INE5644 - Data Mining

Pequena Entrega 3

Alunos:

Mariany Ferreira da Silva - 19200646

Eric de Col Sales - 16202587

Torben Castro Borba - 19100688

PEQUENA ATA

O grupo se reuniu no dia 16/04 e lá foi discutido sobre quais ferramentas poderíamos utilizar para realizar as análises e como poderíamos tratar os dados. A partir disso parte da análise exploratória foi dividida em 3 partes, cada membro do grupo se encarregou de tratar, analisar e desenvolver o dicionário de dados de um conjunto de propriedades do dataset. Após isso, houve uma discussão sobre os resultados obtidos, assim análises mais profundas e testes de hipóteses foram realizados com base nos insights obtidos previamente.

FERRAMENTAS AVALIADAS

- **Python** é uma linguagem de programação de alto nível, interpretada, dinamicamente tipada, orientada a objetos e de propósito geral. Ela é amplamente utilizada na área de ciência de dados, dado que uma ampla biblioteca padrão e uma grande comunidade de desenvolvedores que contribuem com pacotes e bibliotecas adicionais. Além disso, possui sintaxe simples e intuitiva que facilita um desenvolvimento ágil importantíssimo para a ciência de dados.
- **Jupyter Notebook** é uma aplicação web que permite criar e compartilhar documentos interativos que contêm código executável, visualizações, texto e outros elementos. Ele suporta várias linguagens de programação, incluindo Python, R e Julia. Os notebooks Jupyter são organizados em células, onde cada célula pode conter texto formatado, código e saída do código. O Jupyter Notebook é amplamente utilizado em ciência de dados, dado que ele facilita a exploração, a visualização e os testes com os dados.
- **Pandas** é uma biblioteca do Python para análise de dados. Ele oferece estruturas de dados flexíveis e de alto desempenho para manipulação e análise de dados, com recursos de indexação e agregação de dados. Ele permite carregar, limpar, transformar e analisar dados de diversas fontes, incluindo arquivos CSV, bancos de dados SQL e Excel.
- **Numpy** é uma biblioteca do Python para computação científica. Ela oferece suporte para arrays e matrizes multidimensionais, bem como funções matemáticas para manipulação desses arrays. Pelo seu desempenho, ela já é utilizada em outras bibliotecas de análise de dados, como o Pandas e o Matplotlib.
- **Matplotlib** é uma biblioteca de visualização de dados em Python que permite a criação de gráficos e plots de alta qualidade de forma simples e personalizável.
- **Missingno** é uma biblioteca Python que fornece a capacidade de entender a distribuição de valores ausentes por meio de visualizações informativas. As visualizações podem ser na forma de mapas de calor.
- **Seaborn** é uma biblioteca de visualização de dados Python baseada em matplotlib. Ele fornece uma interface de alto nível para desenhar gráficos estatísticos atraentes e informativos.
- **Deepnote** é um notebook colaborativo de ciência de dados que ajuda as equipes a descobrir, entender e compartilhar suas descobertas com qualquer pessoa. É

compatível com **Jupyter Notebook**, é executado na nuvem e funciona com qualquer pilha de dados.

Utilizaremos editores csv em conjunto com nosso dataset pois é onde podemos ter uma visualização primária dos dados, além da facilidade de alterar os tipos dos dados de uma coluna, como números em string para float.

QUESTÕES LEVANTADAS PARA DISCUSSÃO

- Quais são as variáveis e o que elas significam para o nosso cliente?
- Quais são as variáveis com mais dados faltantes no dataset?
- Há alguma relação entre saldo na conta e o valor do empréstimo?
- Há mais empréstimos aprovados para pessoas com maior ou menor saldo?

ANÁLISES REALIZADAS

Quais são as variáveis e o que elas significam para o nosso cliente?

A fim de interpretar e utilizar nossos dados de forma eficaz, criamos um dicionário de dados com as características básicas de cada variável.

Nome do campo no dataset	Descrição do campo determinada pelo cliente	Tipo de variável				
		Tipo	Atributo	Escala dos dados	% Dados faltantes	Afeta variável target?
ID	Número de identificação do cliente	INT	DISCRETO	NOMINAL	0.0	
CHECKING_BALANCE	Saldo em conta corrente do cliente	INT	CONTÍNUO	RACIONAL	63,02	Necessita investigação mais aprofundada
PAYMENT_TERM	Número de dias que o cliente tem para pagar o empréstimo	INT	CONTINUO	INTERVALAR	0,6	sim, inversamente proporcional
CREDIT_HISTORY	Como está a situação de crédito passada do cliente [ALL_CREDITS_PAID_BACK,	TEXTO	CATEGÓRICO	NOMINAL	38,43	Sim, mas necessita de mais investigações

Nome do campo no dataset	Descrição do campo determinada pelo cliente	Tipo de variável				
		Tipo	Atributo	Escala dos dados	% Dados faltantes	Afeta variável target?
	CREDITS_PAID_TO_DATE, PRIOR_PAYMENTS_DELAYED, OUTSTANDING_CREDIT, NO_CREDITS]					
LOAN_PURPOSE	Motivação do empréstimo [CAR_NEW, CAR_USED, FURNITURE, BUSINESS, APPLIANCES, RADIO_TV, EDUCATION, VACATION, REPAIRS, RETRAINING, OTHER]	TEXTO	CATEGÓRICO	NOMINAL	0,5	Sim
LOAN_AMOUNT	Valor do empréstimo	INT	DISCRETO	RACIONAL	0,45	sim, inversamente proporcional
EXISTING_SAVINGS	Saldo de conta poupança	INT	DISCRETO	RACIONAL	42.71	não
EMPLOYMENT_DURATION	Quantos anos o cliente está no último emprego	INT	DISCRETO	INTERVALAR	1.9	não
INSTALLMENT_PERCENT	Em quantas parcelas o empréstimo deve ser pago	INT [1-6]	DISCRETO	INTERVALAR	0.62	Sim, inversamente proporcional
SEX	Sexo [M, F]	TEXTO	CATEGÓRICO	NOMINAL	1.97	Não
OTHERS_ON_LOAN	Se existe um fiador ou outro aplicante para o empréstimo [NONE, CO-APPLICANT, GUARANTOR]	TEXTO	CATEGÓRICO	NOMINAL	0.55	Sim
CURRENT_RESIDENCE_DU	Anos em que o cliente	INT	DISCRETO	INTERVALAR	2.05	Sim,

Nome do campo no dataset	Descrição do campo determinada pelo cliente	Tipo de variável				
		Tipo	Atributo	Escala dos dados	% Dados faltantes	Afeta variável target?
RATION	está vivendo na última casa	[1-6]				inversamente proporcional
PROPERTY	Se o cliente possui alguma propriedade em nome [SAVINGS_INSURANCE, REAL_ESTATE,CAR_OTHER, UNKNOWN]	TEXTO	CATEGÓRICO	NOMINAL	2.02	Sim
AGE	Idade	INT [19-74]	DISCRETO	INTERVALAR	2.02	Sim, inversamente proporcional
INSTALLMENT_PLANS	Plano de financiamento, podendo ser do banco, externo, ou nenhum [NONE, STORES, BANK]	TEXTO	CATEGÓRICO	NOMINAL	0.52	Sim
HOUSING	Tem casa própria ou não: FREE - Moradia grátis OWN - Casa própria RENT - Aluguel	TEXTO	CATEGÓRICO	NOMINAL	2.17	Sim
EXISTING_CREDITS_COUNT	Número de empréstimos que o cliente já tem	INT [1-3]	DISCRETO	INTERVALAR	38.40	Sim, inversamente proporcional
JOB_TYPE	Tipo de emprego: 0 - desempregado 1 - Não qualificado 2 - Autônomo 3 - Qualificado	INT [0-3]	CATEGÓRICO	NOMINAL	2.22	Sim
DEPENDENTS	Número de pessoas com acesso à conta	INT [1-2]	DISCRETO	INTERVALAR	2.10	Sim, inversamente proporcional

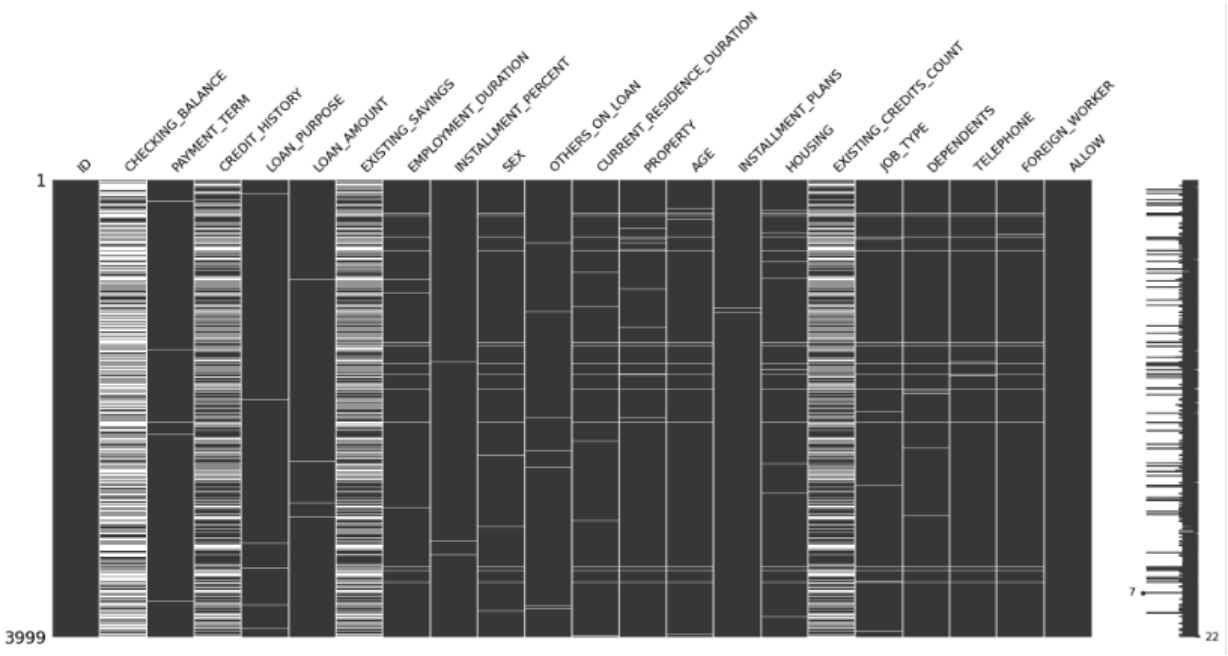
Nome do campo no dataset	Descrição do campo determinada pelo cliente	Tipo de variável				
		Tipo	Atributo	Escala dos dados	% Dados faltantes	Afeta variável target?
TELEPHONE	Se o cliente tem telefone cadastrado ou não	INT [0-1]	BINÁRIO	NOMINAL	2.10	Sim, inversamente proporcional
FOREIGN_WORKER	Se o cliente trabalha num país externo ao do banco ou não	INT [0-1]	BINÁRIO	NOMINAL	2.25	Sim
ALLOW	Se o empréstimo deve ser realizado para o cliente	INT [0-1]	BINÁRIO	NOMINAL		

Quais são as variáveis com mais dados faltantes no dataset?

63.02 % CHECKING_BALANCE: Saldo em conta corrente do cliente
38.43 % CREDIT_HISTORY: Como está a situação de crédito passada do cliente
42.71 % EXISTING_SAVINGS: Saldo de conta poupança
38.40 % EXISTING_CREDITS_COUNT: Número de empréstimos que o cliente já tem

Sabendo que essas variáveis possuem um grande número de dados faltantes, focaremos nossas primeiras análises nas demais variáveis do dataset.

Exemplo de código (utilização da biblioteca Missingno)



Quão balanceado está o dataset?

Executando a função *value_counts* disponível na biblioteca Pandas vimos que temos mais registros onde o empréstimo foi concedido do que registros onde o empréstimo não foi concedido.

Precisaremos balancear o dataset quando realizarmos o treinamento e análise dos modelos.

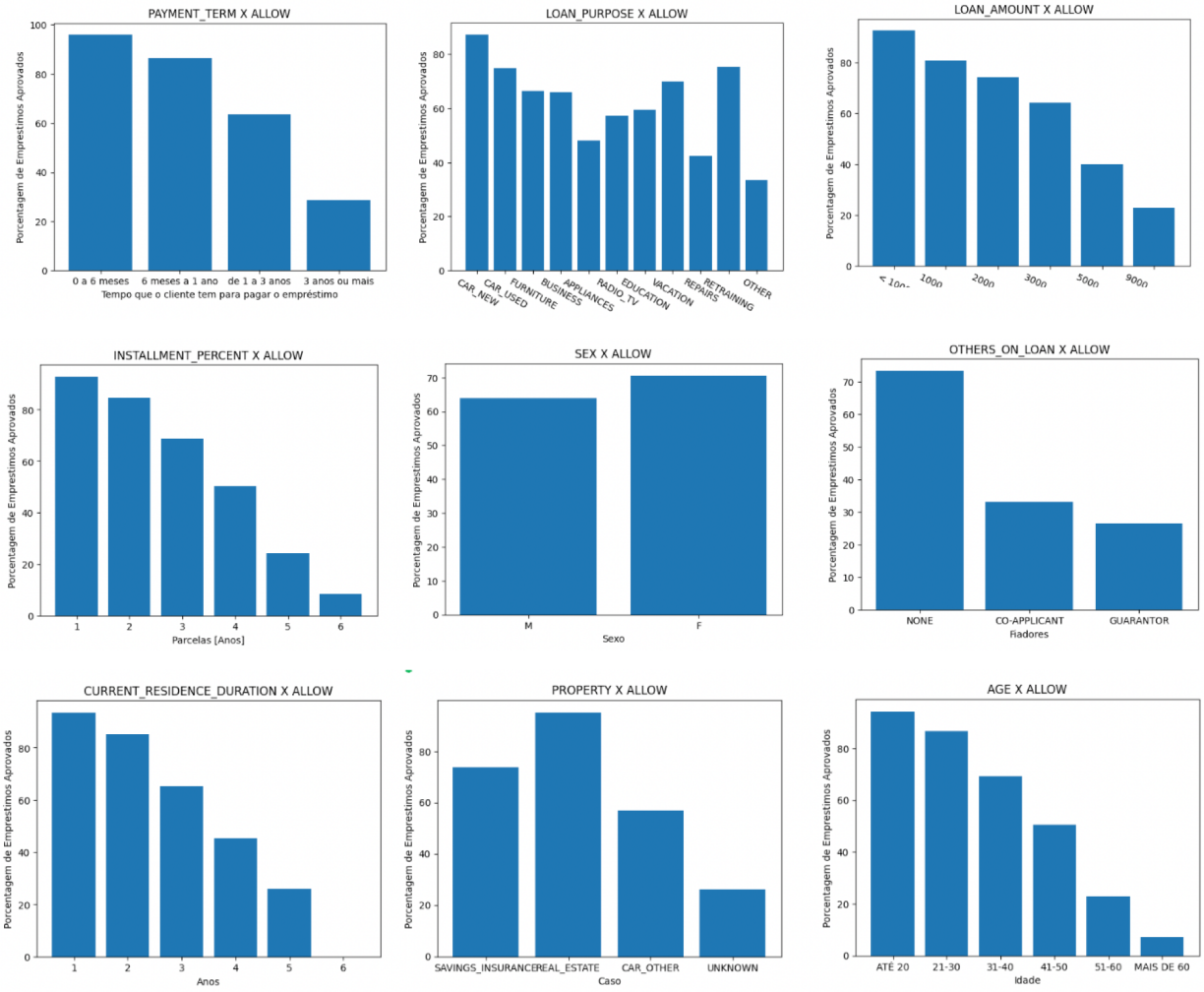
```
df_all['ALLOW'].value_counts()                                    [216]

1    2656
0    1343
Name: ALLOW, dtype: int64
```

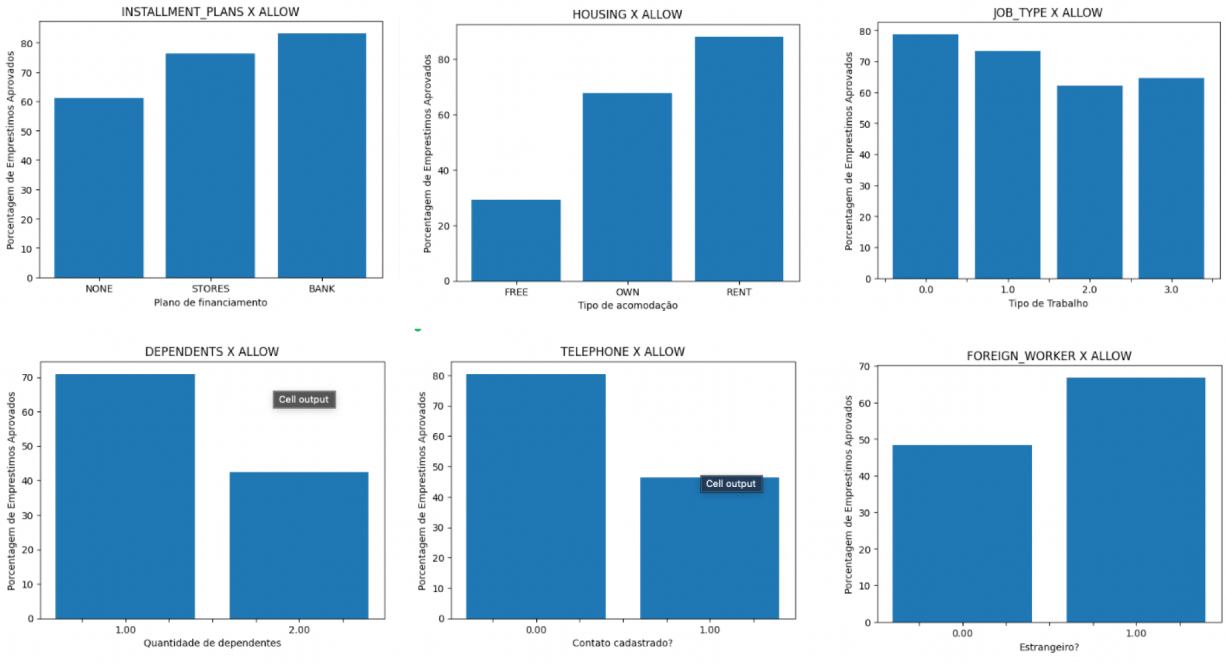
Exemplo de código (utilização da biblioteca Pandas)

Qual a correlação e a porcentagem de empréstimos quando as variáveis assumem os diferentes valores possíveis?

```
[8]
x = ['0 a 6 meses', '6 meses a 1 ano', 'de 1 a 3 anos', '3 anos ou mais',]
y = []
for i,j in [(0,182),(183,364), (365,1095), (1096,1984)]:
    aux_aprov = df_all.query(f"{i} < PAYMENT_TERM < {j}")["ALLOW"].value_counts()[1] # numero de aprovados
    aux_neg = df_all.query(f"{i} < PAYMENT_TERM < {j}")["ALLOW"].value_counts()[0] # numero de negados
    y.append(aux_aprov/(aux_aprov + aux_neg) * 100)
plt.bar(x, y)
plt.title("PAYMENT_TERM X ALLOW")
plt.xlabel("Tempo que o cliente tem para pagar o empréstimo")
plt.ylabel("Porcentagem de Emprestimos Aprovados")

plt.show()
```

Exemplo de código (utilização das bibliotecas Pandas e Matplotlib)

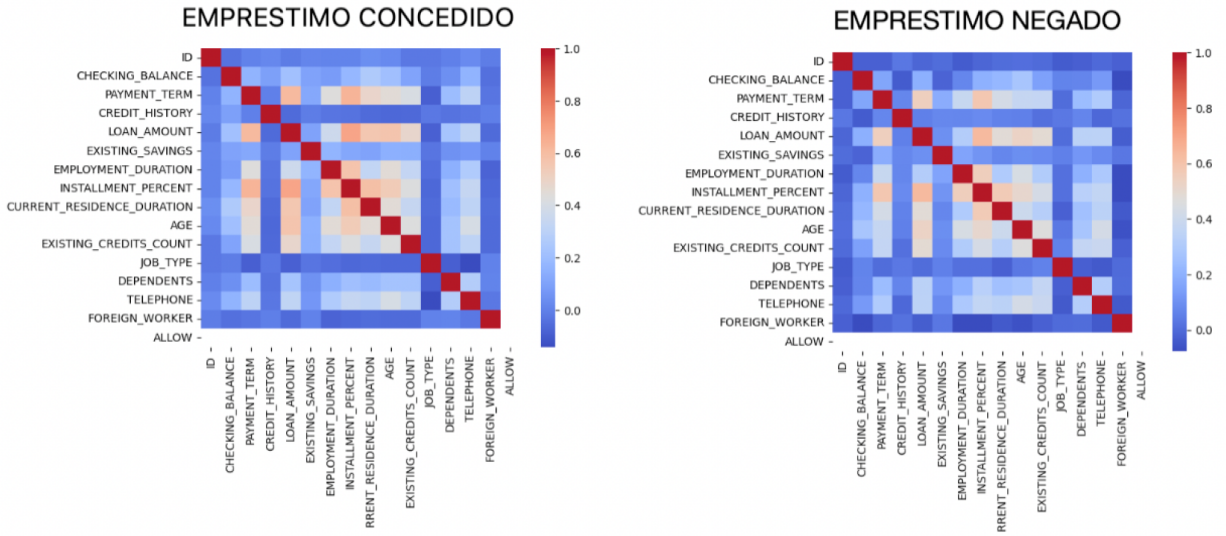




Matriz de correlação entre a variável target e as demais variáveis:

```
[20]
df_allow1 = df_all[df_all["ALLOW"] == 1]

# criar a matriz de correlação
corr_matrix = df_allow1.corr()

# plotar a matriz de correlação
sns.heatmap(corr_matrix, annot=False, cmap="coolwarm")
```

Exemplo de código (utilização das bibliotecas Pandas e Seaborn)

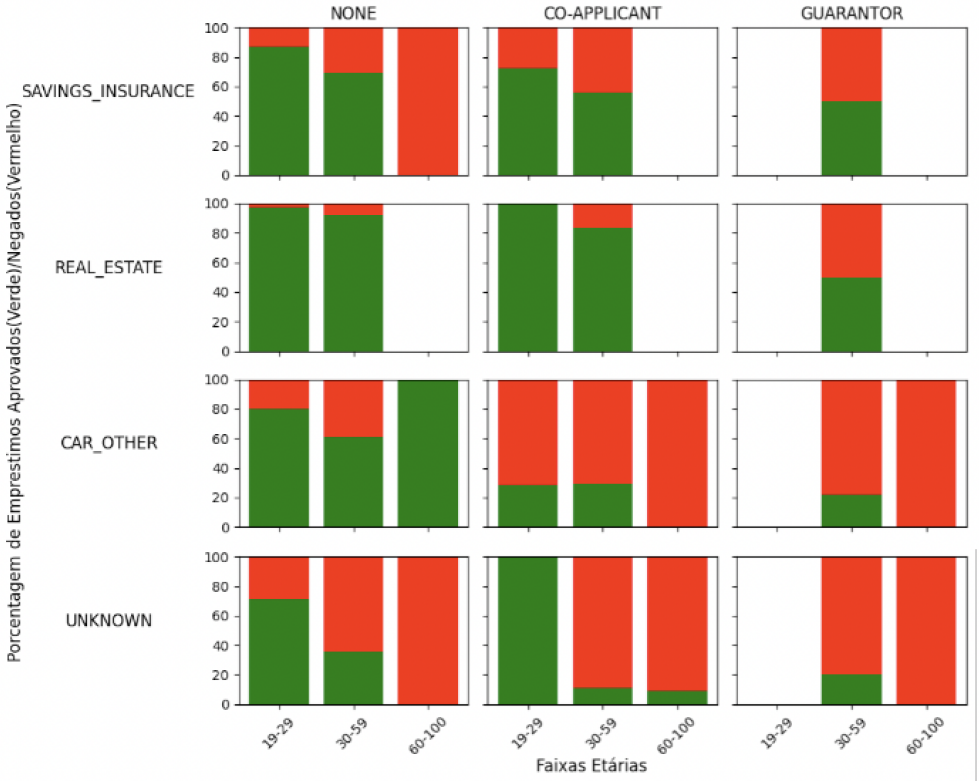


Pela fórmula a correlação é um número entre -1 e 1 e interpretamos da seguinte maneira: quanto mais o r estiver próximo de 1 ou -1 mais forte será a correlação.

Correlação positiva: ambas as variáveis mudam na mesma direção.
Correlação Nula: Nenhuma relação na mudança das variáveis.
Correlação Negativa: as variáveis mudam em direções opostas.

Pessoas jovens com propriedades ou/e fiador têm mais empréstimos aprovados ?

A partir da figura abaixo, observa-se que a existência de um imóvel no nome do cliente é a melhor garantia para aprovação do empréstimo. Entretanto, a presença de um fiador, ou um outro aplicante, não gera vantagens na aprovação do empréstimo.

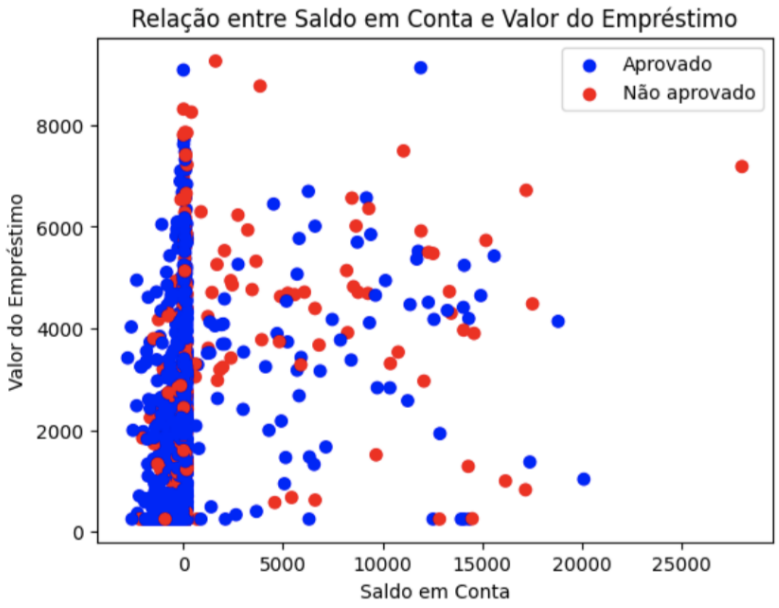


Há alguma relação entre saldo na conta e o valor do empréstimo?

Fica evidente que não há uma correlação entre o saldo na conta e o valor do empréstimo, ao mesmo tempo que o gráfico de dispersão nos mostra que a maior parte dos empréstimos está concentrado em pessoas que possuem pouco saldo na conta ou já estão devendo.

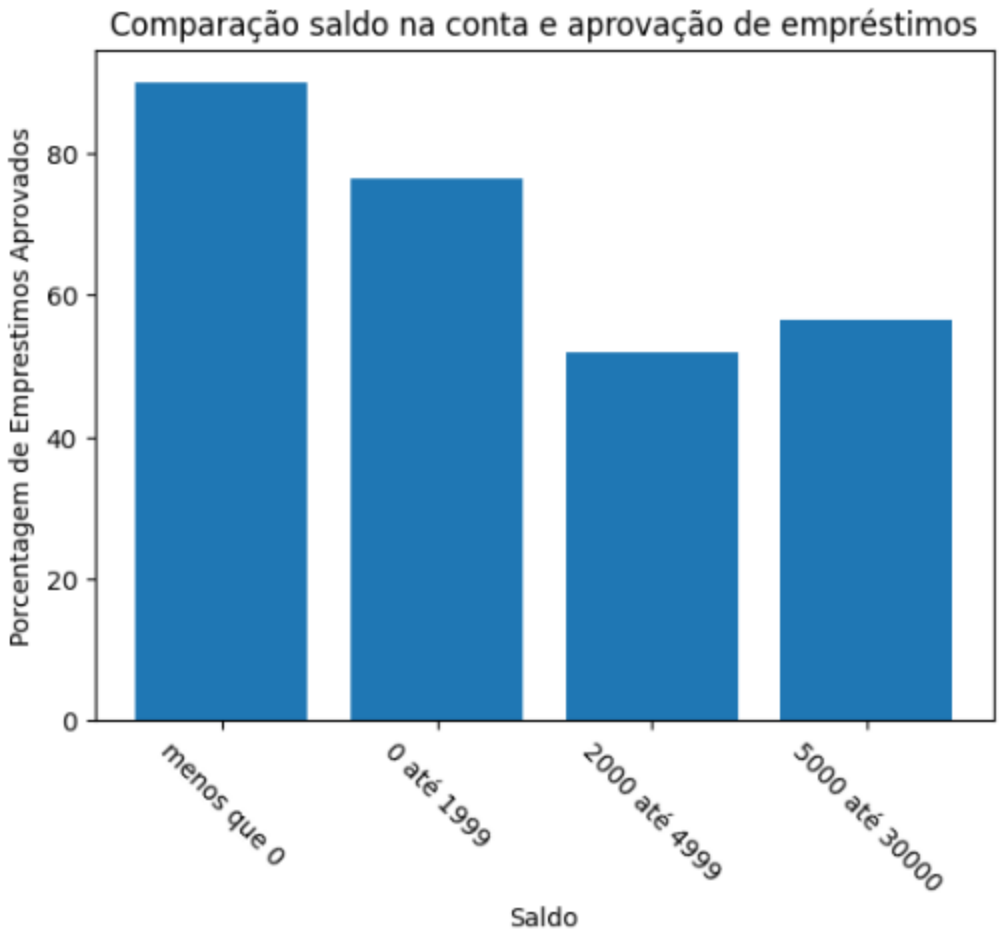
[51]

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

fig, ax = plt.subplots()

x = df_all['CHECKING_BALANCE']
y = df_all['LOAN_AMOUNT']

colors = np.where(df_all['ALLOW'] == 1, 'blue', 'red') # lista de cores

ax.scatter(x, y, c=colors)

ax.set_xlabel('Saldo em Conta')
ax.set_ylabel('Valor do Empréstimo')
ax.set_title('Relação entre Saldo em Conta e Valor do Empréstimo')

blue_patch = plt.scatter([],[], c='blue', label='Aprovado')
red_patch = plt.scatter([],[], c='red', label='Não aprovado')
ax.legend(handles=[blue_patch, red_patch], loc='best')


plt.show()
```



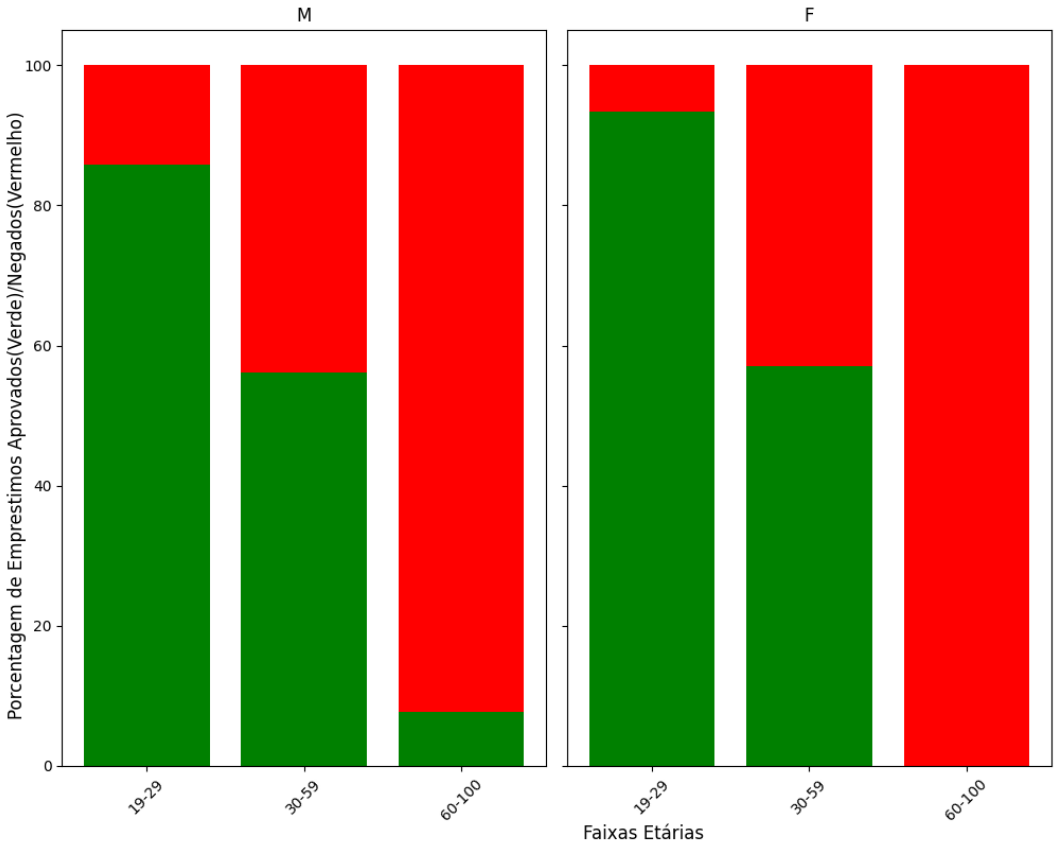
Acima utilizamos o gráfico de dispersão, testando a hipótese de que há uma relação entre o saldo na conta e o valor do empréstimo. Onde é utilizado o Matplotlib e Numpy para realizar análises do Dataframe que foi criado com o Pandas.

Há mais empréstimos aprovados para pessoas com maior ou menor saldo em conta?



O sexo em conjunto com a idade afeta na aprovação dos empréstimos ?

A partir da figura abaixo, observa-se que há um decaimento na aprovação de empréstimos de acordo com a idade do cliente. Entretanto, não pode se afirmar que existe uma relação entre idade e sexo com a aprovação de um empréstimo, nem uma relação entre sexo e aprovação de empréstimos.



REFERÊNCIAS

PYTHON SOFTWARE FOUNDATION. The Python Standard Library. DISPONÍVEL EM: https://docs.python.org/3/library/index.html Acesso em 16 ABR. 2023

JUPYTER TEAM, The Jupyter Notebook. Disponível em: https://jupyter-notebook.readthedocs.io/en/stable/index.html Acesso em 16 ABR. 2023

NUMFOCUS, Inc. Pandas Documentation. Disponível em: https://pandas.pydata.org/docs/ Acesso em 16 ABR. 2023

NUMPY DEVELOPERS. NumPy user guide. Disponível em: https://numpy.org/devdocs/user/index.html Acesso em 16 ABR. 2023

JOHN HUNTER, DARREN DALE, ERIC FIRING, MICHAEL DROETTBOOM AND THE MATPLOTLIB DEVELOPMENT TEAM. Matplotlib 3.7.1 documentation. Disponível em: https://matplotlib.org/stable/index.html Acesso em 16 ABR. 2023

MADHUKAR, Bhoomika. Tutorial On Missingno – Python Tool To Visualize Missing Values Disponível em: https://analyticsindiamag.com/tutorial-on-missingno-python-tool-to-visualize-missing-values/#:~:text=Missingno%20is%20a%20Python%20library,heat%20maps%20or%20bar%20charts Acesso em 16 ABR. 2023

WASKOM, Michael. seaborn: statistical data visualization. Disponível em: https://seaborn.pydata.org/ Acesso em 16 ABR. 2023

DEEPNOTE. Notebooks: a better way for teams to work with data. Disponível em: https://deepnote.com/ Acesso em 16 ABR. 2023