

Kaggle Competition:


House Prices - Advanced Regression Techniques


REPORT WRITTEN BY/KAGGLE NAME: MUHAMMAD FARRUKH KHAN, Username: mfk66666

STUDENT ID: 20200898 – SECTION 2.

9,417 teams – 2581st Position – Top 28%

Link to notebook: <https://www.kaggle.com/mfk66666/housing-mfk>

 Search




Muhammad Farrukh Khan

Add location

Master of Management Analytics at Queen's University @ Smith School of Business

Joined a month ago · last seen in the past day






Competitions
Novice

[Home](#) [Competitions \(1\)](#) [Datasets](#) [Code](#) [Discussion](#) [Followers](#) [Notifications](#) [Account](#) [Edit Profile](#)

Competitions
Novice




Unranked

  
0 0 0

House Prices - ...
Ongoing
Top 28%
2,581st
of 9417

Datasets
Novice




Unranked

  
0 0 0

No dataset results

Notebooks
Novice




Unranked

  
0 0 0

No notebook results

Discussion
Novice

Unranked

  
0 0 0

No discussion results

Bio

Edit

Click to add bio...

Followers

COMPETITION SCREENSHOTS FOR REFERENCE

<div> Search </div>						
Overview	Data	Code	Discussion	Leaderboard	Rules	Team
				My Submissions	Submit Predictions	
2572	Chandrima D	</> House_Price_Predic...			0.12884	1 4mo
2573	Dursun Can Özdemir				0.12884	24 2mo
2574	Ethan Markwalter				0.12885	44 1mo
2575	[Deleted] 0c78213a-d787-41a...				0.12885	8 1mo
2576	Richard Pletan				0.12886	17 1mo
2577	Silpa Noolu				0.12886	15 1d
2578	Benjamin Ai				0.12888	31 1mo
2579	Shakti Sampad Dash				0.12889	25 2d
2580	abcxyz #2				0.12889	14 4mo
2581	Muhammad Farrukh Khan				0.12889	8 ~10s
<div> Your Best Entry </div> <div> Your submission scored 0.12889, which is an improvement of your previous score of 0.13202. Great job! Tweet this! </div>						
2582	Yousra El Alaoui				0.12890	24 1mo
2583	iimaane				0.12890	11 1mo
2584	BUPUNK				0.12890	5 2mo
2585	RISHI BAIJAL (PGP 2016-18)				0.12890	3 3mo
2586	danish bansal				0.12891	6 2mo
2587	soumik saha				0.12893	11 2mo
2588	Akio Onodera				0.12893	37 25d
2589	Aaron Mayzes				0.12893	2 1mo
2594	chataoui hamza				0.12896	2 4mo
2595	Sam Perng				0.12896	8 1mo
2596	jay baek				0.12897	4 1mo
2597	Yamaguchi Shin				0.12898	1 1mo
2598	Lihong Tang				0.12898	1 1mo
2599	Meng-Huan Wu				0.12899	2 2mo
2600	Naruhiko Nakanishi				0.12899	1 3mo
2601	yuya hong				0.12900	9 3mo
2602	Shailesh Dwivedi				0.12901	3 2mo
2603	Russel Abreo				0.12901	14 33m
2604	Marion Hesse				0.12902	16 3mo
2605	NishimiTakeru				0.12902	1 4mo
2606	Tomonari Sadamura				0.12903	1 4mo
2607	Muina_Irina				0.12906	5 24d
2608-9417 Load 6810 More						

COMPETITIONS IDENTIFIED

Identified the following three competitions:

1. COVID Global Forecasting: Forecast daily COVID-19 spread in regions around world
2. Predict Future Sales: Final project for "How to win a data science competition" Coursera course.
3. House Prices – Advanced Regression Techniques

Decided to model House Prices prediction out of the three chosen ones. The most important factor was familiarity with data which I could comprehend easily and be able to make more sense compared to other competitions. For example, it's still not clear which factors are highly valuable to determining number of cases of covid. While for housing prices there is a lot of information available online to determine the main factors. Considering practicality purposes one is also aware of these factors in practice. Also, understanding the variables which are the main drivers of predicting housing prices will be beneficial for me personally, as I am involved in Real Estate business in my family.

There are also a lot of integer variables that can be utilized to build up housing prices and for the purposes of checking correlation. Other competitions had more qualitative variables which always makes it difficult to determine the right ones without some prior knowledge.

The range of variables and their non-collinearity is good for prediction as there are a lot of variables with different nature to each other and they could have an impact on the data as independent variables. Feature Engineering can also be done for several variables, which is also one of the main reasons I chose this dataset. Although, there are missing values, but the data also seems to be clean for making a model, which is an important reason for choosing this competition.

REGRESSION MODEL

Gathering Data

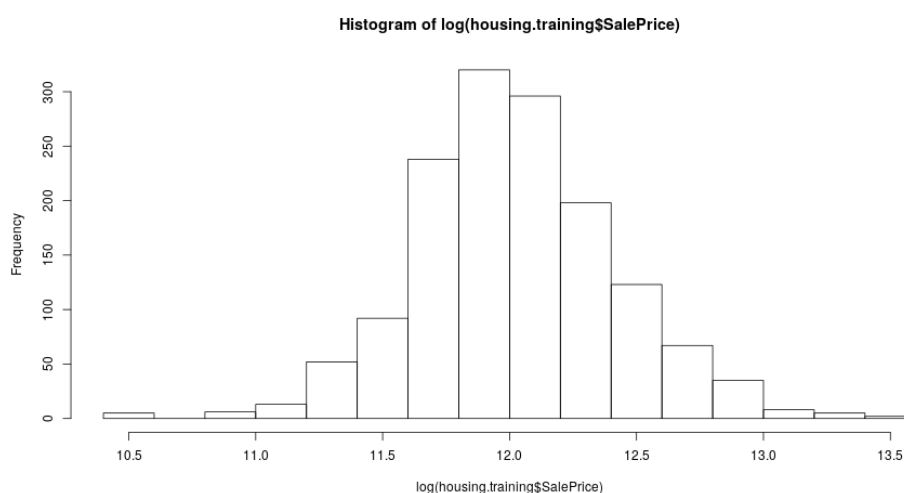
Train and Test data was provided in two separate CSV files. They were imported into the notebook and combined using R bind to create one data frame for further data exploration. A Sale Price column was added to the testing data set for our predictions with a temporary zero value.

Exploratory Data Analysis & Feature Engineering

Since understanding the variables in the data is crucial, we explored the data using a few useful functions of R such as str, head and summary. The data frame resulted in 2919 observations of 81 variables out of which 1460 would be used to train the model, while rest will be used to predict the Sale Price. The data had a mix of categorical and integer variables that can be utilized in intersections for an accurate model.

Using the 'mice' library, we figured that the data has quite a few numbers of rows missing in some variables. In total 13,965 values were missing that need to be dealt with. Some variables like PoolQC, MiscFeature and Alley had more than 90% of data missing. The engineering of these variables will covered under Data Preparation.

A histogram was created for Sale Prices to see their distribution. It was discovered that the data was rightly-skewed which would cause problems in regression, this would be fixed by the Log function to make the data normally distributed as shown below by the histograms below.



Plots were made on Tableau to showcase relationship of different numeric variables to Sale Price. It was also checked if those variables had a variance in the relationship with the effect of categorical variables.

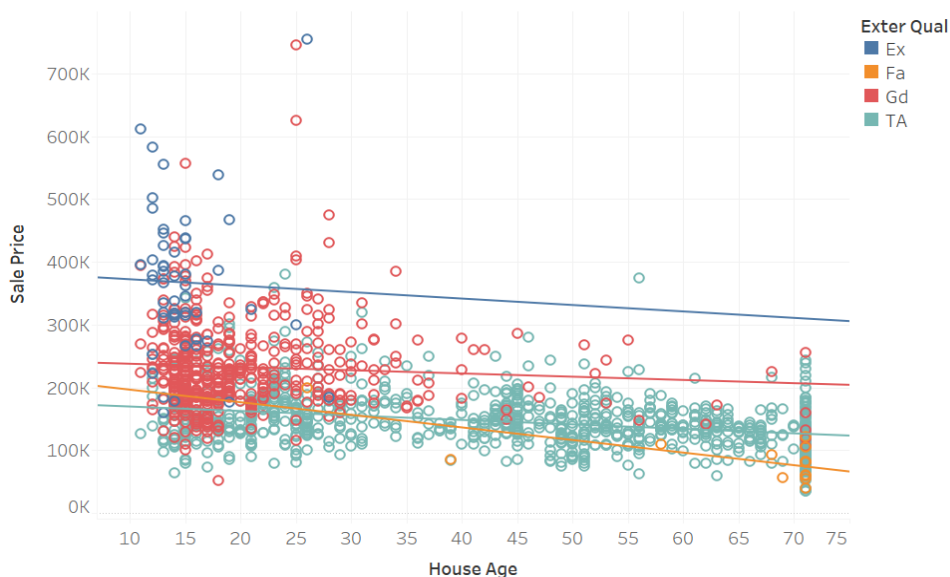
```
cor(housing.train[,unlist(lapply(housing.train, is.numeric))])

options(max.print=9999)
```

VARIABLES GIVING HIGH CORRELATION TO SALE PRICE USING ABOVE CODE WERE PLOTTED ON TABLEAU.

House Age was feature engineered from the YrBuilt variable to determine the age of the houses from current year and this was plotted against Sale Price to check their relationship. The scatterplot showed a negative relationship that means as the houses got older, their Sale Prices declined. One interesting relationship that was discovered that the Exterior Quality of the houses determined the slope of the linear relationship between House Age and Price. As we can see the Excellent Quality line has a higher y-intercept along with less steep slope showcasing a different relationship as other variables. These deductions will be taken into consideration for the model building.

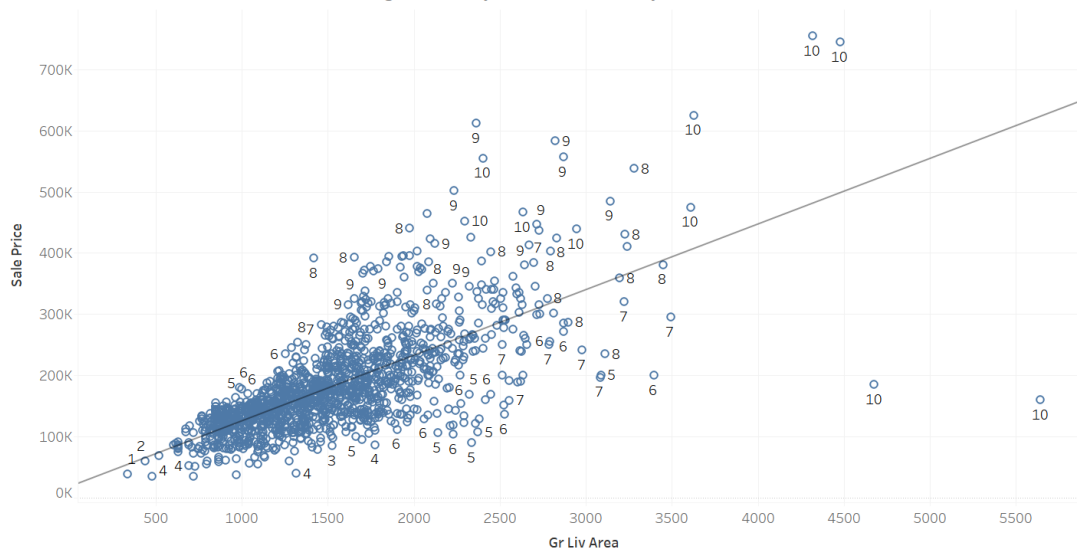
House Age vs Price - Categorized by Exterior Quality



House Age vs. Sale Price. Color shows details about Exter Qual.

Above grade (ground) living area square feet vs Price relationship seemed to be a positive one as shown by the scatterplot below. An Overall Quality that is higher on the 1-10 scale will have a higher Sale Price with similar living area. This interaction will be used in the model.

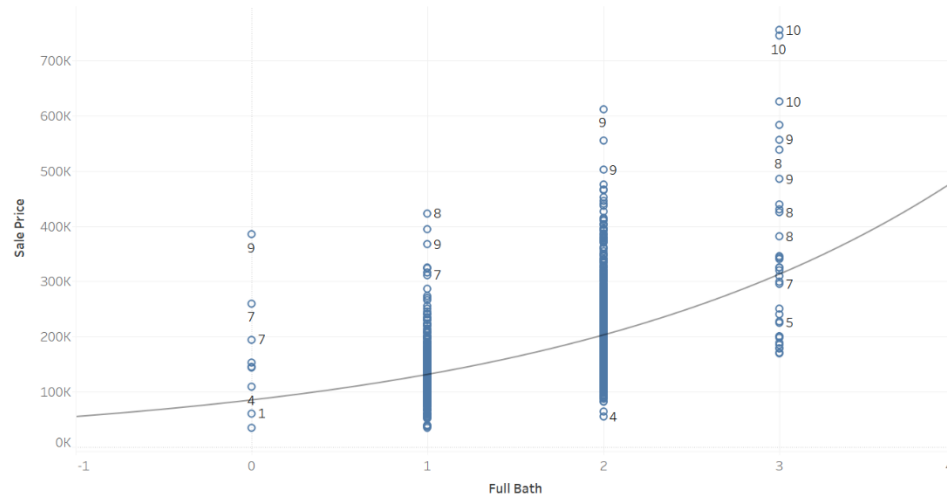
Above Grade Ground vs Price - Categorized by Overall Quality



Although, there were a few outliers with the living area higher than 4000 sq. ft. Outliers were dealt with the log function in the model. There seems to be heteroscedasticity in the ground living area plot above which will also be fixed by using Log. Outliers were also fixed to an extent by regularization and using the Lasso regression model.

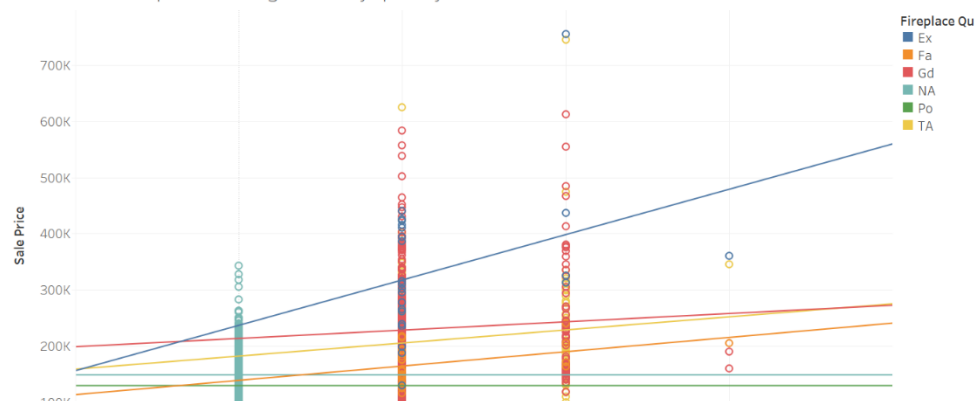
Plots of other variables showing a relationship with Price:

Full Bathrooms vs Price

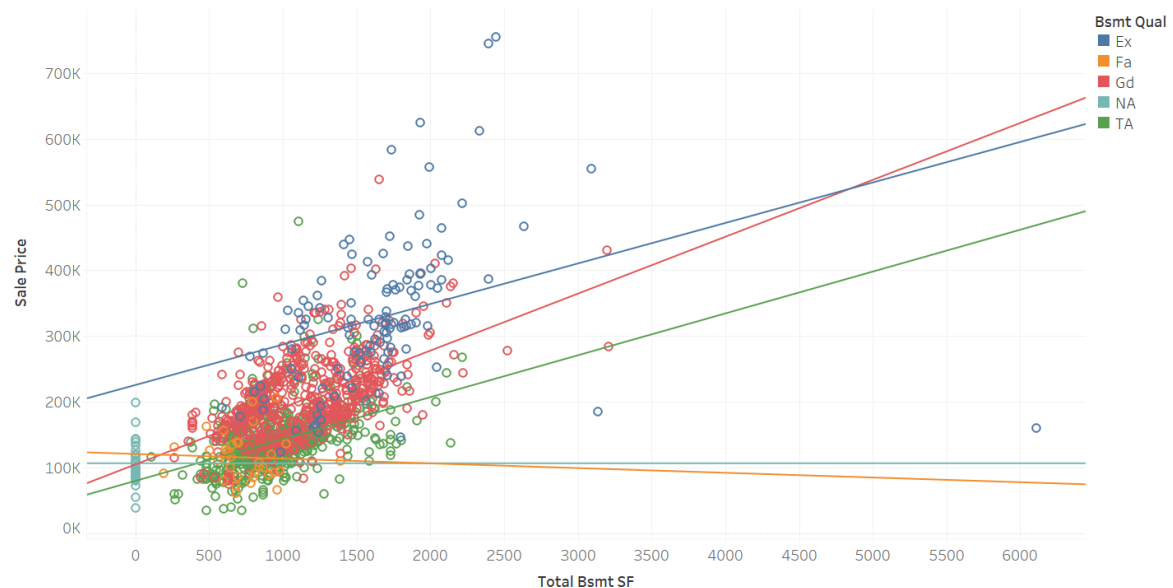


Full Bath vs. Sale Price. The marks are labeled by Overall Qual.

Number of Fireplaces categorized by quality

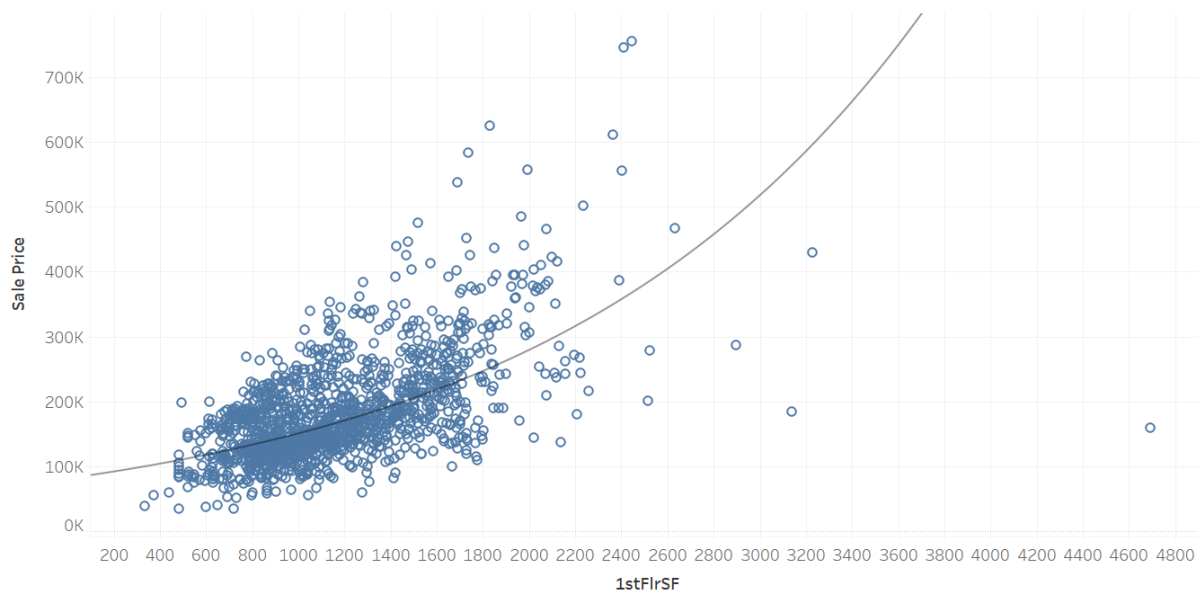


Total Square Feet of basement Area - Categorized by Basement Quality



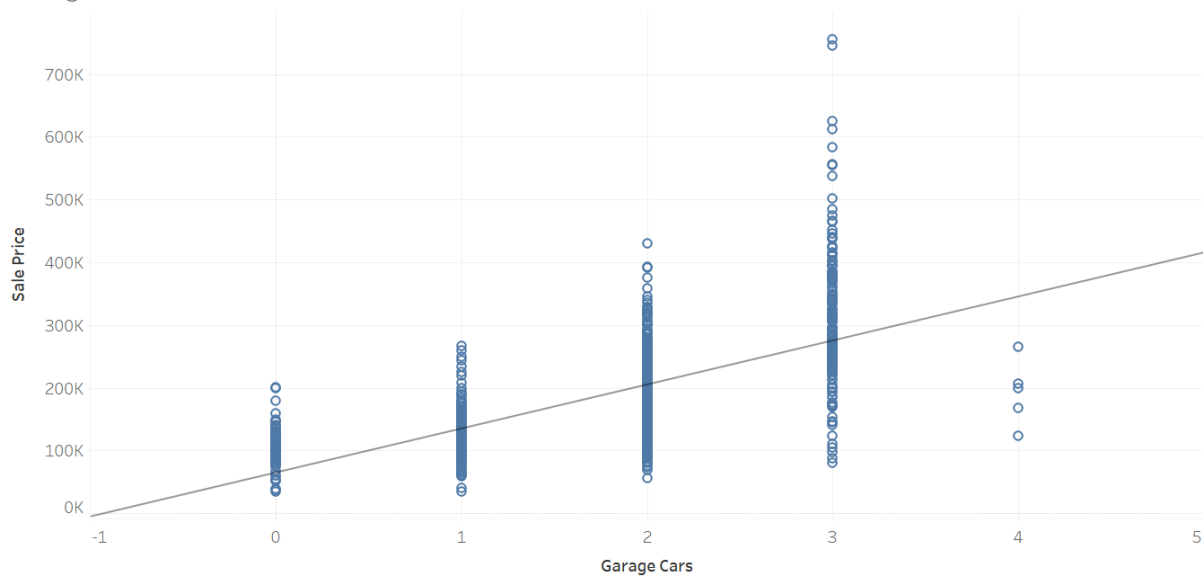
Total Bsmt SF vs. Sale Price. Color shows details about Bsmt Qual.

1st Floor Surface Area vs Price



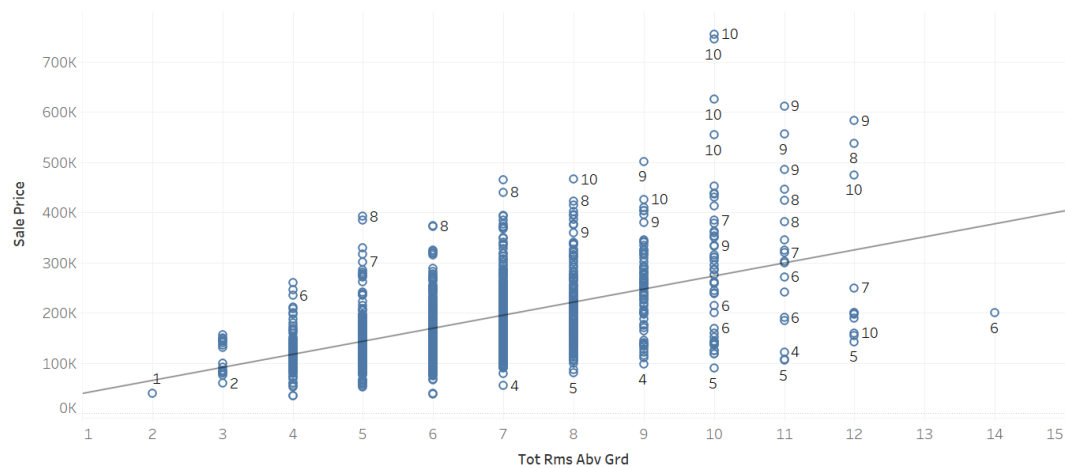
1stFlrSF vs. Sale Price.

Garage Cars



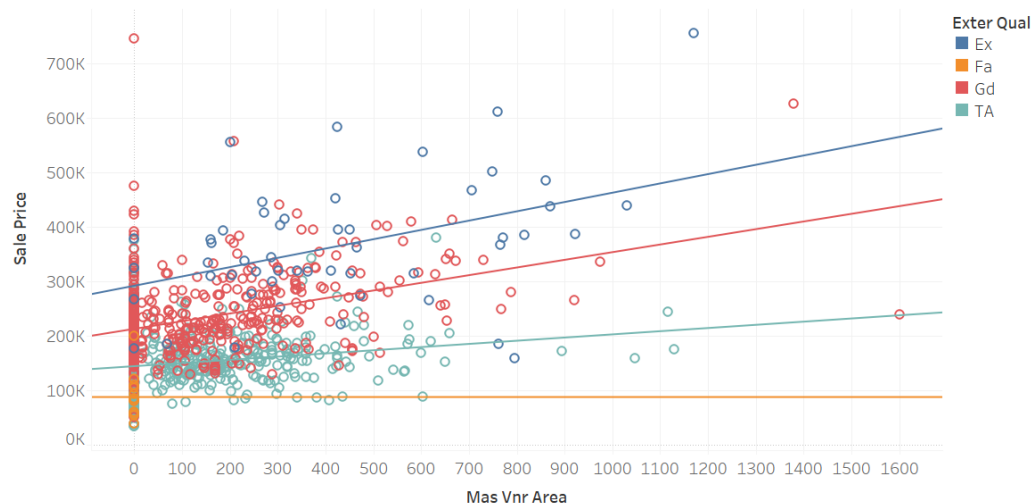
Garage Cars vs. Sale Price.

Total Rooms Above Grade



Tot Rms Abv Grd vs. Sale Price. The marks are labeled by Overall Qual.

Mas Vnr Area vs Sale Price



Mas Vnr Area vs. Sale Price. Color shows details about Exter Qual.

So, the plots of all the integer variables that showed a high correlation with Sale Price are shown above. The correlation of 50% and higher was considered good enough in most cases, as this was checked using the 'cor' functions. Using the Trend lines, it can also be seen that some variables affect Sale Price differently when interacted with categorical variables like Exterior Quality, Basement Quality and Overall Quality.

There is also heteroskedasticity and some outliers for Basement Area, 1st Floor SF Area and Mas Vnr Area. This was dealt with using Logs and square root and a Linear Regularized model.

Data Preparation

Variables with missing data were imputed using different methods according to their class and statistics.

Missing Values -

Columns that had 50% or more of the data missing were categorical variables with factor class. These variables were Alley, FireplaceQu, PoolQC, Fence & MiscFeature. It can be fair to say that the missing data in these features meant they were not part of those housing projects as we could also cross-check a few with their related variables. An additional level of 'None' was added to the existing level of these factors to fill in missing data for our model. (Coding in Notebook).

```
housing.data$Alley = factor(housing.data$Alley, levels=c(levels(housing.data$Alley), "No_Access"))
housing.data$Alley[is.na(housing.data$Alley)] = "No_Access"

levels(housing.data$FireplaceQu)

housing.data$FireplaceQu=factor(housing.data$FireplaceQu,levels=c(levels(housing.data$FireplaceQu), "None"))

housing.data$FireplaceQu[is.na(housing.data$FireplaceQu)]= "None"

housing.data$Fence=factor(housing.data$Fence,levels=c(levels(housing.data$Fence), "None"))

housing.data$Fence[is.na(housing.data$Fence)]= "None"

# %% [markdown]
# PoolQC & MiscFeature will also be replaced with None for housing that doesnt have these features

#PoolQC
housing.data$PoolQC=factor(housing.data$PoolQC,levels=c(levels(housing.data$PoolQC), "None"))
housing.data$PoolQC[is.na(housing.data$PoolQC)]= "None"

#MiscFeature
housing.data$MiscFeature=factor(housing.data$MiscFeature,levels=c(levels(housing.data$MiscFeature), "None"))
housing.data$MiscFeature[is.na(housing.data$MiscFeature)]= "None"
```


Variables having less than 20% of missing data were identified to fill in the missing values. Now, this was a mix of integer and categorical class. Integer variables having more than 50% of zeros in their columns were identified and it was assumed the 2 or 3% of missing values could also be zero. The mean was used to impute Mas Vnr Area and Lot Frontage as they were two important variables for our model that showed a high correlation with Sale Price. Variables like Basement Features and Garage features imply that these characteristics don't exist for the specific house, so they were imputed with zero.

```
housing.data$MasVnrArea[is.na(housing.data$MasVnrArea)] <- mean(housing.data$MasVnrArea, na.rm=TRUE)
```

```
housing.data$LotFrontage[is.na(housing.data$LotFrontage)] <- mean(housing.data$LotFrontage, na.rm=TRUE)
```

```
housing.data$BsmtFinSF1[is.na(housing.data$BsmtFinSF1)] <- 0
housing.data$BsmtFinSF2[is.na(housing.data$BsmtFinSF2)] <- 0
housing.data$BsmtUnfSF[is.na(housing.data$BsmtUnfSF)] <- 0
housing.data$BsmtFullBath[is.na(housing.data$BsmtFullBath)] <- 0
housing.data$BsmtHalfBath[is.na(housing.data$BsmtHalfBath)] <- 0
housing.data$TotalBsmtSF[is.na(housing.data$TotalBsmtSF)] <- 0
housing.data$GarageCars[is.na(housing.data$GarageCars)] <- 0
housing.data$GarageArea[is.na(housing.data$GarageArea)] <- 0
housing.data$GarageYrBlt[is.na(housing.data$GarageYrBlt)] <- 0
```

Feature Engineering

Variable named Area was created by combining all surface area factors to check correlation of all those surface variables with Sale Price, but it did not come out to be too significant. Correlation of other variables was also checked and the significant ones were included in the model.

Two dummy variables house_age and rebuilt_age were created to check the age of the houses and the rebuilding age of the houses, non-surprisingly they came to be highly correlated with Sale Price and were a driving factor in the model.

```
housing.data$house_age<- 2021-housing.data$YearBuilt
housing.data$house_age

housing.data$rebuilt_age<- 2021-housing.data$YearRemodAdd
```

A dummy variable showcasing extra features of house was created to run in the model and check if these extra features have an added-value to the Sale Prices.

```
housing.data$extras<- ifelse(housing.data$WoodDeckSF>0 | housing.data$OpenPorchSF>0 | housing.data$EnclosedPorch>0 | housing.data$ScreenPorch>0,1,0 )
```

Other codes related to imputation & correlations can be viewed in the notebook.

Train & Test datasets were created out of the housing dataset to perform modeling and predicting prices.

Regularization - Lasso Model

There was a total of 81 variables to be used for the model, even though all of them were not crucial to our model, but a lot of them were and using the simple 'lm' model for the prediction of Housing Prices would over-fit the model with this many variables. Hence, Lasso was the best suited regularization method that could be used in devising the model as it shrinks the co-efficient(s) of so many variables which avoided overfitting.

Since we concluded above for multiple variables that would be crucial to our model there existed a lot of outliers. To an extent, the Lasso model also takes care of the outliers.

So, after downloading the 'glmnet' library a model matrix and prediction variable were created by binding the IDs.

As discussed above log of Sale Price was added to our y object because of the visible heteroskedasticity in the distribution of Sale Price data.

Choosing the variables for the final model was hard, but I decided to go through a bit of research on the internet first and also use some of my practical knowledge. So, research showed that Housing Prices were most affected by Location, House-size, Age & Condition of the house, number of buyers in the market and the various features. So, I used this information to build the model and Add/Delete variables from the model. As I kept checking more on my Tableau plots which categorical variables affected my main integer variables and changed their slope vs Sale Price, I kept adding them to my model to improve accuracy.

Since we concluded above for multiple variables that would be crucial to our model there exists outliers and heteroskedasticity.

```
Id ~ house_age * ExterQual * rebuilt_age*ExterCond +
Neighborhood*BldgType + log(GrLivArea)*OverallQual +
KitchenAbvGr*KitchenQual +
Fireplaces*FireplaceQu + sqrt(TotalBsmtSF)*BsmtQual*BsmtExposure +
FullBath*CentralAir +log(X1stFlrSF)*HouseStyle +
OverallQual*OverallCond + sqrt(MasVnrArea)*ExterQual +
GarageArea*GarageCars*GarageType*GarageQual + TotRmsAbvGrd+
log(LotArea)*LotFrontage+ SaleType + SaleCondition + Functional + Foundation
```

^These were the interactions and variables included in the model matrix. House age and rebuilt age had different effects to different exterior qualities and conditions of the houses so they were interacted accordingly. All other variables that showed high correlation with Sale Price were also included.

Log function was also applied to GrLivArea, First Floor SF, and LotArea as they showed heteroskedasticity in their scatterplots too. Square root function was applied to TotalBsmtSF and MasVnrArea to make them normally distributed too.

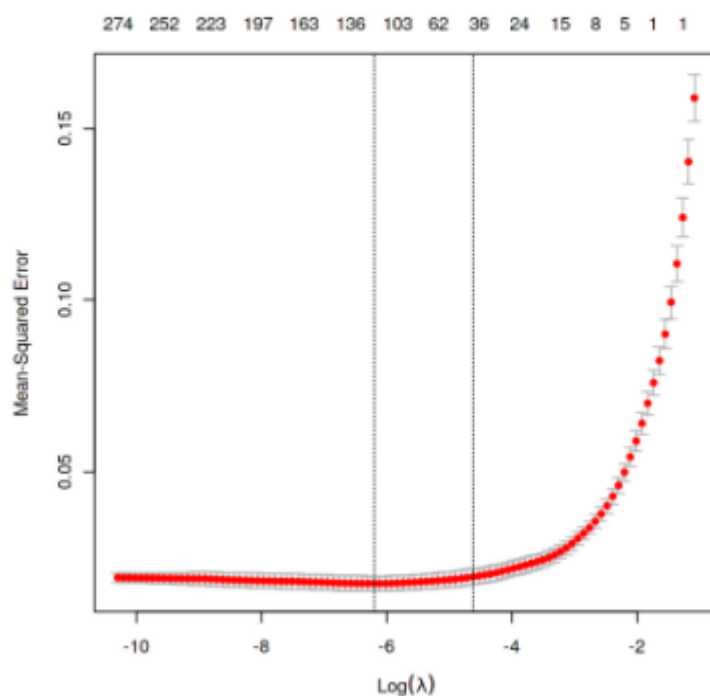
Kitchen, Fireplace, Basement and Garage had different intercepts on the y-axis of the linear model against Sale Price when interacted with their qualitative variables, so they were included accordingly.

SaleType, SaleCondition and foundation were significant drivers of the Sale Price on their own too as discovered in the model.

Since neighbourhood is an important part of house prices, the building type has its say how much the price varies in different neighbourhoods.

Finally, cross-validation was performed using Lasso to find the minimum lamda for our predictions that would give us the lowest error rate out of different combinations of the model.

```
crossval <- cv.glmnet(x = X.training, y = y, alpha = 1)
plot(crossval)
```



The minimum lamda was -6.1984

```
penalty.lasso <- crossval$lambda.min
log(penalty.lasso)
```

-6.19843165580263

Finally, Sale Prices were predicted using the above model which gave us the Top 28% (near 25%?) position in the Kaggle competition.

One con that was learned out of the Lasso model was that the features it selects is very unpredictable and it completely ignores some non-significant variables that could drive the price a bit.

The file named Predicted Sale Price is the predictions and the code is available on the Notebook link shared.