Olá candidata(o),

Bem vinda(o) ao desafio da Data Minds!!!
Agradecemos seu interesse em trabalhar na **Data Minds**.

Nesta fase, solicitamos que realize nosso desafio para avaliarmos o seu potencial dentro do projeto.

Instruções

1. Os cenários apresentados no desafio são **hipotéticos**, portanto, qualquer semelhança com a realidade é mera coincidência.
2. Leia com atenção o enunciado e procure responder as perguntas de acordo com suas **premissas**.
3. Defina objetos ou crie **exemplos de dados simulados** caso necessário.
4. **O desafio foi construído para ser abrangente.** Faça o que você puder sabendo que tudo será levado em consideração na avaliação.
5. Em caso de dúvidas, envie um email com suas dúvidas para desafio@dataminds.com.br com o título **[DESAFIO DA DATA MINDS: DÚVIDAS]**.
6. Envie todos os resultados, assim como códigos, objetos e exemplos de dados construídos para o desafio, para desafio@dataminds.com.br com o título **[DESAFIO DA DATA MINDS: SOLUÇÃO]**.
7. Você tem 7 (sete) dias corridos a partir do recebimento deste arquivo para resolver e devolver os resultados do desafio.
8. Boa sorte!

Desafio 1

Um cliente da Data Minds que recebe diariamente milhares de registros do Cadastro Positivo, contendo informações do histórico de crédito de todos os usuários do sistema financeiro, solicita que você construa uma aplicação, utilizando sua linguagem preferida, para extrair, transformar e carregar registros do tipo Parcelado.

Algumas informações sobre o histórico do cadastro positivo:

- São reportados todos os vencimentos, pagamentos e saldos do mês corrente e dos últimos 12 meses, totalizando 13 meses;
- Parcelas antecipadas e pagas no período são reportadas;
- Parcelas anteriores, com data de vencimento de até 5 anos, da data de pagamento e pagas no período (13 meses) são reportadas;
- Parcelas anteriores, com data de vencimento superior a 5 anos, da data de pagamento e pagas no período (13 meses), não são reportadas.
- Todas as informações sobre o Cadastro Positivo são de domínio público e podem encontradas no site da CIP: https://www.nuclea.com.br/cadastro-positivo/

Especificamente o reporte de Parcelados contém as seguintes informações:

- Dados do contrato:
    - Data de contratação;
    - Indicador: pós-fixado ou pré-fixado;
    - Data de vencimento da última parcela;
    - Valor contratado futuro (Valor original na data da contratação com juros e encargos);
    - Quantidade total de parcelas do contrato;
    - Quantidade de parcelas em aberto.
- Informações das parcelas/prestações:
    - Parcelas anteriores: Parcelas pagas ou não pagas nos últimos 12 meses e parcelas pagas mesmo com data de vencimento futura. Observação: a data de pagamento da parcela deve ser posterior ou igual à data de contratação e anterior ou igual à data de apuração.
    - Número da parcela;
    - Data de vencimento;
    - Valor da parcela;
    - Data de pagamento;
    - Valor pago;
    - Situação da parcela.
- Parcelas futuras:
    - Data de vencimento da próxima parcela a vencer;
    - Valor da próxima parcela a vencer;
    - Quantidade de parcelas a vencer.

Exemplo de reporte de parcelados:

O cliente contratou em 10.01.2019 um financiamento de veículo de R$25.000,00, em 25 parcelas iguais de R$1.200,00. Valor contratado futuro: 25 x R$1.200,00 = R$30.000,00. A primeira parcela foi paga em 09.02.2019 no seu valor total.

- Dados do contrato
    - Data de contratação: 10.01.2019
    - Indicador: pré-fixado
    - Data de vencimento da última parcela: 10.02.2021
    - Valor contratado: R$ 30.000,00 (valor futuro)
    - Quantidade total de parcelas do contrato: 25
    - Quantidade de parcelas em aberto: 24
- Informações das parcelas / prestações
    - Parcelas anteriores (12 meses para trás):
        - Número da parcela: 01
        - Data de vencimento: 10.02.2019
        - Valor da parcela: R$ 1.200,00
        - Data de pagamento: 09.02.2019
        - Valor pago: R$ 1.200,00
        - Situação da parcela: total
    - Parcelas futuras:
        - Data de vencimento da próxima parcela a vencer: 10.03.2019
        - Valor da próxima parcela a vencer: R$ 1.200,00
    - Quantidade de parcelas a vencer: 24
- Data da apuração no sistema legado: 14.02.2019

Considere que os registros de dados sejam enviados ao cliente no formato XML no seguinte layout:

<Tag> / Atributo	Nome do Campo	Tipo	Mult.
<EnvoHstCrd>	EnvioHistoricoCredito	Tag	[1..1]
CnpjIf	Cnpj da Fonte	Atributo	[1..1]
DtRms	DataRemessa	Atributo	[1..1]
<Cli>	Cliente	Tag	[0..n]
IdfcCli	Identificacao	Atributo	[1..1]
<Opr>	Operacao	Tag	[0..n]
NrUnco	NumeroUnico	Atributo	[0..1]
DtCtrc	DataContratacao	Atributo	[0..1]
DtAprc	DataApuracao	Atributo	[0..1]
CdMdld	Modalidade	Atributo	[0..1]
<DetOpr>	DetalheOperacao	Tag	[0..1]
InPreFix	IndicadorPreFixado	Atributo	[0..1]
DtVnctUltPcl	DataVencimentoUltimaParcela	Atributo	[0..1]
VlCtrdFut	ValorContratadoFuturo	Atributo	[0..1]

QtPcl	QuantidadeParcelas	Atributo	[0..1]
</DetOpr> FIM	DetalheOperacao	Tag	[0..1]
<PclAnt>	ParcelaAnterior	Tag	[0..n]
DtVnctPclAnt	DataVencimento	Atributo	[0..1]
VlPclAnt	Valor	Atributo	[0..1]
<PgtoPclAnt>	PagamentoParcelaAnterior	Tag	[0..n]
DtPgtoPclAnt	DataPagamento	Atributo	[0..1]
VlPgtoPclAnt	ValorPagamento	Atributo	[0..1]
</PgtoPclAnt>	PagamentoParcelaAnterior	Tag	[0..n]
</PclAnt> FIM	ParcelaAnterior	Tag	[0..n]
<PclFut>	ParcelasFuturas	Tag	[0..n]
DtVnctPrxPcl	DataVencimentoProximaParcela	Atributo	[0..1]
VlPrxPcl	ValorProximaParcela	Atributo	[0..1]
QtPclVncr	QuantidadeParcelasAVencer	Atributo	[0..1]
QtPclPgr	QuantidadeParcelasAPagar	Atributo	[0..1]
</PclFut> FIM	ParcelasFuturas	Tag	[0..n]
</Opr> FIM	Operacao	Tag	[0..n]
</Cli> FIM	Cliente	Tag	[0..n]
</EnvoHstCrd FIM	EnvioHistoricoCredito	Tag	[1..1]

Questões:

1) Verifique se as informações acima são suficientes para desenvolver a aplicação solicitada. As informações são consistentes? O layout contempla todos os requisitos?
2) Construa uma função para acessar as informações do dado no formato original (XML). Faça suposições e simplificações. Construa exemplos e valide suas premissas.
3) Considerando que o formato apresentado é um documento hierárquico, transforme os dados para um formato mais adequado para fins analíticos. Construa uma função para acessar e validar os campos de dados no novo formato.

Desafio 2

Um cliente da Data Minds que utiliza um banco de dados estruturado, contendo bilhões de registros do Cadastro Positivo (para mais informações veja a descrição da questão anterior), solicita que você faça uma consulta em sua base para entender o valor médio em atraso por usuário do sistema financeiro considerando produtos do tipo parcelado.

Considere que o cliente possui as seguintes entidades em sua base de dados:

Entidade Pessoa

Nome do Campo	Tipo	Descrição
CPF	Alfanumérico	Identificador único de usuário

Entidade Contratos

Nome do Campo	Tipo	Descrição
CNPJ	Alfanumérico	Identificador único do banco
NumeroUnico	Alfanumérico	Número único do contrato
CPF	Alfanumérico	Identificador único do cliente
Modalidade	Caracter	Modalidade do contrato
IndicadorPreFixado	Lógico	Identifica se o contrato é pré-fixado
DataVencimentoUltimaParcela	Data	Data de vencimento da última parcela
ValorContratadoFuturo	Numérico	Valor contratado
QuantidadeParcelas	Numérico	Quantidade de parcelas

Entidade Pagamentos

Nome do Campo	Tipo	Descrição
CNPJ	Alfanumérico	Identificador único do banco
NumeroUnico	Alfanumérico	Número único do contrato
CPF	Alfanumérico	Identificador único do cliente
DataVencimento	Data	Data de vencimento da parcela
Valor	Numérico	Valor da parcela
DataPagamento	Data	Data de pagamento da parcela
ValorPagamento	Numérico	Valor do Pagamento

Observações:

1) Os registros de Contratos e Pagamentos não são obrigatórios. Isso significa que existem pessoas que não possuem contratos assim como existem contratos que não possuem pagamentos.
2) O número único do contrato é único dentro dos contratos de um mesmo banco. Assim, diferentes bancos podem compartilhar o mesmo número único mesmo se tratando de contratos diferentes.
3) O valor do campo Modalidade pode ser “Parcelado”, “Cartão de Crédito”, “Rotativo” ou "Consórcio".

4) O valor do pagamento pode ser igual a zero. Isso significa que ou o pagamento está atrasado ou a parcela ainda não venceu.

Questões:

1) Verifique se as informações acima são suficientes para desenvolver a consulta solicitada. As informações são consistentes?
2) Construa uma consulta em SQL para calcular a quantidade de contratos do tipo parcelado por cliente.
3) Construa uma consulta em SQL para calcular a quantidade de contratos em atraso em contratos do tipo parcelado por cliente.
4) Construa uma consulta em SQL para calcular o valor médio em atraso de contratos do tipo parcelado por cliente.

Desafio 3

Um cliente da Data Minds solicita que você construa um score de crédito customizado para um novo público. Uma amostra analítica contendo 1000 registros foi extraída do banco de dados do cliente contendo os seguintes atributos:

Attribute 1: (qualitative)
Status of existing checking account
A11 : ... < 0 DM
A12 : 0 <= ... < 200 DM
A13 : ... >= 200 DM / salary assignments for at least 1 year
A14 : no checking account

Attribute 2: (numerical)
Duration in month

Attribute 3: (qualitative)
Credit history
A30 : no credits taken/ all credits paid back duly
A31 : all credits at this bank paid back duly
A32 : existing credits paid back duly till now
A33 : delay in paying off in the past
A34 : critical account/ other credits existing (not at this bank)

Attribute 4: (qualitative)
Purpose
A40 : car (new)
A41 : car (used)
A42 : furniture/equipment
A43 : radio/television
A44 : domestic appliances
A45 : repairs
A46 : education
A47 : (vacation - does not exist?)
A48 : retraining
A49 : business
A410 : others

Attribute 5: (numerical)
Credit amount

Attribute 6: (qualitative)
Savings account/bonds

A61 : ... < 100 DM
A62 : 100 <= ... < 500 DM
A63 : 500 <= ... < 1000 DM
A64 : .. >= 1000 DM
A65 : unknown/ no savings account

Attribute 7: (qualitative)
Present employment since
A71 : unemployed
A72 : ... < 1 year
A73 : 1 <= ... < 4 years
A74 : 4 <= ... < 7 years
A75 : .. >= 7 years

Attribute 8: (numerical)
Installment rate in percentage of disposable income

Attribute 9: (qualitative)
Personal status and sex
A91 : male : divorced/separated
A92 : female : divorced/separated/married
A93 : male : single
A94 : male : married/widowed
A95 : female : single

Attribute 10: (qualitative)
Other debtors / guarantors
A101 : none
A102 : co-applicant
A103 : guarantor

Attribute 11: (numerical)
Present residence since

Attribute 12: (qualitative)
Property
A121 : real estate
A122 : if not A121 : building society savings agreement/ life insurance
A123 : if not A121/A122 : car or other, not in attribute 6
A124 : unknown / no property

Attribute 13: (numerical)
Age in years

Attribute 14: (qualitative)
Other installment plans
A141 : bank
A142 : stores
A143 : none

Attribute 15: (qualitative)
Housing
A151 : rent
A152 : own
A153 : for free

Attribute 16: (numerical)
Number of existing credits at this bank

Attribute 17: (qualitative)
Job
A171 : unemployed/ unskilled - non-resident
A172 : unskilled - resident
A173 : skilled employee / official
A174 : management/ self-employed/
highly qualified employee/ officer

Attribute 18: (numerical)
Number of people being liable to provide maintenance for

Attribute 19: (qualitative)
Telephone
A191 : none
A192 : yes, registered under the customers name

Attribute 20: (qualitative)
foreign worker
A201 : yes
A202 : no

Attribute 21: (numerical)
response variable
1: bad
2: good

Observações:

1) O atributo binário “response variable” é a variável resposta do problema em que a categoria “bad” representa clientes inadimplentes (maus pagadores) e “good” clientes que pagam suas contas em dia (bons pagadores).
2) O arquivo contendo a amostra de dados se encontra no seguinte endereço: https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/german.data

Questões:

1) Verifique se as informações acima são suficientes para desenvolver o modelo solicitado.
2) Faça uma análise exploratória dos dados buscando entender a distribuição dos dados e a relação dos atributos com a variável resposta.
3) Estime um modelo de regressão logística para construir um score de crédito em que quanto maior o valor do score menor é a probabilidade do cliente ser um mau pagador.
4) Avalie a performance do modelo estimado em 3).
5) Estime outro modelo utilizando uma técnica de sua escola para construir um score de crédito utilizando os mesmos dados utilizados em 3).
6) Compare a performance dos modelos estimados em 3) e 5). Discuta as qualidades e limitações de cada modelo.