# Statistical Analysis on House Price of Saratoga, New York

## Introduction

Sweta Kumari

MSBA320, Fall 2019

Golden Gate University

**Contents**

# Statistical Analysis on House Price of Saratoga, New York

## Introduction

Saratoga is a county in New York and known for sudden increase in population in the northeast. According to the US census, population is 230,163 in 2018, shows a 4.8% increase in last 8 years (Wikipedia, n.d.). This county appeals to permanent residents due to its unique location with big-city amenities and a small-town feel. Due to this, the house price of Saratoga has been increasing in the past years. After 2000, the average sale price increased in Saratoga by 13% at the end of 2003 (Wood, 2004). This has triggered many interested buyers to think many times before they look for a house. Rising house prices are becoming a major issue for young professionals and working-class people. In June 2003, Saratoga was declared as the state's most expensive housing markets. Seller and buyers both have started analyzing the house data for a good house price prediction that would help better prepare everyone before they think of one of the most important financial decisions in their lives. We have collected data for the year 2006 when house prices rose drastically in Saratoga. To understand the various factors of increasing price, we have performed statistical analyses such as descriptive statistics, multiple regression, ANOVA, correlation, and t-test.

## Data Collection

To perform the statistical analysis, we collected data from DASL website: https://dasl.datadescription.com/datafile/housing-prices-ge19/ . Due to a wide variety of datasets, we chose the DASL website among other websites. This dataset has been originally collected from the Zillow website. With 36 million unique visitors on a monthly average basis, Zillow determines the property value with the help of public and user-submitted data (Bruke,

2019). Saratoga County's house data was captured for the year 2006 to understand how house prices increased with the help of statistical analysis.

**Goals and Variables**

With the statistical analysis of Saratoga house prices, we would like to perform following analysis:

1. To understand the variability and distribution of each variable.
2. To understand how house price has varied for different fuel type, sewer type, and heat type?
3. To understand how house price has risen with the increased number of bedrooms, fireplaces, rooms, and bathrooms.
4. To find a correlation among each variable.
5. To compare two groups via t-test
6. To find the final equation of house price using multiple regression.

Below are the variables with description:

Table1: Data description

| Column Header | Description |
|---|---|
| Price | Home Price (in 1000s of US dollars) |
| Lot.Size | Size of lot (square feet) |
| Waterfront | Whether property includes waterfront( 0: No, 1: Yes) |
| Age | House of Age (in years) |
| Land.Value | Value of land (1000s of US dollars) |
| New.Construct | Whether the property is a new construction (0: No, 1: Yes) |
| Central.Air | Whether the house has central air (0: No, 1: yes) |
| Fuel.Type | Fuel used for heating(Electric/Gas etc.) |
| Heat.Type | Type of heating system (Electric/hot water etc.) |
| Sewer.Type | Type of sewer system(Public/Private/unknown) |
| Living.Area | Living Area (in square feet) |
| Pct.College | Percent of neighborhood that graduated from college. |
| Bedrooms | Number of Bedrooms |
| Fireplaces | Number of fireplaces in house |
| Bathrooms | Number of bathrooms (half bathrooms have no shower or tub) |
| Rooms | Number of Rooms |

## Data Analysis

We have done a few modifications in variables:

1. Changed below numerical variables to factor variables:

   Fireplaces, Bedrooms, Rooms, Bathrooms.

2. Changed below binary variables to factor variables:

   Waterfront, New.Construct, Central.Air

3. Changed the numbers 0 and 1 of binary variables (Waterfront, New.Construct, Central.Air) into “No” and “Yes” respectively.

4. Removed one row which showed the number of bathrooms was zero.

5. Removed the two rows from the dataset:

   Houses that use wood and solar as fuel source had only one row each. It does not give any insight to analyze data.

### Descriptive Statistics

a. Summary Statistics:

   Mean:

```
##          Price       Lot.Size            Age     Land.Value    Living.Area
## 211710.3362218      0.5007972     28.1473137  34549.7515887   1753.6915078
##    Pct.College
##     55.5696129
```

   Median:

```
##       Price    Lot.Size         Age  Land.Value Living.Area Pct.College
##   189900.00        0.37       19.00    25000.00     1632.00        57.00
```

   Five number summaries:

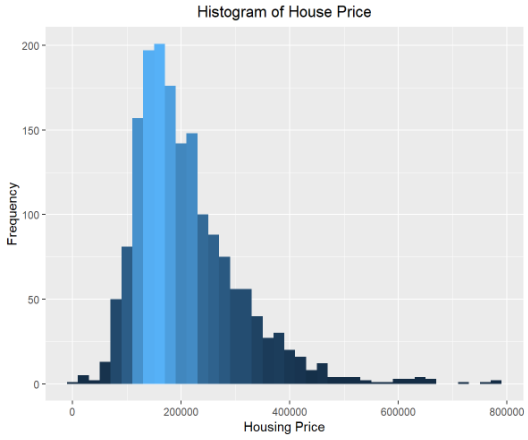
```
##       Price Lot.Size Age Land.Value Living.Area Pct.College
## [1,]   5000     0.00   0        200       616.0          20
## [2,] 145000     0.17  13      15100      1300.0          52
## [3,] 189900     0.37  19      25000      1632.0          57
## [4,] 258193     0.54  34      40200      2135.5          64
## [5,] 775000    12.20 225     412600      5228.0          82
```

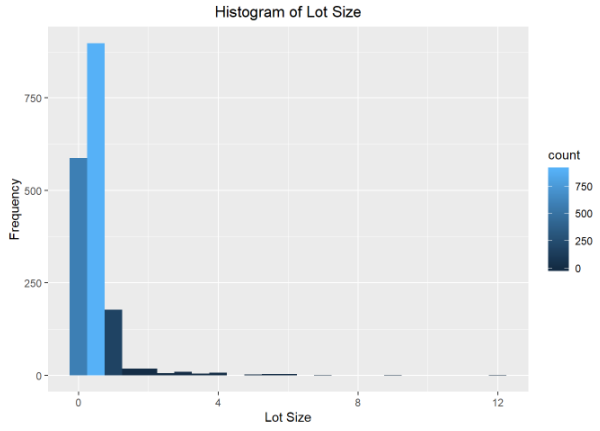
Range of house price:

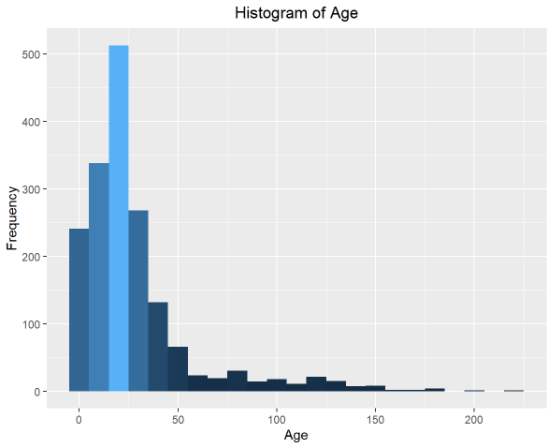
```
## [1]   5000 775000
```

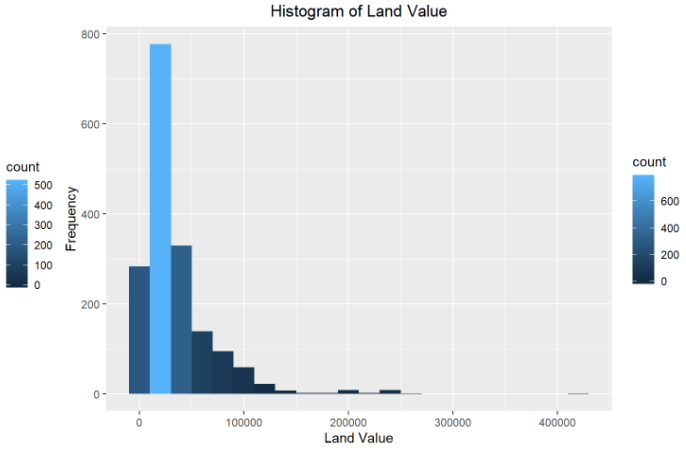
b. Histogram for continuous variables:

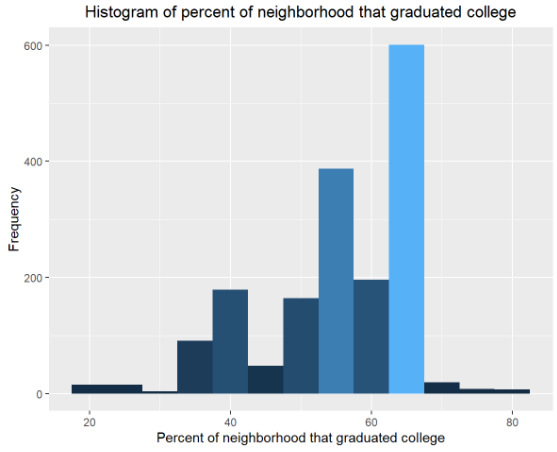
Continuous variables are Price, Lot.Size, Age, Land.Value, Pct.College and Living.Area.

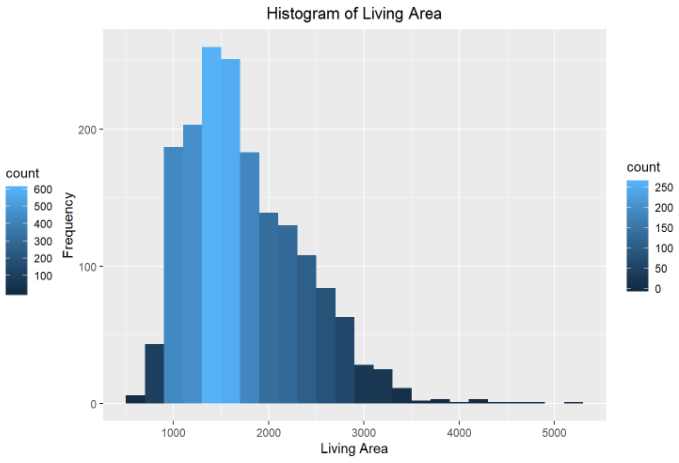


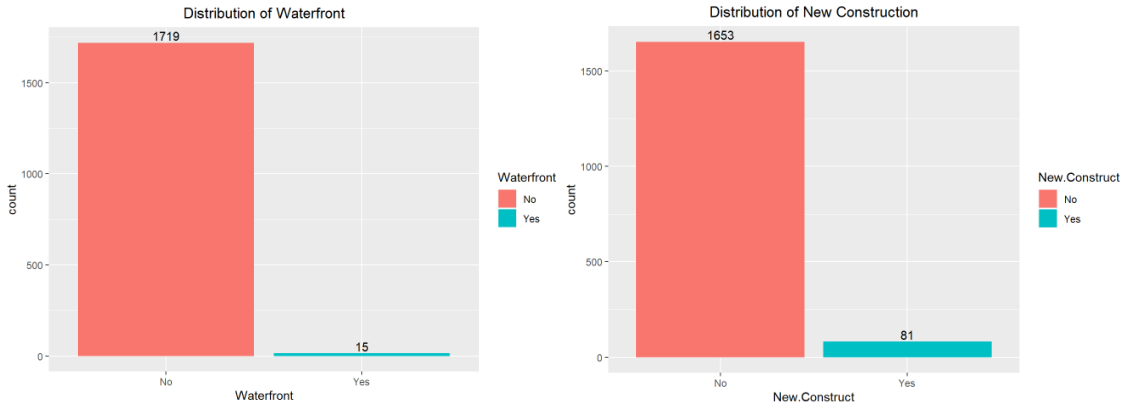



**Interpretation**:

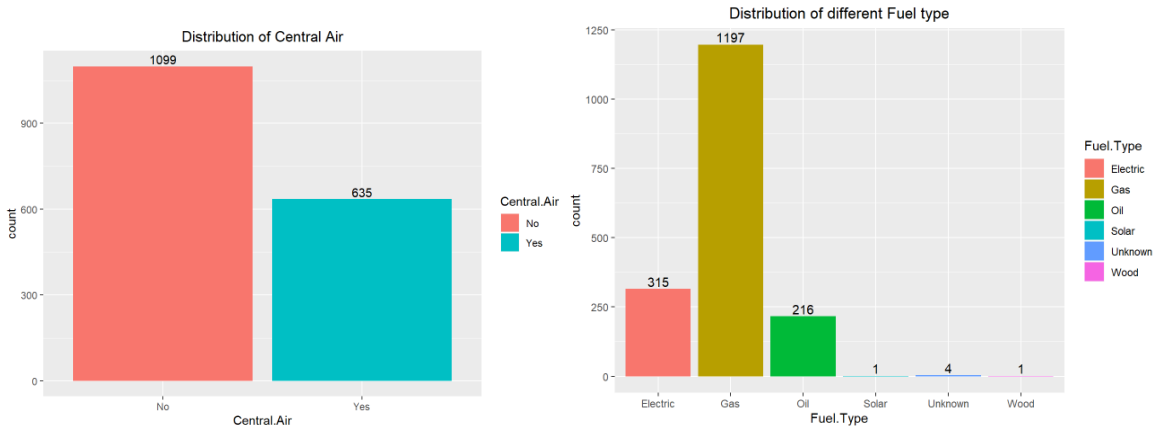
Price, Lot.Size, Age, Land.Value and Living.Area are right-skewed, where mean is greater than the median. Most of the data is on the left side and it has a longer tail on the right side. The average house price in Saratoga is $211710 with average values of lot size 0.5 square feet. The average house age is 28.26 years with land value as $34536 and living area as 1753 square feet. Pct.College data is not normally distributed. It shows a few spikes with the little left skew trend.

c. Bar chart for categorical variables:

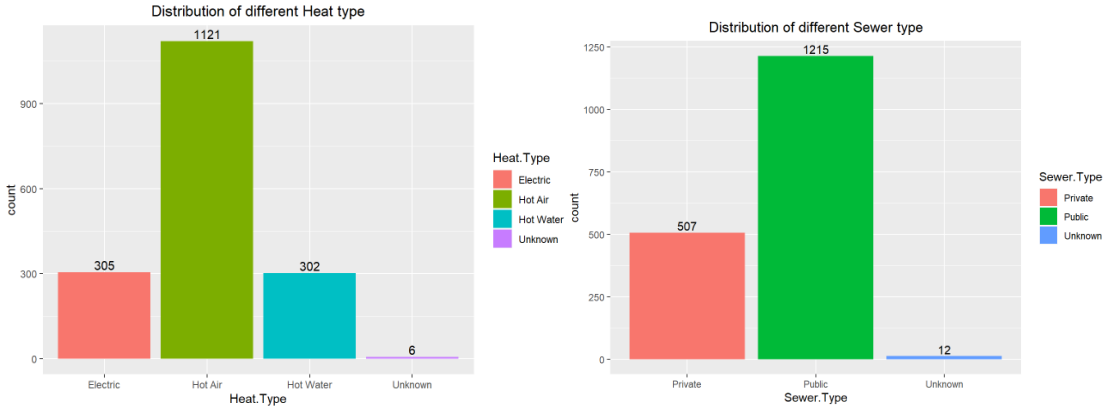
Bar chart for Waterfront and New.Construct,



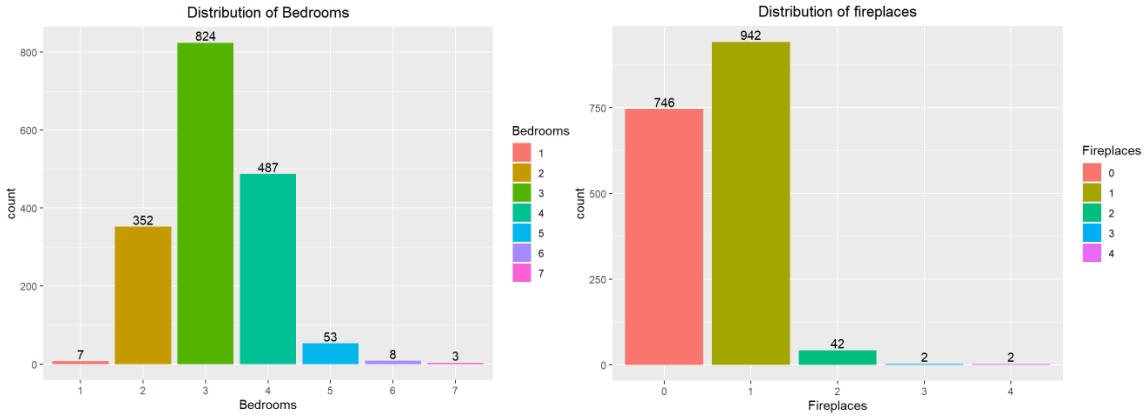

Bar chart for Central.Air and Fuel.Type



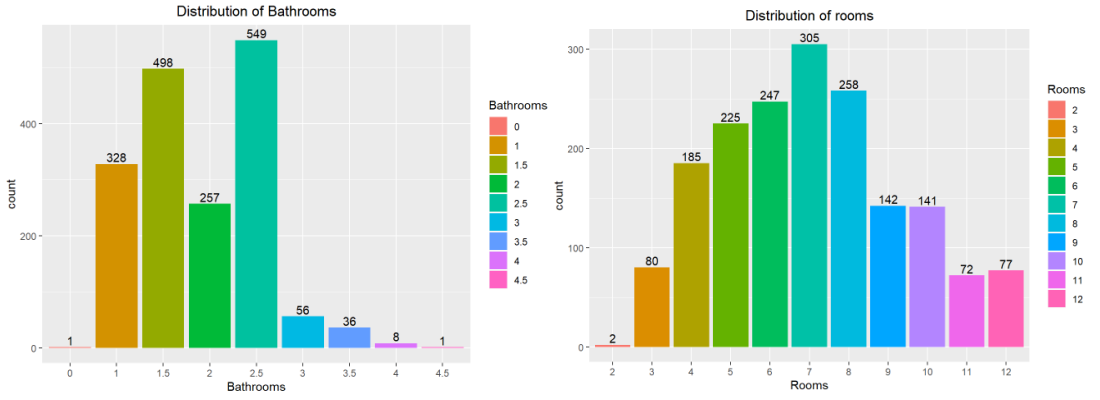

Bar chart for Heat.Type and Sewer.Type




Bar chart for Bedrooms and Fireplaces




Bar chart for Bathrooms and Rooms

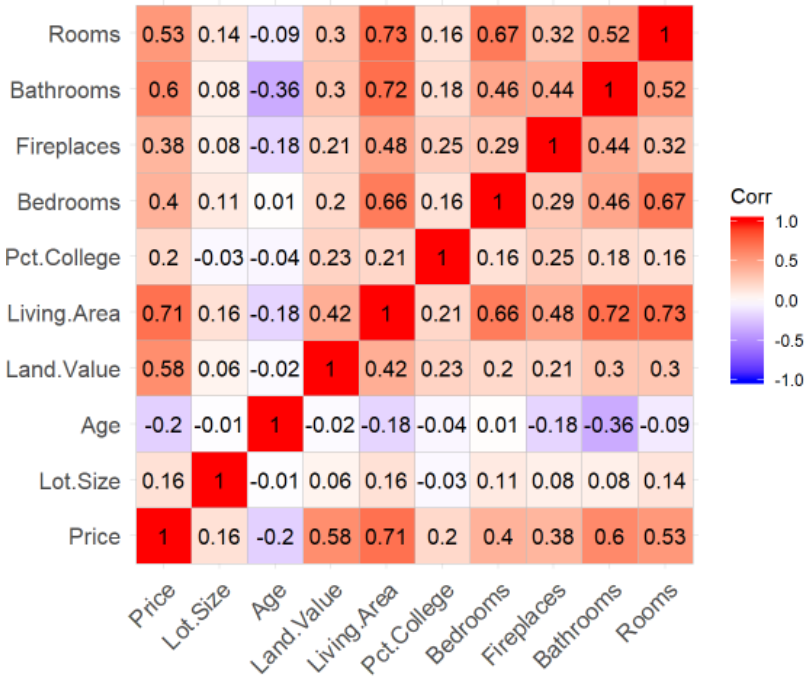



**Interpretation**:

**Waterfront**: In Saratoga, most of the houses (1719 houses) did not have a waterfront. **New.Construct:** Very few houses were newly constructed (81 houses) while many

houses were old (1653 houses). **Central.Air**: More than half of the houses (1099 houses) did not have a central air system. **Fuel.Type & Heat.Type:** Gas was widely used in houses (1197 houses) as fuel sources while hot air was used by a maximum number of houses (1121 houses) as heat sources. **Sewer.type**: Many houses (1215 houses) have public sewer system. **Bedrooms**: Most of the houses (824 houses) had 3 bedrooms followed by 4 and 2 bedrooms. There were a few houses which had 1, 6 or 7 bedrooms. **Fireplaces**: More than half of the houses (942 houses) had only one fireplace. There were quite a few houses with more than one fireplace. **Bathrooms**: In Saratoga, most of the houses (549 houses) had two and a half bathrooms followed by one and a half bathrooms. There were a smaller number of houses which have more than 3 bathrooms. **Rooms:** Many houses (305 houses) have 7 rooms. There is a smaller number of houses which have rooms less than 4 and more than 10.

**Correlation**



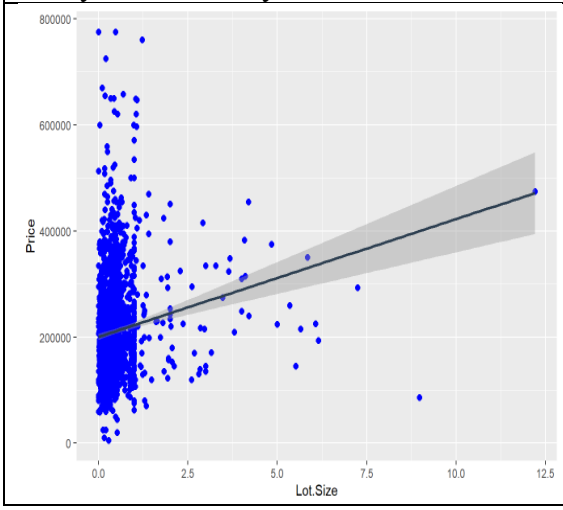

To understand which predictors are correlated with house price, we created the correlation matrix. Few predictors show a strong relationship (correlation more than 0.5) with

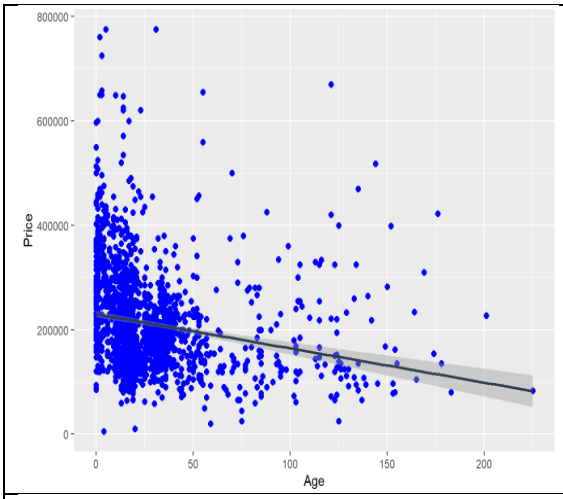
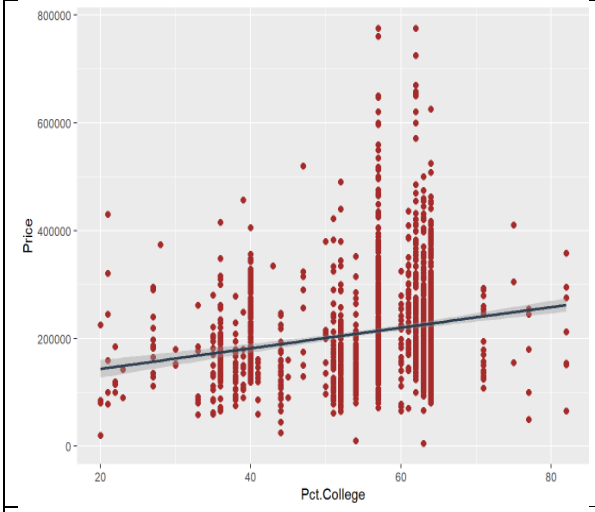
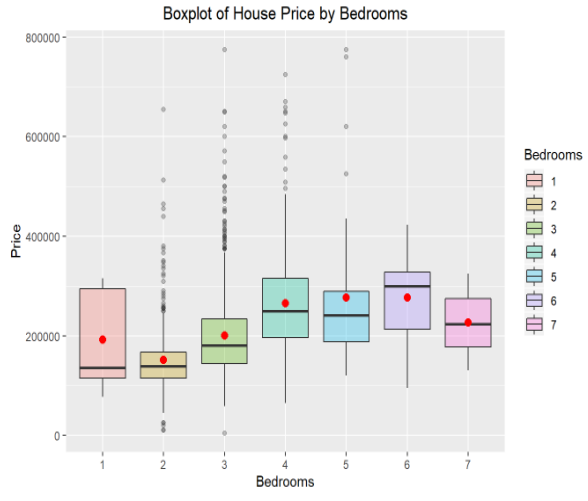
house prices. Living Area (0.71), Land Value (0.58), bathrooms (0.6) and rooms (0.53) showed a positive correlation and likely to be significant predictors. On the other hand, Rooms, Bedrooms, Bathrooms and Living Area are correlated to each other. Let's check the variance inflation factor to assess the multicollinearity level.

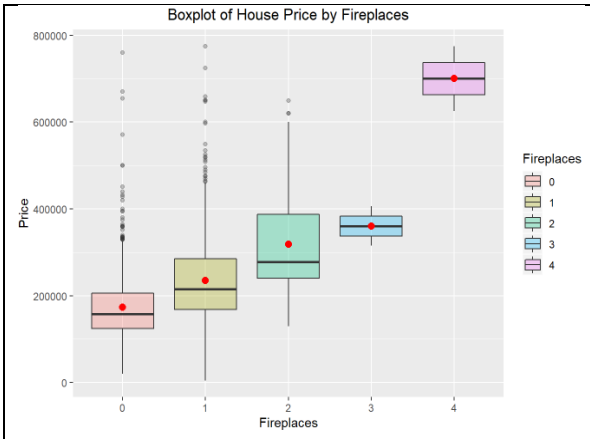
```
##               Lot.Size                    Age             Land.Value
##               1.035819               1.222638               1.276047
##            Living.Area            Pct.College as.numeric(Bedrooms)
##               4.108891               1.114403               2.147644
## as.numeric(Fireplaces) as.numeric(Bathrooms)     as.numeric(Rooms)
##               1.370883               2.402161               2.525071
```

VIF is a bit high for Living.Area. Other variables do not indicate multicollinearity.

Lot.Size, Age, Pct.College, Bedrooms, fireplaces have low correlation with Price. I have analyzed how price has been varied concerning these variables. They showed a low adjusted R square with a very low p-value. Let's look at below table:

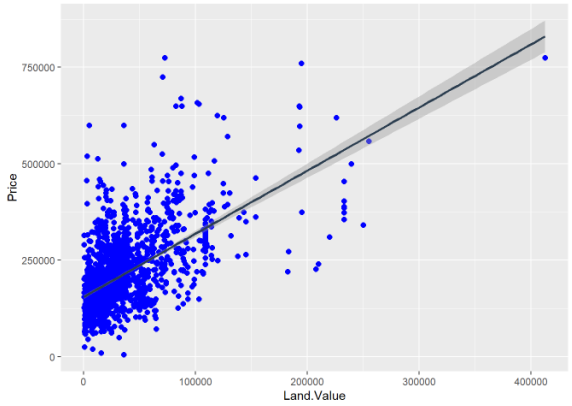
| Analysis of Price by other variables | Analysis and Values | Interpretation |
|---|---|---|
|  | model1=lm (Price~Lot.Size, data=HP)<br><br>Price increases with Lot size. But it does not show linear relation.<br><br>R square =0.024<br>P-value= 0.000000000055<br>Correlation=0.16 | Correlation value is small; Very low p-value; Only 2.4% of variation in house price can be explained by the model. |

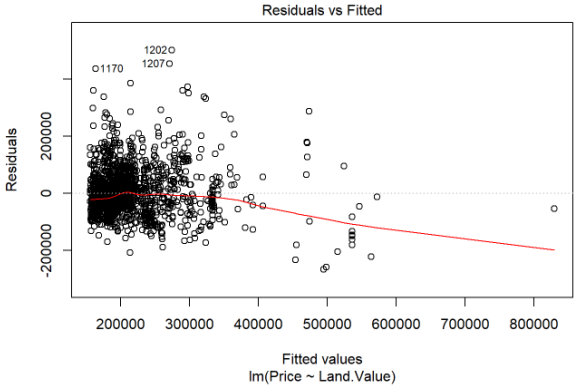
| | | |
|---|---|---|
|  | model2=lm (Price~Age, data=HP)<br><br>Price decreases with increases in Age. But it does not show linear relation.<br><br>R square =0.03744<br>P-value= 0.00000000000000002765<br>Correlation=-0.19 | Correlation value is small; Very low p-value; Only 3.7% of variation in house price can be explained by the model. |
|  | model3=lm (Price~Pct.College, data=HP)<br><br>Price increases with Pct.college. But it does not show linear relation.<br><br>R square =0.03908<br>P-value= 0.00000000000000022<br>Correlation=0.2 | Correlation value is small; Very low p-value; Only 3.9% of variation in house price can be explained by the model. |
|  | model4=lm (Price~Bedrooms, data=HP)<br><br>The most expensive houses are the one which have 5 and 6 bedrooms. 2 bedrooms houses are least expensive.<br><br>R square = 0.1766<br>P-value= 0.00000000000000022<br>Correlation=0.4 | Correlation value is small; Very low p-value; Only 17.66% of variation in house price can be explained by the model. |

| | | |
|---|---|---|
|  | model5=lm (Price~Fireplaces, data=HP)<br><br>House price increased with increase in number of fireplaces.<br><br>R square = 0.1509<br>P-value= 0.00000000000000022<br>Correlation=0.38 | Correlation value is small; Very low p-value; Only 15.09% of variation in house price can be explained by the model. |

Let's check the other variables which shows high correlation with Price.

**Analysis of Price ~ Land.Value**



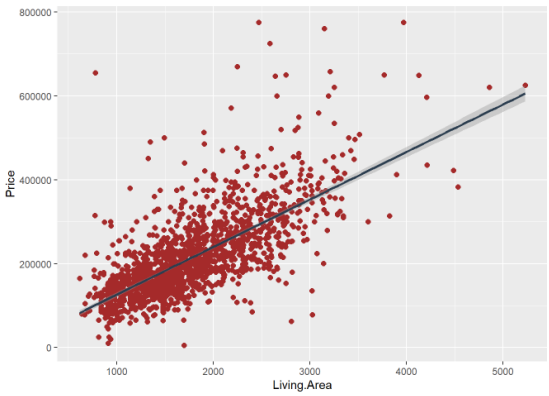

From the Scatter plot, the relationship between land value and house price is not linear. Residuals look random towards the left but start to form a fan-out pattern. The correlation is (0.58) suggests the variation in the house price is well explained by land value compared to other variables.

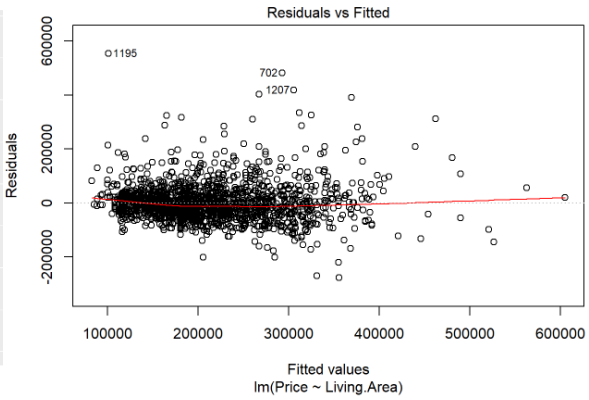
```
## Call:
## lm(formula = Price ~ Land.Value, data = HP)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -267746  -49152  -14017   36236  501092
##
## Coefficients:
##                Estimate   Std. Error t value            Pr(>|t|)
## (Intercept) 155234.41155   2709.83679   57.29 <0.0000000000000002 ***
## Land.Value       1.63463      0.05511   29.66 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 80220 on 1729 degrees of freedom
## Multiple R-squared:  0.3373, Adjusted R-squared:  0.3369
## F-statistic: 879.8 on 1 and 1729 DF,  p-value: < 0.00000000000000022
```

The p-value for intercept as well as Land.Value is very low. The F-statistics (879.8) is high with very low p-value suggest that land value is a significant predictor. On the other side, adjusted R-squared (0.3369) indicates that only 33.69% of variation in house price is explained by land value. Regression equation:

**Price=155234.41155+(1.63463) *Land.Value**

**Analysis of Price ~ Living.Area**



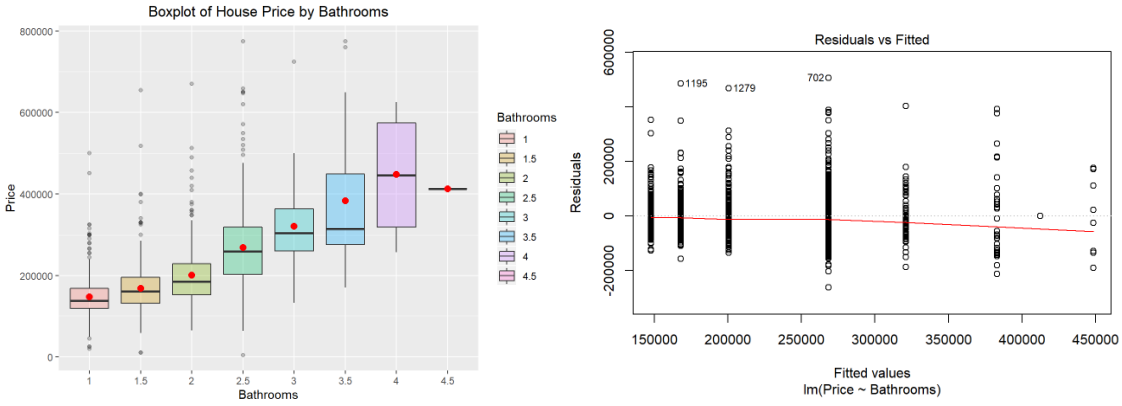

From the Scatter plot, the relationship looks linear. It suggests that if you increase the square feet of the living area, house prices will also increase. Residuals look linear towards the left. The correlation (0.71) suggests the variation of the house price is well explained by the living area.

```
##
## Call:
## lm(formula = Price ~ Living.Area, data = HP)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -277098  -39352   -7638   28354  553580
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 13069.90    4984.18   2.622              0.00881 **
## Living.Area   113.27       2.68  42.271 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69110 on 1729 degrees of freedom
## Multiple R-squared:  0.5082, Adjusted R-squared:  0.5079
## F-statistic:  1787 on 1 and 1729 DF,  p-value: < 0.00000000000000022
```

Both the p-value for intercept and Living.Area is low. The F-statistics is high (1787) with very low p-value suggest that the living area is a highly significant predictor. The Adjusted R-squared (0.5079) indicates that only 50.79% of the house price is explained by the living area. Regression equation:

**Price=13069.90+(113.27) *Living.Area**

**Analysis of Price ~ Bathrooms**

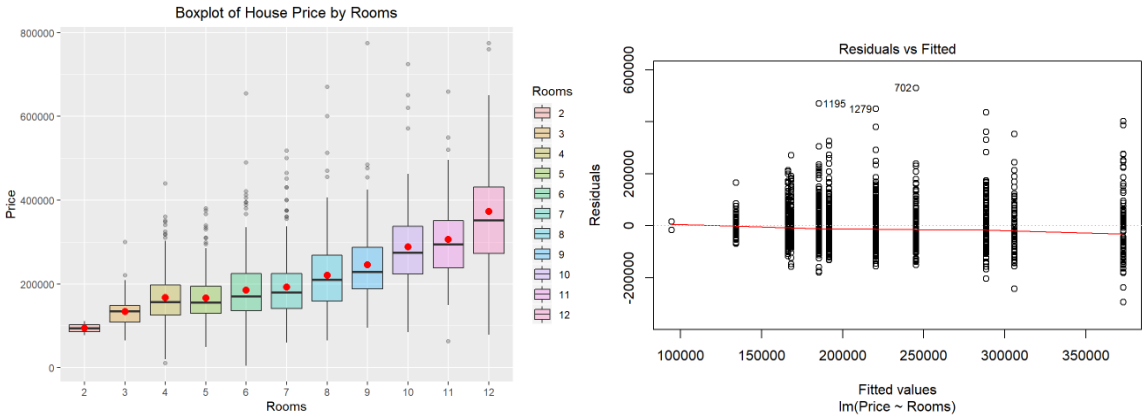



House prices increased with an increase in the number of bathrooms. Houses with the number of bathrooms as 4.5 is a bit cheaper than the houses with 4 bathrooms. Residual plot shows linear relationship with fitted values.

```
## Call:
## lm(formula = Price ~ Bathrooms, data = HP)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -263345  -46979   -9345   31655  506655
##
## Coefficients:
##              Estimate Std. Error t value             Pr(>|t|)
## (Intercept)    147751       4316  34.231 < 0.0000000000000002 ***
## Bathrooms1.5    20278       5558   3.649             0.000272 ***
## Bathrooms2      52861       6507   8.124 0.00000000000000849 ***
## Bathrooms2.5   120594       5452  22.118 < 0.0000000000000002 ***
## Bathrooms3     173228      11288  15.346 < 0.0000000000000002 ***
## Bathrooms3.5   235362      13706  17.172 < 0.0000000000000002 ***
## Bathrooms4     300699      27931  10.766 < 0.0000000000000002 ***
## Bathrooms4.5   264749      78172   3.387             0.000723 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78050 on 1723 degrees of freedom
## Multiple R-squared:  0.3748, Adjusted R-squared:  0.3723
## F-statistic: 147.6 on 7 and 1723 DF,  p-value: < 0.00000000000000022
```

The p-value for intercept and Bathroom is low. The F-statistics (147.6) with very low p-value suggests that the Bathroom is a significant predictor. The Adjusted R-squared (0.3723) indicates that only 37.23 % of the variation in house price is explained by the number of bathrooms.

**Analysis of Price ~Rooms**




House price increased with increase in number of rooms. Residual plot shows linear relationship with fitted values.

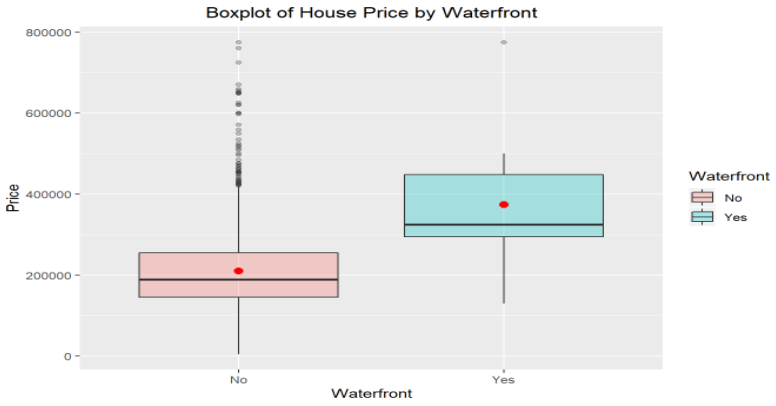
```
## Call:
## lm(formula = Price ~ Rooms, data = HP)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -294719  -50850  -11450   36525  529721
##
## Coefficients:
##              Estimate Std. Error t value   Pr(>|t|)
## (Intercept)     94500      57878   1.633   0.102708
## Rooms3          39656      58597   0.677   0.498651
## Rooms4          73745      58194   1.267   0.205242
## Rooms5          71950      58136   1.238   0.216032
## Rooms6          90814      58112   1.563   0.118300
## Rooms7          97329      58068   1.676   0.093895 .
## Rooms8         126097      58102   2.170   0.030124 *
## Rooms9         150779      58285   2.587   0.009765 **
## Rooms10        194067      58287   3.329   0.000888 ***
## Rooms11        211414      58677   3.603   0.000323 ***
## Rooms12        278719      58625   4.754 0.00000216 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 81850 on 1720 degrees of freedom
## Multiple R-squared:  0.3137, Adjusted R-squared:  0.3097
## F-statistic: 78.61 on 10 and 1720 DF,  p-value: < 0.00000000000000022
```

The p-value for intercept and rooms is more than the significance level. House with rooms8 and above shows low p-value. The F-statistics (78.61) with very low p-value suggests that house with rooms 8,9,10, 11 and 12 is a significant predictor but all rooms are not significant for predicting the house price. The Adjusted R-squared (0.3097) indicates that only 30.97 % of variation in house price is explained by the number of rooms.

**t-test at 0.05 significance level**

Null hypothesis: There is no significant difference between the means of house prices between the two groups. Alternative hypothesis: There is a significant difference between the means of house prices between the two groups.

**t.test(Price~Waterfront)**

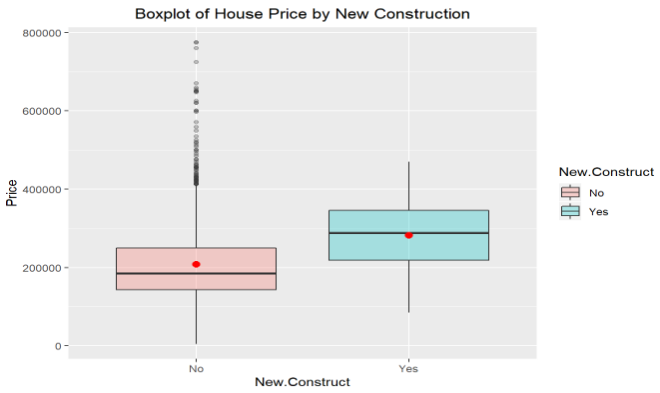


In Saratoga, house with waterfront is costlier (with average price $ 373991.7) than the houses without waterfront (with average price $ 210291.8). We would like to check if the waterfront really makes a significant difference in the house price.

```
##
##  Welch Two Sample t-test
##
## data:  Price by Waterfront
## t = -4.0863, df = 14.096, p-value = 0.001097
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -249803.60  -77908.08
## sample estimates:
## mean in group 0 mean in group 1
##        210135.8        373991.7
```

With the t-test result, we reject the null hypothesis and accept the alternative hypothesis. With t statistics (-4.0863) and p-value (0.001097), we can interpret that means of house price with waterfront is significantly different from the house price where waterfront is not present.

**t.test(Price~New.Construct)**

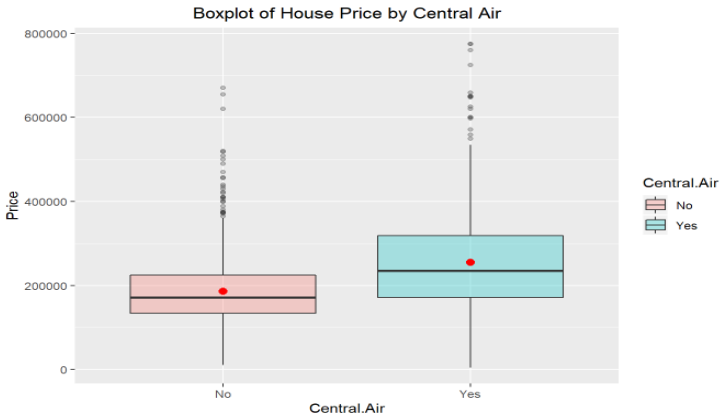



New house in Saratoga more expensive than the old houses. The average price of a new house is $ 282306.8 while the old house has mean value as $ 208244.7.

```
##
##  Welch Two Sample t-test
##
## data:  Price by New.Construct
## t = -7.7077, df = 91.031, p-value = 0.00000000001534
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -93349.61 -55094.07
## sample estimates:
## mean in group 0 mean in group 1
##        208085.0        282306.8
```

With the t-test result, we reject the null hypothesis and accept the alternative hypothesis. With t statistics (-7.7077) and low p-value (less than 0.05), we can interpret that there is a significant difference in means of new and old house price.

**t.test(Price~Central.Air)**



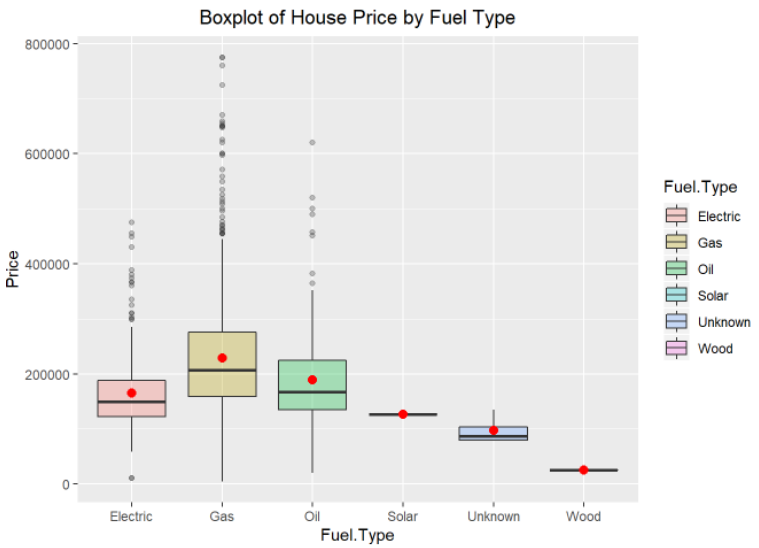

Houses with a central air system have higher price than the houses that do not have central air system. The average house price without central air system is $186684.9 while $254903.8 is the mean price if houses have a central air system.

```
##
##  Welch Two Sample t-test
##
## data:  Price by Central.Air
## t = -13.425, df = 987.42, p-value < 0.0000000000000022
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -78421.04 -58418.71
## sample estimates:
## mean in group 0 mean in group 1
##        186483.9        254903.8
```

With the t-test result, we reject the null hypothesis and accept the alternative hypothesis. The t statistics (-13.425) with very low p-value indicate that mean of house price with a central air system is significantly different from the mean of house price without central air system.

**ANOVA**

**Houses price by fuel type**




Saratoga houses use different types of fuel such as gas, oil, etc. Among all fuel type, the houses which used gas as fuel has the highest price. Oil and electric ones are less expensive than gas. To find whether different fuel type makes a significant impact on house price, we conducted ANOVA test:

Average house price by different fuel type is as follows:

```
##  Electric       Gas       Oil   Unknown
## 164937.57 228562.38 188734.40  97006.25
```

ANOVA result at 0.05 significance level:

```
##                Df         Sum Sq      Mean Sq F value               Pr(>F)
## Fuel.Type       3  1195429652300 398476550767   44.13 <0.0000000000000002
## Residuals    1727 15594921381251   9030064494
```

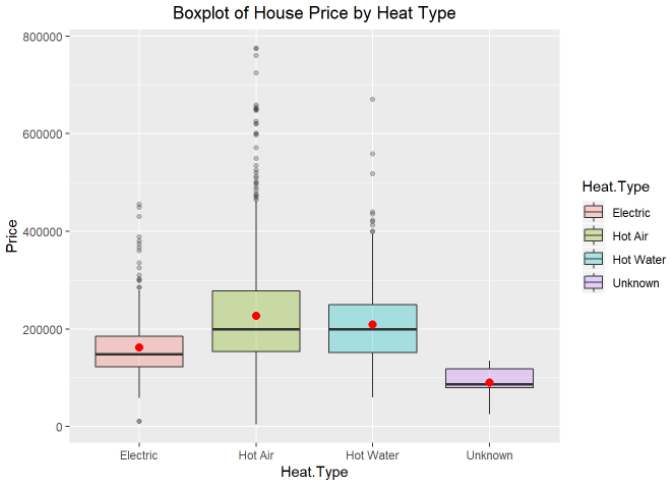
```
##      (Intercept)     Fuel.TypeGas     Fuel.TypeOil Fuel.TypeUnknown
##        164937.57         63624.81         23796.83        -67931.32
```

ANOVA result shows F-statistics as 44.13 with very low p-value. It indicates a statistically significant difference in means of price if fuel sources are different. Regression equation is

**Price =164937.57 + 63624.81(Fuel.TypeGas) + 23796.83(Fuel.TypeOil) - 67931.32 (Fuel.TypeUnknown)**

The price of a house with Gas and Oil fuel is comparatively $63624.81 and $23796.83 more than the house with an Electric fuel. Also, the price of a house with unknown fuel is $67931.32 less than the house with an Electric fuel.

**Houses price by Heat type**




Hot air is extensively used among all heat types with the highest house price followed by Hot water and electric. To find out the statistical significance of average house prices based on different heat type, we conducted ANOVA test:

Average house price by different fuel type is as follows:

```
##  Electric   Hot Air Hot Water   Unknown
## 161888.63 226382.62 209132.46  97006.25
```

ANOVA result at 0.05 significance level:

```
##              Df         Sum Sq      Mean Sq F value              Pr(>F)
## Heat.Type     3  1052815863245 350938621082   38.51 <0.0000000000000002
## Residuals  1727 15737535170305   9112643411
```

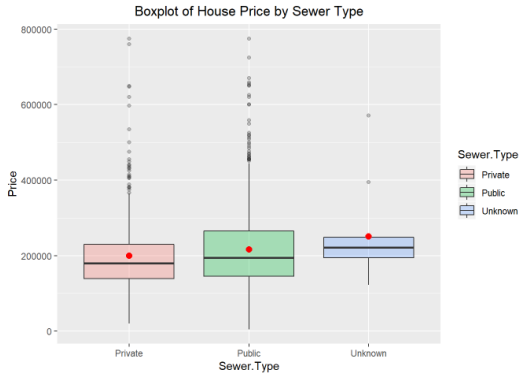
```
##         (Intercept)   Heat.TypeHot Air Heat.TypeHot Water
##           161888.63           64493.99           47243.83
##    Heat.TypeUnknown
##           -64882.38
```

ANOVA result shows F-statistics as 38.15 with very low p-value. It indicates a statistically significant difference in means of price if heat type is different. Regression equation

is **Price =161888.63 + 64493.99(Heat.TypeHot Air) + 47243.83 (Heat.TypeHot Water) - 64882.38 (Heat.TypeUnknown)**

The price of a house with Hot Air and Hot Water is comparatively $64493.99 and $47243.83 more than the house with an Electric heat. Also, the price of a house with Unknown heat is $64882.38 less than the house with an Electric heat.

**Houses price by Sewer type**




Houses with public sewer type are costlier than the other than the type of sewer systems. To find out the statistical significance of house prices based on different sewer types, we conducted an ANOVA test. Average house price by different fuel type is as follows:

```
##  Private   Public  Unknown
## 199597.0 216375.2 250952.3
```

ANOVA result at 0.05 significance level:

```
##              Df         Sum Sq     Mean Sq F value  Pr(>F)
## Sewer.Type    2   119121637669 59560818834   6.174 0.00213 **
## Residuals  1728 16671229395882  9647702197
```

```
##       (Intercept)  Sewer.TypePublic Sewer.TypeUnknown
##         199597.01          16778.16          51355.33
```

ANOVA result shows a low F-statistics as 6.174 with p-value as 0.00213 less than the significance level (0.05). It indicates a statistically significant difference in means of price if

sewer type is different. Regression equation is **Price = 199597.01 + 16778.16 *(Sewer.TypePublic) +51355.33 *(Sewer.TypeUnknown)**

The price of a house with public sewer and unknown sewer system is comparatively $16778.16 and $51355.33 more than the house with private sewer system.

**Linear Regression**

Multiple regression is used to find the final equation to predict house prices. As per correlation analysis, we are using only significant variables that showed a correlation with the response variable above 0.5. Model1 starts with considering Land.Value, Living.Area, Bathrooms, and Rooms. Model2 considers only Land.Value, Living.Area, Bathrooms and eliminate the Rooms variable while Model3 Land.Value, Living.Area, eliminate Bathrooms, Rooms variable. The significance level is 0.05 for all regression analysis.

**Model1: Price~ Land.Value + Living.Area + Bathrooms + Rooms**

```
## Call:
## lm(formula = Price ~ Land.Value + Living.Area + Bathrooms + Rooms,
##     data = HP)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -220736  -35720   -6107   28504  457895
##
## Coefficients:
##                 Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)  24825.25652 43014.12234   0.577               0.5639
## Land.Value       0.95881     0.04629  20.712 < 0.0000000000000002 ***
## Living.Area     67.46644     4.68388  14.404 < 0.0000000000000002 ***
## Bathrooms1.5  1677.97579  4491.69855   0.374               0.7088
## Bathrooms2   26553.62811  5267.60117   5.041      0.0000005123389 ***
## Bathrooms2.5 32673.01750  5553.74474   5.883      0.0000000048343 ***
## Bathrooms3   49127.70068  9916.16941   4.954      0.0000007976127 ***
## Bathrooms3.5 81669.26773 12460.23993   6.554      0.0000000000737 ***
## Bathrooms4   40977.90717 24100.84532   1.700               0.0893 .
## Bathrooms4.5 15921.28166 61701.91746   0.258               0.7964
## Rooms3        9721.93717 43478.88320   0.224               0.8231
## Rooms4       18230.61419 43209.73475   0.422               0.6731
## Rooms5       11032.54585 43198.55445   0.255               0.7985
## Rooms6       19219.74637 43184.54016   0.445               0.6563
## Rooms7       14479.55576 43176.74178   0.335               0.7374
## Rooms8       20716.82258 43259.57426   0.479               0.6321
## Rooms9       10091.37351 43507.16571   0.232               0.8166
## Rooms10      25262.03373 43612.66065   0.579               0.5625
## Rooms11      24218.91394 44021.46186   0.550               0.5823
## Rooms12      24528.64878 44314.69881   0.554               0.5800
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60630 on 1711 degrees of freedom
## Multiple R-squared:  0.6254, Adjusted R-squared:  0.6213
## F-statistic: 150.4 on 19 and 1711 DF,  p-value: < 0.00000000000000022
```

The p-value is low for Intercept, Land.Value, Living.Area. Bathrooms with 1.5, 4 and 4.5 show insignificant p-value. All the rooms have a p-value higher than the significance level. The

adjusted R-square is 0.6213 means 62.13% variation of house price can be explained by this model. We will eliminate Rooms as it is an insignificant predictor.

**Model2: Price~ Land.Value + Living.Area + Bathrooms**

```
## Call:
## lm(formula = Price ~ Land.Value + Living.Area + Bathrooms, data = HP)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -218676  -34780   -6004   28300  464106
##
## Coefficients:
##                 Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)  36449.67197  5450.93655   6.687      0.0000000000307 ***
## Land.Value       0.95584     0.04616  20.708 < 0.0000000000000002 ***
## Living.Area     71.00893     3.69160  19.235 < 0.0000000000000002 ***
## Bathrooms1.5   605.56407  4437.33361   0.136                0.8915
## Bathrooms2   25603.15485  5202.53630   4.921      0.0000009419736 ***
## Bathrooms2.5 32292.06265  5484.66552   5.888      0.0000000046983 ***
## Bathrooms3   48811.90798  9869.02259   4.946      0.0000008315425 ***
## Bathrooms3.5 82548.55485 12226.62423   6.752      0.0000000000199 ***
## Bathrooms4   40523.12361 23641.26148   1.714                0.0867 .
## Bathrooms4.5 15285.53485 61440.36902   0.249                0.8036
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60620 on 1721 degrees of freedom
## Multiple R-squared:  0.6234, Adjusted R-squared:  0.6214
## F-statistic: 316.5 on 9 and 1721 DF,  p-value: < 0.00000000000000022
```

The p-value is low for Intercept, Land.Value, Living.Area. The p-value for Bathrooms with 1.5, 4 and 4.5 is higher than 0.05 (level of significance) . The adjusted R-square is improved and showed value as 0.6214. It means 62.14% variation in house price is explained by this model. We will eliminate Bathrooms as it is an insignificant predictor.

**Model3: Price~ Land.Value + Living.Area**

```
## Call:
## lm(formula = Price ~ Land.Value + Living.Area, data = HP)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -241131  -37208   -6267   28046  465813
##
## Coefficients:
##                Estimate  Std. Error t value             Pr(>|t|)
## (Intercept) 20075.72881  4491.81120   4.469           0.00000835 ***
## Land.Value      0.95713     0.04707  20.333 < 0.0000000000000002 ***
## Living.Area    90.41836     2.65719  34.028 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62100 on 1728 degrees of freedom
## Multiple R-squared:  0.6032, Adjusted R-squared:  0.6027
## F-statistic:  1313 on 2 and 1728 DF,  p-value: < 0.00000000000000022
```

The p-value is very low for Intercept, Land.Value, Living.Area. The adjusted R-square is 0.6027 means 60.27% variation in house price is explained by this model. Though R squared is

less from the previous model, we don't have any insignificant predictors. The regression equation is:

**Price = 20075.72881 + (0.95713) *Land.Value + (90.41836) *Living.Area**

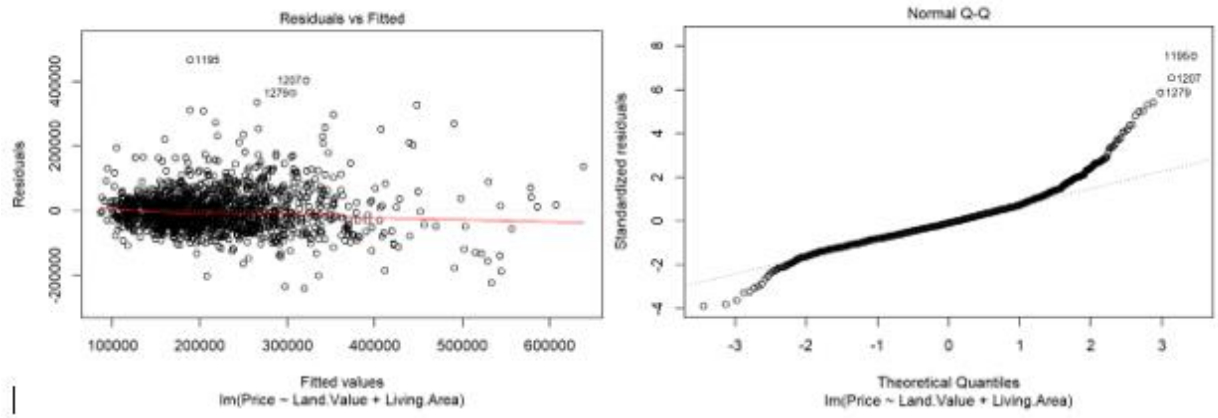
Let's compare Model2 and Model3

```
## Analysis of Variance Table
##
## Model 1: Price ~ Land.Value + Living.Area + Bathrooms
## Model 2: Price ~ Land.Value + Living.Area
##   Res.Df           RSS Df     Sum of Sq      F               Pr(>F)
## 1   1721 6324003469402
## 2   1728 6663044967528 -7 -339041498126 13.181 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA result shows F-statistics (13.181) with very low p-value. We can interpret that Model 3, based on ANOVA results, is statistically an improvement over Model 2. The final regression equation:

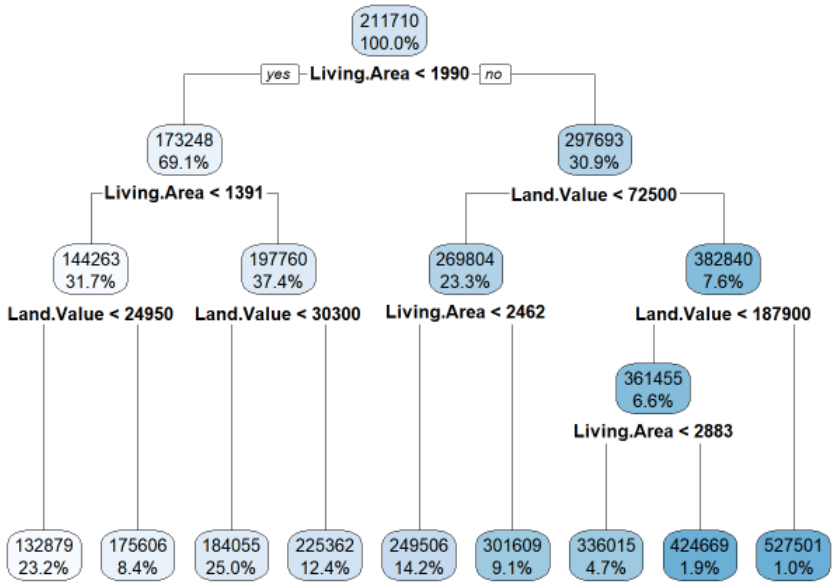
**Price = 20075.72881 + (0.95713) *Land.Value + (90.41836) *Living.Area**

Analysis of Residual plot and Normal Q-Q plot




The fitted line is approximate straight line while Normal Quantile-Quantile is improved but not perfect straight line. It shows errors are not normally distributed.

To visualize the predicted house price, we created the decision tree.

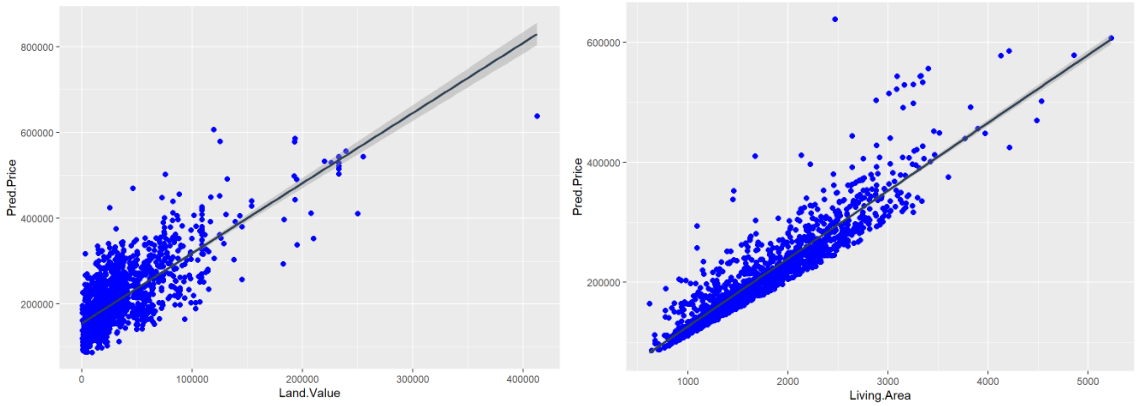



The decision tree clearly explains how the living area and land value can help to predict the price. For instance, if the living area is less than 1990 square feet, house prices would be $173248 with a probability of 69.1%.

To predict the house price with the help of Model3, we took an example of 179th row.

Actual house price: $247000; Predicted house price: $246251.2

To check how predicted house price varied concerning the Land.Value and Living.Area, we drew the scatter plot:



The plot shows that house prices increased with an increase in the living area as well as land value.

**Conclusion**

The statistical analysis shows that the house price of Saratoga County increased where newly constructed houses were available with a waterfront, central air system with the private sewer system. Also, houses that used gas as fuel sources and hot air as a heat source were the most expensive. Besides, the price was increased with an increase in rooms, bathrooms, fireplaces, and bedrooms. Houses with 4 fireplaces, 5 bedrooms, 12 rooms, 4 bathrooms were the most expensive houses.

On the other hand, house price was not affected by graduated people lived around the house area and lot size. Though new houses were expensive, house age was not helpful to predict the house price. In Saratoga County, people can get the idea of the land value and living area. This will be helpful to predict the house price for both sellers and buyers.

References

Burke, J. (2019). *Zillow vs. Realtor vs. Redfin*. Retrived from https://www.investopedia.com/articles/personal-finance/060516/zillow-realtor-redfin-which-best-real-estate-website-z-nws.asp

Wikipedia. (n.d.). *Saratoga County, New York.* Retrieved from https://en.wikipedia.org/wiki/Saratoga_County, New_York

Wood, R. (2004). *Booming real estate market reflects Saratoga County's allure.* Retrieved from https://www.bizjournals.com/albany/stories/2004/06/14/focus1.html