Pontificia Universidad
JAVERIANA
Bogotá

Análisis Exploratorio de Datos

Diplomado en Analítica de Datos

Heriberto Felizzola, MSc.

Ing.heriberto.felizzola@gmail.com

¡¡¡De dónde venimos y hacia donde vamos!!!

Pontificia Universidad JAVERIANA Bogotá

Temas previos

Fundamentos de Ciencia de Datos

- Conceptos básicos, metodología y aplicaciones
- Fundamentos de estadística
- Fundamentos de programación en Python analítica de datos

Temas futuros

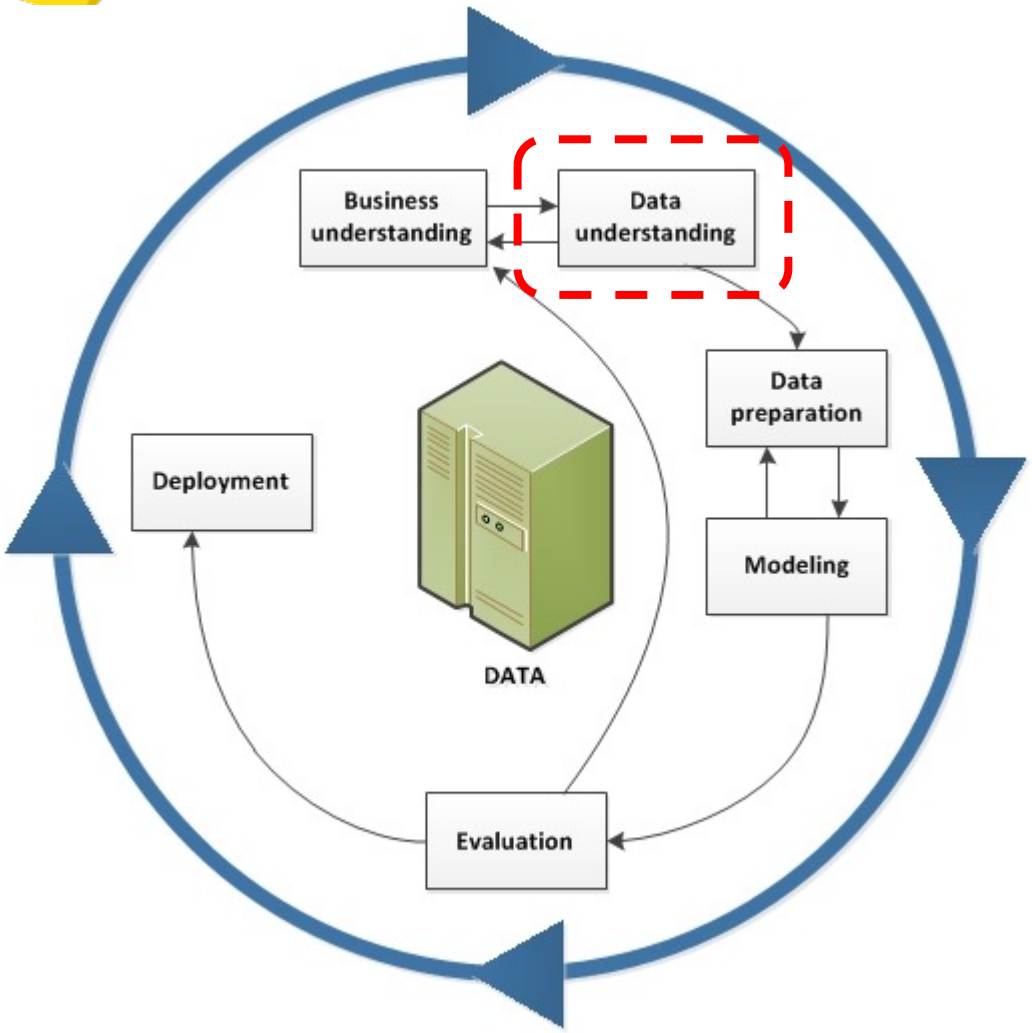
Modelos Predictivos

- Modelos de regresión
- Problema de clasificación: conceptos básicos y métricas
- Regresión logística
- Arboles de decisión
- Ensambles
- Algoritmos basados en vecinos KNN
- Redes neuronales
- Máquinas de soporte

Modelos Descriptivos

- Reglas de asociación
- Reducción de dimensionalidad
- Clustering y anomalías
- Técnicas Inductivas de Aprendizaje de Máquina
- Minería de texto (Introducción)

Cross-Industry Standard Process for Data Mining



Fuente. IBM - Conceptos básicos de ayuda de CRISP-DM.
https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

- Es una metodología estándar que guía el ciclo de vida de un proyecto de minería de datos
- El ***análisis exploratorio de datos (EDA)*** es clave para el ***entendimiento de los datos***
- Note que esto se debe realizar luego de tener un **entendimiento del negocio**: problema, situación, dominio, contexto, objetivos, limitaciones, KPI´s, otros
- El EDA genera insumos claves para la ***preparación de los datos***: limpieza, estandarización, ingeniería de variables, otros.

¿Qué es el EDA?

Definición:

- Es un enfoque de análisis de datos que implica ***explorar y resumir*** nuestros datos de manera visual y estadística.
- Es el proceso de sumergirse en los datos, comprender su estructura y extraer información valiosa."

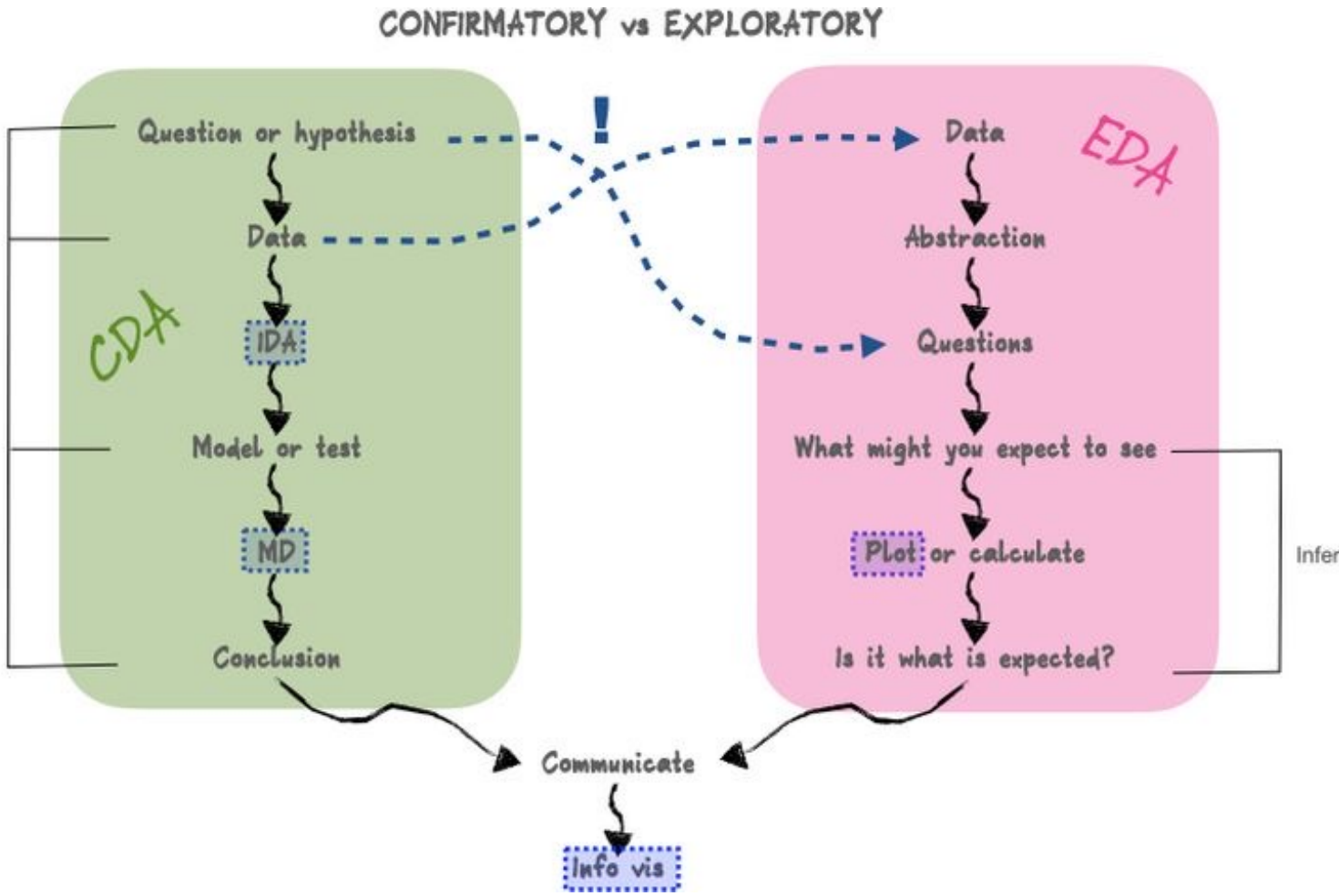
Importancia:

- Es el ***"cimiento de una casa sólida"*** en el análisis de datos.
- Antes de aplicar algoritmos de aprendizaje automático o realizar análisis avanzados, debemos entender nuestros datos en profundidad.
- Permite detectar problemas, identificar tendencias y patrones, y tomar decisiones informadas."

"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

Análisis Confirmatorio vs Exploratorio

- A menudo, se confunde el EDA con el análisis estadístico tradicional.
- La principal diferencia radica en el enfoque.
- Mientras que el *análisis tradicional se centra en probar hipótesis específicas*, el *EDA se enfoca en descubrir información sin prejuicios*.
- Es un proceso más exploratorio y menos orientado a la confirmación

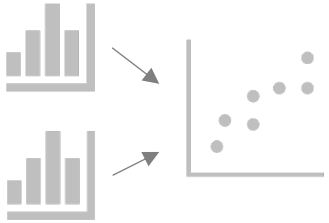


Tareas claves en un EDA

Pontificia Universidad JAVERIANA Bogotá

Datos

VAR1	VAR2	VAR3
2653	F	56.2
5623	M	78.6
1235	M	95.6
1523	F	63.6



Paso 1.
Identificar Tipos de datos

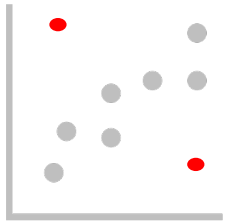
Paso 2.
Análisis univariado

Paso 3.
Análisis bivariado | Multivariado

Paso 4.
Identificar datos Faltantes e Inconsistentes (Errores)

EDAD	GENERO	CARGO
26	M	TEC
23	F	AUX
200	F	NA
34	NA	PROF

Paso 5.
Identificar datos Atípicos



Paso 6.
Ingeniería de variables

VAR1	VAR2	NEW VAR

Ideas Claves

Preguntas

Comunicación

Informe

Tipos de datos: ¿Cómo están estructurados?

Pontificia Universidad JAVERIANA Bogotá

Estructurados

Son archivos que almacenan datos en filas y columnas con títulos. Son datos que pueden ser ordenados y procesados fácilmente por todas las herramientas de minería de datos. *Ej: tablas de Excel.*

Semi-Estructurados

Tienen algunas características consistentes y definidas, no se limitan a una estructura rígida como la necesaria para las bases de datos relacionales. *Ej: código de páginas web.*

No estructurados

No tienen estructura identificada. ***Ej: archivos tipo Word, PDF, e-mails.***

Tipos de variables

Pontificia Universidad JAVERIANA Bogotá

Cuantitativas

- Variables **numéricas** donde tiene sentido realizar operaciones aritméticas tradicionales.
- **Continuas:** Variables cuya escala de valores es continua (peso, altura, distancia).
- **Discretas:** Variables cuya escala solo admite valores enteros, suelen ser resultado de un conteo (hijos, puertas).

Cualitativas

- Sirven para **categorizar** elementos. Suelen ser representadas por **cadenas de texto**. No permiten operaciones aritméticas.
- **Ordinales:** Variables con categorías y un orden implícito en estas categorías (meses, pisos térmicos).
- **Nominales:** Variables con categorías sin un orden para las mismas (departamentos, marcas de celular).

Tipos de Datos:
¿Cómo se representan los atributos o características?

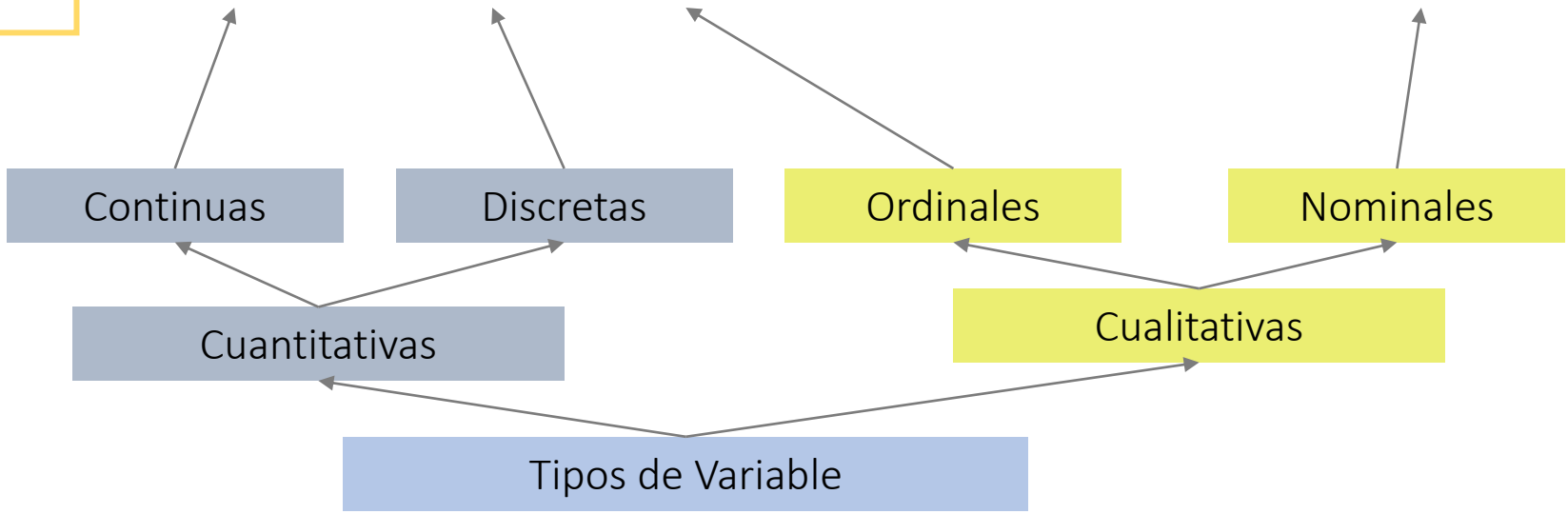
Pontificia Universidad JAVERIANA Bogotá

Las columnas representan las características de cada uno de los préstamos, también denominadas variables.

Cada fila representa un único préstamo. Esto puede ser un caso/instancia/observación.

	loan_amount	interest_rate	term	grade	state	total_income	homeownership
1	7500	7.34	36	A	MD	70000	rent
2	25000	9.43	60	B	OH	254000	mortgage
3	14500	6.08	36	A	MO	80000	mortgage
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	3000	7.96	36	A	CA	34000	rent

Cada celda es un valor único de la variable para una observación particular.



Tipos de Datos: ¿Cómo los representa Python?

Pontificia Universidad JAVERIANA Bogotá

Los tipos de datos **simples** están formados por un solo objeto de un solo tipo

Tipo	Ejemplo	Definición
int	x = 1	Enteros
float	x = 1.0	Punto flotante (decimales)
complex	x = 1 + 2j	Complejos (parte real e imaginaria)
bool	x = True	Booleanos o lógicos: verdadero / falso
str	x = 'abc'	Texto
NoneType	x = None	Tipo especial para indicar valores nulos

Algunas librerías de Python como numpy o pandas agregan tipo de datos que se derivan de estos como int o float con diferentes niveles de precisión o longitud. También se generan los datos categóricos (ordenados y no ordenados) o tipo fecha-tiempo

¿Cómo resumir datos? Análisis Univariado

Pontificia Universidad JAVERIANA Bogotá

		Tipo de variable	
		Numérica	Categórica
Método	Estadísticas	Media, mediana, moda, rango, varianza, desviación estándar, cuártiles, percentiles, sesgo, curtosis, tablas de frecuencia	Conteo, proporción, tablas de frecuencia, tablas de contingencia
	Gráfico	Histograma, gráfico de caja, valores individuales	Gráfico de barras, Pareto, donas, torta

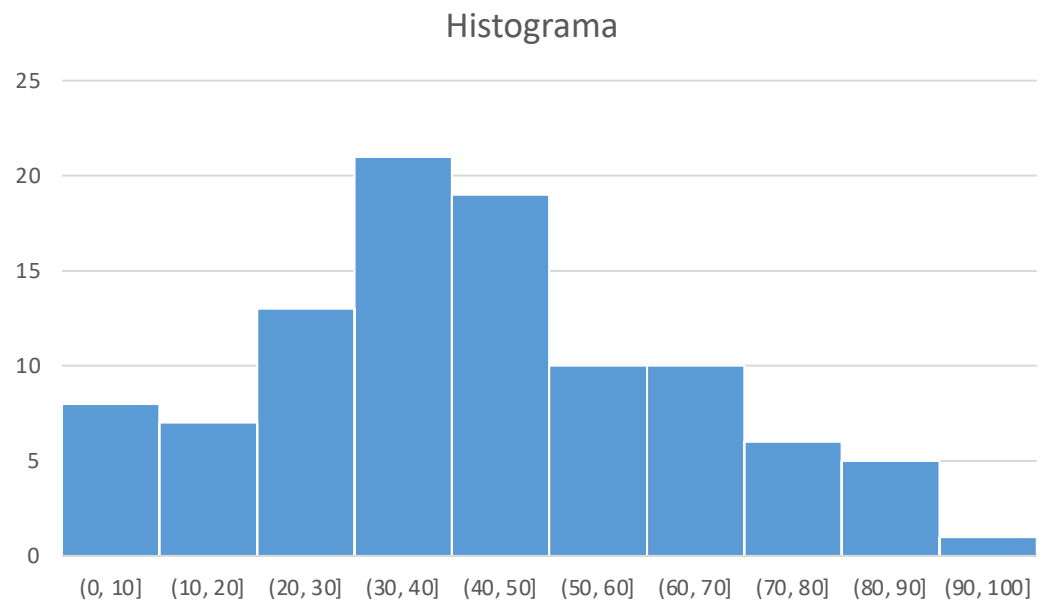
Distribución de frecuencia

La distribución de frecuencia es una tabla resumen en la que los datos están organizados en categorías numéricamente ordenadas.

Condensa datos brutos en una forma más útil.

Permite una rápida interpretación visual de los datos.

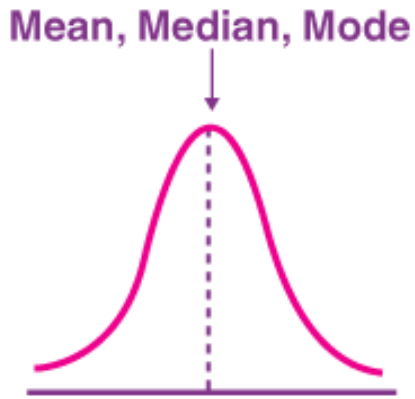
Permite determinación de ciertas características de los datos, incluyendo dónde están concentrados.

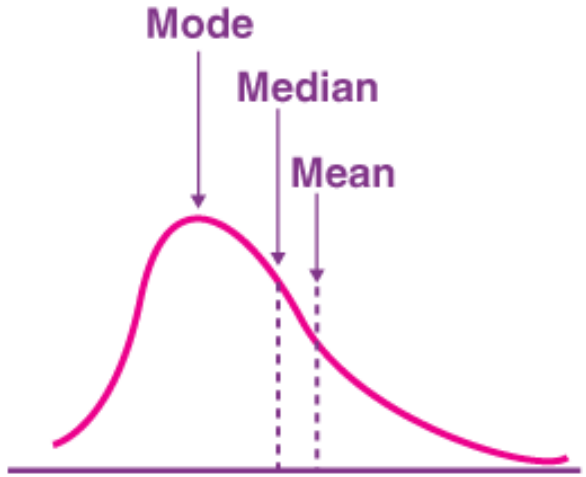


Rango Edad	Frecuencia	Porcentaje	Porcentaje acumulado
(0, 10]	8	8,0%	8,0%
(10, 20]	7	7,0%	15,0%
(20, 30]	13	13,0%	28,0%
(30, 40]	21	21,0%	49,0%
(40, 50]	19	19,0%	68,0%
(50, 60]	10	10,0%	78,0%
(60, 70]	10	10,0%	88,0%
(70, 80]	6	6,0%	94,0%
(80, 90]	5	5,0%	99,0%
(90, 100]	1	1,0%	100%
Total	100		

Medidas de tendencia central

- Las medidas de tendencia central también son conocidas como medidas de localización.
- El objetivo principal de las medidas de tendencia central es dar una idea de cuál es el valor típico o común de una variable determinada.
- Las tres medidas de tendencia central más comunes son la media aritmética, la mediana y la moda.





Medidas de tendencia central

Media Aritmética

Es la suma de todos los datos dividida entre la cantidad de datos. Calculemos el promedio de la nota de 7 estudiantes: 2 3 3 3 4 4 5

$$\overline{X} = \frac{\sum x_i}{n} = \frac{2+3+3+3+4+4+5}{7} = \frac{24}{7} = 3.43$$

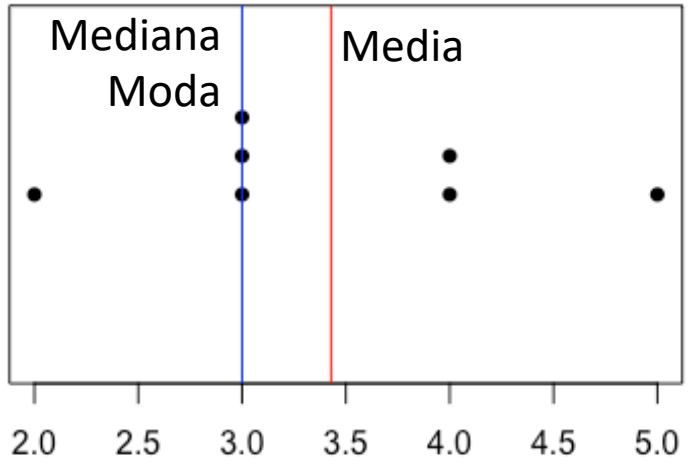
Mediana

Es el dato que divide la muestra en dos partes iguales: 2 3 3 3 4 4 5.

$$\tilde{X} = 3.0$$

La regla general para calcular la mediana:

$$\tilde{X} = \begin{cases} x_{(n+1)/2}, & \text{si } n \text{ es impar}, \\ \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right), & \text{si } n \text{ es par}. \end{cases}$$



Moda

Es el dato de la muestra que más se repite: 2 3 3 3 4 4 5

Moda = 3.0

Media vs mediana

- Existen algunos casos donde la media aritmética “falla” como medida de tendencia central.
- La media es sensible a los sesgos fuertes y los valores extremos.
- Considere los siguientes datos:
  - X = [90, 95, 95, 100, 105]
  - Y = [90, 95, 95, 100, 500]
- Al calcular los promedio de X y Y

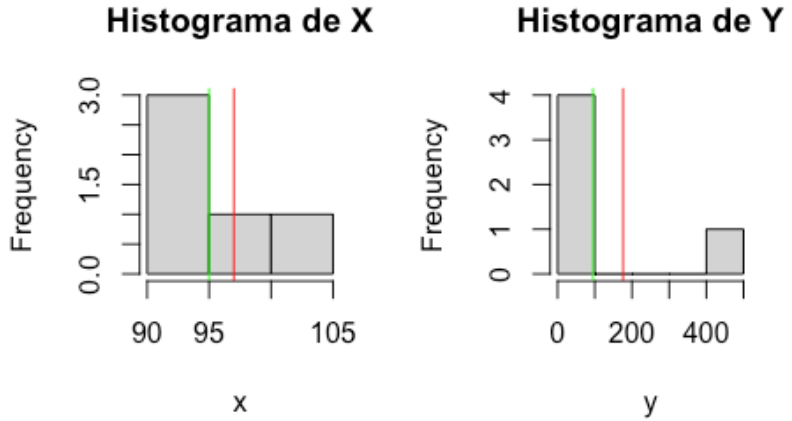
$$\overline{X} = \frac{90 + 95 + 95 + 100 + 105}{5} = \frac{485}{5} = 97$$

$$\overline{Y} = \frac{90 + 95 + 95 + 100 + 500}{5} = \frac{880}{5} = 176$$

- Al calcular la mediana de X y Y

X = [90, 95, 95, 100, 105] -> $\tilde{X} = 95$

Y = [90, 95, 95, 100, 500] -> $\tilde{Y} = 95$



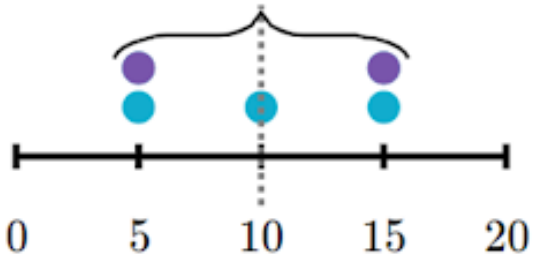
Piense en casos reales donde se pueden presentar sesgo o valores extremos

Medidas de dispersión

- La dispersión se refiere a la discrepancia o diferencia en los valores o resultados de una muestra.
- Por esta razón, las medidas de dispersión se denominan medidas de variabilidad.
- Las medidas de dispersión son:
  - Rango: R = mayor valor – menor valor
  - Varianza: $S^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$
  - Desviación estándar: $S = \sqrt{S^2}$

Medidas de dispersión

Muestra A



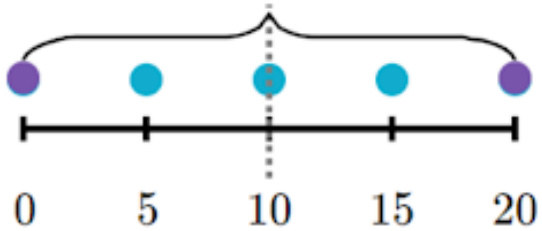
$$\text{Media}: \overline{X}_A = \frac{5+5+10+15+15}{5} = 10$$

$$\text{Rango}: R_A = 15 - 5 = 10$$

$$\text{Varianza}: S_A^2 = \frac{(5-10)^2+(5-10)^2+(10-10)^2+(15-10)^2+(15-10)^2}{5-1} = \frac{100}{4} = 25$$

$$\text{Desv. Estándar}: S_A = \sqrt{25} = 5.0$$

Muestra B



$$\text{Media}: \overline{X}_B = \frac{0+5+10+15+20}{5} = 10$$

$$\text{Rango}: R_B = 20 - 0 = 20$$

$$\text{Varianza}: S_B^2 = \frac{(0-10)^2+(5-10)^2+(10-10)^2+(15-10)^2+(20-10)^2}{5-1} = \frac{250}{4} = 62.5$$

$$\text{Desv. Estándar}: S_B = \sqrt{62.5} = 7.9$$

Observe que en la muestra B los datos están más dispersos en comparación a los datos de la muestra. Esto a su vez se refleja en las medidas de dispersión.

Medidas de posición

- Las medidas de posición se utilizan para determinar puntos de referencias que dividen la muestra ordenada en partes iguales.
- Los más utilizados son los:
  - Cuartiles (4 partes iguales)
  - Quintiles (5 partes iguales)
  - Deciles (10 partes iguales)
  - Percentiles (100 partes iguales).
- En estadística, un cuantil es una puntuación por debajo de la cual cae un determinado porcentaje de los valores de la muestra.
- Existen diferentes formas de calcular estas medidas, a continuación ilustramos una de estas.

Suponga que se tiene una muestra de 10 salarios anuales (millones) en una organización:

28, 30, 35, 38, 39, 46, 50, 52, 53, 70

Los cuartiles de la muestra:

Q2(50%) = (39 + 46)/2 = 42.5

Q1(25%) = 35

Q3(75%) = 52

Los quintiles:

Q1(20%) = (30 + 35)/2 = 32.5

Q2(40%) = (38 + 39)/2 = 38.5

Q3(60%) = (46 + 50)/2 = 48

Q4(80%) = (52 + 53)/2 = 52.5

Medidas de posición

- Los cuantiles se utilizan para posicionar o rankear las entidades o individuos dentro de una muestra o población.

- Esto permite tener una idea de la posición relativa dentro de la muestra.

Pontificia Universidad JAVERIANA Bogotá

GOBIERNO DE COLOMBIA · MINEDUCACIÓN · icfes mejor saber

REPORTE DE RESULTADOS DE ESTUDIANTES
•SABER PRO•

Fecha de aplicación
Fecha de publicación de resultados
Número de registro
Nombres y apellidos
Identificación
Institución: CORPORACION UNIVERSITARIA MINUTO DE DIOS -UNIMINUTO- BOGOTÁ D.C.
Código SNIES: 91237
Programa

PUNTAJE GLOBAL

De 300 puntos posibles, su puntaje global es ▸ **116**

GRUPO REFERENCIA

▸ ADMINISTRACIÓN Y AFINES

¿EN QUÉ PERCENTIL ME ENCUENTRO?

Con respecto a los estudiantes del país, usted está aquí.

MÓDULOS COMPETENCIAS GENÉRICAS

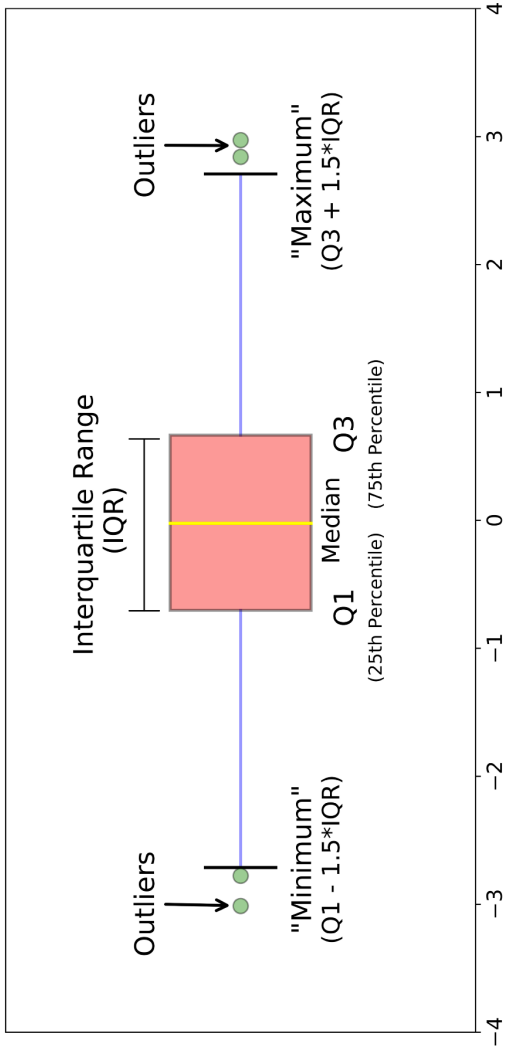
Módulos	De 300 puntos posibles, su puntaje global es	¿EN QUE PERCENTIL ME ENCUENTRO?
Comunicación Escrita[1]	--	
Razonamiento Cuantitativo	123	24 / 22
Lectura Crítica	176	86 / 78
Competencias Ciudadanas	132	46 / 40
Inglés	150	61 / 55

MÓDULOS COMPETENCIAS ESPECÍFICAS

Módulos	De 300 puntos posibles, su puntaje global es	¿EN QUE PERCENTIL ME ENCUENTRO?
Gestión De Organizaciones	159	73
Formulación, Evaluación Y Gestión De Proyectos	173	81
Gestión Financiera	98	6

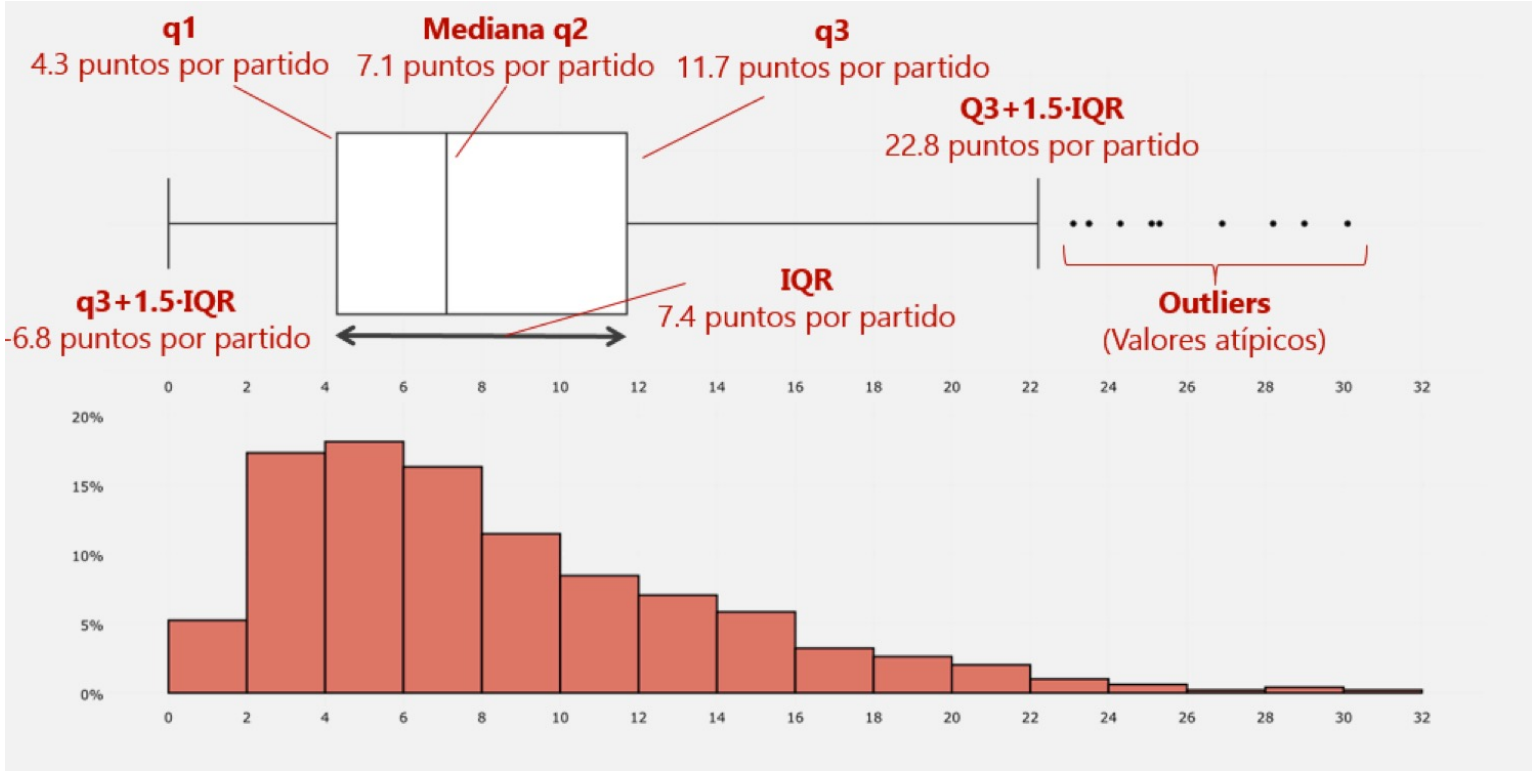
Gráficos de caja

- El diagrama de caja se basa en los cuartiles y divide los datos ordenados en cuatro grupos, cada uno con el 25%.
- De esta forma es posible visualizar dónde termina de acumularse 25% de los datos menores, y a partir de donde se localiza 25% de los datos mayores.
- Entre estos dos cuartiles se ubica el 50% de los datos que están al centro.
- Además de los cuartiles están involucrados los siguientes conceptos:
  - Rango intercuartílico: IQR = Q3 – Q1
  - Barrera interior izquierda: Q1 – 1.5IQR
  - Barrera interior derecha: Q1 + 1.5IQR



Valores extremos o atípicos

- Usualmente los valores atípicos se localizan de la siguiente forma:
  - Puntos por fuera de los bigotes en el gráfico de cajas.
  - Valores abajo y arriba de los percentiles 1% y 99%.
  - Aunque estos criterios pueden variar de acuerdo al contexto y la distribución de los datos.

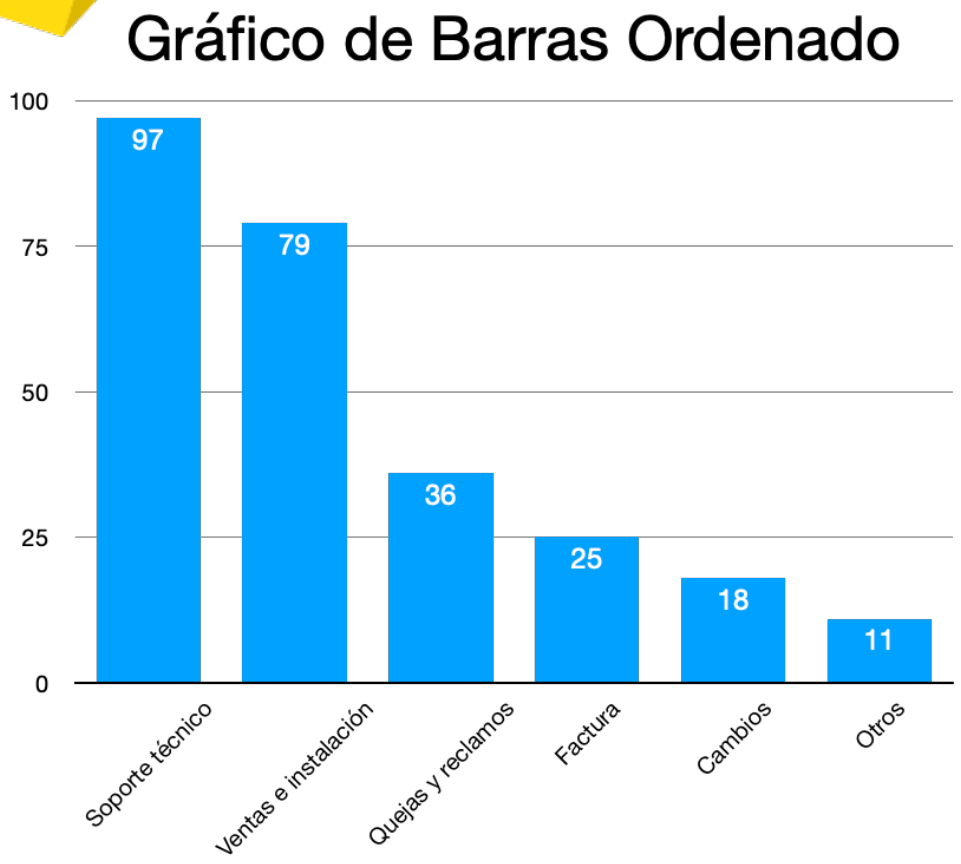


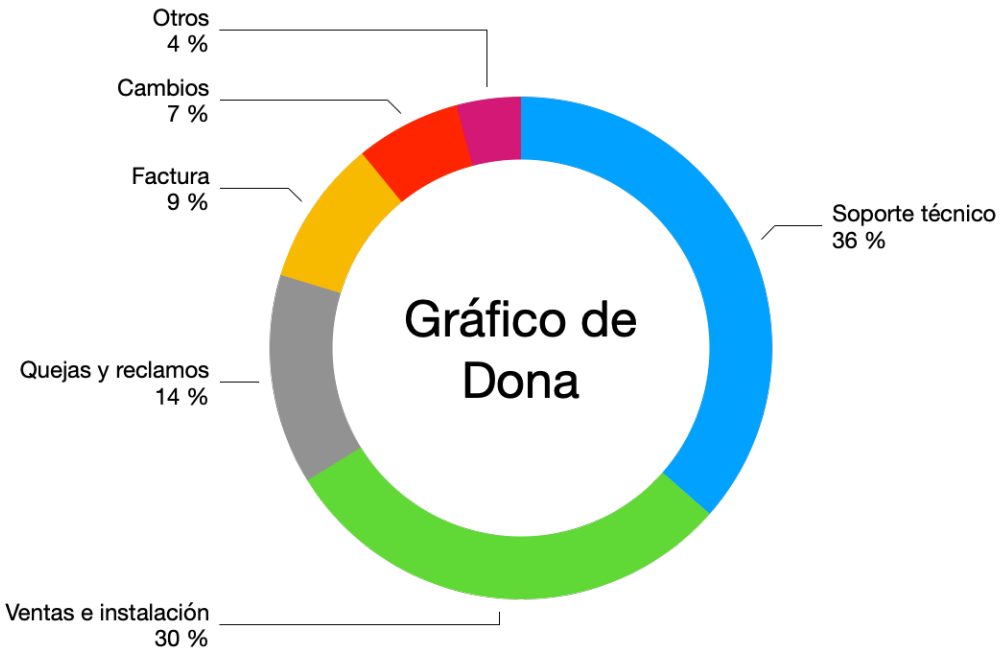
Tablas de frecuencia

- Cuando se tienen variables categóricas se utilizan las tablas de frecuencia para generar estadísticas, tales como:
  - Sumas
  - Sumas acumuladas
  - Porcentajes
  - Porcentajes acumulados
- Las tablas de frecuencia suministran información clave para:
  - Analizar las categorías
  - Su importancia relativa
  - Generar gráficos
- Cuando la variable categórica es nominal, es recomendable generar un ordenamiento según su frecuencia.
- Este tipo de análisis se conoce como Pareto, el cual permite identificar las categorías principales bajo la regla 80%/20%

Requerimiento	Total	Porcentaje	Porcentaje Acumulado
Soporte técnico	97	36,47 %	36,47 %
Ventas e instalación	79	29,70 %	66,17 %
Quejas y reclamos	36	13,53 %	79,70 %
Factura	25	9,40 %	89,10 %
Cambios	18	6,77 %	95,86 %
Otros	11	4,14 %	100,00 %
Total	266	100,00 %	

Gráficos para variables categóricas





Los gráficos de barras se utilizan cuando se quiere presentar las categorías como un ranking de acuerdo a su importancia o peso

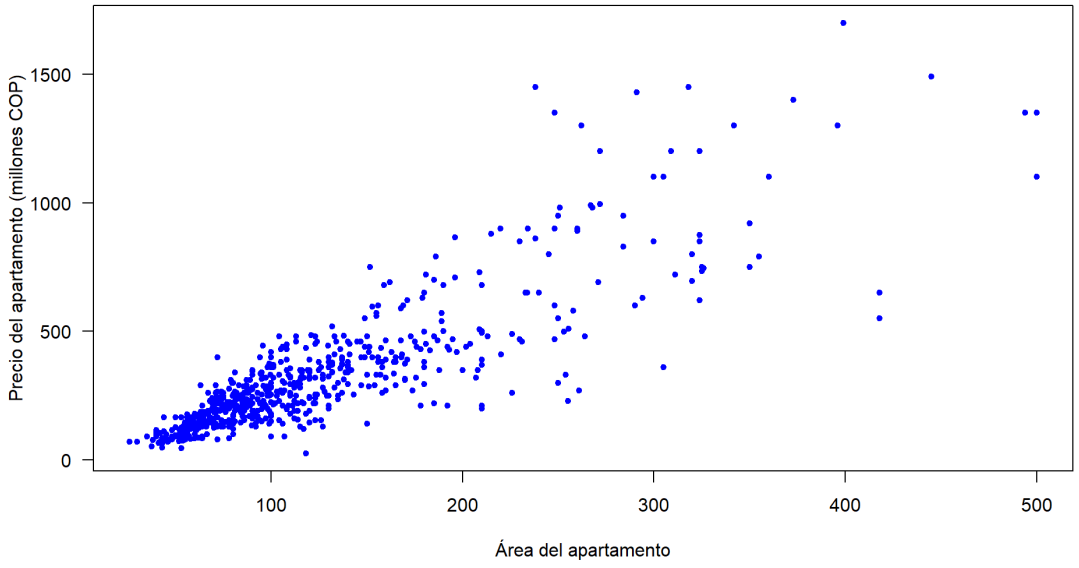
Los gráficos de dona o torta se utiliza cuando se quiere representar el peso o porcentaje de las categorías con respecto al todo.

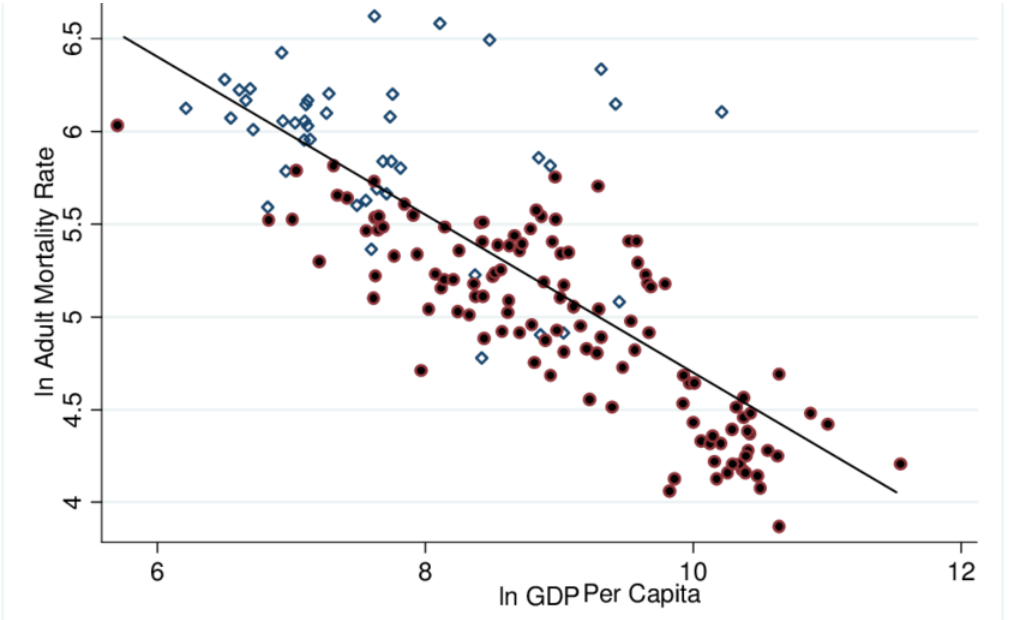
Análisis bivariado

- Variable Numérica vs Variable Numérica:
  - Gráfico de dispersión
  - Covarianza y correlación
- Variable Numérica vs Variable Categórica
  - Gráficos de caja o violin
  - ANOVA, prueba t, prueba kruskal-wallis, prueba Wilcoxon
- Variable Categórica vs Variable Categórica
  - Gráfico de barras agrupados o apilados
  - Tablas de contingencia o tablas cruzadas
- Variable fecha-tiempo vs Variable Numérica
  - Gráfico de serie de tiempo
  - Autocorrelación

Gráfico de dispersión

- Los diagramas de dispersión utilizan las coordenadas cartesianas para representar la relación entre dos variables de interés (X, Y).
- X representa la variable independiente (Eje Horizontal), y por lo general es aquella que puede controlar o manipular el analista.
- Y representa la variable sobre lo cual se analiza la relación de dependencia, por esta razón se le conoce comúnmente como variable dependiente





Coeficiente de Correlación

Pontificia Universidad JAVERIANA Bogotá

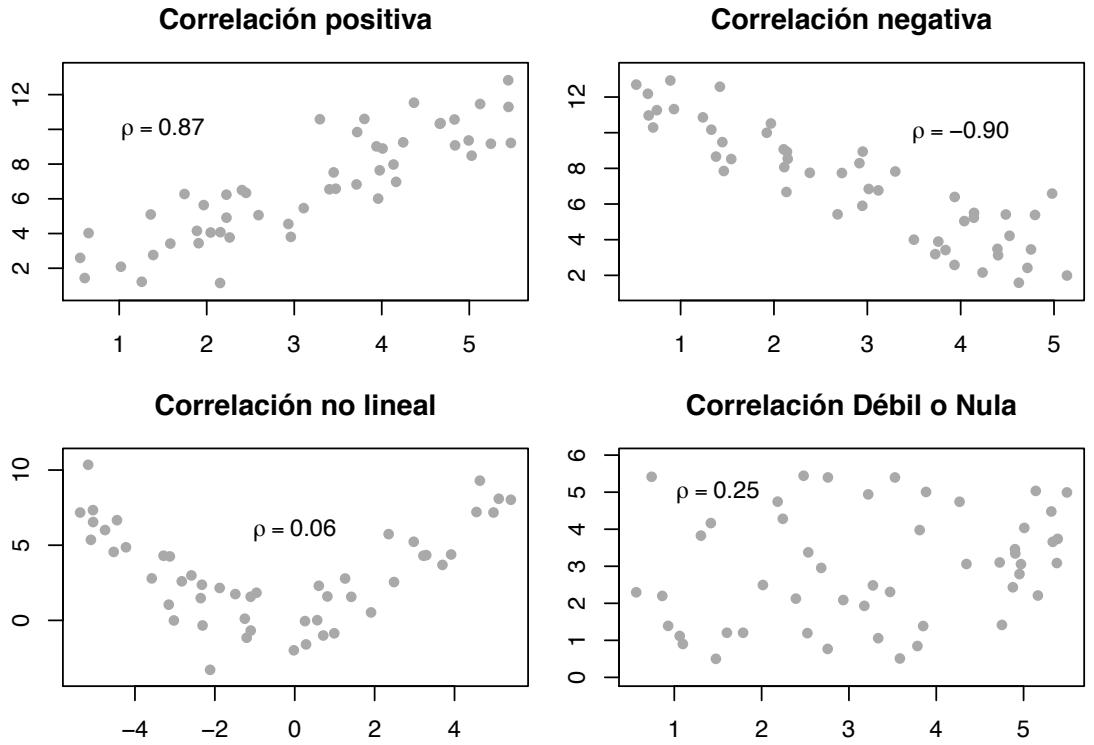
- El coeficiente de correlación permite medir la intensidad y el tipo de asociación lineal entre dos variables.
- El tipo de correlación se clasifica de la siguiente manera:
  - Correlación Lineal Positiva ($0.5 \leq \rho \leq 1$): Si el conjunto de puntos en el plano presentan un comportamiento ascendente .
  - Correlación Lineal Negativa ($-1 \leq \rho \leq -0.5$): Si el conjunto de punto el plano presenta un comportamiento descendente.
  - Correlación Débil o Nula ($-0.5 < \rho < 0.5$): Si no es posible identificar un com- portamiento en el conjunto de puntos.
  - Correlación No Lineal: Si es posible identificar diferentes tendencias en varias secciones del plano.

$$\rho = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$$

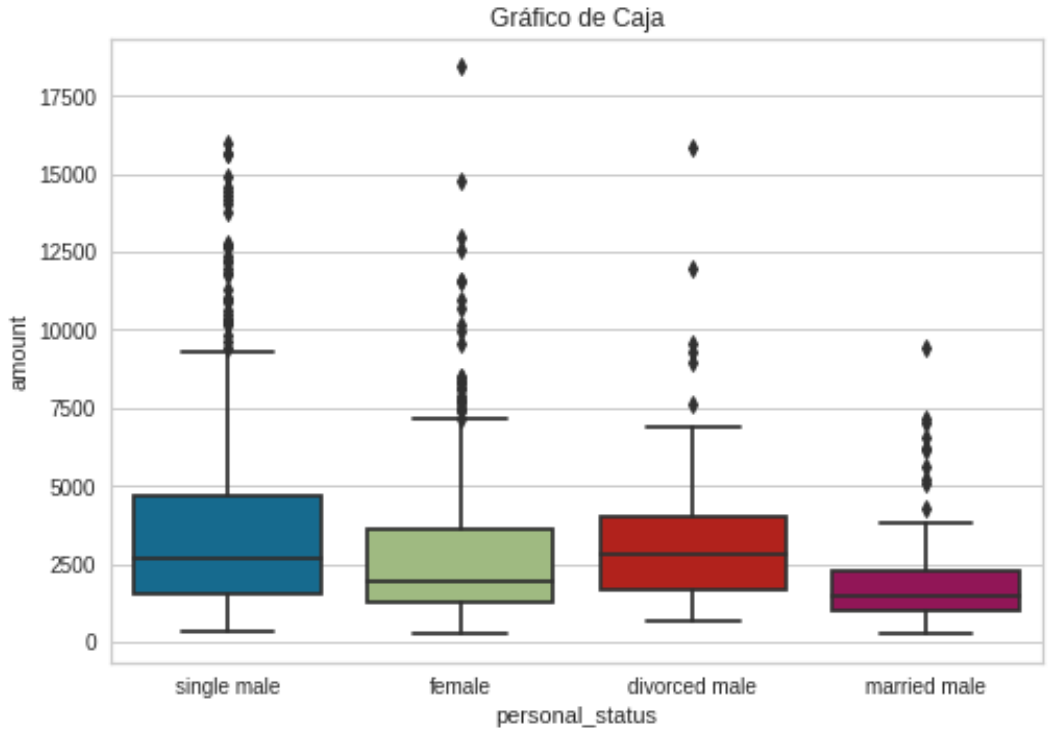
$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\sum_{i=1}^{n} y_i)}{n}$$

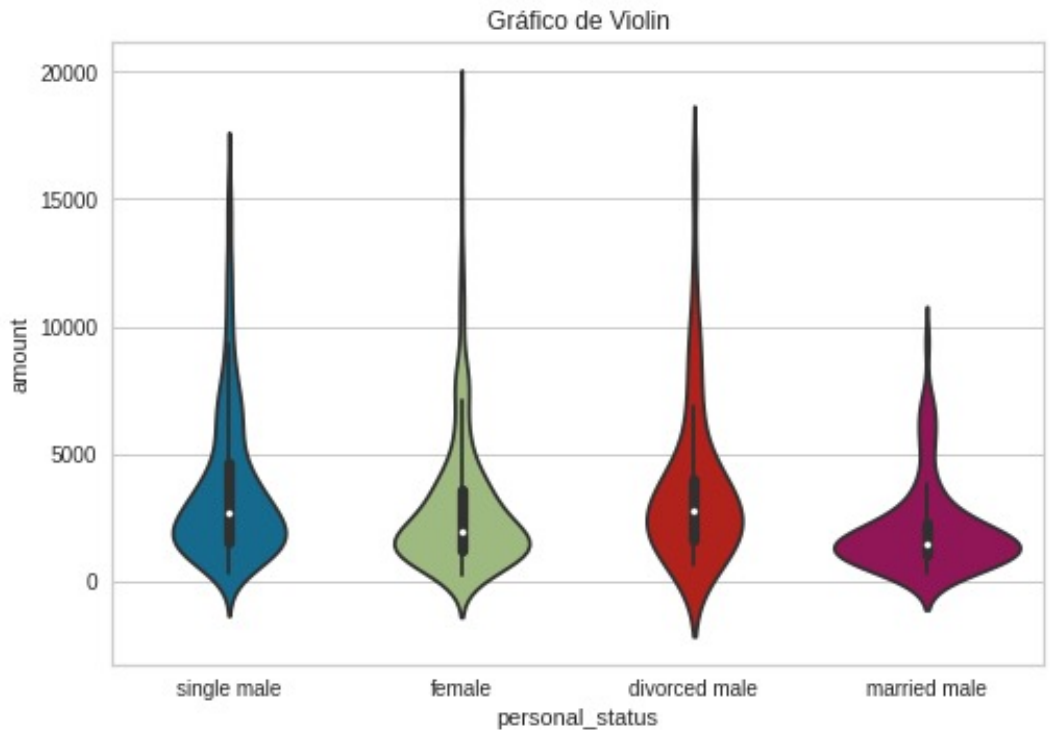


Gráficos de Caja para Comparar Categorías

Pontificia Universidad JAVERIANA Bogotá

Los diagramas de caja y violines son útiles cuando queremos visualizar muchas distribuciones a la vez y/o si estamos interesados principalmente en los cambios generales entre las distribuciones





Tablas Cruzadas

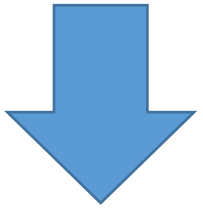
- En algunos casos es interesante analizar la combinación de mas de una categoría.
- En este caso se utilizan las tablas cruzadas, las cuales representan los conteos dentro de las combinación de las categorías, atributos o valores de las variables relacionadas.
- En las tablas cruzadas cada celda representa el conteo de dicha combinación.

Ejemplo:

- Muestra aleatoria de 400 facturas.
- Cada factura se categoriza en monto alto, medio, bajo.
- Cada factura se analiza para buscar errores.
- Se organizan los datos en la tabla de contingencia de la derecha.

	Sin Errores	Errores	Total
Monto Bajo	170	20	190
Monto Medio	100	40	140
Monto Alto	65	5	70
Total	335	65	400

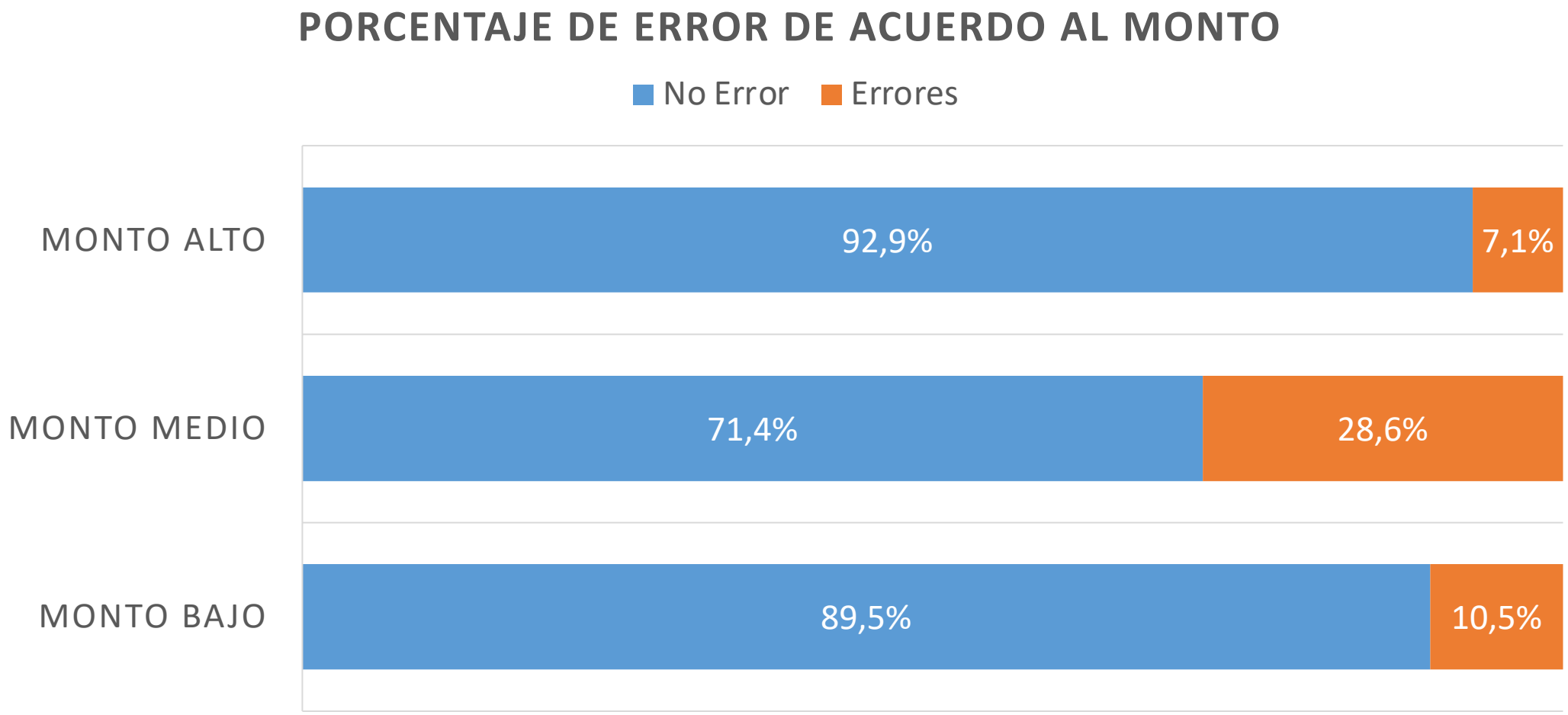
Observe como cambian los porcentajes de error de acuerdo al monto



89.5% = 170 / 190
71.4% = 100 / 140
92.9% = 65 / 70

	Sin Error	Errores	Total
Monto Bajo	89,5%	10,5%	100,0%
Monto Medio	71,4%	28,6%	100,0%
Monto Alto	92,9%	7,1%	100,0%

Gráfico de Barras

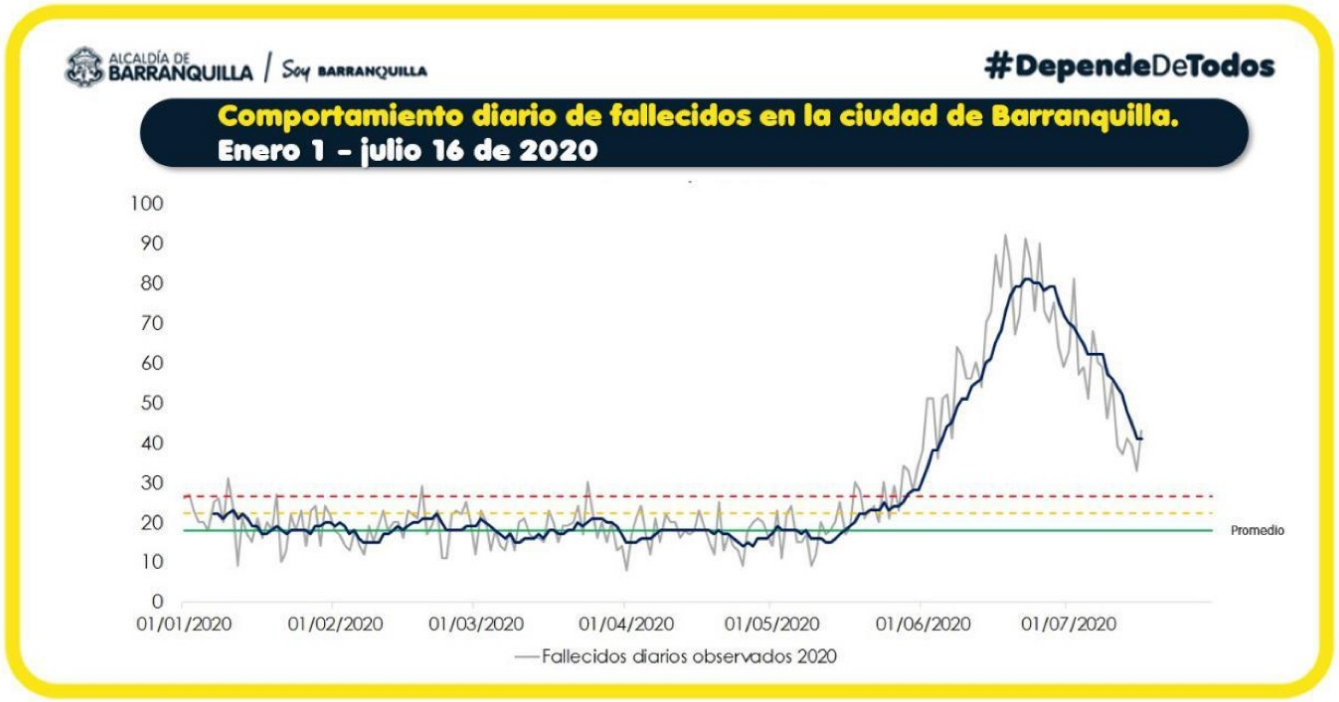


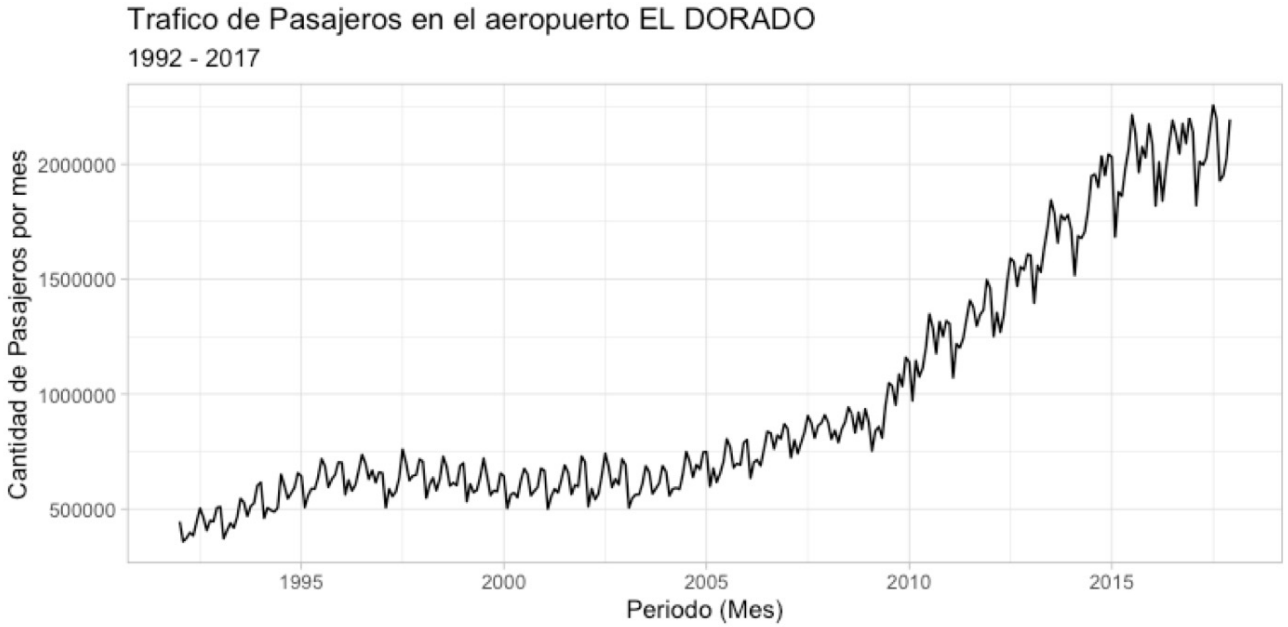
Series de tiempo

Las series de tiempo permite analizar la relación entre una variable numérica y variable tipo fecha-tiempo.

Permite identificar patrones tales:

- Tendencia
- Ciclos
- Estacionalidad





Datos faltantes

Pontificia Universidad JAVERIANA Bogotá

Los datos faltantes (NA, NAN, NULL), también conocidos como valores nulos o valores perdidos, son un componente común en conjuntos de datos del mundo real.

Tipos de datos faltantes	Ejemplo
MCAR (Missing Completely At Random - Faltantes Completamente aleatorios): Los datos faltantes son aleatorios y no hay un patrón discernible.	En una encuesta sobre preferencias musicales en una feria, algunas personas se niegan a responder sobre su género musical favorito.
MAR (Missing At Random - Faltantes aleatorios): La probabilidad de que falten datos depende de otras variables observadas en el conjunto de datos.	En la recopilación de datos académicos, donde se registran calificaciones en matemáticas y participación en actividades extracurriculares, la ausencia más frecuente de datos sobre la participación en actividades extracurriculares para estudiantes con calificaciones más bajas en matemáticas.
MNAR (Missing Not At Random - Faltantes no aleatorios): Los datos faltantes dependen del valor de la variable que se mide. En este caso no es fácil la imputación	En una encuesta sobre ingresos familiares, cuando algunos encuestados optan por no revelar sus ingresos,

¿Cómo tratar los datos faltantes?

Pontificia Universidad JAVERIANA Bogotá

- **Eliminación de Filas o Columnas:** Una estrategia simple pero efectiva que implica eliminar filas o columnas con datos faltantes. Sin embargo, esto puede resultar en una pérdida significativa de información si los datos faltantes son frecuentes.
- **Imputación de Datos:** Llenar valores faltantes con estimaciones basadas en otros datos disponibles. Esto puede incluir el uso de la media, mediana o moda para variables numéricas, o la asignación del valor más común para variables categóricas.
- **Modelado de Datos Faltantes:** Utilizar algoritmos de aprendizaje automático para predecir valores faltantes en función de variables disponibles. Útil cuando los datos faltantes siguen patrones complejos.
- **Creación de Variables Indicadoras:** Crear variables binarias que indiquen la presencia o ausencia de datos en una columna, facilitando el modelado de la influencia de los datos faltantes.
- **Técnicas de Interpolación:** Aplicar técnicas de interpolación para estimar valores faltantes en series temporales u otros contextos, basándose en datos antes y después del punto faltante.