University of Tehran
2/15/2024

STATISTICAL INFERENCE

The Final Project

NAME OF PROJECT:
Denver Crime Data

PRESENTED BY:
Hadiseh Mesbah

Agenda

03	Overview
05	data cleaning
06	Vsualization
09	Hypotesis Test
11	Regresion
14	conclusion

01 This dataset includes criminal offenses in the City and County of Denver for the previous five calendar years and the current year. The data is based on the National Incident-Based Reporting System (NIBRS), which

02 includes all victims of personal crimes and all crimes within an incident. The data is dynamic, which allows for additions, deletions, and modifications at any time, resulting in more accurate

03 information in the database. Due to continuous data entry, the number of records in subsequent extractions is subject to change. Crime data is updated Monday through Friday.

Overview

Introduce the project. Provide a quick background and rationale. Briefly share its overall scope as well as expected outcomes.

Paper

A Comparative Study on Crime in Denver City Based on Machine Learning and Data Mining

Md. Aminur Rab Ratul, Faculty of Engineering, University of Ottawa, mratu076@uottawa.ca

https://arxiv.org/ftp/arxiv/papers/2001/2001.02802.pdf

Data Cleaning

How we deal with missing value!

1

Last Occurrence Date

Replace missing with corresponding values from 'first_occurrence_date'

2

Geo_x Geo_y
Goe_lon Geo_lat

Replace with district geo_location

3

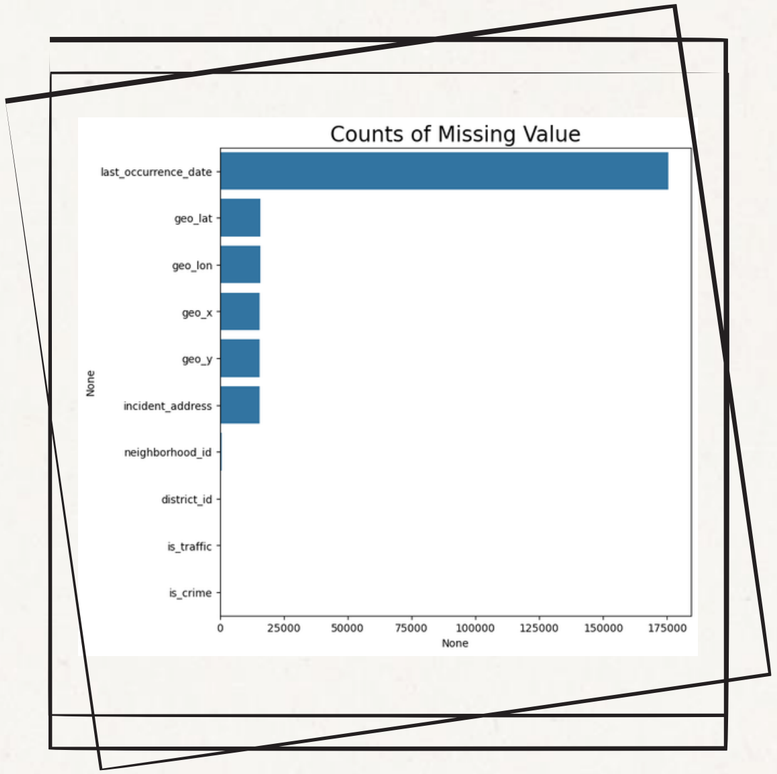
districtd_id

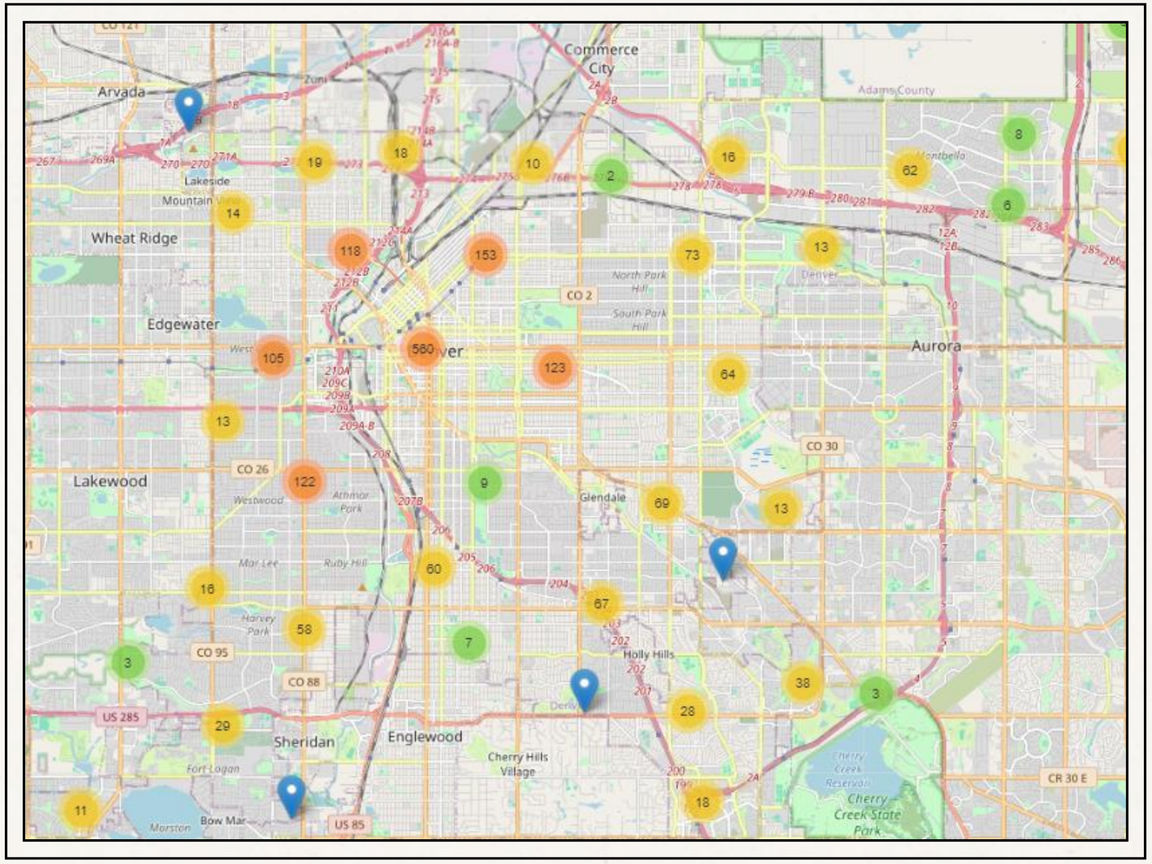
- with geo info
- with precinct_id
- drop

4

neighborhood_id and incident_adress

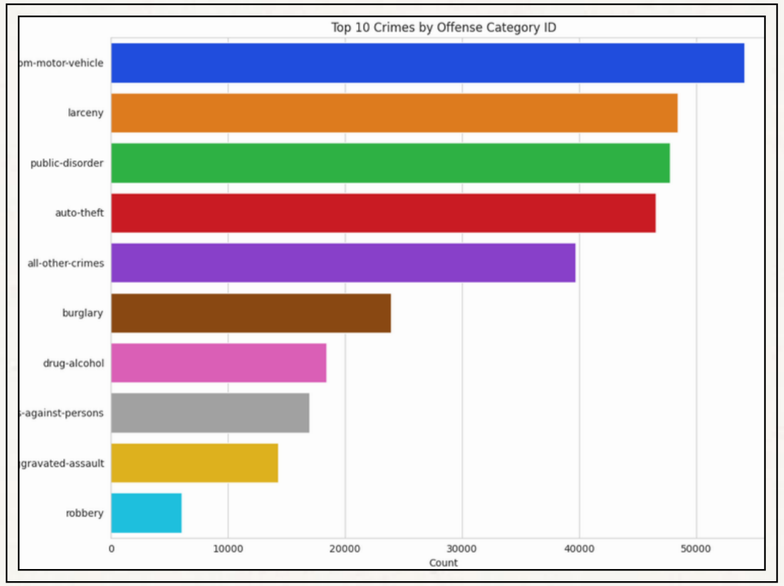
unknown and mix of available data and drop all 2023 data

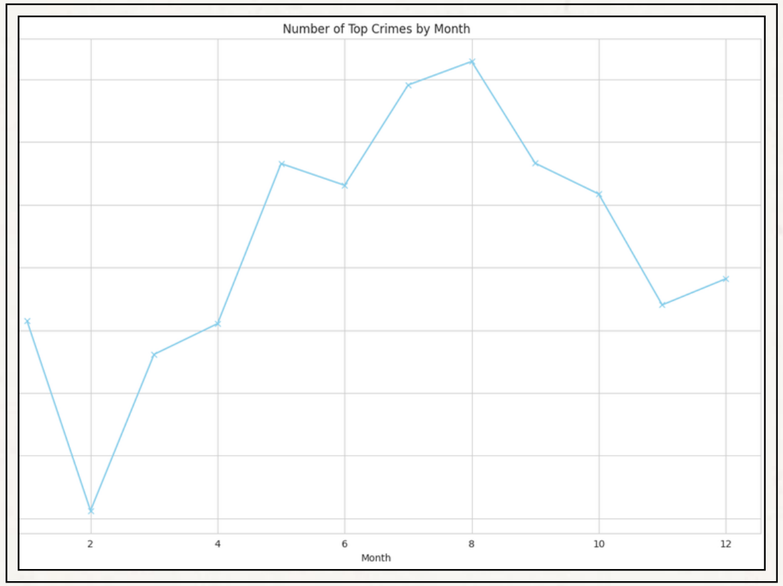


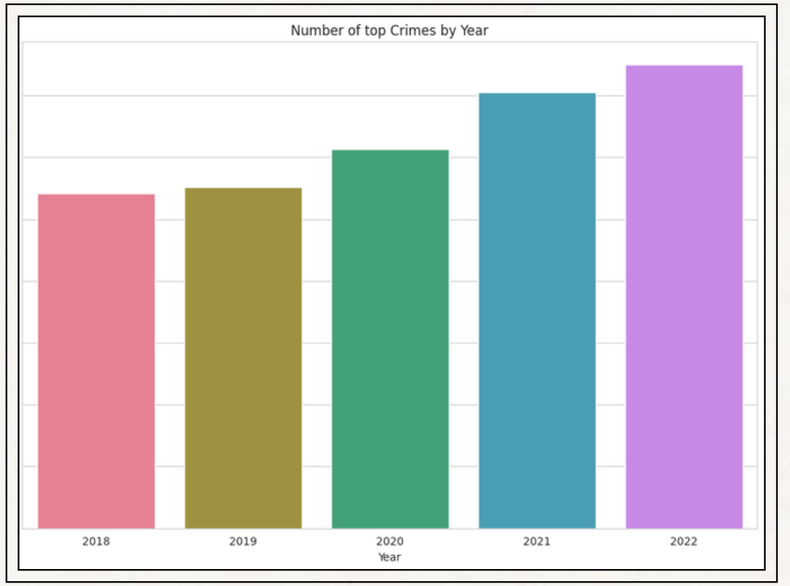


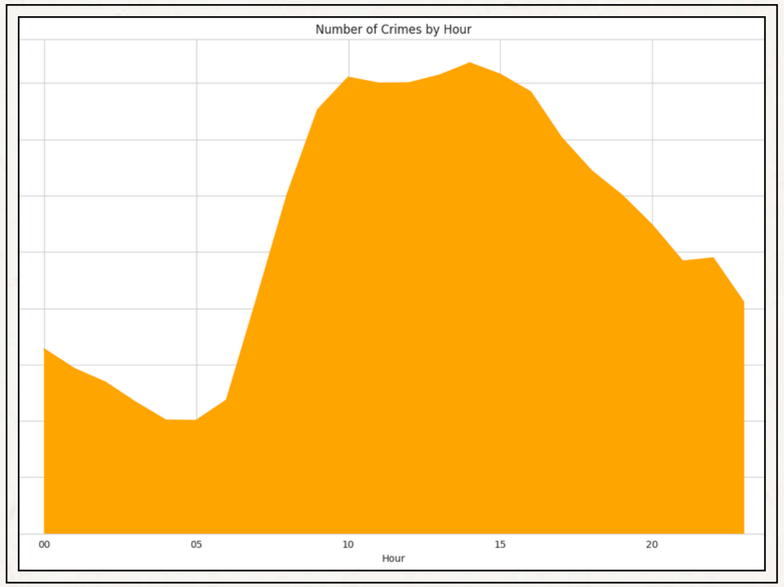
Visualization

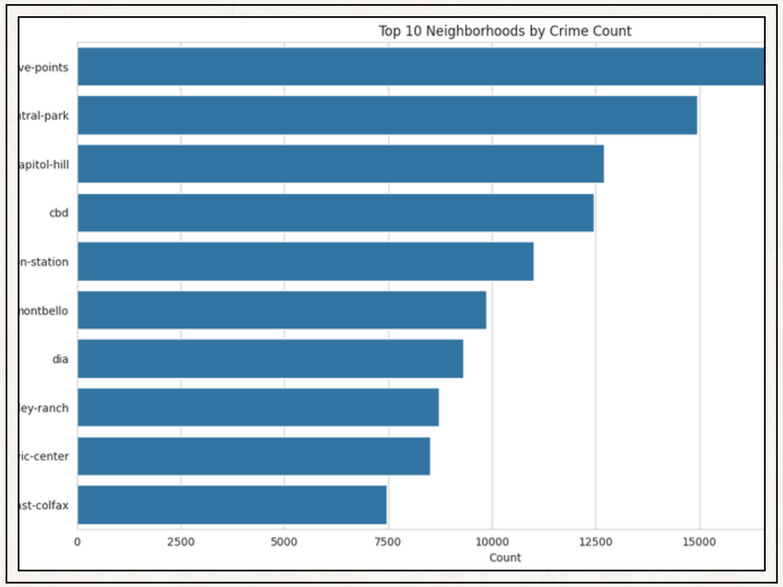
Draw some plots for data for better understanding.

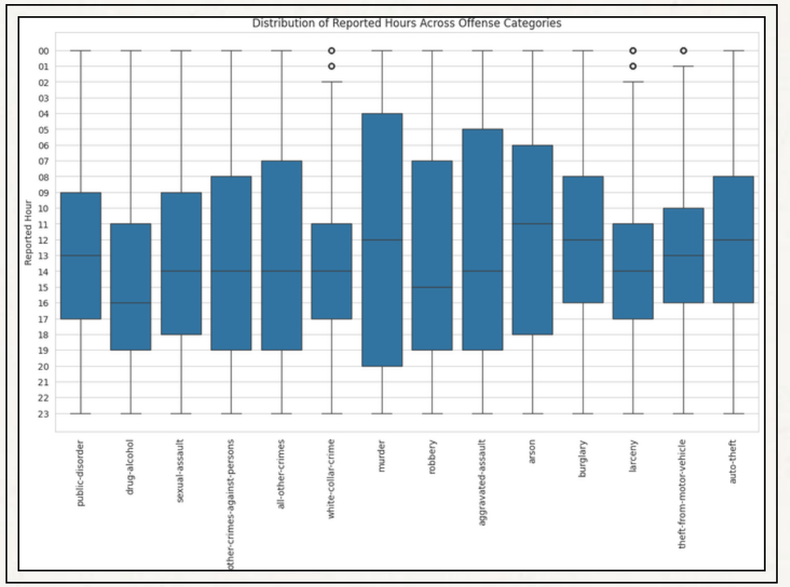

Top 10 Crimes by Offense Category ID
larceny
public-disorder
auto-theft
all-other-crimes
burglary
drug-alcohol
robbery
0
10000
20000
30000
40000
50000
Count

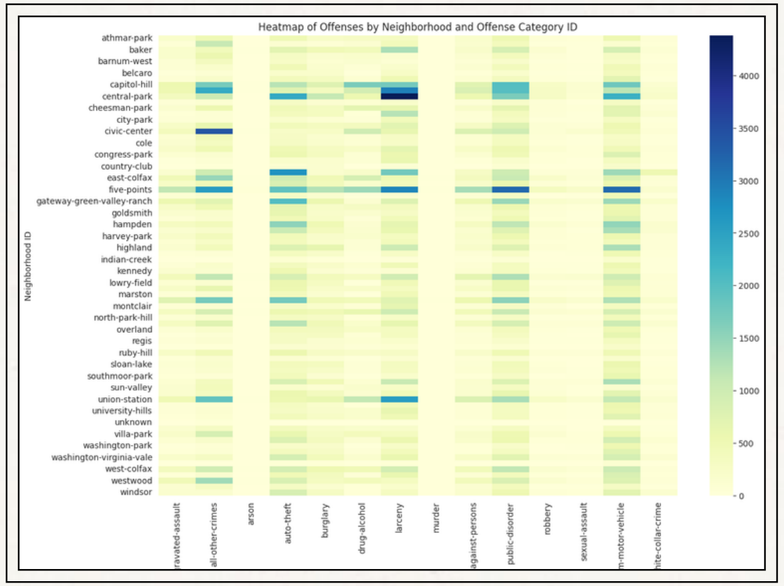

Number of Top Crimes by Month
2
4
6
8
10
12
Month

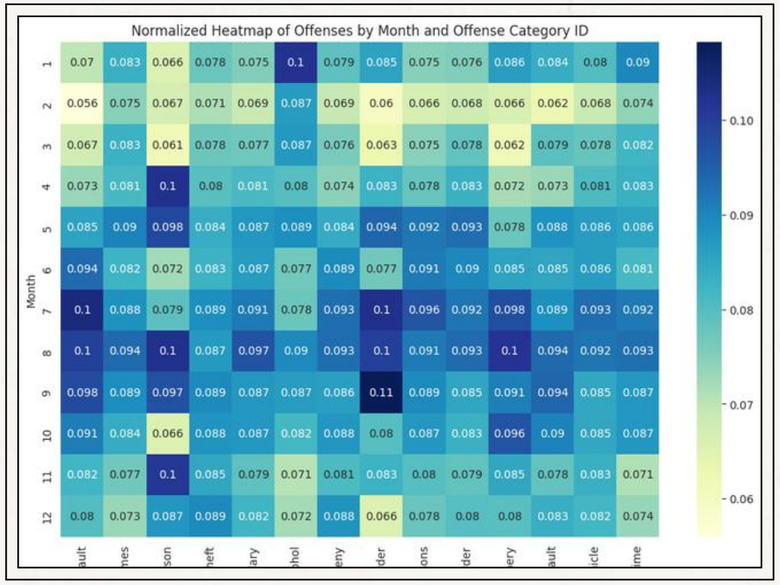

Number of top Crimes by Year
2018
2019
2020
2021
2022
Year

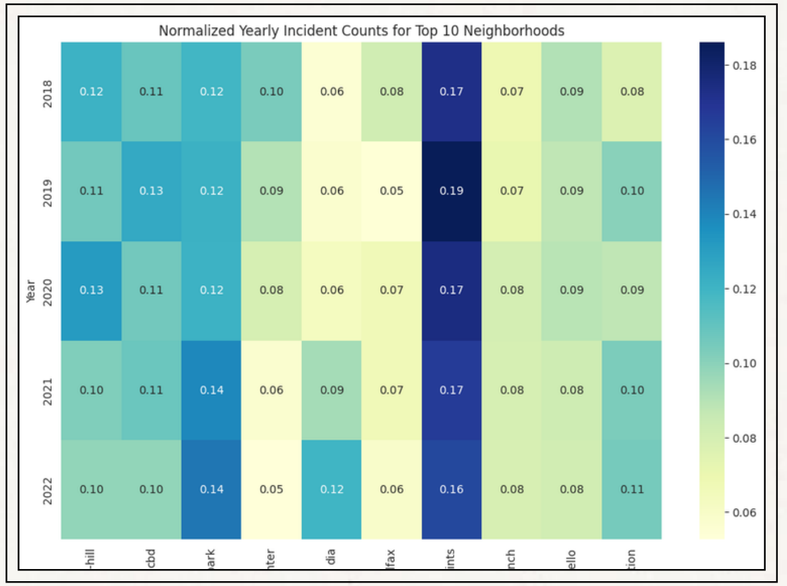

Number of Crimes by Hour
00
05
10
15
20
Hour

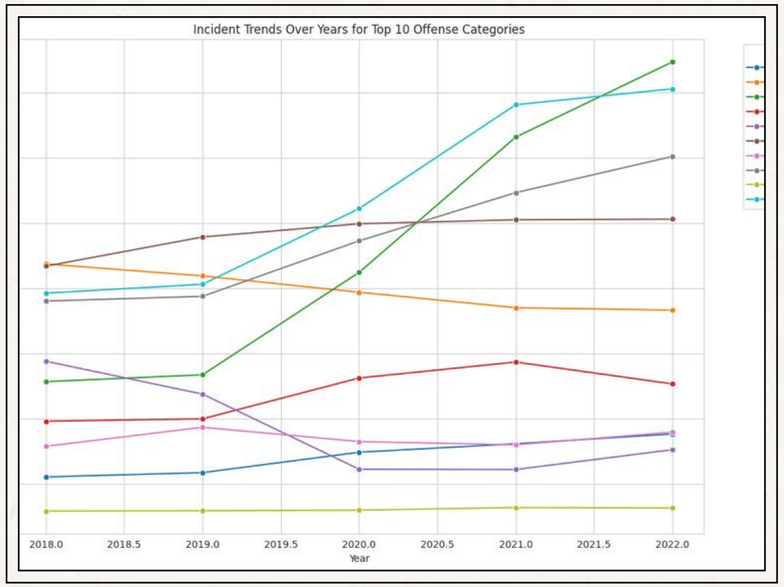

Top 10 Neighborhoods by Crime Count
0
2500
5000
7500
10000
12500
15000
Count

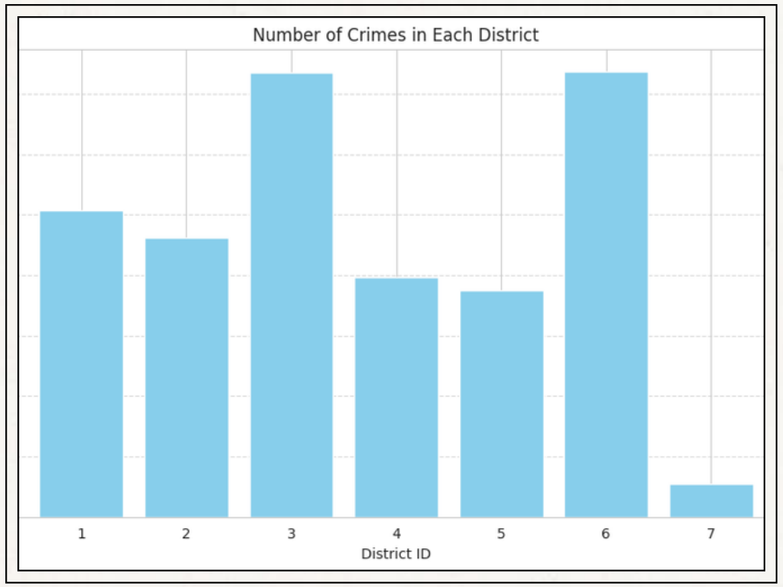

Distribution of Reported Hours Across Offense Categories
Reported Hour
public-disorder
drug-alcohol
sexual-assault
all-other-crimes
white-collar-crime
murder
robbery
arson
burglary
larceny
auto-theft

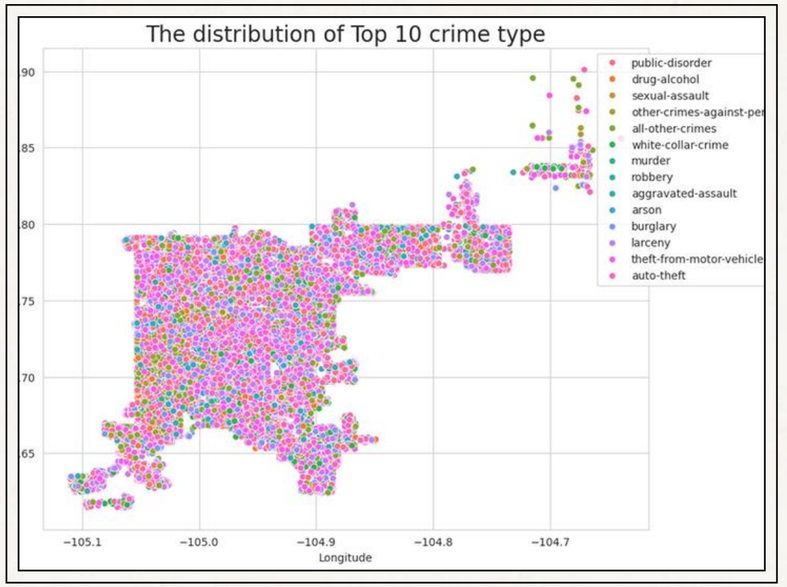

Heatmap of Offenses by Neighborhood and Offense Category ID
Neighborhood ID
athmar-park
baker
barnum-west
belcaro
capitol-hill
central-park
cheesman-park
city-park
civic-center
cole
congress-park
country-club
east-colfax
five-points
gateway-green-valley-ranch
goldsmith
hampden
harvey-park
highland
indian-creek
kennedy
lowry-field
marston
montclair
north-park-hill
overland
regis
ruby-hill
sloan-lake
southmoor-park
sun-valley
union-station
university-hills
unknown
villa-park
washington-park
washington-virginia-vale
west-colfax
westwood
windsor
aggravated-assault
all-other-crimes
arson
auto-theft
burglary
drug-alcohol
larceny
murder
other-crimes-against-persons
public-disorder
robbery
sexual-assault
theft-from-motor-vehicle
white-collar-crime
4000
3500
3000
2500
2000
1500
1000
500
0


Normalized Heatmap of Offenses by Month and Offense Category ID
Month
1 0.07 0.083 0.066 0.078 0.075 0.1 0.079 0.085 0.075 0.076 0.086 0.084 0.08 0.09
2 0.056 0.075 0.067 0.071 0.069 0.087 0.069 0.06 0.066 0.068 0.066 0.062 0.068 0.074
3 0.067 0.083 0.061 0.078 0.077 0.087 0.076 0.063 0.075 0.078 0.062 0.079 0.078 0.082
4 0.073 0.081 0.1 0.08 0.081 0.08 0.074 0.083 0.078 0.083 0.072 0.073 0.081 0.083
5 0.085 0.09 0.098 0.084 0.087 0.089 0.084 0.094 0.092 0.093 0.078 0.088 0.086 0.086
6 0.094 0.082 0.072 0.083 0.087 0.077 0.089 0.077 0.091 0.09 0.085 0.085 0.086 0.081
7 0.1 0.088 0.079 0.089 0.091 0.078 0.093 0.1 0.096 0.092 0.098 0.089 0.093 0.092
8 0.1 0.094 0.1 0.087 0.097 0.09 0.093 0.1 0.091 0.093 0.1 0.094 0.092 0.093
9 0.098 0.089 0.097 0.089 0.087 0.087 0.086 0.11 0.089 0.085 0.091 0.094 0.085 0.087
10 0.091 0.084 0.066 0.088 0.087 0.082 0.088 0.08 0.087 0.083 0.096 0.09 0.085 0.087
11 0.082 0.077 0.1 0.085 0.079 0.071 0.081 0.083 0.08 0.079 0.085 0.078 0.083 0.071
12 0.08 0.073 0.087 0.089 0.082 0.072 0.088 0.066 0.078 0.08 0.08 0.083 0.082 0.074
0.10
0.09
0.08
0.07
0.06


Normalized Yearly Incident Counts for Top 10 Neighborhoods
Year
2018 0.12 0.11 0.12 0.10 0.06 0.08 0.17 0.07 0.09 0.08
2019 0.11 0.13 0.12 0.09 0.06 0.05 0.19 0.07 0.09 0.10
2020 0.13 0.11 0.12 0.08 0.06 0.07 0.17 0.08 0.09 0.09
2021 0.10 0.11 0.14 0.06 0.09 0.07 0.17 0.08 0.08 0.10
2022 0.10 0.10 0.14 0.05 0.12 0.06 0.16 0.08 0.08 0.11
0.18
0.16
0.14
0.12
0.10
0.08
0.06


Incident Trends Over Years for Top 10 Offense Categories
2018.0
2018.5
2019.0
2019.5
2020.0
2020.5
2021.0
2021.5
2022.0
Year


Number of Crimes in Each District
1
2
3
4
5
6
7
District ID


The distribution of Top 10 crime type
90
85
80
75
70
65
−105.1
−105.0
−104.9
−104.8
−104.7
Longitude
public-disorder
drug-alcohol
sexual-assault
other-crimes-against-per
all-other-crimes
white-collar-crime
murder
robbery
aggravated-assault
arson
burglary
larceny
theft-from-motor-vehicle
auto-theft

Hypothesis Tests

Testing some gusess!

Test #1

The Chi-square test
neighborhood_id and offense_category_id
Reject there is no association between the two variables (null hypothesis)

Test #2

Pearson's correlation coefficient.
Reject the correlation between the variables is zero (no linear relationship).

Test #3

The one-way ANOVA test
fails to reject all group means as equal.
District location does not significantly impact the number of victims in reported incidents

Test #4

ANOVA test
fails to reject all group means that are equal.
Based on the victim count data, neighborhood location does not significantly influence the number of victims in reported incidents

Test #5

one-way ANOVA test
Reject H0: There is significant evidence that the mean victim count is different across offense categories

Test #6

t-test
Fail to reject H0: There is no significant evidence that the mean reported hour of crimes is different between weekdays and weekends

Test #7

independent samples t-test
Fail to reject H0: There is no significant evidence that the mean number of reported crimes is different between summer months and winter months

Test #8

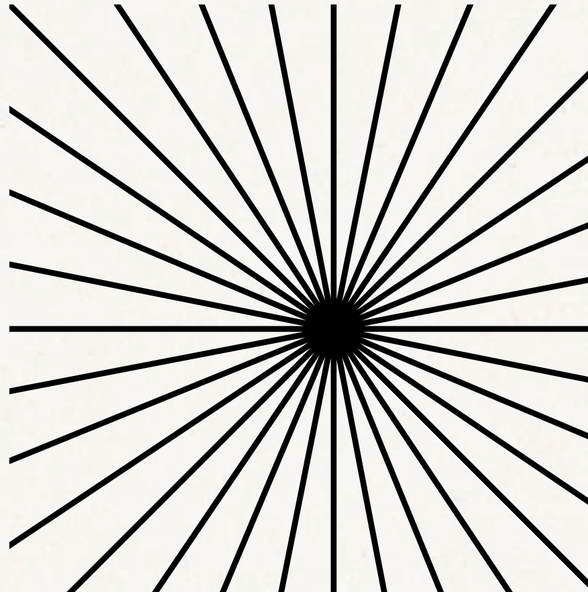
one-way ANOVA test
Fail to reject H0: There is no significant evidence that the mean reported hour of crimes is different across neighborhoods

Test #9

t-test
is that there is no significant evidence to suggest a difference in the latitudinal locations of auto thefts versus white-collar crimes

Test #10

ANOVA test followed by a Tukey's HSD
Reject H0: There is significant evidence that the mean victim count per crime is different across offense types.

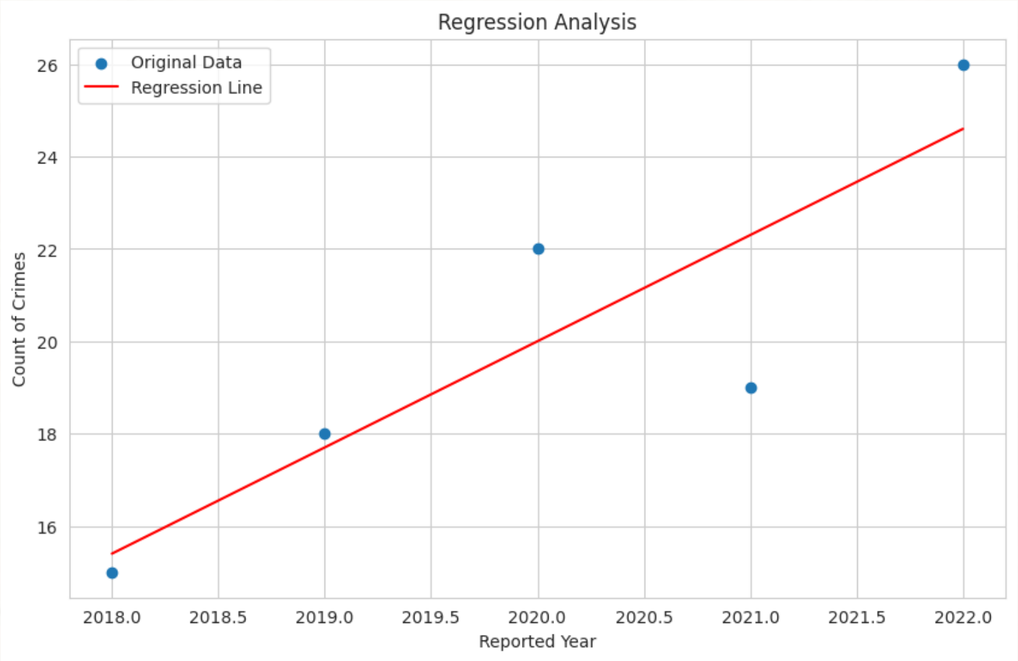


Regression

We do 8 different regressions, but let's see 3 of them!

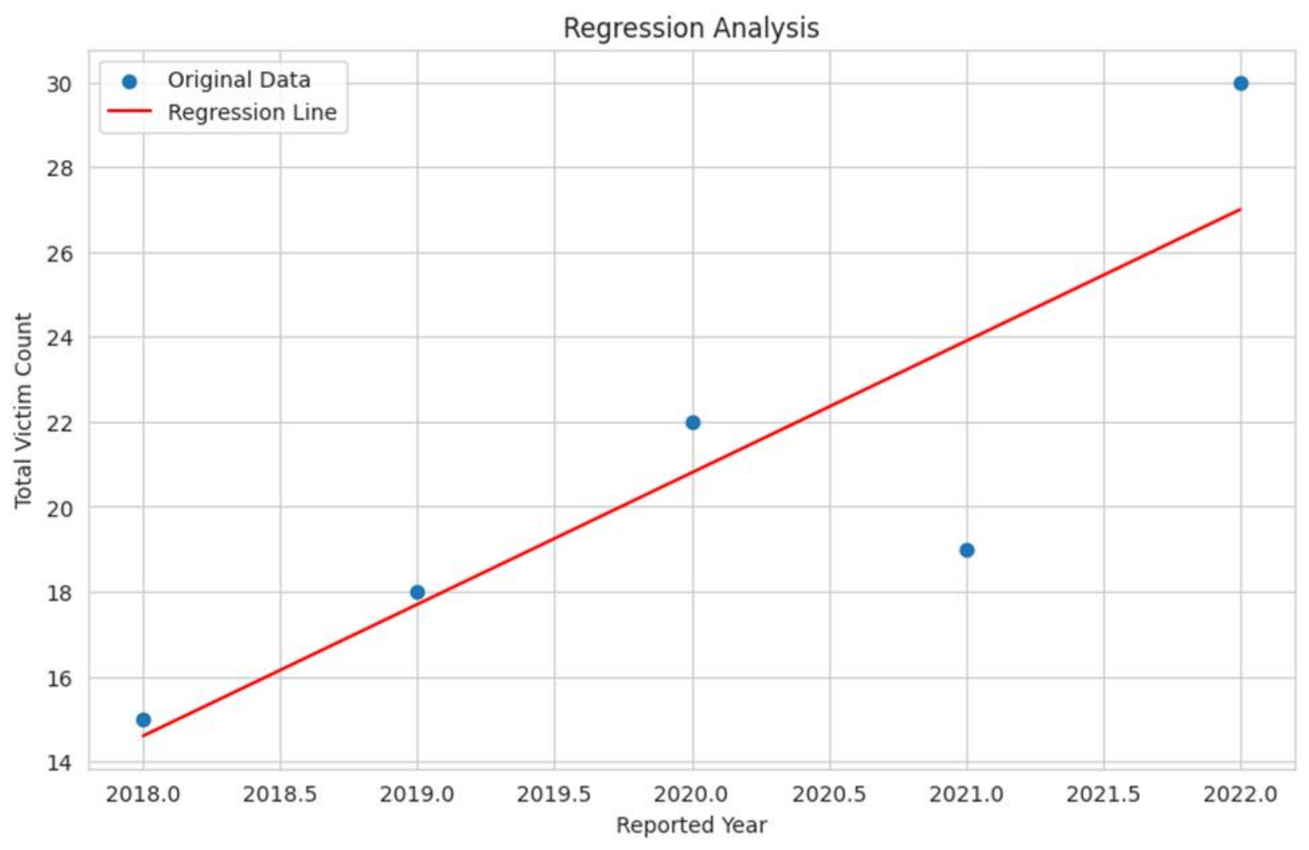
1st

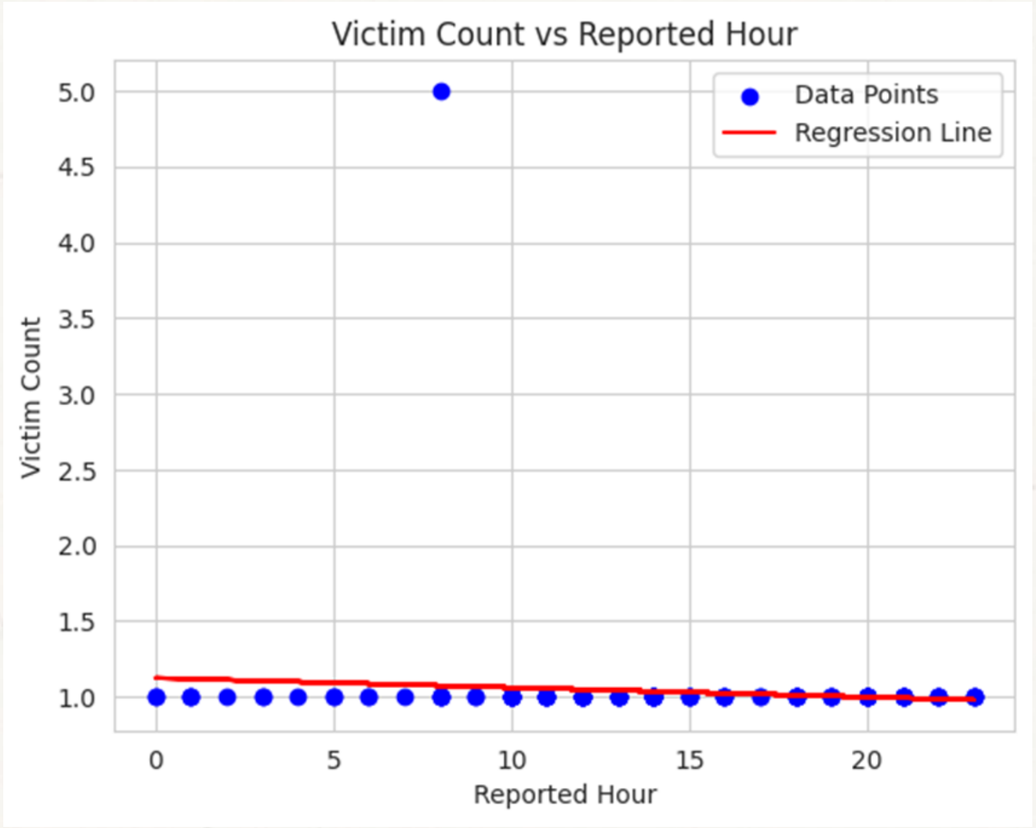
The regression model has an R-squared of 0.756, indicating that the year explains approximately 75.6% of the variability in the crime count.



2nd

The Jarque-Bera test yields a p-value of 0.670, indicating that the residuals could be normally distributed.





3rd

The R-squared is 0.009, indicating that only 0.9% of the variance in the victim count is explained by the reported hour, which is very low.

Thank you

E-mail hadiseh.mesbah@ut.ac.ir

Student ID 810102253

Professor Dr. Mohammadreza A. Dehaqani