

test

Hank Magan

12/5/2021

```
# Hank Magan
# 11.13.2021
# FinalMagan.R (final paper code)

# clean up old stuff and set working directory
rm(list=ls())
setwd("~/GitHub/IE3620/Final (12-15)")
```

```
# import libraries
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(tibble)
library(stringr)
library(ggplot2)
library(cluster)
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.1.2
```

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

```
library(class)
library(gmodels)
```

```
# import data
housingRaw <- read.csv("housing.csv")
```

```
# quick look at data
str(housingRaw)
```

```
## 'data.frame': 505 obs. of 1 variable:
## $ X0.00632..18.00...2.310..0..0.5380..6.5750..65.20..4.0900...1..296.0..15.30.396.90...4.98..24.00:
```

```
glimpse(housingRaw)
```

```
## Rows: 505
## Columns: 1
## $ X0.00632..18.00...2.310..0..0.5380..6.5750..65.20..4.0900...1..296.0..15.30.396.90...4.98..24.00 <
```

```
# data is all one column and unlabeled: clean data
names(housingRaw) <- c("jargon") # rename singular column to something writable to then replace it...
housingRaw$jargon <- str_squish(housingRaw$jargon) # remove extra spaces in data
# i.e. "Hello there how are you"
# becomes "Hello there how are you" (neat function!)

# set split column names (14 in total)
cols <- c("CRIM",      # per capita crime rate by town
          "ZN",        # proportion of residential land zoned for lots over 25,000 sq.ft
          "INDUS",     # proportion of non-retail business acres per town
          "CHAS",      # Charles River dummy variable (1 if tract bounds river; 0 otherwise)
          "NOX",       # nitric oxides concentration (parts per 10 million) [parts/10M]
          "RM",        # average number of rooms per dwelling
          "AGE",       # proportion of owner-occupied units built prior to 1940
          "DIS",       # weighted distances to five Boston employment centres
          "RAD",       # index of accessibility to radial highways
          "TAX",       # full-value property-tax rate per $10,000 [$ /10k]
          "PTRATIO",   # pupil-teacher ratio by town
          "B",         # B=1000(Bk - 0.63)^2 where Bk is the proportion of black people by town
          "LSTAT",     # % lower status of the population
          "MEDV"       # median value of owner-occupied homes in $1000's [k$]
          )
```

```
# separate raw data into new data set
housing <- tidyr::separate(data=housingRaw, col=jargon, sep=" ", into=cols, remove=TRUE)
```

```
# convert chr columns into nums + logic col
housing[, 1:14] <- sapply(housing[, 1:14], as.numeric) # numeric conversion
housing$CHAS <- factor(housing$CHAS, levels=c(0, 1), labels=c(FALSE, TRUE)) # bool conversion
```

```
# check for NAs
clean <- ifelse(complete.cases(housing) == TRUE, 1, 0)
paste("There are ", dim(housing)[1]-sum(clean), " rows with missing data")
```

```
## [1] "There are 0 rows with missing data"
```

```
# take a look at new clean data
dim(housing)
```

```
## [1] 505 14
```

```
glimpse(housing)
```

```
## Rows: 505
## Columns: 14
## $ CRIM    <dbl> 0.02731, 0.02729, 0.03237, 0.06905, 0.02985, 0.08829, 0.14455,~
## $ ZN      <dbl> 0.0, 0.0, 0.0, 0.0, 0.0, 12.5, 12.5, 12.5, 12.5, 12.5, 12.5, 1~
## $ INDUS   <dbl> 7.07, 7.07, 2.18, 2.18, 2.18, 7.87, 7.87, 7.87, 7.87, 7.87, 7.~
## $ CHAS    <fct> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,~
## $ NOX     <dbl> 0.469, 0.469, 0.458, 0.458, 0.458, 0.524, 0.524, 0.524, 0.524,~
## $ RM      <dbl> 6.421, 7.185, 6.998, 7.147, 6.430, 6.012, 6.172, 5.631, 6.004,~
## $ AGE     <dbl> 78.9, 61.1, 45.8, 54.2, 58.7, 66.6, 96.1, 100.0, 85.9, 94.3, 8~
## $ DIS     <dbl> 4.9671, 4.9671, 6.0622, 6.0622, 6.0622, 5.5605, 5.9505, 6.0821~
## $ RAD     <dbl> 2, 2, 3, 3, 3, 5, 5, 5, 5, 5, 5, 5, 4, 4, 4, 4, 4, 4, 4, 4,~
## $ TAX     <dbl> 242, 242, 222, 222, 222, 311, 311, 311, 311, 311, 311, 311, 30~
## $ PTRATIO <dbl> 17.8, 17.8, 18.7, 18.7, 18.7, 15.2, 15.2, 15.2, 15.2, 15.2, 15~
## $ B       <dbl> 396.90, 392.83, 394.63, 396.90, 394.12, 395.60, 396.90, 386.63~
## $ LSTAT   <dbl> 9.14, 4.03, 2.94, 5.33, 5.21, 12.43, 19.15, 29.93, 17.10, 20.4~
## $ MEDV    <dbl> 21.6, 34.7, 33.4, 36.2, 28.7, 22.9, 27.1, 16.5, 18.9, 15.0, 18~
```

```
summary(housing)
```

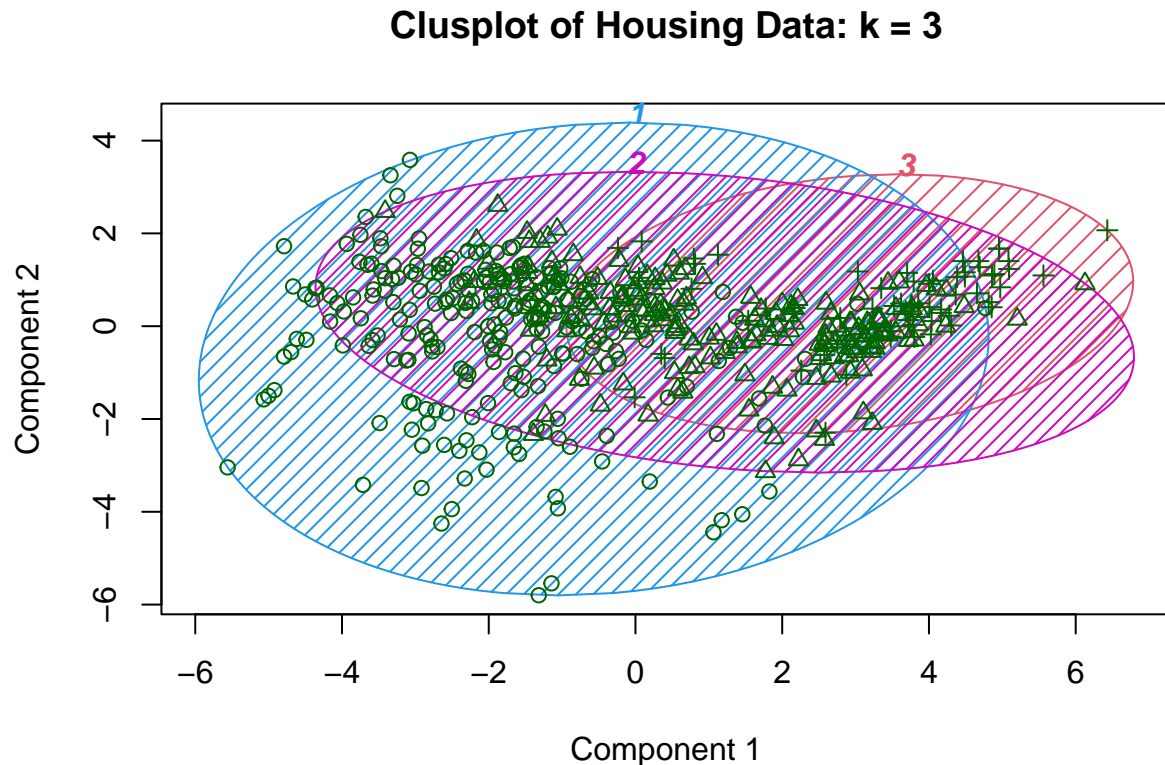
```
##           CRIM           ZN           INDUS           CHAS
## Min.      : 0.00906   Min.      : 0.00   Min.      : 0.46   FALSE:470
## 1st Qu.: 0.08221   1st Qu.: 0.00   1st Qu.: 5.19   TRUE : 35
## Median : 0.25915   Median : 0.00   Median : 9.69
## Mean     : 3.62067   Mean     : 11.35   Mean     :11.15
## 3rd Qu.: 3.67822   3rd Qu.: 12.50   3rd Qu.:18.10
## Max.     :88.97620   Max.      :100.00   Max.     :27.74
##           NOX           RM           AGE           DIS
## Min.      :0.3850   Min.      :3.561   Min.      : 2.90   Min.      : 1.130
## 1st Qu.:0.4490   1st Qu.:5.885   1st Qu.: 45.00   1st Qu.: 2.100
## Median :0.5380   Median :6.208   Median : 77.70   Median : 3.199
## Mean     :0.5547   Mean     :6.284   Mean     : 68.58   Mean     : 3.794
## 3rd Qu.:0.6240   3rd Qu.:6.625   3rd Qu.: 94.10   3rd Qu.: 5.212
## Max.     :0.8710   Max.      :8.780   Max.     :100.00   Max.     :12.127
##           RAD           TAX           PTRATIO           B
## Min.      : 1.000   Min.      :187.0   Min.      :12.60   Min.      : 0.32
## 1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.33
## Median : 5.000   Median :330.0   Median :19.10   Median :391.43
## Mean     : 9.566   Mean     :408.5   Mean     :18.46   Mean     :356.59
## 3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.21
## Max.     :24.000   Max.      :711.0   Max.     :22.00   Max.     :396.90
##           LSTAT           MEDV
## Min.      : 1.73   Min.      : 5.00
## 1st Qu.: 7.01   1st Qu.:17.00
## Median :11.38   Median :21.20
## Mean     :12.67   Mean     :22.53
## 3rd Qu.:16.96   3rd Qu.:25.00
## Max.     :37.97   Max.      :50.00
```

```
set.seed(15587933)
```

```
# analyzing factors affecting house price
```

```
housing.km <- kmeans(housing[,c("RM", "LSTAT")], centers=3, nstart=25)
```

```
clusplot(housing, housing.km$cluster, color=T, shade=T, labels=4, lines=0, main="Clusplot of Housing Data")
```



These two components explain 58.53 % of the point variability.

```
#plot(housing)
```

```
#
```

```
#ggplot(housing, aes(x=RM, y=MEDV)) + geom_point()
```

```
#ggplot(housing, aes(x=LSTAT, y=RM)) + geom_point()
```

```
#ggplot(housing, aes(x=DIS, y=NOX)) + geom_point()
```

```
#ggplot(housing, aes(x=DIS, y=MEDV)) + geom_point()
```

```
#
```

```
#ggplot(housing, aes(x=CRIM, y=MEDV)) + geom_point()
```

```
#ggplot(housing, aes(x=DIS, y=MEDV)) + geom_point()
```

```
#ggplot(housing, aes(x=NOX, y=MEDV)) + geom_point()
```

```
#ggplot(housing, aes(x=LSTAT, y=MEDV)) + geom_point()
```

```
#ggplot(housing, aes(x=AGE, y=MEDV)) + geom_point()
```

```
# =====
```

```
housing.numerical <- select(housing, -CHAS)
```

```
ind <- sample(2, nrow(housing.numerical), replace=TRUE, prob=c(0.7, 0.3))
```

```
data.training <- housing.numerical[ind==1, 1:13]
```

```

data.test <- housing.numerical[ind==2, 1:13]

data.trainLabels <- housing.numerical[ind==1, 13]

data.testLabels <- housing.numerical[ind==2, 13]

data_pred <- knn(train=data.training, test=data.test, cl=data.trainLabels, k=20)

merge <- data.frame(data.testLabels, data_pred)

names <- colnames(data.test)
finaldata <- cbind(data.test, merge)
names(finaldata) <- c("a", "b")

CrossTable(x = data.testLabels, y = data_pred, prop.chisq=FALSE)

```

```

##
##
##      Cell Contents
## |-----|
## |              N |
## |      N / Row Total |
## |      N / Col Total |
## |      N / Table Total |
## |-----|
##
##
## Total Observations in Table:  153
##
##
##      | data_pred
## data.testLabels |      7.2 |      10.5 |      12.3 |      13.4 |      13.8 |      14.1 |      14.4
## -----|-----|-----|-----|-----|-----|-----|-----|
##      5 |      1 |      0 |      0 |      0 |      0 |      0 |      0
##      |      1.000 |      0.000 |      0.000 |      0.000 |      0.000 |      0.000 |      0.000
##      |      0.250 |      0.000 |      0.000 |      0.000 |      0.000 |      0.000 |      0.000
##      |      0.007 |      0.000 |      0.000 |      0.000 |      0.000 |      0.000 |      0.000
## -----|-----|-----|-----|-----|-----|-----|-----|
##      6.3 |      0 |      0 |      0 |      0 |      0 |      0 |      0
##      |      0.000 |      0.000 |      0.000 |      0.000 |      0.000 |      0.000 |      0.000
##      |      0.000 |      0.000 |      0.000 |      0.000 |      0.000 |      0.000 |      0.000
##      |      0.000 |      0.000 |      0.000 |      0.000 |      0.000 |      0.000 |      0.000
## -----|-----|-----|-----|-----|-----|-----|-----|
##      7 |      0 |      0 |      0 |      1 |      0 |      0 |      0
##      |      0.000 |      0.000 |      0.000 |      1.000 |      0.000 |      0.000 |      0.000
##      |      0.000 |      0.000 |      0.000 |      0.100 |      0.000 |      0.000 |      0.000
##      |      0.000 |      0.000 |      0.000 |      0.007 |      0.000 |      0.000 |      0.000
## -----|-----|-----|-----|-----|-----|-----|-----|
##      7.5 |      0 |      0 |      0 |      1 |      0 |      0 |      0
##      |      0.000 |      0.000 |      0.000 |      1.000 |      0.000 |      0.000 |      0.000
##      |      0.000 |      0.000 |      0.000 |      0.100 |      0.000 |      0.000 |      0.000
##      |      0.000 |      0.000 |      0.000 |      0.007 |      0.000 |      0.000 |      0.000
## -----|-----|-----|-----|-----|-----|-----|-----|

```

| | | | | | | | | |
|----|------|-------|-------|-------|-------|-------|-------|-------|
| ## | 8.3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| ## | | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ## | | 0.000 | 0.250 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ## | | 0.000 | 0.007 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ## | 8.4 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| ## | | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 |
| ## | | 0.000 | 0.000 | 0.000 | 0.100 | 0.000 | 0.000 | 0.000 |
| ## | | 0.000 | 0.000 | 0.000 | 0.007 | 0.000 | 0.000 | 0.000 |
| ## | 8.5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| ## | | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ## | | 0.250 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ## | | 0.007 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ## | 8.8 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| ## | | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 |
| ## | | 0.000 | 0.000 | 0.000 | 0.100 | 0.000 | 0.000 | 0.000 |
| ## | | 0.000 | 0.000 | 0.000 | 0.007 | 0.000 | 0.000 | 0.000 |
| ## | 9.6 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| ## | | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 |
| ## | | 0.000 | 0.000 | 0.000 | 0.100 | 0.000 | 0.000 | 0.000 |
| ## | | 0.000 | 0.000 | 0.000 | 0.007 | 0.000 | 0.000 | 0.000 |
| ## | 10.2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| ## | | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ## | | 0.000 | 0.250 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ## | | 0.000 | 0.007 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ## | 10.4 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| ## | | 0.500 | 0.000 | 0.000 | 0.000 | 0.000 | 0.500 | 0.000 |
| ## | | 0.250 | 0.000 | 0.000 | 0.000 | 0.000 | 0.500 | 0.000 |
| ## | | 0.007 | 0.000 | 0.000 | 0.000 | 0.000 | 0.007 | 0.000 |
| ## | 10.9 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| ## | | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 |
| ## | | 0.000 | 0.000 | 0.000 | 0.100 | 0.000 | 0.000 | 0.000 |
| ## | | 0.000 | 0.000 | 0.000 | 0.007 | 0.000 | 0.000 | 0.000 |
| ## | 11.7 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| ## | | 0.000 | 0.000 | 0.000 | 0.500 | 0.000 | 0.500 | 0.000 |
| ## | | 0.000 | 0.000 | 0.000 | 0.100 | 0.000 | 0.500 | 0.000 |
| ## | | 0.000 | 0.000 | 0.000 | 0.007 | 0.000 | 0.007 | 0.000 |
| ## | 11.8 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| ## | | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 |
| ## | | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 |
| ## | | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.007 |
| ## | 11.9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ## | | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ## | | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ## | | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

[illegible]

[illegible]

| | | | | | | | | | | | | | | | |
|----|--------------|--|-------|--|-------|--|-------|--|-------|--|-------|--|-------|--|-------|
| ## | 46 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 |
| ## | | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 |
| ## | | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 |
| ## | | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 |
| ## | ----- | | ----- | | ----- | | ----- | | ----- | | ----- | | ----- | | ----- |
| ## | 48.5 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 |
| ## | | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 |
| ## | | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 |
| ## | | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 |
| ## | ----- | | ----- | | ----- | | ----- | | ----- | | ----- | | ----- | | ----- |
| ## | 50 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | | 0 |
| ## | | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 |
| ## | | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 |
| ## | | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 | | 0.000 |
| ## | ----- | | ----- | | ----- | | ----- | | ----- | | ----- | | ----- | | ----- |
| ## | Column Total | | 4 | | 4 | | 1 | | 10 | | 1 | | 2 | | 1 |
| ## | | | 0.026 | | 0.026 | | 0.007 | | 0.065 | | 0.007 | | 0.013 | | 0.007 |
| ## | ----- | | ----- | | ----- | | ----- | | ----- | | ----- | | ----- | | ----- |
| ## | | | | | | | | | | | | | | | |
| ## | | | | | | | | | | | | | | | |

```

# possibilities:
#
# BIC for determining causality
# clustering for classifying data

```