E-learning Dumitrul Lunic, Andrei-Flavius Vacaru

1. Task 1

Load the data Carseats from the package ISLR

```
# install(packages = c("ggplot2", "ISLR", "car", "lmtest"))
library(ggplot2)
library(ISLR)
library(car)
```

```
## Loading required package: carData
```

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
data <- ISLR::Carseats
head(data)
```

```
##   Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 1  9.50       138     73          11        276   120       Bad  42        17
## 2 11.22       111     48          16        260    83      Good  65        10
## 3 10.06       113     35          10        269    80    Medium  59        12
## 4  7.40       117    100           4        466    97    Medium  55        14
## 5  4.15       141     64           3        340   128       Bad  38        13
## 6 10.81       124    113          13        501    72       Bad  78        16
##   Urban  US
## 1   Yes Yes
## 2   Yes Yes
## 3   Yes Yes
## 4   Yes Yes
## 5   Yes  No
## 6    No Yes
```

2. Task 2

Discover the data using basic statistics and plots (use mean, sd, histogram, boxplot, table, scatter plot, etc.)

```
summary(data) # Basic statistics
```

```
##      Sales          CompPrice       Income        Advertising    
##  Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000  
##  1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000  
##  Median : 7.490   Median :125   Median : 69.00   Median : 5.000  
##  Mean   : 7.496   Mean   :125   Mean   : 68.66   Mean   : 6.635  
##  3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000  
##  Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000  
##    Population        Price        ShelveLoc        Age          Education   
##  Min.   : 10.0   Min.   : 24.0   Bad   : 96   Min.   :25.00   Min.   :10.0  
##  1st Qu.:139.0   1st Qu.:100.0   Good  : 85   1st Qu.:39.75   1st Qu.:12.0  
##  Median :272.0   Median :117.0   Medium:219   Median :54.50   Median :14.0  
##  Mean   :264.8   Mean   :115.8                Mean   :53.32   Mean   :13.9  
##  3rd Qu.:398.5   3rd Qu.:131.0                3rd Qu.:66.00   3rd Qu.:16.0  
##  Max.   :509.0   Max.   :191.0                Max.   :80.00   Max.   :18.0  
##  Urban       US     
##  No :118   No :142  
##  Yes:282   Yes:258  
##                     
##                     
##                     
##                     
```

```
str(data) # Structure of the data
```

```
## '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 ...
```

```
mean(data$Sales) # Mean of the Sales
```

```
## [1] 7.496325
```

```
sd(data$Sales) # Standard deviation of the Sales
```

```
## [1] 2.824115
```

```
table(data$ShelveLoc) # Table of the ShelveLoc
```

```
##
##    Bad   Good Medium
##     96     85    219
```

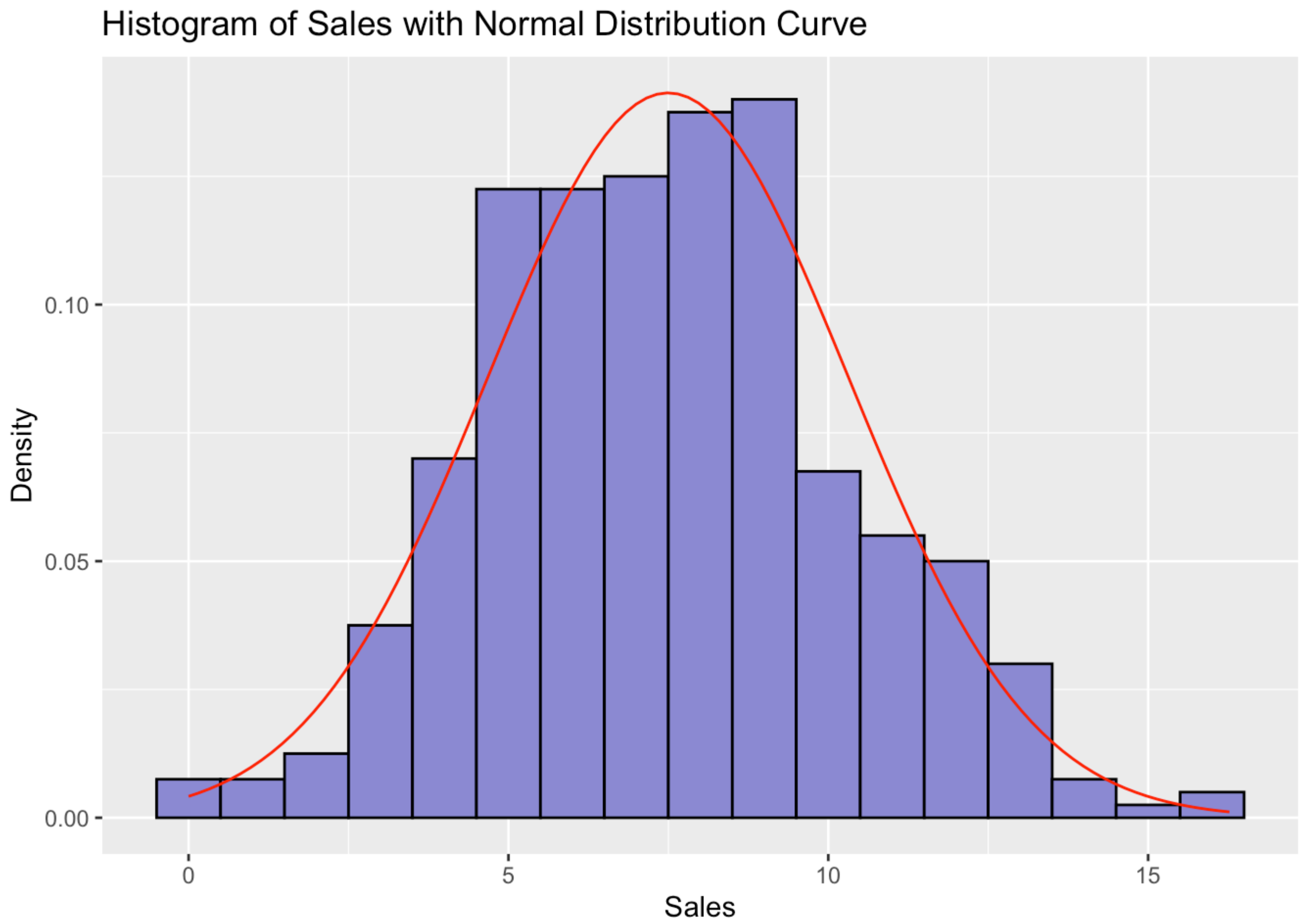
```
table(Carseats$Urban)   # Table of the Urban
```

```
##
##  No Yes
## 118 282
```

Sales

```
# Histogram of sales
ggplot(data, aes(x = Sales)) +
  geom_histogram(aes(y = ..density..), binwidth = 1, fill = "#8d8dd5", color = "blac
k") +
  stat_function(fun = dnorm, args = list(mean = mean(data$Sales), sd = sd(data$Sale
s)), color = "red") +
  labs(title = "Histogram of Sales with Normal Distribution Curve", x = "Sales", y =
"Density")
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## ℹ Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



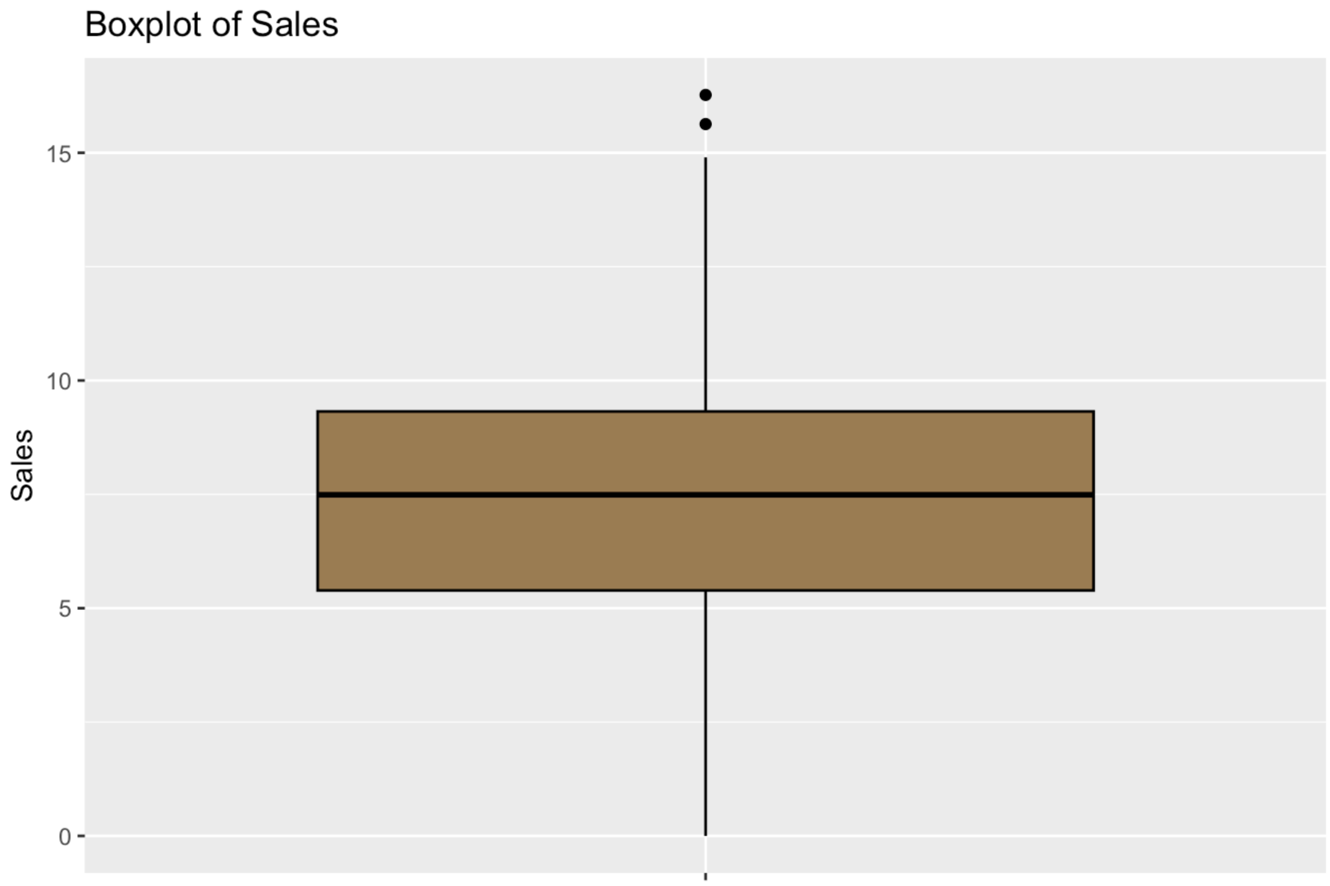
H_0: The Sales are normally distributed H_1: The Sales are not normally distributed

```
shapiro.test(data$Sales)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Sales
## W = 0.9952, p-value = 0.254
```

p-value = 0.254 > 0.05, we fail to reject the null hypothesis. **The Sales are normally distributed.**

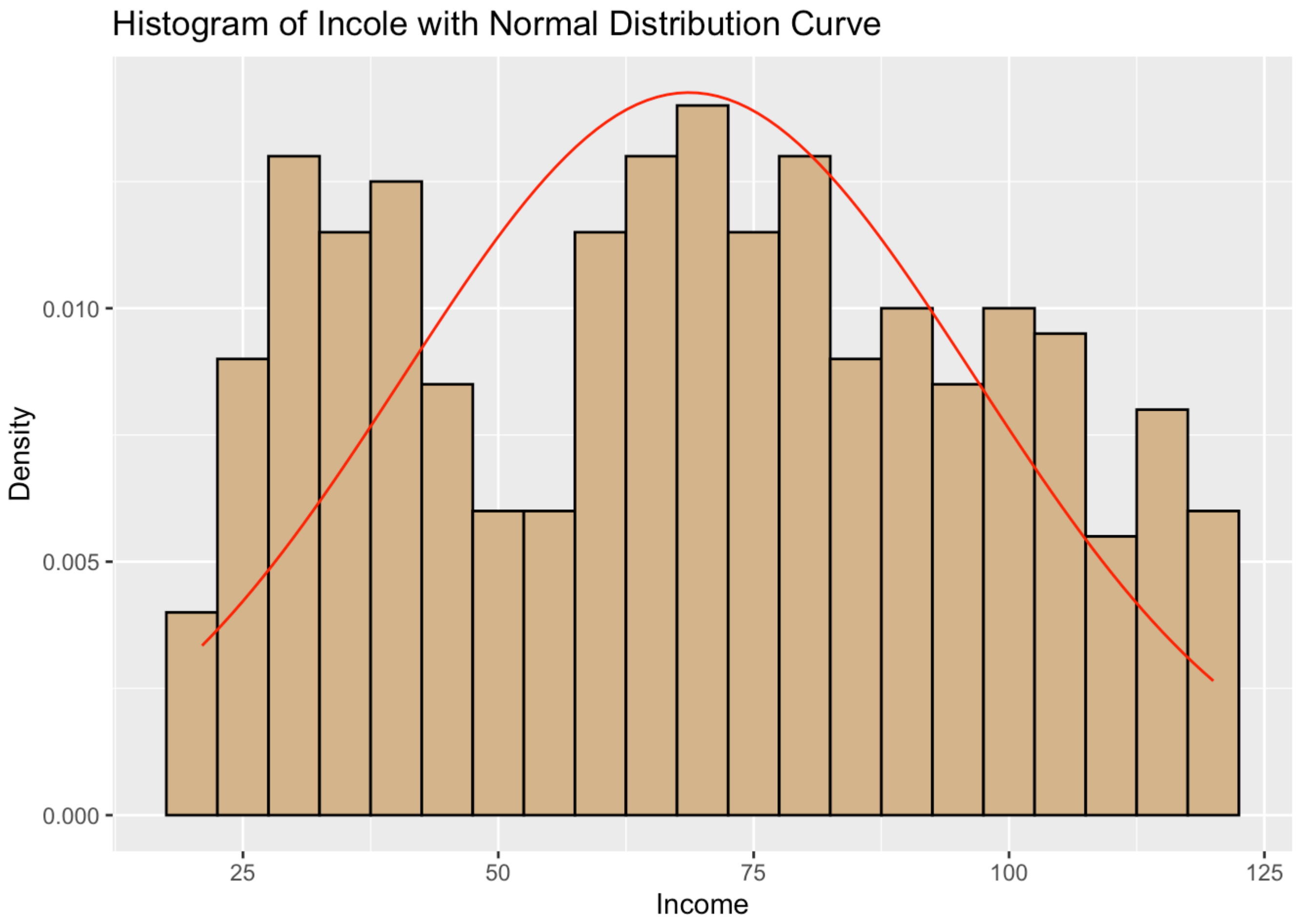
```
# Boxplot of Sales
ggplot(data, aes(x = "", y = Sales)) +
  geom_boxplot(fill = "#9c7d57", color = "black") +
  labs(title = "Boxplot of Sales", x = "", y = "Sales")
```



From boxplot we can conclude: - Data is not skewed, since the median is in the middle of the box - Data spread is uniform, since the box is uniform - Few outliers are present, since there are points outside the whiskers, but only two. So, the data is not heavily skewed.

Income

```
# Histogram of income
ggplot(data, aes(x = Income)) +
  geom_histogram(aes(y = ..density..), binwidth = 5, fill = "#d5b58d", color = "blac
k") +
  stat_function(fun = dnorm, args = list(mean = mean(data$Income), sd = sd(data$Incom
e)), color = "red") +
  labs(title = "Histogram of Incole with Normal Distribution Curve", x = "Income", y
= "Density")
```



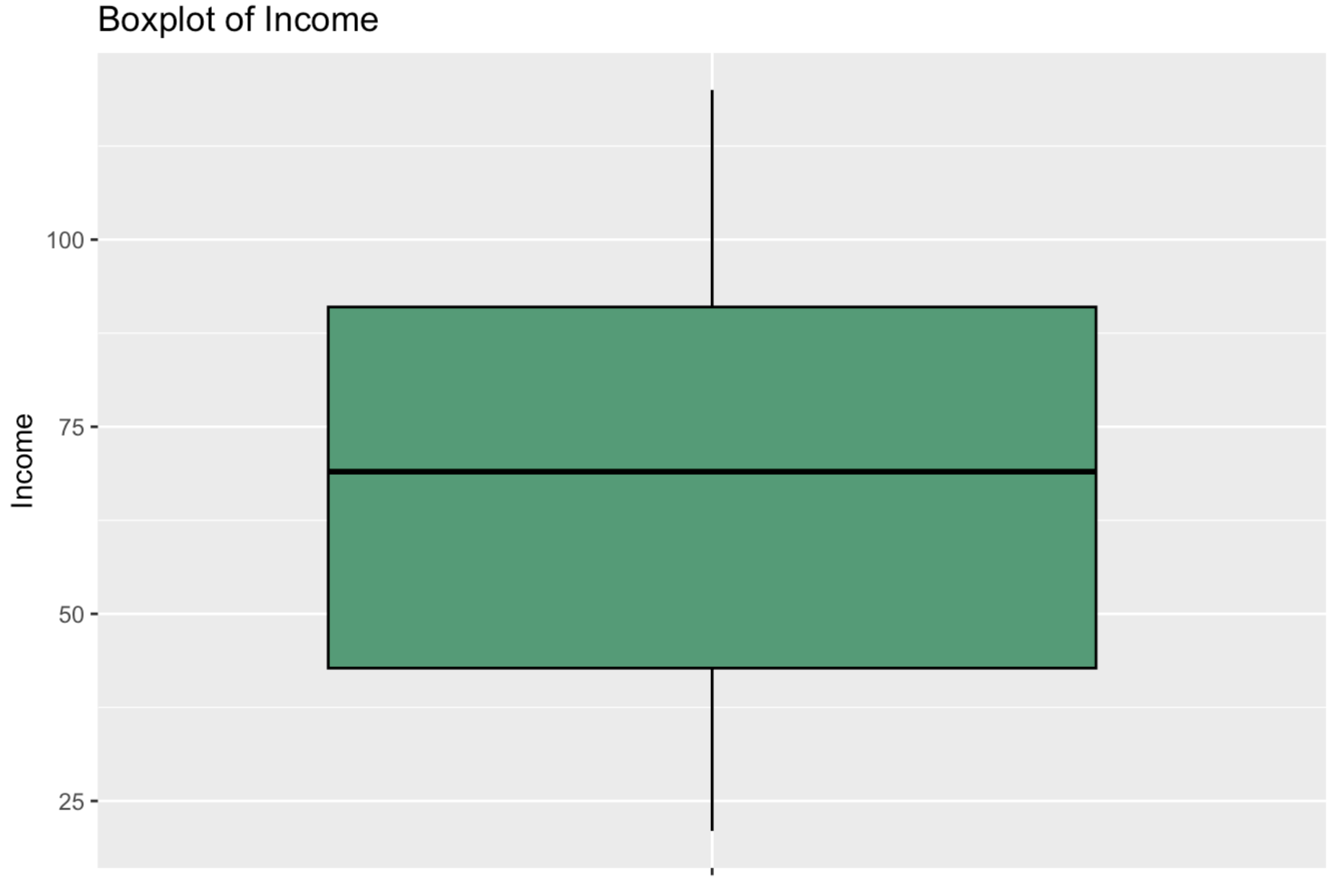
H_0: The Income is normally distributed H_1: The Income is not normally distributed

```
shapiro.test(data$Income)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Income
## W = 0.9611, p-value = 8.396e-09
```

p-value = 0.0008 < 0.05, we reject the null hypothesis. **The Income is not normally distributed.**

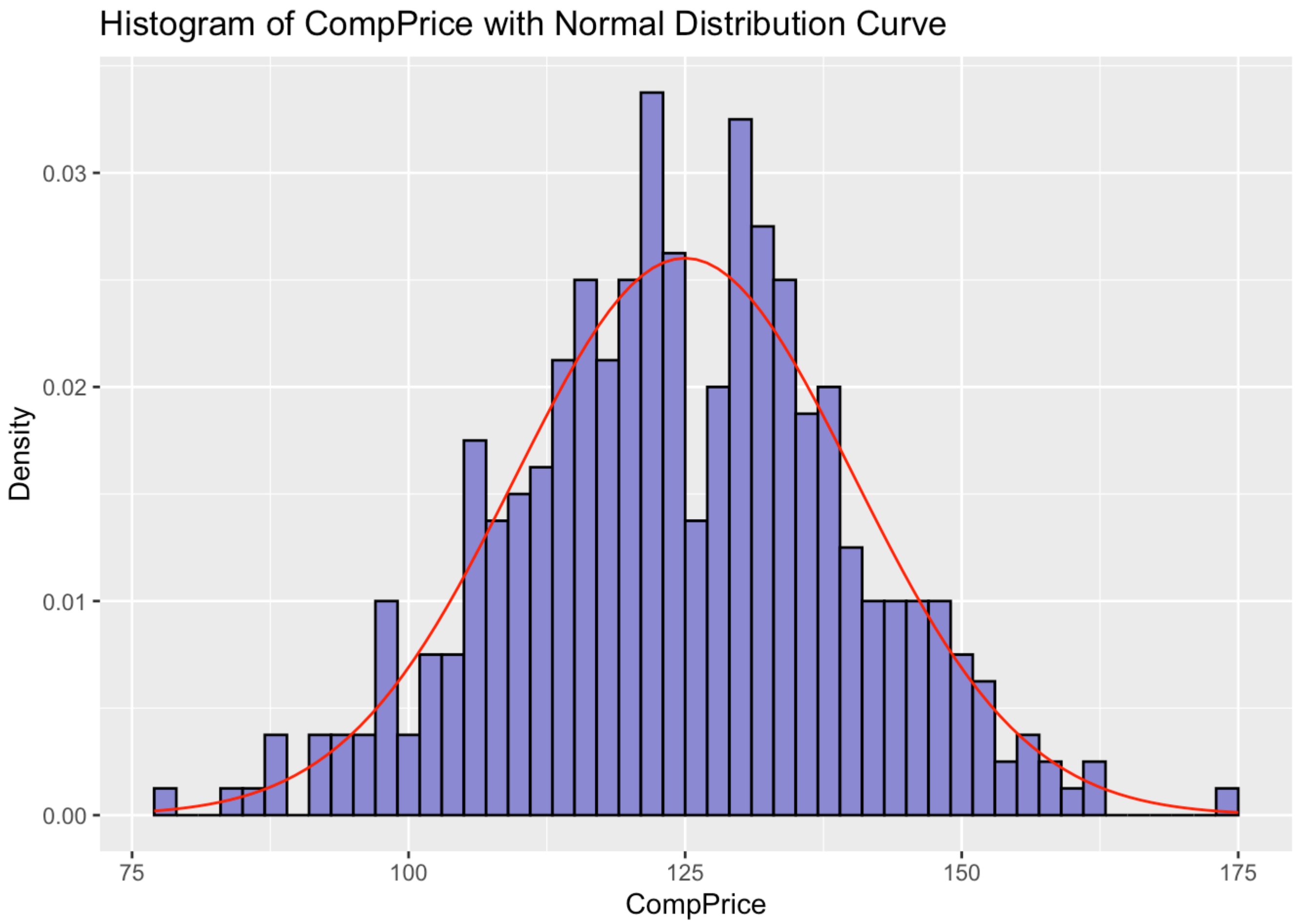
```
# Boxplot of Income
ggplot(data, aes(x = "", y = Income)) +
  geom_boxplot(fill = "#579c77", color = "black") +
  labs(title = "Boxplot of Income", x = "", y = "Income")
```



From boxplot we can conclude: - Data is not skewed, since the median is in the middle of the box - There are no outliers, since there are no points outside the whiskers, so the data is not heavily skewed. - Data spread is uniform, since the box is uniform

Competitors Price

```
# Histogram of compPrice
ggplot(data, aes(x = CompPrice)) +
  geom_histogram(aes(y = ..density..), binwidth = 2, fill = "#8d8dd5", color = "blac
k") +
  stat_function(fun = dnorm, args = list(mean = mean(data$CompPrice), sd = sd(data$Co
mpPrice)), color = "red") +
  labs(title = "Histogram of CompPrice with Normal Distribution Curve", x = "CompPric
e", y = "Density")
```



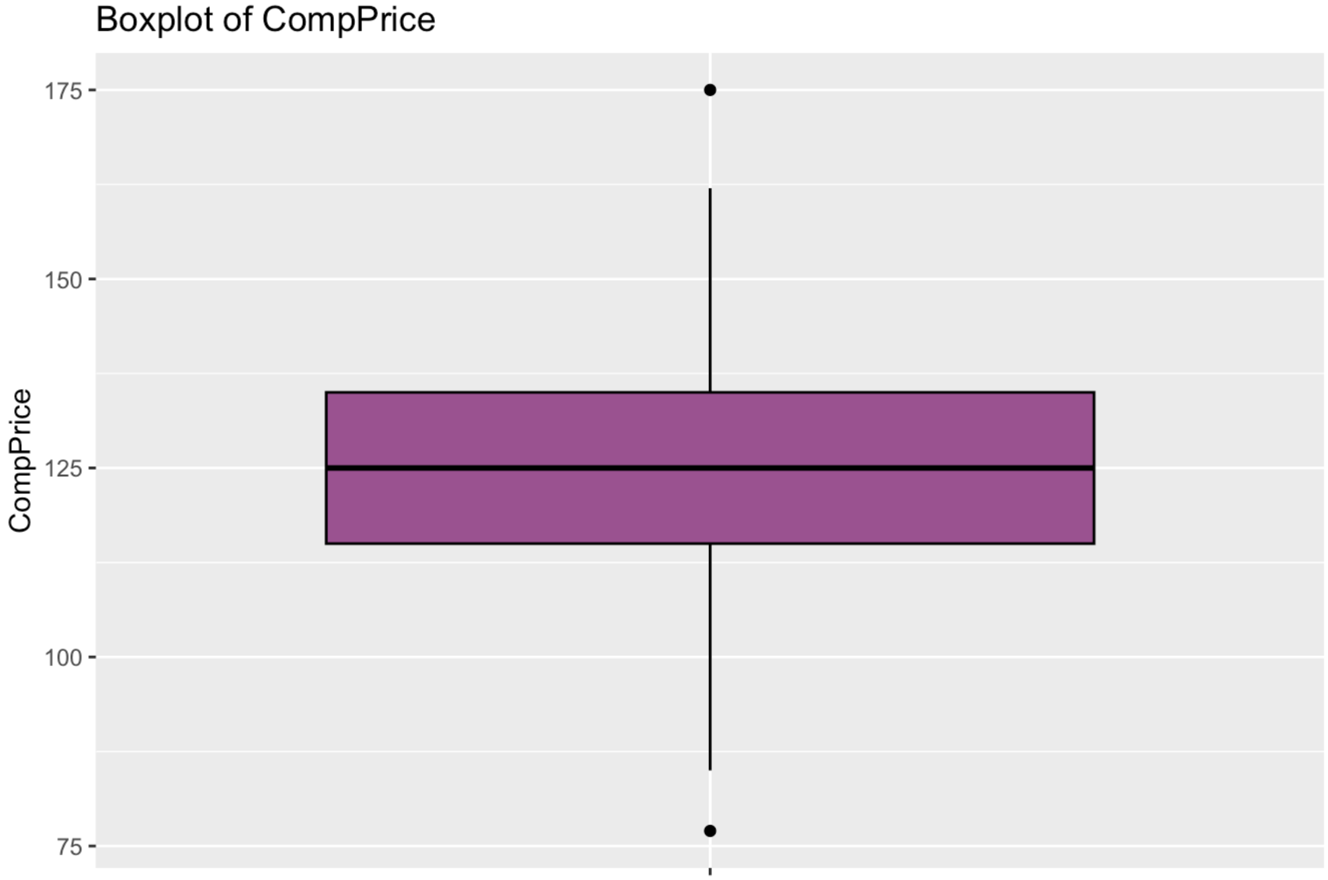
H_0: The CompPrice is normally distributed H_1: The CompPrice is not normally distributed

```
shapiro.test(data$CompPrice)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$CompPrice
## W = 0.99843, p-value = 0.9772
```

p-value = 0.9772 > 0.05, we fail to reject the null hypothesis. **The CompPrice is normally distributed.**

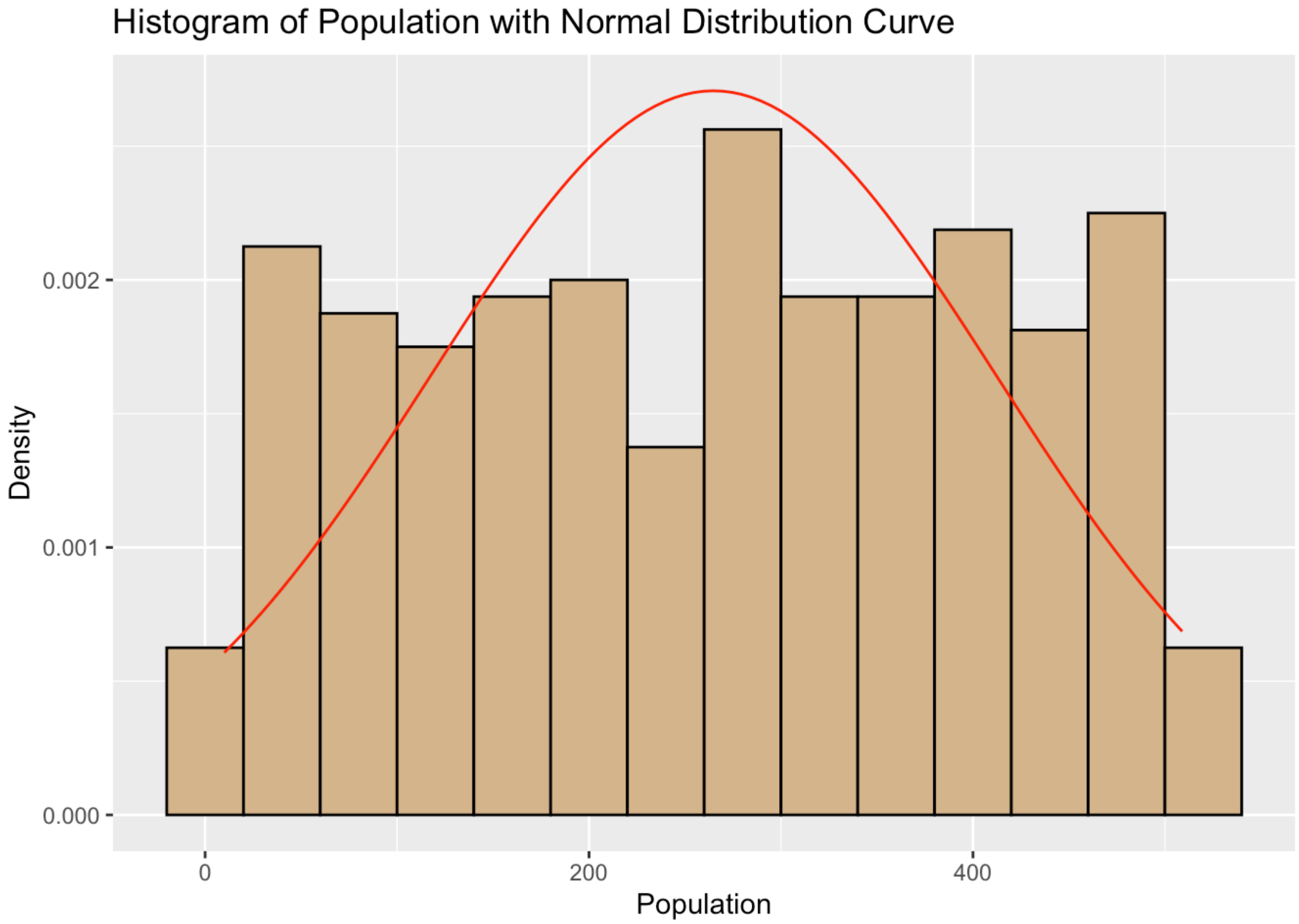
```
# Boxplot of CompPrice
ggplot(data, aes(x = "", y = CompPrice)) +
  geom_boxplot(fill = "#9c5793", color = "black") +
  labs(title = "Boxplot of CompPrice", x = "", y = "CompPrice")
```



From boxplot we can conclude: - Data is not skewed, since the median is in the middle of the box - There are few outliers far away from the whiskers, so the data is not heavily skewed - Data spread is uniform, since the box is uniform

Population

```
# Histogram for population
ggplot(data, aes(x = Population)) +
  geom_histogram(aes(y = ..density..), binwidth = 40, fill = "#d5b58d", color = "blac
k") +
  stat_function(fun = dnorm, args = list(mean = mean(data$Population), sd = sd(data$P
opulation)), color = "red") +
  labs(title = "Histogram of Population with Normal Distribution Curve", x = "Populat
ion", y = "Density")
```



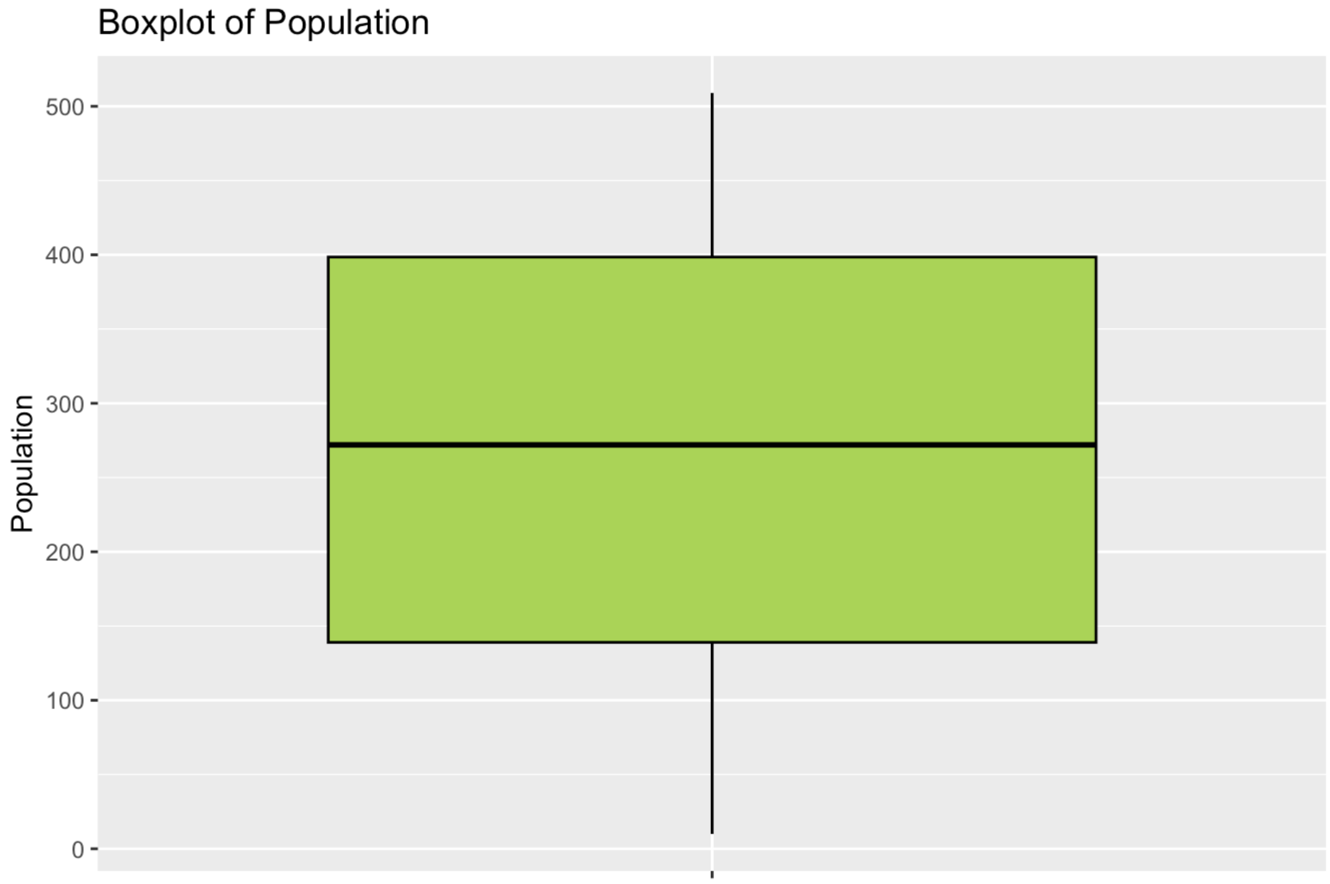
H_0: The Population is normally distributed H_1: The Population is not normally distributed

```
shapiro.test(data$Population)
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  data$Population
## W = 0.95201, p-value = 4.081e-10
```

p-value = 0.0004 < 0.05, we reject the null hypothesis. **The Population is not normally distributed.**

```
# Boxplot of Population
ggplot(data, aes(x = "", y = Population)) +
  geom_boxplot(fill = "#add459", color = "black") +
  labs(title = "Boxplot of Population", x = "", y = "Population")
```

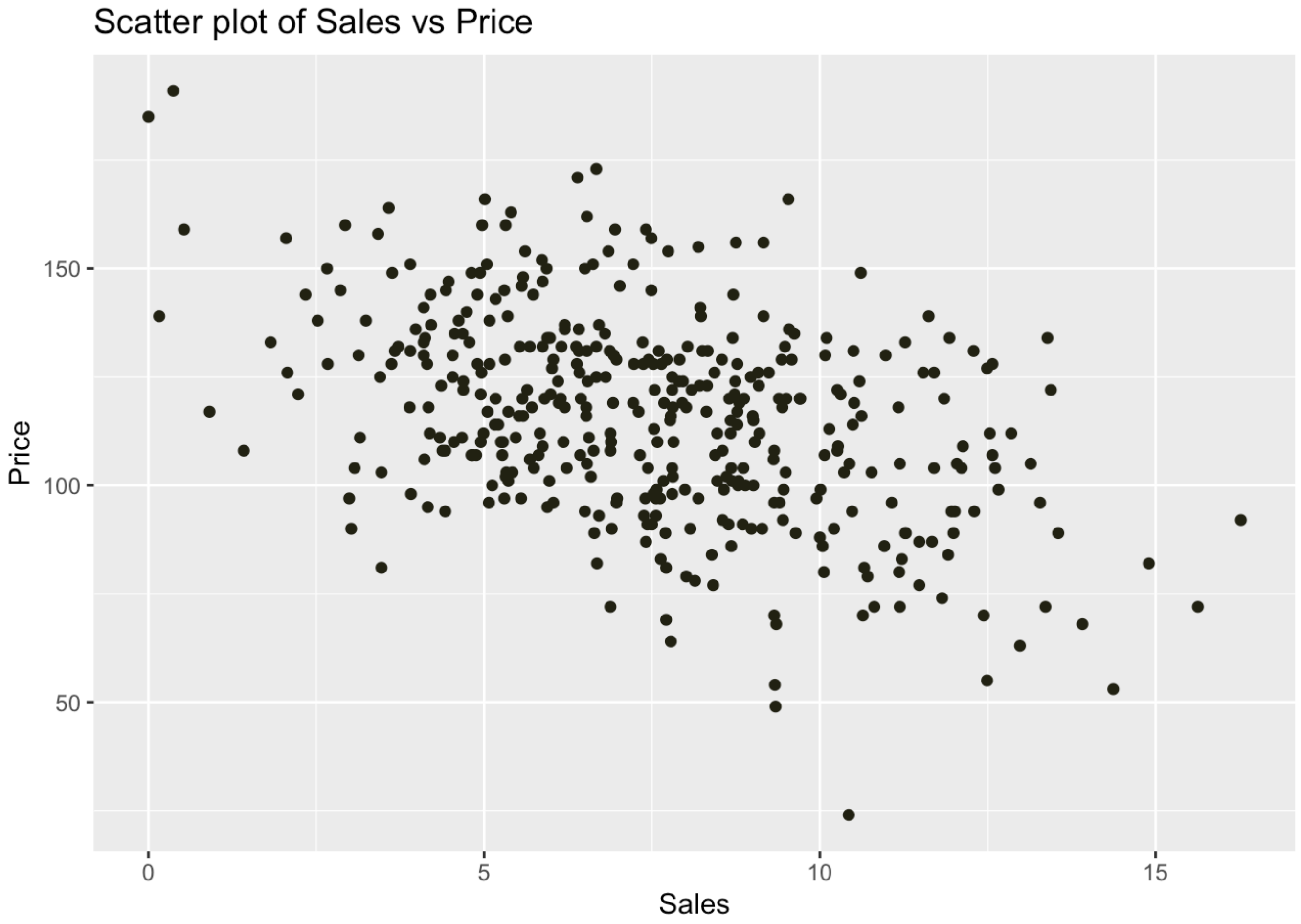


From boxplot we can conclude: - Data is not skewed, since the median is in the middle of the box - There are no outliers, since there are no points outside the whiskers, so the data is not heavily skewed - Data spread is uniform, since the box is uniform

Scatter plots

Sales vs Price

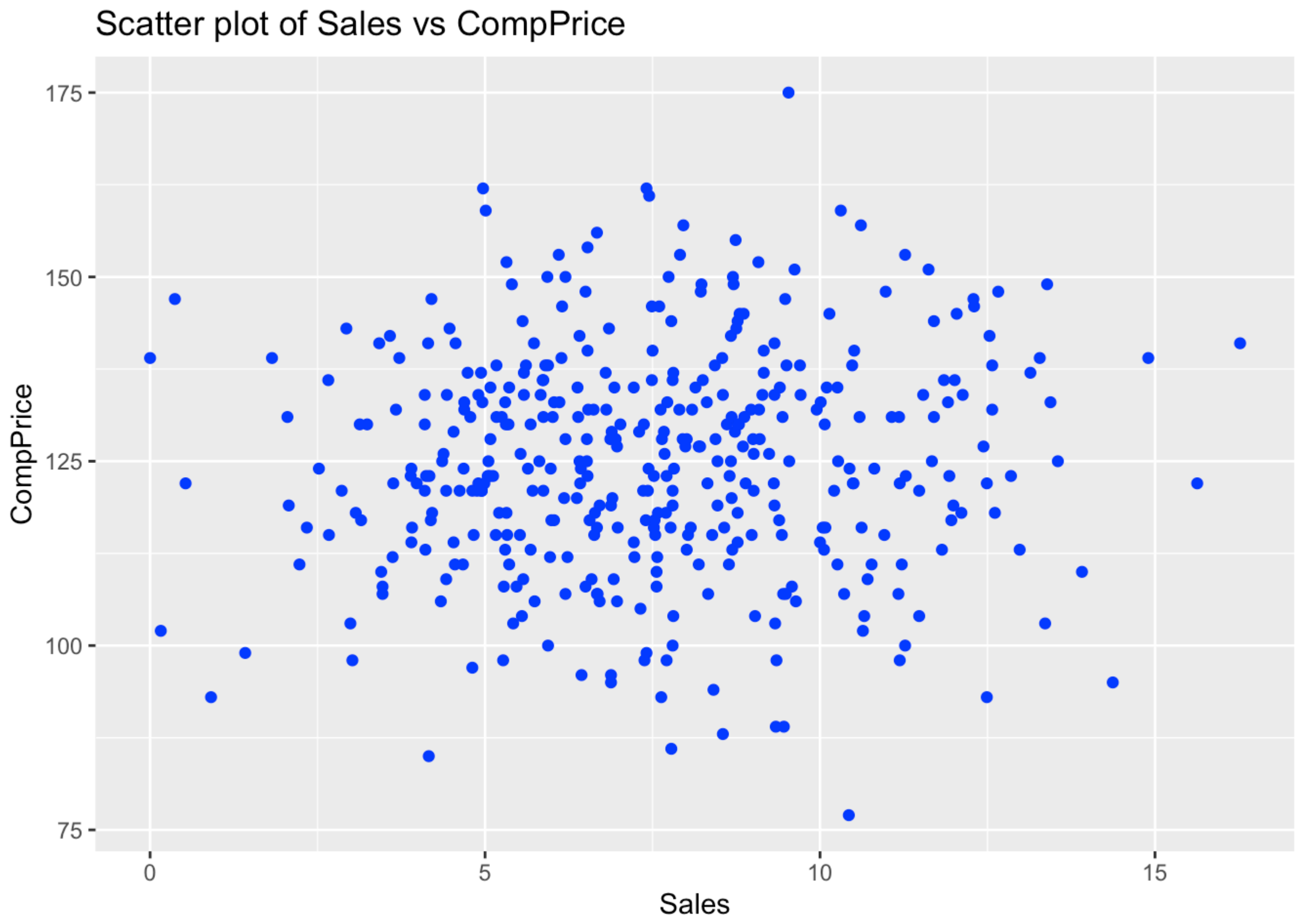
```
ggplot(data, aes(x = Sales, y = Price)) +
  geom_point(color = "#212112") +
  labs(title = "Scatter plot of Sales vs Price", x = "Sales", y = "Price")
```



From scatter plot we can conclude: - Higher prices lead to lower sales (negative correlation) - There are few outliers - Strong relationship between Sales and Price

Sales vs CompPrice

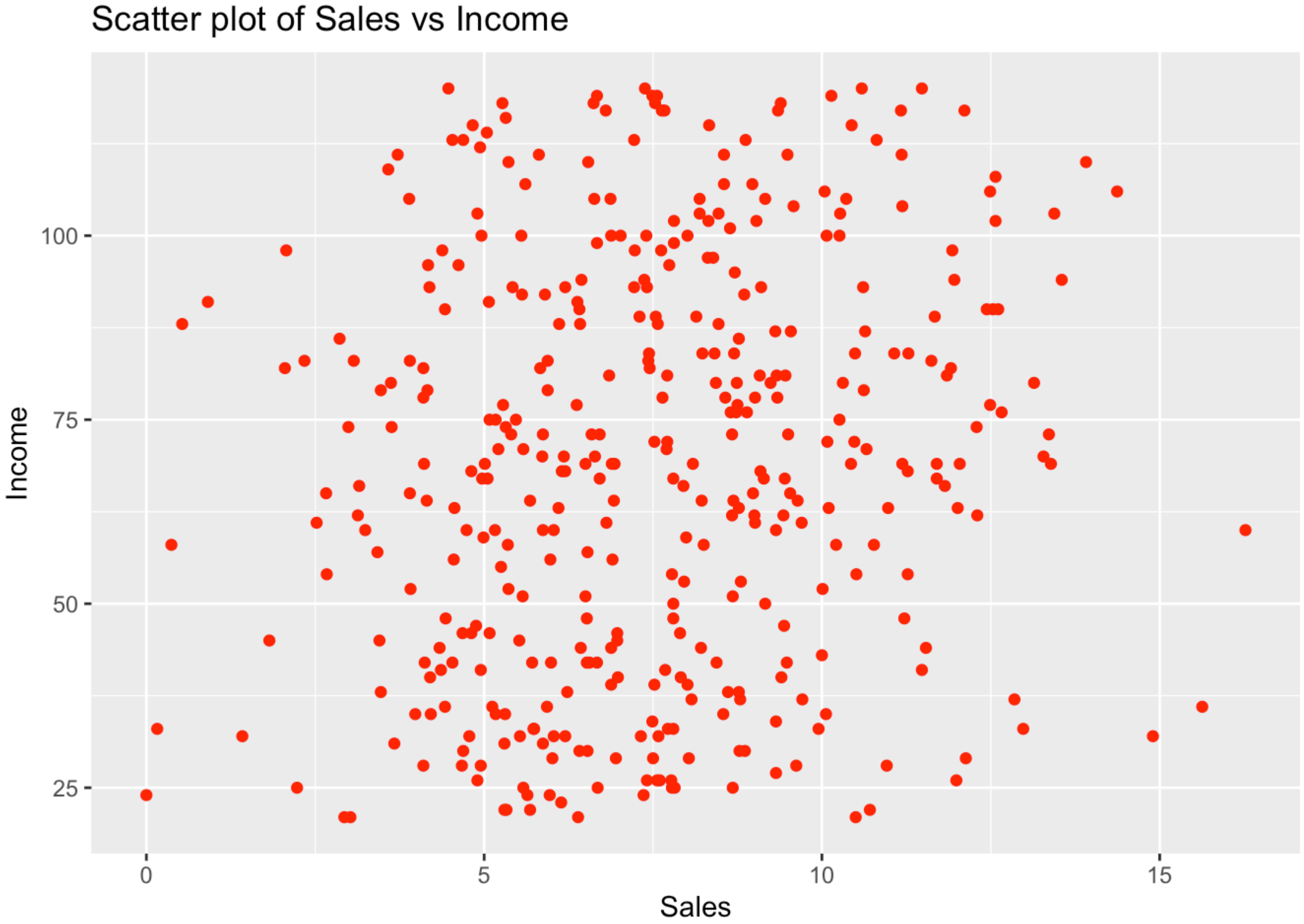
```
ggplot(data, aes(x = Sales, y = CompPrice)) +
  geom_point(color = "#003cff") +
  labs(title = "Scatter plot of Sales vs CompPrice", x = "Sales", y = "CompPrice")
```



From scatter plot we can conclude: - No clear relationship between Sales and CompPrice - Need to check the correlation with a correlation matrix

Sales vs Income

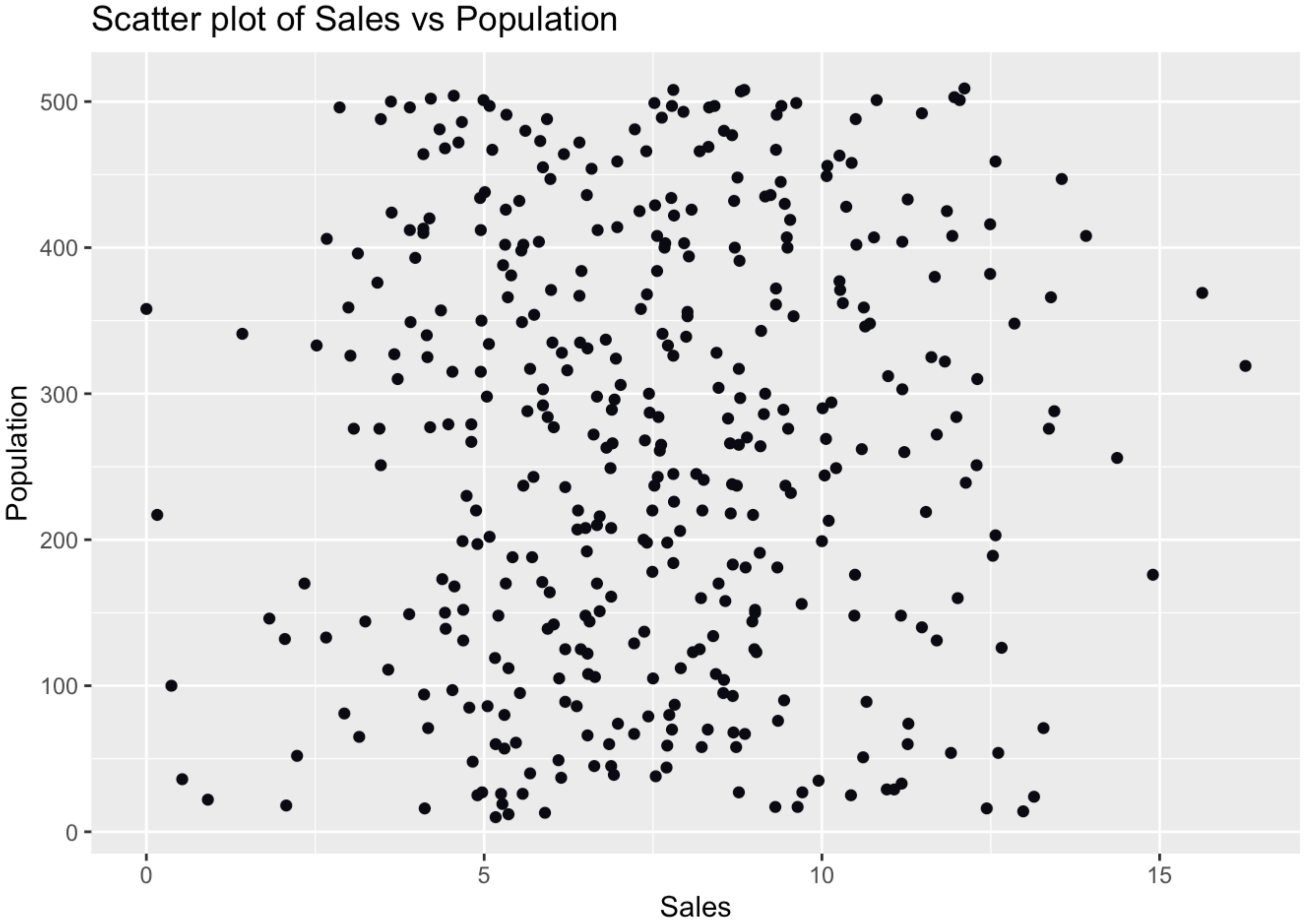
```
ggplot(data, aes(x = Sales, y = Income)) +
  geom_point(color = "#ff0000") +
  labs(title = "Scatter plot of Sales vs Income", x = "Sales", y = "Income")
```



From scatter plot we can conclude: - A weak positive correlation between Sales and Income - There are few outliers

Sales vs Population

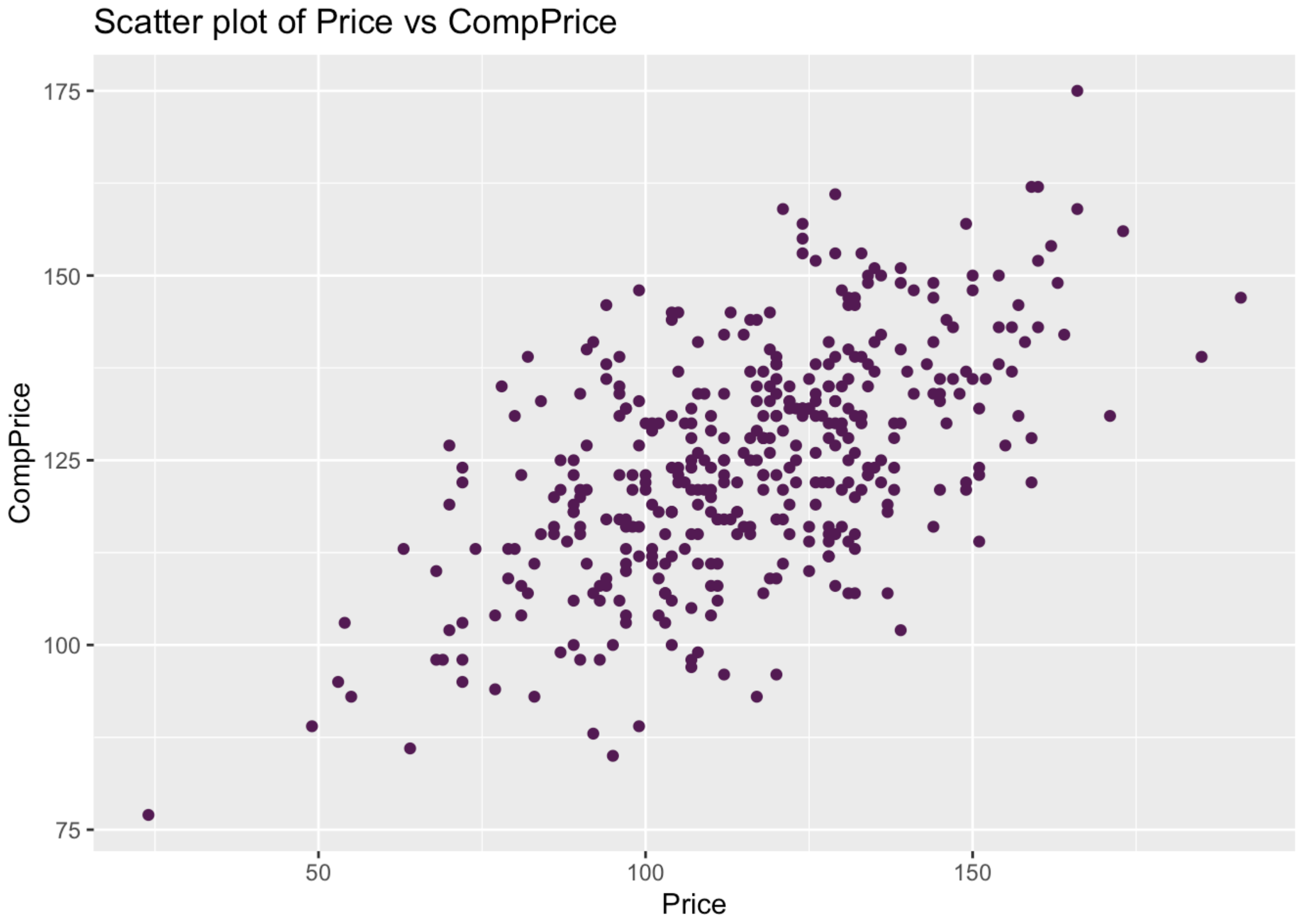
```
ggplot(data, aes(x = Sales, y = Population)) +
  geom_point(color = "#0f0c15") +
  labs(title = "Scatter plot of Sales vs Population", x = "Sales", y = "Population")
```



From scatter plot we can conclude: - Weak positive correlation between Sales and Population - Few outliers present

Price vs CompPrice

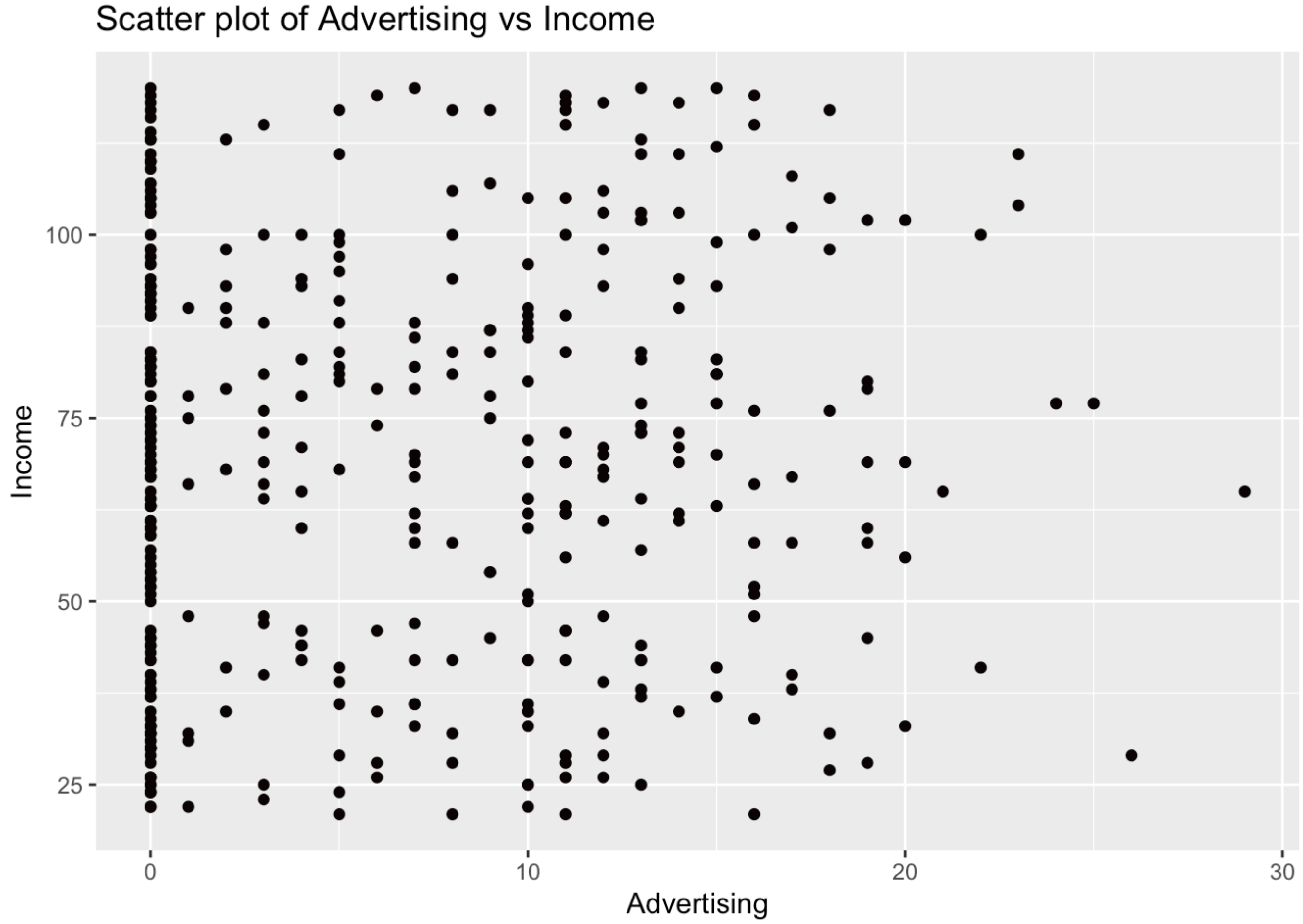
```
ggplot(data, aes(x = Price, y = CompPrice)) +
  geom_point(color = "#531f53") +
  labs(title = "Scatter plot of Price vs CompPrice", x = "Price", y = "CompPrice")
```



From scatter plot we can conclude: - There is a strong positive correlation between Price and CompPrice - Few outliers present - Higher prices are associated with higher CompPrice

Advertising vs Income

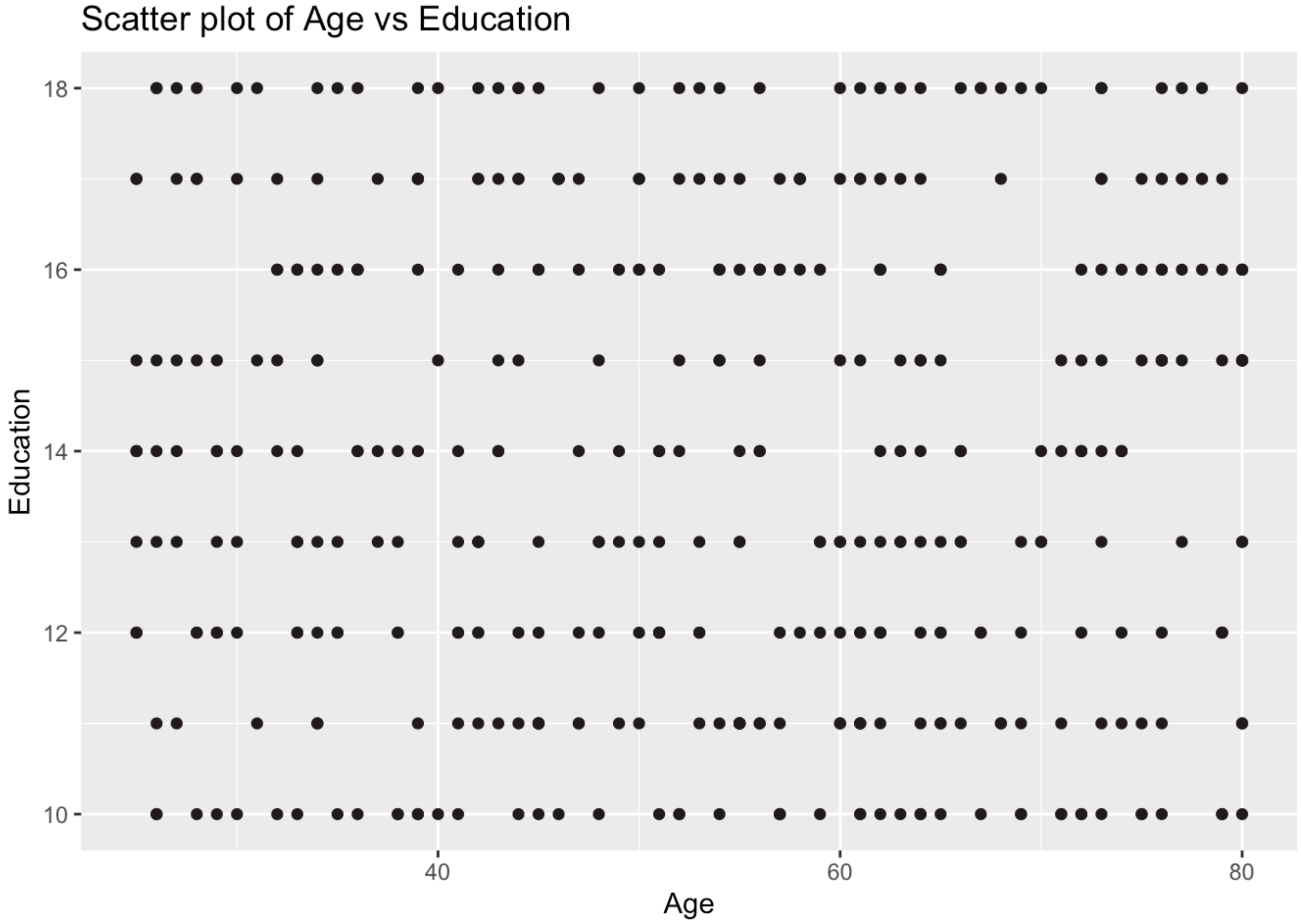
```
ggplot(data, aes(x = Advertising, y = Income)) +
  geom_point(color = "#0a0505") +
  labs(title = "Scatter plot of Advertising vs Income", x = "Advertising", y = "Incom
e")
```



From scatter plot we can conclude: - No clear relationship between Advertising and Income - No advertising also generate low/high income - Few outliers present

Age vs Education

```
ggplot(data, aes(x = Age, y = Education)) +
  geom_point(color = "#241a1a") +
  labs(title = "Scatter plot of Age vs Education", x = "Age", y = "Education")
```



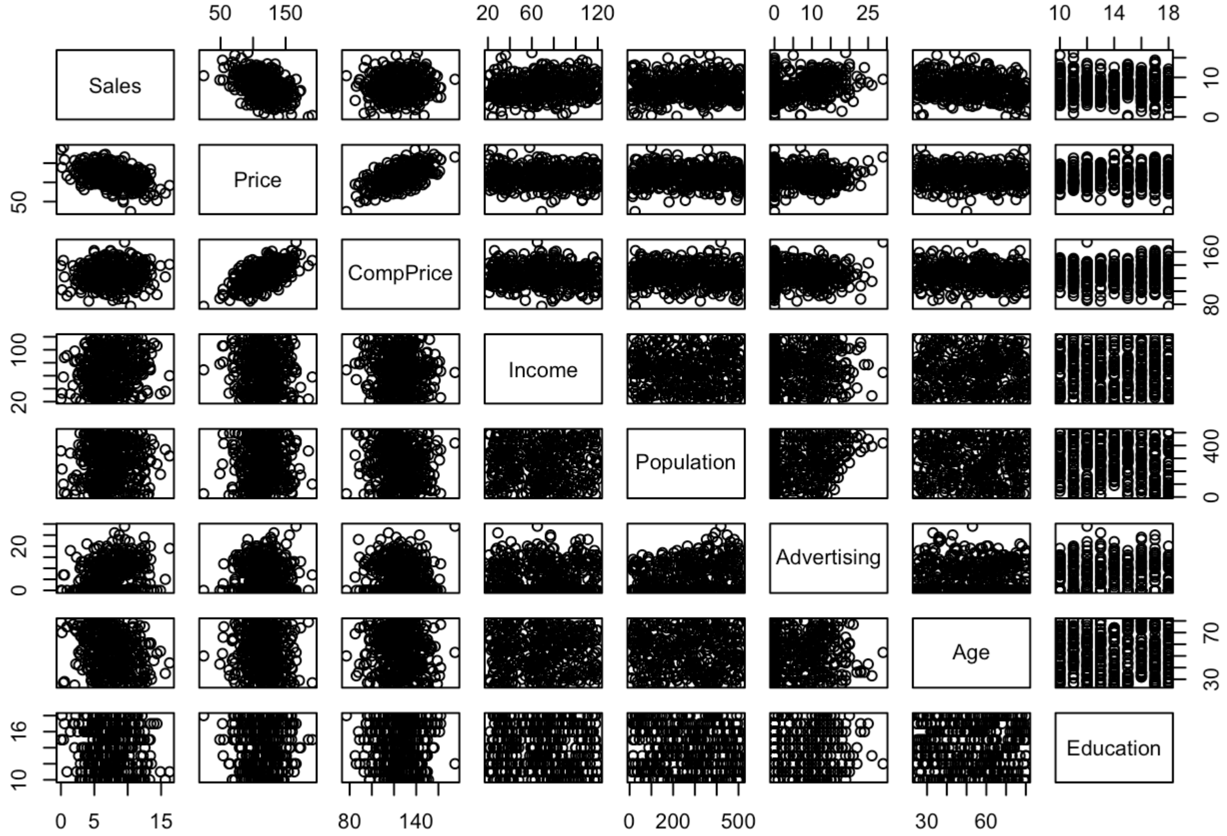
From scatter plot we can conclude: - No clear relationship between Age of customers and their Education level

Correlation matrix

```
cor(data[, c("Sales", "Price", "CompPrice", "Income", "Population", "Advertising", "A
ge", "Education")])
```

```
##                   Sales       Price   CompPrice       Income   Population
## Sales        1.00000000 -0.44495073  0.06407873  0.151950979  0.050470984
## Price       -0.44495073  1.00000000  0.58484777 -0.056698202 -0.012143620
## CompPrice    0.06407873  0.58484777  1.00000000 -0.080653423 -0.094706516
## Income       0.15195098 -0.05669820 -0.08065342  1.000000000 -0.007876994
## Population   0.05047098 -0.01214362 -0.09470652 -0.007876994  1.000000000
## Advertising  0.26950678  0.04453687 -0.02419879  0.058994706  0.265652145
## Age         -0.23181544 -0.10217684 -0.10023882 -0.004670094 -0.042663355
## Education   -0.05195524  0.01174660  0.02519705 -0.056855422 -0.106378231
##              Advertising          Age    Education
## Sales        0.269506781 -0.231815440 -0.051955242
## Price        0.044536874 -0.102176839  0.011746599
## CompPrice   -0.024198788 -0.100238817  0.025197050
## Income       0.058994706 -0.004670094 -0.056855422
## Population   0.265652145 -0.042663355 -0.106378231
## Advertising  1.000000000 -0.004557497 -0.033594307
## Age         -0.004557497  1.000000000  0.006488032
## Education   -0.033594307  0.006488032  1.000000000
```

```
pairs(data[, c("Sales", "Price", "CompPrice", "Income", "Population", "Advertising",
"Age", "Education")])
```



From the correlation matrix and pairs plot we can conclude: - Price and CompPrice have a strong positive correlation - Sales and Price have a strong negative correlation - No other strong correlations are present

3. Task 3

```
# Fit the full model
model <- lm(Sales ~ Price + CompPrice + Income + Advertising + Age + Education + Shel
veLoc + Urban + US, data = Carseats)

# View the summary of the model
summary(model)
```

```
##
## Call:
## lm(formula = Sales ~ Price + CompPrice + Income + Advertising +
##     Age + Education + ShelveLoc + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8799 -0.7015  0.0088  0.6611  3.4268
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.761899   0.575319  10.015  < 2e-16 ***
## Price           -0.095302   0.002667 -35.736  < 2e-16 ***
## CompPrice        0.092609   0.004128  22.436  < 2e-16 ***
## Income           0.015774   0.001843   8.560 2.65e-16 ***
## Advertising      0.125044   0.010558  11.844  < 2e-16 ***
## Age             -0.046119   0.003176 -14.520  < 2e-16 ***
## Education       -0.022411   0.019565  -1.145    0.253
## ShelveLocGood    4.846736   0.152852  31.709  < 2e-16 ***
## ShelveLocMedium  1.952145   0.125732  15.526  < 2e-16 ***
## UrbanYes         0.118853   0.112648   1.055    0.292
## USYes           -0.199075   0.147315  -1.351    0.177
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.018 on 389 degrees of freedom
## Multiple R-squared:  0.8733, Adjusted R-squared:  0.8701
## F-statistic: 268.2 on 10 and 389 DF,  p-value: < 2.2e-16
```

We can make following conclusions from the summary: - Several predictors, including CompPrice, Income, Advertising, Price, ShelveLocGood, ShelveLocMedium, and Age, are significant predictors of Sales, since their p-values are very low - The R-squared value is 0.86, which means that 86% of the variance in Sales is explained by the predictors. - Some predictors, such as Population, Education, UrbanYes, and USYes, are not significant predictors of Sales. Since their p-values are significantly greater then other predictors. - F-statistic p-value is very low, which means that the model is significant and the predictors are significant overall.

4. Task 4

```
# Variance Inflation Factor (VIF) of the model
big_model <- lm(Sales ~ ., data = data)
step_model <- step(big_model)
```

```
## Start:  AIC=26.82
## Sales ~ CompPrice + Income + Advertising + Population + Price + 
##     ShelveLoc + Age + Education + Urban + US
## 
##               Df Sum of Sq     RSS    AIC
## - Population   1      0.33  403.16  25.15
## - Education    1      1.19  404.02  26.00
## - Urban        1      1.23  404.06  26.04
## - US           1      1.57  404.40  26.38
## <none>                      402.83  26.82
## - Income       1     76.16  478.99  94.09
## - Advertising  1    127.14  529.97 134.54
## - Age          1    217.44  620.27 197.48
## - CompPrice    1    519.91  922.74 356.35
## - ShelveLoc    2   1053.20 1456.03 536.80
## - Price        1   1323.23 1726.06 606.85
## 
## Step:  AIC=25.15
## Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc + 
##     Age + Education + Urban + US
## 
##               Df Sum of Sq     RSS    AIC
## - Urban        1      1.15  404.31  24.29
## - Education    1      1.36  404.52  24.49
## - US           1      1.89  405.05  25.02
## <none>                      403.16  25.15
## - Income       1     75.94  479.10  92.18
## - Advertising  1    145.38  548.54 146.32
## - Age          1    218.52  621.68 196.38
## - CompPrice    1    521.69  924.85 355.27
## - ShelveLoc    2   1053.18 1456.34 534.89
## - Price        1   1323.51 1726.67 605.00
## 
## Step:  AIC=24.29
## Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc + 
##     Age + Education + US
## 
##               Df Sum of Sq     RSS    AIC
## - Education    1      1.44  405.76  23.72
## - US           1      1.85  406.16  24.12
## <none>                      404.31  24.29
## - Income       1     76.64  480.96  91.73
## - Advertising  1    146.03  550.34 145.63
## - Age          1    217.59  621.91 194.53
## - CompPrice    1    526.17  930.48 355.69
## - ShelveLoc    2   1053.93 1458.25 533.41
## - Price        1   1322.80 1727.11 603.10
## 
## Step:  AIC=23.72
## Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc + 
##     Age + US
## 
##               Df Sum of Sq     RSS    AIC
## - US           1      1.63  407.39  23.32
## <none>                      405.76  23.72
## - Income       1     77.87  483.62  91.94
## - Advertising  1    145.30  551.06 144.15
## - Age          1    217.97  623.73 193.70
## - CompPrice    1    525.25  931.00 353.92
## - ShelveLoc    2   1056.88 1462.64 532.61
## - Price        1   1322.83 1728.58 601.44
##
## Step:  AIC=23.32
## Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
##     Age
##
##               Df Sum of Sq     RSS    AIC
## <none>                      407.39  23.32
## - Income       1     76.68  484.07  90.30
## - Age          1    219.12  626.51 193.48
## - Advertising  1    234.03  641.42 202.89
## - CompPrice    1    523.83  931.22 352.01
## - ShelveLoc    2   1055.51 1462.90 530.68
## - Price        1   1324.42 1731.81 600.18
```

```
vif(model)  # VIF for full model
```

```
##                 GVIF Df GVIF^(1/(2*Df))
## Price       1.534936  1        1.238925
## CompPrice   1.542439  1        1.241950
## Income      1.023933  1        1.011896
## Advertising 1.898021  1        1.377687
## Age         1.019294  1        1.009601
## Education   1.011972  1        1.005968
## ShelveLoc   1.029565  2        1.007311
## Urban       1.018565  1        1.009240
## US          1.917834  1        1.384859
```

```
vif(step_model)  # VIF for refined model
```

```
##                 GVIF Df GVIF^(1/(2*Df))
## CompPrice   1.534883  1        1.238904
## Income      1.015448  1        1.007694
## Advertising 1.012935  1        1.006447
## Price       1.534425  1        1.238719
## ShelveLoc   1.015139  2        1.003763
## Age         1.016830  1        1.008380
```

For the both the full and refined models, the VIF values are bellow 5 suggesting that there is no multicollinearity in the model.

5. Task 5

```
big_model <- lm(Sales ~ ., data = data)
step_model <- step(big_model)
```

```
## Start:  AIC=26.82
## Sales ~ CompPrice + Income + Advertising + Population + Price + 
##     ShelveLoc + Age + Education + Urban + US
## 
##               Df Sum of Sq     RSS    AIC
## - Population   1      0.33  403.16  25.15
## - Education    1      1.19  404.02  26.00
## - Urban        1      1.23  404.06  26.04
## - US           1      1.57  404.40  26.38
## <none>                      402.83  26.82
## - Income       1     76.16  478.99  94.09
## - Advertising  1    127.14  529.97 134.54
## - Age          1    217.44  620.27 197.48
## - CompPrice    1    519.91  922.74 356.35
## - ShelveLoc    2   1053.20 1456.03 536.80
## - Price        1   1323.23 1726.06 606.85
## 
## Step:  AIC=25.15
## Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc + 
##     Age + Education + Urban + US
## 
##               Df Sum of Sq     RSS    AIC
## - Urban        1      1.15  404.31  24.29
## - Education    1      1.36  404.52  24.49
## - US           1      1.89  405.05  25.02
## <none>                      403.16  25.15
## - Income       1     75.94  479.10  92.18
## - Advertising  1    145.38  548.54 146.32
## - Age          1    218.52  621.68 196.38
## - CompPrice    1    521.69  924.85 355.27
## - ShelveLoc    2   1053.18 1456.34 534.89
## - Price        1   1323.51 1726.67 605.00
## 
## Step:  AIC=24.29
## Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc + 
##     Age + Education + US
## 
##               Df Sum of Sq     RSS    AIC
## - Education    1      1.44  405.76  23.72
## - US           1      1.85  406.16  24.12
## <none>                      404.31  24.29
## - Income       1     76.64  480.96  91.73
## - Advertising  1    146.03  550.34 145.63
## - Age          1    217.59  621.91 194.53
## - CompPrice    1    526.17  930.48 355.69
## - ShelveLoc    2   1053.93 1458.25 533.41
## - Price        1   1322.80 1727.11 603.10
## 
## Step:  AIC=23.72
## Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc + 
##     Age + US
## 
##               Df Sum of Sq     RSS    AIC
## - US           1      1.63  407.39  23.32
## <none>                      405.76  23.72
## - Income       1     77.87  483.62  91.94
## - Advertising  1    145.30  551.06 144.15
## - Age          1    217.97  623.73 193.70
## - CompPrice    1    525.25  931.00 353.92
## - ShelveLoc    2   1056.88 1462.64 532.61
## - Price        1   1322.83 1728.58 601.44
##
## Step:  AIC=23.32
## Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
##     Age
##
##               Df Sum of Sq     RSS    AIC
## <none>                      407.39  23.32
## - Income       1     76.68  484.07  90.30
## - Age          1    219.12  626.51 193.48
## - Advertising  1    234.03  641.42 202.89
## - CompPrice    1    523.83  931.22 352.01
## - ShelveLoc    2   1055.51 1462.90 530.68
## - Price        1   1324.42 1731.81 600.18
```

```
summary(step_model)
```

```
##
## Call:
## lm(formula = Sales ~ CompPrice + Income + Advertising + Price +
##     ShelveLoc + Age, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7728 -0.6954  0.0282  0.6732  3.3292
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.475226   0.505005   10.84   <2e-16 ***
## CompPrice        0.092571   0.004123   22.45   <2e-16 ***
## Income           0.015785   0.001838    8.59   <2e-16 ***
## Advertising      0.115903   0.007724   15.01   <2e-16 ***
## Price           -0.095319   0.002670  -35.70   <2e-16 ***
## ShelveLocGood    4.835675   0.152499   31.71   <2e-16 ***
## ShelveLocMedium  1.951993   0.125375   15.57   <2e-16 ***
## Age             -0.046128   0.003177  -14.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.019 on 392 degrees of freedom
## Multiple R-squared:  0.872,  Adjusted R-squared:  0.8697
## F-statistic: 381.4 on 7 and 392 DF,  p-value: < 2.2e-16
```

Conclusions: - The basic criteria is AIC (Akaikes). The lower the AIC value, the better the model. - Final AIC is 23.32 which is the best result after gradual elimination of predictors. - Each predictor is highly significant, since the p-values are < 0.01. - Price is only predictor with negative coefficient, which means that higher prices lead to lower sales. - ShelveLoc has two categories: Good and Medium. Both are positive coefficients, indicating that products with better shelf locations tend to have higher sales.

- R-squared = 0.872 and Adjusted R-squared = 0.8697: These values suggest that the model explains approximately 87% of the variance in the sales data. This is a relatively high value, indicating that the model is a good fit for the data.

6. Task 6

```
# Refine the model using stepwise selection
step_model <- step(model)
```

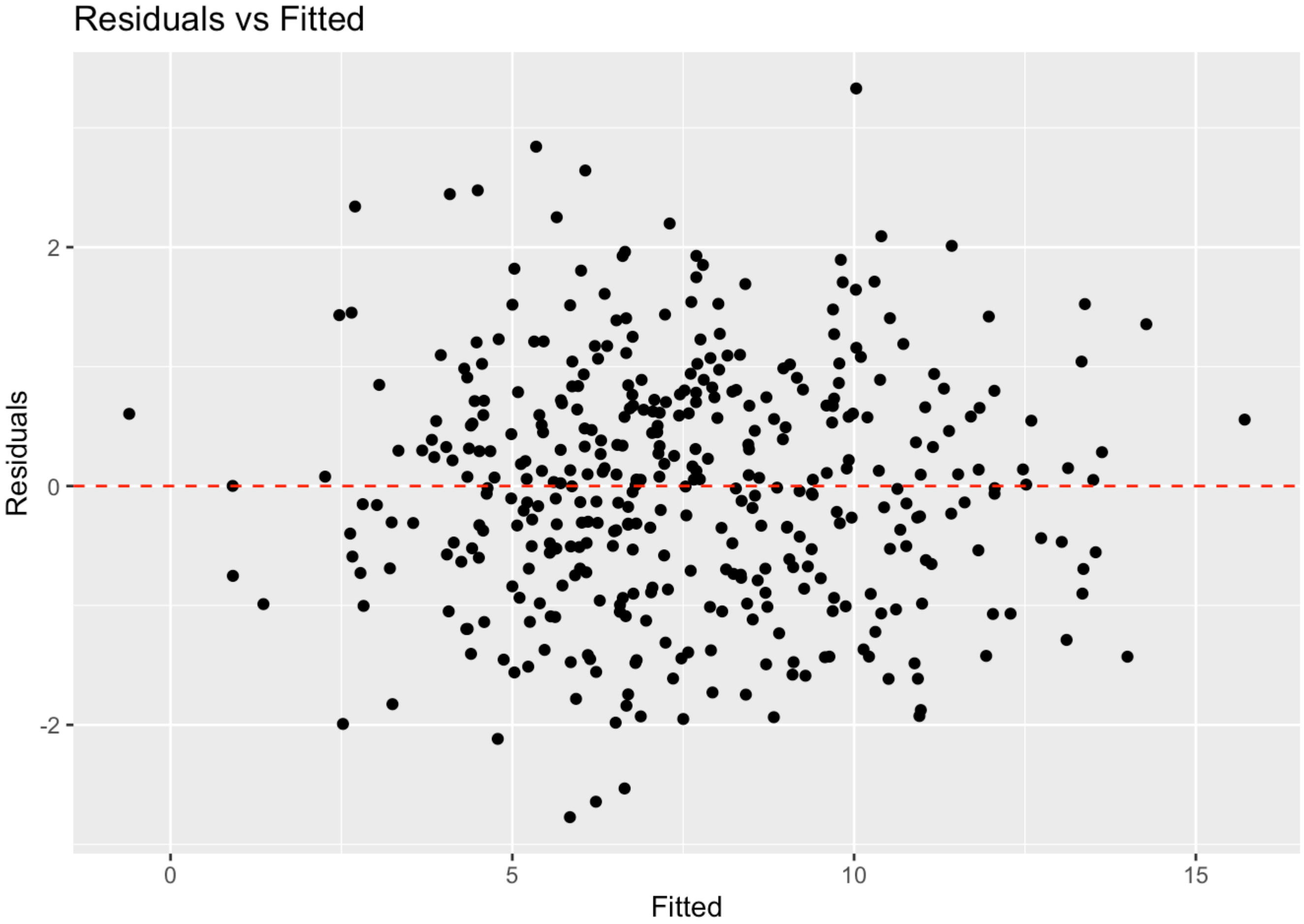
```
## Start:  AIC=25.15
## Sales ~ Price + CompPrice + Income + Advertising + Age + Education + 
##     ShelveLoc + Urban + US
## 
##               Df Sum of Sq     RSS    AIC
## - Urban        1      1.15  404.31  24.29
## - Education    1      1.36  404.52  24.49
## - US           1      1.89  405.05  25.02
## <none>                      403.16  25.15
## - Income       1     75.94  479.10  92.18
## - Advertising  1    145.38  548.54 146.32
## - Age          1    218.52  621.68 196.38
## - CompPrice    1    521.69  924.85 355.27
## - ShelveLoc    2   1053.18 1456.34 534.89
## - Price        1   1323.51 1726.67 605.00
## 
## Step:  AIC=24.29
## Sales ~ Price + CompPrice + Income + Advertising + Age + Education + 
##     ShelveLoc + US
## 
##               Df Sum of Sq     RSS    AIC
## - Education    1      1.44  405.76  23.72
## - US           1      1.85  406.16  24.12
## <none>                      404.31  24.29
## - Income       1     76.64  480.96  91.73
## - Advertising  1    146.03  550.34 145.63
## - Age          1    217.59  621.91 194.53
## - CompPrice    1    526.17  930.48 355.69
## - ShelveLoc    2   1053.93 1458.25 533.41
## - Price        1   1322.80 1727.11 603.10
## 
## Step:  AIC=23.72
## Sales ~ Price + CompPrice + Income + Advertising + Age + ShelveLoc + 
##     US
## 
##               Df Sum of Sq     RSS    AIC
## - US           1      1.63  407.39  23.32
## <none>                      405.76  23.72
## - Income       1     77.87  483.62  91.94
## - Advertising  1    145.30  551.06 144.15
## - Age          1    217.97  623.73 193.70
## - CompPrice    1    525.25  931.00 353.92
## - ShelveLoc    2   1056.88 1462.64 532.61
## - Price        1   1322.83 1728.58 601.44
## 
## Step:  AIC=23.32
## Sales ~ Price + CompPrice + Income + Advertising + Age + ShelveLoc
## 
##               Df Sum of Sq     RSS    AIC
## <none>                      407.39  23.32
## - Income       1     76.68  484.07  90.30
## - Age          1    219.12  626.51 193.48
## - Advertising  1    234.03  641.42 202.89
## - CompPrice    1    523.83  931.22 352.01
## - ShelveLoc    2   1055.51 1462.90 530.68
## - Price        1   1324.42 1731.81 600.18
```

```
summary(step_model)
```

```
##
## Call:
## lm(formula = Sales ~ Price + CompPrice + Income + Advertising +
##     Age + ShelveLoc, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7728 -0.6954  0.0282  0.6732  3.3292
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.475226   0.505005   10.84   <2e-16 ***
## Price           -0.095319   0.002670  -35.70   <2e-16 ***
## CompPrice        0.092571   0.004123   22.45   <2e-16 ***
## Income           0.015785   0.001838    8.59   <2e-16 ***
## Advertising      0.115903   0.007724   15.01   <2e-16 ***
## Age             -0.046128   0.003177  -14.52   <2e-16 ***
## ShelveLocGood    4.835675   0.152499   31.71   <2e-16 ***
## ShelveLocMedium  1.951993   0.125375   15.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.019 on 392 degrees of freedom
## Multiple R-squared:  0.872,  Adjusted R-squared:  0.8697
## F-statistic: 381.4 on 7 and 392 DF,  p-value: < 2.2e-16
```

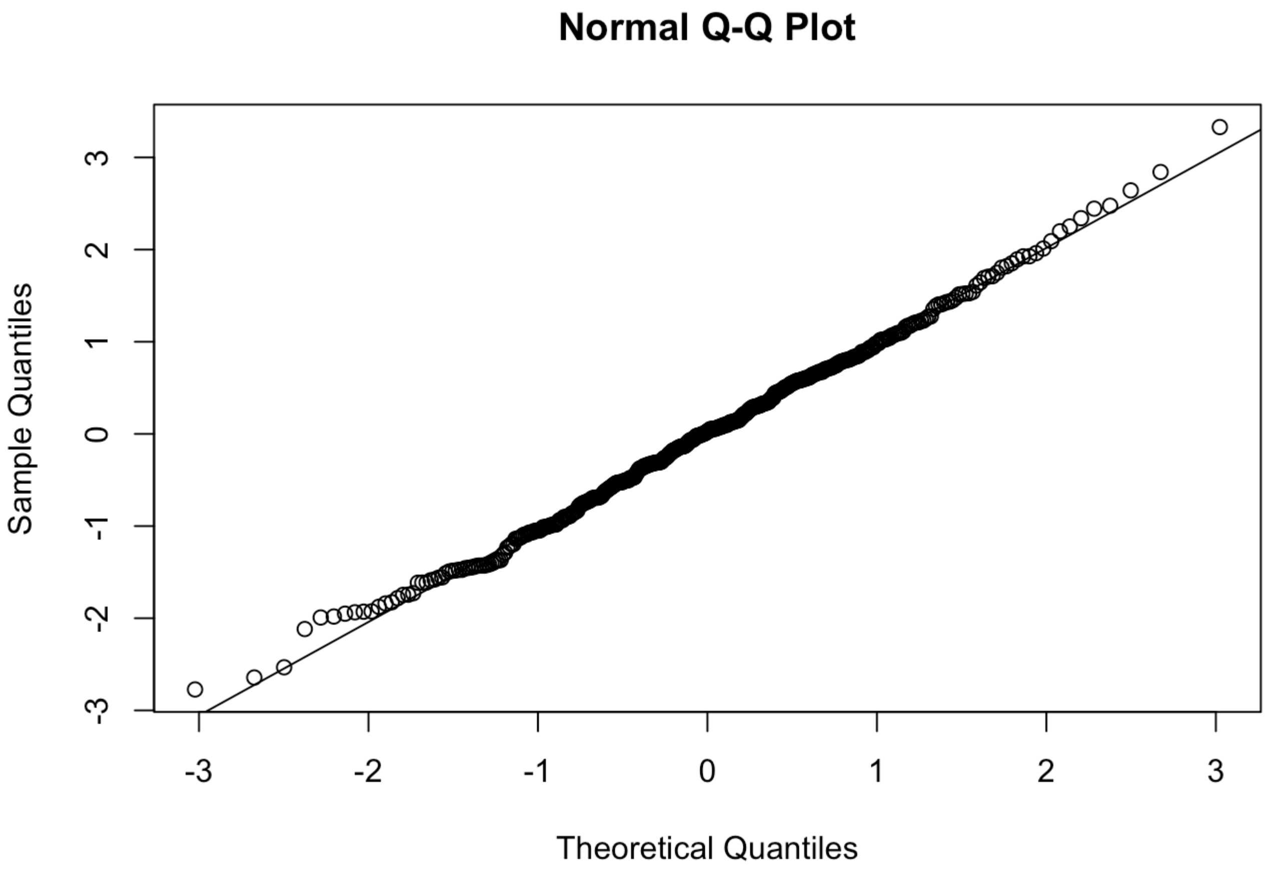
```
# 1. Residuals vs Fitted
residuals_data <- data.frame(
    Fitted = fitted(step_model),  # Get fitted values
    Residuals = residuals(step_model)  # Get residuals
)

# Plot Residuals vs Fitted
ggplot(residuals_data, aes(x = Fitted, y = Residuals)) +
    geom_point() +
    geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
    labs(title = "Residuals vs Fitted", x = "Fitted", y = "Residuals")
```

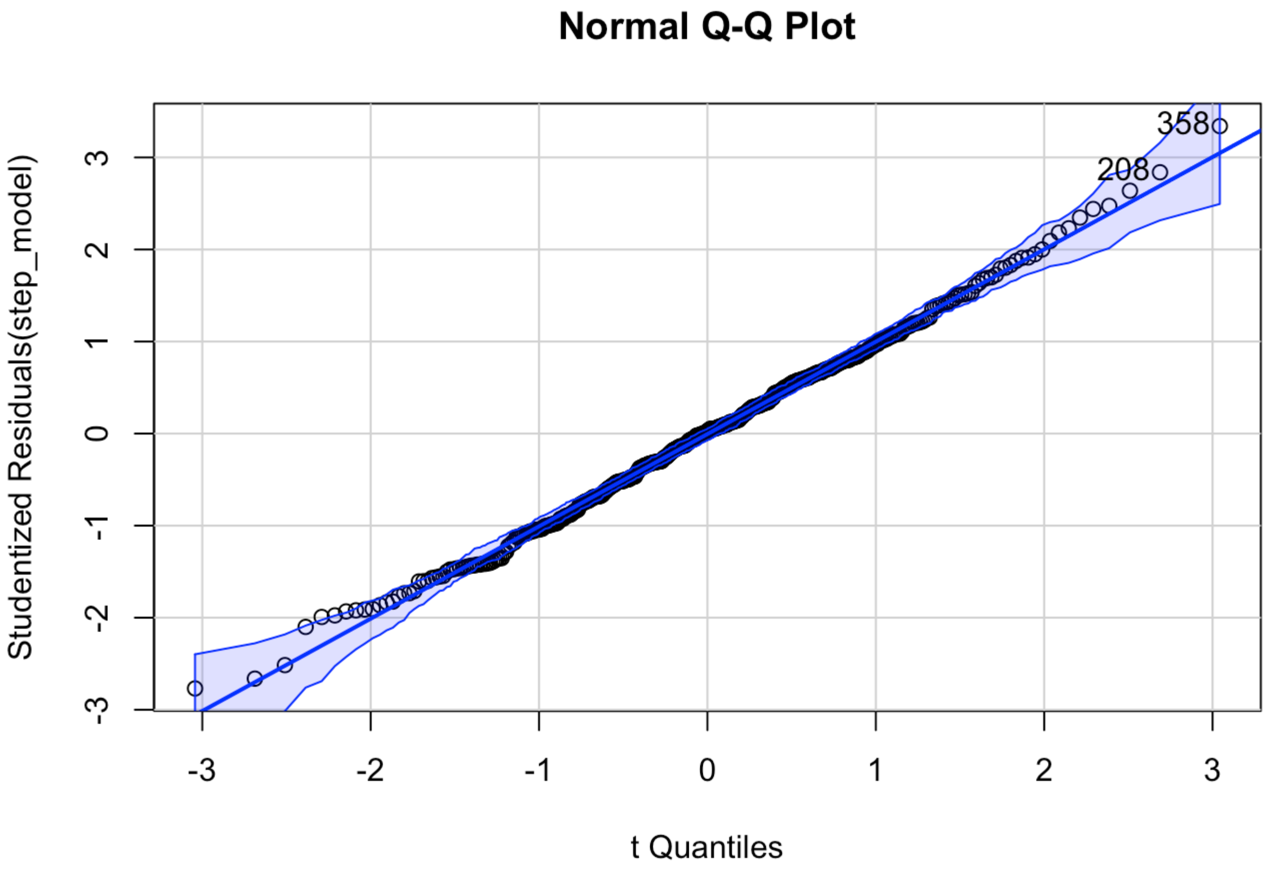


From the residual plot we can conclude: - The residuals are randomly distributed around 0

```
# 2. Normal Q-Q plot
qqnorm(residuals(step_model))  # Normal Q-Q plot
qqline(residuals(step_model))  # Add the line to the plot
```



```
qqPlot(step_model, main = "Normal Q-Q Plot")  # Additional Q-Q plot for verification
```

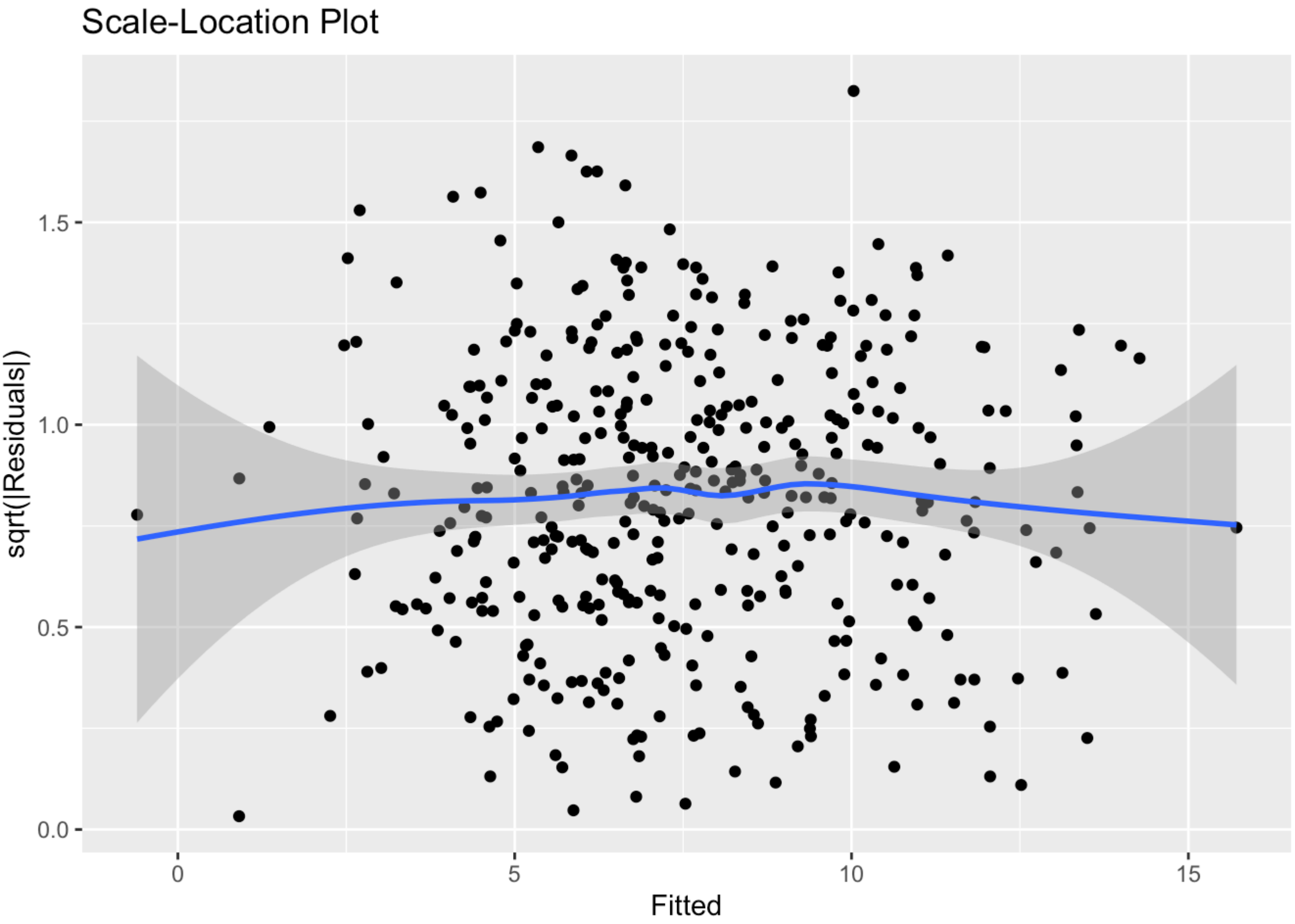


```
## [1] 208 358
```

From the Q-Q plot we can conclude: - The residuals are normally distributed, since the points are close to the line

```
# 3. Scale-Location plot
ggplot(residuals_data, aes(x = Fitted, y = sqrt(abs(Residuals)))) +
    geom_point() +
    geom_smooth() +
    labs(title = "Scale-Location Plot", x = "Fitted", y = "sqrt(|Residuals|)")
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



From the scale-location plot we can conclude: - The residuals are homoscedastic(constant variance), since the points are randomly distributed around the line

H_0: The residuals are homoscedastic

H_1: The residuals are not homoscedastic

```
# 4. Durbin-Watson test for autocorrelation of residuals
dw <- dwtest(step_model)
dw
```

```
## 
##  Durbin-Watson test
## 
## data:  step_model
## DW = 1.9882, p-value = 0.4523
## alternative hypothesis: true autocorrelation is greater than 0
```

p-value = 0.4523 > 0.05, we fail to reject the null hypothesis. **The residuals are not autocorrelated**.