HOUSING PRICE PREDICTION

Submitted by-

AKASHDEEP SINGH MANRAL

BATCH-1834

ACKNOWLEDGEMENT

A US-based housing company named Surprise Housing 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.

I want to thank my intern mentor miss- Swati Mahaseth for providing assistance in solving my queries, with her help and guidance I was able to complete my project successfully

INTRODUCTION

1-Problem Statement-

1-The company is looking at prospective properties to buy houses to enter the market. You 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. For this company wants to know:

- Which variables are important to predict the price of variable?
- How do these variables describe the price of the house?

2-You are required to model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables.

- You need to find important features which affect the price positively or negatively.
- Two datasets are being provided to you (test.csv, train.csv). You will train on train.csv dataset and predict on test.csv file

Analytical Problem Framing

1-Exploratory Data Analysis

A- Categorical Columns-

1- MSZoning: Identifies the general zoning classification of the sale.

A Agriculture

C Commercial

FV Floating Village Residential

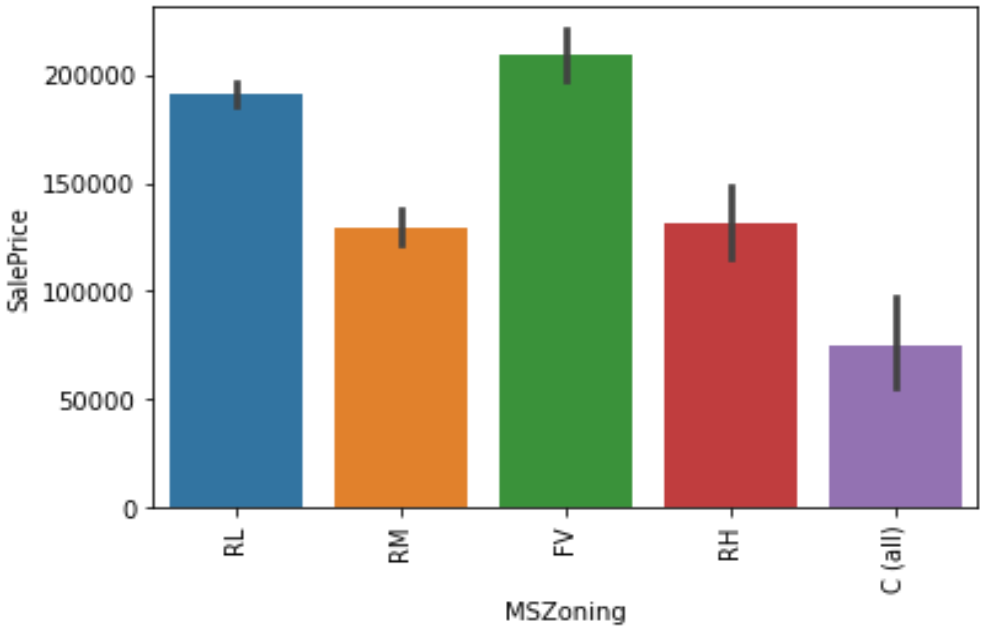
I Industrial

RH Residential High Density

RL Residential Low Density

RP Residential Low Density Park

RM Residential Medium Density



Obs-

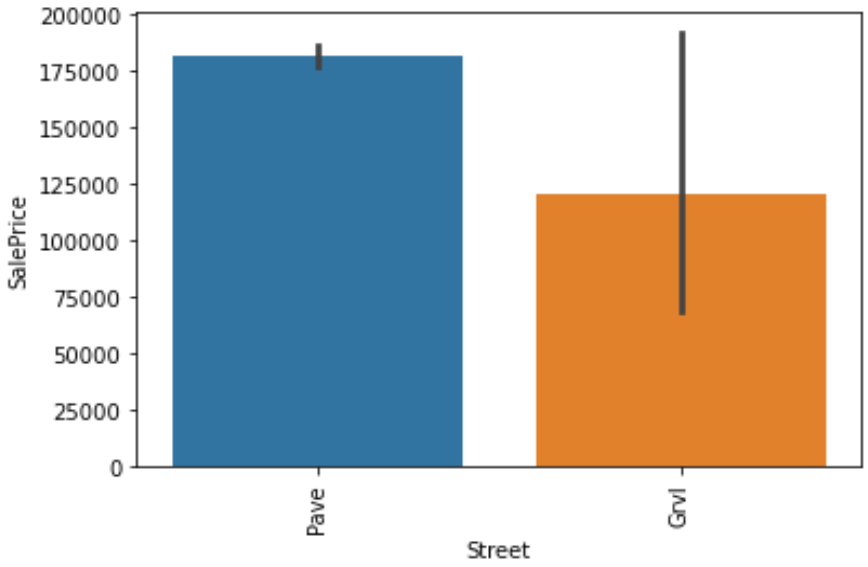
1-Floating Village residential has the highest sale price followed by Residential Low density and Residential High density.

2-C(all) have less sale price as compared to others.

2- Street: Type of road access to property

Grvl Gravel

Pave Paved



Obs-

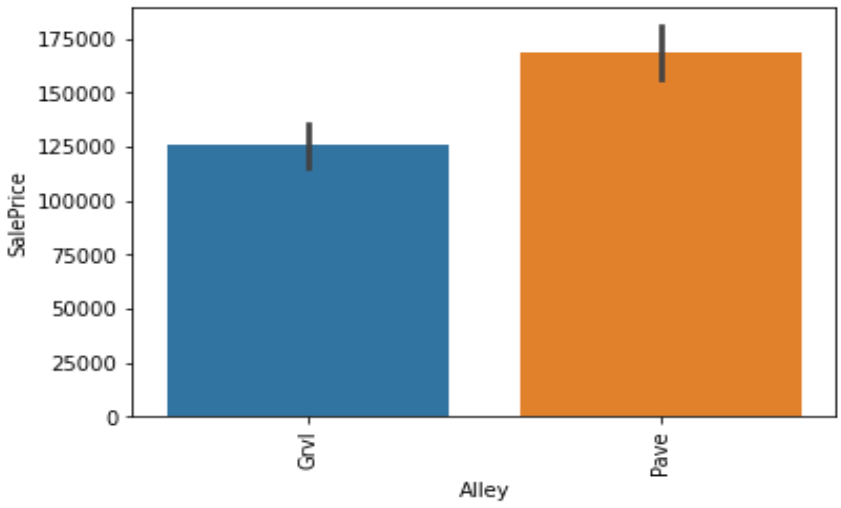
1-Paved road has higher sale prices as compared to gravel

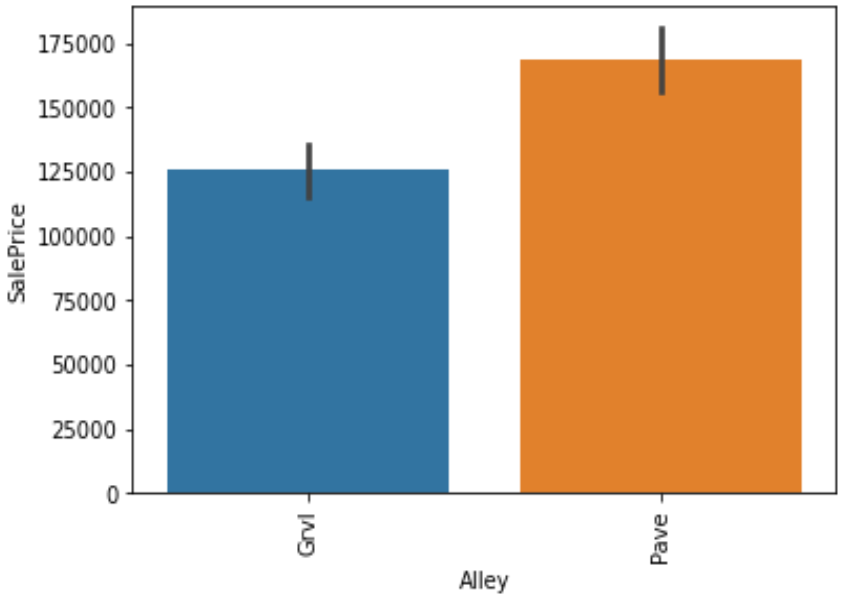
3-Alley: Type of alley access to property

GrvlGravel

Pave Paved

NA No alley access





Obs-

1- Paved alley has higher sale prices than gravel alley

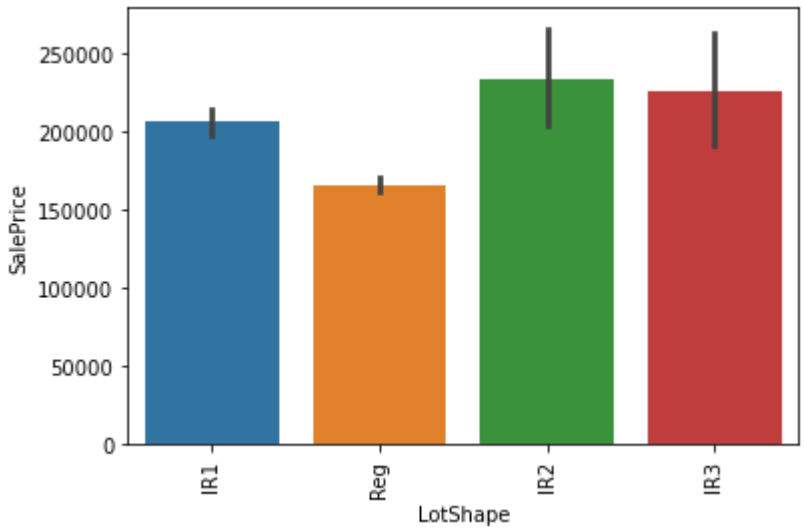
4- LotShape: General shape of property

Reg Regular

IR1 Slightly irregular

IR2 Moderately Irregular

IR3 Irregular



Obs-

1- Moderately irregular have highest sale price

2-Regular shape have lowest sale prices

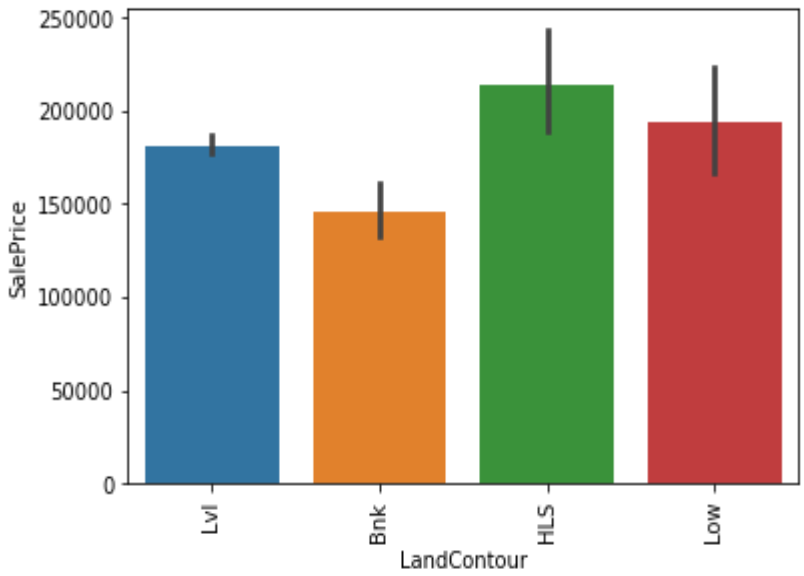
5- LandContour: Flatness of the property

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



Obs-

1- Hill side property have high sale prices

2-Depression and Near flat property have similar prices

3-Banked have the lowest prices

6-LotConfig: Lot configuration

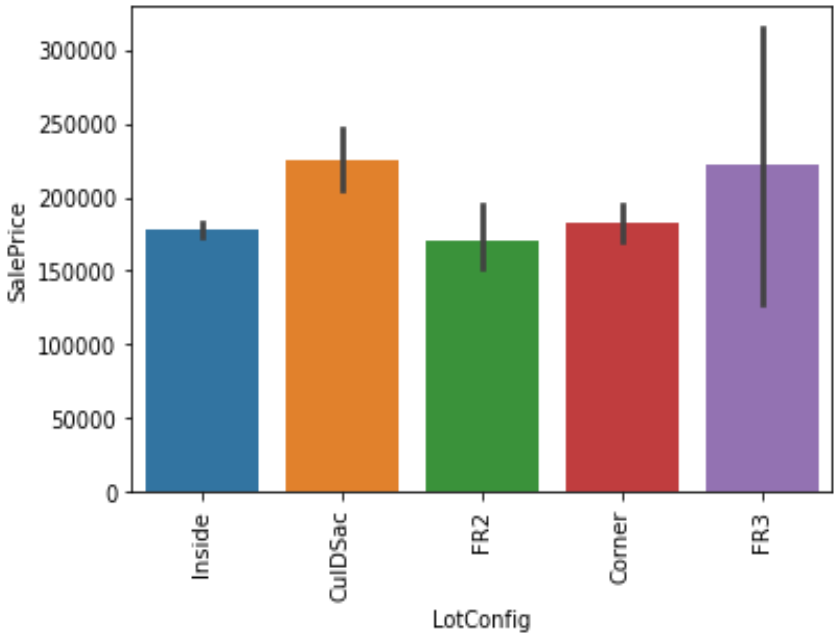
Inside Inside lot

Corner Corner lot

CulDSac Cul-de-sac

FR2 Frontage on 2 sides of property

FR3 Frontage on 3 sides of property



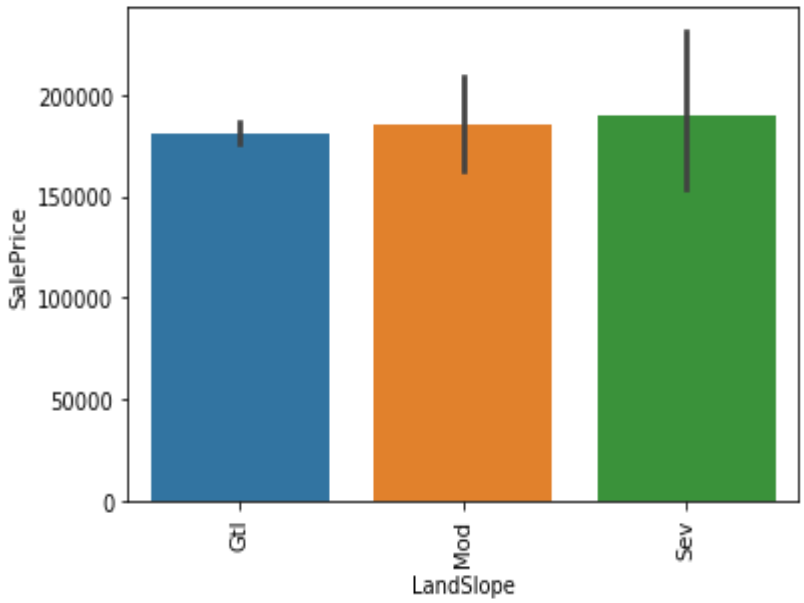
Obs-

1-Fr3 have the highes sale price followed by CuldSac

2-FR2 and Inside have similar house prices

7-LandSlope: Slope of property

Gtl	Gentle slope
Mod	Moderate Slope
Sev	Severe Slope



Obs-

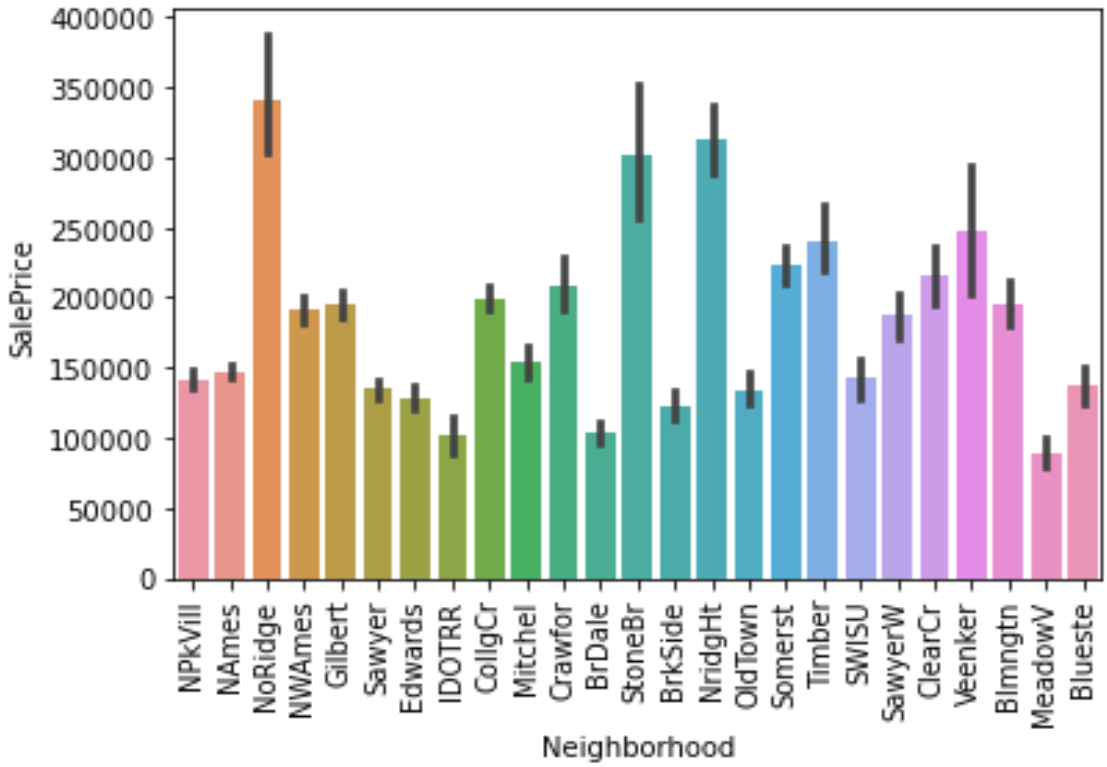
1-Severe slope has higher prices

2-Moderate and Gentle slope have similar prices

8-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



obs-

1- NoRidge has the highest sale prices followed by StoneBr and NridgHt

2- MeadowV has the lowet sale price followed by BrDale, IDOTRR

9- Condition1: Proximity to various conditions

Artery Adjacent to arterial street

Feedr Adjacent to feeder street

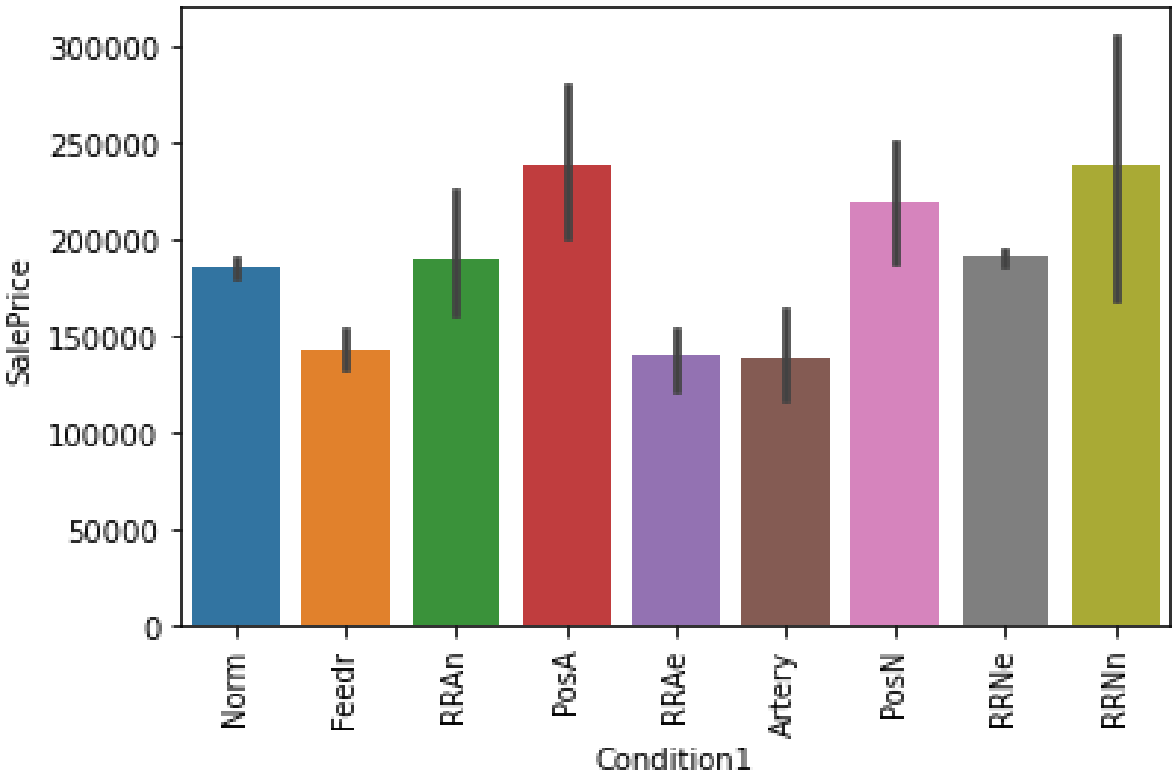
Norm Normal

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



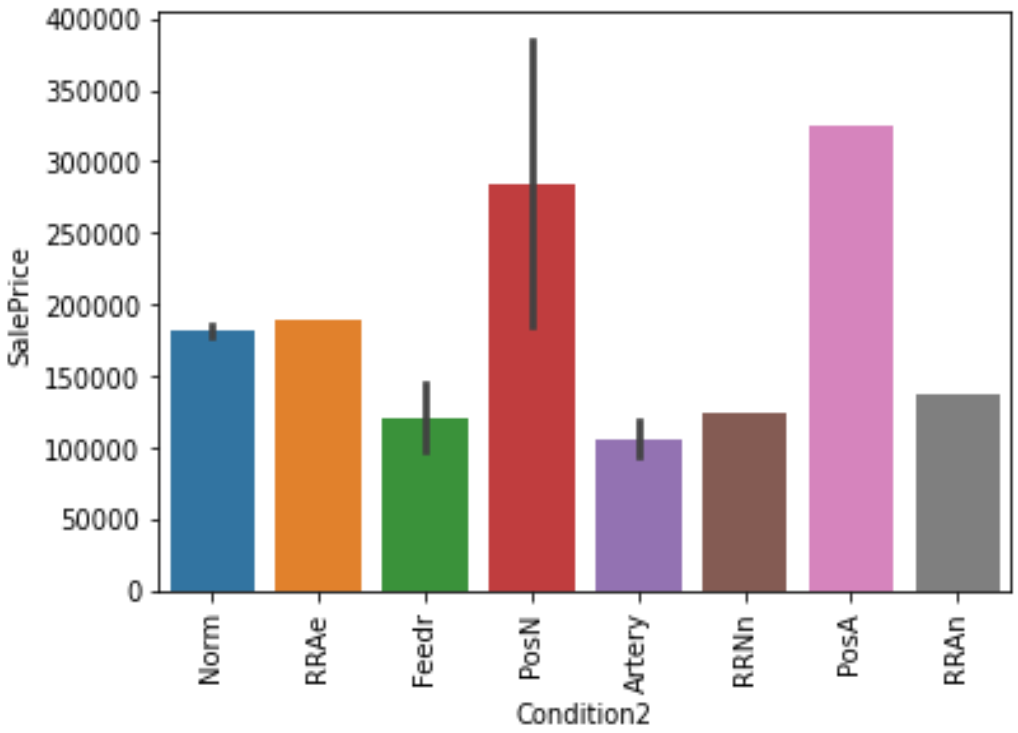
Obs-

1-RRNn (Within 200' of North-South Railroad) has the highest sale prices followed by PosA (Adjacent to postive off-site feature)

2- RRAe (Adjacent to East-West Railroad) and Artery(Adjacent to arterial street) has the lowest sale prices

10-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

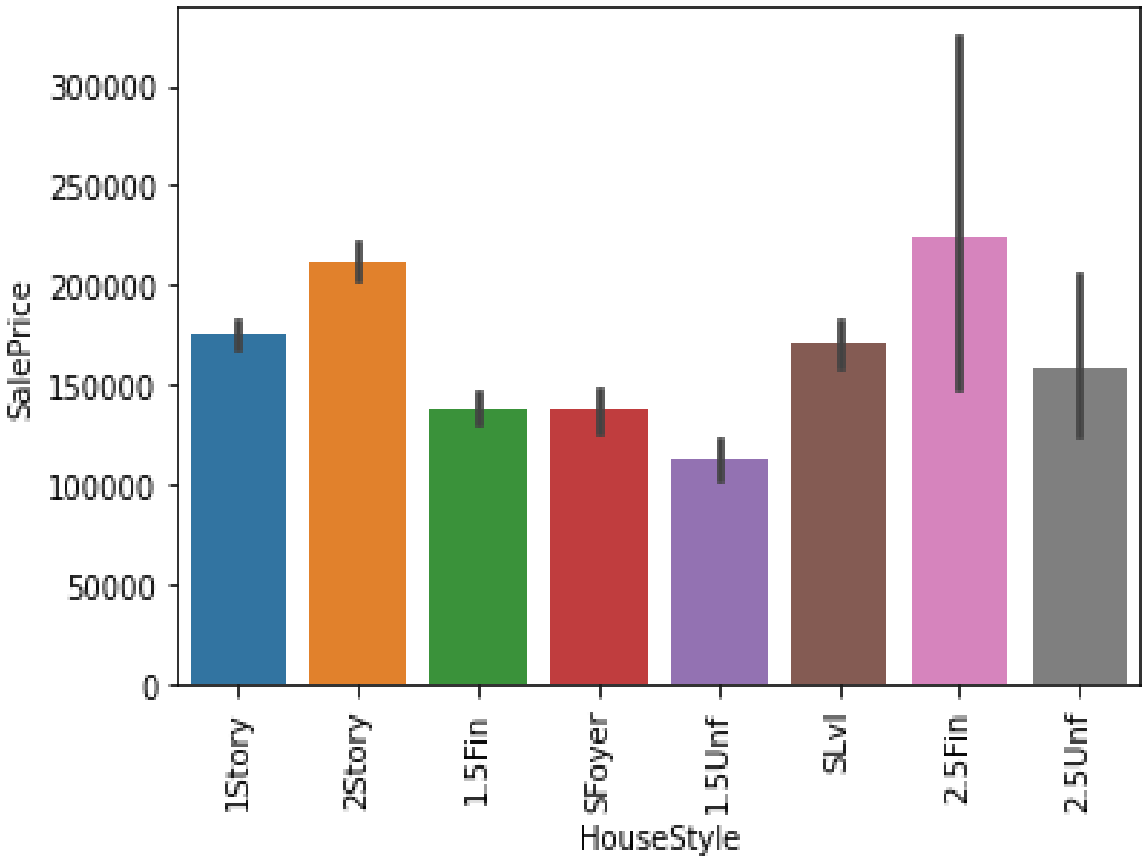


Obs-

1- PosN (Near positive off-site feature--park, greenbelt, etc.) has the highest sale

prices followed by PosA(Adjacent to postive off-site feature
2-Artery(Adjacent to arterial street) followed by Feedr(Adjacent to feeder street) has the lowest sale prices

11-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



Obs-

1-2.5Fin (Two and one-half story: 2nd level finished) has

the highest sale price followed by 2story,hen by 2.5Unf

2-1.5Unf has he lowest sale prices

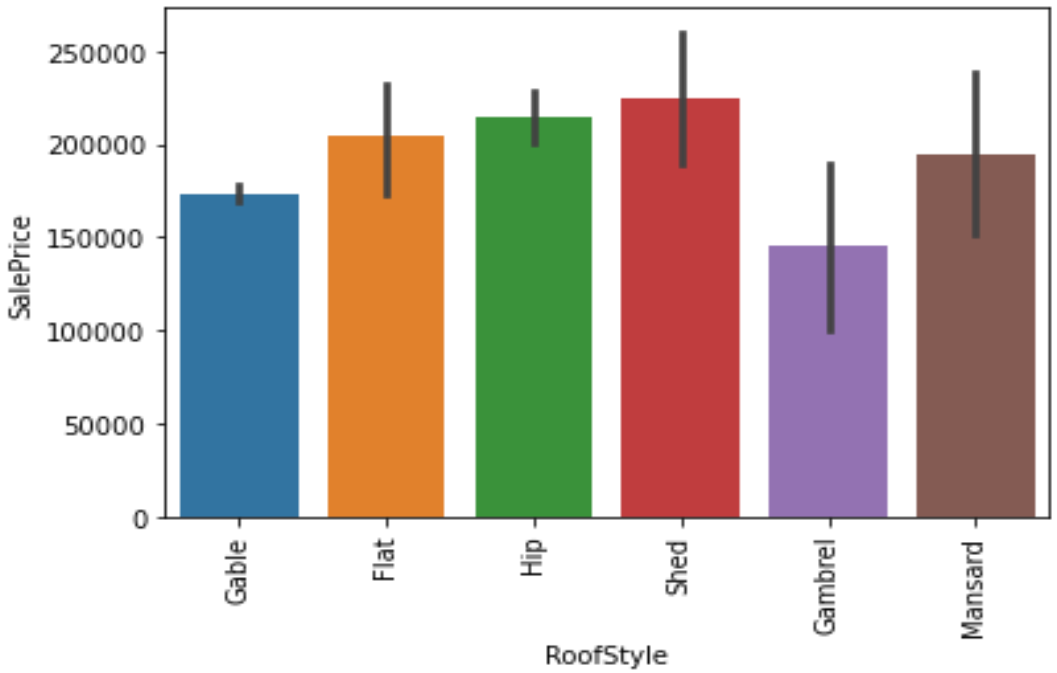
12-RoofStyle: Type of roof

obs-

1- Shed style rrof type has highest sale price

2-Hip, Flat, and Mansard have similar prices

3-Gambrel has the lowest prices



obs-

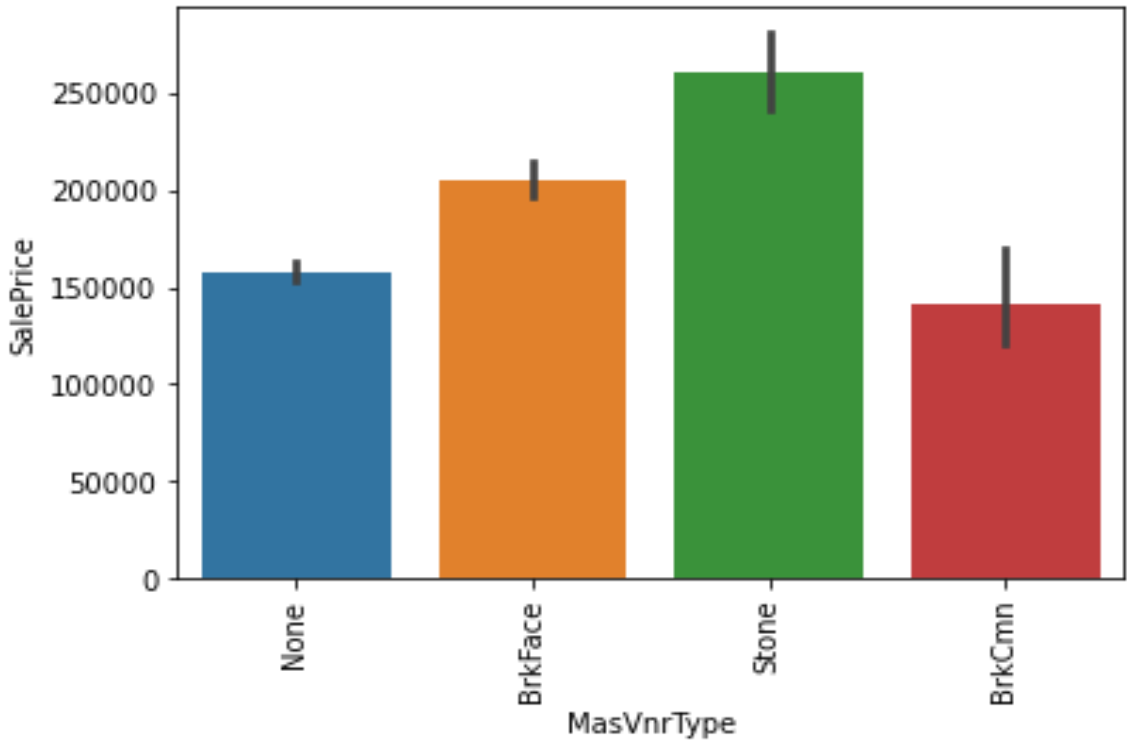
1- Shed style rrof type has highest sale price

2-Hip, Flat, and Mansard have similar prices

3-Gambrel has the lowest prices

13- MasVnrType: Masonry veneer type

BrkCmn	Brick Common
BrkFace	Brick Face
CBlock	Cinder Block
None	None
Stone	Stone



Obs-

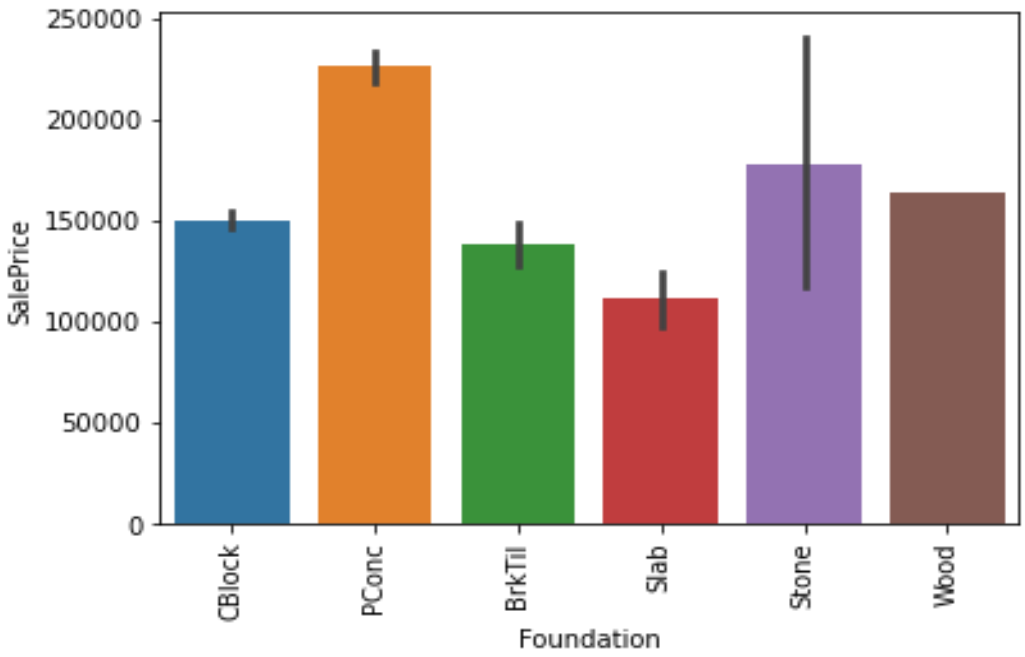
1- Stone masonry have the highest sale price

2- followed by Brick face

3- There are also houses with none of the Masonry type

15-Foundation: Type of foundation

BrkTil	Brick & Tile
CBlock	Cinder Block
PConc	Poured Concrete
Slab	Slab
Stone	Stone
Wood	Wood



Obs-

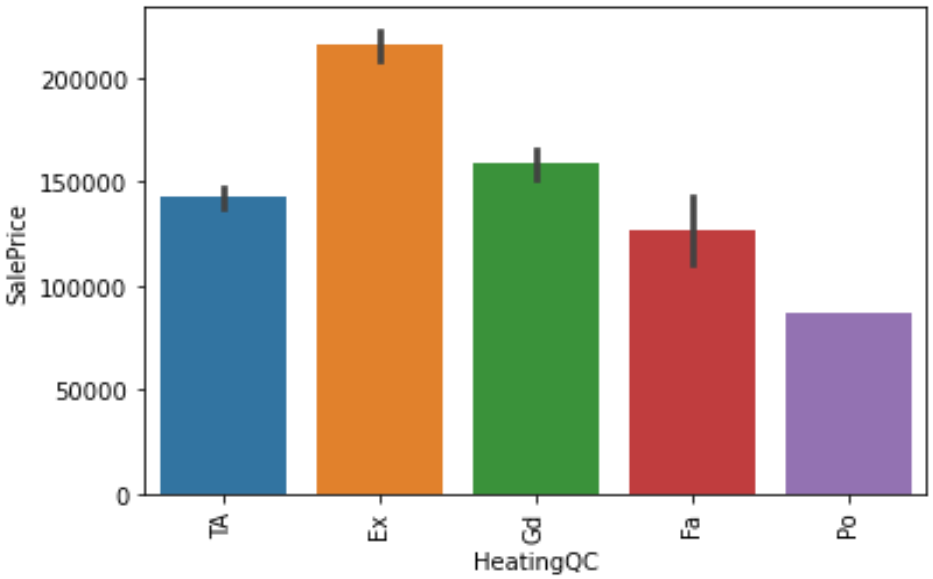
1-Poured Cocrete have the highest sale price Followed by stine foundation.

2-followed by Wood,CBlock ,BrkTil foundation

3- Slab foundation has the least sale prices

16HeatingQC: Heating quality and condition

Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
Po	Poor-



Obs-

1-The better the heating quality condition the higher the prices

17-Electrical: Electrical system

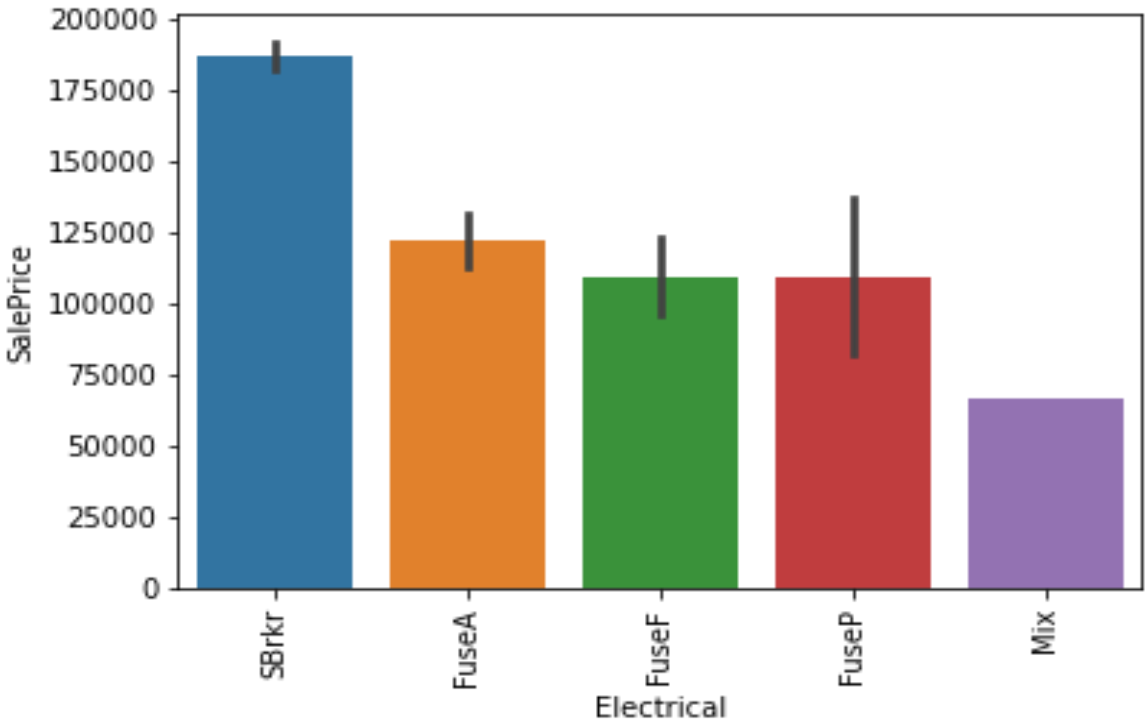
SBrkr Standard Circuit Breakers & Romex

FuseAFuse Box over 60 AMP and all Romex wiring (Average)

FuseF60 AMP Fuse Box and mostly Romex wiring (Fair)

FuseP60 AMP Fuse Box and mostly knob & tube wiring (poor)

Mix Mixed



Obs-

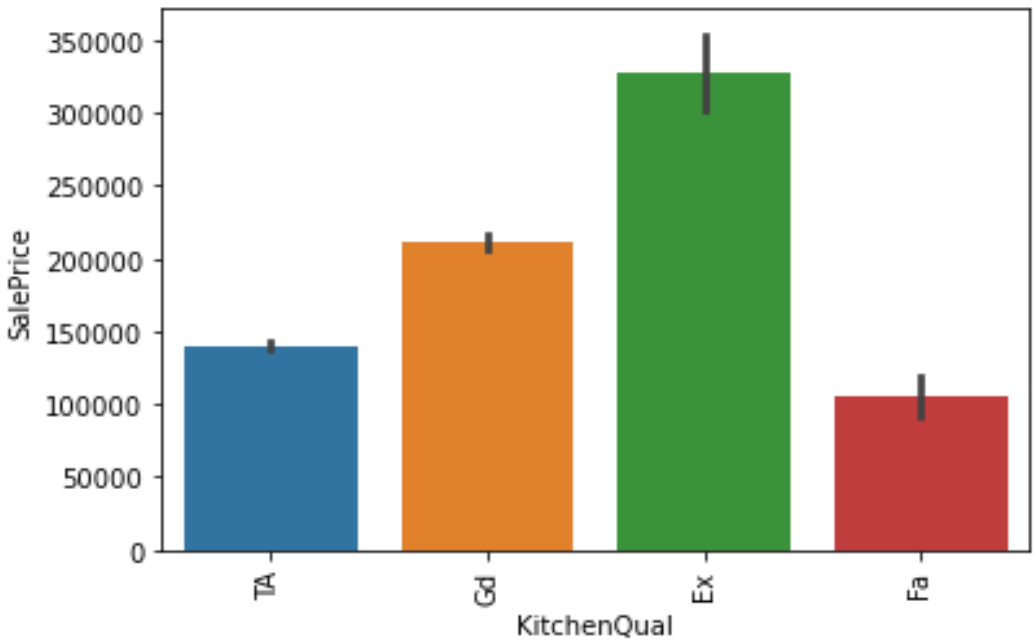
1-house having SBrkr have high sale prices.

2-followed by houses having FuseA

3- Houses with FuseF and FuseP have similar prices.

18-KitchenQual: Kitchen quality

Ex	Excellent
Gd	Good
TA	Typical/Average
Fa	Fair
Po	Poor



1-The better the kitchen quality the higher the prices.

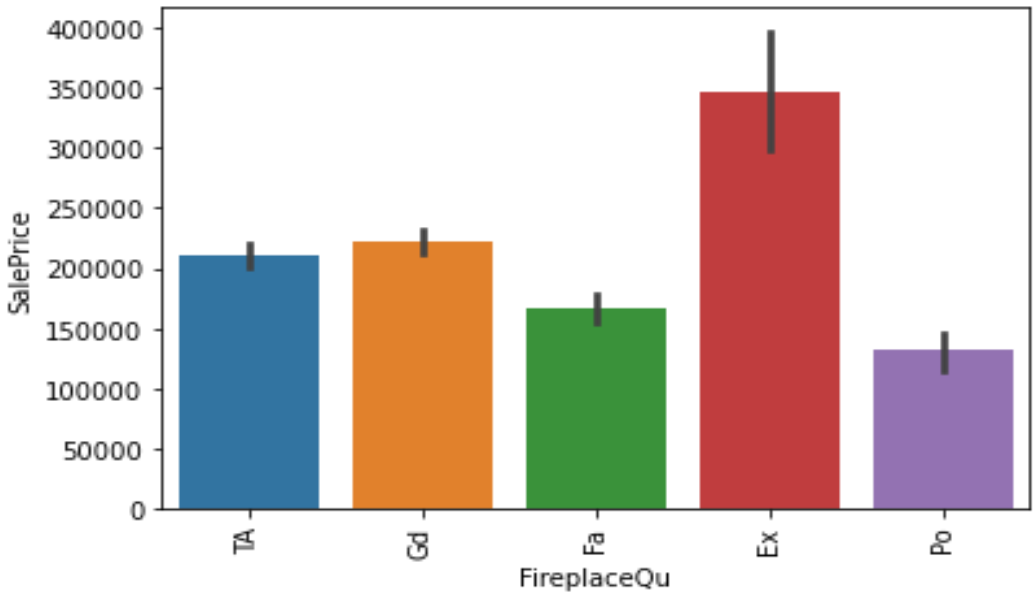
19-FireplaceQu: Fireplace quality

Obs-

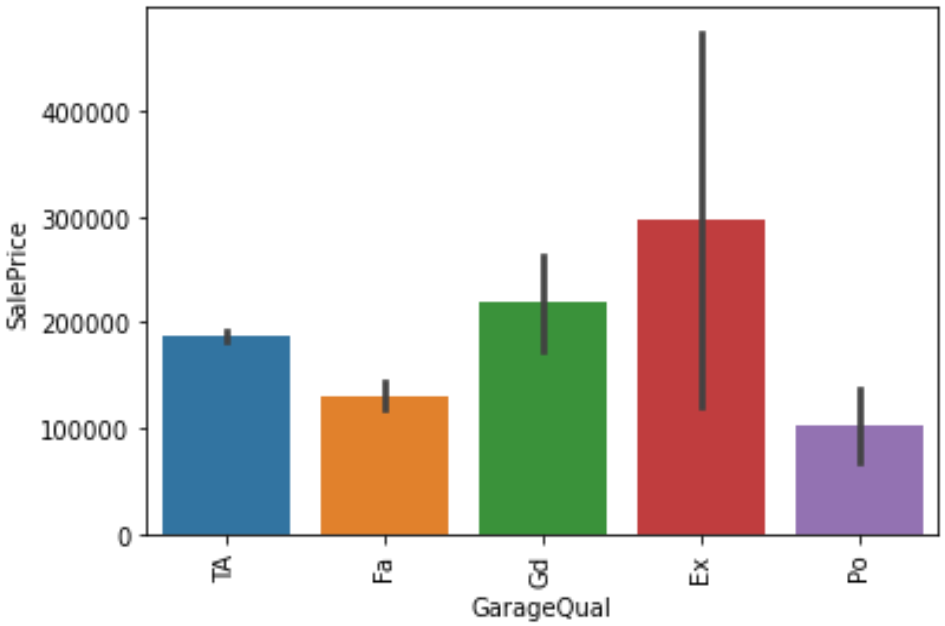
1-The better the FirePlace quality the higher the prices.

2- Some houses even have poor quality, but still sold

for more than 1lakh



20-GarageQual: Garage quality



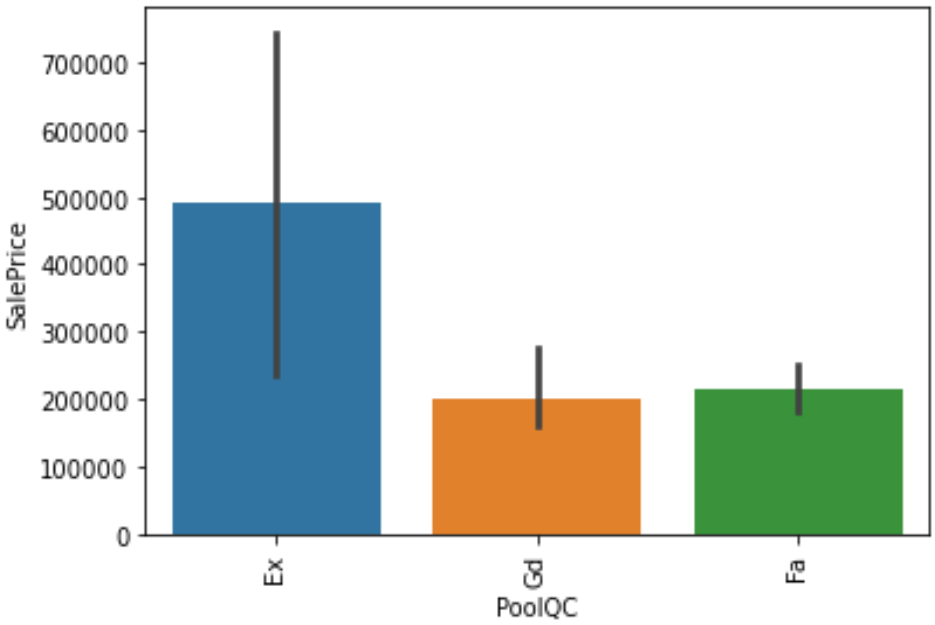
Obs-

1-The better the Garage quality the higher the prices.

2- Some houses even have poor Garage quality, but still sold

for more than 50000

21-PoolQC:Pool Quality



Obs-

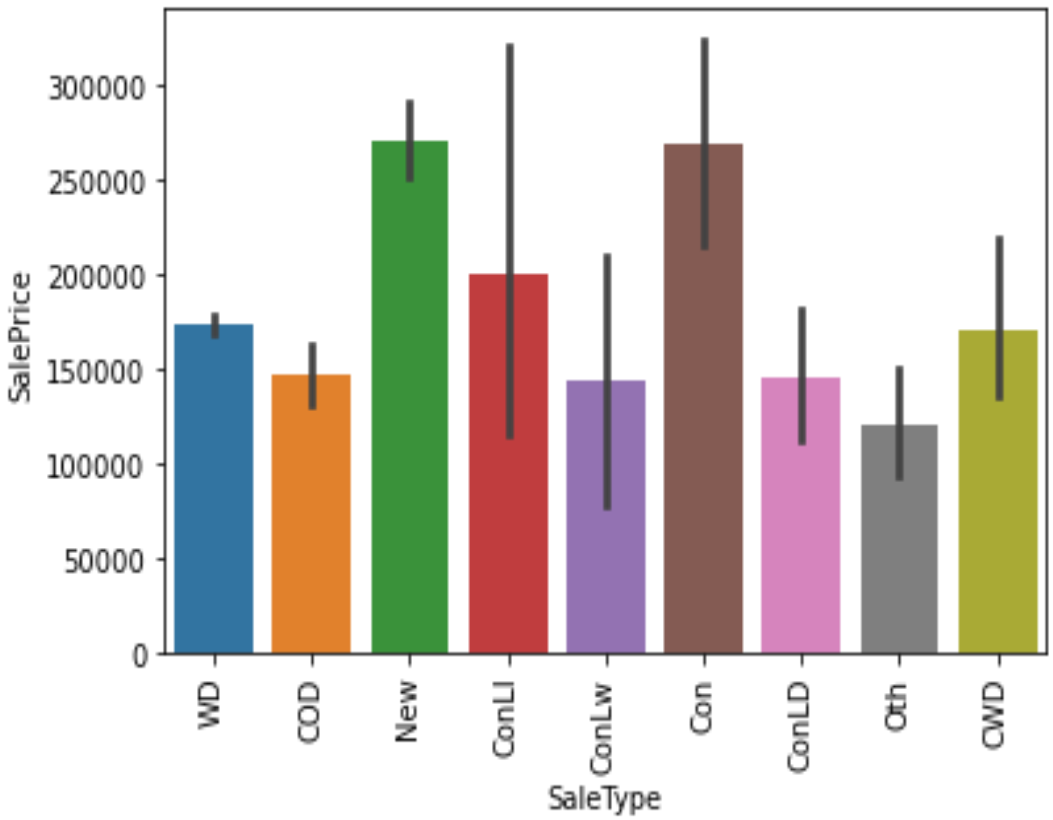
1-The better the Garage quality the higher the prices.

2- Houses with Good and Fair quality pool quality have
similar prices

22-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



Obs-
1-Con(Contract 15% Down payment regular terms) have the highest sale prices, followed by New(Home just constructed and sold), ConLi(Contract Low Interest), CWD Warranty Deed – Cash
2- Other type of houses have less sale price

23-SaleCondition: Condition of sale

Normal Normal Sale

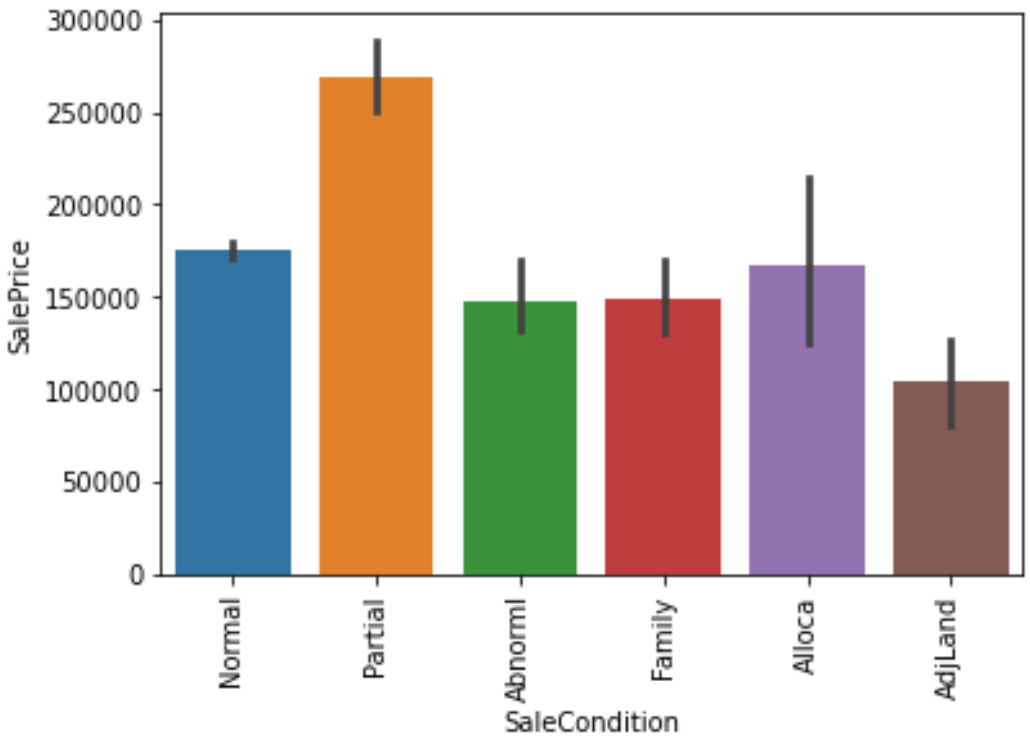
Abnorm Abnormal Sale - trade, foreclosure, short sale

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



Obs-

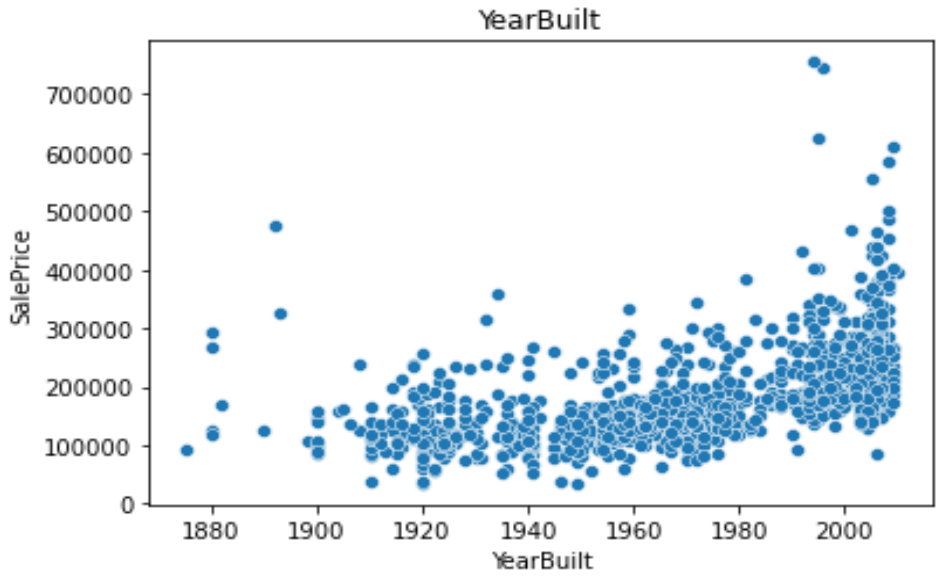
1- Houses with Partial sale condition have the highest sale

prices.

2-followed by Normal and Alloca Sale condition
3-Abnormal and Family type have similar sale prices.
4-Adj Land have the least sailing Prices

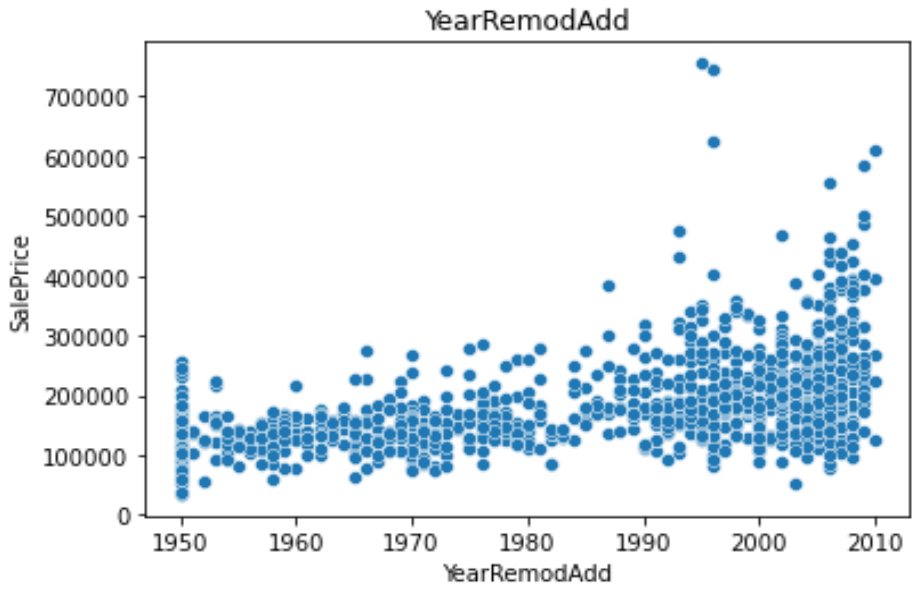
B- Year Features

1-Year Built- In which year he model was built



Obs-
1-The newer the model the better the sale prices

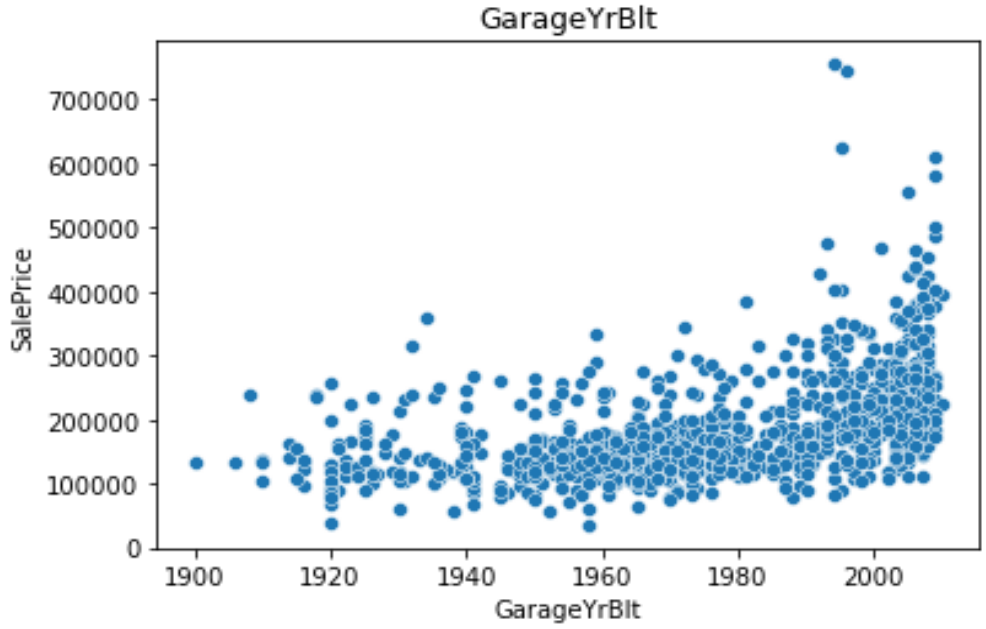
2-YearRemodAdd: Remodel date (same as construction date if no remodeling



Obs-

1-The latest the remodel date, the higher the prices

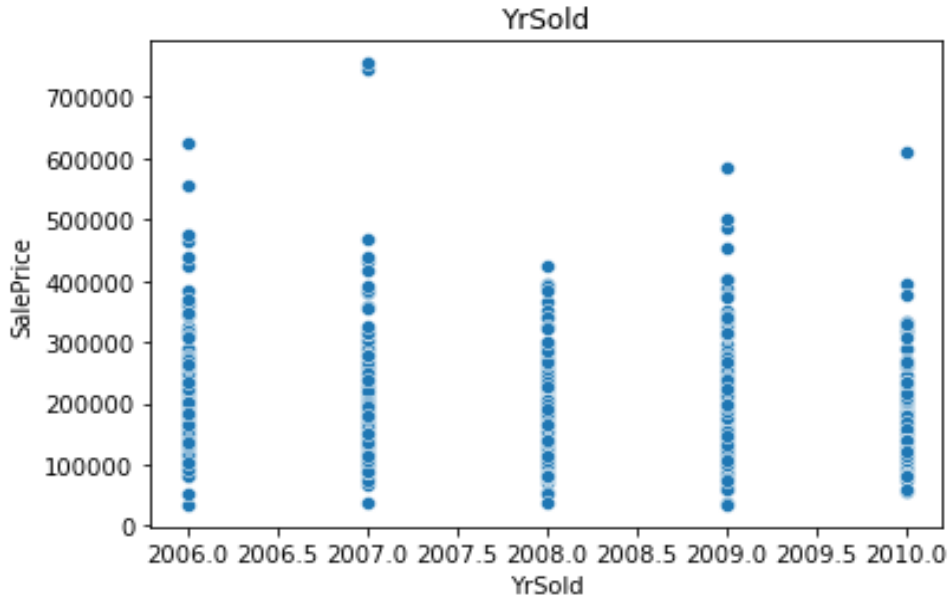
3- GargeYrBuilt: Garage Building year



Obs-

1-The newer the Garage Building date, the higher
the prices

4-Year Sold

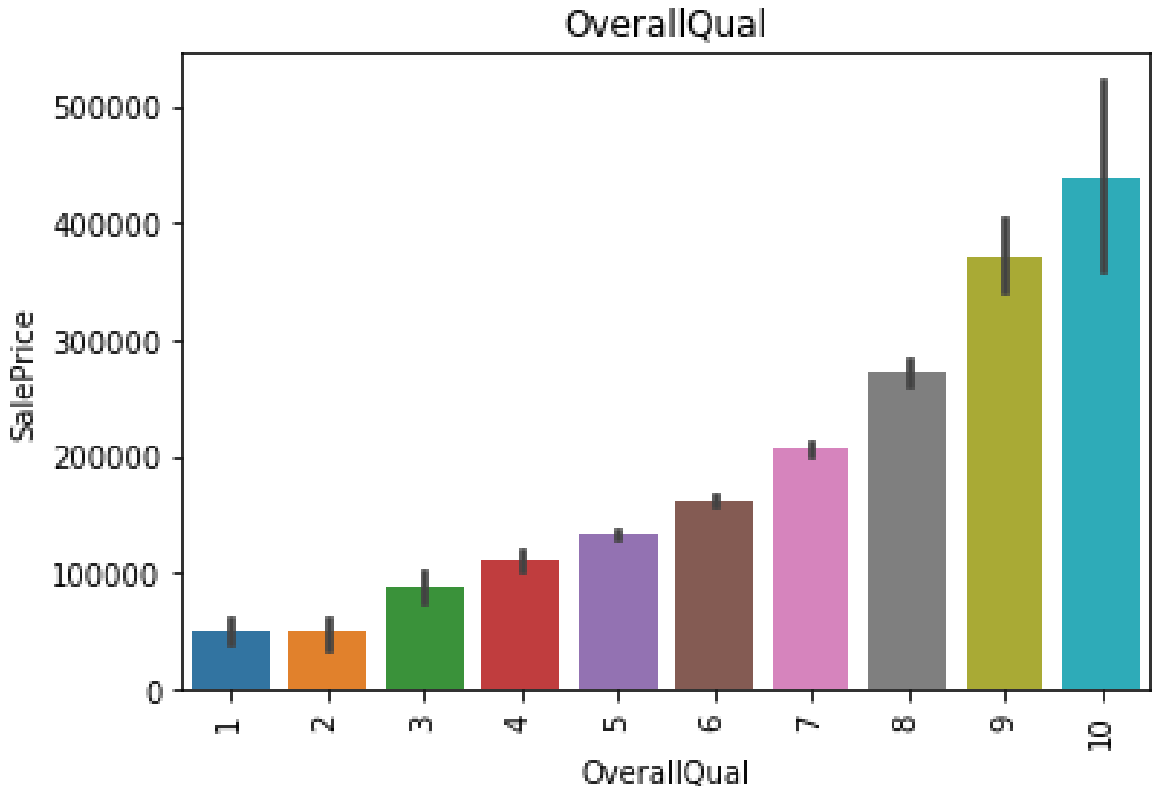


Obs-
1-In 2007 highest price house were sold upto 7lakh
2-In 2008, the prices of the house went below 5lakh

C-Discrete Features-

1-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

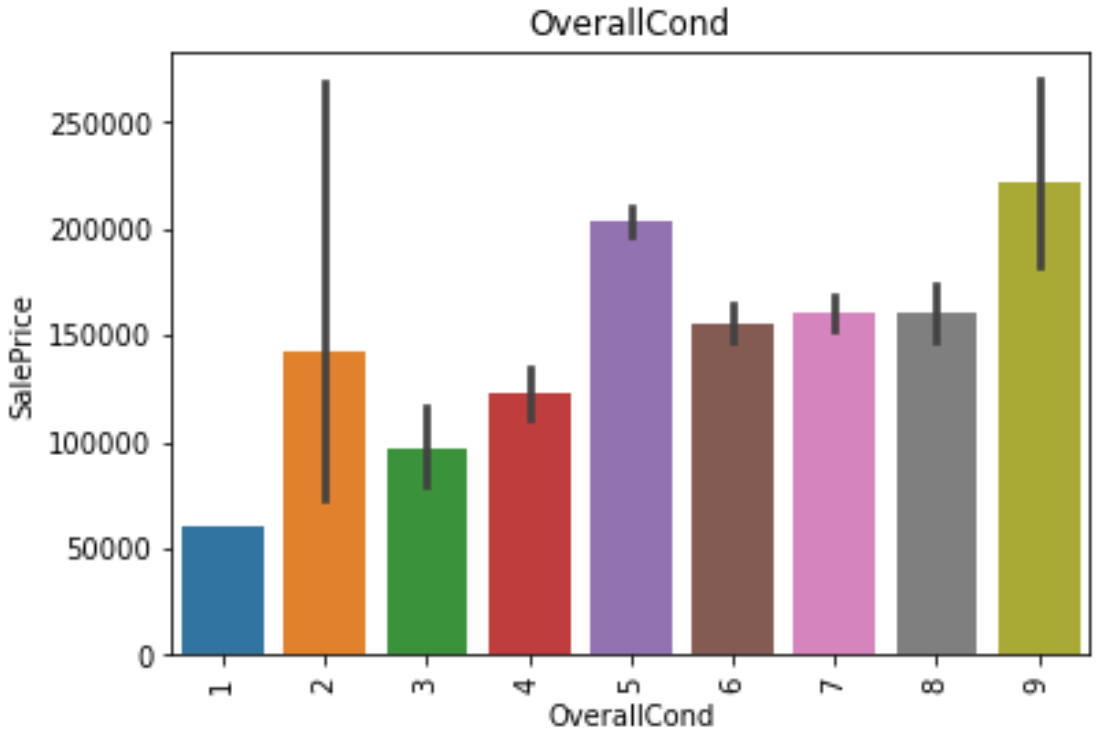


Obs-

1-The better he quality of the house the higher the prices.

2-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



Obs-

1-Excellent conditioned houses have the highest sale prices

2-Average conditioned houses have the 2nd highest sale prices
3-6,7,8 conditioned houses have similar prices
4-Very poor conditioned houses have the least prices

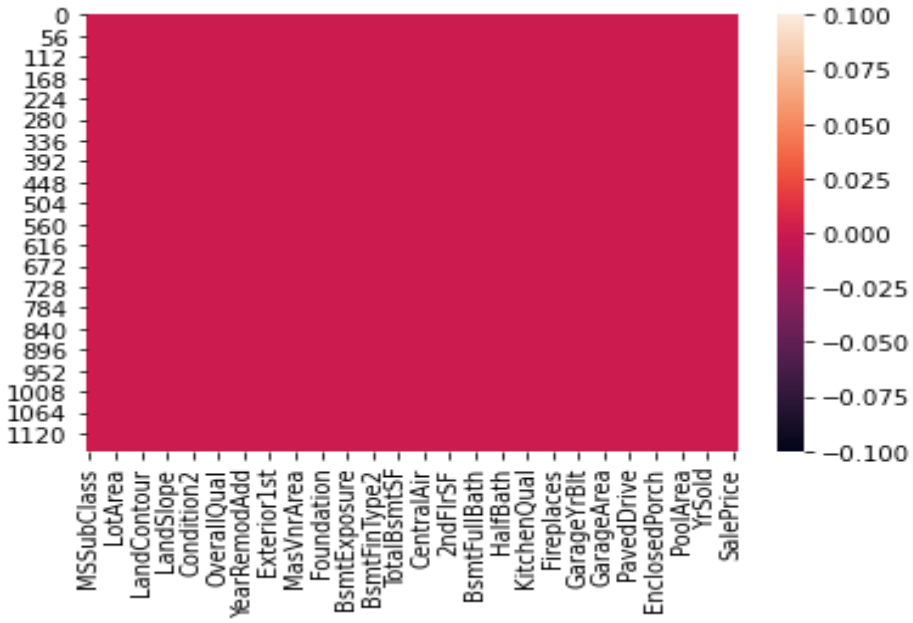
2- **Working on Testing and Training Datasets**

1- finding the null values on the training set.

2- filling the null values, with mean in numerical columns and,with mode in categorical columns.

3-Eliminaing the columns having more than 70% null values

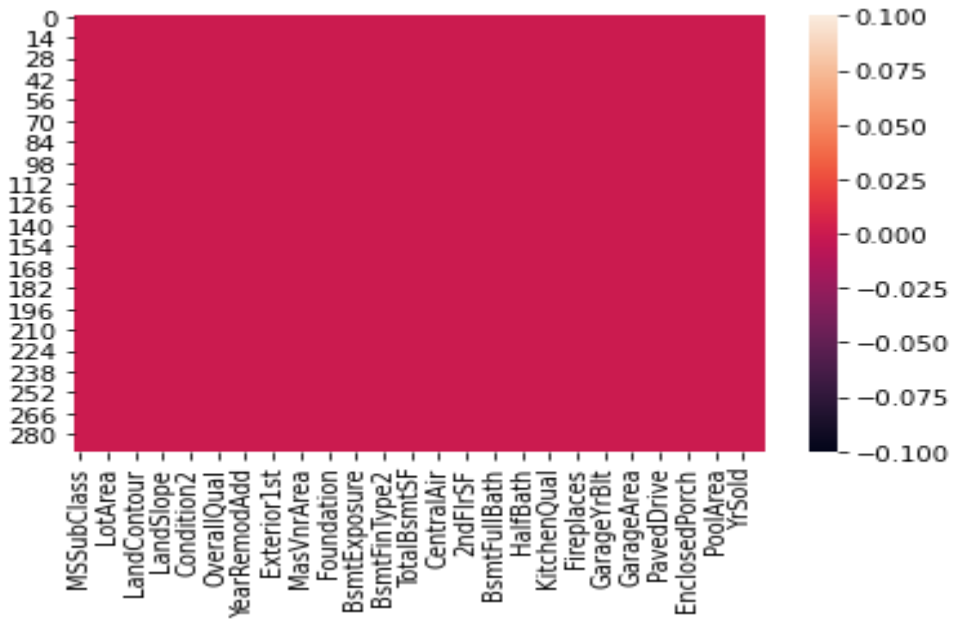
4-



Heatmap showing, no null values in the training dataset.

5- Performing the same feature engineering steps and filling the null values in the testing dataset.

6-



Heatmap showing , no null values present in the testing dataset.

7- Joining both the training and testing datasets.

8- Encoding using, Ordinal Encoder, for encoding the categorical columns in the d(train+test) set.

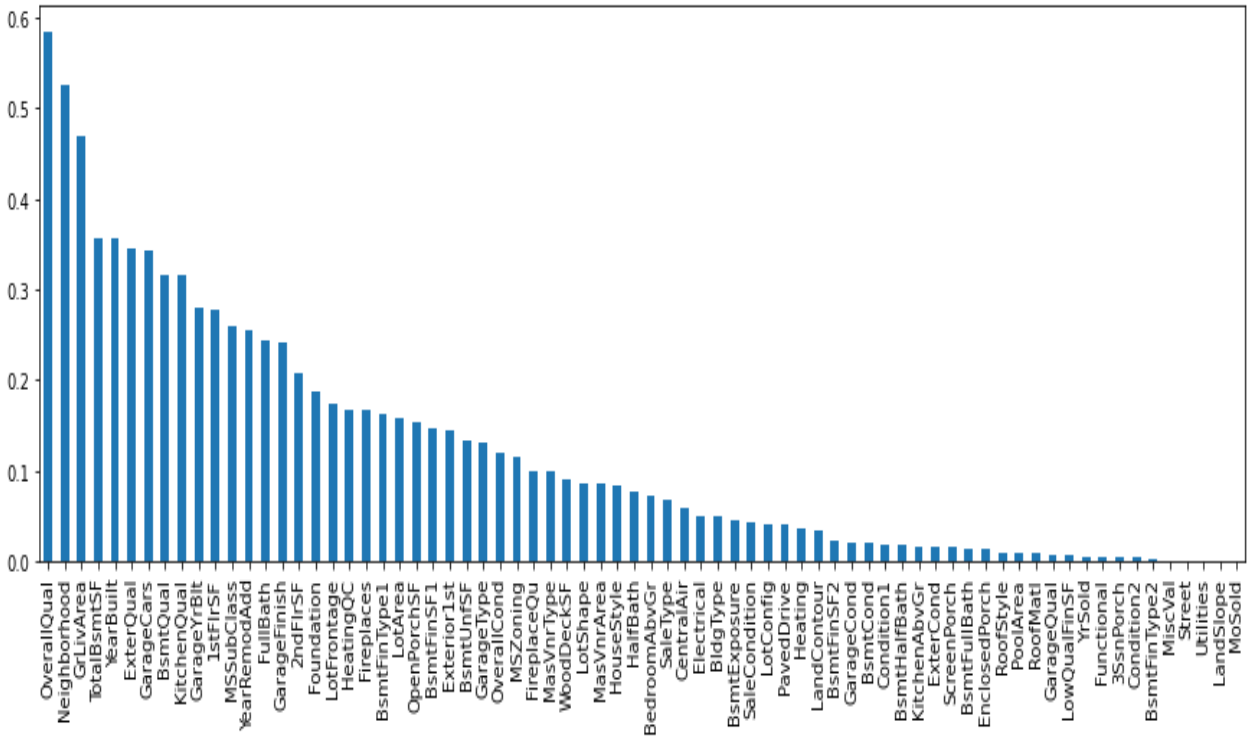
9-Again splitting the two sets, using iloc method, into d_trainset and d_testset

10- Using the Variance Threshold method to find out the feature having only same response,(i.e- Utilities) and removing it from both d_trainset and d_testset

11- Dropping columns based on correlation using Mutual Info .

12- Now separating x(feature) and y(target) variable from the d_trainset.

13- finding out the most contributing feature towards the targe variable and removing the least contributing feature from d_testset and d_trainset.(i.e-



Obs-

1-The histogram shows feature columns in decreasing order of their importance towards the target variable

2-OverallQual followed by Neighbourhood and GrLivArea are the most contributing features for housing price

3-Least Important Feature-MoSold,Condition2,Street,GarageQual,BsmtFinType2

14- using Standard Scaler for scaling the x feature.

Model/s Development and Evaluation

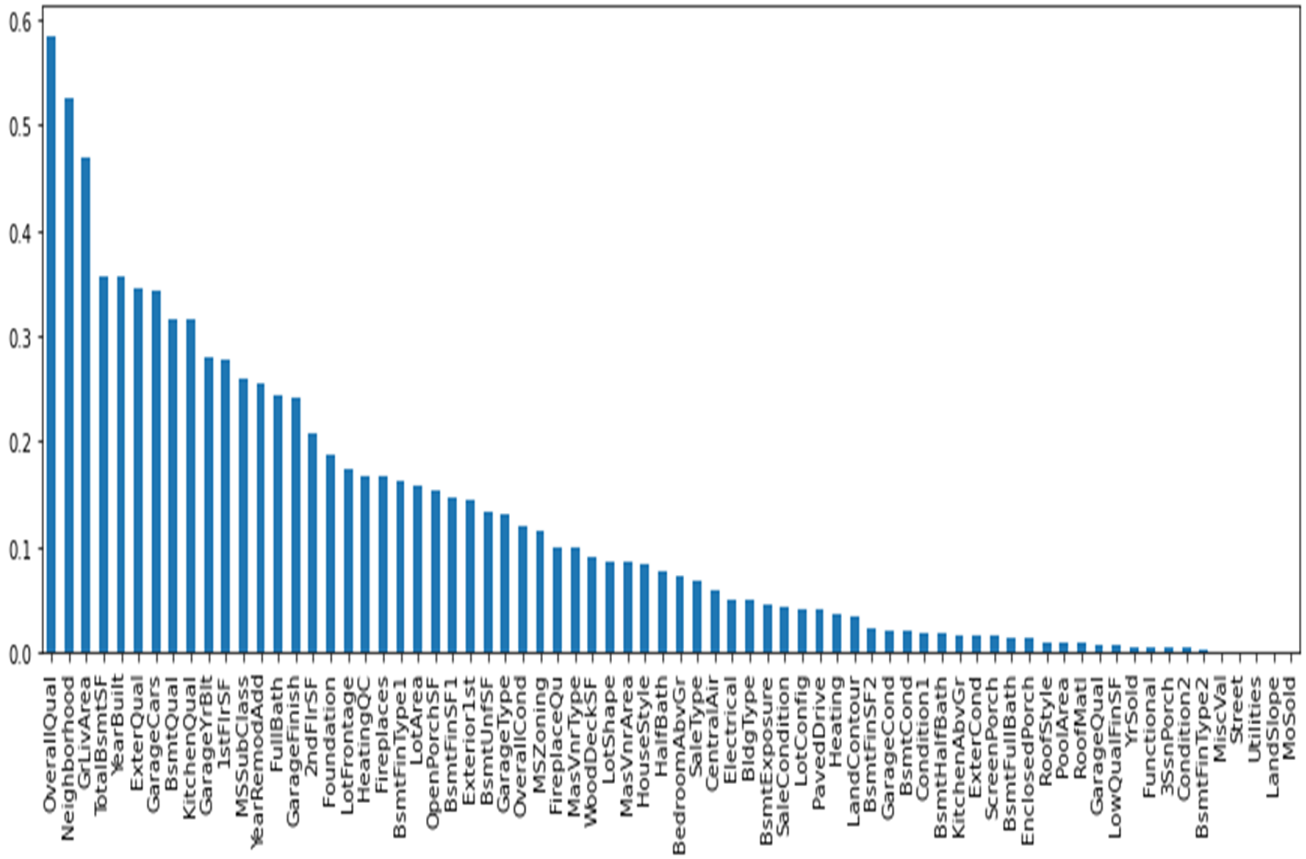
1-After selecting the x and y variable,we then perform the train_test_split and and check the training and testing in various models.
2-Models used – LinearRegression, DecisiontreeRegressor, KNeighborsRegressor, SVR, RandomForestRegressor, AdaBoostRegressor, GradientBoostRegressor,
3-Train_test_split was used to find out the accuracy of each model.
4-After this cross_val_score of each model was calculated.
5-By compairing the accuracy score with cross_val_score, GradientBoostRegressor was found to be most efficient model.
6-After this Hypertest tuning was done through GridSearchCv to find out the best parameters for GradientBoostRegressor

Prediction

1- GradientBoostingRegressor was used for predicting the prices on the d_trainset showed ,84% accuracy.

2- GradientBoostingRegressor was used for predicting the Sale Price on the testing set.

Conclusion



1.OverAllQual,Neighbourhood,GrLivArea,TotalBsmtSf,YearBuilt,,Exernalqual,BSmtQual,KitchenQual,GarageYrBuilt,1sFlrSf,MSClass,YarModAdd,FullBath,GarageFinish,2ndFlrSF,Foundation,LotFronage, HeatingQl,FirePlace,LotArea,OpenPorschSF, are the feaures arranged in the decreasing order of their contribution, respectively to the Sale Prices of the Housing

2.LeastImportantFeatures-MoSold,Condition2,Street,GarageQual,BsmtFinType2

Limitations

1- GradientBoosting regressor is working with 84% accuracy.

2-The predicted values have margin of error of 16%.

Thank You.