Proyecto Final del Curso: Python para Ciencia de Datos

Enunciado y estructura del informe

1. Propósito del proyecto

El proyecto final busca que el equipo aplique, de forma integrada, los contenidos del curso para analizar un conjunto de datos real y comunicar hallazgos de manera clara. El enfoque es práctico: cargar datos, preparar datos, explorar, visualizar y concluir con evidencia.

2. Conformación del equipo

- Trabajo en equipos de mínimo 2 integrantes y máximo 4 integrantes.
- Todos los integrantes deben participar en el análisis y en la redacción.
- El equipo designa un responsable de entrega (solo para el envío, no para "hacer todo").

3. Tema y datos

El equipo debe seleccionar un dataset y un problema de análisis. El tema es libre, siempre que el análisis sea verificable y el dataset tenga suficiente información para explorar (recomendado: al menos 1,000 filas y por lo menos 5 variables relevantes).

El dataset puede provenir de fuentes como Kaggle, datos abiertos (gobierno/municipalidades), repositorios académicos o datos internos simulados.

Reglas para los datos:

- Debe ser legal y ético usar el dataset. Evite datos sensibles (identificadores personales, salud, información financiera personal).
- El dataset debe incluir el enlace de origen o una referencia clara (URL o repositorio).
- Si el dataset requiere licencia, el equipo debe respetarla y citarla.
- Si se usa una muestra, debe indicarse el criterio de muestreo.

4. Alcance mínimo del análisis (requisitos obligatorios)

El proyecto debe incluir, como mínimo, los siguientes componentes:

- Carga de datos en Python (Pandas).

Dr. Eric Gustavo Coronel Castillo
Docente UNI
gcoronel@uni.edu.pe

- Revisión de estructura: dimensiones, tipos de datos y diccionario básico de variables.
- Detección y tratamiento de valores faltantes (nulos) y duplicados, con justificación.
- Limpieza mínima: estandarización de nombres/formatos (por ejemplo, fechas, categorías, textos).
- Análisis exploratorio (EDA): estadísticas descriptivas y análisis por segmentos (categorías o grupos).
- Visualización: mínimo 4 gráficos pertinentes, con título, ejes y breve interpretación debajo de cada gráfico.
- Hallazgos: mínimo 3 hallazgos sustentados con números o gráficos.
- Conclusiones y recomendaciones: mínimo 2 recomendaciones coherentes con los hallazgos.

5. Requisitos técnicos

- Entorno: Jupyter Notebook o Google Colab.
- Lenguaje: Python 3.x.
- Librerías mínimas: numpy, pandas, matplotlib y seaborn. (Opcional: scipy, scikit-learn solo si el equipo lo justifica).
- El notebook debe ejecutarse de inicio a fin sin errores y generar los mismos resultados.

6. Entregables

- Informe final en formato PDF (según la estructura indicada en la Sección 7).
- Notebook (.ipynb) con el análisis completo y comentarios en Markdown.
- Archivo del dataset o enlace de descarga (si el dataset es grande, solo el enlace).
- Carpeta comprimida (.zip) con: informe, notebook y archivos auxiliares (si aplica).

7. Estructura del informe (obligatoria)

El informe debe seguir esta estructura y respetar los títulos:

Sección	Contenido mínimo	Evidencia esperada
Portada	Curso, título del proyecto, integrantes, fecha.	—
Resumen	Objetivo, dataset, método general y 2-3 resultados clave.	1 párrafo (100-150 palabras).

1. Introducción	Contexto del problema, motivación y alcance.	Problema definido en términos claros.
2. Objetivos	Objetivo general y 2-4 objetivos específicos.	Objetivos medibles.
3. Descripción del dataset	Fuente, tamaño, variables principales, diccionario básico.	Tabla breve o lista de variables.
4. Preparación de datos	Nulos, duplicados, transformaciones y justificación.	Tabla/resumen de limpieza y decisiones.
5. Análisis exploratorio (EDA)	Estadísticas, segmentación y patrones.	Tablas y explicaciones breves.
6. Visualizaciones	Al menos 4 gráficos relevantes con interpretación.	Gráficos con títulos, ejes y lectura.
7. Hallazgos	Mínimo 3 hallazgos sustentados con números o gráficos	Cada hallazgo con número o gráfico.
8. Conclusiones y recomendaciones	Conclusiones y 2 recomendaciones accionables.	Coherencia con hallazgos.
Referencias	Lista de referencias según norma APA (7.ª ed.). Incluir fuente del dataset y fuentes externas	Lista con enlaces.
Anexos (opcional)	Código adicional, tablas extensas.	Solo si aporta claridad.

8. Criterios de evaluación (referencial)

El docente evaluará el proyecto considerando, como mínimo, los siguientes aspectos:

- Cumplimiento del alcance mínimo (Sección 4).
- Calidad y coherencia del tratamiento de datos (limpieza y justificación).
- Pertinencia del EDA y de las visualizaciones (no cantidad, sino utilidad).
- Claridad del informe (redacción, orden, títulos, interpretación de resultados).
- Reproducibilidad (notebook ejecutable y resultados consistentes).
- Trabajo en equipo (participación evidenciable y consistencia del documento).

9. Reglas importantes

- No se permite copiar análisis o conclusiones de internet. Toda interpretación debe ser elaborada por el equipo a partir de su evidencia.
- Si se usa IA generativa como apoyo, debe declararse de forma breve en el informe.
- Evite ambigüedades: toda decisión (eliminar filas, imputar, transformar) debe estar justificada en el informe o en el notebook.

Dr. Eric Gustavo Coronel Castillo
Docente UNI
gcoronel@uni.edu.pe

- El informe debe aplicar la norma APA para citas en el texto y lista de referencias. Toda idea o dato externo debe estar citado.

10. Rubrica de evaluación

Criterio (peso)	Logrado (3)	En proceso (2)	En inicio (1)
1. Selección del dataset y planteamiento del problema (10)	Dataset adecuado (≥ 1,000 filas y ≥ 5 variables), problema claro y preguntas medibles.	Dataset o problema aceptable, pero con justificación parcial o preguntas poco precisas.	Dataset insuficiente o problema ambiguo; preguntas no medibles.
2. Preparación y limpieza de datos (20)	Identifica y trata nulos/duplicados; transforma formatos (fechas/categorías/textos) con justificación clara.	Realiza limpieza básica, pero justifica parcialmente o deja decisiones sin sustento.	Limpieza mínima o incorrecta; decisiones sin explicación.
3. EDA: análisis exploratorio y segmentación (15)	Estadísticas y segmentación por grupos relevantes; interpreta patrones con evidencia.	EDA presente, pero con poca segmentación o interpretaciones generales.	EDA incompleto o sin interpretación.
4. Visualizaciones (15)	≥ 4 gráficos pertinentes, con título, ejes y lectura breve bajo cada gráfico; apoyan los hallazgos.	Hay gráficos, pero faltan etiquetas/interpretación o algunos no aportan al análisis.	Gráficos insuficientes o mal construidos; no se interpretan.
5. Hallazgos y recomendaciones (15)	≥ 3 hallazgos con respaldo (números/gráficos) y ≥ 2 recomendaciones accionables coherentes.	Hallazgos o recomendaciones presentes, pero con evidencia parcial o poca coherencia.	Hallazgos sin sustento o recomendaciones vagas/no vinculadas.
6. Informe: estructura, redacción y claridad (10)	Sigue la estructura obligatoria; redacción clara y sin ambigüedades; tablas/figuras bien referidas.	Cumple la estructura con omisiones menores o redacción mejorable.	No respeta la estructura o es difícil de seguir.
7. APA (citas y referencias) (5)	Aplica APA 7: citas en texto y referencias completas (dataset y fuentes externas).	APA parcial: referencias incompletas o citas irregulares.	No aplica APA o referencias ausentes.
8. Reproducibilidad técnica (10)	Notebook ejecuta de inicio a fin sin errores; resultados consistentes; librerías indicadas.	Ejecuta con ajustes menores o falta declarar dependencias claramente.	No ejecuta o presenta errores que impiden reproducir resultados.

Dr. Eric Gustavo Coronel Castillo
Docente UNI
gcoronel@uni.edu.pe

9. Exposición final (última clase) (10)	Presentación clara (8–12 min), explica objetivo, proceso, 3 hallazgos y 2 recomendaciones; responde preguntas.	Presentación entendible, pero incompleta o con lectura excesiva; respuestas parciales.	Presentación desordenada o no evidencia dominio del trabajo.

Reglas claras para la exposición

- Duración: 8–12 minutos por equipo.
- Deben participar mínimo 2 integrantes en la exposición.
- Diapositivas: máximo 10 (título, dataset, limpieza, EDA, gráficos, hallazgos, recomendaciones, cierre).
- Deben mostrar al menos 2 gráficos clave y explicar su lectura.

Anexo A. Fuentes sugeridas de datasets

Perú

- Plataforma Nacional de Datos Abiertos (datos.gob.pe): https://www.datosabiertos.gob.pe/
- INEI – Banco de Información Distrital / estadísticas (según tema): https://www.inei.gob.pe/
- BCRP – Series estadísticas: https://estadisticas.bcrp.gob.pe/estadisticas/series/
- SBS – Estadísticas del sistema financiero (según disponibilidad): https://www.sbs.gob.pe/estadisticas

Internacionales

- Kaggle (datasets variados): https://www.kaggle.com/datasets
- UCI Machine Learning Repository: https://archive.ics.uci.edu/
- World Bank Open Data: https://data.worldbank.org/
- OECD Data: https://data.oecd.org/
- UNData: https://data.un.org/
- Google Dataset Search: https://datasetsearch.research.google.com/

Dr. Eric Gustavo Coronel Castillo
Docente UNI
gcoronel@uni.edu.pe