Questionário

1. **Consideramos valores *outliers* valores que:**
   a) Estão associados a erros na recolha de informação.
   b) Estão desviados da média mais do 1.5 IQR.
   c) Devem ser eliminados, pois impactam nas métricas utilizadas nos modelos (ex.: média).
   d) São atípicos dada a distribuição da variável à qual pertence.

2. **Verdadeiro ou Falso: "Os *outliers* são um problema a resolver antes de iniciar a modelação, seja via eliminação seja via substituição de valores".**
   a) Verdadeiro.
   b) Falso.

3. **As transformações não-lineares, como logaritmo e raiz quadrada,…**
   a) Alteram a distribuição das variáveis e podem ter impacto (positivo ou negativo) na sua capacidade explicativa.
   b) São as melhores a resolver os problemas de outliers.
   c) Devem ser usadas com cautela por reduzem a sua correlação com a target.
   d) São ótimas normalizar as variáveis.

4. **Por vezes pode ser relevante criar novas variáveis adicionais, como por exemplo variáveis Dummy. Que tipo de variáveis são essas e para que servem?**
   a) São variáveis que indicam se o *dataframe* tem *outliers*.
   b) São variáveis que indicam se o dataframe tem *missing values*.
   c) São variáveis que tomam o valor *True* ou *False* e pode ser usadas para tornar quantitativas as variáveis qualitativas.
   d) São variáveis que tomam o valor *True* ou *False* e pode ser usadas para tornar qualitativas as variáveis quantitativas.

5. **Através de um gráfico de *Scatter Plot* que tipo de informação consigo recolher?**
   a) A distribuição da variável, se tem *missing values* e *outliers*.
   b) A correlação entre 2 ou mais variáveis.
   c) Identificar potencial correlação linear entre 2 variáveis.
   d) Todas as anteriores.

6. **Imagina que te é disponibilizado um *dataframe* com um conjunto de dados sobre dados biomédicos de pacientes que estão a ser alvo de um tratamento experimental e que tem como missão perceber qual a dosagem mais adequada para a melhor recuperação dos pacientes.**

1º- Garantir que o conteúdo dataframe está alinhado com o estudo em questão visto que a falta de alinhamento pode inviabilizar o estudo

a) Existe um ID de paciente para comprar o antes e depois do tratamento
b) Existe informação sobre a dosagem atribuída a cada utente, assim como a data da administração
c) Existem variáveis que permitem avaliar a recuperação do utente após a administração da dosagem
d) Dimensão do dataframe é suficiente para tirar conclusões (correlacionado com a diversidade de valores das variáveis, das dosagens que queremos testar, etc.)
e) …

2º- Garantir formatos e consistência de informação para que o estudo considere os valores das variáveis de forma correta e assim aumente a sua eficácia

a) Validar formatos das variáveis e corrigir, caso necessário (ex.: formatos das datas)
b) Normalizar variáveis para garantir consistência (ex.: abreviaturas)
c) …

3º- Identificação e Tratamento de *missing values* por forma a garantir que o dataframe está preparado para ser usado pelas diferentes técnicas que pretendemos utilizar, sem excluir ou incorretamente preencher valores em falta

a) Identificar observações / variáveis com valores em falta ou valores anómalos (ex.: "?" em variáveis numéricas)
b) Determinar dimensão dos casos de valores em falta
c) Compreender motivo para a ausência de valores (tipologia)
d) Tratar casos de *missing values* com respeito à sua tipologia e dimensão

4º- Identificação e Tratamento, caso aplique, de *outliers* por forma minimizar os impactos que possam ter nas métricas usadas pelas técnicas e modelos

a) Análise do comportamento das variáveis via visualização da sua distribuições (ex.: histogramas) e cálculo de métricas de centralidade (ex.: média, mediana, moda) e dispersão (ex.: desvio padrão, IQR)
b) Identificação de *outliers* e potenciais causas (ex.: erros, observações atípicas, etc.)
c) Dependendo das causas para a sua existência, decidir se devem ser mantidos, tratados ou eliminados
d) Caso devam ser tratados, avaliar as várias opções disponíveis como truncar os valores, realizar transformações não lineares, etc.

5º- Ajuste de escalas das variáveis por forma a garantir que discrepâncias entre os domínios das variáveis (ex.: variável "# filhos" e a variável "rendimento anual") não influência os resultados do modelo dando prioridade a variáveis com domínios com valores absolutos mais elevados

a) Analisar as escalas das diferentes variáveis
b) Aplicar transformações lineares (ex.: Min-Max) ou não lineares (ex.: Log)

6º- Transformação e criação de novas variáveis com o objetivo de dar maior informação ao modelo, potenciando assim a sua capacidade explicativa

a) Verificar a existência de variáveis categóricas e transforma-las em Dummies caso eu tenha interesse em usar modelos que não as tomem diretamente
b) Transformar datas em "dias desde o evento x" caso tal possa ser vantajoso
c) Criar novas variáveis Dummies associadas a *missing values* preenchidos, outliers truncados ou outro tipo de manipulação de dados
d) Etc.

7º- Exploração gráfica da relação entre as variáveis, sobretudo entre a variável de interesse e as restantes variáveis de análise por forma a compreender como se relacionam

a) Scatter Plots sobretudo para correlações lineares com a target
b) Headmap para correlações entre todas as variáveis
c) Gráfico com valores médios das variáveis quantitativas para cada valor presente nas variáveis qualitativas
d) Etc.