chicago_2

August 27, 2024

1 Chicago Crime Data Analysis

This comprehensive data analysis script covers the following aspects:

1. Grouping and Aggregating Data:

- Analyzes crime counts by district
- Calculates and visualizes arrest rates by primary type of crime

2. Applying Functions to Data:

- Creates a custom function to categorize time of day
- Applies this function to analyze crime distribution across different times of day

3. Basic Statistical Analysis:

- Performs correlation analysis on numerical columns
- Conducts a chi-square test for independence between 'Primary Type' and 'Arrest'
- Analyzes crime trends over time
- Identifies top locations with highest crime rates

[2]:
```
# Chicago Crime Data Analysis

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
%matplotlib inline

# Load the data (assuming it's already loaded, but including this for␣
 ↪completeness)
file_path = '/Users/YigitAydede/Library/CloudStorage/Dropbox/Documents/Courses/
 ↪MBAN/NLPBootcamp/PythonBC/Crimes_-_2024_20240804.csv'
df = pd.read_csv(file_path)
df['Date'] = pd.to_datetime(df['Date'])

# 1. Grouping and Aggregating Data

# Crime counts by district
crime_by_district = df.groupby('District')['ID'].count().
 ↪sort_values(ascending=False)
```

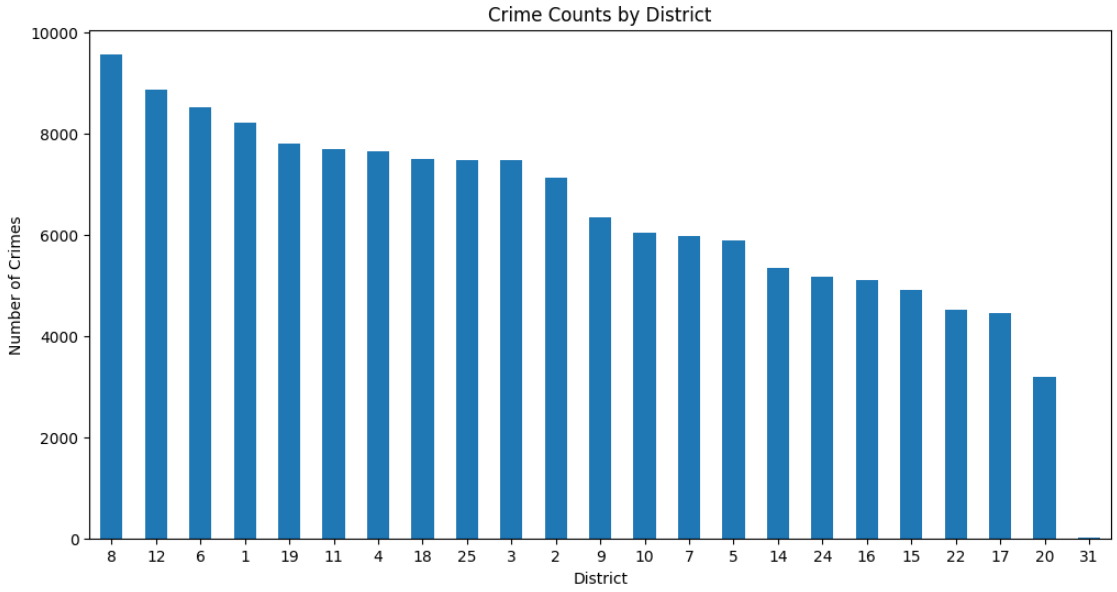
```
print("Crime counts by district:")
print(crime_by_district)

plt.figure(figsize=(12, 6))
crime_by_district.plot(kind='bar')
plt.title('Crime Counts by District')
plt.xlabel('District')
plt.ylabel('Number of Crimes')
plt.xticks(rotation=0)
plt.show()

# Arrest rate by primary type of crime
arrest_rate_by_type = df.groupby('Primary Type')['Arrest'].mean().
 ↪sort_values(ascending=False)
print("\nArrest rate by primary type of crime:")
print(arrest_rate_by_type)

plt.figure(figsize=(12, 6))
arrest_rate_by_type.plot(kind='bar')
plt.title('Arrest Rate by Primary Type of Crime')
plt.xlabel('Primary Type')
plt.ylabel('Arrest Rate')
plt.xticks(rotation=90)
plt.tight_layout()
plt.show()

# 2. Applying Functions to Data

# Custom function to categorize time of day
def categorize_time(hour):
    if 5 <= hour < 12:
        return 'Morning'
    elif 12 <= hour < 17:
        return 'Afternoon'
    elif 17 <= hour < 21:
        return 'Evening'
    else:
        return 'Night'

df['Time of Day'] = df['Date'].dt.hour.apply(categorize_time)

# Crime distribution by time of day
crime_by_time = df['Time of Day'].value_counts()
print("\nCrime distribution by time of day:")
print(crime_by_time)

plt.figure(figsize=(10, 6))
```

```
crime_by_time.plot(kind='pie', autopct='%1.1f%%')
plt.title('Crime Distribution by Time of Day')
plt.ylabel('')
plt.show()

# 3. Basic Statistical Analysis

# Correlation analysis for numerical columns
numeric_columns = df.select_dtypes(include=[np.number]).columns
correlation_matrix = df[numeric_columns].corr()

plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', linewidths=0.5)
plt.title('Correlation Matrix of Numerical Variables')
plt.tight_layout()
plt.show()

# Chi-square test for independence between 'Primary Type' and 'Arrest'
from scipy.stats import chi2_contingency

contingency_table = pd.crosstab(df['Primary Type'], df['Arrest'])
chi2, p_value, dof, expected = chi2_contingency(contingency_table)

print("\nChi-square test results for independence between 'Primary Type' and␣
 ↪'Arrest':")
print(f"Chi-square statistic: {chi2}")
print(f"p-value: {p_value}")

# Additional analysis: Crime trends over time
df['Date'] = pd.to_datetime(df['Date'])
crimes_over_time = df.resample('D', on='Date')['ID'].count()

plt.figure(figsize=(15, 6))
crimes_over_time.plot()
plt.title('Daily Crime Counts Over Time')
plt.xlabel('Date')
plt.ylabel('Number of Crimes')
plt.show()

# Top 5 locations with highest crime rates
top_locations = df['Location Description'].value_counts().head()
print("\nTop 5 locations with highest crime rates:")
print(top_locations)

plt.figure(figsize=(10, 6))
top_locations.plot(kind='bar')
plt.title('Top 5 Locations with Highest Crime Rates')
```

```
plt.xlabel('Location')
plt.ylabel('Number of Crimes')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```

```
/var/folders/b2/gpnsjh9j6bv5prtx7w5lsym80000gp/T/ipykernel_85862/2639725251.py:1
2: UserWarning: Could not infer format, so each element will be parsed
individually, falling back to `dateutil`. To ensure parsing is consistent and
as-expected, please specify a format.
  df['Date'] = pd.to_datetime(df['Date'])
```

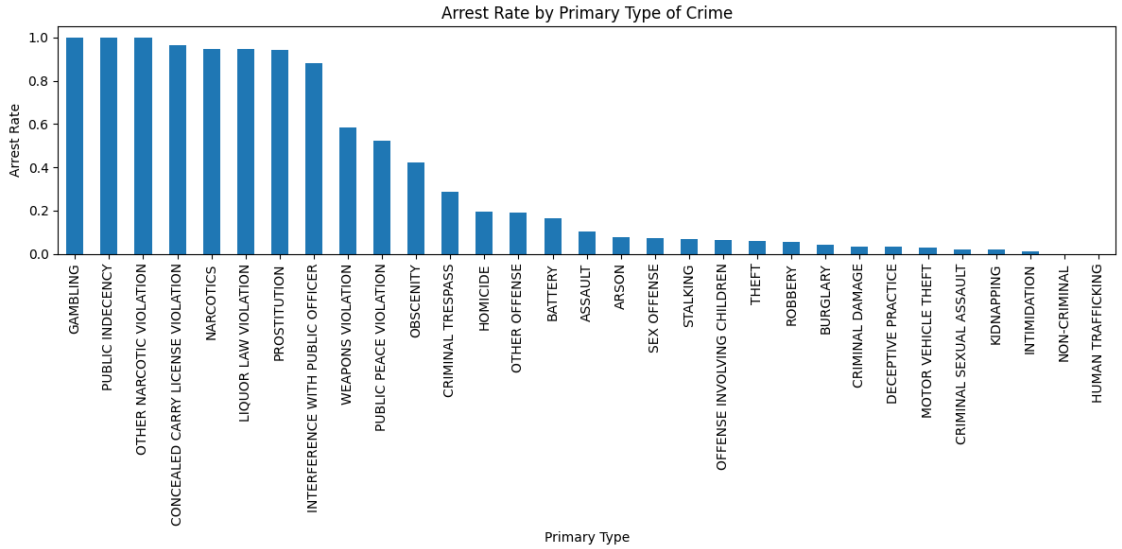
```
Crime counts by district:
District
8     9574
12    8869
6     8520
1     8217
19    7798
11    7690
4     7657
18    7511
25    7489
3     7485
2     7123
9     6346
10    6045
7     5988
5     5898
14    5355
24    5180
16    5115
15    4915
22    4510
17    4451
20    3183
31       6
Name: ID, dtype: int64
```



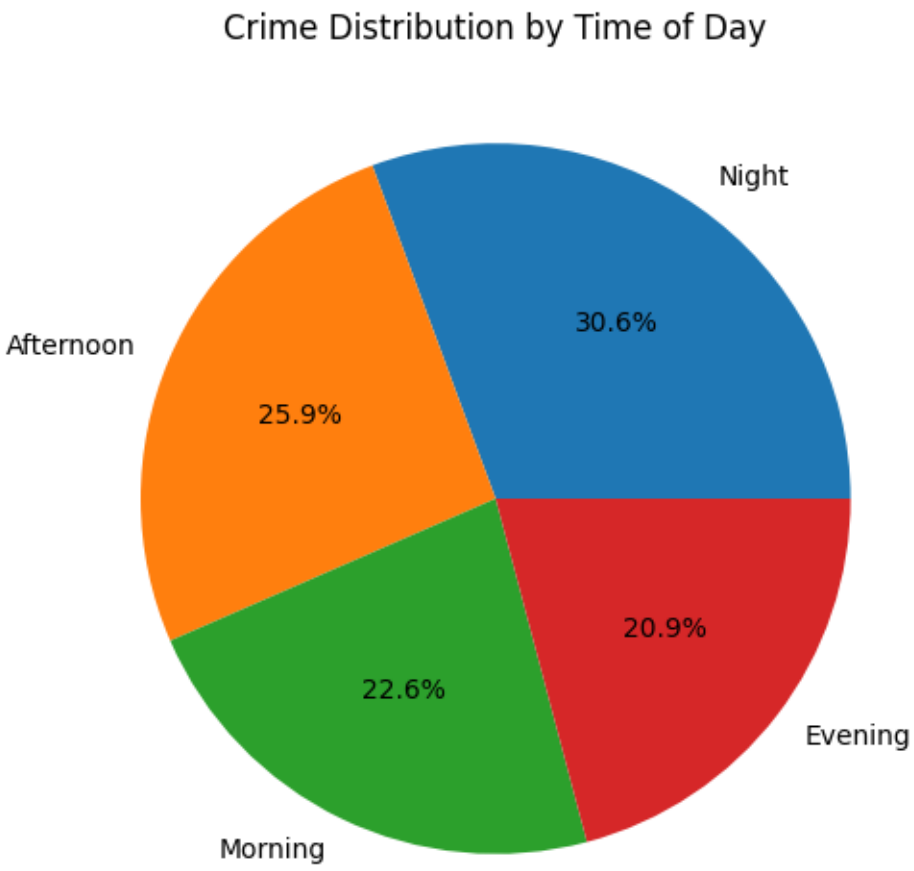
Arrest rate by primary type of crime:

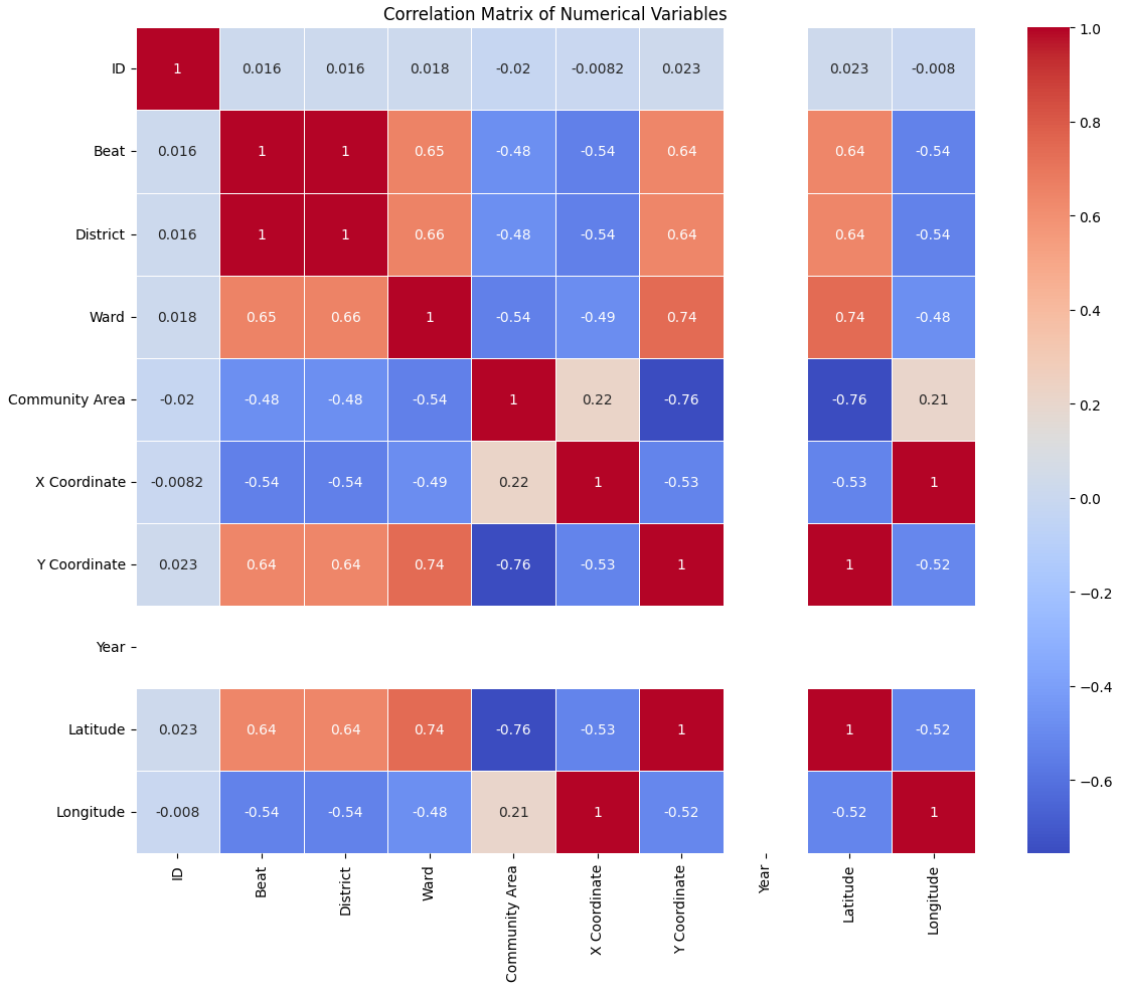
Primary Type	
GAMBLING	1.000000
PUBLIC INDECENCY	1.000000
OTHER NARCOTIC VIOLATION	1.000000
CONCEALED CARRY LICENSE VIOLATION	0.963636
NARCOTICS	0.946956
LIQUOR LAW VIOLATION	0.946429
PROSTITUTION	0.940789
INTERFERENCE WITH PUBLIC OFFICER	0.878866
WEAPONS VIOLATION	0.584312
PUBLIC PEACE VIOLATION	0.520796
OBSCENITY	0.424242
CRIMINAL TRESPASS	0.286237
HOMICIDE	0.194030
OTHER OFFENSE	0.191069
BATTERY	0.163477
ASSAULT	0.101923
ARSON	0.076923
SEX OFFENSE	0.071229
STALKING	0.070312
OFFENSE INVOLVING CHILDREN	0.062807
THEFT	0.061223
ROBBERY	0.054096
BURGLARY	0.040168
CRIMINAL DAMAGE	0.034705
DECEPTIVE PRACTICE	0.032945

```
MOTOR VEHICLE THEFT                 0.029002
CRIMINAL SEXUAL ASSAULT             0.020238
KIDNAPPING                          0.019608
INTIMIDATION                        0.011628
NON-CRIMINAL                        0.000000
HUMAN TRAFFICKING                   0.000000
Name: Arrest, dtype: float64
```

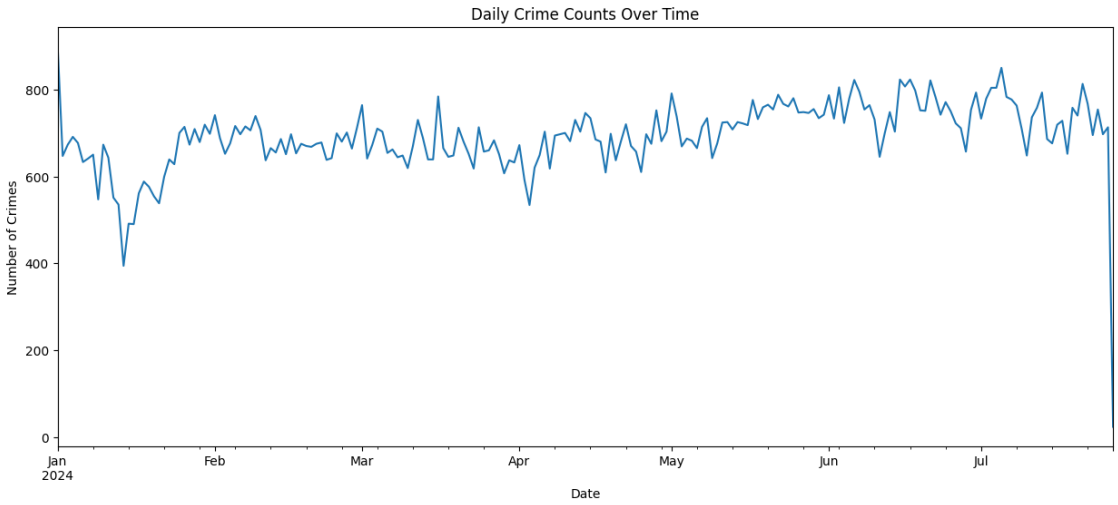



```
Crime distribution by time of day:
Time of Day
Night        44415
Afternoon    37537
Morning      32744
Evening      30229
Name: count, dtype: int64
```


Crime Distribution by Time of Day
Night
30.6%
Afternoon
25.9%
20.9%
22.6%
Evening
Morning



Chi-square test results for independence between 'Primary Type' and 'Arrest':
Chi-square statistic: 39281.53232484505
p-value: 0.0




```
Top 5 locations with highest crime rates:
Location Description
STREET                39705
APARTMENT             27639
RESIDENCE             17026
SIDEWALK               7585
SMALL RETAIL STORE     5442
Name: count, dtype: int64
```