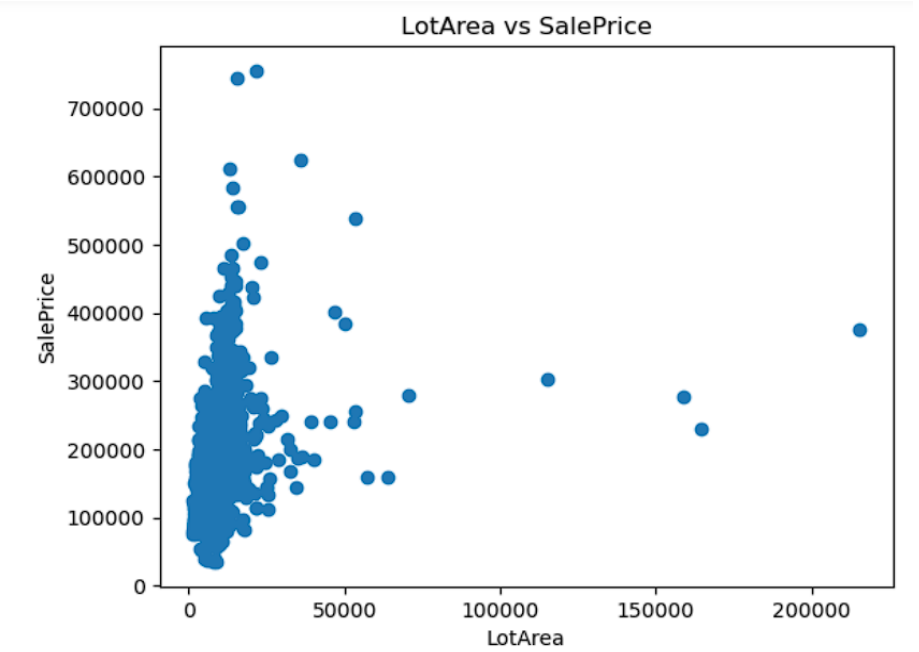
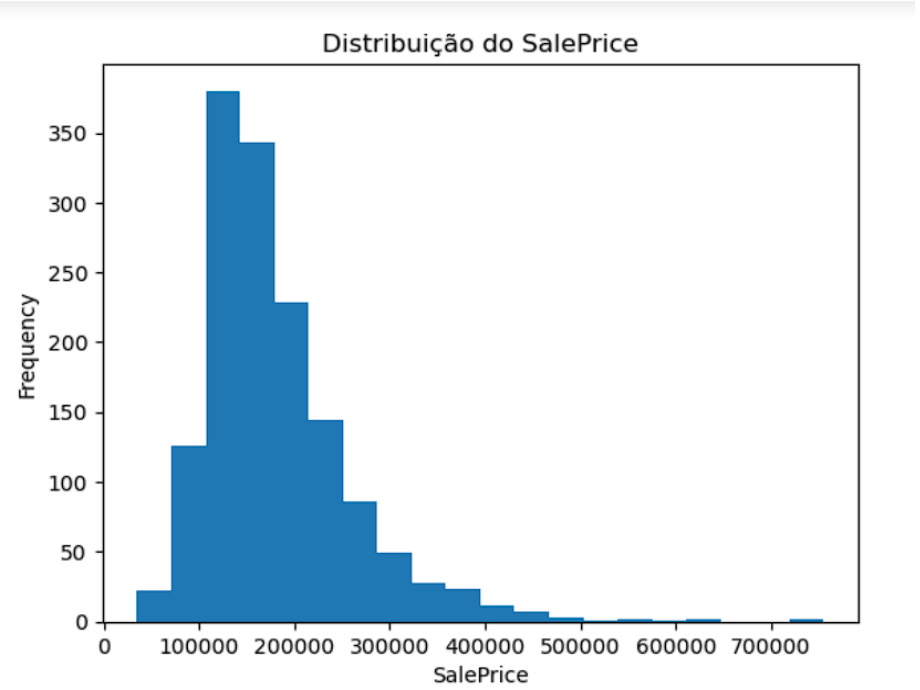
Análise de Dados

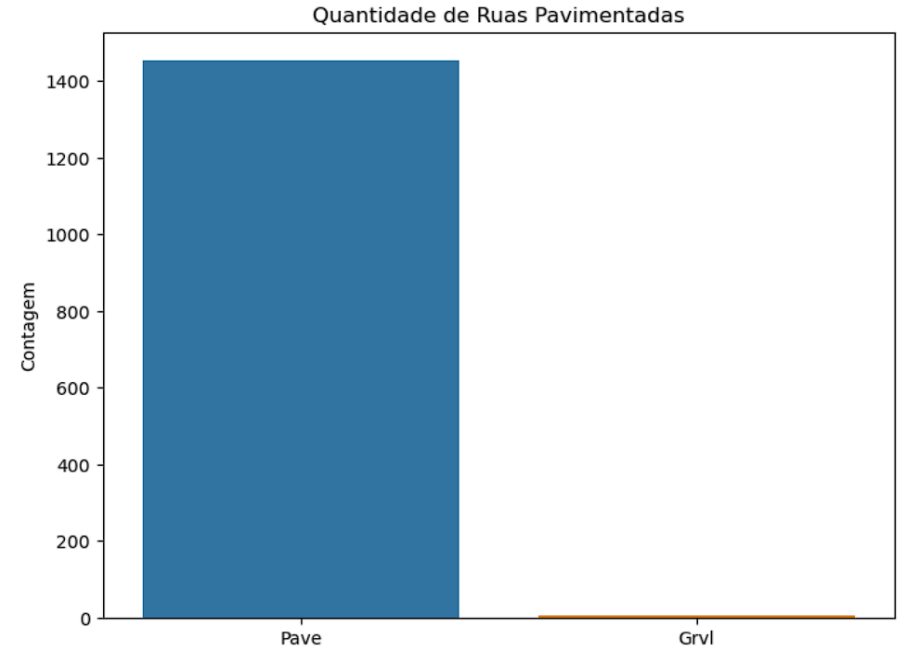
Análise Exploratória:



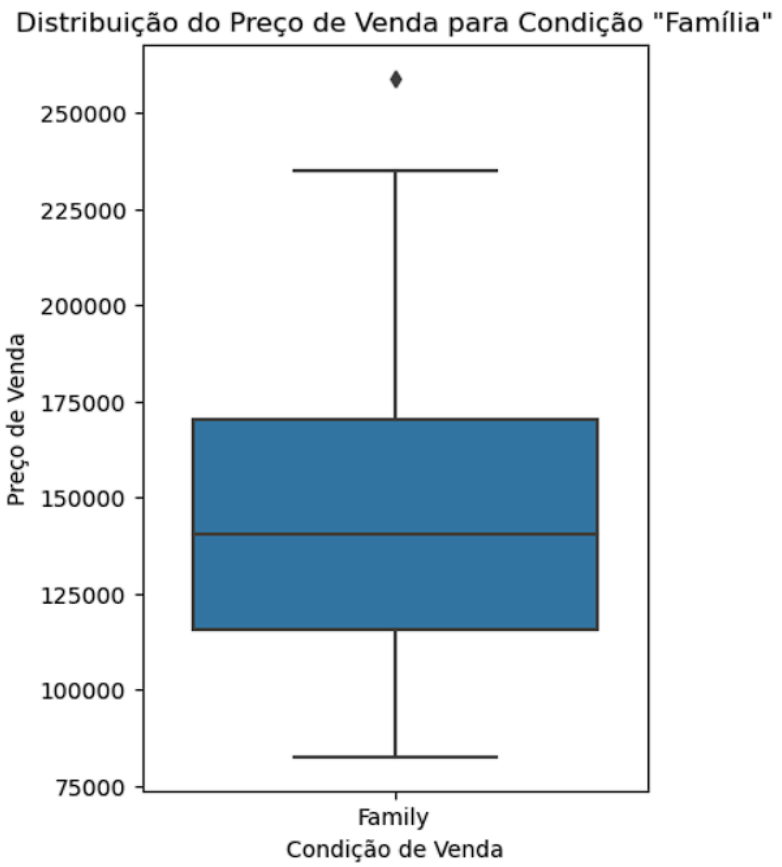
Análise exploratória: Segundo a análise feita em cima do gráfico de dispersão acima, podemos notar que o tamanho do lote não, necessariamente, influencia no preço do terreno já que vários lotes de tamanhos parecidos tem uma grande variância de preço. Além disso, é visível a existência de outliers acima de 50000m^2 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 famíliares 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

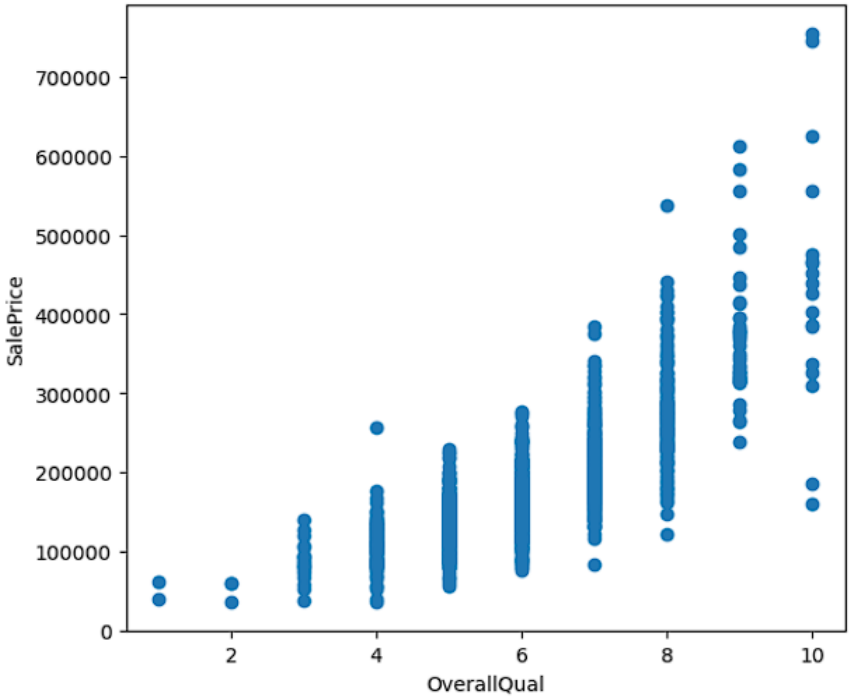
Featuring Engineering:

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 construida
7 YearRemodAdd 504.714855 -> Data de remodelação da casa
```

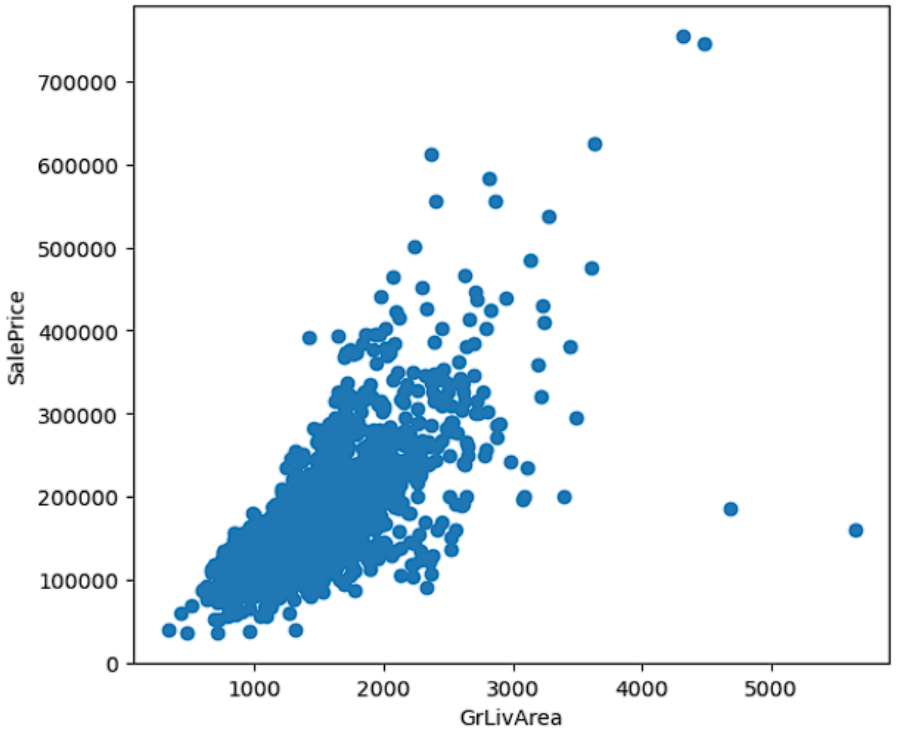
Foi utlizando um mapa de calor para observar as variáveis que tinham maior correlação linear, mostrando que aquelas com maior número de correção linear, afeta de forma maior as variáveis dependentes

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 influência 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 altissimo 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 independente que são importantes para a variável dependente (SalePrice) e chegamos as seguintes conclusões para as variáveis importantes no Feature Enginnering:

	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

TotalBsmtSF: 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 construido
YearRemodAdd: Data de remodelação da casa

Aprendizagem 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 obsevar 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 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).

Métricas de avaliação

- **Comparação entre os modelos de regressão: REGRESSÃO LINEAR X RANDOM FOREST**

Erro Médio Quadrático (MSE):
- Regressão Linear: 1.558.239.575,15
- Random Forest: 877.299.148,12

O modelo de Random Forest apresentou um MSE menor do que o modelo de Regressão Linear. Isso indica que o modelo de Random Forest teve um desempenho melhor na redução do erro médio quadrático em relação aos valores reais.

Erro Absoluto Médio (MAE):
- Regressão Linear: 24.774,22
- Random Forest: 19.033,52

O modelo Random Forest obteve um MAE menor do que o modelo de Regressão Linear. Isso significa que o modelo Random Forest teve uma menor diferença média absoluta entre as previsões e os valores reais em comparação com o modelo de Regressão Linear, sendo superior em relação ao modelo de regressão linear.

Coeficiente de Determinação (R^2):
- Regressão Linear: 0,7968
- Random Forest: 0,8856

Conclusão: O modelo de Random Forest apresentou um R^2 maior do que o modelo de Regressão Linear. Isso mostra que o modelo Random Forest explicou uma maior proporção da variabilidade dos dados em relação à variável dependente

(SalePrice) em comparação com o modelo de Regressão Linear que apresentou um valor um pouco a baixo, cerca de 79% de acerto. Portanto, com base nessas métricas, é visível que o modelo Random Forest obteve um desepenho superior ao modelo de Regressão Linear em todos os quesitos da previsão do preço de venda (SalePrice). Com isso, apresentou um menor erro médio quadrático, um menor erro absoluto médio e uma melhor capacidade de explicar a variância dos dados.

Observações: Alguns pontos devem ser observados, para entender o porque o modelo Random Forest foi superior ao modelo de Regressão linear, são eles:

1. **Capacidade de capturar relações não lineares:** Capacidade de capturar relações não lineares entre a variável dependente e as variáveis independentes, fazendo com que ele gere árvores de decisão, combinando suas previsões e permitindo capturar padrões complexos nos dados. Por outro lado, a regressão linear pressupõe uma relação linear entre as variáveis (dependente e independentes), o que pode limitar a sua capacidade de modelagem dos dados fornecidos.
2. **Segurança a outliers:** O random forest é menos sensível a outliers do que a regressão linear. Isso ocorre porque o random forest divide o seu espaço em várias regiões e toma uma média da previsão para cada árvore criada, reduzindo, assim, o impacto desses outliers.

- **Entendimento sobre as métricas geradas pela Regressão Logística**

Acurácia: Ela permite entender a quantidade de acertos que o modelo obteve, sendo calculado, dividindo o número de predições corretas, ou seja, quantas previsões o modelo acertou, pelo total de instâncias avaliadas, sendo assim a quantidade de dados que foi fornecido para o modelo trabalhar. Com isso, o nosso modelo obteve uma acurácia de 89,4%, tendo uma taxa de acertos bem alta em relação aos dados fornecidos.

Precisão: A precisão é uma métrica que mede a proporção de exemplos classificados como positivos (verdadeiros positivos) que são realmente positivos em relação ao total de exemplos classificados como positivos (verdadeiros positivos + falsos positivos). Por isso, ela tem o papel de avaliar a capacidade do modelo de evitar a classificação incorreta de exemplos negativos como positivos (falsos positivos). Diante disso, o nosso modelo teve uma precisão de 86.2%, indicando que teve uma baixa taxa de falsos positivos e classificou bem as classe de "alta faixa de preço".

Recall: O recall tem o papel de avaliar a capacidade do modelo de identificar corretamente exemplos positivos. Com isso, nosso modelo ovteve um recall de 90.8%, mostrando uma baixa taxa de falsos negativos.

Matriz de confusão: A matriz de confusão mostra que o modelo teve 142 predições corretas para a classe "baixa faixa de preço" (verdadeiros negativos), 119 predições corretas para a classe "alta faixa de preço" (verdadeiros positivos), 19 falsos positivos e 12 falsos negativos.

- **Entendimento sobre as métricas geradas pela Regressão Logística**

n_clusters: Foram utilizadas 2 técnicas conhecidas como Método da Silhueta e Método do cotovelo, para descobrir o melhor valor para n_clusters e foi observado que o melhor valor encontra é de 2, isso significa que a curva das somas dos quadradados intra-cluster apresentou uma queda acentuada de até 2 clusters e que, a partir desse ponto, a diminuição na soma dos quadrados foi menor acentuada.