



Módulo 11

Análisis exploratorio de datos



AGENDA DE LA CLASE

- ✓ La importancia del EDA
- ✓ Estimación de localización
- ✓ Estimación de variabilidad
- ✓ Exploración de la distribución de datos
- ✓ Exploración de datos binarios y categóricos
- ✓ Correlación
- ✓ Probabilidad

01 - La importancia del EDA

El **Análisis Exploratorio de Datos**, o EDA (Exploratory Data Analysis) es el **primer paso** y el más importante en cualquier proyecto basado en datos.



Permite observar, resumir y visualizar los datos, para desarrollar la intuición y **comprensión de los datos** de nuestro proyecto.

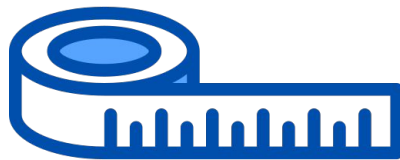
Es un área relativamente nueva de la estadística

1962, John W. Tukey hizo un llamamiento a la reforma de la estadística en su artículo *"El futuro del análisis de datos"*. Propuso una nueva disciplina científica denominada **"Análisis de datos"**

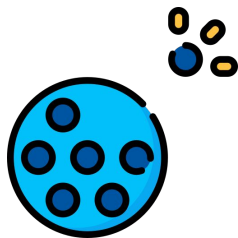
La **estadística clásica** se centraba casi exclusivamente en la inferencia, es decir, extraer conclusiones sobre grandes poblaciones a partir de muestras pequeñas

02 - Estimación de localización

Las variables con **datos medidos** pueden tener miles de valores distintos.



Un paso básico en la exploración de sus datos es obtener un "**valor típico**" para **cada característica** (variable), es decir, una **estimación** de dónde se encuentra la mayoría de los datos (esto es, su **tendencia central**)



Un **valor atípico** (o **valor extremo** u **outlier**) es un valor de un dato que es muy diferente de la mayoría de los valores de datos.

Pueden ser resultado de errores de datos

Combinación de diferentes unidades

Lecturas incorrectas de un sensor

Deben identificarse y merecen una investigación más profunda



02 - Estimación de localización

MEDIA (valor medio o promedio): es la suma de todos los valores (todos los x) dividida por el número de valores (n).

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$



=PROMEDIO

MEDIANA: es el valor central de una lista de datos ordenados de menor a mayor.



=MEDIANA



Si hay un **número par de valores de datos**, el valor medio es uno que no está realmente en el conjunto de datos, sino que es el **promedio** de los dos valores que dividen los datos ordenados en mitades superior e inferior

03 - Estimación de variabilidad

La **variabilidad**, también conocida como **DISPERSIÓN** que mide el grado de agrupación o dispersión de los valores de los datos en torno al valor central.

El **promedio de las desviaciones** respecto al valor típico (medio) no sirve de mucho porque la **suma de los desvíos es cero**.

$$x_i - \bar{x}$$

La **desviación media absoluta** es el promedio de los valores absolutos de las desviaciones de la media.

$$\text{Desviación media absoluta} = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

De este modo vemos que la sumatoria de estas desviaciones ya no nos da cero.

La **VARIANZA** es un promedio del cuadrado de las desviaciones.

$$\text{Varianza} = s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

La **DESVIACIÓN ESTÁNDAR** es la raíz cuadrada de la varianza

$$\text{Desviación estandar} = s = \sqrt{\text{Varianza}}$$



=VAR.S



=DESVEST.M

03 - Estimación de variabilidad

Un enfoque diferente para estimar la dispersión se centra en observar la **distribución de los datos ordenados**.

Los estadísticos que tienen como base los datos **ordenados** (clasificados) se denominan **estadísticos de orden**.

El **RANGO** que es la diferencia entre los números mayor y menor valor

 =MAX; =MIN

Útil para identificar valores atípicos pero extremadamente sensible a ellos.
No es muy útil como medida general de la dispersión de los datos.

El **PERCENTIL**, P, es un valor tal que al menos el P por ciento de los valores toman este valor o un valor inferior y al menos (100-P) por cierto de los valores toman este valor o un valor superior.

 =PERCENTIL.INC

Por ejemplo, tomando un conjunto de datos ordenados, si queremos encontrar el percentil 80, comenzando por el valor más pequeño continuamos hasta el 80% del recorrido.

03 - Estimación de variabilidad

El percentil es esencialmente lo mismo que el **CUANTIL**, con los cuantiles referenciados por porcentajes.



=PERCENTIL.INC

Por ejemplo, el cuantil 0.8 es lo mismo que el percentil 80.

Una medida habitual de la variabilidad es la **diferencia entre el percentil 25 y percentil 75** al que se llama **RANGO INTERCUARTÍLICO** (IQR).

El **CUARTIL**, que es el percentil cada 25% (serían los cuartos).



=CUARTIL

04 - Exploración de la distribución de datos

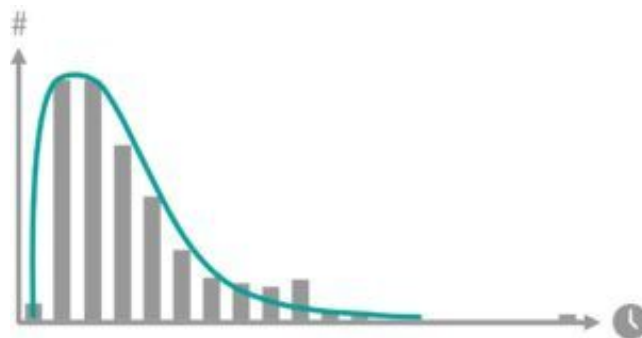


Hasta ahora vimos las estimaciones que resumen los datos en **una sola cifra** para **describir la localización** o la **variabilidad de los datos**.

También es interesante explorar **cómo se distribuyen los datos** en general.

Las **colas**, que son las partes extremas del rango de la distribución.

Los **percentiles** son especialmente indicados para extraer un **resumen de las colas**.

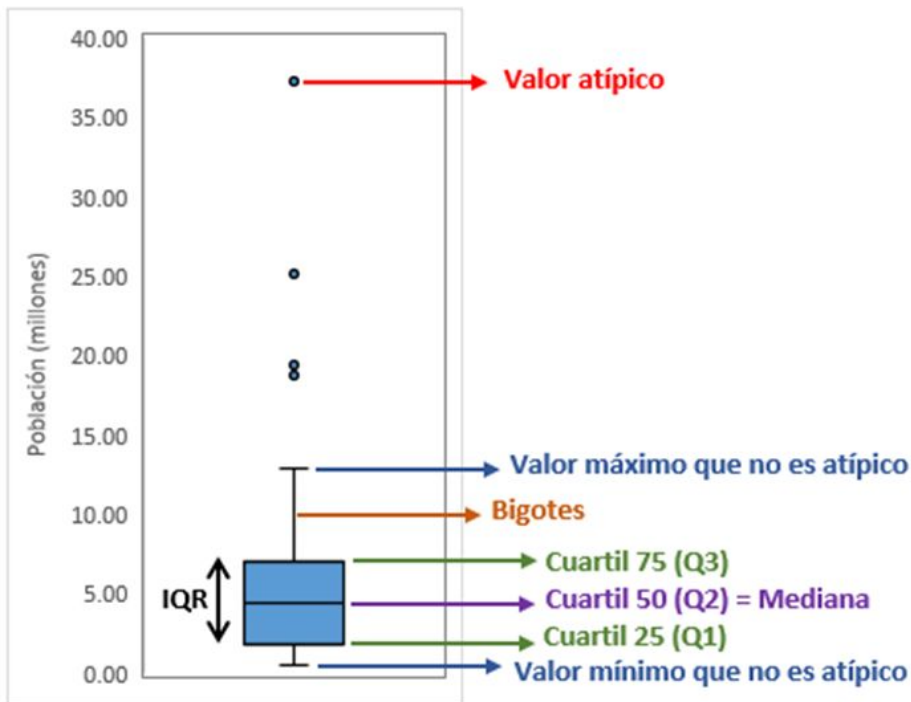


04 - Exploración de la distribución de datos

Los **diagramas de caja** (boxplots o diagrama de caja y bigotes) utiliza percentiles y permite visualizar la distribución de datos de una forma rápida.

En general se considera que un valor es **atípico** cuando el valor se encuentra:

- por debajo de 1.5 veces $Q1$, o
- por encima de 1.5 veces $Q3$



04 - Exploración de la distribución de datos

La **tabla de frecuencias** de una variable divide el rango de la variable en segmentos igualmente espaciados y nos dice **cuántos valores caen dentro de cada segmento**.



=FRECUENCIAS

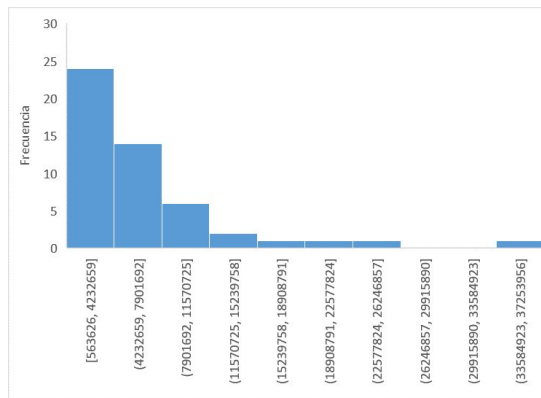
Es una **Función matricial**

Tenemos que seleccionar la matriz donde queremos que se calculen todos los resultados.

Luego de cerrar el paréntesis, debemos hacer Ctrl + Shift + Enter

En la barra de fórmulas veremos que nuestra función está encerrada entre { }

El **histograma** es un modo de visualizar la tabla de frecuencias, con contenedores en el eje x y los valores de los datos en el eje y.



05 - Exploración de datos binarios y categóricos

En el caso de los datos categóricos o una variable binaria, las proporciones simples o porcentajes **cuentan la historia de los datos**.

Los **gráficos de barras** son una de las herramientas visuales más utilizadas para mostrar **una única variable categórica**.

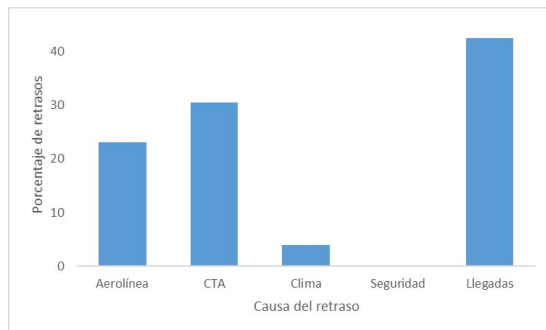


Gráfico de barras *vs* Histograma

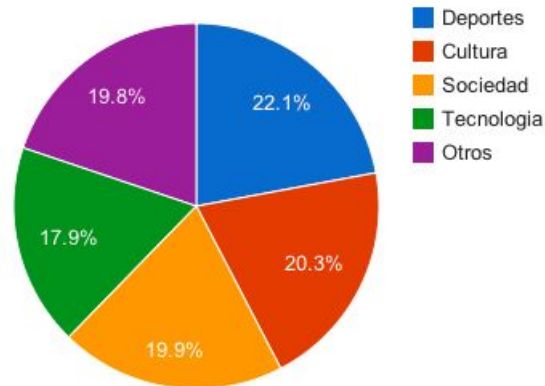
- El **eje x** representa diferentes **categorías** de una variable
- Las **barras** se muestran **separadas entre sí**
- El **eje x** representa los **valores** de una sola **variable** en una escala numérica
- Las **barras** están juntas entre sí. Si hay **espacios**, estos indican **valores ausentes** en los datos

05 - Exploración de datos binarios y categóricos

Los **gráficos de torta** son una alternativa a los gráficos de barras.

Aunque los estadísticos y los expertos en visualización de datos generalmente los evitan por ser menos informativos visualmente.

Visitas a contenidos



La **MODA** es el **valor** (o valores en caso de empate) que aparece con **mayor frecuencia en los datos**.

 =MODA

Es un **resumen estadístico simple para datos categóricos o binarios** y, por lo general, no se usa para datos numéricos

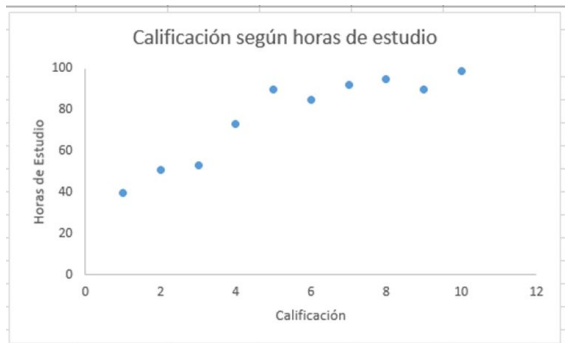
06 - Correlación

La **correlación** entre dos variables es la relación entre dos variables.

El **gráfico de dispersión** (o scatterplots), es el gráfico que se utiliza para mostrar relaciones entre dos variables

Correlación positiva

si los valores altos de x acompañan valores altos de y, y los valores bajos de x acompañan los valores bajos de y.



Correlación negativa

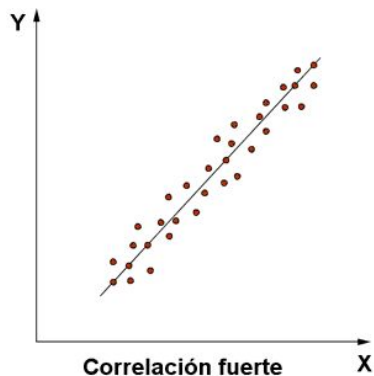
si los valores altos de x acompañan los valores bajos de y, y los valores bajos de x acompañan los valores altos de y.



06 - Correlación

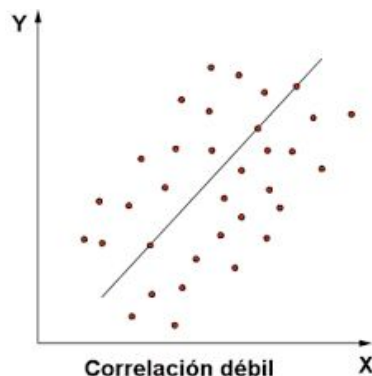
Correlación fuerte

si dibujamos una línea uniendo los puntos nos daría una recta con los valores muy cercanos a ella



Correlación débil

los datos los encontramos más dispersos, pero con una tendencia positiva o negativa



El diagrama de dispersión nos ayuda a determinar, en términos generales, si una correlación es fuerte o débil, positivo o negativo, pero **no nos dice qué tan fuerte es la relación.**

06 - Correlación

Coeficiente de correlación de Pearson o **r de Pearson**, mide la fuerza y la dirección de la relación lineal entre dos variables

Multiplicamos las desviaciones de la media de cada variable y la dividimos por el producto de las desviaciones estándar y por n-1

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)s_x s_y}$$

Este coeficiente siempre se encuentra entre **+1** (correlación positiva perfecta) y **-1** (correlación negativa perfecta). El **0** indica que no hay correlación.



=COEF.DE.CORREL



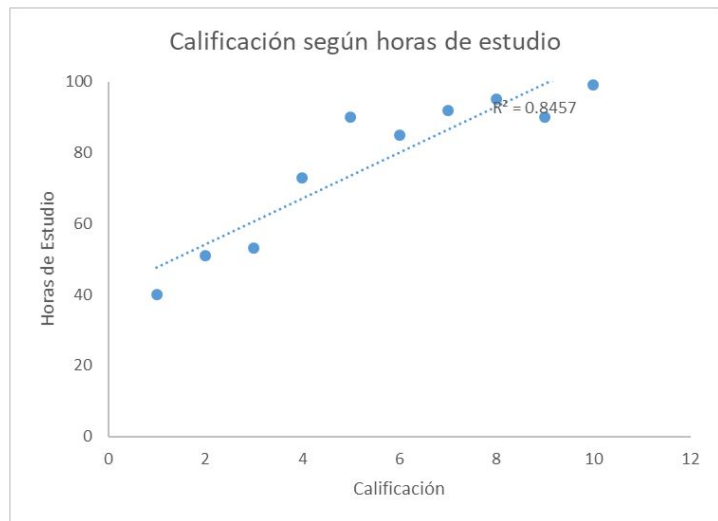
06 - Correlación

La correlación lineal entre dos variables en un gráfico de dispersión se puede mostrar mediante una línea recta llamada **Línea de regresión** o **Línea de tendencia**.

Para saber **cuán bien ajusta esta línea** se calcula el coeficiente r^2 que no es más que el **r de Pearson al cuadrado**.

Este valor es siempre positivo y devuelve valores entre 0 y 1, siendo:

- **0** que no tiene ninguna relación lineal y
- **1** que es una relación lineal perfecta



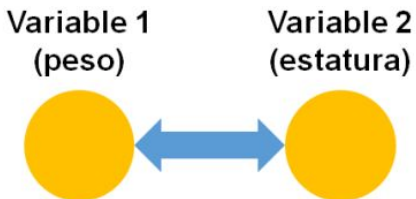
06 - Correlación



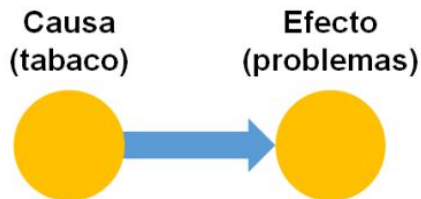
Correlación NO es lo mismo que **causalidad**

Con un **análisis de correlación** NUNCA se demuestra que existe una relación causal entre dos variables

Correlación



Causalidad






07 - Probabilidad

El concepto de probabilidad puede ser motivo de una profunda discusión filosófica a la hora de definirlo. Afortunadamente, para el análisis de datos no necesitamos una definición matemática o filosófica formal. **Para nuestros propósitos:**

La **probabilidad** de que suceda un evento **es la proporción de veces que ocurriría si la situación pudiera repetirse una y otra vez**, innumerables veces.

Ejemplo: Estudio de preferencias de sabores de helado a 100 personas

A		45 personas prefieren A
B		30 personas prefieren B
C		25 personas prefieren C

Concepto de probabilidad:

- Repetir el estudio infinitamente
- Muestras diferentes de 100 personas

Entonces, podemos analizar las **proporciones** en que se eligen las marcas favoritas

La probabilidad de que una persona elija una determinada marca es:

0.45 (45%)

0.30 (30%)

0.25 (25%)

¿PREGUNTAS?





¡Muchas gracias!