Boston-Housing-Analysis

Isha Doshi

2022-11-20

Housing Values in Boston

Checking for nulls

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
data<-read.delim("boston.csv.bz2")
head(data)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

```
cat(any(is.na(data)),"- There are no NULL entries \n")
```

```
## FALSE - There are no NULL entries
```

```
summary(data)
```

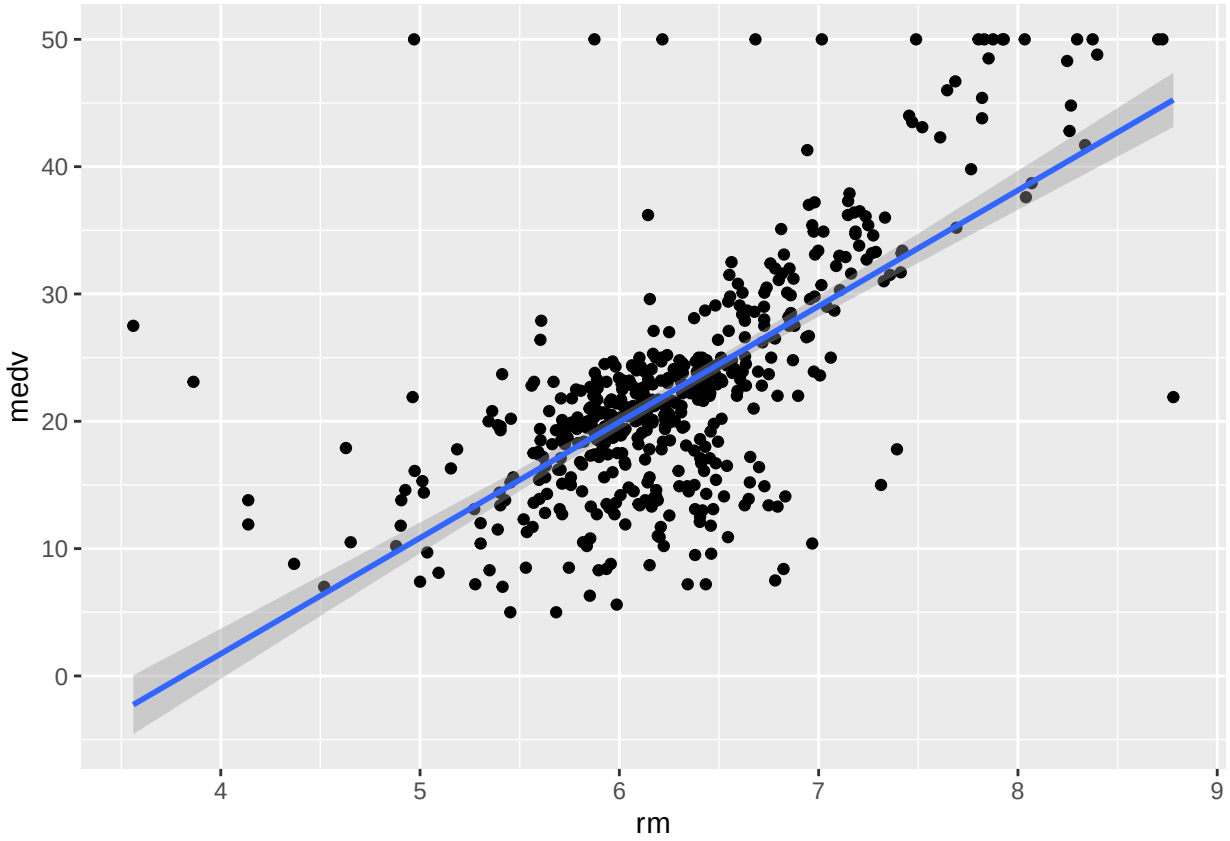
```
##       crim                zn              indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm              age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          black       
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
##      lstat            medv      
##  Min.   : 1.73   Min.   : 5.00  
##  1st Qu.: 6.95   1st Qu.:17.02  
##  Median :11.36   Median :21.20  
##  Mean   :12.65   Mean   :22.53  
##  3rd Qu.:16.95   3rd Qu.:25.00  
##  Max.   :37.97   Max.   :50.00
```

The relationship between median value of owner-occupied homes and average number of rooms per dwelling are directly proportional. There are higher avg of rooms per dwelling in higher values homes. The relationship between median value of owner-occupied homes and lower status of the population is inversely proportional. The cost of houses is lower in the areas where a larger lower status of the population live. The relationship between median values of owner-occupied homes and the age of the house is also inversely proportional. Older houses are less expensive.

medv vs rm

```
library(ggplot2)
ggplot(data,aes(x=rm,y=medv))+geom_point() +geom_smooth(method="lm")
```

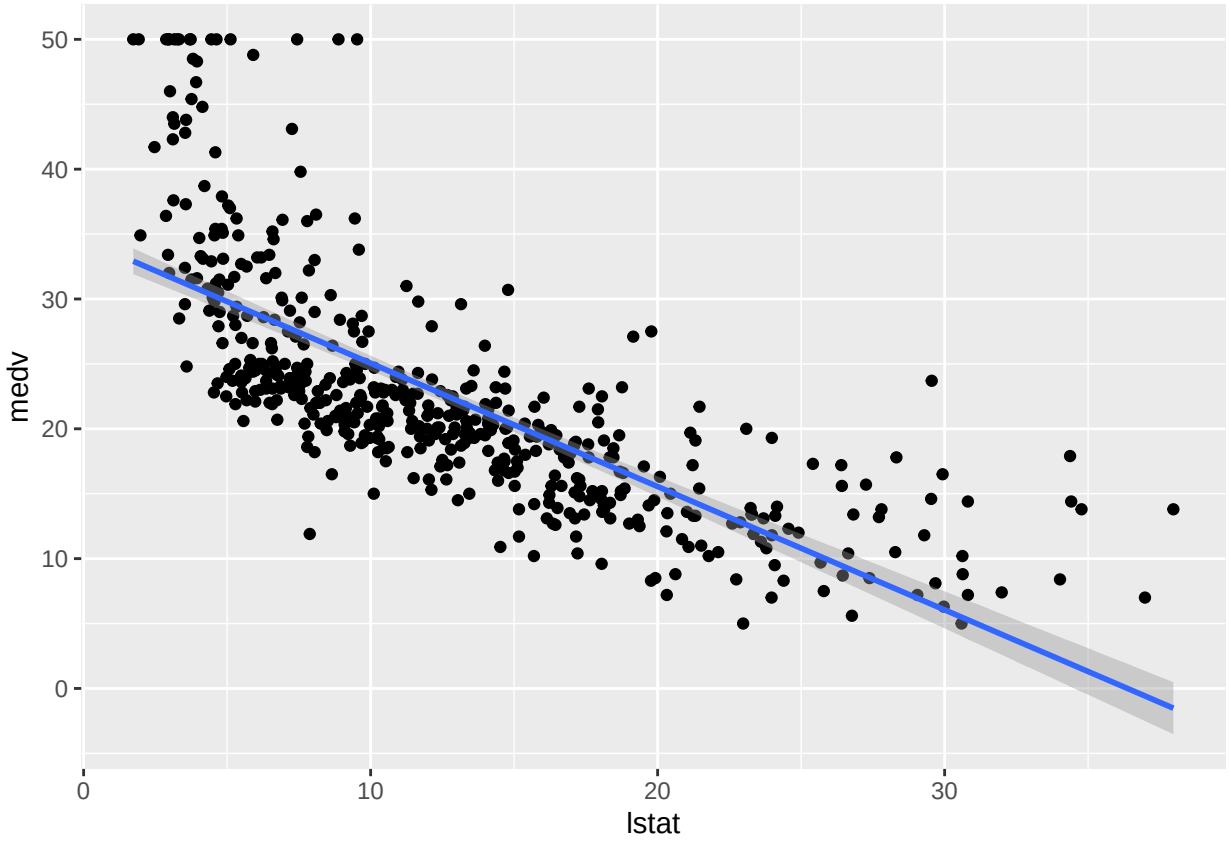
```
## `geom_smooth()` using formula 'y ~ x'
```



medv vs lstat

```
ggplot(data,aes(x=lstat,y=medv))+geom_point() +geom_smooth(method="lm")
```

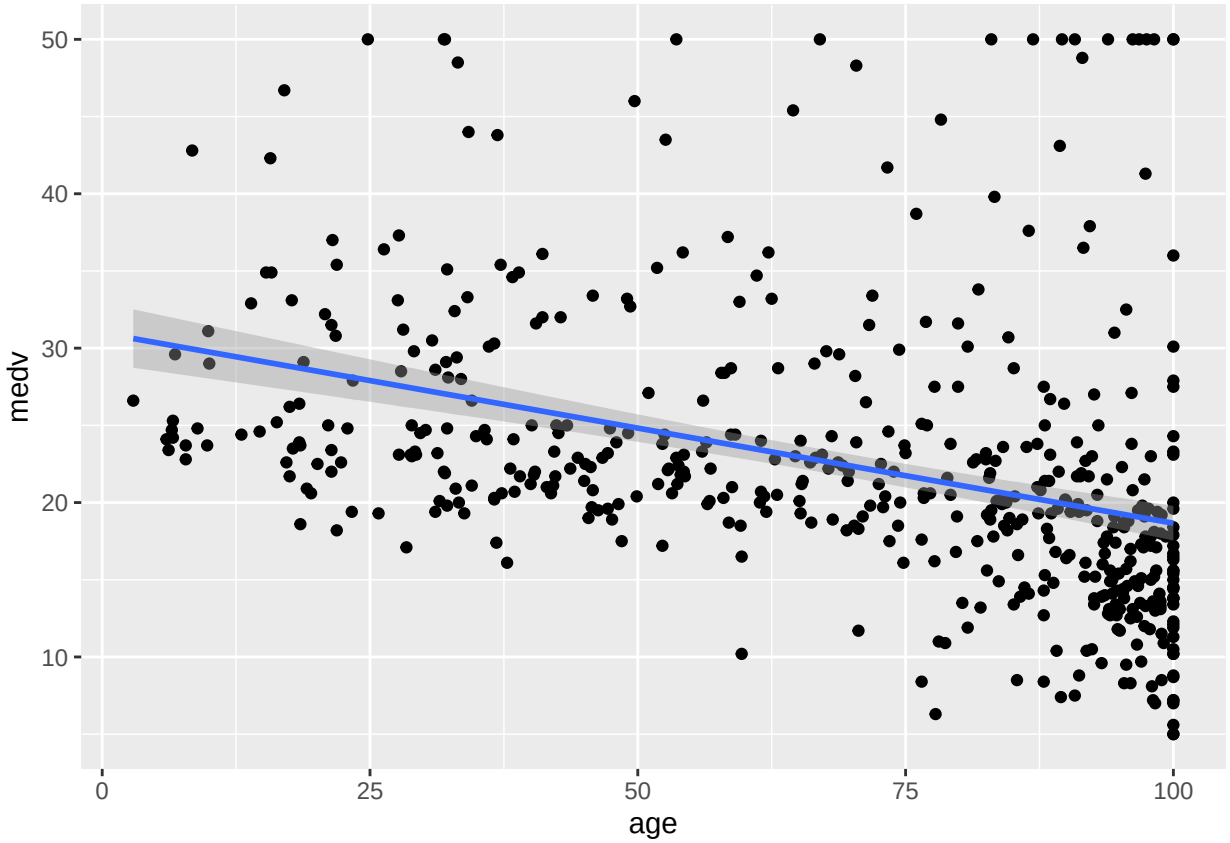
```
## `geom_smooth()` using formula 'y ~ x'
```



medv vs age

```
ggplot(data,aes(x=age,y=medv))+geom_point()+geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Fitting a simple linear regression model to predict the response.

```
model<-lm(medv~rm,data=data)
summary(model)
```

```
##
## Call:
## lm(formula = medv ~ rm, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.346  -2.547   0.090   2.986  39.433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -34.671      2.650  -13.08   <2e-16 ***
## rm             9.102      0.419   21.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.616 on 504 degrees of freedom
## Multiple R-squared:  0.4835, Adjusted R-squared:  0.4825
## F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
model<-lm(medv~lstat,data=data)
summary(model)
```

```
##
## Call:
## lm(formula = medv ~ lstat, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.168  -3.990  -1.318   2.034  24.500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.55384    0.56263   61.41   <2e-16 ***
## lstat       -0.95005    0.03873  -24.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.216 on 504 degrees of freedom
## Multiple R-squared:  0.5441, Adjusted R-squared:  0.5432
## F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
model<-lm(medv~age,data=data)
summary(model)
```

```
##
## Call:
## lm(formula = medv ~ age, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.097  -5.138  -1.958   2.397  31.338
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.97868    0.99911  31.006   <2e-16 ***
## age         -0.12316    0.01348  -9.137   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.527 on 504 degrees of freedom
## Multiple R-squared:  0.1421, Adjusted R-squared:  0.1404
## F-statistic: 83.48 on 1 and 504 DF,  p-value: < 2.2e-16
```

For linear regression, the more significant the slope is from zero, the more significant is the realtionship between dependent and independent variable. The slope for medv vs rm is 9, which is significantly higher than 0, hence it is statistically significant. The slope between medv vs lstat is -0.95, which is not that significant. The slope between medv vs age is -0.123, which is less significant.

I think the reason we see a relationship between the avg no of rooms and the cost is pretty obvious. The value of a home would increase with more number of rooms. I had expected that the crime rate would be inversely proportional to the median value of homes, since higher

valued homes might have better security, but that wasn't the case. There is some relationship between the age of the houses and their costs. Older houses are less expensive than the new ones. This could be because older houses could be considered a liability for people to buy, since they might need renovations and have structural issues which might have to be resolved from time to time. This could be the reason that they are lower priced than the new homes.

The points are more densely packed near the line of regression where there is a higher statistical significance based on the model.

The t-value of rm is 21.72 in this model. This is calculated using this formula:t = (Estimate-null value/std.error). We can show that if H0 is correct, the t value is t-distributed. Here H0 : Estimate = 0. Critical t-value: the critical t value comes from the t-distribution table. Here the critical value tcr = -8.405.

```
model<-lm(medv~rm,data=data)
summary(model)
```

```
##
## Call:
## lm(formula = medv ~ rm, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.346  -2.547   0.090   2.986  39.433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -34.671      2.650  -13.08   <2e-16 ***
## rm             9.102      0.419   21.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.616 on 504 degrees of freedom
## Multiple R-squared:  0.4835, Adjusted R-squared:  0.4825
## F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
qt(p = 2.2e-16, df = 504)
```

```
## [1] -8.405083
```

When we use all predictors, we get adj r square as 0.733, which is a little higher than lstat and rm together. This means that although there might be one or two predictors which are significant stastistcally, rm and lstat are more significant than the others. We can use trial and error to improve the adj r square value. We find ptratio is also statistically significant by trial and error. We can reject the null hypothesis wherever p <0.05. The null hypothesis is rejected for crim, zn, indus, chas, nox, rm, age,rad, dis, ptratio, black, and lstat.

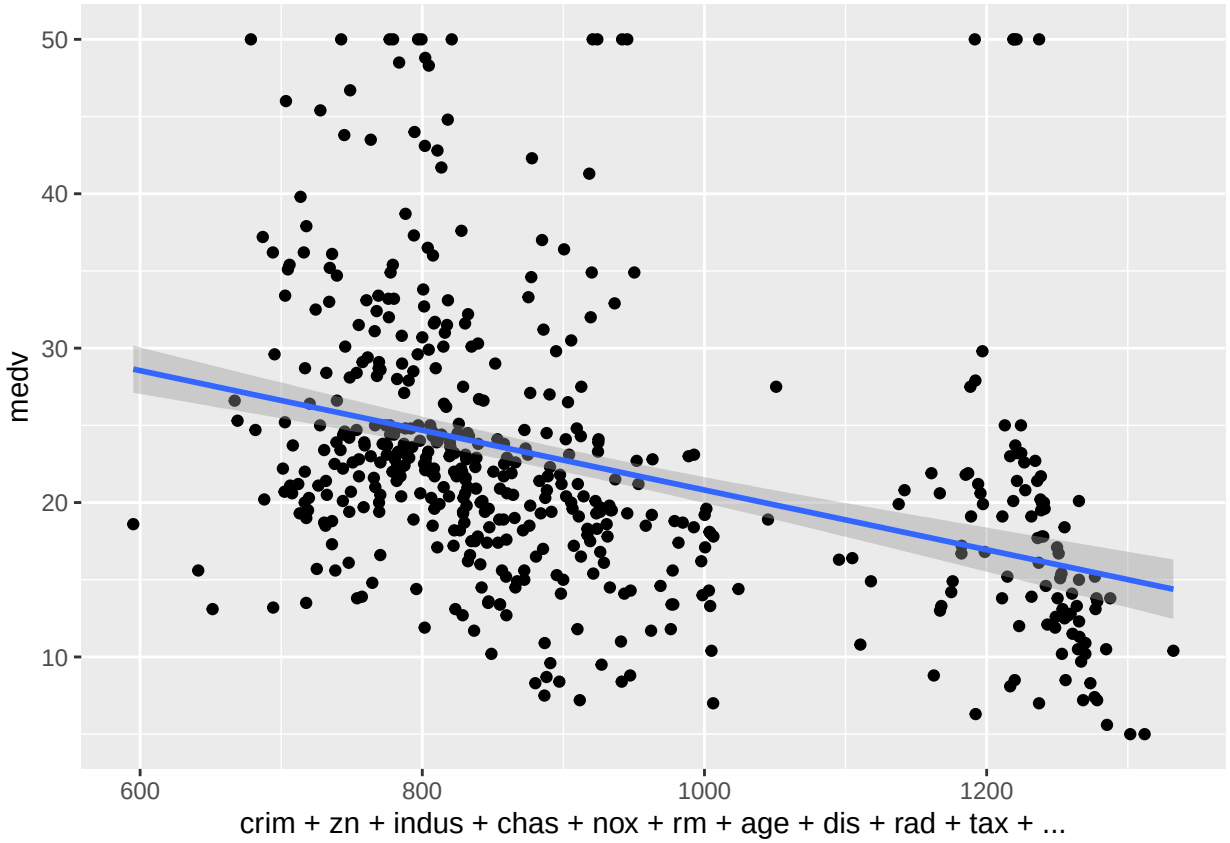
```
##using all predictors
model<-lm(medv~crim+zn+indus+chas+nox+rm+age+dis+rad+tax+ptratio+black+lstat,data=data)
summary(model)
```

```
##
## Call:
```

```
## lm(formula = medv ~ crim + zn + indus + chas + nox + rm + age +
##     dis + rad + tax + ptratio + black + lstat, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.595  -2.730  -0.518   1.777  26.199
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
## crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
## zn           4.642e-02  1.373e-02   3.382 0.000778 ***
## indus        2.056e-02  6.150e-02   0.334 0.738288
## chas         2.687e+00  8.616e-01   3.118 0.001925 **
## nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
## rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
## age          6.922e-04  1.321e-02   0.052 0.958229
## dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
## rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
## tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
## ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
## black        9.312e-03  2.686e-03   3.467 0.000573 ***
## lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.745 on 492 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7338
## F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

```
ggplot(data,aes(x=crim+zn+indus+chas+nox+rm+age+dis+rad+tax+ptratio+black+lstat,y=medv))+geom_point()+g
```

```
## `geom_smooth()` using formula 'y ~ x'
```

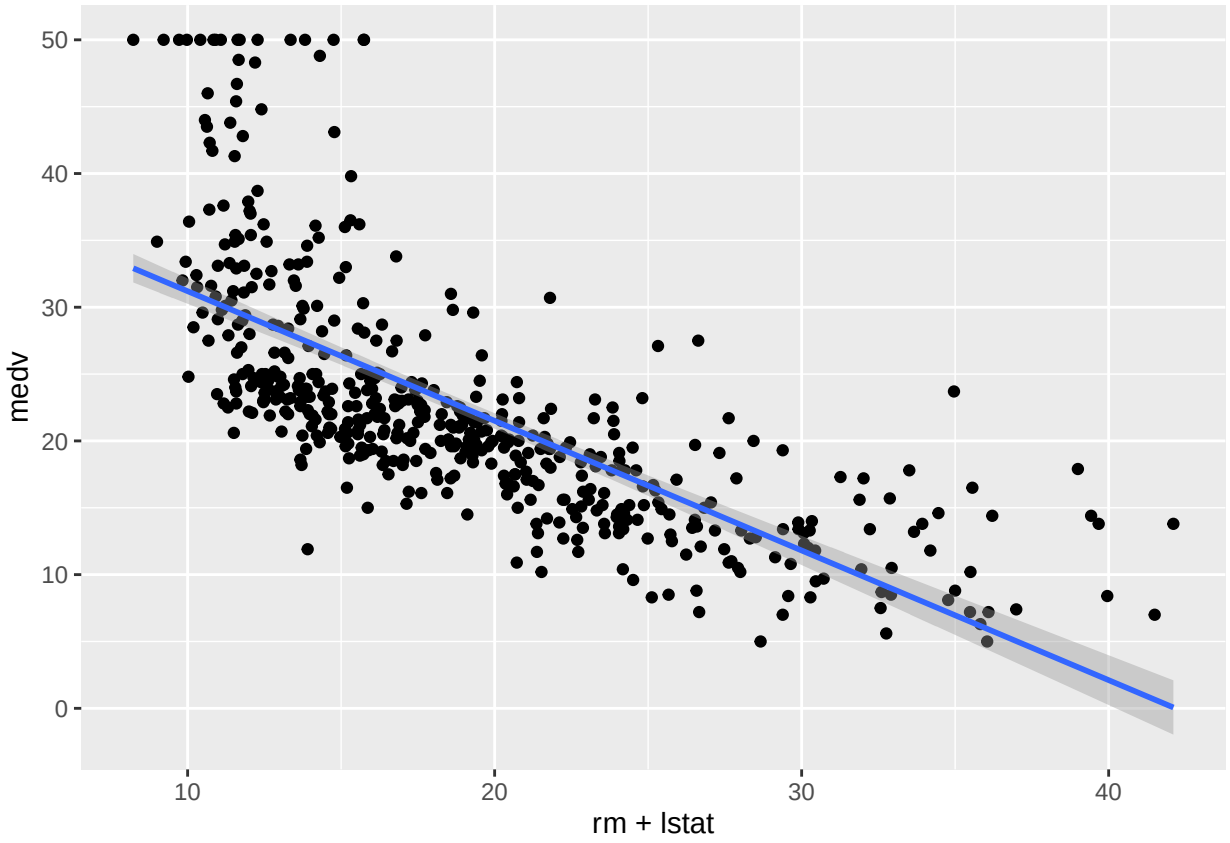



```
model<-lm(medv~rm+lstat,data=data)
summary(model)
```

```
##
## Call:
## lm(formula = medv ~ rm + lstat, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.076  -3.516  -1.010   1.909  28.131
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.35827    3.17283  -0.428    0.669
## rm           5.09479    0.44447  11.463   <2e-16 ***
## lstat       -0.64236    0.04373 -14.689   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.54 on 503 degrees of freedom
## Multiple R-squared:  0.6386, Adjusted R-squared:  0.6371
## F-statistic: 444.3 on 2 and 503 DF,  p-value: < 2.2e-16
```

```
ggplot(data,aes(x=rm+lstat,y=medv))+geom_point()+geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

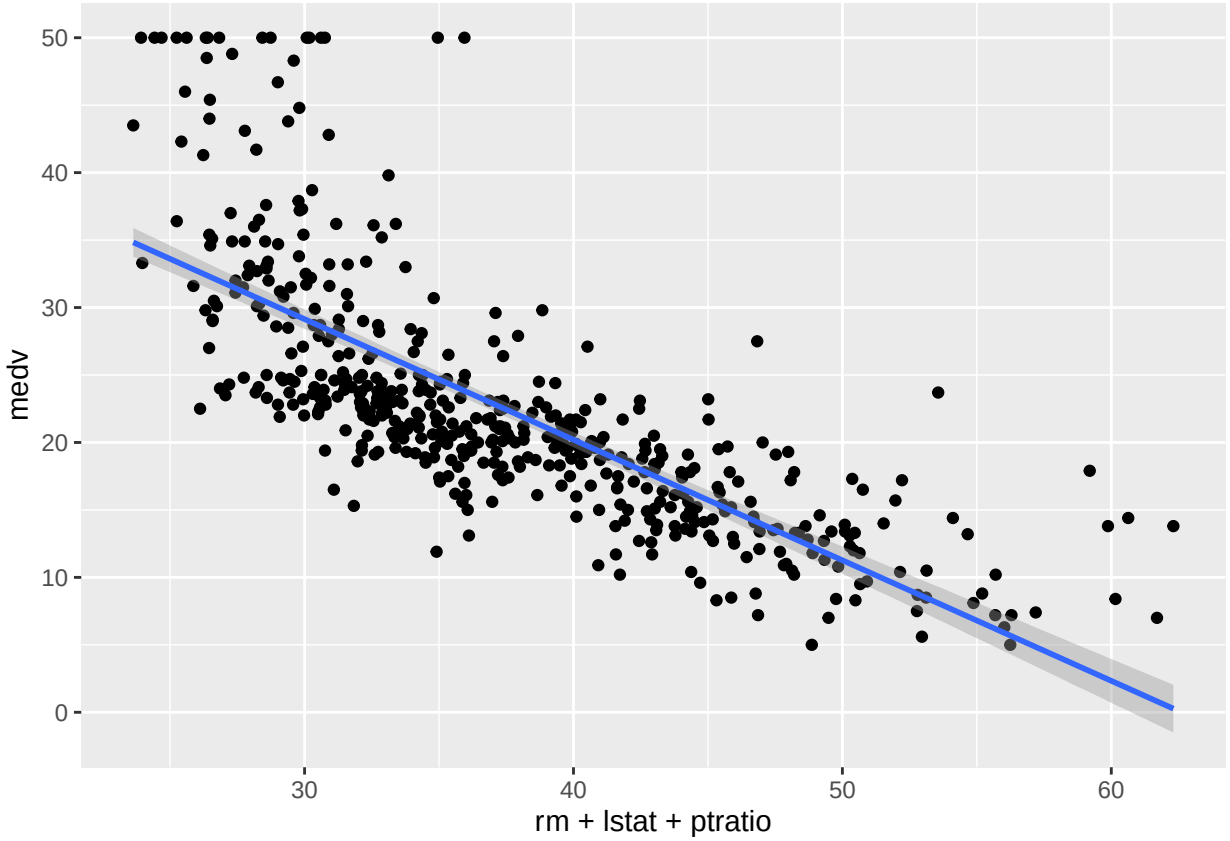



```
model<-lm(medv~rm+lstat+ptratio,data=data)
summary(model)
```

```
##
## Call:
## lm(formula = medv ~ rm + lstat + ptratio, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4871  -3.1047  -0.7976   1.8129  29.6559
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.56711    3.91320   4.745 2.73e-06 ***
## rm           4.51542    0.42587  10.603  < 2e-16 ***
## lstat       -0.57181    0.04223 -13.540  < 2e-16 ***
## ptratio     -0.93072    0.11765  -7.911 1.64e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.229 on 502 degrees of freedom
## Multiple R-squared:  0.6786, Adjusted R-squared:  0.6767
## F-statistic: 353.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
ggplot(data,aes(x=rm+lstat+ptratio,y=medv))+geom_point()+geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
#checking if we can reject the null hypothesis
model<-lm(medv~lstat,data=data)
summary(model)
```

```
##
## Call:
## lm(formula = medv ~ lstat, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.168  -3.990  -1.318   2.034  24.500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.55384    0.56263   61.41   <2e-16 ***
## lstat       -0.95005    0.03873  -24.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.216 on 504 degrees of freedom
```

```
## Multiple R-squared:  0.5441, Adjusted R-squared:  0.5432
## F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

The adjusted r square does not change significantly when indus is added, it is not statistically significant.

```
model<-lm(medv~rm+lstat+indus,data=data)
summary(model)
```

```
##
## Call:
## lm(formula = medv ~ rm + lstat + indus, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.3179  -3.5006  -0.9387   2.0762  28.8816
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.96865    3.18168  -0.304    0.761
## rm           5.07379    0.44428  11.420   <2e-16 ***
## lstat       -0.60671    0.05046 -12.025   <2e-16 ***
## indus       -0.06364    0.04506  -1.412    0.159
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.535 on 502 degrees of freedom
## Multiple R-squared:   0.64,  Adjusted R-squared:  0.6378
## F-statistic: 297.5 on 3 and 502 DF,  p-value: < 2.2e-16
```

When we use multiple regression, the adjusted r square value increases, giving us a better relationship between the predictors and medv. The values differ based on how dependent medv is on the above predictors. The values also differ because a particular predictor could be directly or inversely proportional to medv. It does tell the same story, but multiple regression takes in the values of multiple predictors to answer the question more accurately.

```
model<-lm(medv~lstat,data=data)
summary(model)
```

```
##
## Call:
## lm(formula = medv ~ lstat, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.168  -3.990  -1.318   2.034  24.500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.55384    0.56263   61.41   <2e-16 ***
## lstat       -0.95005    0.03873  -24.53   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.216 on 504 degrees of freedom
## Multiple R-squared:  0.5441, Adjusted R-squared:  0.5432
## F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
model<-lm(medv~rm,data=data)
summary(model)
```

```
##
## Call:
## lm(formula = medv ~ rm, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.346  -2.547   0.090   2.986  39.433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -34.671      2.650  -13.08   <2e-16 ***
## rm             9.102      0.419   21.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.616 on 504 degrees of freedom
## Multiple R-squared:  0.4835, Adjusted R-squared:  0.4825
## F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
model<-lm(medv~ptratio,data=data)
summary(model)
```

```
##
## Call:
## lm(formula = medv ~ ptratio, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.8342  -4.8262  -0.6426   3.1571  31.2303
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.345      3.029   20.58   <2e-16 ***
## ptratio       -2.157      0.163  -13.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.931 on 504 degrees of freedom
## Multiple R-squared:  0.2578, Adjusted R-squared:  0.2564
## F-statistic: 175.1 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
model<-lm(medv~lstat+rm+ptratio,data=data)
summary(model)
```

```
##
## Call:
## lm(formula = medv ~ lstat + rm + ptratio, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4871  -3.1047  -0.7976   1.8129  29.6559
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.56711    3.91320   4.745 2.73e-06 ***
## lstat       -0.57181    0.04223 -13.540  < 2e-16 ***
## rm           4.51542    0.42587  10.603  < 2e-16 ***
## ptratio     -0.93072    0.11765  -7.911 1.64e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.229 on 502 degrees of freedom
## Multiple R-squared:  0.6786, Adjusted R-squared:  0.6767
## F-statistic: 353.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

Interpreting Regression Results

1. Do neighborhoods with more evictions see more or less 311 calls? By how much?

Answer: The number of evictions are correlated to calls for service by 0.048. This demonstrates that the two variables have a positive regression connection. As a result, more 311 calls are made as there are more evictions. With an increase in evictions, there is a noticeable increase in 311 calls.

2. Is the figure statistically significant (at 5% level)?

Answer: Yes, it is statistically significant at 1%, *** means p $<$ .01. The p-value is less than 0.05, which is statistically significant as it lies in the 99% confidence interval.

3. How is poverty rate associated with 311 calls? How much more (or less) calls there are in neighborhoods with 10 pct point more poverty?

Answer: Poverty is negatively correlated to 311 calls, there are less number of calls by 14% in regions where more poverty prevails.

4. What can you tell about association of race (white) and calls?

Answer: There is a negative correlation between the two. In areas where more white people live, there are lesser 311 calls. We can infer that the regression is negative because the p-value for the association between the percentage of white people and 311 calls is negative.Additionally, because the p-value is below 0.05, the model is statistically significant.

5. Is older median age associated with more or less 311 calls? At which level is this statistically significant?

Answer: Older median age is associated with more 311 calls, this is statistically significant at 1% standard error. Because the p-value in this case is less than 0.01 and has a 99% confidence interval, we can say that our model is statistically significant.

6. The value for housing density is -0.13. What does this number mean?

Answer: This means that the housing density is negatively correlated with number of 311 calls. Lesser calls come from homes that have more people staying in them. This value is cross sectional regression value after logging. However, the model is less statistically significant because the absolute p-value is greater than 0.05.

7. The omitted category for city is Austin, TX. Are there more or fewer calls in similar neighborhoods in Philadelphia, compared to Austin? By how much?

Answer: In Philadelphia, there is a negative correlation between the number of calls, compared to Austin. This means more people call 311 in Austin than in Philadelphia by 0.56 and a p-value of less than 0.1, indicating that the model is statistically significant with a 90% confidence interval. Since the absolute p-value is more than 0.05, the relationship is not statistically significant.