**STAT S460F Final Project**

**S12317290**

**CHIU Man Tsun**

**What bring a high cost of living to the world?**

**Introduction**

Hong Kong is known as one of the highest rent cities in the world. Therefore, an impression of high cost of living has always bring in mind. A dataset *Cost-of-Living-Index by City 2020* is published by NUMBEO, which is the largest cost of living database of the world. It compares the cost of basic necessities between 440 cities. According to this dataset, Hong Kong is ranked 44th out of 440 cities in the world which shows that it really does have a high cost of living.

**Problem description**

The data of *Cost-of-Living-Index by City 2020* shows the percentage difference of four actual basic necessities from different city that relative to New York City. A Cost-of-Living-Index was computed by the actual collected data and a rank for 440 cities was generated.

From the dataset, we found that a large gap between different index levels may contained in the same city. For example, Hong Kong has an obviously different between rent index and restaurant price index.

Therefore, this project is aimed to find out the most related variable that correlated to a high cost of living so as to examine which cost of necessity is the most unbalanced for living.

**Aims and Objectives**

In this project, the histograms of Cost-of-Living-Index, Rent-Index, Groceries-Index, Restaurant-Price-Index and Local-Purchasing-Power-Index will be used for display their data distribution. Moreover, the correlation between Cost-of-Living-Index and the other indexes will showed by simple regression models.

In order to discover the correlation between cost of living and cost of necessity, a best regression model will be chosen between Ridge regression and Lasso regression. Two of the best selected indexes will used for test the classification accuracy of Cost-of-Living-Index by using Support Vector Machines.

**Data Description**

In the dataset, total of seven column included the five indexes of 440 observations are showed. Each of them is representing one of the cities in the world.

Cost-of-Living-Index is a relative indicator of consumer goods prices, including groceries, restaurants, transportation and utilities. Rent-Index is an estimation of prices of renting apartments in the city. Groceries-Index is an estimation of grocery prices in the city. Restaurant-Price-Index is a comparison of prices of meals and drinks in restaurants and bars. Local-Purchasing-Power-Index shows relative purchasing power in buying goods and services in a given city for the average net salary in that city.

For the further processes, the data that without consideration of ranking and city name will be separated into two equal size which contain 220 data for each.

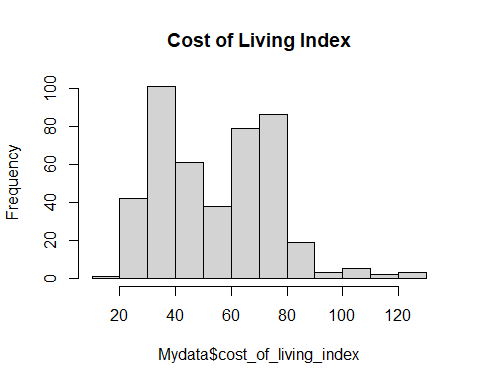
As Cost-of-Living-Index is the target observation of this project, it will be the dependent variable Y. Relatively, Rent-Index, Groceries-Index, Restaurant-Price-Index and Local-Purchasing-Power-Index will be the independent variables X.

**Exploratory and Descriptive Analysis of Data**

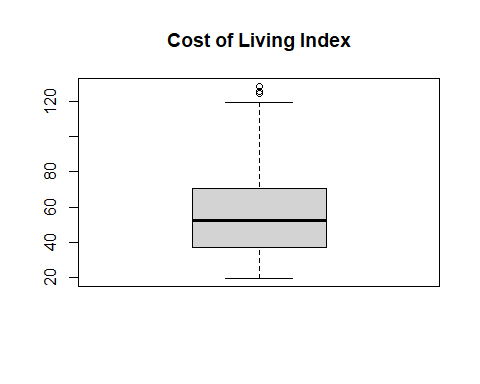
summary(Mydata$cost\_of\_living\_index)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 19.77 37.09 52.45 54.82 70.67 128.29

hist(Mydata$cost\_of\_living\_index, main="Cost of Living Index")



boxplot(Mydata$cost\_of\_living\_index, main="Cost of Living Index")

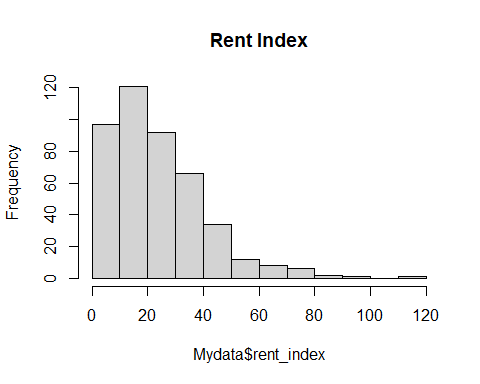


For the Cost-of-Living-Index Y, the histogram shows that the indexes are non-normally distributed. Some outliers can be saw in the upper part of the boxplot.

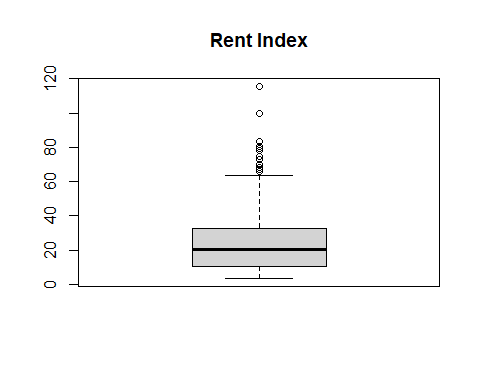
summary(Mydata$rent\_index)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.46 10.62 20.17 23.75 32.50 115.58

hist(Mydata$rent\_index, main="Rent Index")



boxplot(Mydata$rent\_index, main="Rent Index")



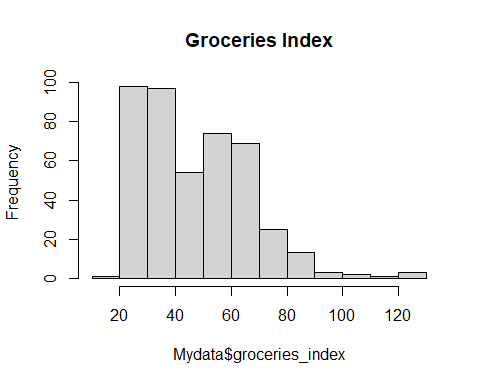
For the Rent-Index X1, the histogram shows that the indexes are positive skewed.

A Lot of outliers can be saw in the upper part of the boxplot.

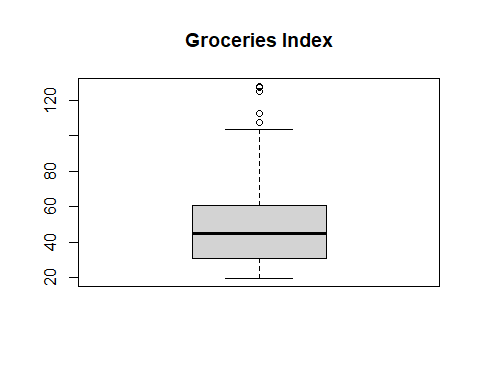
summary(Mydata$groceries\_index)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 19.66 30.82 44.77 47.50 61.03 127.96

hist(Mydata$groceries\_index, main="Groceries Index")



boxplot(Mydata$groceries\_index, main="Groceries Index")

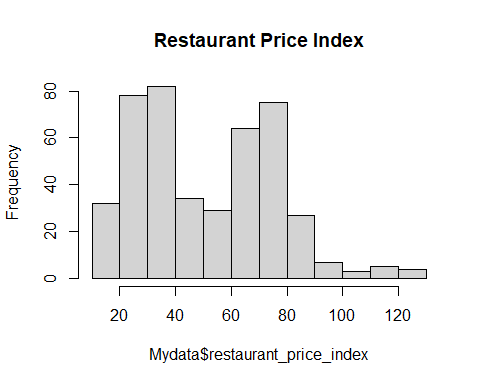


For the Groceries-Index X2, the histogram shows that the indexes are also non-normally distributed. Some outliers can be saw in the upper part of the boxplot.

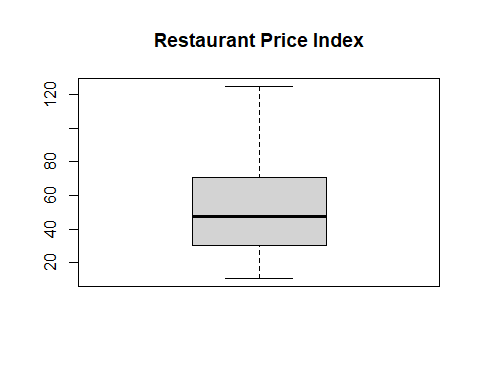
summary(Mydata$restaurant\_price\_index)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 10.66 30.21 47.55 51.21 70.83 124.73

hist(Mydata$restaurant\_price\_index, main="Restaurant Price Index")



boxplot(Mydata$restaurant\_price\_index, main="Restaurant Price Index")

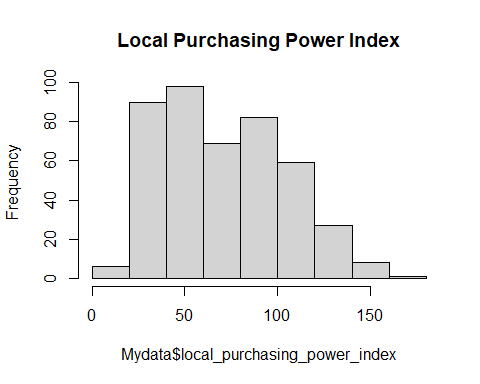


For the Restaurant-Price-Index X3, the histogram shows that the indexes are non-normally distributed with no outlier.

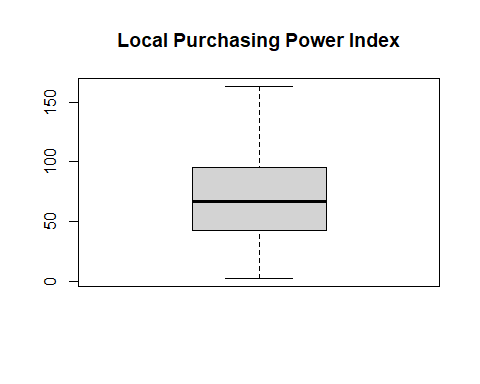
summary(Mydata$local\_purchasing\_power\_index)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.36 42.27 66.64 70.80 95.52 163.27

hist(Mydata$local\_purchasing\_power\_index, main="Local Purchasing Power Index")



boxplot(Mydata$local\_purchasing\_power\_index, main="Local Purchasing Power Index")

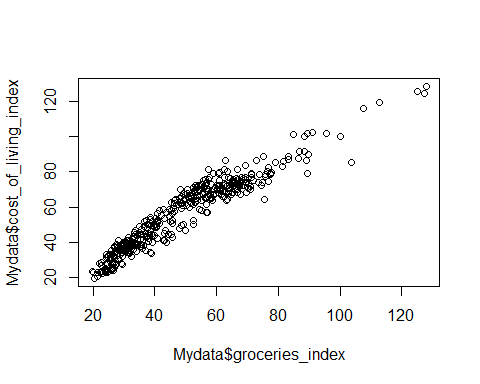
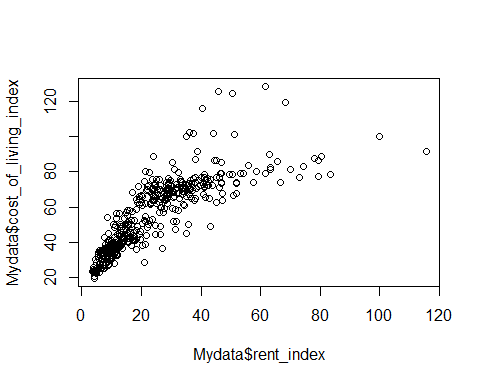


For the Local-Purchasing-Power-Index X4, the histogram shows that the indexes are also non-normally distributed with no outlier.

**Correlation Plots**

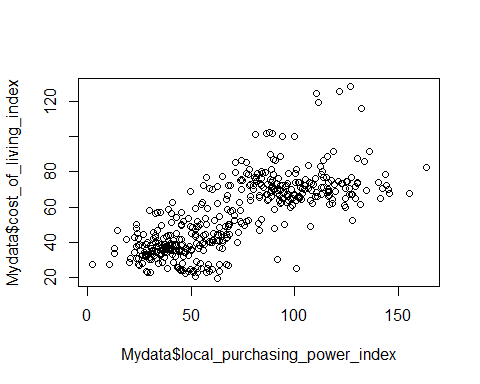
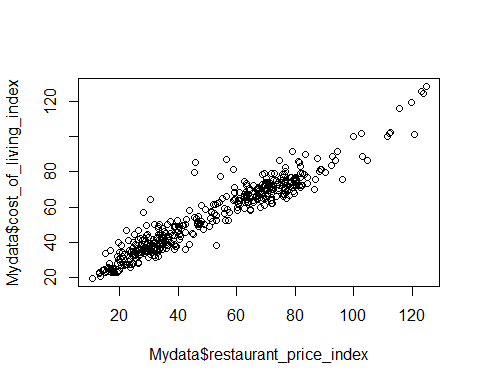
plot(Mydata$rent\_index, Mydata$cost\_of\_living\_index)

plot(Mydata$groceries\_index, Mydata$cost\_of\_living\_index)



plot(Mydata$restaurant\_price\_index, Mydata$cost\_of\_living\_index)

plot(Mydata$local\_purchasing\_power\_index, Mydata$cost\_of\_living\_index)



cor(Mydata, Mydata$cost\_of\_living\_index)

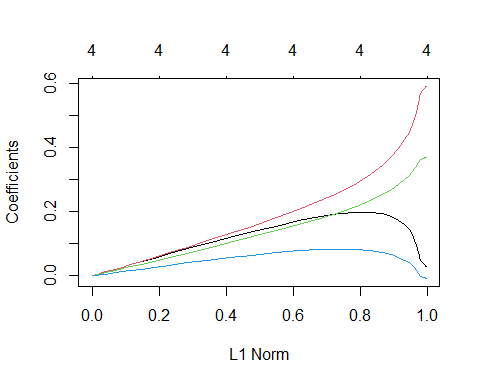
## [,1]  
## cost\_of\_living\_index 1.0000000  
## rent\_index 0.8197904  
## groceries\_index 0.9580292  
## restaurant\_price\_index 0.9560048  
## local\_purchasing\_power\_index 0.7470956

From the correlation plots, all of the independent variables show a positive correlation between Y. Among them, the plots of Groceries-Index and Restaurant-Price-Index show a clear positive linear correlation.

Moreover, the value of correlation R show that linear correlation of Groceries-Index and Restaurant-Price-Index are obviously stronger than the other variables.

**Ridge Regression**

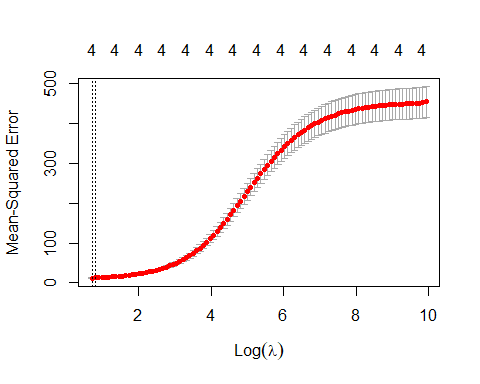
grid = 10^seq(10, -2, length = 100)  
ridge\_mod = glmnet(x\_train, y\_train, alpha=0, lambda = grid, thresh = 1e-12)  
plot(ridge\_mod)



ridge\_pred = predict(ridge\_mod, s = 4, newx = x\_test)  
ridge\_pred = predict(ridge\_mod, s = 1e10, newx = x\_test)  
  
set.seed(1)  
cv.out = cv.glmnet(x\_train, y\_train, alpha = 0)  
bestlam = cv.out$lambda.min  
bestlam

## [1] 2.054318

plot(cv.out)



ridge\_pred = predict(ridge\_mod, s = bestlam, newx = x\_test)  
mean((ridge\_pred - y\_test)^2) #Testing Mean Square Error

## [1] 10.20564

out = glmnet(x, y, alpha = 0)  
predict(out, type = "coefficients", s = bestlam)[1:5,] #Coefficients

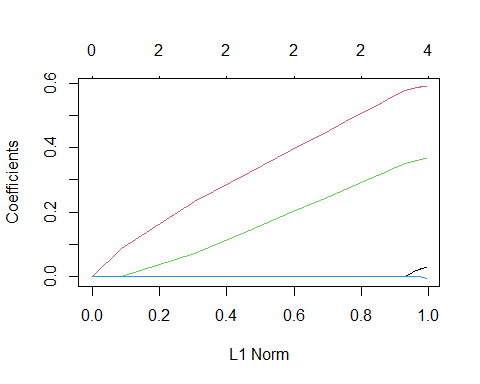
## (Intercept) rent\_index   
## 10.45411301 0.11222359   
## groceries\_index restaurant\_price\_index   
## 0.45631725 0.35561506   
## local\_purchasing\_power\_index   
## 0.02563671

By using Ridge regression, we find a best lambda around 2.05. We further got the value of MSE which is around 10.21, and there is no dimension reduced shown.

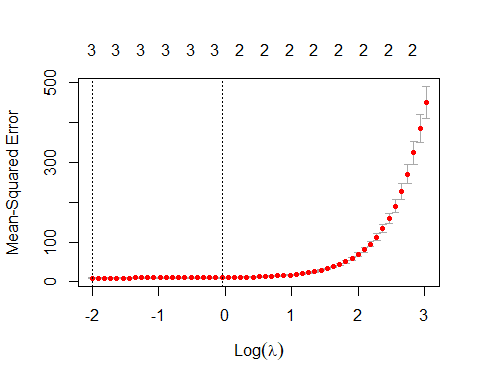
**Lasso Regression**

lasso\_mod = glmnet(x\_train, y\_train, alpha = 1, lambda = grid)  
plot(lasso\_mod)

## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):  
## collapsing to unique 'x' values



set.seed(1)  
cv.out = cv.glmnet(x\_train, y\_train, alpha = 1)  
plot(cv.out)



bestlam = cv.out$lambda.min  
bestlam

## [1] 0.1351604

lasso\_pred = predict(lasso\_mod, s = bestlam, newx = x\_test)  
mean((lasso\_pred - y\_test)^2) #Testing Mean Square Error

## [1] 8.659229

out = glmnet(x, y, alpha = 1, lambda = grid)  
predict(out, type = "coefficients", s = bestlam)[1:5,] #Coefficients

## (Intercept) rent\_index   
## 8.02422234 0.01370848   
## groceries\_index restaurant\_price\_index   
## 0.53571452 0.41054321   
## local\_purchasing\_power\_index   
## 0.00000000

By using Lasso regression, we find a best lambda around 0.14. We further got the value of MSE which is around 8.66, and the variable Local-Purchasing-Power-Index is reduced.

**Model Comparison**

After comparing between Ridge regression and Lasso regression, as the MSE of Lasso regression is lower than Ridge regression. One extra dimension was reduced during Lasso regression. Therefore, we find that Lasso regression will obtain a better performance for this dataset.

**Support Vector Machines**

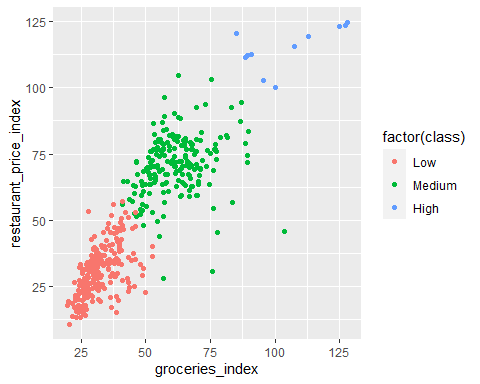
In order to classify the Cost-of-Living-Index by using Support Vector Machines,

This project picked the most correlated independent variables of the above selection methods, X2 Groceries-Index and X3 Restaurant-Price-Index for SVM.

Since there is no standard for separate the level of Cost-of-Living-Index, we will split the index to three relatively levels.

From the dataset, we have got the range of Cost-of-Living-Index which is 19.77 to 128.29. Three equal size of the range will be cut by 56 and 92 to separate the data to Low, Medium and High Cost-of-Living-Index.

groceries\_index = data$groceries\_index  
restaurant\_price\_index = data$restaurant\_price\_index  
class = cut(data$cost\_of\_living\_index, c(0, 56, 92, Inf), labels = c("Low","Medium","High"))  
  
Newdata = data.frame(+groceries\_index, +restaurant\_price\_index)  
  
ggplot(data.frame(data), aes(groceries\_index, restaurant\_price\_index, colour = factor(class)))+  
 geom\_point()



From the above graph, the three levels of data show a clear separation between each other. No obvious overlapping can be seen. Therefore, a linear kernel will be chosen for tunning the SVM model.

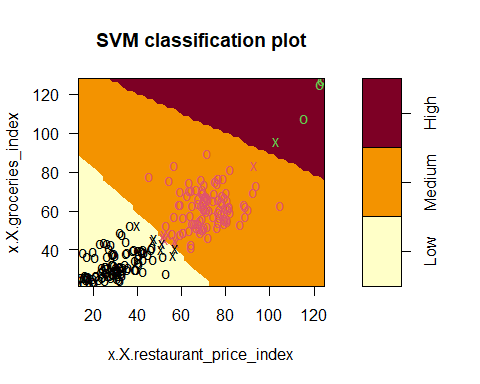
data = data.frame(x=Newdata, class=as.factor(class))  
  
train\_newdata = data %>%  
 sample\_frac(0.5)  
  
test\_newdata = data %>%  
 setdiff(train\_newdata)

set.seed(1)  
tune\_out = tune(svm, class~., data=train\_newdata, kernel="linear", ranges=list(cost=c(0.1, 1, 5, 10, 100)))  
summary(tune\_out)

##   
## Parameter tuning of 'svm':  
##   
## - sampling method: 10-fold cross validation   
##   
## - best parameters:  
## cost  
## 5  
##   
## - best performance: 0.009090909   
##   
## - Detailed performance results:  
## cost error dispersion  
## 1 0.1 0.013636364 0.03067948  
## 2 1.0 0.013636364 0.02195663  
## 3 5.0 0.009090909 0.01916532  
## 4 10.0 0.009090909 0.01916532  
## 5 100.0 0.013636364 0.02195663

bestmod = tune\_out$best.model

plot(bestmod, train\_newdata)



summary(bestmod)

##   
## Call:  
## best.tune(method = svm, train.x = class ~ ., data = train\_newdata,   
## ranges = list(cost = c(0.1, 1, 5, 10, 100)), kernel = "linear")  
##   
##   
## Parameters:  
## SVM-Type: C-classification   
## SVM-Kernel: linear   
## cost: 5   
##   
## Number of Support Vectors: 15  
##   
## ( 7 7 1 )  
##   
##   
## Number of Classes: 3   
##   
## Levels:   
## Low Medium High

By using Ten-fold Cross Validation, we got the best parameters of cost equal to 5.

class\_pred = predict(bestmod, test\_newdata)  
table = table(predicted=class\_pred, true=test\_newdata$class)  
accuracy = (((table[1,1]+ table[2,2]+ table[3,3]))/ count(test\_newdata))  
  
table

## true  
## predicted Low Medium High  
## Low 117 1 0  
## Medium 1 95 0  
## High 0 0 6

accuracy #Accuracy

## n  
## 1 0.9909091

From the above prediction, it shows that SVM have a high accuracy which up to 99.09% on predicting the Level of Cost-of-Living -Index by using the two variables Groceries-Index and Restaurant-Price-Index.

**Conclusions**

From the above result, Lasso regression showed a more suitable result than Ridge regression for this dataset. Also, Support Vector Machines can provide a 99.09% high accuracy classification of Cost-of-Living-Index by using only Restaurant-Price-Index and Groceries-Index.

Moreover, we can conclude that Groceries-Index and Restaurant-Price-Index are the most unbalanced variables that will directly manipulate the level of cost of living for the city.

**Limitations and Recommendations**

However, there is some limitation in this project. As the dataset does not cover all cities in the world and only four variables may not represent all costs of basic necessities, the indexes may just show an approximate result.

**Recommendations**

In order to control the level of cost of living, the government of each city should focus on the policy of manipulating the price of groceries and restaurant.

**Reference**

Cost of Living Index by City 2020 (2020). Numbeo. Retrieved from

https://www.numbeo.com/cost-of-living/rankings.jsp?title=2020