A Complete Tutorial to learn Data Science in R fr Scratch

BUSINESS ANALYTICS (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/BUSINESS-ANALYTICS/) MACHINE LEARNING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/MACHINE-LEARNING/) R (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CA

Introduction

R is a powerful language used widely for data analysis and statistical computing. It was de early 90s. Since then, endless efforts have been made to improve R's user interface. The jo language from a rudimentary text editor to interactive R Studio and more recently Jupyter (http://discuss.analyticsvidhya.com/t/how-to-run-r-on-jupyter-ipython-notebooks/5512) engaged many data science communities across the world.

This was possible only because of generous contributions by R users globally. Inclusion packages in R has made it more and more powerful with time. Packages such as dplyr, data.table, SparkR, ggplot2 have made data manipulation, visualization and computation m

But, what about Machine Learning ?

My first impression of R was that it's just a software for statistical computing. Good thing, I R has enough provisions to implement machine learning algorithms in a fast and simple ma

This is a complete tutorial to learn data science and machine learning using R. By the tutorial, you will have a good exposure to building predictive models using machine learni own.

Note: No prior knowledge of data science / analytics is required. However, prior knowledge algebra and statistics will be helpful.

Table of Contents

Let's get started !

Note: The data set used in this article is from Big Mart Sales Prediction (http://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii).

1. Basics of R Programming

Why learn R ?

I don't know if I have a solid reason to convince you, but let me share what got me starte prior coding experience. Actually, I never had computer science in my subjects. I came to k learn data science, one must learn either R or Python as a starter. I chose the former. Her benefits I found after using R:

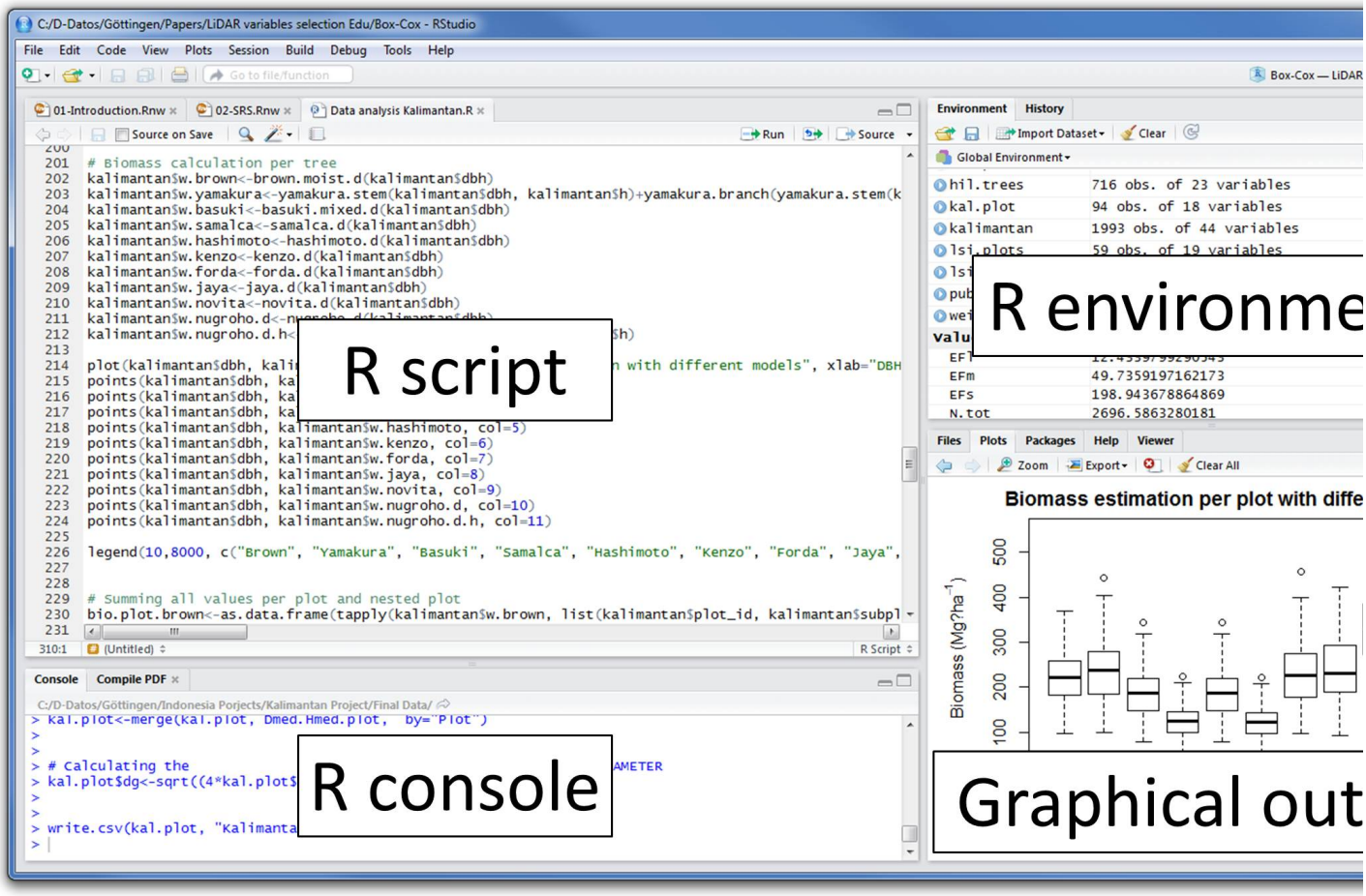
1. The style of coding is quite easy.
2. It's open source. No need to pay any subscription charges.
3. Availability of instant access to over 7800 packages customized for various computation tas
4. The community support is overwhelming. There are numerous forums to help you out.
5. Get high performance computing experience (require packages)
6. One of highly sought skill by analytics and data science companies.

There are many more benefits. But, these are the ones which have kept me going. If you thi exciting, stick around and move to next section. And, if you aren't convinced, you may like Python Tutorial from Scratch (https://www.analyticsvidhya.com/blog/2016/01/comple learn-data-science-python-scratch-2/).

How to install R / R Studio ?

You could download and install the old version (http://ftp.heanet.ie/mirrors/cran.r-project But, I'd insist you to start with RStudio. It provides much better coding experience. For Winc R Studio is available for Windows Vista and above versions. Follow the steps below for Studio:

1. Go to https://www.rstudio.com/products/rstudio/download/
2. In 'Installers for Supported Platforms' section, choose and click the R Studio installer based operating system. The download should begin as soon as you click.
3. Click Next..Next..Finish.
4. Download Complete.
5. To Start R Studio, click on its desktop icon or use 'search windows' to access the program. It this:



Let's quickly understand the interface of R Studio:

1. **R Console:** This area shows the output of code you run. Also, you can directly write codes
Code entered directly in R console cannot be traced later. This is where R script comes to u
2. **R Script:** As the name suggest, here you get space to write codes. To run those codes, si
the line(s) of code and press Ctrl + Enter. Alternatively, you can click on little 'Run' button loc
right corner of R Script.
3. **R environment:** This space displays the set of external elements added. This include
variables, vectors, functions etc. To check if data has been loaded properly in R, always
area.
4. **Graphical Output:** This space display the graphs created during exploratory data analy
graphs, you could select packages, seek help with embedded R's official documentation.

How to install R Packages ?

The sheer power of R lies in its incredible packages. In R, most data handling tasks can be in 2 ways: Using R packages and R base functions. In this tutorial, I'll also introduce you wit handy and powerful R packages. To install a package, simply type:

```
install.packages("package name")
```

As a first time user, a pop might appear to select your CRAN mirror (country serve accordingly and press OK.

Note: You can type this either in console directly and press 'Enter' or in R script and click 'Ru

Basic Computations in R

Let's begin with basics. To get familiar with R coding environment, start with some basic c R console can be used as an interactive calculator too. Type the following in your console:

```
> 2 + 3
> 5

> 6 / 3
>  2

> (3*8)/(2*3)
> 4

> log(12)
> 1.07

> sqrt (121)
> 11
```

Similarly, you can experiment various combinations of calculations and get the results. In want to obtain the previous calculation, this can be done in two ways. First, click in R co press 'Up / Down Arrow' key on your keyboard. This will activate the previously executed c Press Enter.

But, what if you have done too many calculations ? It would be too painful to scroll thr command and find it out. In such situations, creating variable is a helpful way.

In R, you can create a variable using <- or = sign. Let's say I want to create a variable x to c sum of 7 and 8. I'll write it as:

```
> x <- 8 + 7
> x
> 15
```

Once we create a variable, you no longer get the output directly (like calculator), unless variable in the next line. Remember, variables can be alphabets, alphanumeric but not n can't create numeric variables.

2. Essentials of R Programming

Understand and practice this section thoroughly. This is the building block of your R pr knowledge. If you get this right, you would face less trouble in debugging.

R has five basic or 'atomic' classes of objects. Wait, what is an object ?

Everything you see or create in R is an object. A vector, matrix, data frame, even a variable i R treats it that way. So, R has 5 basic classes of objects. This includes:

1. Character
2. Numeric (Real Numbers)
3. Integer (Whole Numbers)
4. Complex
5. Logical (True / False)

Since these classes are self-explanatory by names, I wouldn't elaborate on that. These cl attributes. Think of attributes as their 'identifier', a name or number which aptly identifie object can have following attributes:

1. names, dimension names
2. dimensions
3. class

4. length

Attributes of an object can be accessed using *attributes()* function. More on this coming i section.

Let's understand the concept of object and attributes practically. The most basic object in as vector. You can create an empty vector using *vector()*. Remember, a vector contains obje class.

For example: Let's create vectors of different classes. We can create vector using *c()* or c command also.

```
> a <- c(1.8, 4.5)   #numeric
> b <- c(1 + 2i, 3 - 6i) #complex
> d <- c(23, 44)   #integer
> e <- vector("logical", length = 5)
```

Similarly, you can create vector of various classes.

Data Types in R

R has various type of 'data types' which includes vector (numeric, integer etc), matrices, c and list. Let's understand them one by one.

Vector: As mentioned above, a vector contains object of same class. But, you can mix different classes too. When objects of different classes are mixed in a list, coercion occurs. causes the objects of different types to 'convert' into one class. For example:

```
> qt <- c("Time", 24, "October", TRUE, 3.33)  #character
> ab <- c(TRUE, 24) #numeric
> cd <- c(2.5, "May") #character
```

To check the class of any object, use *class("vector name")* function.

```
> class(qt)
 "character"
```

To convert the class of a vector, you can use *as.* command.

```
> bar <- 0:5
> class(bar)
> "integer"
> as.numeric(bar)
> class(bar)
> "numeric"
> as.character(bar)
> class(bar)
> "character"
```

Similarly, you can change the class of any vector. But, you should pay attention here. If convert a “character” vector to “numeric” , NAs will be introduced. Hence, you should be ca this command.

List: A list is a special type of vector which contain elements of different data types. For exa

```
> my_list <- list(22, "ab", TRUE, 1 + 2i)
> my_list

[[1]]
[1] 22

[[2]]
[1] "ab"

[[3]]
[1] TRUE

[[4]]
[1] 1+2i
```

As you can see, the output of a list is different from a vector. This is because, all the obj different types. The double bracket [[1]] shows the index of first element and so on. Henc easily extract the element of lists depending on their index. Like this:

```
> my_list[[3]]
> [1] TRUE
```

You can use [] single bracket too. But, that would return the list element with its index numl of the result above. Like this:

```
> my_list[3]
> [[1]]
  [1] TRUE
```

Matrices: When a vector is introduced with *row* and *column* i.e. a dimension attribute, it matrix. A matrix is represented by set of rows and columns. It is a 2 dimensional data s consist of elements of same class. Let's create a matrix of 3 rows and 2 columns:

```
> my_matrix <- matrix(1:6, nrow=3, ncol=2)
> my_matrix
[,1] [,2]
[1,] 1 4
[2,] 2 5
[3,] 3 6
```

```
> dim(my_matrix)
[1] 3 2
```

```
> attributes(my_matrix)
$dim
[1] 3 2
```

As you can see, the dimensions of a matrix can be obtained using either *dim()* or *attributes(* command. To extract a particular element from a matrix, simply use the index shown abov example(try this at your end):

```
> my_matrix[,2]   #extracts second column
> my_matrix[,1]   #extracts first column
> my_matrix[2,]   #extracts second row
> my_matrix[1,]   #extracts first row
```

As an interesting fact, you can also create a matrix from a vector. All you need to do dimension *dim()* later. Like this:

```
> age <- c(23, 44, 15, 12, 31, 16)
> age
[1] 23 44 15 12 31 16

> dim(age) <- c(2,3)
> age
[,1] [,2] [,3]
[1,] 23 15 31
[2,] 44 12 16

> class(age)
[1] "matrix"
```

You can also join two vectors using *cbind()* and *rbind()* functions. But, make sure that be have same number of elements. If not, it will return NA values.

```
> x <- c(1, 2, 3, 4, 5, 6)
> y <- c(20, 30, 40, 50, 60)
> cbind(x, y)
> cbind(x, y)
x    y
[1,] 1 20
[2,] 2 30
[3,] 3 40
[4,] 4 50
[5,] 5 60
[6,] 6 70
```

> class(cbind(x, y))
[1] "matrix"

Data Frame: This is the most commonly used member of data types family. It is used to st
data. It is different from matrix. In a matrix, every element must have same class. But, in a
you can put list of vectors containing different classes. This means, every column of a data
like a list. Every time you will read data in R, it will be stored in the form of a data frame.
important to understand the majorly used commands on data frame:

```
> df <- data.frame(name = c("ash","jane","paul","mark"), score = c(67,56,87,91)
> df
name score
1 ash 67
2 jane 56
3 paul 87
4 mark 91

> dim(df)
[1] 4 2

> str(df)
'data.frame': 4 obs. of 2 variables:
$ name : Factor w/ 4 levels "ash","jane","mark",..: 1 2 4 3
$ score: num 67 56 87 91

> nrow(df)
[1] 4

> ncol(df)
[1] 2
```

Let's understand the code above. *df* is the name of data frame. *dim()* returns the dimens
frame as 4 rows and 2 columns. *str()* returns the structure of a data frame i.e. the list of varia
in the data frame. *nrow()* and *ncol()* return the number of rows and number of columns ir

respectively.

Here you see "name" is a factor variable and "score" is numeric. In data science, a variable c categorized into two types: Continuous and Categorical.

Continuous variables are those which can take any form such as 1, 2, 3.5, 4.66 etc. **C variables** are those which takes only discrete values such as 2, 5, 11, 15 etc. In R, categorica represented by factors. In df, name is a factor variable having 4 unique levels. Factor or variable are specially treated in a data set. For more explanation, c (https://www.analyticsvidhya.com/blog/2015/11/easy-methods-deal-categorical-variable predictive-modeling/). Similarly, you can find techniques to deal with continuous vari (https://www.analyticsvidhya.com/blog/2015/11/8-ways-deal-continuous-variables-predi modeling/).

Let's now understand the concept of **missing values** in R. This is one of the most painful part of predictive modeling. You must be aware of all techniques to deal with them. The explanation on such techniques is provided here (https://www.analyticsvidhya.com/blog/2 steps-data-exploration-preparation-building-model-part-2/).

Missing values in R are represented by *NA* and *NaN*. Now we'll check if a data set has mis (using the same data frame *df*).

```
> df[1:2,2] <- NA #injecting NA at 1st, 2nd row and 2nd column of df
> df
name score
1 ash NA
2 jane NA
3 paul 87
4 mark 91

> is.na(df) #checks the entire data set for NAs and return logical output
name score
[1,] FALSE TRUE
[2,] FALSE TRUE
[3,] FALSE FALSE
[4,] FALSE FALSE
```

```
> table(is.na(df)) #returns a table of logical output
FALSE TRUE
6      2
```

> df[!complete.cases(df),] #returns the list of rows having missing values
name score
1 ash NA
2 jane NA

Missing values hinder normal calculations in a data set. For example, let's say, we want t the mean of score. Since there are two missing values, it can't be done directly. Let's see:

```
mean(df$score)
[1] NA
> mean(df$score, na.rm = TRUE)
[1] 89
```

The use of *na.rm = TRUE* parameter tells R to ignore the NAs and compute the mean o values in the selected column (score). To remove rows with NA values in a data frame, y *na.omit*:

```
> new_df <- na.omit(df)
> new_df
name score
3 paul 87
4 mark 91
```

Control Structures in R

As the name suggest, a control structure 'controls' the flow of code / commands written in function. A function is a set of multiple commands written to automate a repetitive coding t

For example: You have 10 data sets. You want to find the mean of 'Age' column present in set. This can be done in 2 ways: either you write the code to compute mean 10 times or create a function and pass the data set to it.

Let's understand the control structures in R with simple examples:

if, else – This structure is used to test a condition. Below is the syntax:

```
if (<condition>){
         ##do something
} else {
         ##do something
}
```

Example

```
#initialize a variable
N <- 10

#check if this variable * 5 is > 40
if (N * 5 > 40){
       print("This is easy!")
} else {
       print ("It's not easy!")
}
[1] "This is easy!"
```

for – This structure is used when a loop is to be executed fixed number of times. It is comr for iterating over the elements of an object (list, vector). Below is the syntax:

```
for (<search condition>){
          #do something
}
```

Example

```
#initialize a vector
y <- c(99,45,34,65,76,23)
```

```
#print the first 4 numbers of this vector
```



```
}
```



```
[1] 99
[1] 45
[1] 34
[1] 65
```


while – It begins by testing a condition, and executes only if the condition is found to be
the loop is executed, the condition is tested again. Hence, it's necessary to alter the con
that the loop doesn't go infinity. Below is the syntax:

```
#initialize a condition
Age <- 12

#check if age is less than 17
while(Age < 17){
        print(Age)
        Age <- Age + 1 #Once the loop is executed, this code breaks the loop
}
[1] 12
[1] 13
[1] 14
[1] 15
[1] 16
```

There are other control structures as well but are less frequently used than explained ab
structures are:

1. repeat – It executes an infinite loop
2. break – It breaks the execution of a loop
3. next – It allows to skip an iteration in a loop
4. return – It help to exit a function

Note: If you find the section 'control structures' difficult to understand, not to worry. R is su various packages to compliment the work done by control structures.

Useful R Packages

Out of ~7800 packages listed on CRAN (https://cran.r-project.org/), I've listed some o powerful and commonly used packages in predictive modeling in this article. Since, I explained the method of installing packages, you can go ahead and install them now. Soo you'll need them.

Importing Data: R offers wide range of packages for importing data available in any form .txt, .csv, .json, .sql etc. To import large files of data quickly, it is advisable to install and use *readr*, *RMySQL*, *sqldf*, *jsonlite*.

Data Visualization: R has in built plotting commands as well. They are good to create sim But, becomes complex when it comes to creating advanced graphics. Hence, you sh *ggplot2*.

Data Manipulation: R has a fantastic collection of packages for data manipulation. These allows you to do basic & advanced computations quickly. These packages are *dplyr*, *lubridate*, *stringr*. Check out this complete (https://www.analyticsvidhya.com/blog/2015/12/faster-data-manipulation-7-packages/) manipulation packages in R.

Modeling / Machine Learning: For modeling, *caret* package in R is powerful enough to ca need for creating machine learning model. However, you can install packages algorithms as *randomForest*, *rpart*, *gbm* etc

Note: *I've only mentioned the commonly used packages. You might like to check this infographic (https://www.analyticsvidhya.com/blog/2015/08/list-r-packages-data-anal complete list of useful R packages.*

Till here, you became familiar with the basic work style in R and its associated components section, we'll begin with predictive modeling. But before you proceed. I want you to practic you've learnt till here.

Practice Assignment: As a part of this assignment, install 'swirl' package in package. *library(swirl)* to initiate the package. And, complete this interactive R tutorial. If you have fc article thoroughly, this assignment should be an easy task for you!

3. Exploratory Data Analysis in R

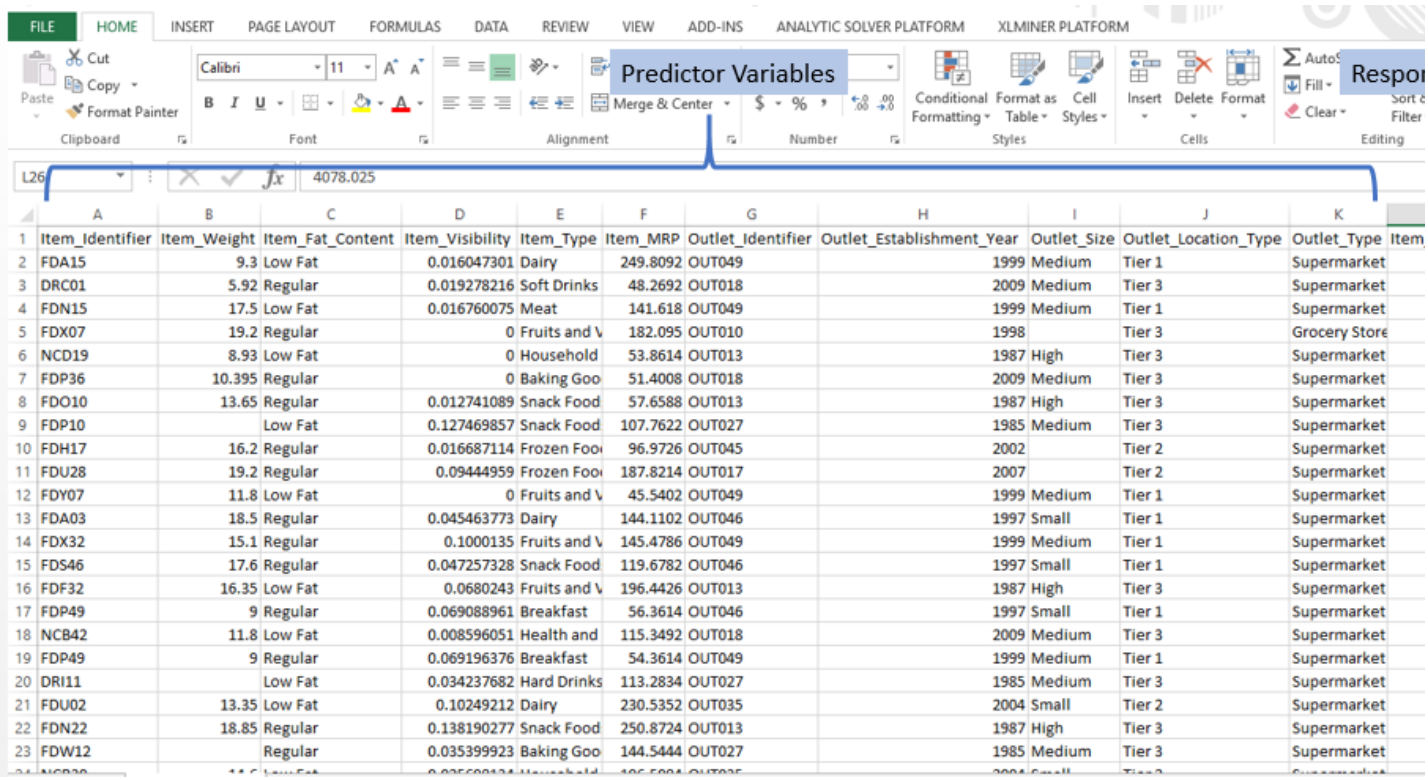
From this section onwards, we'll dive deep into various stages of predictive modeling. He sure you understand every aspect of this section. In case you find anything difficult to unde me in the comments section below.

Data Exploration is a crucial stage of predictive model. You can't build great and practi unless you learn to explore the data from begin to end. This stage forms a concrete fou data manipulation (the very next stage). Let's understand it in R.

In this tutorial, I've taken the data set from Big Mart Sales (https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/). B start, you must get familiar with these terms:

Response Variable (a.k.a Dependent Variable): In a data set, the response variable (y) which we make predictions. In this case, we'll predict 'Item_Outlet_Sales'. (Refer to im below)

Predictor Variable (a.k.a Independent Variable): In a data set, predictor variables (Xi) are t which the prediction is made on response variable. (Image below).



	A	B	C	D	E	F	G	H	I	J	K
1	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outlet_Type
2	FDA15	9.3	Low Fat	0.016047301	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermarket
3	DRC01	5.92	Regular	0.019278216	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermarket
4	FDN15	17.5	Low Fat	0.016760075	Meat	141.618	OUT049	1999	Medium	Tier 1	Supermarket
5	FDX07	19.2	Regular	0	Fruits and V	182.095	OUT010	1998		Tier 3	Grocery Store
6	NCD19	8.93	Low Fat	0	Household	53.8614	OUT013	1987	High	Tier 3	Supermarket
7	FDP36	10.395	Regular	0	Baking Goo	51.4008	OUT018	2009	Medium	Tier 3	Supermarket
8	FDO10	13.65	Regular	0.012741089	Snack Food	57.6588	OUT013	1987	High	Tier 3	Supermarket
9	FDP10		Low Fat	0.127469857	Snack Food	107.7622	OUT027	1985	Medium	Tier 3	Supermarket
10	FDH17	16.2	Regular	0.016687114	Frozen Foo	96.9726	OUT045	2002		Tier 2	Supermarket
11	FDU28	19.2	Regular	0.09444959	Frozen Foo	187.8214	OUT017	2007		Tier 2	Supermarket
12	FDY07	11.8	Low Fat	0	Fruits and V	45.5402	OUT049	1999	Medium	Tier 1	Supermarket
13	FDA03	18.5	Regular	0.045463773	Dairy	144.1102	OUT046	1997	Small	Tier 1	Supermarket
14	FDX32	15.1	Regular	0.1000135	Fruits and V	145.4786	OUT049	1999	Medium	Tier 1	Supermarket
15	FDS46	17.6	Regular	0.047257328	Snack Food	119.6782	OUT046	1997	Small	Tier 1	Supermarket
16	FDF32	16.35	Low Fat	0.0680243	Fruits and V	196.4426	OUT013	1987	High	Tier 3	Supermarket
17	FDP49	9	Regular	0.069088961	Breakfast	56.3614	OUT046	1997	Small	Tier 1	Supermarket
18	NCB42	11.8	Low Fat	0.008596051	Health and	115.3492	OUT018	2009	Medium	Tier 3	Supermarket
19	FDP49	9	Regular	0.069196376	Breakfast	54.3614	OUT049	1999	Medium	Tier 1	Supermarket
20	DRI11		Low Fat	0.034237682	Hard Drinks	113.2834	OUT027	1985	Medium	Tier 3	Supermarket
21	FDU02	13.35	Low Fat	0.10249212	Dairy	230.5352	OUT035	2004	Small	Tier 2	Supermarket
22	FDN22	18.85	Regular	0.138190277	Snack Food	250.8724	OUT013	1987	High	Tier 3	Supermarket
23	FDW12		Regular	0.035399923	Baking Goo	144.5444	OUT027	1985	Medium	Tier 3	Supermarket

(https://www.analyticsvidhya.com/wp-content/uploads/2016/02/PRV.png)

Train Data: The predictive model is always built on train data set. An intuitive way to ident data is, that it always has the 'response variable' included.

Test Data: Once the model is built, it's accuracy is 'tested' on test data. This data always co number of observations than train data set. Also, it does not include 'response variable'.

Right now, you should download the data set. Take a good look at train and test data. Cros information shared above and then proceed.

Let's now begin with **importing and exploring data**.

```
#working directory
path <- ".../Data/BigMartSales"

#set working directory
setwd(path)
```

As a beginner, I'll advise you to keep the train and test files in your working directl unnecessary directory troubles. Once the directory is set, we can easily import the .csv commands below.

```
#Load Datasets
train <- read.csv("Train_UWu5bXk.csv")
test <- read.csv("Test_u94Q5KV.csv")
```

In fact, even prior to loading data in R, it's a good practice to look at the data in Excel. Tl strategizing the complete prediction modeling process. To check if the data set has be successfully, look at R environment. The data can be seen there. Let's explore the data quic

```
#check dimesions ( number of row & columns) in data set
> dim(train)
[1] 8523 12

> dim(test)
[1] 5681 11
```

We have 8523 rows and 12 columns in train data set and 5681 rows and 11 columns in da makes sense. Test data should always have one column less (mentioned above right? deeper in train data set now.

```
#check the variables and their types in train
> str(train)
'data.frame': 8523 obs. of 12 variables:
$ Item_Identifier : Factor w/ 1559 levels "DRA12","DRA24",..: 157 9 663 1122 12
697 739 441 991 ...
$ Item_Weight : num 9.3 5.92 17.5 19.2 8.93 ...
$ Item_Fat_Content : Factor w/ 5 levels "LF","low fat",..: 3 5 3 5 3 5 5 3 5 5
$ Item_Visibility : num 0.016 0.0193 0.0168 0 0 ...
$ Item_Type : Factor w/ 16 levels "Baking Goods",..: 5 15 11 7 10 1 14 14 6 6 .
$ Item_MRP : num 249.8 48.3 141.6 182.1 53.9 ...
$ Outlet_Identifier : Factor w/ 10 levels "OUT010","OUT013",..: 10 4 10 1 2 4 2
...
$ Outlet_Establishment_Year: int 1999 2009 1999 1998 1987 2009 1987 1985 2002 2
```

```
 $ Outlet_Size : Factor w/ 4 levels "","High","Medium",..: 3 3 3 1 2 3 2 3 1 1 .
 $ Outlet_Location_Type : Factor w/ 3 levels "Tier 1","Tier 2",..: 1 3 1 3 3 3 3
...
 $ Outlet_Type : Factor w/ 4 levels "Grocery Store",..: 2 3 2 1 2 3 2 4 2 2 ...
 $ Item_Outlet_Sales : num 3735 443 2097 732 995 ...
```

Let's do some quick data exploration.

To begin with, I'll first check if this data has missing values. This can be done by using:

```
> table(is.na(train))

FALSE TRUE
100813 1463
```

In train data set, we have 1463 missing values. Let's check the variables in which these missing. It's important to find and locate these missing values. Many data scientists have advised beginners to pay close attention to missing value in data exploration stages.

```
> colSums(is.na(train))
Item_Identifier Item_Weight
0               1463
Item_Fat_Content Item_Visibility
0                0
Item_Type        Item_MRP
0                0
Outlet_Identifier Outlet_Establishment_Year
0                0
Outlet_Size      Outlet_Location_Type
0                0
Outlet_Type      Item_Outlet_Sales
0                0
```

Hence, we see that column Item_Weight has 1463 missing values. Let's get more inference data.

```
> summary(train)
```

Here are some quick inferences drawn from variables in train data set:

1. Item_Fat_Content has mis-matched factor levels.
2. Minimum value of item_visibility is 0. Practically, this is not possible. If an item occupies shel grocery store, it ought to have some visibility. We'll treat all 0's as missing values.
3. Item_Weight has 1463 missing values (already explained above).
4. Outlet_Size has a unmatched factor levels.

These inference will help us in treating these variable more accurately.

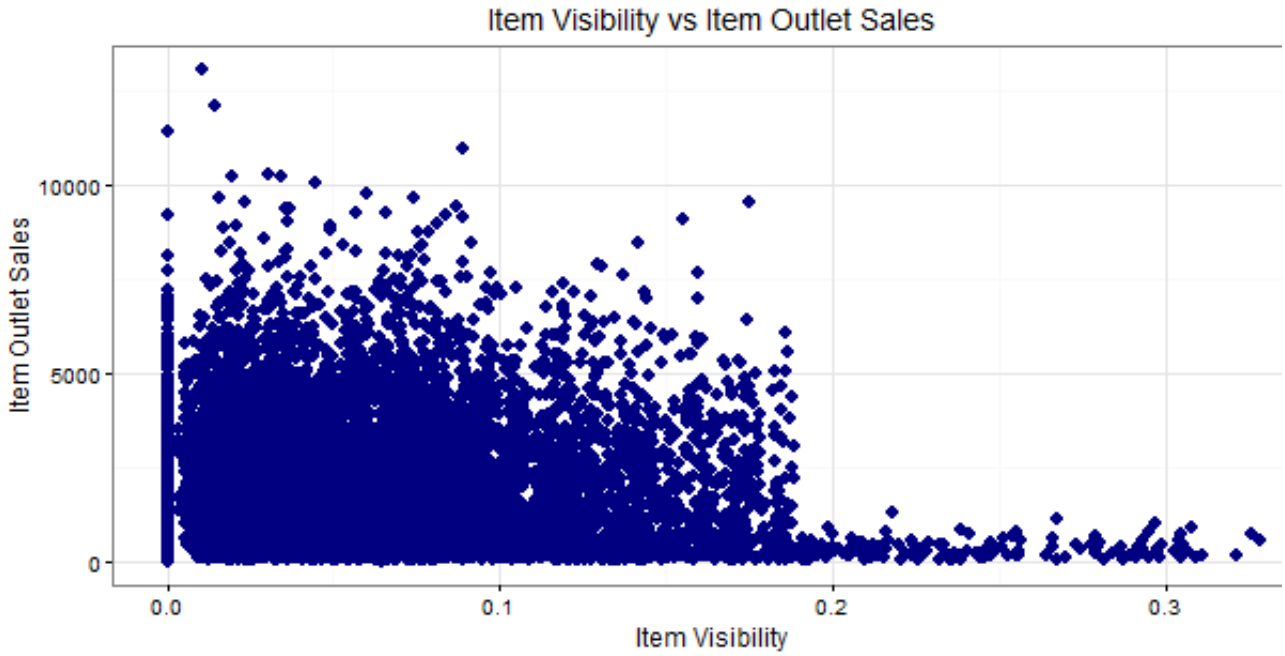
Graphical Representation of Variables

I'm sure you would understand these variables better when explained visually. Using graph analyze the data in 2 ways: Univariate Analysis and Bivariate Analysis.

Univariate analysis is done with one variable. Bivariate analysis is done with two variables analysis is a lot easy to do. Hence, I'll skip that part here. I'd recommend you to try it at you now experiment doing bivariate analysis and carve out hidden insights.

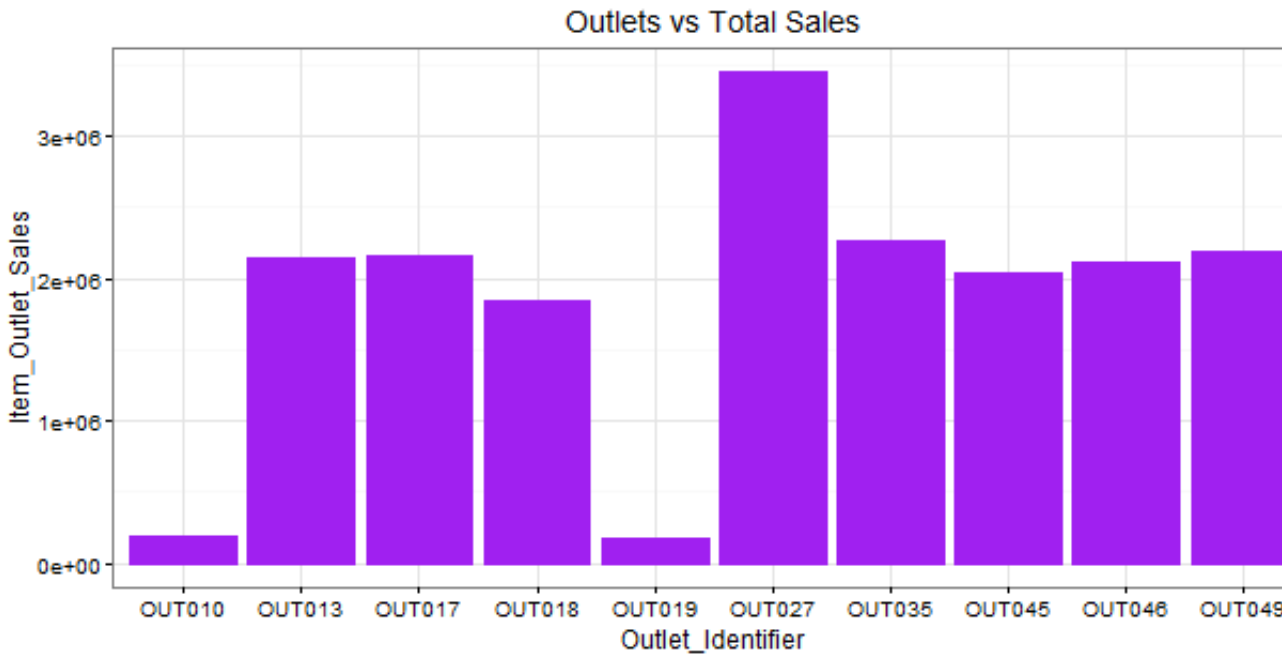
For visualization, I'll use ggplot2 package. These graphs would help us understand the distı and frequency of variables in the data set.

```
> ggplot(train, aes(x= Item_Visibility, y = Item_Outlet_Sales)) + geom_point(si
color="navy") + xlab("Item Visibility") + ylab("Item Outlet Sales") + ggtitle("I
Visibility vs Item Outlet Sales")
```



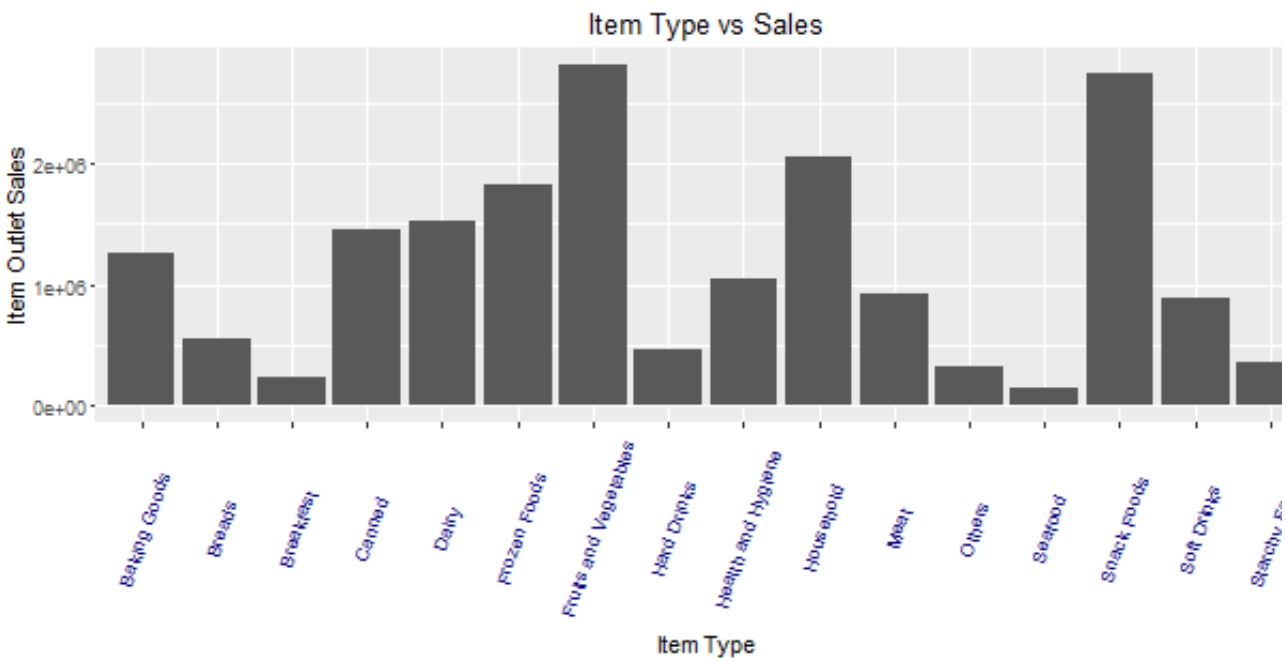
We can see that majority of sales has been obtained from products having visibility less th
suggests that item_visibility < 2 must be an important factor in determining sales. Let's plc
interesting graphs and explore such hidden stories.

```
> ggplot(train, aes(Outlet_Identifier, Item_Outlet_Sales)) + geom_bar(stat = "i
color = "purple") +theme(axis.text.x = element_text(angle = 70, vjust = 0.5, col
"black"))  + ggtitle("Outlets vs Total Sales") + theme_bw()
```



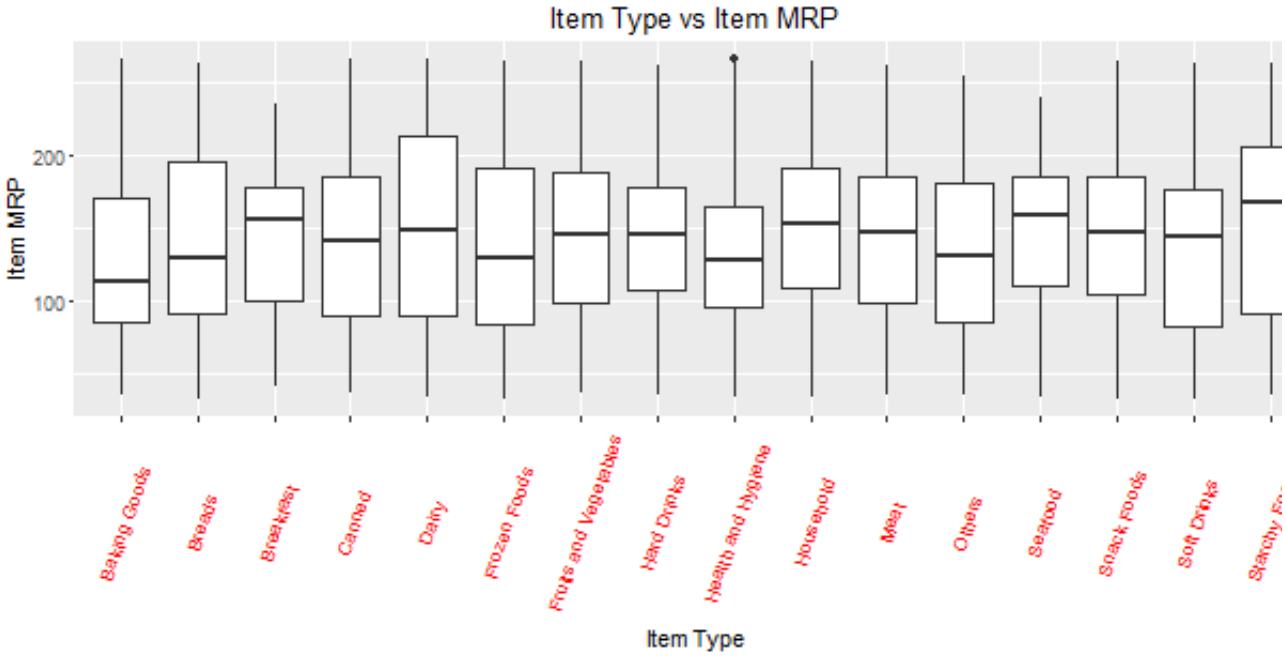
Here, we infer that OUT027 has contributed to majority of sales followed by OUT35. OUT10 a have probably the least footfall, thereby contributing to the least outlet sales.

```
> ggplot(train, aes(Item_Type, Item_Outlet_Sales)) + geom_bar( stat = "identity
+theme(axis.text.x = element_text(angle = 70, vjust = 0.5, color = "navy")) + xl
Type") + ylab("Item Outlet Sales")+ggtitle("Item Type vs Sales")
```



From this graph, we can infer that Fruits and Vegetables contribute to the highest amount c sales followed by snack foods and household products. This information can also be repres using a box plot chart. The benefit of using a box plot is, you get to see the outlier and mea of corresponding levels of a variable (shown below).

```
> ggplot(train, aes(Item_Type, Item_MRP)) +geom_boxplot() +ggtitle("Box Plot")
theme(axis.text.x = element_text(angle = 70, vjust = 0.5, color = "red")) + xlab
Type") + ylab("Item MRP") + ggtitle("Item Type vs Item MRP")
```



The black point you see, is an outlier. The mid line you see in the box, is the mean value of
type. To know more about boxplots, check this tutorial (https://www.analyticsvidhya.com/blog/2015/07/guide-data-visualization-r/).

Now, we have an idea of the variables and their importance on response variable. Let's
back to where we started. Missing values. Now we'll impute the missing values.

We saw variable Item_Weight has missing values. Item_Weight is an continuous variable
this case we can impute missing values with mean / median of item_weight. These ar
commonly used methods of imputing missing value. To explore other methods of this t
check out this tutorial (https://www.analyticsvidhya.com/blog/2016/01/guide-data-explor

Let's first combine the data sets. This will save our time as we don't need to write separat
train and test data sets. To combine the two data frames, we must make sure that they
columns, which is not the case.

```
> dim(train)
[1] 8523 12

> dim(test)
[1] 5681 11
```

Test data set has one less column (response variable). Let's first add the column. We ca column any value. An intuitive approach would be to extract the mean value of sales fron set and use it as placeholder for test variable Item _Outlet_ Sales. Anyways, let's make it now. I've taken a value 1. Now, we'll combine the data sets.

```
> test$Item_Outlet_Sales <-  1
> combi <- rbind(train, test)
```

Impute missing value by median. I'm using median because it is known to be highly robust Moreover, for this problem, our evaluation metric is RMSE (https://www.analyticsvidhya.com/blog/2016/02/7-important-model-evaluation-error-metrics/) which is also highly affected by outliers. Hence, median is better in this case.

```
> combi$Item_Weight[is.na(combi$Item_Weight)] <- median(combi$Item_Weight, na.r
> table(is.na(combi$Item_Weight))

FALSE
14204
```

Trouble with Continuous Variables & Categorical Variables

It's important to learn to deal with continuous and categorical variables separately in a other words, they need special attention. In this data set, we have only 3 continuous variabl are categorical in nature. If you are still confused, I'll suggest you to once again look at tl using *str()* and proceed.

Let's take up *Item_Visibility*. In the graph above, we saw item visibility has zero value als practically not feasible. Hence, we'll consider it as a missing value and once again imputation using median.

```
> combi$Item_Visibility <- ifelse(combi$Item_Visibility == 0,
                            median(combi$Item_Visibility), combi$Item_Visibility
```

Let's proceed to categorical variables now. During exploration, we saw there are mis-mat in variables which needs to be corrected.

```
> levels(combi$Outlet_Size)[1] <- "Other"
> library(plyr)
> combi$Item_Fat_Content <- revalue(combi$Item_Fat_Content,
c("LF" = "Low Fat", "reg" = "Regular"))
> combi$Item_Fat_Content <- revalue(combi$Item_Fat_Content, c("low fat" = "Low F
> table(combi$Item_Fat_Content)
  Low Fat Regular
  9185    5019
```

Using the commands above, I've assigned the name 'Other' to unnamed level in *Outlet_Si*
Rest, I've simply renamed the various levels of Item_Fat_Content.

4. Data Manipulation in R

Let's call it as, the advanced level of data exploration. In this section we'll practically learn a
feature engineering and other useful aspects.

Feature Engineering: This component separates an intelligent data scientist from a
enabled data scientist. You might have access to large machines to run heavy compu
algorithms, but the power delivered by new features, just can't be matched. We create ne
to extract and provide as much 'new' information to the model, to help it make accurate pre

If you have been thinking all this time, great. But now is the time to think deeper. Look at t
and ask yourself, what else (factor) could influence Item_Outlet_Sales ? Anyhow, the answ
But, I want you to try it out first, before scrolling down.

1. **Count of Outlet Identifiers** – There are 10 unique outlets in this data. This variable
information on count of outlets in the data set. More the number of counts of an outlet, c
more will be the sales contributed by it.

```
> library(dplyr)
> a <- combi%>%
            group_by(Outlet_Identifier)%>%
            tally()
```

```
> head(a)
Source: local data frame [6 x 2]  Outlet_Identifier n
(fctr)           (int)
1 OUT010         925
2 OUT013         1553
3 OUT017         1543
4 OUT018         1546
5 OUT019         880
6 OUT027         1559

> names(a)[2] <- "Outlet_Count"
> combi <- full_join(a, combi, by = "Outlet_Identifier")
```

As you can see, dplyr package makes data manipulation quite effortless. You no longer ne long function. In the code above, I've simply stored the new data frame in a variable *a.* Lat column *Outlet_Count* is added in our original 'combi' data set. To know more about dplyr, tutorial (https://rpubs.com/bradleyboehmke/data_wrangling).

2. **Count of Item Identifiers** – Similarly, we can compute count of item identifiers too. practice to fetch more information from unique ID variables using their count. This will understand, which outlet has maximum frequency.

```
> b <- combi%>%
          group_by(Item_Identifier)%>%
          tally()

 > names(b)[2] <- "Item_Count"
> head (b)
Item_Identifier   Item_Count
(fctr)            (int)
1 DRA12             9
2 DRA24             10
3 DRA59             10
```

```
4 DRB01              8
5 DRB13              9
6 DRB24              8
```

> combi <- merge(b, combi, by = "Item_Identifier")

3. **Outlet Years** – This variable represent the information of existence of a particular outlet 2013. Why just 2013? You'll find the answer in problem statem (http://datahack.analyticsvidhya.com/contest/practice-problem-bigmart-sales-prediction) hypothesis is, older the outlet, more footfall, large base of loyal customers and larger the ou

```
> c <- combi%>%
          select(Outlet_Establishment_Year)%>%
          mutate(Outlet_Year = 2013 - combi$Outlet_Establishment_Year)
 > head(c)
Outlet_Establishment_Year  Outlet_Year
1 1999                       14
2 2009                        4
3 1999                       14
4 1998                       15
5 1987                       26
6 2009                        4

> combi <- full_join(c, combi)
```

This suggests that outlets established in 1999 were 14 years old in 2013 and so on.

4. **Item Type New** – Now, pay attention to *Item_Identifiers.* We are about to discover a new carefully, there is a pattern in the identifiers starting with "FD","DR","NC". Now, check the cor *Item_Types* to these identifiers in the data set. You'll discover, items corresponding to "DR", eatables. Items corresponding to "FD", are drinks. And, item corresponding to "NC", are proc can't be consumed, let's call them non-consumable. Let's extract these variables into a ne representing their counts.

Here I'll use *substr()*, *gsub()* function to extract and rename the variables respectively.

```
> q <- substr(combi$Item_Identifier,1,2)
> q <- gsub("FD","Food",q)
> q <- gsub("DR","Drinks",q)
> q <- gsub("NC","Non-Consumable",q)
> table(q)
   Drinks Food  Non-Consumable
   1317   10201 2686
```

Let's now add this information in our data set with a variable name 'Item_Type_New.

```
> combi$Item_Type_New <- q
```

I'll leave the rest of feature engineering intuition to you. You can think of more variables w add more information to the model. But make sure, the variable aren't correlated. Sinc emanating from a same set of variable, there is a high chance for them to be correlate check the same in R using *cor()* function.

Label Encoding and One Hot Encoding

Just, one last aspect of feature engineering left. Label Encoding and One Hot Encoding.

Label Encoding, in simple words, is the practice of numerically encoding (replacing) differe a categorical variables. For example: In our data set, the variable *Item_Fat_Content* has 2 Fat and Regular. So, we'll encode Low Fat as 0 and Regular as 1. This will help us conv variable in numeric variable. This can be simply done using if else statement in R.

```
> combi$Item_Fat_Content <- ifelse(combi$Item_Fat_Content == "Regular",1,0)
```

One Hot Encoding, in simple words, is the splitting a categorical variable into its uni and eventually removing the original variable from data set. Confused ? Here's an example any categorical variable, say, *Outlet_ Location_Type*. It has 3 levels. One hot encoding of th will create 3 different variables consisting of 1s and 0s. 1s will represent the existence of v 0s will represent non-existence of variable. Let look at a sample:

```
> sample <- select(combi, Outlet_Location_Type)
> demo_sample <- data.frame(model.matrix(~.-1,sample))
> head(demo_sample)
Outlet_Location_TypeTier.1 Outlet_Location_TypeTier.2 Outlet_Location_TypeTier.3
1                          1                          0                          0
2                          0                          0                          1
3                          1                          0                          0
4                          0                          0                          1
5                          0                          0                          1
6                          0                          0                          1
```

model.matrix creates a matrix of encoded variables. ~. -1 tells R, to encode all variables in t
frame, but suppress the intercept. So, what will happen if you don't write -1 ? model.matrix v
the first level of the factor, thereby resulting in just 2 out of 3 factor levels (loss of informatio

This was the demonstration of one hot encoding. Hope you have understood the concept r
now apply this technique to all categorical variables in our data set (excluding ID variable).

```
>library(dummies)
>combi <- dummy.data.frame(combi, names =
c('Outlet_Size','Outlet_Location_Type','Outlet_Type', 'Item_Type_New'),  sep='_'
```

With this, I have shared 2 different methods of performing one hot encoding in R. Let's che
encoding has been done.

```
> str (combi)
$ Outlet_Size_Other : int 0 1 1 0 1 0 0 0 0 0 ...
$ Outlet_Size_High : int 0 0 0 1 0 0 0 0 0 0 ...
$ Outlet_Size_Medium : int 1 0 0 0 0 0 1 1 0 1 ...
$ Outlet_Size_Small : int 0 0 0 0 0 1 0 0 1 0 ...
$ Outlet_Location_Type_Tier 1 : int 1 0 0 0 0 0 0 0 1 0 ...
$ Outlet_Location_Type_Tier 2 : int 0 1 0 0 1 1 0 0 0 0 ...
$ Outlet_Location_Type_Tier 3 : int 0 0 1 1 0 0 1 1 0 1 ...
$ Outlet_Type_Grocery Store : int 0 0 1 0 0 0 0 0 0 0 ...
$ Outlet_Type_Supermarket Type1: int 1 1 0 1 1 1 0 0 1 0 ...
$ Outlet_Type_Supermarket Type2: int 0 0 0 0 0 0 0 1 0 0 ...
```

```
$ Outlet_Type_Supermarket Type3: int 0 0 0 0 0 0 1 0 0 1 ...
$ Item_Outlet_Sales : num 1 3829 284 2553 2553 ...
$ Year : num 14 11 15 26 6 9 28 4 16 28 ...
$ Item_Type_New_Drinks : int 1 1 1 1 1 1 1 1 1 1 ...
$ Item_Type_New_Food : int 0 0 0 0 0 0 0 0 0 0 ...
$ Item_Type_New_Non-Consumable : int 0 0 0 0 0 0 0 0 0 0 ...
```

As you can see, after one hot encoding, the original variables are removed automatically fro
data set.

5. Predictive Modeling using Machine Learning

Finally, we'll drop the columns which have either been converted using other variables or a
variables. This can be accomplished using *select* from dplyr package.

```
> combi <- select(combi, -c(Item_Identifier, Outlet_Identifier, Item_Fat_Conten
                        Outlet_Establishment_Year,Item_Type))
> str(combi)
```

In this section, I'll cover Regression, Decision Trees and Random Forest. A detailed exp
these algorithms is outside the scope of this article. These algorithms have been s
explained in our previous articles. I've provided the links for useful resources.

As you can see, we have encoded all our categorical variables. Now, this data set
take forward to modeling. Since, we started from Train and Test, let's now divide the data se

```
> new_train <- combi[1:nrow(train),]
> new_test <- combi[-(1:nrow(train)),]
```

Linear (Multiple) Regression

Multiple Regression is used when response variable is continuous in nature and predictor
Had it been categorical, we would have used Logistic Regression. Before you procee
your basics of Regression here (https://www.analyticsvidhya.com/blog/2015/08/comp

guide-regression/).

Linear Regression takes following assumptions:

1. There exists a linear relationship between response and predictor variables
2. The predictor (independent) variables are not correlated with each other. Presence of collin to a phenomenon known as multicollinearity (https://en.wikipedia.org/wiki/Multicollinearit
3. The error terms are uncorrelated. Otherwise, it will lead to autocorrelation (https://en.wikipedia.org/wiki/Autocorrelation#Regression_analysis).
4. Error terms must have constant variance. Non-constant variance leads to heteroskedasticit (https://en.wikipedia.org/wiki/Heteroscedasticity).

Let's now build out first regression model on this data set. R uses *lm()* function for regressio

```
> linear_model <- lm(Item_Outlet_Sales ~ ., data = new_train)
> summary(linear_model)
```

Adjusted R^2 measures the goodness of fit of a regression model. Higher the R^2, better is Our **R^2 = 0.2085**. It means we really did something drastically wrong. Let's figure it out.

In our case, I could find our new variables aren't helping much i.e. Item count, Outlet Item_Type_New. Neither of these variables are significant. Significant variables are denoted

As we know, correlated predictor variables brings down the model accuracy. Let's find out t of correlation present in our predictor variables. This can be simply calculated using:

```
> cor(new_train)
```

Alternatively, you can also use corrplot package for some fancy correlation plots. Scrolli the long list of correlation coefficients, I could find a deadly correlation coefficient:

```
cor(new_train$Outlet_Count, new_train$`Outlet_Type_Grocery Store`)
[1] -0.9991203
```

Outlet_Count is highly correlated (negatively) with Outlet Type Grocery Store. Here problems I could find in this model:

1. We have correlated predictor variables.
2. We did one hot encoding and label encoding. That's not necessary since linear regres categorical variables by creating dummy variables intrinsically.

3. The new variables (item count, outlet count, item type new) created in feature enginee significant.

Let's try to create a more robust regression model. This time, I'll be using a building a sin without encoding and new features. Below is the entire code:

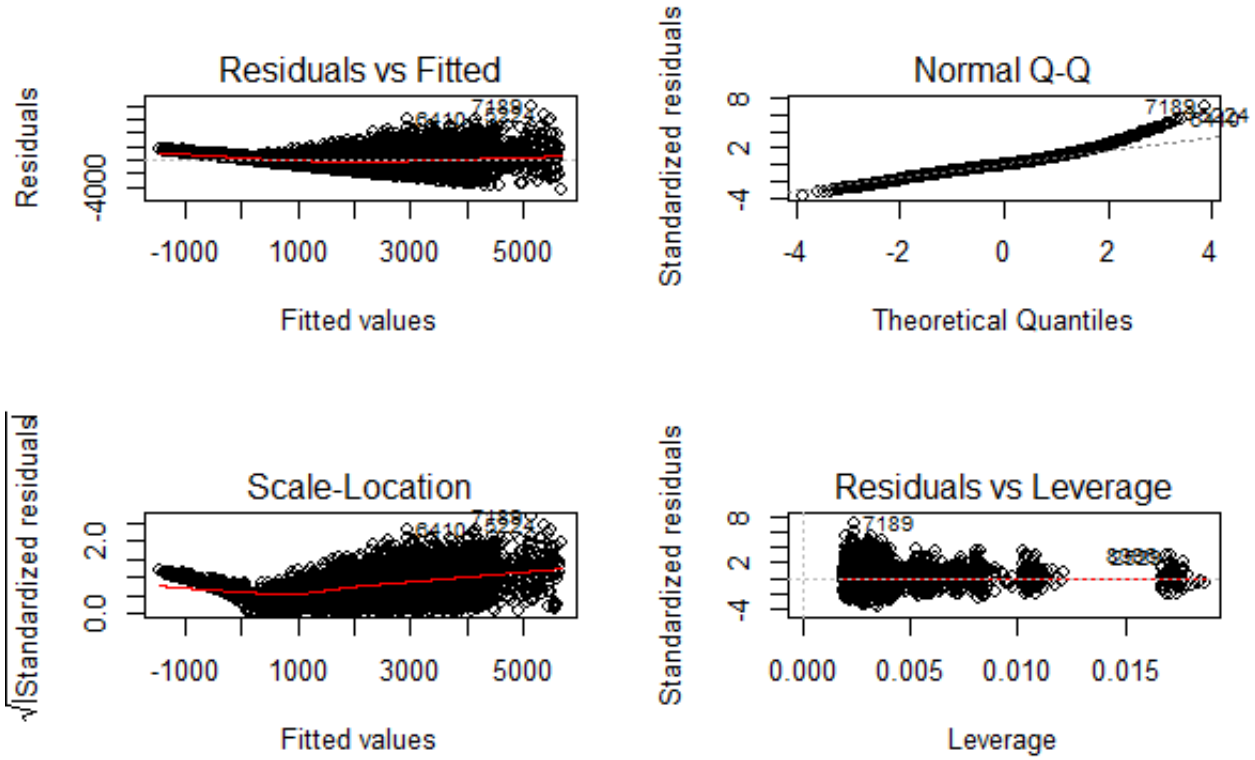
```
#load directory
> path <- "C:/Users/manish/desktop/Data/February 2016"

> setwd(path)

#load data
> train <- read.csv("train_Big.csv")
> test <- read.csv("test_Big.csv")

#create a new variable in test file
> test$Item_Outlet_Sales <- 1

#combine train and test data
> combi <- rbind(train, test)

#impute missing value in Item_Weight
> combi$Item_Weight[is.na(combi$Item_Weight)] <- median(combi$Item_Weight, na.r

#impute 0 in item_visibility
> combi$Item_Visibility <- ifelse(combi$Item_Visibility == 0,
median(combi$Item_Visibility),                        combi$Item_Visibility)

#rename level in Outlet_Size
> levels(combi$Outlet_Size)[1] <- "Other"

#rename levels of Item_Fat_Content
> library(plyr)
> combi$Item_Fat_Content <- revalue(combi$Item_Fat_Content,c("LF" = "Low Fat",
                                "Regular"))
> combi$Item_Fat_Content <- revalue(combi$Item_Fat_Content, c("low fat" = "Low
```

```
#create a new column 2013 - Year
> combi$Year <- 2013 - combi$Outlet_Establishment_Year

#drop variables not required in modeling
> library(dplyr)
> combi <- select(combi, -c(Item_Identifier, Outlet_Identifier,
Outlet_Establishment_Year))

#divide data set
> new_train <- combi[1:nrow(train),]
> new_test <- combi[-(1:nrow(train)),]

#linear regression
> linear_model <- lm(Item_Outlet_Sales ~ ., data = new_train)
> summary(linear_model)
```

Now we have got **R^2 = 0.5623**. This teaches us that, sometimes all you need is simple thoug to get high accuracy. Quite a good improvement from previous model. Next, time when y any model, always remember to start with a simple model.

Let's check out regression plot to find out more ways to improve this model.

```
> par(mfrow=c(2,2))
> plot(linear_model)
```



You can zoom these graphs in R Studio at your end. All these plots have a different story the most important story is being portrayed by Residuals vs Fitted graph.

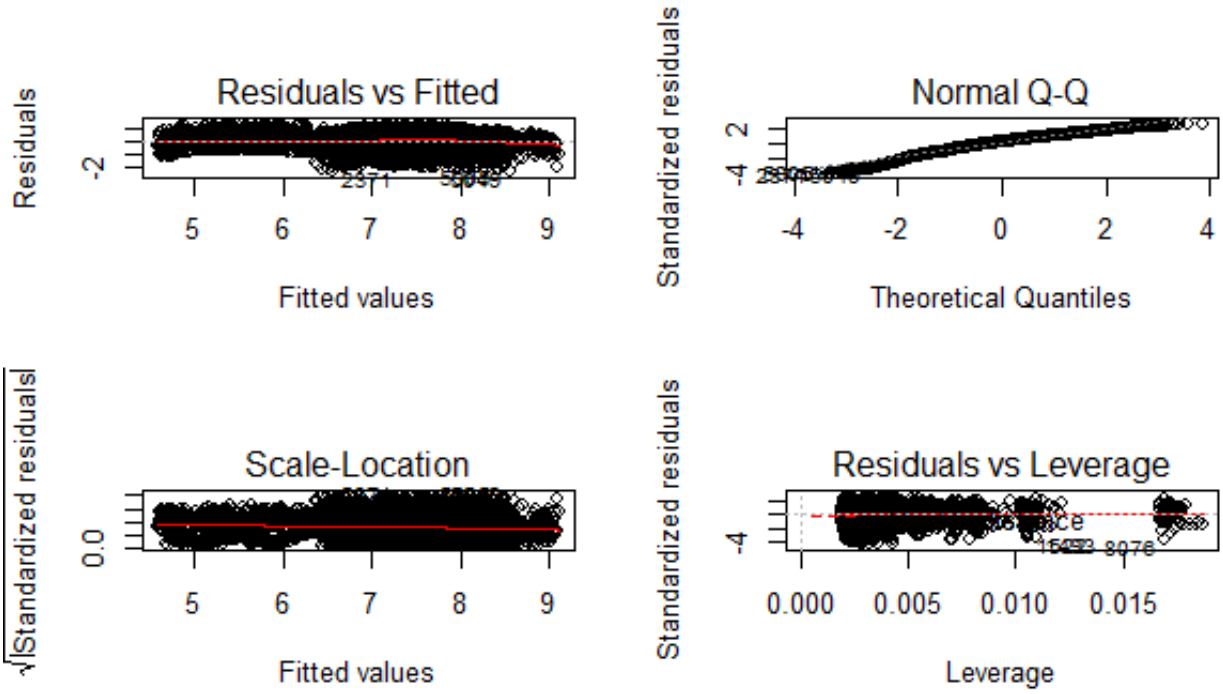
Residual values are the difference between actual and predicted outcome values. Fitted the predicted values. If you see carefully, you'll discover it as a funnel shape graph (from ri The shape of this graph suggests that our model is suffering from heteroskedasticit variance in error terms). Had there been constant variance, there would be no pattern vi graph.

A common practice to tackle heteroskedasticity is by taking the log of response variable and check if we can get further improvement.

```
> linear_model <- lm(log(Item_Outlet_Sales) ~ ., data = new_train)
> summary(linear_model)
```

```
Item_TypeSnack Foods
Item_TypeSoft Drinks
Item_TypeStarchy Foods
Item_MRP                        ***
Outlet_SizeHigh                 **
Outlet_SizeMedium               *
Outlet_SizeSmall                **
Outlet_Location_TypeTier 2
Outlet_Location_TypeTier 3      *
Outlet_TypeSupermarket Type1    ***
Outlet_TypeSupermarket Type2    ***
Outlet_TypeSupermarket Type3    ***
Year                            **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5378 on 8494 degrees of freedom
Multiple R-squared:  0.7214,    Adjusted R-squared:  0.7205
F-statistic: 785.4 on 28 and 8494 DF,  p-value: < 2.2e-16
```

And, here's a snapshot of my model output. Congrats! We have got an improved model wit
Now, we are on the right path. Once again you can check the residual plots (you might zoc
find there is no longer a trend in residual vs fitted value plot.



This model can be further improved by detecting outliers and high leverage points. For r
that part to you! I shall write a separate post on mysteries of regression soon. For now, let's
RMSE so that we can compare it with other algorithms demonstrated below.

To calculate RMSE, we can load a package named *Metrics.*

```
> install.packages("Metrics")
> library(Metrics)
> rmse(new_train$Item_Outlet_Sales, exp(linear_model$fitted.values))
[1] 1140.004
```

Let's proceed to decision tree algorithm and try to improve our RMSE score.

Decision Trees

Before you start, I'd recommend you to glance through the basics of decision tree alg understand what makes it superior than linear regression, check this tutori (https://www.analyticsvidhya.com/blog/2015/01/decision-tree-simplified/) and (https://www.analyticsvidhya.com/blog/2015/01/decision-tree-algorithms-simplified/).

In R, decision tree algorithm can be implemented using *rpart package.* In addition, we'l *package* for doing cross validation. Cross validation is a technique to build robust models not prone to overfitting. Read more about Cross (https://www.analyticsvidhya.com/blog/2015/11/improve-model-performance-cross-valic python-r/).

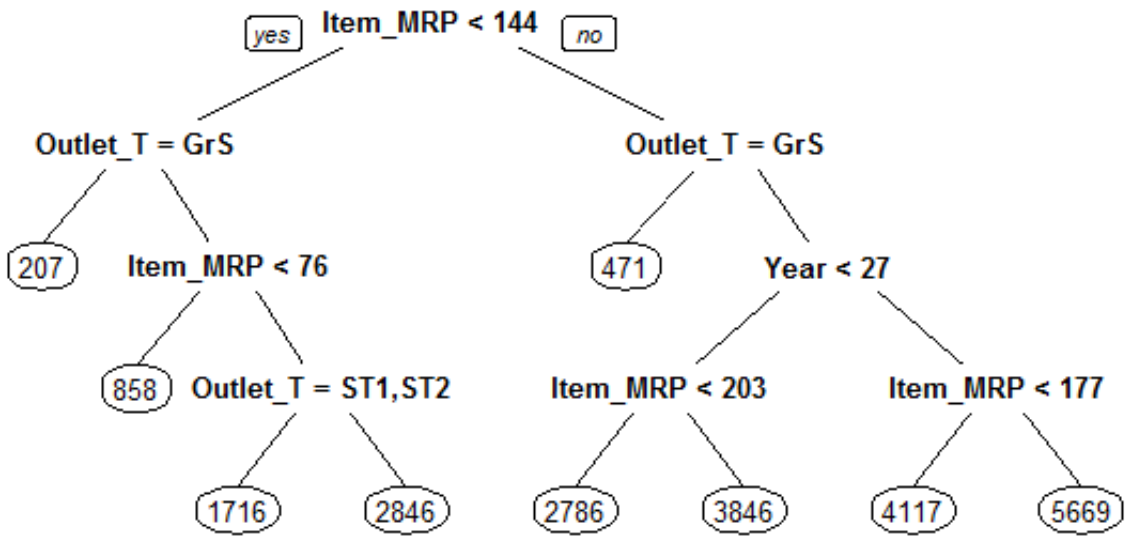
In R, decision tree uses a complexity parameter (*cp*). It measures the tradeoff betw complexity and accuracy on training set. A smaller cp will lead to a bigger tree, which m the model. Conversely, a large cp value might underfit the model. Underfitting occurs model does not capture underlying trends properly. Let's find out the optimum cp value fo with 5 fold cross validation.

```
#loading required libraries
> library(rpart)
> library(e1071)
> library(rpart.plot)
> library(caret)
```

```
#setting the tree control parameters
> fitControl <- trainControl(method = "cv", number = 5)
> cartGrid <- expand.grid(.cp=(1:50)*0.01)

#decision tree
> tree_model <- train(Item_Outlet_Sales ~ ., data = new_train, method = "rpart"
trControl = fitControl, tuneGrid = cartGrid)
> print(tree_model)
```

The final value for cp = 0.01. You can also check the table populated in console for more i
The model with cp = 0.01 has the least RMSE. Let's now build a decision tree with 0.01 as
parameter.

```
> main_tree <- rpart(Item_Outlet_Sales ~ ., data = new_train, control =
rpart.control(cp=0.01))
> prp(main_tree)
```



Here is the tree structure of our model. If you have gone through the basics, you
understand that this algorithm has marked Item_MRP as the most important variable (bei
node). Let's check the RMSE of this model and see if this is any better than regression.

```
> pre_score <- predict(main_tree, type = "vector")
> rmse(new_train$Item_Outlet_Sales, pre_score)
[1] 1102.774
```

As you can see, our RMSE has further improved from 1140 to 1102.77 with decision tree. this score further, you can further tune the parameters for greater accuracy.

Random Forest

Random Forest is a powerful algorithm which holistically takes care of missing values, c other non-linearities in the data set. It's simply a collection of classification trees, hence 'forest'. I'd suggest you to quickly refresh your basics of random forest with tl (https://www.analyticsvidhya.com/blog/2015/09/random-forest-algorithm-multiple-chall

In R, random forest algorithm can be implement using *randomForest* package. Again, we package for cross validation and finding optimum value of model parameters.

For this problem, I'll focus on two parameters of random forest. *mtry* and *ntree*. ntree is the trees to be grown in the forest. *mtry* is the number of variables taken at each node to build we'll do a 5 fold cross validation.

Let's do it!

```
#load randomForest library
> library(randomForest)

#set tuning parameters
> control <- trainControl(method = "cv", number = 5)

#random forest model
> rf_model <- train(Item_Outlet_Sales ~ ., data = new_train, method = "parRF",
=                 control,  prox = TRUE, allowParallel = TRUE)

#check optimal parameters
> print(rf_model)
```

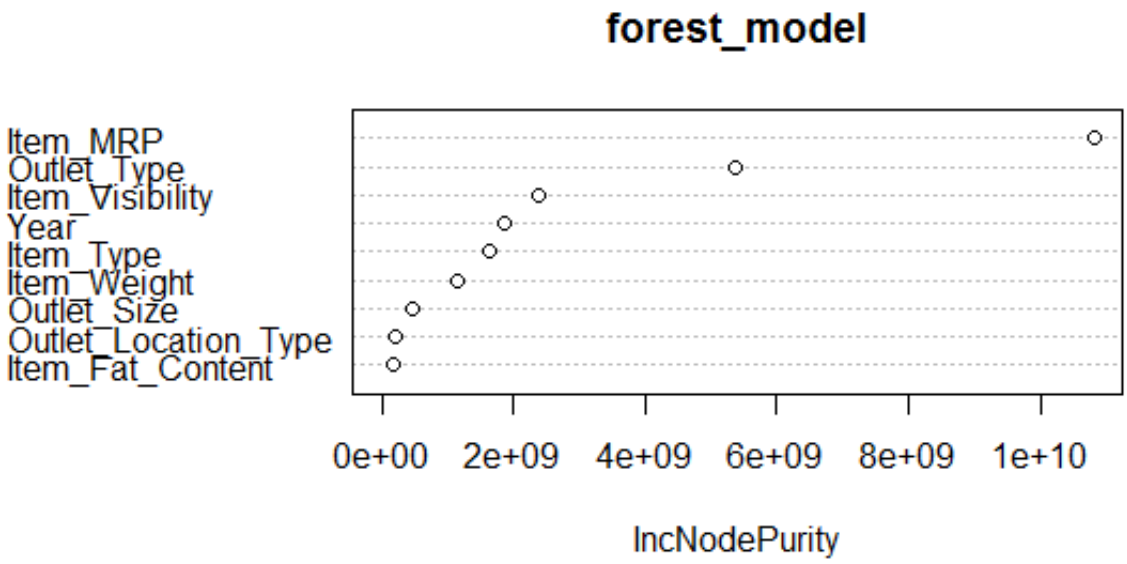
```
Parallel Random Forest

8523 samples
   9 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 6819, 6818, 6818, 6819, 6818
Resampling results across tuning parameters:

  mtry  RMSE      Rsquared   RMSE SD    Rsquared SD
   2    1293.032  0.5157854   7.845358  0.01852388
  15    1122.530  0.5697400  16.109357  0.01307624
  28    1135.780  0.5613481  16.825174  0.01294525

RMSE was used to select the optimal model using  the smallest value.
The final value used for the model was mtry = 15.
```

If you notice, you'll see I've used method = "parRF". This is parallel random forest. This implementation of random forest. This package causes your local machine to take le random forest computation. Alternatively, you can also use method = "rf" as a standard ran function.

Now we've got the optimal value of mtry = 15. Let's use 1000 trees for computation.

```
#random forest model
> forest_model <- randomForest(Item_Outlet_Sales ~ ., data = new_train, mtry =
= 1000)
> print(forest_model)
> varImpPlot(forest_model)
```

This model throws RMSE = 1132.04 which is not an improvement over decision tree mod forest has a feature of presenting the important variables. We see that the most important Item_MRP (also shown by decision tree algorithm).



This model can be further improved by tuning parameters. Also, Let's make out first submis
our best RMSE score by decision tree.

```
> main_predict <- predict(main_tree, newdata = new_test, type = "vector")
> sub_file <- data.frame(Item_Identifier = test$Item_Identifier, Outlet_Identif
test$Outlet_Identifier,       Item_Outlet_Sales = main_predict)
> write.csv(sub_file, 'Decision_tree_sales.csv')
```

When predicted on out of sample data, our RMSE has come out to be 1174.33. Here are s
you can do to improve this model further:

1. Since we did not use encoding, I encourage you to use one hot encoding and label encodir
random forest model.
2. Parameters Tuning will help.
3. Use Gradient Boosting (https://www.analyticsvidhya.com/blog/2015/09/complete-guide-
methods/).
4. Build an ensemble of these models. Read more about Ensemble Modeling
(https://www.analyticsvidhya.com/blog/2015/09/questions-ensemble-modeling/).

Do implement the ideas suggested above and share your improvement in the comme
below. Currently, Rank 1 on Leaderboard (http://datahack.analyticsvidhya.com/contes
problem-big-mart-sales-iii/lb) has obtained RMSE score of 1137.71. Beat it!

End Notes

This brings us to the end of this tutorial. Regret for not so happy ending. But, I've given y hints to work on. The decision to not use encoded variables in the model, turned out to be until decision trees.

The motive of this tutorial was to get your started with predictive modeling in R. We uncanny things such as 'build simple models'. Don't jump towards building a complex mo models give you benchmark score and a threshold to work with.

In this tutorial, I have demonstrated the steps used in predictive modeling in R. I've co exploration, data visualization, data manipulation and building models using Regressio Trees and Random Forest algorithms.

Did you find this tutorial useful ? Are you facing any trouble at any stage of this tutorial ? I mention your doubts in the comments section below. Do share if you get a better score.

Edit: On visitor's request, the PDF version of the tutorial is available for download. You nee a log in account to download the PDF. Also, you can bookmark this page for future Download Here (http://discuss.analyticsvidhya.com/t/download-free-tutorial-to-learn-da in-r-from-scratch/7797/2).

You want to apply your analytical skills and test your potential? Then participate in our Hackathons (http://datahack.analyticsvidhya.com/contest/all) and compete with T Scientists from all over the world.

Share this:

(https://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/?share=linkedin&nb=1) 1K+

(https://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/?share=facebook&nb=1)

(https://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/?share=google-plus-1&nb=1

(https://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/?share=twitter&nb=1)

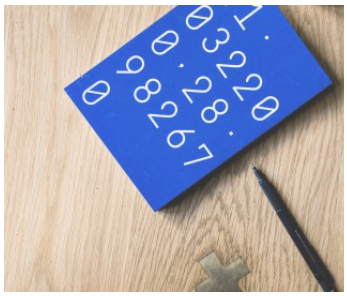
(https://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/?share=pocket&nb=1)

(https://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/?share=reddit&nb=1)

RELATED

(https://www.analyticsvidhya.com/blog/2016/09/most-active-data-scientists-free-books-notebooks-tutorials-on-github/)
Most Active Data Scientists, Free Books, Notebooks & Tutorials on Github
(https://www.analyticsvidhya.com/blog/2016/09/most-active-data-scientists-free-books-notebooks-tutorials-on-github/)
September 30, 2016
In "Machine Learning"

(https://www.analyticsvidhya.com/blog/2015/12/faster-data-manipulation-7-packages/)
Do Faster Data Manipulation using These 7 R Packages
(https://www.analyticsvidhya.com/blog/2015/12/faster-data-manipulation-7-packages/)
December 11, 2015
In "Business Analytics"



(https://www.analyticsvi
blog/2016/05/data-table
frame-work-large-data-se
data.table() vs data.fram
to work on large data set
(https://www.analyticsvi
blog/2016/05/data-table
frame-work-large-data-se
May 3, 2016
In "Machine Learning"

TAGS: AUTOCORRELATION (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/AUTOCORRELATION/), CARET PACKAGE (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/CARET-PACKAGE/), CROSS-VALIDATION (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/CI VALIDATION/), DATA EXPLORATION IN R (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/DATA-EXPLORATION-IN-R/), DATA MANIPULATION IN R (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/DATA-MANIPULATION-IN-R/), DATA MINING IN R (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG MINING-IN-R/), DECISION TREES IN R (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/DECISION-TREES-IN-R/), DPLYR (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/DPLYR/), FEATURE ENGINEERING IN R (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/FEA ENGINEERING-IN-R/), GGPLOT (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/GGPLOT/), HETEROSKEDASTICITY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/HETEROSKEDASTICITY/), HOMOSKEDASTICITY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/HOMOSKEDASTICITY/), LABEL ENCODING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/ ENCODING/), LINEAR REGRESSION IN R (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/LINEAR-REGRESSION-IN-R/), MULTICOLLINEARITY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/MULTICOLLINEARITY/), MULTIPLE REGRESSION IN R (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/MULTIPLE-REGRESSION-IN-R/), ONE HOT ENCODING (HTTPS://WWW.ANALYTICSVIDHYA.COM/B HOT-ENCODING/), OVERFITTING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/OVERFITTING/), PLYR PACKAGE

(HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/PLYR-PACKAGE/), PREDICTIVE MODELING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/ MODELING/), RANDOM FOREST IN R (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/RANDOM-FOREST-IN-R/), UNDERFITTING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/UNDERFITTING/)

Previous Article
Guide to Build Better Predictive Models using Segmentation (https://www.analyticsvidhya.com/blog/2016/02/guide-build-predictive-models-segmentation/)

Next Articl
Complete Guide to Parameter Tuning i XGBoost (with codes in Python (https://www.analyticsvidhya.com/blog/2016 guide-parameter-tuning-xgboost-with-code python/

(https://www.analyticsvidhya.com/blog/author/avcontentteam/)

Author

Analytics Vidhya Content Team (https://www.analyticsvidhya.com/blog/author/avcon

Analytics Vidhya Content team

105 COMMENTS

Steve (http://www.bigewisdom.net/) says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
FEBRUARY 29, 2016 AT 3:46 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN- SCRATCH/#COMMENT-106335)

Thanks for sharing! Can this content be available in a Pdf format?
Thanks,

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
FEBRUARY 29, 2016 AT 6:13 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUT
DATA-SCIENCE-SCRATCH/#COMMENT-106343)

Welcome Steve. I can make that available. I'll email it to you shortly.

Abhijit says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?
REPLYTOCOM=106351#RESPOND)
FEBRUARY 29, 2016 AT 7:45 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/C
TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/#COMMENT-106351)

Please make it(PDF version) available for all the users as well. It wil
in a nutshell.

Hemant says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?
REPLYTOCOM=106366#RESPOND)
FEBRUARY 29, 2016 AT 11:07 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/(
TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/#COMMENT-106366)

Manish nice content for Beginners. Thanks ! I also want this conten
format. Please mail this content in PDF format to me also.

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-S
REPLYTOCOM=106382#RESPOND)
FEBRUARY 29, 2016 AT 3:19 PM
(HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIA
SCIENCE-SCRATCH/#COMMENT-106382)

Hi Hemant
PDF is available for download. Link is added in the tutori
end.

Hemant says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN
SCRATCH/?REPLYTOCOM=106710#RESPOND)
MARCH 6, 2016 AT 1:48 PM
(HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMP
LEARN-DATA-SCIENCE-SCRATCH/#COMMENT-106710)

Manish taht link is not working

http://discuss.analyticsvidhya.com/t/downloa
tutorial-to-learn-data-science-in-r-from-scrat
(http://discuss.analyticsvidhya.com/t/downlo
tutorial-to-learn-data-science-in-r-from-scrat

Please see it.

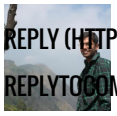
Analytics Vidhya Content Team says:
MARCH 8, 2016 AT 5:50 AM
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=106842#RESPOND) (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2 TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/#COMI

Hi Hemant
Link is working fine. You need to cre
time user login to download the PDI

Ajit Yadav says:
JUNE 20, 2016 AT 1:50 AM
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN SCRATCH/?REPLYTOCOM=112402#RESPOND) (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMP LEARN-DATA-SCIENCE-SCRATCH/#COMMENT-112402)

Hi Manish,

We are looking for R language experts with go
understanding on Data Science. Required an e
write a book on R language using Data Scienc
Interested writers/experts please contact with
profile at alpinessolutions at gmail dot com.



midhun1992 says:
MARCH 10, 2016 AT 11:45 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COM TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/#COMMENT-107001)
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=107001#RESPOND)

Sir, I couldn't find the datasets mentioned in the article. Can you pl
me where can i get the data sets. Thanks.

Elan says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=111173#RESPOND)
MAY 19, 2016 AT 8:59 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLE
LEARN-DATA-SCIENCE-SCRATCH/#COMMENT-111173)

Please advise how to download the data set
Couldn't find the link after having logged in to your site

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-S
REPLYTOCOM=112097#RESPOND)
JUNE 10, 2016 AT 11:07 PM
(HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIA
SCIENCE-SCRATCH/#COMMENT-112097)

Hi Elan
Please download the data from here:
http://datahack.analyticsvidhya.com/contest/practice-p
big-mart-sales-iii
(http://datahack.analyticsvidhya.com/contest/practice-
big-mart-sales-iii)

bgreddy says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=
SEPTEMBER 16, 2016 AT 6:18 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUT
DATA-SCIENCE-SCRATCH/#COMMENT-116143)

plz mail pdf on gangadharreddyb@hotmail.com (mailto:gangadharreddyb@h

Dr.D.K.Samuel says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
FEBRUARY 29, 2016 AT 4:09 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-
SCRATCH/#COMMENT-106336)

Nice writeup useful, thnaks Samue

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
FEBRUARY 29, 2016 AT 6:13 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUT
DATA-SCIENCE-SCRATCH/#COMMENT-106344)

Welcome Samuel !

Himanshu Dhingra (http://www.gutargoo.com) says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
FEBRUARY 29, 2016 AT 6:35 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-
SCRATCH/#COMMENT-106346)

Thanks Manish. You wrote an amazing article for beginners. I was looking for an article l
which clears the basics of R without refering to any books and all.

Even I request you to send me the doc or pdf of this so that i can get it print to make it h
read.

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
FEBRUARY 29, 2016 AT 3:21 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUT
DATA-SCIENCE-SCRATCH/#COMMENT-106384)

Thanks Himanshu ! PDF is available for download. Link is added at the end of

Krishna says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
FEBRUARY 29, 2016 AT 8:25 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-
SCRATCH/#COMMENT-106353)

good one. pl mail me a pdf as well

Devendra Yadav says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
FEBRUARY 29, 2016 AT 9:26 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-
SCRATCH/#COMMENT-106359)

Hi Manish

Could you please share the pdf with me as well. I am a starter in R and this can help as
guide for myself when trying out different things.

Thanks

Rad Mou says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
FEBRUARY 29, 2016 AT 9:33 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-
SCRATCH/#COMMENT-106360)

Hello, when I type log(12) I get 2.484907 as a result. What seems to be the problem ?

Ram says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
FEBRUARY 29, 2016 AT 4:38 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUT
DATA-SCIENCE-SCRATCH/#COMMENT-106391)

@RadMou,

It seems that there is a typo in the article. The fact is: 'log uses base e' ; log10
10' and 'log2 uses base 2'.

You can see that these commands print different values:
log(12) # log to the base e
log10(12) # log to the base 10
log2(12) # log to the base 2

Hope this helps.

Zamin Sherazi says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
FEBRUARY 29, 2016 AT 9:58 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-
SCRATCH/#COMMENT-106361)

Thanks Manish . would be grateful if can be made available in PDF .

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
FEBRUARY 29, 2016 AT 3:19 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUT
DATA-SCIENCE-SCRATCH/#COMMENT-106383)

Hi Zamin
PDF is available for download.

Monil Doshi says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
FEBRUARY 29, 2016 AT 10:23 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN
SCRATCH/#COMMENT-106364)

Hi Manish,
This is very helpful for beginners like me.
Looking forward for more.
Is there any way I can get this in PDF format?
It would be really helpful
My email id is monid@gmail.com (mailto:monid@gmail.com).
Thank you very much!.

Aanish says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
FEBRUARY 29, 2016 AT 11:19 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-
SCRATCH/#COMMENT-106367)

Thanks Manish. This is a great help! I have a questions – I noticed that R automatically t
of the factor variables (by converting them to n or n-1 dummy variables) while performi
regression. Do you recommend that we do it explicitly?

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
FEBRUARY 29, 2016 AT 3:01 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUT
DATA-SCIENCE-SCRATCH/#COMMENT-106379)

Hi Anish

In case of linear regression, decision trees, random forest, kNN, it is not nece
convert categorical variables explicitly as these algorithms intrinsically break
categorical variables with n – 1 levels. However, if you are using boosting alg
(GBM, XGboost) it is recommended to encode categorical variables prior to n
On a similar note, if you have followed this tutorial you'll find that I started wit
encoding and got a terrible regression accuracy. Later, I used the categorical
as it as, and accuracy improved.

---

kishor says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
FEBRUARY 29, 2016 AT 11:55 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-
SCRATCH/#COMMENT-106368)

good presentation. can you please provide it in pdf format.

---

chandrakala (http://) says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
FEBRUARY 29, 2016 AT 12:15 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-
SCRATCH/#COMMENT-106369)

Very helpful for beginners, thanks a lot!!!! keep it up.

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
FEBRUARY 29, 2016 AT 3:17 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUT(
DATA-SCIENCE-SCRATCH/#COMMENT-106381)

Welcome !

Raman says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
FEBRUARY 29, 2016 AT 1:26 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-D
SCRATCH/#COMMENT-106372)

Manish,

Very valuable tutorial. TY. If it is not too much of a trouble. Can you please make a PDF
link on the tutorial, please. Thanks.

Regards
Raman

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
FEBRUARY 29, 2016 AT 3:17 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUT(
DATA-SCIENCE-SCRATCH/#COMMENT-106380)

Hi Raman

I've added the PDF link at the end of this tutorial.

Atul Khairnar says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
FEBRUARY 29, 2016 AT 1:53 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-D
SCRATCH/#COMMENT-106373)

Thanks for sharing this article. This is really help to us. When I ran these script on Rstud
errors for ggplot after I tried " install.packages("ggplot2") AND
"install.packages('ggplot2',dependencies = TRUE) "and I got the following error
> ggplot(train, aes(x= Item_Visibility, y = Item_Outlet_Sales)) + geom_point(size = 2.5, colo
xlab("Item Visibility") + ylab("Item Outlet Sales")
Error: could not find function "ggplot"

And also for merge data
> combi <- merge(b, combi, by = "Outlet_Identifier")
Error in fix.by(by.x, x) : 'by' must specify a uniquely valid column

Can you help me why this happen.

Once again 'Thank You So Much' because I learn new things about R.

Thanks,
Atul

Analytics Vidhya Content Team says:

REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
FEBRUARY 29, 2016 AT 2:42 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUT
DATA-SCIENCE-SCRATCH/#COMMENT-106377)

Hi Atul

After installing the ggplot2 package, you should call the package in the next
library(ggplot2).
Then run the ggplot code, it should work.

merge function is used from package plyr. Have you installed it ? Let me kno

---

Atul Khairnar says:

REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?
REPLYTOCOM=106421#RESPOND)
MARCH 1, 2016 AT 6:39 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMP
LEARN-DATA-SCIENCE-SCRATCH/#COMMENT-106421)

Thanks Manish, I tried manually as well as by syntax through but st
following error

install.packages("plyr")
library(plyr)
combi library(plyr)
Warning message:
package 'plyr' was built under R version 3.1.3
> combi <- merge(b, combi, by = "Outlet_Identifier") ##########E
showing####
Error in fix.by(by.x, x) : 'by' must specify a uniquely valid column

Can you please help me on this…why this error showing…

Arfath says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-S
REPLYTOCOM=106911#RESPOND)
MARCH 9, 2016 AT 8:01 AM
(HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIA
SCIENCE-SCRATCH/#COMMENT-106911)

its not combi library(plyr) but it's only library(plyr) …
1 more thing i want to correct here is in
combi <- merge(b,combi, by = "Outlet_Identifier")
its not Outlet_Identifier but it is Item_identifier..
so correct code is ..
combi <- merge(b,combi, by = "Item_Identifier")
hope this helps you out….

shashi says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
FEBRUARY 29, 2016 AT 2:48 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-
SCRATCH/#COMMENT-106378)

can u share any material of data science

mouradelghissassi1992 says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
FEBRUARY 29, 2016 AT 3:37 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-
SCRATCH/#COMMENT-106388)

Erratum : I'm not sure if the problem is from my computer, but :

– When I execute head(b) I get :
DRA12 9
RA24 10

And not

OUT027 2215.876
OUT035 1463.705

So the command
combi <- merge(b, combi, by = "Outlet_Identifier") should be
combi <- merge(b, combi, by = "Item_Identifier") instead

– Also in head(c) there is a problem with the years, all rows are for 1985.

---

mouradelghissassi1992 says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
FEBRUARY 29, 2016 AT 4:40 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUT
DATA-SCIENCE-SCRATCH/#COMMENT-106392)

"Hence, we see that column Item_Visibility has 1463 missing values. Let's get
inferences from this data." it's the Item_Weight variable that has missing valu

Also in "Label Encoding and One Hot Encoding" : the variable Item_Visibility h
Low Fat and Regular

It's Item_Fat_Content not Item_Visibility

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?
REPLYTOCOM=106410#RESPOND)
MARCH 1, 2016 AT 4:02 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMP
LEARN-DATA-SCIENCE-SCRATCH/#COMMENT-106410)

Hi
Thank you so much! Editing error. Rectified now.

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=
MARCH 1, 2016 AT 4:22 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIA
SCIENCE-SCRATCH/#COMMENT-106411)

Hi
Thanks for pointing out. Made the changes.
In head(c), I wanted to show that using the "mutate" command, count value o
automatically aligned to their particular year value. Hence, I sorted it. For exa
year 1985 would get 25 as count value at all the places in count column. Any
put a better picture of year count now.
Hope this helps.

Balaji says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 1, 2016 AT 11:42 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA
SCRATCH/#COMMENT-106434)

Hi Manish,

I am unable to download the pdf as i get a blank page. Kindly check

---

balajimadhav says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 1, 2016 AT 11:49 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORI
SCIENCE-SCRATCH/#COMMENT-106435)

Thanks. Works now after i relogin

Ambuj Sharma says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 1, 2016 AT 12:29 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DAT
SCRATCH/#COMMENT-106439)

Hii,
When I use full_join for Outlet Years my rowcount increase to 23590924. I did not under
full join is used and why rowcount is increasing.

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 1, 2016 AT 1:09 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIA
SCIENCE-SCRATCH/#COMMENT-106442)

Hi Ambuj
full_join function returns all rows and all columns from the chosen data sets.
value is not present it blatantly returns NA. In your case, you might not have s

the “by” parameter in full_join.

---

ginisk sam says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 4, 2016 AT 1:42 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIA
SCIENCE-SCRATCH/#COMMENT-106546)

What I did was after c which has 14204 rows as flws :

d %
group_by(Outlet_Establishment_Year)%>%
distinct()

then combi <- merge(d, combi, by = "Outlet_Establishment_Year")

combi will now be ready for label encoding…

---

ginisk sam says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 4, 2016 AT 6:25 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIA
SCIENCE-SCRATCH/#COMMENT-106554)

Dear Ambuj,

After generated c .. i created d using distinct

d %
group_by(Outlet_Establishment_Year)%>%
distinct()

Then merge d with combi as flws :
combi <- merge(d, combi, by = "Outlet_Establishment_Year")

Then ready for encoding.

Thanks

Ambuj Sharma says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=106843#RESPOND)
MARCH 8, 2016 AT 5:51 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMP
LEARN-DATA-SCIENCE-SCRATCH/#COMMENT-106843)

Thanks!

gaurav says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 2, 2016 AT 5:01 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA
SCRATCH/#COMMENT-106463)

Hi ,

Can you please send me the pdf file on gauravborse1988@yahoo.co.in (mailto:gauravborse1988@yahoo.co.in) as i am unable to download the file from the link

Thanks in advance

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 2, 2016 AT 9:08 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIA
SCIENCE-SCRATCH/#COMMENT-106476)

Hi Gaurav,

As mentioned, you need to create a one-time user account to download the can find the link in the End Notes.

Jhanak Sharma1 says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?
REPLYTOCOM=106708#RESPOND)
MARCH 6, 2016 AT 11:47 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COM
TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/#COMMENT-106708)

Problem no.1 :
When I execute head(b) I get :
Item_Identifier Item_Count
(fctr) (int)
1 DRA12 9
2 DRA24 10

And not

OUT027 2215.876
OUT035 1463.705

I tried below command but again error:
> combi <- merge(b, combi, by = "Outlet_Identifier")
Error in fix.by(by.x, x) : 'by' must specify a uniquely valid column

Problem No.2 :
When I execute table(q)
I get:
Drinks Food Non-Consumable
2180488 16949063 4461373
and not
Drinks Food Non-Consumable
1317 10201 2686

Problem No.3 :
combi <- dummy.data.frame(combi, names =
+ c('Outlet_Size','Outlet_Location_Type','Outlet_Type', 'Item_Type_N
sep='_')

Error: cannot allocate vector of size 256.0 Mb
In addition: Warning messages:
1: In anyDuplicated.default(row.names) :
Reached total allocation of 3947Mb: see help(memory.size)
2: In anyDuplicated.default(row.names) :
Reached total allocation of 3947Mb: see help(memory.size)

Q. How to deal with Error: "cannot allocate vector of size"?

Please help me for solutions to the problems stated above

Analytics Vidhya Content Team says:
MARCH 8, 2016 AT 7:04 AM
(HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIA
SCIENCE-SCRATCH/#COMMENT-106848)

REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-S
REPLYTOCOM=106848#RESPOND)

Hi Jhanak
Thank you so much for pointing this out.
Answer 1: The code is correct. The output I used requirec
done now. You can check.
Answer 2: I'll require your code to answer it. Because, I've
again at my side, the output of table(q) is

Drinks Food Non-Consumable
1317 10201 2686
Answer 3: Looks like your Problem 2 and Problem 3 are r
After you combine the data set, check the dimension of
set. It should be 14204 rows and 12 columns.Looks like y
data set has too many observations. Usually, memory m
issues are solved using 2 ways. First, by upgrading mach
specifications. Second, by using sparse matrix for compu
Also, while using R and doing computation, it is advisabl
other programs which are not necessary, especially chro
This will allow R to compute faster.

midhun1992 says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-S
REPLYTOCOM=107334#RESPOND)
MARCH 14, 2016 AT 9:13 AM
(HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIA
SCIENCE-SCRATCH/#COMMENT-107334)

Hi Janak, the dataset is not available now. It seems you h
on the dataset. Can you please share the dataset to
v.07.midhun@gmail.com (mailto:v.07.midhun@gmail.com
be of great help. Thanks.

Venu says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 6, 2016 AT 10:43 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DAT
SCRATCH/#COMMENT-106698)

Could you please share the data (…./Data/BigMartSales) that you have used here so th
play it with ?

Venu says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 6, 2016 AT 10:48 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DAT
SCRATCH/#COMMENT-106699)

It seems that your PDF file is missing in the correct link. May I request you to update it.
advance….

Fred says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 6, 2016 AT 11:00 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DAT
SCRATCH/#COMMENT-106702)

I got the PDF file, Thanks…

---

buvana says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=
MARCH 7, 2016 AT 9:27 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA
SCRATCH/#COMMENT-106791)

nice tutorial. I have 2 questions so far
a) how to save my work – for e.g all the data manipulation steps i did are lost the next d
have to start from the setwd(path) command again
b) what is the difference between merge and full_join in the tutorial? when is each com
appropriate?
c) The group by Item_identifier is not working properly. The sample output is wrong

---

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 8, 2016 AT 7:16 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIA
SCIENCE-SCRATCH/#COMMENT-106849)

Hi Buvana
Answer a) Do you directly write codes in console ? Use R Studio. You should
script as they can be saved in .R format and helps you to retrieve codes at lat
more information, check the first section of this tutorial.
Answer b) full_join is used when we wish to combine two columns. It return N
matching value are found. merge is used when we wish to combine two colu
on a column type. In full_join, you don't need to specify "by" parameter.
Answer c) Thank for pointing out. Sorted now.

Guilherme Gadori says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 9, 2016 AT 4:00 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA
SCRATCH/#COMMENT-106897)

Hi,

In the Random Forest section, could you please explain why did you use ntree = 1000 a
mtry = 15?

Cheers,

Analytics Vidhya Content Team says:

REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 10, 2016 AT 7:07 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORI
SCIENCE-SCRATCH/#COMMENT-106982)

Hi Guilherme

If you carefully check random forest section, I've initially done cross validatio
caret package. Cross validation provided the optimal value of mtry and ntree
the RMSE is least (check output). I, then used those parameters in the final ra
forest model. Another method to choose mtry and ntree is hit and trial, which
time consuming and inconsistent. You may try this experiment at your end, a
know if you obtain lesser RMSE than what I've got.

Guilherme Cadori says:

REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?
REPLYTOCOM=107004#RESPOND)
MARCH 10, 2016 AT 12:46 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/CO
TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/#COMMENT-107004)

Hi Manish,

Thank you for your attention. I understood how you got mtry. Howe
output printed in this tutorial, there's no valeu regarding ntree (e.g.
which was the value you used later on). How did you get it?

Thanks,

Arfath says:

REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 9, 2016 AT 12:58 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DAT
SCRATCH/#COMMENT-106925)

Thank you very much for this wonderful and unique post. i came to this site to participa
with your data" competition. i was puzzled looking at the datsets like train,test and sam
have any idea what,and how to solve this. later on i came across this post (thank God i c
really after going through your post i gained confidence & i got a clear picture on how t
these competitions. once agian thanx from bottom of my heart.since i m completely ne
have few doubts…
1) in "linear_model <- lm(Item_Outlet_Sales ~ ., data = new_train)" what does tilde(~) follo

(.) means?
2) what is the best RMSE score for any model?
3) so both train and test datsets are same,only thing is test data doesnt have response v
But, if we do know the response variable value from train dataset, again why we we are
it for test data set? is it because we want to construct a model which predicts the future
but we want to test how good our model predicts value, so thats why we took sample f
dataset and cross check our predicted values with that of main dataset ? correct me if r
understanding is wrong…

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 10, 2016 AT 8:09 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORI
SCIENCE-SCRATCH/#COMMENT-106986)

Hi Arfath
Good to know that you have started learning.
Answer 1: tilde(~) followed by dot (.) tells the model to select all the variables
Otherwise, it would be so much inconvenient to write name of all variables o
Imagine the time which would get wasted if you have got 200 variables to wr
Therefore, use this short sign tilde(~) followed by dot (.)

Answer 2: Ideally, every model strives for achieve RMSE as much as close to
Because, Zero means your model has accurately predicted the outcome. But
possible. Since, every model has got irreducible error which affects the accur
the best RMSE score is the least score you can get.

Answer 3: You are absolutely. Train data set has response variable and a mod
on that. This model gives you a fantastic RMSE score. But, it is worthless until
with same accuracy on out of sample data. The ultimate aim for this model is
future predictions. Right ? Hence, test data is used to check out of sample ac
the model. If the accuracy is not as good as you achieved on train data set, it
that overfitting has taken place.

I would recommend you to read Introduction to Statistical Learning. Downloa
available in my previous article: http://www.analyticsvidhya.com/blog/2016/
read-books-statistics-mathematics-data-science/
(http://www.analyticsvidhya.com/blog/2016/02/free-read-books-statistics-
mathematics-data-science/)

---

vijaynk10 says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 9, 2016 AT 5:43 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA
SCRATCH/#COMMENT-106936)

I am a little late to the game. How do i download the BigMartSales data?

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 10, 2016 AT 6:58 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORI
SCIENCE-SCRATCH/#COMMENT-106981)

Hi Vijay

Link is available in the tutorial.

Idea4Life says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?
REPLYTOCOM=107008#RESPOND)
MARCH 10, 2016 AT 1:59 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COM
TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/#COMMENT-107008)

Sorry Manish. The link i believe you are mentioning is “Big Mart Sal
Prediction”. But when i go into it, it says “The dataset is accessible c
contest is active.” Can you please check and clarify?

Thanks,
Vijay

VK says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?
REPLYTOCOM=107105#RESPOND)
MARCH 11, 2016 AT 7:04 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COM
TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/#COMMENT-107105)

Sorry Manish. Tried from the link “Big Mart Sales Prediction” in the c
But when i go to the link Data Set, it shows up the following messa
“The dataset is accessible only if the contest is active.”

Can you please validate again?

Thanks.

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-S
REPLYTOCOM=107145#RESPOND)
MARCH 12, 2016 AT 5:10 AM
(HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIA
SCIENCE-SCRATCH/#COMMENT-107145)

Hi Vijay

The contest will get active again from tomorrow (13th Ma
Regret the inconvenience caused.

Alfa says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 10, 2016 AT 9:52 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DAT
SCRATCH/#COMMENT-106993)

Thanks for sharing.
I just can not understand what the One Hot Encoding means and how to use it. Becaus
here.
Thanks!

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=
MARCH 12, 2016 AT 5:18 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORI
SCIENCE-SCRATCH/#COMMENT-107148)

Hi Alfa
One Hot Encoding is nothing but, splitting the levels of a categorical variable
variable. The new variables will be encoded with 0s and 1s. 1s represent the
information. 0s represent the absence of information.
For example: Suppose, we have a variable named as Hair Color. It has 3 level
Red Hair, Black Hair, Brown Hair. Doing one hot encoding of this variable, will
different variables namely Red Hair, Black Hair, Brown Hair. And, the original
Color will be removed from data set.
If someone has Red Hair, Red Hair variable will be 1, Black Hair will be 0, Brov
be 0.
If someone has Black Hair, Red Hair variable will be 0, Black Hair will be 1, Br
will be 0.
If someone has Brown Hair, Red Hair variable will be 0, Black Hair will be 0, B
will be 1.

This is One Hot Encoding.

Prateek says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=107263#RESPOND)
MARCH 13, 2016 AT 2:10 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/#COMMENT-107263)

Hi Manish,

Is it advisable to use One hot encoding when there is huge numbe
a categorical variable ?

midhun1992 says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 10, 2016 AT 11:48 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DAT
SCRATCH/#COMMENT-107002)

Can someone please mail me the data sets we need for this article to v.07.midhun@gm
(mailto:v.07.midhun@gmail.com). I couldn't find at the mentioned location. It would be re
helpful. Thanks

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=
MARCH 12, 2016 AT 5:12 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORI
SCIENCE-SCRATCH/#COMMENT-107147)

Hi Midhun
The data set will be available for download from tomorrow onwards (13th Ma
Regret the inconvenience caused.

midhun1992 says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=107323#RESPOND)
MARCH 14, 2016 AT 7:27 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/#COMMENT-107323)

Hi Manish, sorry to bother you but it seems the data set is still unav
it's not too much trouble, can you please mail the data to
v.07.midhun@gmail.com (mailto:v.07.midhun@gmail.com)

maneilakki7 says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 11, 2016 AT 4:18 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA

SCRATCH/#COMMENT-107069)

Hi Manish,
It's a great article & gives a good start for beginner like me. Can you please share the da
download it from the link as the contest is not active.

Thank You

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=
MARCH 12, 2016 AT 5:11 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIA
SCIENCE-SCRATCH/#COMMENT-107146)

Hi Manoj
The data set will be available for download from tomorrow onwards. (13th Ma

---

Roy Basan (http://none) says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=
MARCH 16, 2016 AT 12:08 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DA
SCRATCH/#COMMENT-107491)

Good Day…When I try to instal library(swirl) n R studio console ,,it states its not found in
R.3..2.4.. I got errors which states"Warning in install.packages :
package 'library(swirl)' is not available (for R version 3.2.4)"
Can somebody explain to me this peculiarity and how can I sort it out…
Thanks

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 17, 2016 AT 4:12 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORI/
SCIENCE-SCRATCH/#COMMENT-107544)

Hi Roy
First you should install swirl package and then call it using library function. U
commands below.
> install.packages("swirl")
> library(swirl)

midhun1992 says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 17, 2016 AT 5:36 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DAT

SCRATCH/#COMMENT-107564)

Hi Manish, The datasets are available now. Thank you so much.

---

victorenclinx says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=
MARCH 17, 2016 AT 9:22 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DAT
SCRATCH/#COMMENT-107616)

I encounter problems to log in http://datahack.analyticsvidhya.com/signup (http://datahack.analyticsvidhya.com/signup)… Can you help me ?
I want to log in to then download the data set…

Thanks in advance

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 18, 2016 AT 5:37 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORI
SCIENCE-SCRATCH/#COMMENT-107636)

Hello

There were some technical updates going on at the server. Things are fine no
try again.
Regret the inconvenience caused.

Sourabh1987 says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 19, 2016 AT 7:56 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DAT
SCRATCH/#COMMENT-107799)

trying feature engineering of the outlet _establishment year ,but the code for merging i
lot of rows , i tried both merge as well full join .

JAYMIN says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=
MARCH 21, 2016 AT 11:48 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DAT
SCRATCH/#COMMENT-107918)

hello sir i am a fresher electrical engineer and my maths and logical thinking is good ca
data scientist sir give me some advice thanks

Roy Basan says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
MARCH 22, 2016 AT 7:33 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DAT
SCRATCH/#COMMENT-108023)

I did try to see the link to try the " Big Market Prediction" but unable to open it as it requ
membership. Now when I apply for the analytics Vidhya membership by signing up I go
Invalid Request twice … May I know how I can get over this issue.. Why I can't sign up..sc
continue with my R self tutorial work..

Hulisani says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
APRIL 5, 2016 AT 4:02 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-S
SCRATCH/#COMMENT-109002)

Hi

Thanks for an amazing article. Can you please email me the data used.

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=
JUNE 10, 2016 AT 11:09 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIA
SCIENCE-SCRATCH/#COMMENT-112100)

Hi Hulisani
Please download the data set from here:
http://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sa
(http://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-s

Priyanka Nath says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
APRIL 24, 2016 AT 7:20 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA
SCRATCH/#COMMENT-109960)

Hi,
I am facing a problem in Random Forest execution.
I am using R Studio (R version 3.2.4 Revised)
When I am trying to run the code;
> rf_model print(rf_model), it is returning error in this form :

Error in { : task 1 failed – "cannot allocate vector of size 554.2 Mb" In addition: Warning m

1: executing %dopar% sequentially: no parallel backend registered
2: In eval(expr, envir, enclos) :
model fit failed for Fold1: mtry=15 Error in { : task 1 failed – "cannot allocate vector of size

3: In eval(expr, envir, enclos) :
model fit failed for Fold2: mtry= 2 Error in { : task 1 failed – "cannot allocate vector of size

4: In eval(expr, envir, enclos) :
model fit failed for Fold2: mtry=28 Error in { : task 1 failed – "cannot allocate vector of siz

5: In eval(expr, envir, enclos) :
model fit failed for Fold3: mtry=15 Error in { : task 1 failed – "cannot allocate vector of size

6: In eval(expr, envir, enclos) :
model fit failed for Fold4: mtry= 2 Error in { : task 1 failed – "cannot allocate vector of size

7: In eval(expr, envir, enclos) :
model fit failed for Fold4: mtry=28 Error in { : task 1 failed – "cannot allocate vector of siz

8: In eval(expr, envir, enclos) :
model fit failed for Fold5: mtry=15 Error in { : task 1 failed – "cannot allocate vector of size

9: In nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo, :
There were missing values in resampled performance measures.
10: display list redraw incomplete
Timing stopped at: 1.26 0.3 2.49

Can you please suggest me any way out of this issue?

Priyanka Nath says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
APRIL 24, 2016 AT 7:23 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIA
SCIENCE-SCRATCH/#COMMENT-109961)

The code I am trying to run is :

rf_model <- train(Item_Outlet_Sales ~ ., data = new_train, method = "parRF", tr
control, prox = TRUE, allowParallel = TRUE)

print(rf_model)

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=
JUNE 10, 2016 AT 11:11 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-
SCIENCE-SCRATCH/#COMMENT-112101)

Hi Priyanka
Had I been at your place, I wouldn't have experimented with parallel random
this problem.
Why make things complicated when it can be done in a simple way!
Also, make sure that you drop the ID column before running any algorithm. T
should work fine then.

Raju says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
APRIL 26, 2016 AT 9:00 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA
SCRATCH/#COMMENT-110044)

Hi Manish,

After reading the whole article, I feel u have done a great job and have given more thar
data for a beginner.
I'm thankful to u for sharing all your solutions, this would give us different thought for us
with.

Regards,
Raju.

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
JUNE 10, 2016 AT 11:08 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL
SCIENCE-SCRATCH/#COMMENT-112099)

Glad it helped you. Thanks for your kind words Raju! 🙂

Gregory says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=
MAY 28, 2016 AT 11:50 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-
SCRATCH/#COMMENT-111547)

Good morning
I can not find the data set. Any suggestion?

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
JUNE 10, 2016 AT 11:07 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL
SCIENCE-SCRATCH/#COMMENT-112098)

Hi Gregory
Please download the data from here:
http://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sa
(http://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-s

---

Gregory says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=
MAY 28, 2016 AT 11:52 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-
SCRATCH/#COMMENT-111548)

OK. I've registered and I think it'll be OK.

Thanks

Toddhim says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
JUNE 10, 2016 AT 9:45 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-
SCRATCH/#COMMENT-112093)

I know this is months after this great article was published, but i'm just now working thr
and the BigMart Sales Prediction dataset isn't available. Is it available elsewhere?

Analytics Vidhya Content Team says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
JUNE 10, 2016 AT 11:06 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIA
SCIENCE-SCRATCH/#COMMENT-112096)

Hi Toddim,
The data set is very well available. I've already updated the links.
You can download the data from here:
http://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sa
(http://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-s

---

vipin says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=
JULY 12, 2016 AT 5:22 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-S
SCRATCH/#COMMENT-113378)

Hi Manish,
First of all thanks for a great article.
I encountered with a issue when I was running the code-
combi <- full_join(c, combi, by="Outlet_Establishment_Year")
it is giving me error-
Error: std::bad_alloc
what it is and how to correct this…

---

vipin says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=
JULY 12, 2016 AT 8:36 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-S
SCRATCH/#COMMENT-113383)

2.
combi <- dummy.data.frame(combi, names = c('Outlet_Size','Outlet_Location_Type','Outl
'Item_Type_New'), sep='_')
Error in sort.list(y) : 'x' must be atomic for 'sort.list'
Have you called 'sort' on a list?

---

simar says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=
JULY 31, 2016 AT 3:27 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-S
SCRATCH/#COMMENT-114293)

Hi Manish,

Can you please let me know what do you mean by Item_Fat_Content has mismatched levels?

Parul says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=
AUGUST 3, 2016 AT 5:50 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DAT
SCRATCH/#COMMENT-114413)

Hi Manish. Thanks for this article. Very well written and will help all. I have one query: I c
your post very well before'Graphical representation of Variables', after which I am unabl
out how to write these codes and what do they mean & signify, how to know which con
use & when? I am a beginner in R . Can you please suggest what to do in order for me t
understand all the steps from 'Graphical Representation'. This includes Data manipulati
Predictive modeling as well. Thanks a lot.

KarlWang says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=
AUGUST 4, 2016 AT 8:37 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DAT
SCRATCH/#COMMENT-114457)

Very great article and thank you so much for sharing your knowledge! I am not sure if o
some questions with me, but I list my questions. Hope you have some time to take a lo
Thank you again.

1. About the difference between label encoding and one hot encoding. For label encod
example is convert the 2 levels variables item_Fat_Content into 0 and 1. If I have a varia
state (50 levels = 50 States), is it means I just need simply trans the states to number 1-5
still a one variables, just from category to numerical, am I right?
2. For one hot encoding, I need split into 50 variables (50 States) and marked them as 0
indicate existence or non-existence, am I right?
3. So what is the advantage and disadvantage to convert the category variables into nu
variables? Why do we need to do this transformation?
4. In the article it said, 'We did one hot encoding and label encoding. That's not necessa
linear regression handle categorical variables by creating dummy variables intrinsically
we know which model we need to do the one hot encoding/ label encoding?
5. You mentation correlated variables. What level of correlation we need to remove the
variables? 0.5 or 0.6 or 0.7 ? And if two variables is correlated, how to decide which one
remove? Is there any standard about it?
6. I am running logistic regression, when I remove one of the correlated variables (0.68)
dropped, is it means this level (0.68) correlation is acceptable?
7. The liner regression model with funnel share means heteroscedasticity. So how to ev
logistic regression with Residuals vs Fitted graph?
8. In the article, it is said 'This model can be further improved by detecting outliers and
leverage points.' what is the technical to deal with these points? Just simply remove the
use the average to replace the value or other ways?
9. 'optimum cp value for our model with 5 fold cross validation.' In my mind, cross valida
for evaluate the model stability which is the last step. However, at here, we use cross va
optimum cp value, am I understand right?
10. Why are you using 5 fold cross validation instead of 4 fold or 6 fold or 10 fold?
11. When I running the model, it always have error told me the tree cannot split. Is there
requirement with the decision tree? Such as we cannot use category variables in decisi
12. How to do the Parameters Tuning for random forest? Could you points any arterials?
Thank you !!!

Monish Mathpal says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=

AUGUST 12, 2016 AT 6:53 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DA
SCRATCH/#COMMENT-114761)

In the below excerpts of the article:
""Data Frame: This is the most commonly used member of data types family. It is used t
tabular data. It is different from matrix. In a matrix, every element must have same class
data frame, you can put list of vectors containing different classes. This means, every c
data frame acts like a list. Every time you will read data in R, ""

it seems bit unconvincing that column of a dataframe acts like a list, instead column it s
row as per my understanding:

Vaibhav Gupta says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=
AUGUST 20, 2016 AT 6:59 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DA
SCRATCH/#COMMENT-114970)

Hi
I am beginner in Data Science using R. I was going through your well articulated article
Science using R. I was practicing your Big Mart Predication and got confused with one s
it checks the missing values in train data exploration. As per R and this tutorial , there is
missing values (i assume blank is being considered as missing data) in "Item_Weight" b
also missing in "Outlet_Size" in Train CSV.. But neither R or this tutorial is showing "Outle
missing values observations.

Can you please let me know how and why "Outlet_Size" is not considered as missing va
data exploration of train.

Vaibhav Gupta says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=
AUGUST 20, 2016 AT 8:43 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DA
SCRATCH/#COMMENT-114972)

Hi

I would also like to know what all mathematical concepts like algebra , statics, are requ
learn Data Science using R? Can anybody list down all mathematical concepts requirec
Science?

Thanks
Vaibhav Gupta

Iñigo says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=1
AUGUST 24, 2016 AT 1:28 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DA
SCRATCH/#COMMENT-115093)

Hello, I had an error when launching RStudio. I downloaded it again and installed it aga
when I downloaded for the second time I found this phrase:

"RStudio requires R 2.11.1 (or higher). If you don't already have R, you can download it he
a link)

So, before installing this, it looks like normal R has to be installed first. I write this in case
had the same problem.

Good job with the web, I really like it 🙂

Shuu says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM
AUGUST 25, 2016 AT 1:13 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DA
SCRATCH/#COMMENT-115131)

As someone who came from a non-coding background, you should know that small de
become HUGE hindrances in the learning process of a beginner.

On the Essentials part of the article, this code doesn't work:

> bar class(bar)
> "integer"
> as.numeric(bar)
> class(bar)
> "numeric"
> as.character(bar)
> class(bar)
> "character"

You have to actually set it as 'bar <- as.numeric(bar)' on the 4th line.

Please, keep those small things in mind. It is insanely difficult for someone like me to le
content, if things are any less than perfect, it really becomes impossible (I just spent alr
hour to figure out why I couldn't change the class of the object, and in the end, had to a
external help since I couldn't troubleshoot it myself).

Otherwise, great article, keep the great work up!

Cheers.

Joyce Salil says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-DATA-SCIENCE-SCRATCH/?REPLYTOCOM=
OCTOBER 24, 2016 AT 8:52 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/02/COMPLETE-TUTORIAL-LEARN-D
SCRATCH/#COMMENT-117457)

Thanks you made R programming simpler.
Could you please email the PDF of the same.

LEAVE A REPLY

Connect with:

(https://www.analyticsvidhya.com/wp-login.php?
action=wordpress_social_authenticate&mode=login&provider=Facebook&redirect_to=https%3A%
tutorial-learn-data-science-scratch%2F)

Your email address will not be published.

Comment

Name (required)

Email (required)

Website

SUBM

ABOUT US

For those of you, who are wondering what is "Analytics Vidhya", "Analytics" can be defined as the science of extracting insights from raw data. The spectrum of analytics starts from capturing data and evolves into using insights / trends from this data to make informed decisions. Read More (http://www.analyticsvidhya.com/about-me/)

STAY CONNECTED

8,173 FOLLOWERS (http://www.twitter.com/analyticsvidhya)

26,661 FOLLOWERS (http://www.facebook.com/Analyticsvidhya)

1,596 FOLLOWERS (https://plus.google.com/+Analyticsvidhya)

Email SUBSCRIBE (https://feedburner.google.com/fb/a/mailverify?uri=analyticsvidhya)

LATEST POSTS

(https://www.analyticsvidhya.com/blog.
data-scientist-clustering/)

40 Questions to test a Data S Clustering Techniques (Skill Solution) (https://www.analyticsvidhy 2017/02/test-data-scientist-

SAURAV KAUSHIK , FEBRUARY 5, 201

(https://www.analyticsvidhya.com/blog.
must-know-questions-on-base-sas-for-
data-scientists-out-there-skilltest-soluti

40 must know Questions on Analysts (Skill test Solution) (https://www.analyticsvidhy 2017/02/40-must-know-ques base-sas-for-analysts-data-sc there-skilltest-solution/)

SUNIL RAY , FEBRUARY 5, 2017

Basics of Probability for Dat

(https://www.analyticsvidhya.com/blog/
probability-data-science-with-examples

explained with examples (https://www.analyticsvidhy 2017/02/basic-probability-d with-examples/)

DISHASHREE GUPTA , FEBRUARY 2, 2

(https://www.analyticsvidhya.com/blog/
practical-guide-inferential-statistics-data

Comprehensive & Practical I Statistics Guide for data sci (https://www.analyticsvidhy 2017/01/comprehensive-prac inferential-statistics-data-sc

NSS , JANUARY 31, 2017
