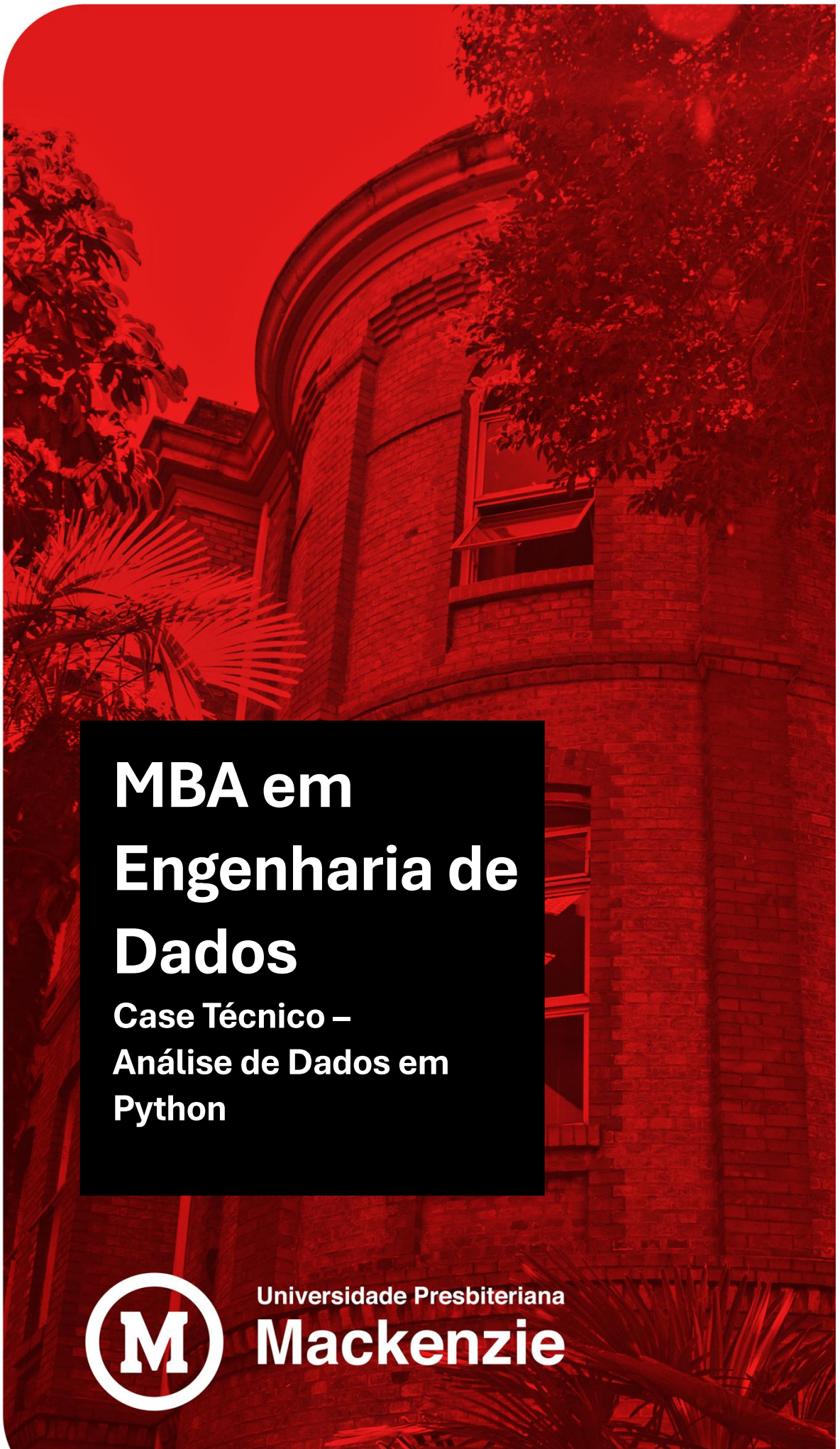

MBA em Engenharia de Dados

Case Técnico – Análise de Dados em Python

Universidade Presbiteriana Mackenzie

Case Técnico – Análise de Dados em Python

Objetivo

O objetivo deste case é aplicar conhecimentos de programação em Python, como importação e exportação de arquivos, manipulação de dados e visualização para realizar uma análise exploratória de dados completa de um dataset à sua escolha, porém **é necessário que tenham variáveis numéricas e categóricas**. Todas as 38 questões abaixo devem ser resolvidas utilizando o mesmo dataset ao longo do trabalho.

IMPORTANTE: Modelo Exemplo do entregável em notebook:
01_projeto_preparação_Analise_Exploratoria_de_Dados

Parte 1 – Leitura de Dados:

1. **Importe o dataset** escolhido diretamente de uma URL ou arquivo local.
2. **Contextualize o problema de negócio relacionado ao seu dataset**
3. **Construa uma análise exploratória de dados completa para o seu problema de negócio (Questão aberta)**
4. **Exiba as 4 primeiras linhas** do conjunto de dados.
5. **Exiba as 3 últimas linhas** do conjunto de dados.
6. **Descreva em poucas palavras as principais variáveis do seu dataset que farão parte das perguntas seguintes.**
7. Verifique e mostre:
    - O **formato** do dataset (shape).
    - Os **tipos de dados** de cada coluna.
    - A existência de **valores ausente** e **duplicações.**

Parte 2 – Estruturas de Dados e Operações em Python

1. Crie **duas listas** em Python utilizando seu dataframe sendo: uma com variáveis **numéricas** e outra com as **categóricas.**
2. Crie um **dicionário** com o nome das colunas como chave e o tipo da variável como valor ('numérica' ou 'categórica').
3. Crie uma **tupla** com os nomes de todas as colunas do seu dataset.
4. Crie uma **tupla** que receba números (uma variável do seu dataset) e retorne:

   a. A soma dos elementos

   b. O maior e o menor valor
5. Crie um conjunto (**set()**) com todos os **valores únicos** de uma variável categórica do seu dataset.
6. Usando seu DataFrame:

   a. Selecione uma coluna usando indexação.
   b. Selecione as primeiras 5 linhas com slicing.

Parte 3 – NumPy

1. Extraia as colunas numéricas do df em um array NumPy.
2. Converta um array NumPy de volta para um DataFrame Pandas, mantendo os nomes das colunas originais.
3. Extraia uma matriz NumPy com as colunas numéricas do seu df.
4. Aplique um reshape() em um subconjunto de dados do seu df.
5. Calcule a **média, mediana** e o **desvio padrão** de cada coluna numérica usando funções do NumPy.

Parte 4 – Pandas

1. Selecione linhas do seu df seguindo alguma condição.
2. Agrupe os dados por uma coluna categórica e calcule a **média de pelo menos 2 variáveis numéricas**.
3. Faça um merge() com um novo DataFrame contendo dados agregados por categoria.
4. Faça um **resumo estatístico** das variáveis numéricas com .describe()
5. Faça um **resumo** das variáveis categóricas com crosstab()

Parte 5 – Visualização com Matplotlib e Seaborn

1. Crie um **gráfico de linha**.
2. Plote um **gráfico de barras** mostrando a contagem de categorias em uma variável.
3. Plote um **histograma** de uma variável numérica.
4. Crie um **boxplot** para comparar a distribuição de uma variável numérica por uma categórica.
5. Plote um **mapa de calor (heatmap)** da matriz de correlação das variáveis numéricas.
6. Use sns.pairplot() para visualizar relações entre variáveis numéricas.

Parte 6 – Exportando DataFrames em CSV

1. Salve o df em um arquivo CSV chamado "dados_trat.csv" no diretório atual.
2. Salve o DataFrame sem incluir o índice no arquivo CSV.
3. Salve apenas as colunas numéricas do DataFrame em um arquivo CSV separado chamado "subset_numericas.csv".