Module 4 CT Option 2

Scott Miner

3/7/2020

# Logistic Regression on Riding Mower Ownership

## Exploratory Analysis

The assignment asks the student to summarize the ***RidingMowers.csv*** file. We begin by reading the file into a data frame and getting the dimensions of the data frame.

### Read the .CSV file

# read the .csv file   
mowers.df <- read.csv("RidingMowers.csv")

### View Dimensions

We see the data frame has 24 rows and 3 columns.

# get the dimensions of the data frame  
dimft <- flextable(as.data.frame(cbind(c("Rows",  
 "Columns"),  
 dim(mowers.df))))  
dimft <- set\_header\_labels(dimft, V1 = "Dimension",  
 V2 = "Totals")  
dimft <- autofit(dimft)  
dimft

| Dimension | Totals |
| --- | --- |
| Rows | 24 |
| Columns | 3 |

### View Classes

Income and Lot\_Size are numeric variables, while Ownership is a factor variable.

# view variable classes  
classesft <- flextable(as.data.frame(cbind(c("Income",  
 "Lot Size",  
 "Ownership"),   
 sapply(mowers.df, class))))  
classesft <- autofit(classesft)  
classesft <- set\_header\_labels(classesft, V1 = "Variable",  
 V2 = "Class")  
classesft

| Variable | Class |
| --- | --- |
| Income | numeric |
| Lot Size | numeric |
| Ownership | factor |

### Unique Values

Income has 22 unique values, Lot\_Size 18, and Ownership 2.

# get the aggregated counts of unique values in each column  
uniqueCountsFt <- flextable(as.data.frame(cbind(c(colnames(mowers.df)),  
 sapply(sapply(mowers.df,  
 table),  
 length))))  
uniqueCountsFt <- set\_header\_labels(uniqueCountsFt,  
 V1 = "Variable",  
 V2 = "Unique Values")  
uniqueCountsFt <- autofit(uniqueCountsFt)  
uniqueCountsFt

| Variable | Unique Values |
| --- | --- |
| Income | 22 |
| Lot\_Size | 18 |
| Ownership | 2 |

### View Records (head() and tail() Functions)

We can view the first 6 and last 6 records using the head() and tail() functions, respectively. #### head() Function

# The head function displays the first 6 rows  
head(mowers.df) %>%  
 flextable() %>%  
 autofit()

| Income | Lot\_Size | Ownership |
| --- | --- | --- |
| 60.0 | 18.4 | Owner |
| 85.5 | 16.8 | Owner |
| 64.8 | 21.6 | Owner |
| 61.5 | 20.8 | Owner |
| 87.0 | 23.6 | Owner |
| 110.1 | 19.2 | Owner |

#### tail() Function

To view that last 6 records, we use the tail() function.

# The tail function displays the last 6 rows  
tail(mowers.df) %>%  
 flextable() %>%  
 autofit()

| Income | Lot\_Size | Ownership |
| --- | --- | --- |
| 59.4 | 16.0 | Nonowner |
| 66.0 | 18.4 | Nonowner |
| 47.4 | 16.4 | Nonowner |
| 33.0 | 18.8 | Nonowner |
| 51.0 | 14.0 | Nonowner |
| 63.0 | 14.8 | Nonowner |

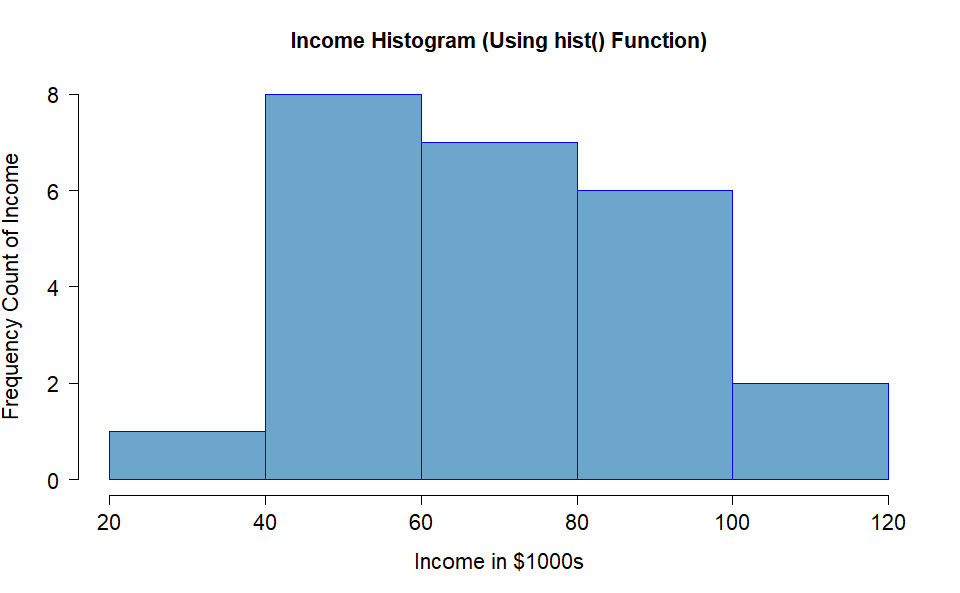
### Histograms

Let’s view histograms for the numeric variables (Income and Lot\_Size).

#### hist for Income

First we’ll use the hist() function.

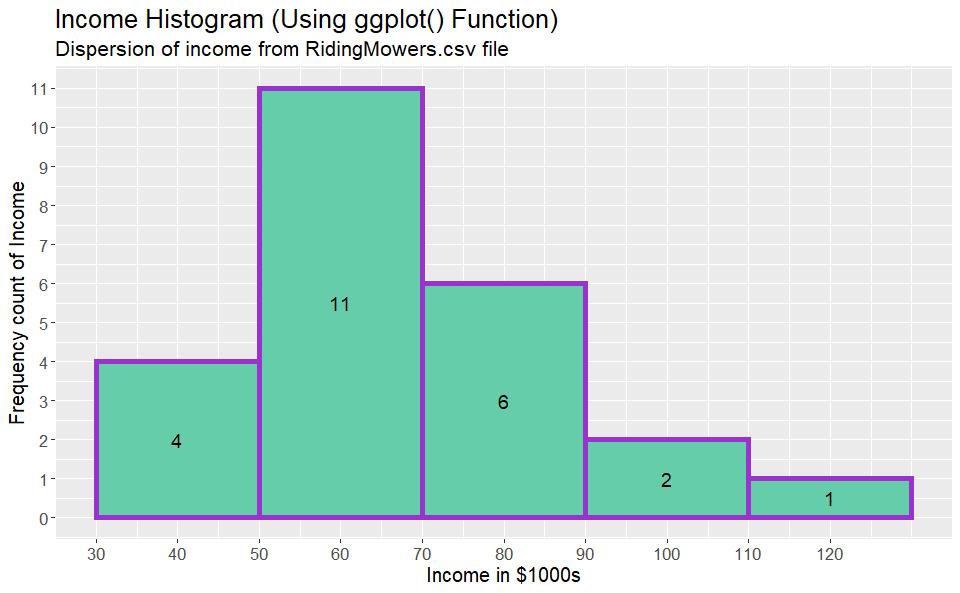
# using the hist() function to examine Income  
hist(mowers.df$Income  
 ,main="Income Histogram (Using hist() Function)"  
 ,xlab ="Income in $1000s"  
 ,ylab = "Frequency Count of Income"  
 ,border = "blue"  
 ,col = "skyblue3"  
 ,breaks = 5  
 ,las = 1  
 ,cex.lab=1.37  
 ,cex.axis=1.37  
 ,cex.main=1.37)



#### ggplot for Income

If we’d prefer, we can also use the ggplot() function combined with the geom\_histogram() function.

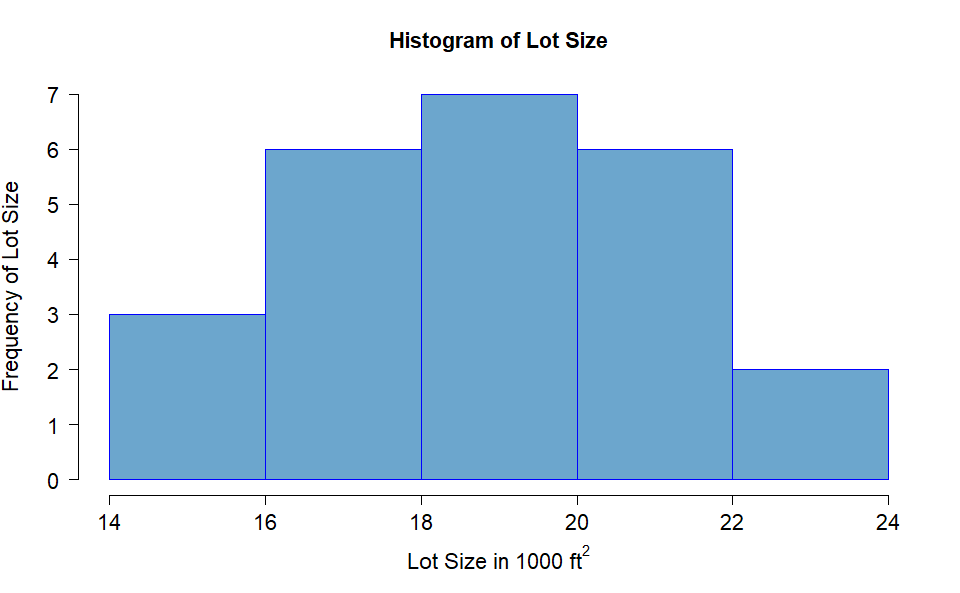
ggplot(mowers.df,aes(Income)) +  
 geom\_histogram(binwidth = 20, fill = "aquamarine3", color = "darkorchid3", size = 2) +  
 scale\_x\_continuous(breaks=seq(20,120,10)) +  
 scale\_y\_continuous(breaks=seq(0,15,1)) +  
 labs(title = "Income Histogram (Using ggplot() Function)",  
 subtitle = "Dispersion of income from RidingMowers.csv file",  
 y = "Frequency count of Income",  
 x = "Income in $1000s") +  
 theme(axis.title = element\_text(size = 15),  
 axis.text = element\_text(size = 13),  
 title = element\_text(size = 16)) +  
 stat\_bin(binwidth = 20, geom="text", color="black", aes(label=..count..), size=5,  
 position=position\_stack(vjust = 0.5), na.rm = TRUE)



#### hist for Lot\_Size

Now, let’s use the hist() function to examine Lot\_Size.

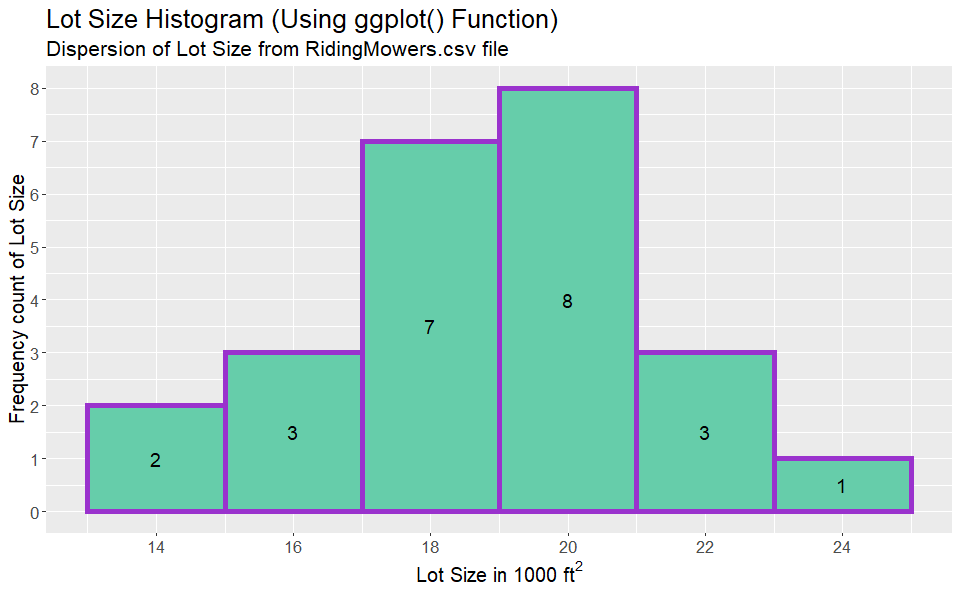
# using the hist() function to examine Lot\_Size  
hist(mowers.df$Lot\_Size  
 ,main="Histogram of Lot Size"  
 ,xlab = TeX("Lot Size in 1000 $ft ^ 2$")  
 ,ylab = "Frequency of Lot Size"  
 ,border = "blue"  
 ,col = "skyblue3"  
 ,breaks = 5  
 ,las = 1  
 ,cex.lab=1.37  
 ,cex.axis=1.37  
 ,cex.main=1.37)



#### ggplot for Lot\_Size

Next, we use ggplot() and geom\_histogram() to plot the Lot\_Size.

ggplot(mowers.df,aes(Lot\_Size)) +  
 geom\_histogram(binwidth = 2, fill = "aquamarine3", color = "darkorchid3", size = 2) +  
 scale\_x\_continuous(breaks=seq(14,24,2)) +  
 scale\_y\_continuous(breaks=seq(0,24,1)) +  
 labs(title = "Lot Size Histogram (Using ggplot() Function)",  
 subtitle = "Dispersion of Lot Size from RidingMowers.csv file",  
 y = "Frequency count of Lot Size",  
 x = TeX("Lot Size in 1000 $ft ^ 2$")) +  
 theme(axis.title = element\_text(size = 15),  
 axis.text = element\_text(size = 13),  
 title = element\_text(size = 16)) +  
 stat\_bin(binwidth = 2, geom="text", color="black", aes(label=..count..), size=5,  
 position=position\_stack(vjust = 0.5), na.rm = TRUE)



### Summarize Data by Ownership

What if we want to compare the average income levels between owners and non-owners?

# requires library(dplyr)  
# compare the averge income of owners vs. nonowners  
mowers.df %>%  
 group\_by(Ownership) %>%  
 summarise(  
 Average\_Income = round(mean(Income, na.rm = TRUE),2),  
 Average\_Lot\_Size = round(mean(Lot\_Size, na.rm = TRUE),2),  
 Count = n()  
 ) %>%  
 flextable() %>%  
 autofit()

| Ownership | Average\_Income | Average\_Lot\_Size | Count |
| --- | --- | --- | --- |
| Nonowner | 57.40 | 17.63 | 12 |
| Owner | 79.47 | 20.27 | 12 |

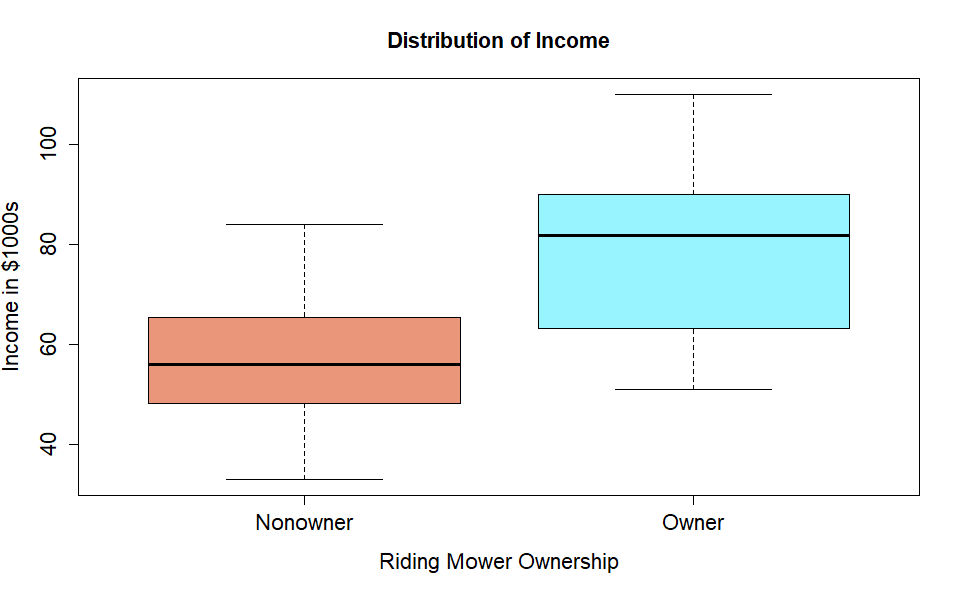
### Vertical Boxplots

Let’s visualize these averages via side-by-side boxplots.

First, we use the boxplot() function.

#### Vert. boxplot for Income

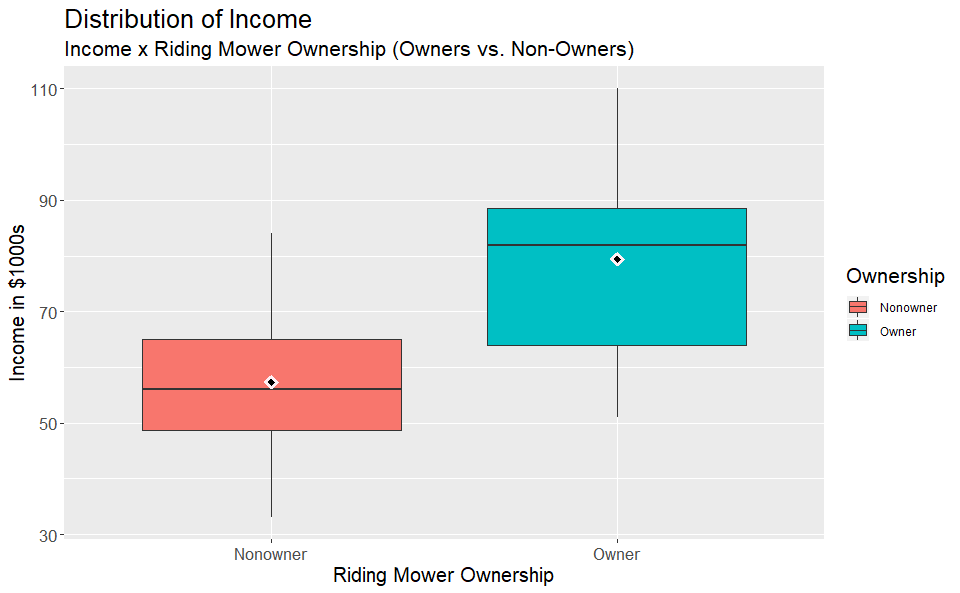
# side-by-side boxplot of Income for owners vs. non-owners  
# vertical boxplot  
  
boxPlotColors <- ifelse(levels(mowers.df$Ownership)=="Owner",  
 "cadetblue1",  
 "darksalmon")  
  
boxplot(mowers.df$Income ~ mowers.df$Ownership,  
 xlab = "Riding Mower Ownership",  
 ylab = "Income in $1000s",  
 col=boxPlotColors,  
 main="Distribution of Income",  
 las = 0,  
 cex.axis = 1.37,  
 cex.main = 1.37,  
 cex.lab = 1.37)



#### Vert. ggplot for Income

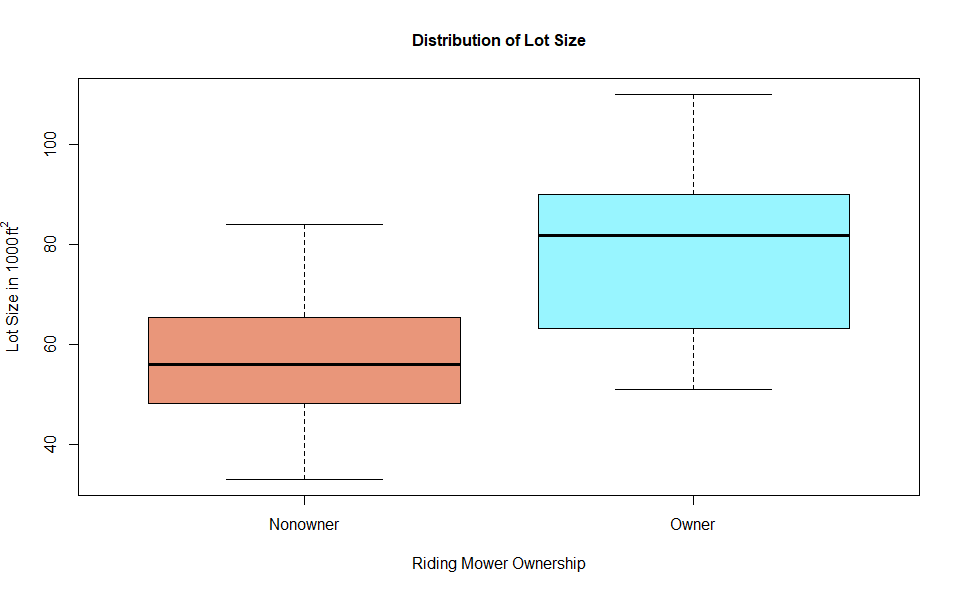
Additionally, we can use the ggplot2 library.

# requires library(ggplot2)  
# alternative boxplot using ggplot (vertical boxplot)  
ggplot(mowers.df, aes(Ownership, Income, fill = Ownership)) +  
 geom\_boxplot() +   
 stat\_summary(fun.y=mean, geom="point", shape=23, size=2,  
 color="lavenderblush", fill="black", stroke=2) +   
 labs(x = "Riding Mower Ownership", title="Distribution of Income",  
 y = "Income in $1000s",  
 subtitle="Income x Riding Mower Ownership (Owners vs. Non-Owners)") +  
 theme(axis.title = element\_text(size = 15),  
 axis.text = element\_text(size = 13),  
 title = element\_text(size = 16))



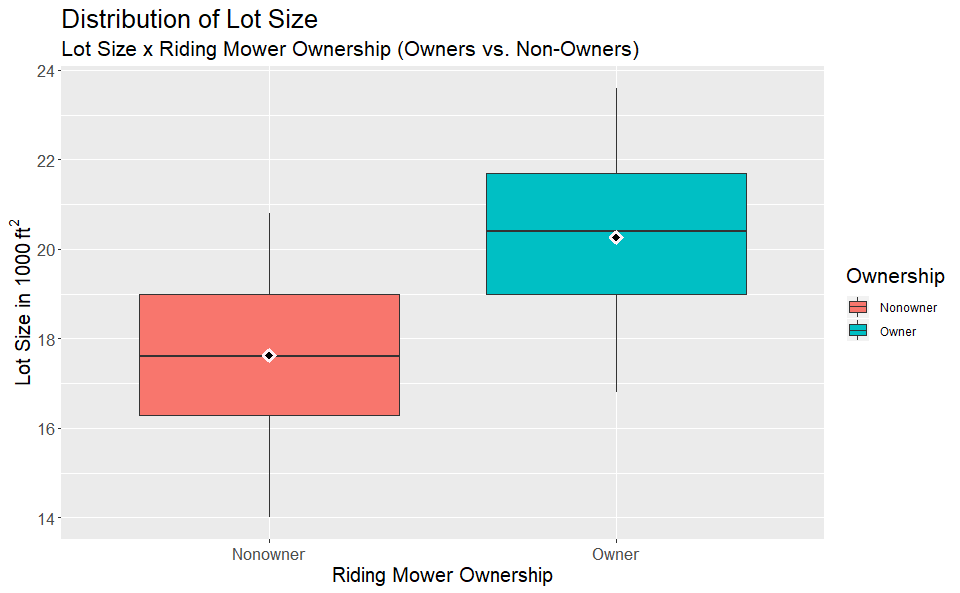
#### Vert. boxplot for Lot\_Size

# side-by-side boxplot of Lot Size for owners vs. non-owners  
# vertical boxplot  
  
boxPlotColors <- ifelse(levels(mowers.df$Ownership)=="Owner",  
 "cadetblue1",  
 "darksalmon")  
  
boxplot(mowers.df$Income ~ mowers.df$Ownership,  
 xlab = "Riding Mower Ownership",  
 ylab = TeX("Lot Size in 1000 $ft ^ 2$"),  
 col=boxPlotColors,  
 main="Distribution of Lot Size",  
 las = 0,  
 cex.axis = 1,  
 cex.main = 1,  
 cex.lab = 1)



#### Vert. ggplot for Lot\_Size

# requires library(ggplot2)  
# alternative boxplot using ggplot (vertical boxplot)  
ggplot(mowers.df, aes(Ownership, Lot\_Size, fill = Ownership)) +  
 geom\_boxplot() +   
 stat\_summary(fun.y=mean, geom="point", shape=23, size=2,  
 color="lavenderblush", fill="black", stroke=2) +   
 labs(x = "Riding Mower Ownership", title="Distribution of Lot Size",  
 y = TeX("Lot Size in 1000 $ft ^ 2$"),   
 subtitle="Lot Size x Riding Mower Ownership (Owners vs. Non-Owners)") +  
 theme(axis.title = element\_text(size = 15),  
 axis.text = element\_text(size = 13),  
 title = element\_text(size = 16))

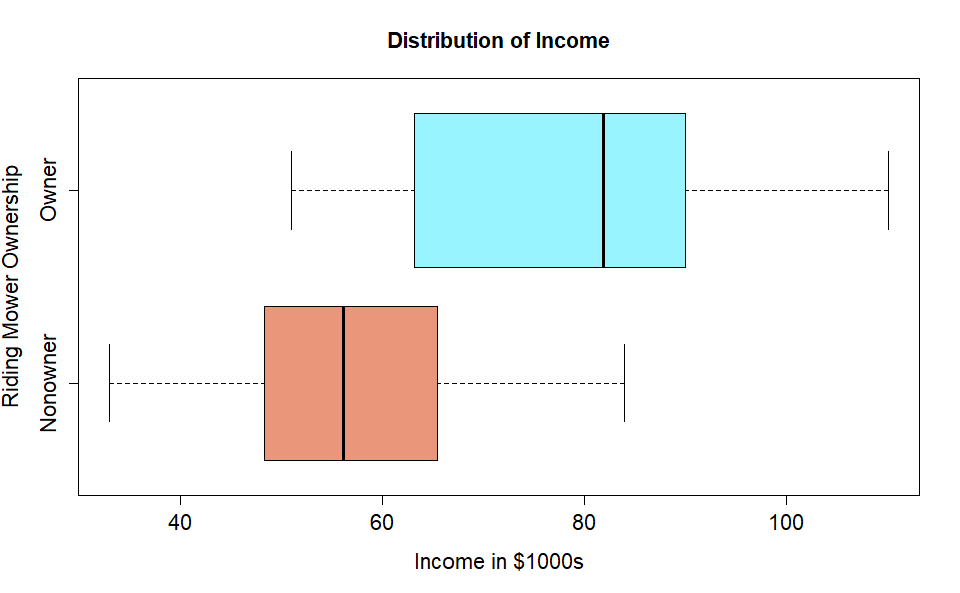


### Horizontal Boxplots

If we prefer, we can plot these boxplots horizontally. Using the base R boxplot() function. We add the horizontal = TRUE argument).

#### Horiz. boxplot for Income

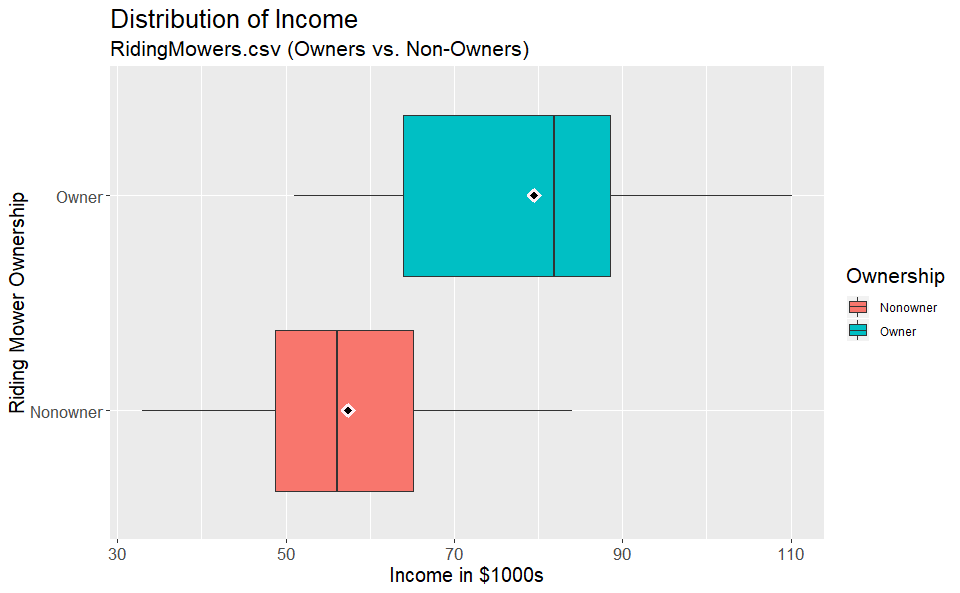
# side-by-side boxplot of Income for owners vs. non-owners (horizontal boxplot)  
  
boxPlotColors <- ifelse(levels(mowers.df$Ownership)=="Owner",  
 "cadetblue1",  
 "darksalmon")  
  
boxplot(mowers.df$Income ~ mowers.df$Ownership,  
 ylab = "Riding Mower Ownership",  
 xlab = "Income in $1000s",  
 col=boxPlotColors,  
 main="Distribution of Income",  
 horizontal = TRUE,  
 las = 0,  
 cex.axis = 1.37,  
 cex.main = 1.37,  
 cex.lab = 1.37)



#### Horiz. ggplot for Income

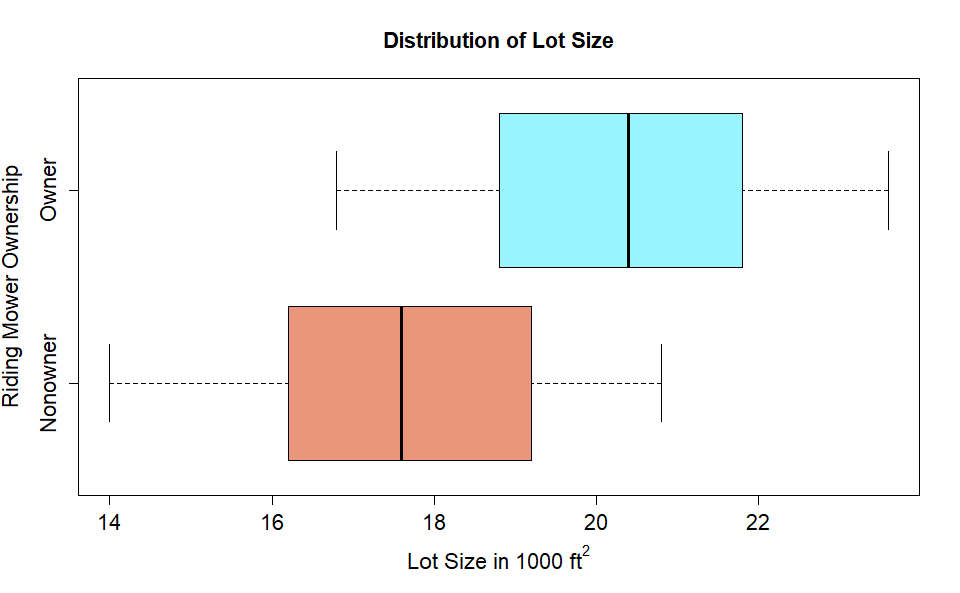
Using the ggplot() function, we add the coord\_flip() argument.

# requires library(ggplot2)  
# alternative boxplot using ggplot (horizontal boxplot)  
ggplot(mowers.df, aes(Ownership, Income, fill = Ownership)) +   
 geom\_boxplot() +   
 stat\_summary(fun.y=mean, geom="point", shape=23, size=2,  
 color="lavenderblush", fill="black", stroke=2) +   
 labs(x = "Riding Mower Ownership", title="Distribution of Income",  
 y = "Income in $1000s",  
 subtitle="RidingMowers.csv (Owners vs. Non-Owners)") +   
 coord\_flip() +  
 theme(axis.title = element\_text(size = 15),  
 axis.text = element\_text(size = 13),  
 title = element\_text(size = 16))



#### Horiz. boxplot for Lot\_Size

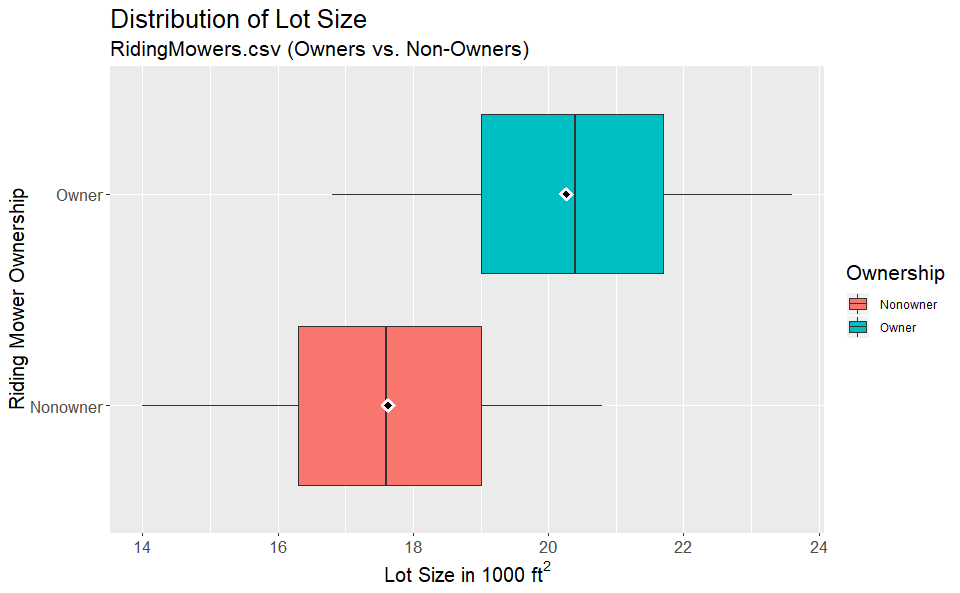
# side-by-side boxplot of Lot Size for owners vs. non-owners (horizontal boxplot)  
  
boxPlotColors <- ifelse(levels(mowers.df$Ownership)=="Owner",  
 "cadetblue1",  
 "darksalmon")  
  
boxplot(mowers.df$Lot\_Size ~ mowers.df$Ownership,   
 ylab = "Riding Mower Ownership",  
 xlab = TeX("Lot Size in 1000 $ft ^ 2$"),  
 col=boxPlotColors,  
 main="Distribution of Lot Size",  
 horizontal = TRUE,  
 las = 0,  
 cex.axis = 1.37,  
 cex.main = 1.37,  
 cex.lab = 1.37)



#### Horiz. ggplot for Lot\_Size

Again, using the ggplot() function, we add the coord\_flip() argument.

# requires library(ggplot2)  
# alternative boxplot using ggplot (horizontal boxplot)  
ggplot(mowers.df, aes(Ownership, Lot\_Size, fill = Ownership)) +   
 geom\_boxplot() +   
 stat\_summary(fun.y=mean, geom="point", shape=23, size=2,  
 color="lavenderblush", fill="black", stroke=2) +   
 labs(x = "Riding Mower Ownership", title="Distribution of Lot Size",  
 y = TeX("Lot Size in 1000 $ft ^ 2$"),  
 subtitle="RidingMowers.csv (Owners vs. Non-Owners)") +   
 coord\_flip() +  
 theme(axis.title = element\_text(size = 15),  
 axis.text = element\_text(size = 13),  
 title = element\_text(size = 16))



### Descriptive Statistics

We’ve yet to view any summarized, descriptive statistics. Let’s view some now.

We see Income has an average of 68.44, a minimum of 33, and a maximum of 110.1, along with other descriptive statistics.

We view the same statistics for Lot\_Size as well.

# summarized, descriptive statistices  
# percent requires package scales  
flextable((data.frame(cbind(c("Income", "Lot\_Size"),  
 mean = round(sapply(Filter(is.numeric, mowers.df),  
 mean, na.rm = TRUE), 2),  
 sd = round(sapply(Filter(is.numeric, mowers.df),   
 sd, na.rm = TRUE), 2),  
 min = sapply(Filter(is.numeric, mowers.df),   
 min, na.rm = TRUE),  
 max = sapply(Filter(is.numeric, mowers.df),   
 max, na.rm = TRUE),  
 median = sapply (Filter(is.numeric, mowers.df),  
 median, na.rm = TRUE),  
 length = sapply(Filter(is.numeric, mowers.df),  
 length),  
 mis.val = sapply(Filter(is.numeric, mowers.df),   
 function(x) sum(is.na(x))),  
 cum.ratio = percent(cumsum(sapply(Filter(is.numeric,mowers.df),   
 function(x) mean(is.na(x))))))))) %>%  
 autofit() %>%  
 set\_header\_labels(V1 = "Variable")

| Variable | mean | sd | min | max | median | length | mis.val | cum.ratio |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Income | 68.44 | 19.79 | 33 | 110.1 | 64.8 | 24 | 0 | 0% |
| Lot\_Size | 18.95 | 2.43 | 14 | 23.6 | 19 | 24 | 0 | 0% |

## Running a Logistic Model

Next, we will run a logistic regression model on the data, following the process outlined in section 10.3 of our text.

We are instructed to not partition the data, and to instead use the whole dataset to fit the model.

We will model the Ownership attribute as a function of Income and Lot\_Size.

### Logistic Regression Model

# create reference categories  
mowers.df$isOwner <- 1 \* (mowers.df$Ownership == "Owner")  
  
# create selected variables for regression model  
selected.var <- c(1:3)  
  
# running logistic regression  
# use glm() (general linear model) with family = "binomial" to fit a logistic regression  
logit.reg <- glm(Ownership ~ ., data = mowers.df[, selected.var], family = "binomial")  
options(scipen=999)  
summary(logit.reg)

##   
## Call:  
## glm(formula = Ownership ~ ., family = "binomial", data = mowers.df[,   
## selected.var])  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.74044 -0.29685 0.00439 0.44750 1.86821   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -25.9382 11.4871 -2.258 0.0239 \*  
## Income 0.1109 0.0543 2.042 0.0412 \*  
## Lot\_Size 0.9638 0.4628 2.083 0.0373 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 33.271 on 23 degrees of freedom  
## Residual deviance: 15.323 on 21 degrees of freedom  
## AIC: 21.323  
##   
## Number of Fisher Scoring iterations: 6

### Equation

Momentarily ignoring -values, a model for the two predictors has the following estimated logistic equation

For the continuous numerical predictors (Income and Lot\_Size), positive coefficients indicate that a higher value on that predictor is associated with a higher probability of owning a riding lawn mower. This makes sense, given that our predictor variables are Income and Lot\_Size. We would expect individuals with more income and larger yards to be more likely to purchase a riding lawn mower.

### Viewing Odds

* By viewing the odds, we can determine the percent increases for a one-unit increase in each predictor variable of being a riding lawnmower owner, holding all other variables at fixed values.
  + For Instance, the coefficient for **income** tells us that, holding **lot\_size** constant, we see a 12% increase in the odds of being a riding lawnmower owner for a one-unit increase in Income because exp(0.110859) = 1.117.
  + Similarly, we see a 162% increase in the odds of being a riding lawnmower owner for a one-unit increase in *lot\_size*, because exp(0.96378) = 2.62.
    - This is supported by the -values in the logistic regression model, which show lot size as having a more significant influence (-value = 0.0373) as compared to income (-value = 0.0412). Both -values are less than the 0.05 significance level.

oddsFt <- flextable(data.frame(cbind(c("Int", "Inc", "LS"), summary(logit.reg)$coefficients), odds = exp(coef(logit.reg))))  
oddsFt <- autofit(oddsFt)  
oddsFt <- set\_header\_labels(oddsFt, V1 = "Predictor")  
  
FitFlextableToPage <- function(ft, pgwidth = 6){  
  
 ft\_out <- ft %>% autofit()  
  
 ft\_out <- width(ft\_out, width = dim(ft\_out)$widths\*pgwidth /(flextable\_dim(ft\_out)$widths))  
 return(ft\_out)  
}  
  
FitFlextableToPage(oddsFt)

| Predictor | Estimate | Std..Error | z.value | Pr...z.. | odds |
| --- | --- | --- | --- | --- | --- |
| Int | -25.9382316052441 | 11.487069350433 | -2.25803734738195 | 0.0239433311139876 | 0.00000000000543462 |
| Inc | 0.110858586332185 | 0.0542962352056294 | 2.04173615191448 | 0.041177710643785 | 1.11723690311449020 |
| LS | 0.963779128889001 | 0.462784201282357 | 2.08256705008167 | 0.0372907052536842 | 2.62158508441392701 |

### Evaluating Model Performance

Popular measures to evaluate model performance include confusion matrices and lift charts.

* The *lift* over the base curve of a **lift chart** indicates for a given number of cases (read on the x-axis) the additional cases we can identify by using the model.
* The **confusion matrix** gives a sense of classification accuracy.

#### Predict Function

The Predict() function uses our model to create predictions from our dataset. We store the predictions in the logit.reg.pred variable.

* We can view our predictions as probabilities side-by-side with the actual ownership of the record.
  + Probabilities that are below the 0.5 cutoff threshold, our model will predict as non-owners.
  + Probabilities above the 0.5 cutoff threshold, our model predicts as owners.
  + When viewing the data side-by-side with actual values, we see our model misclassifies two records in each category.

# R Code for the predict() function.  
# To determine probabilities in logistic regression, set type = "response"  
logit.reg.pred <- predict(logit.reg, mowers.df[selected.var], type = "response")  
  
# View the actual classifications side-by-side with probabilities from our model  
flextable(data.frame(  
 lot\_size = mowers.df$Lot\_Size,  
 income = mowers.df$Income,  
 actual = mowers.df$Ownership,  
 predicted = logit.reg.pred[])) %>%  
 autofit()

| lot\_size | income | actual | predicted |
| --- | --- | --- | --- |
| 18.4 | 60.0 | Owner | 0.174627386 |
| 16.8 | 85.5 | Owner | 0.433316360 |
| 21.6 | 64.8 | Owner | 0.887258019 |
| 20.8 | 61.5 | Owner | 0.716299107 |
| 23.6 | 87.0 | Owner | 0.998424574 |
| 19.2 | 110.1 | Owner | 0.991606411 |
| 17.6 | 108.0 | Owner | 0.952438793 |
| 22.4 | 82.8 | Owner | 0.992072920 |
| 20.0 | 69.0 | Owner | 0.728414536 |
| 20.8 | 93.0 | Owner | 0.988087969 |
| 22.0 | 51.0 | Owner | 0.714776563 |
| 20.0 | 81.0 | Owner | 0.910266763 |
| 19.6 | 75.0 | Nonowner | 0.780097268 |
| 20.8 | 52.8 | Nonowner | 0.490428078 |
| 17.2 | 64.8 | Nonowner | 0.101780659 |
| 20.4 | 43.2 | Nonowner | 0.184215126 |
| 17.6 | 84.0 | Nonowner | 0.583316338 |
| 17.6 | 49.2 | Nonowner | 0.028707298 |
| 16.0 | 59.4 | Nonowner | 0.019213050 |
| 18.4 | 66.0 | Nonowner | 0.291516753 |
| 16.4 | 47.4 | Nonowner | 0.007558038 |
| 18.8 | 33.0 | Nonowner | 0.015354839 |
| 14.0 | 51.0 | Nonowner | 0.001121982 |
| 14.8 | 63.0 | Nonowner | 0.009101169 |

#### Confusion Matrix

We can confirm our results by viewing a confusion matrix, which gives a sense of the classification accuracy.

In the output, we see the two results that our model misclassified from each category.

# Requires Library(caret)  
# Confusion Matrix  
confusionMatrix(as.factor(ifelse(logit.reg.pred > 0.5, 1, 0)),  
 as.factor(mowers.df$isOwner))

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 10 2  
## 1 2 10  
##   
## Accuracy : 0.8333   
## 95% CI : (0.6262, 0.9526)  
## No Information Rate : 0.5   
## P-Value [Acc > NIR] : 0.0007719   
##   
## Kappa : 0.6667   
##   
## Mcnemar's Test P-Value : 1.0000000   
##   
## Sensitivity : 0.8333   
## Specificity : 0.8333   
## Pos Pred Value : 0.8333   
## Neg Pred Value : 0.8333   
## Prevalence : 0.5000   
## Detection Rate : 0.4167   
## Detection Prevalence : 0.5000   
## Balanced Accuracy : 0.8333   
##   
## 'Positive' Class : 0   
##

#### Lift Chart

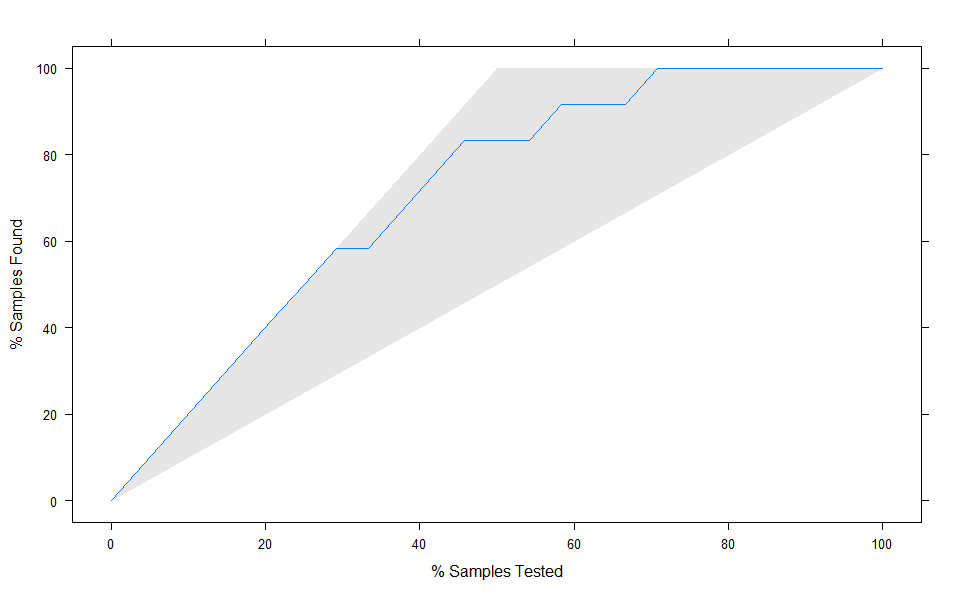
Finally, we plot a lift chart.

* The lift chart serves as a reference line.
  + It provides a benchmark against which we can evaluate the ranking performance of the model.
  + Our model appears to predict better than the random benchmark.
* To read the lift chart, for a given number of records on the x-axis, the lift curve value on the y-axis tells us how much better we are doing compared to random assignment.

#### Lift Chart Using Caret Library

First, we plot a lift chart using the caret library.

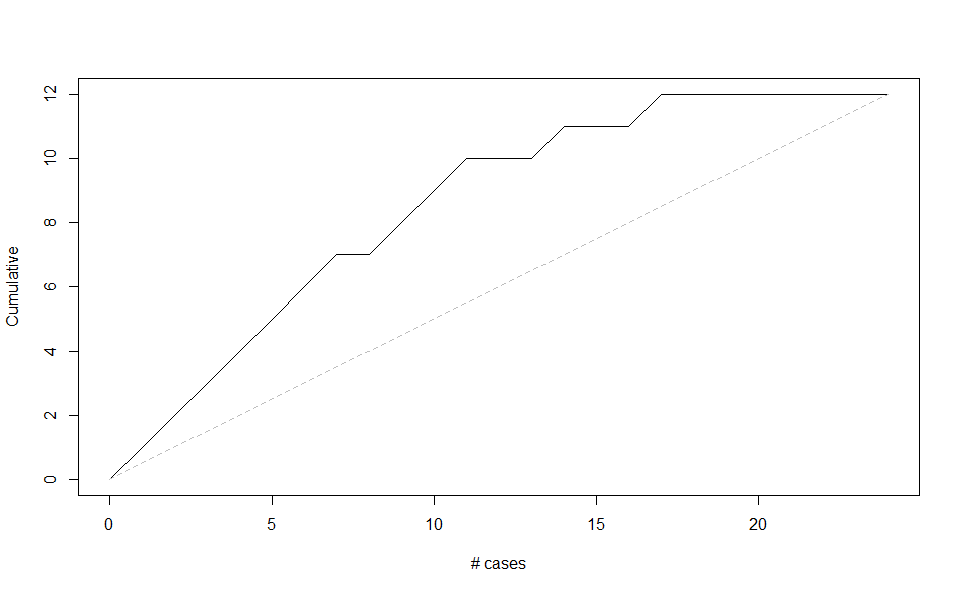
# Plotting a lift chart opt. #1: Caret Library  
# requires caret library  
lift.example <- lift(relevel(as.factor(mowers.df$isOwner),  
 ref="1") ~ logit.reg.pred,  
 data = mowers.df)  
xyplot(lift.example, plt = "gain")



#### Lift Chart Using Gains Library

We can also plot a lift chart using the gains library.

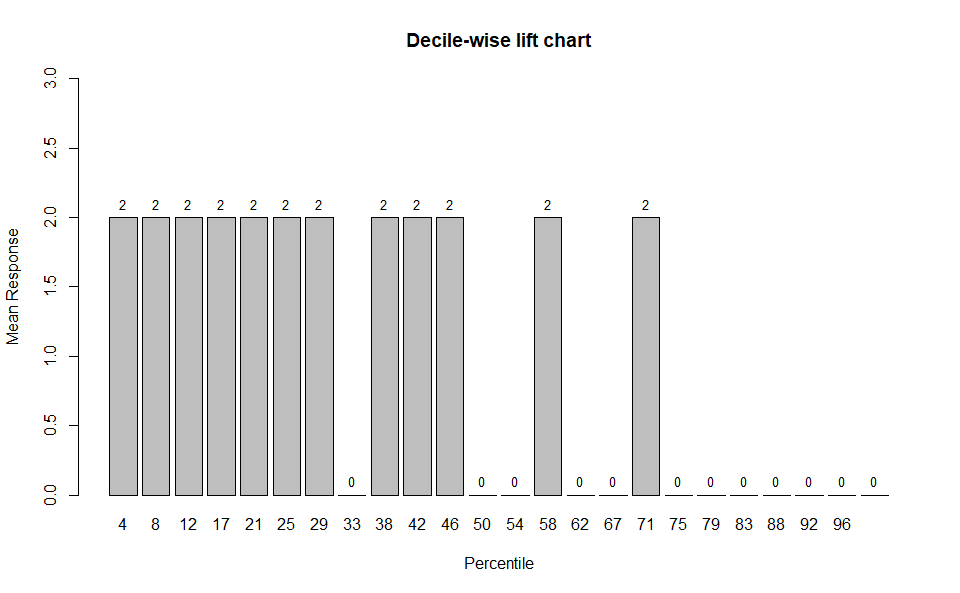
# Plotting a lift chart opt. #2: Gains Library  
# needs library(gains)  
gain <- gains(mowers.df$isOwner,  
 logit.reg.pred,  
 groups = length(logit.reg.pred))  
plot(c(0,gain$cume.pct.of.total\*sum(mowers.df$isOwner))~  
 c(0,gain$cume.obs),  
 xlab ="# cases", ylab = "Cumulative", main="", type="l")  
lines(c(0,sum(mowers.df$isOwner))~c(0, dim(mowers.df)[1]),  
 col="gray", lty=2)



#### Decile Lift Chart

* The decile wise chart aggregates all the lift information into 10 buckets.
  + The y-axis displays the factor by which our model outperforms random assignment of 0’s and 1’s, one decile at a time.
    - Reading from the left, we see that taking 4% of records that our model ranks as “the most probable owners” (those with the highest propensities), yields twice as many owners as would a random selection of 4% of the records.
      * This trend continues until reaching 29% of records, meaning we can use our model to select the top 29% of records with the highest propensities and still perform twice as well as random.
      * We see this trend re-occur when using our model to select the first 71% of records

# Plotting a decile lift chart  
# requires library(gains)  
# when using caret, deciles must be computed manually.  
  
gain.decile <- gains(mowers.df$isOwner, logit.reg.pred, groups=length(logit.reg.pred))  
heights <- gain.decile$mean.resp/mean(mowers.df$isOwner)  
midpoints <- barplot(heights,  
 ylim = c(0,3),  
 names.arg = gain.decile$depth,  
 xlab = "Percentile",  
 ylab = "Mean Response",  
 main = "Decile-wise lift chart")  
# add labels to columns  
text(midpoints, heights+0.1, labels=round(heights,1), cex = 0.8)



**Logistic Regression on Riding Mower Ownership**

The assignment asks the student to create an R Markdown file that outputs to a Microsoft Word document**.** The R Markdown file uses R code, comments, and headings to create the Word document that provides exploratory data analysis of the ***RidingMowers.csv*** file. The dataset contains 24 rows and three columns. Two of the columns (*Income* and *Lot\_Size*) are numeric variables, while *Ownership* is a factor variable that contains whether the individual owns a riding lawnmower. Twelve of the records in the dataset are “*Owners*,” while the other twelve observations are “*Nonowners*.” The average annual income for non-owners is $57,400, while the average yearly salary for owners is $79,470. The average lot size for non-owners is 17,630 ft2, while the average lot size for owners is 20,270 ft2. We use these predictor variables to run a logistic regression model on the data. We model *Ownership* as a function of *Income* and *Lot\_Size*. The estimated logistic equation for the model is as follows:

Positive coefficients for both the *Income* and *Lot\_Size* variables indicate that higher values for these predictors are associated with higher probabilities of owning riding lawnmowers (Shmueli et al., n.d.). We interpret the intercept= -25.9382 as the log odds of an individual owning a riding lawnmower with $0 in income and a lot size of 0 ft2. The coefficient for Income (0.1109) tells us that, holding Lot\_Size constant, we see a 12% increase in the odds of being a riding lawnmower owner for a one-unit increase in Income. Likewise, the coefficient for Lot\_Size (0.9638) tells us that, holding income constant, we see a 162% increase in the odds of being a riding lawnmower owner for a one-unit increase in lot size. The p-values of the coefficients reflect these results. Lot\_Size had a more significant p-value (0.0373) than did Income (0.0412). Both predictors significantly influenced the outcome variable, with p-values < 0.05.

References

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., Lichtendahl, K. C., & Jr. (n.d.). Data Mining for Business Analytics. Retrieved from https://platform.virdocs.com/r/s/0/doc/503437/sp/21743572/mi/74416683?cfi=%2F4%2F2%2F14%2F26%2F6%2F4%2F2%2F4%2F2%2F2%2C%2F1%3A0%2C%2F1%3A0&menu=table-of-contents