Ciggaret Correlation

Introdcution

Eleazar Lopez EID:el28789

The datasets were two ciggaret datasets with some variables including states, income,

packs and cpi. I was interested in these datasets to see if there are any correlations in

variables. I expect income and packs to have an inverse correlation.

I pulled the datasets from github

https://vincentarelbundock.github.io/Rdatasets/datasets.html (https://vincentarelbundock.github.io/Rdatasets/datasets.html)

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

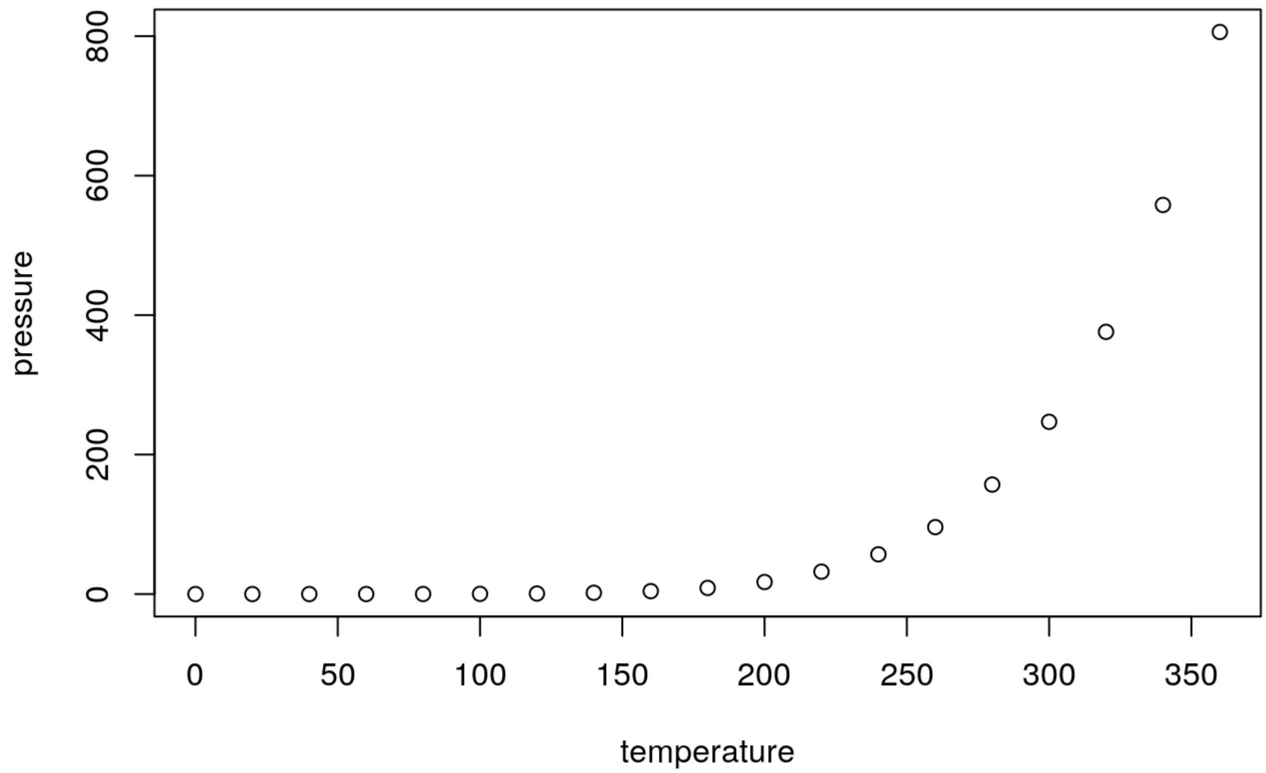
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



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

Tidy

```
#used library to load the tidyverse package
# datasets were already tidy
# used dypyr functions of rename to rename the x1 column to state
# used dyplyr functions of select to get rid of the column of X1 which was useless
#used left_join and joined tidycigA and tidycigB by the variable of state and
#created the joinedcig dataset
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.3     ✓ purrr   0.3.4
## ✓ tibble  3.1.6     ✓ dplyr   1.0.7
## ✓ tidyr   1.2.0     ✓ stringr 1.4.0
## ✓ readr   1.4.0     ✓ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)
CigarettesB <- read_csv("CigarettesB.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
##
## ── Column specification ────────────────────────────────────────────────────────
## cols(
##   X1 = col_character(),
##   packs = col_double(),
##   price = col_double(),
##   income = col_double()
## )
```

```
CigarettesSW <- read_csv("CigarettesSW.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
##
## ── Column specification ────────────────────────────────────────────────────────
## cols(
##   X1 = col_double(),
##   state = col_character(),
##   year = col_double(),
##   cpi = col_double(),
##   population = col_double(),
##   packs = col_double(),
##   income = col_double(),
##   tax = col_double(),
##   price = col_double(),
##   taxs = col_double()
## )
```

```
tidycigB <- CigarettesB
tidycigA <- CigarettesSW
tidycigB <- tidycigB %>%
 rename(state = X1)
tidycigA <- tidycigA %>%
 select(-c(X1))
head(tidycigA)
```

```
## # A tibble: 6 × 9
##   state  year   cpi population packs    income   tax price  taxs
##   <chr> <dbl> <dbl>      <dbl> <dbl>     <dbl> <dbl> <dbl> <dbl>
## 1 AL     1985  1.08    3973000  116.  46014968  32.5 102.   33.3
## 2 AR     1985  1.08    2327000  129.  26210736  37   101.   37
## 3 AZ     1985  1.08    3184000  105.  43956936  31   109.   36.2
## 4 CA     1985  1.08   26444000  100. 447102816  26   108.   32.1
## 5 CO     1985  1.08    3209000  113.  49466672  31    94.3  31
## 6 CT     1985  1.08    3201000  109.  60063368  42   128.   51.5
```

```
joinedcig <- left_join(tidycigA, tidycigB, by = "state")
head(joinedcig)
```

```
## # A tibble: 6 × 12
##   state  year   cpi population packs.x  income.x   tax price.x  taxs packs.y
##   <chr> <dbl> <dbl>      <dbl>   <dbl>     <dbl> <dbl>   <dbl> <dbl>   <dbl>
## 1 AL     1985  1.08    3973000    116.  46014968  32.5   102.   33.3    4.96
## 2 AR     1985  1.08    2327000    129.  26210736  37     101.   37      5.11
## 3 AZ     1985  1.08    3184000    105.  43956936  31     109.   36.2    4.66
## 4 CA     1985  1.08   26444000    100. 447102816  26     108.   32.1    4.50
## 5 CO     1985  1.08    3209000    113.  49466672  31      94.3  31     NA
## 6 CT     1985  1.08    3201000    109.  60063368  42     128.   51.5    4.67
## # … with 2 more variables: price.y <dbl>, income.y <dbl>
```

Exploratory Data Analysis

```
# used select function to get rid of the variable state and year as they were not
# necessary
# used function cor to see the correlation between all variables
# used chart.Correlation to view correlation histagrams and correlation coeeficients
# for all variables
joinedcig <- joinedcig %>%
  select(-c(year,state))
cor(joinedcig)
```

```
##                    cpi  population    packs.x   income.x        tax    price.x
## cpi         1.00000000  0.04758017 -0.4994643  0.2317893  0.6857145  0.9116556
## population  0.04758017  1.00000000 -0.2112834  0.9573113  0.1659856  0.1458604
## packs.x    -0.49946432 -0.21128337  1.0000000 -0.3317847 -0.6421176 -0.6524732
## income.x    0.23178932  0.95731126 -0.3317847  1.0000000  0.3372751  0.3375339
## tax         0.68571446  0.16598557 -0.6421176  0.3372751  1.0000000  0.8993727
## price.x     0.91165558  0.14586043 -0.6524732  0.3375339  0.8993727  1.0000000
## taxs        0.70412144  0.18891721 -0.6574167  0.3582307  0.9853330  0.9203278
## packs.y             NA          NA         NA         NA         NA         NA
## price.y             NA          NA         NA         NA         NA         NA
## income.y            NA          NA         NA         NA         NA         NA
##                  taxs packs.y price.y income.y
## cpi         0.7041214      NA      NA       NA
## population  0.1889172      NA      NA       NA
## packs.x    -0.6574167      NA      NA       NA
## income.x    0.3582307      NA      NA       NA
## tax         0.9853330      NA      NA       NA
## price.x     0.9203278      NA      NA       NA
## taxs        1.0000000      NA      NA       NA
## packs.y            NA       1      NA       NA
## price.y            NA      NA       1       NA
## income.y           NA      NA      NA        1
```

```
library(PerformanceAnalytics)
```

```
## Loading required package: xts
```

```
## Loading required package: zoo
```

```
## 
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
## 
##     as.Date, as.Date.numeric
```

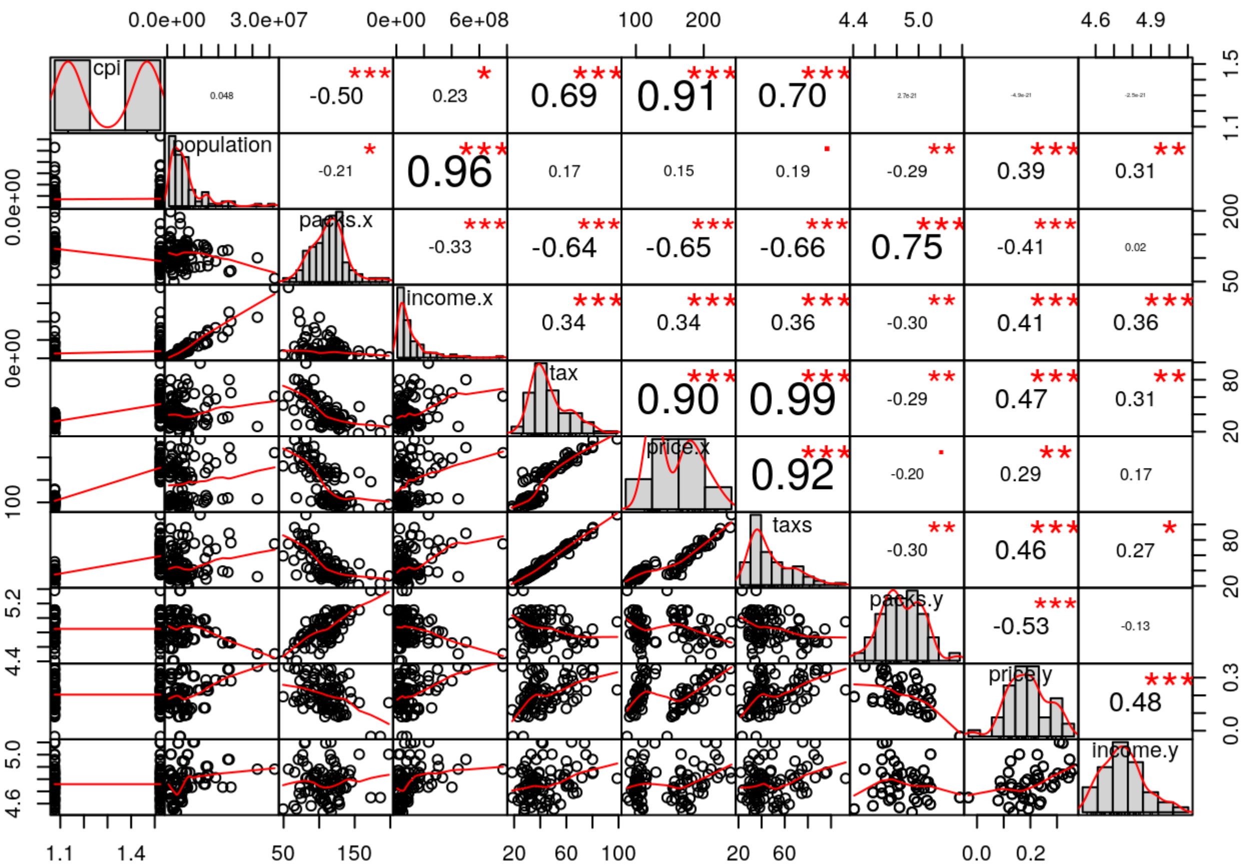
```
## 
## Attaching package: 'xts'
```

```
## The following objects are masked from 'package:dplyr':
## 
##     first, last
```

```
## 
## Attaching package: 'PerformanceAnalytics'
```

```
## The following object is masked from 'package:graphics':
##
##     legend
```

```
chart.Correlation(joinedcig, histogram = TRUE, method = "pearson")
```



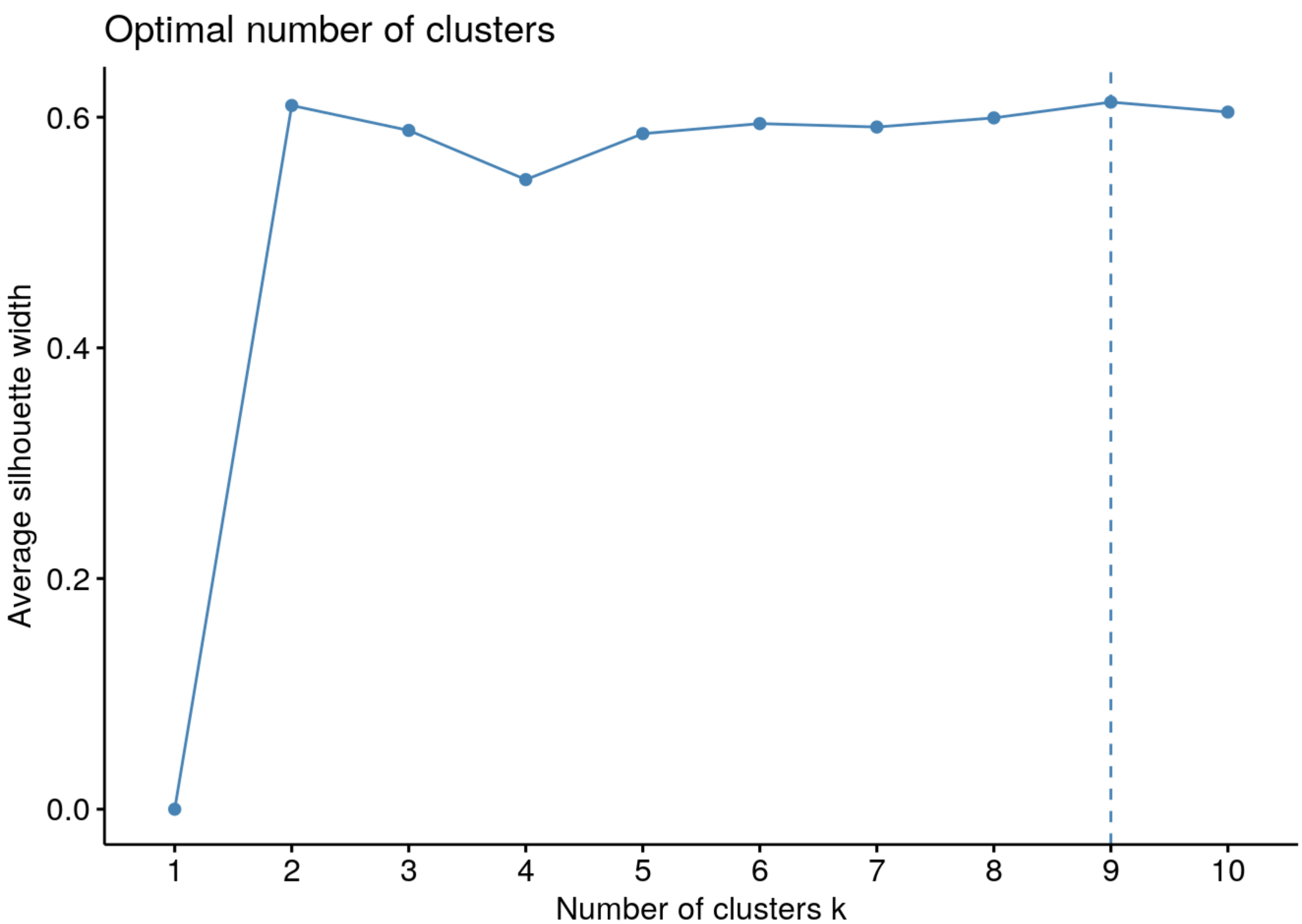

```
# The most correlated variables are tax and taxs but if we ignore that due to them both being a
 tax variable the next most correlated variables is income.x and population. The least correlate
d variables are packs.x and taxs. As packs increase the income tends to decrease.
```

Clustering

```
# used fviz_nbclust with the joinedcig dataframe to see the optimal number of clusters
# used function pam and named it pam_results
# used ggpairs to visualize the clusters
# with this vizualization we can see which variables have the strongest correlation,
# positive and negative
# we can more clearly see how tax and taxs have the strongest correlation between all
# variables
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(cluster)
fixedjoinedcig <- joinedcig %>%
  select(-c(packs.y, price.y, income.y))
fviz_nbclust(fixedjoinedcig, pam, method = "silhouette")
```



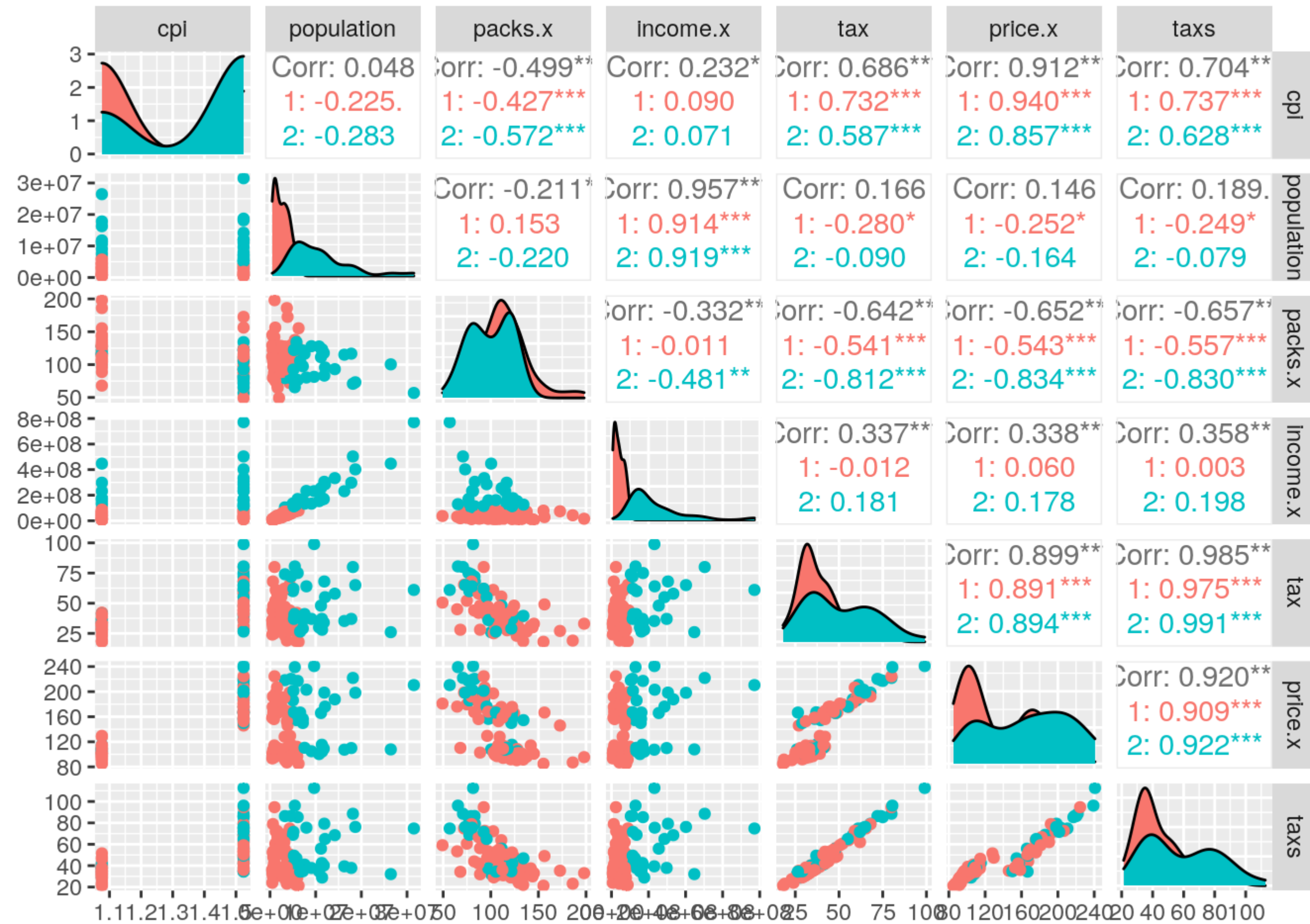

```
pam_results <- pam(fixedjoinedcig,2)
pam_results
```

```
## Medoids:
##      ID   cpi population  packs.x  income.x tax price.x   taxs
## [1,] 10 1.076    2830000 113.7456  37902896  34 101.842 37.917
## [2,]  8 1.076   11352000 122.1811 166919248  37 115.290 42.490
## Clustering vector:
##  [1] 1 1 1 2 1 1 1 2 1 1 1 2 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2 1 1 2 2 1 1 2 1 1
## [39] 1 1 2 1 1 1 1 1 1 1 1 1 1 2 1 2 1 2 2 1 1 2 2 1 1 1 2 2 1 2 2 2 1 1 2 1 1 1
## [77] 2 1 1 2 2 1 1 2 1 1 1 2 2 1 2 1 2 2 1 1
## Objective function:
##    build     swap
## 46995509 44449436
##
## Available components:
##  [1] "medoids"    "id.med"     "clustering" "objective"  "isolation"
##  [6] "clusinfo"   "silinfo"    "diss"       "call"       "data"
```

```r
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
fixedjoinedcig %>% mutate(cluster = as.factor(pam_results$clustering)) %>%
  ggpairs(columns = 1:7, aes(color = cluster))
```



Dimensionality Reduction

```
# used dyplr function of select to get rid of packs.y, price.y, and income.y
# used function prcomp to find principal components
# used the function cut to change cpi to a categorical variable
# used fviz_cluster to seperate the clusters by cpi
fixedjoinedcig <- joinedcig %>%
  select(-c(packs.y, price.y, income.y))
prcomp(fixedjoinedcig)
```

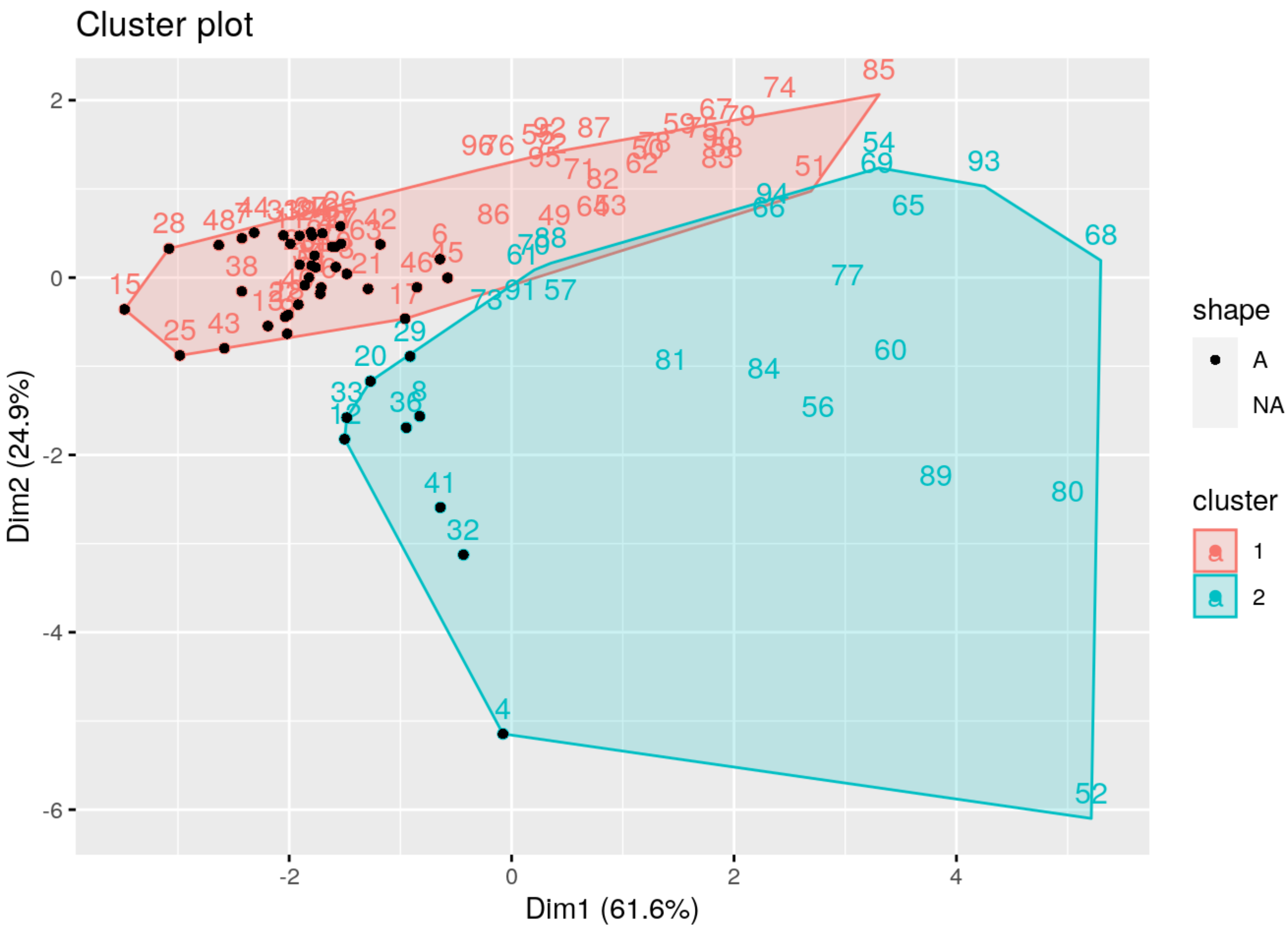
```
## Standard deviations (1, .., p=7):
## [1] 1.206537e+08 1.571692e+06 3.818919e+01 1.792198e+01 8.807293e+00
## [6] 1.993977e+00 4.766130e-02
##
## Rotation (n x k) = (7 x 7):
##                      PC1           PC2           PC3           PC4
## cpi         4.319543e-10 -8.641615e-08 -3.333794e-03 -2.831933e-03
## population  4.318892e-02  9.990669e-01 -1.993289e-05 -2.220790e-06
## packs.x    -7.109888e-08  6.058607e-06  4.028361e-01 -9.137535e-01
## income.x    9.990669e-01 -4.318892e-02  1.021067e-06  8.828357e-08
## tax         4.507149e-08 -5.575065e-06 -2.864985e-01 -9.478971e-02
## price.x     1.226514e-07 -1.712943e-05 -7.938283e-01 -3.774543e-01
## taxs        5.735084e-08 -6.556501e-06 -3.542134e-01 -1.165750e-01
##                      PC5           PC6           PC7
## cpi        -1.023253e-02 -9.556602e-03 -9.998924e-01
## population -9.157836e-07 -2.888839e-07 -1.444217e-09
## packs.x     5.247904e-02  4.813206e-03  6.617956e-04
## income.x    3.791183e-08  1.132105e-08  1.360896e-11
## tax         6.154963e-01 -7.280757e-01  1.883610e-03
## price.x    -4.750573e-01 -4.006410e-02  8.960266e-03
## taxs        6.265989e-01  6.842413e-01 -1.144094e-02
```

```
fixedjoinedcig$cpi <- cut(fixedjoinedcig$cpi,
                       breaks=c(1, 1.5),
                       labels=c("A"))
fviz_cluster(pam_results, data = fixedjoinedcig,
             shape = fixedjoinedcig$cpi) +
  geom_point(aes(shape = fixedjoinedcig$cpi)) +
  guides(shape = guide_legend(title = "shape"))
```

```
## Warning in if (shape %in% colnames(data)) {: the condition has length > 1 and
## only the first element will be used
```

```
## Warning: Removed 48 rows containing missing values (geom_point).

## Warning: Removed 48 rows containing missing values (geom_point).
```



Classification & Cross-Validation

```
# made new data set called ffjoinedcig from fixedjoinedcig by mutating and creating a new
# variable Economic _status rich was over 100,000,000 in income
# made a table with the actual and prediction of ffjoinedcig
# created an ROC plot that showed that economic_status is a perfect indicator of income
library(tidyverse)
ffjoinedcig <- fixedjoinedcig %>%
  mutate(Economic_status = ifelse(income.x > 100000000, "rich", "not rich"))
actual <- ffjoinedcig$income.x
prediction <- ffjoinedcig$Economic_status
table(actual = ffjoinedcig$income.x, prediction = ffjoinedcig$Economic_status) %>%
  addmargins()
```

```
##           prediction
## actual      not rich rich Sum
##   6887097          1    0   1
##   7116756          1    0   1
##   8340000          1    0   1
##   8672948          1    0   1
##   9785230          1    0   1
##   9927301          1    0   1
##   10293195         1    0   1
##   11577261         1    0   1
##   12243384         1    0   1
##   12448607         1    0   1
##   14229156         1    0   1
##   14454129         1    0   1
##   14575292         1    0   1
##   14581495         1    0   1
##   15767469         1    0   1
##   16296835         1    0   1
##   17258916         1    0   1
##   18237436         1    0   1
##   19462380         1    0   1
##   20852964         1    0   1
##   21778072         1    0   1
##   22868920         1    0   1
##   23786644         1    0   1
##   25045934         1    0   1
##   25678534         1    0   1
##   26210736         1    0   1
##   28649564         1    0   1
##   31716160         1    0   1
##   32611268         1    0   1
##   34784360         1    0   1
##   36205164         1    0   1
##   36293064         1    0   1
##   37278220         1    0   1
##   37902896         1    0   1
##   38536176         1    0   1
##   39377292         1    0   1
##   42703144         1    0   1
##   43395580         1    0   1
##   43956936         1    0   1
##   45995496         1    0   1
##   46014968         1    0   1
##   46241956         1    0   1
##   49466672         1    0   1
##   53431900         1    0   1
##   56626672         1    0   1
##   57749668         1    0   1
##   60063368         1    0   1
##   60170928         1    0   1
##   63152360         1    0   1
##   63333300         1    0   1
```

```
##    64846548          1    0   1
##    65732720          1    0   1
##    69341920          1    0   1
##    71209312          1    0   1
##    71751616          1    0   1
##    72050072          1    0   1
##    74079712          1    0   1
##    74851664          1    0   1
##    78364336          1    0   1
##    79104656          1    0   1
##    83903280          1    0   1
##    84572688          1    0   1
##    87361632          1    0   1
##    88870496          1    0   1
##    92946544          1    0   1
##    98328688          1    0   1
##    104315120         0    1   1
##    113216856         0    1   1
##    114259984         0    1   1
##    115959680         0    1   1
##    117639672         0    1   1
##    126525008         0    1   1
##    129680832         0    1   1
##    133549208         0    1   1
##    133728040         0    1   1
##    135115456         0    1   1
##    153455776         0    1   1
##    157633568         0    1   1
##    159800448         0    1   1
##    161441792         0    1   1
##    166919248         0    1   1
##    170033840         0    1   1
##    170051568         0    1   1
##    176786352         0    1   1
##    231003152         0    1   1
##    231594240         0    1   1
##    233208576         0    1   1
##    255312928         0    1   1
##    285923232         0    1   1
##    297728512         0    1   1
##    304767456         0    1   1
##    333525344         0    1   1
##    402096768         0    1   1
##    447102816         0    1   1
##    503163328         0    1   1
##    771470144         0    1   1
##    Sum              66   30  96
```

```
F1 <- function(y_hat, y, positive){
  sensitivity <- mean(y_hat[y == positive] == positive)
  precision <- mean(y[y_hat == positive] == positive)
  2*(sensitivity*precision)/(sensitivity + precision)
}

F1(prediction, actual, "rich")
```

```
## [1] NaN
```

```
library(plotROC)
ROC <- ggplot(ffjoinedcig) +
  geom_roc(aes(d = Economic_status, m = income.x))
ROC
```

```
## Warning in verify_d(data$d): D not labeled 0/1, assuming not rich = 0 and rich =
## 1!
```