

Housing Price Prediction

**A Project Report by:**

**Archit Khanduja**

# Acknowledgement

The project would not have been built without the constant support from **DataTrained** and **Fliprobo** teams.

Following are the research papers, discussions and articles that helped me in completing the project:

<https://stats.stackexchange.com/questions/251708/when-to-use-ridge-regression-and-lasso-regression-what-can-be-achieved-while-us>

<https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.get_dummies.html>

<https://www.tutorialspoint.com/change-data-type-for-one-or-more-columns-in-pandas-dataframe-1>

<https://stackoverflow.com/questions/8420143/valueerror-could-not-convert-string-to-float-id>

# Introduction

## Business problem Framing

In this project, the main problem revolves around a US-based housing company named **Surprise Housing** which has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia.

# Conceptual Background of the Domain Problem

Surprise Housing is looking at prospective properties to buy houses to enter the market. We are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not.

The model is to be built with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

# Review of Literature

While researching more about the problem, it was figured out that in order to build a model to predict the actual value of the prospective properties, it would be more important to know what the data is telling us by visualizing it and knowing the correlations between various features. Also, it would be important to know which features will play the most important role in predicting the value of the properties.

Another thing to notice here is that the number of features in the dataset is very high and we need to decide upon the relevance of some of these features in model building. If the features are not found to be relevant, they have to be dropped.

# Motivation for the Problem Undertaken

Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world’s economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases.

Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company which needs to accurately predict the price of houses.

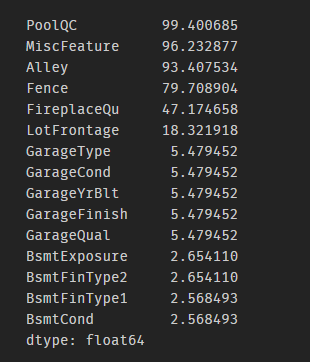
# Analytical Problem Framing

## Mathematical/ Analytical Modelling of the Problem

In this problem, we need to predict the price of properties. Since the target variable is a continuous variable, we have used Regression for our Analysis. The dataset has 1460 entries and 81 features with a mix of categorical and numerical features.

### Presence of Missing Values

There are a few features which are having missing values in them which need to be treated. We checked the percentage of missing values in the dataset for each feature and found the following result:



Given the nature of each feature, we have treated missing values for each feature on a case by case basis.

### Multicollinearity

In our dataset, we have 81 features. These features are important for model building. For a regression model to be successful, all the features should be independent of each other. If any 2 features have a correlation, there would be multicollinearity in our regression model and it would be unstable. The unstable nature of the model may cause **overfitting**. If we apply the model to another sample of data, the accuracy will drop significantly compared to the accuracy of our training dataset.

## Data Sources and their Formats

MSSubClass: Identifies the type of dwelling involved in the sale.

20 1-STORY 1946 & NEWER ALL STYLES

30 1-STORY 1945 & OLDER

40 1-STORY W/FINISHED ATTIC ALL AGES

45 1-1/2 STORY - UNFINISHED ALL AGES

50 1-1/2 STORY FINISHED ALL AGES

60 2-STORY 1946 & NEWER

70 2-STORY 1945 & OLDER

75 2-1/2 STORY ALL AGES

80 SPLIT OR MULTI-LEVEL

85 SPLIT FOYER

90 DUPLEX - ALL STYLES AND AGES

120 1-STORY PUD (Planned Unit Development) - 1946 & NEWER

150 1-1/2 STORY PUD - ALL AGES

160 2-STORY PUD - 1946 & NEWER

180 PUD - MULTILEVEL - INCL SPLIT LEV/FOYER

190 2 FAMILY CONVERSION - ALL STYLES AND AGES

MSZoning: Identifies the general zoning classification of the sale.

A Agriculture

C Commercial

FV Floating Village Residential

I Industrial

RH Residential High Density

RL Residential Low Density

RP Residential Low Density Park

RM Residential Medium Density

LotFrontage: Linear feet of street connected to property

LotArea: Lot size in square feet

Street: Type of road access to property

Grvl Gravel

Pave Paved

Alley: Type of alley access to property

Grvl Gravel

Pave Paved

NA No alley access

LotShape: General shape of property

Reg Regular

IR1 Slightly irregular

IR2 Moderately Irregular

IR3 Irregular

LandContour: Flatness of the property

Lvl Near Flat/Level

Bnk Banked - Quick and significant rise from street grade to building

HLS Hillside - Significant slope from side to side

Low Depression

Utilities: Type of utilities available

AllPub All public Utilities (E,G,W,& S)

NoSewr Electricity, Gas, and Water (Septic Tank)

NoSeWa Electricity and Gas Only

ELO Electricity only

LotConfig: Lot configuration

Inside Inside lot

Corner Corner lot

CulDSac Cul-de-sac

FR2 Frontage on 2 sides of property

FR3 Frontage on 3 sides of property

LandSlope: Slope of property

Gtl Gentle slope

Mod Moderate Slope

Sev Severe Slope

Neighborhood: Physical locations within Ames city limits

Blmngtn Bloomington Heights

Blueste Bluestem

BrDale Briardale

BrkSide Brookside

ClearCr Clear Creek

CollgCr College Creek

Crawfor Crawford

Edwards Edwards

Gilbert Gilbert

IDOTRR Iowa DOT and Rail Road

MeadowV Meadow Village

Mitchel Mitchell

Names North Ames

NoRidge Northridge

NPkVill Northpark Villa

NridgHt Northridge Heights

NWAmes Northwest Ames

OldTown Old Town

SWISU South & West of Iowa State University

Sawyer Sawyer

SawyerW Sawyer West

Somerst Somerset

StoneBr Stone Brook

Timber Timberland

Veenker Veenker

Condition1: Proximity to various conditions

Artery Adjacent to arterial street

Feedr Adjacent to feeder street

Norm Normal

RRNn Within 200' of North-South Railroad

RRAn Adjacent to North-South Railroad

PosN Near positive off-site feature--park, greenbelt, etc.

PosA Adjacent to postive off-site feature

RRNe Within 200' of East-West Railroad

RRAe Adjacent to East-West Railroad

Condition2: Proximity to various conditions (if more than one is present)

Artery Adjacent to arterial street

Feedr Adjacent to feeder street

Norm Normal

RRNn Within 200' of North-South Railroad

RRAn Adjacent to North-South Railroad

PosN Near positive off-site feature--park, greenbelt, etc.

PosA Adjacent to postive off-site feature

RRNe Within 200' of East-West Railroad

RRAe Adjacent to East-West Railroad

BldgType: Type of dwelling

1Fam Single-family Detached

2FmCon Two-family Conversion; originally built as one-family dwelling

Duplx Duplex

TwnhsE Townhouse End Unit

TwnhsI Townhouse Inside Unit

HouseStyle: Style of dwelling

1Story One story

1.5Fin One and one-half story: 2nd level finished

1.5Unf One and one-half story: 2nd level unfinished

2Story Two story

2.5Fin Two and one-half story: 2nd level finished

2.5Unf Two and one-half story: 2nd level unfinished

SFoyer Split Foyer

SLvl Split Level

OverallQual: Rates the overall material and finish of the house

10 Very Excellent

9 Excellent

8 Very Good

7 Good

6 Above Average

5 Average

4 Below Average

3 Fair

2 Poor

1 Very Poor

OverallCond: Rates the overall condition of the house

10 Very Excellent

9 Excellent

8 Very Good

7 Good

6 Above Average

5 Average

4 Below Average

3 Fair

2 Poor

1 Very Poor

YearBuilt: Original construction date

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)

RoofStyle: Type of roof

Flat Flat

Gable Gable

Gambrel Gabrel (Barn)

Hip Hip

Mansard Mansard

Shed Shed

RoofMatl: Roof material

ClyTile Clay or Tile

CompShg Standard (Composite) Shingle

Membran Membrane

Metal Metal

Roll Roll

Tar&Grv Gravel & Tar

WdShake Wood Shakes

WdShngl Wood Shingles

Exterior1st: Exterior covering on house

AsbShng Asbestos Shingles

AsphShn Asphalt Shingles

BrkComm Brick Common

BrkFace Brick Face

CBlock Cinder Block

CemntBd Cement Board

HdBoard Hard Board

ImStucc Imitation Stucco

MetalSd Metal Siding

Other Other

Plywood Plywood

PreCast PreCast

Stone Stone

Stucco Stucco

VinylSd Vinyl Siding

Wd Sdng Wood Siding

WdShing Wood Shingles

Exterior2nd: Exterior covering on house (if more than one material)

AsbShng Asbestos Shingles

AsphShn Asphalt Shingles

BrkComm Brick Common

BrkFace Brick Face

CBlock Cinder Block

CemntBd Cement Board

HdBoard Hard Board

ImStucc Imitation Stucco

MetalSd Metal Siding

Other Other

Plywood Plywood

PreCast PreCast

Stone Stone

Stucco Stucco

VinylSd Vinyl Siding

Wd Sdng Wood Siding

WdShing Wood Shingles

MasVnrType: Masonry veneer type

BrkCmn Brick Common

BrkFace Brick Face

CBlock Cinder Block

None None

Stone Stone

MasVnrArea: Masonry veneer area in square feet

ExterQual: Evaluates the quality of the material on the exterior

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

ExterCond: Evaluates the present condition of the material on the exterior

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

Foundation: Type of foundation

BrkTil Brick & Tile

CBlock Cinder Block

PConc Poured Contrete

Slab Slab

Stone Stone

Wood Wood

BsmtQual: Evaluates the height of the basement

Ex Excellent (100+ inches)

Gd Good (90-99 inches)

TA Typical (80-89 inches)

Fa Fair (70-79 inches)

Po Poor (<70 inches

NA No Basement

BsmtCond: Evaluates the general condition of the basement

Ex Excellent

Gd Good

TA Typical - slight dampness allowed

Fa Fair - dampness or some cracking or settling

Po Poor - Severe cracking, settling, or wetness

NA No Basement

BsmtExposure: Refers to walkout or garden level walls

Gd Good Exposure

Av Average Exposure (split levels or foyers typically score average or above)

Mn Mimimum Exposure

No No Exposure

NA No Basement

BsmtFinType1: Rating of basement finished area

GLQ Good Living Quarters

ALQ Average Living Quarters

BLQ Below Average Living Quarters

Rec Average Rec Room

LwQ Low Quality

Unf Unfinshed

NA No Basement

BsmtFinSF1: Type 1 finished square feet

BsmtFinType2: Rating of basement finished area (if multiple types)

GLQ Good Living Quarters

ALQ Average Living Quarters

BLQ Below Average Living Quarters

Rec Average Rec Room

LwQ Low Quality

Unf Unfinshed

NA No Basement

BsmtFinSF2: Type 2 finished square feet

BsmtUnfSF: Unfinished square feet of basement area

TotalBsmtSF: Total square feet of basement area

Heating: Type of heating

Floor Floor Furnace

GasA Gas forced warm air furnace

GasW Gas hot water or steam heat

Grav Gravity furnace

OthW Hot water or steam heat other than gas

Wall Wall furnace

HeatingQC: Heating quality and condition

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

CentralAir: Central air conditioning

N No

Y Yes

Electrical: Electrical system

SBrkr Standard Circuit Breakers & Romex

FuseA Fuse Box over 60 AMP and all Romex wiring (Average)

FuseF 60 AMP Fuse Box and mostly Romex wiring (Fair)

FuseP 60 AMP Fuse Box and mostly knob & tube wiring (poor)

Mix Mixed

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: Bedrooms above grade (does NOT include basement bedrooms)

Kitchen: Kitchens above grade

KitchenQual: Kitchen quality

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

Functional: Home functionality (Assume typical unless deductions are warranted)

Typ Typical Functionality

Min1 Minor Deductions 1

Min2 Minor Deductions 2

Mod Moderate Deductions

Maj1 Major Deductions 1

Maj2 Major Deductions 2

Sev Severely Damaged

Sal Salvage only

Fireplaces: Number of fireplaces

FireplaceQu: Fireplace quality

Ex Excellent - Exceptional Masonry Fireplace

Gd Good - Masonry Fireplace in main level

TA Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement

Fa Fair - Prefabricated Fireplace in basement

Po Poor - Ben Franklin Stove

NA No Fireplace

GarageType: Garage location

2Types More than one type of garage

Attchd Attached to home

Basment Basement Garage

BuiltIn Built-In (Garage part of house - typically has room above garage)

CarPort Car Port

Detchd Detached from home

NA No Garage

GarageYrBlt: Year garage was built

GarageFinish: Interior finish of the garage

Fin Finished

RFn Rough Finished

Unf Unfinished

NA No Garage

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

GarageQual: Garage quality

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

NA No Garage

GarageCond: Garage condition

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

NA No Garage

PavedDrive: Paved driveway

Y Paved

P Partial Pavement

N Dirt/Gravel

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

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

NA No Pool

Fence: Fence quality

GdPrv Good Privacy

MnPrv Minimum Privacy

GdWo Good Wood

MnWw Minimum Wood/Wire

NA No Fence

MiscFeature: Miscellaneous feature not covered in other categories

Elev Elevator

Gar2 2nd Garage (if not described in garage section)

Othr Other

Shed Shed (over 100 SF)

TenC Tennis Court

NA None

MiscVal: $Value of miscellaneous feature

MoSold: Month Sold (MM)

YrSold: Year Sold (YYYY)

SaleType: Type of sale

WD Warranty Deed - Conventional

CWD Warranty Deed - Cash

VWD Warranty Deed - VA Loan

New Home just constructed and sold

COD Court Officer Deed/Estate

Con Contract 15% Down payment regular terms

ConLw Contract Low Down payment and low interest

ConLI Contract Low Interest

ConLD Contract Low Down

Oth Other

SaleCondition: Condition of sale

Normal Normal Sale

Abnorml Abnormal Sale - trade, foreclosure, short sale

AdjLand Adjoining Land Purchase

Alloca Allocation - two linked properties with separate deeds, typically condo with a garage unit

Family Sale between family members

Partial Home was not completed when last assessed (associated with New Homes)

## Data Preprocessing Done

* A few columns were found irrelevant after going through the data dictionary. Those irrelevant columns were not important for our model building and were removed.
* It was found that a few variables, including the target variable were skewed. To handle the skewness in the target variable, i.e. SalePrice, log transformation was used to make the distribution normal.

In a few skewed categorical variables, some categories had low frequencies. Those categories were clubbed into one ‘others’ category to counter the skewness.

There were a few highly skewed columns which were dropped.

* A new feature ‘YearSinceRemodel’ was derived from ‘YearRemodAdd’ and ‘YearBuilt’. The latter 2 features were dropped as they played no further role in the model.
* Dummy variables were formed from the categorical features.
* Scaling was done for the Numerical features
* All the pre-processing tasks were repeated for the test dataset as well.
* Finally, the data was split into X and Y wherein Y is the target variable and X are the feature variables.

## Data Inputs- Logic- Output Relationships

Here, we will see the univariate and bivariate analysis done during our study of the data

We plotted the target variable SalesPrice on a Distplot and found it to be right skewed.

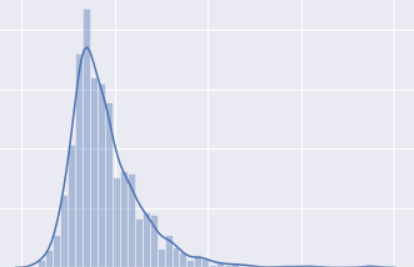


Figure 1 SalesPrice distplot

After the log transformation of sales price, it was normally distributed and this is how it looked like:

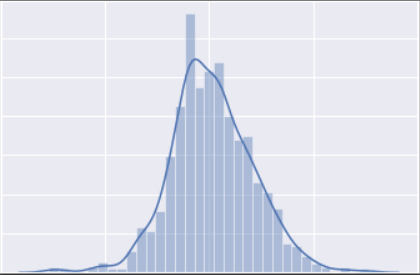


Figure 2 SalesPrice after log transformation

Following are a few countplots associated with Garage of the house. Using these countplots, we replaced missing values with corresponding modes.

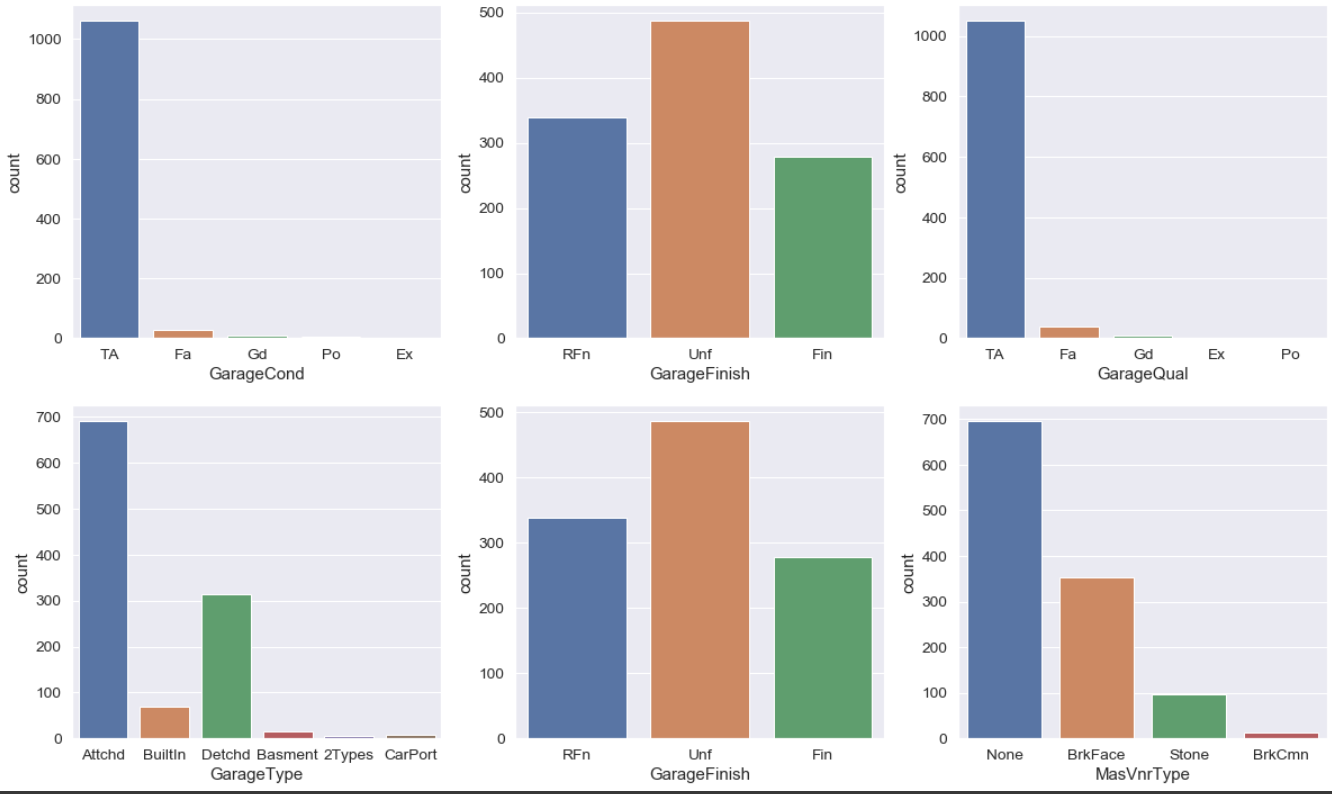


Figure 3 Garage Countplots

Following are a few countplots associated with Basement of the house. Using these countplots, we replaced missing values with corresponding modes.

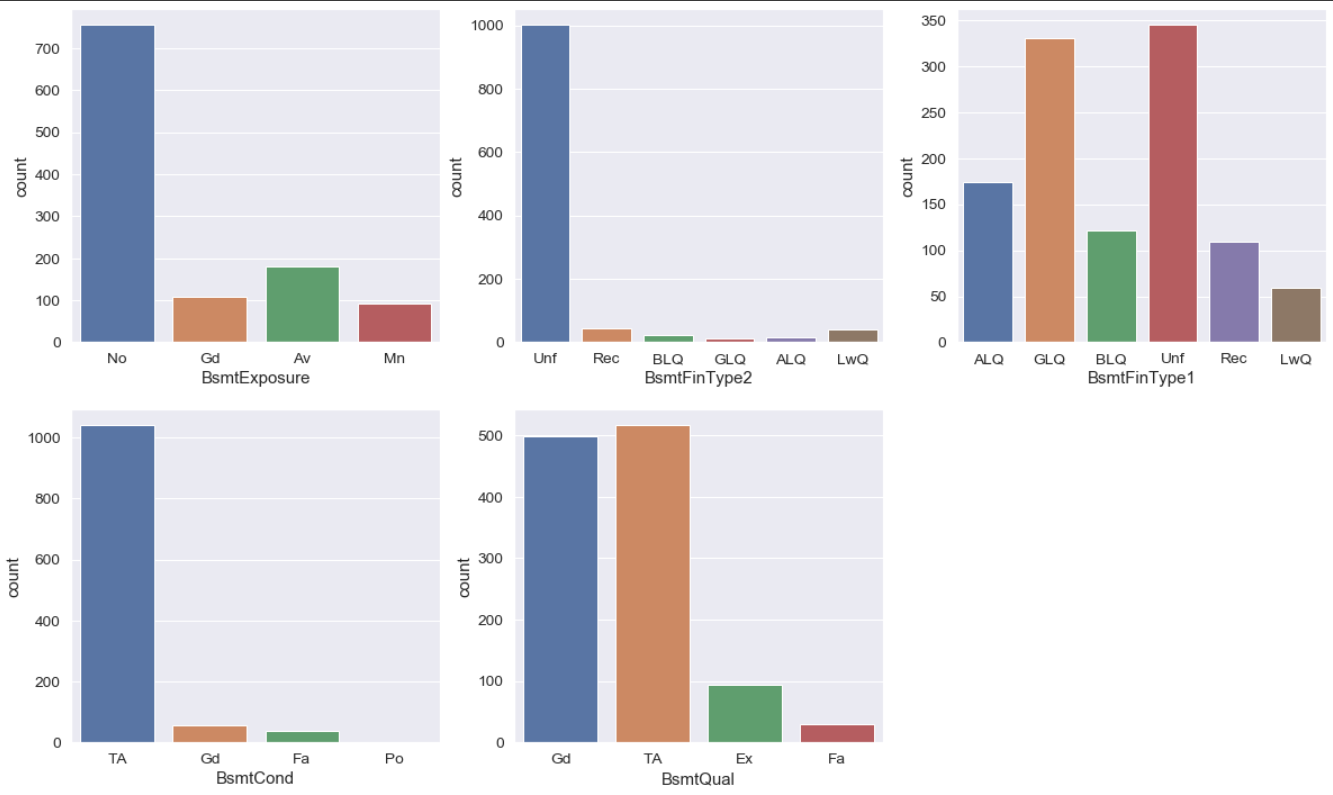
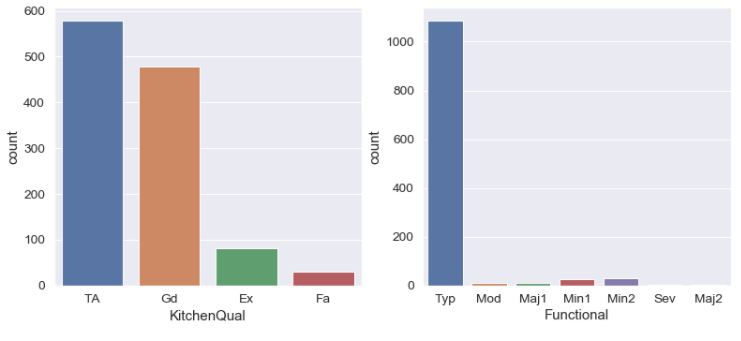


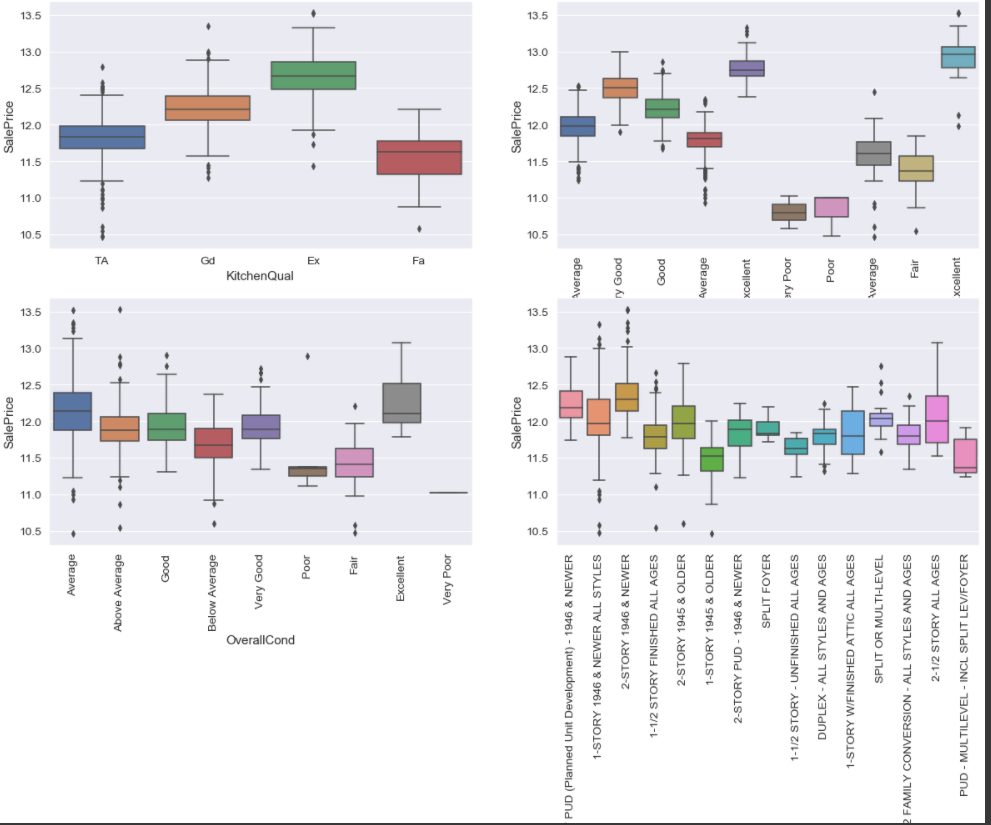
Figure 4 Basement Countplots

Now, we have visualized more countplots associated with other features to check the skewness as shown below:

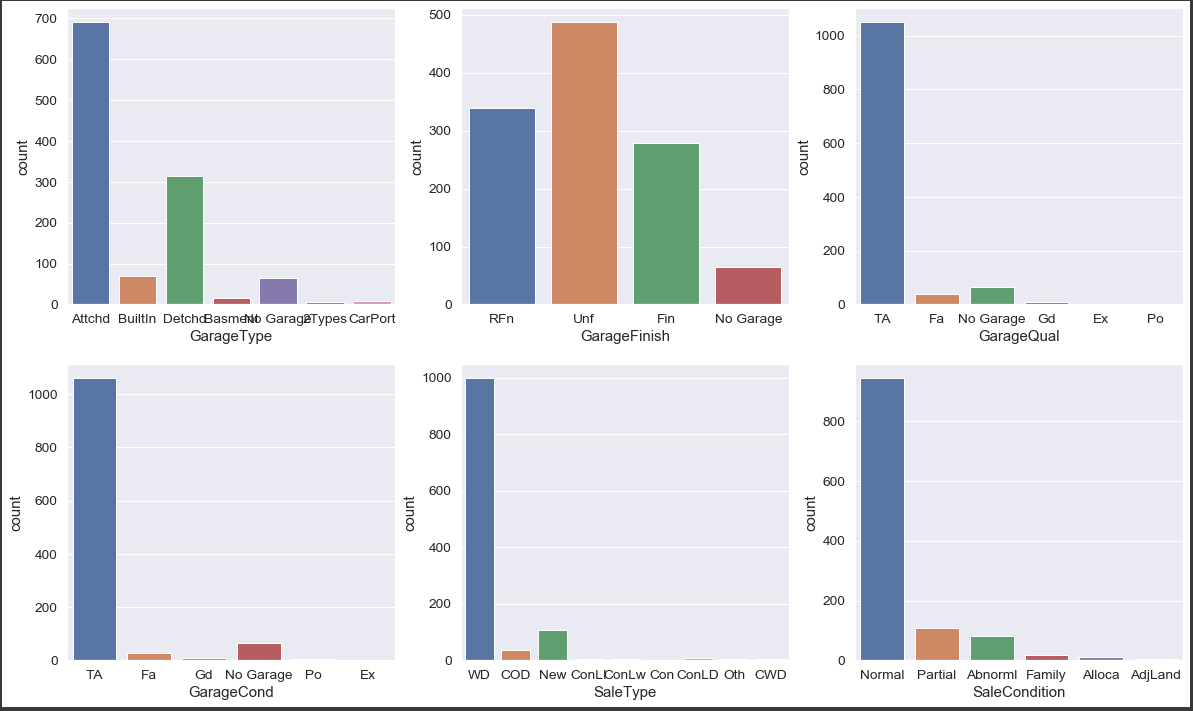


Functional is highly skewed and we have decided to drop it.

Now, we have plotted some box plots to check the skewness



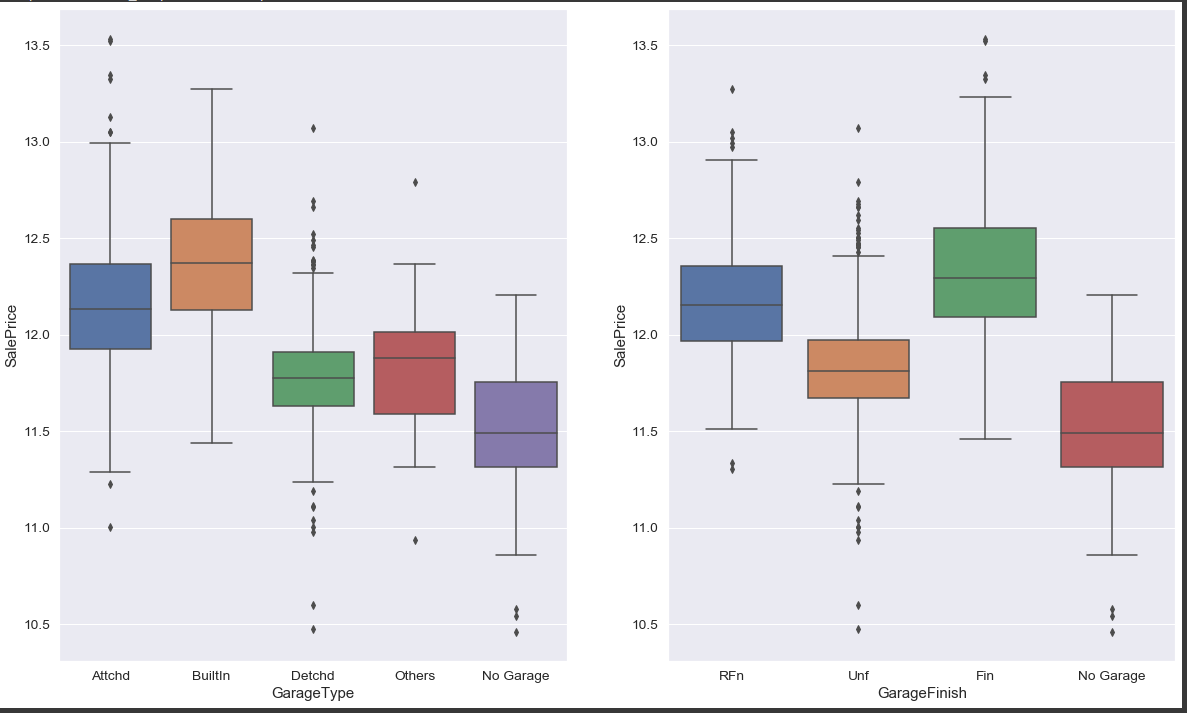
Also, a few countplots were also plotted:



After plotting the above plots, we dropped GarageQual, GarageCond, SaleType columns as they were highly skewed.

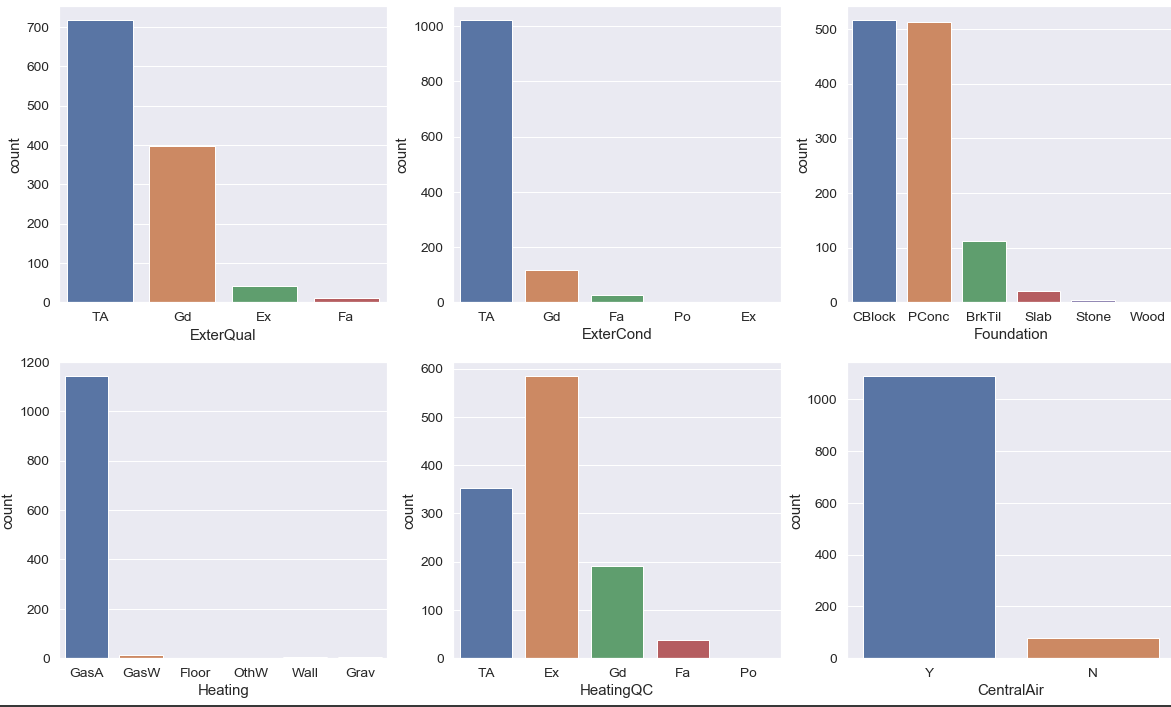
Also, in GarageType variable, 'Basment','CarPort','2Types' categories were clubbed into ‘others’ category to counter the skewness. The same type of clubbing was done in SaleCondition.

Now, let’s see the effect of GarageType and GarageFinish on SalesPrice.

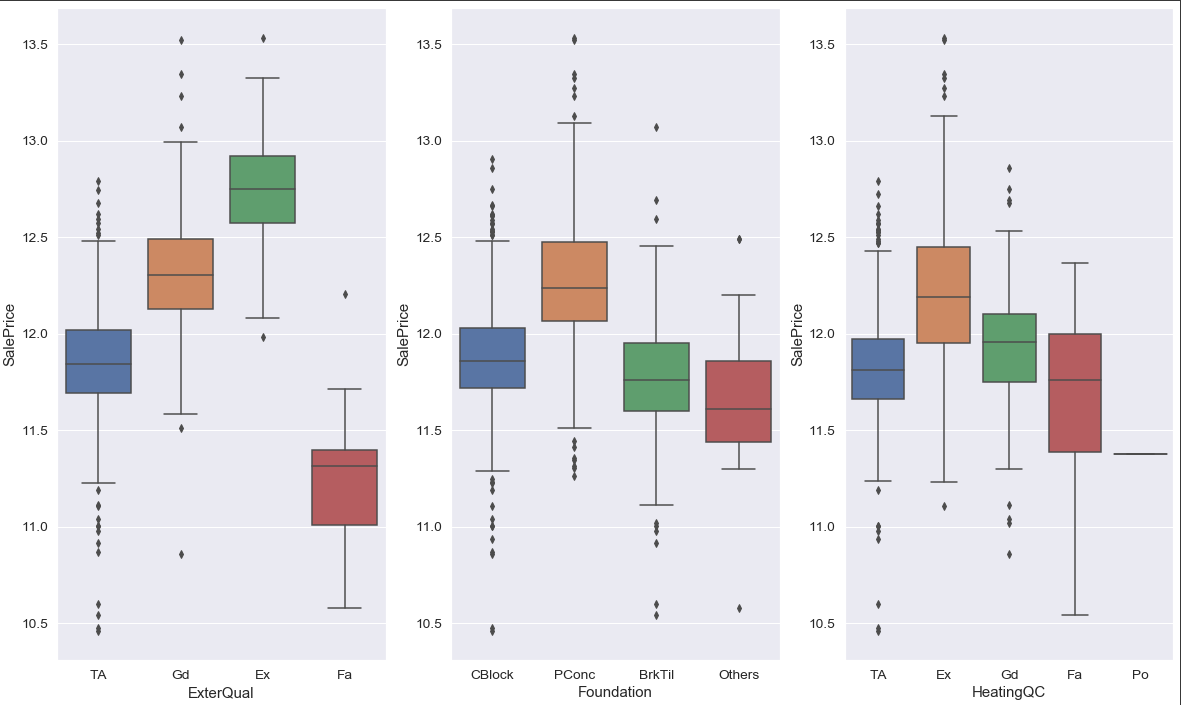


Price of Builtin Garagetype and Finished garage is the highest.

A few more countplots are given below:



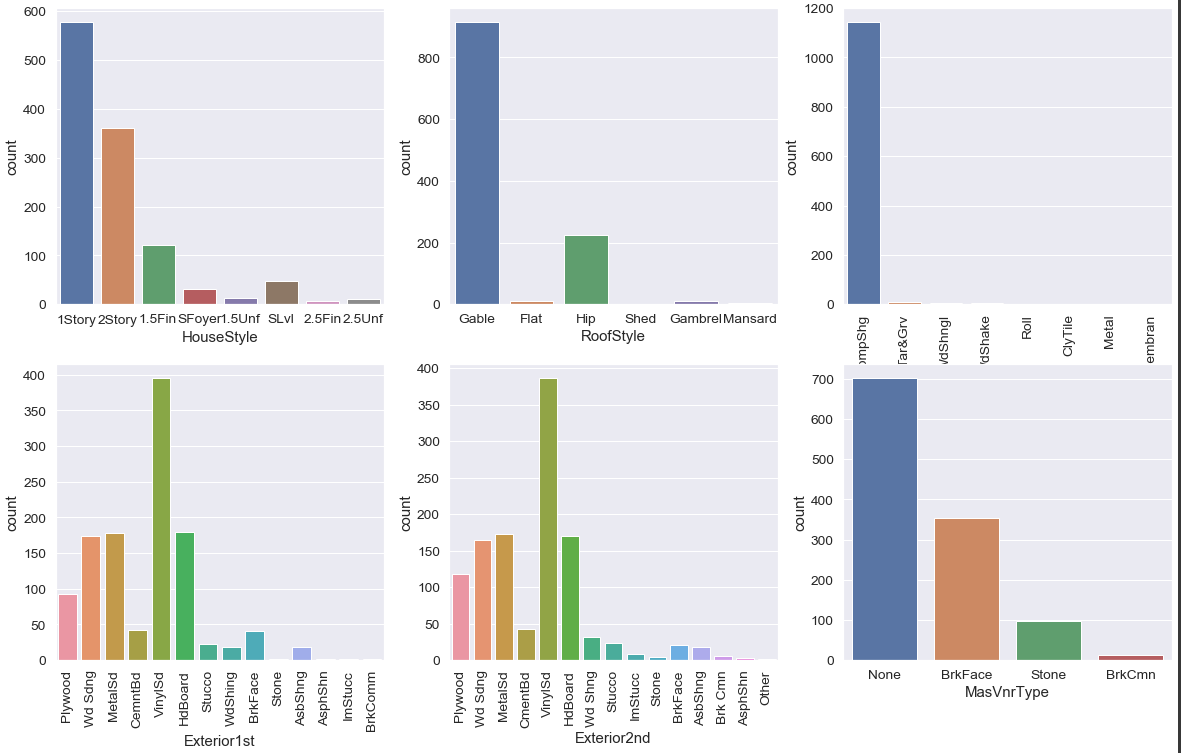
* Majority of ExterQual, ExerCond is TA
* Poured Contrete foundation are the highest in number
* Meanwhile variables like Heating , Central Airand Exter Cond are skewed so would be dropping these variables

Let's see effect of Garage type and GarageFinish on SalePrice

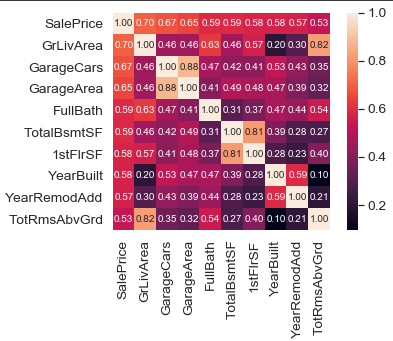
- Price of Excellent ExterQual and HeatingQc is highest

- Price of Poured Contrete Foundation is highest.

Following countplots were plotted to check the skewness:



To counter the skewness in these columns, categories with low frequencies were clubbed into ‘others’ category.

Following is the correlation matrix of sales price:

## Assumptions

1. Attributes are independent of each other (low or no multicollinearity).
2. Slope coefficients of linear regression are always raised to 1.
3. No auto-correlation is there between errors.
4. The error in prediction of each trial is independent of value of ‘X’
5. The ‘X’ should have variance, outliers should not exist.
6. Homoscedasticity: The variance of residual is the same for any value of X.
7. Linearity: The relationship between X and the mean of Y is linear.
8. Normality: For any fixed value of X, Y is normally distributed.

## Hardware and Software Requirements and Tools Used

For the building of this model, an MSI Machine with Intel CORE i7 7th Gen processor and an 8 GB RAM was used.

The programming language used was Python. The compiler that was used was Anaconda Navigator. The programs were run in Jupyter Notebook environments.

The libraries used were as follows: numpy, pandas, matplotlib, seaborn, sklearn, scipy, imblearn.

# Model/s Development and Evaluation

## Identification of possible problem-solving approaches

We have a target variable that is continuous in nature. This arrangement calls for a regression model.

The number of features that we have in our dataset is 81. In case of such large number of features, there is a high chance that there will be multicollinearity amongst the features of our dataset. Multicollinearity may result in overfitting and the model might fail in case another sample is introduced.

To avoid this, we have used the Lasso and Ridge regression and have finalized the model based upon the best R2 Score.

While doing the Lasso and Ridge Regression, we need to select the appropriate value of alpha. Lesser the value of alpha, closer is the model to Linear Regression. Larger the value of alpha, more feature coefficients are 0.

In order to determine the best parameters of alpha, we used best\_params\_ function for both Lasso and Ridge regression.

## Testing of Identified Approaches

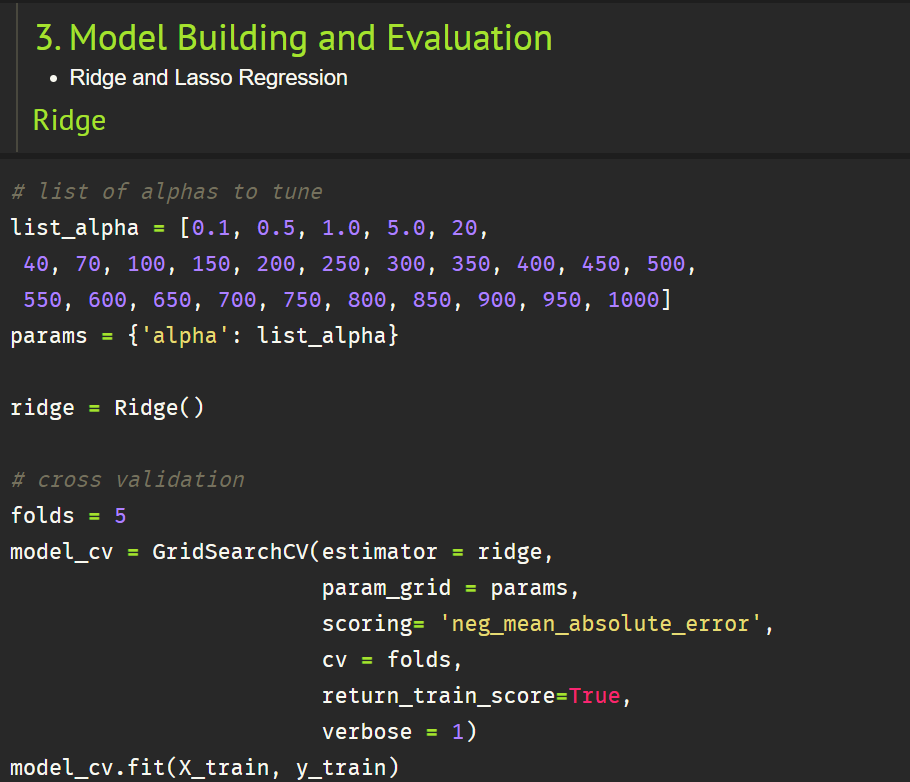
The final model was created and tested on the test dataset by using the predict function.

## Run and Evaluate selected models

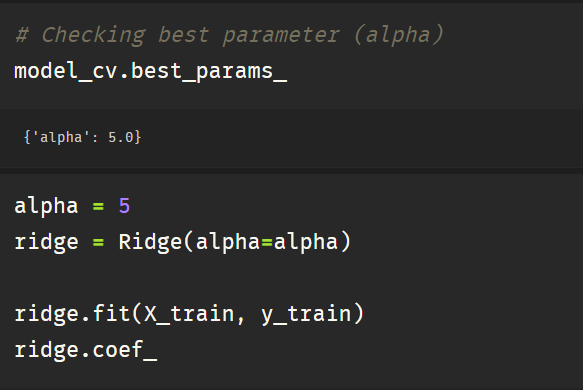
### Ridge Regression

The first model was built on Ridge Regression. As a first step to it, we tuned the parameters using Gridsearch CV. Negative mean absolute error was kept as the scoring parameter. It was put under a 5 folds cross validation and was tested on various values of alpha.

Training and testing sets obtained after the train test split were passed as parameters to the fit function for the CV model to be built.



After this, best parameter for the model was obtained and the alpha value associated with it was found to be **5**. The ridge score was then calculated on this alpha value and was found to be **0.9077**. The same on the test set was found to be **0.8531**.



Our next goal was to find the best parameters that define the price of the property. Following were the parameters which were found to be the best along with their coefficient:

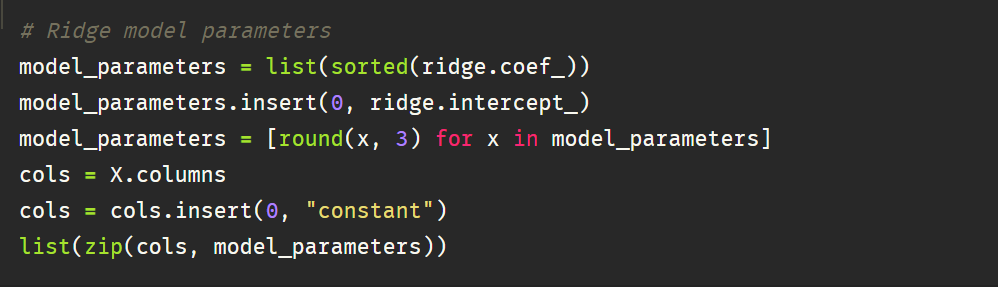
('GarageFinish\_RFn', 0.105),

('GarageFinish\_Unf', 0.107),

('SaleCondition\_Normal', 0.119),

('SaleCondition\_Others', 0.131),

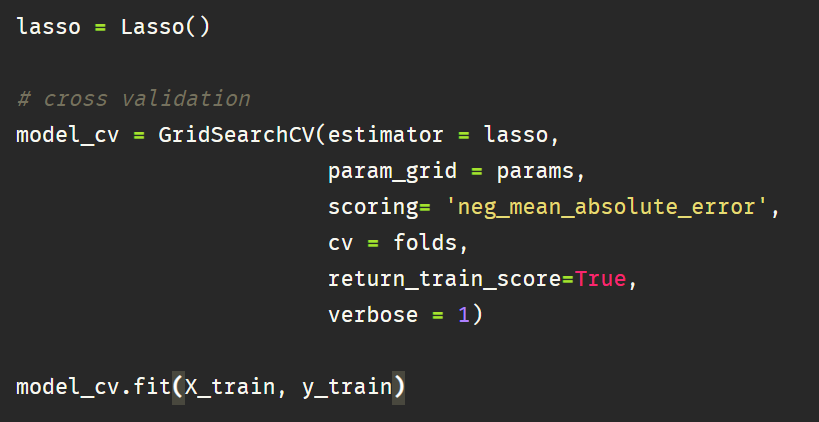
('SaleCondition\_Partial', 0.17)



### Lasso Regression

After making a Ridge Regression model, we moved on to the Lasso Regression. As a first step to it, we tuned the parameters using Gridsearch CV. Negative mean absolute error was kept as the scoring parameter. It was put under a 5 folds cross validation and was tested on various values of alpha.

Training and testing sets obtained after the train test split were passed as parameters to the fit function for the CV model to be built.



After this, best parameter for the model was obtained and the alpha value associated with it was found to be **0.1**. The ridge score was then calculated on this alpha value and was found to be 0.7566. The same on the test set was found to be 0.71.

The score obtained with Lasso was much lesser than the score obtained with Ridge. Hence, we proceeded with Ridge for our final model.

# Key Metrics for success in solving problem under consideration

The key metric used in solving the problem were applied in the same code of model building. **Negative Mean Squared Error** was used as the major key metric.

**Error** in this case means the difference between the observed values y1, y2, y3, … and the predicted ones pred(y1), pred(y2), pred(y3), … We square each difference (pred(yn) – yn)) \*\* 2 so that negative and positive values do not cancel each other out. The error is then multiplied by -1 for the convenience of our generalized convention of, the higher the better.

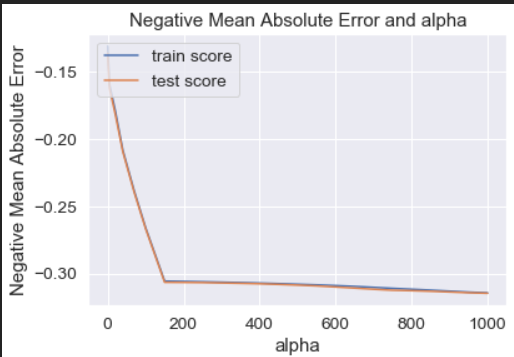
## Visualizations during Model Building

We plotted various alpha values on the X axis and Negative Mean Absolute error on Y axis for both Test and Train datasets. The following are the 2 plots assocated with Ridge and Lasso Regression respectively

Ridge Regression



Lasso Regression



# Conclusion

## Key Findings and Conclusions of the Study

After finding the best parameters of Ridge Regression, following are the major key findings:

1. SaleCondition\_Partial is the most important feature in defining the price of the property. This signifies that if the property is new, the price will be highly impacted.
2. If the garage is unfinished or rough finished, this is going to have a big impact on the price of the property.
3. Above ground living area is another feature that plays an important role in defining the SalesPrice. Clearly, more the area, more the price.
4. Features that are highly correlated can be dropped as they will result in collinearity and affect the model accuracy.

## Learning Outcomes of the Study in respect of Data Science

There were various learning outcomes out of this project.

1. Handling of high number of features

Since the number of features in our dataset is very high, this resulted in multicollinearity and it is highly probable to have this situation. We cannot use sinple linear regression and will have to use Regularisation techniques.

1. Handling Skewed Data

Skewness can be a problem when the distribution is not normal. The skewness was handled by clubbing 3-4 categories in a single one to balance the dataset.

## Limitations of this work and Scope for Future Work

* While doing the encoding, we created dummy variables for categorical features so that our model can interpret it from numbers. This made all the categories in a particular feature equally important for the model. But, there were a few features where one category was weighted more than the other and was ordinal. We can identify those columns and do the label encoding where the model can rank the numbers in a single feature and better results can be obtained.