



Pontificia Universidad
JAVERIANA
Bogotá

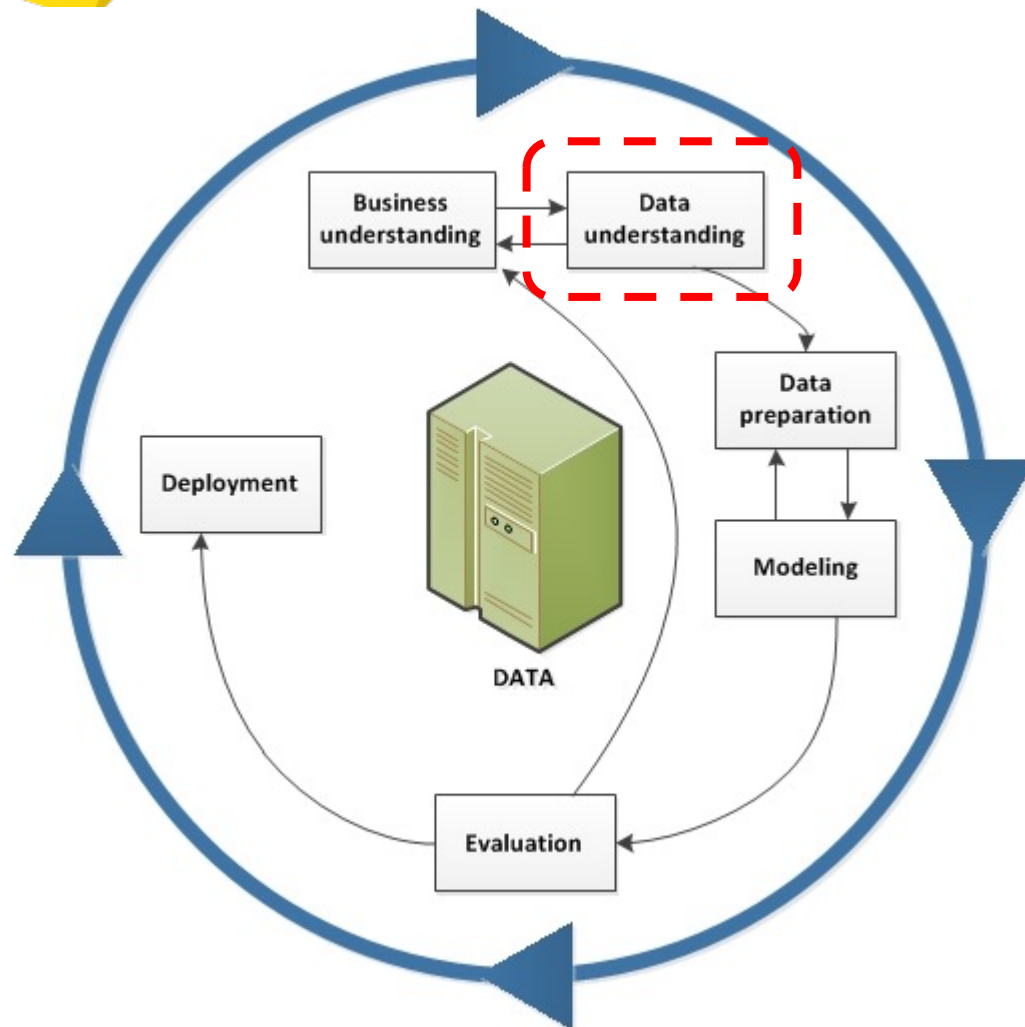
Análisis Exploratorio de Datos

Heriberto Felizzola

Cross-Industry Standard Process for Data Mining



Pontificia Universidad
JAVERIANA
Bogotá



- 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?



Pontificia Universidad
JAVERIANA
Bogotá

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."



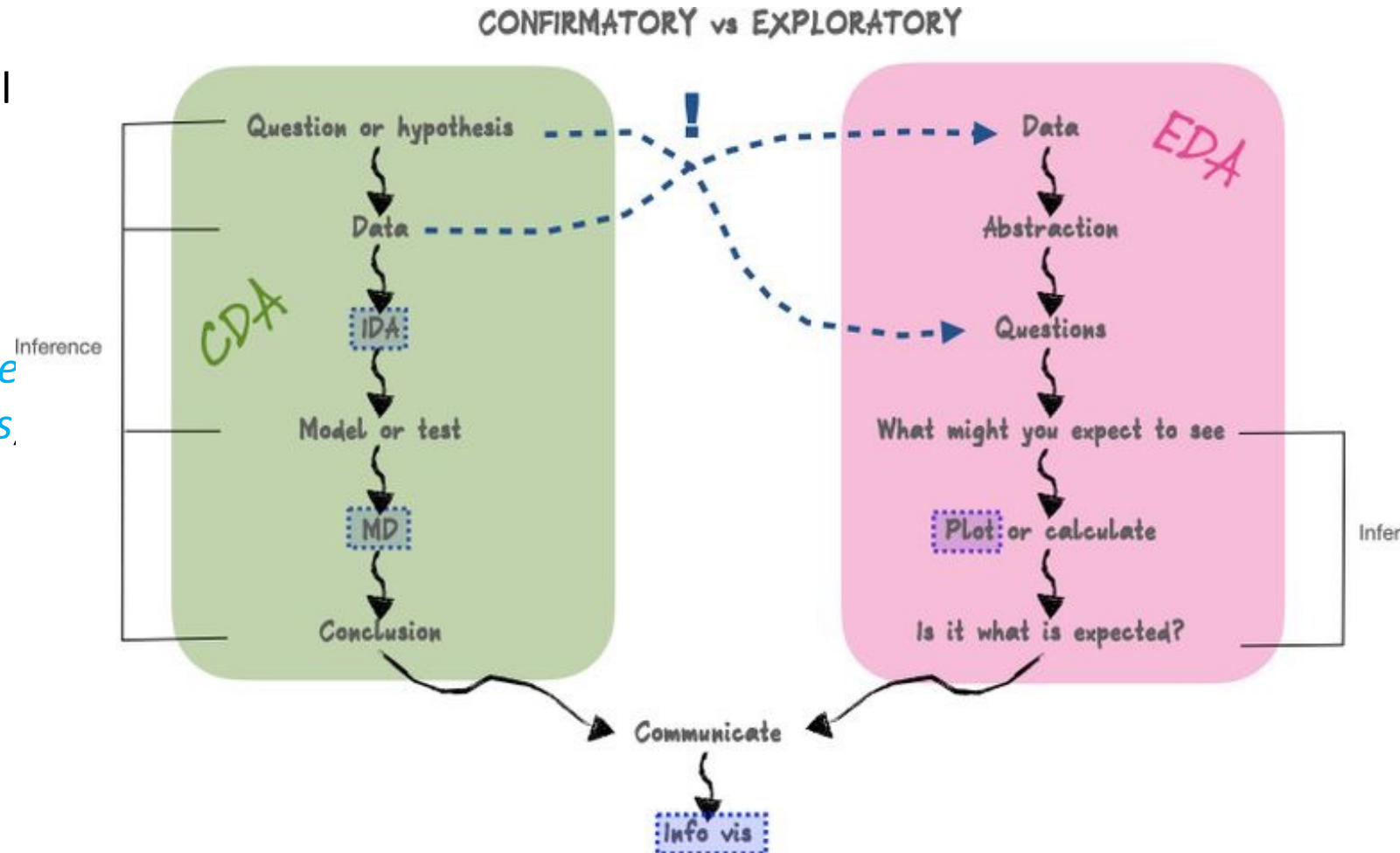
"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

Análisis Confirmatorio vs Exploratorio



Pontificia Universidad
JAVERIANA
Bogotá

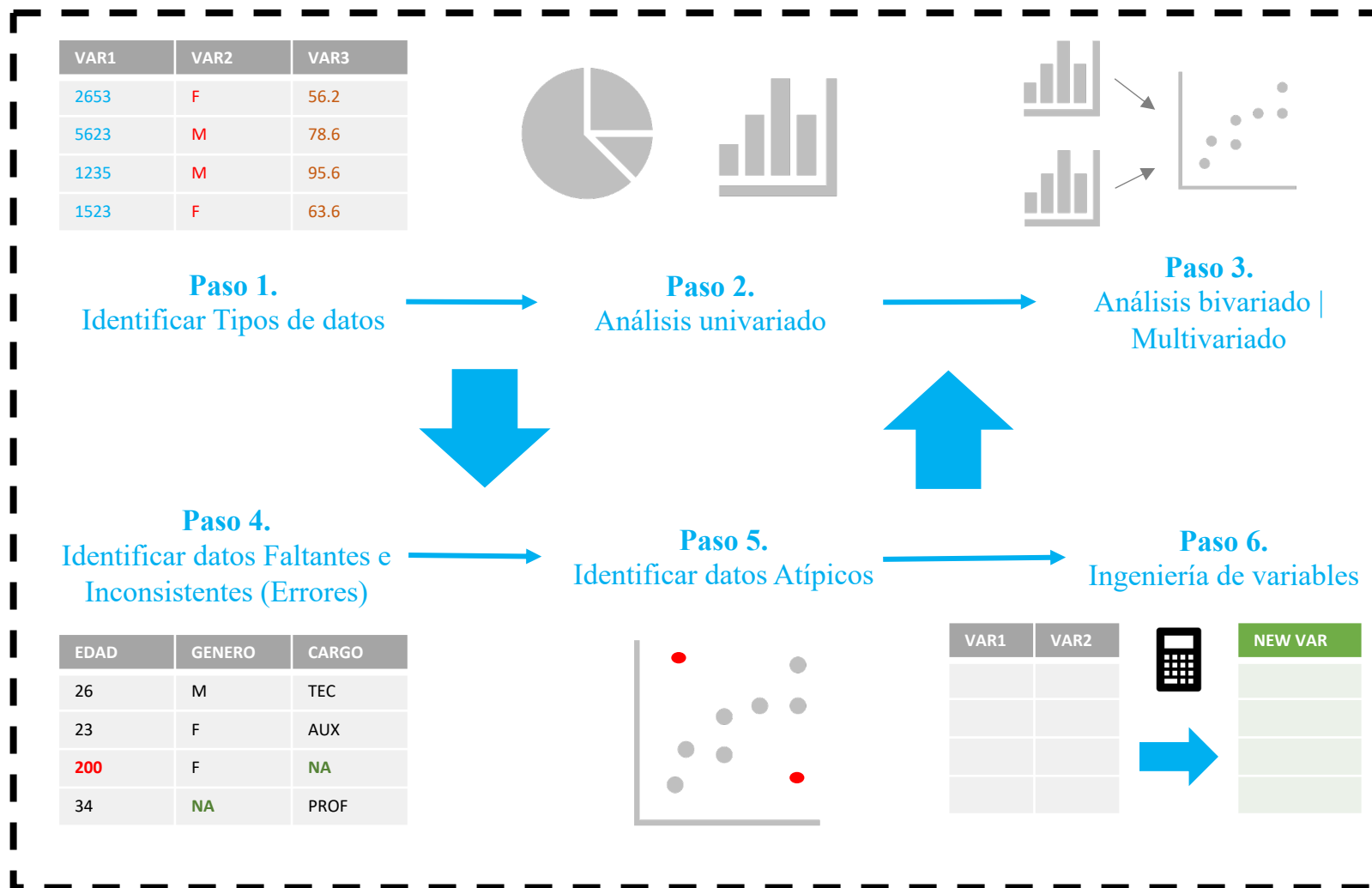
- A menudo, se confunde el EDA con el análisis estadístico tradicional.
- La principal diferencia radica en el enfoque.
- Mientras que el *análisis tradicional se centra en probar hipótesis específicas*, el *EDA se enfoca en descubrir información sin prejuicios*.
- Es un proceso más exploratorio y menos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



Tareas claves en un EDA



Pontificia Universidad
JAVERIANA
Bogotá




Ideas Claves


Preguntas


Comunicación


Informe

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?



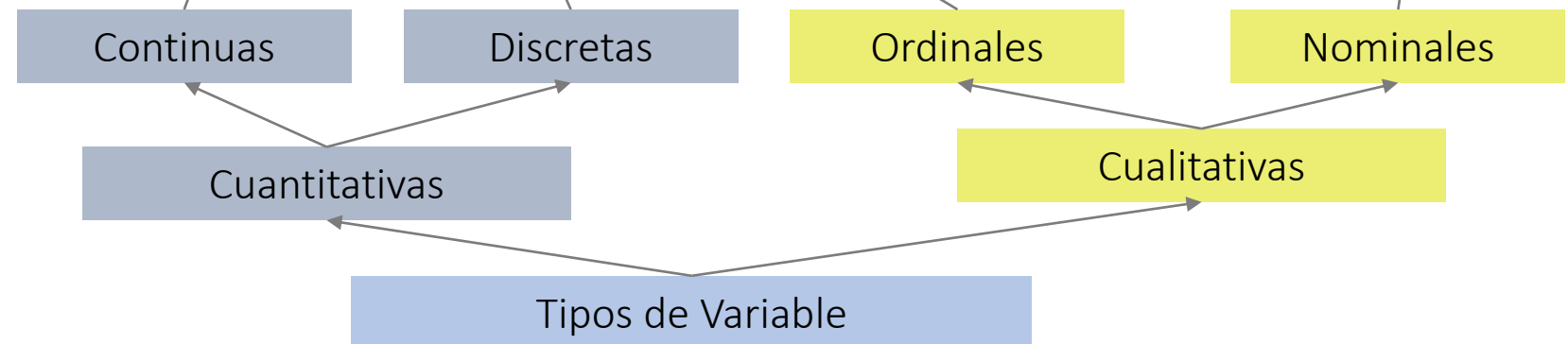
Pontificia Universidad
JAVERIANA
Bogotá

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

Medidas de tendencia central



Pontificia Universidad
JAVERIANA
Bogotá

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

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

Mediana

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

$$\tilde{X} = 3.0$$

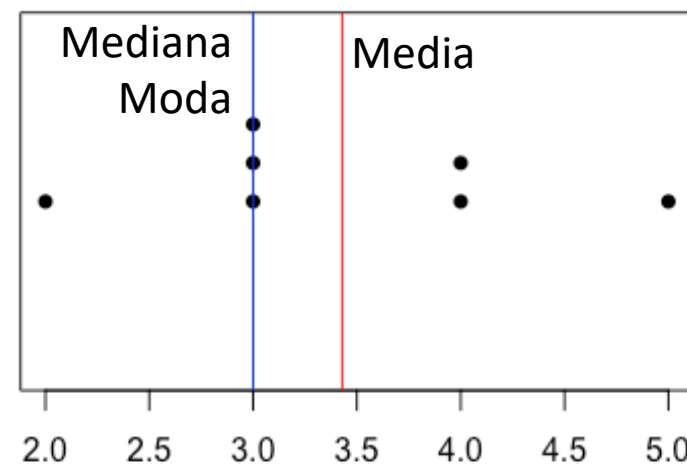
La regla general para calcular la mediana:

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

Moda

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

$$\text{Moda} = 3.0$$

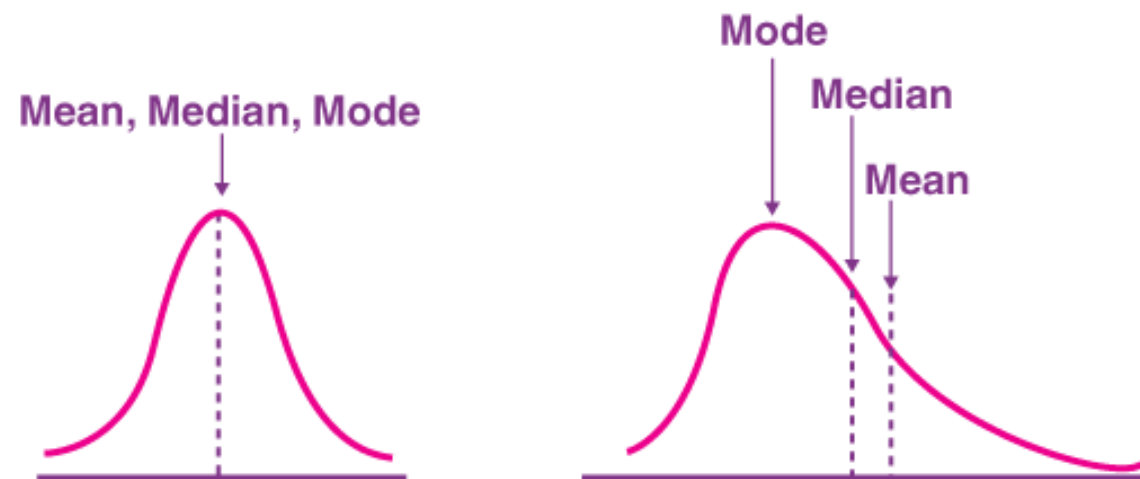


Medidas de tendencia central



Pontificia Universidad
JAVERIANA
Bogotá

- 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.



Media vs mediana



Pontificia Universidad
JAVERIANA
Bogotá

- 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 promedios de X y Y

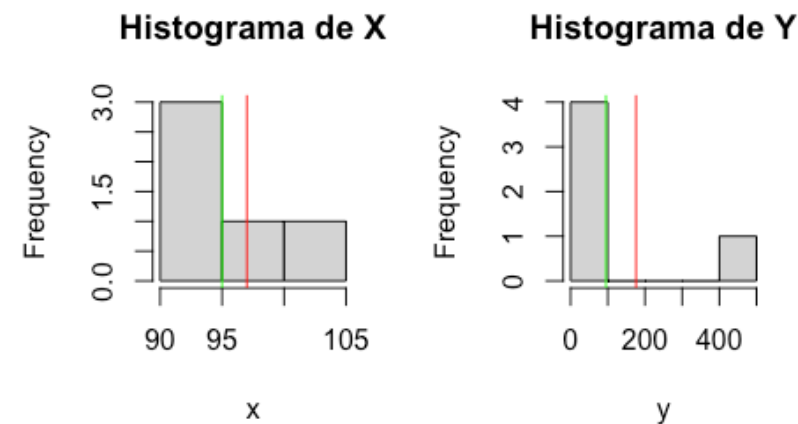
$$\bar{X} = \frac{90 + 95 + 95 + 100 + 105}{5} = \frac{485}{5} = 97$$

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

- Al calcular la mediana de X y Y

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

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



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

Medidas de dispersión



Pontificia Universidad
JAVERIANA
Bogotá

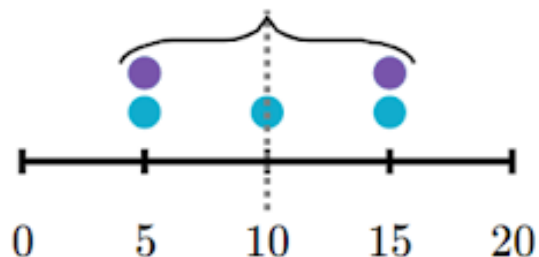
- La dispersión se refiere a la discrepancia o diferencia en los valores o resultados de una muestra.
- Por esta razón, las medidas de dispersión se denominan medidas de variabilidad.
- Las medidas de dispersión son:
 - Rango: $R = \text{mayor valor} - \text{menor valor}$
 - Varianza: $S^2 = \frac{\sum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

Medidas de dispersión



Pontificia Universidad
JAVERIANA
Bogotá

Muestra A



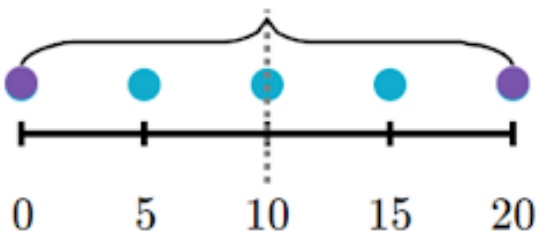
$$\text{Media : } \bar{X}_A = \frac{5 + 5 + 10 + 15 + 15}{5} = 10$$

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

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

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

Muestra B



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

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

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

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

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.

Distribución de frecuencia



Pontificia Universidad
JAVERIANA
Bogotá

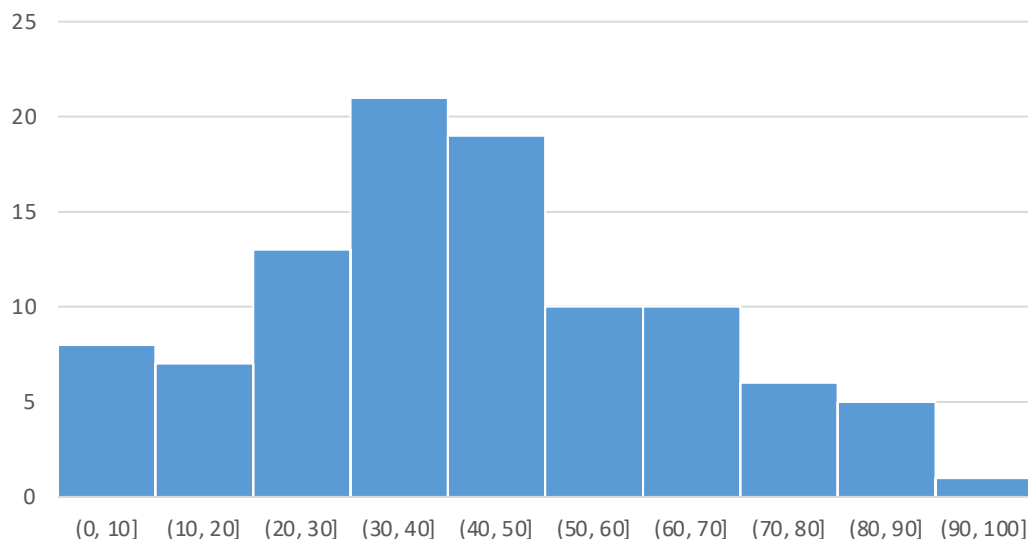
La distribución de frecuencia es una tabla resumen en la que los datos están organizados en categorías numéricamente ordenadas.

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.

Histograma



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 posición



Pontificia Universidad
JAVERIANA
Bogotá

- 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



Pontificia Universidad
JAVERIANA
Bogotá

- 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.



	Fecha de aplicación	_____
	Fecha de publicación de resultados	_____
	Número de registro	_____
	Nombres y apellidos	
Identificación		
Institución <u>CORPORACIÓN UNIVERSITARIA MINUTO DE DIOS-UNIMINUTO-</u>		
<u>BOGOTÁ D.C.</u>		
Código SNIES <u>91237</u>		
Programa _____		

REPORTE DE RESULTADOS DE ESTUDIANTES •SABER PRO.

PUNTAJE GLOBAL

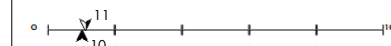
De 300 puntos posibles, su puntaje global es **116**

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	--	0 11 100
Razonamiento Cuantitativo	123	0 24 22 100
Lectura Crítica	176	0 86 78 100
Competencias Ciudadanas	132	0 46 40 100
Inglés	150	0 61 55 100

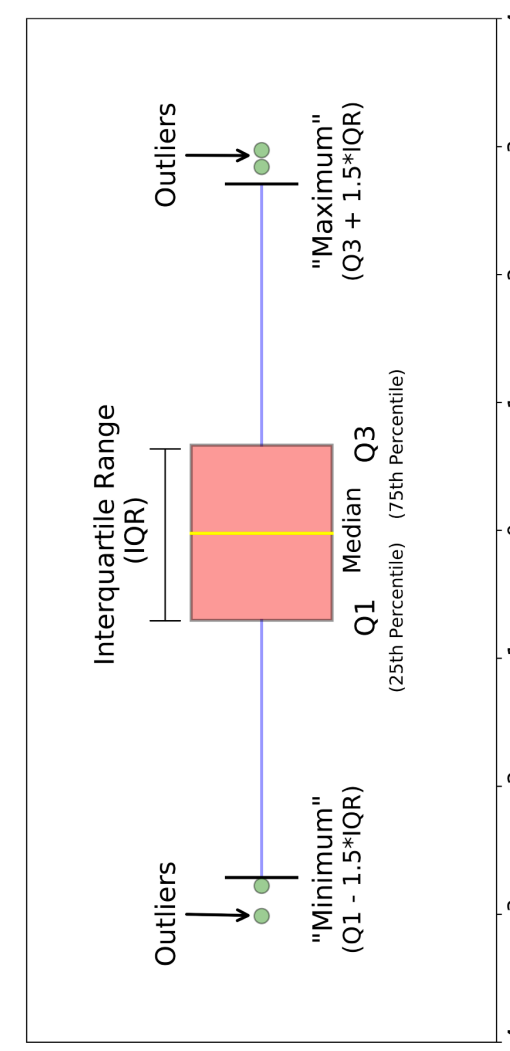
MÓDULOS COMPETENCIAS ESPECÍFICAS		
Módulos	De 300 puntos posibles, su puntaje global es	¿EN QUE PERCENTIL ME ENCUENTRO?
Gestión De Organizaciones	159	0 73 100
Formulación, Evaluación Y Gestión De Proyectos	173	0 81 100
Gestión Financiera	98	0 6 100

Gráficos de caja



Pontificia Universidad
JAVERIANA
Bogotá

- 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: $Q1 + 1.5IQR$

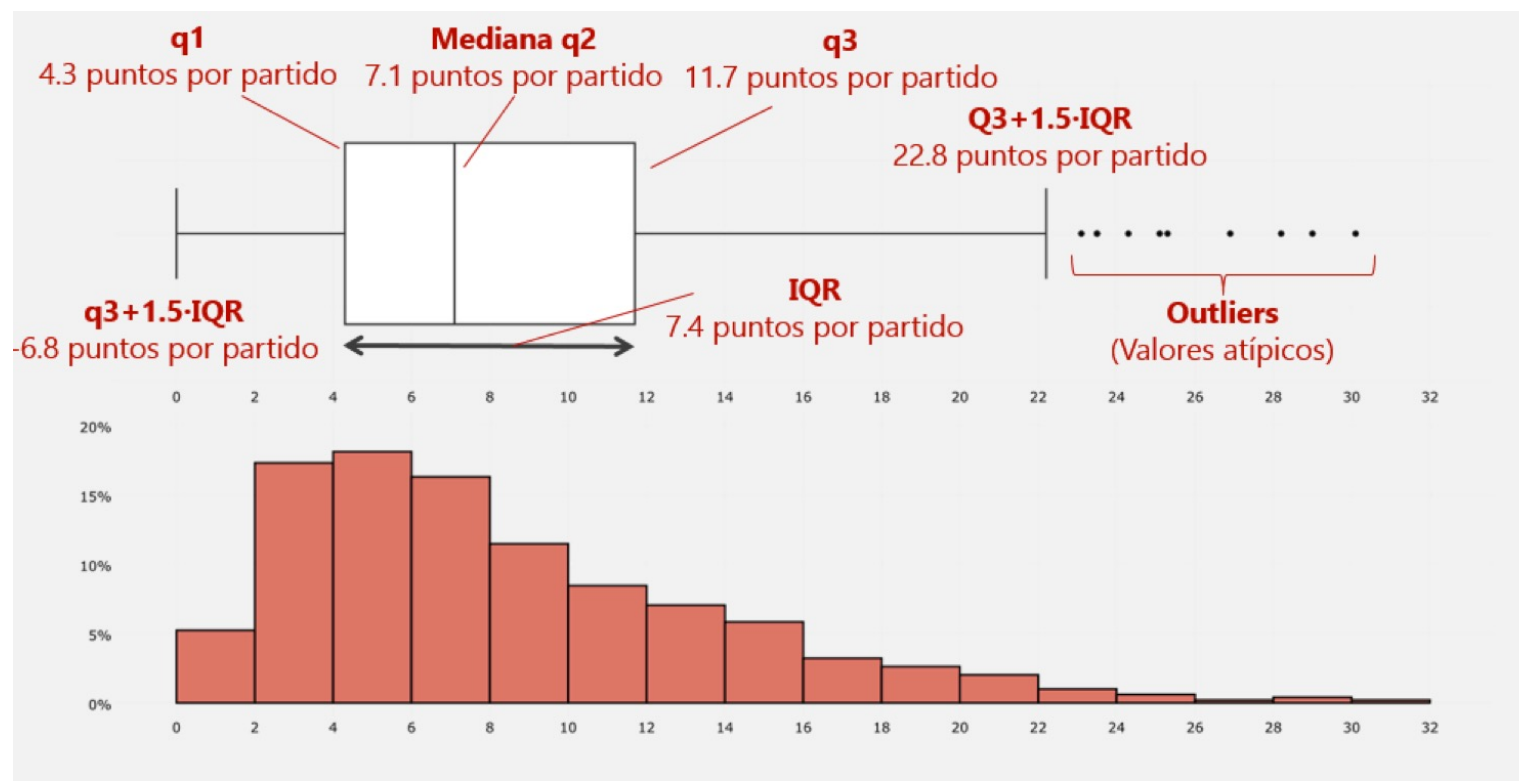


Valores extremos o atípicos



Pontificia Universidad
JAVERIANA
Bogotá

- 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



Pontificia Universidad
JAVERIANA
Bogotá

- 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%

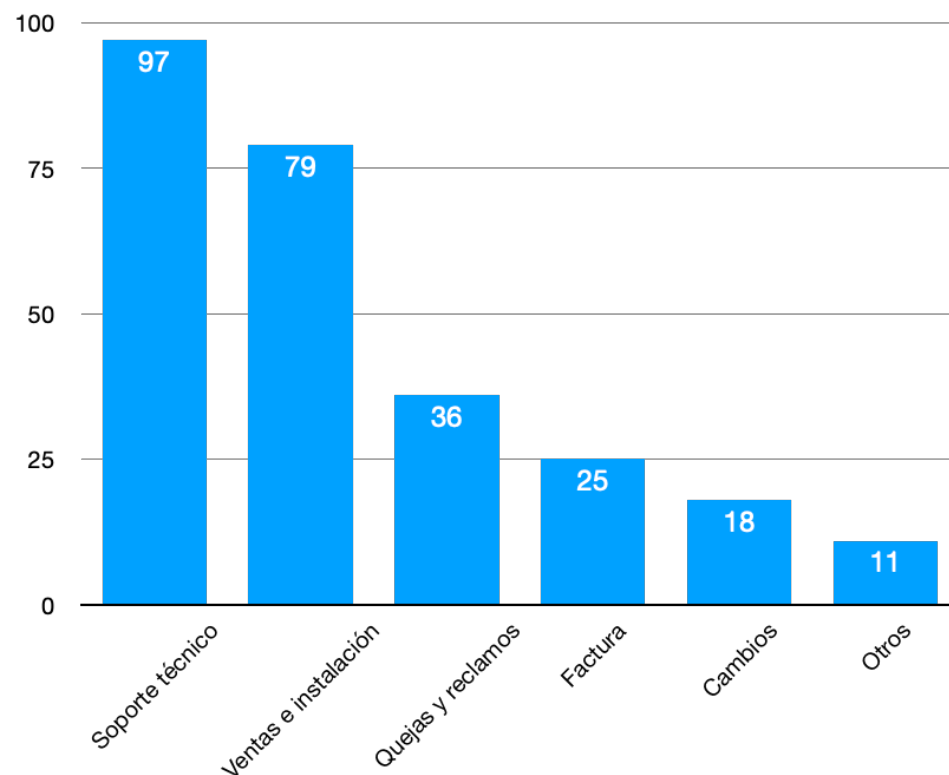
Requerimiento	Total	Porcentaje	Porcentaje Acumulado
Soporte técnico	97	36,47 %	36,47 %
Ventas e instalación	79	29,70 %	66,17 %
Quejas y reclamos	36	13,53 %	79,70 %
Factura	25	9,40 %	89,10 %
Cambios	18	6,77 %	95,86 %
Otros	11	4,14 %	100,00 %
Total	266	100,00 %	

Gráficos para variables categóricas

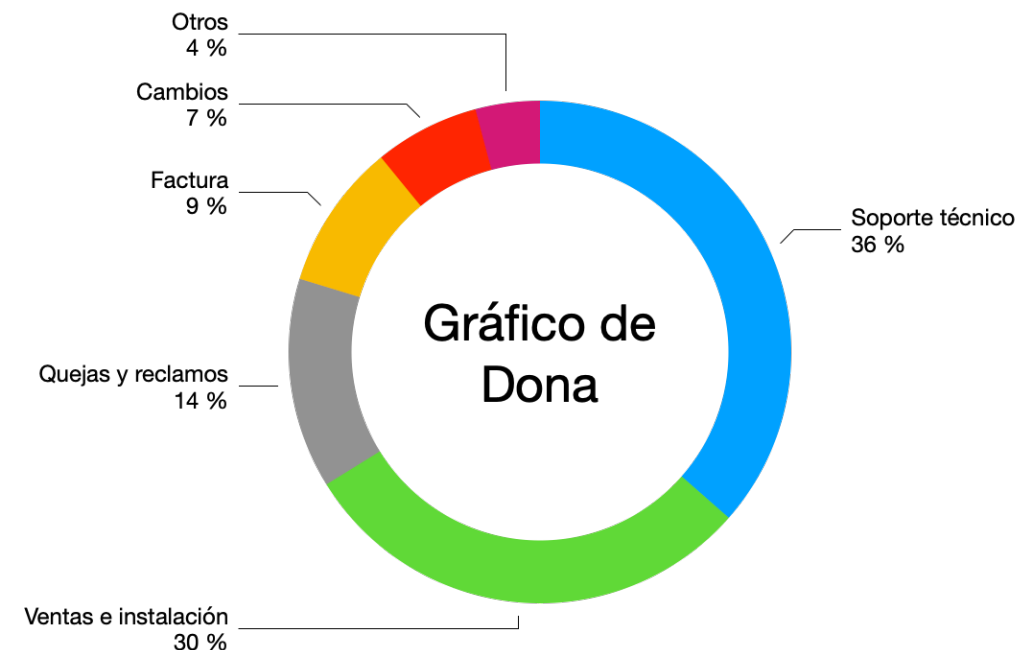


Pontificia Universidad
JAVERIANA
Bogotá

Gráfico de Barras Ordenado



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



Pontificia Universidad
JAVERIANA
Bogotá

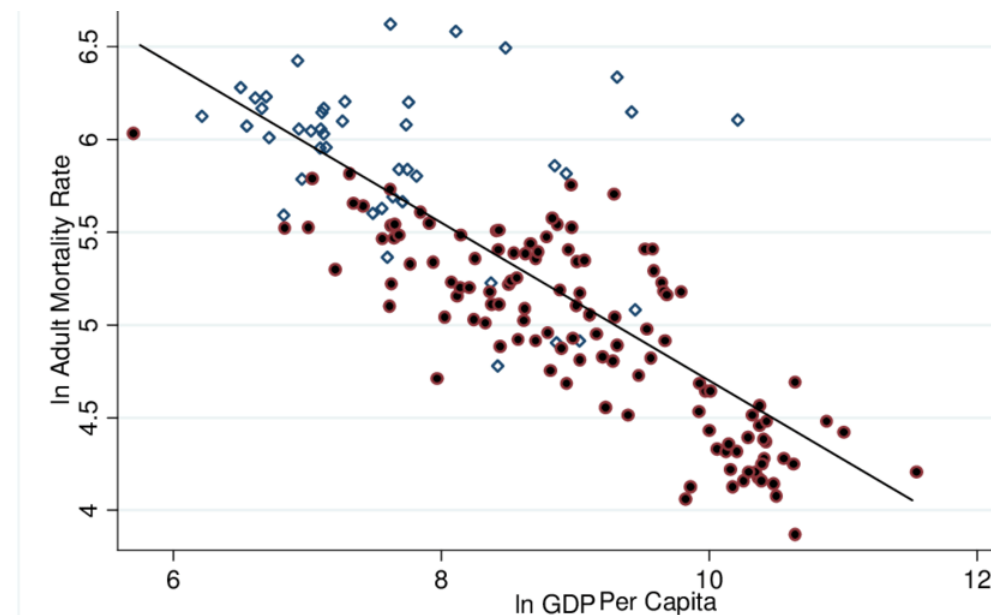
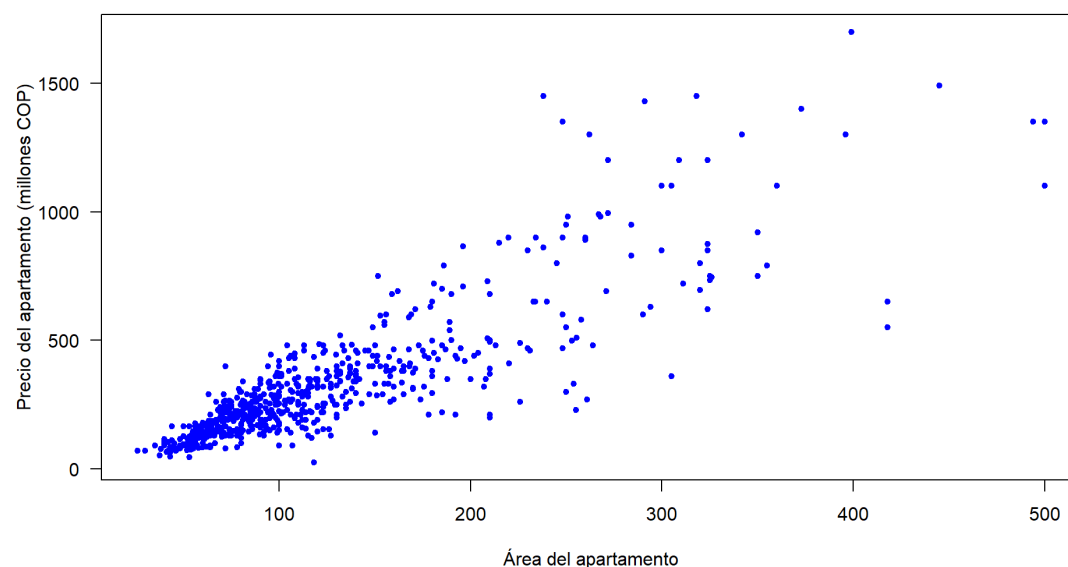
- Variable Numérica vs Variable Numérica:
 - Gráfico de dispersión
 - Covarianza y correlación
- Variable Numérica vs Variable Categórica
 - Gráficos de caja o violin
 - ANOVA, prueba t, prueba kruskal-wallis, prueba Wilcoxon
- Variable Categórica vs Variable Categórica
 - Gráfico de barras agrupados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

Gráfico de dispersión



Pontificia Universidad
JAVERIANA
Bogotá

- 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 puntos en el plano presenta un comportamiento descendente.
 - Correlación Débil o Nula ($-0.5 < \rho < 0.5$): Si no es posible identificar un comportamiento en el conjunto de puntos.
 - Correlación No Lineal: Si es posible identificar diferentes tendencias en varias secciones del plano.

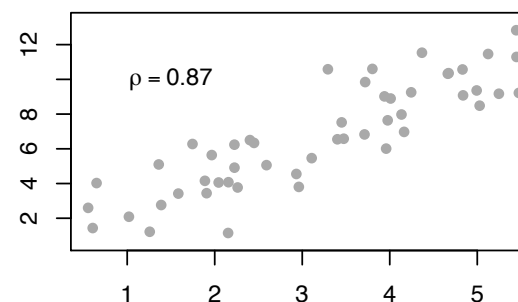
$$\rho = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

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

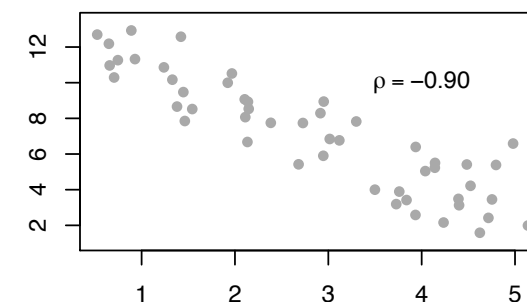
$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}$$

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

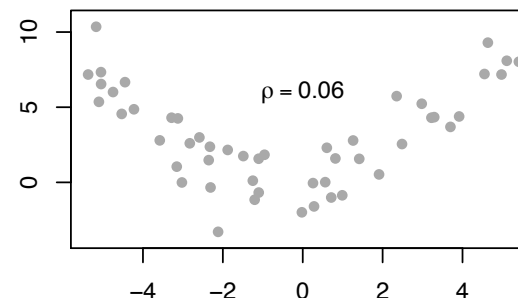
Correlación positiva



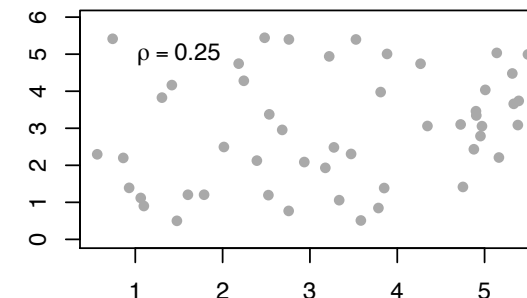
Correlación negativa



Correlación no lineal



Correlación Débil o Nula

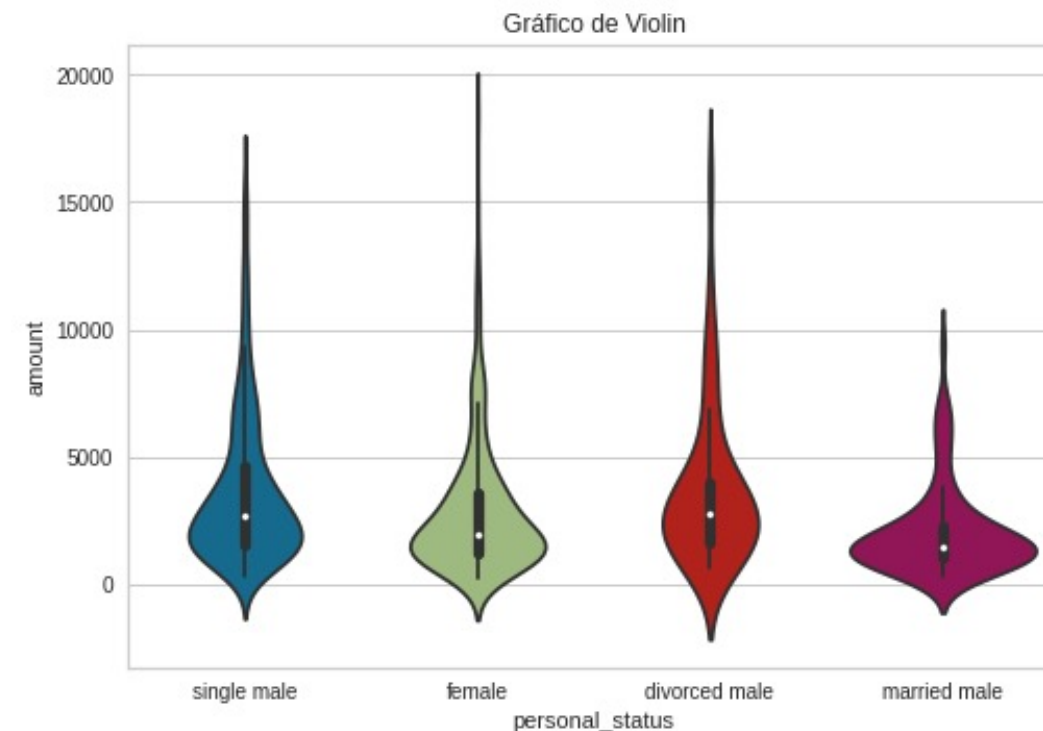
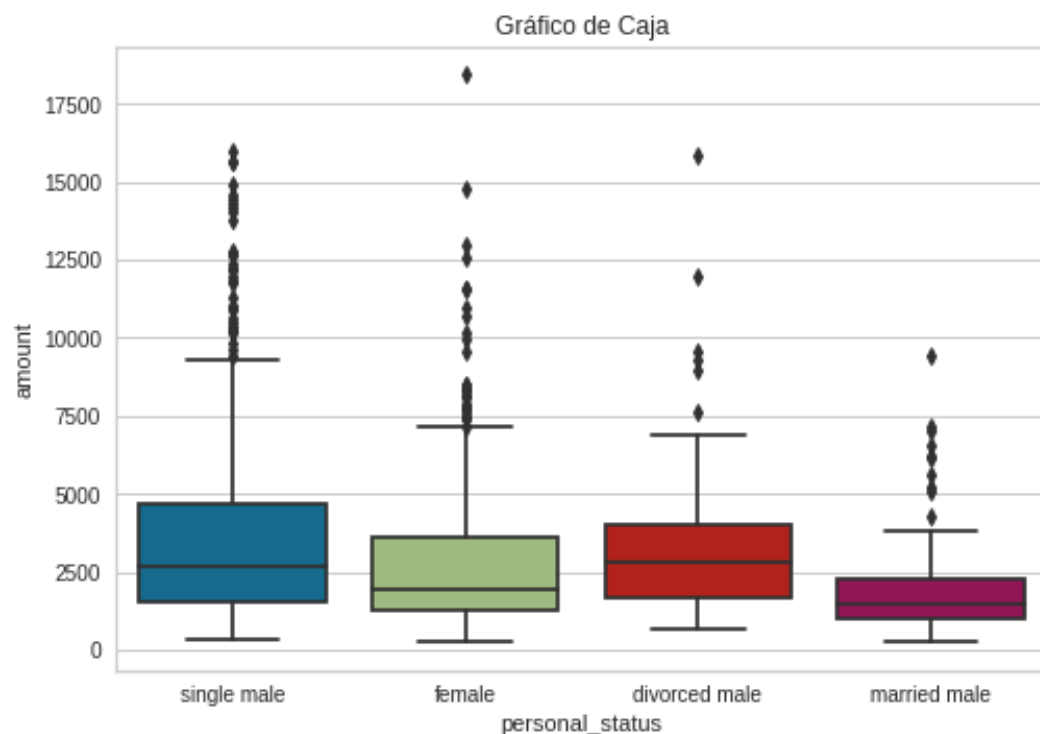


Gráficos de Caja para Comparar Categorías



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



Pontificia Universidad
JAVERIANA
Bogotá

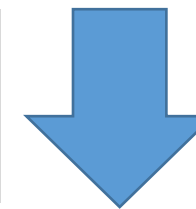
- En algunos casos es interesante analizar la combinación de mas de una categoría.
- En este caso se utilizan las tablas cruzadas, las cuales representan los conteos dentro de las combinación de las categorías, atributos o valores de las variables relacionadas.
- En las tablas cruzadas cada celda representa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

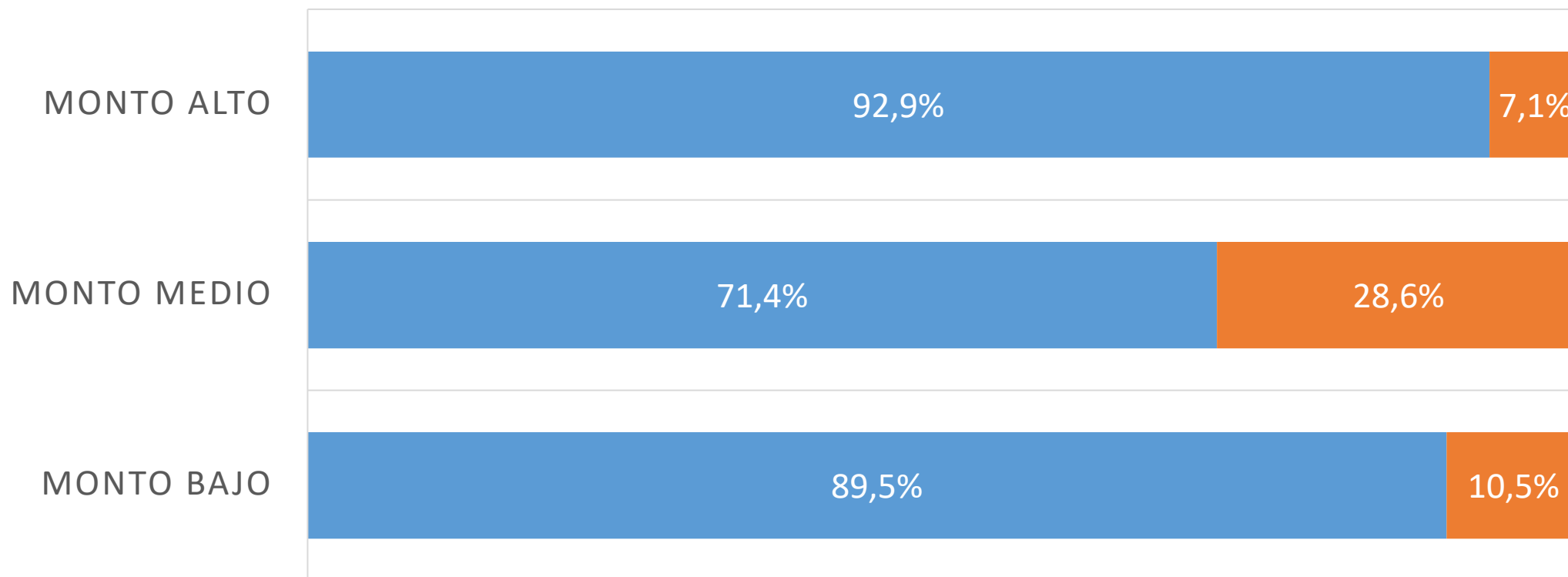
Gráfico de Barras



Pontificia Universidad
JAVERIANA
Bogotá

PORCENTAJE DE ERROR DE ACUERDO AL MONTO

■ No Error ■ Errores



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



Trafico de Pasajeros en el aeropuerto EL DORADO
1992 - 2017

