hw1-Dian-Yu

Dian Yu

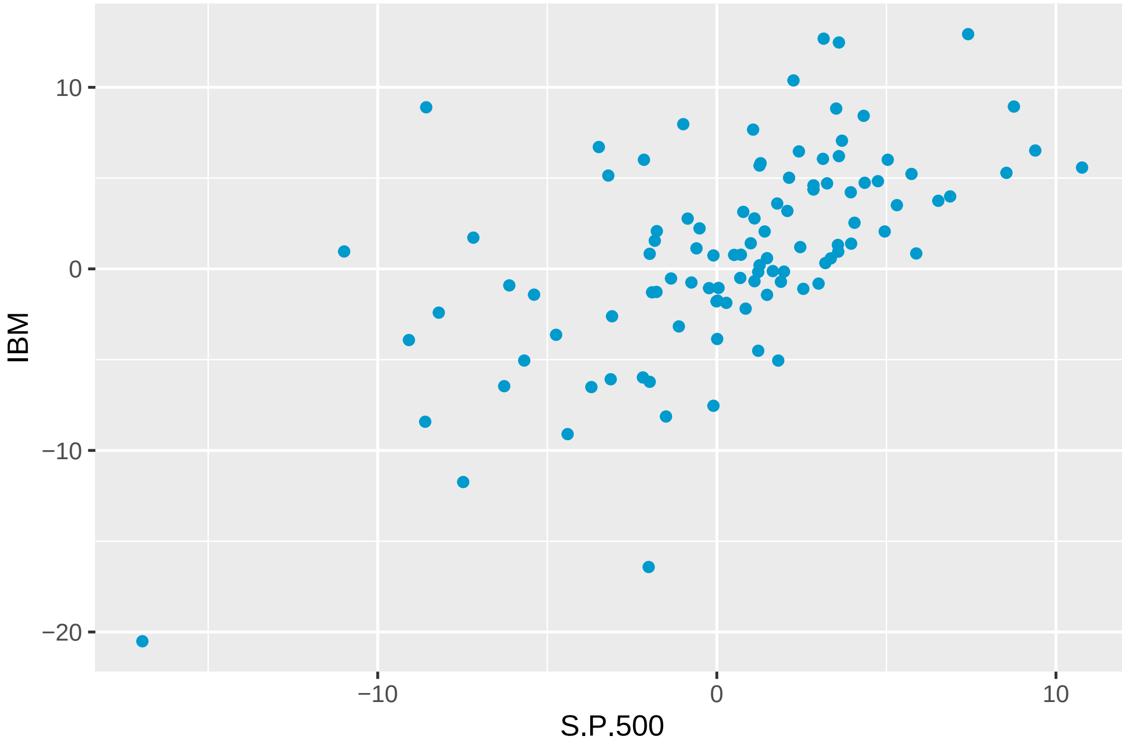
9/26/2020

```
library(tidyverse)
```

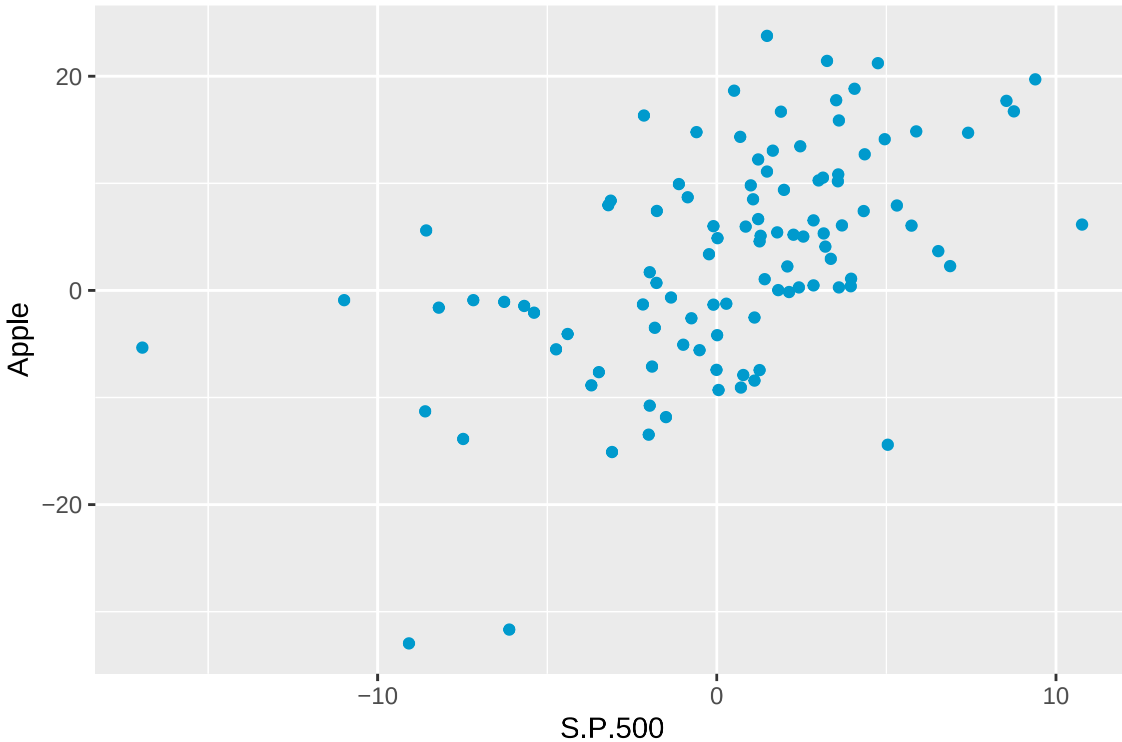
2.9 Beta coefficients of stocks

a)

```
stocks <- read.csv("IBM-Apple-SP500 RR Data.csv", skip = 1, stringsAsFactors = FALSE)


#' Convert a string to a float by removing the percentage sign at the end of the string.
#'
#' @param str A string with a percentage sign at the end.
#'
#' @return A number converted from the string with the percentage sign removed.
#'
rm_percentage_sign <- function(str) {
    num <- as.numeric(substr(str, 1, nchar(str) - 1))
    return(num)
}


stocks <- stocks %>%
    select(1:4) %>%
    mutate(S.P.500 = rm_percentage_sign(S.P.500),
           IBM = rm_percentage_sign(IBM),
           Apple = rm_percentage_sign(Apple))

ggplot(data = stocks, mapping = aes(x = S.P.500, y = IBM)) +
    geom_point(color = "deepskyblue3")
```




```
ggplot(data = stocks, mapping = aes(x = S.P.500, y = Apple)) +
    geom_point(color = "deepskyblue3")
```



The individual stocks appear to be linearly correlated with the S&P 500 index.

b)

```r
ibm_fit = lm(IBM~S.P.500, data = stocks)
summary(ibm_fit)
```

```
##
## Call:
## lm(formula = IBM ~ S.P.500, data = stocks)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5646  -2.4261  -0.6636   2.2188  14.6414
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.64164    0.44136   1.454    0.149
## S.P.500      0.74481    0.09898   7.525 2.15e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.478 on 102 degrees of freedom
## Multiple R-squared:  0.357,  Adjusted R-squared:  0.3507
## F-statistic: 56.63 on 1 and 102 DF,  p-value: 2.15e-11
```

```r
apple_fit = lm(Apple~S.P.500, data = stocks)
summary(apple_fit)
```

```
##
## Call:
## lm(formula = Apple ~ S.P.500, data = stocks)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.5378  -5.9191   0.4677   5.5363  19.4413
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.4863     0.8606   2.889  0.00472 **
## S.P.500       1.2449     0.1930   6.450   3.8e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.732 on 102 degrees of freedom
## Multiple R-squared:  0.2897, Adjusted R-squared:  0.2827
## F-statistic:  41.6 on 1 and 102 DF,  p-value: 3.799e-09
```

The beta for IBM with reference to S&P 500 is 0.74, and the beta for Apple is 1.24. Thus, Apple had a higher expected return relative to S&P 500. When S&P 500 had a return of 1%, the Apple stock had an expected return of 1.24%.

c)

```r
S.P.500_sd = sd(stocks$S.P.500)
ibm_sd = sd(stocks$IBM)
apple_sd = sd(stocks$Apple)
```

```r
cor(stocks[,2:4])
```

```
##           S.P.500       IBM     Apple
## S.P.500 1.0000000 0.5974779 0.5382317
## IBM     0.5974779 1.0000000 0.4147253
## Apple   0.5382317 0.4147253 1.0000000
```

```r
ibm_beta = cor(stocks$IBM, stocks$S.P.500) * (ibm_sd / S.P.500_sd)
ibm_beta
```

```
## [1] 0.7448088
```

```r
apple_beta = cor(stocks$Apple, stocks$S.P.500) * (apple_sd / S.P.500_sd)
apple_beta
```

```
## [1] 1.244856
```

d)

Given the same level of correlation r, the higher is the sample SD (i.e. volatility) of the individual stock, the higher is the beta coefficient. In fact, the sample SD or volatility of a stock is usually a great way to represent the risk. Finance 101 tells us that higher risk is often associated with higher expected return.

2.10 Price elasticities of steaks

a)

```r
steak <- read.csv("steakprices.csv", stringsAsFactors = FALSE)


#' Convert a string to a float by removing the dollar sign in front of the string.
#'
#' @param str A string with a dollar sign in front.
#'
#' @return A number converted from the string with the dollar sign removed.
#'
rm_dollar_sign <- function(str) {
    num <- as.numeric(substr(str, 2, nchar(str)))
    return(num)
}


steak <- steak %>%
    # Remove dollar signs in the price columns
    mutate(Chuck.Price = rm_dollar_sign(Chuck.Price),
           PortHse.Price = rm_dollar_sign(PortHse.Price),
           RibEye.Price = rm_dollar_sign(RibEye.Price)) %>%
    # Log-transformations
    mutate(log_chuck_qty = log(Chuck.Qty),
           log_chuck_price = log(Chuck.Price),
           log_porterhouse_qty = log(PortHse.Qty),
           log_porterhouse_price = log(PortHse.Price),
           log_ribeye_qty = log(RibEye.Qty),
           log_ribeye_price = log(RibEye.Price))
```

```r
chuck_fit = lm(log_chuck_qty~log_chuck_price, data = steak)
summary(chuck_fit)
```

```
##
## Call:
## lm(formula = log_chuck_qty ~ log_chuck_price, data = steak)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32463 -0.12036 -0.01714  0.09430  0.49725
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.8899     0.2871  20.513  < 2e-16 ***
## log_chuck_price  -1.3687     0.3199  -4.278 9.44e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1812 on 46 degrees of freedom
## Multiple R-squared:  0.2846, Adjusted R-squared:  0.2691
## F-statistic:  18.3 on 1 and 46 DF,  p-value: 9.441e-05
```

```r
ribeye_fit = lm(log_ribeye_qty~log_ribeye_price, data = steak)
summary(ribeye_fit)
```

```
##
## Call:
## lm(formula = log_ribeye_qty ~ log_ribeye_price, data = steak)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.54075 -0.21801  0.03995  0.20328  0.70950
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.6627     0.7537  10.167 2.39e-13 ***
## log_ribeye_price  -1.4460     0.3731  -3.876 0.000335 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2943 on 46 degrees of freedom
## Multiple R-squared:  0.2462, Adjusted R-squared:  0.2298
## F-statistic: 15.02 on 1 and 46 DF,  p-value: 0.0003352
```

```r
porterhouse_fit = lm(log_porterhouse_qty~log_porterhouse_price, data = steak)
summary(porterhouse_fit)
```

```
##
## Call:
## lm(formula = log_porterhouse_qty ~ log_porterhouse_price, data = steak)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.57655 -0.23544  0.00317  0.23511  0.49991
##
```

```
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)              9.1123     0.5136  17.742  < 2e-16 ***
## log_porterhouse_price  -2.6565     0.2752  -9.654 1.23e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.283 on 46 degrees of freedom
## Multiple R-squared:  0.6695, Adjusted R-squared:  0.6624
## F-statistic:  93.2 on 1 and 46 DF,  p-value: 1.233e-12
```

The price elasticities of chuck, rib eye, and porter house are respectively -1.37, -1.45, and -2.66. Thus, the order of the elasticities is indeed the same as the order of expensiveness.

b)

```r
summary(chuck_fit)$coefficients[2, 1] * 10
```

```
## [1] -13.68665
```

```r
summary(ribeye_fit)$coefficients[2, 1] * 10
```

```
## [1] -14.46004
```

```r
summary(porterhouse_fit)$coefficients[2, 1] * 10
```

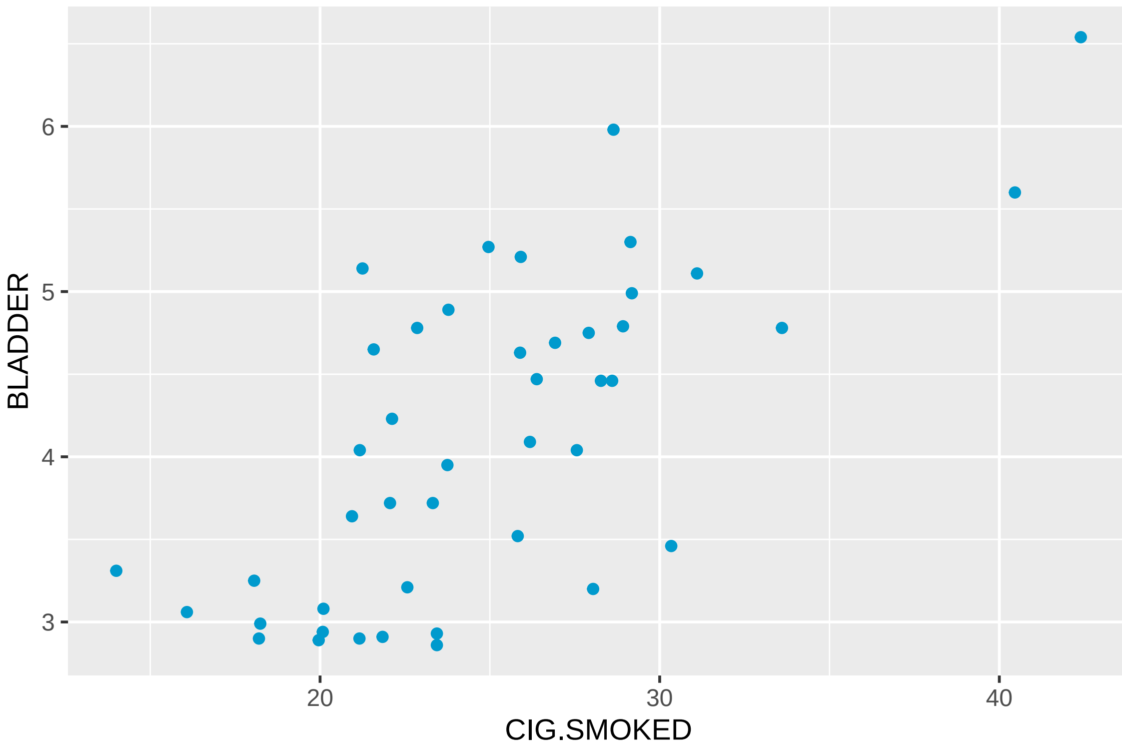
```
## [1] -26.56487
```

The estimated demand change for chuck, rib eye, and porter house are respectively -13.69%, -14.46%, and -26.56%.

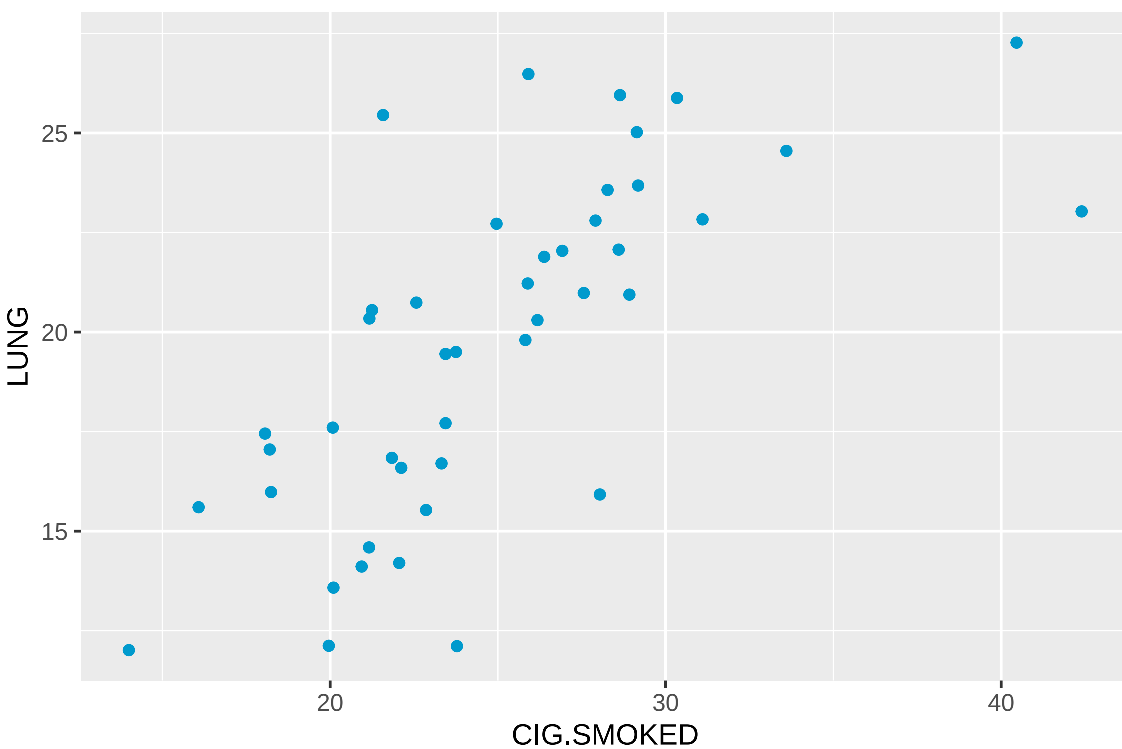
2.11 Smoking versus cancer

a)

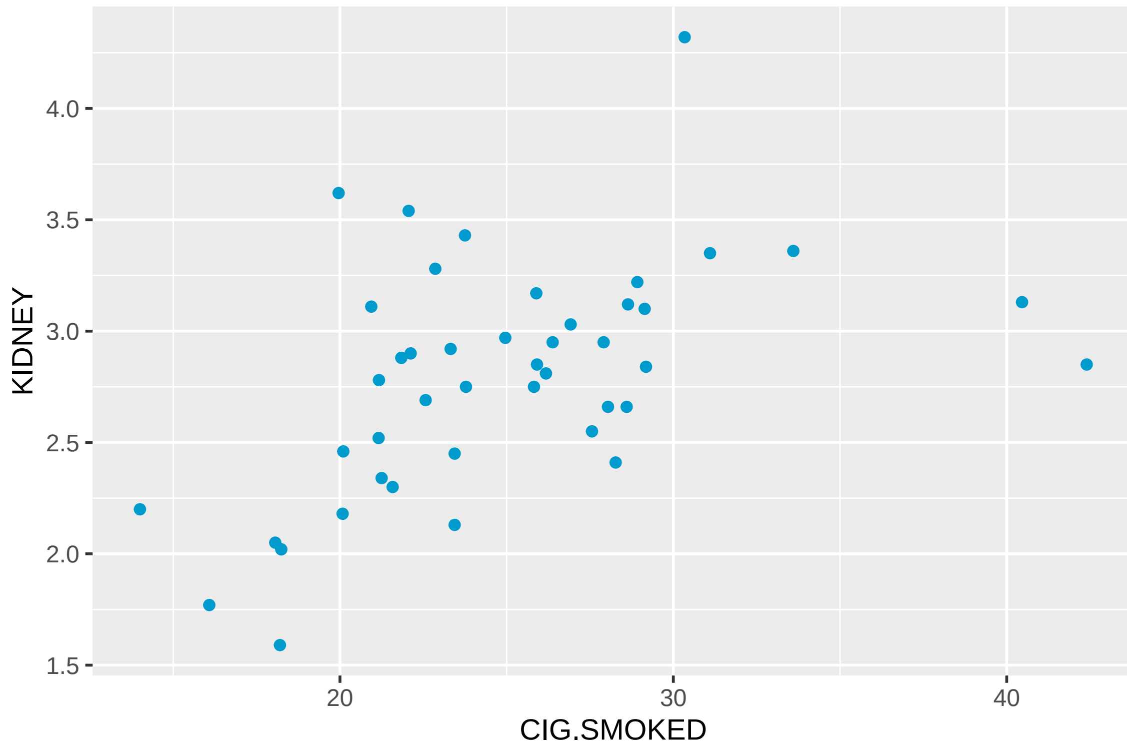
```r
smoking_cancer <- read.csv("Cancer Data.csv", stringsAsFactors = FALSE)

ggplot(data = smoking_cancer, mapping = aes(x = CIG.SMOKED, y = BLADDER)) +
    geom_point(color = "deepskyblue3")
```



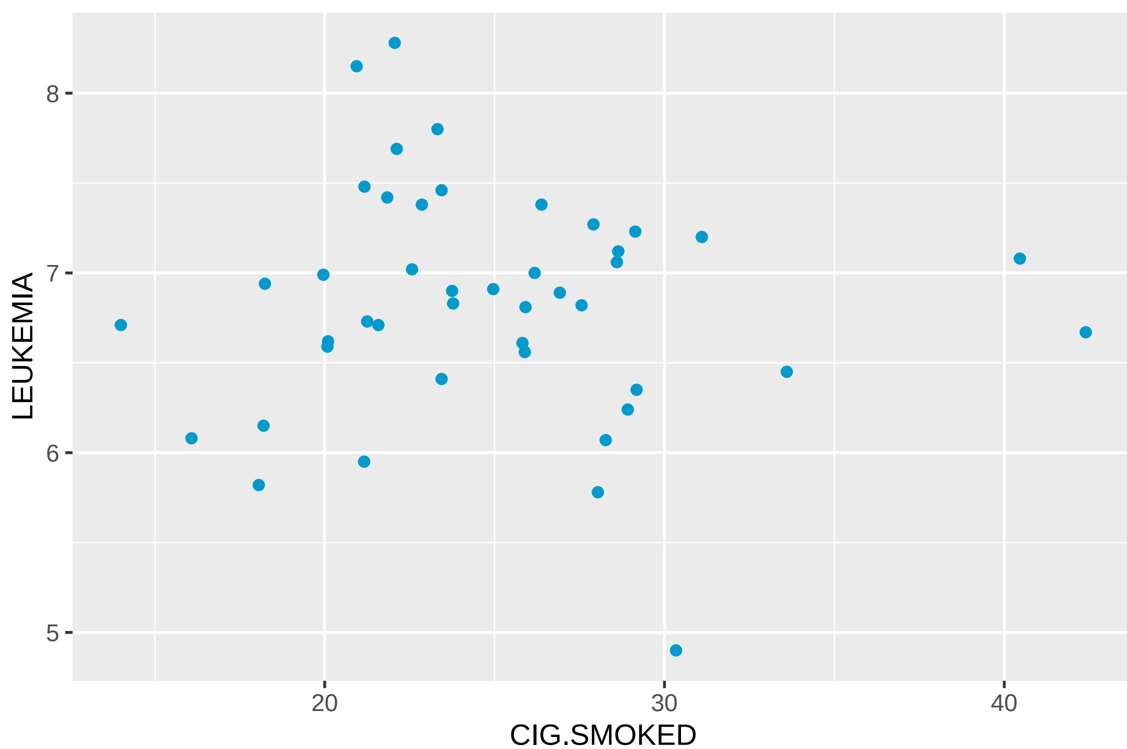

```r
ggplot(data = smoking_cancer, mapping = aes(x = CIG.SMOKED, y = LUNG)) +
    geom_point(color = "deepskyblue3")
```




```r
ggplot(data = smoking_cancer, mapping = aes(x = CIG.SMOKED, y = KIDNEY)) +
    geom_point(color = "deepskyblue3")
```




```
ggplot(data = smoking_cancer, mapping = aes(x = CIG.SMOKED, y = LEUKEMIA)) +
    geom_point(color = "deepskyblue3")
```




```
smoking_cancer[smoking_cancer$CIG.SMOKED > 40,]
```

```
##    STATE CIG.SMOKED BLADDER  LUNG KIDNEY LEUKEMIA
## 8     DC      40.46    5.60 27.27   3.13     7.08
```

```
## 26    NE      42.40    6.54 23.03   2.85     6.67
```

DC and NE has especially high number of cigarettes smoked. When these two observations are removed, the number of deaths due to bladder, lung, and kidney cancer all appear to be linearly correlated with cigarettes smoked. However, the correlation between the number of deaths due to leukemia and cigarettes smoked seems weak.

b)

```r
# Notice that to test the correlation coefficient = 0 is the same as testing the regression coefficient
bladder_fit = lm(BLADDER~CIG.SMOKED, data = smoking_cancer)
summary(bladder_fit)
```

```
##
## Call:
## lm(formula = BLADDER ~ CIG.SMOKED, data = smoking_cancer)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.32213 -0.42488 -0.03275  0.37872  1.46523
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.08608    0.48437   2.242   0.0303 *
## CIG.SMOKED   0.12182    0.01898   6.417 9.96e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6938 on 42 degrees of freedom
## Multiple R-squared:  0.4951, Adjusted R-squared:  0.4831
## F-statistic: 41.18 on 1 and 42 DF,  p-value: 9.964e-08
```

```r
lung_fit = lm(LUNG~CIG.SMOKED, data = smoking_cancer)
summary(lung_fit)
```

```
##
## Call:
## lm(formula = LUNG ~ CIG.SMOKED, data = smoking_cancer)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.943 -1.656  0.382  1.614  7.561
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.4717     2.1407   3.023  0.00425 **
## CIG.SMOKED    0.5291     0.0839   6.306 1.44e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.066 on 42 degrees of freedom
## Multiple R-squared:  0.4864, Adjusted R-squared:  0.4741
## F-statistic: 39.77 on 1 and 42 DF,  p-value: 1.439e-07
```

```r
kidney_fit = lm(KIDNEY~CIG.SMOKED, data = smoking_cancer)
summary(kidney_fit)
```

```
##
## Call:
## lm(formula = KIDNEY ~ CIG.SMOKED, data = smoking_cancer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8998 -0.3122  0.0044  0.2046  1.2792
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.66359    0.32020   5.196 5.63e-06 ***
## CIG.SMOKED   0.04539    0.01255   3.617 0.000792 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4586 on 42 degrees of freedom
## Multiple R-squared:  0.2375, Adjusted R-squared:  0.2194
## F-statistic: 13.09 on 1 and 42 DF,  p-value: 0.0007922
```

```
leukemia_fit = lm(LEUKEMIA~CIG.SMOKED, data = smoking_cancer)
summary(leukemia_fit)
```

```
##
## Call:
## lm(formula = LEUKEMIA ~ CIG.SMOKED, data = smoking_cancer)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.88722 -0.28618  0.03443  0.42240  1.42784
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.025163   0.449835  15.617   <2e-16 ***
## CIG.SMOKED  -0.007843   0.017630  -0.445    0.659
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6443 on 42 degrees of freedom
## Multiple R-squared:  0.00469,    Adjusted R-squared:  -0.01901
## F-statistic: 0.1979 on 1 and 42 DF,  p-value: 0.6587
```

```
cor(smoking_cancer[2:6])[1, 2:5]
```

```
##     BLADDER        LUNG      KIDNEY    LEUKEMIA
##  0.70362186  0.69740250  0.48738962 -0.06848123
```

The correlation between bladder, lung, and kidney cancer deaths and cigarette smoking are all significantly different from 0, but the correlation between leukemia and cigarette smoking is not statistically significant. Among bladder, lung, and kidney cancers, bladder cancer is most significantly correlated with cigarette smoking, with a t value of 6.417 and a correlation coefficient of 0.704.