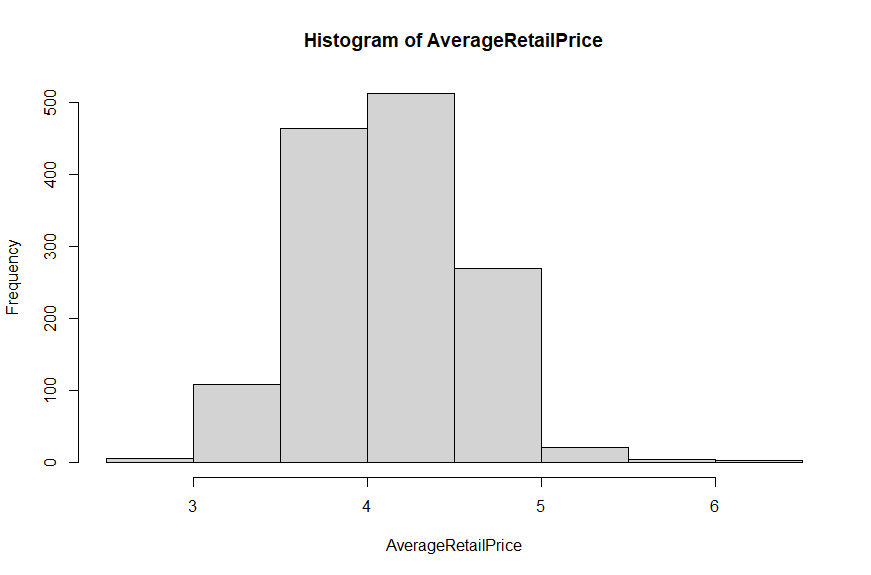
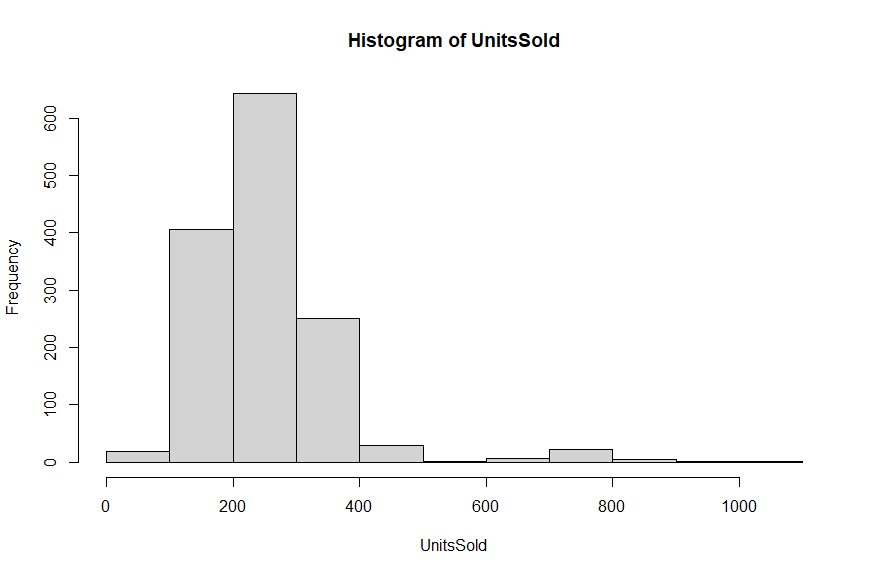
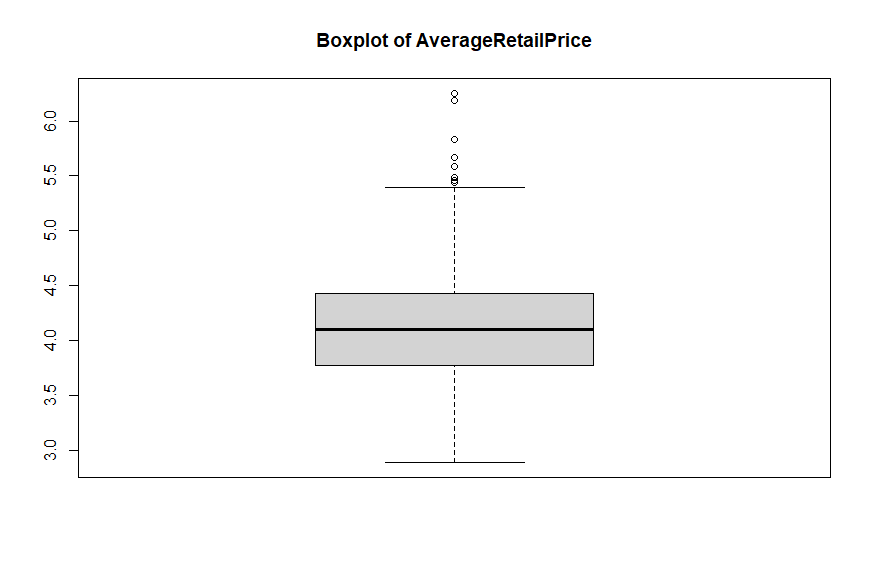
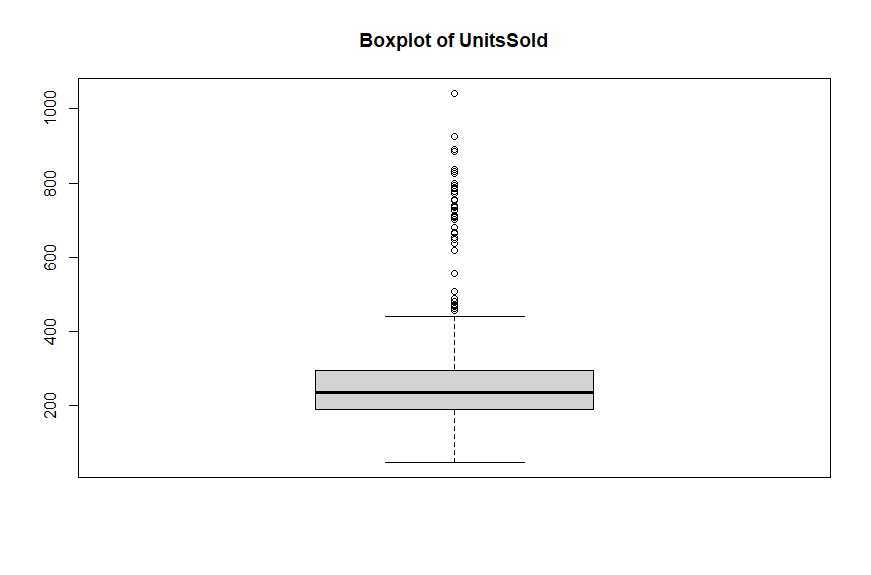
**Business Analytics (110-1)**

**Assignment 2**

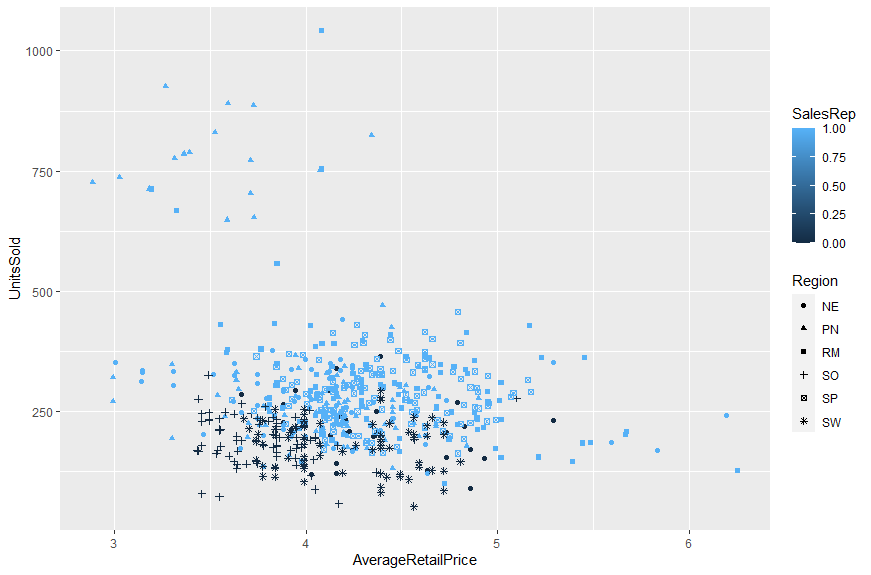
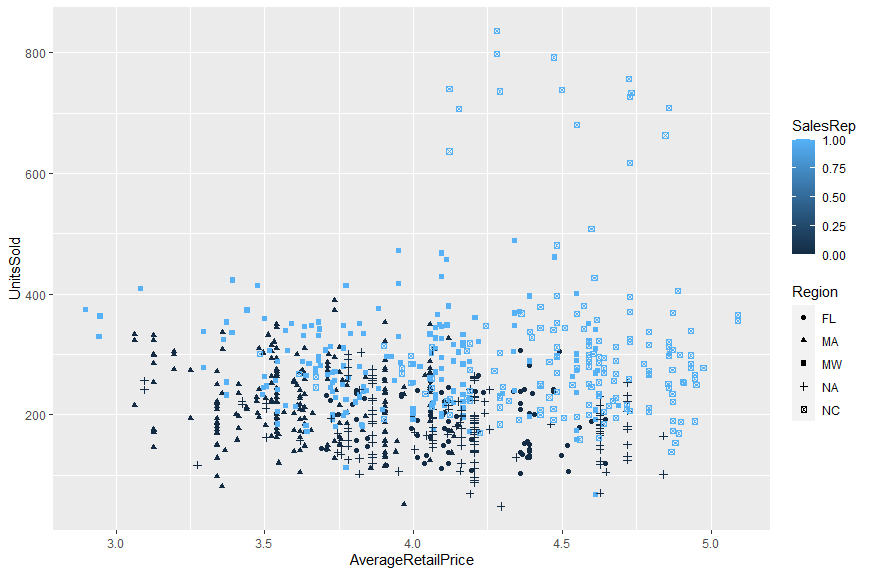
B08701244 工管三 蔡鉎驊

1. Do descriptive statistics, with R, to provide an overview for your retailing business. You may do it from any perspective with any EDA method. You may include some basic summaries as well as some emphases on interesting findings.



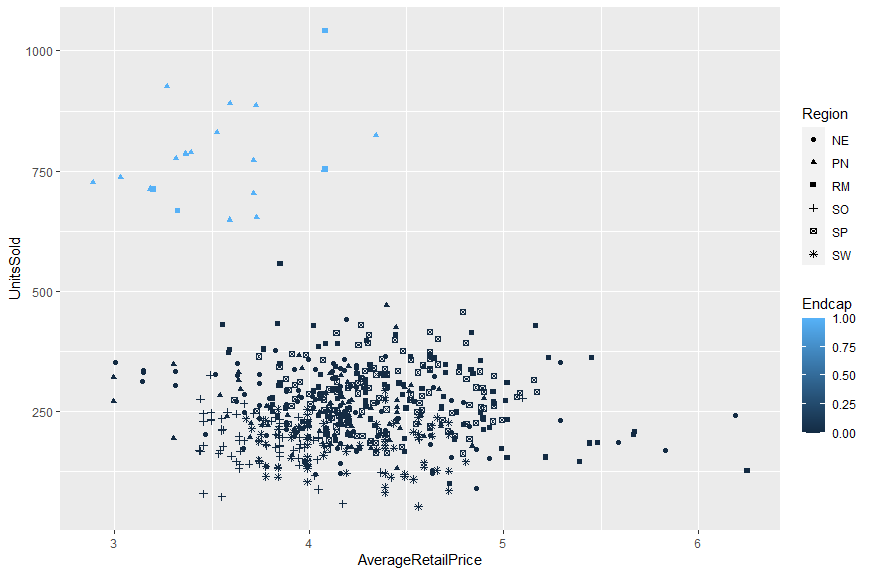
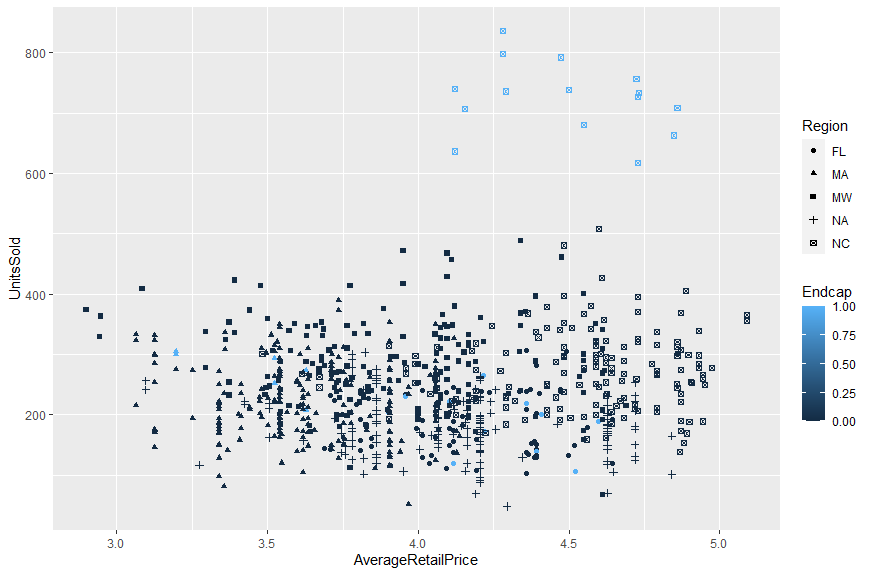


由上面兩張圖可以大致知道UnitsSold與AverageRetailPrice的分布狀況。而相較於AverageRetailPrice，UnitsSold的直方圖呈現右偏，也較多outlier。



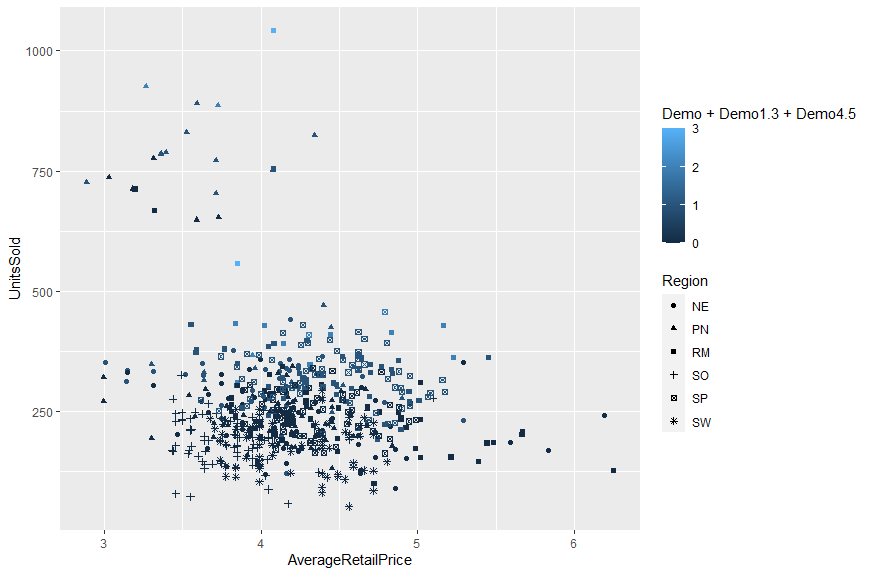
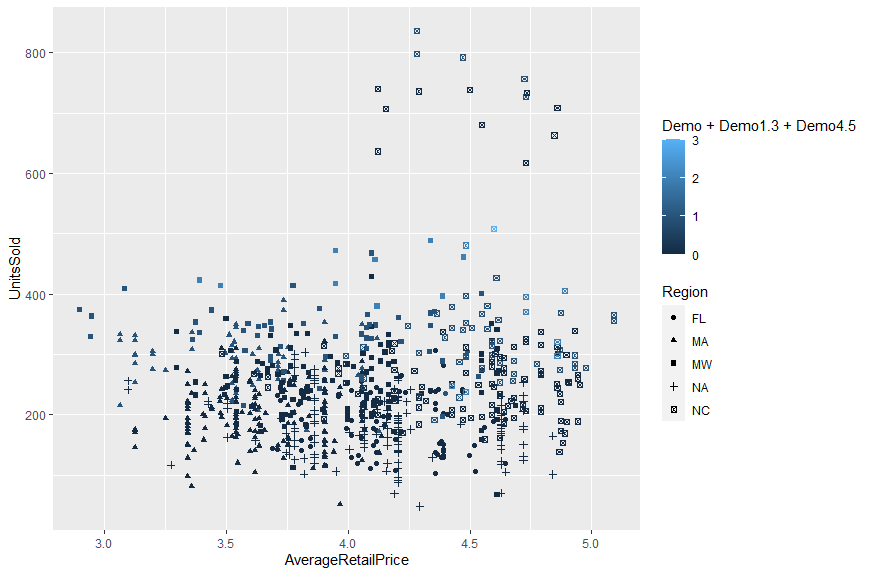
由這兩張圖可以看得出來UnitsSold與AverageRetailPrice都與地區有關係。

而SalesRep與UnitsSold關係比較顯著，與AverageRetailPrice的關係不明顯。



從這兩張圖可以發現在「某些地區」Endcap對於UnitsSold的關係滿顯著的。例如第一張圖，同為NC地區的商店，有無Endcap的UnitsSold差異就滿大的。

但Endcap對於AverageRetailPrice的關係似乎不大，第一張圖有Endcap的商店的AverageRetailPrice偏高，第二張圖有Endcap的商店的AverageRetailPrice偏低，推測是不同地區導致的差異。



由這兩張圖可以發現Demo總次數對於UnitsSold以及AverageRetailPrice的關係，並不比地區因素影響大。

1. Based on your findings from (a), comment on the marketing activities about their effectiveness. Use some graphs and numbers to support your comments. You may comment on all of them, rank them, making suggestions about how to use them. Of course, your comments may be different from region to region, from time to time, or depending on any factor that you find useful.

從上面第3張圖（第一張顏色以Endcap來標記的圖）可以發現，對於NC來說promotion與UnitsSold的關係滿大的，但對於MA與FL來說關係就不顯著。

1. Build a linear model to explain the relationship between sales and promotional efforts, and interpret the regression output.

Call:

lm(formula = UnitsSold ~ Endcap)

Residuals:

Min 1Q Median 3Q Max

-477.51 -52.63 -4.79 54.26 457.28

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 240.696 2.448 98.31 <2e-16 \*\*\*

Endcap 343.229 12.521 27.41 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 89.39 on 1384 degrees of freedom

Multiple R-squared: 0.3519, Adjusted R-squared: 0.3514

F-statistic: 751.5 on 1 and 1384 DF, p-value: < 2.2e-16

線性回歸模型跑出來後發現，Endcap的p-value<0.05（Endcap也就是promotion與UnitsSold也就是Sales有關係），係數為343.229，代表若有promotion，我們預估Sales會上升343.229。

1. Does the in-store demo program boost the sales? If so, for how long does the sales lift last?

Call:

lm(formula = UnitsSold ~ Demo + Demo1.3 + Demo4.5)

Residuals:

Min 1Q Median 3Q Max

-183.59 -48.56 -10.63 29.09 555.29

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 221.003 2.897 76.278 <2e-16 \*\*\*

Demo 150.622 10.754 14.006 <2e-16 \*\*\*

Demo1.3 113.173 6.942 16.303 <2e-16 \*\*\*

Demo4.5 83.110 9.528 8.723 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 93.85 on 1382 degrees of freedom

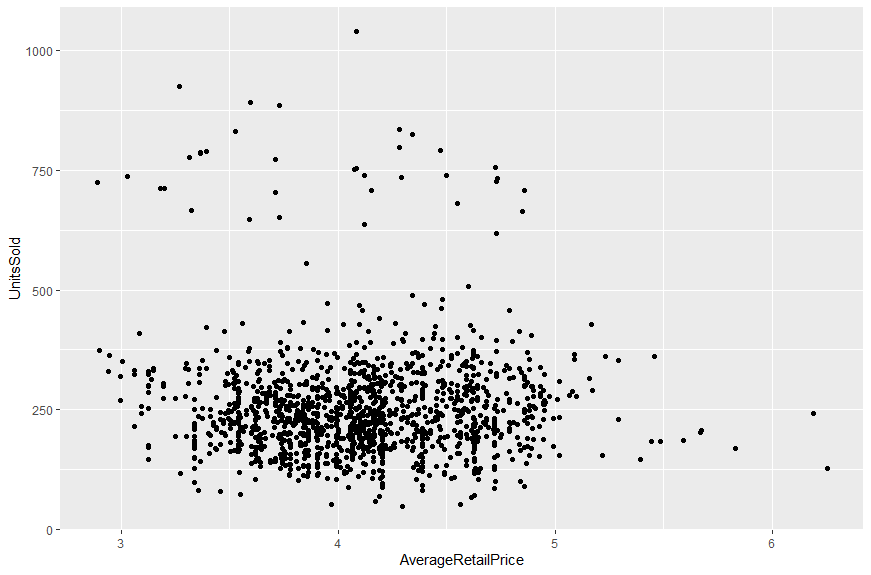
Multiple R-squared: 0.2867, Adjusted R-squared: 0.2852

F-statistic: 185.2 on 3 and 1382 DF, p-value: < 2.2e-16

三個有關Demo的變數的p-value都小於0.05（Demo與Sales有關係），係數分別為150.622, 113.173, 83.110，代表若分別有這三個Demo，則我們預估Sales會上升的幅度分別為150.622, 113.173, 83.110。

但此擁有三個自變數的模型的解釋力（R\_squared）還比上面只有promotion為自變數的模型的解釋力還來的小。此推論也可以拿第3, 4張圖（關於promotion）去與第5, 6張圖（關於demo）比較。

1. Does the placement of the product within the store affect the sales?



lm(formula = UnitsSold ~ AverageRetailPrice)

Residuals:

Min 1Q Median 3Q Max

-205.42 -63.44 -16.86 41.63 787.27

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 272.327 26.583 10.244 <2e-16 \*\*\*

AverageRetailPrice -4.506 6.432 -0.701 0.484

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 111 on 1384 degrees of freedom

Multiple R-squared: 0.0003545, Adjusted R-squared: -0.0003678

F-statistic: 0.4908 on 1 and 1384 DF, p-value: 0.4837

AverageRetailPrice這想變數的p-value高達0.484。由此模型可以發現商店內擺放商品的價格與銷售額關係不大。

1. What other factors affect the sales of Goodbelly’s products? Based on the regression output, what are your recommendations to Goodbelly’s management?

Call:

lm(formula = UnitsSold ~ Region)

Residuals:

Min 1Q Median 3Q Max

-216.14 -54.98 -10.80 40.21 738.81

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 188.486 10.402 18.120 < 2e-16 \*\*\*

RegionMA 33.917 12.400 2.735 0.00632 \*\*

RegionMW 94.039 12.740 7.381 2.70e-13 \*\*\*

RegionNA -7.170 13.221 -0.542 0.58770

RegionNC 123.875 12.881 9.617 < 2e-16 \*\*\*

RegionNE 65.888 13.956 4.721 2.58e-06 \*\*\*

RegionPN 155.315 14.296 10.864 < 2e-16 \*\*\*

RegionRM 113.904 13.956 8.162 7.39e-16 \*\*\*

RegionSO 3.676 15.227 0.241 0.80926

RegionSP 97.302 13.429 7.246 7.15e-13 \*\*\*

RegionSW -10.127 15.227 -0.665 0.50614

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 97.58 on 1375 degrees of freedom

Multiple R-squared: 0.2327, Adjusted R-squared: 0.2271

F-statistic: 41.71 on 10 and 1375 DF, p-value: < 2.2e-16

某些地區與銷售額也有關係，像是若以FL地區為default，則位於MA, MW, NC, NE, PN, RM, SP的預測值就會與原本以FL為default的預測值不同。

建立預測Sales的model時，自變數應包含Region, Store, UnitsSold, AverageRetailPrice, SalesRep, Endcap, Demo, Demo1-3, Demo4-5。

1. Are there any suggestions to improve and refine the model?

**原本模型：**

Call:

lm(formula = UnitsSold ~ Region + AverageRetailPrice + SalesRep +

Endcap + Demo + Demo1.3 + Demo4.5 + Natural + Fitness)

Residuals:

Min 1Q Median 3Q Max

-356.80 -35.22 1.02 37.40 233.90

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 286.8903 21.1406 13.571 < 2e-16 \*\*\*

RegionMA 24.0038 8.5591 2.804 0.005111 \*\*

RegionMW 60.7394 15.9747 3.802 0.000150 \*\*\*

RegionNA 31.6328 8.6316 3.665 0.000257 \*\*\*

RegionNC 81.5800 16.2518 5.020 5.85e-07 \*\*\*

RegionNE 54.4885 13.4711 4.045 5.53e-05 \*\*\*

RegionPN 80.1268 16.4637 4.867 1.27e-06 \*\*\*

RegionRM 64.7195 16.5474 3.911 9.64e-05 \*\*\*

RegionSO 31.0466 10.1072 3.072 0.002170 \*\*

RegionSP 66.4552 16.3813 4.057 5.26e-05 \*\*\*

RegionSW 30.0176 9.9486 3.017 0.002598 \*\*

AverageRetailPrice -32.6520 4.7842 -6.825 1.32e-11 \*\*\*

SalesRep 35.2423 13.5991 2.592 0.009657 \*\*

Endcap 302.7750 9.3902 32.244 < 2e-16 \*\*\*

Demo 112.8824 7.4014 15.251 < 2e-16 \*\*\*

Demo1.3 73.8848 4.9371 14.965 < 2e-16 \*\*\*

Demo4.5 65.8542 6.6019 9.975 < 2e-16 \*\*\*

Natural -1.3787 1.8222 -0.757 0.449412

Fitness -0.1166 1.1465 -0.102 0.919013

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 63.04 on 1367 degrees of freedom

Multiple R-squared: 0.6817, Adjusted R-squared: 0.6775

F-statistic: 162.6 on 18 and 1367 DF, p-value: < 2.2e-16

去掉Natural與Fitness這兩個變數後，做一次nested model test

Model 1: UnitsSold ~ Region + AverageRetailPrice + SalesRep + Endcap +

Demo + Demo1.3 + Demo4.5

Model 2: UnitsSold ~ Region + AverageRetailPrice + SalesRep + Endcap +

Demo + Demo1.3 + Demo4.5 + Natural + Fitness

Res.Df RSS Df Sum of Sq F Pr(>F)

1 1369 5434097

2 1367 5431822 2 2275.1 0.2863 0.7511

Since the p-value of 0.7511 is more than our significance level

(0.05), we cannot reject the null hypothesis.

Removing “Natural” & “Fitness” these two indicators improves the values of

adjusted R2 from 0.6775 to 0.6778

**最後模型：**

Call:

lm(formula = UnitsSold ~ Region + AverageRetailPrice + SalesRep +

Endcap + Demo + Demo1.3 + Demo4.5)

Residuals:

Min 1Q Median 3Q Max

-358.36 -34.99 1.06 37.66 234.50

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 284.862 20.958 13.592 < 2e-16 \*\*\*

RegionMA 23.668 8.542 2.771 0.005668 \*\*

RegionMW 60.092 15.892 3.781 0.000163 \*\*\*

RegionNA 31.590 8.618 3.665 0.000256 \*\*\*

RegionNC 80.995 16.217 4.994 6.66e-07 \*\*\*

RegionNE 54.304 13.450 4.037 5.71e-05 \*\*\*

RegionPN 80.505 16.398 4.909 1.02e-06 \*\*\*

RegionRM 64.888 16.533 3.925 9.11e-05 \*\*\*

RegionSO 31.480 10.065 3.128 0.001800 \*\*

RegionSP 66.179 16.368 4.043 5.56e-05 \*\*\*

RegionSW 30.710 9.901 3.102 0.001963 \*\*

AverageRetailPrice -32.683 4.739 -6.897 8.08e-12 \*\*\*

SalesRep 35.228 13.591 2.592 0.009645 \*\*

Endcap 302.131 9.347 32.325 < 2e-16 \*\*\*

Demo 113.022 7.395 15.284 < 2e-16 \*\*\*

Demo1.3 74.078 4.928 15.032 < 2e-16 \*\*\*

Demo4.5 65.957 6.596 9.999 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 63 on 1369 degrees of freedom

Multiple R-squared: 0.6815, Adjusted R-squared: 0.6778

F-statistic: 183.1 on 16 and 1369 DF, p-value: < 2.2e-16