

Exploración y Estructuración de Datos Abiertos de Salud en Uruguay: Un Primer Paso hacia el Análisis Avanzado.

INFORME DE ACTIVIDADES

1. Introducción

La Sociedad Uruguaya de Salud Digital (SUSADI), a través de su grupo de trabajo de Datos y Analítica de Datos en Salud, ha venido trabajando en **explorar y estructurar los datos abiertos disponibles sobre el sistema de salud en Uruguay**. El objetivo principal es transformar información dispersa y a menudo poco estructurada en recursos **reutilizables y visualizables** para el análisis exploratorio y la toma de decisiones. Este esfuerzo inicial se centró en la revisión de repositorios públicos como **REDATA (ANII)**, **el Catálogo Nacional de Datos Abiertos** y **el Ministerio de Salud Pública (MSP)**. Se identificaron desafíos comunes, como **inconsistencias, falta de metadatos y estructuras orientadas a la presentación** en lugar del análisis automatizado. A pesar de estas limitaciones, se avanzó en el procesamiento de dos datasets clave, buscando aportar en la **mejora de la calidad y usabilidad de los datos sanitarios** en el país.

2. Metodología de Trabajo y Enfoque

El equipo adoptó un **enfoque práctico y colaborativo**, utilizando herramientas como *R*, *RStudio*, *Python (Pandas y Plotly)*, *Jupyter Notebooks* y *Google Colab*. La comunicación y el intercambio de información se facilitaron a través de un espacio de conocimiento en **Discord** (aunque posteriormente se migró a Google Drive y WhatsApp para mayor agilidad).

Las actividades principales incluyeron:

- **Estudio de puntos de dolor y cuellos de botella** en el sistema de salud susceptibles de mejora a través del análisis de datos.
- **Relevamiento y evaluación de fuentes de datos** disponibles, considerando su calidad y actualidad.
- **Selección y procesamiento de datasets relevantes**, priorizando aquellos con mayor potencial analítico.
- **Limpieza, normalización y reestructuración** de los datos para transformarlos en **formato tidy** (ordenado), fundamental para el análisis automatizado.
- **Elaboración de diccionarios de variables** para las bases de datos generadas.

3. Principales Hallazgos y Trabajo Realizado

El relevamiento inicial reveló que, si bien existen datos abiertos, su calidad y estructura presentan **desafíos significativos**. Muchas bases tienen:

- **Intervalos temporales incompletos o irregulares.**
- **Discontinuidades de registro.**
- **Estructuras diseñadas para la presentación visual**, no para el análisis.
- **Inconsistencias internas** y entre diferentes catálogos. (nombres, codificaciones, formatos).
- **Ausencia de metadatos detallados** y documentación técnica.

A pesar de estas limitaciones, el equipo seleccionó y trabajó con dos datasets clave del MSP:

3.1. Dataset "Población afiliada a las IAMC por tipo de afiliación (2012-2024)"

Este dataset, que contiene información sobre la afiliación a instituciones de asistencia médica colectiva (IAMC), fue seleccionado para un procesamiento exhaustivo. Las transformaciones clave realizadas incluyeron:

- **Limpieza y Normalización:** Se estandarizaron los nombres de las instituciones (creando un diccionario de correspondencias) y se homogeneizaron los formatos de fecha a ISO (YYYY-MM-DD) para facilitar el análisis temporal.
- **Reestructuración a Formato Tidy:** Se migró de una estructura orientada a la presentación (varias hojas o celdas combinadas) a un único dataframe relacional. Esto significa que:
 - Cada fila representa una institución en un año dado,
 - Cada variable (total de afiliados, tipo de afiliación) ocupa su propia columna.
 - Se eliminó la estructura multitabla o con celdas combinadas.

3.2. Dataset "Serie de Indicadores Asistenciales IAMC (2005-2023)"

Este dataset, que contiene información histórica sobre consultas, egresos e internaciones, también requirió un procesamiento similar. Los pasos clave fueron:

- **Normalización de Nombres de Instituciones:** Se construyó una tabla de correspondencia para asignar un **nombre estandarizado** a cada institución, eliminando duplicaciones y facilitando la vinculación con otras fuentes. Este proceso se realizó en R Studio utilizando sentencias condicionales.
- **Reestructuración de los Datos:** Los archivos originales, que tenían una hoja distinta para cada atributo y los años como columnas, fueron reorganizados. Se consolidaron en una única tabla en **formato tidy**, donde cada fila representa una combinación única de institución y año, y las variables (consultas, egresos, internaciones) ocupan columnas separadas.

Ambos procesos de reestructuración siguieron los principios del **formato tidy de Hadley Wickham**: "cada variable es una columna, cada observación es una fila, y cada valor es una celda". Este formato es **fundamental para el análisis descriptivo, la construcción de dashboards y la integración** con otras fuentes de datos.

4. Conclusión

Este trabajo representa un **primer paso** para la mejora del ecosistema de datos abiertos en salud en Uruguay. Al procesar y estructurar dos datasets clave, se ha demostrado la viabilidad de transformar información dispersa y compleja en **bases de datos limpias, estandarizadas y listas para el análisis**.

Los principales logros incluyen:

- **Generación de estructuras reutilizables** que permiten desarrollar indicadores longitudinales y análisis exploratorios robustos.
- **Homogeneización de nombres institucionales y formatos de fecha**, facilitando la integración con otras fuentes y mejorando la calidad de los datos.
- **Consolidación de información dispersa** en tablas únicas en **formato tidy**, lo que simplifica significativamente el procesamiento programático y la visualización.
- **Mejora sustantiva en la usabilidad de los datos**, haciéndolos accesibles tanto para públicos técnicos como no técnicos a través de futuras visualizaciones interactivas.

Este esfuerzo promueve una **gestión de la información sanitaria más confiable** y sienta las bases para la **toma de decisiones basada en evidencia**, crucial para el avance del sistema de salud uruguayo.

5. Nuevas Líneas de Trabajo y Recomendaciones

A partir de lo realizado, se abren diversas oportunidades y se sugieren las siguientes líneas de trabajo y recomendaciones:

5.1. Próximos Pasos en el Análisis de Datos

Una vez estructuradas las bases, el siguiente paso es la **explotación analítica y visualización interactiva**:

- **Visualización Interactiva**: Desarrollar gráficos de barras, líneas y mapas de calor utilizando herramientas como Plotly para detectar patrones y tendencias entre instituciones y años.

- **Paneles de Exploración (Dashboards):** Crear interfaces con filtros dinámicos (por institución, año, tipo de afiliación) que permitan a los usuarios no técnicos explorar los datos de manera intuitiva.
- **Mapas Georreferenciados:** Integrar capas del INE para vincular instituciones con su ubicación geográfica y visualizar métricas relevantes espacialmente.
- **Exportación Amigable de Datos:** Habilitar la descarga de datos en formatos comunes (ej. Excel) para facilitar su uso por parte de tomadores de decisión.

5.2. Recomendaciones para la Publicación de Datos Abiertos

Para mejorar el ecosistema de datos abiertos en salud en Uruguay, es crucial que los organismos públicos adopten las siguientes prácticas:

- **Priorizar el Formato Tidy:** Publicar siempre los datasets con una variable por columna, una observación por fila y un valor por celda. Esto maximiza su reutilización por analistas e investigadores.
- **Evitar Formatos Orientados a la Presentación:** No utilizar celdas combinadas, múltiples hojas para un mismo dataset, títulos incrustados en el cuerpo de la tabla o codificación de categorías por colores.
- **Estandarizar Nombres y Codificaciones:** Utilizar nombres consistentes y estandarizados para instituciones y categorías. Preferir el uso de códigos únicos (ID) para vincular bases de datos.
- **Publicar Diccionarios de Datos y Metadatos:** Acompañar cada dataset con una descripción clara de las variables, su fuente, el significado de valores faltantes, la fecha de actualización y la frecuencia de publicación.
- **Mantener Trazabilidad y Versionado:** Indicar la versión del dataset y los cambios respecto a versiones anteriores. Utilizar nombres de archivos claros que incluyan fechas y versiones.
- **Documentar el Proceso de Producción de Datos:** Explicar cómo se originaron los datos y si fueron modificados, imputados o agregados antes de su publicación.
- **Facilitar el Acceso Programático:** Publicar los datos en repositorios accesibles por URL directa (como GitHub o plataformas CKAN) y considerar ofrecer APIs de consulta para bases de datos grandes o dinámicas.
- **Incluir Ejemplos de Uso y Guías de Análisis:** Ofrecer notebooks o scripts de ejemplo (en R o Python) que demuestren cómo cargar y trabajar con los datos, para reducir la barrera de entrada para nuevos usuarios.

Grupo de Trabajo “Datos y Analítica de Datos SUSADI”

Camila Goldman - Marina Ridao - Fernando Cormenzana - Bruno Demoro - Alvaro Sanchez

Junio 2025