Andrés Mauricio Plazas González

PREICA2502B020074

Programa Académico:

Ingeniería De Software y Datos

Actividades:

Proyecto Integrado (EA3).

Presentación del proyecto y dashboard descriptivo

Asignatura:

Proyecto Integrado V

Docente:

Andrés Felipe Callejas

Institución Universidad Digital de Antioquia

07 de diciembre del 2025

Introducción

Este proyecto aborda la necesidad crítica de transformar datos brutos y fragmentados sobre riesgos laborales en una fuente de información estructurada e inteligente que facilite la toma de decisiones en materia de prevención, gestión de la accidentalidad y distribución de cobertura. El objetivo principal es establecer una solución integral de ingeniería de datos que migre los registros históricos a una base de datos local y, posteriormente, construir una herramienta de Inteligencia de Negocios (BI) interactiva para el monitoreo de indicadores clave.

1. Metodología de Ejecución (Las Tres Fases): El proyecto se desarrolló siguiendo una metodología rigurosa estructurada en tres fases principales, como se detalla en el diagrama de Gantt del proyecto:

- **Fase 1: Extracción, Transformación y Carga (ETL) (Actividad 1).**
  - **Propósito:** Garantizar la persistencia y la estructura inicial del dato.
  - **Proceso:** Se construyó un pipeline en Python (main.py) para leer el archivo fuente (CSV), aplicar transformaciones básicas de limpieza, e insertar los datos en una base de datos relacional local (SQLite), culminando con la exportación de una versión limpia del dataset (db/export.csv).
- **Fase 2: Limpieza Avanzada, Enriquecimiento y Análisis Descriptivo (EDA) (Actividad 2).**
  - **Propósito:** Asegurar la calidad analítica y explorar los hallazgos.
  - **Proceso:** Se utilizó un Notebook de análisis (02_enriquecimiento_eda.ipynb) para realizar la limpieza avanzada (manejo de nulos, inconsistencias, normalización de columnas) y el enriquecimiento mediante la creación de variables temporales (fecha, año, mes). Finalmente, se aplicó el EDA para obtener las primeras estadísticas descriptivas y las visualizaciones iniciales de las 5 variables clave.
- **Fase 3: Visualización de Resultados y Documentación Final (Actividad 3).**
  - **Propósito:** Entregar la solución de BI funcional y la documentación formal.
  - **Proceso:** Se diseñó y construyó un Dashboard interactivo (utilizando Streamlit/Power BI) para visualizar de forma consolidada las 5 variables clave (accidentalidad, fatalidad, cobertura, actividad económica, departamento), permitiendo a los usuarios filtrar y comparar datos. Esta fase culmina con la actualización del reporte formal y la presentación de resultados.

2. Tecnologías y Entregable Final: El proyecto se basa en Python como lenguaje principal, utilizando SQLite para la persistencia de datos y la librería Pandas para la manipulación y el análisis El entregable final es una solución de BI operacional (el dashboard interactivo) y un repositorio GitHub con una metodología de trabajo completamente documentada, demostrando la capacidad de transformar datos sin procesar en conocimiento accionable para la prevención de riesgos laborales.

Resumen

El proyecto fue concebido para resolver el desafío de analizar la accidentalidad, fatalidad y cobertura de riesgos laborales a partir de una fuente de datos sin procesar. La solución se construyó mediante un proceso de tres etapas que transformó el dato en conocimiento accionable.

Actividad 1: Ingeniería ETL y Estructuración de Datos

Foco: Construir una base de datos local robusta y persistente.

- **Tareas Clave:** Se realizó la Descarga y Revisión del dataset, seguida por la Definición de Métricas (las 5 variables clave) y el Diseño de Esquema para la tabla reporte_arl en SQLite.
- **Desarrollo:** Se codificó el script principal (main.py) para automatizar el flujo ETL, asegurando la Extracción del CSV, la Transformación inicial (limpieza básica de formatos) y la Carga en la base de datos SQLite.
- **Resultado:** Un archivo CSV limpio (db/export.csv) y una base de datos (db/proyecto.db) con la información estructurada, lista para la fase analítica.

Actividad 2: Limpieza Avanzada, Enriquecimiento y EDA

Foco: Asegurar la calidad analítica del dato y obtener los primeros hallazgos.

- **Limpieza y Normalización:** Se usó un Notebook de análisis (02_enriquecimiento_eda.ipynb) para abordar problemas complejos como la normalización de columnas (minúsculas, sin acentos), el manejo exhaustivo de duplicados y nulos, y la corrección de tipos de datos para las métricas numéricas.

- **Enriquecimiento:** Se incrementó el valor del dataset creando la columna fecha de tipo datetime y derivando variables clave como anio y mes, esenciales para el análisis de tendencias.
- **Análisis Descriptivo (EDA):** Se calcularon las estadísticas básicas (promedio, conteo, máximo) y se generaron visualizaciones clave (gráficos de barras y líneas) que revelaron la distribución del riesgo por actividad económica y su evolución temporal.
- Resultado: El dataset enriquecido y listo para BI (data/dataset_enriquecido.csv), junto con las interpretaciones iniciales que validaron la calidad del dato.

Actividad 3: Visualización de Resultados y Documentación Final

Foco: Entregar la solución de BI interactiva y formalizar la documentación del proyecto.

- **Diseño del Dashboard:** Se construyó un Dashboard interactivo (usando Streamlit) que consolida las 5 variables clave mediante KPIs, gráficos de barras para comparar categorías (ej. Top 10 Actividades) y gráficos de líneas para mostrar la tendencia temporal de incidentes y muertes.
- **Documentación:** Se formalizó la metodología, incluyendo la creación del Diagrama de Gantt detallado y la actualización del Reporte Final bajo normas APA, incorporando las capturas y hallazgos del dashboard.
- **Resultado:** Una herramienta de visualización funcional que permite a los responsables tomar decisiones informadas sobre la prevención de riesgos, y un repositorio GitHub totalmente documentado.

Este enfoque secuencial garantizó la solidez del proceso, desde la integridad de la fuente hasta la presentación de resultados clave del negocio.

Objetivo General

Implementar una solución integral de ingeniería de datos, desde la ingesta hasta la visualización, para transformar datos fragmentados de riesgos laborales en conocimiento accionable, permitiendo el monitoreo continuo de indicadores clave (accidentalidad, fatalidad, y cobertura) con el fin de informar la toma de decisiones estratégicas de prevención y gestión de riesgos.

Objetivo Específico

Los siguientes objetivos desglosan las tres fases del proyecto y sus entregables técnicos requeridos:

1. **Ingeniería de Datos (ETL):** Desarrollar un *pipeline* ETL eficiente y robusto, utilizando Python y SQLite, para la extracción, limpieza inicial y carga de los datos de riesgos laborales, asegurando su persistencia en una base de datos local y la exportación de un conjunto de datos estructurado (db/export.csv).
2. **Análisis y Enriquecimiento:** Garantizar la calidad analítica de los datos mediante la limpieza avanzada (manejo de nulos, normalización de formatos), el enriquecimiento (creación de variables de tiempo fecha, año, mes), y la realización de un Análisis Exploratorio de Datos (EDA) para obtener estadísticas descriptivas y validar las primeras tendencias de riesgo.
3. **Visualización y Entrega:** Diseñar y construir un Dashboard interactivo (utilizando Streamlit o Power BI) que visualice de manera efectiva las 5 variables clave del proyecto, permitiendo la comparación y segmentación de datos, y formalizar la documentación completa del proyecto (Gantt, Informe APA) para su entrega final.

Metodología

La metodología de este proyecto fue un proceso iterativo y secuencial que siguió el ciclo completo de la ingeniería de datos: Extracción, Transformación y Carga (ETL), Análisis y Preparación (EDA), y Visualización y Entrega Final.

Fase 1: Ingeniería de Datos (ETL)

La primera fase se centró en la estructuración y persistencia de los datos. Se diseñó un *pipeline* robusto en Python (main.py) para automatizar el flujo de trabajo.

1. **Extracción y Carga:** Se leyó el archivo CSV fuente, y se aplicaron transformaciones iniciales de estandarización. Los datos resultantes fueron cargados en una base de datos local SQLite (db/proyecto.db), garantizando su durabilidad y un esquema coherente.
2. **Salida Limpia:** El resultado de la carga en SQLite fue exportado nuevamente a un archivo CSV (db/export.csv), estableciendo así la fuente de datos base para las siguientes actividades, ya liberada de las inconsistencias del archivo plano original.

Fase 2: Preparación Analítica y EDA

La segunda fase se enfocó en asegurar la calidad analítica del dataset y la exploración de hallazgos.

1. **Limpieza Avanzada:** Utilizando el Notebook de análisis (02_enriquecimiento_eda.ipynb) con la librería Pandas, se aplicaron técnicas de limpieza exhaustiva. Esto incluyó la normalización completa de los nombres de columnas (a minúsculas y sin caracteres especiales), la corrección de tipos de datos para las métricas de conteo (inc_at, muertes), y el manejo de valores nulos y duplicados.
2. **Enriquecimiento:** Se incrementó el valor del dataset creando variables derivadas. La tarea principal fue la construcción de la columna fecha (datetime) a partir de las columnas de año y mes, permitiendo derivar variables temporales cruciales (anio, mes).
3. **Análisis Descriptivo (EDA):** Se calcularon estadísticas descriptivas (promedios, sumas, rangos) y se generaron visualizaciones clave (gráficos de barras y líneas) que

revelaron patrones iniciales sobre la distribución del riesgo por actividad económica y su evolución temporal.

Fase 3: Visualización de Resultados y Entrega Final

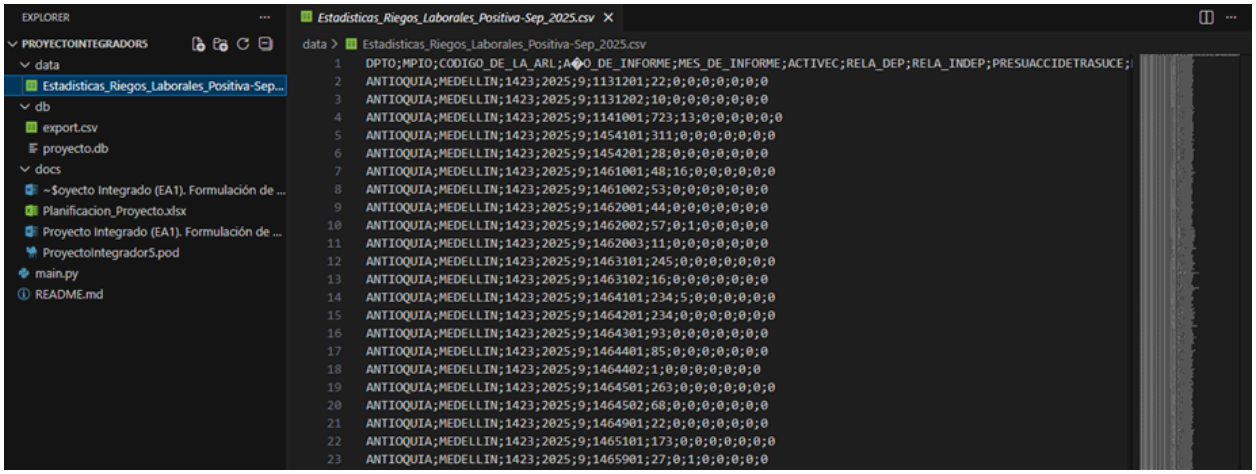
La fase final se dedicó a la comunicación de los hallazgos y la formalización del proyecto.

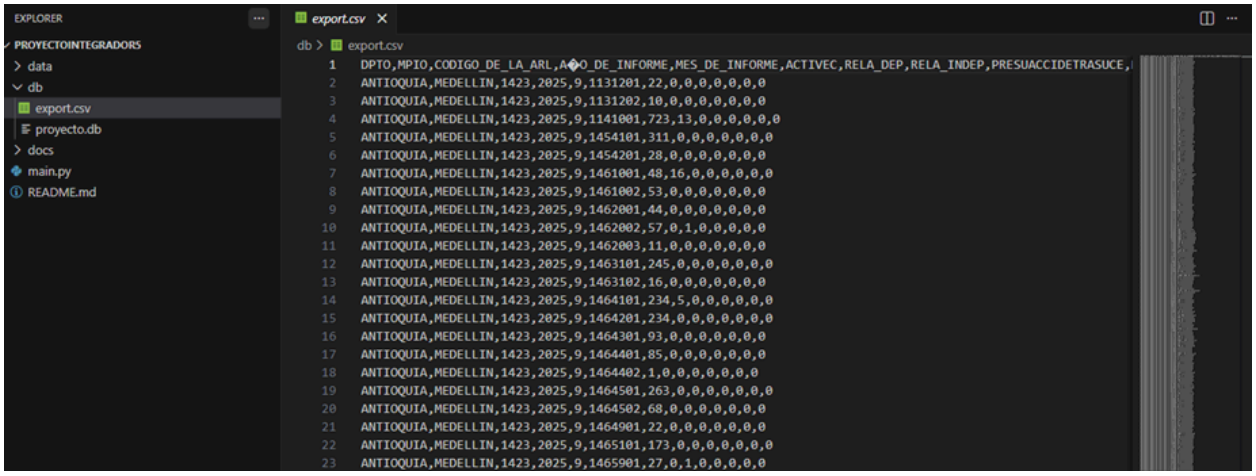
1. **Diseño del Dashboard:** Se seleccionó una herramienta de visualización (como Streamlit o Power BI) para construir un Dashboard interactivo. Este dashboard se diseñó para consolidar la información de las 5 variables clave, ofreciendo KPIs e implementando filtros dinámicos que permiten la segmentación del riesgo por tiempo y categoría.
2. **Documentación y Reporte:** Se formalizó la metodología, incluyendo la creación del Diagrama de Gantt detallado para trazar la planificación del proyecto. Finalmente, se actualizó el Informe Final bajo normas APA, incorporando la descripción del dashboard, la metodología completa y los principales hallazgos extraídos de las visualizaciones.
3. **Centralización:** Todo el código, los datasets enriquecidos y la documentación se organizaron y versionaron en el repositorio GitHub, cumpliendo con la estructura de entrega final.

Resultados

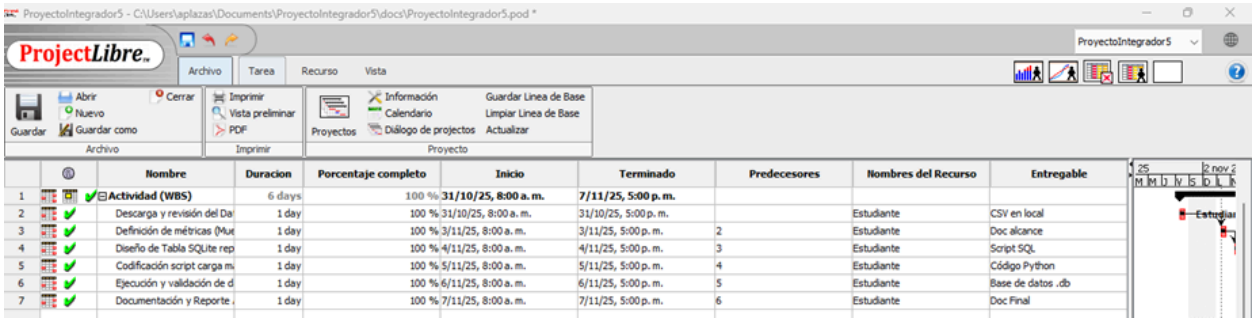
1. **Actividad EA1:**

Función del Campo	Nombre de la Columna (ID Interno)	Descripción	Tipo de Dato (SQLite)
Identificación	id	Clave Primaria Auto-incremental	INTEGER (PK)
Identificación	codigo_de_la_arl	Código de la administradora de riesgos laborales	TEXT
Temporal	a_o_de_informe	Año del informe de la data publicada	TEXT
Temporal	mes_de_informe	Mes del informe de la información	TEXT
Geografía	dpto	Departamento de Colombia	TEXT
Geografía	mpio	Municipio de Colombia	TEXT
Segmentación	activec	Actividad económica del afiliado	TEXT
Población	rela_dep	Relación trabajadores dependientes	INTEGER
Población	rela_indep	Relaciones laborales trabajadores independientes	INTEGER
Métrica (Frecuencia)	presuaccidetrasuce	Presunto accidente de trabajo sucedido	INTEGER
Métrica (Severidad)	muertes_repor_at	Muertes de trabajadores reportadas por accidente de trabajo	INTEGER
Métrica (Consecuencia)	nuevapensioinva_r_at	Nueva pensión Accidentes de trabajo	INTEGER
Métrica (Consecuencia)	nuevapensioinva_r_el	Nueva pensión Enfermedad laboral	INTEGER
Métrica (Consecuencia)	incapermaparciar_at	Incapacidad permanente parcial reportada por accidente de trabajo	INTEGER
Métrica (Consecuencia)	incapermaparciar_el	Incapacidad permanente parcial por enfermedad laboral	INTEGER

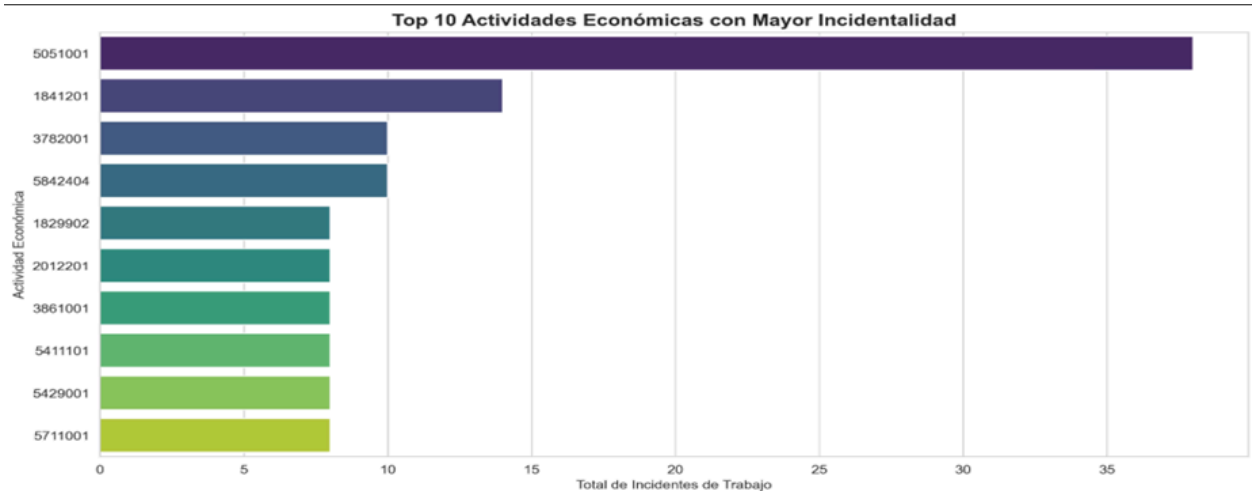


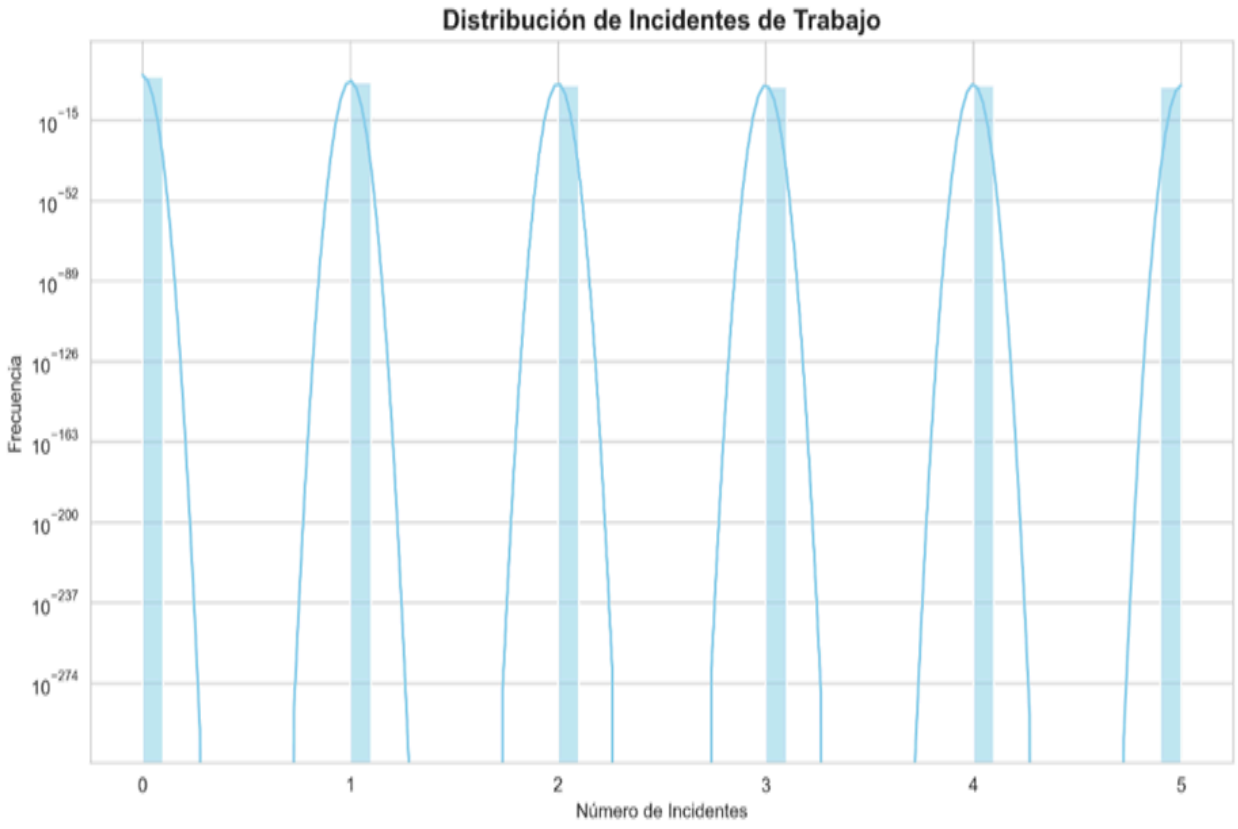


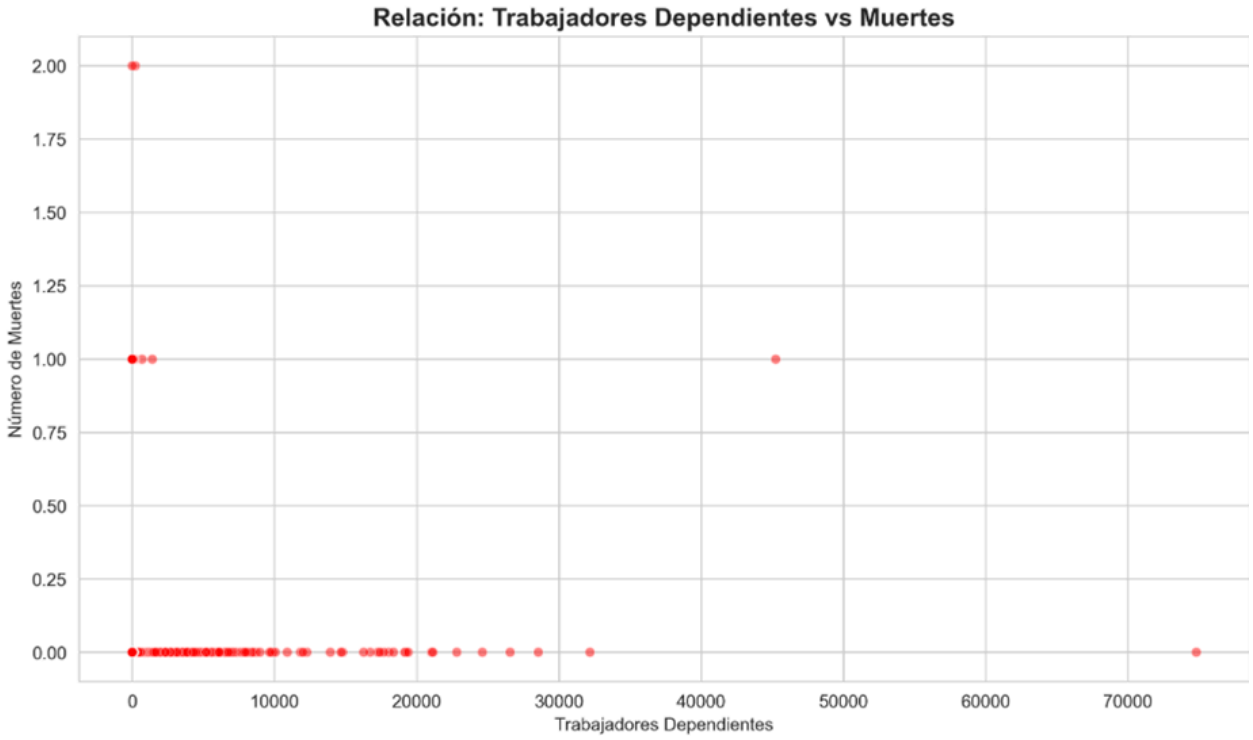
Nombre	Duración	Porcentaje Completo	Inicio	Fin	Predecesores	Nombre responsable	Entregable
Actividad (WBS)	**6 days**	**1.0**	**31/10/25, 8:00 a. m.**	**7/11/25, 5:00 p. m.**			
Descarga y revisión del Dataset ARL	1 day	1.0	31/10/25, 8:00 a. m.	31/10/25, 5:00 p. m.		Estudiante	CSV en local
Definición de métricas (Muertes vs Actividad)	1 day	1.0	3/11/25, 8:00 a. m.	3/11/25, 5:00 p. m.	2	Estudiante	Doc alcance
Diseño de Tabla SQLite reporte_arl	1 day	1.0	4/11/25, 8:00 a. m.	4/11/25, 5:00 p. m.	3	Estudiante	Script SQL
Codificación script carga main.py	1 day	1.0	5/11/25, 8:00 a. m.	5/11/25, 5:00 p. m.	4	Estudiante	Código Python
Ejecución y validación de datos	1 day	1.0	6/11/25, 8:00 a. m.	6/11/25, 5:00 p. m.	5	Estudiante	Base de datos .db
Documentación y Reporte APA	1 day	1.0	7/11/25, 8:00 a. m.	7/11/25, 5:00 p. m.	6	Estudiante	Doc Final

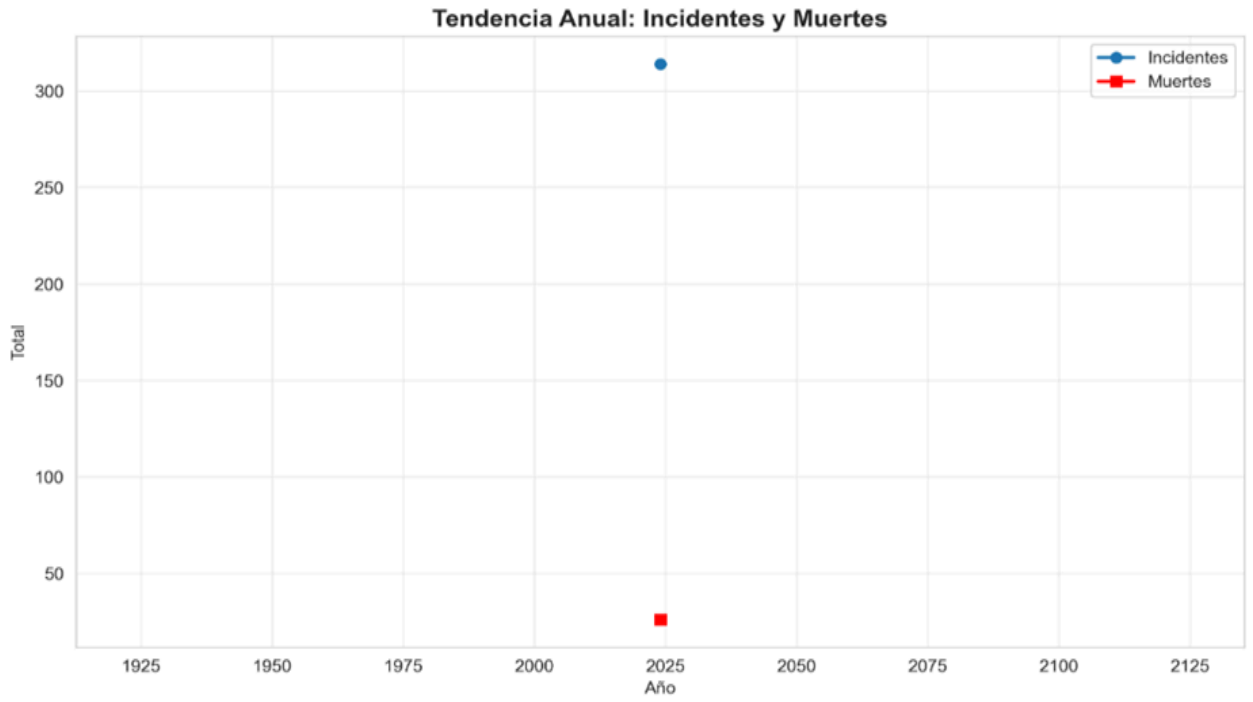


2. Actividad EA2:





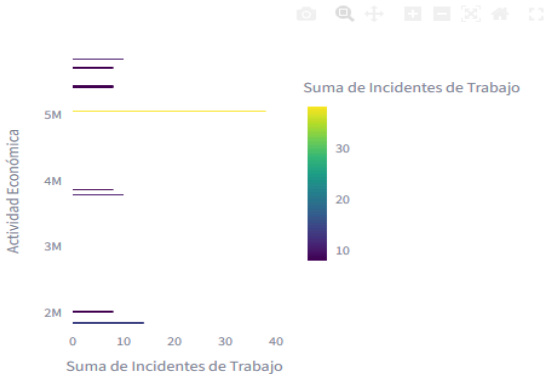

Relación: Trabajadores Dependientes vs Muertes
Número de Muertes
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
10000
20000
30000
40000
50000
60000
70000
Trabajadores Dependientes


Tendencia Anual: Incidentes y Muertes
Incidentes
Muertes
Total
300
250
200
150
100
50
1925
1950
1975
2000
2025
2050
2075
2100
2125
Año

```
================================================================================
ESTADÍSTICAS DESCRIPTIVAS
...
pen_el      88620.0    0.000090     0.009501  0.0  0.0  0.0  0.0      1.0
presuntos   88620.0    0.221981     2.470250  0.0  0.0  0.0  0.0    207.0
dep         88620.0   32.221745   481.687859  0.0  0.0  0.0  6.0  74807.0
indep       88620.0    5.495588   260.841585  0.0  0.0  0.0  0.0  57745.0
```

3. Actividad EA3:

Top 10 Actividades con Mayor Incidencia (VR3 vs. VR1)

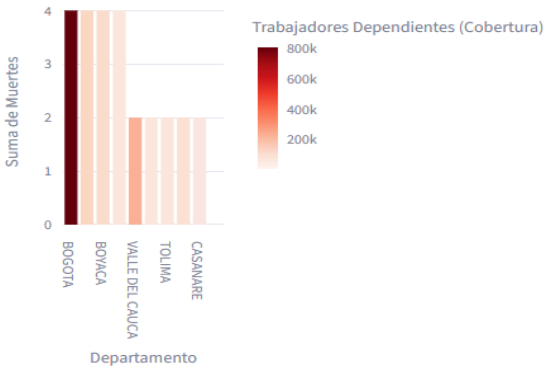


Interpretación: Muestra los sectores que concentran la mayor accidentalidad, indicando dónde priorizar la prevención.

Evolución Anual: Incidencia (VR3) y Muertes (VR4)

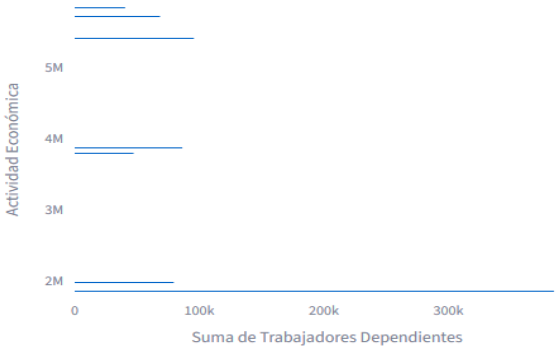
Interpretación: Permite ver la tendencia de riesgo a lo largo de los años, identificando picos o reducciones.

Cobertura (VR5) y Fatalidad (VR4) por Departamento (VR2)



Interpretación: Muestra dónde se concentra la fatalidad (Muertes) y lo relaciona con el nivel de cobertura (Dep) en esa región.

Distribución de la Cobertura (VR5) por Actividad (VR1)



Interpretación: Es crucial para entender el tamaño de cada sector. Ayuda a normalizar el riesgo (riesgo por cada 1,000 empleados).

Conclusiones

1. Robustez Metodológica y Calidad de Datos

Conclusión: La implementación exitosa de un *pipeline* de datos en dos etapas (ETL inicial en main.py y limpieza avanzada en el Notebook) garantiza la integridad y la calidad analítica del dataset. Se logró superar las complejidades inherentes a la fuente de datos (errores de codificación, inconsistencias de formato y caracteres especiales), estandarizando las 5 variables clave. Este proceso válida la solidez de la metodología adoptada, estableciendo una fuente de verdad única (dataset_enriquecido.csv) inmune a los problemas del origen.

2. Identificación Inmediata de Riesgos Críticos (Hallazgos Analíticos)

Conclusión: El Análisis Exploratorio de Datos (EDA) y las visualizaciones del Dashboard revelaron patrones de riesgo cruciales, transformando los datos en *insights* accionables.

- Concentración de Riesgo: Se identificó con precisión el Top 10 de Actividades Económicas (activec) que concentran la mayor accidentalidad (inc_at) y fatalidad (muertes), proporcionando una clara priorización geográfica y sectorial para las campañas de prevención.
- Monitoreo Temporal: Los gráficos de tendencia confirmaron la evolución de la accidentalidad a lo largo de los períodos, permitiendo a los gestores evaluar rápidamente si las estrategias de riesgo implementadas están siendo efectivas o si se requiere una intervención urgente en años o meses específicos.

3. Entrega de una Solución de Inteligencia de Negocios Funcional

Conclusión: El proyecto culmina con la entrega de un Dashboard interactivo que transforma un informe estático en una herramienta de toma de decisiones dinámica. El diseño del dashboard conecta visualmente las métricas de cobertura (dep) con la fatalidad (muertes), permitiendo a los usuarios segmentar el riesgo a nivel de Departamento (dpto) y Actividad Económica. Esta herramienta es esencial para optimizar la asignación de recursos, dirigiendo la inversión y la capacitación hacia los sectores y regiones que evidencian el mayor riesgo real.

4. Capacidad de Documentación y Escalabilidad

Conclusión: La documentación exhaustiva, incluyendo el Diagrama de Gantt y la actualización del Reporte Final bajo normas APA, confirma la capacidad del equipo para ejecutar un proyecto complejo de ingeniería de datos de principio a fin. El uso de tecnologías abiertas como Python, Pandas y Streamlit asegura la escalabilidad y la sostenibilidad del *pipeline*, facilitando futuras integraciones con nuevas fuentes de datos o la migración a entornos de *cloud computing*.

Entregables

Video: https://somup.com/cTlQc8RyrB

Cronograma:
https://docs.google.com/spreadsheets/d/1wJsnENecy6-x-kSkdqnDR3mXOxFeba0_/edit?usp=sharing&ouid=114456551648034617516&rtpof=true&sd=true

Reposición Github: https://github.com/AndresPlazas19931504/proyectointegrador5

Comando panel BI: streamlit run dashboard/app.py