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 time series. 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.

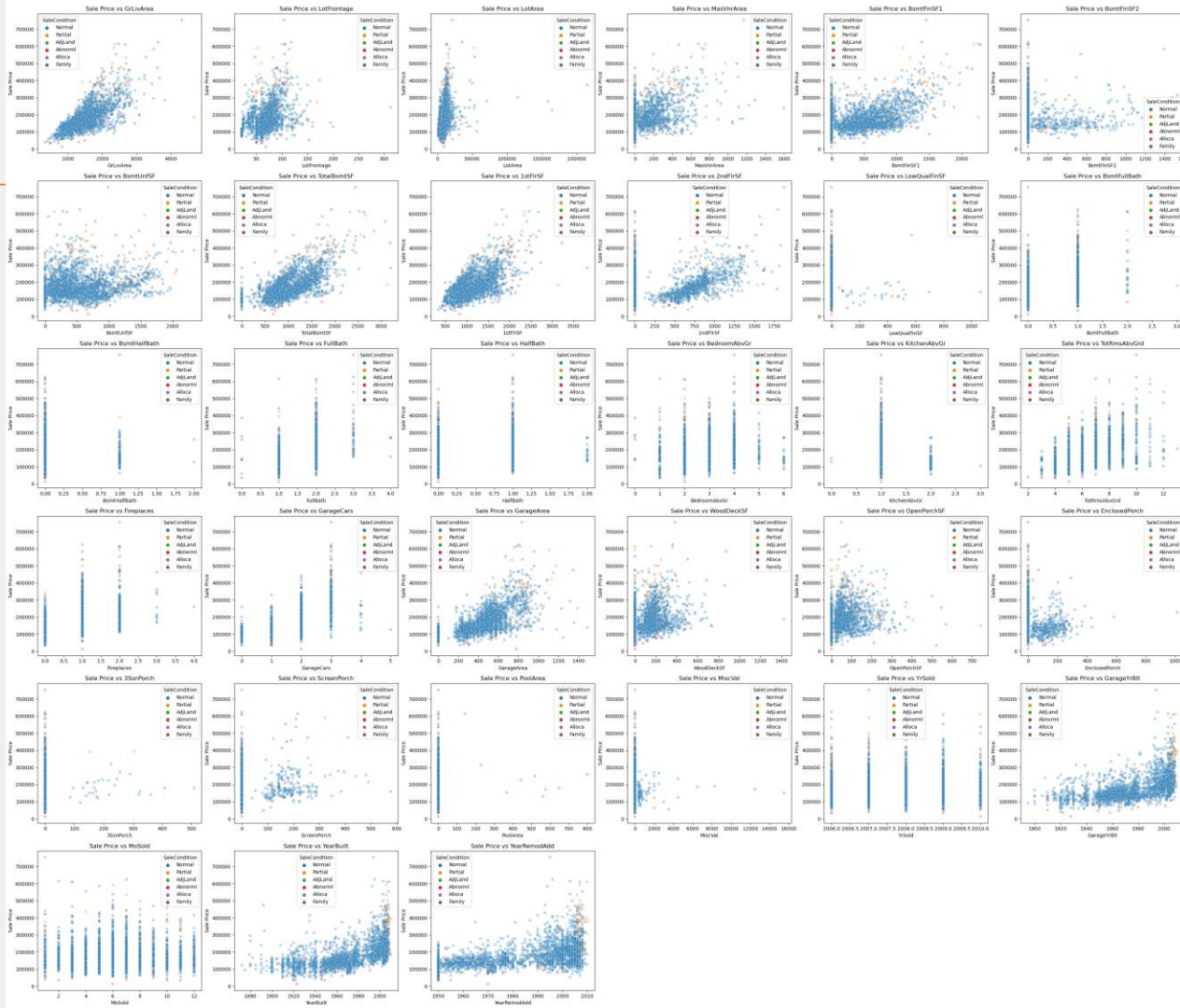
    3. Replace values with most common value or median.

- 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

The insignificant coefficients	
BsmtFinSF2	0.419294
TotalBsmtSF	0.120570
LowQualFinSF	0.610470
BsmtHalfBath	0.939866
FullBath	0.810590
HalfBath	0.652550
GarageArea	0.095607
WoodDeckSF	0.062844
OpenPorchSF	0.958708
3SsnPorch	0.414498
PoolArea	0.084462
MiscVal	0.832196
GarageYrBlt	0.223953
MoSold	0.244385
Alley	0.084301
LandContour	0.372463
Utilities	0.501445
LotConfig	0.370287
LandSlope	0.486086
Neighborhood	0.802727
Condition1	0.870831
RoofStyle	0.335438
Exterior1st	0.320517
Foundation	0.165974
BsmtCond	0.744045
BsmtFinType1	0.101214
BsmtFinType2	0.807077
Heating	0.114950
Electrical	0.605828
FireplaceQu	0.051921
GarageType	0.150826
GarageFinish	0.171744
GarageQual	0.397449
Fence	0.630713
MiscFeature	0.106798
SaleType	0.801740
MixedExterior	0.140310

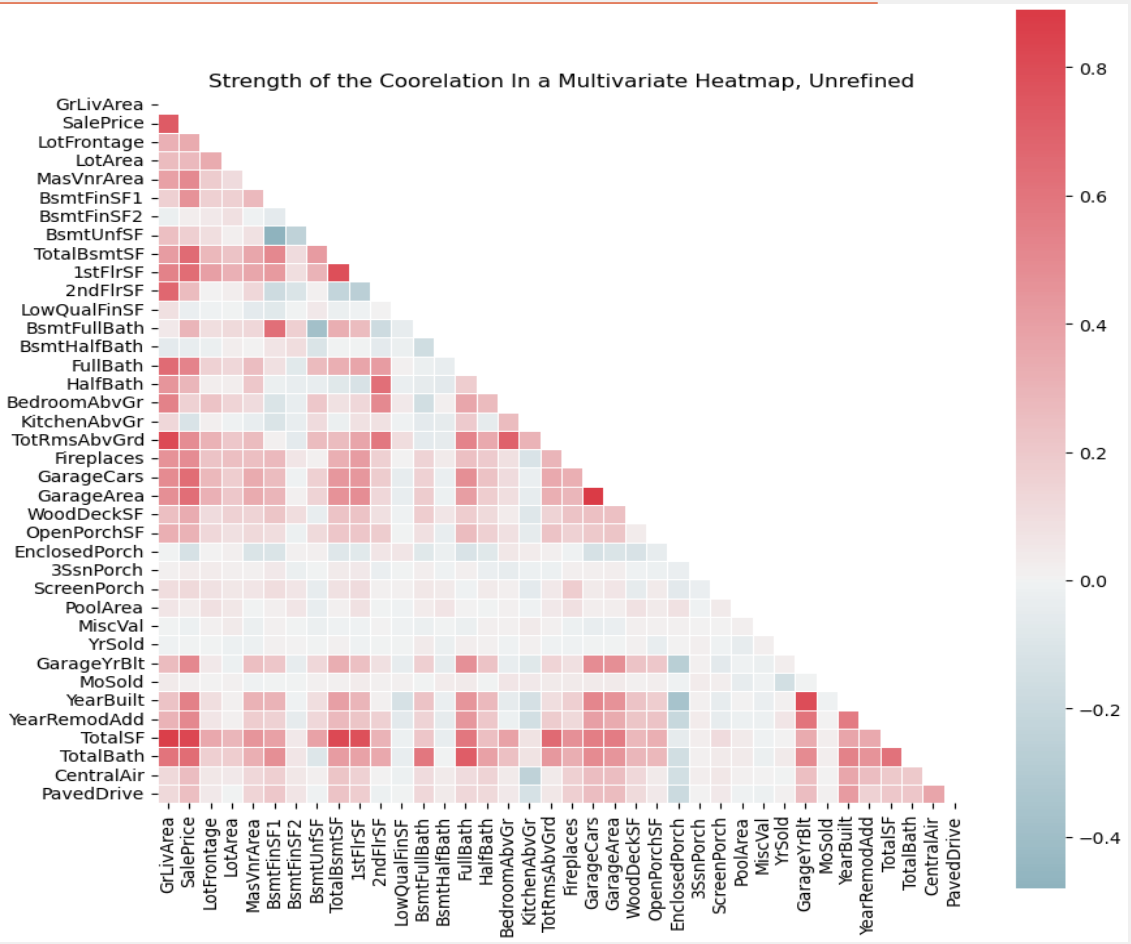
- Since there wasn't a starting point, we ran a model on all columns (with some modifications). Ironically, the score was surprisingly high at .91. However, this is unlikely to be helpful considering how much this model suffers from multi-collinearity.
- Still, this gives us a basis on where to refine the data based on the significant coefficients.

OLS Regression Results

Dep. Variable:	SalePrice	**R-squared:**	0.909
Model:	OLS	**Adj. R-squared:**	0.906
Method:	Least Squares	**F-statistic:**	319.5
Date:	Sat, 01 Jun 2024	**Prob (F-statistic):**	0.00
Time:	04:20:37	**Log-Likelihood:**	1876.0
No. Observations:	2580	**AIC:**	-3594.
Df Residuals:	2501	**BIC:**	-3131.
Df Model:	78		
Covariance Type:	nonrobust		

The significant coefficients	
OverallCond	3.289784e-61
OverallQual	1.291416e-51
TotalSF	4.395247e-26
SaleCondition	7.437340e-19
YearBuilt	1.221690e-18
Fireplaces	8.086013e-14
Functional	3.674940e-12
ExterQual	8.977554e-11
LotArea	1.817810e-09
CentralAir	3.703416e-09
ScreenPorch	4.654287e-09
GrLivArea	1.119912e-08
1stFlrSF	2.399026e-07
KitchenAbvGr	1.052459e-06
PavedDrive	1.706312e-06
EnclosedPorch	4.227983e-06
HeatingQC	8.630677e-06
KitchenQual	1.041197e-05
ExterCond	1.167555e-05
GarageCond	1.237065e-05
BedroomAbvGr	1.880685e-05
BsmtFinSF1	7.123663e-05
const	1.011667e-04
GarageCars	1.552661e-04
BsmtExposure	2.136715e-04
MSZoning	2.693209e-04
YearRemodAdd	3.533935e-04
Exterior2nd	3.964965e-04
BsmtUnfSF	4.434229e-04
TotRmsAbvGrd	8.642359e-04
TotalBath	9.383659e-04
Street	1.228360e-03
2ndFlrSF	4.722108e-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 Quality, is not found.
- From the heat map, we can see that Total Square Footage is not only very statistically significant, it also is strongly correlated with Sale Price. TotalSF is a custom variable, made from combining all SF variables.
- Another note here, Full Bath is highly correlated, yet the unrefined MLR said that it was one of the most insignificant variables, further hinting the data needs to be refined.

Random Forest: As a Model

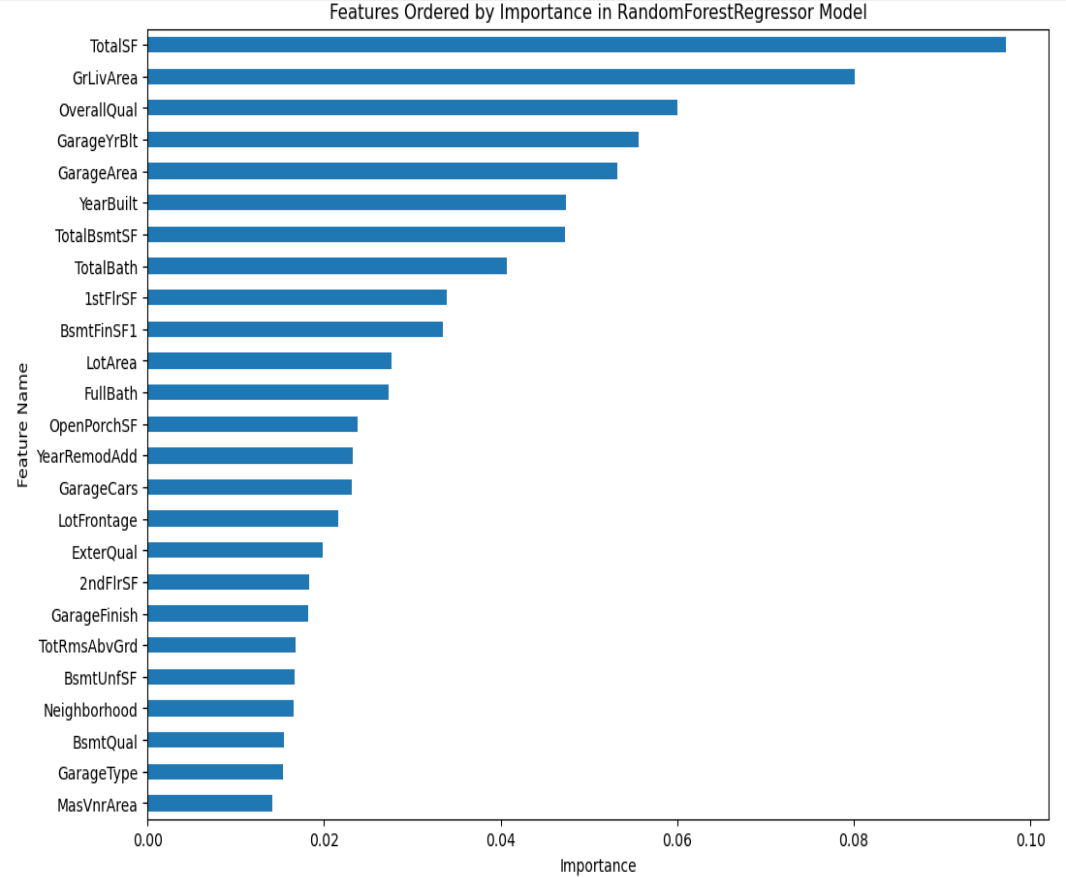
- A random forest model combines multiple decision trees into a single model. In this case, it would determine feature importance rather than MLR's significance and correlation. Feature importance is a degree of dependency on the variables being compared.

- 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".

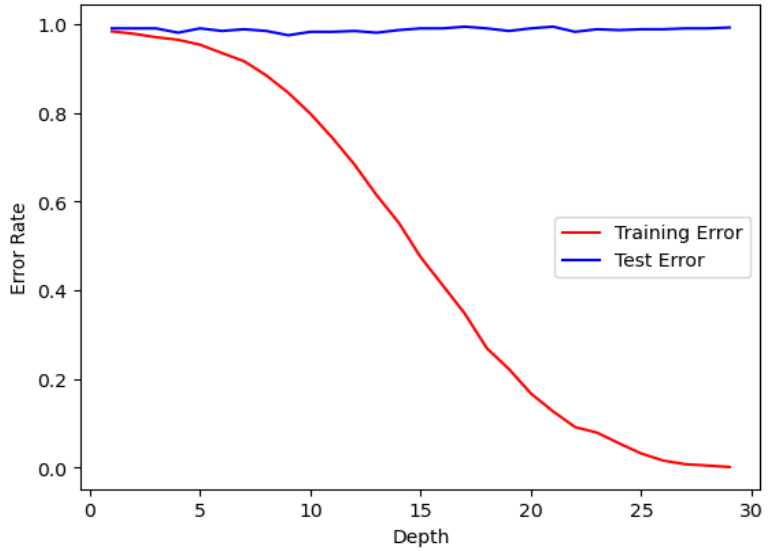
(Image Created Via Stable Diffusion)

RF: Unrefined

- The unrefined RF shows a vague similarity to the unrefined MLR, with OverallQual having high importance, as well as TotalSF having the highest importance.



- However, the decision tree tells us that this graph is overfitting almost immediately. With a score that varies from .67-.87 based on one or two features, this model's predictability is questionable at best.



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 OverallQual is probably important, as are variables like YearBuilt, GrLivArea, FirePlaces, LotArea, etc..

- We also know that several variables can be removed by default. TotalSF, as mentioned, combines three other variables. TotalBath combines four other variables, and MixedExterior combines two other variables, all of which are mentioned as being important. This gets rid of issues where the “parts” were considered insignificant, but the whole was not.

(Image Created Via Stable Diffusion)

MLR: Refined

- After refining the data, the MLR resulted in a lower score, but a 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	1.798756e-125
OverallQual	1.043644e-76
YearBuilt	9.200996e-65
TotalSF	3.355548e-25
GarageArea	2.659971e-24
Fireplaces	9.176173e-21
const	5.236395e-13
GrLivArea	1.009464e-08
LotArea	1.874905e-08
LotFrontage	1.369393e-06
MixedExterior	5.000707e-06
MasVnrType	8.837031e-06
MasVnrArea	1.060431e-04
GarageYrBlt	2.005872e-04
GarageType	1.013372e-03
LotShape	4.684713e-03
BsmtFinSF1	5.976681e-03
GarageFinish	1.168252e-02
BsmtUnfSF	2.114778e-02
LowQualFinSF	4.566449e-02
TotalBath	4.596552e-02

dtype: float64

OLS Regression Results

Dep. Variable:	SalePrice	**R-squared:**	0.879
Model:	OLS	**Adj. R-squared:**	0.878
Method:	Least Squares	**F-statistic:**	808.7
Date:	Thu, 30 May 2024	**Prob (F-statistic):**	0.00
Time:	04:52:57	**Log-Likelihood:**	1513.0
No. Observations:	2580	**AIC:**	-2978.
Df Residuals:	2556	**BIC:**	-2838.
Df Model:	23		
Covariance Type:	nonrobust		

- Some of the results are also identical to the unrefined MLR model, such as with the top two significant coefficients.

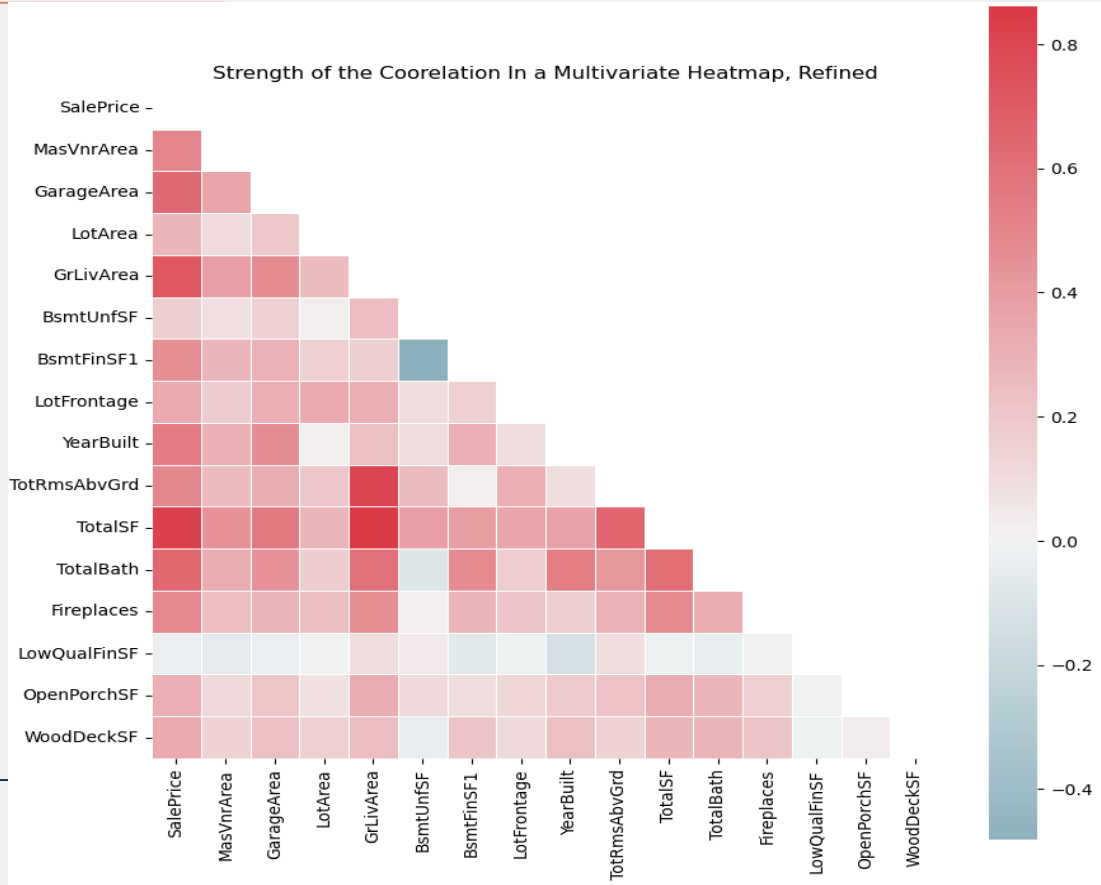
The insignificant coefficients

TotRmsAbvGrd	0.087631
Neighborhood	0.552748
MSSubClass	0.723385

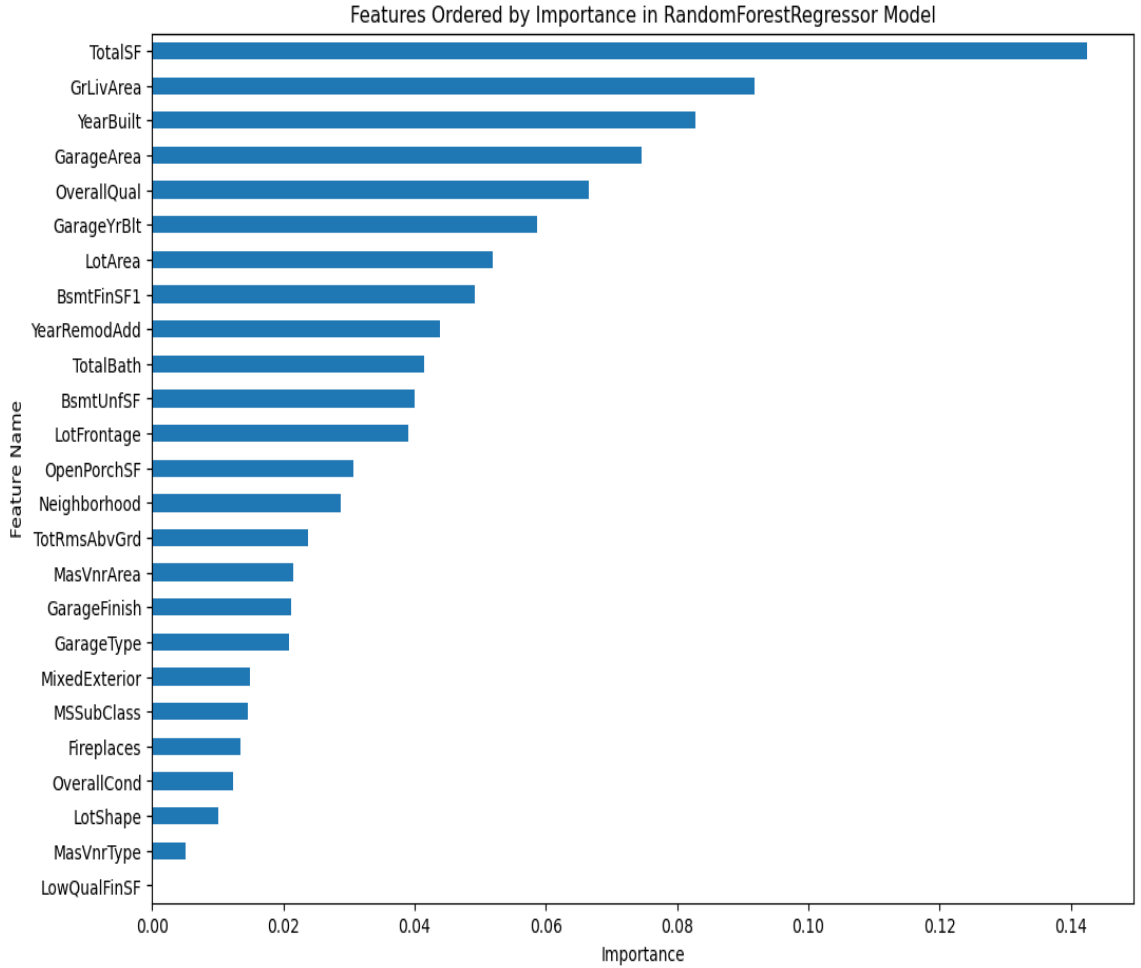
dtype: float64

MLR Refined Correlation Summary

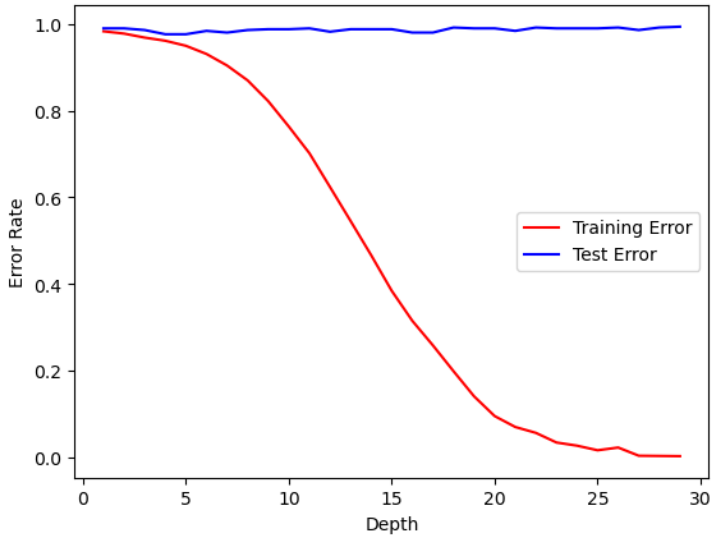
- Here, we can see our results compared to the unrefined findings. The improved F-statistic told us there was a stronger correlation, and here we can see nearly every variable is “strong”, even against each other.
- TotalSF, like with the unrefined model, is still clearly the strongest variable across the board.
- What’s interesting to note is that LowQualFinSF seems to be the weakest of the correlations, yet even the refined MLR model didn’t consider it statistically insignificant.



RF: Refined



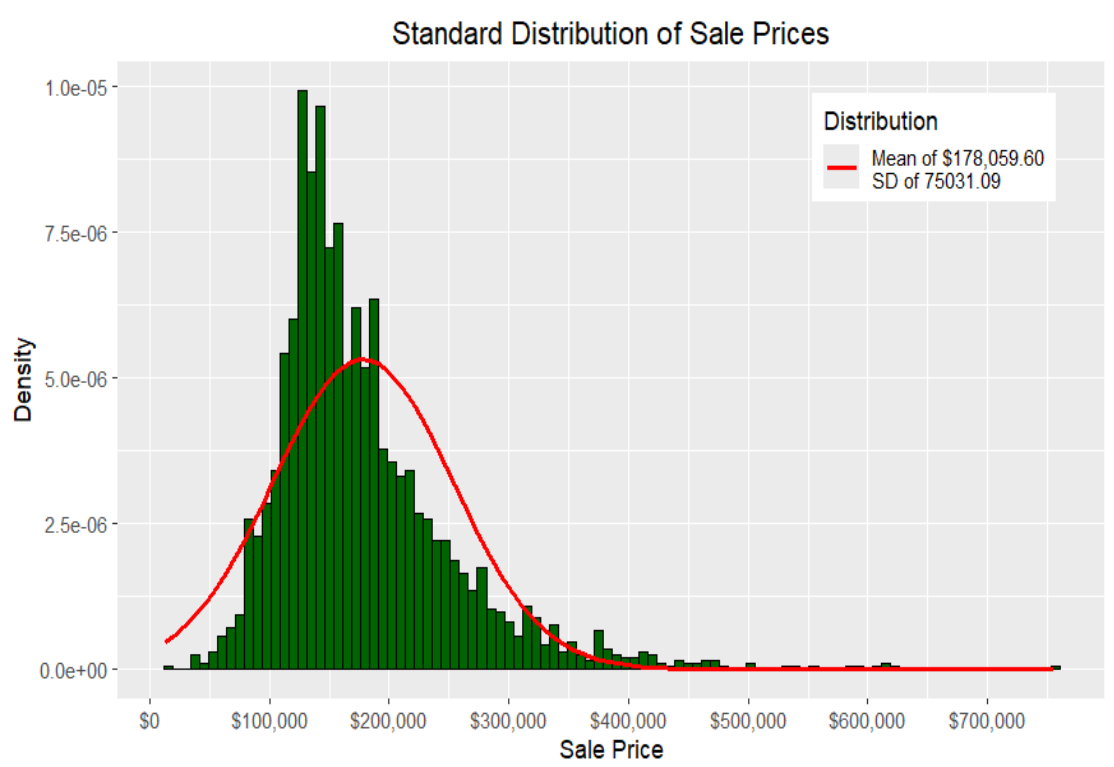
- The results here are more in-line with the correlations, with TotalSF still being the most important variable. Notably, LowQualFinSF also has zero importance here, which matches how it had very low correlation. However, this model still has a drastic overfitting issue.



Breaking Down the Results: The Target

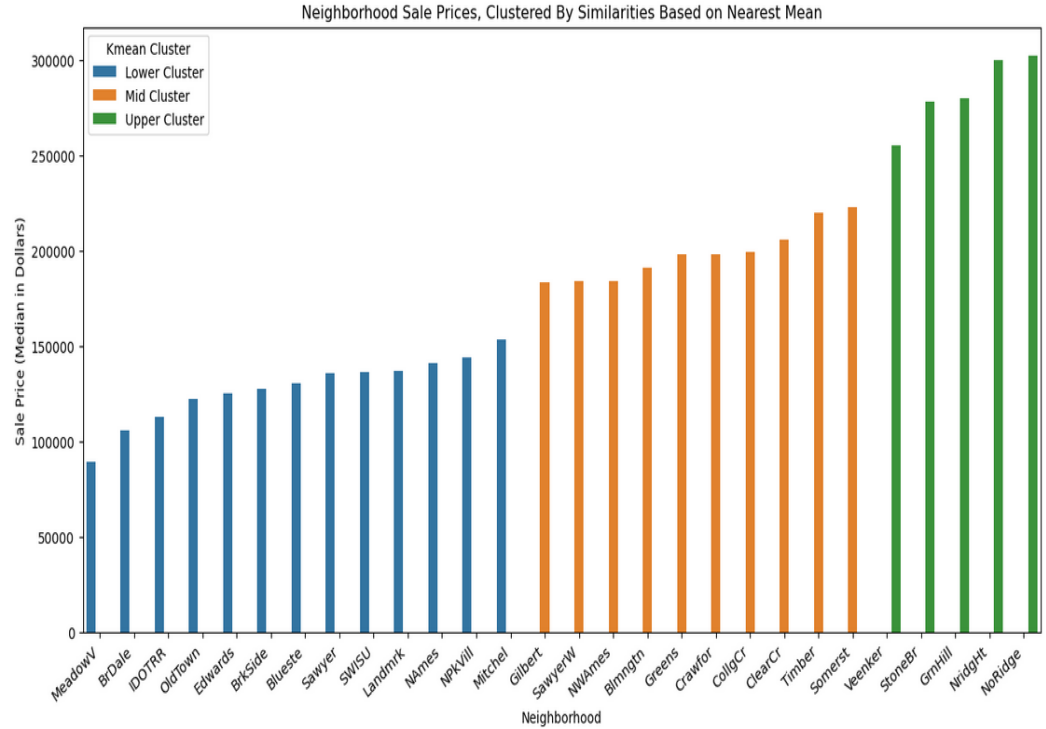
- To start an analysis, we can look at the target of the models: the Sale Price. Using a standard distribution histogram, we can see that most houses have been sold at around $180,000. This is fairly typical, as most people cannot afford more expensive houses.

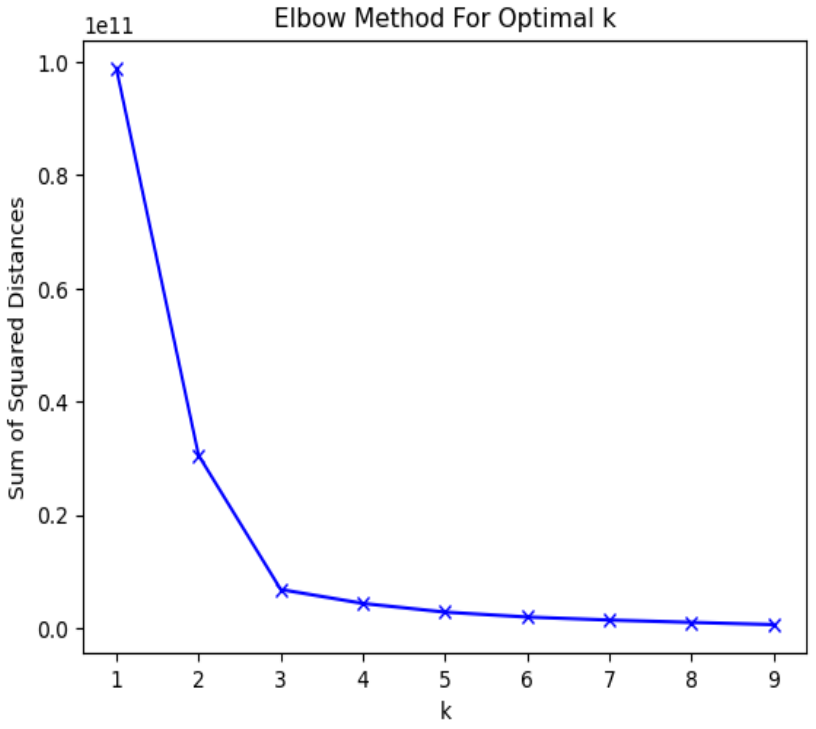
- However, it should also be kept in mind that there are numerous outliers in the data because of this. In fact, one can even be seen here, with a house being sold at over $700,000. This means every variable, such as TotalSF, TotalBath, etc. would likely distort the data compared to what should be the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

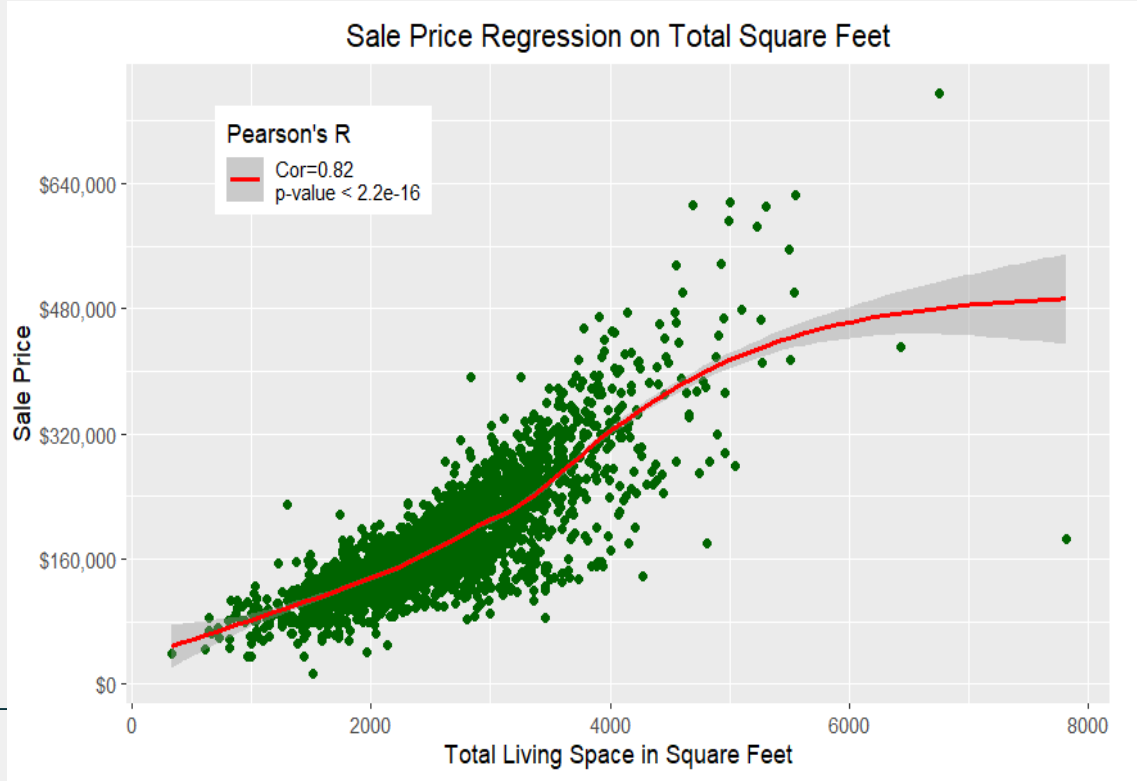




Most Important Feature: Numeric

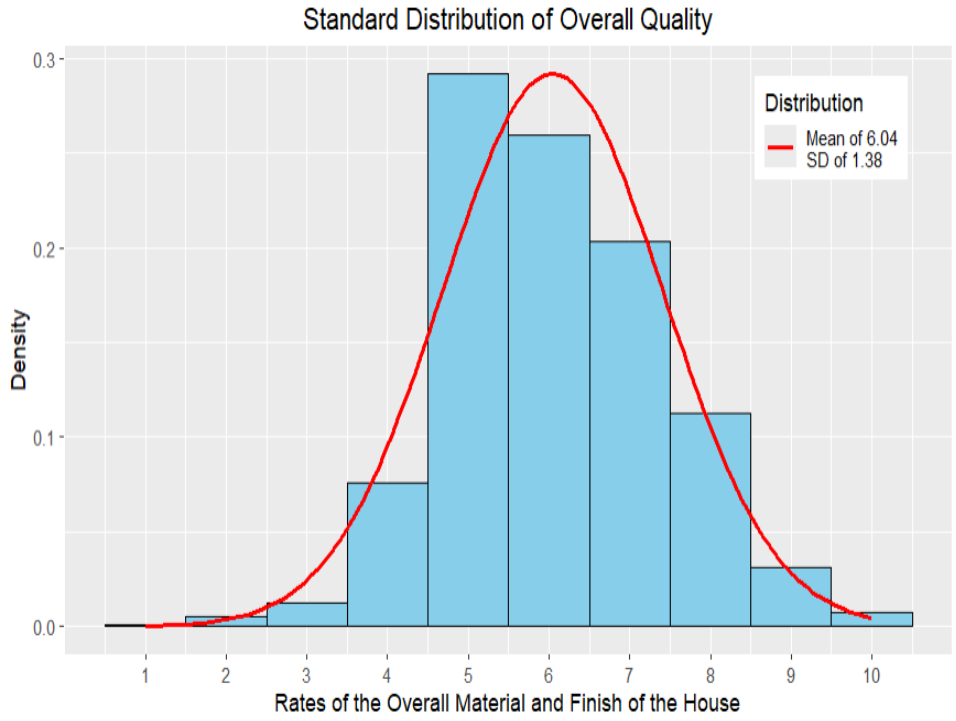
- According to our models, the most important numeric feature is TotalSF, which is the combination of the SF variables (TotalBsmtSF, 1stFlrSF, 2ndFlrSF). From a weighted regression model (loess), we can see it has a nearing-perfect positive correlation with Sale Price. This aligns with what both models said.

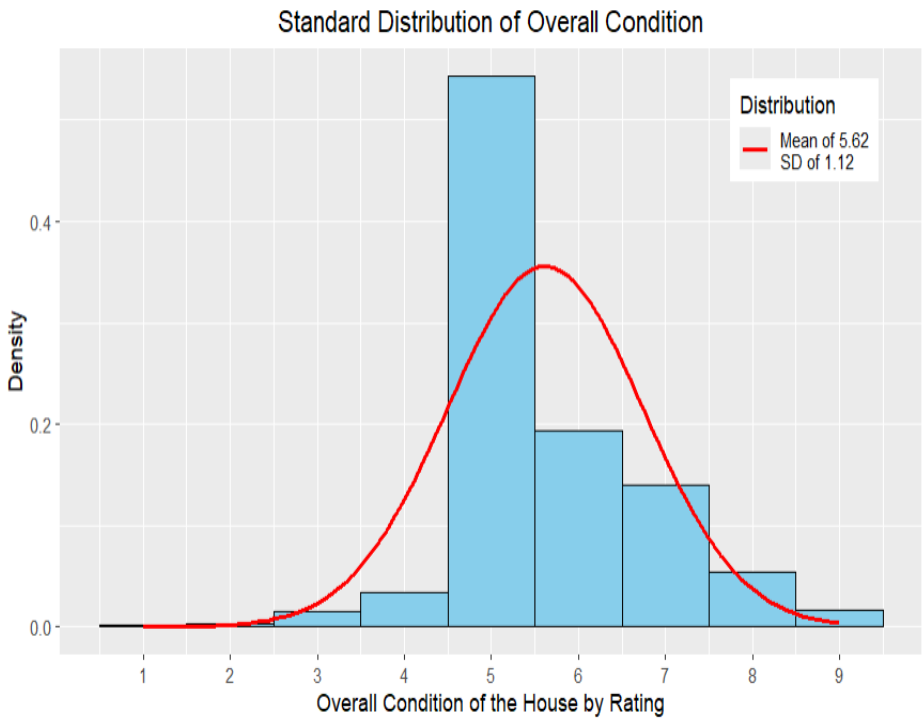
- This has only two visible outliers, but they don't distort the regression line enough to shift it in a meaningful way.



Most Important Feature: Categorical

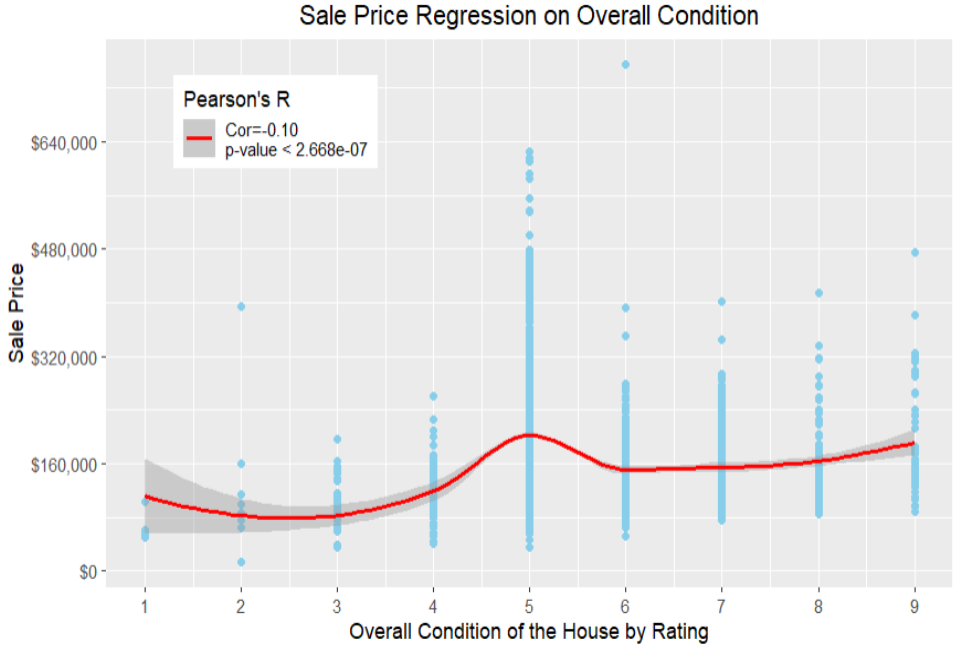
- Our models seem to at least agree that OverallQual is both important and significant, but the random forest model rates OverallCond lower in importance, while the MLR model rates it extremely high in significance. The meaning of this will be examined.
- Right away, we can see smaller variation in the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





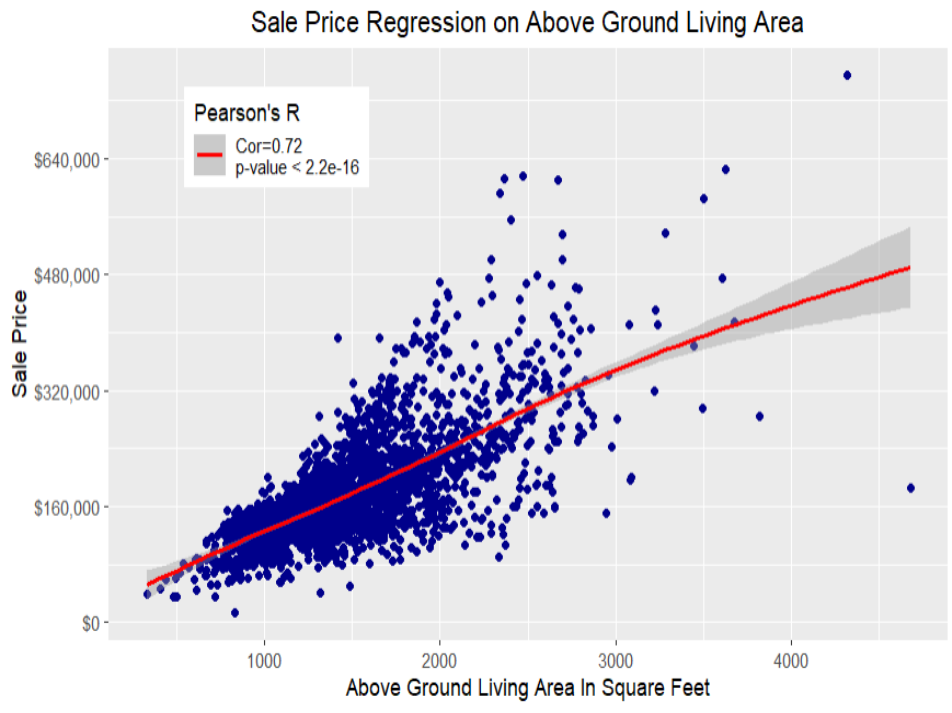
Weighted Regression on Housing Rates

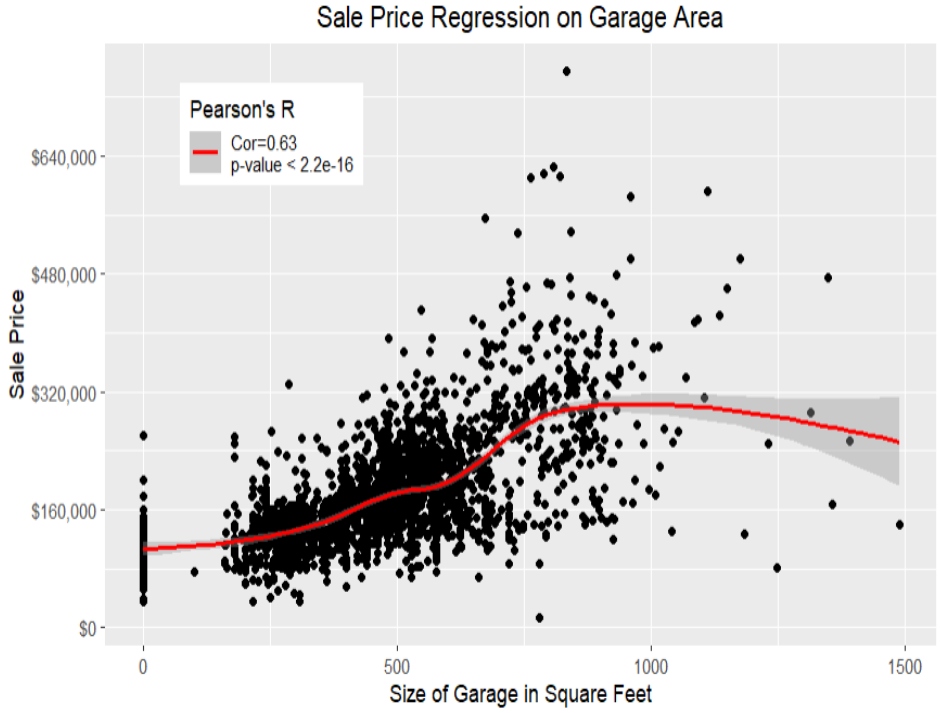
- As expected, the OverallQual variable has a much higher correlation than the OverallCond variable. So much so that OverallCond goes into the negative. This may explain why OverallQual is so much higher in the RF model. As shown in the previous graphs, the smaller variability in OverallCond caused so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



Other Highly “Important” Features

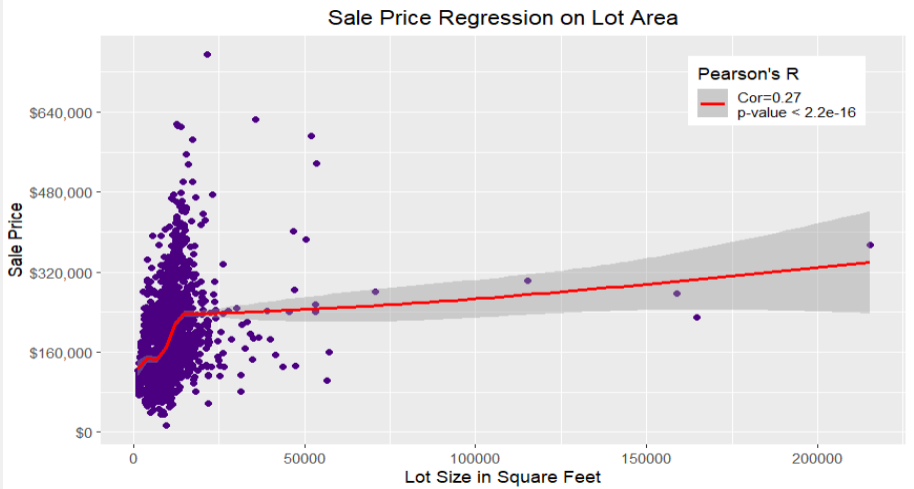
- The other most important features, according to the RF model, are similar to the TotalSF. They are highly correlated, along with being statistically significant. GarageArea was actually considered one of the most significant factors, according to the MLR model, but here we see GrLivArea coinciding 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.

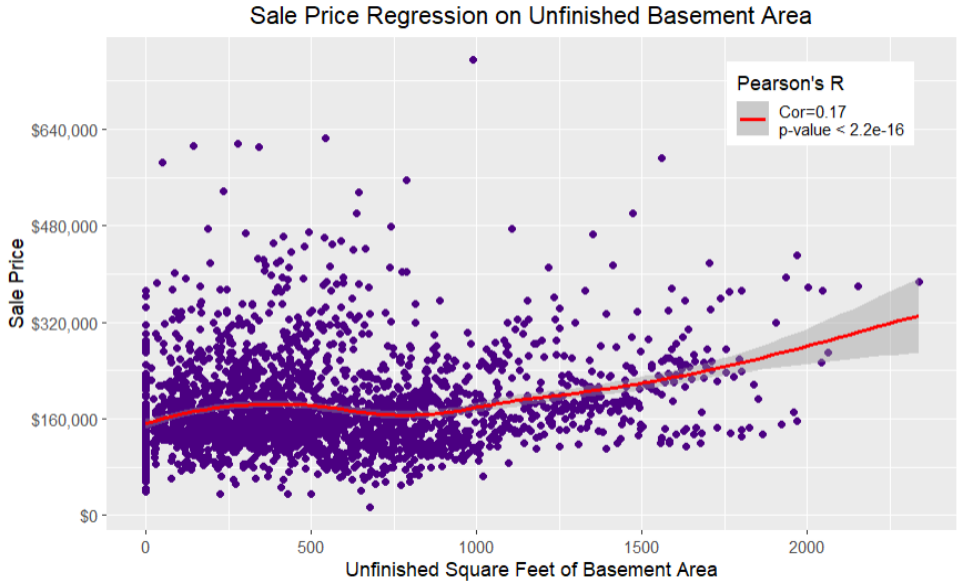


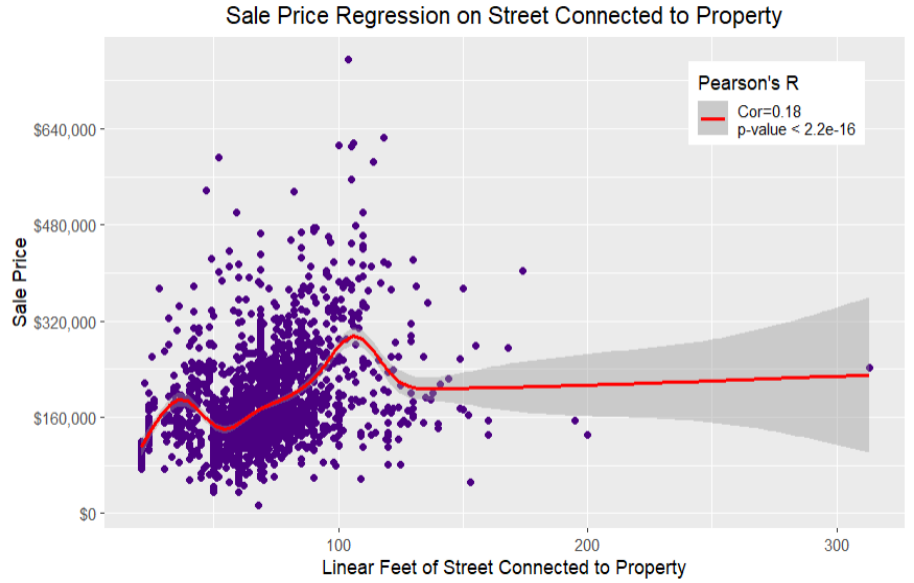


The Weaker Correlations

- The RF model claimed LotArea, BsmtUnfSF, and LotFrontage all had “decent” importance. The MLR model claimed BsmtUnfSF had one of the lowest (as in not good) statistical significances. When actually tested, we see that the correlation for all, except for Lot Area, are particularly poor. Notably, the unrefined model considered LotFrontage and BsmtUnfSF unimportant, which contradicts the refined model.

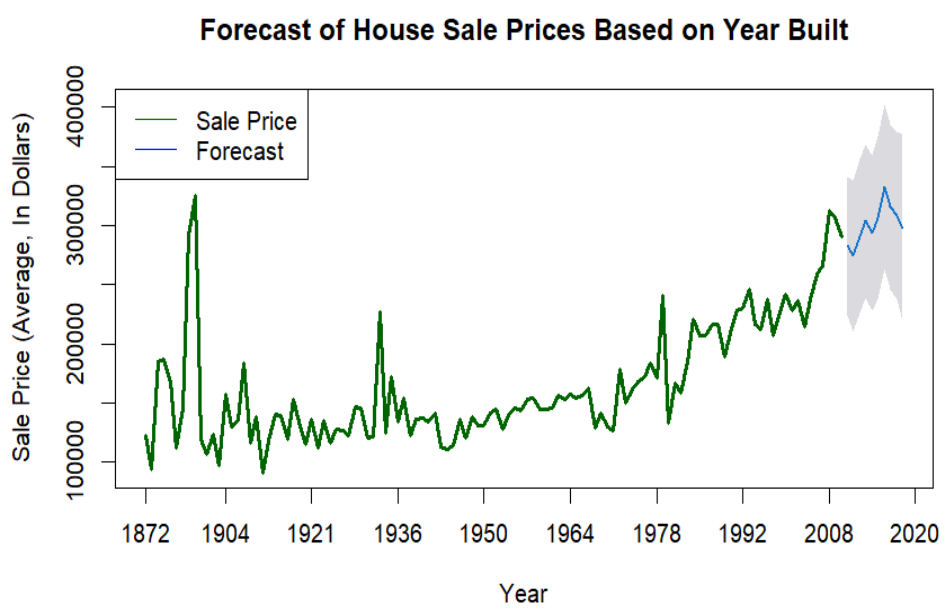


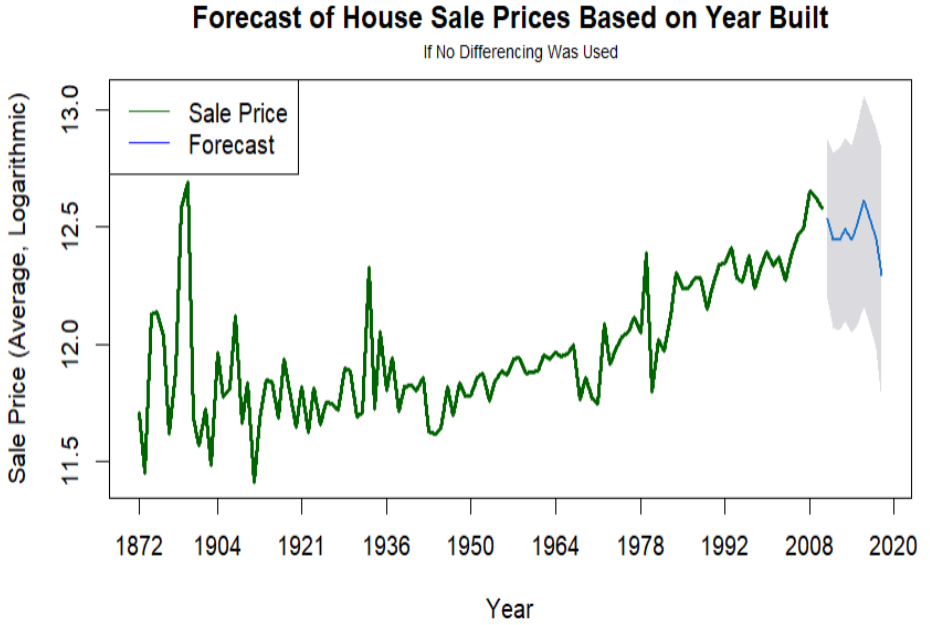




Time Series Model: Year Built

- The models determined that one of the most significant and important features was YearBuilt. This is a special variable, as it is a date variable. To analyze this, a time series model would be used. Here, we can see a clear trend that SalePrice increased, with a drastic increase peaking at around 2006-2008.
- Using an Arima Forecast, the model believes the trend will continue to rise, depending if the series is dependent on time (note that 2010 only had two data points, which caused the data to be non-stationary).





Summary: The Models

- Overall, the models give us mixed results. The unrefined models have a higher score than the refined models, yet the refined models clearly deal with the unique problem of multi-collinearity. Notice how the RF refined score is substantially lower than the unrefined score, yet the included factors do explain importance much more than the unrefined model.

- 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 one of the lowest correlated numeric features, BsmtUnfSF, is the 11th highest feature of importance in the refined model, but the 4th lowest in the unrefined model.

- The immediate overfitting is apparent for both the unrefined and refined RF models, and it should also be noted that the AIC is 500 points lower in the unrefined model. This is supposed to indicate a more optimized model, although the BIC is only around 100 points lower.

MLR Unrefined Score	MLR Refined Score	RF Unrefined Score	RF Refined Score
0.909	0.879	0.857	0.246

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.

- 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.

- We noted that the weakest relationships were inconsistent when displayed in the unrefined and refined models, such as how certain ones were considered more/less important in the refined model, and vice versa. This was also reflected when tested alone.

- The time series model was fairly typical, marking the build-up to the 2008 financial crisis. The forecast was divided between predicting an uptrend and a downtrend, depending on differencing because of such an "odd" event.

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 gradient boosting, lasso regression, gridsearch, and feature engineering.

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?

(This is just a random graph to further demonstrate discrepancies, as TotalBath has a higher correlation and significance here than OverallCond)