Questionário

1. **Qual a diferença entre *missing values* e valores vazios?**
   a) *Missing values* são valores omissos, valores em falta. Valores vazios são valores que não existem no contexto da variável.
   b) São a mesma coisa, valores em falta nos dados.
   c) Os *missing values* são valores que aparecem nos *dataframes* como NaN e os valores vazios aparecem como "?".
   d) *Missing values* são valores que temos de recuperar a todo o custo, valores vazios podemos substituir por 0.

2. **O que significa que um *missing value* ser do tipo MNAR: *Missing Not At Random*?**
   a) São valores em falta devido a erros nos dados.
   b) São valores em falta que têm uma distribuição Normal, em vez de um distribuição aleatória.
   c) São valores em falta facilmente explicados pelas restante informação disponível.
   d) São os *missing values* mais críticos visto que o motivo da sua ausência está relacionada com a variável de interesse do nosso estudo.

3. **Verdadeiro ou Falso: "Sempre que a percentagem de *missing values* é inferior a 0,2% podemos eliminar as observações."**
   a) Verdeiro.
   b) Falso.

4. **Quando devemos recuperar a informação em falta?**
   a) Sempre, visto que é a solução com menor impacto para a nossa análise.
   b) Só em casos extremos, quando a percentagem de *missing* é muito reduzida, por saí muito caro.
   c) Sempre que os valores em causa sejam críticos para o estudo.
   d) Caso haja dinheiro e tempo, é a única solução correta.

5. **Qual o principal objetivo da normalização de escalas?**
   a) Garantir que todas as variáveis têm o mesmo peso no estudo.
   b) Tornar os modelos mais eficientes de executar.
   c) Por todas as variáveis entre [0:1].
   d) Garantir que todas as variáveis têm uma distribuição Normal.

6. Num estudo sobre o aproveitamento escolar dos alunos de um agrupamento de escolas, temos de *missing values* na variável "nível de escolaridade do encarregado de educação".

1º- Compreender o motivo do problema

a) Questionar os responsáveis pela recolha de informação junto dos pais e perguntar se eles sabem qual o motivo da ausência da informação

b) Analisar correlação entre aproveitamento escolar e o nível de escolaridade dos EE no caso em que não há *missing values*

c) Hipóteses:

- Segundo o apurado em a) a ausência deve-se a um erro de inserção de dados → tipologia **MCAR**
- Segundo o apurado em a) a ausência não se deve a erros + segundo b) existe uma relação entre o aproveitamento escolar e a escolaridade, mas não determinístico, ou seja, existe alunos com bom aproveitamento escolar e em que o nível de escolaridade dos EE é baixa + verifica-se uma correlação entre o nível escolar dos EE e outras variáveis disponíveis como profissão, rendimento, localidade, etc. → tipologia **MAR**

Nota: Estamos sempre a assumir que não é possível recuperar os dados, seja por temas de tempo ou custo.

2º- Compreender a dimensão do problema

a) Em ambos casos, **MCAR** e **MAR,** se:
   I. a % de *missing values* é muito baixa (ex.: inferior a 1%) então podemos **descartar as observações**
   II. a % de *missing values* é significativa (ex.: superior a 20%) e assumindo que a variável é relevante para o nosso estudo, então o melhor é **dividir o dataframe em 2 dataframe**, um onde a informação da variável está disponível e outro em que a informação está indisponível.

b) Se a % de *missing values* é baixa (ex.: inferior a 20%) mas não insignificante (ex.: superior a 1%) então, se a tipologia dos *missing values* for…
   i. **MCAR** podemos recorrer à **imputação direta**, recorrendo por exemplo à **moda** ou um **valor fixo como "Sem informação"**
   ii. **MAR** então devemos procurar **estimar**, em função das outras variáveis sobre as quais temos informação, qual o valor em falta

Nota: A classificação da % de *missing values* como muito baixa, baixa, etc… vai depender a dimensão total dos nossos dados e da criticidade dada à representatividade que os dados face à população.