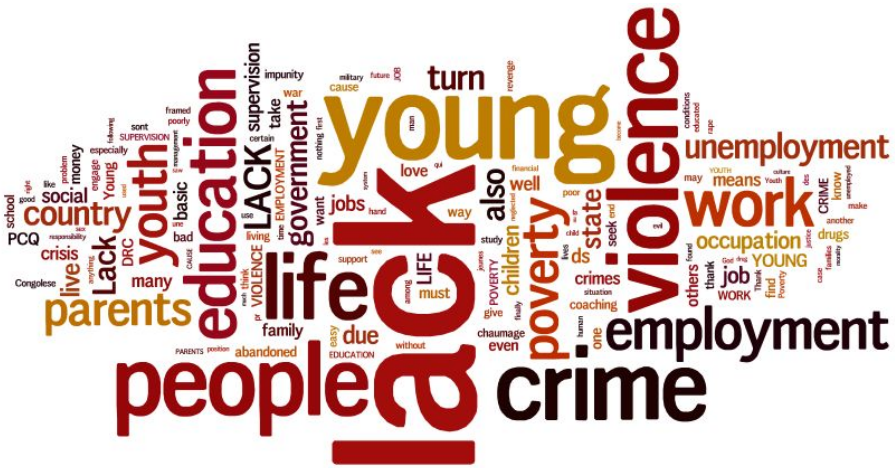
# Do Unemployment Rate and Education Level Play Big Parts in Economic Crime？

-- ANLY 506 Exploratory Data Analytics

-- Members: Yanan Tan, Jinyue Zhu
Yuwei Gao, Nipun Sodhi, Mingyu Wang




# What we are trying to explore…

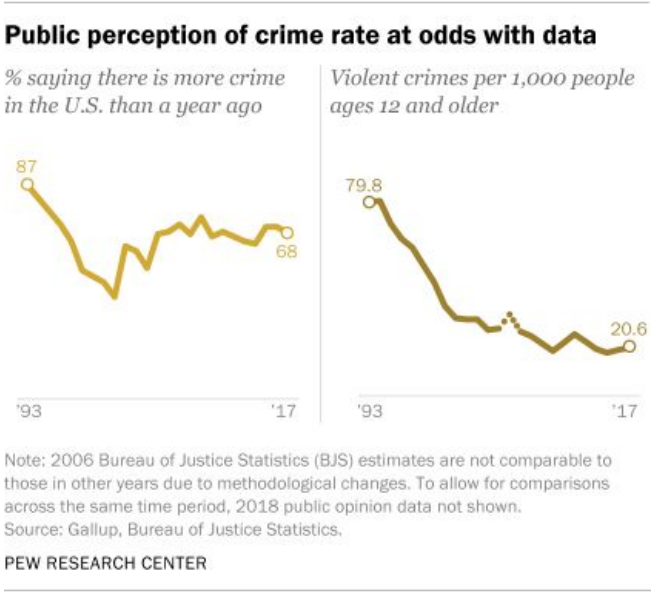
Although the crime rate has decreased sharply since early 1990s (John Gramlich, p.2), various types of crime cases are still happening everyday, putting people's safety into danger.

There are a number of different kinds of crimes have been mentioned in the dataset, but for our project, we will mainly focus on robbery specifically.

We are trying to explore whether certain education level would relate to robbery, or whether certain unemployment level would relate to robbery.

# Background

Interesting phenomenon: people's perception of crime actually increases even though the actual crime rate drops significantly




Therefore, people's perception isn't necessarily the truth. We target to use the data set provided to us to reveal the actual correlation and relationship between unemployment rate, education level and the crime rate.

# Literature Review

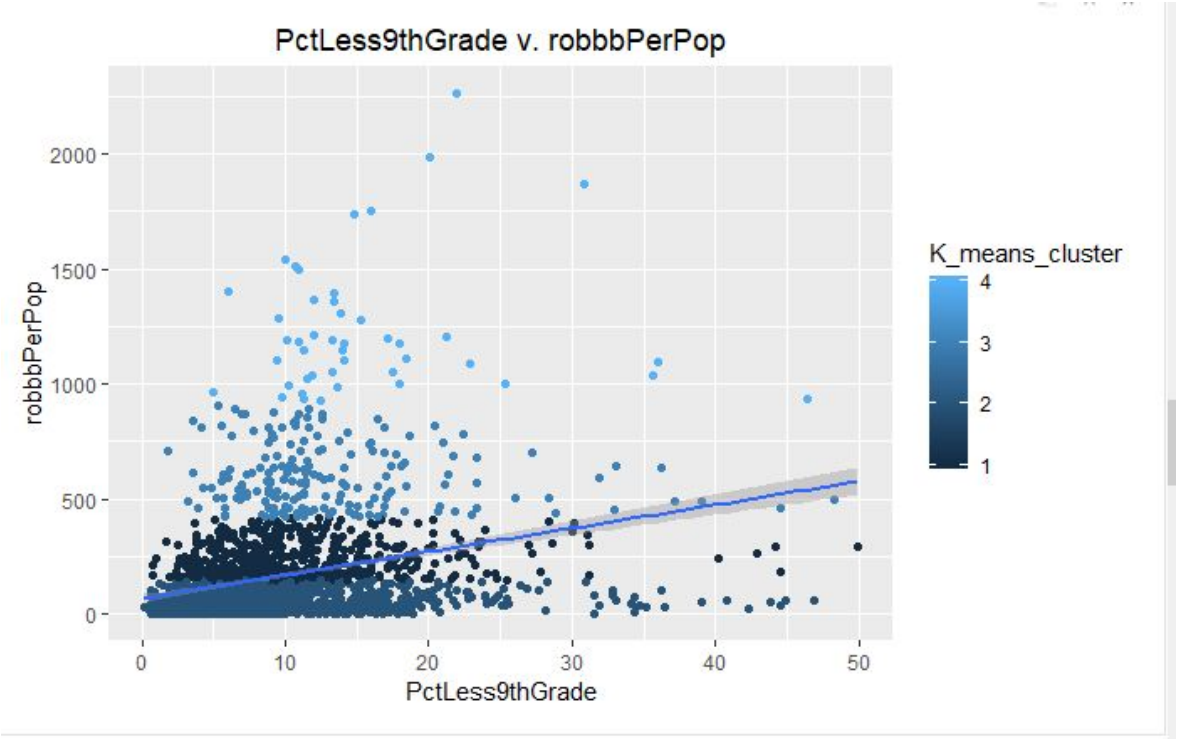
Many people have researched about the connection between unemployment rate, education level and crime.

- Scorcu and Cellini (1998) found unemployment to be a significant explanatory variable for theft in their time-series analysis
- Employed individuals tend to commit fewer crimes than those who are unemployed (Witte and Tauchen, 1994).
- Higher education leads to more opportunity of higher income, which increases the cost of crime (Lochner and Moretti, 2004)
- But high education may decrease the probability of getting caught, which increase the crime rate (Lochner and Moretti, 2004)

# Methodology

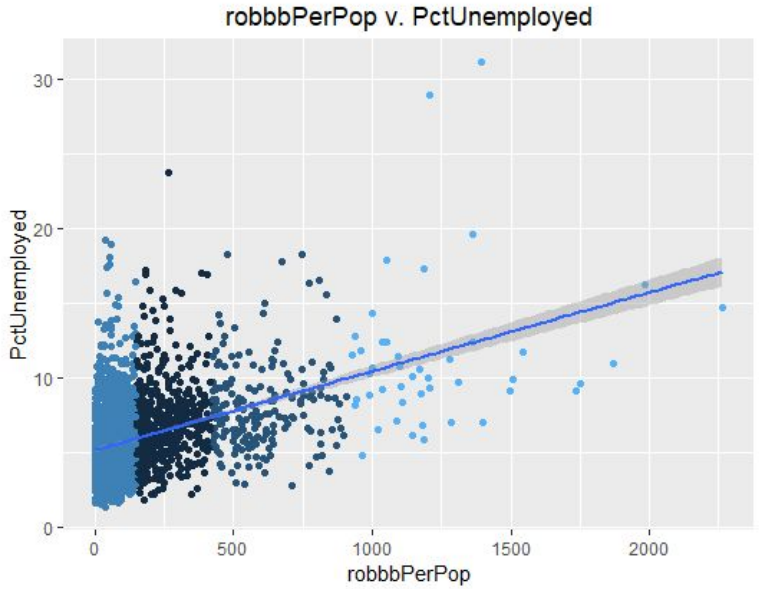
- We have identified the columns of data that are related to our topic, and cleaned the data.
- We have applied principle component analysis on the raw data to test the correlation among different variables.
- We have also used different plots to visualize the data, which could provide good insights on whether there are certain relations among different variables.
- We have applied clustering on the data, to further explore the relationships among the variables.
- If we found possible correlations among the variables, we will try to apply correlation analysis to further explore and quantify the relationships.

# Robberies VS. Education




Robberies VS. Education Less than 9th Grade

# Explore the data (plots)



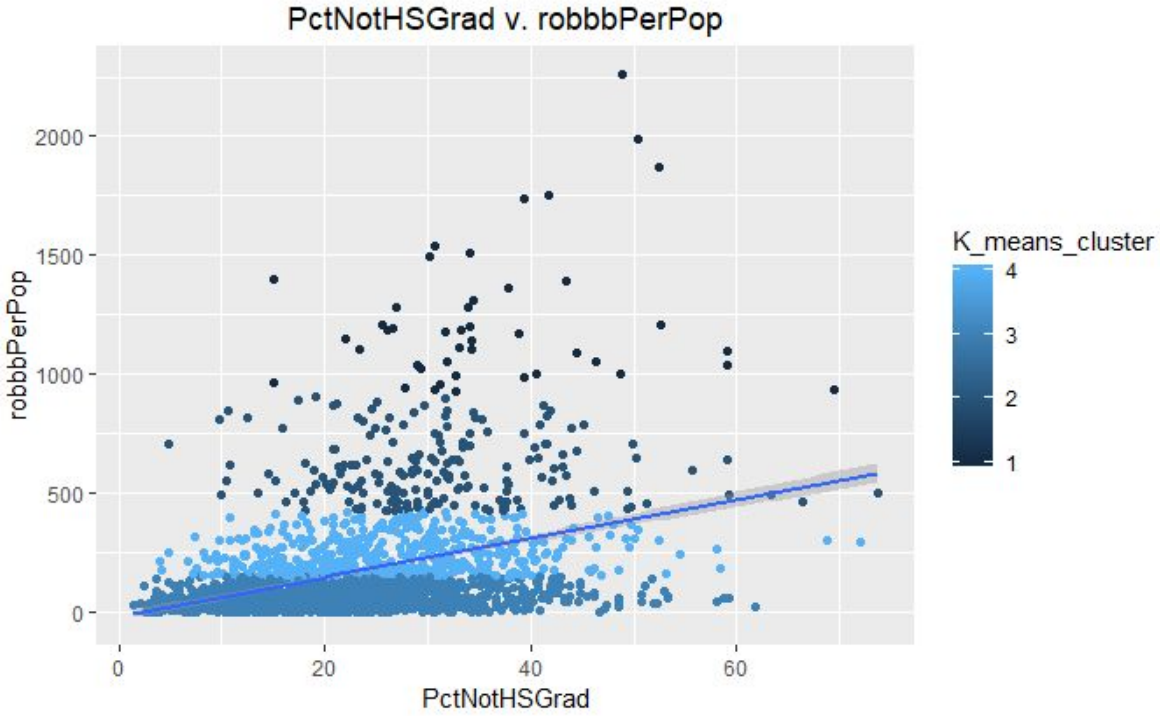

**Robberies VS. Unemployed**

```
K-means clustering with 3 clusters of sizes 83, 1754

Cluster means:
  robbbPerPop PctUnemployed
1  1043.13241      10.278795
2    69.51379       5.520792
3   401.90215       7.547533
```

# Kmeans- Results

```
K-means clustering with 3 clusters of sizes 83, 377, 1754

Cluster means:
  PctLess9thGrade robbbPerPop
1       13.788554  1043.13241
2       12.699204   401.90215
3        8.211038    69.51379


Within cluster sum of squares by cluster:
[1] 7739689 6536493 6353074
 (between_SS / total_SS =  83.1 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
PctLess9thGrade     robbbPerPop
          FALSE           FALSE
```

# Continued..

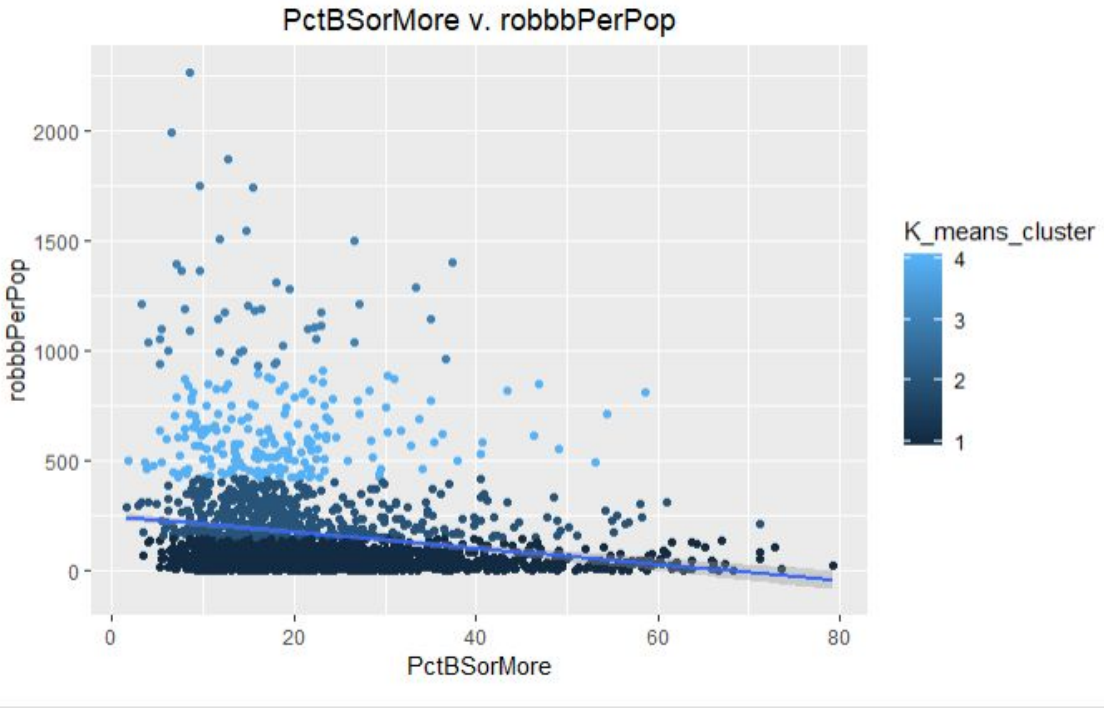



Robberies VS. Education which not graduate from high school

# Kmeans- Results

```
K-means clustering with 3 clusters of sizes 1754, 83, 377

Cluster means:
  PctNotHSGrad robbbPerPop
1     20.25166    69.51379
2     32.98506  1043.13241
3     29.48846   401.90215


Within cluster sum of squares by cluster:
[1] 6465179 7745387 6559407
 (between_SS / total_SS =  83.0 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
PctNotHSGrad  robbbPerPop
       FALSE        FALSE
```

# Continued..




Robberies VS. Education

People who have Bacheler Degree or Higher

# Kmeans- Results

```
K-means clustering with 3 clusters of sizes 378, 1753, 83

Cluster means:
  robbbPerPop PctBSorMore
1   401.46214    18.39439
2    69.41906    24.29362
3  1043.13241    18.14578

Within cluster sum of squares by cluster:
[1] 6579331 6558763 7744334
 (between_SS / total_SS =  82.9 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
robbbPerPop PctBSorMore
      FALSE       FALSE
```

# Principle Component Analysis-2

Importance of components:

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 | PC10 | PC11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Standard deviation | 1.8352 | 1.36008 | 1.29282 | 1.1973 | 1.17305 | 1.1225 | 1.11932 | 1.05600 | 1.02570 | 1.0082 | 0.98611 |
| Proportion of Variance | 0.1203 | 0.06607 | 0.05969 | 0.0512 | 0.04914 | 0.0450 | 0.04475 | 0.03983 | 0.03757 | 0.0363 | 0.03473 |
| Cumulative Proportion | 0.1203 | 0.18635 | 0.24604 | 0.2972 | 0.34638 | 0.3914 | 0.43613 | 0.47595 | 0.51353 | 0.5498 | 0.58456 |

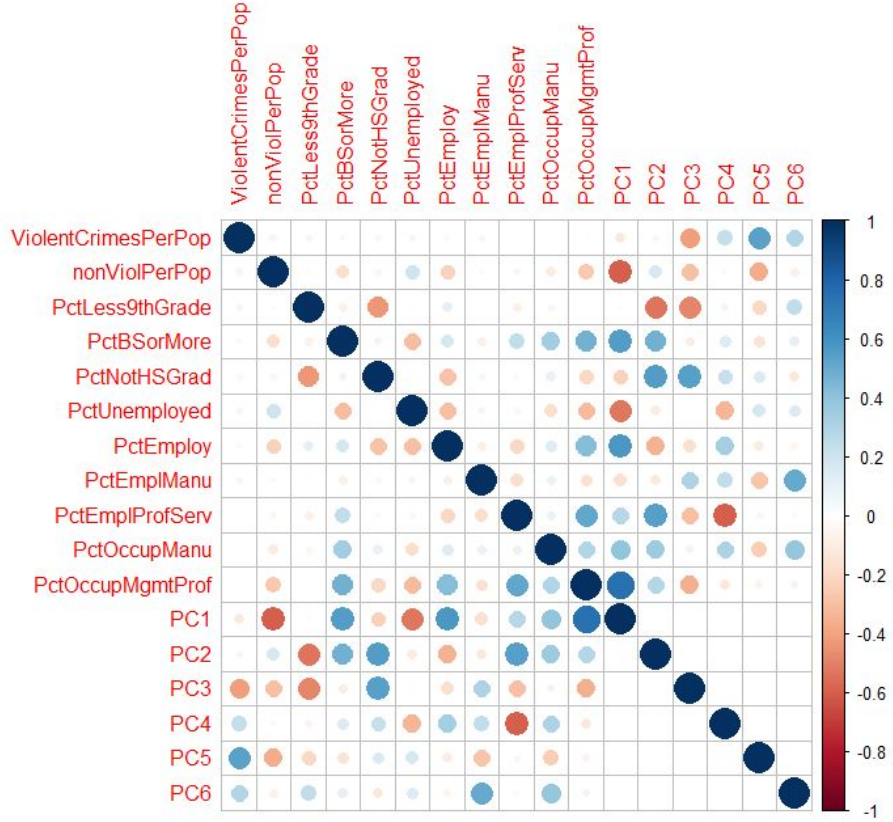
| | PC12 | PC13 | PC14 | PC15 | PC16 | PC17 | PC18 | PC19 | PC20 | PC21 | PC22 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Standard deviation | 0.98295 | 0.96106 | 0.94632 | 0.93790 | 0.91092 | 0.88895 | 0.86547 | 0.85691 | 0.84633 | 0.82686 | 0.82457 |
| Proportion of Variance | 0.03451 | 0.03299 | 0.03198 | 0.03142 | 0.02963 | 0.02822 | 0.02675 | 0.02622 | 0.02558 | 0.02442 | 0.02428 |
| Cumulative Proportion | 0.61907 | 0.65206 | 0.68404 | 0.71546 | 0.74509 | 0.77331 | 0.80006 | 0.82629 | 0.85187 | 0.87629 | 0.90057 |

| | PC23 | PC24 | PC25 | PC26 | PC27 | PC28 |
|---|---|---|---|---|---|---|
| Standard deviation | 0.80316 | 0.74148 | 0.70935 | 0.66900 | 0.66008 | 0.45027 |
| Proportion of Variance | 0.02304 | 0.01964 | 0.01797 | 0.01598 | 0.01556 | 0.00724 |
| Cumulative Proportion | 0.92361 | 0.94324 | 0.96121 | 0.97720 | 0.99276 | 1.00000 |

# Scaling for first 18 components

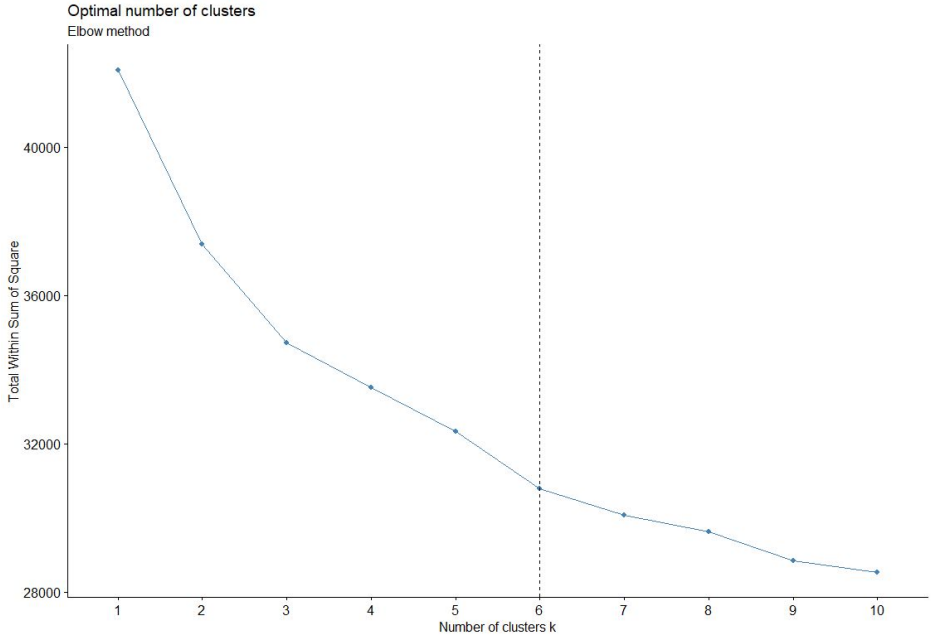
```
> head(pca_results$x[,1:18])
        PC1        PC2         PC3         PC4        PC5        PC6        PC7        PC8        PC9
1 -2.8468979  5.15658469 -2.41611123 -2.34730627 -0.8093929  0.9038685 -2.0508179  2.2884465  2.02194452
2  3.4817868  0.85202056  1.76578966 -0.02425761  0.3727470 -0.1288469 -0.1769745  0.8168969  0.57149665
3  3.0412580  0.62351071 -0.27844950  0.35498014  0.1050545  0.8013031  1.4764980 -0.3924061  0.10085333
4 -0.5459887  2.45245219 -0.04137129  0.23816811 -1.7492462 -0.7626776 -0.4829112 -1.0008768 -0.02143114
5 -0.4651787 -1.42458626  2.96951851 -0.03118387  1.4255871  0.8159805  0.2510937  1.7204827 -0.46923563
6 -0.8091559 -0.01514416  0.86644360  0.93276410 -0.1357640 -0.2617227  4.0134823 -0.7988727  1.28613839
       PC10       PC11       PC12        PC13        PC14       PC15       PC16       PC17       PC18
1 -0.1291788  1.9615303 -1.2594760 -0.20292702 -2.137825076 -0.6457008 0.59752551 -2.6070777 -2.7075733
2  1.0584135 -1.3000204 -0.5578138  0.80628001  1.044984201 -0.3235629 0.05771248  0.5757301  2.7274771
3 -1.1924938 -0.2476364  0.5306965 -0.76727662  1.438910415  0.2187343 1.40849561 -1.0113161  1.0921591
4 -0.7133313  0.4363508 -1.3065290  0.01363394 -0.007180377  1.4803191 0.33568399  1.0559354 -0.6883165
5  0.2006699 -0.7498018  1.2480343 -0.46968717 -0.829480774 -0.4643876 -1.12124573 -0.8997788  0.6004495
6  0.3606721  1.8755494  2.0962035 -1.20737053 -1.833158503 -0.5836582 -0.78139182  0.1335719 -1.7158220
```

# Correlation Plot- PCA




As can be seen for the first 6 components that explain 40% of the variance, the important factors are PctBSorMore, PctEmploy, PctOccupMgmtProf, NonViolPerPop etc.

# Kmeans- Finding the optimal number of clusters..



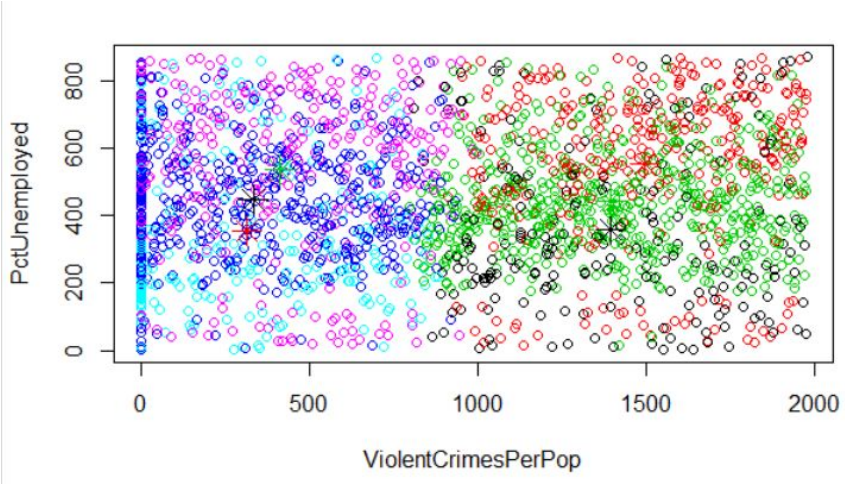

Using the elbow method we can see the 6 is the optimal number of clusters for the given data set containing the following factors

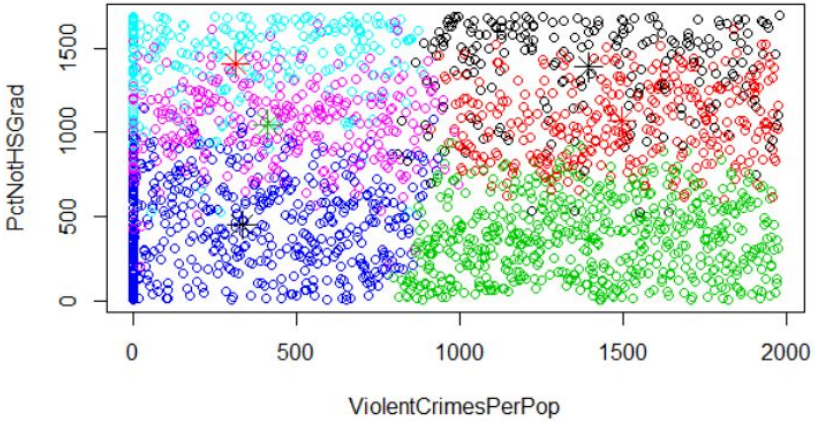
1)ViolentCrimesPerPop
2)PctLess9thGrade
3)PctBSorMore
4)PctNotHSGrad
5) PctUnemployed

# Kmeans- Results

```
Cluster means:
  ViolentCrimesPerPop PctLess9thGrade PctBSorMore PctNotHSGrad PctUnemployed
1           1393.0952         525.5810   1435.6095    1386.0381      357.5810
2           1494.6738         446.2226    339.9939    1066.6220      554.2561
3           1393.0648         935.6620    791.9492     364.2277      442.2119
4            334.1004         978.8803    688.0734     447.0656      446.5560
5            314.5564         434.0350   1397.9183    1405.7276      353.2179
6            413.9217         301.9910    393.9880    1046.4548      534.1777


        Within cluster sum of squares by cluster:
        [1]  80292068 114113858 186528450 185813150  86815562  98394082
         (between_SS / total_SS =  66.7 %)
```

# Kmeans- Results



# Results

- The results show that unemployment rate is positively correlated with the number of robbery.
- Percentage of people whose education level is less than 9th grade is positively correlated with the number of robbery.
- Percentage of people whose education level is lower than 9th grade is positively correlated with the number of robbery.
- Percentage of people whose education level is lower than high school is positively correlated with the number of robbery.
- Percentage of people whose education level is more than sophomore is negatively correlated with the number of robbery.

# Limitation

- Data Cleaning:
  - We experienced a large number of missing or abnormal values in the dataset, which would have negative effects on our analysis.
  - Unable to obtain the average value.
- Solution:
  - We decided to omit the entire row/column.
  - na.omit
  - apply(unemploy, 2, function(x) any(is.na(x)))

# Conclusion

- We successfully applied pre-processing and data exploration techniques that we obtained from class.
- Based on the analysis results, we could conclude that there's relationship between percentage of residents' education levels/unemployment rate and the number of robberies.
- The higher percentage of people in the communities are unemployed, the more robberies
- The higher percentage of people in the communities have received certain level of educations (9th grade, high school and sophomore), the less robberies
- Work in Progress: Status Report, Consolidated Final Report.

# Reference

National Council for Crime Prevention (2001), Brottsutvecklingen i Sverige: 1998–2000 (Trend Analysis of Crime in Sweden, 1998–2000), Stockholm.

Witte, A. D. and Tauchen, H. (1994), Work and Crime: An Exploration Using Paneldata, National Bureau of Economic Research, Cambridge, MA.

Lochner, L. and Moretti, E. (2004), The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports, The American Economic Review, 94(1): 155-189.