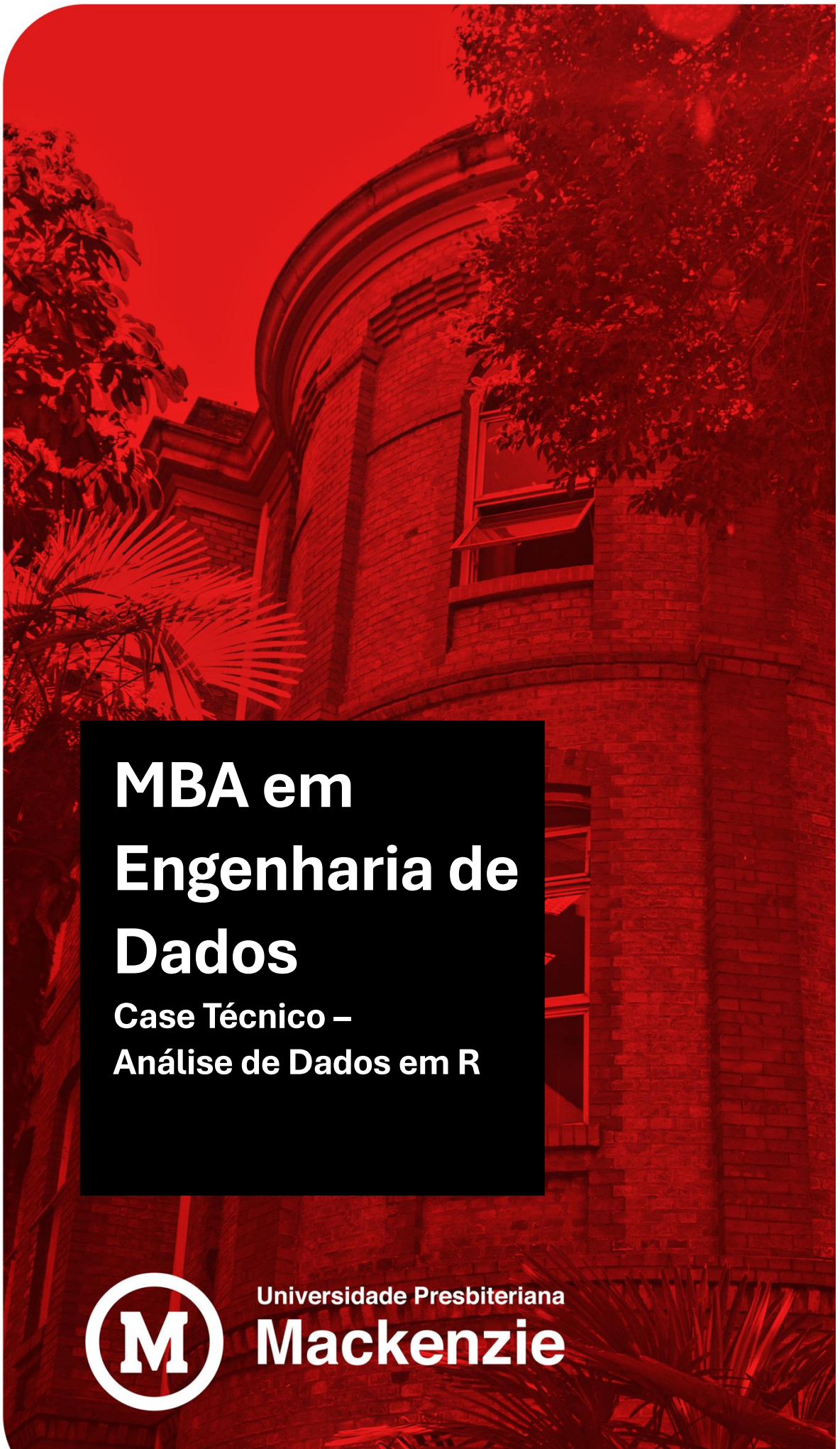

MBA em Engenharia de Dados

Case Técnico – Análise de Dados em R

Universidade Presbiteriana Mackenzie

Case Técnico – Análise de Dados em R

Objetivo

O objetivo deste case é aplicar conhecimentos de programação em R, manipulação de dados, implementação de pipelines e visualização para realizar uma análise completa de um dataset à sua escolha, porém **é necessário que tenham variáveis numéricas e categóricas**. Todas as questões abaixo devem ser resolvidas utilizando o mesmo dataset ao longo do trabalho.

1. Importe o seu dataset para o R.
2. **Contextualize o problema de negócio relacionado ao seu dataset**
3. **Contextualize a solução que seu pipeline deverá resolver (Questão aberta)**
4. Verifique as **primeiras 6 linhas** do dataset.
5. Verifique as **últimas 10 linhas** do dataset.
6. Mostre a quantidade de **linhas e colunas** do dataset.
7. Exiba apenas os nomes das colunas do dataset.
8. **Descreva em poucas palavras as principais variáveis do seu dataset que farão parte dos principais pipelines que irão existir nas perguntas seguintes.**
9. Ao explorar o seu dataset, você percebe que uma coluna que deveria ser categórica está como numérica, ou que uma coluna de datas está como caractere. Verifique o tipo de todas as colunas do dataset e ajuste para o tipo correto
10. Selecione apenas **duas colunas** do dataset.
11. Filtre as linhas onde uma variável numérica seja maior que um valor definido.
12. Ordene o dataset de forma crescente com base em uma coluna numérica.
13. Crie uma nova coluna com base em uma operação entre duas colunas existentes.
14. Remova uma coluna do dataset.
15. Use a função select() para escolher 3 colunas do dataset.

16. Use a função filter() para selecionar linhas que atendam a uma condição.
17. Selecione todas as colunas cujo nome começa com uma letra específica usando select(starts_with()).
18. Renomeie duas colunas do dataset usando rename().
19. Utilize arrange() para ordenar os dados de forma decrescente.
20. Crie uma nova coluna com mutate().
21. Resuma os dados de uma coluna numérica usando summarise().
22. Agrupe os dados por uma variável categórica com group_by().
23. Combine group_by() e summarise() para calcular a média de uma variável por grupo.
24. Use pivot_longer() para transformar colunas em linhas.
25. Utilize um pipeline para: selecionar colunas, filtrar linhas e ordenar os dados.
26. Use pivot_wider() para transformar linhas em colunas.
27. Aplique drop_na() para remover valores ausentes.
28. Substitua valores ausentes por 0 em uma coluna numérica.
29. Crie um gráfico de **dispersão** (scatterplot) com duas variáveis numéricas.
30. Crie um gráfico de **barras** de uma variável categórica.
31. Construa um **histograma** de uma variável numérica.
32. Crie um gráfico de **linha** para visualizar a evolução de uma variável ao longo do tempo.
33. Adicione uma **linha de tendência** a um gráfico de dispersão.
34. Crie um **boxplot** para comparar a distribuição de uma variável numérica entre categorias.
35. Personalize um gráfico com **título, legenda e rótulos nos eixos**.
36. Crie um **mapa de calor (heatmap)** com duas variáveis categóricas.
37. Combine mais de um gráfico em uma mesma visualização usando facet_wrap().
38. Crie uma função chamada resumo_variavel() que receba:
    - um dataframe,
    - o nome de uma coluna numérica,
    - e um parâmetro opcional plot = TRUE.

A função deve:

- ✓ Retornar um resumo estatístico da coluna (mínimo, máximo, média, mediana e desvio padrão).
- ✓ Se plot = TRUE, exibir também um histograma da coluna usando ggplot2.

39. Usando o operador pipe (%>%), faça as seguintes operações no seu dataset:
    - Selecione **três colunas**: duas numéricas e uma categórica.
    - Filtre apenas as linhas em que **não existam valores ausentes (NA)** nessas colunas.
    - Crie uma nova coluna que seja a **razão entre as duas variáveis numéricas**.
    - Agrupe os dados pela variável categórica.
    - Calcule a **média, a mediana e o desvio padrão** da nova coluna criada, para cada grupo.
    - Reorganize os resultados em formato **largo (wide)**, de forma que cada estatística (média, mediana, desvio) vire uma coluna separada.
    - Ordene o resultado pela média em ordem decrescente.

40. Construa um pipeline seguindo as instruções abaixo:
    - Selecione todas as colunas numéricas do dataset.
    - Substitua **valores ausentes por 0**.
    - Crie uma nova coluna categórica com base em uma condição aplicada a uma variável numérica (ex.: “Alto” se > média, “Baixo” se ≤ média).
    - Agrupe pelos valores da nova coluna categórica.
    - Calcule **média, mediana e máximo** de todas as variáveis numéricas agrupadas.
    - Ordene os grupos pela média de uma coluna escolhida.

41. Com o pipeline da questão 38, faça:
    a. Salve este pipeline como **uma função em um arquivo R separado** (ex.: meu_pipeline.R).
    b. Carregue a função do arquivo
    c. Passe o **dataset** como argumento para a função e gere **um dataset final processado**.