Apuntes

diabetes_dataset y clientesmarketing Son dos archivos .zip protegidos con contreaseña hasta el día del examen de la asignatura de Analisis de datos masivos en R.

Es muy probable que contenga excels o csvs inspirados o muy parecidos a los que hay en la plataforma kaggle. Te paso la Guia Maestra de estudio para que entiendas de qué va el curso y trates de decirme que datasets de kaggle muy posiblemente se esten usando para el examen, o que al menos sean muy parecidos según la asignatura. Buscame en kaggle o en github los datasets probables para poder estudiar.

Basado en la "Guía Maestra de Estudio" y los nombres de los archivos que has proporcionado, tengo una certeza muy alta sobre qué datasets son y para qué parte del examen se utilizarán.

Aquí tienes el análisis forense de tus archivos protegidos:

1. Archivo `diabetes_dataset.zip`

Candidato: Pima Indians Diabetes Database

- **Probabilidad:** 99%
- **Por qué:** Es el dataset estándar de oro en la enseñanza de R para **Clasificación y Regresión Logística**.
- **Relación con tu Guía:**
  - Tu guía menciona explícitamente en "Preguntas de Examen Probables": *"¿Cuándo usarías un modelo de clasificación (ej. Regresión Logística)…?"* y *"Dada la confusionMatrix() de dos modelos…"*.

- Este dataset se usa para predecir si un paciente tiene diabetes (variable `Outcome`: 0 o 1) basándose en métricas médicas.
- **Variables clave que encontrarás:** `Glucose`, `BMI` (Índice de Masa Corporal), `Insulin`, `Age`, `Pregnancies`.
- **Trampa típica del examen:** Este dataset tiene valores "0" imposibles (ej. glucosa 0 o BMI 0) que representan valores nulos. Tu guía menciona *"Tratamiento de Datos Nulos: ...reemplazar (imputar) con un valor como cero, la media o la mediana"*. Es casi seguro que te pedirán limpiar estos ceros.

2. Archivo `clientesmarketing.zip`

Candidato: Mall Customer Segmentation Data (o posiblemente "Bank Marketing")

- **Probabilidad:** 90% para Mall Customers / 10% Bank Marketing.
- **Por qué:** El nombre sugiere segmentación de clientes, lo cual encaja perfectamente con el módulo de **Clustering (K-Means)** de tu guía.
- **Relación con tu Guía:**
  - Tu guía tiene una sección entera dedicada a **K-Means**, `NbClust` y *"Interpretación de Clústeres (ej. 'clientes de alto valor')"*.
  - El dataset "Mall Customers" es el estándar para enseñar esto. Contiene datos demográficos y de comportamiento de gasto.
  - **Variables clave esperadas:** `Annual Income (k$)`, `Spending Score (1-100)`, `Age`, `Gender`.
  - **Objetivo del examen:** Te pedirán usar `NbClust` o el "Método del Codo" para decidir si hay 3, 4 o 5 tipos de clientes y luego describir cada grupo (ej. "Jóvenes con bajos ingresos pero alto gasto").

Dónde descargarlos para estudiar (Enlaces)

Para "hackear" tu estudio, descarga estos datasets de Kaggle y practica los ejercicios de tu guía (limpieza de nulos para diabetes, K-Means para marketing) sobre ellos.

Archivo Protegido	Dataset en Kaggle (Fuente Original)	Tarea del Examen (según Guía)
diabetes_dataset	**Pima Indians Diabetes Database**	Limpieza de Nulos (ceros), Regresión Logística, Matriz de Confusión.
clientesmarketing	**Mall Customer Segmentation Data**	Clustering (K-Means), NbClust, Interpretación de grupos (Spending vs Income).

Nota: Si abres el de Marketing y ves muchas variables bancarias (`duration`, `euribor`, `campaign`), entonces es el Bank Marketing Dataset, pero por el contexto de "Análisis Masivo" introductorio, el de Mall Customers es mucho más probable para un ejercicio de clustering limpio.

Recomendación final

Tu guía menciona **Series Temporales (ARIMA)** al final. Ninguno de estos dos nombres de archivo suena a series temporales (que suelen llamarse “sales.csv”, “weather.csv”). Es posible que:

1. Haya un tercer archivo que no has visto.
2. Uno de los excels dentro de los zips contenga una pestaña de histórico de ventas.
3. El examen te pida convertir una columna de fecha del dataset de marketing en una serie temporal, aunque es menos común con esos datasets específicos.

¡Mucha suerte! Si practicas con esos dos enlaces, irás con mucha ventaja.

⁂

---

1. Guia-Maestra-de-Estudio-Analisis-Masivo-de-Datos.md
2. https://www.kaggle.com/datasets/svbstan/customer-marketing-data
3. https://www.youtube.com/watch?v=HTN6rccMu1k
4. https://www.kaggle.com/code/shivashankari06/diabetes-analysis-in-r
5. https://www.kaggle.com/datasets/mathchi/diabetes-data-set
6. https://www.kaggle.com/code/sarcasmos/diabetes-prediction-with-naive-bayes-r
7. https://www.kaggle.com/code/paultimothymooney/predict-diabetes-with-r-starter-kernel
8. https://github.com/vvsbnmit/Diabetes-Prediction-using-R
9. https://github.com/aimee0317/customer-segmentation-R
10. https://www.kaggle.com/code/gatewj/bank-marketing-dataset
11. https://github.com/dmarks84/Ind_Project_Mall-Customer-Clustering–Kaggle