Research project

470535297

Getting the Data files:

First of all we need to read the given three data files which is in CSV format. We want to bring the files in the R data frame in order to read and study the domestic violance happening in the NSW Local LGA. The NSW LGA file is having the Data of domestic violance monthly basis from january 99 to december 15. This code is given bellow:

```
# code for upload the initial data files
df_dv_nsw<- read.csv("C:/Users/golam/Desktop/assesment-data/data/DV_NSW_by_LGA.csv")
```

The NSW Lga file is having the data of various NSW LGA in 2011 data where we can see the data of various catogery is exists. the code is given bellow to read the file:

```
df_nsw_lga<- read.csv("C:/Users/golam/Desktop/assesment-data/data/NSW_LGA.csv")
```

Finally we have the data file For the labels where the description is given about the data pack, such as B1 is the total number of male. the code is given bellow:

```
df_lebel<- read.csv("C:/Users/golam/Desktop/assesment-data/data/labels.csv")
```

Read to analyse the data geting from the NSW Recorded Crime Statistics January to December 2017.

```
library(readxl)
otherdata<- read_excel("C:/Users/golam/Desktop/assesment-data/data/otherdata.xls",
                       sheet="Location")

#df[!is.na(df$B), ]
otherdata[!is.na(otherdata$'Incident Local Government Area'), ]
```

```
## # A tibble: 138 x 4
##    `Incident Local Government~ `Number of inciden~ `Rate per 100,00~ Rank
##    <chr>                                    <dbl> <chr>             <chr>
##  1 Albury                                     278. 532.9             33
##  2 Armidale Regional                          183. 603.7             19
##  3 Ballina                                    104. 244               92
##  4 Balranald                                    6. n.c               n.c
##  5 Bathurst Regional                          216. 509.6             38
##  6 Bayside                                    516. 313               80
##  7 Bega Valley                                107. 315.2             78
##  8 Bellingen                                   34. 263.7             90
##  9 Berrigan                                    13. 151.7             110
## 10 Blacktown                                 2022. 581.7             22
## # ... with 128 more rows
```

```
otherdata[!is.na(otherdata$'Number of incidents'), ]
```

```
## # A tibble: 132 x 4
##    `Incident Local Government~ `Number of inciden~ `Rate per 100,00~ Rank
##    <chr>                                  <dbl> <chr>             <chr>
##  1 Albury                                  278. 532.9             33
##  2 Armidale Regional                       183. 603.7             19
##  3 Ballina                                 104. 244               92
##  4 Balranald                                 6. n.c               n.c
##  5 Bathurst Regional                       216. 509.6             38
##  6 Bayside                                 516. 313               80
##  7 Bega Valley                             107. 315.2             78
##  8 Bellingen                                34. 263.7             90
##  9 Berrigan                                 13. 151.7             110
## 10 Blacktown                              2022. 581.7             22
## # ... with 122 more rows
```

Exploring the data files:

Attribute Information

Here is the attribute information for our three data file which is given from the censes data of NSW government data set:

DV_NSW_by_LGA:

The given DV_NSW_by_LGA file is containing the data of 140 local government areas data of domestic violance as a time series data starting from january 99 till december 2015. Variables are noted bellow:

1) LGA: Name of all the 140 LGA.
2) All other colums containing the number of incedent happen in every month.

NSW_LGA:

This file contain the data of all incedents including the area and catagorical data. Variables are noted bellow:

1) region_id: ID of the corresponding regional area.
2) label:This field contain the name of the Regeon.
3) year: year of the incedent happened.
4) area: size of the area.
5) rest of all other field contains the data of number of people
   based on the catogory of the labels.

label:

This file is mainly important for the description of the labels.

Cleaning the data

Next we have to clean this data. This data is actually almost already cleaned for, But here are some things need to consider doing for other data sets:

Check for NA values

Let's see if there any NA values:

```
# checking is there any NULL values in the initial data files
any(is.na(df_dv_nsw))
```

```
## [1] FALSE
```

```
any(is.na(df_lebel))
```

```
## [1] FALSE
```

```
any(is.na(df_nsw_lga))
```

```
## [1] FALSE
```

we have seen that all three files have no null values or the nill. However it is also important to see the catagorical feature of the data file before further analysis the data and also having a look the visualization of the data.

Now lets see the head of three data files. bellow is the time series data of the NSW lga files.

Bellow is the head of the label file.

```
head(df_lebel)
```

```
##   Sequential        Short                          Long DataPack.file
## 1         B1      Tot_P_M           Total_Persons_Males           B01
## 2         B2      Tot_P_F         Total_Persons_Females           B01
## 3         B3      Tot_P_P         Total_Persons_Persons           B01
## 4         B4 Age_0_4_yr_M   Age_groups_0_4_years_Males           B01
## 5         B5 Age_0_4_yr_F Age_groups_0_4_years_Females           B01
## 6         B6 Age_0_4_yr_P Age_groups_0_4_years_Persons           B01
##   Profile.table Column.heading.description.in.profile
## 1          B01a                                 Males
## 2          B01a                               Females
## 3          B01a                               Persons
## 4          B01a                                 Males
## 5          B01a                               Females
## 6          B01a                               Persons
```

Exploratory Data Analysis

In order to get a total picture of domestic violance happening in nsw local areas based on the data been supplied, we have to do certain type of operation within the data files and have to combine some field for the exploratory data analysis.

Since the data given in the NSW_LGA file is only consedering the data of 2011 so in the CSV file of DV_NSW ONLY 2011 data is added in a sepereate colum, aiming to analyse this data with the NSW_Lga file. the code for combining the fields is given bellow:

Process for combining the data files.

Since the data in the domestic violance file is consist as a monthly data, this has been converted to yearly data for the better iteration.

```
# Combining the data of 2011 in one variable for future iteration


df_dv_nsw$all_month_2011 <- df_dv_nsw$Jan.11 + df_dv_nsw$Feb.11 + df_dv_nsw$Mar.11 +
  df_dv_nsw$Apr.11 + df_dv_nsw$May.11 + df_dv_nsw$Jun.11 + df_dv_nsw$Jul.11 +
  df_dv_nsw$Aug.11+  df_dv_nsw$Sep.11 + df_dv_nsw$Oct.11 +
  df_dv_nsw$Nov.11 + df_dv_nsw$Dec.11



# Combining the data of 99 in one variable for future iteration

df_dv_nsw$all_month_99 <- df_dv_nsw$Jan.99 + df_dv_nsw$Feb.99 + df_dv_nsw$Mar.99 +
  df_dv_nsw$Apr.99 +  df_dv_nsw$May.99 + df_dv_nsw$Jun.99 + df_dv_nsw$Jul.99 +
  df_dv_nsw$Aug.99+  df_dv_nsw$Sep.99 + df_dv_nsw$Oct.99 +
  df_dv_nsw$Nov.99 + df_dv_nsw$Dec.99

# Combining the data of 00 in one variable for future iteration

df_dv_nsw$all_month_00 <- df_dv_nsw$Jan.00 + df_dv_nsw$Feb.00 + df_dv_nsw$Mar.00 +
  df_dv_nsw$Apr.00 +  df_dv_nsw$May.00 + df_dv_nsw$Jun.00 + df_dv_nsw$Jul.00 +
  df_dv_nsw$Aug.00+  df_dv_nsw$Sep.00 + df_dv_nsw$Oct.00 +
  df_dv_nsw$Nov.00 + df_dv_nsw$Dec.00

# Combining the data of 01 in one variable for future iteration

df_dv_nsw$all_month_01 <- df_dv_nsw$Jan.01 + df_dv_nsw$Feb.01 + df_dv_nsw$Mar.01 +
  df_dv_nsw$Apr.01 +  df_dv_nsw$May.01 + df_dv_nsw$Jun.01 + df_dv_nsw$Jul.01 +
  df_dv_nsw$Aug.01+  df_dv_nsw$Sep.01 + df_dv_nsw$Oct.01 +
  df_dv_nsw$Nov.01 + df_dv_nsw$Dec.01

# Combining the data of 02 in one variable for future iteration

df_dv_nsw$all_month_02 <- df_dv_nsw$Jan.02 + df_dv_nsw$Feb.02 + df_dv_nsw$Mar.02 +
  df_dv_nsw$Apr.02 +  df_dv_nsw$May.02 + df_dv_nsw$Jun.02 + df_dv_nsw$Jul.02 +
  df_dv_nsw$Aug.02+  df_dv_nsw$Sep.02 + df_dv_nsw$Oct.02 +
  df_dv_nsw$Nov.02 + df_dv_nsw$Dec.02
```

```r
# Combining the data of 03 in one variable for future iteration

df_dv_nsw$all_month_03 <- df_dv_nsw$Jan.03 + df_dv_nsw$Feb.03 + df_dv_nsw$Mar.03 +
  df_dv_nsw$Apr.03 +  df_dv_nsw$May.03 + df_dv_nsw$Jun.03 + df_dv_nsw$Jul.03 +
  df_dv_nsw$Aug.03+  df_dv_nsw$Sep.03 + df_dv_nsw$Oct.03 +
  df_dv_nsw$Nov.03 + df_dv_nsw$Dec.03

# Combining the data of 04 in one variable for future iteration

df_dv_nsw$all_month_04 <- df_dv_nsw$Jan.04 + df_dv_nsw$Feb.04 + df_dv_nsw$Mar.04 +
  df_dv_nsw$Apr.04 +  df_dv_nsw$May.04 + df_dv_nsw$Jun.04 + df_dv_nsw$Jul.04 +
  df_dv_nsw$Aug.04+  df_dv_nsw$Sep.04 + df_dv_nsw$Oct.04 +
  df_dv_nsw$Nov.04 + df_dv_nsw$Dec.04

# Combining the data of 05 in one variable for future iteration

df_dv_nsw$all_month_05 <- df_dv_nsw$Jan.05 + df_dv_nsw$Feb.05 + df_dv_nsw$Mar.05 +
  df_dv_nsw$Apr.05 +  df_dv_nsw$May.05 + df_dv_nsw$Jun.05 + df_dv_nsw$Jul.05 +
  df_dv_nsw$Aug.05+  df_dv_nsw$Sep.05 + df_dv_nsw$Oct.05 +
  df_dv_nsw$Nov.05 + df_dv_nsw$Dec.05

# Combining the data of 06 in one variable for future iteration

df_dv_nsw$all_month_06 <- df_dv_nsw$Jan.06 + df_dv_nsw$Feb.06 + df_dv_nsw$Mar.06 +
  df_dv_nsw$Apr.06 +  df_dv_nsw$May.06 + df_dv_nsw$Jun.06 + df_dv_nsw$Jul.06 +
  df_dv_nsw$Aug.06+  df_dv_nsw$Sep.06 + df_dv_nsw$Oct.06 +
  df_dv_nsw$Nov.06 + df_dv_nsw$Dec.06

# Combining the data of 07 in one variable for future iteration

df_dv_nsw$all_month_07 <- df_dv_nsw$Jan.07 + df_dv_nsw$Feb.07 + df_dv_nsw$Mar.07 +
  df_dv_nsw$Apr.07 +  df_dv_nsw$May.07 + df_dv_nsw$Jun.07 + df_dv_nsw$Jul.07 +
  df_dv_nsw$Aug.07+  df_dv_nsw$Sep.07 + df_dv_nsw$Oct.07 +
  df_dv_nsw$Nov.07 + df_dv_nsw$Dec.07

# Combining the data of 08 in one variable for future iteration

df_dv_nsw$all_month_08 <- df_dv_nsw$Jan.08 + df_dv_nsw$Feb.08 + df_dv_nsw$Mar.08 +
  df_dv_nsw$Apr.08 +  df_dv_nsw$May.08 + df_dv_nsw$Jun.08 + df_dv_nsw$Jul.08 +
  df_dv_nsw$Aug.08+  df_dv_nsw$Sep.08 + df_dv_nsw$Oct.08 +
  df_dv_nsw$Nov.08 + df_dv_nsw$Dec.08

# Combining the data of 09 in one variable for future iteration

df_dv_nsw$all_month_09 <- df_dv_nsw$Jan.09 + df_dv_nsw$Feb.09 + df_dv_nsw$Mar.09 +
  df_dv_nsw$Apr.09 +  df_dv_nsw$May.09 + df_dv_nsw$Jun.09 + df_dv_nsw$Jul.09 +
  df_dv_nsw$Aug.09+  df_dv_nsw$Sep.09 + df_dv_nsw$Oct.09 +
  df_dv_nsw$Nov.09 + df_dv_nsw$Dec.09

# Combining the data of 10 in one variable for future iteration

df_dv_nsw$all_month_10 <- df_dv_nsw$Jan.10 + df_dv_nsw$Feb.10 + df_dv_nsw$Mar.10 +
  df_dv_nsw$Apr.10 +  df_dv_nsw$May.10 + df_dv_nsw$Jun.10 + df_dv_nsw$Jul.10 +
```

```
  df_dv_nsw$Aug.10+  df_dv_nsw$Sep.10 + df_dv_nsw$Oct.10 +
  df_dv_nsw$Nov.10 + df_dv_nsw$Dec.10

# Combining the data of 12 in one variable for future iteration

df_dv_nsw$all_month_12 <- df_dv_nsw$Jan.12 + df_dv_nsw$Feb.12 + df_dv_nsw$Mar.12 +
  df_dv_nsw$Apr.12 +  df_dv_nsw$May.12 + df_dv_nsw$Jun.12 + df_dv_nsw$Jul.12 +
  df_dv_nsw$Aug.12 +  df_dv_nsw$Sep.12 + df_dv_nsw$Oct.12 +
  df_dv_nsw$Nov.12 + df_dv_nsw$Dec.12

# Combining the data of 13 in one variable for future iteration

df_dv_nsw$all_month_13 <- df_dv_nsw$Jan.13 + df_dv_nsw$Feb.13 + df_dv_nsw$Mar.13 +
  df_dv_nsw$Apr.13 +  df_dv_nsw$May.13 + df_dv_nsw$Jun.13 + df_dv_nsw$Jul.13 +
  df_dv_nsw$Aug.13 +  df_dv_nsw$Sep.13 + df_dv_nsw$Oct.13 +
  df_dv_nsw$Nov.13 + df_dv_nsw$Dec.13

# Combining the data of 14 in one variable for future iteration

df_dv_nsw$all_month_14 <- df_dv_nsw$Jan.14 + df_dv_nsw$Feb.14 + df_dv_nsw$Mar.14 +
  df_dv_nsw$Apr.14 +  df_dv_nsw$May.14 + df_dv_nsw$Jun.14 + df_dv_nsw$Jul.14 +
  df_dv_nsw$Aug.14 +  df_dv_nsw$Sep.14 + df_dv_nsw$Oct.14 +
  df_dv_nsw$Nov.14 + df_dv_nsw$Dec.14

# Combining the data of 15 in one variable for future iteration

df_dv_nsw$all_month_15 <- df_dv_nsw$Jan.15 + df_dv_nsw$Feb.15 + df_dv_nsw$Mar.15 +
  df_dv_nsw$Apr.15 +  df_dv_nsw$May.15 + df_dv_nsw$Jun.15 + df_dv_nsw$Jul.15 +
  df_dv_nsw$Aug.15 +  df_dv_nsw$Sep.15 + df_dv_nsw$Oct.15 +
  df_dv_nsw$Nov.15 + df_dv_nsw$Dec.15
```

So from the above code it is targeting that the 2011 data is been examined along with the corresponding Area codes.

Bellow is the code for selecting new colums.

```
#code to keep the two variables in the new data file df_dv_nsw1
df_dv_nsw1 <- subset(df_dv_nsw, select = c(1, 206, 207, 208, 209, 210, 211, 212,
                                           213, 214, 215, 216, 217, 218,
                                           219, 220, 221, 222))
```

Now new file is become the df_dv_nsw1, which is being usable in future with the combination of the Nsw_lga file. Bellow is the code for LGA wise domestic violance data only for the year of 2011. Nice thing is this data file only contain 2 variables.

Bellow is the head of data which is bacically the total summary of domestic violance year to year.

```
head(df_dv_nsw1)
```

```
##                 LGA all_month_2011 all_month_99 all_month_00 all_month_01
## 1            Albury            235          123          181          192
## 2 Armidale Dumaresq            145           78           86           92
## 3          Ashfield            111           68           80           91
```

```
## 4            Auburn            274          133          138          144
## 5           Ballina            149          109          128          155
## 6         Bankstown            766          339          438          567
##   all_month_02 all_month_03 all_month_04 all_month_05 all_month_06
## 1          236          247          207          222          271
## 2           93           84           93           73          102
## 3          103           94          136          108          116
## 4          155          167          140          133          186
## 5          181          200          155          143          174
## 6          492          488          457          492          621
##   all_month_07 all_month_08 all_month_09 all_month_10 all_month_12
## 1          272          258          239          216          253
## 2          134          130          113          123          158
## 3          110           91           81          101           82
## 4          197          244          277          297          350
## 5          183          112          147          127          150
## 6          638          548          643          674          731
##   all_month_13 all_month_14 all_month_15
## 1          277          304          268
## 2          165          167          162
## 3          108          107          115
## 4          366          350          304
## 5          136          164           93
## 6          893          891          914
```

Summerising the yearly data to see the graph representation.

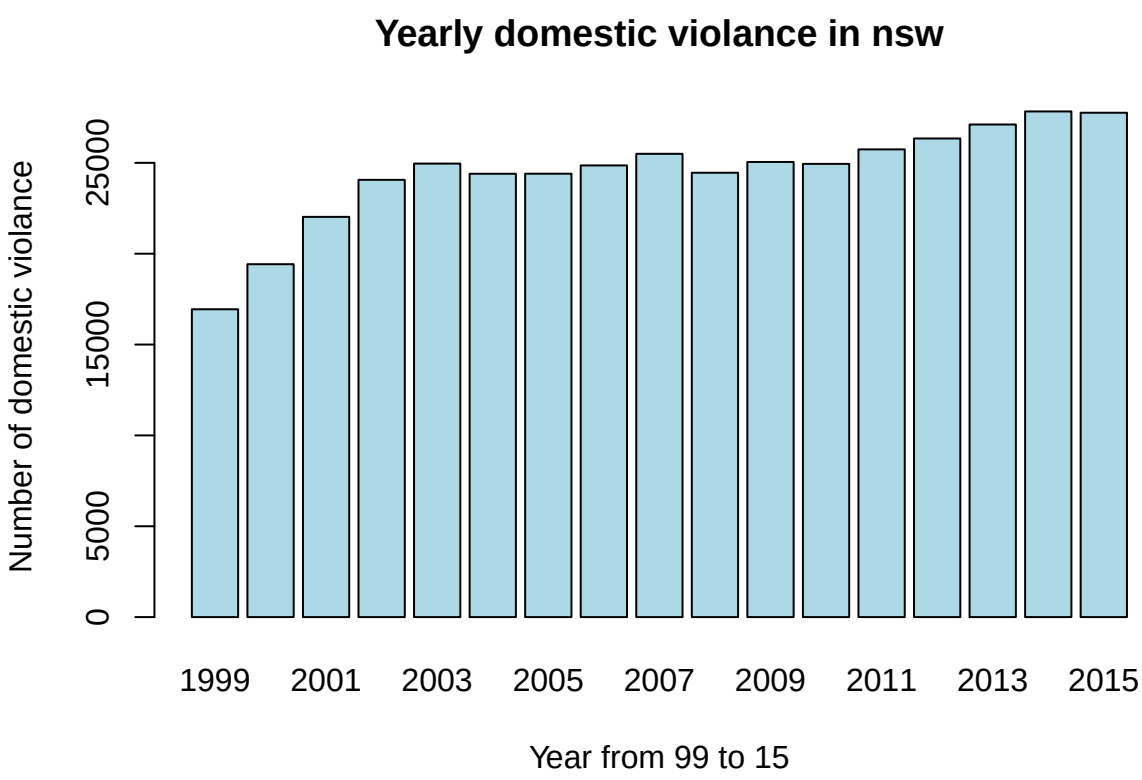
```
dv99<-sum(df_dv_nsw1$all_month_99)
dv00<-sum(df_dv_nsw1$all_month_00)
dv01<-sum(df_dv_nsw1$all_month_01)
dv02<-sum(df_dv_nsw1$all_month_02)
dv03<-sum(df_dv_nsw1$all_month_03)
dv04<-sum(df_dv_nsw1$all_month_04)
dv05<-sum(df_dv_nsw1$all_month_05)
dv06<-sum(df_dv_nsw1$all_month_06)
dv07<-sum(df_dv_nsw1$all_month_07)
dv08<-sum(df_dv_nsw1$all_month_08)
dv09<-sum(df_dv_nsw1$all_month_09)
dv10<-sum(df_dv_nsw1$all_month_10)
dv11<-sum(df_dv_nsw1$all_month_2011)
dv12<-sum(df_dv_nsw1$all_month_12)
dv13<-sum(df_dv_nsw1$all_month_13)
dv14<-sum(df_dv_nsw1$all_month_14)
dv15<-sum(df_dv_nsw1$all_month_15)

dv=c(dv99,dv00, dv01, dv02, dv03, dv04, dv05, dv06, dv07, dv08,
     dv09, dv10, dv11, dv12, dv13, dv14, dv15)
year=c(1999,2000,2001,2002,2003,2004,2005,2006,2007,
       2008,2009,2010,2011,2012,2013,2014,2015)

setNames(dv, year)
```

```
##  1999  2000  2001  2002  2003  2004  2005  2006  2007  2008  2009  2010
## 16941 19421 22027 24065 24959 24399 24403 24858 25498 24455 25048 24942
##  2011  2012  2013  2014  2015
## 25741 26341 27111 27830 27757
```

```
#as.table(setNames(dv, year))
data.frame(setNames(dv, year))
```

```
##      setNames.dv..year.
## 1999              16941
## 2000              19421
## 2001              22027
## 2002              24065
## 2003              24959
## 2004              24399
## 2005              24403
## 2006              24858
## 2007              25498
## 2008              24455
## 2009              25048
## 2010              24942
## 2011              25741
## 2012              26341
## 2013              27111
## 2014              27830
## 2015              27757
```

```
barplot(setNames(dv,year),xlab='Year from 99 to 15', ylab='Number of domestic violance',
        main='Yearly domestic violance in nsw', col='lightblue')
```



From the above graph it has been seen that the trand of domestic violance in nsw is going up which is very significant to mention, and is so alerming. however since from 1999 the popolation in NSW is increasing, however we have to see and examine the corelation with the increasing population density in NSW as well.

Bellow is the visualization of the data file which we called: other data.

```
#head(otherdata)
library(ggplot2)
names(otherdata)[names(otherdata) == "Incident Local Government Area"] <- "Area"
names(otherdata)[names(otherdata) == "Number of incidents"] <- "Incidents"
names(otherdata)[names(otherdata) == "Rate per 100,000"] <- "Rate"
#head(otherdata)

otherdata$Rank <- as.numeric(otherdata$Rank)
```

```
## Warning: NAs introduced by coercion
```

```
attach(otherdata)
newdata <- otherdata[ which(otherdata$Rank<11),]
#head(newdata)

p<-ggplot(data=newdata, aes(x=Area,y=Incidents)) +
   geom_bar(stat="identity", fill="steelblue")
```

```
p
```

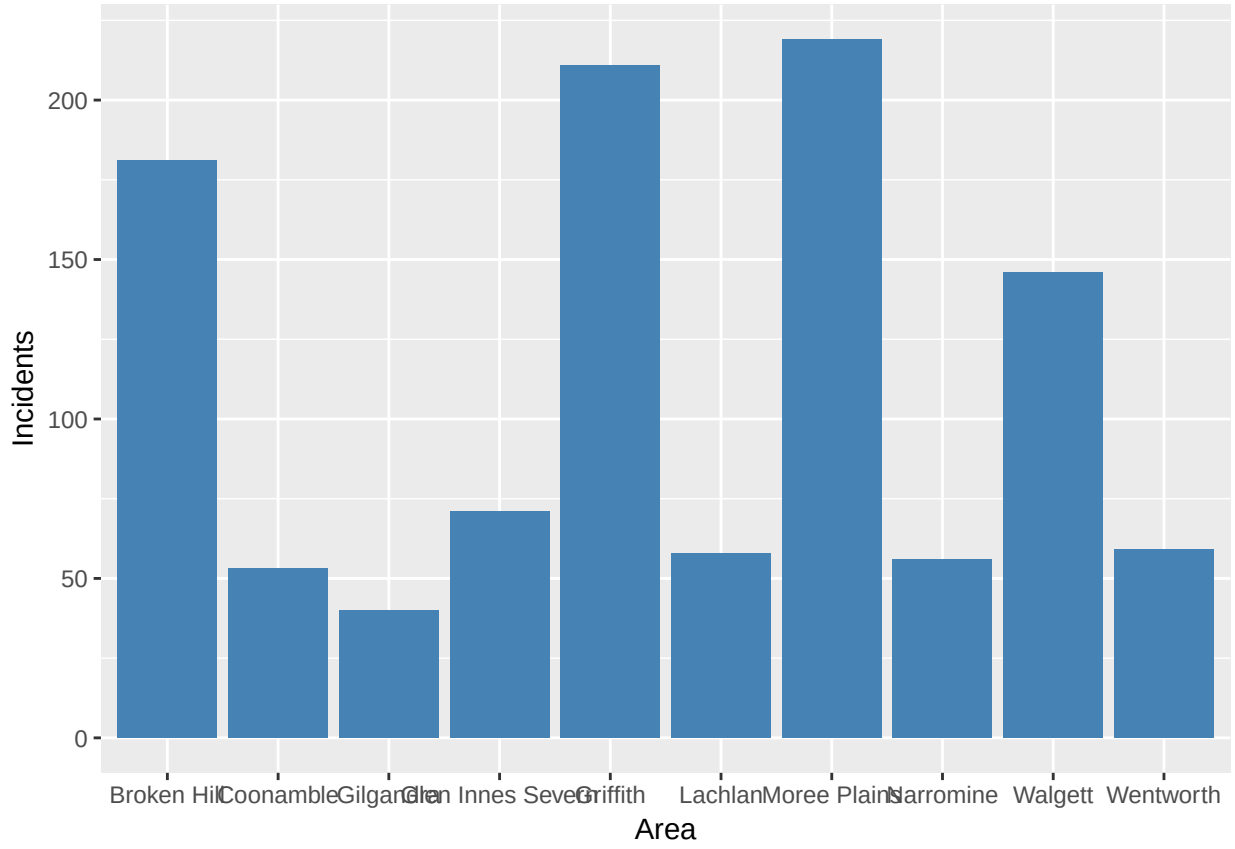
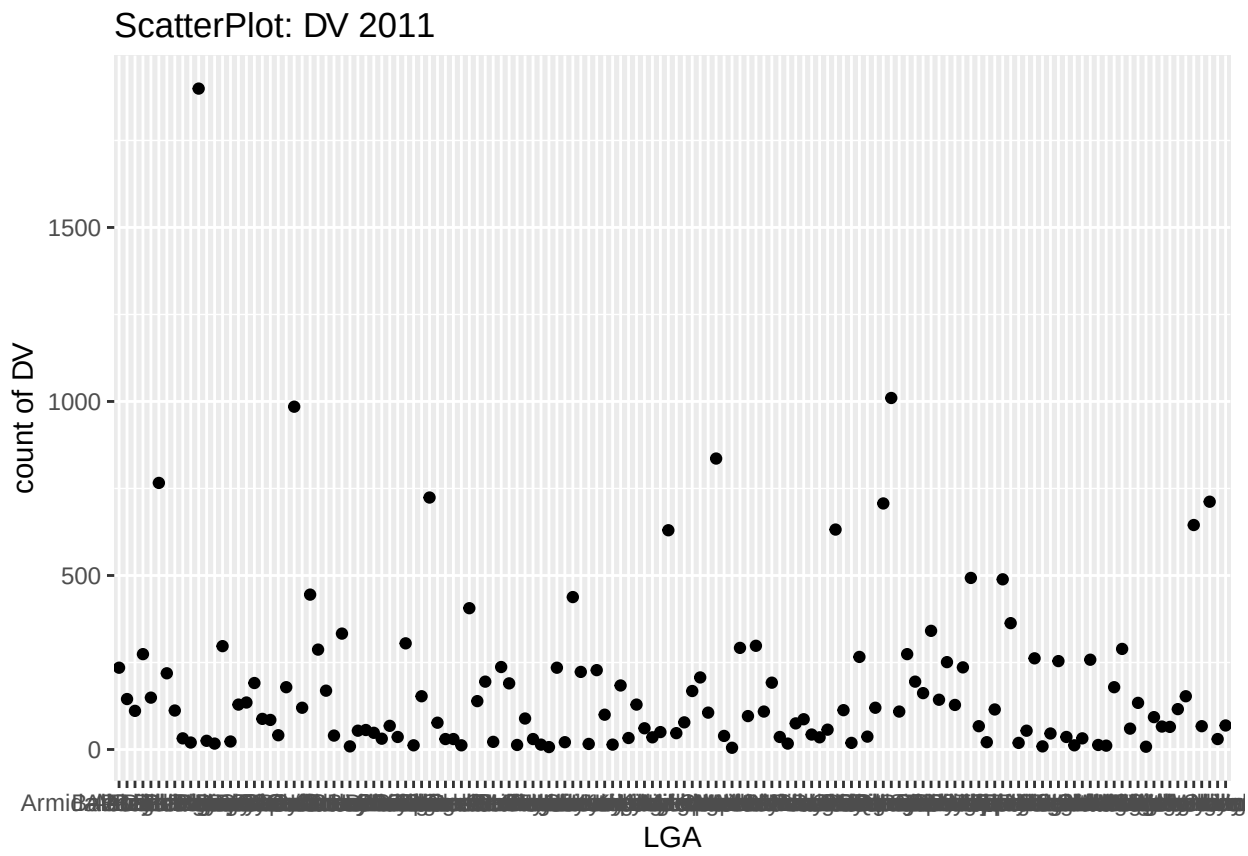


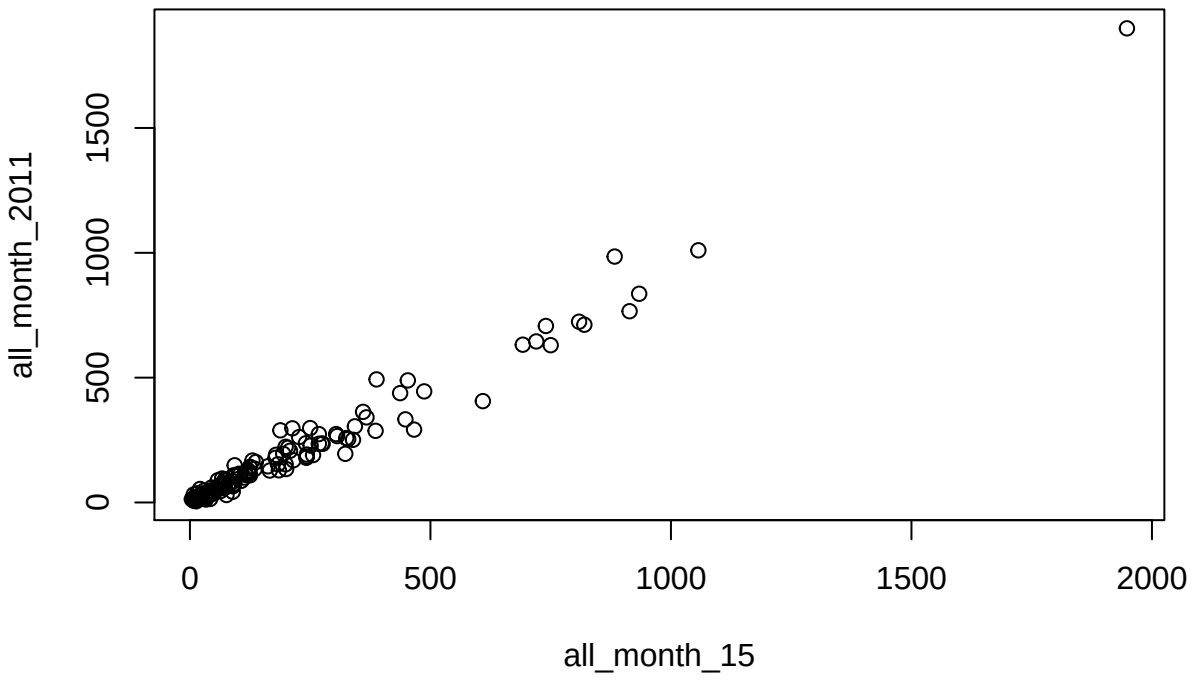
Fig is showing top ten domestic violance rating area in NSW from the data 'NSW Recorded Crime Statistics January to December 2017.'

Now lets visualise some of the data yearly basis rate of domestic violance area wise.

```
#Scatter plot of domestic violance in 2011

library(ggplot2)

plot_dv_2011 <- ggplot(data = df_dv_nsw1, mapping = aes(x = LGA , y = all_month_2011)) +
  geom_point() +
  labs(title="ScatterPlot: DV 2011",x="LGA",y="count of DV")
plot_dv_2011
```




```
plot(all_month_2011~all_month_15, data=df_dv_nsw1)
```



NSW_LGA containing total 7946 variables, which not all them are required to analyse the domestic violance in nsw. Considering the examining year bellow is the variables, been selected to analyse. Bellow is the nine variables been selected for further analysis:

- **region_id** ID of the region
- **label** Name of the area
- **year** 2011
- **area_sqkm** Area of that Region in squire kilometer
- **B1** Total_Persons_Males
- **B2** Total_Persons_Females
- **B3** Total_Persons_Persons
- **B15** Age_groups_20_24_years_Persons
- **B18** Age_groups_25_34_years_Persons

Bellow is the code for new selected data colums, where initially the variables been select is only 9. This new data file and dv_nsw1 file will be combined to get the final workable file to explain the domestic violance in nsw. This is noticable that the LGA variable field in this file has converted to the same shape as the DV_NSW, So that further merging is possible is possible in between the CSV files. The code and final structure of the data file is given bellow:

```
# keeping some selected variables which i have the interest, rather
#than keeping 7942 variables
df_nsw_lga1 <- subset(df_nsw_lga, select = c(1,2,3,4,5,6,7,19,22))
```

```
#setnames(df_nsw_lga1, "label", "LGA")
colnames(df_nsw_lga1)[which(names(df_nsw_lga1) == "label")] <- "LGA"
df_nsw_lga1$LGA <- gsub('.{4}$', '',df_nsw_lga1$LGA )
head(df_nsw_lga1)
```

```
##   region_id               LGA year   area_sqkm    B1    B2    B3  B15
## 1  LGA10050            Albury 2011   305.93100 23072 24738 47810 3445
## 2  LGA10110 Armidale Dumaresq 2011  4230.83000 11515 12590 24105 2487
## 3  LGA10150          Ashfield 2011     8.28119 20032 21182 41214 3190
## 4  LGA10200            Auburn 2011    32.47690 38225 35513 73738 6826
## 5  LGA10250           Ballina 2011   484.71600 18842 20432 39274 1597
## 6  LGA10300         Balranald 2011 21693.12000  1175  1108  2283  113
##     B18
## 1  5975
## 2  2614
## 3  7860
## 4 15120
## 5  3280
## 6   219
```

Bsed on the LGA varible, the data file df_dv_nsw1 and df_dv_lga1 is been combined to the final workable data file in the NSW LGA area, to analyse and explain the domestic violance situation. The domestic violance is been considered is based on the data of the year 2011. The code and final data file head is given bellow:

library sqldf has been used in here, to select the matcheced colums from both of the csv file df_nsw_lga1 and df_dv_nsw1, whicle the matches colum is LGA from both of the file. the result has been stored in the new csv file called data.csv, which has been used in the rest of the analysis of the domestic violance probel. since two colums remain again the same name we need to drom one colum name "LGA". the final gigure of the data file is mentioned bellow.

```
#Combine new two data file into one using the SQL library and statement, and put them in the final
#datafile which have to be obsetrve later.

df1<-df_nsw_lga1
df2<-df_dv_nsw1


library(sqldf )
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Loading required package: RSQLite
```

```
data<-sqldf("SELECT *
      FROM df1, df2
      WHERE df1.LGA==df2.LGA")

data <- subset(data, select = c(1,2,3,4,5,6,7,8,9,11))

head(data)
```

```
##   region_id               LGA year  area_sqkm    B1    B2     B3   B15
## 1  LGA10050            Albury 2011  305.93100 23072 24738  47810  3445
```

```
## 2  LGA10110 Armidale Dumaresq 2011 4230.83000 11515 12590  24105  2487
## 3  LGA10150          Ashfield 2011    8.28119 20032 21182  41214  3190
## 4  LGA10200            Auburn 2011   32.47690 38225 35513  73738  6826
## 5  LGA10250           Ballina 2011  484.71600 18842 20432  39274  1597
## 6  LGA10350         Bankstown 2011   76.80010 89928 92424 182352 12486
##     B18 all_month_2011
## 1  5975            235
## 2  2614            145
## 3  7860            111
## 4 15120            274
## 5  3280            149
## 6 24678            766
```

This new workable data file having the name of some variables, which need to change the name for better understand once observe the structure of the data file.

```
# changing the name of some variables and give the sensable name
names(data)[names(data) == "B1"] <- "maleTotal"
names(data)[names(data) == "B2"] <- "femaleTotal"
names(data)[names(data) == "B3"] <- "populationTotal"
names(data)[names(data) == "B15"] <- "Twenty_twentyfour_Total"
names(data)[names(data) == "B18"] <- "twentyfive_thirtyfive_Total"
```

```
head(data)
```

```
##   region_id              LGA year  area_sqkm maleTotal femaleTotal
## 1  LGA10050           Albury 2011  305.93100     23072       24738
## 2  LGA10110 Armidale Dumaresq 2011 4230.83000     11515       12590
## 3  LGA10150          Ashfield 2011    8.28119     20032       21182
## 4  LGA10200            Auburn 2011   32.47690     38225       35513
## 5  LGA10250           Ballina 2011  484.71600     18842       20432
## 6  LGA10350         Bankstown 2011   76.80010     89928       92424
##   populationTotal Twenty_twentyfour_Total twentyfive_thirtyfive_Total
## 1           47810                    3445                         5975
## 2           24105                    2487                         2614
## 3           41214                    3190                         7860
## 4           73738                    6826                        15120
## 5           39274                    1597                         3280
## 6          182352                   12486                        24678
##   all_month_2011
## 1            235
## 2            145
## 3            111
## 4            274
## 5            149
## 6            766
```

Further explanation from this point

```
# checking is there any NULL value in our new data file
any(is.na(data))
```

```
## [1] FALSE
```

```
# The summary of data of the new data file
summary(data)
```

```
##     region_id       LGA                 year        area_sqkm
##  LGA10050:  1   Length:140         Min.   :2011   Min.   :    5.72
##  LGA10110:  1   Class :character   1st Qu.:2011   1st Qu.:  230.39
##  LGA10150:  1   Mode  :character   Median :2011   Median : 2444.03
##  LGA10200:  1                      Mean   :2011   Mean   : 3985.67
##  LGA10250:  1                      3rd Qu.:2011   3rd Qu.: 4827.40
##  LGA10350:  1                      Max.   :2011   Max.   :45571.01
##  (Other) :134
##    maleTotal       femaleTotal     populationTotal
##  Min.   :  1458   Min.   :  1410   Min.   :  2868
##  1st Qu.:  4010   1st Qu.:  3993   1st Qu.:  7926
##  Median : 12528   Median : 13095   Median : 25909
##  Mean   : 23567   Mean   : 24372   Mean   : 47939
##  3rd Qu.: 31184   3rd Qu.: 32491   3rd Qu.: 63517
##  Max.   :149547   Max.   :151552   Max.   :301099
##
##  Twenty_twentyfour_Total twentyfive_thirtyfive_Total all_month_2011
##  Min.   :   90.0         Min.   :  223.0             Min.   :   5.0
##  1st Qu.:  355.2         1st Qu.:  728.5             1st Qu.:  36.0
##  Median : 1184.0         Median : 2408.5             Median : 109.0
##  Mean   : 3043.4         Mean   : 6294.2             Mean   : 183.9
##  3rd Qu.: 3688.0         3rd Qu.: 7727.2             3rd Qu.: 235.0
##  Max.   :20570.0         Max.   :46564.0             Max.   :1899.0
##
```

Finding the co-orelation in between the numaric colum in this table is given bellow.

Finding the co orelation between the data is one of the important task in statistical learning. bellow is the explanation of the corelation in the data field of this table:

Correlation plots are a great way of exploring data and seeing if there are any interaction terms.

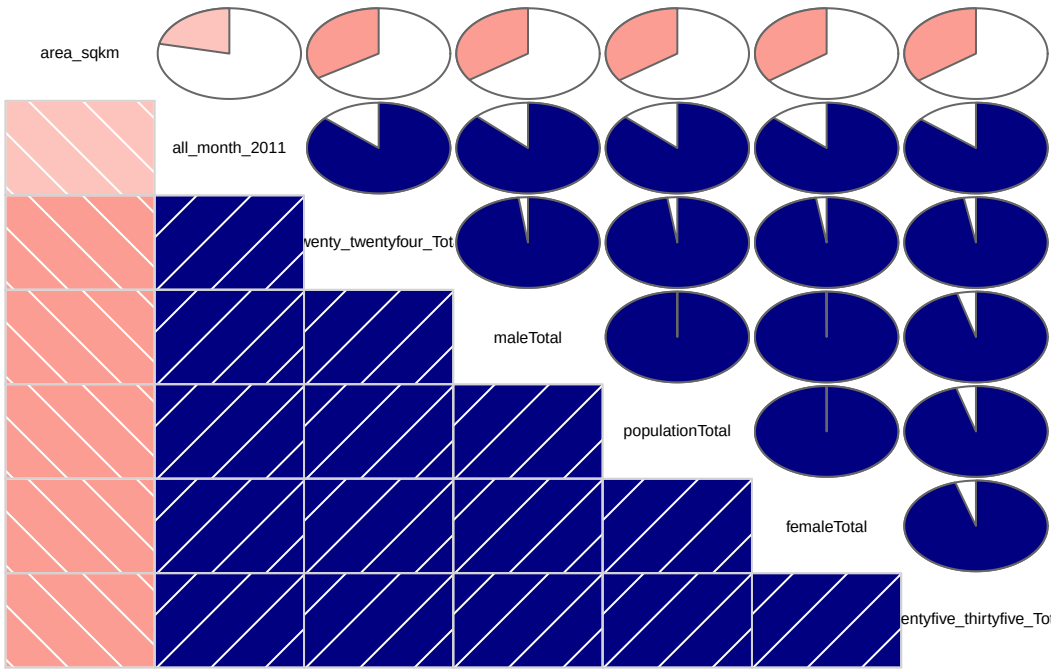
```
data$year <- as.character(data$year)
# Grab only numeric columns
num.cols <- sapply(data, is.numeric)

# Filter to numeric columns for correlation
cor.data <- cor(data[,num.cols])

cor.data
```

```
##                              area_sqkm  maleTotal femaleTotal
## area_sqkm                    1.0000000 -0.3482112  -0.3520409
## maleTotal                   -0.3482112  1.0000000   0.9994673
## femaleTotal                 -0.3520409  0.9994673   1.0000000
## populationTotal             -0.3502042  0.9998624   0.9998712
## Twenty_twentyfour_Total     -0.3404532  0.9789824   0.9761136
## twentyfive_thirtyfive_Total -0.3517381  0.9575261   0.9534249
## all_month_2011              -0.2151984  0.8736348   0.8673730
##                             populationTotal Twenty_twentyfour_Total
## area_sqkm                        -0.3502042              -0.3404532
## maleTotal                         0.9998624               0.9789824
## femaleTotal                       0.9998712               0.9761136
## populationTotal                   1.0000000               0.9776545
## Twenty_twentyfour_Total           0.9776545               1.0000000
## twentyfive_thirtyfive_Total       0.9555689               0.9728675
## all_month_2011                    0.8705682               0.8651148
##                             twentyfive_thirtyfive_Total all_month_2011
## area_sqkm                                   -0.3517381     -0.2151984
## maleTotal                                    0.9575261      0.8736348
## femaleTotal                                  0.9534249      0.8673730
## populationTotal                              0.9555689      0.8705682
## Twenty_twentyfour_Total                      0.9728675      0.8651148
## twentyfive_thirtyfive_Total                  1.0000000      0.8601218
## all_month_2011                               0.8601218      1.0000000
```

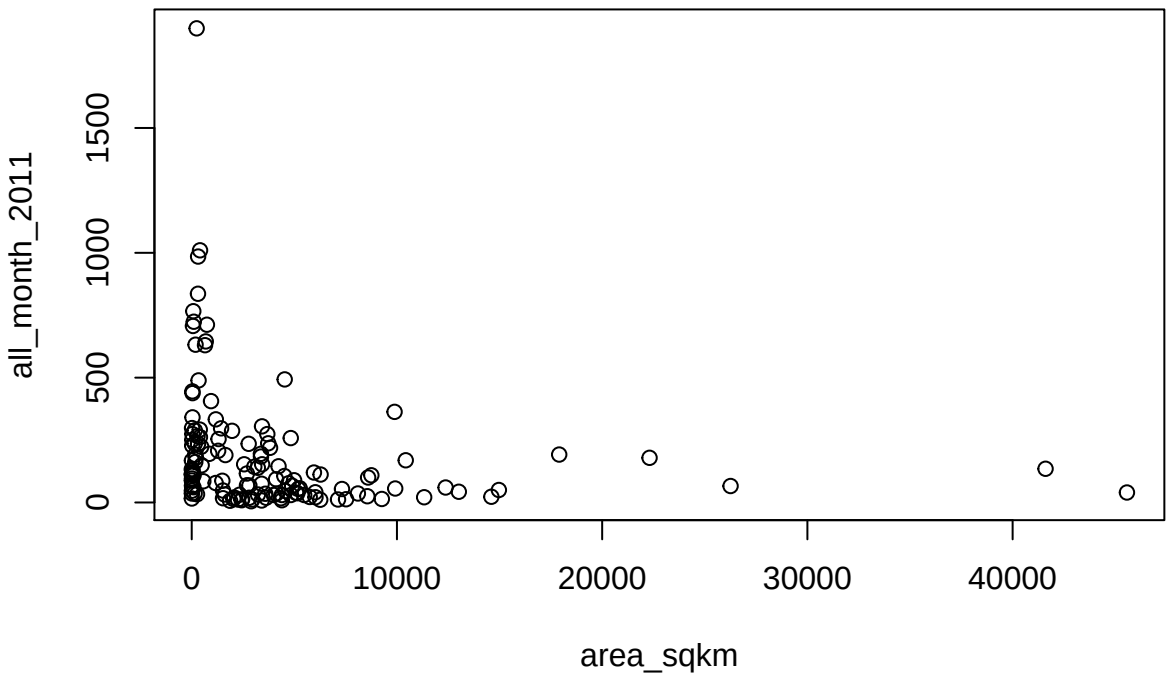
```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

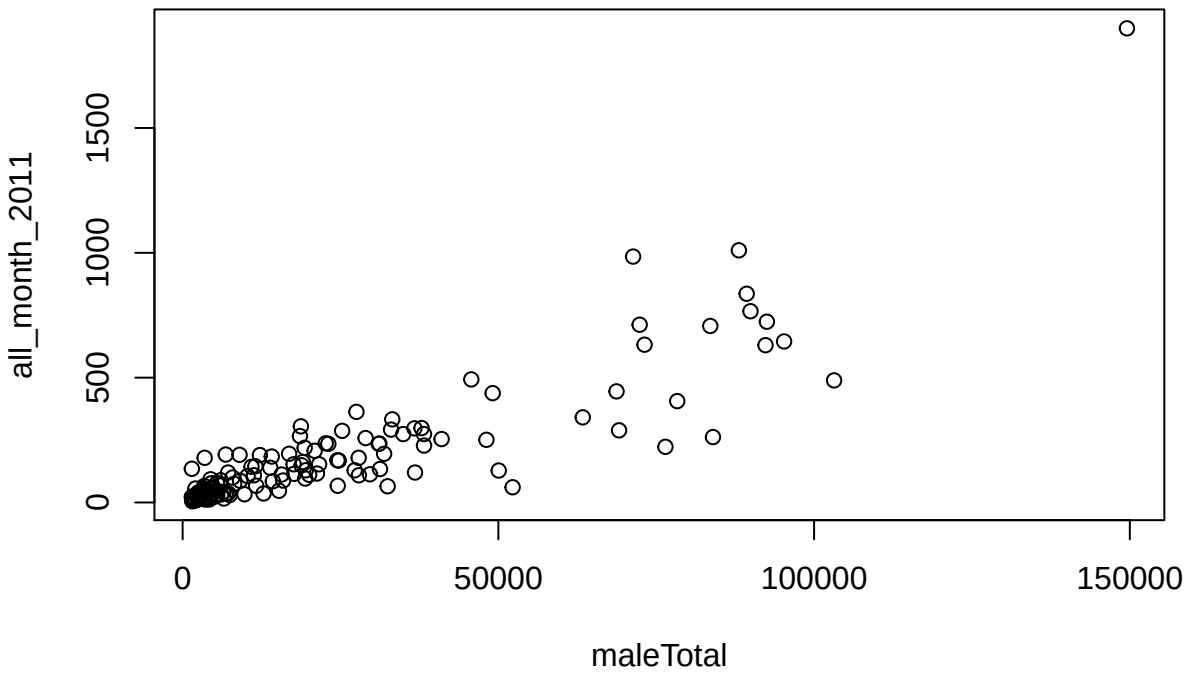
```
library(corrgram)

#Visualization the relation within the variables.

corrgram(data,order=TRUE, lower.panel=panel.shade,
  upper.panel=panel.pie, text.panel=panel.txt)
```



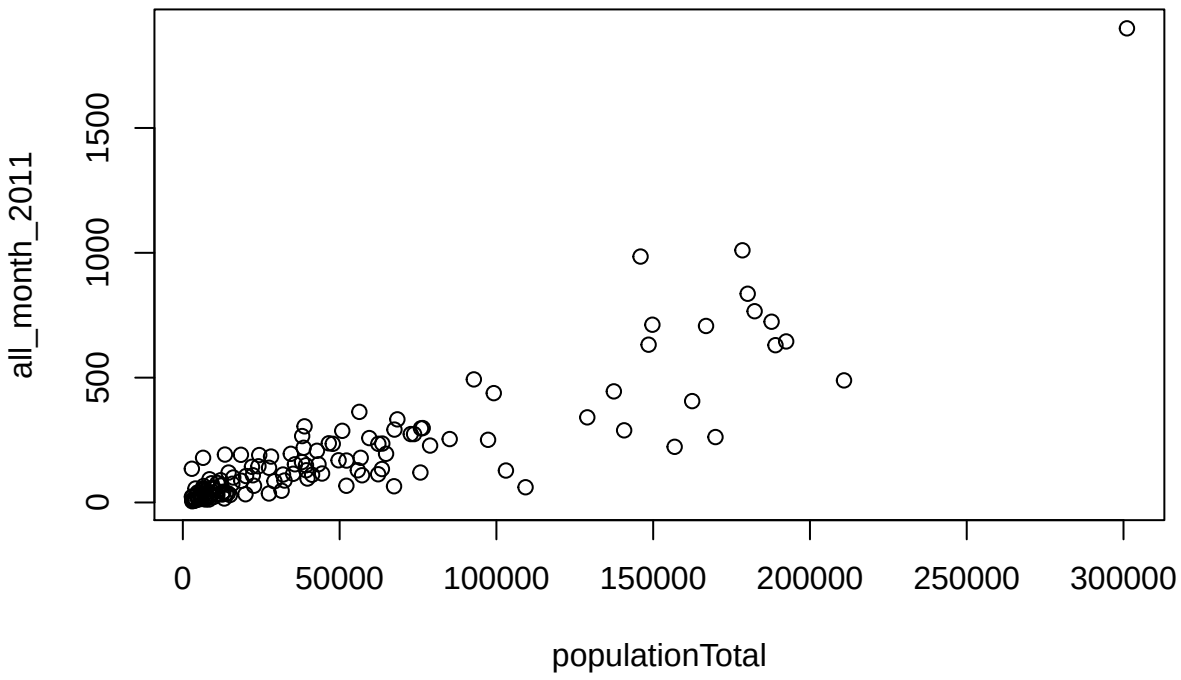

```
plot(all_month_2011~area_sqkm, data=data)
```



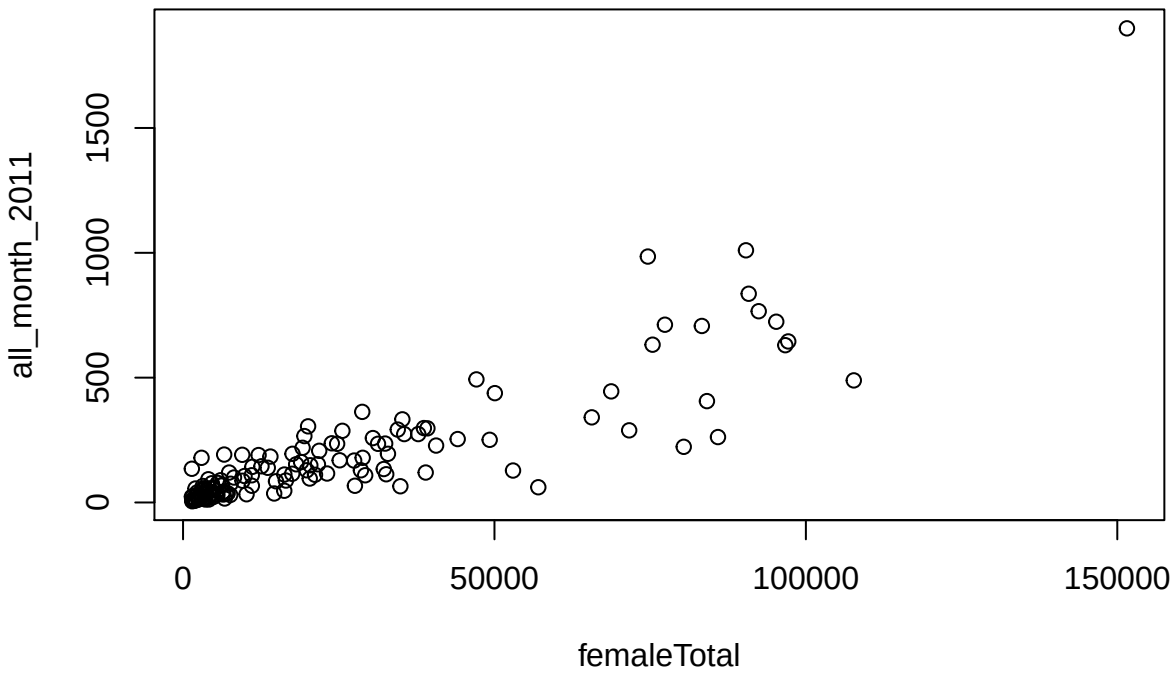

```
plot(all_month_2011~maleTotal, data=data)
```




```
plot(all_month_2011~populationTotal, data=data)
```



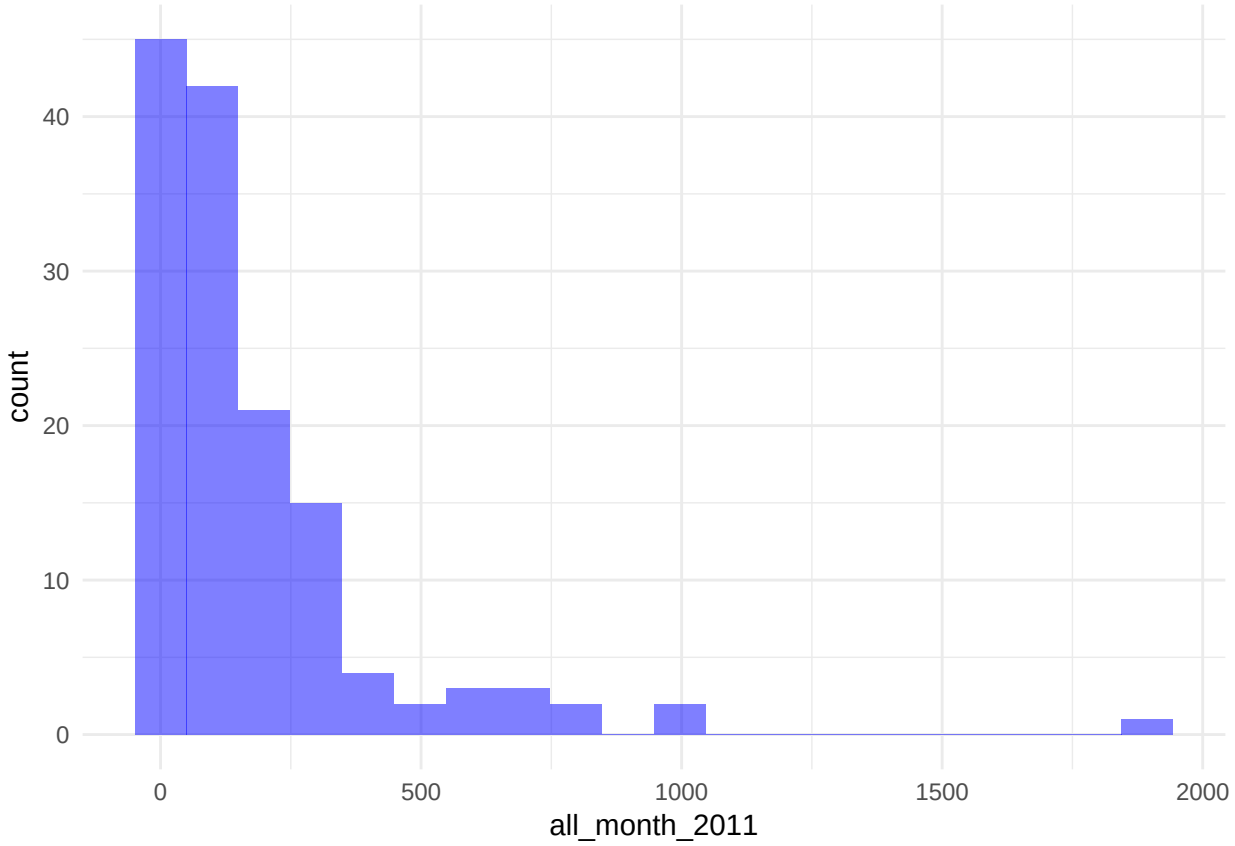

```
plot(all_month_2011~femaleTotal, data=data)
```



The pie in the figure shows that the blue area is perfectly co orelated where as the other color is mean that not perfectly corelated.

Since we're going to eventually try to see the domestic violance in the year of 2011, lets see the histogram of it:

```
library(ggplot2)
ggplot(data,aes(x=all_month_2011)) + geom_histogram(bins=20,alpha=0.5,fill='blue') + theme_minimal()
```



Building model

```
model <- lm(all_month_2011 ~area_sqkm+populationTotal+maleTotal+femaleTotal+
              Twenty_twentyfour_Total+twentyfive_thirtyfive_Total ,data)
```

Model summary

```
summary(model)
```

```
##
## Call:
## lm(formula = all_month_2011 ~ area_sqkm + populationTotal + maleTotal +
##     femaleTotal + Twenty_twentyfour_Total + twentyfive_thirtyfive_Total,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -417.93  -31.24   -4.07   38.54  555.54
##
## Coefficients: (1 not defined because of singularities)
```

```
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -19.346540  16.016704  -1.208 0.229214
## area_sqkm                     0.003164   0.001610   1.965 0.051492 .
## populationTotal              -0.043213   0.011371  -3.800 0.000218 ***
## maleTotal                     0.096153   0.023851   4.031 9.26e-05 ***
## femaleTotal                         NA         NA      NA       NA
## Twenty_twentyfour_Total      -0.011685   0.014950  -0.782 0.435839
## twentyfive_thirtyfive_Total   0.005036   0.005232   0.963 0.337525
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 112.7 on 134 degrees of freedom
## Multiple R-squared:  0.8015, Adjusted R-squared:  0.7941
## F-statistic: 108.2 on 5 and 134 DF,  p-value: < 2.2e-16
```

Model description and analysis/result:

R-Squire shows the model in a good fitting with the variable we are predicting.

From the probability of co efficient we can see, in terms of popolation and the male population the model is denoted very significance. which mean that the number of domesticviolance in NSW is really depends on the populatin or on the male, population. The standered error of the model is significantly low, which denotes that the model predecting a good result,which is very remarkable. the domestic violance is than may be related to the other factor like the income of the house hold or may be the educational levels of the indivisuals. where as we dont have that data available to analyse.

Model visualization

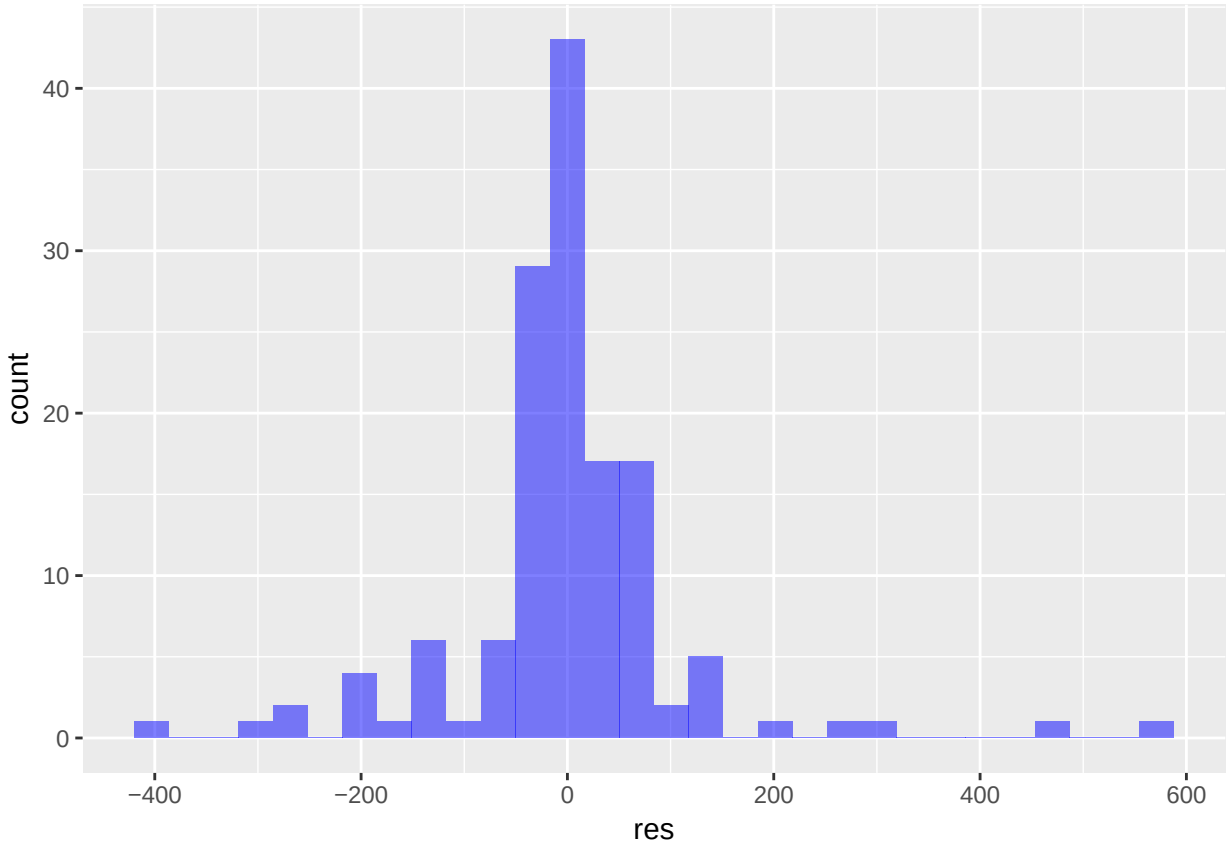
```
# Grab residuals
res <- residuals(model)

# Convert to DataFrame for gglpot
res <- as.data.frame(res)

head(res)
```

```
##          res
## 1  111.12872
## 2  101.31144
## 3  -17.13274
## 4 -192.12675
## 5   54.40066
## 6   39.90753
```

Using ggplot

```
# Histogram of residuals

library(ggplot2)
ggplot(res,aes(res)) +  geom_histogram(fill='blue',alpha=0.5)
```

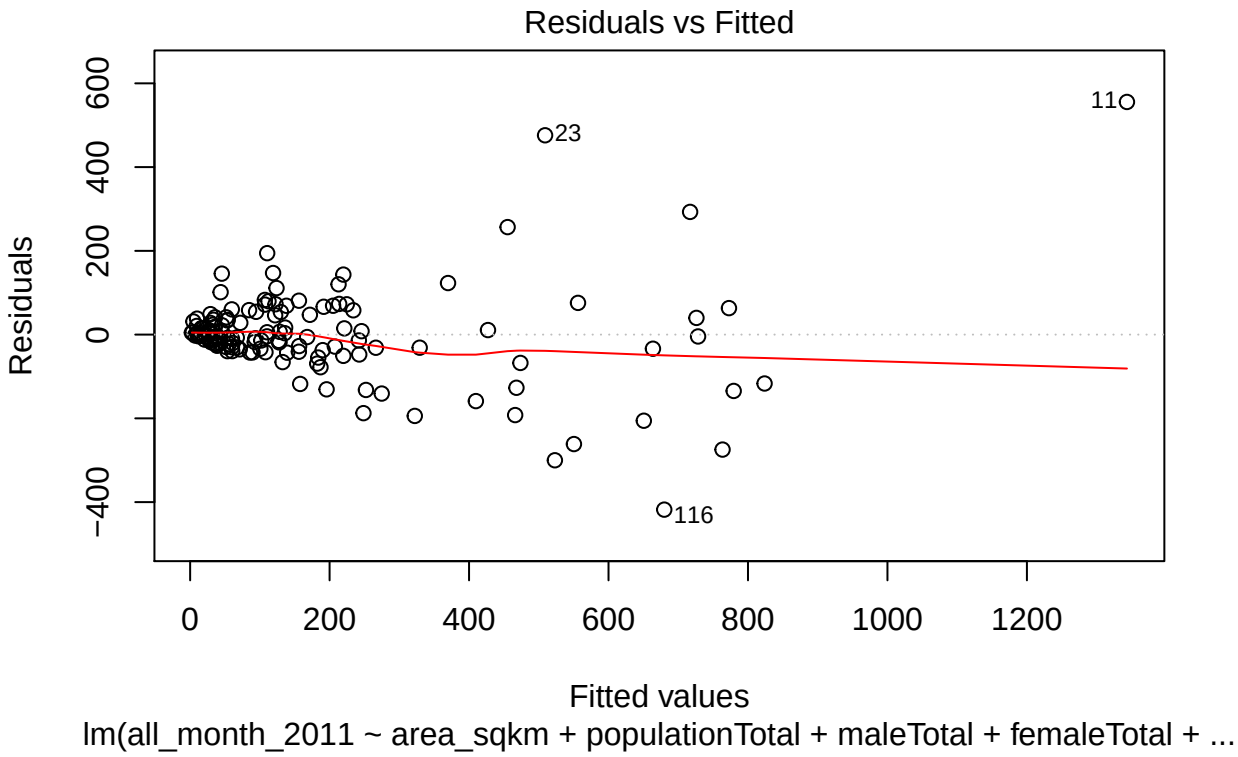
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

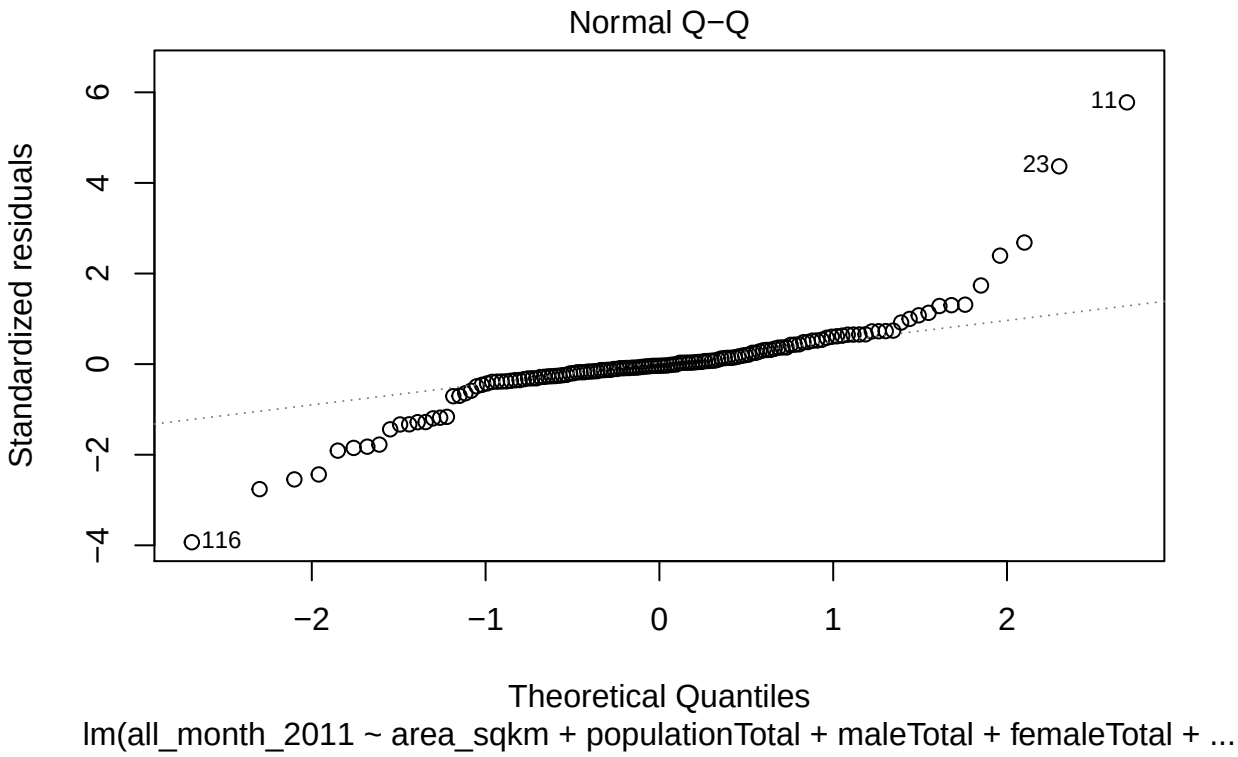


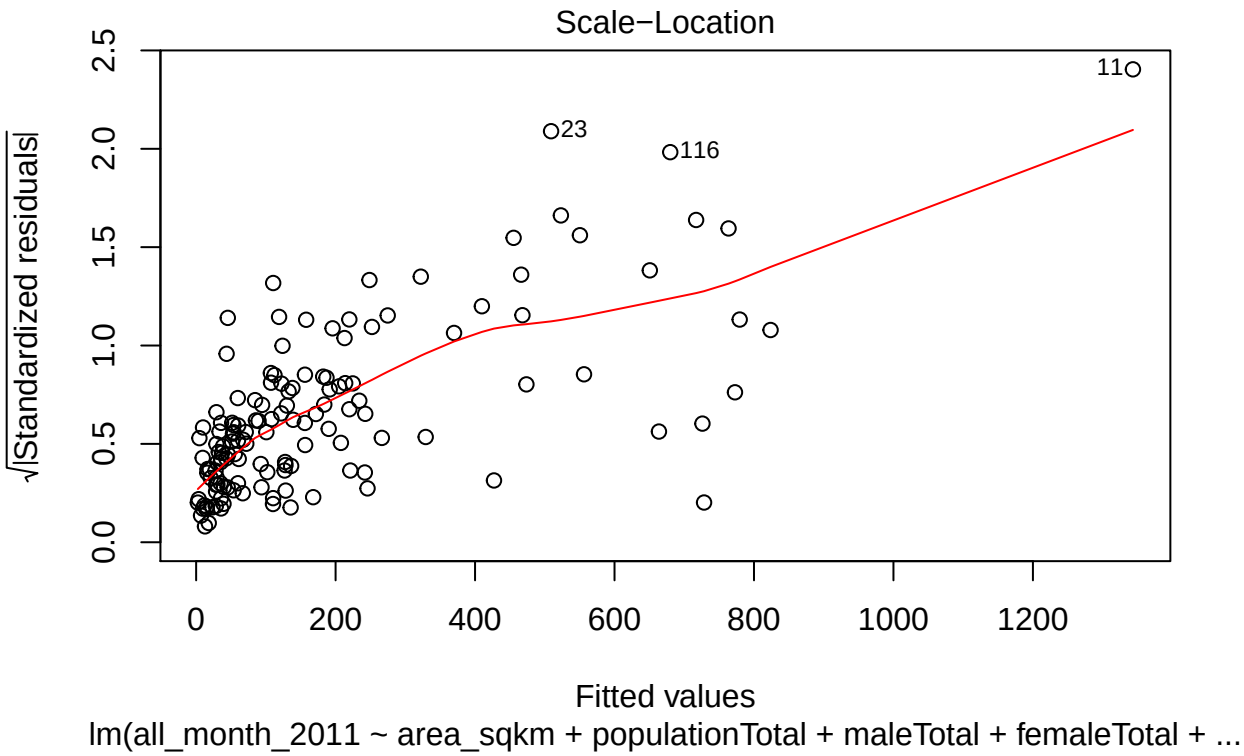
The above distribution of Residual shows that the model we made was preety good as the distribution is almost Gausian or the normal distribution.

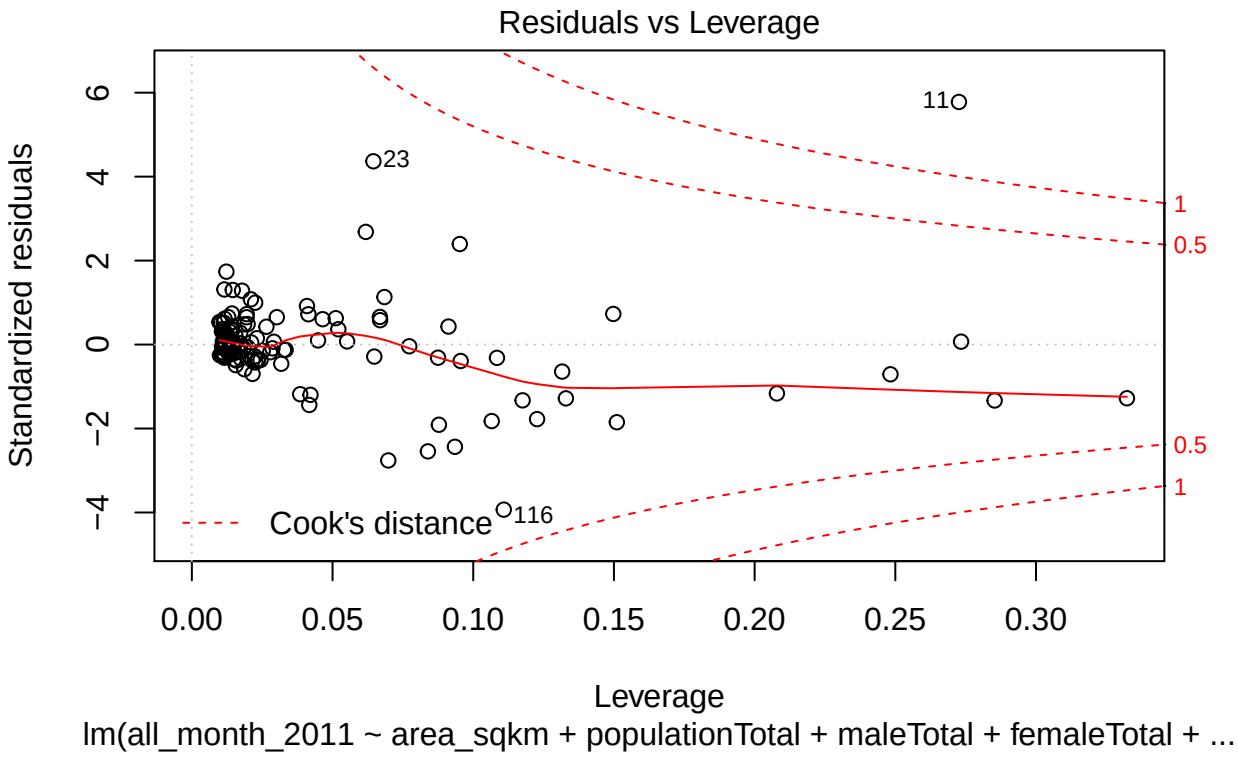
plot model

```
plot(model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(all_month_2011 ~ area_sqkm + populationTotal + maleTotal + femaleTotal + ...
11
23
116
600
400
200
0
−400
0
200
400
600
800
1000
1200


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(all_month_2011 ~ area_sqkm + populationTotal + maleTotal + femaleTotal + ...
11
23
116
−4
−2
0
2
4
6
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
0
200
400
600
800
1000
1200
11
23
116
Fitted values
lm(all_month_2011 ~ area_sqkm + populationTotal + maleTotal + femaleTotal + ...


Residuals vs Leverage
Standardized residuals
6
4
2
0
−2
−4
11
23
116
1
0.5
0.5
1
Cook's distance
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Leverage
lm(all_month_2011 ~ area_sqkm + populationTotal + maleTotal + femaleTotal + ...