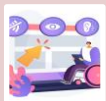


Metodología para visualizar datos

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

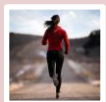
Agenda



Metodologías de visualización



Entender el contexto



Ejercicios



Taller

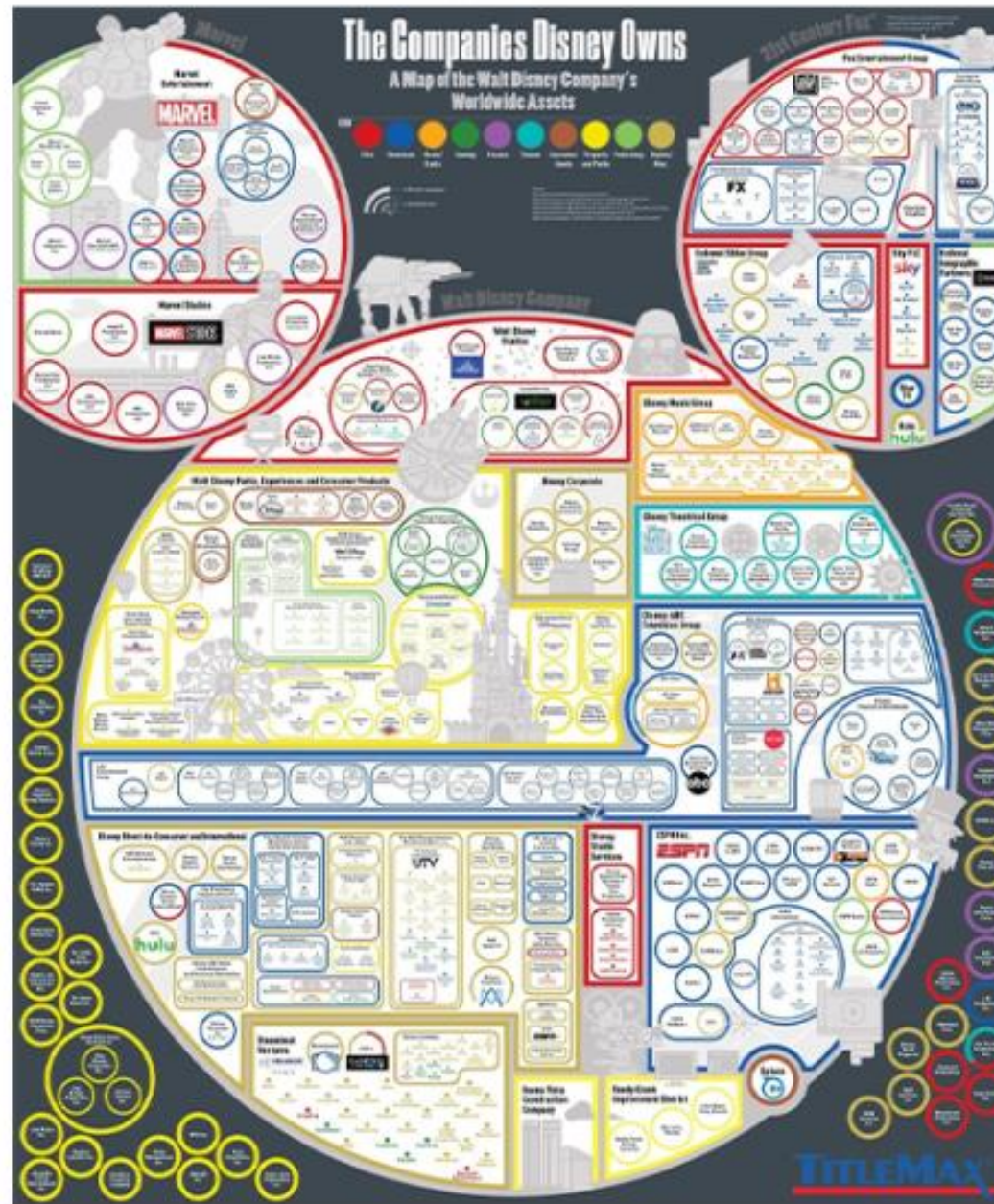
Sydney comes first with 541, some of the most prolific bats. Clarke, Steve Waugh, and Mark Waugh follow with an average of 44.80. This is not far from Boon and George Bailey.

Runs 5416
Avg 45.37

Run
Avg

Sydney

Lauren



ing to power

Suggests PM Modi
NDA to power again

50%

543 seats

fs Credit: My-Axis-India

DiU

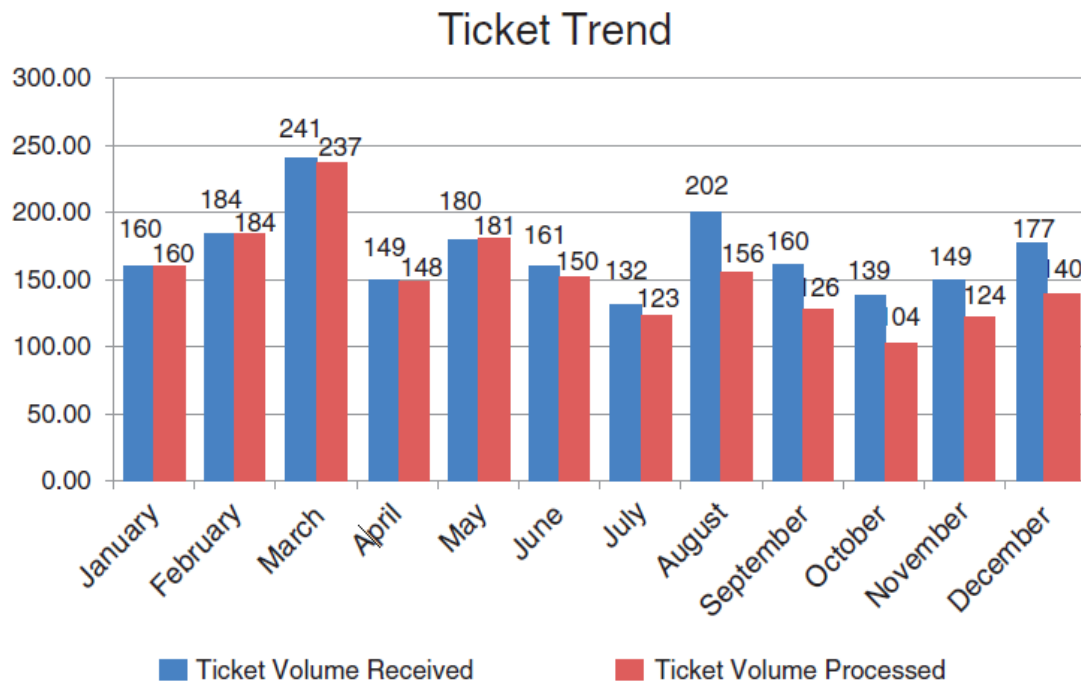
Visualización de

Casi todos podemos hacer un gráfico

**Pero lograr que el gráfico se vea bien,
o que cuente una historia ...**

¿Qué significa contar una historia?

Mostrar datos

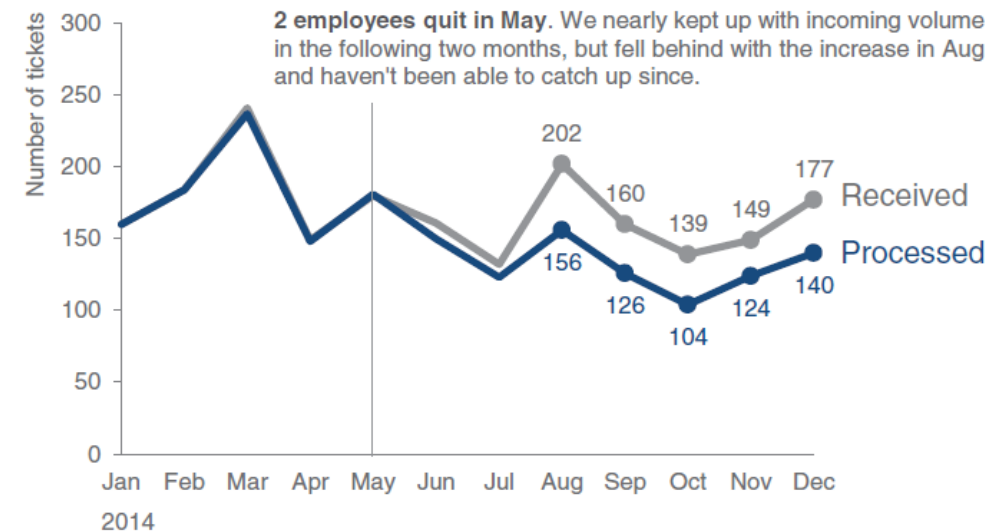


Contar una historia

Please approve the hire of 2 FTEs

to backfill those who quit in the past year

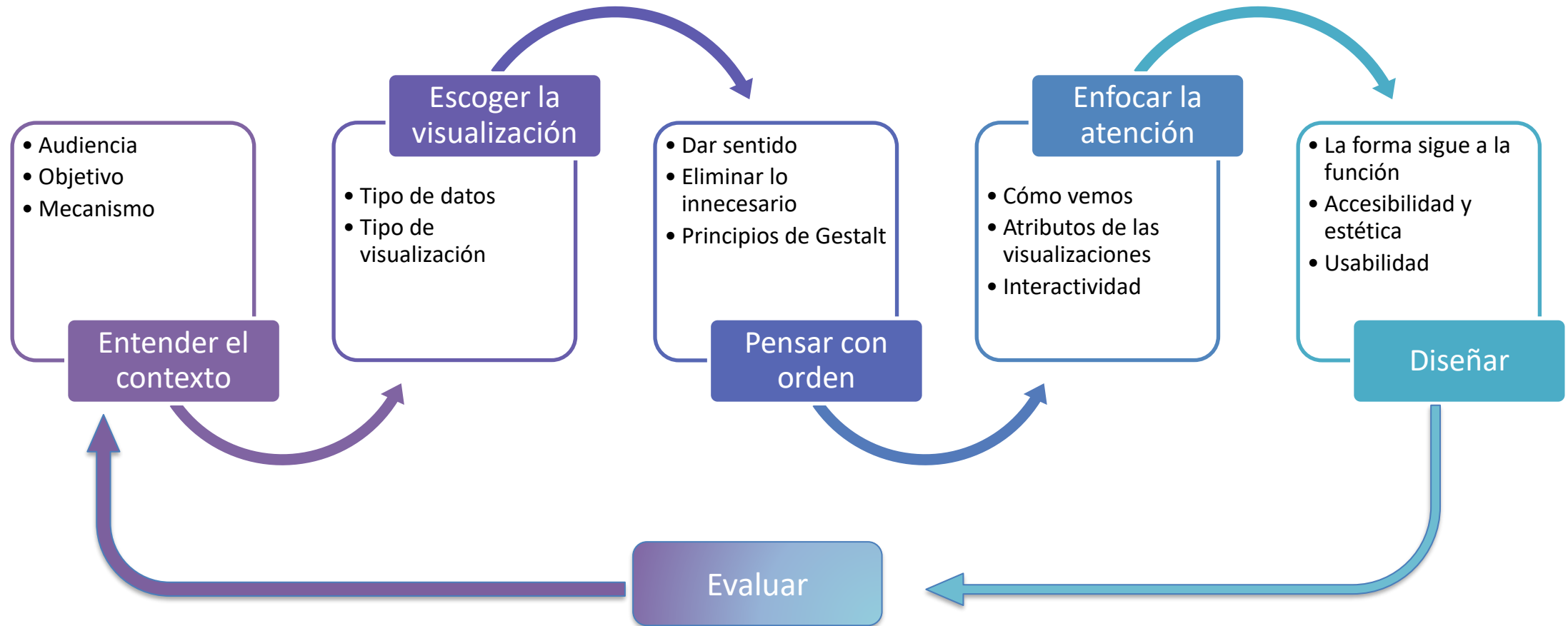
Ticket volume over time



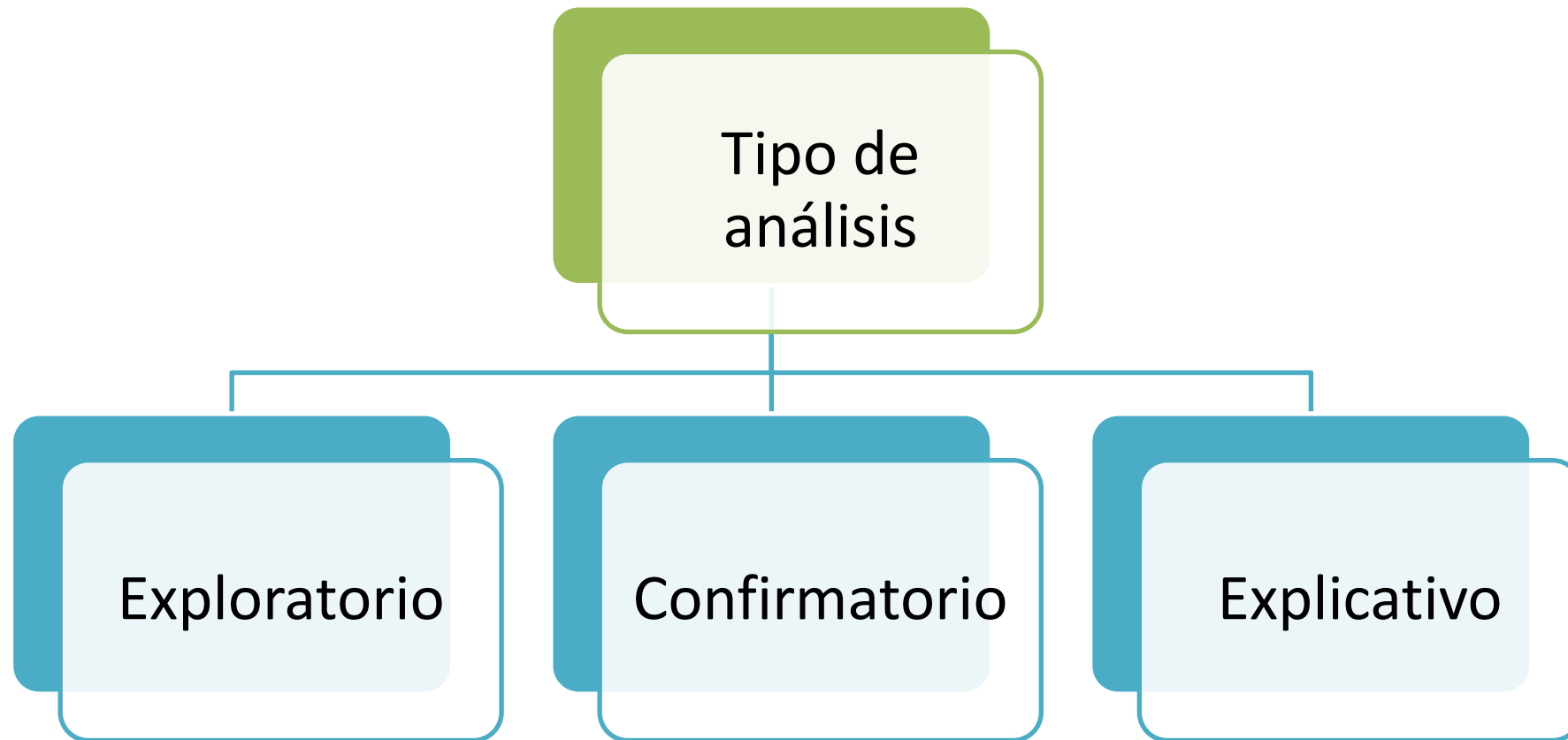
**El analista sabe encontrar datos,
reunirlos, analizarlos, armar un
modelo. Pero ...**

Se le dificulta comunicar su análisis

Metodología para visualizar información



1. Entiende el contexto



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

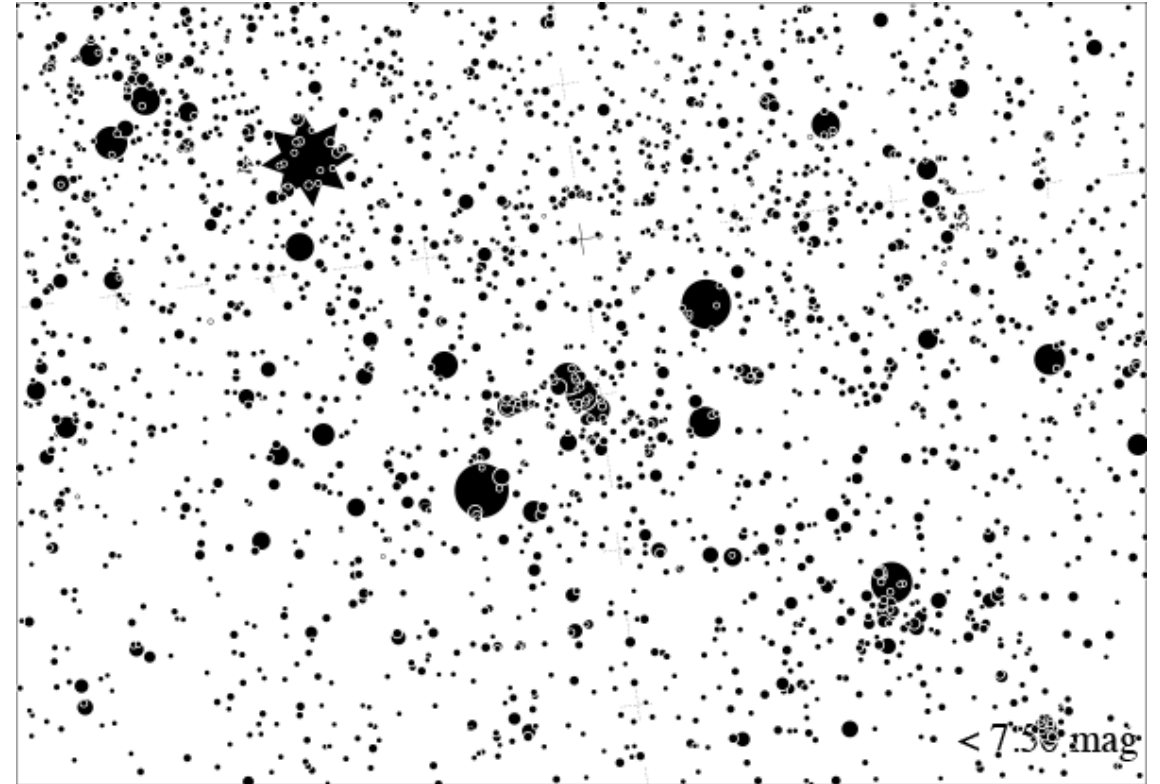


- “Ver” lo datos con el máximo detalle
- Identificar las estructuras básicas
- Seleccionar las variables más importantes
- Detectar las desviaciones y anomalías
- Probar las hipótesis básicas

Contexto

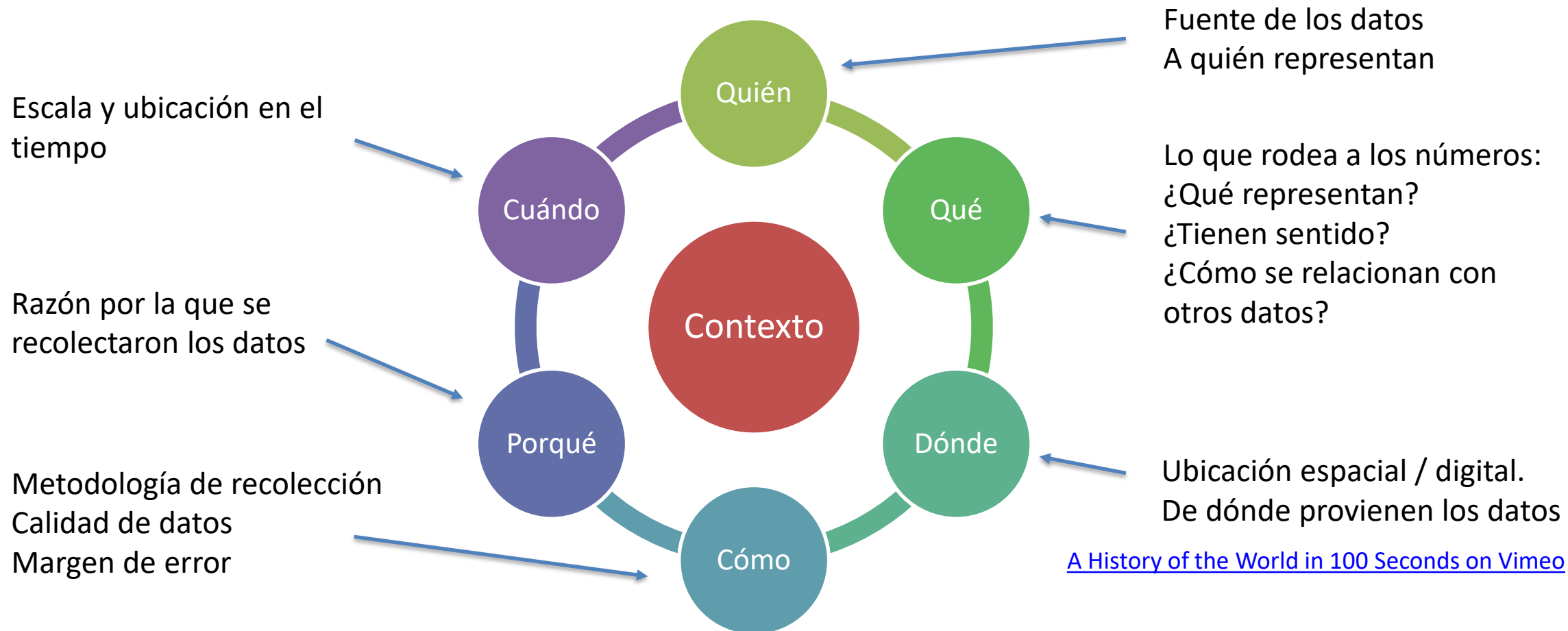
- Cambia la perspectiva de un conjunto de datos
- Ayuda a decidir: qué representan los números y cómo interpretarlos

[Views of the sky #2 \(moebio.com\)](http://moebio.com)



[Globe at Night - Magnitude Charts for Orion](#)

Entendiendo el contexto



Usos del análisis exploratorio

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



reallyreallyreallytrying

“average person eats 3 spiders a year” factoid
actually just statistical error. average person
eats 0 spiders per year. Spiders Georg, who
lives in cave & eats over 10,000 each day, is an
outlier and should not have been counted



dominateeye

World Heritage Post

250,698 notes



¿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

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

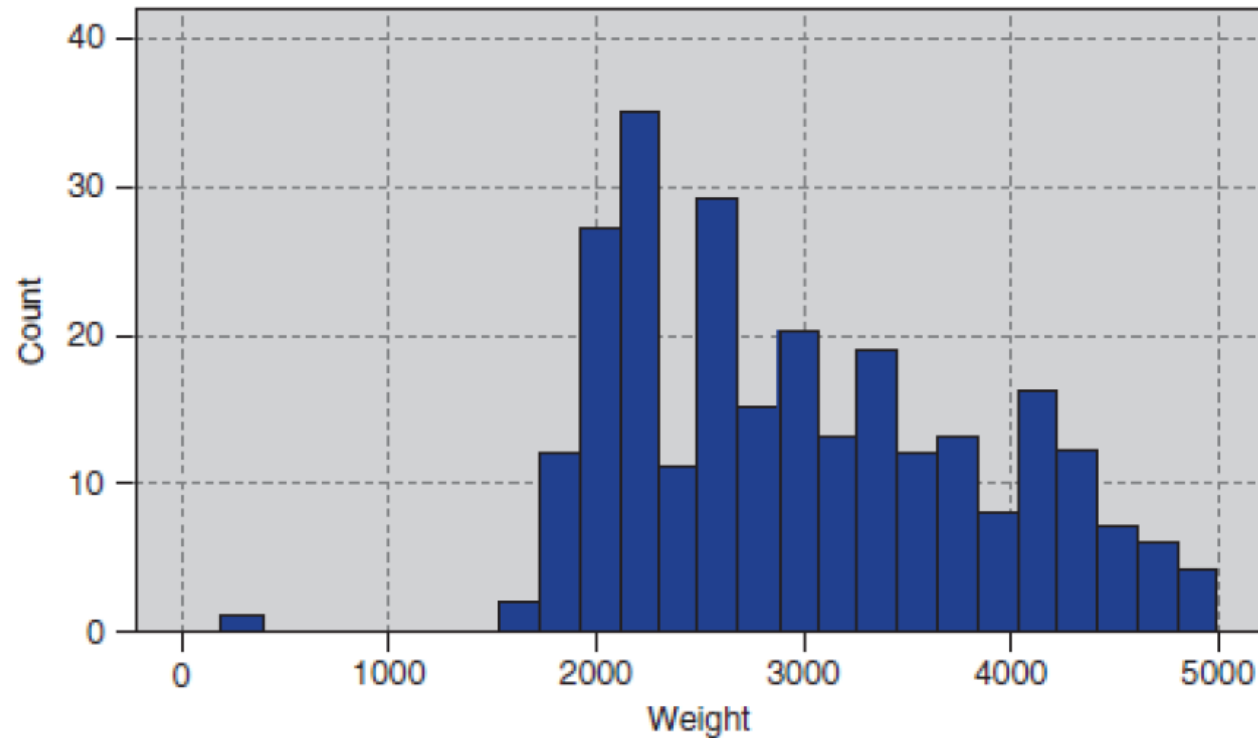
- De lo normal a lo anómalo:
 - Construir un modelo de las observaciones normales o de la población general.
 - Todo ejemplo que se desvíe es una anomalía.
- Construir un modelo de lo que se considera anómalo

Técnicas

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

Anomalías

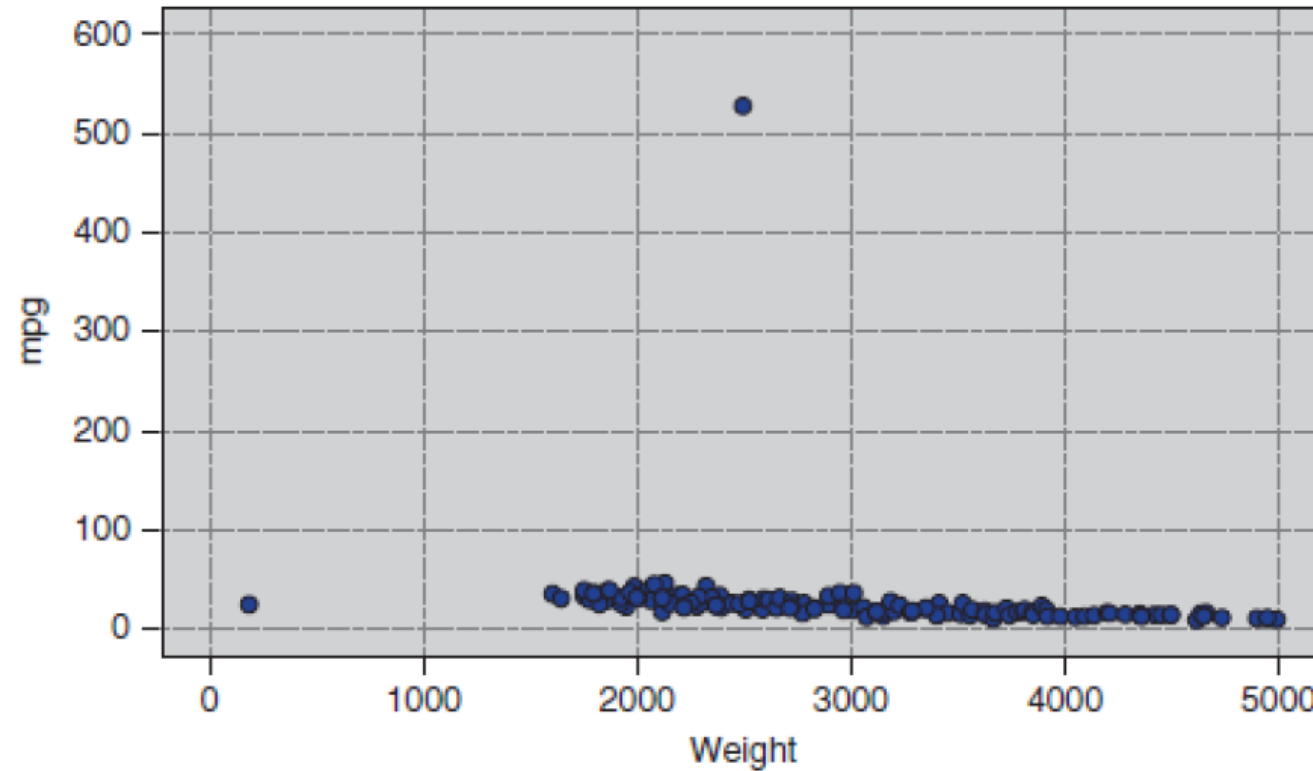
- Con herramientas visuales



Histograma
del peso de
unos
vehículos

Anomalías

- Con herramientas visuales



Scatter de
consumo
MillasPorGalón
según el peso

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

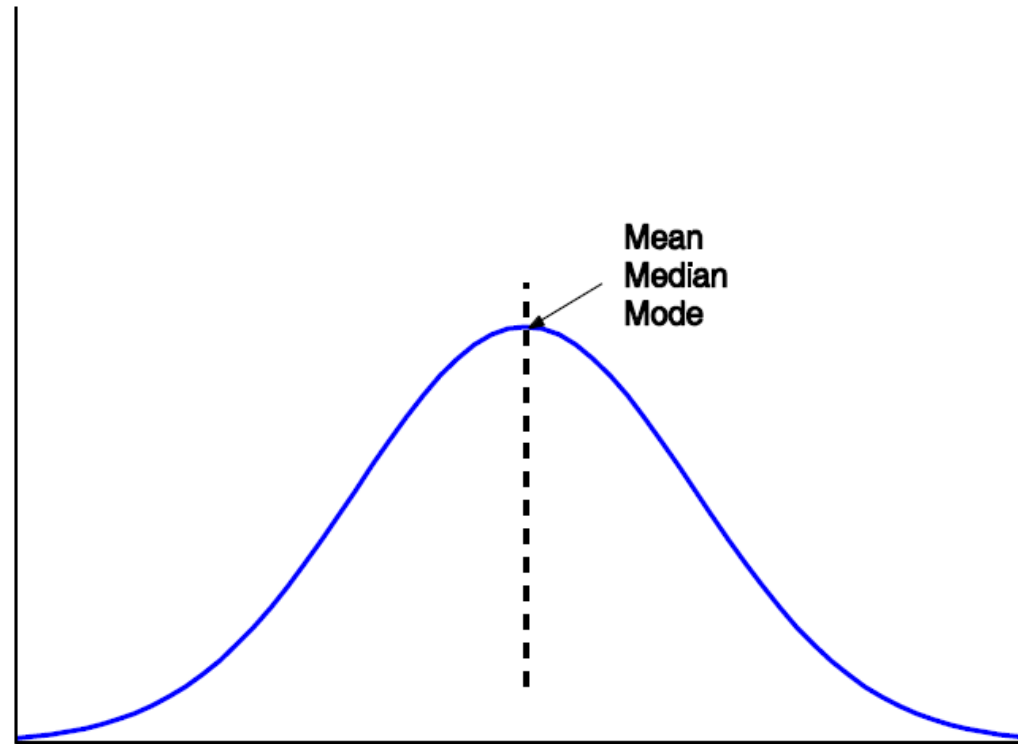
- Objetivo
 - Entender mejor : tendencia central de los datos, variación y dispersión
- Características de la dispersión de los datos
 - media, max, min, cuantiles, atípicos, varianza, ...
- Atributos numéricos que corresponden con intervalos ordenados
 - Dispersión de los datos : analizados con múltiples 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

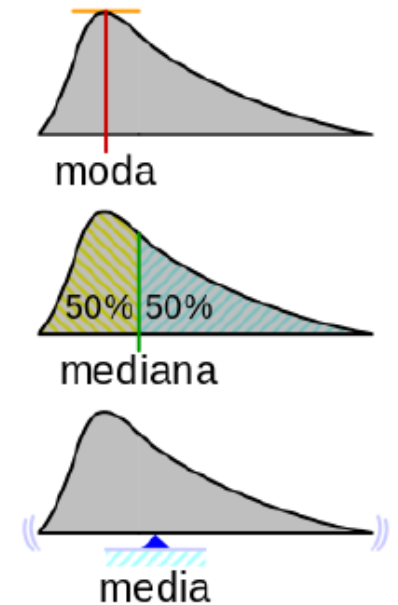
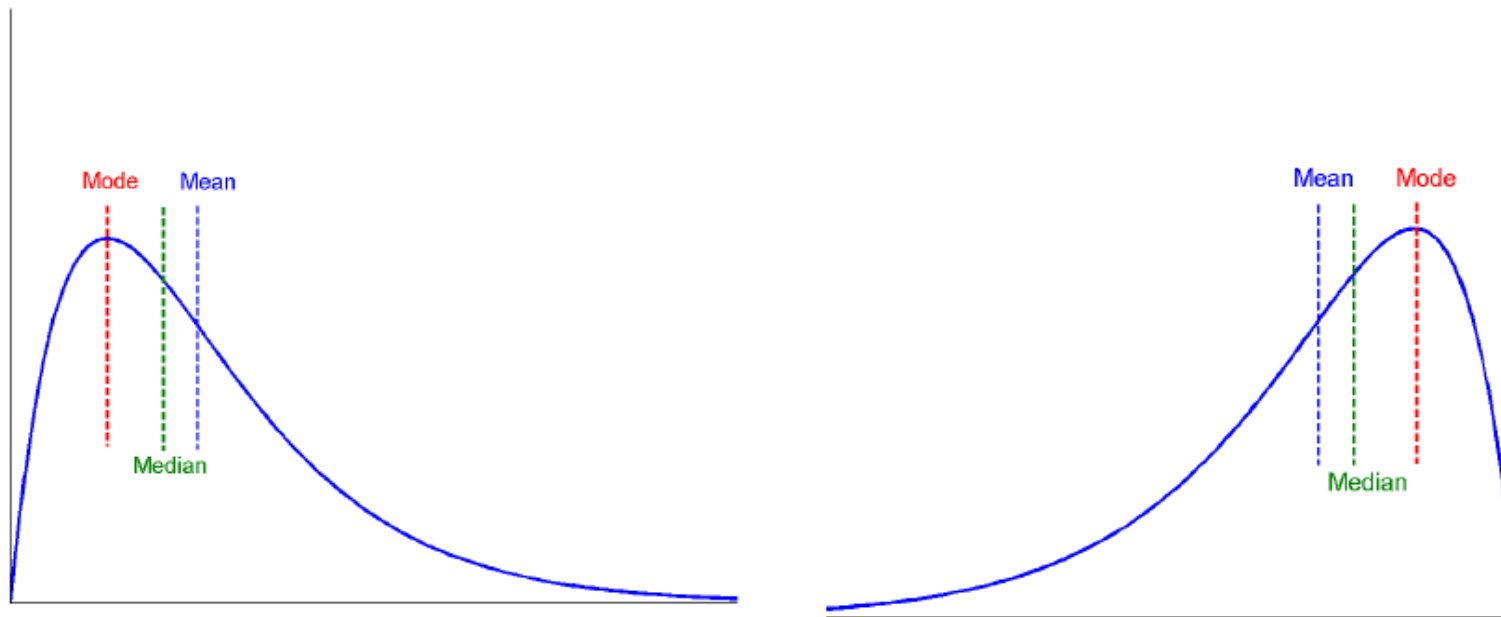
Datos Simétricos Vs. Datos Asimétricos

- 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.

Exploración de datos con Python – Automóviles

TALLER

Exploración de datos con Python

1. Se requiere contar con una cuenta de Gmail para ingresar en Google Collab
2. Abra el archivo “Exploracion de datos.ipynb”:
<https://drive.google.com/file/d/1ajOD6S7tFtkgzo6F7hLQy0VtBralg6b1/view?usp=sharing>
3. Lea con detalle el ejercicio y siga las instrucciones.

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

Ejemplo de visualizaciones exploratorias

Con dataset de colisiones de vehículos en NYC, tomado de [Motor Vehicle Collisions - Crashes - CKAN \(data.gov\)](#)

- ¿Cómo cambia el volumen o la frecuencia de heridos en el tiempo?

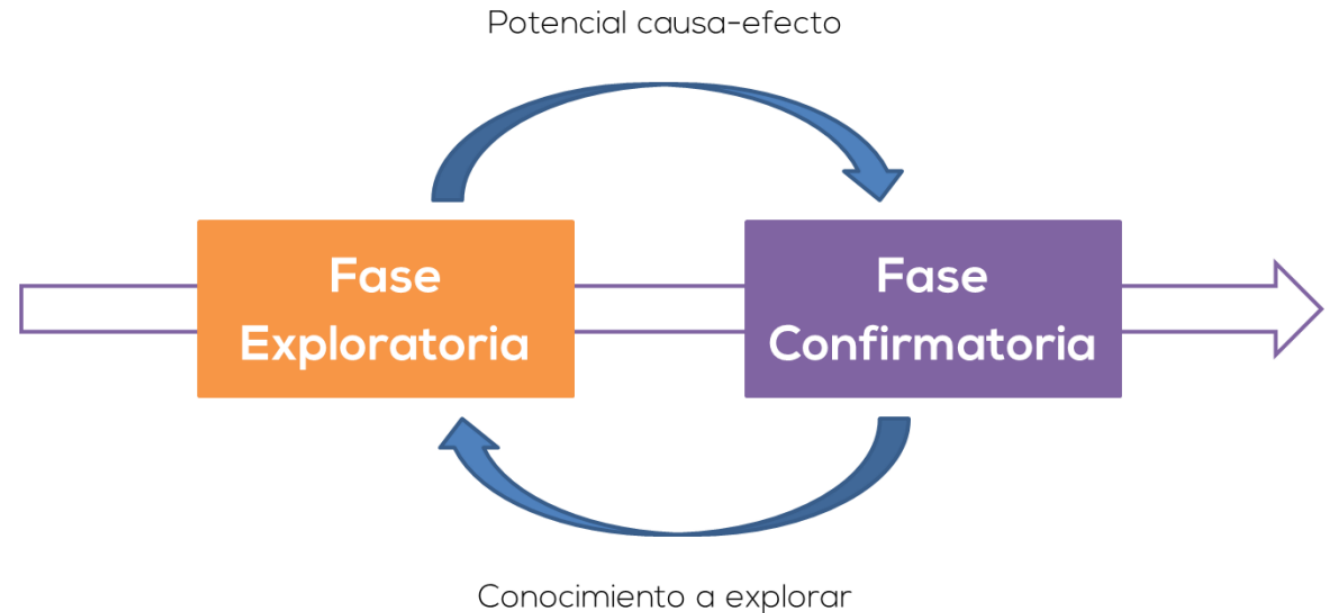
¡¡Tener preguntas es lo que da origen al análisis exploratorio!!

Usualmente la salida de un análisis exploratorio consiste en la generación de nuevas preguntas

Visualización confirmatoria

Visualización Confirmatoria

- Validación de las hipótesis
- Validación de datos
- Validación de gráficos
- Validación estadística
- Simetría de datos



Ejemplo de análisis confirmatorio

Viendo los datos de las colisiones, se observa un patrón en las colisiones.

¿Será que este patrón se repite para diferentes tipos de heridos?
(Ciclistas, motociclistas, peatones y pasajeros)

Pensaría que es más pronunciado para ciclistas y motociclistas

El análisis confirmatorio parte de una hipótesis que se desea verificar

Taller

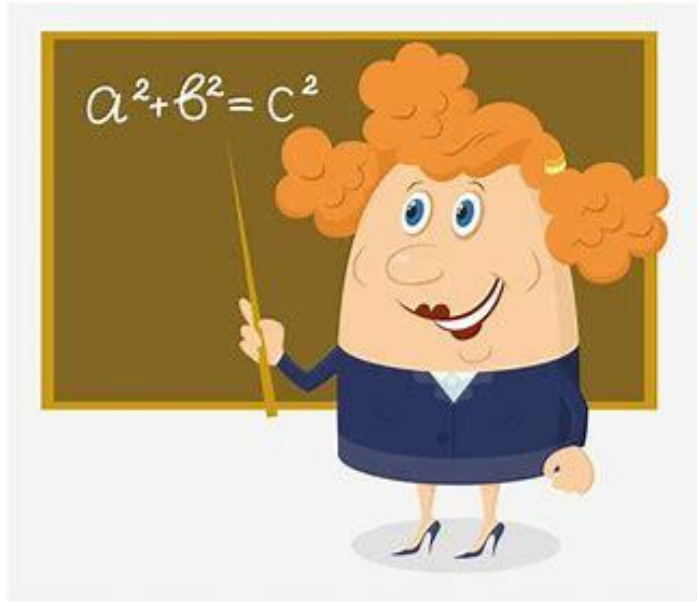
- Utilizando el dataset proporcionado por el US National Center for Health Statistics (NCHS), realice en grupos una exploración de los datos, analizando los valores estadísticos y las correlaciones que puedan existir entre variables.
- Genere un reporte donde incluya todos los pasos realizados, los gráficos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

Análisis Explicativo

Ejemplos de Visualizaciones Explicativas

1. Visualización interactiva de mapa: “[Mapping the Shadows of New York City: Every Building, Every Block](#)” NYT
2. “[Good, Evil, Ugly, Beautiful: Help Us Make a ‘Game of Thrones’ Chart](#)” NYT
3. “[What Happens Next?](#)” Covid-19 por Nicky Case
4. “[Reconstructing seven days of protests in Minneapolis after George Floyd’s death](#)” The Washington Post y Poligraph
5. Calentamiento global explicado por la NASA: [Global Warming \(nasa.gov\)](#)

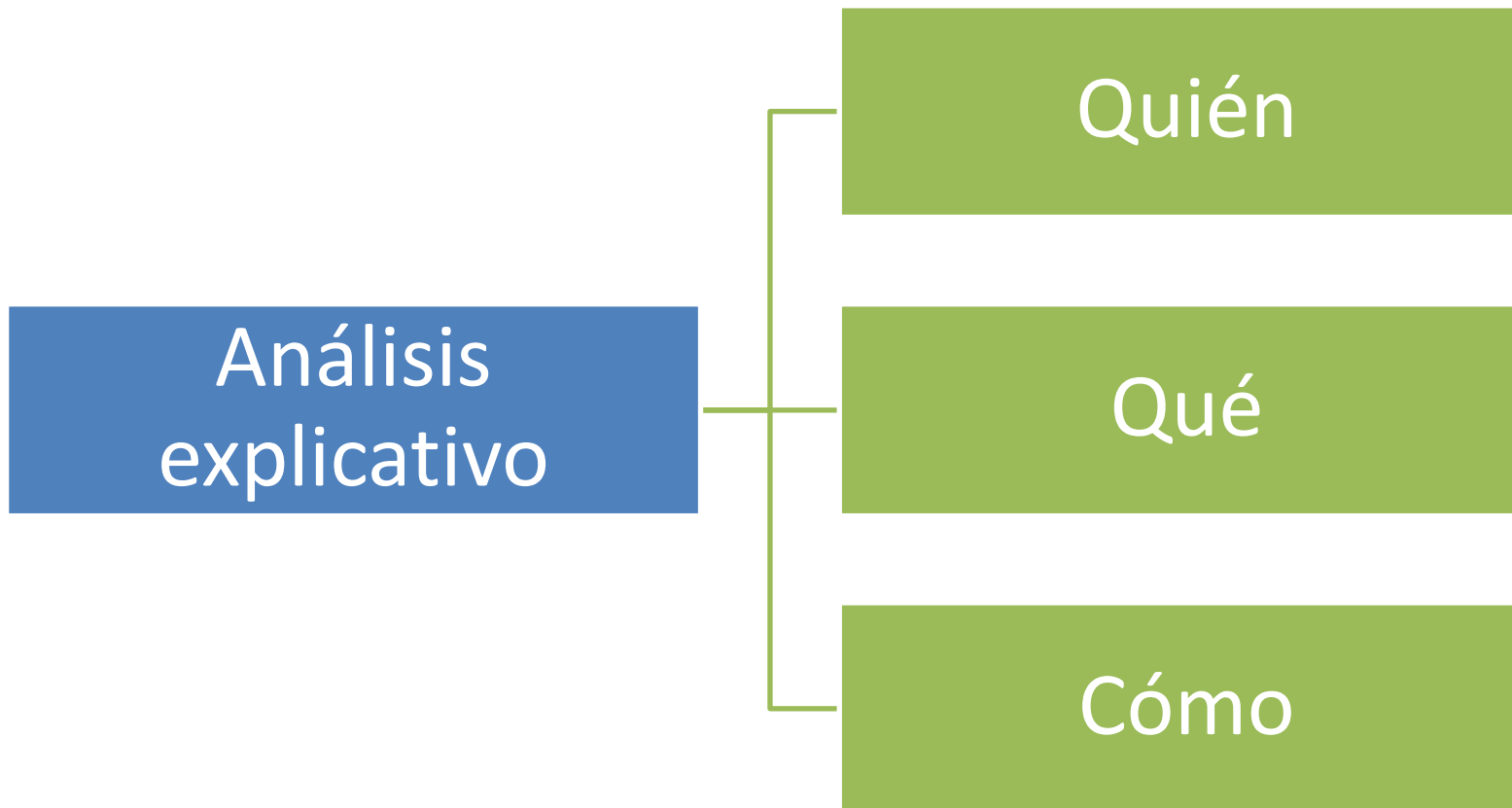
Análisis Explicativo



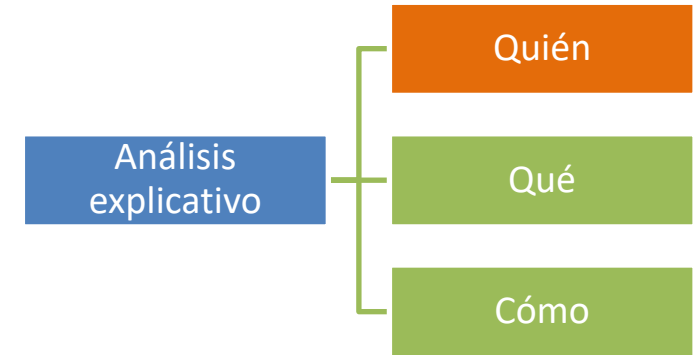
Considera:

- 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

Análisis Explicativo

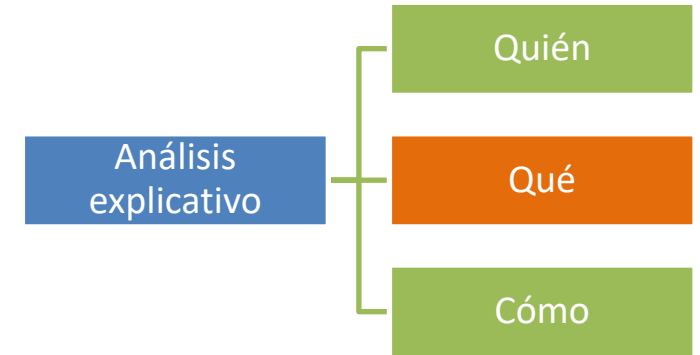


1. Entiende el contexto - Quién

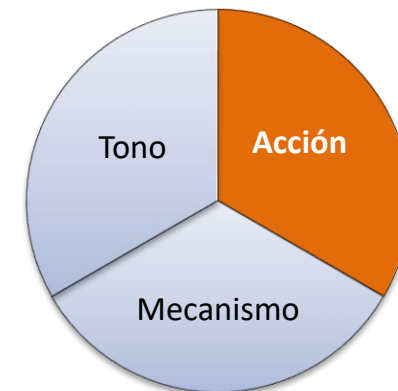


- ¿Quién es su audiencia?
 - ☐ Mejor entre más específica sea
 - ☐ Evite audiencias genéricas: “el que este interesado”
 - ☐ Se pierde efectividad
 - ☐ Identifique el tomador de decisiones
- ¿Quién es usted para su audiencia?
 - ☐ ¿Qué relación tiene con la audiencia?
 - ☐ ¿Cómo lo ven a usted?: ¿lo conocen?, ¿le creen?
 - ☐ Define el orden y la estructura de la historia

1. Entiende el contexto - Qué



1. Entiende el contexto - Qué



- Acción: ¿qué necesita hacerles saber?
 - ☐ Planee cómo hacer la comunicación relevante
 - ☐ Porqué les debería importar
 - ☐ Siempre debería esperar que su audiencia sepa o haga algo, de lo contrario la comunicación no tiene sentido
 - ☐ Recuerde, usted sabe más que la audiencia:
 - ☐ Usted analizó e interpretó los datos
 - ☐ Ellos no los conocen
 - ☐ Debe ayudarles a entenderlos y tomar la mejor decisión.
 - ☐ Haga observaciones y proponga recomendaciones o pasos a seguir.

1. Entiende el contexto – Qué – Acción

Prompting action

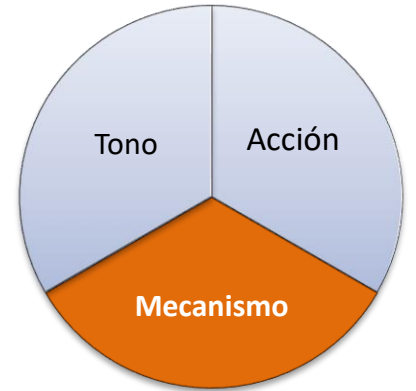
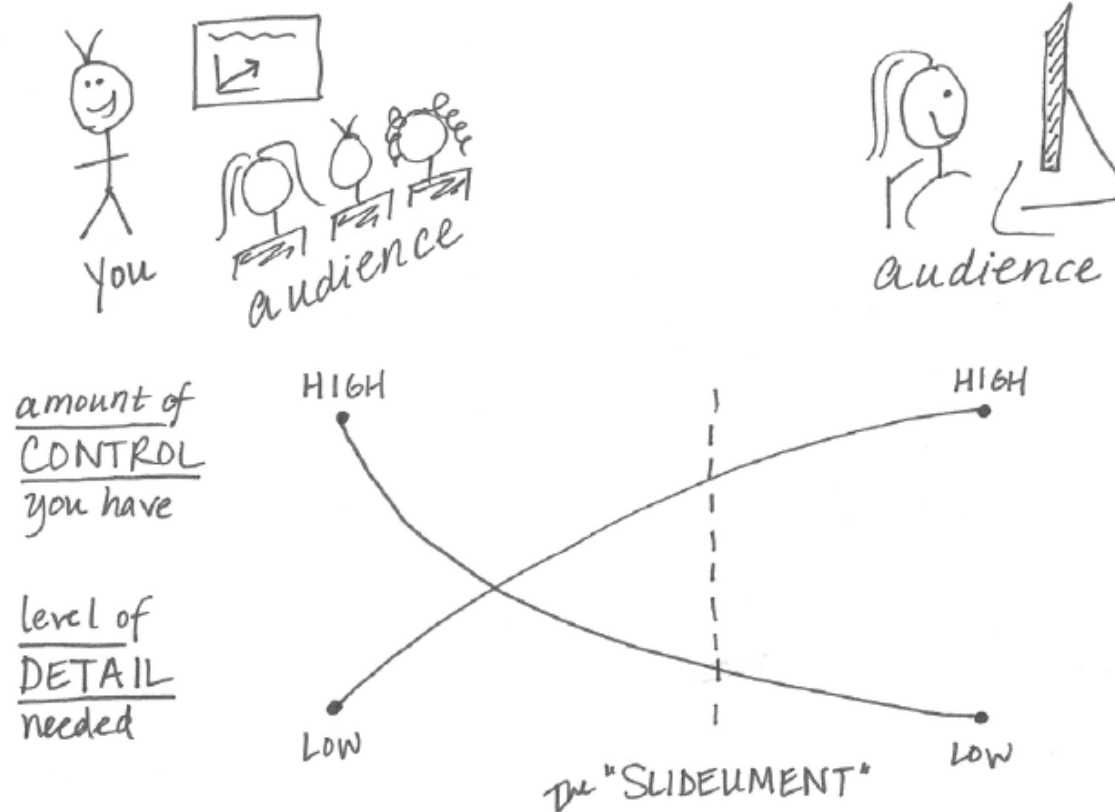
Here are some action words to help act as thought starters as you determine what you are asking of your audience:

accept | agree | begin | believe | change | collaborate | commence
| create | defend | desire | differentiate | do | empathize |
empower | encourage | engage | establish | examine | facilitate
| familiarize | form | implement | include | influence | invest |
invigorate | know | learn | like | persuade | plan | promote
| pursue | recommend | receive | remember | report | respond |
secure | support | simplify | start | try | understand | validate

TECNOLOGÍAS DIGITALES	RECURSOS Y HERRAMIENTAS Flashcards (Tarjetas para memorizar) Definición. Hecho/Dato. Hoja de Trabajo Lista Reproductor	paginas web Explicar Publicaciones Listar Etiquetar Graficas Bosquejar Categorizar	Esculpir Captura de pantalla Entrevistar Audacity Skype Jugar Videojuegos Simuladores	Implicaciones PMI Informar Publishing Usar Hoja de Cálculo Microsoft EXCEL Digitalizadores	TIKES Desktop Investigar Video conferencias Discusiones en cadena Telecomunicaciones Clases virtuales	Hypercomic Programar Game maker Alice Moldear Autocad
------------------------------	---	--	---	--	---	---

Qué – Mecanismo de comunicación

LIVE PRESENTATION WRITTEN DOC or EMAIL

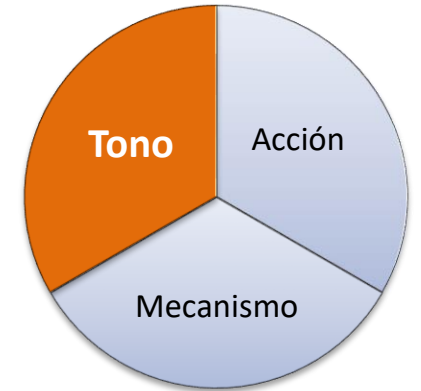


1. Entienda el contexto – Qué - Mecanismo

- Mecanismo: ¿qué va a utilizar para comunicarse con su audiencia?
 - ☐ Implicaciones del método que va a utilizar
 - ☐ Su control sobre la forma como su audiencia tome la información
 - ☐ Nivel de detalle requerido
- Tips:
 - ☐ NO lea sus diapositivas: no es un teleprompter
 - ☐ Conozca el contenido de su presentación
 - ☐ Practique en voz alta
 - ☐ No sobrecargue sus diapositivas. Incluya solo lo que le ayude a reforzar lo que quiere comunicar
 - ☐ Anote los puntos importantes a tratar
 - ☐ Programe la transición entre diapositivas
 - ☐ Practique con un amigo



1. Entiende el contexto – Qué – Tono



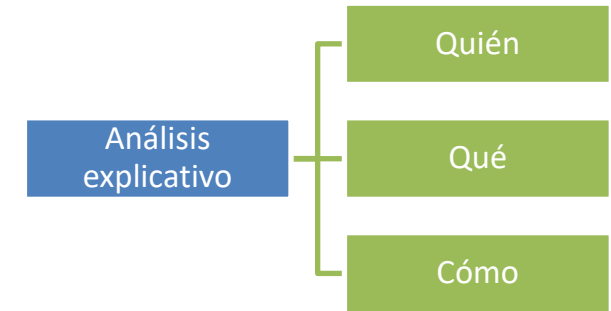
- Tono de la comunicación
 - ¿Están celebrando un logro?
 - ¿Busca “animar” para generar un cambio o una movilización?
 - ¿El tema es ligero o serio?

1. Entiende el contexto – Cómo

- ¿Qué datos hay disponibles?
- ¿Cómo soportar su punto?
 - Mostrar tanto los datos que soportan el punto, como los que no.
- Evidencia



Manos a la obra



Usted participo en una campaña de salud regional, enfocada a la prevención y detección temprana de patologías en mujeres. El proyecto incluyó la toma de mamografías, ecografías, citologías y valoración por ginecología.

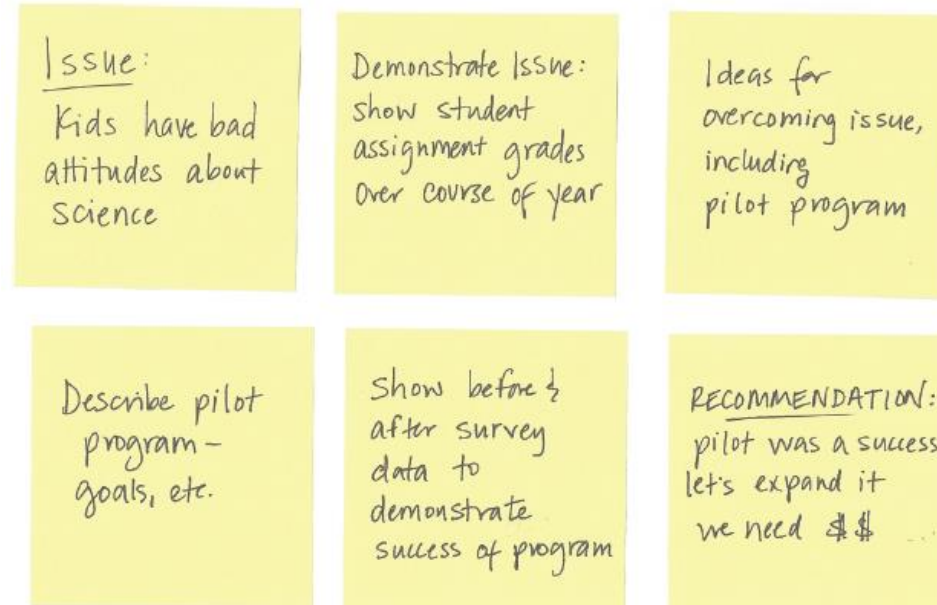
Se tomaron estadísticas de ocurrencia de cáncer y atención de casos, antes y después de la campaña, para ver cómo se afectaron los costos de atención.

Cree que los datos muestran una gran historia de éxito y le gustaría continuar desarrollando campañas similares en otras poblaciones.

Storyboarding - guión

Estructura visual de la comunicación

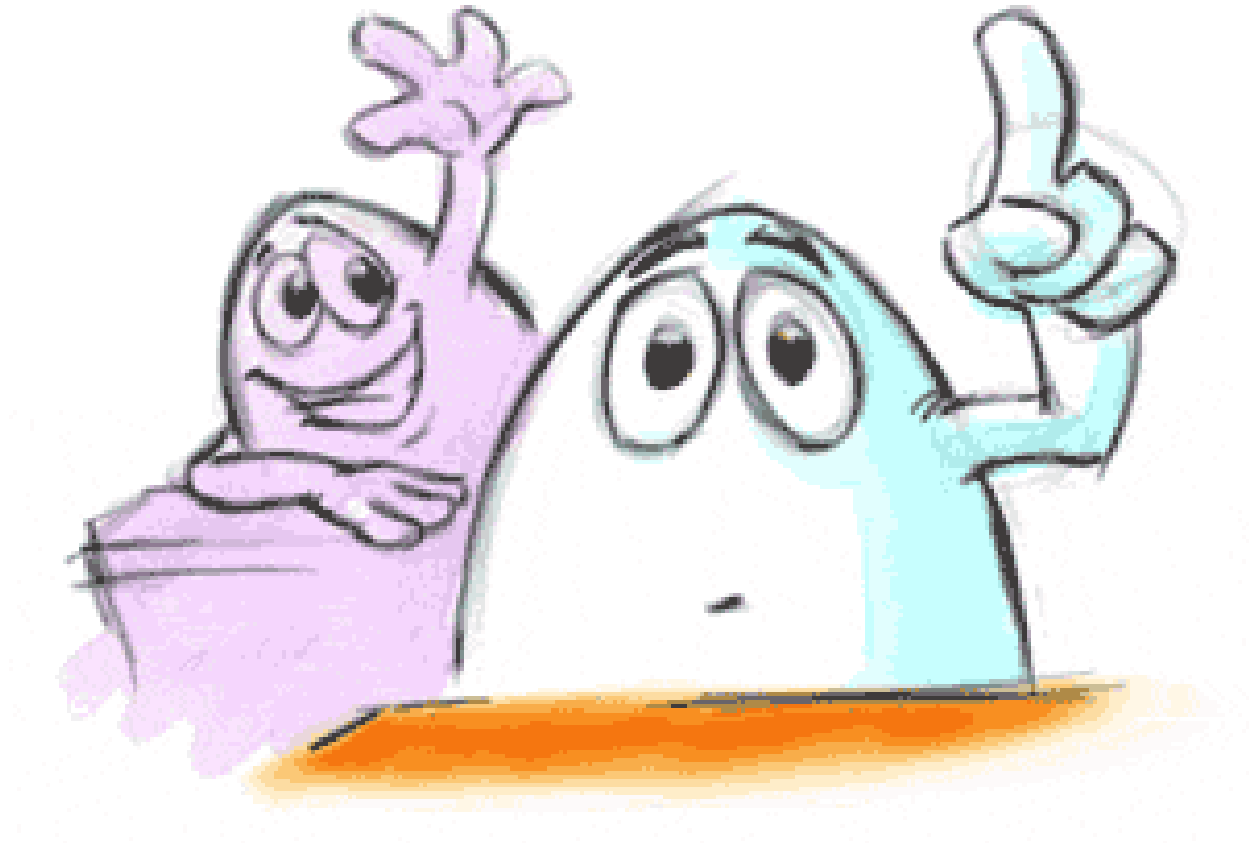
Use tablero, post-its o papel para crear la narrativa



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



¿Preguntas?

Metodología para visualizar datos

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