Setting Up R Assignment

Analyze the role of descriptive statistics in data exploration phase of analytics projects.

1. Install the ISLR library using the install.packages() command. Call the library using the library(ISLR) command to ensure that the library is correctly installed.

```
require("ISLR") #loads the required package
```

```
## Loading required package: ISLR
```

```
library("ISLR") #activates the library

#install.packages("ISLR") - Calling the package off since it's already installed.
```

2. Create a new R-Notebook (.Rmd) file. In the first code chunk, call the ISLR library and then print the summary of the Carseats dataset. How many observations (rows) this dataset contains?

```
library("ISLR")
#activates the ISLR Package
summary(Carseats)#prints out the summary of the Carseats dataset which is a part of the ISLR Package
```

```
##      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
##
##
##
##
```

```
nrow(Carseats)#gives us the count of the total rows presented in the Carseats dataset
```

```
## [1] 400
```

3. Using the summary statistics shown above, what is maximum value of the advertising attribute?

```
max(Carseats$Advertising)
```

```
## [1] 29
```

```
#since we have used summary function above using the max() function exclusively to find the max value
```

4. Calculate the IQR of the Price attribute.

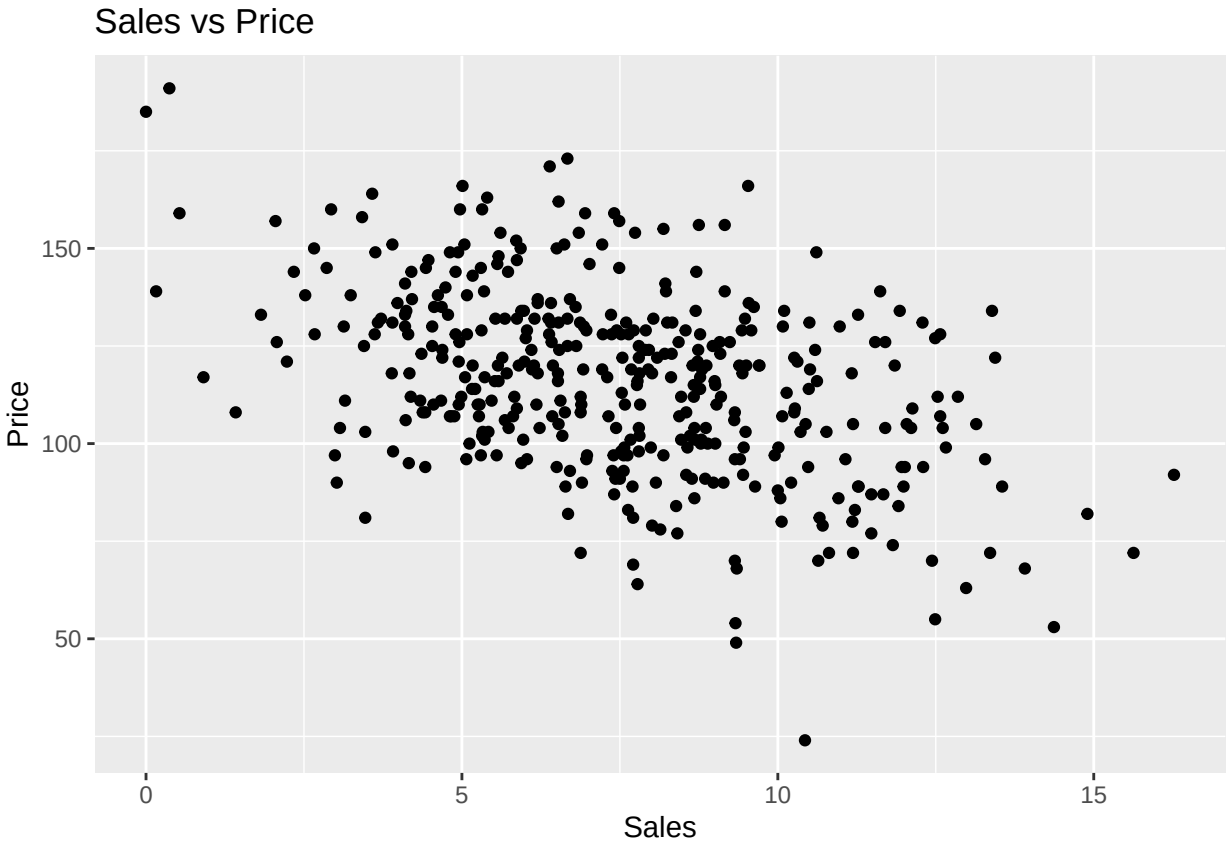
```
IQR(Carseats$Price)
```

```
## [1] 31
```

```
#IQR refers to the inter-quartile range
```

5. Plot the Sales against Price. What do you see in there? Calculate the correlation of the two attributes. What does the sign of the correlation coefficient suggest?

```
#Using qplot function to plot sales vs price
#install.packages("ggplot2") - If the package isn't installed.
library("ggplot2")
#Running the qplot function
qplot(data=Carseats,x=Sales,y=Price, xlab = "Sales", ylab = "Price", main = "Sales vs Price")
```




```
#using cor.test() for calculating the correlation coefficient between Sales and Price.
cor.test(x = Carseats$Sales,y = Carseats$Price,method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  Carseats$Sales and Carseats$Price
## t = -9.912, df = 398, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5203026 -0.3627240
## sample estimates:
##        cor
## -0.4449507
```

Key Findings from the plot Sales vs Price: 1. x and y variables have a negative or a inverse relationship. 2. Strength between the variables - They don't appear to be that strong, if they would have been strong then the points would have been closer helping to form an imaginary line. 3. Relationship - In here it appears that x and y have a linear relationship.

We see that the correlation coefficient value is -0.4449507 which falls under the bucket of -1 indicating that there exists a strong negative correlation or inverse relationship: this means that every time x increases, y decreases and vice-versa.

We are following the "pearson method" which is most widely used. It measures a linear dependence between two variables (x and y). It's also known as a parametric correlation test because it depends to the distribution of the data. Alternatively we can change the method to "kendall" and "spearman" as well.