ZhuojinLyu_midterm_data_preparation.R

lenovo

Sun Oct 18 22:06:37 2015

```
#Analysis of Community Health Status Indicators (CHSI) focused on Worcester, MA
library(rjson)
json_data <- fromJSON(file = "d5190a7b-361e-4f50-9ab4-bf526cc67124")
json_data
```

```
## $`@type`
## [1] "dcat:Dataset"
##
## $accessLevel
## [1] "public"
##
## $bureauCode
## [1] "009:20"
##
## $contactPoint
## $contactPoint$fn
## [1] "admin"
##
## $contactPoint$hasEmail
## [1] "mailto:HealthData@hhs.gov"
##
##
## $description
## [1] "<p>Community Health Status Indicators (CHSI) to combat obesity, heart disease, and cancer are ma
##
## $distribution
## $distribution[[1]]
## $distribution[[1]]$`@type`
## [1] "dcat:Distribution"
##
## $distribution[[1]]$downloadURL
## [1] "ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/CHDI/chsi_dataset.zip"
##
## $distribution[[1]]$format
## [1] "csv"
##
## $distribution[[1]]$mediaType
## [1] "application/unknown"
##
## $distribution[[1]]$title
## [1] "CSV "
##
##
##
## $identifier
## [1] "636fd15d-dd37-4592-a561-eb89dd3f4590"
##
```

```
## $keyword
##  [1] "access"           "behaviors"        "cancer"
##  [4] "chsi"             "community"        "community health"
##  [7] "cost"             "data"             "disease"
## [10] "environments"     "factors"          "health"
## [13] "heart"            "indicators"       "interventions"
## [16] "life expectancy"  "measurable"       "mortality"
## [19] "obesity"          "performance"      "prevalence"
## [22] "quality"          "risk"             "socioeconomic"
## [25] "warehouse"
##
## $language
## [1] "en"
##
## $modified
## [1] "2015-06-01"
##
## $programCode
## [1] "009:000"
##
## $publisher
## $publisher$`@type`
## [1] "org:Organization"
##
## $publisher$name
## [1] "Centers for Disease Control and Prevention"
##
##
## $title
## [1] "Community Health Status Indicators (CHSI) to Combat Obesity, Heart Disease and Cancer"
```

```
#input important packages
library(pracma)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(reshape2)
library(useful)
```

```
##
## Attaching package: 'useful'
##
```

```
## The following objects are masked from 'package:pracma':
##
##     cart2pol, pol2cart
```

```
library(broom)
options(warn = -1)
#1. Demographics--basic analysis of Worcester county
#Part I introduces the basic situation of Worcester
#including its population, age distribution and race distribution.
demo <- read.csv("DEMOGRAPHICS.csv", header = TRUE)
pop <- data.frame(demo$Population_Size)
colnames(pop) <- "Population"
demo.wor <- filter(demo, (demo$CHSI_County_Name == "Worcester")&
                     (demo$CHSI_State_Name == "Massachusetts"))
demo.Wor <- select(demo.wor, c(3,9,12,15,18,21,24,27,30,33,36,39,42))
#get basic indicators about demographics

#get age distribution
age <- select(demo.Wor, c(5:8))
colnames(age) <- c("Under Age 19", "Age 19-64", "Age 65-84", "Age 85+")
age1 <- melt(age, value.name = "Age")
```

```
## No id variables; using all as measure variables
```

```
age1
```

```
##       variable  Age
## 1 Under Age 19 25.5
## 2    Age 19-64 62.2
## 3    Age 65-84 10.1
## 4      Age 85+  2.2
```

```
#get race distribution
race <- select(demo.Wor, c(9:length(demo.Wor)))
race1 <- melt(race, value.name = "Race")
```

```
## No id variables; using all as measure variables
```

```
race1
```

```
##          variable Race
## 1           White 91.2
## 2           Black  3.7
## 3 Native_American  0.3
## 4           Asian  3.7
## 5        Hispanic  7.6
```

```
#2. Risk Factors for Premature Death
#Part II focus on risk factors and medical insurance of Worcester County.
#So I mainly use the RISKFACTORSANDACCESSTOCARE.csv file, and then clean and organize it.
#Finally, use K-means clustering to check whether the organization achieve the goal that
```

```r
#the file is able to used for exploration and simple modeling.

risk <- read.csv("RISKFACTORSANDACCESSTOCARE.csv", header = TRUE)
#combine risk factors and population variable
risk <- cbind(risk, pop)
#check whether there are NA and outliers
head(which((risk == -1111)|(risk == -1111.1)|(risk == -1)))
```

```
## [1] 18849 18850 18852 18856 18858 18860
```

```r
head(which(risk == -9999))
```

```
## integer(0)
```

```r
head(which((risk == -2222)|(risk == -2222.2)|(risk == -2)))
```

```
## [1] 75634 75702 75931 78593 78598 78606
```

```r
#So I need to clean -1111.1, -2222 and -2, those data.


#2.1 data cleaning
risk1 <- subset(risk, (risk$No_Exercise != -1111.1)&(risk$Few_Fruit_Veg != -1111.1)&
                  (risk$Obesity != -1111.1)&(risk$High_Blood_Pres != -1111.1)&
                  (risk$Smoker != -1111.1)&(risk$Diabetes != -1111.1)&
                  (risk$Uninsured != -1111.1)&(risk$Elderly_Medicare != -1111.1)&
                  (risk$Prim_Care_Phys_Rate != -1111.1)&(risk$Dentist_Rate != -1111.1)&
                  (risk$Community_Health_Center_Ind != -1111.1)&(risk$HPSA_Ind != -1111.1)&
                  (risk$No_Exercise != -2222)&(risk$Few_Fruit_Veg != -2222)&
                  (risk$Obesity != -2222)&(risk$High_Blood_Pres != -2222)&
                  (risk$Smoker != -2222)&(risk$Diabetes != -2222)&
                  (risk$Uninsured != -2222)&(risk$Elderly_Medicare != -2222)&
                  (risk$Prim_Care_Phys_Rate != -2222)&(risk$Dentist_Rate != -2222)&
                  (risk$Community_Health_Center_Ind != -2222)&(risk$HPSA_Ind != -2222)&
                  (risk$No_Exercise != -2222.2)&(risk$Few_Fruit_Veg != -2222.2)&
                  (risk$Obesity != -2222.2)&(risk$High_Blood_Pres != -2222.2)&
                  (risk$Smoker != -2222.2)&(risk$Diabetes != -2222.2)&
                  (risk$Uninsured != -2222.2)&(risk$Elderly_Medicare != -2222.2)&
                  (risk$Prim_Care_Phys_Rate != -2222.2)&(risk$Dentist_Rate != -2222.2)&
                  (risk$Community_Health_Center_Ind != -2222.2)&(risk$HPSA_Ind != -2222.2))

# cleaning check
head(which((risk1 == -1111)|(risk1 == -1111.1)|(risk1 == -1)))
```

```
## integer(0)
```

```r
head(which((risk1 == -2222)|(risk1 == -2222.2)|(risk1 == -2)))
```

```
## integer(0)
```

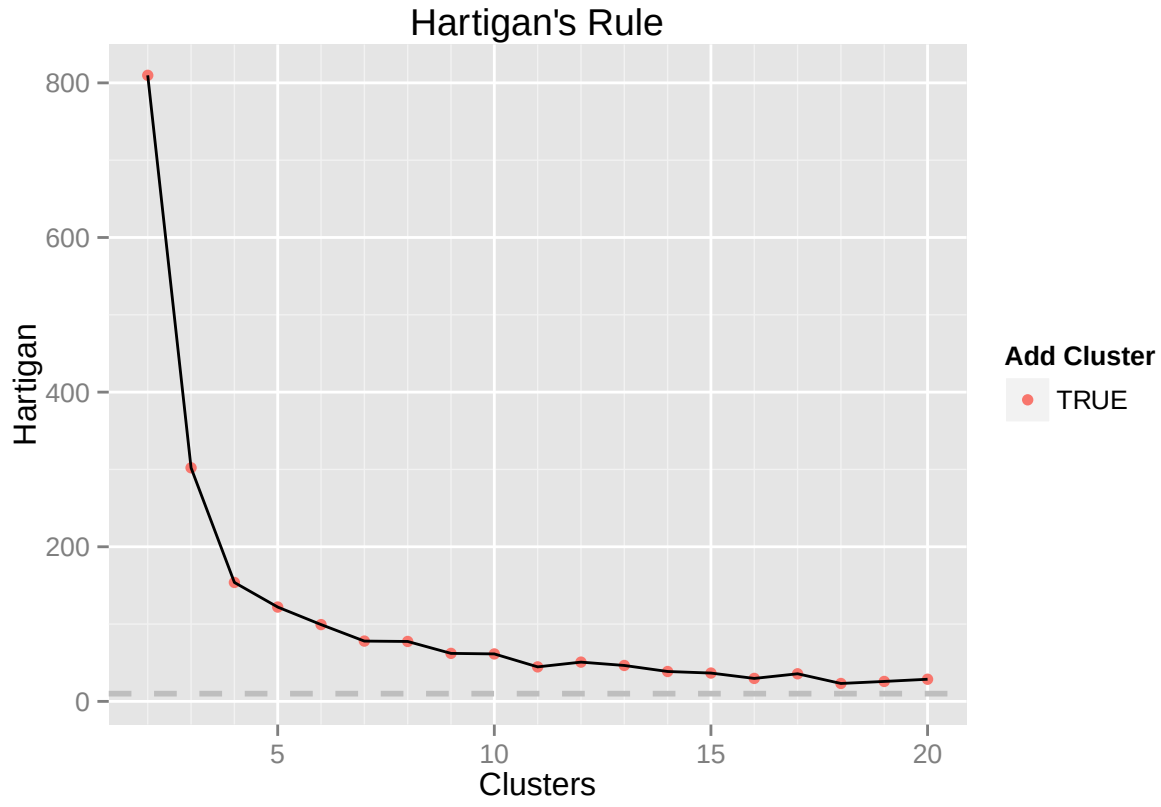
```r
#Cleaning completes


#2.2 data organization
# Because the dataset obtains two parts, risk factors and accessible care, I need to
# select risk factors part.
# And the former six columns are common indicators, so each database should have them
riskFactor <- risk1[, c(1:24)]

aCare <- risk1[, c(1:6,25:length(risk1))]

# Worcester county's risk factors and accessible care

#risk factors
Wor.risk <- filter(riskFactor, (riskFactor$CHSI_County_Name == "Worcester")&
                         (riskFactor$CHSI_State_Name == "Massachusetts"))
Wor.risk <- Wor.risk[,c(7,10,13,16,19,22)]
colnames(Wor.risk) <- c("NoExercise", "Few fruits/vegetables", "Obesity",
                        "Blood prssure", "Smoker", "Diabetes")
Wor.risk <- melt(Wor.risk, value.name = "Risk_Factors")
```

```
## No id variables; using all as measure variables
```

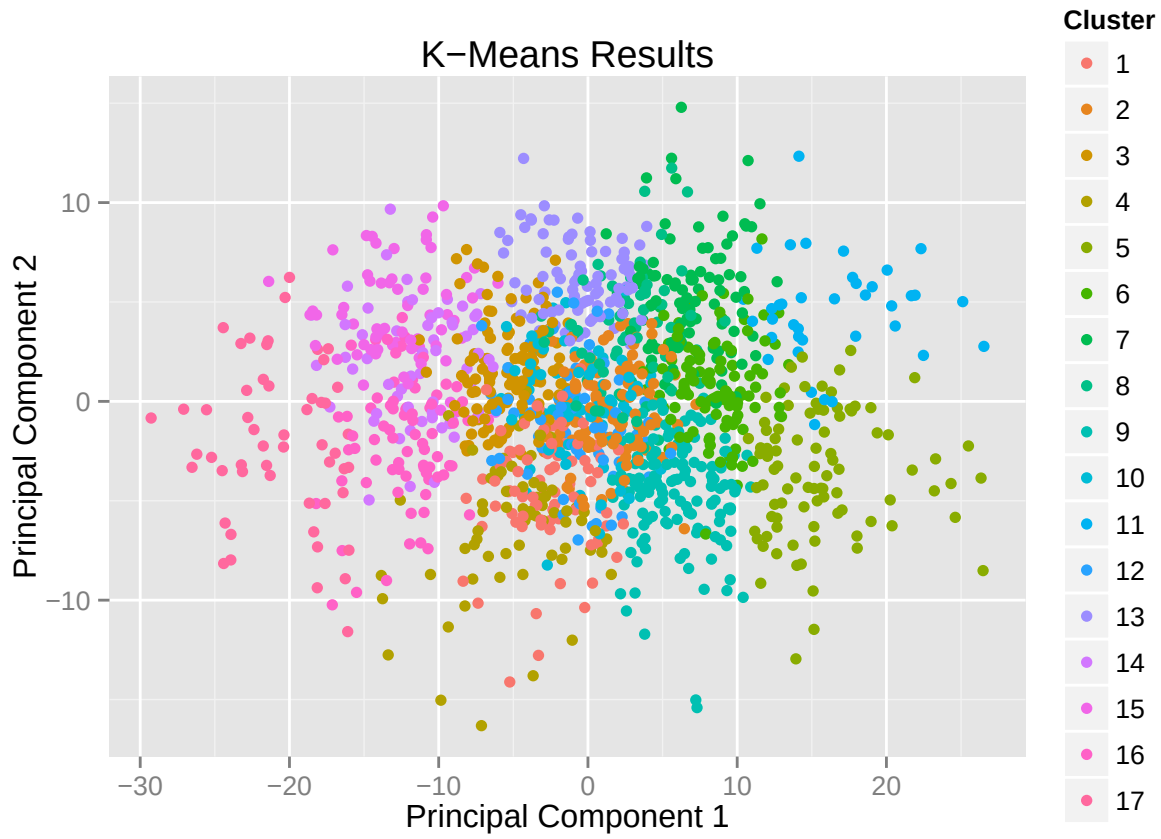
```r
Wor.risk
```

```
##                variable Risk_Factors
## 1            NoExercise         23.8
## 2 Few fruits/vegetables         71.5
## 3               Obesity         21.9
## 4         Blood prssure         24.4
## 5                Smoker         21.0
## 6              Diabetes          6.3
```

```r
#2.3 k-means clustering for risk factors
# Because the former six columns are categorical or county id,
# to avoid correlation I need to delete them before k-means cluster
# And confidence interval do not assist us to do clustering analysis,
# so we can delete confidence interval too.

riskFactor1 <- riskFactor[, c(3,4,7,10,13,16,19,22)]
riskFactor2 <- riskFactor1[, c(3:length(riskFactor1))]

#First, we need to determine the number of centers. Here I choose to ways. The first one
#is Hartigan's Rule and the other is Gap Statistic

#1.Hartigan's Rule
fitBest <- FitKMeans(riskFactor2, max.clusters = 20, nstart = 25, seed = 30000)
PlotHartigan(fitBest)
```




```
# According to Hartigan's Rule, it's appropriate to choose 17 clusters

#2.Gap Statistic: measures difference between reality and expectation
#(This process costs a relatively long time to run, so I make them into comments)

#library(cluster)

#theGap <- clusGap(riskFactor2, FUNcluster = pam, K.max = 20)
#gapDF <- as.data.frame(theGap$Tab)
#gapDF
#ggplot(gapDF, aes(x=1:nrow(gapDF))) +
#   geom_line(aes(y = gap), color = "red") +
#   geom_point(aes(y = gap), color = "red") +
#   geom_errorbar(aes(ymin = gap-SE.sim, ymax = gap+SE.sim), color="red") +
#  labs(x="Number of Clusters", y="Gap")


# The optimal number of clusters is te smallest number producing a gap
# within one standard deviation of the number of clusters that minimizes the gap
# So the number of clusters should be 17

#get K-means clustering of risk factors
seed = 30000
fit <- kmeans(x = riskFactor2, centers = 17, nstart = 25)
#visulization of k-means cluster
plot(fit, data = riskFactor2)
```




```
#glance() summarizes the total within and between sum of squares of the clustering
#totss returns the total sum of squares
#tot.withinss returns the total between-cluster sum of squares
#betweenss returns the total between-cluster sum of squares
glance(fit)
```

```
##      totss tot.withinss betweenss iter
## 1 204740.6     57075.58    147665    6
```

```
#Because the between sum of squares is 72.12%(betweenss/totss) of total sum of squares,
#it means that this is a discernible patter of clustering

#get original data and their cluster
riskFactor.Fin <- augment(fit, riskFactor1)
riskFactor.Wor <- subset(riskFactor.Fin, (riskFactor.Fin$CHSI_County_Name == "Worcester")&
                           (riskFactor.Fin$CHSI_State_Name == "Massachusetts"))
riskFactor.Wor
```

```
##      .rownames CHSI_County_Name CHSI_State_Name No_Exercise Few_Fruit_Veg
## 1229      1229        Worcester   Massachusetts        23.8          71.5
##      Obesity High_Blood_Pres Smoker Diabetes .cluster
## 1229    21.9            24.4     21      6.3        9
```

```r
#So we know the cluster of Worcester MA is 9.

riskFactor.nine <- subset(riskFactor.Fin, riskFactor.Fin$.cluster == 9)
Name.nine <- select(riskFactor.nine, 2:3)
head(Name.nine)
```

```
##     CHSI_County_Name CHSI_State_Name
## 96           Cochise         Arizona
## 102         Maricopa         Arizona
## 105             Pima         Arizona
## 108          Yavapai         Arizona
## 191     Contra Costa      California
## 203      Los Angeles      California
```

```r
#Now we get counties in USA within the same cluster with Worcester MA
#considering risk factors


#ps: Dr. Haviland, I'm sorry I cannot reduce the report into 5 pages because
#there are several graphs taking up too much space. And I keep something important
#and necessary, in my opinion. I hope you can understand. If you have some good ideas
#to better organize the report, it's pleased to hear from you.
```