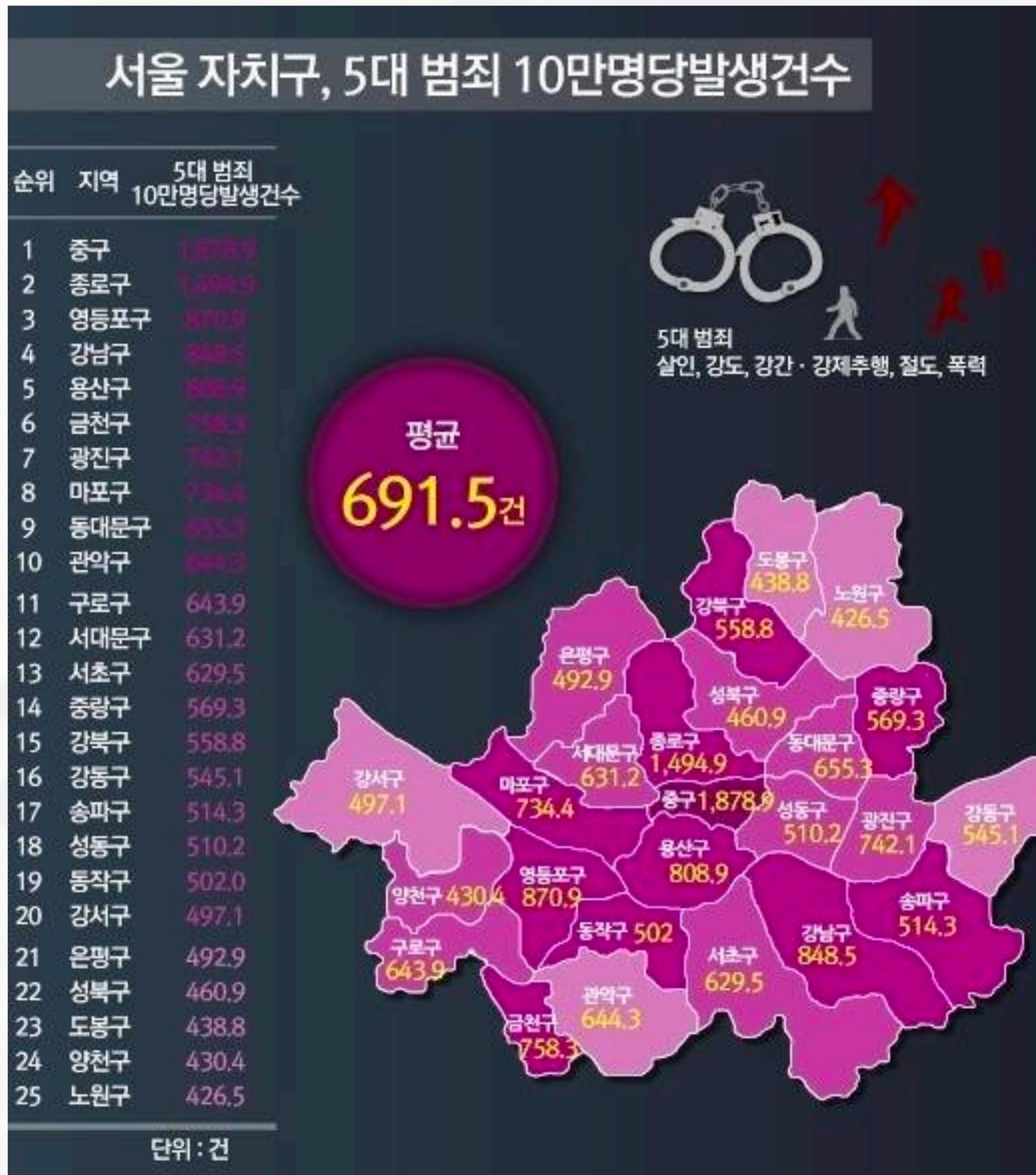


범죄 발생 수에 영향을 미치는 요인 분석

서울시 자치구별 범죄 발생 수에 영향을 미치는 요인들에 대한 연구 및 분석

조가영(ZHAO JIA YING)

연구배경



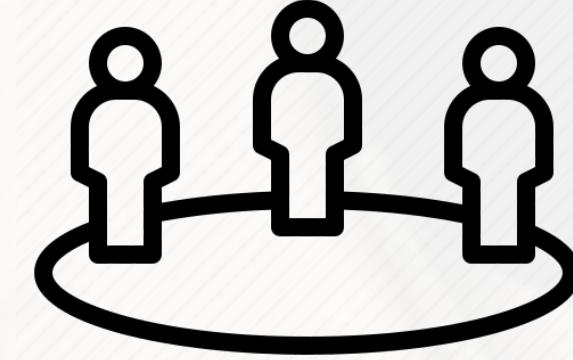
서울시에서 10만명당 범죄 발생 건수만
평균 691.5건이라고 한다.

하지만 자치구에 따라
범죄 발생 수가 차이나는 것을 볼 수 있다.

그렇다면,

서울시에서 범죄 발생 수가 많은 자치구들이
가지고 있는 공통적인 특징이 있을까?

가설설정



인구밀도가 높을 수록
범죄 발생 수가 많다

양(+)의 영향



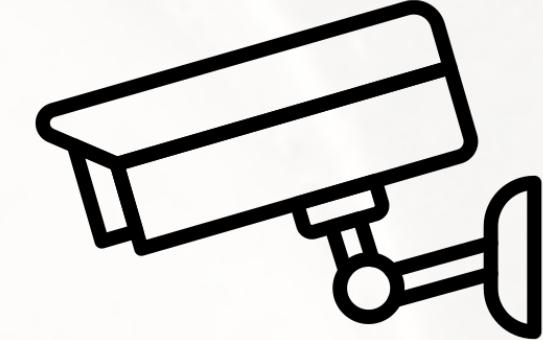
소득수준이 낮을 수록
범죄 발생 수가 많다

음(-)의 영향



실업률이 높을 수록
범죄 발생 수가 많다

양(+)의 영향



CCTV설치 수가 적을 수록
범죄 발생 수가 많다

음(-)의 영향

데이터 수집

서울시 자치구별 인구밀도 데이터셋 (2021년)

출처: 서울 열린데이터 광장

서울시 자치구별 소득 데이터셋 (2020년, 2021년 데이터 부재 / 추후 공개예정)

출처: 국세통계포털TASIS

서울시 자치구별 실업률 데이터셋 (2021년)

출처: 국가통계포털KOSIS

서울시 자치구별 CCTV설치 수 데이터셋 (2021년)

출처: 서울 열린데이터 광장

서울시 자치구별 범죄 발생 수 데이터셋 (2020년, 2021년 데이터 부재 / 추후 공개예정)

출처: 서울 열린데이터 광장

파일 불러오기

```
# 1) 서울 자치구별 인구밀도 데이터셋  
population_density <- read.csv('population_density_2021.csv', header=T, encoding = 'UTF-8')  
  
# 2) 서울 자치구별 소득 데이터셋  
income <- read.csv('income_2020.csv', header = T, encoding = 'UTF-8')  
  
# 3) 서울 자치구별 실업률 데이터셋  
unemployment_rate <- read.csv('unemployment_rate_2021.csv', header = T, encoding = 'UTF-8')  
  
# 4) 서울 자치구별 CCTV설치 수 데이터셋  
cctv <- read.csv('cctv_total_2021.csv', header = T, encoding = 'UTF-8')  
  
# 5) 서울 자치구별 범죄 발생 수 데이터셋  
crime <- read.csv('crime_total_2020.csv', header = T, encoding = 'UTF-8')
```

데이터 전처리 - 데이터 유형 변환

```
# 단계 1: 데이터 유형 변환
population_density$population.density <- gsub('[,]', '', population_density$population.density) # 문자열에 ',', 제거
population_density$population.density <- as.numeric(population_density$population.density) # 문자형 -> 숫자형

income$Income <- as.numeric(gsub('[,]', '', income$Income)) # 문자형 -> 숫자형
income$region <- gsub(' ', '', income$region) # region칼럼에 있는 공백을 없애기

cctv$cctv_total <- as.numeric(gsub('[,]', '', cctv$cctv_total)) # 문자형 -> 숫자형

crime$crime_total <- as.numeric(gsub('[,]', '', crime$crime_total)) # 문자형 -> 숫자형
```

- 인구밀도, 소득, CCTV설치수, 범죄 발생 수에 대한 칼럼들이 모두 문자형으로 구성되어 있으므로, 문자형 값에 있어서 “,” 를 먼저 제거하고 숫자형으로 변환하였습니다.
- 소득 데이터셋의 region칼럼의 경우 일부 값들에 공백(whitespace)이 발견되어 공백을 제거하는 전처리과정을 거쳤습니다.

데이터 전처리 - 데이터프레임 조인

```
# 단계 2: 데이터프레임 조인
```

```
library(dplyr)
```

```
join1 <- left_join(population_density, income, by = 'region')
```

```
join1
```

```
join2 <- left_join(join1, unemployment_rate, by = 'region')
```

```
join2
```

```
join3 <- left_join(join2, cctv, by = 'region')
```

```
join3
```

```
crime_data <- left_join(join3, crime, by = 'region')
```

```
# 'data.frame': 25 obs. of 6 variables:
```

```
# $ region : chr "종로구" "중구" "용산구" "성동구" ...
```

```
# $ population.density: num 6431 13231 10852 17359 20666 ...
```

```
# $ Income : num 20367393 25633311 19471646 8755085 2899545 ...
```

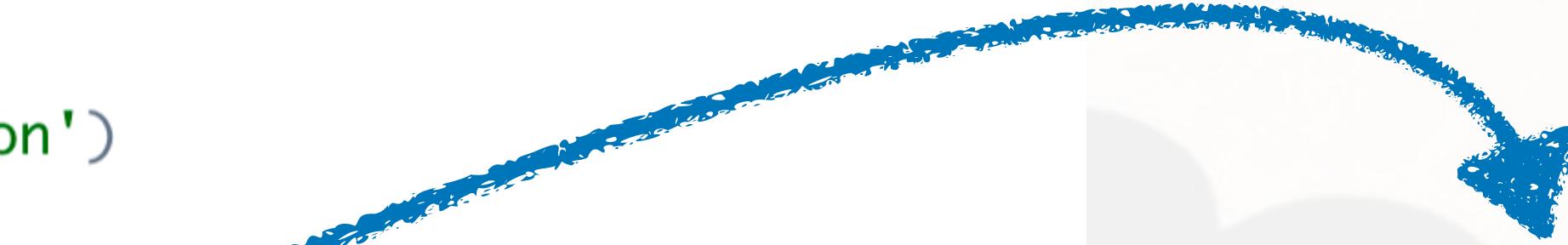
```
# $ unemployment.rate : num 4 4.3 4.8 5 3.8 4.7 5.1 4.6 5.6 5.9 ...
```

```
# $ cctv_total : num 1715 2447 2611 3829 3211 ...
```

```
# $ crime_total : num 3102 3411 2969 2362 3601 ...
```

```
# 독립변수 : population.density, Income, unemployment.rate, cctv_total
```

```
# 종속변수 : crime_total
```



crime_data로 요인 분석 진행

데이터 전처리 - 결측치 및 이상치 확인

```
# 단계 3: 결측치와 이상치 확인
summary(crime_data) # 변수 통계량 확인 -> 결측치 없음

boxplot(crime_data$population.density) # 이상치 확인 -> outlier 없음

boxplot(crime_data$Income) # 이상치 있음
boxplot(crime_data$Income)$stats # 하한값(1196591) ~ 상한값(29073375)
subset(crime_data, Income > 29073375) # 이상치 확인 -> 강남구

boxplot(crime_data$unemployment.rate) # 이상치 확인 -> outlier 없음

boxplot(crime_data$cctv_total) # 이상치 있음
boxplot(crime_data$cctv_total)$stats # 하한값(1715) ~ 상한값(5149)
subset(crime_data, cctv_total > 5149) # 이상치 확인 -> 강남구

boxplot(crime_data$crime_total) # 이상치 있음
boxplot(crime_data$crime_total)$stats # 하한값(2179) ~ 상한값(5410)
subset(crime_data, crime_total > 5410) # 이상치 확인 -> 강남구
```

강남구의
소득, CCTV 설치 수, 범죄 발생 수가
다른 자치구보다 압도적으로 많음

데이터 전처리 - 종속변수 기준으로 독립변수 탐색

```
# 단계 4: 종속변수 기준으로 독립변수 탐색
```

```
library(lattice)
```

```
xyplot(crime_total ~ population.density, data = crime_data)
```

```
# [해석] 인구밀도는 범죄 발생 수와 상관성이 거의 없는 것으로 나타남
```

```
xyplot(crime_total ~ Income, data = crime_data)
```

```
# [해석] 자치구별 소득이 증가함에 따라 범죄 발생 수도 증가하는 경향이 보임
```

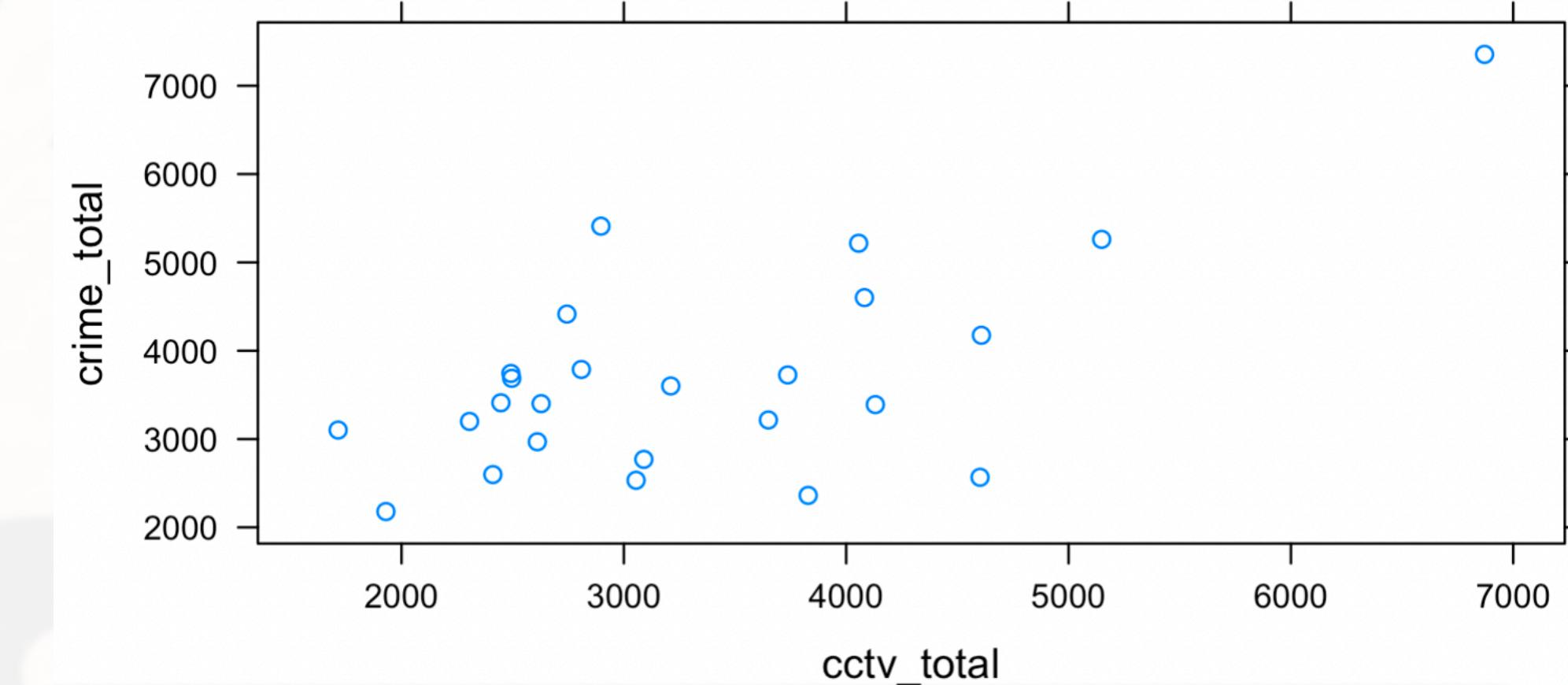
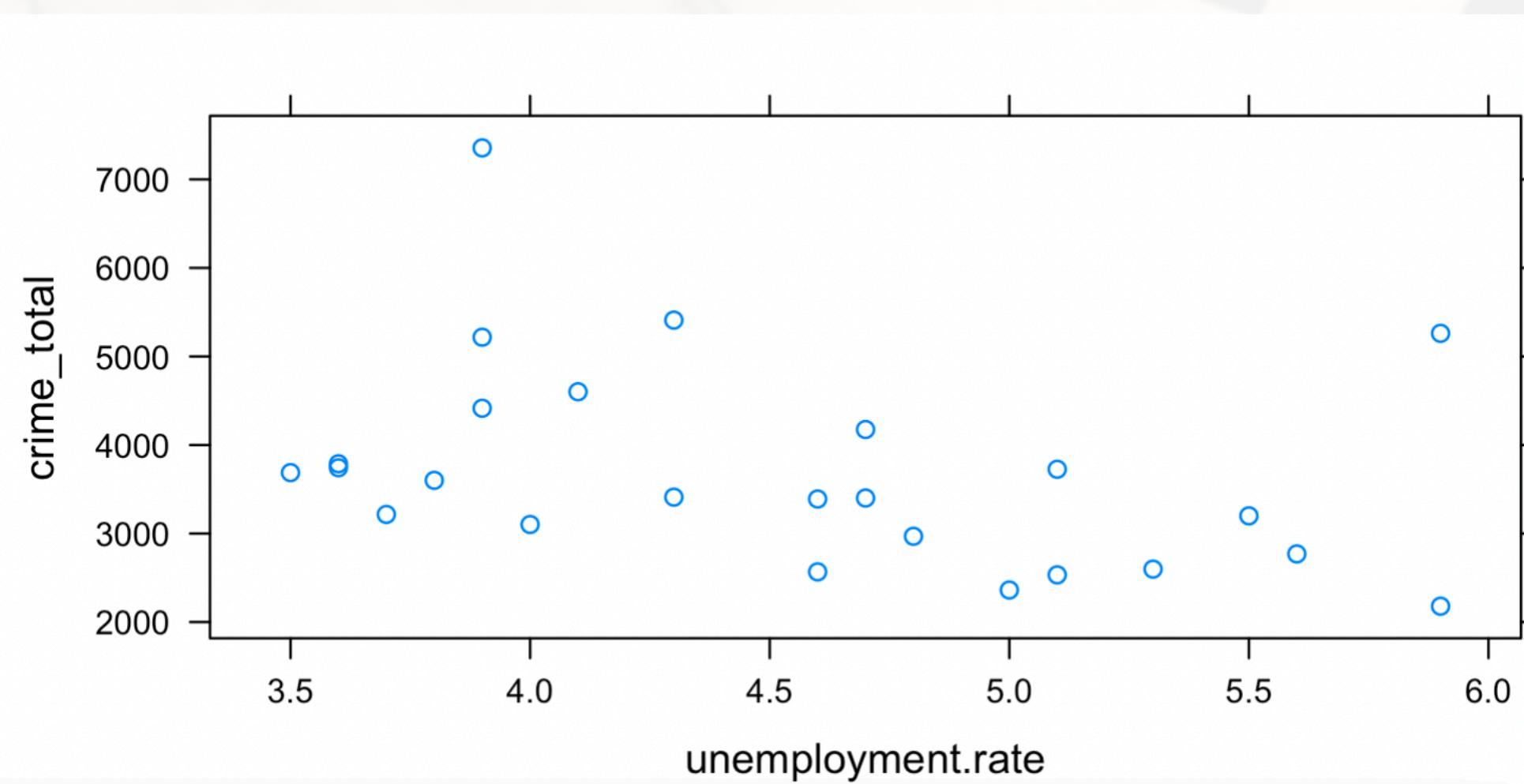
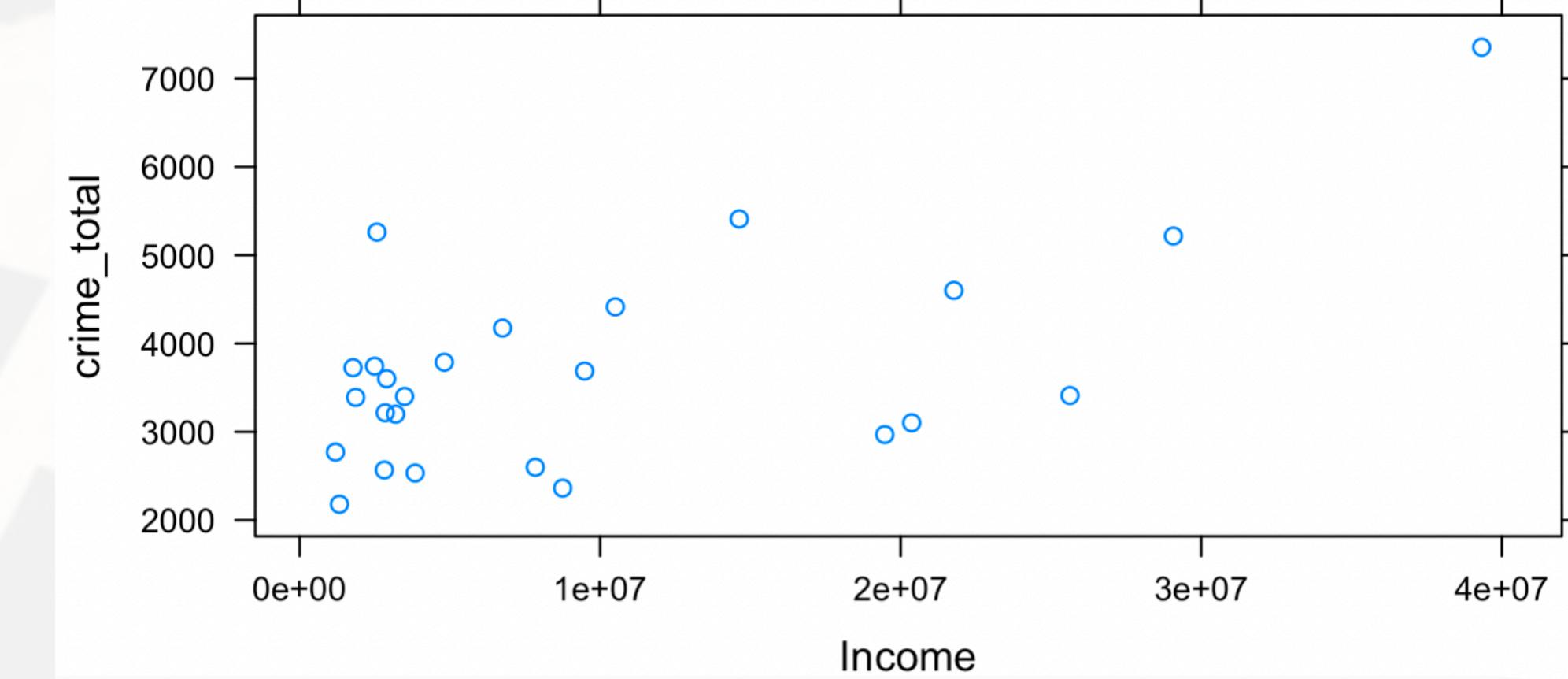
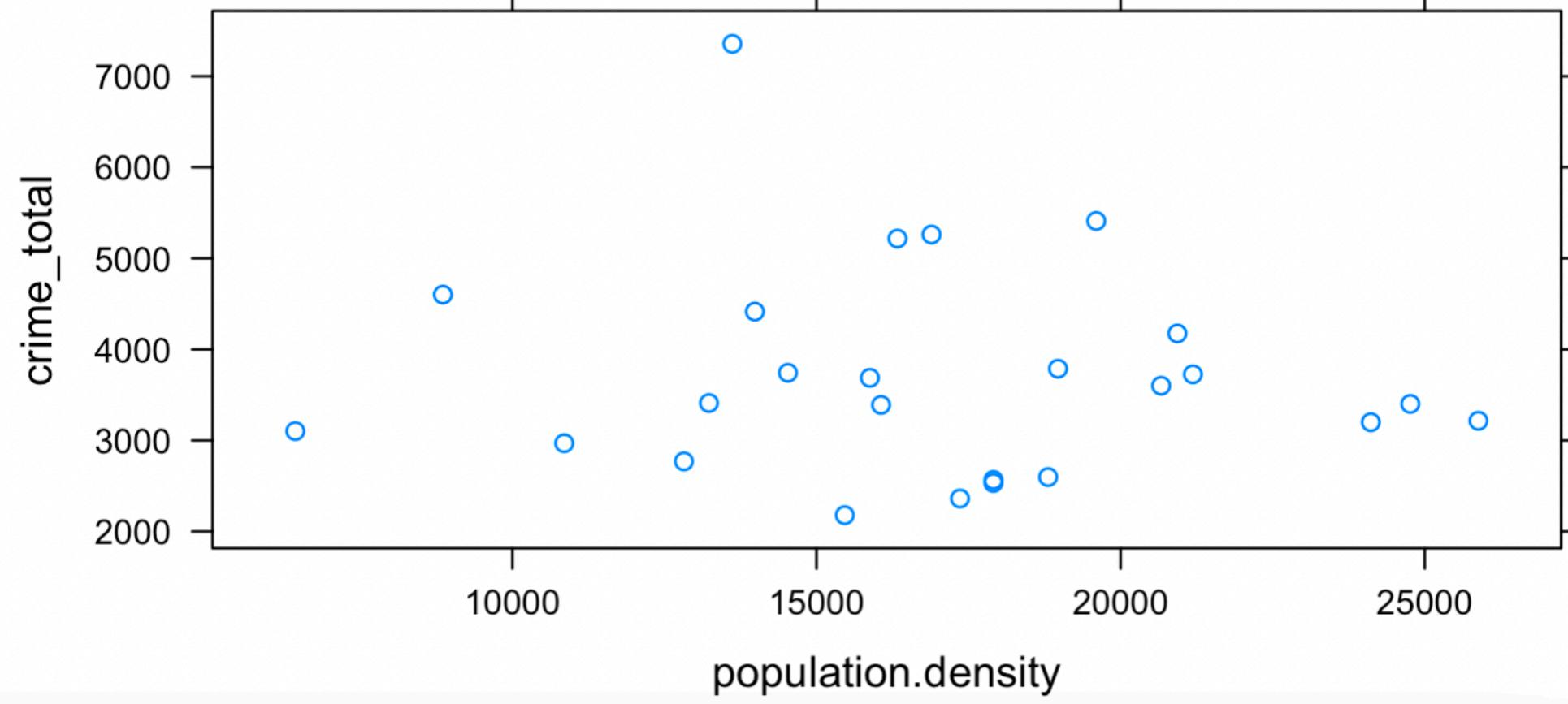
```
xyplot(crime_total ~ unemployment.rate, data = crime_data)
```

```
# [해석] 실업률이 높은 지역일수록 범죄 발생 수가 오히려 적은 경향이 보임
```

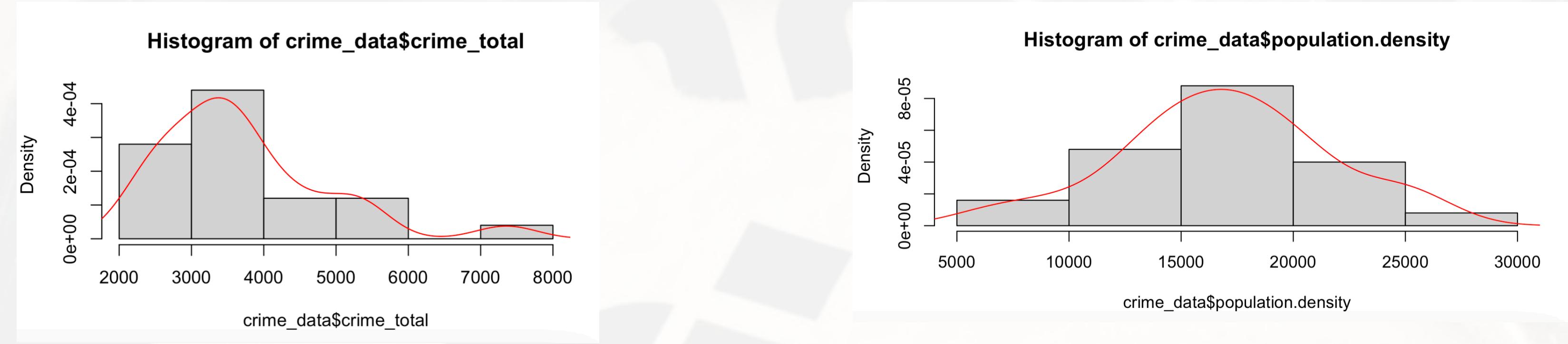
```
xyplot(crime_total ~ cctv_total, data = crime_data)
```

```
# [해석] CCTV 설치 수가 증가함에 따라 범죄 발생 수도 증가하는 경향이 보임
```

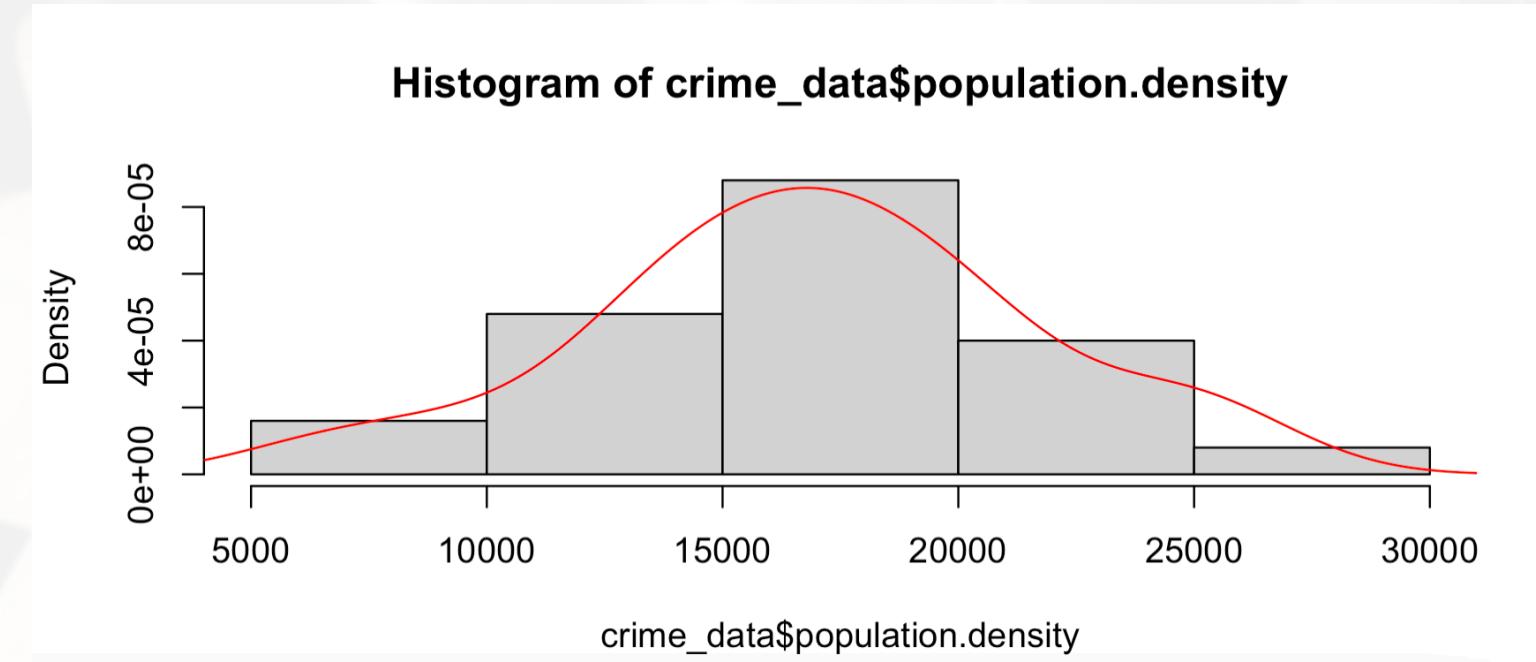
데이터 전처리 - 종속변수 기준으로 독립변수 탐색에 대한 시각화



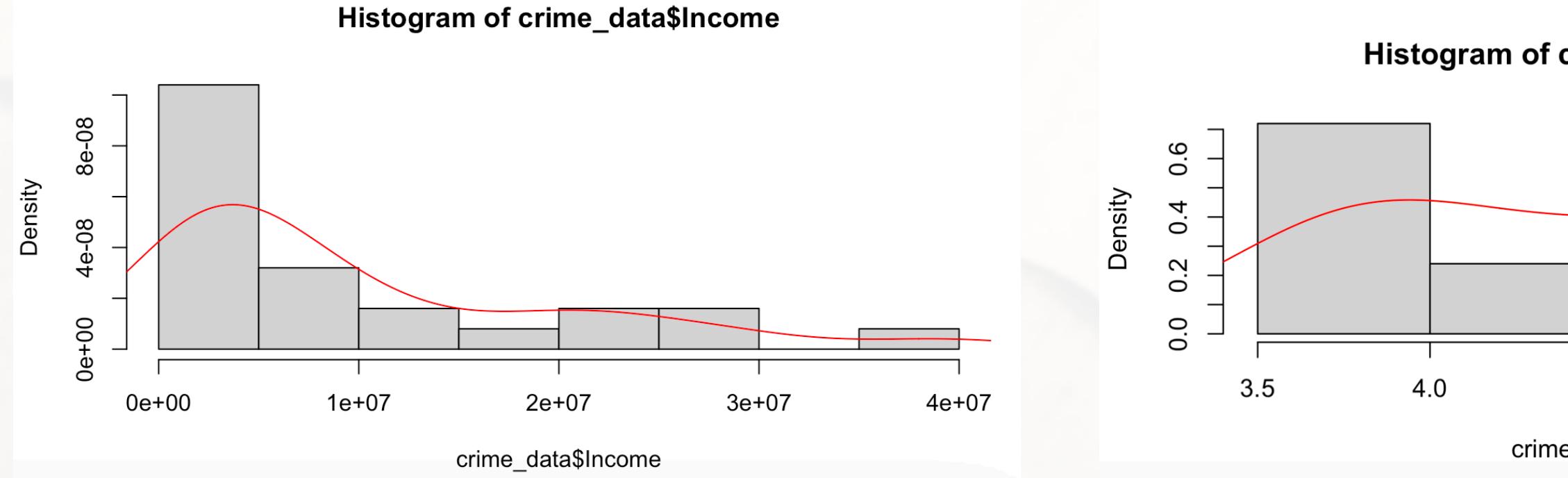
데이터 분석 - 각 변수에 대한 통계량 및 분포 확인



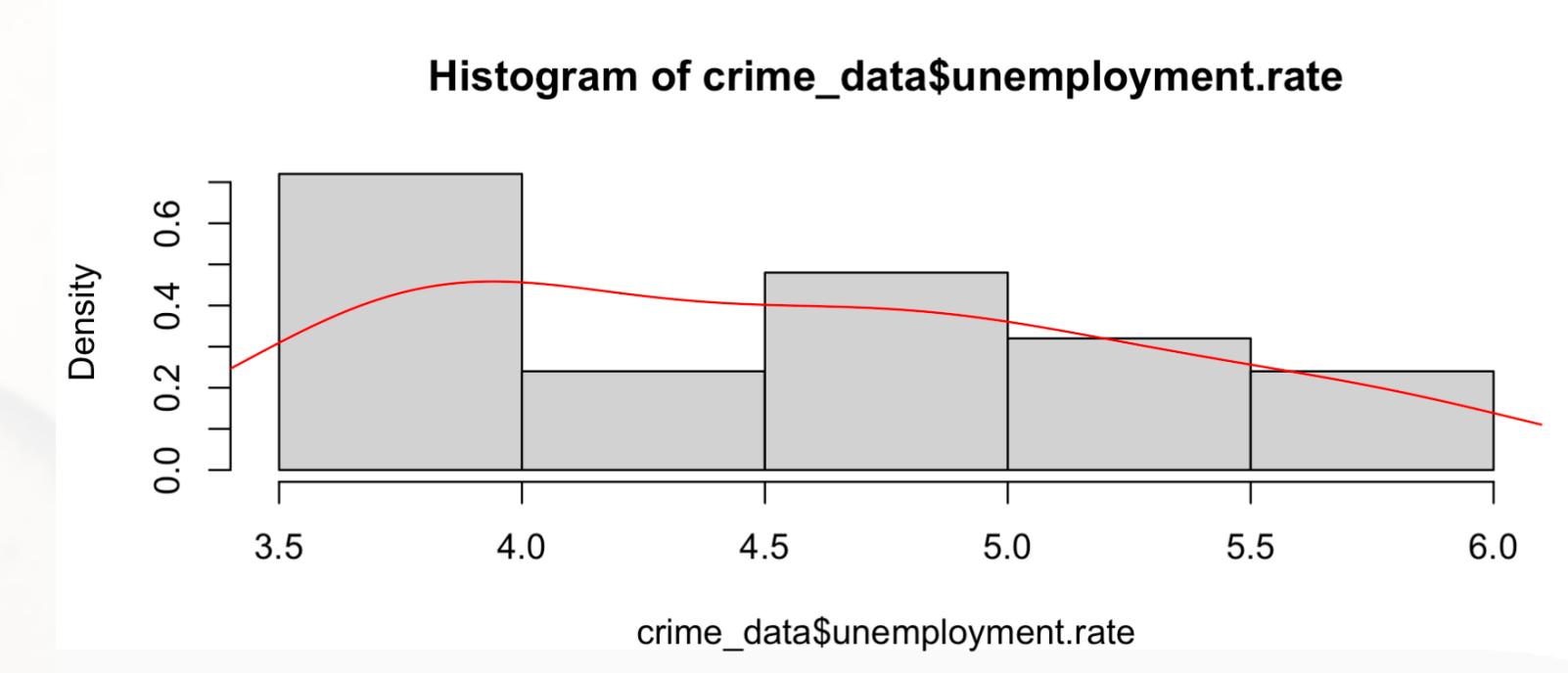
```
# (1) crime_total(범죄 발생 수)
summary(crime_data$crime_total)
# Min. 1st Qu. Median Mean 3rd Qu. Max.
# 2179 2969 3411 3707 4175 7356
hist(crime_data$crime_total, freq = FALSE) # 원쪽으로 치우친 형태
lines(density(crime_data$crime_total), col = 'red')
skewness(crime_data$crime_total) # 왜도 = 1.304917 > 0 (오른쪽꼬리분포)
kurtosis(crime_data$crime_total) # 첨도 = 4.865062 > 3 (정규분포 보다 뾰족한 형태)
```



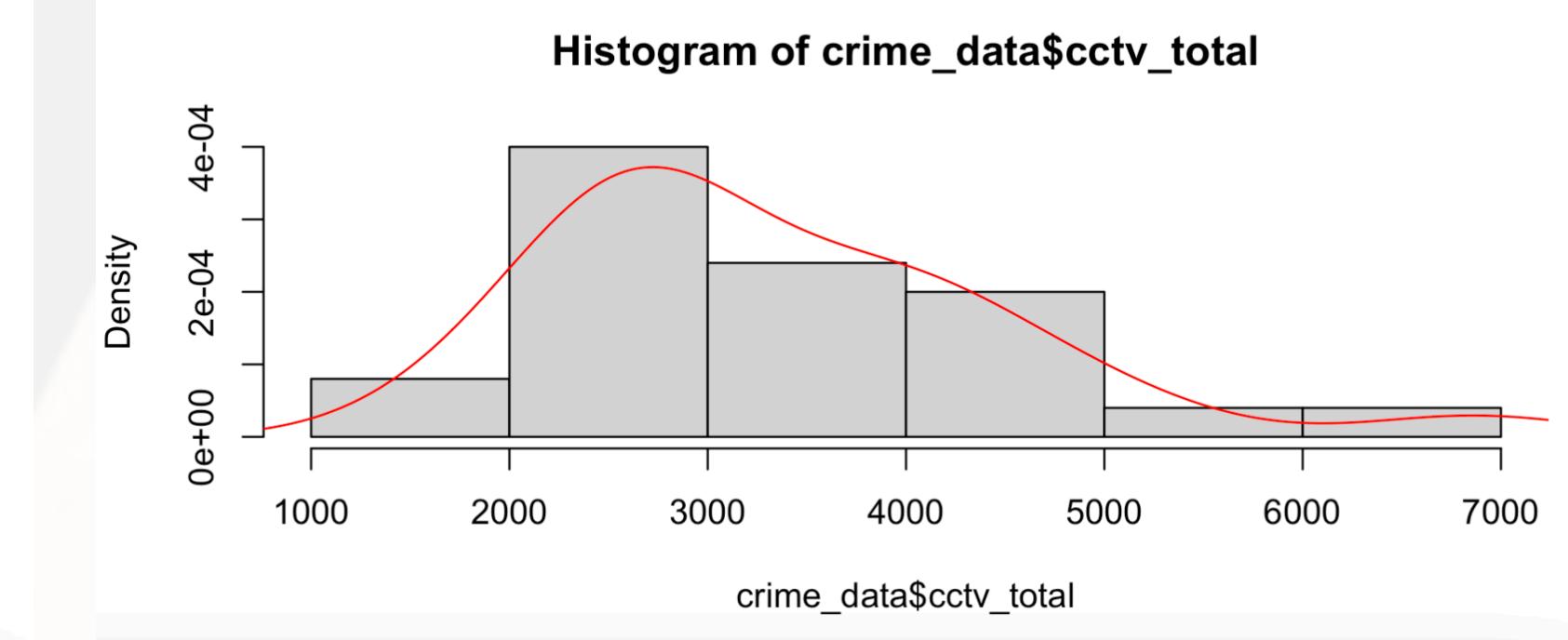
```
# (2) population.density(인구 밀도)
summary(crime_data$population.density)
# Min. 1st Qu. Median Mean 3rd Qu. Max.
# 6431 13986 16891 16922 19599 25882
hist(crime_data$population.density, freq = FALSE)
lines(density(crime_data$population.density), col = 'red')
skewness(crime_data$population.density) # 왜도 = -0.1217695 < 0 (왼쪽꼬리분포)
kurtosis(crime_data$population.density) # 첨도 = 2.854591 < 3 (정규분포 보다 왼만한 형태)
```



```
# (3) Income(소득)
summary(crime_data$Income)
# Min. 1st Qu. Median Mean 3rd Qu. Max.
# 1196591 2823944 4818585 9952570 14630240 39330872
hist(crime_data$Income, freq = FALSE)
lines(density(crime_data$Income), col = 'red')
skewness(crime_data$Income) # 왜도 = 1.349493 > 0 (오른쪽꼬리분포)
kurtosis(crime_data$Income) # 첨도 = 3.926684 > 3 (정규분포 보다 뾰족한 형태)
```



```
# (4) unemployment.rate(실업률)
summary(crime_data$unemployment.rate)
# Min. 1st Qu. Median Mean 3rd Qu. Max.
# 3.500 3.900 4.600 4.536 5.100 5.900
hist(crime_data$unemployment.rate, freq = FALSE)
lines(density(crime_data$unemployment.rate), col = 'red')
skewness(crime_data$unemployment.rate) # 왜도 = 0.3473527 > 0 (오른쪽꼬리분포)
kurtosis(crime_data$unemployment.rate) # 첨도 = 1.967397 < 3 (정규분포 보다 왼만한 형태)
```



```
# (5) cctv_total(CCTV 설치 수)
summary(crime_data$cctv_total)
# Min. 1st Qu. Median Mean 3rd Qu. Max.
# 1715 2496 3055 3342 4056 6871
hist(crime_data$cctv_total, freq = FALSE)
lines(density(crime_data$cctv_total), col = 'red')
skewness(crime_data$cctv_total) # 왜도 = 1.193361 > 0 (오른쪽꼬리분포)
kurtosis(crime_data$cctv_total) # 첨도 = 4.586298 > 3 (정규분포 보다 뾰족한 형태)
```

데이터 분석 - 상관관계분석

2) 상 관 관 계 분 석 & 시 각 화

```
cor <- cor(crime_data[-1]) # 상 관 계 수 보기  
cor['crime_total',]  
# population.density    Income      unemployment.rate      cctv_total      crime_total  
# -0.09975697           0.60665756   -0.36183255        0.63388897     1.00000000
```

종속변수와 독립변수 간의 상관관계 분석 결과:

인구밀도(population.density)는 약 -0.1로 종속변수(crime_total)와 거의 상관관계가 없는 것으로 보임

실업률(unemployment.rate)은 약 -0.36로 종속변수와 낮은 상관관계로 보이며, 소득(Income)과 CCTV설치 수(cctv_total)는 종속 변수와 다소 높은 양(+)의 상관관계로 보임

데이터 분석 - 회귀분석

3) 회귀분석

```
crime_data_new <- crime_data[-1]
model_lm <- lm(crime_total ~ ., data = crime_data_new)
summary(model_lm)

Call:
lm(formula = crime_total ~ ., data = crime_data_new)
```

Residuals:

Min	1Q	Median	3Q	Max
-1398.57	-381.99	62.67	273.12	1513.30

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.308e+03	1.436e+03	1.607	0.12366
population.density	3.459e-02	4.123e-02	0.839	0.41144
Income	5.237e-05	2.136e-05	2.452	0.02352 *
unemployment.rate	-2.961e+02	2.344e+02	-1.263	0.22095
cctv_total	4.895e-01	1.504e-01	3.254	0.00398 **

Signif. codes:	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1			

Residual standard error: 774 on 20 degrees of freedom

Multiple R-squared: 0.6377, Adjusted R-squared: 0.5652

F-statistic: 8.8 on 4 and 20 DF, p-value: 0.0002875

[해석]

- 인구밀도(population.density)와 실업률(unemployment.rate)은 유의하지 않은 변수로 나타남

- 소득(Income)과 CCTV 설치 수(cctv_total)는 유의한 변수로 나타나며, 모두 종속변수에 양(+)의 영향을 끼침

- p-value: 0.0002875 < 0.05이므로 통계적으로 유의함

- Adjusted R-squared: 0.5652로 해당 회귀모델이 대략 57%의 설명력을 가짐

데이터 분석 - 독립변수의 다중공선성 문제 확인 & 유의성 검정 확인

```
# 독립변수 간 다중공선성 확인
library(car)
sqrt(vif(model_lm)) > 2
# population.density      Income      unemployment.rate      cctv_total
# FALSE                   FALSE       FALSE                  FALSE
# [해석] 독립변수 간 다중공선성 확인 결과 문제 없음

# 효과적인 변수선택법으로 독립변수 유의성 검정 확인
library(MASS)
step <- stepAIC(model_lm, direction = 'both')
# Start: AIC=337
# crime_total ~ population.density + Income + unemployment.rate +
#   cctv_total
#
# Df Sum of Sq    RSS    AIC
# - population.density 1 421601 12403176 335.86
# - unemployment.rate  1 956339 12937914 336.92
# <none>                      11981575 337.00
# - Income            1 3601917 15583492 341.57
# - cctv_total        1 6342725 18324300 345.62
#
# Step: AIC=335.86
# crime_total ~ Income + unemployment.rate + cctv_total
#
# Df Sum of Sq    RSS    AIC
# <none>                      12403176 335.86
# - unemployment.rate  1 1154525 13557701 336.09
# + population.density 1 421601 11981575 337.00
# - Income            1 3454211 15857386 340.01
# - cctv_total        1 7875436 20278612 346.16
```

다중공선성 문제 확인 결과:

다중공선성 문제가 없음

변수선택법을 통한 유의성 검정 확인:

변수선택법을 통해서 확인한 결과 인구밀도(population.density) 변수를 제외했을 경우 AIC값이 더 낮게 나타나는 것이 확인됨

따라서, 아래와 같이 인구밀도(population.density)변수를 제외하고 회귀모델을 재생성

데이터 분석 - 회귀모델 재생성

```
new_model_lm <- lm(crime_total ~ Income + unemployment.rate + cctv_total, data = crime_data_new)
summary(new_model_lm)
# Coefficients:
#                                     Estimate Std. Error t value Pr(>|t|)
# (Intercept)                2.998e+03  1.169e+03   2.565  0.01805 *
# Income                     4.218e-05  1.744e-05   2.418  0.02476 *
# unemployment.rate        -3.225e+02  2.306e+02  -1.398  0.17667
# cctv_total                  5.242e-01  1.436e-01   3.652  0.00149 **
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Residual standard error: 768.5 on 21 degrees of freedom
# Multiple R-squared:  0.6249, Adjusted R-squared:  0.5714
# F-statistic: 11.66 on 3 and 21 DF,  p-value: 0.0001035
```

[해석]

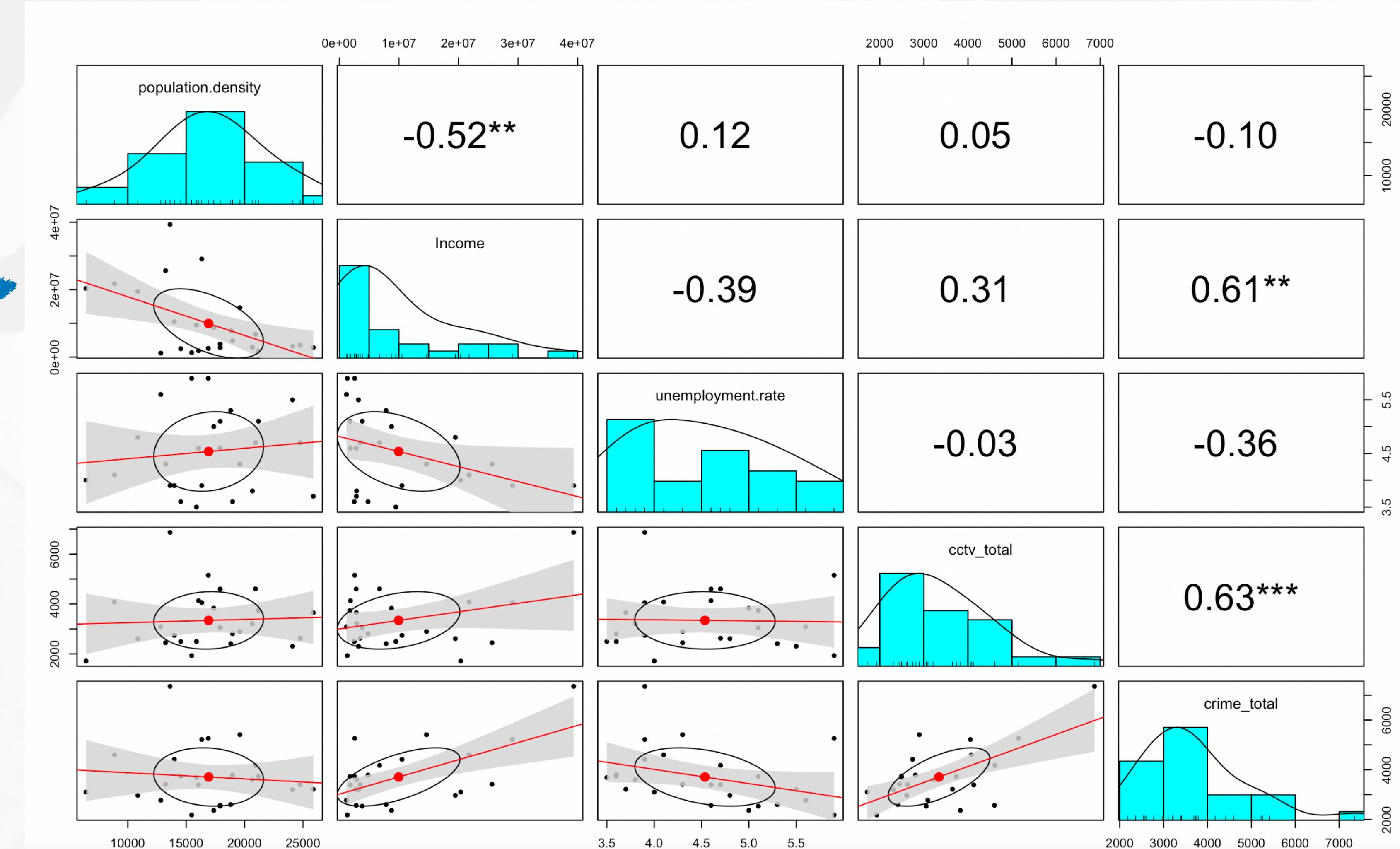
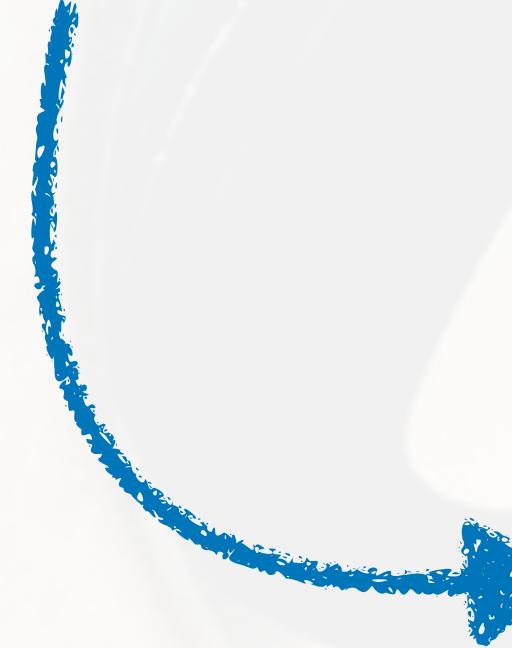
- 실업률(unemployment.rate)은 여전히 유의하지 않은 변수로 나타남
- 소득(Income)과 CCTV설치 수(cctv_total)는 유의한 변수이며, 소득보다 CCTV설치 수의 영향력이 더 크게 나타남
- Adjusted R-squared는 0.5714로 재생성된 모델의 설명력은 기존 모델과 거의 비슷하게 나타남

데이터 분석 - 다중선형회귀분석 회귀선 시각화

다중선형회귀분석 회귀선 시각화

library(psych)

pairs.panels(crime_data_new, stars = TRUE, lm = TRUE, ci = TRUE)



데이터 분석 - 회귀분석 모형진단

회귀분석 모형 진단

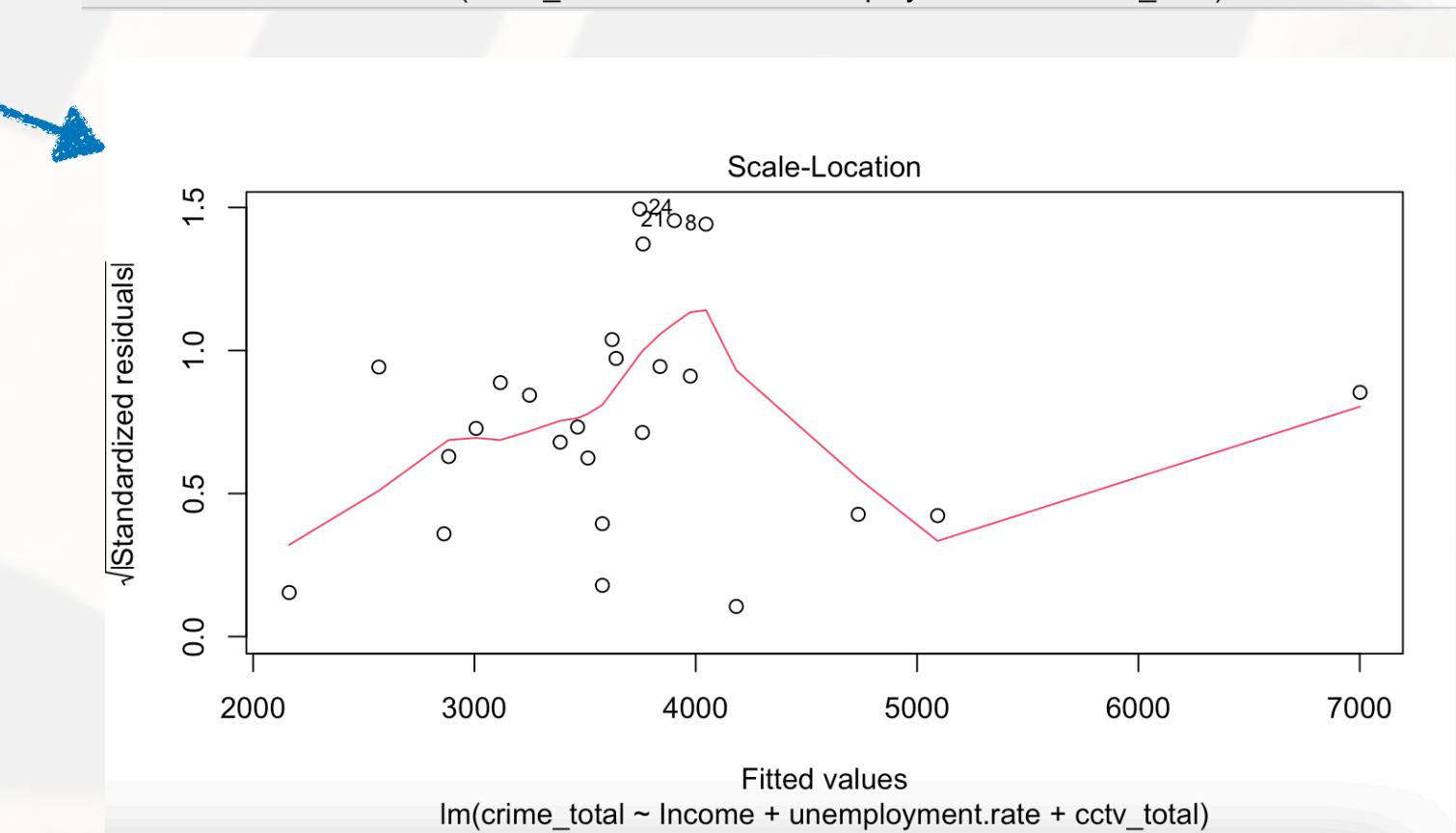
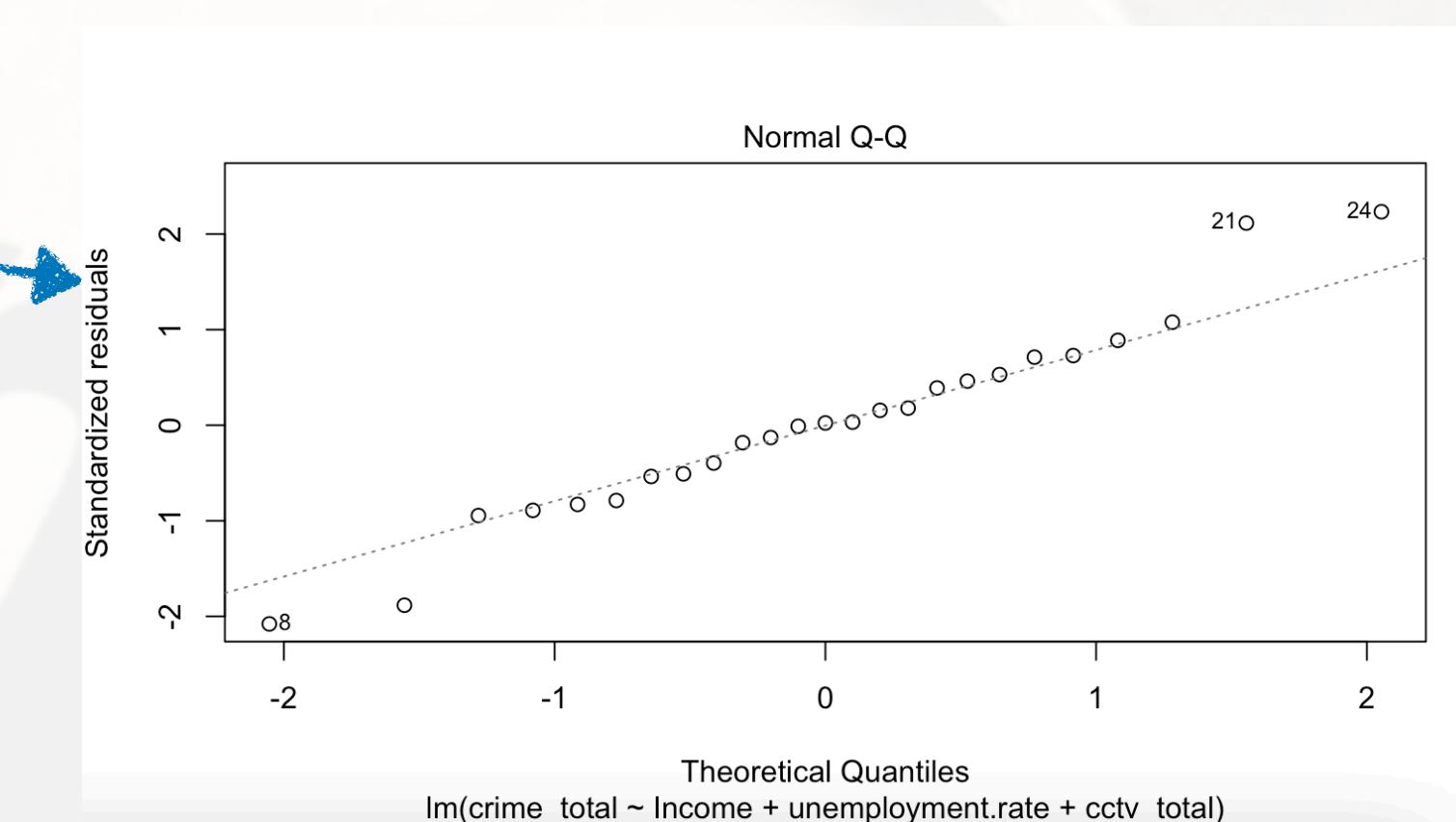
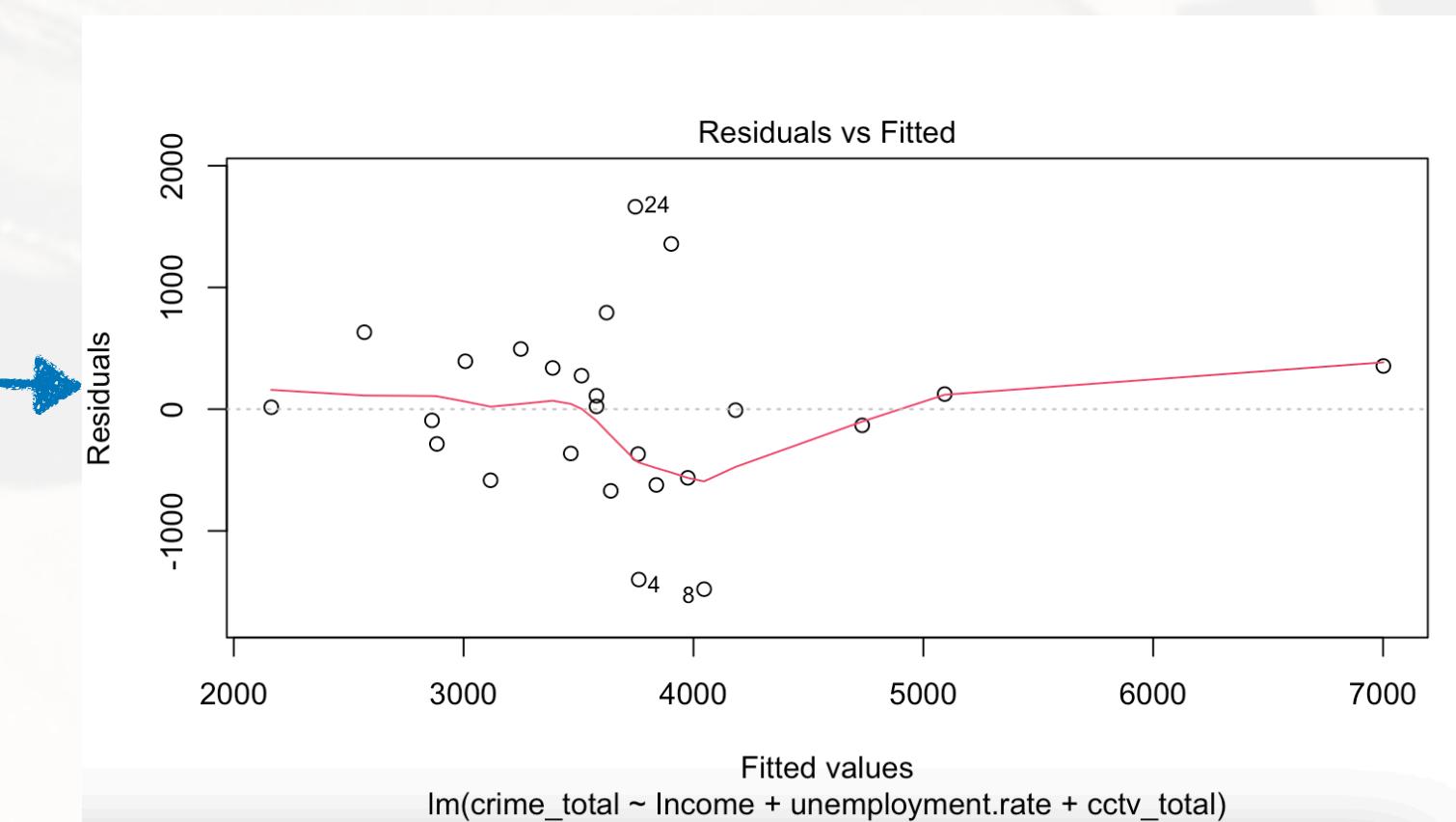
1) 잔차 선형성 검정
plot(new_model_lm, which = 1)

2) 잔차 정규성 검정
plot(new_model_lm, which = 2)

shapiro.test(new_model_lm\$residuals) # p-value = 0.6626 > 0.05 (정규성과 차이 없음)

3) 잔차의 등분산성 검정
plot(new_model_lm, which = 3)

4) 잔차의 독립성
library(lmtest)
dwtest(new_model_lm) # p-value = 0.1121 > 0.05 (독립성 만족)



결론



자치구별 **인구밀도**는
범죄 발생수와 상관성이
거의 없는 것으로 보인다.



자치구별 **소득**은
높은 지역일 수록
범죄 발생 수가
많은 경향이 보인다.



실업률은 범죄 발생 수와
낮은 상관관계로 보이며,
범죄 발생 수에
큰 영향을 미치지 않는다.



CCTV설치 수에 따라 범
죄 발생 수가 증가하는
경향이 보인다.

회고 및 고찰

이번 데이터 분석을 통해서 우리의 선입견과 고정관념이 현실과 다르다는 사실을 깨달을 수 있었습니다.

하지만, 데이터 자체가 한정적이여서 정확한 결론이라고 보기에는 힘들다고 생각합니다.

만약, 서울시 뿐만 아니라 다른 시도의 데이터셋까지 확보해서 분석한다면 더 정확한 분석결과가 나오지 않을까 생각합니다.