Data Profiling

Calidad de Datos y Big Data en Ingeniería de Software

Alejandro Adorjan
adorjan@ort.edu.uy

Abstract

El objetivo de este trabajo es realizar un informe respecto de las siguientes tareas:

1. Data profiling de tres archivos de datos publicados.
2. Especificación de un modelo de calidad de datos que cubra algounos de esos datos. Se deberá priorizar que aspectos y que datos se van a evaluar. Justificación.
3. Especificación de la base de datos donde se almacenarán las medidas de calidad obtenidas en la medición.
4. Ejecución de la medición de calidad.
5. Evaluación final de calidad. Análisis de resultados de medición. Niveles de calidad esperados.

Palabras clave: data profiling, data quality, big data

Índice

1 Introducción

1.1 Data Profiling

"El proceso de descubrimiento de metadatos se conoce como "data profiling". Las actividades de "data profiling" van desde enfoques "ad-hoc", como subconjuntos aleatorios de datos, formulación de consultas de agregación, hasta la inferencia sistemática de información estructural y estadísticas de un conjunto de datos utilizando perfiles dedicados y herramientas [2]. La elaboración de "data profiling" es el conjunto de actividades y procesos para determinar los metadatos sobre un conjunto de datos determinado. Entre los resultados mas frecuentes se encuentran el establecer estadísticas por columna, como la cantidad de valores nulos y valores distintos en una columna, su tipo de datos o los patrones más frecuentes de sus valores de datos. La creación de "data profiling" juega un papel importante en distintos casos de uso, como ser la exploración, integración, y el análisis de datos [1]. El objetivo de la elaboración de"data profiling" mide la consistencia y precisión de los datos, detectando duplicaciones de los mismos para obtener el valor correcto de los datos que podrían usarse en la toma de decisiones [4].

1.2 Modelo de Calidad

ISO/IEC 25012 define modelo de calidad de datos como "un modelo general de calidad de datos para datos retenidos en un formato estructurado dentro de un sistema informático. Se centra en la calidad de los datos como parte de un sistema informático y define las características de calidad de los datos de destino utilizados por los seres humanos y los sistemas". [1].

2 Reporte

En esta sección se describen los principales puntos establecidos por el objetivo del trabajo.

2.1 Data Profiling de los 3 archivos publicados

2.1.1 Fuente de datos. En el siguiente repositorio https://github.com/aadorian/cibse_taller.git están disponibles todos los fuentes correspondientes al trabajo, en particular los reportes de ejecución del data profiling estan reportados en https://bit.ly/softdataprofiling.

Los metadatos del trabajo asumimos que corresponden a los disponibles en el Ministerio de Turismo [2]

[1] https://iso25000.com/index.php/en/iso-25000-standards/iso-25012
[2] https://www.gub.uy/ministerio-turismo/emisivo

Tabla 1. ISO/IEC 25012

Característica	Definición
Exactitud | Accuracy	Grado en el que los datos tienen atributos que permiten ser leídos e interpretados por los usuarios y son expresados utilizando lenguajes, símbolos y unidades apropiados en un contexto de uso específico. Cierta información sobre la comprensibilidad puede ser expresada mediante metadatos.“ **Exactitud Sintáctica**: cercanía de los valores de los datos a un conjunto de valores definidos en un dominio considerado sintácticamente correcto. **Exactitud Semántica**: cercanía de los valores de los datos a un conjunto de valores definidos en un dominio considerado semánticamente correcto.
Completitud | Completeness	Grado en el que los datos asociados con una entidad tienen valores para todos los atributos esperados e instancias de entidades relacionadas en un contexto de uso específico.
Consistencia | Consistency	Grado en el que los datos están libres de contradicción y son coherentes con otros datos en un contexto de uso específico. Puede ser analizada en datos que se refieran tanto a una como a varias entidades comparables.
Comprensibilidad | Understandability	Grado en el que los datos tienen atributos que permiten ser leídos e interpretados por los usuarios y son expresados utilizando lenguajes, símbolos y unidades apropiados en un contexto de uso específico. Cierta información sobre la comprensibilidad puede ser expresada mediante metadatos.

2.1.2 Archivos de Profiling. Los dataset analizados se encuentran disponibles en el siguiente link:
https://github.com/aadorian/cibse_taller/tree/main/profiling

2.2 Categorías del Software de Profiling

En la Tabla 2 se presentan el tipo de alertas de categorias de alertas de profiling provistos por la librería ydata-profiling versión vv4.1.2.

2.2.1 Resumen de resultados del DataProfiling. Los tipos de datos que reconoce ydata-profiling son: boolean , numerical, categorical, time-series, URL, path, file, image, date y datetime.

En las Tabla 5, 6, 7y 8 se presentan los resultados resumen de la realización de los dataProfiling a las fuentes de datos de emisivos, receptivos y operadores. Para la generación del data profiling se utilizó la versión de ydata-profiling vv4.1.2 https://ydata-profiling.ydata.ai/ que permite realizar el análisis de datos exploratorio https://github.com/ydataai/ydata-profiling.

2.2.2 Contexto del origen de los datos. Según la normativa uruguaya se entiende por *turismo emisivo* a la actividad turística que realizan los residentes del país fuera del mismo y por *turismo emisivo* a la actividad turística que realizan los residentes del país fuera del mismo https://www.impo.com.uy/bases/leyes/19253-2014.

El *turismo emisivo* contiene información de residentes en Uruguay con viajes al exterior, gasto y estadía de los mismos por país o destino del viaje, referente a cada trimestre del año[6]. El *turismo receptivo* contiene información correspondientes a los visitantes que ingresaron a Uruguay, gasto y estadía de los mismos, referente a cada trimestre del año [7].

2.3 Especificación Modelo de Calidad

En la Figura 1 se presenta el modelo de calidad. La especificación del modelo de calidad se basa en la propuesta de Etcheverry et. al [5] originalmente basada a su vez en el concepto de GQM (Goal Quality Metrics) propuesto por Basili [3]. Etcheverry et. al [5] proponen la siguiente categorización:

- **Dimensiones de calidad:** se caracteriza mediante múltiples dimensiones, que ayudan a clasificar los datos. Una dimensión captura un aspecto de alto nivel de la calidad.
- **Factores de calidad**: un factor representa un aspecto particular de una dimensión de calidad, por ejemplo, la precisión de los datos implica corrección semántica,

Tabla 2. Tipos de alertas de ydata-profiling

Alerta Calidad	Descripción
Constant	Column only contains one value
Zeros	Column only contains zeros
High Correlation	Correlations
High Cardinality	Column > 50 distinct values.
Imbalance	Column is highly imbalanced
Skewness	Column's univariate distribution
Missing Values	Column has missing values
Infinite Values	Column has infinite values
Unique Values	All values of the column are unique
Seasonal	Column has seasonal pattern
Non Stationary	Column is a time-series non-stationary
Date	Column contains data-datetime format
Uniform	Column follows a uniform distribution
Constant length	For strings/date/datetimes columns
Rejected	Variable has mixed types
Unsupported	Column can't be analysed
Duplicates	Dataset-level warning > 10 records
Empty	Dataset-level warning no data

corrección sintáctica y precisión de los datos. Pueden haber varios factores para la misma dimensión; cada factor se adapta mejor a un problema o tipo de sistema en particular.

- **Métricas de calidad:** una métrica es un instrumento utilizado para medir un factor de calidad específico, por ejemplo, el porcentaje de datos del sistema que coinciden con los datos del mundo real es una métrica para la corrección semántica. Pueden haber varias métricas para el mismo factor de calidad.
- **Métodos de calidad:** un método es un proceso que implementa una métrica de calidad. Se definen dos tipos de métodos: (i) métodos de medición, que calculan la calidad de un objeto midiéndolo directamente (por ejemplo, contando el número de valores nulos en una tupla), y (ii) métodos de agregación, que calculan la calidad de un objeto compuesto mediante la agregación de valores de calidad de las partes del objeto (por ejemplo, calcular la precisión de una tabla promediando la precisión de sus tuplas). Pueden haber varios métodos para implementar la misma métrica.

3

2.3.1 Especificación del modelo de datos. La especificación del modelo cubrirá los datos de **Operadores**. Se evaluaran la dimensión de **exactitud** y **unicidad**. Las preguntas a formular en general son relacionadas con

- ¿Estos datos son lo suficientemente precisos para nuestras necesidades?

[3]Definiciones disponibles en https://eva.fing.edu.uy/

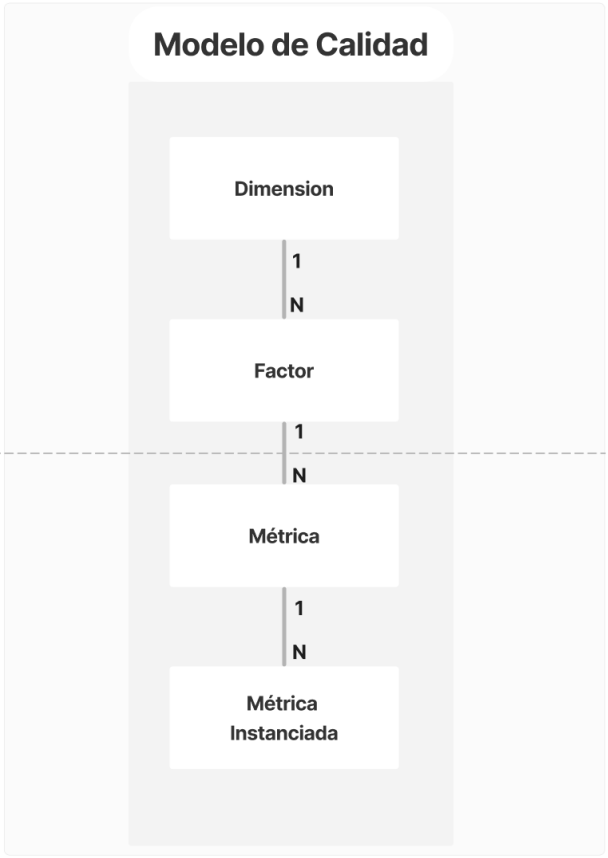

Figura 1. Modelo de Calidad

- ¿El nivel de detalle de los datos es adecuado?
- ¿Estos datos se corresponden con el mundo real?
- ¿Estos datos tienen errores?
- ¿El formato de presentación de los datos es correcto? ¿Es estándar?

2.3.2 Factores a Evaluar. Los tres factores a evaluar en este trabajo son los siguientes:

Factor 1: **Exactitud semántica (Semantic accuracy)**, respondiendo a ¿Los datos se corresponden con la realidad? Eventualmente los datos pueden no corresponden a ningún estado del mundo real y/o a estado equivocado del mundo real y/o con errores en algunos atributos.

Factor 2: **Exactitud sintáctica (Syntactic accuracy)**: respondiendo a ¿Los datos tienen errores sintácticos o de formato?

Eventualmente los datos podrían presentar:

- Errores de valores: Valores fuera de rango, errores ortográficos y de tipeo.
- Errores de estandarización: Valores que no tienen el formato esperado. |
- Valores embebidos: Valores que corresponden a múltiples atributos.

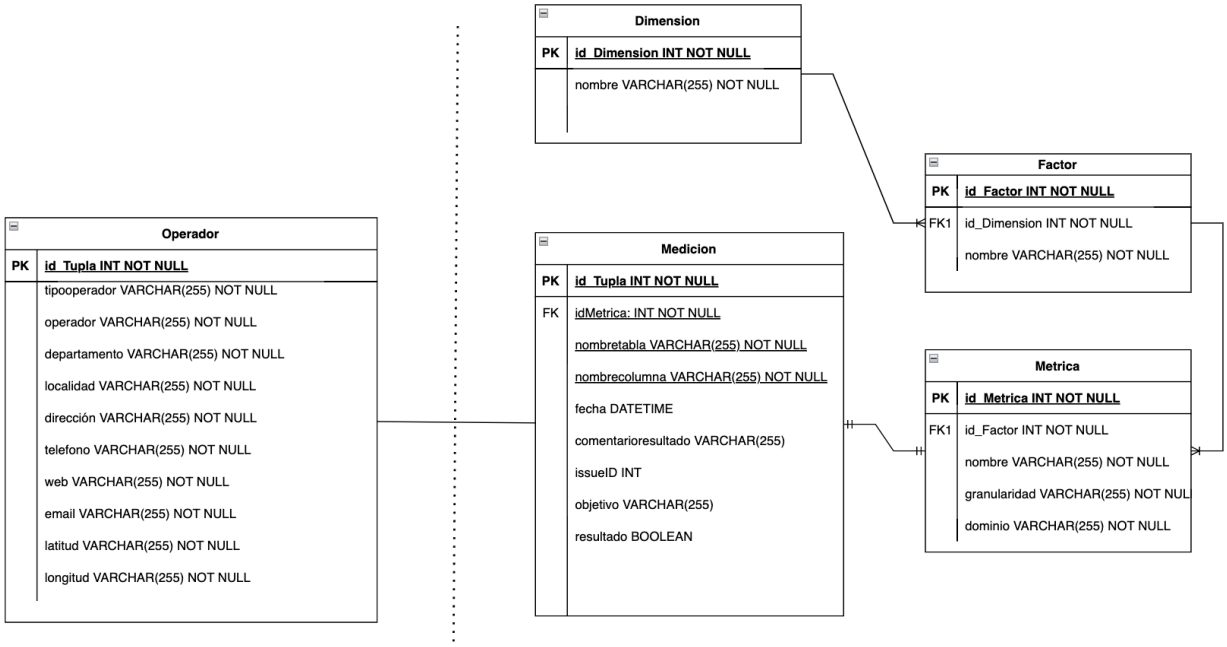

Figura 2. Especificación de diseño de BD de almacenamiento de medidas de calidad obtenidas en la medición

Preguntas asociadas a los factores de calidad

1. ¿Las categorías de los operadores se corresponden con los que estan regulados por el Ministerio? (Exactitud semántica).
2. ¿Las direcciones de mail son correctas? (Exactitud sintáctica).
3. ¿Los telefonos de contacto de los operadores son números válidos? (Exactitud sintáctica).
4. ¿La geolocalización es precisa ? (Precisión)
5. ¿Los operadores estan duplicados?

2.3.3 Aspectos a evaluar y métricas instanciadas. En la Tabla 3 se presentan los aspectos a evaluar. En particular se seleccionaron las métricas M1,M2,M3 de exactitud sintáctica, semántica y precisión correspondientemente y M4 respecto de unicidad (no duplicación). Las métricas variaron desde la verificación de formato, diccionario de datos, cantidades decimales hasta el ratio de no duplicados. Los niveles de granularidad fueron a nivel de celda y tabla.

En la Tabla 4 se presentan las métricas instanciadas. Las métricas M1, M2, M3 y M4 fueron instanciadas sobre el dataset de Operadores. La métrica M1, refería a el mail, url y télefono de los cuales se verifico la correctitud del formato de mail estandar. Las otras métricas instanciadas referían a tipo de operador, departamento, longigud y latitud.

2.4 Especificación de BD de almacenamiento de medidas de Calidad

En la Figura 1 se presenta el diseño de especificación de la base de datos donde se almacenarían las medidas de calidad obtenidas en la medición estableciendo la medición, factor y métrica asociada a la estructura de Operador.

Las medidas de calidad se almacenarán en archivos planos Google Drive y las distintas versiones de cambios se reportarán en el repositorio de GIT [4] y URL de Google Drive [5]

Los archivos fuentes de datos se trabajarán con los CSV que fueron provistos en formato original. Los archivos de trabajos de ejecución de la ejecución se realizarán en DataCleaner versión 5.8.1 (si bien no se realizará la etapa de limpieza, transformación y análisis posterior). La ejecución de los mismos será exploratoria.

Evaluamos los datos de Operadores turísticos [6] por el hecho de su relevancia en el vínculo entre el turismo receptivo y emisivo . La actividad de operadores turísticos esta a su vez regulada por el registro de operadores en https://www.gub.uy/tramites/inscripcion-operador-turistico.

[4] https://github.com/aadorian/cibse_taller
[5] https://docs.google.com/spreadsheets/d/1lcC7mO1O9nn1oqlAxfEHz2H4fmPsvC66tqk4KZLnMg8/edit?usp=sharing
[6] https://www.gub.uy/tramites/inscripcion-operador-turistico

Tabla 3. Modelo de Metricas

MetricaID	Dimensión	Factor	Métrica General	Métrica Definición
M1	Exactitud	Exactitud sintáctica	Verificación de Formato	granularidad: celda, dominio:0,1
M2	Exactitud	Exactitud semántica	Diccionario de Datos	granularidad: celda, dominio:0,1
M3	Exactitud	Precisión	Cantidad Decimales	granularidad: columna, dominio:0,1
M4	Unicidad	No-duplicación	Ratio no-duplicados	granularidad: tabla, dominio%

Tabla 4. Metricas Instanciadas

MetricaID	Metrica Instanciada	Objetivo
M1	Operadores.mail	Verificar que es el formato correcto de email
	Operadores.web	Verificar que es el formato correcto de url
	Operadores.telefono	Verificar si la caracteristica corresponde con el Departamento
M2	Operadores.tipooperador	Verificar si el operador esta en la lista de operadores clasificadas por el Ministerio de Turismo
	Operadores.departamento	Verificar si el departamento esta en la lista de departamentos de Uruguay
M3	Operadores.longitud	Verificar 4 dígitos decimales y > 0
	Operadores.latitud	Verificar 4 dígitos decimales y > 0
M4	Operadores	Verificar no duplicados.

Tabla 5. Resumen del data profiling de emisivos

Estadística del dataset	Frecuencia
Número de variables	43
Número de observaciones	20602
Celdas faltantes	7241
Celdas faltantes (%)	0.8%
Variables numéricas	29
Variables categóricas	14
Alertas	21

Tabla 6. Resumen del data profiling de receptivos

Estadística del dataset	Frecuencia
Número de variables	48
Número de observaciones	48388
Celdas faltantes	25596
Celdas faltantes (%)	% 1.1%
Variables numéricas	29
Variables categóricas	19
Alertas	27

3 Tablas resultados de emisivos, receptivos y operadores

En la Tabla 5 se presenta el resumen correspondiente al data profiling de "`emisivos`"

En la Tabla 6 se presenta el resumen correspondiente al data profiling de "`receptivos`".

Nota: Las alertas de la Tabla 6 corresponden a alta cardinalidad, desvalance, valores faltantes, ceros y alta varianza.

En la Tabla 7 se presenta el resumen correspondiente al data profiling de "`operadopres`" Turísticos.

Nota: Las alertas de la Tabla 5 corresponden a alta cardinalidad, valores faltantes, ceros y alta varianza.

Tabla 7. Resumen del data profiling de Operadores

Estadística del dataset	Frecuencia
Número de variables	10
Número de observaciones	3288
Celdas faltantes	1
Celdas faltantes (%)	< 0.1%
Datos duplicados	98
Datos duplicados(%)	3.0%
Variables numéricas	0
Variables categóricas	10
Alertas	11

Tabla 8. Tipos

Estadística del dataset	Frecuencia
Número de variables	10
Número de observaciones	3288
Celdas faltantes	1
Celdas faltantes (%)	< 0.1%
Datos duplicados	98
Datos duplicados(%)	3.0%
Variables numéricas	0
Variables categóricas	10
Alertas	11

Nota: Las alertas de la Tabla 7 corresponden a alta cardinalidad, duplicados y desvalance.

La creación de un diccionario de datos de categorías correspondientes a las publicadas en el ministerio permitió cotejar con los registros ingresados.

Tabla 9. Nivel de riesgo calidad, umbral establecido en categorías: alto, medio y bajo

Nivel	Rango
alto	>90% de los datos contienen errores
medio	[60,90]% de los datos contienen errores
bajo	< 60% de los datos contienen errores

En las Figuras 2 y 3 se presentan los snapshots de ejemplos de ejecución de ydataprofiling y DataCleaner respectivamente. En las Figuras 4, 5, 6 y 7 presentadas en el anexo se ilustra los templates utilizados en el diseño de medición de calidad.

4 Resultados

4.1 Ejecución de la medición

La ejecución esta disponible en [7]

En la Tabla 10 se presenta el resumen de los resultados de medición.

Nota: las medidas 4.1 y 4.2 correspondientes a latitud, longitud no se realizaron, considerando que el alcance que problema excede las herramientas de data profiling que tenemos disponibles o que pudimos evaluar, sería interesante utilizar el mapeo de una herramienta de geolocalización que permitiera evaluar de forma automática la misma.

Tabla 10. Resumen de resultados de medición.

Medida ID	Total	Nivel Riesgo	Sin Datos	Porcentaje
1.1	1	bajo	827	0.0003%
1.2	3285	alto	1450	99%
2.1	6	bajo	0.001%	
3.1	3269	alto	0	99%
4.1	-	-	-	-
4.2	-	-	-	-
5.1	1283	bajo	0	39%
5.2	2082	medio	0	63%
5.1	2277	medio	0	69%

4.2 Evidencias de ejecución y reporte.

La ejecución de scripts "job" de DataCleaner estan disponibles en https://github.com/aadorian/cibse_taller/tree/main/exec_dataprofiling/jobs

La visualización de la ejecución en formato HTML esta disponible en https://github.com/aadorian/cibse_taller/tree/main/exec_dataprofiling/results

La Figura 6 presenta un snapshot de la ejecución de la medición.

5 Discusión

5.1 Evaluación final de la calidad

La evaluación final de la calidad se presenta en detalle la Tabla 11 y 12 respectivamente. En dichas tablas se presenta un análisis detallado del resultado de cada una de las medidas realizadas y comentarios, así como una posible sugerencia de acciones correctivas.

5.2 Análisis de resultados de medición

Los resultados de medición presentados muestran un grado alto en relación a problemas de calidad. Si bien se estableció un umbral arbitrario, si se considerara que los errores de calidad en un rango de aceptación del nivel del 10% podríamos

[7] https://docs.google.com/spreadsheets/d/1lcC7mO1O9nn1oqlAxfEHz2H4fmPsvC66tqk4KZLnMg8/edit?usp=sharing

decir que todas las métricas estarían en frente a la presencia de un alto grado de problemas de calidad de datos.

El análisis inicial exploratorio de este trabajo dio evidencia del descubrimiento de distintos tipos de errores. Por un lado los correspondiente a datos incompletos, errores de formato y de duplicación entre otros.

El nivel de confianza en la calidad de datos es de bajo nivel en general. Un análisis mas detallado y ejecución de un diseño mas elaborado permitiría identificar en mejor detalle la calidad de los mismos.

5.3 Niveles de Calidad esperados

Los niveles obtenidos de calidad son bajos a nivel global (se establecieron en la Tabla 9). El análisis preliminar de este trabajo permite evidenciar que si los datos básicos de operadores turísticos contienen problemas de calidad, las decisiones basadas en datos que podrían surgir del análisis de los mismos podrían ser incorrectas y eventualmente inexactas.

6 Conclusiones

En este trabajo se realizó un ejemplo de aplicación de "data profiling". Se aplicaron técnicas y métodos de análisis de datos para cumplir con los objetivos del trabajo. Se especificación de un modelo de calidad de datos para luego ser ejecutado. La utilización de dos herramientas como ydata-profiling y Data Cleaner 5.8.1 permitieron realizar la tarea de ejecución del diseño [8]. Se encontraron varios problemas de calidad de datos los cuales no necesariamente son sencillos de identificar a priori. La sistematización de diseño previo y ejecución posterior permitió seguir un proceso interesante. Un análisis mas riguroso y en detalle debería ser necesario para identificar los problemas de calidad de datos existentes.

References

[1] Ziawasch Abedjan. 2018. An introduction to data profiling. In *Business Intelligence and Big Data: 7th European Summer School, eBISS 2017, Bruxelles, Belgium, July 2–7, 2017, Tutorial Lectures 7*. Springer, 1–20.

[2] Ziawasch Abedjan, Lukasz Golab, and Felix Naumann. 2017. Data profiling: A tutorial. In *Proceedings of the 2017 ACM International Conference on Management of Data*. 1747–1751.

[3] Victor R Basili. 1992. *Software modeling and measurement: the Goal/Question/Metric paradigm*. Technical Report.

[4] Bahaa Eddine Elbaghazaoui, Mohamed Amnai, and Abdellatif Semmouri. 2021. Data profiling over big data area: a survey of big data profiling: state-of-the-art, use cases and challenges. In *Intelligent Systems in Big Data, Semantic Web and Machine Learning*. Springer, 111–123.

[5] Lorena Etcheverry, Verónika Peralta, and Mokrane Bouzeghoub. 2008. Qbox-foundation: a metadata platform for quality measurement. In *proceeding of the 4th Workshop on Data and Knowledge Quality (QDC'2008)*.

[6] Ministerio de Turismo de Uruguay. 2022. Turismo Emisivo 2022. https://www.gub.uy/ministerio-turismo/datos-y-estadisticas/estadisticas/turismo-emisivo-2022. Accessed on May 13, 2023.

[7] Ministerio de Turismo de Uruguay. 2023. Turismo Receptivo. https://www.gub.uy/ministerio-turismo/turismoreceptivo. Accessed on May 13, 2023.

[8] https://datacleaner.github.io/downloads

A Snapshot de Herramientas

B Templates de Diseño

C Ejecución de la medición

C.1 Operadores turísticos categorías

1. AGENCIA DE VIAJES
2. TRANSPORTE TURISTICO
3. ALOJAMIENTO TURÍSTICO
4. PRESTADORES DE SERVICIOS TURISTICOS INMOBILIARIOS
5. TURISMO AVENTURA
6. ESTABLECIMIENTO ENOLÓGICO (Prestan servicios de alojamiento).
7. ESTABLECIMIENTO ENOLÓGICO (NO Prestan servicios de alojamiento). SUCURSAL
8. GUIA DE TURISMO
9. ORGANIZADORES PROFESIONALES DE CONGRESOS
10. OBSERVACIÓN DE CETÁCEOS
11. ARRENDADORA DE VEHÍCULOS SIN CONDUCTOR
12. PRESTADORES DE SERVICIOS TURISTICOS RURALES
13. SALAS DE CONVENCIONES instaladas en Establecimientos Rurales.
14. SALAS DE CONVENCIONES instaladas en Alojamientos Turísticos.
15. SALAS DE CONVENCIONES instaladas en establecimientos que no requieren inscripción en el
16. Registro de Prestadores de Servicios Turísticos

Fuente: https://www.gub.uy/tramites/inscripcion-operador-turistico [9]

[9] El registro de ejecución y comentarios de evaluación final esta disponible en la siguiente planilla: https://docs.google.com/spreadsheets/d/1lcC7mO1O9nn1oqlAxfEHz2H4fmPsvC66tqk4KZLnMg8/edit?usp=sharing

Tabla 11. Evaluación Final de Calidad

Medida ID	Nota	Sug. Acción Correctiva	Comentarios
1.1	Si bien cumplen en su mayoría con el formato, varios mails están registrados como listas con delimitadores distintos (ejemplo ; o /). Otros directamente no tienen mail asignado. A su vez, unos con mayúscula.	Unificar a un único formato de minúsculas con una validación de la expresión regular correspondiente al email. En caso de identificar un listado de emails, asignar una estructura adecuada para identificarlos.	Del total de registros 3288, 827 no tienen asignado un valor
1.2	Varios registros independientemente del formato http:// o https:// no cumplen con el formato estándar de acceso (por ejemplo, hay registros con espacios, caracteres especiales).	Unificar el formato y validar el acceso a la URL mediante un mecanismo semiautomático que permita establecer si la URL está disponible en tiempo real y se corresponde con el sitio web del operador.	Del total de registros 1450 no tienen URL
2.1	En general los registros están dentro del dominio especificado, si bien no se realizó un análisis exhaustivo del mismo.	Realizar un análisis de que la característica se corresponda con la localidad en la cual se encuentra el operador.	
3.1	3269 del total de 3288 de los valores NO coincide con la definición formal del tipo de operador definido por el Ministerio de Turismo. Si bien uno podría mapear, por ejemplo, Inmobiliaria se podría eventualmente ser una Agencia de Viajes, pero podría ser catalogado inicialmente como Prestador servicio Turístico inmobiliario. A su vez, un mismo operador podría prestar varios servicios y esto no está contemplado.	Establecer un identificador que permita sin ambigüedad identificar el tipo de operador y que este corresponda con el asignado como categoría del Ministerio en https://www.gub.uy/tramites/inscripcion-operador-turistico. Rediseñar el diccionario de categorías.	El total del registro no se corresponden con la descripción exacta del tipo indicado como categoría en el ministerio. A su vez, por ejemplo, la categorización en sí debería ser no ambigua (SALAS DE CONVENCIONES instaladas en Establecimientos Rurales. SALAS DE CONVENCIONES instaladas en Alojamientos Turísticos. SALAS DE CONVENCIONES instaladas en establecimientos que no requieren inscripción en el Registro de Prestadores de Servicios Turísticos). En el registro se reporta SALA DE CONVENCIÓN y no necesariamente está clara a cuál de las tres posibles se refiere.
4.1	No se realizó una evaluación de latitud / longitud dado el alcance que problema excede las herramientas de data profing que tenemos disponibles		
4.2	No se realizó una evaluación de latitud / longitud dado el alcance que problema excede las herramientas de data profing que tenemos disponibles		

Tabla 12. Continuación evaluación final de calidad.

Medida ID	Nota	Sug. Acción Correctiva	Comentarios
5.1	2005 registros son únicos de operadores, y 1283 no son únicos completando asi los 3288 registros existentes	Establecer un criterio de unicidad en el listado.	Varios registros estan duplicados por tipo de operador teniendo registros como por ejemplo ABITAB que aparece 182 veces.
5.2	Varios mails estan registrados de forma incorrecta	Establecer un criterio de contacto de mail de los operadores	
5.3	1011 de los 3288 son registros únicos. Sin embargo hay 2277 no únicos de los cuales 1450 no tienen descripción	Establecer un criterio de unicidad de URL del operador	

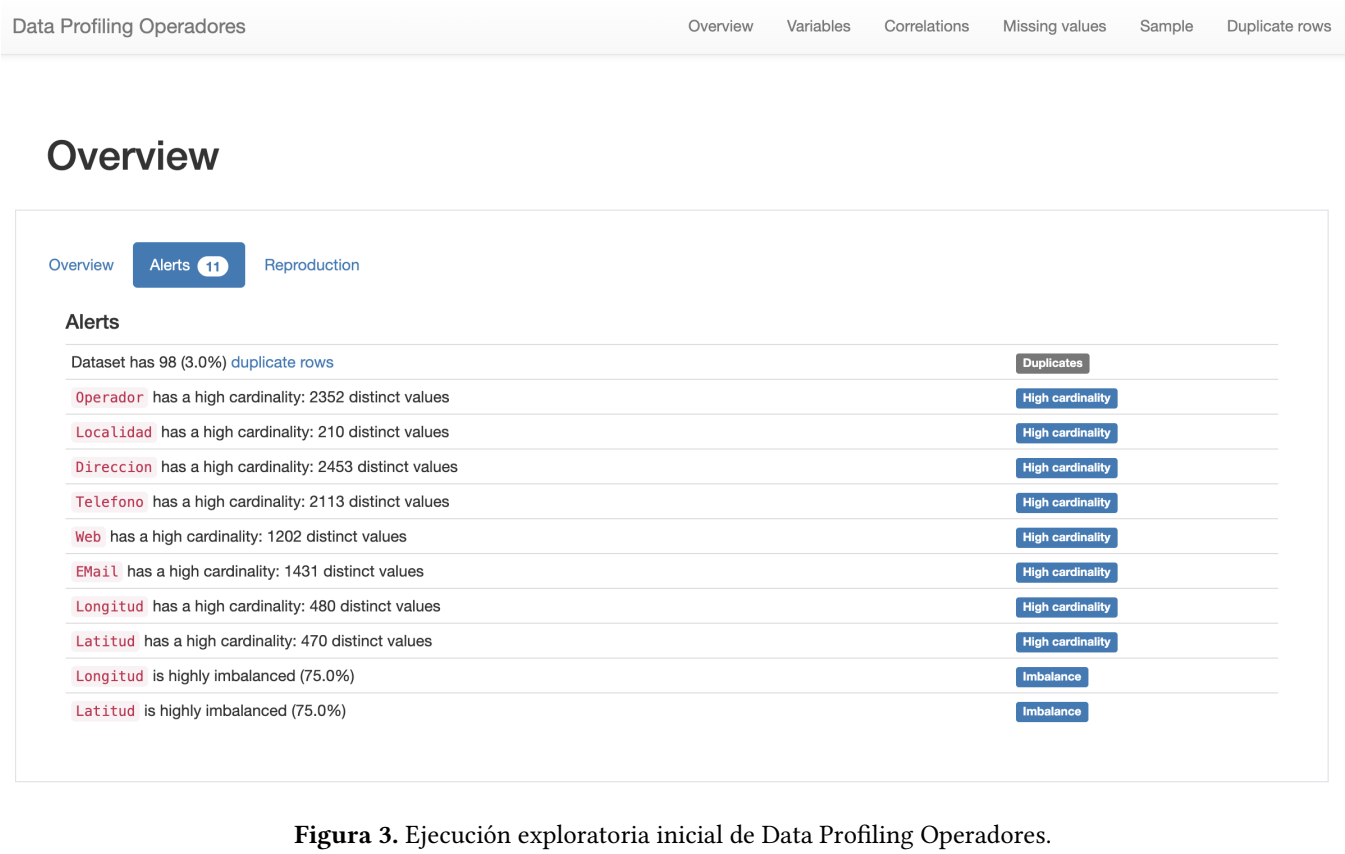

Figura 3. Ejecución exploratoria inicial de Data Profiling Operadores.

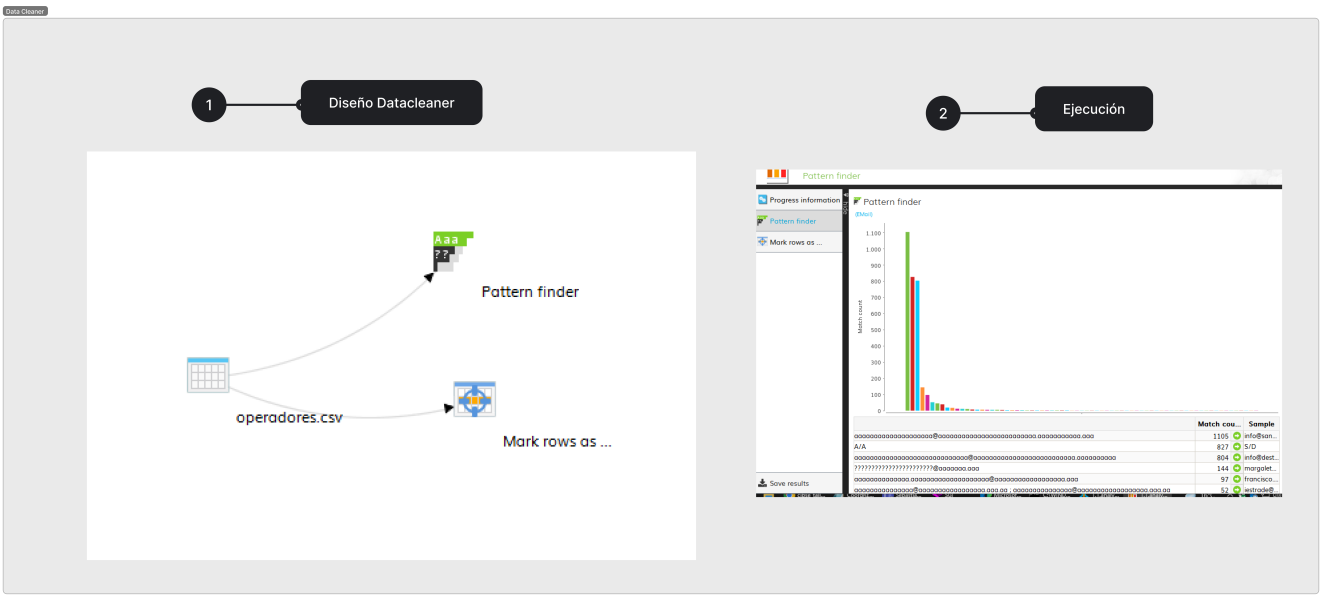

Figura 4. Ejemplo de ejecución de job en DataCleaner

Dimensión	Factor	Metrica ID	Métrica Descripción	Semantica	Unidad	Definición	Granularidad	Nivel de Riesgo
Exactitud	Exactitud Sintáctica	M1	Valor que no cumple formato	Cantidad de valores que no cumplen formato	Booleano o grado	Para booleano: 1esta fuera del rango 0 en caso contrario Para grado: valor entre {0,1}	Celda	Alto
		M2	Valor situado fuera del rango	Cantidad de valores que estan por fuera de un rango establecido	Booleano o grado	Para booleano: 1esta fuera del rango 0 en caso contrario Para grado: valor entre {0,1}	Celda	Alto
	Exactitud Semántica	M3	Valor fuera de un conjunto referencial definido como válido	Cantidad de valores que no son parte del referencial valido	Booleano	1 si pertenece al referencial 0 si no.	Celda	Alto
	Precición	M4	Valor fuera de presición definida	Cantidad de valores que no cumplen la presición definida	Booleano	1 si tiene la presición adecuada 0 en otro caso	Celda	Medio
Unicidad	Duplicación	M5	Dos o mas registros repetidos de manera exacta	Cantidad de registros ingresados por duplicado	Booleano	1 si hay duplicación de registros 0 en otro caso	Tupla	Alto

Figura 5. Diseño de dimensión, factor y métrica.

Operadores.CSV

	A	B	C	D	E	F	G	H	I	J	
	Medida ID	**Objeto**	**Semantica**	**%SinDatos / Total**	**Valores**	**Evaluación Q**	**Estado**	**Sin Datos**	**Porcentaje (Valores obtenidos / Total Registro)**	**Total Registros**	
1	1.1	Operadores.mail	Hallar los valores que no cumplen en formato de mail@server.com	0,2515206813	1	Baja	Realizado	827	0,000304136253	3288	Varios mail listas con d (ejemplo ; o no tienen m unos con m
2	1.2	Operadores.web	Hallar los valores que no cumplen en formato de http:// o https://	0,4409975669	3285	Alta	Realizado	1450	0,9990875912	3288	Varios regis del formato cumplen co acceso (por espcación ,
3	2.1	Operadores.telefono	Hallar los valores que no cumplen con la caracteristica de inicio de un telefono fijo	0	6	Baja	Realizado		0,001824817518	3288	En general l del dominio se realizó ur mismo
4	3.1	Operadores.tipooper	Hallar los valores que no estan dentro del listado aceptado como	0	3269	Alta	Realizado	0	0,9942214112	3288	3269 del tot NO coincid del tipo de c Ministerio d podria mapc Inmobiliaria ser una Age

Ejecución de la medición

Figura 6. Ejecución de la medición.

Metrica ID	Métrica Descripción	Medida ID	Objeto	Semantica	Problema Calidad
M1	Valor que no cumple formato	1.1	Operadores.mail	Hallar los valores que no cumplen en formato de mail@server.com	
		1.2	Operadores.web	Hallar los valores que no cumplen en formato de http:// o https://	
M2	Valor situado fuera del rango	2.1	Operadores.telefono	Hallar los valores que no cumplen con la caracteristica de inicio de un telefono fijo	
M3	Valor fuera de un conjunto referencial definido como válido	3.1	Operadores.tipooperador	Hallar los valores que no estan dentro del listado aceptado como tipos de operador por el ministerio	
M4	Valor fuera de presición definida	4.1	Operadores.latitud	Hallar los valores que no cumplan con una valor >0 y presición de 4 decimales	
		4.2	Operadores.latitud	Hallar los valores que no cumplan con una valor >0 y presición de 4	
M5	Dos o mas registros repetidos de manera exacta	5.1	Operadores	No pueden haber registros duplicados que correspondan con un operador	
		5.2	Operadores.mail	No pueden haber registros de mail duplicados	
		5.3	Operadores.url	No pueden haber registros de url distintas para un mismo operador	

Figura 7. Instanciación de Métricas sobre Operadores.

Medida ID	Objeto	Semantica	% Valores	Resultado	Estado	Issue Reporte
1.1	Operadores.mail	Hallar los valores que no cumplen en formato de mail@server.com				
1.2	Operadores.web	Hallar los valores que no cumplen en formato de http:// o https://				
2.1	Operadores.telefono	Hallar los valores que no cumplen con la caracteristica de inicio de un telefono fijo				
3.1	Operadores.tipooperador	Hallar los valores que no estan dentro del listado aceptado como tipos de operador por el ministerio				
4.1	Operadores.latitud	Hallar los valores que no cumplan con una valor >0 y presición de 4 decimales				
4.2	Operadores.latitud	Hallar los valores que no cumplan con una valor >0 y presición de 4 decimales				
5.1	Operadores	No pueden haber registros duplicados que correspondan con un operador				
5.2	Operadores.mail	No pueden haber registros de mail duplicados				
5.3	Operadores.url	No pueden haber registros de url distintas para un mismo operador				

Figura 8. Diseño de Template de registro de ejecución

Dimensión	Factor	Metrica ID	Métrica Descripción	Semantica	Unidad	Definición	Granularidad	Nivel de Riesgo
Exactitud	Exactitud Sintáctica	M1	Valor que no cumple formato	Cantidad de valores que no cumplen formato	Booleano o grado	Para booleano: 1esta fuera del rango 0 en caso contrario Para grado: valor entre {0,1}	Celda	Alto
		M2	Valor situado fuera del rango	Cantidad de valores que estan por fuera de un rango establecido	Booleano o grado	Para booleano: 1esta fuera del rango 0 en caso contrario Para grado: valor entre {0,1}	Celda	Alto
	Exactitud Semántica	M3	Valor fuera de un conjunto referencial definido como válido	Cantidad de valores que no son parte del referencial valido	Booleano	1 si pertenece al referencial 0 si no.	Celda	Alto
	Precición	M4	Valor fuera de presición definida	Cantidad de valores que no cumplen la presición definida	Booleano	1 si tiene la presición adecuada 0 en otro caso	Celda	Medio
Unicidad	Duplicación	M5	Dos o mas registros repetidos de manera exacta	Cantidad de registros ingresados por duplicado	Booleano	1 si hay duplicación de registros 0 en otro caso	Tupla	Alto

Figura 9. Especificación de Registro