DSC520 Final Project - House Price Prediction

Ganesh Kale

June 4, 2021

Introduction:

House Prices in USA are booming and house prices will continue to race ahead, at nearly twice the pace predicted before this year. This is what we hear or read when talk about housing market in USA. Buying house is very critical job, one should aware of lots of things before buying house and when buying house nobody sure about when is the right time to buy house and wants to have some tool that would consider all the factors determining house price and predict the house price. Predicting the house price is challenging but doable and with help of machine learning algorithms this can be achieved. This is the topic for final project - predicting the house price based on the house features, area features. The data sets I obtained are from houses in Austin TX. Housing market in USA continue to rise in all of the country's major cities. The Austin TX market is also one of the hottest market in 2021.

Data Set Information:

The data set has total 47 unique features and it was uploaded with house images for each house listed in the data set and given the reference of each house to the data set as an additional feature. For this project purpose, I have not included the house images since the research and analysis would be completely based on the features provided for each house instead of the photos. There are three files have been created and main file has all the information about houses, second file has information about the school area where the house is located and third file provides the house features information, such as what are the features available in each listed house.

Few features form the data sets - address of the property, number of garage spaces, year built, number of bedrooms, bathrooms, area of house, area of loft , school zone, appliances, view etc.

Problem Statement:

In order to predict the house prices from given house features and area features, the main business problems are:

1. Identify the features that impact house Price.
2. Predict the house prices based on identified features.

Approach to Solve Business Problems:

To tackle the above mentioned business problems, the approach is -

1. Exploratory Data Analysis - identify missing data, duplicate data, outliers and fix them, check distributions of the features, check correlations of features, transforming the data, finalizing the predictors.
2. Modeling - Fitting the MLR model on house data set and displaying the result summary
3. Result Evaluation - Evaluating the model result by testing model on test dataset and validating the result.

Loading Required Packages:

```
library(dplyr)
library(ggplot2)
library(caret)
library(QuantPsyc)
library(car)
library(plyr)
library(tidyr)
```

Loading the data sets and merging all three data sets into one data frame:

Austin TX housing main data set with property information

```
house_main <- read.csv("project data/austinHousingData.csv")
```

Austin TX housing area school infromation data set

```
house_school <-  read.csv("project data/austinHousingData_school_Info.csv")
```

Austin TX house features data set

```
house_features <- read.csv("project data/austinHousingData_features.csv")
```

Joining all three datasets into one data set and display first 5 rows of final data set

```
##         zpid         city         streetAddress zipcode
## 1  111373431 pflugerville  14424 Lake Victor Dr   78660
## 2  120900430 pflugerville    1104 Strickling Dr   78660
## 3 2084491383 pflugerville   1408 Fort Dessau Rd   78660
## 4  120901374 pflugerville    1025 Strickling Dr   78660
## 5   60134862 pflugerville 15005 Donna Jane Loop   78660
##
```

```
## 1
## 2
## 3 Under construction - estimated completion in August 2019.  The Pioneer features an expansive open
## 4
## 5
##   latitude longitude propertyTaxRate garageSpaces hasAssociation hasCooling
## 1 30.43063 -97.66308            1.98            2           TRUE       TRUE
## 2 30.43267 -97.66170            1.98            2           TRUE       TRUE
## 3 30.40975 -97.63977            1.98            0           TRUE       TRUE
## 4 30.43211 -97.66166            1.98            2           TRUE       TRUE
## 5 30.43737 -97.65686            1.98            0           TRUE       TRUE
##   hasGarage hasHeating hasSpa hasView      homeType parkingSpaces yearBuilt
## 1      TRUE       TRUE  FALSE   FALSE Single Family             2      2012
## 2      TRUE       TRUE  FALSE   FALSE Single Family             2      2013
## 3     FALSE       TRUE  FALSE   FALSE Single Family             0      2018
## 4      TRUE       TRUE  FALSE   FALSE Single Family             2      2013
## 5     FALSE       TRUE  FALSE   FALSE Single Family             0      2002
##   latestPrice numPriceChanges latest_saledate latest_salemonth latest_saleyear
## 1      305000               5          9/2/19                9            2019
## 2      295000               1        10/13/20               10            2020
## 3      256125               1         7/31/19                7            2019
## 4      240000               4          8/8/18                8            2018
## 5      239900               3        10/31/18               10            2018
##                                 latestPriceSource lotSizeSqFt livingAreaSqFt
## 1 Coldwell Banker United, Realtors - South Austin        6011           2601
## 2                                   Agent Provided        6185           1768
## 3                                   Agent Provided        7840           1478
## 4                                   Agent Provided        6098           1678
## 5                                   Agent Provided        6708           2132
##   numOfBathrooms numOfBedrooms numOfStories numOfPhotos
## 1              3             4            2          39
## 2              2             4            1          29
## 3              2             3            1           2
## 4              2             3            1           9
## 5              3             3            2          27
##   numOfAccessibilityFeatures numOfAppliances numOfParkingFeatures
## 1                          0                5                    2
## 2                          0                1                    2
## 3                          0                4                    1
## 4                          0                0                    2
## 5                          0                0                    1
##   numOfPatioAndPorchFeatures numOfSecurityFeatures numOfWaterfrontFeatures
## 1                          1                     3                      0
## 2                          0                     0                      0
## 3                          0                     1                      0
## 4                          0                     0                      0
## 5                          0                     0                      0
##   numOfWindowFeatures numOfCommunityFeatures
## 1                   1                     0
## 2                   0                     0
## 3                   0                     0
## 4                   0                     0
## 5                   0                     0
##                                     homeImage numOfPrimarySchools
```

```
## 1  111373431_ffce26843283d3365c11d81b8e6bdc6f-p_f.jpg                   1
## 2  120900430_8255c127be8dcf0a1a18b7563d987088-p_f.jpg                   1
## 3 2084491383_a2ad649e1a7a098111dcea084a11c855-p_f.jpg                   0
## 4  120901374_b469367a619da85b1f5ceb69b675d88e-p_f.jpg                   1
## 5   60134862_b1a48a3df3f111e005bb913873e98ce2-p_f.jpg                   1
##   numOfElementarySchools numOfMiddleSchools numOfHighSchools avgSchoolDistance
## 1                      0                  1                1          1.266667
## 2                      0                  1                1          1.400000
## 3                      2                  1                1          1.200000
## 4                      0                  1                1          1.400000
## 5                      0                  1                1          1.133333
##   avgSchoolRating avgSchoolSize MedianStudentsPerTeacher
## 1        2.666667          1063                       14
## 2        2.666667          1063                       14
## 3        3.000000          1108                       14
## 4        2.666667          1063                       14
## 5        4.000000          1223                       14
```

Exploratory Data Analysis:

Cleaning data by removing unwanted columns such as property dscription, city name, address, home images, number of photos etc.

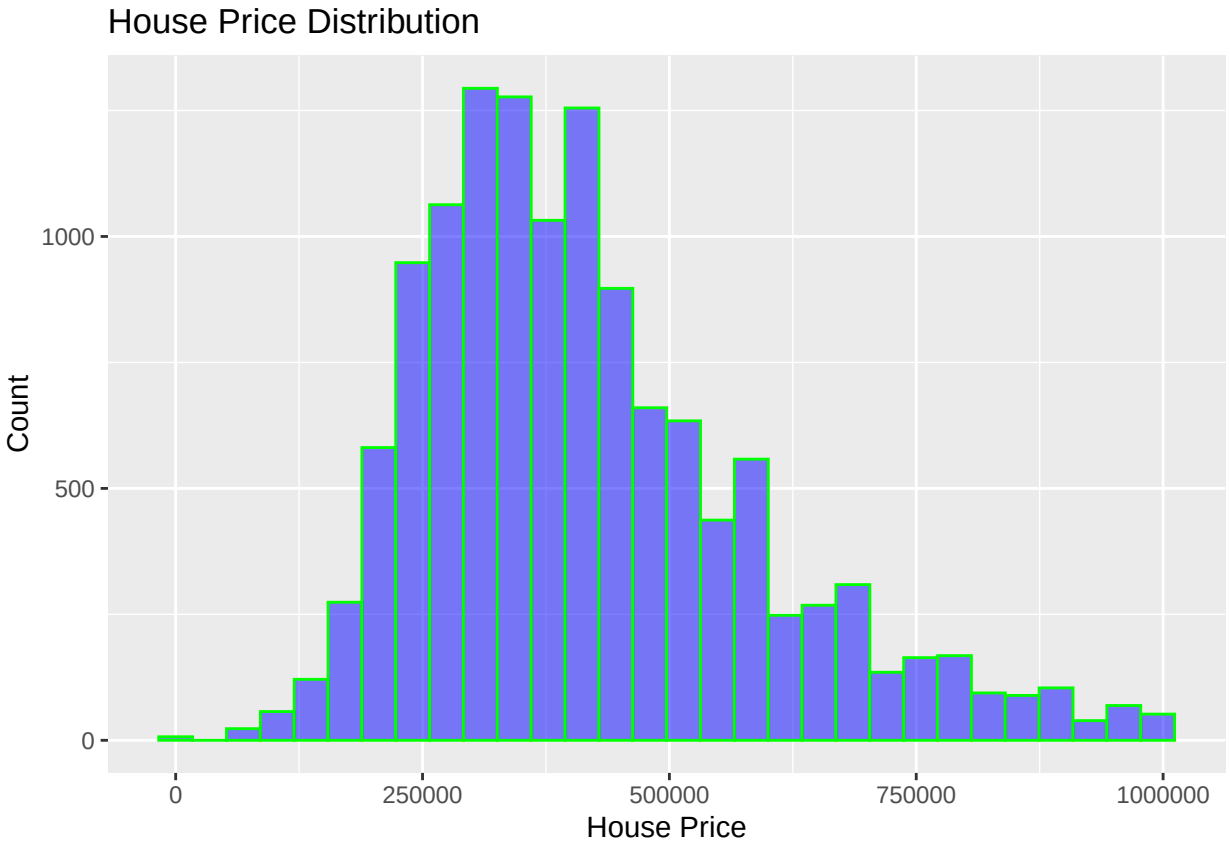
Created new column called property age from year built

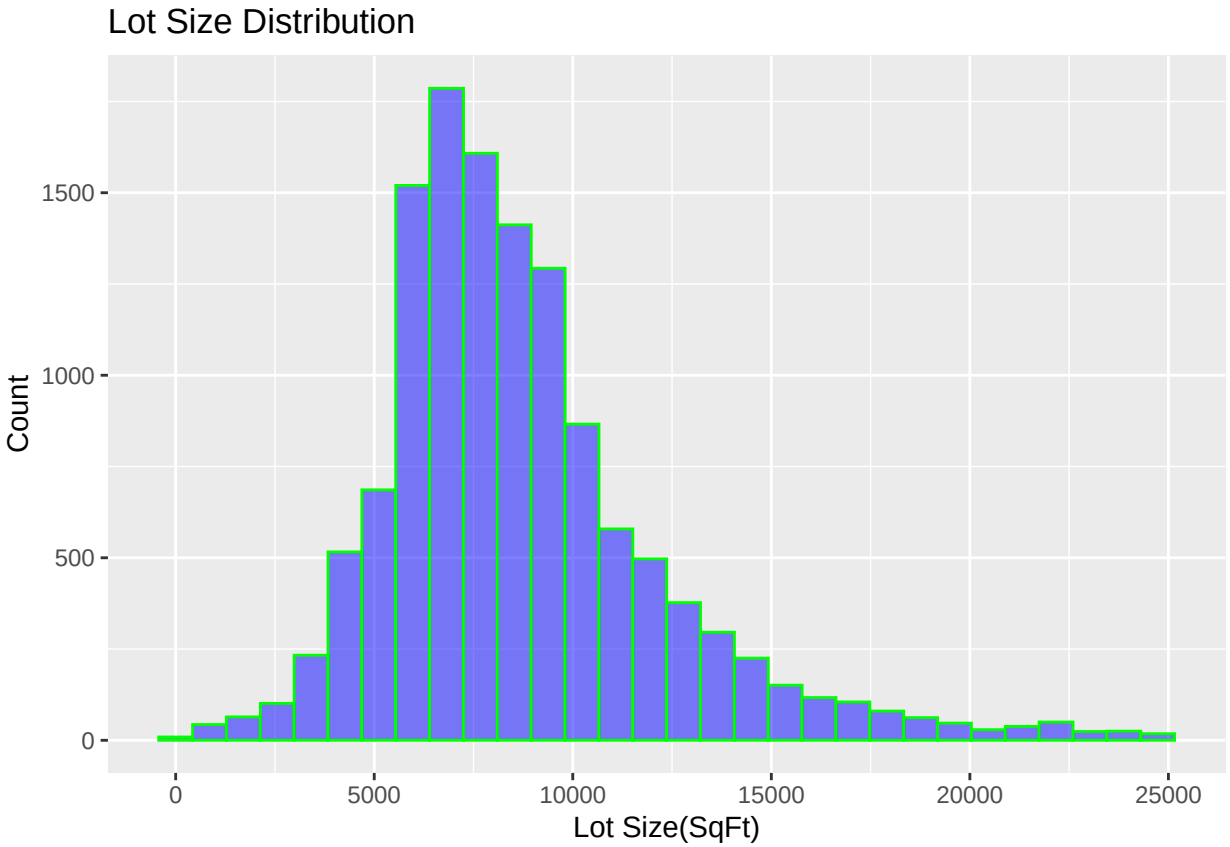
View na or null values in data set

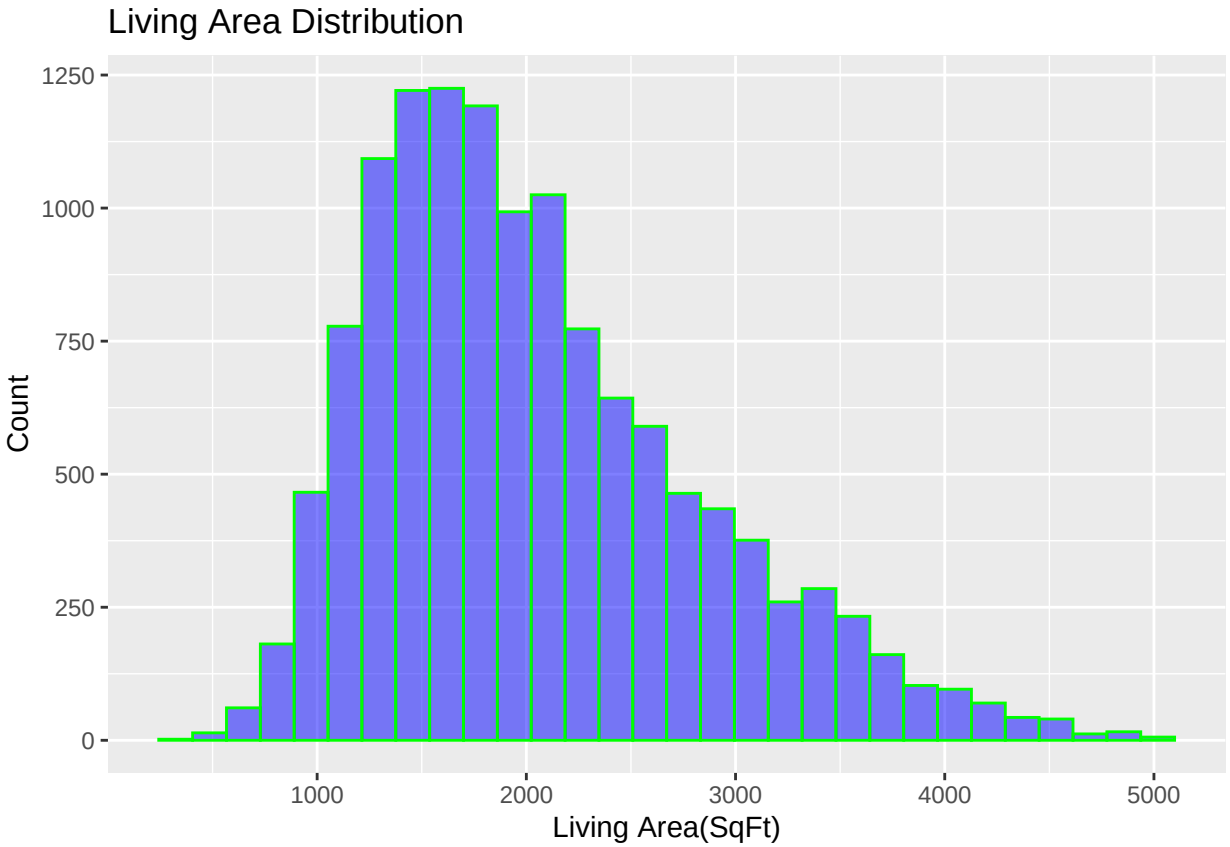
Based on the below, we can see there are no na or null values in the data set

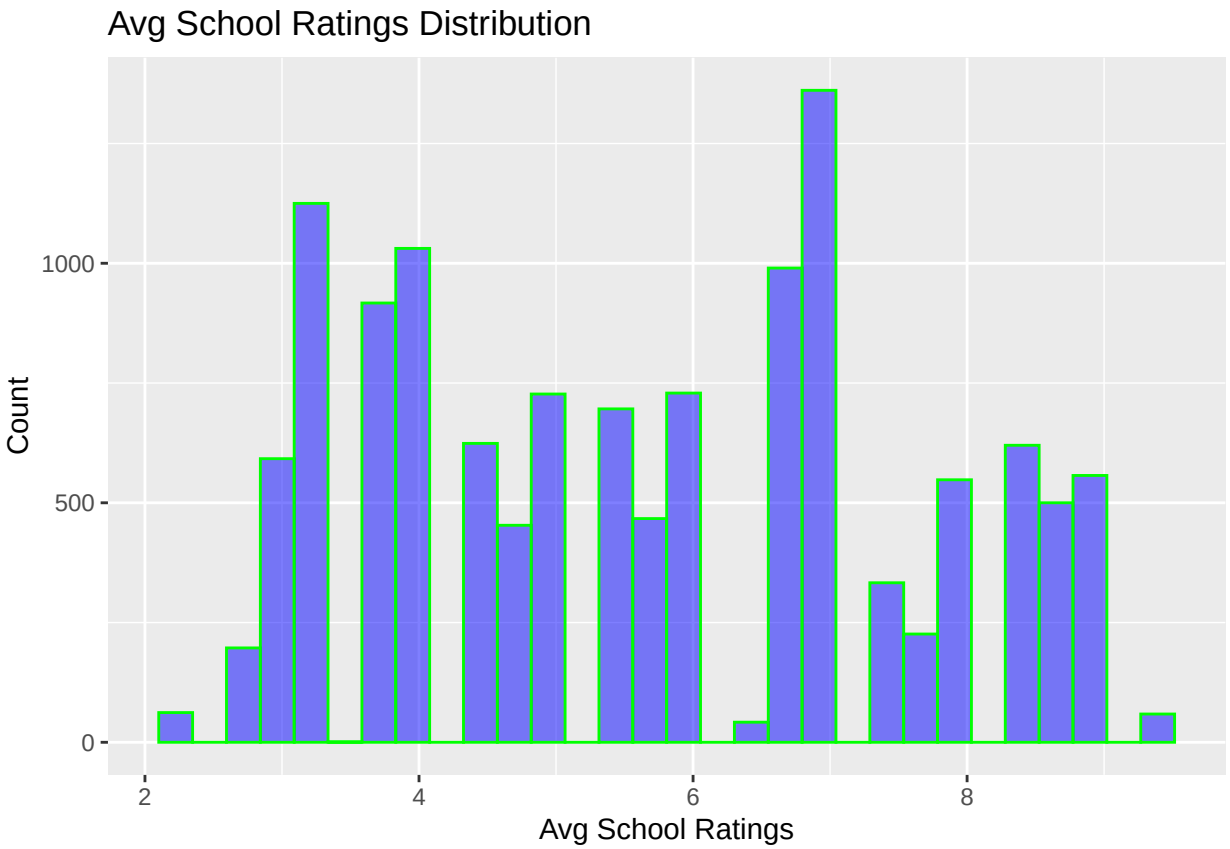
```
## [1] 0
```

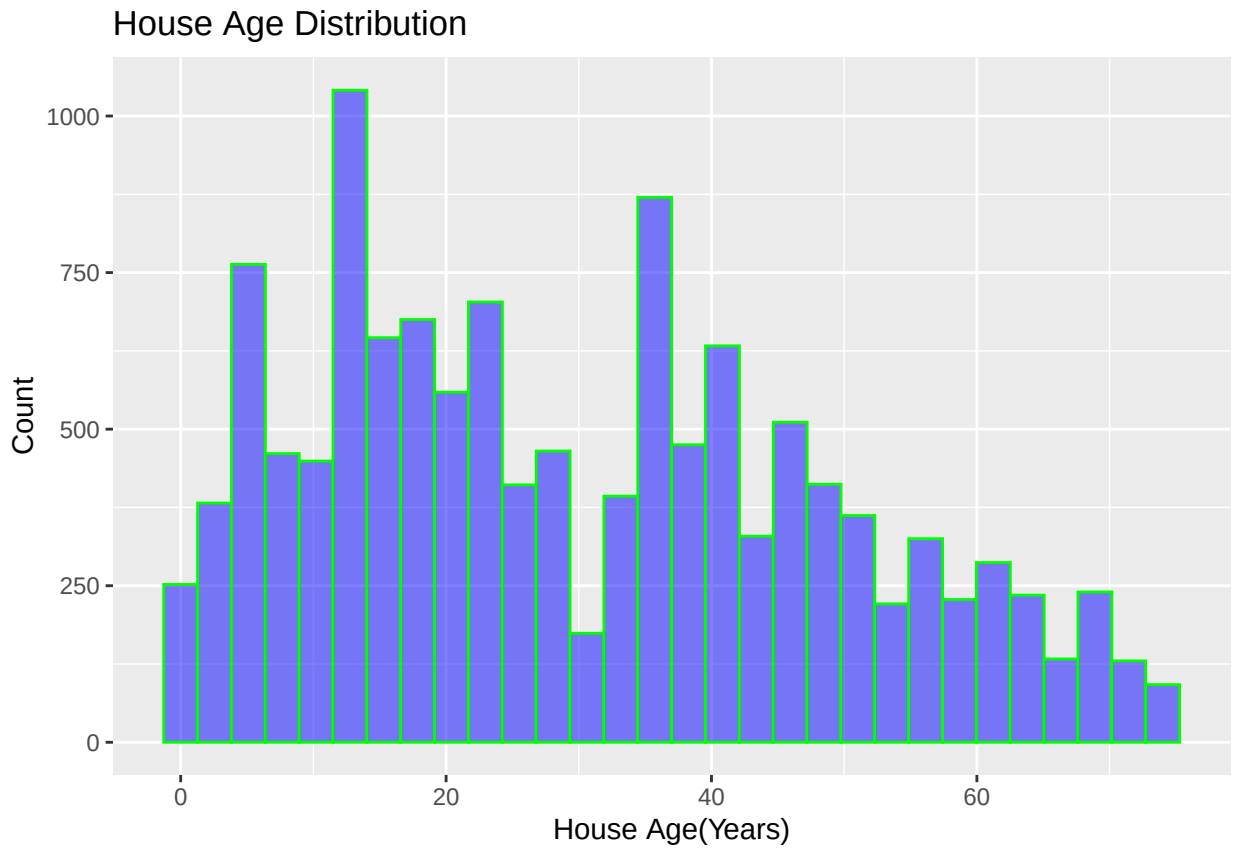
Exploring Features distribution and fixing outliers if any. Removed the ourliers from the house data set.


House Price Distribution
Count
1000
500
0
0
250000
500000
750000
1000000
House Price


Lot Size Distribution
Count
1500
1000
500
0
0
5000
10000
15000
20000
25000
Lot Size(SqFt)


Living Area Distribution
Count
1250
1000
750
500
250
0
1000
2000
3000
4000
5000
Living Area(SqFt)


Avg School Ratings Distribution
Count
1000
500
0
2
4
6
8
Avg School Ratings




```
## [1] 12857    15
```

Predictor Selection:

Since this is Multiple Linear Regression Problem, We are going to test Multicollinearity of predictors. The Predictors should not be perfectly correlated to each other is one of the assumptions in MLR. Finalized the predictors, the list of predictors are:

```
##  [1] "propertyTaxRate"     "garageSpaces"        "hasAssociation"
##  [4] "hasGarage"           "homeType"            "latestPrice"
##  [7] "lotSizeSqFt"         "livingAreaSqFt"      "numOfBathrooms"
## [10] "numOfBedrooms"       "numOfStories"        "numOfPrimarySchools"
## [13] "numOfHighSchools"    "avgSchoolRating"     "houseAge"
```

Modeling:

Fitting model and displaying the result summary:

```
##
## Call:
## lm(formula = formula, data = house)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -875144  -72446  -13402   53694  607240
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     9.934e+05  5.072e+04  19.588  < 2e-16 ***
## propertyTaxRate                -4.118e+05  1.972e+04 -20.885  < 2e-16 ***
## garageSpaces                   -6.828e+03  2.191e+03  -3.117 0.001832 **
## hasAssociationTRUE             -8.977e+04  3.179e+03 -28.237  < 2e-16 ***
## hasGarageTRUE                   3.494e+04  5.040e+03   6.932 4.34e-12 ***
## homeTypeCondo                  -2.049e+03  3.205e+04  -0.064 0.949028
## homeTypeMobile / Manufactured -1.567e+05  4.623e+04  -3.389 0.000703 ***
## homeTypeMultiFamily            -4.426e+04  6.310e+04  -0.701 0.483078
## homeTypeMultiple Occupancy     -8.899e+04  3.531e+04  -2.520 0.011748 *
## homeTypeSingle Family          -5.993e+04  3.154e+04  -1.900 0.057462 .
## homeTypeTownhouse              -7.206e+03  3.296e+04  -0.219 0.826972
## lotSizeSqFt                     2.906e+00  3.868e-01   7.515 6.09e-14 ***
## livingAreaSqFt                  1.258e+02  3.133e+00  40.153  < 2e-16 ***
## numOfBathrooms                  3.107e+04  2.546e+03  12.206  < 2e-16 ***
## numOfBedrooms                  -2.903e+04  2.164e+03 -13.417  < 2e-16 ***
## numOfStories                   -2.591e+04  2.949e+03  -8.784  < 2e-16 ***
## numOfPrimarySchools            -2.938e+03  5.763e+03  -0.510 0.610197
## numOfHighSchools               -3.713e+04  4.394e+03  -8.450  < 2e-16 ***
## avgSchoolRating                 2.355e+04  7.222e+02  32.609  < 2e-16 ***
## houseAge                        6.715e+02  8.376e+01   8.017 1.18e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 121800 on 12837 degrees of freedom
## Multiple R-squared:  0.4913, Adjusted R-squared:  0.4906
## F-statistic: 652.6 on 19 and 12837 DF,  p-value: < 2.2e-16
```

Result Evaluation:

Based on the above result -

1. R-squared value is 0.4913 or 49.13% which tells that how much variability in the outcome is accounted for by the predictors. Here this model accounts for 49% variability in house prices by selected predictors. So, There could be other variables that might impact house price.

2. Predictors estimated values indicates that a unit increase in each holding all other predictors constant will increase or decrease (based on the sign) house price by the value of predictor. For Example - livingAreaSqFt: The living area increases by one square feet, the house value will increase by 125 dollars. avgSchoolRating: The average school rating increase by one unit will increase the house price by 23550 dollars.

3. The Pr(>|t|) value for each predictor shows the estimated value of predictor is statistcally significant or not. For Example - propertyTaxRate: The p-value is less than 0.05 or 5% so this factor will contribute in house price prediction.

4. F-statistics value is 652.6 and corresponding p-value is less than 0.05, which shows that the regression model results in significant better prediction of house prices.

5. Overall the house price prediction model will produce the house prices significantly well.

Implications:

Based on the overall analysis, we can say that the model used for predicting house price will provide better result, which implies that if someone wants to have house information what they are looking for and pass these information to model, it would predict the somewhat accurate price of the house they are looking for.

Limitations:

When talking about the limitation of this analysis and model, There is always improvement since the data we used for house prediction was limited, for example we did not have data related house such as flood zone,crime rate in that area, median income of that neighborhood etc. With the help of more and accurate information will definitely improve the model score and prediction power. Also, model can be improved by implementing the different methods of regression to select right number of predictors such as - Stepwise method, all-subsets method, force entry method etc.

Conclusion:

Overall, by performing the exploratory data analysis and regression analysis, we were able to handle the problem statement mentioned above. With the help of exploratory data analysis the first problem statement was resolved to identify the features that are impacting the overall house prices. Regression analysis helped to tackle second question wherein we need to predict the house price based on given features of house and property area. The Multiple Linear Regression model built to tackle second question on predicting house price and based on the result summary the model we built is statistically significant and predict house price.

After all, the above performed analysis on house price prediction helped to solve both of the business problems.

END