Alla Topp
Statistical Analysis and Experimental Design

Linear Regression

Relationships between car seat sales and other variables from the dataset

Step 1

Our data set contains information about sales of child car seats (a simulated data set containing sales of child car seats at 400 different stores). A format of the data is a data frame with 400 observations on the following 11 variables.

Variables	Description
1. Sales	Unit sales (in thousands) at each location
2. CompPrice	Price charged by competitor at each location
3. Income	Community income level (in thousands of dollars)
4. Advertising	Local advertising budget for company at each location (in thousands of dollars)
5. Population	Population size in region (in thousands)
6. Price	Price company charges for car seats at each site
7. ShelveLoc	A factor with levels Bad, Good and Medium indicating the quality of the shelving location for the car seats at each site
8. Age	Average age of the local population
9. Education	Education level at each location
10. Urban	A factor with levels No and Yes to indicate whether the store is in an urban or rural location
11. US	A factor with levels No and Yes to indicate whether the store is in the US or not

Filter

	Sales	CompPrice	Income	Advertising	Population	Price	ShelveLoc	Age	Education	Urban	US
1	9.50	138	73	11	276	120	Bad	42	17	Yes	Yes
2	11.22	111	48	16	260	83	Good	65	10	Yes	Yes
3	10.06	113	35	10	269	80	Medium	59	12	Yes	Yes
4	7.40	117	100	4	466	97	Medium	55	14	Yes	Yes
5	4.15	141	64	3	340	128	Bad	38	13	Yes	No
6	10.81	124	113	13	501	72	Bad	78	16	No	Yes
7	6.63	115	105	0	45	108	Medium	71	15	Yes	No
8	11.85	136	81	15	425	120	Good	67	10	Yes	Yes
9	6.54	132	110	0	108	124	Medium	76	10	No	No
10	4.69	132	113	0	131	124	Medium	76	17	No	Yes
11	9.01	121	78	9	150	100	Bad	26	10	No	Yes
12	11.96	117	94	4	503	94	Good	50	13	Yes	Yes
13	3.98	122	35	2	393	136	Medium	62	18	Yes	No
14	10.96	115	28	11	29	86	Good	53	18	Yes	Yes
15	11.17	107	117	11	148	118	Good	52	18	Yes	Yes
16	8.71	149	95	5	400	144	Medium	76	18	No	No
17	7.58	118	32	0	284	110	Good	63	13	Yes	No

Alla Topp
Statistical Analysis and Experimental Design

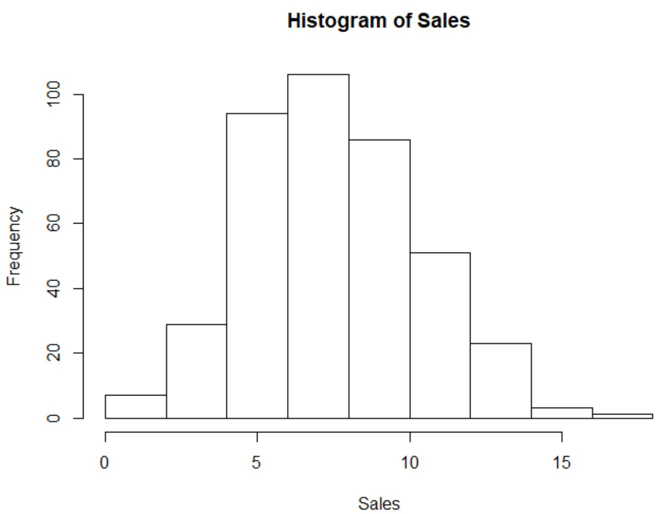
```r
# Alla Topp
# MSDS 660
# Assignment 2

install.packages("ISLR") #install the package for the first time
library("ISLR")   #load the package
data("Carseats") #load specified data set, here "Carseats"
attach(Carseats)

str(Carseats)      # display structure
summary(Carseats) # get summary of your data set/data frame
names(Carseats)   #Lists names of variables in a data.frame
plot(Carseats)

```

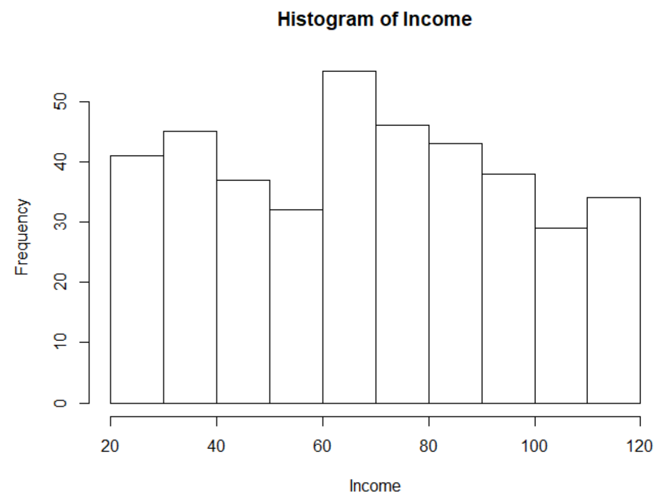
```
> str(Carseats)      # display structure
'data.frame':   400 obs. of  11 variables:
 $ Sales      : num  9.5 11.22 10.06 7.4 4.15 ...
 $ CompPrice  : num  138 111 113 117 141 124 115 136 132 132 ...
 $ Income     : num  73 48 35 100 64 113 105 81 110 113 ...
 $ Advertising: num  11 16 10 4 3 13 0 15 0 0 ...
 $ Population : num  276 260 269 466 340 501 45 425 108 131 ...
 $ Price      : num  120 83 80 97 128 72 108 120 124 124 ...
 $ ShelveLoc  : Factor w/ 3 levels "Bad","Good","Medium": 1 2 3 3 1 1 3 2 3 3 ...
 $ Age        : num  42 65 59 55 38 78 71 67 76 76 ...
 $ Education  : num  17 10 12 14 13 16 15 10 10 17 ...
 $ Urban      : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 2 2 1 1 ...
 $ US         : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 1 2 1 2 ...
> summary(Carseats) # get summary of your data set/data frame
     Sales          CompPrice       Income        Advertising       Population        Price        ShelveLoc        Age
 Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000   Min.   : 10.0   Min.   : 24.0   Bad   : 96   Min.   :25.00
 1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000   1st Qu.:139.0   1st Qu.:100.0   Good  : 85   1st Qu.:39.75
 Median : 7.490   Median :125   Median : 69.00   Median : 5.000   Median :272.0   Median :117.0   Medium:219   Median :54.50
 Mean   : 7.496   Mean   :125   Mean   : 68.66   Mean   : 6.635   Mean   :264.8   Mean   :115.8                Mean   :53.32
 3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000   3rd Qu.:398.5   3rd Qu.:131.0                3rd Qu.:66.00
 Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000   Max.   :509.0   Max.   :191.0                Max.   :80.00
   Education      Urban       US
 Min.   :10.0   No :118   No :142
 1st Qu.:12.0   Yes:282   Yes:258
 Median :14.0
 Mean   :13.9
 3rd Qu.:16.0
 Max.   :18.0
> names(Carseats)   #Lists names of variables in a data.frame
 [1] "Sales"       "CompPrice"   "Income"      "Advertising" "Population"  "Price"       "ShelveLoc"   "Age"         "Education"
[10] "Urban"       "US"
> plot(Carseats)
```

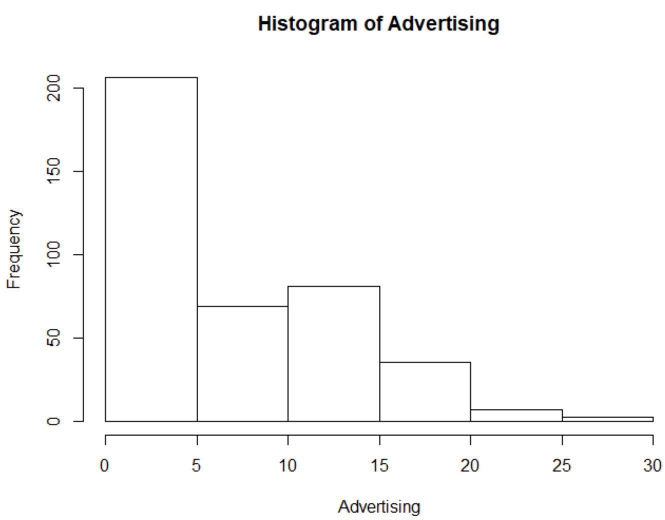
2) Perform data exploration by plotting the graph(s), the distribution, and so on. Then, interpret them.

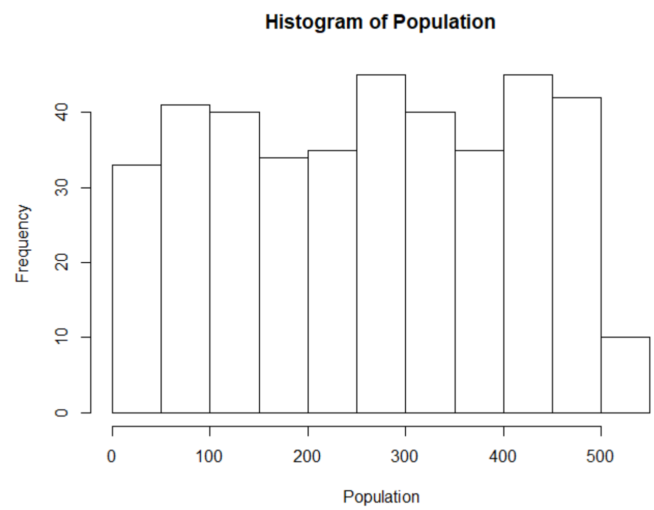
```r
# Data exploration with graphs (hist and density)
hist(Sales) # create histograms where "Sales" values are plotted
hist(Income)    # create histograms where "Income" values are plotted
hist(Advertising) # create histograms where "Advertising" values are plotted
hist(Population)  # create histograms where "Population" values are plotted

plot(density(Sales)) #returns the density data and plots the results of "Sales"
plot(density(Income))
plot(density(Advertising))
plot(density(Population))
```

Alla Topp
Statistical Analysis and Experimental Design









Explanation:

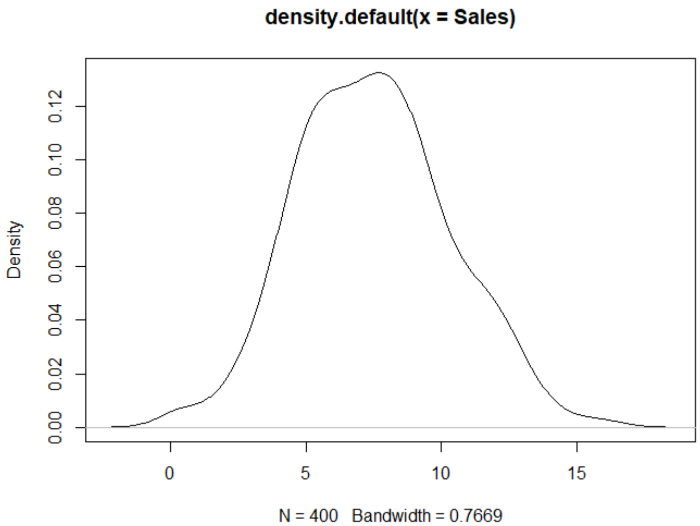
A histogram is a visual representation of the distribution of a dataset. It helps us to easily see where a relatively large amount of the data is situated and where there is very little data to be found. The first histogram shows the distribution of the “Sales”, we can see that the most sales of car seats happened to be between approximately 5 and 7 thousand.

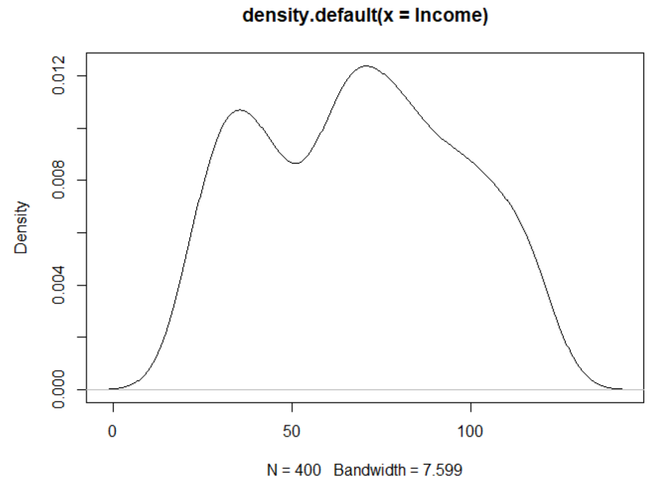
On the 2nd histogram we see the distribution of “Income” values. We can see that the most amount of people are making between 60 and 70 thousand.

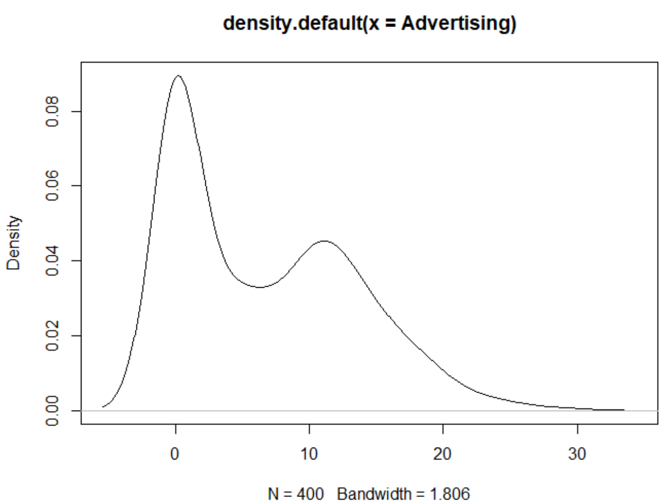
On the 3d histogram we could see the distribution of the “Advertising” values. Mostly, Local advertising budget for company at each location is between 0 and 5 thousand.

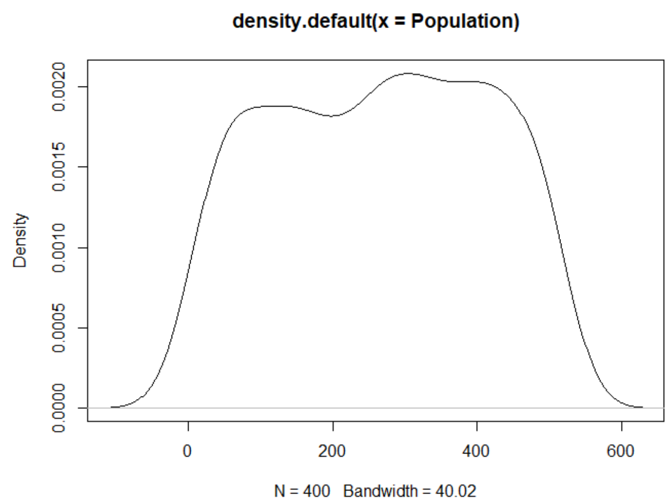
On the 4th histogram we could see the distribution of the “Population” values. The most frequency is located at250-300 and 400-450 thousand people.

Alla Topp
Statistical Analysis and Experimental Design









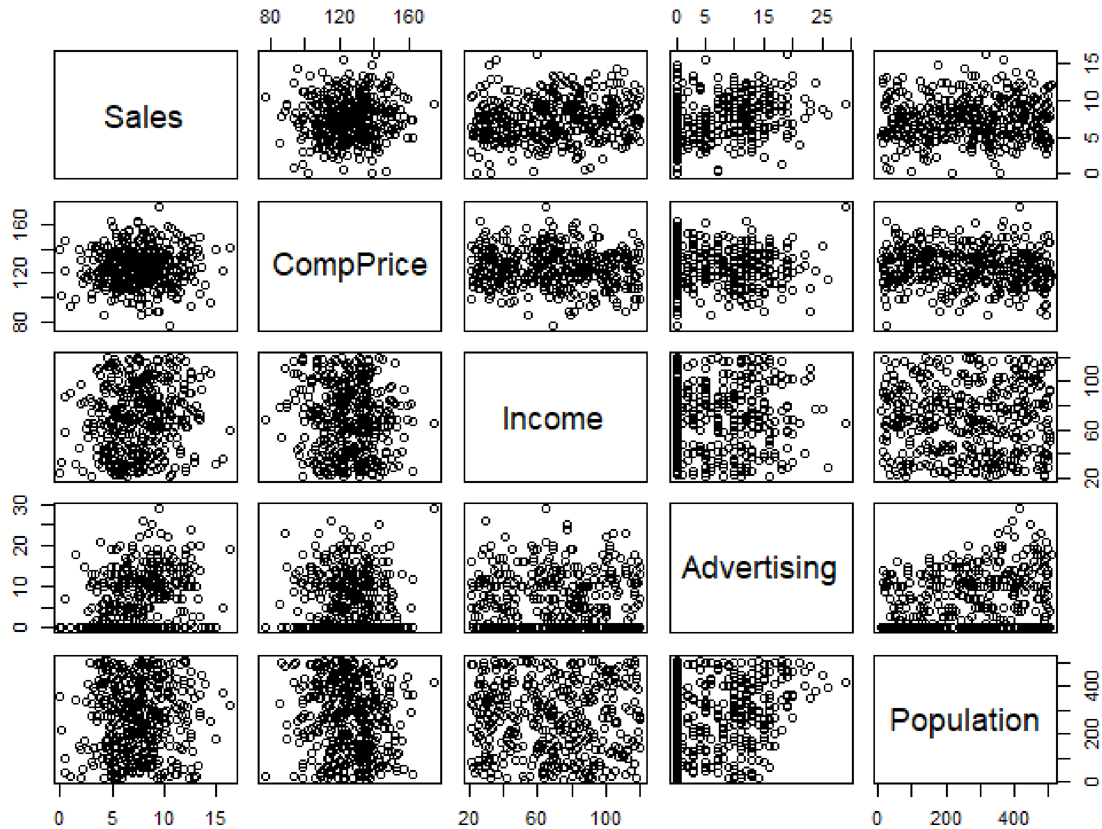
Explanation:

Another way to represent the distribution is with density plots. Kernal density plots are usually a much more effective way to view the distribution of a variable.

3) Perform pairwise scatterplots and describe your findings, such as which pairs show positive correlation, negative correlation, or no correlation?

```
> pairs(Carseats[,1:5]) #visually checks possible correlated variables.
> cor(Carseats[,1:5]) #calculate correlation coefficient between variables
                 Sales   CompPrice       Income Advertising   Population
Sales       1.00000000  0.06407873  0.151950979  0.26950678  0.050470984
CompPrice   0.06407873  1.00000000 -0.080653423 -0.02419879 -0.094706516
Income      0.15195098 -0.08065342  1.000000000  0.05899471 -0.007876994
Advertising 0.26950678 -0.02419879  0.058994706  1.00000000  0.265652145
Population  0.05047098 -0.09470652 -0.007876994  0.26565215  1.000000000
```

Alla Topp
Statistical Analysis and Experimental Design



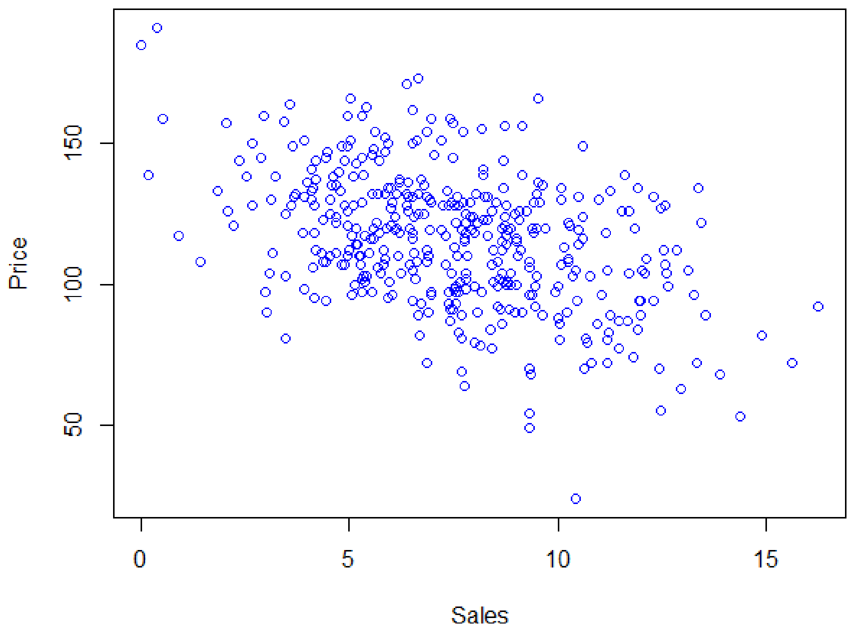
Explanation:

We can see that some pairs have a positive correlation, and some have a negative correlation. For example, pairs Sails and CompPrice, Sales and Income, Sales and Advertising, Sales and Population have positive correlation, Pairs like CompPrice and Income, CompPrice and Advertising, CompPrice and Population have negative correlation.

4) Select one independent (a predictor or X variable) and one dependent variable (a response or Y variable).

I chose x (independent variable) to be “Sales” and dependent variable (y) to be “Price” because I want to predict Sales based on the prices are set. I want to see if prices are higher the sales would increase or decrease.

```
# Select 1 independant(x) and dependent(y), scatterplot of the variables
x <- Carseats$Sales   # independent variable
y <- Carseats$Price     # dependant variable
#produces scatterplot of the variables X and Y with X on the x-axis and Y on the y-axis.
plot(x,y, xlab = "Sales", ylab = "Price", col ="Blue")
```



5) State your null and alternative hypothesis:

H0: <null hypothesis>; H1: <alternative hypothesis>

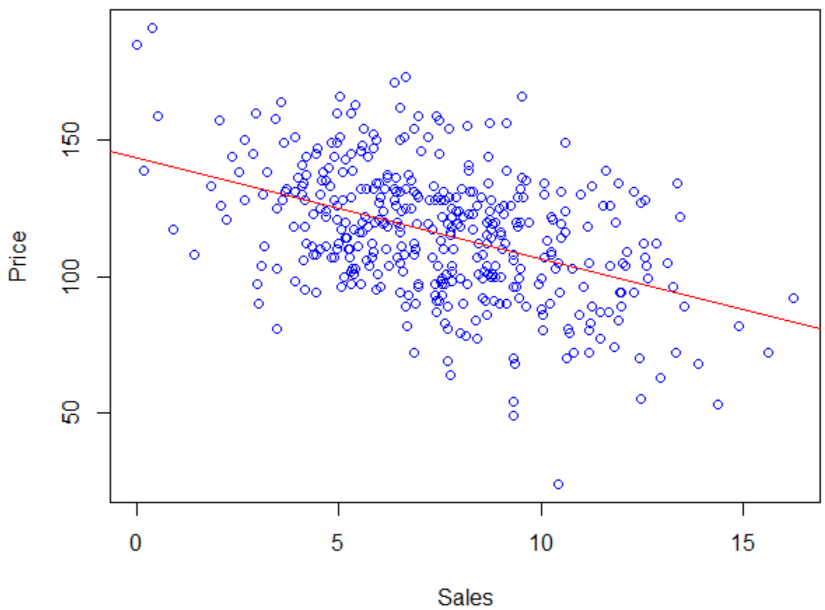
Null hypothesis is that there is no relationship between “Sales” and “Price”, alternative hypothesis would state the opposite – there is relationship between “Sales” and “Price”.

```
        Welch Two Sample t-test

data:  x and y
t = -90.837, df = 410.35, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -110.6423 -105.9550
sample estimates:
 mean of x  mean of y
  7.496325 115.795000
```

6) Show linear regression commands in R and the corresponding results.

```
#Runs a regression of Y on X where Y is a dependent variable and X is an independent variable.
lm_carseats = lm(y ~ x)  # saves all outputs to an object "lm_carseats"
# adds a regression line on the scatterplot graphed earlier
abline(lm(y ~ x), col = "red")

# shows the model formula, residual quartiles, coefficient estimate with std error,
# and a significance test, multiple and adjusted R-square, and F-test for model fit
summary(lm_carseats)
```

Alla Topp
Statistical Analysis and Experimental Design




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

Residuals:
    Min      1Q  Median      3Q     Max 
-80.851 -15.332   0.528  13.315  57.791 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 143.7589     3.0143  47.692   <2e-16 ***
x            -3.7304     0.3763  -9.912   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.23 on 398 degrees of freedom
Multiple R-squared:  0.198,	Adjusted R-squared:  0.196 
F-statistic: 98.25 on 1 and 398 DF,  p-value: < 2.2e-16
```

We can see on the graph that as prices on the car seats increase, sales go down.

7) Explain how you read the result. Do you accept or reject the null hypothesis? Why? Discuss.

If the p-value is less than 0.05, then the null hypothesis is rejected, and we have evidence that the data are not from a normally distributed population – in other words, the lower the p-value, the lower the chance the data came from a normal population.

In this case, as the p-value is much less than 0.05, we reject the null hypothesis that β = 0(that states there is no relationship between x and y) and accept the alternative hypothesis that there is a significant relationship between x(sales) and y(prices) in carseats dataset.

8) Is there any evidence of a linear relationship between the predictor(x) and response(y) variable that you chose? Explain.

Yes, the coefficient p-value has a very low value. The coefficient states a negative relationship between Price and Sales: as Price increases, Sales decreases (the coefficient has a negative value). Another evidence is R2 value that shows us good relationship between sales and price. R2 can be interpreted as the percentage of variance in the dependent variable that can be explained by the predictors(x). The R2 statistic records the percentage of variability in the response that is explained by the predictors. The predictor "Sales" explains 19.8% of the variability in "Price" because R2 = 0.198. It means we have good evidence of the relationship between sales and price.

9) Verify the model assumptions (e.g. linearity, normality, variance) and specify which assumptions do not hold, if there are any.

Based on the scatterplot we provided, linear regression statistics and t-test we can state that there is significant relationship between dependent and independent variables (relationship is linear).

We can see on the first histogram that the distribution of means across samples is normal. For any fixed value of X, Y is normally distributed. The residuals are not skewed, that means that the assumption is satisfied.

It shows how the data points are spread along the range of regression line when we look at the linear regression plot above, it shows equal variance of residuals across the variable range. Residuals are evenly distributed across the range and do not appear spread or narrow at any point.