Tarea Unidad 2 Taller de Software para Data Science

Tema: Aplicación de Pandas y Seaborn para el Análisis y Visualización de Datos

Introducción:

En esta tarea, tendrás la oportunidad de aplicar los conceptos y habilidades de manipulación y análisis de datos utilizando la biblioteca Pandas de Python que has aprendido en el curso. A través de la resolución de un enunciado práctico, podrás demostrar tu capacidad para procesar, limpiar y analizar datos de manera efectiva.

La tarea consiste en tres enunciados diferentes, cada uno de ellos presentando un contexto y objetivos específicos. Los enunciados abarcan el análisis de datos de ventas y clientes de una tienda, el análisis de datos de películas y ratings de usuarios, y el análisis de datos de interacciones en redes sociales. Cada enunciado requerirá que generes archivos CSV con al menos 100 datos aleatorios, realices una limpieza de datos, manipules DataFrames, apliques técnicos de análisis utilizando Pandas y crees visualizaciones básicas con Seaborn.

Para fomentar el trabajo en equipo y la colaboración, deberás formar grupos de 2 o 3 personas para realizar esta tarea. **Es importante que te pongas en contacto con el profesor**, indicando los nombres de los integrantes de tu grupo. Una vez que hayas informado sobre tu grupo, se te asignará aleatoriamente **uno de los tres enunciados** para que trabajes en él junto con tus compañeros. Por tanto, tu grupo debe trabajar **en solo uno** de los proyectos.

Recuerda que, independientemente del enunciado que te asigne, tendrás la oportunidad de aplicar los conceptos clave del análisis de datos con Pandas y demostrar tu comprensión de los temas abordados en el curso. Aprovecha esta oportunidad para fortalecer tus habilidades, colaborar con tus compañeros y obtener experiencia práctica en el manejo y análisis de datos.

Notas:

1) **Asignación de enunciados**: Por favor, contacta al profesor indicando los nombres de los integrantes de tu grupo. Si bien puedes sugerir al profesor el proyecto que te gustaría resolver, el profesor podría aceptarlo o bien te puede asignar uno escogido aleatoriamente entre los tres enunciados disponibles. Esto para que el número de grupos trabajando en cada proyecto sea parejo. Asegúrate de comunicarte con el profesor lo antes posible para recibir tu enunciado y comenzar a trabajar en la tarea.
2) En informe incluya cualquier supuesto extra (con su debida justificación) que haya realizado o dificultad presentada. Por ejemplo, si desea cambiar el delimitador entre etiquetas a valores, explique porque decidió hacerlo.
3) En la tapa del informe debe claramente indicar los nombres (máximo 3) de los alumnos que participaron en el desarrollo de la tarea.
4) El informe debe contener al menos 3 secciones, introducción donde se presenta el proyecto, desarrollo, donde explica el software desarrollado, como se debe ejecutar, incluir capturas de pantallas y una sección de conclusiones donde discuta las dificultades y los aspectos que aprendió del proyecto. Como se distribuyo el trabajo el grupo.

5) Debe crear un archivo zip que incluya el informe de la tarea en formato pdf y los dos programas solicitados. El nombre de este archivo zip DEBE ser: Tarea_2_Taller_Software_DS_Apellido_Nombre.zip (Donde Nombre y Apellido es el nombre y apellido del alumno que envía la tarea)
6) El envío del archivo zip debe ser por email a patricio.galdames@uss.cl. El asunto de su correo DEBE ser Tarea 2 Taller de Software Data Science
7) Tiempo para realizar tarea: 3 Semanas

Rúbrica de evaluación:

1. Generación correcta de los archivos CSV con datos aleatorios (15 puntos)
    a. Creación de los dos archivos CSV requeridos (5 puntos)
    b. Inclusión de las columnas especificadas en cada archivo (5 puntos)
    c. Generación de datos aleatorios utilizando funciones de probabilidad (5 puntos)
2. Carga correcta de los archivos CSV y almacenamiento en DataFrames (10 puntos)
    a. Carga de los datos desde los archivos CSV utilizando Pandas (5 puntos)
    b. Almacenamiento de los datos en DataFrames separados (5 puntos)
3. Limpieza de datos y manejo de valores nulos (15 puntos)
    a. Identificación y manejo adecuado de los valores nulos (5 puntos)
    b. Conversión de los tipos de datos según sea necesario (5 puntos)
    c. Documentación de los pasos realizados para la limpieza de datos (5 puntos)
4. Cruce correcto de los DataFrames utilizando la clave primaria (10 puntos)
    a. Identificación de la columna clave primaria (5 puntos)
    b. Realización del cruce de los DataFrames utilizando la función adecuada (5 puntos)
5. Cálculos y análisis utilizando Pandas (20 puntos)
    a. Aplicación de las funciones de agregación y transformación de Pandas (10 puntos)
    b. Generación de tablas y resúmenes estadísticos (5 puntos)
    c. Utilización correcta de indexación fancy y slicing en los DataFrames (5 puntos)
6. Creación de visualizaciones básicas con Seaborn (10 puntos)
    a. Selección adecuada del tipo de gráfico según los datos (5 puntos)
    b. Configuración y personalización de la visualización (5 puntos)
7. Guardado de los resultados en un archivo Excel (5 puntos)
    a. Exportación de los resultados generados a un archivo Excel (5 puntos)
8. Claridad, estructura y documentación del código (15 puntos)
    a. Organización y legibilidad del código (5 puntos)
    b. Uso de comentarios y docstrings para explicar el propósito del código (5 puntos)
    c. Documentación adecuada de las funciones y clases utilizadas (5 puntos)
9. Calidad del informe y presentación de los resultados (20 puntos)
    a. Estructura y organización del informe (5 puntos)
    b. Explicación clara de los pasos realizados y las decisiones tomadas (5 puntos)
    c. Presentación de los resultados y visualizaciones (5 puntos)
    d. Conclusiones y recomendaciones basadas en los análisis realizados (5 puntos)
10. Envío del archivo zip con el formato solicitado (5 puntos)
    a. Creación de un archivo zip que incluya el informe en formato PDF y los archivos de código (3 puntos)

   b. Nombre del archivo zip siguiendo el formato especificado: Tarea_2_Apellido_Nombre.zip (2 puntos)
11. Envío del correo electrónico con el asunto solicitado (5 puntos)
   a. Envío del archivo zip por correo electrónico a la dirección especificada (3 puntos)
   b. Asunto del correo electrónico siguiendo el formato: Tarea 2 - Apellido Nombre (2 puntos)
12. Bonus: Implementación de técnicas o análisis adicionales (5 puntos extra)
   e. Aplicación de técnicas o análisis adicionales más allá de los requerimientos básicos (5 puntos extra)

Total: 135 puntos (140 puntos con el bonus)

Enunciado 1: Análisis de datos de ventas y clientes de una tienda

Contexto:

Una tienda de electrónica desea realizar un análisis de sus datos de ventas y clientes para obtener información relevante y tomar decisiones estratégicas. Los datos de ventas incluyen información sobre el ID de venta, el ID de cliente, el nombre del producto, la categoría, el precio unitario y la cantidad vendida. Los datos de clientes incluyen información sobre el ID de cliente, el nombre, la edad, el género y la ubicación.

Objetivos:

Aplicar los conceptos de manipulación y análisis de datos utilizando la biblioteca Pandas de Python para procesar los datos de ventas y clientes de la tienda, cruzarlos y generar informes que respondan a las preguntas planteadas.

Requerimientos:

1. Generar dos archivos CSV llamados "ventas.csv" y "clientes.csv" que contengan los datos de ventas y clientes, respectivamente. Los datos deben ser generados aleatoriamente utilizando funciones de probabilidad para simular escenarios realistas (indique en su informe que distribución uso: uniforme, normal u otra). Ejemplo de las primeras líneas de los archivos CSV:

   - *"ventas.csv":*

     ID_Venta,ID_Cliente,Producto,Categoría,Precio,Cantidad
     1,101,Televisor,Electrónica,500.0,2
     2,102,Lavadora,Hogar,800.0,1
     3,101,NULL,Electrónica,1200.0,1

   - *"clientes.csv":*

     ID_Cliente,Nombre,Edad,Género,Ubicación
     101,Juan Pérez,NULL,Masculino,Norte
     102,María González,35,Femenino,Sur
     103,NULL,28,Masculino,Este

   Los valores nulos deben representarse con la palabra "NULL". En el archivo "ventas.csv", la columna "Producto" puede contener hasta un 10% de valores nulos. En el archivo "clientes.csv", las columnas "Nombre" y "Edad" pueden contener hasta un 5% de valores nulos cada una.

2. Cargar los datos de ventas y clientes desde los archivos CSV generados utilizando Pandas y almacenarlos en DataFrames separados.
3. Realizar una limpieza de datos inicial en ambos DataFrames, manejando los valores nulos y convirtiendo los tipos de datos según sea necesario. Documentar en el informe los pasos realizados para la limpieza de datos.

4. Cruzar los DataFrames de ventas y clientes utilizando la columna "ID_Cliente" como clave primaria.
5. Generar un informe que muestre el total de ventas por categoría de producto y género del cliente utilizando una Serie de Pandas.
6. Calcular el promedio de edad de los clientes que compraron cada categoría de producto utilizando las funciones de agregación de Pandas.
7. Identificar los 5 productos más vendidos y los 5 productos menos vendidos utilizando indexación fancy y slicing en el DataFrame cruzado.
8. Crear una visualización básica con Seaborn que muestre la distribución de edades de los clientes por género.
9. Guardar los resultados generados en un archivo Excel.

Enunciado 2: Análisis de datos de películas y ratings de usuarios

Contexto:

Una plataforma de streaming de películas desea analizar los datos de las películas y los ratings de los usuarios para comprender mejor las preferencias de los usuarios y realizar recomendaciones personalizadas. Los datos de las películas incluyen información sobre el ID de la película, el título, el género, el año de lanzamiento y la duración. Los datos de los ratings incluyen información sobre el ID de usuario, el ID de la película y el rating otorgado.

Objetivos:

Utilizar la biblioteca Pandas de Python para analizar los datos de las películas y los ratings de los usuarios, cruzarlos y generar informes estadísticos que permitan a la plataforma de streaming comprender mejor las preferencias de los usuarios y tomar decisiones informadas.

Requerimientos:

1. Generar dos archivos CSV llamados "peliculas.csv" y "ratings.csv" que contengan los datos de las películas y los ratings de los usuarios, respectivamente. Los datos deben ser generados aleatoriamente utilizando funciones de probabilidad para simular escenarios realistas (indique en su informe que distribución uso: uniforme, normal u otra).. Ejemplo de las primeras líneas de los archivos CSV:

   - "peliculas.csv":

   ID_Pelicula,Titulo,Genero,Año,Duracion
   1,The Shawshank Redemption,Drama,1994,142
   2,The Godfather,NULL,1972,175
   3,The Dark Knight,Acción,2008,152

   - "ratings.csv":

   ID_Usuario,ID_Pelicula,Rating
   101,1,4.5
   102,1,4.0
   101,2,NULL
   103,3,4.8

   Los valores nulos deben representarse con la palabra "NULL". En el archivo "peliculas.csv", la columna "Genero" puede contener hasta un 8% de valores nulos. En el archivo "ratings.csv", la columna "Rating" puede contener hasta un 5% de valores nulos.

2. Cargar los datos de las películas y los ratings desde los archivos CSV generados utilizando Pandas y almacenarlos en DataFrames separados.

3. Realizar una limpieza de datos inicial en ambos DataFrames, manejando los valores nulos y convirtiendo los tipos de datos según sea necesario. Documentar en el informe los pasos realizados para la limpieza de datos.
4. Cruzar los DataFrames de películas y ratings utilizando la columna "ID_Pelicula" como clave primaria.
5. Calcular el rating promedio de cada película utilizando una Serie de Pandas.
6. Generar una tabla que muestre la cantidad de películas por género y año de lanzamiento utilizando las funciones de agregación de Pandas.
    - Identificar las 10 películas con el rating promedio más alto y las 10 películas con el rating promedio más bajo utilizando indexación fancy y slicing en el DataFrame cruzado.
7. Crear una visualización básica con Seaborn que muestre la distribución de los ratings por género de película.
8. Guardar los resultados generados en un archivo Excel.

Enunciado 3: Análisis de datos de interacciones en redes sociales

Contexto:

Una empresa de marketing desea analizar los datos de interacción de los usuarios en las redes sociales para comprender mejor el engagement de sus publicaciones. Los datos incluyen información sobre las publicaciones realizadas, el número de likes, comentarios y compartidos, así como detalles demográficos de los usuarios que interactuaron con las publicaciones.

Objetivos:

Utilizar la biblioteca Pandas de Python para analizar los datos de interacciones en redes sociales, cruzar la información de publicaciones y usuarios, y generar informes estadísticos que permitan a la empresa comprender mejor el engagement de sus publicaciones y las características demográficas de los usuarios interactivos.

Requerimientos:

1. Generar dos archivos CSV llamados "publicaciones.csv" y "usuarios.csv" que contengan los datos de las publicaciones y los usuarios que interactuaron con ellas, respectivamente. Los datos deben ser generados aleatoriamente utilizando funciones de probabilidad para simular escenarios realistas. Ejemplo de las primeras líneas de los archivos CSV:

   - "publicaciones.csv":

   ID_Publicacion,Fecha,Contenido,Likes,Comentarios,Compartidos
   1,2023-05-01,Nuevo producto disponible,100,20,10
   2,2023-05-02,Oferta especial,150,30,NULL
   3,2023-05-03,Evento de lanzamiento,200,50,25

   - "usuarios.csv":
   ID_Usuario,Edad,Genero,Ubicacion,ID_Publicacion
   1,25,Femenino,Valdivia,1
   2,30,Masculino,Concepcion,1
   3,35,NULL,Santiago,2
   4,28,Femenino,Puerto Montt,3

   Los valores nulos deben representarse con la palabra "NULL". En el archivo "publicaciones.csv", la columna "Compartidos" puede contener hasta un 8% de valores nulos. En el archivo "usuarios.csv", la columna "Genero" puede contener hasta un 5% de valores nulos.

2. Cargar los datos de las publicaciones y los usuarios desde los archivos CSV generados utilizando Pandas y almacenarlos en DataFrames separados.

3. Realizar una limpieza de datos inicial en ambos DataFrames, manejando los valores nulos y convirtiendo los tipos de datos según sea necesario. Documentar en el informe los pasos realizados para la limpieza de datos.
4. Cruzar los DataFrames de publicaciones y usuarios utilizando la columna "ID_Publicacion" como clave primaria.
5. Calcular el promedio de likes, comentarios y compartidos por publicación utilizando una Serie de Pandas.
6. Generar una tabla que muestre el engagement promedio (likes + comentarios + compartidos) por género y rango de edad (18-25, 26-35, 36-45, 46+) utilizando las funciones de agregación de Pandas.
7. Identificar las 5 publicaciones con mayor engagement y las 5 publicaciones con menor engagement utilizando indexación fancy y slicing en el DataFrame cruzado.
8. Crear una visualización básica con Seaborn que muestre la distribución de engagement por ubicación geográfica de los usuarios.
9. Guardar los resultados generados en un archivo Excel.