Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

IIC2115 – Programación como Herramienta para la Ingeniería (II/2025)

Ejercicio Formativo 2 Capítulo 1

Aspectos generales

- **Objetivos:** aplicar los contenidos de análisis de datos tabulares.
- **Entrega:** lunes 25 de agosto a las 17:30 hrs. en repositorio privado y ticket de salida.
- **Formato de entrega:** archivo `E2.ipynb` con los solicitado, ubicado en la carpeta **C2** del repositorio.
- **ULTRA IMPORTANTE**: todas las celdas utilizadas deben estar ejecutadas al momento de entregar el ejercicio, de modo que las salidas generadas sean visibles. En caso de no cumplir esto, su entrega no será considerada como validación del ticket de salida.

Descripción del problema (versión normal)

En este taller trabajará con el archivo `E2_data.csv`, que contiene series históricas de O_3 y PM2.5 junto con una etiqueta de riesgo ambiental (bajo, medio o alto). Su tarea inicial será explorar y preparar los datos para distintos análisis. A partir de este trabajo, se espera que el sistema que desarrolle le permita responder preguntas como: ¿qué diferencias aparecen según el método de imputación?, ¿cómo se correlacionan O_3 y PM2.5?, o ¿qué patrones temporales emergen al observar los datos a nivel mensual o anual?
Posteriormente, deberá profundizar en el análisis incorporando nuevas perspectivas: proponer un esquema propio para asignar niveles de riesgo a registros sin etiqueta, identificar patrones estacionales a partir de la fecha, detectar valores atípicos con un criterio estadístico y examinar su distribución temporal, y enriquecer el conjunto de datos con nuevas variables que aporten información adicional, como por ejemplo razones entre contaminantes, secuencias de días extremos, entre otros. La clave de esta actividad es que cada paso contribuya a una visión más completa y crítica sobre la calidad del aire, de modo que usted pueda justificar cómo las transformaciones aplicadas afectan la comprensión final de los datos.

Descripción del problema (versión guiada)

Considere el conjunto de datos almacenado en el archivo `E2_data.csv`, que contiene datos obtenidos a lo largo de los años sobre los niveles de ozono (O_3) y material particulado de 2.5 micrómetros (PM2.5). Además de esta información, cada registro está categorizado en tres niveles, en base al riesgo ambiental que presentan las mediciones de O_3 y PM2.5 para la fecha: bajo, medio y alto. En base a toda esta información, complete las misiones indicadas a continuación.

Misión 0: aspectos básicos

Para cumplir las misiones de este taller, es fundamental explorar inicialmente el contenido del archivo y familiarizarse con el formato en que está almacenada la información. Para eso, utilice los comandos `describe` y `head` de `pandas`.

Misión 1: imputación de información faltante

Tanto para O_3 como para PM2.5, el conjunto contiene datos incompletos para algunos días, que fueron generados por motivos desconocidos. Con el fin de facilitar el análisis futuro, deberá ajustar los datos faltantes de 2 formas distintas. Para esto último, cree 2 nuevos `DataFrame`, en el primero de ellos complete los datos faltantes con la media, y en el segundo elimínelos.

Misión 2: descripción y comparación

A continuación, para ambos `DataFrame` generados en el ítem anterior y de manera independiente, imprima en una tabla ordenada los siguientes indicadores para O_3 y PM2.5: media, desviación estándar, máximo, mínimo, curtosis. Además, agregue a esta tabla la correlación entre O_3 y PM2.5.

Misión 3: visualización

Para ambos `DataFrame` obtenidos en el primer ítem y de manera independiente, genere las siguientes visualizaciones:

- Histograma de PM2.5
- Boxplot de O_3 por mes
- Evolución promedio de O_3 y PM2.5 por año.

Misión 4: categorización

En base a todos los análisis realizados anteriormente, proponga e implemente en Python un esquema para asignar un nivel de riesgo medioambiental para cada registro que no tiene esta información. Complete esto para ambos `DataFrame` de manera independiente. Comente y analice los resultados.

Misión 5: análisis temporal avanzado

Identifique patrones estacionales en los datos, utilizando el archivo `data_E2.csv` para crear una nueva columna que indique la estación del año (primavera, verano, otoño, invierno) para cada registro basado en la fecha. Calcule la media de O_3 y PM2.5 para cada estación y cada año y finalmente genere una visualización que muestre la evolución de las medias estacionales de O_3 y PM2.5 a lo largo del tiempo.

Misión 6: detección de anomalías

Identifique registros atípicos que puedan indicar problemas en la recolección de datos o eventos excepcionales. Para esto implemente un método basado en el rango intercuartílico (IQR) para identificar valores atípicos en las columnas de O_3 y PM2.5. Cree un nuevo `DataFrame` que contenga únicamente los registros que fueron considerados atípicos y analiza si hay alguna relación temporal o estacional en los datos atípicos detectados.

Misión 7: transformación y enriquecimiento de datos

Realice manipulaciones avanzadas de datos para crear nuevas variables y enriquecer el conjunto de datos, en base a los siguientes pasos:

- Cree una nueva columna que calcule la razón (*ratio*) entre los niveles de O_3 y PM2.5 para cada registro.
- Implemente un algoritmo para detectar y marcar días consecutivos con niveles extremos de O_3 o PM2.5 (por encima del percentil 95).
- Genere una nueva columna que clasifique cada registro según el día de la semana y determine si es fin de semana o día laboral.
- Cree una tabla dinámica (*pivot table*) que muestre la media de O_3 y PM2.5 por mes y por nivel de riesgo ambiental.
- Utilice `groupby` y `rolling` para calcular la media móvil de 7 días para O_3 y PM2.5, y añada estas columnas al conjunto de datos.
- Analice y comenta cómo estas transformaciones enriquecen la comprensión del conjunto de datos.