Exploratory Data Analysis

- StudentID: 21700442

- Name: KyungTaek Oh

- 1st Major: ICT

- 2nd Major: DataScience

Brief summary of your proposed project idea.

The five most common incidents by the 'incidenttype' variable were analyzed by year, day, and location.

1. Data overview

Descriptives statistics on overall data (sample size, number of variables, data type, data range, distribution, etc.)

The original sample size is 663,249 and the number of columns is 107. However, since I will be analyzing Dollar's cities within the state of Texas, I limited the data to the state of Texas and the city of Dallas. At this time, I changed all 25 types of cities with the wrong dallas, such as dallastx, dallas tx dalllas, ddallas, etc. to dallas and included them in the analysis. Therefore, the total number of samples with Texas as the state and dallas as the city is 652,926 (98.44%). The variable types are all objects except for UCR_ctype, cnt, and year. UCR_ctype and typeofproperty have 420,075 (63.33%) and 486,530 (73.35%) missing values, respectively. In addition, there are 24 columns with the value "", which are more than 80% of the total number of data: community, specialreportprerms, victimbusinessname, victimbusinessaddress, victimbusinessphone, victiminjurydescription, victimcondition, hatecrime, and victimpackage, so I excluded them from the analysis. The crime data is from 2014 to 2020, and most of the data is skewed to the right, meaning that the top data accounts for most of the total data.

2. Univariate analysis

Presentation of key variables from various aspects

2.1 Incident Type

I identified the most common types of incidents using the "incidenttype" variable, which does not have a value of "". We grouped all 1,046 types of incidents and calculated their frequency, and the top five are: burglary of motor vehicle, unauthorized use of motor vehicle, found property, burglary of habitation, and public intoxication. The average per type is 624, but you can see that it is heavily skewed to the right as the

bottom 75% have 81 or fewer, and the top 5, which are 0.47% of the total 1046, account for 26.97% of the total frequency.

typeofincident	count
burglary of motor vehicle	74119
unauthorized use of motor veh - automobile	33863
found property (no offense)	24089
burglary of habitation - forced entry	23377
public intoxication	20709

<Table1> Top5 Incident Types

2.2 NIBRS Crime Category

Additionally, to see if a different criteria would yield similar results new criteria is needed. I chose the National Incident-Based Reporting System (NIBRS) categories as the new classification because NIBRS is known to be more detailed than Uniform Crime Reporting (UCR) and has 62,498 fewer non-empty values than the 'ctype' variable representing UCR.

The Table2 shows the top five most frequent crimes based on NIBRS categories. They account for about 47% of all crimes and are dominated by theft-related offenses too. So when you combine both Table 1 and Table 2, you can see that theft-related crimes, residential burglaries, and vehicle-related crimes are the most common.

nibrscrimecategory	count
MISCELLANEOUS	105775
LARCENY/ THEFT OFFENSES	96276
DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY	38309
MOTOR VEHICLE THEFT	36131
ASSAULT OFFENSES	33919

<Table2> Top5 NIBRS Crime Categories

2.3 Hate Crime Description

With a non-hate crime rate of 96.55%, it is safe to say that hate crimes are extremely rare in the city of Dallas.

hatecrimedescription	count
None	626864
Unknown	21249
Anti White	81
Anti Black Or African American	65
Anti Homosexual (Gays and Lesbians)	54
Anti Hispanic	35
Anti Jewish	21
Anti Male Homosexual (Gay)	21
Anti Other Ethnicity/Natl Origin	16
Anti Female Homosexual (Lesbian)	13

<Table3> Top10 Hate Crimes

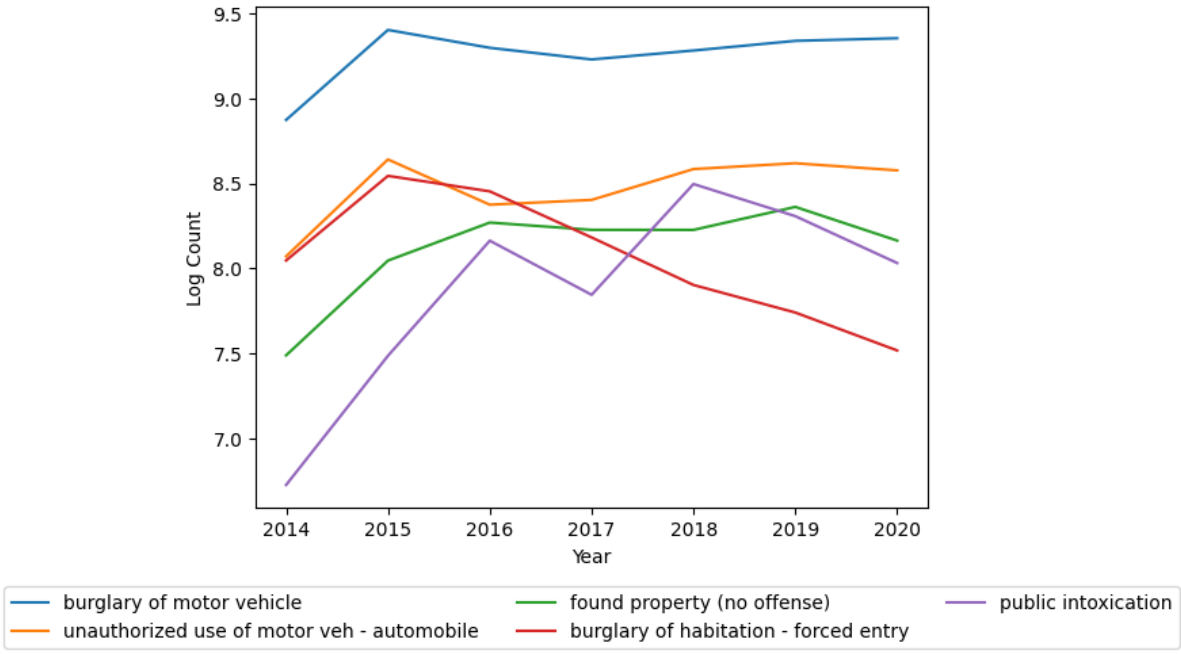
3. Multivariate analysis

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

3.1 Types of Incidents By Year

The below graph shows the top five types and years grouped by incident type. Burglary of motor vehicle, unauthorized use of a motor vehicle, and found property show similar trends. However, burglary of habitation and public intoxication show a different trend. The public intoxication have mostly increased year over year, while burglary of habitation has consistently decreased since 2015.

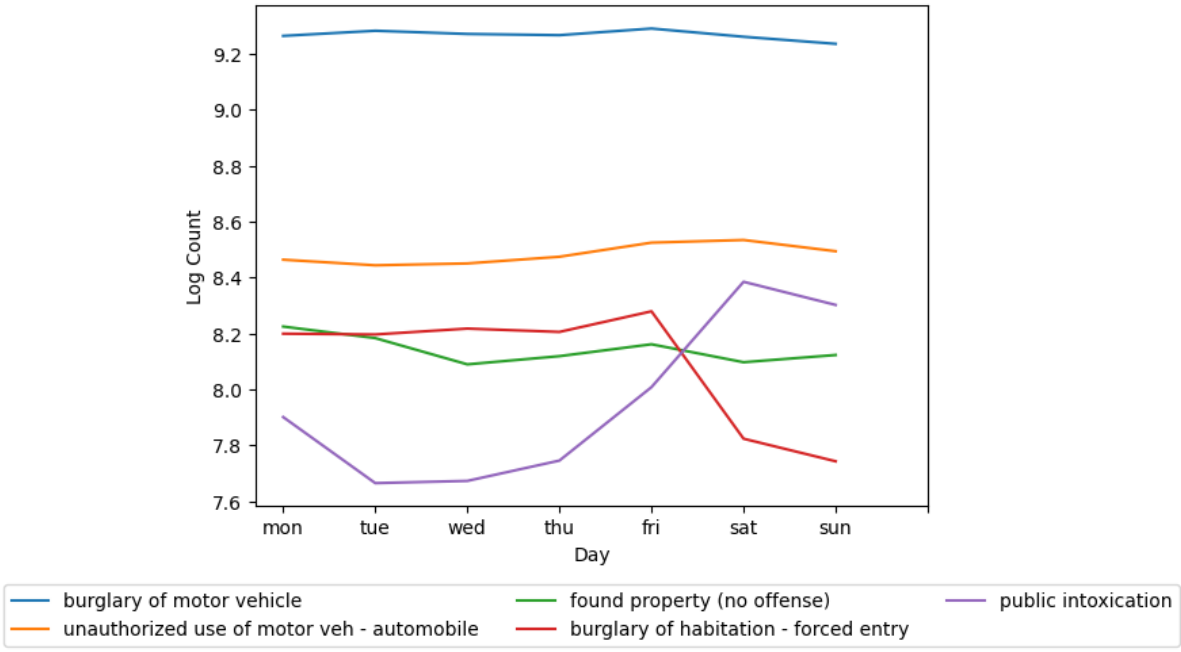
The reason for the log transformation is that when the y-axis is frequency, the lowest and highest values are more than 6 times different, making it difficult to compare.



<Figure1> Top5 Incidents by Year

3.2 Types of Incidents By Day

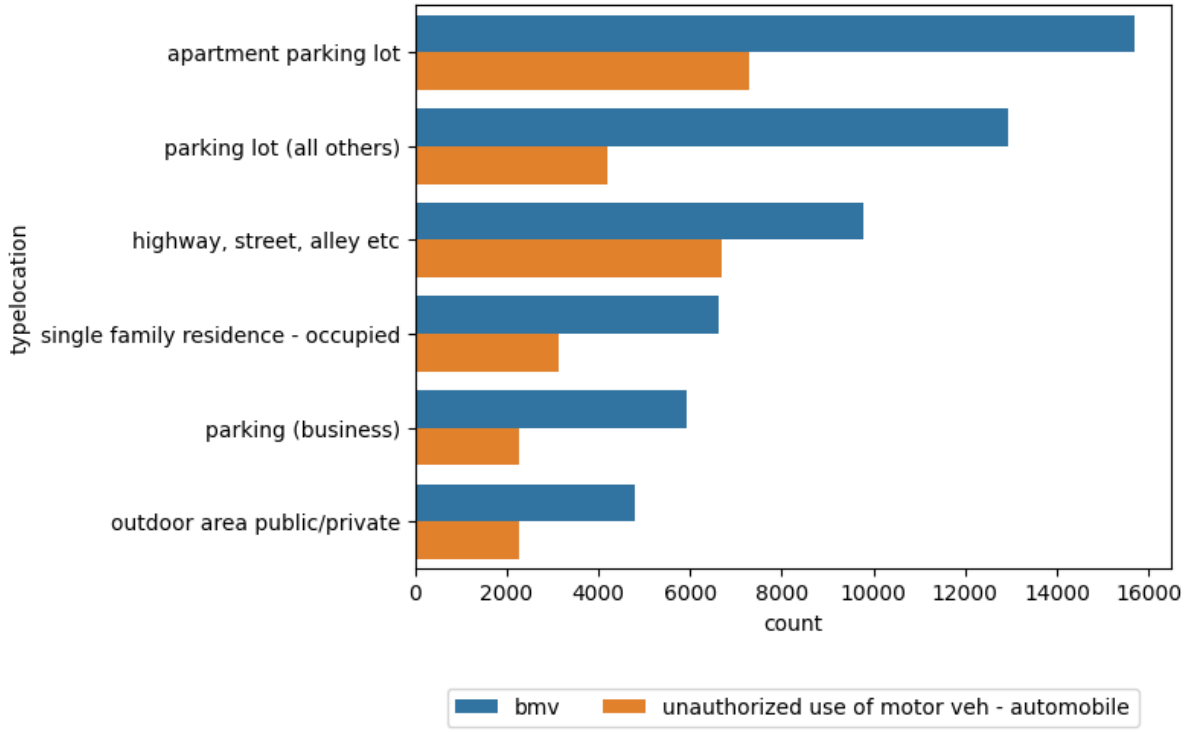
This graph shows the top five incidents and days of the week grouped together. Burglary of motor vehicle, unauthorized use of a motor vehicle, and found property occur similarly regardless of the day of the week, but burglary of habitation and public intoxication are opposite: burglary of habitation occurs mostly on weekdays and decreases on weekends, while public intoxication tends to occur more on weekends than on weekdays.



<Figure2> Top5 Incidents by Day

3.3 Location Type

Figure 3 is a graph comparing the most common vehicle-related crimes, burglary of a motor vehicle and unauthorized use of a motor vehicle, by type of location. Burglary of a motor vehicle and unauthorized use of a motor vehicle both occur mostly in parking lots and roads.



<Figure3> Location of Automobile Related Crimes

4. Suggestion

Based on the insights you obtained from the previous stages, propose the potential project idea.

Of the five most common incidents, burglary of motor vehicle, unauthorized use of a motor vehicle, and found property have similar trends over time (year, day). However, public intoxication and burglary of habitations have different trends, especially on weekends and weekdays. The burglary of motor vehicle and unauthorized use of a motor vehicle are all characterized by the fact that they occur mainly in apartment parking lots, highways, streets, and single family residential areas.