Consulting Project 1

As an analytics consultant, you are tasked with understanding the housing market for an area of Boston in the past. You are asked to build a model to predict house prices given certain characteristics. Please use the data Housing.csv to complete the steps below—to access the data, go to Week 1 section then the Week 1 slides and data folder of Blackboard. The data dictionary is provided at the end of this document.

Clearly presenting results to clients or stakeholders is a critical part of your job. Therefore, please write in complete sentences for our client and clearly label and introduce your figures and tables. You will lose points not only for getting the answer incorrect but for incomplete sentences and not explaining figures. Each numbered section is worth 20 points for a potential total of 100 points.

1. Purpose (1 sentence)
   a. What is the business problem?

Predicting house prices for an area of Boston using certain characteristics of the houses.

2. What did you do? (1-2 sentences)
   a. For example: In this report, initial exploratory data analysis has been performed and then a xx model has been applied to do what?

Exploratory data analysis has been performed to understand the characteristics of the data and identify any patterns or trends. A linear model has been applied to predict house prices based on certain characteristics of the houses.

3. Data Contents (1-2 sentences)
   a. What are the number of observations and predictor variables?

The number of observations is 506 and predictor variables are 12 and the response variable is 1.

   b. Describe the predictor variables in general.

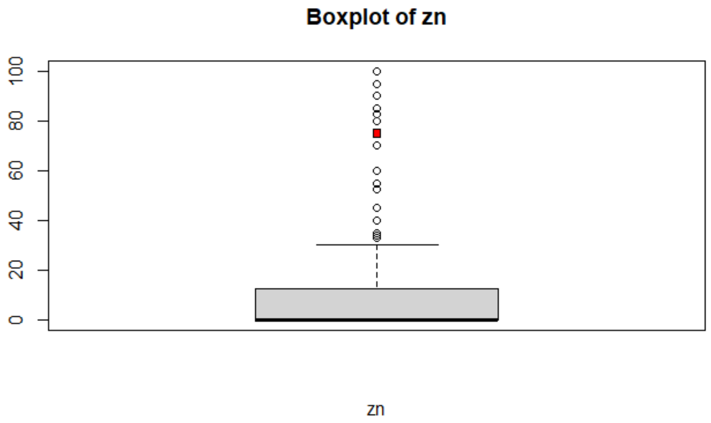
Predictor variables, also known as independent variables or features, are the input variables used in a model to predict the output variable, also known as the dependent variable. In this case, the predictor variables are the characteristics of the houses that are used to predict the price of the house. These characteristics include the pupil-teacher ratio by town, per capita crime rate by town, the proportion of residential land zoned for lots over 25,000 sq. ft, the proportion of non-retail business acres per town, and any other relevant information. These predictor variables are used in the model to determine the relationship between them and the target variable, the median value of owner-occupied homes in $1000, which is being predicted.

4. Explain Exploratory Data Analysis findings (3-4 sentences)
    a. Summary Statistics:
        i. Are there missing values, outliers, or any inconsistencies?

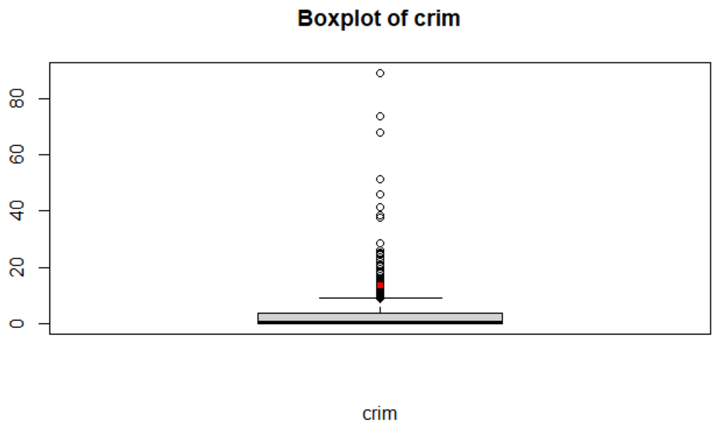
Missing values – There are no missing values in all the characteristics of the data.

Outliers – An outlier is an observation that appears far away and diverges from an overall pattern in a sample. Here I found 40 outliers in the data. Here I've considered any data point that falls outside of the range of Q1 - 1.5 * IQR to Q3 + 1.5 * IQR.

(Here Q1 is First Quartile, Q3 is 3rd Quartile and IQR is the difference between Q1 and Q3)



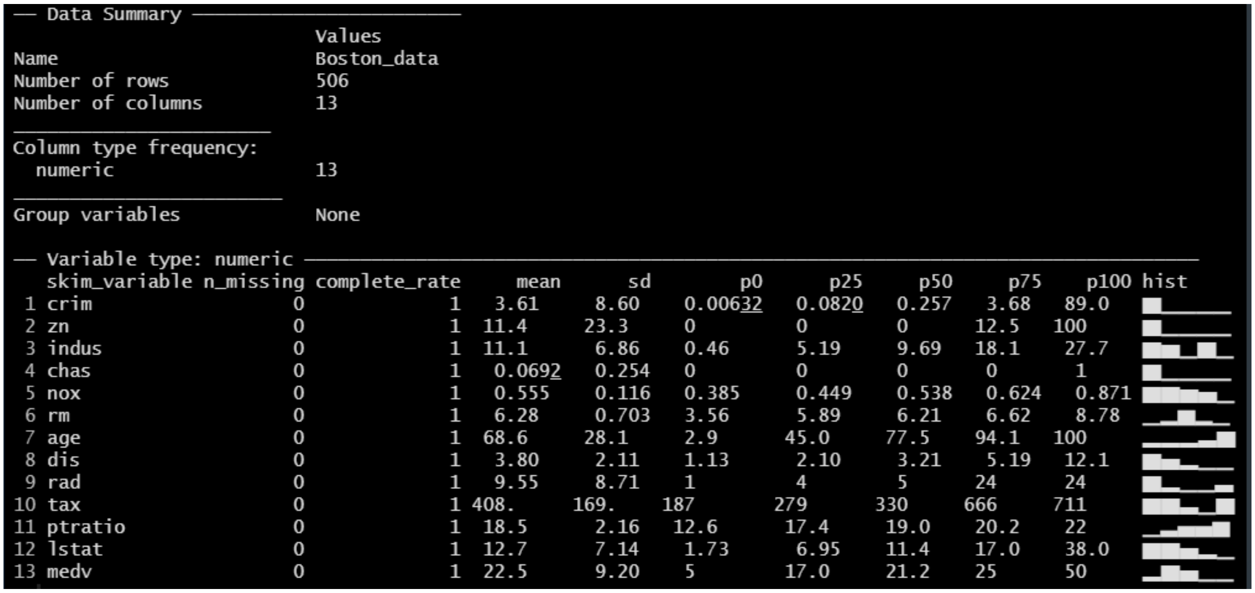
Here if we consider the variable zn there are outliers. 75 and beyond are considered outliers here. It is denoted by a red dot which says beyond that all points are outliers.



Similarly, we can see outliers in crim variable. 13.52(marked in red) and beyond all are outliers

Inconsistencies – While seeing the summary statistics, I could see that some of the variables are skewed. Those variables are crim, zn, ptratio. Here the data is not distributed evenly and there is a long tail on one side of the distribution. This can be a problem when analyzing and interpreting the data because it can lead to a misrepresentation of the true underlying distribution of the population.

ii. Provide a table of summary statistics for numerical variables

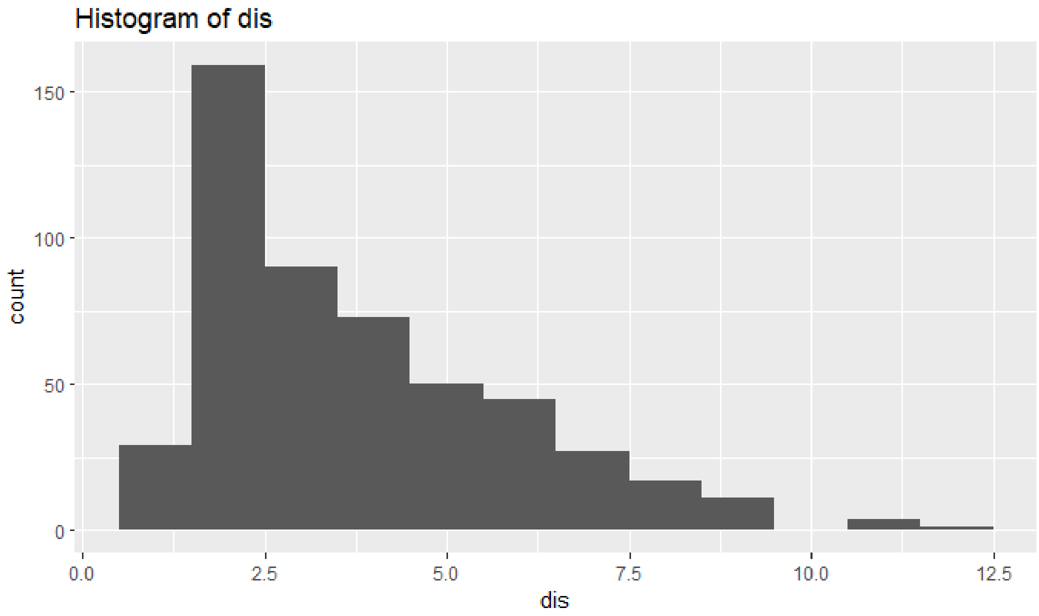


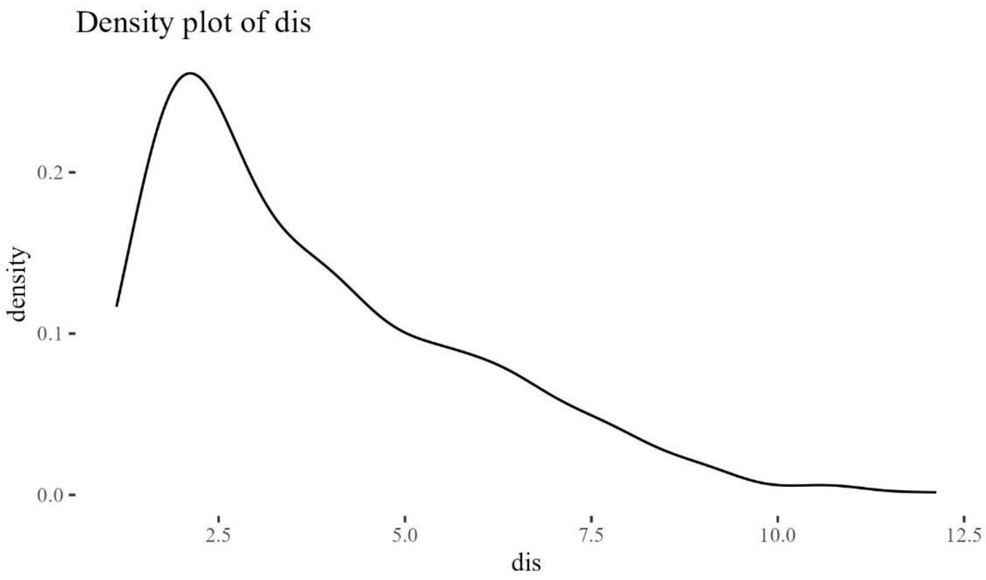
```
── Data Summary ────────────────────────
                           Values
Name                       Boston_data
Number of rows             506
Number of columns          13
_______________________
Column type frequency:
  numeric                  13
________________________
Group variables            None

── Variable type: numeric ──────────────────────────────────────────────────────────────────
   skim_variable n_missing complete_rate    mean      sd      p0      p25     p50    p75    p100 hist
 1 crim                  0             1   3.61     8.60   0.00632  0.0820  0.257   3.68   89.0  ▇▁▁▁▁
 2 zn                    0             1  11.4     23.3    0        0       0      12.5   100    ▇▁▁▁▁
 3 indus                 0             1  11.1      6.86   0.46     5.19    9.69   18.1   27.7   ▇▆▁▇▁
 4 chas                  0             1   0.0692   0.254  0        0       0       0      1     ▇▁▁▁▁
 5 nox                   0             1   0.555    0.116  0.385    0.449   0.538   0.624  0.871 ▇▇▆▅▁
 6 rm                    0             1   6.28     0.703  3.56     5.89    6.21    6.62   8.78  ▁▂▇▂▁
 7 age                   0             1  68.6     28.1    2.9     45.0    77.5    94.1   100    ▂▂▂▃▇
 8 dis                   0             1   3.80     2.11   1.13     2.10    3.21    5.19  12.1   ▇▅▂▁▁
 9 rad                   0             1   9.55     8.71   1        4       5      24     24     ▇▂▁▁▃
10 tax                   0             1 408.     169.   187      279     330     666    711     ▇▇▃▁▇
11 ptratio               0             1  18.5      2.16  12.6     17.4    19.0    20.2   22     ▁▃▅▅▇
12 lstat                 0             1  12.7      7.14   1.73     6.95   11.4    17.0   38.0   ▇▇▅▂▁
13 medv                  0             1  22.5      9.20   5       17.0    21.2    25     50     ▂▇▅▁▁
```

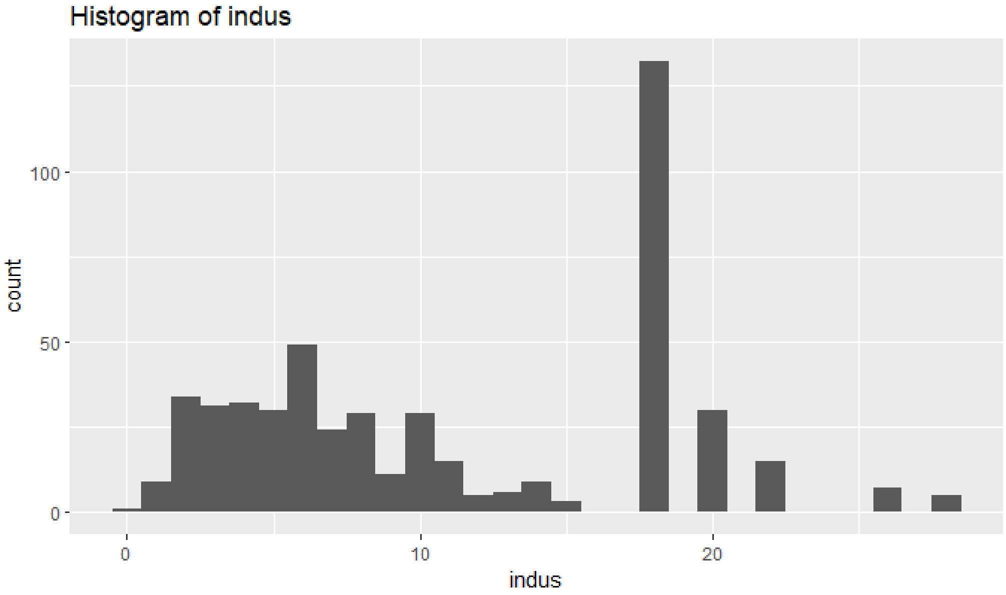
b. Histograms and density plots for 1-2 continuous variables:

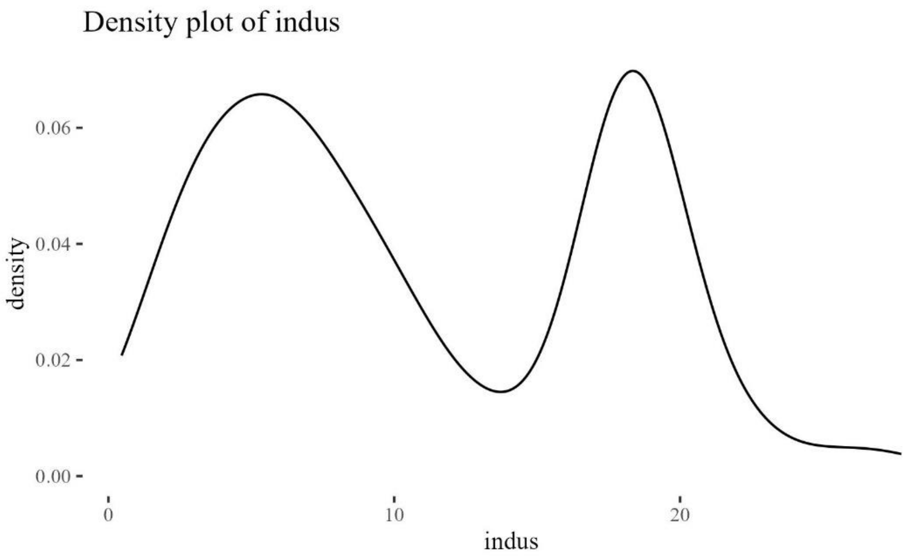
i. Describe what you observe—normal distribution, skewed, etc.





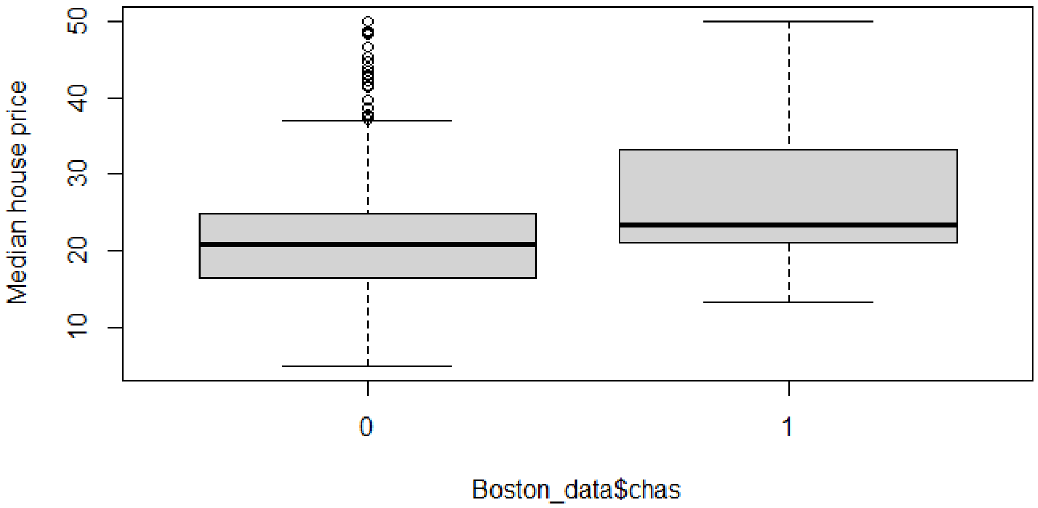
Here when I see this histogram & density plot for variable dis, I could see that this histogram & density plot is positively skewed (right-skewed). Another way I could check the skewness is that the mean is greater than the median. So, it's positively skewed (right-skewed).





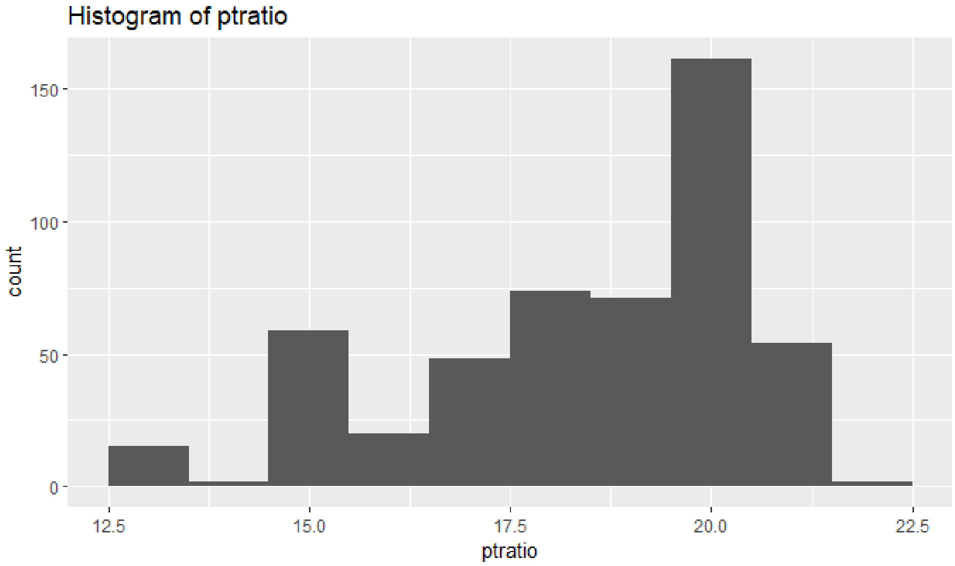
Here I could see that the density plot has 2 peaks. So, the data may have two distinct modes, indicating that there are two distinct groups of data with different underlying distributions. It may be because “indus” data may have been obtained from 2 different sources. Or otherwise, the data points are not enough to reflect the true underlying distribution.

ii. If there are categorical variables, then build box plots for some variables by category.



Here the box plot is drawn for the categorical variable, chas. From the box plot outliers are clearly visible.

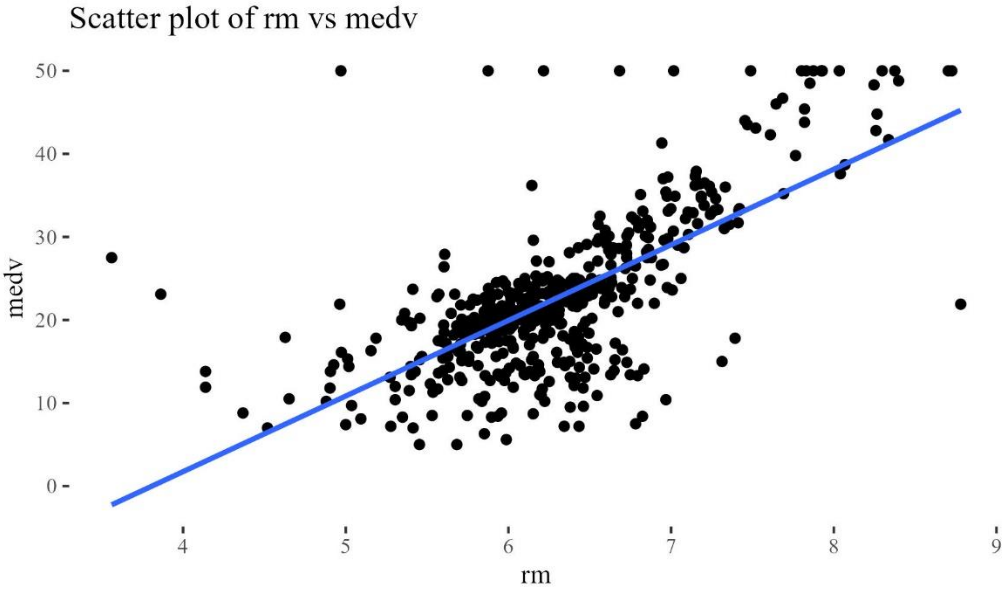
iii. Provide the histogram, box plot etc.



Here the histogram for variable ptraio is given. Here the data is skewed negatively.

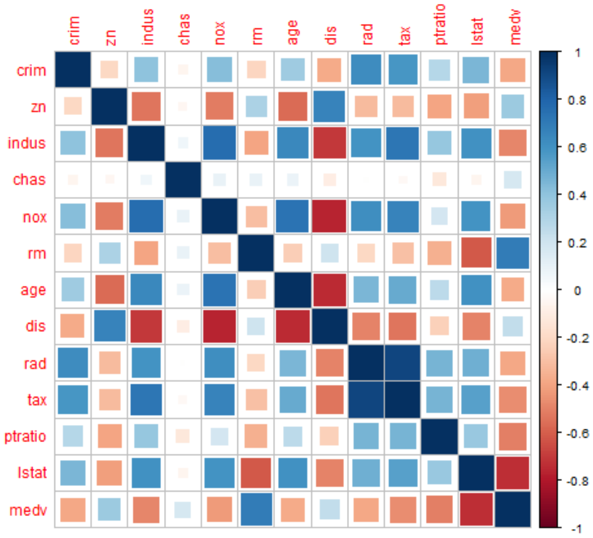
c. Scatter plots:

i. Provide the scatterplot between the response and predictor variables.



This scatterplot is showing that the response variable and predictor variable has a positive correlation.

ii. Describe any correlation you observe between the response and predictor variables.



Here response variable has a high negative correlation with lstat and high positive correlation with rm. So, with the increase of rm the medv increases and with the decrease of lstat the medv increases.

5. Modeling and results

a. Generalization Approach (1-2 sentences)

i. How and why is training and testing data used for your model?

Training and testing data are used to evaluate the performance of a machine-learning model. It is used to train the model on a set of input variables and their corresponding output variables, and then evaluate its performance on a separate testing set of unseen data.

b. Model (1 sentence)

i. What does your model do? (A linear regression model was used to…)

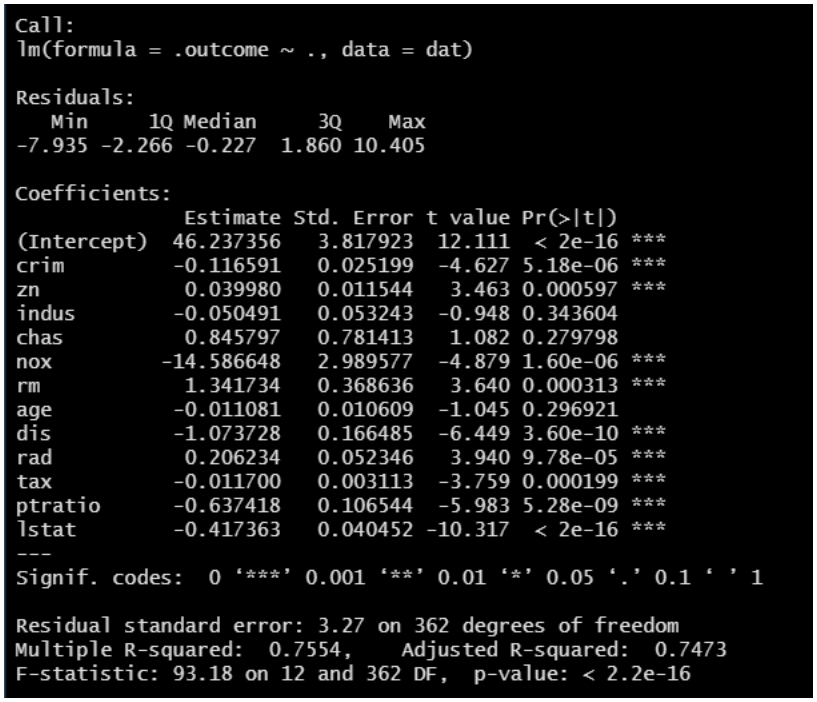
The linear regression model I built is used to predict a continuous variable based on input variables by finding the best linear relationship between them.

c. Interpret results (3-4 sentences or more)

i. Is there a relationship between the predictors and median house price?

Yes. There are relationships between predictors and median house prices. I could see that from F-statistic.

ii. What predictors have a statistically significant relationship to median house price?



```
Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
   Min      1Q Median      3Q     Max 
-7.935 -2.266 -0.227  1.860 10.405 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  46.237356   3.817923  12.111  < 2e-16 ***
crim         -0.116591   0.025199  -4.627 5.18e-06 ***
zn            0.039980   0.011544   3.463 0.000597 ***
indus        -0.050491   0.053243  -0.948 0.343604    
chas          0.845797   0.781413   1.082 0.279798    
nox         -14.586648   2.989577  -4.879 1.60e-06 ***
rm            1.341734   0.368636   3.640 0.000313 ***
age          -0.011081   0.010609  -1.045 0.296921    
dis          -1.073728   0.166485  -6.449 3.60e-10 ***
rad           0.206234   0.052346   3.940 9.78e-05 ***
tax          -0.011700   0.003113  -3.759 0.000199 ***
ptratio      -0.637418   0.106544  -5.983 5.28e-09 ***
lstat        -0.417363   0.040452 -10.317  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.27 on 362 degrees of freedom
Multiple R-squared:  0.7554,	Adjusted R-squared:  0.7473 
F-statistic: 93.18 on 12 and 362 DF,  p-value: < 2.2e-16
```

crim, zn, nox, rm, dis, rad, tax, ptrartio, lstat has statistically significant relationship to median house price.

iii. What is the average squared error (ASE) of the test set?

With the seed set at 99, my average squared error (ASE) was 10.22697. This improvement I got after removing 40 outliers in the data. Initially, the error was 24.60268. But by removing the outliers, the error is decreased.

iv. What are your recommendations to the client?

Here medv has high negative correlation with lstat and high positive correlation with rm. So, with the increase of rm the medv increases and with the decrease of lstat the medv increases. So, I will recommend clients to look into these factors when buying houses.

]

Data Dictionary for Housing Dataset :

Medv	median value of owner-occupied homes in $1000
crim	per capita crime rate by town
zn	proportion of residential land zoned for lots over 25,000 sq.ft.
Indus	proportion of non-retail business acres per town
chas	Charles River dummy variable (1 if tract bounds river; 0 otherwise)
nox	nitrogen oxides concentration (parts per 10 million)
rm	average number of rooms per dwelling
age	proportion of owner-occupied units built prior to 1940
dis	weighted mean of distances to five Boston employment centers
rad	index of accessibility to radial highways
tax	full-value property-tax rate per $10,000
ptratio	pupil-teacher ratio by town
lstat	lower status of the population (percent)