Homework 4

Submitted by: Hang Tian

Submission date: 10/13/23

This .ipynb file is contributed by everyone in group 4 for IST 652 class
Members are Kishan, Babatunde, Kapil, Hemanth Chowdary and Hang

Instructions: You will process and analyze a large data set that contains crimes reported in the city of Chicago from 2018 to 2021.

To load the data set and get the *crimes* dataframe correctly configured, execute the cells with the code provided in this notebook. This could take a few minutes after you start the execution of the code cells.

Once the *crimes* dataframe has been setup proceed to obtain 2 meaningful data analysis results from processing the *crimes* dataframe. Specific cells have been provided for you to describe the results of each of your data analysis procedures. You can add as many code cells as you want to complete each of your analysis and I also recommend that you add some explanatory cells (use Markdown) to provide some additional text with explanations of your analysis.

In [ ]:

```
#EXECUTE THIS CELL to setup the modules you need
%matplotlib inline
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

In [ ]:

```
#Defining location of dataset
filepath="~/datasets/ist652/Crimes/crimes_2018_2021.csv"
localpath='crimes_2018_2021.csv'
```

In [ ]:

```
#EXECUTE THIS CELL to load the dataset into your environment - THIS WILL TAKE 3 TO 5 MINUTES - be patient
# a security warning will appear. You can ignore it.
try:
    crimes=pd.read_csv(filepath,parse_dates=[2])
```

```
except:
    crimes=pd.read_csv(localpath,parse_dates=[2])
```

In [ ]:

```
crimes.head()  #just checking
```

Out[ ]:

	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domestic	Beat	...	Ward	Co
ID													
11646166	JC213529	9/1/2018 0:01	082XX S INGLESIDE AVE	810	THEFT	OVER $500	RESIDENCE	False	True	631	...	8.0	
11645648	JC212959	1/1/2018 8:00	024XX N MONITOR AVE	1153	DECEPTIVE PRACTICE	FINANCIAL IDENTITY THEFT OVER $ 300	RESIDENCE	False	False	2515	...	30.0	
11645959	JC211511	12/20/2018 16:00	045XX N ALBANY AVE	2820	OTHER OFFENSE	TELEPHONE THREAT	RESIDENCE	False	False	1724	...	33.0	
11645557	JC212685	4/1/2018 0:01	080XX S VERNON AVE	1153	DECEPTIVE PRACTICE	FINANCIAL IDENTITY THEFT OVER $ 300	RESIDENCE	False	False	631	...	6.0	
11646293	JC213749	12/20/2018 15:00	023XX N LOCKWOOD AVE	1154	DECEPTIVE PRACTICE	FINANCIAL IDENTITY THEFT $300 AND UNDER	APARTMENT	False	False	2515	...	36.0	

5 rows × 21 columns

Code for data analysis 1

You can place the code for your first data analysis result in this section. Add as many code cells as you need.

Analysis 1. The timely change of crime cases

In [ ]:

```
crimes.info()
# It seems there are lots of crime cases without location information
# I'll drop those just in case there will be location-related analysis.
crimes_ordered=crimes.sort_index(ascending=True)
crimes_ordered.drop(crimes_ordered[np.isnan(crimes_ordered['Latitude'])].index, inplace=True)
# Now set index as date column and sort it
crimes_ordered.set_index("Date",inplace=True)
crimes_ordered=crimes_ordered.sort_index(ascending=True)
crimes_ordered.shape[0]
# In the end, there are 934034 left for analysis
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 946893 entries, 0 to 946892
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   ID                    946893 non-null  int64
 1   Case Number           946893 non-null  object
 2   Date                  946893 non-null  datetime64[ns]
 3   Block                 946893 non-null  object
 4   IUCR                  946893 non-null  object
 5   Primary Type          946893 non-null  object
 6   Description           946893 non-null  object
 7   Location Description  942727 non-null  object
 8   Arrest                946893 non-null  bool
 9   Domestic              946893 non-null  bool
 10  Beat                  946893 non-null  int64
 11  District              946893 non-null  int64
 12  Ward                  946854 non-null  float64
 13  Community Area        946892 non-null  float64
 14  FBI Code              946893 non-null  object
 15  X Coordinate          934034 non-null  float64
 16  Y Coordinate          934034 non-null  float64
 17  Year                  946893 non-null  int64
 18  Updated On            946893 non-null  object
 19  Latitude              934034 non-null  float64
 20  Longitude             934034 non-null  float64
 21  Location              934034 non-null  object
dtypes: bool(2), datetime64[ns](1), float64(6), int64(4), object(9)
memory usage: 146.3+ MB
```

Out[ ]:

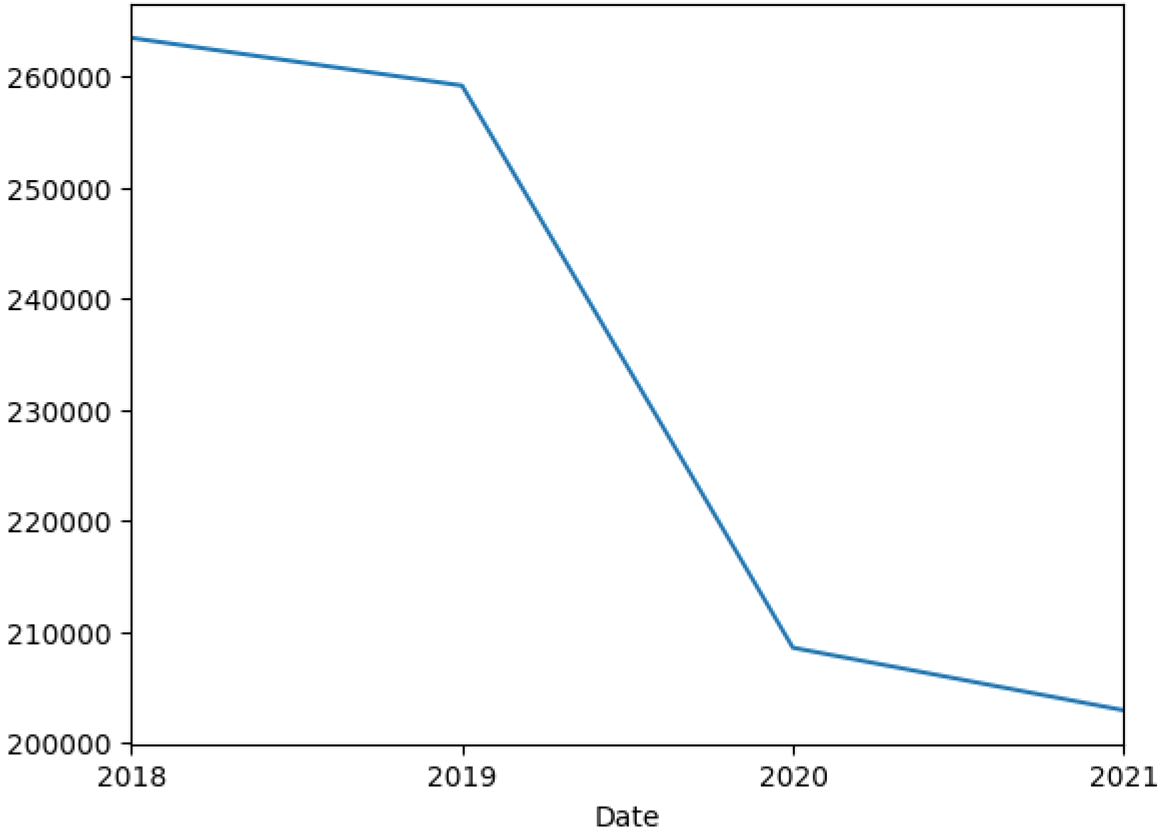
```
934034
```

In [ ]:

```
# Annual change
annual_count=crimes_ordered['Case Number'].resample('Y').count()
annual_count.plot()
# Case numbers through years is getting less. There is a dramatic drop from 2019 to 2020.
```

Out[ ]:

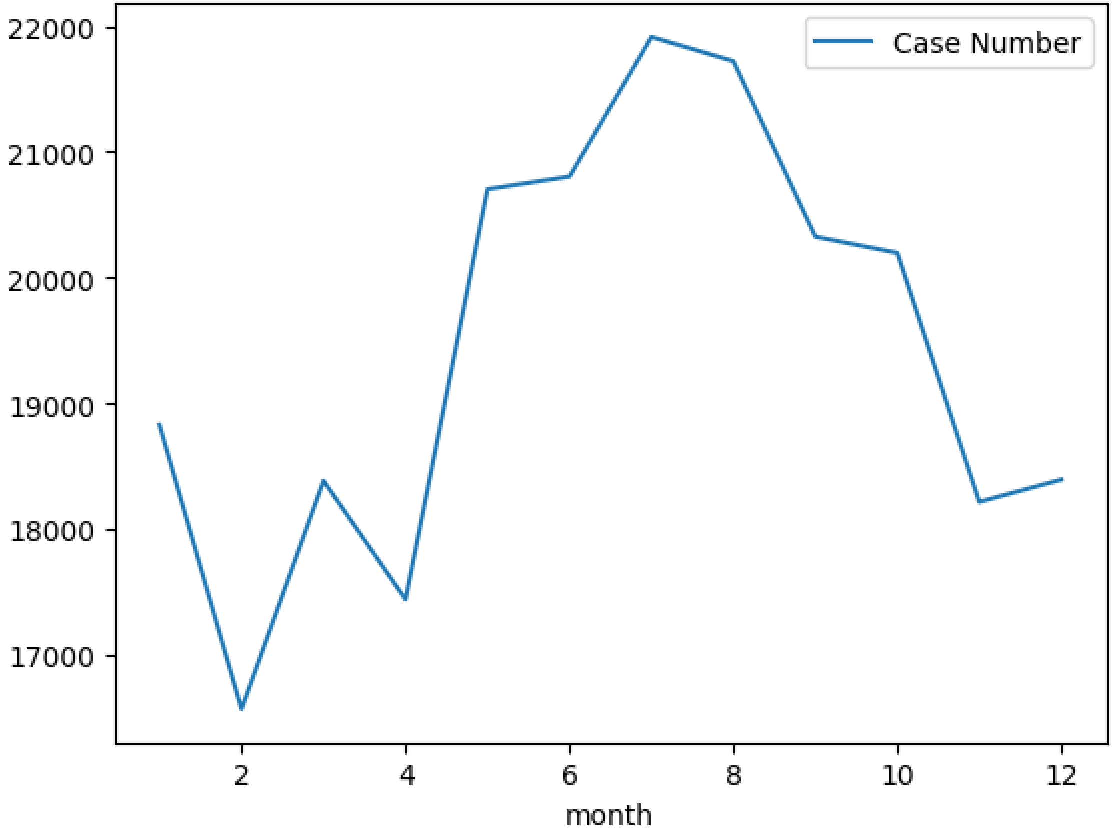
```
<AxesSubplot:xlabel='Date'>
```



In [ ]:

```
# Monthly average change
monthly=crimes_ordered['Case Number'].resample('M').count().to_frame()
monthly.reset_index(inplace=True)
monthly['month']=monthly['Date'].dt.month
month_average=monthly.groupby(['month']).agg('mean')
month_average.plot()
# On average, crime cases count is high from May to October, while summer months (Jult - August) see the pe
# As the weather gets colder, people are less likely to commit crimes.
```

```
Out[ ]: <AxesSubplot:xlabel='month'>
```



Description of data analysis result 1

Use the next cell to describe your data analysis result 1

Result 1

Gnerally speaking, total crime cases decreases through these years. On average, warmer months have more crime cases.

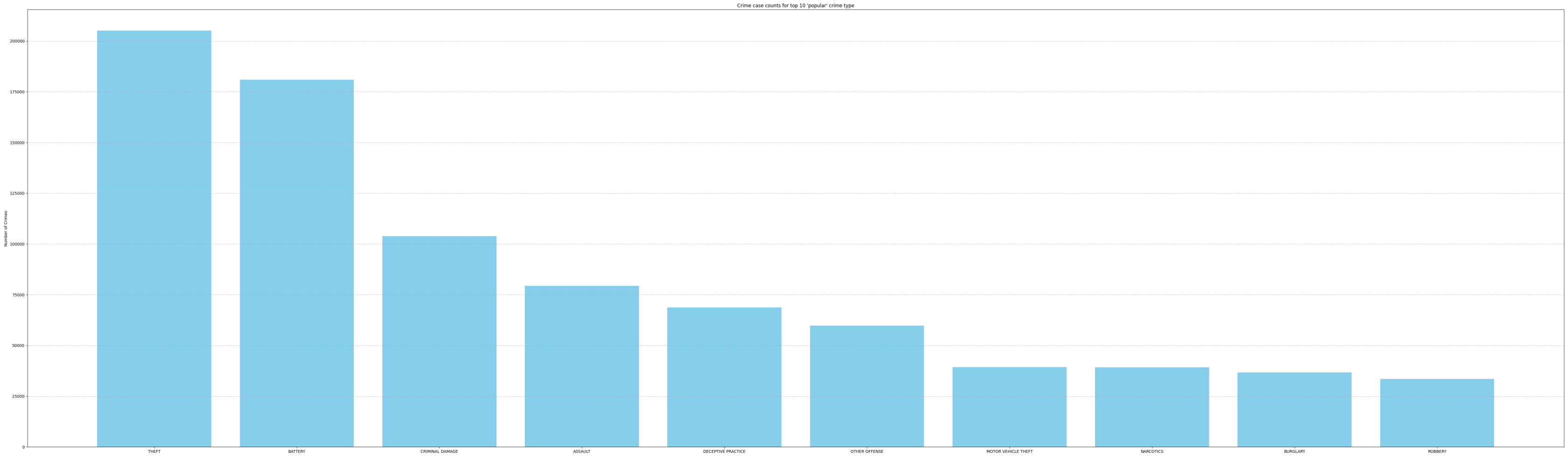
Code for data analysis 2

You can place the code for your second data analysis result in this section. Add as many code cells as you need.

In [ ]:
```
# What are the crime types with more total cases
common_crimes = crimes_ordered['Primary Type'].value_counts().head(10).to_frame()
common_crimes.index
# According to the total crime case counts, 'theft' and 'battery' are the most 'popular' ones,
# followed by 'criminal damage', 'assualt' ...
```

Out[ ]:
```
Index(['THEFT', 'BATTERY', 'CRIMINAL DAMAGE', 'ASSAULT', 'DECEPTIVE PRACTICE',
       'OTHER OFFENSE', 'MOTOR VEHICLE THEFT', 'NARCOTICS', 'BURGLARY',
       'ROBBERY'],
      dtype='object')
```

In [ ]:
```
common_crimes
plt.figure(figsize=(70, 20))
plt.bar(common_crimes['Primary Type'].index, common_crimes['Primary Type'].values, color='skyblue')
plt.title("Crime case counts for top 10 'popular' crime type")
plt.ylabel('Number of Crimes')
plt.grid(axis='y', linestyle='--', alpha=0.8)
plt.show()
```



In [ ]:
```
# What are the crime types that cause higher proportion of people being arrested
total_type_list=crimes_ordered['Primary Type'].unique()
crimes_arrested=crimes_ordered[crimes_ordered['Arrest']==True]
print('The ratio of crime cases having people arrested is',crimes_arrested.shape[0]/crimes_ordered.shape[0
print(f'There are a total of {total_type_list.shape[0]} primary crime types')
```

```
arrested_types=crimes_arrested['Primary Type'].unique()
print(f'There are only {arrested_types.shape[0]} primary crime types involved with people arrested')

type_not_seen_arrested=list(set(total_type_list)-set(arrested_types))[0]
print(f'The only crime type through these records not involved with people arrested is {type_not_seen_arres
```

```
The ratio of crime cases having people arrested is 0.1787536642135083
There are a total of 34 primary crime types
There are only 33 primary crime types involved with people arrested
The only crime type through these records not involved with people arrested is RITUALISM
```

In [ ]:

```
crimes_arrest_by_type_table=crimes_ordered.groupby(['Primary Type','Arrest'])['Case Number'].count().to_fra
crimes_total_by_type_table=crimes_ordered.groupby(['Primary Type'])['Case Number'].count().to_frame()
crimes_arrest_by_type_table.reset_index(inplace=True)
crimes_total_by_type_table.reset_index(inplace=True)
crimes_total_by_type_table.columns=['Primary Type','Total Case Number']
crime_count_table=pd.merge(crimes_arrest_by_type_table,crimes_total_by_type_table,on='Primary Type')
crime_count_table.set_index('Primary Type',inplace=True)
crime_count_table['Proportion']=crime_count_table['Case Number']/crime_count_table['Total Case Number']
crime_arrested_proportion=crime_count_table[crime_count_table['Arrest']==True].sort_values('Proportion',asc
crime_arrested_proportion
# Liquor law violation and public indencency have the highest arrested rate of 100%,
# followed by prostitution, narcotics, gambling, concealed carry license violation and interference with pu
# that have a arrested proportion of more than 90%.
# The bottom 3 is intimidation, deceptive practice and human trafficking
# Though theft is the most 'popular' crime type among these records, the thief is arrested only in 8% of to
# While battery-related crimes have seen 18% arrested rate.
```

Out[ ]:

	Proportion
Primary Type	
LIQUOR LAW VIOLATION	1.000000
PUBLIC INDECENCY	1.000000
PROSTITUTION	0.998303
NARCOTICS	0.996731
GAMBLING	0.994751
CONCEALED CARRY LICENSE VIOLATION	0.972303
INTERFERENCE WITH PUBLIC OFFICER	0.942363
OBSCENITY	0.786611
WEAPONS VIOLATION	0.652416
OTHER NARCOTIC VIOLATION	0.600000
PUBLIC PEACE VIOLATION	0.545435
CRIMINAL TRESPASS	0.484963
NON-CRIMINAL (SUBJECT SPECIFIED)	0.333333
HOMICIDE	0.319496
BATTERY	0.183443
OTHER OFFENSE	0.175829
ASSAULT	0.141387
SEX OFFENSE	0.122492
OFFENSE INVOLVING CHILDREN	0.114381
STALKING	0.093142
NON-CRIMINAL	0.088889
ARSON	0.088410
THEFT	0.081517
CRIM SEXUAL ASSAULT	0.079498

	Proportion
Primary Type	
ROBBERY	0.074279
KIDNAPPING	0.056466
CRIMINAL SEXUAL ASSAULT	0.053818
CRIMINAL DAMAGE	0.052045
BURGLARY	0.050738
MOTOR VEHICLE THEFT	0.048112
INTIMIDATION	0.044335
DECEPTIVE PRACTICE	0.033569
HUMAN TRAFFICKING	0.025000

Description of data analysis result 2

Use the next cell to describe your data analysis result 2

Which crime types have higher arrest rate?

Liquor law violation and public indencency have the highest arrested rate of 100%,
followed by prostitution, narcotics, gambling, concealed carry license violation and interference with public officer
that have a arrested proportion of more than 90%.
The bottom 3 is intimidation, deceptive practice and human trafficking
Though theft is the most 'popular' crime type among these records, the thief is arrested only in 8% of total cases
While battery-related crimes have seen 18% arrested rate.

Note: PLEASE SUBMIT YOUR HOMEWORK IN *ipynb* AND *pdf* formats. Use the "Download as" option in the "File" menu to get your download the notebook in those formats.