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

IIC2115 – Programación como Herramienta para la Ingeniería (I/2024)

Ejercicio Formativo 1 Capítulo 2

Aspectos generales

- **Objetivos:** familiarizarse los elementos básicos del análisis de datos tabulares con Pandas.
- **Entrega:** lunes 01 de abril a las 17:30 hrs. en repositorio privado y ticket de salida.
- **Formato de entrega:** archivo `E1.ipynb` con los solicitado, ubicado en la carpeta **C2** del repositorio.

Descripción del problema

El conjunto de datos Anscombe's quartet, disponible en el archivo `data_E1.csv`, consiste en 4 conjuntos de una serie de pares (x, y), con algunas características particulares. En este ejercicio explorará algunas de ellas mediante el análisis estadístico y visual.
Siga la instrucciones de la lista de a continuación, recordando siempre hacer *commit* regularmente en el repositorio privado, usando comentarios descriptivos:

1. Crea una carpeta de nombre **C2** dentro del repositorio privado y agrega un nuevo **Jupyter Notebook** con nombre `E1.ipynb`.
2. Crea una celda de texto dentro del documento e ingresa el título de la actividad "Ejercicio Formativo 1 Capítulo 2" junto con tu nombre.
3. Cargue el conjunto de datos utilizando Pandas e imprima su contenido. Comente acerca del formato utilizado para los datos.
4. Calcule algunas métricas interesantes para obtener mayor conocimiento de estos datos. Es importante notar que se esperan los valores de las métricas por conjunto, es decir, debe obtener las métricas para el conjunto A, B, C y D de manera separada. Considere inicialmente las siguientes:

- Media.
- Desviación estándar.
- Máximo.
- Mínimo.
- Correlación entre las variables x e y.
- Kurtosis.

5. Comente los resultados, resaltando los aspecto más llamativos.

6. Identifique aquellas métricas que son iguales para los conjuntos. Luego, visualice los conjuntos para caracterizar de mejor manera sus diferencias. Para esto, genere cuatro gráficos de dispersión (*scatter plots*, uno por conjunto), donde se evidencie la relación (x, y) de los pares que corresponden a cada uno.

7. Repita los pasos anteriores, esta vez descartando aleatoriamente el 10% de los datos de cada conjunto, e imputando sus valores utilizando la media de cada coordenada. Comente sobre las diferencias en los resultados.