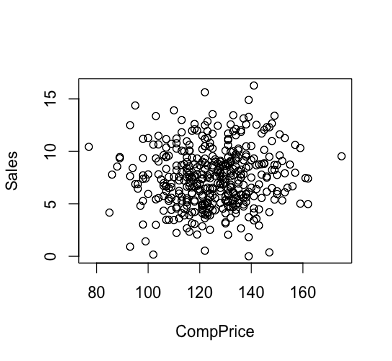
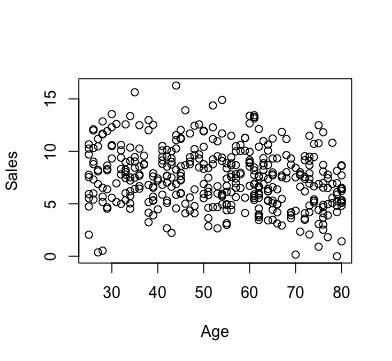
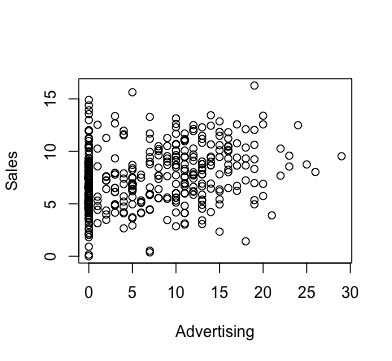
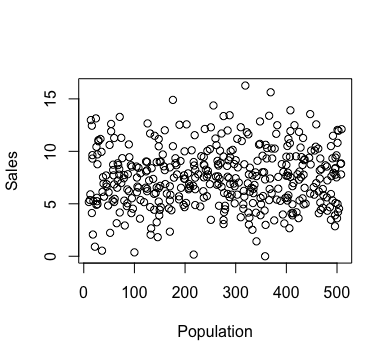
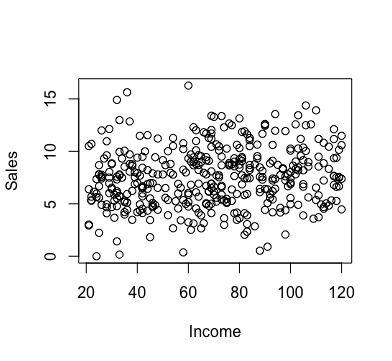
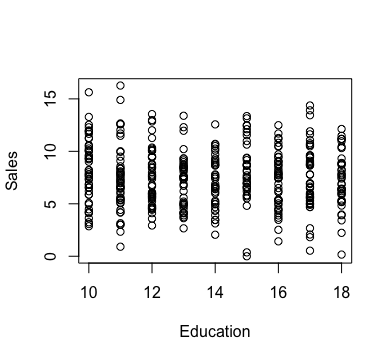
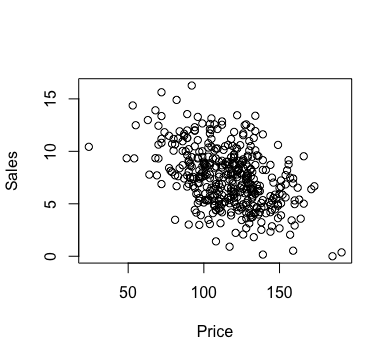
**Question 1: Exploratory Data Analysis**







Trends:

* Sales appears to be positively correlated with Advertising expenditure for non-zero Advertising values.
* Sales appears to be negatively correlated with Price.
* Sales also seems to be slightly negatively correlated with Age and Education.
* Other variables(CompPrice, Income, Population) don’t seem to be correlated with Sales very much.

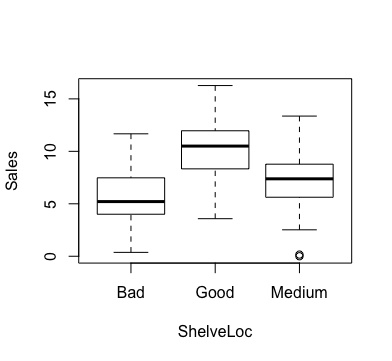
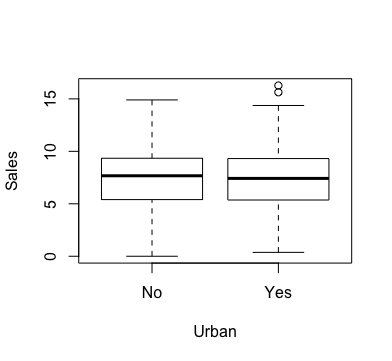
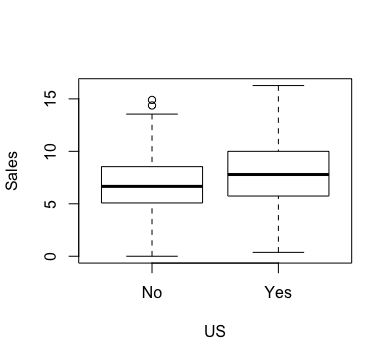
Call:

lm(formula = Sales ~ variable, data = data) for each variable

Coefficients:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | CompPrice | Income | Advertising | Population | Price | Age | Education |
| Coef | 0.0118 | 0.01533 | 0.1144 | 0.0009672 | -0.05307 | -0.04041 | -0.05599 |

Interpretation: With a unit increase of each numeric variable, Sales increases(or decreases if negative) by the coefficient according to each model(involving that variable).

Trends:

* Sales appears to be correlated with better shelving location.
* Whether the location is urban does not seem to be correlated with sales very much.
* Sales in US locations appear to be slightly better.

Call:

lm(formula = Sales ~ as.factor(variable), data = data) for each variable

Coefficients:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | ShelveLocGood | ShelveLocMedium | UrbanYes | USYes |
| Coef | 4.691 | 1.784 | -0.09537 | 1.044 |

Interpretation:

* Sales appear to differ by shelving location. Good and medium shelving locations are expected to increase sales by 4.691 and 1.784, respectively, according to the model.
* Sales will be 0.09537 units lower if the location is urban according to the model.
* Sales will be 1.044 units higher if the location is in the US according to the model.

Based on the exploratory analysis, I think it is reasonable to assume a linear regression model with **select** variables. Advertising, Price, and ShelveLoc appear to have stronger linear relationships with Sales based on the plots. I would start with a model of those three and then see if adding variables with weaker correlations such as Age and Education to the model could help.

**Question 2: Fitting the Linear Regression Model**

(i) What are the model parameters and their estimates?

Call:

lm(formula = Sales ~ ., data = data)

Coefficients:

(Intercept) CompPrice Income Advertising Population Price

5.6606231 0.0928153 0.0158028 0.1230951 0.0002079 -0.0953579

ShelveLocGood ShelveLocMedium Age Education UrbanYes USYes

4.8501827 1.9567148 -0.0460452 -0.0211018 0.1228864 -0.1840928

Variance: 1.0192 = 1.038361

The model parameters are the regression coefficients(betas) and the variance above.

(ii) Write down the equation for the regression line;

y = 5.6606231 + 0.0928153\*CompPrice + 0.0158028\*Income + 0.1230951\*Advertising + 0.0002079\*Population – 0.0953579\*Price + ShelveLocGood\*4.8501827 + 1.9567148\*ShelveLocMedium – 0.0460452\*Age – 0.0211018\*Education + 0.1228864\*UrbanYes – 0.1840928\*USYes

(iii) Interpret the estimated value of the parameters corresponding with Advertising, Shelve-Loc and Price in the context of the problem (include its standard error in your interpretation).

* For each thousand dollars spent on Advertising, Sales is expected to increase by 0.1230951 with a standard error of 0.0111237 controlling for other variables.
* If shelving location is good, Sales is expected to increase by 4.8501827 with a standard error of 0.1531100 compared to bad shelving location controlling for other variables.
* If shelving location is medium, Sales is expected to increase by 1.9567148 with a standard error of 0.1261056 compared to bad shelving location controlling for other variables.
* For each unit the price is higher, Sales is expected decrease by 0.0953579 controlling for other variables.

(iv) Find a 95% confidence interval for the parameters corresponding to all predictors plus the intercept.

2.5 % 97.5 %

(Intercept) 4.4741845403 6.8470615860

CompPrice 0.0846606548 0.1009700295

Income 0.0121751564 0.0194305162

Advertising 0.1012248453 0.1449653319

Population -0.0005204751 0.0009362292

Price -0.1006095157 -0.0901063219

ShelveLocGood 4.5491536828 5.1512117392

ShelveLocMedium 1.7087788955 2.2046507169

Age -0.0523007213 -0.0397896046

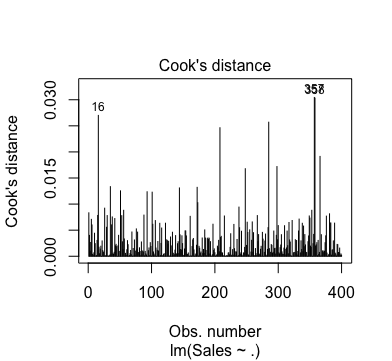
Education -0.0598742386 0.0176705608

UrbanYes -0.0992355421 0.3450083352

USYes -0.4786972882 0.1105116391

**Question 3: Outliers.**

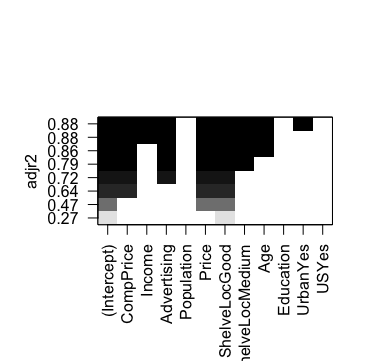
* Using R, I identified 15 data points whose Cook’s distances exceed the threshold of D = 4/n= 4/400 = 0.01.
* However, given the abundance of points that had Cook’s distances slightly below 0.01, I could not justify excluding points that had Cook’s distances slightly above 0.01 as outliers.
* After observing the distribution of Cook’s distances below, I realized that it may be more beneficial to exclude just the 5 outliers whose Cook’s distances exceeded 0.02.



* While it is difficult to know for sure without specific context, I believe that those points may not reflect normal sales data and may influence the regression model.

**Question 4: Variable Selection**

* Based on the previous 2 questions, it could be that not all predictors have a statistically significant association with sales.
* Using the regsubsets function of the leaps package in R, I identified two models that yielded the highest adjusted R2. Of the two, I chose the model with the smaller number of variables.
* The model included CompPrice, Income, Advertising, Price, ShelveLoc, and Age as predictor variables along with the intercept.



**Question 5: Testing the significance of the linear relationship observed in the data.**

(i) What is the P-value of the test?

Call:

lm(formula = Sales ~ CompPrice + Income + Advertising + Price +

ShelveLoc + Age, data = d2)

Residuals:

Min 1Q Median 3Q Max

-2.69939 -0.66400 0.02682 0.69898 2.40632

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.228940 0.490956 10.651 <2e-16 \*\*\*

CompPrice 0.093033 0.003953 23.535 <2e-16 \*\*\*

Income 0.015833 0.001771 8.941 <2e-16 \*\*\*

Advertising 0.115855 0.007414 15.626 <2e-16 \*\*\*

Price -0.094158 0.002569 -36.657 <2e-16 \*\*\*

ShelveLocGood 4.920370 0.146889 33.497 <2e-16 \*\*\*

ShelveLocMedium 1.979897 0.120896 16.377 <2e-16 \*\*\*

Age -0.046204 0.003071 -15.043 <2e-16 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9741 on 387 degrees of freedom

Multiple R-squared: 0.8827, Adjusted R-squared: 0.8806

F-statistic: 416 on 7 and 387 DF, p-value: < 2.2e-16

* The regression coefficient for Price is -0.094158.
* The p-value of the test for Price is 2e-16, approximately zero.

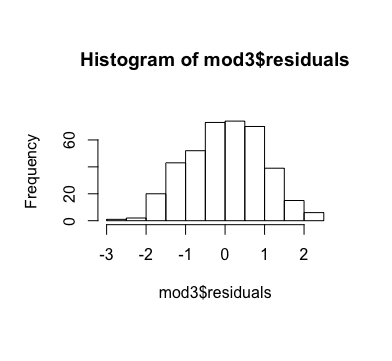
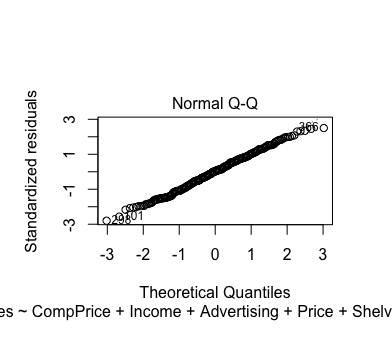
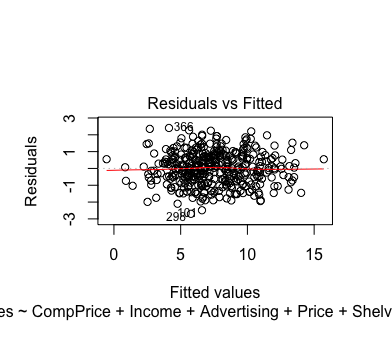
(ii) The p-value tells me the chance of obtaining an estimate more extreme than our outcome if the null hypothesis (that the coefficient was zero) was true.

(iii) As this value is extremely low, we reject the null hypothesis and conclude that Price is a statistically significant predictor of Sales.

**Question 6: Checking Assumptions of the Model**

Assumptions:

* Linearity
* Constant variance
* Independence
* Normality



* In the first plot(residual plot), we can see that there is no particular pattern and the variance is fairly constant across the fitted values. From this, we can see that the linearity and constant variance assumptions hold.
* From the Normal Q-Q plot, we can see a fairly straight diagonal line, confirming the normality assumption. We can see that the normality is clearer after removing the outliers.
* A histogram of the residuals complements this view as well; the residuals have an approximatesymmetric distribution, unimodal and with no gaps in the data.
* Independence cannot be confirmed with this data, as this is largely an observational study. It cannot be determined from the information given that sales data from one store is independent of data from another.