FLIP ROBO

Submitted by:-
Mansi Nagpal

HOUSING: PRICE PREDICTION

1/21/2023

House Price Prediction

Introduction

Background and Approach

Top Three Regression Models

Conclusion

Introduction -

The research project was for a company “Surprise Housing” seeking to do business in Australia w.r.t. Housing Price Prediction. Being a Data Scientist, I have used the Regression Model using multiple algorithms to design and optimize the results.

Libraries Explored

- Statistics, Analytical Modelling
- Hyper Tuning Method
- CV Score
- Predictive Modelling and Regression Model, etc.

Motivation Behind

- Develop a holistic understanding like – use of statistics, past prices, analysis based on past data for better and accurate prediction, etc.

Objective

- Project comprises of eight different test cases where the objective was to train for accuracy and test for accuracy using distribution plot to best understand the linearity of clusters.

Illustration

- Histogram Chart
- Distribution and Box Plot
- Scatter Plot
- Category Plot

Problem Statement

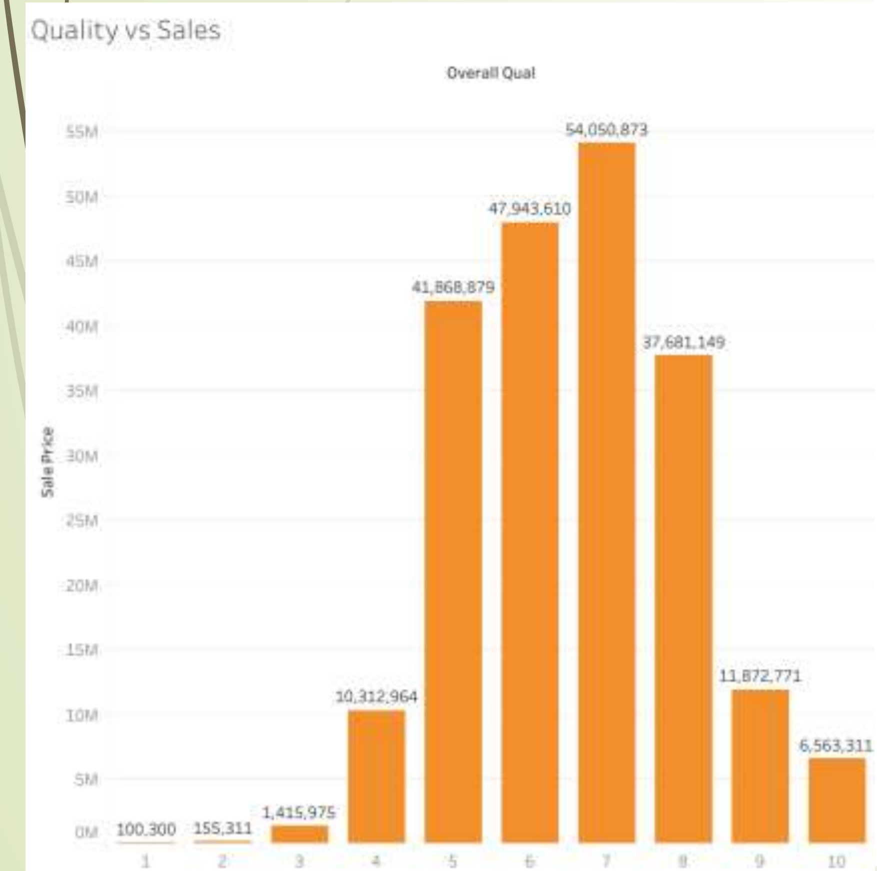
The company is looking at prospective properties to buy houses to enter the market. You are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not. For this company wants to know:

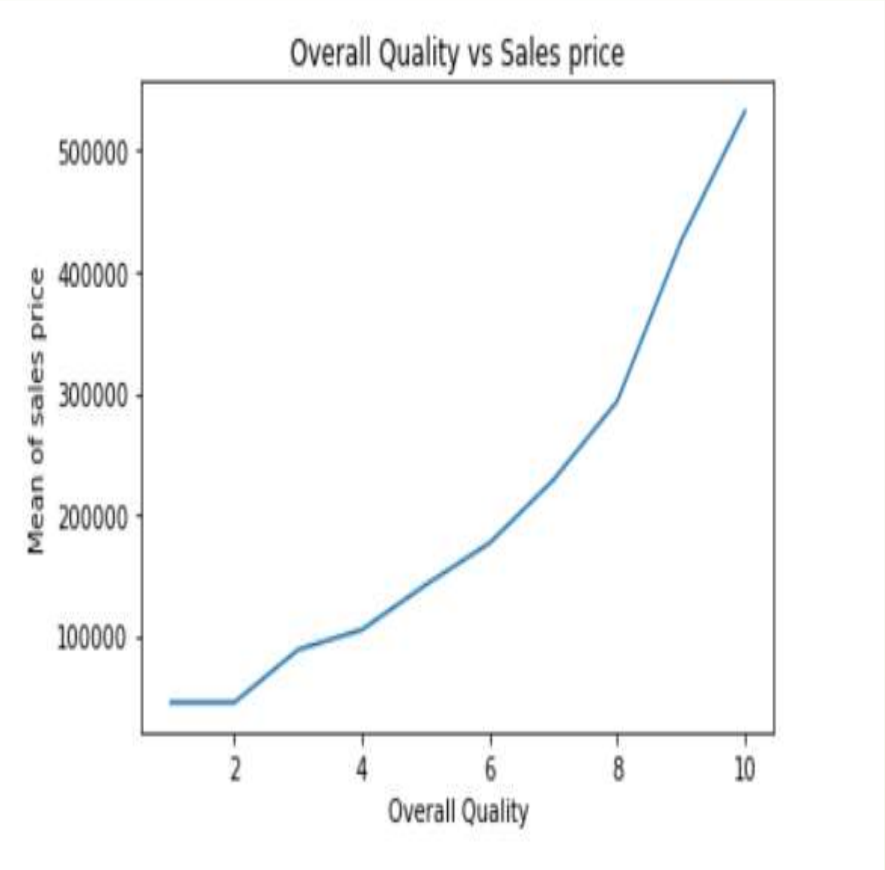
- Which variables are important to predict the price of variable?
- How do these variables describe the price of the house?

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. 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

EXPLORATORY DATA ANALYSIS

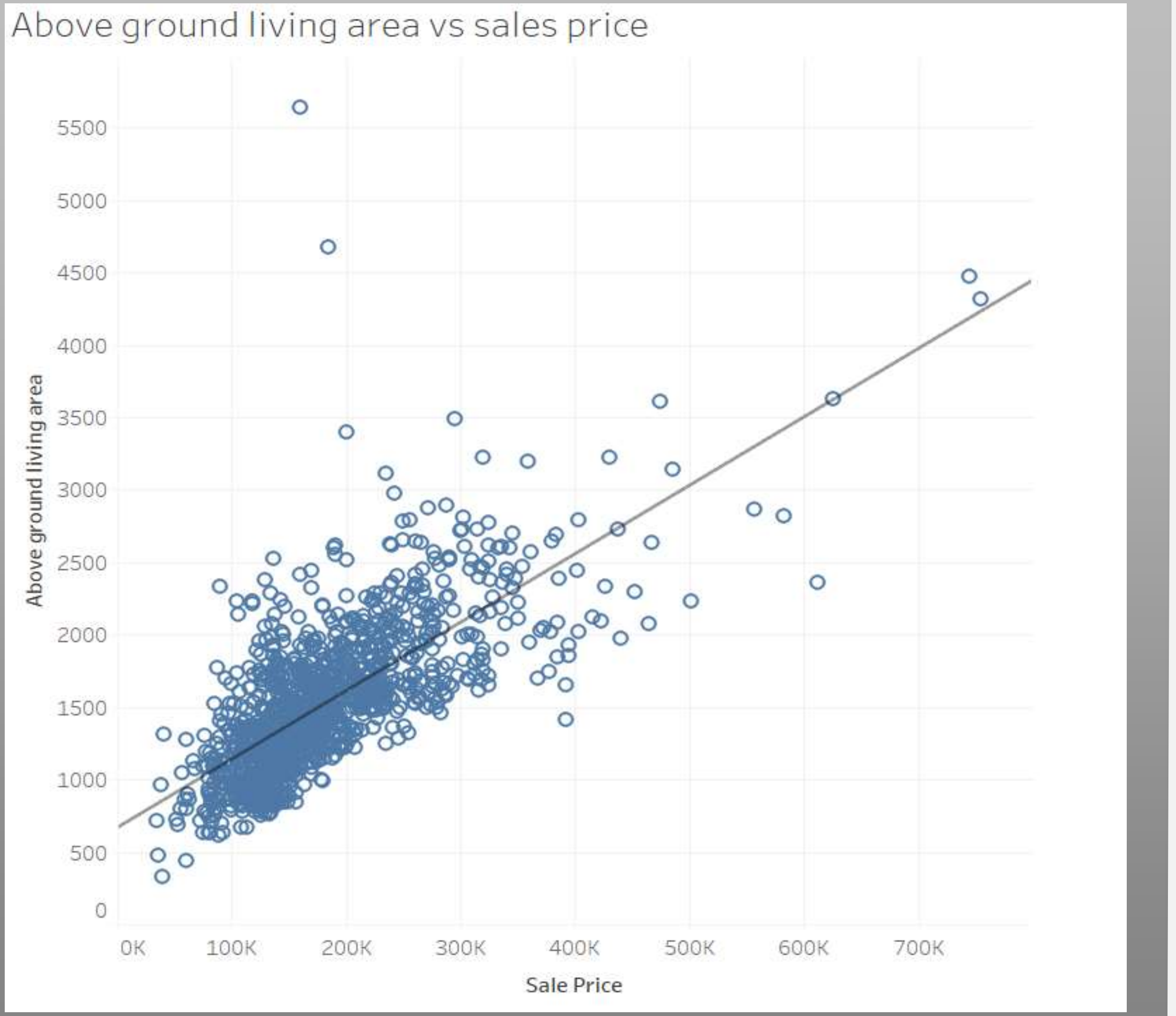
In this given dataset there are 81 columns or parameters, We mainly focused on what are the parameters which decides the sales price, There are 35 parameters out of all which has got a significant effect over deciding the sales price.





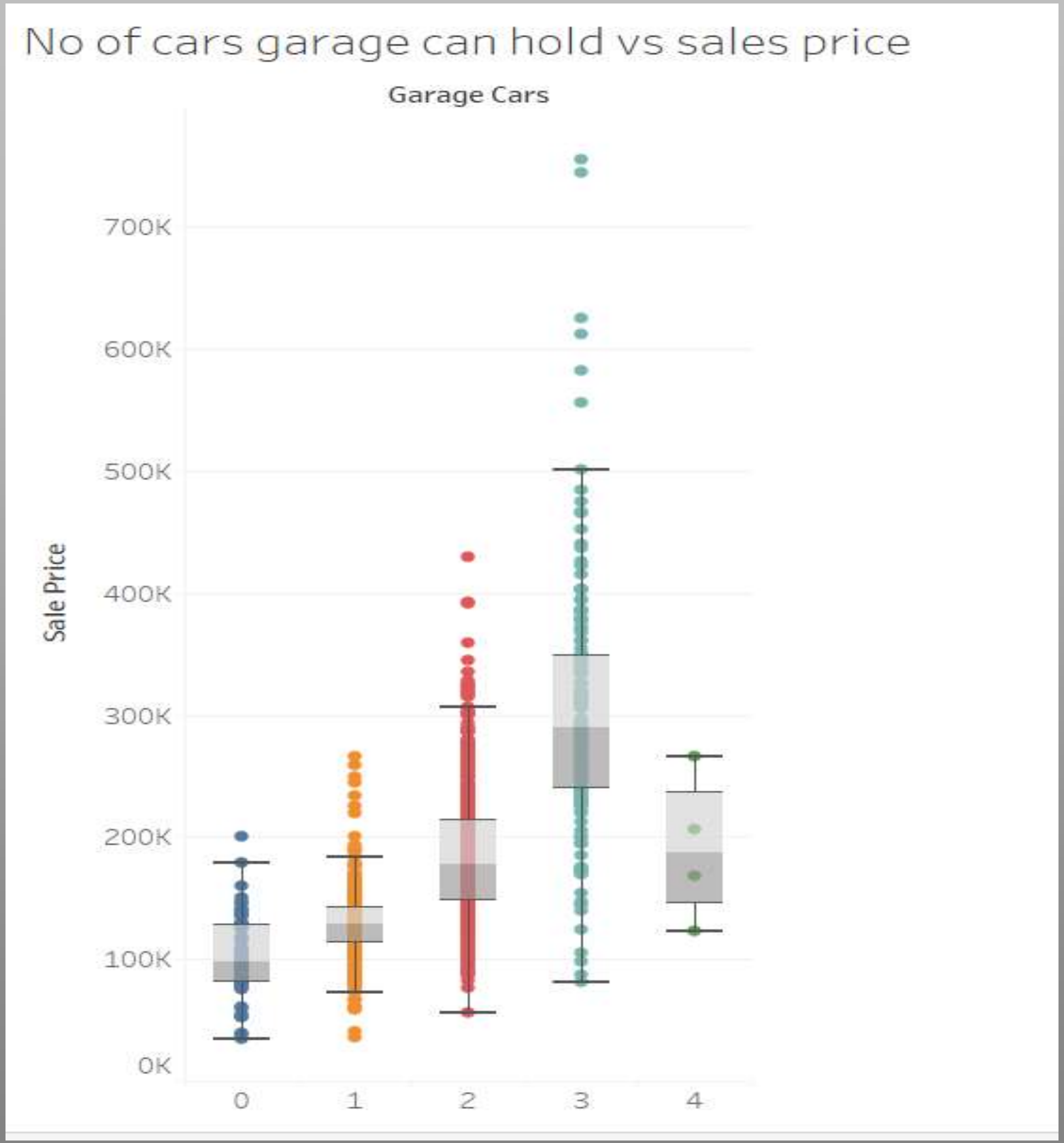
Sl.no	Quality	Range
1	Very poor	30k to 60k
2	Poor	30k to 60k
3	Fair	38k to 139k
4	Below Average	34k to 176k
5	Average	55k to 229k
6	Above average	76k to 277k
7	Good	116k to 341k
8	Very Good	147k to 440k
9	Excellent	239k to 612k
10	Very Excellent	310k to 755k

2.Above grade living area square feet vs Sales Price



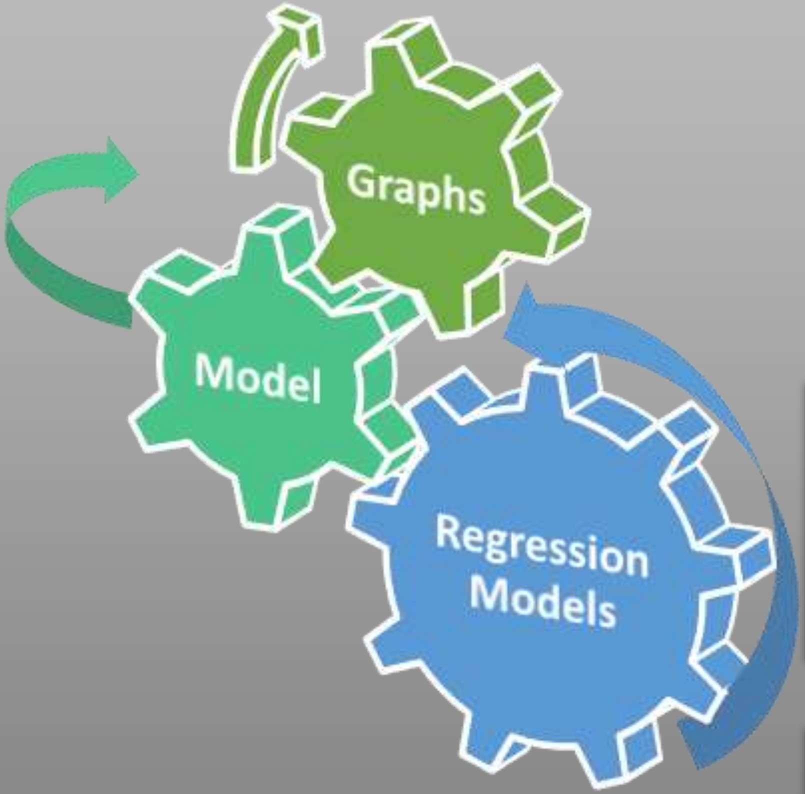
- ❑ The above ground living area and sales price variables vary linearly with each other.

3.No of cars that garage can hold vs Sales Price



- ❑ As the garage size increases the sales price will automatically get increased.

Background and Approach -



The problem statement comprises a housing and Real Estate Company “Surprise Housing” which has decided to enter in Australian market and they seek to buy out properties below the mark-up price.

Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques globally used for achieving the business goals for housing companies.

Currently the assigned project utilises Predictive Modelling algorithm to solve the business statement.

Predective Modelling

Analytical Modelling

CV Score

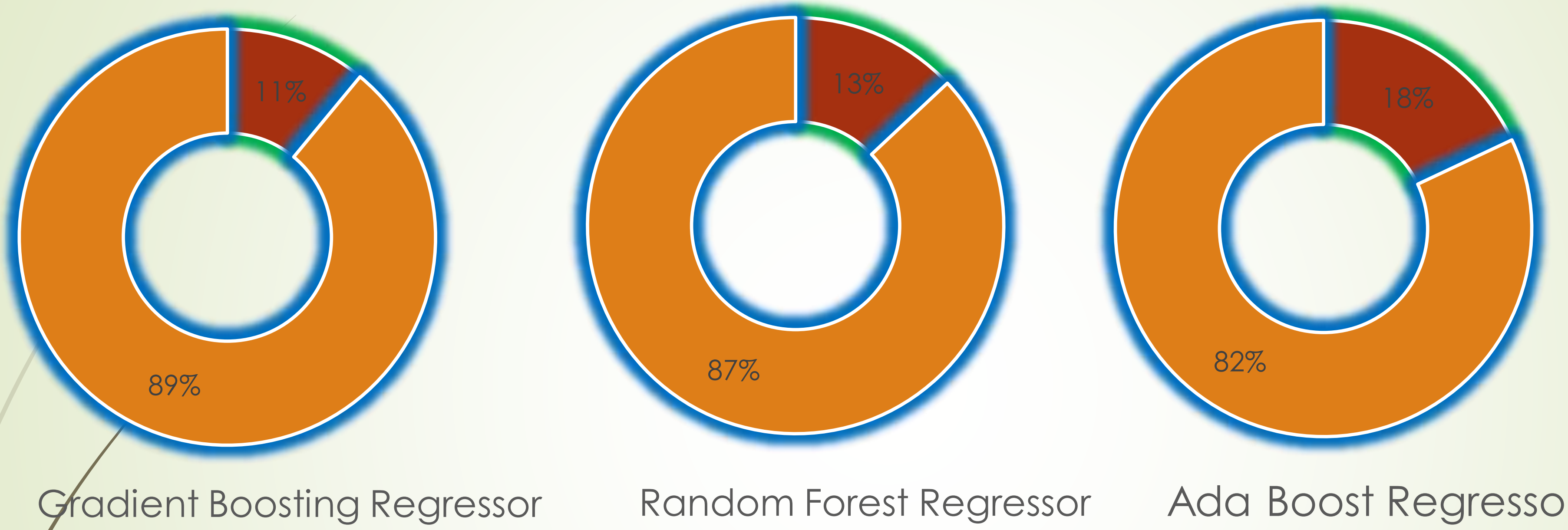
Scikit Learn Liabraries

Regression Modelling

Tuning Method

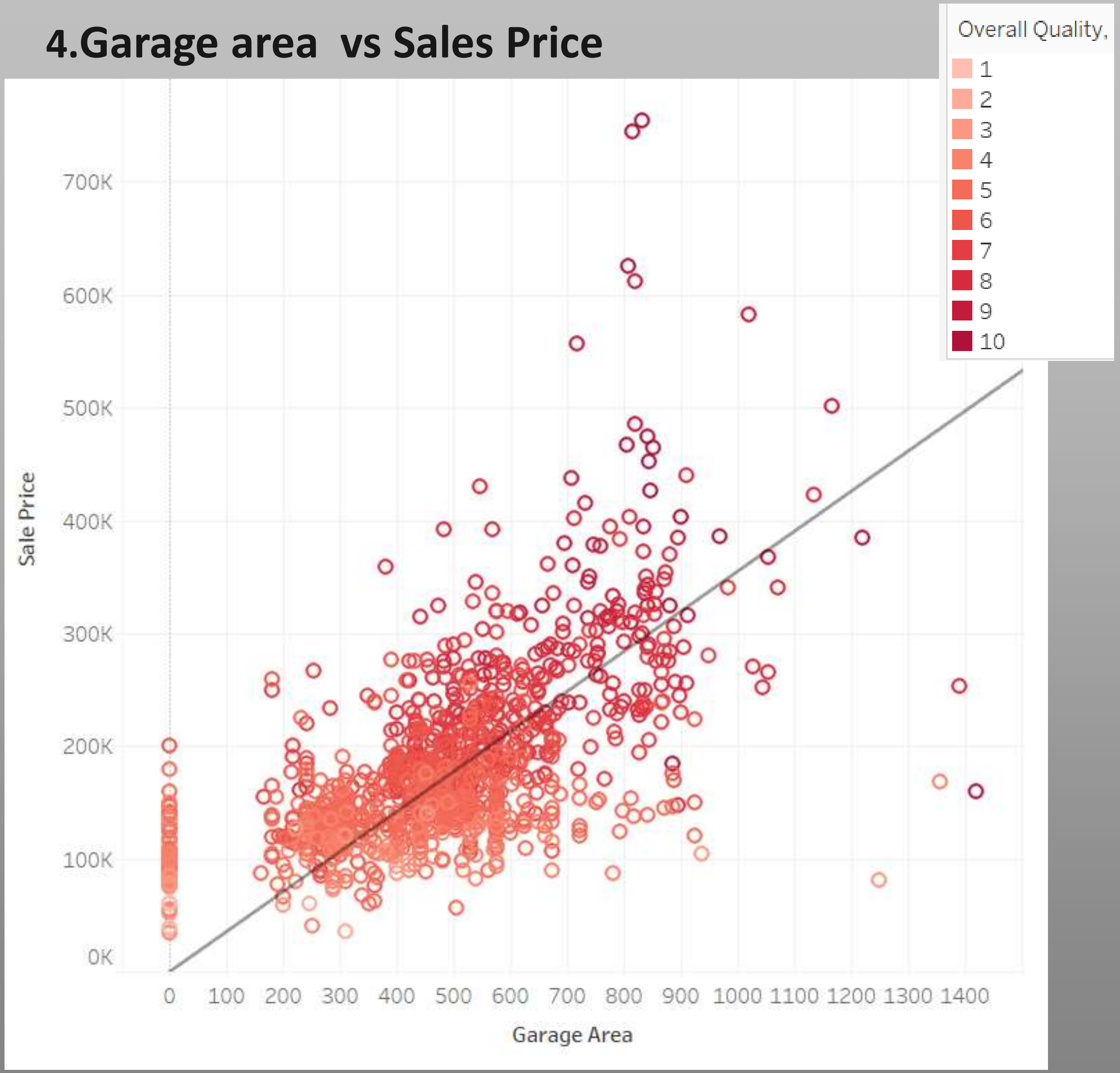
Grid Search

Top Three Regression Models



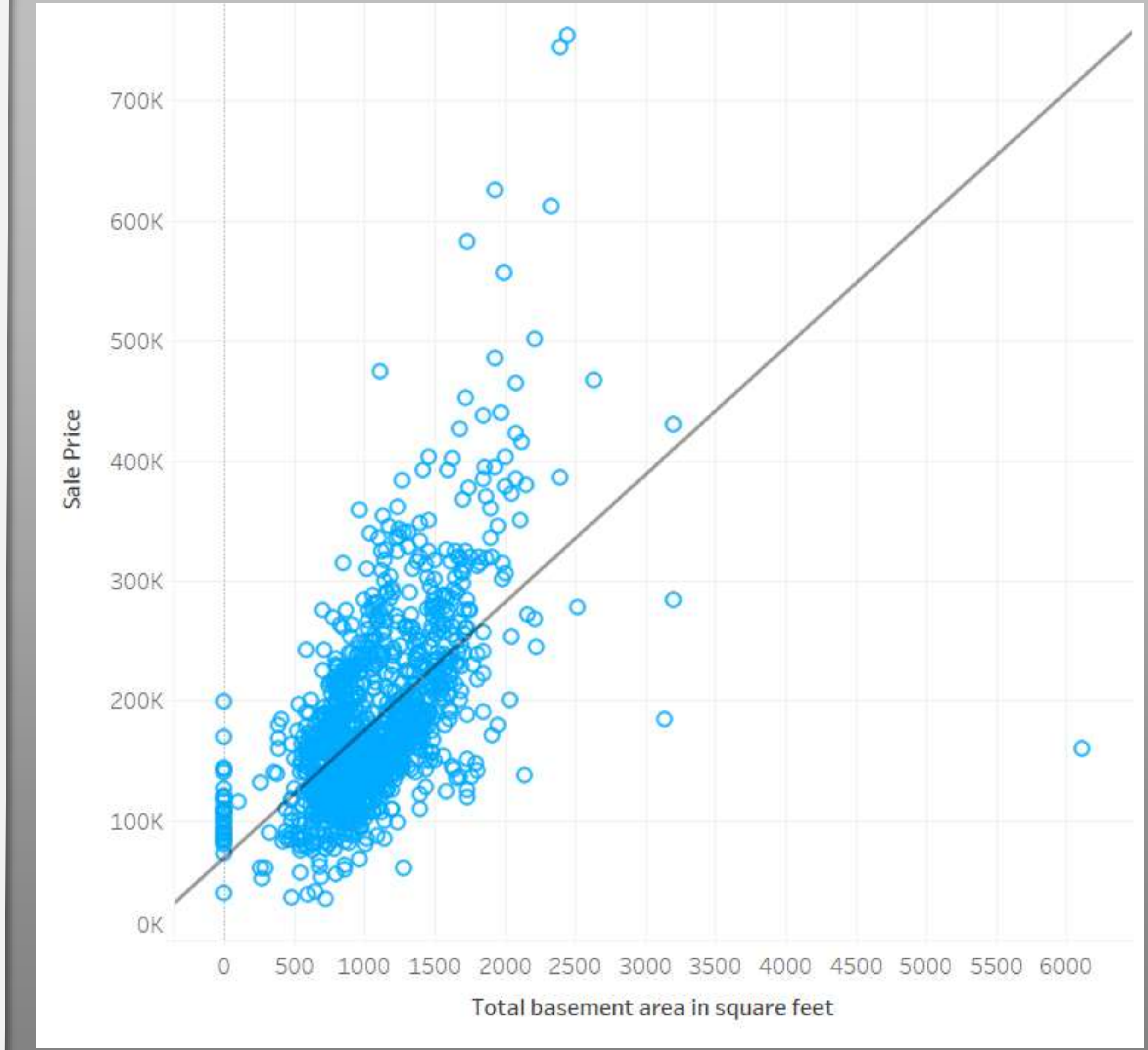
Regression Models	Testing Accuracy (in%)
Ada Boost Regressor	82.24
Random Forest Regressor	78.41
Gradient Boosting Regressor	89.02

4.Garage area vs Sales Price



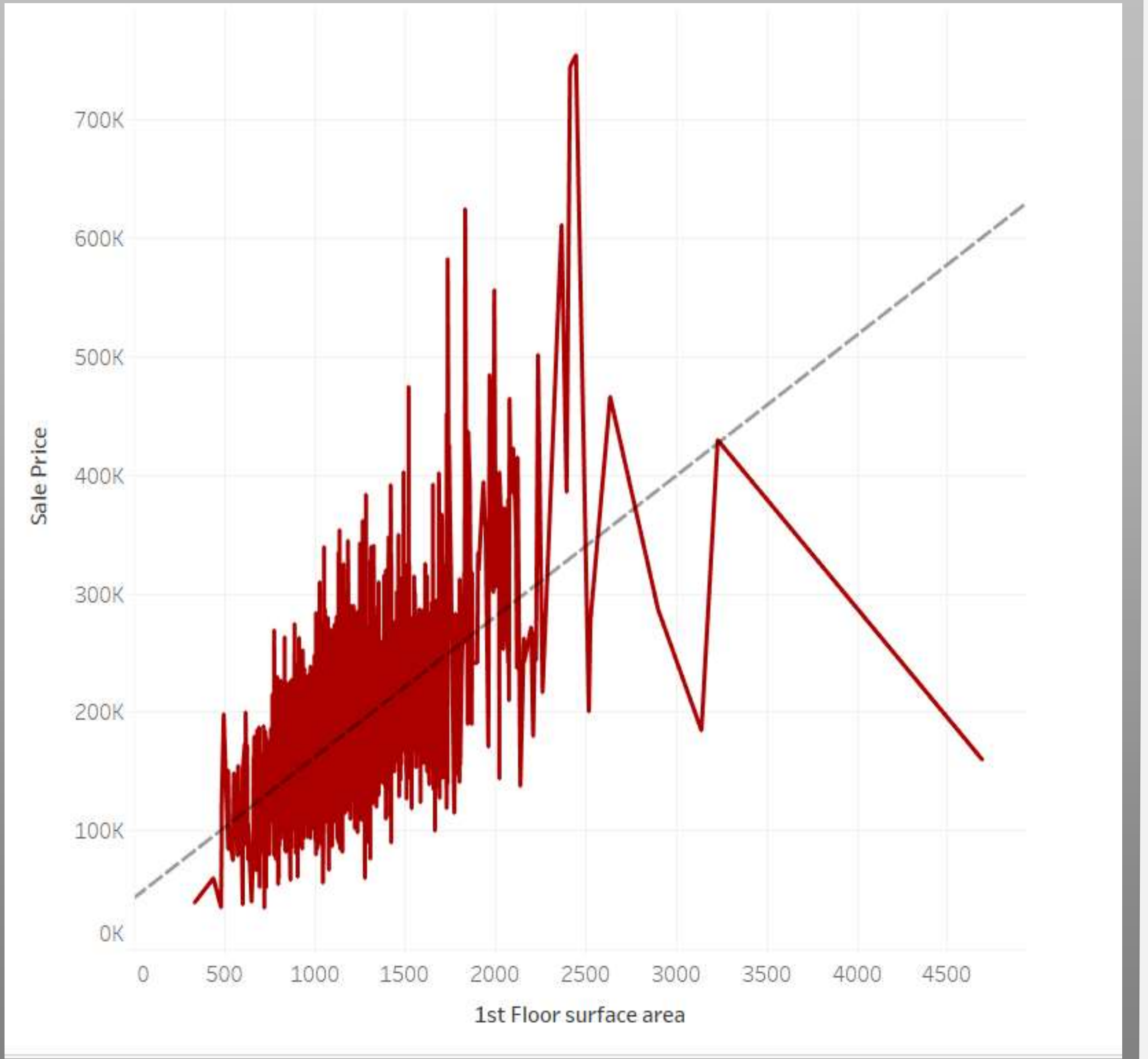
- ❑ The garage and sales price are linearly related.

5. Total basement area in sq ft vs Sales Price



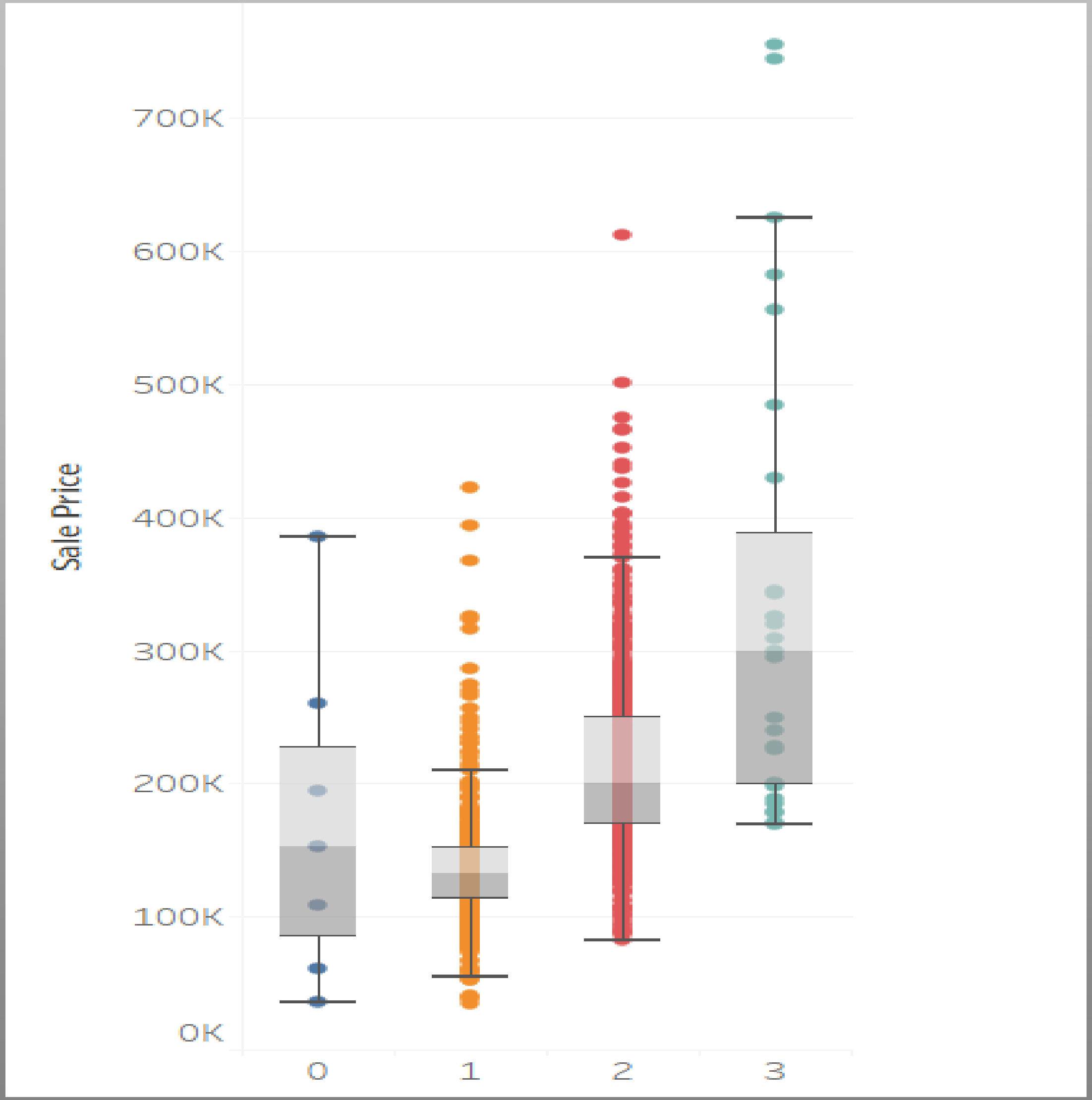
- ❑ The total basement area in sq ft and sales price are linearly related.

6. 1st Floor surface area vs Sales Price



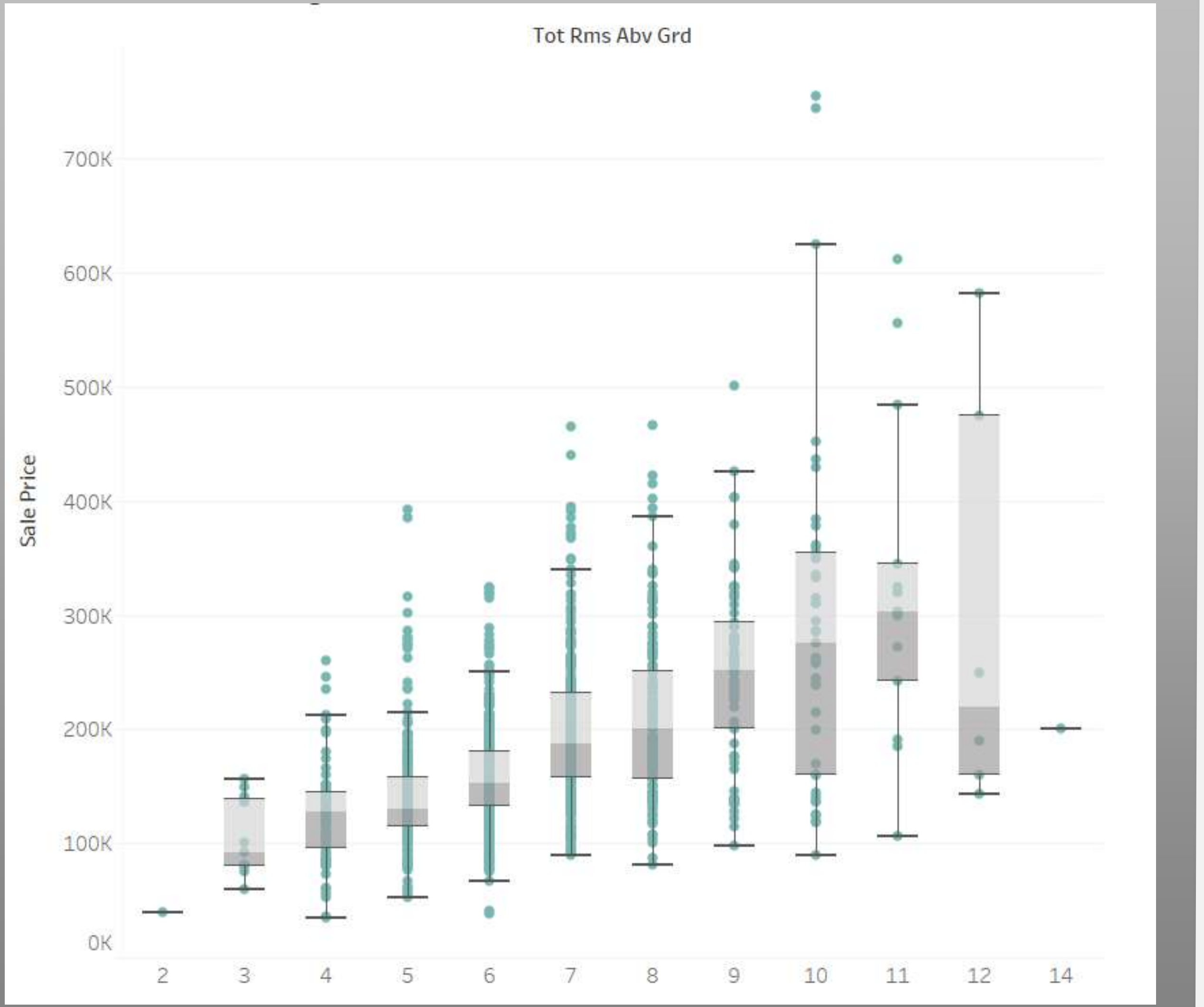
- ❑ There is a positive linear correlation between the 1st floor surface area and sales price.

7.Full bathrooms above ground vs Sales Price



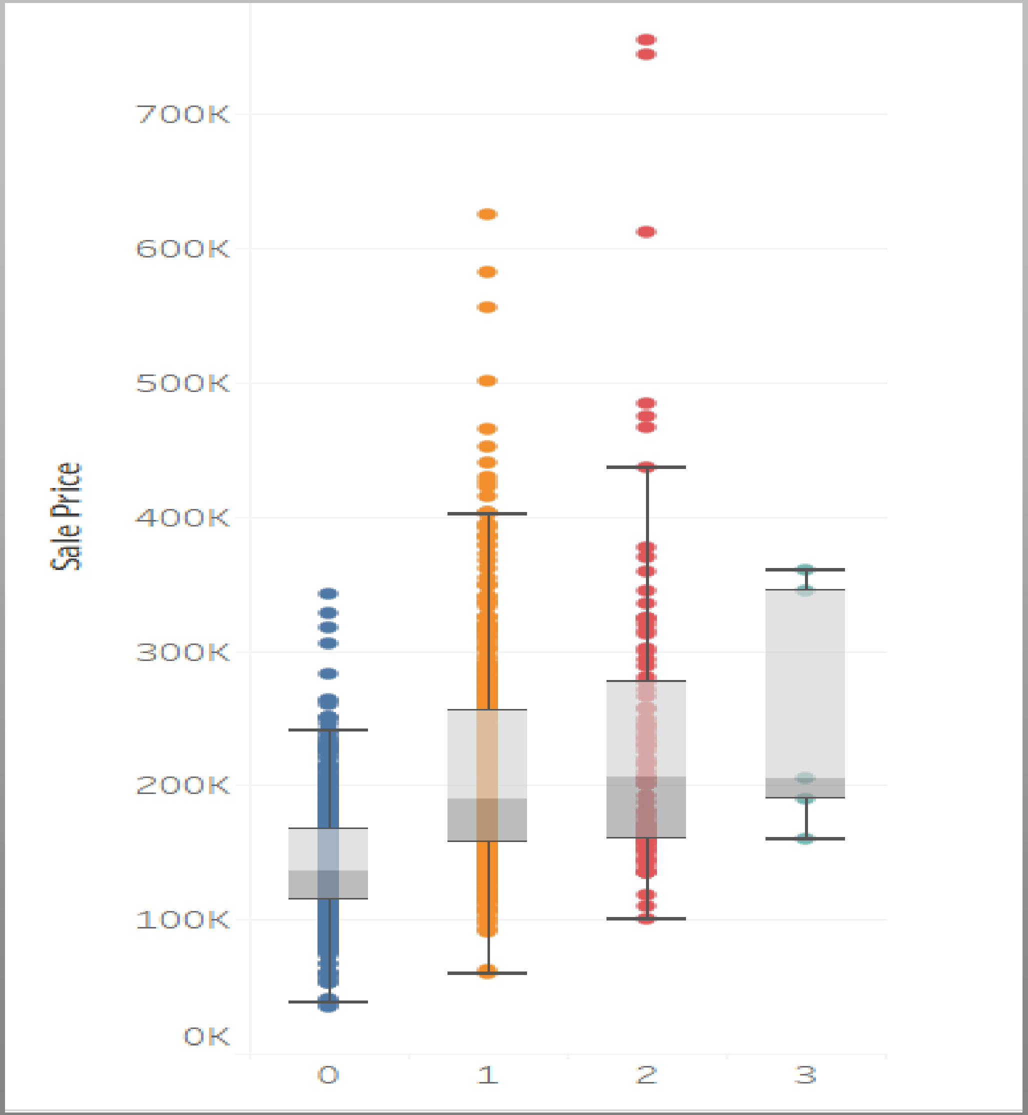
- ❑ Generally as number of full bathroom increases the sales price also going to be increased.

8. Total rooms above ground vs Sales Price



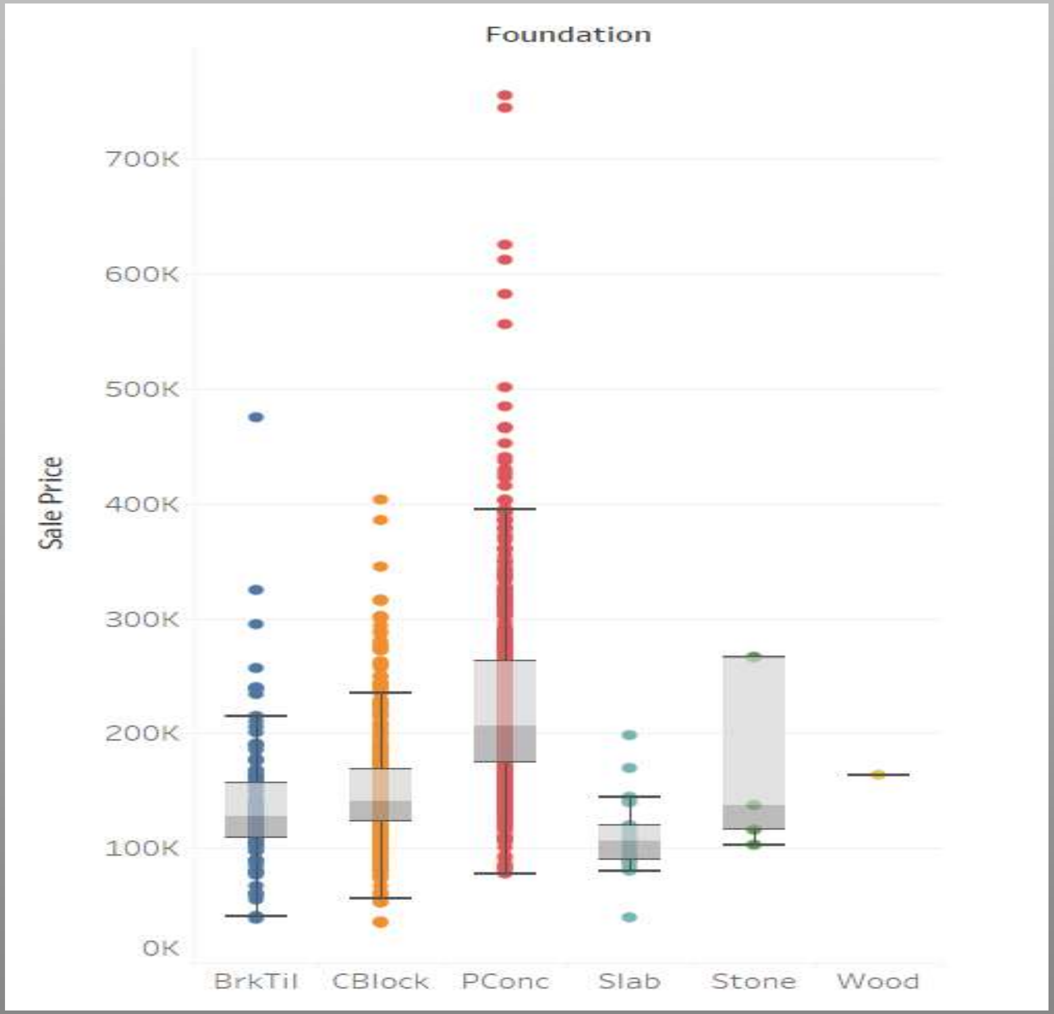
- ❑ As total rooms above the ground increases sales prices also going to increase.

9.Number of fireplaces vs Sales Price



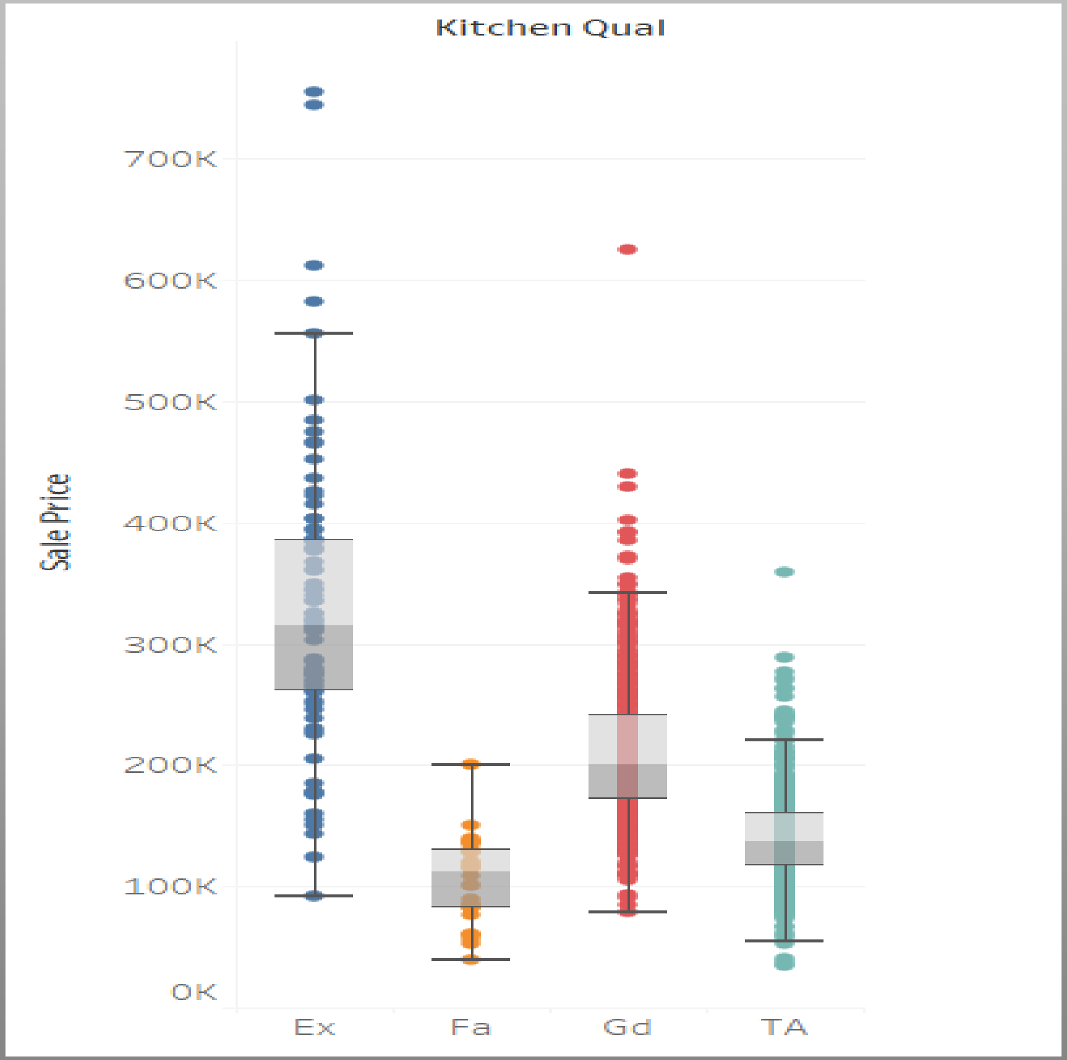
- ❑ Generally as number of fire places increases the sales price also going to be increased.

10.Type of foundation vs Sales Price



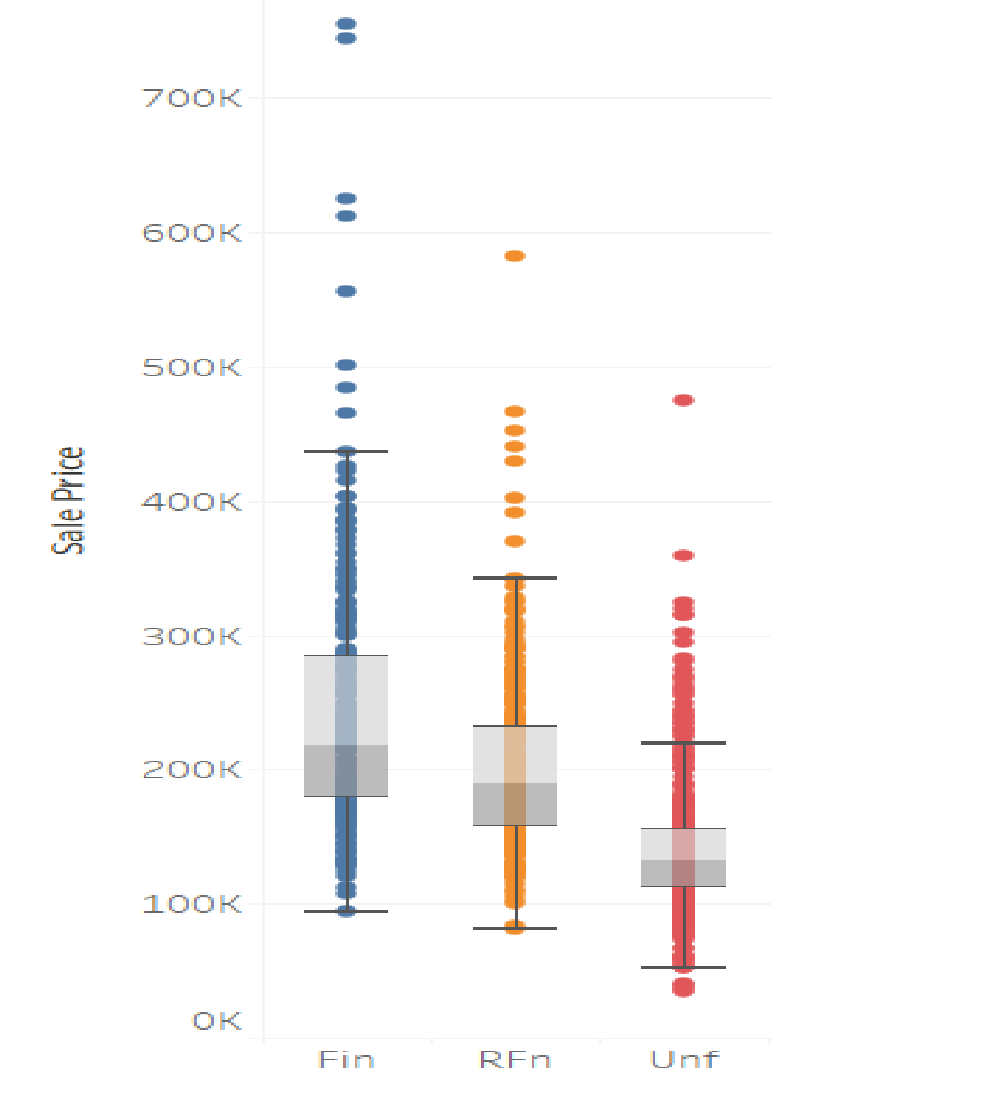
❑ If its poured concrete foundation then its sales price will be higher than any other type of foundation followed by Cinder block,stone,brick and tail and wood.

11.Kitchen quality vs Sales Price



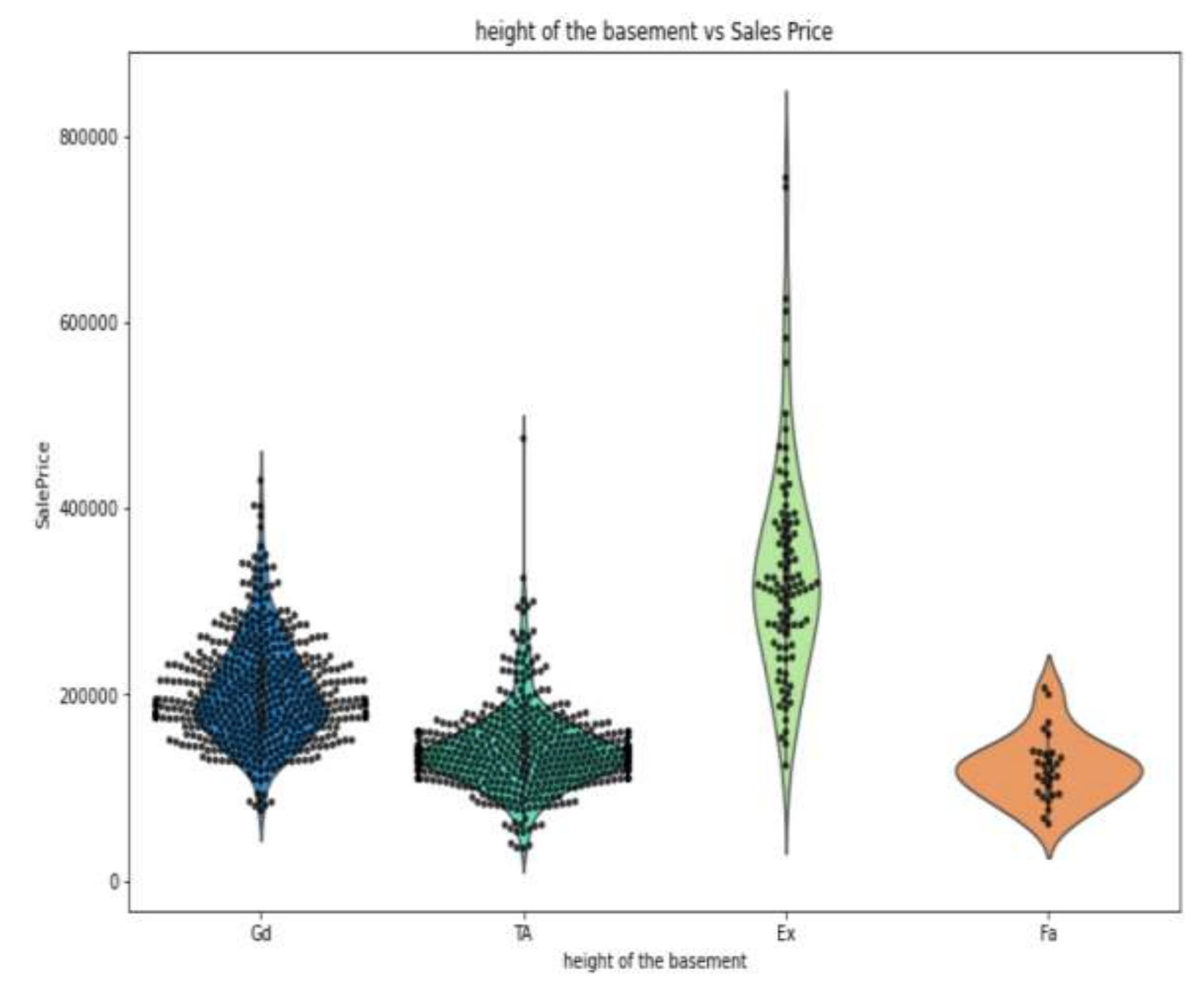
❑ The excellent kitchen quality will get sold for higher price,followed by good,average/typical,fair and poor.

12. Interior finish of the garage vs Sales Price



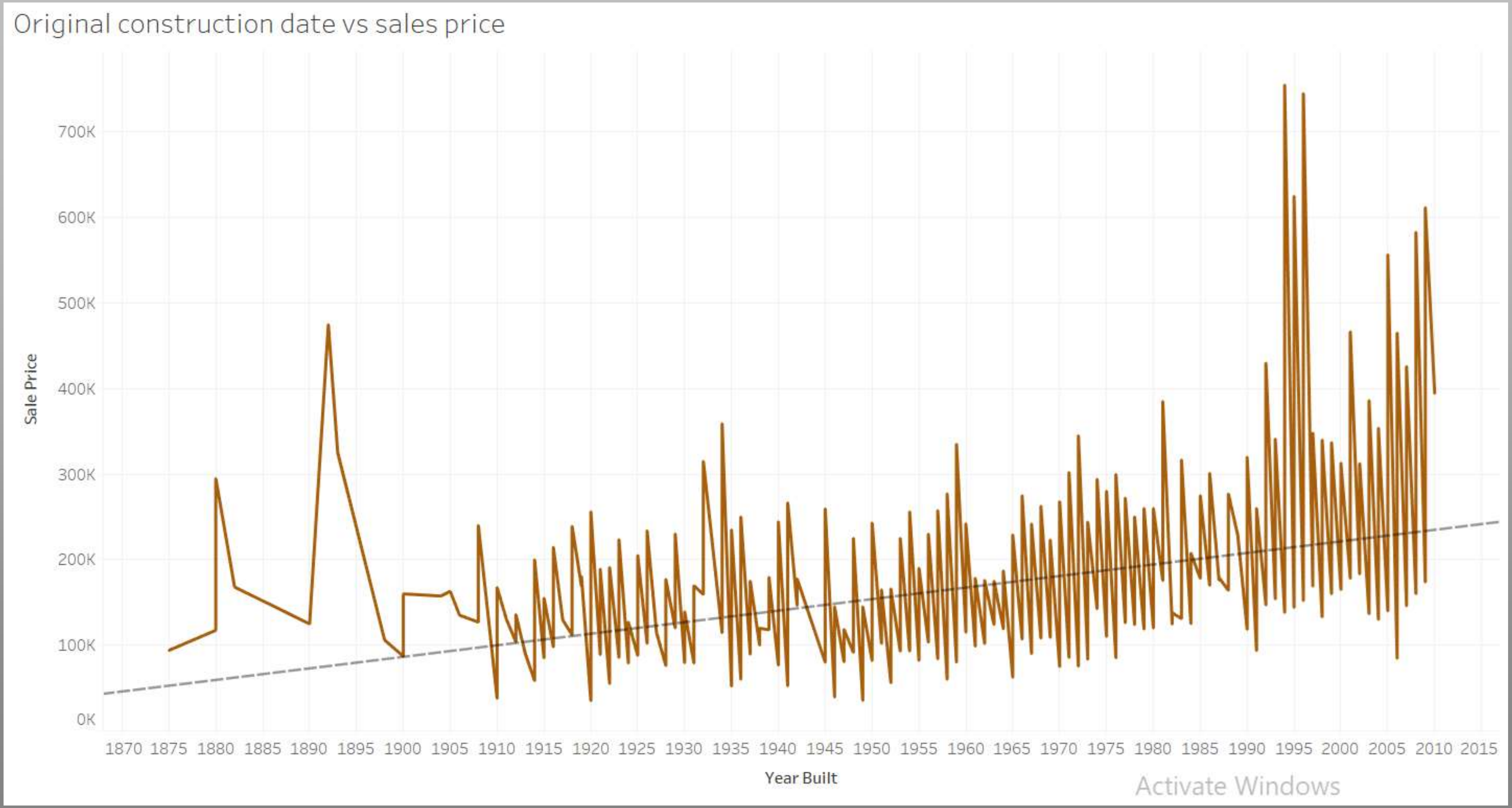
- ❑ If interior of garage is finished then it will have higher sales price followed by rough finish and unfinished.

13. Evaluates the height of the basement vs Sales Price



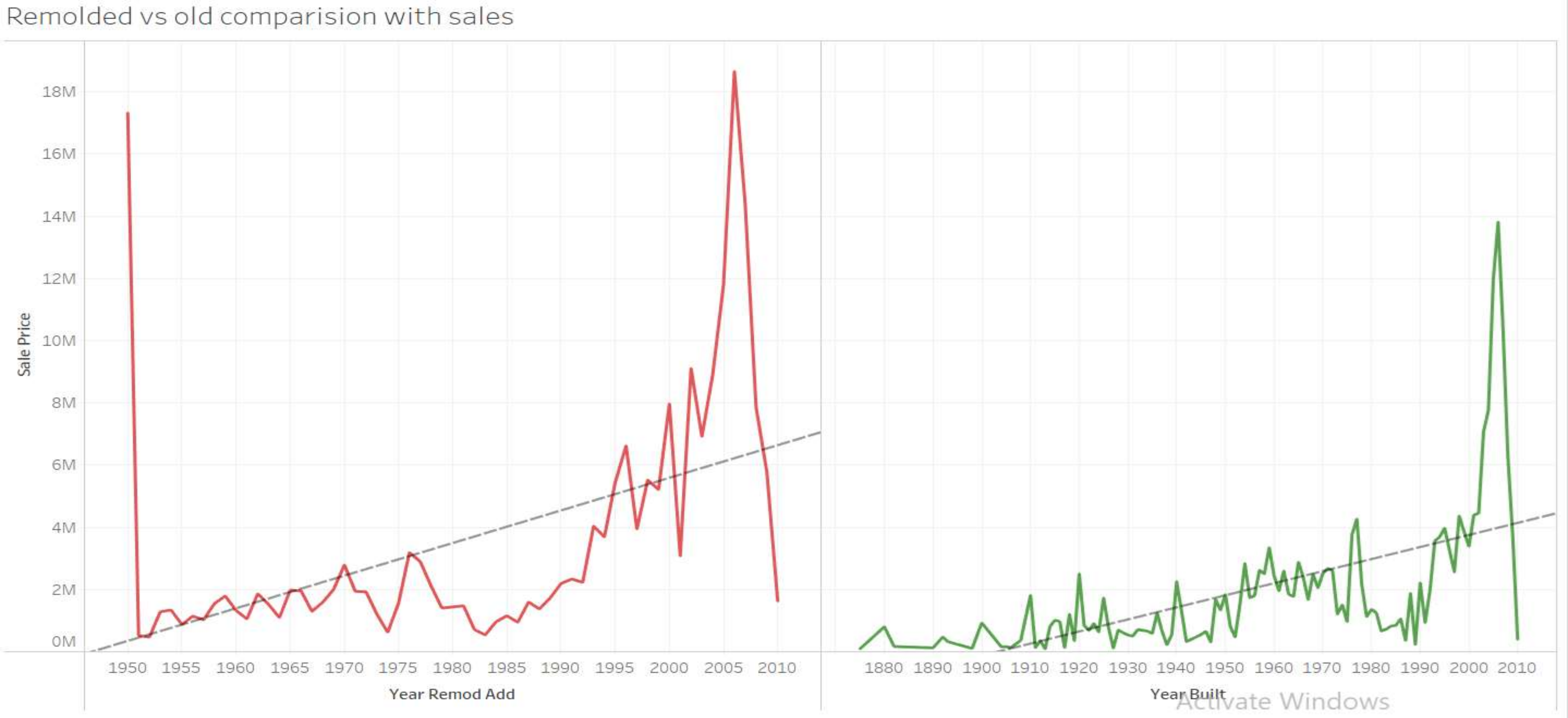
- ❑ The Excellent (100+ inches) basement will get sold for significantly high price followed by Gd-Good (90-99 inches),TA-Typical (80-89 inches)

14. Year in which house was built vs Sales Price



❑ We can see the trend that if the house was built in recent years then it will have high sales price.

15. Comparing the remodeled and non remodeled houses against sales price



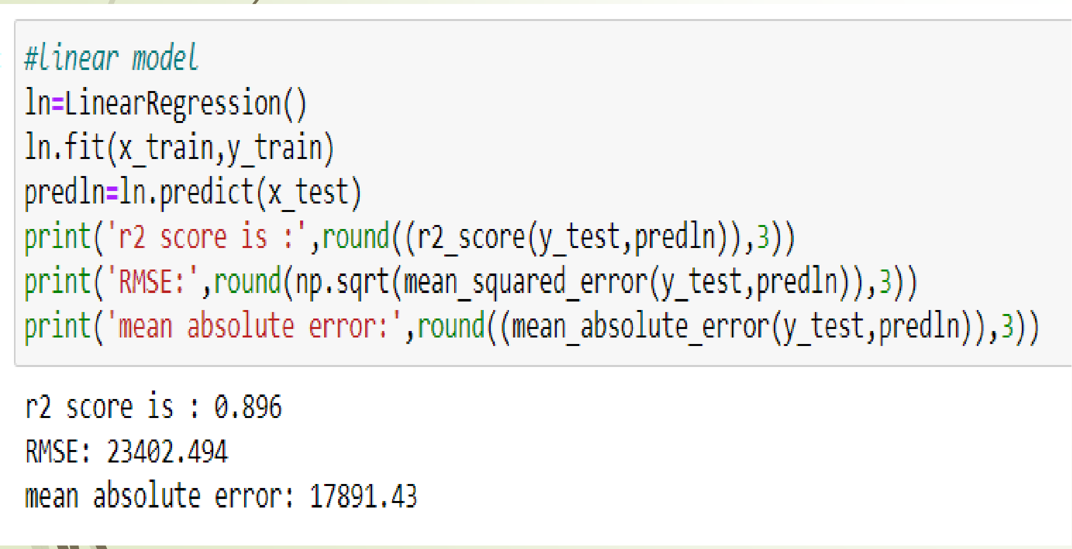
❑ The remolded house will get sold for better price than non remolded house

Building the model

- ❑ In order to fetch the best suited model for this dataset we need to evaluate all major parameters regarding the linear regression models, we have to find the difference between cross validation score and accuracy , the least of the difference is considered as best model and we had hyper parameter tuned that.

- ❑ The accuracy ,cross validation score and their difference are follows.

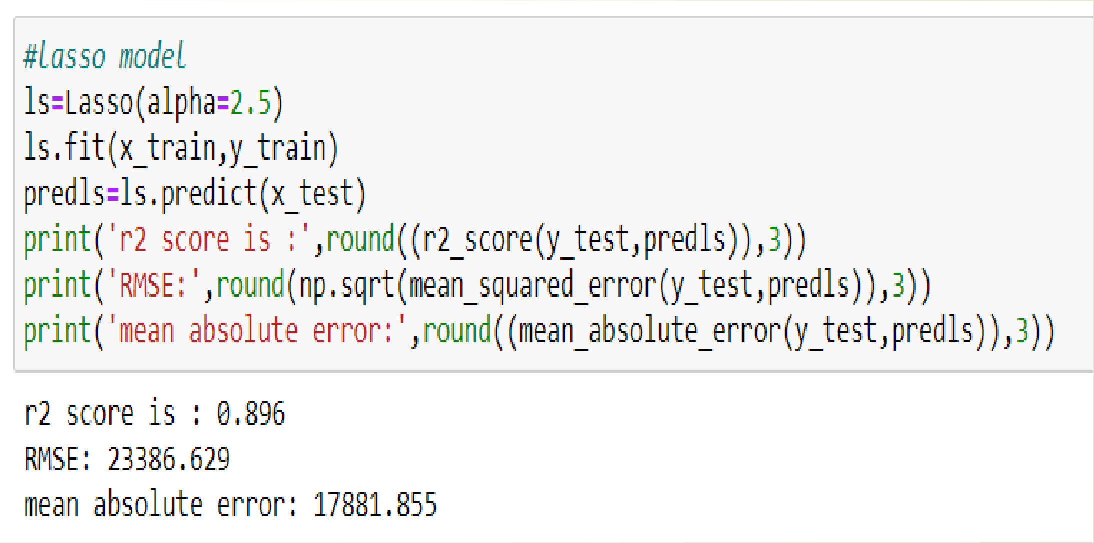
Linear Model



```
#linear model
ln=LinearRegression()
ln.fit(x_train,y_train)
predln=ln.predict(x_test)
print('r2 score is :',round((r2_score(y_test,predln)),3))
print('RMSE:',round(np.sqrt(mean_squared_error(y_test,predln)),3))
print('mean absolute error:',round((mean_absolute_error(y_test,predln)),3))
```

```
r2 score is : 0.896
RMSE: 23402.494
mean absolute error: 17891.43
```

Lasso Model



```
#lasso model
ls=Lasso(alpha=2.5)
ls.fit(x_train,y_train)
predls=ls.predict(x_test)
print('r2 score is :',round((r2_score(y_test,predls)),3))
print('RMSE:',round(np.sqrt(mean_squared_error(y_test,predls)),3))
print('mean absolute error:',round((mean_absolute_error(y_test,predls)),3))
```

```
r2 score is : 0.896
RMSE: 23386.629
mean absolute error: 17881.855
```

Conclusion -

- I have tested out the prediction over the best three models. Out of the three, Gradient Boosting Regressor is the top model as it giving the testing accuracy as almost 89% which is higher than that of Random Forest Regressor as well as AdaBoost Regressor which are giving testing accuracy as 82% and 78% respectively.

- Also Random Forest Regressor and Gradient Boosting Regressor training accuracy is more than 93% as well and it indicates that biasness and variance are optimal and model is regularized as well.

- Hence Gradient Boosting Regressor is my top accurate model in predicting the Sale Price of the house.

- Also as you can see in the original and predicted Sale Price row, almost all the corresponding data's are equal mostly for Gradient boosting and Random forest model and the best fit line is containing the most data points as well and the distribution is normal distribution

Thank you