Projeto de Disciplina

Análise exploratória de dados [21E1_3]

Wagner Oliveira

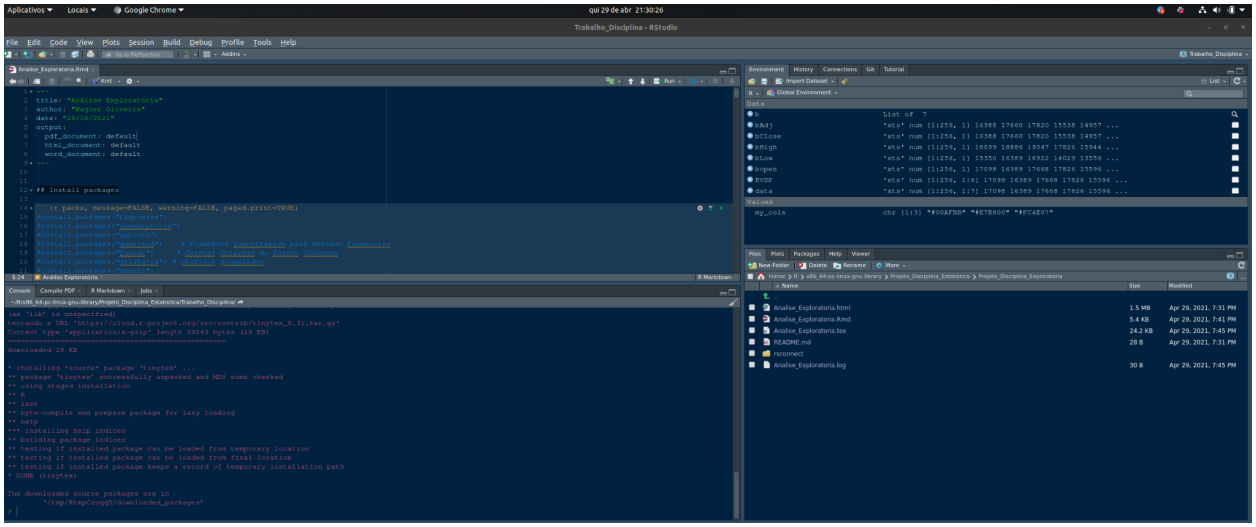
Sumário

1 O relatório final deve ser apresentado utilizando RMarkdown. Nesse relatório devem haver:

a. imagens estáticas ("prints" de tela, imagens da internet - com a devida fonte mencionada - ou figuras criadas pelo aluno fora do ambiente do R);

b. imagens geradas através do ambiente R, particularmente com a biblioteca ggplot;

c. links clicáveis (como fontes e referências).



2 Escolha uma base de dados para realizar esse projeto.

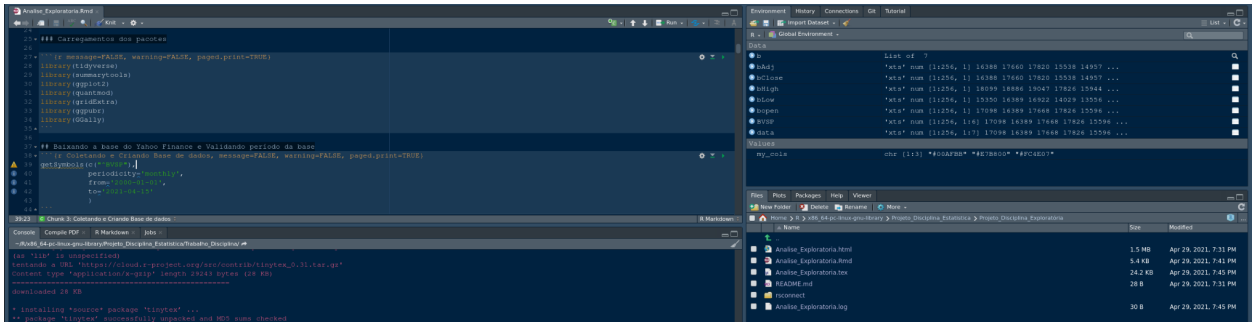
Essa base de dados será utilizada durante toda sua análise. Essa base necessita ter 4 (ou mais) variáveis de interesse, onde todas são numéricas (confira com o professor a possibilidade de utilização de dados categóricos). Observe que é importante que haja dados faltantes em pelo menos uma variável para 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.

Base dados escolhida são as cotações do índice Bovespa

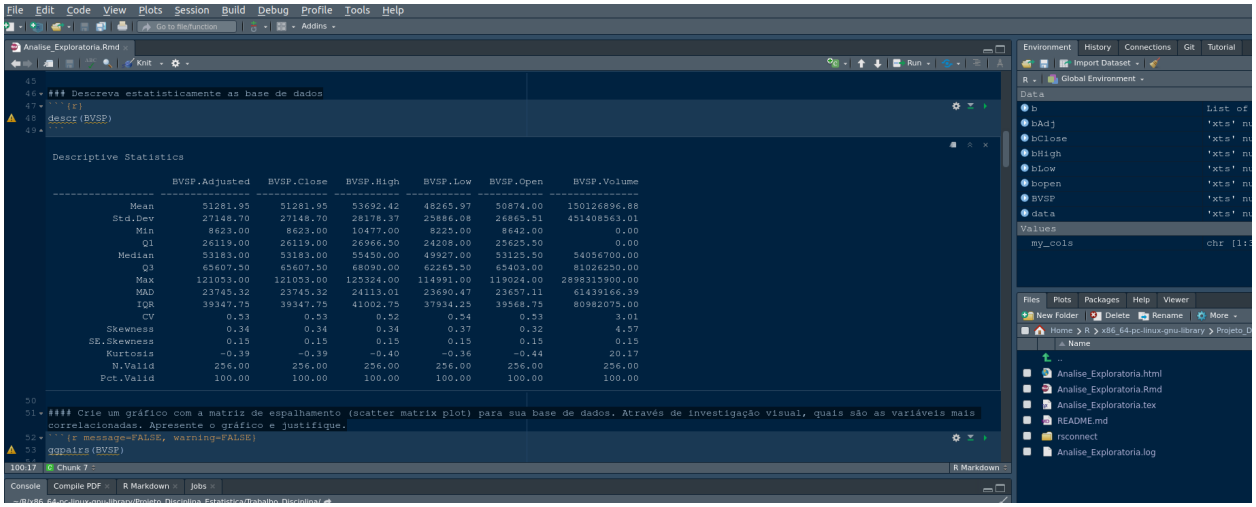
Variáveis
BVSP.Adjusted BVSP.Close BVSP.High BVSP.Low BVSP.Open BVSP.Adjusted

Motivo: Atualmente estou trabalhando com projeto particular envolvendo Mercado Financeiro e gostaria de explorar as possibilidades de criar um ambiente de análise baseado em dados históricos.

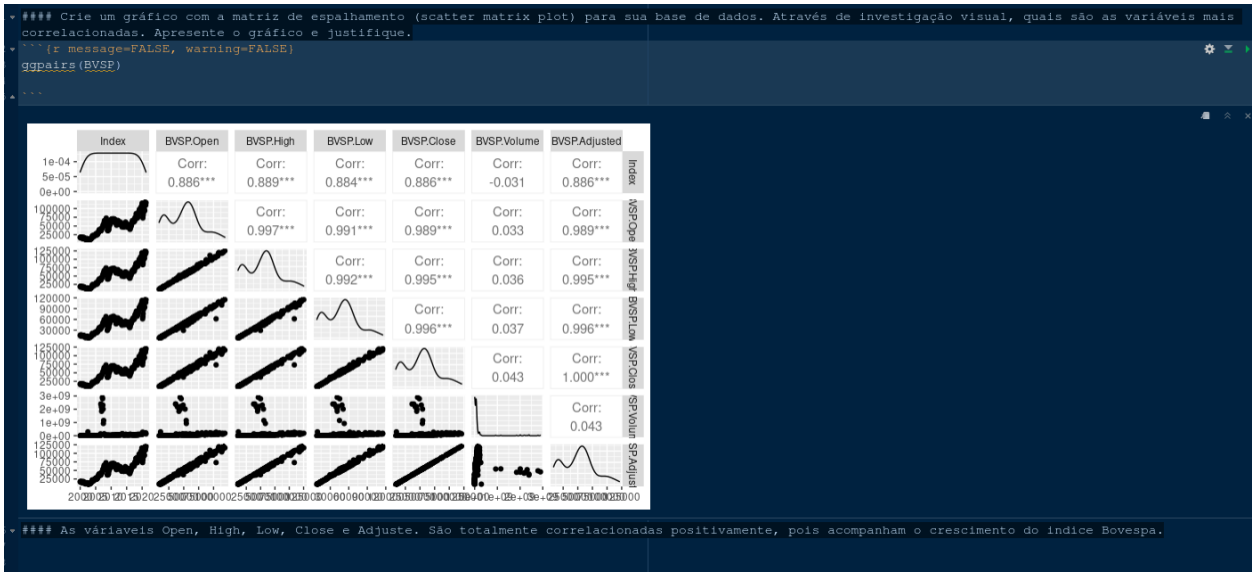
Resultado Esperado: Conseguir explorar a correlação entre os dados de abertura, fechamento, máxima, mínima e Ajuste do índice Bovespa



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.



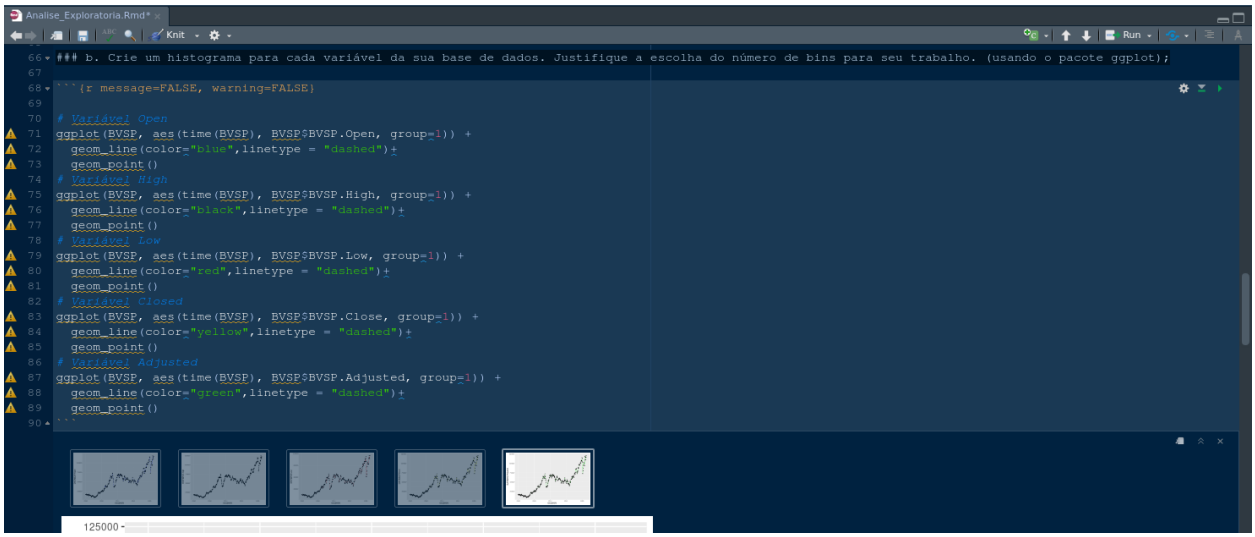
As variáveis Open, High, Low, Close e Adjust. São totalmente correlacionadas positivamente, pois acompanham o crescimento do índice Bovespa.

5 Sobre a normalidade das variáveis:

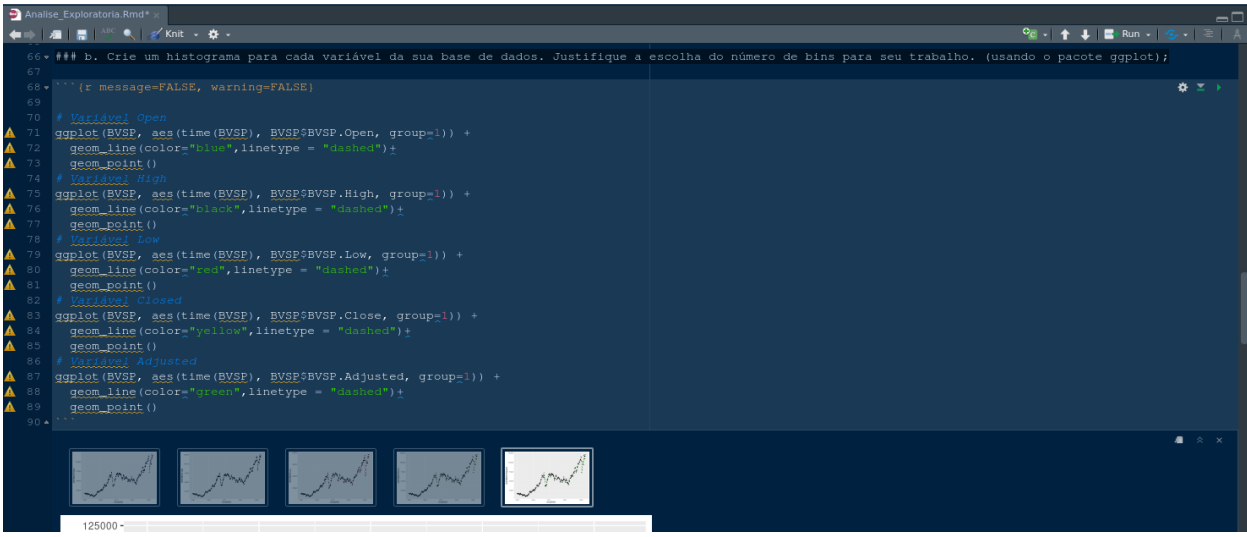
a. Descreva o que é uma distribuição normal;

Distribuição Normal é a mais importante distribuição contínua. Importante para diversas práticas, destaco o Teorema Central do Limite, para garantir a média de um conjunto de dados que converge para uma distribuição normal conforme o número de dados cresce.

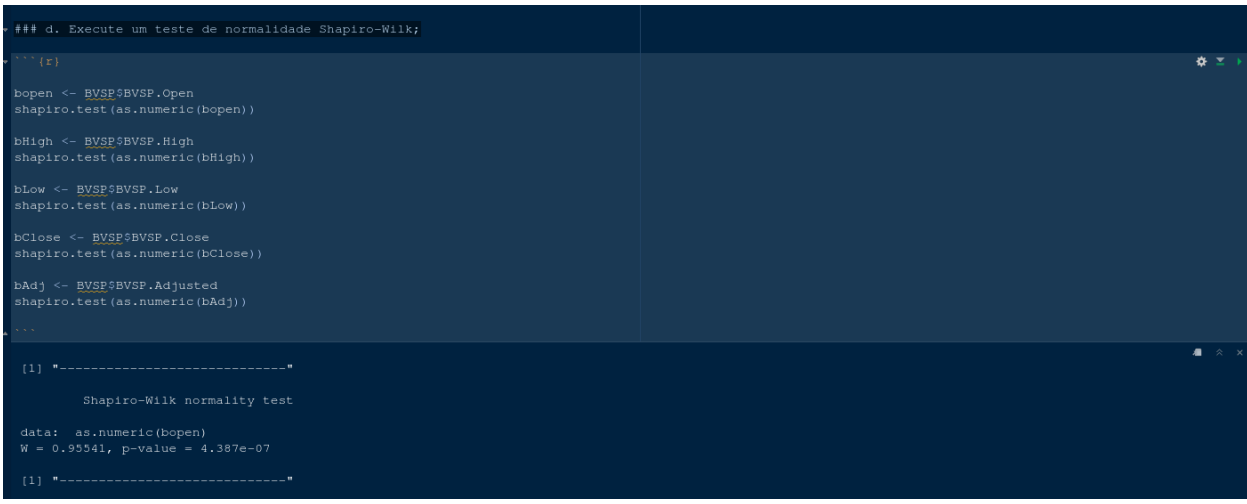
b. Crie um histograma para cada variável da sua base de dados. Justifique a escolha do número de bins para seu trabalho. (usando o pacote ggplot);



c. Crie um gráfico Q-Q para cada variável de sua base de dados. (use as funções presentes no pacote ggpubr);



d. Execute um teste de normalidade Shapiro-Wilk;



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 está relacionado à completude de dados. Em suas palavras, como é definido a completude? Qual o impacto em uma análise exploratória de dados?

Nesta disciplina meu entendimento é baseado no percentual de registros ou campos preenchidos em uma base, ou seja, poucos dados como NA(null)

Atualmente a análise exploratória desempenha um grande papel fornecendo entendimento sobre os dados e identificando padrões e revelando informações ocultas em uma estrutura de dados. Muitas pesquisas atualmente utilizam a análise exploratória para formular hipóteses antes mesmos da realização da modelagem dos dados, sendo assim os pesquisadores podem decidir qual técnica melhor será aplicada para tratamento dos 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):

a. uma variável da sua base de dados e um gráfico em linha seja mostrado na tela;

b. escolher a cor da linha do gráfico;

c. selecionar o limite inferior e superior do eixo X do gráfico;

d .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)

Código github

https://github.com/Wagner85/analise_exploratorio_BVSP

https://rpubs.com/wagner85/AnaliseExploratoria_Bovespa