Adv Data Mining and Analytics - Assignment 1

msasnur@kent.edu

28/02/2020

Initializing all the required libraries.

```
library(ISLR)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

Assigning Carseats Dataset to safeBabies Variable by selecting only Sales, Price, and Shelf Location columns.

```
safeBabies <- Carseats %>% select("Sales", "Price", "ShelveLoc")
```

Splitting dataset into three parts based on Shelf Location

```
good <- filter(safeBabies,ShelveLoc == 'Good')
medium <- filter(safeBabies,ShelveLoc == 'Medium')
bad <- filter(safeBabies,ShelveLoc == 'Bad')
```

Question 1:

Building Linear Regression model between Sales(dependent) and Price(independent) for Good Shelf Location.

```
good_ordered <- good[order(good$Price),]
model_1 <- lm(Sales ~ Price, data = good_ordered)
summary(model_1)
```

```
##
## Call:
## lm(formula = Sales ~ Price, data = good_ordered)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.721 -1.351 -0.098  1.483  4.353
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.968864   0.988008  18.187  < 2e-16 ***
## Price       -0.065785   0.008199  -8.023 5.85e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.888 on 83 degrees of freedom
## Multiple R-squared:  0.4368, Adjusted R-squared:   0.43
## F-statistic: 64.37 on 1 and 83 DF,  p-value: 5.848e-12
```

Finding out optimal price for Good Shelf Location by subsituting values obtained from above regression model in the following equation.

```
good_optimal_price = (-0.065785 *55 - 17.968864)/(2 * -0.065785)
print(good_optimal_price)
```

```
## [1] 164.0727
```

Varying production cost from $40 to $85, to find optimal price

```
result <- vector("numeric", 40)
for(cost in 40:86) {
  good_optimized_result <- (-0.065785 *cost - 17.968864)/(2 * -0.065785)
  result[cost - 40] <- good_optimized_result
}
```

Using cbind to create dataframe by binding Optimized price and change in cost. Naming the columns accordingly.

```
price<- c(40:85)
good_optimized_price<-cbind.data.frame(result,price)
names(good_optimized_price)<-c('Optimized_Price','Change_in_Cost')
```

Question 2

Similar to above solution for 1st question. Repeating all the codes for Bad Shelf Location to find optimal price.

```r
bad_ordered<-bad[order(bad$Price),]
model_2 <- lm(Sales ~ Price, data = bad_ordered)
summary(model_2)
```

```
##
## Call:
## lm(formula = Sales ~ Price, data = bad_ordered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4622 -1.0617 -0.2014  1.2050  4.6412
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.832984   0.990317  11.949  < 2e-16 ***
## Price       -0.055220   0.008486  -6.507  3.7e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.967 on 94 degrees of freedom
## Multiple R-squared:  0.3105, Adjusted R-squared:  0.3032
## F-statistic: 42.34 on 1 and 94 DF,  p-value: 3.702e-09
```

```r
bad_optimal_price = (-0.05522 * 55 - 11.832984) / (2 * -0.05522)
print(bad_optimal_price)
```

```
## [1] 134.644
```

```r
result_1 <- vector("numeric", 40)
for(cost in 40:86) {
  bad_optimized_result <- (-0.05522 *cost - 11.832984)/(2 * -0.05522)
  result_1[cost - 40] <- bad_optimized_result
}

bad_optimized_price<-cbind.data.frame(result_1,price)
names(bad_optimized_price)<-c('Optimized_Price','Change_in_Cost')
```

Plotting above two results and comparing them to see the variations in optimal price using ggplot

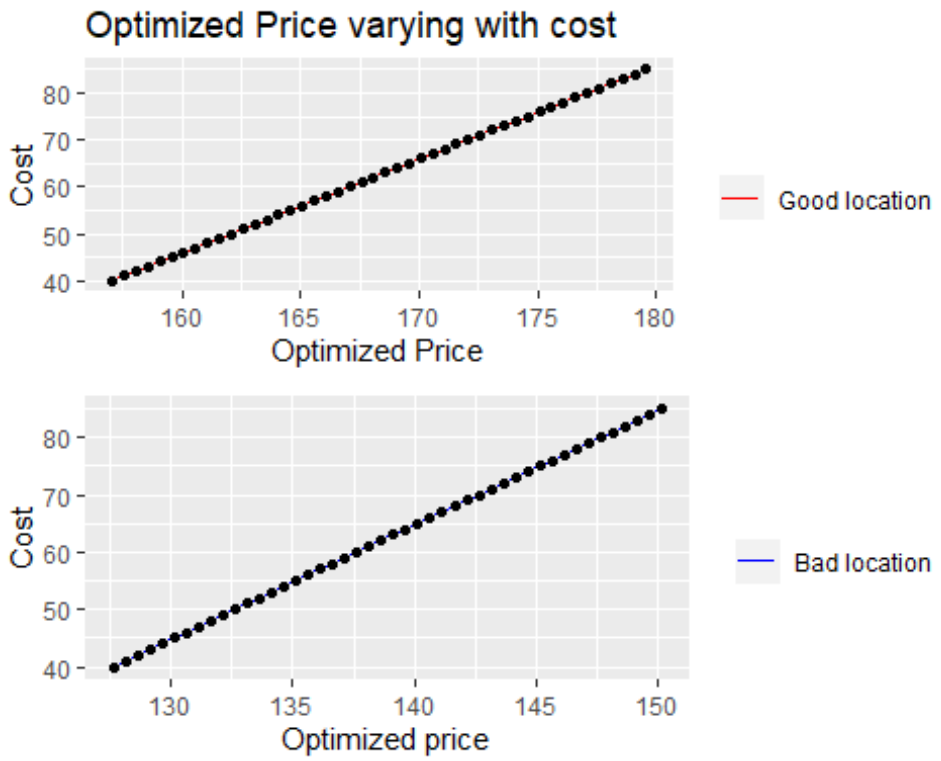
```r
plot_good <- ggplot(good_optimized_price, aes(Optimized_Price,
Change_in_Cost, colour='Good location')) +
             labs(title = 'Optimized Price varying with cost',x='Optimized
Price',y='Cost') +
             geom_line() +
             scale_color_manual("", values = ("Good Price" = "red")) +
             geom_point(colour='black')

plot_bad <- ggplot(bad_optimized_price, aes(Optimized_Price, Change_in_Cost,
colour='Bad location')) +
            labs(x='Optimized price',y='Cost') +
```

```
            geom_line() +
            scale_color_manual("", values = ("Bad Price" = "blue")) +
            geom_point(colour='black')
```

Using GridExtra library to display the plots together.

```
grid.arrange(plot_good, plot_bad, ncol=1)
```



We have obtained optimal price to sell carseats when production cost is $55 - for good shelf location as 164.07 - for bad shelf location as 134.64

Looking at the graph we can observe that optimal price varies from $120 to $150 for Bad shelf location and from $150 to $180 for good shelf location when we vary production cost from $40 to $85