ANALISIS EXPLORATORIO

Elizabeth León Guzmán

Research Group on Data Mining – MIDAS
Universidad Nacional de Colombia, Bogotá D.C., Colombia

2025

Definición AED

Definición

Es una aproximación para el análisis de datos que emplea una variedad de técnicas en su mayoría **gráficas** para:

- Maximizar el entendimiento de un conjunto de datos
- Descubrir estructuras subyacentes
- Extraer variables importantes
- Detectar valores atípicos y anomalías
- Probar suposiciones subyacentes

AED

Se puede realizar antes del procesamiento de datos para "entender comportamientos" y "naturaleza" de los datos, permite:

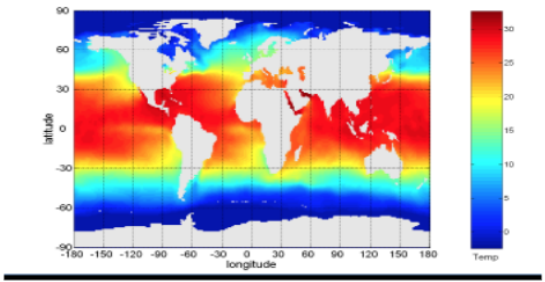
- Ayudar a seleccionar la herramienta adecuada para el preprocesamiento o Análisis
- Hacer uso de las habilidades humanas para el reconocimiento de patrones
- Las personas pueden reconocer patrones no capturados por las herramientas de análisis

Una imagen vale más que mil palabras

AED

Ejemplo: Temperatura de la superficie del mar

La siguiente imagen muestra la temperatura de la superficie del mar para julio de 1982. Decenas de miles de puntos de datos son resumidos en una figura simple que permite entender y analizar la temperatura.



Source: Bravo (2013)

AED

Enfoque de AED

- Énfasis en la técnicas gráficas para ganar entendimiento, lo que se opone a la aproximación clásica de las pruebas cuantitativas.
- La mayoría de los analistas de datos usan una mezcla de técnicas gráficas y clásicas cuantitativas para encaminar sus problemas.

AED

Objetivo de AED

Maximizar el entendimiento del analista de un conjunto de datos, capturando la estructura subyacente del conjunto de datos teniendo en cuenta los ítems que un analista quiere extraer del conjunto de datos como:

- Lista de valores atípicos
- Interpretación robusta de conclusiones
- Estimados para parámetros
- Incertidumbres para esos estimados
- Lista ordenada de los factores importantes

Historia del AED

El trabajo original en AED es:

Exploratory Data Analysis, Tukey, (1977).

Trabajos destacados:

- Data Analysis and Regression, Mosteller and Tukey (1977),
- Interactive Data Analysis, Hoaglin (1977),
- El ABC's de EDA, Velleman y Hoaglin (1981).

Técnicas de AED

- Graficar datos crudos como: histogramas, diagramas de dispersión.
- Graficar estadísticas simples como: gráficas de promedio, gráficas de desviación estándar, gráficas de cajas, etc.
- Organización de tales diagramas para maximizar la habilidad de reconocimiento natural de patrones, como puede ser la utilización de múltiples gráficas por página.

Estadísticas Descriptivas

Son números que resumen las propiedades de los conjuntos de datos: la Ubicación/Localización y la Dispersión. Existen medidas para calcular estas estadísticas

- **Ubicación o Localización:** Representan un *centro* en torno al cual se encuentra ubicado el conjunto de los datos. Las medidas de *centralidad* resumen en un solo valor un conjunto de valores de una variable. Las medidas de tendencia central más utilizadas son: media, mediana y moda.
- **Dispersión:** miden el grado de dispersión de los valores de la variable. Las medidas de dispersión pretenden evaluar en qué medida los datos difieren entre sí. La más utilizada es la desviación estándar.

La mayoría de las estadísticas descriptivas pueden ser calculadas en un solo paso a través de los datos.

Medidas de Centralidad - Media

Media

- Resume los valores observados en un único valor asociado al valor localizado en el centro.
- La **media** es la medida mas común de tendencia central para una variable numérica.
- Es un **promedio** de los datos. Si tenemos m observaciones se calcula como la media aritmética o promedio

$$mean(x) = \overline{x} = \frac{1}{m} \sum_{i=1}^{m} x_i$$

Problema: sensible a outliers o valores atípicos.

Medidas de Centralidad - Mediana

Mediana

- La mediana representa de posicion central de la variable que separa la mitad inferior y la mitad superior de las observaciones.
- El valor donde para una mitad todos los valores son mayores que este, y para la otra mitad todos son menores.
- La mediana es más robusta que la media

$$median(x) = \begin{cases} x_{r+1} & si\ m\ es\ impar \\ \frac{1}{2}(x_r + x_{r+1})\ si\ m\ es\ par \end{cases}$$

r es el valor entero de la mitad de la longitud (m) del vector x

$$r = entero(\frac{m}{2})$$

Medidas de Centralidad - Moda

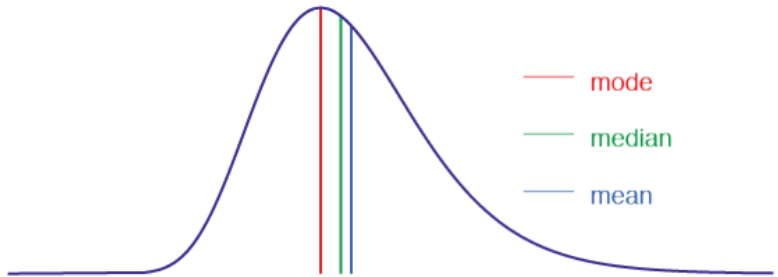
Frecuencia

La frecuencia es el número de veces que se repite un dato.

Moda

La moda es el dato que tiene mayor frecuencia.

Medidas de Centralidad

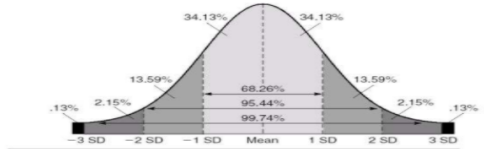


Source: Bravo (2013)

Percentiles

Los percentiles son valores de la variable que dividen la distribución en 100 partes iguales. El k-esimo percentil de una variable numérica es un valor tal que el k% de las observaciones se encuentran debajo del percentil y el (100 − k) % se encuentran sobre este valor.

Ejemplo: Si se tienen los datos de la varibale edad, y si el percentil 80 (P80) es igual a 35 años de edad, significa que el 80% de los casos tiene edad igual o inferior a 35 años.



z-scores	−3	−2	−1	0	1	2	3
Percentile ranks	0.1	2	16	50	84	98	99.9
IQ	55	70	85	100	115	130	145
SAT	200	300	400	500	600	700	800

FIGURE 15.8 Percentile ranks and standard scores in relation to the normal curve. SD = standard deviation

Percentiles

- En estadística se usan los **cuartiles** que son equivalentes a los percentiles expresados en fracciones en vez de porcentajes.
- Los **cuartiles** son tres percentiles específicos:
  - El primer cuartil Q1 (lower quartile) es el percentil con k = 25.
  - El segundo cuartil Q2 es con k = 50 equivale a la mediana.
  - El tercer cuartil Q3 (upper quartile) es con k = 75.

Ejercicio:

Usar el comando `tapply`. Analizar la media, la mediana y los cuartiles para las tres especies de Iris para las cuatro variables. ¿Hay diferencias en las distintas especies?

```
tapply(iris$Petal.Length,iris$Species,summary)
tapply(iris$Petal.Width,iris$Species,summary)
tapply(iris$Sepal.Length,iris$Species,summary)
tapply(iris$Sepal.Width,iris$Species,summary)
```

Medidas de Dispersión

Rango

Diferencia entre el valor máximo y mínimo

```
max(sepal.length)-min(sepal.length)
```

Desviación estándar

Es la raíz cuadrada de la **varianza** que mide las diferencias cuadráticas promedio de todas las observaciones con respecto a la **media**.

$$var(x) = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \overline{x})^2$$
$$sd(x) = \sqrt{var(x)}$$

```
var(sepal.length)
sd(sepal.length)
```

Medidas de Dispersión

Desviación Media Absoluta - AAD

Es una medida de dispersión más robusta a `outliers` que la Desviación Estándar.

$$AAD(x) = \frac{1}{m} \sum_{i=1}^{m} |(x_i - m(x)|$$

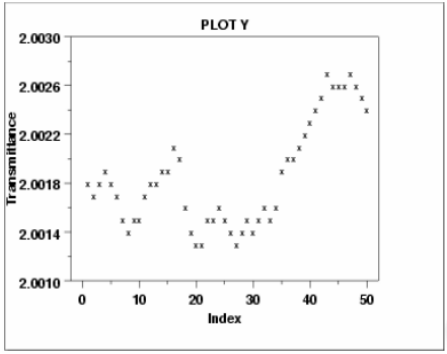
$m(x)$ es una medida de tendencia central de x (puede ser la mediana).

Gráfica de Ejecución de Secuencia

Definición: $Y(i)$ Vs. i

La gráfica se constituye por:

- Eje vertical: variable de respuesta $Y(i)$
- Eje horizontal: índice i ($i = 1, 2, 3, \ldots$)



Gráfica de Ejecución de Secuencia

La Gráfica de Ejecución de Secuencia puede ser usada para responder:

- ¿Hay cambios de localización?
- ¿Hay cambios en la variación?
- ¿Hay cambios en la escala?
- ¿Hay cambios en los valores atípicos?

Estadísicas Multivariadas

Comparar como varía una variable con respecto a otra.

Covarianza

Mide el grado de variación lineal conjunta de variables.

$$cov(x, y) = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \overline{x})(y_i - \overline{y})$$

- Donde $cov(x, x) = var(x)$
- Si las variables son independientes la covarianza es 0
- El signo de la covarianza indica la tendencia en la relación lineal entre las variables.

En R se calcula: `cov(sepal.length,sepal.width)`

El parámetro de cov puede ser una matriz o un data.frame, lo que calcula la **matriz de covarianzas:** `cov(iris[,1:4])`

Estadísicas Multivariadas

Correlacion lineal o Coefiente de Correlación de Pearson

Es una medida de relación que no depende de la escala de cada variable. Se define como r(x,y):

$$r(x, y) = \frac{cov(x,y)}{sd(x)sd(y)}$$

- Varía entre -1 a 1. Un valor cercano a 1 indica que mientras una variable crece la otra tambien lo hace en una proporción lineal.
- Un valor cercano a -1 indica una relacion inversa (una crece la otra decrece).
- Si la correlación es cercana a 0 hay independencia lineal. Sin embargo, no implica que no pueda haber una relacion no-lineal entre las variables.

En R se calcula: `cor(iris[,1:4])`

Estadísicas Multivariadas

Tablas de Contingencia

Se usan para analizar variables **categóricas**
La tabla contiene las frecuencias marginales de todos los pares de valores entre dos variables categóricas.

En R se calcula con `table`

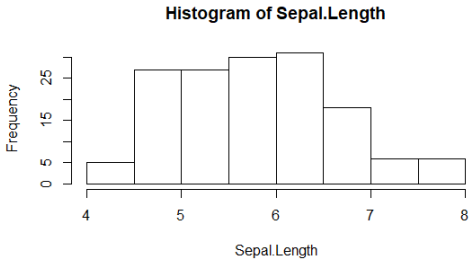
Ejemplo:

```
> origen<-c("Bogota","Cali","Bogota", "Cartagena","Bogota", "Cali")
> colegio<-c("Privado","Privado","Publico","Privado","Publico","Publico")
> table(origen,colegio)
           colegio
origen      Privado Publico
  Bogota          1       2
  Cali            1       1
  Cartagena       1       0
```

Histogramas

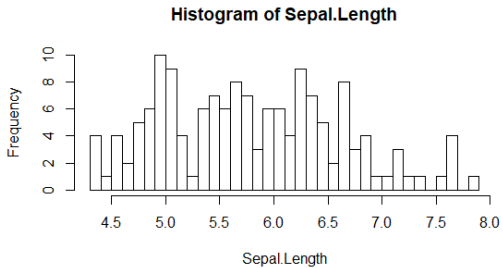
- Gráfica de la distribucion de los valores de una variable.
- Los valores se dividen en bins y se crea una gráfica de barra por cada bin.
- La altura de cada barra indica el numero de elementos o frecuencia del bin.

En R se crea con el comando `hist:` `hist(sepal.length)`



Histogramas

- En R se puede definir el número de bins con `nclass`
  `hist(sepal.length, nclass=50)`

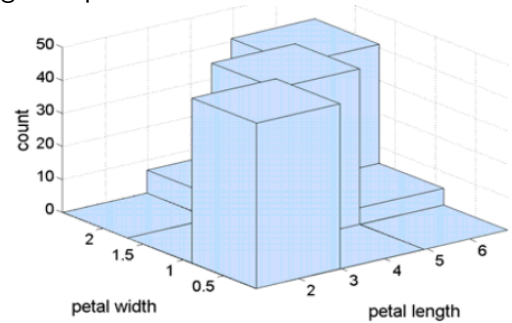


Histogramas

Histograma de dos dimensiones

Muestra la unión de distribución de valores de dos dimensiones o variables.

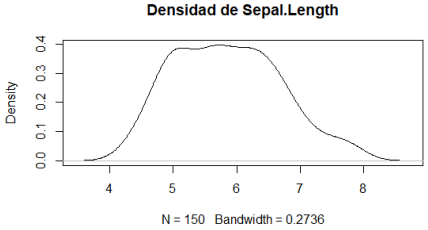
Ejemplo: Anchura del pétalo y largo del pétalo



¿Qué nos está diciendo?

Estimación de Densidad

- Otra forma de visualizar los datos.
- Se usan técnicas estadísticas no paramétricas: estimación de densidad de **kernel**.
- Versión suavizada del histograma, permite determinar si los datos observados se comportan como una densidad conocida.
- En R se crea con el comando `density`, y se visualiza con el comando `plot`.
  `plot(density(iris$Sepal.Length),main="Densidad de Sepal.Length")`



Histogramas

El histograma puede ser usado para responder las siguientes preguntas:

- ¿De qué tipo es la distribución de la población de donde vienen los datos?
- ¿Dónde están ubicados los datos?
- ¿Son los datos simétricos o asimétricos?
- ¿Hay valores atípicos?

Diagramas de Caja

- Otra forma de mostrar la distribución de los datos.
- Se construyen a partir de los percentiles

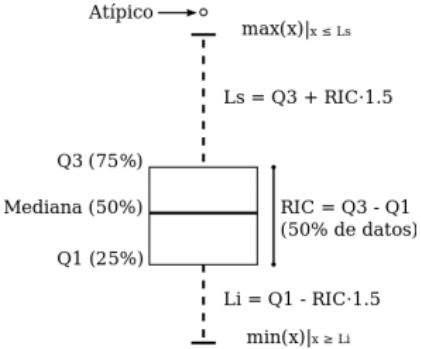
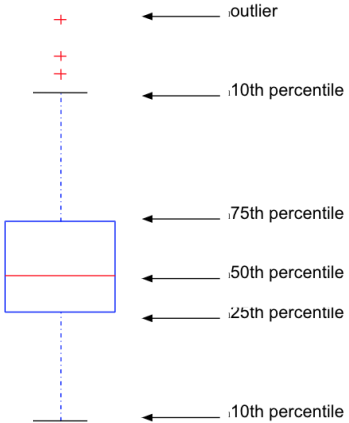


Figura: Fuente:

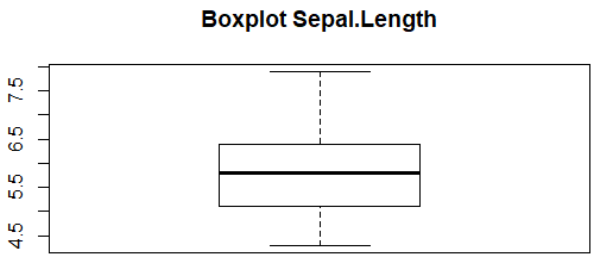
Diagramas de Caja

- Inventadas por J. Tukey
- La siguiente figura muestra la parte básica de una gráfica de caja:



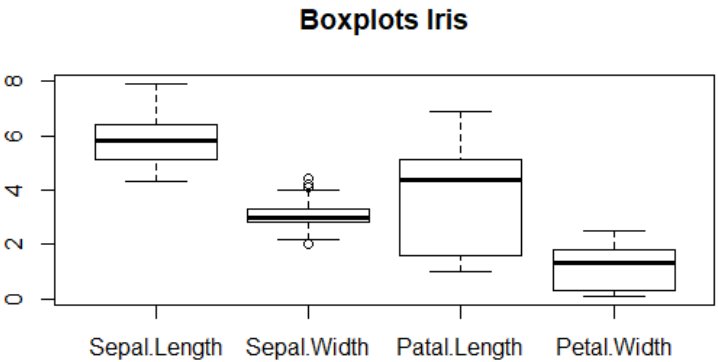
Diagramas de Caja

- En R se grafican con el comando `boxplot`:
  `boxplot(Sepal.Length,main="Boxplot Sepal.Length")`



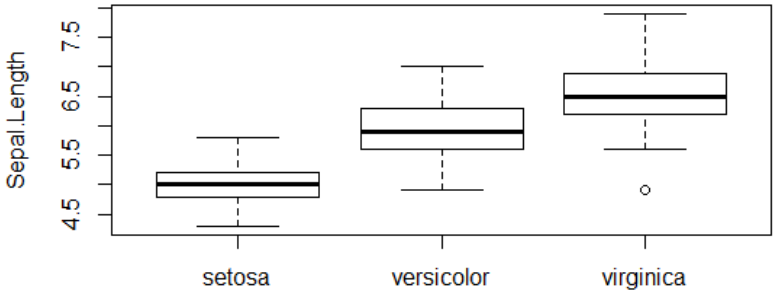
Diagramas de Caja

- En R se pueden analizar varias variables al mismo tiempo:
  `boxplot(x=iris[,1:4],main="Boxplots Iris")`



Diagramas de Caja

- En R se puede crear un boxplot por cada "clase" o categoría:
  `boxplot(sepal.length~species,ylab="Sepal.Length")`



Diagramas de Dispersión

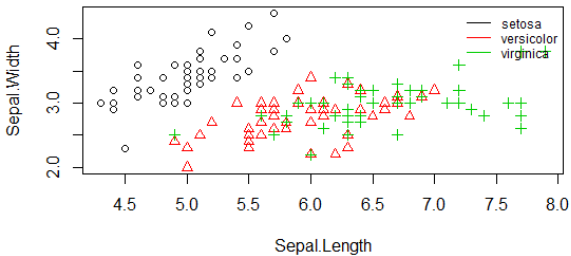
- Los diagramas de dispersion o **scatter plots** usan coordenadas cartesianas para mostrar los valores de dos variables del mismo largo.
- Los valores de los atributos determinan la posicion de los elementos. Se pueden usar atributos para definir tamano, forma o color de los objetos.
- En R se grafica un scatterplot de dos variables numéricas usando el comando `plot(x,y)`.
- scatterplots de data.frame o matriz numerica usando el comando `pairs(x)`.

Diagramas de Dispersión

```
plot(Sepal.Width Sepal.Length, col=Species)
plot(Sepal.Length, Sepal.Width,col=Species, pch=as.numeric(Species))
```

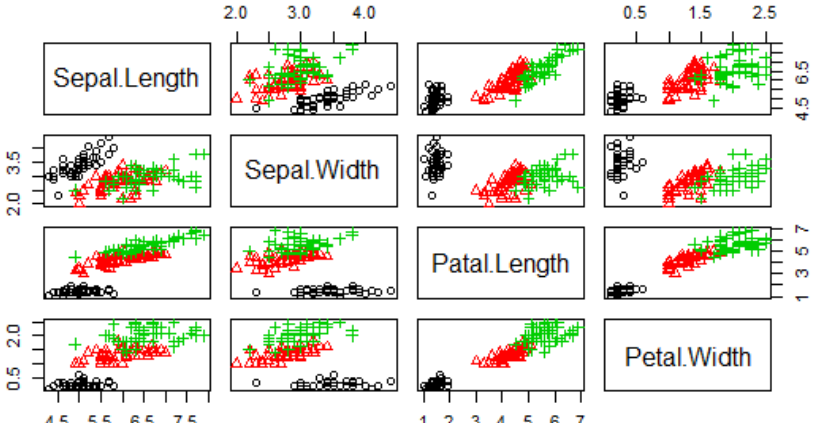
Agrega una leyenda

```
legend('topright', levels(Species),lty=1, col=1:3, bty='n', cex=.75)
```



Diagramas de Dispersión

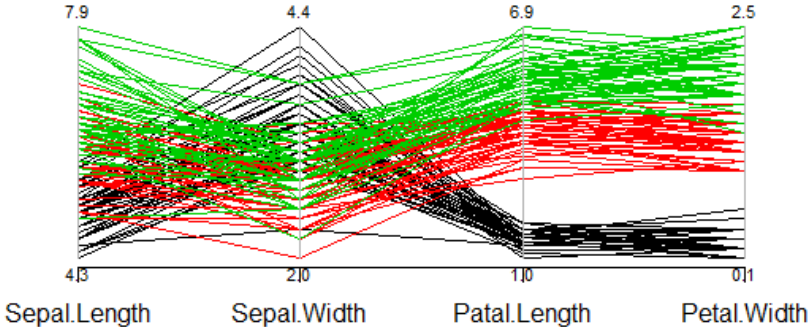
```
pairs(iris[,1:4],pch=as.numeric(iris$Species),col=iris$Species)
```



Coordenadas Paralelas

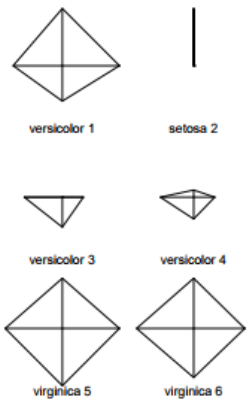
En 3D: `library(MASS)`

```
parcoord(iris[1:4], col=iris$Species,var.label=T)
```



Diagramas de Estrella

```
iris_sample1<-iris[sample(1:dim(iris)[1],size=6,replace=F),]
rownames(iris_sample1)<-paste(as.character(iris_sample1$Species),1:6)
stars(iris_sample1[1:4])
```



Star plots

Caras de Chernoff

```
library("aplpack")
iris_sample<-iris[sample(1:dim(iris)[1],size=16,replace=F),]
faces(iris_sample[1:4],face.type=1,labels=iris_sample$Species)
```

Chernoff Faces

References I

Bravo, F. (2013). *Análisis exploratorio en r*.