Case Study on Graphs and Charts using R

Fraud is the misappropriation of funds . This case study deals with understanding the data thru graphs and charts to pin point possible fraud cases.

About the data :- Council contracts register; Contract expenditure over $5000

What is this data?

This dataset lists details of all procurement spending over $5,000 including purchase orders and tenders.

Definitions:- The dataset contains the following fields that you should use :

|  |  |
| --- | --- |
| Effective Date | Date of data pull |
| Contract Reference Number | Contract id |
| Service responsible | Area of contract |
| Contract Start Date | Start date |
| End Date | End date . If missing, assume ongoing |
| Review Date | Date of supervisory review |
| Extension Period | No of days |
| Contract Value | Amount |

Problem statement :-

Do an exploratory data analysis to understand the data better. Describe the date and number fields and understand the distributions .

Create contingency tables, graphs and charts to better understand the data.

Do some projects finish significantly earlier that the date expected ?

What is the per day rate at an average for projects ?

Note :- Fraud in these contracts can be of 2 types .

1. If the project can be completed in say 10 days but the contractor quotes 25 days . He then finished the work in 10 days and pockets the money for the 15 days .

2. If the project can be completed @ $1 per day of work and the contractor charges $5 per day of work

Solution in R

1. Define :- The project is to create graphs and charts to understand cases which could signify fraud defined as

1.1 Lower than normal time to complete the project

1.2 Higher than normal cost per delivery day of project

2. Collect :- The data is given. No other data is required to be referenced

3. Organise :-

3.1 Missing Values

3.2 Outliers

3.3 Non numeric to numeric data

4. Visualise the Y variables across significant segments of 'Service Responsible'

**R Code and Solution**

1. Import data :-

setwd('H:/springer book/Case study/CaseStudy3')

#IMPORT DATA

mydata <- read.table("H:/springer book/Case study/CaseStudy3/Contacts register Aug 2015 and purchase order over 5000 April to June 2015.csv", header=TRUE, sep=',')

#DIMENSIONS OF DATA (rows and obs)

dim(mydata)

> dim(mydata)

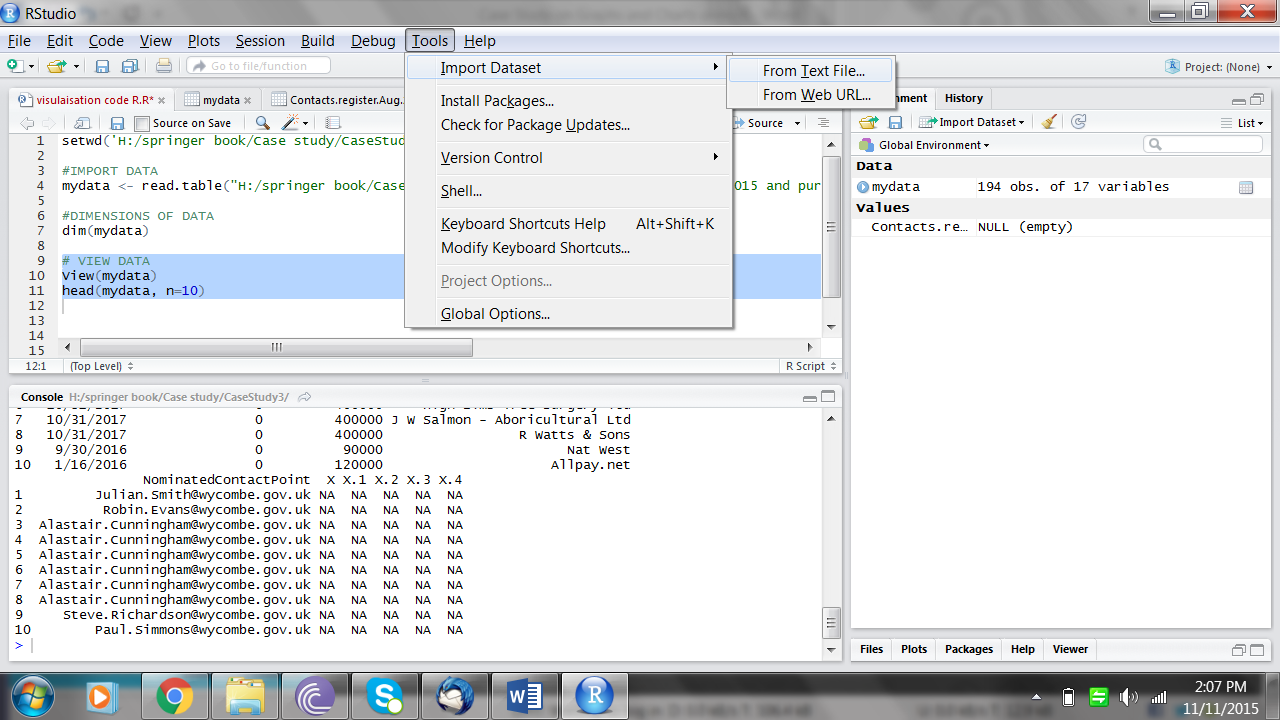
[1] 194 17

# VIEW DATA

View(mydata)

head(mydata, n=10)

NOTE ALTERNATE IMPORT METHOD:- Use the button Tools <Import Dataset< From Text file



1. View the contents of the R file and understand the variables

# VIEW CONTENTS

str(mydata)

> str(mydata)

'data.frame': 194 obs. of 17 variables:

$ Effective.Date : chr "August 2015" "August 2015" "August 2015" "August 2015" ...

$ Contract.Reference.Number : chr "342" "228" "289" "290" ...

$ Title.of.agreement : chr "Air quality Equipment Maintenance" "ANPR for car parks" "Arboricultural Services" "Arboricultural Services" ...

$ Service.responsible : chr "Environment" "Environment" "Planning & Sustainability" "Planning & Sustainability" ...

$ Description.of.goods.and.services: chr "Air quality Equipment Maintenance" "ANPR for car parks" "Arboricultural Services" "Arboricultural Services" ...

$ Contract.Start.Date : chr "1/1/2015" "2/26/2013" "11/1/2014" "11/1/2014" ...

$ End.Date : chr "12/31/2015" "2/25/2018" "10/31/2018" "10/31/2018" ...

$ Review.Date : chr "11/30/2015" "10/1/2017" "10/31/2017" "10/31/2017" ...

$ Extension.Period : int 0 2 0 0 0 0 0 0 0 0 ...

$ Contract.Value : num 5944 1200000 400000 400000 400000 ...

$ SupplierName : chr "ESU1 ltd" "APT Control Ltd - Veri-Park" "Advance Tree Services Ltd" "Glendale Countryside Ltd" ...

$ NominatedContactPoint : chr "Julian.Smith@wycombe.gov.uk" "Robin.Evans@wycombe.gov.uk" "Alastair.Cunningham@wycombe.gov.uk" "Alastair.Cunningham@wycombe.gov.uk" ...

$ X : logi NA NA NA NA NA NA ...

$ X.1 : logi NA NA NA NA NA NA ...

$ X.2 : logi NA NA NA NA NA NA ...

$ X.3 : logi NA NA NA NA NA NA ...

$ X.4 : logi NA NA NA NA NA NA ...

ls(mydata)

> ls(mydata)

[1] "Contract.Reference.Number" "Contract.Start.Date"

[3] "Contract.Value" "Description.of.goods.and.services"

[5] "Effective.Date" "End.Date"

[7] "Extension.Period" "NominatedContactPoint"

[9] "Review.Date" " Service.responsible"

[11] "SupplierName" "Title.of.agreement"

[13] "X" "X.1"

[15] "X.2" "X.3"

[17] "X.4"

1. Keep only relevant variables

# DROP VARIABLES 13, 14,15, 16,17

newdata1 <- mydata[c( -13, -14,-15, -16,-17)]

head(newdata1, n=10)

ls(newdata1)

newdata1$Effective.Date<-NULL

newdata1$Description.of.goods.and.services<-NULL

newdata1$Extension.Period<-NULL

newdata1$NominatedContactPoint<-NULL

newdata1$Review.Date<-NULL

newdata1$SupplierName<-NULL

newdata1$Title.of.agreement<-NULL

dim(newdata1)

> dim(newdata1)

[1] 194 5

str(newdata1)

|  |
| --- |
| > str(newdata1)  'data.frame': 194 obs. of 5 variables:  $ Contract.Reference.Number: chr "342" "228" "289" "290" ...  $ Service.responsible : chr "Environment" "Environment" "Planning & Sustainability" "Planning & Sustainability" ...  $ Contract.Start.Date : chr "1/1/2015" "2/26/2013" "11/1/2014" "11/1/2014" ...  $ End.Date : chr "12/31/2015" "2/25/2018" "10/31/2018" "10/31/2018" ...  $ Contract.Value : num 5944 1200000 400000 400000 400000 ... |
|  |
| |  | | --- | |  | |

1. Create required Fraud indicator variables (y)

CREATE Y VARIABLES

TAT = END DATE - START DATE

PERDAY = CONTACT AMOUNT / TAT

# CONVERT DATE VARIABLES INTO DATE FORMAT

newdata1$Contract.Start.Date1<- as.Date(newdata1$Contract.Start.Date, "%m/%d/%Y")

newdata1$End.Date1<- as.Date(newdata1$End.Date,"%m/%d/%Y")

str(newdata1)

#DROP REDUNDANT VARIABLES

newdata1$Contract.Start.Date<-NULL

newdata1$End.Date<-NULL

str(newdata1)

> str(newdata1)

'data.frame': 165 obs. of 5 variables:

$ Contract.Reference.Number: chr "342" "228" "289" "290" ...

$ Service.responsible : chr "Environment" "Environment" "Planning & Sustainability" "Planning & Sustainability" ...

$ Contract.Value : num 5944 1200000 400000 400000 400000 ...

$ Contract.Start.Date1 : Date, format: "2015-01-01" "2013-02-26" "2014-11-01" ...

$ End.Date1 : Date, format: "2015-12-31" "2018-02-25" "2018-10-31" ...

# CREATE NEW VARIABLES

newdata1$tat<- newdata1$End.Date1-newdata1$Contract.Start.Date1

newdata1$tat1<- as.numeric(newdata1$tat)

newdata1$perday<- newdata1$Contract.Value/newdata1$tat1

str(newdata1)

> str(newdata1)

'data.frame': 194 obs. of 8 variables:

$ Contract.Reference.Number: chr "342" "228" "289" "290" ...

$ Service.responsible : chr "Environment" "Environment" "Planning & Sustainability" "Planning & Sustainability" ...

$ Contract.Value : num 5944 1200000 400000 400000 400000 ...

$ Contract.Start.Date1 : Date, format: "2015-01-01" "2013-02-26" "2014-11-01" ...

$ End.Date1 : Date, format: "2015-12-31" "2018-02-25" "2018-10-31" ...

$ tat :Class 'difftime' atomic [1:194] 364 1825 1460 1460 1460 ...

.. ..- attr(\*, "units")= chr "days"

$ tat1 : num 364 1825 1460 1460 1460 ...

$ perday : num 16.3 657.5 274 274 274 ...

1. C and O :-

# C - NO ACTIVITY REQUIRED

# O - MISSING VALUES

# CHECK MISSING VALUES – in R the missing values are imported as NA values

sapply(newdata1, function(x) sum(is.na(x)))

> sapply(newdata1, function(x) sum(is.na(x)))

Contract.Reference.Number Service.responsible Contract.Value

0 0 0

Contract.Start.Date1 End.Date1 tat

0 110 110

tat1 perday

110 110

# DROP OBS WHERE END DATE IS MISSING ; KEEP OBS WHERE DATA IS COMPLETE

newdata2<- newdata1[complete.cases(newdata1[,5:8]),]

# CHECK FREQUENCY TO UNDERSTAND SEGMENTS IN THE Service.responsible VARIABLE# INSTALL PACKAGE

install.packages('plyr')

library('plyr')

table1<- count(newdata2, 'Service.responsible')

table1

> table1

Service.responsible freq

1 Community 3

2 Democratic Legal & Policy 2

3 Environment 11

4 Finance & Commercial 31

5 Human Resources ICT/CSC & Shared Support Services 25

6 Planning & Sustainability 11

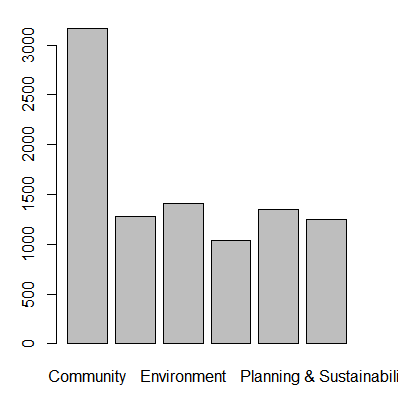
Note :- R is not case sensitive for obs within a variables (unlike SAS )

6. Visualisation :-

a. Bar Chart :- To understand Average value of TAT across different segments of Service\_responsible

# To understand Average value of TAT across different segments of Service\_responsible

barplot(with(newdata2, tapply(tat, Service.responsible, mean) ))



Note :- You should also explore the barplot functions in the package ggplot2

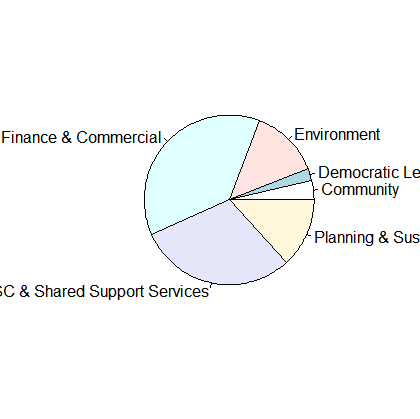
#CREATE PIE CHART TO UNDERSTAND CONTRIBUTION OF EACH SEGMENT

OF Service\_responsible IN THE DATA

library(MASS)

serv.freq = table(newdata2$Service.responsible)

pie(serv.freq)

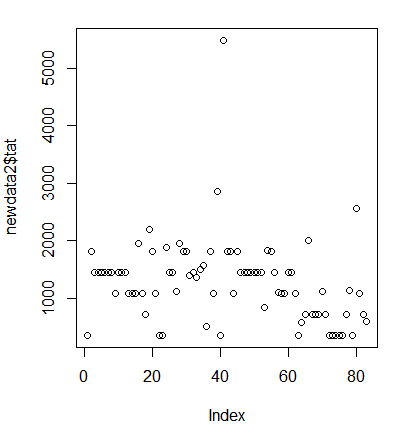


#DRAW A SCATTE PLOT TO SEE IF SOME CASES HAVE SIGNIFICANTLY HIGHER TAT

OR HIGHER RATE PER DAY VALUES

Note – I am doing it for TAT . You should do it for Per Day Rate as your exercise

plot(newdata2$tat)



#scatter plot1

newdata3 <- subset(newdata2,newdata2$Service.responsible=='Community')

plot(newdata3$tat)

#scatter plot2

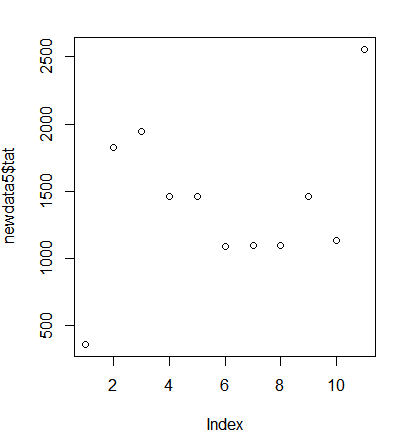
newdata4 <- subset(newdata2,newdata2$Service.responsible=='Democratic Legal & Policy')

plot(newdata4$tat)

#scatter plot3

newdata5 <- subset(newdata2,newdata2$Service.responsible=='Environment')

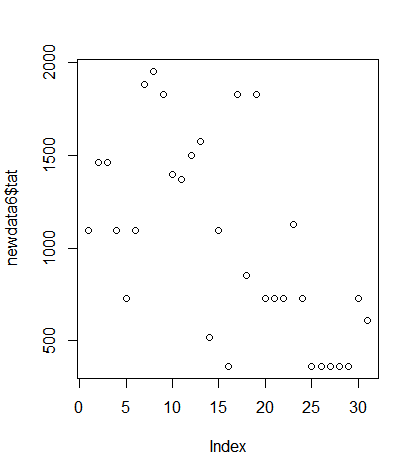
plot(newdata5$tat)



#scatter plot4

newdata6 <- subset(newdata2,newdata2$Service.responsible=='Finance & Commercial')

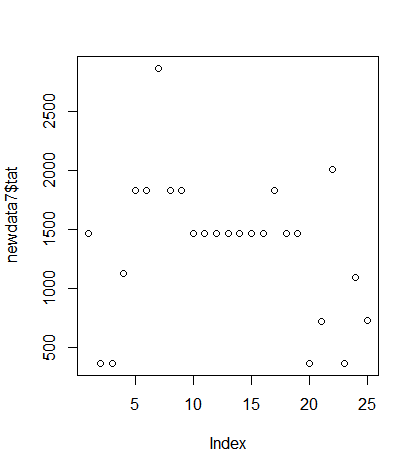
plot(newdata6$tat)



#scatter plot5

newdata7 <- subset(newdata2,newdata2$Service.responsible=='Human Resources ICT/CSC & Shared Support Services')

plot(newdata7$tat)



#scatter plot6

newdata8 <- subset(newdata2,newdata2$Service.responsible=='Planning & Sustainability')

plot(newdata8$tat)

