ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN PROFESIONAL

Calidad de datos

Gestión de Datos

Vicente Calisto

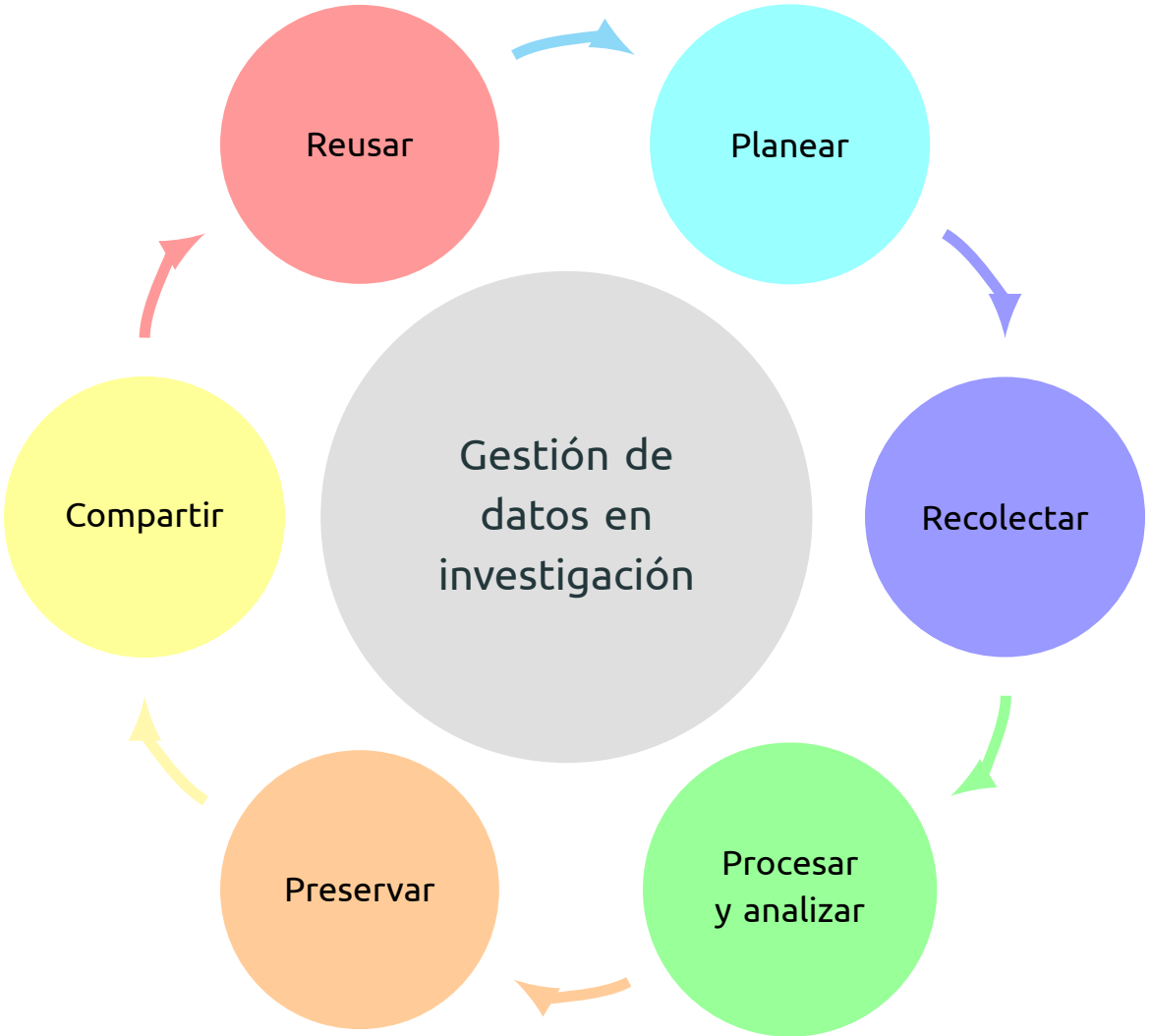
Educación Profesional - Escuela de Ingeniería

Tabla de Contenidos

Tabla de Contenidos

ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

¿Qué hemos visto hasta ahora?



Ciclo de vida de los datos

Planificación:

- ¿Dónde obtener los datos? Datos abiertos, generar propios datos, alianzas, ...
- ¿Dónde guardar los datos? Nube, disco duro personal, ¿papel?, ...
- ¿Qué voy a guardar como datos? Creación y/o modificación del esquema, codificación, ...

Ciclo de vida de los datos

Recolección:

- Traer los datos de las diversas fuentes a una sola ubicación.
- Inicio de documentación y metadatos
- Evaluar la calidad de los datos

Ciclo de vida de los datos

Procesamiento y análisis:

- Localizar y reparar los datos con errores.
- Llevar los datos a un formato que sea compatible con lo que queremos hacer.
- Anonimizar para asegurar acceso seguro de terceros
- Finalizar la documentación

Ciclo de vida de los datos

Preservar: Elegir qué datos compartiremos, dónde los guardaremos y llevarlos a un formato accesible.

Compartir: Escoger cómo compartirlos (repositorio), entregar documentación y elegir licencia.

Reusar: Luego que los datos están disponibles, un tercero que se encuentre en la etapa de planificación puede elegirlos como fuente de información. En este punto pueden obtenerse citas, manejar permisos de acceso, entre otras.

Hasta el momento nos hemos centrado en la parte técnica:

1. Ciclo de vida de los datos
2. Datos estructurados: Bases de datos relacionales.
3. R para manejo de los datos.
4. Lenguaje de consulta SQL.
5. Modelo E/R

Pero... ¿Cómo esto se integra en una investigación o proyecto de políticas públicas?

Un proyecto o una investigación por lo general involucra recolectar, combinar y transformar datos. Las investigaciones en políticas publicas tienen fuentes ricas en datos ya que provienen de instituciones publicas que por ley deben compartirlos (luego de quitarles la información sensible), los problemas:

- Esquema variable
- Poca consistencia
- Datos de gran tamaño
- Falta de documentación

Por esto gran parte de las investigaciones en el área involucran **big data**. En base a esto en las clases hemos visto:

1. Estructura de una investigación.
2. Datos bien estructurados (relacionales), su esquema y formas de tratarlos.
3. Diferencias entre R y SQL, como usar R para tratar y visualizar datos pequeños, principales comandos y ejemplos.
4. Lenguaje de consulta SQL, similitudes con R y ventajas de uso. Principales comandos y ejemplos.
5. Modelo entidad-relación para un buen manejo de los datos.

Pero... ¿Cómo sé que mis datos están limpios?

Tabla de Contenidos

ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

Documentación de los dato

La **documentacion de los datos (su ciclo de vida)** se refiere al proceso de reportar los principales aspectos de los datos utilizados en una investigación.

Cabe destacar que no existe un estándar para la documentación, sin embargo esta debe ser capaz de solucionar las principales preguntas de un externo a la investigación.

- ¿Cuáles son los objetivos de la recolección de datos?
- ¿Cómo se generaron o recolectaron?
- ¿Cómo se procesaron?
- ¿Cómo se organizan y cual es su formato y unidades de medida?
- Etc...

Documentación de los datos

En la documentación podemos encontrar:

- Información básica: Título, Creador, Resumen, Fechas importantes, Auspiciadores
- Estructura de los datos, directorios y archivos. Formato de los datos, requisitos de uso.
- Metodología con la cual los datos fueron generados (o la fuente)
- Procesamiento al que se sometieron, técnicas de validación y limpieza
- Historial de cambios, comparativas.
- Accesos a los datos, condiciones de uso o confidencialidad

¿Qué nos gustaría encontrar al momento de incorporar una fuente de datos?

Documentación de los datos

Por otro lado los **metadatos** es un archivo (que puede ser encontrado en la documentación) cuyo objetivo es describir los datos de una manera técnica. En este documento podemos encontrar:

- Definiciones de las tablas (esquema), restricciones y tuplas de ejemplo.
- Descripción de cada una de los atributos en las tablas.
- Diagramas Entidad-Relación (como se relacionan las tablas entre ellas)
- Reportes de calidad de los datos
- Seguridad, accesos

Documentación de los datos

El objetivo de los tanto de la documentación como los metadatos es introducir a un usuario nuevo a la base de datos, por esto es que se busca que sea corta pero con gran nivel de detalle. Ejemplos de esto es el esquema del registro en las bases de datos del MINEDUC o de la encuesta de su Tarea 2.

Si bien esto es fundamental para ayudar a la comprensión, no acaba acá. También es fundamental entregar datos de buena calidad para facilitar su uso y minimizar los posibles errores, para esto estudiaremos las dimensiones de la calidad de los datos que nos servirá de guía y métrica. [1]

Una forma de organizar y documentar los datos de una forma más fácil, rápida y cumpliendo los estándares de calidad de datos a través del **esquema entidad-relación**.

[1] Para más información pueden visitar la pagina: https://www.axiomdatascience.com/best-practices/index.html

Tabla de Contenidos

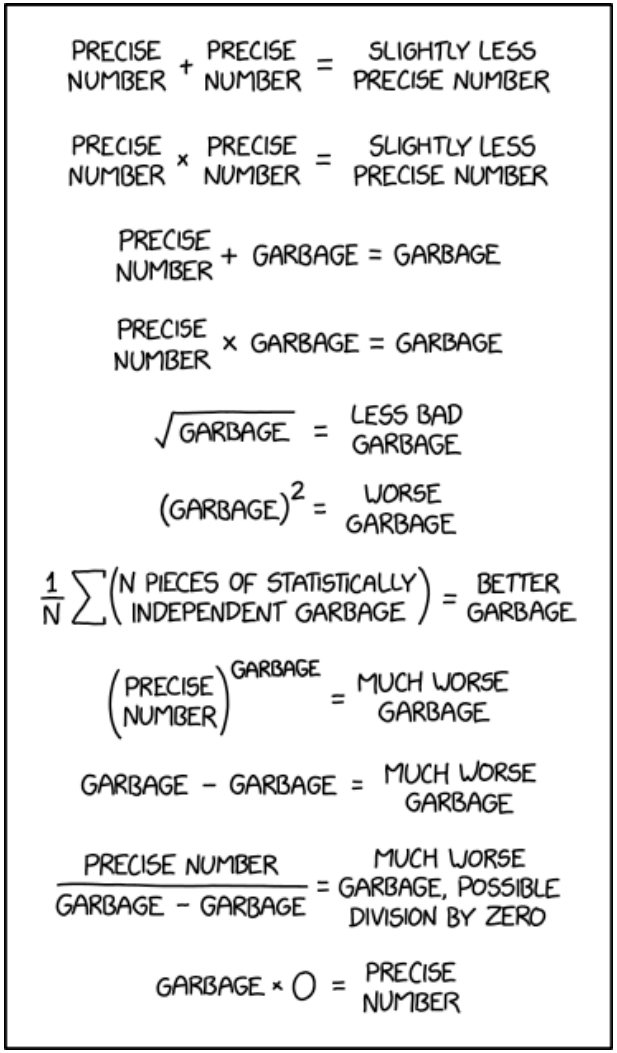
ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

Dimensiones de la calidad de los datos

”Garbage in, garbage out.”

Investigaciones precisas necesitan valores precisos, sin embargo tener datos libres de errores es prácticamente imposible.

La propagación de errores tanto a nivel numérico, como computacional y humano debe ser correctamente manejado para obtener resultados confiables.



Calidad de los datos

Existen iniciativas gubernamentales relacionadas a la calidad de los datos como por ejemplo Gobierno Digital (e-Government), Datos Abiertos, entre otras.

- Si bien los avances en esta área han sido bastantes, aun falta mucho. Temas como la consistencia de los datos y el acceso aun suele generar problemas.
- La digitalización de tramites es de los principales beneficios a la ciudadania.

Calidad de datos

- ¿A qué se le puede llamar datos de calidad?
- ¿Cómo diagnosticar la falta de calidad?
- ¿Cómo evitar esta falta?
- ¿Cómo aumentar la calidad de mis datos?

¿A qué se le puede llamar datos de calidad?

Dividimos la calidad de datos en dimensiones, entre las cuales se encuentran:

- Exactitud
- Validez
- Completitud
- Consistencia
- Uniformidad
- Entre otras...

Diremos que un dato es de mejor calidad mientras más dimensiones de calidad de datos satisfaga.

¿A qué se le puede llamar datos de calidad?

Las dimensiones de la calidad de los datos nos ayudaran a identificar, reparar y reducir la incerteza y confusión cuando tratemos con los datos. Si bien no es una practica obligatoria, sí puede ayudar como el primer paso de la fase de análisis exploratorio de datos.

Las principales dimensiones son las mencionadas anteriormente, pero también existen otras dimensiones ligadas al tiempo de acceso, tamaño de los datos, etc, sin embargo dependen del caso particular de uso.

¿Cómo diagnosticar la falta de calidad?

Para diagnosticar la calidad de un conjunto de datos podemos evaluar las diferentes dimensiones en ellos.

- Exactitud: Contrastar valor real con valor en DB.
- Validez: Contrastar un conjunto de reglas de la definición de los datos en la DB.
- Completitud: Contar el porcentaje de datos faltantes.
- Consistencia: Revisar si las distintas fuentes de datos hacen sentido entre si.
- Uniformidad: Verificar que los datos se encuentren en la misma escala.

¿Cómo evitar esta falta?

Es posible evitar este problema en el momento justo de la creación del dato definiendo un esquema de datos adecuado para organizar la información. Para esto nos ayudaba el modelo entidad relación y SQL.

Ventajas:

- Datos ingresados al modelo verificados por el esquema por lo que gran parte de las dimensiones de calidad de datos pueden cubrirse si es que se diseña de manera adecuada.
- Optimización de espacio en almacenamiento
- Fácil de entender para un externo.

¿Cómo evitarlo?

Desventajas:

- Puede ser difícil ingresar los datos para un externo al equipo de diseño de la base de datos.
- Los modelos mientras más simples son más restrictivos, se deben tomar decisiones al diseñarlo.
- Cambiar el modelo de datos puede ser muy complicado.

¿Cómo aumentar la calidad de mis datos?

La realidad:

Por lo general los datos que utilizamos no los recolectamos nosotros, así que no podemos evitar tratar con problemas de calidad de datos.

Debemos identificar los errores evaluando los datos en las dimensiones de calidad de datos para repararlos. Esto se conoce como limpieza o depuración de datos y tiene el objetivo principal de procesar los datos para dejarlos listos para su utilización.

Exactitud

Definimos la **Exactitud** como la cercanía entre un valor v y un valor v' considerado la representación correcta del fenómeno de la vida real que v intenta representar.

Como ejemplo, si el nombre de una persona es Maximiliano entonces el valor $v' = \texttt{Maximiliano}$ es correcto, mientras que $v' = \texttt{Maximilino}$ es incorrecto.

Pueden identificarse dos tipos de exactitud:

- Exactitud semántica
- Exactitud sintáctica

ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

Exactitud

Llamaremos **exactitud semántica** a la cercanía entre v y su verdadero valor v'.

Exactitud

Llamaremos **exactitud semántica** a la cercanía entre v y su verdadero valor v'.

Hablaremos de **exactitud sintáctica** a cercanía de algún valor v con los elementos en el dominio de ese valor. En este caso el dominio es el conjunto de valores posibles para v.

Exactitud

Llamaremos **exactitud semántica** a la cercanía entre *v* y su verdadero valor *v'*.

Hablaremos de **exactitud sintáctica** a cercanía de algún valor *v* con los elementos en el dominio de ese valor. En este caso el dominio es el conjunto de valores posibles para *v*.

Estas características también pueden ser analizadas a nivel de conjunto (atributo, tupla y relación o tabla). En esos casos nos haremos la pregunta: ¿Qué tan cercanos son los valores de este conjunto con sus valores reales?

Exactitud

Es importante notar que en la exactitud sintáctica no estamos tratando de comparar con el verdadero valor de *v*, entonces en el caso anterior $v' = \texttt{Maximilino}$ podríamos decir que es sintácticamente correcto porque el valor más cercano dentro del **dominio de nombres** es *Maximiliano*. sin embargo es semánticamente incorrecto.

Exactitud

Es importante notar que en la exactitud sintáctica no estamos tratando de comparar con el verdadero valor de *v*, entonces en el caso anterior $v' = \texttt{Maximilino}$ podríamos decir que es sintácticamente correcto porque el valor más cercano dentro del **dominio de nombres** es *Maximiliano*. sin embargo es semánticamente incorrecto.

Pregunta: ¿qué es cercanía? ¿Como definimos que 'Maximilino' y 'Maximiliano' son cercanas?

Exactitud

Id	Nombre	Director	Año	Genero	id_secuela
1	Interstelar	C. Nolan	2014	SciFi	5
2	Avengers: Endgame	Hermanos Russo	2019	SciFi	NULL
3	Avengers: Infinity War	Joe Russo	2018	Sci-Fi	2
3	Avengers: Infinity War	Anthony Russo	2018	Sci-Fi	2
4	Django	C. Nolan	202	Drama	0

Surgen varias preguntas:

- ¿Es exacto el valor *`Interstelar`*?

Respuesta: Depende... de la documentación, del esquema, etc.

ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

Exactitud

Id	Nombre	Director	Año	Genero	id_secuela
1	Interstelar	C. Nolan	2014	SciFi	5
2	Avengers: Endgame	Hermanos Russo	2019	SciFi	NULL
3	Avengers: Infinity War	Joe Russo	2018	Sci-Fi	2
3	Avengers: Infinity War	Anthony Russo	2018	Sci-Fi	2
4	Django	C. Nolan	202	Drama	0

Surgen varias preguntas:

- ¿Es exacto el valor *`Interstelar`*?
- ¿Es exacto el valor *`Hermanos Russo`*?

Respuesta: Depende... de la documentación, del esquema, etc.

Exactitud

Id	Nombre	Director	Año	Genero	id_secuela
1	Interstelar	C. Nolan	2014	SciFi	5
2	Avengers: Endgame	Hermanos Russo	2019	SciFi	NULL
3	Avengers: Infinity War	Joe Russo	2018	Sci-Fi	2
3	Avengers: Infinity War	Anthony Russo	2018	Sci-Fi	2
4	Django	C. Nolan	202	Drama	0

Surgen varias preguntas:

- ¿Es exacto el valor *`Interstelar`*?
- ¿Es exacto el valor *`Hermanos Russo`*?
- ¿Es exacto el valor *`SciFi`* y *`Sci-Fi`*?

Respuesta: Depende... de la documentación, del esquema, etc.

Exactitud

Id	Nombre	Director	Año	Genero	id_secuela
1	Interstelar	C. Nolan	2014	SciFi	5
2	Avengers: Endgame	Hermanos Russo	2019	SciFi	NULL
3	Avengers: Infinity War	Joe Russo	2018	Sci-Fi	2
3	Avengers: Infinity War	Anthony Russo	2018	Sci-Fi	2
4	Django	C. Nolan	202	Drama	0

Surgen varias preguntas:

- ¿Es exacto el valor *`Interstelar`*?
- ¿Es exacto el valor *`Hermanos Russo`*?
- ¿Es exacto el valor *`SciFi`* y *`Sci-Fi`*?
- ¿Es exacto el valor repetido para *`Avengers: Infinity War`*?

Respuesta: Depende... de la documentación, del esquema, etc.

ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

Exactitud

Id	Nombre	Director	Año	Genero	id_secuela
1	Interstelar	C. Nolan	2014	SciFi	4
2	Avengers: Endgame	Hermanos Russo	2019	SciFi	3
3	Avengers: Infinity War	Joe Russo	2018	Sci-Fi	2
3	Avengers: Infinity War	Anthony Russo	2018	Sci-Fi	2
4	Django	C. Nolan	202	Drama	0

Ciertamente:

- Deberíamos elegir un lenguaje (y listarlo en la documentación y metadata), por lo que o *`Interstelar`* o *`Endgame`* y *`Infinity War`* están incorrectos.
- *`Hermanos Russo`* es correcto sin embargo existe un problema de exactitud a nivel atributo (al igual que *`SciFi`* y *`Sci-Fi`*) al considerarse un valor ambiguo con *`Joe Russo`* y *`Anthony Russo`*. Esto también toca el concepto de validez y consistencia que veremos a continuación.

Exactitud

Podemos medir la exactitud de los datos en una tabla contando los errores sintácticos y semánticos, desde el punto de vista de las tuplas podemos hablar de:

- Error de exactitud débil: porcentaje de tuplas que tienen errores sintácticos pero no afectan a la identificación de la tupla con respecto a sus valores reales
- Error de exactitud fuerte: porcentaje de tuplas que tienen errores sintácticos que afectan a la identificación de la tupla con respecto a sus valores reales

Validez

Definimos **validez** como la dimensión de los datos que captura la violación de reglas y definiciones sobre los mismos. Entre ellas encontramos:

- Restricciones de tipo de datos, ej.: entero, texto, decimal
- Restricciones de rango, ej.: $edad \in [0, 120]$
- Restricciones de unicidad, ej: RUT es único para cada persona
- Restricciones de pertenencia, ej: carrera de estudios debe estar dentro de una lista posible de carreras.
- Restricciones de dependencia, ej: el titulo y año de una película determinan al director.

Validez

Si bien hemos pinceleado un poco de como se logra implementar algunas de estas restricciones en una base de datos relacional (SQL). Sin embargo en las bases de datos no relacionales igual pueden ser implementadas.

Esto lo hacemos mediante reglas semánticas, estas las podemos escribir mediante restricciones lógicas en un lenguaje de programación para que nos ayude a encontrar y reparar estos errores. Este problema de localizar y arreglar errores se conoce como *edit-imputation problem*

Validez

Ejemplos de esto son:

- *Si una película A es secuela de B entonces el año de A debe ser mayor o igual que el año de B*. Luego la pelicula con id = 1 y 2 tiene mal este atributo.
- Año de películas debería ser mayor que la fecha de la primera película registrada. Con esto id=4 esta mal.
- Si id_secuela referencia a la tabla de películas entonces la de id=4 esta mal porque no existe el id=0.
- Hermanos Russo y SciFi podrian estar restringidas por pertenencia a otra tabla.

Para buscar tuplas que no cumplan estas condiciones usamos filter en R y WHERE en SQL, en conjunto con Group By.

Consistencia

Diremos que **consistencia** es la ausencia de diferencias cuando se compara dos o mas representaciones de un objeto mediante datos. Esto se puede dar en la misma tabla mediante un cambio no explicitado en la forma de registro del dato, o entre distintas tablas.

Al igual que con la validez podemos escribir restricciones lógicas en un lenguaje de programación para que nos ayude a encontrar y consolidar estos datos en uno solo.

Este dimensión de la calidad de los datos se mezcla con la validez (de hecho en muchos textos se enseñan como uno sola).

Consistencia

Ejemplos:

- Una base de datos de registro de los teléfonos residenciales desde el año 2000 al 2020. Notar el cambio en los códigos de ciudad durante estos años.
- Cambio de nombre de algún colegio de un año a otro.
- Cambios en los códigos de región en la base de datos de colegios. Notar la incorporación de 2 regiones nuevas el año 2007.

Completitud

Cuando hablamos de **completitud** nos referimos al porcentaje de los datos que están almacenados con respecto al total posible. Hay 3 grandes nociones de completitud

- Completitud del esquema: que tan completo es el esquema, es decir tablas y atributos, con respecto a la realidad.
- Completitud de las columnas: que tantos datos no tengo disponibles para cada atributo o columna de una tabla
- Completitud de la población: que tantos datos tengo con respecto a la población total.

Completitud

Para datos relacionales podemos caracterizar la completitud de los datos en funcion del porcentaje de valores 'nulos' que se tienen, podemos definir:

- Completitud de un valor
- Completitud de una tupla
- Completitud de un atributo
- Completitud de una tabla

Completitud

El valor faltante puede aparecer de muchas maneras:

- NA
- ”
- 999
- Etc...

Todo depende de como la documentación defina el valor faltante y como lo lee R o SQL

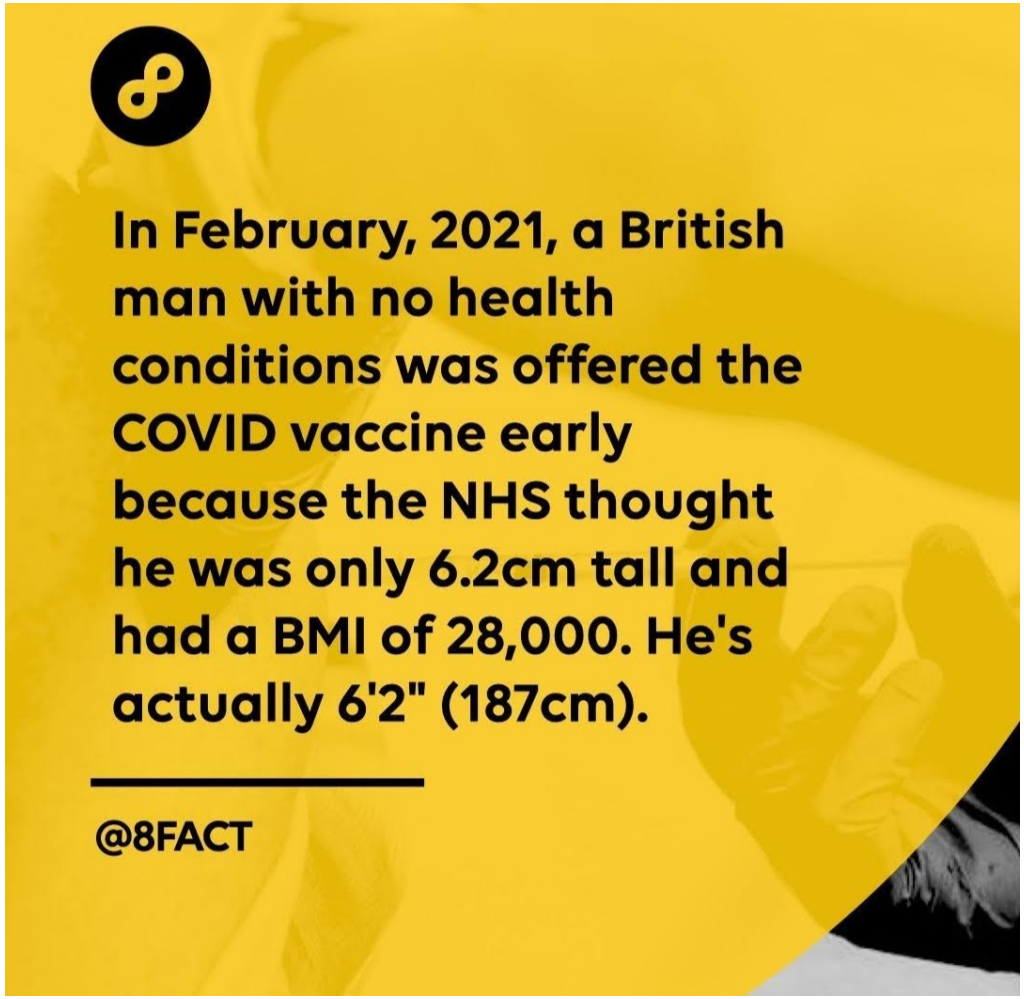
Completitud

ID	Nombre	Apellido	Asistencia	F. Nacimiento
6754	Mike	Collins	29	07/17/2004
8907	Anne	Herbert	18	12/31/9999
6578	Julianne	Merrals	NULL	17/17/2004
0987	Robert	Archer	NULL	NULL
1243	Mark	Taylor	26	09/30/2004
2134	Bridget	Abbott	30	09/30/2004

Otras medidas de calidad de datos

- Uniformidad: ¿Están los datos en la misma unidad de medida?
- Unicidad: ¿Están presentes datos duplicados?
- Actualidad: ¿Se actualizan los datos a tiempo?
- Puntualidad: ¿Están disponibles los datos cuando son necesarios?
- Volatilidad: ¿Qué tan confiables son los datos más actuales?

Tabla de Contenidos


In February, 2021, a British man with no health conditions was offered the COVID vaccine early because the NHS thought he was only 6.2cm tall and had a BMI of 28,000. He's actually 6'2" (187cm).
@8FACT

Ahondaremos en algunos problemas relacionados a la calidad de datos y las herramientas para analizar y preparar estos errores.

- Exactitud: vincular dos valores entre si según alguna medida de similitud.
- Completitud: trabajar con datos faltantes.
- Consistencia: como consolidar datos que se encuentran duplicados con diferente información en uno o mas atributos.
- Validez: filtrar y reparar valores que no cumplan un formato o reglas del esquema.
- Uniformidad: como encontrar y llevar los datos a una escala correcta.

Antes de ver los problemas en profundidad, analizar las características de cada uno y ver la solución especifica veremos un par de técnicas que nos ayudarán a tocar el tema de limpieza de datos de manera mas sencilla.

- Medidas de similitud y vecinos cercanos
- Interpolación
- Regresión Lineal
- Normalizar

Medidas de similitud

Distancia: para un conjunto de elementos X se define distancia o métrica como cualquier función matemática o aplicación $d(a, b)$ de $X \times X$ en $\mathbb{R}$ que verifique las siguientes condiciones:

- No negatividad: la distancia es siempre positiva y solo es 0 cuando se calcula la distancia de un punto consigo mismo, es decir $d(a, a) = 0$ para todo valor a
- Simetría: es lo mismo medir de a a b que de b a a
- Desigualdad triangular: se cumple la siguiente desigualdad

$$\forall a, b, c \in X : \quad d(a, b) \leq d(a, c) + d(c, b)$$

La más clásica es la distancia euclidiana de 2 puntos en el plano (d_2), si $x, y \in \mathbb{R}^2$ y $x = (x_1, x_2), y = (y_1, y_2)$ entonces:

$$d_2(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

Sin embargo ya hemos mencionado algunas otras medidas en el curso:

- Distancia absoluta (Manhattan o d_1) $|x_1 - y_1| + |x_2 - y_2|$
- Distancia del máximo (d_∞) $\max(|x_1 - y_1|, |x_2 - y_2|)$
- Distancia de Levenshtein (distancia de edición): número mínimo de operaciones requeridas para transformar una cadena de caracteres en otra. Implementada en R bajo el comando $adist(s_1, s_2)$ con s_1 y s_2 strings.

Medidas de similitud

Con esto podemos tener una noción de cercanía en el tipo de espacio que estemos usando. Luego podemos definir los vecinos cercanos a un punto (o tupla) a las tuplas que se encuentran cerca de otra.

Vecindad de un punto: diremos que la vecindad de un punto p son todos los puntos que se encuentran a una distancia menor a r, matemáticamente escribimos:

$$V(p) = \{v \in X : d(p, v) < r\}$$

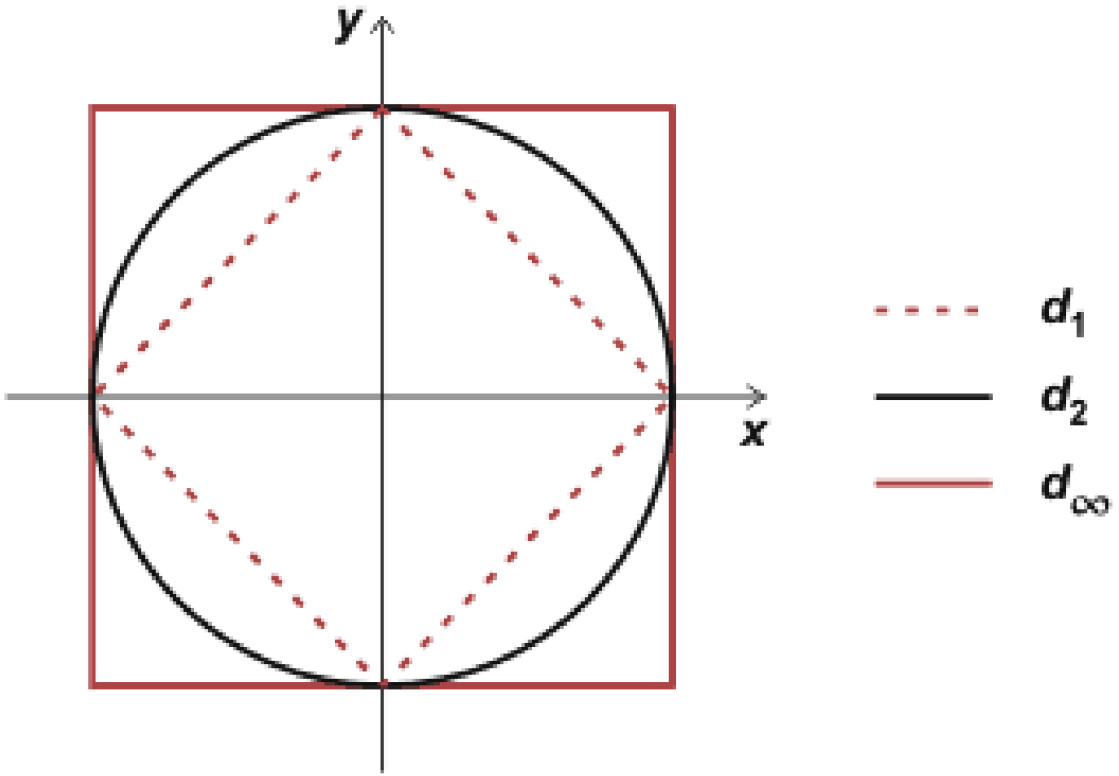


Figura 1: Vecindades a un punto

Ojo: A menos que digamos lo contrario siempre trabajaremos con la distancia euclidiana.

Exactitud

Recordemos que con exactitud nos referimos al grado en el cual los datos describen correctamente algun objeto o evento del mundo real. El problema que podríamos llegar a reparar era el de exactitud semántica, en el cual los datos son contrastados con una lista de valores posibles y se intenta encontrar el match correcto.

Acá el concepto de **Distancia** se hace presente porque nos puede ayudar a:

- Elimina la barrera entre la exactitud semántica y sintáctica.
- Luego de revisar los candidatos a match de cada valor, podemos identificar errores, mejorar la medida de similitud para luego consolidar los valores nuevos

Exactitud

Con las medidas de similitud también podemos definir Join's inexactos (Rolling Join para data.table, sin embargo el concepto no se encuentra masificado).

Esto es fundamental al momento de unir dos bases de datos ya que por lo general aunque estas posean el mismo atributo para hacer el Join, el formato en que el dato se presenta puede ser distinto y eso anularía toda posibilidad de unificar esas tablas.

Exactitud

Problema:

...	Nombre
...	Maximilano
...	Francisc0
...	camila

Cuadro 1: Caption

Nombres
Maximiliano
Rodrigo
Camila
Francisco
Francisca

Cuadro 2: Lista de nombres correctos

Interpolación

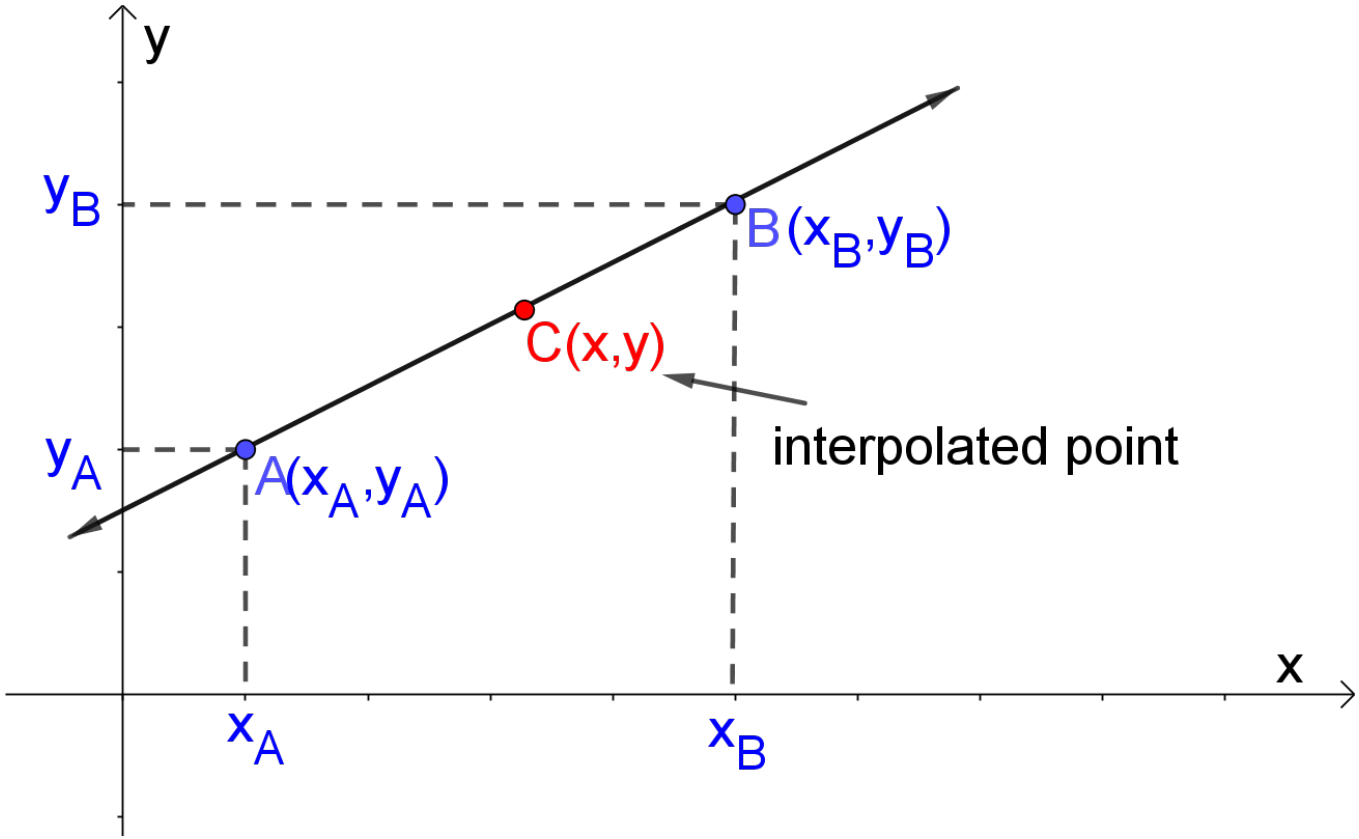


Figura 2: Interpolación en 2 dimensiones

Es mas complicado si tratamos de extenderlo a datos como por ejemplo *strings*.

interpolacion en $\mathbb{R}$

Es el problema de aproximar el valor de una función para un punto no dado en algún espacio cuando se le da el valor de esa función en puntos alrededor (vecinos) de ese punto.

En $\mathbb{R}$

Para el caso de una serie de puntos en $\mathbb{R}$ una de las formas mas sencilla de estimar un punto intermedio seria directamente usar el valor mas cercano, sin embargo esto causa problemas porque vuelve a la secuencia de valores en una función discreta o por partes.

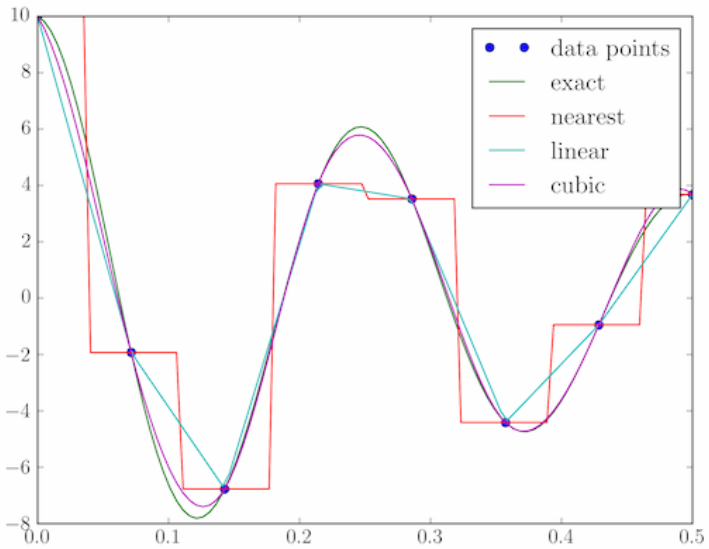


Figura 3: Distintos tipos de interpolación

El segundo método mas sencillo es la interpolación lineal. En la interpolación lineal se utilizan dos puntos, (x_A, y_A) y (x_B, y_B), para obtener un tercer punto interpolado (x, y) a partir de la siguiente fórmula:

$$y = y_A + (x - x_A)\frac{(y_B - y_A)}{(x_B - x_A)}$$

Que viene directamente de la ecuación de la recta. Este metodo es muy simple, genera una serie continua de valores, pero en algunos casos puede no ser muy preciso.

interpolacion en $\mathbb{R}^n$

En 2 dimensiones o más el problema se vuelve mas complejo. Es necesario relacionarlo con el concepto de distancia. El algoritmo de vecino más cercano (**Nearest Neighbor Algorithm**) selecciona el valor del punto más cercano y no considera los valores de los demás puntos vecinos en absoluto, lo que produce un interpolante constante por partes.

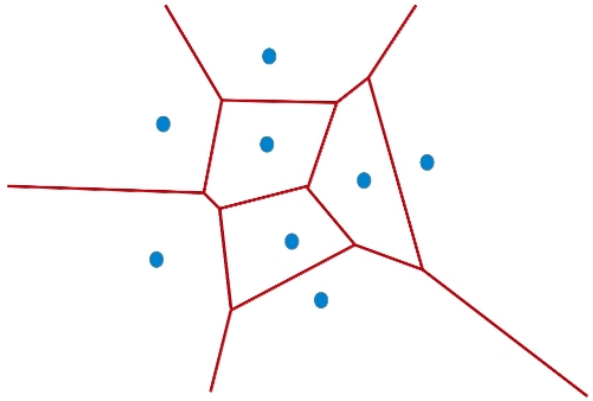


Figura 4: Nearest neighbor classification

El algoritmo es muy simple de implementar y puede utilizarse para asignar variables categóricas.

Pensemos en la región formada por 3 puntos en una región en 2 dimensiones:

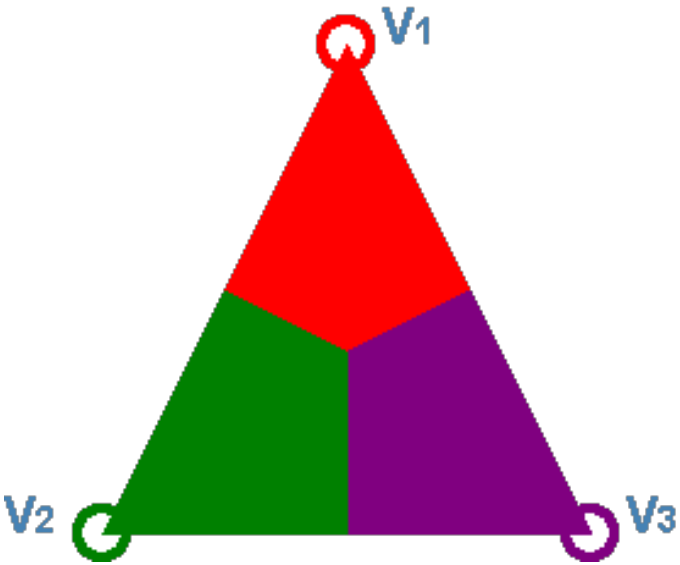


Figura 5: Nearest neighbor classification 3 puntos

El algoritmo del vecino mas cercano nos daría una disposición como la de la figura entre cada trio de puntos cercanos. Es decir, un punto nuevo sin un valor será asignado al valor correspondiente a la region en la que se ubique.

Este algoritmo nos da zonas escalonadas (no es un mapeo suave) que es lo que uno busca en variables categóricas o con dominio discreto (finito). El mapeo que genera se le llama diagrama de Voronoi.

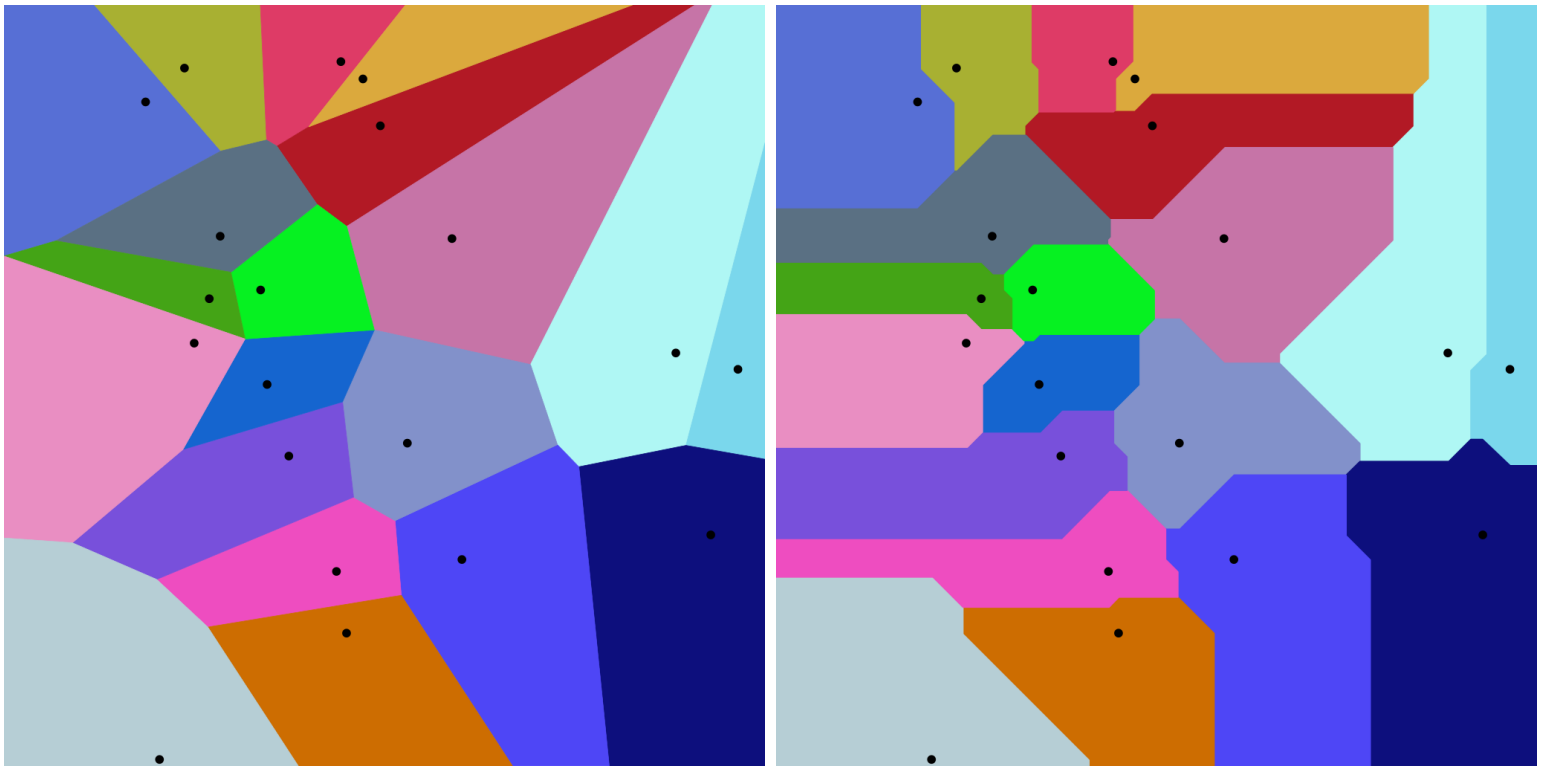


Figura 6: Diagrama de Voronoi para la distancia Euclidiana y la Manhattan

Coordenadas Baricentricas Hacer una Interpolación continua en 2 o más dimensiones en el punto (P_x, P_y) que se parezca a lo que hicimos en $\mathbb{R}$ involucra resolver el siguiente sistema de ecuaciones:

$$\begin{aligned} P_y &= W_{v1} Y_{v1} + W_{v2} Y_{v2} + W_{v3} Y_{v3} \\ P_x &= W_{v1} X_{v1} + W_{v2} X_{v2} + W_{v3} X_{v3} \\ 1 &= W_{v1} + W_{v2} + W_{v3} \end{aligned}$$

Para obtener algun valor $Z_p = W_{v1} Z_{v1} + W_{v2} Z_{v2} + W_{v3} Z_{v3}$

Reordenando podemos obtener los W_{vi} de manera directa:

$$W_{v1} = \frac{(Y_{v2} - Y_{v3})(P_x - X_{v3}) + (X_{v3} - X_{v2})(P_y - Y_{v3})}{(Y_{v2} - Y_{v3})(X_{v1} - X_{v3}) + (X_{v3} - X_{v2})(Y_{v1} - Y_{v3})}$$

$$W_{v2} = \frac{(Y_{v3} - Y_{v1})(P_x - X_{v3}) + (X_{v1} - X_{v3})(P_y - Y_{v3})}{(Y_{v2} - Y_{v3})(X_{v1} - X_{v3}) + (X_{v3} - X_{v2})(Y_{v1} - Y_{v3})}$$

$$W_{v3} = 1 - W_{v1} - W_{v2}$$

Sin embargo buscaremos aplicar algo mas simple.

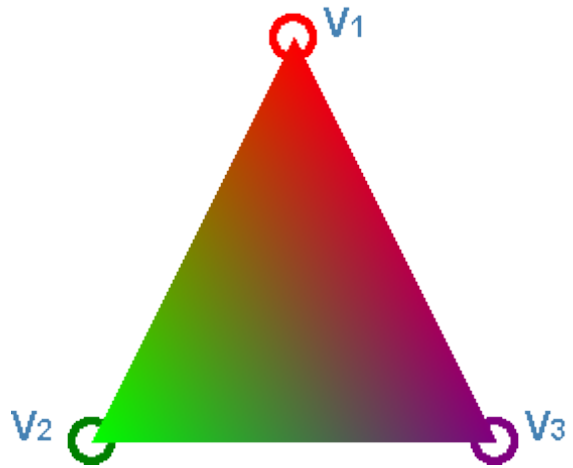


Figura 7: Interpolación baricentrica con 3 puntos

Interpolación de la distancia inversa

Este es un metodo de interpolacion directo que requiere saber la distancia a los puntos cercanos del punto que se quiera estimar, diremos que:

$$W_{vi} = \frac{1}{dist((X_{vi}, Y_{vi}), (P_x, P_y))}$$

Luego podemos calcular:

$$Z_p = \frac{W_{v1}Z_{v1} + W_{v2}Z_{v2} + W_{v3}Z_{v3}}{W_{v1} + W_{v2} + W_{v3}}$$

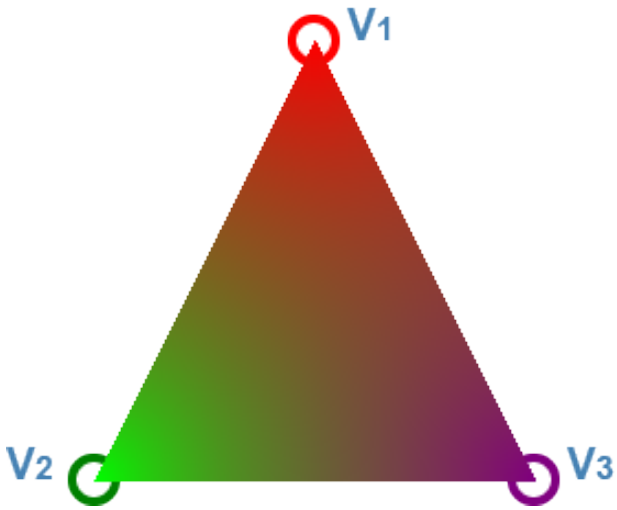


Figura 8: Interpolación de la distancia inversa con 3 puntos

Completitud

	Lat	Lon	Mes	Precipitación
...	-17.7719	-69.7244	1	22.3
...	-18.0808	-69.1383	1	23.4
...	-18.23	-69.6594	1	25.6
...	-18.034	-69.555	1	NULL
...	-18.034	NULL	2	27.6
...	...	...		

ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

Completitud

	Lat	Lon	Mes	Precipitación
...	-17.7719	-69.7244	4	22.3
...	-17.7719	-69.7244	5	23.1
...	-17.7719	-69.7244	6	17.8
...	-17.7719	-69.7244	7	NULL
...	-17.7719	-69.7244	8	12.4
...	-17.7719	-69.7244	9	2.3
...	...	...		

Completitud

Existen muchas maneras para corregir esto y todo dependerá de la persona que este analizando y obviamente del problema en sí. Es responsabilidad de ustedes ver cual es la mejor opcion a aplicar. Las opciones con los datos faltantes podrian resumirse a las siguientes:

- Verificar si es posible encontrar el dato
- Eliminar la tupla (o serie) donde se encuentre el dato faltante.
- Reemplazar
- Estimar o interpolar

Completitud

El ultimo caso es de nuestro particular interés. Se han hecho algunos estudios donde se demuestra que es mejor estimar el dato que dejarlo como nulo, en especial si este esta involucrado en una serie temporal.

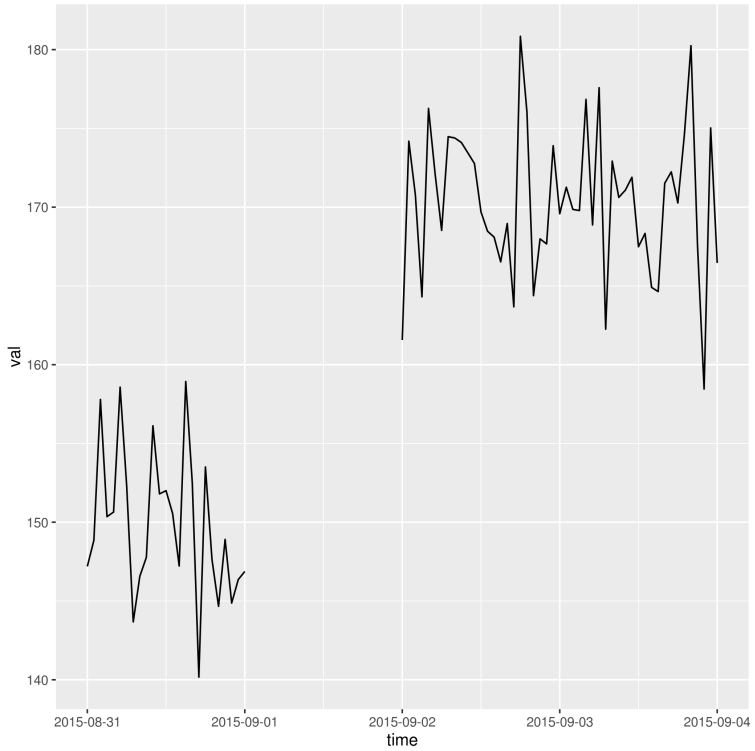


Figura 9: Serie de tiempo con datos faltantes

Completitud

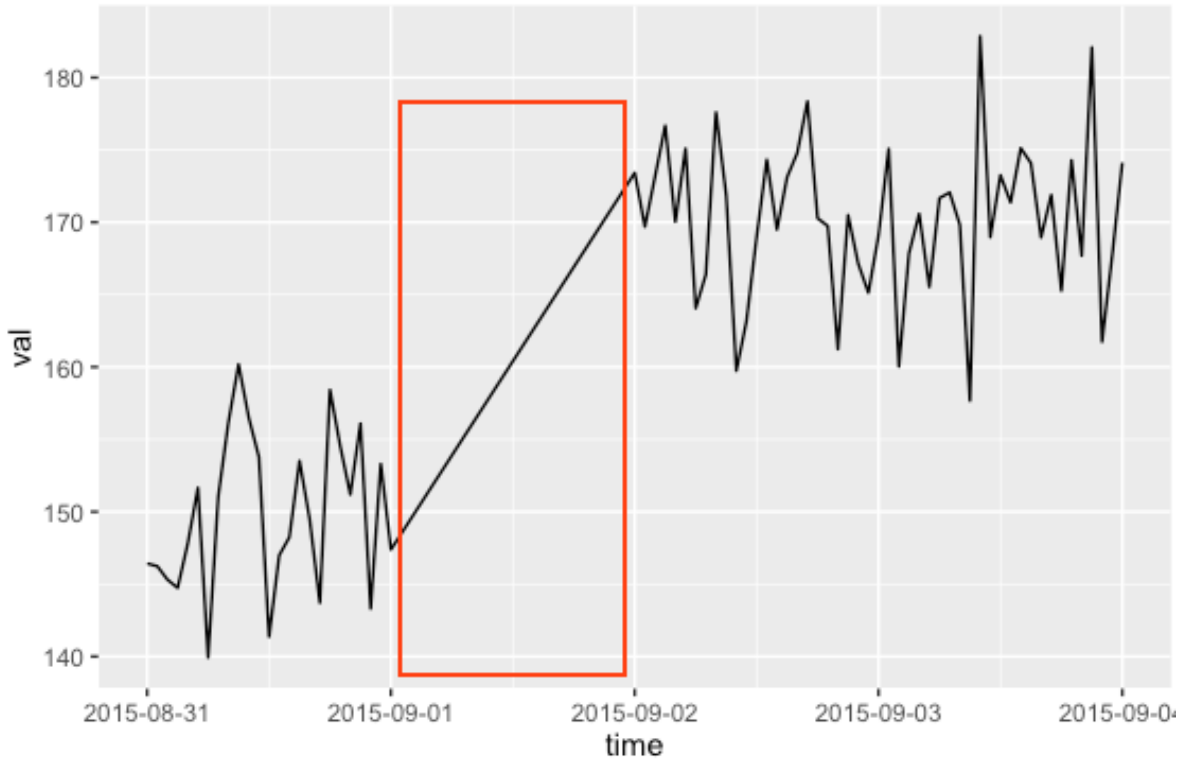


Figura 10: Interpolación lineal

Completitud

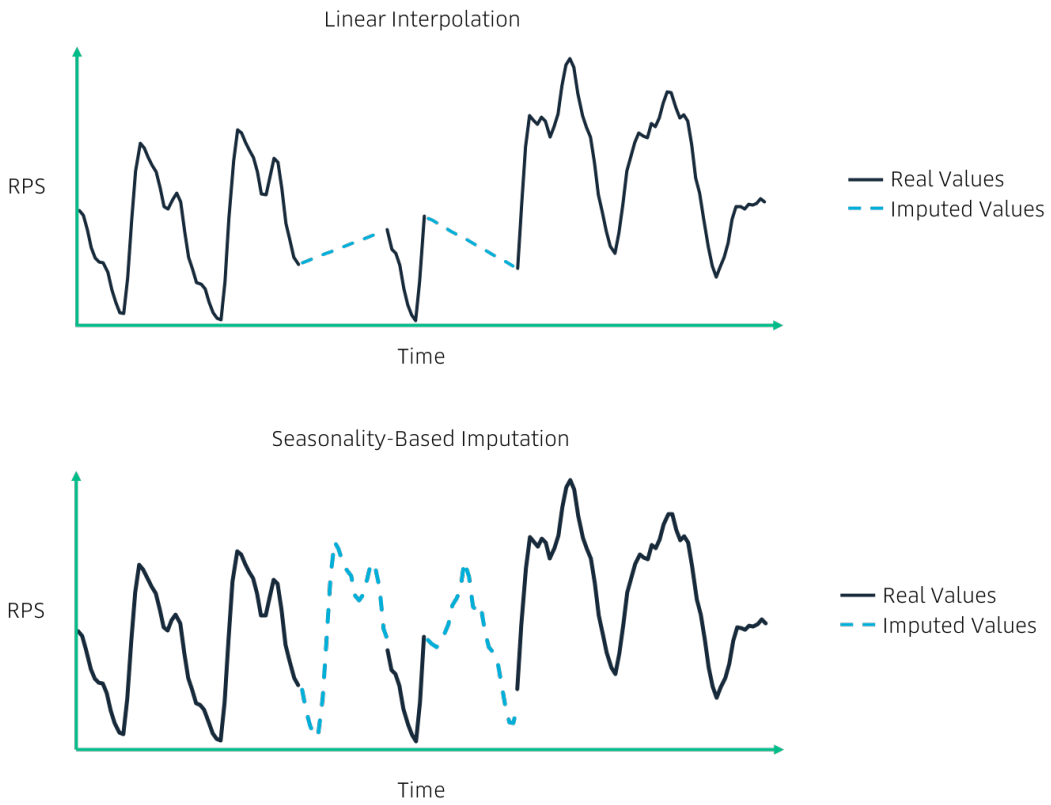


Figura 11: Interpolación lineal en serie periodica

Acá es posible desarrollar un modelo para ingresar los datos faltantes. Tenemos muchos periodos de datos y todos distribuyendo de manera similar por lo que podemos hacer un modelo de regresión lineal que nos entregue el valor esperado del dato faltante dado su pasado.

Normalización

Lo ultimo que veremos será la **normalización**, con esto nos referiremos a justar los valores medidos en diferentes escalas respecto a una escala común. Tomemos como X una columna de datos a normalizar, Podemos distinguir 2 casos importantes:

- Puntaje estándar: si se asume que los datos deberían seguir la distribución de una variable aleatoria normal con media μ y varianza σ podemos estandarizar que nos lleva a una distribución $\mathcal{N}(0, 1)$.

$$f(X) = \frac{X - \mu}{\sigma}$$

Normalización

- Rescalamiento Min-Max (o (0,1)): este escalamiento lleva cualquier conjunto de datos al invervalo (0,1) ideal para datos que ya se encuentran acotados, por ejemplo: evaluaciones, porcentajes, etc.

$$f(X) = \frac{X - \text{mín } X}{\text{máx } X - \text{mín } X}$$

  Gran parte de los modelo de Machine Learning asumen que los datos vienen en el formato Min-Max, por esto el reescalamiento se vuelve algo fundamental para preparar los datos para el uso.

Consistencia

Recordemos que la consistencia se refiere a que no existan diferencias para dos o mas representaciones del mismo objeto o evento en una o mas bases de datos.

A las diferencias entre tuplas que se refieran al mismo objeto las llamaremos conflictos, estos los podemos dividir en:

- Conflictos de atributo
- Conflictos de llave

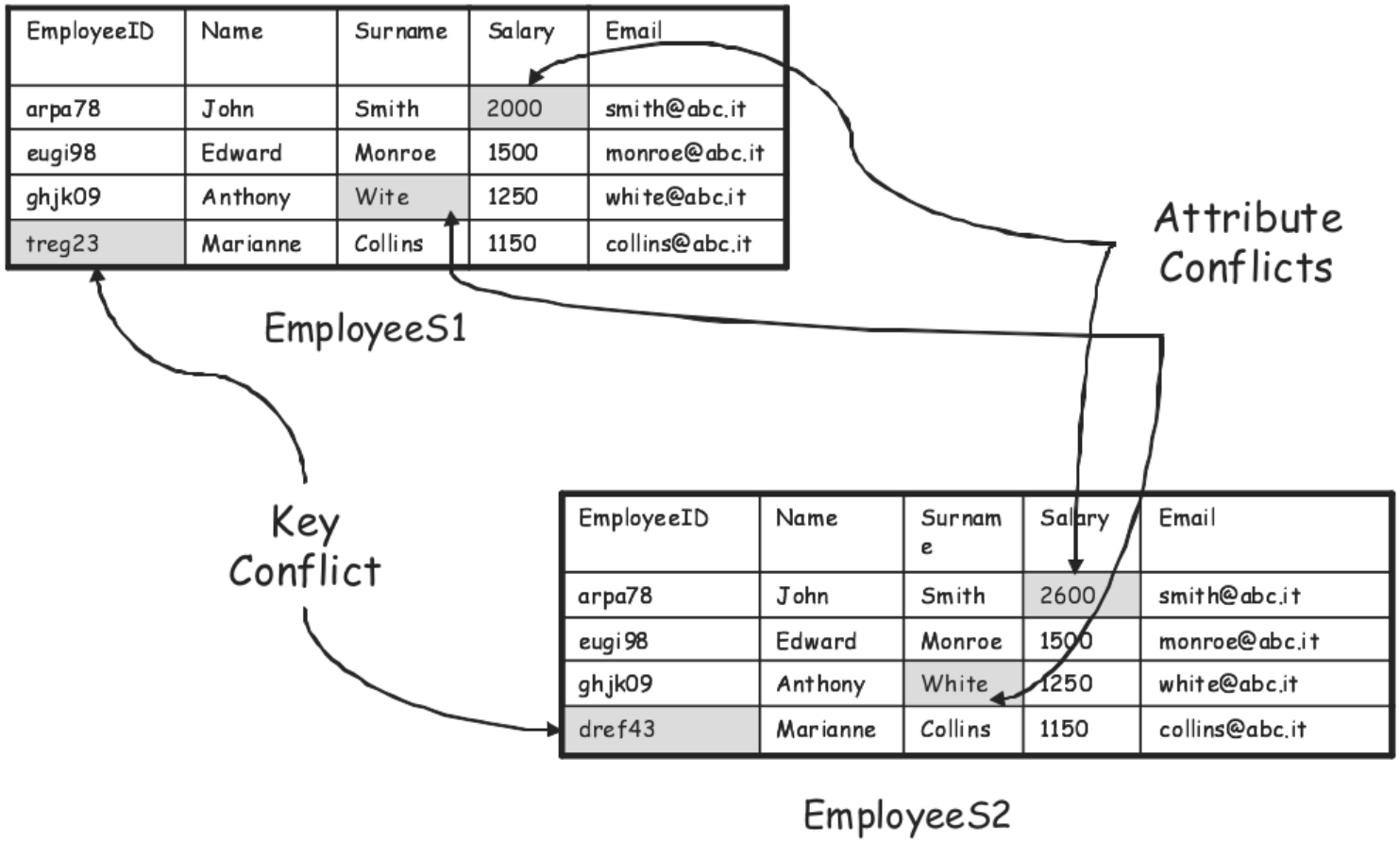


Figura 12: Tipos de inconsistencias

Para resolver los conflictos podemos definir funciones que lo hagan por nosotros. En particular la metodologia consiste en:

- Juntar las tablas en una sola y eliminar los valores duplicados
- Agrupar los datos según alguna llave (o candidato a llave, solo necesitamos que en teoría sea un identificador único de la tupla)
- Contar la cantidad de tuplas en cada grupo. Como seleccionamos en base a una llave las tuplas conflictivas tendrán grupos de mas de un elemento.
- Dados los valores conflictivos, seleccionar o asignar un valor en base a una función de resolución: `max`, `min`, `mean`, `sum`, `random`, `shortest`, etc.

Notese que también podemos agrupar según valores cercanos, sin embargo cuidado con eliminar tuplas distintas entre si.

Validez

Recordemos que los datos son validos si están conformes a su definición en rango, formato, tipo, etc.

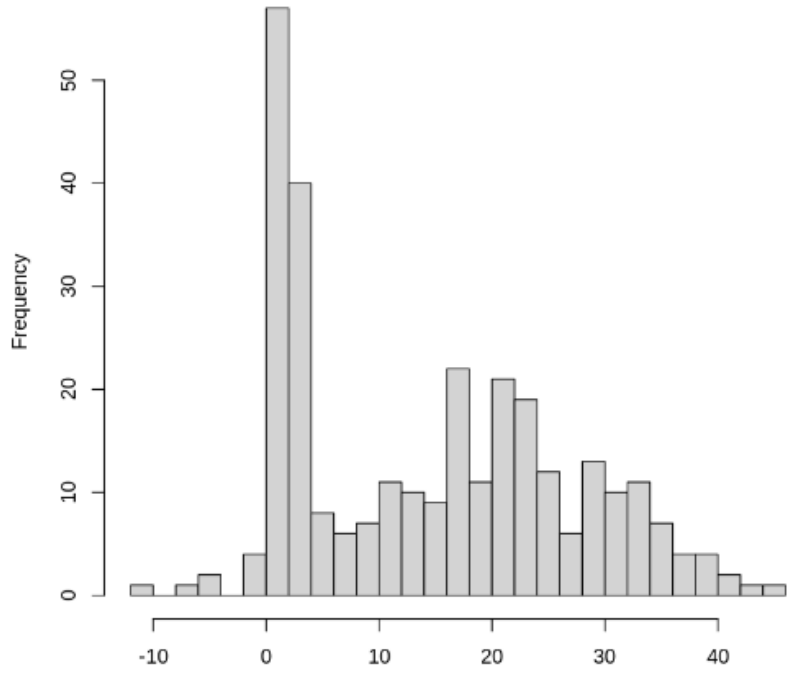
Similar al caso pasado, no existe una herramienta que nos enumere todos los posibles problemas de definición, porque estos son específicos del problema o proyecto. Sin embargo:

- Las medidas de similitud nos ayudan a reparar problemas de dominio (un valor que debe pertenecer a ciertos valores específicos)
- Podemos definir y filtrar según reglas como mencionamos la clase pasada para ver si la validez se mantiene
- En R podemos reemplazar los valores según alguna función de resolución de problemas o analizando los errores en detalle.

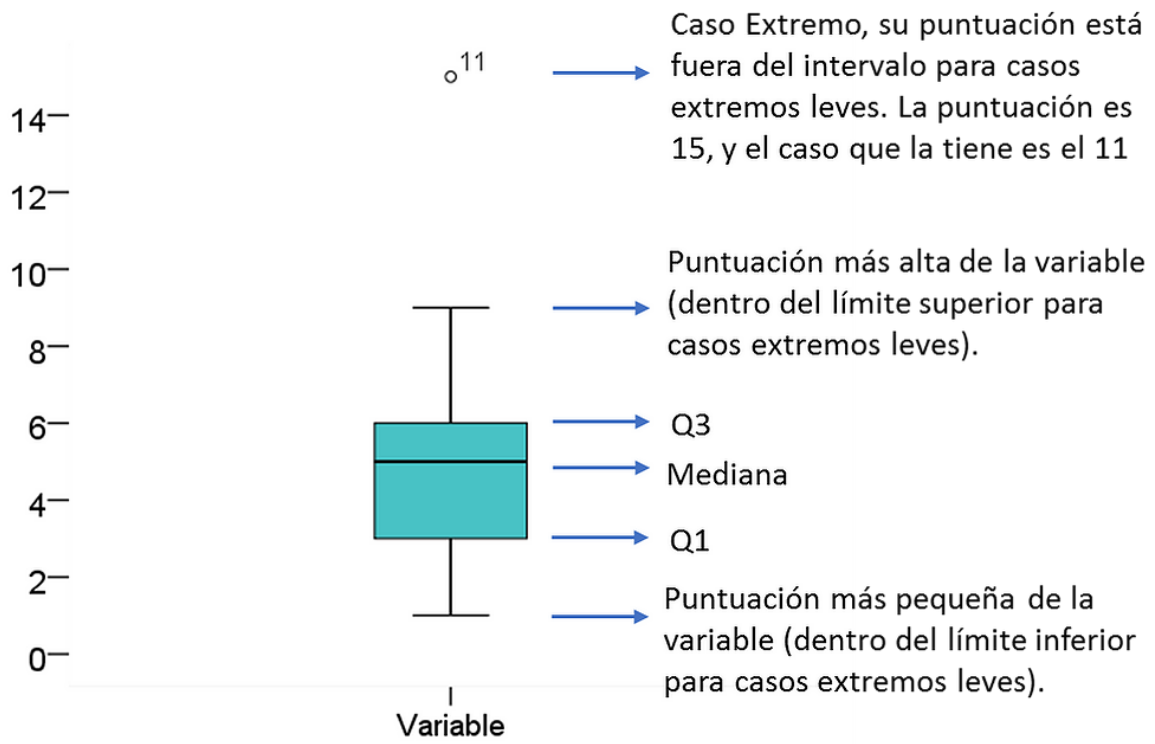
Uniformidad

Recordemos que la uniformidad se refiere al hecho de que todos los atributos se encuentran en la misma escala o unidad de medida.

Notar que un error en la escala debería volver una parte del conjunto de datos proporcionalmente mas grande o pequeño según la escala. Si los datos se encuentran distribuidos según alguna ley de probabilidad, entonces con el histograma (que cuenta la frecuencia de cada rango de valores) podríamos detectar 2 grandes masas de valores.



Uniformidad



También podemos guiarnos por el gráfico de caja que nos habla sobre la distribución de los datos.
Mientras mayores sean los errores en escala mas outliers (valores atípicos) aparecerán.

Uniformidad

Los problemas de escala los resolveremos normalizando los datos de manera adecuada. Dependiendo del problema y de las fuentes de datos la normalización se podría llegar a realizar según grupos, por ejemplo:

- Dos fuentes de datos distintas cuyos valores distribuyen normal podemos estandarizarlos tomando las medias y varianzas locales para cada tabla.
- Si los encargados de registrar algún valor son distintos y el criterio de cada uno no es claro puede resultar conveniente normalizar según las notas que ponen