EDA Report - Dallas Data Analysis

Information

- StudentID: 21700365

- Name: Son Juchan

- 1st Major: Social Welfare

- 2nd Major: Data Science

Abstract

This report presents an exploratory data analysis (EDA) of crime data in Dallas. The data is imported using Python libraries such as pandas, numpy, matplotlib, and seaborn. The report provides an overview of the data, including its size, variable names, and information. The analysis includes the distribution of crimes by year, month, day, and time. The report also examines the crimes by location, victim race, age, and type. The study uses visualizations to present the findings, such as bar charts and histograms. This report conducts a multivariate analysis based on previous analysis. The topics of multivariate analysis are logistic regression and k-means clustering. The report also provides suggestions for future research.

1. Data overview

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

Dallas = pd.read_stata('raw_Dallas.dta')

Dallas.info()
```

First, the module that used for the EDA was imported.
Pandas and numpy were used to handle data. And matplotlib and seaborn were used to visualize the data.
The `pd.read_stata` code was used to import data to be used for analysis. This allowed the dta file to be read from Python into a data frame. This data frame was saved as a variable named Dallas.

Dallas is data with 663249 rows. And this data has 107 column names, and the names of the columns are not duplicated. Based on this, it can be seen that the size of this data is 663249 rows and has 107 variables.

```
# Check missing values
Dallas.isnull().sum()

# Check the names of variables with missing values
vars_miss = Dallas.columns[Dallas.isnull().any()].tolist()
print(vars_miss)
```

As a result of checking the overall missing value of the data, there was a missing value only in the crime type code 'UCR_ctype'. The outliers will be identified later when analyzing each variable.

2. Univariate analysis

There are many variables in Dallas' crime data.
In order to execute EDA suitable for the purpose using this data, the analysis must be conducted with the necessary variables.
Therefore, variables were classified according to The Six Ws(When, Where, Who, What, Why, How).
Some of the variables were not clearly distinguished among The Six Ws, but the report was prepared according to this criterion as much as possible. Several analyses were run on this, but the report will focus on the variables needed for a multivariate analysis.

2.1 Variable 1 - When

The variables about the 'when' are a collection of time-related variables. The year, month, and hour of the day can tell us something about what time of day a crime usually occurs.The columns used here are named as follows.

- 'year1ofoccurrence' : The year of the crime is occurred.
- 'month1ofoccurence' : The month of the crime is occurred.
- 'day1oftheweek' : The day of the crime is occurred.
- 'time1ofoccurrence' : The time of the crime is occurred.

2.1.1 The number of crimes by time

The number of crimes for this variable by time of day appears to be relatively low during the day, starts to increase at 4pm, and decreases between 11pm and 3am. With the exception of the peak in the number of crimes at 24 hours, we can see that crimes occur mainly in the evening hours.

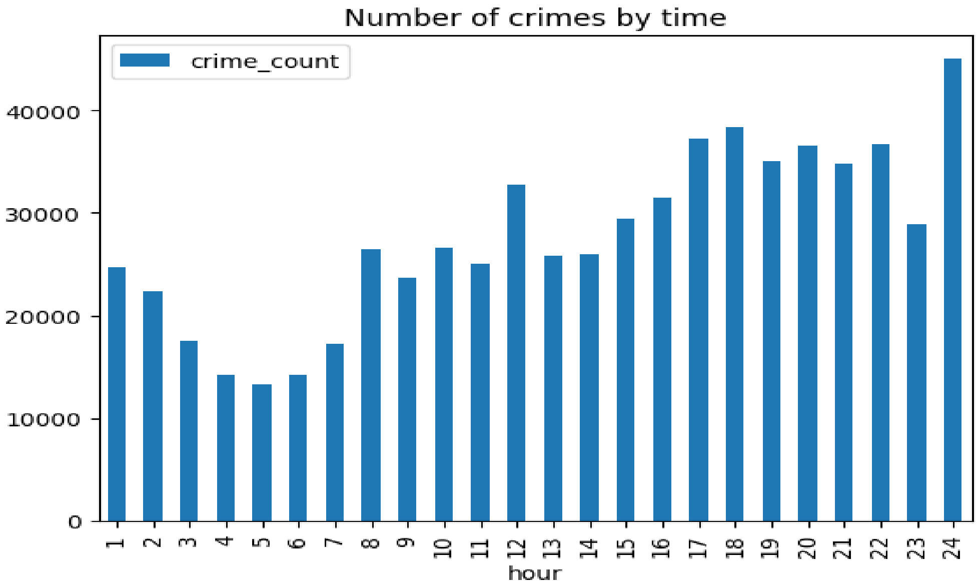


Fig. 1. The number of crimes by time

2.2 Variable 2 - Where

The variables about the 'where' are a collection of location-related variables. The location of the crime can tell us something about the characteristics of the crime. The columns used here are named as follows.

- 'division': The division of the crime is occurred.
- 'victimcity': The city of the crime is occurred.
- 'community': The community of the crime is occurred.
- 'typelocation': The type of location that the crime is occurred

2.2.1 The number of crimes by location

The number of crimes by location is shown in Table 1. The most common location for crimes is the highway, street, or alley. The second most common location is a single-family residence. The third most common location is an apartment parking lot. The fourth most common location is a parking lot. The fifth most common location is an apartment complex or building. Based on this, the data suggests that residential neighborhoods are vulnerable to crime, except for the fact that most crimes occur on the street.

location	count
Highway, Street, Alley ETC	113,395
Single Family Residence - Occupied	79,869
Apartment Parking Lot	59,192
Parking Lot (All Others)	45,627
Apartment Complex/Building	45,045

Table. 1. The number of crimes by location

2.3 Variable 3 - Who

The variables for "who" are a set of victim-related variables. The report uses information about the victims of a crime to infer the characteristics of the crime. The names of the columns used are

- 'victimgender' = The victim's gender information.
- 'victimage' = The victim's age information.
- 'victimrace' = The victim;s race information.
- 'victimtype' = This variable records the type of victim of the crime, which can be either an individual or a specific group of people.

2.3.1 The number of victim age

According to Fig 2, the group most exposed to crime in Dallas is people between the ages of 20 and 30. As you can see from the data, the number of victim continues to increase until age 30 and then gradually decreases after age 30.

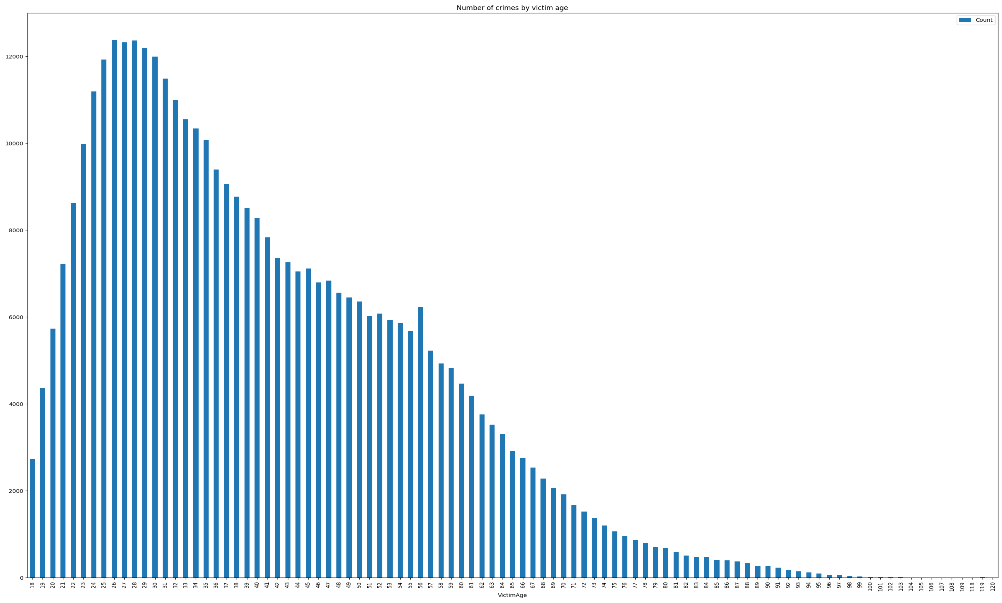


Fig. 2. The number of crimes by age

2.4 Variable 4 - What

The variables for "what" are a set of crime-related variables. This variable contains information about what type of crime was committed.

- 'typeofincident' = This variable records the type of crime that occurred, categorized into types such as violent crime or property crime. Represents the number of incidents per type

The most common type is "BMV," with 75,861 recorded incidents. "BMV" stands for "Burglary of Motor Vehicle," which means theft of a motor vehicle. The next most common crimes are "UNAUTHORIZED USE OF MOTOR VEH - AUTOMOBILE", "FOUND PROPERTY (NO OFFENSE)", "BURGLARY OF HABITATION - FORCED ENTRY", and "PUBLIC INTOXICATION", in that order. The data suggests that these are mainly theft-related crimes.

2.5 Variable 5 - How

The variables for "how" are a set of crime-related variables. This variable contains information about how the crime was committed. Specifically, this category includes variables for what weapon was used in the crime and what damage was done.

- 'weaponused' = This variable records the type of weapon used in the crime, categorized into types such as a gun or a knife. Represents the number of incidents per type
- 'victiminjurydescription' = This variable records the type of injury suffered by the victim, categorized into types such as a gunshot wound or a stab wound. Represents the number of incidents per type

2.6 Variable 6 - Why

The variables for "why" are a set of crime-related variables. This variable contains information about why the crime was committed. Specifically, this category includes variables for what the motive of the crime was and what the relationship between the victim and the perpetrator was.

- 'hatecrime' = This variable records whether the crime was a hate crime or not. Represents the number of incidents per type
- 'familyoffense' = This variable records whether the crime was a family offense or not. Represents the number of incidents per type

2.6.1 The number of family offense

Of the total data, those who responded that domestic violence occurred were categorized as 1 and those who did not were categorized as 0. Of the total data, those who responded that domestic violence occurred were categorized as 1 and those who did not were categorized as 0. As a result, 23,203 data were identified as domestic violence cases.

FamilyOffense	Count
0	640,046
1	23,203

Table. 2. The number of family offense

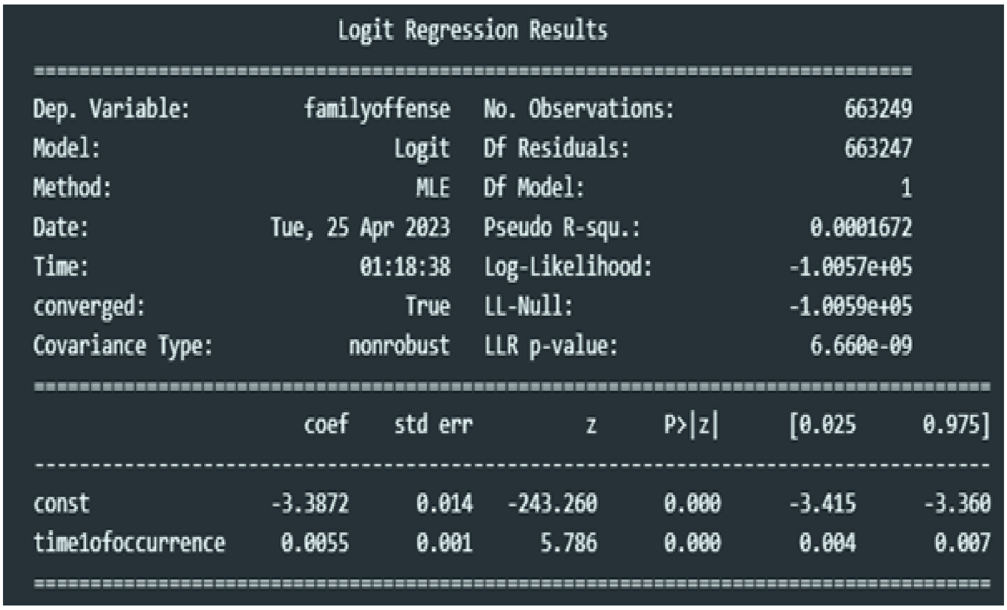
3. Multivariate analysis

Presenation of hidden patterns between variables (correlation, clustering, etc.)

3.1 Correlation Analysis of the Effect of Time on the Occurrence of Domestic Violence

```
import statsmodels.api as sm

# Correlation Analysis of the Effect of Time on the Occurrence of Domestic
Violence
Dallas['familyoffense'] = Dallas['familyoffense'].replace('', 1)
Dallas['familyoffense'] = Dallas['familyoffense'].replace('false', 0)
Dallas['familyoffense'] = Dallas['familyoffense'].astype(int)

# Dallas 에서 familyoffense열만 추출해서 df_familyoffense에 저장
df_familyoffense = Dallas['familyoffense']

# Dallas에서 time1ofoccurrence열만 추출해서 df_time에 저장
df_time = Dallas['time1ofoccurrence']

# df_time를 시간으로 변환해서 df_time_hour에 저장
df_time_hour = pd.to_datetime(df_time, format='%H:%M').dt.hour

# df_familyoffense와 df_time_count를 합쳐서 df_familyoffense_time에 저장
df_familyoffense_time = pd.concat([df_familyoffense, df_time_hour], axis=1)

# 설명 변수와 종속 변수를 분리합니다.
X = df_familyoffense_time[['time1ofoccurrence']]
y = df_familyoffense_time['familyoffense']

# 상수항 추가
X = sm.add_constant(X)

# 로지스틱 회귀모델 적합
logit_model = sm.Logit(y, X)
result = logit_model.fit()

# 결과 출력
print(result.summary())
```

The results of the logistic regression indicate that there is a statistically significant relationship between the occurrence of domestic violence (familyoffense) and the time of occurrence (time1ofoccurrence). The coefficient (coef) of time1ofoccurrence is 0.0055, which means that for every 1 increase in time, the log odds of familyoffense increases by 0.0055, indicating that the later the time of occurrence, the slightly higher the probability of violence. The Pseudo R-squared value is 0.0001672, which is a very low value, meaning that the variation in the independent variable explains very little of the variation in the dependent variable. However, the LLR p-value is less than 0.05, so we can conclude that the model is significant. This is illustrated in Table 3.



```
                           Logit Regression Results
==============================================================================
Dep. Variable:          familyoffense   No. Observations:               663249
Model:                          Logit   Df Residuals:                   663247
Method:                           MLE   Df Model:                            1
Date:                Tue, 25 Apr 2023   Pseudo R-squ.:               0.0001672
Time:                        01:18:38   Log-Likelihood:            -1.0057e+05
converged:                       True   LL-Null:                   -1.0059e+05
Covariance Type:            nonrobust   LLR p-value:                 6.660e-09
====================================================================================
                       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------
const               -3.3872      0.014   -243.260      0.000      -3.415      -3.360
time1ofoccurrence    0.0055      0.001      5.786      0.000       0.004       0.007
====================================================================================
```

Table 3. Result

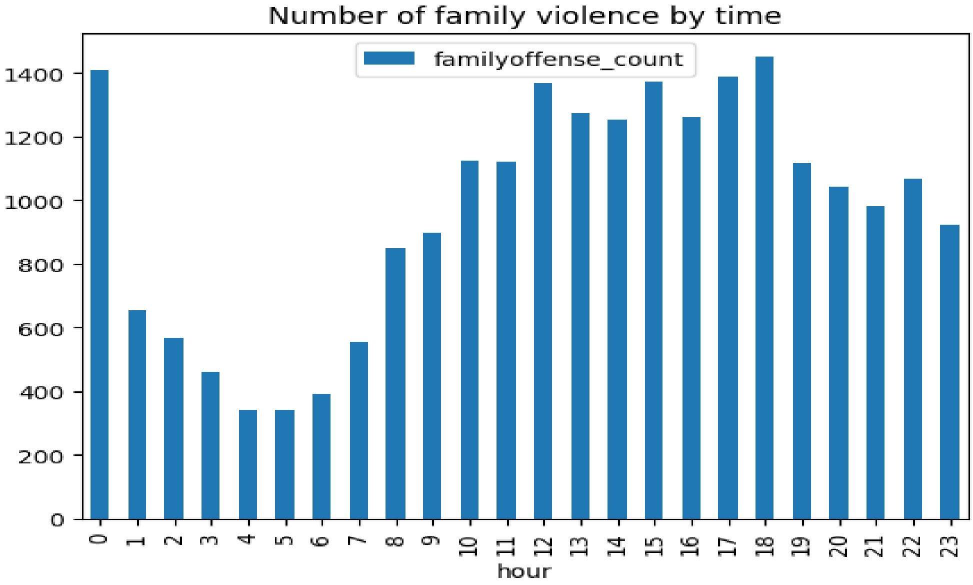


Fig. 3. Number of family violence by time

3.2 Clustering Analysis of the crime type and the victim age

```
# Clustering of location type by victim age
# Dallas에서 victimage와 typelocation열만 추출해서 df_victimage_location에 저장
df_victimage_location = Dallas[['victimage', 'typelocation']]
# victimage열의 값이 ''인 행을 삭제
df_victimage_location =
df_victimage_location.drop(df_victimage_location[df_victimage_location['victimage'
] == ''].index)
# victimage열의 값이 200 이상인 행을 삭제
df_victimage_location =
df_victimage_location.drop(df_victimage_location[df_victimage_location['victimage'
].astype(int) >= 200].index)
# victimage열의 값을 정수형으로 변환
df_victimage_location['victimage'] =
df_victimage_location['victimage'].astype(int)
# victimage열의 값을 오름차순으로 정렬
df_victimage_location = df_victimage_location.sort_values(by='victimage',
ascending=True)
# typelocation열의 값을 범주형으로 변환
df_victimage_location['typelocation'] =
df_victimage_location['typelocation'].astype('category')
# typelocation열의 범주형을 숫자형으로 변환
df_victimage_location['typelocation'] =
df_victimage_location['typelocation'].cat.codes
# typelocation열의 값을 오름차순으로 정렬
df_victimage_location = df_victimage_location.sort_values(by='typelocation',
ascending=True)

# K-means clustering
from sklearn.cluster import KMeans

# K-means clustering을 위한 데이터를 준비합니다.
X = df_victimage_location[['victimage', 'typelocation']]
# K-means clustering을 수행합니다.
kmeans = KMeans(n_clusters=4, random_state=0).fit(X)
# K-means clustering의 결과를 데이터프레임에 저장합니다.
df_victimage_location['cluster_id'] = kmeans.labels_
# K-means clustering의 결과를 시각화합니다.
plt = df_victimage_location.plot(kind='scatter', x='victimage', y='typelocation',
c='cluster_id', cmap='Set1', colorbar=False)
```

The code above extracts the 'victimage' and 'typelocation' columns from the dataframe df_victimage_location, performs K-means clustering, and visualizes the results. K-means clustering is an unsupervised learning algorithm that clusters data so that the data within each cluster has similar characteristics to each other. Before processing the data, we removed blank values and outliers. In this code, we clustered the data into four clusters based on the 'victimage' and 'typelocation' columns, and visualized the clustering results in a scatter plot. Looking at the clustering results, we can see that four clusters were formed, which are separated by different colors. The reason for the four clusters is that we created criteria based on similar types of places. The data is divided into four characteristics based on the type of place. Public, Commercial, Residential, and Other are the criteria for the clusters. That's why the variable has four clusters.

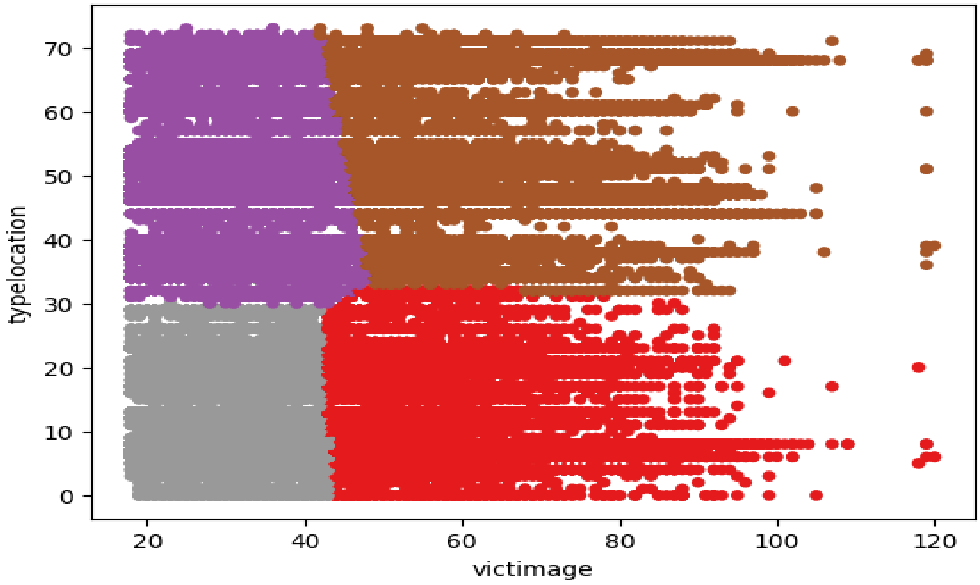


Fig. 4. Clustering result

The different clusters represent clusters of data with different characteristics in the 'victimage' and 'typelocation' columns. So, for example, if cluster 0 contains victims under the age of 10 located in commercial locations, we can infer from the characteristics of that cluster that crimes of that type are more likely to occur in commercial or public areas than in older or residential neighborhoods.

4. Suggestions

4.1 Suggestions based on the results from the logistic regression

The logistic regression analysis showed a statistically significant relationship between the occurrence of domestic violence and the timing of its occurrence. Therefore, from this data, we can conclude that the later the timing of the occurrence, the higher the probability of violence. To combat this, the government could increase the number of police patrols during those hours, or deploy more domestic violence counselors.

4.2 Suggestions based on the results from the clustering analysis

By understanding the crime characteristics of each cluster, you can create crime prevention policies for specific clusters. For example, certain clusters may contain a high number of younger victims, so the government can apply additional precautions to these clusters, such as school education programs or installing CCTV in public facilities.