Exploratory Data Analysis

- StudentID: 21700741

- Name: Jong hyun Choi

- 1st Major: Life Science

- 2nd Major: Data Science

This report identifies the correlation between the types of crime and the types of place where the crime occurred. Before identifying the correlation, the overall information of data set is checked, and the types of the crime variable and the types of the location of crime variable is examined in detail. In addition, mutual information and chi-square tests are conducted to find out the correlation between the two variables.

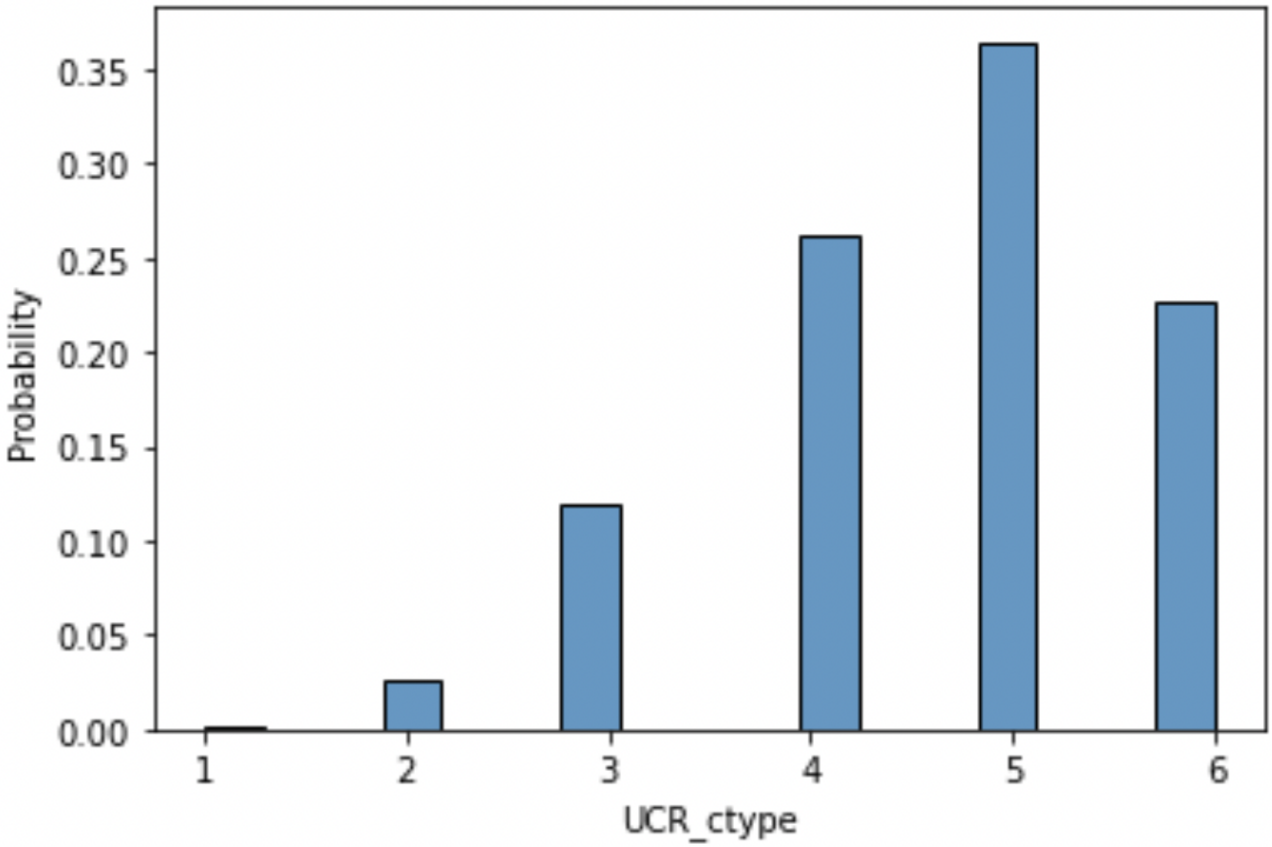
1. Data explanation

1-1. Brief information of Dallas data

- Sample size: 663249
- Number of variables: 107
- Data type: float64(1), int16(1), int8(1), object(104) : A data frame with 1 column with real number data, 2 columns with integer data, and 104 columns with string data.

1-2. UCR_ctype Variable
: Categorical data (1 to 6)

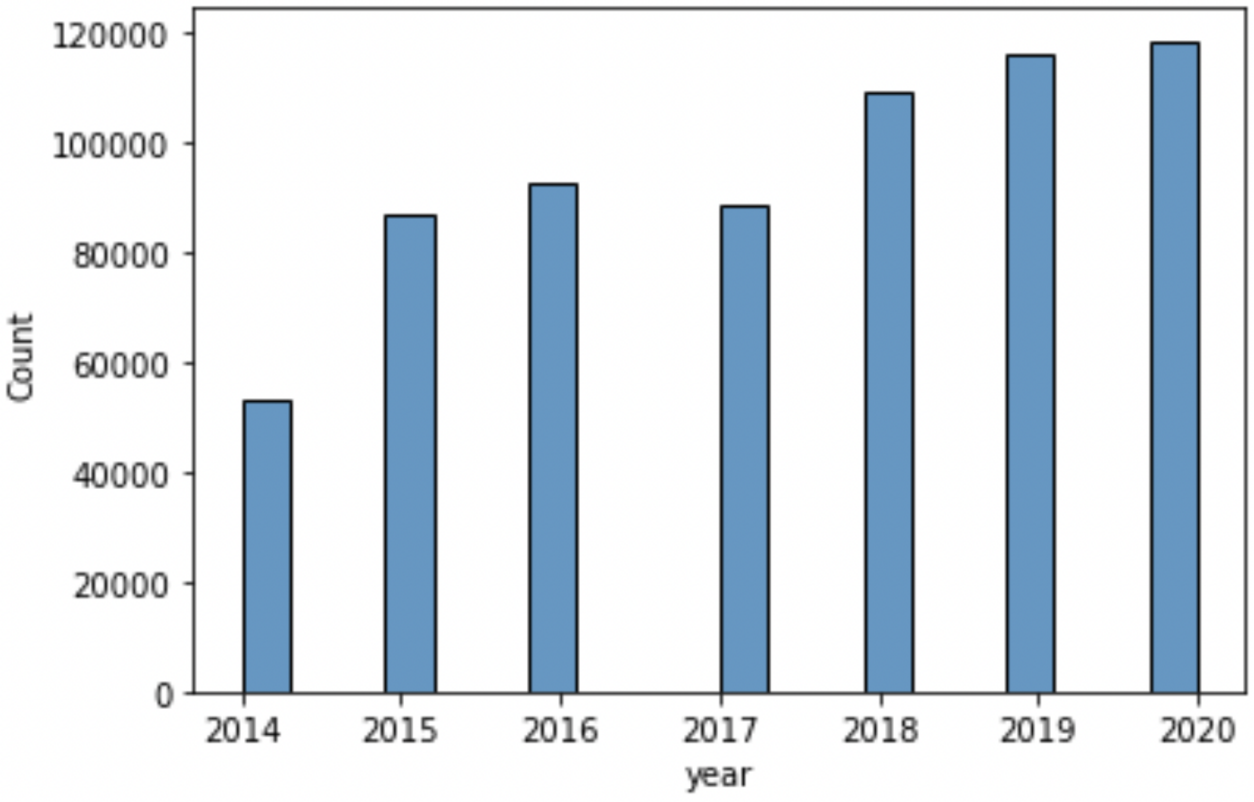
1: Crimes against people: crimes committed directly against others (ex. murder, robbery, assault, etc.)
2: Crime against Property: Crime against property (ex. theft, intrusion, destruction, etc.)
3: Drug/Narcotic Offenses: Drug and Drug Related Crimes (ex. Drug Possession, Distribution, Manufacturing, etc.)
4: Sex Offenses: Sexual crimes (ex. sexual violence, rape, sexual exploitation of children, etc.)
5: White-Collar Crimes: Nonviolent Economic Crimes (ex. Fraud, Financial Crimes, Fraud, etc.)
6: Others: Types of crimes that do not fall under the above category (ex. animal abuse, environmental crimes, etc.)



While relatively few crimes occur against the types of Crime Against People and Crime Administrator Property, Sex Offenses and White-Collar Crimes are relatively common types of crimes.

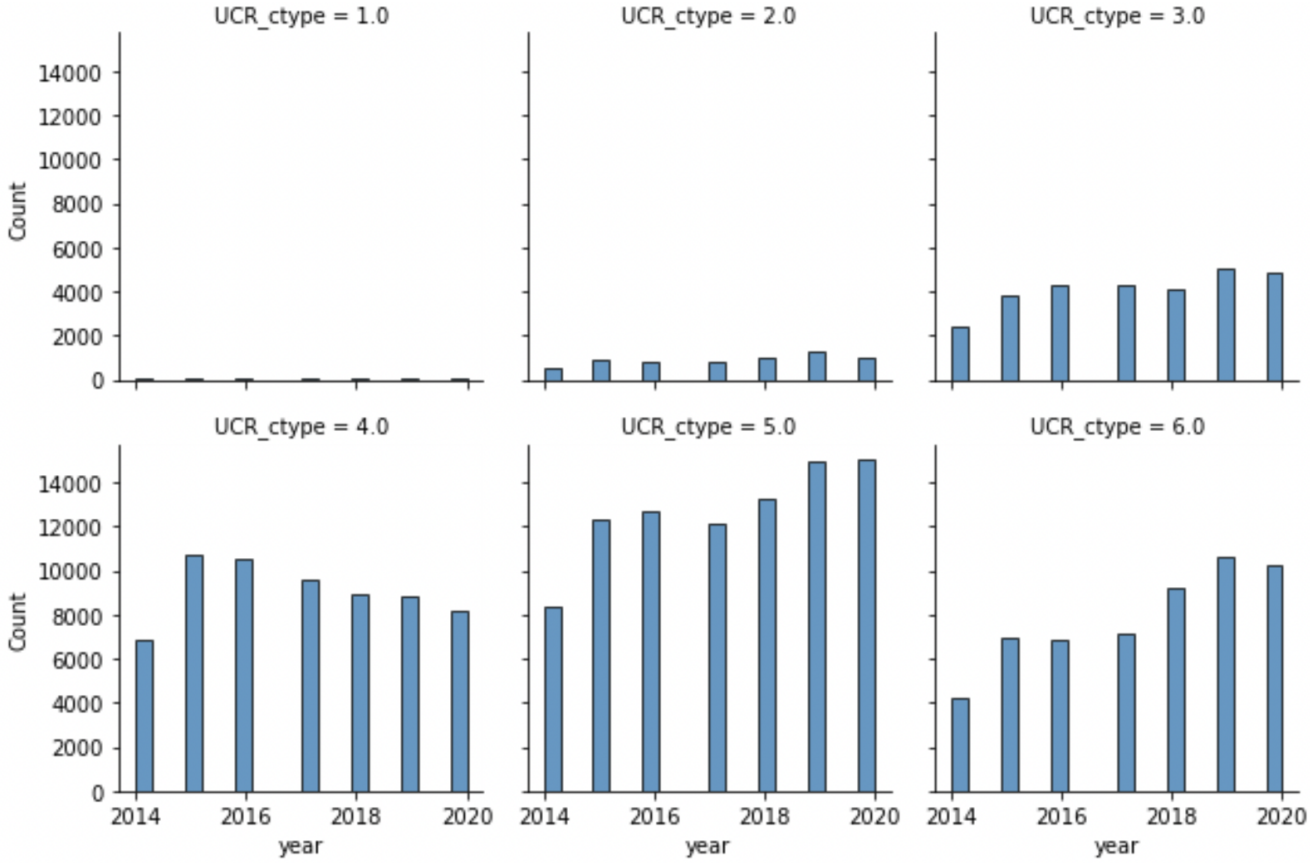
1-3. Year Variable

: Categorical data (2014 ~ 2020)



In order to explain the tendency of graph, the Pearson correlation coefficient for the year and the frequency of crime is calculated, the Pearson correlation coefficient was 0.925, and the p-value was 0.003. Therefore, it is represented that The frequency of crimes increases over the years.

The frequency of crime by year for each UCR_ctype(1~6)



UCR_ctype	Pearson correlation	p-value
Crimes against people (UCR_ctype = 1)	0.668	0.101
Crime against Property (UCR_ctype = 2)	0.789,	0.035
Drug/Narcotic Offenses (UCR_ctype = 3)	0.857	0.014
Sex Offenses: (UCR_ctype = 4)	-0.079	0.865
White-Collar Crimes (UCR_ctype = 5)	0.891	0.007
Others (UCR_ctype = 6)	0.946	0.001

In the table, the values obtained by Pearson correlation and p-value are summarized for each graph. As a result of Pearson correlation, it was concluded that the 'Crime Against Property', 'Drug/Narcotic Offenses', 'White-Collar Crimes', and 'Others' types increased the frequency of crimes over year.

1-4. Summary of other important variables

- Crime Basic Information:

incidentnum: crime occurrence number

ctype: Crime type

UCR_ctype: Crime type (based on UCR)

type of incident: Crime type details

- Crime location Information:

typelocation: crime location type

type of property: type of criminal property

division: Districts

sector: competent sector

geo_lat: latitude of the location of the crime

geo_long: longitude of location of crime

address: the address where the crime

- Crime Time Information:

date1ofoccurrence: date of crime

Year 1 of incidence: Year of crime

month1ofoccurrence: The month of the crime

day1oftheweek: the day of the crime

Time1ofoccurrence: Time of crime

day1oftheyear: number of days of crime year

dateofreport: Date of report creation

dateincident created: Date of the crime report

call receiveddatetime: date and time of receipt of a report

calldatetime: report date and time

callcleareddatetime: reporting resolution date and time

2. Univariate analysis

There are two variables that I would like to investigate intensively.
: Type location, Type of incident

2-1. Type location
: type of location where crime occurred

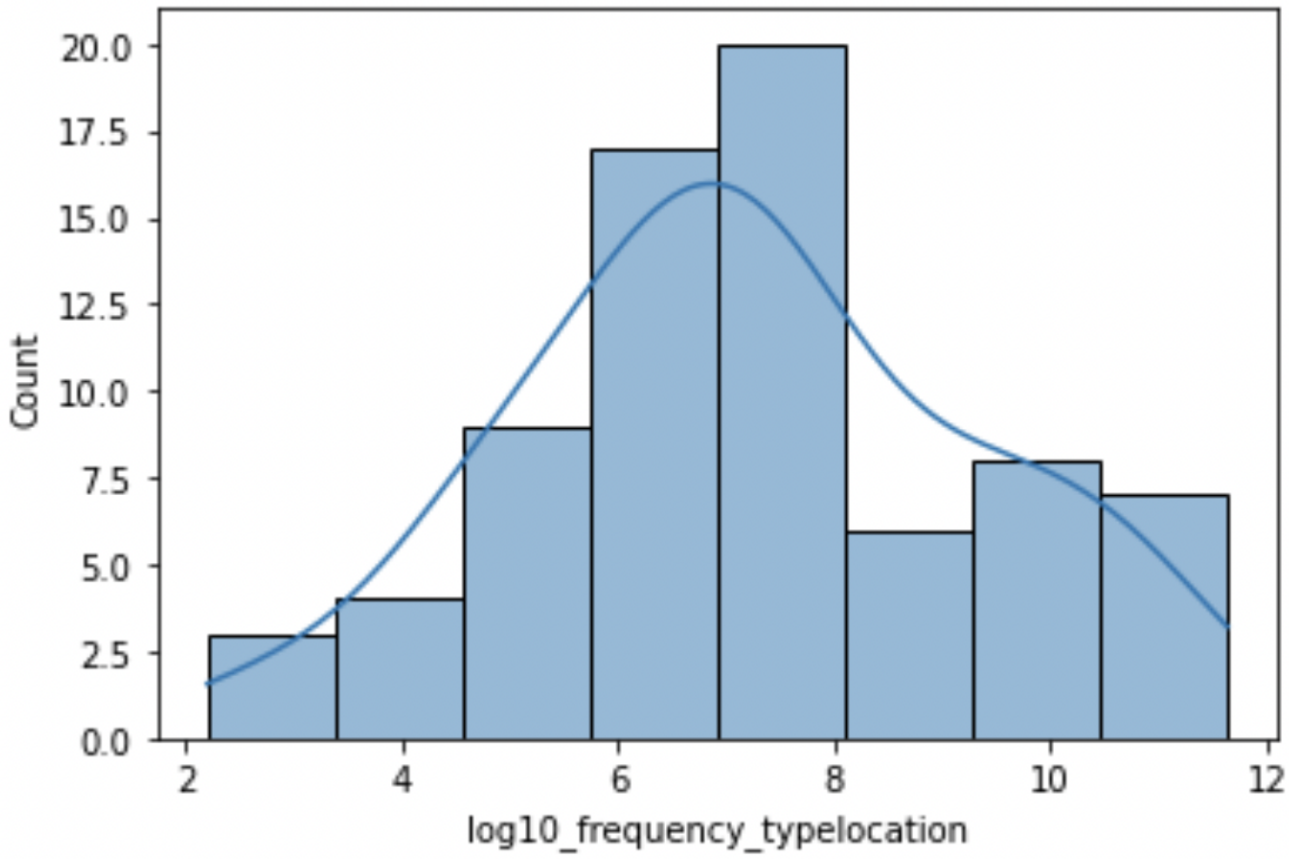
In the values of the Type location variable, there were 742 missing values, and there were 74 unique values.

1. Table of frequency of Type location

Type location	Frequency
Highway, Street, Alley ETC	113395
Single Family Residence - Occupied	79869
Apartment Parking Lot	59192
Parking Lot (All Others)	45627
.....	.....
Playground	21
Military Installation	16
Tribal Lands	9

The frequency of the values in the type location is summarized in the table. Roads were the most common types of crime places, followed by Single Family Residence and Apartment Parking Lot.

2. Histogram of the frequency table



To check how much the frequency of the type location value was distributed, the histogram of the frequency table above was drawn. However, the frequency variation was very severe, therefore log10 was applied and the distribution was examined. As a result, it was confirmed that the distribution of data applied log10 exists in a form similar to the normal distribution.

2-2. Type of Incident
: details information of crime type

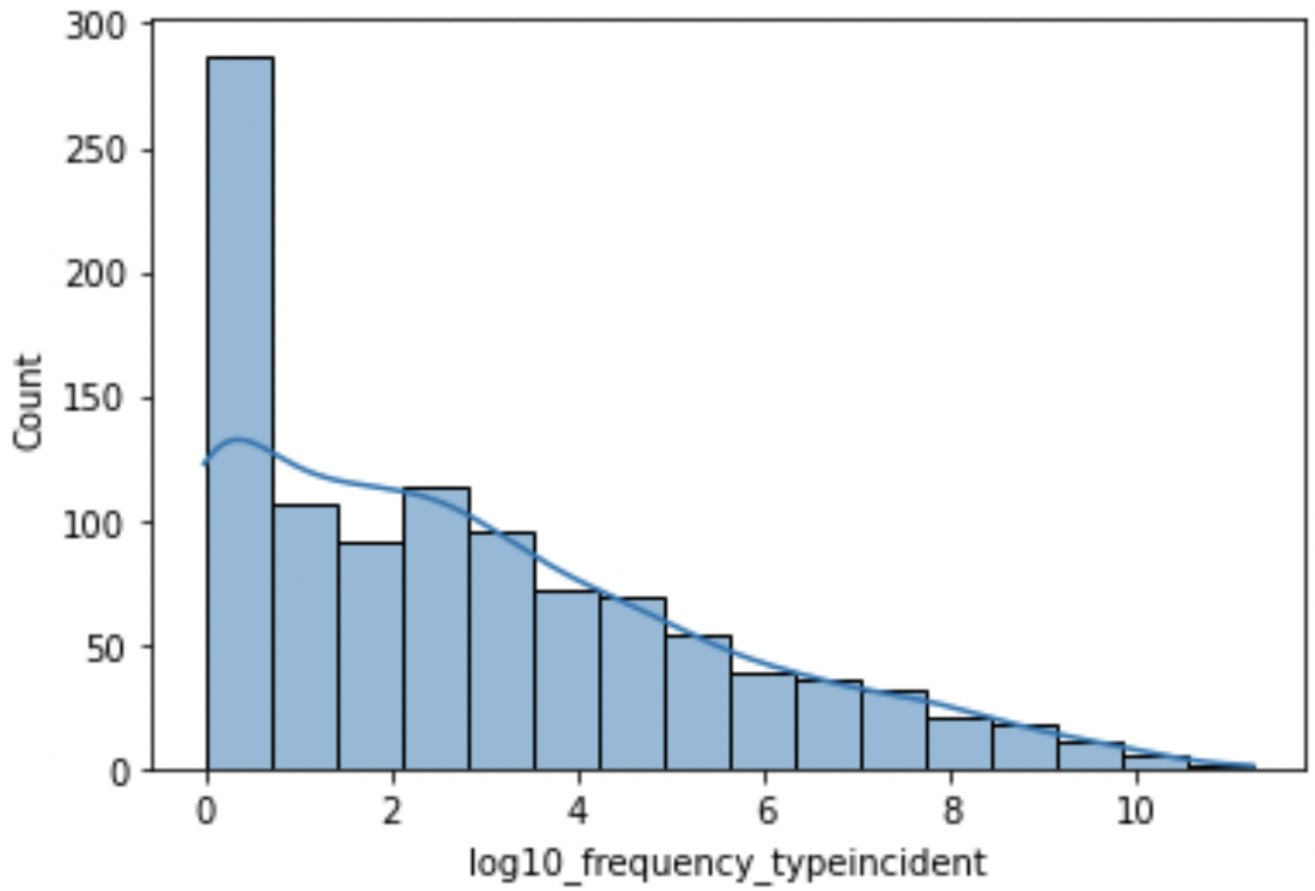
In the values of the Type of Incident column, there were no missing values, and there were 1155 unique values.

1. Table of frequency of Type of incident

Type of incident	Frequency
BMV	75861
UNAUTHORIZED USE OF MOTOR VEH - AUTOMOBILE	34169
FOUND PROPERTY (NO OFFENSE)	24282
BURGLARY OF HABITATION - FORCED ENTRY	23850
.....	.....
MISAPP FIDUC/FINAN PROP > OR EQUAL $30K<$150K	1
INSURANCE FRAUD RISKING BODILY INJ/DEATH	1

The frequency of the values in the type of incident is summarized in the table. BMV value is most frequent value, followed by "UNAUTHORIZED USE OF MOTOR VEH", "FOUND PROPERTY".

2. Histogram of the frequency table



To check how much the frequency of the type of incident value was distributed, the histogram of the frequency table above was drawn. Like type location variable, the frequency variation was very severe, therefore log10 was applied and the distribution was examined. As a result, the data applied log10 showed a distribution biased to the left.

3. Multivariate analysis

The purpose is to figure out that verify correlation between the "type location" and the "type of crime". Since both variables are string data, mutual information and chi square tests were conducted on the frequency of unique values in each column. The results showed that the Mutual information was 0.6171 and the P-value was 0. Therefore, it was confirmed that the correlation between these two variables was very high.

* 'chi2_contingency' function in 'scipy.stats' package is used to calculate p-value, and 'mutual_info_ score' function is used to calculate mutual information.

4. Suggestion

Identifying the correlation between the types of crimes and the types of places where crimes occurred is important in developing crime prevention and response strategies. This is because certain types of crime occur more often in certain places, so you can come up with strategies to prevent and respond to crime in those places.

Drug crimes, for example, are typically prone to occur in certain local types for illegal trade or use. This allows you to analyze data related to drug crime to identify where these crimes are likely to occur and to develop strategies for preventing and responding to crime in your area. This enables efficient distribution of resources for crime prevention and response.

Another example is the lack of healthy public and educational facilities in certain areas, which makes it easier for teenagers to spend time on the streets, which can easily lead to crime. To solve this problem, public facilities and educational facilities can be provided in the area to expand the place of youth activities and prevent crimes.

Therefore, analyzing the correlation between the types of crimes and the types of places where crimes took place is critical to developing crime prevention and response strategies.