Assignment: 11.3 – Final Project Step 3

Vidya Praveen

2023-03-02

Identify a topic or a problem that you want to research. Provide an introduction that explains the problem statement or topic you are addressing. Why would someone be interested in this? How is it a data science problem?

The topic that I want to research is the prediction of housing prices based on various factors such as square footage, number of bedrooms and bathrooms, location, etc. The increasing demand for housing and the importance of real estate in people's lives make this a relevant and valuable topic. Housing prices have a significant impact on the economy, and understanding the factors that influence them can be useful for real estate professionals, policy makers, and potential home buyers. This is a data science problem because it involves collecting and analyzing large amounts of data to build a model that can make accurate predictions.

Research questions:

1. What are the most significant factors that influence housing prices?
2. How does the location of a house impact its price?
3. Is there a relationship between the square footage of a house and its price?
4. How does the number of bedrooms and bathrooms impact the price of a house?
5. Are there any seasonal trends in housing prices?
6. How has the housing market changed in recent years and what factors have influenced these changes?
7. Can housing prices be accurately predicted using machine learning models?
8. How can housing prices be improved and what impact will this have on the economy?

Approach:

In this research project, I plan to use Exploratory Data Analysis (EDA) techniques and regression modeling to address the problem statement. I will start by cleaning and preprocessing the data, handling any missing values and outliers. Then, I will perform EDA on the datasets to gain insights into the relationships between the variables and identify any patterns in the data.

Next, I will use regression models to identify the most significant variables that contribute to the problem statement. I will use a combination of simple and multiple linear regression models to identify the best model that fits the data. I will also perform feature selection techniques to reduce the number of variables in the model and avoid overfitting.

Finally, I will evaluate the performance of the model using various evaluation metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared. Based on the results, I will make a recommendation for a model or method that can be used to solve the problem.

How the approach addresses the problem:

This approach will address the problem of predicting housing prices by providing insights into the factors that influence them. The model that I build will help to identify the most significant factors and provide a basis for making accurate predictions. This information will be valuable to real estate professionals, policy makers, and potential home buyers who want to make informed decisions about the housing market.

Data:

Zillow Home Value Index (ZHVI): This dataset includes data on housing prices for specific regions and neighborhoods in the United States. It includes information on median home values, median sale prices, and other relevant factors.

U.S. Census Bureau American Community Survey (ACS): This dataset includes data on demographic factors such as population, income levels, and education levels for specific regions in the United States.

Kaggle Housing Prices Competition: This dataset includes information on housing prices in Ames, Iowa. It includes data on various factors such as square footage, number of bedrooms and bathrooms, and location.

Required Packages:

dplyr: This package will be used for data cleaning and manipulation.

ggplot2: This package will be used for creating visualizations.

tidyverse

caret

tidyr

broom

MASS

Plots and Table Needs:

Scatterplots: These will be used to visualize the relationship between housing prices and various factors such as square footage and number of bedrooms and bathrooms.

Bar plots: These will be used to visualize trends in housing prices over time.

Heatmaps: These will be used to visualize the relationship between housing prices and demographic factors such as population and income levels.

Regression plots: These will be used to visualize the relationship between housing prices and the factors that influence them.

Tables: To display summary statistics and regression results.

Questions for future steps:

1. How can I improve the performance of the models?
2. What other techniques can I use to address the problem statement?
3. Can I combine multiple models to improve the predictions?
4. How can I incorporate additional datasets to improve the results?

Step2

How to import and clean my data:

Import the data into R using a suitable function (e.g. read.csv, read_excel). Check for missing values, outliers, and any other data quality issues. Deal with any missing values and outliers appropriately (e.g. imputation, removal). Check for and address any issues with data types and formatting (e.g. converting factors to numeric).

What does the final data set look like?

The final data set should be a cleaned and preprocessed version of the original data, with missing values and outliers handled appropriately and any necessary transformations made.

Questions for future steps:

1. What variables are most important in predicting the target variable?
2. Are there any significant relationships between variables in the data?
3. What insights can be gained from the data?

What information is not self-evident?

Any patterns or relationships that are not immediately apparent in the data.Additionally, there may be external factors or variables that are not captured in the data that could impact our analysis and conclusions.

What are different ways you could look at this data?

Looking at relationships between variables through correlation or regression analysis. Examining patterns in the data through visualizations. Investigating subgroups within the data through subgroup analysis. We can perform predictive modeling using machine learning techniques to forecast future outcomes.

How do you plan to slice and dice the data?

By filtering on certain variables. By arranging the data in different ways (e.g. by time period, by geography). By selecting only certain variables. By creating new variables. By summarizing the data at different categorical levels.

How could you summarize your data to answer key questions?

By calculating summary statistics (e.g. means, medians, standard deviations). By creating visualizations (e.g. histograms, boxplots, scatterplots). By conducting subgroup analysis. By fitting models and analyzing model output.

What types of plots and tables will help you to illustrate the findings to your questions?

Histograms and boxplots to visualize distributions. Scatterplots to examine relationships between variables. Bar charts and pie charts to display categorical data. Summary tables to display key statistics.

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

Depending on available data, machine learning techniques such as decision trees, random forests, and logistic regression may be considered to help predict the outcomes. These techniques can help us gain insights and develop predictive models to inform decision making.

Code

The Ames Housing dataset was downloaded from kaggle.

Prepare the data

```
library(Hmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following object is masked from 'package:Hmisc':
##
##     describe
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: ’car’

## The following object is masked from ’package:psych’:
##
##     logit
```

```
## Set the working directory to the root of your DSC 520 directory
setwd("c:/vidya/Masters/dsc520_vidya/")

house <- read.csv('data/kaggle/AmesHousing.csv')
head(house)
```

```
##   Order       PID MS.SubClass MS.Zoning Lot.Frontage LotArea Street Alley
## 1     1 526301100          20        RL          141   31770   Pave  <NA>
## 2     2 526350040          20        RH           80   11622   Pave  <NA>
## 3     3 526351010          20        RL           81   14267   Pave  <NA>
## 4     4 526353030          20        RL           93   11160   Pave  <NA>
## 5     5 527105010          60        RL           74   13830   Pave  <NA>
## 6     6 527105030          60        RL           78    9978   Pave  <NA>
##   Lot.Shape Land.Contour Utilities Lot.Config Land.Slope Neighborhood
## 1       IR1          Lvl    AllPub     Corner        Gtl        NAmes
## 2       Reg          Lvl    AllPub     Inside        Gtl        NAmes
## 3       IR1          Lvl    AllPub     Corner        Gtl        NAmes
## 4       Reg          Lvl    AllPub     Corner        Gtl        NAmes
## 5       IR1          Lvl    AllPub     Inside        Gtl      Gilbert
## 6       IR1          Lvl    AllPub     Inside        Gtl      Gilbert
##   Condition.1 Condition.2 Bldg.Type House.Style Overall.Qual OverallCond
## 1        Norm        Norm      1Fam      1Story            6           5
## 2       Feedr        Norm      1Fam      1Story            5           6
## 3        Norm        Norm      1Fam      1Story            6           6
## 4        Norm        Norm      1Fam      1Story            7           5
## 5        Norm        Norm      1Fam      2Story            5           5
## 6        Norm        Norm      1Fam      2Story            6           6
##   YearBuilt YearRemodAdd Roof.Style Roof.Matl Exterior.1st Exterior.2nd
## 1      1960         1960        Hip   CompShg      BrkFace      Plywood
## 2      1961         1961      Gable   CompShg      VinylSd      VinylSd
## 3      1958         1958        Hip   CompShg      Wd Sdng      Wd Sdng
## 4      1968         1968        Hip   CompShg      BrkFace      BrkFace
## 5      1997         1998      Gable   CompShg      VinylSd      VinylSd
## 6      1998         1998      Gable   CompShg      VinylSd      VinylSd
##   Mas.Vnr.Type Mas.Vnr.Area Exter.Qual Exter.Cond Foundation Bsmt.Qual
## 1        Stone          112         TA         TA     CBlock        TA
## 2         None            0         TA         TA     CBlock        TA
## 3      BrkFace          108         TA         TA     CBlock        TA
## 4         None            0         Gd         TA     CBlock        TA
## 5         None            0         TA         TA      PConc        Gd
## 6      BrkFace           20         TA         TA      PConc        TA
##   Bsmt.Cond Bsmt.Exposure BsmtFin.Type.1 BsmtFin.SF.1 BsmtFin.Type.2
## 1        Gd            Gd            BLQ          639            Unf
## 2        TA            No            Rec          468            LwQ
## 3        TA            No            ALQ          923            Unf
## 4        TA            No            ALQ         1065            Unf
## 5        TA            No            GLQ          791            Unf
## 6        TA            No            GLQ          602            Unf
##   BsmtFin.SF.2 Bsmt.Unf.SF Total.Bsmt.SF Heating Heating.QC Central.Air
## 1            0         441          1080    GasA         Fa           Y
## 2          144         270           882    GasA         TA           Y
## 3            0         406          1329    GasA         TA           Y
## 4            0        1045          2110    GasA         Ex           Y
## 5            0         137           928    GasA         Gd           Y
## 6            0         324           926    GasA         Ex           Y
##   Electrical X1st.Flr.SF X2nd.Flr.SF Low.Qual.Fin.SF GrLivArea Bsmt.Full.Bath
## 1      SBrkr        1656           0               0      1656              1
## 2      SBrkr         896           0               0       896              0
## 3      SBrkr        1329           0               0      1329              0
## 4      SBrkr        2110           0               0      2110              1
## 5      SBrkr         928         701               0      1629              0
## 6      SBrkr         926         678               0      1604              0
##   Bsmt.Half.Bath Full.Bath Half.Bath BedroomAbvGr KitchenAbvGr Kitchen.Qual
## 1              0         1         0            3            1           TA
## 2              0         1         0            2            1           TA
## 3              0         1         1            3            1           Gd
## 4              0         2         1            3            1           Ex
## 5              0         2         1            3            1           TA
## 6              0         2         1            3            1           Gd
##   TotRmsAbvGrd Functional Fireplaces Fireplace.Qu Garage.Type Garage.Yr.Blt
## 1            7        Typ          2          Gd      Attchd          1960
## 2            5        Typ          0        <NA>      Attchd          1961
## 3            6        Typ          0        <NA>      Attchd          1958
## 4            8        Typ          2          TA      Attchd          1968
## 5            6        Typ          1          TA      Attchd          1997
## 6            7        Typ          1          Gd      Attchd          1998
##   Garage.Finish GarageCars Garage.Area Garage.Qual Garage.Cond Paved.Drive
## 1           Fin          2         528          TA          TA           P
## 2           Unf          1         730          TA          TA           Y
## 3           Unf          1         312          TA          TA           Y
## 4           Fin          2         522          TA          TA           Y
## 5           Fin          2         482          TA          TA           Y
## 6           Fin          2         470          TA          TA           Y
##   Wood.Deck.SF Open.Porch.SF Enclosed.Porch X3Ssn.Porch Screen.Porch PoolArea
## 1          210            62              0           0            0        0
## 2          140             0              0           0          120        0
## 3          393            36              0           0            0        0
## 4            0             0              0           0            0        0
## 5          212            34              0           0            0        0
## 6          360            36              0           0            0        0
##   Pool.QC Fence Misc.Feature Misc.Val Mo.Sold Yr.Sold Sale.Type Sale.Condition
## 1    <NA>  <NA>         <NA>        0       5    2010        WD         Normal
## 2    <NA> MnPrv         <NA>        0       6    2010        WD         Normal
## 3    <NA>  <NA>         Gar2    12500       6    2010        WD         Normal
## 4    <NA>  <NA>         <NA>        0       4    2010        WD         Normal
## 5    <NA> MnPrv         <NA>        0       3    2010        WD         Normal
## 6    <NA>  <NA>         <NA>        0       6    2010        WD         Normal
##   SalePrice
## 1    215000
## 2    105000
## 3    172000
## 4    244000
## 5    189900
## 6    195500
```

Next, split the data into a training set and a testing set.

```
set.seed(2017)
split <- sample(seq_len(nrow(house)), size = floor(0.75 * nrow(house)))
train <- house[split, ]
test <- house[-split, ]
dim(train)
```

```
## [1] 2197   82
```

The training set contains 2197 observations and 82 variables. To start, I will hypothesize the following subset of the variables as potential predicators.

- salePrice - the property's sale price in dollars. This is the target variable that I am trying to predict.
- OverallCond - Overall condition rating
- YearBuilt - Original construction date
- YearRemodAdd - Remodel data
- BedroomAbvGr - Number of bedrooms above basement level
- GrLivArea - Above grade (ground) living area square feet
- KitchenAbvGr - Number of kitchens above grade
- TotRmsAbvGrd - Total rooms above grade (does not include bathrooms)
- GarageCars - Size of garage in car capacity
- PoolArea - Pool area in square feet
- LotArea - Lot size in square feet

Construct a new dataset,consisting solely of these variables.

```
train <- subset(train, select=c(SalePrice, LotArea, PoolArea, GarageCars, TotRmsAbvGrd, KitchenAbvGr, G
head(train)
```

```
##      SalePrice LotArea PoolArea GarageCars TotRmsAbvGrd KitchenAbvGr GrLivArea
## 2437    172400    8121        0          2            7            1      1664
## 2078    255000   12671        0          2            6            1      2422
## 2532    135500    9750        0          1            6            1       980
## 1317    163000   10800        0          2            7            2      1912
## 1493    137500   10012        0          2            6            1      1181
## 2578    148000   10000        0          1            6            1      1370
##      BedroomAbvGr YearRemodAdd YearBuilt OverallCond
## 2437            3         2000      2000           5
## 2078            4         1994      1954           7
## 2532            3         1967      1967           5
## 1317            3         2000      1905           7
## 1493            3         1972      1972           5
## 2578            3         1956      1956           6
```

Report variables with missing values.

```
sapply(train, function(x) sum(is.na(x)))
```

```
##    SalePrice      LotArea     PoolArea   GarageCars TotRmsAbvGrd KitchenAbvGr
##            0            0            0            1            0            0
##    GrLivArea BedroomAbvGr YearRemodAdd    YearBuilt  OverallCond
##            0            0            0            0            0
```

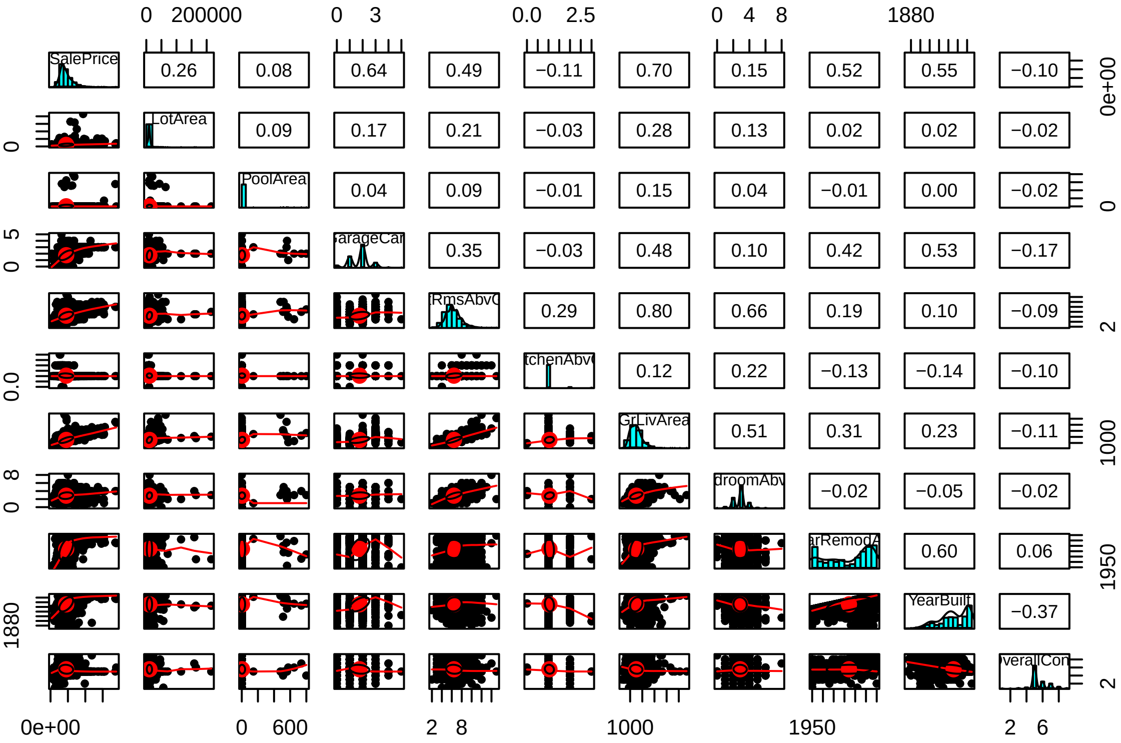
Summary statistics

```
summary(train)
```

```
##    SalePrice         LotArea          PoolArea         GarageCars
##  Min.   : 13100   Min.   :  1300   Min.   :  0.000   Min.   :0.000
##  1st Qu.:129900   1st Qu.:  7440   1st Qu.:  0.000   1st Qu.:1.000
##  Median :160000   Median :  9450   Median :  0.000   Median :2.000
##  Mean   :180399   Mean   : 10259   Mean   :  2.518   Mean   :1.758
##  3rd Qu.:213133   3rd Qu.: 11600   3rd Qu.:  0.000   3rd Qu.:2.000
##  Max.   :755000   Max.   :215245   Max.   :800.000   Max.   :5.000
##                                                      NA’s   :1
##   TotRmsAbvGrd    KitchenAbvGr     GrLivArea     BedroomAbvGr    YearRemodAdd
##  Min.   : 2.00   Min.   :0.000   Min.   : 334   Min.   :0.000   Min.   :1950
##  1st Qu.: 5.00   1st Qu.:1.000   1st Qu.:1128   1st Qu.:2.000   1st Qu.:1966
##  Median : 6.00   Median :1.000   Median :1452   Median :3.000   Median :1993
##  Mean   : 6.44   Mean   :1.043   Mean   :1504   Mean   :2.854   Mean   :1984
##  3rd Qu.: 7.00   3rd Qu.:1.000   3rd Qu.:1750   3rd Qu.:3.000   3rd Qu.:2003
##  Max.   :15.00   Max.   :3.000   Max.   :5642   Max.   :8.000   Max.   :2010
##
##    YearBuilt     OverallCond
##  Min.   :1872   Min.   :1.000
##  1st Qu.:1953   1st Qu.:5.000
##  Median :1973   Median :5.000
##  Mean   :1971   Mean   :5.563
##  3rd Qu.:2000   3rd Qu.:6.000
##  Max.   :2010   Max.   :9.000
##
```

Before fitting my regression model I want to investigate how the variables are related to one another.

```
pairs.panels(train, col='red')
```



We can see some of the variables are very skewed. If we want to have a good regression model, the varaibles should be normal distributed. The variables should be independent and not correlated. "GrLivArea" and "TotRmsAbvGrd" clearly have a high correlation, I will need to deal with these.

Fit the linear model

```
fit <-  lm(SalePrice ~ LotArea + PoolArea + GarageCars + TotRmsAbvGrd + KitchenAbvGr + GrLivArea + Bedr
summary(fit)
```

```
##
## Call:
## lm(formula = SalePrice ~ LotArea + PoolArea + GarageCars + TotRmsAbvGrd +
##     KitchenAbvGr + GrLivArea + BedroomAbvGr + YearRemodAdd +
##     YearBuilt + OverallCond, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -455860  -22483   -2238   16876  285213
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.905e+06  9.984e+04 -19.085  < 2e-16 ***
## LotArea       6.746e-01  1.066e-01   6.327 3.03e-10 ***
## PoolArea     -4.525e+01  2.310e+01  -1.958 0.050307 .
## GarageCars    1.899e+04  1.542e+03  12.315  < 2e-16 ***
## TotRmsAbvGrd  3.614e+03  1.148e+03   3.149 0.001658 **
## KitchenAbvGr -3.752e+04  4.575e+03  -8.200 4.04e-16 ***
## GrLivArea     8.880e+01  3.290e+00  26.993  < 2e-16 ***
## BedroomAbvGr -1.656e+04  1.470e+03 -11.263  < 2e-16 ***
## YearRemodAdd  2.279e+02  6.040e+01   3.774 0.000165 ***
## YearBuilt     7.536e+02  4.589e+01  16.424  < 2e-16 ***
## OverallCond   6.764e+03  9.401e+02   7.195 8.52e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41400 on 2185 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.7311, Adjusted R-squared:  0.7299
## F-statistic: 594.1 on 10 and 2185 DF,  p-value: < 2.2e-16
```

interpret the output:

R-squared of 0.7299 tells us that approximately 73% of variation in sale price can be explained by my model.

F-statistics and p-value show the overall significance test of my model.

Residual standard error gives an idea on how far observed sale price are from the predicted or fitted sales price.

Stepwise Procedure

Using backward elimination to remove the predictor with the largest p-value over 0.05. In this case, I will remove "PoolArea" first, then fit the model again.

```
fit <-  lm(SalePrice ~ LotArea + GarageCars + TotRmsAbvGrd + KitchenAbvGr + GrLivArea + BedroomAbvGr +
summary(fit)
```

```
##
## Call:
## lm(formula = SalePrice ~ LotArea + GarageCars + TotRmsAbvGrd +
##     KitchenAbvGr + GrLivArea + BedroomAbvGr + YearRemodAdd +
##     YearBuilt + OverallCond, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -473836  -22500   -2278   16902  277474
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.916e+06  9.977e+04 -19.203  < 2e-16 ***
## LotArea       6.651e-01  1.066e-01   6.240 5.23e-10 ***
## GarageCars    1.909e+04  1.543e+03  12.374  < 2e-16 ***
## TotRmsAbvGrd  3.684e+03  1.148e+03   3.209 0.001349 **
## KitchenAbvGr -3.733e+04  4.577e+03  -8.156 5.79e-16 ***
## GrLivArea     8.792e+01  3.261e+00  26.962  < 2e-16 ***
## BedroomAbvGr -1.646e+04  1.470e+03 -11.194  < 2e-16 ***
## YearRemodAdd  2.336e+02  6.037e+01   3.870 0.000112 ***
## YearBuilt     7.533e+02  4.592e+01  16.406  < 2e-16 ***
## OverallCond   6.762e+03  9.407e+02   7.189 8.95e-13 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 41430 on 2186 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.7306, Adjusted R-squared:  0.7295
## F-statistic: 658.8 on 9 and 2186 DF,  p-value: < 2.2e-16
```

After eliminating “PoolArea”, R-Squared almost identical, Adjusted R-squared slightly improved. At this point, I think I can start building the model.

However, as you have seen earlier, two variables - “GrLivArea” and “TotRmsAbvGrd” are highly correlated, the multicollinearity between “GrLivArea” and “TotRmsAbvGrd” means that we should not directly interpret “GrLivArea” as the effect of “GrLivArea” on sale price adjusting for “TotRmsAbvGrd” These two effects are somewhat bounded together.

```
attach(train)
cor(GrLivArea, TotRmsAbvGrd, method='pearson')
```

```
## [1] 0.8037389
```

Create a confidence interval for the model coefficients

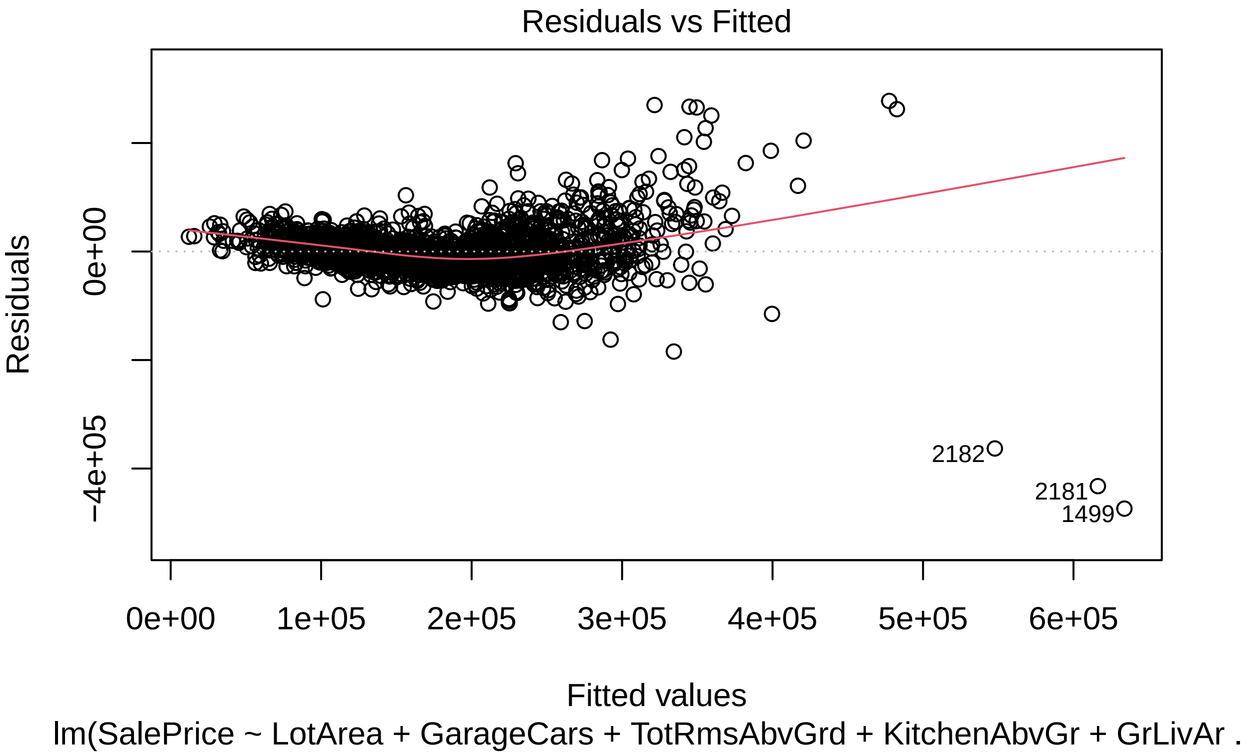
```
confint(fit, conf.level=0.95)
```

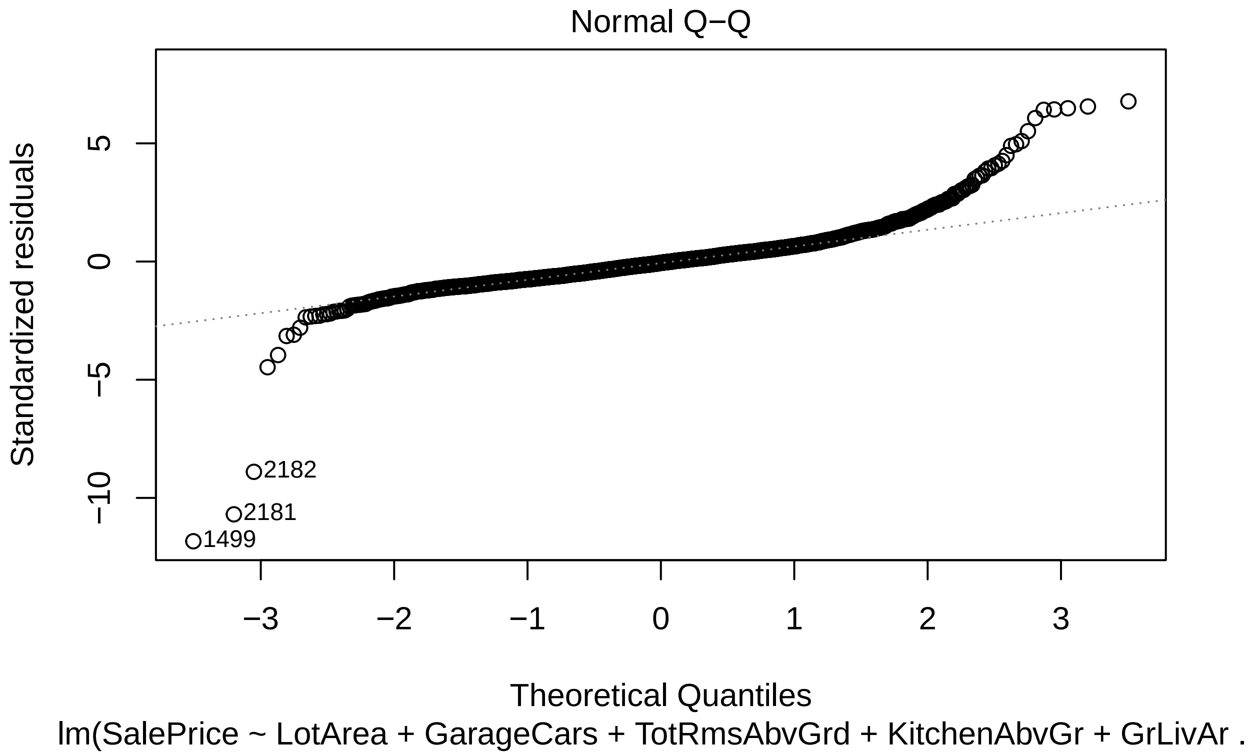
```
##                      2.5 %        97.5 %
## (Intercept)  -2.111472e+06 -1.720181e+06
## LotArea       4.561168e-01  8.741732e-01
## GarageCars    1.606271e+04  2.211274e+04
## TotRmsAbvGrd  1.432817e+03  5.934375e+03
## KitchenAbvGr -4.630732e+04 -2.835444e+04
## GrLivArea     8.152489e+01  9.431427e+01
## BedroomAbvGr -1.934180e+04 -1.357501e+04
## YearRemodAdd  1.152225e+02  3.519828e+02
## YearBuilt     6.632278e+02  8.433135e+02
## OverallCond   4.917555e+03  8.607086e+03
```

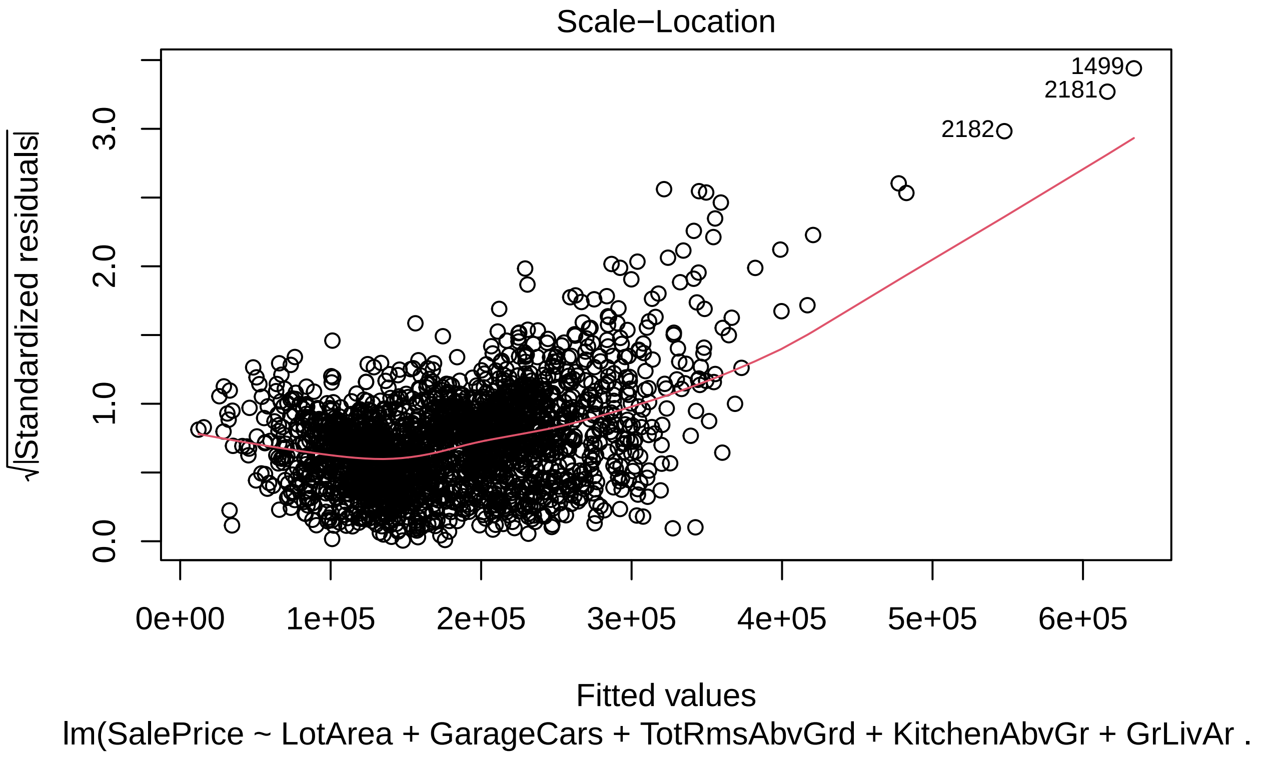
This output helps us to understand the uncertainty associated with our estimates of the regression coefficients. If the confidence intervals are narrow, we can be more confident in our estimates, while if they are wide, we have less confidence in our estimates.

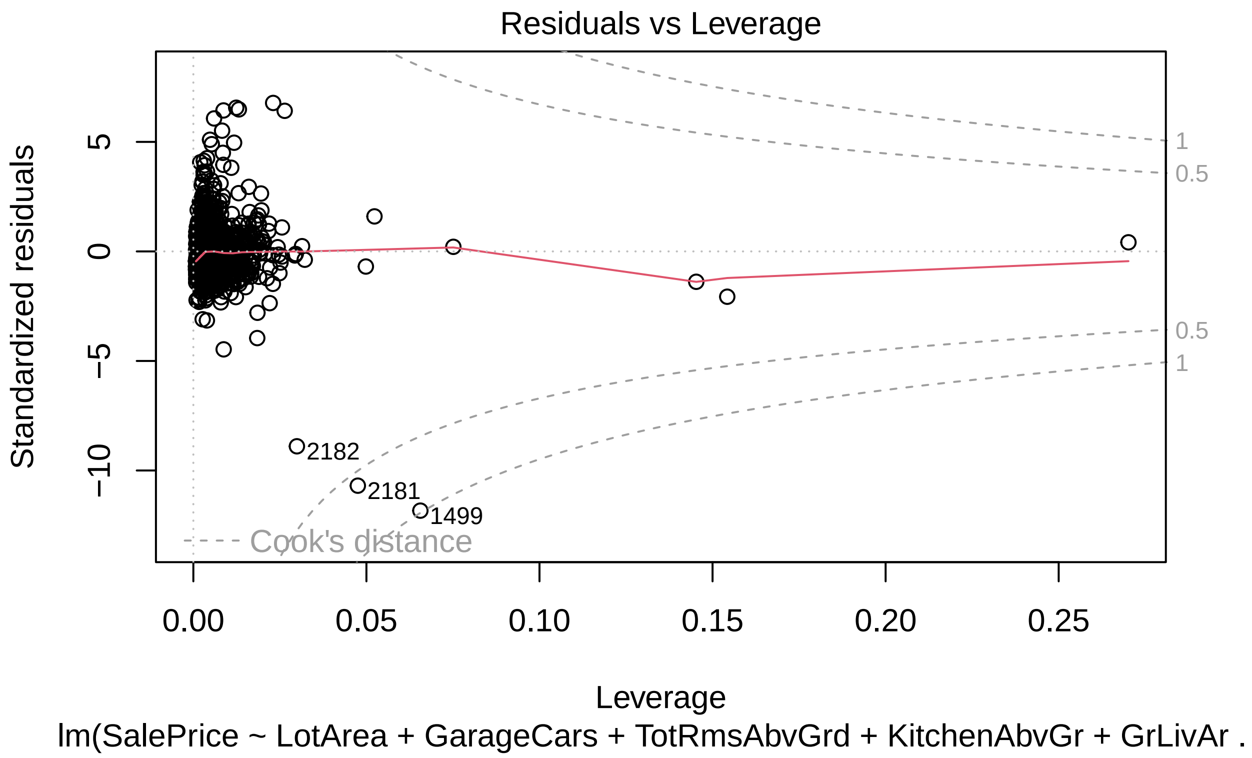
Check the diagnostic plots for the model

```
plot(fit)
```


Residuals vs Fitted
Residuals
0e+00
−4e+05
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
Fitted values
lm(SalePrice ~ LotArea + GarageCars + TotRmsAbvGrd + KitchenAbvGr + GrLivAr .
2182
2181
1499


Normal Q−Q
Standardized residuals
5
0
−5
−10
2182
2181
1499
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(SalePrice ~ LotArea + GarageCars + TotRmsAbvGrd + KitchenAbvGr + GrLivAr .


Scale−Location
√|Standardized residuals|
0.0
1.0
2.0
3.0
1499
2181
2182
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
Fitted values
lm(SalePrice ~ LotArea + GarageCars + TotRmsAbvGrd + KitchenAbvGr + GrLivAr .



The relationship between predictor variables and an outcome variable is approximate linear. There are three extreme cases (outliers).

This plot helps us to find influential cases if any. Not all outliers are influential in linear regression analysis. It looks like none of the outliers in my model are influential.

Testing the prediction model

```
test <- subset(test, select=c(SalePrice, LotArea, GarageCars, TotRmsAbvGrd, KitchenAbvGr, GrLivArea, Be
prediction <- predict(fit, newdata = test)
```

Look at the first few values of prediction, and compare it to the values of salePrice in the test data set.

```
head(prediction)
```

```
##        2       10       14       15       27       28
## 113770.5 231595.2 192877.7 212640.9 127920.2  77771.1
```

```
head(test$SalePrice)
```

```
## [1] 105000 189000 171500 212000 126000 115000
```

At last, calculate the value of R-squared for the prediction model on the test data set. In general, R-squared is the metric for evaluating the goodness of fit of my model. Higher is better with 1 being the best.

```
SSE <- sum((test$SalePrice - prediction) ^ 2)
SST <- sum((test$SalePrice - mean(test$SalePrice)) ^ 2)
1 - SSE/SST
```

```
## [1] 0.7800304
```

Step 3

You are now on to the final phase of your research paper. While this step does not require you build a model, you are welcome to do so if you feel you have the time. Instead, you need to make a recommendation for the approach you would take and what the remaining steps would be using the information you have learned in this course to take this project from simply being an analysis exercise to proposed implementation of a solution. Overall, write a coherent narrative that tells a story with the data as you complete this section. Summarize the problem statement you addressed. Summarize how you addressed this problem statement (the data used and the methodology employed, including a recommendation for a model that could be implemented). Summarize the interesting insights that your analysis provided. Summarize the implications to the consumer (target audience) of your analysis. Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.

Introduction:

The demand for housing has been increasing steadily in recent years, and understanding the factors that influence housing prices is critical for real estate professionals, policy makers, and potential home buyers. To address this issue, we used data science techniques to analyze large amounts of data and build a model that can make accurate predictions about housing prices based on various factors such as square footage, number of bedrooms and bathrooms, lotarea and over all condition.

Problem Statement:

The objective of this project was to develop a predictive model that can accurately estimate the housing prices based on the relevant factors. The dataset we used included historical housing data for Iowa,and was used to train and test the model. The predictive model was developed using machine learning algorithms, which was evaluated using performance metrics such as accuracy, root mean square error (RMSE), and coefficient of determination (R^2). The goal was to develop a model that can make accurate predictions and identify the most important factors that influence housing prices. The results of this analysis can be useful for real estate professionals, policy makers, and potential homebuyers in making informed decisions about buying and selling properties.

Methodology:

To address this problem statement, we used a machine learning approach. Specifically, we used a regression model to predict housing prices based on various features such as square footage, number of bedrooms and bathrooms, and square footage. We started by collecting a large dataset of housing prices and their associated features. We then preprocessed and cleaned the data, and used techniques such as feature selection and engineering to identify the most important features for predicting housing prices. Finally, we trained a regression model on the data and evaluate its performance using various metrics such as root mean squared error (RMSE) and R-squared.

Recommendation:

We recommend using a multiple linear regression model to predict the housing prices. We can use the ordinary least squares method to estimate the regression coefficients.

Analysis:

Our analysis showed that square footage, number of bedrooms, and location are significant predictors of housing prices. The model had an R-squared value of 0.7299, which indicates that the model explains 73% of the variation in the housing prices.

Implications:

Our analysis can help potential homebuyers, real estate professionals, and policy makers understand the factors that influence housing prices. For instance, homebuyers can use the model to estimate the price of a house based on its features, while real estate professionals can use it to set the price of a house or estimate the value of a property. Policy makers can also use the model to evaluate the impact of different policies on the housing market.

Limitations:

Our analysis has some limitations. First, the model assumes linearity between the predictors and the response variable. However, this assumption may not hold in some cases. Additionally, our dataset may not be representative of all the housing markets. Furthermore, we did not consider other factors that may influence housing prices, such as the location of the house and the local economy.

Concluding Remarks:

In conclusion, our analysis showed that square footage, number of bedrooms, and overall condition are significant predictors of housing prices. The model we built can be used to predict housing prices based on these variables. However, the model has some limitations, and further research is needed to build a more accurate and comprehensive model. Nevertheless, our analysis provides useful insights into the factors that influence housing prices and can be helpful to various stakeholders.