SÃO VICENTE – REPÚBLICA DE CABO VERDE
DEPARTAMENTO DE ENGENHARIA E RECURSOS DO MAR

CURSO DE LICENCIATURA em ENGENHARIA INFORMÁTICA e SISTEMA E COMPUTACIONAIS
"IMPLEMENTAÇÃO PRATICA DO ALGORITMO K-MEANS NO DATASET PKDD99"

CADEIRA:ENGENHARIA DO CONHECIMENTO

DOCENTE: DR. ESTANISLAU LIMA

Autor: Hernâni Baptista nº 4933

Mindelo, 2020

Indice

1.Resumo

O trabalho proposto pelo professor de engenharia de conhecimento é fazer um estudo de mineração de dados (PKDD'99 Discovery Challenge -Guia para o conjunto de dados financeiros. Esse banco de dados foi preparado por Petr Berka e Marta Sochorova) de um banco que oferecia serviços particulares. Os serviços incluem gerenciamento de contas , oferta de empréstimos,etc. O banco deseja melhorar seus serviços encontrado grupos interessantes de clientes (por exemplo, para diferenciar entre bons e maus clientes). Os gerentes dos bancos têm apenas uma vaga ideias de quem é um bom cliente (a quem oferecer alguns serviços adicionais) e quem é cliente ruim (a quem observar atentamente para minimizar as perdas do banco).Felizmente, o banco armazena dados sobre seus clientes , as contas(transações dentro vários meses),os empréstimos ja concedidos, os cartões de créditos emitidos. Assim , os gerentes dos bancos esperam encontrar algumas respostas(e perguntas também) analisando esses dados.

2.Descrição de dados das tabelas

2.1.Account

- constituida por 4500 linhas e 4 colunas (account_id,district_id,frequency e account_date) , *account_id* é o identificador de cada conta do banco , *district_id* é identificador de localização do filial da conta , frequency é uma frequência de emissão de declarações em que caracterizada em 3 tipos : emissão mensal, emissão semanal e emissão após transação.

2.2.Client

- Constituida por 5369 linhas e 3 colunas(client_id,bith_number,district_id),

  client_id é o identificador do cliente do banco , *birth_number* é data de nascimento do cliente em que esta inserida o sexo do cliente ,isto é, para o sexo masculino a data de nascimento esta formatando ano-mes-dia (YY/MM/DD) para o sexo feminino a data de nascimento esta formatado ano-mes-+50dia(YY/MM/+50DD) e o *district_id* é endereço do cliente.

2.3.Disp

- Constituida por 5369 linhas e 4 colunas(disp_id,client_id,account_id,type), *disp_id* é o identificador da provisão do cliente, *client_id* é o identificação do cliente , *account_id* é identificação da conta bancaria e *type* é o tipo de provisão de cada cliente caso é dono da conta(unico de pode pedir empréstimos no banco) ou usuário da conta.

2.4.Order

- constituida por 6471 linhas e 6 colunas (order_id ,account_id,bank_to,account_to,amount,k_symbol) , *order_id* é identificador de ordem de pagamento, *account_id* é identificador da conta do cliente , *bank_to* é o banco destinatário que representada por código de duas letras ,account_to é identificador da conta que vai receber o saldo , amount é valor debitado na conta selecionado e K_symbol é descrição de pagamento que pode ser pagamento de seguro, suporte para uso domestico,pagamento de renda ou pagamento de empréstimo.

2.5.Loan

- constituida por 682 linhas e 7 colunas (loan_id,account_id,date,amount,duration,payments,status), loan_id é identificador de empréstimo feito pelo dono da conta , account_id é identificador de conta do cliente , date é data que quando foi pedido o emprestimo representada YY/MM/DD , amount é quantidade de dinheiro pedido pelo conta do cliente , duration é duração do empréstimo feito pela conta do cliente , payments é o pagamento mensal que devera ser feito pela conta do cliente e status é caraterização do pagamento do empréstimo que caraterizada por 4 categorias **A**(significada contrato concluído,sem problemas),**B** (significa contrato concluído,empréstimo não pago), **C** (significa contrato em execução e ate agora esta tudo bem com os empréstimos) e **D** (significa contrato em execução e esta com dividas sem pagar ainda).

2.6.Transaction

- constituida por 1056320 linhas e 10 colunas (trans_id,account_id,trans_date,type,operation,amout,balance,k_symbol,bank,account), trans_id é o identificador da transferência , account_id é identificador da conta que efetuou a transferência , date é data que foi feita a transferência , type é caracterizada

que 2 tipos de transferência credito ou retirado de saldo , operation é modo de transação que caracterizada por 5 tipos : deposito de saldo pelo cartão de credito , depositado em dinheiro , cobrança de outro banco , levantamento em dinheiro e remessa para outra banco . O amount é montante em dinheiro , balance é o saldo depois da transferência , k_symbol é caracterização da transferência em 6 tipos : pagamento seguro , pagamentos por extrato, juros de sanção se saldo negativo , para família , aposentadoria e pagamento de empréstimo .O bank é banco do parceiro e account é a conta do parceiro .

2.7. Card

- constituída por 892 linhas e 4 colunas (card_id,disp_id,card_type,issued), card_id é identificador do cartao de credito , disp_id é identificador da disposição de uma conta do cliente , type é caracterizado pelo 3 tipo de cartao: junior, classic e gold. O issued é a data de emissão do cartão em formato YY/MM/DD.

2.8.District

- constituida por 77 linhas e 16 colunas(A1,A2,A3,A4,A5,A6,A7,A8,A9,A10,A11,A12,A13,A14,A15,16),A1(district_id) é o código de um distrito ,A2 é nome do distrito,A3 é região que se encontra , A4 é o numero de habitantes , A5 é numero de municípios com habitantes menor que 499, A6 é numero de municípios com habitantes entre 500 e 1999, A7 é numero de municípios com habitantes habitantes entre 2000 e 9999 , A8 é numero de municípios com habitantes superior a 10000, A9 é o numero de cidades , A10 é proporção de habitantes urbanos , A11 é salário medio , A12 é a taxa de desemprego ‘95 ,A13 é taxa de desemprego ‘96, A14 é numero de empresários por 1000 habitantes , A15 é numero de crimes cometidos ‘95 e A16 é numero de crimes cometidos ‘96.

3.Ferramentas Utilizadas:

- O processo de pré-processamento foi feito no notepad++.

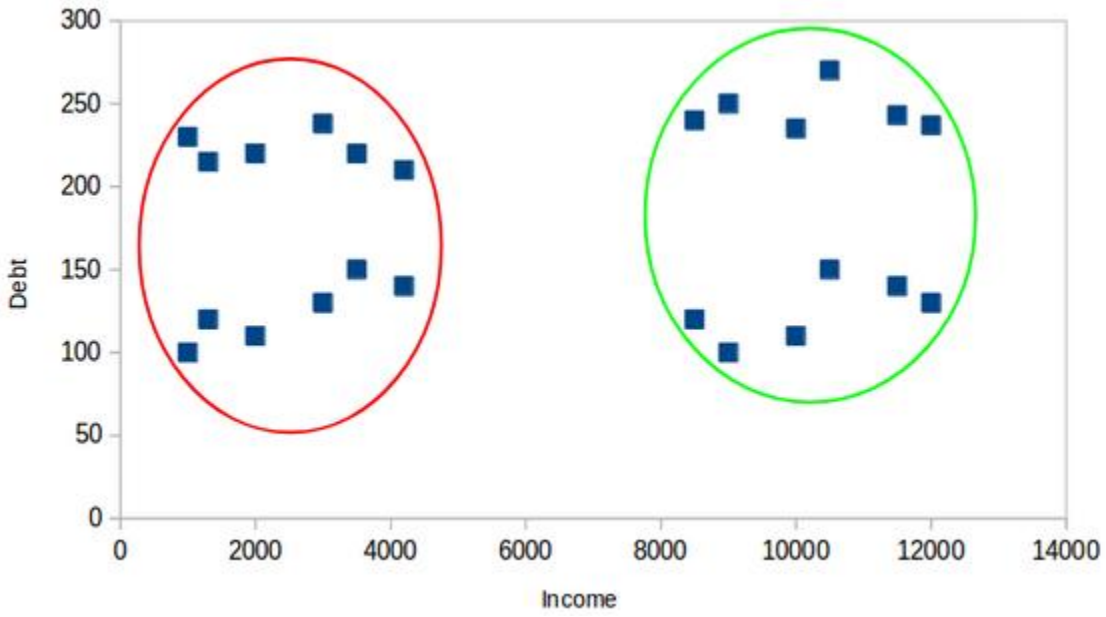
- O processo de pré-processamento foi feito em python(Anaconda-JupyterNotebook).

- O merge de uma tabela foi feita no sql.

-A implementação do algoritmo k-means foi feito no matlab.

4.K-means

K-means clustering é provavelmente um dos primeiros algoritmos de aprendizado não supervisionado que a maioria das pessoas encontra quando inicia um curso de aprendizado de máquina. É fácil de usar e intuitivo. No entanto, se nos dedicarmos à teoria probabilística por trás do K-means, torna-se aparente que o algoritmo faz suposições muito gerais sobre a distribuição dos dados. O algoritmo K-means tenta detectar clusters dentro do conjunto de dados sob os critérios de otimização de que a soma das variações entre os cluster é minimizada. Portanto, o algoritmo de agrupamento K-Means produz uma estimativa mínima de variância (MVE) do estado dos clusters identificados nos dados.



A intuição por trás do algoritmo reside no fato de que, em média, a distância do centróide do cluster (μk) aos elementos dentro do cluster deve ser homogênea entre todos os clusters identificados. Embora o método funcione bem na detecção de clusters homogêneos, inevitavelmente fica aquém devido à suposição simplista sobre a natureza esférica dos clusters inerentes à sua função de otimização (ou seja, assume que todos os clusters têm matrizes de covariância iguais).

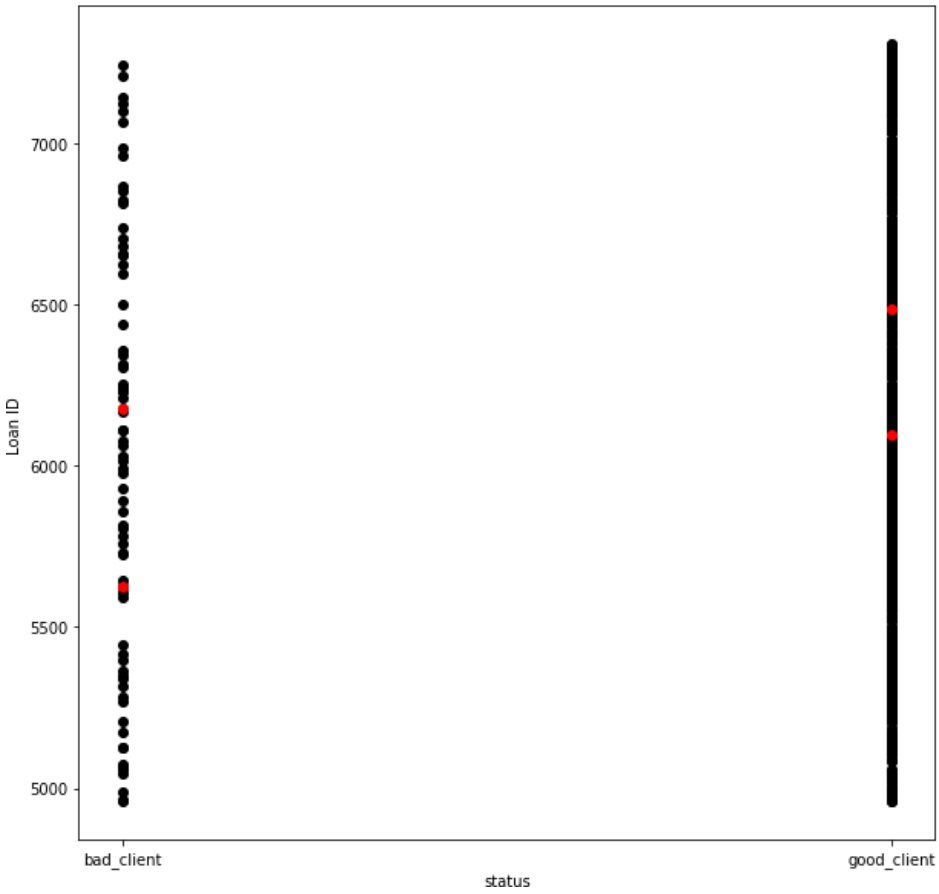
5.Pre-processamento

Primeiramente abri os ficheiros no notepad++ todos e substitui todos os “;” para “,” , assim tive 100% extração dos dados disponíveis . Com isso não foi necessário fazer uma limpeza de dados.

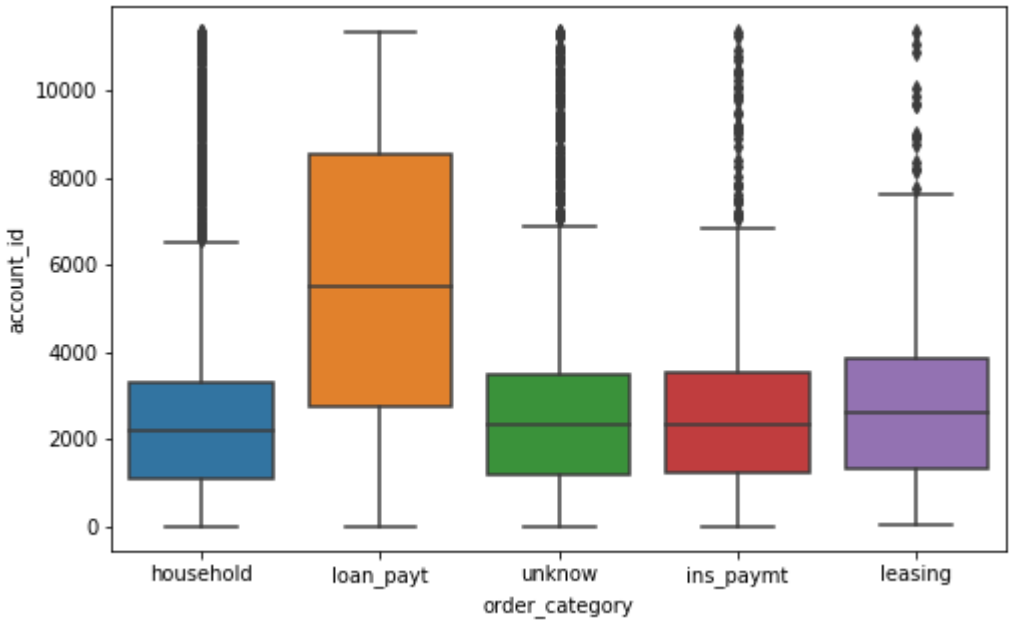
Passei diretamente para transformação de dados que comecei por traduzir todo o que estava em tcheco para inglês .Depois continuei formatando as datas de cada tabela.

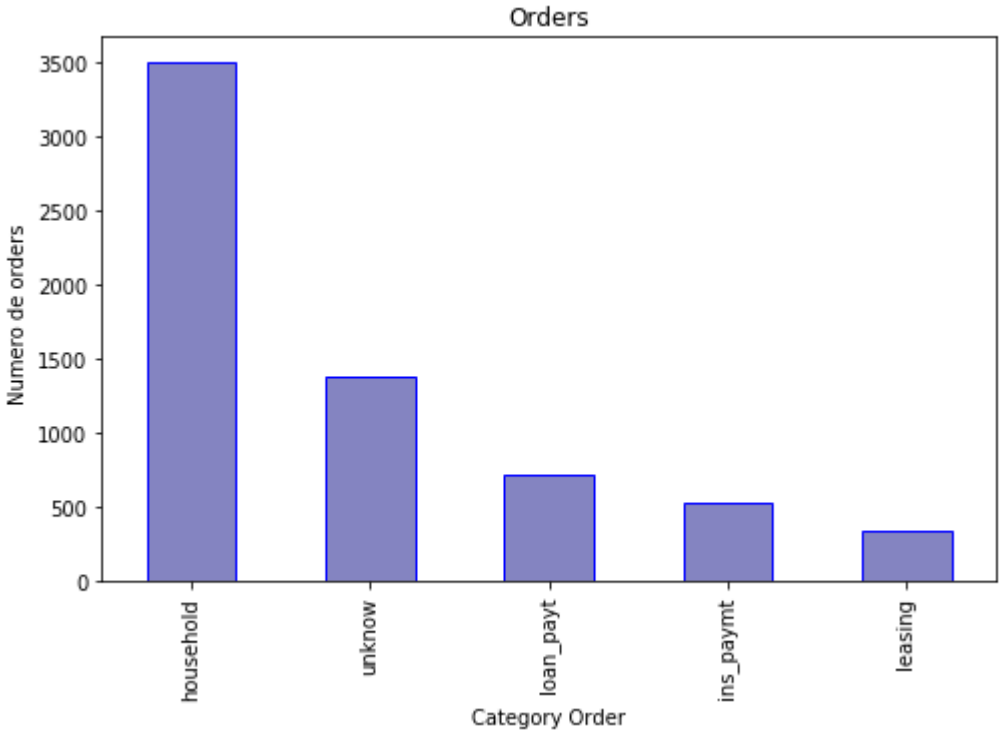
E finalmente começei a fazer redução de dados .

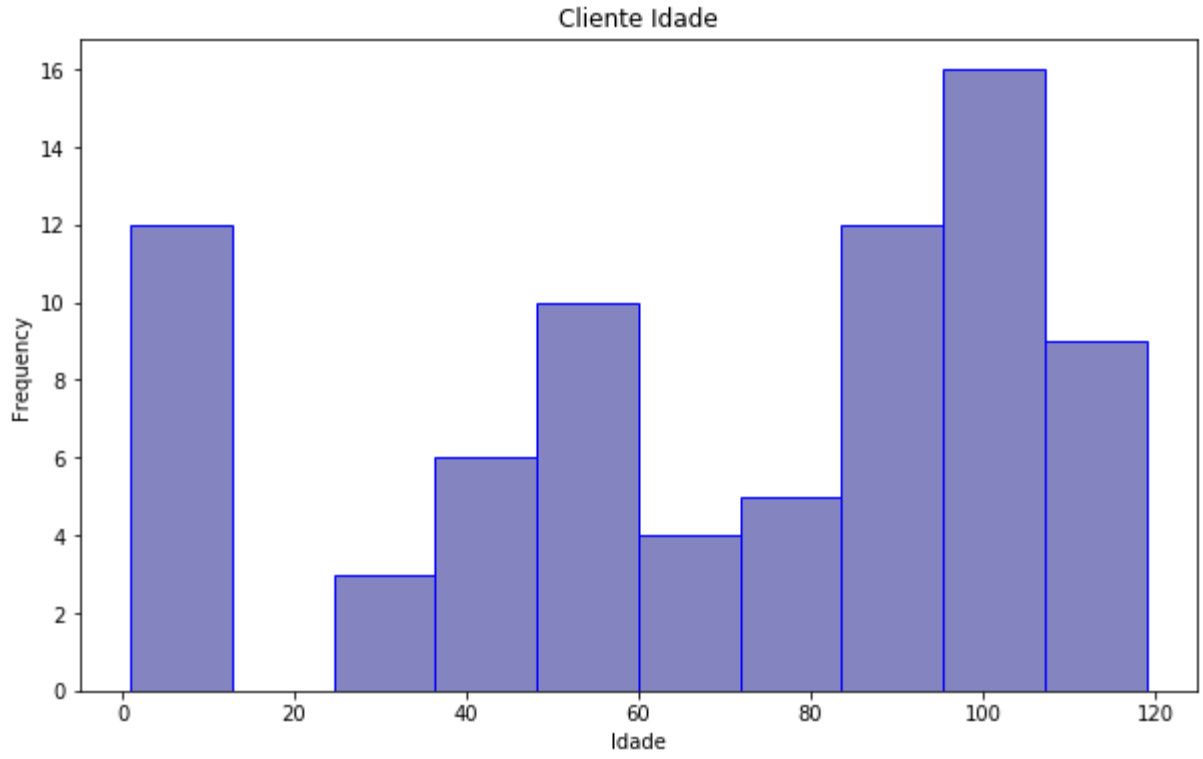
6.Implementação do K-means

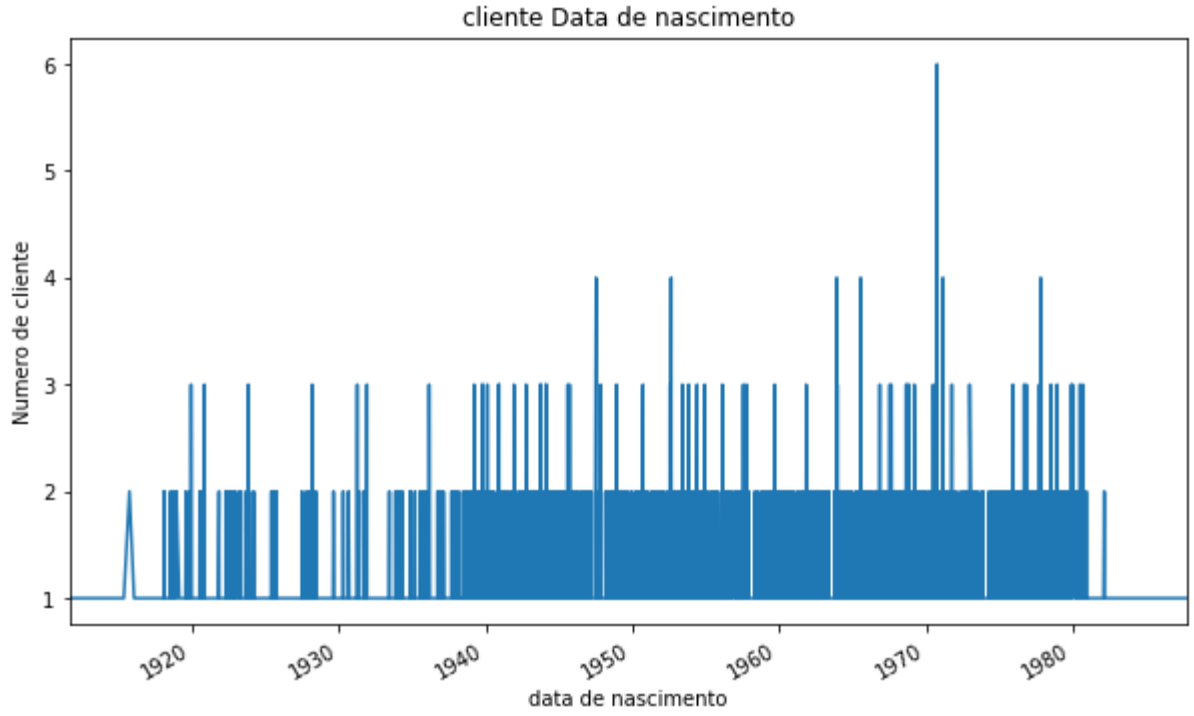


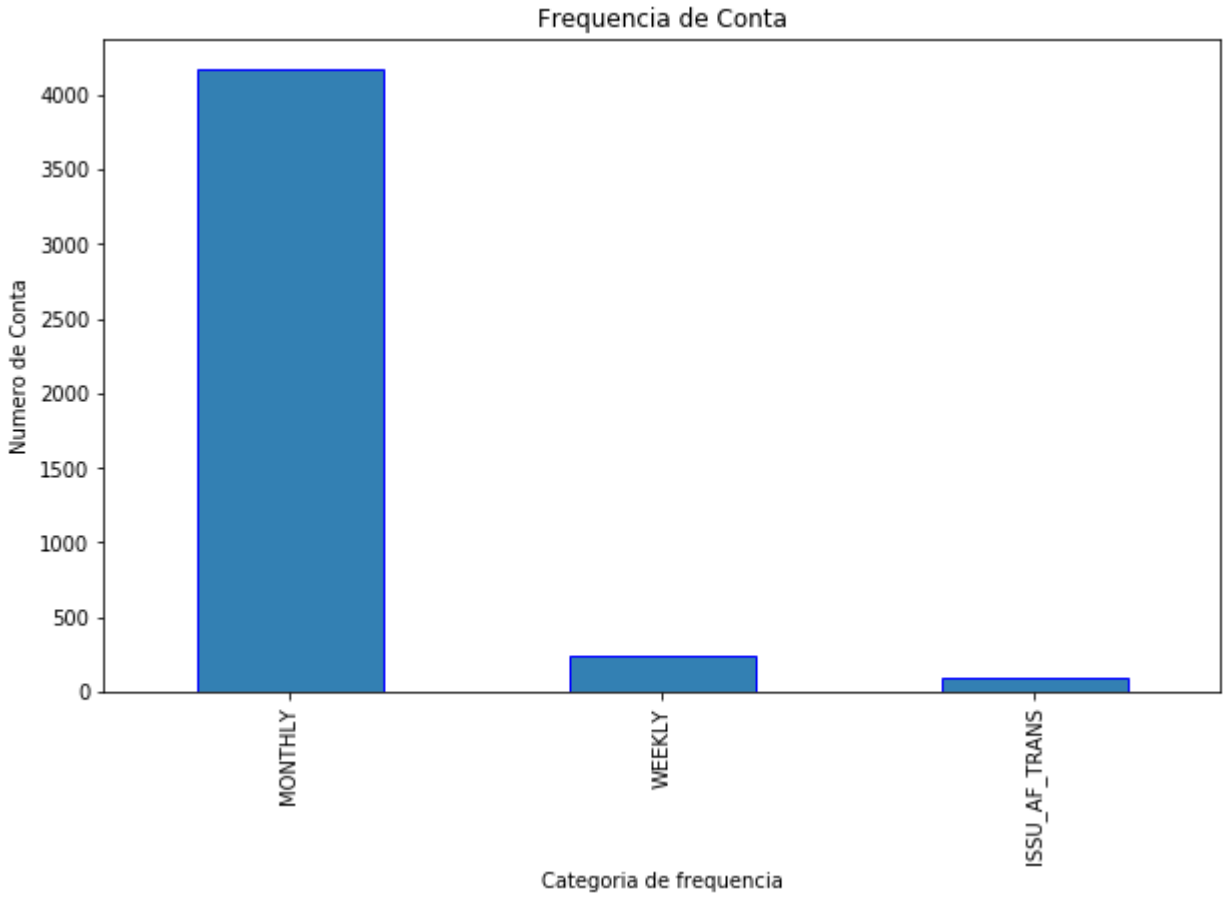
7.Visualização e interpretação

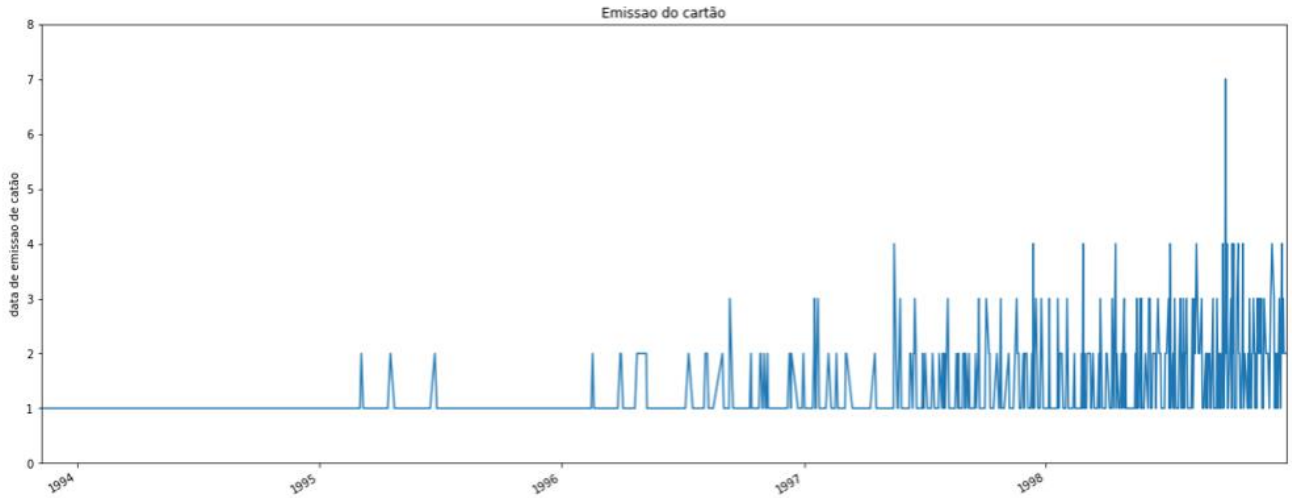


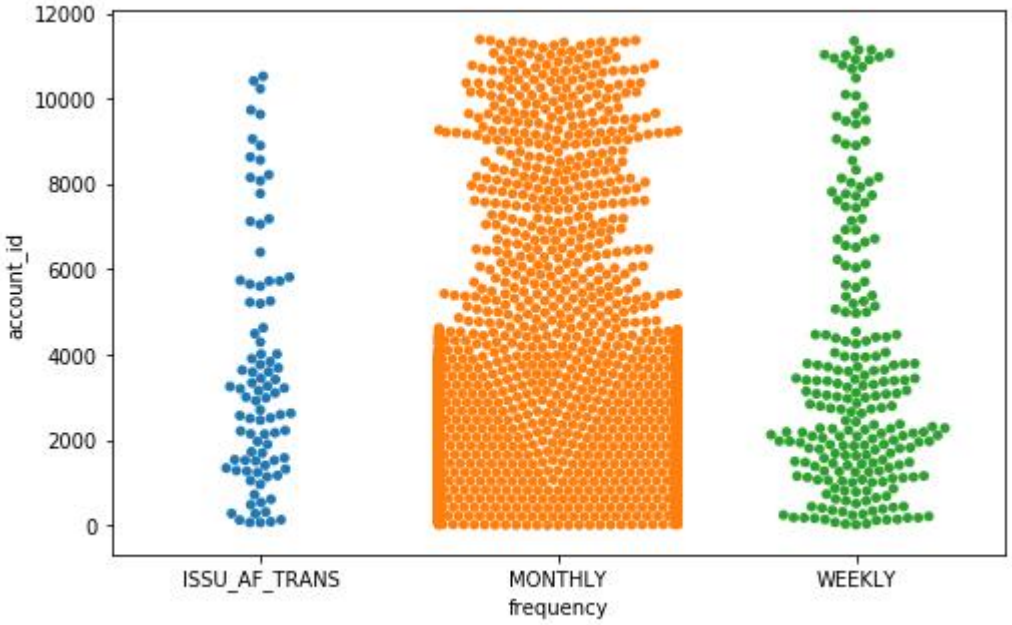

Orders
Numero de orders
3500
3000
2500
2000
1500
1000
500
0
household
unknow
loan_payt
ins_paymt
leasing
Category Order

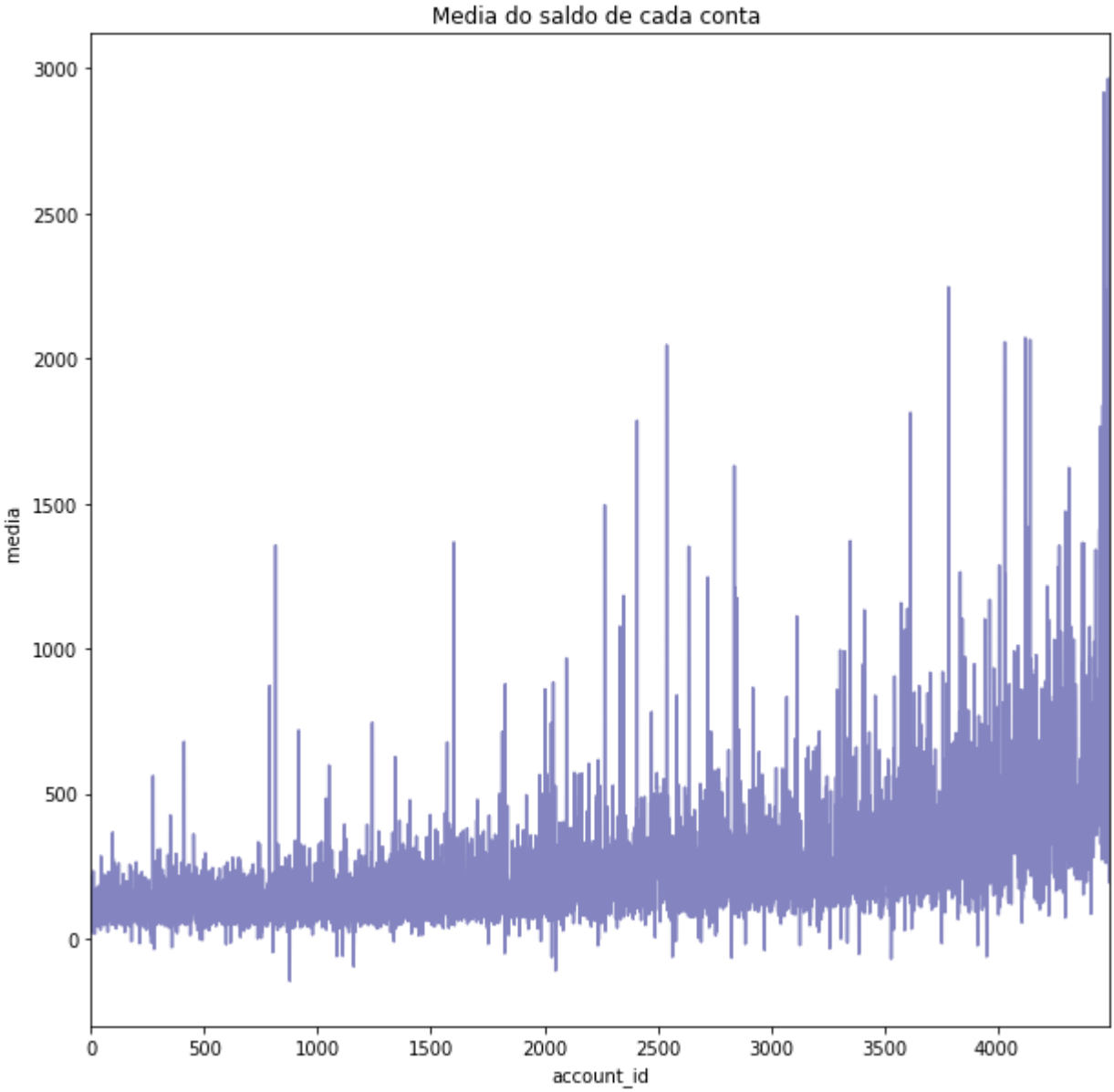

Cliente Idade
Frequency
Idade
0
2
4
6
8
10
12
14
16
0
20
40
60
80
100
120

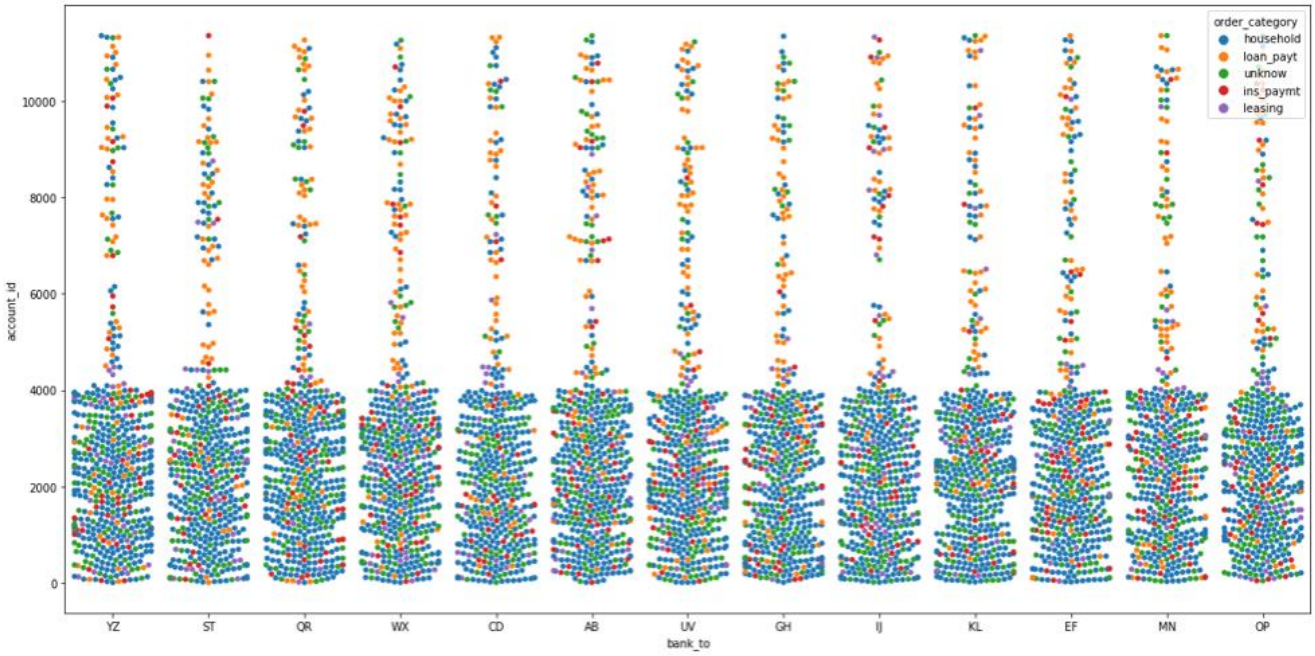

cliente Data de nascimento
Numero de cliente
data de nascimento
1
2
3
4
5
6
1920
1930
1940
1950
1960
1970
1980


Frequencia de Conta
Numero de Conta
4000
3500
3000
2500
2000
1500
1000
500
0
MONTHLY
WEEKLY
ISSU_AF_TRANS
Categoria de frequencia


Emissao do cartão
data de emissao de catão
8
7
6
5
4
3
2
1
0
1994
1995
1996
1997
1998


12000
10000
8000
6000
4000
2000
0
account_id
ISSU_AF_TRANS
MONTHLY
WEEKLY
frequency


Media do saldo de cada conta
media
account_id
3000
2500
2000
1500
1000
500
0
0
500
1000
1500
2000
2500
3000
3500
4000


order_category
household
loan_payt
unknow
ins_paymt
leasing
account_id
0
2000
4000
6000
8000
10000
YZ
ST
QR
WX
CD
AB
UV
GH
IJ
KL
EF
MN
OP
bank_to

Conclusão

Este trabalho foi interessante ver como muitas cadeiras estão interligadas entre si para alcançar melhorias nesta base de dados.

Tendo o objectivo fazer o aluno fazer um estudo independe para resolver o problema proposto pelo professor . Com intensão do aluno implementar um algoritmo de machine learning (k-means) .

Mesmo não tenha terminado todo o trabalho consegui ver o meus conhecimento adquirido em outras cadeiras não foram invão .

https://towardsdatascience.com/a-comparison-between-k-means-clustering-and-expectation-maximization-estimation-for-clustering-8c75a1193eb7