US crime

ESC 5조

강동인 신예진 이청파 한인욱 정유진

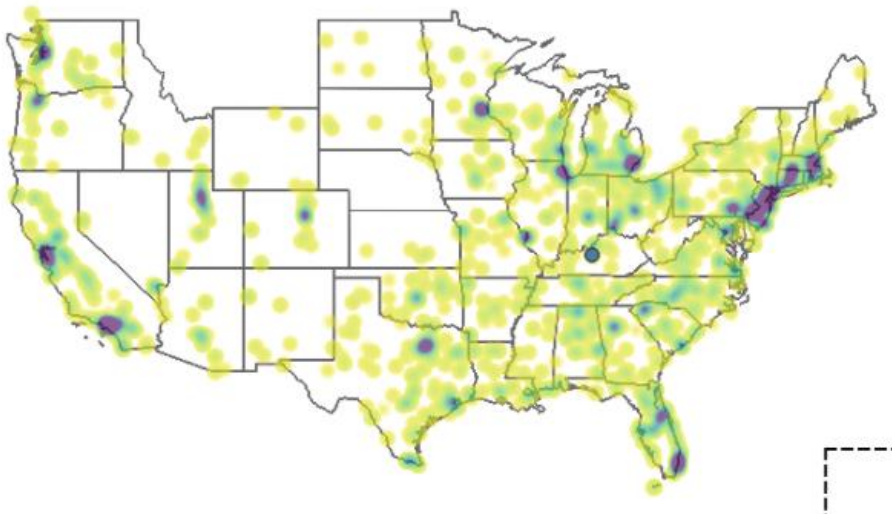
Table of Contents

1. Data

US crime data

- 미국 8개의 범죄율 및 다양한 지표를 포함하고 있는 자료
- murders 살인 / rapes 강간/ robberies 강도/ assaults 폭행 / burglaries (절도, 강도 목적의) 주거침입/ larcenies 절도 / autoTheft 차량탈취 / arsons 방화, 8종류의 범죄에 대한 범죄지도 완성하기
- 변수별 scale이 다름->scailing
- Skewed->transformation
- Correlation 높은 변수들->제거 및 변환

변수 제거		결측치 84.5% 이상
PctOccupManu	data description에 따르면 PctEmplManu와 유사	NumKindsDrugsSeiz
OwnOccQrange	1, 3분위수 변수가 있음. 3분위 - 1분위 변수 제거	LemasSwornFT
RentQrange	1, 3분위수 변수가 있음. 3분위 - 1분위 변수 제거	LemasSwFTPerPop
communityCode	필요없는 지역정보 코드&결측치	LemasSwFTFieldOps
countyCode	필요없는 주 구별 코드&결측치	LemasSwFTFieldPerPop
fold	필요없는 cv 를 나누는 코드	LemasTotalReq
NumKidsBornNeverMar	같은 의미의 비율 변수 존재	LemasTotReqPerPop
murders		PolicReqPerOffic
rapes		PolicPerPop
robberies		RacialMatchCommPol
assaults		PctPolicBlack
burglaries	y변수 인구	PolicAveOTWorked
larcenies		PctPolicAsian
autoTheft		PolicBudgPerPop
arsons		LemasGangUnitDeploy
ViolentCrimesPerPop		LemasPctPolicOnPatr
nonViolPerPop		PolicOperBudg
		PolicCars
		OfficAssgnDrugUnits
		PctPolicWhite
		PctPolicHisp



인구	경제	경제	집	개인특성	이민	환경
Population	numbUrban	PctPopUnderPov	PctHousOccup	agePct12t29	NumImmig	MalePctNevMarr
Householdsize	pctUrban	PctUnemployed	PctHousOwnOcc	agePct65up	PctImmigRecent	TotalPctDiv
PctLargHouseFam	medIncome	PctEmploy	HousVacant	PctLess9thGrade	PctRecentImmig	PctFam2Par
PopDens	pctWWage	PctHousNoPhone	racePctHisp	PctSpeakEnglOnly		PctWorkMom
LandArea	pctWFarmSelf	NumInShelters	PersPerOccupHous	PctNotSpeakEnglWell		PctKidsBornNeverMar
PctSameState85	medFamInc	NumStreet	PersPerOwnOccHous	Racepctblac		LemasPctOfficDrugUn
PctSameCity85	pctWRetire		PersPerRentOccHous	racePctWhite		PctForeignBorn
PctSameHouse85	whitePerCap		PctPersOwnOccup	racePctAsian		
PctBornSameState	blackPerCap		PctPersDenseHous	racePctHisp		
	indianPerCap		PctHousLess3BR	agePct22t29		
	AsianPerCap		MedNumBR			
	OtherPerCap		PctVacantBoarded			
	HispPerCap		PctVacMore6Mos			
	NumUnderPov					

2. Freq vs Bayes

Frequentist vs Bayesian

Frequentist : Backward selection

1. Forward Selection

- 변수를 하나씩 늘려가며 매단계마다 가장 성능이 좋은 변수를 선택한 뒤에 유의미한 성능 향상이 없을 때까지 반복

2. Backward Elimination

- 모든 변수를 가지고 시작하며, 영향이 가장 작은 변수부터 하나씩 제거

3. Ridge Regression

$$\sum_{i=1}^{M}(y_i - \hat{y_i})^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_i \times x_{ij}\right)^2 + \lambda \sum_{j=0}^{p} w_j^2$$

4. Lasso Regression

$$\sum_{i=1}^{M}(y_i - \hat{y_i})^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_i \times x_{ij}\right)^2 + \lambda \sum_{j=0}^{p} |w_j|$$

1. Ridge는 변수 선택 측면에서 적합하지 않음.
2. Bayesian linear regression과 비교하기 위해선 Ordinary Least Squares 방법을 쓰는 게 적합하다고 판단 → Ridge, Lasso 제외.
3. 최대한 많은 변수를 Bayesian 모델과 비교하기 위해 모든 변수를 가지고 시작하는 Backward Elimination을 최종 선택.

Frequentist vs Bayesian

Bayesian : Bayesian linear regression

- Single optimal model이 아닌 many possible model을 파악가능
- 모델의 uncertainty를 반영 (averaging over all possible combinations of predictors)
- **'분포'**를 보여주는 것이 Bayesian estimate

$$P(\beta|D) = \sum_{k} \mathrm{P(B|M_k, D)\,P(M_K|D)}$$

Average of posterior distributions under each model weighted by the corresponding posterior model probabilities.
=> Weighted means of beta in all possible models

Model의 uncertainty를 underestimate할 수 있는 기존의 Frequentist 방법을 보완!

3. Modeling

Backward seletion

Variable selection

```
import feature_selection as fsel
```

```
backward_col={}
for i in y_col:
    cols=fsel.backwardSelection(dic[i]['Train_X'], dic[i]['Train_y'], model_type='linear', elimination_criteria='aic')
    if 'intercept' in cols:
        cols.remove('intercept')
    if 'idx' in cols:
        cols.remove('idx')
    backward_col[i]=cols

forward_col={}
for i in y_col:
    cols=fsel.forwardSelection(dic[i]['Train_X'], dic[i]['Train_y'], model_type='linear', elimination_criteria='aic')
    if 'intercept' in cols:
        cols.remove('intercept')
    if 'idx' in cols:
        cols.remove('idx')
    forward_col[i]=cols
```

Backward Selection

```
df_backward=df_final.copy()
backward_coef=pd.DataFrame(columns=x_col)

for i, j, k in zip(y_col, nan_col, crime) :
    mlr=LinearRegression()
    mlr.fit(dic[i]['Train_X'][backward_col[i]], dic[i]['Train_y'])
    y_predict=mlr.predict(dic[i]['Pred_X'][backward_col[i]])
    df_backward[j][df_backward[j]==k]=y_predict

    coef={'Intercept':mlr.intercept_}
    for x in range(len(backward_col[i])):
        coef[backward_col[i][x]]=mlr.coef_[x]

    backward_coef= pd.concat([backward_coef,pd.DataFrame(coef, index=[i])])

backward_coef=backward_coef.T
```

```
Character Variables (Dummies Generated, First Dummies Dropped): []
Eliminated : MedOwnCostPctInc
Eliminated : MedNumBR
Eliminated : RentHighQ
Eliminated : LemasPctOfficDrugUn
Eliminated : OwnOccLowQuart
Eliminated : PctKidsBornNeverMar
Eliminated : PctHousNoPhone
Eliminated : idx
Eliminated : PctSameState85
Eliminated : indianPerCap
Eliminated : OtherPerCap
Eliminated : blackPerCap
Eliminated : PersPerFam
Eliminated : agePct65up
Eliminated : householdsize
Eliminated : racePctAsian
Eliminated : PersPerOwnOccHous
Eliminated : PctFam2Par
Eliminated : AsianPerCap
```

Backward_coefficient

medFamInc		46.71106					
whitePerC	-4.74578	17.35161		152.5707	483.3436		-5.20747
blackPerC	-1.20895				-57.479		
indianPerCap							
AsianPerCap					-71.4157		-1.46838
OtherPerC	1.56896		17.01747	18.98613	75.90884		1.333255
HispPerCap							
NumUnderPov					468.6732	84.93714	
PctPopUnderPov		-46.8153			328.1667	-108.923	-8.1604
PctLess9th	-5.85707	-28.4908		-94.8694	-290.753	-78.7332	-4.18968
PctBSorMore		-39.7344			-348.949	-87.8228	
PctUnemp	2.066396		26.77987			34.07585	
PctEmploy		46.80944	51.07658	133.7079	290.0771	95.21105	
PctEmplManu		-4.57272			-78.232		1.71319
PctEmplProfServ					-159.986		
PctOccupMgmtProf		15.97148		48.9887	382.6756	78.46833	
MalePctNevMarr		58.97967	34.86083	251.94	297.0571	223.6165	5.024239
TotalPctD	13.8478	15.23969		133.4894	464.1492	161.7945	7.786219
PersPerFam		-61.8388		-179.09			-10.2393
PctFam2P	7.131981				-524.367	57.34192	
PctWorkM	-1.38732	-26.5738	-20.4837	-29.4217		-67.9151	-1.96803

Bayesian linear regression

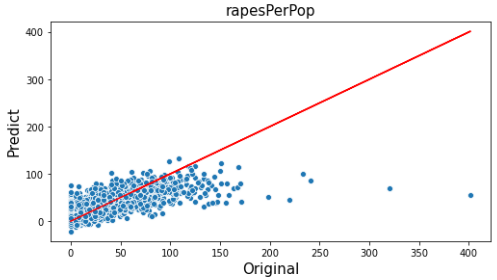
R BAS package 사용
-model prior-> uniform
-method="MCMC"-> p가 큰 경우 MCMC방법이 좋음
-각 8개의 범죄에 대해 시행

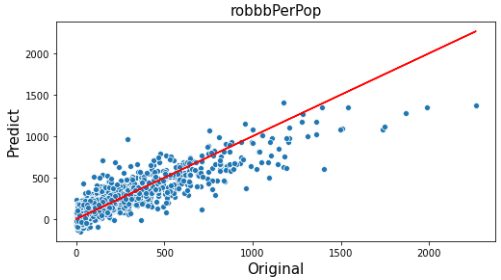
```
fit.gprior = bas.lm(f,
                    data = data,
                    method = "MCMC", # better than default "BAS"
                                     # for large p
                    prior = "ZS-null", # default
                                      # "JZS" also can use
                    modelprior = uniform(),
                    include.always = ~1,
                    MCMC.iterations = 100000)
```

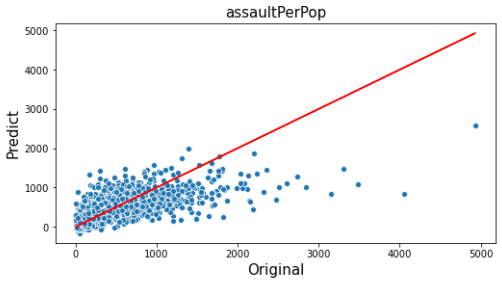
Bayesian regression (leechungpa.github.io)

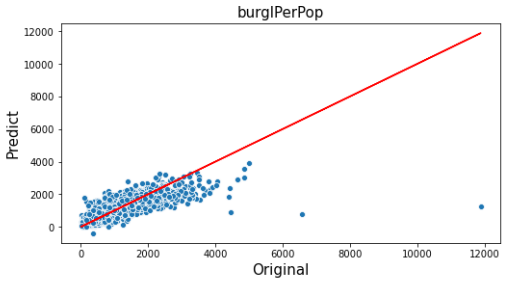
Backward elimination 결과

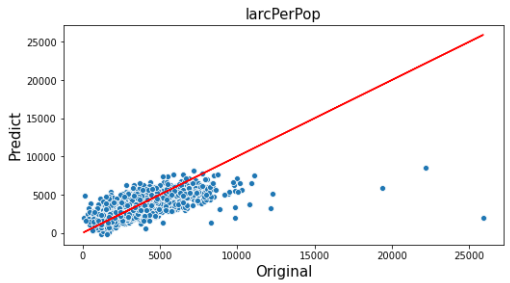
Backward Elimination

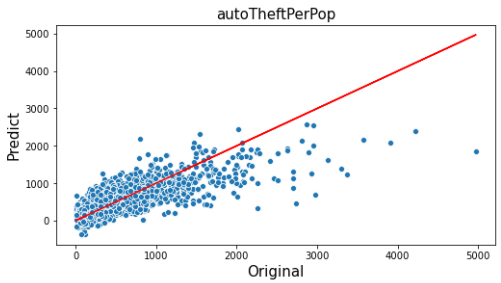


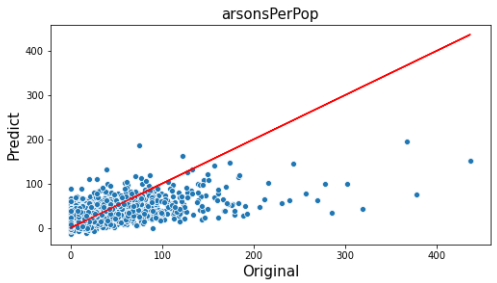




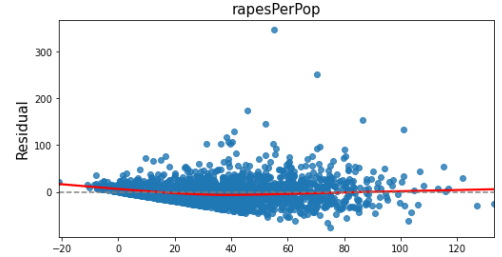


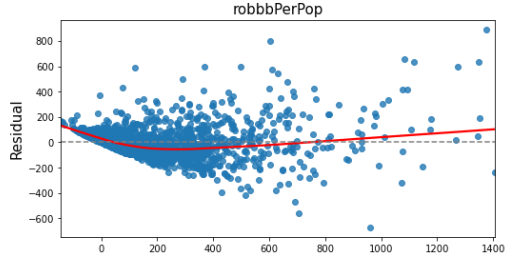


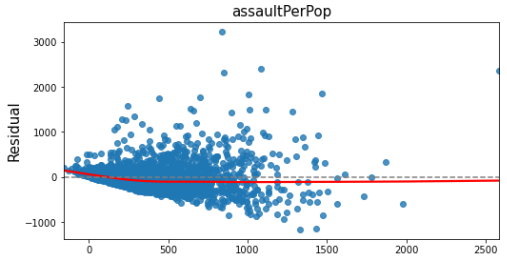


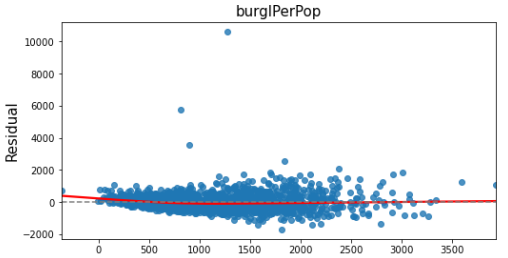


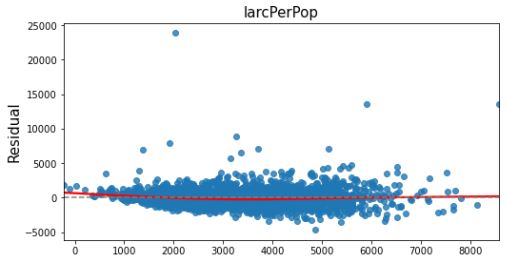
Backward Elimination

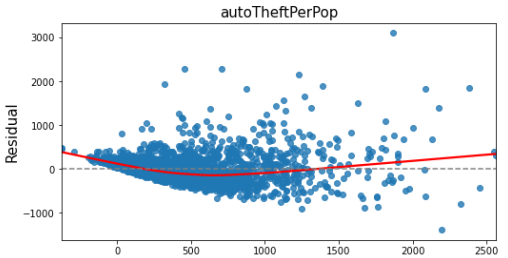


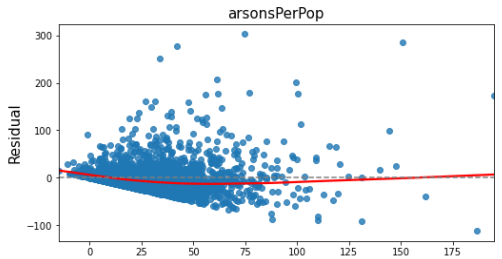




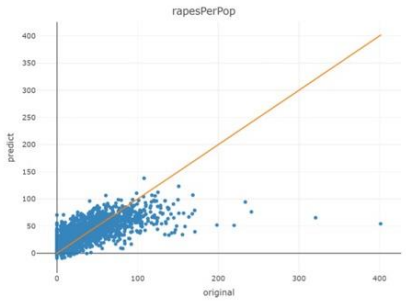


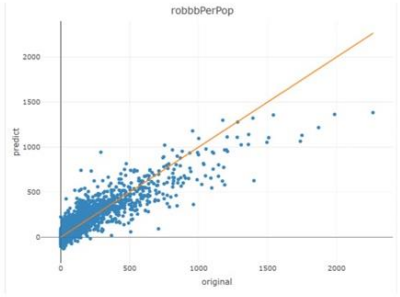


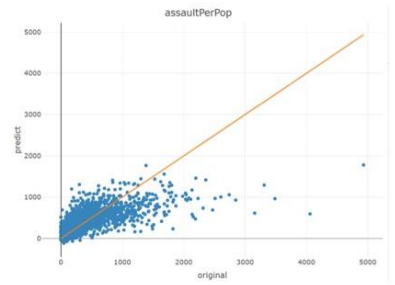


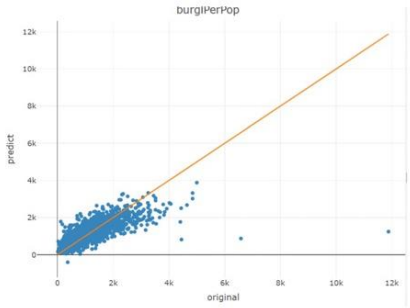


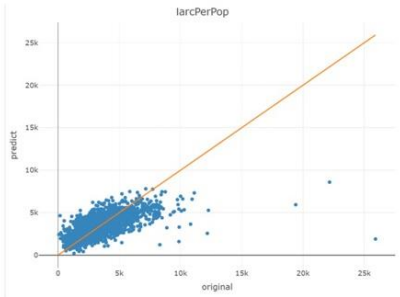
Bayesian linear regression 결과

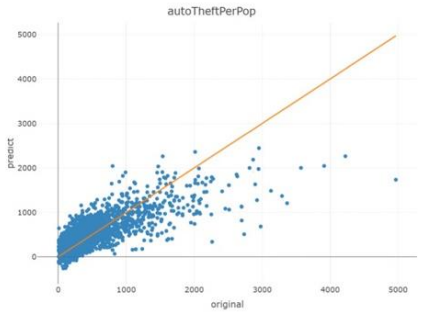


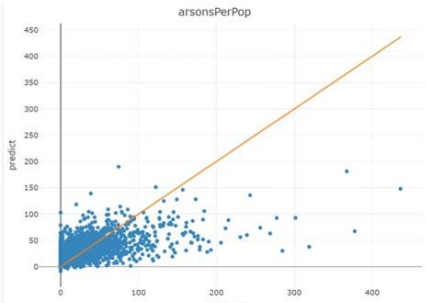


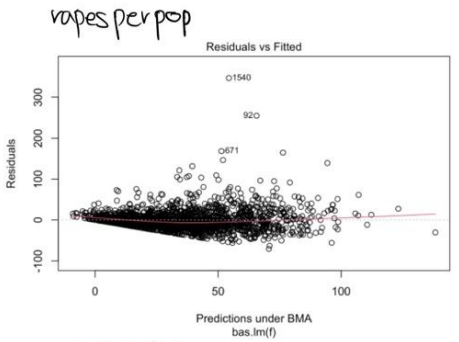


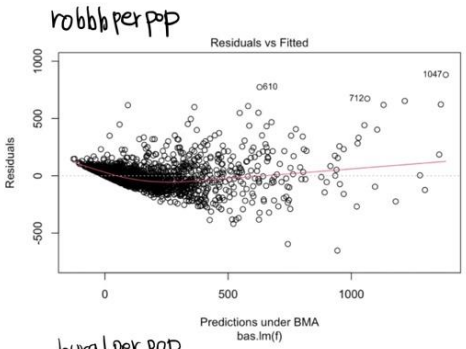


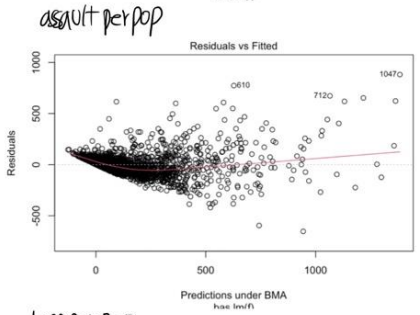


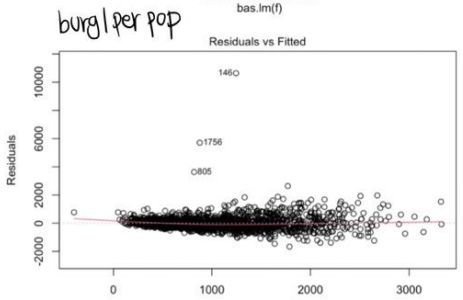


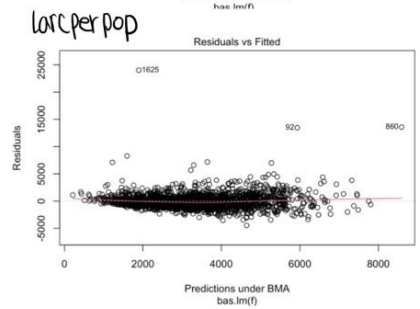


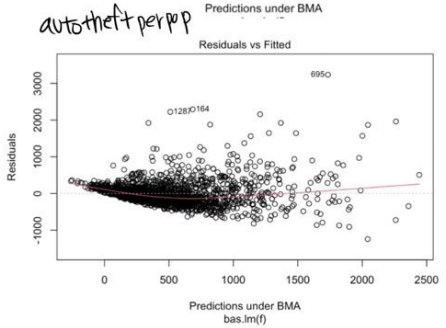


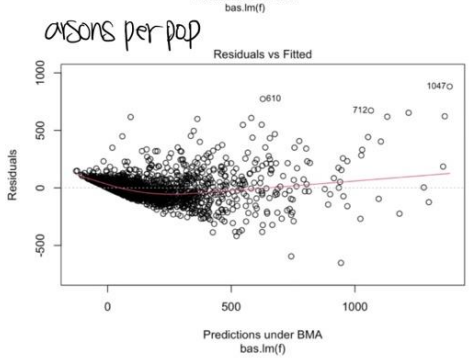




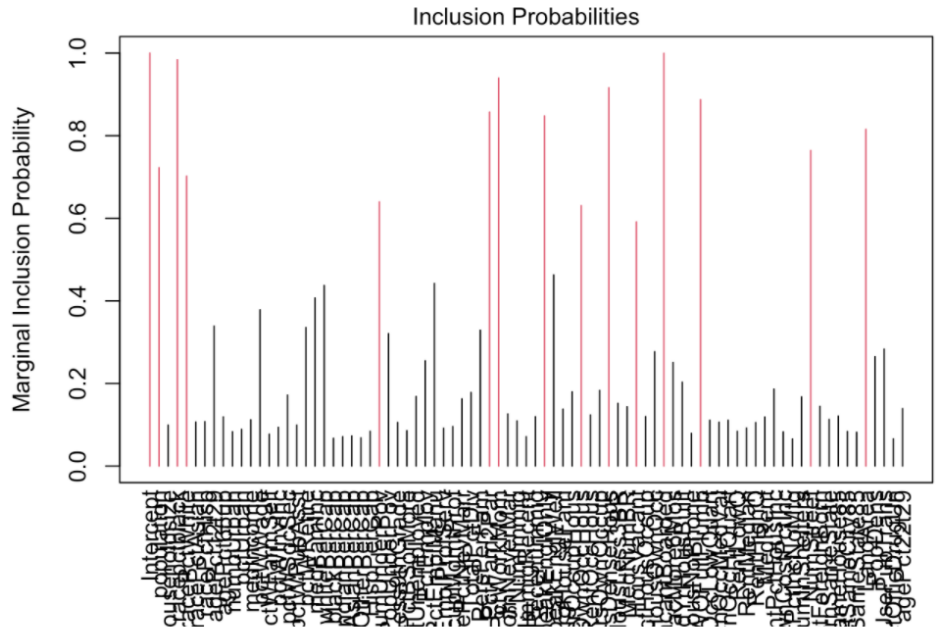


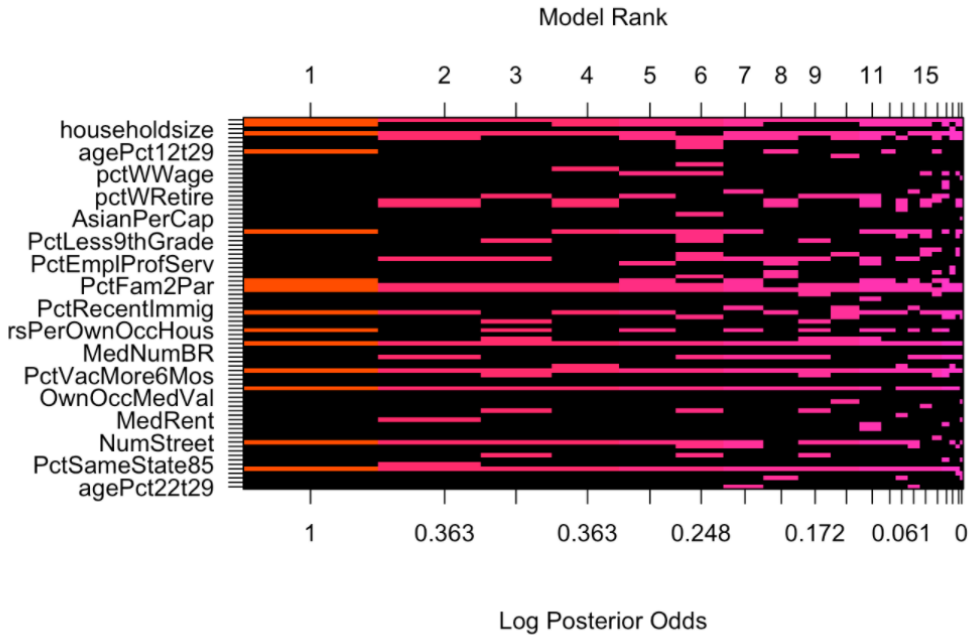




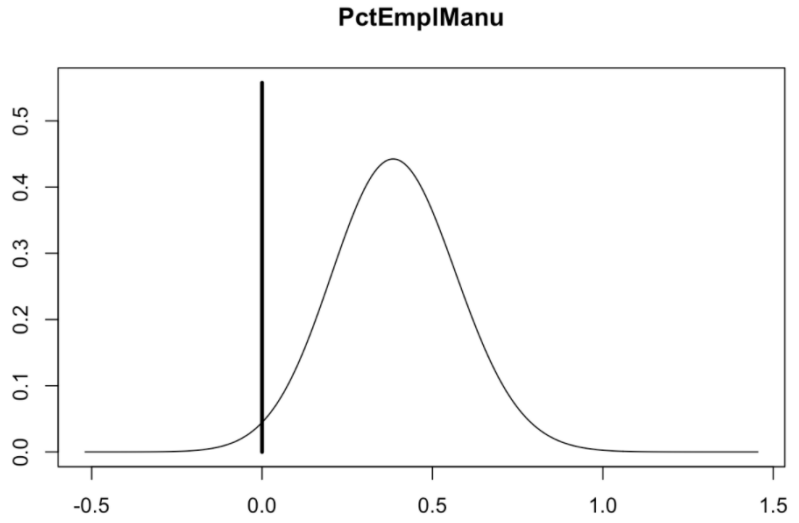


Bayesian linear regression 결과

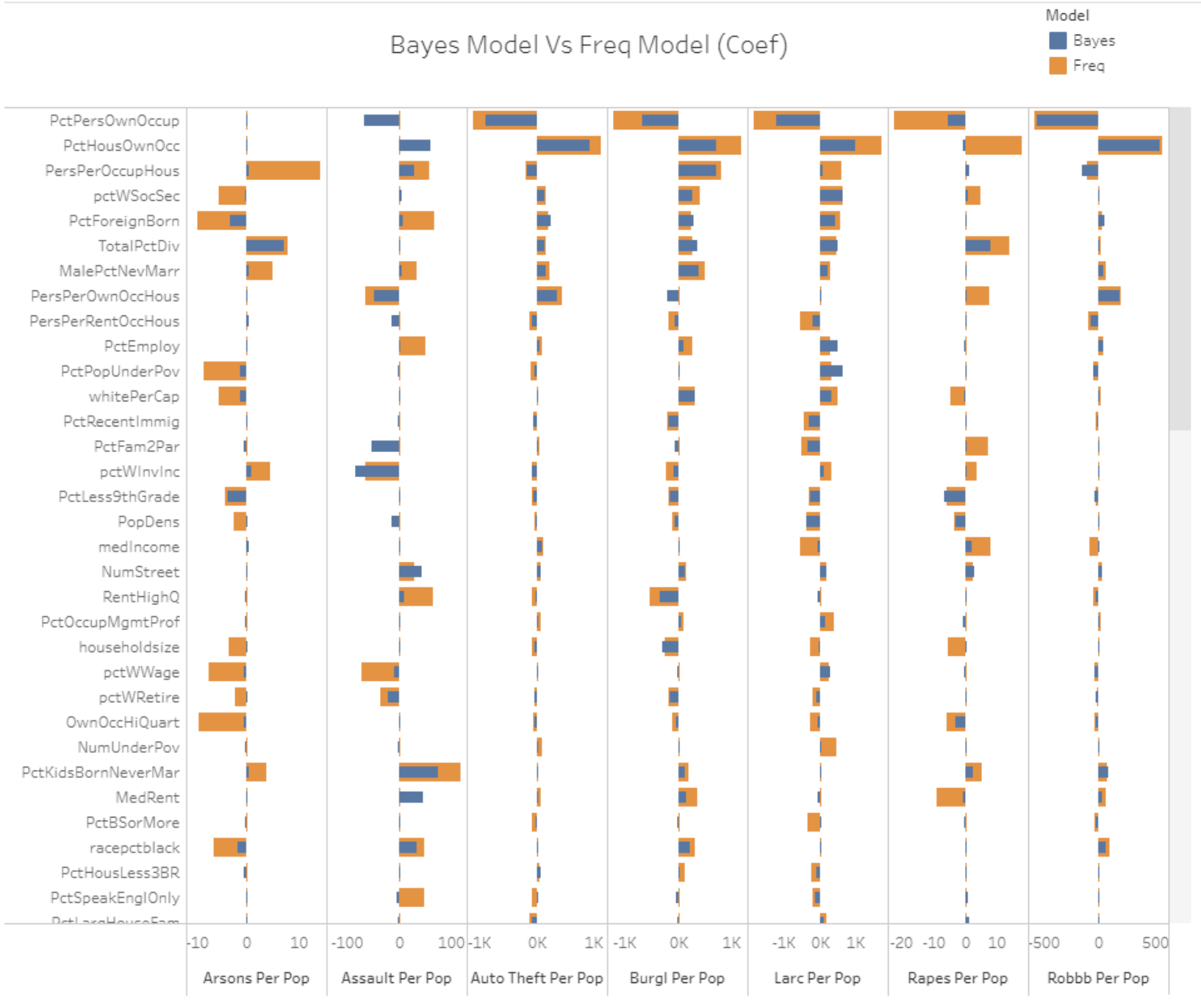




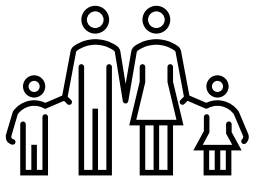
-> Inclusion probability가 높고, log posterior odds가 높은 model에 속한 변수들을 유의미하다고 판단하여 해석!



4. Interpretation



[Bayes Vs Freq 모델 계수비교 | Tableau Public]



가정환경

Property crime

-totalpctdiv: 모든 범죄율 높음

-**PctKidsBornNeverMar**: burglary, robbery에 영향 끼침

특히 가정 환경의 영향이 property crime에 많은 영향을 끼침!

Violent crime

-**TotalPctDiv**
-PctKidsBornNeverMar
모든 범죄율 높음

-> 부모가 이혼하였거나 미혼 부모 가정에서 태어난 자녀들은 상대적으로 범죄와 연관되어 있음. 모든 범죄에서 범죄율이 높음.

사회-경제

Property crime

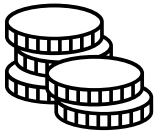
-pctemploy(고용률): property crime 모두 범죄율 높음
고용률 좋으면-> 경제 좋음 -> target attractive!

Violent crime

딱히 사회 경제와 상관 없는 듯!

-> Property crime은 특히 사회 경제와 관련 있는듯. 중요한 것은 고용률 좋음->경제 좋음->사람들의 property crime 줄어드는 것이 아니라, 오히려 사람들이 밖에 많이 돌아다니고, 경제력을 갖춰 범죄자들에게 attractive해짐!

경제가 안 좋다고 property 범죄가 늘어나는 것은 아님.



소득

Property crime

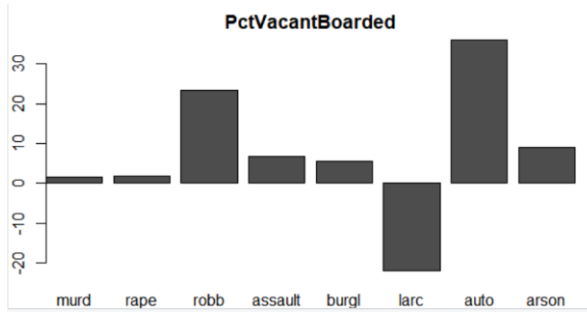
-**pctWretire:** 은퇴소득이 있는 가구의 비율 -> 범죄율 낮음(확실히 property crime범죄율 낮음)
-**pctWsocsec:** 사회보장소득이 있는 가구의 비율-> nonviolent crime범죄율이 높음-> 생계를 위한 범죄가 많음
-**NumStreet :** 범죄율 높음(특히 property crime에 큰 영향!)
-**whitepercap**: larceny, burglary 범죄율 높음-> 타겟?
-**pctVacantBoarded:** 파손 후 나무로 수리하고 비어있는 집
->larceny 제외하고 다른 범죄율 높음->target attractive-지역의 치안 관련 있는듯

Violent crime

마찬가지로
-**pctWsocsec:** arson을 제외한 나머지 범죄율 높음-> Federal poverty threshold보다 적은 수입을 가지고 있는 사람들 소득이 많은 사람들보다 violent범죄 피해율 2배 이상
-**NumInShelters, NumStreet**=매우매우 유의미한 변수. 이들이 많을수록 치안도 안 좋고 위험함
-**PctVacantBoarded:** violent crime 범죄율도 높음. 치안과 관련있는듯.
-> vacancy는 범죄를 증가하게 하는 요인 중 하나임. 실제로 비어있는 집이 많을수록 강력범죄율 증가함.

->Property crime에서 소득은 범죄와 직접적인 관련이 있어 보임. 소득이 낮을수록 property crime 범죄율 증가. 그러나 violent crime과 소득은 다른 방식으로 관련되어있는 것 같다. 소득은 치안을 나타내는 지표가 되어, 소득이 낮을수록 ->치안이 좋지 않음 -> 범죄율 증가 이런식인듯!

특히 larceny는 target-attractive한 곳에 많음



인종

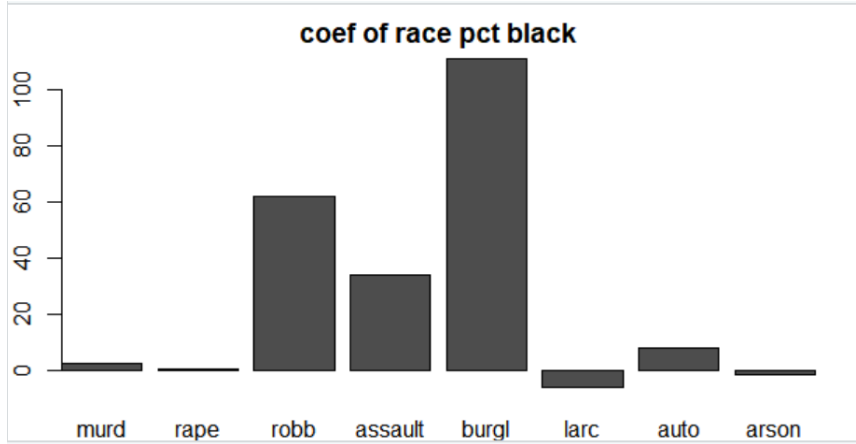
Property crime

-**racepctblack**: burglary, robbery에서 흑인의 범죄율 높음
-> 실제로 robbers 중 흑인이 50%를 차지함
-> Larceny에서는 흑인의 범죄율이 낮았음.

Violent crime

-**racepctblack**: murder, assault에서 범죄율 높음.
->murder victime은 흑인의 비율이 높음.
-racePctwhite: violent crime에서 범죄율 대체로 낮음

->흑인의 범죄율은 특정 범죄에서만 높았음. Racepctwhite는 property crime에서는 범죄율이 높았으나, violent crime에서는 낮음.



이민

Property crime

-Numimmig: burglary**,** Auto theft범죄율 높음
-**Pctforeighborn**: 특히 property crime에서 범죄율 높음!

Violent crime

-Numimmig: violent crime 범죄율도 높음**.**

-> 실제로 Strong social organization, youth job opportunities and residential stability는 낮은 violent crime 범죄율과 관련된 neighborhood 특징임. 이민자들은 이런 특성이 부족하여 violent crime 범죄율 높게 나타나는듯. Property crime에서 높은 범죄율 나타나는 것은 생계를 위한 것 일거라고 추정.

나이

Property crime

-**Agepct65up: l**arceny, autotheft 에서 범죄율 높음->범죄의 대상

Violent crime

- **agePct12to29:** 대체적으로 rape를 제외한 나머지 모든 범죄 범죄율 낮았음. Rape만 범죄율이 높았던 이유는 피해자일듯

-> 실제로 나이가 많고 소득이 많은 사람보다 나이가 어리고 소득이 적은 사람이 범죄의 피해자가 되는 비율이 높음. 그러나 auto theft, larceny와 같은 property crime의 경우, 상대적으로 약한 노인이 범죄의 target이 되기 쉬움.

F1:	PersPerOccupHous
Model:	Bayes
Arsons Per Pop:	0.32
Assault Per Pop:	29.7
Auto Theft Per Pop:	-162
Burgl Per Pop:	356
Larc Per Pop:	77
Rapes Per Pop:	1.10
Robbb Per Pop:	-152

F1:	PersPerOccupHous
Model:	Freq
Arsons Per Pop:	13.94
Assault Per Pop:	57.3
Auto Theft Per Pop:	-201
Burgl Per Pop:	399
Larc Per Pop:	605
Rapes Per Pop:	0.00
Robbb Per Pop:	-109

가구

Property crime

-**persperoccuphous:** 가구 당 평균 인원
->auto theft, robbery에서는 낮았지만, burglary, larceny에서는 높았음.
-**persperOwnOccuhous:** owner occupied household 당 평균 거주자
-> auto theft, robbery, larceny 범죄율 높음
-**PctPersOwnOccup**:부유함
-> 모든 범죄 범죄율 낮음.

Violent crime

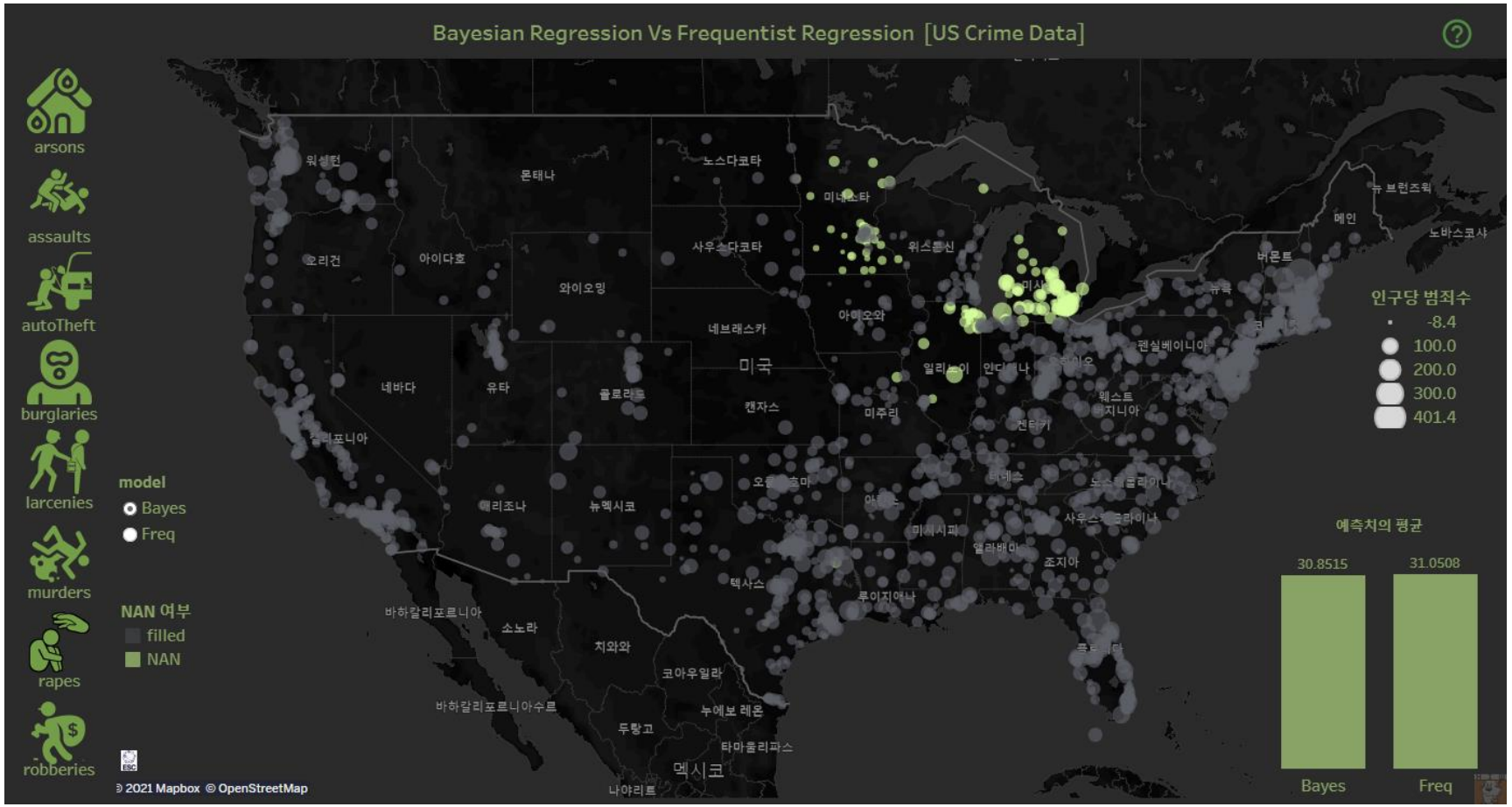
-persperoccuphous: 마찬가지로 violent crime에서도 범죄율 높음.
-**persperOwnOcchous**: assault, burglary 제외 범죄율 높음.
-**pctpersDenseHous**: 가난함
-> burglary, larcency제외 모든 범죄의 범죄율 높음.

->큰 family size는 일반적으로 열악한 역할 모델(부모, 형제), 부적절한 자녀 양육태도, 신체적인 경쟁(낮은 소득)과 심리적인 경쟁 (애정 부족)과 관련이 있음. -> 이는 가정환경과 연결되어 -> 높은 범죄율에 기여하는 것이 아닐까 추정

-> 청소년 범죄를 예방하고 치료하는데 있어서 가족의 역할이 중요하나, 이러한 역할 잘 수행하지 못할 것이라고 생각

+ **MalePctNeverMar**: 거의 모든 범죄에서 범죄율 높게 나타남. 범죄를 막아주는 가정이 없어서 그런 것이 아닐까..?

5. visualization



[미국 범죄 시각화 2 | Tableau Public]

감사합니다!