Documentación Sobre Las Etapas De Procesamiento De Datos De Mortalidad Materna

Versión 1.0

Desarrollado por Dalton Costa

Índice

1. Introducción

1.1 Objetivo

El principal objetivo de este documento es describir detalladamente las etapas de procesamiento de datos utilizadas para alimentar una herramienta de análisis de la mortalidad materna en diferentes países. Sirviendo como una guía esencial, este documento está destinado a profesionales que buscan entender en profundidad cómo se procesaron y prepararon los datos antes de integrar la herramienta.

1.2 Contexto del Proyecto

El propósito de la herramienta de análisis de la mortalidad materna es integrar de manera sistematizada y estandarizada varias bases de datos sobre muertes maternas. Esta herramienta está diseñada para funcionar como un insumo analítico fundamental, proporcionando métricas dinámicas de mortalidad materna para una variedad de usuarios, incluyendo la Organización Panamericana de Salud (OPS), equipos de ministerios de salud nacionales, institutos de estadística y académicos.

En 2020, aproximadamente 287.000 mujeres fallecieron globalmente debido a complicaciones durante o después del embarazo y el parto, de las cuales cerca de 9.200 estaban en la Región de las Américas. Esta región registró un promedio de 25 muertes maternas diarias, manteniendo una tasa de mortalidad materna de 68 por 100.000 habitantes entre los años 2000 y 2020. La pandemia de COVID-19 exacerbó esta situación, aumentando el número de muertes maternas en varios países de la región e impactando la tendencia de reducción de la mortalidad materna - un objetivo clave del Objetivo de Desarrollo Sostenible 3 (ODS 3) para 2030.

Ante este escenario, se vuelve crucial fortalecer la capacidad analítica de equipos nacionales en diferentes áreas e instituciones, como programas de salud materna, estadísticas e información, monitoreo y evaluación. El contexto post-pandémico exige un esfuerzo redoblado en el desarrollo de herramientas analíticas que proporcionen datos para una toma de decisiones eficiente en salud materna.

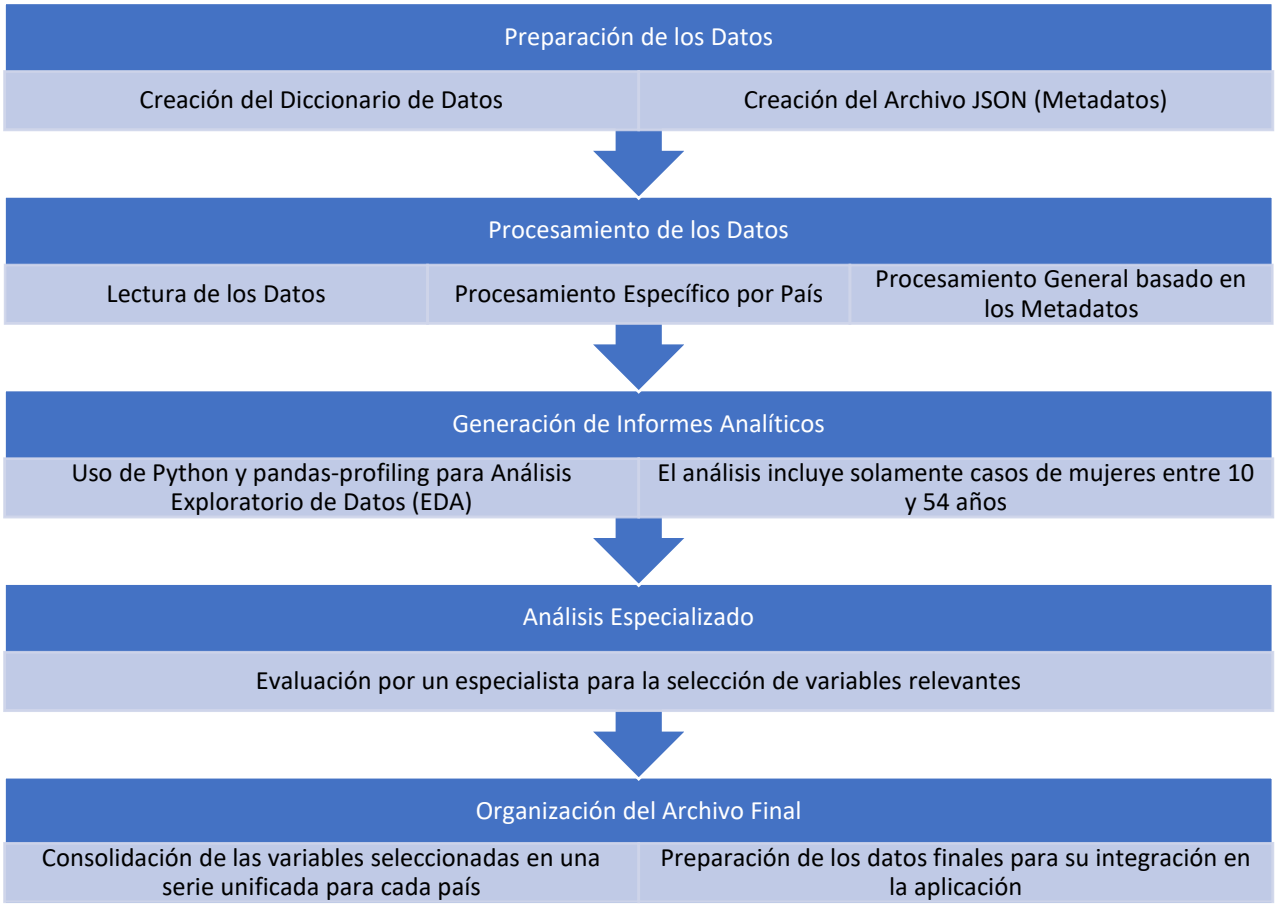
1.3 Países y Período Contemplados

El proyecto abarca un total de 18 países, cada uno aportando datos cruciales para el análisis de la mortalidad materna desde 2015 hasta el último año disponible. Los países participantes, así como el período contemplado son:

1. Argentina: 2015 a 2021
2. Brasil: 2015 a 2020
3. Chile: 2015 a 2020
4. Colombia: 2015 a 2021
5. Costa Rica: 2015 a 2019
6. Cuba: 2015 a 2020
7. Ecuador: 2015 a 2022
8. El Salvador: 2015 a 2018
9. Estados Unidos: 2015 a 2021
10. Guatemala: 2015 a 2022
11. México: 2015 a 2022
12. Nicaragua: 2015 a 2022
13. Panamá: 2015 a 2020
14. Paraguay: 2015 a 2020
15. Perú: 2015 a 2020
16. República Dominicana: 2015 a 2018
17. Uruguay: 2015 a 2021
18. Venezuela: 2015 a 2016

1.4 Resumen de las Etapas de Procesamiento de los Datos

El esquema a continuación ofrece una visión general y estructurada del proceso que se describirá en detalle más adelante. Cada etapa aporta significativamente a la creación de un conjunto de datos robusto e informativo, listo para ser utilizado en la aplicación analítica.



2. Preparación de los Datos

La preparación de los datos es un paso crucial en el proceso de análisis y procesamiento de información relacionada con la mortalidad materna en los 18 países involucrados en el proyecto. Esta sección describe las metodologías y herramientas utilizadas para estructurar y estandarizar los datos, destacando que los procedimientos descritos a continuación se aplicaron en todos los países y en todos los años de los datos recopilados y se basaron en la documentación de interpretación de los datos proporcionada por cada país.

2.1 Creación del Diccionario de Datos

La primera etapa en la preparación de los datos fue la creación manual de un único diccionario de datos para cada país que contempla todos los años disponibles, con la ayuda del lenguaje de programación R. Este proceso implicó el análisis de las características de cada

variable, documentándolas en un archivo Excel con una estructura específica. El diccionario incluyó información como el número de identificación de la variable (N), el nombre de la variable (Variable), una descripción oficial proporcionada en la documentación de los datos por el país (Descripción), el tipo de variable (Tipo), el rango de códigos para variables categóricas (Rango Clave), y los códigos o valores que representan datos ignorados o ausentes (Ignorado/Missing).

Cabe destacar que, durante este proceso, se normalizaron los nombres de las variables de los datos de cada país para garantizar la estandarización, incluyendo la conversión a minúsculas y la eliminación de espacios y caracteres especiales. Esto significa que los nombres en el diccionario pueden diferir ligeramente de los nombres originales en los datos. El diccionario de datos se creó de forma unificada para cada país, abarcando todas las variables de todos los años, lo que puede resultar en variaciones en la disponibilidad de variables a lo largo de los años.

Ejemplo de diccionario de datos:

N	Variable	Descripcion	Tipo	Rango Clave	Ignorado/Missing
1	ano	Año de registro	Numérico	-	-
2	provres	Jurisdicción de residencia del fallecido	Categorías	02 --98	99
3	depres	Departamento de residencia del fallecido	Categorías	02001 -- 98999	99999
4	sexo	Sexo del fallecido	Categorías	1 -- 3	-

2.2 Creación del Archivo JSON (Metadatos)

Tras la creación del diccionario de datos, el siguiente paso fue la creación de un archivo JSON que contiene metadatos para cada país. Este archivo se elaboró manualmente basándose en la información recopilada en Excel y en las documentaciones oficiales proporcionadas por cada país. La creación de este archivo de metadatos tuvo como objetivo facilitar la estandarización de los valores y etiquetas (labels) de los datos a lo largo de todos los años proporcionados.

El archivo JSON se construyó siguiendo una estructura jerárquica que permite asignar diferentes características a cada variable. Dentro de esta estructura, cada variable se especificó con un tipo ("type"), además de otras propiedades relevantes. A continuación, se describen todas las posibilidades de “type”:

- **"numeric"**: se asignó a las variables que son de tipo numérico. Además del atributo "type", cuando la variable era del tipo "numeric", también era posible especificar el

parámetro "missing" donde se pueden proporcionar una lista de valores que deberían considerarse como missing.

- **"categorical_numeric"**: se asignó a las variables que son de tipo categórica y tienen como valores en la base de datos códigos numéricos. Además del atributo "type", cuando la variable era del tipo "categorical_numeric", también era posible especificar el parámetro "labels", donde se pueden proporcionar los valores de los códigos y las etiquetas asociadas de acuerdo con la documentación oficial proporcionada.
- **"categorical_character"**: se asignó a las variables que son de tipo categórica y tienen como valores en la base de datos textos, como, por ejemplo, el nombre completo del municipio en lugar de un código que lo represente.
- **"categorical_character_to_character"**: se asignó a las variables que son de tipo categórica y tienen como valores en la base de datos códigos no numéricos y que no representan el valor de la categoría de acuerdo con la documentación.
- **"date"**: se asignó a las variables que son de tipo fecha. Además del atributo "type", cuando la variable era del tipo "date", también era posible especificar el parámetro "format" donde se puede proporcionar el formato en el que la fecha debería ser configurada durante el procesamiento de los datos.
- **"character"**: se asignó a las variables que son de tipo texto y no pudieron ser encuadradas en los otros tipos de datos, como por ejemplo el nombre del paciente.

Ejemplo de estructura JSON para algunas variables:

```
{
 "edad": {
  "type": "numeric",
  "missing": [999]
 },
 "sexo": {
  "type": "categorical_numeric",
  "labels": {
   "1": "Hombre",
   "2": "Mujer",
   "3": "Indeterminado",
   "9": "No Especificado"
  }
 },
 "codmuer": {
  "type": "categorical_character"
 },
 "nac_dif": {
  "type": "categorical_character_to_character",
  "labels": {
   "C": "Chileno",
   "N": "Nacionalizada",
   "E": "Extranjero",
   "Vacios": "Desconocido"
  }
 },
 "fecha_def": {
  "type": "date",
  "format": "%d/%m/%Y"
 },
 "nommadr": {
  "type": "character"
 }
}
```

Esta estructura de metadatos juega un papel fundamental en el procesamiento y análisis de datos, asegurando la consistencia y la estandarización en todos los países y años involucrados en el estudio. Cabe destacar que todas las etiquetas de las variables "categorical_numeric" fueron normalizadas para garantizar la estandarización, incluyendo la conversión a titlecase (primera letra de cada palabra en mayúscula) y la eliminación de espacios y caracteres especiales para evitar conflictos en softwares.

3. Procesamiento de los Datos

3.1 Lectura de los Datos

El proceso de lectura de datos es un paso fundamental en el análisis y procesamiento de la información sobre mortalidad materna de los 18 países participantes. Para cada país y cada año, los datos presentan diferentes formatos y características, requiriendo un enfoque flexible y eficaz. La función **read_file** se desarrolló específicamente para satisfacer esta necesidad, facilitando la lectura y el procesamiento inicial de los datos, independientemente del formato del archivo fuente. Este procedimiento de lectura de datos se aplicó uniformemente en todos los países y para todos los años de datos disponibles.

La función **read_file** se puede acceder en este enlace: https://github.com/daltonbc96/dataOPS/blob/master/R/read_file.R

3.1.1 Descripción de la Función Construida para la Lectura de los Datos

La función **read_file** se diseñó para leer archivos de diferentes formatos y procesarlos en un objeto **data.table**, una estructura de datos eficiente y ampliamente utilizada en el lenguaje R. Esta función soporta la lectura de archivos con extensiones como CSV, TXT, DBF, SAV, XLS y XLSX. El código de la función está estructurado de la siguiente manera:

- **Identificación de la Extensión del Archivo**: La función comienza identificando la extensión del archivo proporcionado a través del camino del archivo (file_path).
- **Lectura Condicional del Archivo**: Dependiendo de la extensión del archivo, se usa un paquete R apropiado para leer los datos:
    - Archivos CSV y TXT se leen usando data.table::fread.
    - Archivos DBF se leen usando foreign::read.dbf.
    - Archivos SAV se leen usando foreign::read.spss.
    - Archivos XLS y XLSX se leen usando readxl::read_excel.

- **Procesamiento de Valores Ausentes**: La función reemplaza una serie de cadenas predefinidas (como "NA", "NULL", etc.) por NA para estandarizar la representación de valores ausentes.
- **Post-procesamiento de los Datos**: Después de la lectura, la función realiza algunas operaciones de limpieza, como la estandarización de los nombres de las columnas (convirtiéndolos a minúsculas y eliminando caracteres especiales) y la eliminación de espacios en blanco innecesarios.

3.1.2 Ejemplos Prácticos de Uso de la Función

Para ilustrar cómo se usa la función **read_file** en la práctica, considere el siguiente ejemplo. Supongamos que tenemos un archivo de datos disponible en un camino específico (file_path). El siguiente código muestra cómo se aplica la función para leer estos datos:

```
# Ejemplo de cómo usar la función read_file
# Suponga que file_path sea el camino para su archivo de datos
data <- read_file(file_path, sheet = 1)
```

En este ejemplo, la función **read_file** se llama con el camino del archivo y, opcionalmente, con el número de la hoja (en el caso de archivos XLS o XLSX). El resultado es un objeto **data.table** que contiene los datos leídos y procesados, listo para ser utilizado en las siguientes etapas de análisis.

Este método de lectura de datos asegura que los conjuntos de datos de diferentes países y años se traten de manera consistente, estableciendo una base sólida para los análisis posteriores de mortalidad materna.

3.2 Procesamiento Específico por País

El procesamiento específico de datos para cada país es una etapa esencial para asegurar que las variables de cada conjunto de datos estén alineadas y estandarizadas para análisis posteriores. Dada la diversidad en los formatos de datos y las peculiaridades de cada país, se realizaron procedimientos de adecuación y corrección según las necesidades específicas. Estos procedimientos variaron de país a país, afectando diferentes variables y requiriendo una serie de ajustes detallados, como por ejemplo la estandarización de la nomenclatura de la variable a lo largo de los años (renombramiento) y ajustar el número de dígitos para estar de acuerdo con la documentación (como añadir ceros a la izquierda para que el número tenga 3 dígitos). A

continuación, se describirá los procedimientos de procesamiento de datos específico realizados en cada país.

3.2.1 Argentina

- **Variables ‘provres’, ‘provoc’, ‘finstruc’**: Formateo de números con dos dígitos.
- **Variables ‘depres’, ‘depoc’**: Formateo de números con tres dígitos y unión de números (join_numbers).
- **Variable ‘mesdef’**: Extracción de mes y año, con sustitución del mes '00' por NA.

3.2.2 Brasil

- **Eliminación de la variable ‘d_r’**: Por no tener valores.
- **Limpieza de columnas (‘codificado’, ‘stcodifica’, ‘tppos’,’tpnivelinv’)**: Sustitución de ‘NULL’ por cadenas vacías y eliminación de caracteres especiales.
- **Eliminación del símbolo ‘*’**: En las variables linhaii, linhab, linhac, linhad, linhaa.
- **Adecuación de fechas**: Transformación de las variables 'dtatestado', 'dtcadastro', 'dtinvestig', 'dtnasc', 'dtobito','dtrecebim', 'dtrecorig', dtregcart', 'dtressele', 'dtconcaso', 'dtcadinf', 'dtconinv', 'dtcadinv', 'dtrecoriga', al formato ‘%d-%m-%Y’.
- **Decodificación de la variable 'edad'**: Extracción de edad y tipo de edad.
- **Variables 'escfalagr1' y 'escmaeagr1'**: Adecuación de los dígitos conforme a la documentación.

3.2.3 Chile

- **Renombramiento de Variables**:
    - 'ano1_nac' a 'ano_nac1'
    - 'ano2_nac' a 'ano_nac2'
    - 'c_medico' a 'cal_medico'
    - 'ocupacion' a 'ocupa'
- **Unión de Datos**: Variables 'ocupa' y 'activ' combinadas para formar 'ocupa'.

3.2.4 Colombia

- **Renombramiento de 'barriofall' a 'barriofal'**.
- **Combinaciones de Códigos de Localidades Conforme Documentación**:
    - 'código' formado por 'cod_dpto', 'cod_munic', 'codigo'.

    - 'cod_insp' formado por 'cod_dpto', 'cod_munic', 'cod_insp'.
    - 'cod_munic' formado por 'codptore', 'cod_munic'.
    - 'codmunre' formado por 'cod_dpto', 'codmunre'.
    - 'mm_exp' formado por 'dd_exp', 'mm_exp'.
- **Formateo de Dígitos**:
    - 'doc_id', 'doc_idm', 'localocuhe', 'tipoidcer', 'gru_ed2', 'cod_region', 'codmunoc', 'niv_edum', 'nivel_edu', 'cau_homol' ajustados a dos o tres dígitos.
- **Recodificación de Edad**: Transformación de la 'edad' (edad) con 'recodificar_edad' y 'codigo_para_edad'.

3.2.5 Costa Rica

- **Ningún ajuste específico realizado en los datos originales**.

3.2.6 Cuba

- **Eliminación de d_r**.
- **Renombramiento de Variables para Estandarización Entre Años**:
    - 'munocu' a 'lugoc'.
    - 'edadmad' a 'edadm'.
    - 'nacm' a 'nac_m'.
    - 'actm' a 'act_m'.
    - 'nacv' a 'nac_v'.
    - 'actv' a 'act_v'.
    - 'codcpiel' a 'raza'.
    - 'codocup' a 'ocup'.
- **Creación de la Variable 'ecivil2'**: en algunos años la codificación de 'ecivil' cambió al punto de no ser posible combinar entre los años y se configuró como variables separadas.
- **Configuración de 'ocup'**: Eliminación de ceros a la izquierda.
- **Procesamiento de 'letra' y 'ecant'**: Generación de letra de edad y edad numérica conforme orientación de la documentación.
- **Formateo de Fechas**: 'fecn' y 'fecd' ajustadas al formato de fecha.

3.2.7 Ecuador

- **Renombramiento de 'mor_mat' a 'mort_mat'**.
- **Conversión de Fechas**: Transformación de los formatos de 'fecha_fall', 'fecha_nac', 'fecha_insc' a un formato de fecha.
- **Limpieza de Columnas**: 'prov_insc', 'cant_insc', 'parr_insc' limpias y sustituidas por 'NA' cuando necesario.

3.2.8 El Salvador

- **Renombramiento de 'edadminu' a 'edadminuto'**.
- **Renombramiento de 'fecha_regi' a 'fecha_reg_'**.
- **Formateo de Fecha**: 'fecha_reg_' ajustada al formato de fecha.

3.2.9 Estados Unidos

- **Renombramiento de Variables**:
    - 'bridged_race' a 'race'.
    - 'hispanic_origin_bridged_race_recode' a 'hispanic_origin_race_recode'.
    - 'bridged_race_recode_5' a 'race_recode_5'.
    - 'bridged_race_recode_3' a 'race_recode_3'.
- **Limpieza de Columnas**: Eliminación de espacios y sustitución de cadenas vacías por 'NA'. Incluyen:

‘record_condition_1’, ‘record_condition_2’, ‘record_condition_3’, ‘record_condition_4’, ‘record_condition_5’, ‘record_condition_6’, ‘record_condition_7’, ‘record_condition_8’, ‘record_condition_9’, ‘record_condition_10’, ‘record_condition_11’, ‘record_condition_12’, ‘record_condition_13’, ‘record_condition_14’, ‘record_condition_15’, ‘record_condition_16’, ‘record_condition_17’, ‘record_condition_18’, ‘record_condition_19’, ‘record_condition_20’, ‘entity_condition_1’, ‘entity_condition_2’, ‘entity_condition_3’, ‘entity_condition_4’, ‘entity_condition_5’, ‘entity_condition_6’, ‘entity_condition_7’, ‘entity_condition_8’, ‘entity_condition_9’, ‘entity_condition_10’, ‘entity_condition_11’, ‘entity_condition_12’, ‘entity_condition_13’, ‘entity_condition_14’, ‘entity_condition_15’, ‘entity_condition_16’, ‘entity_condition_17’, ‘entity_condition_18’, ‘entity_condition_19’, ‘entity_condition_20’, ‘130_infant_cause_recode’, ‘occupation_recode’, ‘industry_4_digit_code’, ‘industry_recode’,

‘population_size_of_county_of_residence’, ‘population_size_of_county_of_occurrence’, ‘state_country_birth_recode’, ‘state_of_residence_fips’, ‘state_country_of_residence_recode’, ‘county_of_residence_fips’, ‘state_of_occurrence_fips’, ‘county_of_occurrence_fips’, ‘record_type’, ‘hispanic_origin_race_recode’, ‘race_recode_3’, ‘race_recode_5’,’race’.

- **Formateo de Números**: 'education_1989_revision' formateada a dos dígitos.
- **Procesamiento de Edad**: Decodificación de 'detail_age' conforme documentación.

3.2.10 Guatemala

- **Ningún ajuste específico realizado en los datos originales**.

3.2.11 México

- **Renombramiento de ‘loc_ocur’ a ‘loc_ocurr’ y ‘tipo_defun’ a ‘presunto’**.
- **Formateo de Números**: ‘ocupacion’ formateada a tres dígitos y ‘cve_lengua’ a cuatro dígitos.
- **Creación de la variable ‘ocupacion2’**: a lo largo de los años hay una variación en las categorías de ‘ocupacion’ que impide juntar las dos informaciones.
- **Procesamiento de la variable 'grupo' del CIE**: Basado en 'capitulo' y 'grupo' conforme documentación.
- **Unión de Códigos de Localización**: Conjunción de códigos para localización conforme documentación para las variables: ‘ent_resid’, ‘mun_resid’, ‘ent_ocurr’, ‘mun_ocurr’, ‘ent_ocules’, ‘mun_ocules’, ‘ent_regis’, ‘mun_regis’, ‘dis_re_oax’.

3.2.12 Nicaragua

- **Renombramiento de ‘dato’ a ‘datos’**.
- **Renombramiento de ‘fechaemision’ a ‘fecha_emision’**.

3.2.13 Panamá

- **Procesamiento de ‘def_edad’**: Decodificación de tipo y valor de edad conforme documentación.
- **Renombramiento de ‘grupo2’ a ‘grupo2_edad’, de ‘causabasica’ a ‘causa_basica’, ‘defuncion_causa_descripciona’ a ‘def_causa_descripciona’, ‘defuncion_causa_descripcionb’ a ‘def_causa_descripcionb’,**

‘defuncion_causa_descripcionc’ a ‘def_causa_descripcionc’, ‘defuncion_causa_descripcionp2’ a ‘def_causa_descripcionp2’ y ‘regionesdesalud’ a ‘regiondesalud’.

- **Conversión de Fecha**: ‘def_fecha’ ajustada al formato de fecha.

3.2.14 Paraguay

- **Renombramiento de Variables**:

'30departamentoderesidenciadelfallecidoa' a 'coddptor', '31distritoderesidenciadelfallecidoa' a 'coddistr', '13sexo' a 'sexo', '10fechadedefuncion' a 'fechadef', '27grupodeedad' a 'grupoed', '28valordelaedad' a 'valedad', '29niveleducativo' a 'nivedu', '21codigodeetnia' a 'codetnia', '20etnia' a 'esetnia', '26estadocivildelfallecidoa' a 'estciv', '15departamentodeocurrencia' a 'coddpto', '16distritodeocurrenciadeladefuncion' a 'coddist', '22tipodeinstitucion' a 'tipoinst', '68recibioasistenciaduranteelprocesoquelollevoalamuerte' a 'asistencia', '55edaddelamadre' a 'edadmadr', '60tiempodegestacion' a 'tgesta', '42tipodeparto' a 'tipopart', '58nodehijosnv' a 'nhijviv', '59nodehijosquenacieronyhanmuerto' a 'nhijmue', '69causaa' a 'causaa', '73causab' a 'causab', '77causac' a 'causac', '81causad' a 'causad', '90muertesincertificacionmedica' a 'msincer', '36probablemanerademuerte' a 'pmanmue', '64estuvoembarazadacuandofallecio' a 'embaraz', '65estuvoembarazadaenlasultimas6semanas' a 'emb6sem', '66estuvoembarazadaenlosultimos12meses' a 'emb12m', '93quienexpideelcertificado' a 'quienexcert'.

- **Creación de la variable ‘estciv2’**: a lo largo de los años hay una variación en las categorías de ‘estciv’ que impide juntar las dos informaciones.
- **Variable ‘coddist’ y ‘coddistr’ referentes a localización combinadas conforme documentación**.
- **Formateo de Fechas**: Conversión del formato de ‘fechadef’.

3.2.15 Perú

- **Renombramiento de Variables**: ‘causa_basica’ a ‘causa_basi’, ‘causabasic’ a ‘causa_basi’, ‘t_edad’ a ‘da32tiem’, ‘edad_a’ a ‘da32edad’, ‘tiempoedad’ a ‘da32tiem’, ‘tiempo_edad’ a ‘da32tiem’, ‘edad’ a ‘da32edad’, ‘idd’ a ‘id’, ‘sexo’ a ‘da31sexo’, ‘edad_final’ a ‘da32edad’, ‘estadocivil’ a ‘da33conyu’, ‘niveldeinstruccion’ a ‘da34codniv’, ‘ocupacion’ a ‘da35codocu’, ‘fecha’ a ‘lf56fecha’, ‘hora’ a ‘lf56hora’,

‘tipolugar’ a ‘lf57codsit’, ‘caubasf’ a ‘cd63caubas’, ‘muerteviolenta’ a ‘cd64codvio’, ‘ncolegio’ a ‘dp74cmp’, ‘fecharegistro’ a ‘fecregistr’, ‘tipo_edad_final’ a ‘da32tiem’.

- **Formateo de Fechas**: Conversión de ‘fecregistr’ al formato de fecha.

3.2.16 República Dominicana

- **Ningún ajuste específico realizado en los datos originales**.

3.2.17 Uruguay

- **Renombramiento de Variables**: ‘lugardeladefuncion_new’ a ‘lugardeladefuncion’, ‘pais_de_nacimiento_original’ a ‘paisdenacimiento’, ‘tramoetario2’ a ‘tramoetario’, ‘enfprincipalmadreafectaalfetoc’ a ‘enfmadreafectofetoc’, ‘otrasenfmadreafectaalfetod’ a ‘otraenfmadreaffetod’, ‘otrascircunstanciase’ a ‘otrascircunstpertinentese’, ‘enfmadreafectafetocie10’ a ‘enfmadreafectofetocie10’, ‘edadmadre’ a ‘edad_madre’, ‘edad_de_la_madre’ a ‘edad_madre’, ‘fechadedefuncion’ a ‘fechadeladefuncion’, ‘departamentodeocurrencia_original’ a ‘departamentodeocurrencia’, ‘otrosestadosmorbidos’ a ‘otrosestadosmorbidosd’, ‘enfmadreaffetoc’ a ‘enfmadreafectofetoc’, ‘efermedadcausantea’ a ‘enfermedadcausantea’, ‘lugardeladefuncion_old’ a ‘lugardeladefuncion’.
- **Creación de la variable ‘etnia2’**: a lo largo de los años hay una variación en las categorías de ‘etnia’ que impide juntar las dos informaciones.
- **Creación de la variable ‘mayornivelalcanzado2’**: a lo largo de los años hay una variación en las categorías de ‘mayornivelalcanzado’ que impide juntar las dos informaciones.
- **Creación de la variable ‘ocupacion2’**: a lo largo de los años hay una variación en las categorías de ‘ocupacion’ que impide juntar las dos informaciones.
- **Creación de la variable ‘lugardeladefuncion2’**: a lo largo de los años hay una variación en las categorías de ‘lugardeladefuncion’ que impide juntar las dos informaciones.
- **Procesamiento de Fechas**: Conversión del formato de fechas para variables ‘horadeladefuncion’ y ‘horadenacimiento’.

3.2.18 Venezuela

- **Eliminación de 'd_r'**.

- **Unión de Códigos de Localización**: Conjunción de códigos para localización conforme documentación para las variables: ‘mcpocurr’, ‘munpres’, ‘dtoocurr’, ‘dtores’.
- **Formateo de Fechas**: Conversión de los formatos de ‘fecnac’ y ‘fecmte’.

3.3 Procesamiento General con Base en los Metadatos

Después del ajuste individual de cada país, ahora procederemos al procesamiento general de los datos utilizando las informaciones de los metadatos que están en el archivo JSON. Para ello, se desarrolló la función **'process_mortality'** que contiene un pipeline de procesamiento de datos. Diseñada para funcionar juntamente con las informaciones detalladas contenidas en un archivo de metadatos JSON, esta función garantiza la estandarización y precisión en la manipulación de los datos. Este procedimiento se aplica uniformemente a todos los países y años de datos, utilizando reglas y formatos definidos en el JSON para adaptar los datos a las necesidades específicas de análisis. Este sistema de procesamiento ofrece flexibilidad y eficiencia, permitiendo que los cambios en los requisitos de datos se gestionen de manera centralizada y consistente.

La función **'process_mortality'** está disponible públicamente y se puede acceder a través del enlace: https://github.com/daltonbc96/dataOPS/blob/master/R/process_mortality.R

3.3.1 Detalle del funcionamiento de la función 'process_mortality'

La función **'process_mortality'** utiliza las informaciones de un archivo de metadatos JSON para procesar conjuntos de datos de mortalidad. Cada tipo de dato ('type') en el JSON determina un conjunto específico de operaciones que se aplican a la columna correspondiente en el conjunto de datos. Vamos a detallar el procesamiento para cada 'type':

1. Type "**character**":
   - **Limpieza de Strings**: Elimina puntuaciones, convierte caracteres a ASCII, transforma todo en minúsculas y capitaliza strings.
   - **Tratamiento de NAs**: Sustituye strings vacías por valores ‘NA’.
2. Type "**numeric**":
   - **Tratamiento de Valores Ausentes**: Si el metadato especifica valores que deben considerarse como ausentes ('missing'), estos son sustituidos por ‘NA’.
   - **Conversión a Numérico**: Transforma la columna en tipo numérico.
3. Type "**date**":

- **Conversión de Formatos de Fecha**: Las fechas se convierten al formato estándar ("%d/%m/%Y") o a un formato especificado en el JSON.
- **Tratamiento de Formatos Diversos**: La función maneja varias representaciones de fecha, incluyendo fechas de SPSS y Excel.

4. Type "**time**":
   - **Tratamiento de 'NAs'**: Sustituye strings vacías por 'NA'.
   - **Conversión a Tiempo**: Transforma strings en objetos de tiempo.
5. Type "**categorical_character**":
   - **Limpieza de Strings**: Elimina puntuaciones, convierte caracteres a ASCII, transforma todo en minúsculas y capitaliza strings.
   - **Conversión a Factor**: Cada valor único se convierte en un nivel de un factor.
6. Type "**categorical_numeric**":
   - **Mapeo de Valores a Etiquetas**: Los valores numéricos se mapean a etiquetas textuales específicas, conforme definido en el JSON.
   - **Tratamiento de Valores Ausentes y Adición de Etiquetas Faltantes**: Valores ausentes se asignan valores NA, y se añaden etiquetas para valores que no tienen una etiqueta especificada en el JSON, se añade el mensaje "Sin Etiqueta Para El Valor X".
7. Type "**categorical_character_to_character**":
   - **Mapeo de Códigos a Etiquetas**: Similar al 'type' "categorical_numeric", pero para códigos que no son numéricos.

   Además de estas operaciones, hay funcionalidades opcionales:
   - Crea la Variable **'maternal_mortality_cases'**: basada en una columna específica donde haya códigos de CIE-10 de la causa de defunción.
   - Añade una Variable **'reference_year'**: Si se proporciona un año de referencia, se añade al conjunto de datos.

Cada una de estas etapas se ejecuta iterativamente para todas las variables listadas en el archivo JSON, asegurando que el conjunto de datos final esté limpio, estandarizado y listo para análisis.

3.3.2 Ejemplos del Procesamiento Realizado

Ejemplo de uso de la función **'process_mortality'**:

```
# Carga de los paquetes necesarios
library(data.table)
library(jsonlite)

# Ejemplo de uso de la función 'process_mortality'
# Suponiendo que `mortality_data` sean los datos y `metadata.json` el camino del archivo de metadatos
processed_data <- process_mortality(mortality_data,
                                    "metadata.json",
                                    set_reference_year = 2020,
                                    var_maternal_mortality_cases = "icd_code_column")
```

Este ejemplo demuestra cómo se utiliza la función 'process_mortality' con datos de mortalidad ('mortality_data'), un archivo JSON de metadatos ('metadata.json'), un año de referencia opcional (2020) y una variable opcional para casos de mortalidad materna ('icd_code_column'). El resultado es un objeto data.table procesado conforme a las definiciones del archivo JSON.

Para la lectura inicial de los datos, se debe utilizar la función 'read_file' para asegurar que los datos estén en el formato adecuado antes de aplicar el procesamiento con 'process_mortality'.

4. Generación del Reporte y EDA (Análisis Exploratorio de Datos)

Después de la estandarización y la verificación de las variables en todos los países y años, las bases de datos anuales de cada país se consolidan en una serie única. Esta unificación es posible gracias a la estandarización previa de los nombres de las variables y las etiquetas. El objetivo principal en esta fase fue analizar la calidad y las características de las informaciones contenidas en cada variable, a través de un Análisis Exploratorio de Datos (EDA).

4.1 ¿Qué es EDA (Análisis Exploratorio de Datos)?

El EDA es un enfoque estadístico que busca explorar y analizar conjuntos de datos para resumir sus características principales, a menudo utilizando métodos visuales. El proceso de EDA implica el examen minucioso de patrones, anomalías, tendencias y relaciones dentro de los datos. La idea es usar técnicas visuales y estadísticas para obtener *insights* iniciales sobre

los datos, informar hipótesis futuras y preparar los datos para modelado más avanzado. El EDA es una etapa crucial en el proceso de análisis de datos, ya que ayuda a comprender mejor el contexto y la estructura de los datos, facilitando la identificación de posibles direcciones o áreas para investigación más profunda.

4.2 Procedimientos para EDA

Antes de iniciar el EDA, se realiza un filtrado de los datos basado en criterios específicos:

- **Inclusión de Mujeres**: Solo se incluyen casos de muerte de mujeres en el análisis.
- **Rango de Edad**: La edad de las mujeres en el momento de la muerte se restringe a 10 a 54 años, correspondiendo a la edad fértil.

4.3 Herramientas Utilizadas para EDA

Para realizar el EDA, se utiliza el lenguaje Python y la biblioteca '**pandas-profiling**'. Esta biblioteca ofrece una manera eficiente de generar informes de EDA, proporcionando visiones amplias de los datos a través de estadísticas descriptivas, distribuciones de variables, correlaciones y otros aspectos relevantes. Los informes generados ofrecen una visión detallada de cada variable, facilitando la identificación de patrones, valores ausentes, distribuciones de frecuencia, entre otros aspectos cruciales para la comprensión de los datos.

4.4 Acceso al Reporte

Al utilizar '**pandas-profiling'** para generar EDA, se guardó el informe de las variables para cada país en un archivo .html. Estos archivos están en la carpeta de Reports del SharePoint.

5. Análisis Especializado y Selección de Variables

En la fase posterior al procesamiento de datos, filtrado y generación de informes descriptivos, la serie unificada y filtrada de datos sobre mortalidad materna se somete a un análisis especializado. Este proceso crucial se lleva a cabo en todos los países participantes del proyecto. El objetivo de esta etapa es contar con la experiencia de un especialista en el dominio de la salud materna para identificar y seleccionar las variables más relevantes que serán incorporadas en la herramienta analítica final. Este procedimiento asegura que la herramienta no solo contenga información completa, sino que también se centre en las variables más críticas e informativas para el análisis de la mortalidad materna.

5.1 Proceso de Análisis y Selección

- **Análisis Especializado**: un especialista en el área de salud materna, con profundo conocimiento y experiencia en cuestiones relacionadas con la mortalidad materna, analiza la serie de datos. Este análisis se informa tanto por los informes de EDA generados anteriormente como por el conocimiento especializado del analista.
- **Criterios de Selección**: el especialista utiliza múltiples criterios para la selección de variables, incluyendo:
    - **Relevancia Clínica y Epidemiológica**: importancia de las variables en el contexto de la salud materna y su relación con los factores de riesgo y desenlaces de la mortalidad materna.
    - **Calidad de los Datos**: evaluación de la integridad, precisión y consistencia de las variables basada en los informes de EDA y las series de los países.
    - **Disponibilidad y Comparabilidad**: verificación de la disponibilidad consistente de las variables a lo largo de los años y entre los diferentes países.
- **Selección Final de Variables**: basándose en este análisis, el especialista identifica un conjunto de variables clave. Estas variables son seleccionadas por su capacidad de proporcionar *insights* significativos y por su aplicabilidad en el análisis transnacional y longitudinal de la mortalidad materna.
- **Retroalimentación e Iteración**: el proceso de selección puede ser iterativo, con el especialista proporcionando retroalimentación para ajustes adicionales en los datos o solicitando más información. Esta colaboración continua ayuda a refinar la selección de variables, asegurando que la herramienta final sea robusta y relevante.

5.2 Variables Finales Seleccionadas

El resultado de esta etapa es una lista refinada de variables que ofrecen el mayor potencial para análisis significativos y accionables en el contexto de la mortalidad materna. La selección cuidadosa de estas variables garantiza que la herramienta analítica final no solo sea completa e informativa, sino también enfocada y relevante para los desafíos y necesidades específicos en el campo de la salud materna.

6. Organización del Archivo Final

Después de la meticulosa selección de variables por un especialista, como se describió en la etapa anterior, procedemos a la fase final de organización de los datos. Esta etapa es

esencial para garantizar que la aplicación final se alimente de un conjunto de datos refinado, que refleje las elecciones informadas por el especialista. Este proceso se realizó uniformemente para todos los países y años de datos, asegurando consistencia y fiabilidad en la aplicación final.

6.1 Proceso de Refinamiento de la Serie Única

- **Consolidación de las Variables Seleccionadas**: Basándose en la lista de variables definidas por el especialista, la serie unificada de datos de cada país se refina. Esto implica la eliminación de columnas que no fueron seleccionadas, manteniendo solo aquellas consideradas más relevantes para el análisis.
- **Creación de la Versión Final del Conjunto de Datos**: La serie unificada y refinada constituye la versión final del conjunto de datos. Esta versión contiene solo las informaciones esenciales y pertinentes, asegurando que la aplicación final se alimente de datos precisos y relevantes.
- **Preparación para la Integración en la Aplicación**: Los datos refinados se preparan para ser integrados en la aplicación final. Este preparo implica asegurar que los datos estén en el formato correcto y que sean compatibles con los sistemas y tecnologías utilizados en la aplicación.

El resultado de este proceso es un archivo de datos final, optimizado para su uso en la aplicación de análisis de mortalidad materna. Este archivo contiene informaciones cruciales y validadas, listas para ser utilizadas en análisis posteriores y en la toma de decisiones informadas en el campo de la salud materna. La organización cuidadosa y el refinamiento de los datos son pasos decisivos para garantizar la eficacia y la eficiencia de la aplicación final.

7. Conclusión

El proyecto de procesamiento y análisis de datos de mortalidad materna abarcó varias fases críticas, cada una contribuyendo significativamente a la construcción de una herramienta analítica robusta e informativa. Aquí está un resumen de los pasos realizados:

- **Desarrollo de un Pipeline de Procesamiento de Datos**: Implementamos un pipeline utilizando R para procesar datos de 18 países, abarcando los años desde 2015 hasta el más reciente disponible.
- **Creación de Diccionario de Datos y Archivo JSON de Metadatos**: Anotamos manualmente las variables en un archivo Excel y las convertimos en un archivo JSON para estandarizar el procesamiento de los datos.

- **Lectura y Procesamiento Específico por País**: Desarrollamos funciones específicas para cada país para ajustar variables según las necesidades locales.
- **Procesamiento General con Base en JSON**: Utilizamos la función **'process_mortality'** para estandarizar y procesar los datos basándonos en las especificaciones del JSON.
- **Generación de Reporte y Análisis Exploratorio de Datos (EDA)**: Realizamos EDA usando Python y '**pandas-profiling'** para investigar la calidad y características de cada variable.
- **Análisis Especializado y Selección de Variables**: Un especialista revisó la serie de datos unificada para seleccionar las variables más relevantes para el análisis de mortalidad materna.
- **Organización del Archivo Final**: Los datos fueron refinados para incluir solo las variables seleccionadas, resultando en un conjunto de datos optimizado para la aplicación final.

Este proyecto representa un esfuerzo amplio y detallado para abordar la compleja cuestión de la mortalidad materna en diversos contextos nacionales. A través de una combinación de técnicas avanzadas de procesamiento de datos, análisis exploratorio y experiencia especializada, fue posible sintetizar un gran volumen de datos en un recurso informativo y accesible. El resultado es una herramienta analítica refinada, capaz de ofrecer *insights* valiosos para la salud materna y orientar intervenciones eficaces y políticas basadas en evidencia. Este proyecto no solo resalta la importancia del análisis de datos en la salud pública, sino que también establece un modelo para futuras iniciativas de análisis de datos en otras áreas críticas de la salud y el desarrollo social.

8. Enlaces Útiles

- Repositorio con Códigos de Procesamiento de Datos: https://github.com/daltonbc96/dataOPS
- Repositorio con códigos de la Herramienta de Mortalidad Materna: https://github.com/daltonbc96/maternal-mortality-analysis-tool