EDA and Various Analysis report using Dallas Crime Data

Junsuk Seo

2023 4 16

Exploratory Data Analysis

- StudentID: 21700356

- Name: Junsuk Seo

- 1st Major: ICT convergence

- 2nd Major: Data Science

Brief Summary:

Developing a crime occurrence prediction model.
Identifying major types of crimes and clustering them by type to understand the correlations between types.

1. Data overview

Basic data description:

Crime data for Dallas TX from 2014 to 2020
663249 samples and **107** variables

Key variables used:

ctype, offensetype, ucroffensedescription, dateincidentcreated

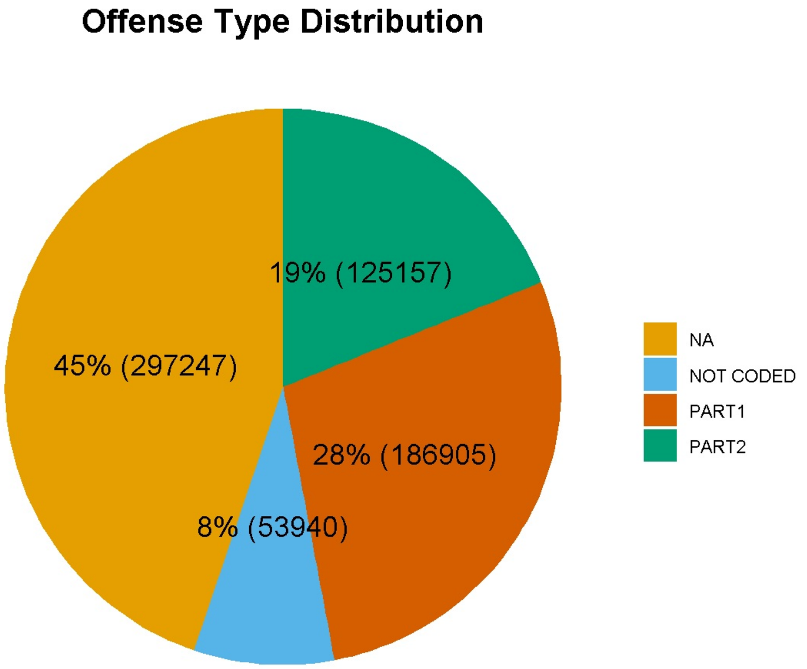
ctype: String type, 53 unique values, 297247 missing values, **Criminal classification term** used in Uniform Crime Reporting(UCR), the US crime statistics

offensetype: String type, 3 unique values, 297247 missing values, **Criminal severity classifications** used by UCR
- **“PART1”**: The most **severely classified** crime in the UCR, includes murder, robbery, rape, theft, assault, etc.
- **“PART2”**: Relatively **less serious crimes** than the “Part 1” category, and also includes statistical information such as people’s illegal behavior, includes Traffic Violation, Other Theft, Assault, Inappropriate Conduct, and Gambling.
- **“NOT CODED”**: Classification process **fails to classify** a specific incident into an appropriate category.Ex) Missing info about the exact category

ucroffensedescription: String type, 46 unique values, 297247 missing vlaues, **a description of each crime type** in the crime classification and record system used by UCR

dateincidentcreated: String type, 0 missing values, **date and time** the actual crime occurred

Sample count graph by **“offensetype”**



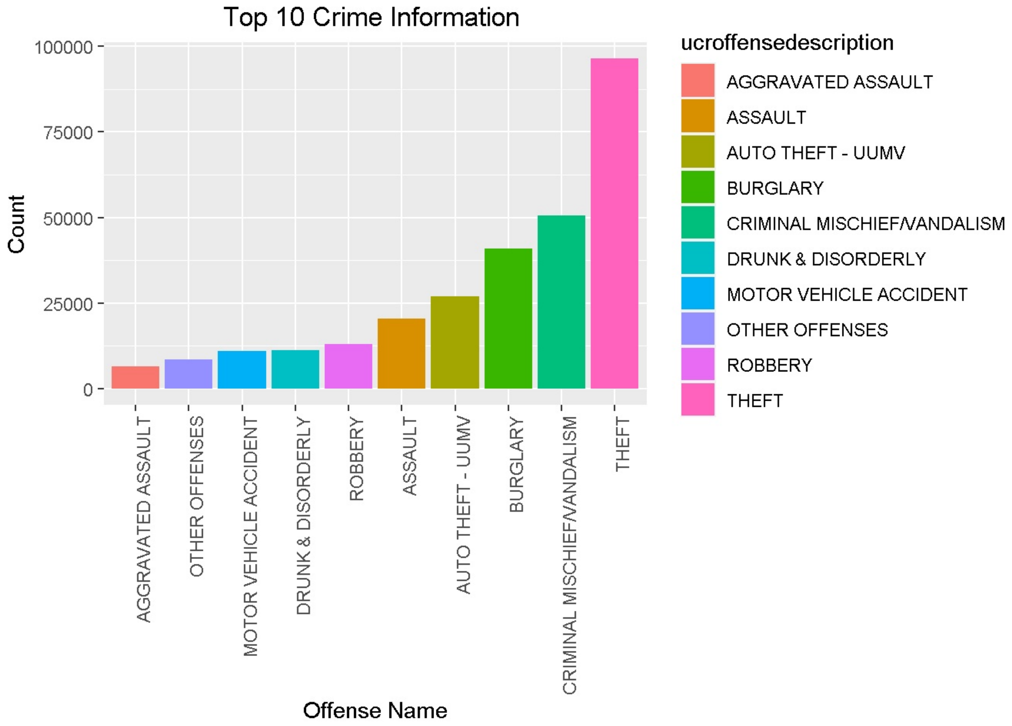
PART1: 125157 samples
PART2: 186905 samples
NOT CODED: 53940 samples
NA: 297247 samples

As of June 30, 2018, there are only **73** data without records of crime types.
However, since July 1, 2018, there are **297174** data without a crime type record.

2.Univariate Analysis

The 10 Most Crimes in Dallas

Except when there is no offensetype or when the offensetype is **“NOT CODED”**
In the case of **“NOT CODED”**, there are many cases that cannot be regarded as crimes (ex: “Found”, case of receiving lost property), so it was excluded.



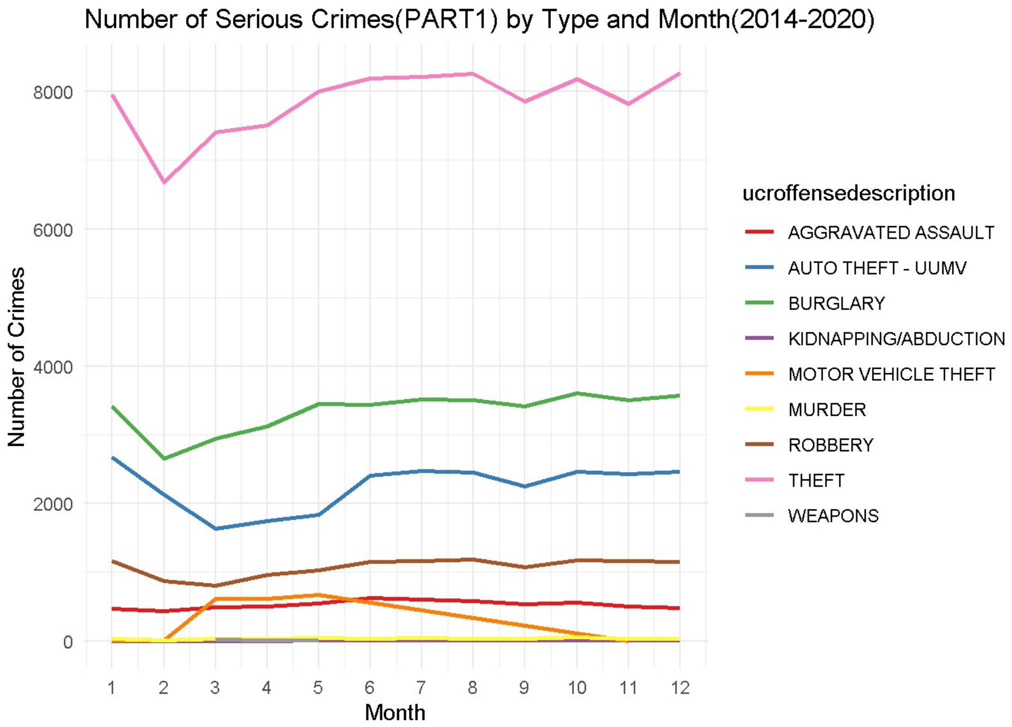
Top 10 crime information: THEFT, CRIMINAL MISCHIEF/VANDALISM, BURGLARY, AUTO THEFT - UUMV, ASSAULT, ROBBERY, DRUNK &

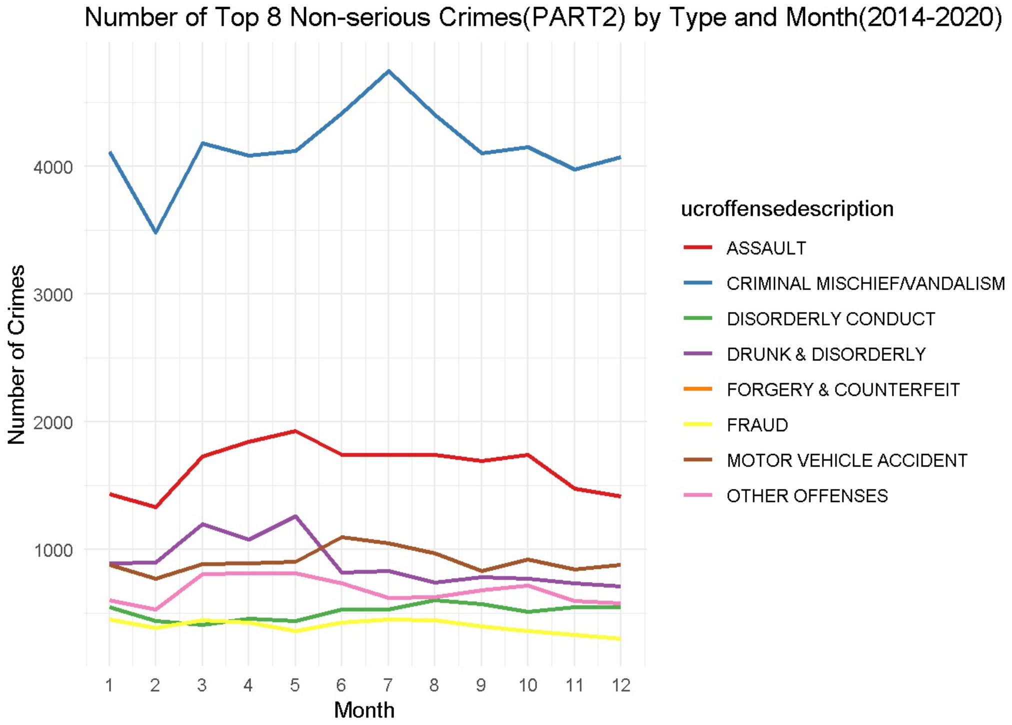
DISORDERLY, MOTOR VEHICLE ACCIDENT, OTHER OFFENSES, AGGRAVATED ASSAULT

Total number of incidents per crime (by month)

Measurement period: 2014.06 ~ 2018.05 Total 4 years
Since the data starts from June 2014 and there is almost no record of crime types from July 2018, data after June 2018 is **not available.**
Therefore, only data from 4 years from 14.06 to 18.05 were selected.

Sum the total number of major crimes per month divided into felonies (Part 1) and misdemeanors (Part 2)





Note. A total of 12 murders and a total of 15 weapons occurred over the past 4 years.
In the case of **PART1**, crimes tended to be small in **February and March**, and the number of cases was similar in all cases except for February and March.
Also, unusually, in the case of **motorcycle theft**, it rarely happened in winter.

Even in the case of major misdemeanors in **Part2**, the number of crimes was the lowest in **February**.
In the case of **drunk & disorderly**, it happened a lot in **March, April, and May** when the weather was good and it was good to go outside, and in the case of **criminal mischief/vandalism**, it happened a lot in the **hot summer** when the weather was hot and the discomfort index was high.

3.Multivariate Analysis

Time Series Data Clustering

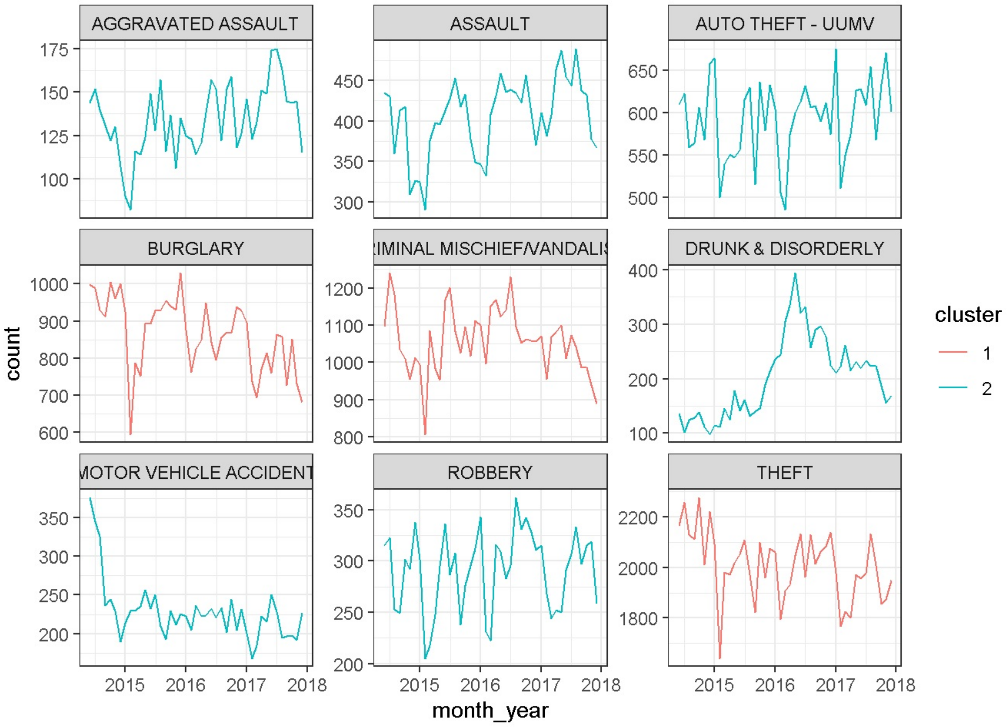
DTW(Dynamic Time Warping) based instead of Euclidean distance

Dynamic Time Warping (DTW): A technique used to measure the similarity between two time series with different lengths. It is often used in signal processing and pattern recognition.
DTW calculates the distance between two time series by measuring the minimum amount of warping and it is more flexible than other distance metrics, such as Euclidean distance.

Clustering the 9 most common crimes(excluding **“OTHER OFFENSES”** from the previous top 10 crime information chart)

```
##             [,1]       [,2]       [,3]        [,4]        [,5]
## Sil    0.6592874 0.37519459 0.09947318 -0.02486481 -0.02474222
## SF     0.0000000 0.00000000 0.00000000  0.00000000  0.00000000
## CH     7.1329804 5.45204680 3.84254567  2.40565072  1.44088748
## DB     0.1226696 0.76445156 0.66003485  0.61396057  1.96793643
## DBstar 0.1226696 2.28509690 1.97344666  3.20927833  4.89340325
## D      0.9360721 0.05561898 0.03159436  0.03385287  0.04334700
## COP    0.2051091 0.17455639 0.16512137  0.16065912  0.10848536
```

This table is an indicator table that helps determine how many clusters are most appropriate when the number of clusters is 2 to 6.
Since the silhouette index is highest when cluster = 2, 2 clusters were selected.



Burglary, criminal mischief/vandalism, theft are grouped in cluster 1
Aggravated assault, auto theft - UUMV, drunk & disorderly, motor vehicle accident, robbery are grouped in cluster 2

In the case of **cluster 1**, the peaks and troughs of the wave tend to decrease, showing a large seasonality of one year cycle.
In the case of **cluster 2**, there are relatively many residual waves.

Overall, the least crime occurred at the beginning of **2015**, and seasonality was confirmed in all patterns except for **drunk & disorderly.**

4. Suggestion

Time forecasting model with LSTM

Since the data after June 2018 is not appropriate to use, there is no way to determine how many serious crimes have occurred since then.
Therefore, I would like to propose a time series model to predict how many serious crimes have occurred.
As a time series prediction model, I would like to propose a model using the **LSTM** technique, one of the deep learning techniques.

Data description:
Only data from June 2014 to June 18 was used.
Extract only the sample whose **offensetype** is **“PART1”** and **“ctype”** is not blank
Train and validation period: 2014.06 ~ 2017.12
Test period: 2018.01 ~ 2018.06
A model to predict the number of serious crimes

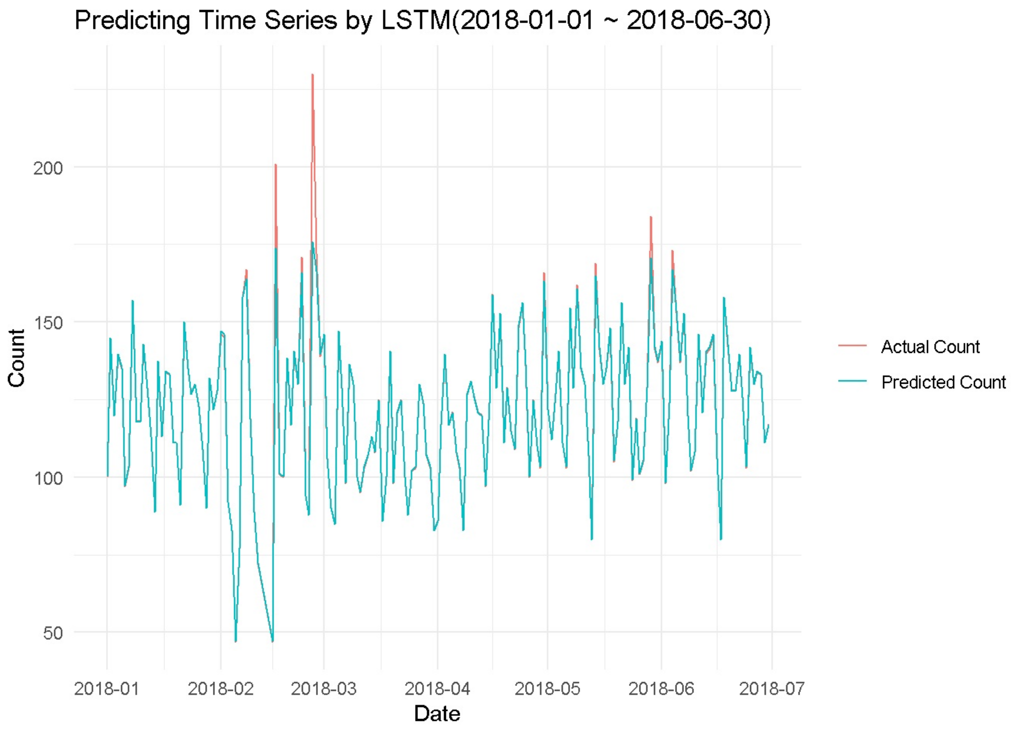
model structure:

```
## Model
## Model: "sequential"
## ________________________________________________________________________________
##  Layer (type)                       Output Shape                      Param #
## ================================================================================
##  lstm (LSTM)                        (None, 50)                        10400
##
##  dense (Dense)                      (None, 1)                         51
##
## ================================================================================
## Total params: 10,451
## Trainable params: 10,451
## Non-trainable params: 0
## ________________________________________________________________________________
```

Units: 50
Time step size:1
Batch size: 32(default)
Epochs: 300
Learning rate: 0.001(default)
Dropout rate: X
Train loss: 6.0036 **Validation loss**: 33.5284

LSTM prediction result

```
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
```



Calculating error metrics

```
##        MSE     RMSE       MAE R_SQUARED
## 1 22.5408 4.747715 0.9888422 0.9654223
```

It did not predict the high volatility part well, but it showed a pretty good prediction overall.
Therefore, through the LSTM model, it is possible to predict serious crime cases after June 2018.