EDA 1조

Final Project with Crimedata

최우현 전은지 이솔희 엄상준 김윤환

목차

1. 도입
2. 변수 제거
3. Outlier
4. Skewness
5. Clustering
6. 변수 선택
7. 향후 계획 및 역할 배분

도입

	CZ	DA	DB	DC	DD	DE	DF	DG	DH	DI	DJ
1	LemasSwc	LemasSwF	LemasSwF	LemasSwF	LemasTota	LemasTotl	PolicReqP	PolicPerPc	RacialMatc	PctPolicW	PctPolicBla
2	?	?	?	?	?	?	?	?	?	?	?
3	?	?	?	?	?	?	?	?	?	?	?
4	?	?	?	?	?	?	?	?	?	?	?
5	?	?	?	?	?	?	?	?	?	?	?
6	?	?	?	?	?	?	?	?	?	?	?
7	?	?	?	?	?	?	?	?	?	?	?
8	?	?	?	?	?	?	?	?	?	?	?
9	?	?	?	?	?	?	?	?	?	?	?
10	?	?	?	?	?	?	?	?	?	?	?
11	198	183.53	187	173.33	73432	68065.1	370.9	183.5	89.32	78.28	11.11
12	?	?	?	?	?	?	?	?	?	?	?
13	?	?	?	?	?	?	?	?	?	?	?
14	?	?	?	?	?	?	?	?	?	?	?
15	111	189.09	89	151.61	39900	67969.3	359.5	189.1	63.67	81.98	18.02
16	?	?	?	?	?	?	?	?	?	?	?
17	?	?	?	?	?	?	?	?	?	?	?
18	?	?	?	?	?	?	?	?	?	?	?

- row 2215
- column 147
- 44592의 NA
- 특정 변수에 밀집된 NA

변수 제거

변수 제거

```
> sum(cd$LemasPctOfficDrugUn==0)
[1] 1882
```

```
##                          n naratio nacatg
## 1              LemasSwornFT   0.845    Bad
## 2            LemasSwFTPerPop   0.845    Bad
## 3          LemasSwFTFieldOps   0.845    Bad
## 4       LemasSwFTFieldPerPop   0.845    Bad
## 5             LemasTotalReq   0.845    Bad
## 6          LemasTotReqPerPop   0.845    Bad
## 7           PolicReqPerOffic   0.845    Bad
## 8                PolicPerPop   0.845    Bad
## 9          RacialMatchCommPol   0.845    Bad
## 10            PctPolicWhite   0.845    Bad
## 11            PctPolicBlack   0.845    Bad
## 12             PctPolicHisp   0.845    Bad
## 13            PctPolicAsian   0.845    Bad
## 14            PctPolicMinor   0.845    Bad
## 15       OfficAssgnDrugUnits   0.845    Bad
## 16         NumKindsDrugsSeiz   0.845    Bad
## 17          PolicAveOTWorked   0.845    Bad
## 18                PolicCars   0.845    Bad
## 19            PolicOperBudg   0.845    Bad
## 20      LemasPctPolicOnPatr   0.845    Bad
## 21      LemasGangUnitDeploy   0.845    Bad
## 22           PolicBudgPerPop   0.845    Bad
## 23            communityCode   0.553    Bad
## 24               countyCode   0.551    Bad
```

- NA가 0으로 적혀 있는 LemasPctOfficDrugUn
- NA 비율이 80% 이상인 변수 22개
- 분석에 불필요할 것으로 생각되는 변수들 (communityname, State, communityCode, countryCode, fold)

→ 28개 변수 삭제

변수 제거

우리의 목표는 ?

Bayes Normal Model을 이용한 Imputation!

Reponse Variable 없는 row도 삭제!

이제 NA 안녕!

변수 제거

```
> dtx <- subset(dt, select = -c(102:119))
> dty <- subset(dt, select = c(102:119))
> str(dty)
'data.frame':	1901 obs. of  18 variables:
 $ murders            : int  0 0 3 7 0 8 0 29 1 12 ...
 $ murdPerPop         : num  0 0 8.3 4.63 0 ...
 $ rapes              : chr  "0" "1" "6" "77" ...
 $ rapesPerPop        : chr  "0" "4.25" "16.6" "50.98" ...
 $ robberies          : chr  "1" "5" "56" "136" ...
 $ robbbPerPop        : chr  "8.2" "21.26" "154.95" "90.05" ...
 $ assaults           : chr  "4" "24" "14" "449" ...
 $ assaultPerPop      : chr  "32.81" "102.05" "38.74" "297.29" ...
 $ burglaries         : chr  "14" "57" "274" "2094" ...
 $ burglPerPop        : chr  "114.85" "242.37" "758.14" "1386.46" ...
 $ larcenies          : chr  "138" "376" "1797" "7690" ...
 $ larcPerPop         : chr  "1132.08" "1598.78" "4972.19" "5091.64" ...
 $ autoTheft          : chr  "16" "26" "136" "454" ...
 $ autoTheftPerPop    : chr  "131.26" "110.55" "376.3" "300.6" ...
 $ arsons             : chr  "2" "1" "22" "134" ...
 $ arsonsPerPop       : chr  "16.41" "4.25" "60.87" "88.72" ...
 $ ViolentCrimesPerPop: chr  "41.02" "127.56" "218.59" "442.95" ...
 $ nonViolPerPop      : chr  "1394.59" "1955.95" "6167.51" "6867.42" ...
```

요로코롬 X와 Y 분리도 했구!

이제 정규분포 가정만 만족하면!

넣을 수 있어!

Outlier

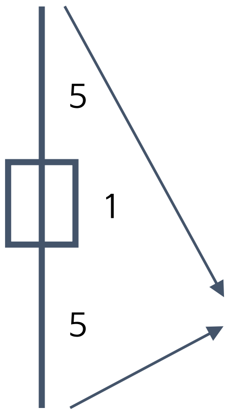
Outlier

```r
dtx_q1 <- c()
dtx_q3 <- c()
for(i in 1:ncol(dtx)) {
  dtx_q1[i] <-  quantile(dtx[,i])[1]
  dtx_q3[i] <-  quantile(dtx[,i])[3]
}

dtx_q <- as.data.frame((cbind(dtx_q1, dtx_q3)))

dtx_q <- dtx_q %>%
  mutate(dtx_out1 = dtx_q1 - 5*(dtx_q3-dtx_q1)) %>%
  mutate(dtx_out2 = dtx_q3 + 5*(dtx_q3-dtx_q1))
```

Outlier의 기준?



1사분위수와 3사분위수로부터
이 두 수의 간격보다 5배 이상
떨어진 친구들!

이런 친구들을

요 친구들로 대체!

Outlier

```
> stem(dtx[,92])

  The decimal point is 3 digit(s) to the right of the |

   0 | 000000000000000000000000000000000000000000000000000000000000000000000+1803
   1 | 001223367
   2 | 0248
   3 | 4
   4 | 067
   5 |
   6 |
   7 |
   8 |
   9 |
  10 |
  11 |
  12 |
  13 |
  14 |
  15 |
  16 |
  17 |
  18 |
  19 |
  20 |
  21 |
  22 |
  23 | 4
```

```
> stem(dtx[,93])

  The decimal point is 3 digit(s) to the right of the |

   0 | 000000000000000000000000000000000000000000000000000000000000000000000+1816
   1 | 16
   2 | 1
   3 | 1
   4 |
   5 |
   6 |
   7 |
   8 |
   9 |
  10 | 4
```

여전히 잡히지 않는 Outlier들,,

NA가 0으로 적혀 있는 것으로 판단해

→ NumInShelters
NumStreet 삭제!

Skewness

Skewness

```
[1] "agePct12t21"

  The decimal point is at the |

   4 | 673699
   6 | 014901266799
   8 | 0001111222233446666778888889990001111123333344455555556666666677777778
  10 | 00000000000000111111122222222222223333333333344444444444455555555555+280
  12 | 000000000000000000000000000000111111111111111111111111111111122222222+710
  14 | 0000000000000000000000000000000000000000000000001111111111111111111+453
  16 | 00000000000000000111111111111111111111111122222222222222233333344444444+126
  18 | 000111122222233334455555666666777777777888888899900000000001111122333+8
  20 | 000001112333334445555556678888900122234557 99
  22 | 012456812234567 7789
  24 | 02244568899000256679
  26 | 00011378904
  28 | 0135822223668
  30 | 3077
  32 | 055
  34 | 09568
  36 | 335638
  38 | 029
  40 | 088
  42 | 3884
  44 |
  46 | 899
  48 | 73
  50 | 8
  52 |
  54 | 4
```

왼쪽과 같이 skew된
X 변수들이 있네? !

Normal Model을 사용하려면
변환을 해줘야겠다! ! !

Skewness

```r
colnames(dtx)[71]
```

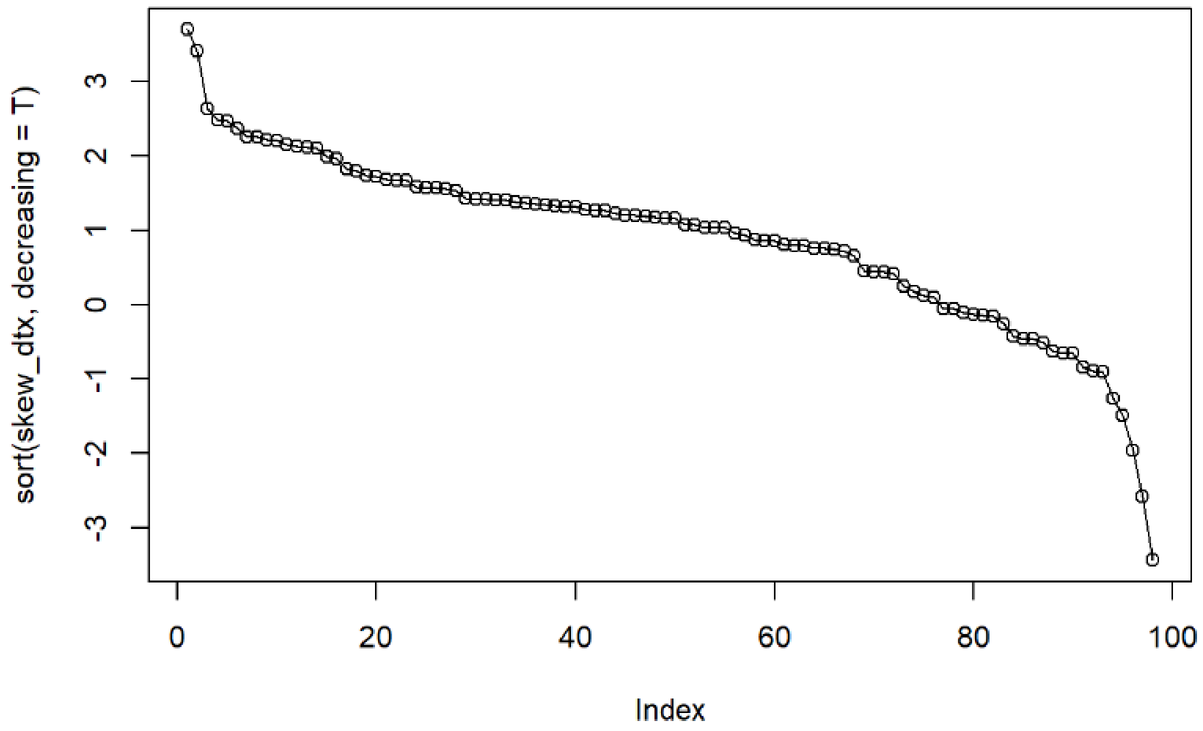
```
## [1] "MedNumBR"
```

```r
conti_dtx <- dtx[,-71]
cate_dtx <- dtx[,71]
```

Skewness함수를 쓰기에 앞서,
범주형 변수인 “MedNumBR”
를 구분

Skewness

```r
skew_dtx=c()
for (i in 1:ncol(conti_dtx)){
  skew_dtx[i]=skewness(conti_dtx[,i])
}
```



- Skewness함숫값이 -4부터 4까지 분포
- 이 값들을 기준으로 변환을 해주자!

Skewness

```
trans_conti_dtx=conti_dtx
for(i in 1:ncol(conti_dtx)) {
  if(skew_dtx[i]>2){
    for(j in 1:dim(conti_dtx)[1]){
      if (conti_dtx[j,i]==0){
        conti_dtx[j,i]=0.03}
    }
    trans_conti_dtx[,i]=log(conti_dtx[,i])
    colnames(trans_conti_dtx)[i]=paste('log',colnames(trans_conti_dtx)[i])
  }
  if(skew_dtx[i]<(-2)){
    for(k in 1:dim(conti_dtx)[1]){
      if (conti_dtx[k,i]==0){
        conti_dtx[k,i]=0.03}
    }
    trans_conti_dtx[,i]=(conti_dtx[,i])^2
    colnames(trans_conti_dtx)[i]=paste('sq',colnames(trans_conti_dtx)[i])}
}
```

Skewness 〉 2

=〉 0을 0.03으로 대체하고

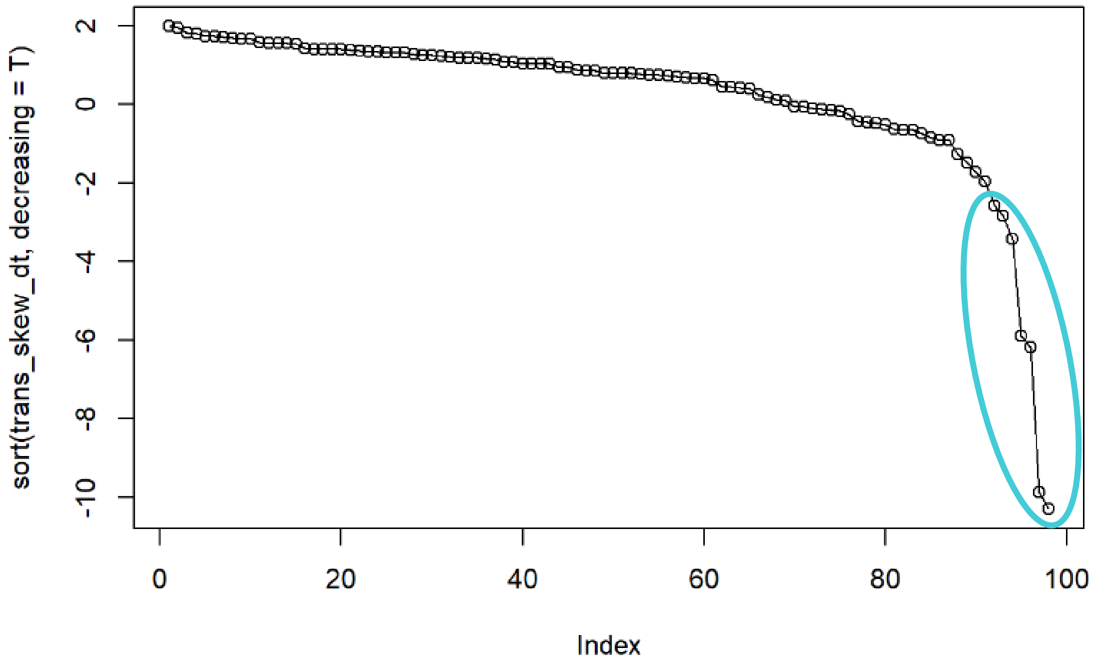
=〉 Log 변환

Skewness 〈 -2

=〉 Square 변환

Skewness

```
trans_skew_dt=c()
for (i in 1:ncol(trans_conti_dtx)){
  trans_skew_dt[i]=skewness(trans_conti_dtx[,i])
}
```

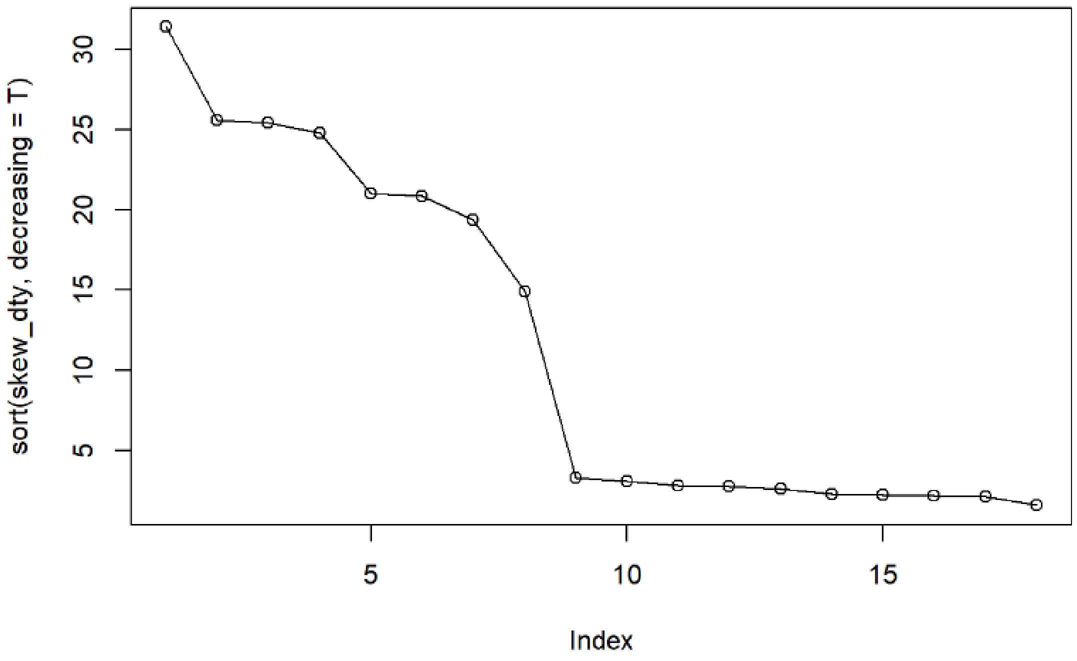



```
##      a
## [1,] "log HispPerCap"      "27" "-10.2934214020652"
## [2,] "log OwnOccQrange"    "82" "-9.87742635877414"
## [3,] "log AsianPerCap"     "25" "-6.17693510840391"
## [4,] "log blackPerCap"     "23" "-5.89841276251884"
## [5,] "log OtherPerCap"     "26" "-3.41634952946269"
## [6,] "log indianPerCap"    "24" "-2.83272299998629"
## [7,] "sq PctHousOccup"     "72" "-2.57681111345612"
```

변환 후에도 여전히 높은 왜도값

삭제!

Skewness

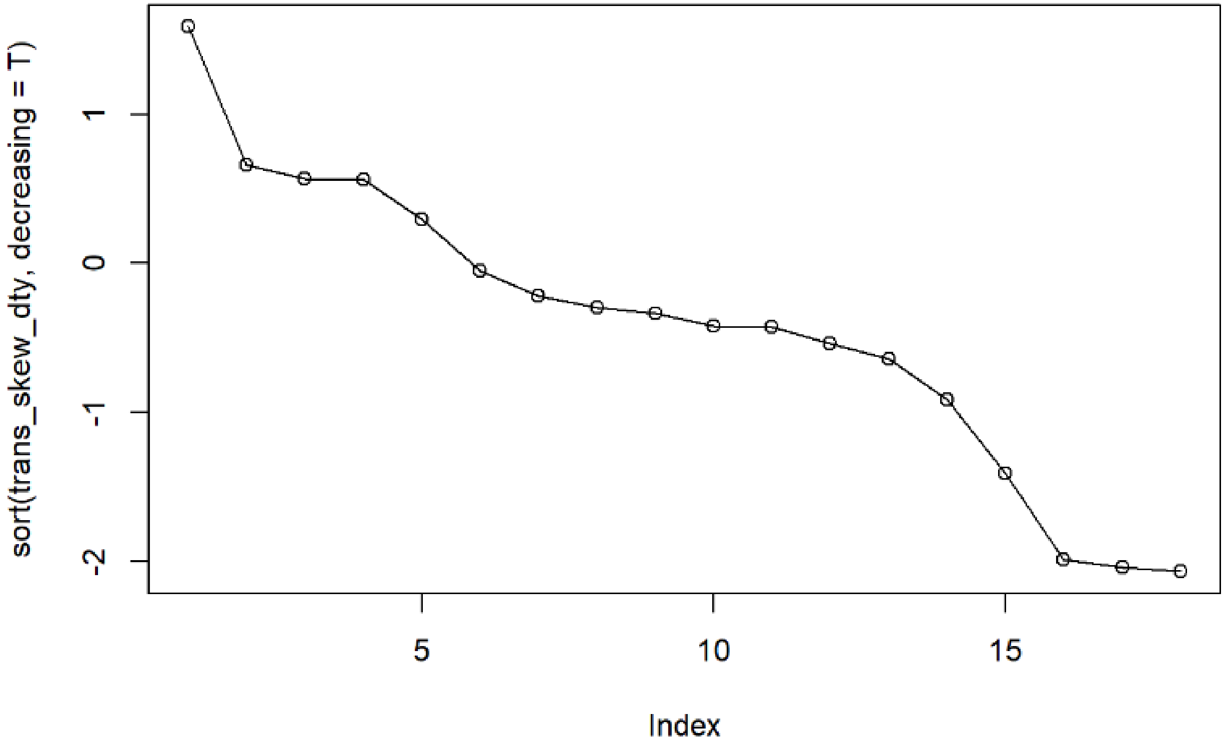


Reponse Variables의 Skewness

: 대체로 높게 나타난다.

=〉 **변환 필요!**

Skewness



같은 방법으로 변환 후 skewness 값

: −2와 2 사이로 안정

Clustering

Clustering

2	others	Population	PctSpeakE	PctNotSpe	LandArea	PopDens	PctUsePubTrans							
3	race	racepctblack	racePctWh	racePctAsi	racePctHisp									
4	House	householdsize	PersPerOc	PctLargHc	PctLargHc	PersPerOv	PersPerRe	PctPersOv	PctPersDe	PctHousLe	MedNumE	HousVaca	PctHousO	PctHousO
5	Age	agePct12t21	agePct12t	agePct16t	agePct65up									
6	urban	numbUrban	pctUrban											
7	Income	medIncome	pctWWag	pctWFarm	pctWInvIn	pctWSocS	pctWPubA	pctWRetir	medFamIr	perCapInc				
8	Race Income	whitePerCap	blackPerCa	indianPerC	AsianPerC	OtherPerC	HispPerCap							
9	Economic	NumUnderPov	PctPopUnderPov											
10	Education	PctLess9thGrad	PctNotHS	PctBSorMore										
11	Employment	PctUnemployed	PctEmploy	PctEmplM	PctEmplPr	PctOccupM	PctOccupMgmtProf							
12	Marital State	MalePctDivorce	MalePctNe	FemalePct	TotalPctDiv									
13	Family Form	PersPerFam	PctFam2Pa	PctKids2Pa	PctYoungK	PctTeen2P	PctWorkM	PctWorkM	NumKidsB	PctKidsBornNeverMar				
14	Immigrant	NumImmig	PctImmigR	PctImmigR	PctImmigR	PctImmigR	PctRecent	PctRecImm	PctRecImm	PctRecImmig10				
15	Ownership	PctHousNoPho	PctWOFullPlumb											
16	Rent	RentLowQ	RentMedia	RentHighQ	RentQrang	MedRent	MedRentP	MedOwnC	MedOwnCostPctIncNoMtg					
17	Poverty	NumInShelters	NumStreet											
18	Population Chang	PctForeignBorn	PctBornSa	PctSameH	PctSameC	PctSameState85								

➡ Description을 바탕으로 한 직관적 Clustering

Clustering

Others

Urban

Education

Immigrant

Race

Income

Employment

Ownership

House

Race Income

Marital State

Rent

Age

Economy

Family Form

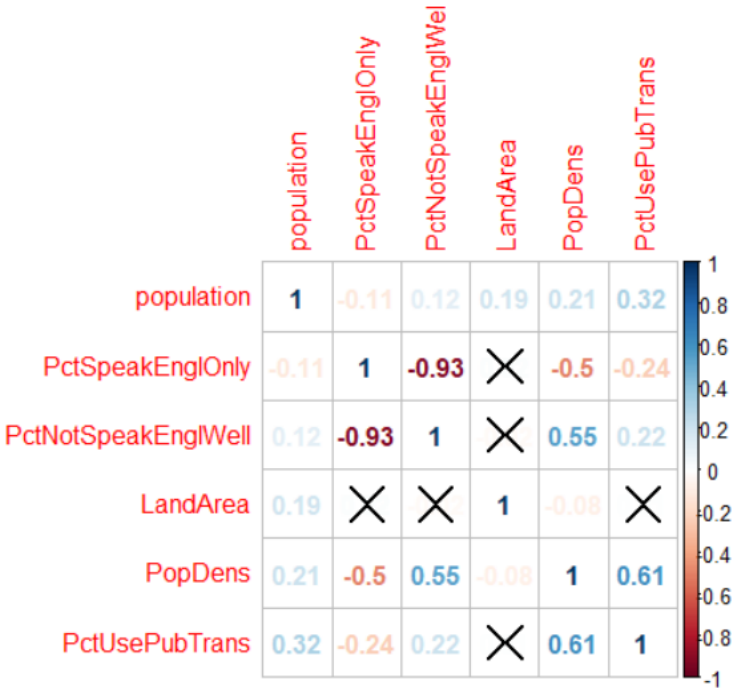
Poverty

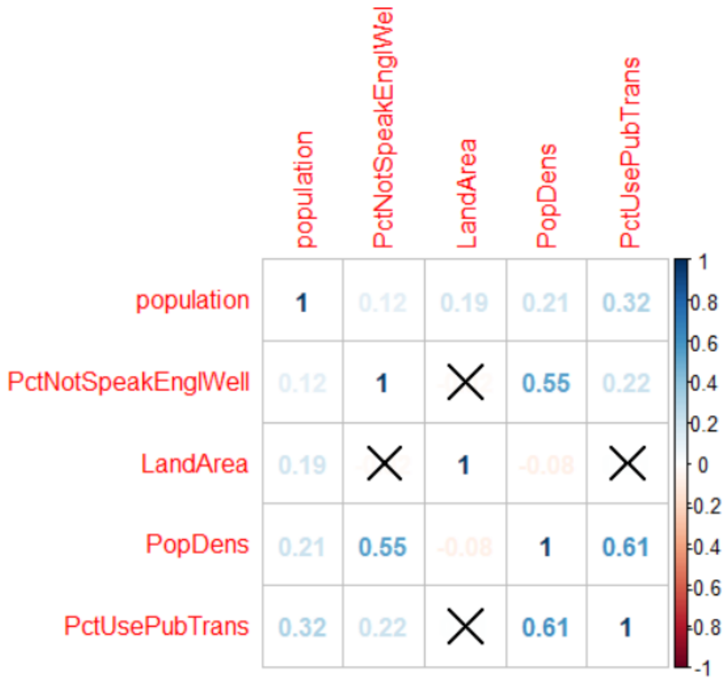
Population Change

총 17 개의 그룹

변수 선택

변수 선택





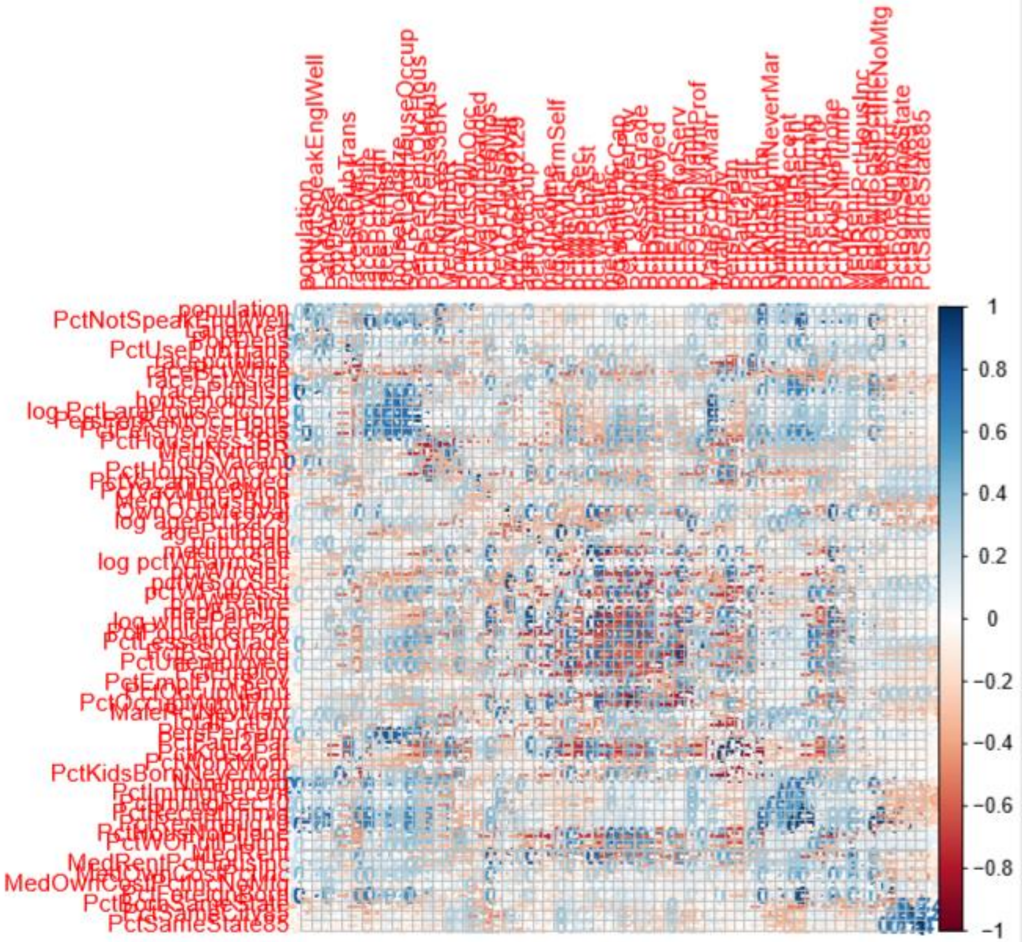
Group 별로 corrplot 그려서 **“상관관계 높은 변수 삭제”**

변수 선택

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	others	race	House	Age	urban	Income	Race Inco	Economic	Education	Employme	Marital St	Family For	Immigrant	Ownershi	Rent	Population Change	
2	populatior	racepctbla	householc	log agePc	pctUrban	medIncor	log whiteF	PctPopUn	PctLess9tl	PctUnemp	MalePctN	PersPerFa	NumImmi	PctHousN	MedRent	PctForeignBorn	
3	LandArea	racePctWh	log PctLar	agePct65up		log pctWF	OtherPerCap		PctBSorM	PctEmploy	TotalPctDi	PctFam2P	PctImmigF	PctWOFul	MedRentF	PctBornSameState	
4	PopDens	racePctAsi	PersPerRentOccHous			pctWInvInc				PctEmplProfServ		PctWorkM	PctRecentImmig		MedOwn(	PctSameCity85	
5	PctUsePul	racePctHis	PctPersDenseHous			pctWSocSec				PctOccupMgmtProf		PctKidsBornNeverMar			MedOwn(	PctSameState85	
6			PctHousLess3BR			pctWPubAsst											
7			MedNumBR			pctWRetire											
8			HousVacant			medFamInc											
9			PctHousOwnOcc														
10			PctVacantBoarded														
11			PctVacMore6Mos														
12			MedYrHousBuilt														
13			OwnOccMedVal														

1차로 58개 변수 선택

변수 선택



끔찍..

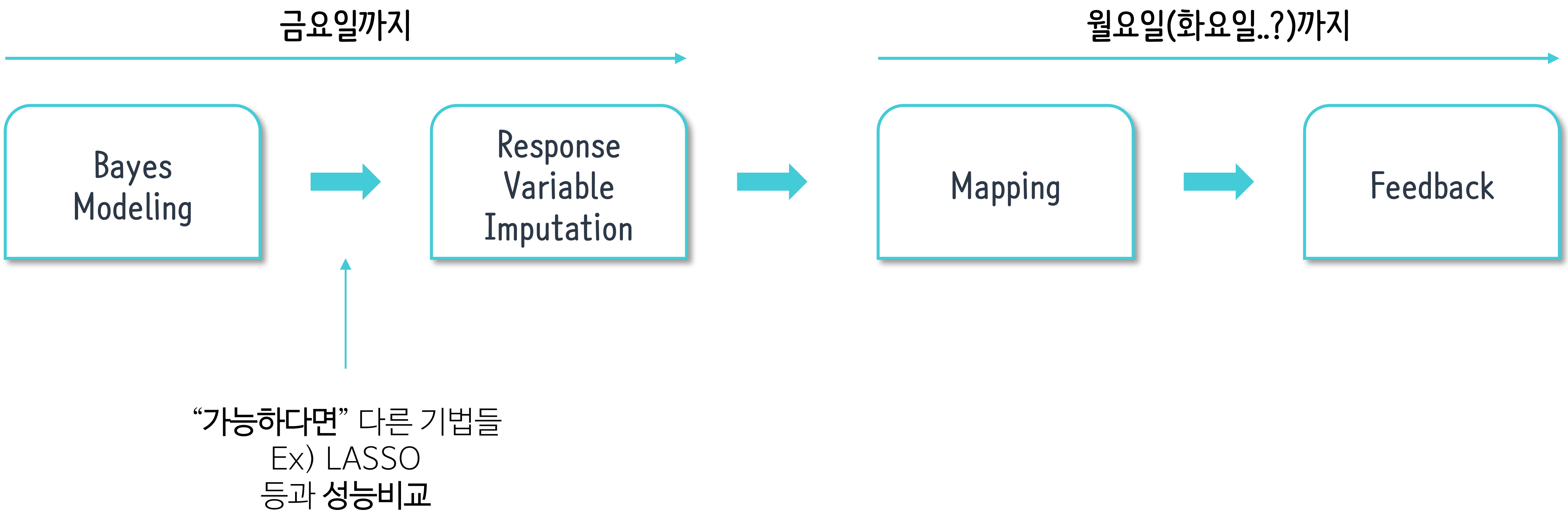
전체 corrplot그리고,
correlation 구해서 **“2차로 변수 삭제”**

변수 선택

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	others	race	House	Age	urban	Income	Race Incor	Economic	Education	Employme	Marital Sta	Family For	Immigrant	Ownership	Rent	Population Change	
2	populatior	racepctbla	household	log agePct	pctUrban	medIncom	OtherPerC	PctPopUn	PctLess9th	PctUnemp	MalePctNe	PctKids2Pa	NumImmig	PctWOFull	MedRentP	PctBornSameState	
3	LandArea	racePctWh	PersPerRe	agePct65up		log pctWFarmSelf			PctBSorMo	PctEmploy	TotalPctDi	PctWorkM	PctImmigRecent		MedOwnC	PctSameCity85	
4	PopDens	racePctAsi	PctPersDenseHous			pctWInvInc				PctEmplProfServ		PctKidsBor	PctImmigRec10		MedOwnC	PctSameState85	
5	PctUsePub	racePctHis	PctHousLess3BR			pctWRetire				PctOccupManu			PctRecentImmig				
6			MedNumBR														
7			HousVacant														
8			PctHousOwnOcc														
9			PctVacantBoarded														
10			PctVacMore6Mos														
11			MedYrHousBuilt														

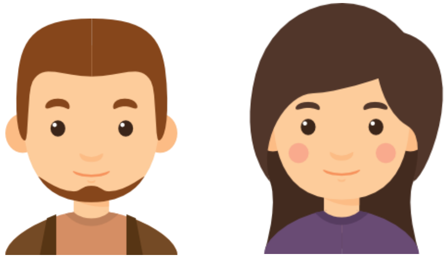
2차로 49개 변수 선택

향후 계획


금요일까지
월요일(화요일..?)까지
Bayes Modeling
Response Variable Imputation
Mapping
Feedback
"가능하다면" 다른 기법들
Ex) LASSO
등과 성능비교

역할 배분

Bayes Modeling

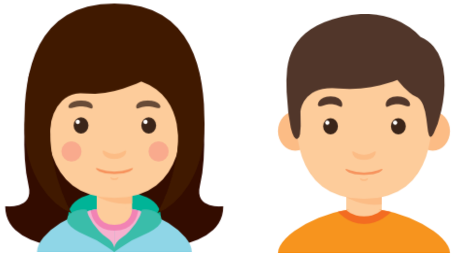


상준　　우현

Response Variable Imputation

은지

Mapping



솔희　　윤환

Feedback

다같이

Q & A