

projeto 2
simulacao

ITCAI escola
britânica de
artes criativas
& tecnologia

Profissão: Cientista de
Dados

Por: [Rafael Rosa Alves](#)
Arquivos: [Pagina GitHub](#)

Bibliotecas/Pacotes

Previsão de renda

Etapa 1 CRISP - DM: Entendimento do negócio

Uma instituição financeira está interessada em aprofundar sua compreensão do perfil de renda de seus novos clientes para várias finalidades, como ajustar de forma mais precisa os limites de crédito dos cartões dos novos clientes, sem a necessidade de solicitar olerites ou documentação que possa afetar a experiência do cliente.

Com esse objetivo, realizou um estudo com alguns clientes, validando suas rendas por meio de olerites e outros documentos, e tem a intenção de desenvolver um modelo preditivo para prever essas rendas com base em algumas variáveis já presentes em seu banco de dados.

Etapa 2 Crisp-DM: Entendimento dos dados

Dicionário de dados

Variável	Descrição	Tipo
data_ref	Data de referência de coleta das variáveis	object
id_cliente	Código identificador exclusivo do cliente	int
sexo	Sexo do cliente (M = 'Masculino'; F = 'Feminino')	object (binária)
posse_de_veiculo	Indica se o cliente possui veículo (True = 'Possui veículo'; False = 'Não possui veículo')	bool (binária)
posse_de_imovel	Indica se o cliente possui imóvel (True = 'Possui imóvel'; False = 'Não possui imóvel')	bool (binária)
qtd_filhos	Quantidade de filhos do cliente	int
tipo_renda	Tipo de renda do cliente (Empresário, Assalariado, Servidor público, Pensionista, Bolsista)	object
educacao	Grau de instrução do cliente (Primário, Secundário, Superior incompleto, Superior completo, Pós graduação)	object
estado_civil	Estado civil do cliente (Solteiro, União, Casado, Separado, Viúvo)	object
tipo_residencia	Tipo de residência do cliente (Casa, Governamental, Com os pais, Aluguel, Estúdio, Comunitário)	object
idade	Idade do cliente em anos	int
tempo_emprego	Tempo no emprego atual	float
qt_pessoas_residencia	Quantidade de pessoas que moram na residência	float
renda	Valor numérico decimal representando a renda do cliente em reais	float

Carregando os dados

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15000 entries, 0 to 14999
Data columns (total 15 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   Unnamed: 0    15000 non-null   int64  
 1   data_ref     15000 non-null   object  
 2   id_cliente   15000 non-null   int64  
 3   sexo         15000 non-null   object  
 4   posse_de_veiculo  15000 non-null   bool   
 5   posse_de_imovel  15000 non-null   bool   
 6   qtd_filhos   15000 non-null   int64  
 7   tipo_renda   15000 non-null   object  
 8   educacao    15000 non-null   object  
 9   estado_civil 15000 non-null   object  
 10  tipo_residencia 15000 non-null   object  
 11  idade        15000 non-null   int64  
 12  tempo_emprego 12427 non-null   float64 
 13  qt_pessoas_residencia 15000 non-null   float64 
 14  renda        15000 non-null   float64 
dtypes: bool(2), float64(3), int64(4), object(6)
memory usage: 1.5+ MB
```

	Unnamed: 0	data_ref	id_cliente	sexo	posse_de_veiculo	posse_de_imovel	qtd_filhos	tipo_renda	educacao	estado_civil	tipo_residencia	idade	tempo_emprego	qt_pessoas_residencia	re
0	0	2015-01-01	15,056	F	<input type="checkbox"/>	<input checked="" type="checkbox"/>	0	Empresário	Secundário	Solteiro	Casa	26	6.6027	1	8,
1	1	2015-01-01	9,968	M	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	0	Assalariado	Superior completo	Casado	Casa	28	7.1836	2	1,
2	2	2015-01-01	4,312	F	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	0	Empresário	Superior completo	Casado	Casa	35	0.8384	2	2,
3	3	2015-01-01	10,639	F	<input type="checkbox"/>	<input checked="" type="checkbox"/>	1	Servidor público	Superior completo	Casado	Casa	30	4.8466	3	6,
4	4	2015-01-01	7,064	M	<input checked="" type="checkbox"/>	<input type="checkbox"/>	0	Assalariado	Secundário	Solteiro	Governamental	33	4.2932	1	6,
5	5	2015-01-01	10,581	F	<input type="checkbox"/>	<input checked="" type="checkbox"/>	0	Assalariado	Superior completo	Casado	Casa	39	4.3452	2	1,
6	6	2015-01-01	7,129	F	<input type="checkbox"/>	<input checked="" type="checkbox"/>	0	Empresário	Superior completo	Viúvo	Casa	55	6.3781	1	1,
7	7	2015-01-01	9,952	F	<input type="checkbox"/>	<input checked="" type="checkbox"/>	0	Empresário	Secundário	Casado	Casa	36	3.1041	2	2,
8	8	2015-01-01	883	F	<input type="checkbox"/>	<input checked="" type="checkbox"/>	0	Assalariado	Secundário	Casado	Casa	50	18.6055	2	3,
9	9	2015-01-01	8,070	M	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	0	Assalariado	Superior completo	Casado	Casa	60	10.5589	2	12,

Variável	Valores únicos
0	Unnamed: 0
1	data_ref
2	id_cliente
3	sexo
4	posse_de_veiculo
5	posse_de_imovel
6	qtd_filhos
7	tipo_renda
8	educacao
9	estado_civil

10	tipo_residencia	6
11	idade	47
12	tempo_emprego	2589
13	qt_pessoas_residencia	9
14	renda	9786

Quantidade total de linhas: 15888

Quantidade de linhas duplicadas: 487

Quantidade após remoção das linhas duplicadas: 14593

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14593 entries, 0 to 14592
Data columns (total 13 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   data_ref          14593 non-null   object  
 1   sexo              14593 non-null   object  
 2   posse_de_veiculo 14593 non-null   bool    
 3   posse_de_imovel   14593 non-null   bool    
 4   qtd_filhos        14593 non-null   int64  
 5   tipo_renda         14593 non-null   object  
 6   educacao          14593 non-null   object  
 7   estado_civil      14593 non-null   object  
 8   tipo_residencia   14593 non-null   object  
 9   idade              14593 non-null   int64  
 10  tempo_emprego    12890 non-null   float64 
 11  qt_pessoas_residencia 14593 non-null   float64 
 12  renda              14593 non-null   float64 
dtypes: bool(2), float64(3), int64(2), object(6)
memory usage: 1.3+ MB
```

Entendimento dos dados - Univariada

Pandas Profiling – Relatório interativo para análise exploratória de dados

Overview

Overview Alerts 0 Reproduction

Dataset statistics

Number of variables	13
Number of observations	14593
Missing cells	2503
Missing cells (%)	1.3%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	6.5 MiB
Average record size in memory	467.7 B

Variable types

DateTime	1
Categorical	5
Boolean	2
Numeric	5

Variables

Select Columns ▾

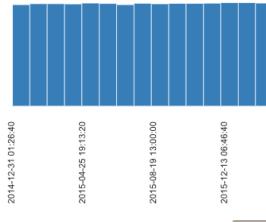
data_ref

Date

Distinct	15
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	114.1 KiB

Minimum 2015-01-01 00:00:00

Maximum 2016-03-01 00:00:00



More details

SEXO

Categorical

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	826.7 KiB

F 9851

M 4742

posse_de_veiculo

Boolean

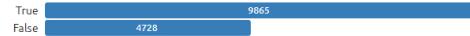
Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	14.4 KIB


[More details](#)

posse_de_imovel

Boolean

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	14.4 KIB


[More details](#)

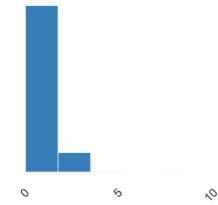
qtd_filhos

Real number (⌚)

ZEROS

Distinct	8
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	0.43328993

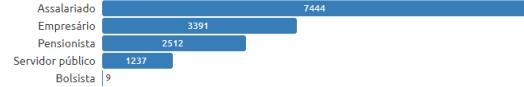
Minimum	0
Maximum	14
Zeros	10080
Zeros (%)	69.1%
Negative	0
Negative (%)	0.0%
Memory size	114.1 KIB


[More details](#)

tipo_renda

Categorical

Distinct	5
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	1.1 MiB


[More details](#)

educacao

Categorical

Distinct	5
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	1.3 MiB


[More details](#)

estado_civil

Categorical

Distinct	5
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	954.2 KIB


[More details](#)

tipo_residencia

Categorical

Distinct	6
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	883.8 KIB

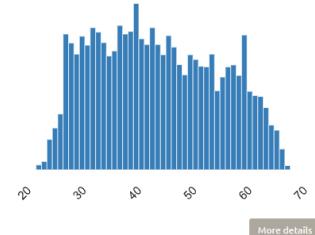

[More details](#)

idade

Real number (⌚)

Distinct	47
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	43.86691

Minimum	22
Maximum	68
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	114.1 KIB



[More details](#)

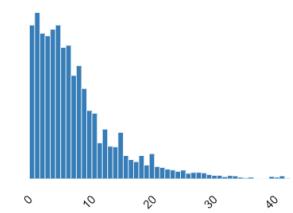
tempo_emprego

Real number ⓘ

MISSING

Distinct	2589
Distinct (%)	21.4%
Missing	2503
Missing (%)	17.2%
Infinite	0
Infinite (%)	0.0%
Mean	7.7245671

Minimum	0.11780822
Maximum	42.906849
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	114.1 KIB



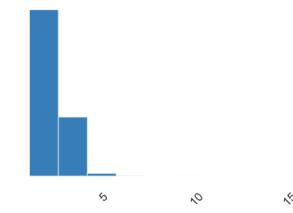
[More details](#)

qt_pessoas_residencia

Real number ⓘ

Distinct	9
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	2.2070856

Minimum	1
Maximum	15
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	114.1 KIB



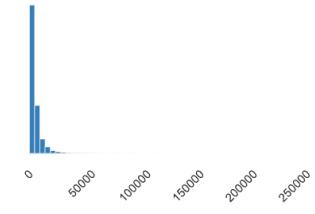
[More details](#)

renda

Real number ⓘ

Distinct	9786
Distinct (%)	67.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	5698.1406

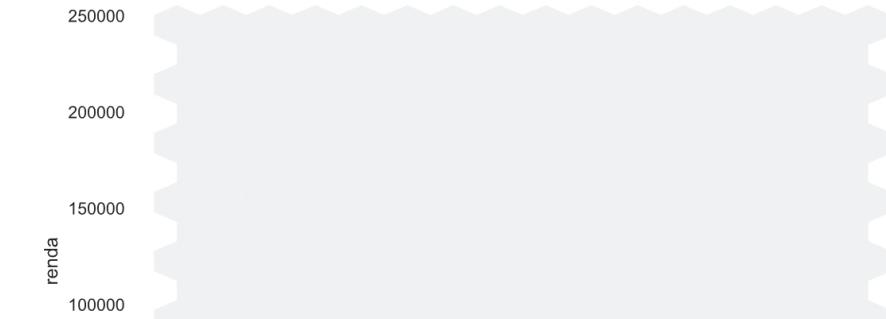
Minimum	118.71
Maximum	245141.67
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	114.1 KIB



[More details](#)

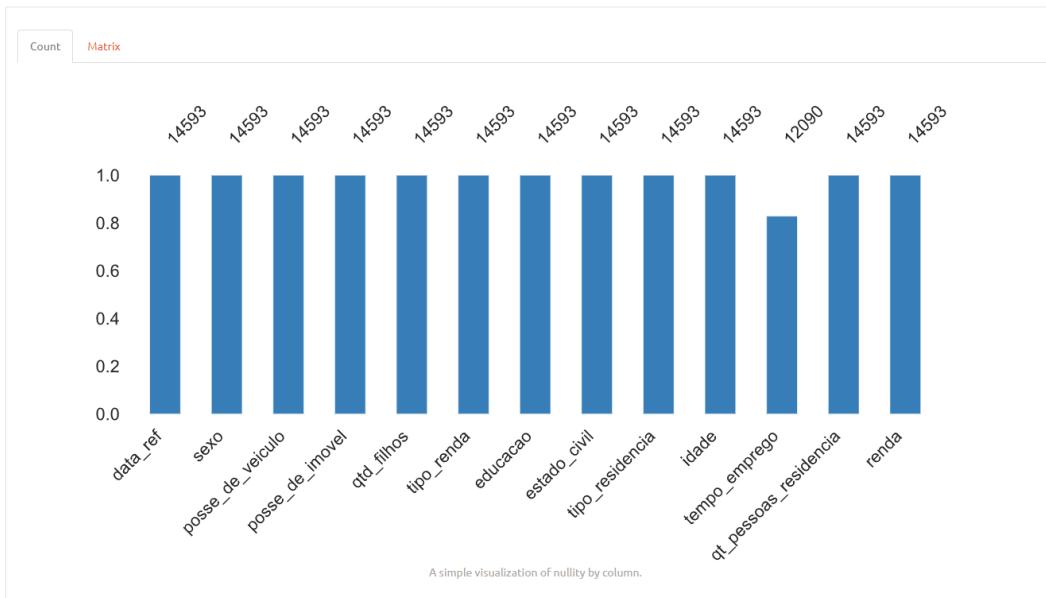
Interactions

qtd_filhos	idade	tempo_emprego	qt_pessoas_residencia	renda
renda	qtd_filhos	idade	tempo_emprego	qt_pessoas_residencia





Missing values



Sample

First rows Last rows

	data_ref	sexo	posse_de_veiculo	posse_de_imovel	qtd_filhos	tipo_renda	educacao	estado_civil	tipo_residencia	idade	tempo_emprego	qt_pessoas_r
0	2015-01-01	F	False	True	0	Empresário	Secundário	Solteiro	Casa	26	6.602740	1.0
1	2015-01-01	M	True	True	0	Assalariado	Superior completo	Casado	Casa	28	7.183562	2.0
2	2015-01-01	F	True	True	0	Empresário	Superior completo	Casado	Casa	35	0.838356	2.0
3	2015-01-01	F	False	True	1	Servidor público	Superior completo	Casado	Casa	30	4.846575	3.0
4	2015-01-01	M	True	False	0	Assalariado	Secundário	Solteiro	Governamental	33	4.293151	1.0
5	2015-01-01	F	False	True	0	Assalariado	Superior completo	Casado	Casa	39	4.345205	2.0
6	2015-01-01	F	False	True	0	Empresário	Superior completo	Viúvo	Casa	55	6.378082	1.0
7	2015-01-01	F	False	True	0	Empresário	Secundário	Casado	Casa	36	3.104110	2.0
8	2015-01-01	F	False	True	0	Assalariado	Secundário	Casado	Casa	50	18.605479	2.0
9	2015-01-01	M	True	True	0	Assalariado	Superior completo	Casado	Casa	60	10.558904	2.0

Report generated by YData.

Estatísticas descritivas das variáveis quantitativas

	count	mean	std	min	25%	50%	75%	max
qtd_filhos	14,593	0.4333	0.746	0	0	0	1	14
idade	14,593	43.8689	11.2766	22	34	43	53	68
tempo_emprego	12,090	7.7246	6.718	0.1178	3.0062	6.0137	10.1233	42.9068
qt_pessoas_residencia	14,593	2.2071	0.9091	1	2	2	3	15
renda	14,593	5,698.1406	8,314.0212	118.71	2,018.88	3,488.41	6,379.57	245,141.67

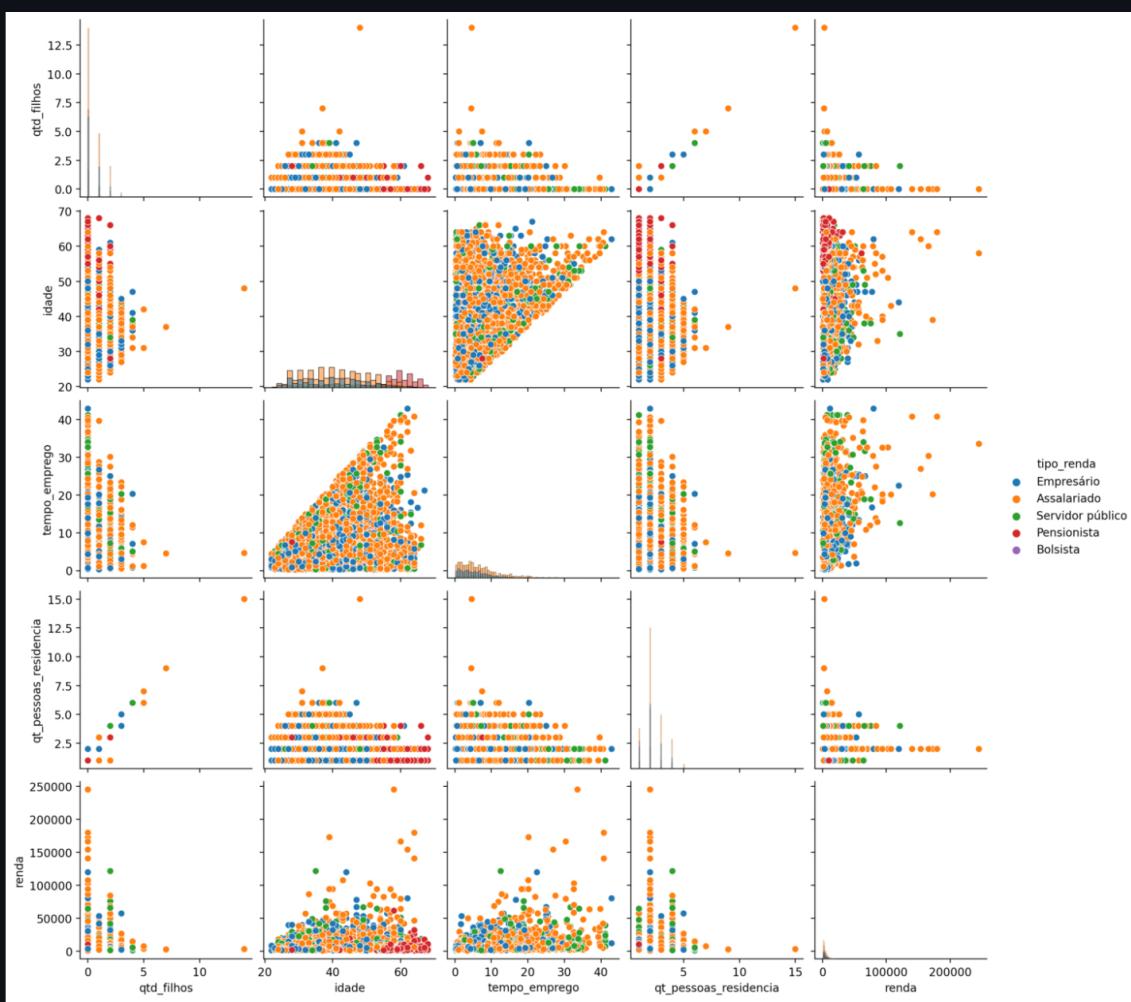
Entendimento dos dados - Bivariadas

Matriz de correlação

	posse_de_imovel	qtd_filhos	idade	tempo_emprego	qt_pessoas_residencia	renda
renda	0.0042	0.0303	0.0234	0.385	0.0486	1

Com base na matriz de correlação, nota-se que a variável `tempo_emprego` exibe a maior associação com a variável `renda`, apresentando um coeficiente de correlação de 38,5%.

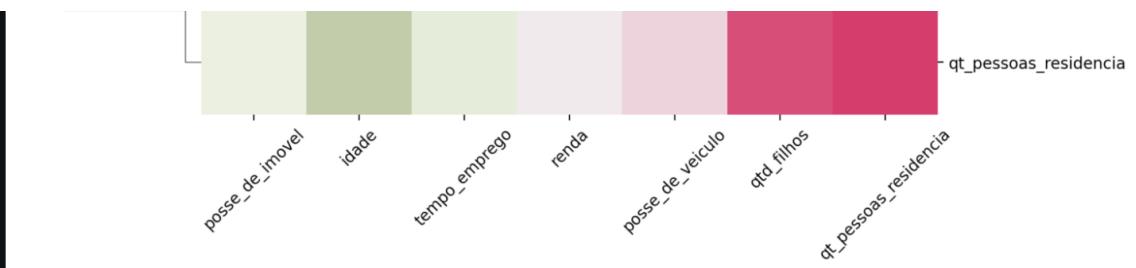
Matriz de dispersão



Ao examinar o *pairplot*, que é uma representação gráfica na forma de matriz de dispersão, é perceptível a presença de alguns *outliers* na variável `renda`, os quais têm o potencial de influenciar os resultados da análise de tendência, embora sejam pouco frequentes. Ademais, nota-se uma correlação fraca entre praticamente todas as variáveis quantitativas, corroborando os achados da matriz de correlação.

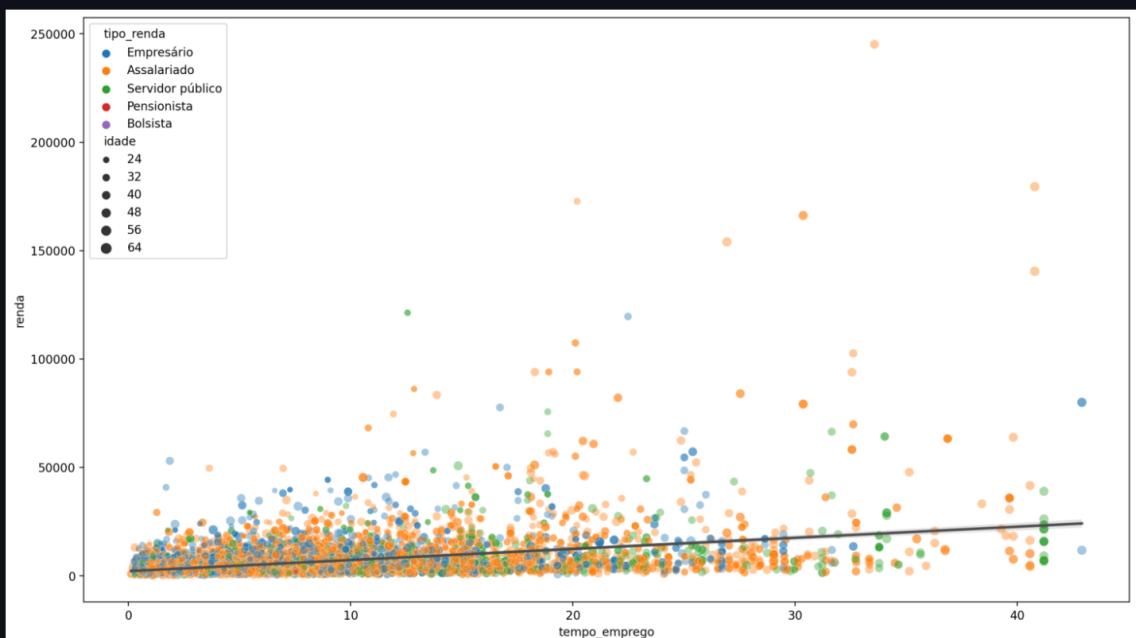
Clustermap





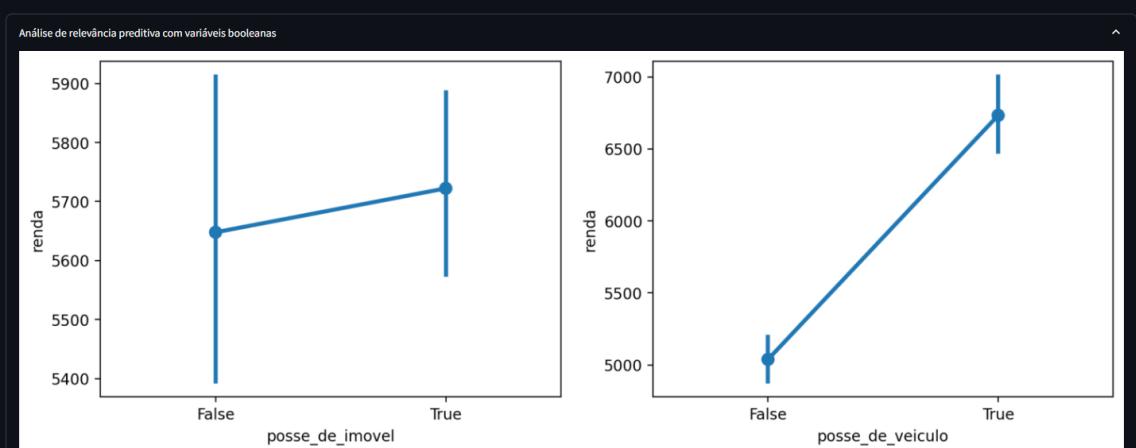
Ao analisar o *pairplot*, uma representação gráfica na forma de matriz de dispersão, é evidente a presença de alguns *outliers* na variável `renda`, os quais possuem o potencial de impactar os resultados da análise de tendência, apesar de sua baixa frequência. Além disso, observa-se uma correlação fraca entre praticamente todas as variáveis quantitativas, o que confirma os resultados obtidos na matriz de correlação.

Linha de tendência



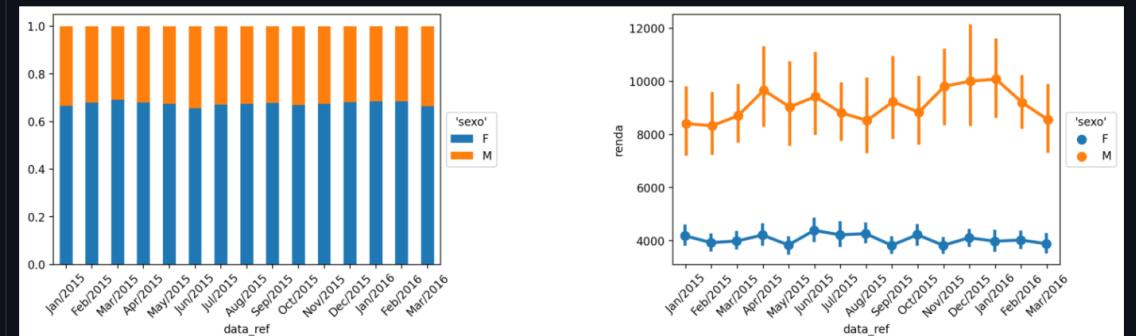
Apesar de não ser tão alta, a correlação entre a variável `tempo_emprego` e a variável `renda` revela claramente uma covariância positiva, evidenciada pela inclinação da linha de tendência.

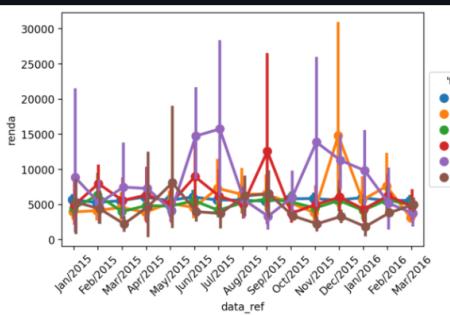
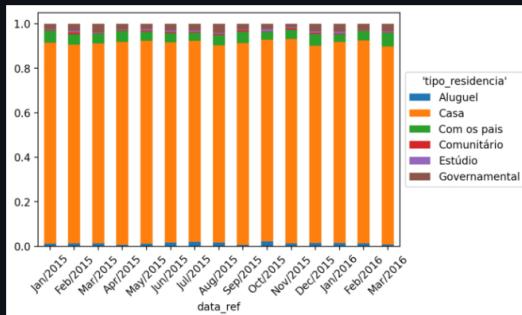
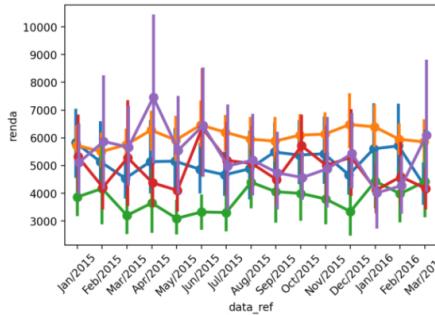
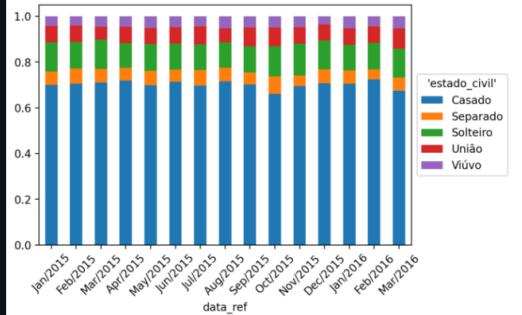
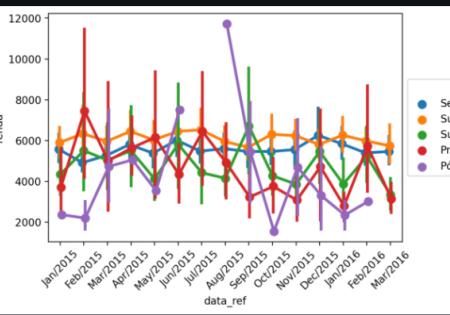
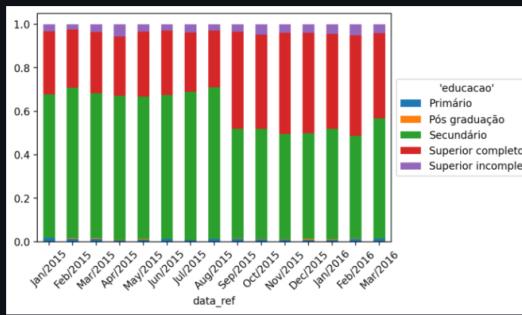
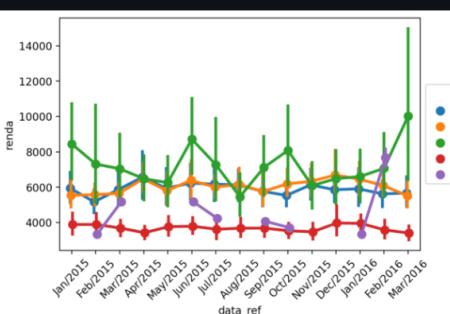
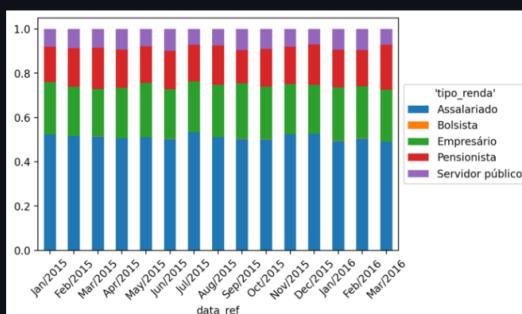
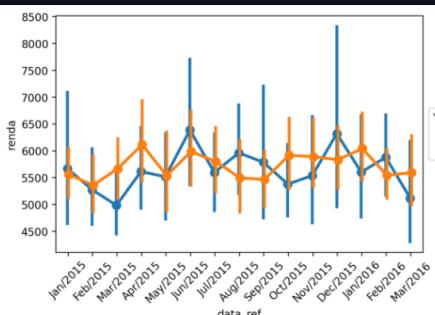
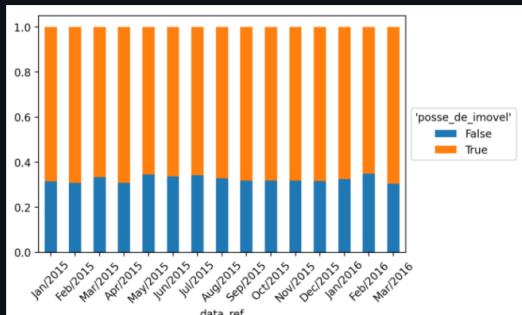
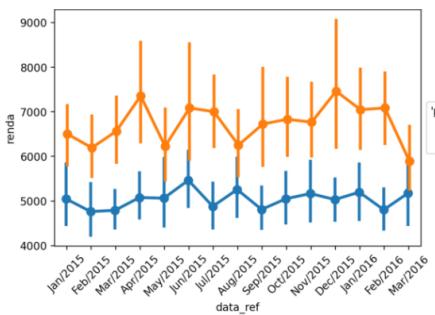
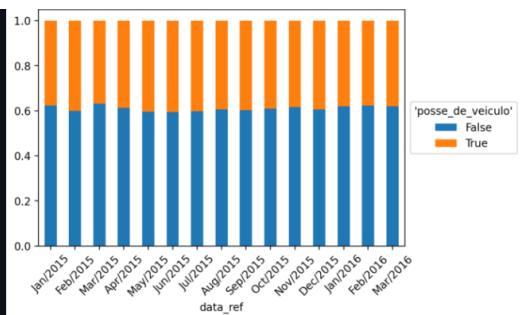
Análise das variáveis qualitativas



Ao contrastar os gráficos acima, é evidente que a variável `posse_de_veiculo` é mais relevante na predição de renda, como indicado pela maior disparidade entre os intervalos de confiança para aqueles que possuem e não possuem veículo. Em contrapartida, a variável `posse_de_imovel` não mostra diferença significativa entre as diferentes condições de posse imobiliária.

Análise das variáveis qualitativas ao longo do tempo



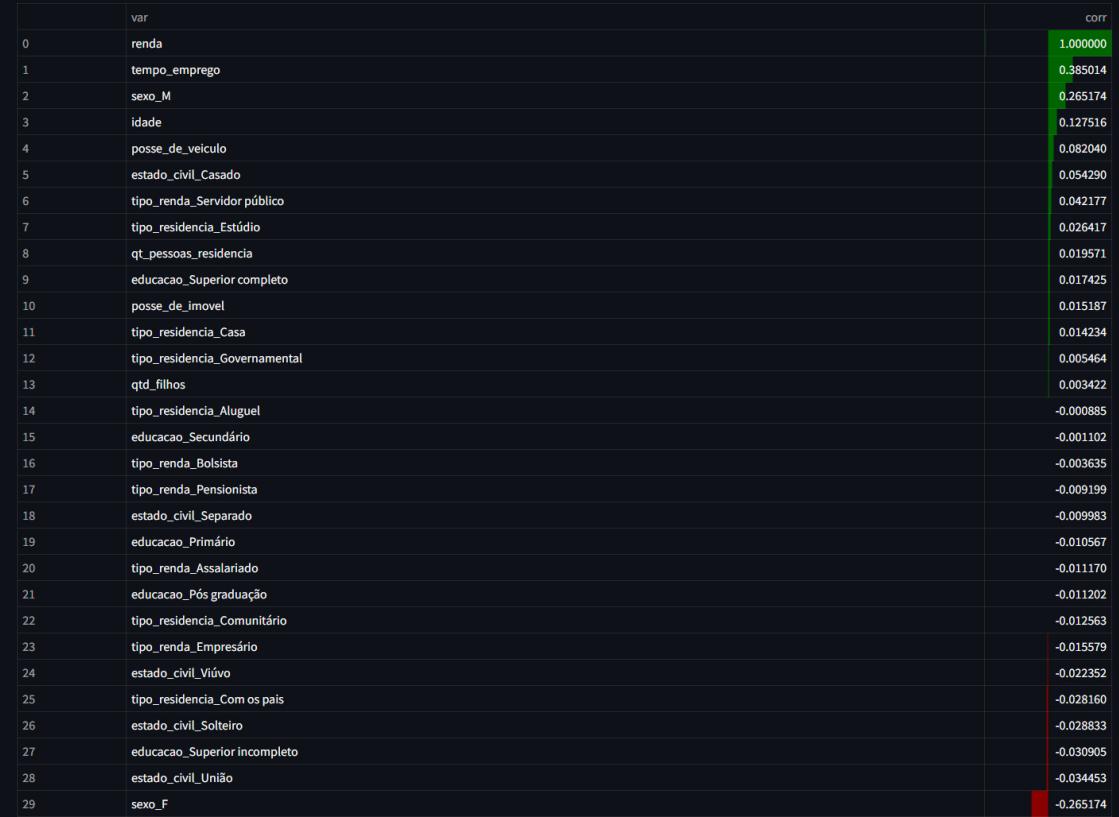


Etapa 3 Crisp-DM: Preparação dos dados

	tipos_dados	qtd_valores	qtd_categorias
sexo	object	12090	2
posse_de_veiculo	bool	12090	2
posse_de_imovel	bool	12090	2
qtd_filhos	int64	12090	8
tipo_renda	object	12090	5
educacao	object	12090	5
estado_civil	object	12090	5
tipo_residencia	object	12090	6
idade	int64	12090	46
tempo_emprego	float64	12090	2589
qt_pessoas_residencia	float64	12090	9
renda	float64	12090	8126

Conversão das variáveis categóricas em variáveis numéricas (dummies)

```
<class 'pandas.core.frame.DataFrame'>
Index: 12090 entries, 0 to 14592
Data columns (total 30 columns):
 #   Column          Non-Null Count Dtype  
---- 
 0   posse_de_veiculo    12090 non-null  bool    
 1   posse_de_imovel     12090 non-null  bool    
 2   qtd_filhos          12090 non-null  int64  
 3   idade               12090 non-null  int64  
 4   tempo_emprego       12090 non-null  float64 
 5   qt_pessoas_residencia 12090 non-null  float64 
 6   renda               12090 non-null  float64 
 7   sexo_F              12090 non-null  bool    
 8   sexo_M              12090 non-null  bool    
 9   tipo_renda_Assalariado 12090 non-null  bool    
 10  tipo_renda_Bolsista   12090 non-null  bool    
 11  tipo_renda_Empresario 12090 non-null  bool    
 12  tipo_renda_Pensionista 12090 non-null  bool    
 13  tipo_renda_Servidor público 12090 non-null  bool    
 14  educacao_Prímário    12090 non-null  bool    
 15  educacao_Pós graduação 12090 non-null  bool    
 16  educacao_Secundário   12090 non-null  bool    
 17  educacao_Superior completo 12090 non-null  bool    
 18  educacao_Superior incompleto 12090 non-null  bool    
 19  estado_civil_Casado   12090 non-null  bool    
 20  estado_civil_Separado  12090 non-null  bool    
 21  estado_civil_Solteiro  12090 non-null  bool    
 22  estado_civil_União    12090 non-null  bool    
 23  estado_civil_Viúvo    12090 non-null  bool    
 24  tipo_residencia_Aluguel 12090 non-null  bool    
 25  tipo_residencia_Casa   12090 non-null  bool    
 26  tipo_residencia_Com os pais 12090 non-null  bool    
 27  tipo_residencia_Comunitário 12090 non-null  bool    
 28  tipo_residencia_Estúdio  12090 non-null  bool    
 29  tipo_residencia_Governamental 12090 non-null  bool    
dtypes: bool(25), float64(3), int64(2)
memory usage: 861.9 KB
```



Etapa 4 Crisp-DM: Modelagem

Optamos pelo DecisionTreeRegressor como técnica, dada sua aptidão para lidar com problemas de regressão, como a previsão de renda dos clientes. Além disso, as árvores de decisão são de fácil interpretação e possibilitam a identificação dos atributos mais relevantes para a previsão da variável-alvo, o que a torna uma escolha sólida para o projeto.

Divisão da base em treino e teste

Quantidade de linhas e colunas de X: (12896, 29)

Quantidade de linhas de y: 12896

X_train: (9606, 29)

X_test: (3283, 29)

y_train: (9606,)

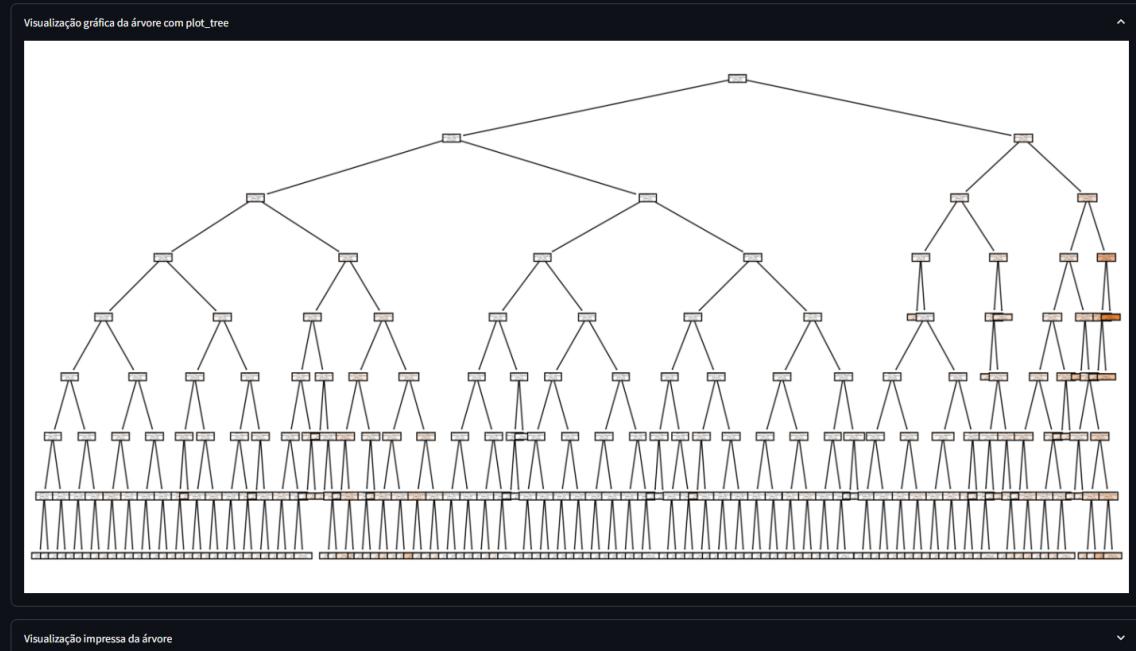
y_test: (3283,)

Seleção de hiperparâmetros do modelo com for loop

	max_depth	min_samples_leaf	score
213	8	4	0.4171
543	19	4	0.4011
243	9	4	0.401
393	14	4	0.4002
245	9	6	0.3994
276	10	7	0.3986
483	17	4	0.3984
275	10	6	0.3984
453	16	4	0.3979
363	13	4	0.3974

Rodando o modelo

```
DecisionTreeRegressor(max_depth=8, min_samples_leaf=4, random_state=42)
```



Etapa 5 Crisp-DM: Avaliação dos resultados

O coeficiente de determinação (R^2) da árvore com profundidade = 8 para a base de treino é: 0,60

O coeficiente de determinação (R^2) da árvore com profundidade = 8 para a base de teste é: 0,42

	renda	renda_predict
0	8,060,34	3,127,79
1	1,852,15	7,146,41
2	2,253,89	2,465,37
3	6,600,77	3,654,23
4	6,475,97	5,465,31
5	1,445,87	3,654,23
6	1,726,03	3,654,23
7	2,515,98	2,512,87
8	3,420,34	4,992,07
9	12,939,14	30,365,34

Etapa 6 Crisp-DM: Implantação

Neste cenário foi desenvolvida uma calculadora simples para a simulação das condições de crédito, baseando-se em todo o entendimento que foi feito sobre este modelo ao longo da trajetória da análise.

