Homework 1 - Predictive Modeling in Finance and Insurance

By: Dennis Goldenberg

```
# import packages
library(ggplot2)
library(patchwork)
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:patchwork':
##
##     area
```

```
library(magrittr)
```

1. Automobile Insurance Claims

```
# read in Data
auto <- read.table(file = 'AutoClaims-1.csv', header = TRUE, sep = ',')
auto$state <- factor(auto$state, ordered = TRUE)
auto$gender <- factor(auto$gender)
auto$class <- factor(auto$class)
```

1a.

I compute the descriptive statistics for the “PAID” variable:

```
summary(auto$paid)
```

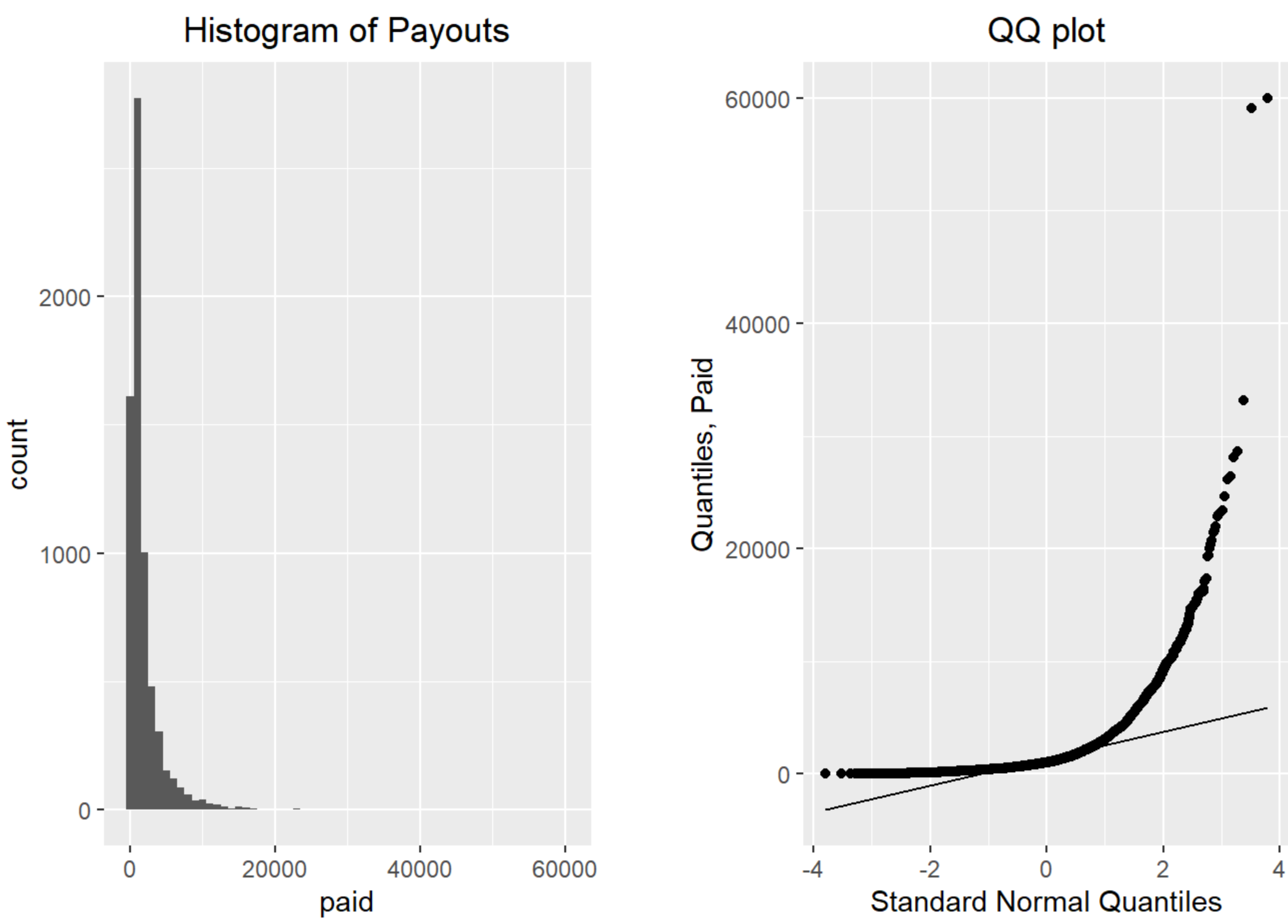
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.5   523.7  1001.7  1853.0  2137.4 60000.0
```

So, the mean paid is $1,853.00 and the median paid is $1,001.70.

1b.

I graph the histogram and qqplot of “paid”, comparing quantiles:

```
hist <- ggplot(data = auto) +
        geom_histogram(aes(paid), binwidth = 1000) +
        labs(title = "Histogram of Payouts") +
        theme(plot.title = element_text(hjust = 0.5))
qq <- ggplot(data = auto) + geom_qq(aes(sample = paid)) +
      geom_qq_line(aes(sample = paid)) +
      labs(x = "Standard Normal Quantiles", y = 'Quantiles, Paid') +
      labs(title = 'QQ plot') +
      theme(plot.title = element_text(hjust = 0.5))
hist + plot_spacer() + qq + plot_layout(widths = c(6,0.5,6))
```

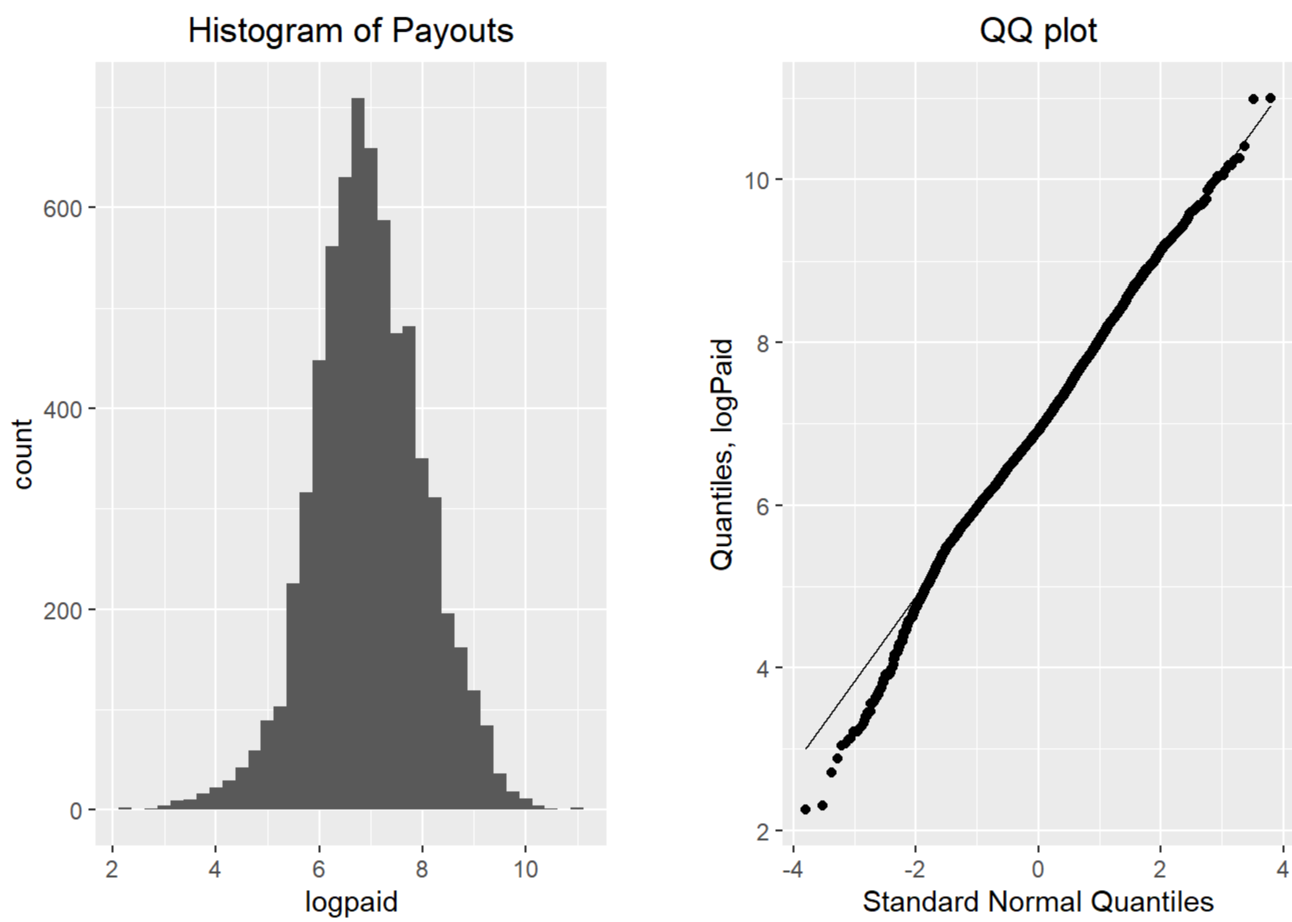


The qqplot suggests that the normal distribution is not a good fit. Based on the histogram, the distribution of paid looks right-skewed, and potentially $\mathbf{lognormal}$.

1c.

I graph the histogram and qqplot of “logpaid”, comparing quantiles:

```
hist <- ggplot(data = auto) +
        geom_histogram(aes(logpaid), binwidth = 0.25) +
        labs(title = "Histogram of Payouts") +
        theme(plot.title = element_text(hjust = 0.5))
qq <- ggplot(data = auto) + geom_qq(aes(sample = logpaid)) +
      geom_qq_line(aes(sample = logpaid)) +
    labs(x = "Standard Normal Quantiles", y = 'Quantiles, logPaid') +
      labs(title = 'QQ plot') +
      theme(plot.title = element_text(hjust = 0.5))
hist + plot_spacer() + qq + plot_layout(widths = c(6,0.5,6))
```

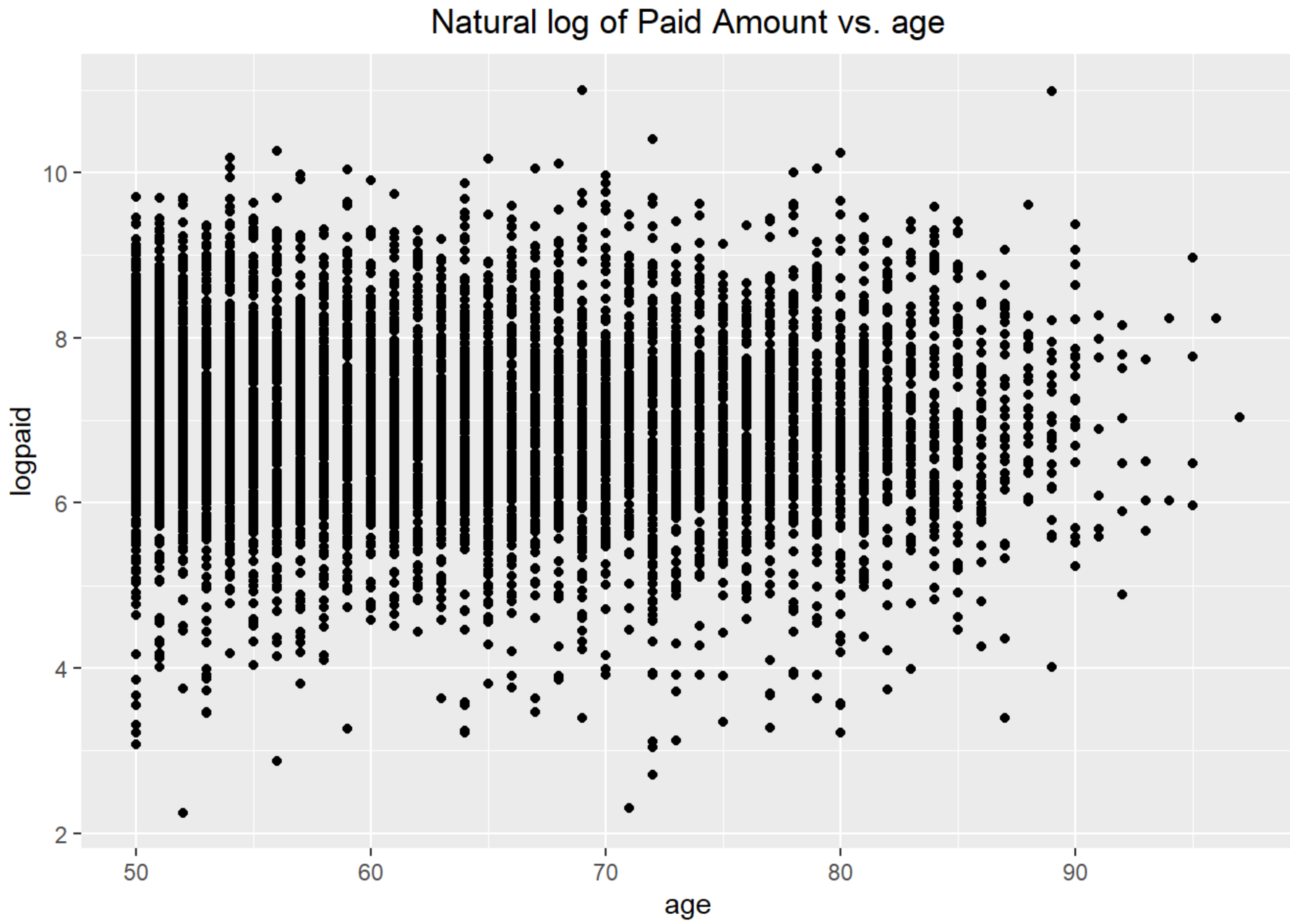


Based on the good fit in the qqplot, and the general shape in the histogram, the distribution of “logpaid” seems to be $\textbf{approximately normal}$. This would suggest that the distribution of paid is $\textbf{lognormal}$.

1d.

I graph the scatterplot of “logpaid” against “age”:

```
scat <- ggplot(data = auto) + geom_point(aes(x = age, y = logpaid)) +
        labs(title = "Natural log of Paid Amount vs. age") +
        theme(plot.title = element_text(hjust = 0.5))
scat
```

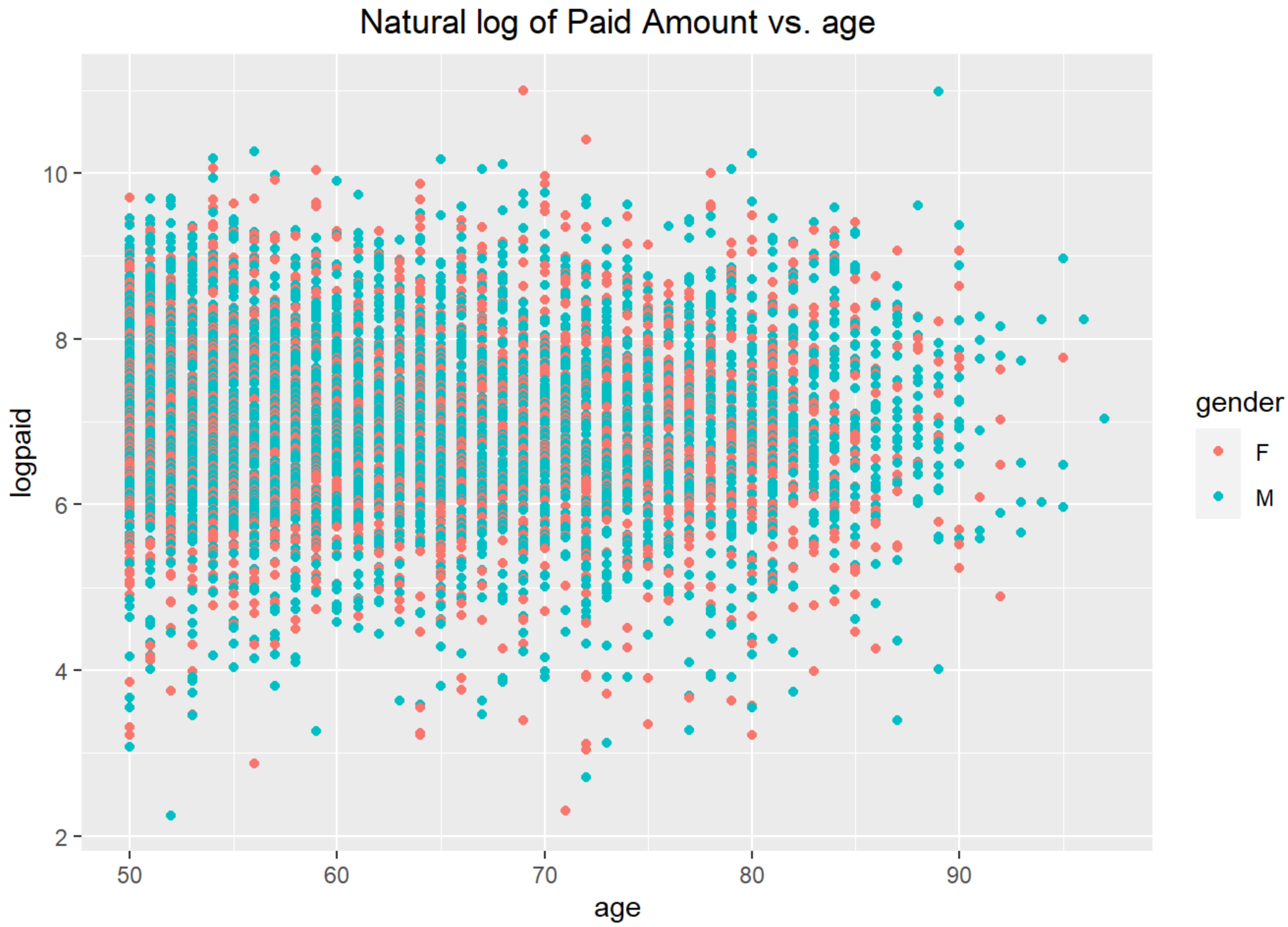


This plot suggests that the variance of log payments for individuals on the younger side (closer to 50) is higher than the variance of payments for older individuals; outside of this, however, there appears to be **no discernable relationship** between the natural log of the paid amount for an automobile insurance claim and the age of an individual.

1e.

I account for gender in the next scatter plot:

```
scat <- ggplot(data = auto) +
        geom_point(aes(x = age, y = logpaid, color = gender)) +
        labs(title = "Natural log of Paid Amount vs. age") +
        theme(plot.title = element_text(hjust = 0.5))
scat
```



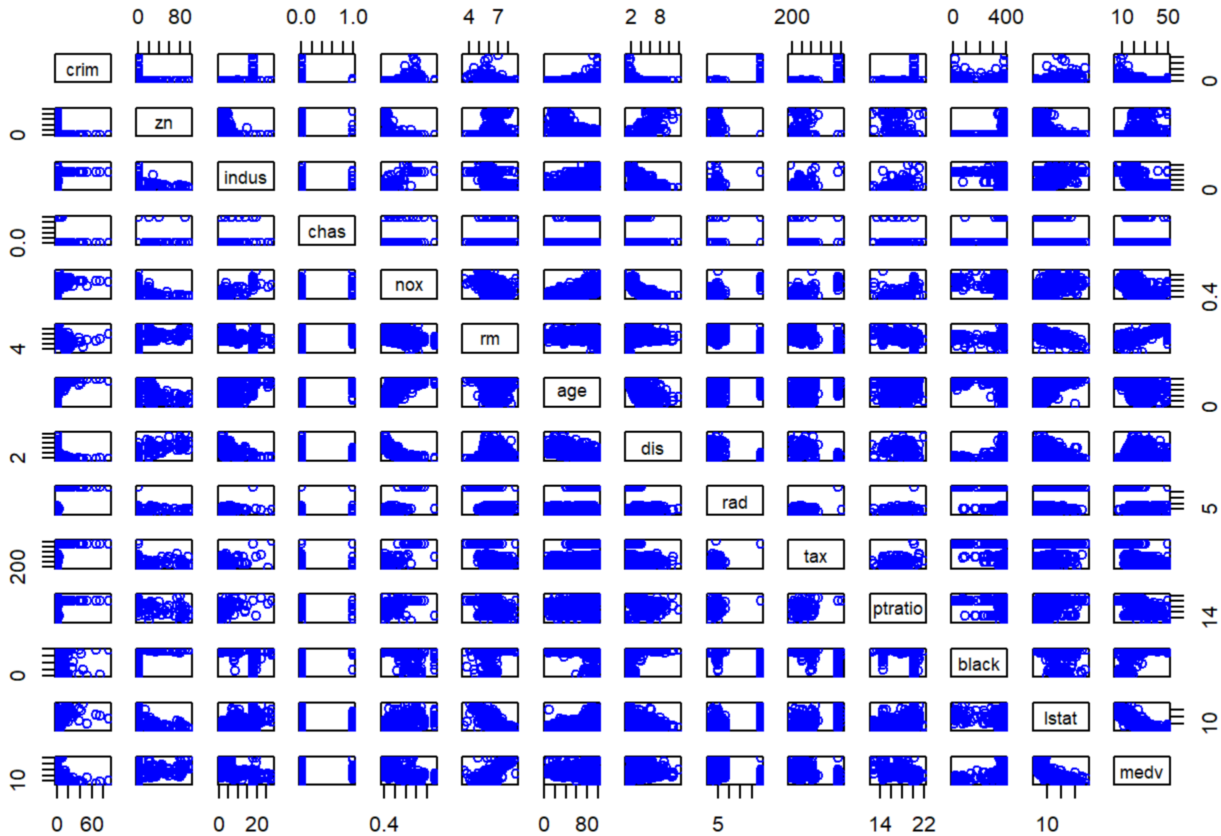
Based on the scatter plot above, there is **no discernable relationship** between the natural log of the paid amount for an automobile insurance claim and the gender of an individual.

2. Boston Housing Dataset

2a.

I first show to pairwise scatter plots of all combinations of variables:

```
pairs(Boston, col = 'blue')
```



This output is too small to discern any patterns; I look at the correlation matrix to see which scatter plots may have a strong relationship. I find where the correlations are strongest by generating a correlation matrix and breaking it into two to see all columns:

```
cormatrix <- cor(Boston)
twocor <- matrix(as.numeric(sprintf(cormatrix, fmt = '%#.2f')), nrow = dim(cormatrix)[1])
rownames(twocor) <- rownames(cormatrix)
colnames(twocor) <- colnames(cormatrix)
twocor[,1:7]
```

```
##          crim    zn indus  chas   nox    rm   age
## crim     1.00 -0.20  0.41 -0.06  0.42 -0.22  0.35
## zn      -0.20  1.00 -0.53 -0.04 -0.52  0.31 -0.57
## indus    0.41 -0.53  1.00  0.06  0.76 -0.39  0.64
## chas    -0.06 -0.04  0.06  1.00  0.09  0.09  0.09
## nox      0.42 -0.52  0.76  0.09  1.00 -0.30  0.73
## rm      -0.22  0.31 -0.39  0.09 -0.30  1.00 -0.24
## age      0.35 -0.57  0.64  0.09  0.73 -0.24  1.00
## dis     -0.38  0.66 -0.71 -0.10 -0.77  0.21 -0.75
## rad      0.63 -0.31  0.60 -0.01  0.61 -0.21  0.46
## tax      0.58 -0.31  0.72 -0.04  0.67 -0.29  0.51
## ptratio  0.29 -0.39  0.38 -0.12  0.19 -0.36  0.26
## black   -0.39  0.18 -0.36  0.05 -0.38  0.13 -0.27
## lstat    0.46 -0.41  0.60 -0.05  0.59 -0.61  0.60
## medv    -0.39  0.36 -0.48  0.18 -0.43  0.70 -0.38
```

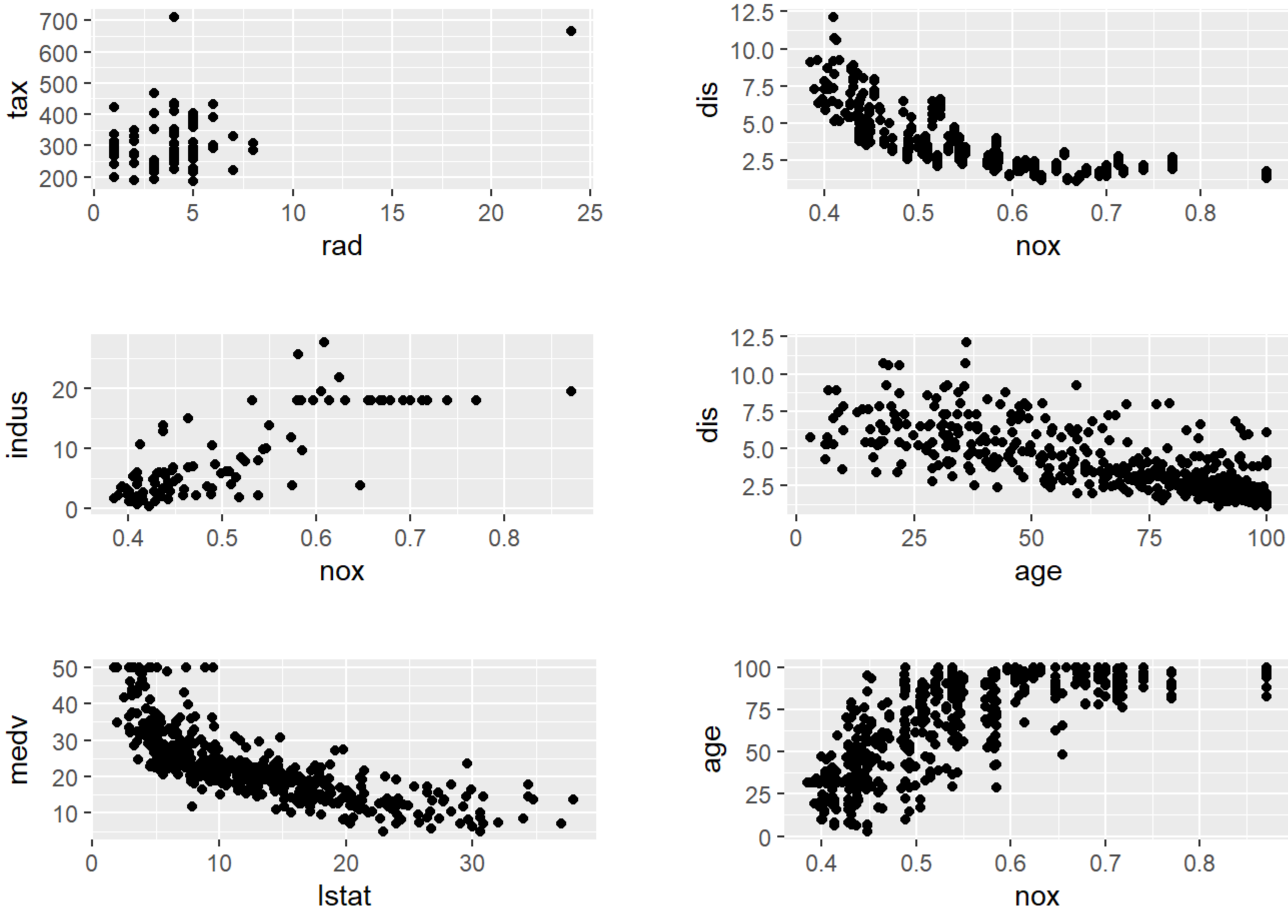
```
twocor[,8:14]
```

```
##           dis   rad   tax ptratio black lstat  medv
## crim    -0.38  0.63  0.58    0.29 -0.39  0.46 -0.39
## zn       0.66 -0.31 -0.31   -0.39  0.18 -0.41  0.36
## indus   -0.71  0.60  0.72    0.38 -0.36  0.60 -0.48
## chas    -0.10 -0.01 -0.04   -0.12  0.05 -0.05  0.18
## nox     -0.77  0.61  0.67    0.19 -0.38  0.59 -0.43
## rm       0.21 -0.21 -0.29   -0.36  0.13 -0.61  0.70
## age     -0.75  0.46  0.51    0.26 -0.27  0.60 -0.38
## dis      1.00 -0.49 -0.53   -0.23  0.29 -0.50  0.25
## rad     -0.49  1.00  0.91    0.46 -0.44  0.49 -0.38
## tax     -0.53  0.91  1.00    0.46 -0.44  0.54 -0.47
## ptratio -0.23  0.46  0.46    1.00 -0.18  0.37 -0.51
## black    0.29 -0.44 -0.44   -0.18  1.00 -0.37  0.33
## lstat   -0.50  0.49  0.54    0.37 -0.37  1.00 -0.74
## medv     0.25 -0.38 -0.47   -0.51  0.33 -0.74  1.00
```

I pick the 6 pairs with the highest correlation to plot:

```
radtax <- ggplot(Boston) + geom_point(aes(x = rad, y = tax))
noxdis <- ggplot(Boston) + geom_point(aes(x = nox, y = dis))
noxindus <- ggplot(Boston) + geom_point(aes(x = nox, y = indus))
agedis <- ggplot(Boston) + geom_point(aes(x = age, y = dis))
lstatmedv <- ggplot(Boston) + geom_point(aes(x = lstat, y = medv))
noxage <- ggplot(Boston) + geom_point(aes(x = nox, y = age))

(radtax+plot_spacer()+noxdis+plot_layout(widths = c(6,0.5,6)))/
  plot_spacer()/
  (noxindus+plot_spacer()+agedis+plot_layout(widths = c(6,0.5,6)))/
  plot_spacer()/
  (lstatmedv+plot_spacer()+noxage+plot_layout(widths = c(6,0.5,6))) +
  plot_layout(heights = c(6, 0.5, 6, 0.5, 6))
```



Here, it is notable that “rad”, or accessibility to highways, is $\textbf{directly related}$ with “tax”, or property tax rate per \$10,000. Further, “nox”, or nitric oxide concentration, is $\textbf{inversely related}$ with “dis”, or distance to employment centers with the relationship looking like exponential decay, but $\textbf{directly related}$ with “indus”, or proportion of non-retail business acres, and “age”, or proportion of buildings built before 1940. Finally, “dis” is $\textbf{inversely related}$ with “age”, and “medv”, or the median value of owner occupied homes, is $\textbf{inversely related}$ to “lstat”, or percent of the population in lower status.

2b.

I check the correlation matrix with regards to crime rate:

```
twocor[1:7,'crim']
```

```
##  crim    zn indus  chas   nox    rm   age
##  1.00 -0.20  0.41 -0.06  0.42 -0.22  0.35
```

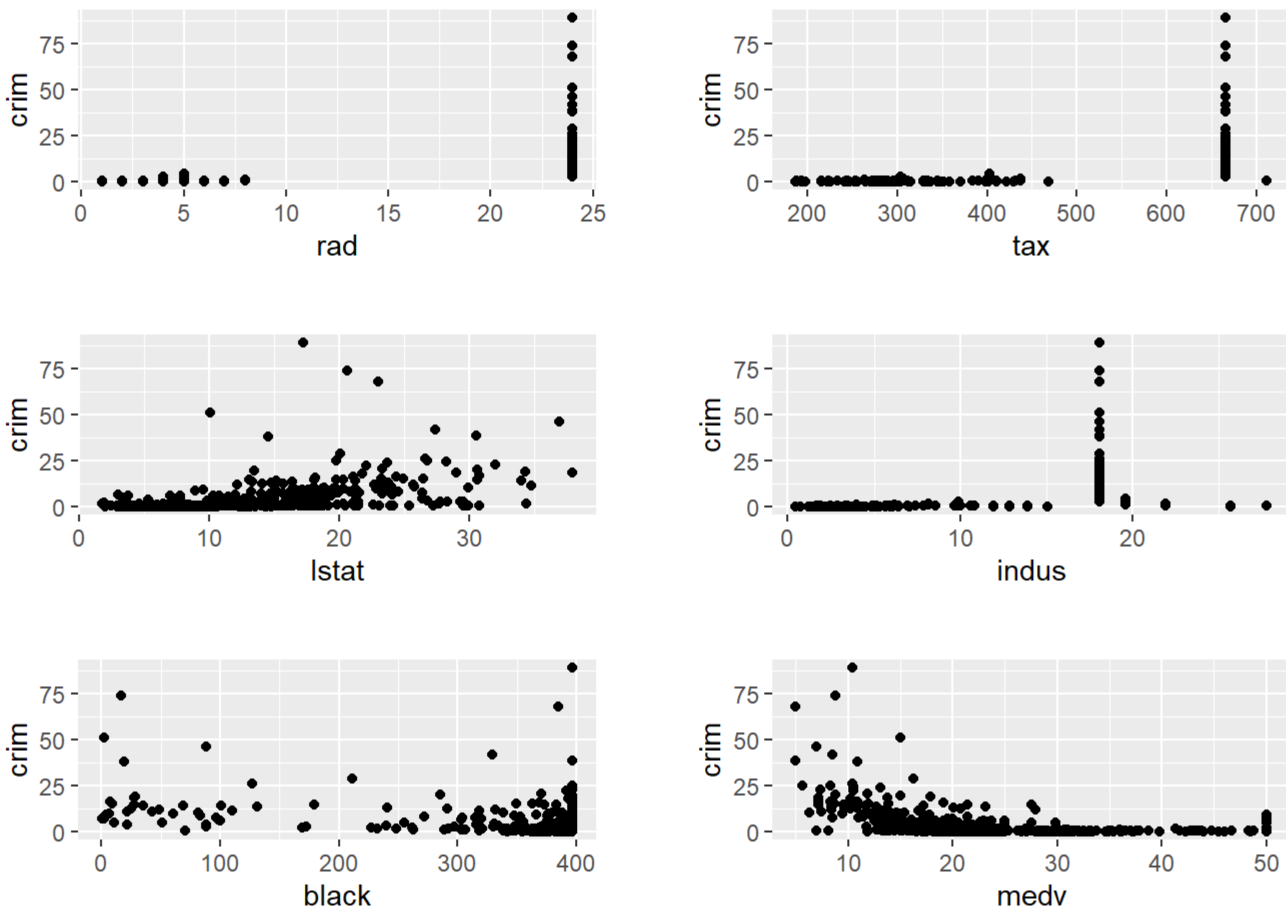
```
twocor[8:14,'crim']
```

```
##     dis     rad     tax ptratio   black   lstat    medv
##   -0.38    0.63    0.58    0.29   -0.39    0.46   -0.39
```

I pick the 6 variables with the strongest correlation (excluding nox, to be evaluated later):

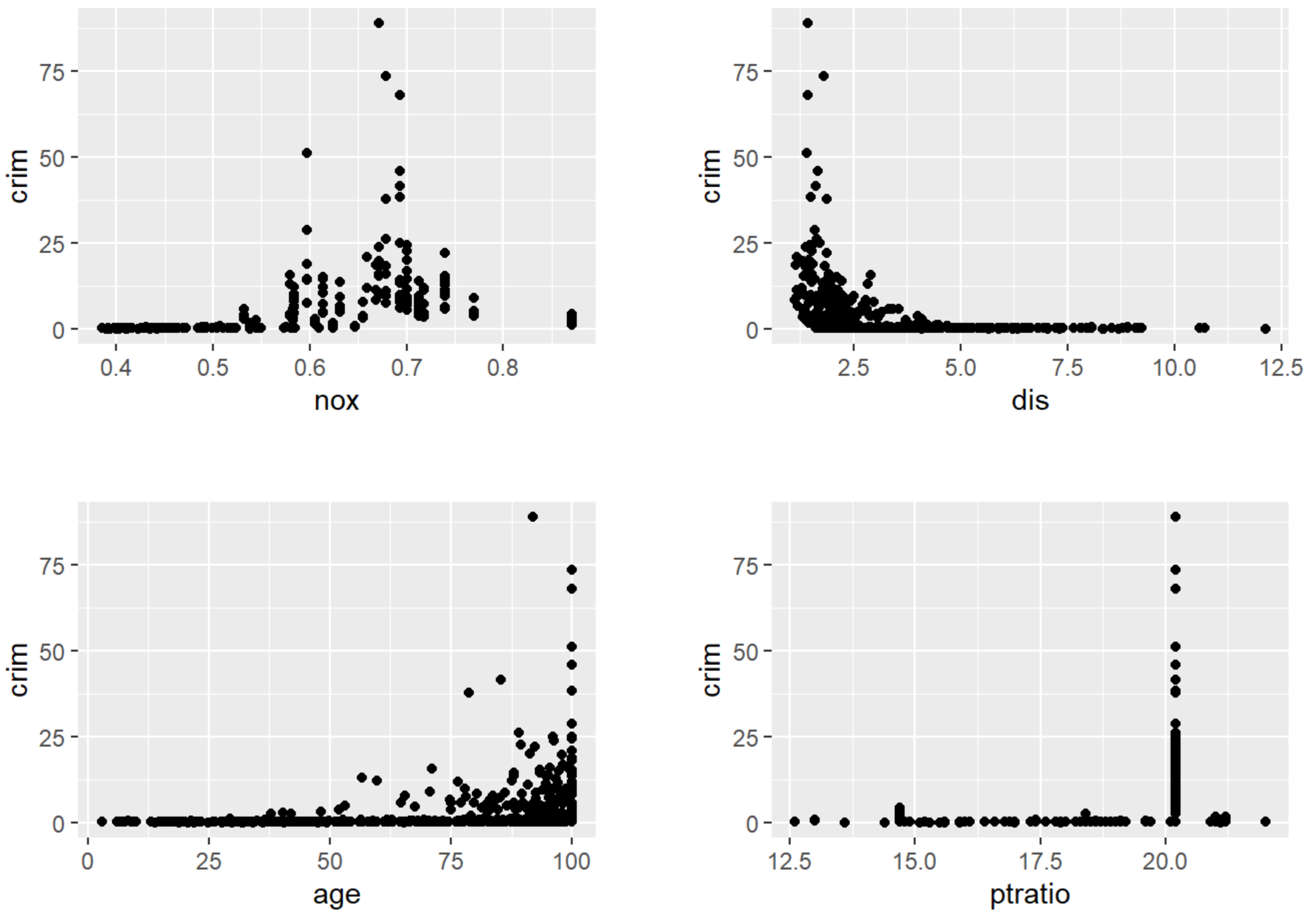
```
crimrad <- ggplot(Boston) + geom_point(aes(x = rad, y = crim))
crimtax <- ggplot(Boston) + geom_point(aes(x = tax, y = crim))
crimlstat <- ggplot(Boston) + geom_point(aes(x = lstat, y = crim))
crimindus <- ggplot(Boston) + geom_point(aes(x = indus, y = crim))
crimblack <- ggplot(Boston) + geom_point(aes(x = black, y = crim))
crimmedv <- ggplot(Boston) + geom_point(aes(x = medv, y = crim))

(crimrad+plot_spacer()+crimtax+plot_layout(widths = c(6,0.5,6)))/
  plot_spacer()/
(crimlstat+plot_spacer()+crimindus+plot_layout(widths = c(6,0.5,6)))/
  plot_spacer()/
(crimblack+plot_spacer()+crimmedv+plot_layout(widths = c(6,0.5,6))) +
  plot_layout(heights = c(6, 0.5, 6, 0.5, 6))
```



The strongest relationships with per capital crime rate, according to these graphs, are ‘lstat”, ’black’, and ‘medv’. It seems as though there is a **direct relationship** between ‘lstat’ and the crime rate; that is, the higher the proportion of a population is of lower status, the higher there are crime rates. Meanwhile, though these relationships appear weaker, there is an **inverse relationship** between crime rates and the proportion of population that is black and the median value of owner occupied homes. I next graph some of the other variables with relation to crim to see if there is a relationship:

```
crimnox <- ggplot(data = Boston) + geom_point(aes(x = nox, y = crim))
crimdis <- ggplot(data = Boston) + geom_point(aes(x = dis, y = crim))
crimage <- ggplot(data = Boston) + geom_point(aes(x = age, y = crim))
crimr <- ggplot(data = Boston) + geom_point(aes(ptratio,crim))
((crimnox+plot_spacer()+crimdis+plot_layout(widths = c(6,0.5,6)))/
  plot_spacer()/
  (crimage+plot_spacer()+crimr+plot_layout(widths = c(6,0.5,6))))+
  plot_layout(heights = c(6, 0.5, 6))
```



Here, it appears as though crime rate has a **direct relationship** with nitric oxide concentration; the areas with the highest crime rates tend to have higher crime rates; it is also true that the areas with the highest crime rates tend to have higher proportions of buildings built before 1940. Alternatively, the crime rate seems to be **inversely related** to distance to employment centers; the areas closest to the employment centers tend to have the highest crime rates.

2c.

The maximum, minimum, mean, and range of each variable is listed below:

```
max_min <- cbind(apply(Boston, 2, min), apply(Boston, 2, mean),
                 apply(Boston, 2, max))
max_min2 <- cbind(max_min, max_min[,3] - max_min[,1])
colnames(max_min2) <- c("Min", "Mean", "Max", "Range")
max_min2
```

```
##               Min         Mean      Max     Range
## crim      0.00632   3.61352356  88.9762  88.96988
## zn        0.00000  11.36363636 100.0000 100.00000
## indus     0.46000  11.13677866  27.7400  27.28000
## chas      0.00000   0.06916996   1.0000   1.00000
## nox       0.38500   0.55469506   0.8710   0.48600
## rm        3.56100   6.28463439   8.7800   5.21900
## age       2.90000  68.57490119 100.0000  97.10000
## dis       1.12960   3.79504269  12.1265  10.99690
## rad       1.00000   9.54940711  24.0000  23.00000
## tax     187.00000 408.23715415 711.0000 524.00000
## ptratio  12.60000  18.45553360  22.0000   9.40000
## black     0.32000 356.67403162 396.9000 396.58000
## lstat     1.73000  12.65306324  37.9700  36.24000
## medv      5.00000  22.53280632  50.0000  45.00000
```

The most notable results from this are the following: there is a wide range of tax rates and crime rates, because there is no theoretical ceiling as to how high they can go. Meanwhile, nitric oxide concentration has a much smaller range, but smaller movements in this predictor could be potentially more important due to the environmental changes caused. The black variable is a formula with a squared term, contributing to its large range. However, the age, zn, and indus variables are all proportions, so their theoretical range can only go from 0 to 100. Finally, though it seems that the median value of homes, or medv, has a smaller range than other variables, that it is because it was recorded in terms of $1000s; in reality, it actually has the highest range, even higher than tax rates or crime rates. Next, I calculate the number of suburbs with one of the following: a crime rate per capita above 60, a tax rate per $10,000 above 600, or a pupil-teacher ratio above 20:

```
hc <- as.numeric(Boston['crim'] > 60) %>% sum
ht <- as.numeric(Boston['tax'] > 600) %>% sum
hpt <- as.numeric(Boston['ptratio'] > 20) %>% sum
highs <- as.matrix(c(hc, ht, hpt))
rownames(highs) <- c("High Crime","High Tax", "High PT")
highs
```

```
##            [,1]
## High Crime    3
## High Tax    137
## High PT     201
```

There are $\mathbf{3}$ suburbs with a crime rate above 60, $\mathbf{137}$ with a tax rate about 600, and $\mathbf{201}$ with a pupil-teacher ratio above 20.

2d.

The variable representing whether or not a suburb bounds the Charles river is ‘chas’; it is a simple binary variable, with 1 indicating that the suburb bounds the river. I check how many satisfy this in total by summing the column:

```
Boston[['chas']] %>% sum
```

```
## [1] 35
```

So, $\mathbf{35}$ suburbs bound the Charles river.

2e.

I check the summary of the ptratio:

```
summary(Boston['ptratio'])
```

```
##     ptratio
##  Min.   :12.60
##  1st Qu.:17.40
##  Median :19.05
##  Mean   :18.46
##  3rd Qu.:20.20
##  Max.   :22.00
```

The median pupil-teacher ratio is $\mathbf{19.05}$ students per teacher.

2f.

```
index <- which.min(Boston[['medv']])
as.vector(Boston[index,])
```

```
## $crim
## [1] 38.3518
##
## $zn
## [1] 0
##
## $indus
## [1] 18.1
##
## $chas
## [1] 0
##
## $nox
## [1] 0.693
##
## $rm
## [1] 5.453
##
## $age
## [1] 100
##
## $dis
## [1] 1.4896
##
## $rad
## [1] 24
##
## $tax
## [1] 666
##
## $ptratio
## [1] 20.2
##
## $black
## [1] 396.9
##
## $lstat
## [1] 30.59
##
## $medv
## [1] 5
```

The 399th suburb (value of index) has the lowest median value of owner-occupied homes, at $5,000. It achieves maximum possible values for the “age” (100% of homes in this suburb were built before 1940), black, and “rad” (accessibility to highways) variables. It has an above average nitric oxide concentration, tax rate, pupil-teacher ratio, non-retail business acre use rate (‘indus’), weighted distance to employment centers (‘dis’), and percent of inhabitants in lower status (‘lstat’). It is below average for number of rooms per dwelling, and has no residential area zoned for lots above 25,000 square feet. It seems like a difficult area to live in.

2g.

I calculate the desired quantity by looking at the amounts asked about:

```
g7 <- Boston[Boston['rm'] > 7,]
g8 <- Boston[Boston['rm'] > 8,]
counts <- c(dim(g7)[1], dim(g8)[1])
counts
```

```
## [1] 64 13
```

There are 64 suburbs that average more than seven rooms per dwelling, and 13 that average more than eight rooms per dwelling. I then examine some of the characteristics of these subsets of the data. I first examine the minimum, mean, max, and range for those with more than seven rooms per dwelling:

```
mmg7 <- cbind(apply(g7, 2, min), apply(g7, 2, mean),
                 apply(g7, 2, max))
mmg72 <- cbind(mmg7, mmg7[,3] - mmg7[,1])
colnames(mmg72) <- c("Min", "Mean", "Max", "Range")
mmg72
```

```
##               Min        Mean      Max     Range
## crim      0.00906   0.9791089  19.6091  19.60004
## zn        0.00000  28.1718750  95.0000  95.00000
## indus     0.46000   5.7756250  19.5800  19.12000
## chas      0.00000   0.1250000   1.0000   1.00000
## nox       0.39400   0.5044547   0.7180   0.32400
## rm        7.00700   7.5700937   8.7800   1.77300
## age       8.40000  60.6406250 100.0000  91.60000
## dis       1.20240   4.1996172   9.2229   8.02050
## rad       1.00000   5.9843750  24.0000  23.00000
## tax     193.00000 312.2343750 666.0000 473.00000
## ptratio  12.60000  16.2593750  20.2000   7.60000
## black   354.31000 388.2751563 396.9000  42.59000
## lstat     1.73000   5.4740625  16.7400  15.01000
## medv     15.00000  38.3968750  50.0000  35.00000
```

On average, these suburbs contain lower amounts of the following: crime (and none of the highest crime suburbs have more than 7 rooms per dwelling), proportion of the land set aside for non-residential business, proportion of buildings built pre 1940, pupil-teacher ratio, proportion of lower status people, nitric oxide concentration, accessibility to highways, and tax rate. On the other hand, these suburbs are higher in the following: proportion of land set aside for >25,000 acre plots, likelihood of bounding the Charles River, variable associated with proportion of the population that is black, and median value of owner occupied homes. Most of these results are in line with more valuable properties being in said suburbs, as the average dwelling is bigger than average among all suburbs. I repeat this for suburbs that average more than 8 rooms per dwelling:

```
mmg8 <- cbind(apply(g8, 2, min), apply(g8, 2, mean),
                 apply(g8, 2, max))
mmg82 <- cbind(mmg8, mmg8[,3] - mmg8[,1])
colnames(mmg82) <- c("Min", "Mean", "Max", "Range")
mmg82
```

```
##               Min        Mean       Max     Range
## crim      0.02009   0.7187954   3.47428   3.45419
## zn        0.00000  13.6153846  95.00000  95.00000
## indus     2.68000   7.0784615  19.58000  16.90000
## chas      0.00000   0.1538462   1.00000   1.00000
## nox       0.41610   0.5392385   0.71800   0.30190
## rm        8.03400   8.3485385   8.78000   0.74600
## age       8.40000  71.5384615  93.90000  85.50000
## dis       1.80100   3.4301923   8.90670   7.10570
## rad       2.00000   7.4615385  24.00000  22.00000
## tax     224.00000 325.0769231 666.00000 442.00000
## ptratio  13.00000  16.3615385  20.20000   7.20000
## black   354.55000 385.2107692 396.90000  42.35000
## lstat     2.47000   4.3100000   7.44000   4.97000
## medv     21.90000  44.2000000  50.00000  28.10000
```

Among these suburbs, certain predictors vary in the same way but to a more extreme amount than those that average more than 7 rooms: on average, crime is even lower, the likelihood of bounding the Charles river is higher, distance to employment centers is even lower, the proportion of lower status people is even lower. Certain predictors vary in the same way but to a less extreme amount: average proportion of land for >25,000 acre plots, proportion of land, tax rate, the feature corresponding to the proportion of black population, and the proportion of non-retail business acres is greater than global average but not by as much. Similarly, average nitric oxide concentration, pupil-teacher ratio, and accessibility to highways is lower than global average but not by as much. The remaining variables reversed polarity: average distance to employment centers is lower instead of higher and average proportion of pre-1940 building is higher instead of lower.

3. KNearestNeighbors

```
# create data set
x_1 <- c(0,2,0,0,-1,1,-1)
x_2 <- c(3,0,1,1,0,1,2)
x_3 <- c(0,1,3,2,1,1,-1)
y <- c('Red','Red','Red','Green','Green','Red','Green')
KNN_data <- as.data.frame(cbind(x_1,x_2,x_3,y))
```

3a.

To calculate the euclidean distance between each point and $(0,0,0)$, I simply need the square root of the sum of the squared features of each individual observation:

```
obs_data <- KNN_data[c('x_1','x_2','x_3')]
vec_data <- obs_data %>% as.matrix %>% as.numeric
mdata <- vec_data %>% as.vector
dim(mdata) <- c(7,3)
colnames(mdata) <- c('x_1','x_2','x_3')
sqdist <- mdata^2 %>% rowSums
dist <- sqdist^(0.5)
distdata <- as.data.frame(cbind(1:7,dist))
colnames(distdata) <- c('Obs','dist')
distdata
```

```
##   Obs     dist
## 1   1 3.000000
## 2   2 2.236068
## 3   3 3.162278
## 4   4 2.236068
## 5   5 1.414214
## 6   6 1.732051
## 7   7 2.449490
```

3b.

The closest observation by Euclidean distance is observation 5. Observation 5 has the label 'Green', so if K = 1, the prediction would be $\textbf{Green}$.

3c.

The closest two observations are 5 and 6. Observation 5 is labeled 'Green' and 6 is labeled 'Red'; however, there is a tie for the next closest between observations 2 and 4, which have opposing labels. If the tiebreaker goes to observation 2, then there are two 'Red' labels and one 'Green' label closest to the origin, causing the prediction to be 'Red', but if observation 4 wins the tiebreaker, the prediction would be 'Green'. Consequently, if K = 3, the origin point $\textbf{exists on the decision boundary}$.

3d.

If the decision boundary is non-linear, this reflects that the algorithm must produce flexible results. Therefore, we would expect the best value of K in this case to be $\textbf{small}$, as it takes less points to sway predictions one direction or the other. Therefore, a smaller K would produce a more flexible boundary.