Analysis and Data Mining of two large data sets using various Visualization Techniques

Rohit A. Khadse(L20407353)
Saurabh Mahajan (L20398116)

Guided by
Prof Sujing Wang

Department of Computer Science

Lamar University

Beaumont, Texas

Dataset

We analysed two data sets from the link [http://catalog.data.gov/dataset/campus-safety-and-security-survey-2013]

1. OnCampusArrest (*oncampusarrest101112.xls*)
2. NonCampusArrest (*noncampusarrest101112.xls*)

We used R Programming and MySQL for this and tried to find where Arrest Rate is higher. OnCampusArrest data set has 24 attributes and 11064 tuple entries while as NonCampusArrest data set has 24 attributes and 11064 tuple entries. OnCampusArrest data set and NonCampusArrest data set shows the tabular data of on campus and non-campus arrests that took place in various colleges. Following is data format in both data sets[1] :

Variables in Creation Order						
#	Variable	Type	Len	Format	Informat	Label
1	UNITID_P	Num	8			Unitid_plus
2	INSTNM	Char	93	$93.	$93.	Institution Name
3	BRANCH	Char	89	$89.	$89.	Branch Name
4	Address	Char	92	$92.	$92.	
5	City	Char	28	$28.	$28.	
6	State	Char	2	$2.	$2.	
7	Zip	Char	14	$14.	$14.	
8	sector_cd	Num	8			
9	sector_desc	Char	36	$36.	$36.	
10	men_total	Num	8			Total Men
11	women_total	Num	8			Total Women
12	Total	Num	8			Grand Total
13	Weapon10	Num	8			Weapons: carrying, possessing, etc. 2010
14	Drug10	Num	8			Drug Law Violations 2010
15	Liquor10	Num	8			Liquor Law Violations 2010
16	Weapon11	Num	8			Weapons: carrying, possessing, etc. 2011
17	Drug11	Num	8			Drug Law Violations 2011
18	Liquor11	Num	8			Liquor Law Violations 2011
19	Weapon12	Num	8			Weapons: carrying, possessing, etc. 2012

20	Drug12	Num	8			Drug Law Violations 2012
21	Liquor12	Num	8			Liquor Law Violations 2012
22	FILTER10	Num	8			Data_year = 2010 (FILTER)
23	FILTER11	Num	8			Data_year = 2011 (FILTER)
24	FILTER12	Num	8			Data_year = 2012 (FILTER)

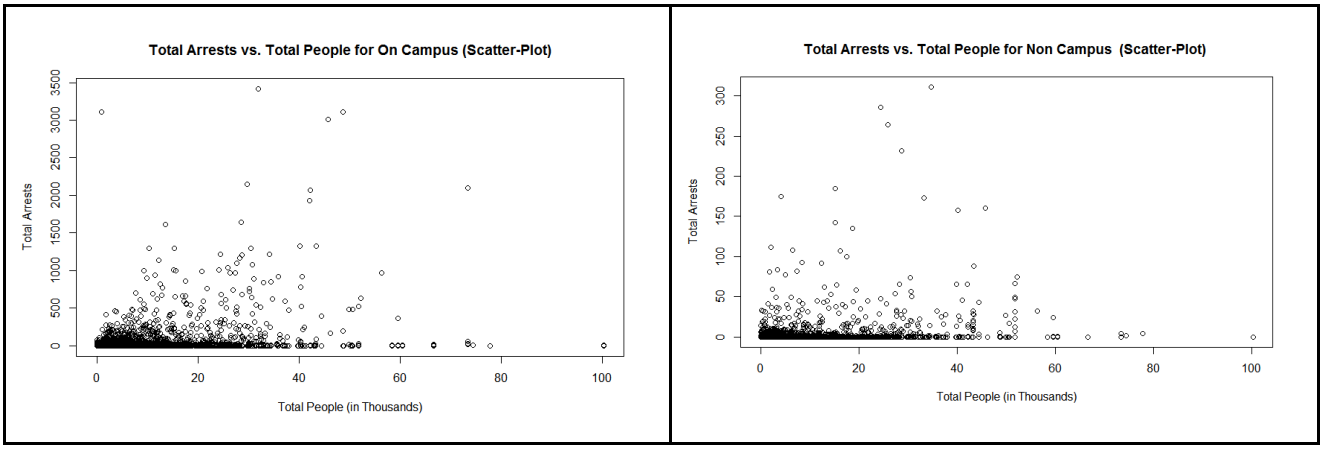
We are going to analyse the two data sets and interpret the areas where the arrest rate is high i.e. either On Campus or Non Campus. We will also use different visualization techniques to analyse datasets.

Pre-processing and Implementation

We had to pre-process data since most of Arrests/Violation related rows were blank (and not numeric). We imported files in MySQL database into OnCampus and NonCampus table using built in import tool in MySQL workbench. We applied DML queries to modify and clean both tables so we can query and perform analysis and visualization with it.

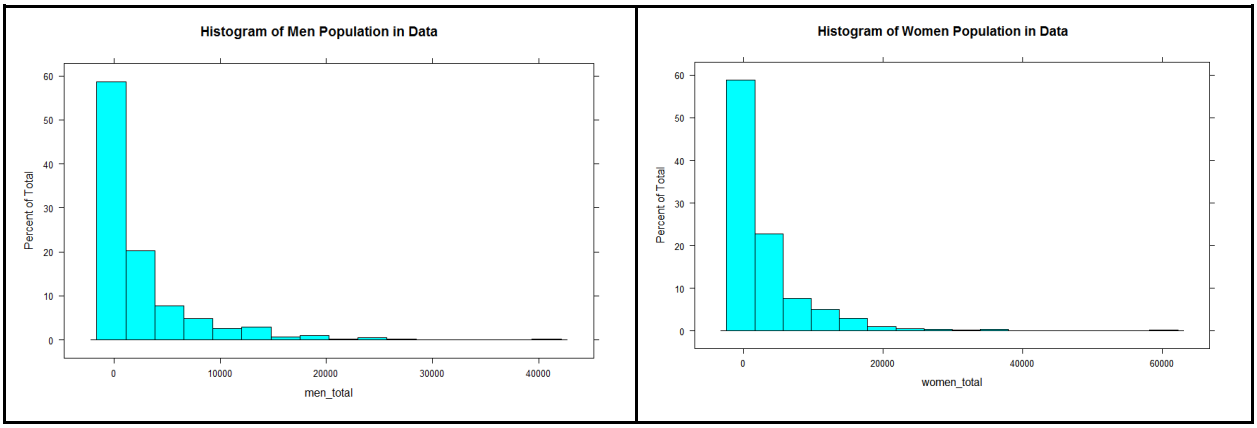
Results

Scatter Plot



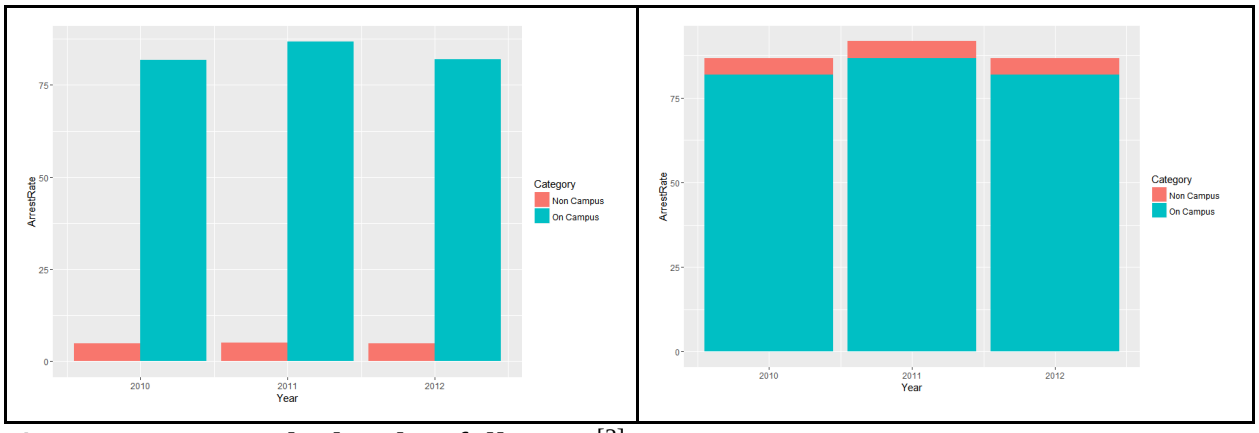
Above is Scatter Plots for Total Number of Arrests Vs Total number of People who live on-campus and non-campus. After comparing both scatterplots we can say that distribution of number of arrests and total number of people is similar. Hence Data set is ideal for comparison of Arrest Rate between two datasets.

Histogram



Histogram shows us distribution of combined women and men population in both data sets. We can deduce that most of colleges and universities have 2000~5000 population of men and 2000-10000 women.

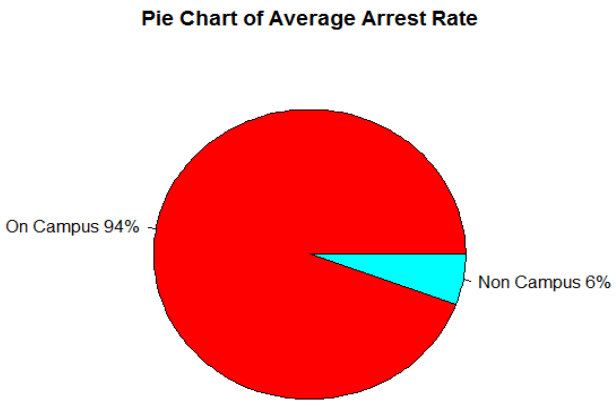
Bar Graph



Arrest Rate can calculated as follows : [2]
= Total Number of Arrests/Total Population * 100000
For example. If arrest rate is 678, then we say that on an average 678 people got arrested **per 100,000** people.

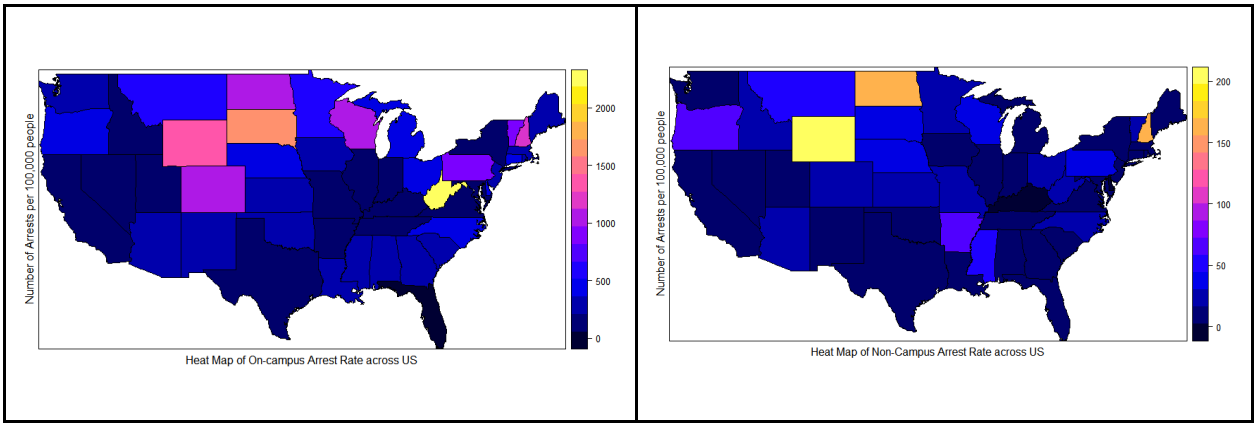
Above Bar Graph shows year wise (for three years: 2010, 2011, 2012) on-campus and non-campus arrest rate in datasets. We can clearly see on-campus arrest rate is significantly higher than non-campus arrest rate.

Pie Chart

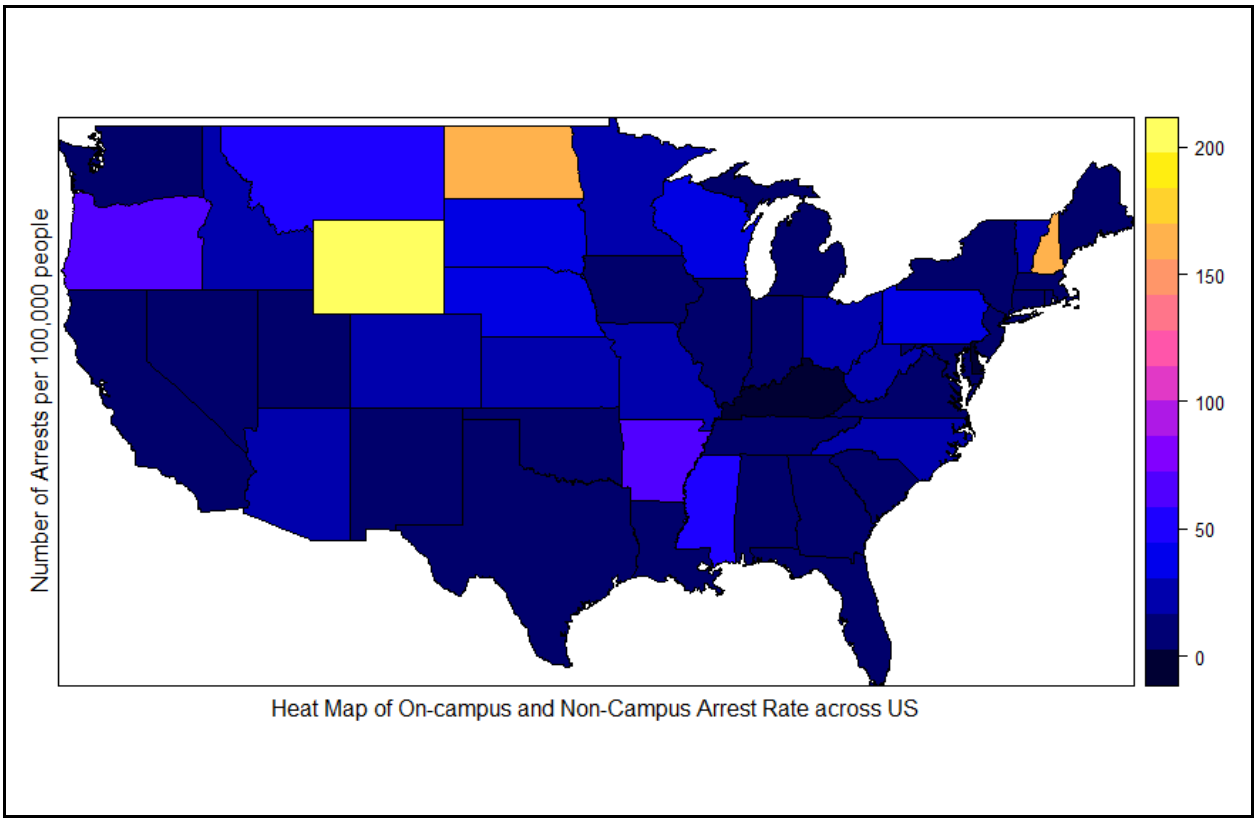


Pie Chart show arrest rate difference for On - Campus and Non-Campus

Heat Map

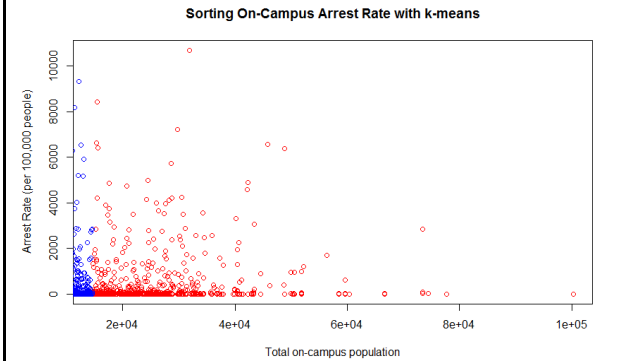


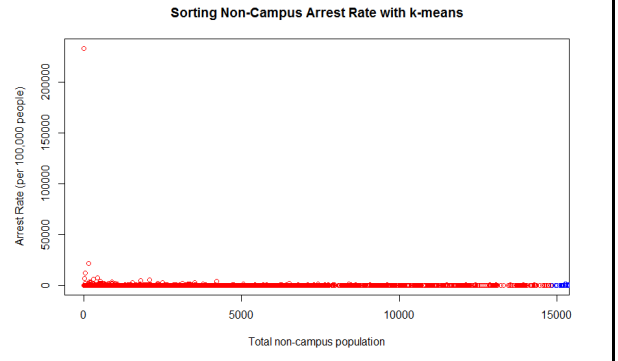
Heat map of On-campus Arrest Rate shows West Virginia, South Dakota and Wyoming has highest on-campus arrest rate in US. While as heat map of Non-Campus Arrest Rate shows North Dakota, Vermont and Wyoming has highest Non-campus arrest rate in US.



Above heatmap shows us that North Dakota, Vermont and Wyoming has highest Non-campus arrest rate in US.

K-means Plot Graph





We choose 2 centres. Hence number of clusters = 2.
We selected the two attributes- Total Campus Population and the Arrest rate. The Arrest Rate is taken from the Views that were created during the Pre-Processing. Since there are many nominal attributes in the data set, we preferred using just the two attributes that would yield the required results.

K-means plot graph for on-campus:

As Arrest Rate clusters towards positive zero we can conclude that On campus arrest rate is higher where population is low. Two Dividing clusters clearly shows that small private and public institutions (with low on-campus population) lack security infrastructure and planning and hence on-campus arrest rate is higher in less populated colleges.

K-means plot graph for Non-campus:

As Arrest Rate clusters towards positive infinity we can conclude that Non campus arrest rate is higher when population is high. Two Dividing clusters clearly shows that private and public institutions (with large non-campus population) lack enough security infrastructure and planning to handle large population and hence Non-campus arrest rate is higher in largely populated colleges.

Conclusion

From above visualized data diagrams, we can safely conclude that on-campus arrest rate is higher compared to non-campus arrest rate.

References

[1] http://catalog.data.gov/dataset/campus-safety-and-security-survey-2013
[2] https://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2012/crime-in-the-u.s.-2012/persons-arrested/persons-arrested
[3] https://cran.r-project.org/web/packages/RMySQL/RMySQL.pdf
[4] https://cran.r-project.org/web/packages/maptools/maptools.pdf
[5] https://cran.r-project.org/web/packages/maps/maps.pdf

Source Code

Pre-processing, Tables and Views

OnCampus Table

```
CREATE TABLE OnCampusArrest
(
ID int NOT NULL primary KEY AUTO_INCREMENT,
UNITID_P int,
INSTNM varchar(255),
BRANCH varchar(255),
Address varchar(255),
City varchar(255),
State varchar(255),
ZIP int,
sector_cd int,
Sector_desc varchar(255),
men_total int,
women_total int,
Total int,
WEAPON10 int,
DRUG10 int,
LIQUOR10 int,
WEAPON11 int,
DRUG11 int,
LIQUOR11 int,
WEAPON12 int,
DRUG12 int,
LIQUOR12 int,
FILTER10 int,
FILTER11 int,
FILTER12 int
);
```

Non Campus Table

```
CREATE TABLE NonCampusArrest
(
ID int NOT NULL primary KEY AUTO_INCREMENT,
UNITID_P int,
INSTNM varchar(255),
BRANCH varchar(255),
Address varchar(255),
City varchar(255),
State varchar(255),
ZIP int,
sector_cd int,
Sector_desc varchar(255),
men_total int,
women_total int,
Total int,
WEAPON10 int,
DRUG10 int,
LIQUOR10 int,
WEAPON11 int,
DRUG11 int,
LIQUOR11 int,
WEAPON12 int,
DRUG12 int,
LIQUOR12 int,
FILTER10 int,
FILTER11 int,
FILTER12 int
);
```

```
CREATE VIEW `vwOnCampusTotalArrestRate` AS
SELECT        ID, SUM(WEAPON10 + DRUG10 +
LIQUOR10 + WEAPON11 + DRUG11 + LIQUOR11 +
WEAPON12 + DRUG12 + LIQUOR12) AS TotalArrests,
SUM(Total) AS Total,
                        CASE WHEN SUM(Total)
= 0 THEN 0 ELSE SUM(WEAPON10 + DRUG10 +
LIQUOR10 + WEAPON11 + DRUG11 + LIQUOR11 +
WEAPON12 + DRUG12 + LIQUOR12)
                        * 100000 / SUM(Total)
END AS ArrestRate
FROM            rkhadse.OnCampusArrest
GROUP BY ID
```

```
CREATE VIEW `vwNonCampusTotalArrestRate` AS
SELECT        ID, SUM(WEAPON10 + DRUG10 +
LIQUOR10 + WEAPON11 + DRUG11 + LIQUOR11 +
WEAPON12 + DRUG12 + LIQUOR12) AS TotalArrests,
SUM(Total) AS Total,
                        CASE WHEN SUM(Total)
= 0 THEN 0 ELSE SUM(WEAPON10 + DRUG10 +
LIQUOR10 + WEAPON11 + DRUG11 + LIQUOR11 +
WEAPON12 + DRUG12 + LIQUOR12)
                        * 100000 / SUM(Total)
END AS ArrestRate
FROM            rkhadse.NonCampusArrest
GROUP BY ID
```

Scatter Plot

```
#Install Package only if it is not installed
#RMySQL connects to MySQL Database
if("RMySQL" %in% rownames(installed.packages()) == FALSE)
{install.packages("RMySQL")}
library(RMySQL)

#Generate a Connection String so that we can connect to database
con <- dbConnect(MySQL(),
                 user = 'rohitsaurabh',
                 password = 'rohit1991',
                 host = '50.62.209.88',
                 dbname='rkhadse')
#Query for On Campus
tmp <- sprintf("SELECT SUM(WEAPON10 + DRUG10+ LIQUOR10+ WEAPON11+ DRUG11+ LIQUOR11+ WEAPON12+ DRUG12+
LIQUOR12) as TotalArrests,SUM(Total)/1000 AS TotalPeople
FROM rkhadse.OnCampusArrest
GROUP BY ID ")

result <- dbGetQuery(con, tmp)
head(result)

plot(
    x = result$TotalPeople,
    y = result$TotalArrests,
    main = "Total Arrests vs. Total People for On Campus (Scatter-Plot)",
    xlab = "Total People (in Thousands)",
    ylab = "Total Arrests")

#Query for Non Campus
tmp <- sprintf("SELECT SUM(WEAPON10 + DRUG10+ LIQUOR10+ WEAPON11+ DRUG11+ LIQUOR11+ WEAPON12+ DRUG12+
LIQUOR12) as TotalArrests,SUM(Total)/1000 AS TotalPeople
FROM rkhadse.NonCampusArrest
GROUP BY ID ")

sqlquery<-dbEscapeStrings(con, tmp)

result <- dbGetQuery(con, tmp)
dbDisconnect(con)
head(result)

plot(
  x = result$TotalPeople,
  y = result$TotalArrests,
  main = "Total Arrests vs. Total People for Non Campus  (Scatter-Plot) ",
  xlab = "Total People (in Thousands)",
  ylab = "Total Arrests")
```

Histogram

```
if("RMySQL" %in% rownames(installed.packages()) == FALSE)
{install.packages("RMySQL")}

if("lattice" %in% rownames(installed.packages()) == FALSE)
{install.packages("lattice")}

library(lattice)
library(RMySQL)

con <- dbConnect(MySQL(),
                 user = 'rohitsaurabh',
                 password = 'rohit1991',
                 host = '50.62.209.88',
                 dbname='rkhadse')

tmp <- sprintf("
               SELECT
               men_total,women_total
               FROM rkhadse.OnCampusArrest
               UNION ALL
               SELECT
               men_total,women_total
               FROM rkhadse.NonCampusArrest
            ")

sqlquery<-dbEscapeStrings(con, tmp)

result <- dbGetQuery(con, tmp)
dbDisconnect(con)
head(result)
histogram(
    x = ~women_total,
    data = result,
    main = "Histogram of Women Population in Data")

histogram(
  x = ~men_total,
  data = result,
  main = "Histogram of Men Population in Data")
```

Bar Graph

```
if("RMySQL" %in% rownames(installed.packages()) == FALSE)
{install.packages("RMySQL")}
if("ggplot2" %in% rownames(installed.packages()) == FALSE)
{install.packages("ggplot2")}
library(RMySQL)
library(ggplot2)
con <- dbConnect(MySQL(), user = 'rohitsaurabh', password = 'rohit1991', host = '50.62.209.88',
dbname='rkhadse')
sqlOnCampusArrest <- sprintf(" SELECT
                Case when SUM(Total)=0 then 0   # To Take care of Divide by zero error
                Else
                SUM( WEAPON10 + DRUG10 + LIQUOR10) * 100000/SUM(Total) END AS '2010',
                Case when SUM(Total)=0 then 0
                Else
                SUM( WEAPON11 + DRUG11  + LIQUOR11) * 100000/SUM(Total) END AS '2011',
                Case when SUM(Total)=0 then 0
                Else
                SUM(WEAPON12 + DRUG12 + LIQUOR12) * 100000/SUM(Total) END AS '2012'
                FROM rkhadse.OnCampusArrest")
sqlNonCampusArrest <- sprintf(" SELECT
                Case when SUM(Total)=0 then 0   # To Take care of Divide by zero error
                Else
                SUM( WEAPON10 + DRUG10 + LIQUOR10) * 100000/SUM(Total) END AS '2010',
                Case when SUM(Total)=0 then 0
                Else
                SUM( WEAPON11 + DRUG11  + LIQUOR11) * 100000/SUM(Total) END AS '2011',
                Case when SUM(Total)=0 then 0
                Else
                SUM(WEAPON12 + DRUG12 + LIQUOR12) * 100000/SUM(Total) END AS '2012'
                FROM rkhadse.NonCampusArrest")
OnCampusArrest <- dbGetQuery(con, sqlOnCampusArrest)
NonCampusArrest <- dbGetQuery(con, sqlNonCampusArrest)
# Disconnect SQL Database
dbDisconnect(con)
OnCampusArrest
NonCampusArrest
Category = c("On Campus", "On Campus", "On Campus","Non Campus", "Non Campus", "Non Campus")
Year = c(2010, 2011, 2012,2010, 2011, 2012)
ArrestRate = c(OnCampusArrest[1, 1] , OnCampusArrest[1, 2] ,OnCampusArrest[1, 3] ,NonCampusArrest[1, 1]
, NonCampusArrest[1, 2] ,NonCampusArrest[1, 3] )
result <- data.frame( Category, Year,  ArrestRate )
result
# Bar graph, time on x-axis, color fill grouped by  Category (NonCampus,Campus)
ggplot(data=result, aes(x=Year, y=ArrestRate, fill=Category)) +
  geom_bar(stat="identity", position=position_dodge())

# Stacked bar graph
ggplot(data=result, aes(x=Year, y=ArrestRate, fill=Category)) +
  geom_bar(stat="identity")
```

Pie Chart

```
if("RMySQL" %in% rownames(installed.packages()) == FALSE)
{install.packages("RMySQL")}

library(RMySQL)

con <- dbConnect(MySQL(), user = 'rohitsaurabh', password = 'rohit1991', host = '50.62.209.88',
dbname='rkhadse')

#Query for Pie Chart
tmp <- sprintf("SELECT  'On Campus' as Category,
               CASE
               WHEN SUM(Total) = 0 THEN 0
               ELSE SUM(WEAPON10 + DRUG10 + LIQUOR10 + WEAPON11 + DRUG11 + LIQUOR11 + WEAPON12 + DRUG12
+ LIQUOR12) * 100000 / SUM(Total)
               END AS ArrestRate
               FROM
               rkhadse.OnCampusArrest
               union ALL
               SELECT
               'Non Campus' as Category,
               CASE
               WHEN SUM(Total) = 0 THEN 0
               ELSE SUM(WEAPON10 + DRUG10 + LIQUOR10 + WEAPON11 + DRUG11 + LIQUOR11 + WEAPON12 + DRUG12
+ LIQUOR12) * 100000 / SUM(Total)
               END AS ArrestRate
               FROM
               rkhadse.NonCampusArrest")

sqlquery<-dbEscapeStrings(con, tmp)

result <- dbGetQuery(con, tmp)
dbDisconnect(con)
head(result)

# Pie Chart with Percentages
slices <- result$ArrestRate
lbls <- result$Category
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=rainbow(length(lbls)),
    main="Pie Chart of Average Arrest Rate")
```

Heat Map

```
if("RMySQL" %in% rownames(installed.packages()) == FALSE) {install.packages("RMySQL")}
if("maps" %in% rownames(installed.packages()) == FALSE) {install.packages("maps")}
if("maptools" %in% rownames(installed.packages()) == FALSE) {install.packages("maptools")}
if("sp" %in% rownames(installed.packages()) == FALSE) {install.packages("sp")}
library(RMySQL)
library(maps)
library(maptools)
library(sp)
con <- dbConnect(MySQL(), user = 'rohitsaurabh', password = 'rohit1991', host = '50.62.209.88',
dbname='rkhadse')
tmp <- sprintf("SELECT State, Case when SUM(Total)=0 then 0 Else FLOOR(SUM( WEAPON10 + DRUG10 + LIQUOR10
+ WEAPON11 + DRUG11 + LIQUOR11 + WEAPON12 + DRUG12 + LIQUOR12) * 100000/SUM(Total))END AS ArrestRate
FROM rkhadse.OnCampusArrest UNION ALL SELECT State, Case when SUM(Total)=0 then 0 Else FLOOR(SUM(
WEAPON10 + DRUG10 + LIQUOR10 + WEAPON11 + DRUG11 + LIQUOR11 + WEAPON12 + DRUG12 + LIQUOR12) *
100000/SUM(Total)) END AS ArrestRate FROM rkhadse.NonCampusArrest where State != '' AND State IS NOT
NULL GROUP BY State ORDER BY State")
sqlquery<-dbEscapeStrings(con, tmp)
result <- dbGetQuery(con, tmp)
dbDisconnect(con)
head(result)
textstate <- paste(result$State,collapse=" ")
textarrestrate <- paste(result$ArrestRate,collapse=" ")
combinedmapdata <- c(textstate, textarrestrate)
txt <- paste(combinedmapdata,collapse=" \n ")
txt
#Library needs following format state (newline char \n) data
#txt <- "AB  AK  AL  AN  AR  AZ  CA  CO  CT  DC  DE  EN  FL  GA  HI  IA  ID  IL  IN  KS
#    1  21  31   1  12  56 316  53  31  16   7   1 335  63  11  42  29  73  40  2"

dat <- stack(read.table(text = txt,  header = TRUE))
#Inbuilt List of Abbrevations of States
names(dat)[2] <-'state.abb'
#Match it with our data and eliminate non-state values
dat$states <- tolower(state.name[match(dat$state.abb,  state.abb)])

mapUSA <- map('state',  fill = TRUE,  plot = FALSE)
nms <- sapply(strsplit(mapUSA$names,  ':'),  function(x)x[1])
USApolygons <- map2SpatialPolygons(mapUSA,  IDs = nms,  CRS('+proj=longlat'))

idx <- match(unique(nms),  dat$states)
dat2 <- data.frame(value = dat$value[idx], state = unique(nms))
row.names(dat2) <- unique(nms)

USAsp <- SpatialPolygonsDataFrame(USApolygons,  data = dat2)

spplot(USAsp['value'],  xlab = "Heat Map of On-campus and Non-Campus Arrest Rate across US", ylab =
"Number of Arrests per 100,000 people")
```

K-Means Plot Graph

Source Code for On Campus Cluster

```
if("RMySQL" %in% rownames(installed.packages()) == FALSE)
{install.packages("RMySQL")}
if("psych" %in% rownames(installed.packages()) == FALSE)
{install.packages("psych")}
library(RMySQL)
library(psych)
con <- dbConnect(MySQL(), user = 'rohitsaurabh', password = 'rohit1991', host = '50.62.209.88',
dbname='rkhadse')

tmp <- sprintf("SELECT
               Total,ArrestRate  AS ArrestRate
               FROM rkhadse.vwOnCampusArrestVsPopulation AS OnCampusArrests ")

result <- dbGetQuery(con, tmp)
dbDisconnect(con)
head(result)
plot(ArrestRate~Total,result)
result.kmeans <- kmeans(result,centers = 2);
result.kmeans$centers
result.kmeans$cluster

plot(
  result[result.kmeans$cluster==1,],col="red",main="Sorting On-Campus Arrest Rate with k-
means",xlab="Total on-campus population",ylab ="Arrest Rate (per 100,000 people)")
points(result[result.kmeans$cluster==2,], col="blue")
```

Source Code for Non-Campus Cluster

```
if("RMySQL" %in% rownames(installed.packages()) == FALSE)
{install.packages("RMySQL")}
if("psych" %in% rownames(installed.packages()) == FALSE)
{install.packages("psych")}
library(RMySQL)
library(psych)
con <- dbConnect(MySQL(), user = 'rohitsaurabh', password = 'rohit1991', host = '50.62.209.88',
dbname='rkhadse')
tmp <- sprintf("SELECT
               Total,ArrestRate  AS ArrestRate
               FROM rkhadse.vwNonCampusArrestVsPopulation AS NonCampusArrests ")
result <- dbGetQuery(con, tmp)
dbDisconnect(con)
head(result)
plot(ArrestRate~Total,result)
result.kmeans <- kmeans(result,centers = 2);
result.kmeans$centers
result.kmeans$cluster
plot(
  result[result.kmeans$cluster==1,],col="red",main="Sorting Non-Campus Arrest Rate with k-
means",xlab="Total Non-campus population",ylab ="Arrest Rate (per 100,000 people)")
points(result[result.kmeans$cluster==2,], col="blue")
```