Homework 1

The goal is to get a good insight into a dataset by mean of summary statistics and visualizations. For this exercise set choose one alternative below. Regardless of your choice, your submission must comply with the guidelines at the end of this document.

Data selection

Para a seleção do conjunto de dados será considerado a alternativa 7.

Alternative 7 - Your choice: You have a set of data of your own interest. The dataset should comprise of a certain number of observations, each observation consists of a certain number of predictors and corresponding class label

Os dados escolhidos para a realização da atividade podem ser encontrados no seguinte link: <https://data.seattle.gov/browse?q=eui&sortBy=newest&utf8=✓>

Data analysis

Documentation Guidelines

A atividade detalhada no documento requer que os participantes realizem uma análise exploratória de dados abrangente, dividida em várias etapas. Aqui está um resumo completo do que a atividade exige:

### **Seleção de Dados**

Os participantes devem escolher um dos seguintes conjuntos de dados:

1. **Abalone** - Características de abalones (UC Irvine ou biblioteca R).
2. **Gapminder** - Expectativa de vida, PIB per capita e população por país (site do Gapminder ou biblioteca R).
3. **Wine Quality** - Dados de vinhos tintos e brancos de Portugal (UC Irvine).
4. **Gas Sensor Array Drift** - Medições de sensores químicos (UC Irvine).
5. **Dados ambientais em Melbourne** - Consumo de energia, clima e características de águas residuais (Mendeley).
6. **Concrete Mixture** - Dados sobre resistência de misturas de concreto (UC Irvine).
7. **Escolha própria** - Qualquer conjunto de dados que atenda aos critérios do exercício.

### **Análise de Dados**

Independentemente do conjunto escolhido, as etapas obrigatórias incluem:

1. **Descrição dos Dados**
   * Quantidade de observações (NNN), variáveis preditoras (DDD), classes (LLL) e distribuição das classes.
2. **Análise Monovariada Incondicional**
   * Plotar histogramas e box-plots.
   * Calcular média (μ\muμ), desvio padrão (σ\sigmaσ) e assimetria (γ\gammaγ) para todas as DDD variáveis.
3. **Análise Monovariada Condicional às Classes**
   * Plotar histogramas e box-plots condicionais às classes.
   * Calcular médias, desvios padrão e assimetrias condicionais (μd∣l\mu\_d|lμd​∣l, σd∣l\sigma\_d|lσd​∣l, γd∣l\gamma\_d|lγd​∣l).
   * Tabular resultados e comentar padrões significativos.
4. **Análise Bivariada Incondicional**
   * Plotar scatter plots entre pares de preditores, indicando as classes por cores ou símbolos.
   * Avaliar correlação linear entre pares de variáveis (ρ\rhoρ).
   * Apresentar uma matriz de correlação e comentar os resultados.
5. **Análise Multivariada Incondicional**
   * Realizar análise de componentes principais (PCA).
   * Plotar os dados projetados nos dois primeiros componentes principais, indicando as classes.
   * Avaliar a separação entre as classes e a complexidade das fronteiras.

### **Entrega**

Os participantes devem produzir:

1. **Artigo Científico** (máximo de 6 páginas, seguindo template IEEE):
   * **Título**: Resuma o trabalho em uma frase.
   * **Resumo**: Introduza os objetivos e os resultados principais.
   * **Introdução**: Contextualize os dados e métodos.
   * **Métodos**: Descreva o conjunto de dados e as técnicas usadas.
   * **Resultados**: Discuta os principais achados.
   * **Referências**: Liste as fontes bibliográficas usadas.
2. **Código Fonte**
   * Deve ser funcional e apresentado no apêndice do artigo ou em um arquivo zipado.

### **Prazos e Penalidades**

* Entrega até **30 de novembro de 2024**.
* Atrasos são penalizados:
  + <24h: 20%.
  + <48h: 40%.
  + Etc.

Essa descrição se refere à alternativa 7 de uma proposta de atividade ou projeto, na qual você pode utilizar um conjunto de dados que seja de seu próprio interesse para realizar uma análise ou tarefa específica. Aqui está a explicação detalhada dos termos mencionados:

1. **"You have a set of data of your own interest"**:  
   Isso significa que você deve selecionar ou fornecer um conjunto de dados que considere interessante ou relevante para a análise.
2. **"The dataset should comprise a certain number of observations"**:  
   O conjunto de dados precisa conter **observações**, que são registros ou amostras. Por exemplo, se você estiver analisando dados de clientes, cada linha pode ser uma observação representando um cliente.
3. **"Each observation consists of a certain number of predictors"**:  
   Cada observação no conjunto de dados deve ter **preditores** (também chamados de variáveis independentes ou características). Esses preditores são as informações ou atributos usados para prever ou classificar algo. Por exemplo, em um conjunto de dados sobre casas, os preditores poderiam ser o tamanho da casa, o número de quartos e a localização.
4. **"And corresponding class label"**:  
   Além dos preditores, cada observação deve ter um **rótulo de classe** (ou variável dependente), que é a saída ou categoria associada àquela observação. Por exemplo, em um problema de classificação de e-mails, os rótulos de classe poderiam ser "spam" ou "não spam".

### Exemplo Prático:

Um conjunto de dados para prever o tipo de fruta com base em atributos poderia ser assim:

* **Preditores**: Peso, Cor, Tamanho.
* **Classe**: Tipo de fruta (maçã, laranja, banana).

Se você tiver um dataset de interesse, você poderia utilizá-lo para treinar e testar modelos de machine learning ou realizar análises exploratórias com base nesses conceitos.

4o