2015 SKKU ISS Data Science Project

Team : Mean Girls

|  |  |
| --- | --- |
|  | Woongkyu Ham, |
|  | Hyunji Kim, |
|  | Aaron Wan |
|  | Siobhan Zhang |
|  | Yeeun Yang |

Preliminary Analysis

Team: Mean\_Girls

Names: Woongkyu Ham, Hyunji Kim, Aaron Wan

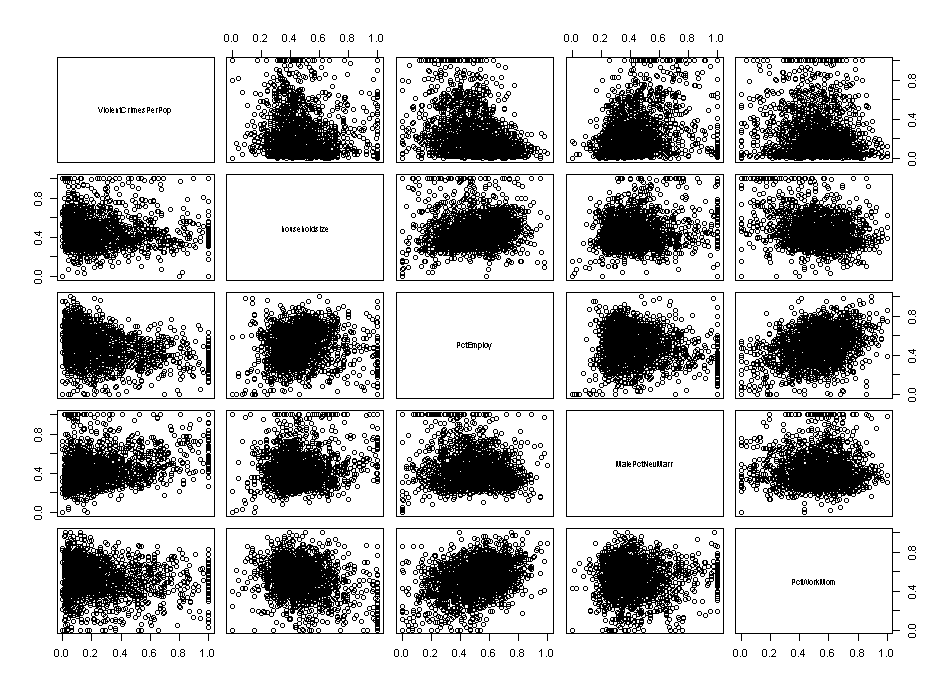
Siobhan Zhang, Yeeun Yang

1. Summary

The sources are the US Census in 1990, LEMAS(Law Enforcement Management and Administrative Statistic) in 1990 and the FBI UCR(Uniform Crime Reporting)in 1995. There are 1995 observed communities defined by 5 indexes. The response variable is violent crimes per 100k population, and there are 122 explanatory variables. 101 explanatory variables and one responsible variable. However, there are some variables that has missing values more then 1000, so we removed those 21 variables. Also, variable named OtherPerCap has one missing value. We replace it with mean of this value.

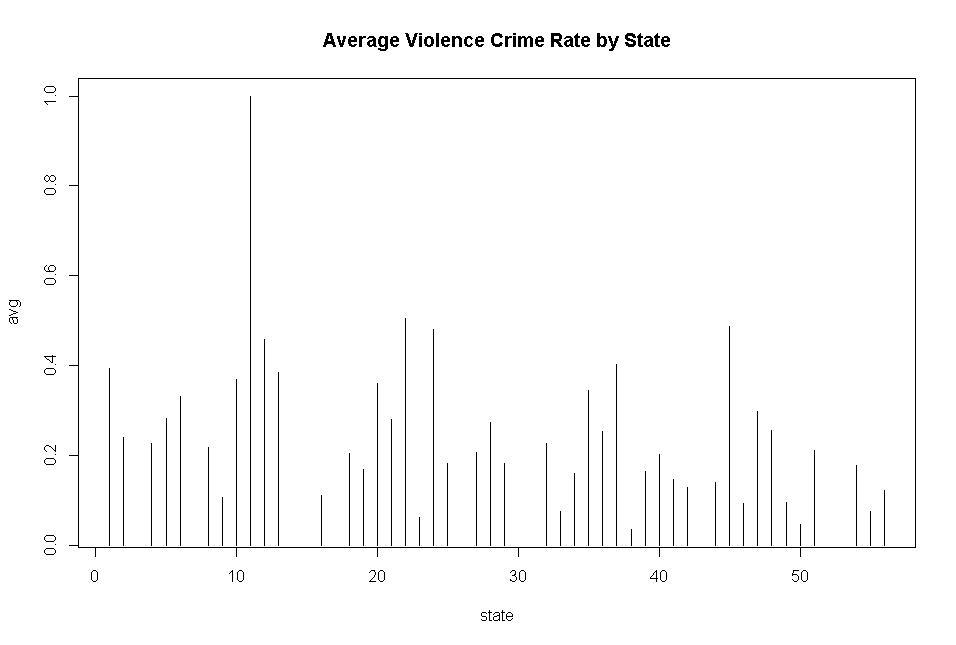
1. Graph of scatter plot with some of variables

We can find some non-linearity between explanatory variables and response variable. And we can also find strong linearity within some explanatory variables. Therefore, we have to consider multicolliniarity. For example, in the red box above, there is a multicolliniarity between “MalePctNevMar” and “PctWorkMom”.



1. Basic statistics and histogram for data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **variable** | **Max** | **median** | **min** | **mean** | **SD** |
| population | 1 | 0.02 | 0 | 0.057593 | 0.126906 |
| householdsize | 1 | 0.44 | 0 | 0.463395 | 0.163717 |
| racepctblack | 1 | 0.06 | 0 | 0.179629 | 0.253442 |
| racePctWhite | 1 | 0.85 | 0 | 0.753716 | 0.244039 |
| racePctAsian | 1 | 0.07 | 0 | 0.153681 | 0.208877 |
| racePctHisp | 1 | 0.04 | 0 | 0.144022 | 0.232492 |
| agePct12t21 | 1 | 0.4 | 0 | 0.424218 | 0.155196 |
| agePct12t29 | 1 | 0.48 | 0 | 0.493867 | 0.143564 |
| agePct16t24 | 1 | 0.29 | 0 | 0.336264 | 0.166505 |
| agePct65up | 1 | 0.42 | 0 | 0.423164 | 0.179186 |
| numbUrban | 1 | 0.03 | 0 | 0.064072 | 0.128256 |
| pctUrban | 1 | 1 | 0 | 0.696269 | 0.444811 |
| medIncome | 1 | 0.32 | 0 | 0.361123 | 0.209363 |
| pctWWage | 1 | 0.56 | 0 | 0.558154 | 0.182913 |
| pctWFarmSelf | 1 | 0.23 | 0 | 0.29157 | 0.204108 |
| pctWInvInc | 1 | 0.48 | 0 | 0.495687 | 0.178071 |
| pctWSocSec | 1 | 0.475 | 0 | 0.471133 | 0.173619 |
| pctWPubAsst | 1 | 0.26 | 0 | 0.317778 | 0.222137 |
| pctWRetire | 1 | 0.47 | 0 | 0.479248 | 0.167564 |
| medFamInc | 1 | 0.33 | 0 | 0.375677 | 0.198257 |
| perCapInc | 1 | 0.3 | 0 | 0.350251 | 0.191109 |
| whitePerCap | 1 | 0.32 | 0 | 0.368049 | 0.186804 |
| blackPerCap | 1 | 0.25 | 0 | 0.291098 | 0.171593 |
| indianPerCap | 1 | 0.17 | 0 | 0.203506 | 0.164775 |
| AsianPerCap | 1 | 0.28 | 0 | 0.322357 | 0.195411 |
| OtherPerCap | 1 | 0.25 | 0 | 0.284739 | 0.19096 |
| HispPerCap | 1 | 0.345 | 0 | 0.386279 | 0.183081 |
| NumUnderPov | 1 | 0.02 | 0 | 0.055507 | 0.127941 |
| PctPopUnderPov | 1 | 0.25 | 0 | 0.303024 | 0.228474 |
| PctLess9thGrade | 1 | 0.27 | 0 | 0.315807 | 0.21336 |
| PctNotHSGrad | 1 | 0.36 | 0 | 0.38333 | 0.202508 |
| PctBSorMore | 1 | 0.31 | 0 | 0.361675 | 0.209193 |
| PctUnemployed | 1 | 0.32 | 0 | 0.363531 | 0.202171 |
| PctEmploy | 1 | 0.51 | 0 | 0.501073 | 0.174036 |
| PctEmplManu | 1 | 0.37 | 0 | 0.396384 | 0.202386 |
| PctEmplProfServ | 1 | 0.41 | 0 | 0.440597 | 0.175457 |
| PctOccupManu | 1 | 0.37 | 0 | 0.391224 | 0.198922 |
| PctOccupMgmtProf | 1 | 0.4 | 0 | 0.441339 | 0.186292 |
| MalePctDivorce | 1 | 0.47 | 0 | 0.461244 | 0.18246 |
| MalePctNevMarr | 1 | 0.4 | 0 | 0.434453 | 0.175437 |
| FemalePctDiv | 1 | 0.5 | 0 | 0.487568 | 0.17517 |
| TotalPctDiv | 1 | 0.5 | 0 | 0.494273 | 0.183607 |
| PersPerFam | 1 | 0.47 | 0 | 0.487748 | 0.154594 |
| PctFam2Par | 1 | 0.63 | 0 | 0.610918 | 0.201976 |
| PctKids2Par | 1 | 0.64 | 0 | 0.620657 | 0.206353 |
| PctYoungKids2Par | 1 | 0.7 | 0 | 0.664032 | 0.218749 |
| PctTeen2Par | 1 | 0.61 | 0 | 0.582884 | 0.191507 |
| PctWorkMomYoungKids | 1 | 0.51 | 0 | 0.501449 | 0.168612 |
| PctWorkMom | 1 | 0.54 | 0 | 0.52669 | 0.175241 |
| NumIlleg | 1 | 0.01 | 0 | 0.036294 | 0.108671 |
| PctIlleg | 1 | 0.17 | 0 | 0.249995 | 0.229946 |
| NumImmig | 1 | 0.01 | 0 | 0.03006 | 0.087189 |
| PctImmigRecent | 1 | 0.29 | 0 | 0.320211 | 0.219088 |
| PctImmigRec5 | 1 | 0.34 | 0 | 0.360622 | 0.210924 |
| PctImmigRec8 | 1 | 0.39 | 0 | 0.399077 | 0.201498 |
| PctImmigRec10 | 1 | 0.43 | 0 | 0.427879 | 0.19497 |
| PctRecentImmig | 1 | 0.09 | 0 | 0.181364 | 0.235792 |
| PctRecImmig5 | 1 | 0.08 | 0 | 0.182126 | 0.236333 |
| PctRecImmig8 | 1 | 0.09 | 0 | 0.184774 | 0.236739 |
| PctRecImmig10 | 1 | 0.09 | 0 | 0.182879 | 0.234822 |
| PctSpeakEnglOnly | 1 | 0.87 | 0 | 0.785903 | 0.226869 |
| PctNotSpeakEnglWell | 1 | 0.06 | 0 | 0.150587 | 0.219716 |
| PctLargHouseFam | 1 | 0.2 | 0 | 0.267608 | 0.196567 |
| PctLargHouseOccup | 1 | 0.19 | 0 | 0.251891 | 0.190709 |
| PersPerOccupHous | 1 | 0.44 | 0 | 0.462101 | 0.169551 |
| PersPerOwnOccHous | 1 | 0.48 | 0 | 0.494428 | 0.157924 |
| PersPerRentOccHous | 1 | 0.36 | 0 | 0.404097 | 0.189301 |
| PctPersOwnOccup | 1 | 0.56 | 0 | 0.562598 | 0.197087 |
| PctPersDenseHous | 1 | 0.11 | 0 | 0.186264 | 0.209956 |
| PctHousLess3BR | 1 | 0.51 | 0 | 0.495186 | 0.172508 |
| MedNumBR | 1 | 0.5 | 0 | 0.314694 | 0.255182 |
| HousVacant | 1 | 0.03 | 0 | 0.076815 | 0.150465 |
| PctHousOccup | 1 | 0.77 | 0 | 0.719549 | 0.194024 |
| PctHousOwnOcc | 1 | 0.54 | 0 | 0.548686 | 0.185204 |
| PctVacantBoarded | 1 | 0.13 | 0 | 0.204529 | 0.21777 |
| PctVacMore6Mos | 1 | 0.42 | 0 | 0.433335 | 0.188986 |
| MedYrHousBuilt | 1 | 0.52 | 0 | 0.494178 | 0.232467 |
| PctHousNoPhone | 1 | 0.185 | 0 | 0.264478 | 0.242847 |
| PctWOFullPlumb | 1 | 0.19 | 0 | 0.243059 | 0.206295 |
| OwnOccLowQuart | 1 | 0.18 | 0 | 0.264689 | 0.224425 |
| OwnOccMedVal | 1 | 0.17 | 0 | 0.26349 | 0.231542 |
| OwnOccHiQuart | 1 | 0.18 | 0 | 0.268942 | 0.235252 |
| RentLowQ | 1 | 0.31 | 0 | 0.346379 | 0.219323 |
| RentMedian | 1 | 0.33 | 0 | 0.372457 | 0.209278 |
| RentHighQ | 1 | 0.37 | 0 | 0.422964 | 0.248286 |
| MedRent | 1 | 0.34 | 0 | 0.384102 | 0.213404 |
| MedRentPctHousInc | 1 | 0.48 | 0 | 0.490125 | 0.1695 |
| MedOwnCostPctInc | 1 | 0.45 | 0 | 0.449754 | 0.187274 |
| MedOwnCostPctIncNoMtg | 1 | 0.37 | 0 | 0.403816 | 0.192593 |
| NumInShelters | 1 | 0 | 0 | 0.029438 | 0.102607 |
| NumStreet | 1 | 0 | 0 | 0.022778 | 0.1004 |
| PctForeignBorn | 1 | 0.13 | 0 | 0.215552 | 0.231134 |
| PctBornSameState | 1 | 0.63 | 0 | 0.608892 | 0.204329 |
| PctSameHouse85 | 1 | 0.54 | 0 | 0.53505 | 0.181352 |
| PctSameCity85 | 1 | 0.67 | 0 | 0.626424 | 0.200521 |
| PctSameState85 | 1 | 0.7 | 0 | 0.65153 | 0.198221 |
| LandArea | 1 | 0.04 | 0 | 0.065231 | 0.109459 |
| PopDens | 1 | 0.17 | 0 | 0.232854 | 0.203092 |
| PctUsePubTrans | 1 | 0.07 | 0 | 0.161685 | 0.229055 |
| LemasPctOfficDrugUn | 1 | 0 | 0 | 0.094052 | 0.240328 |
| ViolentCrimesPerPop | 1 | 0.15 | 0 | 0.237979 | 0.232985 |



1. Figure out VIF for multico-linearity

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **VIF** | **Variable** | **VIF** | **Variable** | **VIF** |
| population | 290.9776 | PctEmploy | 21.58122 | PersPerRentOccHous | 26.80105 |
| householdsize | 22.82612 | PctEmplManu | 4.817687 | PctPersOwnOccup | 570.0064 |
| racepctblack | 19.23216 | PctEmplProfServ | 5.880017 | PctPersDenseHous | 28.83801 |
| racePctWhite | 23.55274 | PctOccupManu | 13.65021 | PctHousLess3BR | 11.73361 |
| racePctAsian | 5.892148 | PctOccupMgmtProf | 29.61476 | MedNumBR | 2.813338 |
| racePctHisp | 17.67551 | MalePctDivorce | 233.7418 | HousVacant | 13.79781 |
| agePct12t21 | 30.88968 | MalePctNevMarr | 16.2524 | PctHousOccup | 4.126627 |
| agePct12t29 | 57.63464 | FemalePctDiv | 336.3859 | PctHousOwnOcc | 550.7007 |
| agePct16t24 | 85.54458 | TotalPctDiv | 1037.397 | PctVacantBoarded | 2.486415 |
| agePct65up | 39.29788 | PersPerFam | 77.6415 | PctVacMore6Mos | 2.5761 |
| numbUrban | 282.1059 | PctFam2Par | 118.9204 | MedYrHousBuilt | 5.18069 |
| pctUrban | 5.534107 | PctKids2Par | 117.3204 | PctHousNoPhone | 8.343939 |
| medIncome | 149.4782 | PctYoungKids2Par | 12.7405 | PctWOFullPlumb | 1.999546 |
| pctWWage | 30.51778 | PctTeen2Par | 7.631893 | OwnOccLowQuart | 237.7907 |
| pctWFarmSelf | 1.933174 | PctWorkMomYoungKids | 7.177748 | OwnOccMedVal | 577.7458 |
| pctWInvInc | 16.56558 | PctWorkMom | 10.18121 | OwnOccHiQuart | 171.9009 |
| pctWSocSec | 39.54922 | NumIlleg | 15.91257 | RentLowQ | 24.74712 |
| pctWPubAsst | 11.9649 | PctIlleg | 13.66819 | RentMedian | 122.807 |
| pctWRetire | 4.356984 | NumImmig | 5.285234 | RentHighQ | 52.34628 |
| medFamInc | 115.6617 | PctImmigRecent | 9.254639 | MedRent | 87.45803 |
| perCapInc | 148.8476 | PctImmigRec5 | 22.50084 | MedRentPctHousInc | 3.482764 |
| whitePerCap | 92.8436 | PctImmigRec8 | 27.66875 | MedOwnCostPctInc | 4.774147 |
| blackPerCap | 2.181873 | PctImmigRec10 | 15.47813 | MedOwnCostPctIncNoMtg | 2.542412 |
| indianPerCap | 1.170327 | PctRecentImmig | 95.06307 | NumInShelters | 4.966828 |
| AsianPerCap | 1.564361 | PctRecImmig5 | 313.3035 | NumStreet | 2.559466 |
| OtherPerCap | 1.462605 | PctRecImmig8 | 478.2197 | PctForeignBorn | 49.42096 |
| HispPerCap | 2.374444 | PctRecImmig10 | 302.2416 | PctBornSameState | 8.263344 |
| NumUnderPov | 35.69575 | PctSpeakEnglOnly | 29.17244 | PctSameHouse85 | 12.47054 |
| PctPopUnderPov | 23.46987 | PctNotSpeakEnglWell | 25.78802 | PctSameCity85 | 6.678506 |
| PctLess9thGrade | 23.94657 | PctLargHouseFam | 226.0116 | PctSameState85 | 8.196368 |
| PctNotHSGrad | 43.16726 | PctLargHouseOccup | 233.1431 | LandArea | 3.300686 |
| PctBSorMore | 30.01291 | PersPerOccupHous | 205.105 | PopDens | 4.350645 |
| PctUnemployed | 7.768672 | PersPerOwnOccHous | 80.19035 | PctUsePubTrans | 3.206867 |
|  |  |  |  | LemasPctOfficDrugUn | 1.578417 |

-There are 65 variables that has VIF bigger then 10. So we can conclude that our data has multicolliniaty.