Metodología para visualizar datos

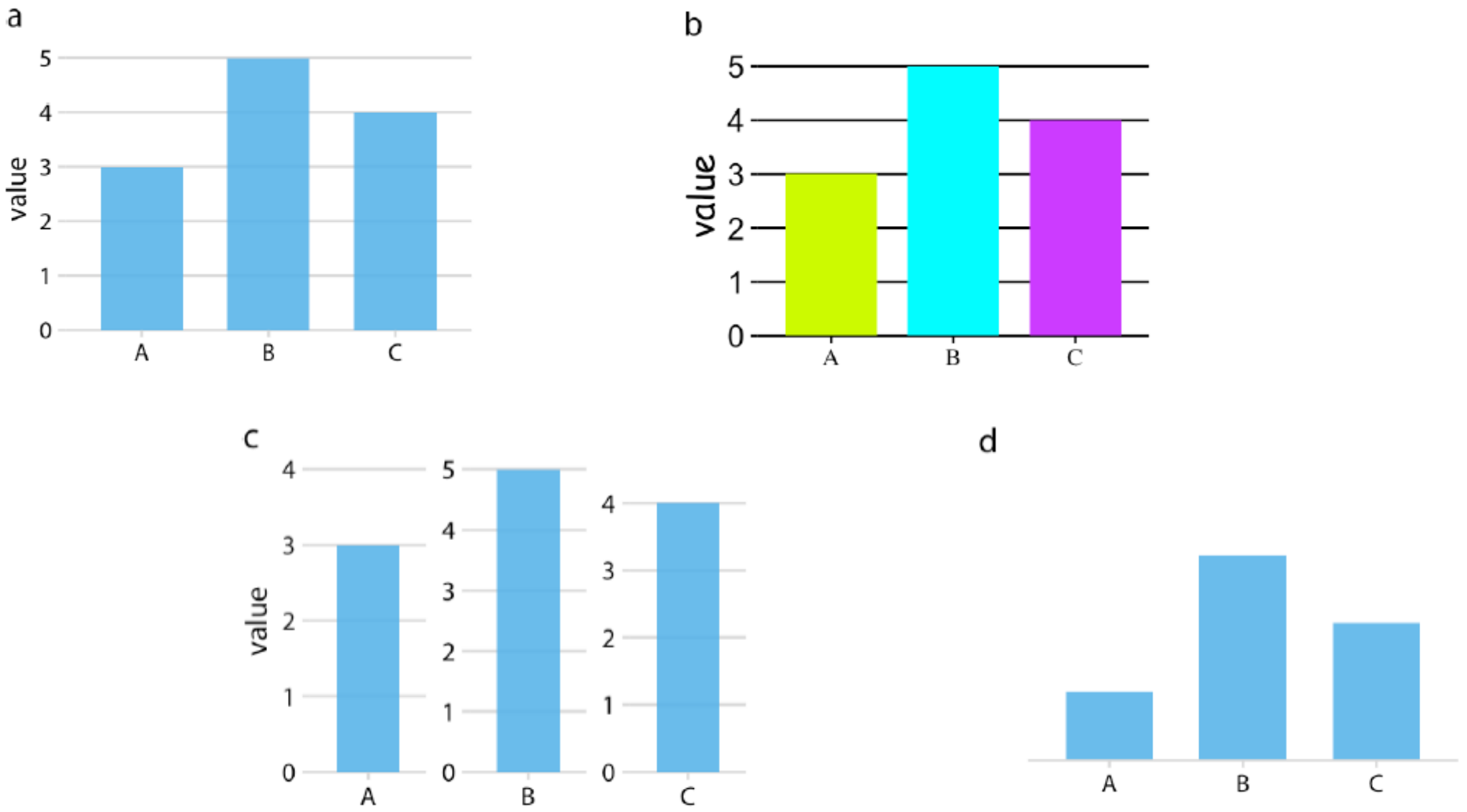
Maria Isabel Serrano G.
maria-serrano@javeriana.edu.co

Agenda

Metodologías de visualización

Exploración de datos

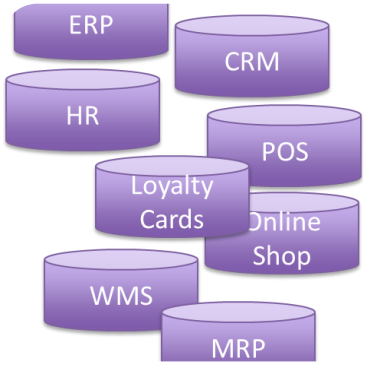
Ejercicio


a
value
5
4
3
2
1
0
A
B
C
b
value
5
4
3
2
1
0
A
B
C
c
value
4
3
2
1
0
A
5
4
3
2
1
0
B
4
3
2
1
0
C
d
A
B
C

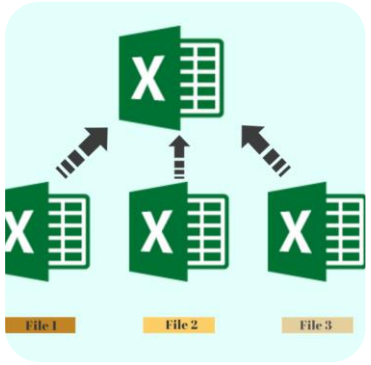
METODOLOGÍAS DE VISUALIZACIÓN

Realidad actual de las empresas

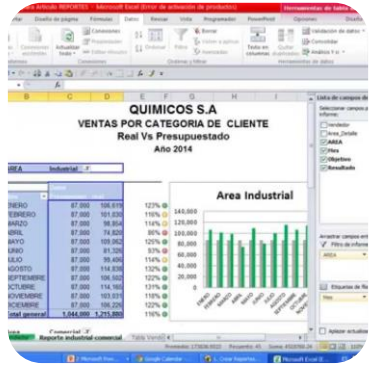
- Silos de datos
- Inexactitud e inconsistencia
- No auditable



Fuentes de datos



Análisis

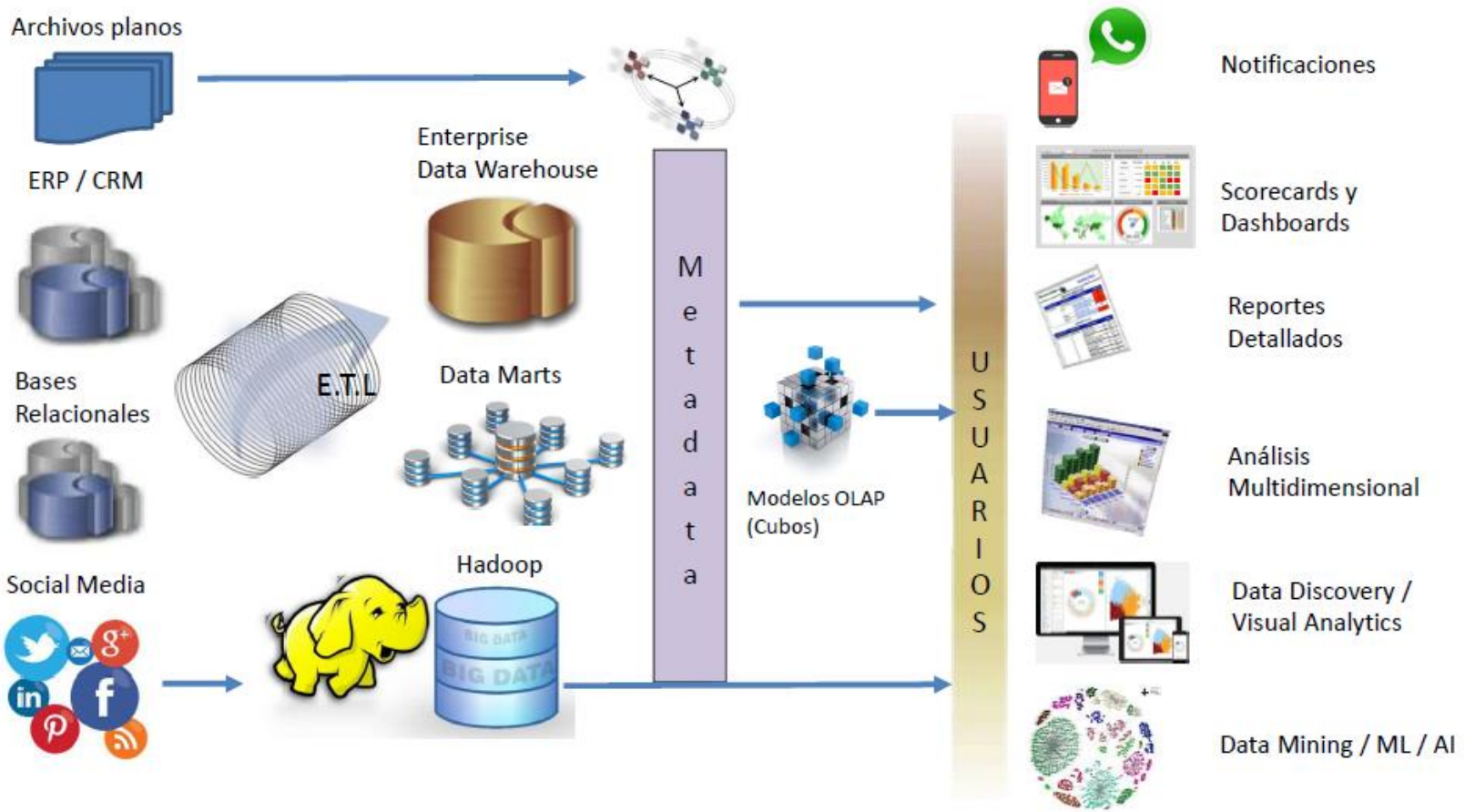


Presentación

¿Ventaja competitiva?

- No integridad de datos
- Conocimiento traducido en fórmulas
- Baja eficiencia y efectividad

Arquitectura soluciones de BI

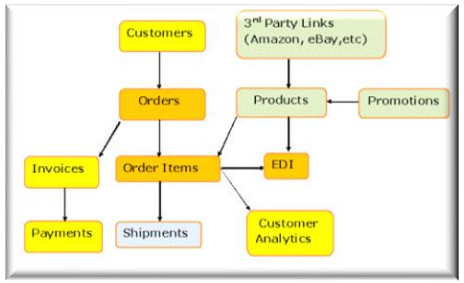


Metodología de implementación

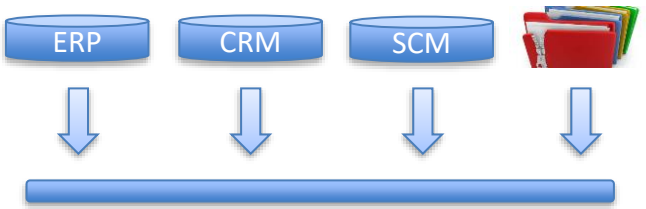
1. Requerimientos del negocio

- ¿Cuál es nuestra participación?
- ¿Somos rentables?
- ¿Vamos creciendo?
- ¿Qué, quién, cuándo, dónde, cuánto vendo?

2. Modelo Conceptual



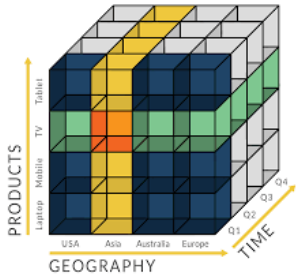
3. Fuentes operacionales



4. Proceso de ETL

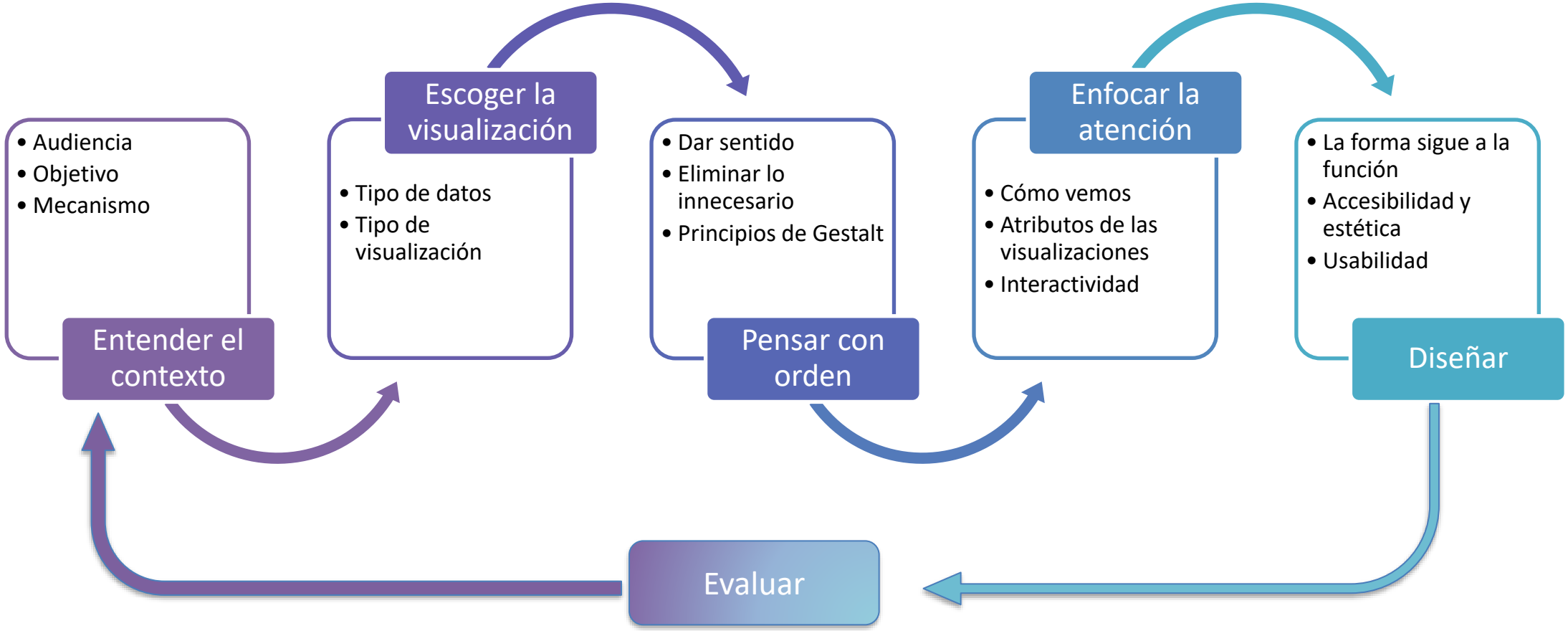
5. Data Warehouse

6. Modelos analíticos



6. Solución de BI

Metodología para visualizar información



Cole Nussbaumer Knaflic - Storytelling with Data

Contexto: Explorar vs Explicar

Explorar

- Entender los datos.
- Descubrir información importante o interesante para uno o para otros.

¡Buscar las perlas!

Explicar

- Comunicar nuestro análisis
- Explicar los puntos importante o la historia específica

¡Mostrar las perlas!

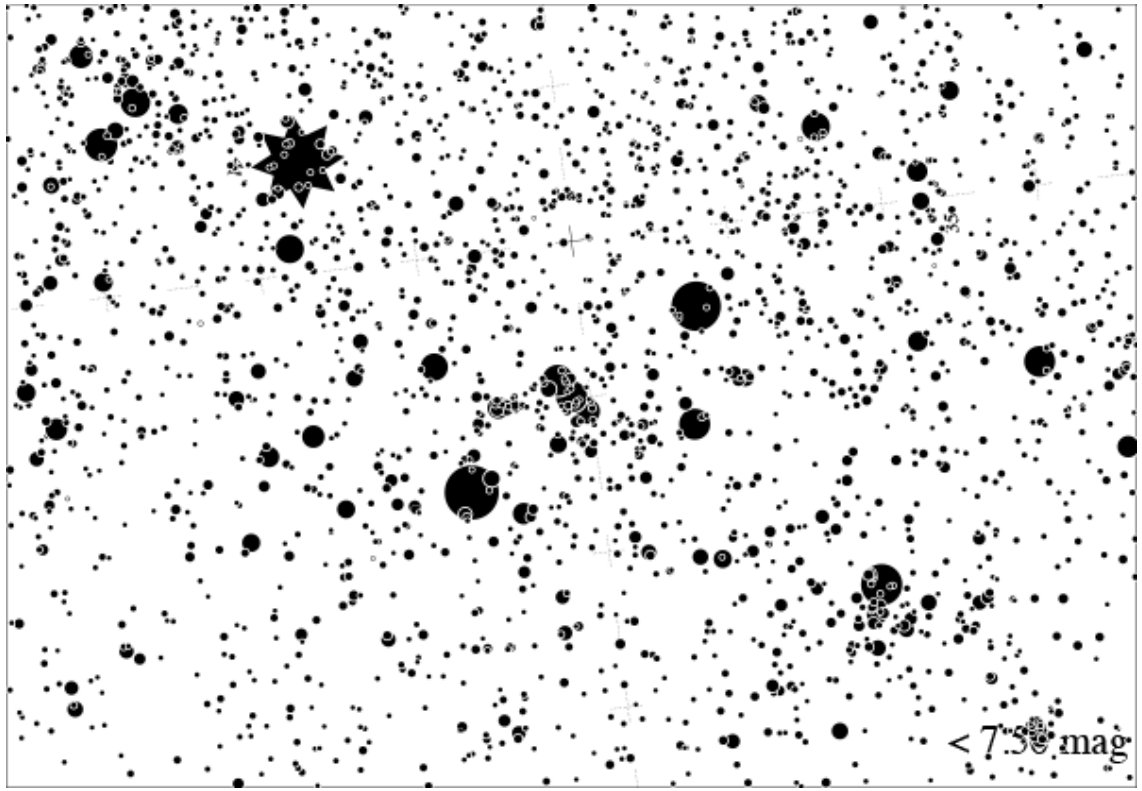
Análisis Exploratorio

- “Ver” lo datos con el máximo detalle
- Identificar las estructuras básicas
- Seleccionar las variables más importantes
- Detectar las desviaciones y anomalías
- Probar las hipótesis básicas

Contexto

- Cambia la perspectiva de un conjunto de datos
- Ayuda a decidir: qué representan los números y cómo interpretarlos

Views of the sky #2 (moebio.com)



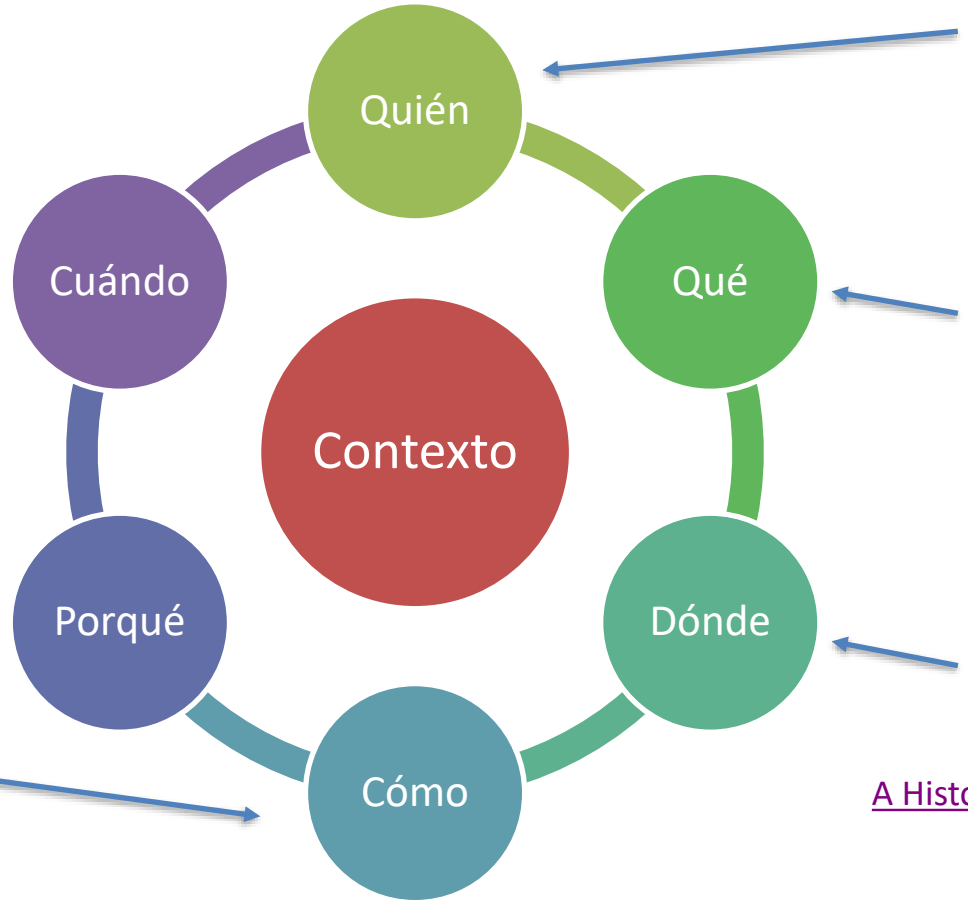
Globe at Night - Magnitude Charts for Orion

Entendiendo el contexto

Escala y ubicación en el tiempo

Razón por la que se recolectaron los datos

Metodología de recolección
Calidad de datos
Margen de error



Fuente de los datos
A quién representan

Lo que rodea a los números:
¿Qué representan?
¿Tienen sentido?
¿Cómo se relacionan con otros datos?

Ubicación espacial / digital.
De dónde provienen los datos

A History of the World in 100 Seconds on Vimeo

Usos del análisis exploratorio

- Encontrar errores y anomalías
- Obtener nuevos conocimientos sobre los datos
- Detectar valores atípicos en los datos
- Probar supuestos
- Identificar factores importantes en los datos
- Entender relaciones

Anomalías

- Es una observación de un fenómeno que es tan diferente de las demás observaciones del mismo, que lleva a pensar que fue generado por un mecanismo diferente.
- Los objetos de datos que son diferentes o inconsistentes con el conjunto restante se llaman valores atípicos o *outliers.*
- Los valores atípicos pueden ser causados por errores de medición o ejecución, o representan algún tipo de actividad fraudulenta.
- En ocasiones, un valor atípico puede ofrecer información interesante.
- Durante el análisis de datos una de las tareas consiste en detectar el valor anómalo e identificar el mecanismo que creó dicho valor.

Anomalías

Formas de detectar anomalías

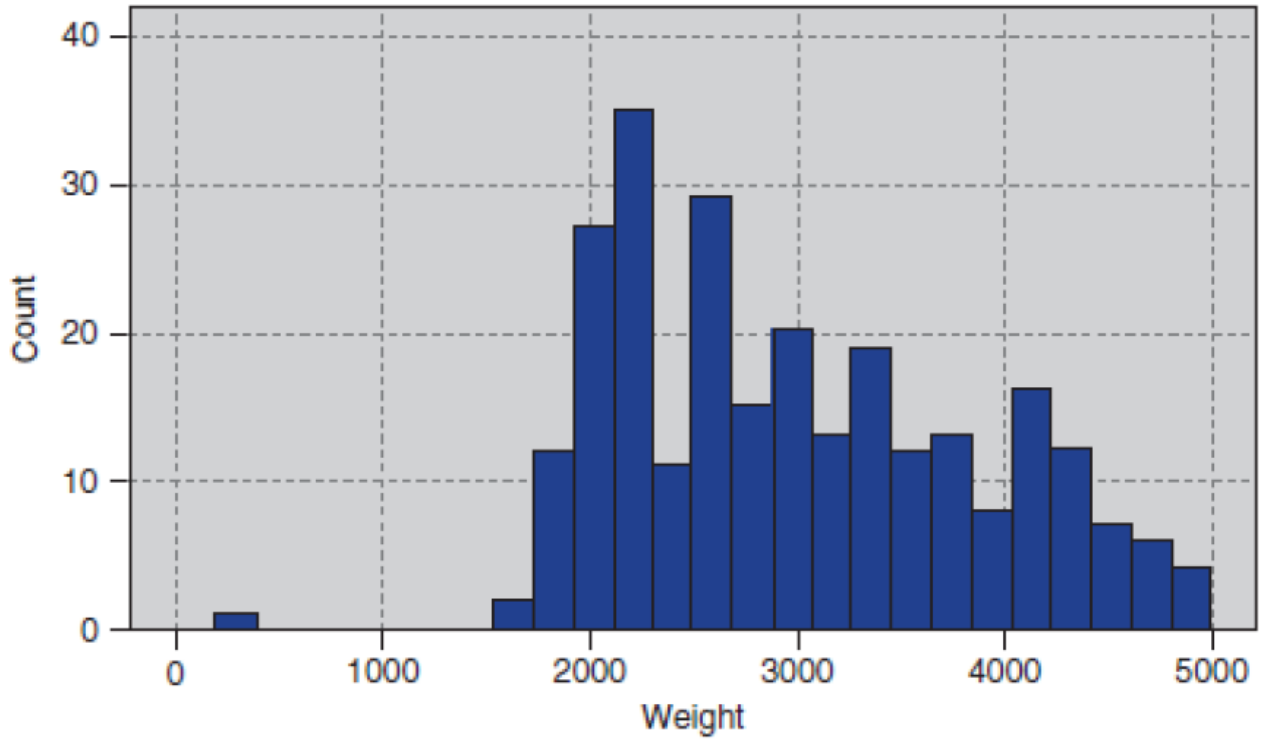
- De lo normal a lo anómalo:
  - Construir un modelo de las observaciones normales o de la población general.
  - Todo ejemplo que se desvíe es una anomalía.
- Construir un modelo de lo que se considera anómalo

Técnicas

- Con herramientas visuales.
- Con técnicas estadísticas.
- Con técnicas basadas en distancias.
- Con técnicas basadas en densidad.

Anomalías

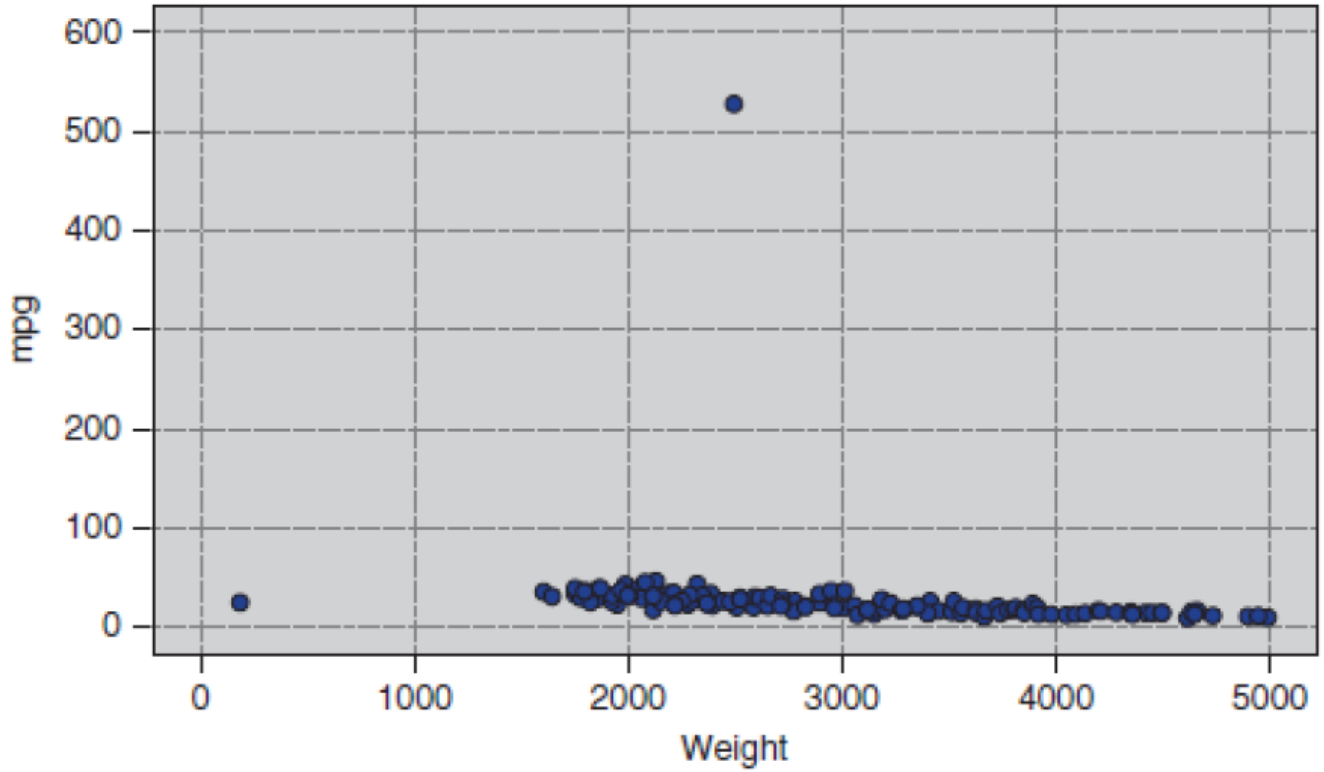
- Con herramientas visuales



Histograma del peso de unos vehículos

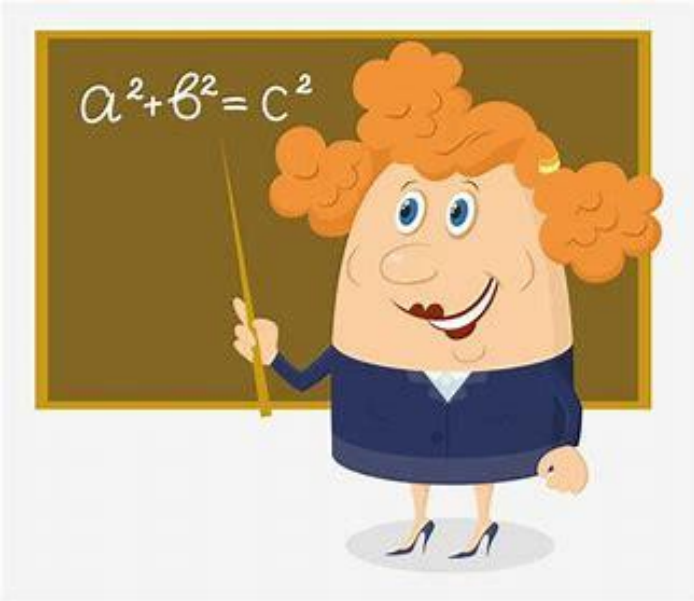
Anomalías

- Con herramientas visuales



Scatter de consumo MillasPorGalón según el peso

Análisis Explicativo



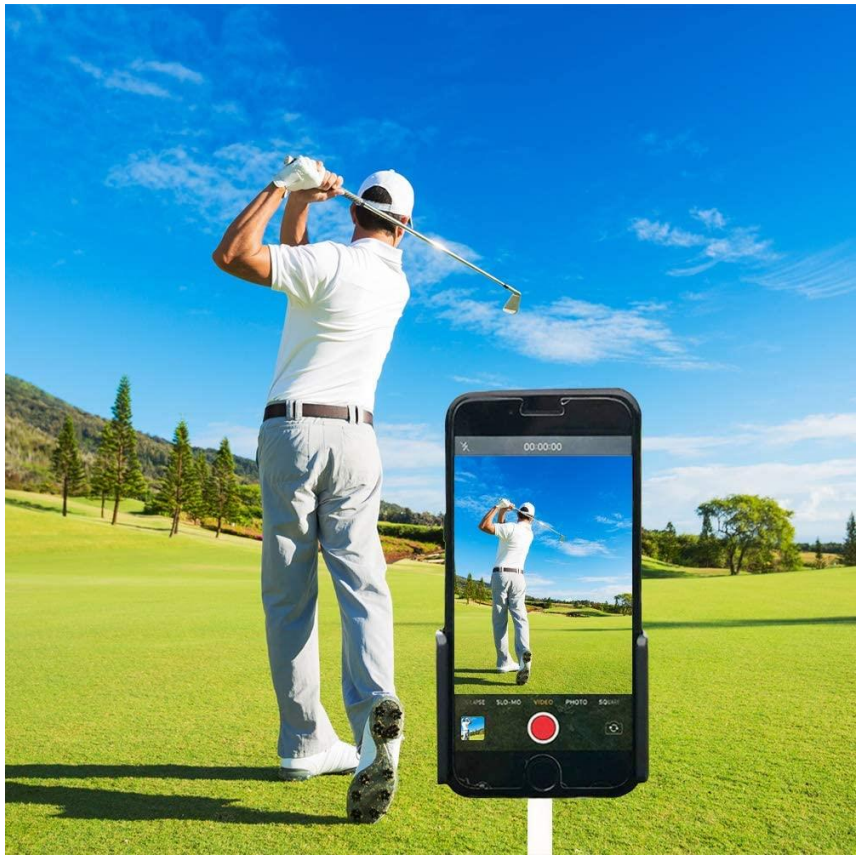
Considera:

- A quién le comunico
  - Audiencia
  - Como me perciben
- Qué quiero hacerles saber
  - Como quiero que respondan
  - Tono del mensaje
- Cómo uso los datos para fortalecer mi punto

Quién

- Audiencia
  - La comunicación cambia según la audiencia
  - Identifique al tomador de decisiones
- Relación con la audiencia
  - Cómo espera que lo perciban
  - Hay una relación
  - Confían o es necesario construir la credibilidad
  - Afecta el orden y flujo de la historia

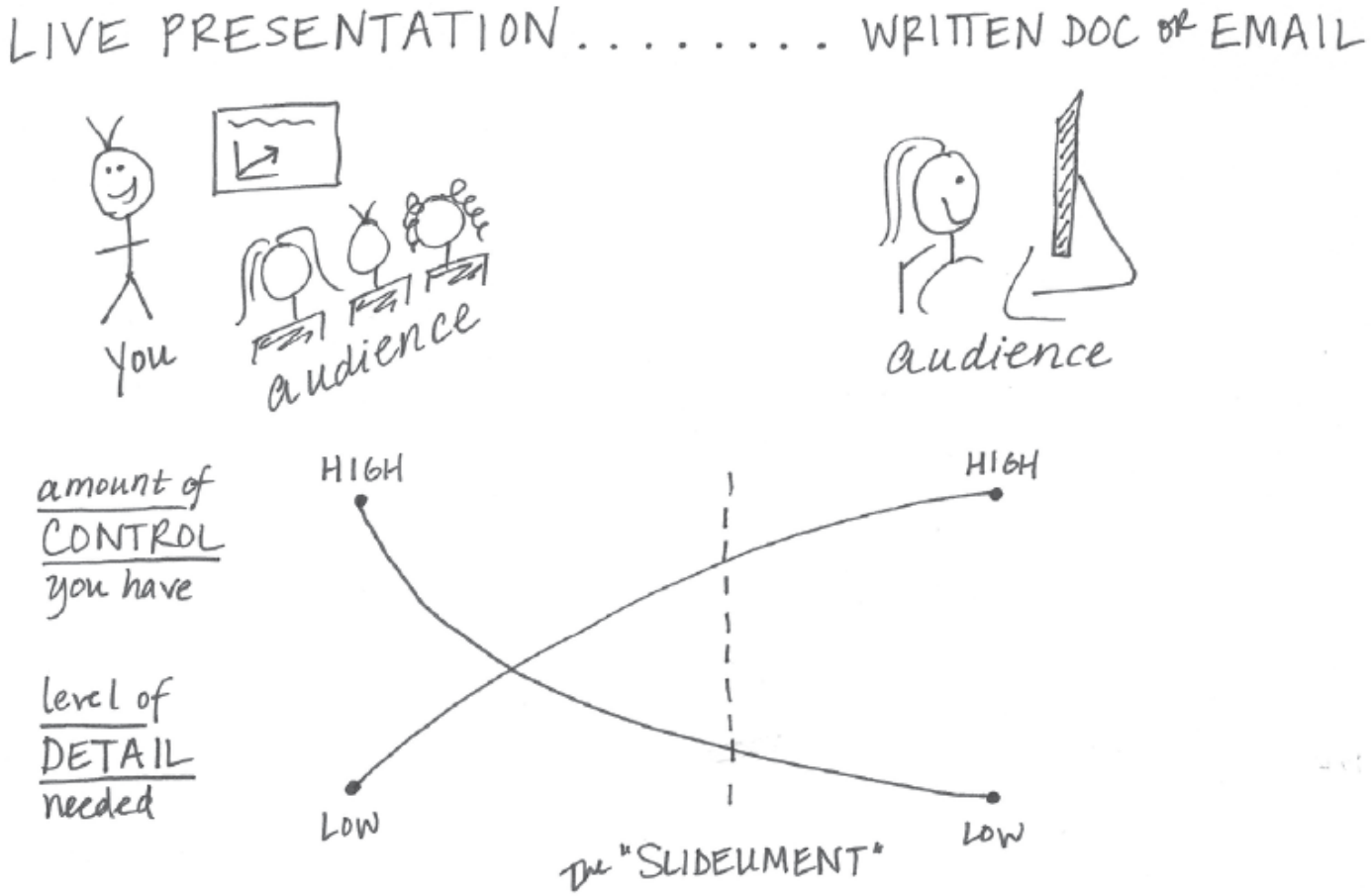
Qué



- Qué debe saber o hacer la audiencia
  - Porque les debe importar lo que digo
  - Sugiera pasos a seguir
- Cómo hacer relevante y clara mi información
  - Mecanismo y tono de la comunicación

¡Usted es el experto en el asunto. Lo analizó!

Qué – Mecanismo de comunicación



Cole Nussbaumer Knaflic - Storytelling with Data

Cómo

- ¿Qué datos tengo disponibles?
    - Evidencia
    - Mostrar tanto los datos que la soportan como los que no
    - Muestre el contexto

Ejercicio - Especifique el Quién, Qué y Cómo

Usted es un profesor de ciencias de cuarto año. Acaba de concluir un programa piloto vacacional sobre ciencias, que tenía como objetivo trabajar con los niños un tema poco popular.

Realizó unas encuestas con los niños al principio y al final del programa para comprender si cambiaron las percepciones hacia la ciencia y cómo lo hicieron.

Cree que los datos muestran una gran historia de éxito y le gustaría continuar ofreciendo el programa vacacional sobre ciencias en el futuro.

PREPARAR LOS DATOS

¿Por qué es necesario preparar los datos?

- Los datos originales pueden tener diferentes problemas
  - Incompletos : valores vacíos o datos resumidos

    Diagnóstico =NA, Venta =50.000
  - Con errores : valores que no coinciden con el dominio de la empresa, registros con valores atípicos

    Salario = “-10”
  - Inconsistentes : Discrepancias entre atributos o entre fuentes de datos

    Edad =“42”, FechaNacimiento =“03/07/

    Calificación “1,2,3”, Calificación “A, B, C”

    Dirección : Cra 7 No 40 - 62 y en otra fuente Cra 7 No 45 - 39

¿Por qué es necesario preparar los datos?

¿Qué problemas pueden encontrar en este DataSet?

Customer ID	Zip	Gender	Income	Age	Marital Status	Transaction Amount
1001	10048	M	75,000	C	M	5000
1002	J2S7K7	F	–40,000	40	W	4000
1003	90210		10,000,000	45	S	7000
1004	6269	M	50,000	0	S	1000
1005	55101	F	99,999	30	D	3000

EXPLORACIÓN DE DATOS

Exploración inicial

- Realizar ingeniería inversa para perfilamiento de datos
- Identificar variación y calidad de cada atributo de la fuente de datos

- ¿Para atributos numéricos qué podemos explorar?
    - Min, max , media, desviación, moda
    - Distribución
    - Atípicos
    - Faltantes

- ¿Para atributos categóricos qué podemos explorar?
    - Valores
    - Frecuencia de cada valor, moda
    - Faltantes

- Relaciones entre atributos
    - Tabulaciones cruzadas: análisis de tablas de contingencia.
    - Correlaciones

Exploración inicial

Tabla resumen de todas las variables de mi set de datos

Atributo	Tabla	Tipo de dato Almacenamiento	Tipo de dato conceptual	# nulos	# distintos	Media	Desviación Estandar	Moda	Min	Max	Valores
Edad	Cliente	Integer	Numérico	10	40	38	12	34	18	87	-
Sexo	Cliente	String	Nominal	0	3	-	-	F	-	-	F,M
VIP	Cliente	Integer	Nominal	0	2	-	-	0	-	-	0,1

Exploración de datos

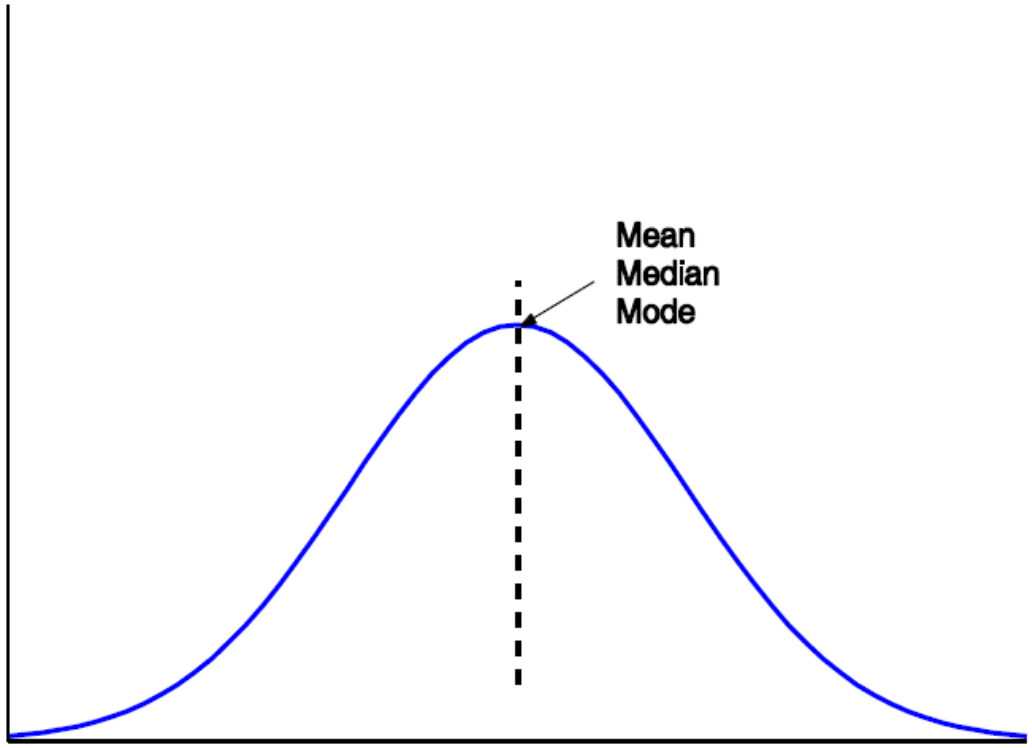
- Objetivo
  - Entender mejor : tendencia central de los datos, variación y dispersión
- Características de la dispersión de los datos
  - media, max, min, cuantiles, atípicos, varianza, …
- Atributos numéricos que corresponden con intervalos ordenados
  - Dispersión de los datos : analizados con múltiples granularidades de precisión
  - Análisis Boxplot o cuantiles en intervalos ordenados
- Análisis de nominales
  - Frecuencias, distribución

Medir la tendencia central

- Media – Promedio (medida algebraica)
- Mediana
- Moda

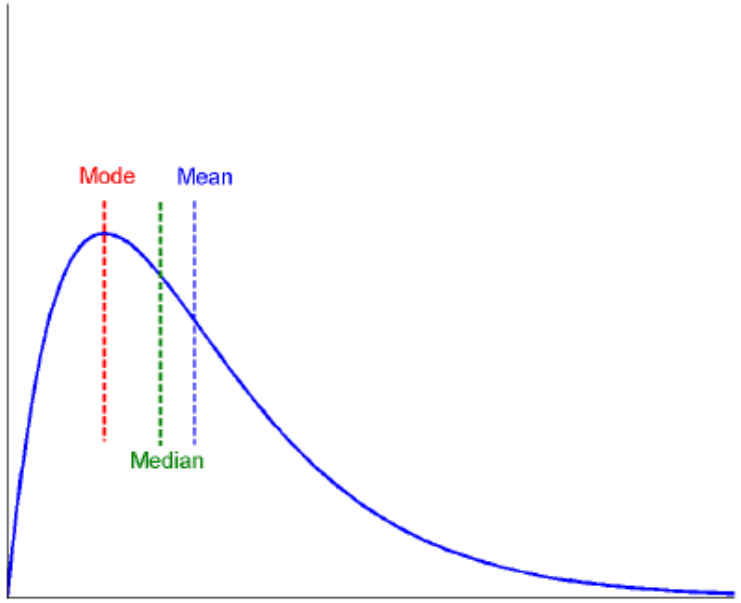
Datos Simétricos Vs. Datos Asimétricos

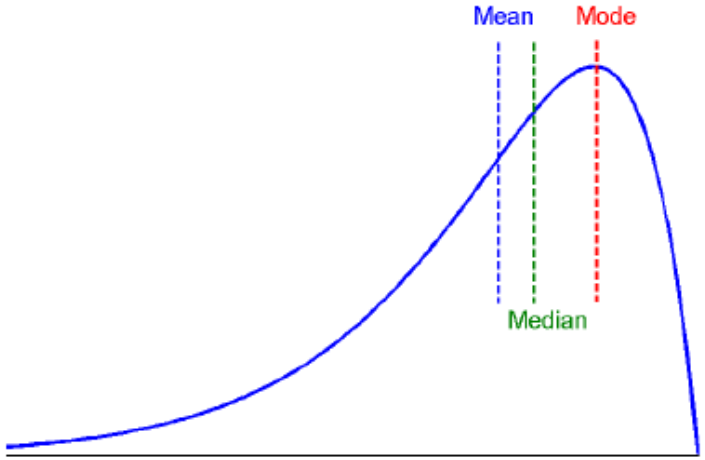
- La media, mediana y moda de datos simétricos, coinciden

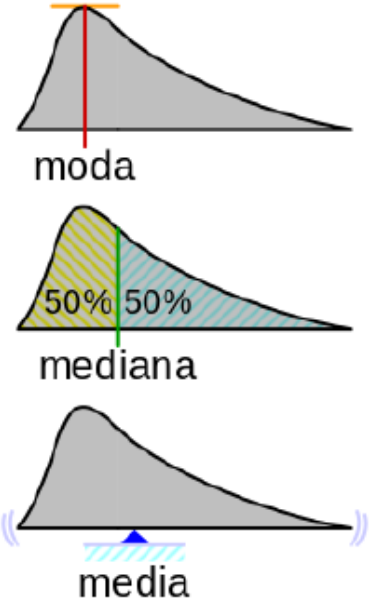


Datos Simétricos Vs. Datos Asimétricos

- La media, mediana y moda de datos asimétricos
    - Asimétricos positivos (sesgo positivo), asimétricos negativos (sesgo negativo)







Dispersión de los datos

- Varianza: define que tan lejos están los valores de la media
  - Su unidad de medida corresponde al cuadrado de la unidad de medida de la variable.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \bar{x}^2$$

- Desviación estándar: qué tan alejados están los valores de la media.
  - La desviación está en las unidades de la variable a la cual se le está midiendo la dispersión.

AMPLIANDO EL CONOCIMIENTO

Lecturas

- “Visualization Analysis and Design” de Tamara Munzner.
- “Storytelling with data”, Cole Nussbaumer
- “Data Points”, Nathan Yau
- “Visual Thinking for design”, Colin Ware
- “Fundamentals of data visualization”, Claus O. Wilke

Exploración de datos con Python – Automóviles

TALLER

Pontificia Universidad JAVERIANA Bogotá

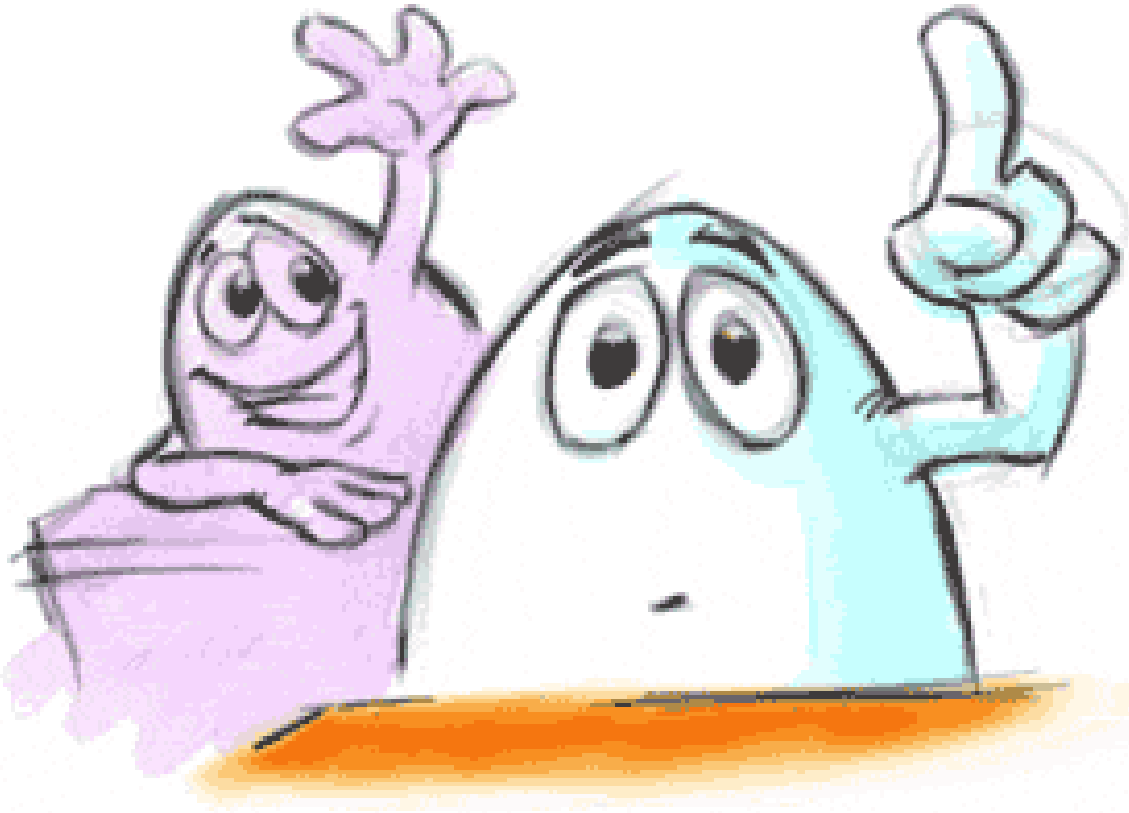
Exploración de datos con Python

1. Se requiere contar con un cuaderno de Jupyter, que se puede conseguir descargando el navegador de Anaconda de: Anaconda | Individual Edition o en línea en Cocalc
2. Descargue en su computador el archivo “Exploracion de datos.ipynb”
3. Abra anaconda y despliegue un JupyterLab para Python 3, o cree un proyecto en Cocalc
4. Cargue el archivo, lea con detalle el ejercicio y siga las instrucciones.

Nota: al usar Cocalc, se debe subir el archivo .data al sitio de Cocalc

Taller

- Utilizando el dataset proporcionado por el US National Center for Health Statistics (NCHS), realice en grupos una exploración de los datos, analizando los valores estadísticos y las correlaciones que puedan existir entre variables.
- Genere un reporte donde incluya todos los pasos realizados, los gráficos y sus conclusiones.



¿Preguntas?

Metodología para visualizar datos

Maria Isabel Serrano G.
maria-serrano@javeriana.edu.co