IS621_hw1

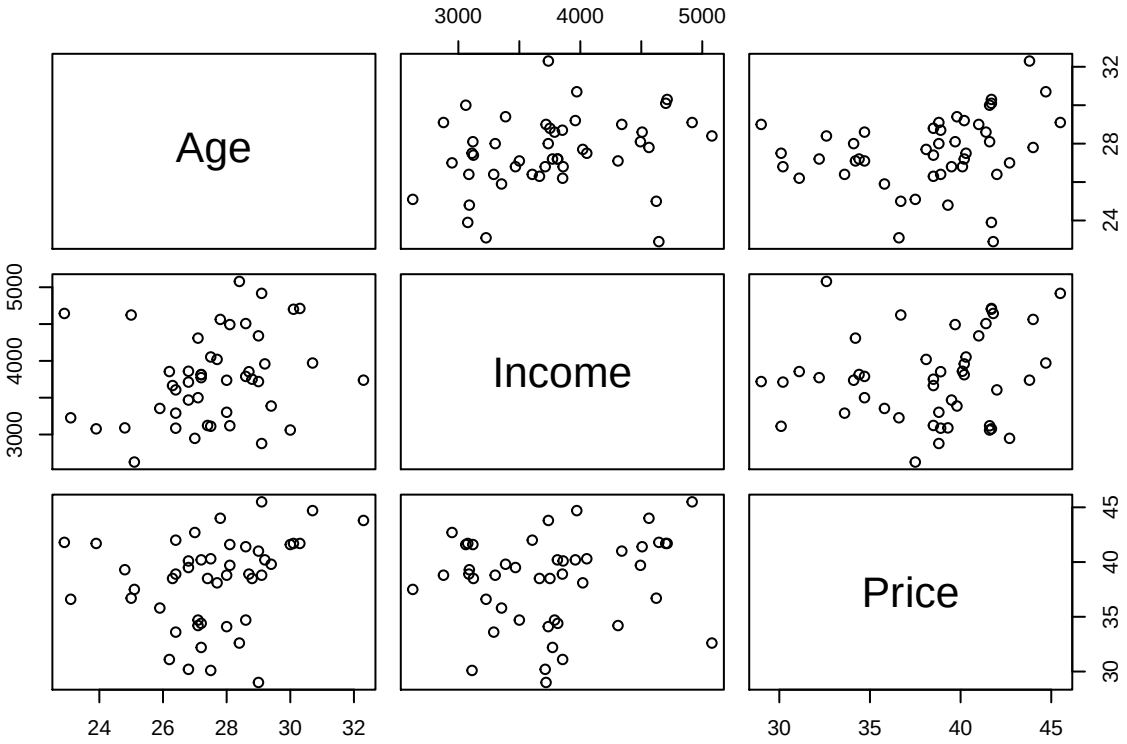
Charley Ferrari

February 3, 2016

```
library(ggplot2)

library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
setwd("/Users/Charley/Downloads/cuny/IS 621 Business Analytics and Data Mining/Homework 1")

cigarettes <- read.csv("cigarette-training-data.csv")

pairs(select(cigarettes, -c(State, Sales)))
```



```r
cor(select(cigarettes, -c(State, Sales)))
```

```
##              Age    Income     Price
## Age    1.0000000 0.2489590 0.2375821
## Income 0.2489590 1.0000000 0.1465515
## Price  0.2375821 0.1465515 1.0000000
```

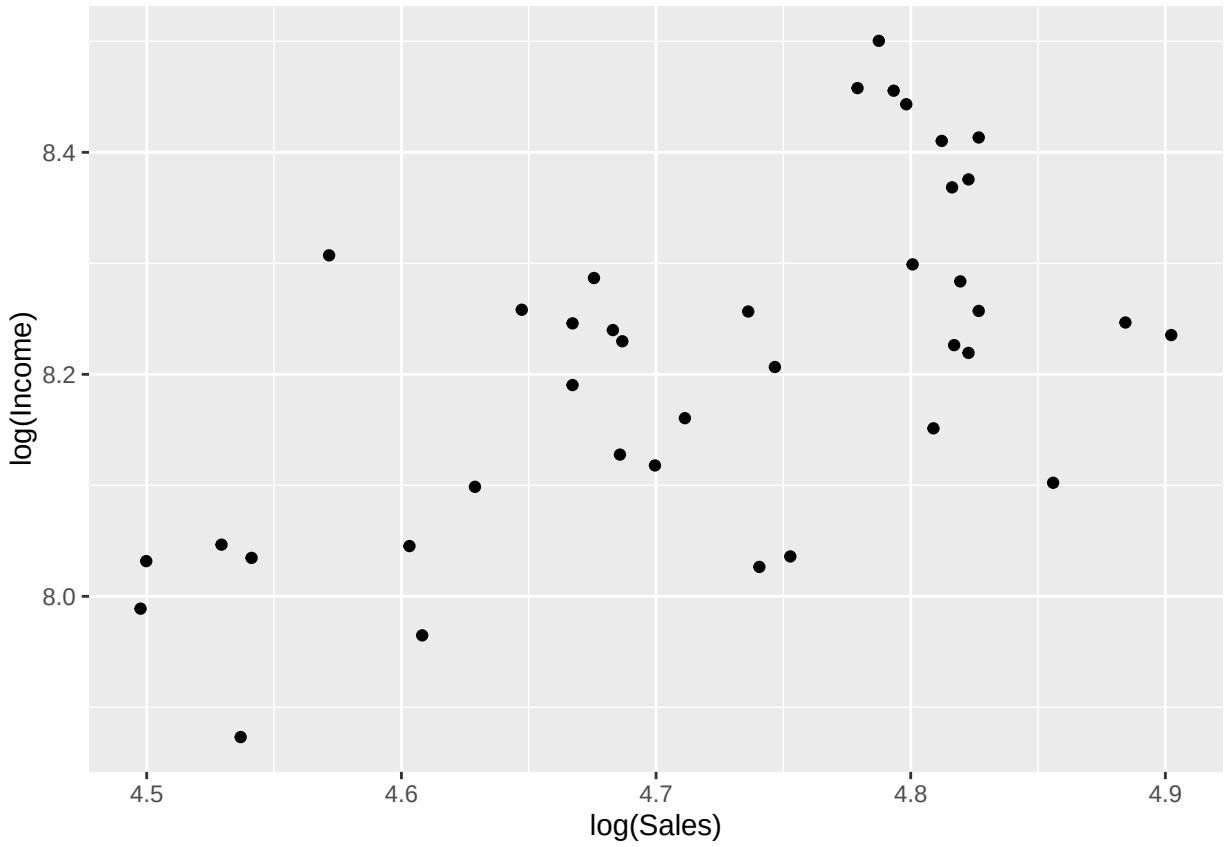
```r
summary(lm(Sales ~ Price + Income + Age, data = cigarettes))
```

```
##
## Call:
## lm(formula = Sales ~ Price + Income + Age, data = cigarettes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.329 -13.420  -4.093   3.348 130.767
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.594723  66.672107   0.699  0.48849
## Price       -3.198441   1.061814  -3.012  0.00438 **
## Income       0.017851   0.007263   2.458  0.01819 *
## Age          4.667748   2.327559   2.005  0.05139 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.08 on 42 degrees of freedom
## Multiple R-squared:  0.2989, Adjusted R-squared:  0.2488
## F-statistic: 5.969 on 3 and 42 DF,  p-value: 0.001744
```

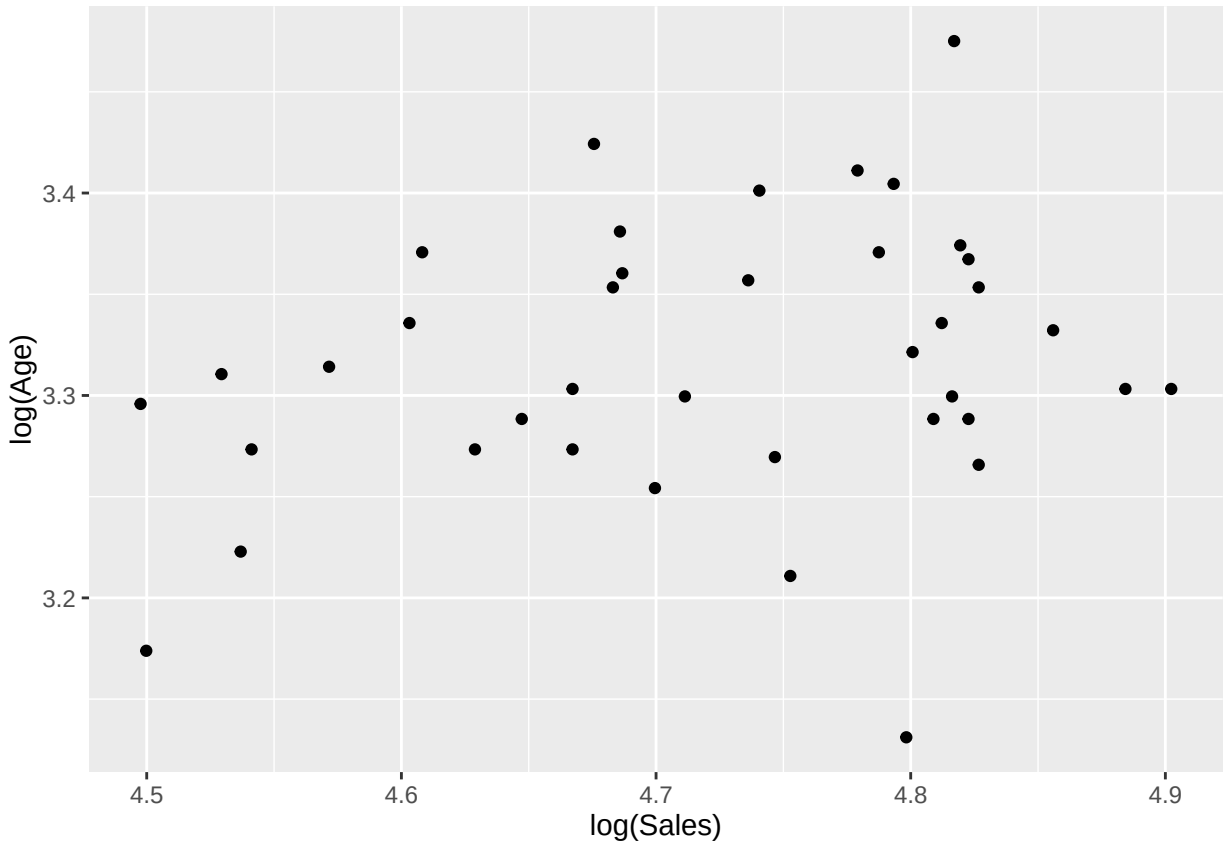
```r
summary(lm(Sales ~ Price + Income, data = cigarettes))
```

```
##
## Call:
## lm(formula = Sales ~ Price + Income, data = cigarettes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.279 -11.257  -4.287   2.147 134.859
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 145.828361  46.227232   3.155  0.00293 **
## Price        -2.751481   1.074018  -2.562  0.01400 *
## Income        0.021097   0.007325   2.880  0.00617 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.05 on 43 degrees of freedom
## Multiple R-squared:  0.2318, Adjusted R-squared:  0.1961
## F-statistic: 6.487 on 2 and 43 DF,  p-value: 0.00345
```

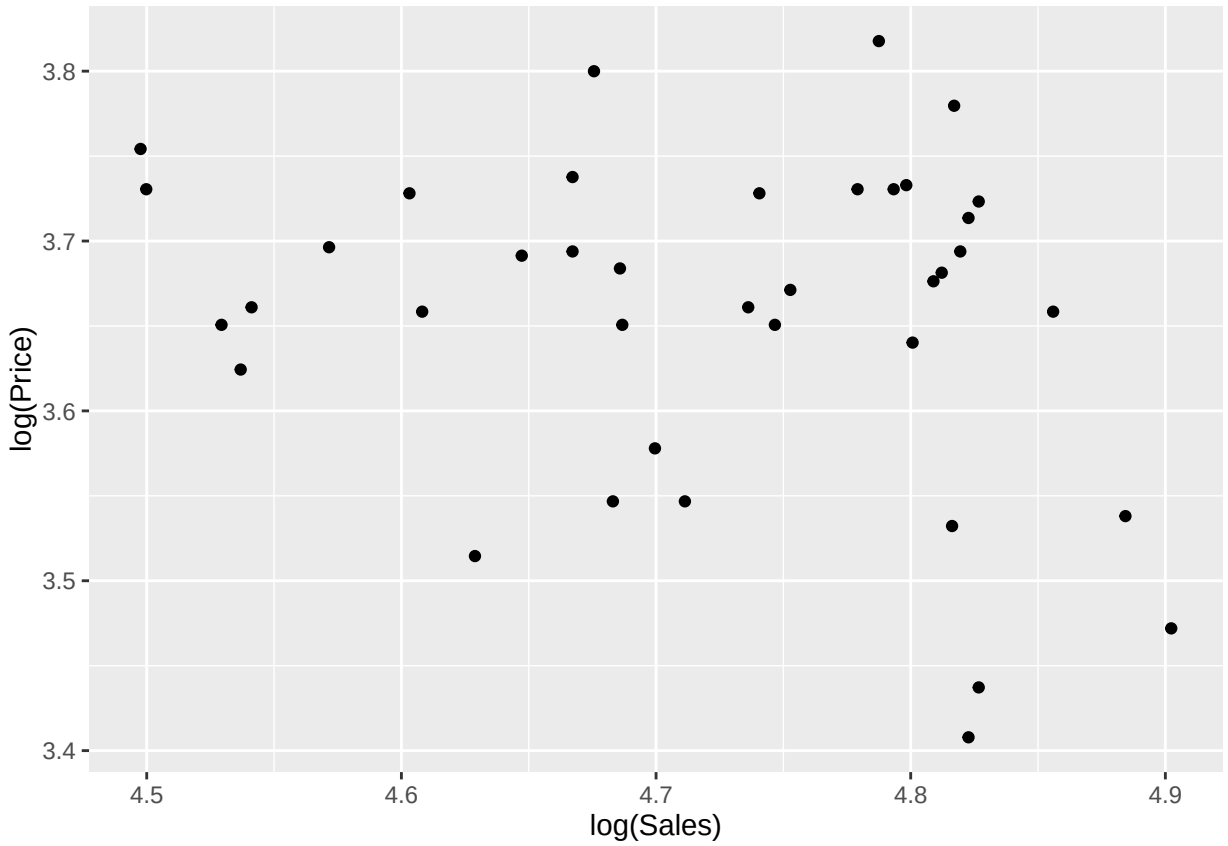
```
test <- cigarettes

test <- filter(cigarettes, Sales < 150 & Sales > 85)

ggplot(test, aes(x=log(Sales), y=log(Income))) + geom_point()
```



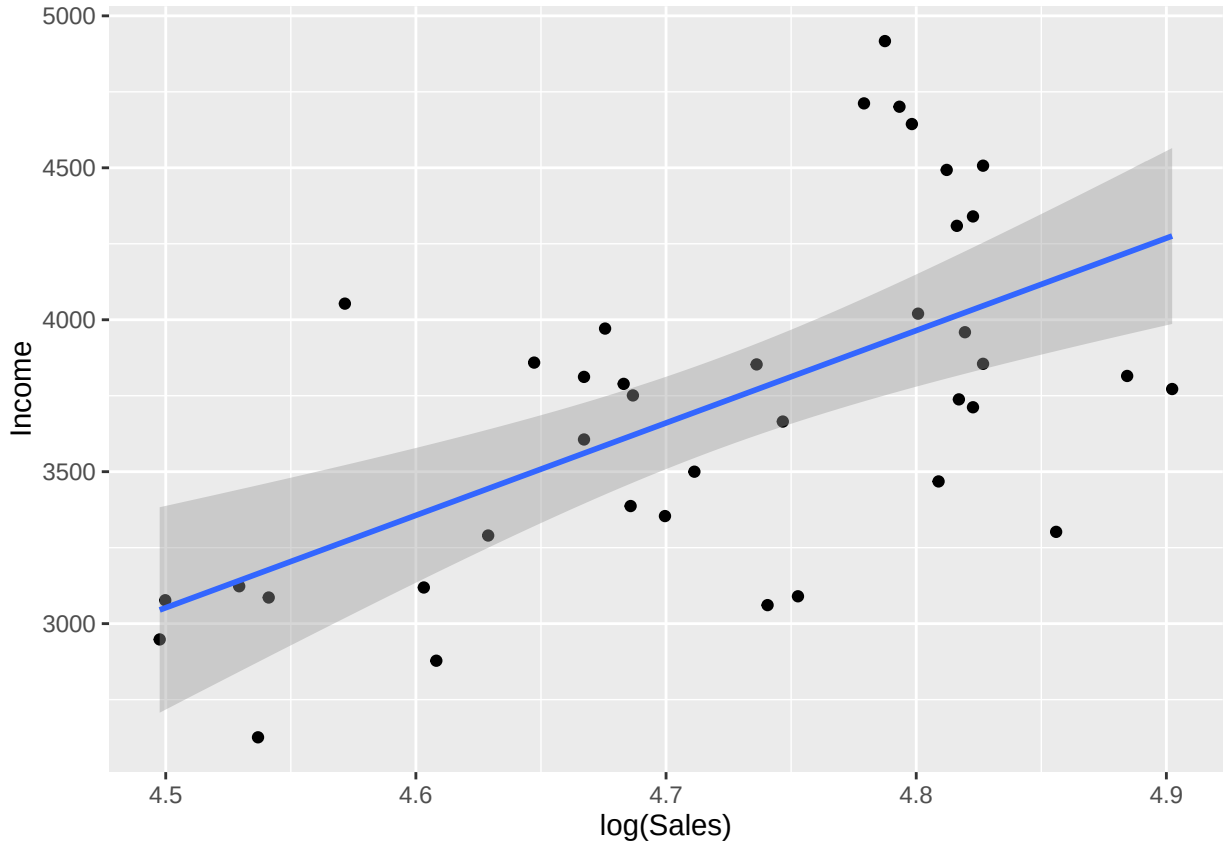

```
ggplot(test, aes(x=log(Sales), y=log(Age))) + geom_point()
```



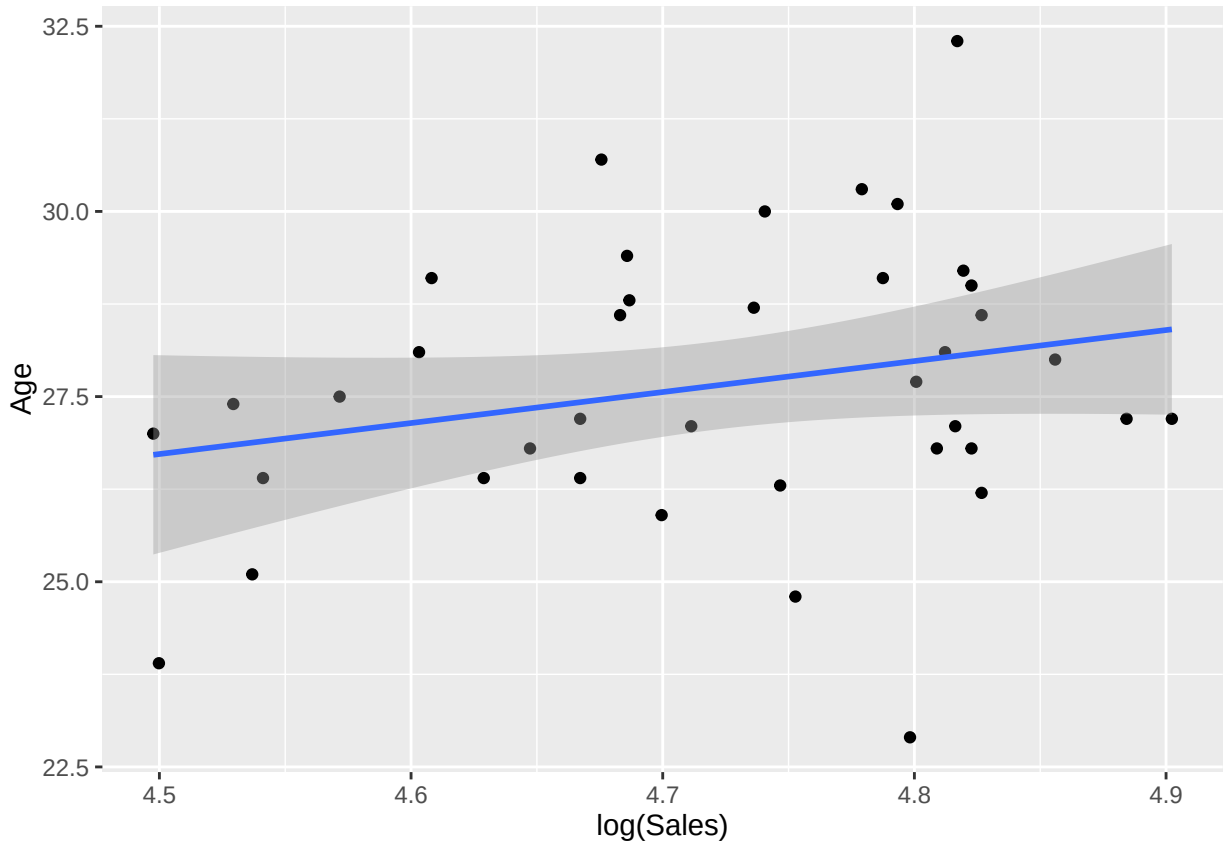

```
ggplot(test, aes(x=log(Sales), y=log(Price))) + geom_point()
```



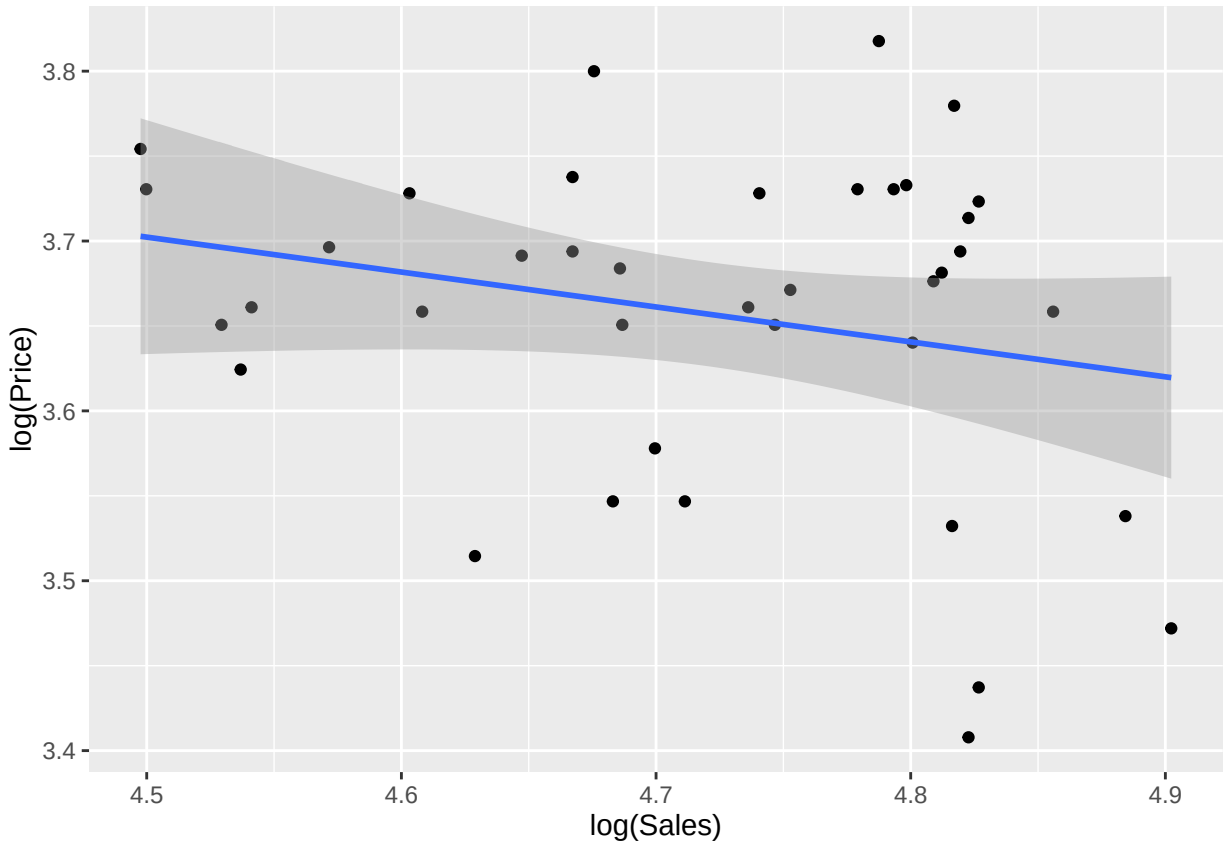

```
ggplot(test, aes(x=log(Sales), y=Income)) + geom_point() +
  geom_smooth(method = "lm")
```




```
ggplot(test, aes(x=log(Sales), y=Age)) + geom_point() +
  geom_smooth(method = "lm")
```




```
ggplot(test, aes(x=log(Sales), y=log(Price))) + geom_point() +
  geom_smooth(method = "lm")
```




```
northeast <- c("ME", "NH", "VT", "MA", "RI", "CT", "NY", "PA", "NJ")
midwest <- c("ND", "SD", "NE", "KS", "MN", "IA", "MO", "IL", "WI", "MI",
             "IN", "OH")
south <- c("DE", "MD", "DC", "WV", "VA", "KY", "NC", "TN", "SC", "GA",
           "AL","MS","AR","LA","OK","TX")
west <- c("WA", "OR", "ID", "MT", "WY", "CA", "NV", "UT", "CO", "AZ", "NM")

regionlookup <- rbind(data.frame(Region = "Northeast", State = northeast),
                      data.frame(Region = "Midwest", State = midwest),
                      data.frame(Region = "South", State = south),
                      data.frame(Region = "West", State = west))

cigarettesregion <- merge(cigarettes,regionlookup, by="State")

summary(lm(Sales ~ Price + Income + Age + Region, data = cigarettesregion))
```

```
##
## Call:
## lm(formula = Sales ~ Price + Income + Age + Region, data = cigarettesregion)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.878 -14.920  -1.593   3.405 112.951
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   137.494568 103.091630   1.334  0.19143
## Price          -3.938789   1.180731  -3.336  0.00211 **
## Income          0.020311   0.009304   2.183  0.03625 *
## Age             2.630944   3.636058   0.724  0.47443
## RegionMidwest -22.595138  15.638524  -1.445  0.15793
## RegionSouth   -12.687118  14.588463  -0.870  0.39076
## RegionWest    -20.274393  15.653935  -1.295  0.20425
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.64 on 33 degrees of freedom
## Multiple R-squared:  0.4034, Adjusted R-squared:  0.295
## F-statistic: 3.719 on 6 and 33 DF,  p-value: 0.006194
```

```
summary(lm(log(Sales) ~ log(Price) + Income + Age, data = cigarettes))
```

```
##
## Call:
## lm(formula = log(Sales) ~ log(Price) + Income + Age, data = cigarettes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39430 -0.08775 -0.02037  0.05141  0.71313
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.268e+00  9.398e-01   6.670 4.32e-08 ***
## log(Price)  -8.685e-01  2.587e-01  -3.358  0.00168 **
## Income       1.363e-04  4.832e-05   2.821  0.00728 **
## Age          4.132e-02  1.544e-02   2.676  0.01058 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.187 on 42 degrees of freedom
## Multiple R-squared:  0.3785, Adjusted R-squared:  0.3341
## F-statistic: 8.527 on 3 and 42 DF,  p-value: 0.000154
```

```
summary(lm(log(Sales) ~ log(Price) + Income, data = cigarettes))
```

```
##
## Call:
## lm(formula = log(Sales) ~ log(Price) + Income, data = cigarettes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51980 -0.07619 -0.02601  0.03802  0.74745
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.811e+00  9.811e-01   6.942 1.56e-08 ***
## log(Price)  -7.354e-01  2.714e-01  -2.709  0.00964 **
## Income       1.657e-04  5.031e-05   3.294  0.00198 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1999 on 43 degrees of freedom
## Multiple R-squared:  0.2726, Adjusted R-squared:  0.2387
## F-statistic: 8.056 on 2 and 43 DF,  p-value: 0.001068
```

```
summary(lm(log(Sales) ~ log(Price) + Age, data = cigarettes))
```

```
##
## Call:
## lm(formula = log(Sales) ~ log(Price) + Age, data = cigarettes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38077 -0.11435 -0.02555  0.06032  0.71132
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.28856    1.01294   6.208 1.82e-07 ***
## log(Price)  -0.80812    0.27785  -2.908  0.00573 **
## Age          0.05124    0.01621   3.161  0.00288 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2015 on 43 degrees of freedom
## Multiple R-squared:  0.2608, Adjusted R-squared:  0.2264
## F-statistic: 7.584 on 2 and 43 DF,  p-value: 0.00151
```

```
summary(lm(log(Sales) ~ Income + Age, data = cigarettes))
```

```
##
## Call:
## lm(formula = log(Sales) ~ Income + Age, data = cigarettes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37667 -0.09950 -0.04440  0.08945  0.81261
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.4327894  0.4591763   7.476 2.65e-09 ***
## Income      0.0001229  0.0000536   2.293   0.0268 *
## Age         0.0313465  0.0168679   1.858   0.0700 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2081 on 43 degrees of freedom
## Multiple R-squared:  0.2117, Adjusted R-squared:  0.175
## F-statistic: 5.774 on 2 and 43 DF,  p-value: 0.00601
```

Data Exploration