```r
setwd("C:\\Users\\Swapnil bandekar\\Downloads\\Swapnil\\Data Analytics\\My
Work\\R\\Datasets")

getwd()


## EnY : Analytics Company

## Reliance : Client

## Client wants to increase the sales
## which channel to invest in TV or Radio ??
## Age group target
## Male/Female Ratio


## Business Problem

# Client wants to increase the sales by allocating money in Right Channels


## Defining Analytical Problem ( IMP Step / Converting the Business Problem to
Analytical Problem )

# Create a " Linear Regrrssion Model " which predicts the sales using all the
marketing inputs as independent variables
# so that model will give all the ' Beta ' and then Reliance can optimize marketing
budget using various combinations



## In Linear Regression Model , Target variable is continuous

# e.g. Sales Prediction , Income Prediction

# Input variables can be continuous or categorical

# Target variable = Dependent Variable



## In Logistics Regression Model , Target Variable is Categorical ( Binary most of
the times )

# Insurance claims ( Yes / No ) , Defaulting on Loan



### Simple Linear Regression
```

```
## y = A + Bx ( Equation denotes a linear relationship between X and Y )

# where A = Intercept and B = Slope


## Intercept : It's the value of Y when X = 0

# If A = 0 , line passes through Origin and Y is directly proportional to X

## Slope : It's the rate of change of Y when X changes , or the magnitude of impact
of changes in X on Y .

# If B = 0 , Y is constant so there is no relationship between X and Y ; because
however X changes Y does not change



### Fitted values and Residuals


# The line we choose will not touch every points

# The difference between the line and the actual points are called residuals or
error " e "

## OLS ( Ordinary Least Square Regression )

# OLS finds the line by looking at the residuals ( or the difference between the
points on each line and actual Y ) and
# minimizing the sum of their squares

# Residuals capture the error in the estimated line ( difference between line
estimated and real life values )

# The algorithm will minimize the residuals ( sum of square of distance )


## Sales = Intercept Estimate + Beta coefficient * Amount spent + error

## Sales = b0 + b1 * Amount spent + e


# The OLS will chooose the line which is having least error ( sum of the squares of
distances )
```

```r
# Prediction is best when the error is low

# Modelling usually takes 20% of time

# 80% of the time is spent on data visualization , missing value treatment and
outlier treatment



### Missing value Treatment ( why ??? )


# Algorithm specific need

# Regression models don't take missing values

# Decision trees can work with the missing values



#### Linear Regression Sample



MMix <- read.csv( "MMix.csv" , header = TRUE , stringsAsFactors = FALSE )

View( MMix )

dim( MMix )
str( MMix )
head( MMix )
tail( MMix )



# In this dataset , dependent variable is NewVolSales


summary( MMix )
summary( MMix$NewVolSales )


## Checking the outliers

X <- boxplot( MMix$NewVolSales )

Out <- X$out
```

```
Out

# Storing the outliers in Out object


## Outlier Treatment

Index <- which(MMix$NewVolSales %in% Out)

Index

# Storing the indexes of outliers in Index object


length(Index)

Non_Outlier <- MMix[ -Index , ]

# Removing the outliers and storing the data in Non_Outlier Object

dim(Non_Outlier)

summary(Non_Outlier)

MMix$NewVolSales[Index] <- 23000

# Replacing the outlier values with value 23000 in MMix data

summary(MMix$NewVolSales)


## Checking missing values

is.na(MMix)

colSums(is.na(MMix))

summary(MMix)

## Treating missing values

MMix$NewVolSales[ is.na(MMix$NewVolSales) ] <- mean( MMix$NewVolSales , na.rm =
TRUE )
```

```
# Replacing the missing value with mean
# "na.rm = TRUE" : NA values are excluded while calculating the mean
summary(MMix$Base.Price)

### Exploratory Analysis

## Univariate Analysis
library(ggplot2)
qplot(MMix$NewVolSales)
hist(MMix$NewVolSales)
hist(MMix$Base.Price)

## Bivariate Analysis

qplot( MMix$Base.Price , MMix$NewVolSales )
# same piece of code can be written with the use of with command
with(MMix , qplot( Base.Price , NewVolSales ))
with(MMix , qplot( InStore ,NewVolSales ))
with(MMix , qplot( Radio ,NewVolSales ))

## finding correlations

cor( MMix$NewVolSales , MMix$Base.Price )
with( MMix , cor( NewVolSales , InStore ))
with( MMix , cor( NewVolSales , Radio ))

## Use of Log variables
```

```
## Log variable is used to make a variable in scale with other variable

with( MMix , qplot( log(NewVolSales) , InStore ))


## Creating indicator variables or dummy variables

## why indicator variables or dummy variables ?? => Because model does not accept
character values

## model accept only numbers or factors

unique( MMix$Website.Campaign )

table( MMix$Website.Campaign )

MMix$wc <- ifelse( MMix$Website.Campaign == "Website Campaign " , 1 , 0 )

MMix$fb <- ifelse( MMix$Website.Campaign == "Facebook" , 1 , 0 )

MMix$tw <- ifelse( MMix$Website.Campaign == "Twitter" , 1 , 0 )


table( MMix$wc )

table( MMix$fb )

table( MMix$tw )


# If their are 4 variables then we have to create only 3 dummy variables



## Creating New Variables

## Data Transformations


MMix$LnSales <- log( MMix$NewVolSales )

MMix$LnPrice <- log( MMix$Base.Price )

MMix$OfflineSpend <- MMix$Radio + MMix$TV + MMix$InStore


## Creating Price Buckets => Converting numerical data into categorical

MMix$PriceBucket[ MMix$Base.Price < 15.03 ] <- "Low"
```

```
MMix$PriceBucket[ MMix$Base.Price >= 15.03 & MMix$Base.Price < 15.33 ] <- "Average"

MMix$PriceBucket[ MMix$Base.Price >=15.33 & MMix$Base.Price < 15.64 ] <- "High"

MMix$PriceBucket[ MMix$Base.Price > 15.64 ] <- "Very High"

head(MMix)

####----------------------------------------------Building
MOdels------------------------------------------------------


### Simple Linear Regression Model


?lm


attach(MMix)

Reg <- lm( NewVolSales~Base.Price , data = MMix )
Reg

# Y = a + b*X
# Sales = a + b * Price
# NewVolSales = 50356 - 1976 * Base.Price

summary(Reg)

# Checking the summary of the Regression object "Reg"


## Residuals:
##     Min       1Q   Median       3Q      Max
##  -2422.36  -589.16   -10.34   693.20  2141.41

# Errors of good model are normally distributed with 0 mean
# Median should be close to 0
# 1Q and 3Q should be sort of similar
# Some errors are positive and some errors are negative
# Error = vertical distance between line and data point

##  Coefficients:
##  Estimate Std. Error         t      value   Pr(>|t|)
##   (Intercept)  50355.9     2805.9   17.95   <2e-16 ***
##   Base.Price   -1976.0      183.2  -10.79   <2e-16 ***

##  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# intercept = a = 50356
# beta coefficient = b = -1976
# 1 unit change in x will change Y by beta (b)


# Null Hypothesis , beta (b) = 0 ( Always : for all linear regerssion model )
# No relationship between sales and price
# if p value is less than my preassumed significance level, reject the null
hypothesis that means keep the variable in model
# generalize p value is assumed to be 5% (0.05)

# * represents the lower p value and significance level of variable
# higher the no of * , lower the p value and higher is the significance level of
the variable



## Metrices to Access a Model

# 1. R square
# 2. Coefficients
# 3. p values : significance level of the individuals
# 4. Residuals Distribution


## Factor Variables as individuals
# One of the factor variables becomes the baseline. The estimates of the other
# Types of factor are only given by the model

Reg1 <- lm( NewVolSales~as.factor(PriceBucket) , data = MMix )
summary(Reg1)


# "Low" Price is most significant variable , so creating a dummy variable for it

MMix$PriceBucketLow <- ifelse( MMix$PriceBucket == "Low" , 1 , 0 )


Reg1 <- lm( NewVolSales~PriceBucketLow , data = MMix )
summary(Reg1)
formula(Reg1)


## Multi Co-linearity

# Higher the R2 ( R square ) better  is the model
# If the variable is not relevant to model , Adjusted R2 will go down
```

```r
# VIF ( Variance Inflation Factor ) = 1 / (1 - R2)
# The qualification of multi co-linearity



## Multivariate Regression Model

#Iteration 1

MulReg <- lm( NewVolSales~Base.Price + InStore , data = MMix )
MulReg
summary( MulReg )

# Significance of the variable is decided by the p value
# lower the p variable more significant variable is
# Base.Price is more significant than InStore



# Iteration 2


MulReg <- lm( NewVolSales~Base.Price + InStore + WebCamp , data = MMix )
MulReg
summary( MulReg )

# Beta value for WebCamp is negative
# that means spend on webcamp resulting in decrease in sales
# which in not significant to the model
# as webcamp campaign may increase the sales in 2nd or 3rd week (lag impact )
# beta value and p value may become unstable due to multi co-linearity


MulReg1 <- lm( NewVolSales~WebCamp , data = MMix )
MulReg1
summary(MulReg1)

# beta value is negative in its own as well
# no chance of co-linearity


# Iteration 3

MulReg <- lm( NewVolSales~Base.Price + InStore + TV , data = MMix )
MulReg
summary( MulReg )

# Getting the formula

formula( MulReg )
```

```
## Getting the predicted values

PredSales <- predict( MulReg , data = MMix )
PredSales

# predict command is used to predict the values


## Plotting Actual vs Predicted values

plot( MMix$NewVolSales , col = "blue" , type = "l" )

lines( PredSales , col = "red" ,  type = "l" )

# Plotting the Actual sales values by using the plot fn
# and then adding the Predicted sales values using lines fn




## Finding Residuals

ResSales <- resid(MulReg)
ResSales

plot(ResSales)



## Plotting Residuals vs Predicted Values

# Checking Heteroskedastcity - exists if there is a pattern between predicted
values and error


plot( PredSales , ResSales , abline(0,0) )

plot( NewVolSales , ResSales , abline(0,0) )



## Trying different validation data


sampling <- sort( sample( nrow(MMix) , nrow(MMix)*0.7 ))

head(sampling)

length(sampling)
```

```
View(sampling)


## Dividing the original dataset into Train and Test Data (Train : 70% , Test : 30%
)

Train <- MMix [ sampling , ]
Test <- MMix [ -sampling , ]

dim(MMix)
dim(Train)
dim(Test)

PredSales1 <- predict( MulReg , data = Test )
PredSales1


# What is multicollinearity , a great article!
#
http://blog.minitab.com/blog/understanding-statistics/handling-multicollinearity-in
-regression-analysis

# The function is vif .It belongs to car package

library(HH)
vif(MulReg)

# vif= variation inflation factor
## If the variables in the model have a vif>=2, then you can exclude them from the
model.

# What is vif?:
#
http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/regre
ssion-and-correlation/model-assumptions/what-is-a-variance-inflation-factor-vif/



#####-------------------Summary-------------------



# Used to Predict the variable which is continuous in nature

# Only 1 Target variable is present

# Hypothesis on independent variable which can affect the Target Variable

# Percentile of independent variable should be considered
```

```
# EDA ( Exploratory Data Analysis )

# Univariate Analysis on independent variables

# Dummy variable creation for categorical variables

# Never Impute values of Target Variable ( No missing value Treatment on Target
Variable )

# Bivariate Analysis ( Independent vs Target Variable ) - to check if linear
relationship exist between the two, if not
# then Transformation of independent variable is required ( for e.g log , exp,
square root etc..)

# Model Iterations

# Remove the variable with high P values

# Variables Shortlisting - It can be done by checking the correlation between
target and independent variable , the
# independent variable which is having higher correlation can be taken in model....
Also, can be done by the automation
# of model iteration of each independent variable and then choosing the variable on
the basis of p value

# Model should contain Max 10 variables

# All variables should be unique in nature

# 70 : 30 Rule Train : Test

### MAPE ( Mean Avearge Percentage Error ) can also be used to check if the is good
fit or not.
#   It can be calulated by calculating the absolute % error 100*( Actual Value -
Predicted Value)/Actual Value and then
#   finding the mean of all the % error ( if the mape is <5% then the model is good
fit )


## F-Test


# Null Hypothesis

# B1 = B2 = B3 = B4 = B5 = 0

# Alternate Hypothesis

# Atleast one of the B is non zero
```

```
# F-test is used to compare output of different samples


## Adjusted R2


# Addition of new variable to model will increase the R2

# But , adjusted R2 will increase only if new variable is significant


## Degrees of freedom


# Adjusted R2 take into account degrees of freedom

# If degrees of freedom doesn't change new variable is not significant



## If Adjusted R2 and R2 are having a difference of > 2% ( Adjusted R2 is low and
R2 is high )
#  there lies multi co-linearity

## If Adjusted R2 and R2 are having a difference of > 2% ( Adjusted R2 is high and
R2 is low )
#  there lies heteroskedasticity


## Heteroskedasticity : It refers to the circumstances in which variability of a
variable is unequal across the range of
#  values of a second variable that predicts it
```