Documentação da Análise Exploratória

Entendimento do Dataset

Tecnologias

Utilizei durante o projeto todo **Python** para a análise de dados. Principalmente a biblioteca **pandas** para a manipulação dos dados, **seaborn** e **matplotlib** para criação de gráficos e **numpy** para algumas funções matemáticas especificas.

Limpeza

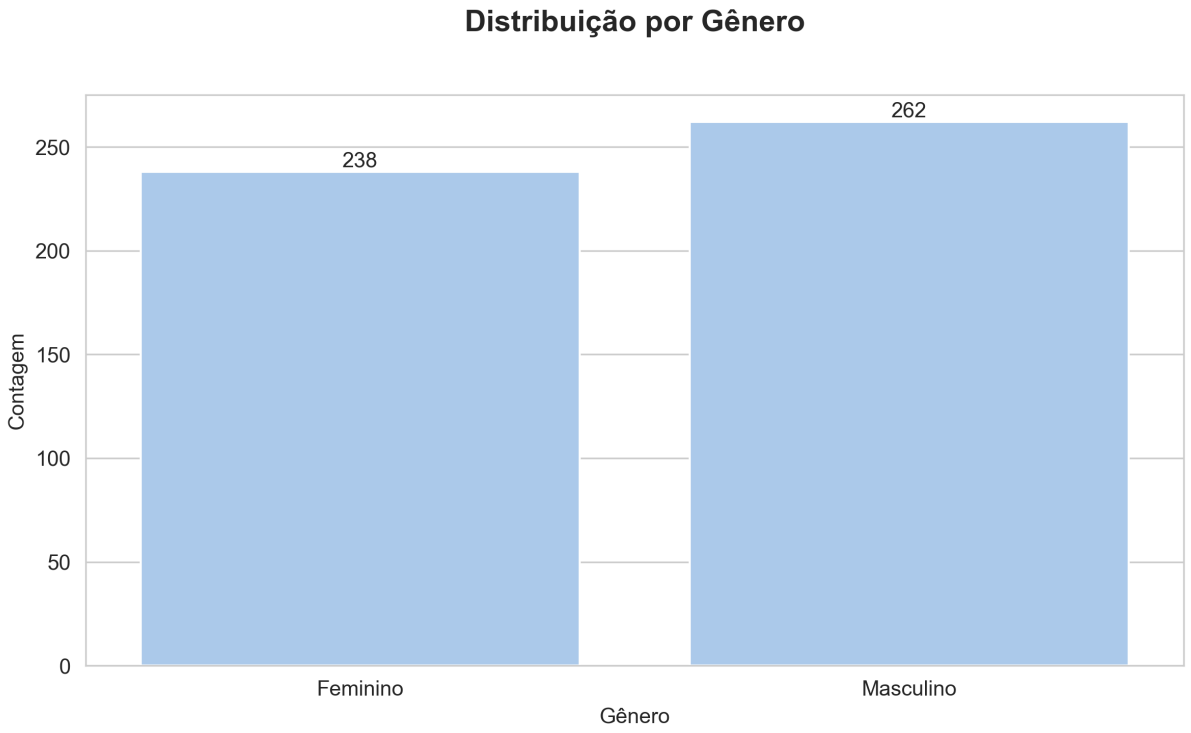
Comecei olhando se não haviam valores nulos e não havia. Depois me certifiquei que as colunas eram do tipo correto (se uma coluna numérica era do tipo inteiro por exemplo). Depois me certifiquei que não haviam IDs duplicados, o que significaria que a base tem pessoas repetidas ou algum erro. Por fim, verifiquei os Outliers com boxplot, não encontrei nada anormal. Como não havia nada disso, os dados já estavam prontos para ser trabalhados.

Análise

Análise de Clientes

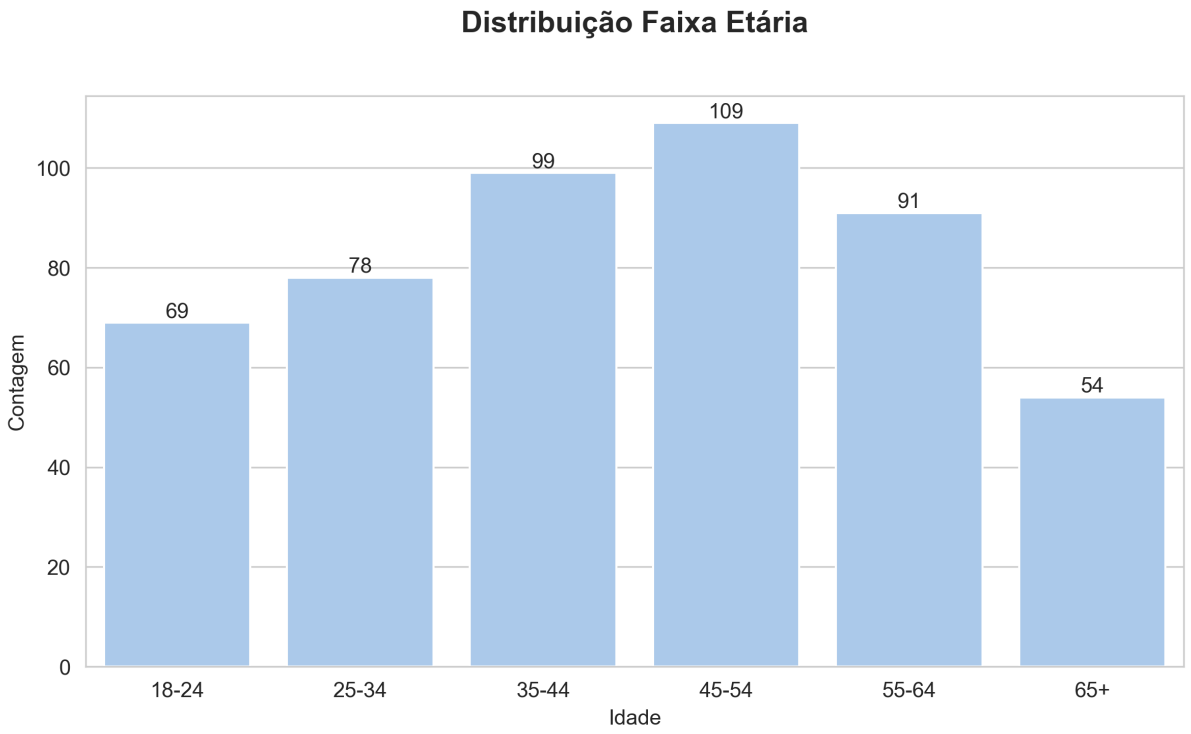
Qual é a distribuição dos clientes por gênero, faixa etária e faixa de renda?

gênero: Homens financiam um pouco mais



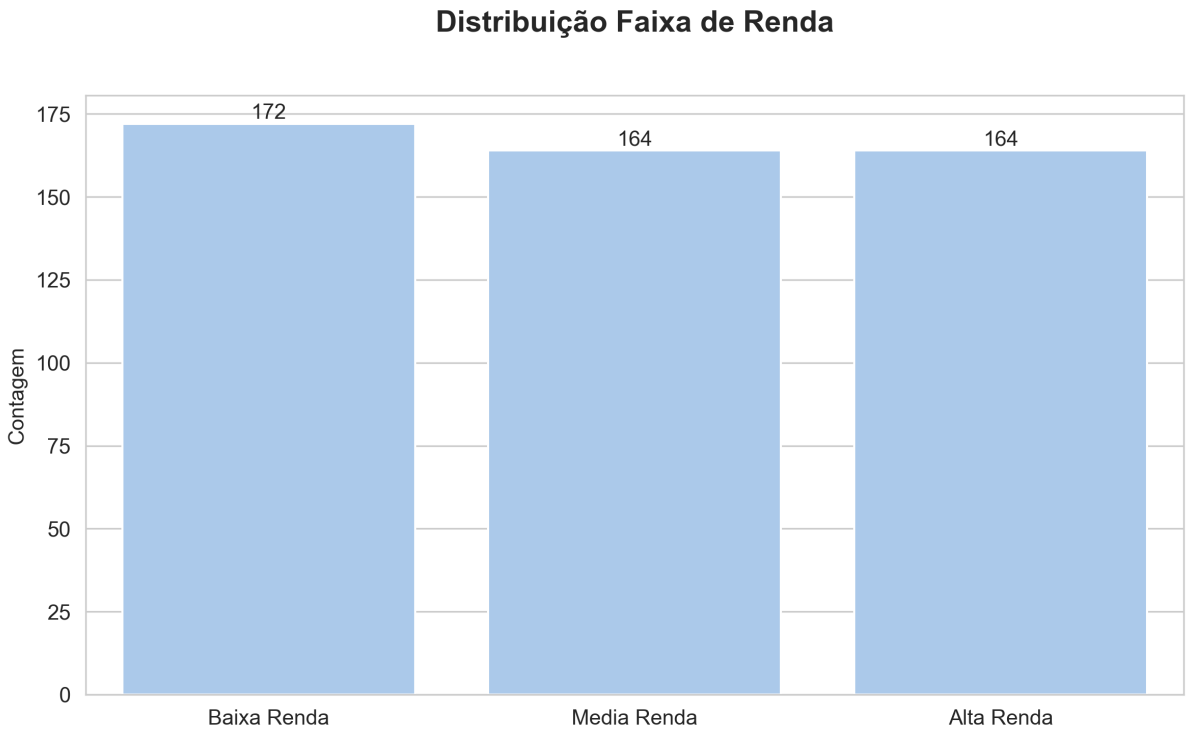
Detalhe: Usei uma contagem com pandas e fiz um gráfico de barras.

Faixa Etária: O maior volume de financiamento se concentra na média idade, formando uma espécie de gráfico em U. Onde os mais novos e mais velhos tem menos financiamentos. Os financiamentos se concentram um pouco mais entre 35 e 64 anos.



Detalhe: Decidi fazer uma faixas de idade para facilitar a análise. Fiz uma função que categoriza a idade a sua respectiva faixa de idade e apliquei em uma nova coluna ao data frame.

Faixa de Renda: Sem distinção relevante



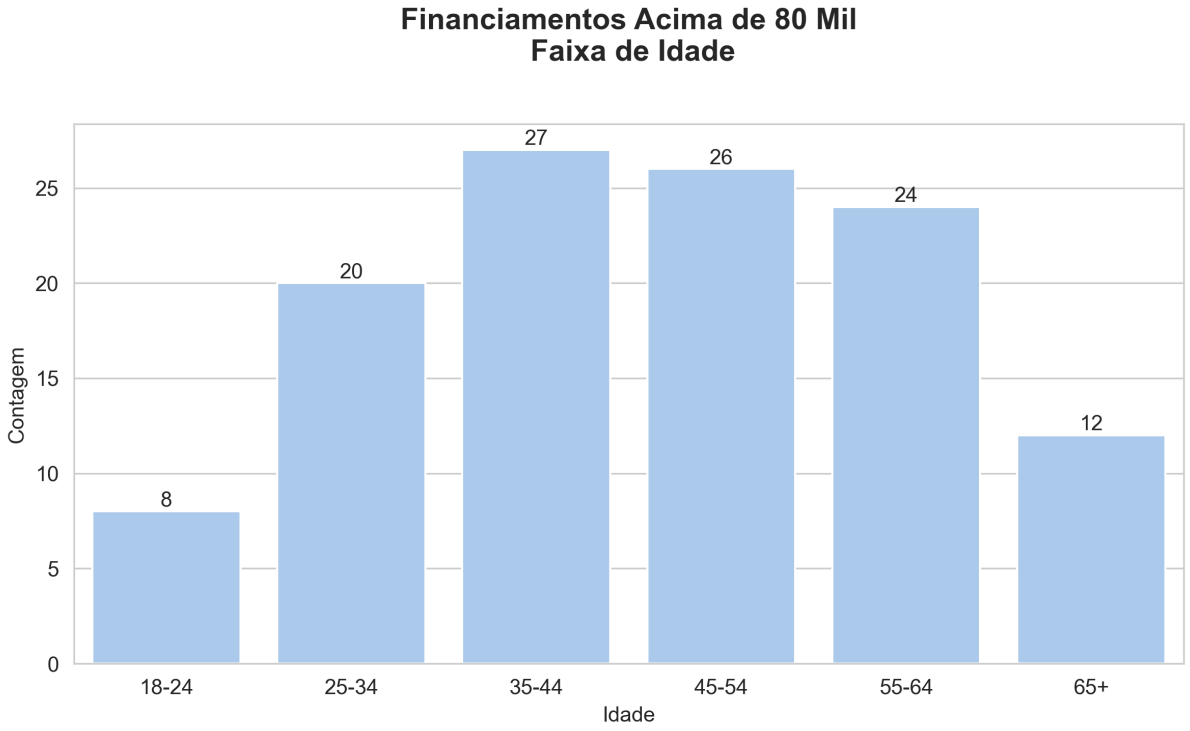
Detalhe: Usei uma contagem com pandas e fiz um gráfico de barras.

Características dos clientes que tomam empréstimos maiores

Os parâmetros que eu vou avaliar são:
Faixa_Idade; Genero; Regiao; Faixa_de_Renda;

parâmetro final é **Valor_Financia**

Faixa_Idade: Na média quem tem entre 18 e 24 faz financiamentos um pouco menores. Mas, quem tem 18 e 24 anos e mais de 65 anos, faz poucos financiamentos acima de 80 mil. (80 mil é um corte dos 25% dos financiamentos mais caros)



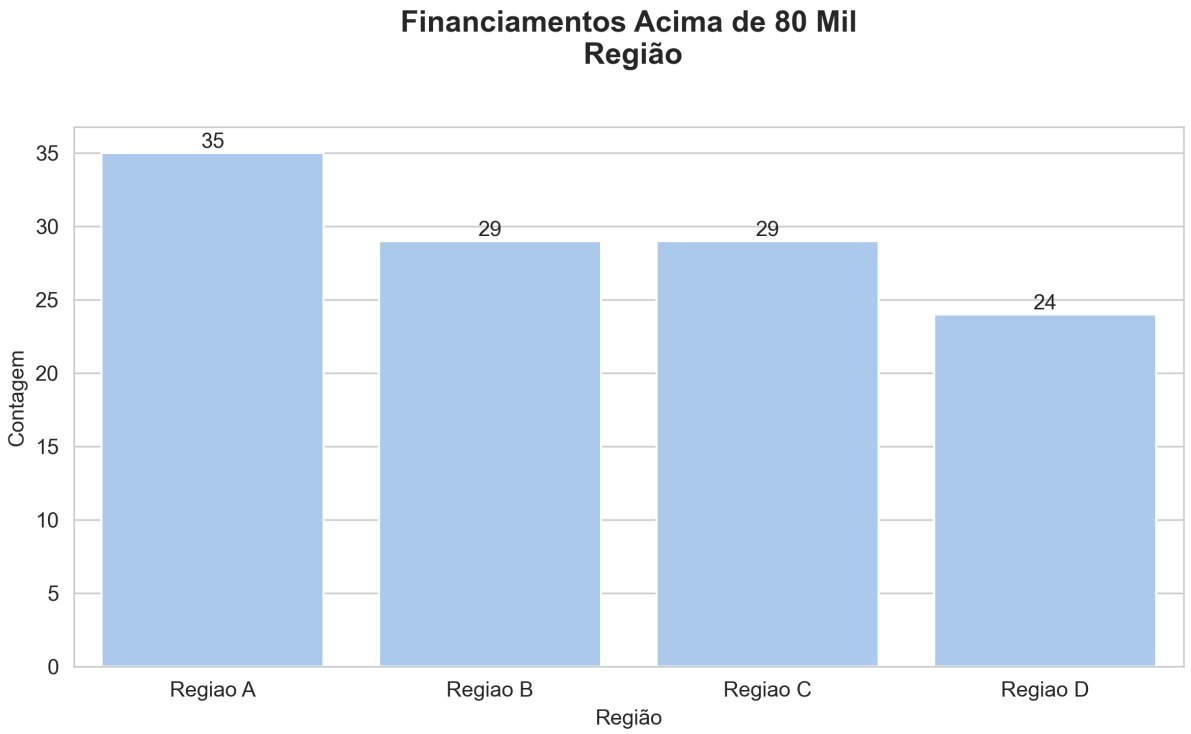
Detalhe: Agrupei as faixas de idade olhando pelo Valor_Financiado. fiz uma tabela com a média, mediana, desvio padrão, mínimo, máximo e quantidade. Depois fiz um boxplot, percebi que o ultimo quartil era acima de 80 mil, e fiz uma análise que com gráfico de barras com o filtro de 80 mil, para ver qual perfil faz empréstimos maiores.

Gênero: Mulheres fazem empréstimos levemente mais baixo.

Genero	mean	median	std	min	max
Feminino	53985.53	53496.5	26532.26	10281	99393
Masculino	56067.61	55756.5	25415.98	10526	99899

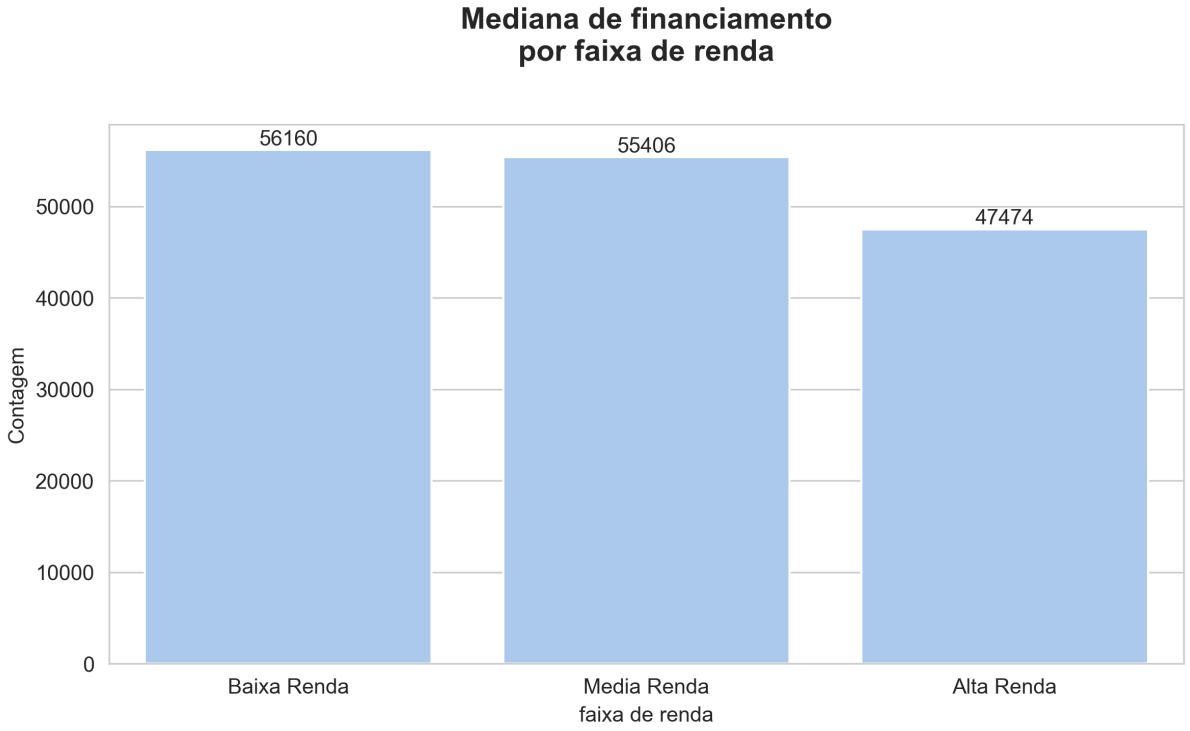
Detalhe: Agrupei e fiz uma tabela com estatísticas gráfico de barra, boxplot e análise acima dos 80 mil.

Região: Quanto mais alto a faixa de renda da região, maior a chance de fazer empréstimos mais altos.



Detalhe: Agrupei e fiz uma tabela com estatísticas. Fiz gráfico de barras com a mediana, boxplot e gráfico de barras para financiamentos de 80 mil.

Faixa de Renda: Quanto mais baixo é a faixa de renda, mais alta é o financiamento em média.



Detalhe: Agrupei e fiz uma tabela com estatísticas. Fiz gráfico de barras com a mediana, boxplot e gráfico de barras para financiamentos acima de 80 mil.

Análise de Financiamentos e Veículos

Quais são as regiões com maior volume de financiamentos?

Quanto maior a faixa de renda da região mais volume de financiamentos.

Há diferenças significativas nas taxas de inadimplência entre regiões com diferentes rendas médias?

Quanto maior a renda média da região, menor a taxa de Inadimplência.

Detalhe: Agrupei e fiz um gráfico de barras com a média.

Qual o perfil dos veículos financiados (valor médio dos veículos, modelo e ano mais financiados)?

Não tem uma relação entre os veículos de anos mais recentes serem mais financiados. No entanto, 2011, 2013 e entre 2016 e 2020 são anos com menos financiamentos que os demais. Sobre os valores:

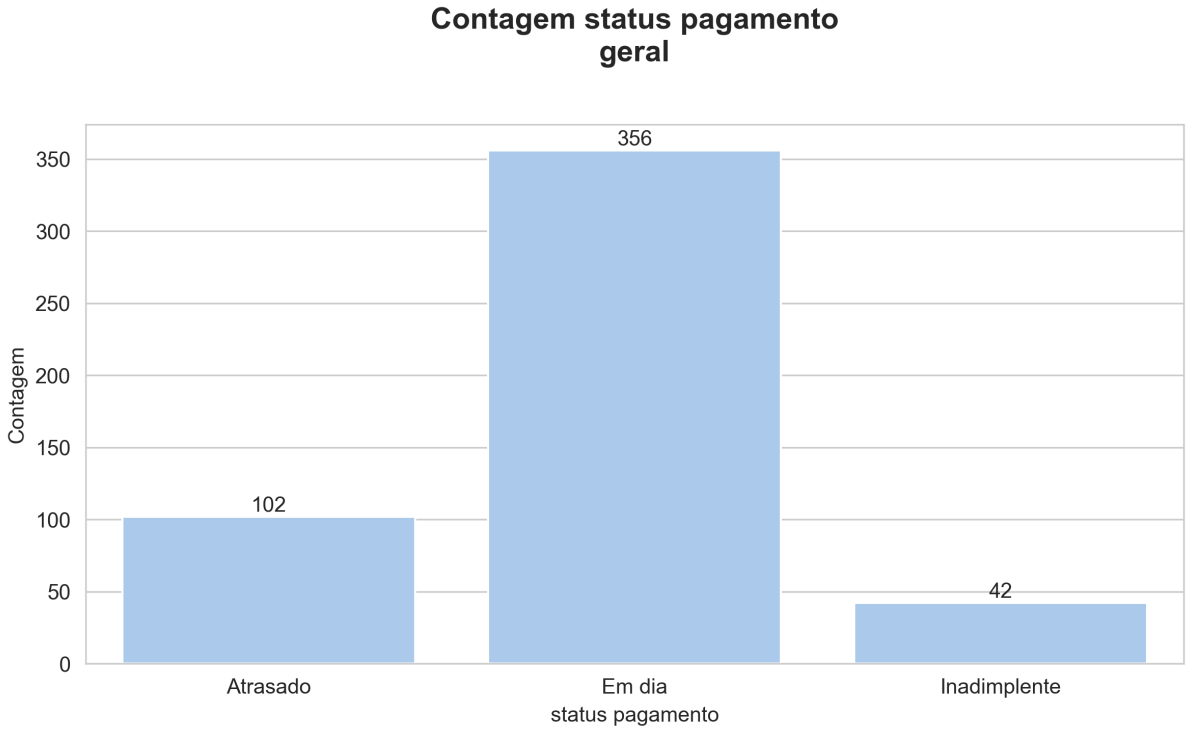
```
min     11.845
25%     46.155
50%     79.670
75%    115.524
max    234.517

Média: 84.813
```

Detalhe: Agrupei e fiz uma tabela com estatísticas. Fiz um gráfico de barras para ver quais são os modelos mais financiados. Fiz um boxplot para ver a média de valores de forma mais detalhada.

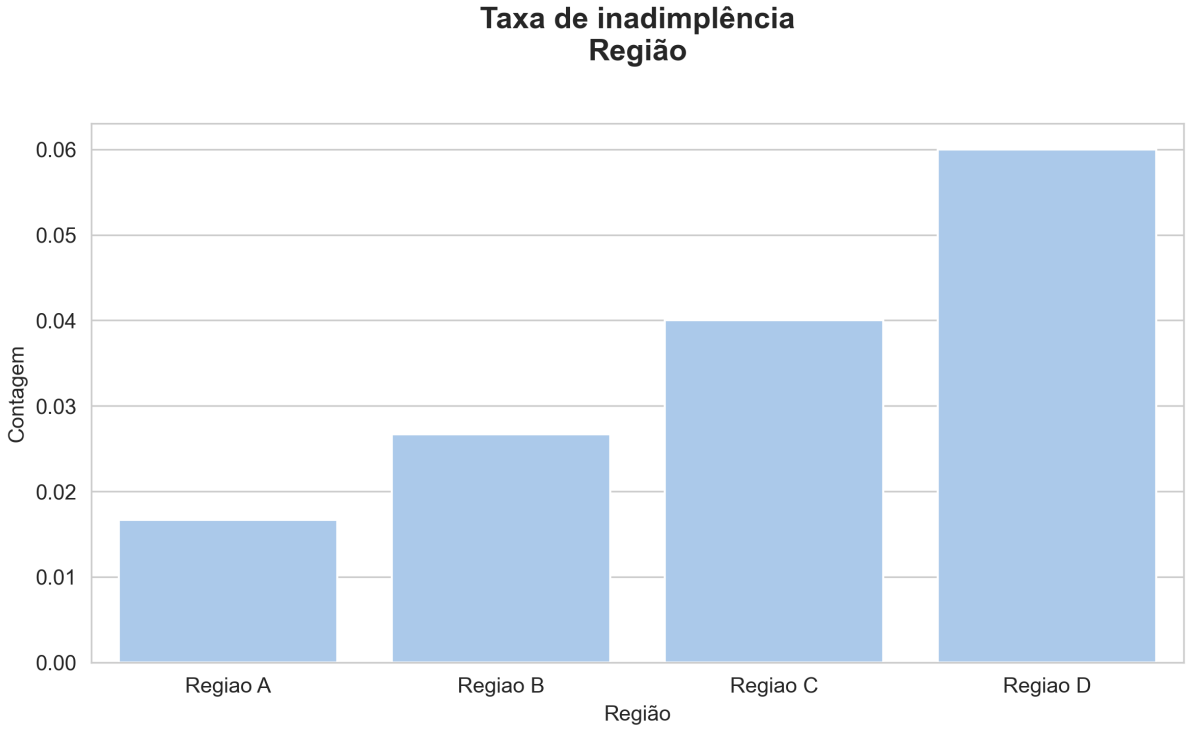
Análise de Inadimplência

Qual é a taxa de inadimplência geral e como ela varia de acordo com a renda da região e as características dos clientes?

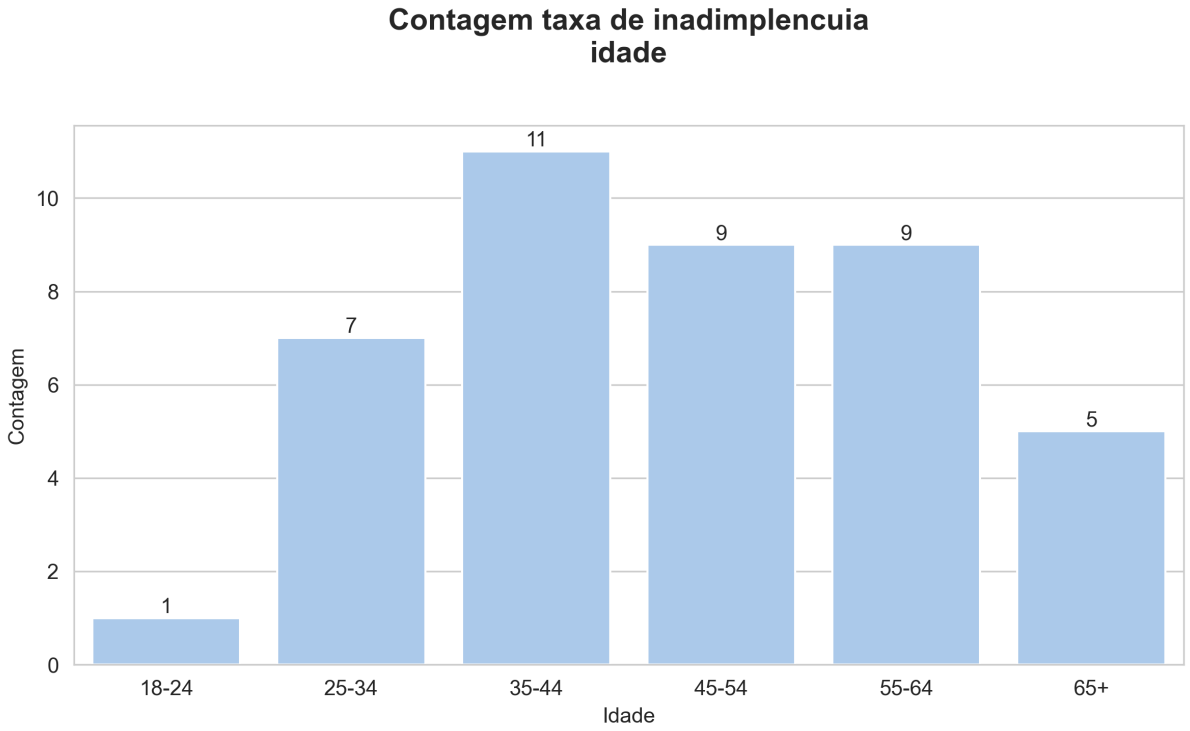
Taxa de inadimplência geral:

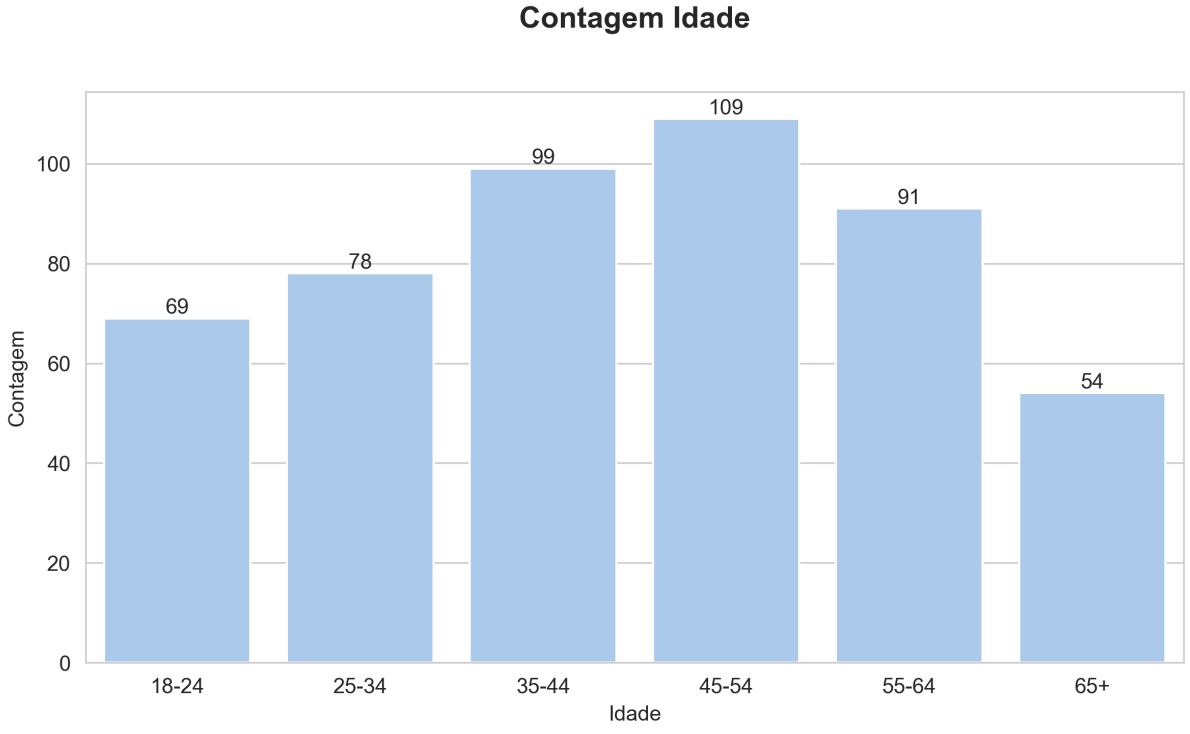


Taxa de inadimplência região: Quanto menor a renda da região maior a taxa de inadimplência.



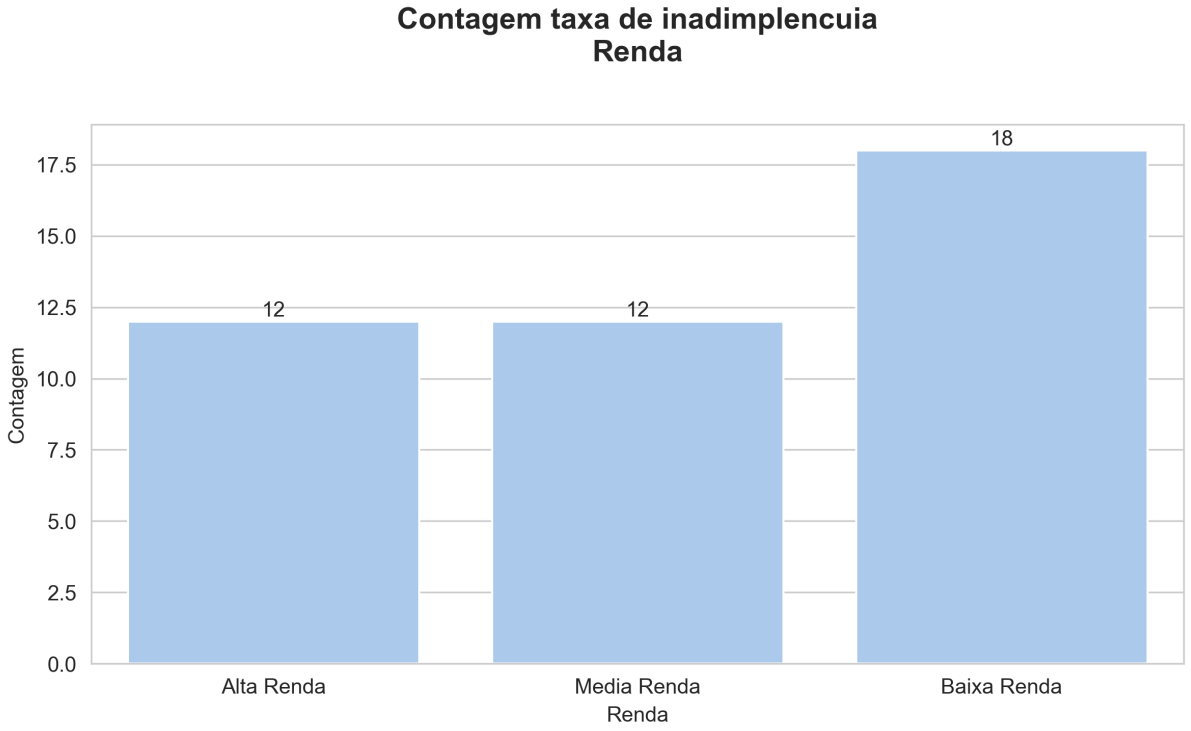
Taxa de inadimplência Faixa_Idade: O que se destaca é que entre 18-24 a taxa é extremamente baixa e entre 35-44 é alta





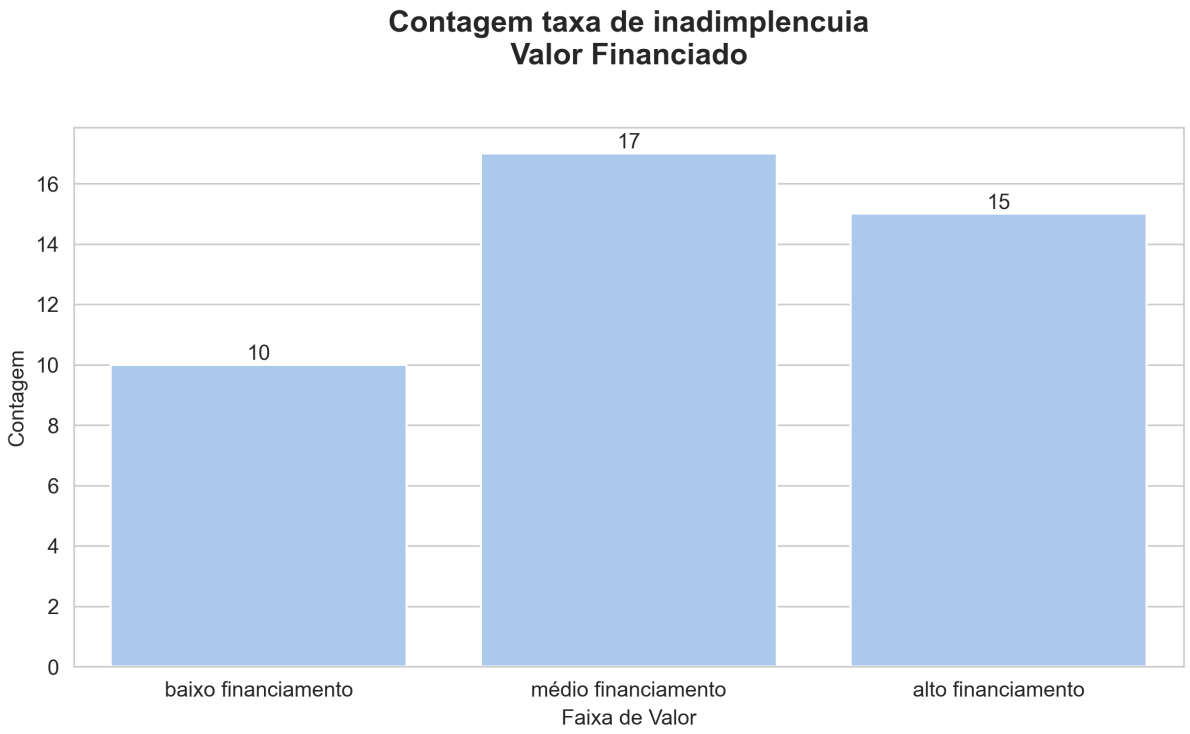
Taxa de inadimplência Gênero: Sem diferenças relevantes

Taxa de inadimplência Faixa_de_Renda: Baixa renda tem muito mais taxa de inadimplência.



Taxa de inadimplência Valor_Financiado: Taxa de Inadimplencia é menor em financiamentos de baixo porte

Detalhe: Categorizei por quartis baixo, médio e alto financiamento.



Taxa de inadimplência Valor_Financiado x Faixa_de_Renda: com exceção da média renda, financiamentos categorizados como mais altos que a categoria de renda da pessoa tende a causar mais inadimplência.

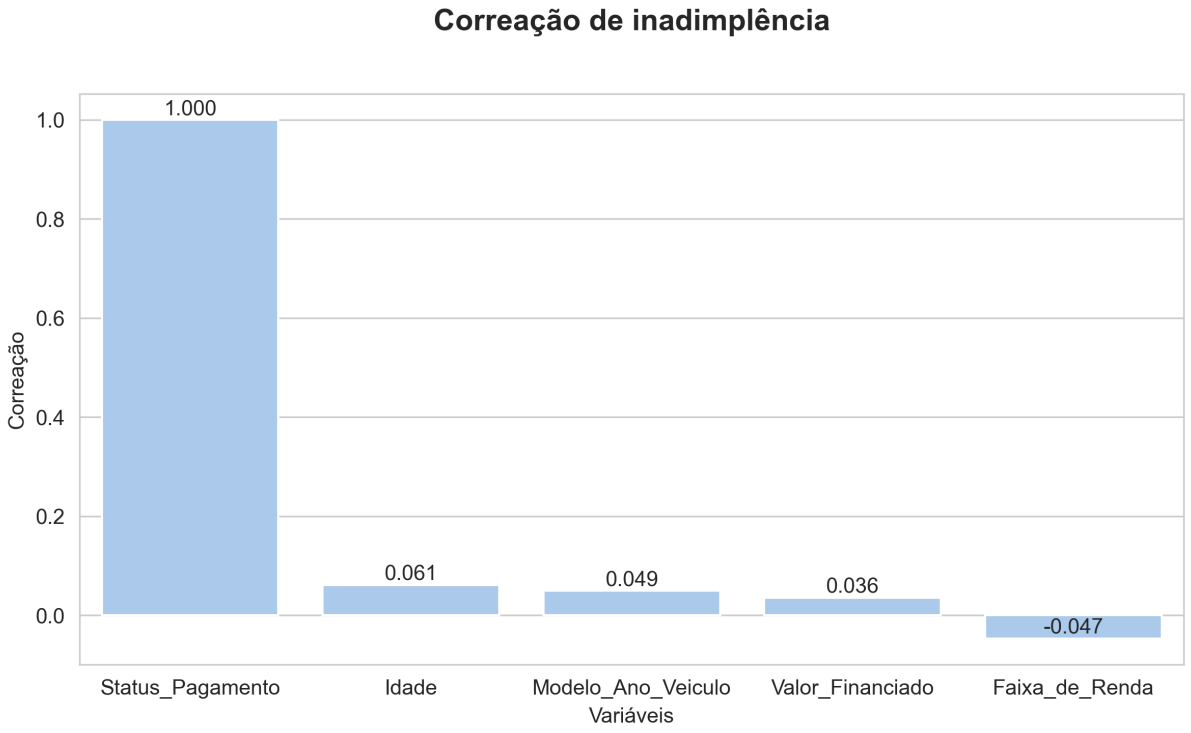
Detalhe: Agrupei e fiz uma terceira coluna, juntando faixa de renda e faixa do financiamento. Avaliei cada uma pela quantidade de inadimplência.

Quais variáveis (idade, renda, tipo de veículo, valor financiado) parecem estar mais associadas à inadimplência?

Apliquei correlação linear, mas nenhuma variável apresenta forte correlação com inadimplência. No entanto, é possível observar que temos um leve indicio de que quanto maior a idade mais chance tem de inadimplência. P mesmo do modelo do veiculo, quanto mais recente, maior a chance de inadimplência. Já quanto menor a faixa de renda, maior é a taxa de inadimplência. Se eu tiver que apontar o que mais tem correlação, seria a

idade. Mas ainda sim, por esse dados, é uma baixa correlação que pode inclusive estar associada a outros fatores.

Isso não invalida toda análise, podem ter outros fatores que influenciam ou outros tipos de correlação. Visto isso, acredito que um modelo de árvore de decisão seria mais eficiente.



Detalhe: Tive que agrupar e transformar tudo em numérico, então apliquei a função do pandas corr(method='pearson')