

Analítica con Exploración de Base de Datos.

« Divide las dificultades que examinas en tantas partes como sea posible , para su mejor solución»



Agenda



- Introducción al AED.
- Etapas del AED.
- Imputación de Variables.
- Transformación de Variables.

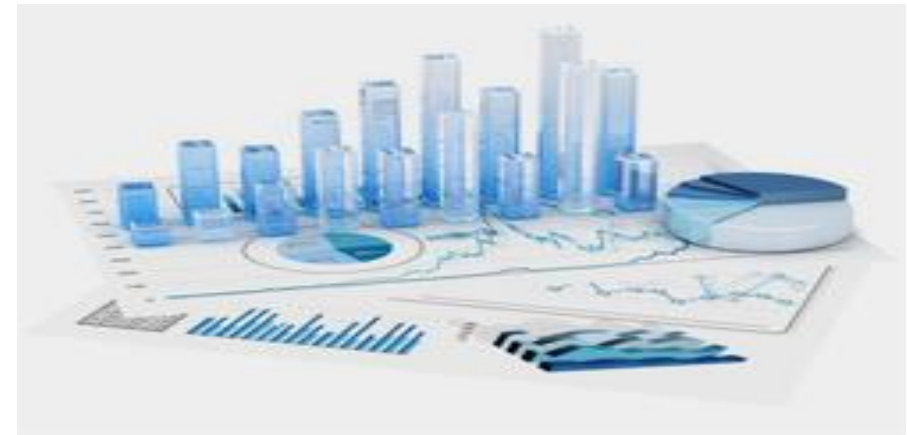


Análisis Exploratorio de Datos



I. Introducción

- La finalidad del AED es examinar a detalle los datos previamente a la aplicación de cualquier técnica o algoritmo.
- De esta forma el analista consigue un entendimiento básico de sus datos y de las relaciones existentes entre las variables analizadas.
- El examen previo de los datos es un paso necesario, que lleva tiempo, y que habitualmente se descuida por parte de los analistas de datos.



¿Qué es el AED?

- El Análisis Exploratorio de Datos es un conjunto de técnicas estadísticas cuya finalidad es conseguir un entendimiento básico de los datos y de las relaciones existentes entre las variables analizadas.
- Para conseguir este objetivo el A.E.D. proporciona métodos sistemáticos sencillos para organizar y preparar los datos, detectar fallas en el diseño y recogida de los mismos, tratamiento y evaluación de datos ausentes (missing), identificación de casos atípicos (outliers) y comprobación de algunos supuestos.



ETAPAS DEL AED

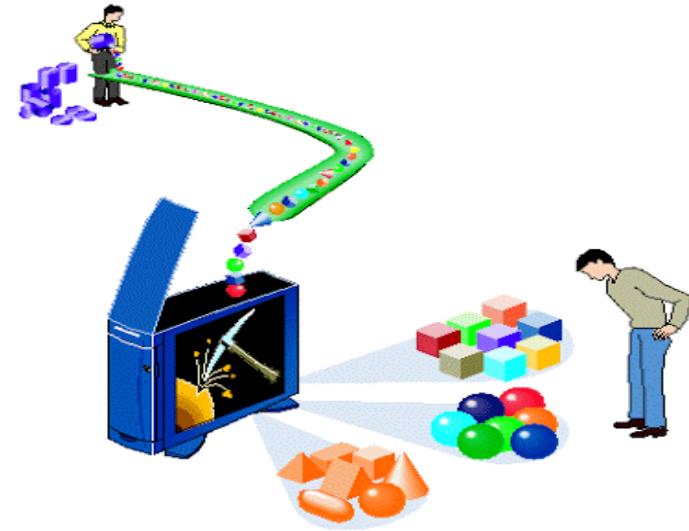
- Entendimiento del contexto del problema que uno pretende resolver.
- Preparación y Valor Agregado de los Datos.
- Examen gráfico para evaluar la naturaleza de las variables individuales y un análisis descriptivo numérico que permita cuantificar algunos aspectos de los datos.
- Examen gráfico de las asociaciones entre las variables analizadas , tanto para las variables cuantitativas como cualitativas.
- Identificar los posibles casos atípicos (outliers) y evaluar el impacto potencial que puedan ejercer en análisis estadísticos posteriores.
- Evaluar, si fuera necesario, el impacto potencial que pueden tener los datos ausentes (missing) sobre la representatividad de los datos analizados.

ETAPA 1: entendimiento contextual del problema a desarrollar



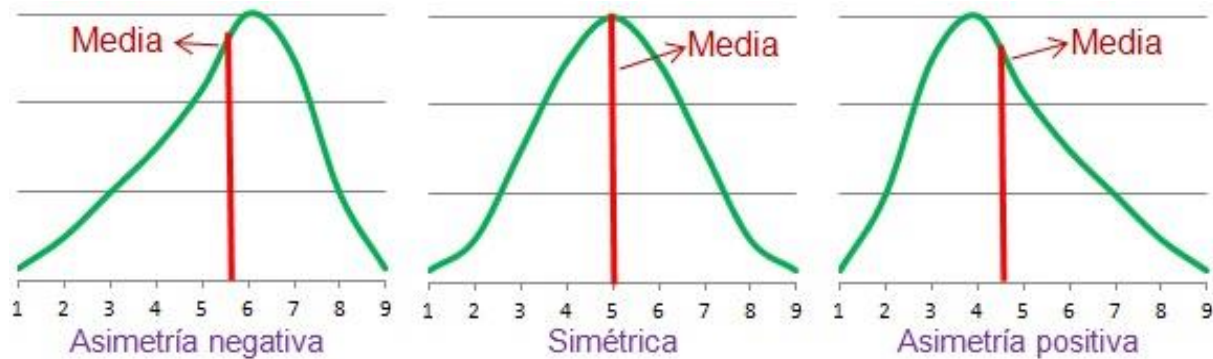
ETAPA 2: PREPARACIÓN Y VALOR AGREGADO DE LOS DATOS

- ✓ Combinar conjuntos de datos de dos archivos distintos
- ✓ Seleccionar subconjuntos de los datos
- ✓ Dividir el archivo de los datos en varias partes
- ✓ Transformar variables
- ✓ Ordenar casos
- ✓ Agregar nuevos datos y/o variables
- ✓ Eliminar datos y/o variables
- ✓ Guardar datos y/o resultados

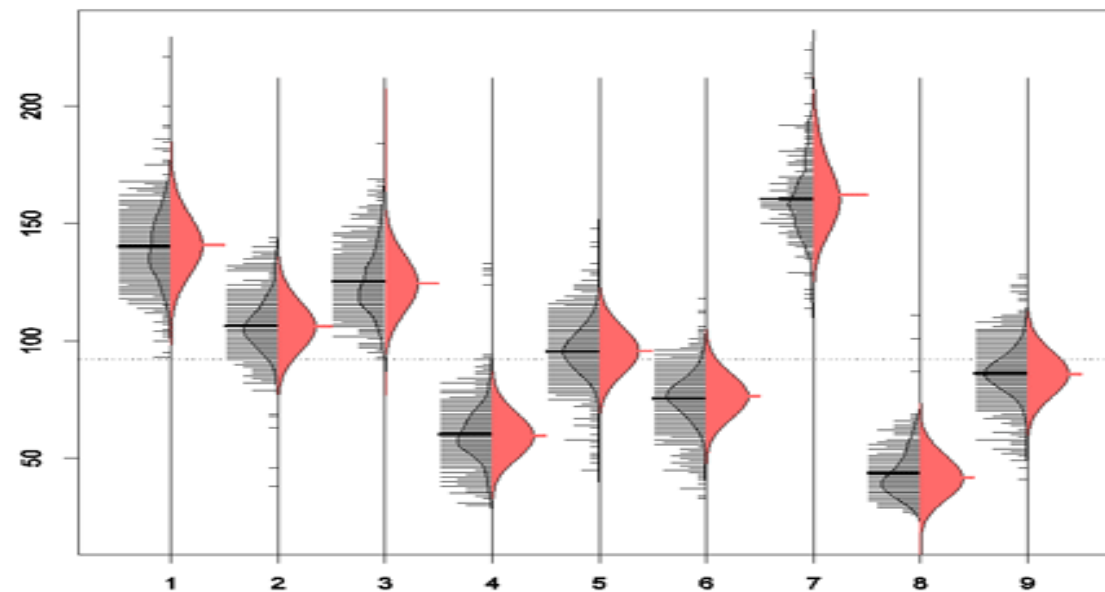


ETAPA 3: examen gráfico y descriptivo numérico

ESCALA DE MEDIDA	REPRESENTACIÓN GRÁFICA	MEDIDA DE TENDENCIA CENTRAL	MEDIDA DE DISPERSIÓN
NOMINAL	Diagrama de barras, líneas , pictograma y sectores	Moda	
ORDINAL	Gráficos de cajas ,violín	Moda, Mediana, Media trunca	Rango intercuartílico,
			CVQ
INTERVALO	Histograma, polígonos de frecuencias, Gráficos de cajas ,violín	Moda, Media, Mediana	Desviación estándar, Rango intercuartílico,
			CVQ
RAZÓN	Histograma, polígonos de frecuencias, Gráficos de cajas ,violín	Moda, Mediana, Media, Media geométrica	Desviación estándar,
			Coeficiente de variación, Rango intercuartílico,
			CVQ

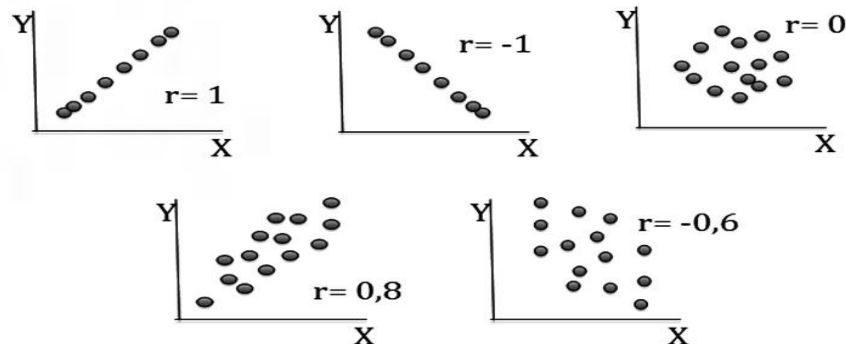


<u>Simetría</u>	<u>Relación</u>
Simétrica o insesgada	$\text{Moda} = \text{Mediana} = \text{Media}$
Asimétrica o con sesgo positivo a la derecha	$\text{Moda} > \text{Mediana} > \text{Media}$
Asimétrica o con sesgo negativo a la izquierda	$\text{Moda} < \text{Mediana} < \text{Media}$



ETAPA 4: asociaciones cualitativas y cuantitativas

Variable X	Variable Y	Coefficiente de correlación
Cualitativa	Cualitativa	Chi-cuadrado Contingencia Phi
Ordinal	Ordinal	Spearman
Cualitativa	Cuantitativa	Biserial-puntual
Cuantitativa	Cuantitativa	Pearson



		EDAD				Total
		MENOS DE 30 AÑOS	ENTRE 30 Y 45	ENTRE 45 Y 60	MÁS DE 60 AÑOS	
IMPRESIÓN	MUY BUENA	40,4%	43,2%	43,2%	46,9%	42,1%
	BUENA	45,6%	43,3%	44,1%	38,9%	44,3%
	NORMAL	12,5%	12,3%	11,5%	12,8%	12,3%
	MALA	1,5%	1,3%	1,2%	1,4%	1,4%
Total		100,0%	100,0%	100,0%	100,0%	100,0%

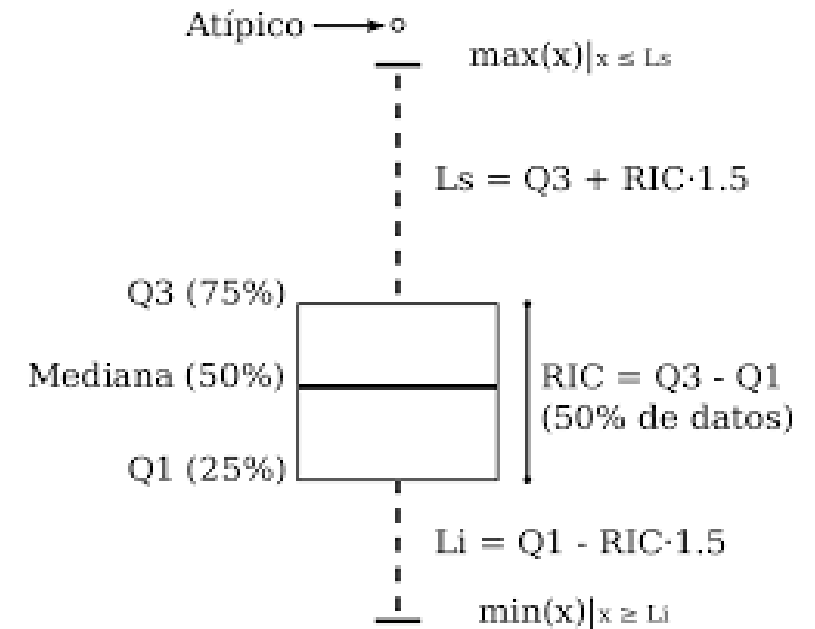
ETAPA 5: casos atípicos univariados y multivariados

- Los casos atípicos son observaciones con características diferentes de las demás.
- Este tipo de casos no pueden ser caracterizados categóricamente como benéficos o problemáticos sino que deben ser contemplados en el contexto del análisis y debe evaluarse el tipo de información que pueden proporcionar.
-
- Por otra parte, aunque diferentes a la mayor parte de la muestra, pueden ser indicativos de las características de un segmento válido de la población.

Tipos de Atípicos

Los casos atípicos pueden clasificarse en 2 categorías:

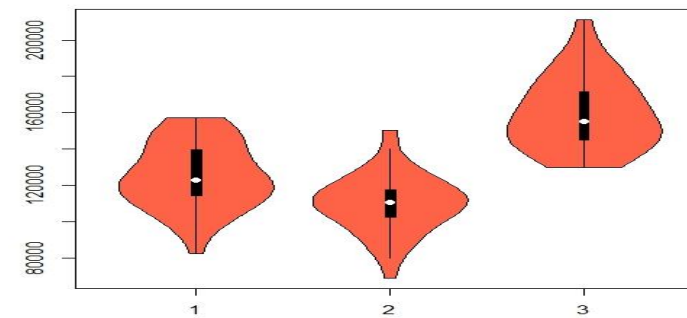
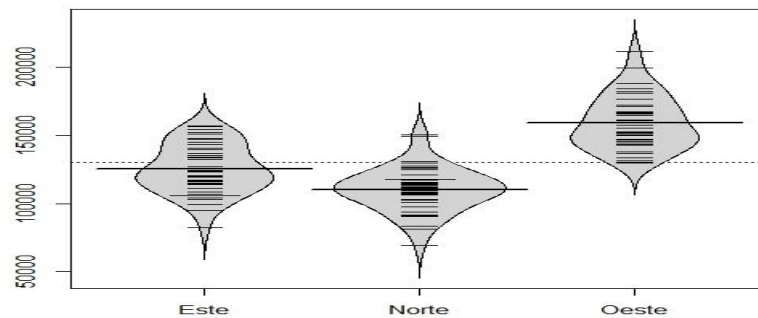
- **La primera categoría** contiene aquellos casos atípicos que surgen de un error de procedimiento, tales como la entrada de datos o un error de codificación. Estos casos atípicos deberían subsanarse en el filtrado de los datos, y si no se puede, deberían eliminarse del análisis o recodificarse como datos ausentes.
- **La segunda clase** es la observación que ocurre como consecuencia de un acontecimiento extraordinario. En este caso, el outlier no representa ningún segmento válido de la población y puede ser eliminado del análisis.



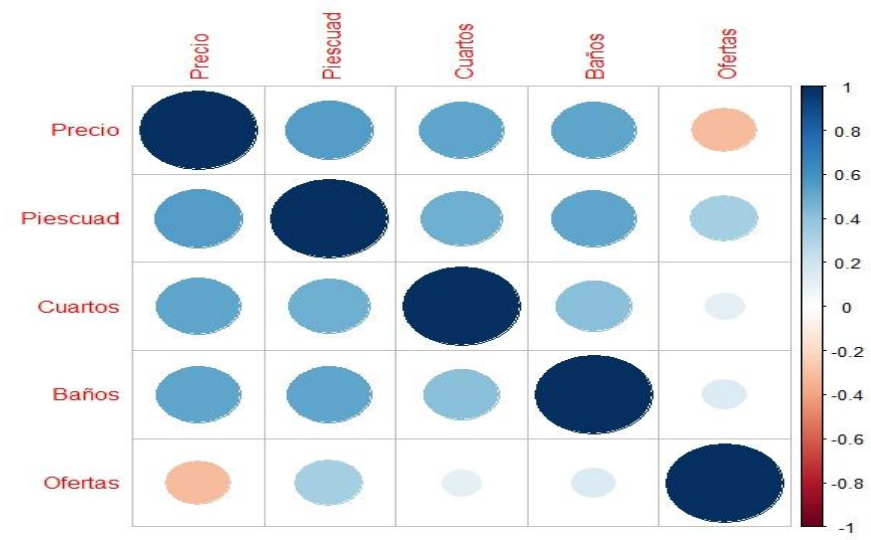
GRÁFICOS ESPECIALES

El **beanplot** es una combinación de un gráfico de densidad (doble) con las marcas de todos los datos. Dichas marcas cambian de color si se salen del interior de la doble densidad y se alargan si coinciden algunos datos con el mismo valor.

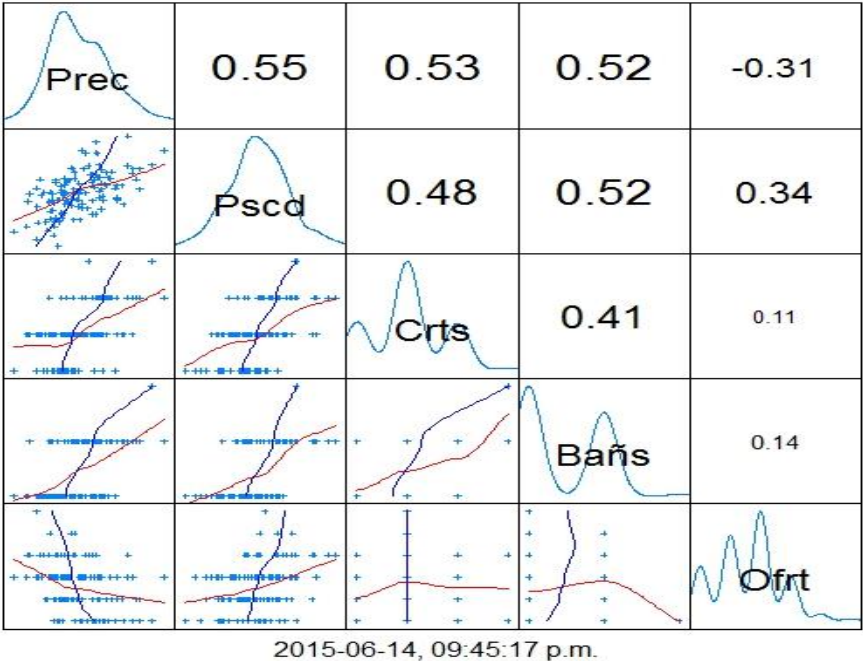
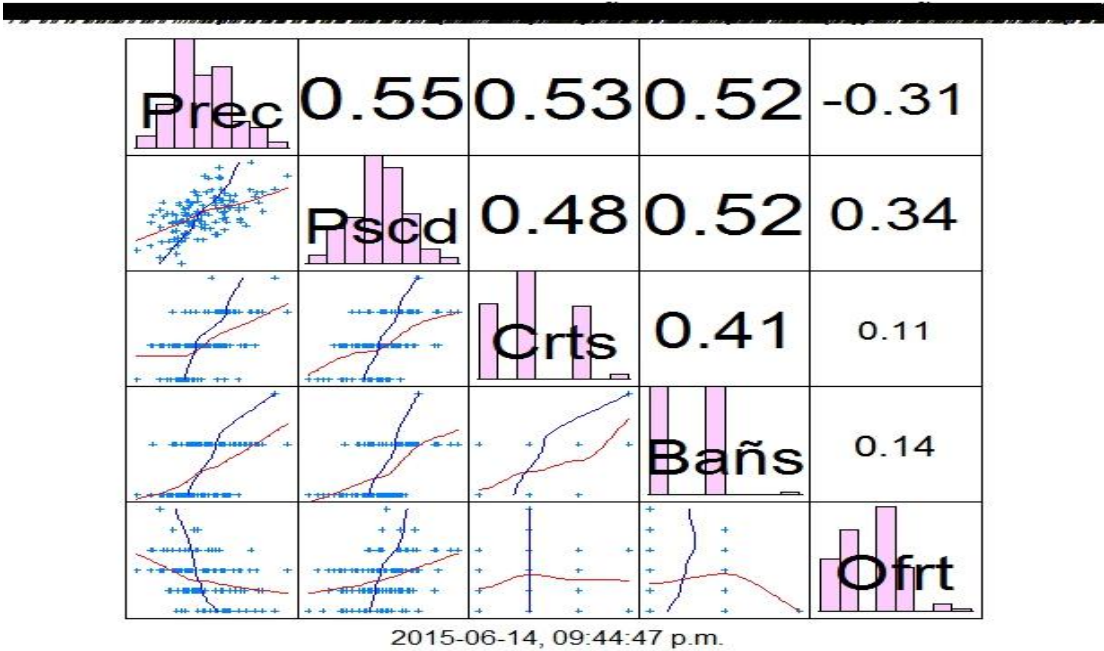
Para poder comparar los grupos, se señalan las medias de cada grupo y la media general. Por otra parte, si en la población general hay un factor con dos niveles, se puede considerar un beanplot asimétrico con dos densidades distintas en función del factor.



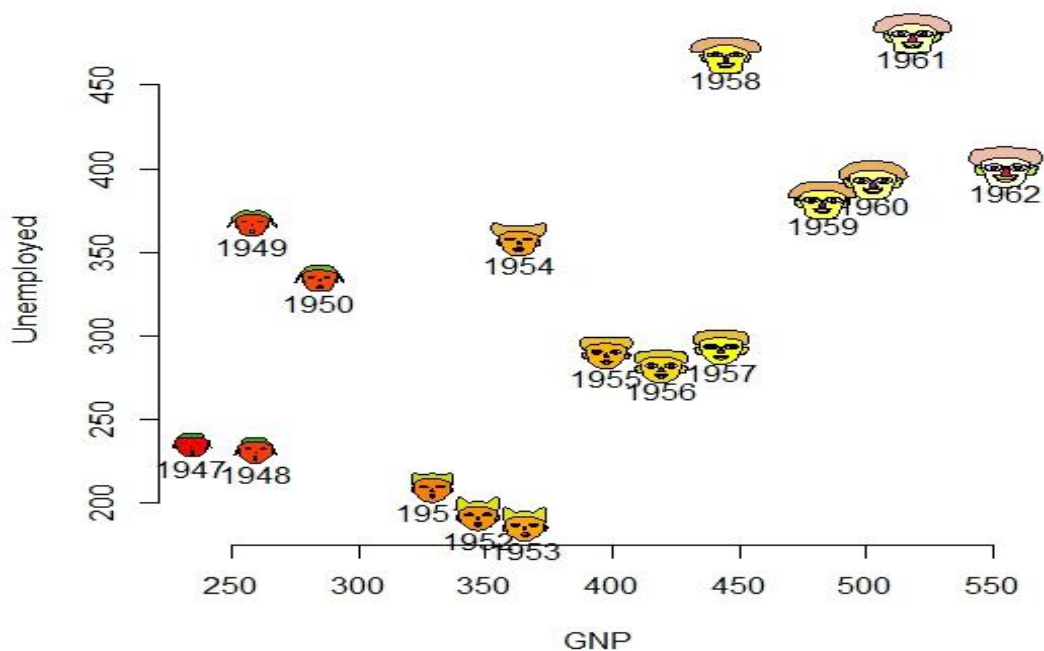
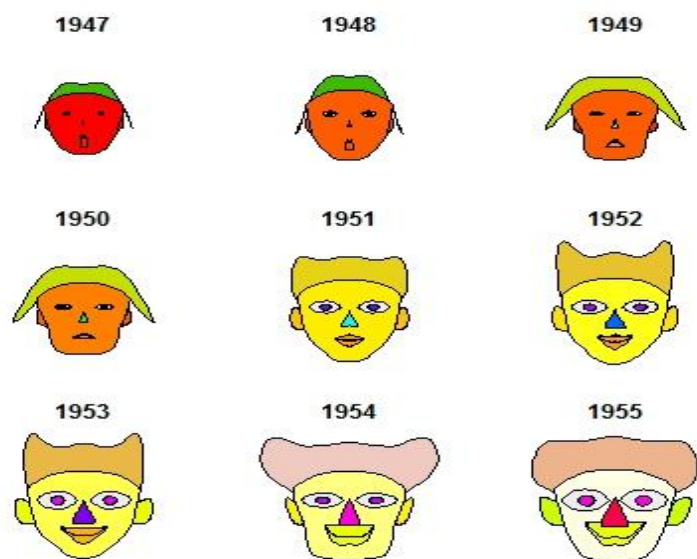
El paquete **corrplot** es un dispositivo gráfico de una matriz de correlación e intervalos confidenciales.



La sentencia **SplomT** crea una matriz de diagramas de dispersión con **a)** covarianzas (con tamaño de la escritura proporcional al tamaño) en el triángulo superior, **b)** histogramas (con suavización) y los nombres de las variables en la diagonal, y **c)** diagramas de dispersión con suavizadores en la dirección y y x en el triángulo inferior, resaltando altas correlaciones por líneas casi paralelas.



Las **caras de Chernoff** Gráfica que muestra datos multivariados en forma de caras humanas. Los rasgos característicos individuales como los ojos, la boca, la nariz representan valores de las variables a través de su forma, tamaño, lugar en la cara donde se encuentran ubicados y orientación. La idea que sustenta este tipo de representación gráfica es la facilidad con que el observador reconocerá cualquier variación por imperceptible que parezca. Ya que los rasgos de la cara cambian muy considerablemente, se debe escoger cuidadosamente la forma en que se grafican los rasgos.



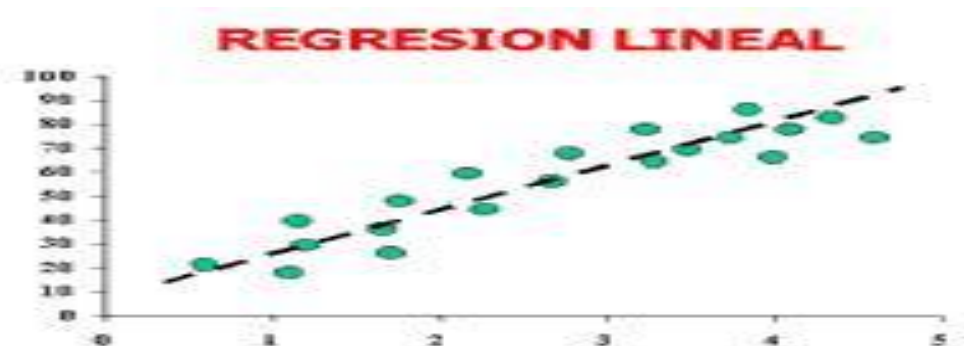
Imputación de datos

- Técnicas de imputación En esta investigación se describen de forma sucinta algunas de las técnicas de imputación existentes que luego serán utilizadas para comparar los resultados por medio de simulaciones.



REGRESIÓN

Consiste en asignar, a los campos a imputar, valores en función del modelo $y_k = \alpha + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k + e$ donde y_k es la variable dependiente a imputar, x_j con $j = 1, 2, \dots, n$ variables independientes.



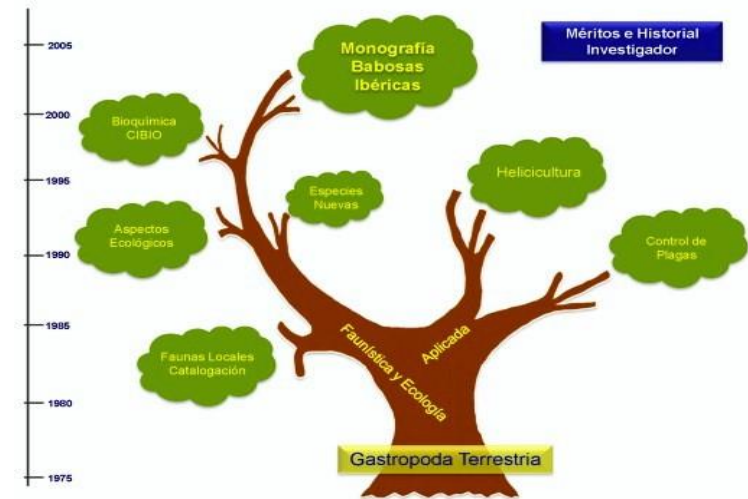
IMPUTACIÓN POR LA MEDIA (MEDIA)

Consiste en un caso particular del método de regresión en el cual se considera solamente una constante en el modelo. De esta manera, se le asigna el valor medio de la variable a todos los valores faltantes de la población

CART (CLASSIFICATION AND REGRESSION TREE)

Se parte de un nodo inicial con n observaciones $H = \{(x_1, y_1), \dots, (x_n, y_n)\}$, donde $x_i \in \mathbb{R}^d$. Se realizan sucesivas particiones binarias en el espacio de las variables X , de forma tal que en cada partición se formen grupos homogéneos con respecto a la variable Y , hasta que un criterio de parada es conformado.

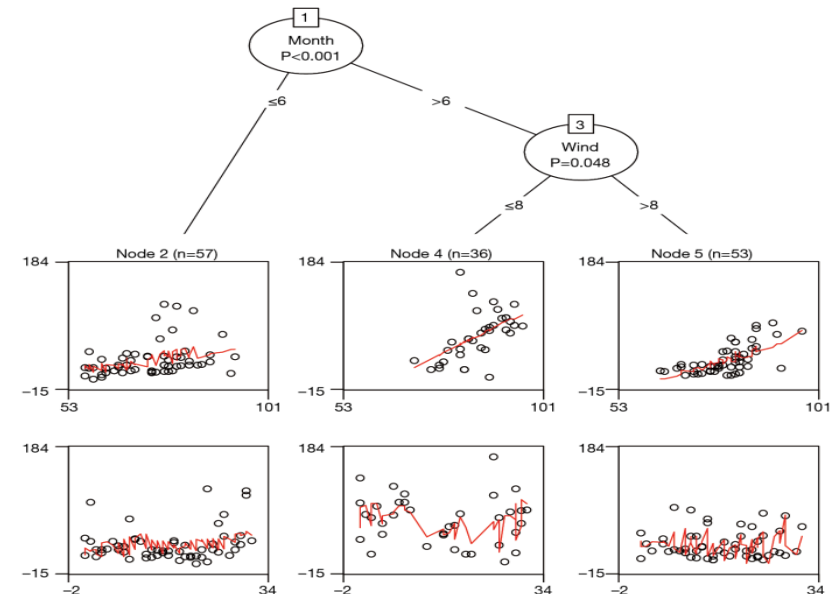
Cuando la variable Y es categórica se denomina árbol de clasificación y se asigna la moda de las categorías como predicción; mientras que si es continua se denomina árbol de regresión y se asigna la media como predicción



MOB (Model-based Recursive Partitioning)

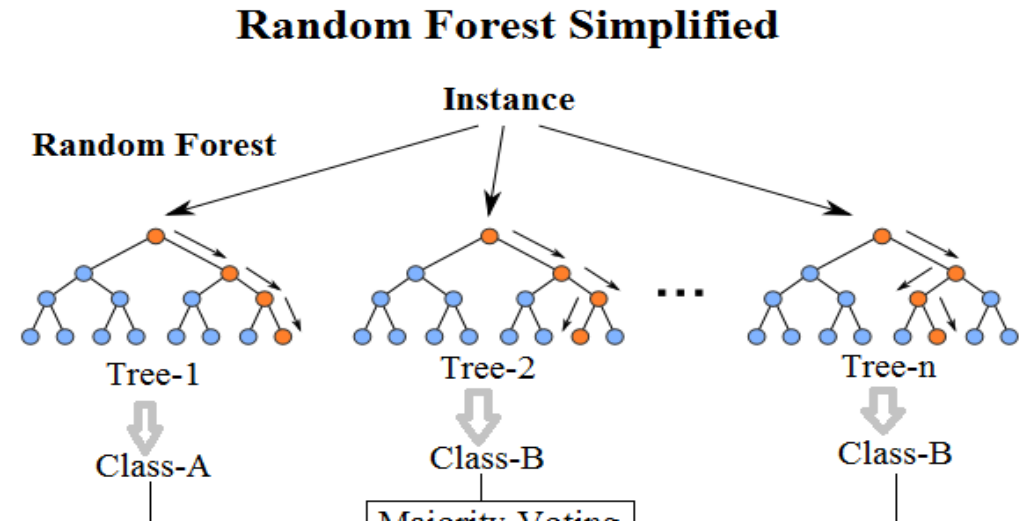
Método similar al CART en el sentido de que también se genera un árbol. Las sucesivas particiones en este caso se realizan ajustando un modelo paramétrico al conjunto de datos, luego se busca dentro del set de variables X cuál proporciona la mejor partición.

La mejor partición será aquella que genere la mayor «inestabilidad» en las estimaciones de los parámetros del modelo dentro de cada hoja y además reduzca la suma de los residuos al cuadrado. El proceso se repite hasta que se cumpla algún criterio de parada.



RANDOM FOREST (RF)

Se generan N árboles de clasificación en forma aleatoria. La aleatoriedad se logra construyendo cada árbol a partir de una muestra bootstrap y eligiendo al azar $q < q_0$ variables

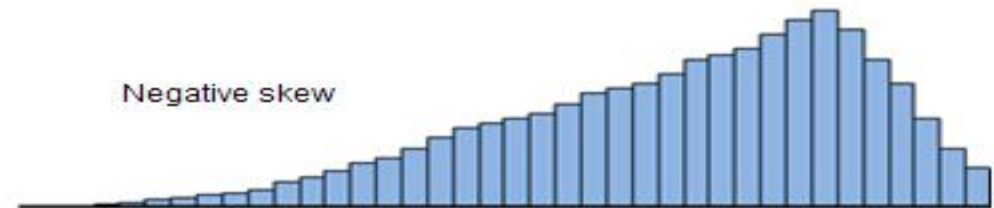
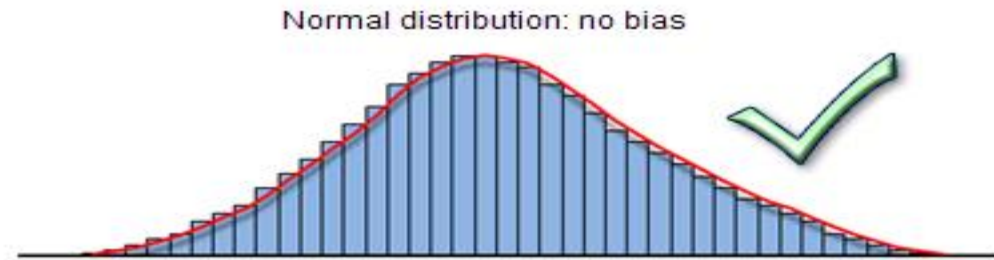
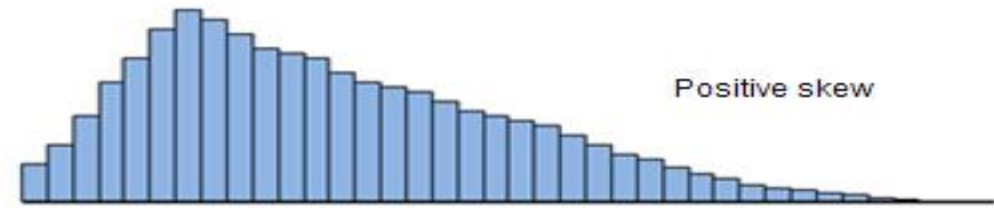


IMPUTACIÓN MÚLTIPLE (IM)

Se imputan m valores para cada una de las celdas vacías en la matriz de datos generando m conjuntos de datos «completos». Los datos faltantes son imputados mediante simulación. Para cada uno de los m conjuntos de datos completos se realiza el análisis que se desee, para luego combinarlos en un sólo resultado

TRANSFORMACIONES DE VARIABLES

En el modelado de datos, la transformación se refiere al reemplazo de una variable por una función. Por ejemplo, reemplazar una variable x por la raíz cuadrada / cubo o logaritmo x es una transformación. En otras palabras, la transformación es un proceso que cambia la distribución o relación de una variable con otras.



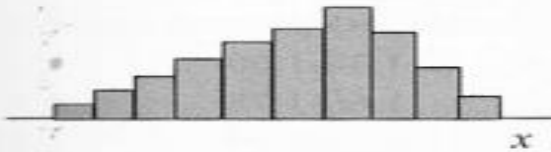
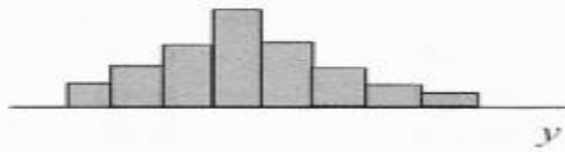
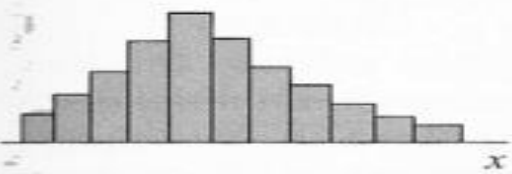
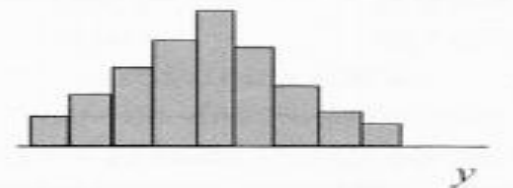
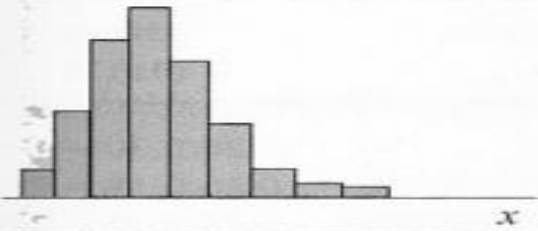
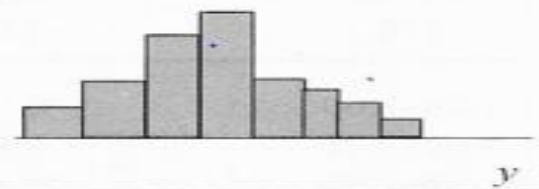
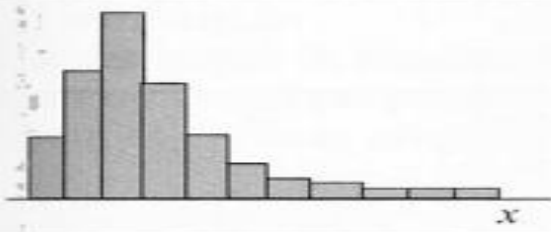
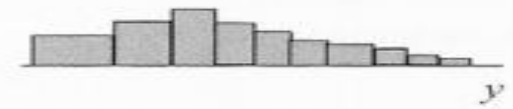
TIPIFICACIÓN DE VARIABLES

La tipificación de variables resulta muy útil para eliminar su dependencia respecto a las unidades de medida empleadas. En realidad, una tipificación equivale a una transformación lineal

$$Z = \frac{X - \bar{x}}{\sigma} = \frac{1}{\sigma}X - \frac{\bar{x}}{\sigma}$$

TRANSFORMACIONES NO LINEALES MÁS FRECUENTES

Cuando se tienen distribuciones de frecuencias con asimetría negativa (frecuencias altas hacia el lado derecho de la distribución), es conveniente aplicar la transformación $y = x^2$. Esta transformación comprime la escala para valores pequeños y la expande para valores altos. Para distribuciones asimétricas positivas se usan las transformaciones \sqrt{x} , $\ln(x)$ y $1/x$, que comprimen los valores altos y expanden los pequeños. El efecto de estas transformaciones está en orden creciente: menos efecto \sqrt{x} , más $\ln(x)$ y más aún $1/x$.

Histograma inicial	Transformación	Histograma transformado
	$y = x^2$	
	$y = \sqrt{n}$	
	$y = \ln x$	
	$y = \frac{1}{x}$	

CREACIÓN DE VARIABLES

Es un proceso para generar nuevas variables / características basadas en variables existentes.

- ✓ Creación de variables derivadas: Se refiere a la creación de nuevas variables a partir de variables existentes utilizando un conjunto de funciones o métodos diferentes. Métodos tales como tomar registro de variables, binning variables y otros métodos de transformación de la variable también se puede utilizar para crear nuevas variables.
- ✓ Creación de variables ficticias: Una de las aplicaciones más comunes de la variable dummy es convertir variables categóricas en variables numéricas. Las variables ficticias también se denominan variables de indicador. Es útil tomar variable categórica como predictor en modelos estadísticos. La variable categórica puede tomar los valores 0 y 1.

