DA Lab 1

Name: Dheeraj Chaudhary
Roll: 17BCS009

```
library(tabulizer)
library(dplyr)
library(ggplot2)
library(reshape2)

########### Reading the pdf file and Extracting the data from 2000 to 2016##################
#original_data <- "/home/dheeraj/Desktop/Lecture/6th_sem_Academics/DataScience/CRS_2016.pdf"
#read_pdf <- extract_areas(original_data,pages = 11,output = "data.frame",header = F)
#read_pdf #read_pdf <- as.data.frame(read_pdf) #colnames(read_pdf) <-
c("year","event_Reg_LBirth","event_Reg_SBirth","event_Reg_Deaths","CRS_Births","CRS_deaths",
"percent_ofCRS_SRS_Births",
       # "percent_ofCRS_SRS_Deaths") #read_pdf #m <- tail(read_pdf, -11) #write.csv(m,
"/home/dheeraj/Desktop/Lecture/6th_sem_Academics/DataScience/new_data.csv") #writing the
extracted data into an csv file
nw_read <- read.csv("/home/dheeraj/Desktop/Lecture/6th_sem_Academics/DataScience/
new_data.csv") #reading the csv file
nw_read nw_read$X <- NULL nw_read
```

	A	B	C	D	E	F	G	H	I
1		year	event_Reg_LBirth	event_Reg_SBirth	event_Reg_Deaths	CRS_Births	CRS_deaths	percent_ofCRS_SRS_Births	percent_ofCRS_SRS_Deaths
2	12	2011	1108562	6940	384745	18.72	6.5	99.47	91.55
3	13	2012	1124490	6524	407015	18.62	6.74	100	94.93
4	14	2013	1068671	5708	413635	17.54	6.79	95.85	97
5	15	2014	1087530	5685	411533	17.21	6.51	94.04	93
6	16	2015	1053248	5067	393731	16.44	6.15	92	95
7	17	2016	1107258	4477	420774	16.42	6.28	92	95
8									

```
#################### Basic statistics for column 2 i.e, event_Reg_LBirth #################
min(nw_read[["event_Reg_SBirth"]]) max(nw_read[["event_Reg_SBirth"]])
mean(nw_read[["event_Reg_SBirth"]]) median(nw_read[["event_Reg_SBirth"]])
mode(nw_read[["event_Reg_SBirth"]]) var(nw_read[["event_Reg_SBirth"]])
sd(nw_read[["event_Reg_SBirth"]]) IQR(nw_read[["event_Reg_SBirth"]])
```

```
> min(nw_read[["event_Reg_LBirth"]])
[1] 1053248
> max(nw_read[["event_Reg_LBirth"]])
[1] 1124490
> mean(nw_read[["event_Reg_LBirth"]])
[1] 1091626
```

```
> median(nw_read[["event_Reg_LBirth"]])
[1] 1097394
> mode(nw_read[["event_Reg_LBirth"]])
[1] "numeric"
> var(nw_read[["event_Reg_LBirth"]])
[1] 725562028
> sd(nw_read[["event_Reg_LBirth"]])
[1] 26936.26
> IQR(nw_read[["event_Reg_LBirth"]])
[1] 34850.25
```

```
##################### Basic statistics for column 3 i.e, event_Reg_SBirth #################

min(nw_read[["event_Reg_SBirth"]])
max(nw_read[["event_Reg_SBirth"]])
mean(nw_read[["event_Reg_SBirth"]])
median(nw_read[["event_Reg_SBirth"]])
mode(nw_read[["event_Reg_SBirth"]])
var(nw_read[["event_Reg_SBirth"]])
sd(nw_read[["event_Reg_SBirth"]])
IQR(nw_read[["event_Reg_SBirth"]])
```

```
> min(nw_read[["event_Reg_SBirth"]])
[1] 4477
> max(nw_read[["event_Reg_SBirth"]])
[1] 6940
> mean(nw_read[["event_Reg_SBirth"]])
[1] 5733.5
> median(nw_read[["event_Reg_SBirth"]])
[1] 5696.5
> mode(nw_read[["event_Reg_SBirth"]])
[1] "numeric"
> var(nw_read[["event_Reg_SBirth"]])
[1] 821309.9
> sd(nw_read[["event_Reg_SBirth"]])
[1] 906.2615
> IQR(nw_read[["event_Reg_SBirth"]])
[1] 1098.5
```

```
##################### Basic statistics for column 4 i.e, event_Reg_Deaths #################

min(nw_read[["event_Reg_Deaths"]])
max(nw_read[["event_Reg_Deaths"]])
mean(nw_read[["event_Reg_Deaths"]])
median(nw_read[["event_Reg_Deaths"]])
mode(nw_read[["event_Reg_Deaths"]])
var(nw_read[["event_Reg_Deaths"]])
sd(nw_read[["event_Reg_Deaths"]])
IQR(nw_read[["event_Reg_Deaths"]])
```

```
> min(nw_read[["event_Reg_Deaths"]])
```

```
[1] 384745
> max(nw_read[["event_Reg_Deaths"]])
[1] 420774
> mean(nw_read[["event_Reg_Deaths"]])
[1] 405238.8
> median(nw_read[["event_Reg_Deaths"]])
[1] 409274
> mode(nw_read[["event_Reg_Deaths"]])
[1] "numeric"
> var(nw_read[["event_Reg_Deaths"]])
[1] 181407151
> sd(nw_read[["event_Reg_Deaths"]])
[1] 13468.75
> IQR(nw_read[["event_Reg_Deaths"]])
[1] 16057.5
```

```
##################### Basic statistics for column 5 i.e, CRS_Births #################

min(nw_read[["CRS_Births"]])
max(nw_read[["CRS_Births"]])
mean(nw_read[["CRS_Births"]])
median(nw_read[["CRS_Births"]])

mode(nw_read[["CRS_Births"]])
var(nw_read[["CRS_Births"]])
sd(nw_read[["CRS_Births"]])
IQR(nw_read[["CRS_Births"]])
```

```
> min(nw_read[["CRS_Births"]])
[1] 16.42
> max(nw_read[["CRS_Births"]])
[1] 18.72
> mean(nw_read[["CRS_Births"]])
[1] 17.49167
> median(nw_read[["CRS_Births"]])
[1] 17.375
> mode(nw_read[["CRS_Births"]])
[1] "numeric"
> var(nw_read[["CRS_Births"]])
[1] 1.023617
> sd(nw_read[["CRS_Births"]])
[1] 1.011739
> IQR(nw_read[["CRS_Births"]])
[1] 1.7175
```

```
##################### Basic statistics for column 6 i.e, CRS_deaths #################

min(nw_read[["CRS_deaths"]])
max(nw_read[["CRS_deaths"]])
mean(nw_read[["CRS_deaths"]])
```

```
median(nw_read[["CRS_deaths"]])
mode(nw_read[["CRS_deaths"]])
var(nw_read[["CRS_deaths"]])
sd(nw_read[["CRS_deaths"]])
IQR(nw_read[["CRS_deaths"]])
```

```
> min(nw_read[["CRS_deaths"]])
[1] 6.15
> max(nw_read[["CRS_deaths"]])
[1] 6.79
> mean(nw_read[["CRS_deaths"]])
[1] 6.495
> median(nw_read[["CRS_deaths"]])
[1] 6.505
> mode(nw_read[["CRS_deaths"]])
[1] "numeric"
> var(nw_read[["CRS_deaths"]])
[1] 0.06251
> sd(nw_read[["CRS_deaths"]])
[1] 0.25002
> IQR(nw_read[["CRS_deaths"]])
[1] 0.3475
```

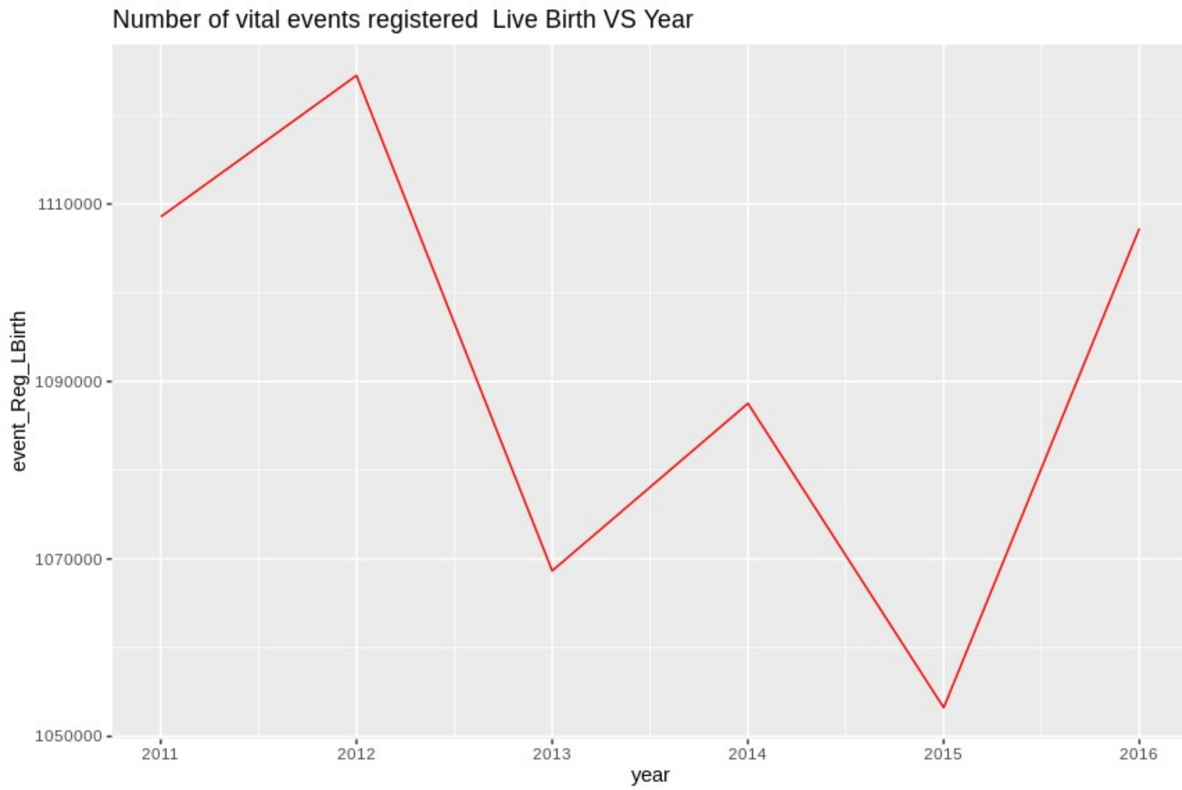
```
############ Basic statistics for column 7 i.e, percent_ofCRS_SRS_Births ##########

min(nw_read[["percent_ofCRS_SRS_Births"]])
max(nw_read[["percent_ofCRS_SRS_Births"]])
mean(nw_read[["percent_ofCRS_SRS_Births"]])
median(nw_read[["percent_ofCRS_SRS_Births"]])
mode(nw_read[["percent_ofCRS_SRS_Births"]])
var(nw_read[["percent_ofCRS_SRS_Births"]])
sd(nw_read[["percent_ofCRS_SRS_Births"]])
IQR(nw_read[["percent_ofCRS_SRS_Births"]])
```

```
> min(nw_read[["percent_ofCRS_SRS_Births"]])
[1] 92
> max(nw_read[["percent_ofCRS_SRS_Births"]])
[1] 100
> mean(nw_read[["percent_ofCRS_SRS_Births"]])
[1] 95.56
> median(nw_read[["percent_ofCRS_SRS_Births"]])
[1] 94.945
> mode(nw_read[["percent_ofCRS_SRS_Births"]])
[1] "numeric"
> var(nw_read[["percent_ofCRS_SRS_Births"]])
[1] 12.54868
> sd(nw_read[["percent_ofCRS_SRS_Births"]])
[1] 3.542412
> IQR(nw_read[["percent_ofCRS_SRS_Births"]])
[1] 6.055
```

```
########### Basic statistics for column 8 i.e, percent_ofCRS_SRS_Deaths ##########

min(nw_read[["percent_ofCRS_SRS_Deaths"]])
max(nw_read[["percent_ofCRS_SRS_Deaths"]])
mean(nw_read[["percent_ofCRS_SRS_Deaths"]])
median(nw_read[["percent_ofCRS_SRS_Deaths"]])
mode(nw_read[["percent_ofCRS_SRS_Deaths"]])
var(nw_read[["percent_ofCRS_SRS_Deaths"]])
sd(nw_read[["percent_ofCRS_SRS_Deaths"]])
IQR(nw_read[["percent_ofCRS_SRS_Deaths"]])
```

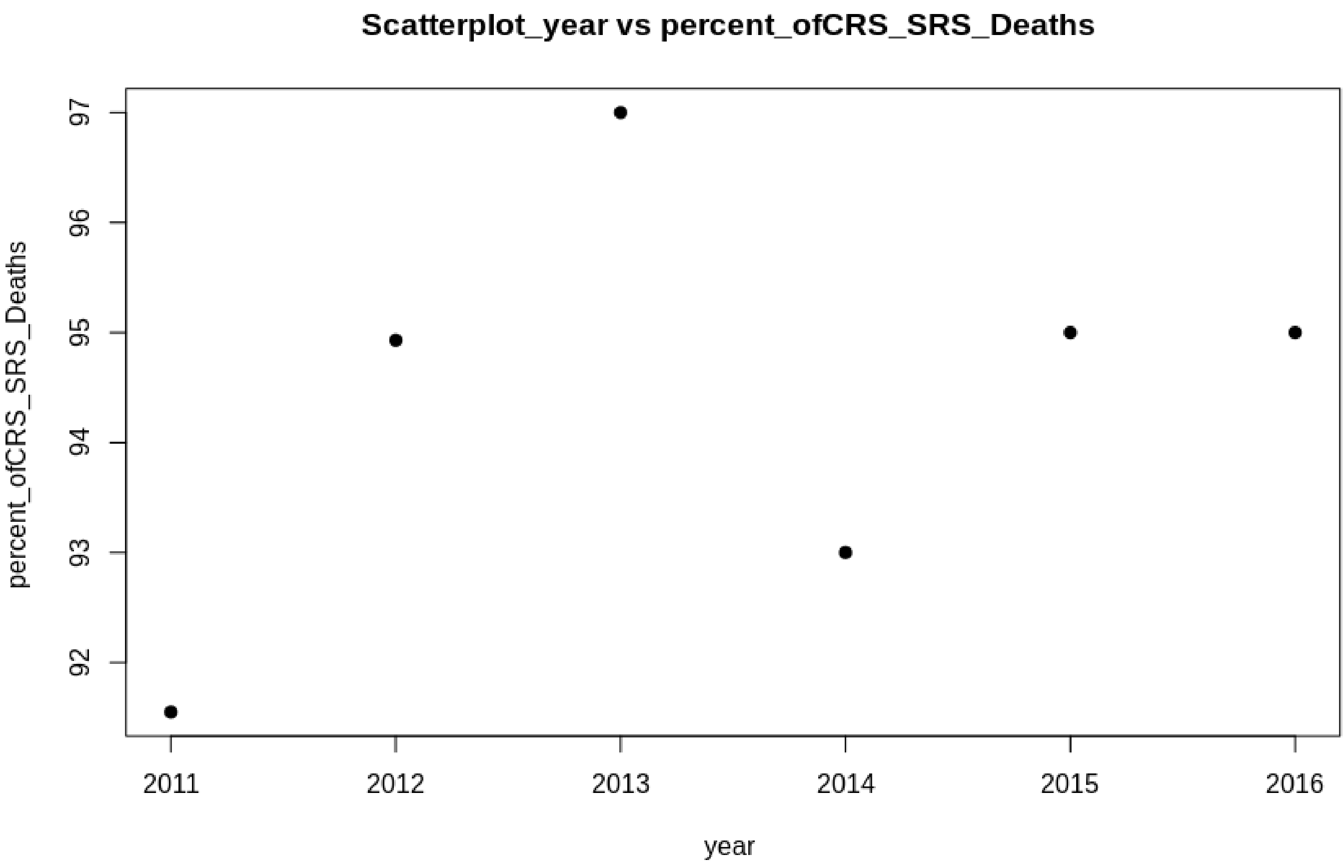
```
> min(nw_read[["percent_ofCRS_SRS_Deaths"]])
[1] 91.55
> max(nw_read[["percent_ofCRS_SRS_Deaths"]])
[1] 97
> mean(nw_read[["percent_ofCRS_SRS_Deaths"]])
[1] 94.41333
> median(nw_read[["percent_ofCRS_SRS_Deaths"]])
[1] 94.965
> mode(nw_read[["percent_ofCRS_SRS_Deaths"]])
[1] "numeric"
> var(nw_read[["percent_ofCRS_SRS_Deaths"]])
[1] 3.568467
> sd(nw_read[["percent_ofCRS_SRS_Deaths"]])
[1] 1.889039
> IQR(nw_read[["percent_ofCRS_SRS_Deaths"]])
[1] 1.5175
```

```
############### PLOT 1: Line plot for number of vital event registered for various years
ggplot(nw_read,aes(x = year,y = event_Reg_LBirth)) + geom_line(color = "red") +
ggtitle("Number of vital events registered Live Birth VS Year")

######### DESCRIPTION OF PLOT 1: It can be clearly seen from the plot that in year 2012
highest number of vital event registered
```




```
############### PLOT 2: Scatter plot for visualizing Death rates percentage of CRS vs SRS
attach(nw_read)


plot(year, percent_ofCRS_SRS_Deaths, main="Scatterplot_year vs percent_ofCRS_SRS_Deaths ",
xlab="year ", ylab="percent_of_CRS_SRS_Deaths ", pch=19) ############### DESCRIPTION OF PLOT
2: It can be clearly seen from the plot that in year 2013 highest percentage rate of CRS vs
SRS registered
```

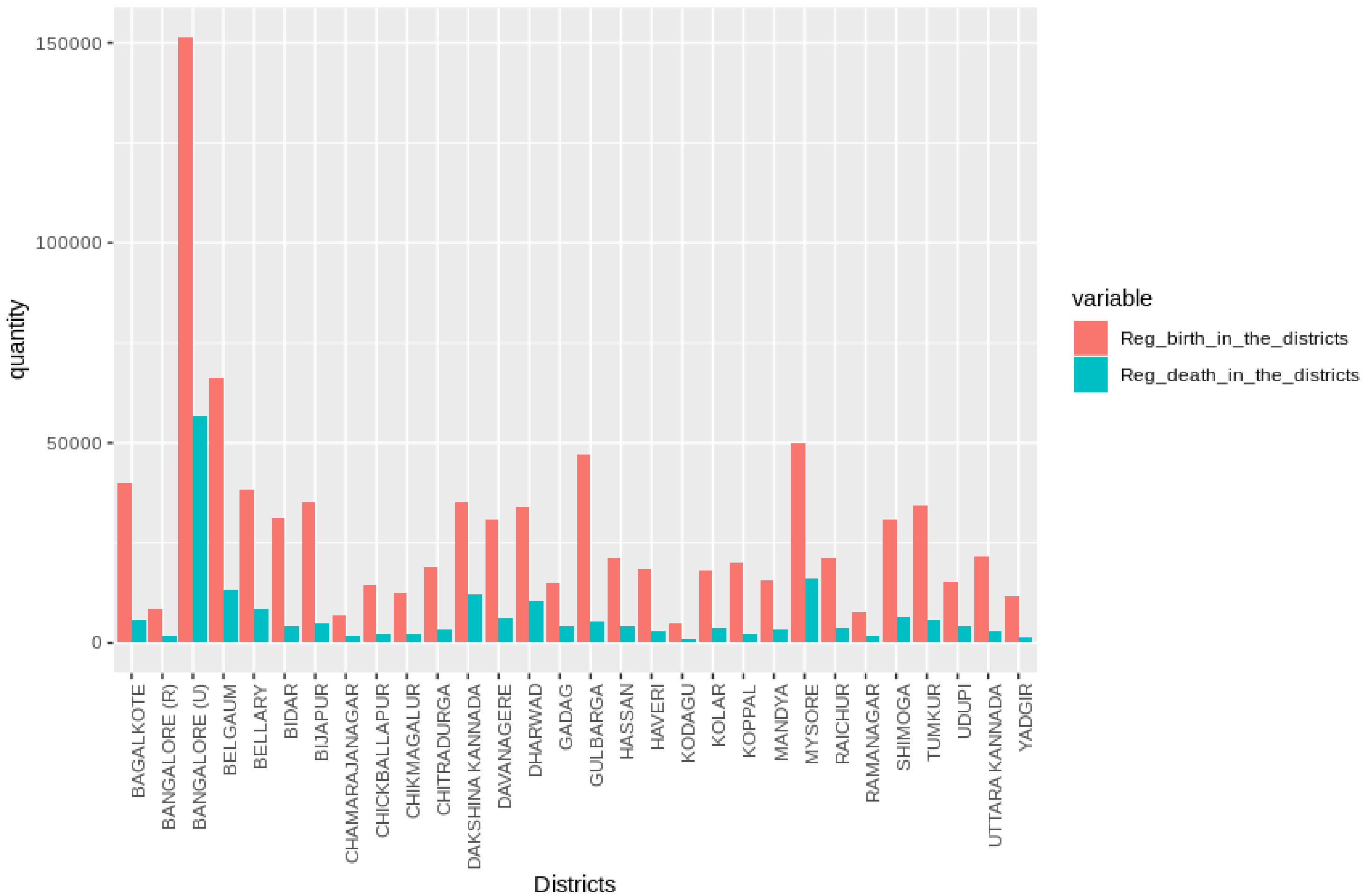
############# Reading the pdf file and Extracting districtwise aalysis for URBAN ##########




```
original_data <- "/home/dheeraj/Desktop/Lecture/6th_sem_Academics/DataScience/CRS_2016.pdf" #

District_Urban <- extract_areas(original_data,pages = 16,output = "data.frame",header = F)

District_Urban District_Urban <- as.data.frame(District_Urban) colnames(District_Urban) <- c("Districts","Reg_birth","Birth_rate","Reg_death","Death_rate","Reg_infant_death","Reg_still_birth","Still_birth_rate") District_Urban write.csv(District_Urban, "/home/dheeraj/Desktop/Lecture/6th_sem_Academics/DataScience/district_urbn.csv", row.names = FALSE) dist_urban_read <- read.csv("/home/dheeraj/Desktop/Lecture/6th_sem_Academics/DataScience/district_urbn.csv") dist_urban_read

############### PLOT 3: Bar plot for visualizing registered death and birth for urban districts

Urban_analysis <- data.frame(District_Urban$Districts,District_Urban$Reg_birth,District_Urban$Reg_death) colnames(Urban_analysis) <- c("Districts","Reg_birth_in_the_districts","Reg_death_in_the_districts") Urban_analysis <- melt(Urban_analysis,id.vars = "Districts") ggplot(Urban_analysis,aes(x = Districts , y =
```

```
value,fill = variable))+ylab("quantity") + geom_bar(stat = "identity",position = "dodge") +
theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

```
############### DESCRIPTION OF PLOT 3: By seeing the plot we can conclude that Banglore
distrist have highest number of registered birth and death, and all the district have higher
registered birth then death
```



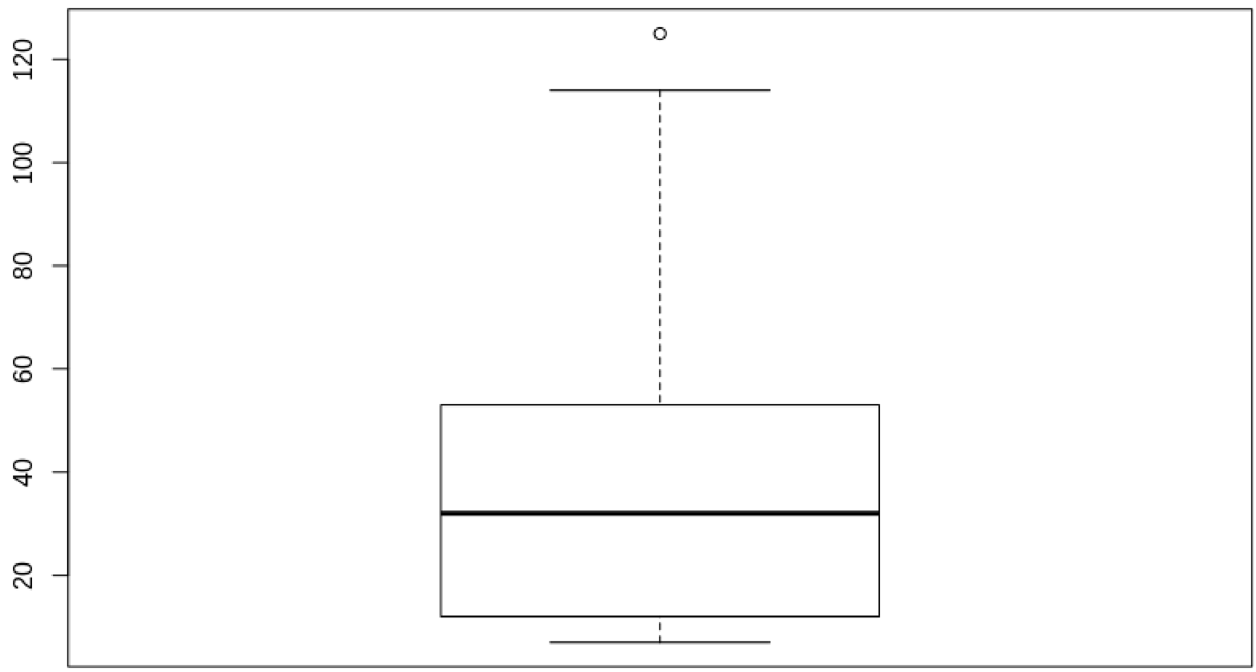

```
    ############### Reading the pdf file and Extracting districtwise aalysis for
    RURAL##################

original_data <- "/home/dheeraj/Desktop/Lecture/6th_sem_Academics/DataScience/CRS_2016.pdf"
District_Rural <- extract_areas(original_data,pages = 17,output = "data.frame",header = F)
District_Rural District_Rural <- as.data.frame(District_Rural) colnames(District_Rural) <-
c("Districts","Birth_reg","Birth_rate","Death_reg","Death_rate","Reg_infant_death","Still_bi
```

```
rth_reg","Still_birth_rate") District_Rural write.csv(District_Rural,
"/home/dheeraj/Desktop/Lecture/6th_sem_Academics/DataScience/district_rural.csv", row.names =
FALSE) dist_rural_read <-
read.csv("/home/dheeraj/Desktop/Lecture/6th_sem_Academics/DataScience/district_rural.csv")
dist_rural_read
```

############### PLOT 4: Box plot for visualizing registered infant death RuRAL District with OUTLIERS

```
boxplot(District_Rural$Reg_infant_death) num = as.numeric(District_Rural$Reg_infant_death)
outvalues = boxplot(num)$out which(District_Rural$Reg_infant_death %in% outvalues)
```

############### DESCRIPTION OF PLOT 4: By seeing the plot we got the median as 27 that means 75 percent of district have registered infant deaths above 27 and 25% have below 27, and a dot is shown which is outlier



############### PLOT 5: Box plot for visualizing registered infant death RuRAL District

```
without OUTLIERS removed = District_Rural$Reg_infant_death[!(District_Rural$Reg_infant_death
%in% outvalues)] boxplot(removed)
```

############### DESCRIPTION OF PLOT 5: we can see that the outlier is removed and we are grtting the desired output