

# **HOUSE PRICE PREDICTION**

Submitted by:

**MOHAMMED MINHAJ**

## **ACKNOWLEDGMENT**

Wikipedia, Youtube.com, google.com. The data were collected from the used car website cardheko.com. Referred in study materials of machine learning which I am studying.

# INTRODUCTION

- **Business Problem Framing**

Population is increasing in the world. Humans want shelter to live, it is the basic feature of humans. So this feature can make a business and earn profit. But the house want some features to get price demands. Quality of the house, is it suitable for living, age of the house and condition of the house. These are basic features of the house there are some other features that demands, the firm or organisation want to focuses on that feature which is and try to give that features for the house which they are selling. As the firm is starting their business in Australia, want to ensure that the house can survive the climate of the country, which impacts the price of the house.

- **Conceptual Background of the Domain Problem**

As we want to predict the house price within features first of all we want to build a regression model. Try to relate all features with price how they are related to price by visualization.

- **Review of Literature**

Every country have different culture, climates so want to choose the right feature for this. Like Australia have a different climate to other countries. So want to select the features to the house which have the ability to survive the climate. We can take an example that in such country have the possibility of heavy snow falling, in such cases a house with flat roof cannot survive that, so there need a roof which have curve. So same possibility is not going to happen in the Africa. So when country changes the needs and feature should change. So first of all focus on the country's climate and what are features want to the house for selling, to get demands in price.

- **Motivation for the Problem Undertaken**

To help those people who have no houses or shelter for living. The firm want to know the business possibilities in Australia with the past data. And analyse and predict how much it is possible the business In Australia. Objective among this project to help a firm of house selling to start their business in Australia.

## **Analytical Problem Framing**

- **Mathematical/ Analytical Modelling of the Problem**

Predictions are done by analysing the given data. The given data will be the past data. With this past data we want to predict future. So for analysing the past data we need to use some techniques for that. We want to relate each features, how they are related, how they related to the output. In this project the features year when the house built, garage, Year of re-modification, basement. These features are related to the label price and some of them were will give a high impact in output. There are so many features have house, it id the up to the perspective of the customer who buys it.

- **Data Sources and their formats**

In this project contains of 1460 entries with 81 variable. Train and test data are given separately. 1168 entries in train dataset and 292 entries in test dataset.

Desktop/Python/

house\_pricing - Jupyter Notebook

localhost:8889/notebooks/Desktop/Python/house\_pricing.ipynb

Apps Gmail YouTube Maps Mushroom - Jupyter...

Reading list

jupyter

house\_pricing (autosaved)

Python

Logout

File

Edit

View

Insert

Cell

Kernel

Widgets

Help

Not Trusted

Python 3 C

Save

+

Undo

Redo

Run

Stop

Restart

Markdown

In [3]:

1 train.head()

Out[3]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities
127	120	RL	NaN	4928	Pave	NaN	IR1	Lvl	AllPub	
889	20	RL	95.0	15865	Pave	NaN	IR1	Lvl	AllPub	
793	60	RL	92.0	9920	Pave	NaN	IR1	Lvl	AllPub	
110	20	RL	105.0	11751	Pave	NaN	IR1	Lvl	AllPub	
422	20	RL	NaN	16635	Pave	NaN	IR1	Lvl	AllPub	

Type here to search

29°C Partly sunny

09:30 AM 09-02-2022

By analysing the data I understand the overall quality of house have play a major role in price. The price demands when the house have great overall quality. So try to maintain the quality of the house. Then another feature the year built, when the house built and difference between the year of built and the selling price plays a important impacts in price. When the difference increases price gets low. But also sure that year built and selling year is increases also try to concentrate on the modifying the house. This also makes the price to get high. Houses are built of materials which have only guaranty for few years, within that the material got destroyed and it may affect for the houses. So try to modify the house frequently for better price and safety for the people living in the house.

Another features is roof style and material used to make roof. Roof style hip, flat, gable it changes up to the climate condition of the place. Roof material is important feature, how strong is roof get safety. Masonry veneer type another feature which impacts the price. as we mention overall quality of house also the exterior quality of the house also impacts the price. It impacts negatively to price. Condition of the exterior also impacts the house price. Foundation house represents how the house is strong and it also

plays a major impacts in price. Quality of basement or in this project it evaluates the height of the basement by analysing the data I can understand the height of the basement plays a major role in impacts the price of the house. Basement finishing area rating is also an important feature for house price. Square feet of basement area, heating quality, air conditioning of the house, electrical system of the house, square feet of first floor, square feet of second floor, living area square feet, full bathrooms in first floor, half bathrooms in first floor these features also important to house pricing. When the garage built? Year of garage built is also important feature. Quality of kitchen, length of each bedroom in first floor, number of rooms in first floor, fireplaces, quality of fireplace, how much car can the garage store so these all are features that impacts to the price of the house in my perspective and analysing the data.

- **Data Pre-processing Done**

The first step I analyse in the data is there missing values in the data. As we want to predict price of the house we need built a regression model. First of all cleaning the data, I treated missing values. When we come to the data types of the features we can see half of them are categorical variables and remaining are numerical features. So as we know we want to use different techniques for different data types. So here in this project, I treated categorical features with mode and numerical features with mean. While we have some features which is not needed for prediction or the features which have above 80% of data are missing. We want sufficient data for prediction so I decide to drop such features have that much missing values and some features will not affect for our prediction also dropped. As usual, I encoded categorical variables with label encoder. These steps of data cleaning are also done in test dataset as well.

- **Data Inputs- Logic- Output Relationships**

We can assume that we are the customer of to buy the house, every human have different approaches and different perspectives. We look clear the basic features of the house. In this project, the main feature that impacts the price is year, which year the house built. We know that the materials built for house can be get destroyed. For example, in a business firm the assets like machineries get depreciated. So same case can happen in house materials which is built. So while selling that is the most important feature. Unless the business, safety is the first matter. Modifying such houses can get prices high. Overall quality of the house can get high demands in market. We can define overall quality in 10 and if a house gets overall quality, price increases. I assume that the overall quality can be defined or evaluated the quality of all area of the house it may be the kitchens, bedrooms living rooms or any other feature of the house.

- **State the set of assumptions (if any) related to the problem under consideration**

When trying to sell a house basic features like path to house, how much car can garage store, length of the basement, when last the house modified. Quality of kitchen, roof these are features which affects the price.

- **Hardware and Software Requirements and Tools Used**

About 739.2 Kb of memory were used by this project. This usage is only for the train dataset. Test dataset may be less. As usual pandas were imported for visualization importing of the dataset. Dataset is in CSV format. Numpy module imported. Typical visualization packages matplotlib and seaborn. Warnings packages imported for avoiding warnings. From sklearn packages import four algorithms to

perform in model. From tree import decision tree, from linear model linear regression is imported. And also regularization technique lasso regression imported. Bagging and boosting technique used. Ensemble technique random forest is used. And boosting technique, xg boost is imported. Metrics R2 score, RMSE, MSE were imported. From model selection cross validation score imported for checking Overfitting and Underfitting. Grid search is imported for tuning the parameters after finalising the model. From pre-processing packages one of the scaling technique Standard scaler is imported for feature scaling. For encoding label encoder is imported.

## **Model/s Development and Evaluation**

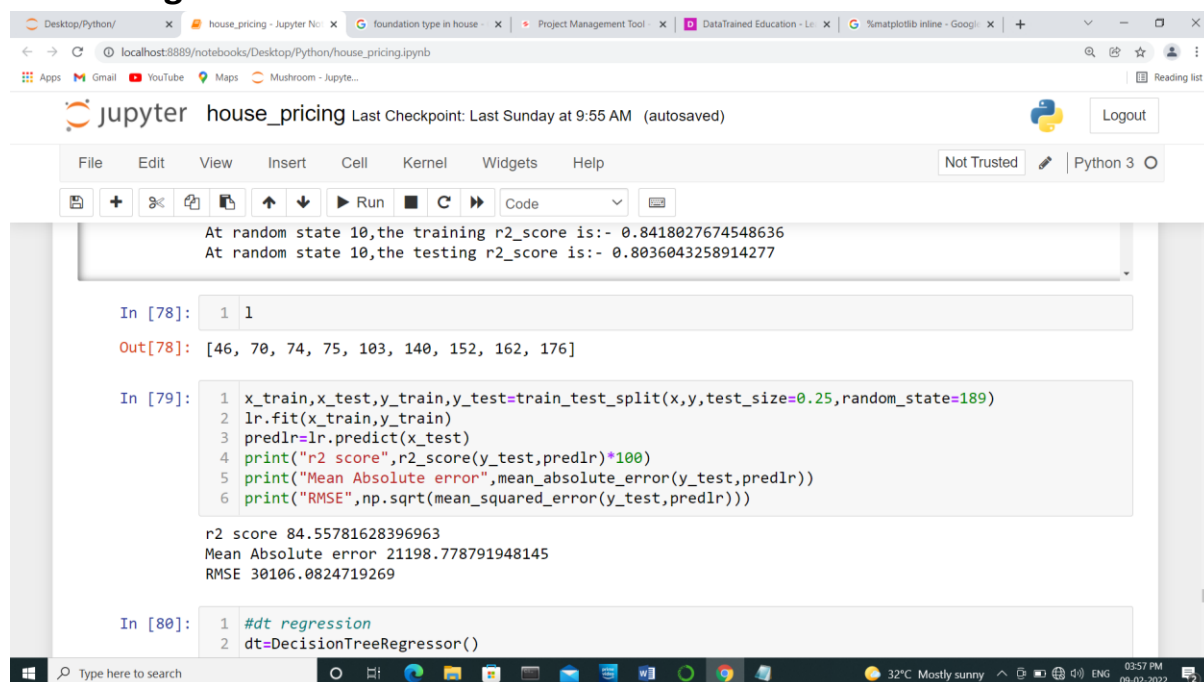
- Identification of possible problem-solving approaches (methods)

Our objective in this project is to predict the house price. For the prediction we need some features related to the label. Here for the price of the house is related to some features of the house. We can analyse the data by visualizations for better understanding of data.

- Testing of Identified Approaches (Algorithms)
  - a. Linear Regression
  - b. Decision Tree Regressor.
  - c. XGBoost Regressor
  - d. Random Forest Regressor
  - e. Lasso Regression(Regularzation)
- Run and Evaluate selected models



# Linear Regression



A screenshot of a Jupyter Notebook titled 'house\_pricing'. The notebook shows the results of a linear regression model. At the top, it states: 'At random state 10, the training r2\_score is:- 0.8418027674548636' and 'At random state 10, the testing r2\_score is:- 0.8036043258914277'. Below this, there are three code cells. The first cell (In [78]) contains the number 1. The second cell (In [79]) contains code to split the data and fit a linear regression model, with output showing an r2 score of 84.55781628396963, a Mean Absolute error of 21198.778791948145, and an RMSE of 30106.0824719269. The third cell (In [80]) contains code to create a DecisionTreeRegressor object.

```
At random state 10, the training r2_score is:- 0.8418027674548636
At random state 10, the testing r2_score is:- 0.8036043258914277

In [78]: 1 1

Out[78]: [46, 70, 74, 75, 103, 140, 152, 162, 176]

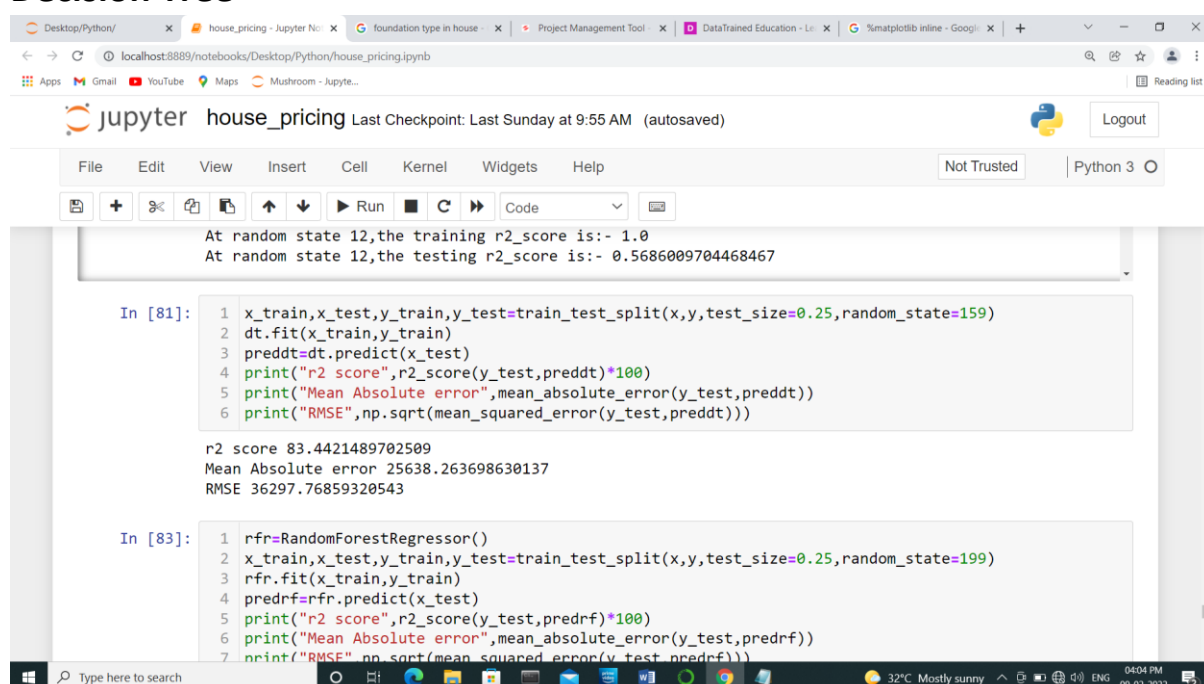
In [79]: 1 x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.25,random_state=189)
2 lr.fit(x_train,y_train)
3 predlr=lr.predict(x_test)
4 print("r2 score",r2_score(y_test,predlr)*100)
5 print("Mean Absolute error",mean_absolute_error(y_test,predlr))
6 print("RMSE",np.sqrt(mean_squared_error(y_test,predlr)))

r2 score 84.55781628396963
Mean Absolute error 21198.778791948145
RMSE 30106.0824719269

In [80]: 1 #dt regression
2 dt=DecisionTreeRegressor()
```

Here in the above picture we can observe that r2 score of linear regression is 84% at the random state 189. RMSE and MSE are also given that is little bit higher.

# Decision Tree



A screenshot of a Jupyter Notebook titled 'house\_pricing'. The notebook shows the results of a decision tree model. At the top, it states: 'At random state 12, the training r2\_score is:- 1.0' and 'At random state 12, the testing r2\_score is:- 0.5686009704468467'. Below this, there are three code cells. The first cell (In [81]) contains code to split the data and fit a decision tree model, with output showing an r2 score of 83.4421489702509, a Mean Absolute error of 25638.263698630137, and an RMSE of 36297.76859320543. The second cell (In [83]) contains code to create a RandomForestRegressor object.

```
At random state 12, the training r2_score is:- 1.0
At random state 12, the testing r2_score is:- 0.5686009704468467

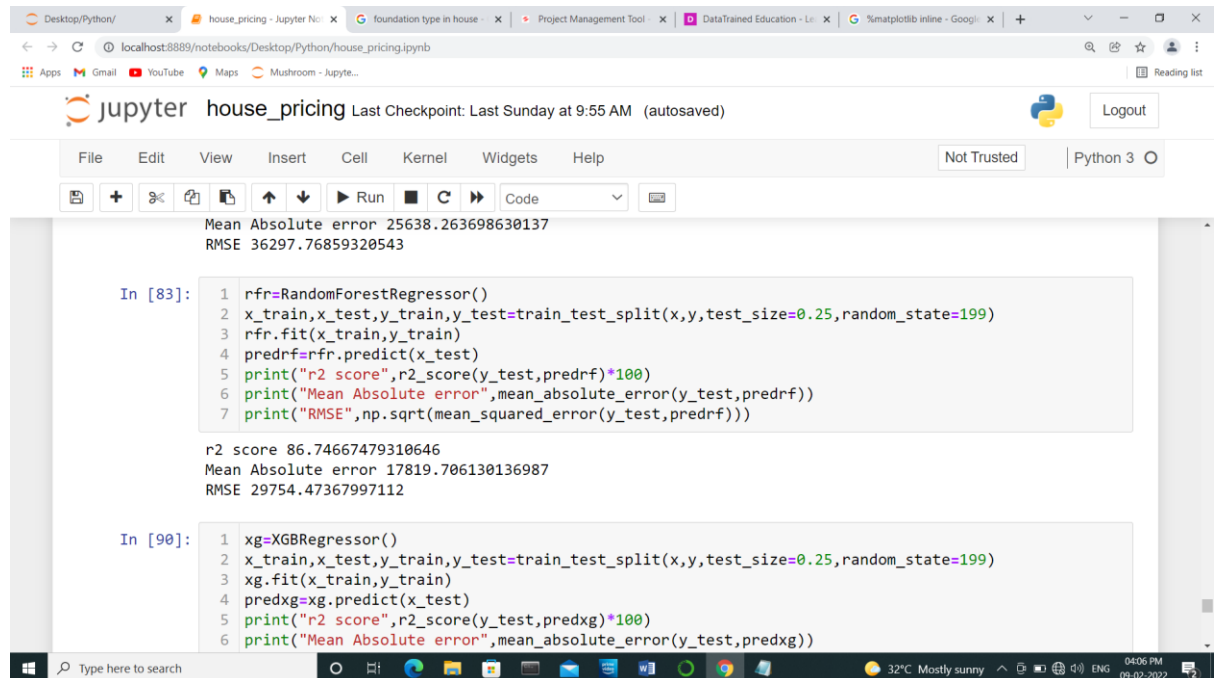
In [81]: 1 x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.25,random_state=159)
2 dt.fit(x_train,y_train)
3 preddt=dt.predict(x_test)
4 print("r2 score",r2_score(y_test,preddt)*100)
5 print("Mean Absolute error",mean_absolute_error(y_test,preddt))
6 print("RMSE",np.sqrt(mean_squared_error(y_test,preddt)))

r2 score 83.4421489702509
Mean Absolute error 25638.263698630137
RMSE 36297.76859320543

In [83]: 1 rfr=RandomForestRegressor()
2 x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.25,random_state=199)
3 rfr.fit(x_train,y_train)
4 predrfr=rfr.predict(x_test)
5 print("r2 score",r2_score(y_test,predrfr)*100)
6 print("Mean Absolute error",mean_absolute_error(y_test,predrfr))
7 print("RMSE",np.sqrt(mean_squared_error(y_test,predrfr)))
```

Here in the above picture we can observe the r2 score of decision tree is 83% at random state 159. RMSE and MSE are given.

## Random Forest



```
Mean Absolute error 25638.263698630137
RMSE 36297.76859320543

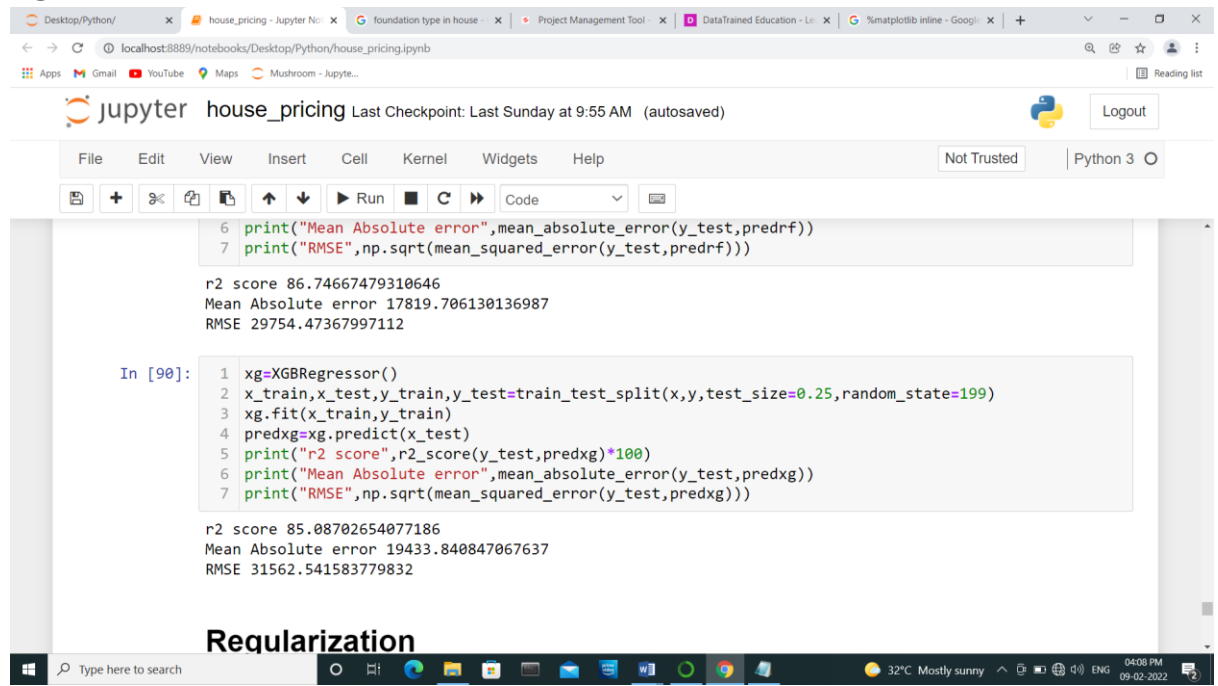
In [83]: 1 rfr=RandomForestRegressor()
2 x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.25,random_state=199)
3 rfr.fit(x_train,y_train)
4 predrfr=rfr.predict(x_test)
5 print("r2 score",r2_score(y_test,predrfr)*100)
6 print("Mean Absolute error",mean_absolute_error(y_test,predrfr))
7 print("RMSE",np.sqrt(mean_squared_error(y_test,predrfr)))

r2 score 86.74667479310646
Mean Absolute error 17819.706130136987
RMSE 29754.47367997112

In [90]: 1 xg=XGBRegressor()
2 x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.25,random_state=199)
3 xg.fit(x_train,y_train)
4 predxg=xg.predict(x_test)
5 print("r2 score",r2_score(y_test,predxg)*100)
6 print("Mean Absolute error",mean_absolute_error(y_test,predxg))
```

Here from the above picture we can observe the r2 score of random forest is 86% at random state 199. RMSE and MSE are given.

## XgBoost

A screenshot of a Jupyter Notebook titled 'house\_pricing' in a web browser. The notebook shows two code cells. The first cell contains two lines of Python code: `print("Mean Absolute error", mean_absolute_error(y_test, predrf))` and `print("RMSE", np.sqrt(mean_squared_error(y_test, predrf)))`. The output for this cell is: `r2 score 86.74667479310646`, `Mean Absolute error 17819.706130136987`, and `RMSE 29754.47367997112`. The second cell starts with `In [90]:` and contains code to create an `XGBRegressor`, split the data, fit the model, and print the `r2 score`, `Mean Absolute error`, and `RMSE`. The output for this cell is: `r2 score 85.08702654077186`, `Mean Absolute error 19433.840847067637`, and `RMSE 31562.541583779832`. The notebook interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help), a toolbar with icons for file operations and execution, and a status bar at the bottom showing the system clock and weather.

```
6 print("Mean Absolute error", mean_absolute_error(y_test, predrf))
7 print("RMSE", np.sqrt(mean_squared_error(y_test, predrf)))

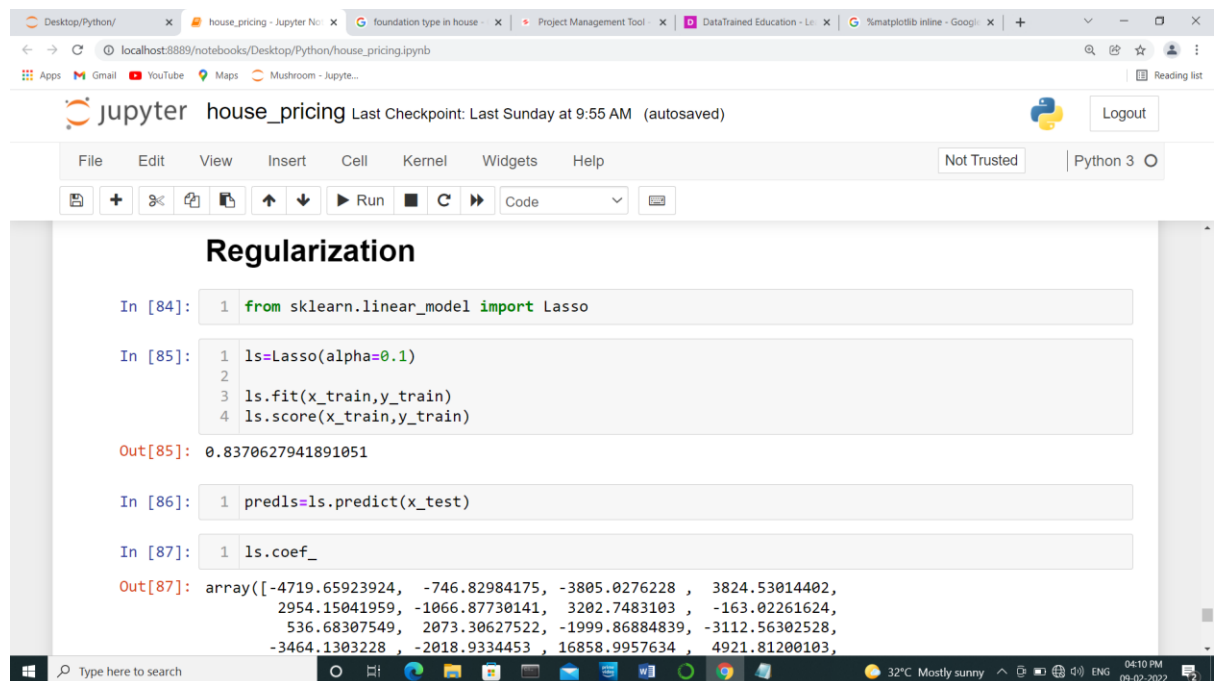
r2 score 86.74667479310646
Mean Absolute error 17819.706130136987
RMSE 29754.47367997112

In [90]: 1 xg=XGBRegressor()
2 x_train, x_test, y_train, y_test=train_test_split(x, y, test_size=0.25, random_state=199)
3 xg.fit(x_train, y_train)
4 predxg=xg.predict(x_test)
5 print("r2 score", r2_score(y_test, predxg)*100)
6 print("Mean Absolute error", mean_absolute_error(y_test, predxg))
7 print("RMSE", np.sqrt(mean_squared_error(y_test, predxg)))

r2 score 85.08702654077186
Mean Absolute error 19433.840847067637
RMSE 31562.541583779832
```

Here we can observe that the r2 score of xg boost is 85% at random state 199. RMSE and MSE are also given.

## Lasso

A screenshot of a Jupyter Notebook titled 'house\_pricing' in a web browser. The notebook shows three code cells. The first cell contains `from sklearn.linear_model import Lasso`. The second cell starts with `In [85]:` and contains code to create a `Lasso` object with `alpha=0.1`, fit it to the training data, and print the `score`. The output for this cell is `Out[85]: 0.8370627941891051`. The third cell starts with `In [86]:` and contains `predls=ls.predict(x_test)`. The fourth cell starts with `In [87]:` and contains `ls.coef_`. The output for this cell is a large array of coefficients: `Out[87]: array([-4719.65923924, -746.82984175, -3805.0276228, 3824.53014402, 2954.15041959, -1066.87730141, 3202.7483103, -163.02261624, 536.68307549, 2073.30627522, -1999.86884839, -3112.56302528, -3464.1303228, -2018.9334453, 16858.9957634, 4921.81200103, ...])`. The notebook interface is the same as the previous one.

```
Regularization

In [84]: 1 from sklearn.linear_model import Lasso

In [85]: 1 ls=Lasso(alpha=0.1)
2
3 ls.fit(x_train, y_train)
4 ls.score(x_train, y_train)

Out[85]: 0.8370627941891051

In [86]: 1 predls=ls.predict(x_test)

In [87]: 1 ls.coef_

Out[87]: array([-4719.65923924, -746.82984175, -3805.0276228, 3824.53014402,
2954.15041959, -1066.87730141, 3202.7483103, -163.02261624,
536.68307549, 2073.30627522, -1999.86884839, -3112.56302528,
-3464.1303228, -2018.9334453, 16858.9957634, 4921.81200103,
```

As we know Lasso is a regularization technique. Here we can observe that alpha 0.1 the score is 83%.

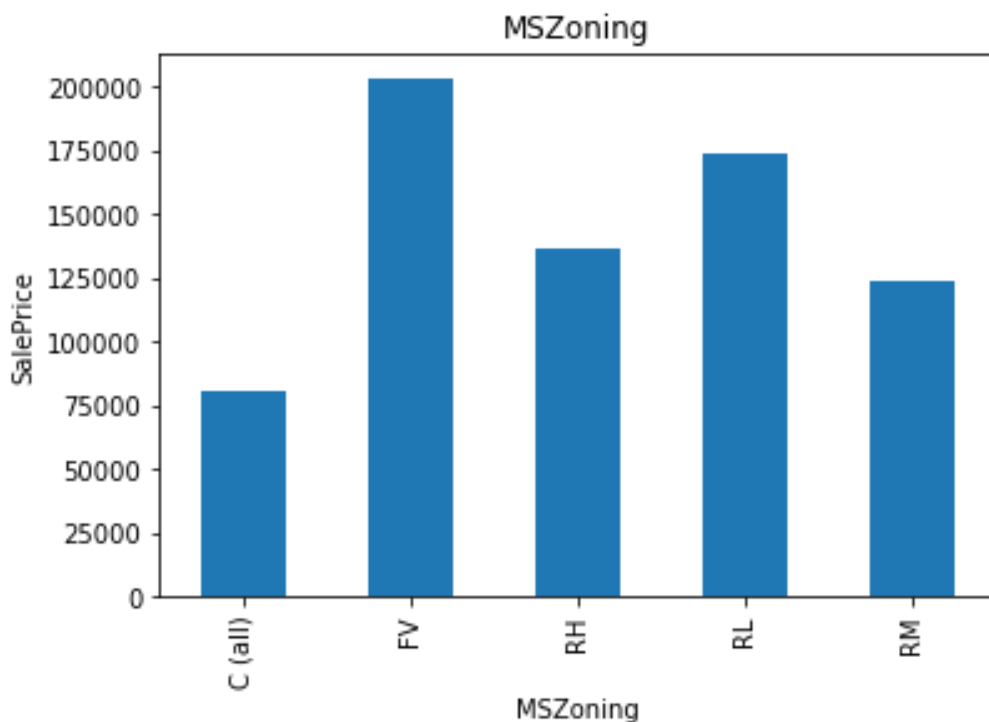
- Key Metrics for success in solving problem under consideration

R<sup>2</sup> score, Mean absolute error (MSE) and Root Mean Squared error (RMSE) were used in this project. Here the R<sup>2</sup> score of the project using random forest regression is 86%. As we know the Mean squared error told us that how close the regression line with set of points. Least difference makes a good regression line. While RMSE is the standard deviation of the residuals or errors. These are the key metrics used in this project.

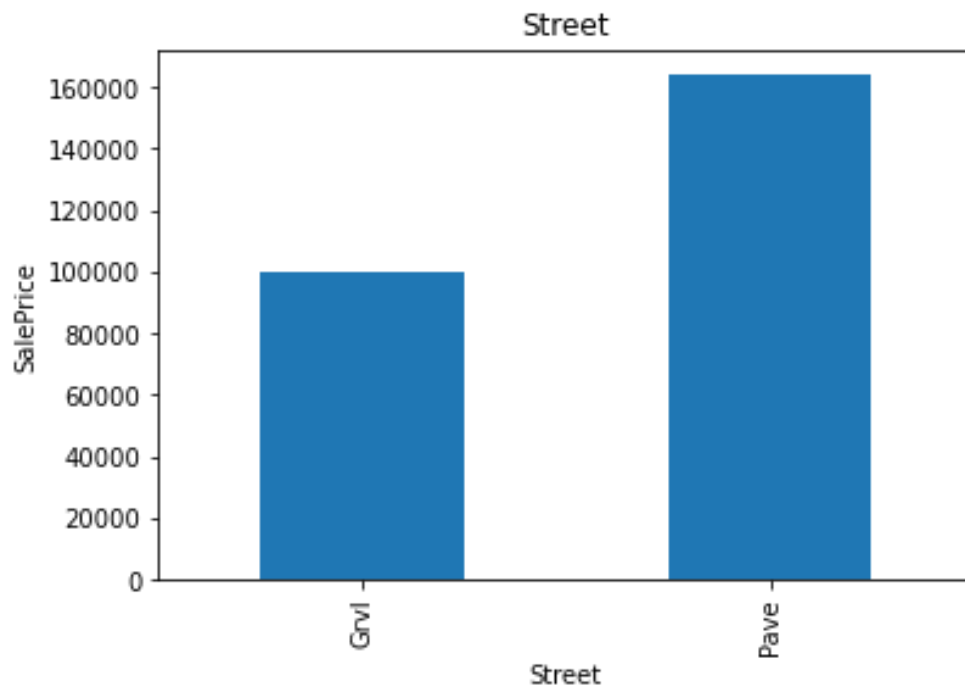
- Visualizations

Visualization is one of the best tool for understanding and analysing the data. Here in this project I analyse the categorical and numerical features separately.

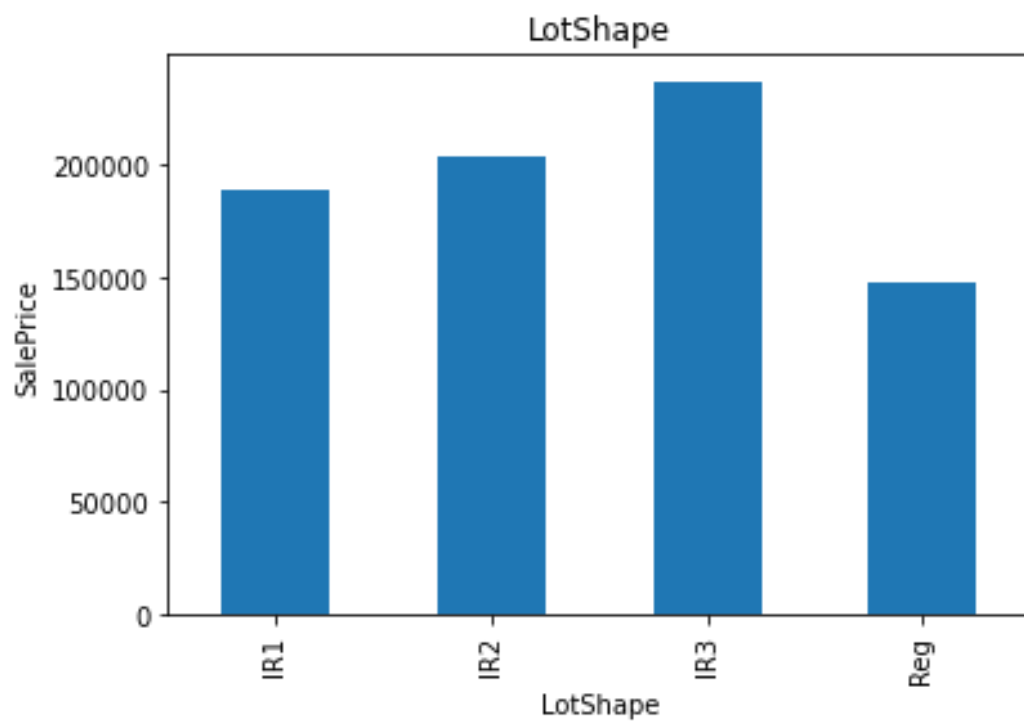
So first of all let's analyse categorical features with the dependent variable.



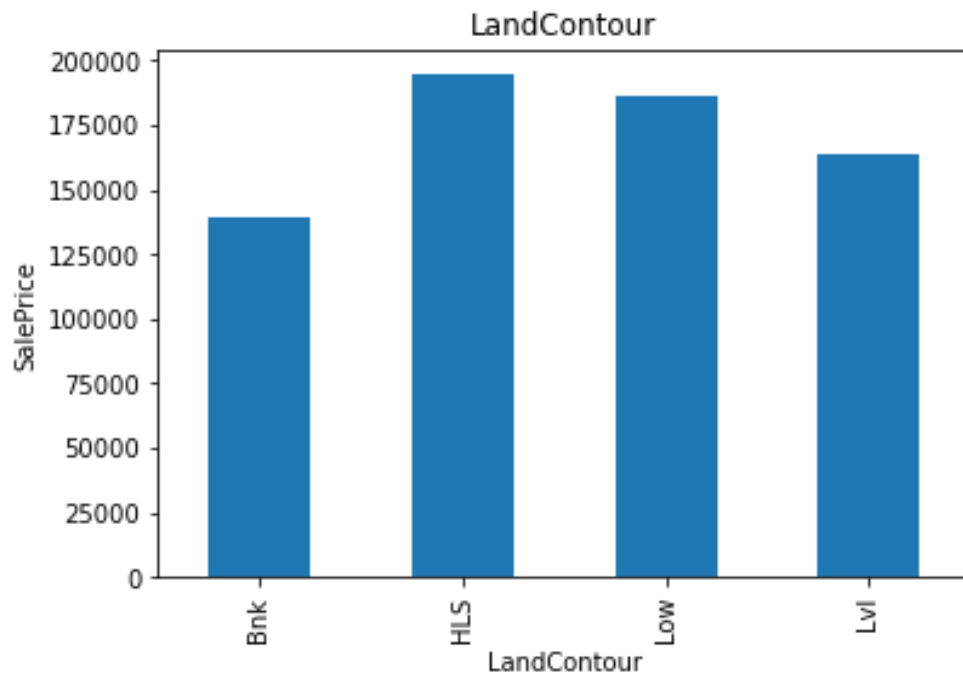
The house which is in floating village residential have high prices.



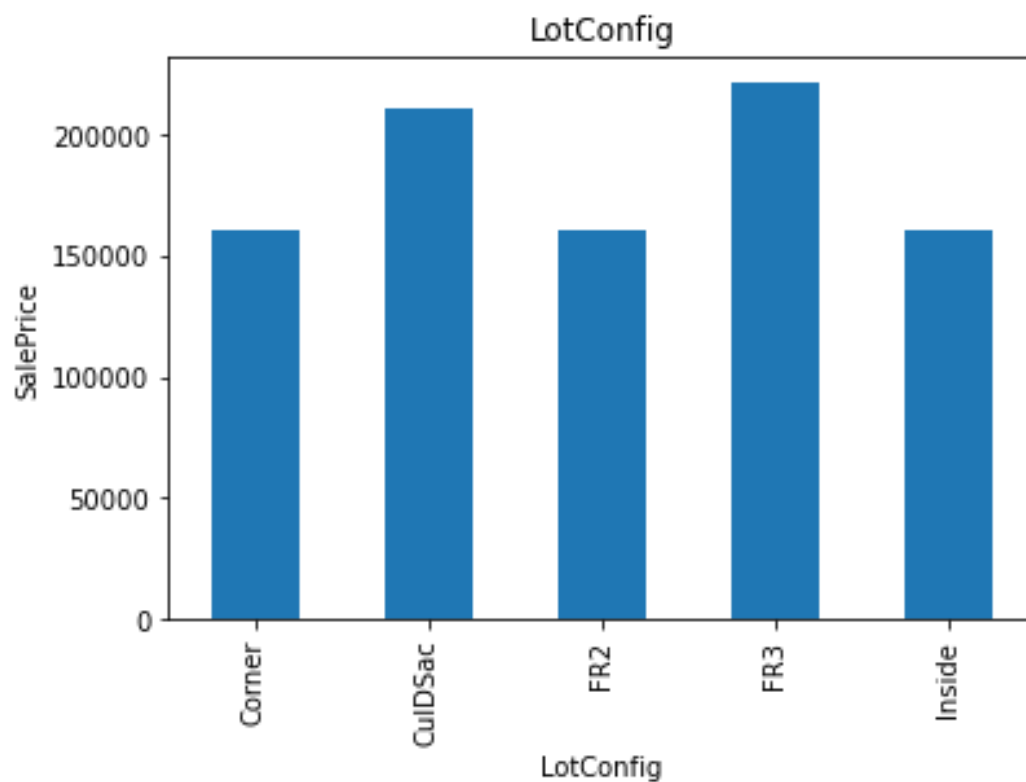
Pave road access property have high prices.



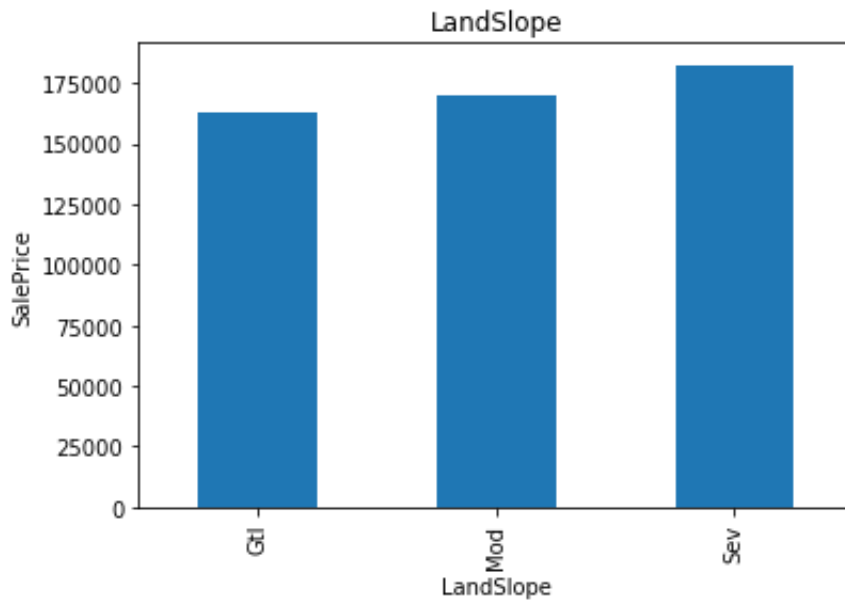
Shape of property, here we can observe that irregular shape of property have high prices.



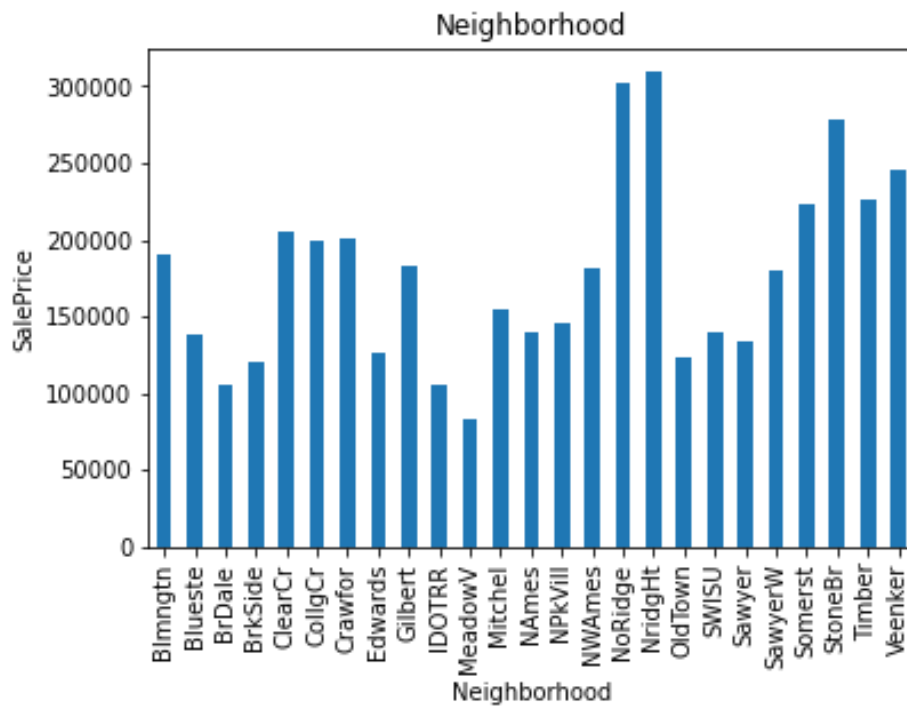
Fatness of property we can observe that hill side that is significant slope from side to side to have high prices than others.



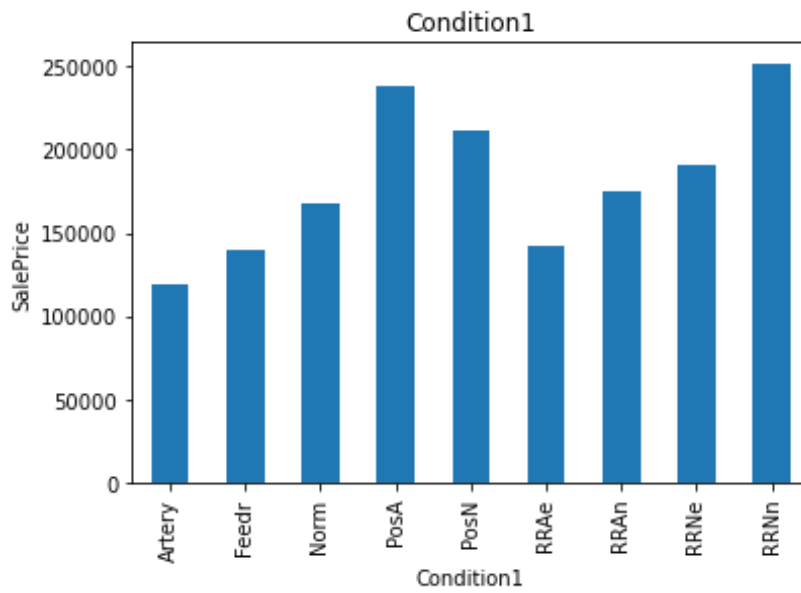
The configuration of lot Frontage on 3 sides of property have high prices.



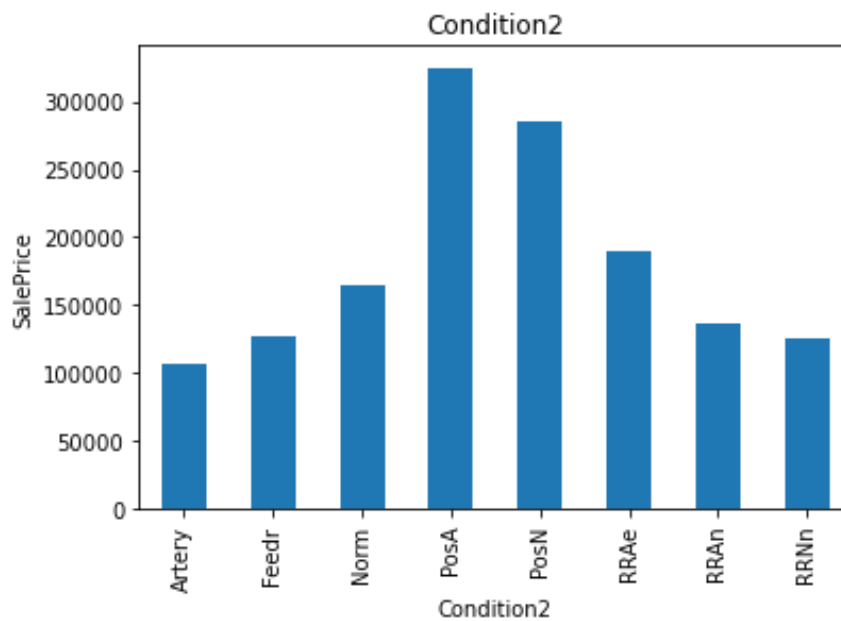
The property which have severe slope have high prices.



Here we can observe that Northridge Heights have higher prices.

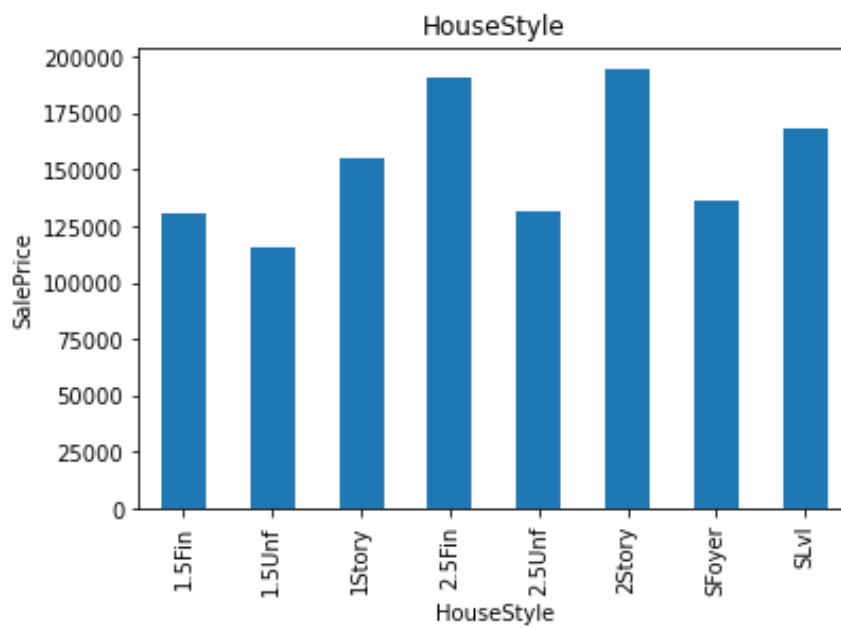
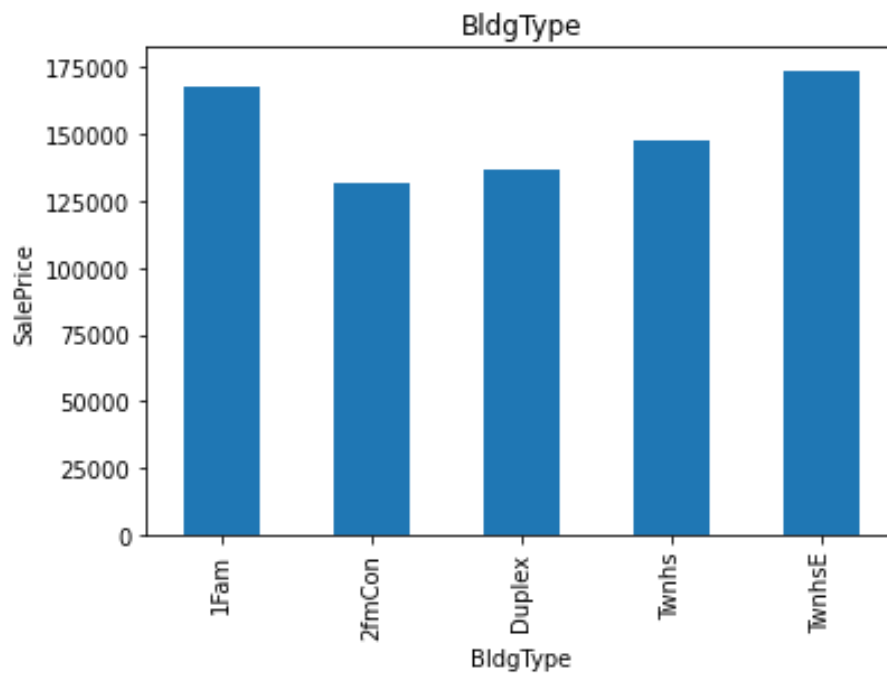


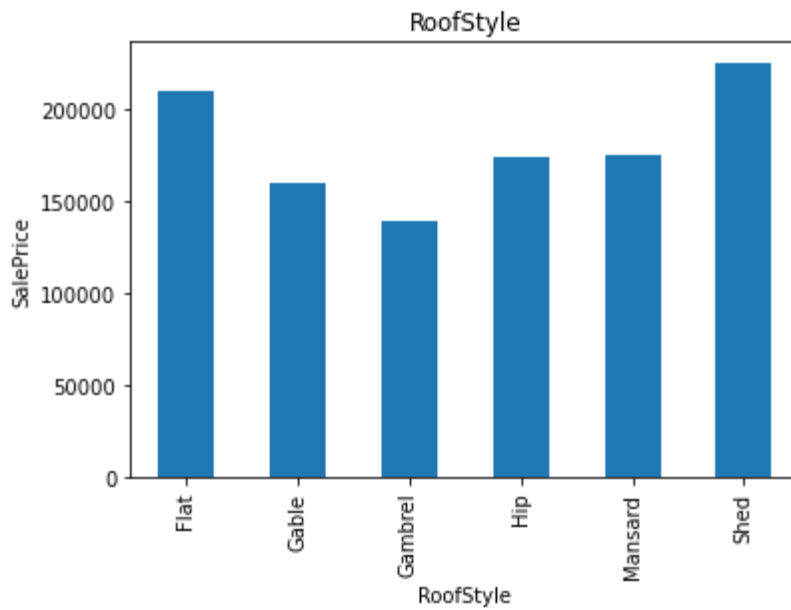
Within 200' of North-South Railroad proximity condition have high prices.



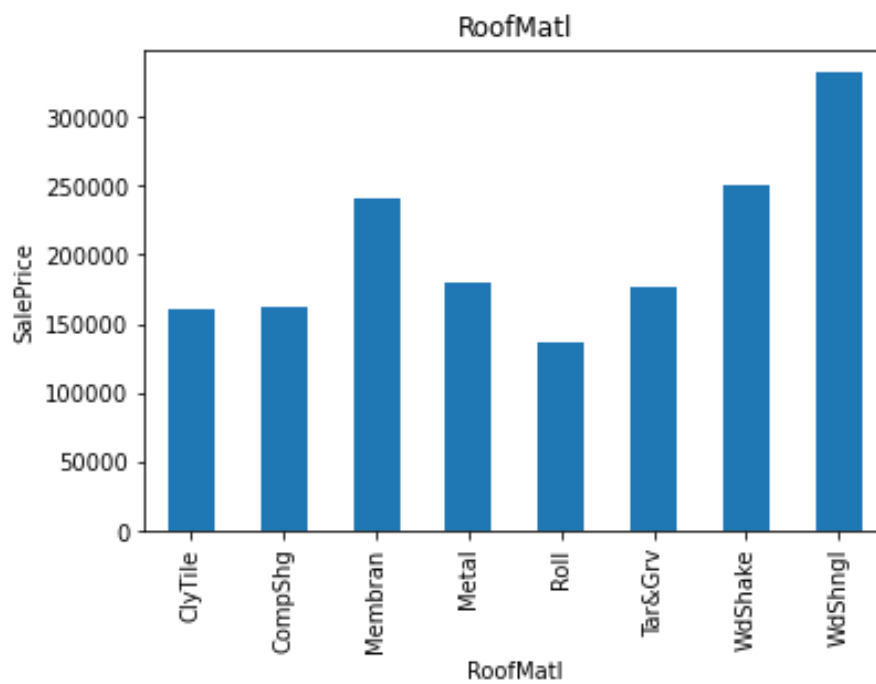
Adjacent to positive off-site feature have high prices.



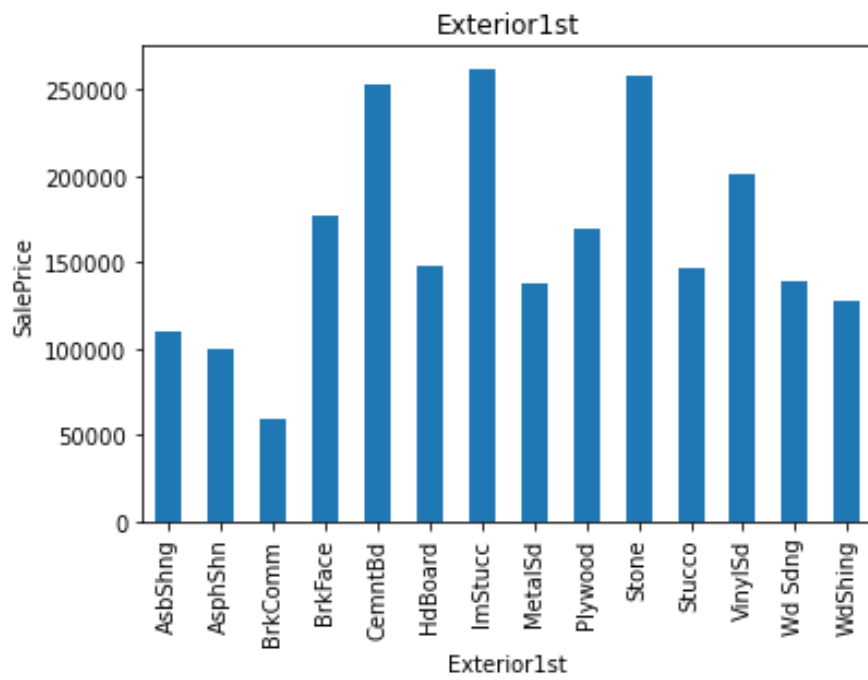




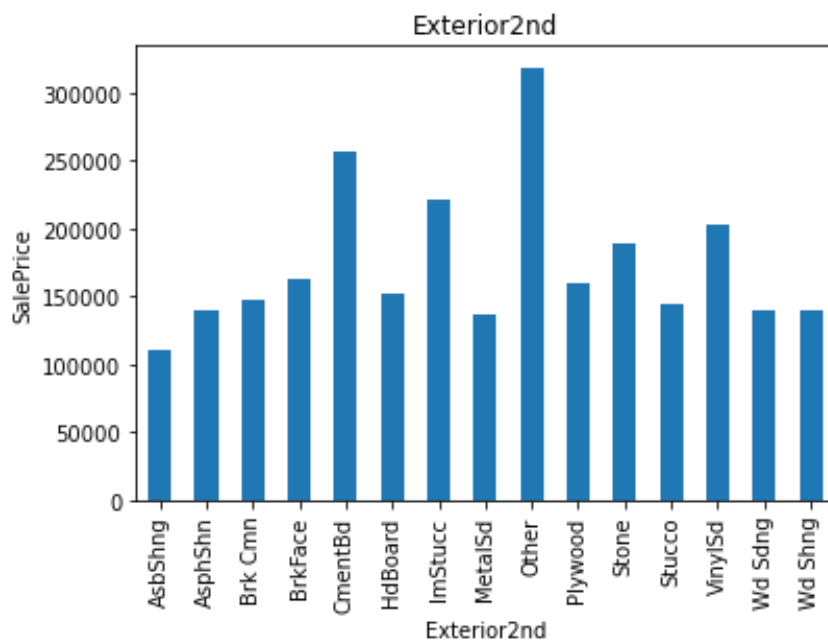
Shed roof style have high prices.

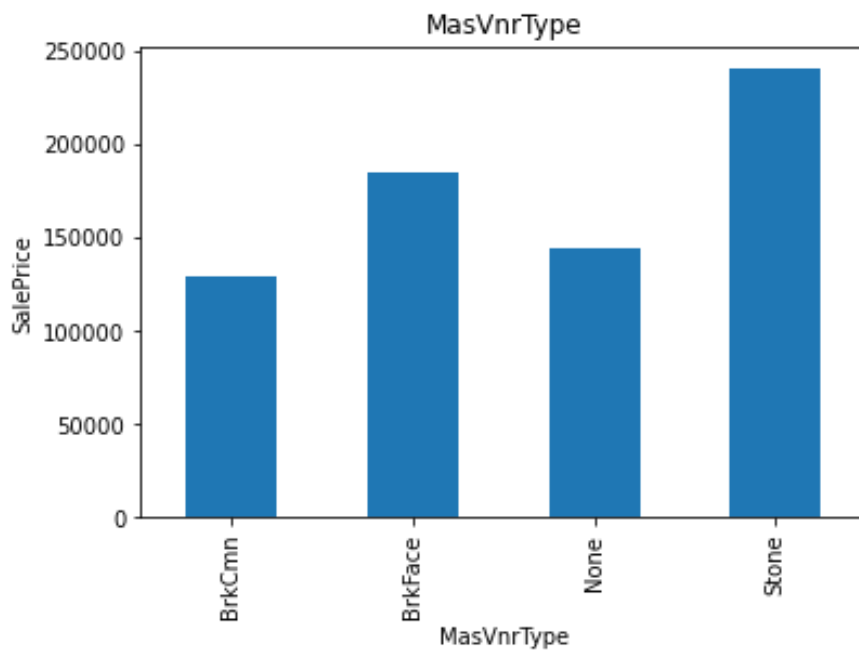


Wood shingles material have high prices.

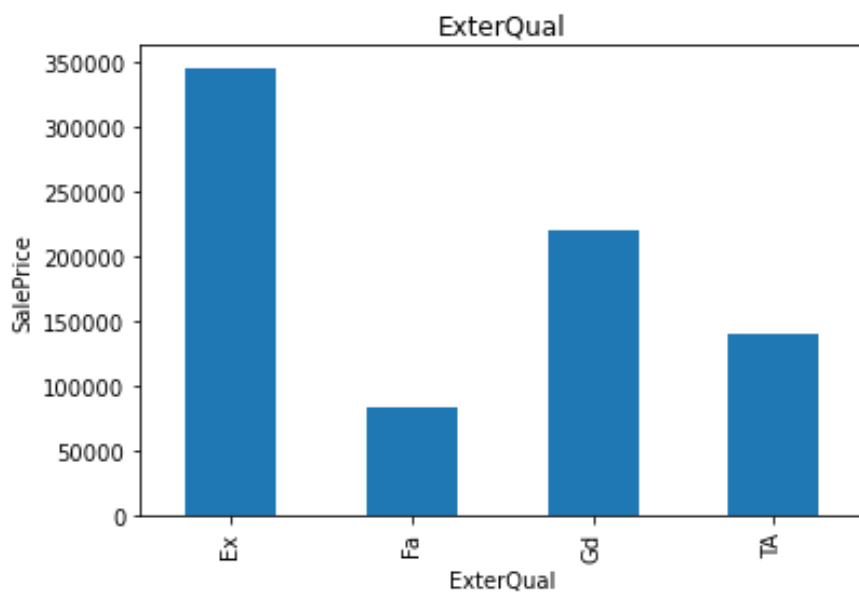


Imitation Stucco, Stone exterior covering have high prices.

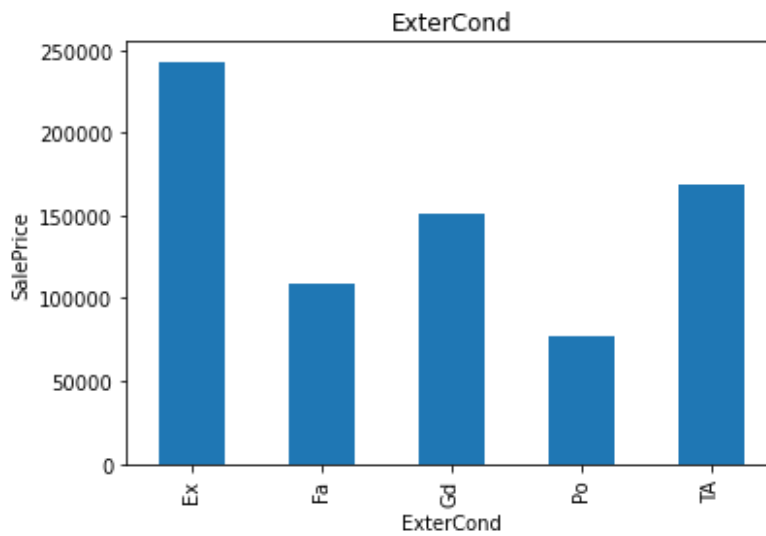




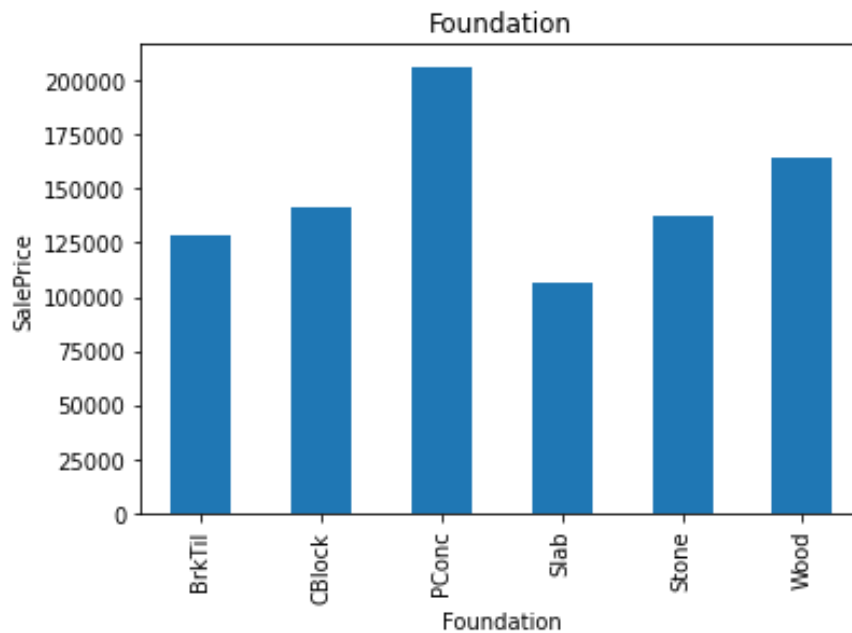
The houses with stone Masonry veneer type have high demand and price.



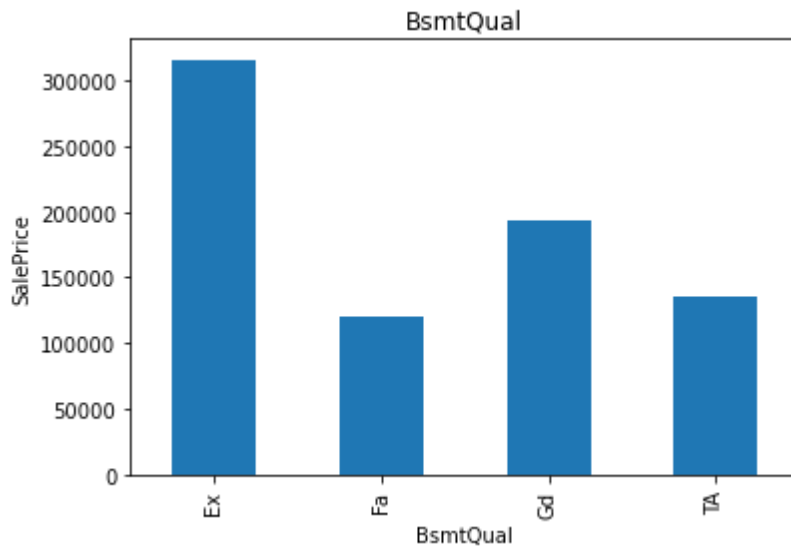
Excellent quality of material have high prices.



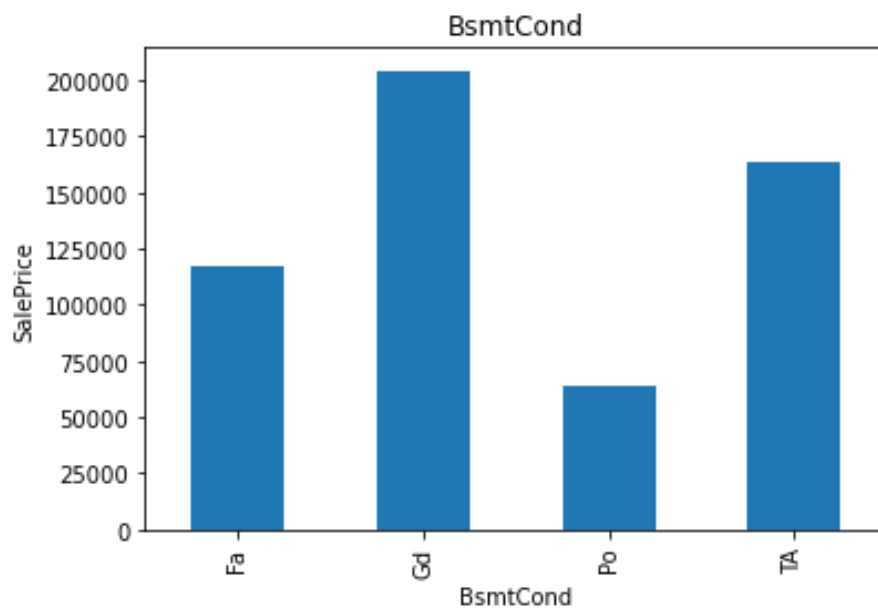
Also excellent exterior condition demands high prices.



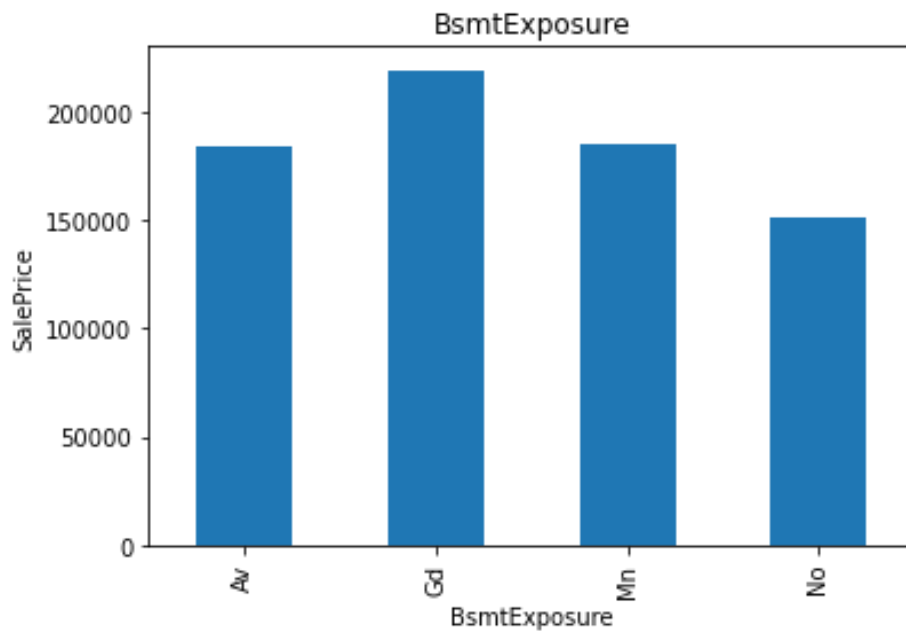
The house with poured concrete have high demands high prices.



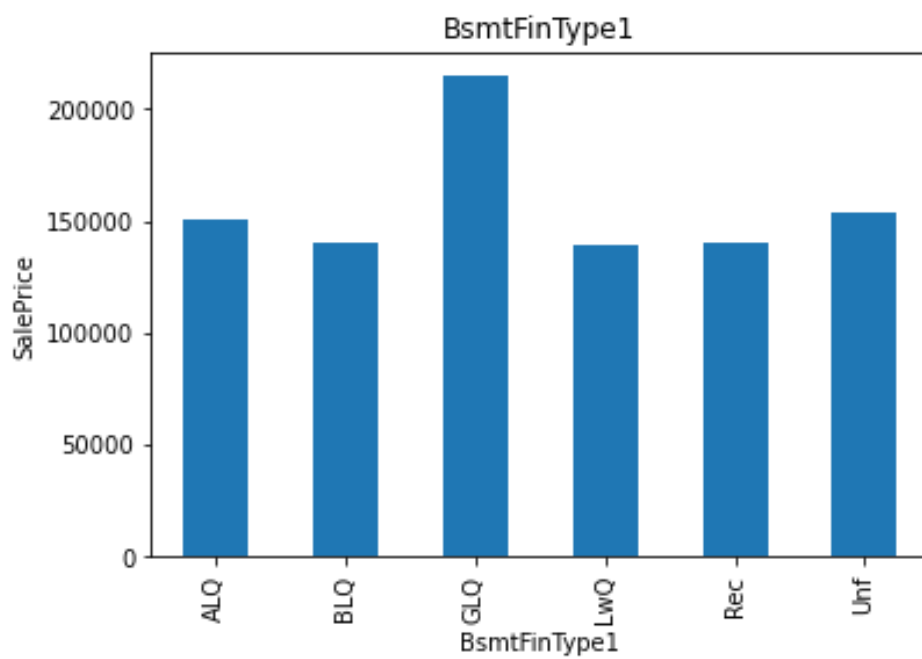
It evaluates height of the basement here excellent basement that is 100+ inches basements have high prices.



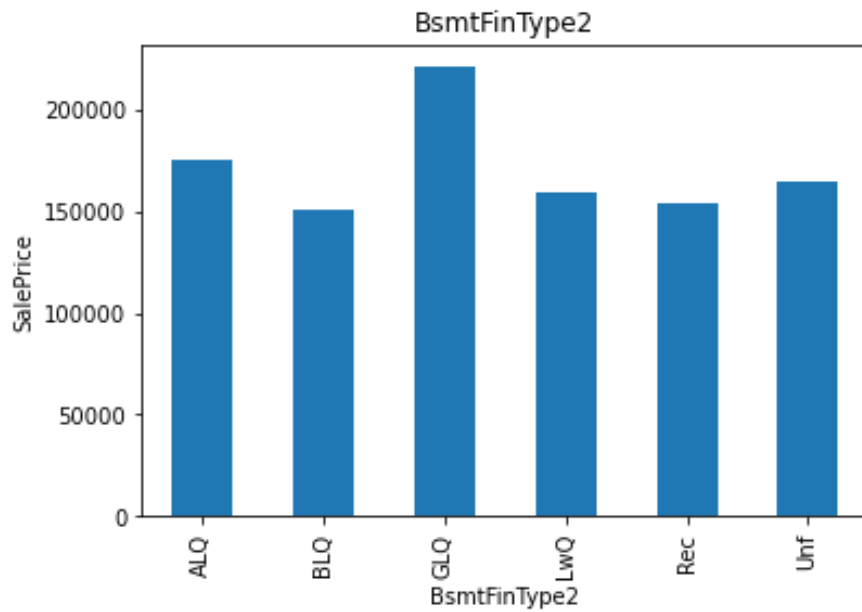
Condition of the basement at the time of selling. Here good condition basement have high prices.



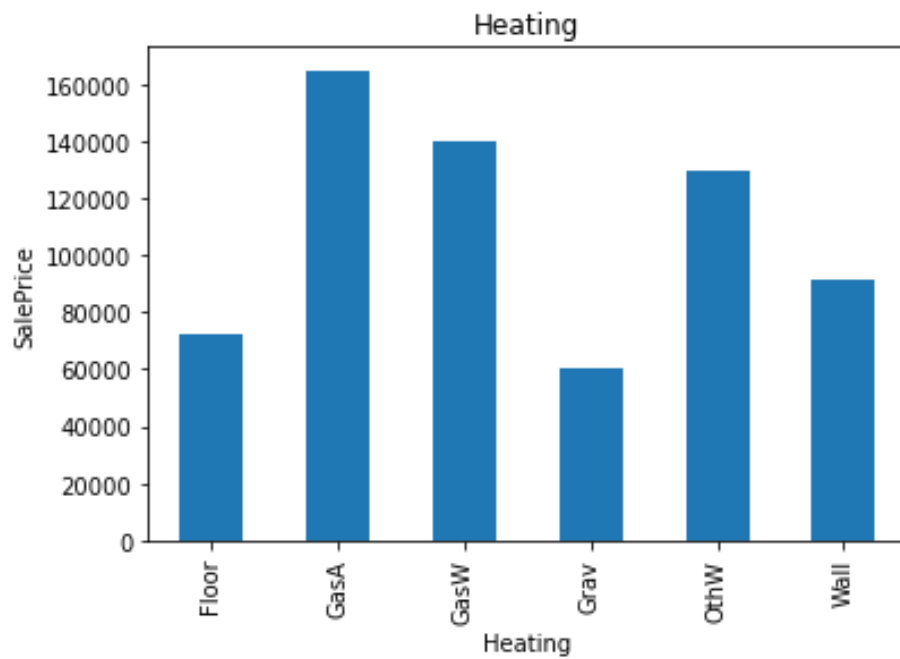
Good basement exposure have best prices.



Rating of basement finished area with good living quarters have high prices.

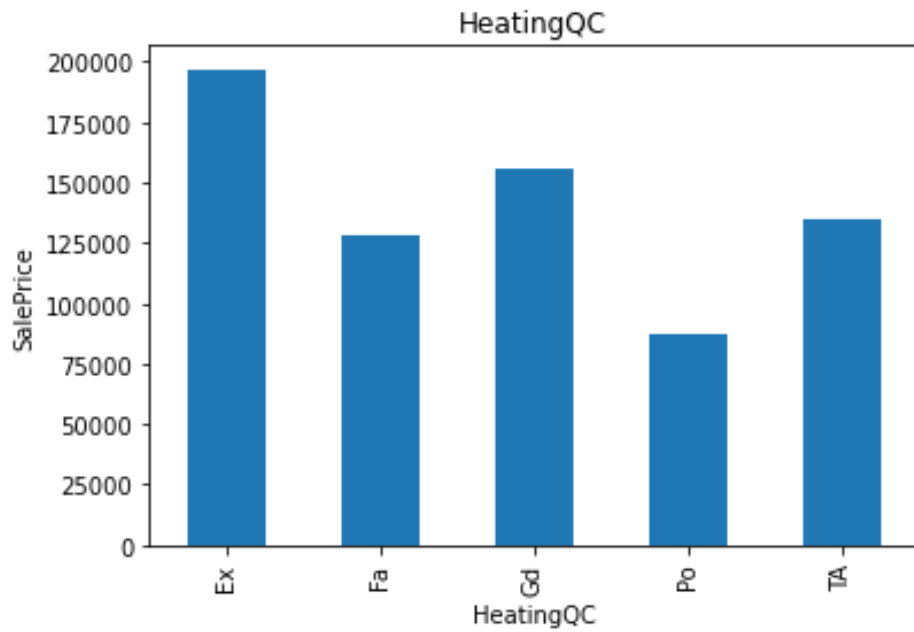


Rating of basement finished area (if multiple types) with good living quarters have high prices.

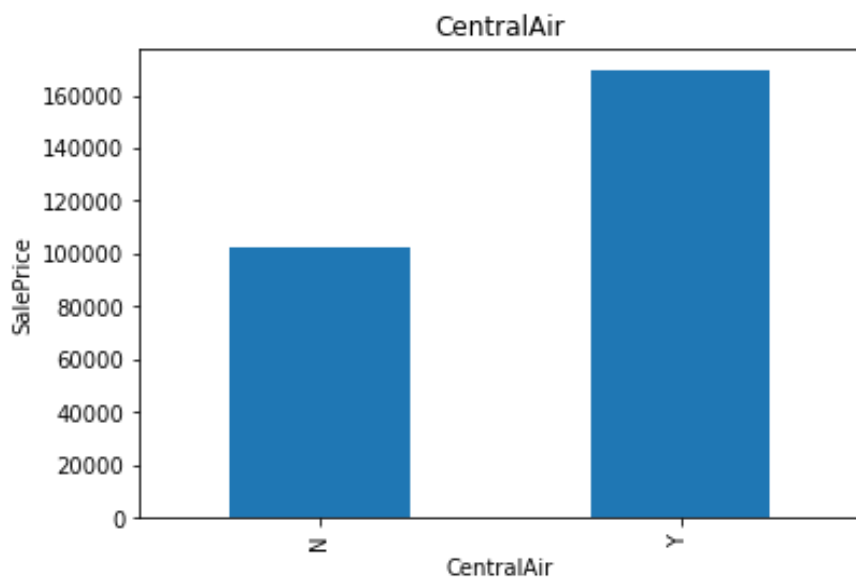


Gas forced warm air furnace type of heating have high prices.

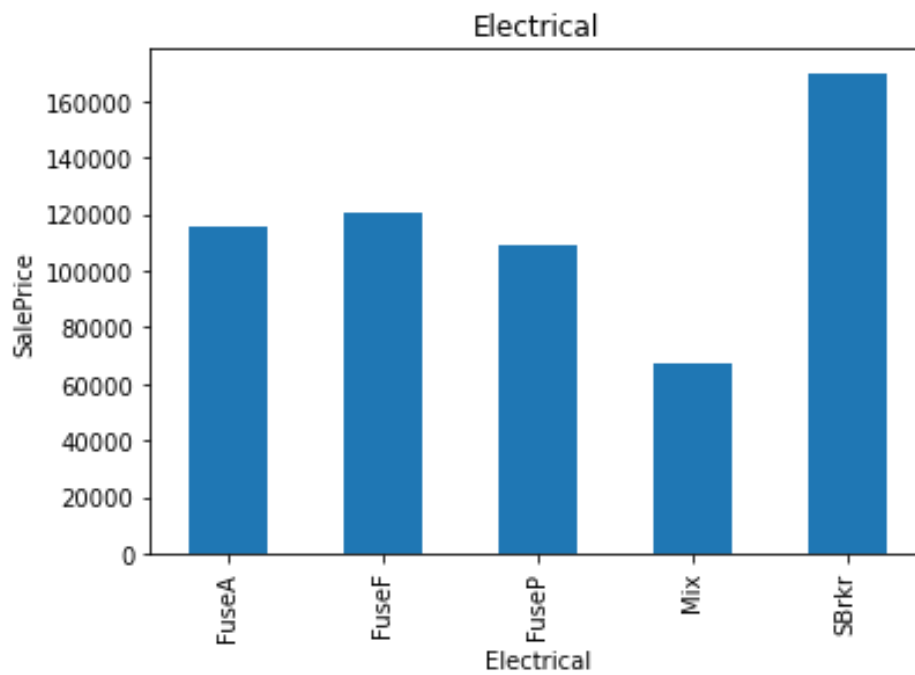




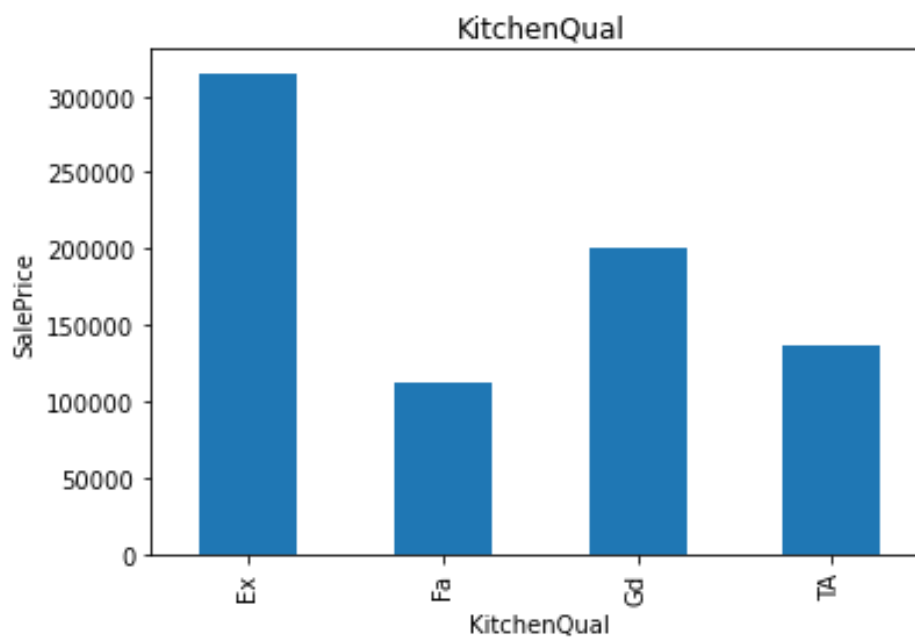
Excellent heating quality and condition have high prices.



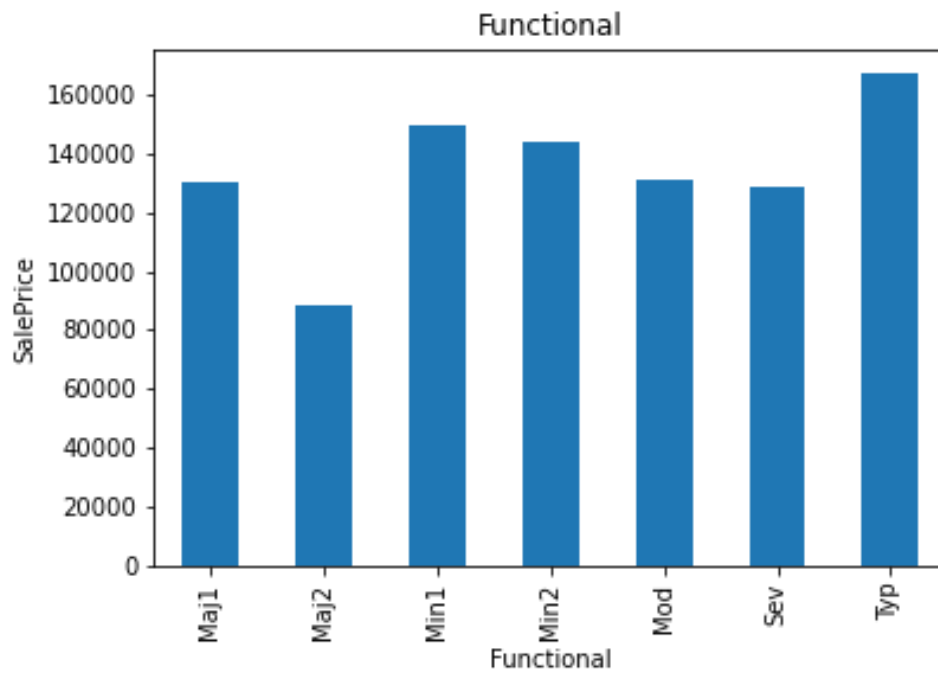
The houses with central air conditioning have high prices.



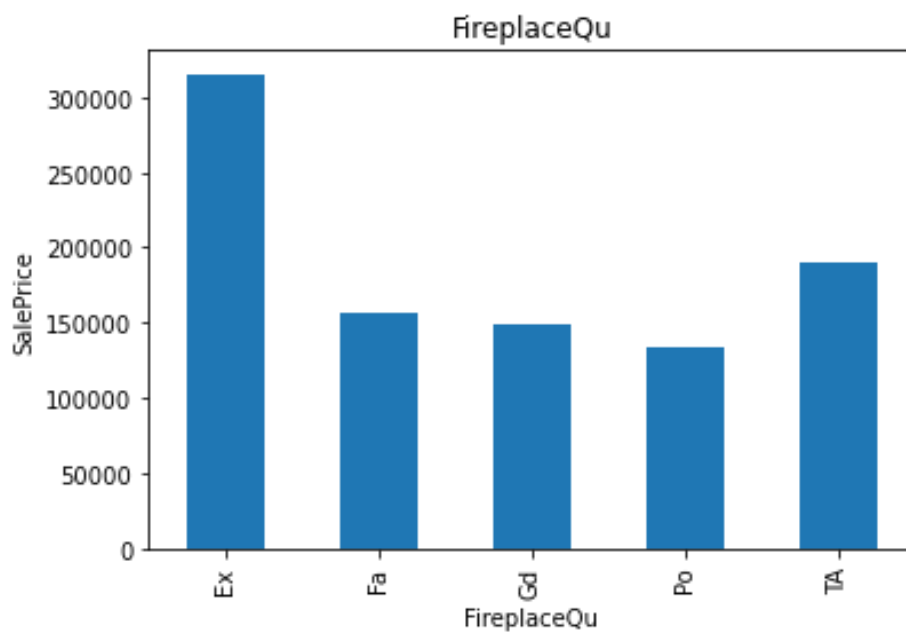
The houses with Standard Circuit Breakers and roblox electrical system have high prices.



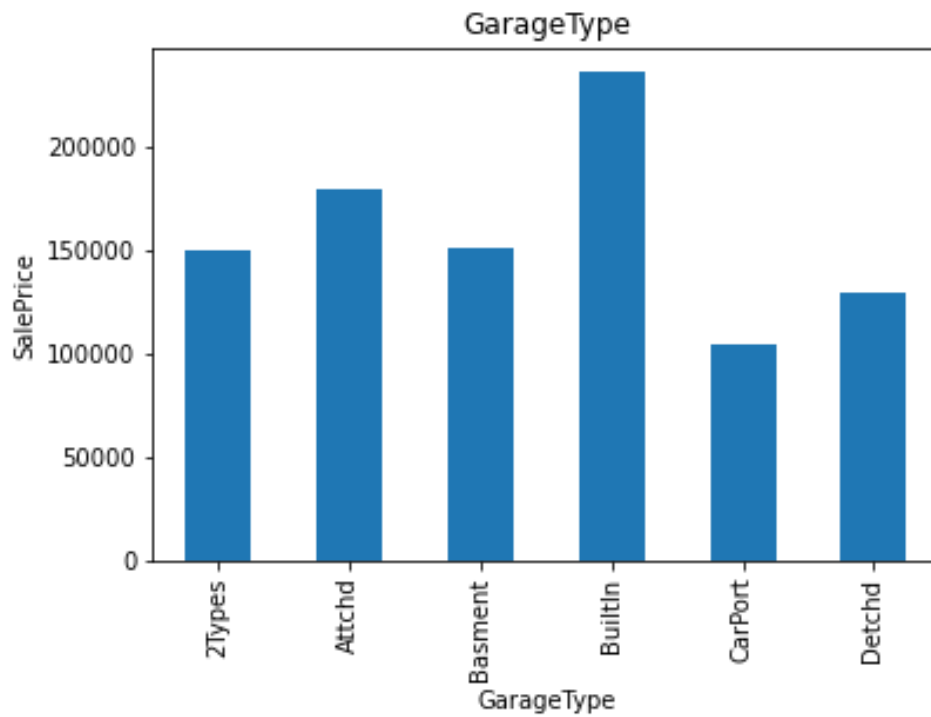
Excellent kitchen quality have high prices.



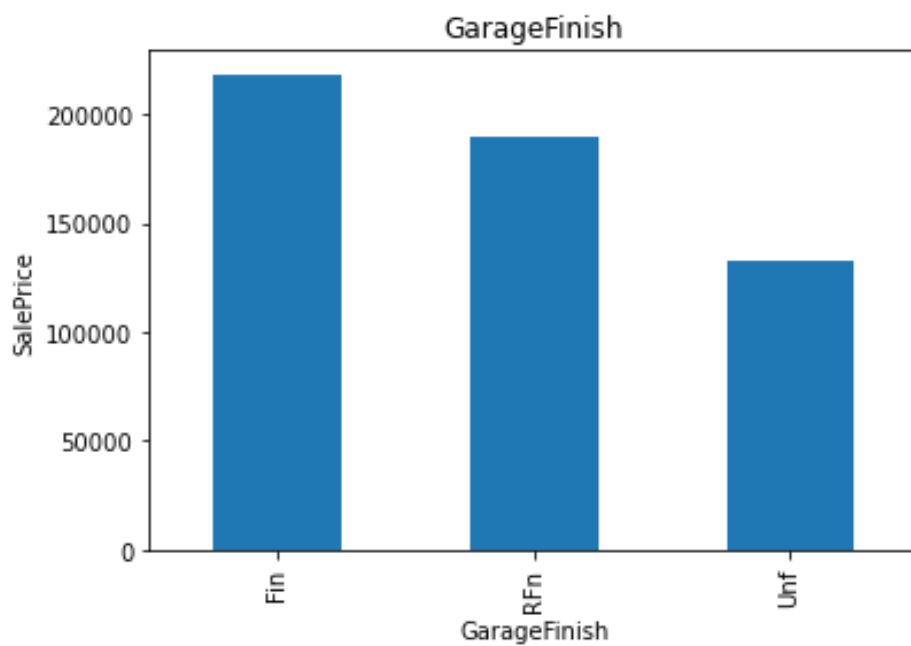
Typical functionality houses are high prices.



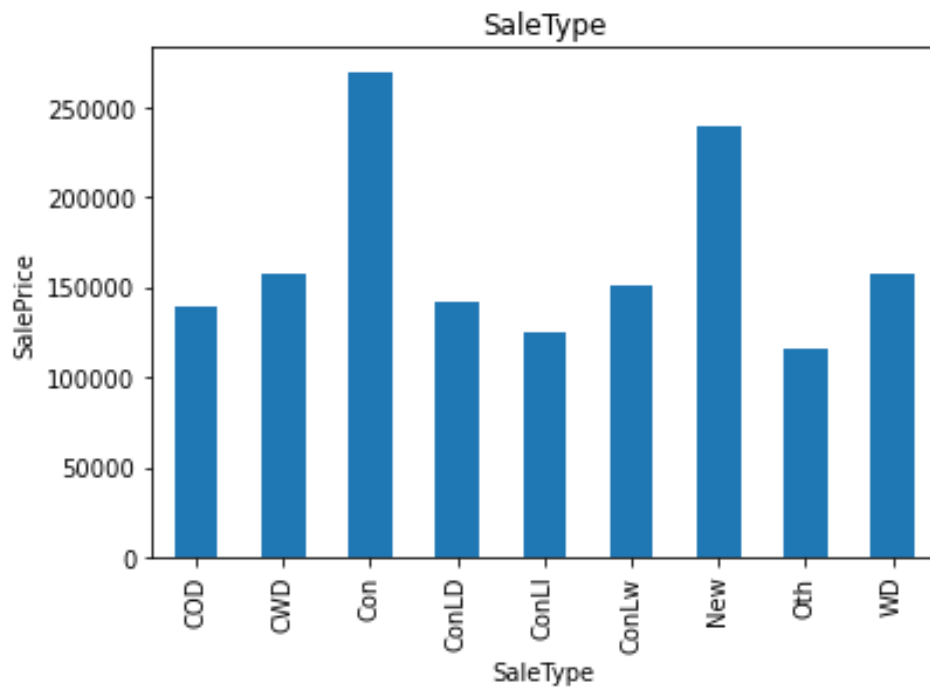
Excellent - Exceptional Masonry Fireplace have high prices.



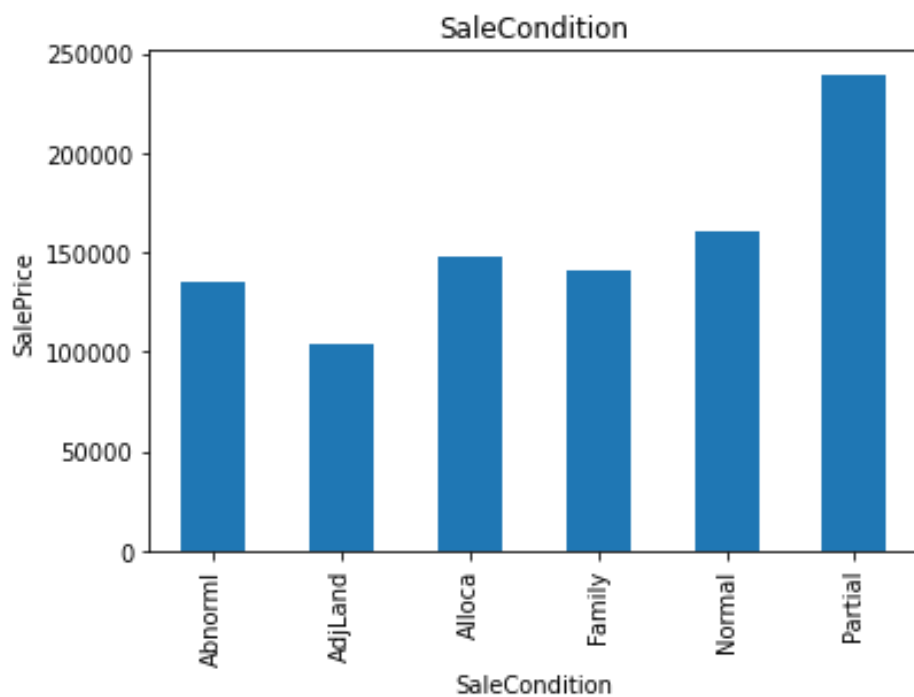
Built-In (Garage part of house - typically has room above garage) have high prices.



Finished interior finished garage have high prices.

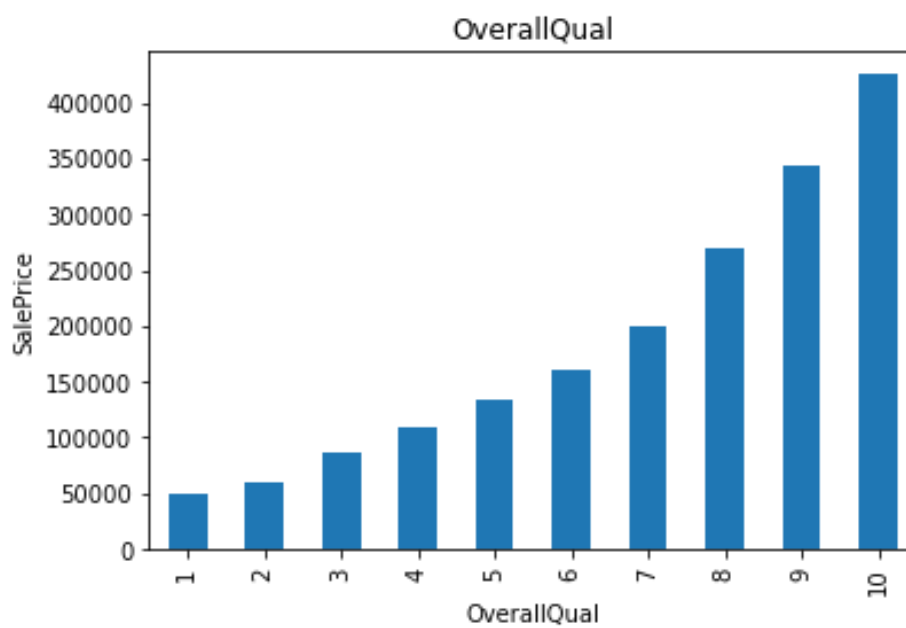
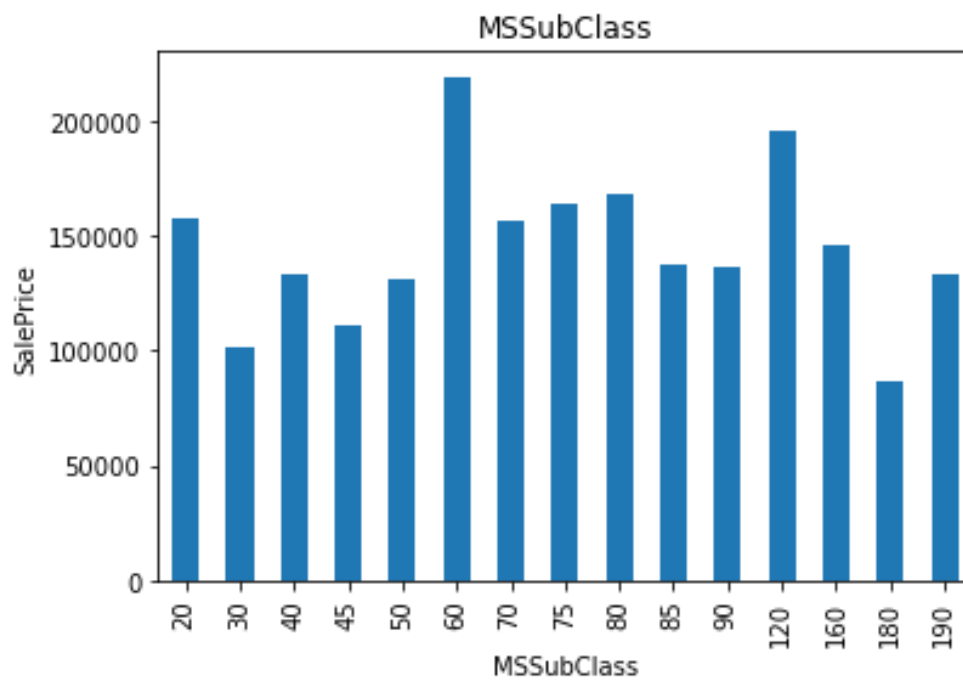


Sal type contract 15% down payment regular term have high prices.

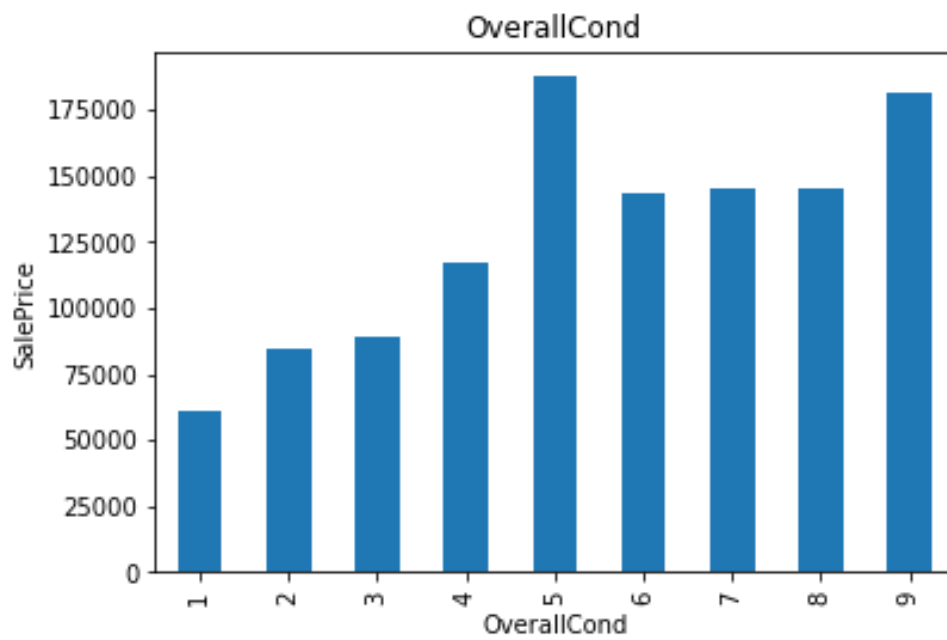


The partial work completed home or a new house have high prices.

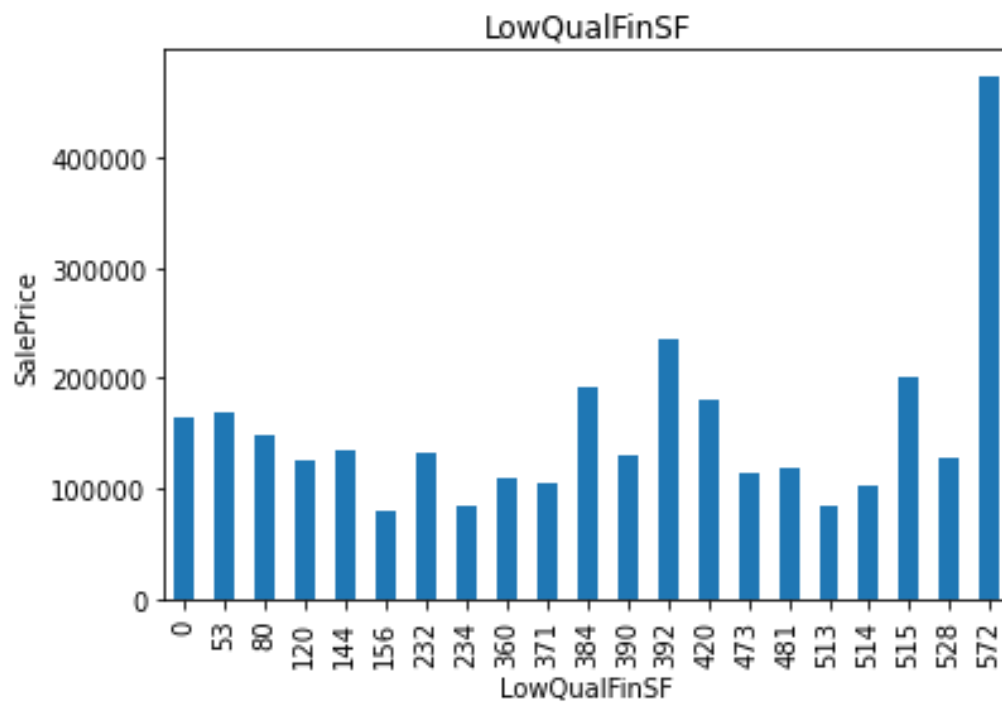
Now let's analyse the numerical features



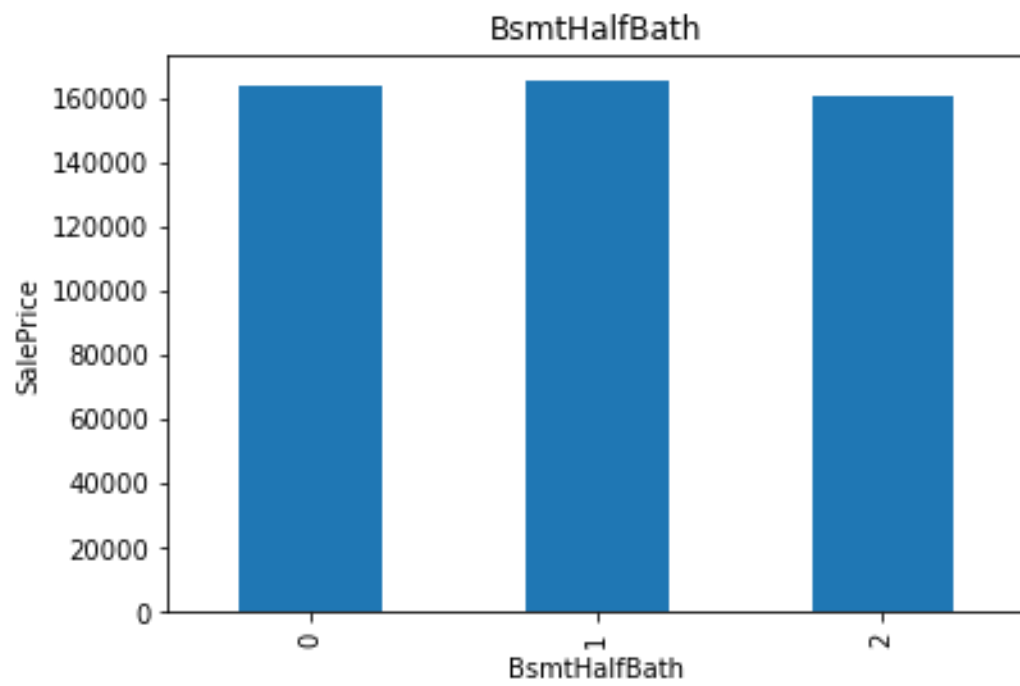
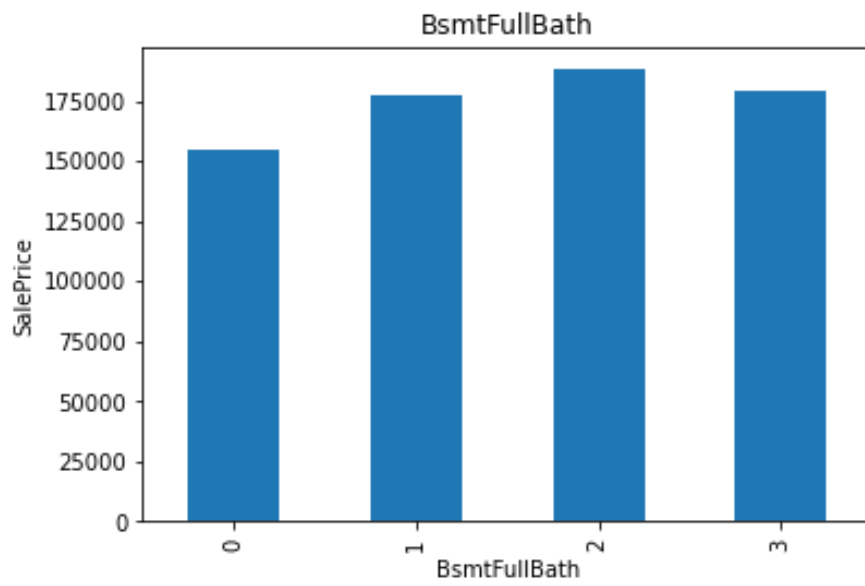
Overall quality 10 have high prices.



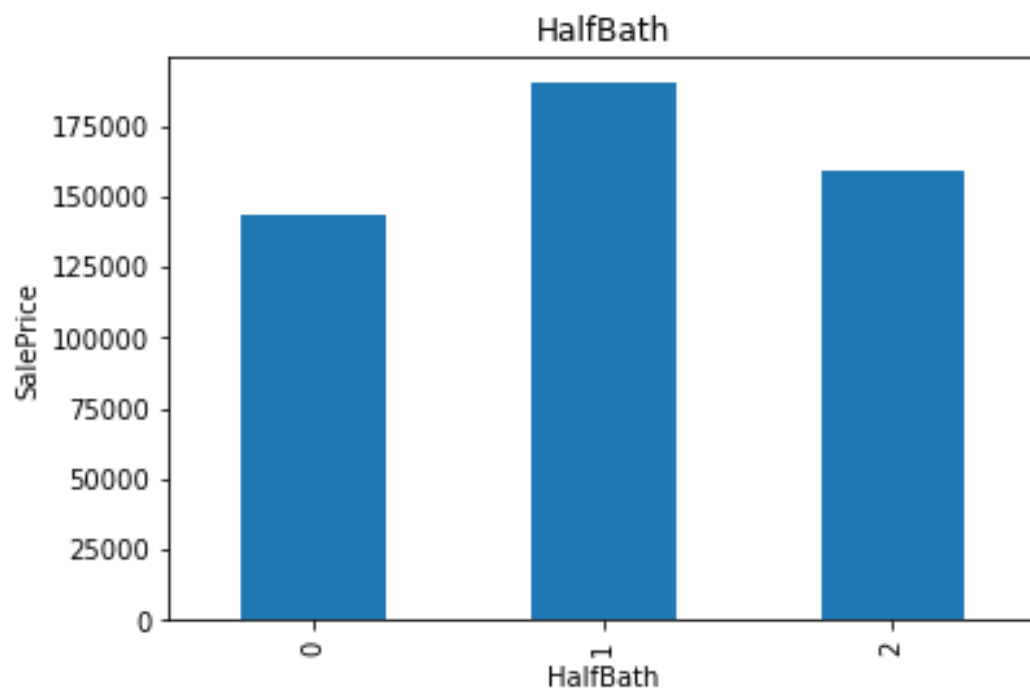
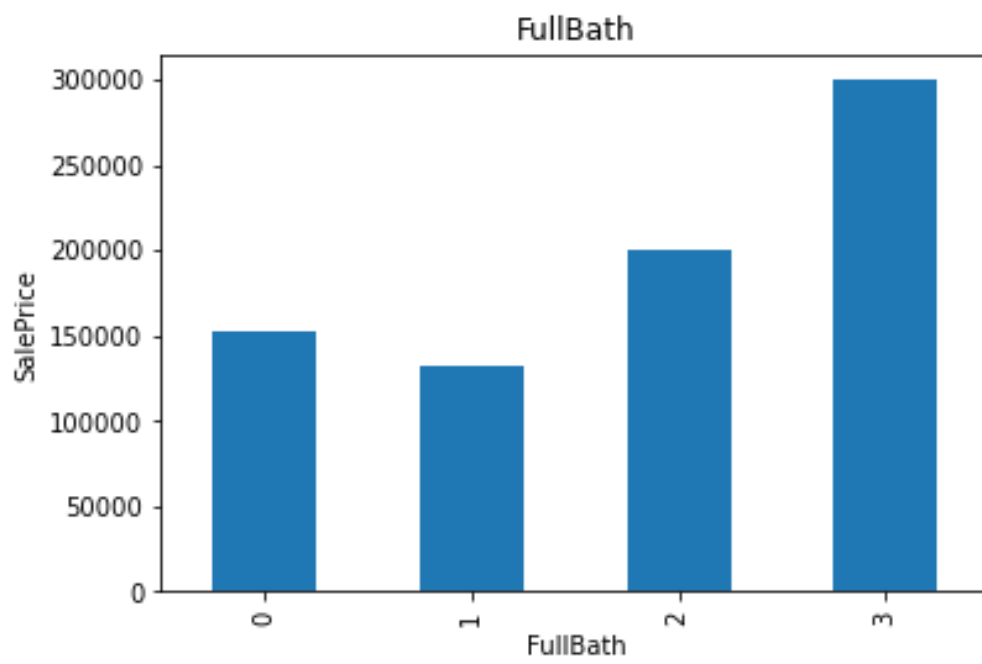
Overall condition of the house.

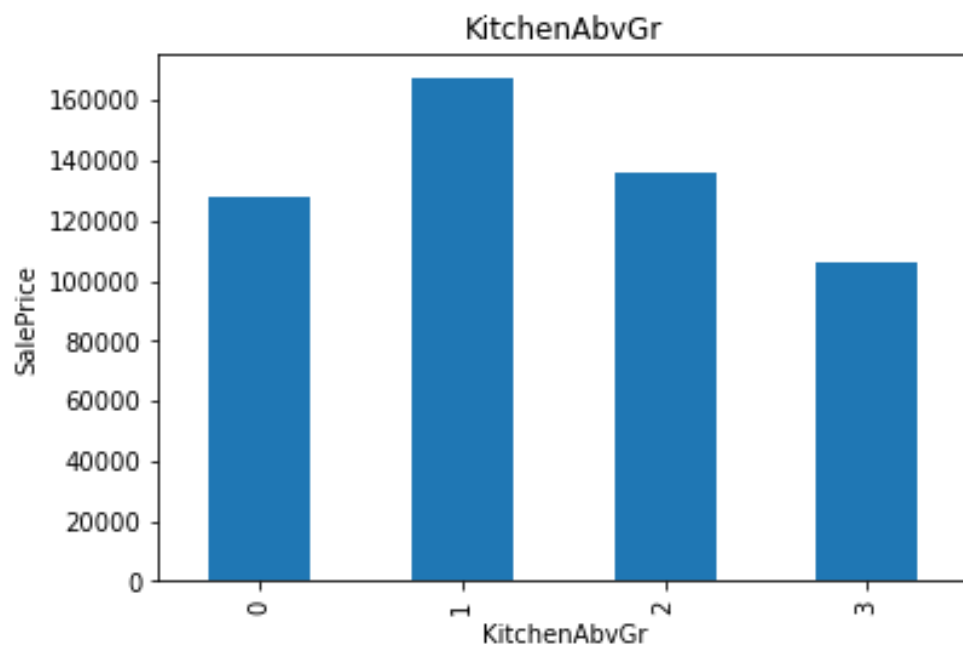
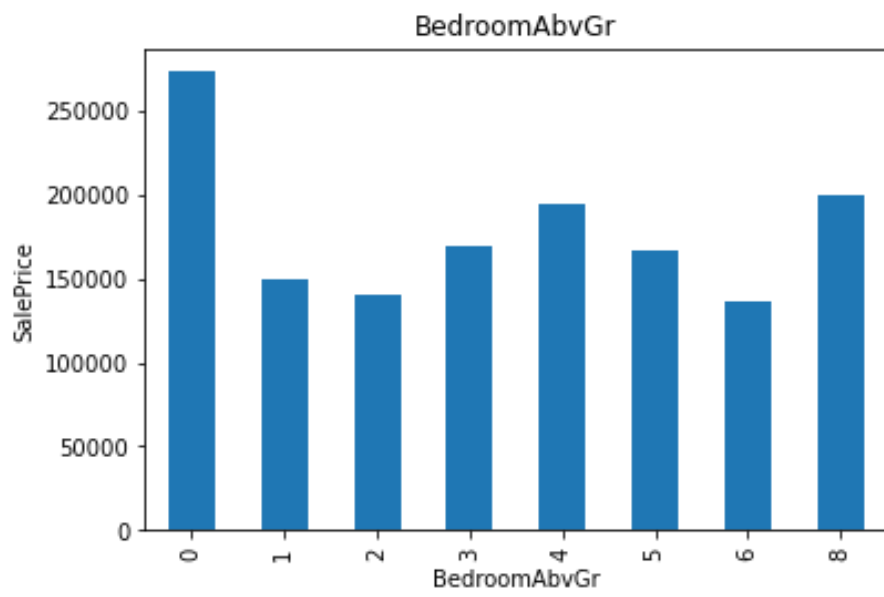


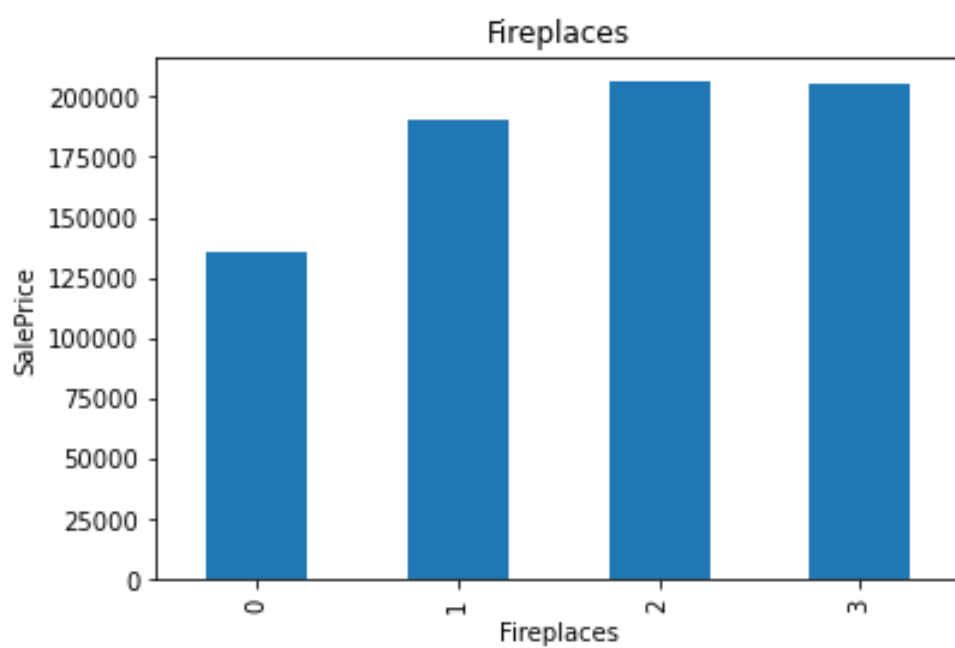
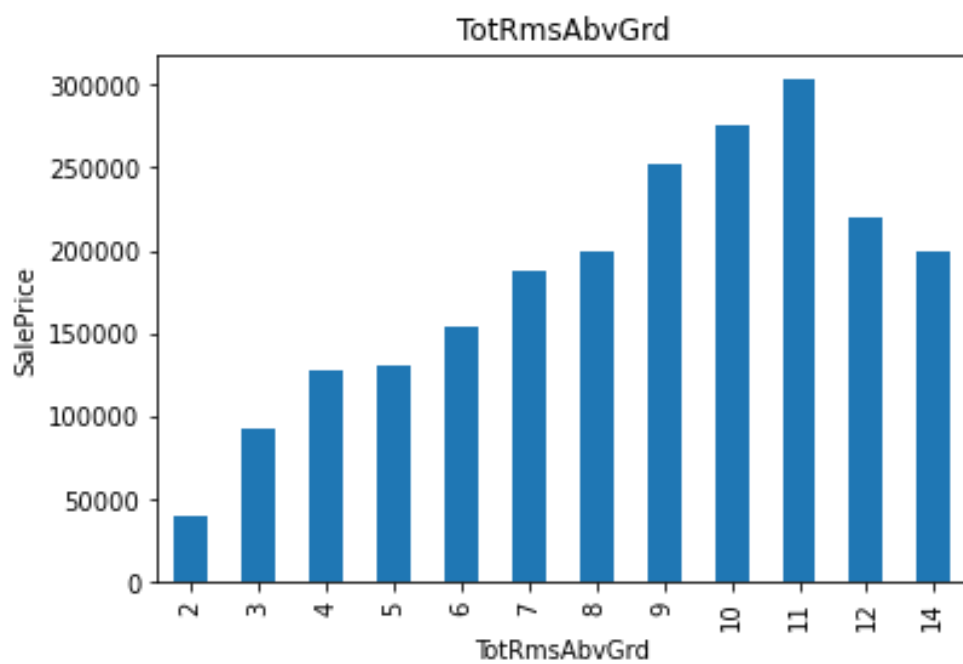
Low quality finished square feet.

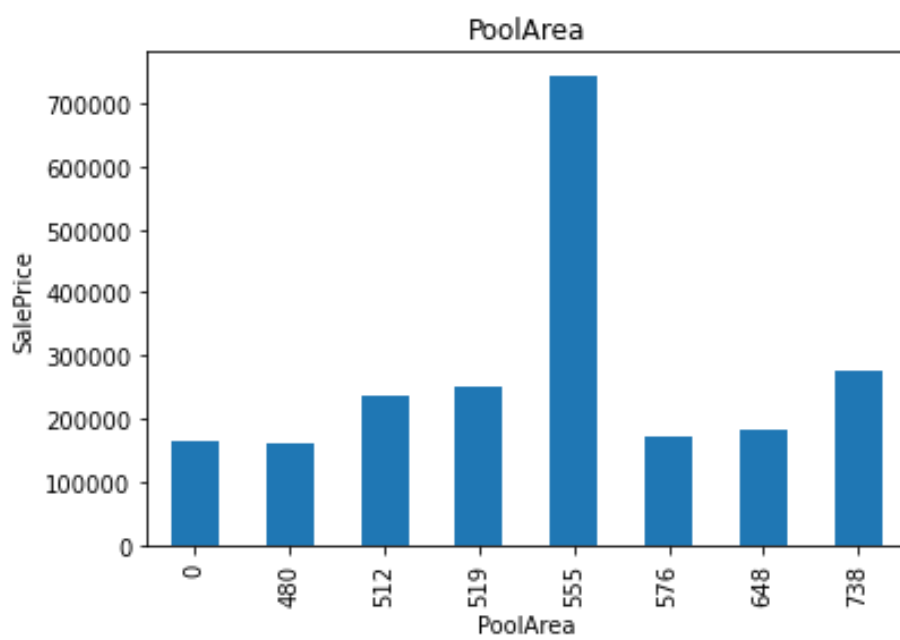
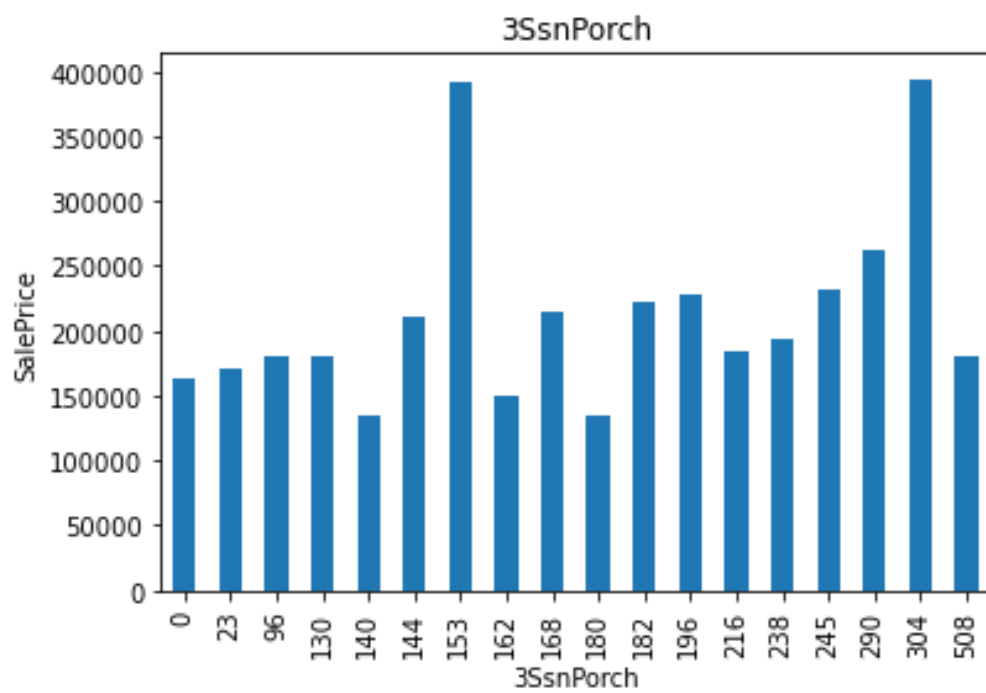


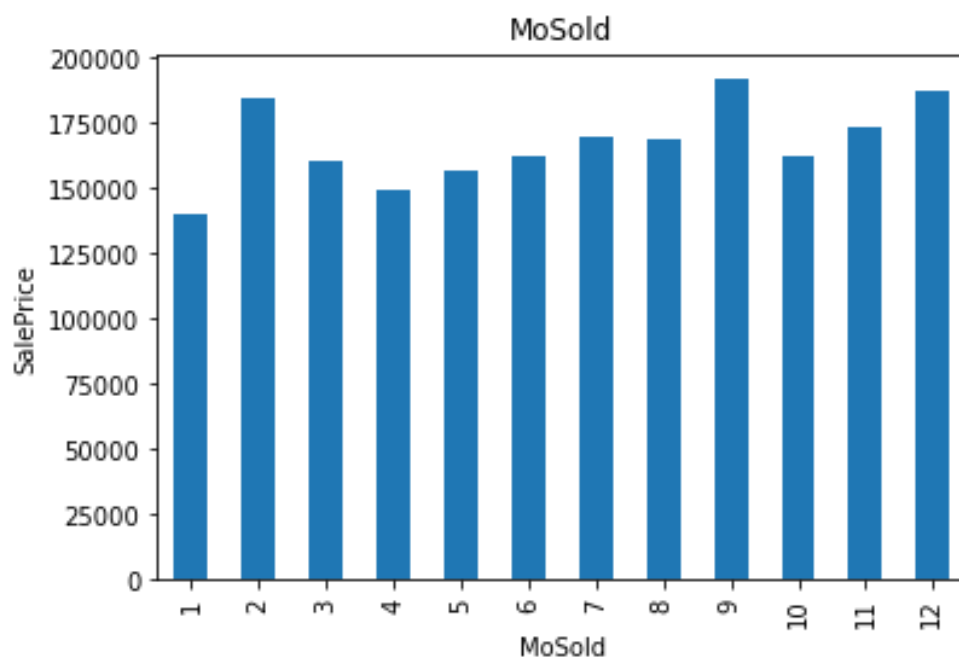
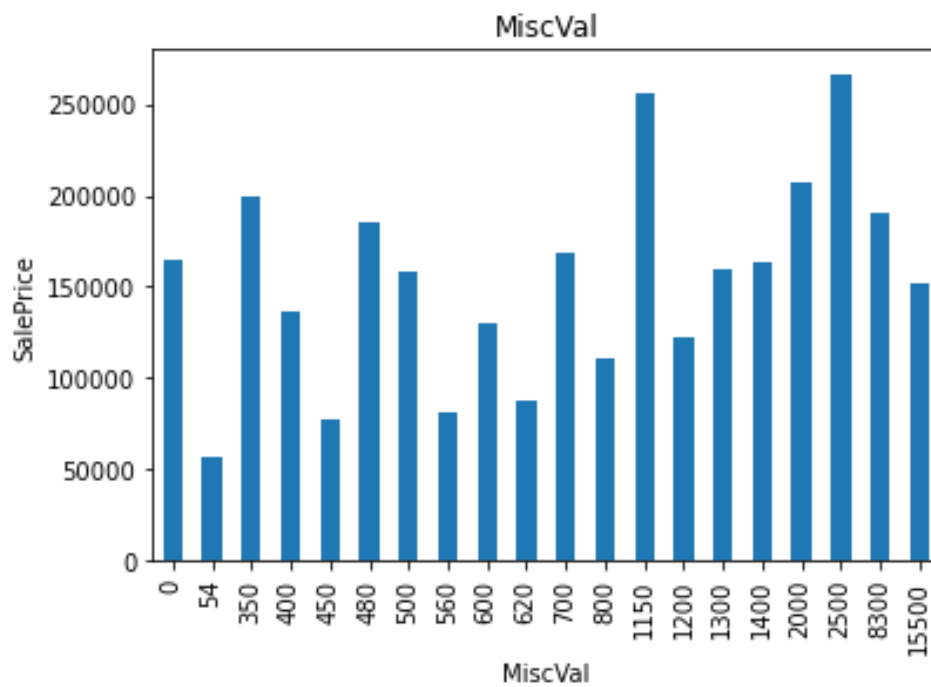




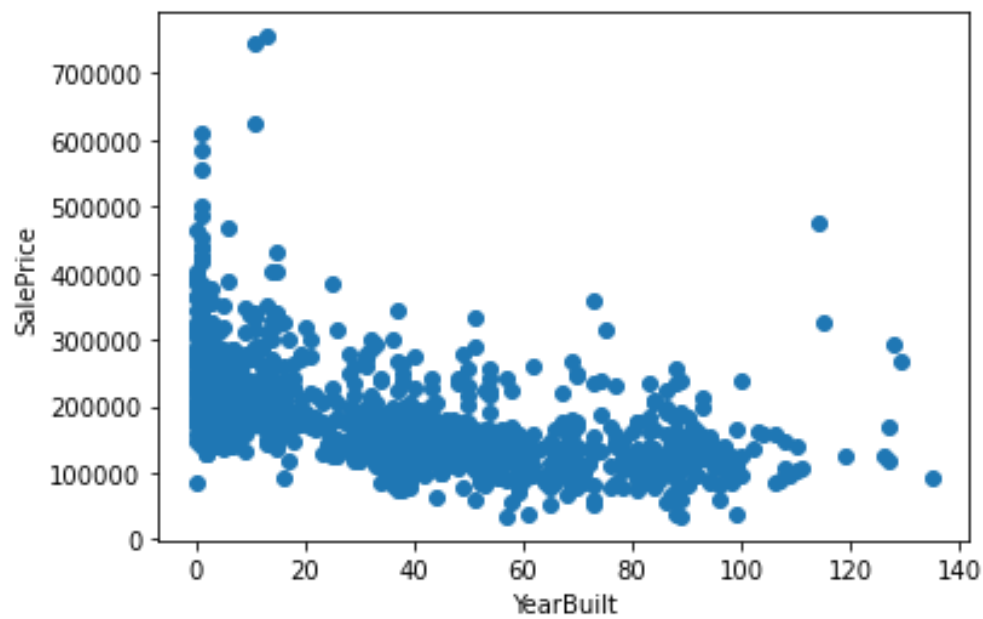




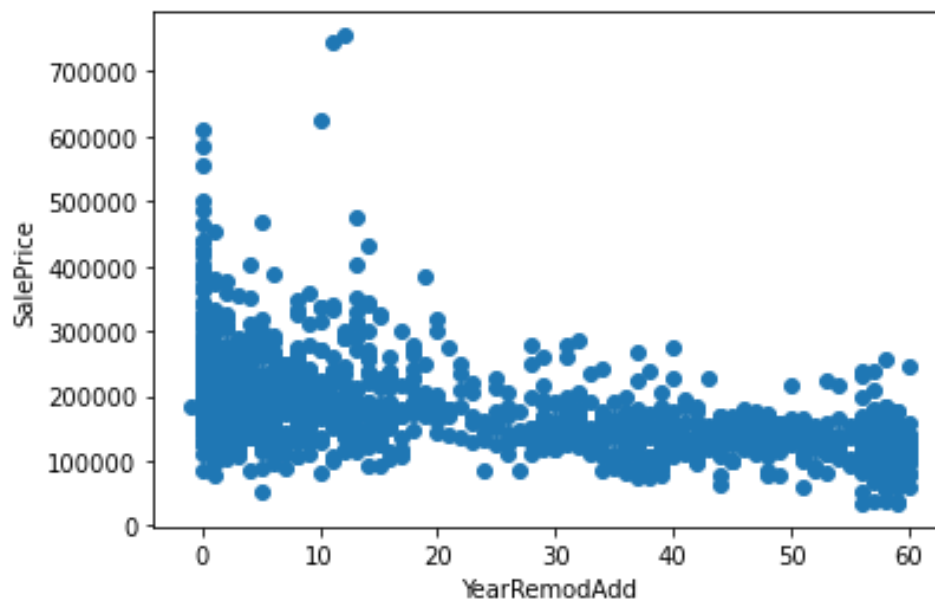




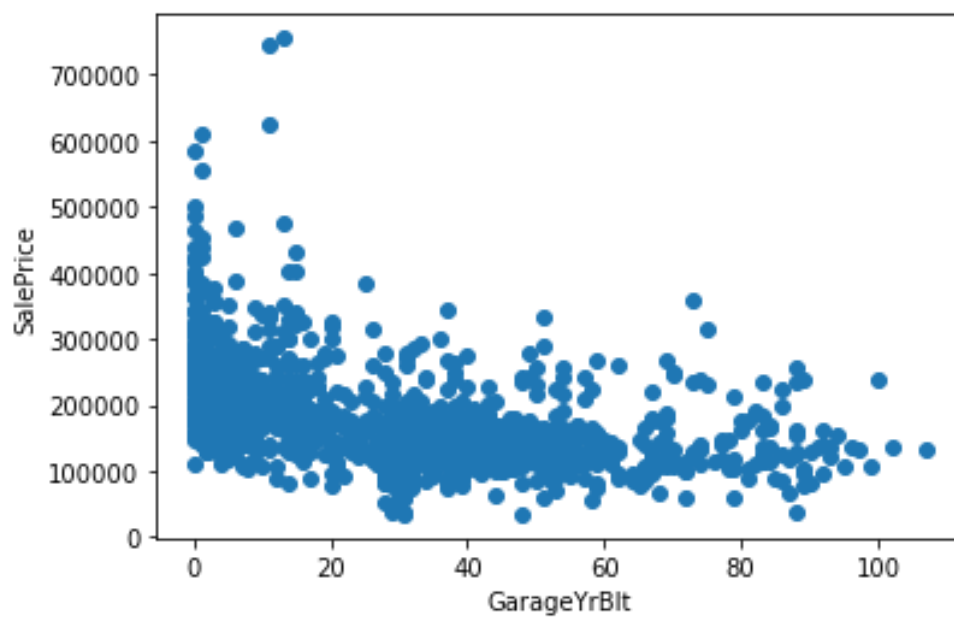
So these are discrete features, we can totally observe that features contains the garage features, kitchen quality, number of bedrooms, living room, number of bath rooms. These are basic features that affects the price.



We can observe more the house gets old price gets low.



Try to modify frequently, otherwise price gets low.



Price gets lowing when garage built year and selling year difference high.

- **Interpretation of the Results**

After visualization I understand that there are so many features that may affect the price of the house. The basement features that is quality of basement interior works in basement, basement length, basement quality and condition so we can understand that basement is an important feature of the house. Then there is overall quality we can assume that this overall quality is evaluated by evaluating all the features.

## **CONCLUSION**

- **Key Findings and Conclusions of the Study**

House is an essential factor for humans for living. But different countries have different culture and different needs. So understanding that needs in different country makes success to the business. A good path to house, when the house built, overall quality of house these type of features are basic features of the house.

- **Learning Outcomes of the Study in respect of Data Science**

Visualization is very helpful for analysing and understanding the data. How the features related to the label, we can easily understand and evaluate it by visualization. We don't know which algorithm is perfect for different project. Some algorithms good for small data and while other algorithms are perfect for large data. So first of all we want to perform multiple models and evaluate its performance and select best algorithm which performs well. Machine wants data to learn about the data. More data gives high accuracy. In this project it is relatively small project. So the performance should be not low may be average.



- Limitations of this work and Scope for Future Work

By analysing the data of this project, I understand that selling the old house or we can say that used houses or the house used to stay will give low prices. There is a reason for that, the materials may be depreciated, the overall quality of house will gets low. It affects the price. So try to sell new house and make deal with customers while the house were partially worked. It will take advantage that if customer want to add anything apart from the house built it is so convenient to customers and business as well.

