Multivariate Analysis

—GROUP PROJECT

Analysis of Boston House-Price

何钰佳　1530005008
彭博雅　1530005027
唐一菡　1530012043
王萌萌　1530005030
杨　璐　1530005042

Contents

PART ONE

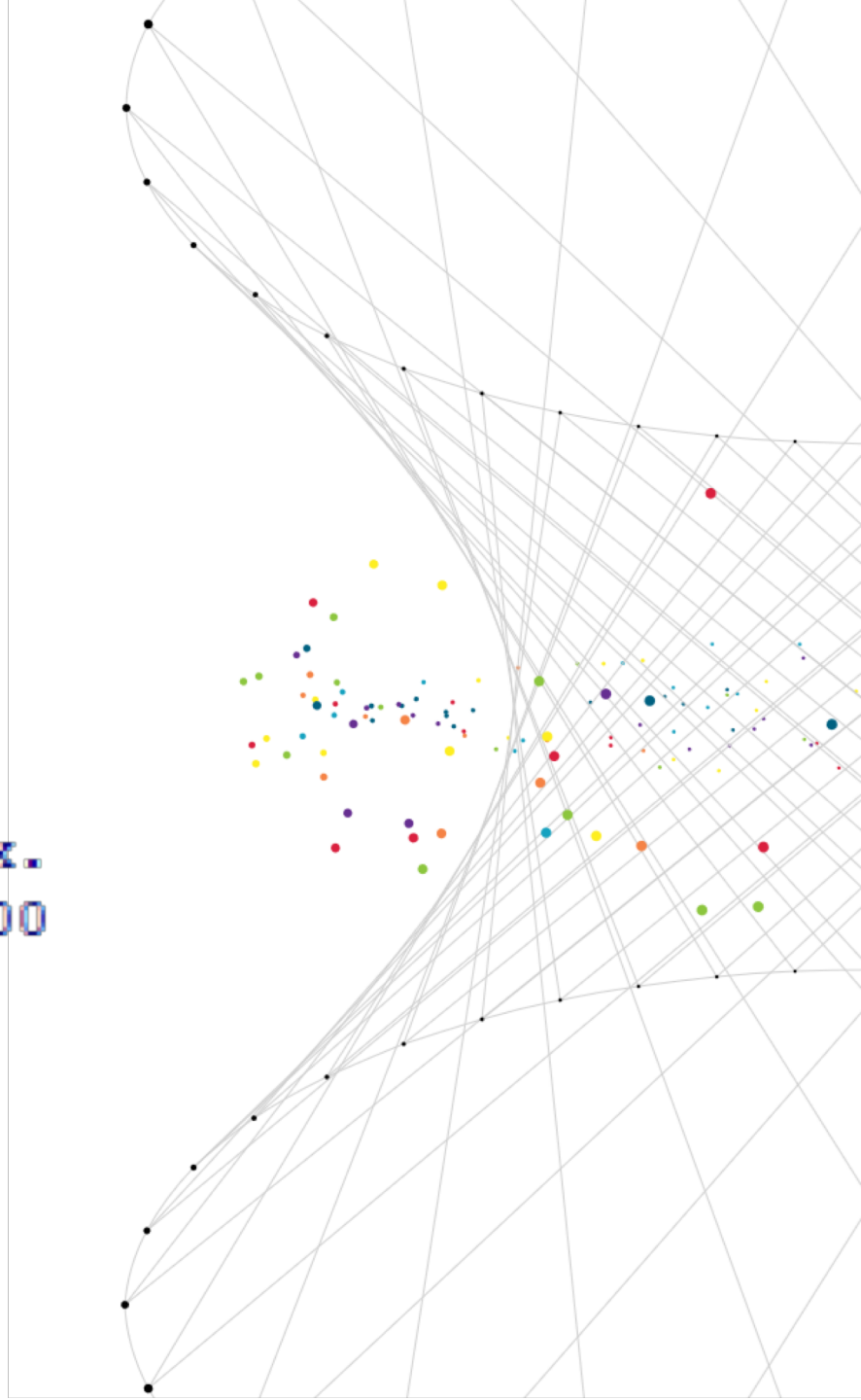
Introduction

Introduction of variables

The dependent variable

- MEDV: *median value* of owner-occupied homes in $1000's

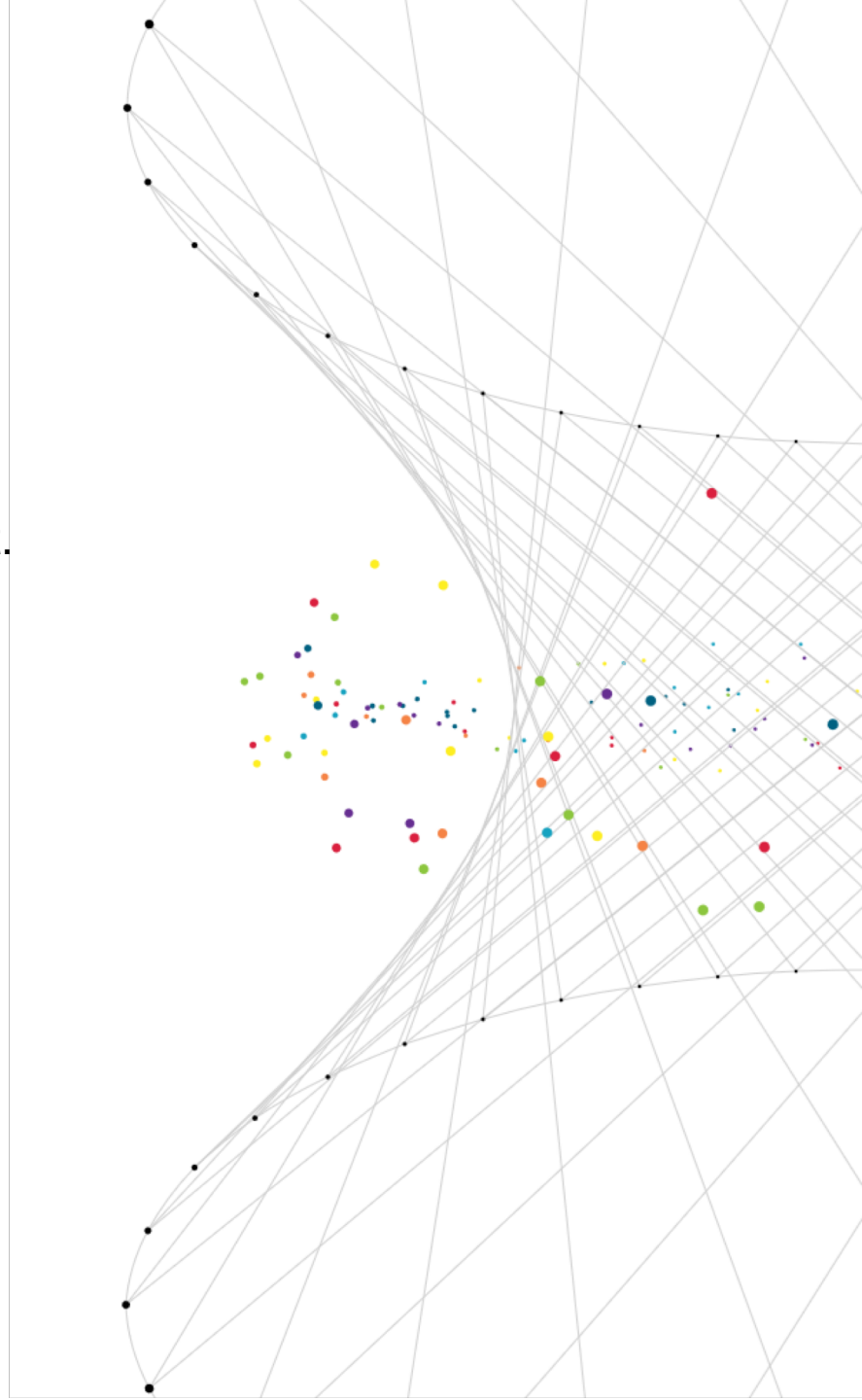
```
> summary(MEDV)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.00   17.02   21.20   22.53   25.00   50.00
```



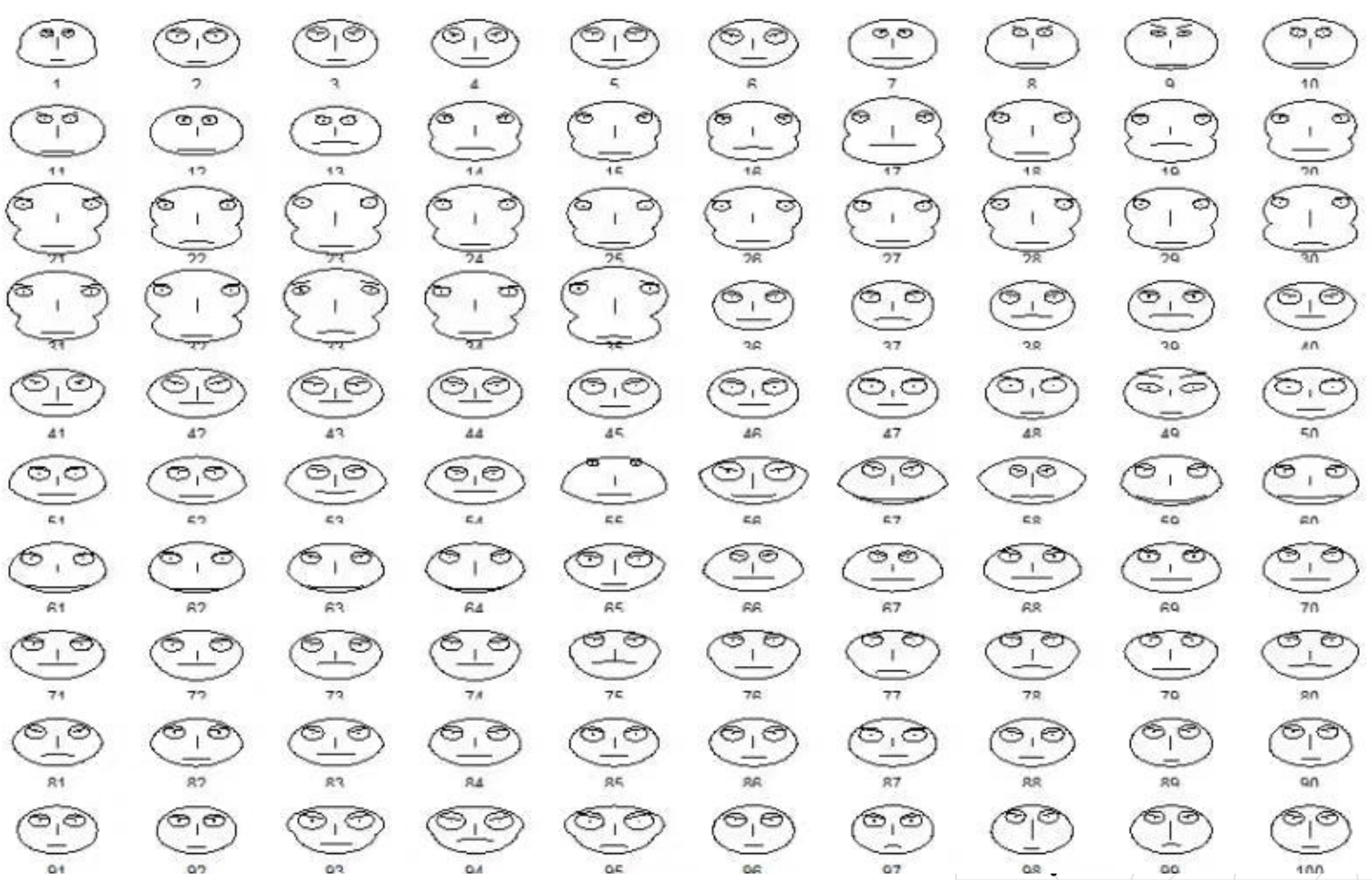
Introduction of variables

The independent variable

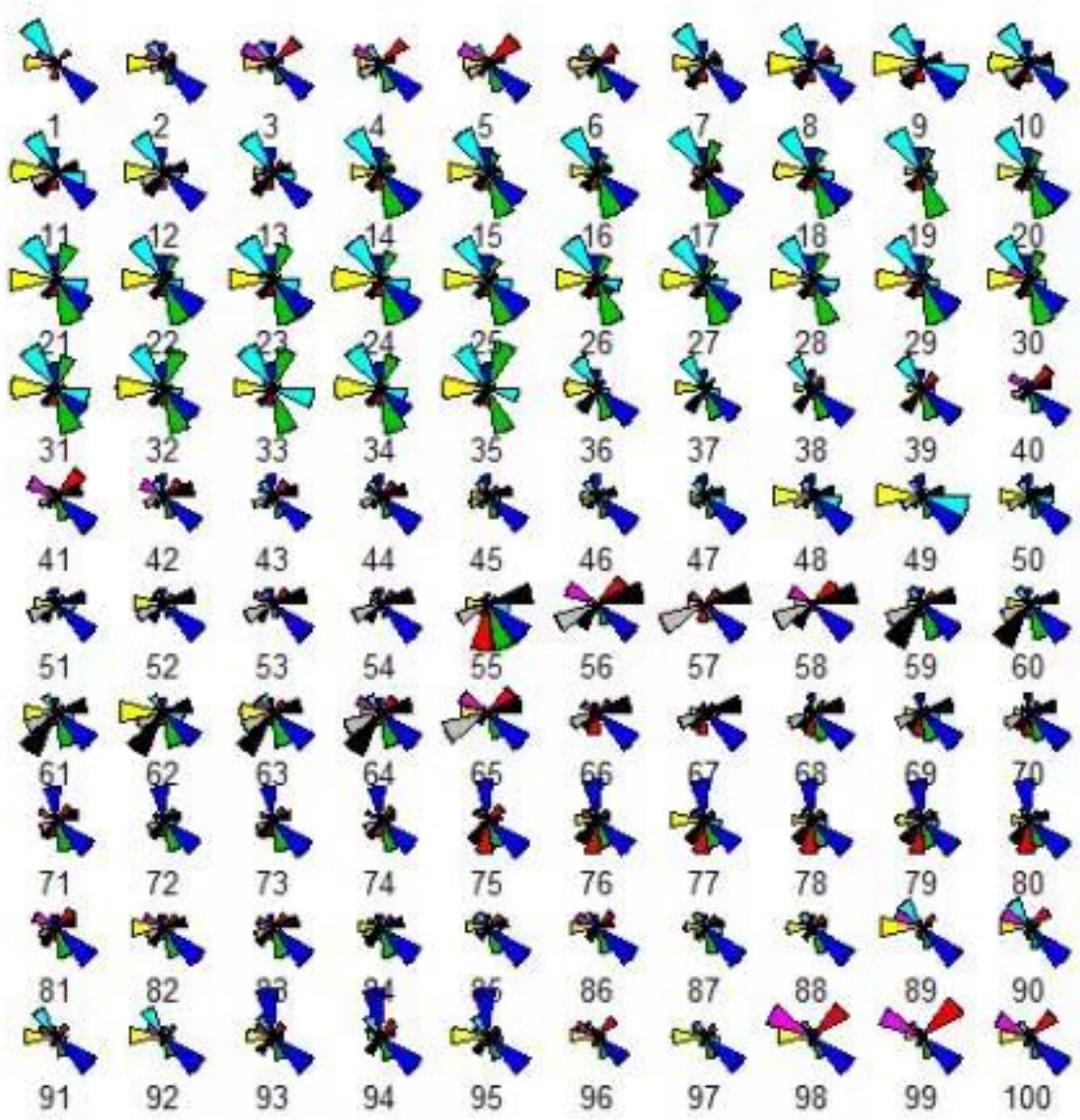
- TRACT: census of tract
- CRIM: per capita crime rate by town
- ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS: proportion of non-retail business acres per town
- CHAS: close to Charles river or not
- NOX: nitric oxides concentration
- RM: average number of rooms per dwelling
- 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
- PTRATIO: pupil-teacher ratio by town
- B: proportion of blacks by town
- LSTAT: lower status of the population

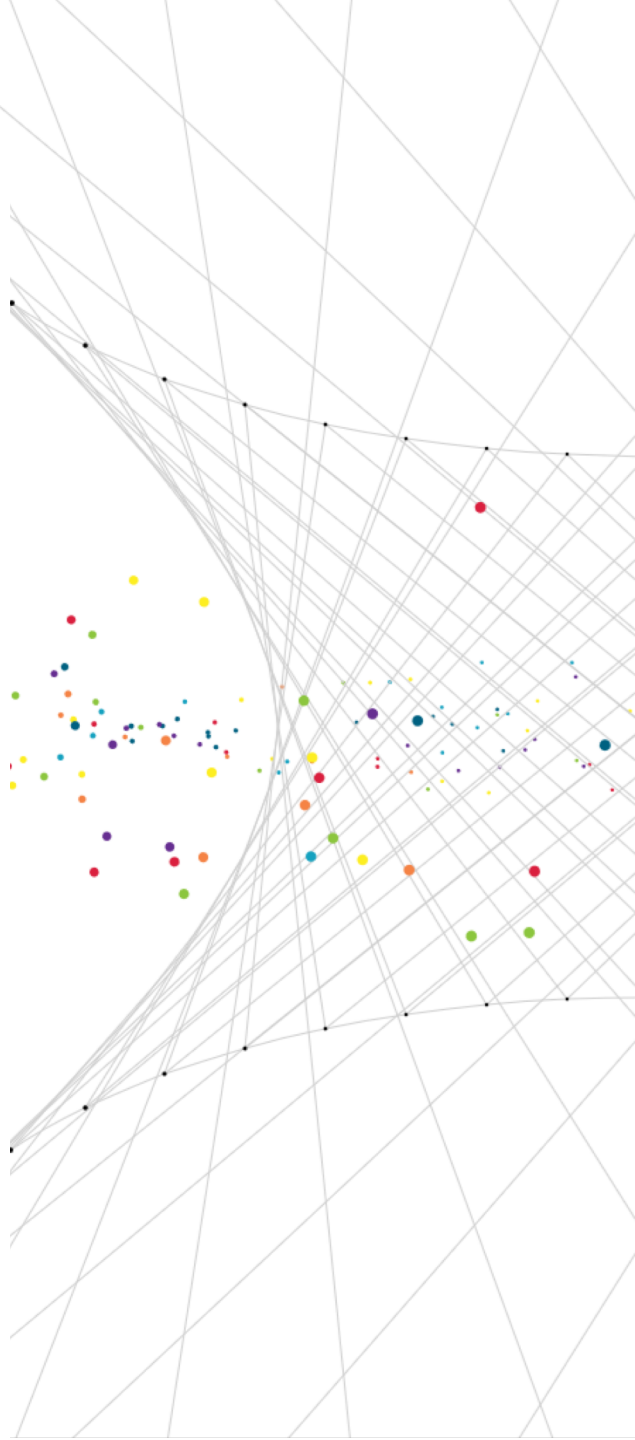


Data Visualization



Data Visualization





Canonical Correlation Analysis

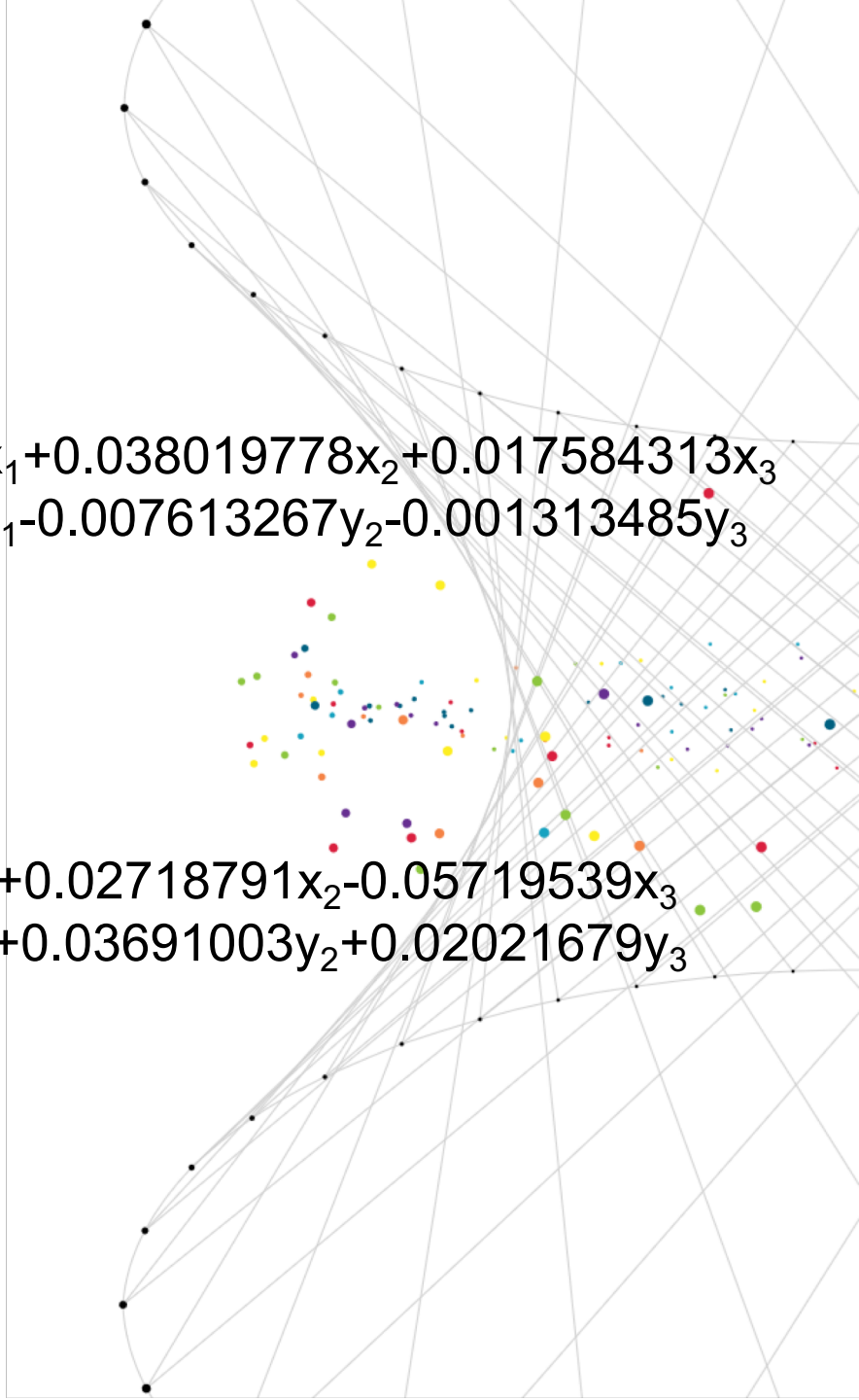
```
> cancor.test(xy[,1:3],xy[,4:6],plot=T)

$`cor`
[1] 0.81787903 0.21490036 0.04835866

$xcoef
              [,1]        [,2]        [,3]
RM    -0.002927027  0.02718791  0.05562524
DIS    0.038019778  0.02314537 -0.01162946
MEDV   0.017584313 -0.05719539 -0.01856723

$ycoef
              [,1]        [,2]        [,3]
NOX   -0.039856085 -0.03524798 -0.01854652
CRIM  -0.007613267  0.03691003 -0.04295722
RAD   -0.001313485  0.02021679  0.06229590

$xcenter
            RM             DIS            MEDV
-2.035729e-17 -1.966477e-17 -3.651147e-18

$ycenter
           NOX            CRIM             RAD
-4.031689e-18 -6.707483e-18  1.755293e-17

cancor test:
                r           Q             P
[1,] 0.81787903 579.251339 0.000000e+00
[2,] 0.21490036  24.836748 5.425859e-05
[3,] 0.04835866   1.169479 2.795084e-01
> |
```

$U1=-0.002927027x_1+0.038019778x_2+0.017584313x_3$

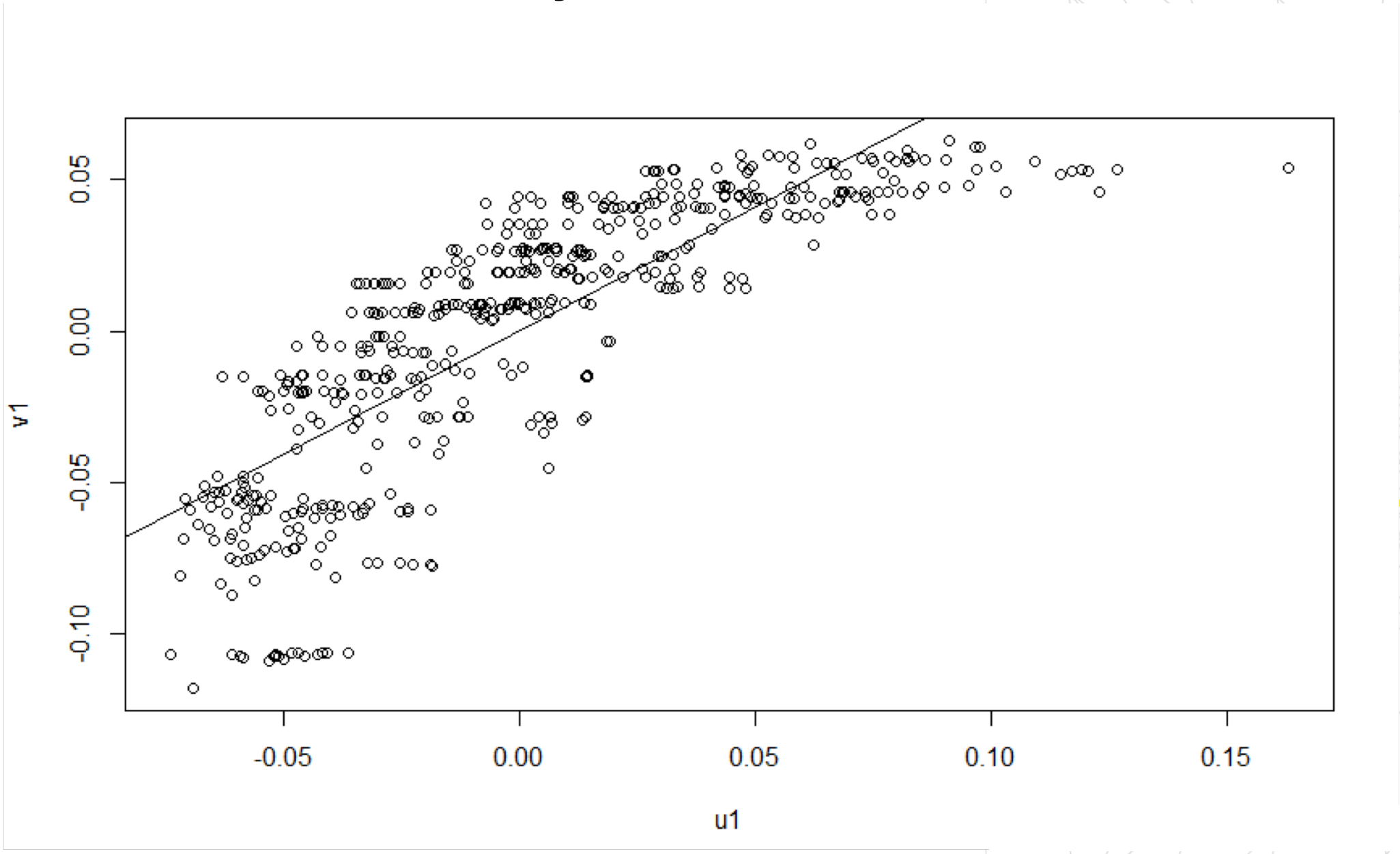
$V1=-0.039856085y_1-0.007613267y_2-0.001313485y_3$

$U2=-0.02718791x_1+0.02718791x_2-0.05719539x_3$

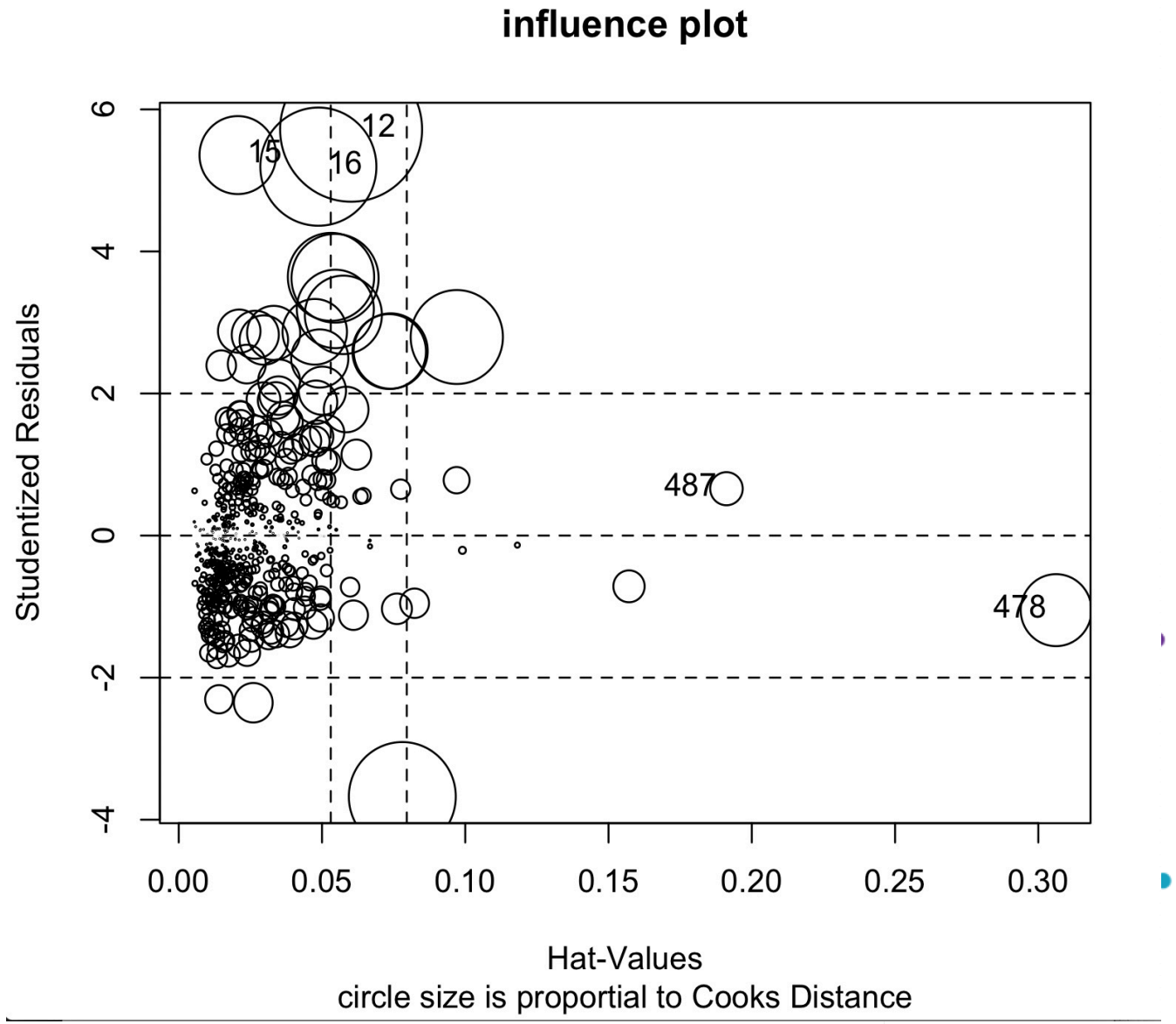
$V2=-0.03524798y_1+0.03691003y_2+0.02021679y_3$

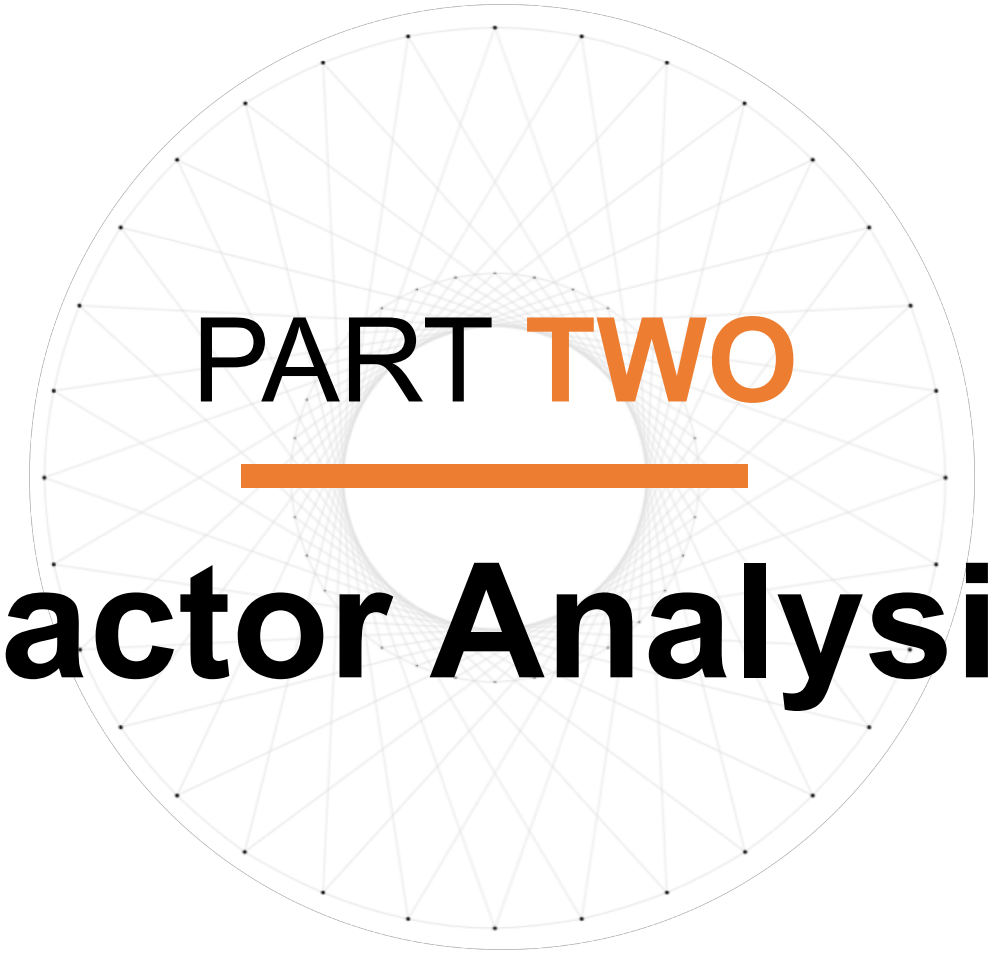


Canonical Correlation Analysis



Outlier Detection





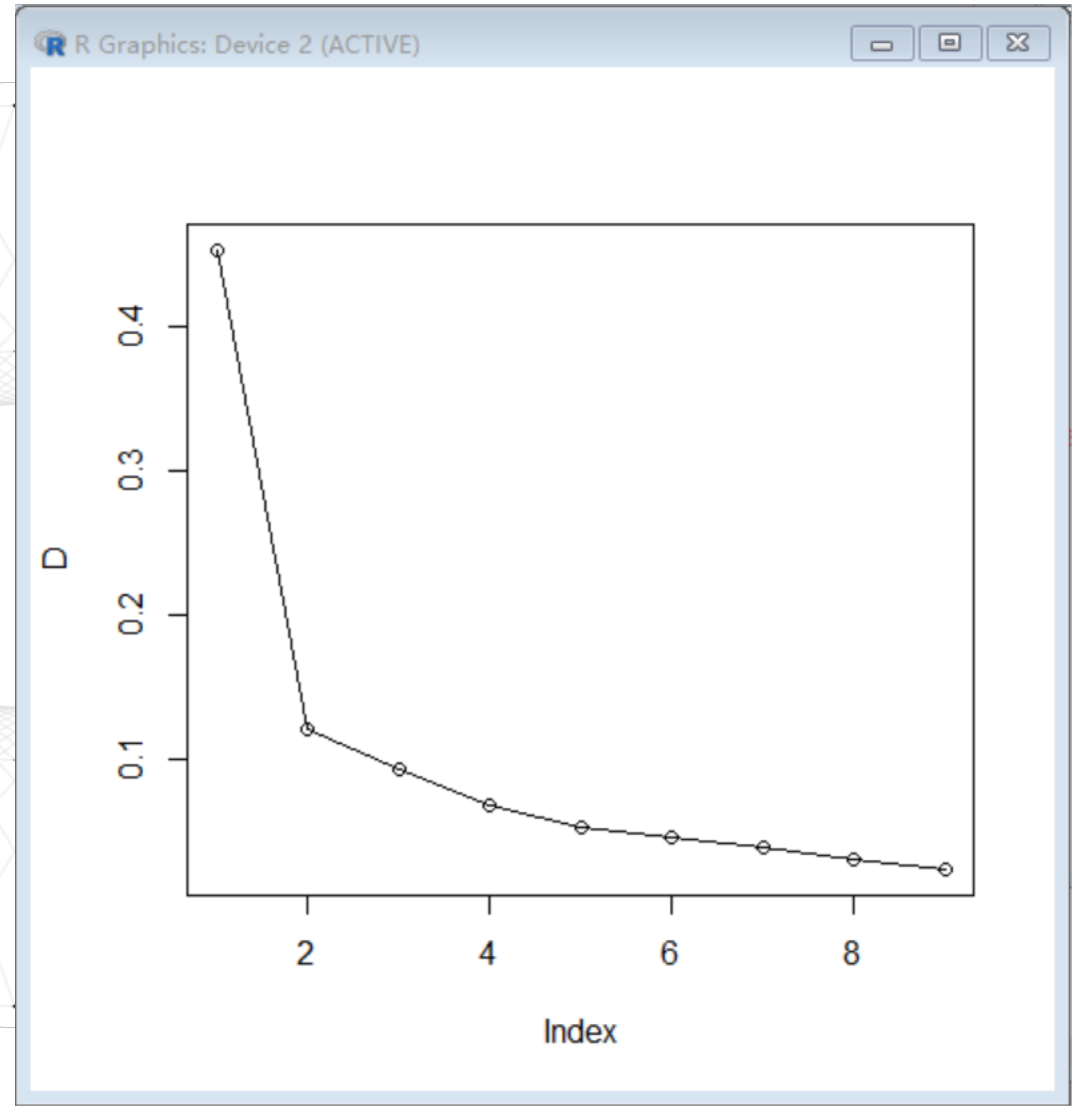
PART TWO

Factor Analysis

Factor Analysis

```
library(mvstats)
Boston=read.csv(file.choose(),header=T)
```

```
Fac1=factpc(Boston,9)
D=Fac1$Vars[,2]
plot(D,type="o")
```



Factor Analysis

```
library(mvstats)
Boston=read.csv(file.choose(),header=T)
Fac1=factpc(Boston,4,rot="varimax")
Fac1$Vars
```

```
Fac1$loadings
```

```
> Fac1$Vars
          Vars Vars.Prop Vars.Cum
Factor1 3.857    22.689    22.69
Factor2 3.803    22.371    45.06
Factor3 3.422    20.130    65.19
Factor4 1.397     8.217    73.41
```

```
> Fac1$loadings
          Factor1   Factor2  Factor3   Factor4
TRACT   -0.80698 -0.321199  0.14332 -0.347145
LON      0.04785 -0.003434 -0.36588  0.584524
LAT     -0.15321  0.134685  0.08680  0.755569
MEDV    -0.27084 -0.173518  0.90651 -0.046787
CMEDV   -0.27292 -0.174226  0.90641 -0.048076
CRIM     0.68857  0.176579 -0.19196 -0.085345
ZN      -0.07043 -0.712561  0.22924 -0.278940
INDUS    0.42403  0.698986 -0.29114 -0.070439
NOX      0.38094  0.784300 -0.20908 -0.103602
RM      -0.05007 -0.169672  0.81291 -0.099493
AGE      0.23099  0.815494 -0.18184  0.094693
DIS     -0.26694 -0.895493  0.01927 -0.026172
RAD      0.89864  0.298479 -0.10526 -0.077915
TAX      0.83398  0.369866 -0.19997 -0.095138
PTRATIO  0.52167  0.021012 -0.40873  0.392010
B       -0.48240 -0.191856  0.20153  0.264885
LSTAT    0.30551  0.468878 -0.66784 -0.008278
```

Factor Analysis

varimax(Fac1$loadings)

```
> varimax(Fac1$loadings)
$loadings

Loadings:
        Factor1 Factor2 Factor3 Factor4
TRACT   -0.807  -0.321   0.143  -0.347
LON                     -0.366   0.584
LAT     -0.153   0.135           0.756
MEDV    -0.271  -0.174   0.907
CMEDV   -0.273  -0.174   0.906
CRIM     0.689   0.177  -0.192
ZN              -0.713   0.229  -0.279
INDUS    0.424   0.699  -0.291
NOX      0.381   0.784  -0.209  -0.104
RM              -0.170   0.813
AGE      0.231   0.816  -0.182
DIS     -0.267  -0.896
RAD      0.899   0.299  -0.105
TAX      0.834   0.370  -0.200
PTRATIO  0.522           -0.409   0.392
B       -0.482  -0.192   0.202   0.265
LSTAT    0.305   0.469  -0.668

               Factor1 Factor2 Factor3 Factor4
SS loadings      3.857   3.804   3.422   1.397
Proportion Var   0.227   0.224   0.201   0.082
Cumulative Var   0.227   0.451   0.652   0.734

$rotmat
             [,1]         [,2]         [,3]         [,4]
[1,]  1.000e+00  7.750e-05 -2.962e-05 -0.0003089
[2,] -7.747e-05  1.000e+00  2.310e-05  0.0001049
[3,]  2.967e-05 -2.312e-05  1.000e+00  0.0001625
[4,]  3.089e-04 -1.049e-04 -1.625e-04  0.9999999
```

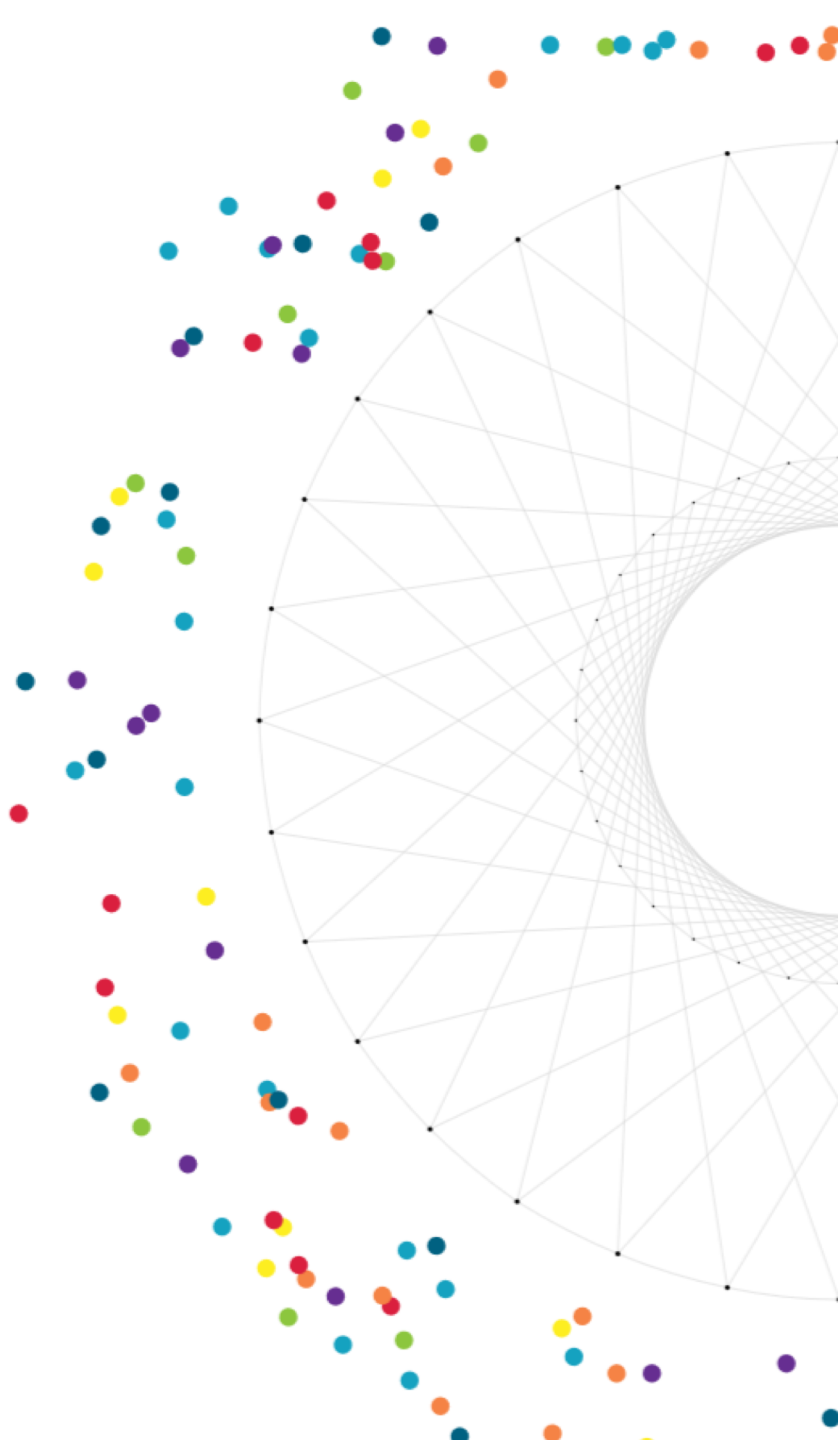
Factor Analysis

Variables	Factor1	Factor2	Factor3	Factor4	Cummunalities	Specific variances
TRACT	-0.807	-0.321	0.143	-0.347	0.895148	0.104852
LON			-0.366	0.584	0.475012	0.524988
LAT	-0.153	0.135		0.756	0.61317	0.38683
MEDV	-0.271	-0.174	0.907		0.926366	0.073634
CMEDV	-0.273	-0.174	0.906		0.925641	0.074359
CRIM	0.689	0.177	-0.192		0.542914	0.457086
ZN		-0.713	0.229	-0.279	0.638651	0.361349
INDUS	0.424	0.699	-0.291		0.753058	0.246942
NOX	0.381	0.784	-0.209	-0.104	0.814314	0.185686
RM		-0.17	0.813		0.689869	0.310131
AGE	0.231	0.816	-0.182		0.752341	0.247659
DIS	-0.267	-0.896			0.874105	0.125895
RAD	0.899	0.299	-0.105		0.908627	0.091373
TAX	0.834	0.37	-0.2		0.872456	0.127544
PTRATIO	0.522		-0.409	0.392	0.593429	0.406571
B	-0.482	-0.192	0.202	0.265	0.380217	0.619783
LSTAT	0.305	0.469	-0.668		0.75921	0.24079
	Human Environment	Environment	Quality of House	Geographical Position		
eig	7.697674	2.052783	1.578281	1.150672		
cumulative	0.452804353	0.573556294	0.666396353	0.734082941		

PART THREE

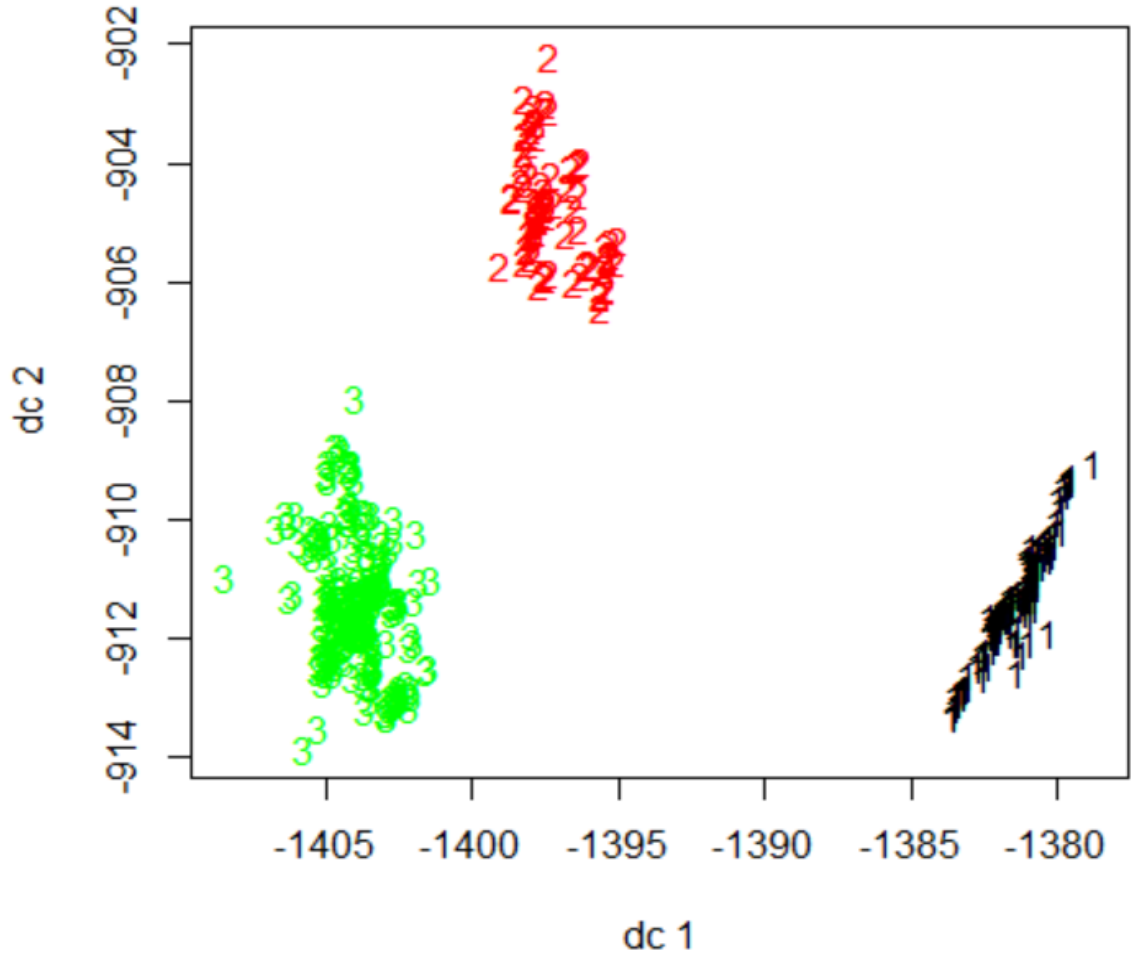
Models

3.1 Cluster



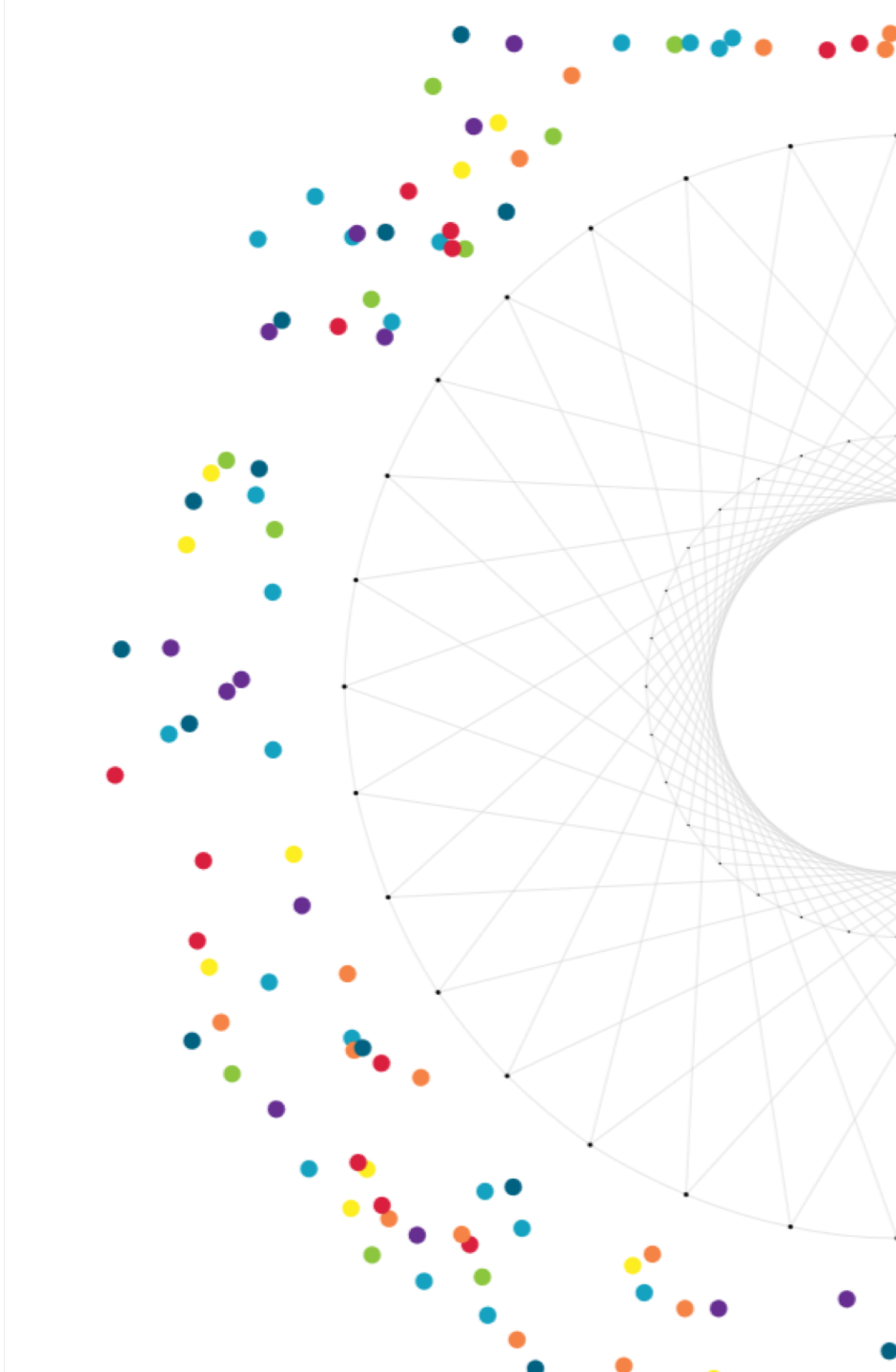
Cluster

Cluster– k-means

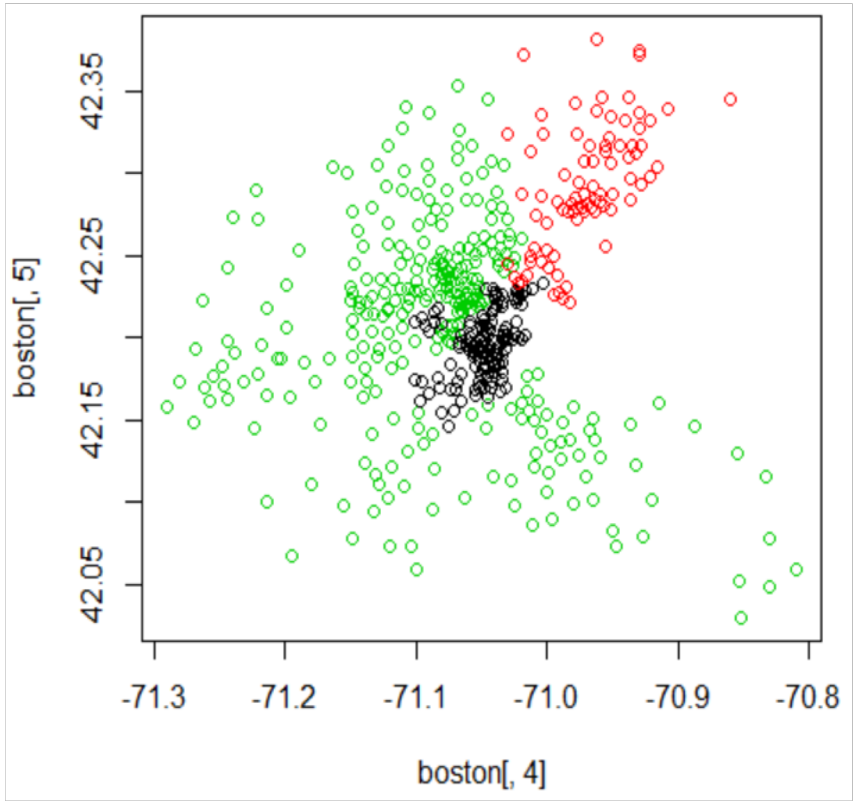


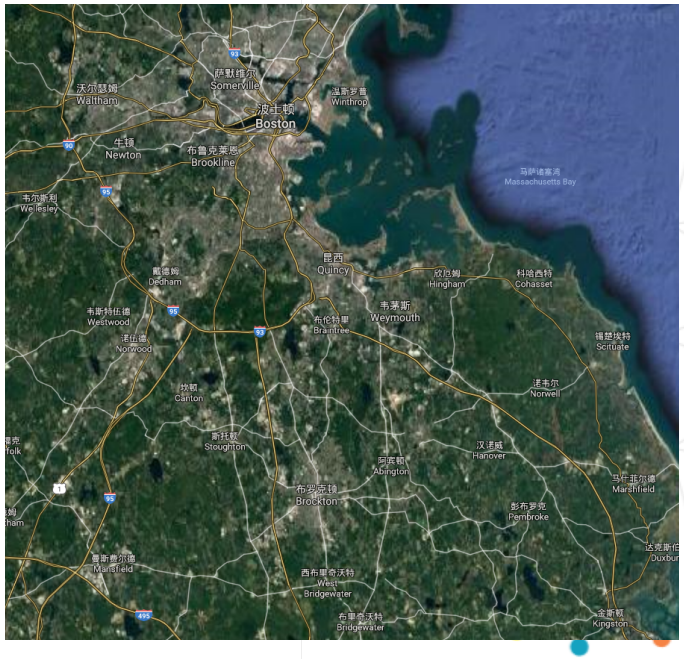
Cluster

```
> summary(first)
                     TOWN        TOWN.         TRACT             LON
 Boston Savin Hill      :23   83     :23   Min.   :   1.0   Min.   :-71.10
 Boston Roxbury         :19   82     :19   1st Qu.: 508.8   1st Qu.:-71.06
 Boston South Boston    :13   80     :13   Median : 810.5   Median :-71.05
 Boston East Boston     :12   79     :12   Mean   : 736.1   Mean   :-71.05
 Boston Dorchester      :11   84     :11   3rd Qu.: 923.2   3rd Qu.:-71.04
 Boston Allston-Brighton: 8   74     : 8   Max.   :1404.0   Max.   :-71.00
 (Other)                :46   (Other):46
      LAT             MEDV            CMEDV            CRIM
 Min.   :42.15   Min.   : 5.00   Min.   : 5.00   Min.   : 2.379
 1st Qu.:42.18   1st Qu.:11.22   1st Qu.:11.22   1st Qu.: 5.686
 Median :42.20   Median :14.40   Median :14.35   Median : 9.085
 Mean   :42.20   Mean   :16.40   Mean   :16.37   Mean   :12.759
 3rd Qu.:42.21   3rd Qu.:19.90   3rd Qu.:19.90   3rd Qu.:14.334
 Max.   :42.23   Max.   :50.00   Max.   :50.00   Max.   :88.976

       ZN          INDUS            NOX              RM
 Min.   :0   Min.   :18.1   Min.   :0.5320   Min.   :3.561
 1st Qu.:0   1st Qu.:18.1   1st Qu.:0.6140   1st Qu.:5.713
 Median :0   Median :18.1   Median :0.6930   Median :6.176
 Mean   :0   Mean   :18.1   Mean   :0.6724   Mean   :6.022
 3rd Qu.:0   3rd Qu.:18.1   3rd Qu.:0.7130   3rd Qu.:6.419
      AGE              DIS              RAD             TAX           PTRATIO
 Min.   : 40.30   Min.   :1.130   Min.   :24   Min.   :666   Min.   :20.2
 1st Qu.: 85.92   1st Qu.:1.589   1st Qu.:24   1st Qu.:666   1st Qu.:20.2
 Median : 94.40   Median :1.943   Median :24   Median :666   Median :20.2
 Mean   : 89.81   Mean   :2.061   Mean   :24   Mean   :666   Mean   :20.2
 3rd Qu.: 98.83   3rd Qu.:2.431   3rd Qu.:24   3rd Qu.:666   3rd Qu.:20.2
 Max.   :100.00   Max.   :4.098   Max.   :24   Max.   :666   Max.   :20.2

       B              LSTAT
 Min.   :  0.32   Min.   : 2.96
 1st Qu.:167.38   1st Qu.:14.18
 Median :373.74   Median :17.91
 Mean   :288.09   Mean   :18.60
 3rd Qu.:395.29   3rd Qu.:23.05
 Max.   :396.90   Max.   :37.97
```

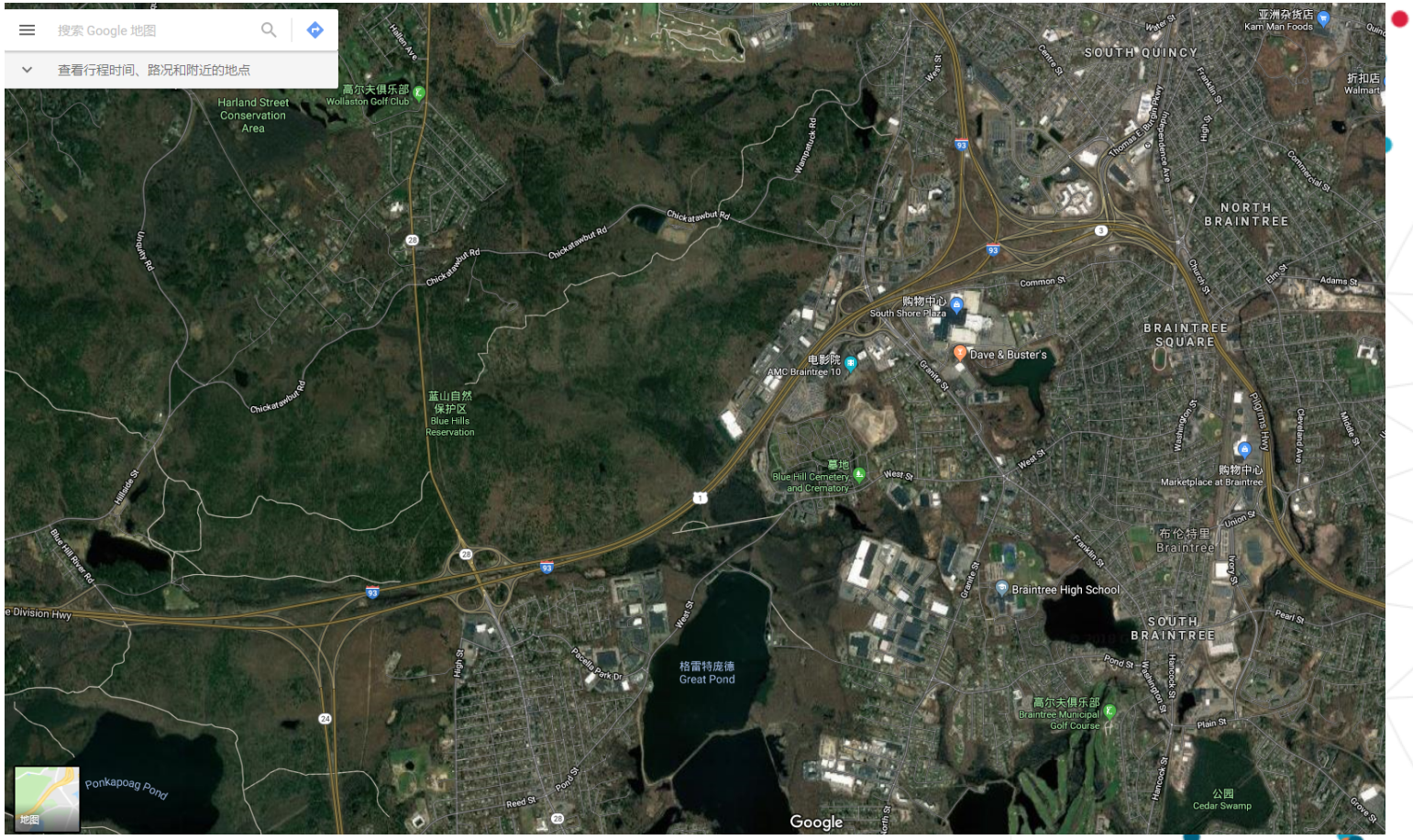


Cluster


boston[, 5]
42.35
42.25
42.15
42.05
-71.3
-71.2
-71.1
-71.0
-70.9
-70.8
boston[, 4]


Somerville
Boston
Winthrop
Waltham
Newton
Brookline
Massachusetts Bay
Quincy
Hingham
Cohasset
Dedham
Weymouth
Braintree
Westwood
Norwood
Scituate
Canton
Norwell
Stoughton
Hanover
Abington
Marshfield
Brockton
Pembroke
Duxbury
Bridgewater
Kingston

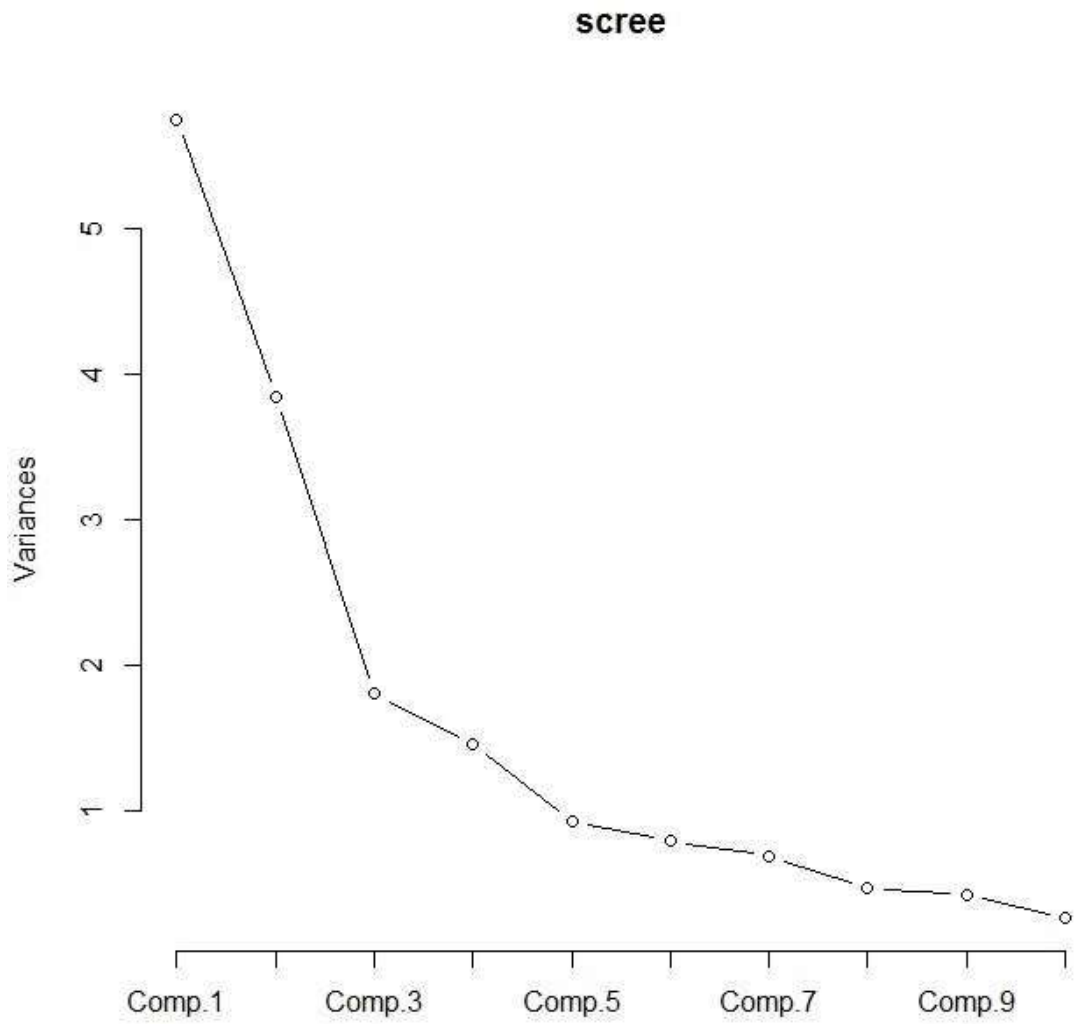
Cluster



The areas with lowest house price are the southwest of Quincy and the north of Braintree and Randolph where are suburbs in Norfolk County, Massachusetts, United States.

3.2 PCA+Multiple Linear Regression

Principle Component Analysis



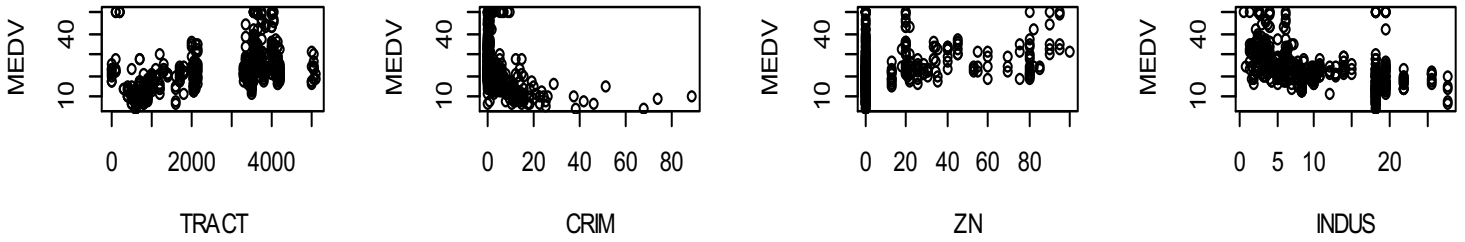
Principle Component Analysis

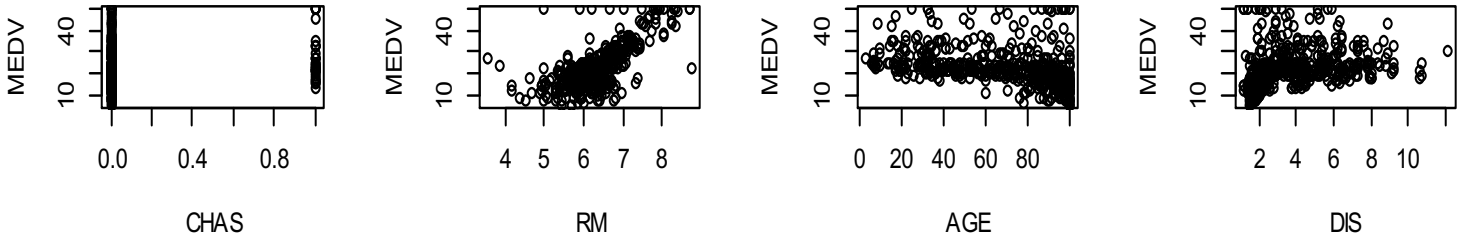
PC<-princomp(first)

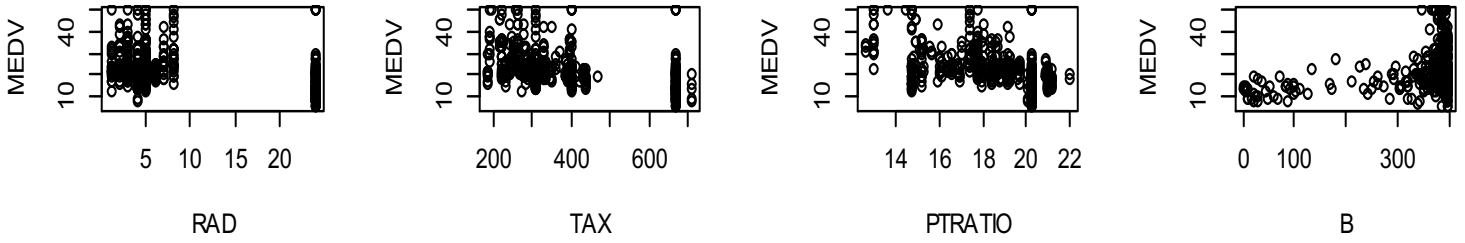
```
Importance of components:
                             Comp.1       Comp.2      Comp.3
Standard deviation     703.7547264 122.53134312 81.12801895
Proportion of Variance   0.9567936   0.02900481  0.01271503
Cumulative Proportion    0.9567936   0.98579840  0.99851342
```

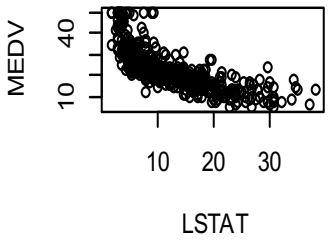
PCA$Loading for first three components

	[,1]	[,2]	[,3]
TRACT	9.708627e-01	0.1186403611	0.2077346449
LON	5.059332e-05	-0.0000656776	-0.0001713802
LAT	5.822420e-05	-0.0001091204	-0.0002579559
CMEDV	2.192098e-03	-0.0214660051	-0.0213880350
CRIM	-8.409911e-03	0.0165521818	-0.0015320211
ZN	7.533344e-03	-0.0081258753	-0.0137283101
INDUS	-6.827497e-03	0.0087614647	0.0327251926
NOX	-1.210349e-04	0.0000635785	0.0001337233
RM	1.270295e-04	0.0002075651	-0.0001822706
AGE	-1.901247e-02	0.0162376272	0.0347177267
DIS	2.097574e-03	-0.0016307838	-0.0035032315
RAD	-1.249246e-02	0.0128317250	0.0259969036
TAX	-2.325699e-01	0.2823096794	0.9278829199
PTRATIO	-7.592746e-04	0.0013222202	0.0029258863
B	5.118526e-02	-0.9511420712	0.3033867887
LSTAT	-3.940483e-03	0.0156245938	0.0146533558

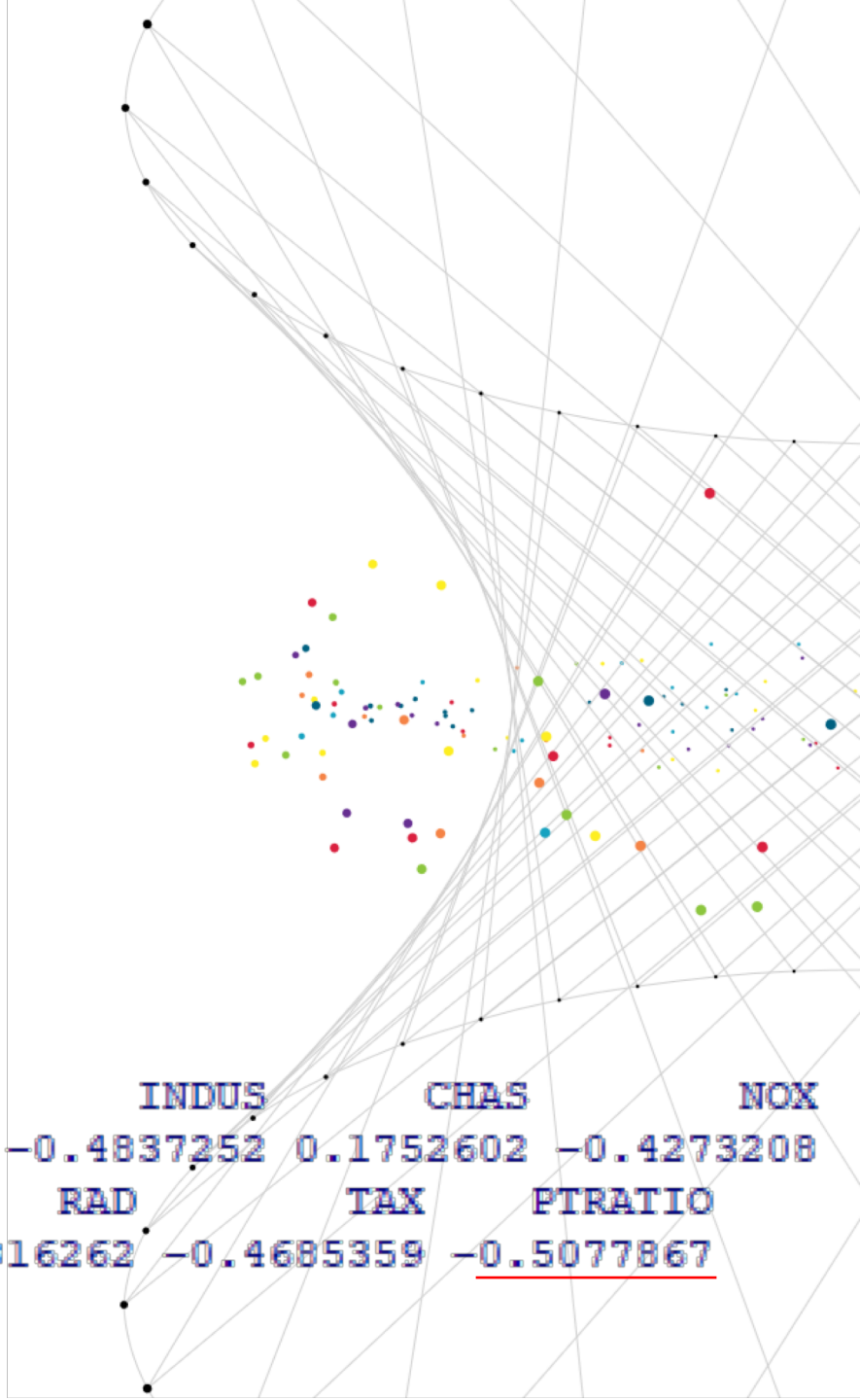









```
> cor(MEDV,a)
           TRACT MEDV       CRIM         ZN      INDUS       CHAS        NOX
[1,] 0.4263792    1 -0.3883046 0.3604453 -0.4837252 0.1752602 -0.4273208
             RM        AGE       DIS        RAD       TAX   PTRATIO
[1,] 0.6953599 -0.3769546 0.2499287 -0.3816262 -0.4685359 -0.5077867
             B      LSTAT
[1,] 0.3334608 -0.7376627
```



Multiple Linear Regression

16
VARIABLES

Crim
Zn
Indus
Nox
Rm
Age
Dis
Rad
Tax
Ptratio
B
Lstat
Cmedv
Tract
Lon
Lat

Variable press

V1

V2

V3

MEDV(Housing price)

Quadratic Model for 1st cluster

```
Call:
lm(formula = y ~ V1 + V2 + V3 + V1 * V2 + V1 * V3 + V2 * V3 +
    V1 * V1 + V2 * V2 + V3 * V3, data = A)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8911 -0.6150  0.0376  0.5897  5.6443

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.190e+01  3.679e+00  -3.236  0.00141 **
V1          -3.799e-03  3.059e-03  -1.242  0.21564
V2          -1.198e-01  2.121e-02  -5.647 5.31e-08 ***
V3           4.036e-02  4.472e-03   9.024  < 2e-16 ***
V1:V2       -5.096e-06  3.793e-06  -1.343  0.18062
V1:V3       -4.158e-06  3.713e-06  -1.120  0.26400
V2:V3        1.622e-04  2.690e-05   6.031 7.35e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.709 on 208 degrees of freedom
Multiple R-squared:  0.9185,    Adjusted R-squared:  0.9161
F-statistic: 390.5 on 6 and 208 DF,  p-value: < 2.2e-16
```

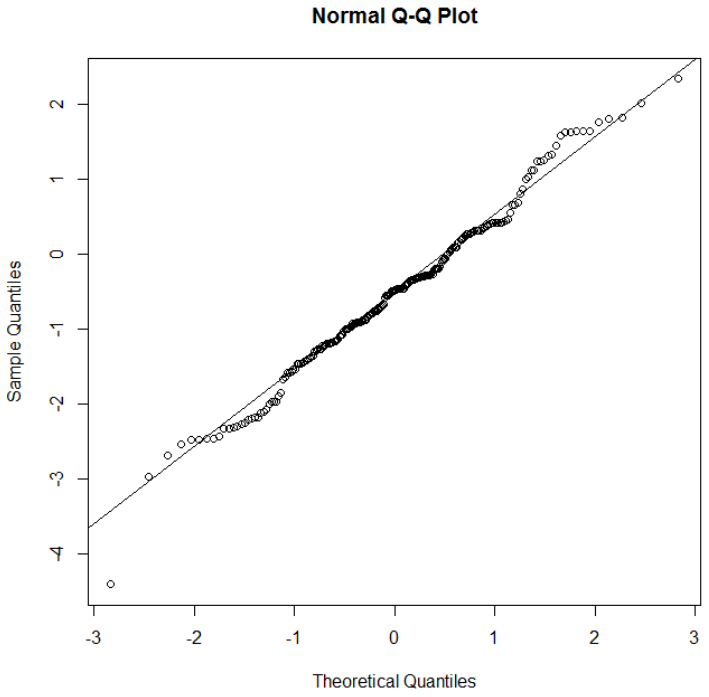
y=V1+V2+V3+V1*V2+V1*V3+V2*V3+V1*V1+V2*V2+V3*V3

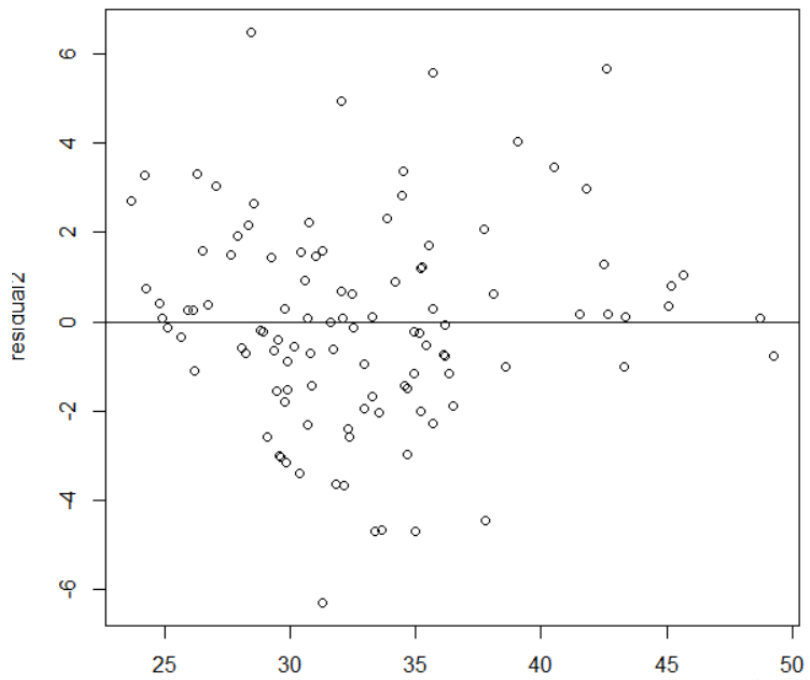
Stepwise Model

```
Call:
lm(formula = y ~ V1 + V2 + V3 + V1 * V2 + V1 * V3 + V2 * V3 +
    V1 * V1 + V2 * V2 + V3 * V3, data = A)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8911 -0.6150  0.0376  0.5897  5.6443

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.190e+01  3.679e+00  -3.236  0.00141 **
V1          -3.799e-03  3.059e-03  -1.242  0.21564
V2          -1.198e-01  2.121e-02  -5.647 5.31e-08 ***
V3           4.036e-02  4.472e-03   9.024  < 2e-16 ***
V1:V2       -5.096e-06  3.793e-06  -1.343  0.18062
V1:V3       -4.158e-06  3.713e-06  -1.120  0.26400
V2:V3        1.622e-04  2.690e-05   6.031 7.35e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.709 on 208 degrees of freedom
Multiple R-squared:  0.9185,    Adjusted R-squared:  0.9161
F-statistic: 390.5 on 6 and 208 DF,  p-value: < 2.2e-16
```

$$y = V1+V2+V3+V1*V2 +V2*V3$$

Model Diagnostic





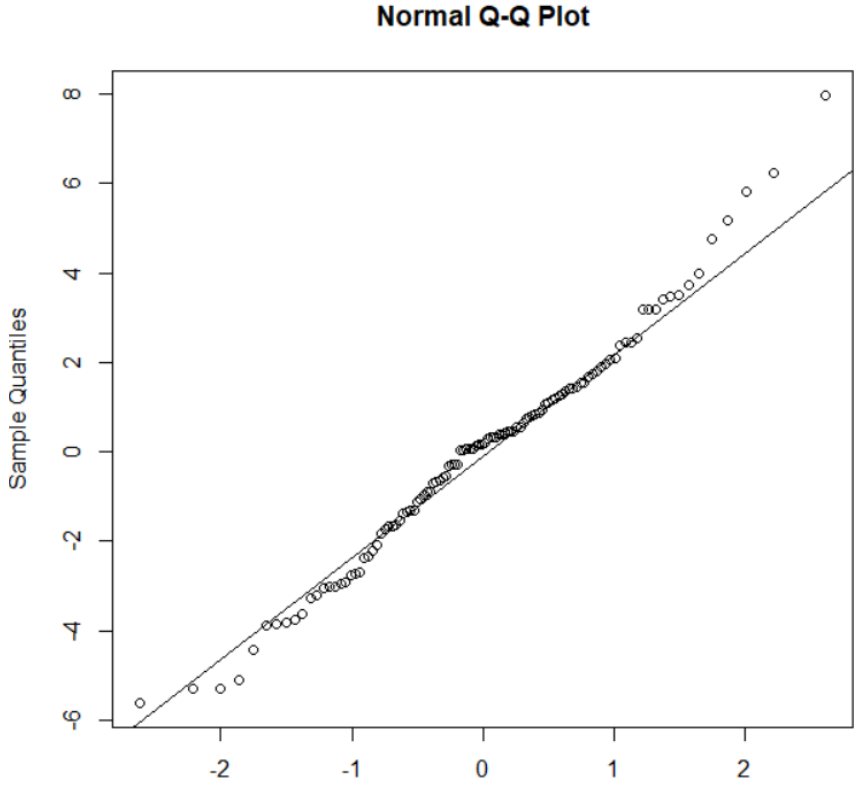

```
> durbinWatsonTest(model1)
 lag Autocorrelation D-W Statistic p-value
   1       0.9050302     0.1894857       0
 Alternative hypothesis: rho != 0
>  ncvTest(model1)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.5065077, Df = 1, p = 0.47665
```

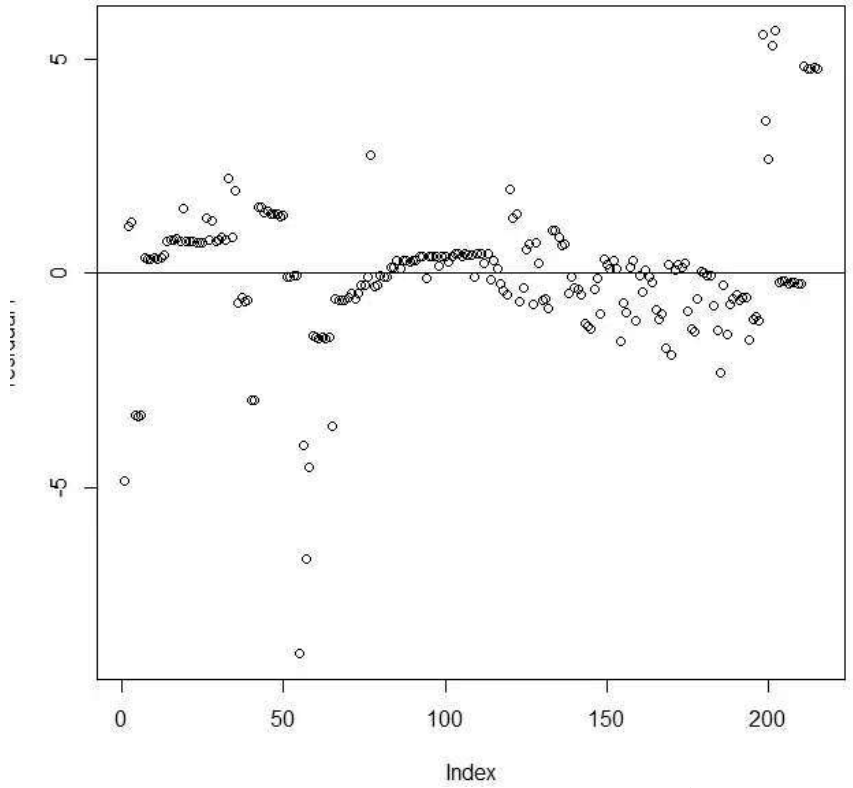
Quadratic Model for 2nd cluster

```
Call:
lm(formula = y ~ V1 + V2 + V3 + V1 * V2 + V1 * V3 + V2 * V3 +
    V1 * V1 + V2 * V2 + V3 * V3, data = A)

Residuals:
     Min       1Q   Median       3Q      Max
-21.8650  -6.2263   0.7453   5.2402  20.6301

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.209e+02  8.703e+01  -1.389  0.16822
V1           6.055e-02  2.006e-02   3.018  0.00329 **
V2           3.749e-01  2.762e-01   1.358  0.17793
V3           1.858e-01  4.614e-01   0.403  0.68814
V1:V2       -7.210e-06  5.578e-05  -0.129  0.89743
V1:V3       -2.095e-04  1.053e-04  -1.989  0.04966 *
V2:V3       -1.410e-03  7.372e-04  -1.913  0.05888 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.236 on 92 degrees of freedom
Multiple R-squared:  0.8994,    Adjusted R-squared:  0.8929
F-statistic: 137.1 on 6 and 92 DF,  p-value: < 2.2e-16
```

$$y=V1+V2+V3+V1*V2+V1*V3+V2*V3+V1*V1+V2*V2+V3*V3$$

Model Diagnostic

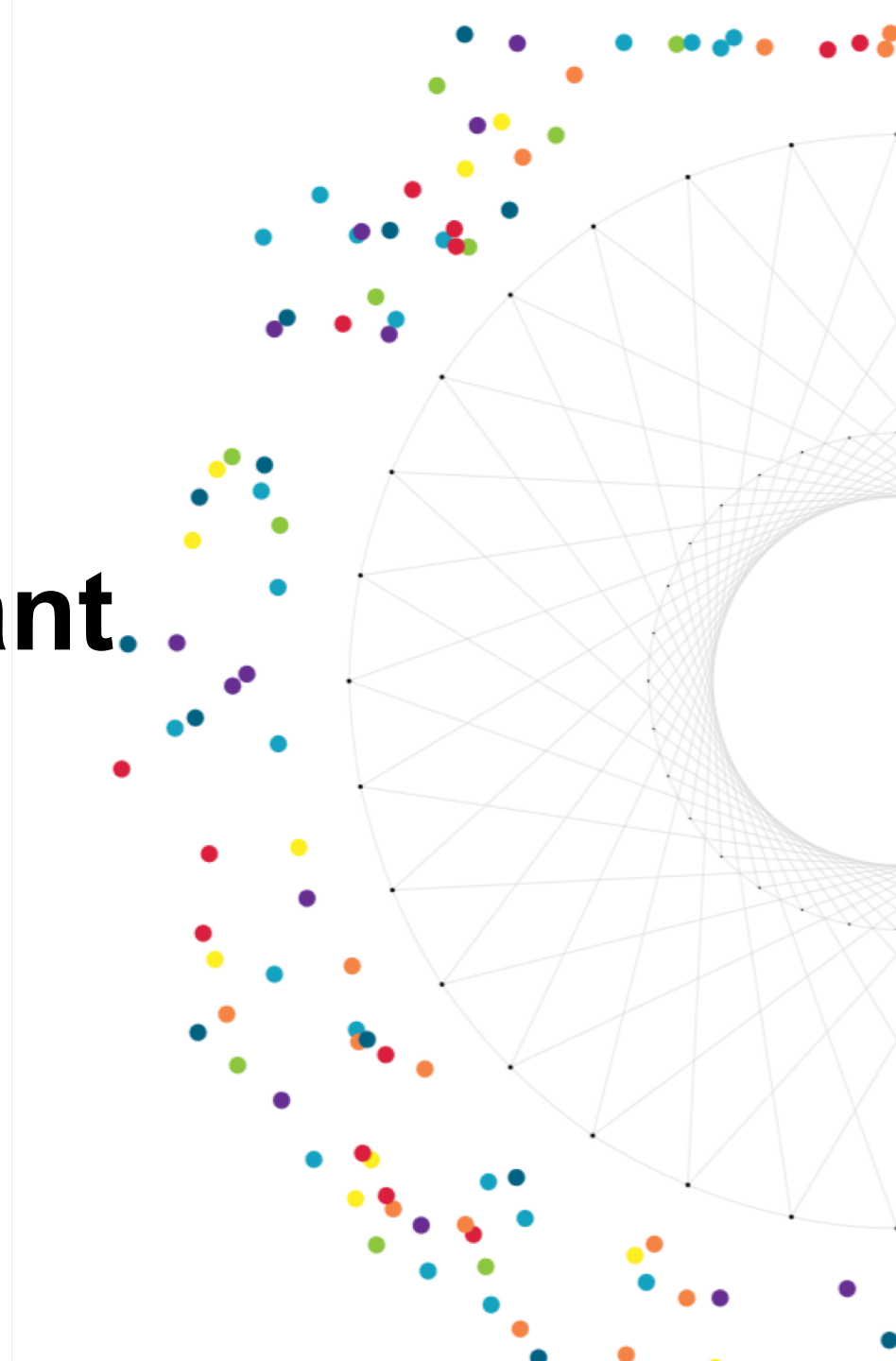





```
> durbinWatsonTest(model1)
 lag Autocorrelation D-W Statistic p-value
   1       0.6233566      0.7353559       0
 Alternative hypothesis: rho != 0
```

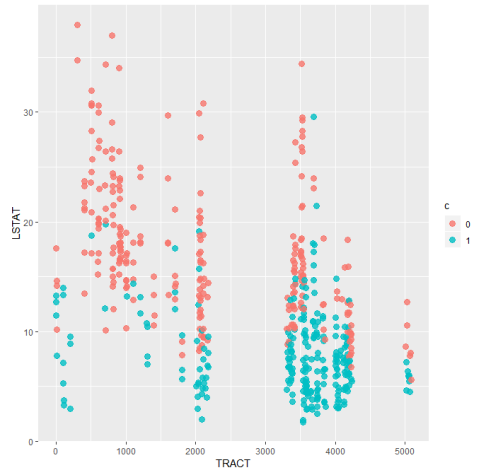
```
>  ncvTest(model1)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.505102, Df = 1, p = 0.47727
```

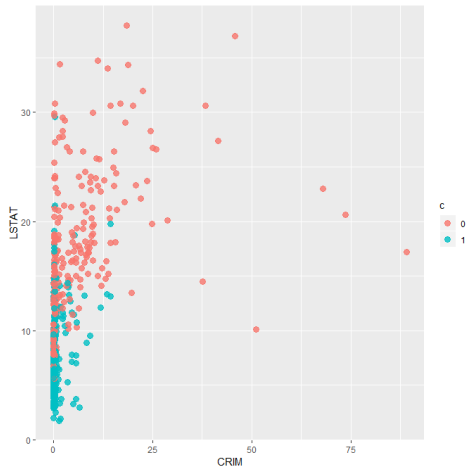
3.3 Discriminant



Discriminant

The relationship between parameter



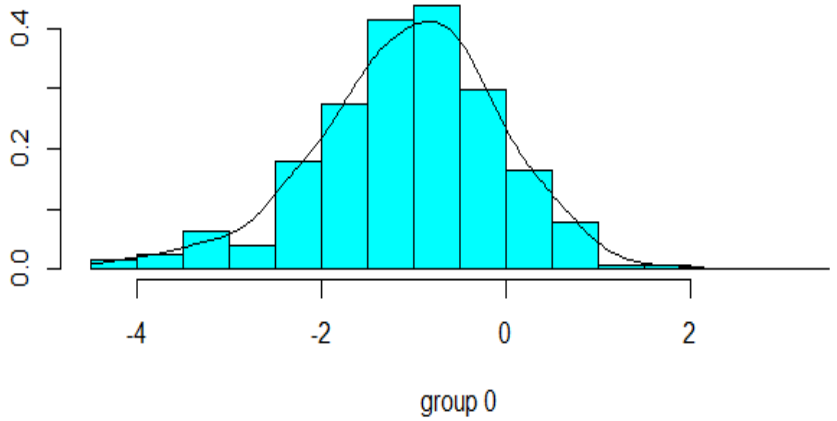


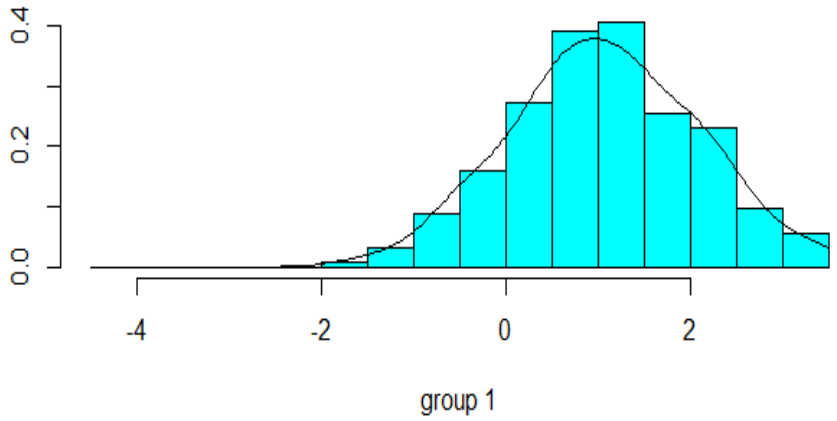
Distance Discriminant

Establish distance discrimination model

```
> fit
Call:
lda(c ~ LON + LAT + NOX + AGE + TRACT + CRIM + LSTAT + RM + PTRATIO,
    data = newdata)

Prior probabilities of groups:
        0         1
0.5059289 0.4940711

Group means:
        LON      LAT       NOX      AGE    TRACT     CRIM    LSTAT       RM
0 -71.03263 42.21599 0.6061750 81.77305 2143.023 6.279721 17.32945 5.931141
1 -71.08072 42.21690 0.5019796 55.06000 3271.064 0.883337  7.86444 6.646612
   PTRATIO
0 19.43906
1 17.44840

Coefficients of linear discriminants:
                  LD1
LON     -3.3867200444
LAT      1.3030471439
NOX     -1.2312291768
AGE     -0.0066111091
TRACT   -0.0000976262
CRIM     0.0016687617
LSTAT   -0.1193820254
RM       0.3381218047
PTRATIO -0.2023099479
```

Model plot



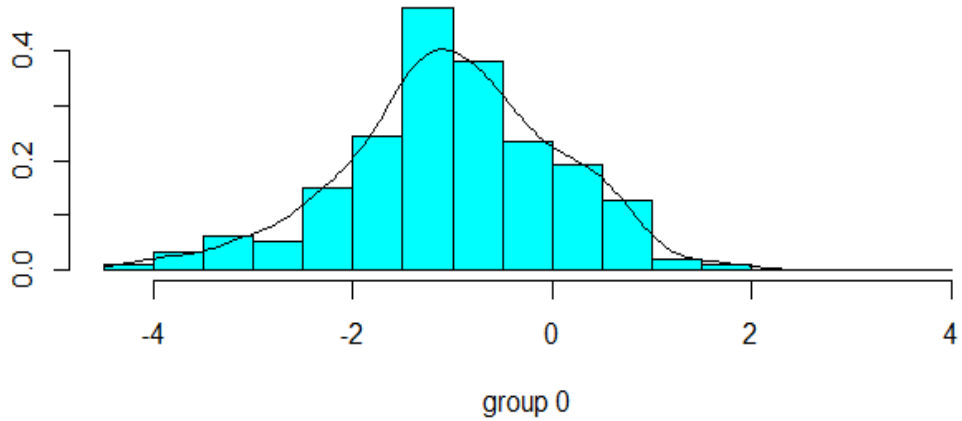


Fisher Discriminant Model

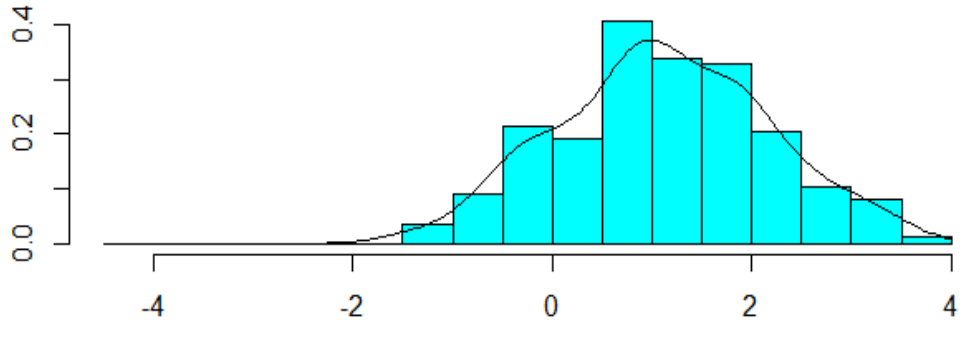
Establish distance discrimination model

```
> fisher_model
Call:
lda(c ~ ., data = train_data)

Prior probabilities of groups:
        0         1
0.5185185 0.4814815

Group means:
     TRACT       LON     LAT     MEDV    CMEDV      CRIM        ZN     INDUS
0 2167.659 -71.03118 42.2156 16.04725 16.08077 6.7373323  4.461538 13.759451
1 3258.929 -71.08702 42.2169 28.72899 28.72604 0.8826983 19.014793  7.775385
        NOX       RM      AGE      DIS       RAD      TAX  PTRATIO        B
0 0.5945044 5.911478 80.78956 3.302527 12.423077 476.5604 19.54341 323.5019
1 0.5036219 6.637970 56.66450 4.402639  6.461538 334.0118 17.46272 385.9707
      LSTAT
0 17.249066
1  8.123787

Coefficients of linear discriminants:
                LD1
TRACT    0.0001108355
LON     -3.1628932466
LAT      1.9735932715
MEDV     0.1907376123
CMEDV   -0.1127629275
CRIM     0.0050223919
ZN      -0.0046606272
INDUS   -0.0106042304
NOX     -0.0710421819
RM       0.2311995601
AGE     -0.0159588104
DIS     -0.1252047409
RAD      0.0939747643
TAX     -0.0041732272
PTRATIO -0.1313722579
B        0.0018302863
LSTAT   -0.0436094245
```

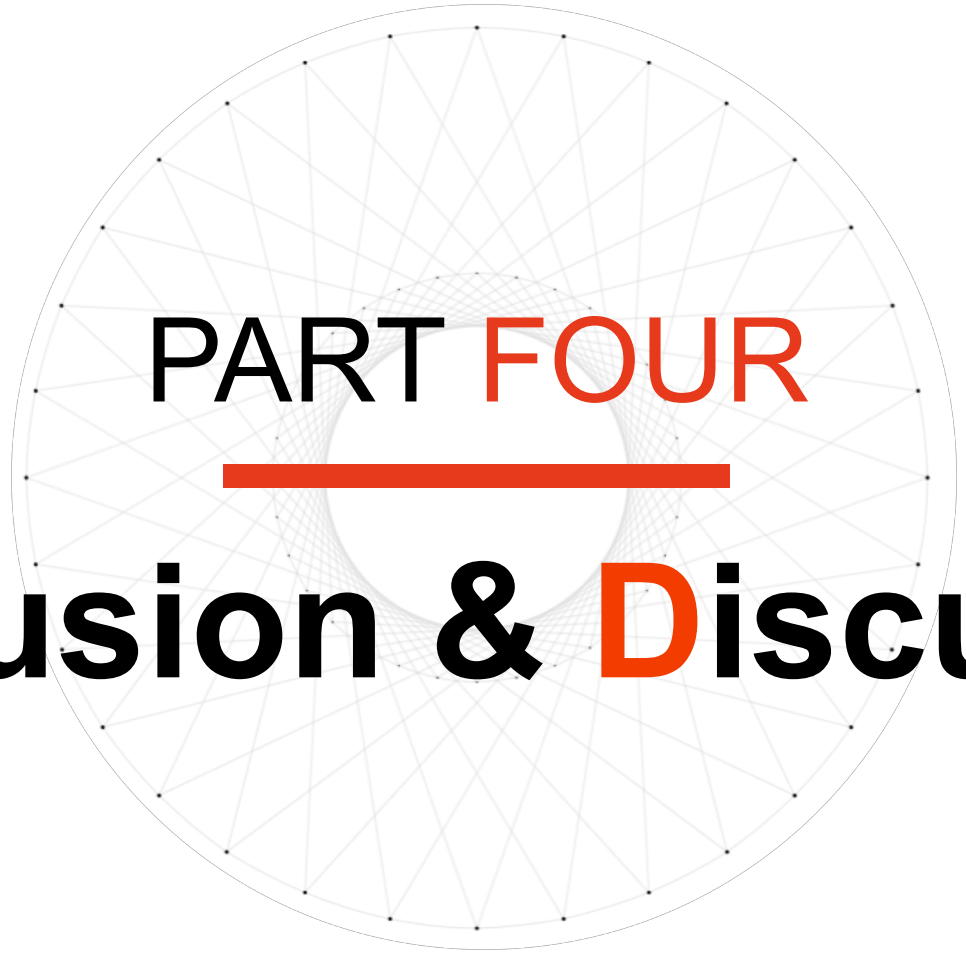
Model plot



group 0



group 1



PART FOUR

Conclusion & Discussion

Conclusion

1. By CCA and FA result, there exists significant relationship among factors.
2. The cluster is pretty good among this dataset.
3. The model we built is very efficient in fit and prediction.
4. Classify data according to its features, then build the model to predict.

Further improvement

1. Stepwise discriminant method
2. Collect time series data

PART FIVE

Reference

Reference

张淑玲，庞进丽，最小二乘法原理在计量测试中的应用[J]，商丘职业技术学院学报，2008(05)

代大山，测量系统的线性分析[J]，电子质量.，2004(06)

王敏，残差分析在统计中的应用[J]，江苏统计， 2000(08)

沈其君，SAS统计分析[M]，东南大学出版社，2001

张宇山，多元线性回归分析的实例研究[J]，科技信息,2009(09):54-56.

王学仁,王松桂，实用多元统计分析[M]. 上海科学技术出版社，1990

回归分析及其试验设计[M]. 上海教育出版社 ，上海师范大学数学系概率统计教研组 编, 1978

赵玉新.多元线性回归分析中自变量的筛选[J].中国城市经济,2011(27):31-32+34.

数据来源： ‘Hedonic prices and the demand for clean air’, J. Environ. Economics & Management, vol.5, 81-102, 1978.

Q&A

Thank you