Car Seat Sales

Celeste Guerrero 4/14/2021

Celeste Guerrero cbg928

```
# Introduction

# Installing appropriate packages
library(readxl)
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(tidyverse)



## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.5     v stringr 1.4.0
## v tidyr   1.1.2     v forcats 0.5.0
## v readr   1.4.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()



library(ggplot2)
library(cluster)
```

```
library(plotROC)
```

```
## Warning: package 'plotROC' was built under R version 4.0.5
```

```
# Dataset
library(readxl)
Carseats <- read_excel("Carseats.xlsx")
```

```
## New names:
## * `` -> ...1
```

```
View(Carseats)
```

The data set ‘Carseats’ was obtained from https://vincentarelbundock.github.io/Rdatasets/datasets.html. The dataset contains 400 observations across 11 variables. The data set focuses on child car seat sales among 400 different stores using simulated data.

This data set will be manipulated to include the variables, unit sales(in thousands), advertising budget of company for each store (in thousands), population size (in thousands), average age of the local population, urban (regarding location type), quality of shelving location across 3 levels (good, medium, bad) and US consisting of only stores in the United States. After selecting for stores in the United States, only 258 of the original 400 observations were included in the final set. This data set is being used because it can explore the relationships between car seat sales based on location demographics being in the United States and whether the store is present in an urban or rural area.

I expect there to be a significant difference in car seat sales and advertising costs based the effect on stores in urban and rural locations. In addition, I expect there to be a significant difference in car seat sales across different qualities of shelving location.

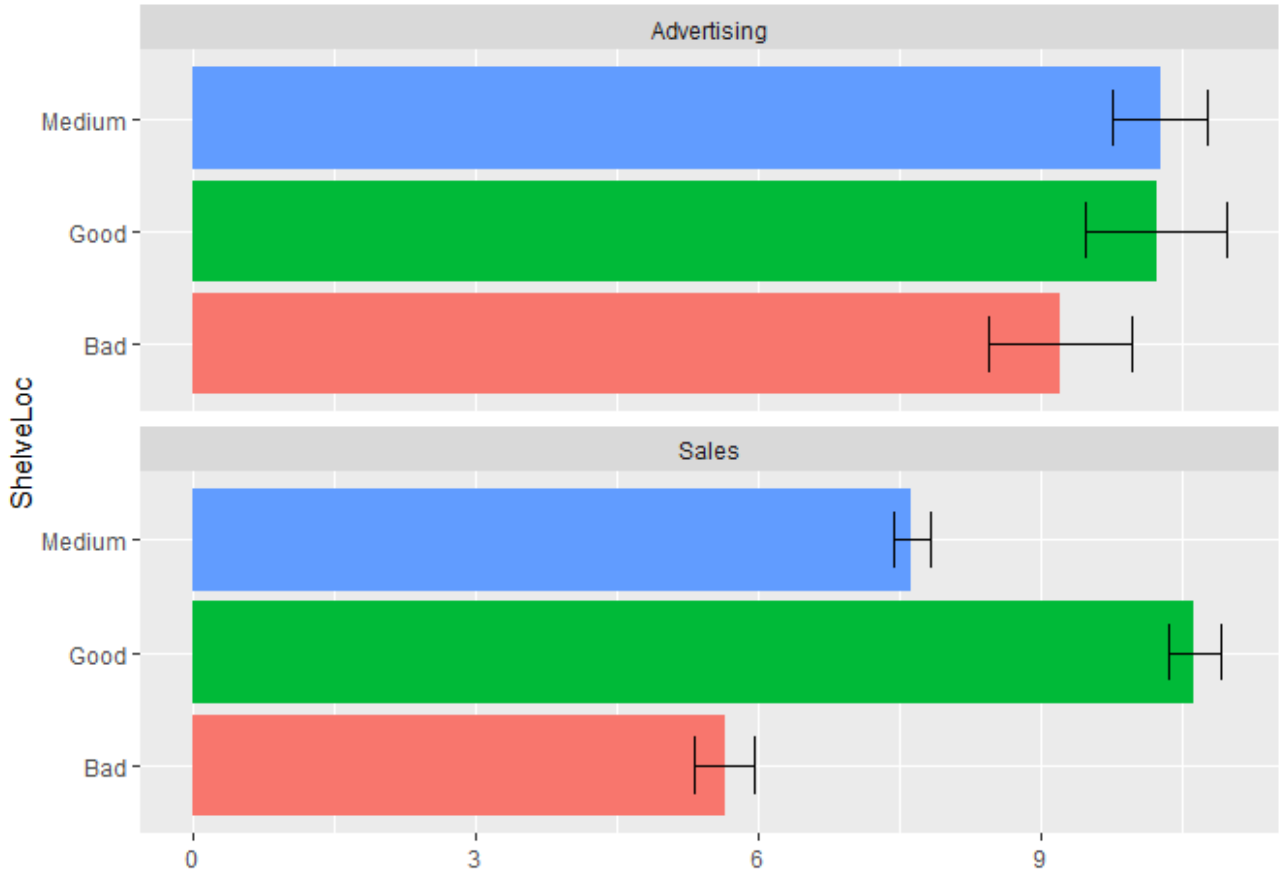
Project 2

```
# Selecting ideal variables
carseat1 <- Carseats %>% select(Sales, Advertising, Age, ShelveLoc, Population, Urban
view(carseat1)
```

```
# Univariate Statistics
carseat1 %>%
  group_by(ShelveLoc) %>%
  summarize(mean(Sales), mean(Advertising),mean(Population))  #mean of sales, adverti
```

```
## # A tibble: 3 x 4
##   ShelveLoc `mean(Sales)` `mean(Advertising)` `mean(Population)`
## * <chr>             <dbl>               <dbl>              <dbl>
## 1 Bad                5.65                9.21               299.
## 2 Good              10.6                10.2                280.
## 3 Medium             7.63               10.3                255
```

```
# Visual
carseat1 %>%
  select(ShelveLoc,Sales,Advertising) %>%
  pivot_longer(-1,names_to='DV', values_to='measure') %>%
  ggplot(aes(ShelveLoc,measure,fill=ShelveLoc)) +
  geom_bar(stat="summary", fun = "mean") +
  geom_errorbar(stat="summary", fun.data = "mean_se", width=.5) +
  facet_wrap(~DV, nrow=2) +
  coord_flip() +
  ylab("") +
  theme(legend.position = "none")
```




```
# Bivariate statistics

# Removing non numeric variables
carseat3 <- as.data.frame(carseat1) %>% select(-ShelveLoc, -Urban, -US)
head(carseat3)
```

```
##   Sales Advertising Age Population
## 1  9.50          11  42        276
## 2 11.22          16  65        260
## 3 10.06          10  59        269
## 4  7.40           4  55        466
## 5 10.81          13  78        501
## 6 11.85          15  67        425
```

```
# Correlation matrix
library(corrplot)
```

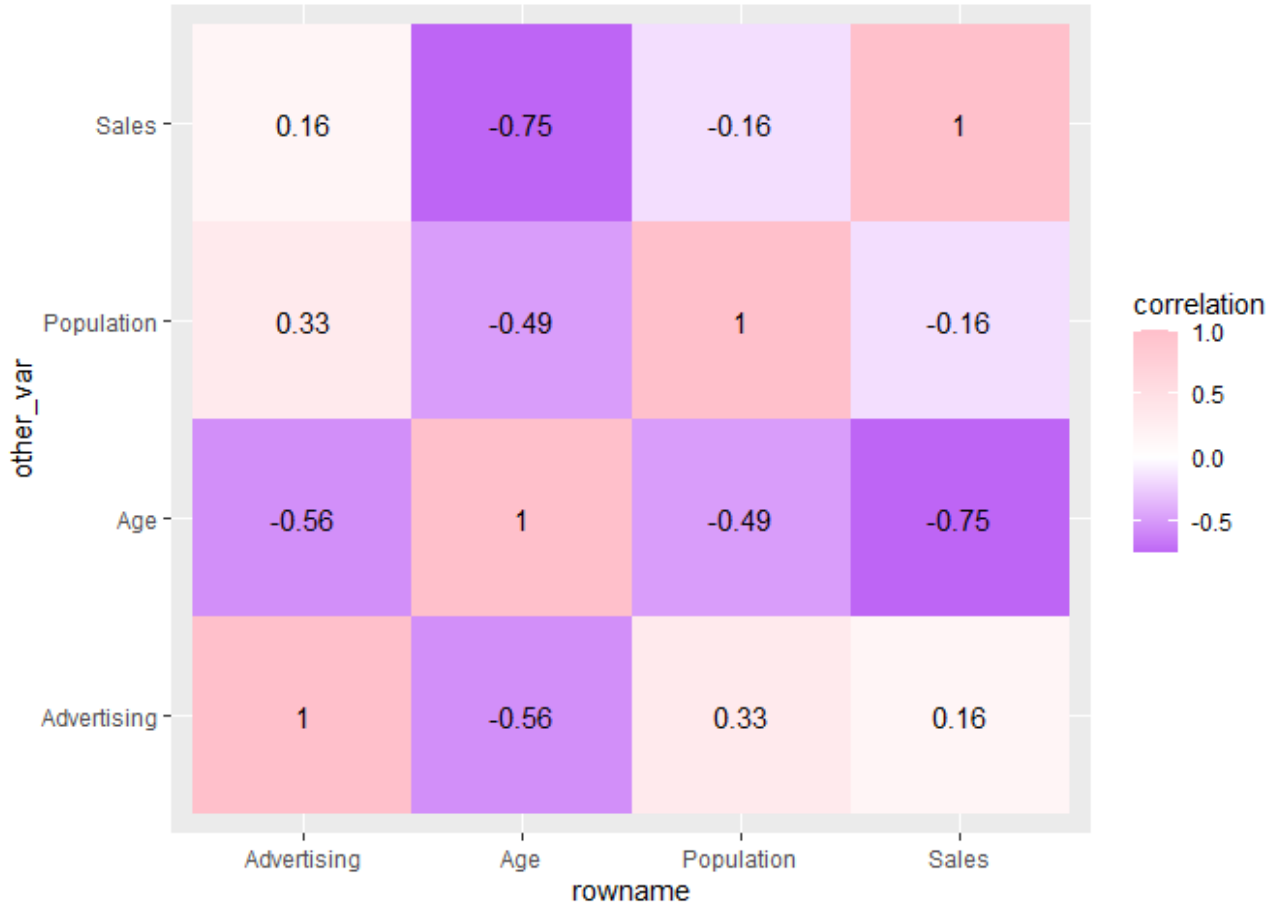
```
## Warning: package 'corrplot' was built under R version 4.0.4
```

```
## corrplot 0.84 loaded
```

```
cor2 <- carseat3 %>% select_if(is.numeric) %>%
cor(carseat3, use = "pairwise.complete.obs")
cor2
```

```
##                   Sales Advertising         Age  Population
## Sales        1.00000000  0.25573321 -0.23963491  0.06850591
## Advertising  0.25573321  1.00000000 -0.03111748  0.36519417
## Age         -0.23963491 -0.03111748  1.00000000 -0.06867497
## Population   0.06850591  0.36519417 -0.06867497  1.00000000
```

```
#Correlation heat map
cor(cor2, use = "pairwise.complete.obs") %>%
  # Save as a data frame
  as.data.frame %>%
  # Convert row names to an explicit variable
  rownames_to_column %>%
  # Pivot so that all correlations appear in the same column
  pivot_longer(-1, names_to = "other_var", values_to = "correlation") %>%
  ggplot(aes(rowname, other_var, fill=correlation)) +
  # Heatmap with geom_tile
  geom_tile() +
  # Change the scale to make the middle appear neutral
  scale_fill_gradient2(low="purple",mid="white",high="pink") +
  # Overlay values
  geom_text(aes(label = round(correlation,2)), color = "black", size = 4)
```



The univariate statistics demonstrate the 'Good' quality shelving locations have the highest mean sales, but these stores did not, on average, spend the most on Advertising. The stores with 'bad' quality shelving locations have the highest mean local population, but the lowest mean unit sales. The correlation matrix showed the highest correlation between local population and the amount spent on advertising. The second highest correlation was between the amount spent on advertising and the revenue generated from car seat sales.

```
# MANOVA

manova_carseat <- manova(cbind(Sales, Advertising) ~ ShelveLoc, data = carseat1)

# OUtput of MANOVA
summary(manova_carseat)
```

```
##             Df  Pillai approx F num Df den Df    Pr(>F)
## ShelveLoc    2 0.38012   29.919      4    510 < 2.2e-16 ***
## Residuals 255
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Based on signficant results
summary.aov(manova_carseat)
```

```
##  Response Sales :
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## ShelveLoc     2  777.95  388.98  73.503 < 2.2e-16 ***
## Residuals   255 1349.46    5.29
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Advertising :
##              Df Sum Sq Mean Sq F value Pr(>F)
## ShelveLoc     2   52.1  26.032   0.742 0.4772
## Residuals   255 8945.9  35.082
```

```r
# Post hoc analysis
    pairwise.t.test(carseat1$Sales,carseat1$ShelveLoc, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  carseat1$Sales and carseat1$ShelveLoc
##
##        Bad     Good
## Good   < 2e-16 -
## Medium 5.1e-08 2.3e-15
##
## P value adjustment method: none
```

```r
# Probability of Type 1 error
1-0.95^(6)
```

```
## [1] 0.2649081
```

```r
# Bonferroni correction based on number of tests run
0.05/6
```

```
## [1] 0.008333333
```

The tests conducted include 1 MANOVA, 2 ANOVAs, and 3 t-tests. A one way MANOVA was performed to determine the effect of shelf location quality on number of car seat sales and the advertising budget for each store. The results reflected significant differences among the three different qualities of location for at least one of the dependent variables. (Pillai's trace = 0.38012, pseudo F(2,255)= 29.919, p < 0.0001).

Univariate ANOVAs were performed for each dependent variable as follow up tests to the MANOVA results. The dependent variable, sales of car seat units, was significant (F(2,255)= 73.50, p < 0.0001), but advertising budget did not have significant results (F(2,255)=0.742, p > 0.0001).

A post hoc analysis was conducted using pairwise comparisons to determine which type of shelving location significantly differed. Based on the results, all 3 types of shelving location (good, medium, bad) were significantly different in terms of the units of car seats sold.

The probability of a type 1 error was 0.264. The Bonferroni correction adjusted the significance level to 0.008, so the probability of type 1 errors is kept at 0.008.

```
# Randomization Test

# Observed F statistic
summary(aov(Sales ~ Urban, data = carseat1))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Urban         1   16.2  16.206   1.965  0.162
## Residuals   256 2111.2   8.247
```

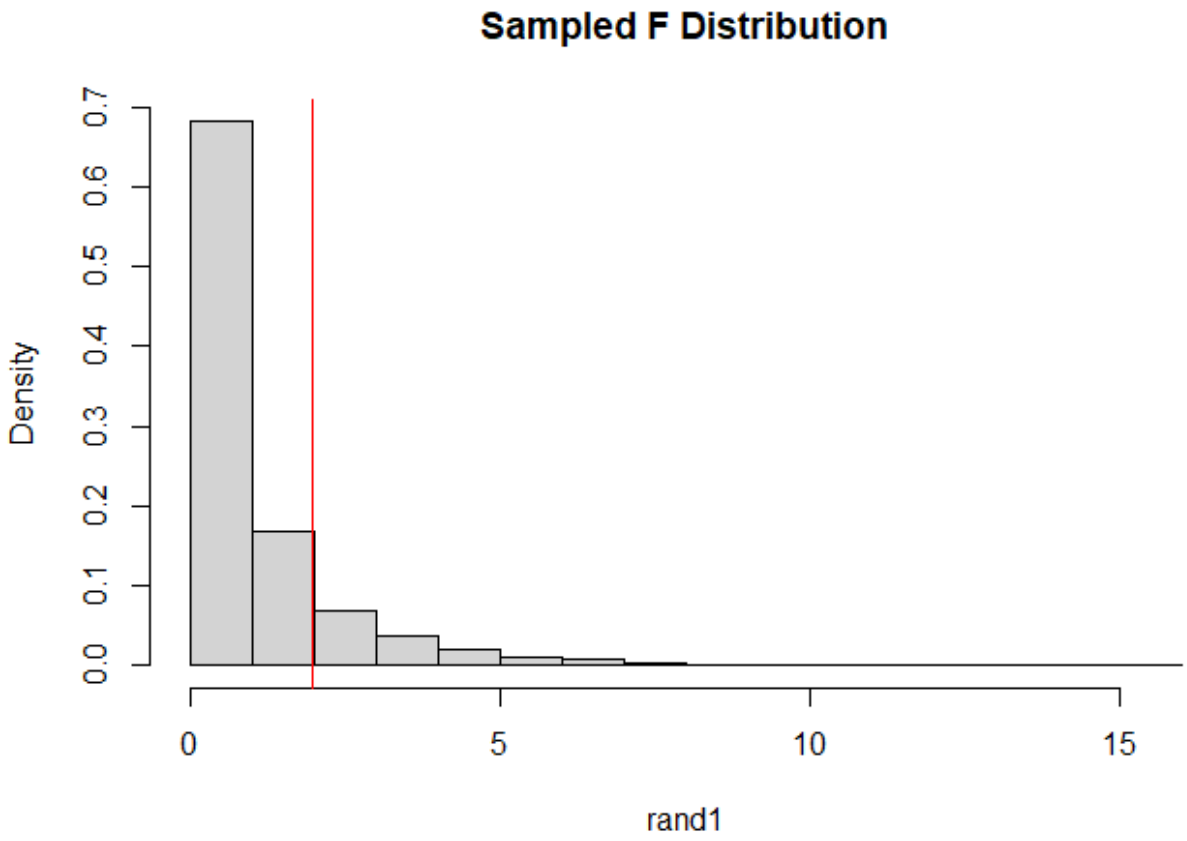
```
obs_F <- 1.965

#
rand1 <- replicate(5000,{
  # Randomly permute the response variable across Sales
  newcarseat <- carseat1 %>%
    mutate(Sales = sample(Sales))
  # Compute variation within groups
  SSW <- newcarseat %>%
    group_by(Urban) %>%
    summarize(SSW = sum((Sales - mean(Sales))^2)) %>%
    summarize(sum(SSW)) %>%
    pull
  # Compute variation between groups
  SSB <- newcarseat %>%
    mutate(mean = mean(Sales)) %>%
```

```
    group_by(Urban) %>%
    mutate(groupmean = mean(Sales)) %>%
    summarize(SSB = sum((mean - groupmean)^2)) %>%
    summarize(sum(SSB)) %>%
    pull
  # Compute the F-statistic (ratio of MSB and MSW)
  # df for SSB is 3 groups - 1 = 2
  # df for SSW is 258 observations - 2 groups = 256
  (SSB/1)/(SSW/256)
})

hist(rand1, prob=T, main = 'Sampled F Distribution'); abline(v = obs_F, col="red",add
```

```
## Warning in int_abline(a = a, b = b, h = h, v = v, untf = untf, ...): "add" is
## not a graphical parameter
```



Updated: The F statistic from the ANOVA test was 1.965. The histogram shows a left-skewed sampled null F distribution as opposed to the observed F statistic. The null hypothesis is unit sales are the same among both urban and rural locations. The alternative hypothesis is sales are significantly different among urban and rural locations. The p value

is 0.162 > 0.05 which means we fail to reject the null hypothesis. There is no significant difference between unit sales (in thousands) among urban and rural locations.

```
# Correlation coefficient
cor(carseat1$Sales, carseat1$Population)
```

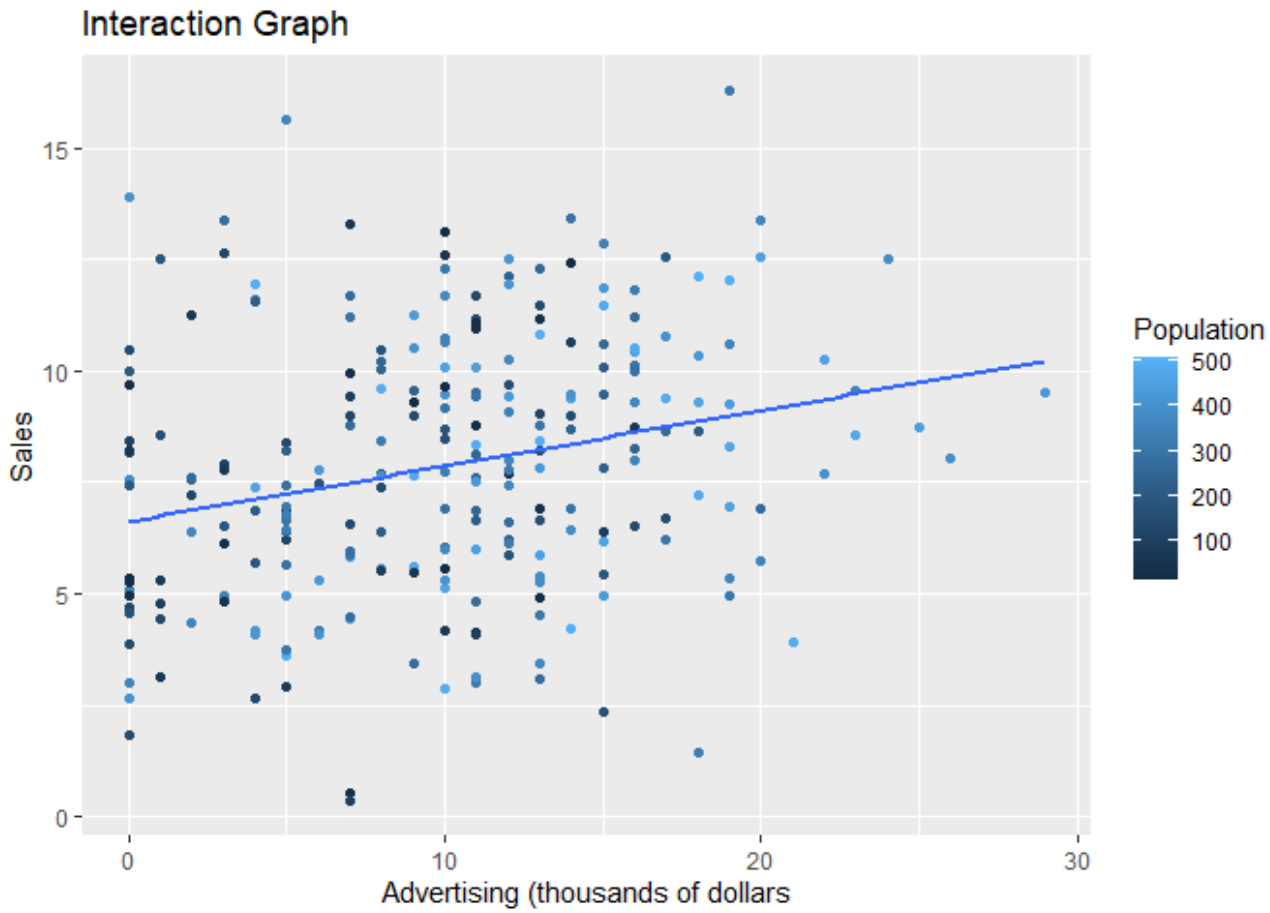
```
## [1] 0.06850591
```

```
#MLR
fit <- lm(Sales ~ Advertising * Population, data = carseat1)
summary(fit)
```

```
##
## Call:
## lm(formula = Sales ~ Advertising * Population, data = carseat1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4379 -1.9649 -0.1273  1.8044  8.4447
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)             6.641e+00  6.775e-01   9.802   <2e-16 ***
## Advertising             1.395e-01  7.033e-02   1.984   0.0483 *
## Population             -2.429e-04  2.427e-03  -0.100   0.9204
## Advertising:Population -3.449e-05  2.149e-04  -0.160   0.8726
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.797 on 254 degrees of freedom
## Multiple R-squared:  0.06621,    Adjusted R-squared:  0.05518
## F-statistic: 6.003 on 3 and 254 DF,  p-value: 0.0005755
```

```
#Visualization
ggplot(carseat1, aes(x = Advertising, y = Sales, color= Population)) +
  geom_point() +
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE) + xlab("Advertising (thousands of
```

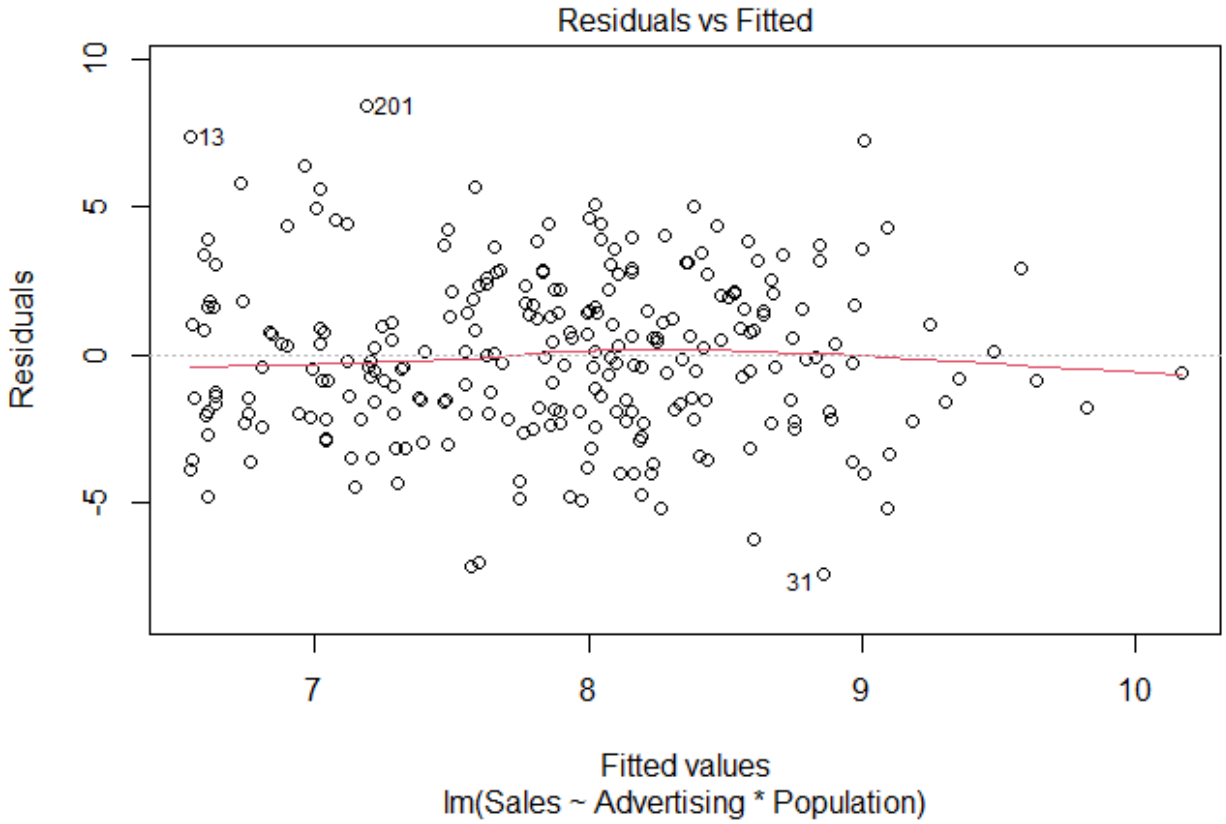
```
## `geom_smooth()` using formula 'y ~ x'
```



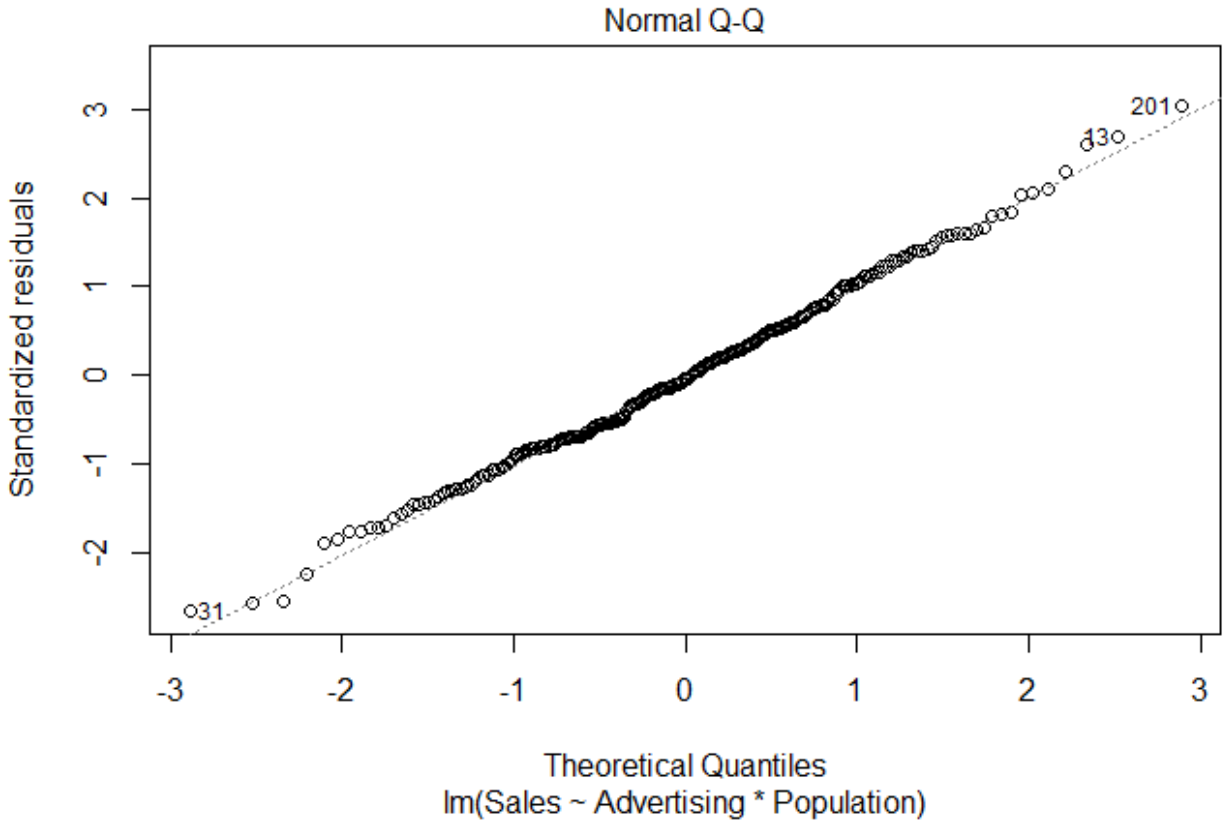
Sales = 6.641 + (0.1395 * Advertising) + (-0.0002429 * Population) + (0.00003449 * Advertise*Pop)

Based on the interaction plot between advertising and population, there appears to be a slightly negative interaction between advertising and population on the amount of sales. Sales (thousands of dollars) increased as the interaction between advertisement () decreased. The model explains a proportion of variance in Sales equal to 0.05518.

```
# Residuals against fitted values plot to check for any problematic patterns (nonline
plot(fit, which = 1)
```




```
# Q-Q plot to check for normality
plot(fit, which = 2)
```




```
library(sandwich)
```

```
## Warning: package 'sandwich' was built under R version 4.0.5
```

```
# Install a new package
# install.packages("lmtest")
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.0.5

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.0.5

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
# Breusch-Pagan test
# H0: The residuals have constant variance
bptest(fit)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit
## BP = 1.5997, df = 3, p-value = 0.6595
```

The residuals versus fitted values plot demonstrates a slight upwards funnel between 8-9. However, the red line follows the dashed line well indicating linearity. The Q-Q plot demonstrates linearity. The p value is greater than 0.05 indicating the null hypothesis is not rejected and there is homoscedasticity.

```
# Uncorrected standard errors
summary(fit)$coef
```

```
##                             Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)             6.640902e+00 0.6775236879  9.8017270 1.910296e-19
## Advertising             1.395251e-01 0.0703264794  1.9839626 4.833644e-02
## Population             -2.428770e-04 0.0024267934 -0.1000814 9.203586e-01
## Advertising:Population -3.448539e-05 0.0002148788 -0.1604877 8.726245e-01
```

```
# Robust Standard Errors
# install.packages("sandwich")
library(sandwich)
coeftest(fit, vcov = vcovHC(fit))
```

```
##
## t test of coefficients:
##
##                            Estimate  Std. Error t value Pr(>|t|)
## (Intercept)             6.6409e+00  6.5605e-01 10.1225  < 2e-16 ***
## Advertising             1.3953e-01  6.6182e-02  2.1082  0.03599 *
## Population             -2.4288e-04  2.5397e-03 -0.0956  0.92389
## Advertising:Population -3.4485e-05  2.0810e-04 -0.1657  0.86851
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Bootstrap SEs
samp_SEs <- replicate(5000, {
  # Bootstrap your data (resample observations)
  cs_data <- sample_frac(carseat1, replace = TRUE)
  # Fit regression model
  fitcar <- lm(Sales ~ Advertising * Population, data = carseat1)
  # Save the coefficients
  coef(fitcar)
})
```

The uncorrected standard errors for advertising, population and the interaction are 0.070, 0.002 and 0.0002 respectively. 0.6775 as the uncorrected intercept SE. The bootstrapped intercept SE is 0.656 which is less than the uncorrected. These were calculated to correct for any failed assumptions.

```
# Convert binary categorical variable into numeric values
carseat2 <- carseat1 %>% mutate(Loc_type=ifelse(Urban == "Yes",1,0))

view(carseat2)

# Log regression
cslogreg <- glm(Loc_type ~ Sales + Advertising, data = carseat2, family = binomial(li
summary(cslogreg)
```

```
##
## Call:
## glm(formula = Loc_type ~ Sales + Advertising, family = binomial(link = "logit"),
##     data = carseat2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7715  -1.4710   0.7616   0.8266   1.0494
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.41390    0.44139   3.203  0.00136 **
## Sales       -0.07501    0.05033  -1.490  0.13618
## Advertising  0.01347    0.02448   0.550  0.58206
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 305.51  on 257  degrees of freedom
## Residual deviance: 303.24  on 255  degrees of freedom
## AIC: 309.24
```

```
##
## Number of Fisher Scoring iterations: 4
```

```
# odds
exp(coef(cslogreg))
```

```
## (Intercept)       Sales Advertising
##   4.1119559   0.9277364   1.0135641
```

Based on the logistic regression results, both Sales and Advertising are statistically significant ($p < 0.05$). Every one-unit (in thousands) increase in Sales and Advertising (in thousands of dollars) increases the log odds of an urban location type by 0.93.

Every one-unit increase in Sales multiplies the odds of an urban location type by 0.928 and Every one-unit increase in Advertising multiplies the odds of an urban (i.e., the odds of malignancy increase by 153% for every additional unit of clump thickness).

```
# Predicted logit (log-odds) for sales
predict(cslogreg, newdata = data.frame(Sales = 10, Advertising = 10), type = "link")
```

```
##        1
## 0.798552
```

```
# Predicted odds
exp(predict(cslogreg, newdata = data.frame(Sales = 10, Advertising = 10), type = "lin
```

```
##        1
## 2.222321
```

```
# Predicted probability for Sales = 10 and Advertising =10
predict(cslogreg, newdata = data.frame(Sales = 10, Advertising = 10), type = "respons
```

```
##         1
## 0.6896646
```

For a store with car seat Sales amounting to $10000 and spending $10000 on Advertising, the log odds of it being in an Urban location are 0.799 while holding other predictors constant.

For a store with Sales and Advertising being $10000, the odds of it being in an Urban area are 2.222.

For a store with Sales and Advertising being $10000, the probability of the store being in an Urban area is 68.9%.

```
# Prediction variables to dataset
carseat2$prob <- predict(cslogreg, type = "response")

# Predicted outcome
# if the probability is greater than 0.7, the location is urban
carseat2$predicted <- ifelse(carseat2$prob > .7, "urban", "rural")
carseat2$logit <- predict(cslogreg)
view(carseat2)

# Confusion Matrix
table(truth = carseat2$Loc_type, prediction = carseat2$predicted)
```

```
##      prediction
## truth rural urban
##     0    27    45
##     1    46   140
```

```
# Accuracy
(27 + 140)/258
```

```
## [1] 0.6472868
```

```
# Sensitivity (TPR)
140/186
```

```
## [1] 0.7526882
```

```
# Specificity (TNR)
27/72
```

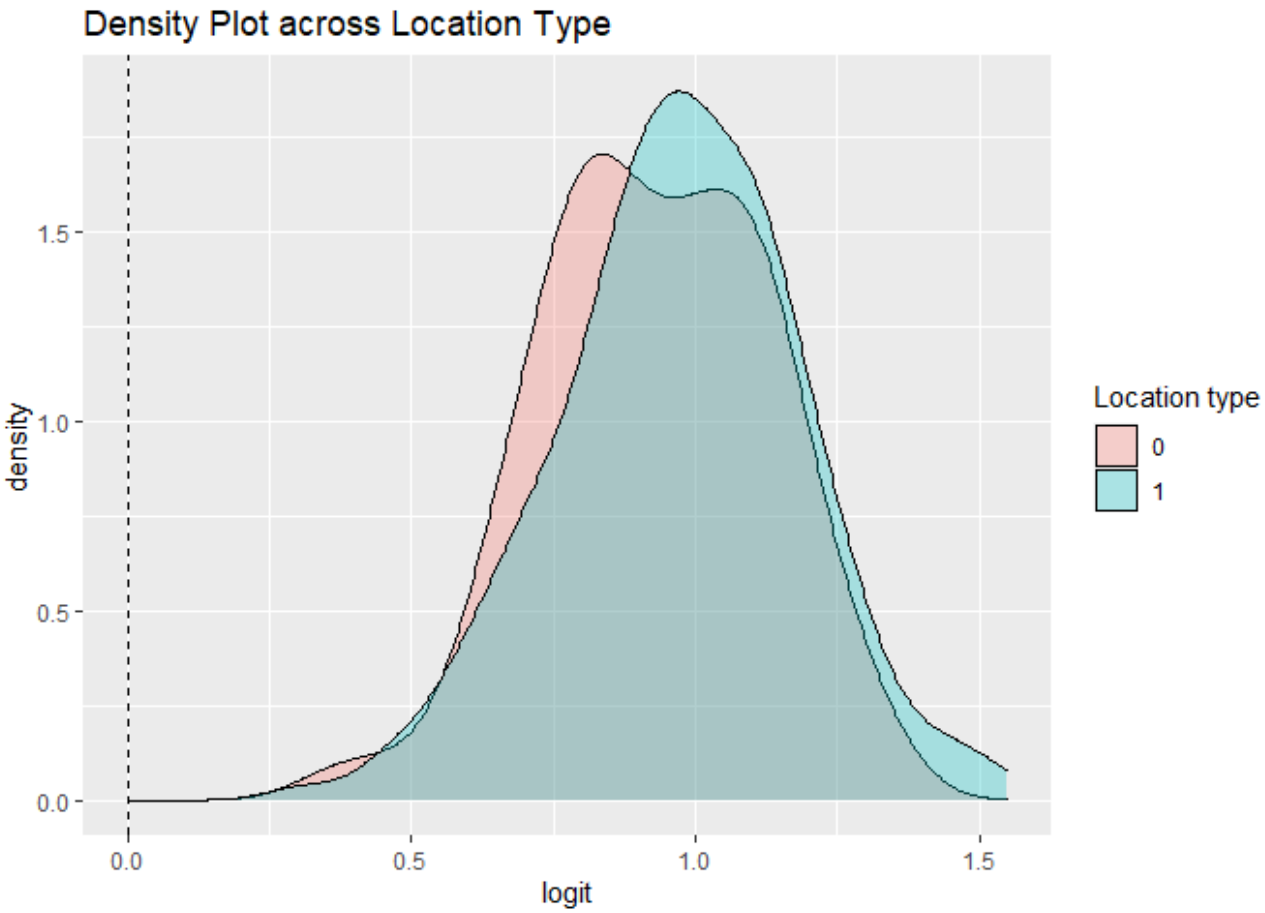
```
## [1] 0.375
```

```
# Precision
140/186
```

```
## [1] 0.7526882
```

The confusion matrix results demonstrate the logistic model has a 64% accuracy, the proportion of true positives is 0.75, the proportion of true negatives is 0.375 and the proportion of a true positive prediction is 0.753.

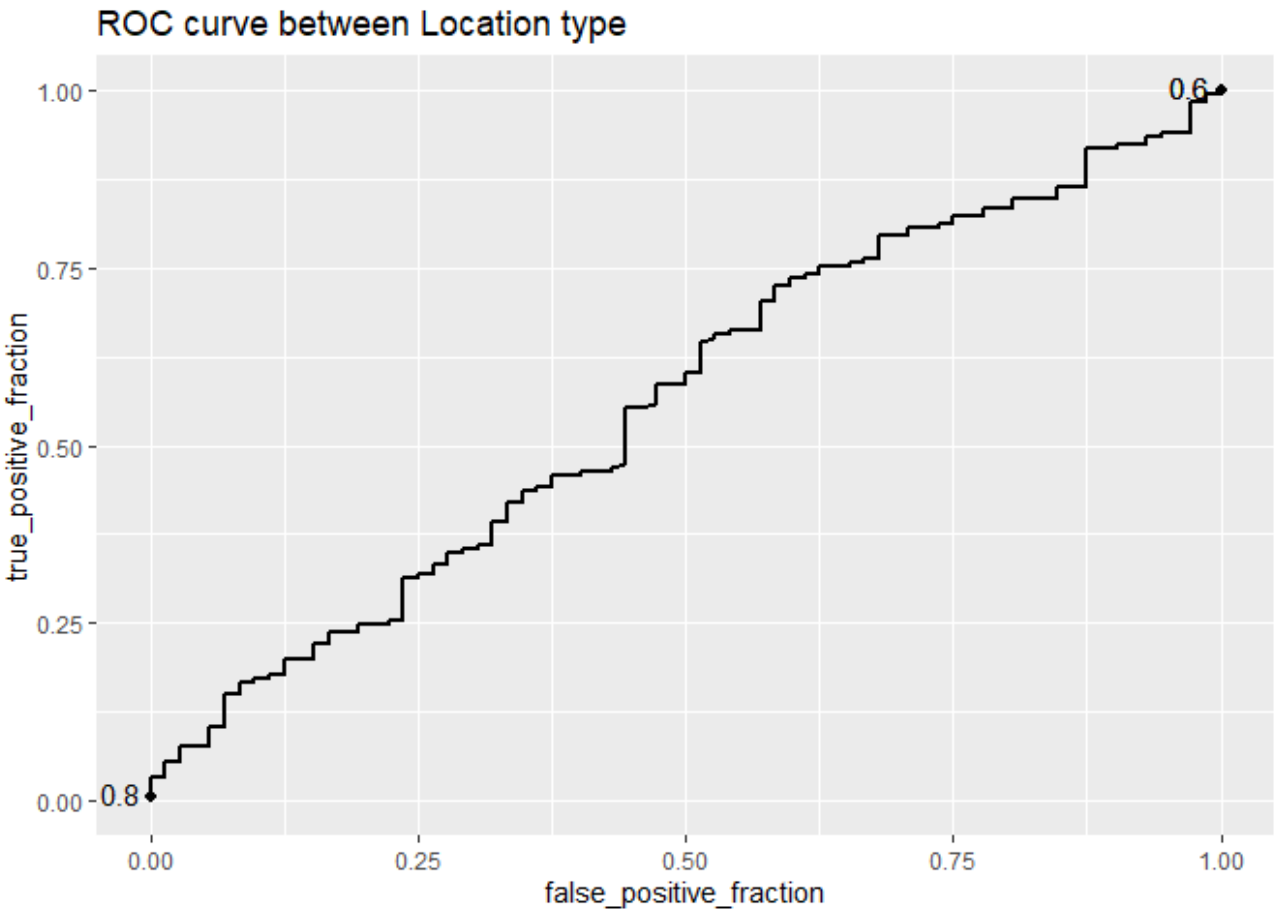
```
# Logit Model
ggplot(carseat2, aes(logit, fill = as.factor(Loc_type))) +
  geom_density(alpha = .3) +
  geom_vline(xintercept = 0, lty = 2) +
  labs(fill = "Location type") + ggtitle('Density Plot across Location Type')
```




```
# ROC curve

# Call the library plotROC
library(plotROC)

# Plot ROC depending on values of y and its probabilities displaying some cutoff valu
ROCplot1 <- ggplot(carseat2) +
  geom_roc(aes(d = Loc_type, m = prob), cutoffs.at = list(0.1, 0.5, 0.9)) + ggtitle('
ROCplot1
```




```
# AUC
calc_auc(ROCplot1)
```

```
##   PANEL group       AUC
## 1     1    -1 0.5611932
```

The density plot demonstrates significant overlap which is representative of the stores that were misclassified into urban or rural location types.

The ROC curve demonstrates the model relatively follows the equation FPR = TPR. The slope of the curve follows an accuracy proportion of roughly 0.64. However, this is consistent with the test not being the best model for predicting true positives or true negatives for location type.

The AUC was found to be 0.561 which is consistent with a model that does not separate the classes well. In this case, there is a 56% chance the model will be able to distinguish between urban and rural location types.