Evaluación N°2 – Informe de análisis con R

Objetivo

Aplicar técnicas de procesamiento y análisis de datos con R, mediante la limpieza, transformación y visualización del conjunto de datos histórico de titulados 2007–2020, con el propósito de desarrollar una comprensión analítica sobre la distribución, evolución y características de los titulados en Chile, a través de la generación de tablas resumen, gráficos estadísticos y conclusiones interpretativas.

Contexto del archivo

El archivo "titulados_historico 2007–2020.csv" contiene información oficial sobre los titulados de educación superior en Chile entre los años 2007 y 2020.
Este conjunto de datos recopila información proveniente del Ministerio de Educación, y permite analizar tendencias de titulación por año, género, institución, sede, carrera, área del conocimiento y región, entre otros criterios relevantes.

Cada registro representa una combinación específica de carrera, sede e institución, junto con la cantidad total de titulados (mujeres, hombres y totales), su rango de edad y promedios de edad de titulación.
Este tipo de información es clave para comprender el comportamiento histórico de la educación superior chilena y facilita el desarrollo de análisis descriptivos y visuales mediante herramientas de minería de datos.

Entrega:

Esta entrega es de tipo individual

Nombre de archivo: cada archivo deberá tener su nombre y apellido

- codido_nombre_apellido.r
- archivo_limpio_nombre_apellido.csv
- análisis_nombre_apellido.pdf

Debe entregar un zip con los 3 documentos, en caso de que cuyo archivo supere el máximo de tamaño permitido por la plataforma debe subir el archivo a **OneDrive(Microsoft)** y compartir link para acceder.

Fecha de entrega: miercoles 29 de octubre 23:30 hrs, vía plataforma

Ponderación: 30%

No se recibirán trabajos por correos, si el trabajo no está en la plataforma en la fecha y plazo máximo se le asignara nota mínima.

Parte I – Limpieza del archivo (90 pts)

Como base, debe realizar la limpieza del archivo original titulados_historico_2007_2020.csv.

Tareas obligatorias:

I. Renombrar columnas (15 pts):
Utilizar los nuevos nombres indicados en la tabla siguiente:

Nombre antiguo	Nuevo nombre
ANNIO	ANNIO
TOTAL TITULADOS	TOTAL_TITULADOS
TITULADOS MUJERES POR PROGRAMA	TITULADOS_MUJERES_POR_PROGRAMA
TITULADOS HOMBRES POR PROGRAMA	TITULADOS_HOMBRES_POR_PROGRAMA
CLASIFICACION INSTITUCION NIVEL 1	CLASIFICACION_NIVEL_1
CLASIFICACION INSTITUCION NIVEL 2	CLASIFICACION_NIVEL_2
CLASIFICACION INSTITUCION NIVEL 3	CLASIFICACION_NIVEL_3
CODIGO INSTITUCION	CODIGO_INSTITUCION
NOMBRE INSTITUCION	NOMBRE_INSTITUCION
COMUNA	COMUNA
PROVINCIA	PROVINCIA
REGION	REGION
NOMBRE SEDE	NOMBRE_SEDE
NOMBRE CARRERA	NOMBRE_CARRERA
AREA DEL CONOCIMIENTO	AREA_CONOCIMIENTO
CINE-F_97 AREA	AREA
CINE-F_97 SUBAREA	SUBAREA
AREA CARRERA GENERICA	AREA_CARRERA_GENERICA
NIVEL GLOBAL	NIVEL_GLOBAL
CARRERA CLASIFICACION NIVEL 1	CARRERA_CLASIFICACION_NIVEL_1
CARRERA CLASIFICACION NIVEL 2	CARRERA_CLASIFICACION_NIVEL_2
MODALIDAD	MODALIDAD
JORNADA	JORNADA
TIPO DE PLAN DE LA CARRERA	TIPO_PLAN_CARRERA
DURACION ESTUDIO CARRERA	DURACION_ESTUDIO_CARRERA
DURACION TOTAL DE LA CARRERA	DURACION_TOTAL_CARRERA
CODIGO CARRERA	CODIGO_CARRERA
TOTAL RANGO EDAD	TOTAL_RANGO_EDAD
RANGO DE EDAD 15 A 19 AÑOS	RANGO_15_A_19_ANNIOS
RANGO DE EDAD 20 A 24 AÑOS	RANGO_20_A_24_ANNIOS
RANGO DE EDAD 25 A 29 AÑOS	RANGO_25_A_29_ANNIOS
RANGO DE EDAD 30 A 34 AÑOS	RANGO_30_A_34_ANNIOS
RANGO DE EDAD 35 A 39 AÑOS	RANGO_35_A_39_ANNIOS

RANGO DE EDAD 40 Y MAS AÑOS	RANGO_40_Y_MAS_ANNIOS
RANGO DE EDAD SIN INFORMACION	RANGO_SIN_INFORMACION
PROMEDIO EDAD CARRERA	PROMEDIO_EDAD_CARRERA
PROMEDIO EDAD MUJER	PROMEDIO_EDAD_MUJER
PROMEDIO EDAD HOMBRE ,,,,,,,,,	PROMEDIO_EDAD_HOMBRE

II. Reemplazar valores vacíos, Null y NaN por 0 (15 pts)
Asegúrese de convertir correctamente las celdas vacías en valores numéricos válidos.

III. Corregir registros en PROMEDIO_EDAD_HOMBRE (15 pts)
Elimine las comas o símbolos sobrantes al final de los registros.

IV. Modificar columna ANNIO (15 pts)
Elimine el prefijo “TIT_” para dejar únicamente el año numérico.

V. Eliminar columnas innecesarias (15 pts)
- CODIGO_INSTITUCION
- CODIGO_CARRERA
- RANGO_SIN_INFORMACION

VI. Guardar DataFrame limpio (15 pts)
Guarde el resultado como:
archivo_limpio_nombre_apellido.csv

Parte II – Desarrollo de análisis y visualización (100 pts)

Mediante código en R, deberá responder a las siguientes 10 preguntas, cada una con:

- Una tabla resumen (4 pts)
- Un gráfico asociado (4 pts)
- Un breve análisis personal que interprete los resultados (2–3 líneas/ 2pts)

N°	Pregunta / Requerimiento	Tipo de gráfico sugerido
1	Cantidad total de titulados por año	Gráfico de líneas
2	Top 10 regiones con más titulados (acumulado)	Barras horizontales
3	Evolución anual de titulados para Top 5 regiones	Líneas
4	Distribución de titulados por género	Torta o barras apiladas
5	Comparación mujeres vs hombres por año	Líneas
6	Top 15 instituciones con más titulados	Barras horizontales
7	Top 15 sedes con más titulados	Barras horizontales
8	Titulados por área del conocimiento	Barras horizontales
9	Top 20 carreras con más titulados	Barras horizontales
10	Clasificación de instituciones por niveles 1, 2 y 3	Barras agrupadas

Nota: Cada gráfico debe incluir un título claro, ejes etiquetados y fuente:

Fuente: Elaboración propia, datos de titulados 2007–2020.

Parte III – Informe de análisis

Esta parte se evalúa en conjunto con la anterior

Deberá **documentar y explicar su proceso analítico** utilizando el enfoque **KDD (Knowledge Discovery in Databases)**, integrando los resultados obtenidos en el código.

Estructura mínima del documento PDF:

1. **Portada:**
    - Nombre del estudiante
    - Asignatura y evaluación
    - Fecha
    - Título: *"Análisis KDD – Titulados Educación Superior 2007–2020"*
2. **Introducción:**
   Breve descripción del propósito del análisis y del conjunto de datos utilizado.
3. **Desarrollo del proceso KDD:**
   Describa brevemente las etapas aplicadas:
    - **Selección de datos:** Identificación del archivo y sus variables principales.
    - **Limpieza de datos:** Transformaciones realizadas y correcciones aplicadas.
    - **Transformación:** Creación de nuevas variables, agrupaciones o conversiones.
    - **Minería de datos:** Resumen de los cálculos o consultas generadas (las 10 preguntas).
    - **Interpretación y evaluación:** Principales hallazgos y observaciones relevantes.
4. **Evidencias gráficas:**
   Incluya los **gráficos generados en el código R** con su respectivo número de pregunta, título, breve descripción e interpretación personal.
   Ejemplo:

Figura 1: Titulados por año (2007–2020).
Se observa una tendencia creciente hasta 2018, con una leve baja en 2020.

5. **Conclusiones:**
    - Principales descubrimientos y patrones observados.
    - Reflexión sobre la utilidad del análisis de datos en la educación superior.
    - Posibles líneas de análisis futuro (ejemplo: por provincia o edad promedio).