**Applied Statistic – Turma 9DTS**

Projeto Integrado – Parte 01

**A *QuantumFinance* está acompanhando um crescimento de inadimplência entre seus clientes e solicita a consultoria para desenvolver uma análise com base na sua carteira atual de clientes.**

**Para que a *QuantumFinance* tome decisões mais precisas sobre concessões de crédito, ela precisa aprimorar seu modelo de crédito.**

**Com o objeto de trazer novos clientes com o perfil de baixo risco de crédito desenvolva um modelo de Credit Scoring.**

**Desafio: Desenvolver o modelo preditivo mediante uso do valor target disponível na base de dados “Base\_ScoreCredito\_QuantumFinance.csv” e criar um simulador do modelo para os analistas de créditos e gerentes de conta.**

**1ª etapa: Preencher o quadro**

| **Tópico** | **O que é?** | **Como está?**  **(Mercado)** |
| --- | --- | --- |
| **Crédito ao consumidor** | Crédito ao consumidor é uma forma de troca/comércio onde uma pessoa ou empresa obtém dinheiro com a promessa de pagar por isso futuramente..  Fonte: <https://www.serasa.com.br/credito/blog/credito-direto-consumidor/> | Hoje, na terceira década do século XXI, no Brasíl, já temos uma indústria de crédito mais madura apesar de sua recente história de utilização em massa em meados dos anos 90, após a estabilidade alcançada com a implantação do plano real. Para 2024, dados divulgados pela Febraban indicam que há estimativa de alta para carteira de crédito, com revisão positiva que chega a 10%.  Fonte: <https://portal.febraban.org.br/noticia/4152/pt-br/#:~:text=A%20proje%C3%A7%C3%A3o%20de%20crescimento%20para,avan%C3%A7o%20de%209%2C3%25> |
| **Risco de crédito** | O risco de crédito indica se as chances de uma pessoa honrar um compromisso financeiro são altas ou baixas.  Fonte: <https://www.spcbrasil.org.br/blog/risco-de-credito> | Desde o início dos anos 2000, houve um significativo aumento no risco de crédito para instituições financeiras, dado que entre 2004 e 2012 houve um aumento de 209,9% no volume de crédito, quando comparado a um aumento de apenas 1.6% entre 1995 e 2003. Em 2024, com o atual amadurecimento de coleta de dados sobre os consumidores, e soluções mais robustas para análise do risco de crédito, há direcionamento positivo para carteiras de crédito ao consumidor com instituições financeiras propensas a correr mais risco.  Fonte:  <https://portal.febraban.org.br/noticia/4064/pt-br/> |
| **Inadimplência** | É a falha do devedor em cumprir o compromisso feito em cima do crédito assegurado.  Fonte:  <https://www.serasa.com.br/limpa-nome-online/blog/o-que-e-inadimplencia/> | Para 2024, a projeção reforça a tese de que a inadimplência deve seguir em queda. Para 2025, a estimativa ficou em 4,3%, sugerindo continuidade de tal movimento no próximo ano.  Fonte:  <https://portal.febraban.org.br/noticia/4064/pt-br/> |
| **Endividamento da população** | Se refere ao aumento das dívidas da população de um país em relação à sua capacidade de honrar com estas dívidas.  Fonte:  <https://www.serasa.com.br/limpa-nome-online/blog/endividamento-no-brasil/> | O número percentual de famílias brasileiras endividadas em abril de 2024 foi de 78.5%, representando alta ante Março, de acordo com os dados do levantamento da Pesquisa de Endividamento e Inadimplência do Consumidor da CNC (Confederação Nacional do Comércio de Bens, Serviços e Turismo).  Fonte:  <https://www.poder360.com.br/economia/endividamento-das-familias-sobe-para-785-em-abril/#:~:text=O%20percentual%20de%20fam%C3%ADlias%20brasileiras,m%C3%AAs%20de%20consecutivo%20de%20alta>. |

**2ª etapa: Preencher o quadro conceitual estatístico**

| COMPONENTES | DESCRIÇÃO |
| --- | --- |
| 1.Tema | Crédito |
| 2.Problema | Aumento das taxas de inadimplência entre os clientes da QuantumFinance |
| 3.Hipóteses conceituais | 1. O modelo atual de análise de crédito está defasado 2. As features/características selecionadas para a atual análise de crédito estão subutilizadas |
| 4.Objetivo Principal | Desenvolver um modelo de Credit Scoring que considere o panorama atual do mercado e aplicá-lo na população de clientes da Quantum Finance, tendo conhecimento da situação delicada que o país se encontra com mais de 70% das famílias endividadas. Assim, objetivamos dar dar continuidade na oferta de carteiras de crédito com qualidade, enquanto se reduz as possibilidades risco de crédito e taxas de inadimplência. |
| 1. População de Estudo | Clientes da QuantumFinance |

**3ª etapa: Conhecer os dados**

Atividade: Listar as variáveis qualitativas.

| Sexo, Estado Civíl, Escola (escolaridade), Trabalha (boolean), Região de moradia, Casa própria (boolean) |
| --- |

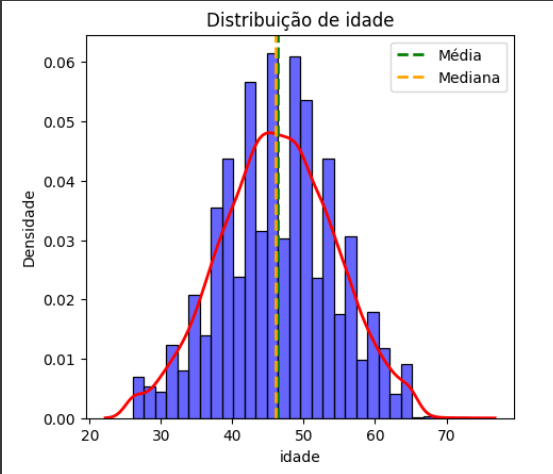
Atividade: Listar as variáveis quantitativas.

| Idade, Quantidade de dependentes, Tempo no último serviço/trabalho, Salário anual em milhares, Valor do imóvel em milhares, Quantidade de cartões, Quantidade de carros, Score de crédito. |
| --- |

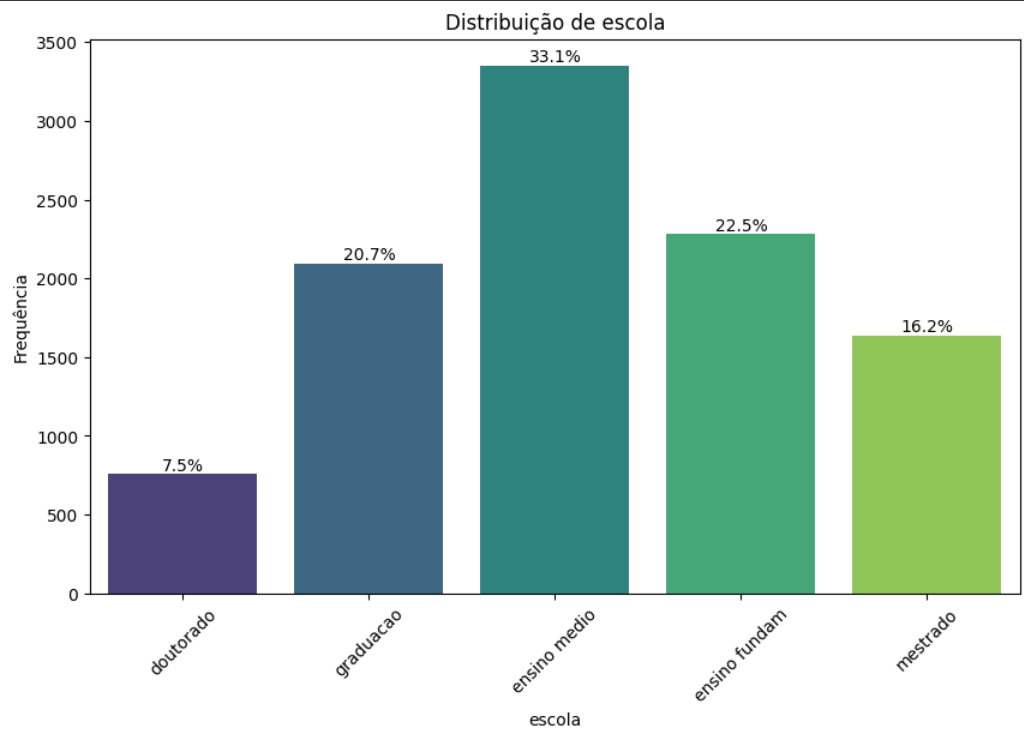
**4ª etapa: Preencher o quadro conceitual estatístico.**

| COMPONENTES | DESCRIÇÃO |
| --- | --- |
| 1. Plano Básico de Análise | Com o objetivo de aprimorar a precisão e a capacidade preditiva do nosso modelo de score de crédito para reduzir as taxas de inadimplência, planejamos e realizamos uma análise estatística detalhada dos dados disponíveis.  **Passo 1: Coleta e Preparação dos Dados**  Iniciamos nossa análise a partir de uma amostra representativa da base de clientes da QuantumFinance.  Em seguida, realizamos uma análise exploratória inicial utilizando as funções info() e describe() da biblioteca pandas para obter um entendimento geral do dataset e identificar possíveis inconsistências ou valores ausentes.  **Passo 2: Categorização e Análise das Variáveis**  Categorizamos as variáveis em numéricas e categóricas para facilitar a análise.  As variáveis numéricas, como idade, quantidade de dependentes, cartões, carros, valor da renda e valor do imóvel foram analisadas através de histogramas e boxplots, nos permitindo visualizar a distribuição, simetria dos dados e presença de outliers.  Já para variáveis categóricas, como sexo, estado civil, nível de escolaridade, região da moradia, casa própria, e se a pessoa atualmente trabalha, analisamos através de gráficos de barras, permitindo-nos visualizar a distribuição das categorias e identificar as mais frequentes.  **Próximos Passos**  Com base nos resultados da análise exploratória, os próximos que sugeriríamos para a construção de um modelo para a QuantumFinance são:   1. Maior tratamento dos dados para a remoção de outliers 2. Engenharia de features para criar novas variáveis que possam agregar valor ao modelo 3. Seleção de variáveispara identificar variáveis mais relevantes para a predição do score de crédito. 4. E a construção e validação do modelo para interpretar os resultados   Concluindo, a análise exploratória dos dados nos proporcionou um entendimento mais profundo sobre o perfil dos nossos clientes e as variáveis podem influenciar no risco de crédito.  **Insights:**   * Observamos uma concentração de clientes na faixa etária entre 40 e 50 anos. * A maioria dos clientes possui escolaridade completa até o ensino médio. * Identificamos a presença de outliers na variável 'valor do imóvel', que podem influenciar os resultados da análise * O público alvo é em sua maioria feminino * Mais de 90% dos indivíduos atualmente trabalham * Apenas 35% dos indivíduos possuem casa própria |

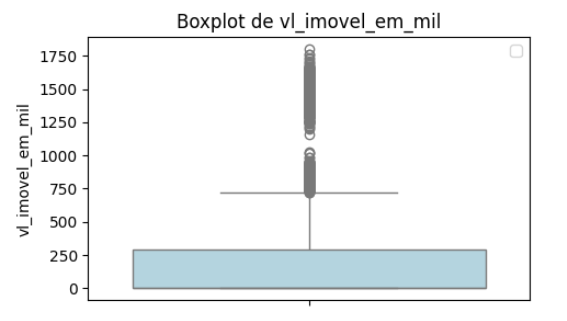
**5ª etapa: Faça a análise descritiva das variáveis. Apresente os gráficos e as medidas resumos.**

****

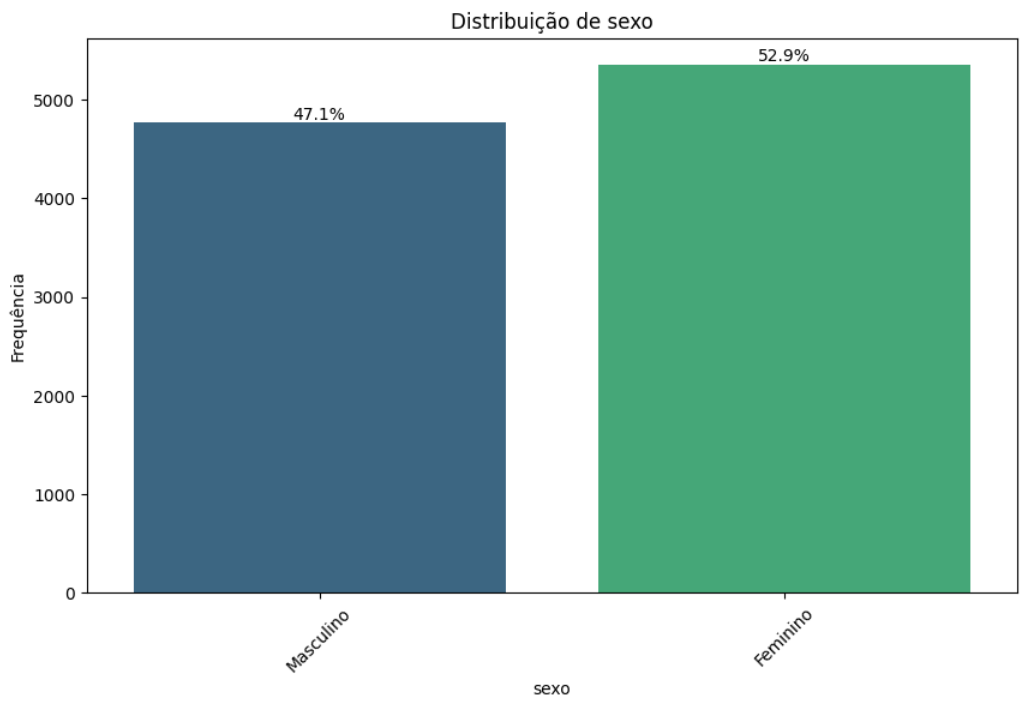
* Observamos uma concentração de clientes na faixa etária entre 40 e 50 anos.



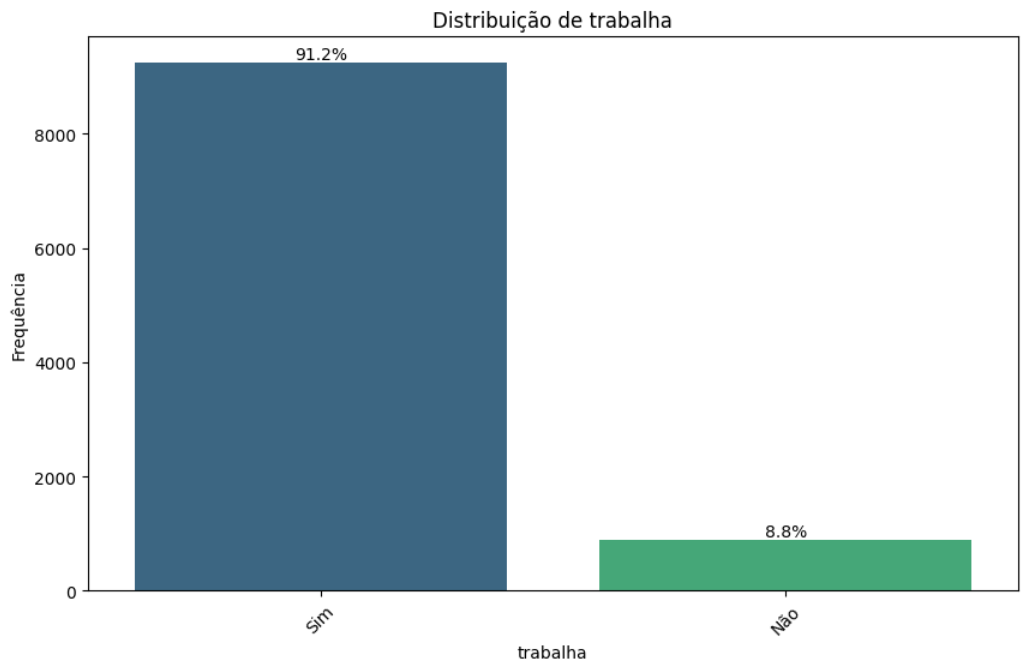
* A maioria dos clientes possui escolaridade completa até o ensino médio.



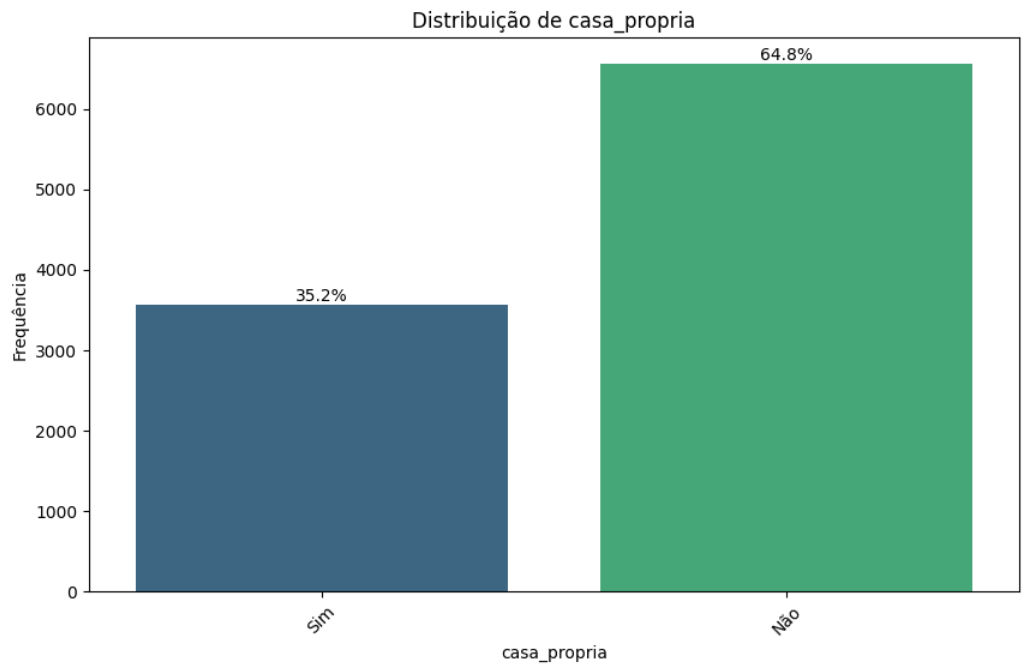
* Identificamos a presença de outliers na variável 'valor do imóvel', que podem influenciar os resultados da análise



* O público alvo é em sua maioria feminino



* Mais de 90% dos indivíduos atualmente trabalham



* Apenas 35% dos indivíduos possuem casa própria

# Segunda parte

**6ª etapa: Faça a análise bivariada das variáveis qualitativas e interprete os resultados.**

**a)** **Tabela de frequência bivariada.**

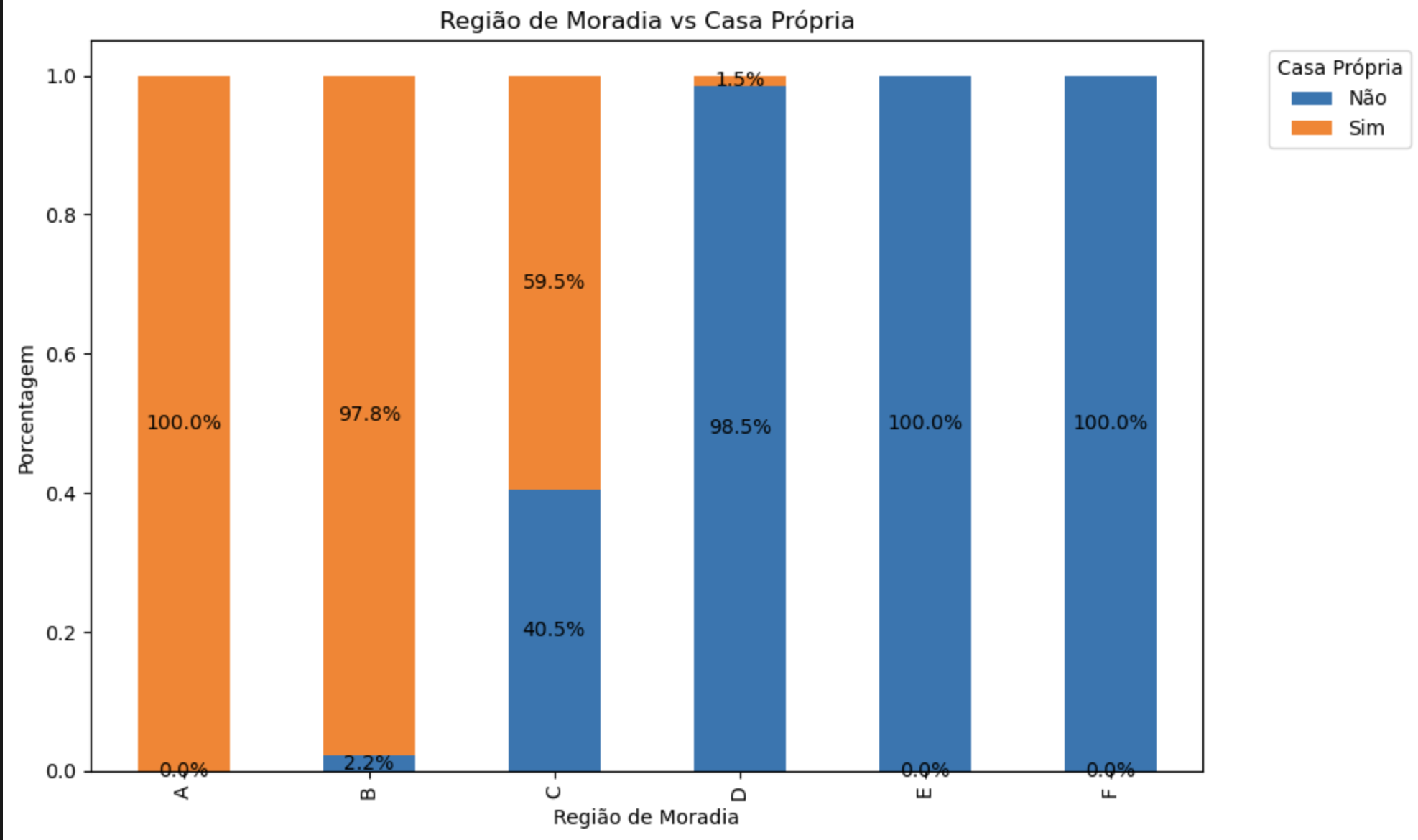
**b)** **Teste Qui-quadrado.**

**c)** **Gráfico 100% empilhado.**

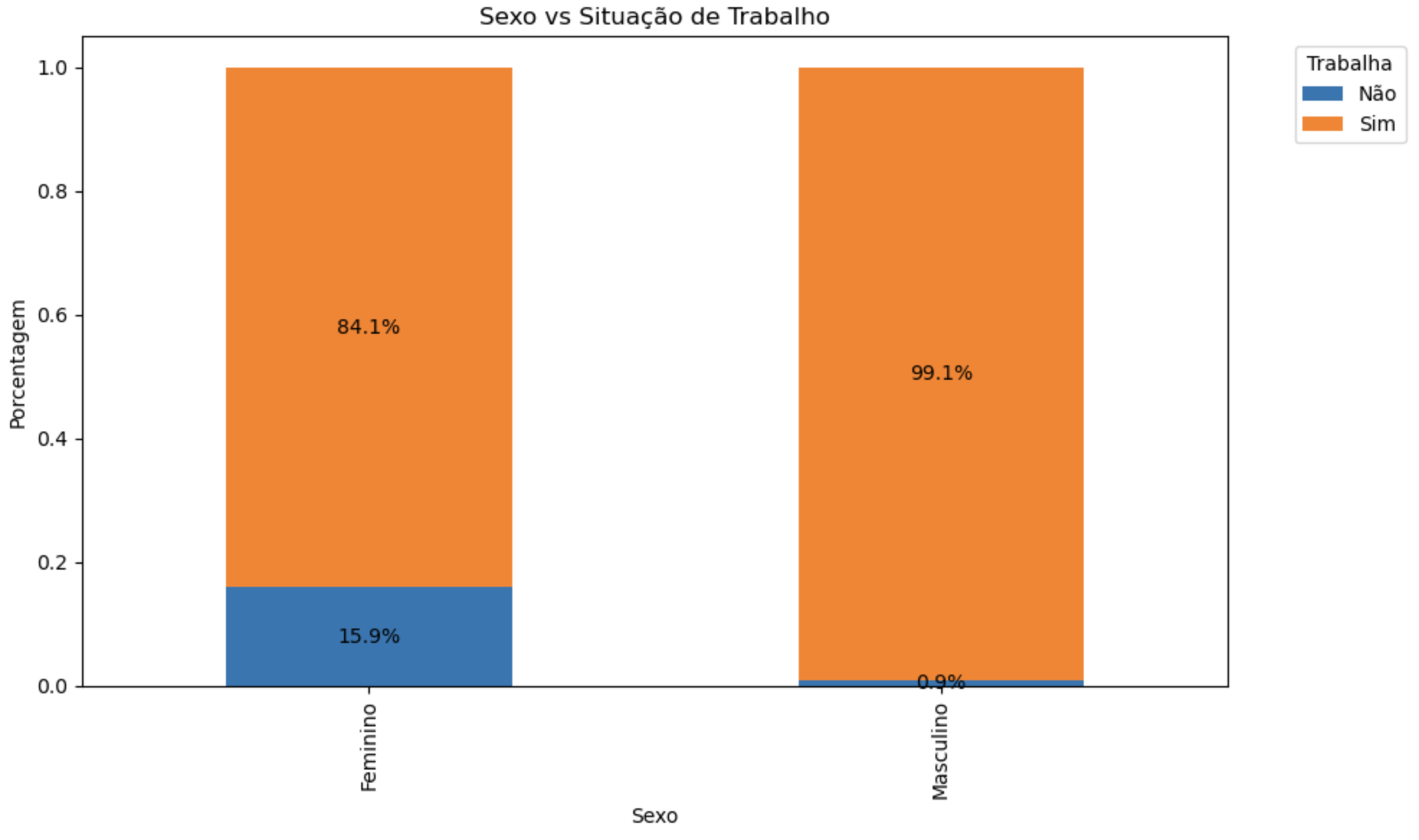
**Resposta:**

Analisando os resultados dos testes qui-quadrado para todas as relações bivariadas possíveis, identificamos algumas associações com maior significância estatística. A seguir, fizemos a seleção destas associações de forma ordenada pelos valores qui-quadráticos, da associação mais forte à mais fraca, também considerando o valor p para todas as análises. Os valores comparativos tabelados para todas as outras possibilidades combinatórias podem ser encontrados no arquivo .ipynb.

**Região de moradia vs Casa própria**

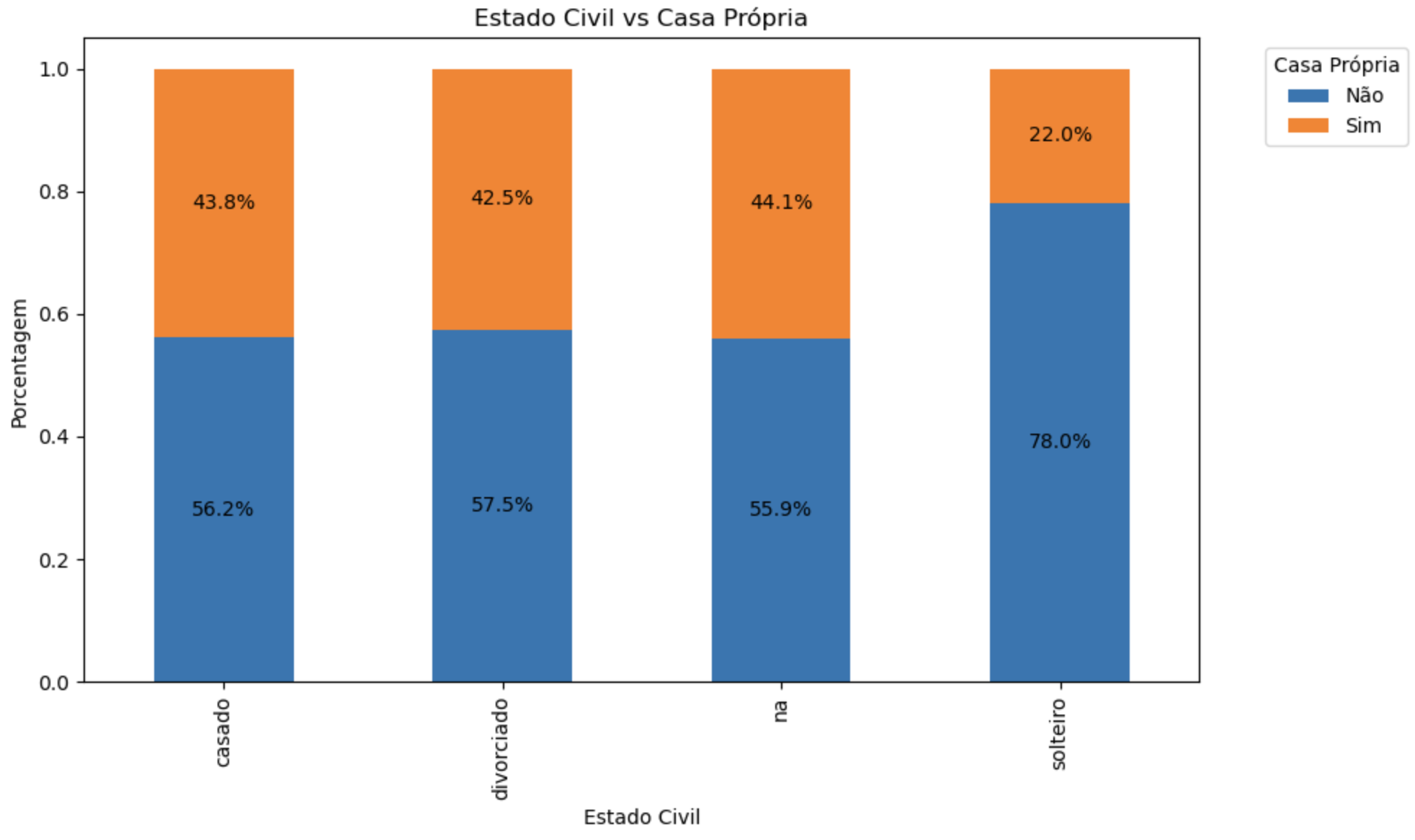
* Qui-quadrado: 7521.46
* Valor-p: 0.00
* ****
* Existe uma clara variação na posse de casa própria entre as regiões. As regiões "A" e "B" têm as maiores taxas de casa própria, com 100% e 97.8%, respectivamente. O teste qui-quadrado significativo confirma a associação entre região e posse de casa própria.

**Sexo vs Trabalha**

* Qui-quadrado: 703.36
* Valor-p: 0.00
* ****
* Existe disparidade na situação de trabalho entre os sexos. 99,1% dos homens trabalham, comparado a 84,1% das mulheres. O teste qui-quadrado indica uma associação considerável.

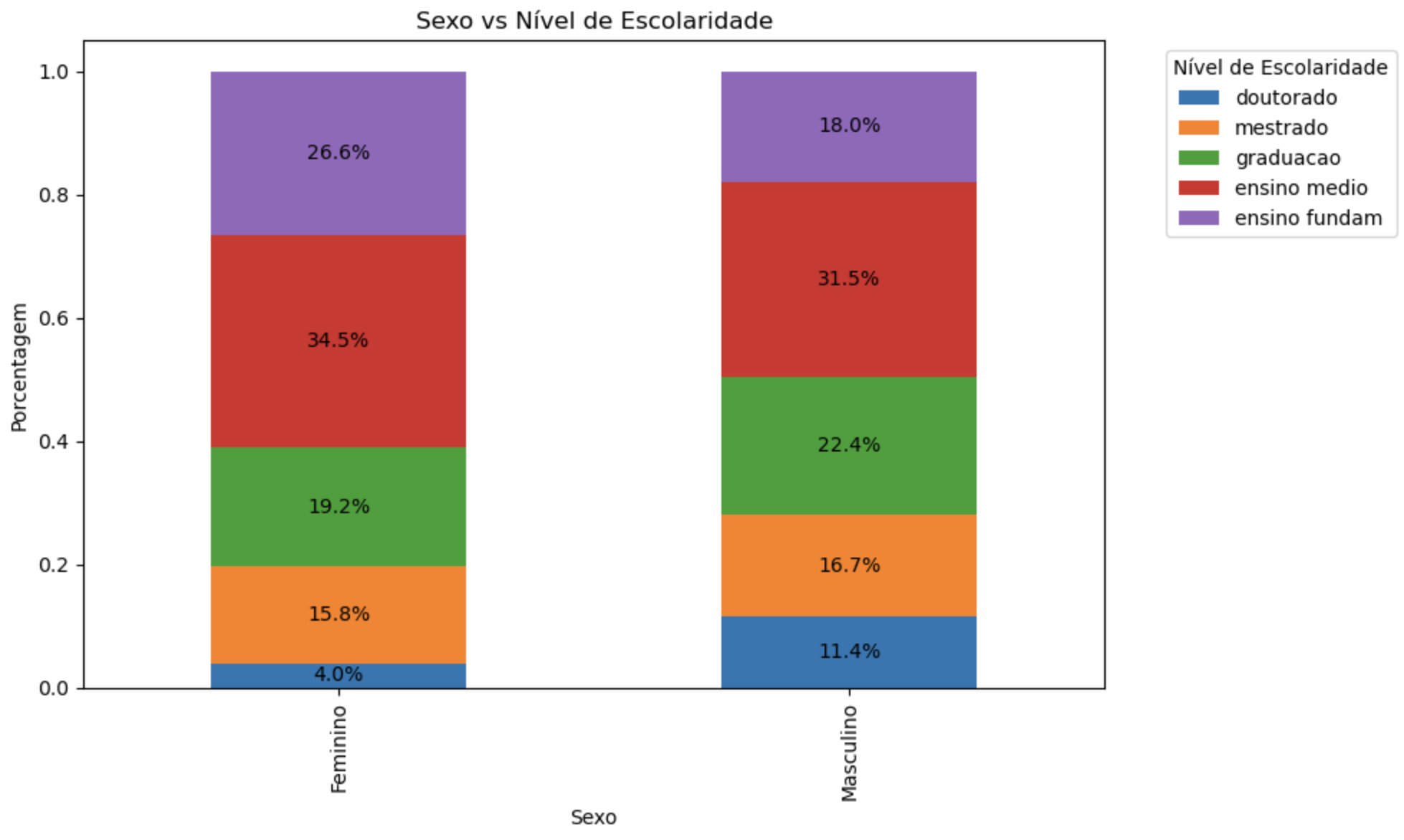
**Estado civíl vs Casa própria**

* Qui-quadrado: 497.56
* Valor-p: 0.00

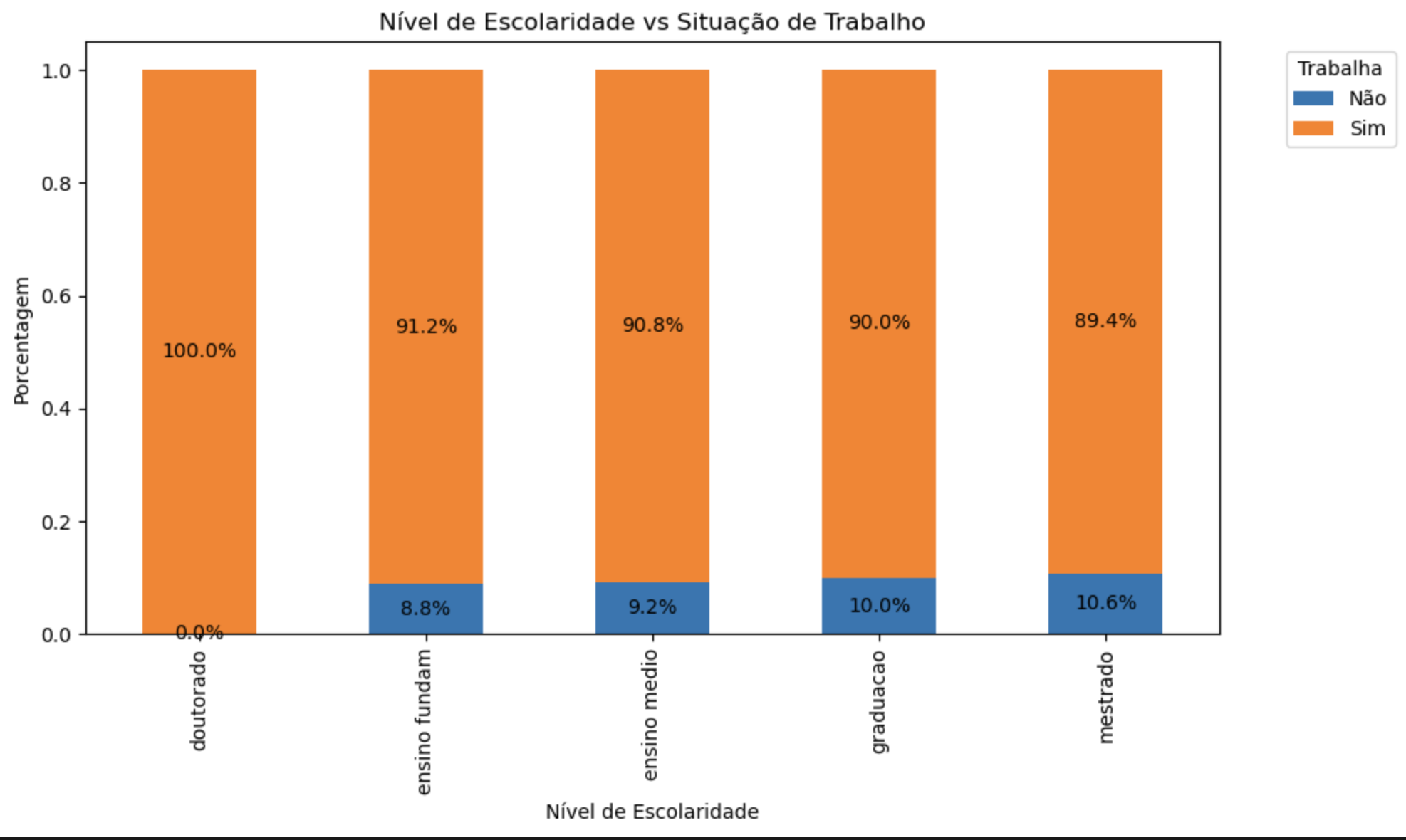
****

* ​​Solteiros tem a menor taxa de casa própria (22%), enquanto casados tem a maior (44,1%)
* Existe uma relação entre estado civil solteiro e não possuir a casa própria. Quando observamos o teste de Qui-Quadrado, obtemos um valor de *497.5635*, podemos concluir que existe uma associação significativa entre o estado civil e a posse de casa própria. Ou seja, a probabilidade de uma pessoa ser proprietária de uma casa varia de acordo com seu estado civil.

**Sexo vs Escolaridade**

* Qui-quadrado: 291.10
* Valor-p: 0.00
* 
* É possível notar diferenças entre os niveis educacionais entre homens e mulheres. Enquanto as mulheres tem maiores taxas de escolaridade do que os homens nos níveis fundamental e médio, há maior presença percentual de homens com o ensino superior completo (graduação, mestrado e doutorado). O teste qui-quadrado indica associação entre sexo e nível educaciona.
* Com base nos resultados de Qui-Quadrado e significância, podemos concluir que existe uma associação significativa entre o sexo e o nível de escolaridade. Ou seja, a probabilidade de uma pessoa estar em determinado nível de escolaridade varia de acordo com seu sexo.

**Escolaridade vs Trabalha**

* Qui-quadrado: 84.1714
* Valor-p: 0.00
* 
* Com um teste qui-quadrado de menor valor, apesar de 100% dos indivíduos com doutorado trabalharem, vemos uma relação fraca mais fraca no que tange às diferenças entre taxas de emprego entre os diferentes níveis de escolaridade, quando comparado à relação do sexo com a taxa de emprego.

**7ª etapa: Faça a análise bivariada das variáveis quantitativas e interprete os resultados.**

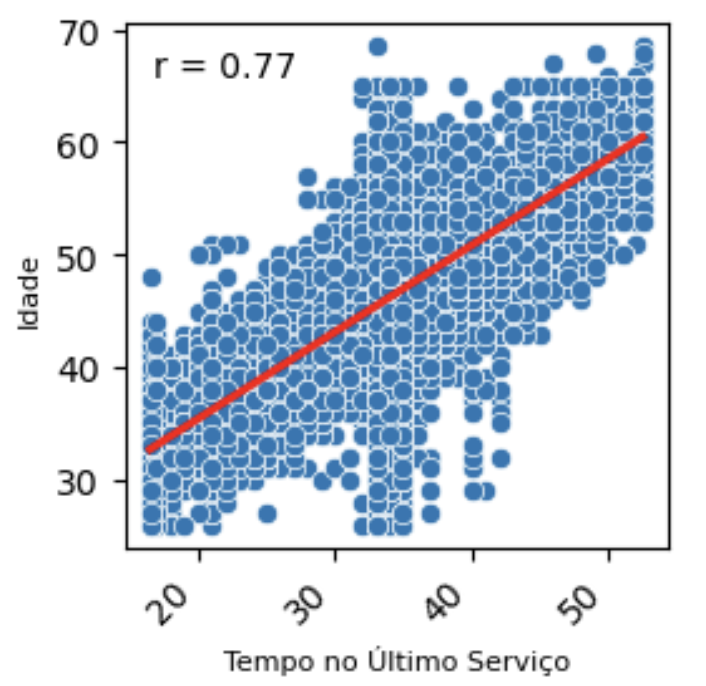
**d)** **Gráfico de dispersão.**

**e)** **Análise de correlação de Pearson.**

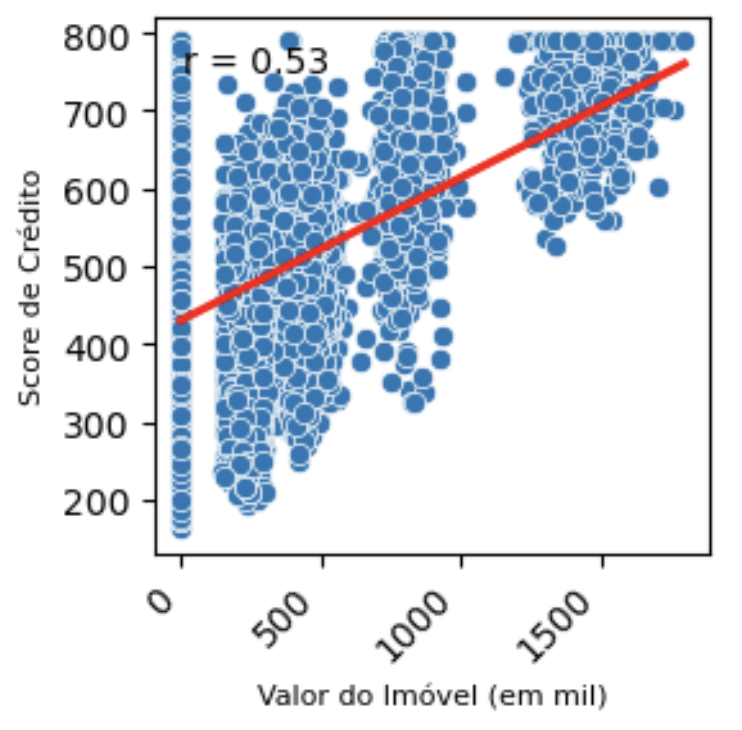
**f)** **Matriz de correlação de Pearson.**

Analisando a matriz de dispersão apresentada, podemos identificar os pares de variáveis que apresentam as maiores correlações. Vou listar as correlações mais fortes em ordem decrescente de magnitude:

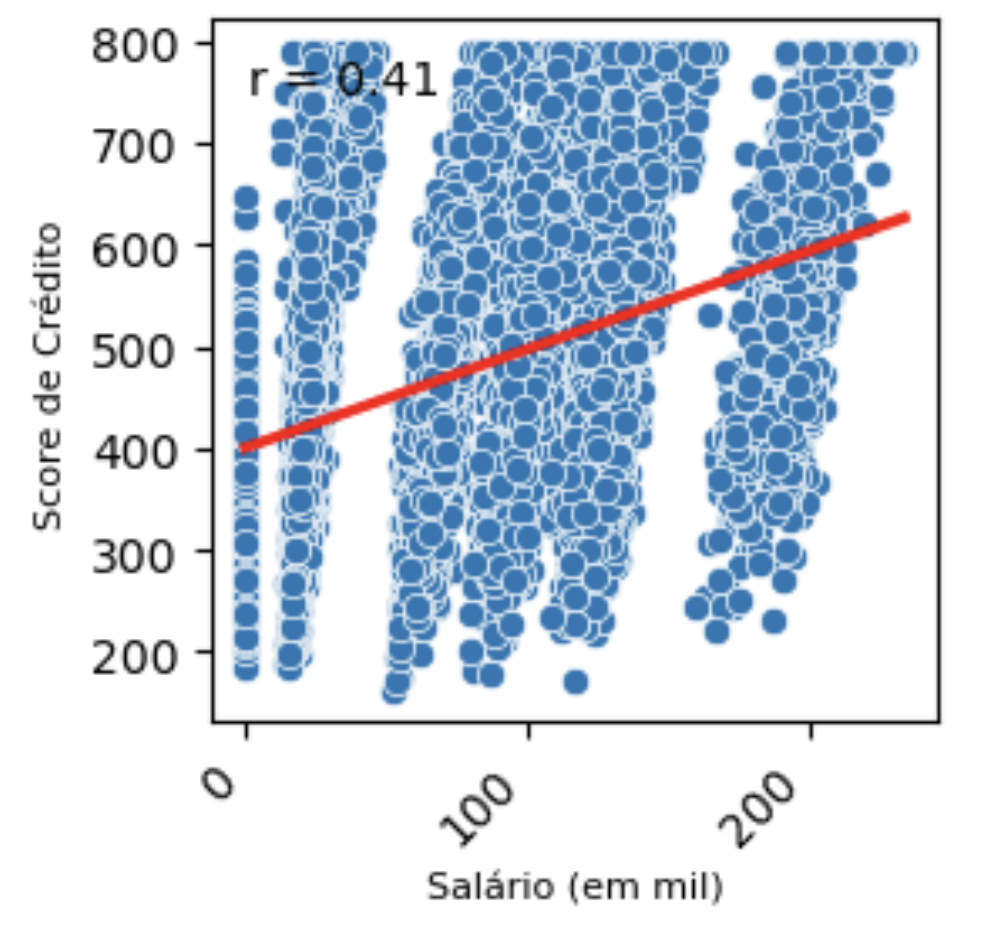
1. Tempo no Último Serviço vs Idade: r = 0.77. Esta é a correlação mais forte observada, mostrando uma relação positiva forte entre a idade da pessoa e o tempo no último serviço.

****

1. Score de Crédito vs Valor do Imóvel (em mil): r = 0.53 Também uma correlação positiva moderada, sugerindo que pessoas com scores de crédito mais altos tendem a ter imóveis de maior valor.



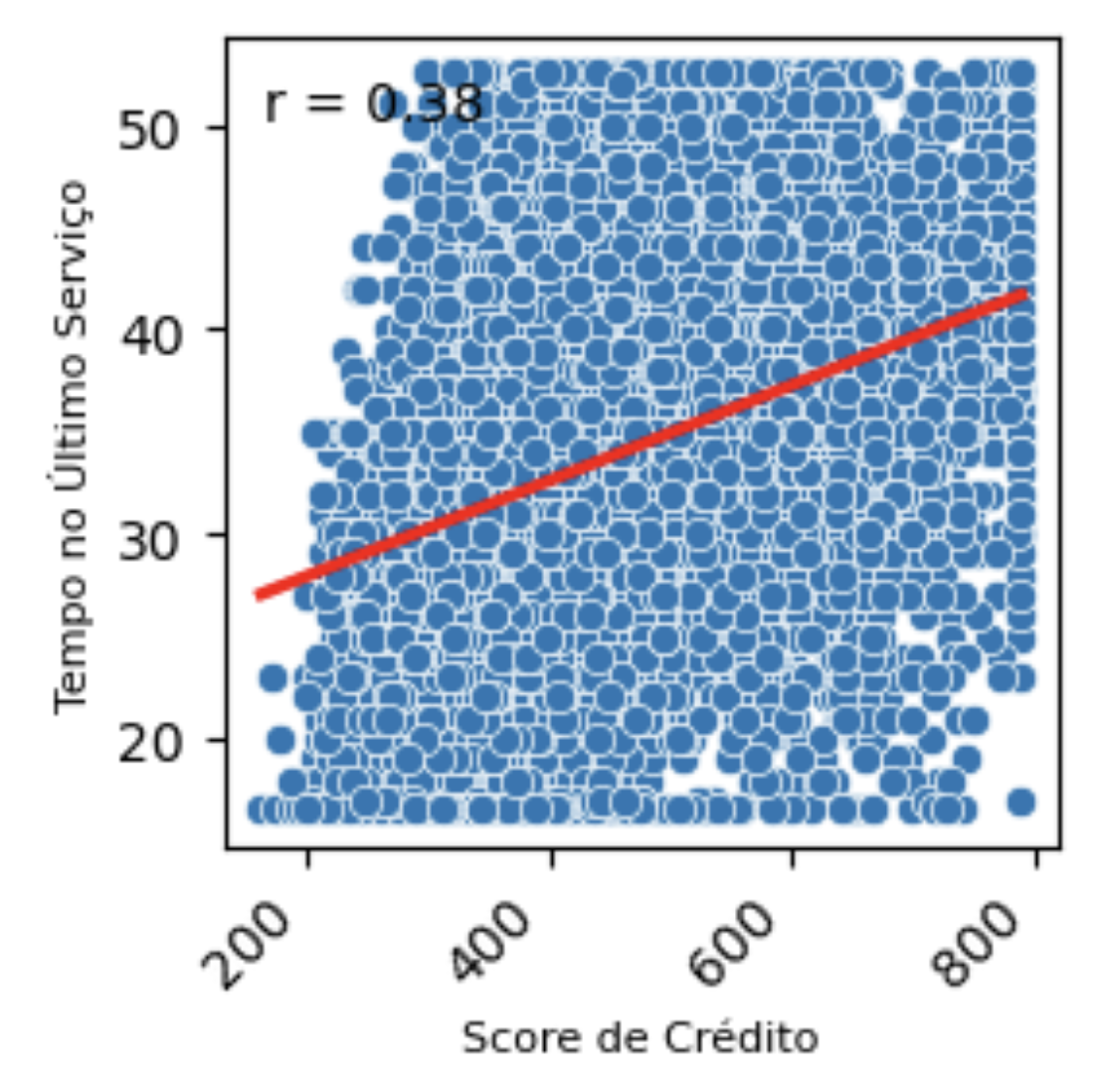
1. Score de Crédito vs Salário (em mil): r = 0.41 Há uma correlação positiva moderada, indicando que scores de crédito mais altos tendem a estar associados a salários mais altos.



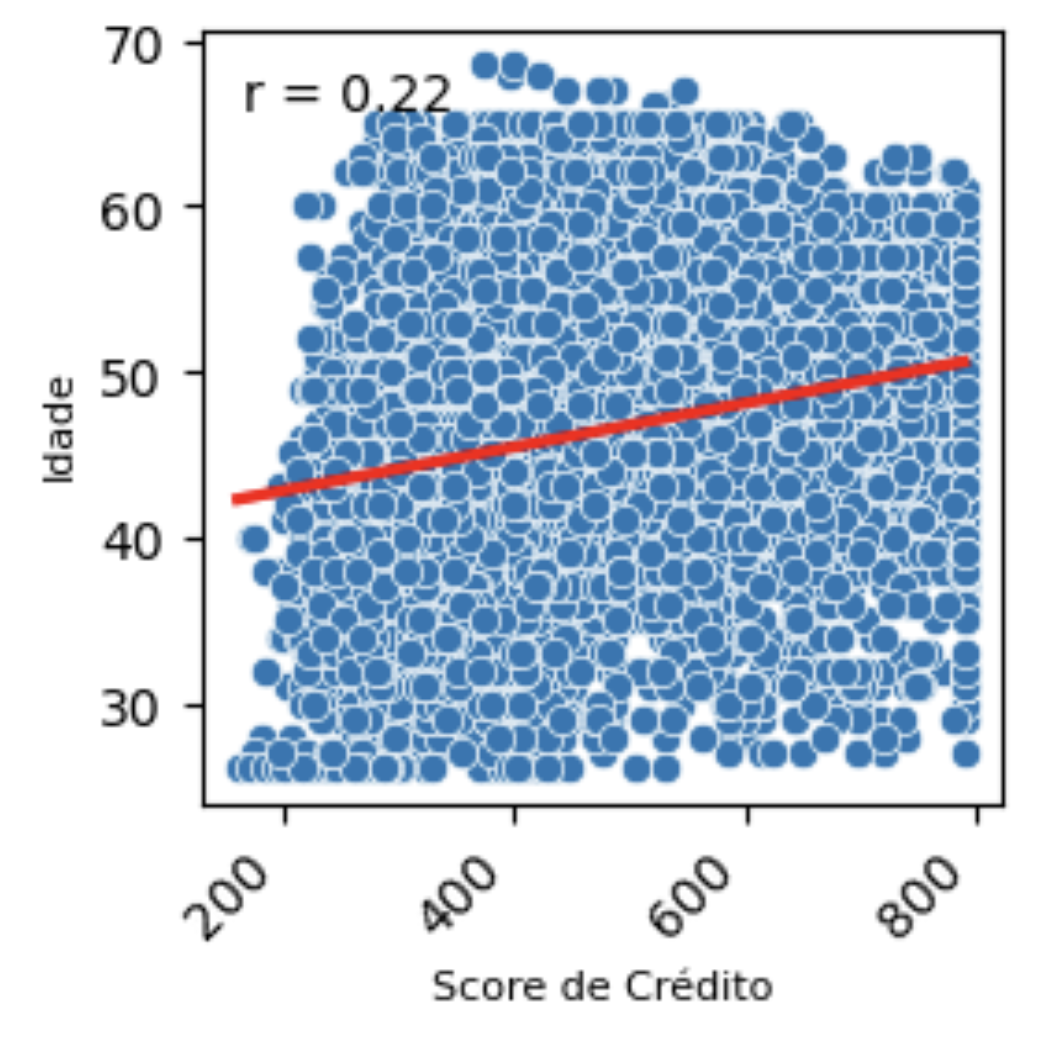
1. Salário (em mil) vs Valor do Imóvel (em mil): r = 0.41 Uma correlação positiva moderada, indicando que salários mais altos estão associados a imóveis de maior valor.

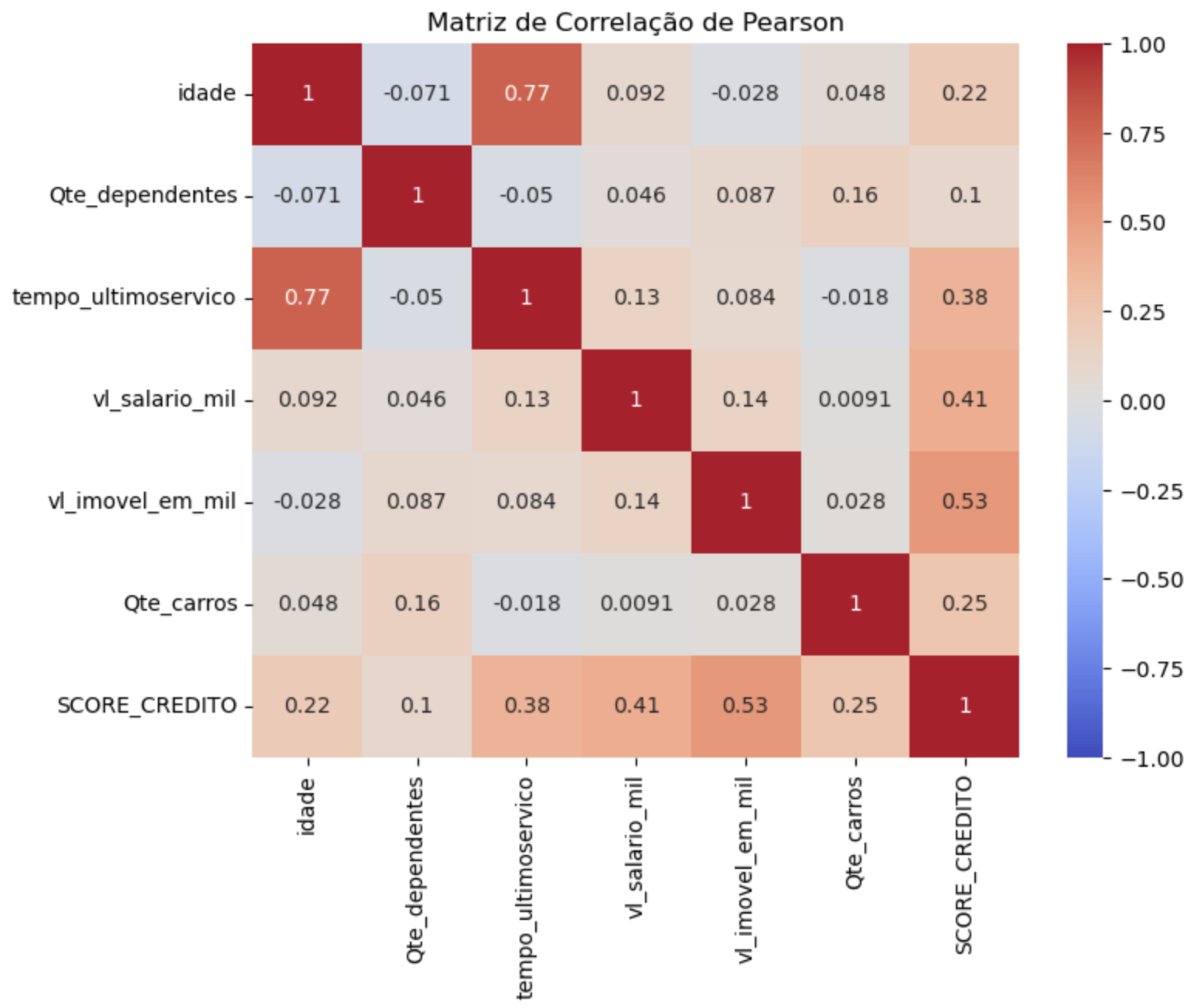


1. Score de Crédito vs Tempo no Último Serviço: r = 0.38 Semelhante à correlação com idade, indicando uma tendência de scores de crédito mais altos para pessoas com mais tempo no último serviço.

****

1. Score de Crédito vs Idade: r = 0.22 Uma correlação positiva fraca a moderada, sugerindo que o score de crédito tende a aumentar ligeiramente com a idade.



****

**8ª etapa: Preencher o quadro conceitual estatístico.**

| COMPONENTES | DESCRIÇÃO |
| --- | --- |
| 1. Técnica Estatística | Após a análise descritiva e testes de relação entre variáveis, utilizando o método Qui-Quadrado para avaliar associações entre variáveis categóricas, entramos na etapa de modelagem.  Aqui iremos:   1. Realizar a remoção de outliers nas variáveis numéricas onde for necessário 2. Aplicar o método de Hot-Encoding para variáveis categóricas de interesse 3. Selecionar apenas as variáveis numéricas com alta correlação com a variável-alvo (Score de crédito) 4. Analisar erros e resíduos e realizar testes práticos com os resultados atingidos |
| 1. Resultados Estatístico Principal | O resultado estatístico principal do teste Qui-Quadrado envolve dois componentes principais:   1. **Estatística Qui-Quadrado**: A medida numérica que compara as frequências observadas e esperadas para determinar se há uma associação entre as variáveis categóricas. Por exemplo, no teste **sexo vs trabalha**, a estatística Qui-Quadrado foi 703.3681, indicando uma grande discrepância entre os valores observados e esperados. 2. **p-value**: O p-value indica a probabilidade de que as diferenças entre as frequências observadas e esperadas sejam devidas ao acaso. Se o p-value for menor que o nível de significância (geralmente 0.05), rejeitamos a hipótese nula e concluímos que há uma associação significativa entre as variáveis. Por exemplo, para **sexo vs escola**, o p-value foi 0.000000, sugerindo uma associação estatisticamente significativa.   O resultado principal do Qui-Quadrado, portanto, é a verificação de se existe uma associação significativa entre as variáveis categóricas analisadas, com base na estatística Qui-Quadrado e no p-value. |

**9ª etapa: Construção do modelo preditivo e interprete os resultados.**

**g)** **Selecionar as variáveis preditoras.**

**h)** **Definir a variável resposta.**

**i)** **Rodar o modelo de Regressão Linear Múltipla.**

**j)** **Análise de resíduos.**

**k)** **Calcular as medidas de erros do modelo na amostra de desenvolvimento.**

**l)** **Calcular as medidas de erros do modelo na amostra de validação.**

**m)** **Construir o simulador do modelo.**

As variáveis selecionadas foram: *vl\_salario\_mil, vl\_imovel\_em\_mil, estado\_civil, escola, casa\_propria.*

A variável resposta é *SCORE\_CREDITO*

Análise de erros indica que o modelo pode não ser um bom preditor para a variável resposta, conforme consta abaixo:

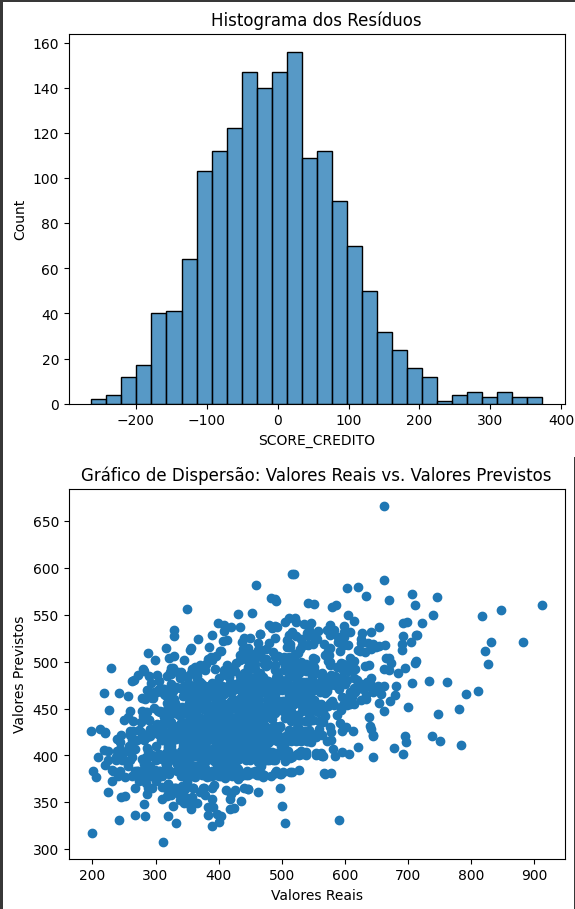
***Mean Squared Error: 9064.962812573625***

***Root Mean Squared Error: 95.21009826995046***

***R-squared: 0.20514872663583528***

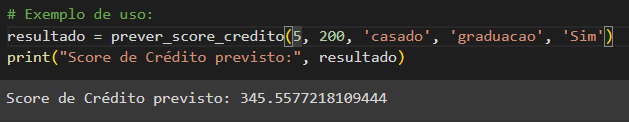
R-quadrado de 0.2 indica que o modelo não está capturando grande quantidade da variabilidade dos dados.

Para otimizar esse dado, poderíamos realizar etapas de *feature-engineering*, otimizar o processo de seleção de variáveis ou, preferencialmente, utilizar outro tipo de modelo que melhor trabalhe com dados categóricos, como *random forest* ou outros.



O histograma dos resíduos apresenta normalidade, a relação entre valores reais e previstos, porém, não apresenta uma linearidade.

Com o modelo apresentado, uma possível função simuladora nos traria o seguinte resultado para um indivíduo que receba 5 mil reais, com uma casa própria de 200 mil, casado e com graduação completa:



Entrega do script Python e o relatório com as interpretações dos resultados.

Data de entrega: 20/09/2024

Regina Bernal

26/08/2024