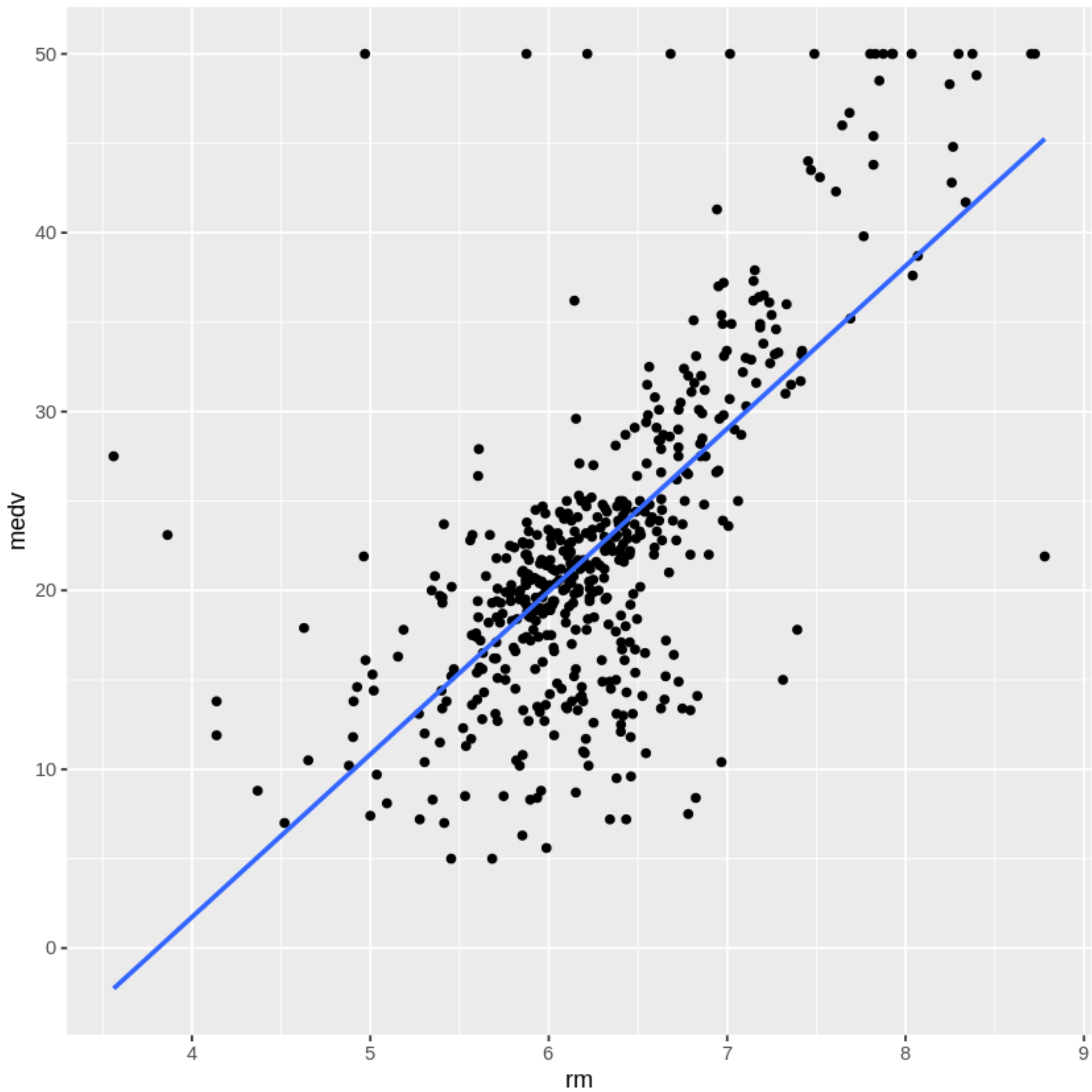
Ezra Cohen lab 8

August 1, 2021

task 1

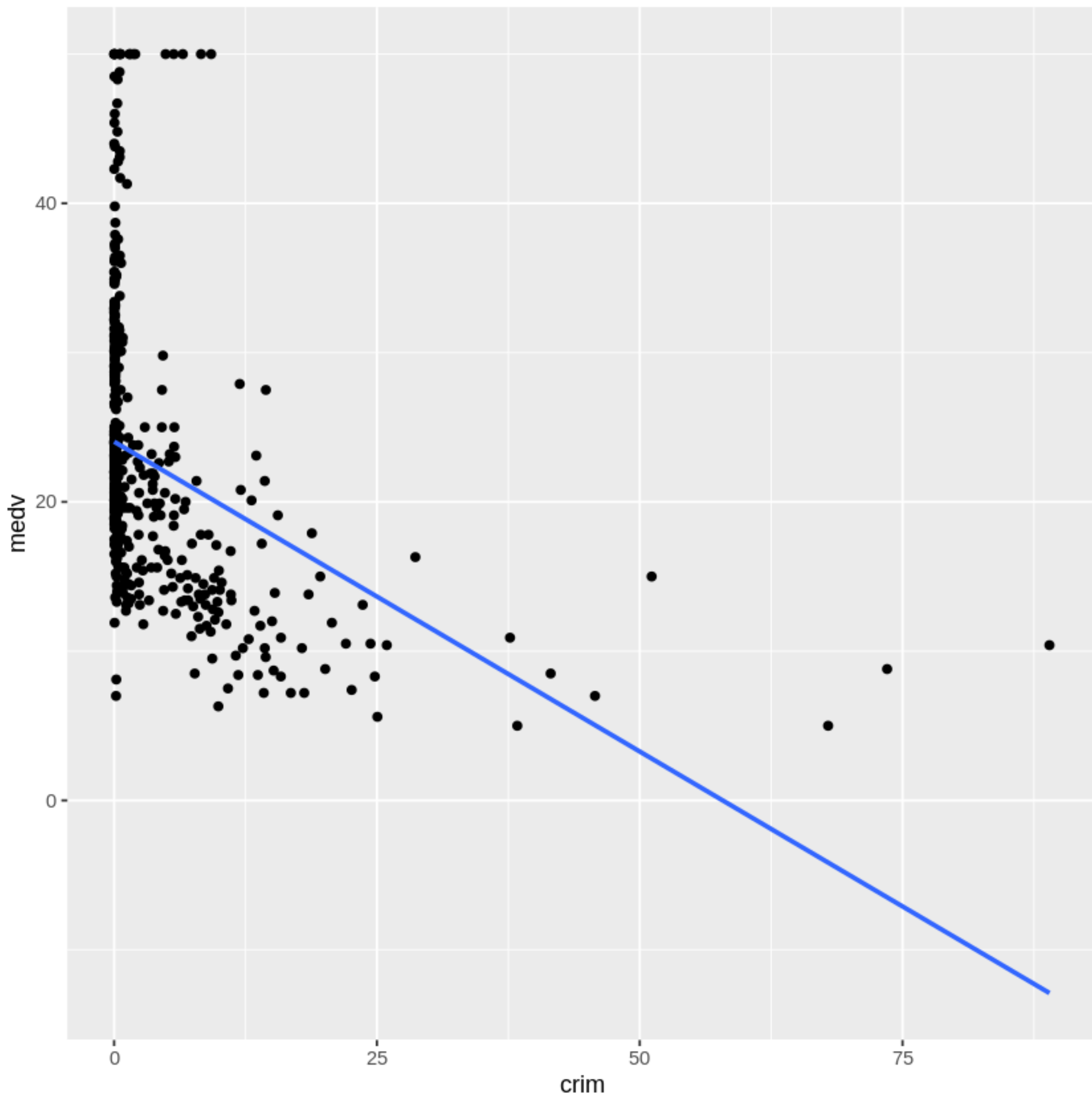
[5]:
```
library(ggplot2)
library(MASS)
ggplot(data=Boston) + aes(x=rm, y=medv) + geom_point() + 
↪geom_smooth(method="lm", se=FALSE)
```

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



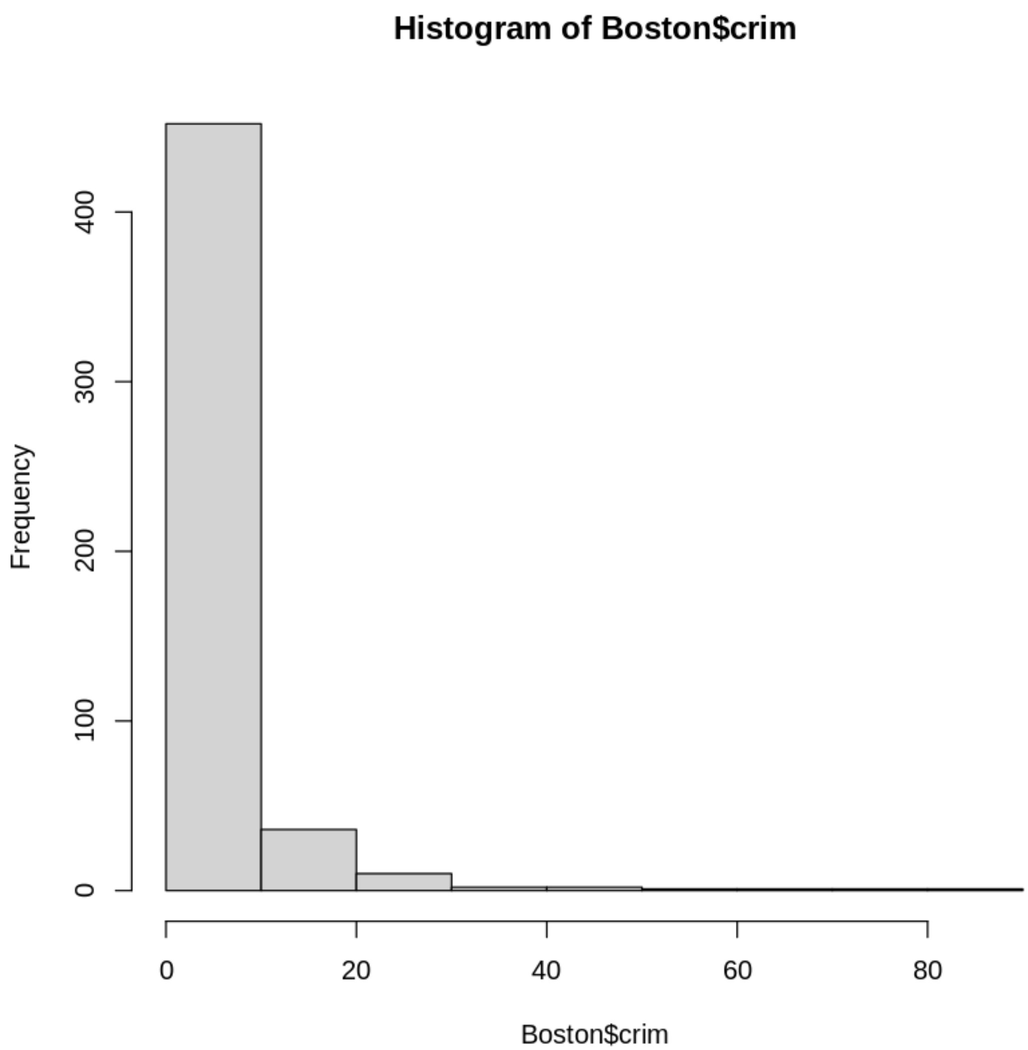
[6]:
```
ggplot(data=Boston) + aes(x=crim, y=medv) + geom_point() + ␣
↪geom_smooth(method="lm", se=FALSE)
```

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



task 2

[7]:
```
hist(Boston$crim)
#In most places the crime rate is very low from 0 to 10, there are a couple of␣
↪places with the crime rate being from 10 to 30, but after that from 30 to 80␣
↪there are incredibly few places with crime rates that high and these are the␣
↪outliers, especially those on the high end from 50 to 80 where there are␣
↪even less places then there are from 30 to 50 and the crime rate is␣
↪incredibly high in those places
```



task 3

[27]:
```
lmout<-lm(formula=medv ~ crim,data=Boston)
summary(lmout)
```

```
Call:
lm(formula = medv ~ crim, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-16.957  -5.449  -2.007   2.512  29.800

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.03311    0.40914   58.74   <2e-16 ***
crim        -0.41519    0.04389   -9.46   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.484 on 504 degrees of freedom
Multiple R-squared:  0.1508,	Adjusted R-squared:  0.1491
F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16
```

task 4

[28]:
```r
lmout2<-lm(formula=medv ~ crim + rm + dis,data=Boston)
summary(lmout2)
```

```
Call:
lm(formula = medv ~ crim + rm + dis, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-21.247  -2.930  -0.572   2.390  39.072

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -29.45838    2.60010 -11.330  < 2e-16 ***
crim         -0.25405    0.03532  -7.193 2.32e-12 ***
rm            8.34257    0.40870  20.413  < 2e-16 ***
dis           0.12627    0.14382   0.878     0.38
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.238 on 502 degrees of freedom
Multiple R-squared:  0.5427,	Adjusted R-squared:  0.5399
F-statistic: 198.6 on 3 and 502 DF,  p-value: < 2.2e-16
```

task 5

[17]:
```r
```

```
#The adjusted r-squared value is only about .54 which is about 55%, so that␣
↪indicates that there is some correlation but not necessarily that it is a␣
↪very high correlation send it is only just above 50%, but the r-squared␣
↪value needs to be examined on a case-by-case basis since there is no set␣
↪amount that is a good value, the P value was 2.2e-16 Which is quite good (I␣
↪did look up on Google if this was a good P value since I didn't know exactly␣
↪what it meant by 2.2e - 16) and it is below .05 which shows that there is␣
↪correlation and that this would be unlikely for the data to have been like␣
↪this if we knew that there was no factor that was causing them to be this␣
↪correlated, the coefficients are the intercep, crim rm and dis, we can see␣
↪that the P value for all of them except distance the last 1 are extremely␣
↪significant meaning that we can reject the null hypothesis, the book really␣
↪didn't go over the rest of the stuff for the coefficients so I Googled it,␣
↪the estimate gives the intercept and slopes, The standard error is about how␣
↪much the predicted values differ from the actual values and it seems like␣
↪that should be low if we want to say there is probably correlation, and all␣
↪the standard errors other than the intercept Are quite low so it again␣
↪supports the fact that there is a likely correlation, the t value is used to␣
↪calculate the P value and while I don't fully understand how to interpret it␣
↪since I know that the P values are very significant for each of them I think␣
↪I can safely say that the T values are all good enough to reject the null␣
↪hypothesis again with the exclusion of distance which has a bad p value
```

task 6

[29]:
```
predDF <- data.frame(crim = 0.26, dis=3.2, rm=6.2)
```

task 7

[31]:
```
predict(lmout2, predDF)
```

1: 22.6035475000185