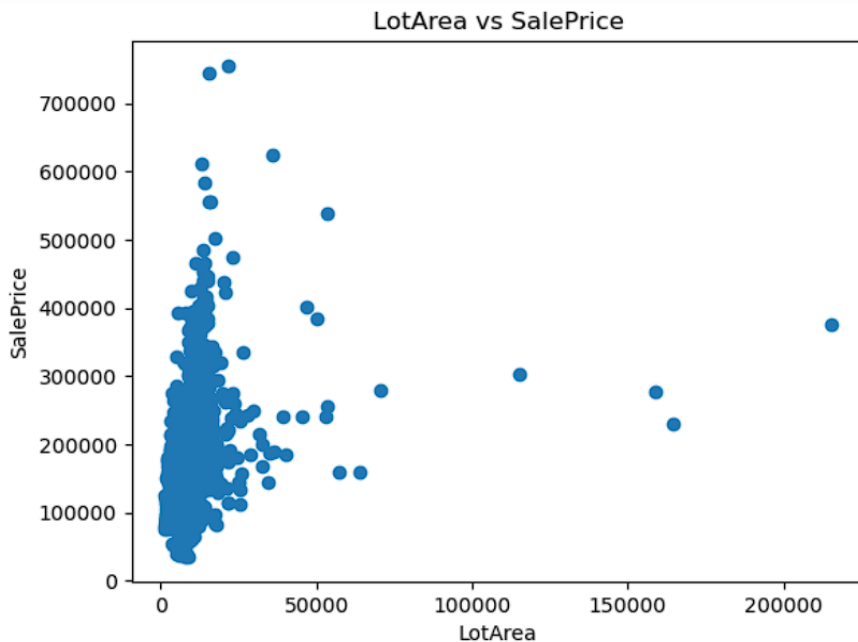
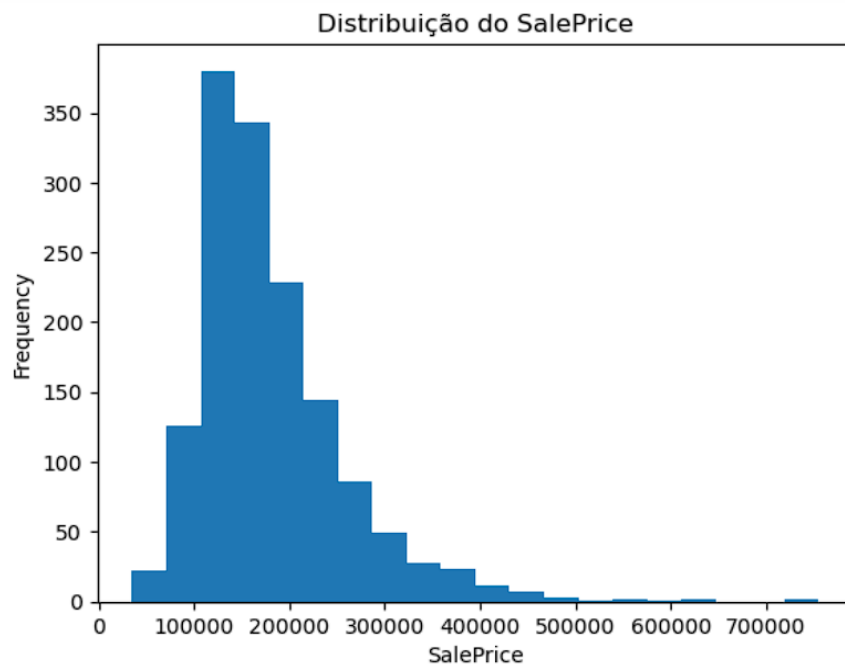


Análise de Dados

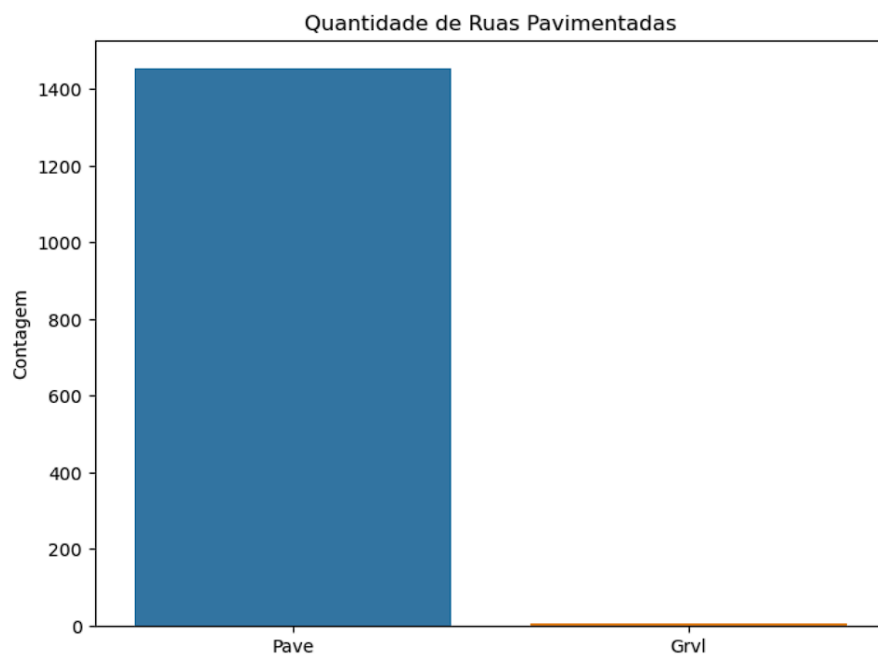
Análise Exploratória:



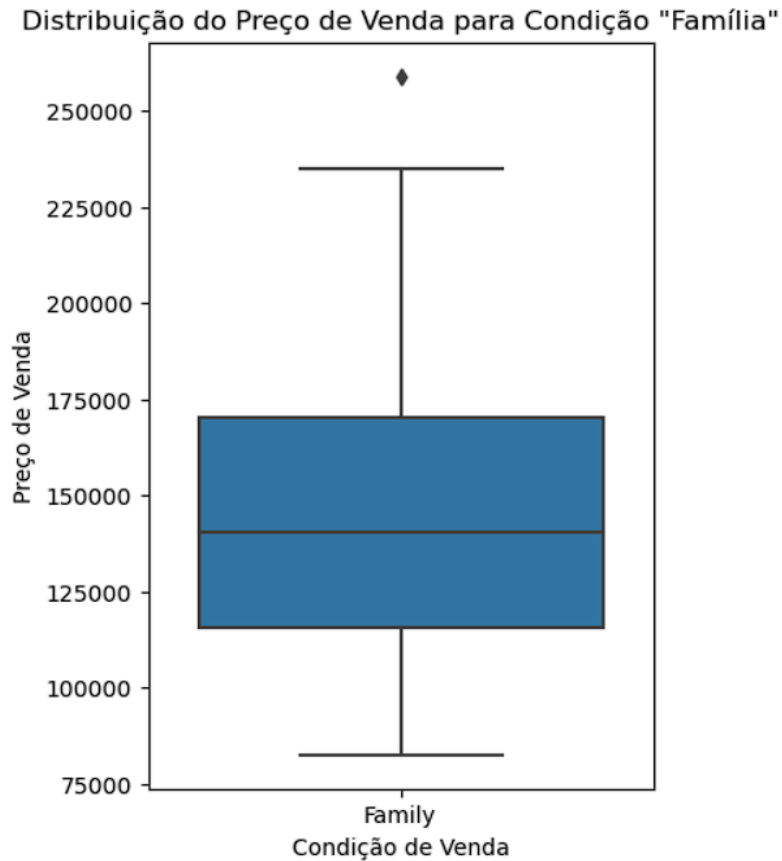
Análise exploratória: Segundo a análise feita em cima do gráfico de dispersão acima, podemos notar que o tamanho do lote não, necessariamente, influencia no preço do terreno já que vários lotes de tamanhos parecidos tem uma grande variância de preço. Além disso, é visível a existência de outliers acima de 50000m² e também nos lotes acima de \$500,000.



Análise exploratória: De acordo com o histograma acima, é notável que grande parte do valor dos lotes se concentram entre \$100,000 e \$250,000. Por fim, percebe-se a existência de outliers a partir do valor de \$500,000.



Análise exploratória: De acordo com o gráfico de barras acima, percebe-se que a grande parte dos lotes se encontram em ruas pavimentadas.



Análise exploratória: Através do boxplot gerado, percebe-se que o preço mais alto vendido entre familiares foi um pouco acima de \$250,000 e o mais baixo um pouco acima de \$75,000, contendo 1 outlier com uma venda entre familiares acima de \$250,000

Featuring Engineering:

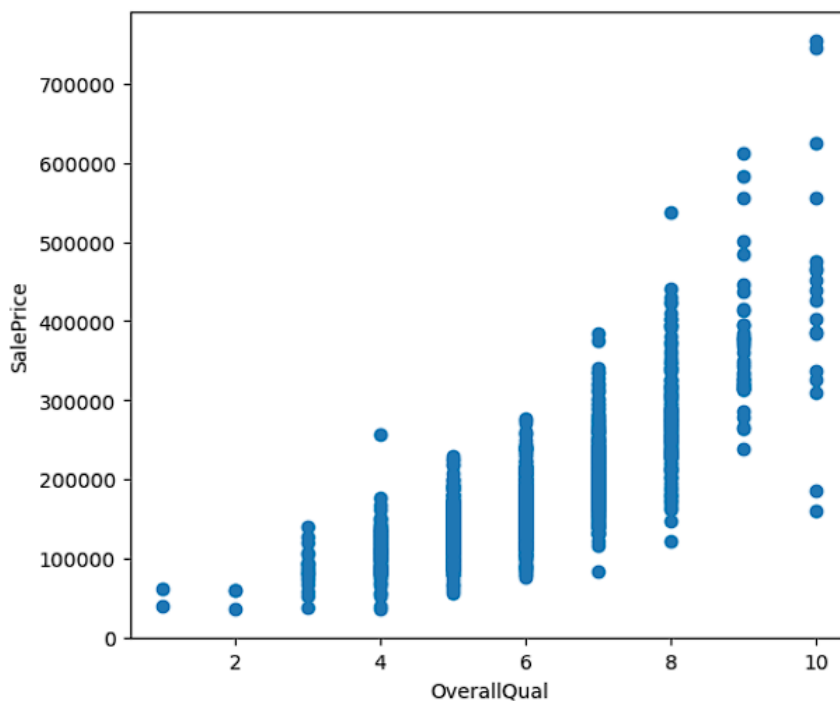
Variáveis mais importantes, para o treinamento de máquina:

```
id Feature Score
4 OverallQual 2436.770591 -> Qualidade da casa e o acabamento geral
16 GrLivArea 1470.585010 -> Área habitavel acima do solo (em pés quadrados)
26 GarageCars 1013.705666 -> Quantidade de carros que cabem na garagem
27 GarageArea 926.951287 -> Tamanho da garagem (em metros quadrados)
12 TotalBsmtSF 880.341282 -> Tamanho do porão (em metros quadrados)
13 1stFlrSF 845.524488 -> Tamanho do primeiro andar (em pés quadrados)
19 FullBath 668.430296 -> Banheiros completos acima do nível
23 TotRmsAbvGrd 580.762801 -> Total de quartos acima do nível
6 YearBuilt 548.665821 -> Data que a casa foi construída
7 YearRemodAdd 504.714855 -> Data de remodelação da casa
```

Foi utilizando um mapa de calor para observar as variáveis que tinham maior correlação linear, mostrando que aquelas com maior número de correção linear, afeta de forma maior as variáveis dependentes

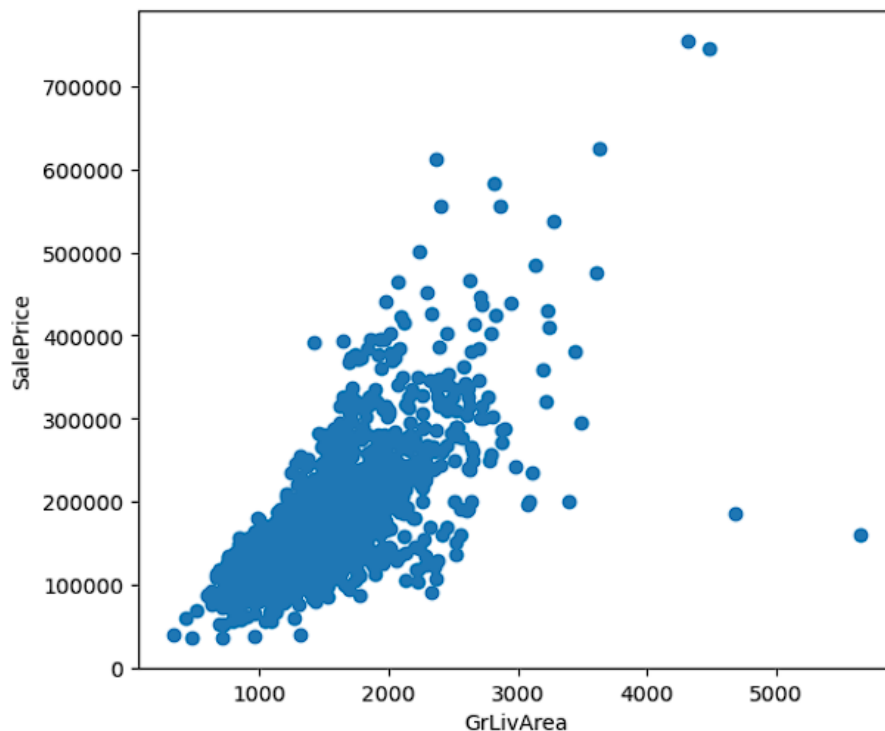
Além disso, utilizamos um gráfico de dispersão com a variável dependente SalePrice e a variável independente OverallQual que classifica o material geral e o acabamento da casa em notas de 1 a 10, sendo 10 muito boa e 1 muito ruim.

Com isso, é possível chegar a conclusão de que a qualidade do material geral da casa e o acabamento influenciam de certa forma no preço da casa, ainda mais quando a nota começa a chegar de 6 para cima. Além disso, o seu SCORE no mapa de calor foi bem alto, chegando a 0.79 de 1, mostrando que é uma variável muito importante.



Em segundo plano, utilizamos de outro gráfico de dispersão para mostrar o quão importante é a variável GrLivArea, que indica a área habitável acima do solo (pés quadrados) de uma casa, no qual é nitidamente perceptível, normalmente, que o nível da casa em relação ao solo influencia também diretamente no preço da casa, tornando ela, uma variável independente muito importante. Porém, existem alguns outliers tanto positivamente quanto negativamente, podemos perceber que existem 2 pontos fora da curva para a altura da casa, que mesmo ela sendo alta, continua com o preço baixo e servindo também para os 2 pontos outliers que, chegaram num

nível altíssimo de preço para casa que ultrapassem 3000 pés quadrados, chegando a mais de \$700,000.



Por fim, foram plotados outros gráficos de dispersão, junto com o mapa de calor, mostrando as variáveis independentes que são importantes para a variável dependente (SalePrice) e chegamos as seguintes conclusões para as variáveis importantes no Feature Engineering:

	Variável	Pontuação
4	OverallQual	2436.770591
16	GrLivArea	1470.585010
26	GarageCars	1013.705666
27	GarageArea	926.951287
12	TotalBsmtSF	880.341282
13	1stFlrSF	845.524488
19	FullBath	668.430296
23	TotRmsAbvGrd	580.762801
6	YearBuilt	548.665821
7	YearRemodAdd	504.714855

OverallQual: Classifica o material geral e o acabamento da casa em notas de 1 a 10, sendo 10 muito boa e 1 muito ruim.

GrLivArea: Indica a área habitável acima do solo (pés quadrados) de uma casa

GarageCars: Quantidade de carros que a garagem comporta

GarageArea: Tamanho da garagem em metros quadrados

TotalBsmstSF: Tamanho da área do porão em pés quadrados

1stFlrSF: Tamanho do primeiro andar em pés quadrados

FullBath: Banheiros completos acima do nível do solo

TotRmsAbvGrd: Total de quartos acima do primeiro nível

YearBuilt: Ano em que a casa foi construído

YearRemodAdd: Data de remodelação da casa

Aprendizagem não supervisionada:

Primeiro foram selecionadas as variáveis independentes mais importantes para começar o treinamento de máquina utilizando a Regressão Linear, em seguida dividimos o conjunto de dados em treino e teste, por fim, geramos as métricas MSE e R^2

Erro Médio Quadrático (MSE): 1558239575.1510978
Coeficiente de Determinação (R^2): 0.7968483304281087

O coeficiente de determinação diz que em 79% dos casos o modelo conseguiu prever o preço das casas de forma confiável, sendo um bom número.

Contudo, logo em seguida usamos o Random Forest, para observar se ele iria prever melhor os preços das casas o resultados obtidos foram:

Métricas de avaliação do modelo:
Erro Médio Quadrático (MSE): 877299148.1199831
Coeficiente de Determinação (R^2): 0.8856242714556324

É nítido que o R^2 foi maior, mostrando que o random forest, nesse caso, possui um maior acerto nos preços das casas.

Ainda na aprendizagem não supervisionada, usamos o modelo de classificação Regressão Logística e, obtivemos os seguintes resultados:

1. A acurácia do modelo é de aproximadamente 89.4%. Isso significa que o modelo de classificação tem uma boa capacidade de prever corretamente se uma casa será vendida por um preço alto ou baixo, com uma taxa de acerto de cerca de 89.4%.

2. A matriz de confusão mostra que os valores na diagonal principal representam os acertos do modelo, enquanto os valores fora da diagonal são os erros. Portanto, na classe dos preços baixos, o modelo classificou corretamente 142 casas como preços baixos (verdadeiros negativos) e 19 casas como preços altos (falsos positivos). Por outro lado, na classe dos preços altos, o modelo classificou corretamente 119 casas como preços altos (verdadeiros positivos) e 12 casas como preços baixos (falsos negativos).