CALIDAD DE DATOS E INFORMACIÓN:
aplicación a la música y películas

Lic. NAVADIAN | Lic. ROBAINA
Grupo de posgrado 10

Noviembre 2023

Índice general

INTRODUCCIÓN

En este informe, se realiza la tarea de análisis de calidad de datos bajo dos formas de trabajo diferentes sobre dos conjunto de datos distintos.

En una primera parte del informe, se realiza un análisis básico y no sistematizado sobre un dataset con información de películas; comenzando por un data profiling, seguido por el planteamiento de un modelo de calidad de datos y finalizando con el diseño de una base de metadatos de calidad con la ejecución de las mediciones.

En la segunda parte, se procede por aplicar una metodología más ordenada llamada CaDQM (context-aware data quality management), la cual sistematiza de cierta forma los pasos aplicados en la primera parte, agregando componentes de contexto que sirven para el análisis e instancias de retroalimentación para ampliar el contexto de los datos.

Como conclusiones, se evalúa la metodología de trabajo en cada instancia, comparando y reconociendo fortalezas y debilidades de la aplicación de cada una.

Capítulo 1

PARTE 1

1.1. Introducción

En esta primera parte de la tarea, se analizará la calidad de datos de dos datasets sobre películas. A través del data profiling, se dará cuenta de algunos errores predominantes en el conjunto de datos, que servirán de guía para la especificación del modelo de calidad de datos. Posteriormente, se presentará el diseño de la base de metadatos de calidad y se analizarán los resultados obtenidos.

1.2. Descripción general de la fuente de datos

La fuente de datos para este trabajo son dos bases de datos relacionales (dataset 2 A y dataset 2 B), presentadas en formato "csv" que incluyen datos estructurados referente a películas y su calificación según los espectadores. Se podría suponer que los datos han sido extraídos de sitios web de *streaming* o *reviews*. La información que se posee de los datos se resume en la siguiente tabla:

Columna	Descripción
ID	Es el identificador de las películas.
Title	Es el título de las películas.
Original Language	Es el idioma en el que se construyó originalmente la película.
Release Date	Son las fechas de lanzamiento de las películas en los países respectivos.
Vote Average	Es la calificación que las personas otorgan, en una escala de 0 a 10.
Popularity	Indica cuán popular es la película.
Adult	Indica si la película está destinada a personas mayores de 18 años.
Overview	Es un resumen de la película.
Vote Count	Es la cantidad de votos que ha recibido la película.

1.2.1. Data Profiling

Dataset *a*

En el dataset a observamos que hay 8 variables con distinta información sobre las películas:

- **Unnamed: 0**:

  Esta columna parece haber sido introducida por error, coincide con el índice de la tabla. Cada fila de la base de datos tiene un numero distinto, podría ser interpretado como un índice pero se observan datos repetidos que tienen valores distintos en esta columna.

- **Title**:

  Indica el título. Aparece texto, pocos caracteres (entre 6 y 40)

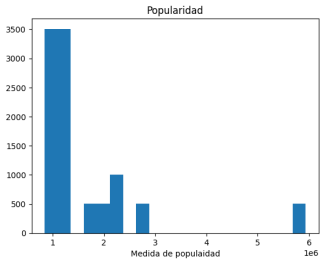


Figura 1.1:

- **Overview**: Incluye un párrafo corto con una descripción, entre 70 y 640 caracteres (En su mayoría menos de 450)
- **Original_language**:

  Lenguaje original, en este dataset aparece solamente aparece ingles como idioma original.
- **Release_date**:

  Fecha en la que se estrenó, tiene distintos formatos de fecha según la pelicula. La mitad aparecen registradas con el año primero y terminan indicando una hora y la otra mitad aparecen como MM/DD/AAAA
- **Popularity**: Popularidad de la pelicula, no indica en que unidades se encuentra pero ronda entre 847514 y 5935577. Sin embargo en su mayoría los datos están por debajo de 3000000.
- **Vote_average**:

  Promedio de evaluaciones con puntajes del 1 al 10.
- **Vote_count**:

  Cantidad de evaluaciones que recibió.

Parecería haber algunas inconsistencias con los metadatos, ya que en estos no aparece indicado la existencia de columna *Unnamed: 0*. Además figura la existencia de una columna booleana que indica si la película es para adultos, y esta no se observa en el dataset cargado.

Dataset *b*

- **Unnamed: 0**:

  Esta columna parece haber sido introducida por error, coincide con el índice de la tabla. Cada fila de la base de datos tiene un numero distinto, podría ser interpretado como un índice pero se observan datos repetidos que tienen valores distintos en esta columna.
- **Id**:

  Parece ser una *forgein key* en esta tabla.
- **Title**:

  Se cuenta con algunos problemas de formato, ya que algunos títulos tienen formato numérico, otros fecha u hora y el resto son strings.
- **Original_language**:

  Lenguaje original, en este dataset se observan 46 distintas lenguas.
- **Release_date**:

  Fecha en la que se estrenó, se observa una única observación con problemas de formato.
- **Popularity**: Popularidad de la pelicula, no indica en que unidades se encuentra.

- **Vote_average**:

  Promedio de evaluaciones con puntajes del 1 al 10.

- **Adult**:

  Dummy que indica si la película es apta para personas mayores de 18 o no. Todos los valores son *False*.

Ambas bases cuentan con algunas entradas en común como lo son el título de la película, la fecha de lanzamiento de la misma, el lenguaje original, la popularidad y la calificación promedio asignada por los espectadores, sin necesidad de que coincidan las películas ni la información contenida sobre las mismas entre un dataset y el otro.

A su vez, existen entradas distintas entre una base y la otra; el dataset A cuenta con una descripción de la película y un recuento de los votos, mientras que el dataset B cuenta con un número identificatorio de cada película y una variable binaria que indica si es apta sólo para adultos o no.

dataset A	dataset B
unnamed 0	unnamed 0
	ID
title	title
overview	
original language	original language
release date	release date
popularity	popularity
vote count	
vote average	vote average
	adult

En una aproximación inicial al sistema de información, se detectó de forma rápida un problema en la repetición de registros en ambas bases de datos:

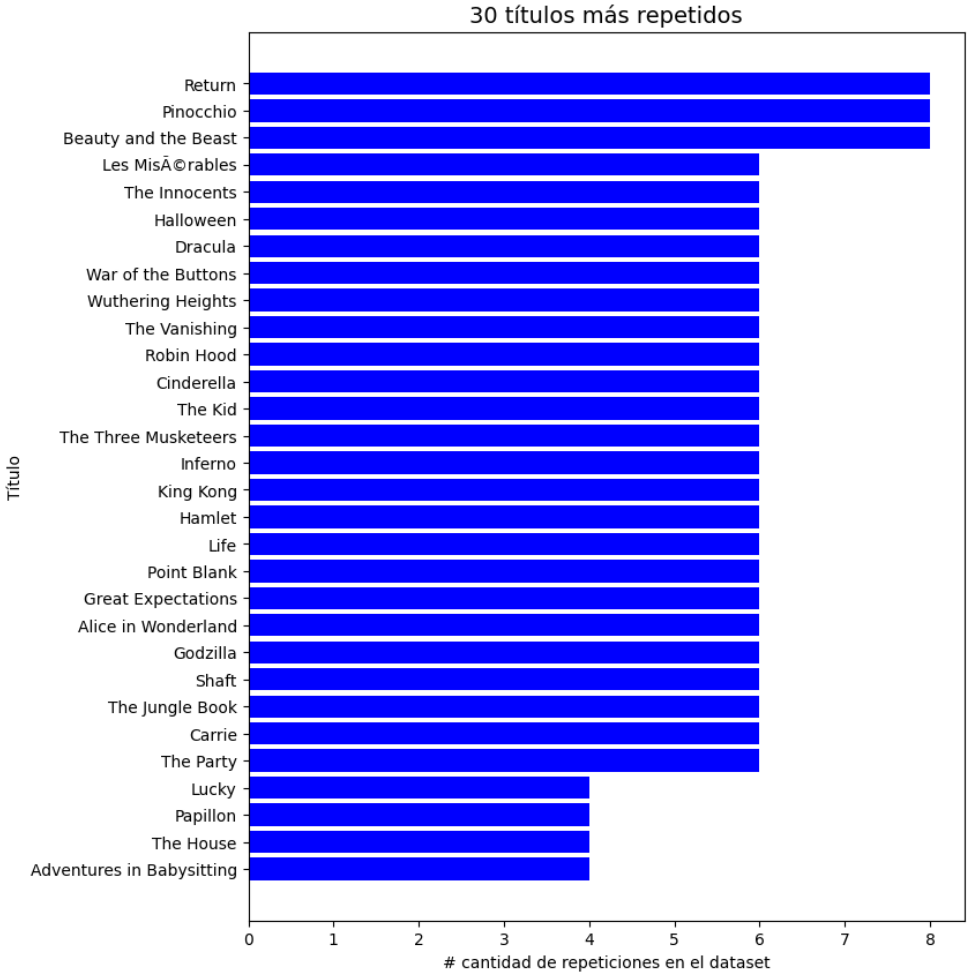
Película	Cantidad de apariciones
Guardians of the Galaxy Vol. 3	501
Fast X	501
Mortal Kombat Legends: Scorpion's Revenge	501
Insidious: The Red Door	501
Extraction 2	501
War of the Worlds: The Attack	501
Elemental	501
A Good Day to Die Hard	501
The Super Mario Bros. Movie	501
John Wick: Chapter 4	501
Spider-Man: Across the Spider-Verse	501
Insidious: The Last Key	501
The Darkest Minds	501
Sheroes	501
Confidential Informant	501
Sound of Freedom	501
San Andreas	501
Transformers: Rise of the Beasts	501
Knights of the Zodiac	501
The Conjuring: The Devil Made Me Do It	501

Cuadro 1.1: "Count" de películas en el dataset A

Claramente, en el dataset A se presenta una incidencia, ya que en lugar de ser 10,020 registros de distintas películas, son los mismos 20 títulos repetidos 501 veces. Se corroboró y no hay diferencias en los títulos u otras variables que justifiquen la reiteración de las observaciones, a excepción de

la variable “original_language”, para la cual el 65.6 % de las observaciones repetidas tienen algún valor y el restante están vacíos.

Respecto al dataset B el patrón no es tan claro a simple vista, si bien todas las películas están duplicadas, hay algunos títulos que se repiten hasta 8 veces, contándose con 9661 títulos únicos. A su vez, se encuentran algunas diferencias entre los registros de un mismo título para los cuales se deberá corroborar la consistencia o si refieren efectivamente a la misma película. Algunos ejemplos de esto pueden ser distintos “original language”, distintas fechas de lanzamiento para el título en un mismo idioma, o diferentes calificaciones.

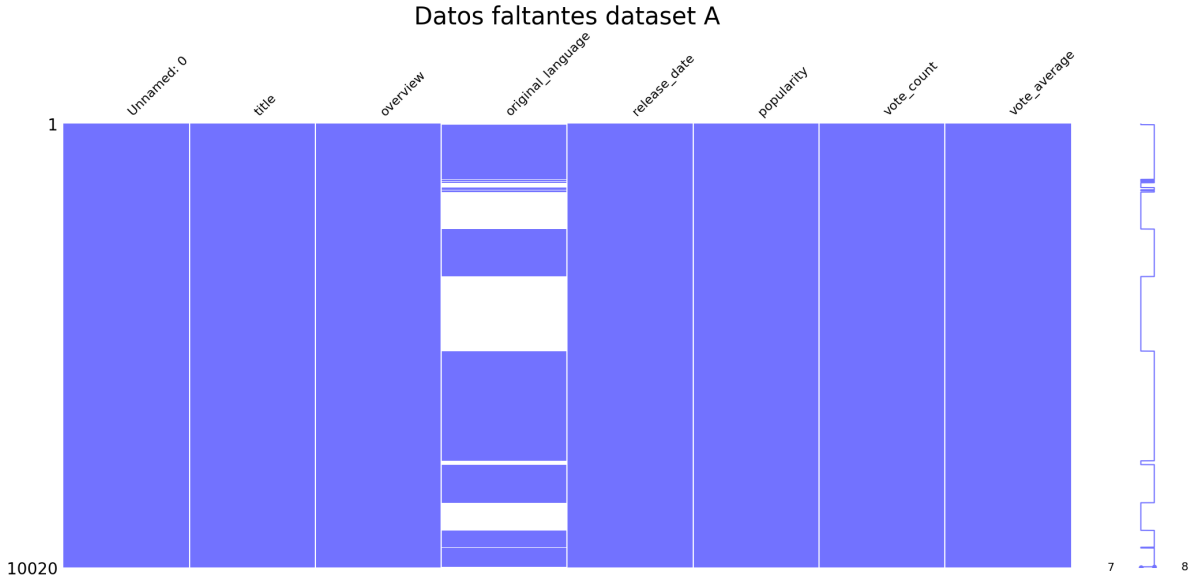


Se observan en la figura anterior los títulos con más repeticiones del conjunto de datos B, y a continuación en la siguiente tabla, se toma el caso particular de “Pinocchio” para ver diferencias entre registros para un mismo título.

Title	Original Language	Release Date	Vote Average
Pinocchio	en	23/2/1940	7.1
Pinocchio	en	23/2/1940	7.1
Pinocchio	en	7/9/2022	6.5
Pinocchio	en	7/9/2022	6.5
Pinocchio	it	11/10/2002	5.8
Pinocchio	it	11/10/2002	5.8
Pinocchio	it	19/12/2019	6.6
Pinocchio	it	19/12/2019	6.6

Volumen y problemas a priori

El dataset a cuenta con **10020 datos**, aparecen **3450 datos ausentes** (34,4 %) en la columna del lenguaje original.

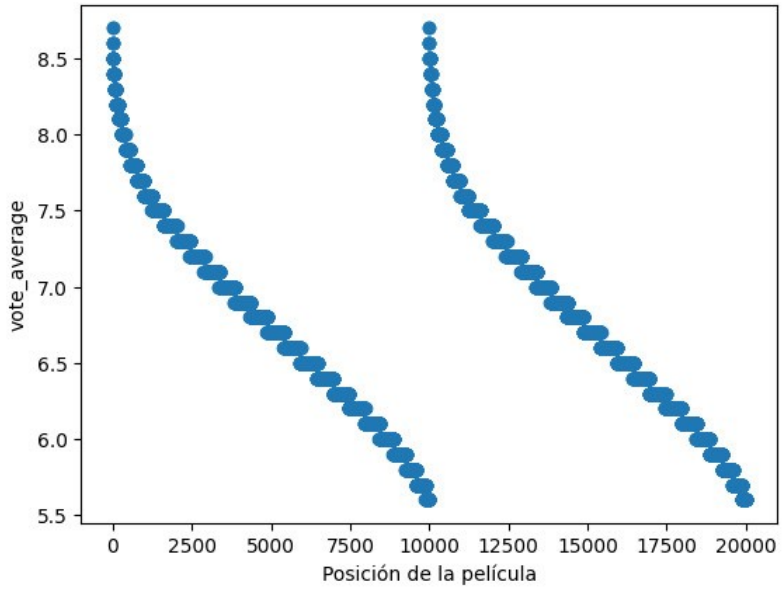


A pesar de que se tenga esta gran cantidad de datos, cuando entramos en detalle a ver cada fila se observa que hay solamente **20 datos únicos**, y cada uno aparece exactamente 501 veces. El problema central del dataset es que está compuesto esencialmente de datos repetidos.

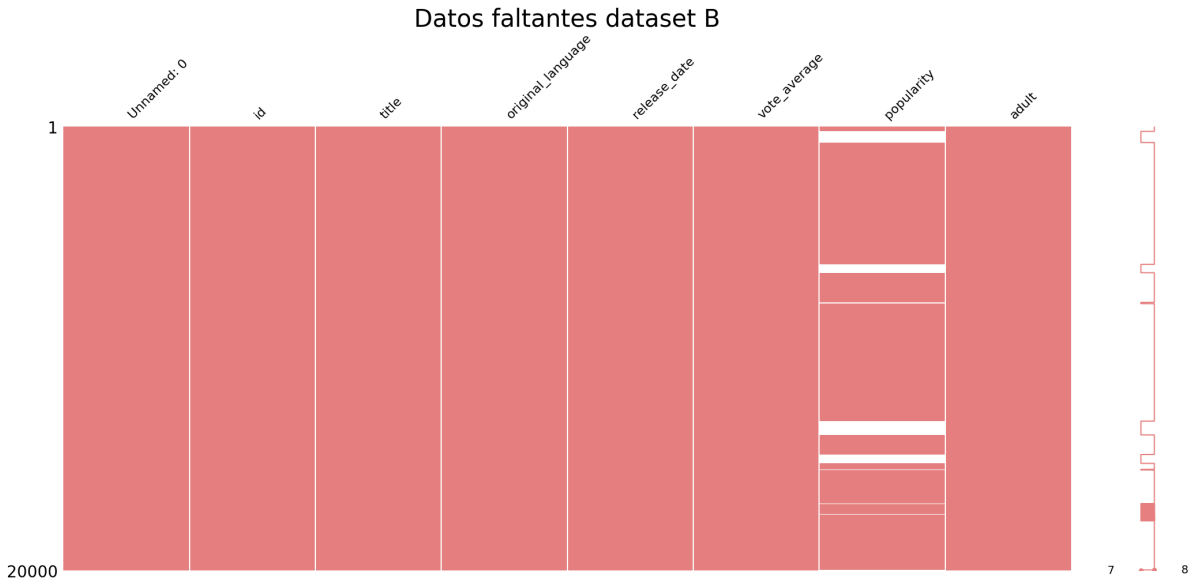
Otro problema que se observa rápidamente es el de la columna de fecha de estreno aprecen dos formatos distintos, y no es posible distinguir cual es la entrada en la que se marca el día y en cual el mes.

Como se ve en la figura a continuación para el dataset b, donde se ven los registros ordenados por su posición en el dataset (número de fila) contra la calificación de la misma, vemos que hasta el registro 10000 se da una ordenación y luego desde el 10001 hasta el 20000 la misma se repite, fuerte indicio de que el dataset está duplicado.

Posición de la película y calificación promedio (dataset b)



Además, presenta un 10.5 % de datos faltantes en la variable *popularity* como se ve en el siguiente gráfico.



En términos generales, se concluye del data profiling que el sistema de información presenta problemas relevantes en la duplicación de registros en los títulos, a su vez, con algunos missing values.

1.3. Especificación del Modelo de Calidad de Datos

Del data profiling, se observó la predominancia de ciertos problemas en el dataset, lo que sirvó de guía para especificar un modelo de calidad de datos acorde a los inconvenientes presentes. Se decidió evaluar las dimensiones exactitud, unicidad, consistencia y completitud, con los factores y granularidad detallados en la tabla:

Dimensión	Factor	Métrica	Granularidad
Exactitud	Precisión	Granularidad	Celda
	Exactitud sintáctica	Bool de exactitud sintáctica	Celda
	Exactitud semántica	Bool de exactitud	Columna
Unicidad	No duplicación	Bool no duplica exacta	Tabla
	No duplicación	Bool no duplica exacta	Tabla
	No duplicación	Bool duplica con Nan	Tabla
	No duplicación	Bool duplica exacta	Columna
	No contradicción	Bool no contradicción	SI
Consistencia	Integridad intra-relación	Bool de consistencia	Tabla
	Consistencia en la representación	Comparación de formato	Celda
		Comparación de tipo de dato	Celda
Completitud	Densidad	Ratio de densidad	Columna

1.3.1. Exactitud

La dimensión exactitud refiere a que los registros no tengan errores y que sean de utilidad como fuente de información. Aplicado al dataset de películas, se optó por los factores de precisión, exactitud sintáctica y exactitud semántica.

Factor precisión

- Métrica: Granularidad

- Granularidad: Celda
- Métrica instanciada: celdas de ds2a[release_date"]

El método se implemento iterando en cada celda de la columna y registrando en la base de metadatos según si el dato tuviera granularidad fecha o fecha - hora.

Factor exactitud sintáctica

- Métrica: Bool de exactitud sintáctica
- Granularidad: Celda
- Métrica instanciada: celdas de ds2a["title"] y ds2b["title"]

Para la medición de esta métrica se utilizó una base de datos del sitio Kaggle[1], que cuenta con 45.000 títulos de películas como diccionario para evaluar la exactitud sintáctica de las celdas de la columna *"title"* . Se itero en cada celda comparando cada título con todos los títulos incluidos en el diccionario. Se registó *True* cuando la pelicula aparecía en el registro y *False* cuando no estaba. Al no tratarse de un diccionario que capture todo el universo es probable que no sea la métrica más relevante a la hora de evaluar la exactitud, pero de todas formas se incluyó por la extensión de la base de datos y la evaluación de dicha base en la página.

Factor exactitud semántica

- Métrica: Bool de exactitud semántica
- Granularidad: Columna
- Métrica instanciada: columnas ds2a["Unnamed: 0"] y ds2b["Unnamed: 0"]

Para la medición de esta métrica se utilizó como referencial también la base de datos descrita anteriormente, comparando esta columna con las columnas incluidas en el referencial.

1.3.2. Unicidad

Para todo el análisis de la unicidad se descartó la columna *Unnamed: 0* ya que contaba con un valor distinto para cada fila y esto naturalmente generaba que cada fila fuese única.

Factor no duplicación

- Métrica: Bool duplica exacta
- Granularidad: Fila
- Métrica instanciada: Filas ds2a y ds2b

Para esta métrica se utilizó la función de pandas en python pd.duplicate() que devuelve *True* para cada fila que aparece repetida en la tabla y *False* en otro caso.

Factor no duplicación con Nan

- Métrica: Bool duplica con Nan
- Granularidad: Fila
- Métrica instanciada: Filas ds2a

Se observo trabajando con los datos que la métrica de duplica exacta no daba una imagen completa de lo que ocurría en el dataset ya que se observaban algunas filas que eran casi duplicas de otras filas de la tabla con la excepción de que en una columna tenían un valor faltante, se entiende que estos datos no son contradictorios sino copias. Para medir esta métrica se utilizó la función

[1] kaggle.com/datasets/rounakbanik/the-movies-dataset/

duplicate pero considerando las columnas que no contaban con datos faltantes, es decir todas menos "*original_language*".

Factor no duplicación

- Métrica: Bool duplica exacta
- Granularidad: celda
- Métrica instanciada: celdas de ds2b["id"]

Para evaluar esta métrica se utilizó la función duplicate aplicada a la base de datos teniendo en cuenta solamente la columna id, es decir que se obtuvo *True* cada vez que un id aparecía repetido.

Factor no contradicción

- Métrica: Bool no contradicción
- Granularidad: conjunto de filas
- Métrica instanciada: Filas con títulos que aparecen ds2a y ds2b en simultáneo

Para esta métrica se tomaron aquellas películas que aparecían en ambas tablas, como una de las tablas cuenta con id y la otra no, se considero la comparación por título. Y dentro de está se evaluó una de los atributos que tienen en común para buscar contradicciones. Para ello se iteró cada título de la tabla a con una comparación booleana con los títulos de la tabla b. Y dentro de las películas que aparecen en ambas tablas se hizo una comparación booleana del atributo *release_date"*.

1.3.3. Completitud

Factor densidad

- Métrica: ratio de densidad
- Granularidad: columna
- Métrica instanciada: aplicando la función isna() de pandas

Para controlar la completitud del conjunto de datos, se toman las filas null o na de cada columna y se calcula la proporción de datos faltantes en el sistema de información.

1.4. Diseño de la base de Metadatos de Calidad

Para la base de metadatos de calidad se consideró una tabla de nombre *dim_fac_met* que contiene los atributos:

- id_metrica
- Dimensión
- Factor
- Métrica
- Granularidad

	id_metrica	Dimensión	Factor	Métrica	Granularidad
0	0	Unicidad	No duplicación	Bool no duplica exacta	celda
1	1	Exactitud	Precisión	Granularidad	celda
2	2	Exactitud	Exactitud sintáctica	Bool de exactitud sintáctica	celda
3	3	Consistencia	Consistencia en la representación	Tipo de dato	celda
4	4	Exactitud	Exactitud semantica	Bool de exactitud	columna
5	5	Completitud	Densidad	Ratio de densidad	columna
6	6	Unicidad	No duplicación	Bool no duplica exacta	fila
7	7	Unicidad	No duplicación	Bool duplica con Nan	fila
8	8	Unicidad	No contradicción	Bool no contradicción	fila
9	9	Consistencia	Integridad intra-relación	Bool de consistencia	fila

Además de una tabla por cada medición de la métrica separada por granularidad.

medicion_celda:

- id_tupla: Que indica la tupla sobre la que se realiza la medición
- atributo: atributo al que corresponde la celda
- id_metrica: Clave foránea a la tabla antes descrita.
- medicion: Resultado del procedimiento

medicion_fila:

- id_tupla: Que indica la tupla sobre la que se realiza la medición
- id_metrica: Clave foránea a la tabla antes descrita.
- medicion: Resultado del procedimiento

medicion_columna:

- id_metrica: Clave foránea a la tabla antes descrita.
- atributo: Atributo a medir
- medicion: Resultado del procedimiento

agregaciones:

- id_metrica: Clave foránea a la tabla antes descrita.
- granularidad: Granularidad luego de la agregación
- descripcion: Breve descripción de la agregación medida
- medicion: Resultado del procedimiento

1.5. Análisis de resultados

1.5.1. Exactitud

Se corroboró la existencia de los títulos para un 10 % del dataset a y un 65 % del dataset b respecto al dataset de referencia. Estos valores son extremanadamente bajos para un dataset que más allá de los problemas que presenta pueda ser utilizable para muchos fines. Teniendo en cuenta que la coincidencia caractér a caractér dificulta el éxito de la medición y hace al método utilizado poco fiable, no se eliminarían títulos por arrojar “False” en esta medición (no coincidir con el dataset de referencia).

1.5.2. Completitud

El dataset a tenía la variable original language con datos faltantes, concluyendose en un 34.4 % de registros vacíos para esta columna. En el caso del dataset b, donde la única variable con missings era popularity, se llegó a que un 10.5 % de los registros tenían datos faltantes. En el sistema de información fue un común denominador que los datos faltantes estén en alguna otra fila que se consideró la "original" del dataset, por lo que no se tomaron otras consideraciones para el tratamiento de estos NULL.

1.5.3. Unicidad

Una vez validada la exactitud de los títulos y precisión de alguna variable relevante, se procedió por analizar el principal problema de este dataset que es la condición de unicidad de los datos. Si bien tanto en el dataset a como en el b se cuenta con un atributo llamado *Unnamed: 0*, el mismo no parece ser una clave primaria razonable, por lo que se decidió eliminar esta columna como se describió en la definición del modelo de calidad.

De un primer control de no duplicación para el dataset a, se desprendió que hay 40 filas originales y 9980 copias de alguna de estas filas. Cuando se procedió a analizar el mismo control de duplicas pero sin tener en cuenta los NA's (que correspondían a la columna *original language*) en realidad eran 20 títulos originales y 20 repeticiones con null en el idioma. Se decidió agregar esta medida como el recuento de la cantidad de filas originales, obteniendo un resultado del 0.2 %.

Por otra parte en el dataset 2b, se llegó a que había 12093 filas originales y 7907 copias exactas de alguna fila. A su vez, como en esta tabla contaba con una columna ID, se analizó la unicidad y no contradicción de los datos, es decir, que no haya un mismo ID con distintos datos. Y aunque a aparecen algunos datos igual id que parecerían estar en contradicción se observa que ocurre como se mencionó anteriormente, estas filas eran copias exactas de otras pero con un nan en algún atributo. Por lo tanto se concluyó que los id son únicos si no tenemos en cuenta este error en la completitud los datos.

Si en el dataset 2b se realiza un recuento por títulos, se ve que se da una contradicción en la fecha de lanzamiento, lenguaje original o incluso la calificación para un mismo film. Sin embargo, realizando un recuento de los 20000 ID, se observa que 10000 son únicos, entonces el problema de no contradicción no sería real ya que se referirían a distintas películas con el mismo título. El resultado de esta métrica agregada como el porcentaje de filas únicas incluyendo contradicciones es 50 %.

Otro control de no contradicción se realizó a través de los dataset a y b, cruzando las películas que coincidían entre uno y otro y verificando si la información contenida era la misma. Se llegó a que 20 películas

Capítulo 2

PARTE 2

2.1. Introducción

En esta sección del informe, se analizó la calidad de los datos de un dataset de canciones provenientes de Spotify. Para ello se utilizó la metodología Context-aware Data Quality Management (CaDQM), metodología que parte del contexto de los datos para un mejor modelo de calidad de los datos enfocado en las necesidades de los usuarios.

A partir del contexto dado, se realizó el data profiling centrado en las reglas de negocio y requerimientos de los usuarios, para luego de identificados los principales problemas de los datos, tener un intercambio con los usuarios de estos datos en Spotify, identificando así nuevos componentes de contexto y ordenando las prioridades de los problemas hallados para luego pasar al modelo de calidad de datos.

Cabe destacar que **no** se incluyeron en el modelo de calidad de datos **todos** los problemas hallados, sino los que se consideraron más relevantes dado el contexto y el orden de prioridad. Se tuvo en cuenta la 'materialidad' de los errores observando qué porcentaje del dataset presentaba dicha incidencia y definiendo umbrales.

Por último, se realizó la ejecución de las mediciones sobre las dimensiones y factores planteados en el modelo de calidad, guardando los resultados en la base de metadatos de calidad, y se analizándose los mismos.

2.2. Descripción general de la fuente de datos

La fuente de los datos de este trabajo es un gran acervo de 114 mil canciones presentado como una base de datos relacional en formato '.xlsx' proveniente de Spotify, la plataforma de transmisión de música online más grande del mundo. El dataset contiene distintos atributos de las canciones, tanto generados internamente como información conocida de las canciones externa a la empresa. Para el mejor entendimiento del sistema de información, se cuenta con los metadatos del dataset resumidos en la tabla a continuación:

Campo	Descripción
track_id	ID de Spotify para la canción
artists	Nombres de artistas (separados por ;)
album_name	Nombre del álbum de la canción
track_name	Nombre de la canción
popularity	Popularidad de la canción (0-100). Calculada por algoritmo basado en reproducciones
duration_ms	Duración de la canción en ms
explicit	Letras explícitas (true/false/desconocido)
danceability	Idoneidad para bailar (0.0 a 1.0)
energy	Intensidad y actividad (0.0 a 1.0)
key	Tonalidad de la canción (número o -1 si no se detecta)
loudness	Sonoridad en decibelios (dB)
mode	Modo (mayor=1, menor=0)
speechiness	Detecta palabras habladas (0.0-1.0)
acousticness	Probabilidad de ser acústica (0.0-1.0)
instrumentalness	Predicción de contenido vocal (0.0-1.0)
liveness	Probabilidad de actuación en vivo (0.0-1.0)
valence	Positividad musical (0.0-1.0)
tempo	Tempo estimado en BPM
time_signature	Firma de tiempo estimada (3-7)
track_genre	Género de la canción

Por otra parte, se cuenta con otros datos, como la información de los Grammy Awards entregados entre los años 1958 y 2019 (el dataset grammy_awards.csv con 4810 registros de la categoría, nominación de cancioón, artista y colaboradores de la misma para cada año del período mencionado) y un referencial de géneros musicales presentado en formato '.txt'.

Por último, como parte de la metodología de este trabajo, se consulto a los usuarios de estos datos (gerentes de Spotify, técnicos, entre otros) para la generación de un contexto de los mismos. Abarcando desde el dominio de los datos, las reglas de negocio, las características y tareas de cada tipo de usuario; y los requerimientos del sistema para cada uno de ellos, desde necesidades de filtrado hasta requerimientos de calidad de los datos.

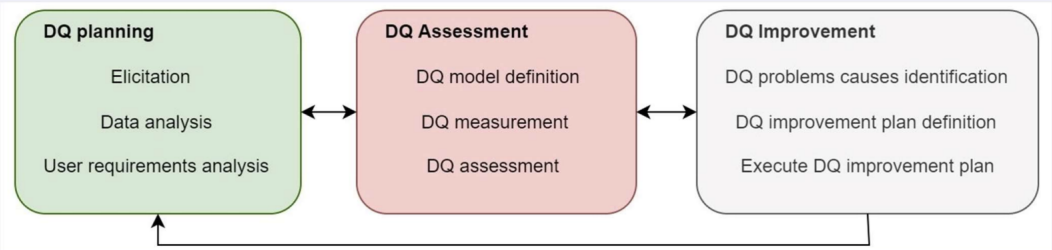
2.3. Ejecución de CaDQM

La metodología Context-aware Data Quality Management (CaDQM) está inspirada en la metodología Comprehensive Data Quality (CDQ) de Batini, la cual se basa en dos fases centrales que son la evaluación (assessment) y mejora (improvement) con tres etapas cada una, haciendo leve incapié en el contexto en una fase inicial llamada reconstrucción del estado (state reconstruction) y que se ejcuta una sola vez.

Lo que la metodología CaDQM implementa es una mayor consideración del contexto de los datos y la posibilidad de actualización del mismo en etapas posteriores, incorporando una fase inicial de planeamiento (planning) la cual tiene tres etapas:

- Stage 1: data context
- Stage 2: data analysis report
- Stage 3: user requirement report

En este trabajo nos centraremos en las primeras 6 etapas del CaDQM que consisten en los 3 stages antes mencionadso de la fase de planeamiento y los 3 stages de la fase de evaluación (data quality model definition, measurument and assessment).



2.3.1. ST1: Data Context

En esta etapa de la metodología CaDQM, se define el contexto de los datos proveniente de la fuente de los datos previo al análisis de los mismos o el data profiling. El Data Context en este caso fue dado e incluyó requerimientos de calidad informados por los usuarios así como filtros necesarios en los datos, reglas de negocio y otros datos que aplican para todos los usuarios, y los metadatos que informan acerca de los atributos del dataset.

Las reglas de negocio informadas en el contexto en principio fueron 6. Por otra parte, se dividió a los usuarios en dos tipos:

- **User 1:** abarcó las necesidades de roles gerenciales y no informáticos, con el punto en común que sus tareas consisten en la generación de reportes para tomar decisiones basadas en datos.
  - **data filtering needs:** canciones que no sean explícitas
  - **data quality requirements:** sobre las variables 'instrumentalness', 'track_id', 'artist', 'album_name' y 'track_genre'
- **User 2:** incluyó roles técnicos e informáticos, con tareas de clasificación de canciones basadas en características de audio y género.
  - **data filtering needs:** canciones que tengan menor nivel de ruido
  - **data quality requirements:** sobre las variables 'track_genre' y 'tempo'

2.3.2. ST2: Data Analysis Report

Siendo las reglas de negocio entendidas como directrices y restricciones que regulan las operaciones de la empresa, se evaluaron como punto de partida en el data profiling, analizando en primer lugar el cumplimiento o ruptura de las mismas.

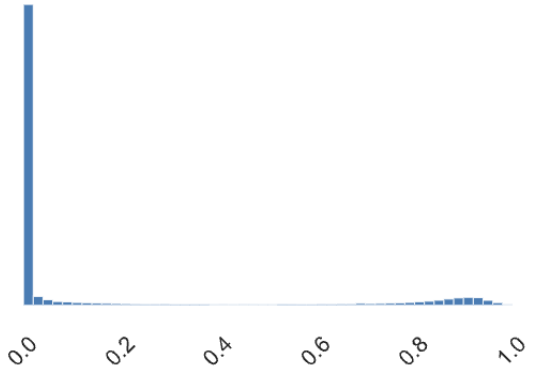
Se presenta a continuación, los casos en los que se notó una violación en las reglas a partir de las distribuciones de los atributos, ya que esto se traduce en un problema en la calidad de los datos.

Posteriormente se analizan los filtros de datos necesarios para cada usuario, los requerimientos de calidad de datos de los mismos, y por último, se presentan otras incidencias en la calidad no asociadas al contexto.

Bussiness Rules

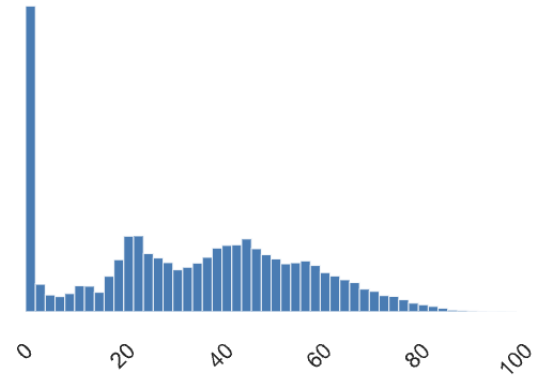
Se desprendió de la distribución, una concentración en 0 para la variable Instrumentalness, lo que estaría rompiendo la Bussines Rule 6 que indica que la misma debe ser mayor a 0.

Instrumentalness



Popularity también presenta una concentración en 0, de suceder lo mismo para valores de danceability ≥ 0.5, se estaría quebrantando la BR4.

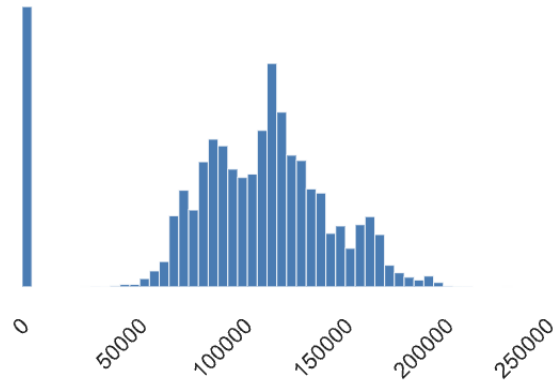
Popularity



Data Quality Requirements

Por otra parte, para el atributo 'tempo' que se asocia al DQR7, se observaron acumulaciones sospechosas en 0, a pesar de que nada indique que no puedan tomar este valor.

Tempo



Otras apreciaciones a partir del data profiling

Por último, y si bien no involucra ninguna regla de negocio, data filerting, ni requerimiento de calidad, se observó que un track tiene duración = 0 lo cual debe ser un error, dado que esto no sería una canción. Aunque en la generalidad esta variable tiene un comportamiento razonable.

Duration

Esta etapa dió una aproximación a la realidad de la calidad de los datos, siendo incluídos en este informe solamente algunos de los atributos y componentes del contexto analizados. Se puede acceder al análisis completo de este stage, en la hoja ST2 del reporte CaDQM adjunto.

Observando el Data Profiling y procesos ejecutados en el ST2, se estima que esta base de datos cuenta con varios problemas asociados a las reglas de negocio y a los requerimientos de calidad especificados en el Data Context.

2.3.3. ST3: User Requirement Report

En la tercera etapa de esta metodología, a raíz del análisis de los datos y el contexto de los mismos, surgieron algunas preguntas para realizarle a los usuarios, actualizar el contexto y validar los problemas.

A pesar de que el data context no incluía ninguna regla ni requerimiento para la variable 'tempo', se vió en los metadatos de la base que la unidad de medida son beats por minuto (bpm). Contrastando esto con los datos, los valores que tomaba el atributo no eran razonables, ya que rondaba entre los 50.000 y 20.000 bpm. Tomando como referencia el metrónomo[1] de una conocida aplicación móvil[2], se observa que una canción puede variar entre 30 y 240 bpm dependiendo de algunas características como pueden ser el género.

El usuario entrevistado estuvo de acuerdo con agregarlo como nueva regla de negocio en el Data Context.

Se incorporaron entonces, a partir de esta etapa, nuevos componentes de contexto que pueden encontrarse junto a las respuestas del usuario entrevistado en la etapa 3 del reporte CaDQM.

2.3.4. ST4: Data Quality Model

Esta etapa consiste en la definición del modelo de calidad de datos, a partir del data profiling (data analysis report) basado en el contexto de los datos.

En primer lugar, se procedió a identificar los problemas y ordenarlos por prioridad como se puede observar en el Data Quality Problem Priorization Report del Reporte CaDQM. Las entradas para esta etapa intermedia fueron los resultados de los stages 1, 2 y 3. Durante la ejecución, se analizó

[1] Instrumento de medición del 'tempo'.
[2] GuitarTuna: `https://yousician.com/guitartuna`

qué porcentaje de los registros rompían una regla o requerimiento, definiendo distintos umbrales para el otorgamiento de prioridades:

- Se asignó prioridad baja si habían menos de 20 % de los registros con problemas de calidad
- Prioridad media si había entre 20 % y 30 % de registros problemáticos
- Prioridad alta si más del 30 % de los resgitros rompían una regla de negocio o requerimiento de calidad

La salida de esta etapa es la identificación, descripción, componentes de contexto, etapa, fecha de reportaje y nivel prioridad de cada problema, que se puede apreciar en el CaDQM como fue antes mencionado. Sólo se evacuará el DQP15, ya que no refiere directamente a ningún componente de contexto en particular.

El DQP15 trata acerca de la contradicción que puede darse a partir de track_id repetidos que tengan algún campo de diferencia entre registros. El id es la clave identificatoria de las canciones en este dataset por lo que, si dos registros con mismo id (que ya es un problema) tienen distinta información, se estaría ante un grave problema de inconsistencia y contradicción, y sería muy costoso reconstruir para cada uno de estos registros cuál es el atributo correcto.

Una vez identificados y priorizados todos los problemas de calidad, esta etapa sirvió de insumo para el Data Quality Model Definition, donde se especifican las dimensiones, factores y métricas a evaluar (ST4 del reporte CaDQM para ver asociaciones entre el Modelo de Calidad y DQP identificados).

Los problemas de rupturas de bussiness rules se englobó en la dimensión de consistencia, debido a que correspondían a lineamientos claros y claves que ponen a prueba la integridad del dominio de las variables. Por su parte, el DQP15 que se detalló en la etapa anterior referente al track_id, es claramente un problema en la dimensión unicidad por los factores de no duplicación y no contradicción.

Por último, el problema referente a 'tempo', correspondió a la dimensión de exactitud, tanto por el factor de exactitud sintáctica en lo que refiere a la unidad de medida de acuerdo a lo identificado en el stage 3, y también contemplando en el contexto del stage 1 específicamente el DQR7, donde se requiere que tenga 3 decimales siendo un claro problema de precisión.

A continuación se resume las dimensiones, factores y métricas resultantes de esta etapa.

Dimensión	**Factor**	**Métrica**	**Granularidad**
Consistencia	Integridad de dominio	Bool de consistencia	Tabla
Unicidad	No duplicación No contradicción	Bool no duplica exacta Bool no contradicción	Tabla SI
Exactitud	Exactitud sintáctica Precisión	Bool de exactitud sintáctica Granularidad	Celda Celda

2.3.5. ST5: Data Quality Measurement

Una vez con el modelo de calidad de datos ya especificado se pasó a diseñar la base de metadatos de calidad. Este paso es fundamental para el almacenamiento de las mediciones de las distintas métricas. Una vez con los resultados obtenidos y guardados es posible hacer varias iteraciones del proceso a lo largo del tiempo y así llevar una trazabilidad de la calidad de estos datos a lo largo del tiempo, lo que nos permite tener una noción real del estado en el que se encuentran además de darnos la posibilidad de mejorar los procesos que los generan.

El diseño de la base de datos cuenta con varias tablas que se relacionan las unas con las otras a traves los id.

La tabla **DQmodel** contiene:

- **id**
- **nombre**

- **descripción**

Contiene información sobre el modelo a utilizar

La tabla **DQdimension** contiene:

- **id_dimension**
- **nombre**
- **descripción**
- **arises_from** : Nos indica de que elemento del contexto surge
- **id_modelo**

Con la información de las dimensiones que se utilizaron. Esta cuenta con la clave **id_modelo** que indica al modelo que corresponde la dimensión.

La tabla **DQfactor** contiene:

- **id_factor**
- **nombre**
- **descripción**
- **arises_from** : Nos indica de que elemento del contexto surge
- **id_dimension**

Cuenta con la información de los factores que serán evaluados de cada dimensión, está vinculado a la tabla de dimensiones a través de **id_dimension**.

La tabla **DQmetrica** contiene:

- **id_metrica**
- **nombre**
- **descripción**
- **granularidad**
- **result domain**
- **id_factor**

Incluye los datos de cada métrica, está vinculada a la tabla de factores a través de **id_factor**

La tabla **DQmetodo** contiene:

- **id_measure**
- **nombre**
- **descripción**
- **tipo de input**
- **tipo de output**
- **algoritmo**
- **id_metrica**

Incluye la información de los métodos a utilizar, está vinculada a la tabla de métricas a través de **id_metrica**

La tabla **DQAppliedMethod** contiene:

- **id app method**

- **type**: Measurement o agregación
- **description**: aplica solo para agregación
- **Use to**: Datos del contexto que se utilizan
- **Applied to**
- **id_measure**

Contiene la información de los métodos a aplicar, con los datos sobre los que deben ser ejecutados los algoritmos de medición. Se vincula con la tabla **DQmetodo** a través de **id_measure**.

Por úlitmo existen las tablas **DQMedicionCeldas**, **DQMedicionesFilas** y **DQAgregaciones**:

- **id_celda**
- **id app method**
- **atributo**: aplica solo para la tabla de celdas y agregaciones
- **medición**

También tenemos **DQAgregaciones**:

- **id_celda**
- **id app method**: Measurement o agregación
- **atributo**: aplica solo para la tabla de celdas
- **medición**

En estas registramos los valores de las mediciones de los métodos referenciados en **id app method**. En caso de que hubiera más de una tabla de datos sería necesario incluir un identificador de la tabla, pero en este caso tenemos cada dato identificado con un id único y eso es suficiente.

Una vez hechas las mediciones se actualiza el contexto de los datos ya que los metadatos de calidad son nuevos metadatos ingresados.

2.3.6. ST6: Data Quality Assessment

Para está etapa se tomará como entrada las mediciones de calidad y los datos del contexto actualizados con los metadatos de calidad. Con toda esta información se agruparán las métricas evaluadas y se dará un valor cuantitativo de calidad.

En este caso se eligió optar por agrupar las mediciones según el contexto de los datos provisto, es decir se juntaron las reglas de negocio para evaluarlas en conjunto al igual que los data requirements. A estos últimos también se le agregó la evaluación de no contradicción ya que su información complementa la de unicidad de los id.

Se observó que las reglas de negocio reportaron un error bastante grande y se evaluaron con un valor cualitativo *bastante malo*. Los requerimientos de calidad también reportaron un error grande, el único del conjunto que tuvo un resultado razonable es la evaluación de ids no duplicados con 79 % pero mirando el resultado de no contradicción vemos que obtenemos el mismo valor, por lo tanto se puede concluir que los id repetidos están todos (o casi) en contradicción, lo cual es un problema bastante grave. Por este motivo se concluyó que el valor cualitativo sería *bastante malo*.

En el global se ve una base de datos con algunos errores bastante importantes tomando en cuenta el uso que se les da tanto por los usuarios técnicos como los gerenciales.

Capítulo 3

CONCLUSIONES GENERALES

En lo que respecta a la primera parte de la tarea, donde se contó con dos datasets sobre películas, nos pareció que los mismos presentaban problemas de calidad relacionados a la fuente generadora de los datos, como pueden ser la duplicación de registros y a la falta de columnas en los dataset respecto a los metadatos presentados. Personalmente, como equipo tuvimos problemas de tiempos, lo que nos retrasó fuertemente todo el trabajo y no se pudo llegar a impactar la totalidad del análisis realizado en el informe, pudimos haber logrado un trabajo más detallado e incluir algunas dimensiones que llegamos a considerar e incluso programar.

En cuanto a la segunda parte, los problemas de calidad y el abordaje de los mismos fue mucho más claro y ordenado gracias a los datos de contexto incorporados. En lo personal, al habernos atrasado con la primera parte, contamos con menos tiempo para la segunda, pero el hecho de que la metodología nos ordenara y guiara la forma de trabajar, logramos aprovechar de mejor manera el poco tiempo y lograr mejores resultados que se ven reflejados en el hecho de que el informe cubre la totalidad de las ideas que fueron implementadas en código y que en general logramos tener un orden más claro a la hora de realizarlo.

En esta última instancia de la tarea también se cometieron algunos errores, como el hecho de no haber realizado y diseñado la base de metadatos de calidad en MySQL. Esto se debió principalmente a que no hicimos ni conocíamos la materia de bases de datos que está en la grilla de la carrera en computación, ya que venimos de otras carreras y esta materia la estamos haciendo en el marco de la Maestría en Ciencia de datos. Identificamos este paso como uno fundamental ya que es lo que permite tener un registro histórico de los datos y de esa forma poder metódicamente evaluar y mejorar la calidad.

En términos generales, creemos que de la primera a la segunda parte de la tarea, tuvimos un profundo aprendizaje y entendemos que ambas instancias de trabajo fueron fundamentales para sacar provecho al curso. Por su parte, en la tarea 1 se pusieron en práctica los conceptos claves de la Calidad de Datos en los Sistemas de Información como los son el data profiling, el modelo de calidad de datos y diseño de la base de metadatos. Al aplicarlo de manera no sistematizada, consideramos que nos empujó a replantearnos y entender los conceptos por separado. Para cerrar el curso, las distintas metodologías aprendidas y la aplicación de la metodología CaDQM en esta segunda parte de la tarea, nos acercó al mundo de la gestión de datos y afianzó los conocimientos básicos, pudiendo sistematizar de manera eficiente el trabajo de un analista de calidad de datos.

Capítulo 4

ANEXO

A continuación se presentan las respuestas al cuestionario respecto a la implementación de la metodología CaDQM.

1. Sí, como se mencionaba constantemente en Introducción a la Ciencia de Datos, el conocimiento del dominio es una parte fundamental de cualquier tarea de análisis de datos y el contexto en el marco de la calidad de datos viene siendo algo similar al conocieminto del dominio.
2. De acuerdo con la definición de contexto la parte nos pareció más importante fue la de conocer las reglas de negocio, en el primer caso cuando nos enfrentamos a la base de datos solamente con nuestras intuiciones fue muy difícil ordenar las ideas y buscar puntos que resultaran interesantes para considerar a la hora de generar el modelo de datos. En cambio tener reglas de negocio nos dio un marco desde el cual acercarnos a los datos y tener algunas ideas más ordenadas.
3. No supimos darle un uso demasiado profundo a los datos de los premios, aunque reconocemos que en otros casos es de suma utilidad tener datos extra que complementen. Algo similar nos ocurrión con la información de Data filtering, no nos quedó demasiado claro como utilzarla más que para poner en contexto el trabajo de los usuarios.
4. Tal vez sería importante precisar un poco más las tareas de los usuarios, entendemos que eso es algo que en un contexto real sería útil para poder interactuar sobre todo el ST3.
5. Nos resultó fundamental el contexto a la hora de la etapa 3 y 4 ya que casi toda la construcción de las intuiciones de los errores surgieron del contexto y las búsquedas que este trajo.
6. La parte que nos resultó menos clara fue la de completar la tabla del modelo de calidad de datos, fue confuso la diferencia entre Method y applied Method además de los campos Inlcudes, Use To (no nos pareció un nombre muy representativo, parece que pide describir para qué se usó el método). También al quedar tan particionado era difícil de entender qué era lo que se buscaba medir cuando en el método aplicado, ya que el algoritmo y la descripción se encontraba en una columna, la granularidad en otra, y los datos a los que sería aplicada en otro distinto.
7. La metodología fue muy útil, como ya indicamos en las conclusiones nos ayudó a tener un uso mucho más eficiente del tiempo y logramos obtener resultados más prolijos, además de la salud mental que aporta no tener que estar enredándose con los errores y tratando de inventar métricas y dimensiones de la nada.
8. Si, claramente el proceso es mucho mas ordenado al aplicar una metodología, ofreciendo además la ventaja de actualización del contexto en etapas posteriores
9. Si, en lo personal nos sucedió que trabajando con la metodología CaDQM, logramos avanzar mucho mas rápido y orientados que en la primera parte de la tarea, donde no teníamos tantas

herramientas para saber qué problemas de los datos atacar primero o con mayor prioridad.

10.
- i. Verdadero. Sin el contexto uno quizás querría atacar ciertos problemas que en realidad para los usuarios de los datos no son relevantes.
- ii. Falso. Se puede hacer un modelo de calidad mucho más conciso sin gastar recursos en problemas que no interesa solucionar
- iii. Verdadero. Aunque la profundidad y detalle puede variar por decisión de los analistas de datos, si se compara con la opción sin contexto, se puede dar el lujo de ir mas a fondo con los problemas centrales
- iv. Falso, dado a que hay un contexto para cada dataset, difícilmente el modelo de calidad aplicado en un contexto sea reproducible para otro dataset en otro contexto.