Universidad Católica de la Santísima Concepción
Facultad de Ingeniería
Ingeniería Civil Informática

GESTIÓN DE DATOS
PROYECTO SEMESTRAL 2025-II

"Análisis y Visualización de Tendencias Epidemiológicas Globales con Python y Pandas"

Duración: 4 semanas
Modalidad: Grupal (3 estudiantes)
Repositorio base: CSSEGISandData/COVID-19 – Johns Hopkins University

El proyecto tiene como objetivo aplicar los conocimientos adquiridos en el curso para **procesar, analizar y visualizar grandes volúmenes de datos reales**, usando Python y librerías como **Pandas, NumPy, Matplotlib y Seaborn**.

Se trabajará con los **reportes diarios de COVID-19** publicados por la **Johns Hopkins University (JHU CSSE)**, disponibles en formato CSV dentro del repositorio.

El trabajo debe demostrar un dominio progresivo en **limpieza, análisis exploratorio, visualización, optimización y desarrollo de dashboard**.

ETAPA 1: Limpieza y preparación de datos

Objetivo: Comprender la estructura del dataset y generar una base consolidada y limpia.
Rango temporal: *1 mes de datos (ejemplo: enero 2020)*.
Dificultad: Básica – operaciones individuales sobre un solo archivo CSV.

Responde los siguientes ítems **utilizando Python y Pandas**:

1. Cargar y visualizar los primeros 5 registros del archivo 01-22-2020.csv.
2. Mostrar el número total de filas y columnas del DataFrame.
3. Describir los tipos de datos (dtypes) y convertir las columnas necesarias (por ejemplo, fechas).
4. Detectar y mostrar valores nulos o faltantes por columna.
5. Eliminar columnas irrelevantes (por ejemplo, códigos FIPS o coordenadas si no se usarán).
6. Estandarizar nombres de columnas (usar formato snake_case).
7. Homogeneizar nombres de países (ej. "US" → "United States").
8. Convertir la columna Last_Update al formato YYYY-MM-DD.
9. Crear una columna active_cases = Confirmed - Deaths - Recovered.
10. Guardar el DataFrame limpio como covid_clean_enero2020.csv e indicar su tamaño en MB.

Resultado esperado: 01_limpieza_datos.ipynb con comentarios explicativos.

ETAPA 2: Análisis exploratorio y perfilado

Objetivo: Explorar la evolución de los datos y realizar análisis comparativos.
Rango temporal: *6 meses de datos (enero–junio 2020)*.
Dificultad: Media – combinar múltiples archivos y realizar cálculos agregados.

Responde las siguientes **10 preguntas guiadas**, utilizando Pandas, Matplotlib y Seaborn:

1. ¿Cuáles son los **10 países con más casos confirmados** acumulados durante el semestre?
2. ¿Qué países presentan **mayor tasa de letalidad** (Deaths / Confirmed * 100)?
3. ¿Cuántos países **no registran recuperados** en los datos analizados?
4. ¿Qué país latinoamericano presenta la **mayor cantidad de casos activos** en junio 2020?
5. ¿Cómo evolucionaron los **casos confirmados en Chile** entre enero y junio? (gráfico de líneas).
6. ¿Cuál fue la **fecha con más nuevos casos** a nivel mundial durante este período?
7. ¿Existe **correlación entre casos confirmados y fallecidos**? (gráfico de dispersión + regresión).
8. Mostrar el **Top 10 de países** con mayor crecimiento porcentual de casos entre mayo y junio.
9. Identificar países con **rebrote** (un día sin casos y luego un incremento posterior).
10. Generar un **reporte de perfilado automático** (ydata-profiling o pandas_profiling) que incluya distribuciones, correlaciones y resumen de calidad de datos.

Resultado esperado: 02_analisis_exploratorio.ipynb + perfilado.html.

ETAPA 3: Visualizaciones avanzadas

Objetivo: Profundizar en el análisis visual, comparando regiones y tendencias.
Rango temporal: *2 años de datos (2020–2021)*.
Dificultad: Media-alta – manipulación de datasets grandes y generación de gráficos agregados.

Crear al menos **5 visualizaciones avanzadas** que aborden los siguientes puntos:

1. Evolución temporal global de casos confirmados, activos y fallecidos (líneas).
2. Comparativa Top 10 países con más casos confirmados (barras).
3. Heatmap de correlaciones entre columnas relevantes (confirmados, fallecidos, activos, ratio).
4. Gráfico de barras horizontales comparando tasas de letalidad por continente.
5. Mapa o gráfico geográfico que muestre la incidencia por continente o país (opcional).

Resultado esperado: 03_visualizaciones.ipynb con análisis e interpretación de resultados.

ETAPA 4: Dashboard final

Objetivo: Integrar análisis, visualizaciones y optimización en una herramienta interactiva.
Rango temporal: *Datos globales completos (2020–2022)*.
Dificultad: Alta – integración total y automatización del flujo de análisis.

Construir un **dashboard interactivo** en **Streamlit, Dash o Panel**, que cumpla con:

- Filtros por continente, país y rango de fechas.
- Indicadores principales: casos confirmados, activos, recuperados, fallecidos.
- Visualizaciones dinámicas que se actualicen al cambiar los filtros.
- Sección de "Insight" o conclusiones automáticas.
- Indicador de rebrote y tasa de crecimiento.

Resultado esperado: carpeta /dashboard con app funcional + README con instrucciones.

ETAPA 5: Optimización y documentación (transversal para todas las etapas anteriores)

Objetivo: Mejorar la eficiencia del código y demostrar dominio técnico.

Implementar al menos **3 optimizaciones**, por ejemplo:

- Lectura eficiente (chunksize o dask).
- Conversión de tipos (astype para reducir memoria).
- Uso de índices y operaciones vectorizadas.
- Medición de tiempos de carga y comparación antes/después.

Documentar las mejoras con evidencia (tiempos y memoria).

ENTREGABLES

1. **Repositorio GitHub** del equipo, con:
   - Notebooks 01 a 04.
   - Dashboard funcional.
   - Carpeta /data con datasets limpios.
   - README.md con instrucciones de ejecución.
2. **Informe técnico (PDF)** con:
   - Introducción y objetivos.
   - Desarrollo por etapas.
   - Resultados principales y gráficos.
   - Evidencias de optimización.
   - Conclusiones y aprendizajes.
3. **Presentación oral (5–7 minutos)** mostrando el dashboard final.