Business Analytics (110-1)

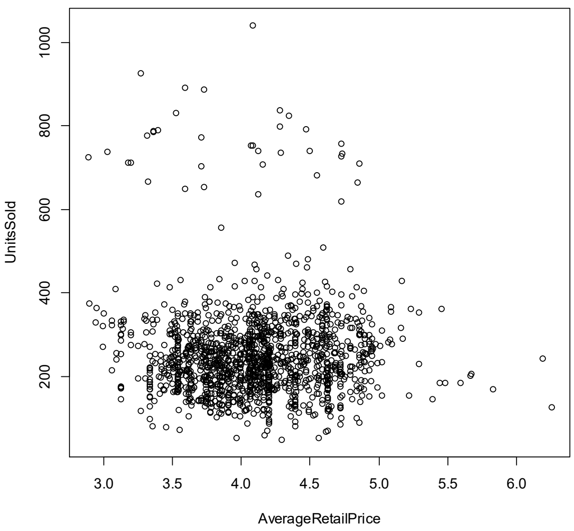
Assignment 2 – Reference Solutions

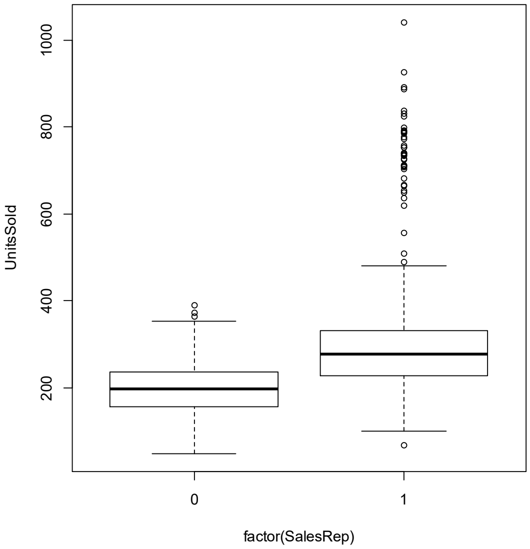
1.

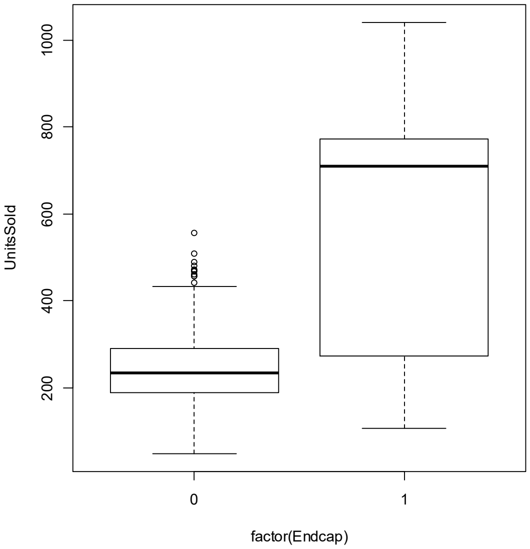
(a) & (b)

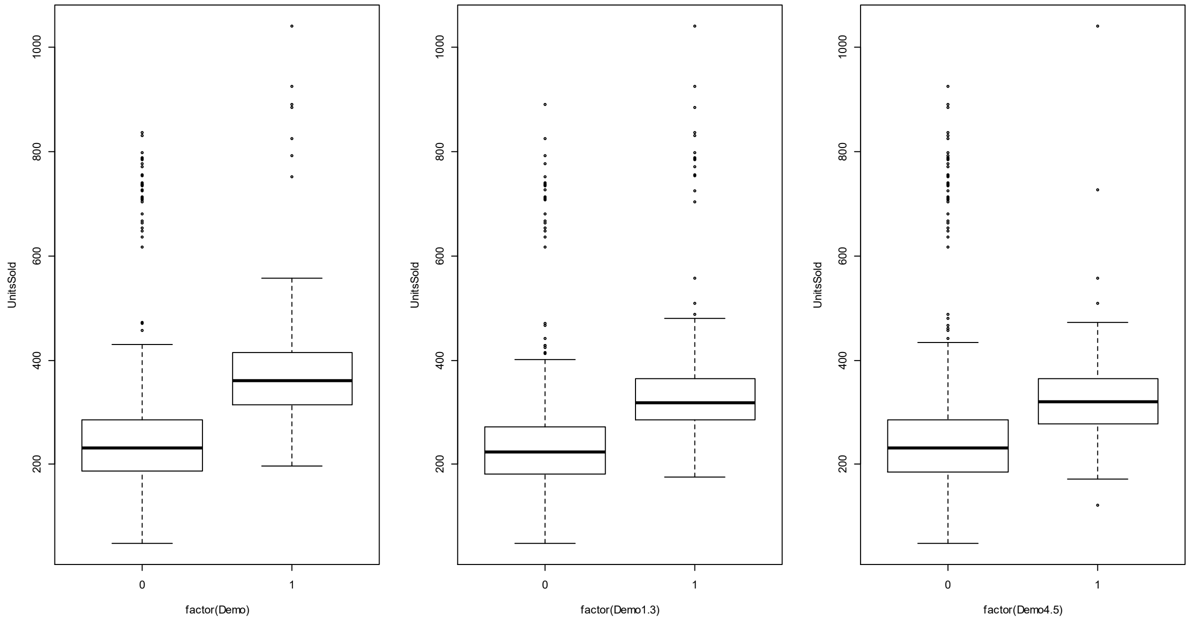
In the GoodBelly dataset, we have 12 variables. But only 2 of them are numerical/continuous; the other 10 are categorical/discrete. Thus, the EDA can be conducted with scatter plot and boplots, as follows.

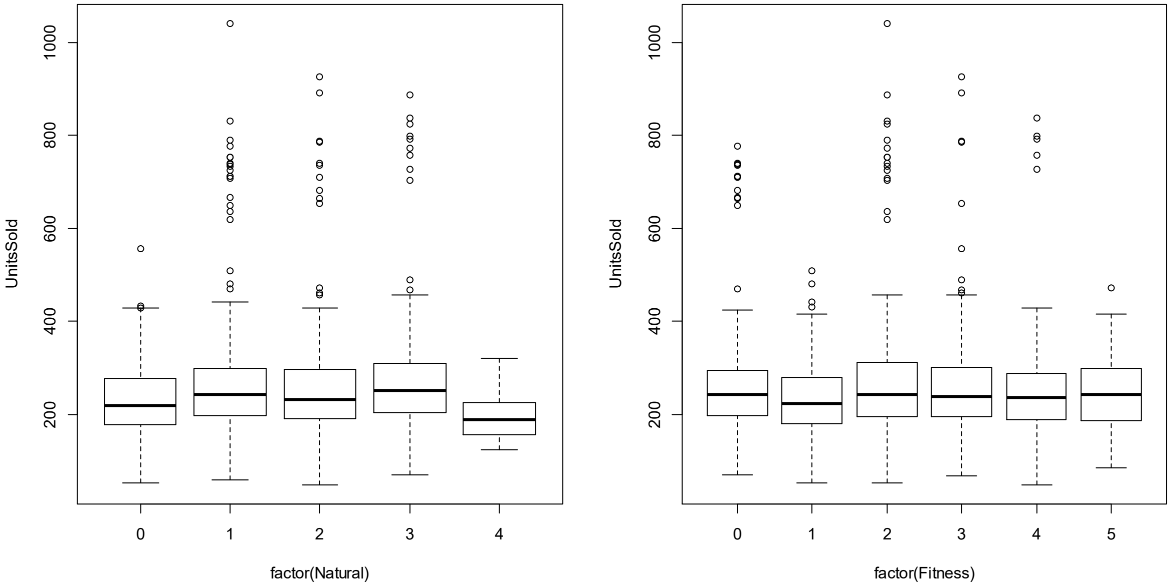
```
goodbelly <- read.csv("GoodBelly_data.csv", header=TRUE)
summary(goodbelly);     attach(goodbelly)
```







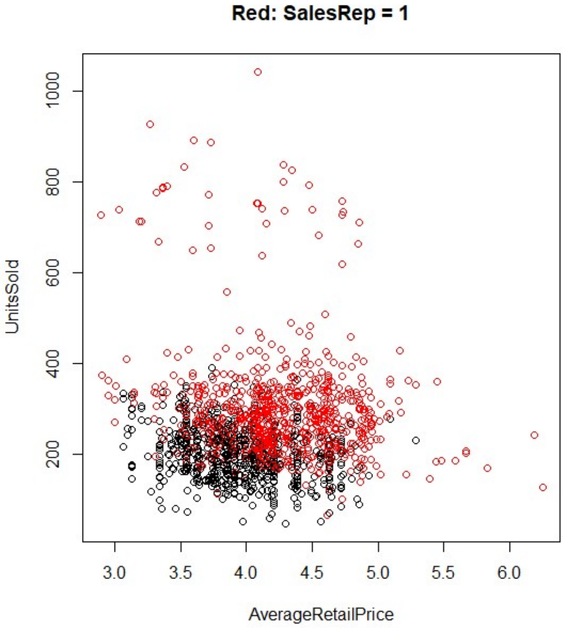


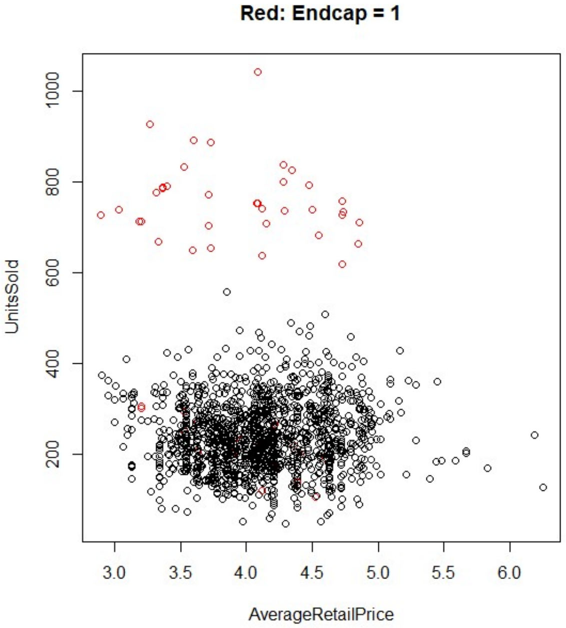


Some categorical variables have some interesting associations with `UnitsSold`, and that `UnitsSold` and `AverageRetailPrice` is negatively corelated.

The scatterplot can be more informative with color-coded categorical variable, as follows.
You should have tried various combinations.

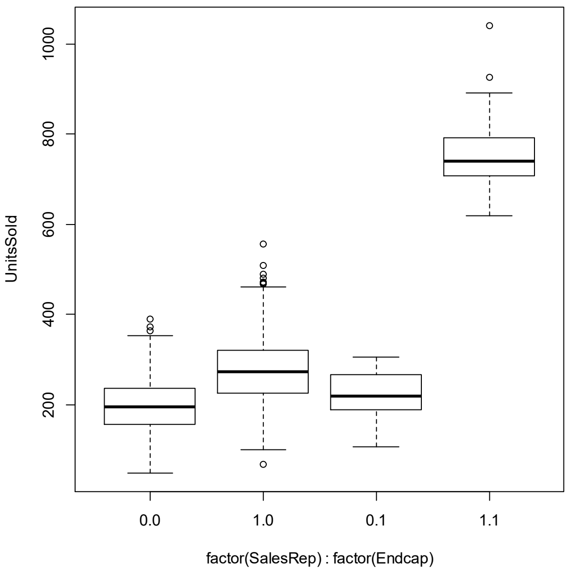
Obviously both `SalesRep` and `EndCap` highly influence the distribution pattern of `UnitsSold`, observed from the scatterplots below.

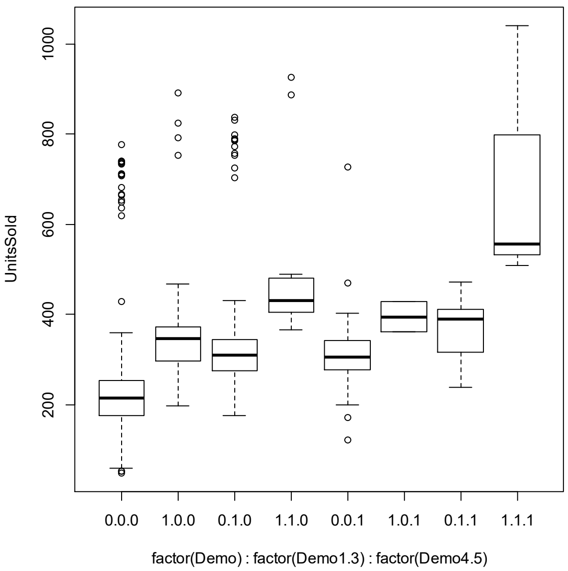




The interaction effect between `SalesRep` and `EndCap` can be graphically explored, as follows.

```
boxplot(UnitsSold ~ factor(SalesRep) + factor(Endcap))
boxplot(UnitsSold ~ factor(Demo) + factor(Demo1.3) + factor(Demo4.5))
```





What conclusion may you draw from the plots above?

(c) – (g)

```
m01 <- lm(UnitsSold ~ AverageRetailPrice + SalesRep + Endcap + Demo + Demo1.3 +
Demo4.5 + Natural + Fitness)
summary(m01)
```

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         298.488     16.183  18.444  < 2e-16 ***
AverageRetailPrice  -28.535      3.952  -7.220 8.56e-13 ***
SalesRep             77.437      3.864  20.038  < 2e-16 ***
Endcap              305.102      9.056  33.692  < 2e-16 ***
Demo                111.133      7.404  15.010  < 2e-16 ***
Demo1.3              73.517      4.895  15.018  < 2e-16 ***
Demo4.5              67.570      6.542  10.329  < 2e-16 ***
Natural              -1.594      1.776  -0.897    0.370
Fitness              -1.020      1.084  -0.941    0.347
---
Residual standard error: 63.69 on 1377 degrees of freedom
Multiple R-squared:  0.6726,   Adjusted R-squared:  0.6707
F-statistic: 353.7 on 8 and 1377 DF,  p-value: < 2.2e-16
```

See if `Natural` and `Fitness` can be dropped:

```
m02 <- update(m01, .~.- Natural - Fitness)
summary(m02)
anova(m01, m02, test="F")
```

```
Model 1: UnitsSold ~ AverageRetailPrice + SalesRep + Endcap + Demo + Demo1.3 +
    Demo4.5 + Natural + Fitness
Model 2: UnitsSold ~ AverageRetailPrice + SalesRep + Endcap + Demo + Demo1.3 +
    Demo4.5
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1   1377 5586216
2   1379 5592534 -2     -6318 0.7787 0.4592
```

Thus, all the promotional efforts are significantly related to sales, and either `Natural` nor `Fitness` is influential.   Let's consider the interaction effect between `SalesRep` and `EndCap`.

```
m05 <- update(m02, .~. + SalesRep:Endcap)
```

```
summary(m05)
```

```
Coefficients:
                    Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)          276.5735    12.1686   22.728     < 2e-16 ***
AverageRetailPrice -22.0664     3.0446   -7.248    7.04e-13 ***
SalesRep              59.4618     3.0112   19.747    < 2e-16 ***
Endcap                 0.6204    12.0769    0.051       0.959
Demo                 106.7527     5.7002   18.728     < 2e-16 ***
Demo1.3               73.3698     3.7660   19.482     < 2e-16 ***
Demo4.5               74.5520     5.0397   14.793     < 2e-16 ***
SalesRep:Endcap      453.8033    14.7372   30.793     < 2e-16 ***
---
Residual standard error: 49.03 on 1378 degrees of freedom
Multiple R-squared:  0.8059,    Adjusted R-squared:  0.8049
F-statistic: 817.1 on 7 and 1378 DF,  p-value: < 2.2e-16
```

How would you interpret the results of this model **`m05`***?*

The sequence of the models explored by the lecturer is as follows:

- `m02 <- update(m01, .~.- Natural - Fitness)`
- `m03 <- update(m02, .~. + Region)`
- `m04 <- update(m02, .~. + Endcap:(Demo+Demo1.3+Demo4.5))`
- `m04a <- update(m04, .~. - Endcap:Demo4.5)`
- `m034 <- update(m04, .~. + Region)`
- `m05 <- update(m02, .~. + Sales.Rep:Endcap)`
- `m035 <- update(m05, .~. + Region)`
- `m045 <- update(m05, .~. + Endcap:(Demo+Demo1.3))`
- `m05f <- update(m05, .~. + as.factor(Fitness))`
- `m05d <- update(m05, .~. + Demo:Demo1.3 + Demo:Demo4.5 + Demo1.3:Demo4.5)`
- `m05r <- update(m05, .~. + Sales.Rep:Demo)`

Among these, **`m05`** turns out to be the best. *Why?*

NOTE: This is NOT saying that **`m05`** is the best model from all aspects and that **`m05`** is "the standard solution". There are still some other possibilities that the lecturer did not explore. What is your model search plan and path? Have you discovered some other models better than **`m05`**? In what aspect?