МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО”
Факультет прикладної математики
Кафедра програмного забезпечення комп’ютерних систем


**Лабораторна робота №2**

з дисципліни: «Технології оброблення великих даних»

на тему: «Аналіз та візуалізація даних у Python»

Виконав
студент ІІІ курсу каф. ПЗКС ФПМ
групи КП-82
Мельничук Олексій Геннадійович

Перевірила
доц. каф. ПЗКС ФПМ
Олещенко Л.М.

Київ 2021

## 1. Індивідуальне завдання

**Мета**: продемонструвати свої знання про життєвий цикл аналізу даних, використовуючи заданий набір даних та вказані інструменти.

У цій лабораторній роботі ви імпортуєте деякі пакети Python, необхідні для аналізу набору даних, що містить інформацію про злочини в Сан-Франциско. Потрібно використати засоби Python та Jupyter, щоб підготувати ці дані до аналізу, проаналізувати їх, побудувати графіки та повідомити про свої результати.

## 2. Хід роботи

### Частина 1: Імпорт пакетів Python

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import folium
import os

from matplotlib import colors
```

### Частина 2: Завантажте дані

**Крок 1: Завантажте дані про злочини Сан-Франциско у дата фрейм.**

```
#PART2
#step1
folder_path = os.path.dirname(os.path.abspath(__file__))

dataset_path = folder_path + '\\res\Map-Crime_Incidents-Previous_Three_Months.csv'

SF = pd.read_csv(dataset_path)
```

**Крок 2: Перегляньте імпортовані дані.**

```
#step2

pd.set_option('display.max_rows', 10)
print(SF)
print('VARIABLES:', SF.columns)
print('AMOUNT OF COLUMNS:', len(SF.columns))
print('AMOUNT OF ROWS:', len(SF))
```

```
C:\Users\datru\Desktop\study2021\big_data\lab2>py main.py
       IncidntNum        Category                                  Descript  ...           X         Y                                  Location
0             NaN   LARCENY/THEFT         GRAND THEFT FROM UNLOCKED AUTO  ... -122.417393  37.790974  (37.7909741243888, -122.417392830334)
1             NaN   LARCENY/THEFT           GRAND THEFT FROM LOCKED AUTO  ... -122.404418  37.796302  (37.7963018736036, -122.404417620748)
2             NaN   LARCENY/THEFT           GRAND THEFT FROM LOCKED AUTO  ... -122.406959  37.789435  (37.7894347630337, -122.406958660602)
3             NaN   DRUG/NARCOTIC         POSSESSION OF METH-AMPHETAMINE  ... -122.419672  37.765050  (37.7650501214965, -122.419671780296)
4             NaN   DRUG/NARCOTIC                   POSSESSION OF COCAINE  ... -122.417904  37.785167  (37.7851670875814, -122.417903977564)
...           ...             ...                                     ...  ...         ...        ...                                    ...
30755         NaN   LARCENY/THEFT                PETTY THEFT SHOPLIFTING  ... -122.408052  37.783957  (37.7839574642528, -122.408051765969)
30756         NaN  OTHER OFFENSES  DRIVERS LICENSE, SUSPENDED OR REVOKED  ... -122.418601  37.780261  (37.7802607511488, -122.418600974625)
30757         NaN         ASSAULT                                 BATTERY  ... -122.412122  37.781379  (37.7813786419025, -122.412121608136)
30758         NaN         ASSAULT         ASSAULT WITH CAUSTIC CHEMICALS  ... -122.407434  37.787494  (37.7874944447786, -122.407434204569)
30759         NaN  OTHER OFFENSES  DRIVERS LICENSE, SUSPENDED OR REVOKED  ... -122.426391  37.733675  (37.7336749150401, -122.426391018521)

[30760 rows x 12 columns]
VARIABLES: Index(['IncidntNum', 'Category', 'Descript', 'DayOfWeek', 'Date', 'Time',
       'PdDistrict', 'Resolution', 'Address', 'X', 'Y', 'Location'],
      dtype='object')
AMOUNT OF COLUMNS: 12
AMOUNT OF ROWS: 30760

C:\Users\datru\Desktop\study2021\big_data\lab2>
```

У фреймі знаходиться 12 змінних та 30760 рядків.

## Частина 3: Підготовка даних

### Крок 1: Витягніть місяць і день із поля Дата.

```python
#PART3
#step1

SF['Month'] = SF['Date'].apply(lambda row: int(row[0:2]))
SF['Day'] = SF['Date'].apply(lambda row: int(row[3:5]))

print(SF['Month'][0:2])
print(SF['Day'][0:2])

print(type(SF['Month'][0]))
```

```
C:\Users\datru\Desktop\study2021\big_data\lab2>py main.py
0    8
1    8
Name: Month, dtype: int64
0    31
1    31
Name: Day, dtype: int64
<class 'numpy.int64'>

C:\Users\datru\Desktop\study2021\big_data\lab2>
```

**Крок 2: Видаліть змінні з дата фрейму SF.**

```
#step2
del SF['IncidntNum']

SF.drop('Location', axis=1, inplace=True )
print('DATABASE: ', SF)
print('COLUMNS:', SF.columns)
```

```
C:\Users\datru\Desktop\study2021\big_data\lab2>py main.py
DATABASE:              Category                              Descript DayOfWeek                      Date  ...           X          Y Month Day
0       LARCENY/THEFT          GRAND THEFT FROM UNLOCKED AUTO   Sunday  08/31/2014 07:00:00 AM +0000  ... -122.417393  37.790974     8  31
1       LARCENY/THEFT            GRAND THEFT FROM LOCKED AUTO   Sunday  08/31/2014 07:00:00 AM +0000  ... -122.404418  37.796302     8  31
2       LARCENY/THEFT            GRAND THEFT FROM LOCKED AUTO   Sunday  08/31/2014 07:00:00 AM +0000  ... -122.406959  37.789435     8  31
3       DRUG/NARCOTIC          POSSESSION OF METH-AMPHETAMINE   Sunday  08/31/2014 07:00:00 AM +0000  ... -122.419672  37.765050     8  31
4       DRUG/NARCOTIC                    POSSESSION OF COCAINE   Sunday  08/31/2014 07:00:00 AM +0000  ... -122.417904  37.785167     8  31
...               ...                                     ...      ...                           ...  ...         ...        ...   ...  ..
30755   LARCENY/THEFT                  PETTY THEFT SHOPLIFTING   Sunday  06/01/2014 07:00:00 AM +0000  ... -122.408052  37.783957     6   1
30756  OTHER OFFENSES  DRIVERS LICENSE, SUSPENDED OR REVOKED   Sunday  06/01/2014 07:00:00 AM +0000  ... -122.418601  37.780261     6   1
30757         ASSAULT                                 BATTERY   Sunday  06/01/2014 07:00:00 AM +0000  ... -122.412122  37.781379     6   1
30758         ASSAULT          ASSAULT WITH CAUSTIC CHEMICALS   Sunday  06/01/2014 07:00:00 AM +0000  ... -122.407434  37.787494     6   1
30759  OTHER OFFENSES  DRIVERS LICENSE, SUSPENDED OR REVOKED   Sunday  06/01/2014 07:00:00 AM +0000  ... -122.426391  37.733675     6   1

[30760 rows x 12 columns]
COLUMNS: Index(['Category', 'Descript', 'DayOfWeek', 'Date', 'Time', 'PdDistrict',
       'Resolution', 'Address', 'X', 'Y', 'Month', 'Day'],
      dtype='object')

C:\Users\datru\Desktop\study2021\big_data\lab2>
```

## Частина 4: Аналіз даних

**Крок 1: Узагальнення змінних для отримання статистичної інформації.**

```
#PART4
#step1
print('Most frequent offence:', SF['Category'].value_counts()[0:1])
print('Most dangerous district:', SF['PdDistrict'].value_counts()[0:1])
```

```
C:\Users\datru\Desktop\study2021\big_data\lab2>py main.py
Most frequent offence: LARCENY/THEFT    8205
Name: Category, dtype: int64
Most dangerous district: SOUTHERN    6185
Name: PdDistrict, dtype: int64

C:\Users\datru\Desktop\study2021\big_data\lab2>
```

Найбільш скоєний злочин: Крадіжка, 8205 разів

Найнебезпечніший район: Південний, 6185 злочинів

**Крок 2. Підгрупуйте дані у менші дата фрейми (кадри даних).**

```
#step2
AugustCrimes = SF[SF['Month'] == 8]
print('Crimes in August:', len(AugustCrimes))
BurglaryAugust = SF[SF['Category'] == 'BURGLARY']
print('Burglaries of August:', len(BurglaryAugust))

Crime0704 = SF.query('Month == 7 and Day == 4')
print('Crimes as of July 4th:', len(Crime0704))
```

```
C:\Users\datru\Desktop\study2021\big_data\lab2>py main.py
Crimes in August: 9720
Burglaries of August: 1257
Crimes as of July 4th: 341

C:\Users\datru\Desktop\study2021\big_data\lab2>
```

Кількість злочинів скоєних в Серпні: 9720

Кількість зламів скоєних в Серпні: 1257

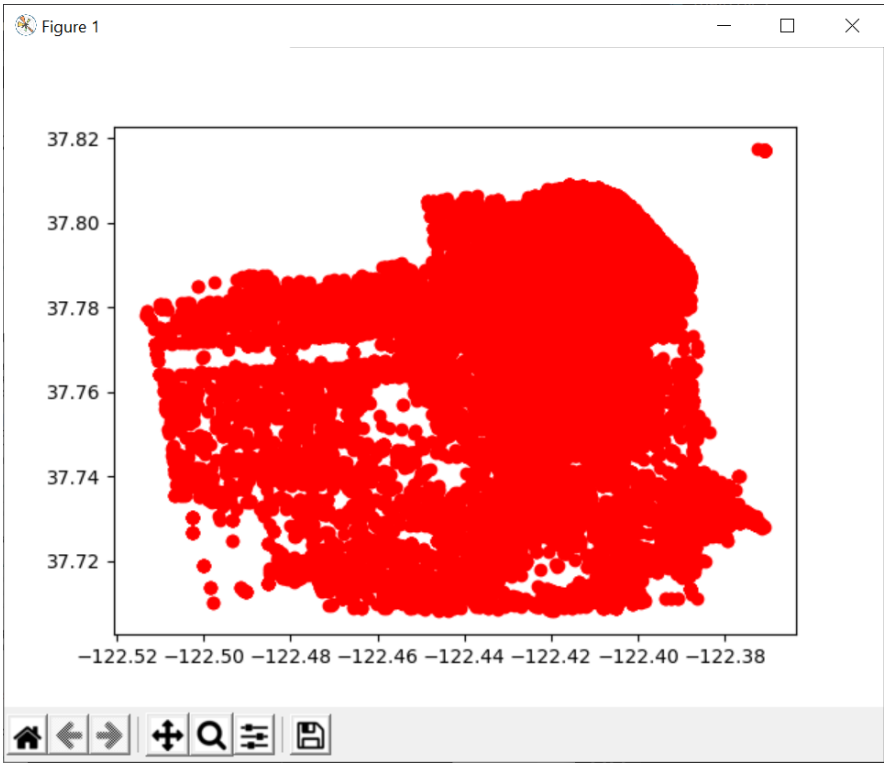
Кількість злочинів скоєних 4 Липня: 341

## Частина 5: Представлення даних

**Крок 1: Побудуйте графік дата фрейму SF, використовуючи змінні X та Y.**

а) Використовуйте функцію plot()для побудови кадру даних SF.

```
#PART5
#step1
plt.plot(SF['X'],SF['Y'], 'ro')
plt.show()
```




б) Визначте номери відділів поліції, складіть словник pd_districts, щоб зв’язати їх рядок із цілим числом.

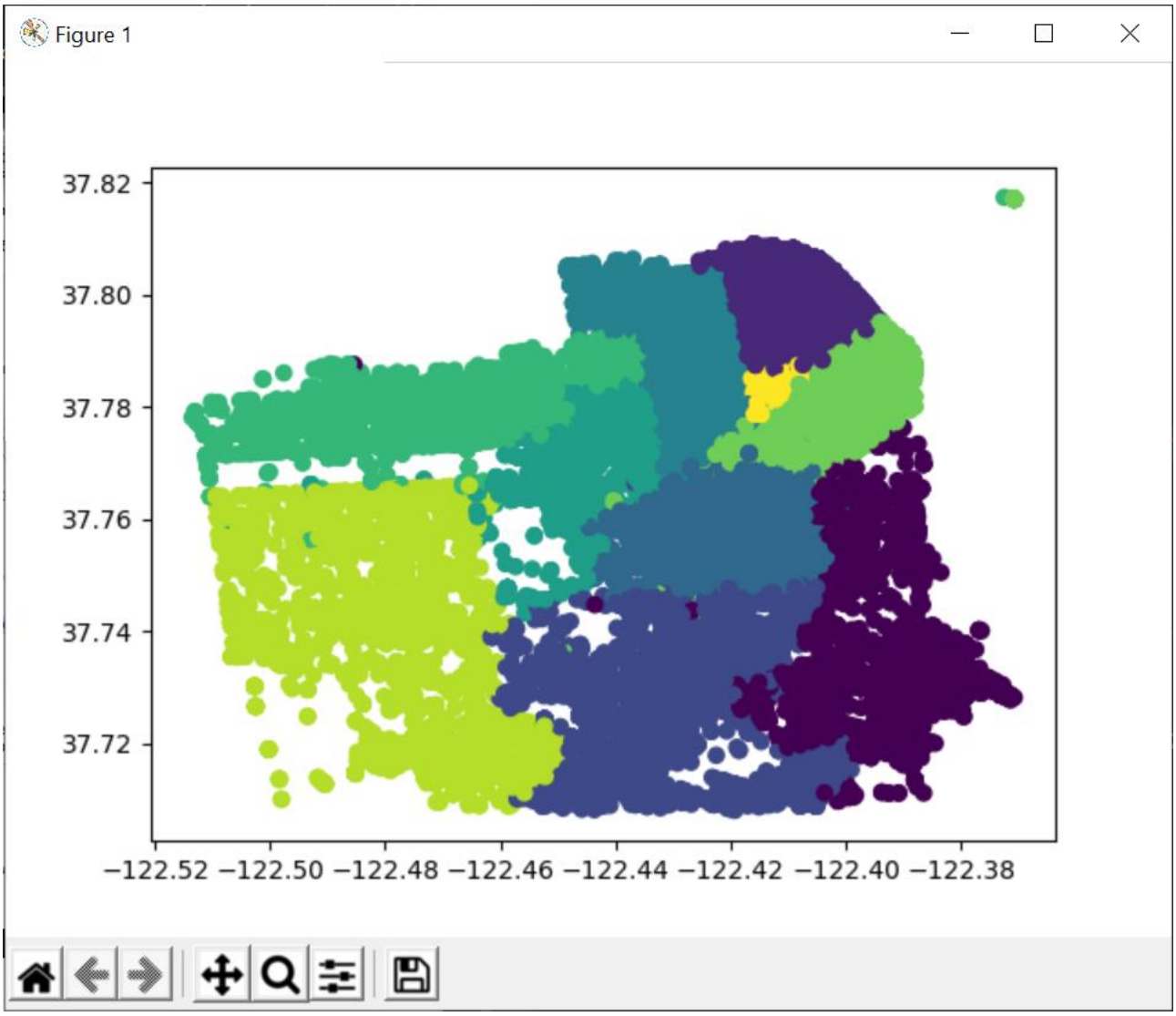
в) Використовуйте apply та, lambda, щоб додати ціле число поліцейського відділу до нового стовпця DataFrame

```
pd_districts = np.unique(SF['PdDistrict'])
pd_districts_levels = dict(zip(pd_districts, range(len(pd_districts))))

SF['PdDistrictCode'] = SF['PdDistrict'].apply(lambda row: pd_districts_levels[row])
print(pd_districts_levels)
```

```
C:\Users\datru\Desktop\study2021\big_data\lab2>py main.py
{'BAYVIEW': 0, 'CENTRAL': 1, 'INGLESIDE': 2, 'MISSION': 3, 'NORTHERN': 4, 'PARK': 5, 'RICHMOND
': 6, 'SOUTHERN': 7, 'TARAVAL': 8, 'TENDERLOIN': 9}

C:\Users\datru\Desktop\study2021\big_data\lab2>
```

г) Використовуйте щойно створений PdDistrictCode для автоматичної зміни кольору

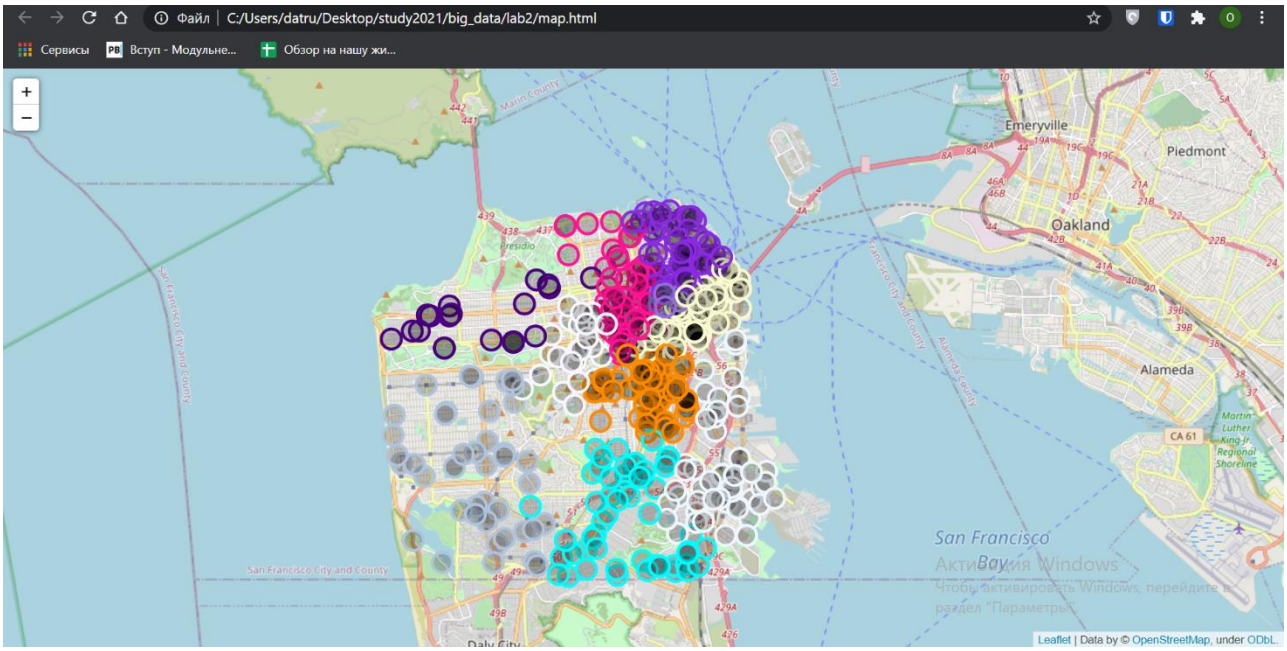
```
plt.scatter(SF['X'], SF['Y'], c=SF['PdDistrictCode'])
plt.show()
```




**Крок 2: Додайте пакети для побудови карт, щоб покращити сюжет.**

```
#step2
districts = np.unique(SF['PdDistrict'])
print('Colors:', list(colors.cnames.values())[0:len(districts)])
color_dict = dict(zip(districts, list(colors.cnames.values())[0:-1:len(districts)]
))
print(color_dict)

map_osm = folium.Map(location=[SF['Y'].mean(), SF['X'].mean()], zoom_start = 12)
plotEvery = 50
obs = list(zip( SF['Y'], SF['X'], SF['PdDistrict']))
for el in obs[0:-1:plotEvery]:
    folium.CircleMarker(el[0:2], color=color_dict[el[2]], fill_color=el[2],radius=10).add_to(map_osm)

map_osm.save("map.html")
```

```
C:\Users\datru\Desktop\study2021\big_data\lab2>py main.py
Colors: ['#F0F8FF', '#FAEBD7', '#00FFFF', '#7FFFD4', '#F0FFFF', '#F5F5DC', '#FFE4C4', '#000000', '#FFEBCD', '#0
000FF']
{'BAYVIEW': '#F0F8FF', 'CENTRAL': '#8A2BE2', 'INGLESIDE': '#00FFFF', 'MISSION': '#FF8C00', 'NORTHERN': '#FF1493', 'PARK': '#F8F8FF', 'RI
CHMOND': '#4B0082', 'SOUTHERN': '#FAFAD2', 'TARAVAL': '#B0C4DE', 'TENDERLOIN': '#9370DB'}

C:\Users\datru\Desktop\study2021\big_data\lab2>
```



## Висновки

В ході виконанні лабораторної роботи були використані навички обробки набору даних про злочини в Сан-Франциско з допомогою бібліотек для мови Python, а саме Numpy, Pandas, Matplotlib та Folium. Ці бібліотеки були використані для зручної побудови графіків для візуалізації даних з набору.

Найбільше злочинів було скоєно в Південному районі, також більше всього було скоєно крадіжок.