---
title: "4300 final r code v2"
author: "cheng chung yu 20516184"
date: "12/21/2020"
output:
  word_document: default
  pdf_document: default

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = FALSE)
```

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>.

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:

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

```{r data}
load("~/Desktop/4300/Final Proj/HKC2011.RData")
```

Below commend performed a scatterplot include fit lines, marginal box plots, conditioning on a factor, and interactive point identification of income and occupation. This examine the relationship between income and education level.
```{r edu }
edu <- tibble(HKC2011$mearn, HKC2011$educnh)
linearMod <- lm(edu$`HKC2011$mearn`~edu$`HKC2011$educnh`, data=edu)
summ(linearMod, confint  = "TRUE", digits= 3)
```

```{r edu_income}
scatterplot(edu$`HKC2011$educnh`, edu$`HKC2011$mearn`,xlab="education", ylab="income" , col = 1, pch = 15,regLine = list(col ="green", lwd = 3),smooth = list(col.smooth = "red",col.spread = "blue"))
```

```{r income_o}
income_o<- tibble(HKC2011$mearn, HKC2011$occup)
linearMod2 <- lm(income_o$`HKC2011$mearn`~income_o$`HKC2011$occup`, data = income_o)
summ(linearMod2, confint  = "TRUE", digits= 3)
```

```{r income_occup}
scatterplot(income_o$`HKC2011$occup`, income_o$`HKC2011$mearn`,xlab="occupation", ylab="income" , col = 1, pch = 15,regLine = list(col ="green", lwd = 3),smooth = list(col.smooth = "red",col.spread = "blue"))
```
k-mean of income, occupation, room
```{r income_room}
income_room <- tibble(HKC2011$mearn, HKC2011$occup,  HKC2011$room)
set.seed(123)
K<- kmeans(income_room,3, nstart = 50)
aggregate(income_room, by=list(cluster=K$cluster), mean)
income_room$cluster <- K$cluster
```

```{r K_mean}
fviz_cluster(K, data = income_room[, -3],
palette = c("#2E9FDF", "#00AFBB", "#E7B800"),
geom = "point",
xlab= "Dim", ylab = "Dim 50%")
```

```
## supervised learning
```{r room_room}
room_norm <- as_tibble(lapply(income_room[1:3], normalize))
set.seed(1234)
ind2 <- sample(2, nrow(room_norm), replace=TRUE, prob=c(0.67, 0.33))
training <- income_room[ind2  ==1, 1:3]
test <- income_room[ind2==2, 1:3]
trainLabels <- income_room[ind2==1,4]
testLabels<- income_room[ind2==2, 4]
cl<- trainLabels
room_pred <- knn(train = training, test = test, cl , k=133)
TestLabels <- data.frame(testLabels)
Rmerge <- data.frame(room_pred, TestLabels)
names(Rmerge) <- c("Predicted ", "Observed ")
Rmerge
```

```{r F1_score}
F1_Score(y_pred  =Rmerge$`Predicted `, y_true = Rmerge$`Observed `, positive = NULL)
```

```{r cross_table}
CrossTable(x = Rmerge$`Predicted `, y = Rmerge$`Observed `, prop.chisq=FALSE)
```

```{r error}
error <- mean(Rmerge$`Predicted ` != Rmerge$`Observed `)
error
```
```