EM-624

PROJECT BY:

SWARAJ KATE

CWID: 10454088

TITLE:

DC CRIME VISUALIZATION AND PREDICTION

INSTRUCTOR:

PROFESSOR CARLO LIPIZZI

PROJECT OBJECTIVE

Data visualization is an efficient tool to detect areas which should be concentrated upon. Using this model, we can help tell which block is most affected and where the police department should concentrate.

We can use geographical maps to pinpoint the exact areas to concentrate on for better visualization.

Using bar graphs and heat maps we can deduce areas of crime and categories of the crime i.e. murder, sexual abuse, etc.

DATA COLLECTION

Here we pull the data from an open source provided by the Open Data DC and Metropolitan Police Department (MPD).

Link: https://opendata.dc.gov/search

Using the above link, I have downloaded the data sets for DC crimes from 2012-2018.

Each data contains an approximate of 35,000 lines in a csv file.

Problem Areas Identified and areas we can work on after visualization and analysis:

Crime related:

- Using this data:
  - We can determine if there are other factors that predict a combination of offenses

- DC makes available crime location data. Using that data:
  - Is it possible to build a model to predict or explain crime?
  - Is it possible to also merge this dataset with another dataset to explore the relationship between crime and, e.g. house prices, parking tickets, 311 requests, or certain city infra.

-**Property-Related**:

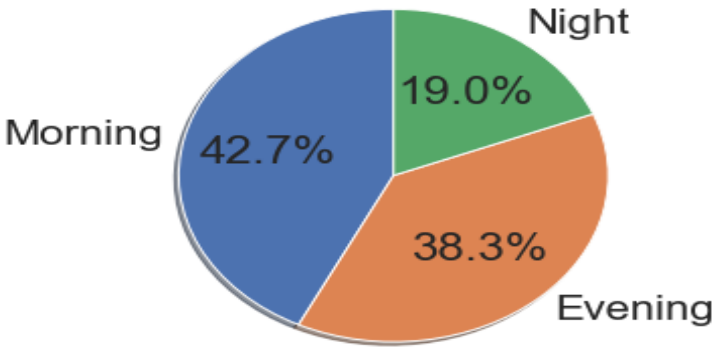
- Is it possible to identify of different features in order to predict whether a house is vacant, or identify the type of property?

- **Forecasting**:
  - We can forecast service needs around DC based on 311 calls and other data.
  - We can also forecast traffic accidents, moving violations, and parking tickets. Based on these forecasts, can we make policy recommendations, forecast revenue.

Visualization of the data: Output and its analysis

The following pie chart shows total crime recorded from various shifts: day, night and midnight. In the open data package, was given the specific time span of the three classifications. In the midnight time range, we can still observe a relatively small amount of crime.

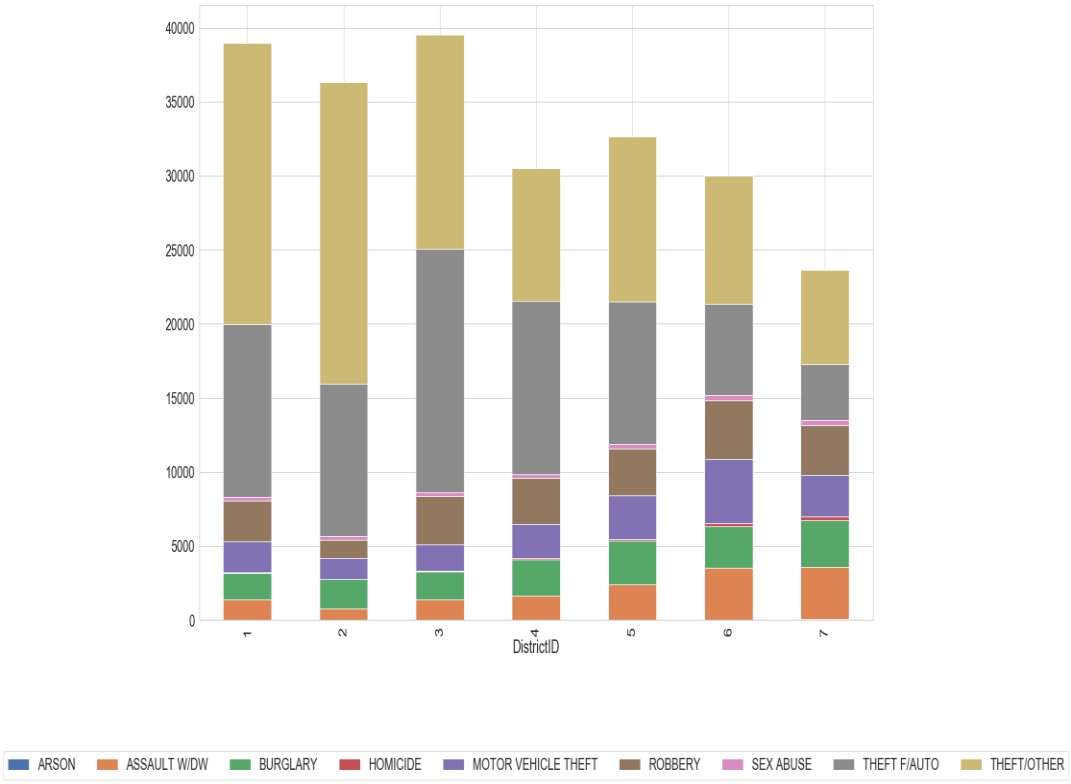
```
#pie chart
labels = 'Morning', 'Evening', 'Night'
#sizes = []
#explode = (0, 0.1, 0, 0)  # only "explode" the 2nd slice (i.e. 'Hogs')

fig1, ax1 = plt.subplots()
ax1.pie(df.SHIFT.value_counts(), labels=labels, autopct='%1.1f%%',
        shadow=True, startangle=90)
ax1.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.

plt.show()
```



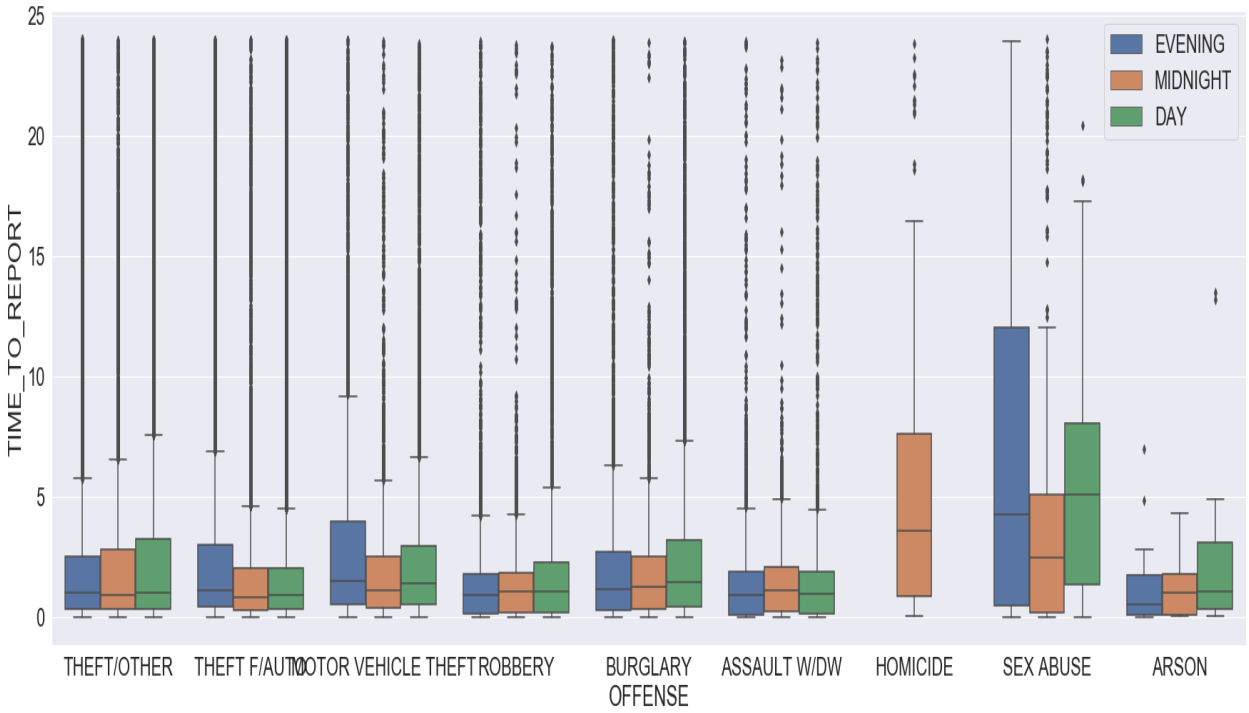
Total Crimes by Shift:

The stacked bar chart below offers information on various types of crime commonly occurring in the Washington DC area clustered using the city's police districts. For each district, a stacked bar plot was chosen to visualize the crimes. This bar plot will count the number of instances for each type of crime and display the results in stacked vertical bar heights.

```
temp_var = pd.crosstab(df.DistrictID, df.OFFENSE)
temp_var.plot(kind='bar', stacked=True, figsize=(25, 15))
plt.legend(loc=9, bbox_to_anchor=(.5, -.2), ncol=9)
```



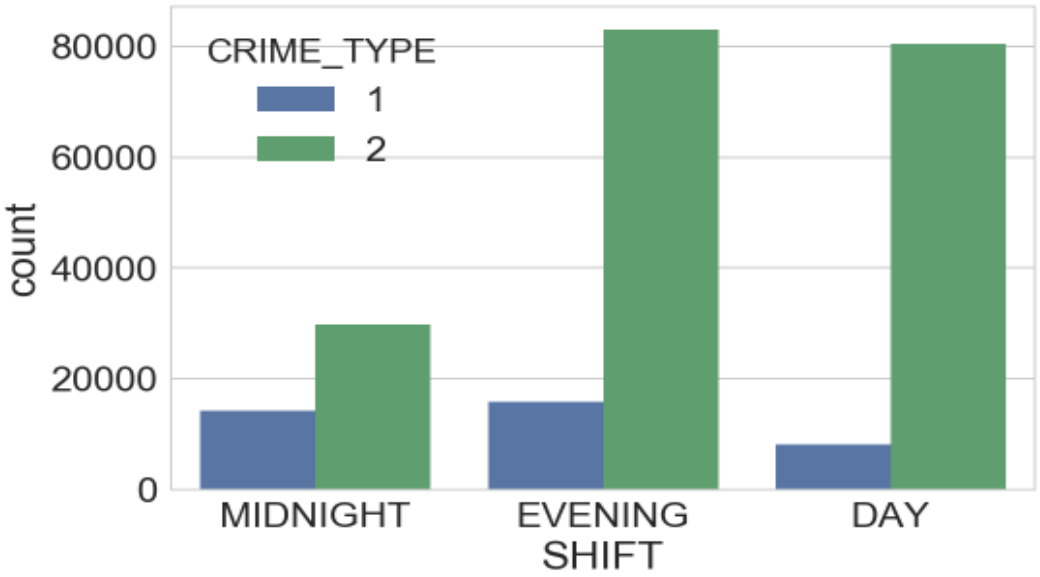
Analysis of crime (type) occurred during the time of the day:



Crimes by type per shift:

Here we can visualize the type of crime occurring during any of the shifts (time of the day) using bar graph.

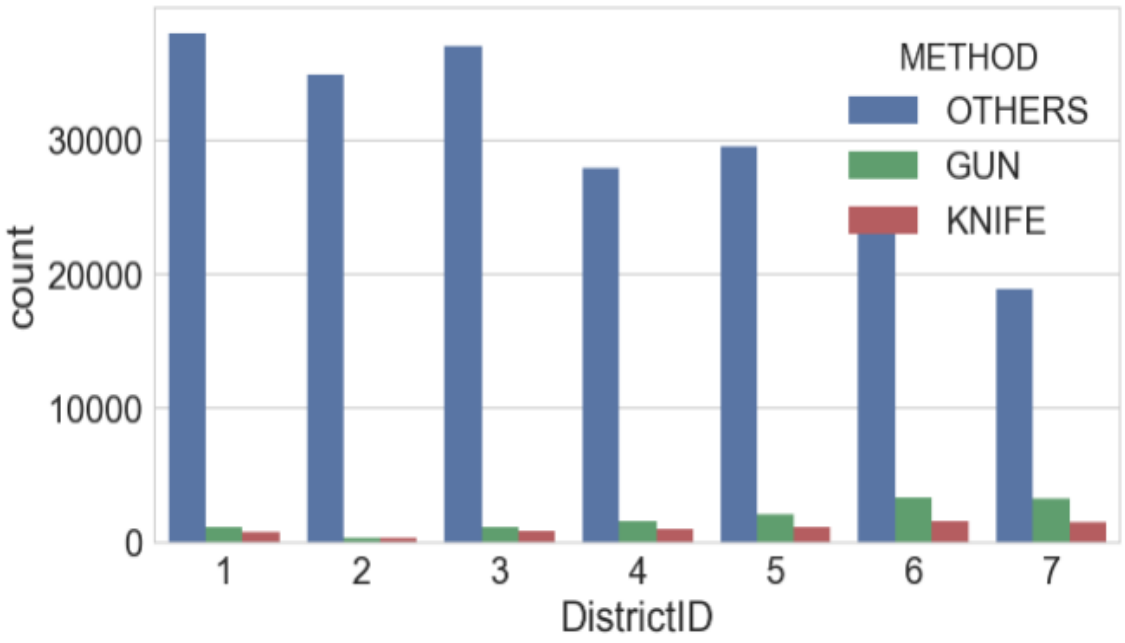
```
temp_var = df[['SHIFT', 'CRIME_TYPE']]
plt.figure(figsize=(8, 5))
sns.countplot(x='SHIFT', hue='CRIME_TYPE', data=temp_var)
```



Method of crime performed analysis:

Here we can visualize the method of crime occurred in the districts.

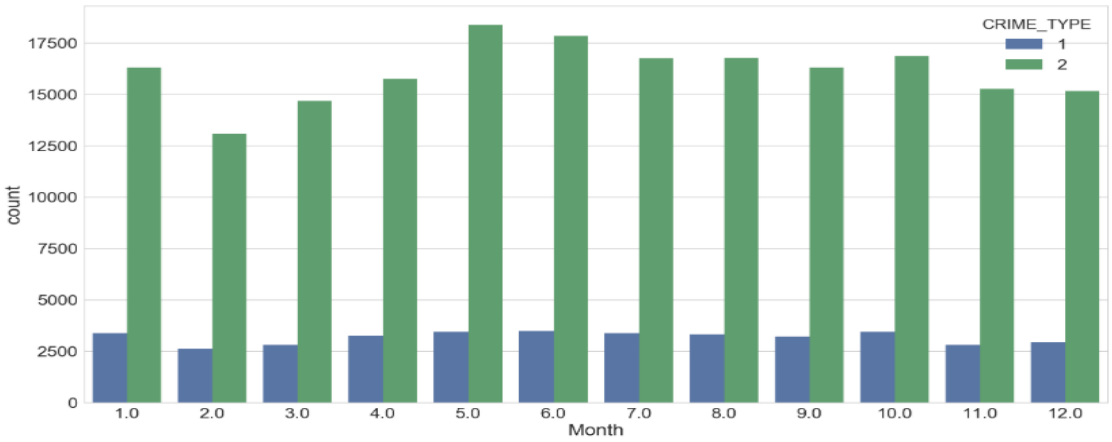
```
temp_var1 = df[['DistrictID', 'METHOD']]
temp_var1 = temp_var1[temp_var1['METHOD'] != "OTHERS"]

plt.figure(figsize=(10, 5))
sns.countplot(x='DistrictID', hue='METHOD', data=temp_var)
plt.figure(figsize=(10, 5))
sns.countplot(x='DistrictID', hue='METHOD', data=temp_var1)

#  1 = Others, 2 = Gun, 3 =  Knife
```



Visualization by crime type per month:

Here we can visualize the type of crime occurred in the months.

```
temp_var = df[['Month', 'CRIME_TYPE']]
plt.figure(figsize=(20, 10))
sns.countplot(x='Month', hue='CRIME_TYPE', data=temp_var)
```

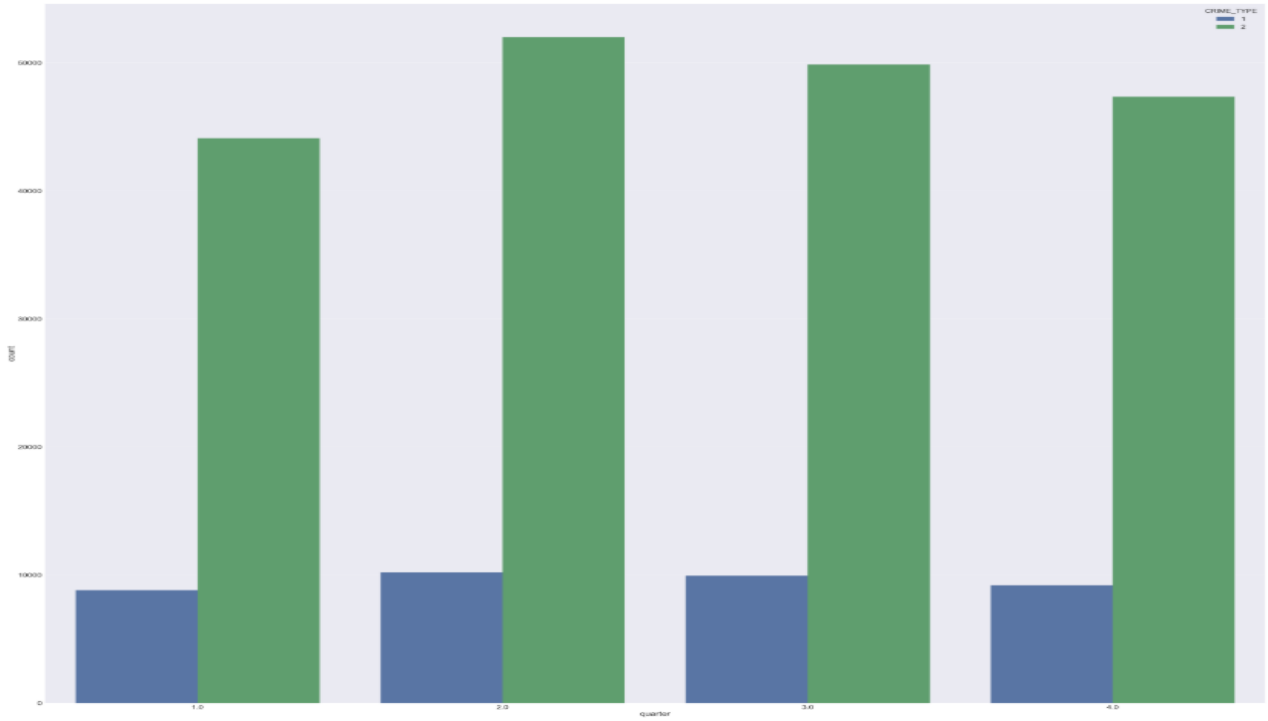


The second and third Quarters of the year show significant increase in the number of violent and nonviolent crimes in the Metropolitan district, which is basically intuitive to understand that as the weather condition improves.

Analysis of Crime type:

Here we analyze the type of crime vs the quarter.

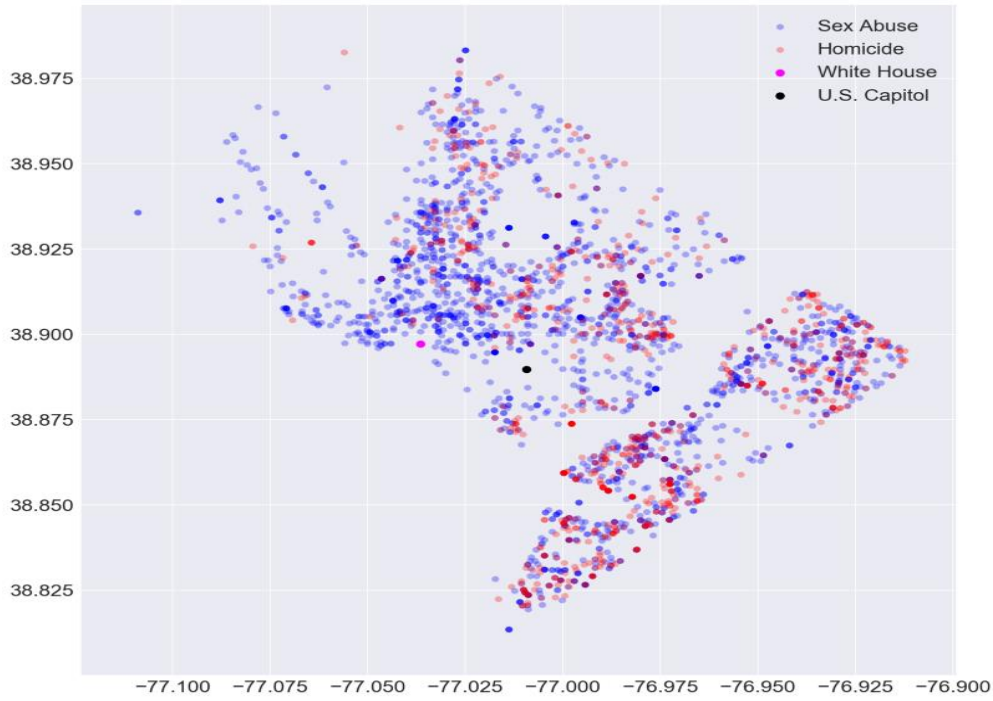
```
temp_var = df[['quarter', 'CRIME_TYPE']]
plt.figure(figsize=(50, 45))
sns.countplot(x= 'quarter', hue='CRIME_TYPE', data=temp_var)
```



Distribution of crime based on time using table graph:

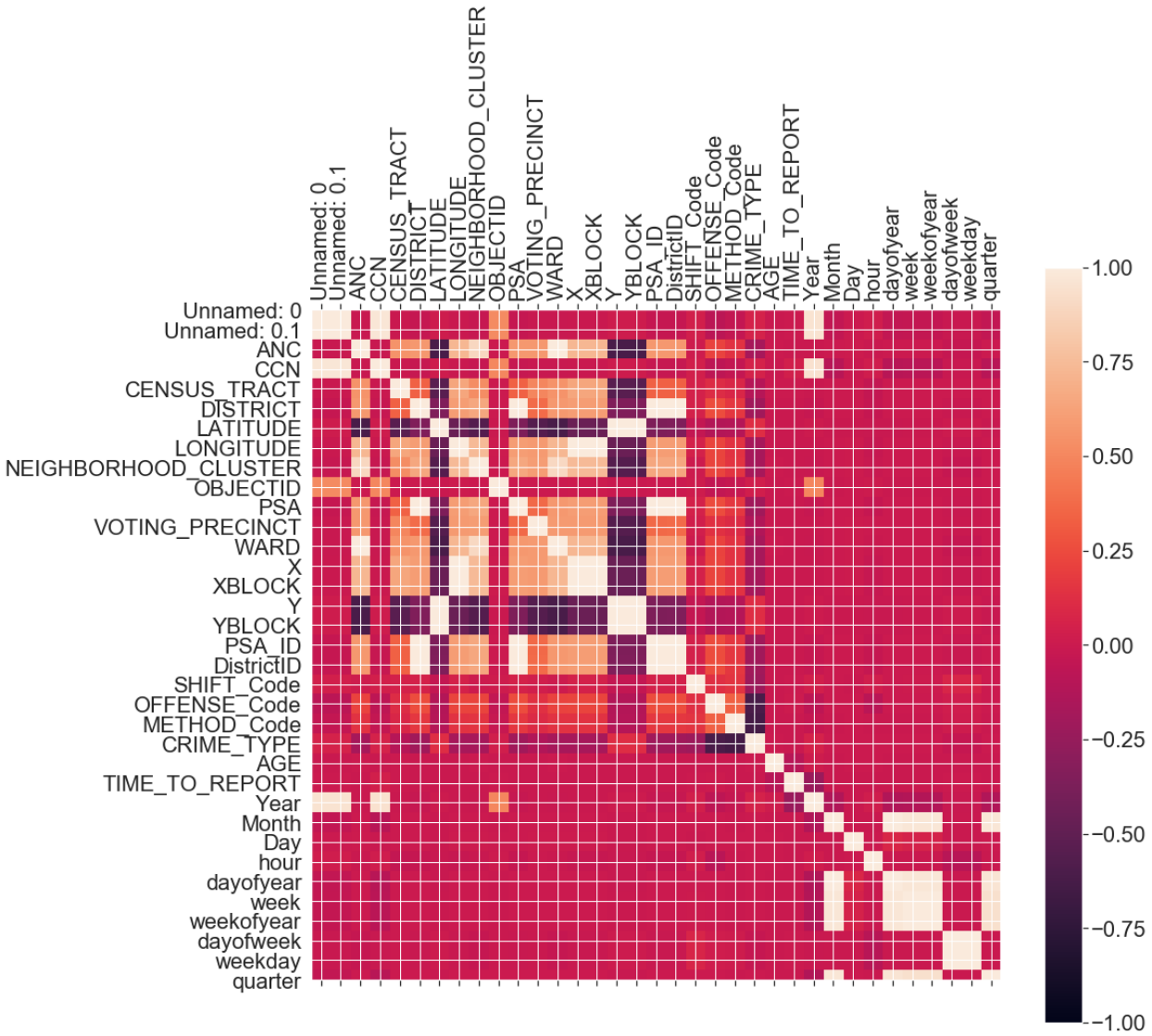
During the holidays, crimes are said to escalate. New Year, Independence Day, Th anksgiving, and Christmas is the definition of holidays for this plot. The analysis ac tually shows a relatively small percentage of the total amount of reported crimes that occur over holidays.

```
plt.figure(figsize=(15, 15))
sns.set(font_scale=2)
plt.scatter(df['LONGITUDE'][df['OFFENSE']=='SEX ABUSE'],
df['LATITUDE'][df['OFFENSE']=='SEX ABUSE'], s=50, alpha=0.3, color='b', lw=0,
label='Sex Abuse')
plt.scatter(df['LONGITUDE'][df['OFFENSE']=='HOMICIDE'],
df['LATITUDE'][df['OFFENSE']=='HOMICIDE'], s=50, alpha=0.3, color='r', lw=0,
label='Homicide')
plt.scatter(-77.03654,38.89722,s=60, color=[1,0,1], lw=1, label='White House')
plt.scatter(-77.00937,38.88968,s=60, color=[0,0,0], lw=1, label='U.S. Capitol')
plt.legend(loc='upper right')
plt.show()
```



The map aboveis an attempt to see if the numerous violent crimes across the distr ict were geographically dispersed. Although there seem to be a few clumps of dangerous areas, it seems to be spreading across D.C. in general. — Except for the Northwest area (District 2 of the Police). The density of violent crimes near the downtown DC are surprising.

Correlation between the data:



Conclusion/Understanding the data from the above data analysis:

- We have two potential reaction factors: Crime Type (peaceful wrongdoing versus Brutal wrongdoing), and Offense Code (The more explicit sort of offense: Homicide, Robbery, Theft, Arson, and so on.). The objective is to furnish the police Department with a model that can foresee or arrange a wrongdoing dependent on the accessible illustrative factors gathered from the 'opendatadc' just as from the site.

- The informative factors for this dataset center around time and areas. We accept that the recognition/grouping of a Violent wrongdoing would be founded fundamentally on the injured individual's qualities, and not solely on the area or time.

- Our investigation of the factors appears to demonstrate that time is one of the more noteworthy factors. At the point when we separated the time into singular hours of the day, we saw an articulated cyclic impact, where evening violations were undeniably more probable than daytime crimes.

    - Area additionally seemed to have some impact. Distinctive political regions (Wards and the subordinate Association Neighborhood Committees) demonstrated an unexpected pattern in comparison to utilizing worldwide areas (Latitude and Longitude). Police regions (and their subordinate Police Service Areas (PSA) demonstrated an unexpected pattern in comparison to the Ward gathering. This reveals to us that there are some area impacts, however it is hard to isolate them out because of the relationship between geo-physical regions and the unique political mappings.