Visualización de datos

Calidad y pre-procesamiento de datos

Contenido

Calidad de datos

¿Cómo saber cuándo sus datos son lo suficientemente buenos?
¿Cómo saber cuándo hay que preocuparse y cuándo no, y sobre qué?
¿Hay que desechar los datos de la encuesta porque un par de personas no han respondido a ciertas preguntas?

Calidad de datos

La intuición es buena pero…

Operar desde un marco de calidad de datos le permite:

- **Dejar de preocuparse por lo que cree saber o no saber sobre los datos.**
- **Salir de la sabiduría convencional sobre lo que necesita y no necesita, y establecer establecer nuevos conceptos sobre los datos y sus problemas.**
- **Desarrollar y reutilizar herramientas para la gestión de la calidad de los datos en una mayor variedad de de escenarios y aplicaciones.**

Un enfoque sistemático del análisis de la calidad de los datos de los datos puede guiarle de forma eficaz y coherente hacia un mayor grado de conocimiento de las características y la calidad de sus datos antes de dedicar un tiempo excesivo a la toma de decisiones personales o empresariales.

Framework: Las cuatro C de la calidad de los datos

Completitud
¿Está todo lo que debe estar aquí?

Coherencia
¿Todos los datos "cuadran"?

Correctitud
¿Son, de hecho, los valores correctos?

aCcesibilidad
¿Podemos rastrear los datos?

Completitud/Exhaustividad

Cliente: "es un conjunto de datos que lo tiene todo“

Uds: ¿cómo definiríamos "todo" aquí? ¿Es simplemente el número de registros? ¿El número de campos previstos? Quizás deberíamos incluir si todos los campos tienen valores-asumiendo que la "falta de valor" no es en sí mismo, un valor.

Respuesta: ¿Tengo todos los datos necesarios para responder a mi(s) pregunta(s)?

Entienda la pregunta a la que desea responder.

Completitud

¿Qué acciones puedo hacer para evaluar integridad?

- Evaluar los atributos que necesita para responder la pregunta
- Análisis de datos perdidos. (Ojo, perdidos no es erróneos)
- Evaluar el porcentaje “aceptable” de datos perdidos basados en reglas de negocios.

¿Qué acción(es) debe(n) tomar cuando haya(n) evaluado la integridad?

- Eliminar los registros que faltan. Esta es una buena opción si su conjunto de consultas debe ser internamente coherentes entre sí.
- Corregir los datos que faltan. Si se puede, es una buena opción, aunque no siempre es posible o práctico.
- Marque los registros infractores y anote su carácter incompleto con los resultados de la consulta.
- No hacer nada. Dependiendo de sus consultas, puede seguir con su análisis. Aunque esto suele ser una mala idea la mayoría de las veces…

Coherencia

¿Sus datos tienen sentido en relación con ellos mismos?
¿Son los propios valores internamente coherentes donde tienen que serlo?

En concreto, se trata de determinar si los registros se relacionan entre sí de forma coherente y siguen la lógica interna del conjunto de datos.

<u>¿Qué acciones puedo hacer para evaluar la coherencia?</u>

- **Análisis de datos duplicados**
- **Si hay más de una fuente de datos con relaciones, evalúe como es la relación de éstas fuentes de datos.**

Coherencia

¿Qué acción(es) debe(n) tomar cuando haya(n) evaluado la integridad?

- Determine qué nivel y forma de coherencia necesita realmente.
¿Se trata de una validación de la integridad referencial? ¿Una simple validación de integridad del valor en uno o más valores de campo? ¿O requiere una validación de integridad más compleja?
- Determine cuán completa debe ser la validación y cuáles son sus limitaciones de rendimiento y tiempo.
¿Necesita validar todos los registros o relaciones? ¿O puede aplicar el muestreo estadístico para elegir un subconjunto significativo pero manejable de registros para evaluar?
- Como siempre, es fundamental que la evaluación se adapte a sus necesidades y equilibre adecuadamente la tiempo/calidad.

Consistencia

¿Tus datos son lo suficientemente correctos para la pregunta que intentas responder?

Al igual que la comprobación de la coherencia, la consistencia requiere un cierto grado de conocimiento del ámbito, es decir, reglas de negocio.
El objetivo último de los datos es impulsar las decisiones, por lo que su riesgo máximo es ser incorrectos.

¿Qué acciones puedo hacer para evaluar la consistencia?
Antes de… ¡Pedir el diccionario de datos!
¿No hay diccionario de datos? Ir directamente al experto de los datos. Ojo: No siempre el experto de datos es el mismo experto en reglas de negocio.

- **Determine cuáles son importantes para validar, en el contexto de lo que le interesa.**
- **Detalla que la información brindada (diccionario o experto) se refleje en los datos.**

Consistencia

¿Qué acción(es) debe(n) tomar cuando haya(n) evaluado la consistencia?

- **Comprenda qué parte de sus datos debe ser correcta. ¿puede (debe) comprobar todos los registros, o es suficiente algún tipo de muestreo?**

- **Decida qué hará con los datos incorrectos. ¿Es posible "arreglar" los registros de alguna manera, o debe trabajar sin ellos?**

Accesibilidad

¿Quién es el responsable de sus datos?

A la hora de evaluar la calidad de sus datos, ¿puede hacer afirmaciones similares de trazabilidad?

Accesibilidad

¿Qué acciones puedo hacer para evaluar la accesibilidad?

- Mantenga un registro de sus fuentes de datos. Cómo acceder, responsables, tipo de permisos y privacidad.
- Guardar todo. Desde los datos originales hasta las modificaciones.
- Auditoría de los datos en cada proceso.
- Observe el flujo de datos.

¿Cómo registrar el proceso de calidad de datos?

Documentar TODO en un informe de calidad de datos

Tips:

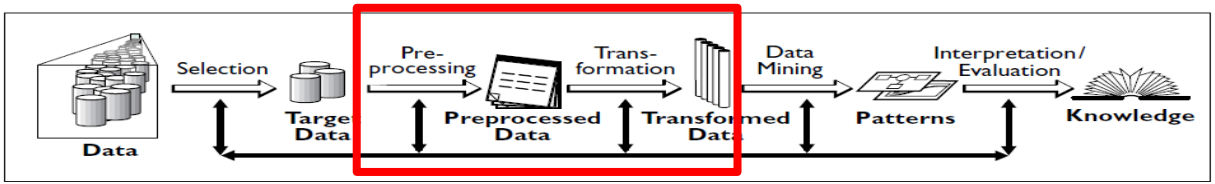
- Registrar cada hallazgo que se encontró en cada uno de los procesos del marco de calidad de datos (Completitud, Coherencia, Consistencia, Accesibilidad)
- Soportar en anexos el procedimiento que se realizó
- Diccionario de datos de las fuentes y atributos seleccionados
- Acuerdos de confidencialidad
- Habeas data

Ejemplo DANE:
https://sitios.dane.gov.co/revista_ib/html_r3/articulo12_r3.html

Preprocesamiento de datos

Pre-procesamiento de datos

El preprocesamiento de datos es una etapa esencial del proceso de descubrimiento de información o KDD. Esta etapa se encarga de la limpieza de datos, su integración, transformación y reducción para la siguiente fase de minería de datos.



Debido a que normalmente el uso de datos de baja calidad implica un proceso de minería de datos con pobres resultados, se hace necesaria la aplicación de técnicas de preprocesamiento.

Preprocesamiento de datos

La limpieza de datos es el proceso de corregir o eliminar datos incorrectos, corruptos, formateados incorrectamente, duplicados o incompletos dentro de un conjunto de datos.

- Eliminación de errores cuando están en juego múltiples fuentes de datos.
- Menos errores generan clientes más felices y empleados menos frustrados.
- Capacidad para mapear las diferentes funciones y lo que se pretende hacer con sus datos.
- Supervisión de errores y mejores informes para ver de dónde vienen los errores, lo que facilita la corrección de datos incorrectos o corruptos para aplicaciones futuras.

Datos basura = análisis basura

Pre-procesamiento de datos

Después de la aplicación de la fase de preprocesamiento, el conjunto resultante puede ser visto como una fuente consistente y adecuada de datos de calidad para la aplicación de algoritmos de minería de datos. El preprocesamiento incluye un rango amplio de técnicas que podemos agrupar en dos áreas: preparación de datos y reducción de datos.

Preparación de datos (obligatoria):

- Análisis valores perdidos
- Análisis de datos duplicados
- Consistencia de datos
- Normalización de datos
- Discretización de datos
- Integración de datos

Reducción de datos (opcional):

- Selección de atributos
- Selección de instancias
- Discretización

Preprocesamiento de datos

Consideraciones importantes:

- Si desea presentar sus resultados y publicar sus datos, querrá publicar la versión depurada.
- Al documentar nuestro proceso, nos aseguramos de poder reproducirlo cuando aparezcan nuevos datos.
- La limpieza de los datos facilita el almacenamiento, la búsqueda y la reutilización.

Manos a la obra

- Github
- Python

Gracias

¿Preguntas?