Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 45 No. 3
March 2024

An Analysis of Social Variables for Preventing Big 5 Crimes in Seoul

Seung-Yeon Hwang [1], Jeong-Joon Kim [2*]

[1] Dept. of Computer Engineering, Anyang University, Anyang-si, Gyeonggi-do, Republic of Korea
[2] Corresponding Author, Dept. of Software, Anyang University, Anyang-si, Gyeonggi-do, Republic of Korea

Abstract

It is common to see that most social activities occur intensively in Seoul in Korea. As most of you know, Seoul is the center stage of Korean issues, so the population density and social issues, which are much higher than other regions, are very influential. Therefore, based on this social context, it can be fully inferred that the assumption that the Seoul Metropolitan Government's crime rate and criminal sensitivity will also be very high will be reasonable. Therefore, this paper analyzes various social variables that may affect the five major crime rates in Seoul using R, a tool for big data analysis, using simple regression analysis that correlates with the total crime rate and multiple regression analysis that correlates each social variable. In addition, we would like to present policies from various perspectives to prevent crime in Seoul by checking the correlation.

Keywords: Big Data, R, Multiple regression, Crime, Preventing Crimes.

1. Introduction

Among the various social problems that exist in recent years, centralism in the Seoul metropolitan area, called the "Seoul Republic" has emerged. 'centrism' in a particular region refers to a phenomenon in which social issues are centered around a particular region where more active social activities occur within a particular space with a large unit of culture that represents the identity of a particular nation. In other words, the "specific region" mentioned in the above sentence will be Seoul, the capital area of Korea.

The background of this phenomenon is as follows. As Korea began implementing economic development policies in the 1960s and 1970s that aim for full-fledged growth and prioritize technological development, the gap between development demand and underprivileged areas naturally arose. As a result, large-scale population movements have continued to occur in metropolitan areas and provincial cities, especially in Seoul and nearby areas, resulting in unbalanced development centered on Seoul.

In response, the government is implementing several policies to address the concentration of the metropolitan area, but the effect is insignificant and it does not help to improve the problem. The most representative case is policies to relocate public institutions to various non-metropolitan cities, such as Sejong City, but it has not moved public officials voluntarily, and only encouraged the competition rate for civil servants exam in Seoul. In addition, various policies such as policies to regulate development in the metropolitan area and policies to develop innovative cities in the non-metropolitan area are being implemented, but there are no dramatic effects.

Derived from this centralism in the Seoul metropolitan area, various problems arise, and among them, there is a "crime" problem that is frequently issued and dealt with profoundly by the media, causing the most socially sensitive response. As mentioned earlier, Seoul has a more diverse and dangerous impact on a wider social network, even if it is the same crime, as it has a much greater impact on social issues than other regions, causing fears among Seoul citizens [1]. Also, because of the high population density, crimes occur relatively more frequently.

Therefore, in order to come up with specific solutions to prevent the five major crimes in the Seoul Metropolitan Government [2], research was conducted to analyze various social factors that affect the crime rate. The reason for dealing with the five major crimes among crimes is that they are heavier, more dangerous than other minor crimes and have the potential to cause a great stir in society, so they should be prevented more thoroughly.

In this paper, we investigate various social variables that seem to be related to the five major crimes,

analyze the direct correlation with the five major crime rates, and select the variables that indirectly affect the five major crime rates. The goal is to lower Seoul's five major crime rates by devising crime prevention policies in areas that include the relevant variables and presenting various solutions.

The social variables to investigate are largely divided into social variables related to social activities and urban variables related to spatial and geographical characteristics. First of all, the former social variables include property tax indicating economic indicators, resident population indicating the size of the population that is always living in an area, population density showing the number of people relative to area, floating population indicating the degree of urban revitalization, percentage of youth population to total population, percentage of highly educated population representing the number of college graduates to the total population, number of pubs that are indicators of entertainment and the number of police officers indicating the degree of security. The latter urban variables include housing supply rate that can determine whether it is a residential area and the number of accommodations that can be an indirect indicator of the inflow rate of outsiders who are not regular residents. Data quantifying these social variables were collected from the Seoul Public Data Portal, the National Police Agency, and the National Statistical Portal and analyzed how each factor was correlated through simple regression and multiple regression analyses. And we would like to present policies for crime prevention.

2. Related Technologies

1. Big Data

Currently, the amount of data is doubling in zettabytes per year. As such, big data includes not only numbers but also technologies that store large amounts of data to include letters and images, and analyze them with various big data tools to extract meaningful content [3]. Steps to deal with big data can be largely divided into stages of collection, storage, processing, analysis, and visualization.

2. Distributed File System

Distributed file system is a network connection between physically different computers, allowing users to use multiple file systems as a single file system. Furthermore, with the development and active spread of network technology, several small computers are networked to each other, making it easy to share information between users and efficiently use storage space without time and place constraints.

3. Map Reduce

MapReduce is a software framework for parallel/distributed processing of big data, released in 2004 and a prime example is Hadoop [4]. It is now widely used by companies dealing with big data such as Amazon, Yahoo, and Facebook. The Map phase is a stage in which scattered unstructured data is grouped among related data and transformed into a key/value pair in a structured form. In the Reduce phase, it is a task of extracting useful data, such as removing redundant data, by performing parallel processing on a set key/value pair created in the Map phase. By default, Map Reduce is effective in dealing with big data with tasks such as clustering by performing batch-based processing [5].

4. Statistical Language R

R is a program originally built on S language and used as a statistical tool, but it is easier to use than languages such as C and JAVA and is open source, making it easier to use packages created by other users. In addition, the latest data mining techniques can be implemented through packages that provide visual analysis results, making it an appropriate data processing environment for statistical calculations and result generation graphics. R is the language adopted by companies that seek high-performance big data analysis such as IBM and TerraData. R can easily apply analysis techniques appropriate for data by downloading packages using the Internet [6].

3. Related Works

1. A Study on the Effects of Community Conditions on Crime Rates: Longitudinal Analyses Using the Growth Curve Model

This study examines social variables affecting crime rates in human interactions with the physical and social environment.

Based on the ecological theory, it was intended to help establish effective crime prevention measures by identifying the causal relationship between structural variables and crime rates in the community. To this end, a growth curve model was established and analyzed for 162 police stations nationwide using various demosocial data, including the National Police

Agency's crime data (2000-2004) and the National Statistical Office's population housing survey.

To summarize the analysis, first, the initial crime rate (2000 crime rate) and the crime growth rate are statistically significant across regions, suggesting that research on the socio-structural factors that cause this is necessary. Second, it was found that efforts to resolve economic imbalances in middle and large urban areas, which are jurisdictions of the first class, and measures are needed to strengthen voluntary control among residents, namely informal social control, rather than strengthening official social control. Third, in the small urban area, which is under the jurisdiction of the second class, efforts are needed to prevent the dissolution of families and actively support defective families, while strengthening informal social control as in the middle and large urban areas. Finally, it was found that in the farming and fishing villages in the jurisdiction of Grade 3, the strengthening of official social control by protecting and controlling foreigners and increasing the number of police officers will be effective crime prevention measures.

Although it required attention to interpret the results due to limitations such as not being analyzed by type of crime and about 30% of samples were excluded from the analysis, it is meaningful in that it is an ecological study conducted for the first time in Korea [7].

2. Building Crime Prevention System Utilizing Big Data

This study presents the direction of the use of big data technologies to prevent crimes while minimizing the risk of human rights violations and privacy violations of big data technologies in the face of the importance of prediction-based crime prevention activities through valuable big data production. In particular, for rapidly changing and evolving information and communication technologies and criminal phenomena caused by changes in social structure, the need and limitations of crime prevention policies through big data are reviewed from a criminal policy perspective, and the basis for legislative policy discussions that take into account the risks of infringement of information privacy inherent in big data [8].

4. Data Analysis

The purpose of this paper is to analyze how much influence various social variables have on the five major crime rates and how each social variable has a correlation. Simple regression and multiple regression are required to achieve the objectives of the study. At this point, simple regression creates a regression model to predict how one independent variable affects one dependent variable. Multiple regression is the creation of a regression model for predicting dependent variables with multiple independent variables, which must be a continuous variable by default [9].

We also use the p-value to illustrate the results of regression data analysis. The p-value is an indicator of the significance of several social variables, independent of the five major crime rates, which are dependent variables, and can be confident that they are meaningful only if they are less than 0.05. Depending on the degree of involvement, the number of asterisks is marked from 0 to 3, and the more asterisks there are, the more significant the significance increases.

The order of each data regression is as follows:

First, refine the data accordingly to the analysis. Second, extract the data sheet for regression. Third, perform regression. Fourth, determin the contribution by the p-value resulting from regression analysis. Fifth, interpret graph directions and degrees of involvement specifically.

Table 1. Execution Result Source Code

Execution Result Source Code
total_Crime_police_num<- cbind(total_Crime,police_num)
total_Crime_police_num<- total_Crime_police_num[-c(1,3)]
colnames(total_Crime_police_num)=c("total_crime" ,"police_num")
write.csv(total_Crime_police_num, " \total_Crime_police_num.csv", row.names=F)
total_Crime_police_num_result<- lm(total_crime_rate~police_num,data=total_Crime _police_num)
plot(total_Crime_police_num$police_num,total_Cri me_police_num$total_crime_rate)
abline(total_Crime_police_num_result,col="red")

After refining the data on the total number of crimes provided by the National Police Agency and the

number of hired police officers provided by the National Statistical Portal, the following source code was used to visualize the correlation of the big 5 crime rates according to the number of police officers as shown in Figure 1. Figure 1 shows that the higher the number of police officers, the higher the crime rate. It can be inferred that the result was the deployment of more police forces in areas with higher crime rates.

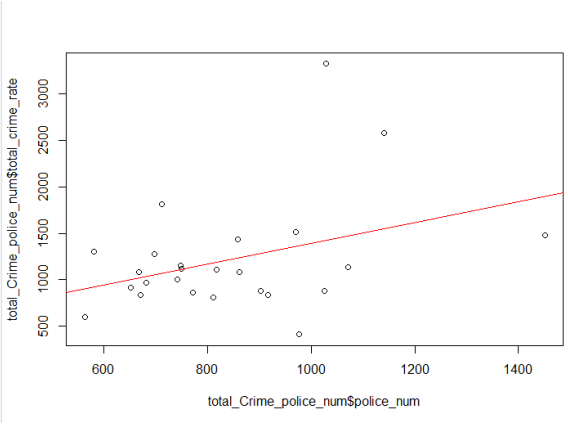

Fig. 1. Number of Police Officers - Big 5 Crimes Correlation Visualization

Table 2. Execution Result Source Code

Execution Result Source Code
> summary(total_Crime_police_num_result) Call: lm(formula = total_crime_rate ~ police_num, data = total_Crime_police_num) Residuals: Min 1Q Median 3Q Max -957.58 -337.13 -85.51 156.16 1900.03 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 269.5075 511.3166 0.527 0.6032 police_num 1.1229 0.5911 1.900 0.0701 . --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 581.2 on 23 degrees of freedom Multiple R-squared: 0.1356, Adjusted R-squared: 0.09806 F-statistic: 3.609 on 1 and 23 DF, p-value: 0.07007

Based on the following results, the p-value was 0.07, which was higher than 0.05, significance was lower, and contribution was lower due to lack of significance.

It can be seen that the big 5 crime rates and police deployments do not show statistical significance to each other. This suggests that the deployment of police forces is not effective in crime prevention. Therefore, it is not a reasonable policy to increase the number of police officers unconditionally in crime prevention policies.

Table 3. Execution Result Source Code

Execution Result Source Code
total_Crime_residentPopulation<-cbind(total_crime_rate,residentPopulation[,2]) colnames(total_Crime_residentPopulation)=c("total_crime","residentPopulation") total_Crime_residentPopulation<-cbind(total_crime_rate,residentPopulation[,2]) colnames(total_Crime_residentPopulation)=c("total_crime","residentPopulation") write.csv(total_Crime_residentPopulation, "total_Crime_residentPopulation.csv", row.names=F) total_Crime_residentPopulation_result<-lm(total_crime~residentPopulation,data=total_Crime_residentPopulation) plot(total_Crime_residentPopulation$residentPopulation,total_Crime_residentPopulation$total_crime) abline(total_Crime_residentPopulation_result,col="red")

After refining the data on the number of permanent residents in each administrative district of Seoul provided by the National Statistical Portal, the following source code was executed to visualize the correlation of the big 5 crime rates according to the number of resident populations by x-axis and y-axis. Figure 2 shows that the higher the number of business owners, the lower the crime rate. As a result, the higher the number of people, the lower the big 5 crime rate.

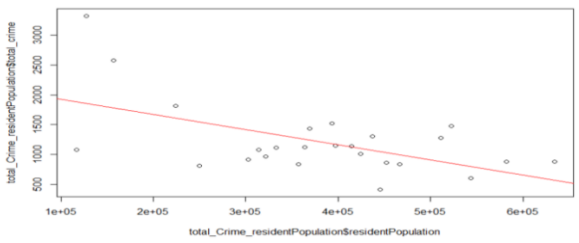

Fig. 2. Resident Population - Big 5 Crimes Correlation Visualization

Table 4. Execution Result Source Code

Execution Result Source Code
> summary(total_Crime_residentPopulation_result) Call: lm(formula = total_crime ~ residentPopulation, data = total_Crime_residentPopulation) Residuals: Min 1Q Median 3Q Max -801.01 -298.90 -94.37 239.42 1472.46 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 2.178e+03 3.153e+02 6.908 4.83e-07 *** residentPopulation -2.544e-03 7.870e-04 -3.232 0.00368 ** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 518.4 on 23 degrees of freedom Multiple R-squared: 0.3124, Adjusted R-squared: 0.2825 F-statistic: 10.45 on 1 and 23 DF, p-value: 0.003683

Based on the following results, the p-value was 0.00368, which is lower than 0.05, and the contribution was determined to be high, considering that there were two asterisks.

In other words, the big 5 crime rates and the number of perpetrators show high statistical significance to each other. This suggests a lower incidence of crime in residential complexes than in non-residential complexes. Therefore, it can be said that the housing complex revitalization policy is a helpful policy in crime prevention.

Table 5. Execution Result Source Code

Execution Result Source Code
> summary(total_Crime_residentPopulation_result) Call: lm(formula = total_crime ~ residentPopulation, data = total_Crime_residentPopulation) Residuals: Min 1Q Median 3Q Max -801.01 -298.90 -94.37 239.42 1472.46 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 2.178e+03 3.153e+02 6.908 4.83e-07 *** residentPopulation -2.544e-03 7.870e-04 -3.232 0.00368 ** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 518.4 on 23 degrees of freedom Multiple R-squared: 0.3124, Adjusted R-squared: 0.2825 F-statistic: 10.45 on 1 and 23 DF, p-value: 0.003683

After refining the data on the number of accommodations provided by the National Statistical Portal, the following source code is used to visualize the correlation of the big 5 crime rates according to the number of accommodations with the x-axis and the big 5 crime rates as y-axis as shown in Figure 3. What Figure 3 shows is that the higher the number of accommodations, the higher the crime rate. Unlike residential complexes, it can be seen that the higher the inflow rate of outsiders who use short-term accommodation, the higher the big 5 crime rates.

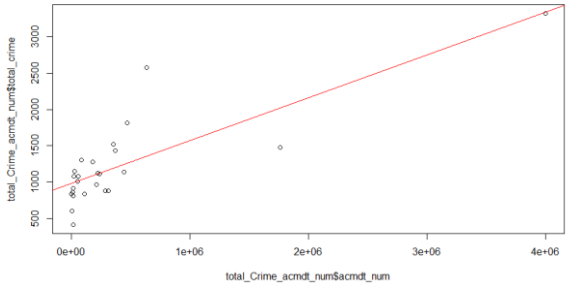

Fig. 3. Number of Accommodations - Big 5 Crimes Correlation Visualization

Table 6. Execution Result Source Code

Execution Result Source Code
> summary(total_Crime_acmdt_num_result) Call: lm(formula = total_crime ~ acmdt_num, data = total_Crime_acmdt_num) Residuals: Min 1Q Median 3Q Max -581.7 -186.8 -14.7 153.1 1223.5 Coefficients:

Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.827e+02 8.290e+01 11.853 2.82e-11 ***
acmdt_num 5.890e-04 9.136e-05 6.447 1.41e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 373.1 on 23 degrees of freedom
Multiple R-squared: 0.6438, Adjusted R-squared: 0.6283
F-statistic: 41.56 on 1 and 23 DF, p-value: 1.405e-06

The following results determined that the p-value was 1.41e-06, which is lower than 0.05, and that the contribution was very high, considering that there were three asterisks.

That is, the total number of big 5 crime rates and accommodations shows very high statistical significance to each other. This suggests that the more pervasive the environment in which there is no face-to-face relationship between each other, the higher the crime rate. Therefore, it can be said that strengthening security near accommodations is a helpful policy in crime prevention.

Table 7. Execution Result Source Code

Execution Result Source Code
total_Crime_floatingopulation<-cbind(total_crime_rate,floatingopulation[,2])
colnames(total_Crime_floatingopulation)=c("total_crime","floatingopulation")
total_Crime_floatingopulation<-read.csv("total_Crime_floatingopulation.csv",header=TRUE)
total_Crime_floatingopulation_result<-lm(total_crime~floatingopulation,data=total_Crime_floatingopulation)
plot(total_Crime_floatingopulation$floatingopulation,total_Crime_floatingopulation$total_crime)
abline(total_Crime_floatingopulation_result,col="red")
summary(total_Crime_floatingopulation_result)

After refining the data on the number of floating populations provided by the National Statistical Portal, the following source code was executed to visualize the correlation of the big 5 crime rates according to the number of floating populations, as shown in Figure 4. Figure 4 shows that the higher the floating population, the higher the crime rate. As a result, it can be seen that the higher the commercial activity and the higher the inflow rate of outsiders, the higher the big 5 crime rates.

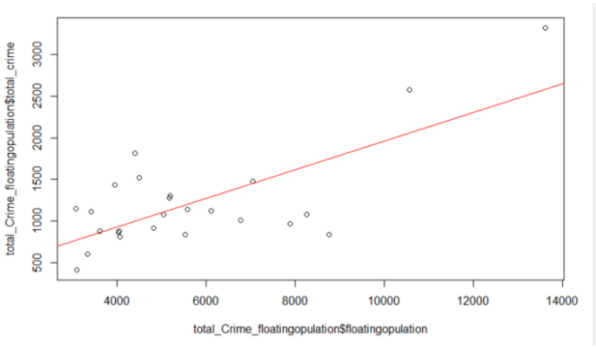

Fig. 4. Floating Population - Big 5 Crimes Correlation Visualization

Table 8. Execution Result Source Code

Execution Result Source Code
> summary(total_Crime_floatingopulation_result)
Call:
lm(formula = total_crime ~ floatingopulation, data = total_Crime_floatingopulation)
Residuals:
Min 1Q Median 3Q Max
-913.79 -211.13 -52.66 283.72 816.48
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 240.72492 217.43835 1.107 0.28
floatingopulation 0.17185 0.03509 4.897 6.03e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 437.4 on 23 degrees of freedom
Multiple R-squared: 0.5104, Adjusted R-squared: 0.4891
F-statistic: 23.98 on 1 and 23 DF, p-value: 6.027e-05

Based on the following results, the p-value was 6.03e-05, which was lower than 0.05, and the contribution

was very high, considering that there were three asterisks.

That is, the big 5 crime rates and the number of floating populations show very high statistical significance to each other. This suggests that the more pervasive and commercial-enabled environments are, the higher the crime rate is. Therefore, it can be said that strengthening security, such as strengthening patrols near commercial districts with large floating populations, is a helpful policy in preventing crime.

Table 9. Execution Result Source Code

Execution Result Source Code
total_Crime_tax<-cbind(total_crime_rate,tax[,2])
colnames(total_Crime_tax)=c("total_crime","tax")
total_Crime_tax<-cbind(total_crime_rate,tax[,2])
colnames(total_Crime_tax)=c("total_crime","tax")
write.csv(total_Crime_tax, "total_Crime_tax.csv",
row.names=F□)
total_Crime_tax_result<-lm(total_crime~tax,data=total_Crime_tax)
plot(total_Crime_tax$tax,total_Crime_tax$total_crime)
abline(total_Crime_tax_result,col="red")

After refining the average paid property tax data for each administrative district of Seoul provided by the National Statistical Portal, the following source code was executed to visualize the correlation of the big 5 crime rates according to property taxes as shown in Figure 5. Figure 5 shows that the more property taxes are paid, the higher the crime rate. This suggests that the prejudice that higher economic levels will result in lower crime rates is wrong.

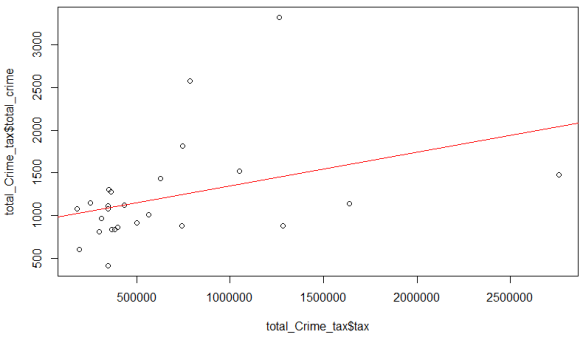

Fig. 5. Property Tax - Big 5 Crimes Correlation Visualization

Table 10. Execution Result Source Code

Execution Result Source Code
> summary(total_Crime_tax_result)
Call:
lm(formula = total_crime ~ tax, data = total_Crime_tax)
Residuals:
Min 1Q Median 3Q Max
-682.6 -273.2 -109.9 147.5 1872.4
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.546e+02 1.791e+02 5.329 2.07e-05 ***
tax 3.940e-04 2.059e-04 1.913 0.0683 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 580.6 on 23 degrees of freedom
Multiple R-squared: 0.1373, Adjusted R-squared: 0.09979
F-statistic: 3.66 on 1 and 23 DF, p-value: 0.06825total_Crime_floatingopulation_result<-lm(total_crime~floatingopulation,data=total_Crime_floatingopulation)
plot(total_Crime_floatingopulation$floatingopulation,total_Crime_floatingopulation$total_crime)
abline(total_Crime_floatingopulation_result,col="red")
summary(total_Crime_floatingopulation_result)

Based on the following results, it was determined that the p-value was 0.0683, which is higher than 0.05, significance is lower, and contribution is lower due to lack of significance.

The five major crime rates and economic levels show no statistical significance to each other. Therefore, it can be predicted that regional economic level balancing policies will not be of much help in crime prevention policies.

Table 11. Execution Result Source Code

Execution Result Source Code
total_Crime_pubs_num<-cbind(total_crime_rate,pubs_num[,2])
colnames(total_Crime_pubs_num)=c("total_crim

```
e","pubs_num")
total_Crime_pubs_num_result<-lm(total_crime~pubs_num,data=total_Crime_pubs_num)
plot(total_Crime_pubs_num$pubs_num,total_Crime_pubs_num$total_crime)
abline(total_Crime_pubs_num_result,col="red")
```

After refining the data on the number of entertainment bars provided by the National Statistical Portal, the following source code was executed to visualize the correlation of the big 5 crime rates according to the number of entertainment bars by x-axis and y-axis. Figure 6 shows that the higher the number of entertainment bars, the higher the crime rate. Thus, the higher the entertainment industry, the higher the crime rate.

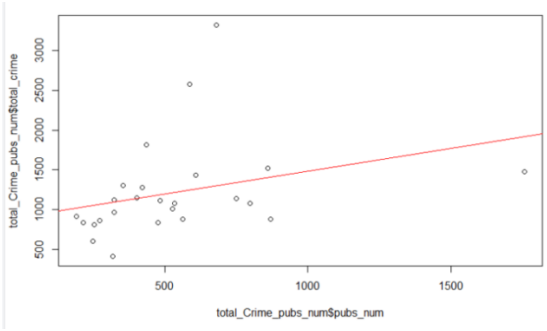

Fig. 6. Number of Pubs - Big 5 Crimes Correlation Visualization

Table 12. Execution Result Source Code

Execution Result Source Code
> summary(total_Crime_pubs_num_result) Call: lm(formula = total_crime ~ pubs_num, data = total_Crime_pubs_num) Residuals: Min 1Q Median 3Q Max -684.3 -288.8 -137.5 111.2 2023.6 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 910.5404 232.0101 3.925 0.000678 *** pubs_num 0.5748 0.3753 1.532 0.139275 --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 595.5 on 23 degrees of freedom Multiple R-squared: 0.09255, Adjusted R-squared: 0.05309 F-statistic: 2.346 on 1 and 23 DF, p-value: 0.1393

Based on the following results, it was determined that the p-value was 0.139275, which was higher than 0.05, less significant, and less contribution due to lack of significance.

The five major crime rates and entertainment scores show no statistical significance to each other. Therefore, it can be predicted that the policy of cracking down on entertainment bars will not be very helpful in crime prevention policies.

Table 13. Execution Result Source Code

Execution Result Source Code
total_Crime_pop_density<-cbind(total_crime_rate,pop_density[,2]) colnames(total_Crime_pop_density)=c("total_crime","pop_density") total_Crime_pop_density_result<-lm(total_crime~pop_density,data=total_Crime_pop_density) plot(total_Crime_pop_density$pop_density,total_Crime_pop_density$total_crime) abline(total_Crime_pop_density_result,col="red")

After refining the population density data for each administrative district of Seoul provided by the National Statistical Portal, the following source code was executed to visualize the five major crime rate correlations according to population density with the x-axis and the five major crime rates as y-axis as shown in Figure 7. Figure 7 shows that crime rates decrease as population density increases. Unlike residential complexes, it can be seen that the higher the inflow rate of outsiders who use short-term accommodation, the higher the big 5 crime rates.

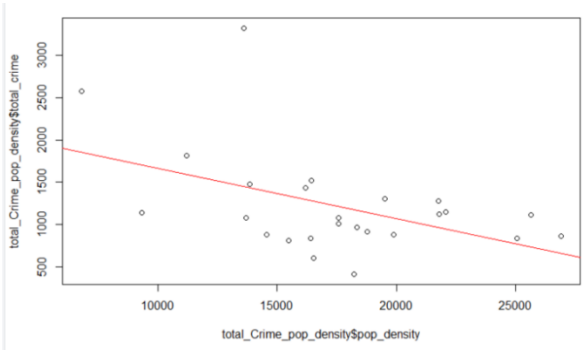

Fig. 7. Population Density - Big 5 Crimes Correlation Visualization

Table 14. Execution Result Source Code

Execution Result Source Code
> summary(total_Crime_pop_density_result) Call: lm(formula = total_crime ~ pop_density, data = total_Crime_pop_density) Residuals: Min 1Q Median 3Q Max -762.98 -360.62 45.84 210.84 1879.51 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 2255.18720 421.66198 5.348 1.97e-05 *** pop_density -0.05946 0.02327 -2.555 0.0177 * --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 551.7 on 23 degrees of freedom Multiple R-squared: 0.221, Adjusted R-squared: 0.1872 F-statistic: 6.527 on 1 and 23 DF, p-value: 0.01771

Based on the following results, the p-value was 0.0177, which is lower than 0.05, and the contribution was considered appropriate, considering that there was only one asterisk.

In other words, the big 5 crime rates and population density show statistical significance to each other. This suggests that the more densely populated the residential complex, the lower the crime rate. Therefore, it can be said that the housing complex activation policy and the household migration activation policy are helpful policies in crime prevention.

Table 15. Execution Result Source Code

Execution Result Source Code
total_Crime_housing_rate<-cbind(total_crime_rate,housing_rate[,2]) colnames(total_Crime_housing_rate)=c("total_crime","housing_rate") total_Crime_housing_rate_result<-lm(total_crime~housing_rate,data=total_Crime_housing_rate) plot(total_Crime_housing_rate$housing_rate,total_Crime_housing_rate$total_crime) abline(total_Crime_housing_rate_result,col="red")

After refining the data on the housing penetration rate for each administrative district of Seoul provided by the National Statistical Portal, the following source code was executed to visualize the correlation of the big 5 crime rates according to the housing penetration rate as shown in Figure 8. What Figure 8 shows is that housing penetration and five major crime rates are not very relevant. As we saw earlier, the number of resident populations affects crime rates, but the low level of involvement suggests that it is necessary to investigate other characteristics of households, such as types of households, rather than quantitative households.

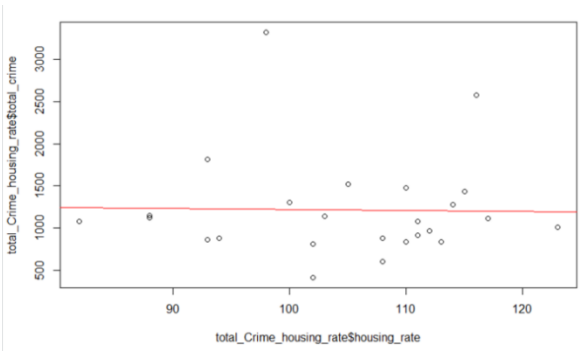

Fig. 8. Housing Penetration Rate - Big 5 Crimes Correlation Visualization

Table 16. Execution Result Source Code

Execution Result Source Code
> summary(total_Crime_housing_rate_result)

```
Call:
lm(formula = total_crime ~ housing_rate, data =
total_Crime_housing_rate)

Residuals:
    Min      1Q  Median      3Q     Max
-809.45 -348.43 -127.33   85.31 2102.06

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1333.039   1281.003   1.041    0.309
housing_rate   -1.123     12.184  -0.092    0.927

Residual standard error: 625 on 23 degrees of
freedom
Multiple R-squared:  0.0003696,     Adjusted R-
squared:  -0.04309
F-statistic: 0.008503 on 1 and 23 DF,  p-value:
0.9273
```

Based on the following results, it was determined that the p-value was 0.927 which was significantly higher than 0.05, significance was lower, and contribution was very low due to lack of significance.

The five major crime rates and housing penetration rates show no statistical significance to each other. Therefore, it can be predicted that a random increase in the number of dwellings in crime prevention policies will not be of much help.

Table 17. Execution Result Source Code

Execution Result Source Code
total_Crime_graduation_num<- cbind(total_crime_rate,graduation_num[,2]) colnames(total_Crime_graduation_num)=c("total_crime","graduation_num") total_Crime_graduation_num_result<- lm(total_crime~graduation_num,data=total_Crime_graduation_num) plot(total_Crime_graduation_num$graduation_num,total_Crime_graduation_num$total_crime) abline(total_Crime_graduation_num_result,col="red")

After refining the data on the number of university graduates by administrative district of Seoul provided by the National Statistical Portal, the following source code was executed to visualize the correlation of the big 5 crime rates according to the number of floating populations as shown in Figure 9. Figure 9 shows that the higher the number of highly educated people, the lower the crime rate.

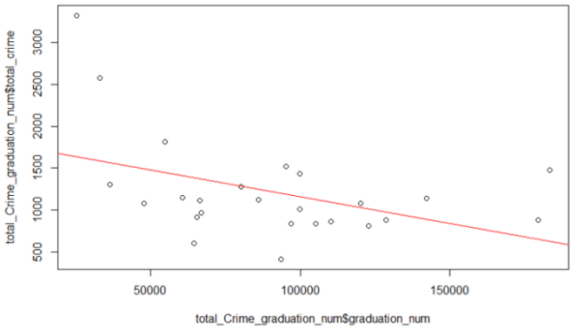

Fig. 9. Population of College Graduates – Big 5 Crimes Correlation Visualization

Table 18. Execution Result Source Code

Execution Result Source Code

```
> summary(total_Crime_graduation_num_result)

Call:
lm(formula = total_crime ~ graduation_num,
data = total_Crime_graduation_num)

Residuals:
   Min     1Q Median     3Q    Max
-787.7 -289.6 -152.2  249.1 1692.0

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.796e+03  2.785e+02   6.448
1.4e-06 ***
graduation_num -6.404e-03  2.810e-03  -2.279
0.0323 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1
Residual standard error: 564.6 on 23 degrees
of freedom
Multiple R-squared:  0.1843,          Adjusted R-
squared:  0.1488
F-statistic: 5.195 on 1 and 23 DF,  p-value:
0.03225
```

Based on the following results, the p-value was 0.323, which is lower than 0.05, and the contribution was considered appropriate, considering that there was one asterisk.

In other words, the big 5 crime rates and the number of highly educated people show statistical significance to each other. This suggests that the higher the education level is guaranteed, the lower the crime rate. Therefore, it can be said that education revitalization policies to raise the overall level of education for residents, regardless of whether they graduate from college, are helpful policies in crime prevention.

Table 19. Execution Result Source Code

Execution Result Source Code
total_Crime_student_num<-cbind(total_crime_rate,student_num[,2]) colnames(total_Crime_student_num)=c("total_crime","student_num") total_Crime_student_num_result<-lm(total_crime~student_num,data=total_Crime_student_num) plot(total_Crime_student_num$student_num,total_Crime_student_num$total_crime) abline(total_Crime_student_num_result,col="red")

After refining the data on the number of youth population by administrative district of Seoul provided by the National Statistical Portal, the following source code was executed to visualize the correlation of the big 5 crime rates according to the number of youth as shown in Figure 10. Figure 10 shows that the higher the number of teenagers, the lower the crime rate.

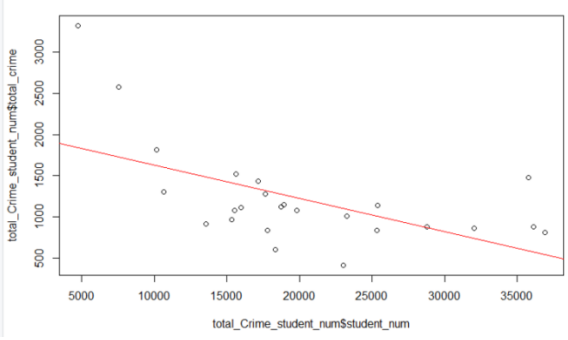

Fig. 10. Population of Youth - Big 5 Crimes Correlation Visualization

Table 20. Execution Result Source Code

Execution Result Source Code
> summary(total_Crime_student_num_result) Call: lm(formula = total_crime ~ student_num, data = total_Crime_student_num) Residuals: Min 1Q Median 3Q Max -691.81 -292.98 -86.18 129.15 1488.86 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 2027.52040 264.16997 7.675 8.67e-08 *** student_num -0.04029 0.01208 -3.336 0.00287 ** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 513.2 on 23 degrees of freedom Multiple R-squared: 0.3261, Adjusted R-squared: 0.2968 F-statistic: 11.13 on 1 and 23 DF, p-value: 0.00287

In other words, the total number of big 5 crime rates and adolescents shows high statistical significance to each other. This suggests that the higher the occupancy rate at the household unit, as well as the number of adolescents, the lower the crime rate. Therefore, it can be said that policies related to the activation of family-level residential policies are helpful in crime prevention in order to create a cooperative atmosphere.

Table 21. Execution Result Source Code

Execution Result Source Code
library(psych) integrated_data_result <-lm(total_crime_rate~.,data = integrated_data) pairs.panels(integrated_data)

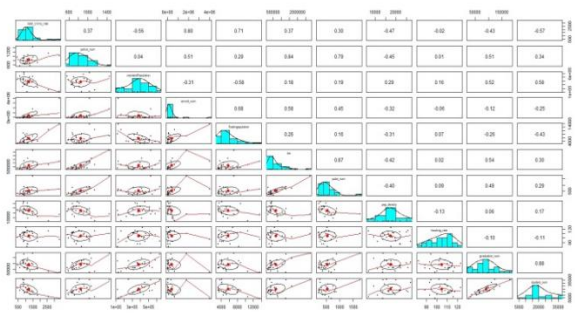

Fig. 11. Multiple Regression Analysis Results for Social Variables

Table 21 and Figure 11 show multiple regression results with each social variable as a variable.

Table 22. Execution Result Graph Description

Execution Result Source Code
> summary(integrated_data_result) Call: lm(formula = total_crime_rate ~ ., data = integrated_data) Residuals: Min 1Q Median 3Q Max -554.89 -136.87 -16.94 135.83 341.71 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 1.756e+03 8.315e+02 2.112 0.0532 . police_num 5.858e-01 6.278e-01 0.933 0.3666 residentPopulation -5.384e-04 7.198e-04 -0.748 0.4669 acmdt_num 2.652e-04 1.500e-04 1.768 0.0988 . floatingopulation 2.544e-02 4.169e-02 0.610 0.5515 tax 1.828e-04 3.245e-04 0.563 0.5821 pubs_num 9.734e-02 3.897e-01 0.250 0.8064 pop_density -8.025e-03 1.636e-02 -0.490 0.6314 housing_rate -3.546e+00 6.412e+00 -0.553 0.5890 graduation_num -5.279e-03 4.588e-03 -1.151 0.2692 student_num -1.302e-02 1.801e-02 -0.723 0.4818 --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 289 on 14 degrees of freedom Multiple R-squared: 0.8699, Adjusted R-squared: 0.777 F-statistic: 9.361 on 10 and 14 DF, p-value: 0.0001286

Table 22 shows that the p-value is 0.0001, which is less than 0.05, and the overall social variable has a significant correlation with the five major crime rates. It can also be seen that each social variable has a pair of mutants that exceed 70 percent of the correlation. Therefore, it is possible to analyze the net effect of each of the five major crime rates, as well as to analyze the correlation between the social variables to select the variables that indirectly affect the five major crime rates.

In other words, multiple regression results in the selection of pairs of variables with more than 70% correlation is the following.

Table 23. Pairs of Social Variables with a Degree of Involvement with more than 70%

Correlation	Social Variable 1	Social Variable 2
0.80	Big 5 Crimes	Number of Accommodations
0.71	Big 5 Crimes	Floating Population
0.84	Number of Police Officers	Property Tax
0.79	Number of Police Officers	Number of Pubs
0.87	Property Tax	Number of Pubs
0.88	Population of Youth	Population of College Graduates

In Table 23, the first and second pairs of variables were previously described as variables directly related to the big 5 crime rates, so we performed a simple regression of the remaining four pairs of variables except for the two pairs of variables.

Table 24. Simple Regression Analysis Source Code

Execution Result Source Code
tax_police_num_result<-

```
lm(tax~police_num,data=tax_police_num)
plot(tax_police_num$police_num,tax_police_num$tax)
abline(tax_police_num_result,col="red")
summary(tax_police_num_result)
```

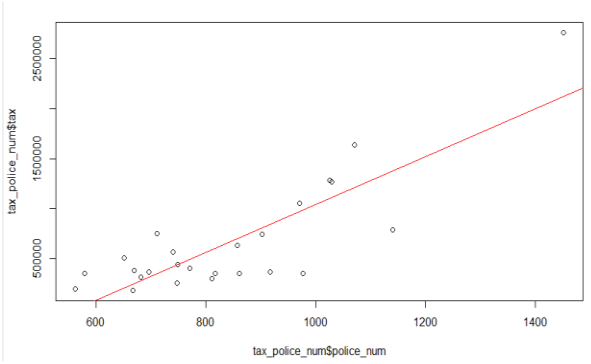

Fig. 12. Property Tax - Number of Police Officers Correlation Visualization

Table 25. Execution Result Graph Description

Execution Result Source Code
Call: lm(formula = tax ~ police_num, data = tax_police_num) Residuals: Min 1Q Median 3Q Max -636977 -183248 35145 180224 639248 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) -1356497.9 284075.9 -4.775 8.15e-05 *** police_num 2396.3 328.4 7.297 2.00e-07 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 322900 on 23 degrees of freedom Multiple R-squared: 0.6984, Adjusted R-squared: 0.6853 F-statistic: 53.25 on 1 and 23 DF, p-value: 2e-07

Figure 12 shows that the higher the number of police officers, the higher the property tax. In addition, in Table 40, the p-value was 2.00e-07, which was lower than 0.05, and the contribution was very high considering that there were three asterisks. However, since property taxes and police numbers are not much correlated with the total crime rate, they are not indirect social variables that affect crime prevention.

Table 26. Simple Regression Analysis Source Code

Execution Result Source Code
pubs_num_police_num_result<-lm(police_num~pubs_num,data=pubs_num_police_num) plot(pubs_num_police_num$pubs_num,pubs_num_police_num$police_num) abline(pubs_num_police_num_result,col="red") summary(pubs_num_police_num_result)

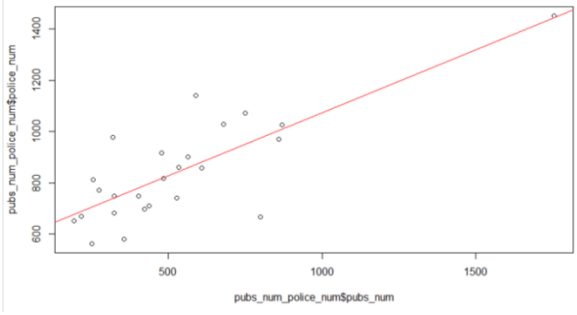

Fig. 13. Number of Pubs - Number of Police Officers Correlation Visualization

Table 27. Execution Result Graph Description

Execution Result Source Code
Call: lm(formula = police_num ~ pubs_num, data = pubs_num_police_num) Residuals: Min 1Q Median 3Q Max -307.054 -58.675 -3.102 54.349 269.398 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 582.31091 48.85040 11.920 2.52e-11 *** pubs_num 0.49029 0.07902 6.205 2.49e-06 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 125.4 on 23 degrees of freedom Multiple R-squared: 0.626, Adjusted R-squared: 0.6098

F-statistic: 38.5 on 1 and 23 DF, p-value: 2.488e-06

Figure 13 shows that the higher the number of entertainment bars, the higher the number of police officers. In Table 44, the p-value was 2.49e-06, which is lower than 0.05, and the contribution was very high, considering that there were three asterisks. However, since the number of entertainment bars and the number of police do not correlate much with the total crime rate, it is not an indirect social variable that affects crime prevention.

Table 28. Simple Regression Analysis Source Code

Execution Result Source Code
tax_pubs_num_result<- lm(tax~pubs_num,data=tax_pubs_num) plot(tax_pubs_num$pubs_num,tax_pubs_num$tax) abline(tax_pubs_num_result,col="red") summary(tax_pubs_num_result)

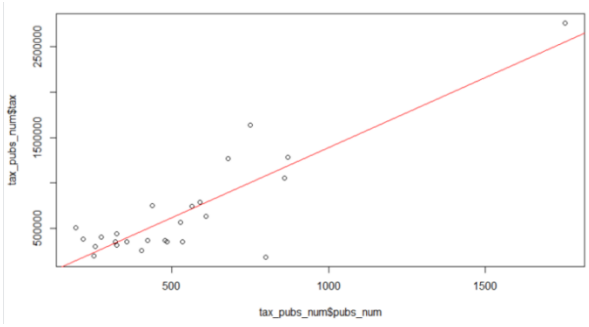

Fig. 14. Number of Pubs – Property Tax Correlation Visualization

Table 29. Execution Result Graph Description

Execution Result Source Code
Call: lm(formula = tax ~ pubs_num, data = tax_pubs_num) Residuals: Min 1Q Median 3Q Max -892523 -128592 13037 133931 637093 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) -155348.4 114035.4 -1.362 0.186 pubs_num 1541.0 184.5 8.355 2.02e-08 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 292700 on 23 degrees of freedom Multiple R-squared: 0.7522, Adjusted R-squared: 0.7414 F-statistic: 69.8 on 1 and 23 DF, p-value: 2.021e-08

Figure 14 shows that the higher the number of entertainment bars, the higher the property tax. In addition, the p-value in Table 48 was 2.02e-08, which was lower than 0.05, and the contribution was very high, considering that there were three asterisks. However, since the number of entertainment bars and property taxes do not have much to do with the total crime rate, it is not an indirect social variable that affects crime prevention.

Table 30. Simple Regression Analysis Source Code

Execution Result Source Code
student_num_graduation_num_result<- lm(graduation_num~student_num,data=student_num_graduation_num) plot(student_num_graduation_num$student_num,student_num_graduation_num$graduation_num) abline(student_num_graduation_num_result,col="red") summary(student_num_graduation_num_result)

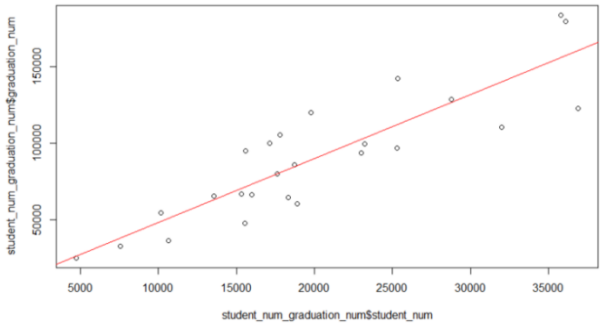

Fig. 15. Population of College Graduates - Population of Youth Correlation Visualization

Table 31. Execution Result Graph Description

Execution Result Source Code
Call: lm(formula = graduation_num ~ student_num, data = student_num_graduation_num) Residuals: Min 1Q Median 3Q Max -37470 -14622 -1196 21848 31056 Coefficients:

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)  6877.262  10301.814   0.668    0.511
student_num     4.155      0.471   8.821 7.71e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 20010 on 23 degrees of freedom
Multiple R-squared:  0.7719,    Adjusted R-squared:  0.7619
F-statistic: 77.81 on 1 and 23 DF,  p-value: 7.708e-09
```

Figure 15 shows that the higher the number of teenagers, the higher the number of highly educated people. In addition, in Table 31, the p-value was 7.71e-09, which was lower than 0.05, and the contribution was very high, considering that it had three asterisks. Also, since the number of teenagers and highly educated people was correlated with the total crime rate earlier, the data shows that the city needs to focus more on improving its education policy to prevent crime.

Accordingly, the social variables are divided by the number of asterisks based on the p-values as shown in Table 32.

Table 32. Social Variables Classified Based on P-Value

Three Asterisks	Two Asterisks	One Asterisks
0.80	Big 5 Crimes	Number of Accommodations
0.71	Big 5 Crimes	Floating Population
0.84	Number of Police Officers	Property Tax
0.79	Number of Police Officers	Number of Pubs
0.87	Property Tax	Number of Pubs
0.88	Population of Youth	Population of College Graduates

5. Conclusion

1. Social Variants - Policies on the Number of Accommodations (3 Asterisks)

First of all, the greater the density of accommodations, the greater the crime rate. Therefore, the first step is to strengthen the patrol of accommodations in order to prevent crimes occurring in dense accommodations. Second, the crackdown on illegal accommodation businesses is strengthened to provide safer accommodation. Third, add cctv to dense accommodations. We propose a total of three policies.

2. Social Variants – Policies on the Number of Floating Populations (3 Asterisks)

The higher the floating population, the greater the crime rate. Therefore, the first step is to strengthen patrols around commercial districts with large floating populations in order to prevent crimes occurring in places with large floating populations. Second, police personnel are deployed to police stations near the commercial district, considering that the previous indiscriminate expansion of police personnel was inefficient. Third, commercial district-centered CCTV expansion. We propose a total of three policies.

3. Social Variants – Policies on the Number of Residents (2 Asterisks)

The higher the resident population, the lower the crime rate. Therefore, first, actively implementing residential activation policies in non-residential areas. Second, a policy proposal that encourages each region to actively participate in new town development policies. Third, a policy proposal that actively utilizes non-mainstream areas and develops amenities to reduce barriers to entry into residence. Fourth, a policy proposal to activate public transportation. We propose a total of four policies.

4. Social Variants – Policies on the Number of Youth Populations (2 Asterisks)

The higher the youth population, the lower the crime rate. This really suggests that the higher the occupancy rate of not only teenagers but also families, the lower the crime rate. Therefore, first, gradually expand the school district. Second, promoting the redevelopment of old-fashioned housing complexes for a pleasant residential environment. Third, the activation of shopping malls and amenities to create a peaceful, family-friendly and mutual atmosphere. Fourth, activation of family-based residential policies. Fifth, the policy of initiative and improvement for the middle class. We propose a total of five policies.

5. Social Variants – Policies on Population Density (1 Asterisks)

The higher the population density, the lower the crime rate. Thus, similarly to what we saw earlier, the first is the expansion of the business to improve the

residential environment. Second, adjustment and improvement of housing policies; third, strengthening security near residential environments. A total of three policies are established to reduce barriers to entry into residence.

6. Social Variants – Policies on the Number of Highly Educated Persons (1 Asterisks)

Based on the phenomenon that the higher the number of highly educated people, the lower the crime rate, the higher the education level of the residents should be established. Therefore, the government should actively pursue policies to reduce the number of students per teacher to improve the quality of education, or create an environment where incumbent and students can exchange active feedback to improve the educational environment.

Acknowlegement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2022R1F1A1062953).

References

[1] Park J. H., Lee. K. J., Lee. S. W., “A Study on Fear of crime of people in Seoul”, Seoul Studies, Vol. 19, No. 4, pp 111-129, 2018.

[2] Uh. S. K., "A Study on Perception for Public Safety of Seoul Citizens using Multiple Regression Analysis", The Journal of The Institute of Internet, Broadcasting and Communication, Vol. 18, No. 1, pp. 195-201, 2018.

[3] Park. J. H., “Study on R Parallel Package for Big Data Processing”, Master Degree Thesis, Hannam University, 2017.

[4] Park. J. H., “Hadoop and MapReduce”, Journal of the Korean Data And Information Science Sociaty, Vol. 24, No. 5. pp. 1013-1027, 2013.

[5] Kim. K. H. et al., “Cloud based Clustering System using MapReduce”, Korea Information Processing Society, The KIPS Spring Conference, pp. 159-160, 2013.

[6] Kim. K. S., Lee. K. W., “A Web Application for Open Data Visualization Using R”, Journal of the Korean Association of Geographic Information Studies, Vol. 17, No. 2, pp. 72-81, 2014.

[7] Cheong. J. S., "A Study on the Effects of Community Conditions on Crime Rates: Longitudinal Analyses Using the Growth Curve Model", Korean Criminological Review, Vol. 75, pp. 251-290, 2008.

[8] Yoon. H. S., "Building Crime Prevention System Utilizing Big Data", Korean Institute of Criminology, 2015.

[9] Lee. Y. J., "A Study on the Verification of the Main Effect in Multiple Regression Analysis including Interaction Effect", anagement Research, Vol. 23, No. 4, pp. 183-210, 1994.