma

Kurtosis: Describe la forma de la distribución de la variable y que tan pesadas son sus colas de la distribución. Ayuda a identificar si hay valores extremos.



Covarianza y correlación

Covarianza y correlación

Covarianza: Da información sobre cómo X y Y están estadísticamente relacionados. Nos permite saber cómo se comporta una variable en función de lo que hace otra variable.

$$\operatorname{Cov}(X,Y) = E\big[(X - EX)(Y - EY)\big] = E[XY] - (EX)(EY).$$

$$Cov(X,Y) = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Economipedia – Covarianza: https://economipedia.com/definiciones/covarianza.html
Probability course - Covariance and Correlation https://www.probabilitycourse.com/chapter5/5_3_1_covariance_correlation.php

Covarianza y correlación

Coeficiente de correlación:

$$\rho_{xy} = \frac{Cov_{xy}}{\sigma_x \sigma_y}$$

Interpretación:

- Si ρ(X,Y)=0 decimos que X e Y no están correlacionados.
- Si ρ(X,Y)>0 decimos que X e Y están correlacionados positivamente.
- Si ρ(X,Y)<0 decimos que X e Y están negativamente correlacionados.

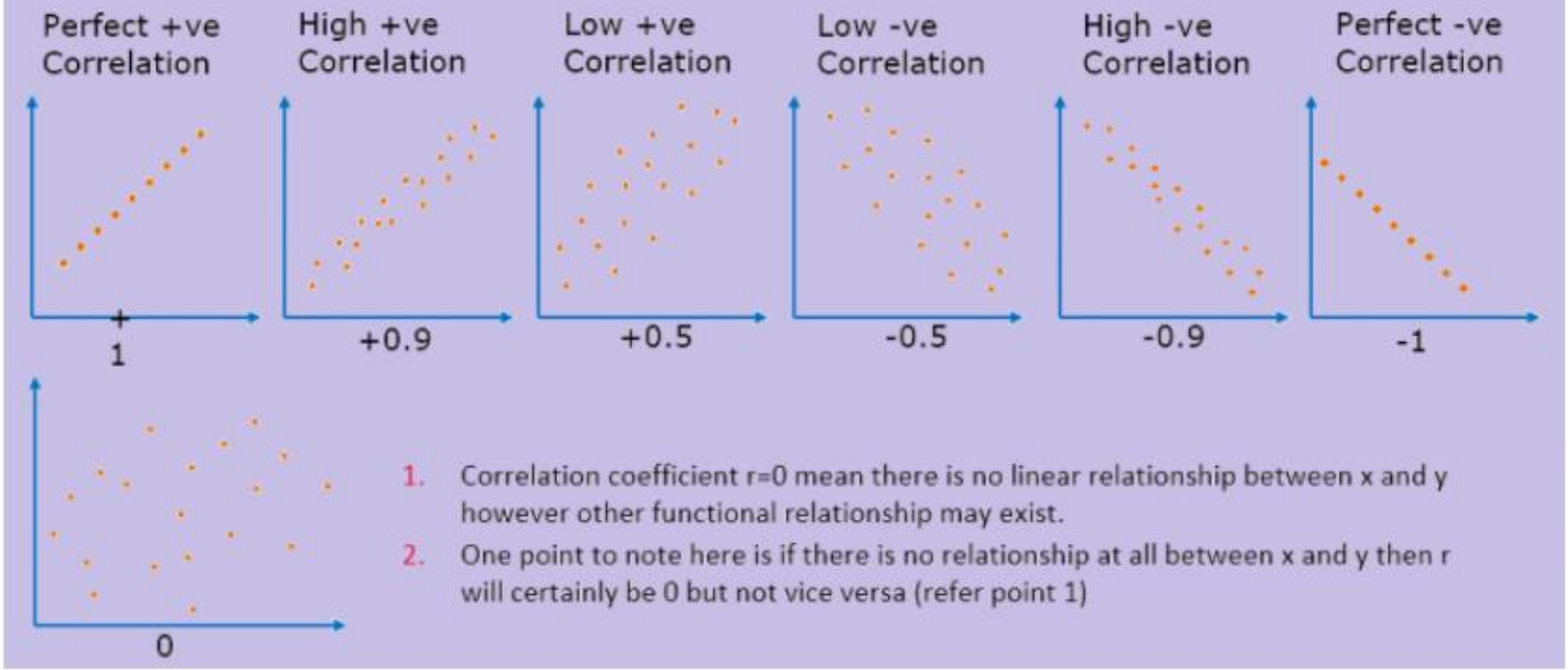
Probability course - Covariance and Correlation https://www.probabilitycourse.com/chapter5/5_3_1_covariance_correlation.php

Correlation coefficient r is number between -1 to +1 and tells us how well a regression line fits the data and defined by

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where,

- s_{xy} is the covariance between x and y
- s_x and s_y are the standard deviations of x and y respectively.

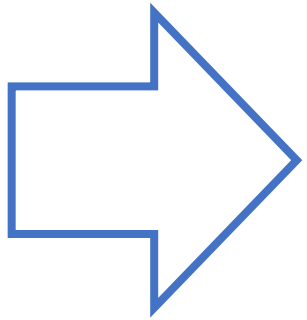


1. Correlation coefficient r=0 mean there is no linear relationship between x and y however other functional relationship may exist.
2. One point to note here is if there is no relationship at all between x and y then r will certainly be 0 but not vice versa (refer point 1)

Covarianza y correlación: comprendiendo su utilidad https://sitiobigdata.com/2019/10/26/covarianza-y-correlacion/#

Ejemplo:

	ALTURA (cm)	PESO (Kg)
1	162.9	63.8
2	173.1	71.5
3	178.6	76.2
4	154.3	58.8
5	150.1	45.8
6	172.1	64.6
7	159.6	52.5
8	177.9	69.4
9	173.6	65.2
10	157.5	61.5
11	165.8	60.0
12	150.3	53.4
13	171.3	62.1
14	151.3	56.5
15	150.2	54.0
16	155.0	58.0
17	155.7	50.4
18	172.6	73.0
19	157.7	56.8
20	174.7	66.5



Covarianza

66.8

Desviación Altura

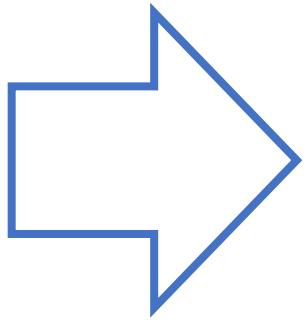
9.9

Desviación Peso

7.8

Ejemplo:

	ALTURA (cm)	PESO (Kg)
1	162.9	63.8
2	173.1	71.5
3	178.6	76.2
4	154.3	58.8
5	150.1	45.8
6	172.1	64.6
7	159.6	52.5
8	177.9	69.4
9	173.6	65.2
10	157.5	61.5
11	165.8	60.0
12	150.3	53.4
13	171.3	62.1
14	151.3	56.5
15	150.2	54.0
16	155.0	58.0
17	155.7	50.4
18	172.6	73.0
19	157.7	56.8
20	174.7	66.5



Coeficiente de correlación

0.8

QUIZ TIME

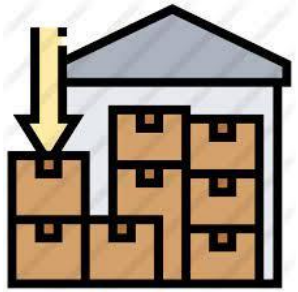
Taller Estadística Aplicada

Análisis de Datos en E-commerce

Sitio Web

Pago

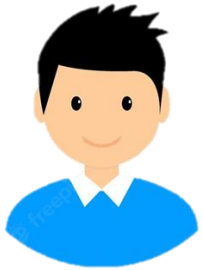
Información en la base de datos



Almacén

Delivery

Entrega del producto

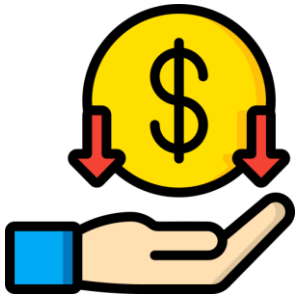


Cliente

¿Cómo nos puede ayudar el análisis de datos?

Sitio web
¿Cantidad de visitantes en la web?
¿Qué productos son los más visitados?
¿Qué marcas son las más visitadas?

Ventas
¿Qué productos son los más vendidos?
¿De qué lugar provienen las ventas?



Reducción de costos
¿Qué productos son rentables o no?

Nuevas oportunidades

Fidelización de clientes

CASO PRÁCTICO

Caso 1

¿Cuál es la cantidad promedio de productos vendidos por transacción?

¿Cuál es el promedio de precio total por transacción?

Caso 2

¿Cuál es la cantidad de ventas por país?

¿En qué país genera la mayor cantidad de ventas?

CASO PRÁCTICO

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/01/2010 08:26	2.55	17850	United Kingdom
536365	71053	WHITE METAL LANTERN	6	12/01/2010 08:26	3.39	17850	United Kingdom
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/01/2010 08:26	2.75	17850	United Kingdom
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/01/2010 08:26	3.39	17850	United Kingdom
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/01/2010 08:26	3.39	17850	United Kingdom
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/01/2010 08:26	7.65	17850	United Kingdom
536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/01/2010 08:26	4.25	17850	United Kingdom
536366	22633	HAND WARMER UNION JACK	6	12/01/2010 08:28	1.85	17850	United Kingdom
536366	22632	HAND WARMER RED POLKA DOT	6	12/01/2010 08:28	1.85	17850	United Kingdom
536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	12/01/2010 08:34	1.69	13047	United Kingdom
536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	12/01/2010 08:34	2.1	13047	United Kingdom
536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	12/01/2010 08:34	2.1	13047	United Kingdom
536367	22749	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	12/01/2010 08:34	3.75	13047	United Kingdom
536367	22310	IVORY KNITTED MUG COSY	6	12/01/2010 08:34	1.65	13047	United Kingdom
536367	84969	BOX OF 6 ASSORTED COLOUR TEASPOONS	6	12/01/2010 08:34	4.25	13047	United Kingdom
536367	22623	BOX OF VINTAGE JIGSAW BLOCKS	3	12/01/2010 08:34	4.95	13047	United Kingdom
536367	22622	BOX OF VINTAGE ALPHABET BLOCKS	2	12/01/2010 08:34	9.95	13047	United Kingdom
536367	21754	HOME BUILDING BLOCK WORD	3	12/01/2010 08:34	5.95	13047	United Kingdom
536367	21755	LOVE BUILDING BLOCK WORD	3	12/01/2010 08:34	5.95	13047	United Kingdom
536367	21777	RECIPE BOX WITH METAL HEART	4	12/01/2010 08:34	7.95	13047	United Kingdom
536367	48187	DOORMAT NEW ENGLAND	4	12/01/2010 08:34	7.95	13047	United Kingdom

Error Nº 1

Asumir que tus datos están listos para utilizarse

Error Nº 2

No explorar tu conjunto de datos antes de empezar a trabajar con ellos.

Ejercicios 1:

Resolver los 10 ejercicios del notebook “Taller de Estadística Aplicada_Análisis_Ecommerce.ipynb”

Enviarlo por correo con el asunto: Ejercicios 1 – Módulo Estadística Descriptiva con Python – [Apellidos y nombres]

Correo: team@dataanalitica.net
Fecha entrega: 22-marzo hasta las 10 pm

Challenge 1:

- Formar sus grupos. Indicar el listado de integrantes.
- Buscar y elegir una data para el proyecto final
- Plantear la temática y al menos 4 preguntas a responder.

Enviar su presentación (PPT) por correo con el asunto: Challenge 1 – Módulo Estadística Descriptiva con Python – [Nombre de grupo]

Subirlo también al drive: https://drive.google.com/drive/folders/1kry4hSOEBqhHMqar6CnFvy8KUbOc2sZ7?usp=sharing

Correo: team@dataanalitica.net
Fecha entrega: 22-marzo hasta las 10 pm

Créditos y recursos importantes:

Data Original:

https://www.kaggle.com/carrie1/ecommerce-data

Notebooks revisados:

https://www.kaggle.com/code/sercanyesiloz/crm-analytics/notebook

https://www.kaggle.com/code/pierrelouisdanieau/recommender-system-associations-rules/notebook

12 errores de Data Science que puedes evitar… explicados con memes ;)

https://keyrus.com/sp/es/insights/12-errores-de-data-science-que-puedes-evitar-explicados-con-memes

GRACIAS

DATA ANALÍTICA