Alumni

Gestión de datos

Extraerlos, limpiarlos, procesarlos y comunicarlos

1

Las preguntas y las respuestas

1.1. Las preguntas

Detrás de un análisis de datos hay claras preguntas que condicionan el camino y el procesamiento de los datos.

¿Qué porcentaje de nuestros consumidores han registrado su número de teléfono?

¿Son las recomendaciones de productos en nuestra web efectivas?

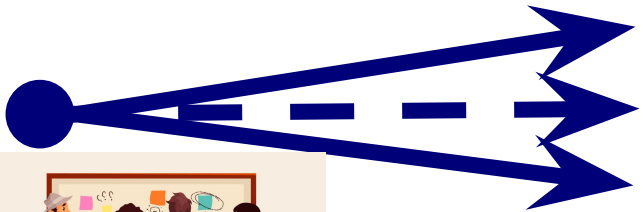
1.1. Las preguntas

- Preguntas específicas

¿Qué porcentaje de nuestros consumidores han registrado su número de teléfono?

- Preguntas generalistas

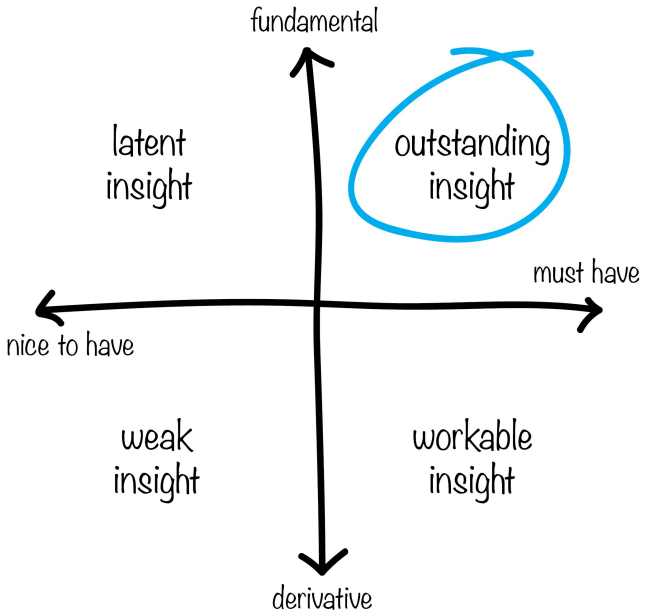
¿Son las recomendaciones de productos en nuestra web efectivas?



desglose en preguntas específicas o hipótesis

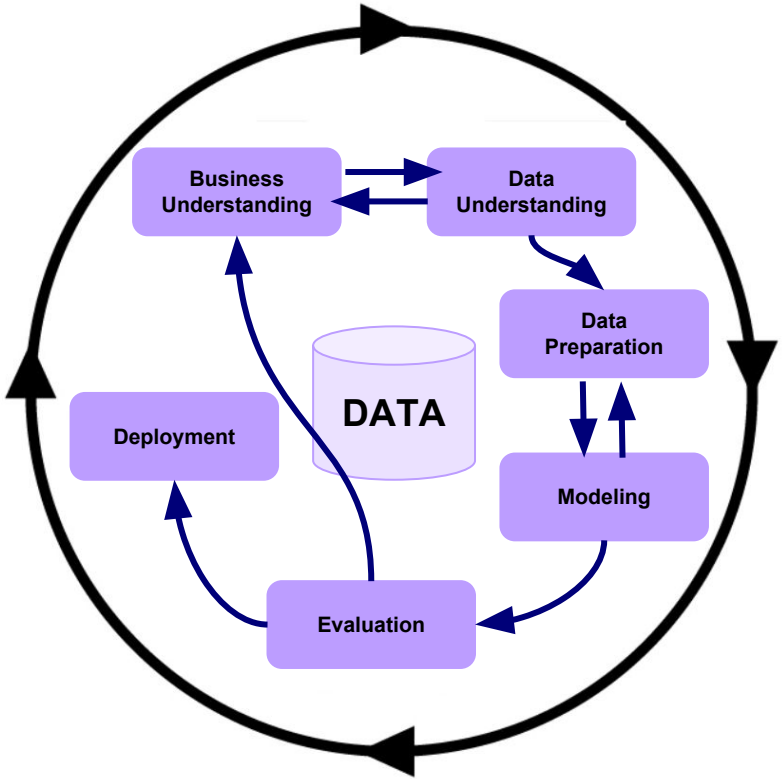
1.2. Las respuestas

- Poner escenarios a posibles soluciones y acciones detrás de la respuesta que puedes obtener ayudará en tus análisis.
- Mantener foco y distinguir entre:



1.3. CRISP Methodology

CRISP-DM:
CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING

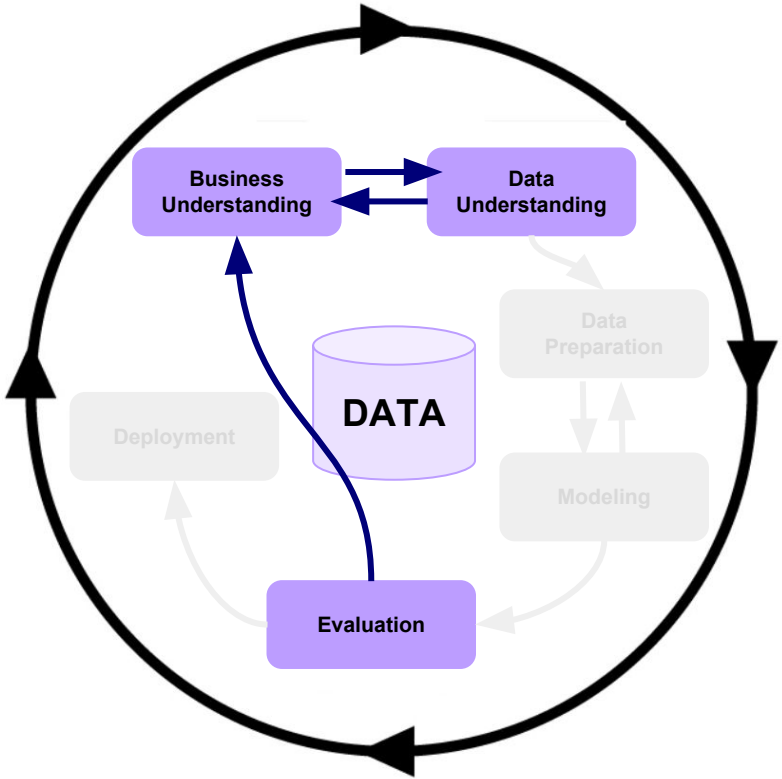


La metodología detrás de este proceso ayudará a tener éxito en el análisis

1.3. CRISP Methodology

Business Understanding:

- Recopilación información: definición objetivos, criterios de éxito
- Evaluación de la situación: requisitos, suposiciones, limitaciones, riesgos
- Creación de un plan preliminar

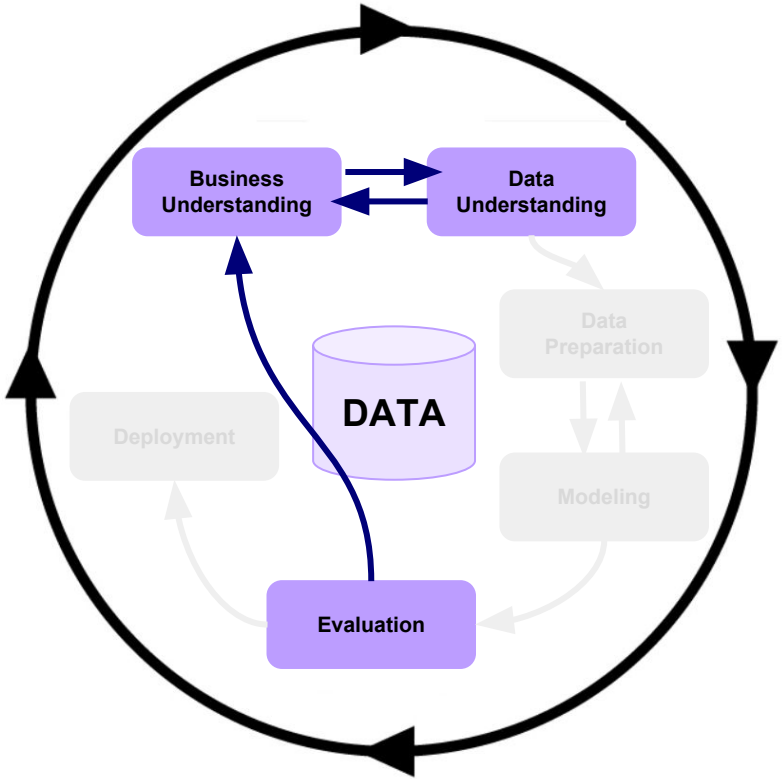


Ejemplo: *¿Son las recomendaciones de productos en nuestra web efectivas?*

1.3. CRISP Methodology

EXAMPLE OF THE PROJECT PLAN

Phase	Time	Resources	Risks
Business understanding	1 week	All analysts	Economic change
Data understanding	3 weeks	All analysts	Data problems, technology problems
Data preparation	5 weeks	Data scientists, DB engineers	Data problems, technology problems
Modeling	2 weeks	Data scientists	Technology problems, inability to build adequate model
Evaluation	1 week	All analysts	Economic change, inability to implement results
Deployment	1 week	Data scientist, DB engineers, implementation team	Economic change, inability to implement results



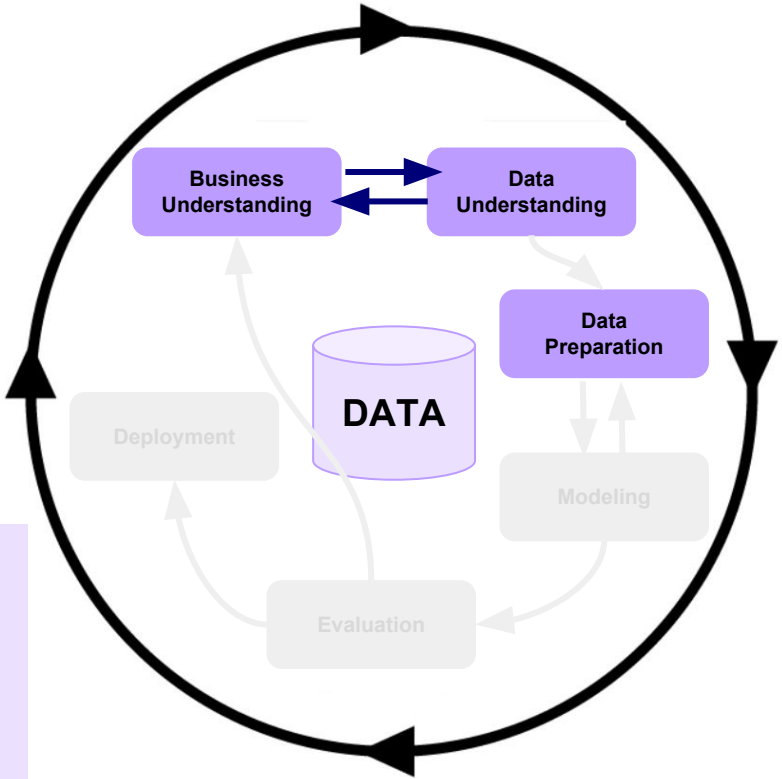
1.3. CRISP Methodology

Data Understanding:

- ¿Qué tipo de análisis o técnicas de Minería de datos pueden ayudar al problema.
- ¿Cómo garantizar que los resultados son precisos?
- ¿Cómo implementar los resultados?

Ejemplo: *¿Son las recomendaciones de productos en nuestra web efectivas?*

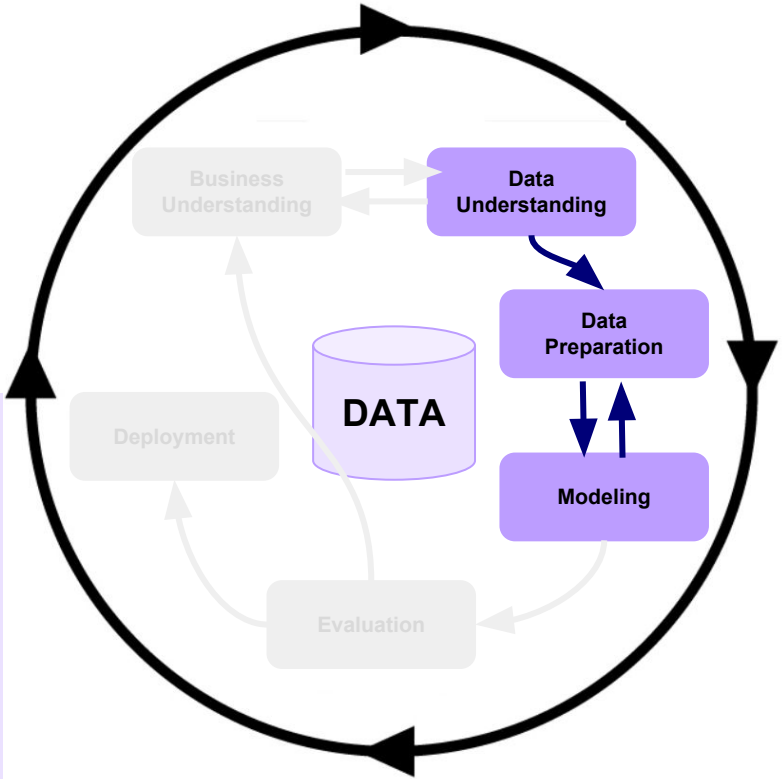
- *¿Dónde están? ¿Cómo? ¿Acceso? ¿Restricciones?*
- *¿Qué campos son clave? ¿Qué histórico? ¿Suficientes datos para conclusiones significativas?*



1.3. CRISP Methodology

Data Preparation:

- Selección de los datos
- Selección de las variables o atributos adecuados
- Analizar la calidad de nuestros datos
- Creación de código: Consistente, Reproducible, Escalable



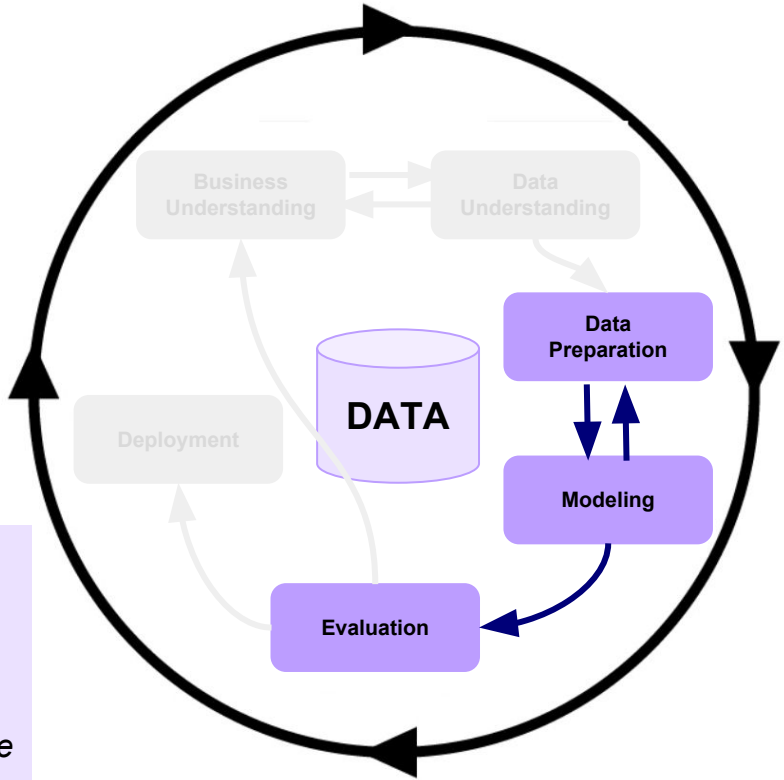
Ejemplo: *¿Son las recomendaciones de productos en nuestra web efectivas?*

- *Creación de un dataset donde cada fila son agregaciones de información de un consumidor en el último trimestre.*
- *Las variables, por ejemplo:*
  - *variables asociadas al customer, edad, género*
  - *# recomendaciones clickeadas per categoria,*
  - *# productos comprados por cada categoría,*
  - *# productos comprados que fueron recomendados por categoría.*

1.3. CRISP Methodology

Modeling:

- Selección de las técnicas o modelos que vamos a usar.
- Evaluar tus técnicas
- Validación de los resultados
- Interpretación de tus resultados en el problema de negocio.



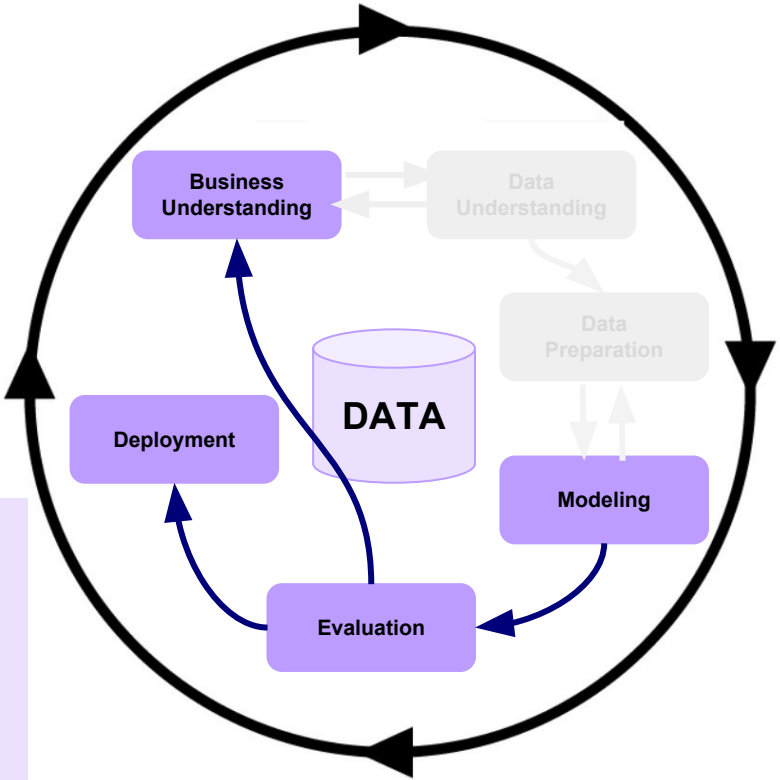
Ejemplo: *¿Son las recomendaciones de productos en nuestra web efectivas?*

En este caso en esta primera fase es hacer un análisis exploratorio de los datos de la conversión de productos que han sido comprados por una recomendación.

1.3. CRISP Methodology

Evaluación y Deployment:

- ¿Hemos encontrado la respuesta que necesitamos?
- ¿Tenemos que redirigir de nuevo el problema y realizar ajustes en nuestros datos?
- En caso de tener una solución. ¿Está está presentada clara? ¿Tienes claras acciones al respecto?

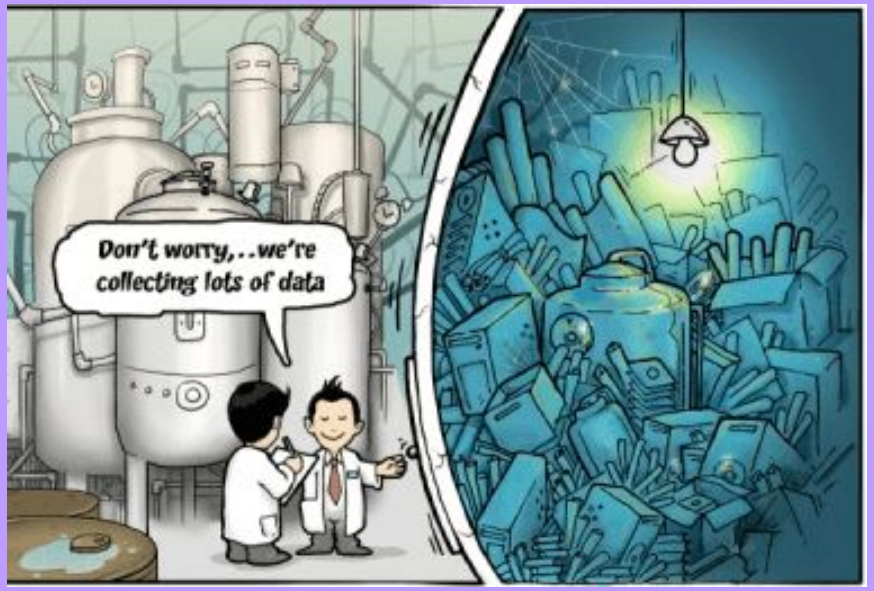


Ejemplo: *¿Son las recomendaciones de productos en nuestra web efectivas?*

Evaluación si la métrica es buena o no. En caso positivo el deployment será un reporte con toda la información. En caso negativo, iremos de nuevo a empezar el proceso, donde pasaremos por investigar qué está fallando del recomendador actual.

2

Extracción, procesamiento y calidad de datos



2.1. ¿Qué fuentes puedo usar para extracción de datos?

FUENTES DE DATOS

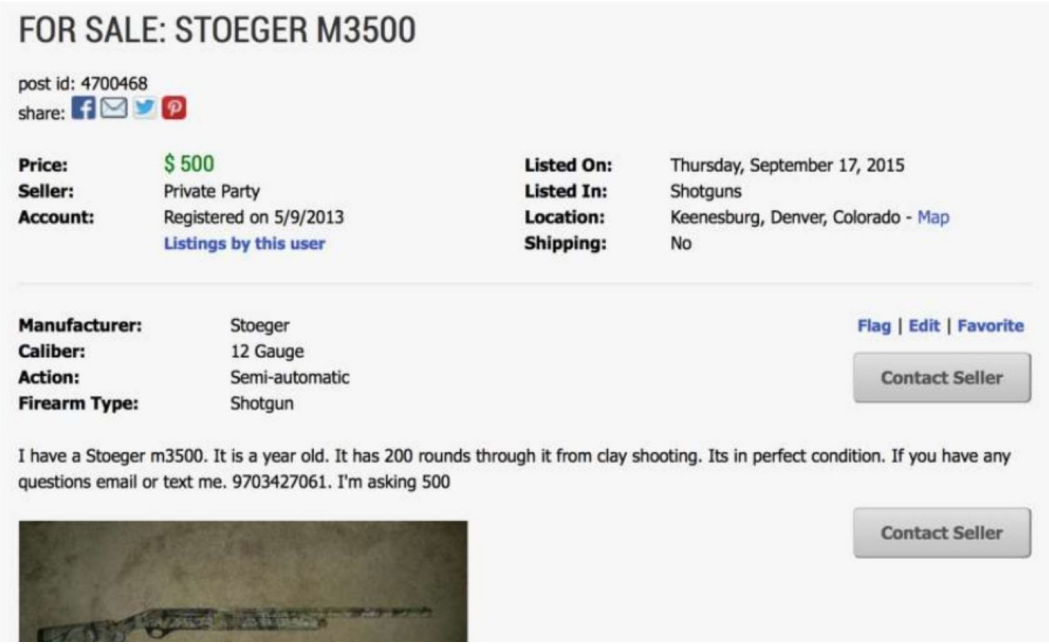
Tracking en páginas web
Tracking en aplicaciones
Configuraciones de sistemas
Estados de usuarios
Estados de sistema
APIs
...

ALMACENAMIENTO DE DATOS

Archivos de texto: XML, JSON, CSV, TXT
Archivos procesados: Excel
Base Datos relacionales: MySQL, Postgres, Hive
Bases de datos no relacionales: MongoDB
...

2.2. Procesamiento de datos

2.3. Ejemplo



5,000 Vistas de Página /Hour
24 Horas
7 Days

This is a total of

1.000.000 pages
every week

2.3. Ejemplo

FOR SALE: STOEGER M3500

post id: 4700468
share:

Price:	$ 500	**Listed On:**	Thursday, September 17, 2015
Seller:	Private Party	**Listed In:**	Shotguns
Account:	Registered on 5/9/2013	**Location:**	Keenesburg, Denver, Colorado - Map
	Listings by this user	**Shipping:**	No

Manufacturer:	Stoeger
Caliber:	12 Gauge
Action:	Semi-automatic
Firearm Type:	Shotgun

Flag | Edit | Favorite

Contact Seller

I have a Stoeger m3500. It is a year old. It has 200 rounds through it from clay shooting. Its in perfect condition. If you have any questions email or text me. 9703427061. I'm asking 500

Contact Seller

FOR SALE: STOEGER M3500

post id: 4700468
share:

Price:	$ 500	**Listed On:**	Thursday, September 17, 2015
Seller:	Private Party	**Listed In:**	Shotguns
Account:	Registered on 5/9/2013	**Location:**	Keenesburg, Denver, Colorado - Map
	Listings by this user	**Shipping:**	No

Manufacturer:	Stoeger
Caliber:	12 Gauge
Action:	Semi-automatic
Firearm Type:	Shotgun

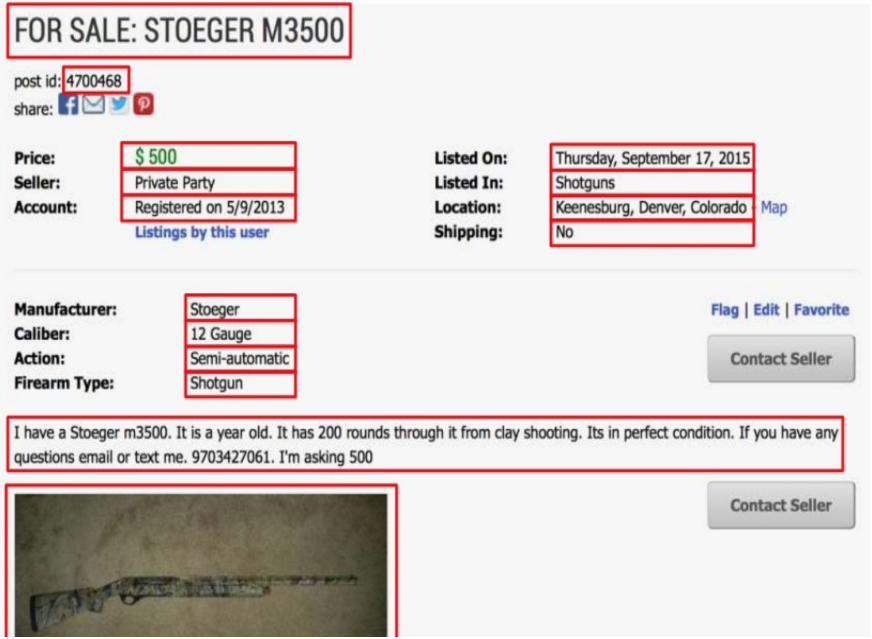
Flag | Edit | Favorite

Contact Seller

I have a Stoeger m3500. It is a year old. It has 200 rounds through it from clay shooting. Its in perfect condition. If you have any questions email or text me. 9703427061. I'm asking 500

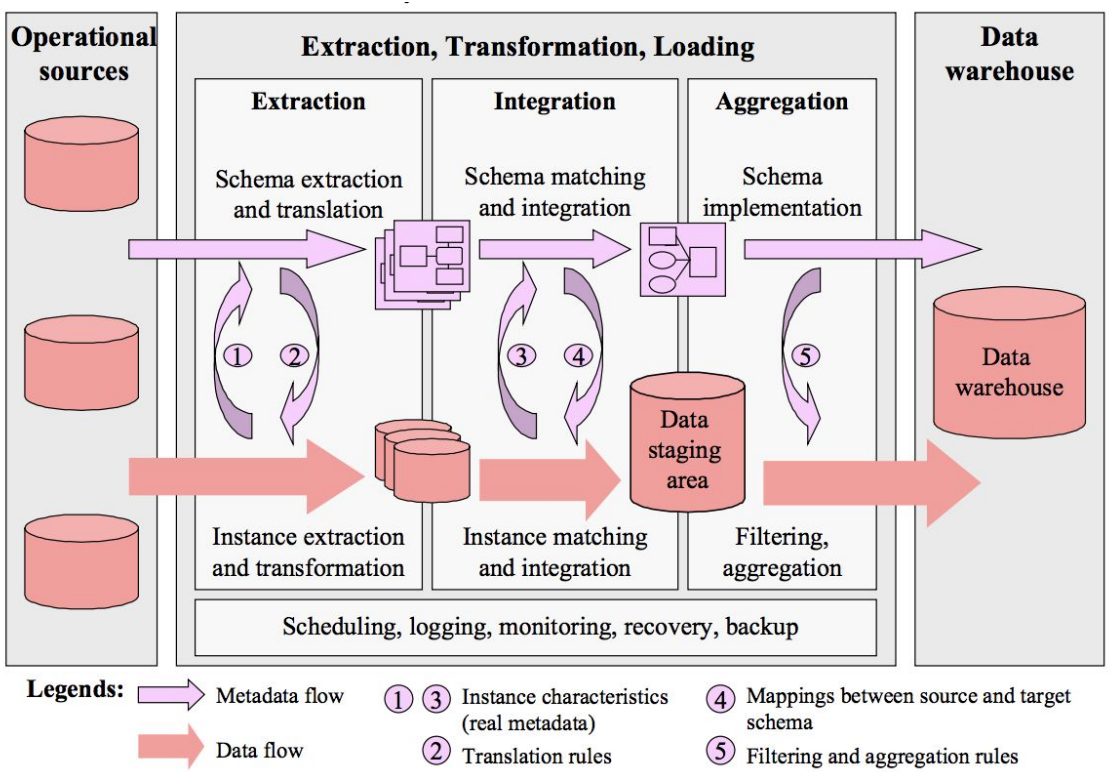
Contact Seller

2.3. De Raw Data to Structured Data



- Title
- Description
- Seller
- Post Date
- Expiry Date
- Price
- Location
- Category
- Member Since
- Num Views
- Post ID

2.4. Calidad de los datos



2.5. Típicos problemas en calidad de los datos

Problema	Ejemplo	Razón
Valores ilegales	bdate = 30.13.70	un valor fuera del rango
Dependencias irregulares	edad=22, bdate = 12.02.70	la edad debería ser: *edad = hoy-bdate*
Violación de Unicidad	(nombre ‘María Sanz, SNN=’1367’) (nombre ‘Luca Pérez, SNN=’1367’)	Un valor único por usuario
Violación de referencias en la fuente de datos	(nombre ‘Luca Pérez’, depno=’136’)	referencia de departamento 127 no existe
Missing values	(nombre ‘María Sanz, SNN=’1367’) (nombre ‘Luca Pérez, SNN=null)	Valores necesarios que no están
Duplicación	(nombre ‘María Sanz, SNN=’1367’) (nombre ‘Mara Sanz, SNN=’1367’)	Misma persona registrada varias veces
Errores Escritos	(city=’Maaalaga’)	Problemas ortográficos, Abreviaturas, convivencias de distintas nomenclaturas

2.6. Integrando fuentes de datos distintas

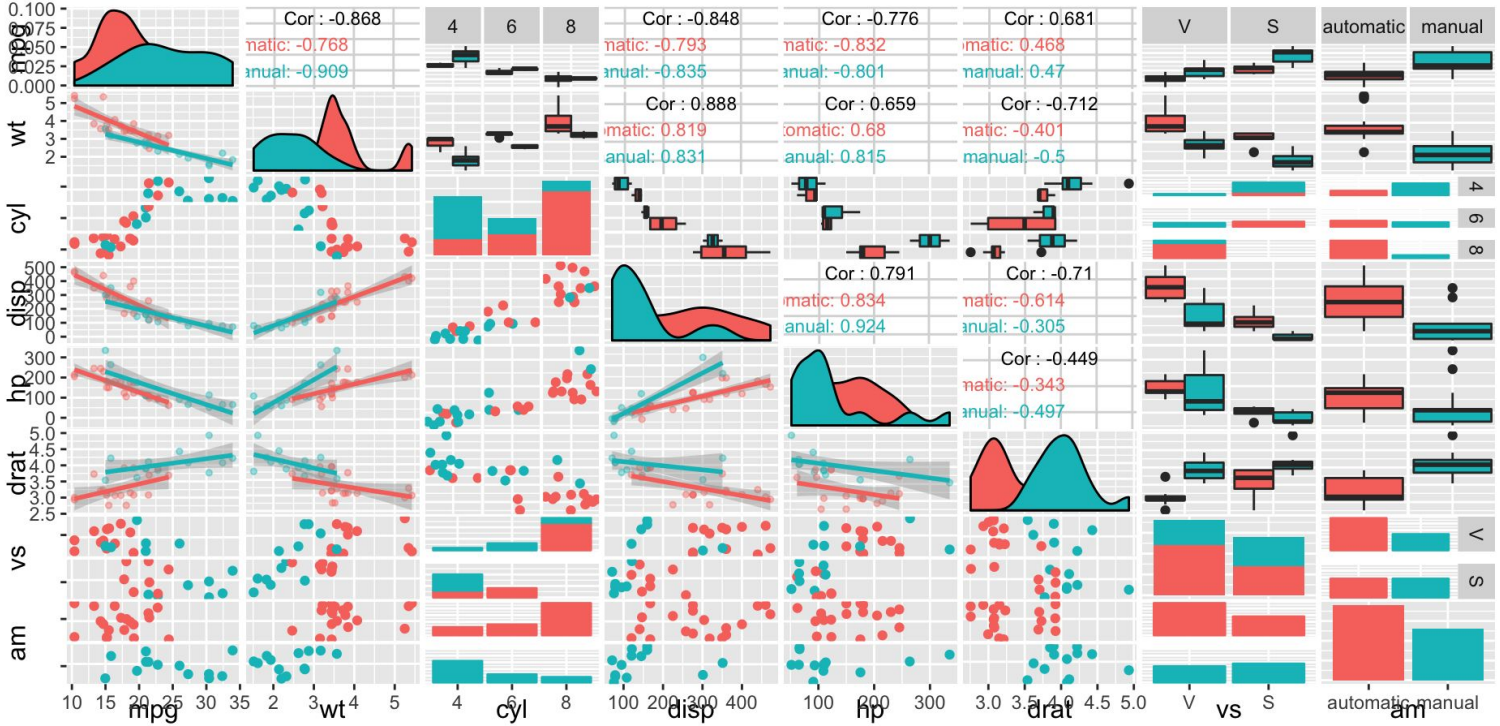
CID	Nombre	Calle	Ciudad	Género
11	V García	2 Avellana Pl.	Barcelona 08904	0
24	Manuel García	Avellana 21	Barcelona	1

Number ID	Nombre	Apellido	Dirección	Teléfono/email	Género
24	Luis	García	Virgen del Socorro, Sevilla, 05535	626803069/lgarcia@gmail.com	Hombre
345	Vanessa	García	Plaza de la Avellana, 2, Barcelona, 08904	vgarcia@yahoo.es	Mujer

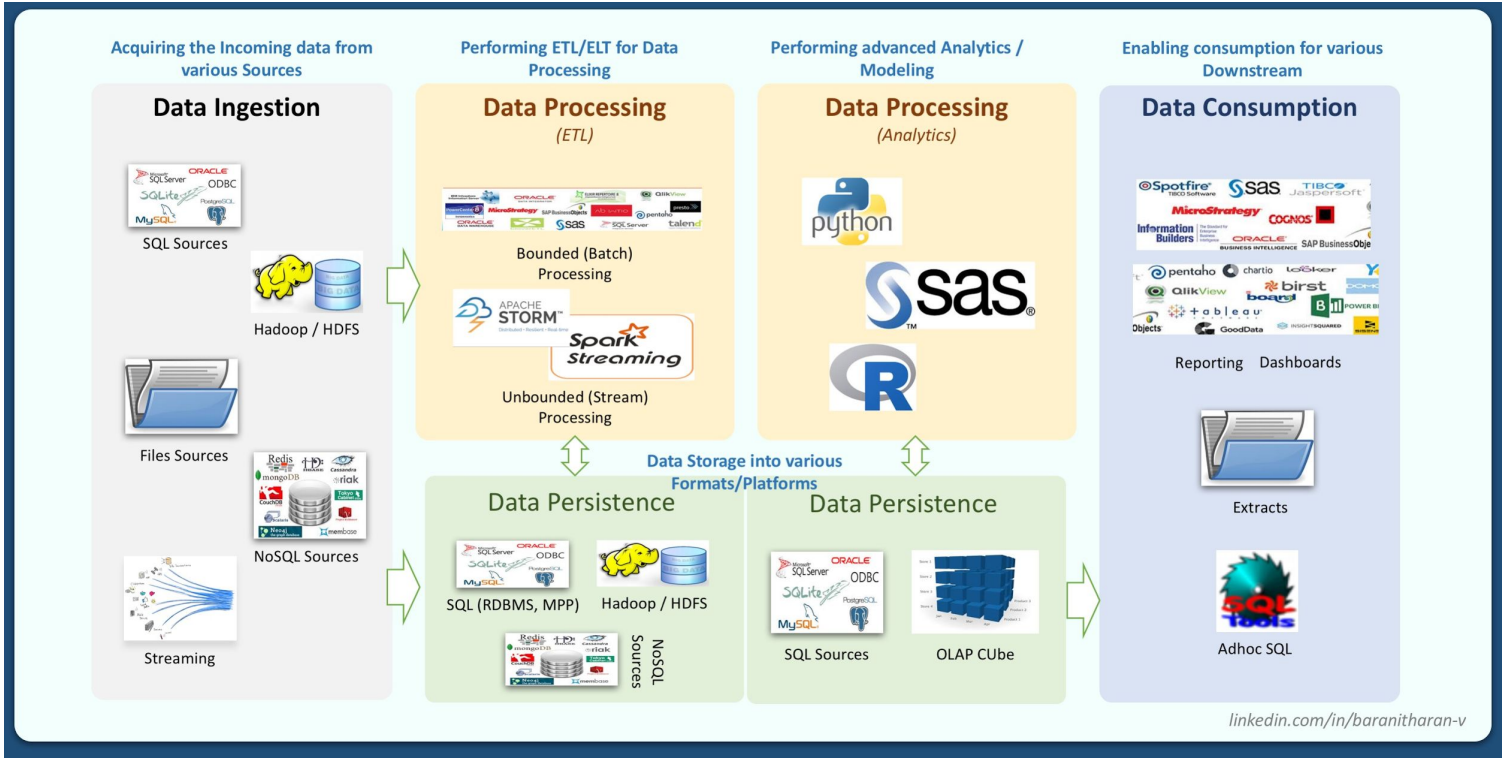
2.6. Integrando fuentes de datos distintas

Id	Nombre	Apellido	Calle	Numero	Ciudad	Código Postal	Teléfono	Mail	Género	CID	Number ID
1	Vanesa	García	Plaza Avellana	2	Barcelona	08904	626803069	vgarcia@yahoo.es	Mujer	11	345
2	Luis	García	Virgen del Socorro		Sevilla	05535		lgarcia@gmail.com	Hombre		24
3	Manuel	García	Avellana	21	Barcelona				Hombre	24	

2.7. Mirando tus datos

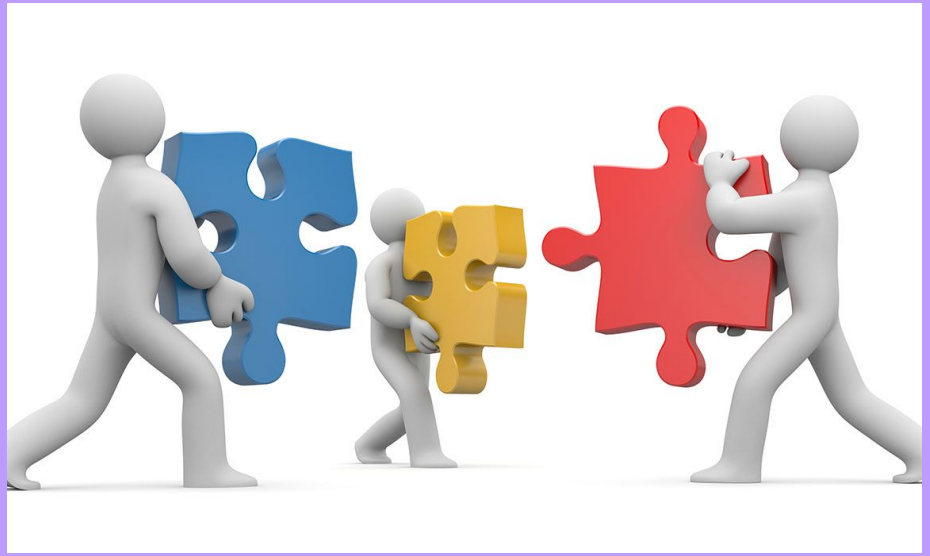


2.4. Tools


Acquiring the Incoming data from various Sources
Performing ETL/ELT for Data Processing
Performing advanced Analytics / Modeling
Enabling consumption for various Downstream
Data Ingestion
SQL Sources
Hadoop / HDFS
Files Sources
NoSQL Sources
Streaming
Data Processing
(ETL)
Bounded (Batch) Processing
Unbounded (Stream) Processing
Data Processing
(Analytics)
Data Storage into various Formats/Platforms
Data Persistence
SQL (RDBMS, MPP)
Hadoop / HDFS
NoSQL Sources
Data Persistence
SQL Sources
OLAP CUbe
Data Consumption
Reporting
Dashboards
Extracts
Adhoc SQL
linkedin.com/in/baranitharan-v

3

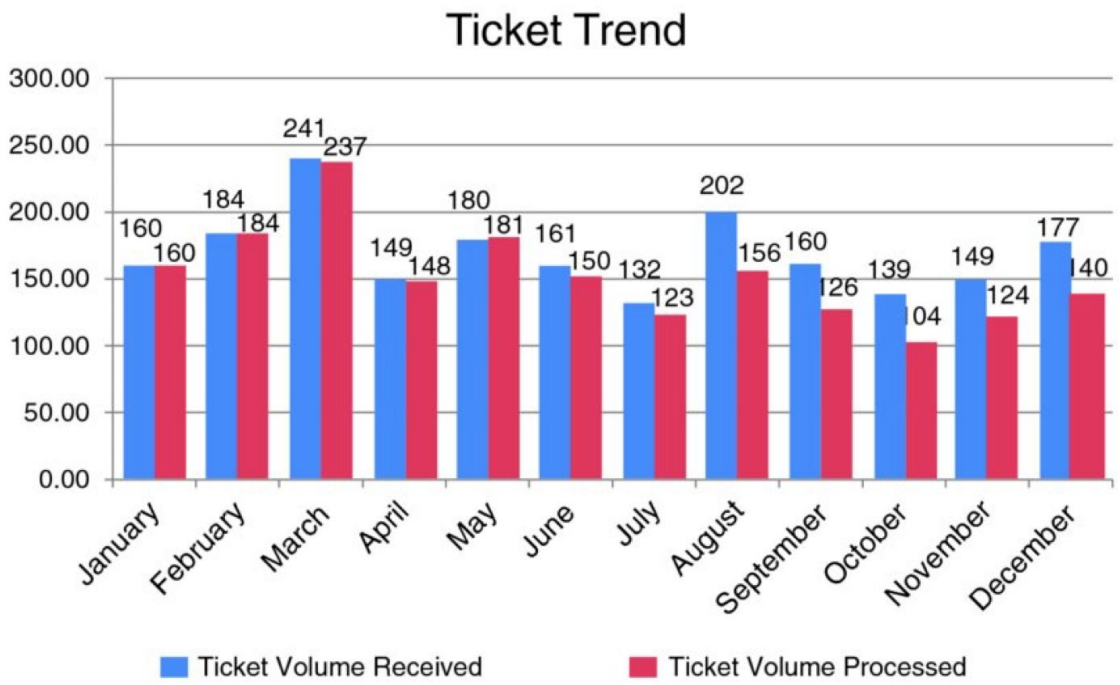
Presentación y comunicación de tus resultados



3.1. Tips generales a tener en cuenta en tu discurso o reporte:

- Adaptarse a tu audiencia objetivo
- Elección del mejor camino para mostrar resultados: ppt, reporte, mail, dashboard…
- Entender el contexto en el que se comunica
- Claro+Conciso+Sencillo
- Claridad en conclusiones y en acciones a hacer para tu audiencia

3.2. Comunicar resultados no es solo visualizar



¿Qué vemos aquí?

3.2. Comunicar resultados no es solo visualizar

“no dejes que otros interpreten tus resultados, tú eres el que más sabe de tú análisis”

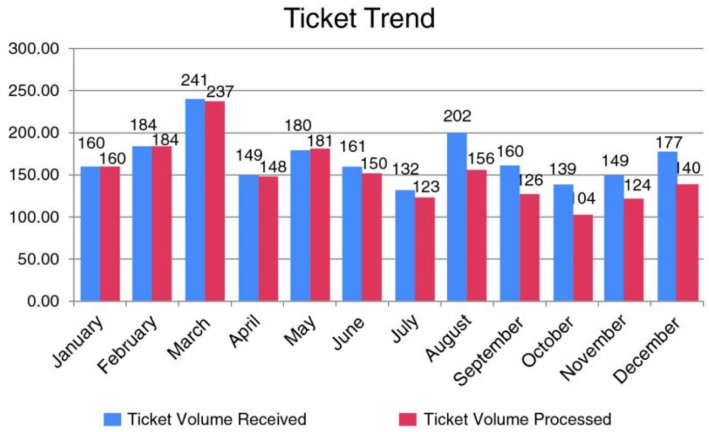
3.3. Tip 1: cuenta una historia

3.3. ¿Te imaginas Caperucita Roja explicado así?

- Caperucita Roja (CR) camina 554 metros del punto A (Casa) al punto B (Casa la Abuela)
- CR se encuentra al lobo, quien (1) corre a punto B, (2) come abuela, (3) pone sus ropas
- CR llega a punto B a las 14:00 pm. Hace tres preguntas.
- Identificación del problema: después de tercera pregunta, Lobo come a CR.
- Solución: el cazador usa una herramienta (el hacha)
- Resultado Esperado: Abuela y CR vicen, el lobo no

3.3. Tip 1: cuenta una historia

Antes



Después

Please approve the hire of 2 FTEs

to backfill those who quit in the past year

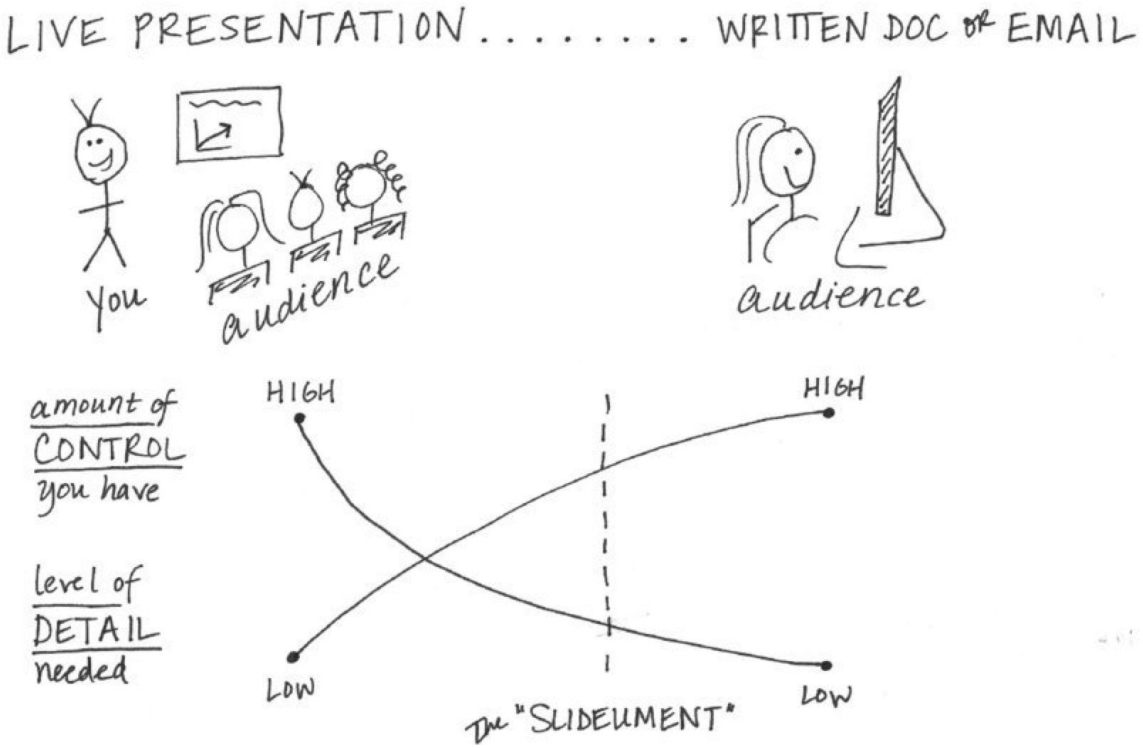
Acción

Ticket volume over time

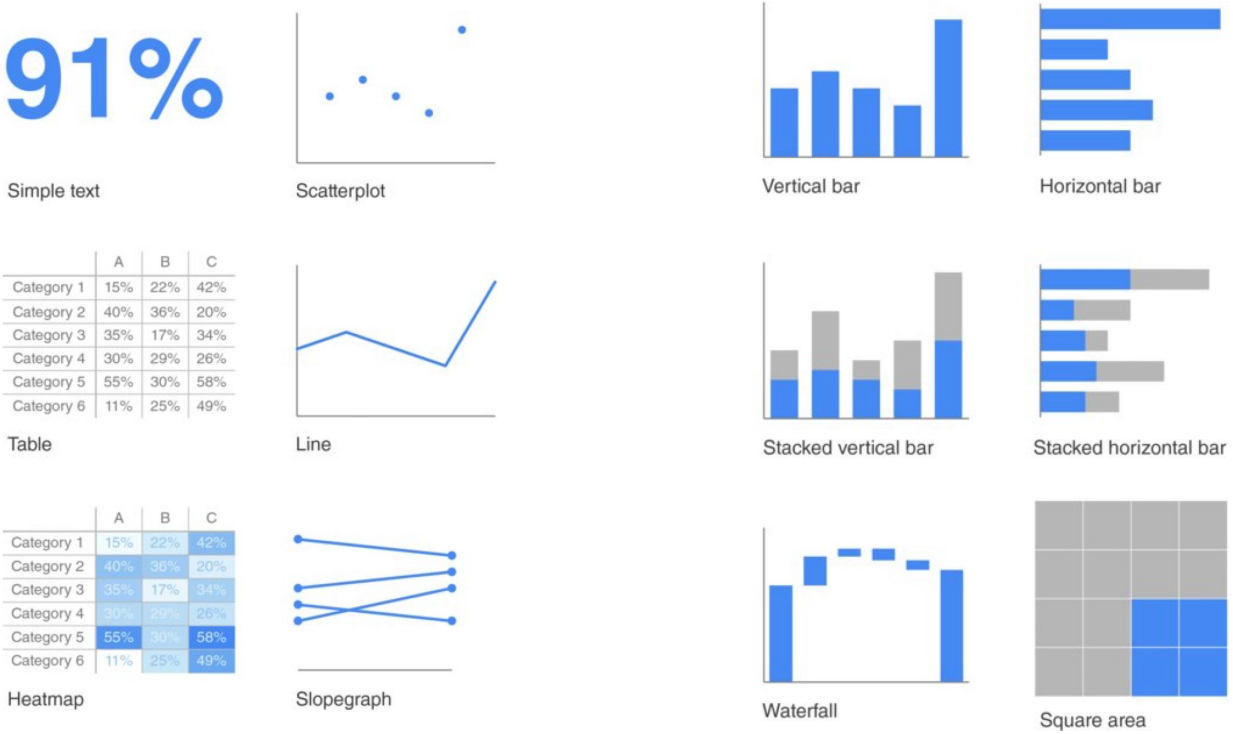
Contexto

Atención

3.4. Tip 2: el control que tú tienes dependiendo del formato elegido



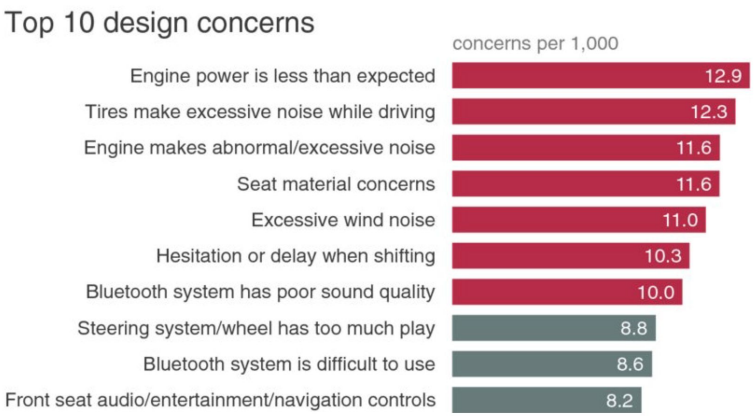
3.5. Tip 3: no necesitamos visualizaciones complicadas



3.6. Tip 4: Ayuda a que tu audiencia se focalice en lo que tú quieres, usa contraste o elementos clave (tamaño, colores, tipos de letra…)

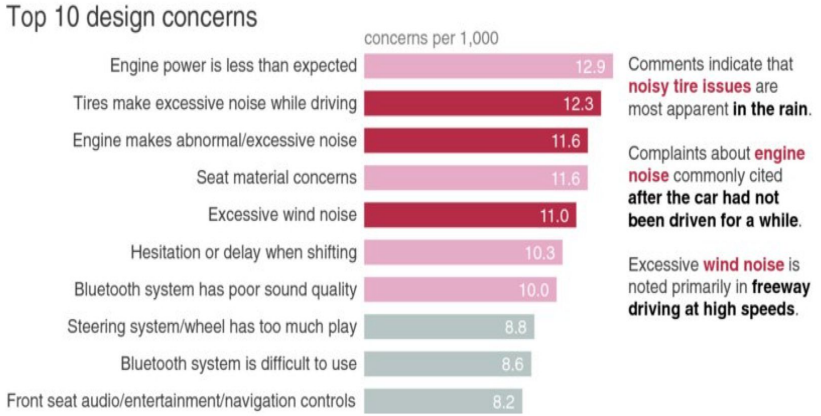
Foco 1

7 of the top 10 design concerns have 10 or more concerns per 1,000.
Discussion: is this an acceptable default rate?

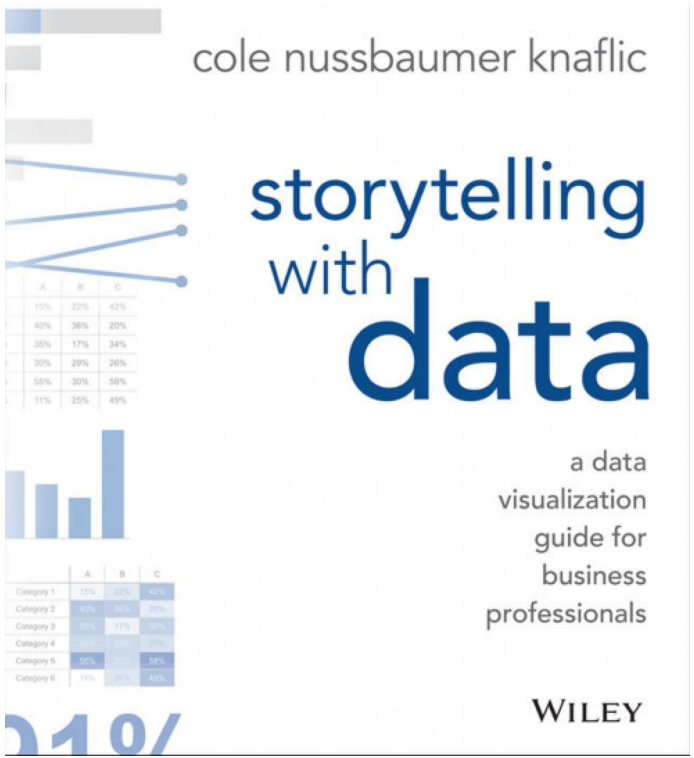


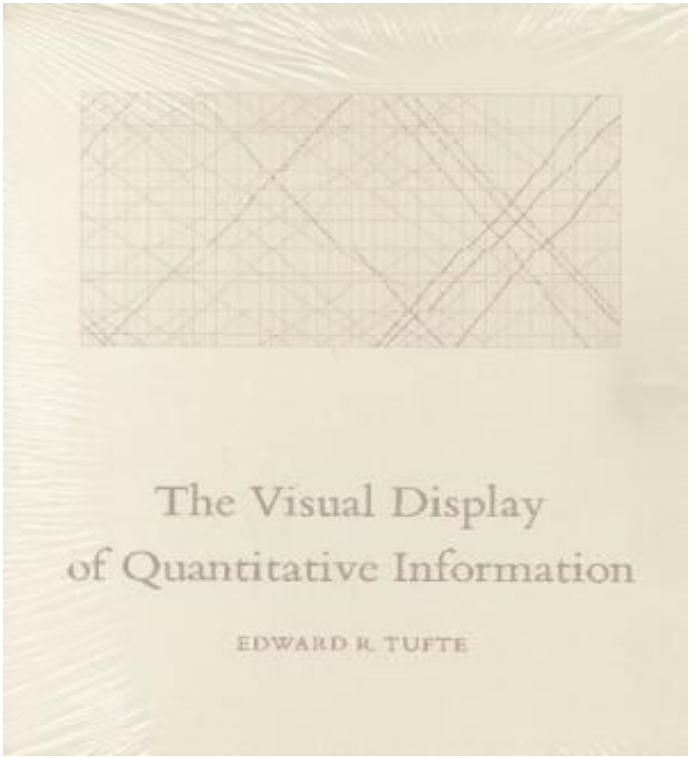
Foco 2

Of the top design concerns, three are noise-related.



3.7. Algunas referencias





Alumni

AlumniUOC
@UOCalumni