Antonio,

Nessa disciplina, aprofundamos nossos conhecimentos na linguagem R e de estatística, para realizar análises descritivas de bases de dados, tarefa que é extremamente importante para o dia-a-dia de um cientista de dados. Agora iremos validar nosso conhecimento.

1. O relatório final deve ser apresentado utilizando RMarkdown. Nesse relatório devem haver:
   1. imagens estáticas ("prints" de tela, imagens da internet - com a devida fonte mencionada - ou figuras criadas pelo aluno fora do ambiente do R);
   2. imagens geradas através do ambiente R, particularmente com a biblioteca ggplot;
   3. links clicáveis (como fontes e referências).
2. Escolha uma base de dados para realizar esse projeto. Essa base de dados será utilizada durante toda sua análise. Essa base necessita ter 4 (ou mais) variáveis de interesse, onde todas são numéricas (confira com o professor a possibilidade de utilização de dados categóricos). Observe que é importante que haja dados faltantes em pelo menos uma variável para executar esse projeto. Caso você tenha dificuldade para escolher uma base, o professor da disciplina irá designar para você. Explique qual o motivo para a escolha dessa base e aponte os resultados esperados através da análise.
3. Utilizando o pacote summarytools (função descr), descreva estatisticamente a sua base de dados.
4. Crie um gráfico com a matriz de espalhamento (*scatter matrix plot*) para sua base de dados. Através de investigação visual, quais são as variáveis mais correlacionadas. Apresente o gráfico e justifique.
5. Sobre a normalidade das variáveis:
   1. Descreva o que é uma distribuição normal;
   2. Crie um histograma para cada variável da sua base de dados. Justifique a escolha do número de bins para seu trabalho. (usando o pacote ggplot);
   3. Crie um gráfico Q-Q para cada variável de sua base de dados. (use as funções presentes no pacote ggpubr);
   4. Execute um teste de normalidade Shapiro-Wilk;
   5. Baseado nos itens anteriores, é possível afirmar que algumas das variáveis se aproximam de uma distribuição normal? Justifique.
6. Qualidade de dados tem sido um dos temas mais abordados nos projetos de estruturação em data analytics, sendo um dos principais indicadores do nível de maturidade das organizações. Um dos problemas mais comuns de qualidade é relacionado à completude de dados. Em suas palavras, como é definido completude? Qual o impacto em uma análise exploratória de dados?
7. Qual a completude para cada uma das variáveis do seu banco de dados?
8. Realize uma operação de imputação de dados usando o pacote MICE.
9. Crie um dashboard Shiny onde seja possível selecionar (tire um print-screen da tela final do sistema):
   1. uma variável da sua base de dados e um gráfico em linha seja mostrado na tela;
   2. escolher a cor da linha do gráfico;
   3. selecionar o limite inferior e superior do eixo X do gráfico;
   4. selecionar o limite inferior e superior do eixo Y do gráfico.
10. Disponibilize os códigos (RMarkdown e Shiny) em uma plataforma de compartilhamento de códigos (sugestão GitHub)

Assim que terminar, salve o seu arquivo PDF e poste no Moodle. Utilize o seu nome para nomear o arquivo, identificando também a disciplina no seguinte formato: “nomedoaluno\_nomedadisciplina\_pd.PDF”.

Tabela

Descrição gerada automaticamente

Italy, France, Germany

Brazil, Argentina,Paraguay

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente