Crime Rate Analysis in Los Angeles Using Predictive Modeling And Visualization

Dataset used: https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8
We used the Los Angeles Crime Data from 2020 to the present(csv format). The LAPD owns this data and has 608,592 records with 28 attributes. The attributes include a unique record number(DR_NO), the date and time the crime was reported and occurred, location details(area codes, address, latitudes, longitudes, etc.), victim details, weapon and crime details, the status of the case, etc. This data has multiple blank/unknown entries which would be dealt with in the pre-processing stage.

MODEL

We created a crime database that uses the csv data. The record number (DR_NO) serves as the primary key for the tables. We divided the data into six tables:

1. CrimeDetails
2. CrimeType
3. VictimData
4. CrimeStatus
5. CrimeLocation
6. AdditionalCrimeCodes

CrimeDetails

DR_NO	INT NOT NULL PK
CrimeCode	INT NOT NULL
CrimeCodeDesc	TEXT NOT NULL
Weapon	TEXT
WeaponDesc	TEXT

CrimeDates

DR_NO	INT NOT NULL PK
DateReported	DATE NOT NULL
DateOccured	DATE NOT NULL
TimeOccured	INT NOT NULL

VictimData

DR_NO	INT NOT NULL PK
VictimAge	INT
VictimSex	TEXT
VictimDescent	TEXT

CrimeLocations

DR_NO	INT NOT NULL PK
AreaCode	INT NOT NULL
AreaName	TEXT NOT NULL
DistrictNumber	INT NOT NULL
PremiseCode	INT
PremiseDesc	TEXT
Location	TEXT NOT NULL
CrossStreet	TEXT
Latitude	FLOAT NOT NULL
Longitude	FLOAT NOT NULL

CrimeStatus

DR_NO	INT NOT NULL PK
Status	INT NOT NULL
StatusDesc	TEXT NOT NULL

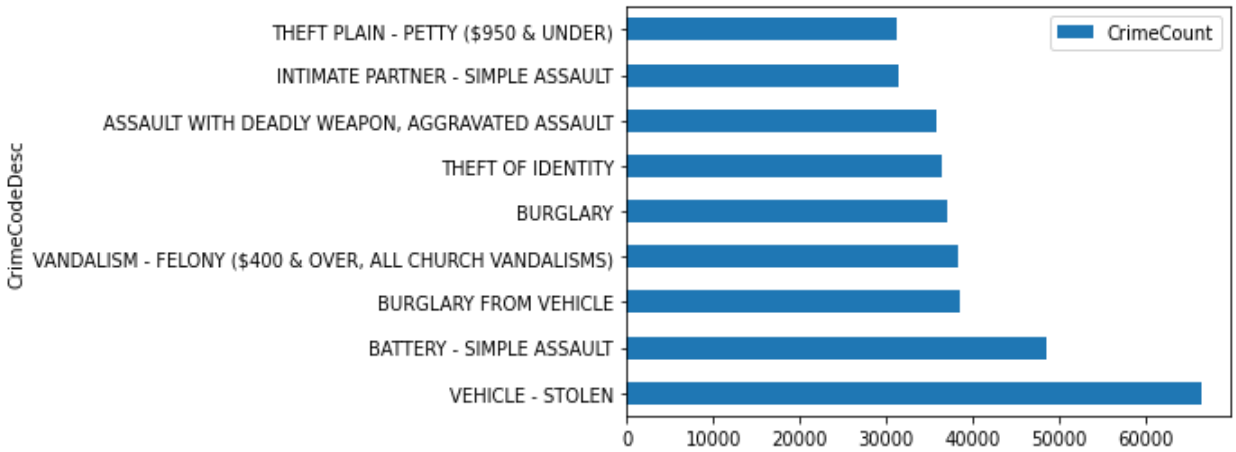
AdditionalCrimeCodes

DR_NO	INT NOT NULL PK
CrimeCode	INT NOT NULL
CrimeCode1	INT
CrimeCode2	INT
CrimeCode3	INT
CrimeCode4	INT

ANALYSIS

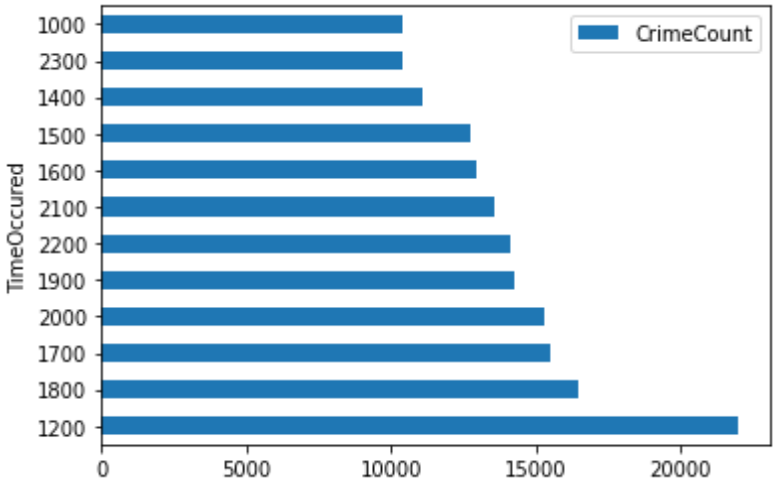
1. Data analysis by plotting using bar plots.

- Plot: CrimeCount v/s CrimeCodeDesc



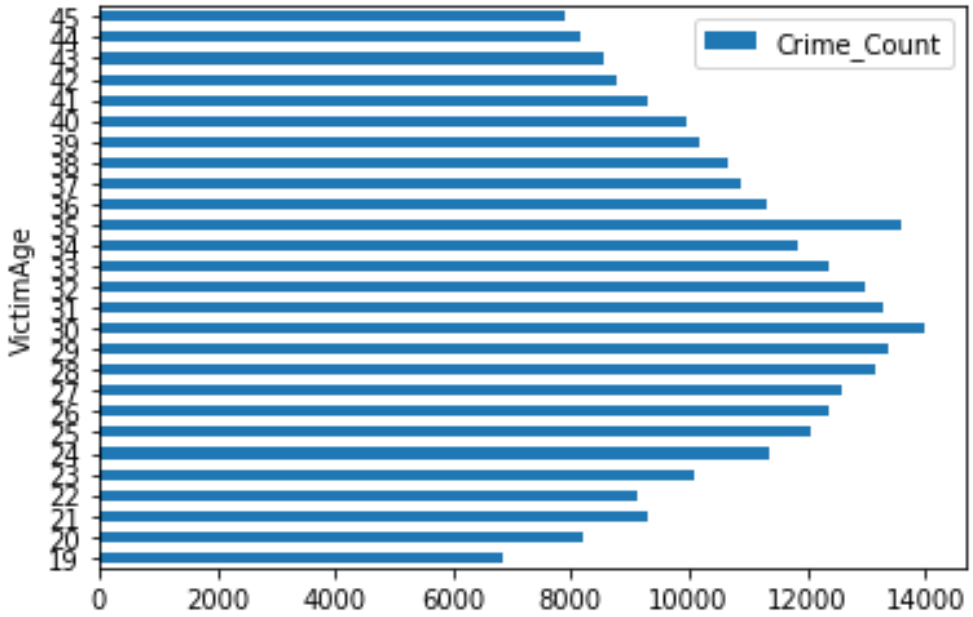
Inference: Stolen vehicles are reported the most.

- Plot: CrimeCount v/s TimeOccured



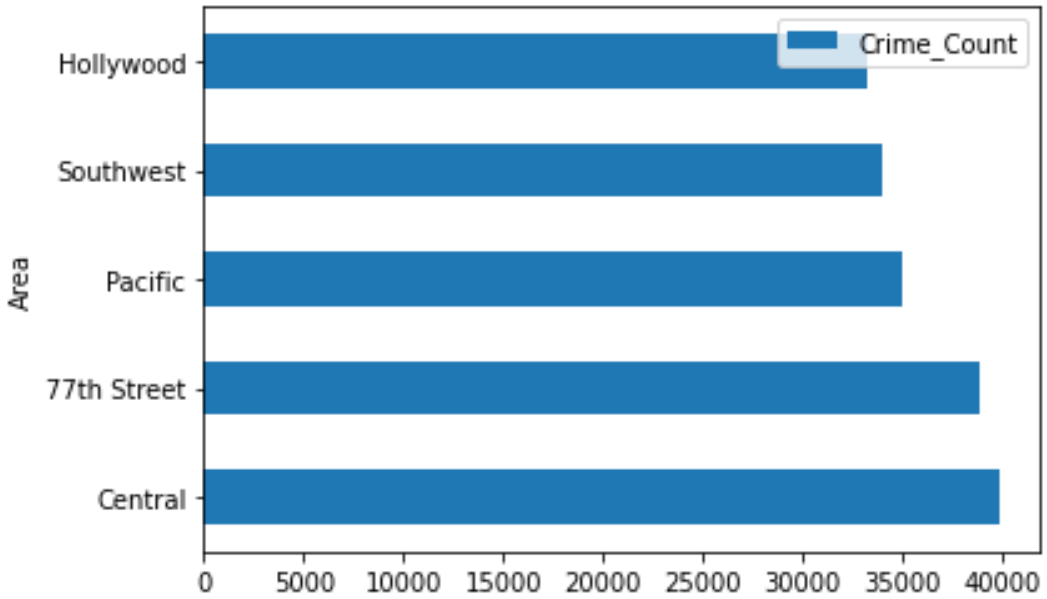
Inference: Maximum number of crimes are reported in the evenings.

- Plot: CrimeCount v/s VictimAge



Inference: People of age between 30 and 35 are affected the most. It peaks at 30 and then keeps decreasing.

- Plot: CrimeCount v/s Area



Inference: Central is the area where the most number of crimes have been reported.

2. Feature correlation.

- Used matplotlib, seaborn, and numpy.
- Converted VictimDescent and VictimSex data from categorical to numeric.

- Dropped the rows which have None and NA values.
- Used Pearson's correlation to find the correlation between features: CrimeCode, Weapon, TimeOccured, AreaCode, DistrictNumber, PremiseCode, VictimAge, VictimSex, and VictimDescent.

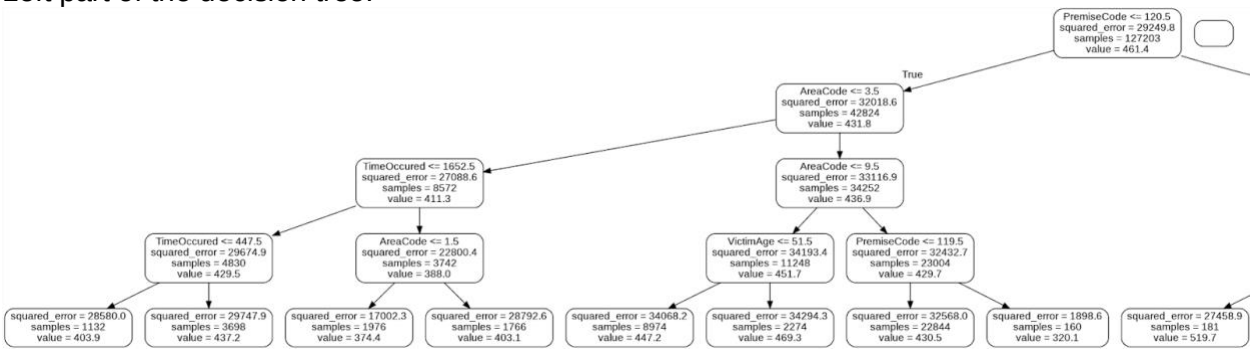
- Inferences:
  - CrimeCode is mostly correlated with PremiseCode.
  - District number and Victim age are the least correlated.

3. Random Forest Classification.

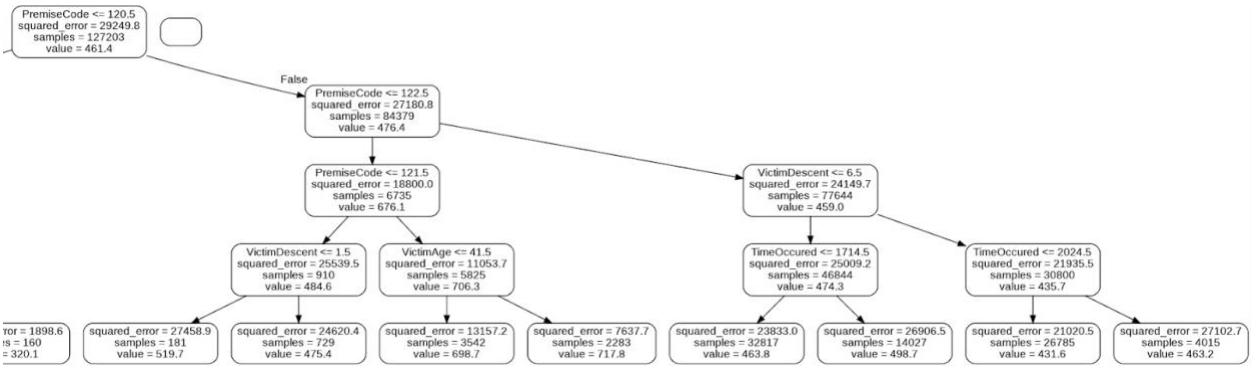
- Created a classification model utilizing the random forest classifier to generate a decision tree that categorizes crimes based on AreaCode, PremiseCode, TimeOccured, VictimAge, and VictimDescent.

- These steps were carried out using the sklearn package.
- The steps carried out for classification are as follows:
    - Collect the data for certain highly occurring CrimeCodes- 510, 624, 330, 740, 310, 354, 230, 626, 440.
    - Extract the attributes using JOIN between CrimeDetails, CrimeLocations, CrimeDates, VictimData tables.
    - Convert categorical VictimDescent data to numeric.
    - Split the train:test data by ratio 75:25.
    - Fit the random forest model and calculate mean absolute error and accuracy.
- Result: Achieved an accuracy of 64.13%

Left part of the decision tree:



Right part of the decision tree:

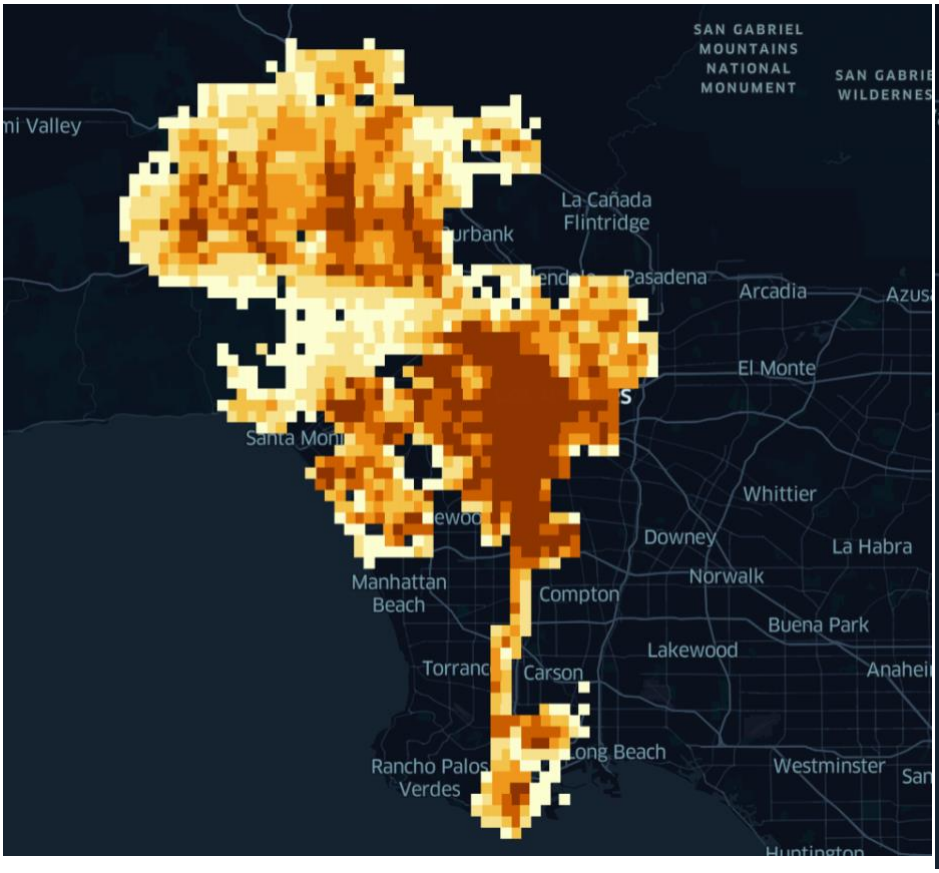


VISUALIZATION

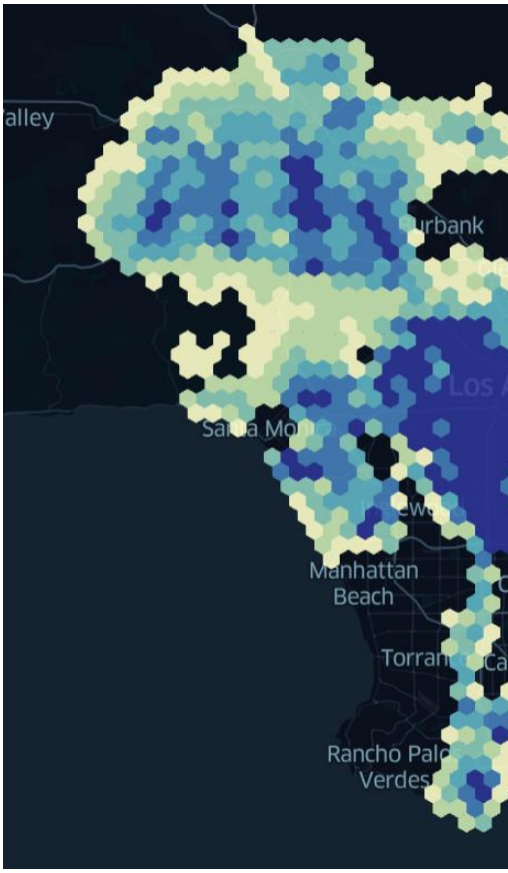
- Used Keplergl for geospatial visualization of crimes in LA city.
- Features:
    - Uses latitude and longitude to build visualizations on an interactive map.
    - Displays the crimes as heatmap, hexbin, grids, and points.
    - Added the date on which the crime occurred to display the crimes happening in a time playback video.
    - Also displays the crime description, and time of crime when hovered over. Other labels can be added as well.

- Can visualize the change in crime intensity over a period of time at any place using the time playback video and filters.

Cluster map



Heat Map



Hexbin Map

RESULTS

- CrimeCode is correlated more with PremiseCode.
- Stolen vehicles are reported the most.
- The maximum number of crimes are reported in the evenings.
- People of ages 30 and 35 are affected the most. It peaks at 30 and then keeps decreasing.
- Central is the area where the most number of crimes have been reported.
- We could recognize the areas with the highest number of crimes which would help to devise ways to reduce violence in these areas.

CONCLUSION

The inferences derived can be used to understand crime distribution to a large extent and list out which people and places need more support to fight crime. Random forest classification can be used to predict the crime codes that can occur in a particular time or place and are strong indicators to identify patterns. Once there is an understanding of how these patterns work, more police forces can be dispersed to places where there is a crime prediction.