Abstract:
The project titled "A Comprehensive Study on Crime Rates and Trends" involves an analysis and interpretation of various socio-economic, environmental, and demographic trends in the city using data-driven techniques. This study aims to demarcate the major patterns that define living in Los Angeles through the employment of various public and private data sources that include census reports, economic indicators, environmental metrics, and social media activity. It will apply various methods of Data Science, including statistical analysis, machine learning, and data visualization to shed light on such topical issues as urban development, air quality, criminality, traffic congestion, and population shift. The current research will, therefore, be in a position to outline the main trends by comparing historical data and by predicting futuristic scenarios, hence making it highly valuable for the needs of policymakers, business, and residents alike. This research work, therefore, intends to point out challenges and opportunities that surround growth and sustainability of the city. The outcome of this project will provide real policy insights that could help in better urban planning, environmental sustainability, and economic development in Los Angeles.

DATA SET:

The dataset for this crime analysis project was obtained from the Los Angeles Police Department (LAPD) website *https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8/about_data.* The data reflects incidents of crime in the City of Los Angeles dating back to 2020. The LAPD transcribes this data from original crime reports, which are initially written on paper and then manually entered into the system. This data is released publicly by the LAPD to allow for transparency and analysis of crime patterns within the city.

It's important to note that starting from March 7th, 2024, the LAPD will be transitioning to a new crime reporting system that complies with the FBI's National Incident-Based Reporting System (NIBRS). This shift will improve the granularity and accuracy of crime data moving forward.

However, due to manual entry from paper reports, some inaccuracies or missing data may be present in the dataset. For instance, some location fields with missing data are noted as (0°, 0°), and addresses are anonymized to the nearest hundred block for privacy reasons.

Data cleaning:

The Data Cleaning has been done by Handling Missing values, Encoding, Standardization, Dimensionality reduction, removing outliers.

Code snippet:

```
crime_data['DATE OCC'] = pd.to_datetime(crime_data['DATE OCC'])
# Filter data for November 2023
nov_2023_data = crime_data[(crime_data['DATE OCC'] >= '2024-01-01') & (crime_dat

nov_2023_data = nov_2023_data.sort_values(by='DATE OCC')

# Save the filtered data to a new CSV file
nov_2023_data.to_csv('crimedata_2024.csv', index=False)

print("November 2023 data extracted and saved as 'crime_data_nov_2023.csv'.")
```

In [ ]:
```
# Check for missing values in the dataset
missing_values = df.isnull().sum()
```



```
# Display columns with missing values
print(missing_values[missing_values > 0])

# Unwanted Columns
unwanted_columns = ['DR_NO', 'Mocodes', 'Cross Street', 'Crm Cd 1', 'Crm Cd 2',

# Drop unwanted columns
crime_data_cleaned = df.drop(columns=unwanted_columns)

# Verify the cleaning process
crime_data_cleaned.info()
```

```
Mocodes            24864
Vict Sex           24181
Vict Descent       24184
Premis Cd              4
Premis Desc           52
Weapon Used Cd     86537
Weapon Desc        86537
Status                 1
Crm Cd 2          104672
Crm Cd 3          109412
Crm Cd 4          109544
Cross Street       95754
```

Replacing the numerical columns with the mean of the data i.e., in this case we are replacing the missing values in the 'Vict Age' column with the mean of the column:

In [ ]:
```
crime_data_cleaned['Vict Age'].mean()
```

Out[ ]:
```
23.723905939057566
```

Replacing the null values from the categorical columns 'Vict Sex' and 'Vict Descent' with the mode of the data

In [ ]:
```
categorical_columns = ['Vict Sex', 'Vict Descent','Premis Cd']
for column in categorical_columns:
    crime_data_cleaned[column].fillna(crime_data_cleaned[column].mode()[0], inpl
```

Replacing the missing values from the other columns based on the analysis and context:

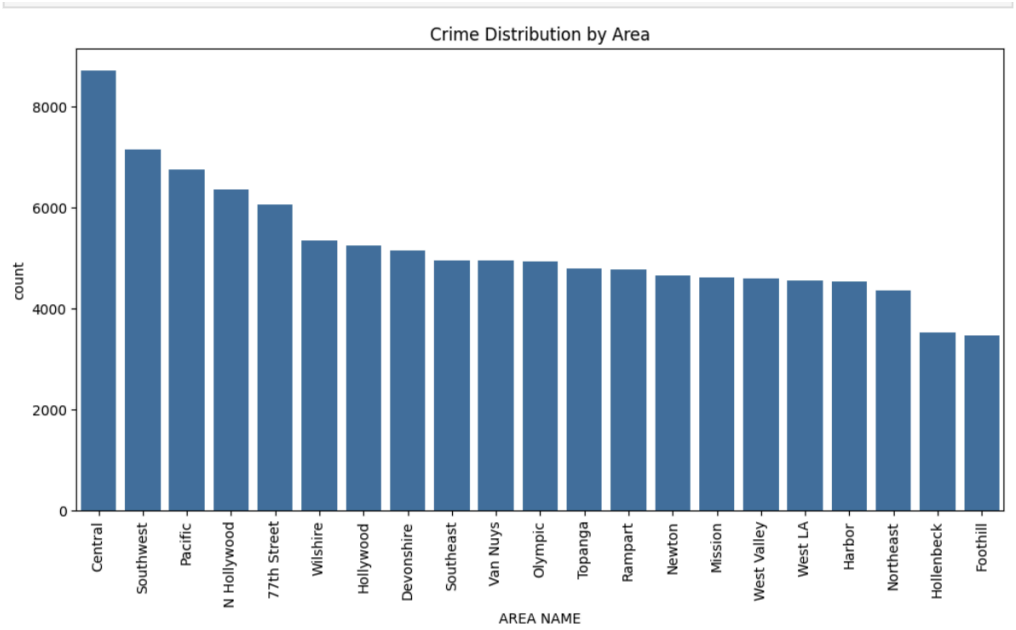
In [ ]:
```
# Replace missing values in 'Weapon Desc' with "Unknown"
crime_data_cleaned['Weapon Desc'].fillna('Unknown', inplace=True)

# Replace missing values in 'Weapon Used Cd' with "No Weapon Used"
crime_data_cleaned['Weapon Used Cd'].fillna("No Weapon Used", inplace=True)

# Replace missing values in 'Premis Desc' with a placeholder like "Unknown Premi
crime_data_cleaned['Premis Desc'].fillna("Unknown Premise", inplace=True)

# Replace missing values in 'Premis Desc' with a placeholder like "Unknown Premi
crime_data_cleaned['Status'].fillna("Unknown Premise", inplace=True)
```

In [ ]:
```
crime_data_cleaned.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 109546 entries, 0 to 109545
Data columns (total 21 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   Date Rptd       109546 non-null  object
 1   DATE OCC        109546 non-null  object
```

Data visualization on crime data set:

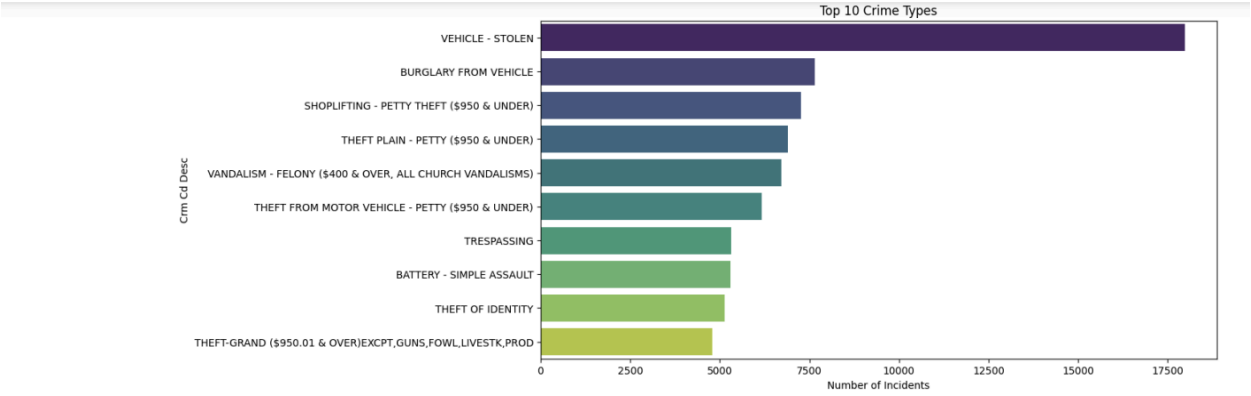
1.Crime distribution by area:



The bar chart shows the distribution of crime incidents across various areas, with the Central area having the highest crime count (over 1400 incidents). Other areas like 77th Street and

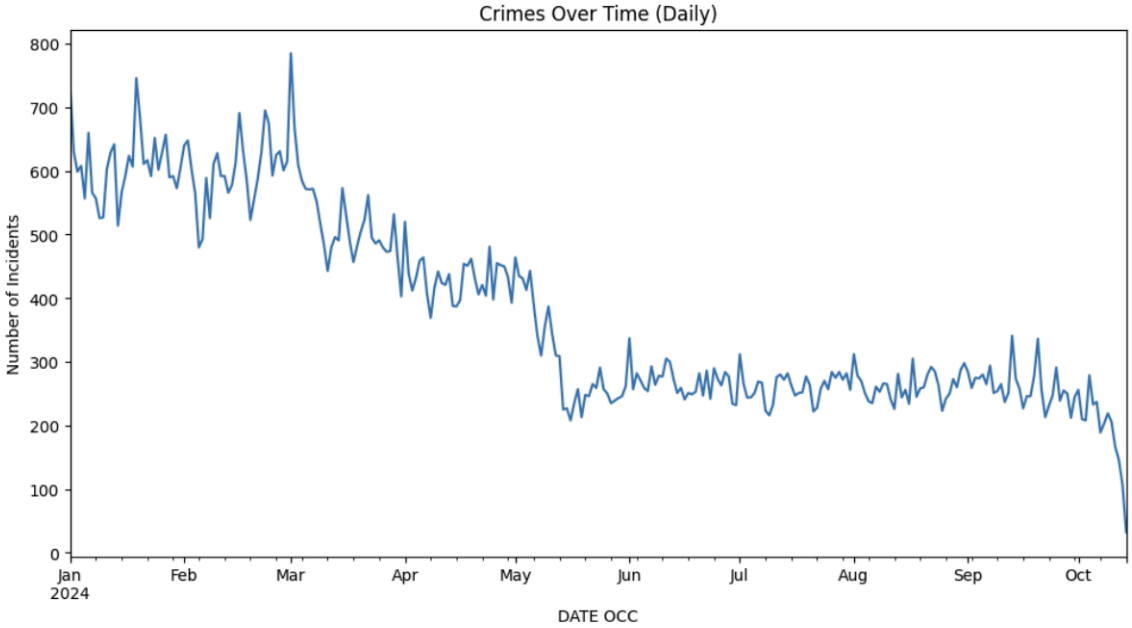
Southwest also report high crime rates, each with over 1000 incidents. The areas with the lowest crime counts include Foothill, Hollenbeck, and Harbor, indicating fewer reported incidents in these regions. The chart highlights significant variation in crime distribution across different areas.

2.top 10 crime types:



The bar chart shows the Top 10 Crime Types by number of incidents, with Vehicle Stolen being the most frequent, followed by Battery - Simple Assault and Burglary from Vehicle. The chart uses a gradient color scheme to differentiate the categories, making it clear that vehicle-related crimes are the most common. Incidents range from around 500 to over 2000

3. Crime over times

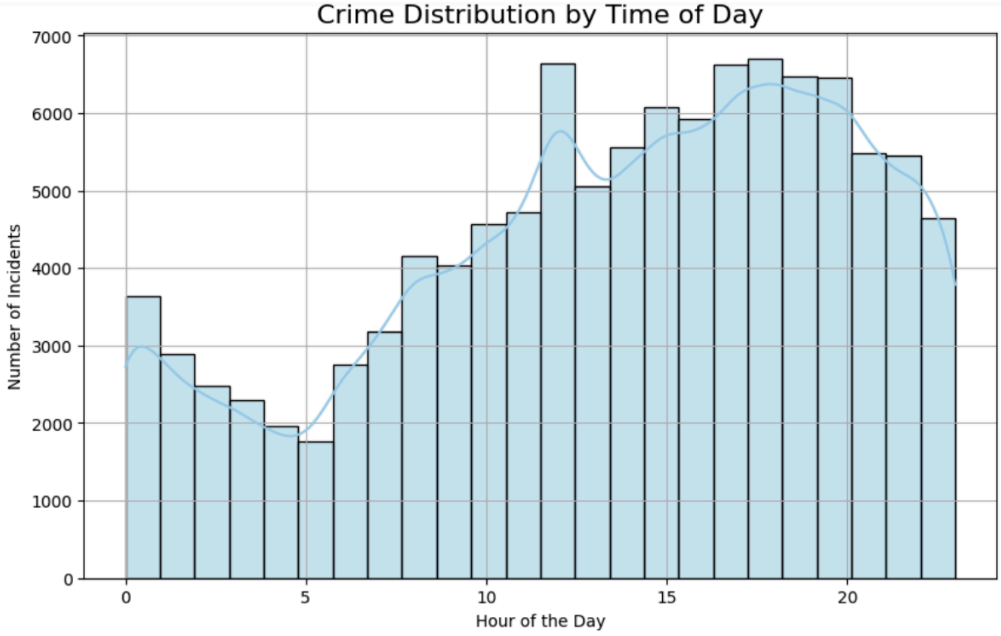


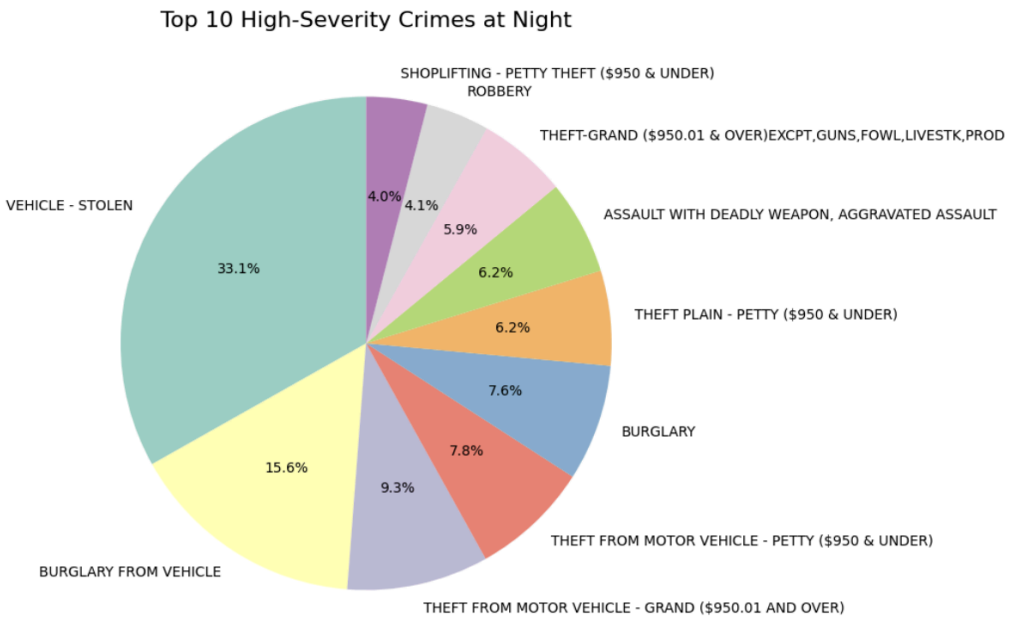
The line chart shows the Crimes Over Time (Daily) for November 2023. The Y-axis represents the number of incidents, ranging from approximately 500 to 675. The X-axis displays the dates throughout November. Crime rates fluctuate throughout the month, with noticeable peaks and

drops. The highest number of incidents occurs at the start of the month, and a general downward trend is observed toward the end.

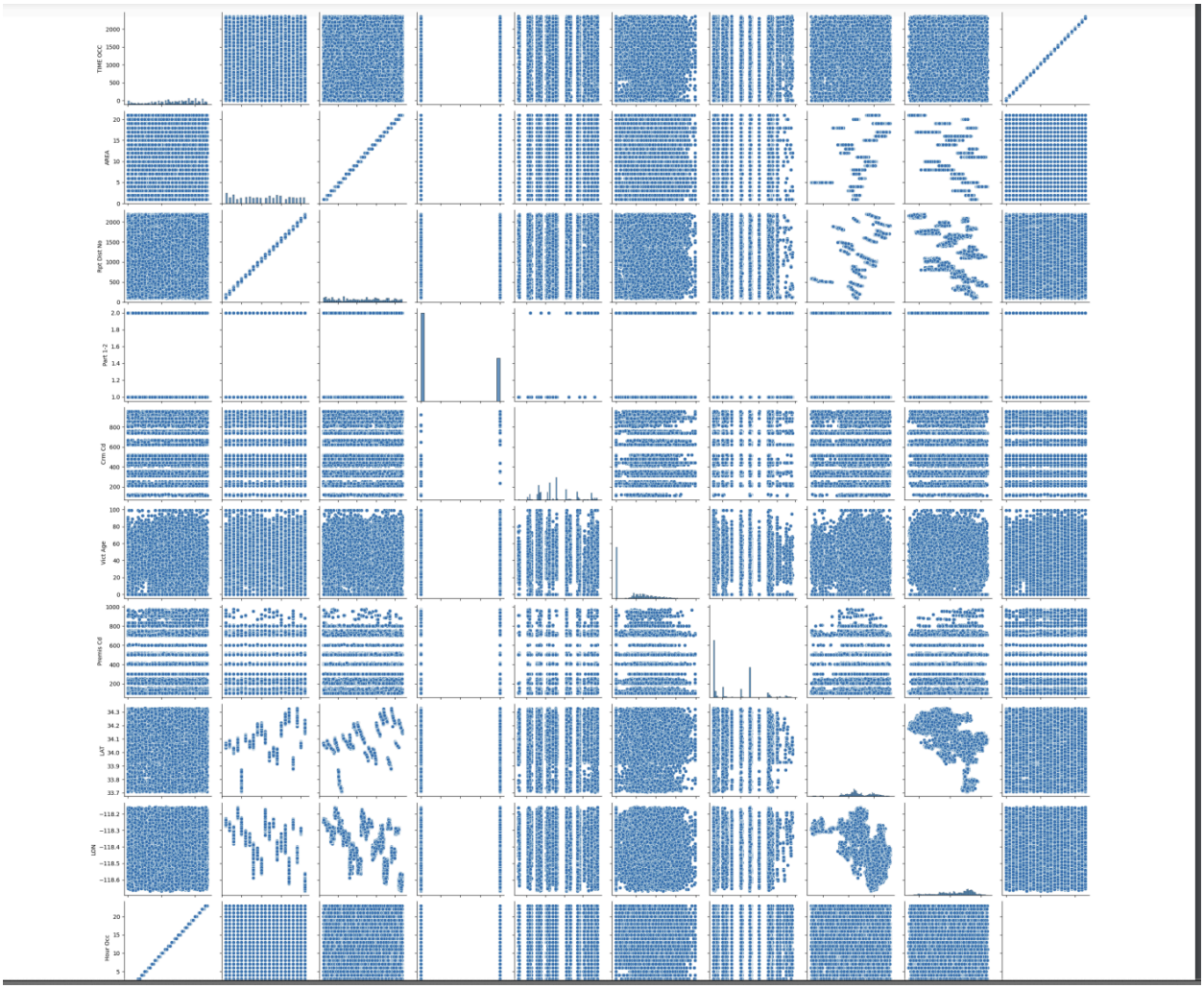
4**. Crime Distribution by time of day**

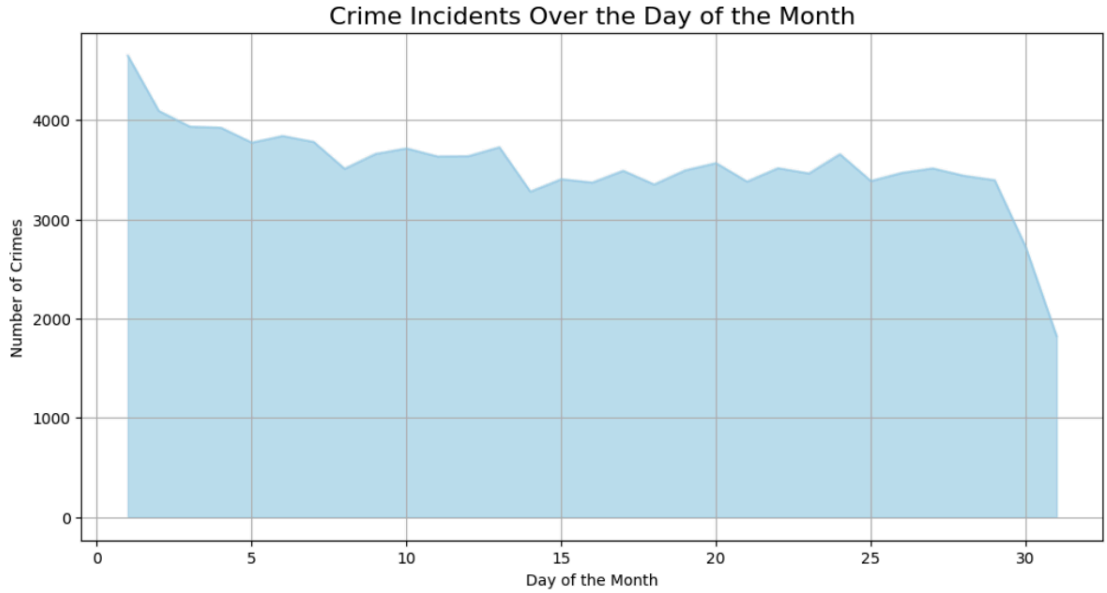
The histogram shows the Crime Distribution by Time of Day, with the X-axis representing the hour of the day (from 0 to 24), and the Y-axis indicating the number of incidents. Crimes are more frequent between 12 PM and 8 PM, peaking around 3 PM. There is a noticeable dip in crime activity between 2 AM and 6 AM, indicating fewer incidents during the early morning hours. The graph also includes a KDE (Kernel Density Estimation) curve to smooth out the distribution pattern.



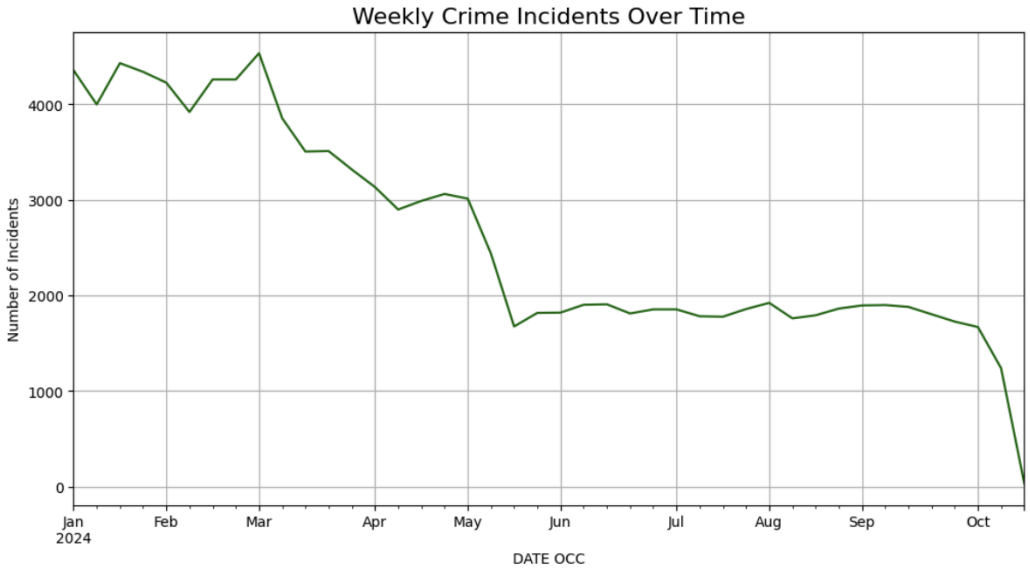


The pie chart shows the Top 10 High-Severity Crimes at Night. The largest portion of crimes at night is Vehicle - Stolen, accounting for 26.7% of the total, followed by Burglary from Vehicle at 16% and Burglary at 13.4%. Other notable crimes include Assault with Deadly Weapon, Aggravated Assault and Theft from Motor Vehicle. Smaller crime categories, such as Intimate Partner - Aggravated Assault, make up a smaller portion of the total incidents.


TIME OCC
AREA
Rpt Dist No
Part 1-2
Crm Cd
Vict Age
Premis Cd
LAT
LON
Hour Occ


Crime Incidents Over the Day of the Month
Number of Crimes
Day of the Month
0
1000
2000
3000
4000
0
5
10
15
20
25
30

The above plot represents the crime incidents over the day of the month



The above graph shows the weekly crime incidents over time.