**Practical No.:- 6**

Data Preprocessing Techniques

This reference Material is created for Mumbai university MCA Course for ADBMS.

Topics Covered are Implementation of Data preprocessing techniques like,

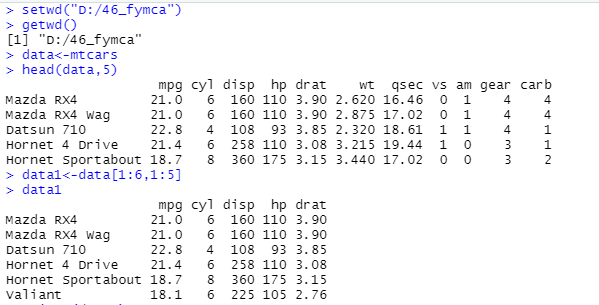
1. Naming and Renaming variables, adding a new variable.

2. Dealing with missing data.

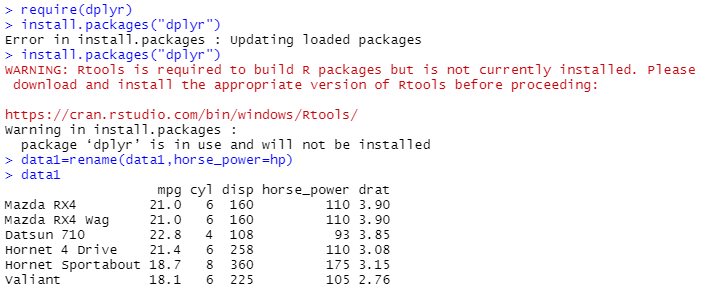
3. Dealing with categorical data.

4. Data reduction using sub setting.

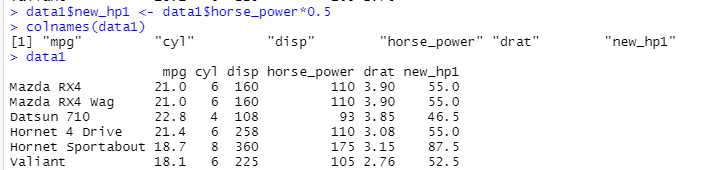
#my\_data



## Renaming columns with dplyr::rename()

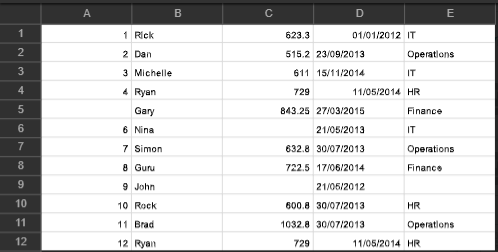


## Adding new variable



#naming variable

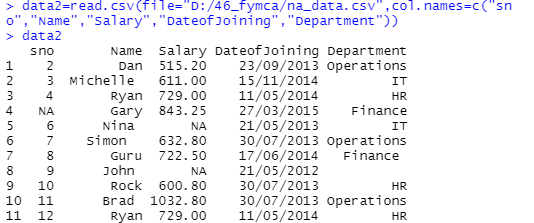
Create missing\_col1.csv file with following data.



#Reading with read.table() assumes no headers by default. First few lines :



#V1, V2, V3.. are given as default names (titles) by R



Error Detection and Correction NA: Not Available - Known as missing values

Works as a place holder for something that is ‘missing’

Most basic operations (addition, subtraction, multiplication, etc.) in R deal with it without crashing and return NA if one of the inputs is NA

is.na(VALUE) is used to check if the input value is NA or not. Returns a TRUE/FALSE vector Whereas in case of Excel like utilities for numeric computations it’s assumed to be 0.

# Operation with NA

NA+4

[1]NA

# Create a vector V with 1 NA value

v <- c(1,2,NA,3)

# Median with and without NA (remove NA)



# On removing NAs



# Apply is.na() to vector

>is.na(V)

[1] FALSE FALSE TRUE FALSE

# Removing the NA values by using logical indexing



# Get values that are not NA

V[!naVals]

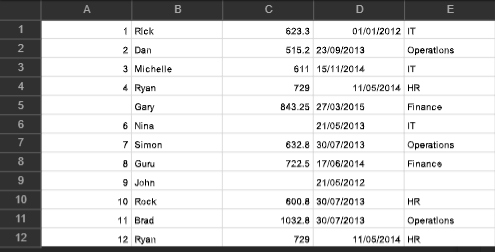
[1] 1 2 3

# Subsetting with complete cases - values that are not NA

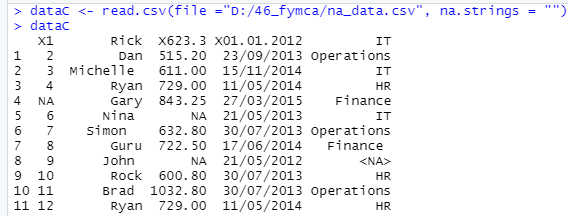
V[complete.cases(V)]

[1] 1 2 3

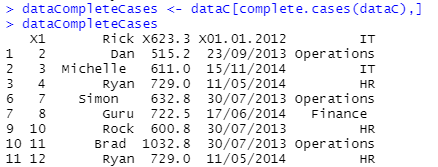
#Create na\_data.csv file with following data.



# Subsetting a data frame with complete cases # Complete Data of Prime Ministers. Notice NAs



# Subset only the rows without NA



Imputation

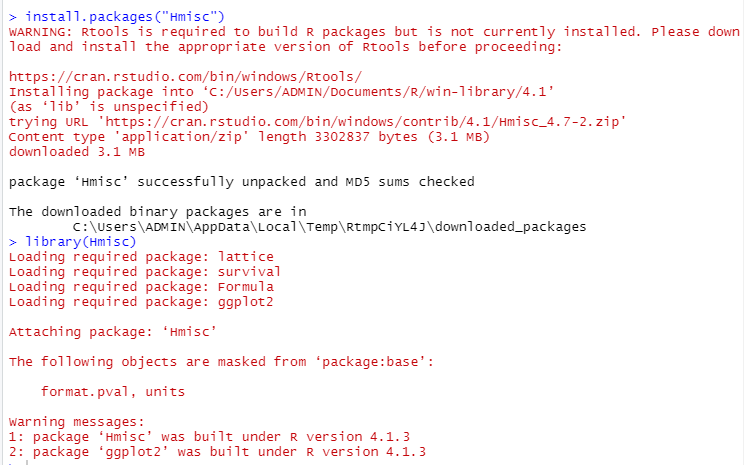
The process of estimating or deriving missing values There are various methods for imputation

– Imputation of the mean

– Imputation of the median

– Imputation using linear regression models

• Package Hmisc implments many imputation methods, few examples :



## create a vector



# mean imputation - from package, mention name of function to be used



#median imputation



\*\* Categorical Data \*\*

## Factors are variables in R which take on a limited number of different values; such variables are often referred to as categorical variables.

#Convert Character into Factor(categorical data)

# Create gender vector



# Convert gender\_vector to a factor



# Create Ordinal categorical vector

****

# Convert `day\_vector` to a factor with ordered level

****

# Print the new variable

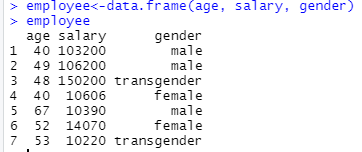


# Convert Numeric to Factor

# Creating vectors



# Creating data frame named employee



# Creating a factor corresponding to age with labels