group project

Import packages, set seeds.

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.0     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
set.seed(1)
```

Initial Data Preprocessing

Read data, rename columns

```
df = read.csv("CommViolPredUnnormalizedData.txt", header = F)

variables = c(
 'communityname',
 'state',
 'countyCode',
 'communityCode',
 'fold',
```

```
'population',
'householdsize',
'racepctblack',
'racePctWhite',
'racePctAsian',
'racePctHisp',
'agePct12t21',
'agePct12t29',
'agePct16t24',
'agePct65up',
'numbUrban',
'pctUrban',
'medIncome',
'pctWWage',
'pctWFarmSelf',
'pctWInvInc',
'pctWSocSec',
'pctWPubAsst',
'pctWRetire',
'medFamInc',
'perCapInc',
'whitePerCap',
'blackPerCap',
'indianPerCap',
'AsianPerCap',
'OtherPerCap',
'HispPerCap',
'NumUnderPov',
'PctPopUnderPov',
'PctLess9thGrade',
'PctNotHSGrad',
'PctBSorMore',
'PctUnemployed',
'PctEmploy',
'PctEmplManu',
'PctEmplProfServ',
'PctOccupManu',
'PctOccupMgmtProf',
'MalePctDivorce',
'MalePctNevMarr',
'FemalePctDiv',
```

```
'TotalPctDiv',
'PersPerFam',
'PctFam2Par',
'PctKids2Par',
'PctYoungKids2Par',
'PctTeen2Par',
'PctWorkMomYoungKids',
'PctWorkMom',
'NumKidsBornNeverMar',
'PctKidsBornNeverMar',
'NumImmig',
'PctImmigRecent',
'PctImmigRec5',
'PctImmigRec8',
'PctImmigRec10',
'PctRecentImmig',
'PctRecImmig5',
'PctRecImmig8',
'PctRecImmig10',
'PctSpeakEnglOnly',
'PctNotSpeakEnglWell',
'PctLargHouseFam',
'PctLargHouseOccup',
'PersPerOccupHous',
'PersPerOwnOccHous',
'PersPerRentOccHous',
'PctPersOwnOccup',
'PctPersDenseHous',
'PctHousLess3BR',
'MedNumBR',
'HousVacant',
'PctHousOccup',
'PctHousOwnOcc',
'PctVacantBoarded',
'PctVacMore6Mos',
'MedYrHousBuilt',
'PctHousNoPhone',
'PctWOFullPlumb',
'OwnOccLowQuart',
'OwnOccMedVal',
'OwnOccHiQuart',
```

```
'OwnOccQrange',
'RentLowQ',
'RentMedian',
'RentHighQ',
'RentQrange',
'MedRent',
'MedRentPctHousInc',
'MedOwnCostPctInc',
'MedOwnCostPctIncNoMtg',
'NumInShelters',
'NumStreet',
'PctForeignBorn',
'PctBornSameState',
'PctSameHouse85',
'PctSameCity85',
'PctSameState85',
'LemasSwornFT',
'LemasSwFTPerPop',
'LemasSwFTFieldOps',
'LemasSwFTFieldPerPop',
'LemasTotalReq',
'LemasTotReqPerPop',
'PolicReqPerOffic',
'PolicPerPop',
'RacialMatchCommPol',
'PctPolicWhite',
'PctPolicBlack',
'PctPolicHisp',
'PctPolicAsian',
'PctPolicMinor',
'OfficAssgnDrugUnits',
'NumKindsDrugsSeiz',
'PolicAveOTWorked',
'LandArea',
'PopDens',
'PctUsePubTrans',
'PolicCars',
'PolicOperBudg',
'LemasPctPolicOnPatr',
'LemasGangUnitDeploy',
'LemasPctOfficDrugUn',
```

```
 'PolicBudgPerPop',
 'murders',
 'murdPerPop',
 'rapes',
 'rapesPerPop',
 'robberies',
 'robbbPerPop',
 'assaults',
 'assaultPerPop',
 'burglaries',
 'burglPerPop',
 'larcenies',
 'larcPerPop',
 'autoTheft',
 'autoTheftPerPop',
 'arsons',
 'arsonsPerPop',
 'ViolentCrimesPerPop',
 'nonViolPerPop'
 )

names(df) = variables
```

Filter out data where the target is missing:

```
df = df[df["ViolentCrimesPerPop"] != "?", ]
```

Filter out non-predictive features

The first five are non-predictive features

```
df = df[, -(1:5)]
```

The last 18 are target variables. We are only interested in 1 of them.

```
df = df[, -(ncol(df) + c(0, -2:-17))]
```

Explore data, model building

Initial train test split

```
# shuffled index of test set, 20% percent of the data
idx_test = sample(1:nrow(df))[1:floor(nrow(df) * 0.2)]
# shuffled index of train set
idx_train = sample((1:nrow(df))[-idx_test])

train_set = df[idx_train, ]
test_set = df[idx_test, ]
```

Look at missing data

```
missing_pct = apply(df == "?", 2, \(x) sum(x) / length(x))

train_set = train_set[, !(names(train_set) %in% names(missing_pct[missing_pct > 0]))]
test_set = test_set[, !(names(test_set) %in% names(missing_pct[missing_pct > 0]))]

train_x = train_set[, -ncol(train_set)]
train_y = train_set[, ncol(train_set)]
```

PCA

```
library(nFactors)
```

```
Loading required package: lattice


Attaching package: 'nFactors'

The following object is masked from 'package:lattice':

    parallel
```

```
library(EFA.dimensions)
```

```
**************************************************************************************
EFA.dimensions 0.1.8.1

Please contact Brian O'Connor at brian.oconnor@ubc.ca if you have questions or suggestions.
**************************************************************************************
```

```r
library(GPArotation)
library(psych)
```

```
Attaching package: 'psych'

The following objects are masked from 'package:GPArotation':

    equamax, varimin

The following objects are masked from 'package:ggplot2':

    %+%, alpha
```

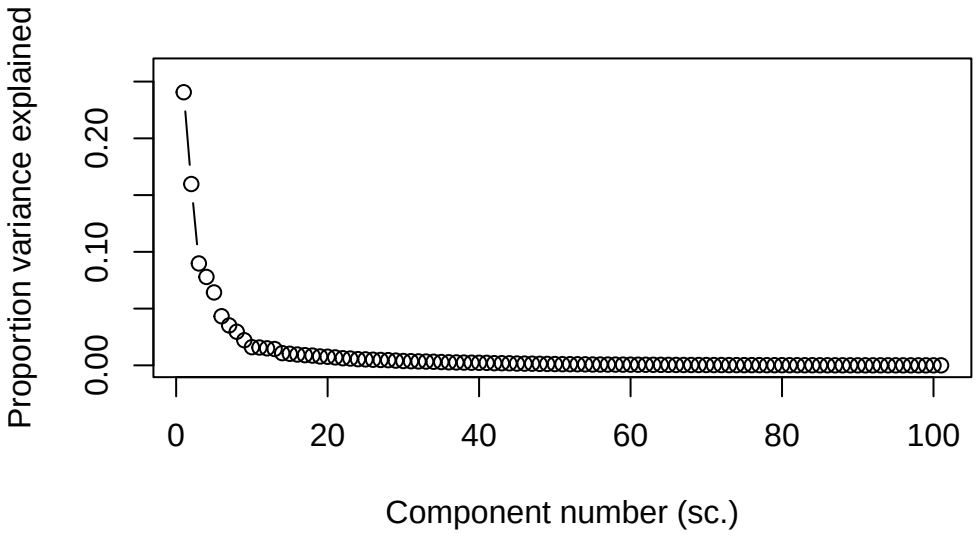
```r
pca_result = prcomp(train_x, scale = T)
```

Determine the number of components

Scree plot

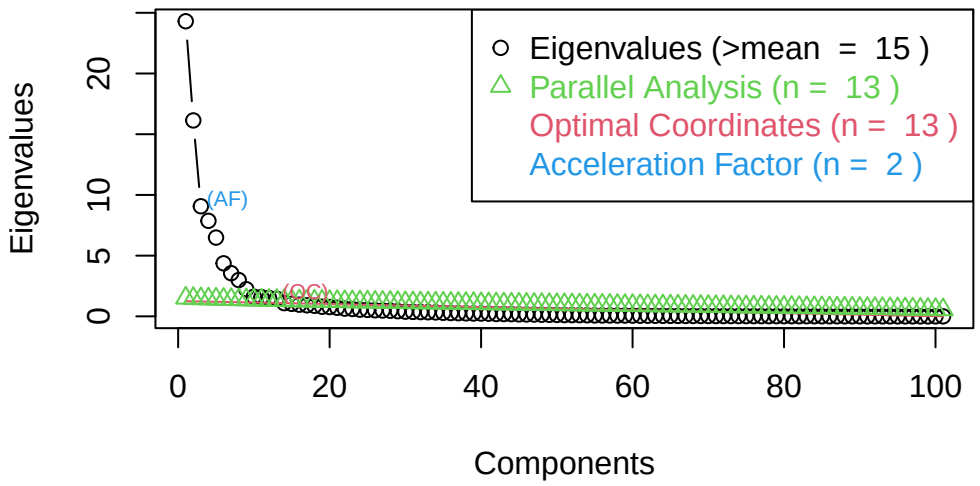
```r
prop_var_expl = summary(pca_result)$importance[2,]

plot(prop_var_expl, type="b", xlab="Component number (sc.)",
ylab="Proportion variance explained", ylim=c(0, 0.26) )
```



Parallel analysis

```
ev = eigen(cor(train_x))
ap = parallel(subject = nrow(train_x), var = ncol(train_x), rep = 1000)
nS = nScree(x = ev$values, aparallel = ap$eigen$qevpea)
plotnScree(nS, main = "")
```



MAP test

```
map_res = MAP(train_x)
```

```
MINIMUM AVERAGE PARTIAL (MAP) TEST

Number of cases = 1596

Number of variables = 101

Specified kind of correlations for this analysis: Pearson

Total Variance Explained (Initial Eigenvalues):
```

	Eigenvalues	Proportion of Variance
Factor 1	24.30	0.24
Factor 2	16.14	0.16
Factor 3	9.07	0.09
Factor 4	7.87	0.08
Factor 5	6.49	0.06
Factor 6	4.37	0.04
Factor 7	3.55	0.04
Factor 8	2.99	0.03
Factor 9	2.24	0.02
Factor 10	1.62	0.02
Factor 11	1.59	0.02
Factor 12	1.51	0.01
Factor 13	1.45	0.01
Factor 14	1.09	0.01
Factor 15	1.03	0.01
Factor 16	0.96	0.01
Factor 17	0.91	0.01
Factor 18	0.87	0.01
Factor 19	0.79	0.01
Factor 20	0.77	0.01
Factor 21	0.71	0.01
Factor 22	0.64	0.01
Factor 23	0.61	0.01
Factor 24	0.55	0.01
Factor 25	0.53	0.01
Factor 26	0.50	0.00
Factor 27	0.48	0.00
Factor 28	0.46	0.00
Factor 29	0.42	0.00
Factor 30	0.39	0.00
Factor 31	0.37	0.00
Factor 32	0.35	0.00
Factor 33	0.34	0.00
Factor 34	0.32	0.00
Factor 35	0.29	0.00
Factor 36	0.27	0.00
Factor 37	0.26	0.00
Factor 38	0.25	0.00
Factor 39	0.23	0.00
Factor 40	0.23	0.00
Factor 41	0.22	0.00
Factor 42	0.19	0.00

Factor 43	0.19	0.00
Factor 44	0.18	0.00
Factor 45	0.16	0.00
Factor 46	0.15	0.00
Factor 47	0.15	0.00
Factor 48	0.14	0.00
Factor 49	0.13	0.00
Factor 50	0.12	0.00
Factor 51	0.11	0.00
Factor 52	0.10	0.00
Factor 53	0.10	0.00
Factor 54	0.09	0.00
Factor 55	0.08	0.00
Factor 56	0.07	0.00
Factor 57	0.07	0.00
Factor 58	0.07	0.00
Factor 59	0.07	0.00
Factor 60	0.06	0.00
Factor 61	0.06	0.00
Factor 62	0.06	0.00
Factor 63	0.05	0.00
Factor 64	0.05	0.00
Factor 65	0.04	0.00
Factor 66	0.04	0.00
Factor 67	0.03	0.00
Factor 68	0.03	0.00
Factor 69	0.03	0.00
Factor 70	0.03	0.00
Factor 71	0.03	0.00
Factor 72	0.03	0.00
Factor 73	0.02	0.00
Factor 74	0.02	0.00
Factor 75	0.02	0.00
Factor 76	0.02	0.00
Factor 77	0.02	0.00
Factor 78	0.02	0.00
Factor 79	0.02	0.00
Factor 80	0.01	0.00
Factor 81	0.01	0.00
Factor 82	0.01	0.00
Factor 83	0.01	0.00
Factor 84	0.01	0.00
Factor 85	0.01	0.00

Factor 86	0.01	0.00
Factor 87	0.01	0.00
Factor 88	0.01	0.00
Factor 89	0.01	0.00
Factor 90	0.00	0.00
Factor 91	0.00	0.00
Factor 92	0.00	0.00
Factor 93	0.00	0.00
Factor 94	0.00	0.00
Factor 95	0.00	0.00
Factor 96	0.00	0.00
Factor 97	0.00	0.00
Factor 98	0.00	0.00
Factor 99	0.00	0.00
Factor 100	0.00	0.00
Factor 101	0.00	0.00

	Cumulative Prop. Variance
Factor 1	0.24
Factor 2	0.40
Factor 3	0.49
Factor 4	0.57
Factor 5	0.63
Factor 6	0.68
Factor 7	0.71
Factor 8	0.74
Factor 9	0.76
Factor 10	0.78
Factor 11	0.79
Factor 12	0.81
Factor 13	0.82
Factor 14	0.83
Factor 15	0.84
Factor 16	0.85
Factor 17	0.86
Factor 18	0.87
Factor 19	0.88
Factor 20	0.89
Factor 21	0.89
Factor 22	0.90
Factor 23	0.91
Factor 24	0.91
Factor 25	0.92
Factor 26	0.92

Factor 27	0.93
Factor 28	0.93
Factor 29	0.94
Factor 30	0.94
Factor 31	0.94
Factor 32	0.95
Factor 33	0.95
Factor 34	0.95
Factor 35	0.96
Factor 36	0.96
Factor 37	0.96
Factor 38	0.96
Factor 39	0.97
Factor 40	0.97
Factor 41	0.97
Factor 42	0.97
Factor 43	0.97
Factor 44	0.98
Factor 45	0.98
Factor 46	0.98
Factor 47	0.98
Factor 48	0.98
Factor 49	0.98
Factor 50	0.98
Factor 51	0.99
Factor 52	0.99
Factor 53	0.99
Factor 54	0.99
Factor 55	0.99
Factor 56	0.99
Factor 57	0.99
Factor 58	0.99
Factor 59	0.99
Factor 60	0.99
Factor 61	0.99
Factor 62	0.99
Factor 63	0.99
Factor 64	0.99
Factor 65	1.00
Factor 66	1.00
Factor 67	1.00
Factor 68	1.00
Factor 69	1.00

```
Factor 70                          1.00
Factor 71                          1.00
Factor 72                          1.00
Factor 73                          1.00
Factor 74                          1.00
Factor 75                          1.00
Factor 76                          1.00
Factor 77                          1.00
Factor 78                          1.00
Factor 79                          1.00
Factor 80                          1.00
Factor 81                          1.00
Factor 82                          1.00
Factor 83                          1.00
Factor 84                          1.00
Factor 85                          1.00
Factor 86                          1.00
Factor 87                          1.00
Factor 88                          1.00
Factor 89                          1.00
Factor 90                          1.00
Factor 91                          1.00
Factor 92                          1.00
Factor 93                          1.00
Factor 94                          1.00
Factor 95                          1.00
Factor 96                          1.00
Factor 97                          1.00
Factor 98                          1.00
Factor 99                          1.00
Factor 100                         1.00
Factor 101                         1.00


Velicer's Average Squared Correlations

   root   Avg.Corr.Sq.   Avg.Corr.power4
      0        0.09917           0.03918
      1        0.07690           0.02731
      2        0.06206           0.01955
      3        0.05606           0.01678
      4        0.05890           0.01724
```

5	0.04537	0.01071
6	0.04439	0.00992
7	0.03142	0.00646
8	0.02830	0.00602
9	0.02720	0.00518
10	0.02653	0.00502
11	0.02421	0.00446
12	0.02226	0.00409
13	0.01995	0.00368
14	0.01879	0.00348
15	0.01853	0.00342
16	0.01846	0.00338
17	0.01886	0.00337
18	0.01926	0.00340
19	0.01970	0.00345
20	0.01924	0.00336
21	0.01896	0.00333
22	0.01892	0.00338
23	0.01962	0.00340
24	0.02059	0.00356
25	0.02123	0.00370
26	0.02124	0.00366
27	0.02191	0.00374
28	0.02275	0.00393
29	0.02343	0.00412
30	0.02387	0.00413
31	0.02478	0.00423
32	0.02545	0.00420
33	0.02605	0.00432
34	0.02567	0.00408
35	0.02584	0.00403
36	0.02650	0.00431
37	0.02729	0.00443
38	0.02801	0.00463
39	0.02938	0.00505
40	0.02961	0.00514
41	0.03035	0.00517
42	0.03075	0.00547
43	0.03035	0.00564
44	0.03108	0.00568
45	0.03186	0.00579
46	0.03271	0.00570
47	0.03224	0.00590

48	0.03219	0.00611
49	0.03438	0.00696
50	0.03259	0.00622
51	0.03387	0.00640
52	0.03369	0.00654
53	0.03485	0.00688
54	0.03507	0.00715
55	0.03542	0.00688
56	0.03619	0.00716
57	0.03691	0.00722
58	0.03914	0.00793
59	0.04209	0.00887
60	0.04408	0.00960
61	0.04479	0.00995
62	0.04558	0.01044
63	0.04716	0.01062
64	0.04867	0.01084
65	0.04947	0.01178
66	0.04854	0.01136
67	0.04716	0.01080
68	0.05060	0.01176
69	0.05257	0.01247
70	0.05670	0.01398
71	0.06034	0.01583
72	0.06328	0.01725
73	0.06632	0.01922
74	0.07075	0.02045
75	0.07181	0.02082
76	0.07999	0.02462
77	0.07771	0.02258
78	0.08191	0.02534
79	0.09023	0.03014
80	0.09488	0.03354
81	0.10064	0.03540
82	0.11857	0.04282
83	0.11865	0.04156
84	0.12167	0.04464
85	0.11114	0.03919
86	0.11839	0.04256
87	0.12837	0.04891
88	0.14036	0.05682
89	0.15886	0.07128
90	0.19909	0.09945

```
 91        0.24482          0.13467
 92        0.23692          0.13104
 93        0.26339          0.15088
 94        0.34013          0.21884
 95        0.36310          0.24250
 96        0.42021          0.29943
 97        0.57566          0.46447
 98             NA               NA
 99             NA               NA
100             NA               NA
```

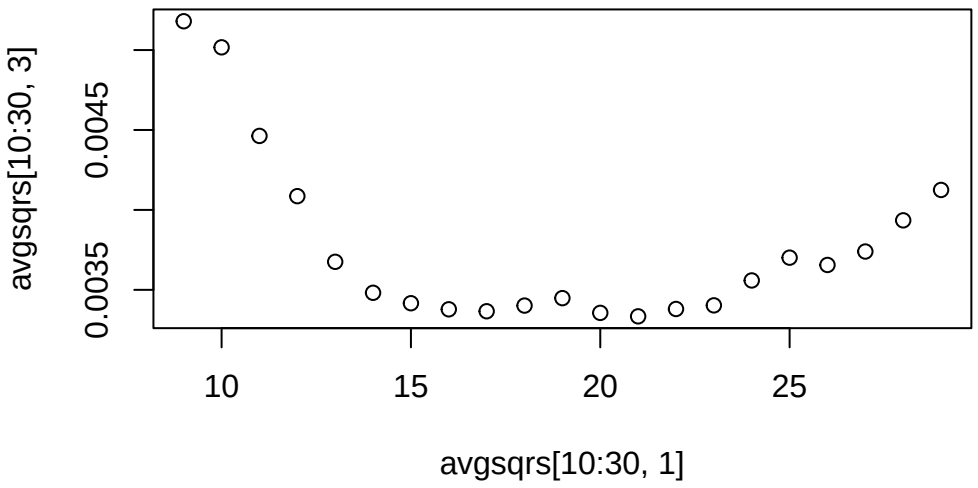
```
The smallest average squared correlation is 0.01846
```

```
The smallest average 4rth power correlation is 0.00333
```

```
The number of components according to the original (1976) MAP Test is = 16
```

```
The number of components according to the revised (2000) MAP Test is = 21
```

```
avgsqrs = map_res$avgsqrs
plot(avgsqrs[10:30, 1], avgsqrs[10:30, 3])
```



Oblimin 21

```
n_axis = 21
# PCA
train_x_scaled = scale(train_x, center = F, scale = T)

eigen_res = eigen(cov(train_x_scaled))

l = eigen_res$values
q = eigen_res$vectors

sum(eigen_res$values[1:n_axis]) / sum(eigen_res$values)
```

```
[1] 0.9372965
```

```
# create correlation loadings from principal axis, then rotate
principal_axis_rot = oblimin(q[, 1:n_axis], maxit = 10000)$loadings

scores = scale(train_x_scaled %*% principal_axis_rot)
colnames(scores) = paste("PC", 1:n_axis, sep = "")
```

```
values = round(principal_axis_rot, 1)
variable_names = colnames(train_x_scaled)

for (i in 1:n_axis) {
  values_i = values[, i]

  ord_gt0 = order(values_i, decreasing = T)
  ord_lt0 = order(values_i, decreasing = F)

  values_i_gt0 = values_i[ord_gt0]
  values_i_lt0 = values_i[ord_lt0]

  variables_gt0 = variable_names[ord_gt0][values_i_gt0 > 0]
  variables_lt0 = variable_names[ord_lt0][values_i_lt0 < 0]

  message("PC", i)
  message("gt 0")
  message(paste(variables_gt0, values_i_gt0[values_i_gt0 > 0], " "))
  message("lt 0")
  message(paste(variables_lt0, values_i_lt0[values_i_lt0 < 0], " "), "\n")
}
```

```
PC1

gt 0

population 0.4  numbUrban 0.4  NumUnderPov 0.4  NumKidsBornNeverMar 0.4  NumImmig 0.4  NumIn

lt 0



PC2

gt 0

pctWPubAsst 0.1  MalePctDivorce 0.1  FemalePctDiv 0.1  TotalPctDiv 0.1  PctHousNoPhone 0.1

lt 0
```

PctRecentImmig -0.4 PctRecImmig5 -0.4 PctRecImmig8 -0.4 PctRecImmig10 -0.4 PctNotSpeakEng

PC3

gt 0

pctUrban 0.1

lt 0

OwnOccLowQuart -0.4 OwnOccMedVal -0.4 OwnOccHiQuart -0.4 OwnOccQrange -0.4 perCapInc -0.

PC4

gt 0

racepctblack 0.8 PctKidsBornNeverMar 0.5 pctUrban 0.1 whitePerCap 0.1 HispPerCap 0.1 Pct

lt 0

racePctHisp -0.2 racePctWhite -0.1 PctKids2Par -0.1 PctTeen2Par -0.1 PctImmigRecent -0.1

PC5

gt 0

racePctHisp 0.1 PctNotSpeakEnglWell 0.1 PctVacMore6Mos 0.1 PctForeignBorn 0.1

lt 0

PctImmigRecent -0.6 PctImmigRec5 -0.5 PctImmigRec8 -0.4 PctImmigRec10 -0.3 pctUrban -0.1

PC6

gt 0

```
lt 0

LandArea -1  HousVacant -0.1

PC7

gt 0

LemasPctOfficDrugUn 1  pctUrban 0.1

lt 0



PC8

gt 0

MalePctDivorce 0.2  HousVacant 0.2  agePct65up 0.1  PctBSorMore 0.1  FemalePctDiv 0.1  Totall

lt 0

PctLargHouseOccup -0.6  PctLargHouseFam -0.5  PctPersDenseHous -0.3  racePctHisp -0.2  housel

PC9

gt 0

indianPerCap 1

lt 0



PC10

gt 0
```

```
agePct16t24 0.1  pctUrban 0.1  PctEmplManu 0.1  NumImmig 0.1  PctPersDenseHous 0.1  NumStreet

lt 0

PctUsePubTrans -0.9  HousVacant -0.1  PctVacMore6Mos -0.1  RentQrange -0.1  MedOwnCostPctIncl

PC11

gt 0

agePct12t21 0.1  pctUrban 0.1  medIncome 0.1  pctWInvInc 0.1  medFamInc 0.1  perCapInc 0.1  l

lt 0

racePctHisp -0.4  PctHousNoPhone -0.4  pctWPubAsst -0.3  PctPopUnderPov -0.2  PctUnemployed -

PC12

gt 0

racePctAsian 0.9  pctWPubAsst 0.1  PctRecImmig8 0.1  PctRecImmig10 0.1  PctForeignBorn 0.1

lt 0

racePctHisp -0.1  PctLess9thGrade -0.1  PctNotHSGrad -0.1  PctKidsBornNeverMar -0.1  PctNotSp

PC13

gt 0

blackPerCap 1  racePctHisp 0.1  medIncome 0.1  perCapInc 0.1  HispPerCap 0.1  PctLess9thGrade

lt 0

OwnOccQrange -0.1

PC14
```

```
gt 0

PctEmplManu 0.6  PctOccupManu 0.4  PctLess9thGrade 0.2  PctNotHSGrad 0.2  PctKidsBornNeverMar

lt 0

pctUrban -0.4  racePctHisp -0.2  PctBSorMore -0.2  PctEmplProfServ -0.2  PctOccupMgmtProf -0

PC15

gt 0

NumImmig 0.1  PctPersDenseHous 0.1  NumInShelters 0.1  NumStreet 0.1

lt 0

PctVacantBoarded -0.9  PctVacMore6Mos -0.2  pctUrban -0.1  PctUnemployed -0.1  PctKidsBornNev

PC16

gt 0

pctWFarmSelf 0.9  perCapInc 0.1  whitePerCap 0.1  HispPerCap 0.1  PctLess9thGrade 0.1  PctBS

lt 0

pctUrban -0.2  agePct16t24 -0.1  pctWPubAsst -0.1  pctWRetire -0.1  PctUnemployed -0.1  PctEm

PC17

gt 0

PctHousLess3BR 0.1  RentQrange 0.1

lt 0

AsianPerCap -0.9  HispPerCap -0.2  medIncome -0.1  medFamInc -0.1  perCapInc -0.1  whitePerCa
```

```
PC18

gt 0

pctWRetire 0.1  PctKidsBornNeverMar 0.1  PctLargHouseFam 0.1  HousVacant 0.1

lt 0

PctWOFullPlumb -1  racePctHisp -0.1  HispPerCap -0.1  PctLess9thGrade -0.1  PctBSorMore -0.1

PC19

gt 0

agePct65up 0.5  pctWSocSec 0.4  pctWRetire 0.3  pctWPubAsst 0.2  PctPopUnderPov 0.2  PctLess9

lt 0

racePctHisp -0.1  agePct12t21 -0.1  agePct12t29 -0.1  agePct16t24 -0.1  medIncome -0.1  pctW

PC20

gt 0

agePct16t24 0.5  agePct12t21 0.4  PctPopUnderPov 0.3  MalePctNevMarr 0.3  agePct12t29 0.2  Pc

lt 0

pctUrban -0.2  MalePctDivorce -0.2  FemalePctDiv -0.2  TotalPctDiv -0.2  agePct65up -0.1  med

PC21

gt 0

PopDens 0.8  pctUrban 0.3  agePct12t29 0.1  agePct16t24 0.1  pctWPubAsst 0.1  PctEmplManu 0.1

lt 0

HousVacant -0.2  PctVacMore6Mos -0.2  racePctHisp -0.1  medIncome -0.1  medFamInc -0.1  white
```

```r
# library(mclust)
#
# mod <- Mclust(scores)
#
# plot(mod, what = "BIC")
# summary(mod, parameters = TRUE)
```

```r
fit = lm(train_y ~ scores)
summary(fit)
```

```
Call:
lm(formula = train_y ~ scores)

Residuals:
     Min       1Q   Median       3Q      Max
-1662.98  -181.73   -37.81   121.36  2316.44

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  575.059      9.143  62.894  < 2e-16 ***
scoresPC1     41.248      9.947   4.147 3.55e-05 ***
scoresPC2     -1.942     22.689  -0.086 0.931788
scoresPC3    -41.090     18.046  -2.277 0.022920 *
scoresPC4    268.697     13.347  20.132  < 2e-16 ***
scoresPC5    -19.904     12.529  -1.589 0.112354
scoresPC6     -9.086      9.627  -0.944 0.345407
scoresPC7     38.668     10.066   3.841 0.000127 ***
scoresPC8     -6.209     16.140  -0.385 0.700528
scoresPC9     -5.011      9.464  -0.529 0.596550
scoresPC10    12.498     12.849   0.973 0.330868
scoresPC11  -263.269     21.229 -12.401  < 2e-16 ***
scoresPC12     6.691     12.661   0.528 0.597238
scoresPC13   -20.734     11.302  -1.835 0.066752 .
scoresPC14   -69.134     13.413  -5.154 2.87e-07 ***
scoresPC15   -70.216     12.436  -5.646 1.94e-08 ***
scoresPC16   -27.142     10.770  -2.520 0.011833 *
scoresPC17   -12.615     11.238  -1.122 0.261823
scoresPC18    28.749     11.904   2.415 0.015845 *
scoresPC19    20.543     12.694   1.618 0.105778
scoresPC20   -80.581     12.060  -6.682 3.27e-11 ***
```

```
scoresPC21    19.768     15.462   1.278 0.201281
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 365.3 on 1574 degrees of freedom
Multiple R-squared:  0.6406,    Adjusted R-squared:  0.6358
F-statistic: 133.6 on 21 and 1574 DF,  p-value: < 2.2e-16
```

Varimax 21

```r
n_axis = 21
# PCA
train_x_scaled = scale(train_x, center = F, scale = T)

eigen_res = eigen(cov(train_x_scaled))

l = eigen_res$values
q = eigen_res$vectors

sum(eigen_res$values[1:n_axis]) / sum(eigen_res$values)
```

```
[1] 0.9372965
```

```r
# create correlation loadings from principal axis, then rotate
principal_axis_rot = varimax(q[, 1:n_axis])$loadings

scores = scale(train_x_scaled %*% principal_axis_rot)
colnames(scores) = paste("PC", 1:n_axis, sep = "")
```

```r
values = round(principal_axis_rot, 1)
variable_names = colnames(train_x_scaled)

for (i in 1:n_axis) {
  values_i = values[, i]

  ord_gt0 = order(values_i, decreasing = T)
  ord_lt0 = order(values_i, decreasing = F)

  values_i_gt0 = values_i[ord_gt0]
  values_i_lt0 = values_i[ord_lt0]
```

```
  variables_gt0 = variable_names[ord_gt0][values_i_gt0 > 0]
  variables_lt0 = variable_names[ord_lt0][values_i_lt0 < 0]

  message("PC", i)
  message("gt 0")
  message(paste(variables_gt0, values_i_gt0[values_i_gt0 > 0], " "))
  message("lt 0")
  message(paste(variables_lt0, values_i_lt0[values_i_lt0 < 0], " "), "\n")
}
```

```
PC1

gt 0

population 0.4  numbUrban 0.4  NumUnderPov 0.4  NumKidsBornNeverMar 0.4  NumImmig 0.4  NumIn

lt 0



PC2

gt 0

pctWPubAsst 0.1  PctHousNoPhone 0.1  NumInShelters 0.1  NumStreet 0.1  PctBornSameState 0.1

lt 0

PctRecentImmig -0.4  PctRecImmig5 -0.4  PctRecImmig8 -0.4  PctRecImmig10 -0.4  PctNotSpeakEng

PC3

gt 0

pctWFarmSelf 0.1  AsianPerCap 0.1  PctPopUnderPov 0.1  PctHousNoPhone 0.1

lt 0
```

OwnOccLowQuart -0.4 OwnOccMedVal -0.4 OwnOccHiQuart -0.4 OwnOccQrange -0.4 perCapInc -0.2

PC4

gt 0

racepctblack 0.8 PctKidsBornNeverMar 0.4 pctUrban 0.1 whitePerCap 0.1 HispPerCap 0.1

lt 0

racePctHisp -0.3 racePctWhite -0.1 PctKids2Par -0.1 PctNotSpeakEnglWell -0.1 PopDens -0.1

PC5

gt 0

racePctHisp 0.1 PctNotSpeakEnglWell 0.1 PctVacMore6Mos 0.1 PctForeignBorn 0.1 PopDens 0.1

lt 0

PctImmigRecent -0.6 PctImmigRec5 -0.5 PctImmigRec8 -0.4 PctImmigRec10 -0.3 PctBSorMore -0

PC6

gt 0

lt 0

LandArea -1 HousVacant -0.1

PC7

gt 0

LemasPctOfficDrugUn 1 pctUrban 0.1

```
lt 0


PC8

gt 0

pctUrban 0.2  pctWRetire 0.1  AsianPerCap 0.1  PctEmplManu 0.1  HousVacant 0.1  RentLowQ 0.1

lt 0

pctWFarmSelf -0.9  PctHousNoPhone -0.2  PopDens -0.2  whitePerCap -0.1  PctPopUnderPov -0.1

PC9

gt 0

indianPerCap 1

lt 0


PC10

gt 0

pctUrban 0.1  pctWFarmSelf 0.1  HispPerCap 0.1  PctEmplManu 0.1  NumImmig 0.1  PctPersDenseH

lt 0

PctUsePubTrans -0.9  PopDens -0.2  PctKidsBornNeverMar -0.1  HousVacant -0.1  PctVacMore6Mos

PC11

gt 0
```

PctLargHouseFam 0.5 PctLargHouseOccup 0.5 PctPersDenseHous 0.4 racePctHisp 0.3 households

lt 0

agePct65up -0.2 racePctWhite -0.1 pctWInvInc -0.1 pctWSocSec -0.1 perCapInc -0.1 blackPe

PC12

gt 0

racePctAsian 0.9 pctWPubAsst 0.1 OwnOccLowQuart 0.1

lt 0

racePctHisp -0.2 PctNotSpeakEnglWell -0.2 pctUrban -0.1 PctLess9thGrade -0.1 PctNotHSGrad

PC13

gt 0

blackPerCap 1 racePctHisp 0.1 HispPerCap 0.1 PctLess9thGrade 0.1 PctUnemployed 0.1 PctNo

lt 0

OwnOccQrange -0.1

PC14

gt 0

NumImmig 0.1 PctPersDenseHous 0.1 NumInShelters 0.1 NumStreet 0.1

lt 0

PctVacantBoarded -0.9 HousVacant -0.2 PctVacMore6Mos -0.2 pctUrban -0.1 PctUnemployed -0

PC15

```
gt 0

pctUrban 0.3  racePctHisp 0.2  PctBSorMore 0.2  PctEmplProfServ 0.2  PctOccupMgmtProf 0.2  H

lt 0

PctEmplManu -0.5  PctOccupManu -0.4  PctLess9thGrade -0.2  PctNotHSGrad -0.2  PctBornSameSta

PC16

gt 0

agePct65up 0.2  pctUrban 0.2  medIncome 0.1  pctWFarmSelf 0.1  pctWSocSec 0.1  pctWRetire 0.

lt 0

agePct16t24 -0.5  agePct12t21 -0.4  agePct12t29 -0.3  PctPopUnderPov -0.3  MalePctNevMarr -0

PC17

gt 0

pctUrban 0.1  RentQrange 0.1

lt 0

AsianPerCap -0.9  pctWPubAsst -0.1  perCapInc -0.1  whitePerCap -0.1  HispPerCap -0.1  PctPo

PC18

gt 0

HousVacant 0.1

lt 0

PctWOFullPlumb -0.9  racePctHisp -0.1  pctWPubAsst -0.1  HispPerCap -0.1  PctPopUnderPov -0.
```

```
PC19

gt 0

PctPersOwnOccup 0.2  PctHousOwnOcc 0.2  PctVacMore6Mos 0.2  householdsize 0.1  agePct12t21 0

lt 0

MalePctDivorce -0.3  FemalePctDiv -0.3  TotalPctDiv -0.3  PctHousNoPhone -0.3  PopDens -0.3

PC20

gt 0

PctEmploy 0.2  PctEmplManu 0.2  agePct12t29 0.1  medIncome 0.1  pctWWage 0.1  pctWFarmSelf 0

lt 0

agePct65up -0.5  pctWSocSec -0.4  pctWPubAsst -0.3  pctWRetire -0.3  PctPopUnderPov -0.2  Pct

PC21

gt 0

PopDens 0.7  pctUrban 0.5  pctWInvInc 0.1  pctWRetire 0.1  HispPerCap 0.1  PctLargHouseFam 0

lt 0

PctHousNoPhone -0.3  HousVacant -0.2  racePctHisp -0.1  PctPopUnderPov -0.1  PctLess9thGrade
```

```r
# library(mclust)
#
# mod <- Mclust(scores)
#
# plot(mod, what = "BIC")
# summary(mod, parameters = TRUE)
```

```r
fit = lm(train_y ~ scores)
summary(fit)
```

```
Call:
lm(formula = train_y ~ scores)

Residuals:
     Min       1Q   Median       3Q      Max
-1662.98  -181.73   -37.81   121.36  2316.44

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  575.059      9.143  62.894  < 2e-16 ***
scoresPC1     27.152     10.292   2.638  0.00842 **
scoresPC2    -33.986     21.052  -1.614  0.10665
scoresPC3    -32.738     17.460  -1.875  0.06098 .
scoresPC4    224.091     12.772  17.546  < 2e-16 ***
scoresPC5    -16.984     11.932  -1.423  0.15483
scoresPC6     -4.949      9.641  -0.513  0.60776
scoresPC7     40.792     10.001   4.079 4.76e-05 ***
scoresPC8     19.964     10.572   1.888  0.05916 .
scoresPC9     -5.597      9.382  -0.597  0.55092
scoresPC10     7.971     12.983   0.614  0.53933
scoresPC11    76.374     16.565   4.611 4.34e-06 ***
scoresPC12   -15.825     11.495  -1.377  0.16882
scoresPC13    -9.904     10.412  -0.951  0.34163
scoresPC14   -70.577     12.114  -5.826 6.86e-09 ***
scoresPC15    54.311     13.122   4.139 3.68e-05 ***
scoresPC16    64.135     11.377   5.637 2.04e-08 ***
scoresPC17   -20.170     10.568  -1.909  0.05648 .
scoresPC18    15.857     13.871   1.143  0.25316
scoresPC19  -264.446     17.213 -15.363  < 2e-16 ***
scoresPC20   -36.020     12.320  -2.924  0.00351 **
scoresPC21   -13.015     14.536  -0.895  0.37073
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 365.3 on 1574 degrees of freedom
Multiple R-squared:  0.6406,    Adjusted R-squared:  0.6358
F-statistic: 133.6 on 21 and 1574 DF,  p-value: < 2.2e-16
```

Violent crime, defined by FBI Why choose 21 ? Something that is related, something that are not related?