Multivariate Analysis on a Census Data of district in India

VIGNESH J MURALIDHARAN

May 7, 2018

CONTENTS

1. INTRODUCTION

The purpose of this project is to analyze how people in India with the particular area are more related to different types of education and worker population with each city in a district The district name is Madras/Chennai with 7.78 million people in 270 SQ miles. Having said it is second most literate distict in the country with fifth highest employment creating district it has 35 cities as rows in the data and all the population subgroups as based on the total population in the group.

In this scenario its better to see which variable really makes cities different from each other in the particular district. Though geographical locations are not much apart, the people in each location have different sets of subgroups dependent on each other either with workculture or the education systems. All the population subgroups has been converted to proportions to make sense with the each locations.

2. EXPLANATION OF VARIABLES

1 - Name of City block

2 - Number of Households

3,4,5 - Total population persons, Male & Female

6,7,8 - Population age groups based on education (0-6),(7-13),(14-20)

9 to 14 - Caste based (Religion based) & Tribal based education systems

15 - 22 - Literates & Illiterate in Total with Male & Female

23 - Total working population

24 - 29 - Main worker Main Agriculture Main Household Main Cultivator Main Otherworking people

30 - 34 - Marginal worker Marginal Agriculture Marginal Household Marginal Cultivator Marginal otherworking people

35 - Non working population

CASTE BASED –

Though people are educated in differnt types like schools, college and medicine ..etc.. India has a special based of eduction systems for people who are interested in religious based education. For example: person who wanted to learn more about bible or quran or vedas(hindu) they have seperate education systems and they are alloted with different schools and they are considered as literate population in the whole total population subgroups.

TRIBAL BASED –

These people are having tribal based education systems like forest studies, wild animal activities etc.. but since this is a city we have very less tribal population considred in this district.

Main & Marginal –

Main worker are population who are working a job as there main profession Marginal is something like a part-time job

3. TOTAL COMPARITIVE ANALYSIS

A. PRINCIPLE COMPONENT ANALYSIS

The PCA for the total population has been done checking what variables are really making difference in the analysis the PC1 explains almost 62% of the needed analysis of the data and rest 19% from the PC2

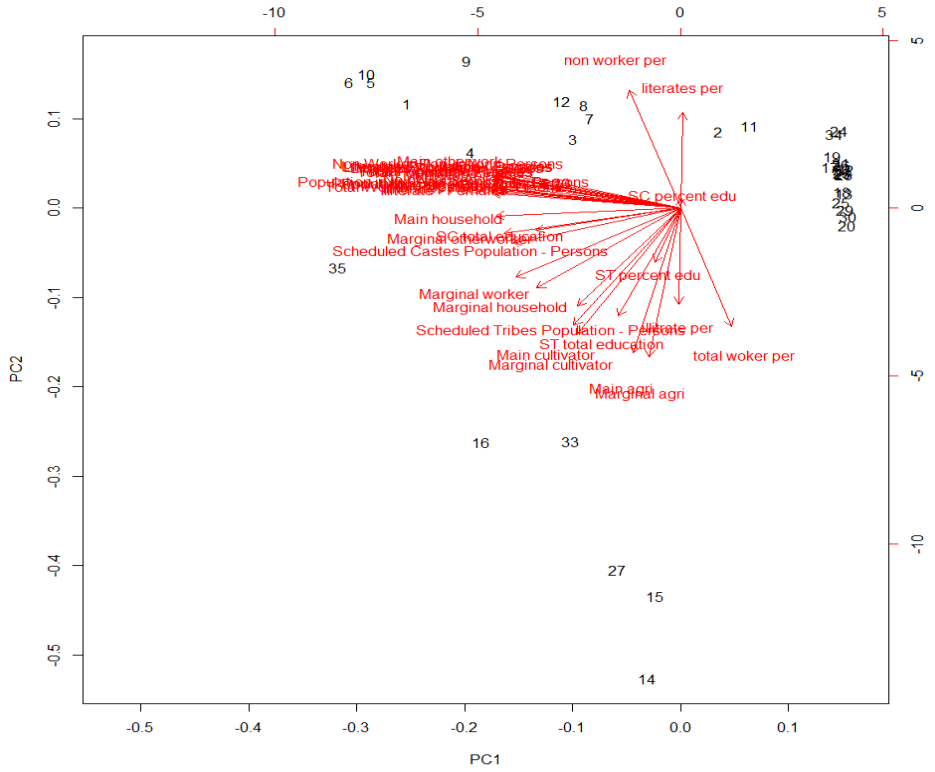
```
library(tidyverse)
## Warning: package 'tidyverse' was built under R version 3.4.4
## -- Attaching packages ----------------------------------------------------------- ti
dyverse 1.2.1 --
## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.4.2     v dplyr   0.7.4
## v tidyr   0.8.0     v stringr 1.3.0
## v readr   1.1.1     v forcats 0.3.0
## Warning: package 'ggplot2' was built under R version 3.4.4
## Warning: package 'tibble' was built under R version 3.4.4
## Warning: package 'tidyr' was built under R version 3.4.4
## Warning: package 'readr' was built under R version 3.4.4
## Warning: package 'purrr' was built under R version 3.4.4
## Warning: package 'dplyr' was built under R version 3.4.3
## Warning: package 'stringr' was built under R version 3.4.4
## Warning: package 'forcats' was built under R version 3.4.4
## -- Conflicts -------------------------------------------------------------- tidyvers
e_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
library(googlesheets)
## Warning: package 'googlesheets' was built under R version 3.4.4
data.url = "https://docs.google.com/spreadsheets/d/1zbVfT_St4rlKpD-yXYMbcnbhqV69SJon-sGip
iQoQss/edit#gid=2138857311"

#my_sheets = gs_ls()
data = data.url %>%
  gs_url() %>%
  gs_read()
## Sheet-identifying info appears to be a browser URL.
## googlesheets will attempt to extract sheet key from the URL.
## Putative key: 1zbVfT_St4rlKpD-yXYMbcnbhqV69SJon-sGipiQoQss
## Sheet successfully identified: "editpurpose.xlsx"
## Accessing worksheet titled 'Sheet1'.
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   Name = col_character(),
##   `SC percent edu` = col_double(),
##   `ST percent edu` = col_double(),
##   `literates per` = col_double(),
##   `illitrate per` = col_double(),
##   `total woker per` = col_double(),
##   `non worker per` = col_double()
## )
## See spec(...) for full column specifications.
editpurpose=data
```

```r
attach(editpurpose)
row.names(editpurpose)<-editpurpose$Name
## Warning: Setting row names on a tibble is deprecated.
census.pca=prcomp(editpurpose[,-1],scale=TRUE)
summary(census.pca)
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6
## Standard deviation     4.6676 2.5837 1.35959 1.22345 0.93183 0.84106
## Proportion of Variance 0.6225 0.1907 0.05281 0.04277 0.02481 0.02021
## Cumulative Proportion  0.6225 0.8132 0.86600 0.90877 0.93358 0.95379
##                           PC7     PC8     PC9    PC10    PC11    PC12
## Standard deviation     0.68816 0.57765 0.45627 0.39644 0.37296 0.31414
## Proportion of Variance 0.01353 0.00953 0.00595 0.00449 0.00397 0.00282
## Cumulative Proportion  0.96732 0.97685 0.98280 0.98729 0.99126 0.99408
##                           PC13   PC14    PC15    PC16    PC17    PC18
## Standard deviation     0.27525 0.2214 0.18012 0.17428 0.10989 0.06389
## Proportion of Variance 0.00216 0.0014 0.00093 0.00087 0.00035 0.00012
## Cumulative Proportion  0.99625 0.9977 0.99858 0.99944 0.99979 0.99990
##                           PC19    PC20    PC21     PC22     PC23     PC24
## Standard deviation     0.05263 0.01769 0.01015 0.009603 0.006783 0.003331
## Proportion of Variance 0.00008 0.00001 0.00000 0.000000 0.000000 0.000000
## Cumulative Proportion  0.99998 0.99999 1.00000 1.000000 1.000000 1.000000
##                             PC25      PC26      PC27      PC28      PC29
## Standard deviation     1.451e-15 4.484e-16 4.484e-16 4.484e-16 4.484e-16
## Proportion of Variance 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
## Cumulative Proportion  1.000e+00 1.000e+00 1.000e+00 1.000e+00 1.000e+00
##                             PC30      PC31      PC32      PC33      PC34
## Standard deviation     4.484e-16 4.484e-16 4.484e-16 4.484e-16 4.484e-16
## Proportion of Variance 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
## Cumulative Proportion  1.000e+00 1.000e+00 1.000e+00 1.000e+00 1.000e+00
##                             PC35
## Standard deviation     2.312e-16
## Proportion of Variance 0.000e+00
## Cumulative Proportion  1.000e+00
```

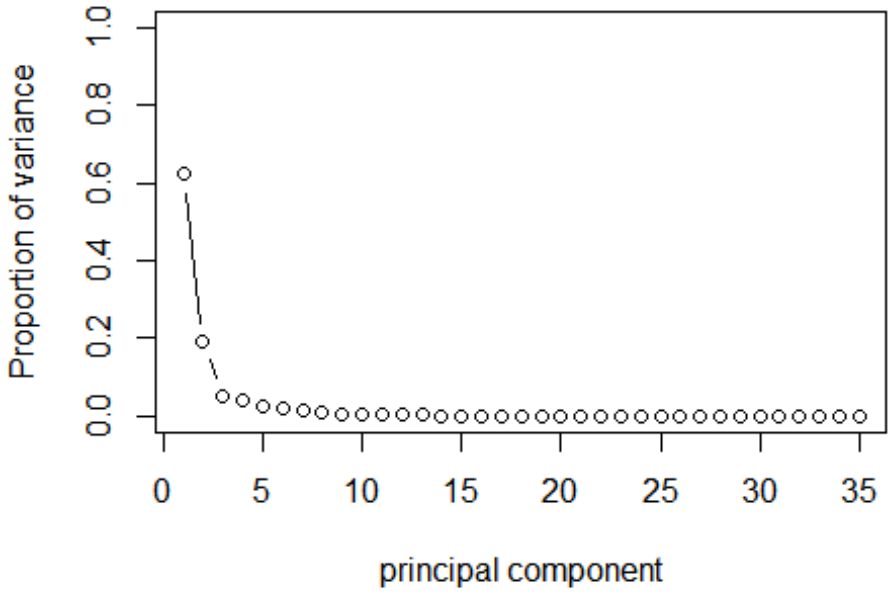
after doing PCA its better to look over in the biplot graph

```r
#standard deviation of each component
census.sd=census.pca$sdev
census.var=census.pca$sdev^2
census.var
##  [1] 2.178626e+01 6.675298e+00 1.848495e+00 1.496818e+00 8.683016e-01
##  [6] 7.073749e-01 4.735641e-01 3.336814e-01 2.081787e-01 1.571663e-01
## [11] 1.391022e-01 9.868487e-02 7.576372e-02 4.900523e-02 3.244487e-02
## [16] 3.037338e-02 1.207615e-02 4.081445e-03 2.769999e-03 3.128263e-04
## [21] 1.030722e-04 9.221207e-05 4.600778e-05 1.109462e-05 2.104541e-30
## [26] 2.010770e-31 2.010770e-31 2.010770e-31 2.010770e-31 2.010770e-31
## [31] 2.010770e-31 2.010770e-31 2.010770e-31 2.010770e-31 5.343808e-32
#proportion of variance explained
pve=census.var/sum(census.var)
pve
##  [1] 6.224644e-01 1.907228e-01 5.281414e-02 4.276623e-02 2.480862e-02
##  [6] 2.021071e-02 1.353040e-02 9.533754e-03 5.947963e-03 4.490465e-03
## [11] 3.974349e-03 2.819568e-03 2.164678e-03 1.400149e-03 9.269962e-04
## [16] 8.678108e-04 3.450328e-04 1.166127e-04 7.914282e-05 8.937895e-06
## [21] 2.944919e-06 2.634631e-06 1.314508e-06 3.169891e-07 6.012975e-32
```

```
## [26] 5.745058e-33 5.745058e-33 5.745058e-33 5.745058e-33 5.745058e-33
## [31] 5.745058e-33 5.745058e-33 5.745058e-33 5.745058e-33 1.526802e-33
#biplot
biplot(census.pca)
```




```
#proportion of variance explained
plot(pve,xlab="principal component",ylab="Proportion of variance"
     , ylim=c(0,1), type='b')
```



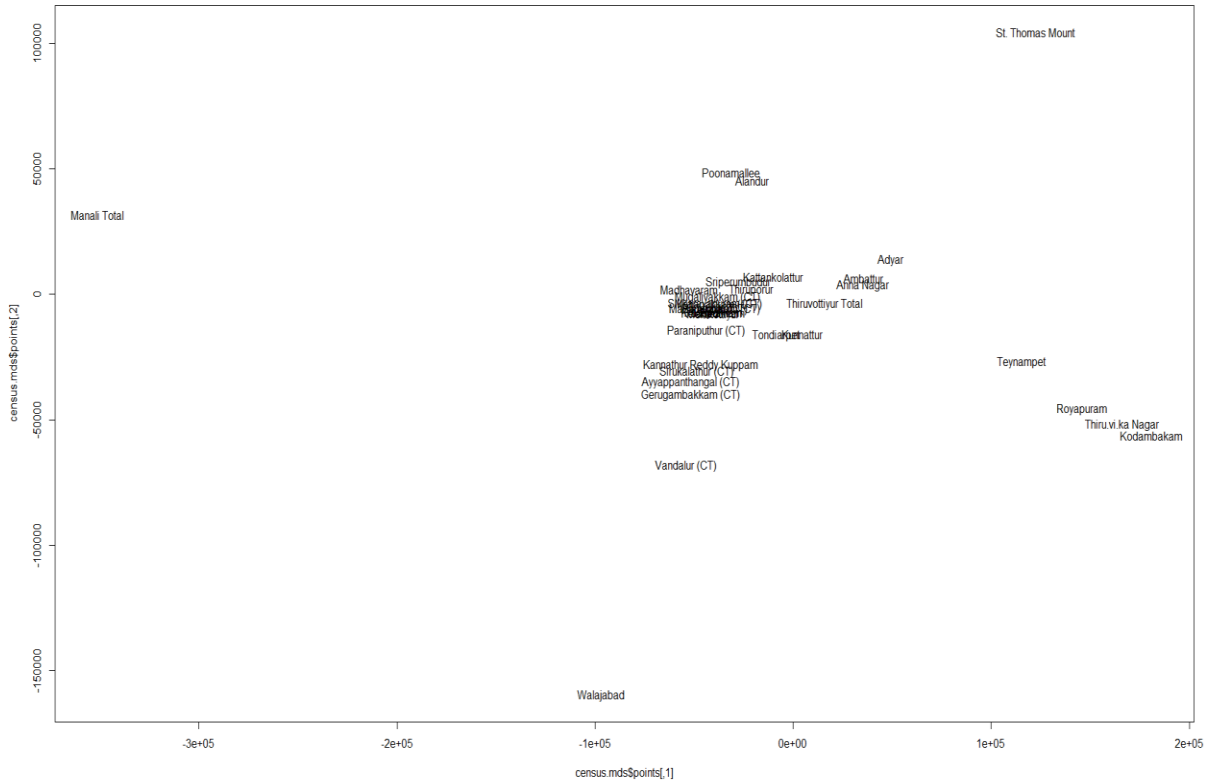
The principle component graph condenses to total of 2 PCA's and while analyzing the biplot a condensed cluster is formed making it few locations away from most of the variables with both worker and the education people. but the locations like St.Thomas mount and walajabad makes scattered locations in the biolopt which makes interest in see what really makes these locations different. While considering difference the most scattered locations are taken and the most clustered locations are taken to see what variables make even these locations different within each groups.

B. MULTIDIMENTIONAL SCALLING

Here the same analysis as PCA was checked to see if the locations are scattered in the same way PCA has analyzed the data.

```
########################
#Total - multidimentinal scalling
########################
library(MASS)
## Warning: package 'MASS' was built under R version 3.4.4
##
## Attaching package: 'MASS'
## The following object is masked from 'package:dplyr':
##
##     select
census<-as.matrix(editpurpose[,-1])
census.mds<-cmdscale(census,k=2,eig=TRUE)
census.mds$points
##               [,1]        [,2]
##  [1,]   16060.163    3681.996
##  [2,] -351215.389  -31730.527
##  [3,]  -52285.745   -2184.344
##  [4,]   -8261.386   16188.362
##  [5,]  145976.562   45711.068
##  [6,]  166190.634   51845.483
##  [7,]   35659.598   -6572.301
##  [8,]   35299.869   -3716.880
##  [9,]  115509.891   26840.259
## [10,]  181097.087   56337.796
## [11,]  -20490.817  -45506.751
## [12,]   49498.471  -13781.113
## [13,]  -31349.609  -48793.969
## [14,]  -96808.923  159789.032
## [15,]  -27658.160   -4872.438
## [16,]    4678.821   15639.141
## [17,]  -51768.607   35046.568
## [18,]  -43631.830   14338.336
## [19,]  -37928.248    1141.920
## [20,]  -39246.662    3724.654
## [21,]  -40025.049    5738.212
## [22,]  -51621.889   40009.421
## [23,]  -39137.172    4087.918
## [24,]  -48457.119   30836.662
## [25,]  -39687.782    5759.671
## [26,]  -39734.081    5789.143
## [27,]  -21078.347   -1916.546
## [28,]  -46630.312   28095.246
## [29,]  -40381.853    7863.671
```

```
## [30,]  -40203.657    7243.698
## [31,]  -40218.520    7252.095
## [32,]  -40222.252    7251.621
## [33,]  -10065.058   -7110.452
## [34,]  -54157.953   68215.639
## [35,]  122559.409 -104497.481
plot(census.mds$points,type="n",col='black')
#text(census.mds$points,rownames(census.mds$points),col=rainbow(18))
text(census.mds$points,rownames(editpurpose))
```



Here in MDS as in PCA the locations are splitted according to the same analysis and this proves that both techniqes helps us to visualize the same thing. So here the most scattered locations are taken analyzed with 2 different groups.

This census data has Education based groups and worker based groups. so, we would first pick the locations and analyzed based on the education first and then with worker in both MDS, PCA and Chi-square analysis.

4. EDUCATION - LOCATION BASED ANALYSIS

Most scattered locations 1. THIRVOTIUR 2. MANALI 3. MADAVARAM 4. TEYNAMPET 5. KODAMBAKAM 6. ADAYAR 7. POONAMALI 8. WALAJABAD 9. ST.THOMAS MOUNT 10. SRIPERAMBATUR Most clustered locations 1. ROYAPURAM 2. THIRUVIKANAGAR 3. AMBATHUR 4. ANNA NAGAR 5. ALANDUR 6. AYYAPANTHANGAL 7. KELAMBAKAM 8. KATTANGALATURE 9. VANDALUR 10 . TONDIRAPET

```
#functionforchisquare
chisqD <- function(x) {
  r <- nrow(x)
  c<- ncol(x)
  row.sums <- apply(x,1,sum)
  col.sums <- apply(x,2,sum)
  N<-sum(row.sums)
  pijrow <- matrix(0,nrow=r,ncol=c)
  pijcol <- matrix(0,nrow=r,ncol=c)
  distx.row <- matrix(0,nrow=r,ncol=r)
  distx.col <- matrix(0,nrow=c,ncol=c)
  for (i in 1:r){
    pijrow[i,] <- x[i,]/row.sums[i]
  }
  for (j in 1:c){
    pijcol[,j] <- x[,j]/col.sums[j]
  }

  for (i in 1:r){
    for (ii in 1:(i-1)) {
      d.row<- sum( (N/col.sums)*(pijrow[i,]-pijrow[ii,])^2 )
      distx.row[i,ii] <- d.row
      distx.row[ii,i] <- d.row
    }
  }

  for (j in 1:c){
    for (jj in 1:(j-1)) {
      d.col <- sum( (N/row.sums)*(pijcol[,j]-pijcol[,jj])^2 )
      distx.col[j,jj] <- d.col
      distx.col[jj,j] <- d.col
    }
  }
  return((list(dist.row=distx.row, dist.col=distx.col)))
}
```

4.1.A. EDUCATION ANALYSIS ON MOST SCATTERED AREAS IN CHI-SQUARE ANALYSIS

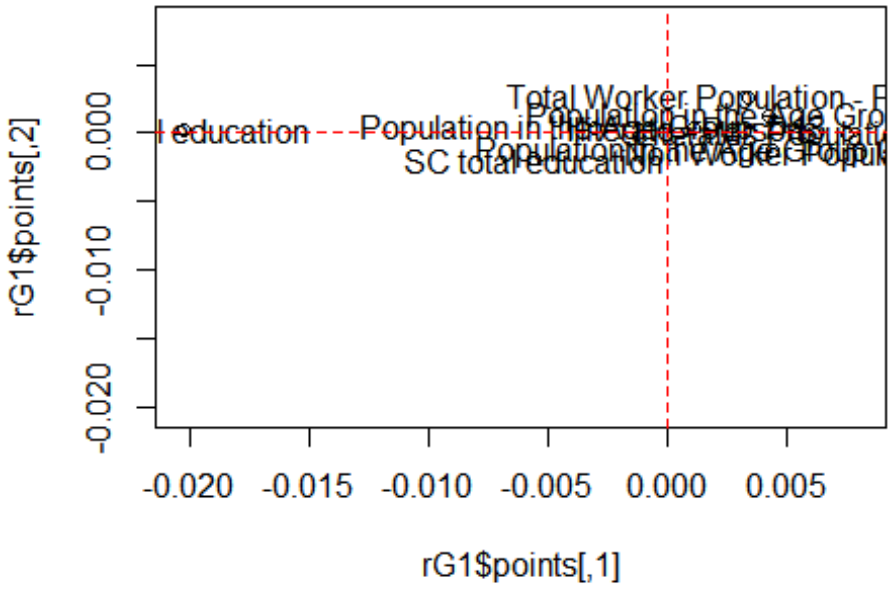
In the most scattered areas the tribal education system is very less which makes sense for sure because its a city based census. But comparitively the literates population is very high and at the same time non working population is also very high. Unfortunatelly though the illiterate population is very less the number of working populatin is very high is understable because even if people are not studying they can work something in their life. Here in India caste based education is considered as the population who are literates so these kind of population is higher in there areas .

```
data.url = "https://docs.google.com/spreadsheets/d/1mF1QWTnOEmMugQ_u30rY-4AAFkoLOgGRbie1AeHWHwI/edit#gid=920190064"
```

```
#my_sheets = gs_ls()
data = data.url %>%
  gs_url() %>%
  gs_read()
## Sheet-identifying info appears to be a browser URL.
## googlesheets will attempt to extract sheet key from the URL.
## Putative key: 1mF1QWTnOEmMugQ_u30rY-4AAFkoLOgGRbie1AeHWHwI
## Sheet successfully identified: "chisquare1to9.xlsx"
## Accessing worksheet titled 'Sheet1'.
## Parsed with column specification:
## cols(
##   Name = col_character(),
##   `Population in the Age Group 0-6 - Persons` = col_double(),
##   `Population in the Age Group 7-13` = col_double(),
##   `Population in the Age Group 14-20` = col_double(),
##   `SC total education` = col_double(),
##   `ST total education` = col_double(),
##   `Literates Population - Persons` = col_double(),
##   `Illiterate - Persons` = col_double(),
##   `Total Worker Population - Persons` = col_double(),
##   `Non Worker Population - Persons` = col_double()
## )
chisquare1to9=data
chisquare1to9
## # A tibble: 9 x 10
##   Name     `Population in the Ag~ `Population in the ~ `Population in the~
##   <chr>                     <dbl>                <dbl>                <dbl>
## 1 Thiruvo~                  11.9                 12.0                 11.8
## 2 Manali ~                  10.1                  9.98                 8.12
## 3 Madhava~                  21.1                 10.5                  7.82
## 4 Teynamp~                   8.76                 4.39                 7.21
## 5 Kodamba~                   9.99                 8.16                 6.59
## 6 Adyar                      9.80                 6.34                 4.83
## 7 Poonama~                  12.3                 20.3                  3.06
## 8 Walajab~                  11.0                 14.2                 10.4
## 9 St. Tho~                  11.6                  5.89                 9.64
## # ... with 6 more variables: `SC total education` <dbl>, `ST total
## #   education` <dbl>, `Literates Population - Persons` <dbl>, `Illiterate
## #   - Persons` <dbl>, `Total Worker Population - Persons` <dbl>, `Non
## #   Worker Population - Persons` <dbl>
attach(chisquare1to9)
## The following objects are masked from editpurpose:
##
##     Illiterate - Persons, Literates Population - Persons, Name,
##     Non Worker Population - Persons, Population in the Age Group
##     0-6 - Persons, Population in the Age Group 14-20, Population
##     in the Age Group 7-13, SC total education, ST total education,
##     Total Worker Population - Persons
dim(chisquare1to9)
## [1]  9 10
cities<-as.matrix(chisquare1to9[,c(2:10)])
t(cities)
##                                                    [,1]        [,2]
## Population in the Age Group 0-6 - Persons 11.89313974 10.09590862
```

```
## Population in the Age Group 7-13          11.95051377  9.98383563
## Population in the Age Group 14-20         11.79284812  8.11738927
## SC total education                         3.30194326  1.47634613
## ST total education                         0.01852233  0.01580517
## Literates Population - Persons            70.24501422 73.36829628
## Illiterate - Persons                      29.75532184 26.63170372
## Total Worker Population - Persons         33.18334598 63.09637559
## Non Worker Population - Persons           66.81702879 36.10000000
##                                                 [,3]        [,4]
## Population in the Age Group 0-6 - Persons 21.10000000  8.759087271
## Population in the Age Group 7-13          10.48956117  4.389741116
## Population in the Age Group 14-20          7.82377392  7.214114190
## SC total education                         2.97056255  1.070625774
## ST total education                         0.01594544  0.003947412
## Literates Population - Persons            71.18007110 80.585532740
## Illiterate - Persons                      28.81992890 19.414467260
## Total Worker Population - Persons         31.82818066 35.755874520
## Non Worker Population - Persons           68.17181934 64.244125480
##                                                 [,5]      [,6]       [,7]
## Population in the Age Group 0-6 - Persons  9.98736575  9.797215 12.3148148
## Population in the Age Group 7-13           8.16378025  6.335451 20.3472222
## Population in the Age Group 14-20          6.59197860  4.828974  3.0555556
## SC total education                         1.21998915  0.000000  6.8518519
## ST total education                         0.01153484  0.000000  0.1388889
## Literates Population - Persons            78.63009403 77.960801 70.7638889
## Illiterate - Persons                      21.36990597 22.039199 27.1527778
## Total Worker Population - Persons         35.92326450 35.948869 35.2083333
## Non Worker Population - Persons           64.07673550 64.051131 62.7083333
##                                                 [,8]        [,9]
## Population in the Age Group 0-6 - Persons 11.0089051 11.57638210
## Population in the Age Group 7-13          14.2439082  5.89438965
## Population in the Age Group 14-20         10.3699542  9.64211809
## SC total education                         5.5048718  3.90508304
## ST total education                         0.4075198  0.05126104
## Literates Population - Persons            65.7943283 77.61615143
## Illiterate - Persons                      34.2056717 22.38384857
## Total Worker Population - Persons         46.9838501 40.39391170
## Non Worker Population - Persons           53.0161499 59.60608830
G<-t(cities)%*%cities
chisqD(G)$dist.col
##              [,1]        [,2]         [,3]        [,4]        [,5]
##  [1,] 0.0000000000 0.002475020 0.0012238075 0.005699845 0.024278472
##  [2,] 0.0024750205 0.000000000 0.0040577654 0.001105904 0.012705381
##  [3,] 0.0012238075 0.004057765 0.0000000000 0.007878504 0.023913200
##  [4,] 0.0056998447 0.001105904 0.0078785043 0.000000000 0.008379075
##  [5,] 0.0242784722 0.012705381 0.0239132002 0.008379075 0.000000000
##  [6,] 0.0008460577 0.004067204 0.0007217291 0.008596657 0.027772158
##  [7,] 0.0005530421 0.001608638 0.0006350500 0.004696795 0.019823523
##  [8,] 0.0023274524 0.003925533 0.0008884812 0.008359167 0.022845051
##  [9,] 0.0004685326 0.003843170 0.0013276837 0.007803714 0.028391372
##              [,6]         [,7]         [,8]         [,9]
##  [1,] 0.0008460577 0.0005530421 0.0023274524 0.0004685326
##  [2,] 0.0040672035 0.0016086384 0.0039255328 0.0038431696
##  [3,] 0.0007217291 0.0006350500 0.0008884812 0.0013276837
##  [4,] 0.0085966570 0.0046967949 0.0083591670 0.0078037139
```

```
##  [5,] 0.0277721583 0.0198235227 0.0228450514 0.0283913720
##  [6,] 0.0000000000 0.0007622039 0.0010487246 0.0003785120
##  [7,] 0.0007622039 0.0000000000 0.0009594504 0.0010517264
##  [8,] 0.0010487246 0.0009594504 0.0000000000 0.0024817638
##  [9,] 0.0003785120 0.0010517264 0.0024817638 0.0000000000
rG1 <- cmdscale(chisqD(G)$dist.col, eig = TRUE)
cG1 <- cmdscale(chisqD(G)$dist.row, eig = TRUE)
plot(rG1$points, xlim = range(rG1$points[,1], cG1$points[,1]) ,
     ylim = range(rG1$points[,1], cG1$points[,1]))
text(rG1$points, labels = colnames(G), cex = 1)
abline(h = 0, lty = 2,col='red')
abline(v = 0, lty = 2,col='red')
```

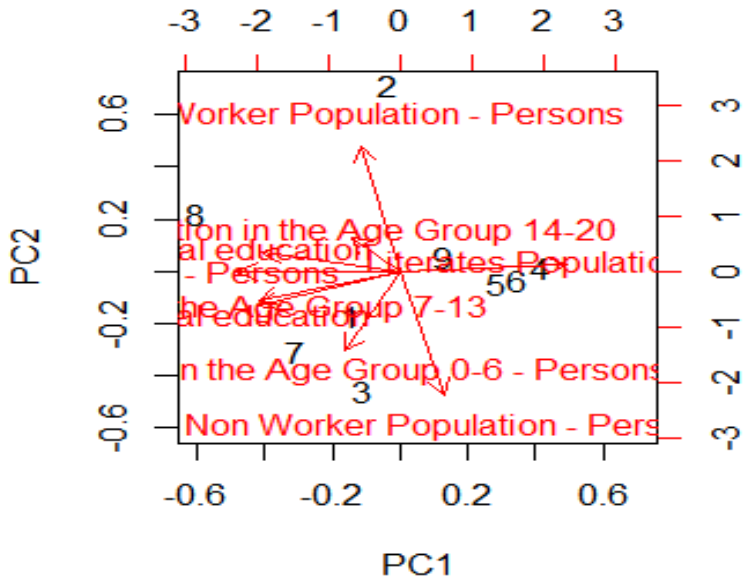


4.1.B. EDUCATION ANALYSIS ON MOST SCATTERED AREAS IN PRINCIPLE COMPONENT ANALYSIS

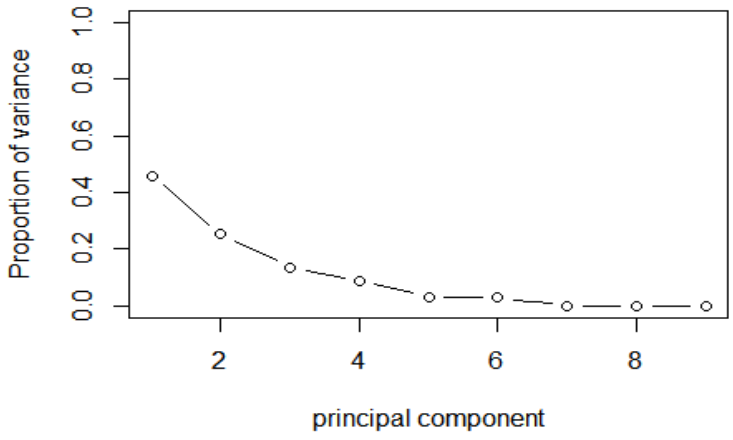
In the PCA here we can clearly see that chi-square method proves us the same for the literates population is higher in areas like st.Thomas mount , Teynampet, kodambakam and Adyar. rather than the caste based education is seems higher in poonamali and population age group of (0to6) is higher in madavaram while thirvotur seems like having all the variables in common.

```
row.names(chisquare1to9)<-chisquare1to9$Name
## Warning: Setting row names on a tibble is deprecated.
census1to9.pca=prcomp(chisquare1to9[,-1],scale=TRUE)
summary(census1to9.pca)
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6
## Standard deviation     2.0326 1.5091 1.1083 0.89478 0.53659 0.51364
## Proportion of Variance 0.4591 0.2530 0.1365 0.08896 0.03199 0.02931
## Cumulative Proportion  0.4591 0.7121 0.8486 0.93753 0.96952 0.99884
##                            PC7      PC8      PC9
## Standard deviation     0.10191 0.009806 1.42e-16
```

```
## Proportion of Variance 0.00115 0.000010 0.00e+00
## Cumulative Proportion  0.99999 1.000000 1.00e+00
#Standard deviation of each component
census1to9.sd=census1to9.pca$sdev
census1to9.var=census1to9.pca$sdev^2
census1to9.var
## [1] 4.131408e+00 2.277327e+00 1.228398e+00 8.006351e-01 2.879258e-01
## [6] 2.638233e-01 1.038612e-02 9.615149e-05 2.015825e-32
#proportion of variance explained
pve=census1to9.var/sum(census1to9.var)
#biplot
biplot(census1to9.pca)
```




```
#proportion of variance explained
plot(pve,xlab="principal component",ylab="Proportion of variance" , ylim=c(0,1), type='b'
)
```



4.1.C. EDUCATION ANALYSIS ON MOST SCATTERED AREAS IN MULTIDIMENTIONAL SCALLING

Areas in MDS proves the same scattered locations in PCA has the same adjustments in the MDS scale. While manali (2) is totally different with respect to total worker population but in MDS scale it seems to share a bit with other nearby locations like teynampet(4) and kodambakam(5).

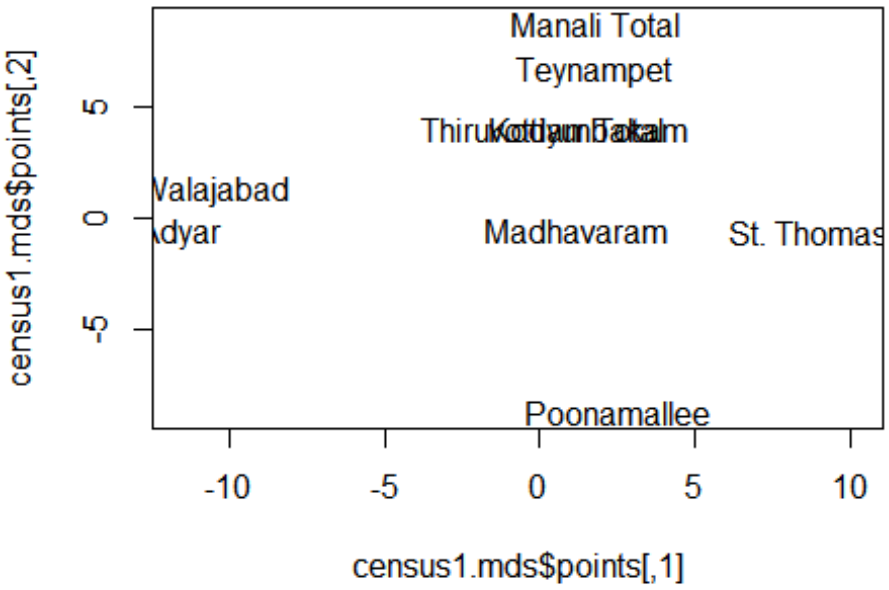
```
census1<-as.matrix(chisquare1to9[,-1])
census1.mds<-cmdscale(census1,k=2,eig=TRUE)
census1.mds$points
```

```
##             [,1]       [,2]
##  [1,]   0.128444  3.9058399
##  [2,]   1.787125  8.7722862
##  [3,]   1.249892 -0.5061966
##  [4,]   1.796214  6.6732891
##  [5,]   1.639583  4.0837996
##  [6,] -11.589253 -0.6105775
##  [7,]   2.540093 -8.6852253
##  [8,] -10.437213  1.3278652
##  [9,]  10.209571 -0.5273670
```

```
row.names(chisquare1to9)<-chisquare1to9$Name
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```
plot(census1.mds$points,type="n",col='red')
#text(census.mds$points,rownames(census.mds$points),col=rainbow(18))
text(census1.mds$points,rownames(chisquare1to9))
```



4.2.A EDUCATION ANALYSIS ON CLOSELY SCATTERED AREAS IN CHI-SQUARE ANALYSIS

Here the caste based education is most higher compared to the normal literate population in the cities. Which means that caste based population is also more here who are insisting the yonger generations on these types of education systems in the society. At the same time you can also see the comparison of the population education age group of (7-13) is very less in the total population which clearly shows that entry level school education is completely not insisted in the society here. Otherwise all other population groups occur in similar way for all locations.

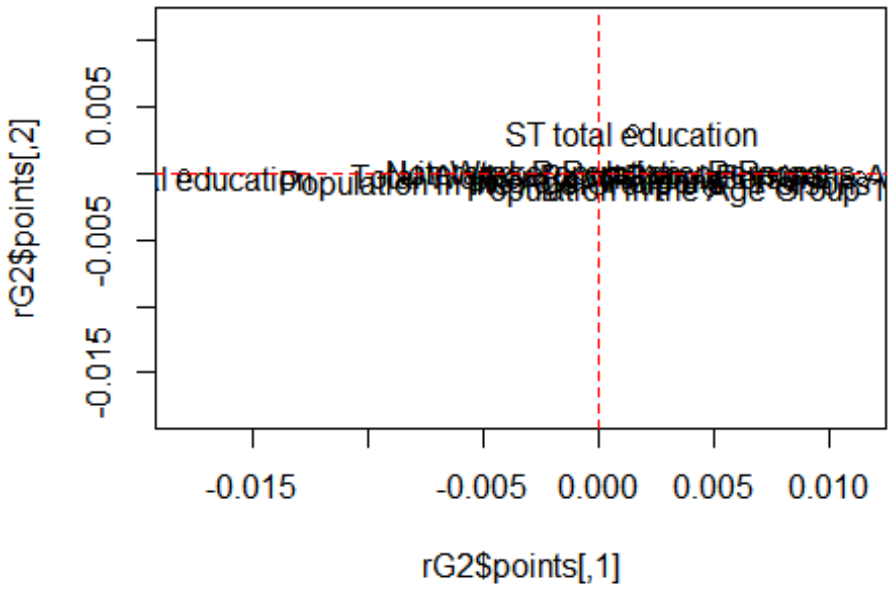
```
data.url = "https://docs.google.com/spreadsheets/d/1awOEtfoFg7r7tVMx-gc5G5xez1z107gvfzG1t7bY-t8/edit#gid=1676744121"

#my_sheets = gs_ls()
data = data.url %>%
  gs_url() %>%
  gs_read()
## Sheet-identifying info appears to be a browser URL.
## googlesheets will attempt to extract sheet key from the URL.
## Putative key: 1awOEtfoFg7r7tVMx-gc5G5xez1z107gvfzG1t7bY-t8
## Sheet successfully identified: "chisquaresecond9.xlsx"
## Accessing worksheet titled 'Sheet1'.
## Parsed with column specification:
## cols(
##   Name = col_character(),
##   `Population in the Age Group 0-6 - Persons` = col_double(),
##   `Population in the Age Group 7-13` = col_double(),
##   `Population in the Age Group 14-20` = col_double(),
##   `SC total education` = col_double(),
##   `ST total education` = col_double(),
##   `Literates Population - Persons` = col_double(),
##   `Illiterate - Persons` = col_double(),
##   `Total Worker Population - Persons` = col_double(),
##   `Non Worker Population - Persons` = col_double()
## )
chisquaresecond9=data
chisquaresecond9
## # A tibble: 9 x 10
##   Name      `Population in the Ag~ `Population in the~ `Population in the~
##   <chr>                      <dbl>               <dbl>               <dbl>
## 1 Royapuram                   9.97                9.81                9.75
## 2 Thiru.vi~                   9.76               10.5                 9.45
## 3 Ambattur                    8.75                6.79                5.23
## 4 Anna Nag~                   9.18                6.93                6.90
## 5 Alandur                     8.30                4.19                6.70
## 6 Ayyappan~                  11.8                 6.30                8.25
## 7 Kelambak~                  11.3                 0.713               0.867
## 8 Kattanko~                  11.4                 5.83                5.60
## 9 Vandalur~                  10.9                11.8                 5.38
## # ... with 6 more variables: `SC total education` <dbl>, `ST total
## #   education` <dbl>, `Literates Population - Persons` <dbl>, `Illiterate
## #   - Persons` <dbl>, `Total Worker Population - Persons` <dbl>, `Non
## #   Worker Population - Persons` <dbl>
attach(chisquaresecond9)
```

```
## The following objects are masked from chisquare1to9:
##
##     Illiterate - Persons, Literates Population - Persons, Name,
##     Non Worker Population - Persons, Population in the Age Group
##     0-6 - Persons, Population in the Age Group 14-20, Population
##     in the Age Group 7-13, SC total education, ST total education,
##     Total Worker Population - Persons
## The following objects are masked from editpurpose:
##
##     Illiterate - Persons, Literates Population - Persons, Name,
##     Non Worker Population - Persons, Population in the Age Group
##     0-6 - Persons, Population in the Age Group 14-20, Population
##     in the Age Group 7-13, SC total education, ST total education,
##     Total Worker Population - Persons
dim(chisquaresecond9)
## [1]  9 10
cities1<-as.matrix(chisquaresecond9[,c(2:10)])
dim(cities1)
## [1] 9 9
t(cities1)
##                                              [,1]        [,2]
## Population in the Age Group 0-6 - Persons  9.97379739  9.76127585
## Population in the Age Group 7-13           9.81337721 10.51164661
## Population in the Age Group 14-20          9.74518397  9.45120377
## SC total education                         1.40098988  2.59106638
## ST total education                         0.01055139  0.01715449
## Literates Population - Persons            77.99156280 79.79458877
## Illiterate - Persons                      22.00843720 20.20541123
## Total Worker Population - Persons         32.77300251 35.09790973
## Non Worker Population - Persons           67.22699749 64.90209027
##                                              [,3]        [,4]
## Population in the Age Group 0-6 - Persons  8.75367502  9.18104484
## Population in the Age Group 7-13           6.78575727  6.93131311
## Population in the Age Group 14-20          5.23184843  6.89509263
## SC total education                         3.59781992  3.37967523
## ST total education                         0.08527758  0.03418942
## Literates Population - Persons            77.80616543 79.19996750
## Illiterate - Persons                      22.19383457 20.80003250
## Total Worker Population - Persons         34.13269604 35.42768937
## Non Worker Population - Persons           65.86730396 64.57231063
##                                              [,5]      [,6]
## Population in the Age Group 0-6 - Persons  8.29793425 11.832157
## Population in the Age Group 7-13           4.18637516  6.300403
## Population in the Age Group 14-20          6.69687019  8.245128
## SC total education                         1.85294484 14.986559
## ST total education                         0.01330064  0.000000
## Literates Population - Persons            79.17286670 76.499496
## Illiterate - Persons                      20.82713330 23.500504
## Total Worker Population - Persons         36.30242321 38.726478
## Non Worker Population - Persons           63.69757679 61.273522
##                                              [,7]         [,8]
## Population in the Age Group 0-6 - Persons 11.33166313 11.431949720
## Population in the Age Group 7-13           0.71304683  5.834454300
## Population in the Age Group 14-20          0.86721912  5.597495416
## SC total education                         9.01907882  2.618036640
```

```
## ST total education                          0.03854307  0.002872229
## Literates Population - Persons             80.05396030 74.704758800
## Illiterate - Persons                       19.94603970 25.295241200
## Total Worker Population - Persons          38.19618424 40.741130800
## Non Worker Population - Persons            61.80381576 59.258869200
##                                                    [,9]
## Population in the Age Group 0-6 - Persons 10.906717
## Population in the Age Group 7-13          11.784951
## Population in the Age Group 14-20          5.376216
## SC total education                         1.471635
## ST total education                         0.000000
## Literates Population - Persons            80.815333
## Illiterate - Persons                      19.184667
## Total Worker Population - Persons         36.873962
## Non Worker Population - Persons           63.126038
G1<-t(cities1)%*%cities1
G1
##                                           Population in the Age Group 0-6 - Persons
## Population in the Age Group 0-6 - Persons                                942.585755
## Population in the Age Group 7-13                                         636.120683
## Population in the Age Group 14-20                                        584.136383
## SC total education                                                       442.668476
## ST total education                                                         1.913034
## Literates Population - Persons                                          7169.718238
## Illiterate - Persons                                                    1977.303237
## Total Worker Population - Persons                                       3353.725029
## Non Worker Population - Persons                                         5793.296446
##                                           Population in the Age Group 7-13
## Population in the Age Group 0-6 - Persons                          636.12068
## Population in the Age Group 7-13                                   531.54188
## Population in the Age Group 14-20                                  454.89337
## SC total education                                                 230.05182
## ST total education                                                   1.19944
## Literates Population - Persons                                    4939.84016
## Illiterate - Persons                                              1346.29235
## Total Worker Population - Persons                                 2263.19103
## Non Worker Population - Persons                                   4022.94148
##                                           Population in the Age Group 14-20
## Population in the Age Group 0-6 - Persons                           584.13638
## Population in the Age Group 7-13                                    454.89337
## Population in the Age Group 14-20                                   433.02633
## SC total education                                                  246.63084
## ST total education                                                    1.08543
## Literates Population - Persons                                     4550.38134
## Illiterate - Persons                                               1260.24443
## Total Worker Population - Persons                                  2095.78326
## Non Worker Population - Persons                                    3714.84251
##                                           SC total education
## Population in the Age Group 0-6 - Persons        442.6684757
## Population in the Age Group 7-13                 230.0518165
## Population in the Age Group 14-20                246.6308438
## SC total education                               351.4368794
## ST total education                                 0.8613813
## Literates Population - Persons                  3193.3118630
## Illiterate - Persons                             898.4687642
```

```
## Total Worker Population - Persons              1532.4572821
## Non Worker Population - Persons                2559.3233451
##                                          ST total education
## Population in the Age Group 0-6 - Persons         1.91303410
## Population in the Age Group 7-13                  1.19943965
## Population in the Age Group 14-20                 1.08542967
## SC total education                                0.86138130
## ST total education                                0.01051752
## Literates Population - Persons                   15.88782205
## Illiterate - Persons                              4.30106030
## Total Worker Population - Persons                 7.14195503
## Non Worker Population - Persons                  13.04692733
##                                          Literates Population - Persons
## Population in the Age Group 0-6 - Persons                     7169.71824
## Population in the Age Group 7-13                              4939.84016
## Population in the Age Group 14-20                             4550.38134
## SC total education                                            3193.31186
## ST total education                                              15.88782
## Literates Population - Persons                               55417.36588
## Illiterate - Persons                                         15186.50410
## Total Worker Population - Persons                            25736.26366
## Non Worker Population - Persons                              44867.60631
##                                          Illiterate - Persons
## Population in the Age Group 0-6 - Persons           1977.30324
## Population in the Age Group 7-13                    1346.29235
## Population in the Age Group 14-20                   1260.24443
## SC total education                                   898.46876
## ST total education                                     4.30106
## Literates Population - Persons                     15186.50410
## Illiterate - Persons                                4209.62592
## Total Worker Population - Persons                   7090.88393
## Non Worker Population - Persons                    12305.24609
##                                          Total Worker Population - Persons
## Population in the Age Group 0-6 - Persons                       3353.725029
## Population in the Age Group 7-13                                2263.191027
## Population in the Age Group 14-20                               2095.783257
## SC total education                                              1532.457282
## ST total education                                                 7.141955
## Literates Population - Persons                                 25736.263663
## Illiterate - Persons                                            7090.883931
## Total Worker Population - Persons                              12022.178411
## Non Worker Population - Persons                                20804.969183
##                                          Non Worker Population - Persons
## Population in the Age Group 0-6 - Persons                       5793.29645
## Population in the Age Group 7-13                                4022.94148
## Population in the Age Group 14-20                               3714.84251
## SC total education                                              2559.32335
## ST total education                                                13.04693
## Literates Population - Persons                                 44867.60631
## Illiterate - Persons                                           12305.24609
## Total Worker Population - Persons                              20804.96918
## Non Worker Population - Persons                                36367.88322
chisqD(G1)$dist.col
##               [,1]         [,2]         [,3]       [,4]        [,5]
##  [1,] 0.000000e+00 0.0032771323 0.0016382051 0.01432944 0.002307910
```

```
##  [2,] 3.277132e-03 0.0000000000 0.0006176106 0.02913877 0.006784045
##  [3,] 1.638205e-03 0.0006176106 0.0000000000 0.02245169 0.005299729
##  [4,] 1.432944e-02 0.0291387746 0.0224516886 0.00000000 0.018995043
##  [5,] 2.307910e-03 0.0067840452 0.0052997288 0.01899504 0.000000000
##  [6,] 1.179040e-04 0.0024416624 0.0011794891 0.01688439 0.002010602
##  [7,] 5.796726e-05 0.0027279018 0.0012591840 0.01591966 0.002170082
##  [8,] 3.200668e-05 0.0030038642 0.0015050029 0.01548825 0.002016774
##  [9,] 1.656377e-04 0.0022347481 0.0010383778 0.01737737 0.002079945
##               [,6]         [,7]         [,8]         [,9]
##  [1,] 1.179040e-04 5.796726e-05 3.200668e-05 1.656377e-04
##  [2,] 2.441662e-03 2.727902e-03 3.003864e-03 2.234748e-03
##  [3,] 1.179489e-03 1.259184e-03 1.505003e-03 1.038378e-03
##  [4,] 1.688439e-02 1.591966e-02 1.548825e-02 1.737737e-02
##  [5,] 2.010602e-03 2.170082e-03 2.016774e-03 2.079945e-03
##  [6,] 0.000000e+00 2.735440e-05 3.855021e-05 5.902918e-06
##  [7,] 2.735440e-05 0.000000e+00 1.480479e-05 4.402986e-05
##  [8,] 3.855021e-05 1.480479e-05 0.000000e+00 7.138076e-05
##  [9,] 5.902918e-06 4.402986e-05 7.138076e-05 0.000000e+00
rG2 <- cmdscale(chisqD(G1)$dist.col, eig = TRUE)
cG2 <- cmdscale(chisqD(G1)$dist.row, eig = TRUE)
plot(rG2$points, xlim = range(rG2$points[,1], cG2$points[,1]) ,
     ylim = range(rG2$points[,1], cG2$points[,1]))
text(rG2$points, labels = colnames(G1), cex = 1)
abline(h = 0, lty = 2,col='red')
abline(v = 0, lty = 2,col='red')
```



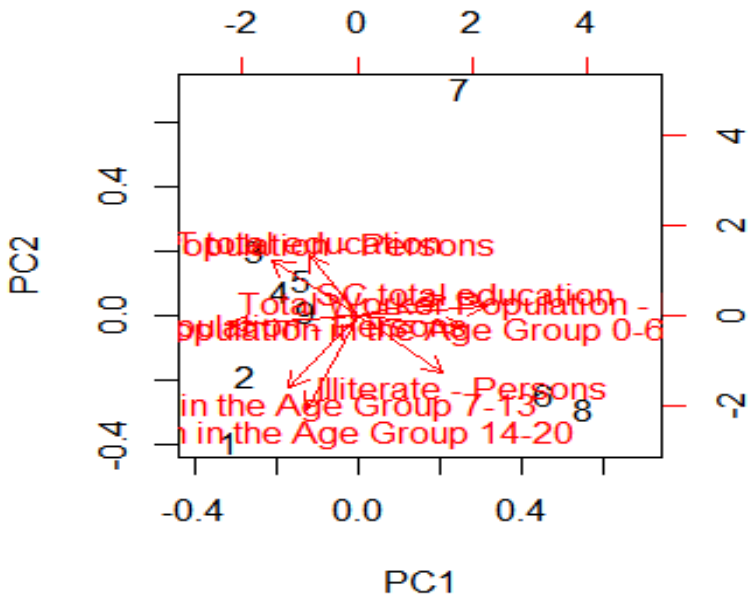
4.2.B. EDUCATION ANALYSIS ON CLOSELY SCATTERED AREAS IN PRINCIPLE COMPONENT ANALYSIS

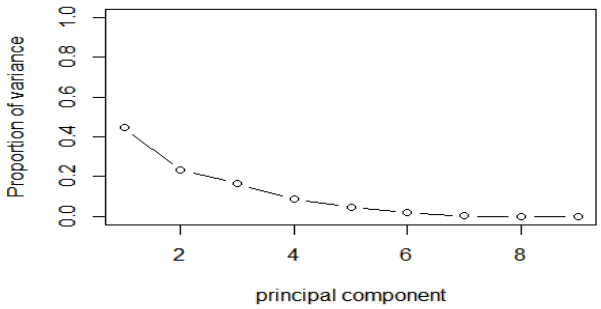
Here we can clearly see that the illitrate population is drastically gone down in towns like Ayyapandhangal(6) & Kattangalatur(8). But very high literate population in Annanagar(4), Alandur(5), vandalur(9). With all this even interesting things like population age group of education for (14-20) is

also more in Royapuram. Which means compared to Royapuram(1) & Ayyapandhangal (6) we should have more elder people who are illiterate but yonger generations are giving importance to education.

```
row.names(chisquaresecond9)<-chisquaresecond9$Name
## Warning: Setting row names on a tibble is deprecated.
censussecond9.pca=prcomp(chisquaresecond9[,-1],scale=TRUE)
summary(censussecond9.pca)
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6
## Standard deviation     2.0025 1.4429 1.2168 0.88910 0.65921 0.42371
## Proportion of Variance 0.4456 0.2313 0.1645 0.08783 0.04828 0.01995
## Cumulative Proportion  0.4456 0.6769 0.8414 0.92924 0.97753 0.99748
##                            PC7       PC8       PC9
## Standard deviation     0.15072 8.107e-16 4.247e-16
## Proportion of Variance 0.00252 0.000e+00 0.000e+00
## Cumulative Proportion  1.00000 1.000e+00 1.000e+00
print(censussecond9.pca)
## Standard deviations (1, .., p=9):
## [1] 2.002489e+00 1.442921e+00 1.216845e+00 8.891016e-01 6.592075e-01
## [6] 4.237128e-01 1.507150e-01 8.107234e-16 4.246922e-16
##
## Rotation (n x k) = (9 x 9):
##                                                 PC1         PC2
## Population in the Age Group 0-6 - Persons  0.3878471 -0.07540409
## Population in the Age Group 7-13          -0.2511728 -0.45818872
## Population in the Age Group 14-20         -0.1815610 -0.59730969
## SC total education                         0.3133079  0.12135711
## ST total education                        -0.1758653  0.38531646
## Literates Population - Persons            -0.3081798  0.35851970
## Illiterate - Persons                       0.3081798 -0.35851970
## Total Worker Population - Persons          0.4660915  0.06053668
## Non Worker Population - Persons           -0.4660915 -0.06053668
##                                                  PC3         PC4
## Population in the Age Group 0-6 - Persons -0.339971051  0.24658905
## Population in the Age Group 7-13          -0.293046462 -0.01839789
## Population in the Age Group 14-20          0.007693579  0.26356388
## SC total education                        -0.004580942  0.83711059
## ST total education                         0.539841405  0.12465374
## Literates Population - Persons            -0.480531917  0.07025125
## Illiterate - Persons                       0.480531917 -0.07025125
## Total Worker Population - Persons         -0.150354068 -0.26760631
## Non Worker Population - Persons            0.150354068  0.26760631
##                                                 PC5         PC6
## Population in the Age Group 0-6 - Persons -0.53345291  0.4614555
## Population in the Age Group 7-13          -0.57914106 -0.4360271
## Population in the Age Group 14-20          0.34863214 -0.2209111
## SC total education                         0.12775715 -0.2621717
## ST total education                        -0.46986046 -0.3612807
## Literates Population - Persons             0.08522531 -0.1074563
## Illiterate - Persons                      -0.08522531  0.1074563
## Total Worker Population - Persons          0.05857398 -0.4036001
## Non Worker Population - Persons           -0.05857398  0.4036001
##                                                 PC7           PC8
## Population in the Age Group 0-6 - Persons -0.4122975  0.000000e+00
## Population in the Age Group 7-13           0.3394652 -5.881787e-16
## Population in the Age Group 14-20         -0.6085931  9.265680e-16
```

```
## SC total education                            0.3182446  5.787457e-17
## ST total education                           -0.4029178 -1.952452e-17
## Literates Population - Persons               -0.1477582 -4.646072e-01
## Illiterate - Persons                          0.1477582 -4.646072e-01
## Total Worker Population - Persons            -0.1362006 -5.330480e-01
## Non Worker Population - Persons               0.1362006 -5.330480e-01
##                                                       PC9
## Population in the Age Group 0-6 - Persons  0.000000e+00
## Population in the Age Group 7-13           4.999519e-16
## Population in the Age Group 14-20         -2.166617e-16
## SC total education                        -3.440318e-16
## ST total education                         7.623582e-16
## Literates Population - Persons             5.330480e-01
## Illiterate - Persons                       5.330480e-01
## Total Worker Population - Persons         -4.646072e-01
## Non Worker Population - Persons           -4.646072e-01
#Standard deviation of each component
censussecond9.sd=censussecond9.pca$sdev
censussecond9.var=censussecond9.pca$sdev^2
censussecond9.var
## [1] 4.009963e+00 2.082021e+00 1.480713e+00 7.905017e-01 4.345546e-01
## [6] 1.795325e-01 2.271503e-02 6.572725e-31 1.803635e-31
#proportion of variance explained
pve=censussecond9.var/sum(censussecond9.var)
pve
## [1] 4.455514e-01 2.313356e-01 1.645236e-01 8.783352e-02 4.828384e-02
## [6] 1.994806e-02 2.523892e-03 7.303028e-32 2.004038e-32
biplot(censussecond9.pca) #biplot
```



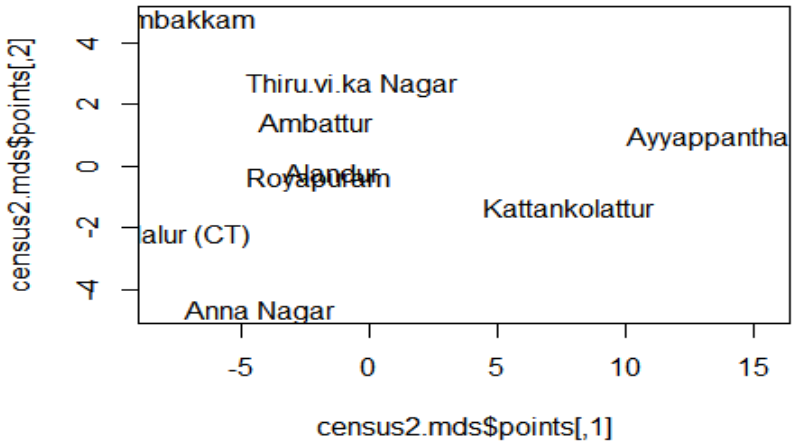


4.2.C. EDUCATION ANALYSIS ON CLOSELY SCATTERED AREAS IN MULTIDIMENTIONAL SCALLING

Here in MDS we have a good comparison scatter with regard to PCA also having kelambakam(7) spotted all alone. And all the points taken into analysis have a similar effect which was shown in the PCA analysis. So this proves exactly better way of my work.

```
census2<-as.matrix(chisquaresecond9[,-1])
census2.mds<-cmdscale(census2,k=2,eig=TRUE)
census2.mds$points
```

```
##              [,1]       [,2]
##  [1,] -1.9702934 -0.4243603
##  [2,] -0.6650249  2.6353795
##  [3,] -2.0694431  1.4476625
##  [4,] -4.2782377 -4.6956332
##  [5,] -1.3958641 -0.1994377
##  [6,] 15.4901138  0.9057278
##  [7,] -7.9977838  4.7837113
##  [8,]  7.8417015 -1.3298734
##  [9,] -8.0908547 -2.2497800
row.names(chisquaresecond9)<-chisquaresecond9$Name
## Warning: Setting row names on a tibble is deprecated.
plot(census2.mds$points,type="n",col='red')
#text(census.mds$points,rownames(census.mds$points),col=rainbow(18))
text(census2.mds$points,rownames(chisquaresecond9))
```



5. WORKER - LOCATION BASED ANALYSIS

5.1.A. WORKER ANALYSIS ON MOST SCATTERED AREAS IN CHI-SQUARE ANALYSIS

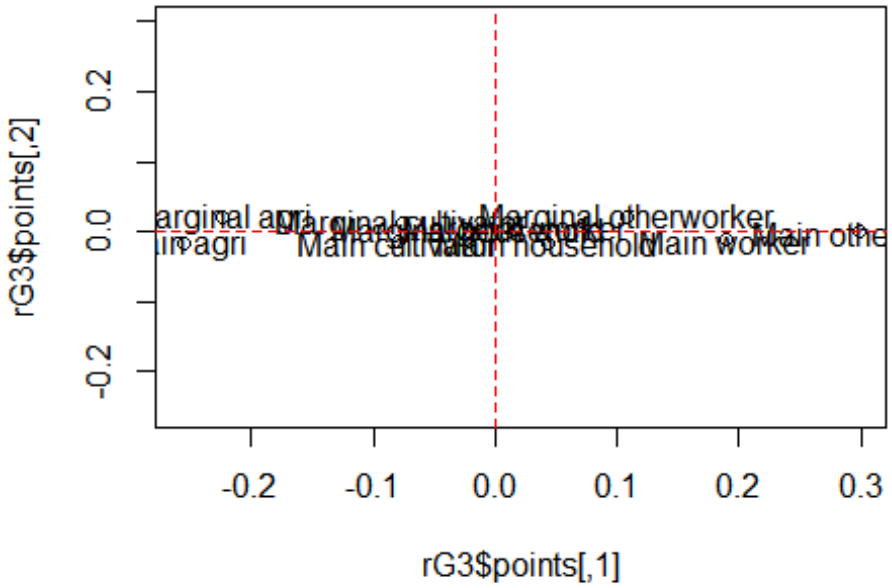
With otherworking population more here and agriculture population is very less we can clearly see that when comparing with the education of most scattered the literate population is more so the other working population and worker and household population is more. The cultivation is suprizingly getting almost touched the margin which means that in these areas the cultivators are having agricultural lands in these areas. In a city this is very rare to see.

```
data.url = "https://docs.google.com/spreadsheets/d/1EfJamAr67K1KZun8wzYAwVGxjtmWZZYSUqPgIXRz3c0/edit#gid=1094962871"

#my_sheets = gs_ls()
data = data.url %>%
  gs_url() %>%
  gs_read()
## Sheet-identifying info appears to be a browser URL.
## googlesheets will attempt to extract sheet key from the URL.
## Putative key: 1EfJamAr67K1KZun8wzYAwVGxjtmWZZYSUqPgIXRz3c0
## Sheet successfully identified: "chisquareworker1to9.xlsx"
## Accessing worksheet titled 'Sheet1'.
## Parsed with column specification:
## cols(
##   Name = col_character(),
##   `Main worker` = col_double(),
##   `Main cultivator` = col_double(),
##   `Main agri` = col_double(),
##   `Main household` = col_double(),
##   `Main otherwork` = col_double(),
##   `Marginal worker` = col_double(),
##   `Marginal cultivator` = col_double(),
##   `Marginal agri` = col_double(),
##   `Marginal household` = col_double(),
##   `Marginal otherworker` = col_double()
## )
chisquareworker1to9=data
chisquareworker1to9
## # A tibble: 10 x 11
##    Name        `Main worker` `Main cultivato~ `Main agri` `Main household`
##    <chr>               <dbl>            <dbl>       <dbl>            <dbl>
##  1 Thiruvotti~          31.0            0.102      0.0998            0.878
##  2 Manali Tot~          29.5            0.337      0.162             0.539
##  3 Madhavaram           29.1            0.348      0.180             0.670
##  4 Teynampet            33.8            0.389      0.161             0.552
##  5 Kodambakam           33.4            0.365      0.119             0.462
##  6 Adyar                33.0            0.455      0.140             0.555
##  7 Poonamallee          31.7            0.567      0.0946            0.118
##  8 Walajabad            36.5            4.94      11.7               1.73
##  9 Sriperumbu~          31.4            2.36       5.31              1.36
## 10 St. Thomas~          34.8            0.393      0.503             0.675
## # ... with 6 more variables: `Main otherwork` <dbl>, `Marginal
## #   worker` <dbl>, `Marginal cultivator` <dbl>, `Marginal agri` <dbl>,
## #   `Marginal household` <dbl>, `Marginal otherworker` <dbl>
attach(chisquareworker1to9)
```

```
## The following object is masked from chisquaresecond9:
##
##     Name
## The following object is masked from chisquare1to9:
##
##     Name
## The following objects are masked from editpurpose:
##
##     Main agri, Main cultivator, Main household, Main otherwork,
##     Main worker, Marginal agri, Marginal cultivator, Marginal
##     household, Marginal otherworker, Marginal worker, Name
dim(chisquareworker1to9)
## [1] 10 11
worker1<-as.matrix(chisquareworker1to9[,c(2:11)])
dim(worker1)
## [1] 10 10
t(worker1)
##                            [,1]        [,2]        [,3]        [,4]
## Main worker          30.97949217 29.53482525 29.11600668 33.78589678
## Main cultivator       0.10209867  0.33693739  0.34826285  0.38903935
## Main agri             0.09983985  0.16236215  0.18047336  0.16118598
## Main household        0.87777750  0.53881246  0.67043317  0.55197974
## Main otherwork       29.89977615 28.49671324 27.91683729 32.68369171
## Marginal worker       2.20347903  2.34203815  2.71217398  1.96997774
## Marginal cultivator   0.01219763  0.03232875  0.07356645  0.03004419
## Marginal agri         0.02891290  0.02227091  0.02319336  0.01842125
## Marginal household    0.12649393  0.21408815  0.08262636  0.06447439
## Marginal otherworker  2.03587458  2.07335034  2.53278781  1.85703791
##                            [,5]        [,6]        [,7]       [,8]
## Main worker          33.39767609 32.98302179 31.67848700 36.4671552
## Main cultivator       0.36478919  0.45495639  0.56737589  4.9380335
## Main agri             0.11913323  0.13955897  0.09456265 11.7208070
## Main household        0.46211437  0.55501754  0.11820331  1.7315400
## Main otherwork       32.45163930 31.83348889 30.89834515 18.0767747
## Marginal worker       2.52558841  2.96584755  4.27895981 10.5166949
## Marginal cultivator   0.06830786  0.04359389  0.07092199  0.6490131
## Marginal agri         0.02415106  0.04388647  0.30732861  5.6490969
## Marginal household    0.07119157  0.15711356  0.04728132  0.7135790
## Marginal otherworker  2.36193792  2.72125363  3.85342790  3.5050060
##                           [,9]      [,10]
## Main worker          31.3742804 34.8070883
## Main cultivator       2.3624637  0.3928613
## Main agri             5.3144613  0.5025263
## Main household        1.3591308  0.6752172
## Main otherwork       22.3382245 33.2364836
## Marginal worker      14.2847249  5.5868234
## Marginal cultivator   0.9583171  0.1609261
## Marginal agri         6.0615504  0.1271022
## Marginal household    0.8864650  0.2102963
## Marginal otherworker  6.3783924  5.0884988
G2<-t(worker1)%*%worker1
G2
##                      Main worker Main cultivator Main agri Main household
## Main worker          10553.80682      349.431968 641.81962     248.048939
## Main cultivator        349.43197       31.178311  70.98145      13.234110
```

```
## Main agri              641.81962       70.981447 166.01084      28.386184
## Main household         248.04894       13.234110  28.38618       7.651926
## Main otherwork        9314.50629      234.038100 376.44115     198.776718
## Marginal worker       1626.82983       95.297241 204.51398      50.818006
## Marginal cultivator     69.79115        5.666497  12.82628       2.693106
## Marginal agri          415.44829       42.494404  98.54144      18.240841
## Marginal household      85.03756        5.963719  13.28806       3.025467
## Marginal otherworker  1056.55283       41.172621  79.85820      26.858591
##                       Main otherwork Marginal worker Marginal cultivator
## Main worker              9314.50629      1626.82983           69.791152
## Main cultivator           234.03810        95.29724            5.666497
## Main agri                 376.44115       204.51398           12.826280
## Main household            198.77672        50.81801            2.693106
## Main otherwork           8505.25033      1276.20060           48.605269
## Marginal worker          1276.20060       400.92834           22.380421
## Marginal cultivator        48.60527        22.38042            1.384592
## Marginal agri             256.17160       148.42877            9.524371
## Marginal household         62.76032        23.32162            1.378024
## Marginal otherworker      908.66342       206.79753           10.093434
##                       Marginal agri Marginal household Marginal otherworker
## Main worker              415.448288          85.037560           1056.55283
## Main cultivator           42.494404           5.963719             41.17262
## Main agri                 98.541444          13.288057             79.85820
## Main household            18.240841           3.025467             26.85859
## Main otherwork           256.171600          62.760317            908.66342
## Marginal worker          148.428772          23.321619            206.79753
## Marginal cultivator        9.524371           1.378024             10.09343
## Marginal agri             68.770014           9.465833             60.66855
## Marginal household         9.465833           1.444047             11.03372
## Marginal otherworker      60.668554          11.033716            125.00183
chisqD(G2)$dist.col
##              [,1]       [,2]        [,3]        [,4]        [,5]
##  [1,] 0.000000000 0.24447663 0.439471129 0.042305184 0.006325135
##  [2,] 0.244476627 0.00000000 0.028394018 0.083889967 0.329446806
##  [3,] 0.439471129 0.02839402 0.000000000 0.209701736 0.551241195
##  [4,] 0.042305184 0.08388997 0.209701736 0.000000000 0.081280172
##  [5,] 0.006325135 0.32944681 0.551241195 0.081280172 0.000000000
##  [6,] 0.078531395 0.05353570 0.156735834 0.007534347 0.128239531
##  [7,] 0.242478487 0.01096476 0.043789163 0.084817748 0.325431808
##  [8,] 0.412214509 0.02791929 0.008676197 0.191466664 0.519733859
##  [9,] 0.147252194 0.01839563 0.086168804 0.032954963 0.213674504
## [10,] 0.010158943 0.16950651 0.335530395 0.016037592 0.030204829
##              [,6]       [,7]        [,8]       [,9]      [,10]
##  [1,] 0.078531395 0.24247849 0.412214509 0.14725219 0.01015894
##  [2,] 0.053535698 0.01096476 0.027919291 0.01839563 0.16950651
##  [3,] 0.156735834 0.04378916 0.008676197 0.08616880 0.33553040
##  [4,] 0.007534347 0.08481775 0.191466664 0.03295496 0.01603759
##  [5,] 0.128239531 0.32543181 0.519733859 0.21367450 0.03020483
##  [6,] 0.000000000 0.04512965 0.132656661 0.01110884 0.03496276
##  [7,] 0.045129646 0.00000000 0.024113591 0.01203956 0.15905801
##  [8,] 0.132656661 0.02411359 0.000000000 0.06715299 0.30382077
##  [9,] 0.011108838 0.01203956 0.067152989 0.00000000 0.08539673
## [10,] 0.034962763 0.15905801 0.303820771 0.08539673 0.00000000
rG3 <- cmdscale(chisqD(G2)$dist.col, eig = TRUE)
cG3 <- cmdscale(chisqD(G2)$dist.row, eig = TRUE)
```

```
plot(rG3$points, xlim = range(rG3$points[,1], cG3$points[,1]) ,
     ylim = range(rG3$points[,1], cG3$points[,1]))
text(rG3$points, labels = colnames(G2), cex = 1)
abline(h = 0, lty = 2,col='red')
abline(v = 0, lty = 2,col='red')
```

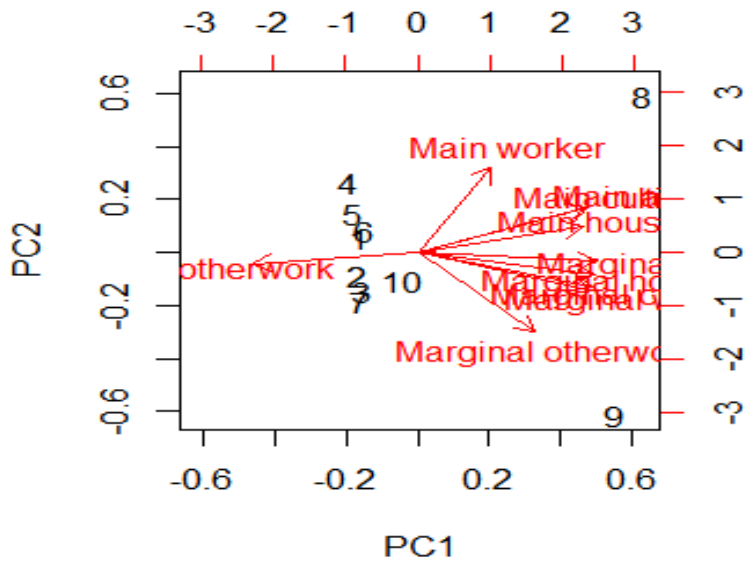


5.1.B. WORKER ANALYSIS ON MOST SCATTERED AREAS IN PRINCIPLE COMPONENT ANALYSIS

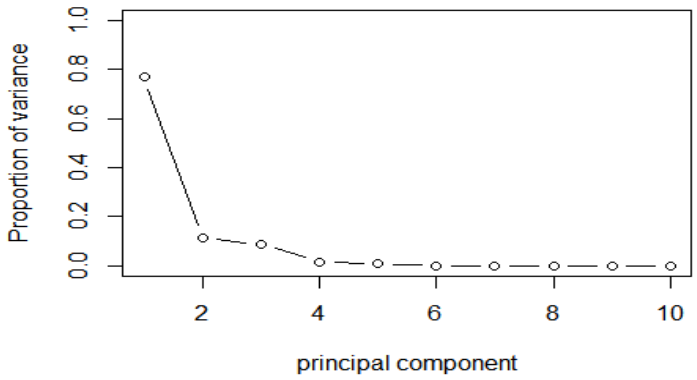
Here in this PCA most of the population is based on the otherworking groups than rest over subgroups. Only St Thomas mount(8) have a main worker based population and Sripermbadur(9) have marginal other worker population. Other distribution of the locations i think its based on the the divitions of the main working groups. In this scenario Though we have more literate population in the most scattered areas the people who work for other working is more which means on aveage it can be said that competitions in getting a dream job is very difficult though being a literate. Or this can also be because of the illiterate pop who needs other working income might be the way this population is divided. In either case its really a very diverse not census group.

```
row.names(chisquareworker1to9)<-chisquareworker1to9$Name
## Warning: Setting row names on a tibble is deprecated.
workerfirst9.pca=prcomp(chisquareworker1to9[,-1],scale=TRUE)
summary(workerfirst9.pca)
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6
## Standard deviation     2.7716 1.0799 0.94070 0.40849 0.27884 0.13994
## Proportion of Variance 0.7682 0.1166 0.08849 0.01669 0.00778 0.00196
## Cumulative Proportion  0.7682 0.8848 0.97330 0.98998 0.99776 0.99972
##                           PC7     PC8       PC9      PC10
## Standard deviation     0.04885 0.02083 2.436e-10 1.464e-16
## Proportion of Variance 0.00024 0.00004 0.000e+00 0.000e+00
## Cumulative Proportion  0.99996 1.00000 1.000e+00 1.000e+00

print(workerfirst9.pca)
```

```
## Standard deviations (1, .., p=10):
##  [1] 2.771621e+00 1.079897e+00 9.407020e-01 4.084866e-01 2.788417e-01
##  [6] 1.399446e-01 4.885106e-02 2.083325e-02 2.436066e-10 1.463553e-16
##
## Rotation (n x k) = (10 x 10):
##                              PC1         PC2         PC3         PC4
## Main worker            0.1460765  0.57759794  0.70889971 -0.09304148
## Main cultivator        0.3345767  0.30493291 -0.05657440  0.40977364
## Main agri              0.3373410  0.30446049 -0.07469918  0.24572124
## Main household         0.3269736  0.17653105 -0.18946290 -0.75897345
## Main otherwork        -0.3258023 -0.07541428  0.42428912 -0.28720190
## Marginal worker        0.3424903 -0.25960535  0.14479111  0.05730621
## Marginal cultivator    0.3457445 -0.23104311  0.03605754 -0.09329522
## Marginal agri          0.3565967 -0.06334589 -0.05343315  0.05493635
## Marginal household     0.3495618 -0.15498513 -0.04857285 -0.26284702
## Marginal otherworker   0.2316651 -0.54317666  0.49526164  0.14293261
##                              PC5         PC6         PC7          PC8
## Main worker           -0.10496446  0.02270657 -0.01429738 -0.090863044
## Main cultivator        0.07545254  0.06146841 -0.12194308  0.742156032
## Main agri              0.17872014  0.04610559 -0.05445029 -0.615641915
## Main household         0.41800708 -0.21288162  0.05624274  0.148758379
## Main otherwork        -0.24979370 -0.02374244  0.06742473  0.192288273
## Marginal worker       -0.09426803 -0.12740137  0.30022230  0.001126173
## Marginal cultivator   -0.40629660 -0.37991541 -0.70786950 -0.050213748
## Marginal agri         -0.40710259 -0.29071523  0.61798596 -0.002130660
## Marginal household    -0.27683460  0.83744627 -0.03192626  0.001054796
## Marginal otherworker   0.54579151  0.03604988 -0.01061622  0.017354781
##                               PC9         PC10
## Main worker            0.336483100 -0.025145001
## Main cultivator       -0.220342872  0.016465974
## Main agri             -0.556569548  0.041591810
## Main household        -0.067981817  0.005080204
## Main otherwork        -0.719919135  0.053798741
## Marginal worker        0.061261798  0.819787610
## Marginal cultivator   -0.004749735 -0.063559588
## Marginal agri         -0.035800684 -0.479074357
## Marginal household    -0.004322369 -0.057840681
## Marginal otherworker  -0.021850659 -0.292399177

#Standard deviation of each component
workerfirst9.sd=workerfirst9.pca$sdev
workerfirst9.var=workerfirst9.pca$sdev^2
workerfirst9.var
##  [1] 7.681883e+00 1.166177e+00 8.849203e-01 1.668613e-01 7.775271e-02
##  [6] 1.958448e-02 2.386426e-03 4.340244e-04 5.934416e-20 2.141986e-32
#proportion of variance explained
pve=workerfirst9.var/sum(workerfirst9.var)
pve
##  [1] 7.681883e-01 1.166177e-01 8.849203e-02 1.668613e-02 7.775271e-03
##  [6] 1.958448e-03 2.386426e-04 4.340244e-05 5.934416e-21 2.141986e-33
#biplot
biplot(workerfirst9.pca)
```




```
#proportion of variance explained
plot(pve,xlab="principal component",ylab="Proportion of variance" , ylim=c(0,1), type='b'
)
```



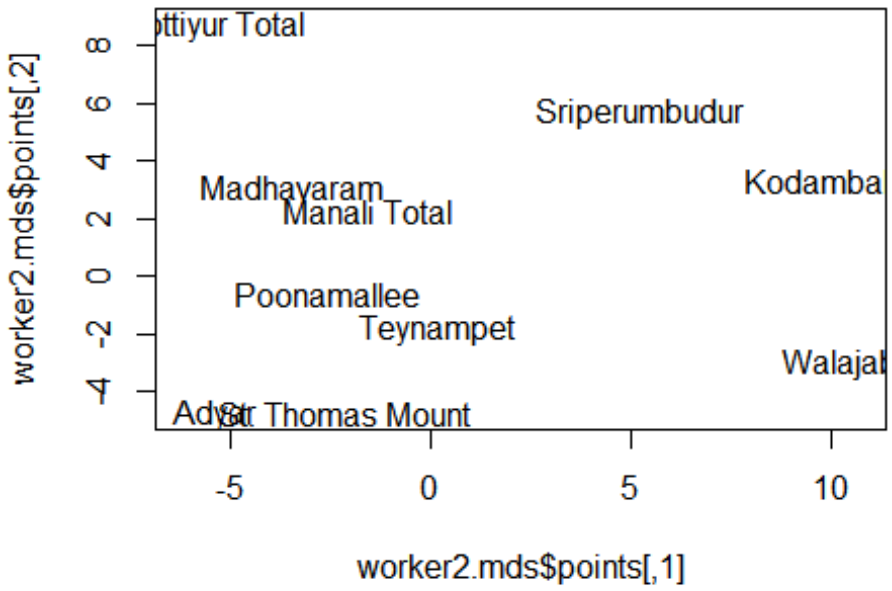
5.1.C. WORKER ANALYSIS ON MOST SCATTERED AREAS IN MULTIDIMENTIONAL SCALE ANALYSIS

As explained in the PCA though the group are similar in few things in MDS it is totally differnt story of showing various points for the locations. But still it proves a little bit for few locations compared to the PCA.

```
worker2<-as.matrix(chisquareworker1to9[,-1])
worker2.mds<-cmdscale(worker2,k=2,eig=TRUE)
worker2.mds$points
```

```
##             [,1]       [,2]
##  [1,] -6.2087817  8.7128220
##  [2,] -1.6058419  2.2898538
##  [3,] -3.4560195  3.1494198
##  [4,]  0.1859184 -1.7994260
##  [5,] 10.3403263  3.3217822
##  [6,] -5.3641970 -4.7623769
```

```
##  [7,] -2.6000468 -0.5850115
##  [8,] 10.7094812 -3.0222729
##  [9,]  5.2303887  5.6921700
## [10,] -2.1439102 -4.7442156

row.names(chisquareworker1to9)<-chisquareworker1to9$Name

## Warning: Setting row names on a tibble is deprecated.

plot(worker2.mds$points,type="n",col='red')
#text(census.mds$points,rownames(census.mds$points),col=rainbow(18))
text(worker2.mds$points,rownames(chisquareworker1to9))
```



5.2.A. WORKER ANALYSIS ON CLOSELY SCATTERED AREAS IN CHI-SQUARE ANALYSIS

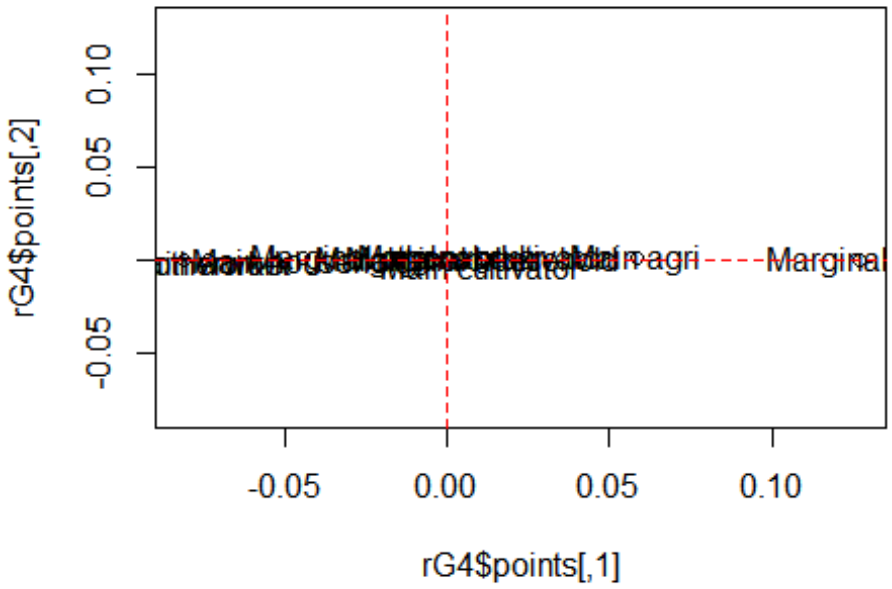
Intrestingly, marginal agri workers are more in this population groups. that's would be a real suprize because, when we compare with the PCA of the education of closely scatterd group I mentioned that elder people were not much educated but only the yonger. So in this scenario the elder people are considered doing agriculture and cultivators and only the yonger generations are having the household and other working interest.

```
data.url = "https://docs.google.com/spreadsheets/d/1581Srpvo9_TSFPr6Oc5sWtYQD4eUav1rJCTEsyzxSH8/edit#gid=2041523407"

#my_sheets = gs_ls()
data = data.url %>%
  gs_url() %>%
  gs_read()
## Sheet-identifying info appears to be a browser URL.
## googlesheets will attempt to extract sheet key from the URL.
## Putative key: 1581Srpvo9_TSFPr6Oc5sWtYQD4eUav1rJCTEsyzxSH8
## Sheet successfully identified: "chisquareworkersecond9.xlsx"
## Accessing worksheet titled 'Sheet1'.
```

```
## Parsed with column specification:
## cols(
##   Name = col_character(),
##   `Main worker` = col_double(),
##   `Main cultivator` = col_double(),
##   `Main agri` = col_double(),
##   `Main household` = col_double(),
##   `Main otherwork` = col_double(),
##   `Marginal worker` = col_double(),
##   `Marginal cultivator` = col_double(),
##   `Marginal agri` = col_double(),
##   `Marginal household` = col_double(),
##   `Marginal otherworker` = col_double()
## )
chisquareworkersecond9=data
chisquareworkersecond9
## # A tibble: 10 x 11
##    Name        `Main worker` `Main cultivato~ `Main agri` `Main household`
##    <chr>               <dbl>            <dbl>       <dbl>            <dbl>
##  1 Tondiarpet           28.2            0.415      0.165             0.799
##  2 Royapuram            30.2            0.339      0.150             0.527
##  3 Thiru.vi.k~          32.7            0.443      0.147             0.489
##  4 Ambattur             31.7            0.191      0.0684            0.467
##  5 Anna Nagar           33.5            0.328      0.0921            0.683
##  6 Alandur              34.9            0.589      0.135             0.558
##  7 Ayyappanth~          36.3            0.143      0.214             0.273
##  8 Kelambakkam          33.4            0.463      0.501             0.771
##  9 Kattankola~          32.2            1.89       2.54              0.820
## 10 Vandalur (~          31.6            0.326      0.255             0.510
## # ... with 6 more variables: `Main otherwork` <dbl>, `Marginal
## #   worker` <dbl>, `Marginal cultivator` <dbl>, `Marginal agri` <dbl>,
## #   `Marginal household` <dbl>, `Marginal otherworker` <dbl>
attach(chisquareworkersecond9)
## The following objects are masked from chisquareworker1to9:
##
##     Main agri, Main cultivator, Main household, Main otherwork,
##     Main worker, Marginal agri, Marginal cultivator, Marginal
##     household, Marginal otherworker, Marginal worker, Name
## The following object is masked from chisquaresecond9:
##
##     Name
## The following object is masked from chisquare1to9:
##
##     Name
## The following objects are masked from editpurpose:
##
##     Main agri, Main cultivator, Main household, Main otherwork,
##     Main worker, Marginal agri, Marginal cultivator, Marginal
##     household, Marginal otherworker, Marginal worker, Name
dim(chisquareworkersecond9)
## [1] 10 11
worker3<-as.matrix(chisquareworkersecond9[,c(2:11)])
dim(worker3)
## [1] 10 10
t(worker3)
```

```
##                            [,1]        [,2]        [,3]        [,4]
## Main worker           28.19280082 30.19240652 32.71140608 31.72876227
## Main cultivator        0.41476966  0.33940299  0.44251216  0.19084298
## Main agri              0.16472741  0.14986879  0.14682771  0.06842838
## Main household         0.79922304  0.52717862  0.48862639  0.46661967
## Main otherwork        26.81408071 29.17595612 31.63343983 31.00287124
## Marginal worker        3.58188213  2.58059600  2.38650366  2.40393377
## Marginal cultivator    0.05633999  0.04494110  0.07507397  0.02028781
## Marginal agri          0.04721828  0.03165416  0.04021161  0.01341059
## Marginal household     0.20362883  0.18093677  0.12063483  0.06774066
## Marginal otherworker   3.27469503  2.32306397  2.15058325  2.30249471
##                            [,5]        [,6]        [,7]        [,8]
## Main worker           33.455873640 34.91998836 36.31552419 33.39757179
## Main cultivator        0.328015355  0.58855314  0.14280914  0.46251686
## Main agri              0.092074486  0.13466894  0.21421371  0.50105993
## Main household         0.683111441  0.55779542  0.27301747  0.77086144
## Main otherwork        32.352672360 33.63897086 35.68548387 31.66313355
## Marginal worker        1.971815729  1.38243485  2.41095430  4.79861245
## Marginal cultivator    0.031819859  0.04904610  0.02940188  0.28907304
## Marginal agri          0.008462728  0.01828837  0.09240591  0.07708614
## Marginal household     0.081242193  0.05652770  0.09660618  0.21198690
## Marginal otherworker   1.850290949  1.25857268  2.19254032  4.22046637
##                            [,9]      [,10]
## Main worker           32.1771016 31.5511512
## Main cultivator        1.8884905  0.3263708
## Main agri              2.5385716  0.2551626
## Main household         0.8200214  0.5103252
## Main otherwork        26.9300181 30.4592927
## Marginal worker        8.5640292  5.3228103
## Marginal cultivator    0.3599860  0.1424163
## Marginal agri          1.3011197  0.1008782
## Marginal household     0.6759312  0.2788986
## Marginal otherworker   6.2269923  4.8006171
G3<-t(worker3)%*%worker3
G3
##                      Main worker Main cultivator  Main agri Main household
## Main worker          10586.92305     165.6942795 138.174205    189.7172778
## Main cultivator        165.69428       4.8806857   5.446370      3.4787405
## Main agri              138.17420       5.4463700   6.908853      3.1089810
## Main household         189.71728       3.4787405   3.108981      3.7525741
## Main otherwork       10093.33729     151.8884832 122.710001    179.3769822
## Marginal worker       1138.80436      25.8108138  27.878848     22.7254238
## Marginal cultivator     35.62078       0.9792285   1.139293      0.7627016
## Marginal agri           55.92891       2.6031472   3.409731      1.2993914
## Marginal household      63.12845       1.7514558   2.012078      1.3221084
## Marginal otherworker   984.12621      20.4769823  21.317746     19.3412224
##                      Main otherwork Marginal worker Marginal cultivator
## Main worker             10093.33729     1138.804357          35.6207828
## Main cultivator           151.88848       25.810814           0.9792285
## Main agri                 122.71000       27.878848           1.1392928
## Main household            179.37698       22.725424           0.7627016
## Main otherwork           9639.36182     1062.389272          32.7395599
## Marginal worker          1062.38927      167.277124           5.9752808
## Marginal cultivator        32.73956        5.975281           0.2489597
## Marginal agri              48.61664       12.693465           0.5162921
```

```
## Marginal household         58.04281       10.408754           0.3825580
## Marginal otherworker      922.99026      138.199623           4.8274711
##                       Marginal agri Marginal household Marginal otherworker
## Main worker              55.9289125         63.1284530           984.126209
## Main cultivator           2.6031472          1.7514558            20.476982
## Main agri                 3.4097313          2.0120776            21.317746
## Main household            1.2993914          1.3221084            19.341222
## Main otherwork           48.6166426         58.0428112           922.990258
## Marginal worker          12.6934653         10.4087544           138.199623
## Marginal cultivator       0.5162921          0.3825580             4.827471
## Marginal agri             1.7230044          0.9556935             9.498475
## Marginal household        0.9556935          0.6920787             8.378424
## Marginal otherworker      9.4984753          8.3784242           115.495253
chisqD(G3)$dist.col
##               [,1]         [,2]        [,3]        [,4]         [,5]
##  [1,] 0.000000e+00 0.0383223178 0.124766041 0.001829844 8.844981e-05
##  [2,] 3.832232e-02 0.0000000000 0.024930682 0.024023483 4.209011e-02
##  [3,] 1.247660e-01 0.0249306819 0.000000000 0.097405785 1.314963e-01
##  [4,] 1.829844e-03 0.0240234832 0.097405785 0.000000000 2.699354e-03
##  [5,] 8.844981e-05 0.0420901121 0.131496343 0.002699354 0.000000e+00
##  [6,] 1.740410e-02 0.0051977358 0.050376930 0.008108942 1.993815e-02
##  [7,] 4.207635e-02 0.0013662064 0.023482439 0.026598165 4.598503e-02
##  [8,] 1.983718e-01 0.0624489350 0.008554935 0.163564884 2.068317e-01
##  [9,] 4.056460e-02 0.0007302509 0.023738393 0.025485156 4.442434e-02
## [10,] 1.128441e-02 0.0095091105 0.063053401 0.004157277 1.331833e-02
##               [,6]         [,7]        [,8]         [,9]        [,10]
##  [1,] 0.0174040992 0.0420763468 0.198371826 0.0405645958 0.0112844106
##  [2,] 0.0051977358 0.0013662064 0.062448935 0.0007302509 0.0095091105
##  [3,] 0.0503769296 0.0234824386 0.008554935 0.0237383931 0.0630534007
##  [4,] 0.0081089419 0.0265981651 0.163564884 0.0254851562 0.0041572767
##  [5,] 0.0199381511 0.0459850264 0.206831718 0.0444243354 0.0133183299
##  [6,] 0.0000000000 0.0054393222 0.100382975 0.0049842354 0.0007112713
##  [7,] 0.0054393222 0.0000000000 0.060212540 0.0001926009 0.0100128540
##  [8,] 0.1003829749 0.0602125399 0.000000000 0.0607062455 0.1179863690
##  [9,] 0.0049842354 0.0001926009 0.060706246 0.0000000000 0.0094585517
## [10,] 0.0007112713 0.0100128540 0.117986369 0.0094585517 0.0000000000
rG4 <- cmdscale(chisqD(G3)$dist.col, eig = TRUE)
cG4 <- cmdscale(chisqD(G3)$dist.row, eig = TRUE)
plot(rG4$points, xlim = range(rG4$points[,1], cG4$points[,1]) ,
     ylim = range(rG4$points[,1], cG4$points[,1]))
text(rG4$points, labels = colnames(G3), cex = 1)
abline(h = 0, lty = 2,col='red')
abline(v = 0, lty = 2,col='red')
```

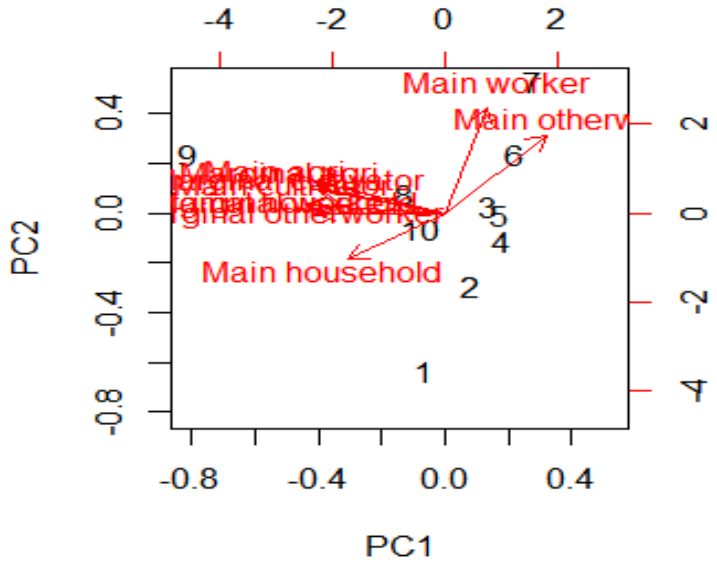


5.2.B. WORKER ANALYSIS ON CLOSELY SCATTERED AREAS IN PRINCIPLE COMPONENT ANALYSIS

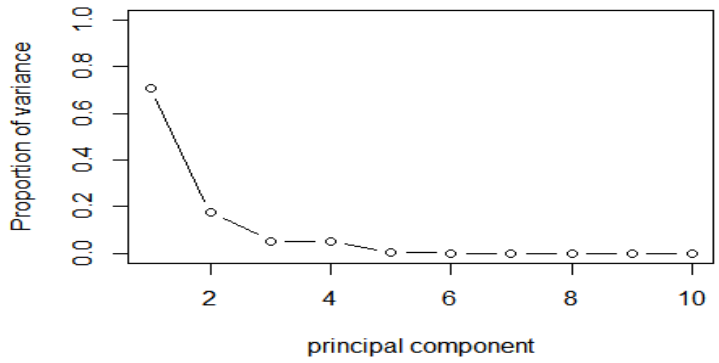
Tondiapet(1) has shown very less main worker population which makes it way different from the other cities. But otherwise the central cluster cities showing very les of other working groups and making all the Kelambakam (7), Ayyapandangal (6) as the working population group. which is very good to see with respect to the eduction analysis comparison having illiterate population. So though people are illiterte the working population is always higher.

```
row.names(chisquareworkersecond9)<-chisquareworkersecond9$Name
## Warning: Setting row names on a tibble is deprecated.
workersecond10.pca=prcomp(chisquareworkersecond9[,-1],scale=TRUE)
summary(workersecond10.pca)
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6
## Standard deviation     2.6592 1.3358 0.73396 0.72121 0.23785 0.13984
## Proportion of Variance 0.7071 0.1784 0.05387 0.05201 0.00566 0.00196
## Cumulative Proportion  0.7071 0.8856 0.93945 0.99146 0.99712 0.99907
##                           PC7     PC8       PC9      PC10
## Standard deviation     0.09538 0.01249 4.647e-10 6.143e-17
## Proportion of Variance 0.00091 0.00002 0.000e+00 0.000e+00
## Cumulative Proportion  0.99998 1.00000 1.000e+00 1.000e+00
print(workersecond10.pca)
## Standard deviations (1, .., p=10):
##  [1] 2.659200e+00 1.335827e+00 7.339589e-01 7.212081e-01 2.378534e-01
##  [6] 1.398444e-01 9.538366e-02 1.248577e-02 4.646551e-10 6.142534e-17
##
## Rotation (n x k) = (10 x 10):
##                              PC1          PC2          PC3        PC4
## Main worker           0.1111157  0.701182528  0.190066283 -0.1629585
## Main cultivator      -0.3409931  0.157333953  0.354137491  0.3189140
## Main agri            -0.3521015  0.213524816  0.106468563  0.2418340
## Main household       -0.2549012 -0.308970318  0.693221036 -0.4442323
```

```
## Main otherwork        0.2652153  0.514785330  0.021168169 -0.2302527
## Marginal worker      -0.3602148  0.050669115 -0.329272814 -0.1767329
## Marginal cultivator  -0.3328391  0.168080981  0.005649724 -0.5114752
## Marginal agri        -0.3439150  0.210092251  0.068587663  0.3849933
## Marginal household   -0.3689338  0.050667286 -0.160711078  0.1389443
## Marginal otherworker -0.3422886 -0.002268351 -0.455123687 -0.3202019
##                             PC5         PC6         PC7         PC8
## Main worker           0.1872659 -0.07118408 -0.02306295  0.10013205
## Main cultivator      -0.2660809 -0.63764625  0.36003407  0.08200500
## Main agri             0.0174496  0.52282630 -0.03436068  0.66701338
## Main household        0.3902132  0.02361303 -0.07484183 -0.03206497
## Main otherwork        0.1739983 -0.08720364 -0.07006028 -0.11054249
## Marginal worker       0.2243287 -0.06101109  0.18358407 -0.03832341
## Marginal cultivator  -0.7196329  0.19644919 -0.08309886 -0.18440242
## Marginal agri         0.2046864  0.36658722  0.13059271 -0.68530212
## Marginal household    0.1005427 -0.32583767 -0.83344685 -0.02585764
## Marginal otherworker  0.3018045 -0.15478914  0.32535599  0.13897322
##                              PC9          PC10
## Main worker          -0.6184876256 -0.0097335945
## Main cultivator       0.1345534621  0.0021175663
## Main agri             0.2021459357  0.0031813195
## Main household        0.0470214169  0.0007400107
## Main otherwork        0.7456854665  0.0117354005
## Marginal worker      -0.0125768760  0.7991540171
## Marginal cultivator   0.0006956936 -0.0442054337
## Marginal agri         0.0023171333 -0.1472348014
## Marginal household    0.0010679268 -0.0678576408
## Marginal otherworker  0.0090798463 -0.5769474633
#Standard deviation of each component
workersecond10.sd=workersecond10.pca$sdev
workersecond10.var=workersecond10.pca$sdev^2
workersecond10.var
##  [1] 7.071344e+00 1.784435e+00 5.386957e-01 5.201411e-01 5.657424e-02
##  [6] 1.955646e-02 9.098042e-03 1.558945e-04 2.159044e-19 3.773073e-33
#proportion of variance explained
pve=workersecond10.var/sum(workersecond10.var)
pve
##  [1] 7.071344e-01 1.784435e-01 5.386957e-02 5.201411e-02 5.657424e-03
##  [6] 1.955646e-03 9.098042e-04 1.558945e-05 2.159044e-20 3.773073e-34
#biplot
biplot(workersecond10.pca)
```




```
#proportion of variance explained
plot(pve,xlab="principal component",ylab="Proportion of variance" , ylim=c(0,1), type='b')
```

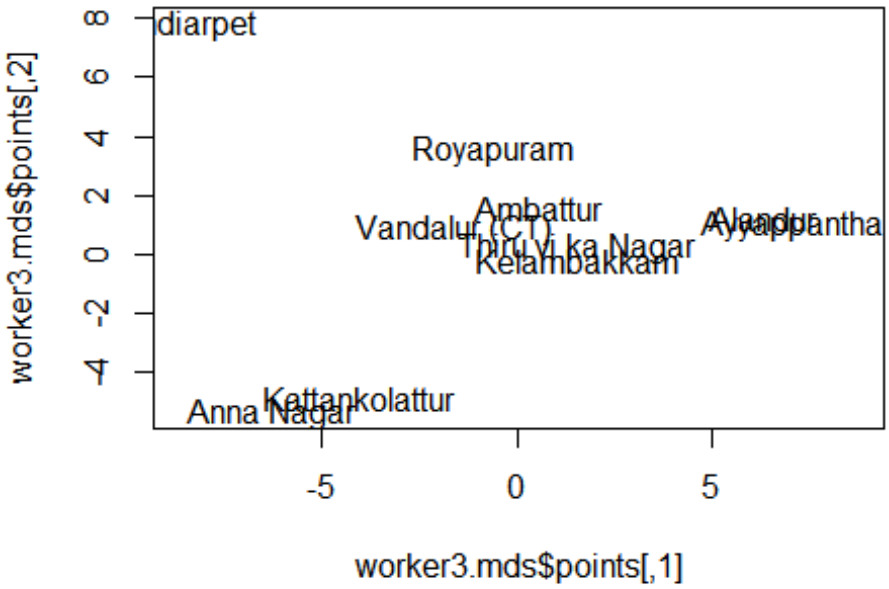


5.2.C. WORKER ANALYSIS ON CLOSELY SCATTERED AREAS IN MULTIDIMENTIONAL SCALE ANALYSIS

The analysis in the PCA with all he scattered locations exactly matches in the MDS except for the Vandalur(9) which shows all the marginal and the main cultivator making it differnt from the other locations.

```
worker3<-as.matrix(chisquareworkersecond9[,-1])
worker3.mds<-cmdscale(worker3,k=2,eig=TRUE)
worker3.mds$points
##             [,1]       [,2]
##  [1,] -8.6290284  7.8214275
##  [2,] -0.5990936  3.5696676
##  [3,]  1.5594700  0.2409979
##  [4,]  0.5922903  1.5770380
##  [5,] -6.2712916 -5.3534207
##  [6,]  6.3415587  1.2896159
##  [7,]  8.7349575  1.0256414
##  [8,]  1.5956601 -0.2131454
```

```
##  [9,] -4.0113292 -4.8214419
## [10,] -1.5884440  0.8752349
row.names(chisquareworkersecond9)<-chisquareworkersecond9$Name
## Warning: Setting row names on a tibble is deprecated.
plot(worker3.mds$points,type="n",col='red')
#text(census.mds$points,rownames(census.mds$points),col=rainbow(18))
text(worker3.mds$points,rownames(chisquareworkersecond9))
```



Since the locations are not based on the Regions I did’nt use the Hirachial Clustering. And moreover since its a one part of location based in few distance partition clustering seems like a best option for this data analysis.

6. PARTITION CLUSTERING

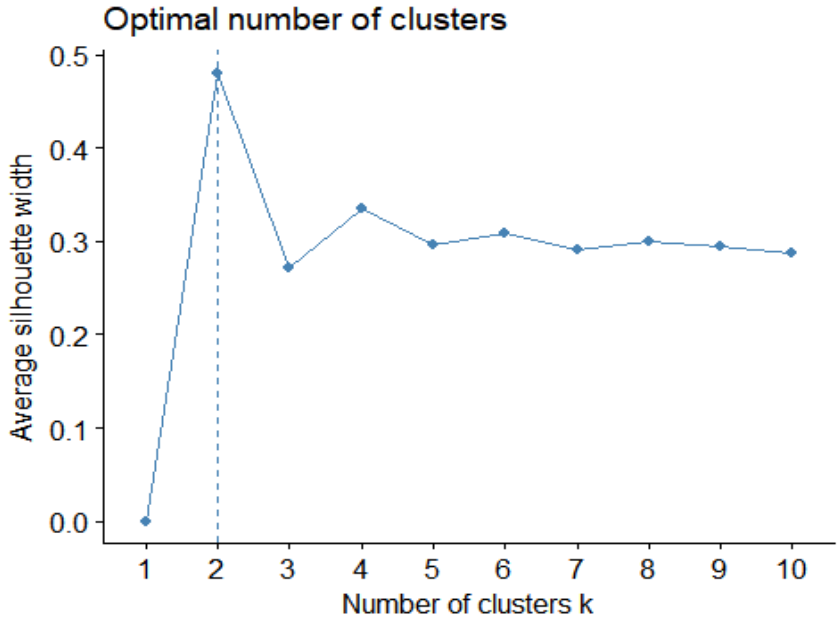
Having K-means is the simplest learning algorithms that will solve the clustering problem. The main features of k-means which make it efficient are often regarded as its Euclidean distance is used as a metric and variance as a measure of cluster scatter. The number of K is an input parameter and thats the reason its important to run diagnostic checks for determining he number of clusters in the dataset.Since it is a cluster model, the concept is based on spherical cluster center.

6.A. K-MEANS CLUSTERING FOR THE WHOLE DATASET FOR ALL LOCATIONS.

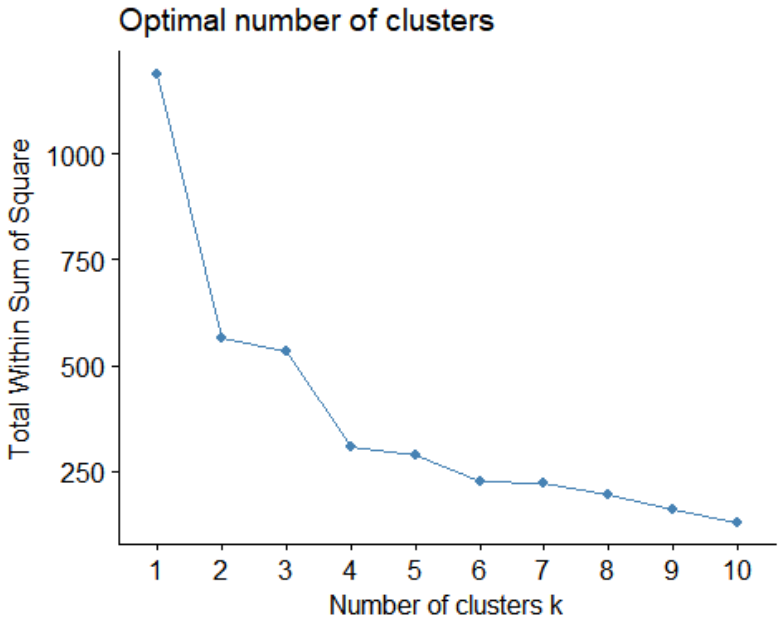
partition clustering On the whole dataset for k-means we have 2 clusters with Dim 1 having 62% and Dim 2 having 19% the two clusters fromed clearly shows that the closely related locations are grouped here on the right with blue shade and yellow shade having almost similar locations available in the PCA for making as clusters. K=2 seems like a best option from the silhouettte test. When we see the Distance graph also we can clearly see that similar locations are close to one another. blue color corresponds to small distance and red color indicates big distance between observation for the total census dataset.Also thsi plot helps us to clearly see also the difference between the cluster locations clearly mentions here as the smaller cluster (blue) is in the blue color of smaller distance in the 4th graph.

```
library(mclust)
## Warning: package 'mclust' was built under R version 3.4.4
## Package 'mclust' version 5.4
## Type 'citation("mclust")' for citing this R package in publications.
##
## Attaching package: 'mclust'
## The following object is masked from 'package:purrr':
##
##     map
library(MASS)
library(factoextra)
## Warning: package 'factoextra' was built under R version 3.4.4
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/1
3EFCZ
library(ggdendro)
## Warning: package 'ggdendro' was built under R version 3.4.4
library(dendextend)
## Warning: package 'dendextend' was built under R version 3.4.4
##
## ---------------------
## Welcome to dendextend version 1.8.0
## Type citation('dendextend') for how to cite the package.
##
## Type browseVignettes(package = 'dendextend') for the package vignette.
## The github page is: https://github.com/talgalili/dendextend/
##
## Suggestions and bug-reports can be submitted at: https://github.com/talgalili/dendexte
nd/issues
## Or contact: <tal.galili@gmail.com>
##
##  To suppress this message use:  suppressPackageStartupMessages(library(dendextend))
## ---------------------
##
## Attaching package: 'dendextend'
```

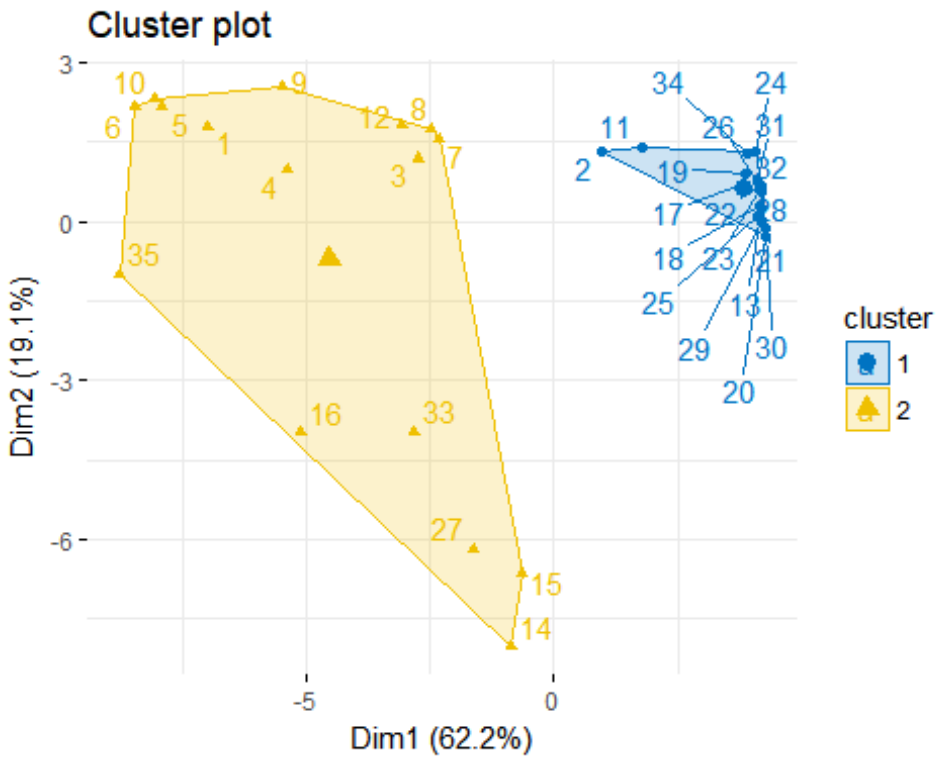
```
## The following object is masked from 'package:ggdendro':
##
##     theme_dendro
## The following object is masked from 'package:stats':
##
##     cutree
censusdata.s2<-scale(editpurpose[,c(-1)])
set.seed(123)
fviz_nbclust(censusdata.s2,kmeans,method="silhouette")
```



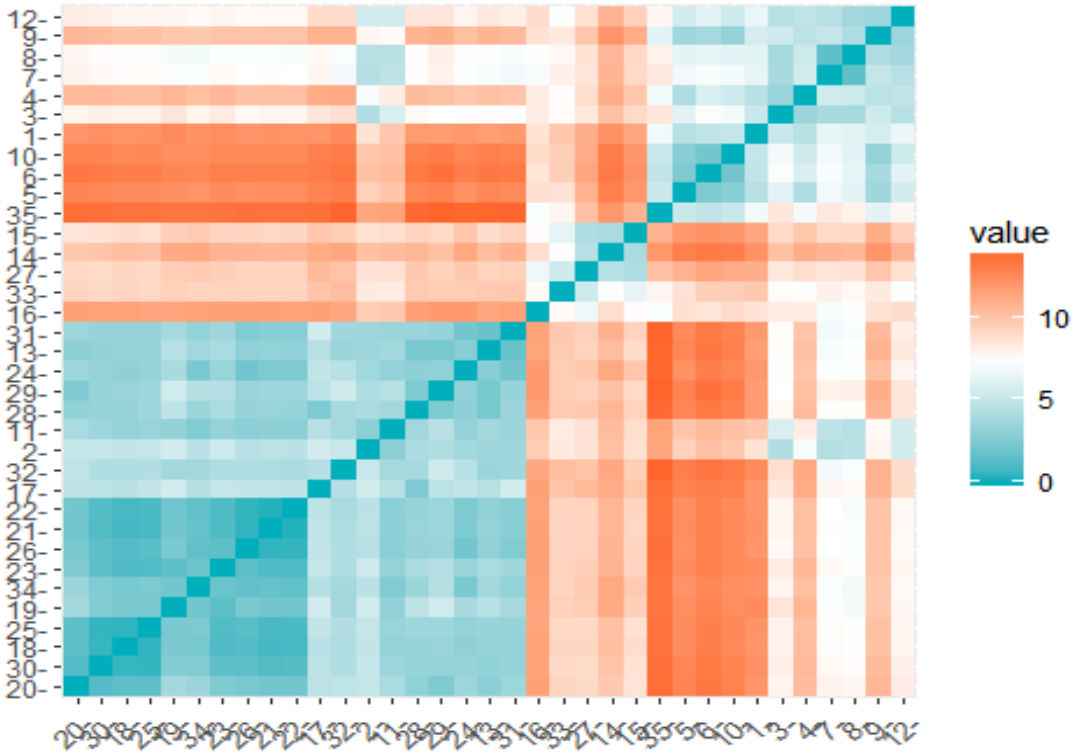

```
fviz_nbclust(censusdata.s2,kmeans,method="wss")
```




```
censusdata.k42<-kmeans(censusdata.s2,centers=2,iter.max = 100,nstart = 25)
fviz_cluster(censusdata.k42,data=censusdata.s2,
             ellipse.type="convex",palette="jco",repel=TRUE,
             ggtheme=theme_minimal())
```



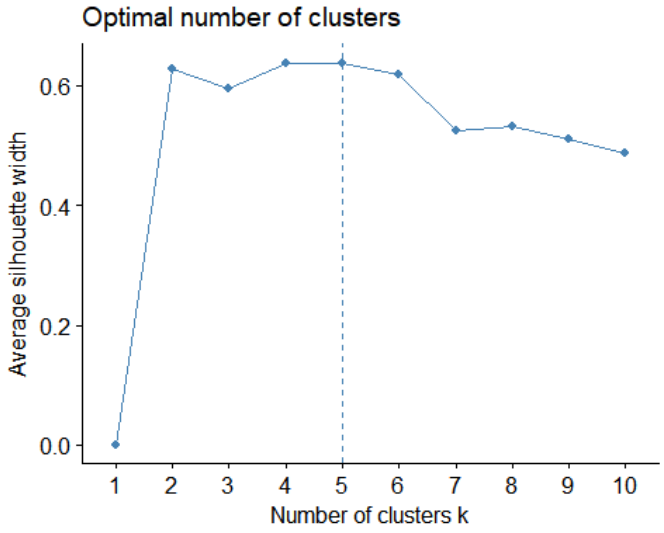

```
distance <- get_dist(censusdata.s2)
fviz_dist(distance, gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"))
```



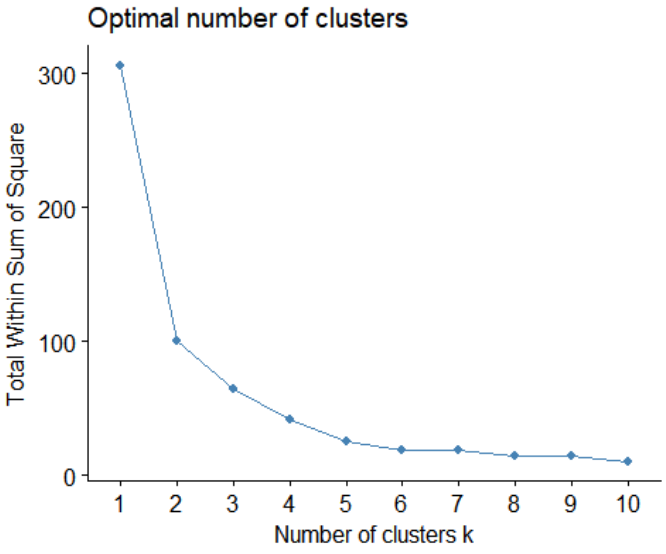
6.B. K-MEANS CLUSTERING FOR THE EDUCATION POPULATION OF THE WHOLE DATASET FOR ALL LOCATIONS.

Partition clustering On the education population for k-means we have 5 clusters with Dim 1 having 82% and Dim 2 having 12% the five clusters fromed clearly shows different proves for us that the closely related locations are grouped here on the right with blue shade and yellow shade having almost similar locations available in the PCA for making as clusters. K=2 seems like a best option from the silhouettte test. 1. Considering number 16 in the K-means which as formed a seperate cluster if we compare with the total PCA earlier it is the only place which has the more “Tribal education” in the all locations making it as a seperate cluster 2. Points 14,27 and 15 in K-means which has seperate cluster when compared with the total PCA earlier there are the locations with highest illitrate population has been seen. 3. Points 1,6,10,35 in k-means which has seperte cluster when compared with the total PCA we can see that the total caste based eduction is higher in these areas making it as a seperate cluster. 4. Most of the clusters seperation has been proved in the distance graph also with blue being the smaller distance and from 35 poin the the graph it is of longer distance.

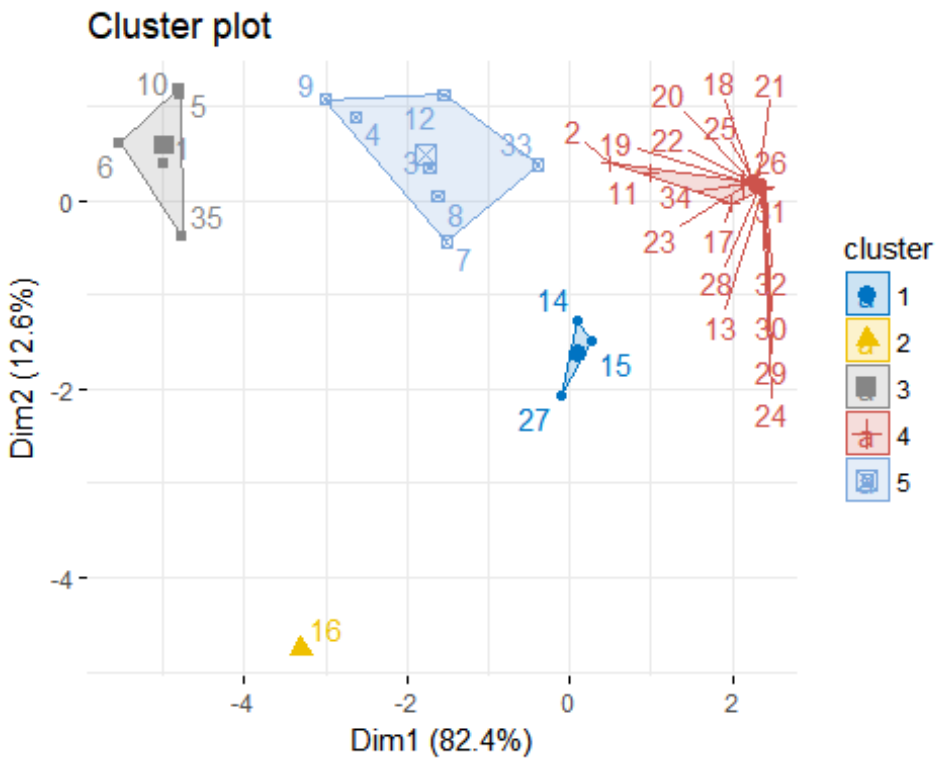
```
censusdata.s1<-scale(editpurpose[,c(6,7,8,10,13,15,19,23,35)])
set.seed(123)
fviz_nbclust(censusdata.s1,kmeans,method="silhouette")
```



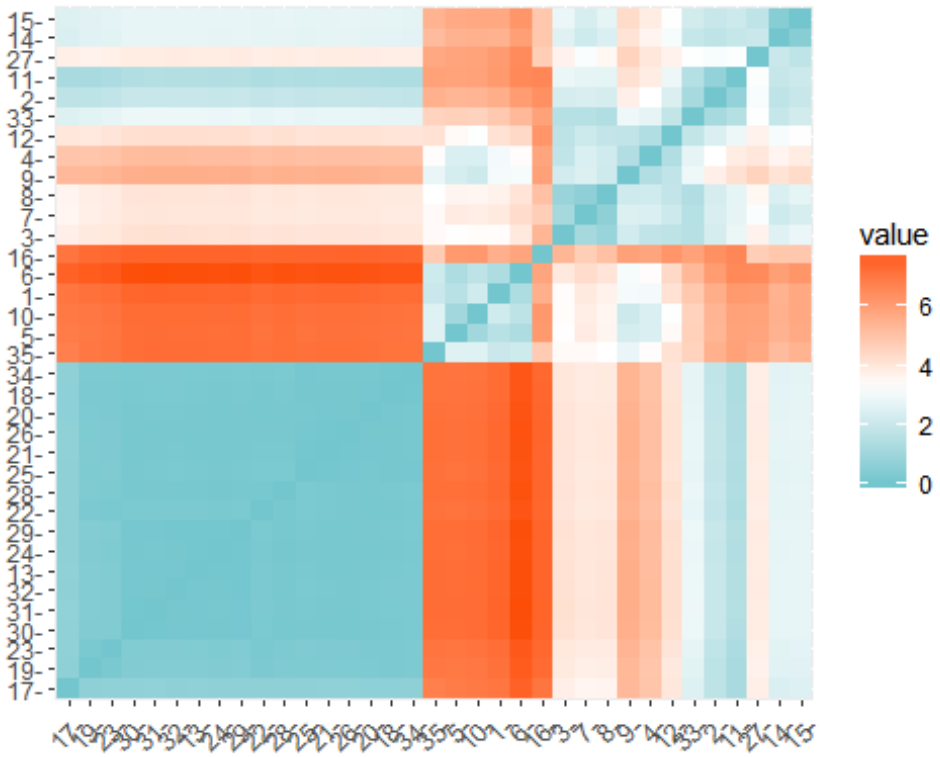

```
fviz_nbclust(censusdata.s1,kmeans,method="wss")
```



```
censusdata.k41<-kmeans(censusdata.s1,centers=5,iter.max = 100,nstart = 25)
fviz_cluster(censusdata.k41,data=censusdata.s1,
             ellipse.type="convex",palette="jco",repel=TRUE,
             ggtheme=theme_minimal())
```



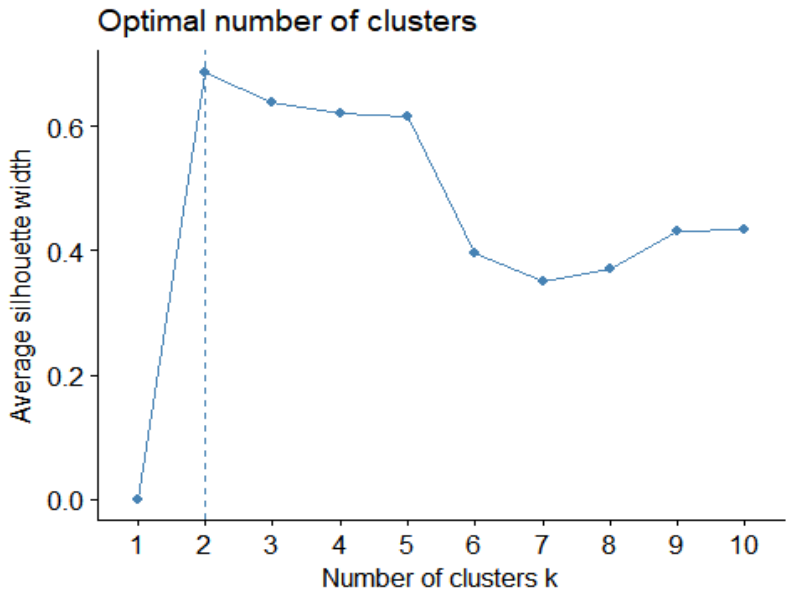

```
distance <- get_dist(censusdata.s1)
fviz_dist(distance, gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"))
```



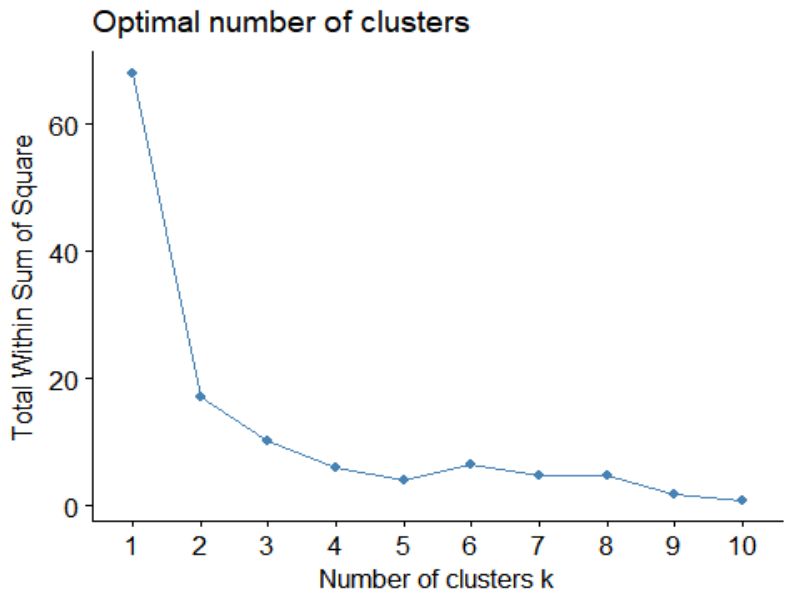
6.C. K-MEANS CLUSTERING FOR THE WORKER POPULATION OF THE WHOLE DATASET FOR ALL LOCATIONS.

Partition clustering On the worker population for k-means we have 2 clusters. clusters fromed clearly shows different proves for us that the scattered related locations are grouped here on the right with yellow shade and blue shade having almost similar locations available in the PCA for making as clusters. K=2 seems like a best option from the silhouettte test. 1. Considering number 14(Walajabad) &15(Sriprampudrur) was barely about to join the right(yellow) cluster when compared to the total k-means partition. This might be because these both locations have wide range of difference in the PCA earlier and that makes it close during the cluster part.

```
censusdata.s<-scale(editpurpose[,c(25,34)])
set.seed(123)
fviz_nbclust(censusdata.s,kmeans,method="silhouette")
```

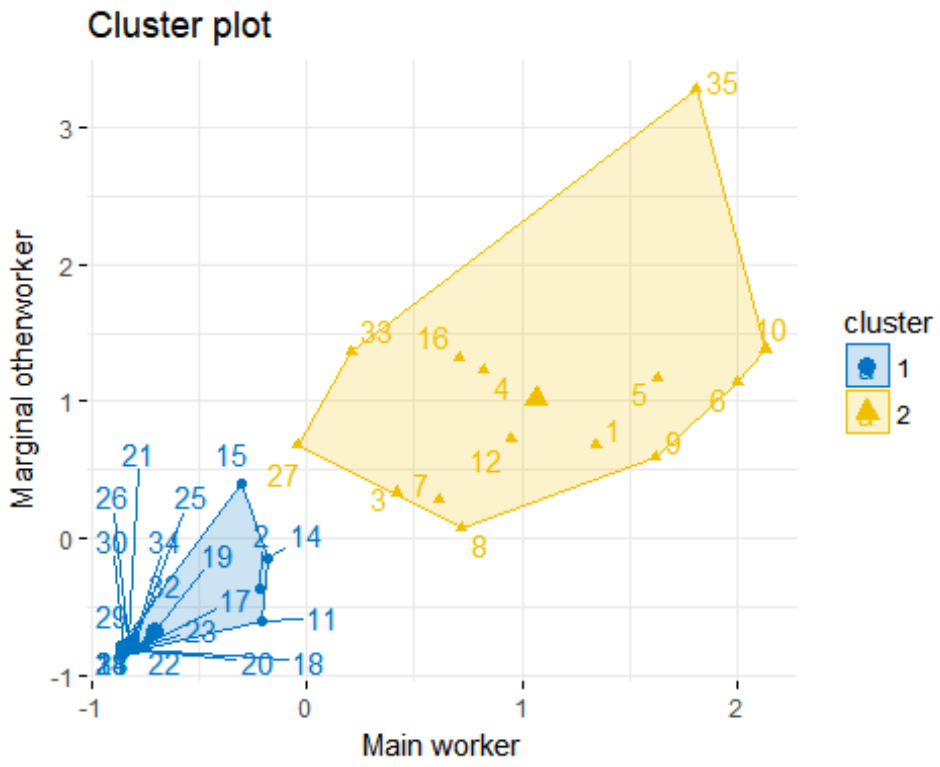



```
fviz_nbclust(censusdata.s,kmeans,method="wss")
```

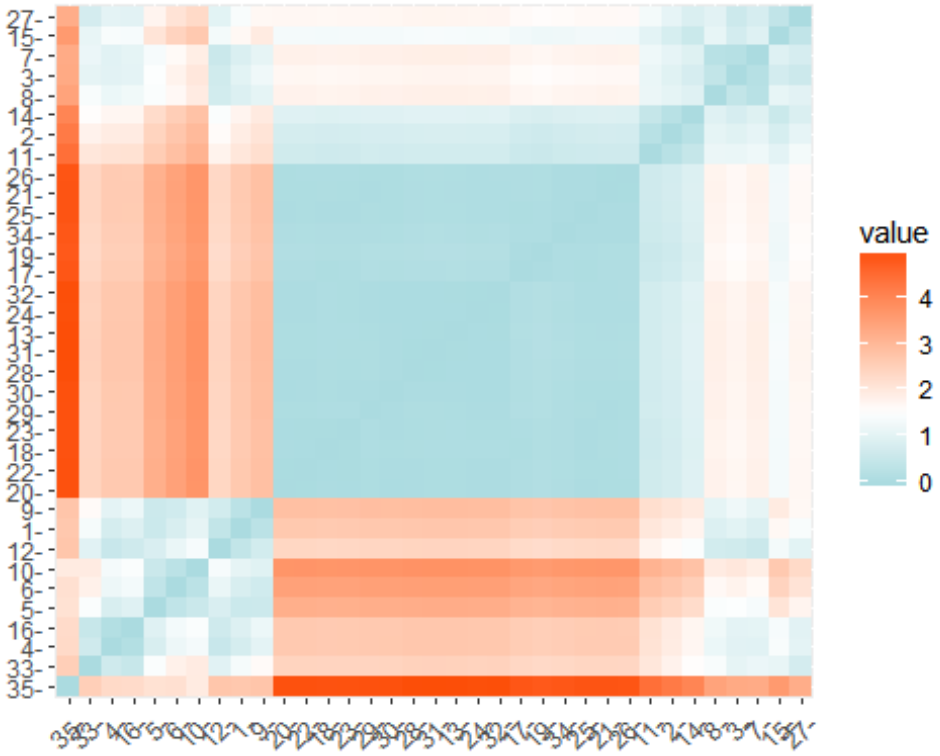



```
censusdata.k4<-kmeans(censusdata.s,centers=2,iter.max = 100,nstart = 25)
fviz_cluster(censusdata.k4,data=censusdata.s,
```

```
            ellipse.type="convex",palette="jco",repel=TRUE,
            ggtheme=theme_minimal())
```




```
distance <- get_dist(censusdata.s)
fviz_dist(distance, gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"))
```



7. SUMMARY

In this dataset for census the Y-parameter has taken as only one variable that is the Total Population Persons in the district. Which means analyzing with MANOVA is not possible. But in this scenario we can either conduct the ANOVA or regression analysis. Since all the locations is based on population parameters I have considered to convert all the subgroups adding to to the total population parameter and converted into the proportions.

So ANOVA is done with the parameters for all the population first and seen for the significant variables affecting the total population difference in the district Those are 1. Y2 - Number of Households 2. Y4 - Total Population Male 3. Y5 - Total Population Female 4. Y6 - Population education age group (0-6) 5. Y7 - Population education age group (7-13) 6. Y9 - Caste based People 7. Y13 - Caste based education for the people 8. Y15 - Literate Population 9. Y28 - Main Household industry working groups 10. Y32 - Marginal Agriculture working groups

```
data.url = "https://docs.google.com/spreadsheets/d/16Ju0LWxCntYuFHMFHB4d7fknE-I6gEi2-EY02
Cm7PXs/edit#gid=1139834335"

#my_sheets = gs_ls()
data = data.url %>%
  gs_url() %>%
  gs_read()
## Sheet-identifying info appears to be a browser URL.
## googlesheets will attempt to extract sheet key from the URL.
## Putative key: 16Ju0LWxCntYuFHMFHB4d7fknE-I6gEi2-EY02Cm7PXs
## Sheet successfully identified: "editpurposeys.xlsx"
## Accessing worksheet titled 'Sheet1'.
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   y1 = col_character(),
##   y11 = col_double(),
##   y14 = col_double(),
##   y16 = col_double(),
##   y20 = col_double(),
##   y24 = col_double(),
##   y36 = col_double()
## )
## See spec(...) for full column specifications.
editpurposeys=data
editpurposeys
## # A tibble: 35 x 36
##    y1         y2     y3     y4     y5    y6    y7    y8     y9   y10   y11
##    <chr>   <int>  <int>  <int>  <int> <int> <int> <int>  <int> <int> <dbl>
##  1 Thiru~  98413 442709 224393 218316 52652 52996 52308  53815 14618 27.2
##  2 Manal~  29078 139195  70015  69180 14053 13897 11299   9665  2055 21.3
##  3 Madha~  53910 275941 143945 131996 29450 28945 21589  71033  8197 11.5
##  4 Tondi~  78152 372737 189565 183172 41820 31437 30258 103463  1534  1.48
##  5 Royap~ 114760 511781 259640 252141 51044 50223 49874  69678  7170 10.3
##  6 Thiru~ 124848 542132 277378 264754 52919 56987 51238  61025 14047 23.0
##  7 Ambat~  56158 290815 153998 136817 25457 19734 15215  43464 10463 24.1
##  8 Anna ~  61560 295413 149485 145928 27122 20476 20369  42967  9984 23.2
##  9 Teyna~ 104303 455995 230856 225139 39941 20017 32896  57885  4882  8.43
## 10 Kodam~ 130893 554841 286794 268047 55414 45296 36575  46127  6769 14.7
## # ... with 25 more rows, and 25 more variables: y12 <int>, y13 <int>,
```

```
## #   y14 <dbl>, y15 <int>, y16 <dbl>, y17 <int>, y18 <int>, y19 <int>,
## #   y20 <dbl>, y21 <int>, y22 <int>, y23 <int>, y24 <dbl>, y25 <int>,
## #   y26 <int>, y27 <int>, y28 <int>, y29 <int>, y30 <int>, y31 <int>,
## #   y32 <int>, y33 <int>, y34 <int>, y35 <int>, y36 <dbl>
attach(editpurposeys)
fit<-aov(y3~y2+y4+y5+y6+y7+y8+y9+y10+y11+y12+y13+y14+y15+y16+y17+y18+y19+y20
         +y21+y22+y23+y24+y25+y26+y27+y28+y29+y30+y31+y32+y33+y34+y35+y36,data=editpurpos
eys)
summary(fit)
##             Df    Sum Sq   Mean Sq   F value   Pr(>F)
## y2           1 1.207e+12 1.207e+12 1.375e+33  < 2e-16 ***
## y4           1 1.010e+10 1.010e+10 1.151e+31  < 2e-16 ***
## y5           1 1.969e+08 1.969e+08 2.243e+29  < 2e-16 ***
## y6           1 0.000e+00 0.000e+00 2.174e+01 0.000891 ***
## y7           1 0.000e+00 0.000e+00 7.700e+00 0.019621 *
## y8           1 0.000e+00 0.000e+00 6.300e-02 0.807400
## y9           1 0.000e+00 0.000e+00 4.226e+01 6.89e-05 ***
## y10          1 0.000e+00 0.000e+00 2.037e+00 0.183994
## y11          1 0.000e+00 0.000e+00 3.000e-03 0.954794
## y12          1 0.000e+00 0.000e+00 2.226e+00 0.166538
## y13          1 0.000e+00 0.000e+00 7.159e+00 0.023266 *
## y14          1 0.000e+00 0.000e+00 2.255e+00 0.164053
## y15          1 0.000e+00 0.000e+00 1.602e+01 0.002510 **
## y16          1 0.000e+00 0.000e+00 6.800e-01 0.428753
## y17          1 0.000e+00 0.000e+00 1.207e+00 0.297602
## y23          1 0.000e+00 0.000e+00 8.450e-01 0.379580
## y24          1 0.000e+00 0.000e+00 1.528e+00 0.244705
## y25          1 0.000e+00 0.000e+00 3.000e-02 0.865096
## y26          1 0.000e+00 0.000e+00 3.721e+00 0.082564 .
## y27          1 0.000e+00 0.000e+00 3.600e-01 0.561936
## y28          1 0.000e+00 0.000e+00 1.908e+01 0.001403 **
## y31          1 0.000e+00 0.000e+00 4.070e-01 0.537924
## y32          1 0.000e+00 0.000e+00 8.507e+00 0.015387 *
## y33          1 0.000e+00 0.000e+00 3.191e+00 0.104346
## Residuals   10 0.000e+00 0.000e+00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After analyzing this I felts it will be better to analyze which varibles really effect the eduction groups in these locations. Here all the variables with regard to eduction was significant.

```
attach(editpurposeys)
## The following objects are masked from editpurposeys (pos = 3):
##
##     y1, y10, y11, y12, y13, y14, y15, y16, y17, y18, y19, y2, y20,
##     y21, y22, y23, y24, y25, y26, y27, y28, y29, y3, y30, y31,
##     y32, y33, y34, y35, y36, y4, y5, y6, y7, y8, y9
fitedu<-aov(y3~y6+y7+y8+y10+y11+y15+y19+y23+y35,data=editpurposeys)
summary(fitedu)
##             Df    Sum Sq   Mean Sq   F value Pr(>F)
## y6           1 1.196e+12 1.196e+12 3.240e+32 <2e-16 ***
## y7           1 1.278e+07 1.278e+07 3.462e+27 <2e-16 ***
## y8           1 1.221e+09 1.221e+09 3.307e+29 <2e-16 ***
## y10          1 1.310e+09 1.310e+09 3.549e+29 <2e-16 ***
## y11          1 1.048e+08 1.048e+08 2.839e+28 <2e-16 ***
## y15          1 1.811e+10 1.811e+10 4.907e+30 <2e-16 ***
## y19          1 7.689e+08 7.689e+08 2.083e+29 <2e-16 ***
```

```
## y23          1 0.000e+00 0.000e+00 5.200e-01  0.477
## Residuals   26 0.000e+00 0.000e+00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparitive for the education its better to analyze the working variables also with all these locations. Here only Main worker(y25), Main household(y28) & Marginal agri (y32) makes significant difference within these population locations of this district.

```
attach(editpurposeys)
## The following objects are masked from editpurposeys (pos = 3):
##
##     y1, y10, y11, y12, y13, y14, y15, y16, y17, y18, y19, y2, y20,
##     y21, y22, y23, y24, y25, y26, y27, y28, y29, y3, y30, y31,
##     y32, y33, y34, y35, y36, y4, y5, y6, y7, y8, y9
## The following objects are masked from editpurposeys (pos = 4):
##
##     y1, y10, y11, y12, y13, y14, y15, y16, y17, y18, y19, y2, y20,
##     y21, y22, y23, y24, y25, y26, y27, y28, y29, y3, y30, y31,
##     y32, y33, y34, y35, y36, y4, y5, y6, y7, y8, y9
fitwork<-aov(y3~y25+y26+y27+y28+y29+y30+y31+y32+y33+y34,data=editpurposeys)
summary(fitwork)
##             Df    Sum Sq   Mean Sq   F value   Pr(>F)
## y25          1 1.210e+12 1.210e+12 10454.350  < 2e-16 ***
## y26          1 1.703e+08 1.703e+08     1.471   0.2360
## y27          1 2.291e+08 2.291e+08     1.979   0.1713
## y28          1 3.059e+09 3.059e+09    26.426 2.32e-05 ***
## y30          1 2.246e+08 2.246e+08     1.940   0.1754
## y31          1 5.810e+07 5.810e+07     0.502   0.4850
## y32          1 6.472e+08 6.472e+08     5.591   0.0258 *
## y33          1 2.337e+08 2.337e+08     2.019   0.1673
## Residuals   26 3.010e+09 1.158e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

8. CONCLUSION

In conclusion I would say that though the city blocks are scattered in a near distance here geographical location doesn't seem to like have an effect in the employment nor the education for this population group. It is certain categorized people in each areas making the difference. Categorized people may be based on Poor or language based or religion based or also work based. So in total compared to real statistics in the government websites we can see tha only normal census in india doen't really help in analyzing the whole population based on the employment and education. There is some other significant variables which is affecting these groups making it different. This could be a vital analyzis between the group of this location. I have not done the hiracheal clustering because if its divided into regions i would have done it. I have not done MANOVA instead ANOVA because there is only one response variable total population. The general understanding of the data graphs can be done here,but in these kind of data creating Pairs graph with more variables took lot of time to export the image out so having more in these cases that was avoided to jump directly into the analysis.

9. FUTURE WORK

In future analyzing between the two disticts with each of same locations and groups will be a best way to use these multivariate analysis which helps us to analyze different population group of each district. I will make sure this will be done very soon and try to keep these techniques to find differences.

THE END

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.