Ames, Iowa Housing Data Analysis

By: Daniel Immediato
Date: 5/03/24

Preliminary: What is the Ames Dataset?

- Ames, Iowa is a college town of Iowa State University. The Ames dataset consists of the housing sale records between 2006-2010, including features like their attributes and sale prices. The goal of this project was twofold:

1. Provide a data analysis of the included features.

    - This includes descriptive models of key features (such as linear and lasso regressions). The highlight being to find the strongest correlations and those values with highest statistical significance.

2. Develop machine learning algorithms for the sale prices.

    - This includes multiple linear regression, random forest, and gradient boosting. The focus would be on high accuracy and high variance (R Squared).

    - Ideally, one would find the most significant contributing factors using multiple linear regression, then one would find the most important indicators using random forest.

Modifying the Data: Data Types

- Before beginning any sort of analysis, one must examine whether or not the data “can” be anyalyzed properly to begin with. The typical dataset is split between numeric and categorical values. When coding, one has to ensure values are consistently classified throughout.
    - As a few examples, taken plainly, YearBuilt and YearRemodAdd are both dates and not integers. We also know PID (a classification) is a string, not an integer. MSSubClass is also a category, not an integer.
    - It is necessary to convert these to their proper data types, otherwise errors may occur or results may appear different than expected.

PID	int64
GrLivArea	int64
SalePrice	int64
MSSubClass	int64
MSZoning	object
LotFrontage	float64
LotArea	int64
Street	object
Alley	object
LotShape	object
LandContour	object
Utilities	object
LotConfig	object
LandSlope	object
Neighborhood	object
Condition1	object
Condition2	object
BldgType	object
HouseStyle	object
OverallQual	int64
OverallCond	int64
YearBuilt	int64
YearRemodAdd	int64

Modifying the Data: Nulls and NA

- An NA value is likely to cause the most issues when dealing with data. When running functions, depending on the program, many can’t process them and may produce an error or show results incorrectly. To deal with them you have several options, such as:

    1. Assume what the value was going to be based on related values.

    2. Replace values as a “0” or “none” if the value is legitimately lacking, or if there is no value.

    3. Replace values with most common value or median.

- Shown is an untransformed version of the “count” of all missing values in the dataset.

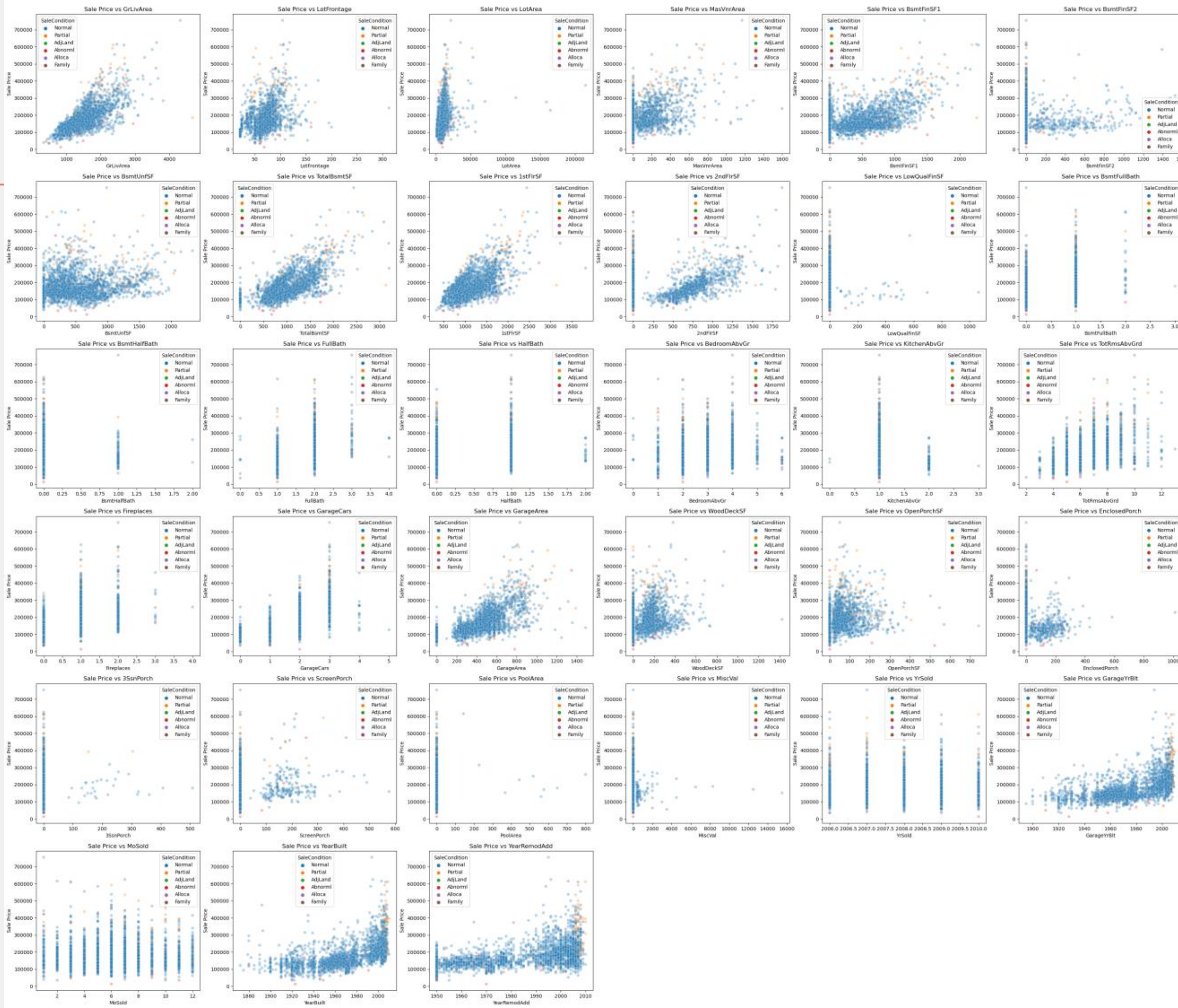
LotFrontage	462
Alley	2412
MasVnrType	1573
MasVnrArea	14
BsmtQual	69
BsmtCond	69
BsmtExposure	71
BsmtFinType1	69
BsmtFinSF1	1
BsmtFinType2	70
BsmtFinSF2	1
BsmtUnfSF	1
TotalBsmtSF	1
Electrical	1
BsmtFullBath	2
BsmtHalfBath	2
FireplaceQu	1241
GarageType	127
GarageYrBlt	129
GarageFinish	129
GarageCars	1
GarageArea	1
GarageQual	129
GarageCond	129
PoolQC	2571
Fence	2055
MiscFeature	2483

Overview: Numeric

- Here is every numeric variable tested against the target, which is SalePrice. We can see many of the variables have some sort of correlation just from eyeing them, and we can also tell which might satisfy the features of linear regression.

- We can also see which variables are ordinal judging by the spacing of the plot points.

- However, having so much data displayed like this is not very helpful.



Multiple Linear Regression: As a Model


(Image Created Via Stable Diffusion)

- A MLR is used to determine the strength of relationships, as well as statistical significance. This type of model can help us find the more important variables when compared to our target.

- One must ensure features are linear to the target, there is constant variance, normality of errors, there is independence of errors, and that there is as little multi-collinearity as possible.

MLR: Unrefined

- Since there wasn't a starting point, we ran a model on all columns (with some modifications). The score was surprisingly high at .91. However, this is unlikely to be helpful considering how much this model suffers from multi-collinearity.

- Still, this gives us a basis on where to refine the data based on the significant coefficients.

The insignificant coefficients

BsmtUnfSF	0.988478
OpenPorchSF	0.944456
Condition1	0.862291
BsmtFinType2	0.836090
SaleType	0.813722
MiscVal	0.810322
Neighborhood	0.781728
BsmtCond	0.731701
Fence	0.654655
Electrical	0.612990
BsmtHalfBath	0.580295
Utilities	0.471880
LandSlope	0.458453
3SsnPorch	0.427188
GarageQual	0.393263
LotConfig	0.378711
LandContour	0.372921
LowQualFinSF	0.365857
RoofStyle	0.357785
MoSold	0.228199
GarageYrBlt	0.201522
GarageFinish	0.179811
HalfBath	0.171371
GarageType	0.168582
Foundation	0.155015
Heating	0.119663
BsmtFinType1	0.112230
MiscFeature	0.104077
GarageArea	0.098933
PoolArea	0.088707
Exterior1st	0.080238
Alley	0.079807
WoodDeckSF	0.065995
FireplaceQu	0.053153
RoofMatl	0.053090

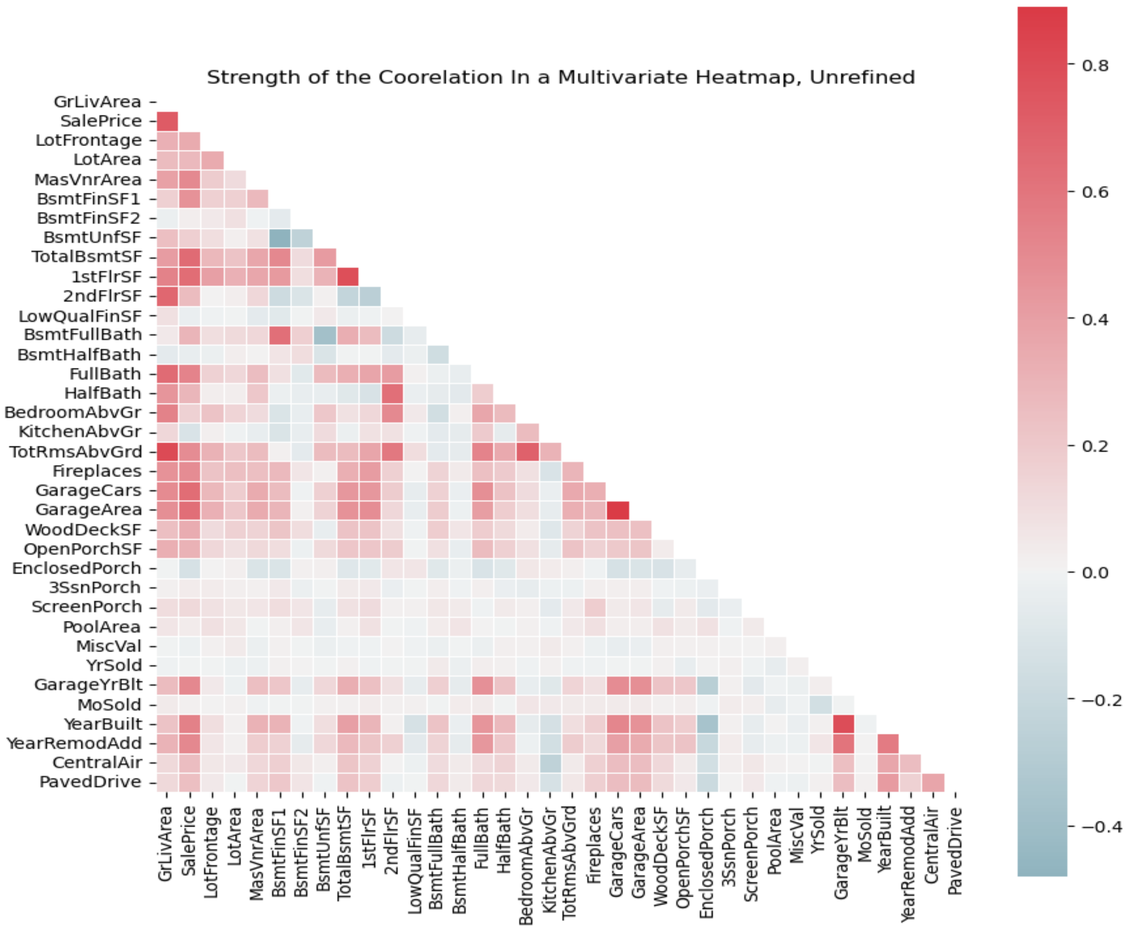
OLS Regression Results

Dep. Variable:	SalePrice	**R-squared:**	0.909
Model:	OLS	**Adj. R-squared:**	0.906
Method:	Least Squares	**F-statistic:**	323.5
Date:	Sat, 15 Jun 2024	**Prob (F-statistic):**	0.00
Time:	17:06:35	**Log-Likelihood:**	1874.8
No. Observations:	2580	**AIC:**	-3594.
Df Residuals:	2502	**BIC:**	-3137.
Df Model:	77		
Covariance Type:	nonrobust		

The significant coefficients

OverallCond	5.321098e-62
OverallQual	1.772705e-51
TotalBsmtSF	7.504272e-28
GrLivArea	5.649917e-26
YearBuilt	3.077108e-19
SaleCondition	6.186082e-19
BsmtFinSF1	5.993099e-15
Fireplaces	1.361580e-13
1stFlrSF	3.270336e-13
Functional	3.939083e-12
ExterQual	1.253207e-10
CentralAir	1.570432e-09
LotArea	2.581841e-09
ScreenPorch	6.320603e-09
2ndFlrSF	1.770946e-08
KitchenAbvGr	6.475868e-07
PavedDrive	1.599109e-06
EnclosedPorch	4.495701e-06
HeatingQC	8.556424e-06
GarageCond	1.031048e-05
KitchenQual	1.059962e-05
ExterCond	1.649634e-05
BedroomAbvGr	1.965074e-05
const	1.069183e-04
GarageCars	1.254324e-04
MSZoning	2.224002e-04
BsmtFullBath	2.295378e-04
BsmtExposure	2.488707e-04
YearRemodAdd	3.054915e-04
TotRmsAbvGrd	8.191396e-04
Exterior2nd	1.232505e-03
Street	1.550811e-03
LotFrontage	4.897234e-03
Condition2	8.903051e-03
BsmtFinSF2	8.978047e-03

MLR: Summarizing Unrefined Results



- To the left, we see the correlations of what the previous MLR told us. Here, it only tells us the *numeric* correlations, which is why the most significant result, Overall Condition, is not found.
- For example, from the heat map, we can see that Ground Living Area is not only very statistically significant (MLR model said 5.64x10^-26), it also is strongly correlated with many other variables, including SalePrice.
- Enclosed Porch is another note here, as it is clearly not very correlated with anything, but it was marked as highly statistically significant (4.5x10^-6).

Random Forest: As a Model

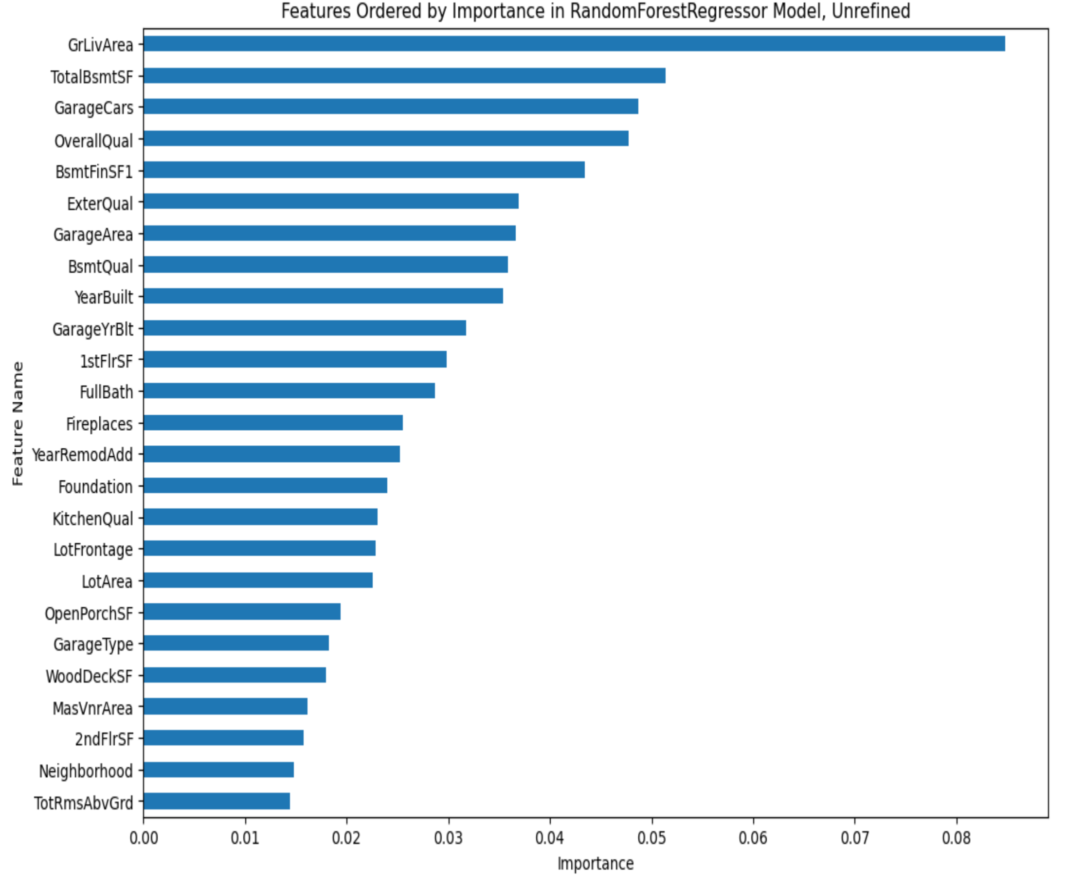
- A random forest model combines multiple decision trees into a single model. In this case, it would determine feature importance rather than MLR's significance and correlation. Feature importance is a degree of dependency on the variables being compared.

- A decision tree is a hierarchical structure where each "node" represents branches based on different conditions being satisfied. Decision trees are subject to "overfitting", which can lead to "poor predictive performance".

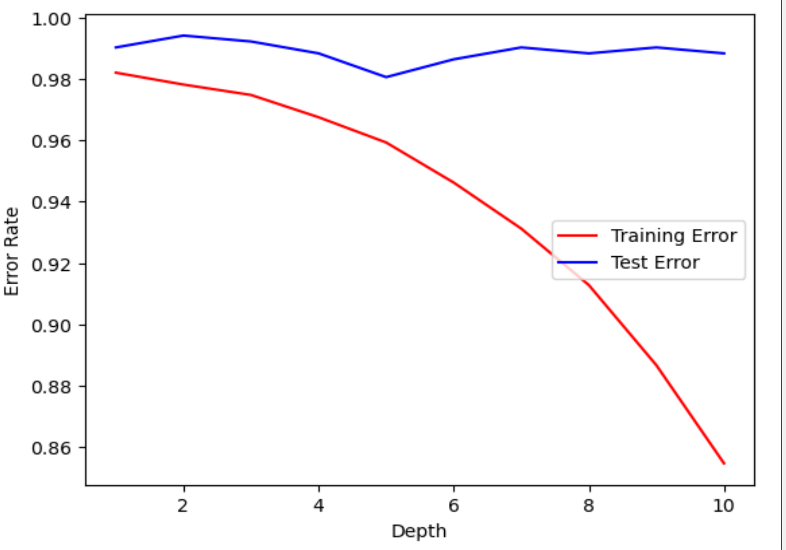
(Image Created Via Stable Diffusion)

RF: Unrefined

- The unrefined RF more represents the correlated heat map, while only vaguely representing the MLR model. For example, OverallCond, the most significant, isn't even in the Top 25 as a feature.



- However, the decision tree tells us that this graph is overfitting at the depth of five. It's score is also under .90, at .85, which means the model is potentially unreliable.



Constructing a Refined Model

- In order to refine the models, we have to take what was discovered from feature importance, as well as statistical significance from the MLR model, and combine the results. This would hopefully create a smaller, more accurate model.

- For example, from both models we know GrLivArea is probably important, as are variables like OverallQual, YearBuilt, FirePlaces, BsmtFinSF1, etc..

- Some important variables, like 1stFlrSF can also be combined. We can therefore create Total Square Footage (TotalBsmtSF, 1stFlrSF, and 2ndFlrSF), Total Bath (BsmtFullBath, BsmtHalfBath, FullBath, and HalfBath), and Mixed Exterior (Exterior1st, Exterior2nd). This gets rid of issues where the “parts” may not be considered insignificant, but the whole could be.

- Finally, we need to adjust the parameters of our model to be more ideal in a form of hyperparameter tuning.

(Image Created Via Stable Diffusion)

MLR: Refined

- After refining the data, the MLR resulted in a slightly lower score, but a much higher F-statistic. This means it has a stronger correlation, but it is less predictable. We can also see that the results much more align with the importance scale, just based on eyeing it.

The significant coefficients

OverallCond	1.814289e-93
OverallQual	3.270585e-63
YearBuilt	1.677427e-49
TotalSF	6.382417e-45
Fireplaces	3.860264e-25
GarageArea	5.097195e-24
BsmtFinSF1	6.307434e-21
GrLivArea	1.419740e-19
const	1.774230e-15
LotArea	1.142200e-10
ExterQual	3.148558e-09
KitchenQual	1.399883e-07
GarageType	4.279913e-06
LotFrontage	4.831742e-06
BsmtExposure	2.095022e-04
YearRemodAdd	3.781898e-04
MixedExterior	5.670407e-03
BsmtQual	9.672514e-03

dtype: float64

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.886
Model:	OLS	Adj. R-squared:	0.885
Method:	Least Squares	F-statistic:	796.2
Date:	Wed, 12 Jun 2024	Prob (F-statistic):	0.00
Time:	03:58:37	Log-Likelihood:	1591.2
No. Observations:	2580	AIC:	-3130.
Df Residuals:	2554	BIC:	-2978.
Df Model:	25		
Covariance Type:	nonrobust		

- Some of the results are also identical to the unrefined MLR model, while our new variable, TotalSF, has now become the most significant numeric variable.

The insignificant coefficients

Neighborhood	0.984448
OpenPorchSF	0.965231
WoodDeckSF	0.773077
TotRmsAbvGrd	0.523517
MSSubClass	0.229175
GarageYrBlt	0.220629
MasVnrArea	0.151747
TotalBath	0.105733

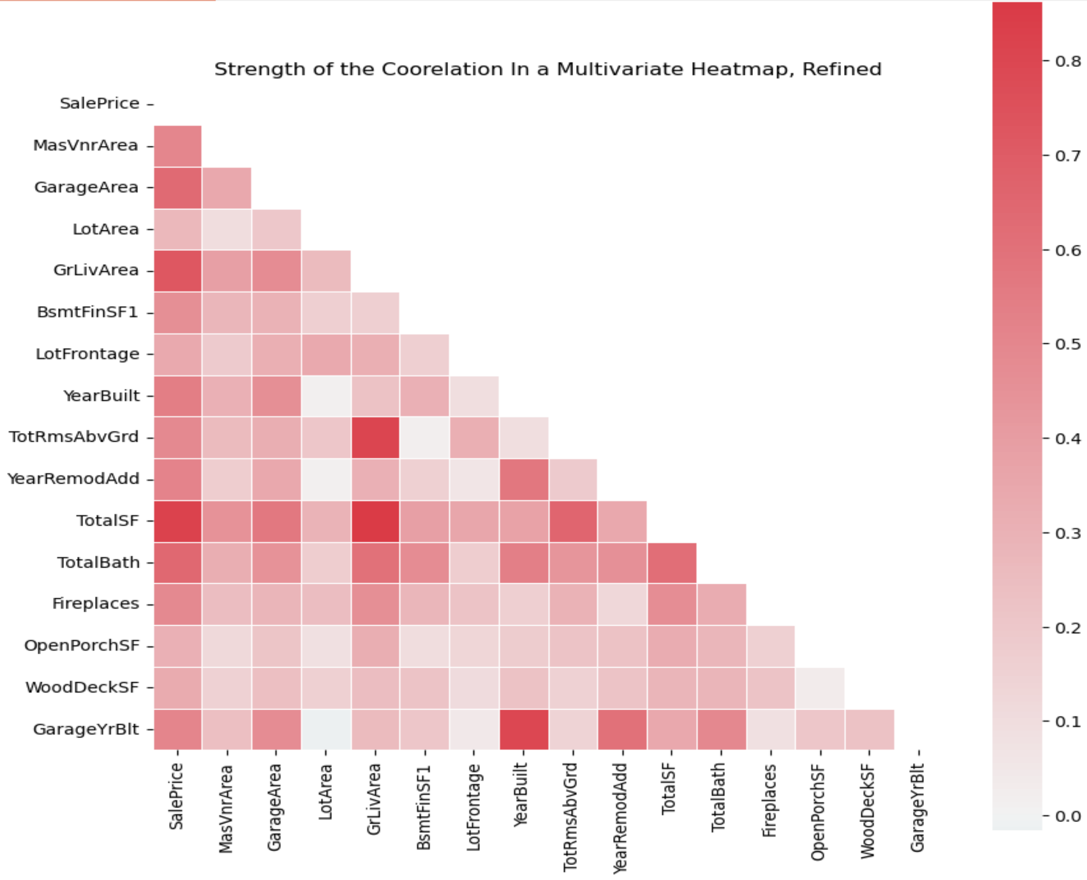
dtype: float64

MLR Refined Correlation Summary

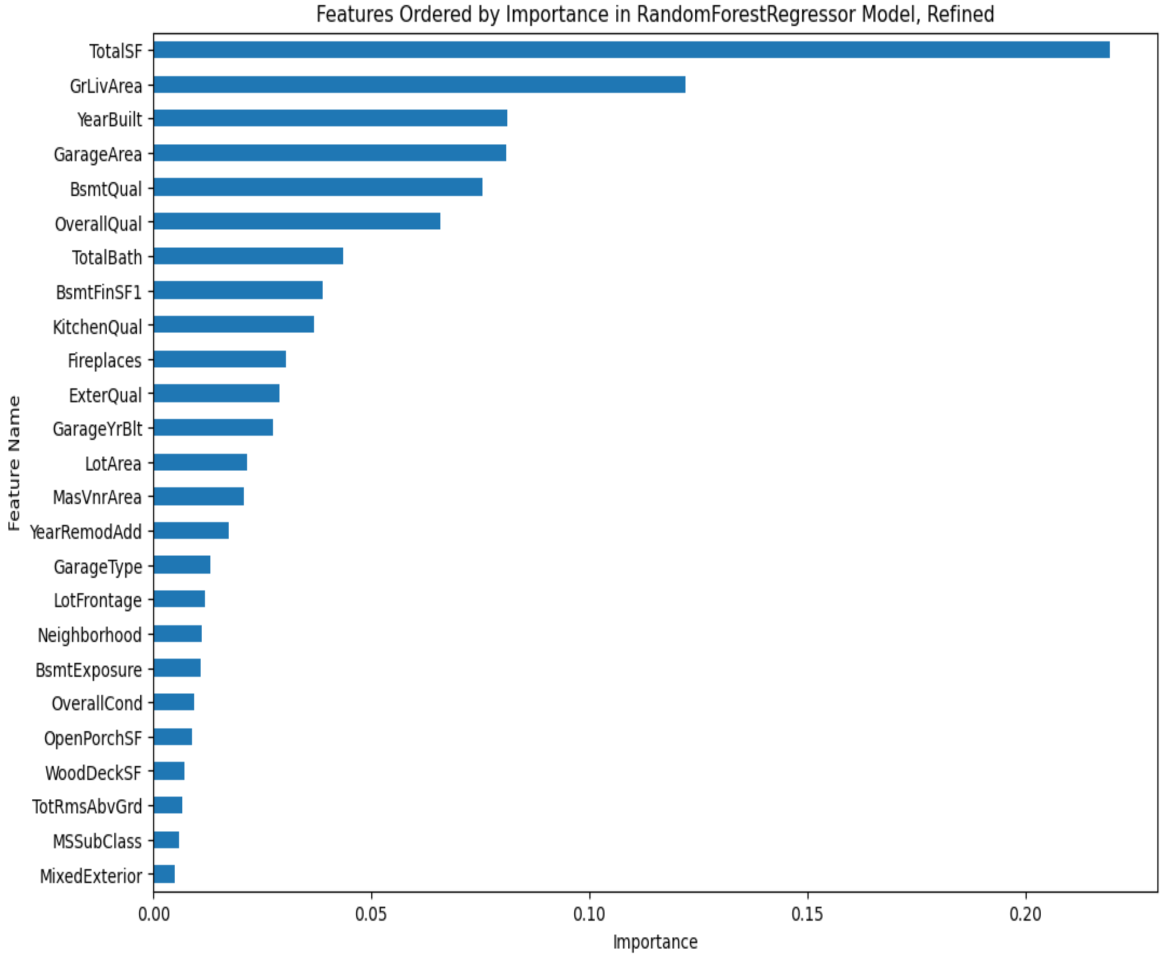
- Here, we can see our results compared to the unrefined findings. The improved F-statistic told us there was a stronger correlation, and here we can see nearly every variable is “strong”, even against each other.

- Total SF seems to not only have the highest significance, but also the highest inter-correlation across the board, followed by GrLivArea.

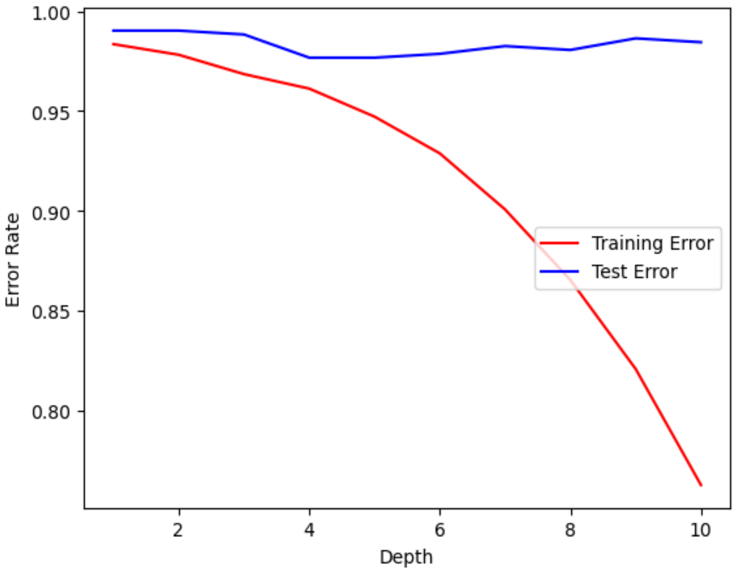
- Our weakest correlation seems to be LotArea, which was considered statistically significant. This is of note because, GarageYrBlt has an, overall, higher correlation yet was insignificant.



RF: Refined



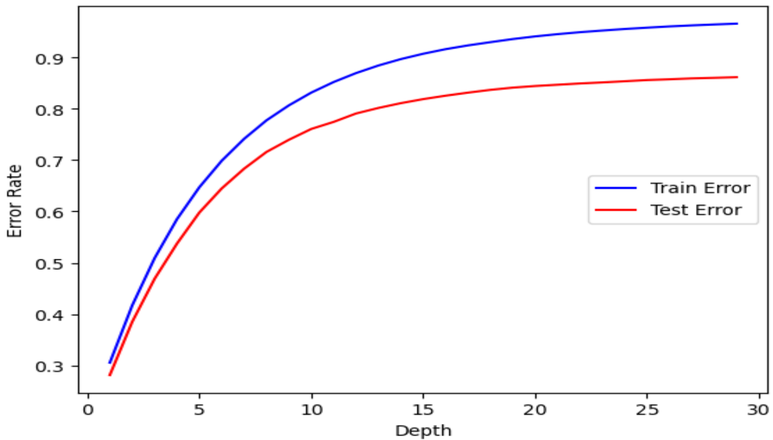
- The results here are more in-line with the correlations, with TotalSF still being the most important variable. One of the new variables, MixedExterior, is also now the least important variable, yet it was considered more significant than BsmtQual, which is in the top five.



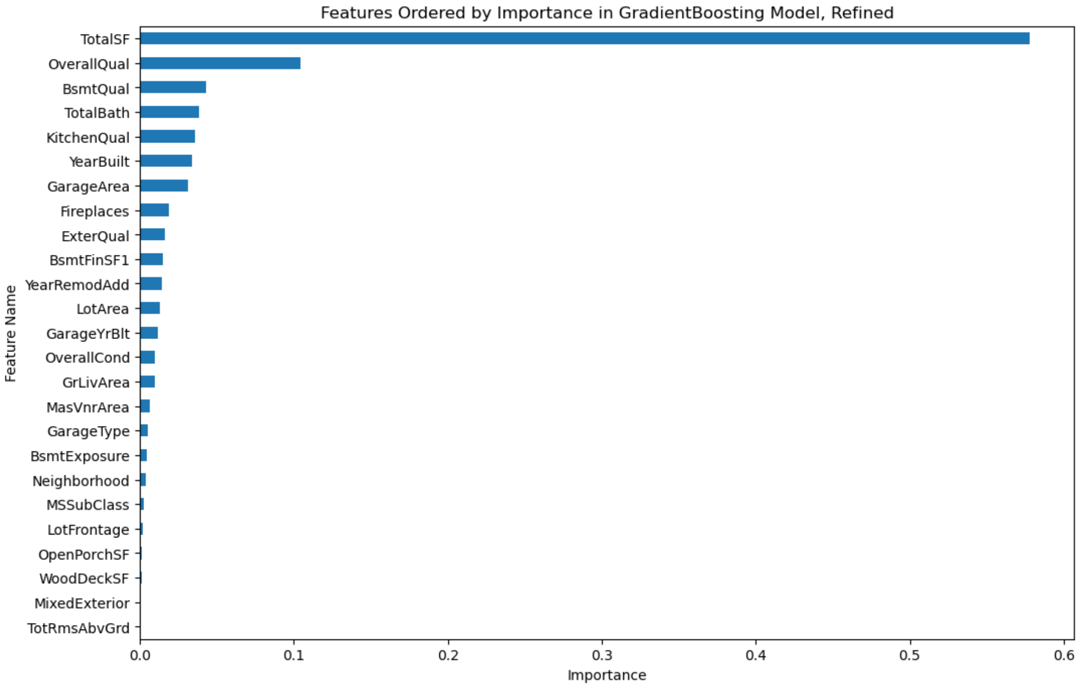
Gradient Boosting: Improving Accuracy

- Gradient boosting, like random forests, combines multiple decision trees to create a model. However, the main difference is that, while random forests are constructed independently, gradient boosting is linear and constructed sequentially, so that it corrects itself.

- Because of this self-correction, the expectation, at least, is a collective increase in the score at the risk of increased overfitting.



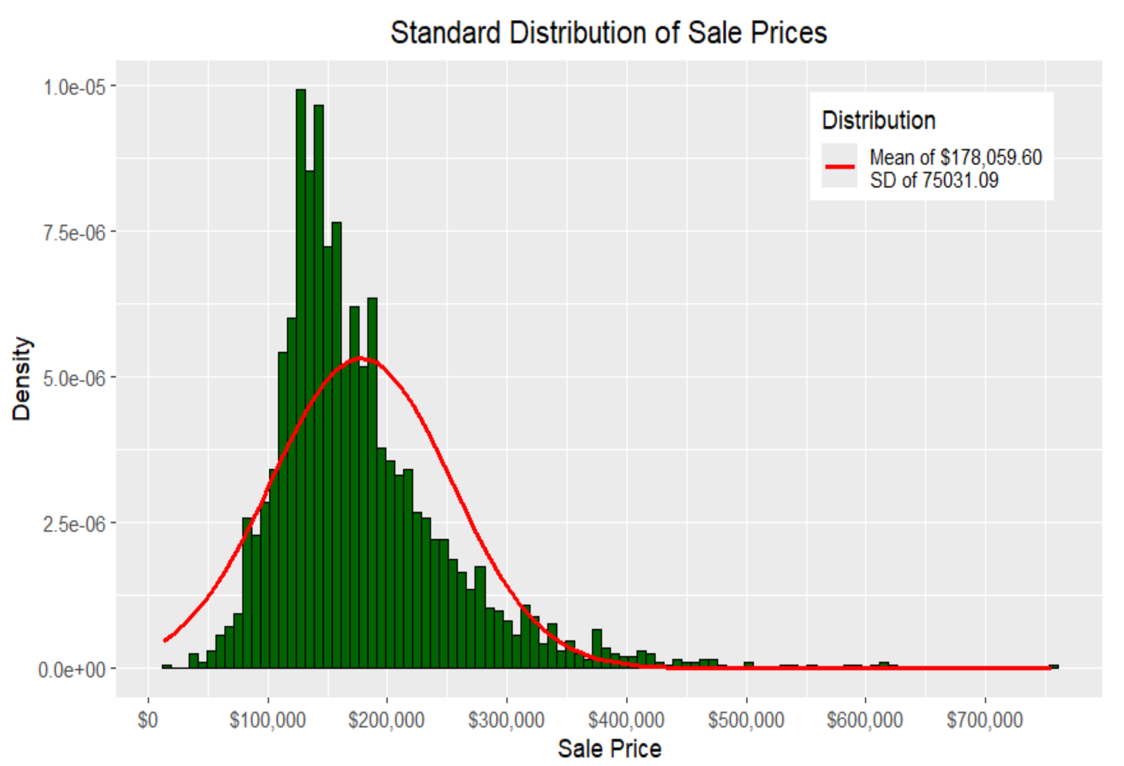
- This time overall importance was reduced significantly, although TotalSF is still considered to be the most important factor. Most of the other factors are off by one or two places compared to the RF. This could be explained by the different decision tree, which is now underfitting instead of overfitting.



Breaking Down the Results: The Target

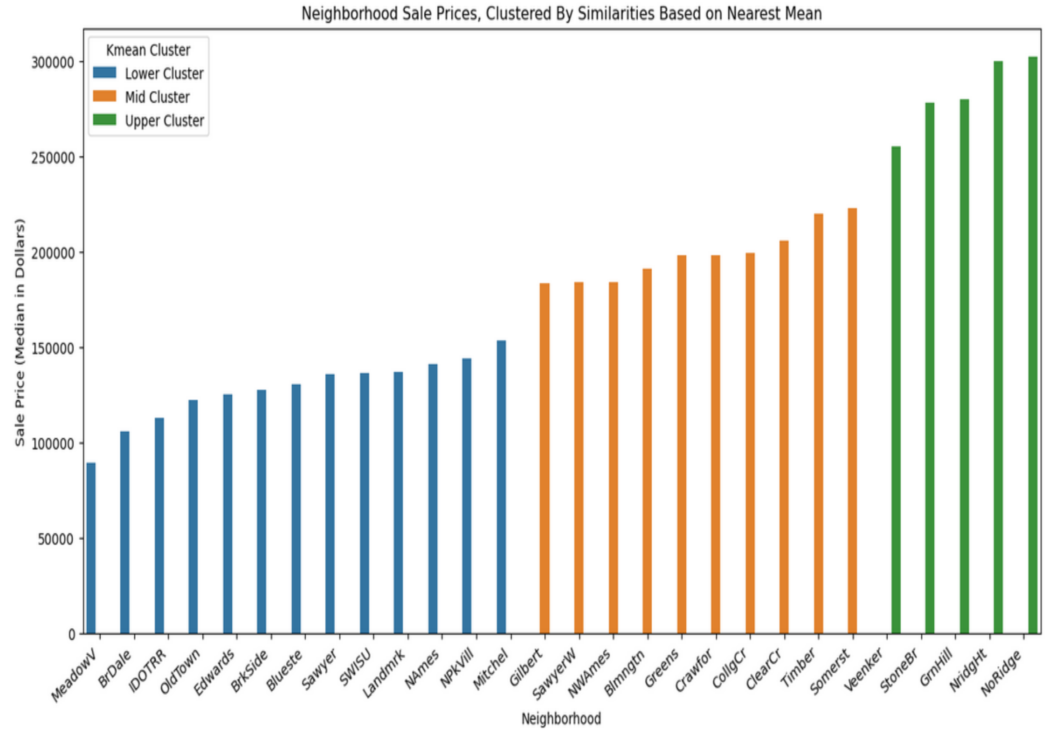
- To start an analysis, we can look at the target of the models: the Sale Price. Using a standard distribution histogram, we can see that most houses have been sold at around $180,000. This is fairly typical, as most people cannot afford more expensive houses.

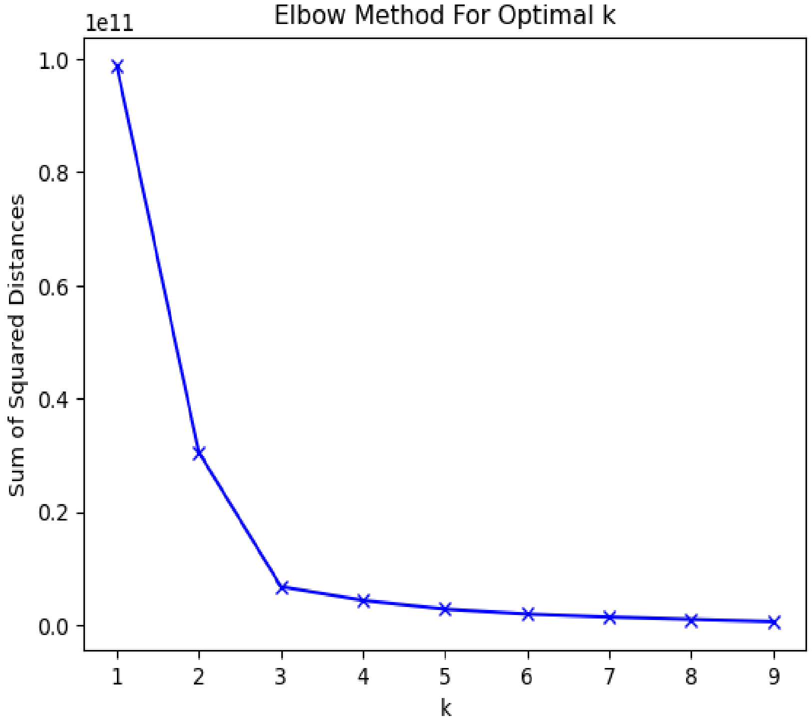
- However, it should also be kept in mind that there are numerous outliers in the data because of this. In fact, one can even be seen here, with a house being sold at over $700,000. This means every variable, such as TotalSF, TotalBath, etc. would likely distort the data compared to what should be the average.



Binning the Neighborhoods

- To get a better understanding of which neighborhoods produced the most significant sales, we decided to “bin” them using k-mean clustering. This is a grouping method where each value is grouped together based on similarities according to the nearest mean. To find said mean, one would use the “elbow method”, where the “elbow” is how many groups there are. So, here we can see there are three clusters of neighborhoods based on sale price, with the green cluster possibly being the most valuable.

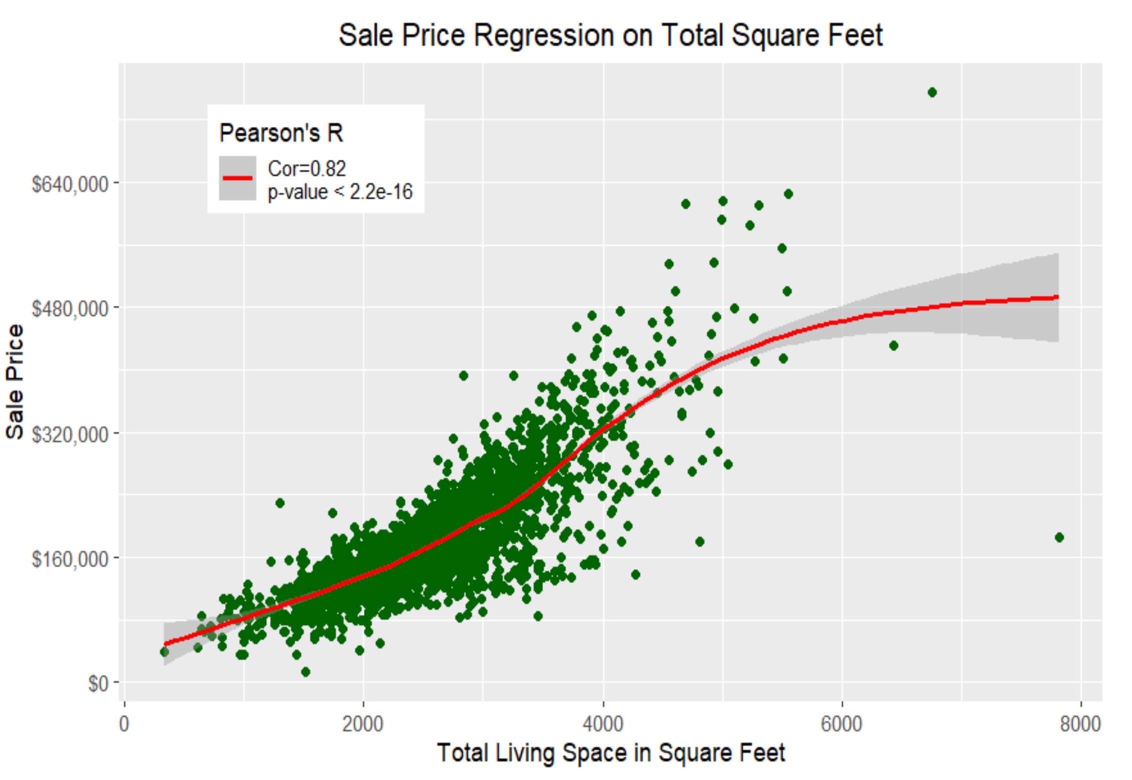




Most Important Feature: Numeric

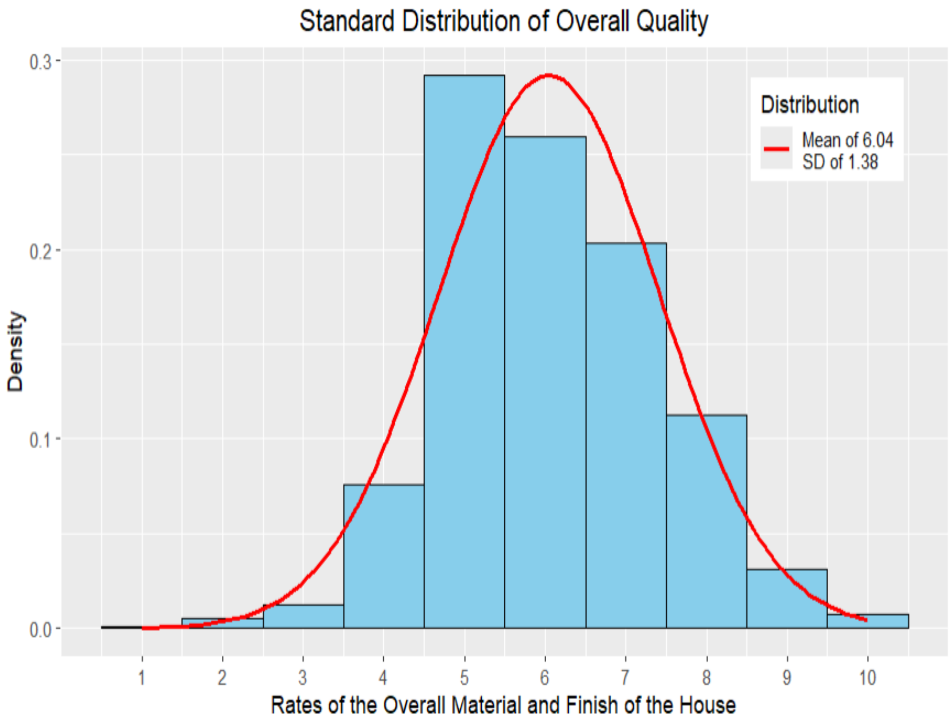
- According to our models, the most important numeric feature is TotalSF, which is the combination of the SF variables. From a weighted regression model (loess), we can see it has a nearing-perfect positive correlation with Sale Price. This aligns with what both models said.

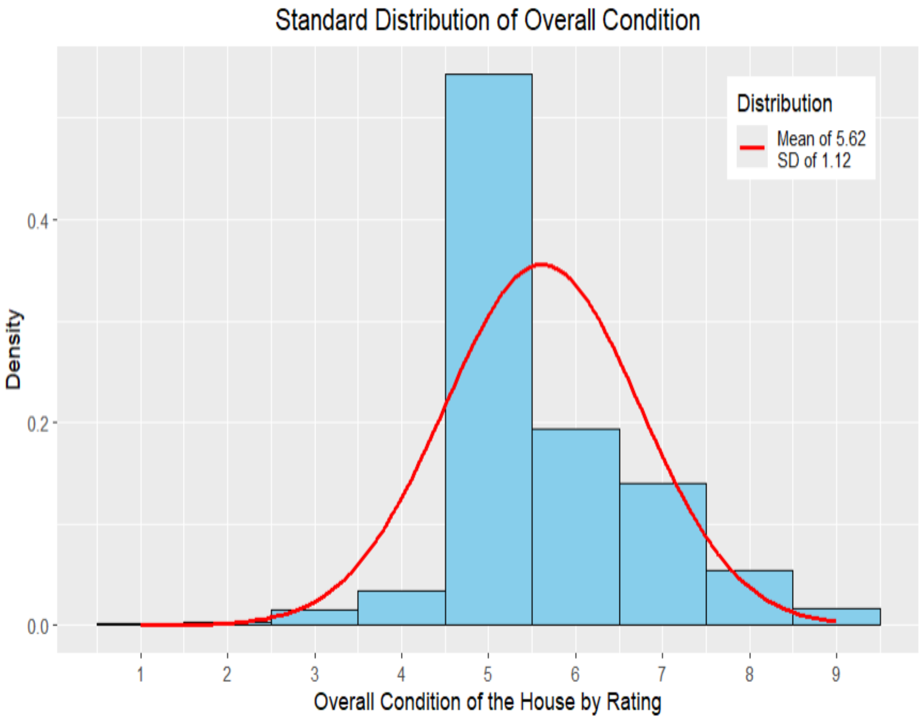
- This has only two visible outliers, but they don’t distort the regression line enough to shift it in a meaningful way.



Most Important Feature: Categorical

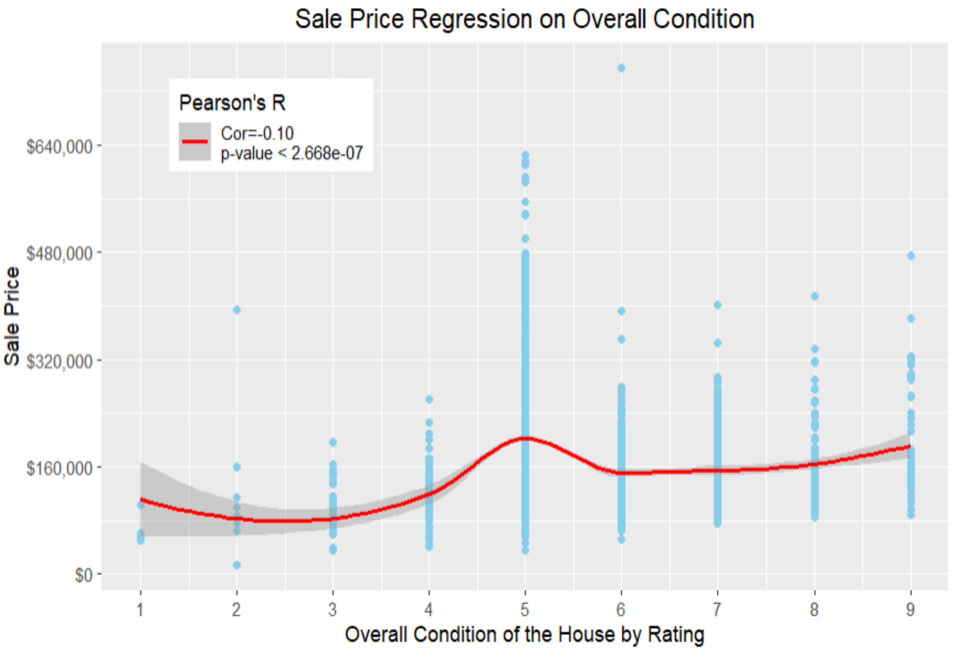
- Our models seem to at least agree that OverallQual is both important and significant, but the random forest model rates OverallCond lower in importance, while the MLR model rates it the highest in significance. The meaning of this will be examined.
- Right away, we can see smaller variation in the data. While less variability means less extreme data, this also looks like the regression is going to be more effected by weight.





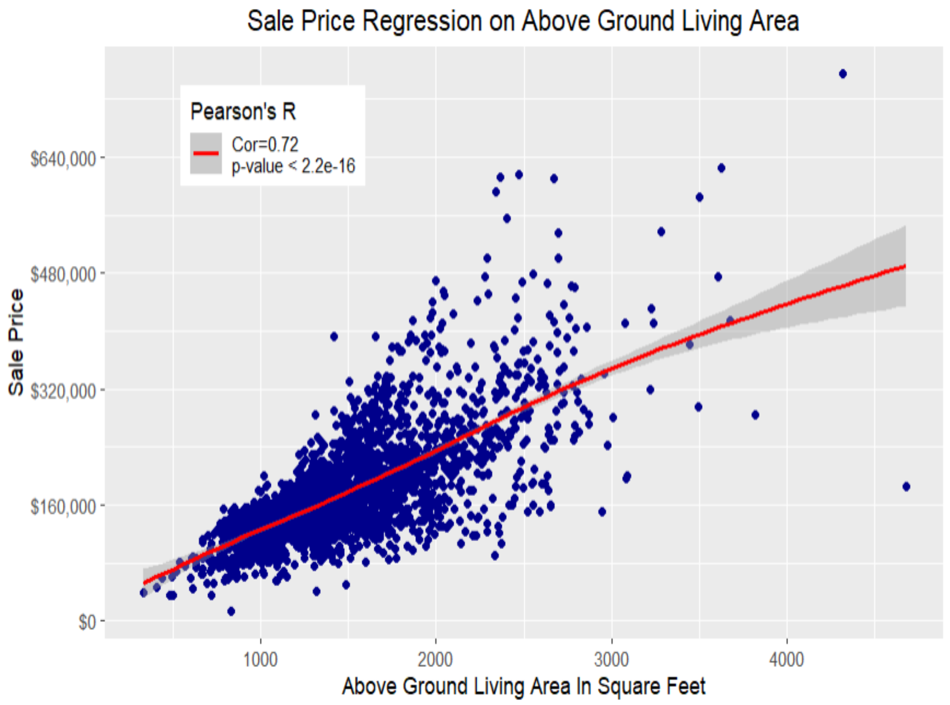
Weighted Regression on Housing Rates

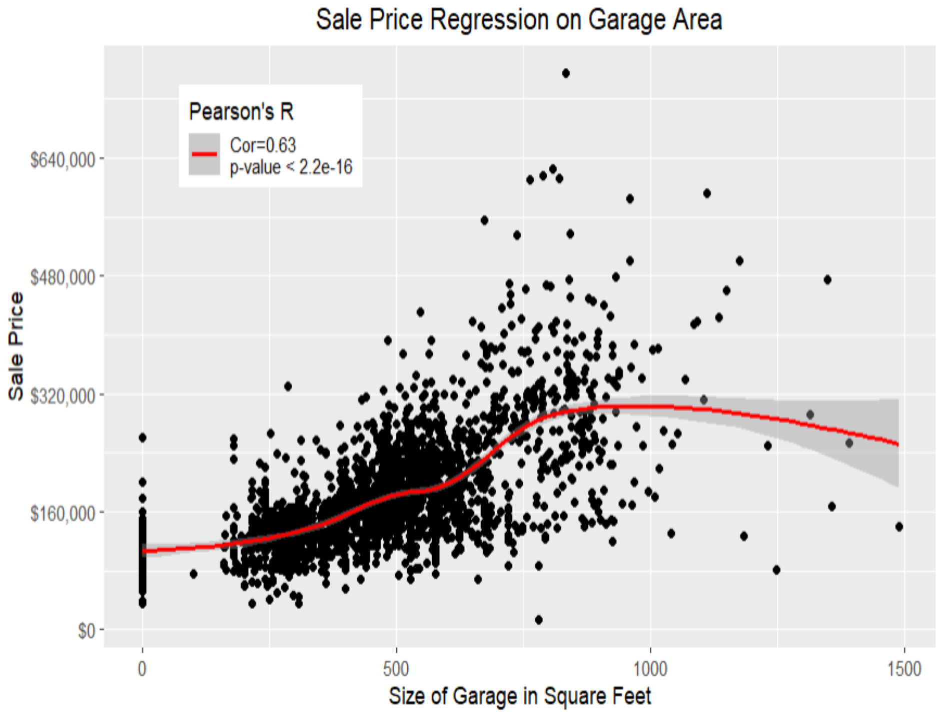
- As expected, the OverallQual variable has a much higher correlation than the OverallCond variable. So much so that OverallCond goes into the negative. This may explain why OverallQual is so much higher in the RF model. As shown in the previous graphs, the smaller variability in OverallCond caused so much focus on one rating (5) that the loess regression produced a “hump” at that point.



Other Highly "Important" Features

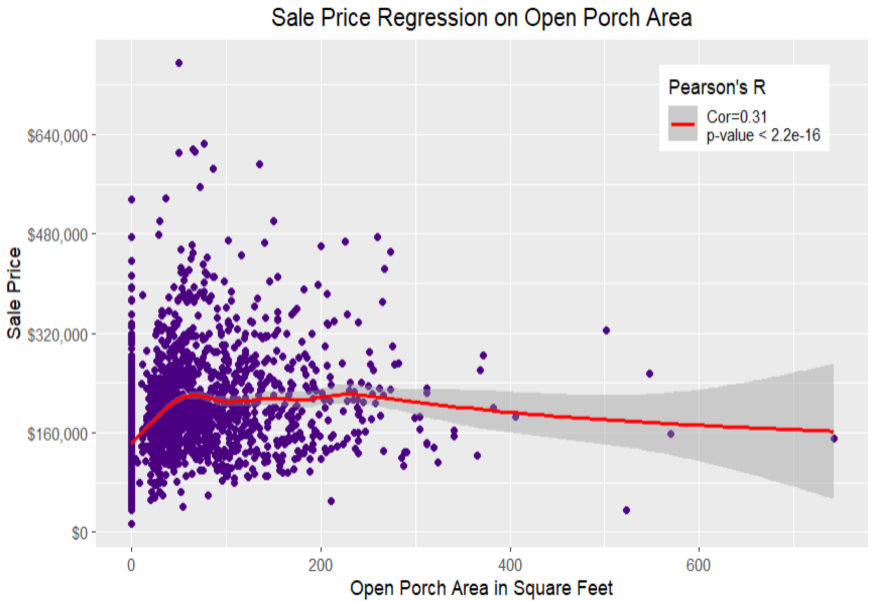
- The other most important features, according to the RF model, are similar to the TotalSF. They are highly correlated, along with being statistically significant. GarageArea was actually considered one of the most significant factors, according to the MLR model, but here we see GrLivArea coinciding with linearity more. GarageArea also has a numerous amount of "0" values, presumably those houses that do not have garages, which could distort the data.

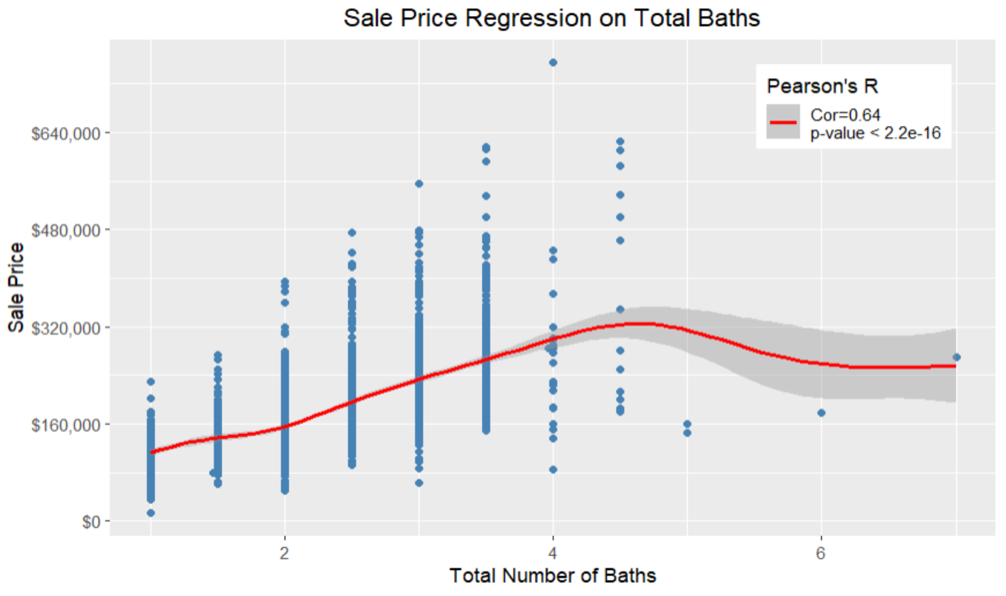




The Weaker Correlations

- The RF said that TotRmsAbvGrd was the lowest numerical factor by importance. OpenPorchSF not only had low importance, but it was the numerically most insignificant factor. TotalBath, one of the new variables, was considered important, but insignificant. There are inconsistencies here, such as how TotRmsAbvGrd has a higher correlation than LotArea (0.27), yet is less important. In the MLR OpenPorchSF and TotalBath were insignificant, but when analyzed by themselves, they're significant.





Summary: The Models

- Overall, the models give us mixed results. The refined models increased the score with the exception of the MLR model, although the unrefined MLR model has the unique problem of multi-collinearity.

- The refined models focus-in on the strength of the relationships, while still leaving room for similarity between the models (such as how TotalSF is the most important feature, and how OveralCond is the most significant feature in both models).

- Yet still, there are many discrepancies to be made, such as how TotalBath is one of the more important factors, yet the MLR model considered it insignificant. The GB model also, despite having the highest score, had a severe underfitting issue instead of overfitting. Its features were also mixed compared to the refined RF, having a lowered importance overall.

- While immediate overfitting is apparent for both the unrefined and refined RF models, it should also be noted that the unrefined BIC is higher than the refined model, while having a lower AIC. This gives two different answers to the question of “is it a better-fit model”.

MLR Unrefined Score	MLR Refined Score	RF Unrefined Score	RF Refined Score	GB Refined Score
0.909	0.886	0.850	0.903	0.960

Summary: The Data

- The refined models can be used to pick out the features that have A) the strongest correlations and B) those that have the highest statistical significance. We can then test these features by themselves against the unrefined and refined models to see if the assumption still holds true.

- It would appear that TotalSF in all cases has the strongest correlation. GrLivArea was considered the second most important feature, but its strength was outmatched by OverallQual, which was consistently one of the features with the highest statistical significance.

- We noted that the weakest relationships had inconsistent results with the models, such as how certain ones were significant in one area or more/less important. This was also reflected when tested alone.

- OverallCond notably had a very low correlation, even though it was one of the most statistically significant relationships according to the models. However, when tested alone in Pearson’s r (note the MLR were Ordinary Least Squares models) it was actually had a lower significance than the other relationships.

Areas For Improvement

Modeling

1. **Predictive Reliability:** The score for all models was lower than desired. Ideally, they should all be above .90

2. **Effects Size:** One of the notable problems with the models was that the correlations went against the significance. A feature can be statistically significant, but not statistically meaningful. A model should, ideally, reflect both.

3. **Other Models:** Other models can be used to address various issues. This includes classification modeling (regressors were used), lasso regression, gridsearch, and feature engineering. Additionally the parameters can be adjusted

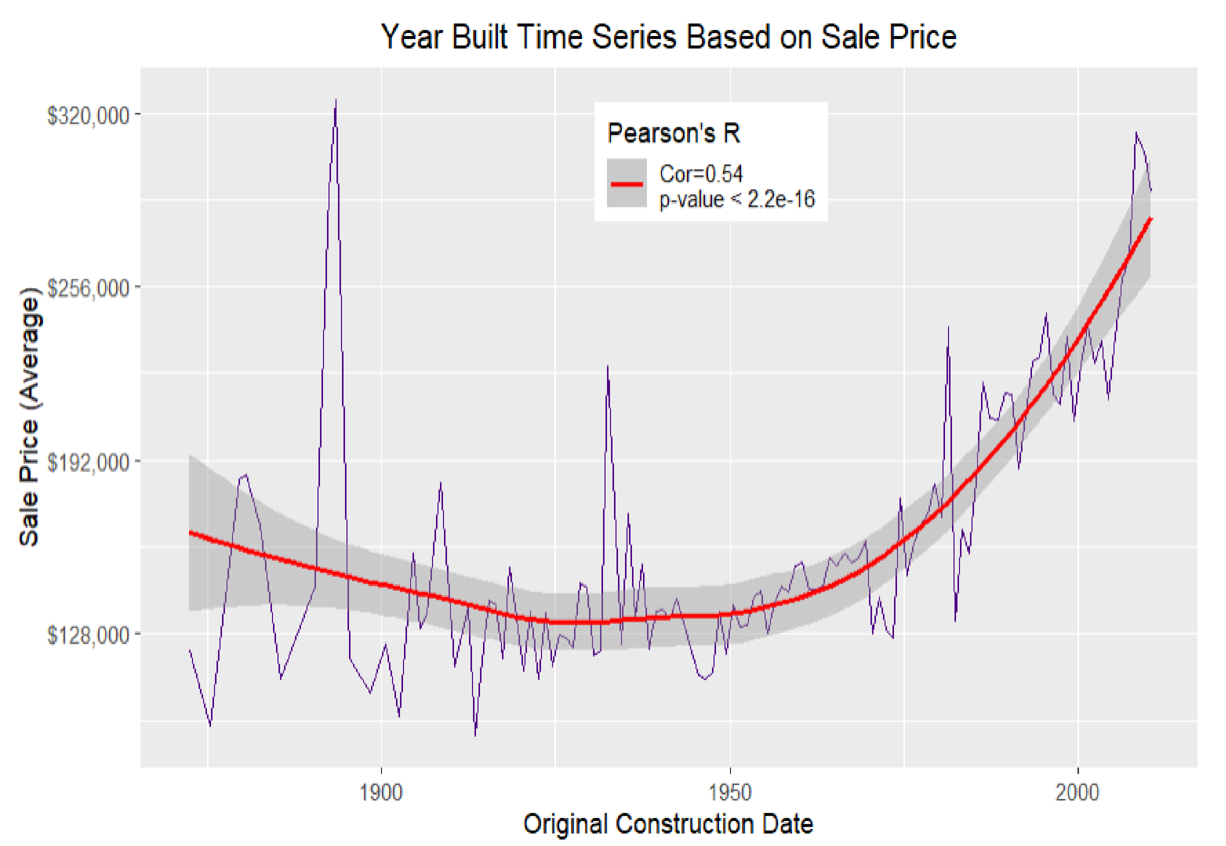
Data

1. **Feature Selection:** We refined the results arbitrarily based on the feature importance combined with the statistical significance. Judging by what was mentioned with “best score”, the slightest changes with what was selected could have drastically affected the results.

2. **Outliers and Missing Values:** Most of the missing values were replaced with “mean”, unless context said otherwise. This is not always wise and can distort the data. Furthermore, no outliers were removed. This can severely effect linearity when it comes to a probability plot. This is noticible in several of the graphs.

3. **Normalization:** Data can be normalized to reduce problems with overfitting.

Questions? Comments? Suggestions?



- YearBuilt was in the top five for the most important factors, and was notably also statistically significant according to the MLR model.