Final Project

김민정 김윤전 최은성 박대한 이재상

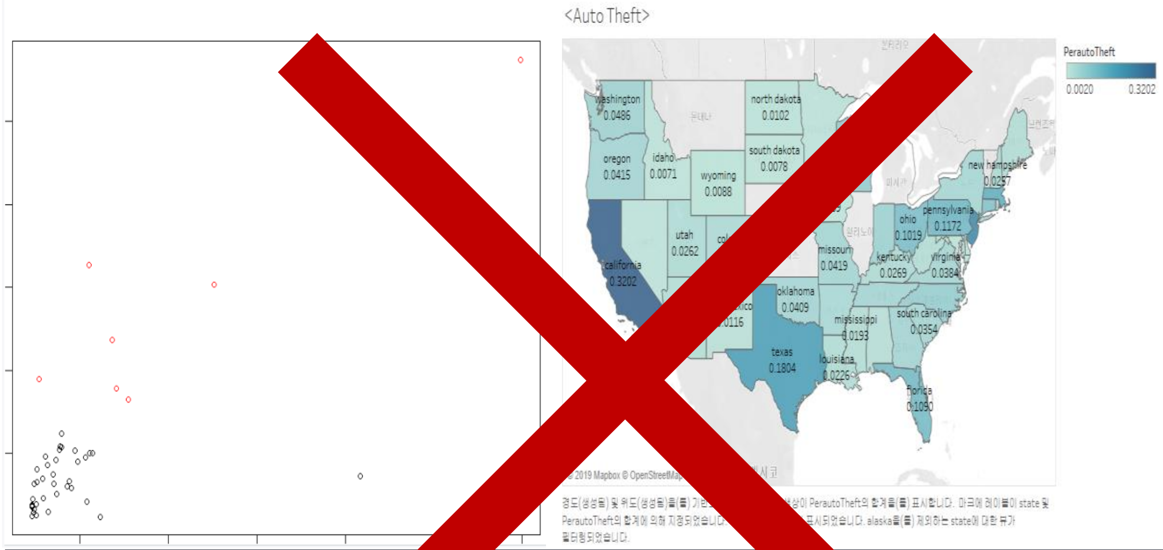
목차

1. EDA Review

EDA Review - 지난 시간 Review

- ✓ 프로젝트 목적 및 범죄 데이터 확인
- ✓ 주요 변수 설명
- ✓ 결측치 비율 84% 이상인 22개 X변수 삭제
- ✓ EDA 결과를 범죄별로 시각화
- ✓ K-Means Clustering을 통해 범죄율이 높게 나온 도시들 간의 상관관계 확인

EDA Review – 추가 진행내역(1/2)




```
# A tibble: 2 x 9
  kmeans Permurders Perrapes Per…    …es Perassaults Perburglaries Per…    …s PerautoTheft Perarsons
   <dbl>      <dbl>    <dbl>        <dbl>       <dbl>         <dbl>                <dbl>     <dbl>
1      1   0.000710   0.0057…      0.0152      0.0221        0.0358    0.06…        0.0270   0.00531
2      2   0.00332    0.03…        0.104       0.147         0.243     0.386        0.171    0.0339
```

1. K-Means Clustering 철회:

- ✓ 이유1: 동일 주 내 극단적으로 범죄율이 높은 일부 outlier 도시들 존재로 clustering에 문제 발생
- ✓ 이유2: community/county별로 clustering시 NA값을 채우지 못함

EDA Review – 추가 진행내역(2/2)

2. Murder에 대해 clustering:

- ✓ Murder에 NA가 존재하지 않아 없어서 x변수로 활용
- ✓ Murder가 다른 도시 범죄 비율 murderperpop으로 4개 도시별 그룹핑

3. X 변수 관련 추가 변경 내역:

- ✓ X 변수들 간의 Correlation 확인 후 X 변수 추가 삭제 및 파생 변수 추가 (X 변수 총 102개 → 71개: 31개 감소)

[참고] X 변수 관련 추가 변경 내역

Deleted 29 X Variables

- ✓ agePct12t21
- ✓ agePct12t29
- ✓ agePct16t24
- ✓ numbUrban
- ✓ whitePerCap
- ✓ blackPerCap
- ✓ indianPerCap
- ✓ AsianPerCap
- ✓ OtherPerCap
- ✓ HispPerCap
- ✓ NumUnderPov
- ✓ MalePctDivorce
- ✓ FemalePctDiv
- ✓ PersPerFam
- ✓ PctKids2Par
- ✓ PctYoungKids2Par
- ✓ PctTeen2Par
- ✓ PctWorkMomYoungKids
- ✓ NumKidsBornNeverMar
- ✓ PctImmigRec5
- ✓ PctImmigRec8
- ✓ PctRecImmig5
- ✓ PctRecImmig8
- ✓ PctSpeakEnglOnly
- ✓ PctLargHouseFam
- ✓ OwnOccLowQuart
- ✓ OwnOccHiQuart
- ✓ RentLowQ
- ✓ RentHighQ

Other Changes

- ✓ racePctOther <- racePctAsian + racePctHisp
- ✓ Numhomeless <- NumInShelters + NumStreet
- ✓ pctUrban: 50 이상이면 1 이하면 0 으로 변환

2. Exploratory Factor Analysis

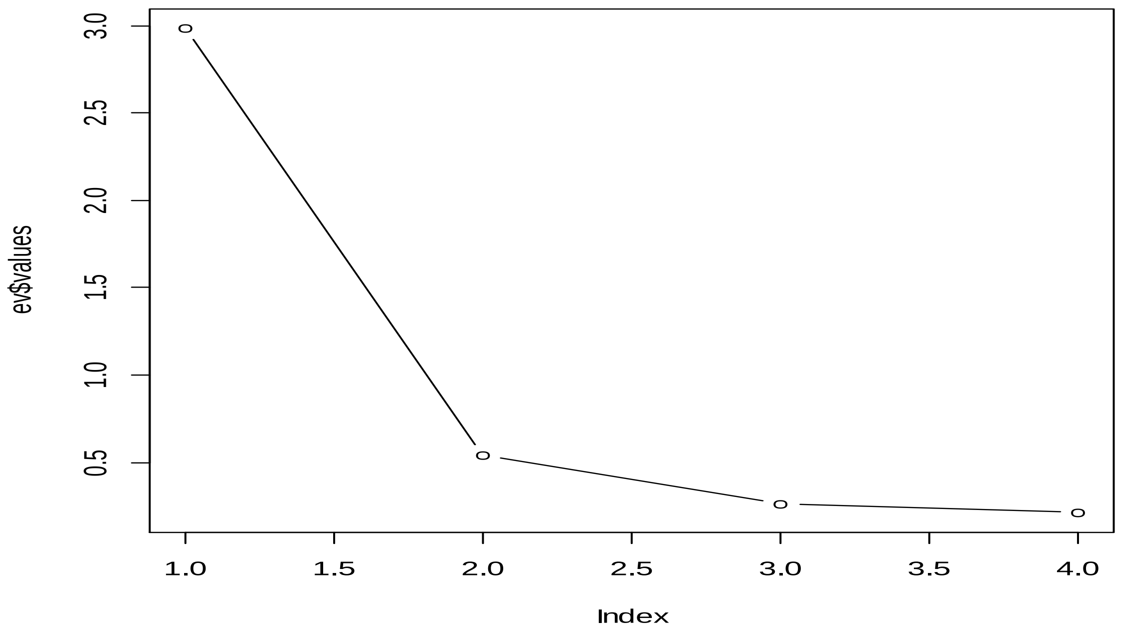
Exploratory Factor Analysis (1/4)

```
cor_mat<-cor(df_hier)
ev <- eigen(cor_mat) # Eigenvalue decomposition
ev$values

ev$values[ev$values>1]
ev$values/length(ev$values)
cumsum(ev$values/length(ev$values)) ## Cum % >0.6: 4~7 factors
windows()
plot(ev$values,type="b") ## Scree plot: 5 factors
fit <- factanal(df_hier,2, rotation="varimax", scores="regression")
prod<-data.frame(df_hier,fit$score)
dim(prod)
head(prod)
names(prod)[6:7]<-c("F_violent","F_nonviolent")
```

Exploratory Factor Analysis 사용 : model 28개

Exploratory Factor Analysis (2/4)



Scree plot of factors

- 두개의 factor를 사용하여 5개의 변수 요약 가능

Exploratory Factor Analysis (3/4)

```
names(prod)[6:7]<-c("F_violent","F_nonviolent")
scaled_factor<-scale(df_hier)
d<-dist(scaled_factor,method='euclidean')
dend <- hclust(d, method="ward.D")
plot(dend)
rect.hclust(dend, k=4, border="red")
cutree(dend, k=4)->group
aggregate(scaled_factor,by=list(cutree(dend, k=4)),mean)
```

Factor 1: Violent- murderPerPop ,larcPerPop, burglPerPop

Factor 2: Non violent- autoTheftPerPop, robbbPerPop

Exploratory Factor Analysis (4/4)

```
> aggregate(scaled_factor,by=list(cutree(dend, k=4)),mean)
  Group.1 murdPerPop robbbPerPop burglPerPop larcPerPop autoTheftPerPop
1       1  0.7349536   0.5811006  -0.7669493  0.5694618      -0.4839867
2       2 -0.3938060   0.4998432   0.7005863 -0.6035829      -0.2853774
3       3 -0.3145813  -1.0275028   0.8382689 -0.3443277       0.3108500
4       4  0.1701831  -0.5536695  -0.9851527  0.6610478       0.6489531
```

X변수 Scaling (1900개)

– col 별로 -> 비슷한 eigenvalue 들로 factor analysis
- murder에 한 열 추가 - cluster
- cluster 별로 groupby
- Final data : scale + group

-> Na가 없는 1900개 data 대해 x변수 scale

3. Bayesian Model Selection

Model Selection goals

df_x : grouped and scaled

MURDER를 X 변수로 설정

- ✓ Cutoff Value별 Factor화시켜 Murder별 다른 모델을 생성 (7가지 범죄 X 4개 Factor = 총 28개 모델)
- ✓ murder 약한 상관관계 나와서 factor 처리함

df_y: log transformed

7개 범죄별 PerPop을 Y 변수로 두고 예측

- ✓ Violent / NonViolent로 더하여 값을 찾음
- ✓ NA가 존재하는 총 16개의 열 중, 7개 예측 시
  나머지 9개 열에 대한 NA를 찾을 수 있음
- ✓ Y는 log 취해서 표준화
- ✓ Log 변환 후 y 변수 변환
- ✓ Log 변환을 통해 normal 분포를 만족하도록
- ✓ Log에서 0값으로 인해 –inf 되는 경우 로그 변환 후의 min 값으로 대체

Model Selection goals

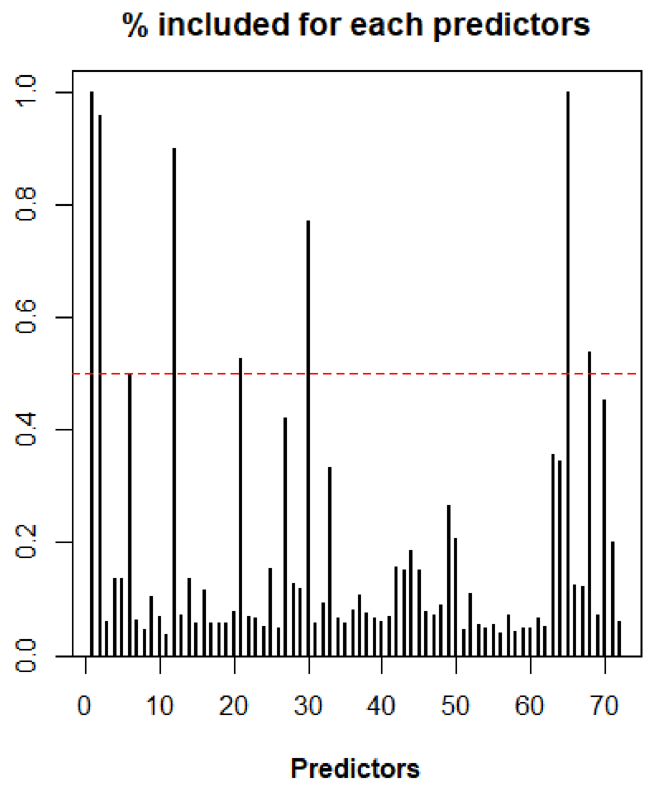
```
crime_x1 <- crime_x %>%
  filter(group==1)
crime_x2 <- crime_x %>%
  filter(group==2)
crime_x3 <- crime_x %>%
  filter(group==3)
crime_x4 <- crime_x %>%
  filter(group==4)
crime_y1 <- crime_y[which(crime_x$group==1),]
crime_y2 <- crime_y[which(crime_x$group==2),]
crime_y3 <- crime_y[which(crime_x$group==3),]
crime_y4 <- crime_y[which(crime_x$group==4),]
```

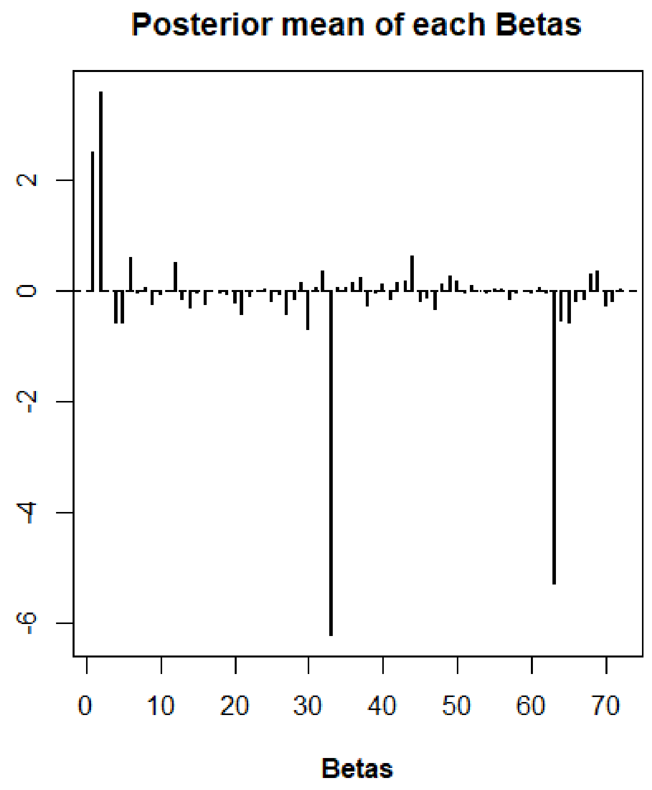
Bayesian regression 을 할 때

EFA에서 만들어진 4가지 그룹으로 나누어서 추정하고자 하는 7가지 변수에 대해 4번씩 모델을 만듦

Bayesian Model Selection - Method

Example:





- ✓ Gibbs Sampling을 반복하여 변수들이 선택된 비율이 특정 비율 이상(0.25)인 변수들을 추출하였음
- ✓ 그 결과, 7가지 범죄와 Factor별 중요한 변수들이 상이하였음

Bayesian Model Selection Results- Autotheft

Autotheft 1

	V1
beta interc	-0.44594
beta robb	0.088772
beta assau	0.058572
beta burg	0.349264
beta larcP	0.069409
beta autoT	-0.04248
beta arsor	0.072695
beta pctU	-0.02687
beta medI	0.491729
beta pctW	-0.1244
beta medI	-0.62299
beta Male	0.029533
beta PctRe	-0.07697
beta PctRe	0.496907
beta PctLa	0.038375
beta PersP	-0.06963
beta PctVa	-0.02979
beta Rent	0.052573
beta Med	0.036525
beta PctFo	-0.66642
beta PctSa	-0.08686
beta Lema	0.06551

Autotheft 2

	V1
beta interc	2.269746
beta robbl	0.125042
beta assau	0.160341
beta burgl	0.024659
beta larcPe	-0.07366
beta autoT	-0.06419
beta arsor	0.122539
beta pctUr	0.005094
beta medI	-0.14728
beta Total	0.269897
beta PctVa	0.001648
beta PctVa	-0.03879
beta MedY	-0.03774
beta PctFo	-0.09737

Autotheft 3

	V1
beta interc	-0.47587
beta robb	0.088149
beta assau	0.056748
beta burgl	0.352797
beta larcPe	0.076111
beta autoT	-0.04754
beta arsor	0.071521
beta pctUr	-0.02385
beta medI	0.495599
beta pctW	-0.12754
beta medF	-0.63013
beta Male	0.028228
beta PctRe	-0.09359
beta PctRe	0.520178
beta PctLa	0.038631
beta PersP	-0.07445
beta PctVa	-0.02933
beta Rent	0.057977
beta Med	0.034463
beta PctFo	-0.66927

Autotheft 4

	V1
beta interc	0.885091
beta robbl	0.177184
beta assau	0.120174
beta burgl	-0.25766
beta larcPe	0.232602
beta autoT	0.105206
beta arsor	0.061646
beta pctUr	0.016006
beta medI	-0.25505
beta pctW	0.022023
beta PctLe	-0.10608
beta PctUr	0.082951
beta PctEr	0.018835
beta Total	0.071009
beta PctW	-0.0131
beta PctHo	0.043982
beta PctVa	0.036643

Bayesian Model Selection Results- Arson

Arson 1

	V1
beta interc	2.263006
beta popu	5.208182
beta raceP	0.72498
beta pctW	0.488629
beta PctBS	-0.16635
beta PctO	-0.27743
beta PctFa	-0.62111
beta Num	-4.92842
beta PctHo	0.128919
beta Num	-5.54719
beta PctFo	-0.2634
beta PctBo	-0.62037
beta PctSa	0.182061
beta PopD	-0.25404

Arson 2

	V1
beta interc	3.091655
beta popu	1.149677
beta racep	-0.22009
beta perCa	-0.11945
beta PctLe	-0.13608
beta PctEr	-0.10143
beta PctO	0.091709
beta Total	0.174156
beta MedY	-0.0506
beta PctSa	0.10917

Arson 3

	V1
beta interc	2.756402
beta popu	2.22869
beta raceP	0.231462
beta pctUr	-0.13534
beta medI	0.808314
beta perCa	-0.64393
beta PctLe	-0.34414
beta PctO	0.111148
beta Total	0.225954
beta Num	-2.28383
beta PctLa	-0.48794
beta PctPe	0.938203
beta MedY	-0.20408
beta Rent	-0.12138
beta PctFo	-0.12784

Arson 4

	V1
beta interc	3.269862
beta pctW	0.354035
beta PctPo	-0.4077
beta PctNo	-0.14627
beta PctEr	-0.24441
beta PctEr	0.213259
beta PctEr	0.137982
beta Total	0.1929
beta Hous	0.0382
beta PctVa	0.163959
beta PctSa	-0.28806
beta LandA	0.243108

Bayesian Model Selection Results- Rapes

Rapes 1

	V1
beta interc	3.104466
beta popu	0.414678
beta pctUı	0.137905
beta pctW	-0.21709
beta PctPc	0.31072
beta Total	0.196731
beta Hous	1.56629
beta PctVa	0.146423
beta PctVa	-0.18525
beta PctFc	-0.23107

Rapes 2

	V1
beta interc	3.281872
beta popu	0.398173
beta raceF	-0.17354
beta pctW	0.120553
beta PctLe	-0.22222
beta PctO	-0.15434
beta Total	0.297189
beta PctFa	-0.12015
beta Hous	0.077431
beta MedI	-0.17896
beta MedI	0.11945

Rapes 3

	V1
beta interc	3.113675
beta perCa	-0.28015
beta PctLe	-0.24551
beta Total	0.22301
beta PctPe	-0.28746
beta Hous	0.205026
beta PctH	0.124446
beta PctH	0.109124
beta Num	0.959701
beta PopD	-0.21727

Rapes 4

	V1
beta interc	3.493572
beta medI	-0.31855
beta pctW	-0.1027
beta PctLe	-0.29828
beta Total	0.182168
beta PctKi	0.051647
beta PctLa	0.194753
beta PctPe	0.202509
beta PctH	-0.38782
beta PctH	0.113383
beta Own	-0.1657
beta RentI	0.221143
beta Med	-0.15377
beta PctSa	-0.05181
beta Land	0.088917
beta PopD	-0.05911

Bayesian Model Selection Results- Robberies

Robberies 1

	V1
beta interc	3.929177
beta popu	3.984832
beta hous	-0.14395
beta racep	0.023489
beta raceP	-0.74812
beta pctUi	0.223378
beta pctW	-0.10129
beta PctUi	-0.19477
beta PctO	-0.17121
beta Total	0.042201
beta PctFa	-0.19855
beta Num	-4.29643
beta Own(	0.143567
beta Rent(	-0.10474
beta PopD	0.197848

Robberies 2

	V1
beta interc	4.760565
beta popu	0.404349
beta racep	0.198145
beta raceP	-0.19891
beta pctUi	0.159463
beta PctPc	-0.14756
beta PctLe	-0.30186
beta PctN	0.352758
beta Total	0.120306
beta PctFa	-0.09089
beta PctKi	0.169347
beta PctVa	-0.08287
beta PctFc	0.201969

Robberies 3

	V1
beta interc	4.331637
beta racep	0.646897
beta ageP	0.249421
beta pctUi	0.228639
beta medI	-0.03927
beta PctPc	-0.23252
beta PctLe	-0.2991
beta PctN	0.536401
beta PctO	-0.13036
beta PctO	0.183885
beta Male	0.304979
beta Total	0.41244
beta PctW	-0.1086
beta PctN	0.315833
beta PctLa	-0.42015
beta PersF	0.472452
beta PctW	-0.10689

Robberies 4

	V1
beta interc	5.394944
beta raceP	-0.10327
beta pctUi	0.127852
beta perCa	0.128078
beta Total	0.236132
beta PctKi	0.22665
beta PctH	0.041728
beta MedY	-0.08602
beta Med(	-0.08288
beta PctFc	0.04886
beta PctBc	-0.10796
beta Lema	0.038323

Bayesian Model Selection Results- Larcenies

Larcenies 1

	V1
beta interc	-1.64452
beta robbl	0.002054
beta assau	0.211202
beta burgl	0.223532
beta larcP	0.23749
beta autoT	0.017858
beta arsor	0.067244
beta popu	0.722698
beta hous	0.040312
beta PctPc	0.224332
beta PctUı	-0.07712
beta PctO	-0.10978
beta Total	0.139499
beta Rent(	-0.00932
beta PctSa	-0.07193

Larcenies 2

	V1
beta interc	1.30325
beta robbl	0.164039
beta assau	0.178592
beta burgl	0.074176
beta larcP	-0.04991
beta autoT	-0.03563
beta arsor	0.152133
beta pctW	-0.06283
beta PersF	-0.15812
beta Own(	-0.06066
beta Rentl	-0.11187

Larcenies 3

	V1
beta interc	0.128943
beta robbl	0.041866
beta assau	0.077841
beta burgl	0.330052
beta larcP	0.08151
beta autoT	-0.11991
beta arsor	0.06834
beta pctW	0.13926
beta pctW	-0.10445
beta Total	0.18786
beta Rentl	0.027431
beta Medl	-0.21919

Larcenies 4

	V1
beta interc	0.39558
beta robbl	0.200781
beta assau	0.220583
beta burgl	-0.14602
beta larcP	0.058969
beta autoT	0.14179
beta arsor	0.121084
beta PersF	-0.20164
beta PersF	0.087064
beta PctH	0.262648
beta PctBc	0.021605
beta PopD	-0.09501

Bayesian Model Selection Results- Assaults

Assaults 1

	V1
beta interc	4.836175
beta raceP	0.434151
beta pctUr	0.121206
beta pctW	-0.70599
beta PctO	-0.18262
beta PctW	0.09757
beta PctH	-0.06654
beta Rent	0.083834
beta PctSa	0.112769

Assaults 2

	V1
beta interc	5.573422
beta popu	0.38223
beta pctW	0.087409
beta pctW	-0.48781
beta PctPo	-0.12917
beta PctN	0.231801
beta PctUr	-0.09355
beta PctEr	-0.1718
beta Male	-0.0794
beta PctKi	0.295399
beta PersP	-0.15881
beta Hous	0.208707
beta Own	0.147987

Assaults 3

	V1
beta interc	5.316738
beta raceP	-0.15109
beta raceP	0.10059
beta pctW	-0.33724
beta Total	0.149071
beta PctFa	-0.2315
beta PctH	-0.1028
beta RentI	-0.24973
beta Rent	0.095486
beta MedI	0.360769
beta MedI	0.120263

Assaults 4

	V1
beta interc	6.001709
beta pctW	-0.16758
beta Total	0.109744
beta PctFa	-0.2507
beta PersP	0.069055
beta PctBo	-0.1409
beta PopD	-0.09509

Y – NA 범죄 종류별로 뽑은 결과

NA arson

communit	state	countyCoc	communit
Gloversvill	NY	35	29443
Amsterdar	NY	57	2066
SanAngelc	TX	NA	NA
Bloomingt	MN	53	6616
NewIberia	LA	NA	NA
Selmacity	AL	NA	NA
RapidCityc	SD	103	52980
Wichitacity	KS	173	79000
Corningcit	NY	101	18256
Dothancity	AL	NA	NA
Sparkscity	NV	NA	NA
Scottsborc	AL	NA	NA
FortDodge	IA	187	91370
Hyattsville	MD	NA	NA
Oneidacity	NY	53	54837
Genevacity	NY	69	28640
Longviewc	WA	NA	NA

NA assault

communit	state	countyCoc	communit
Laurelcity	MS	NA	NA
Lancasterc	OH	45	41720
UniversalC	TX	NA	NA
Dumascity	TX	NA	NA
Allencity	TX	NA	NA
Worcester	MA	27	82000
Houstonci	TX	NA	NA
Fairfieldto	CT	1	26620
Garlandcit	TX	NA	NA
Akroncity	OH	153	1000
Springfielc	MA	13	67000
Anaheimci	CA	NA	NA
Bristoltow	CT	3	8490

NA autotheft

communit	state	countyCoc	communit
Lawrencec	MA	9	34550
NewBedfo	MA	5	45000
Saugustov	MA	9	60015

NA burglaries

communit	state	countyCoc	communit
Mesquitec	TX	NA	NA
Lamesacity	TX	NA	NA
BayCitycity	TX	NA	NA

Y – NA 범죄 종류별로 뽑은 결과

NA robberies

communit	state	countyCoc	communit
Bemidjicity	MN	7	5068
NewUlmci	MN	15	46042
Maplewoc	MN	123	40382
Plymouthc	MN	53	51730
Pontiaccity	MI	125	65440
Wyomingc	MI	81	88940
Hastingsci	MN	37	27530
ParkForest	IL	NA	NA
Bloomingt	MN	53	6616
Wheatonc	IL	NA	NA
WhiteLake	MI	125	86860
Pittsfieldtc	MI	161	64560
Algonquin	IL	NA	NA
Doltonvill	IL	NA	NA
Sturgiscity	MI	149	76960
Richfieldci	MN	53	54214
Monroecit	MI	115	55020

NA larcenies

communit	state	countyCoc	communit
Dumascity	TX	NA	NA
Lawrencec	MA	9	34550
Lamesacity	TX	NA	NA

NA rapes

communit	state	countyCoc	communit
Bemidjicity	MN	7	5068
NewUlmci	MN	15	46042
Maplewoc	MN	123	40382
Plymouthc	MN	53	51730
Pontiaccity	MI	125	65440
Wyomingc	MI	81	88940
Hastingsci	MN	37	27530
ParkForest	IL	NA	NA
Bloomingt	MN	53	6616
Wheatonc	IL	NA	NA
WhiteLake	MI	125	86860
Pittsfieldtc	MI	161	64560
Algonquin	IL	NA	NA
Doltonvill	IL	NA	NA
Sturgiscity	MI	149	76960
Richfieldci	MN	53	54214
Monroecit	MI	115	55020

→ y값 exp 변환
→ NA 그룹 분류

4. Conclusion

결론 및 개선점 제언

- 나중에 RMSE 값 소개 해야?
- 최종 모델 선정 - 예측에 이러이러한 모델 최적화 : ~ 소개 – hierichial? (베이지안 모델 적합)

- 범죄 모델 구성 시, 다른 범죄 관련 수치들을 x 변수로 적극 활용하지 못함 (NA값이 겹치는 등의 이유)
- Train set과 Test set을 구분하여 진행하지 못함
- 지역 코드 관련 데이터 활용에 한계
  → 범죄 예측보다 시각화에 중점을 두었음

Thank you!