What would you conclude based on this? How much would you bet that you are right?

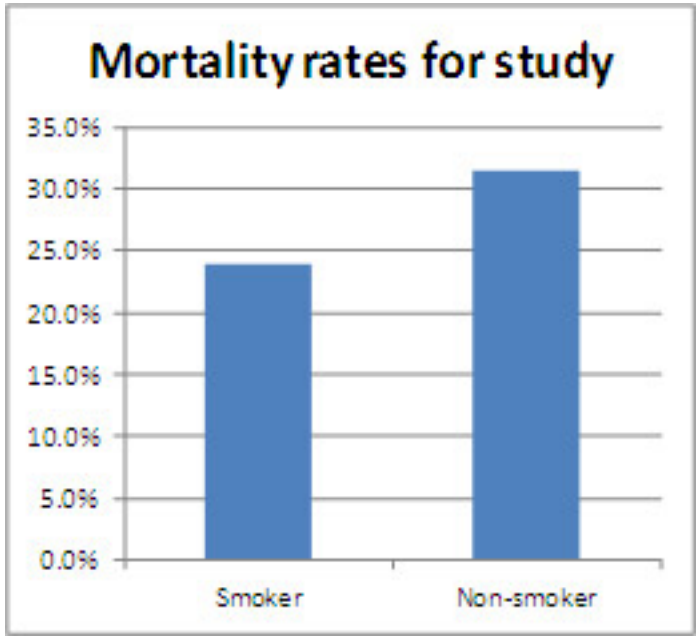
```
Call:
lm(formula = y ~ x, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-96.031 -21.346   0.634  22.624 103.108 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -10.1590     5.5832   -1.82   0.0719 .  
x            30.1596     0.1579  191.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39.44 on 98 degrees of freedom
Multiple R-squared:  0.9973,	Adjusted R-squared:  0.9973 
F-statistic: 3.65e+04 on 1 and 98 DF,  p-value: < 2.2e-16
```

It seems to suggest that X is positively correlated with Y. The few examples should make you skeptical of that conclusion.

```
Call:
lm(formula = y ~ x, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-96.031 -21.346   0.634  22.624 103.108 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -10.1590     5.5832   -1.82   0.0719 .  
x            30.1596     0.1579  191.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39.44 on 98 degrees of freedom
Multiple R-squared:  0.9973,	Adjusted R-squared:  0.9973 
F-statistic: 3.65e+04 on 1 and 98 DF,  p-value: < 2.2e-16
```

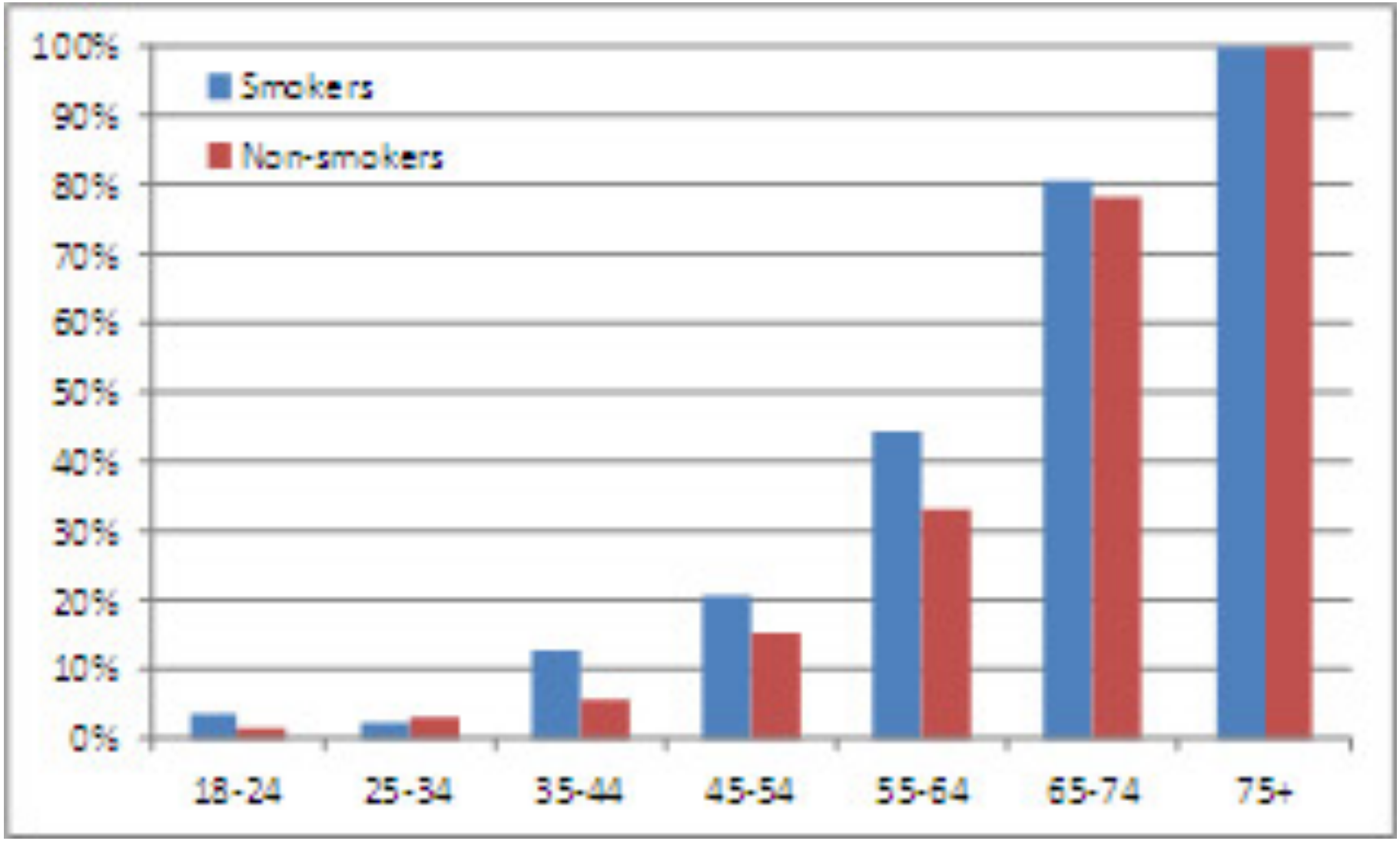
Results of a 20 year study that followed women who smoke and non-smokers

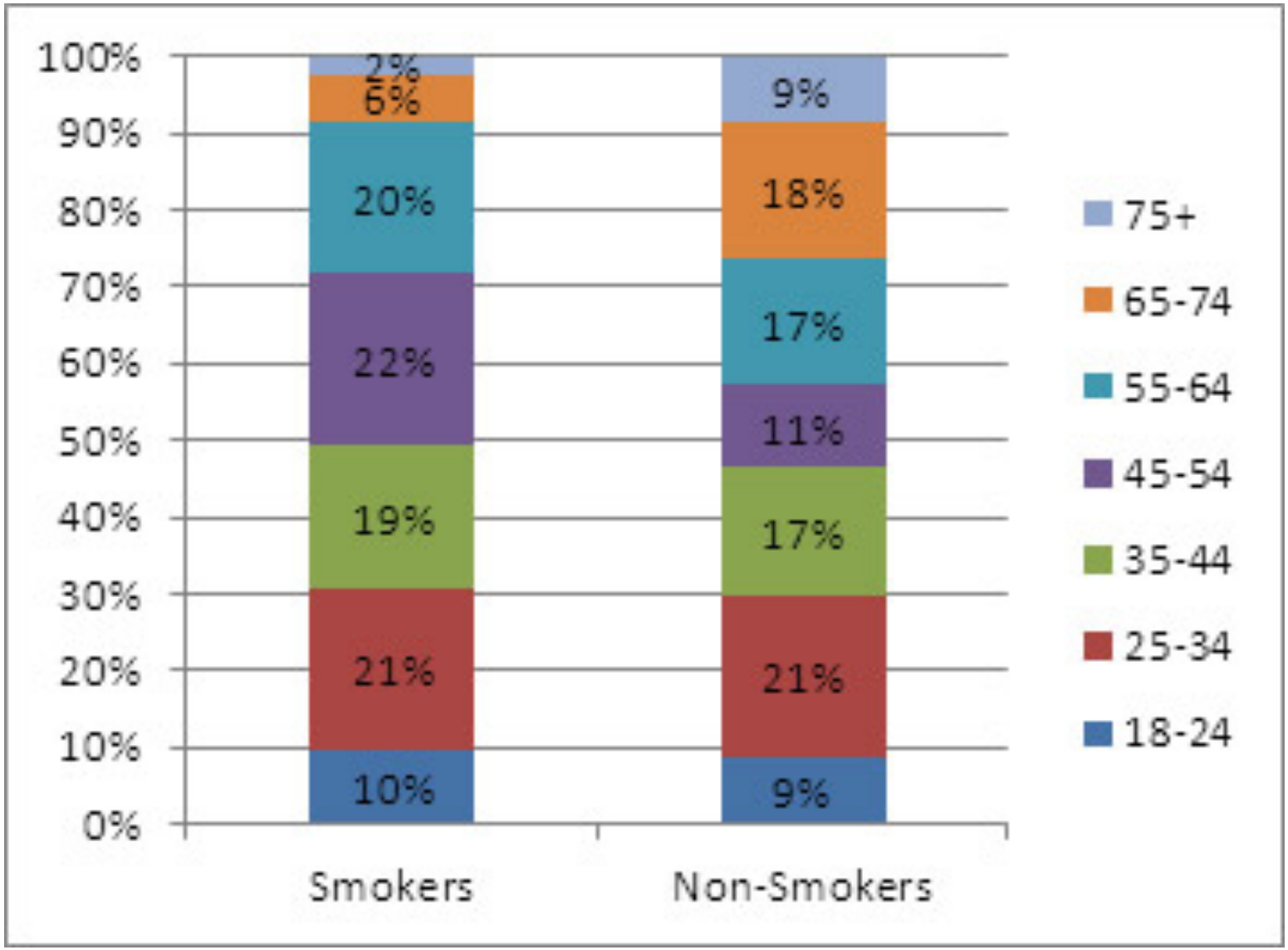
	Died	Survived	Total	Mortality rate
Smoker	139	443	582	23.9%
Non-smoker	230	502	732	31.4%
Total	369	945	1,314	28.1%



Overall, death rates are lower for smokers than non-smokers, but…

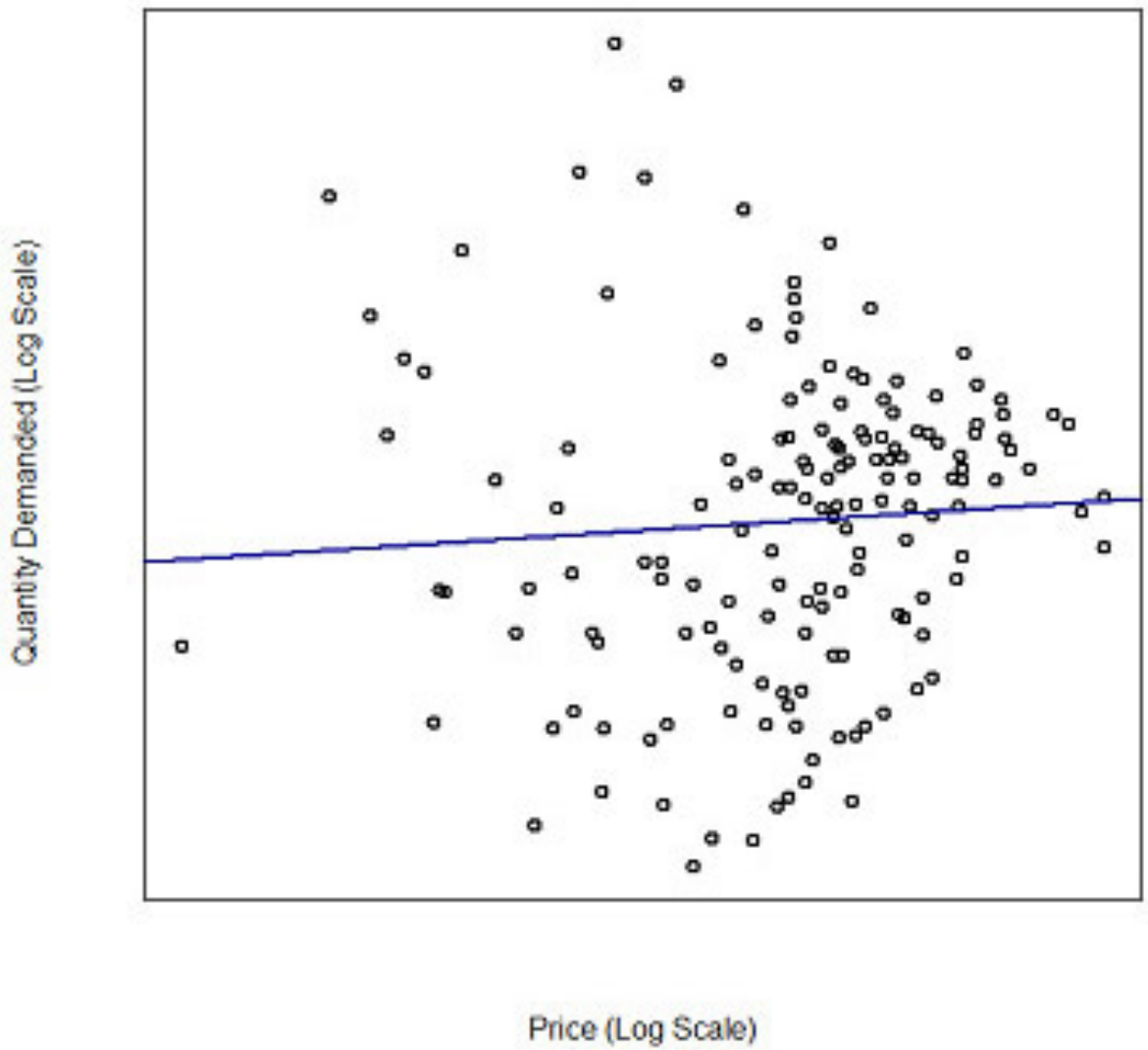
In every age group, death rates are higher for smokers than non-smokers!




100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
2%
6%
20%
22%
19%
21%
10%
9%
18%
17%
11%
17%
21%
9%
Smokers
Non-Smokers
75+
65-74
55-64
45-54
35-44
25-34
18-24

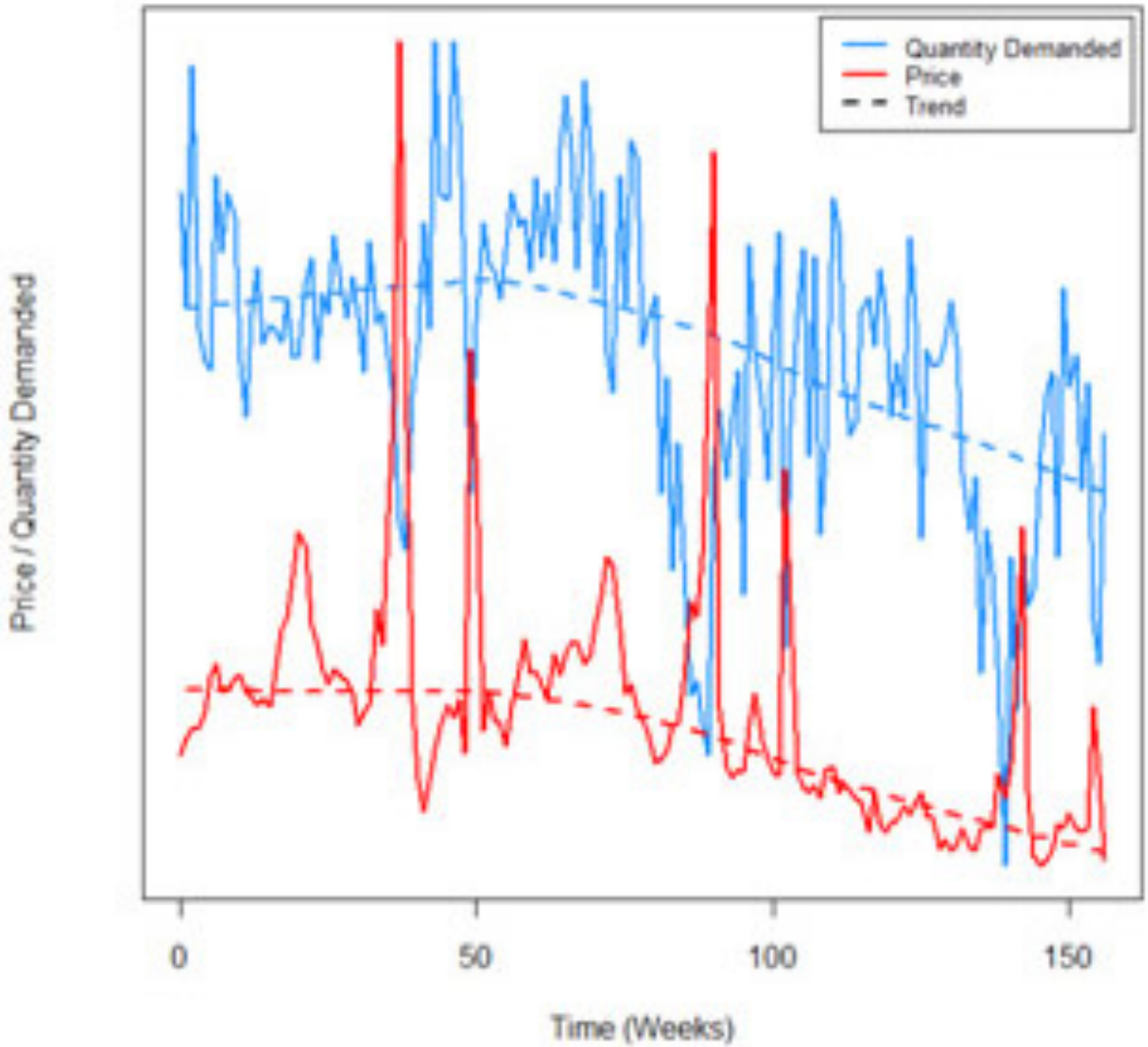
Economic Example

Microeconomics suggests that demand should be lower for higher priced items.
Consider following data for 1 commodity.

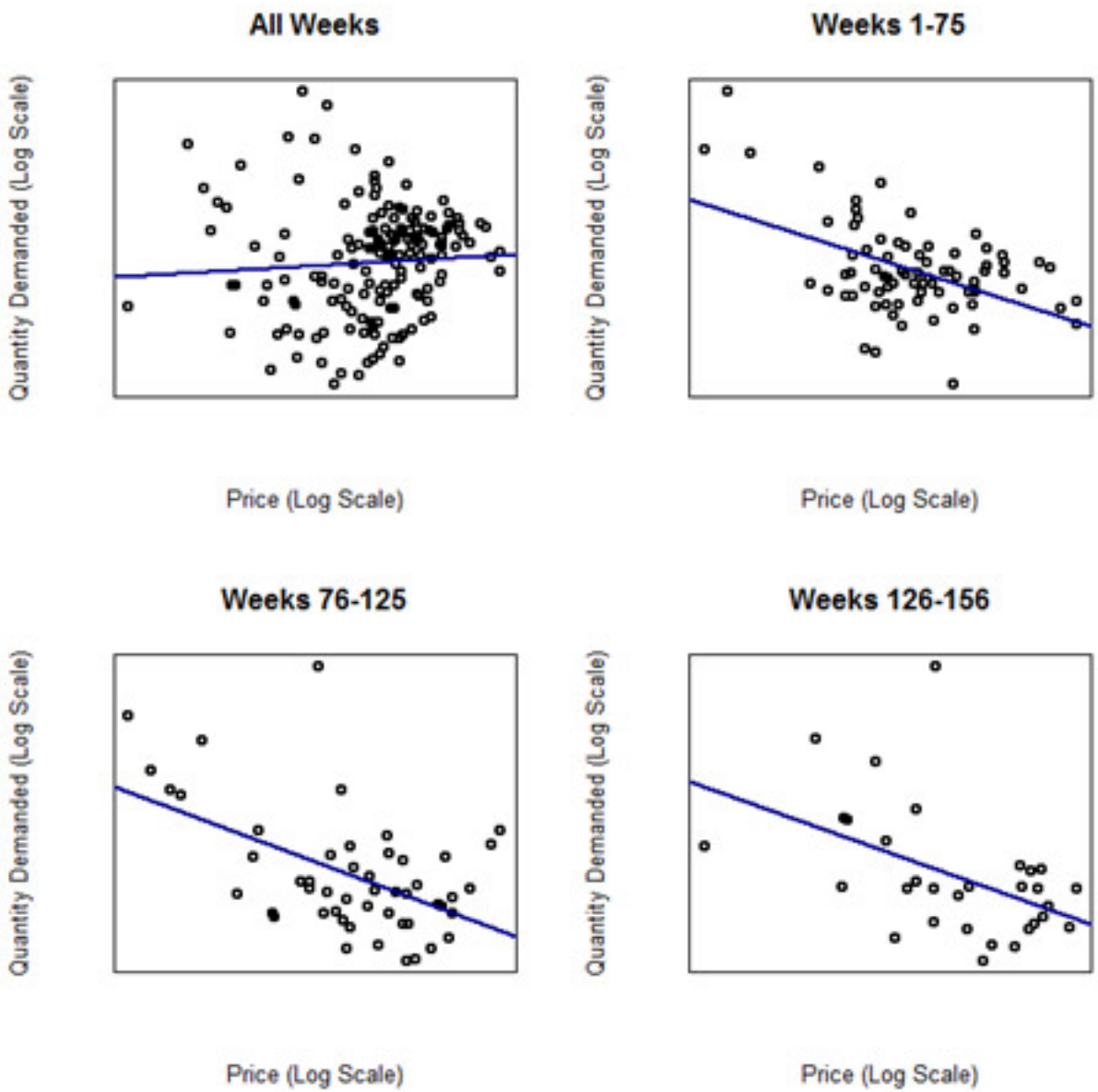


Relationship is Statistically Significantly Positive!

Higher price = Higher demand!



Time matters too!

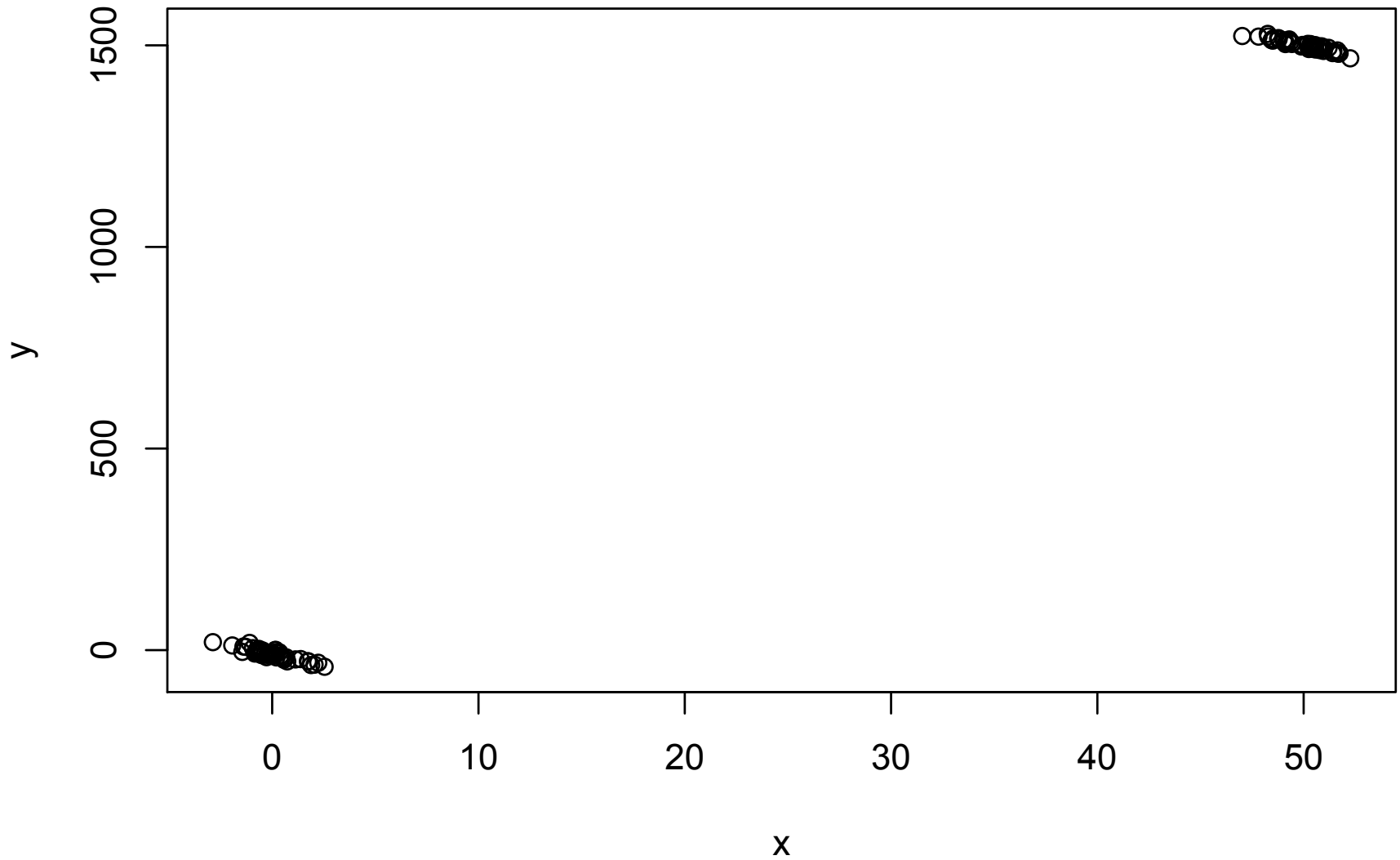


Need to interact
With time period!

Back to introductory example.

```
Call:
lm(formula = y ~ x, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-96.031 -21.346   0.634  22.624 103.108 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -10.1590     5.5832   -1.82   0.0719 .  
x            30.1596     0.1579  191.05   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.44 on 98 degrees of freedom
Multiple R-squared:  0.9973,	Adjusted R-squared:  0.9973 
F-statistic: 3.65e+04 on 1 and 98 DF,  p-value: < 2.2e-16
```

An example of why you should plot your data!



Add another covariate and interact it with X

```
Call:
lm(formula = y ~ x * c, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.0522  -3.3453   0.2285   3.6188  11.9944 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -8.9987     0.7014 -12.829   <2e-16 ***
x             -9.6484     0.7258 -13.293   <2e-16 ***
c           1955.0471    35.0416  55.792   <2e-16 ***
x:c            0.7136     1.0088   0.707    0.481    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.948 on 96 degrees of freedom
Multiple R-squared:      1,   Adjusted R-squared:      1 
F-statistic: 7.752e+05 on 3 and 96 DF,  p-value: < 2.2e-16
```

Simpson's (sic) Paradox (sic) (Stat 525 formulation)

- The correlation between a covariate and a response can change sign after interacting with a second covariate!
- Interaction: lm(y~x*c,data=data)

  (show what it fits)

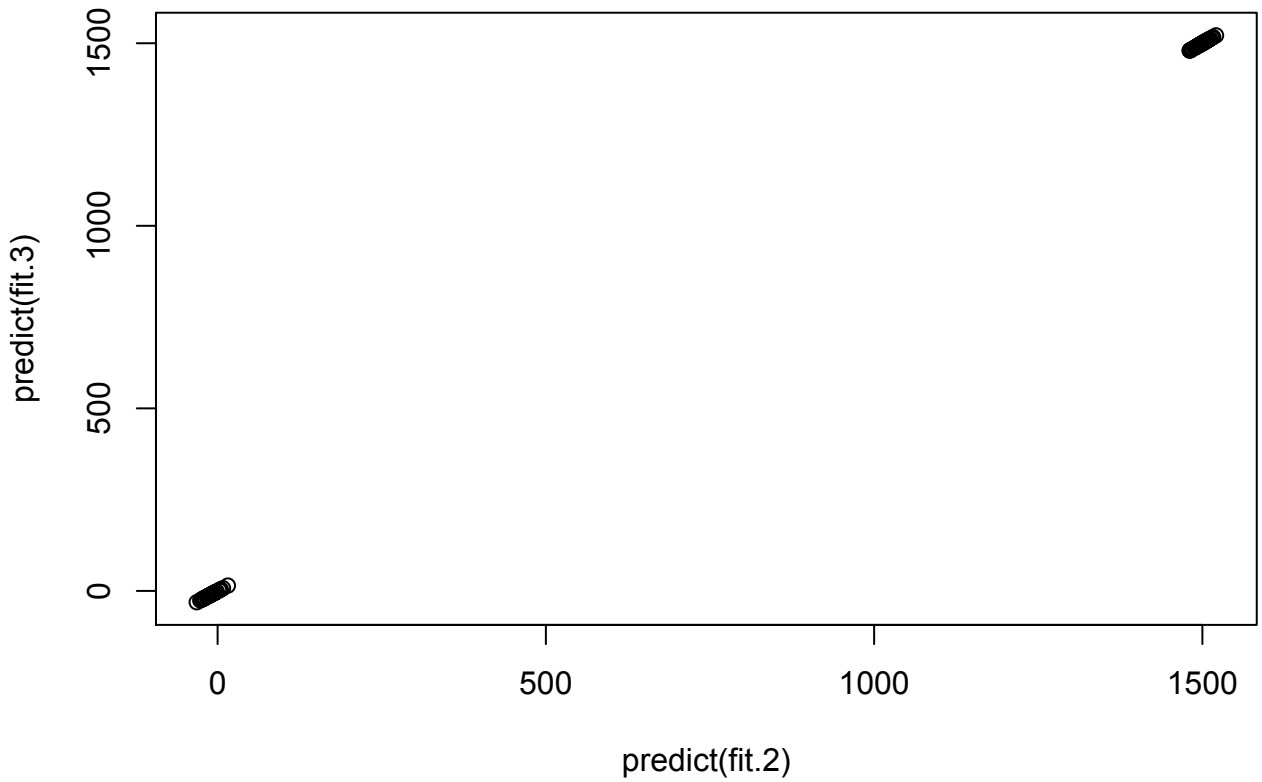
Interaction term looks non-significant

```
Call:
lm(formula = y ~ x * c, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.0522  -3.3453   0.2285   3.6188  11.9944 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -8.9987     0.7014 -12.829   <2e-16 ***
x             -9.6484     0.7258 -13.293   <2e-16 ***
c           1955.0471    35.0416  55.792   <2e-16 ***
x:c            0.7136     1.0088   0.707    0.481    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.948 on 96 degrees of freedom
Multiple R-squared:      1,   Adjusted R-squared:      1 
F-statistic: 7.752e+05 on 3 and 96 DF,  p-value: < 2.2e-16
```

Is this model "better"? (next 2 figures assess that)

```
Call:
lm(formula = y ~ x + c, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.2037  -3.5147   0.2745   3.3868  12.2643 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -9.0236     0.6987  -12.91   <2e-16 ***
x             -9.2790     0.5028  -18.46   <2e-16 ***
c           1972.2790    25.1232   78.50   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.935 on 97 degrees of freedom
Multiple R-squared:      1,   Adjusted R-squared:      1 
F-statistic: 1.169e+06 on 2 and 97 DF,  p-value: < 2.2e-16
```

Estimated ys look pretty similar, but the scale hides differences.



Better way: residuals!