```
#################################################################
#In this assignment, we will work with the Carseats
#data which is included in the  ISLR library which you
#installed last time. We will formulate a data science model
#the number of predictors in the dataset. Don't forget to
#create a temporary version of this dataset to work
#with as we did in lecture.
################################################################


######import installed packages########
library(ISLR)
# temporary version of the dataset
carseat_df <- Carseats
#column Names
colnames(carseat_df)

##############################################################################
############################################################
#  A data frame with 400 observations on the following 11 variables.That are given below:
#
#  1)Sales-Unit sales (in thousands) at each location
#  2)CompPrice-Price charged by competitor at each location
#  3)Advertising-Local advertising budget for company at each location (in thousands of dollars)
#  4)Population-Population size in region (in thousands)
#  5)Price-Price company charges for car seats at each site
#  6)ShelveLoc-A factor with levels Bad, Good and Medium indicating the quality of the shelving
location for the car seats at each site
#  7)Age-Average age of the local population
#  8)Education-Education level at each location
#  9)Urban-A factor with levels No and Yes to indicate whether the store is in an urban or rural
location
#  10)US-A factor with levels No and Yes to indicate whether the store is in the US or not
#  11)Income-Community income level (in thousands of dollars)
#
##############################################################################
############################################################

##############################################################################
#############
#(Q1) What are the predictors in the dataset?
#
```

```
# There are 10 predictors in the dataset and their names are given below
#"CompPrice" ,"Income","Advertising","Population",
#"Price","ShelveLoc","Age","Education","Urban","US"
#
###############################################################################
#############


###############################################################################
#########
#(Q2)Use the command "str" dataset, where dataset refers to the name of the copy
#of the dataframe you created. What are the datatypes of thepredictor variables?
#As mentioned above the temporary copy of the dataframe is -carseat_df
#str function  shows the data type of the dataframe.This can be extremely useful
#when we are not sure of all the data that is within our data frame and to get a quick
#look at the data and its structure.
str(carseat_df)
###############################################################################
########



###############################################################################
###########
#Q3) Create a multiple regression model that predicts Sales.
#here Sales is the target variable and others variable are the predictors ,so
# we are doing multiple linear regression so we use all predictors & so we use "." in our code.
# Also we use lm mean linear model built in function in R.
lm1 <- lm(Sales~., data = carseat_df)
lm1
###############################################################################
##################################




###############################################################################
###########
#(Q4)Create a summary of this model. What are the summary statistics? Further,
#which variables are found to be most significant?
summary(lm1)

##################################
#  Residual summary statistics:- #
```

```
##################################

#The symmetry of the residual distribution. The median should be
#close to 0 and in our carseat_df dataframe the median is 0.0211,
#as the mean of the residuals is 0,and symmetric distributions
#each other in magnitude,yes theyare 0.6636 & -0.6908 respectively.
#They would be equal under a symmetric 0 mean distribution.
#The max and min should also have similar magnitude. However,
#in our case, not holding may indicate an outlier rather than
#a symmetry violation.I investigate this further with a boxplot
#of the residuals.The code is given below:-
#####################################################################

dev.off()#To make Rstudio avoid figure margin  error in plot
boxplot(lm1['residuals'],main='Boxplot: Residuals',ylab='residual value')

#####################################################################
#Conclusion from the box plot Residuals.
#We can see that the median is close to 0.
# Further, the 25 and 75 percentile look approximately the
#same distance from 0,and the non-outlier min and max also
#look about the same distance from 0. All of this is good as
#it suggests correct model specification
######################################################################



####################
#  Coefficients:-  #
####################


##############################################################################
##############################################################################
########################
#1)Estimates(The intercept tells us that when all the features are at 0,the expected response is
the intercept.
#2)standard errors(The standard error is the standard error of our estimate, which allows us to
construct marginal confidence intervals for the estimate of that particular feature.),
#3)t statistics(it tells us about how far our estimated parameter is from a hypothesized 0 value,
scaled by the standard deviation of the estimate)
#4)p-values(This is the probability value for the individual coefficient).
```

```
##############################################################################
##############################################################################
##########################


############################
#     More Results         #
############################


#1)Residual standard error( It gives the standard deviation of the residuals, and tells us about how large the prediction error is carseat dataset i.e 1.019)
#2)Multiple R-squared:  0.8734 & Adjusted R-squared:  0.8698 (It tells us about how well our model fits the carset data.)
#3)F-statistic: 243.4 on 11 and 388 DF(degree of freedom),p-value: < 2.2e-16

#################################################
# Variables that are found more significants  #
#################################################

#CompPrice,Price,Age,Education,Income.Advertising



#################################################
#Q5)   Qualitative variables in the dataset?  #
#################################################

#We have three qualitative predictors in this datasets:
#When I used contrast function it converted all qualitative into the
#dummy variable as shown below.

contrasts(carseat_df$ShelveLoc)
contrasts(carseat_df$US)
contrasts(carseat_df$Urban)
```