# Statistics - Week 1

Jimmy Tsz Ming Yue*

University of Sydney

jyue6728@uni.sydney.edu.au

Semester 2 Statistics 2018

## Contents



## 1 Importing Files and Important Functions

To do analysis of files, we first need to import into a the R program. We can use use the code:

```
> #read.delim()
> #scan
> #read.table
```

We can find help on any function in R by placing a ? mark in front of a command: for example

```
> ?read.table()
```

This is equivalent to the help function:

*440159151

```
> help("read.table")
```

We can save a workspace in R using the function:

```
> save.image()
```

which saves parameters and command history.

## 1.1 Basics: Directories

To get the current working directory:

```
> getwd()
```

```
[1] "/home/jyue/Documents/MDS/STAT/Data_w1"
```

Furthermore we can change directory using the setwd() command:

```
> setwd("/home/jyue/Documents/MDS/STAT/Data_w1/")
```

We can also save image with a specific name as:

```
> save.image("week1.Rdata")
```

## 1.2 Search paths and packages

In R there exists a base package, which exsists with an installation enviroment of R. There also exists community built "contributed" packages which are available for installation. The search funcction gives a list of attached backages and R objects.

```
> search()
```

```
[1] ".GlobalEnv"        "package:stats"     "package:graphics"
[4] "package:grDevices" "package:utils"     "package:datasets"
[7] "package:methods"   "Autoloads"         "package:base"
```

```
> library(cluster)
> search()
```

```
 [1] ".GlobalEnv"        "package:cluster"   "package:stats"
 [4] "package:graphics"  "package:grDevices" "package:utils"
 [7] "package:datasets"  "package:methods"   "Autoloads"
[10] "package:base"
```

as we can see we have that the third entry of the search function is the cluster library that we have loaded.

# 2 Basic elements in R

## 2.1 Concatenate and is.datatype

Normally in R we work with vectors or Matricies in R. (Another one is a list but that shall be in later sections) The simplest data structure is a numeric vector, which is a singular entity consisting of an ordered collection of numbers. To generate a vector we use the concatenate function:

```
> x = c(10.4, 5.6, 3.1, 6.4, 21.7)
> x

[1] 10.4  5.6  3.1  6.4 21.7

> x <- c(10.4, 5.6, 3.1, 6.4, 21.7)
> x

[1] 10.4  5.6  3.1  6.4 21.7
```

which are numeric vectors. We can verify that it is numeric through the is.numeric() function:

```
> is.numeric(x)

[1] TRUE
```

if we have strings included the boolean returned from is.numeric will return false:

```
> x = c(10.4, "5.6", 3.1, 6.4, 21.7)
> is.numeric(x)

[1] FALSE
```

similarly we have the is.character command for strings and is.logical for boolean results:

```
> X <- c("a", "b", "c3", "4")
> is.character(X)

[1] TRUE

> X <- c(FALSE, FALSE, TRUE, FALSE)
> is.logical(X)

[1] TRUE
```

## 2.2 Package Installations

We can install packages using the install.package(packagename)

We load packages using the forementioned libarary function:

```
> library(e1071)
```

# 3 Matrices

WE can create matrices using the matrix function:

```
> mymatrix <- matrix(1:20,5,4)
> mymatrix

     [,1] [,2] [,3] [,4]
[1,]    1    6   11   16
[2,]    2    7   12   17
[3,]    3    8   13   18
[4,]    4    9   14   19
[5,]    5   10   15   20
```

we can find subsets of matrices through indexing:

```
> mymatrix[1,2]

[1] 6

> #first row second column
> mymatrix[1,]

[1]  1  6 11 16

> #frist row
> mymatrix[,1]

[1] 1 2 3 4 5

> #first column
> mymatrix[1:2,]

     [,1] [,2] [,3] [,4]
[1,]    1    6   11   16
[2,]    2    7   12   17

> #first and second column
> mymatrix[c(1,3),]

     [,1] [,2] [,3] [,4]
[1,]    1    6   11   16
[2,]    3    8   13   18

> #first and third column
```

# 4 R markdown

R markdown is shit so we skip this section

# 5 Review of Statistical Concepts

## 5.1 Population and Samples

**Definition.** Population: The set of data corresponding to the entire collection of units about which information is sought

Examples of populations include:

- Blood Presure: blood pressure readings of all people in Australia
- The number of languages spoken from ALL currently enrolled students in University of Sydney

**Definition.** Sample: A subset of population data that are actually collected in the course of a study.

Examples of samples include:

- Blood pressure readings of 1000 randomly selected people in Australia
- The number of lanugages spoken from 500 randomly selected students currently enrolled in University of Sydney

In most studies, it is difficult to obtain information about the whole population. That is why we rely on samples to make estimates and inferences related to the whole population.

## 5.2 Parameters vs Statistics

A parameter is a number that describes a population A statistic is a number that describes a sample We often estsimate parameters thorough looking at statistics. Population parameters are notationally denoted using Greek letters such as $\mu, \sigma$ whereas statistics we use roman letters such as : $x, s$ or we can put hats on greek letters such as: $\hat{\mu}, \hat{\sigma}$ A parameter is a fixed number usually unknown. A statistic is a variable whose value varies from sample to sample

## 5.3 Descriptive Statistics

Many methods are available for summarising data in both nnumeric and graphical form

### 5.3.1 Numeric

For measures of location we use Mean, Mode, Median For measures of Spread we use: Standard Deviation, Median absolute deviation, IQR (Inter quartile Range) we can also use Min, Max Quartle, Five num summaries

### 5.3.2 Mean

Consider a sample of data drawn from some population

$$\{x_1, \dots x_n\} \tag{1}$$

Definition of sample mean: The sum of all observations divided by the number of observations. It is written in symbols as:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2}$$

Example: Consider the following data set: 23, 34, 32, 33, 34, 22, 32, 29, 29, 34, 32, 31
Sample mean = 365/12 = 30.4

### 5.3.3 Median

The median of a set of data is a value $\tilde{x}$ such that at least one half of the observations are less than or equal to $\tilde{x}$ and at least one half of the observations are greater than or equal to $\tilde{x}$. Definition of Sample median is:

(a) The $(n+1)/2$ largest observation if $n$ odd

(b) The average of the $n/2$ and the $n/2+1$ if $n$ even

### 5.3.4 Mode

The mode is the most frequently occuring value amongst all the observations in a sample

### 5.3.5 Mode or Median

Both the median and the mean are measures of location, but which is preferable?.
For symmetric data, the mean is usually less variable from sample to sample than the median.
For skewed data, the median is a better measure of location.
The median does not reacted as much as the mean by outliers. This property of the median is known as 'robustness'.

### 5.3.6 Range

The range of a list is the largest value minus the smallest value. This gives a quick feeling for the overall spread – but is misleading because it is solely influenced by two most extreme values.

### 5.3.7 Graphical

# 6 Tutorial 1

1. Download Communities and Crime dataset from
https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime

2. Read in the data using RStudio

**Solution.** We use the read table command to import the data:

```
> data <- read.csv("communities.data", na.strings = "?", header = FALSE)
> #head(data)
> name <- read.table("communities.names", sep = " ", header = FALSE)
> #name[,2]
> df <- data.frame(data)
> colnames(df) <- name[,2]
> head(df)

  state county community       communityname fold population householdsize
1     8     NA        NA         Lakewoodcity    1       0.19          0.33
2    53     NA        NA          Tukwilacity    1       0.00          0.16
3    24     NA        NA         Aberdeentown    1       0.00          0.42
4    34      5     81440 Willingborotownship    1       0.04          0.77
5    42     95      6096   Bethlehemtownship    1       0.01          0.55
6     6     NA        NA   SouthPasadenacity    1       0.02          0.28
  racepctblack racePctWhite racePctAsian racePctHisp agePct12t21 agePct12t29
1         0.02         0.90         0.12        0.17        0.34        0.47
2         0.12         0.74         0.45        0.07        0.26        0.59
3         0.49         0.56         0.17        0.04        0.39        0.47
4         1.00         0.08         0.12        0.10        0.51        0.50
5         0.02         0.95         0.09        0.05        0.38        0.38
6         0.06         0.54         1.00        0.25        0.31        0.48
  agePct16t24 agePct65up numbUrban pctUrban medIncome pctWWage pctWFarmSelf
1        0.29       0.32      0.20      1.0      0.37     0.72         0.34
```

```
2         0.35        0.27       0.02       1.0       0.31      0.72          0.11
3         0.28        0.32       0.00       0.0       0.30      0.58          0.19
4         0.34        0.21       0.06       1.0       0.58      0.89          0.21
5         0.23        0.36       0.02       0.9       0.50      0.72          0.16
6         0.27        0.37       0.04       1.0       0.52      0.68          0.20
  pctWInvInc pctWSocSec pctWPubAsst pctWRetire medFamInc perCapInc whitePerCap
1       0.60       0.29        0.15       0.43      0.39      0.40        0.39
2       0.45       0.25        0.29       0.39      0.29      0.37        0.38
3       0.39       0.38        0.40       0.84      0.28      0.27        0.29
4       0.43       0.36        0.20       0.82      0.51      0.36        0.40
5       0.68       0.44        0.11       0.71      0.46      0.43        0.41
6       0.61       0.28        0.15       0.25      0.62      0.72        0.76
  blackPerCap indianPerCap AsianPerCap OtherPerCap HispPerCap NumUnderPov
1        0.32         0.27        0.27        0.36       0.41        0.08
2        0.33         0.16        0.30        0.22       0.35        0.01
3        0.27         0.07        0.29        0.28       0.39        0.01
4        0.39         0.16        0.25        0.36       0.44        0.01
5        0.28         0.00        0.74        0.51       0.48        0.00
6        0.77         0.28        0.52        0.48       0.60        0.01
  PctPopUnderPov PctLess9thGrade PctNotHSGrad PctBSorMore PctUnemployed
1           0.19            0.10         0.18        0.48          0.27
2           0.24            0.14         0.24        0.30          0.27
3           0.27            0.27         0.43        0.19          0.36
4           0.10            0.09         0.25        0.31          0.33
5           0.06            0.25         0.30        0.33          0.12
6           0.12            0.13         0.12        0.80          0.10
  PctEmploy PctEmplManu PctEmplProfServ PctOccupManu PctOccupMgmtProf
1      0.68        0.23            0.41         0.25             0.52
2      0.73        0.57            0.15         0.42             0.36
3      0.58        0.32            0.29         0.49             0.32
4      0.71        0.36            0.45         0.37             0.39
5      0.65        0.67            0.38         0.42             0.46
6      0.65        0.19            0.77         0.06             0.91
  MalePctDivorce MalePctNevMarr FemalePctDiv TotalPctDiv PersPerFam PctFam2Par
1           0.68           0.40         0.75        0.75       0.35       0.55
2           1.00           0.63         0.91        1.00       0.29       0.43
3           0.63           0.41         0.71        0.70       0.45       0.42
4           0.34           0.45         0.49        0.44       0.75       0.65
5           0.22           0.27         0.20        0.21       0.51       0.91
6           0.49           0.57         0.61        0.58       0.44       0.62
  PctKids2Par PctYoungKids2Par PctTeen2Par PctWorkMomYoungKids PctWorkMom
1        0.59             0.61        0.56                0.74       0.76
2        0.47             0.60        0.39                0.46       0.53
3        0.44             0.43        0.43                0.71       0.67
4        0.54             0.83        0.65                0.85       0.86
5        0.91             0.89        0.85                0.40       0.60
6        0.69             0.87        0.53                0.30       0.43
  NumIlleg PctIlleg NumImmig PctImmigRecent PctImmigRec5 PctImmigRec8
1     0.04     0.14     0.03           0.24         0.27         0.37
2     0.00     0.24     0.01           0.52         0.62         0.64
```

```
3     0.01     0.46     0.00            0.07           0.06           0.15
4     0.03     0.33     0.02            0.11           0.20           0.30
5     0.00     0.06     0.00            0.03           0.07           0.20
6     0.00     0.11     0.04            0.30           0.35           0.43
  PctImmigRec10 PctRecentImmig PctRecImmig5 PctRecImmig8 PctRecImmig10
1          0.39           0.07         0.07         0.08          0.08
2          0.63           0.25         0.27         0.25          0.23
3          0.19           0.02         0.02         0.04          0.05
4          0.31           0.05         0.08         0.11          0.11
5          0.27           0.01         0.02         0.04          0.05
6          0.47           0.50         0.50         0.56          0.57
  PctSpeakEnglOnly PctNotSpeakEnglWell PctLargHouseFam PctLargHouseOccup
1             0.89                0.06            0.14              0.13
2             0.84                0.10            0.16              0.10
3             0.88                0.04            0.20              0.20
4             0.81                0.08            0.56              0.62
5             0.88                0.05            0.16              0.19
6             0.45                0.28            0.25              0.19
  PersPerOccupHous PersPerOwnOccHous PersPerRentOccHous PctPersOwnOccup
1             0.33              0.39               0.28            0.55
2             0.17              0.29               0.17            0.26
3             0.46              0.52               0.43            0.42
4             0.85              0.77               1.00            0.94
5             0.59              0.60               0.37            0.89
6             0.29              0.53               0.18            0.39
  PctPersDenseHous PctHousLess3BR MedNumBR HousVacant PctHousOccup
1             0.09           0.51      0.5       0.21         0.71
2             0.20           0.82      0.0       0.02         0.79
3             0.15           0.51      0.5       0.01         0.86
4             0.12           0.01      0.5       0.01         0.97
5             0.02           0.19      0.5       0.01         0.89
6             0.26           0.73      0.0       0.02         0.84
  PctHousOwnOcc PctVacantBoarded PctVacMore6Mos MedYrHousBuilt PctHousNoPhone
1          0.52             0.05           0.26           0.65           0.14
2          0.24             0.02           0.25           0.65           0.16
3          0.41             0.29           0.30           0.52           0.47
4          0.96             0.60           0.47           0.52           0.11
5          0.87             0.04           0.55           0.73           0.05
6          0.30             0.16           0.28           0.25           0.02
  PctWOFullPlumb OwnOccLowQuart OwnOccMedVal OwnOccHiQuart RentLowQ RentMedian
1           0.06           0.22         0.19          0.18     0.36       0.35
2           0.00           0.21         0.20          0.21     0.42       0.38
3           0.45           0.18         0.17          0.16     0.27       0.29
4           0.11           0.24         0.21          0.19     0.75       0.70
5           0.14           0.31         0.31          0.30     0.40       0.36
6           0.05           0.94         1.00          1.00     0.67       0.63
  RentHighQ MedRent MedRentPctHousInc MedOwnCostPctInc MedOwnCostPctIncNoMtg
1      0.38    0.34              0.38             0.46                  0.25
2      0.40    0.37              0.29             0.32                  0.18
3      0.27    0.31              0.48             0.39                  0.28
```

```
4      0.77    0.89              0.63             0.51                  0.47
5      0.38    0.38              0.22             0.51                  0.21
6      0.68    0.62              0.47             0.59                  0.11
  NumInShelters NumStreet PctForeignBorn PctBornSameState PctSameHouse85
1          0.04         0           0.12             0.42           0.50
2          0.00         0           0.21             0.50           0.34
3          0.00         0           0.14             0.49           0.54
4          0.00         0           0.19             0.30           0.73
5          0.00         0           0.11             0.72           0.64
6          0.00         0           0.70             0.42           0.49
  PctSameCity85 PctSameState85 LemasSwornFT LemasSwFTPerPop LemasSwFTFieldOps
1          0.51           0.64         0.03            0.13              0.96
2          0.60           0.52           NA              NA                NA
3          0.67           0.56           NA              NA                NA
4          0.64           0.65           NA              NA                NA
5          0.61           0.53           NA              NA                NA
6          0.73           0.64           NA              NA                NA
  LemasSwFTFieldPerPop LemasTotalReq LemasTotReqPerPop PolicReqPerOffic
1                 0.17          0.06              0.18             0.44
2                   NA            NA                NA               NA
3                   NA            NA                NA               NA
4                   NA            NA                NA               NA
5                   NA            NA                NA               NA
6                   NA            NA                NA               NA
  PolicPerPop RacialMatchCommPol PctPolicWhite PctPolicBlack PctPolicHisp
1        0.13               0.94          0.93          0.03         0.07
2          NA                 NA            NA            NA           NA
3          NA                 NA            NA            NA           NA
4          NA                 NA            NA            NA           NA
5          NA                 NA            NA            NA           NA
6          NA                 NA            NA            NA           NA
  PctPolicAsian PctPolicMinor OfficAssgnDrugUnits NumKindsDrugsSeiz
1           0.1          0.07                0.02              0.57
2            NA            NA                  NA                NA
3            NA            NA                  NA                NA
4            NA            NA                  NA                NA
5            NA            NA                  NA                NA
6            NA            NA                  NA                NA
  PolicAveOTWorked LandArea PopDens PctUsePubTrans PolicCars PolicOperBudg
1             0.29     0.12    0.26           0.20      0.06          0.04
2               NA     0.02    0.12           0.45        NA            NA
3               NA     0.01    0.21           0.02        NA            NA
4               NA     0.02    0.39           0.28        NA            NA
5               NA     0.04    0.09           0.02        NA            NA
6               NA     0.01    0.58           0.10        NA            NA
  LemasPctPolicOnPatr LemasGangUnitDeploy LemasPctOfficDrugUn PolicBudgPerPop
1                 0.9                 0.5                0.32            0.14
2                  NA                  NA                0.00              NA
3                  NA                  NA                0.00              NA
4                  NA                  NA                0.00              NA
```

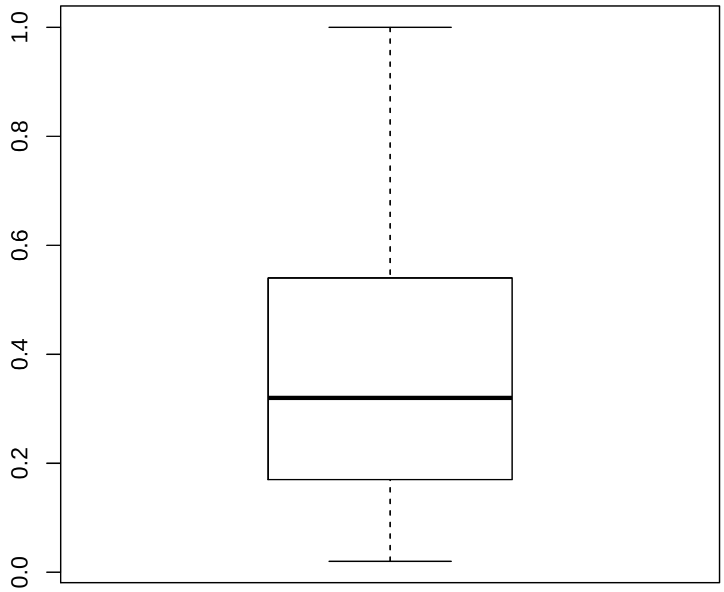
```
5                  NA                   NA                 0.00               NA
6                  NA                   NA                 0.00               NA
  ViolentCrimesPerPop
1                0.20
2                0.67
3                0.43
4                0.12
5                0.03
6                0.14

> df <- na.omit(df)
```

□

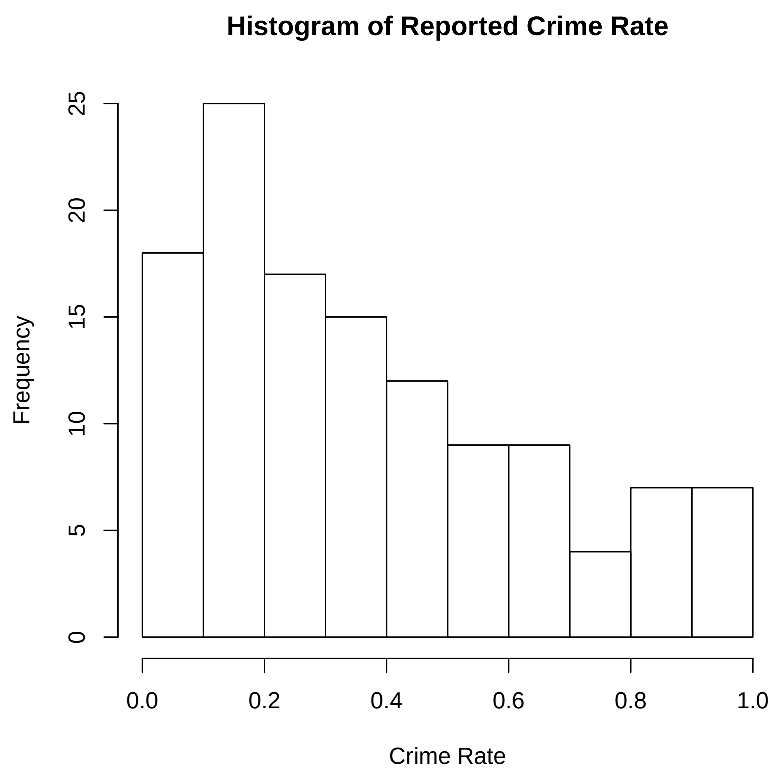
3. Create R Markdown report and use descriptive statistics to summarise data

**Solution.** Let us first look at the dependent variable of response (crime rate per pop)

```
> response <- df["ViolentCrimesPerPop"]
> summary(response)

 ViolentCrimesPerPop
 Min.   :0.0200
 1st Qu.:0.1700
 Median :0.3200
 Mean   :0.3825
 3rd Qu.:0.5400
 Max.   :1.0000

> boxplot(response)
>
```




```
> numericresp <- as.numeric(unlist(response))
> hist(numericresp, xlab ="Crime Rate", main="Histogram of Reported Crime Rate")
```




```
> dfnumeric <- subset(df, select= -c(communityname))
```

□

4. Identify the top 9 most predictive variable with respect to response (remove instances with missing values and/or categorical variables if necessary)

**Solution.** We generate a loop over the above data frame and create a correlation vector with respect to response

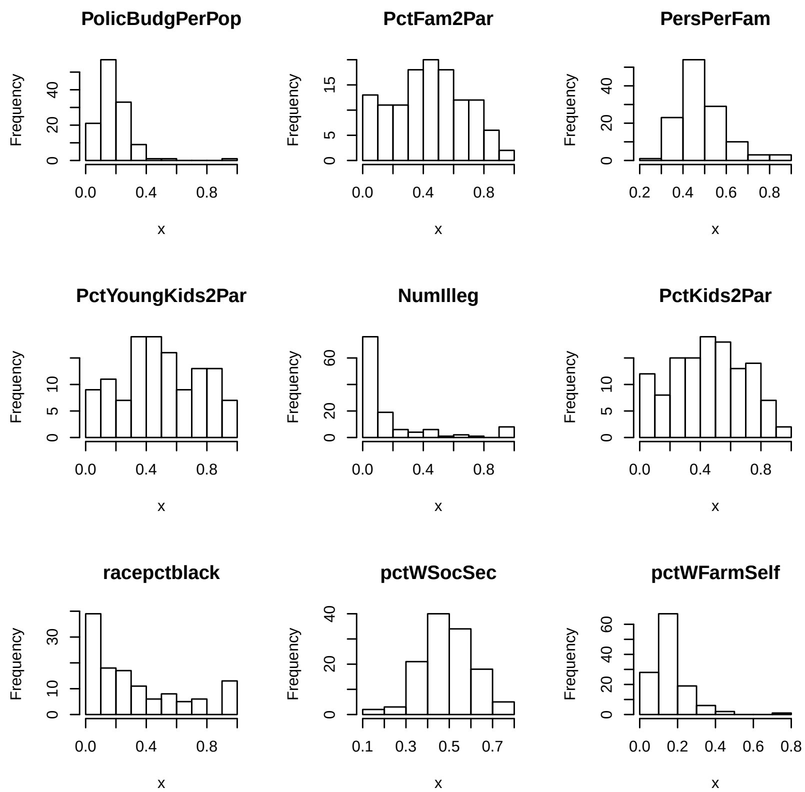
```
> correlationVector <- c()
> for(i in 1:ncol(dfnumeric)) {
+   correlationVector <- c(correlationVector, cor(dfnumeric[,i], response))
+ }
> names(correlationVector) <- colnames(dfnumeric)
> #correlationVector
> sortnames <- name[,2][order(abs(correlationVector), decreasing = TRUE)[1:9]]
> sortnames

[1] PolicBudgPerPop  PctFam2Par       PersPerFam       PctYoungKids2Par
[5] NumIlleg         PctKids2Par      racepctblack     pctWSocSec
[9] pctWFarmSelf
128 Levels: agePct12t21 agePct12t29 agePct16t24 agePct65up ... whitePerCap
```

□

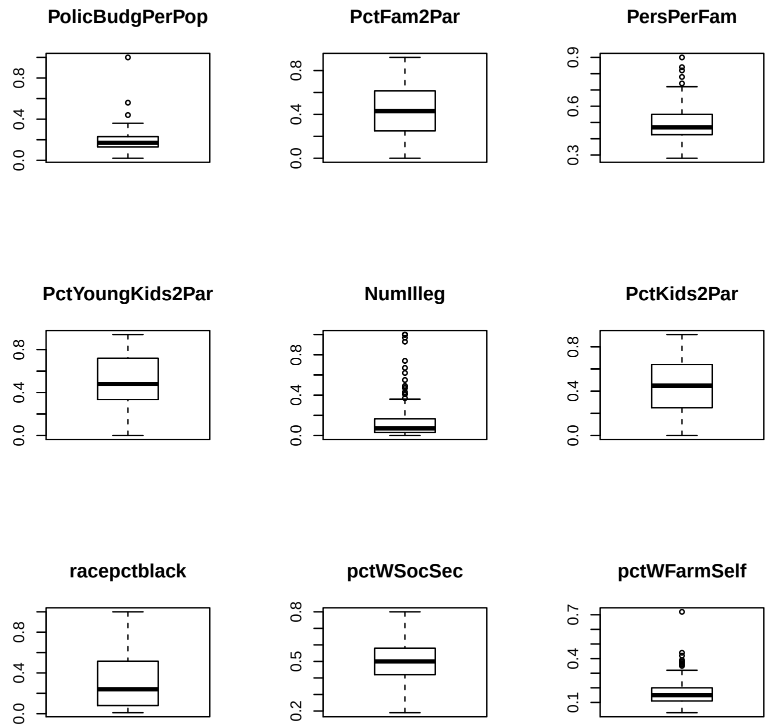
5. Generate histogram, estimate density, and boxplot for each of these predictive variables

**Solution.** We generate a new dataframe composed of the 9 highly correlated variables:

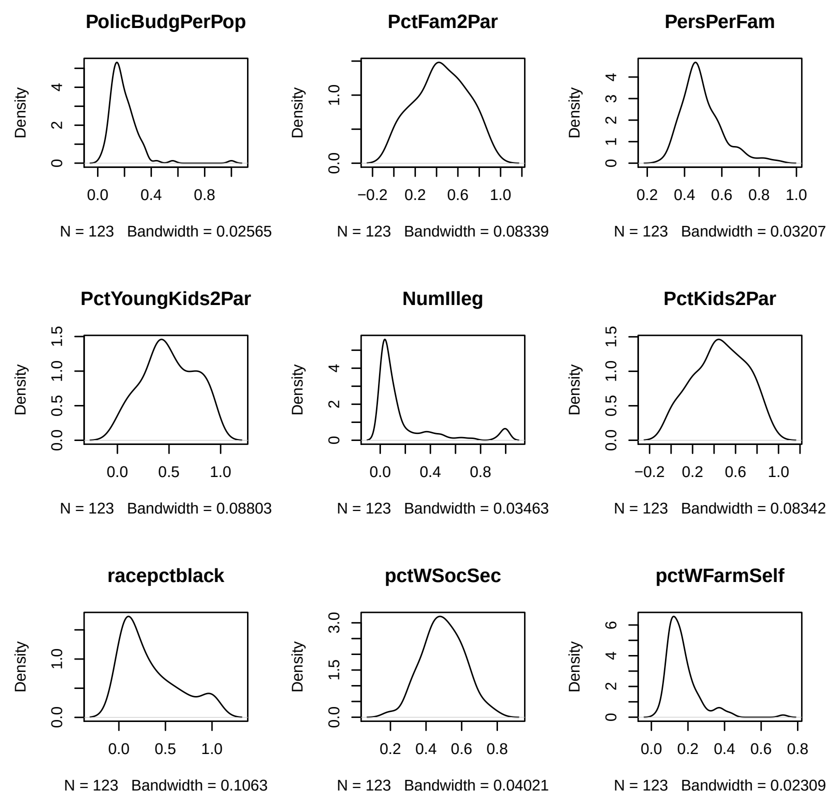
```
> df9 <- subset(dfnumeric, select = c("PolicBudgPerPop", "PctFam2Par", "PersPerFam", "PctYoun
> par(mfrow = c(3, 3))
> for (i in names(df9)) {
+   x <- df9[,i]
+   hist(x, main = i)
+ }
>
```




```
> par(mfrow = c(3, 3))
> for (i in names(df9)) {
+   x <- df9[,i]
+   boxplot(x, main = i)
+
+ }
```




```
> par(mfrow = c(3, 3))
> for (i in names(df9)) {
+   x <- density(df9[,i])
+   plot(x, main = i)
+
+ }
```



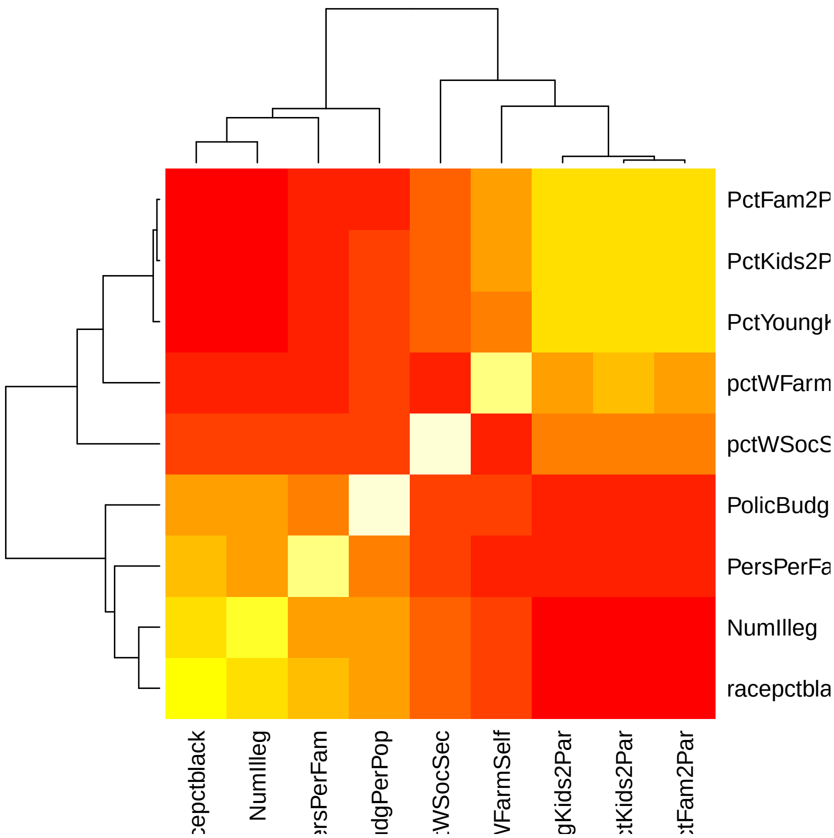

□

6. Are these highly predictive variables correlated with each other?

**Solution.**

```
> res <- cor(df9)
> round(res, 2)

                 PolicBudgPerPop PctFam2Par PersPerFam PctYoungKids2Par
PolicBudgPerPop             1.00      -0.32       0.07            -0.23
PctFam2Par                 -0.32       1.00      -0.34             0.97
PersPerFam                  0.07      -0.34       1.00            -0.35
PctYoungKids2Par           -0.23       0.97      -0.35             1.00
NumIlleg                    0.27      -0.65       0.29            -0.63
PctKids2Par                -0.31       0.99      -0.39             0.96
racepctblack                0.31      -0.75       0.44            -0.68
pctWSocSec                 -0.11       0.07      -0.15             0.05
pctWFarmSelf               -0.11       0.37      -0.29             0.35
                 NumIlleg PctKids2Par racepctblack pctWSocSec pctWFarmSelf
PolicBudgPerPop      0.27       -0.31         0.31      -0.11        -0.11
PctFam2Par          -0.65        0.99        -0.75       0.07         0.37
PersPerFam           0.29       -0.39         0.44      -0.15        -0.29
PctYoungKids2Par    -0.63        0.96        -0.68       0.05         0.35
NumIlleg             1.00       -0.65         0.67      -0.14        -0.23
```

```
PctKids2Par         -0.65       1.00       -0.78      0.09       0.38
racepctblack         0.67      -0.78        1.00     -0.20      -0.33
pctWSocSec          -0.14       0.09       -0.20      1.00      -0.25
pctWFarmSelf        -0.23       0.38       -0.33     -0.25       1.00

> heatmap(res)
```




□