Housing Prices

Jacob Thomas

Goals

- Predict the sales price of a home
- Find meaningful features through data exploration and feature selection
- Find optimal hyperparameters
- Build multiple regression models – simple and advanced; Evaluate using RMSE and MAE

About the data

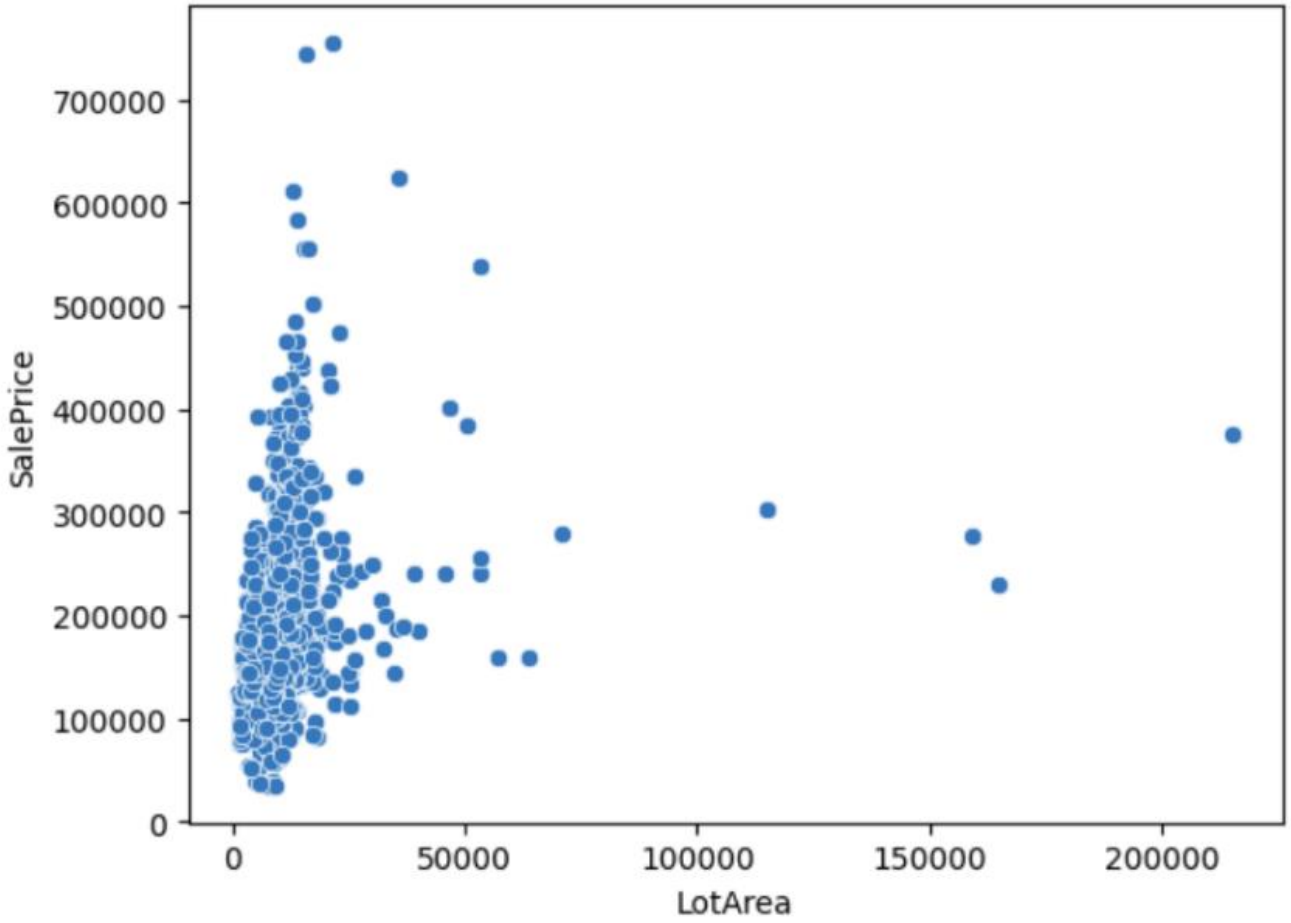
- 1460 rows
- 81 columns

- SalePrice - the property's sale price in dollars. This is the target variable.
- MSSubClass: The building class
- MSZoning: The general zoning classification
- LotFrontage: Linear feet of street connected to property
- LotArea: Lot size in square feet
- Street: Type of road access
- Alley: Type of alley access
- LotShape: General shape of property
- LandContour: Flatness of the property
- Utilities: Type of utilities available
- LotConfig: Lot configuration
- LandSlope: Slope of property
- Neighborhood: Physical locations within Ames city limits
- Condition1: Proximity to main road or railroad
- Condition2: Proximity to main road or railroad (if a second is present)
- BldgType: Type of dwelling
- HouseStyle: Style of dwelling
- OverallQual: Overall material and finish quality
- OverallCond: Overall condition rating
- YearBuilt: Original construction date
- YearRemodAdd: Remodel date
- RoofStyle: Type of roof
- RoofMatl: Roof material
- Exterior1st: Exterior covering on house
- Exterior2nd: Exterior covering on house (if more than one material)
- MasVnrType: Masonry veneer type
- MasVnrArea: Masonry veneer area in square feet

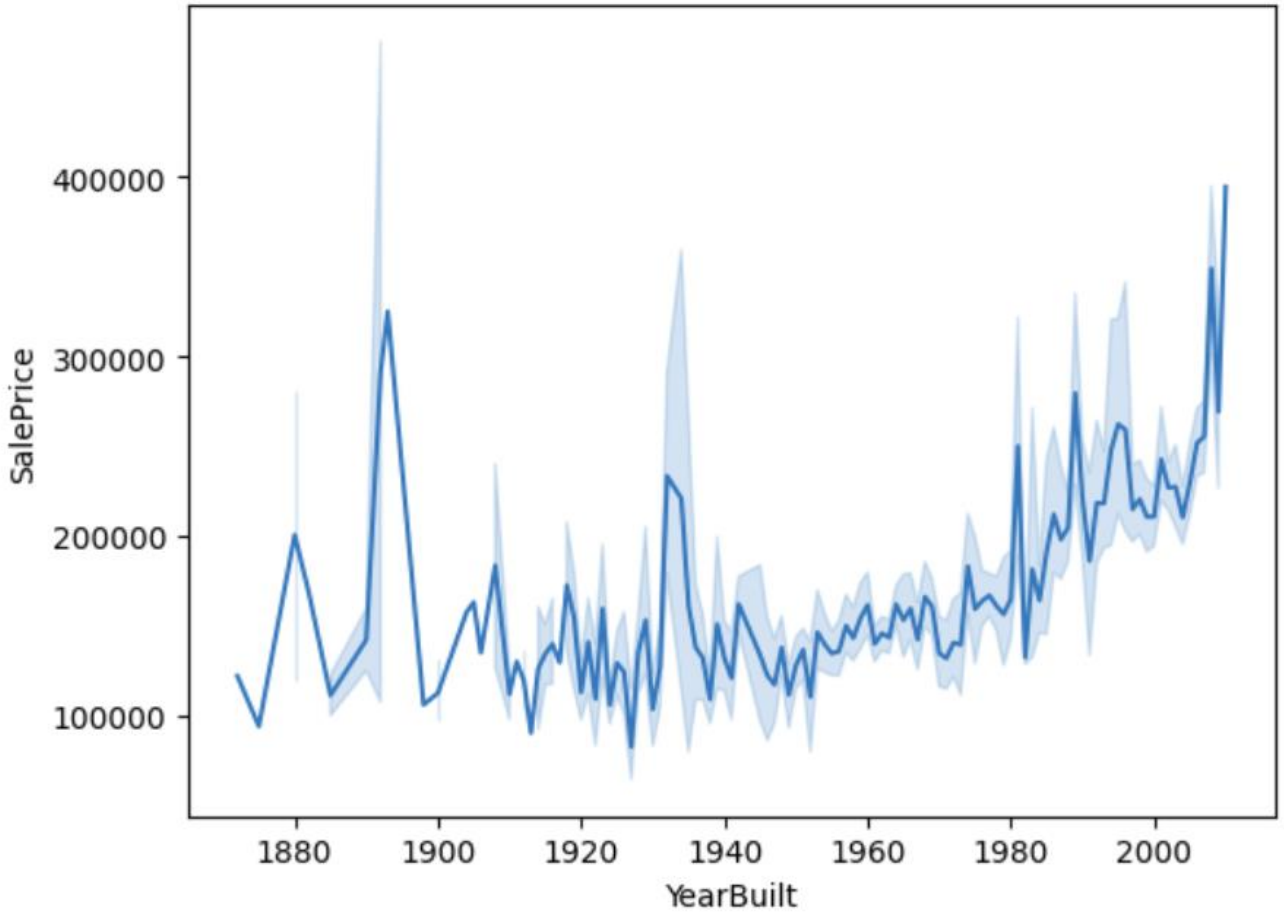
- [illegible]: Present condition of the material on the exterior
- Foundation: Type of foundation
- BsmtQual: Height of the basement
- BsmtCond: General condition of the basement
- BsmtExposure: Walkout or garden level basement walls
- BsmtFinType1: Quality of basement finished area
- BsmtFinSF1: Type 1 finished square feet
- BsmtFinType2: Quality of second finished area (if present)
- BsmtFinSF2: Type 2 finished square feet
- BsmtUnfSF: Unfinished square feet of basement area
- TotalBsmtSF: Total square feet of basement area
- Heating: Type of heating
- HeatingQC: Heating quality and condition
- CentralAir: Central air conditioning
- Electrical: Electrical system
- 1stFlrSF: First Floor square feet
- 2ndFlrSF: Second floor square feet
- LowQualFinSF: Low quality finished square feet (all floors)
- GrLivArea: Above grade (ground) living area square feet
- BsmtFullBath: Basement full bathrooms
- BsmtHalfBath: Basement half bathrooms
- FullBath: Full bathrooms above grade
- HalfBath: Half baths above grade
- Bedroom: Number of bedrooms above basement level
- Kitchen: Number of kitchens
- KitchenQual: Kitchen quality
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

- Fireplaces: Number of fireplaces
- FireplaceQu: Fireplace quality
- GarageType: Garage location
- GarageYrBlt: Year garage was built
- GarageFinish: Interior finish of the garage
- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage in square feet
- GarageQual: Garage quality
- GarageCond: Garage condition
- PavedDrive: Paved driveway
- WoodDeckSF: Wood deck area in square feet
- OpenPorchSF: Open porch area in square feet
- EnclosedPorch: Enclosed porch area in square feet
- 3SsnPorch: Three season porch area in square feet
- ScreenPorch: Screen porch area in square feet
- PoolArea: Pool area in square feet
- PoolQC: Pool quality
- Fence: Fence quality
- MiscFeature: Miscellaneous feature not covered in other categories
- MiscVal: $Value of miscellaneous feature
- MoSold: Month Sold
- YrSold: Year Sold
- SaleType: Type of sale
- SaleCondition: Condition of sale

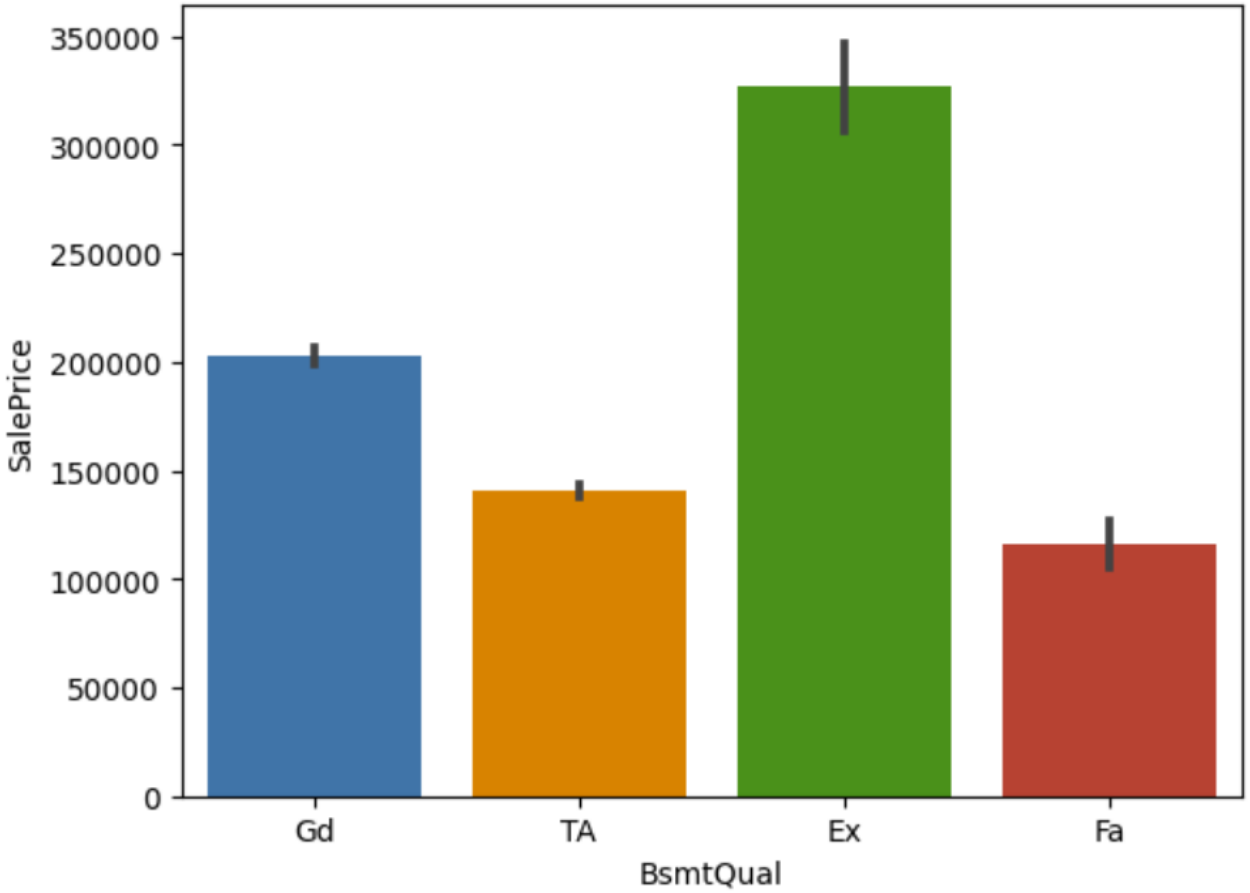
Lot area square feet vs Sale Price



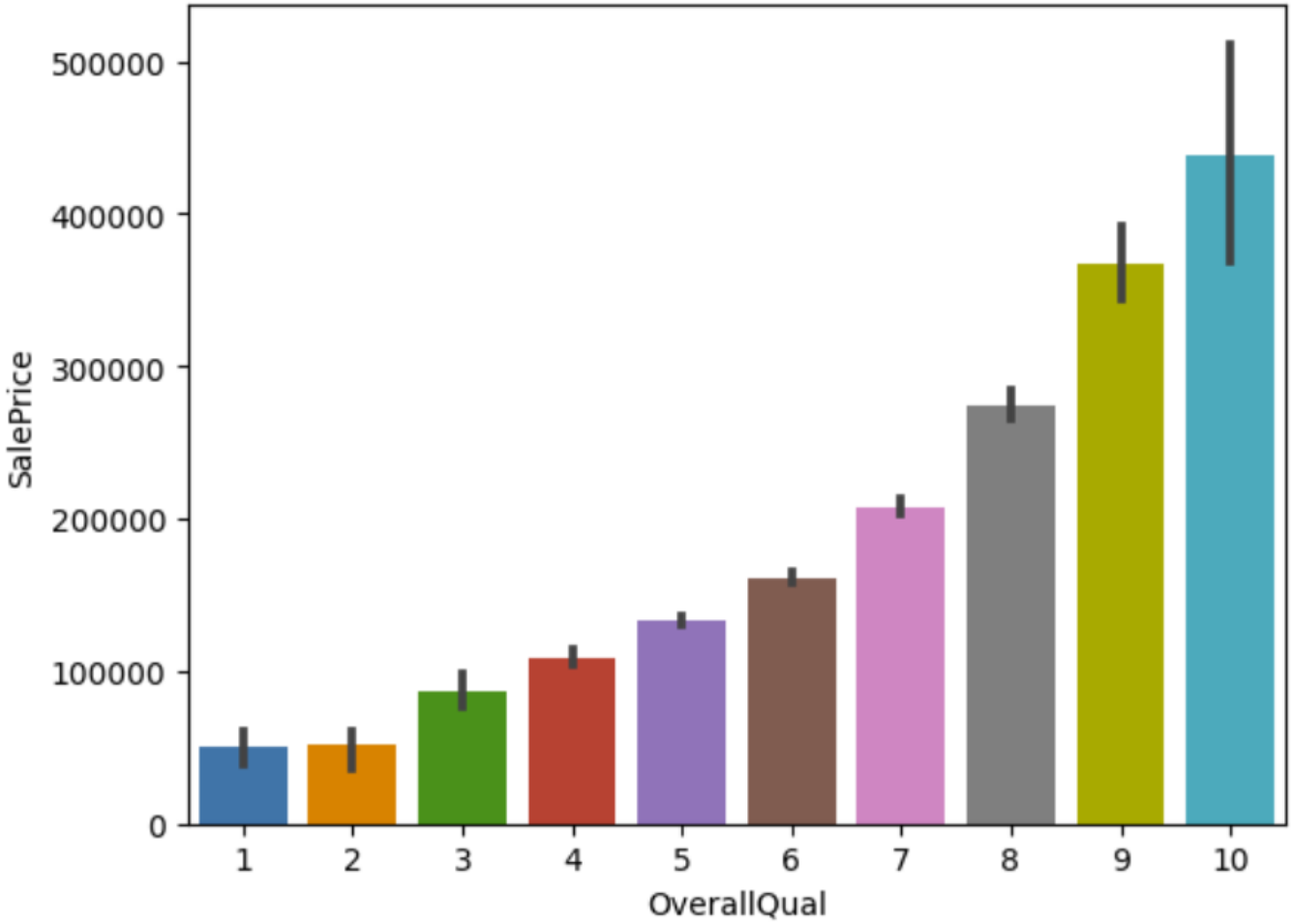
Year Built vs Sale Price



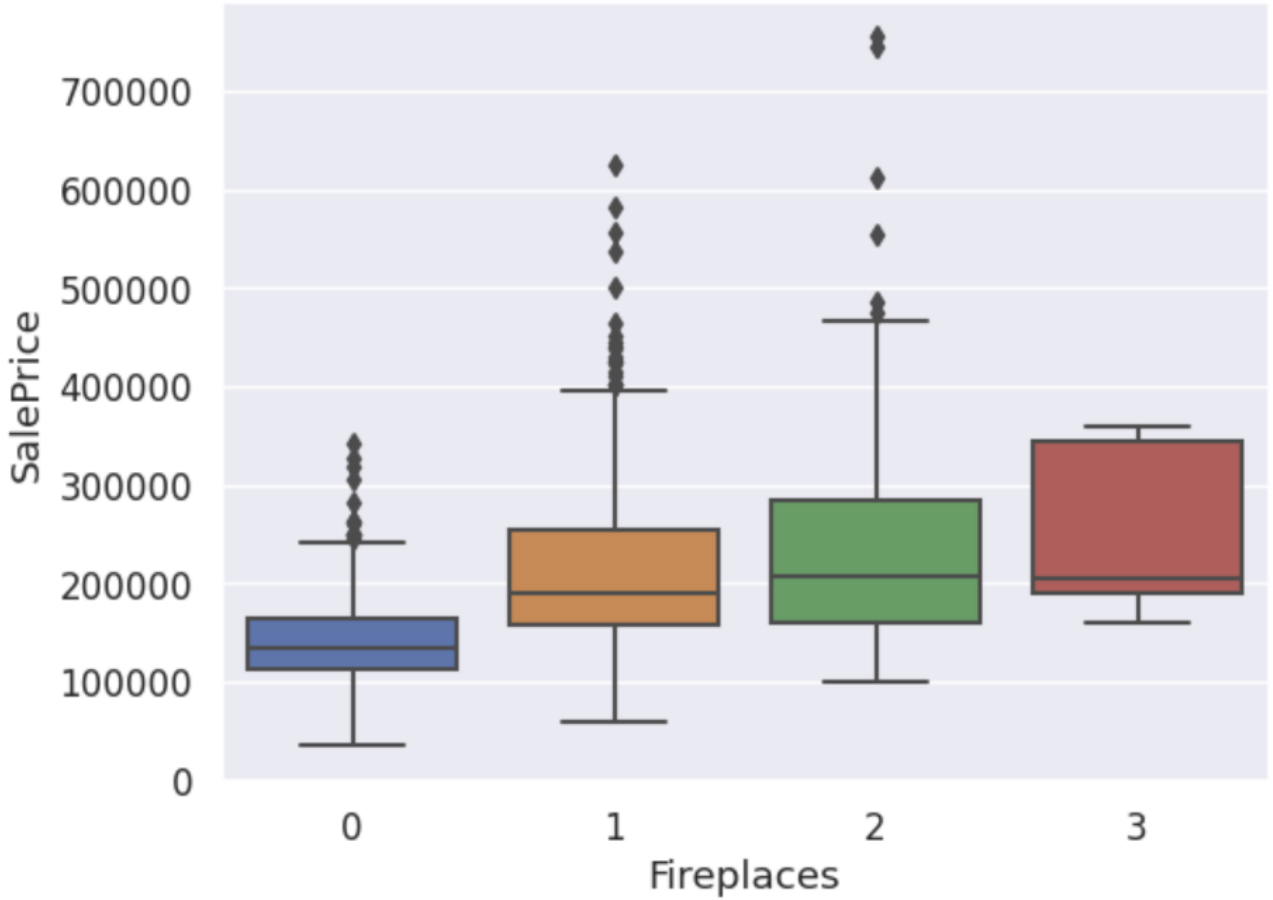
Height of basement vs Sale Price



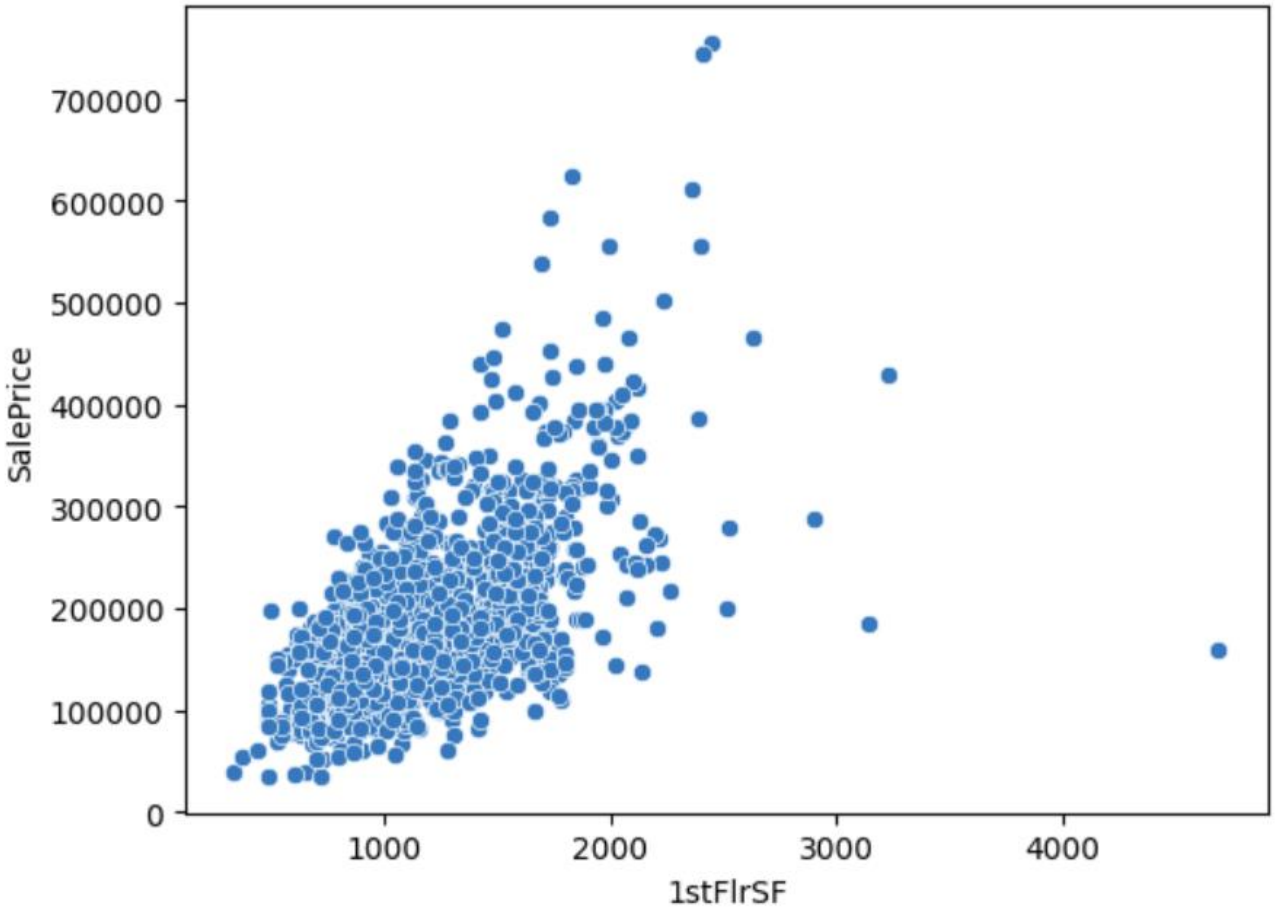
Overall quality of house vs Sale Price



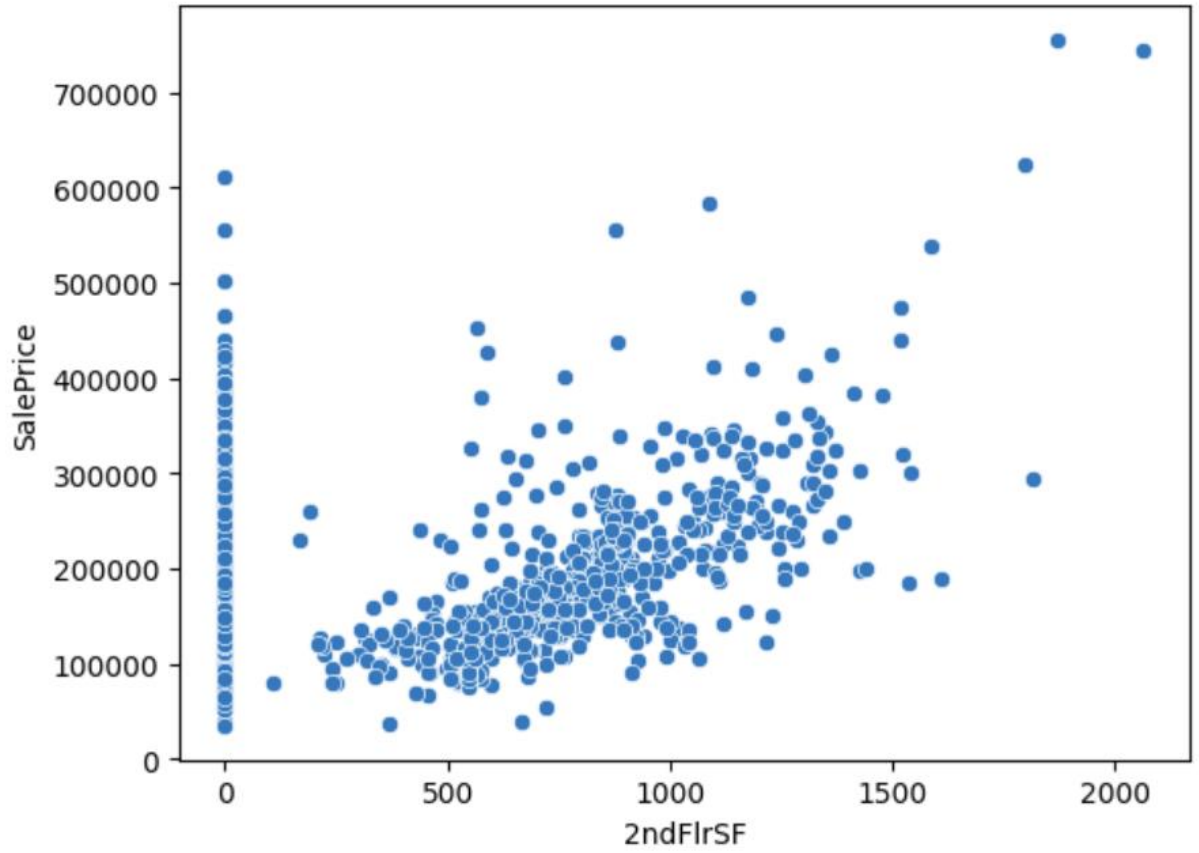
Number of fireplaces vs Sale Price



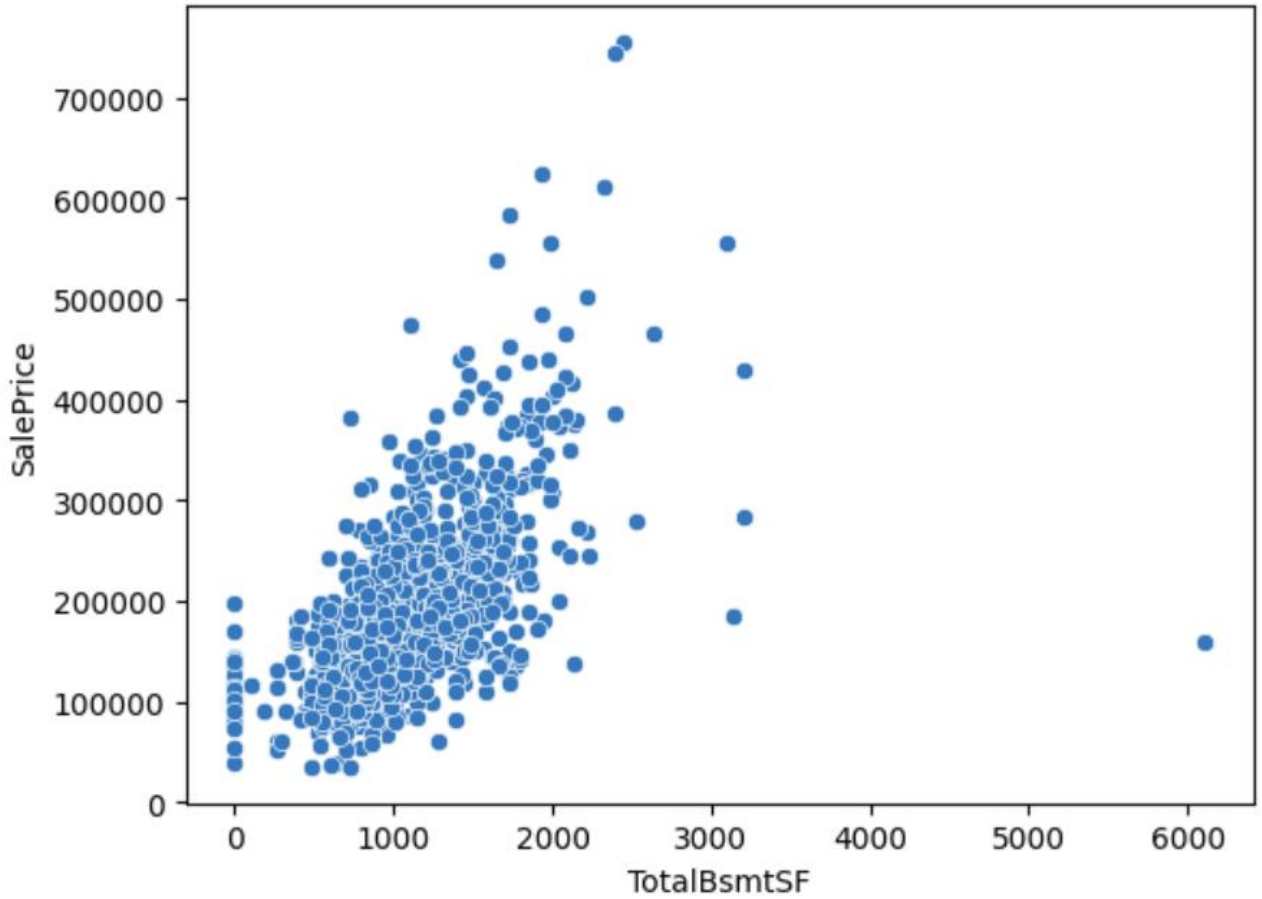
1st floor square feet vs Sale Price



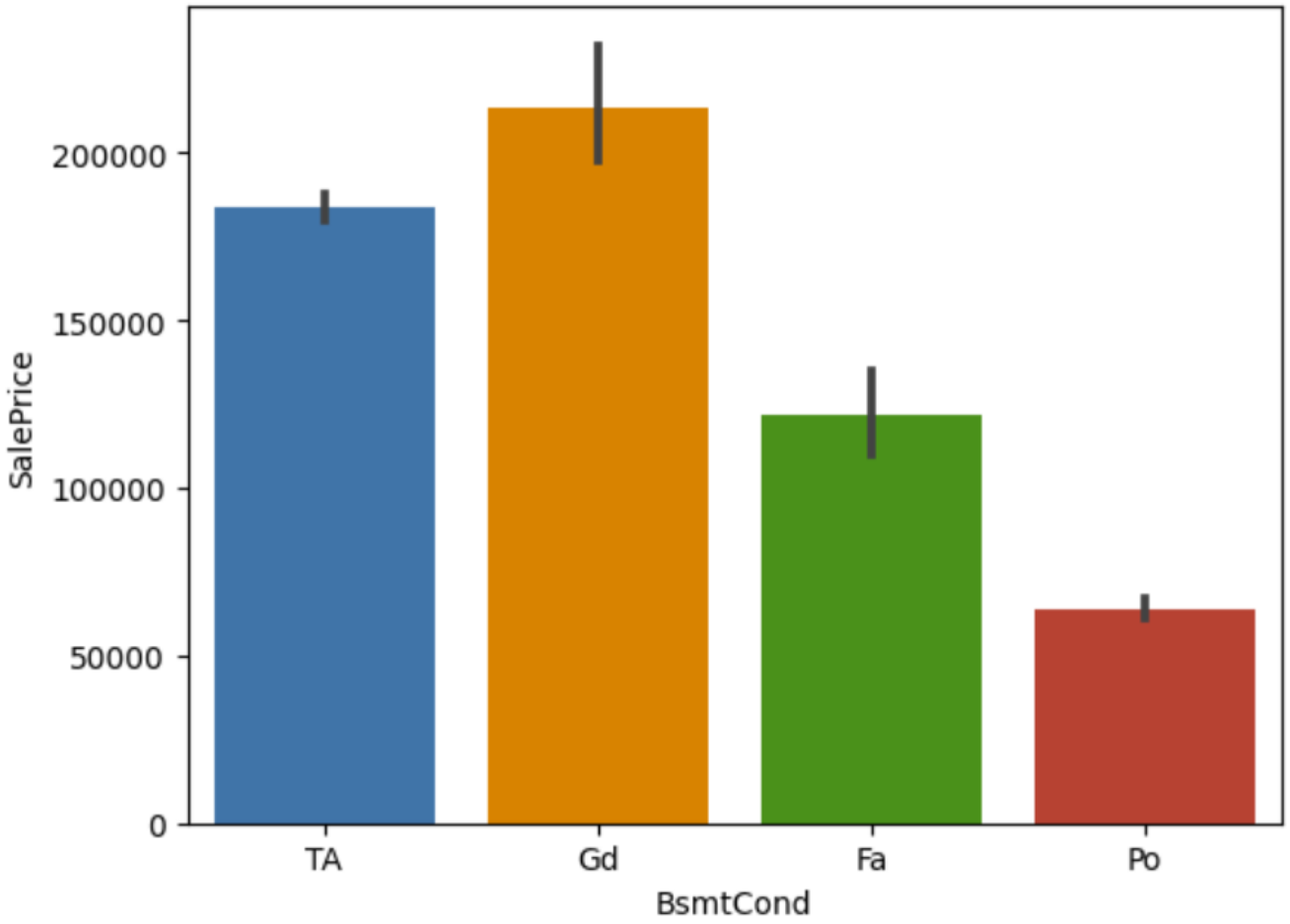
2nd floor square feet vs Sale Price



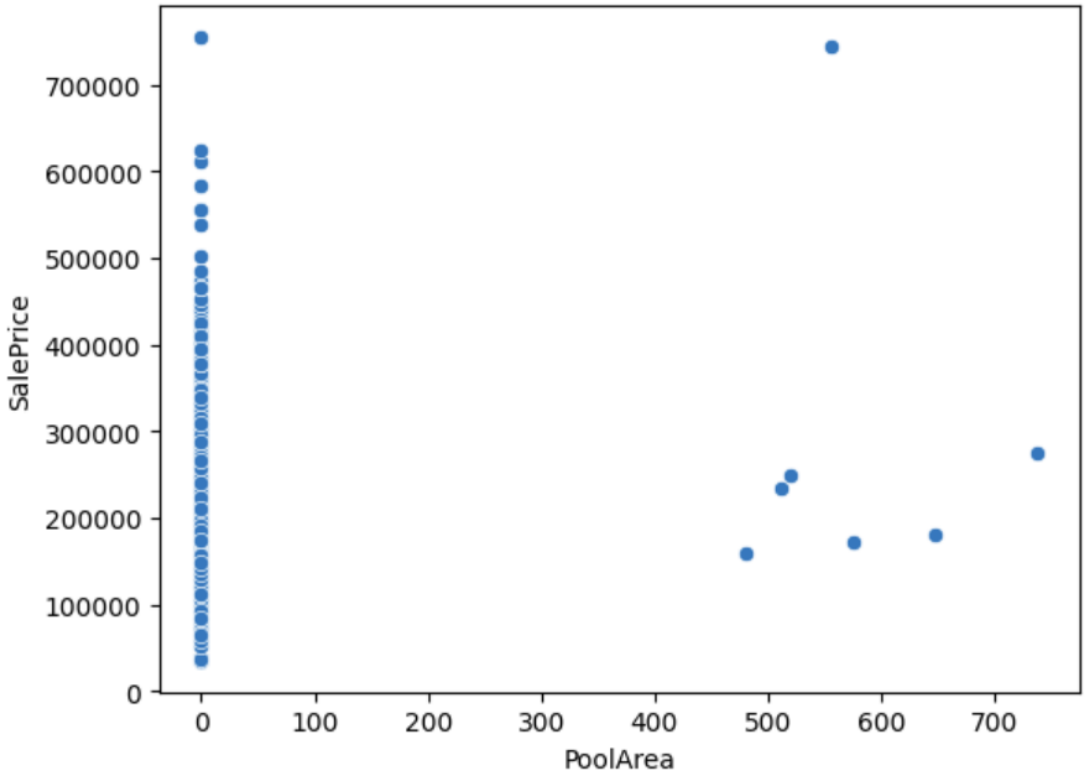
Basement square feet vs Sale Price

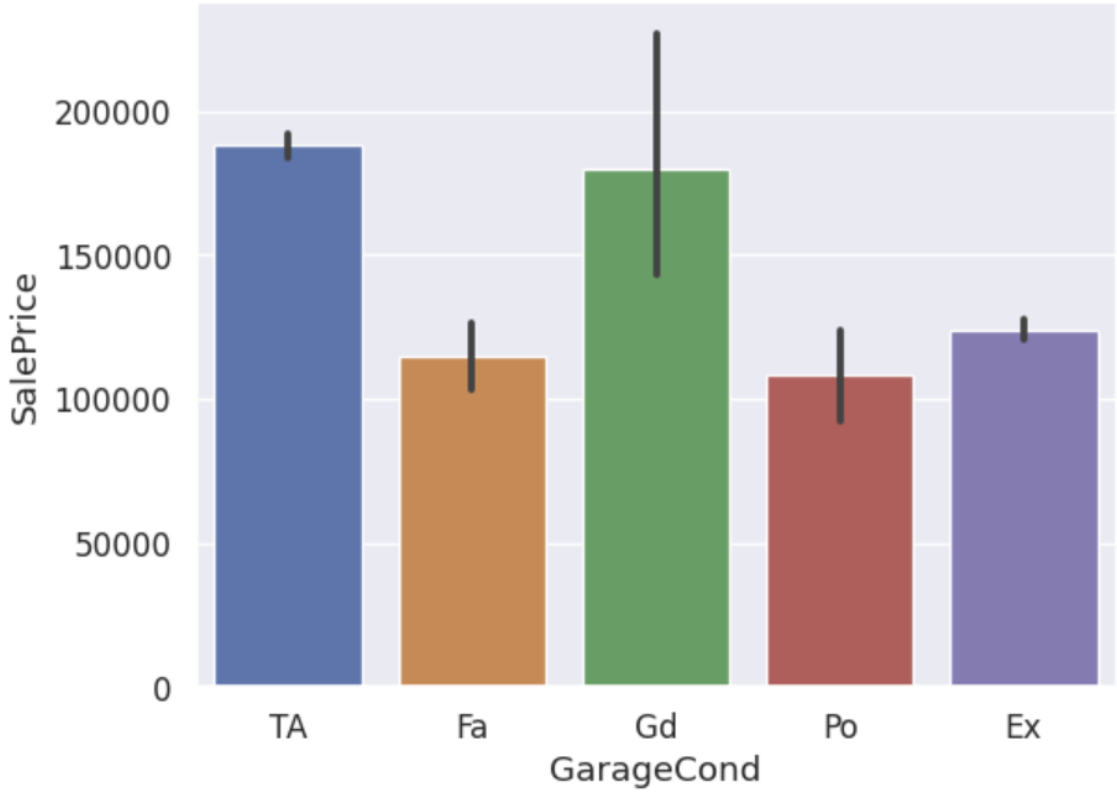


Basement condition vs Sale Price



Not good predictors





Modeling – Version 1

```
The MSE of the gb model is 29377.712223096038
The MAE of the gb model is 19494.582560415183

The MSE of the xgb model is 31472.14500881965
The MAE of the xgb model is 19744.879454730308

The MSE of the Decision Tree model is 40088.6296067423
The MAE of the Decision Tree model is 26362.77397260274

The MSE of the ada model is 37641.267194955624
The MAE of the ada model is 26453.934049700423

The MSE of the knn model is 56061.868756054646
The MAE of the knn model is 34496.572602739725

The MSE of the et model is 28114.369108441155
The MAE of the et model is 18224.241952054792

The MSE of the Random Forest model is 30578.81927422495
The MAE of the Random Forest model is 19412.62626712329
```

Modeling – Version 2

- Hyperparameter tuning
- Feature selection
- Missing values

Missing Values

LotFrontage
Alley
MasVnrType
MasVnrArea
BsmtQual
BsmtCond
BsmtExposure
BsmtFinType1
BsmtFinType2
Electrical
FireplaceQu
GarageType
GarageYrBlt
GarageFinish
GarageQual
GarageCond
PoolQC
Fence
MiscFeature

Feature Selection

	MSSubClass	LotArea	OverallQual	OverallCond	YearBuilt	YearRemodAdd	BsmtFinSF1	BsmtFinSF2	BsmtUnfSF	TotalBsmtSF	1stFlrSF	2ndFlrSF	LowQualFinSF	GrLivArea	BsmtFullBath	BsmtHalfBath	FullBath	HalfBath	BedroomAbvGr	KitchenAbvGr	TotRmsAbvGrd	Fireplaces
0	0.001178	0.013571	0.426864	0.003731	0.010095	0.010013	0.026143	0.000987	0.003535	0.038551	0.018127	0.043565	0.00031	0.095024	0.002742	0.000474	0.013521	0.002795	0.003666	0.005606	0.009696	0.010182

```
['MSSubClass', 'LotArea', 'OverallQual', 'OverallCond', 'YearBuilt', 'YearRemodAdd', 'BsmtFinSF1',
'BsmtFinSF2', 'BsmtUnfSF', 'TotalBsmtSF', '1stFlrSF', '2ndFlrSF', 'LowQualFinSF', 'GrLivArea',
'BsmtFullBath', 'BsmtHalfBath', 'FullBath', 'HalfBath', 'BedroomAbvGr', 'KitchenAbvGr',
'TotRmsAbvGrd', 'Fireplaces', 'GarageCars', 'GarageArea', 'WoodDeckSF', 'OpenPorchSF',
'EnclosedPorch', '3SsnPorch', 'ScreenPorch', 'PoolArea', 'MiscVal', 'MoSold', 'YrSold', 'MSZoning',
'LotShape', 'LandContour', 'LotConfig', 'LandSlope']])
```

Hyperparameter tuning

n_estimators

max_depth

learning_rate

rf, et, gb, xgb hyperparameters

	max_depth	n_estimators	MSE	MAE	mse_mae
0	10	125	31482.517730	18562.596929	50045.114660
1	10	500	31841.073280	18651.694223	50492.767503
2	10	1000	31598.092310	18661.276450	50259.368761
3	10	250	31456.653498	18684.843651	50141.497148
4	10	50	32177.527552	18721.670196	50899.197748

	max_depth	n_estimators	MSE	MAE	mse_mae
0	10	125	33958.240704	17880.019550	51838.260254
1	9	125	32205.030525	17952.186372	50157.216897
2	10	200	33215.166919	18089.292253	51304.459172
3	10	1000	33698.166440	18182.618575	51880.785015
4	10	250	33494.756924	18278.960801	51773.717726

	max_depth	n_estimators	learning_rate	MSE	MAE	mse_mae
0	3	500	0.05	27148.015299	17191.174329	44339.189628
1	3	1000	0.05	27059.842935	17239.267694	44299.110629
2	3	250	0.05	27628.712590	17381.292191	45010.004781
3	3	200	0.10	27845.266541	17442.741743	45288.008284
4	3	100	0.25	26668.171962	17478.766886	44146.938848

	max_depth	n_estimators	eta	MSE	MAE	mse_mae
0	5	1000	0.01	23682.626412	15197.281411	38879.907822
1	5	500	0.10	23710.701187	15232.845034	38943.546222
2	5	1000	0.10	23755.540561	15257.220288	39012.760850
3	5	250	0.10	23773.907370	15317.504120	39091.411490
4	5	200	0.10	23862.391712	15419.747726	39282.139438

Model 2

```
The MSE of the rf model is 34670.69428995096
The MAE of the rf model is 19127.288432784444

The MSE of the et model is 24833.873710586413
The MAE of the et model is 16481.381664312394

The MSE of the gb model is 33414.58654066964
The MAE of the gb model is 17019.61061834213

The MSE of the xgb model is 25123.569594107586
The MAE of the xgb model is 16397.58201787243
```

Model 1

```
The MSE of the Random Forest model is 30578.81927422495
The MAE of the Random Forest model is 19412.62626712329

The MSE of the et model is 28114.369108441155
The MAE of the et model is 18224.241952054792

The MSE of the gb model is 29377.712223096038
The MAE of the gb model is 19494.582560415183

The MSE of the xgb model is 31472.14500881965
The MAE of the xgb model is 19744.879454730308
```

Best Model

- XGBoost
  - n_estimators = 1000
  - max_depth = 5
  - eta (learning_rate) = 0.01
  - MAE = 16,397

Why is this important?

- Home buyers
- Home sellers
- Real estate agents
- Investors