Mental Health and Suicide Rates ETL

Link al repositorio de Github: https://github.com/ramonmorillx/ETL-s

Autores: Ramón Morillo Barrera
José Carlos Monescillo Calzado

1. Objetivos del Proyecto:

Este proyecto tiene como propósito principal realizar un proceso de ETL (Extracción, Transformación y Carga) aplicado a un conjunto de datos titulado 'Mental Health and Suicide Rates'. El conjunto de datos a tratar incluye información acerca de factores relacionados con la salud mental, tasas de suicidios, facilidades para los pacientes, etc.

El objetivo real de este proyecto es convertir estos datos en un valioso recurso analítico que facilite el estudio de tendencias y patrones subyacentes, con la finalidad de apoyar investigaciones en la salud y la formulación de políticas preventivas a esta problemática social.

Se garantizará que los datos mantengan su consistencia, estén limpios y preparados para poder realizar un análisis posterior en herramientas de business inteligence (inteligencia de negocio) y creación de modelos predictivos.

Objetivos específicos:

1. **Extracción de datos:** Obtención de información de fuentes relacionadas con el tema a tratar, en nuestro caso, el dataset se encontraba en Kaggle por lo que nos facilita el proceso de extracción de los datos.
   Tendremos que traspasar los datos de nuestros archivos .csv a tablas ya creadas en SQL y verificar si los valores de cada fila y columna de nuestras tablas de SQL se corresponden con el número de valores en cada fila y columna de cada archivo .csv

2. **Transformación de datos:** Este paso en nuestro caso, es el más importante ya que incluye la estandarización de formato, manejo de datos faltantes y reducción de inconsistencias, además del cálculo de las métricas establecidas. En este proceso transformaremos los datos con SQL con el objetivo de dejarlos preparados para posteriores análisis.

3. **Carga de datos:** Depositar los datos transformados y procesados en un sistema de datos estructurado, como SQL. En este caso lo cargaremos como

una base de datos relacional que estará optimizada para recibir consultas acerca de los datos.

Justificación de la importancia de realizar un proceso ETL
Un proyecto de ETL es primordial para asegurar tanto la calidad como la accesibilidad y la utilidad de un conjunto de datos. Si los datos no están correctamente transformados y limpios, tras su análisis en un proyecto real, las conclusiones podrían estar sesgadas o ser completamente erróneas.

Es por ello que el enfoque de la ETL asegura que nuestro dataset sea transformado y suponga un valioso recurso para investigadores interesados en abordar la problemática de la salud mental.

2. Dataset:

Enlace a los datasets utilizados:

https://www.kaggle.com/datasets/twinkle0705/mental-health-and-suicide-rates?select=Crude+suicide+rates.csv

Explicación de las fuentes de datos (Diccionario de datos)
Nuestro dataset se compone de 4 ficheros .csv (comma separated values), por tanto, el diccionario de datos se compone de 4 tablas, cada una asociada a su correspondiente fichero.

- Age-standardized suicide rates

Country	VARCHAR	Nombre del país, 183 valores únicos
Sex	VARCHAR	Sexo, 3 valores únicos (Male, Female, Both sexes)
2016	DECIMAL	Tasa de suicidio estandarizada para el año 2016
2015	DECIMAL	Tasa de suicidio estandarizada para el año 2015
2010	DECIMAL	Tasa de suicidio estandarizada para el año 2010
2000	DECIMAL	Tasa de suicidio estandarizada para el año 2000

- Crude suicide rates

Country	VARCHAR	Nombre del país, 183 valores únicos
Sex	VARCHAR	Sexo, 3 valores únicos (Male, Female, Both sexes)

80_above	DECIMAL	Tasas de suicidio (año 2016) por cada 100 000 habitantes en rango de edad 80+
70to79	DECIMAL	Tasas de suicidio (año 2016) por cada 100 000 habitantes en rango de edad 70 a 79
60to69	DECIMAL	Tasas de suicidio (año 2016) por cada 100 000 habitantes en rango de edad 60 a 69
50to59	DECIMAL	Tasas de suicidio (año 2016) por cada 100 000 habitantes en rango de edad 50 a 59
40to49	DECIMAL	Tasas de suicidio (año 2016) por cada 100 000 habitantes en rango de edad 40 a 49
30to39	DECIMAL	Tasas de suicidio (año 2016) por cada 100 000 habitantes en rango de edad 30 a 39
20to29	DECIMAL	Tasas de suicidio (año 2016) por cada 100 000 habitantes en rango de edad 20 a 29
10to19	DECIMAL	Tasas de suicidio (año 2016) por cada 100 000 habitantes en rango de edad 10 a 19

- Facilities

Country	VARCHAR	Nombre del país, 112 valores únicos
Year	VARCHAR	año 2016
Mental_hospitals	DECIMAL	Hospitales mentales (por cada 100 000 habitantes)
Mental_units	DECIMAL	Unidades de salud mental en hospitales generales (por cada 100 000 habitantes)
outpatient_facilities	DECIMAL	Centros ambulatorios de salud mental (por cada 100 000 habitantes)
day_treatment	DECIMAL	Centros de tratamiento diurno de salud mental (por cada 100 000 habitantes)
residential_facilities	DECIMAL	Instalaciones residenciales comunitarias (por cada 100 000 habitantes)

- Human Resources

Country	VARCHAR	Nombre del país, 107 valores únicos
Year	VARCHAR	año 2016

Psychiatrists	DECIMAL	Psiquiatras que trabajan en el sector de la salud mental (por cada 100.000 habitantes)
Nurses	DECIMAL	Enfermeros que trabajan en el sector de la salud mental (por cada 100.000 habitantes)
Social_workers	DECIMAL	Trabajadores sociales que trabajan en el sector de la salud mental (por cada 100.000 habitantes)
Psychologists	DECIMAL	Psicólogos que trabajan en el sector de la salud mental (por cada 100.000 habitantes)

Frecuencia de actualización de los datos.

Según la fuente de nuestros datos (Kaggle), la frecuencia de actualización de los mismos es 'nunca', es decir, se publicaron por primera vez en el año 2016 y no se han vuelto a actualizar. Observamos que no hay frecuencia de actualización para ninguno de nuestros datos, por tanto, no lo tendremos en cuenta en las métricas de data quality ya que no lo podemos medir y además evitaremos distorsionar los resultados de calidad del dato.

3. Características de los Datos:

Descripción de los tipos de datos manejados.

Nuestros datos son datos estructurados, pues son datos organizados en un formato rígido, en filas y columnas, distribuidos en varias tablas que se ajustan bien a una base de datos relacional (SQL).

Los datos están organizados en 4 ficheros .csv, es un formato simple de archivo de texto donde los valores están separados por comas. Las ventajas que esto supone son que este tipo de formato es ampliamente utilizado, sencillo de manipular y compatible con casi todas las herramientas.

Nos encontramos con 1 tipo de dato que se repite en las 4 tablas, de los cuáles podremos beneficiarnos para realizar nuestro modelo de datos e interconectar las tablas entre sí.

- Country

Es de tipo VARCHAR, por lo que contiene texto. En nuestro caso, el país (Country), será nuestra columna óptima para relacionar las tablas.

Además, cada tabla contiene datos de carácter DECIMAL que nos proporcionan bastante información. Su formato en las diferentes tablas son:

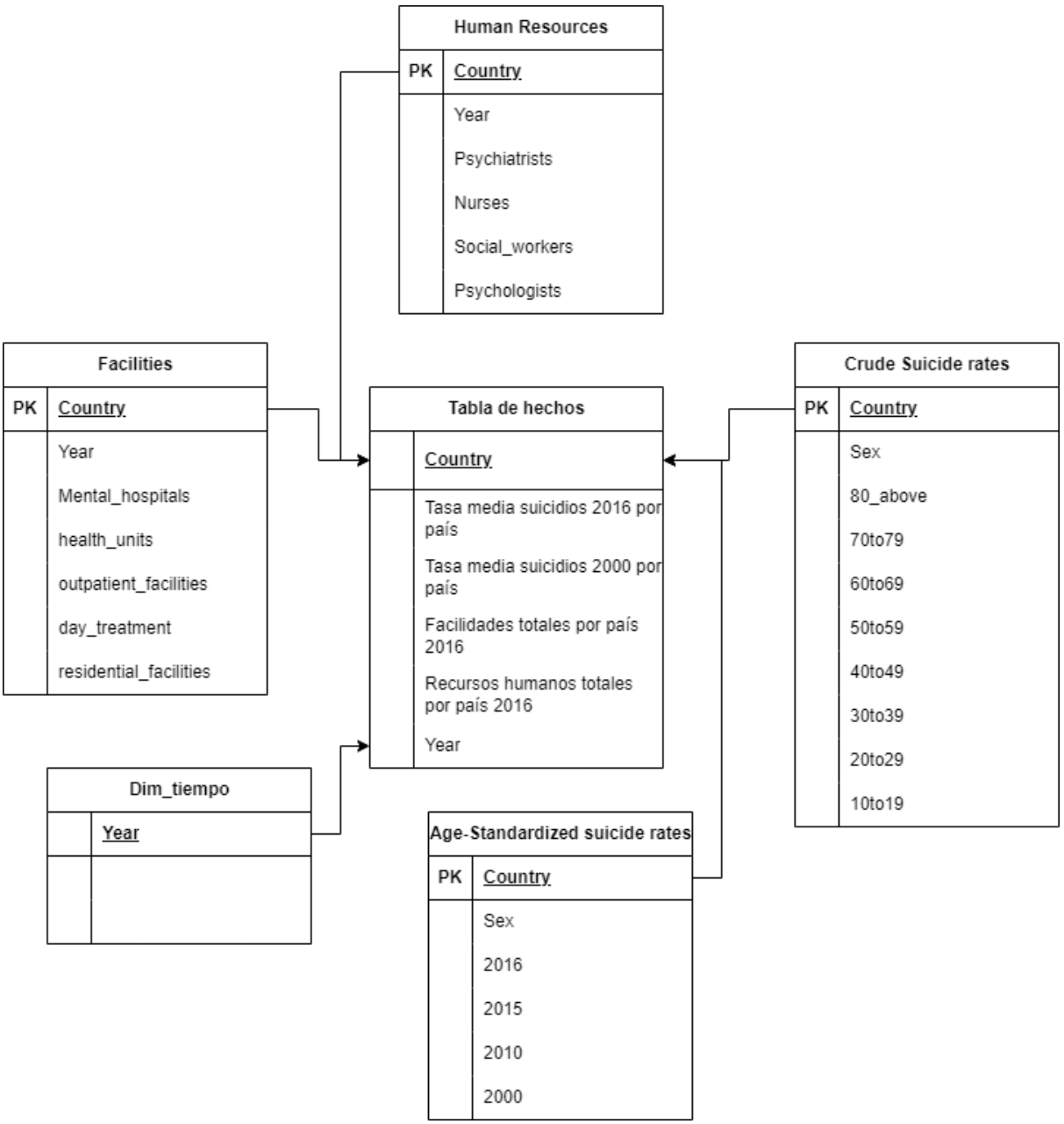
- ‘Age-standardized suicide rates’ contiene datos estandarizados, es decir, toman valores de 0 a 100
- ‘Crude suicide rates’ , ‘Facilities’ y ‘Human Resources’ presentan datos de tasas por cada 100 000 habitantes.

Definición del Modelo de Datos.

Hemos optado por definir un modelo relacional de estrella en el que las diferentes tablas se unen mediante el país, ya que el país es un registro común en todas las tablas de los archivos disponibles, siendo un registro único e identificable. El país será la clave primaria en todas las tablas que se unen a nuestra tabla de hechos, pues no puede haber dos registros con el mismo país ni registros nulos.

Añadimos la tabla de dimensión de tiempo ya que en nuestros datos tenemos columnas con los años 2016, 2015, 2010 y 2000.

Las métricas que calcularemos son las siguientes:
- Tasa media de suicidios en 2016 por país
- Tasa media de suicidios en 2000 por país
- Facilidades totales en 2016 por país (por cada 100,000 habitantes)
- Recursos humanos totales en 2016 por país (por cada 100,000 habitantes)



4. Calidad de los Datos:

En cuanto a la calidad de los datos, solo se medirán aquellas métricas de las que tengamos suficiente información para calcular, debido a que, si damos por hecho valores para métricas que no podemos calcular, distorsionará nuestro valor real de calidad del dato.

Vamos a realizar un estudio de calidad de los datos, teniendo en cuenta las siguientes métricas de calidad del dato:

- **Precisión**: Evalúa la exactitud en la representación de los valores registrados, especialmente en términos numéricos y de decimales. En nuestro caso, que tengan el mismo número de decimales en cada tabla.
- **Estructura:** Mide si los datos siguen un formato o esquema estructural definido, como el tipo de dato y la longitud esperada.
- **Completitud**: Evalúa el porcentaje de datos presentes frente al total de datos esperados. En nuestro caso que no hayan datos faltantes.
- **Razonabilidad**: Evalúa si los datos tienen sentido y cumplen con ciertas reglas lógicas o umbrales, basándose en el contexto. En nuestro caso que se ajusten las tasas porcentuales entre [0, 100].
- **Semántica**: Evalúa el significado de los datos en el contexto en el que se encuentran. En nuestro caso que los datos tipo Varchar estén bien escritos y que los datos numéricos se ajusten a rangos de valores lógicos.
- **Identificabilidad**: Evalúa si cada dato puede ser identificado de manera única. En nuestro caso que los países no se repitan en una misma tabla, ya que debería ser clave primaria.
- **Consistencia:** Mide la coherencia entre datos que deberían ser iguales o relacionados entre diferentes tablas.

En este caso no mediremos:

- **Linaje:** no sabemos qué transformaciones previas han recibido nuestros datos. Necesitaríamos la información sobre el origen de los datos, qué transformaciones se realizaron antes de que el dataset fuese publicado y un historial de flujo de los datos desde su recolección hasta su publicación.
- **Puntualidad:** no tenemos una frecuencia de actualización disponible. Únicamente sabemos que se publicaron en 2016.
- **Integridad:** no tenemos datos que referencian directamente a otros datos.

Evaluación de calidad de datos

En nuestro caso, el valor de las métricas de calidad calculadas son:

- Precisión: 87,12%
- Estructura: 100%
- Completitud: 90%
- Razonabilidad: 100%

- Semántica: 90,30%
- Identificabilidad: 100%
- Consistencia: 71,57%

Por lo que obtenemos una **calidad total** de los datos de un **91,28%.**

Con nuestro proceso de ETL tendremos que acercarnos lo máximo posible al 100%, con el objetivo de que estos datos puedan ser completamente utilizables para futuros proyectos.

5. Limpieza de los Datos:

En la limpieza y preparación de datos realizaremos los siguientes cambios para estandarizar los datos y facilitar la interpretabilidad y el análisis de los mismos.

En primer lugar, como no sabemos el valor de los datos NaN, vamos a imputar dichos valores por 0 en caso de datos numéricos y por el valor 'Desconocido' en el caso de datos textuales. Si contamos con algún país que presente un valor nulo, directamente eliminaremos dicha fila, ya que la información de dicho registro carece de sentido.

En segundo lugar, estandarizamos todas las columnas numéricas, añadiendo 1 o 3 decimales, dependiendo de la mayoría de registros en cada tabla. Así tendremos todas las columnas con la misma longitud.

En tercer lugar, ajustar todas las tasas para que se encuentren entre los valores [0, 100], ya que no pueden ser mayor de 100% si es el total, ni menor que un 0%. En cualquier caso, los valores mayores que 100 serán imputados por 100 y los menores de 0 se imputarán por 0.

6. Problemas y Próximos Pasos:

Una de las principales problemáticas que surgieron fue la de calcular métricas como el linaje, en la que necesitamos una frecuencia de actualización de los datos. Nuestros datos no se actualizan desde 2016 por lo que únicamente disponemos de información hasta ese mismo año. Esto supone que el cálculo de nuestras métricas se ajuste a ese mismo año.

Tampoco pudimos calcular la puntualidad de los datos, pues no tenemos una frecuencia de actualización disponible. Únicamente sabemos que se publicaron en 2016 y que no han sufrido actualizaciones posteriores, por lo que decidimos no tener en cuenta esa métrica de calidad.

Una gran mejora a futuro sería incluir un registro de actualizaciones de los datos, con los que poder calcular métricas y obtener valores de periodos anuales recientes. Estos datos facilitan la obtención de conclusiones que se ajusten al tiempo y a la realidad, promoviendo la investigación y la creación de políticas preventivas para esta problemática social.

Como propuesta para futuros pasos me gustaría recomendar el uso de estos datos para la realización de estudios acerca del impacto del suicidio en los distintos países del mundo, comparándolo con la proporción de instituciones o facilidades que tienen los pacientes en dichos países, pienso que podría dar conclusiones significativas y mostrar aspectos en los que un país debe incidir y mejorar para reducir esta gran problemática mundial.