Homework 1

Tim

January 29, 2018

Chapter 2.4

8 (c)

```
library(ISLR)
college <- ISLR::College
```

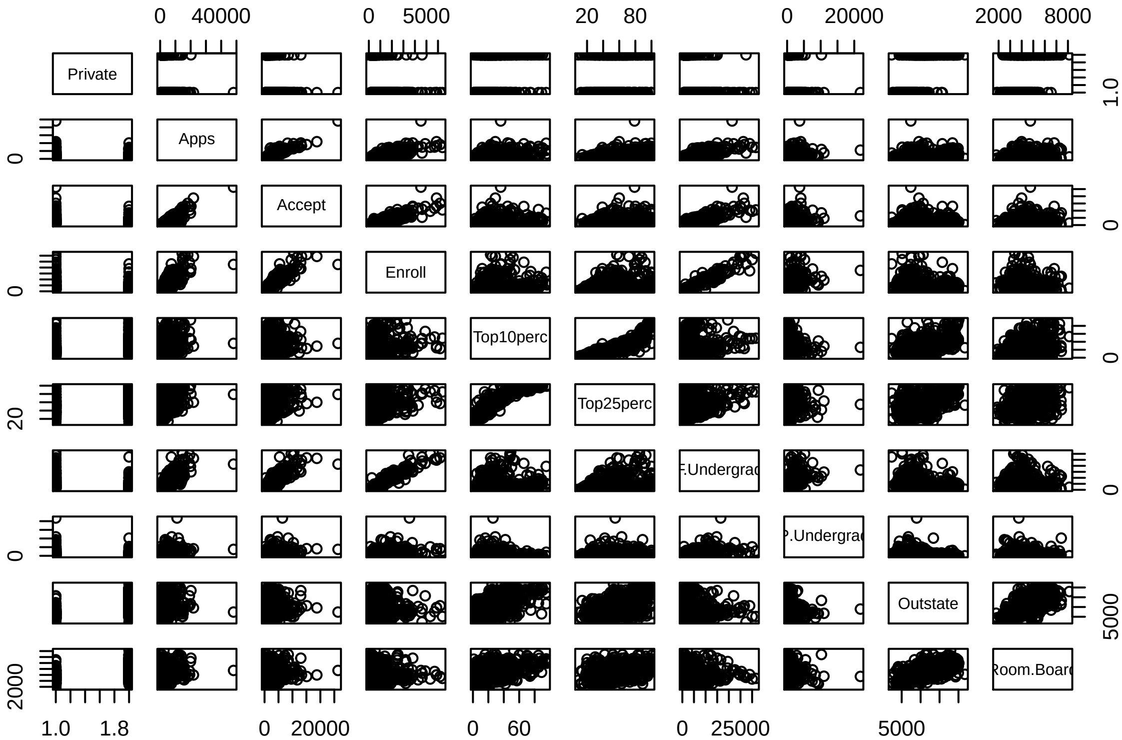
(i)

```
summary(college)
```

```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00
##            Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56
##            3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00
##            Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00
##    Top25perc      F.Undergrad     P.Undergrad         Outstate
##  Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340
##  1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320
##  Median : 54.0   Median : 1707   Median :  353.0   Median : 9990
##  Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441
##  3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925
##  Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700
##    Room.Board       Books           Personal         PhD
##  Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00
##  1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00
##  Median :4200   Median : 500.0   Median :1200   Median : 75.00
##  Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66
##  3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00
##  Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00
##     Terminal       S.F.Ratio      perc.alumni        Expend
##  Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186
##  1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751
##  Median : 82.0   Median :13.60   Median :21.00   Median : 8377
##  Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660
##  3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830
##  Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233
##    Grad.Rate
##  Min.   : 10.00
##  1st Qu.: 53.00
##  Median : 65.00
##  Mean   : 65.46
##  3rd Qu.: 78.00
##  Max.   :118.00
```

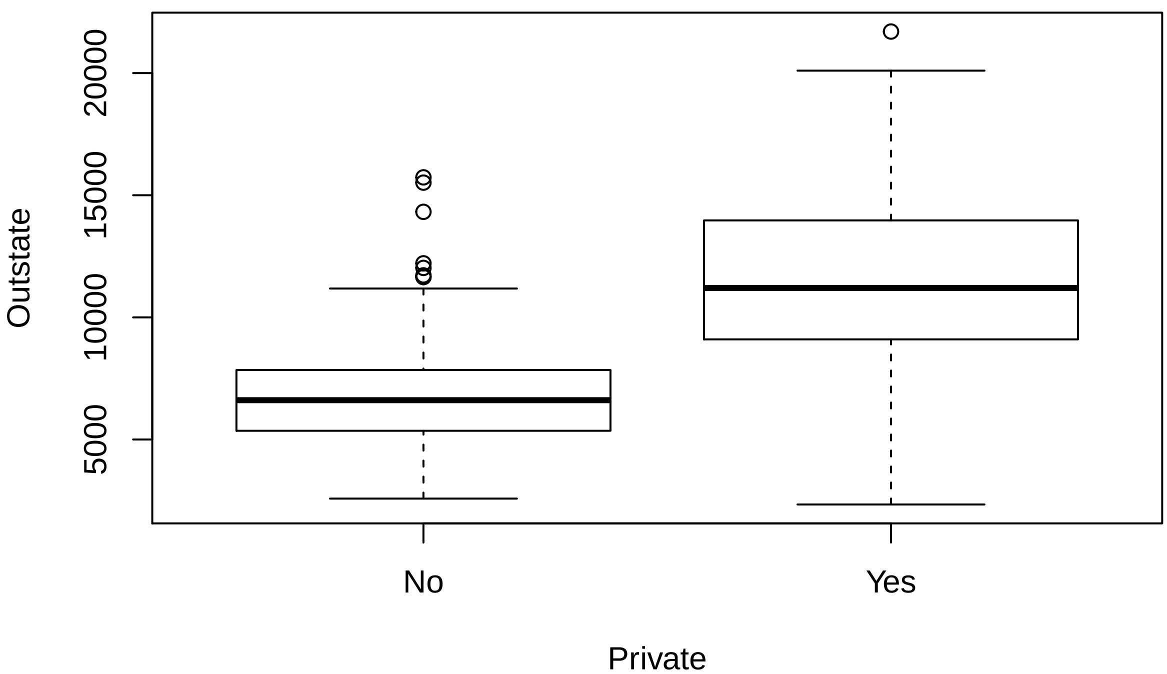
(ii)

```
pairs(college[,1:10])
```



(iii)

```
plot(Outstate ~ Private, data = college)
```



(iv)

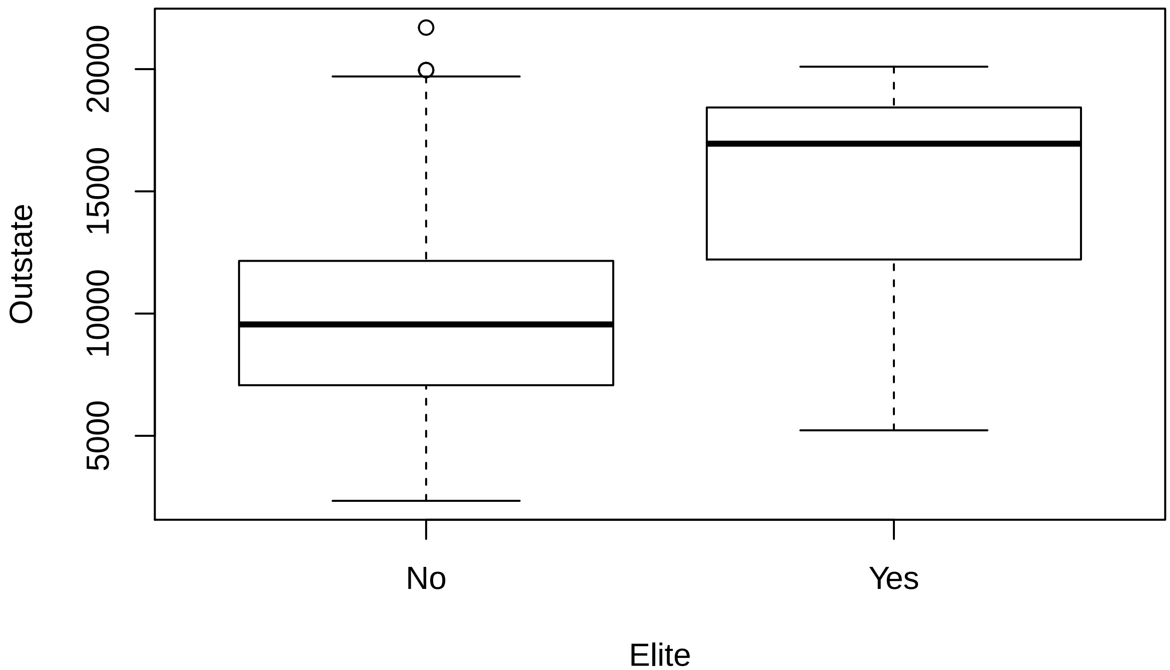
```r
Elite = rep("No", nrow(college))
Elite[college$Top10perc > 50] = "Yes"
Elite = as.factor(Elite)
college = data.frame(college, Elite)
```

```r
summary(Elite)
```

```
##  No Yes
## 699  78
```

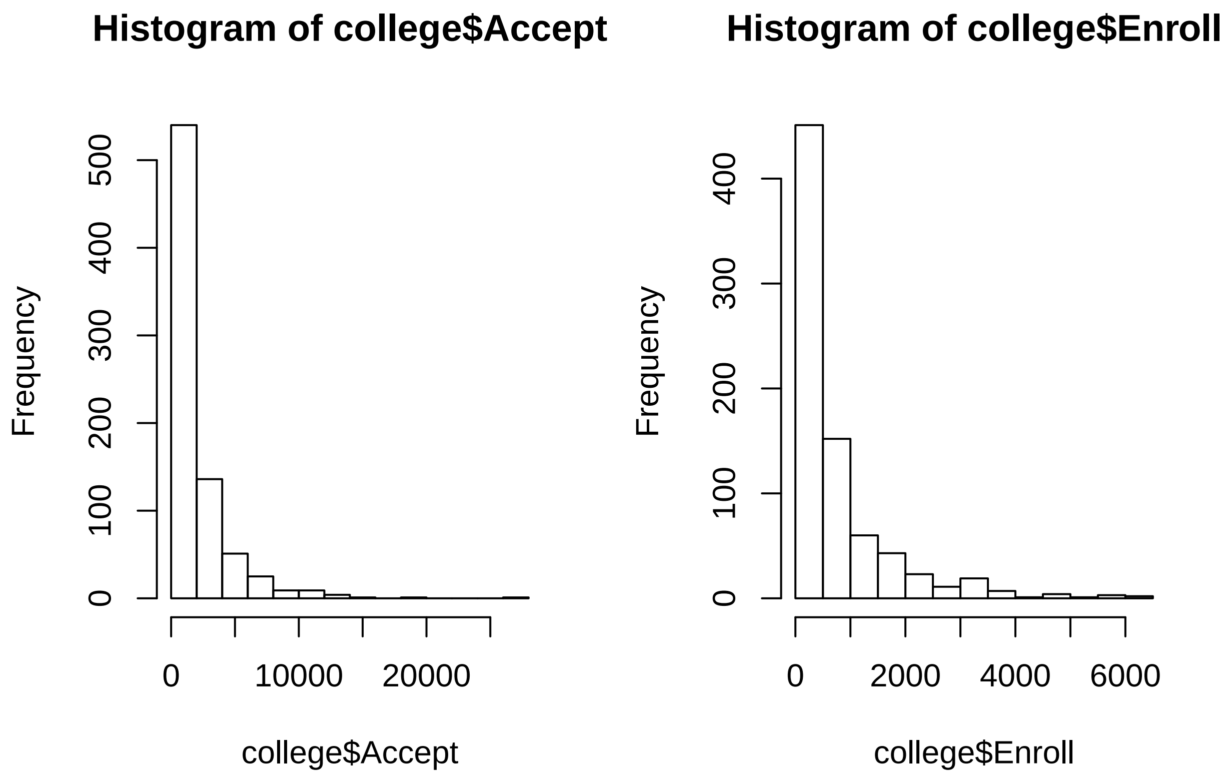
There are 78 elite universities.

```r
plot(Outstate ~ Elite, data = college)
```



(v)

```r
par(mfrow = c(1, 2))
hist(college$Accept)
hist(college$Enroll)
```

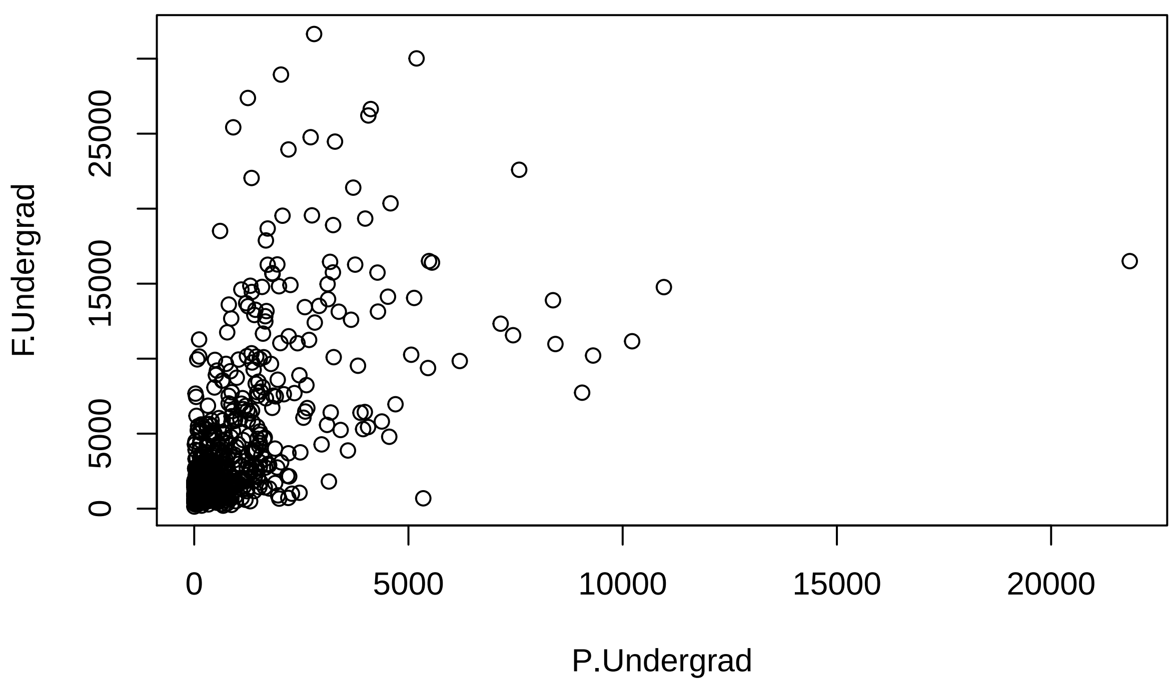


(vi)

I found that as the number of part-time undergraduates increase, the number of full-time undergraduates increases less, or not at all in some cases.

See this scatter plot, for example.

```
plot(F.Undergrad ~ P.Undergrad, data = college)
```



This matches the idea that different universities cater to differ groups of students. For example, I would guess that the proportion of part-time undergraduates at UNO is higher than the proportion at UNL.

9

```r
autos <- ISLR::Auto
```

(a)

`str` will give us a list of all the variables and tell us whether they are quantitative or qualitative:

```r
str(autos)
```

```
## 'data.frame':    392 obs. of  9 variables:
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : num  8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : num  130 165 150 150 140 198 220 215 225 190 ...
##  $ weight      : num  3504 3693 3436 3433 3449 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ year        : num  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ name        : Factor w/ 304 levels "amc ambassador brougham",..: 49 36 231 14 161 141 54 223 241 2
```

The class is misleading on the `origin` variable, however. According to the documentation: "Origin of car (1. American, 2. European, 3. Japanese)." Additionally, I would say `cylinders` is also qualitative, since it is describing the type of engine and we would treat it as a factor. Lastly, you could treat `year` as a quantitative or qualitative variable. I'll say quantitative, since we might want to see if `mpg` improves over time, for example. If we treated the variable qualitatively we would loss the ordering.

(b)

```r
quan_autos <- autos[,-c(2, 8, 9)]
sapply(quan_autos, range)
```

```
##       mpg displacement horsepower weight acceleration year
## [1,]  9.0           68         46   1613          8.0   70
## [2,] 46.6          455        230   5140         24.8   82
```

(c)

```r
sapply(quan_autos, function (x) {c(mean = mean(x), sd = sd(x))})
```

```
##            mpg displacement horsepower    weight acceleration      year
## mean 23.445918      194.412  104.46939 2977.5842    15.541327 75.979592
## sd    7.805007      104.644   38.49116  849.4026     2.758864  3.683737
```

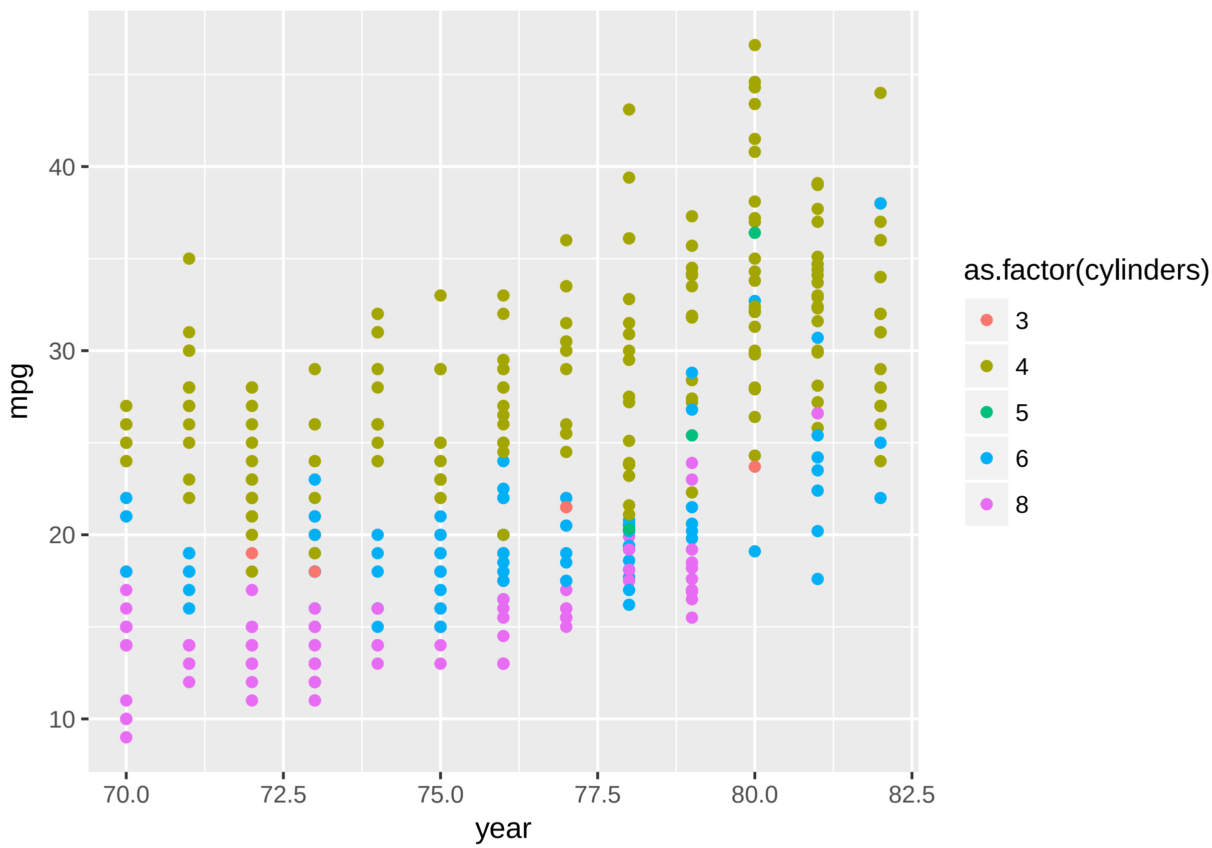
(d)

```r
quan_autos <- quan_autos[-c(10:85),]
sapply(
  quan_autos,
  function (x) {c(mean = mean(x), sd = sd(x), range = range(x))}
)
```

```
##              mpg displacement horsepower    weight acceleration      year
## mean   24.404430    187.24051  100.72152 2935.9715    15.726899 77.145570
## sd      7.867283     99.67837   35.70885  811.3002     2.693721  3.106217
## range1 11.000000     68.00000   46.00000 1649.0000     8.500000 70.000000
## range2 46.600000    455.00000  230.00000 4997.0000    24.800000 82.000000
```

(e)

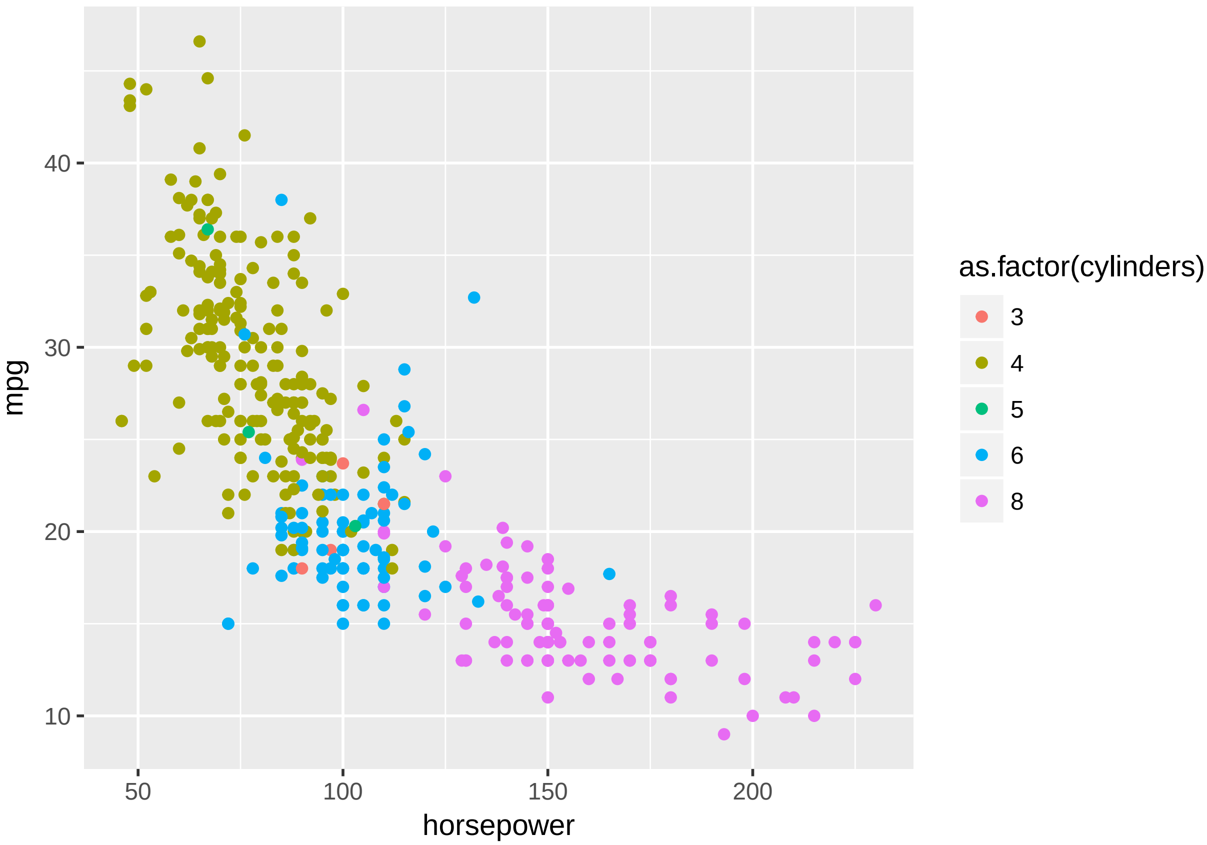
```r
library(tidyverse)
ggplot(autos) +
  geom_point(aes(x = year, y = mpg, color = as.factor(cylinders)))
```



This scatter plot shows average improvement of `mpg` over the years and highlights the different mpg possible with different cylinders. As expected, high cylinders have worse mpg than fewer cylinders, on average.

You would also expect horsepower to trade off with mpg. Let's take a look:

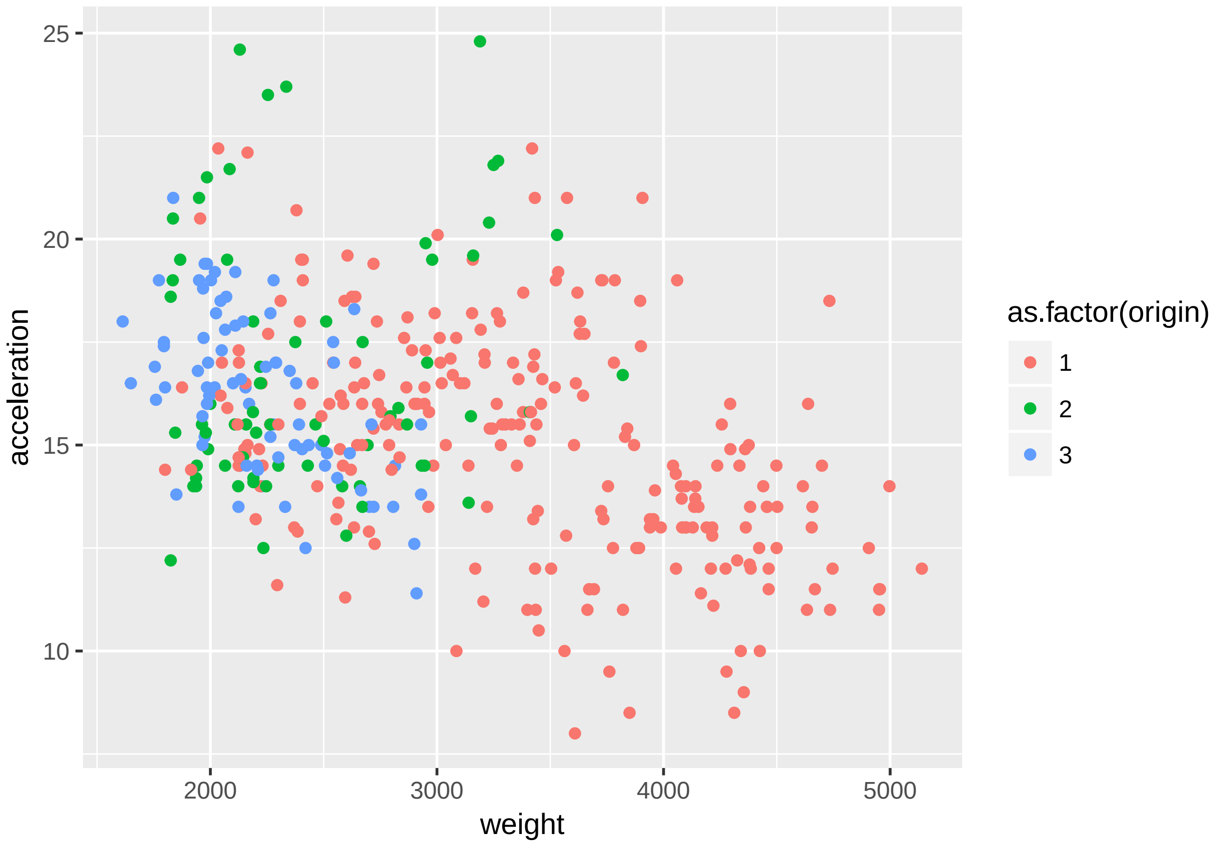
```r
ggplot(autos) +
  geom_point(aes(x = horsepower, y = mpg, color = as.factor(cylinders)))
```



This time we have a non-linear relationship, somewhat surprising to me. However, the cylinders are grouped as expected, with higher cylinders having both more horsepower and less mpg.

Mathematically, we might also expect a trade-off between weight and acceleration:

```
ggplot(autos) +
  geom_point(aes(x = weight, y = acceleration, color = as.factor(origin)))
```



This trend is much more variable, but the heavier cars do seem to have less acceleration. Moreover, almost all the cars weighing over 3000 lbs. are American made (`origin == 1`). The US has larger roads and we do more driving than any other country, so this is not surprising. Likewise, this fits into the American stereotype of large trucks and SUVs.

(f)

Both year and horsepower would do well in predicting mpg. However, horsepower appears to show a much more well-defined non-linear relationship than year, which might be most useful for prediction.

Chapter 3.7

8

(a)

```
m <- lm(mpg ~ horsepower, data = autos)
summary(m)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower, data = autos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

As we saw from the scatter plot, there clearly is a relationship between mpg and horsepower: as horsepower goes down, mpg goes down. This is confirmed by the negative slope coefficient and by the low p-values. However, visually we know the relationship is not linear, which is reflected in the R-squared value of 0.6059.

```r
predict(
  m, newdata = data.frame(horsepower = 98),
  interval = 'confidence',
  level = 0.95
)
```

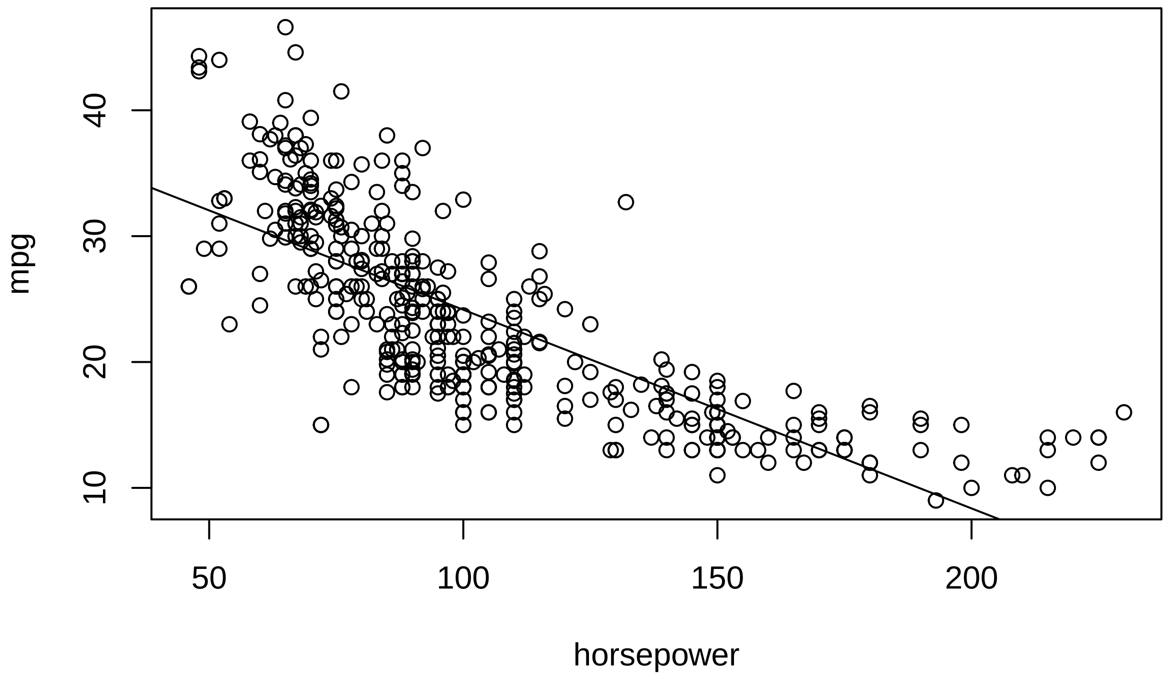
```
##        fit      lwr      upr
## 1 24.46708 23.97308 24.96108
```

```r
predict(
  m, newdata = data.frame(horsepower = 98),
  interval = 'prediction',
  level = 0.95
)
```

```
##        fit     lwr      upr
## 1 24.46708 14.8094 34.12476
```

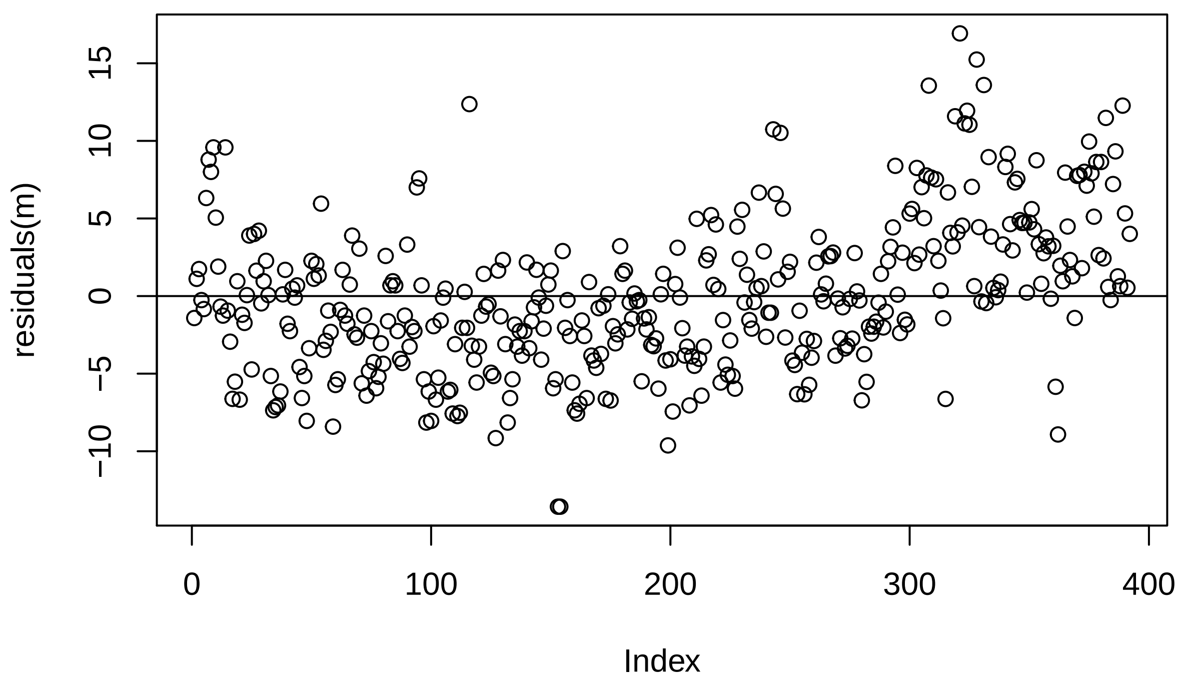
(b)

```r
plot(mpg ~ horsepower, data = autos)
abline(a = coef(m)[1], b = coef(m)[2])
```



(c)

```
plot(residuals(m))
abline(h = 0)
```



The residuals are not normally distributed around zero. The residuals are too negative around low values of horsepower and too positive for large horsepower. That is consistent with the quadratic shape we see in the scatter plot.

Graduate Problem

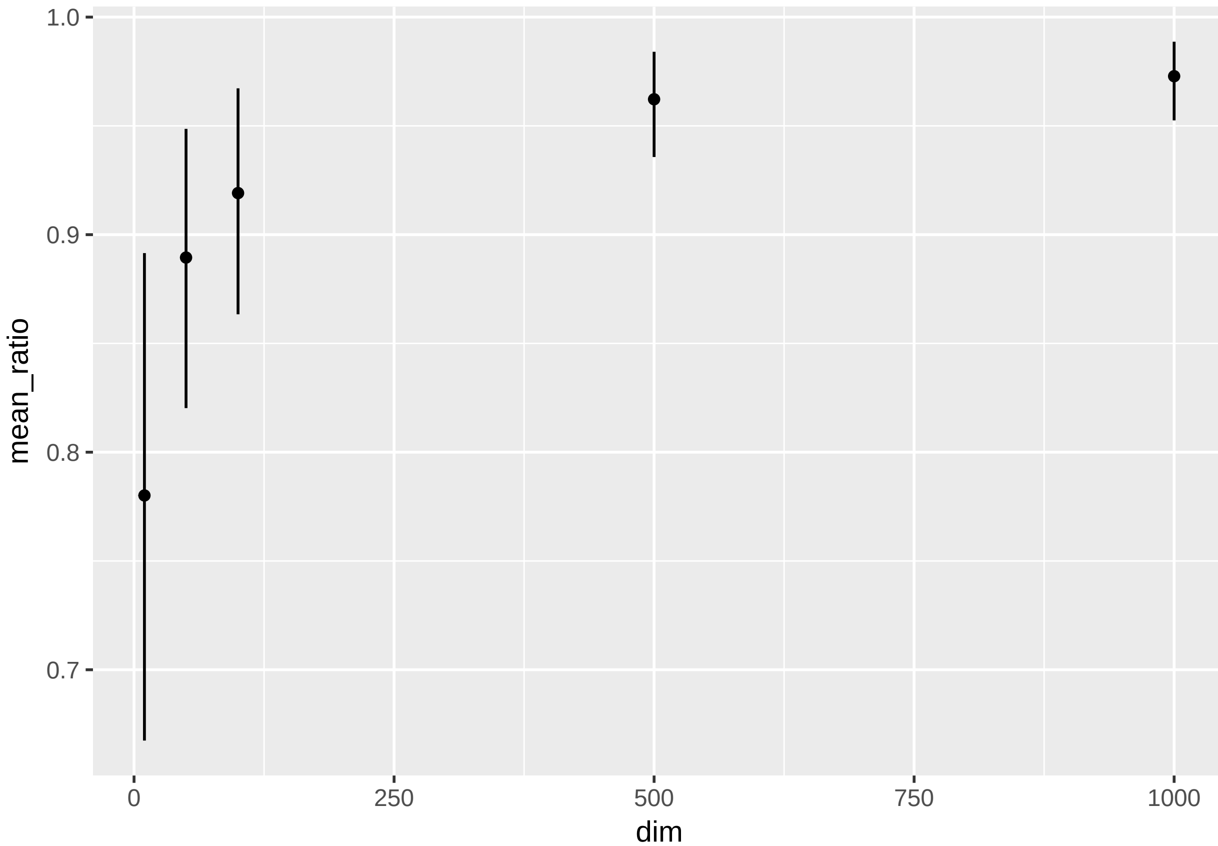
```
point_dist <- function(dims) {
  sapply(
    dims,
    function(dim) { return(sqrt(sum((runif(dim) - runif(dim))^2))) }
  )
}
```

```
d <- tibble(
  dim = c(rep(10, 1e4), rep(50, 1e4), rep(100, 1e4), rep(500, 1e4), rep(1000, 1e4))
)

set.seed(123456)
d <- d %>%
  group_by(dim) %>%
  mutate(distance = point_dist(dim)) %>%
  mutate(dim_sample = sample(1e3, 1e4, replace = TRUE)) %>%
  group_by(dim, dim_sample) %>%
  summarise(ratio = mean(distance)/max(distance)) %>%
  summarise(
    mean_ratio = mean(ratio),
    low_ratio = quantile(ratio, probs = 0.025)[[1]],
    high_ratio = quantile(ratio, probs = 0.975)[[1]]
  )
d
```

```
## # A tibble: 5 x 4
##      dim mean_ratio low_ratio high_ratio
##    <dbl>      <dbl>     <dbl>      <dbl>
## 1   10.0      0.780     0.667      0.892
## 2   50.0      0.889     0.820      0.949
## 3  100        0.919     0.863      0.967
## 4  500        0.962     0.936      0.984
## 5 1000        0.973     0.953      0.989
```

```
d %>%
  ggplot(aes(x = dim, y = mean_ratio)) +
    geom_point() +
    geom_linerange(aes(ymin = low_ratio, ymax = high_ratio))
```


1.0
0.9
0.8
0.7
mean_ratio
0
250
500
750
1000
dim