2021應用多變數統計方法期末報告

動科五 李童 B06606026

# 資料簡介

Kaggle房價預測練習賽，此競賽的目的為使用房子的各種變數來預測房價，此資料包含了81個變數與1460筆房屋資料，雖說預測是用機器學習的模型，但是觀察資料是首要步驟，當初在自學時難以理解網站上分析的原理與解釋意義，因此想藉由此課堂練習分析這筆資料，觀察各種導致房價變化的變因與他們的相關程度，進而挑選適合訓練模型的變數。

資料來源: House Prices Advanced Regression Techniques on Kaggle competition platform.

檔案路徑：/kaggle/input/house-prices-advanced-regression-techniques/train.csv

# 資料整理

因為這筆資料包含遺漏變數以及文字類別變數，因此先以Python套件將文字類別變數轉為數字，並將有包含遺漏變數的變數去除，以便後續分析。

1. #計算各column遺漏變數的量
2. count\_missing <- data.frame(Name='name', count=0)
3. n <- 0
4. newdata <- data
5. drop <- c()
6. chr <- c()
7. for(i in names(data)){
8. n = n+1
9. missing <- sum(is.na(data[,n]))
10. if(missing>0){
11. count\_missing<-rbind(count\_missing,c(i, missing))
12. drop <- c(drop,i)
13. }
14. }
15. count\_missing <- count\_missing[-c(1),]
16. #原始train去除有 missing value 的 columns
17. newdata = newdata[,!(names(newdata) %in% drop)]
18. #numerate 後的資料去除有 missing value 的 columns
19. outdata <- read.csv(file="C:/5th/multivariate/final\_report/outdata.csv", header=T)
20. #以上是已經處理過類別變數的資料(因為不會用R處理，所以先用python處理過)
21. #去除有 missing value 的 columns
22. donedata <- outdata[,!(names(outdata) %in% drop)]
23. #去除第一行
24. donedata <- donedata[,-1]

而去除有遺漏變數的行後，仍有62個變數，為了留下較有相關性的變數，我留下與價錢較為相關的前11個變數做此報告的資料。

1. #之前做分析時這些 column 對 price 的相關性較高
2. newcol <- c('OverallQual','Neighborhood','GrLivArea','GarageCars','TotalBsmtSF','GarageArea','YearBuilt','KitchenQual','ExterQual','BsmtQual','X1stFlrSF','GarageYrBlt','MSSubClass','SalePrice')#,'SalePrice'
3. newdata2 <- subset(outdata, select = newcol)
4. donedata2 <- newdata2[,!(names(newdata2) %in% drop)]
5. names(donedata2)

OverallQual: 整體評價 (1-10)

Neighborhood: Physical locations within Ames city limits (categorical)

GrLivArea: 地面生活區面積(平方英尺)

GarageCars: 車庫中的汽車容量大小（平方英尺）

TotalBsmtSF: 地下室面積(平方英尺)

GarageArea: 車庫面積（平方英尺）

YearBuilt: 原建設日期

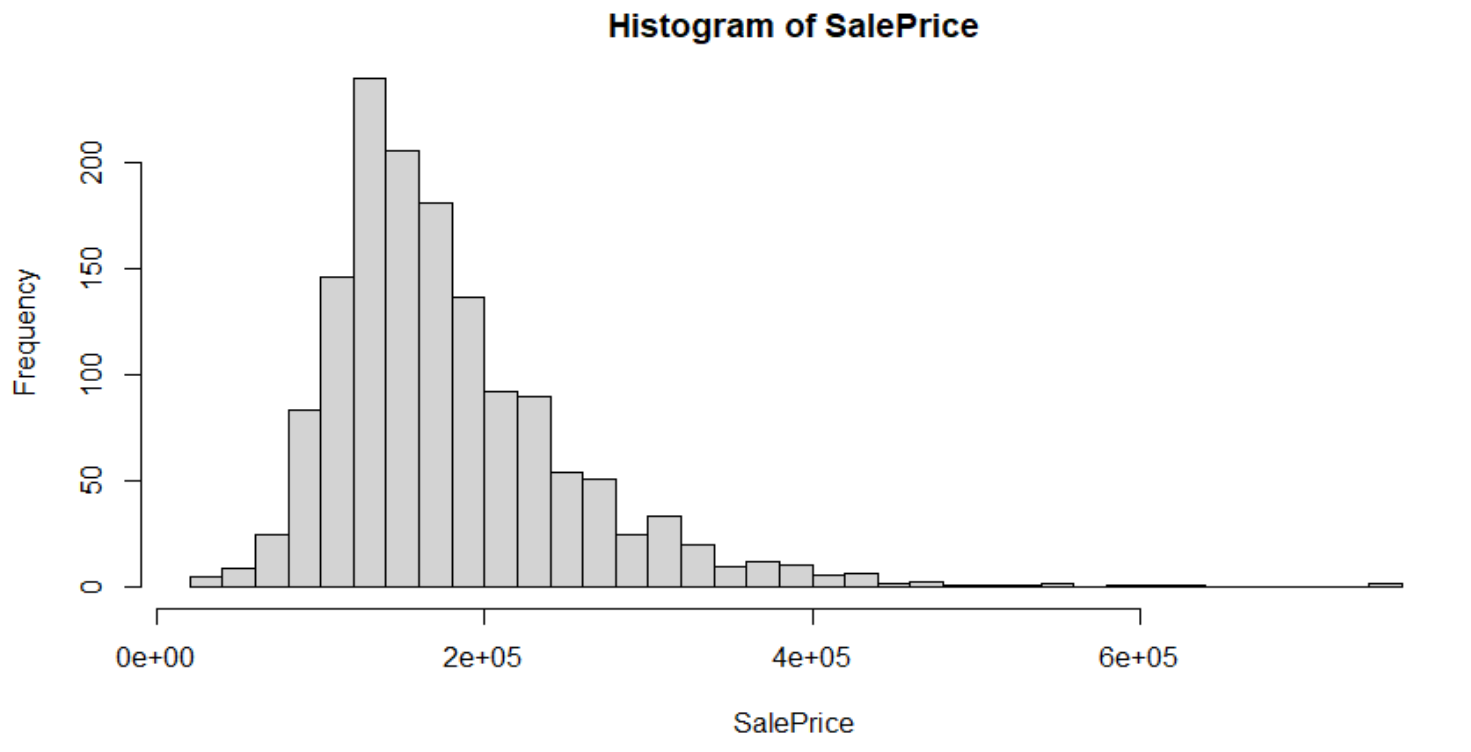
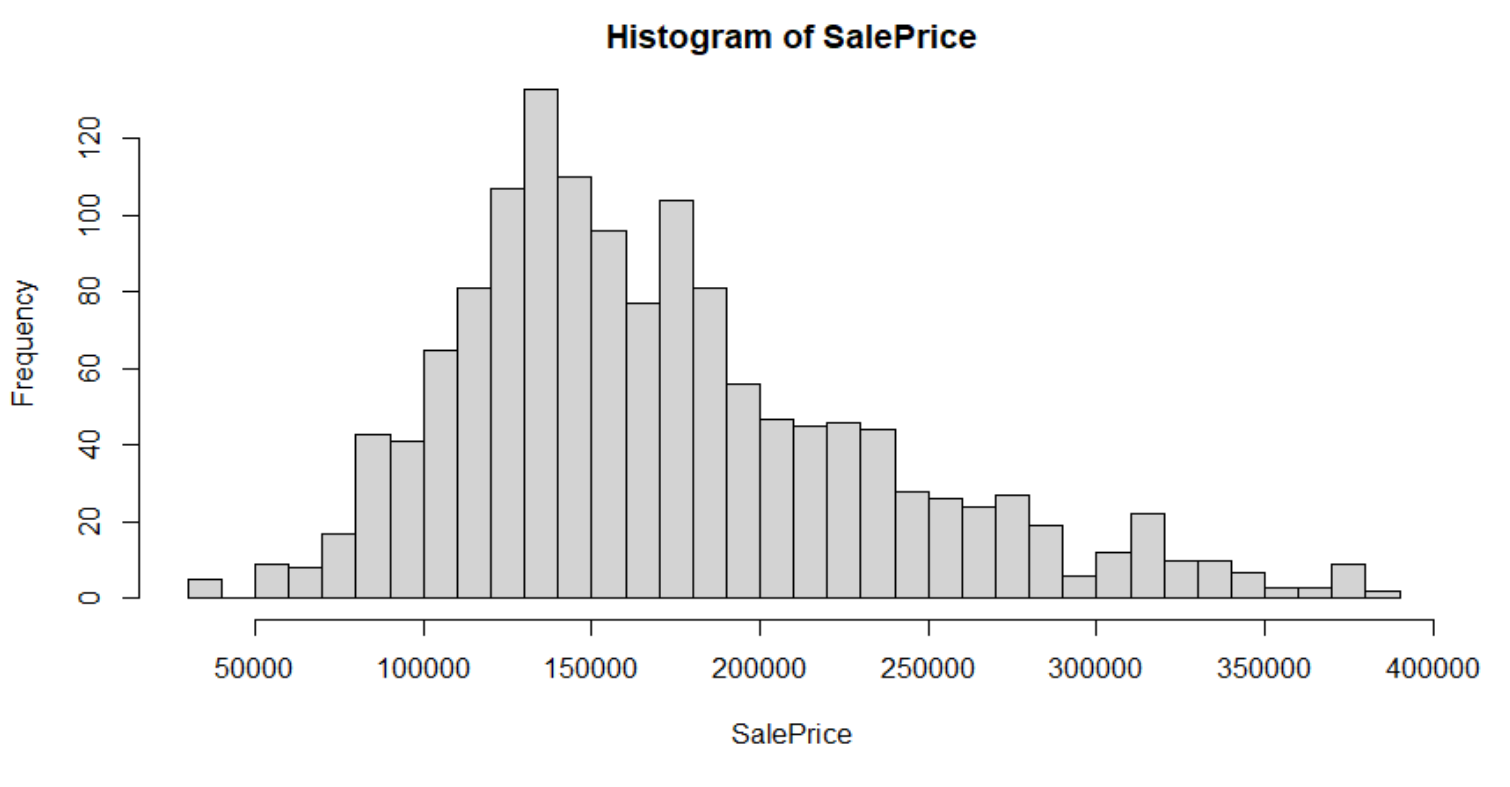
KitchenQual: 廚房品質 (Excellent, Good, Typical/Average, Fair, Poor)

ExterQual: 評估外部材料的品質 (Excellent, Good, Typical/Average, Fair, Poor)

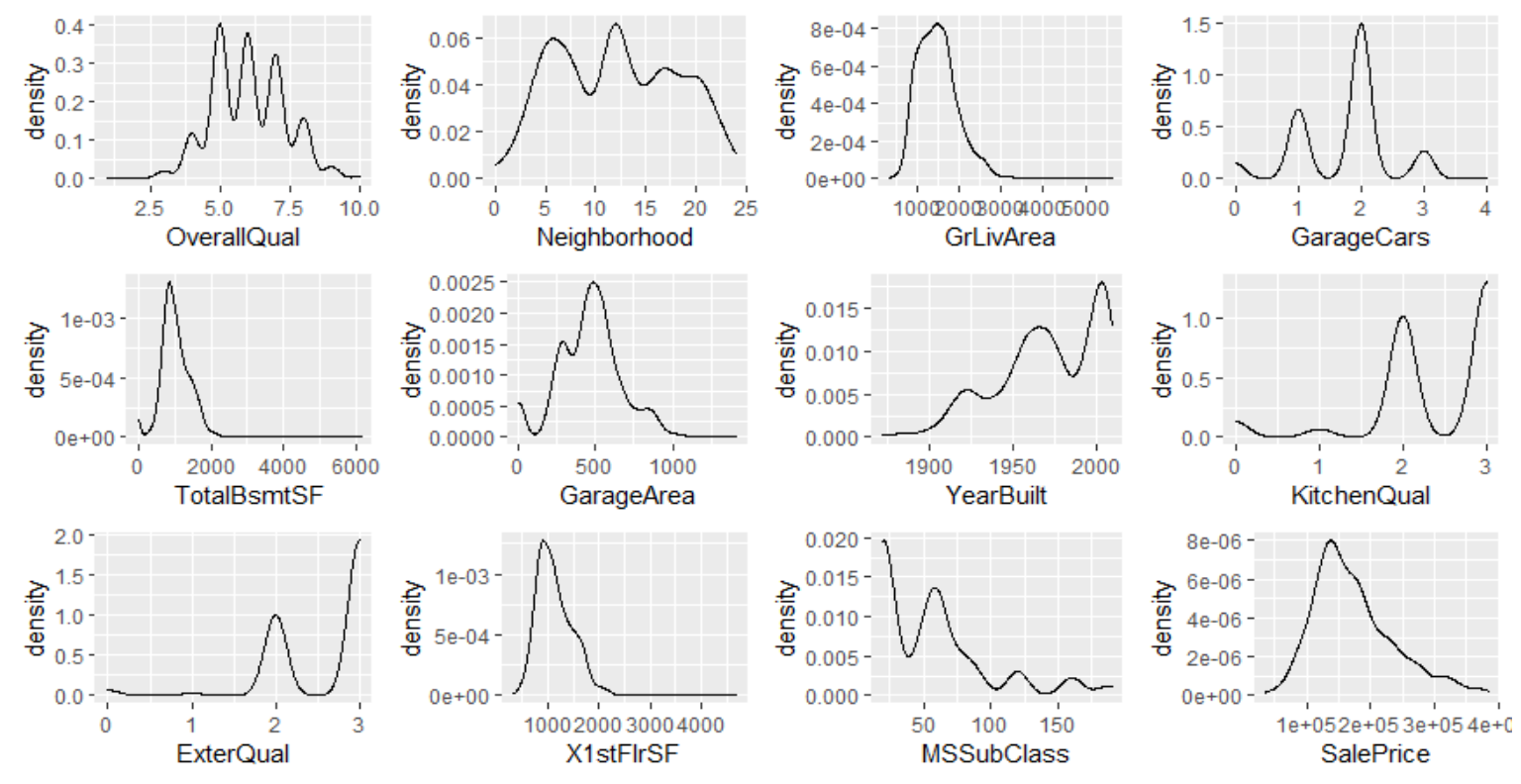
X1stFlrSF: 一樓面積

MSSubClass: 標識銷售中涉及的住宅類型。(categorical)

# Data Observation

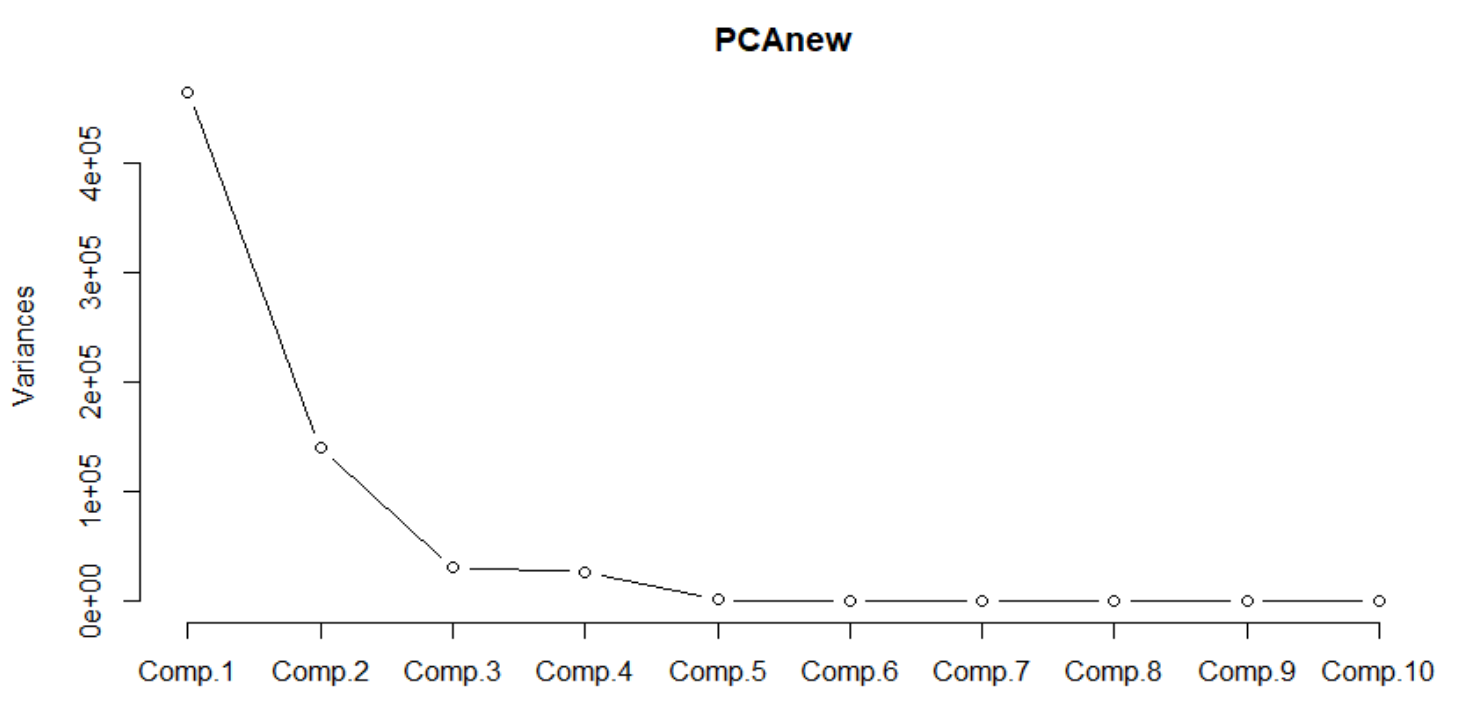
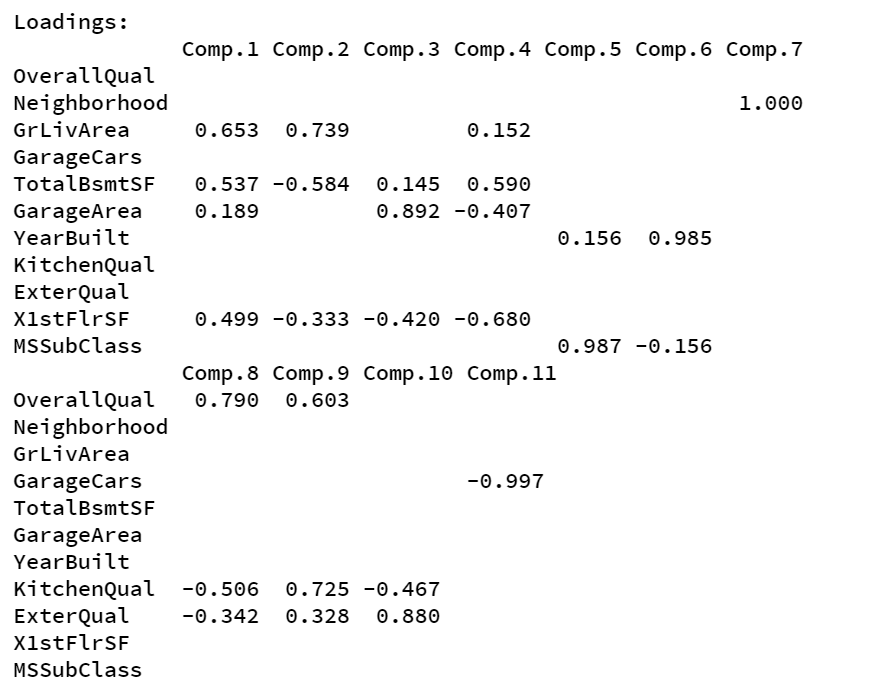
有outliers 去除outliers



整體分布如上，整體評價、鄰居型態、車庫品質、車庫面積、以及價格的分佈有接近常態分布，其他的變量都有偏鋒，可能也有outliers也可能是資料的收集太雜亂。

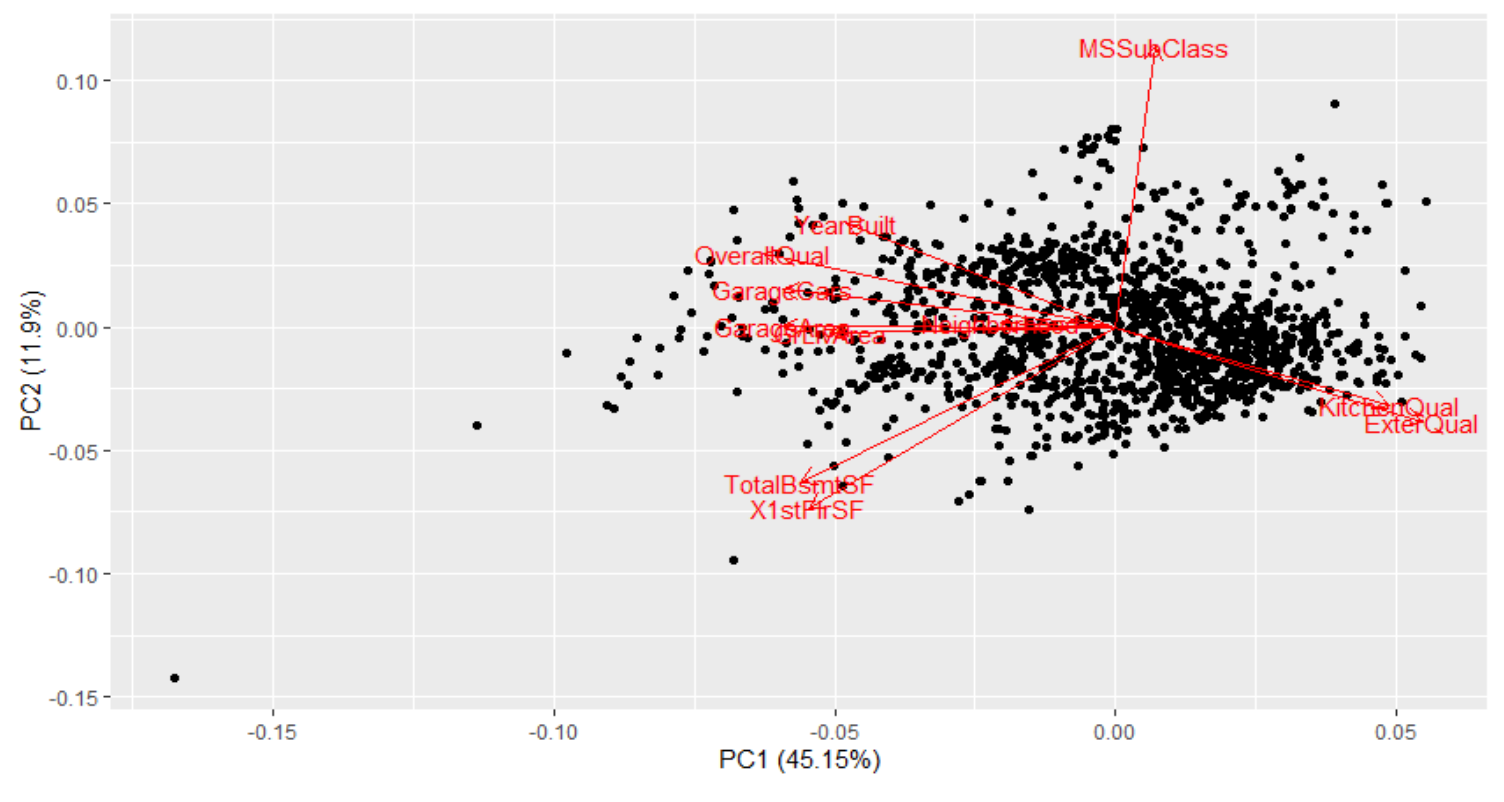
1. library(ggplot2)
2. library(cowplot)
3. my\_plots <- lapply(names(donedata2), function(var\_x){
4. p <-
5. ggplot(donedata2) +
6. aes\_string(var\_x)
7. if(is.numeric(donedata2[[var\_x]])) {
8. p <- p + geom\_density()
9. } else {
10. p <- p + geom\_bar()
11. }
12. })
13. plot\_grid(plotlist = my\_plots)
14. hist(donedata2$SalePrice,
15. xlab = "SalePrice",
16. main = "Histogram of SalePrice",
17. breaks = sqrt(nrow(donedata2))
18. ) # set number of bins
19. upper\_bound <- quantile(donedata2$SalePrice, 0.975)
20. outlier\_ind <- which(donedata2$SalePrice > upper\_bound)
21. donedata2[outlier\_ind, ]
22. donedata2 <- donedata2[-outlier\_ind,]

# PCA



圖一、主成分的分布 圖二、個主成分可解釋的變異比例

第一個主成分包含地面生活區面積、地下室面積、車庫面積、一樓面積，而第二個主成分包含地面生活區面積、地下室面積、一樓面積。主要都是面積大小的變數。



圖三、biplot

圖三顯示的向量當中，可以分為四個方向：

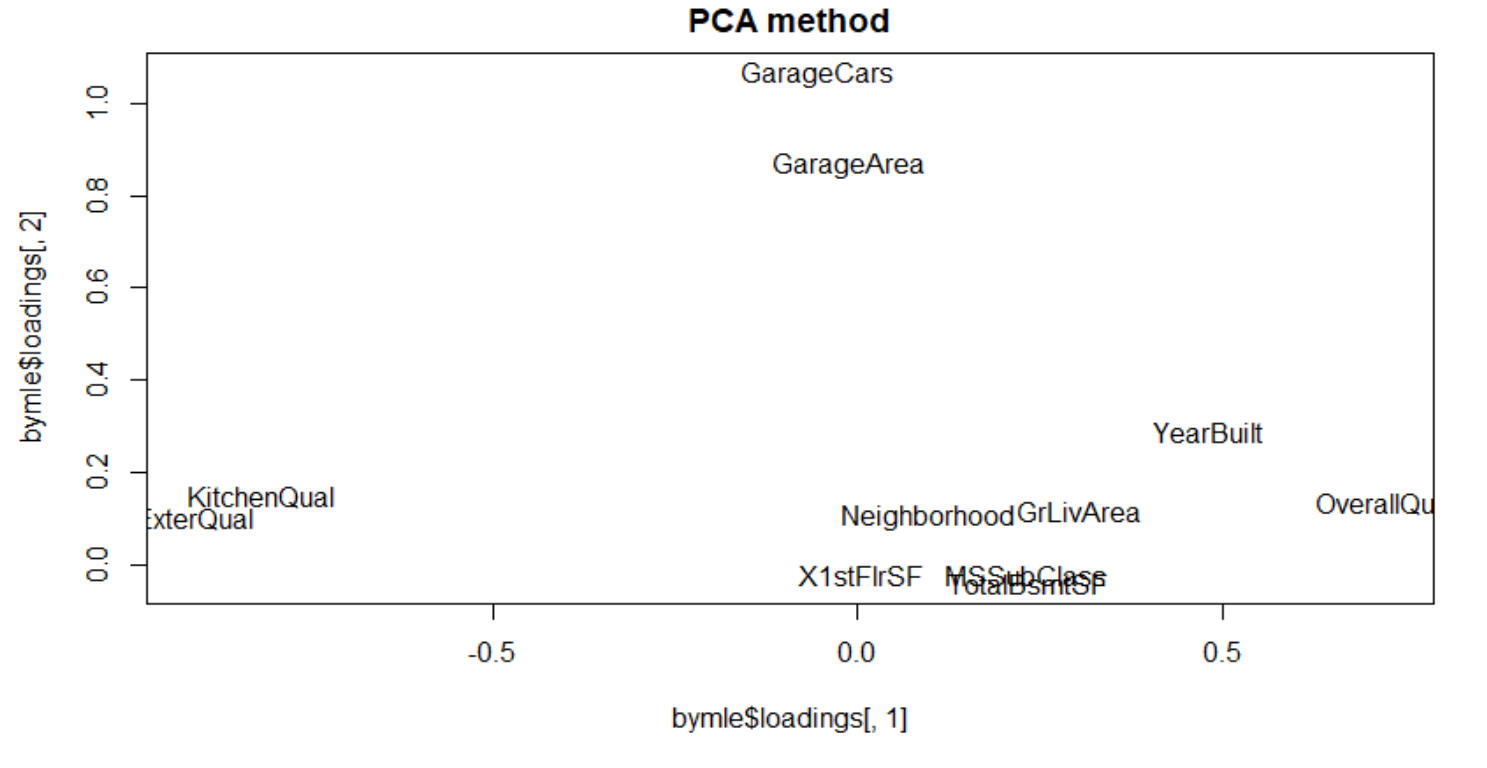
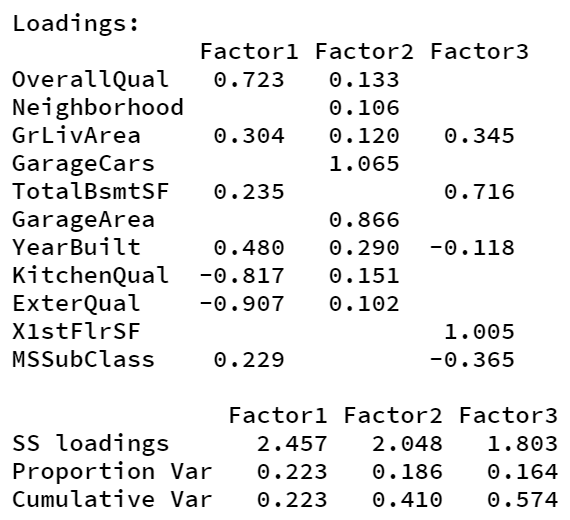
1. 標識銷售中涉及的住宅類型
2. 廚房品質 、外部材料的品質
3. 地下室面積與一樓面積
4. 原建設日期、整體評價、地面生活區面積、車庫中的汽車容量大小、車庫面積以及鄰居型態。

總結而言，主成分分析可以將11個變數分成四個類別，並且能得知不同類別可能得像似之處，比如：第二類都是品質相關的類別變數，也都分為五類 (Excellent, Good, Typical/Average, Fair, Poor)，因此可能會有類似的variation分布；而地下室面積與一樓面積都與屋內生活空間有關；與汽車相關的變量也在同一個方向中。因此此主成分分析可以比較個變數之間的相似之處。

1. #直接用donedata2執行
2. PCAnew <- princomp(donedata2[,1:11], scale = TRUE)
3. PCAnew$loadings
4. screeplot(PCAnew, type = 'l')
5. #繪圖(比較好看)
6. devtools::install\_github("sinhrks/ggfortify")
7. library(ggfortify)
8. ggplot2::autoplot(stats::prcomp(donedata2[,1:11], scale=TRUE), label = FALSE, loadings.label = TRUE)

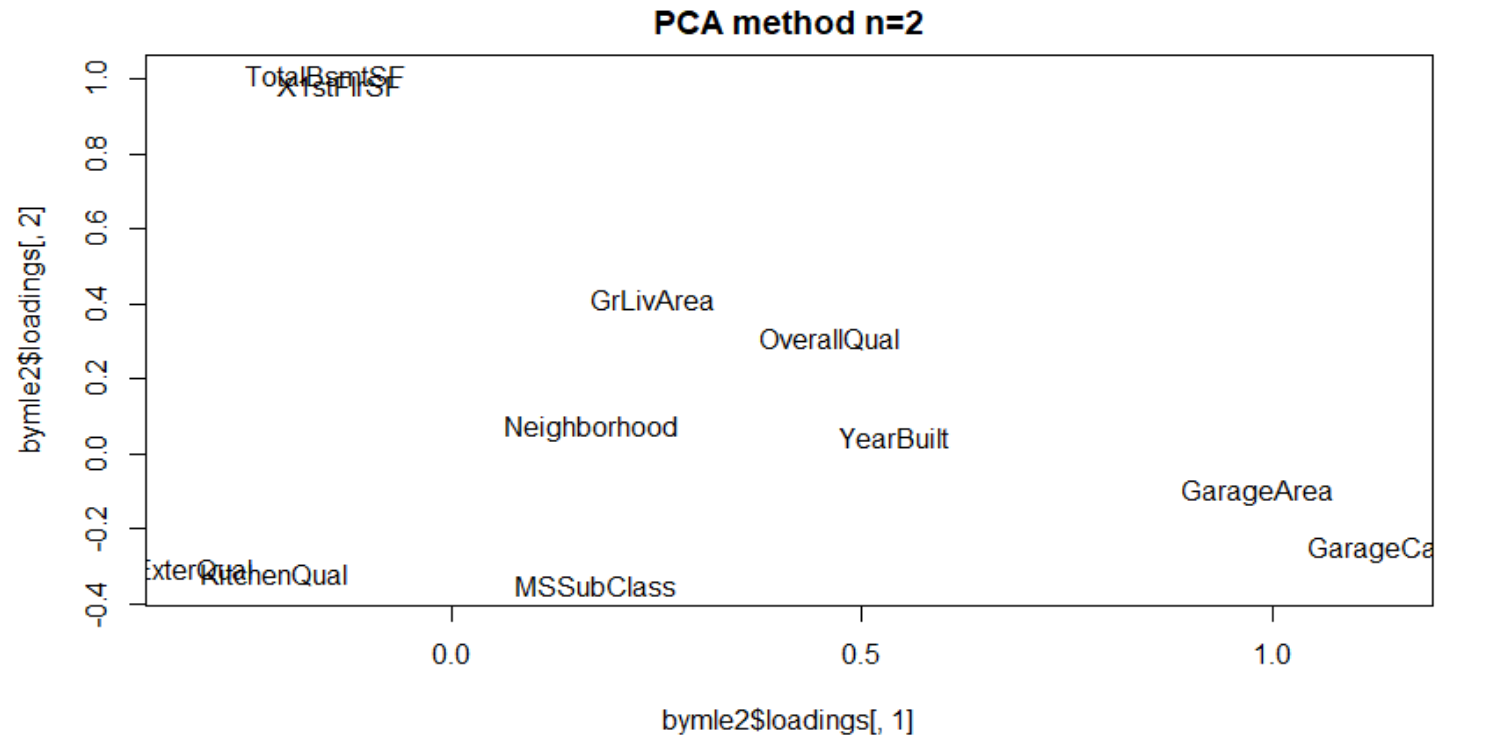
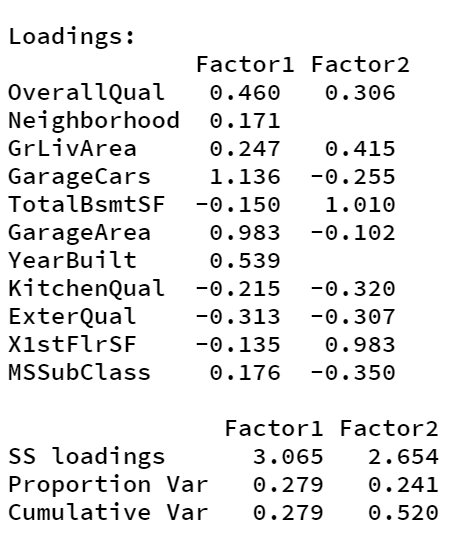
# Factor Analysis

## 三因子分析：



圖四 圖五

## 二因子分析：



圖六 圖七

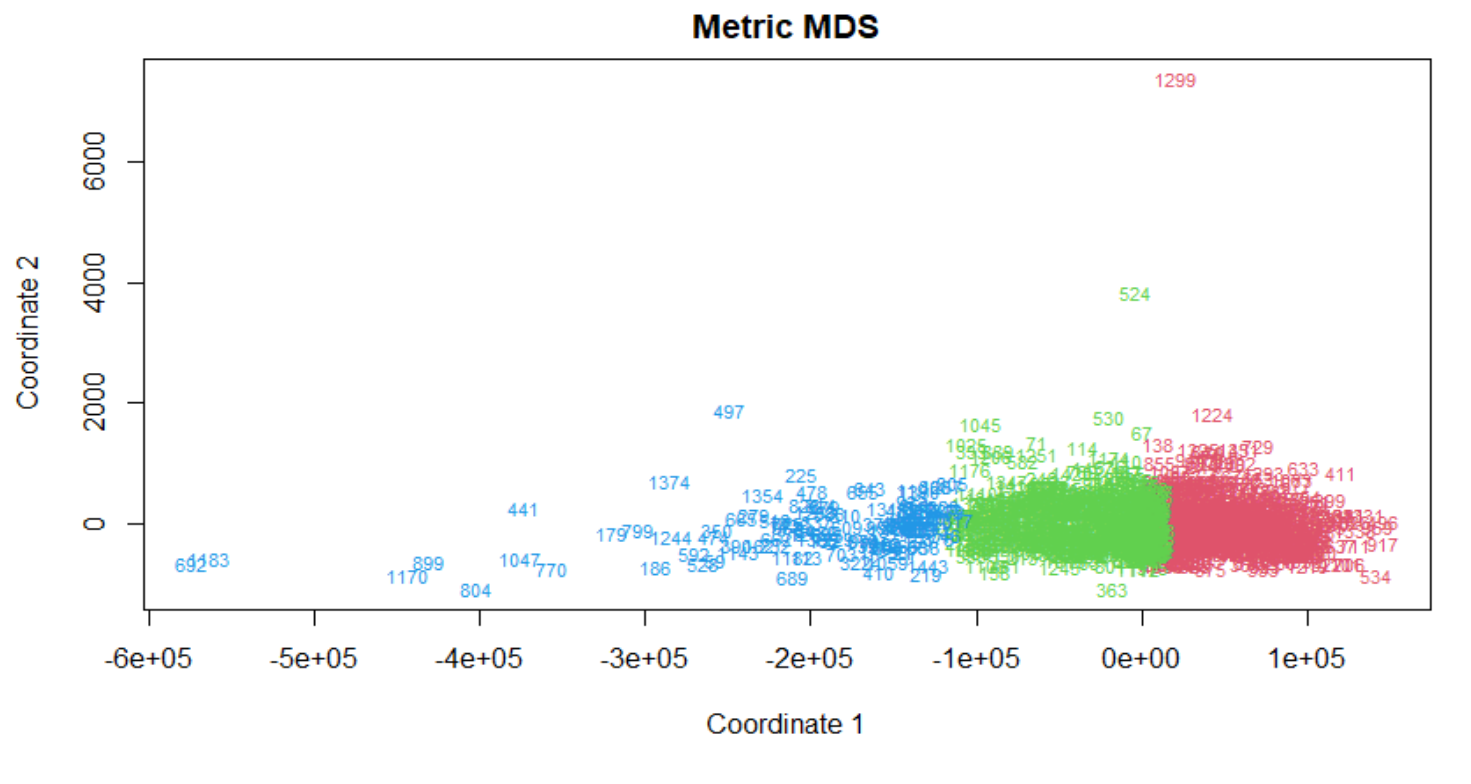
由因素分析可以將11個變量分為三個主要的類別，分別為：

1. 廚房品質 、外部材料的品質
2. 車庫中的汽車容量大小、車庫面積
3. 地下室面積與一樓面積

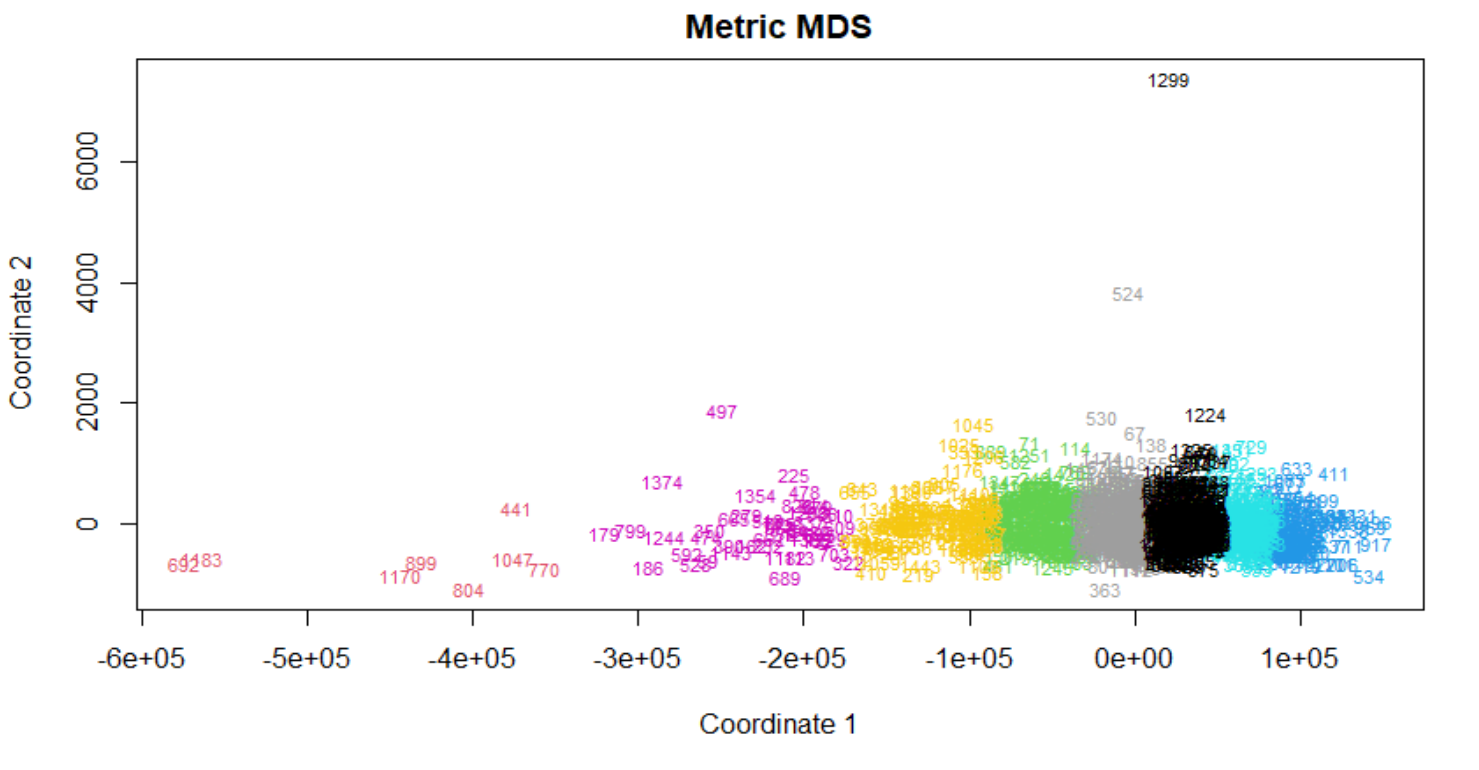
而這三類的分組也都有與主成分分析時相符，因此可以確定此三組裡所含的變量為相似變化的變量。

1. #n=3
2. library(psych)
3. cor <- cor(donedata2[,1:11])
4. #bypca <- fa(donedata2[,1:11], 4, fm="pa", rotate="varimax")#跑不出來
5. bymle <- factanal(donedata2[,1:11], factors = 3, rotation="promax", scores = "regression")
6. print(bymle$loadings)
7. plot(bymle$loadings[,1], bymle$loadings[,2], type="n", main = 'PCA method')
8. text(bymle$loadings[,1], bymle$loadings[,2], labels=rownames(bymle$loadings))
9. #n=2
10. bymle2 <- factanal(donedata2[,1:11], factors = 2, rotation="promax", scores = "regression")
11. print(bymle2$loadings)
12. plot(bymle2$loadings[,1], bymle2$loadings[,2], type="n", main = 'PCA method n=2')
13. text(bymle2$loadings[,1], bymle2$loadings[,2], labels=rownames(bymle2$loadings))

# Kmeans Clasification



圖八、k=3，三類分界清楚，但是藍色的範圍太大

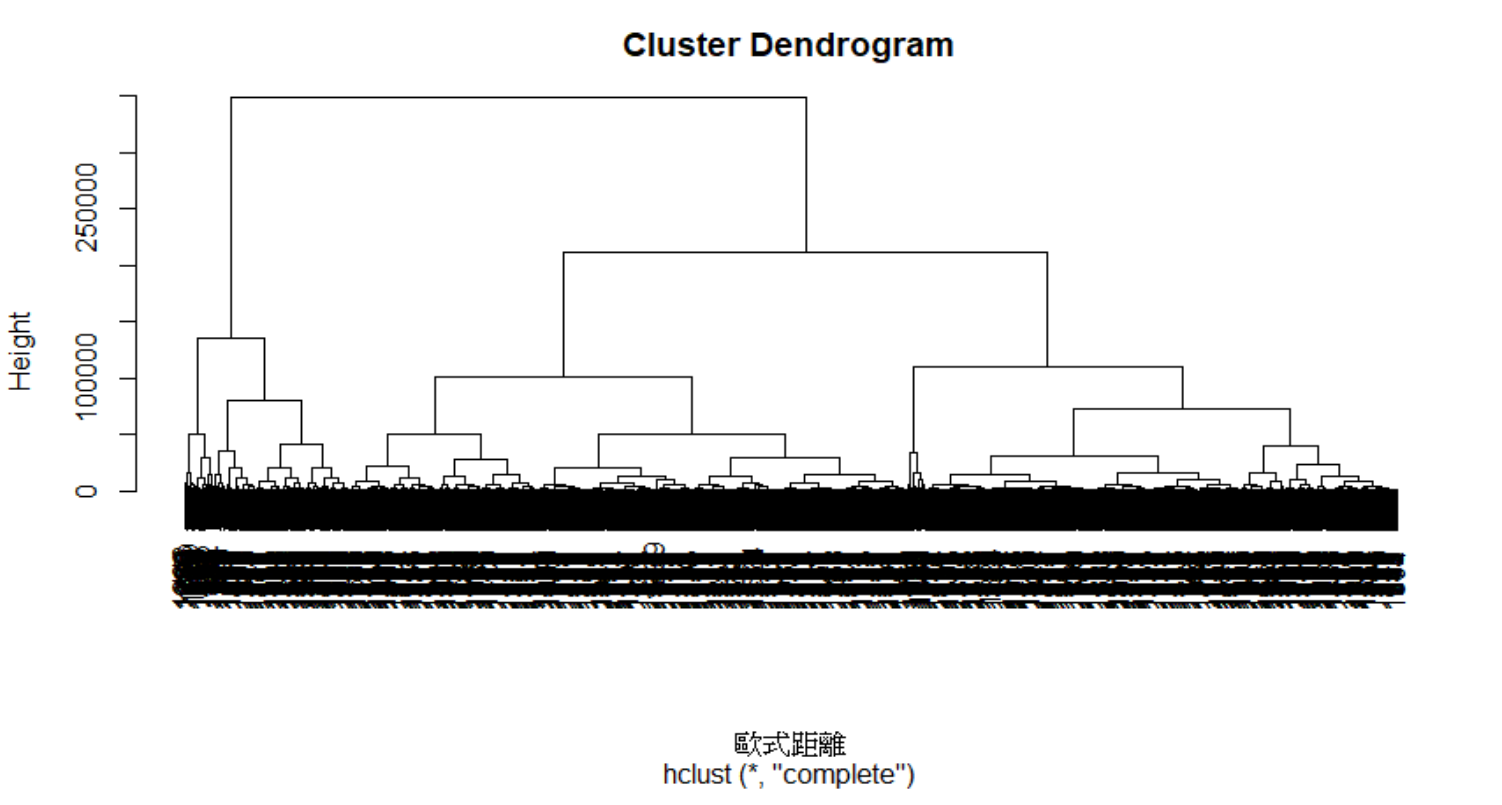


圖九、k=8，界線還算清楚，不過黑、藍、深藍那裏重疊有點多。

用Kmean可以將1640間房子做分類，不過因為目的並不是要分類房子，因此只做簡單的分類看看。不同的顏色反應不同類型的房子，而同顏色的房屋性質應較相似，因此可用於房屋性質的比較。

1. kms <- kmeans(donedata2,centers=6)
2. #dimnames(donedata2)[[1]][kms$cluster==1]
3. #row.names(donedata2)[kms$cluster==1]
4. dat\_dist <- dist(donedata2,method='euclidean')
5. mds <- cmdscale(dat\_dist,k=2)
6. plot(mds[,1],mds[,2],xlab = 'Coordinate 1',ylab = 'Coordinate 2',main='Metric MDS',type='n')
7. text(mds[,1],mds[,2],labels=row.names(donedata2),cex=.7)
8. #用圖顯示N個區域
9. plot(mds[,1],mds[,2],xlab = 'Coordinate 1',ylab = 'Coordinate 2',main='Metric MDS',type='n')
10. for(i in 1:N){
11. kms\_cl <- kms$cluster==i
12. text(mds[,1][kms\_cl],mds[,2][kms\_cl],
13. labels=row.names(donedata2)[kms\_cl],cex=.5,col=i+1)
14. }

# Hierarchical Clustering

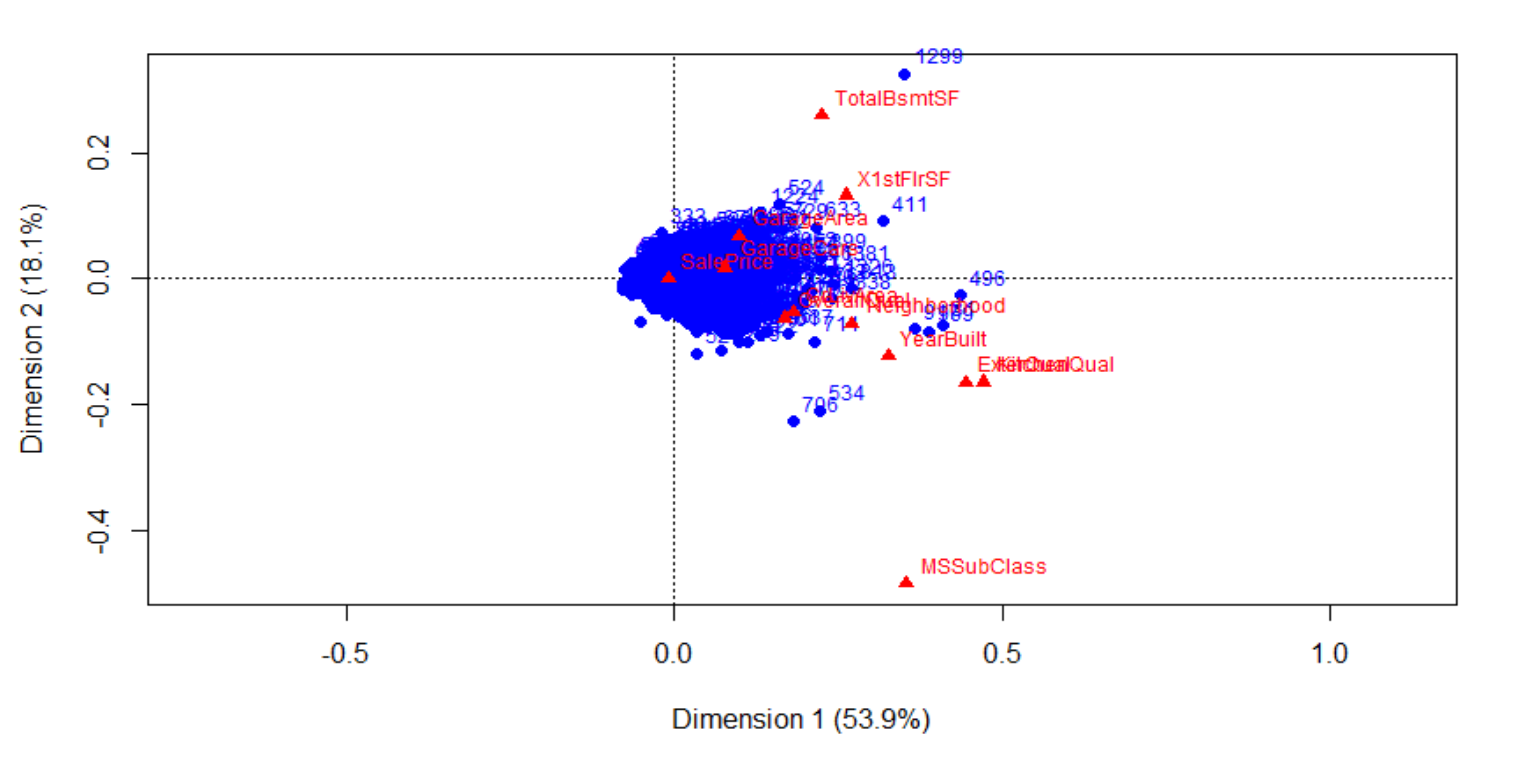


圖十

長的很可怕，因為要將1000多個房子分群會分成太多類別。

1. donedata2\_noprice <- donedata2[,-12] #去除房價
2. E.dist <- dist(donedata2\_noprice, method="euclidean") # 歐式距離
3. h.E.cluster <- hclust(E.dist)
4. plot(h.E.cluster, xlab="歐式距離")

# 相關性分析



圖十一

看各房屋與房屋的各種變數的關係似乎也沒什麼意義。

1. library(ca)
2. fit <- ca(donedata2)
3. plot(fit)

# 結論

因為此資料的目的是要預測房價，因此在主成分與因素分析上比要有意義，在分群上面並沒有什麼意義，分群應比較適合用在種類的分類，像是以物種、以房子類型來收集的資料等等，才能做分群分析。之後有機會的話可以將某類別變數整理到rows，就可以以hierarchical clustering 來做類別分析。

而以主成分與因素分析可以得知，有三個類別是互相類似的，分別是廚房品質與外部材料的品質、車庫中的汽車容量大小與車庫面積、以及地下室面積與一樓面積，若是之後在建模型的時候可能可以增加這三個類別的模擬數據。