Trabalho de conclusão Rodrigo Cunha

1. 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.

R: Sobre o conjunto de dados:
Este caso requer o desenvolvimento de uma segmentação de clientes para definir a estratégia de marketing. O conjunto de dados de amostra resume o comportamento de uso de cerca de 9.000 portadores de cartão de crédito ativos durante os últimos 6 meses. O arquivo está no nível do cliente variáveis comportamentais.
BALANCE : Saldo valor restante em sua conta para fazer compras

PURCHASES : Quantidade de compras feitas na conta

ONEOFFPURCHASES : Valor máximo de compra feito de uma só vez

CREDITLIMIT : Limite de cartão de crédito por usuário

PAYMENTS : Valor do pagamento feito pelo usuário

MINIMUM_PAYMENTS : Valor mínimo de pagamentos feitos pelo usuário

MONTH: Mês referente dos dados. Todos os dados preenchidos

Sex : Definição se é Masculino ou Feminino. Todos os dados preenchido

2. Utilizando o pacote summarytools (função descr), descreva estatisticamente a sua base de dados.

3. 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.

Variáveis mais correlacionadas

Credit_Limit X Month : Significa que temos limite de crédito em todos os meses

One Of Purchase X Moth : Significa que temos compras de uma só vez em todos os meses

One of Purchase X Credit Limit : Significa que as compras realizadas de uma só vez já preenchem todo o limite de crédito

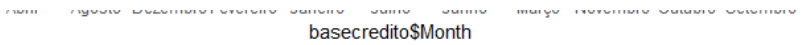
Purchases X Credit Limit : Significa que todo mês praticamente compram todo o limite de crédito

4. Sobre a normalidade das variáveis:
    a. Descreva o que é uma distribuição normal;

R: A distribuição normal (gaussiana) é uma curva simétrica em torno do seu ponto médio, apresentando um formato de sino. Uma das distribuições de probabilidade mais utilizadas para modelar fenômenos naturais.

    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);

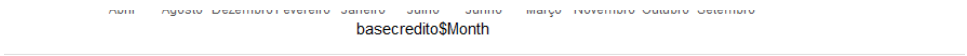
Balance



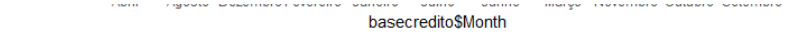

Purchases



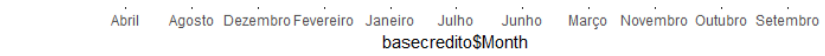

ONEOFFPURCHASES



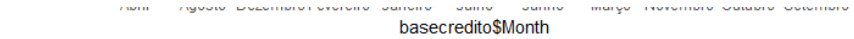

CREDIT_LIMIT




PAYMENTS



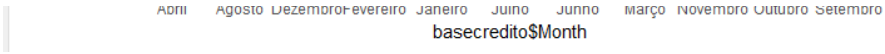

MINIMUM_PAYMENTS




Month




Sex




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

Balance



PURCHASES




ONEOFF_PURCHASES


CREDIT_LIMIT




PAYMENTS




MINIMUM_PAYMENTS

d. Execute um teste de normalidade Shapiro-Wilk;

```
W = 0.034112, p-value < 2.2e-16
```

e. Baseado nos itens anteriores, é possível afirmar que algumas das variáveis se aproximam de uma distribuição normal? Justifique.

Não, porque a média, mediana e modados dados não possuem o mesmo valor.

5. 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?

R: Completude de dados é o percentual de registros ou campos preenchido em uma base de dados. O impacto do não preenchimento pode mudar toda assertividade do relatório. Por exemplo, se faço uma entrevista com 100 pessoas perguntando se “gostam de futebol?” e somente 50% respondem os 50% restantes (completude de registro) mudam todo o contexto porque desses 50%, 45% poderiam dizer sim, impactando uma análise. Uma observação importante é que 0 não é nulo podendo confundir em uma análise comparado com completude.

6. Qual a completude para cada uma das variáveis do seu banco de dados?

R: Em relação a cada variável

BALANCE : Saldo valor restante em sua conta para fazer compras (não existe porque quando for zerado significa que não existe saldo)

PURCHASES : Quantidade de compras feitas na conta (não existe porque significa que não realizou compras)

ONEOFFPURCHASES : Valor máximo de compra feito de uma só vez (dependendo da análise 4302/8950)

CREDITLIMIT : Limite de cartão de crédito por usuário  (não existe porque todo cartão possui um limite)

PAYMENTS : Valor do pagamento feito pelo usuário  (não existe porque estar zerado significa inadimplência)

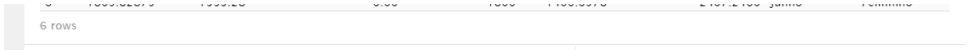
MINIMUM_PAYMENTS : Valor mínimo de pagamentos feitos pelo usuário. (240/8950) os nulos significam aqueles que não pagaram e existe um valor mínimo.

PRCFULLPAYMENT : Percentual do pagamento integral pago pelo usuário. (240/8950) o mesmo valor do anterior porque se não existe valor pago não existe percentual tbm.

MONTH: Mês referente dos dados. Todos os dados preenchidos

Sex : Definição se é Masculino ou Feminino. Todos os dados preenchido

7. Realize uma operação de imputação de dados usando o pacote MICE.




8. Crie um dashboard Shiny onde seja possível selecionar (tire um print-screen da tela final do sistema):
    f. uma variável da sua base de dados e um gráfico em linha seja mostrado na tela;
    g. escolher a cor da linha do gráfico;
    h. selecionar o limite inferior e superior do eixo X do gráfico;
    i. selecionar o limite inferior e superior do eixo Y do gráfico.

9. Disponibilize os códigos (RMarkdown e Shiny) em uma plataforma de compartilhamento de códigos (sugestão GitHub)