Question 10

a) I created a simple lm() using the Carseats dataset with the data requested

```
# a)
Carseats$US = as.factor(Carseats$US)
model = lm(Sales ~ Price + Urban + US, data = Carseats)
```

b) Here is the summary of model. The intercept of 13.04, Price estimate of -.054, and US factor with Estimate of 1.2 are all considered statistically significant. The Urban factor has an incredibly high p-value, and is therefore considered not significant towards the overall Sales of the carseats prediction

```
Call:
lm(formula = Sales ~ Price + Urban + US, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9206 -1.6220 -0.0564  1.5786  7.0581

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
Price       -0.054459   0.005242 -10.389  < 2e-16 ***
UrbanYes    -0.021916   0.271650  -0.081    0.936
USYes        1.200573   0.259042   4.635 4.86e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.472 on 396 degrees of freedom
Multiple R-squared:  0.2393,	Adjusted R-squared:  0.2335
F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

c) Sales approx =
13.043 - .054 * Price -.022 * UrbanYes + 1.20 * USYes
, where if Urban or US = yes, then UrbanYes or USYes = 1, else 0

d) We can reject the null hypothesis of Bj = 0 for each predictor except UrbanYes, since all of them have p-values that are at the *** significance level (below .001)

e) I repeated part a, but left out the Urban predictor.

```
# e)
model2 = lm(Sales ~ Price + US, data = Carseats)
```

f) I extracted the R^2 variable from both model and model2 using summary(model)$r.squared

```
[1] "R^2 for model 1:"
[1] 0.2392754
[1] "R^2 for model 2:"
[1] 0.2392629
```

Both of these models perform pretty poorly, however the R^2 for model 1 is ever so slightly better, probably due to its added predictor, even if it's not statistically significant

g) Shown is are the confidence intervals for the intercept, Price, and USYes coefficients for the second model from part e. Each of these are at a 95% confidence level

```
                  2.5 %      97.5 %
(Intercept) 11.79032020 14.27126531
Price       -0.06475984 -0.04419543
USYes        0.69151957  1.70776632
```

Supp

a) I wrote the model using the following screenshot, and got the summary results of the second. Adding the interaction term did not help the model much at all, if anything. The R^2 statistic is still the same

```
# a)
model3 = lm(Sales ~ Price + US + Price:US, data = Carseats)
model4 = lm(Sales ~ Price*US, data = Carseats)
```

```
[1] "---------------------"

Call:
lm(formula = Sales ~ Price * US, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.9299 -1.6375 -0.0492  1.5765  7.0430 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 12.974798   0.953079  13.614  < 2e-16 ***
Price       -0.053986   0.008163  -6.613 1.22e-10 ***
USYes        1.295775   1.252146   1.035    0.301    
Price:USYes -0.000835   0.010641  -0.078    0.937    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.472 on 396 degrees of freedom
Multiple R-squared:  0.2393,	Adjusted R-squared:  0.2335 
F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Sales ~ Price + US + I(Price^2), data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.8801 -1.6077 -0.0547  1.5830  7.0528 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 14.2586982  1.8822098   7.576 2.55e-13 ***
Price       -0.0770603  0.0330277  -2.333   0.0201 *  
USYes        1.2126624  0.2593130   4.676 4.01e-06 ***
I(Price^2)   0.0000987  0.0001425   0.692   0.4890    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.471 on 396 degrees of freedom
Multiple R-squared:  0.2402,	Adjusted R-squared:  0.2344 
F-statistic: 41.73 on 3 and 396 DF,  p-value: < 2.2e-16

[1] "---------------------------------------"

Call:
lm(formula = Sales ~ US + poly(Price, 2), data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.8801 -1.6077 -0.0547  1.5830  7.0528 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)       6.7142     0.2079  32.289  < 2e-16 ***
USYes             1.2127     0.2593   4.676 4.01e-06 ***
poly(Price, 2)1 -25.7719     2.4752 -10.412  < 2e-16 ***
poly(Price, 2)2   1.7157     2.4775   0.692    0.489    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.471 on 396 degrees of freedom
Multiple R-squared:  0.2402,	Adjusted R-squared:  0.2344 
F-statistic: 41.73 on 3 and 396 DF,  p-value: < 2.2e-16
```

b) In the summaries, you can see how I called each function under the Call: header. The estimates are all slightly between the two models, but the RSE and R^2 are the same. They produce slightly better results than the simple linear model.

c) The difference between poly() and I() is that the poly() command by default picks Coefficients for X and X^2 that are not correlated with each other. One source online said that the raw=FALSE (the default option) produces coefficients that are orthogonal to each other, while there is typically collinearity among the coefficients when raw=TRUE. In this second screenshot, I set raw=TRUE, which matches that of the I(Price^2) model: (See next page)

```
Call:
lm(formula = Sales ~ Price + US + I(Price^2), data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.8801 -1.6077 -0.0547  1.5830  7.0528 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 14.2586982  1.8822098   7.576 2.55e-13 ***
Price       -0.0770603  0.0330277  -2.333   0.0201 *  
USYes        1.2126624  0.2593130   4.676 4.01e-06 ***
I(Price^2)   0.0000987  0.0001425   0.692   0.4890    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.471 on 396 degrees of freedom
Multiple R-squared:  0.2402,	Adjusted R-squared:  0.2344 
F-statistic: 41.73 on 3 and 396 DF,  p-value: < 2.2e-16

[1] "--------------------------------------------"

Call:
lm(formula = Sales ~ US + poly(Price, 2, raw = TRUE), data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.8801 -1.6077 -0.0547  1.5830  7.0528 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 14.2586982  1.8822098   7.576 2.55e-13 ***
USYes                        1.2126624  0.2593130   4.676 4.01e-06 ***
poly(Price, 2, raw = TRUE)1 -0.0770603  0.0330277  -2.333   0.0201 *  
poly(Price, 2, raw = TRUE)2  0.0000987  0.0001425   0.692   0.4890    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.471 on 396 degrees of freedom
Multiple R-squared:  0.2402,	Adjusted R-squared:  0.2344 
F-statistic: 41.73 on 3 and 396 DF,  p-value: < 2.2e-16
```