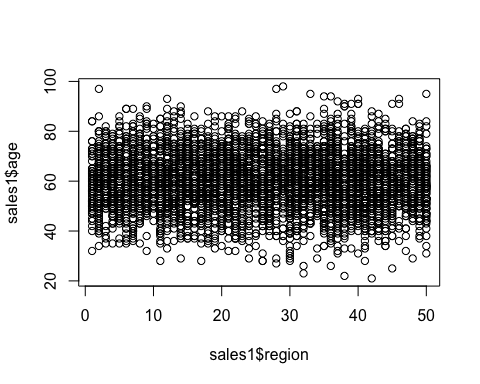
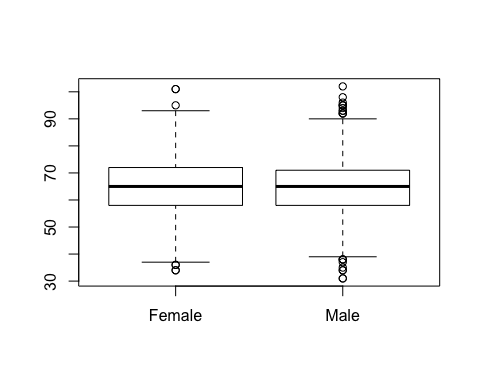
Kaiyue Wang

sales1 <- read.csv("/Users/Karen/Downloads/customer\_sales1.csv")  
sales1 <- sales1[sales1$sale\_amount>0,]  
plot(sales1$region, sales1$age)



plot(sales1$gender, sales1$income)



salessg <-lm(sales1$sale\_amount~sales1$gender)  
summary(salessg)

##   
## Call:  
## lm(formula = sales1$sale\_amount ~ sales1$gender)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -570.86 -196.82 -51.12 144.67 1539.18   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 580.862 6.065 95.78 <2e-16 \*\*\*  
## sales1$genderMale -285.395 8.029 -35.55 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 274.3 on 4763 degrees of freedom  
## Multiple R-squared: 0.2097, Adjusted R-squared: 0.2095   
## F-statistic: 1264 on 1 and 4763 DF, p-value: < 2.2e-16

There is a difference in the amount of sales by gender because the P-value is 2.2e^-16 which is less than 0.01, so it’s significant at 0.1% level, and t-statistic is also significant at 0.1% level, we are confident that it is different by gender.

sales1$age.cat <- cut(sales1$age,  
 right = FALSE,  
 breaks = c(0,45,65,100),  
 labels = c("Young", "Middle", "Old"))

data1 <- lm(sales1$sale\_amount~gender, data = sales1)  
summary(data1)

##   
## Call:  
## lm(formula = sales1$sale\_amount ~ gender, data = sales1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -570.86 -196.82 -51.12 144.67 1539.18   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 580.862 6.065 95.78 <2e-16 \*\*\*  
## genderMale -285.395 8.029 -35.55 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 274.3 on 4763 degrees of freedom  
## Multiple R-squared: 0.2097, Adjusted R-squared: 0.2095   
## F-statistic: 1264 on 1 and 4763 DF, p-value: < 2.2e-16

data2 <- lm(sales1$sale\_amount~gender+income, data = sales1)  
summary(data2)

##   
## Call:  
## lm(formula = sales1$sale\_amount ~ gender + income, data = sales1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -581.57 -196.30 -51.92 143.59 1528.98   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 542.5688 26.6361 20.370 <2e-16 \*\*\*  
## genderMale -285.3086 8.0280 -35.539 <2e-16 \*\*\*  
## income 0.5904 0.3999 1.476 0.14   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 274.3 on 4762 degrees of freedom  
## Multiple R-squared: 0.21, Adjusted R-squared: 0.2097   
## F-statistic: 633 on 2 and 4762 DF, p-value: < 2.2e-16

data3 <- lm(sales1$sale\_amount~gender+income+age, data = sales1)  
summary(data3)

##   
## Call:  
## lm(formula = sales1$sale\_amount ~ gender + income + age, data = sales1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -579.9 -196.3 -48.6 144.7 1547.2   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 477.0329 36.8759 12.936 <2e-16 \*\*\*  
## genderMale -275.7921 8.8376 -31.207 <2e-16 \*\*\*  
## income 0.5849 0.3996 1.464 0.1434   
## age 1.0054 0.3915 2.568 0.0103 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 274.1 on 4761 degrees of freedom  
## Multiple R-squared: 0.2111, Adjusted R-squared: 0.2106   
## F-statistic: 424.7 on 3 and 4761 DF, p-value: < 2.2e-16

data4 <- lm(sales1$sale\_amount~gender+income+age+region, data = sales1)  
summary(data4)

##   
## Call:  
## lm(formula = sales1$sale\_amount ~ gender + income + age + region,   
## data = sales1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -587.1 -194.8 -50.0 145.1 1548.5   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 496.3626 37.4399 13.258 <2e-16 \*\*\*  
## genderMale -275.1308 8.8336 -31.146 <2e-16 \*\*\*  
## income 0.6046 0.3994 1.514 0.1301   
## age 1.0076 0.3912 2.576 0.0100 \*   
## region -0.8188 0.2812 -2.912 0.0036 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 273.9 on 4760 degrees of freedom  
## Multiple R-squared: 0.2125, Adjusted R-squared: 0.2119   
## F-statistic: 321.2 on 4 and 4760 DF, p-value: < 2.2e-16

BIC(data1,data2,data3,data4)

## df BIC  
## data1 3 67050.50  
## data2 4 67056.78  
## data3 5 67058.66  
## data4 6 67058.64

Data1 has lowest BIC, therefore better fitting model than the others

Customer\_sales1New <- read.table('/Users/Karen/Downloads/customer\_sales2.csv', sep = ',', header = T, stringsAsFactors = F)  
Customer\_sales1Predict <- predict(data1, newdata = Customer\_sales1New, se.fit =T,   
 interval = 'prediction', level = .95)  
head(Customer\_sales1Predict$fit)

## fit lwr upr  
## 1 580.8623 42.91592 1118.8086  
## 2 295.4670 -242.44677 833.3808  
## 3 580.8623 42.91592 1118.8086  
## 4 295.4670 -242.44677 833.3808  
## 5 295.4670 -242.44677 833.3808  
## 6 580.8623 42.91592 1118.8086

head(Customer\_sales1Predict$se.fit)

## 1 2 3 4 5 6   
## 6.064871 5.261018 6.064871 5.261018 5.261018 6.064871