Homework4_Fall21

October 5, 2021

1 Homework 4

Submitted by: Vinit Horakeri

Submission date: 09/05/2021

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

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

Once the *crimes* dataframe has been setup you could should proceed to obtain 3 meaningful data analysis results from processing the *crimes* dataframe. Four 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 cells) to provide some additional text with explanations of what you are doing.

[1]:
```
#EXECUTE THIS CELL to setup the modules you need
%matplotlib inline

import pandas as pd
import numpy as np
import requests
from io import StringIO
```

[8]:
```
#EXECUTE THIS CELL to load the dataset into your environment - a security␣
 ↪warning will appear. You can ignore it.
url="https://gitlab.gitlab.svc.cent-su.org/ccaicedo/652public/-/raw/master/
 ↪crimes_2018.csv"
csvdata=requests.get(url,verify=False).text  #this will generate a warning but␣
 ↪you can proceed
```

```
/opt/conda/lib/python3.9/site-packages/urllib3/connectionpool.py:1013:
InsecureRequestWarning: Unverified HTTPS request is being made to host
'gitlab.gitlab.svc.cent-su.org'. Adding certificate verification is strongly
advised. See: https://urllib3.readthedocs.io/en/1.26.x/advanced-usage.html#ssl-
warnings
  warnings.warn(
```

[ ]:

[3]:
```
#EXECUTE THIS CELL to setup the crimes dataframe with the data from dataset␣
↪correctly formatted
crimes=pd.read_csv(StringIO(csvdata),parse_dates=[0], index_col=[0])
```

[46]:
```
crimes.head()
```

[46]:

```
                          ID Case Number                 Block  IUCR  \
Date
2018-09-01 00:01:00  11646166    JC213529  082XX S INGLESIDE AVE  0810
2020-03-17 21:30:00  12014684    JD189901   039XX N LECLAIRE AVE  0820
2018-01-01 08:00:00  11645648    JC212959    024XX N MONITOR AVE  1153
2019-09-24 08:00:00  11864018    JC476123   022XX S MICHIGAN AVE  1154
2019-10-13 20:30:00  11859805    JC471592    024XX W CHICAGO AVE  0860

                            Primary Type  \
Date
2018-09-01 00:01:00              THEFT
2020-03-17 21:30:00              THEFT
2018-01-01 08:00:00  DECEPTIVE PRACTICE
2019-09-24 08:00:00  DECEPTIVE PRACTICE
2019-10-13 20:30:00              THEFT

                                              Description  \
Date
2018-09-01 00:01:00                              OVER $500
2020-03-17 21:30:00                          $500 AND UNDER
2018-01-01 08:00:00      FINANCIAL IDENTITY THEFT OVER $ 300
2019-09-24 08:00:00  FINANCIAL IDENTITY THEFT $300 AND UNDER
2019-10-13 20:30:00                            RETAIL THEFT

                           Location Description  Arrest  Domestic  Beat  \
Date
2018-09-01 00:01:00                   RESIDENCE   False      True   631
2020-03-17 21:30:00                      STREET   False     False  1634
2018-01-01 08:00:00                   RESIDENCE   False     False  2515
2019-09-24 08:00:00  COMMERCIAL / BUSINESS OFFICE   False     False   132
2019-10-13 20:30:00           GROCERY FOOD STORE   False     False  1221

                     …  Ward  Community Area  FBI Code X Coordinate  \
Date                 …
2018-09-01 00:01:00  …   8.0            44.0        06          NaN
2020-03-17 21:30:00  …  45.0            15.0        06    1141659.0
2018-01-01 08:00:00  …  30.0            19.0        11          NaN
2019-09-24 08:00:00  …   3.0            33.0        11    1177560.0
2019-10-13 20:30:00  …  26.0            24.0        06    1160005.0
```

```
                     Y Coordinate  Year              Updated On   Latitude  \
Date
2018-09-01 00:01:00           NaN  2018  04/06/2019 04:04:43 PM        NaN
2020-03-17 21:30:00     1925649.0  2020  03/25/2020 03:45:43 PM  41.952052
2018-01-01 08:00:00           NaN  2018  04/06/2019 04:04:43 PM        NaN
2019-09-24 08:00:00     1889548.0  2019  10/20/2019 03:56:02 PM  41.852248
2019-10-13 20:30:00     1905256.0  2019  10/20/2019 04:03:03 PM  41.895732

                     Longitude                      Location
Date
2018-09-01 00:01:00        NaN                           NaN
2020-03-17 21:30:00 -87.754660  (41.952051946, -87.754660372)
2018-01-01 08:00:00        NaN                           NaN
2019-09-24 08:00:00 -87.623786  (41.852248185, -87.623786256)
2019-10-13 20:30:00 -87.687784  (41.895732399, -87.687784384)

[5 rows x 21 columns]
```

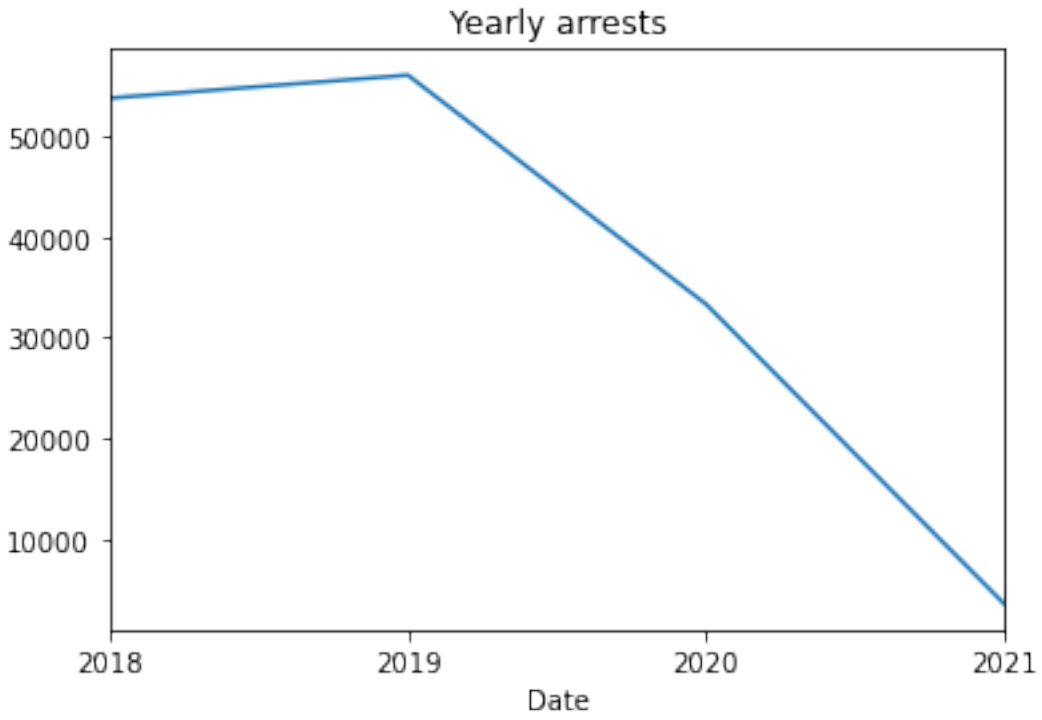
2 Code for data analysis 1

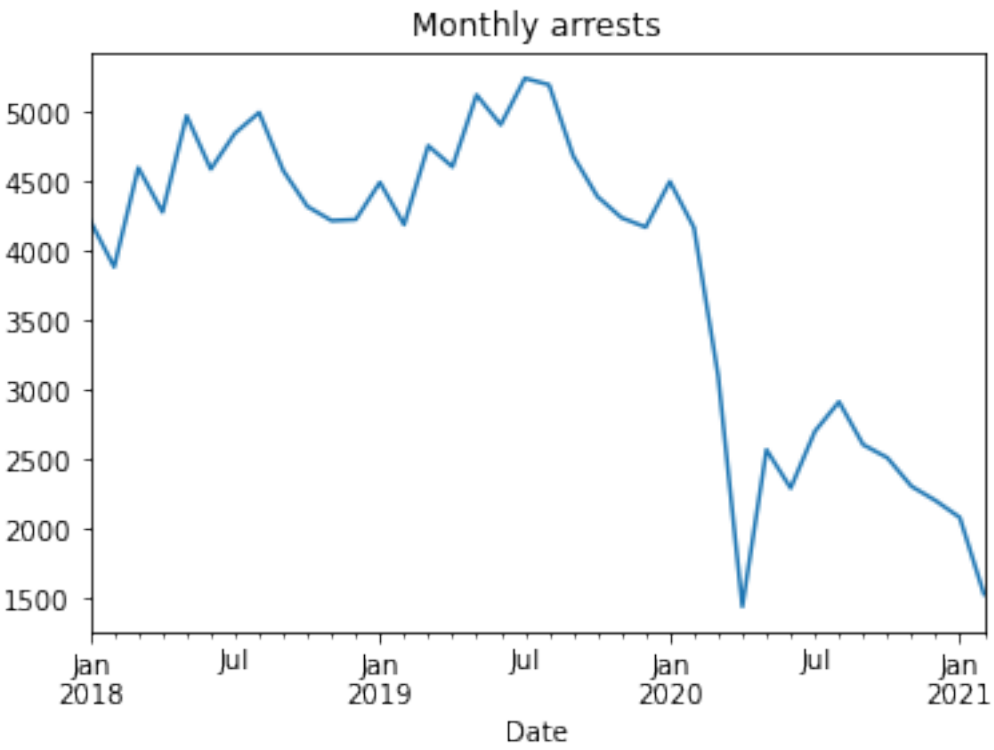
You can place the code for your first data analysis result in this section. Add as many code cells as you need.

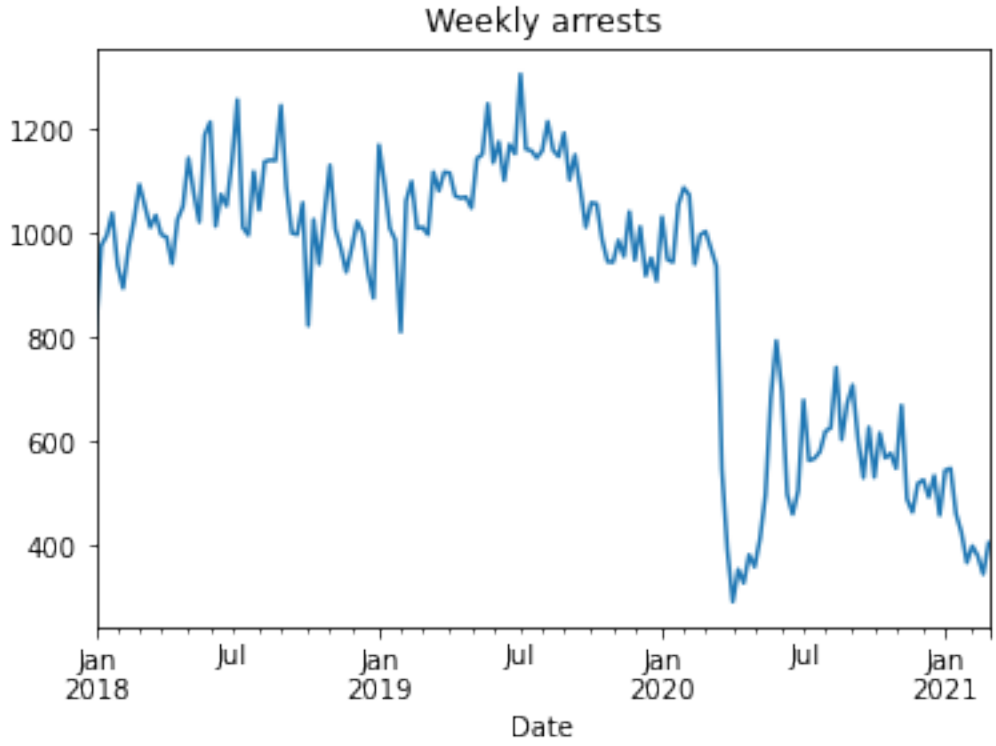
[ ]:
```
crimes_2018 = crimes.loc['2018'] # store all the crimes that occured in 2018␣
↪into a df
crimes_2019 = crimes.loc['2019'] # store all the crimes that occured in 2018␣
↪into a df
crimes_2020 = crimes.loc['2020'] # store all the crimes that occured in 2018␣
↪into a df

arrest_yearly = crimes[crimes['Arrest'] == True]['Arrest'] ## storing all the␣
↪arrests into arrest yearly df
```

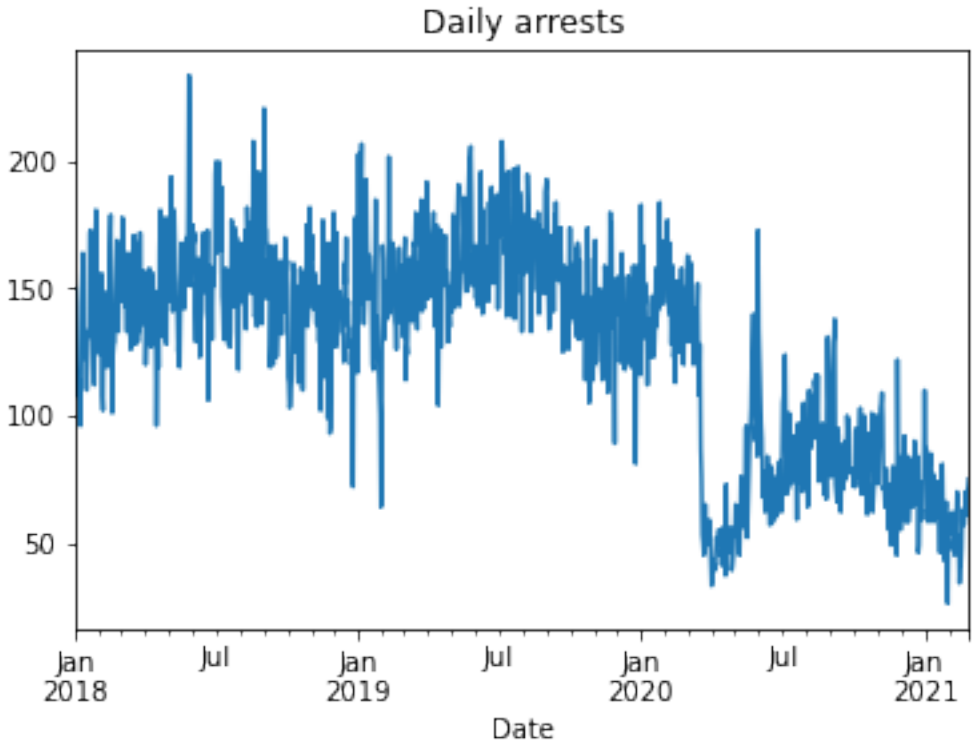
[47]:
```
import matplotlib.pyplot as plt
plt.subplot()
# yearly arrest
arrest_yearly.resample('A').sum().plot() ## resample by year
plt.title('Yearly arrests')
plt.show()
# Monthly arrest
arrest_yearly.resample('M').sum().plot()  ## resample by M
plt.title('Monthly arrests')
plt.show()
# Weekly arrest
arrest_yearly.resample('W').sum().plot()  # resample by W
```

```
plt.title('Weekly arrests')
plt.show()
# daily arrest
arrest_yearly.resample('D').sum().plot() # resample by D
plt.title('Daily arrests')
plt.show()
plt.show()
```




Monthly arrests
5000
4500
4000
3500
3000
2500
2000
1500
Jan
2018
Jul
Jan
2019
Jul
Jan
2020
Jul
Jan
2021
Date


Weekly arrests
1200
1000
800
600
400
Jan
2018
Jul
Jan
2019
Jul
Jan
2020
Jul
Jan
2021
Date



2.1 Description of data analysis result 1

Use the next cell to describe your data analysis result 1

2.1.1 1. Yearly Arrests - From the plot it is clear that the yearly arrests have gone down from 2018 to 2021, In 2020 it was around 20k and in 2021 it is below 15000.

2.1.2 2. Monthly Arrests - In march 2020, there was a huge dip in monthly cases, The main reason for this can be the COVID - 19 lockdown restrictions.

2.1.3 3. Daily Arrests - The daily arrests are also in downtrend from the plot

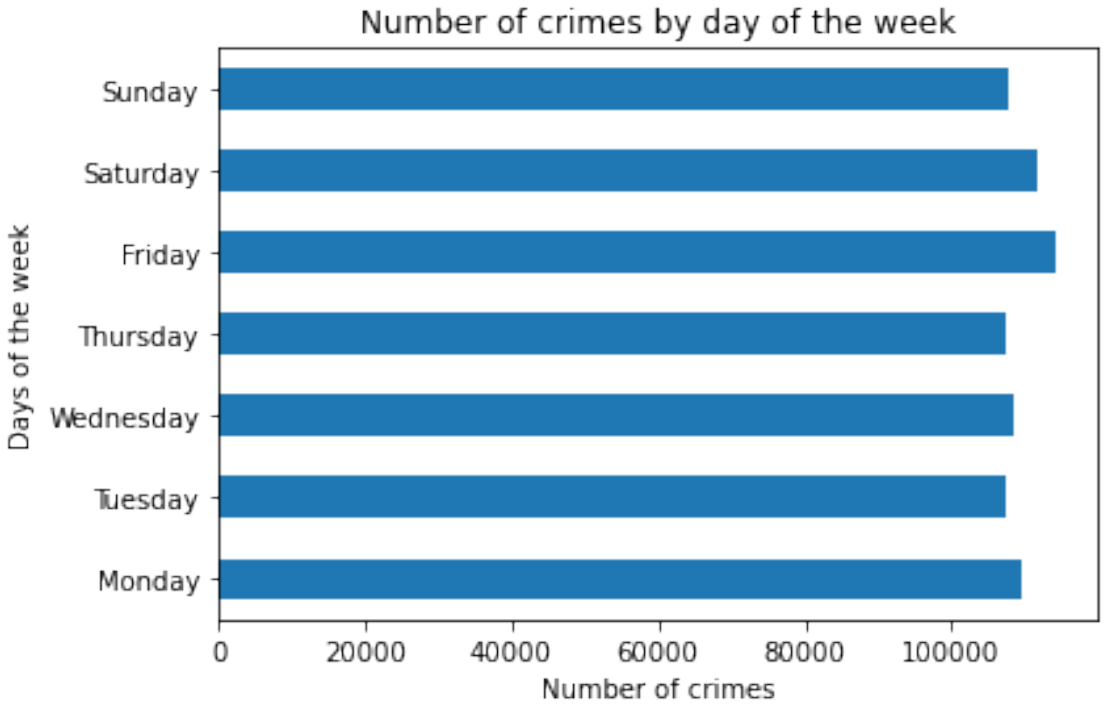
3 Code for data analysis 2

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

The first thing we are going to look at is if there is a difference in the number of crimes during specific days of the week. Are there more crimes during weekdays or weekend?

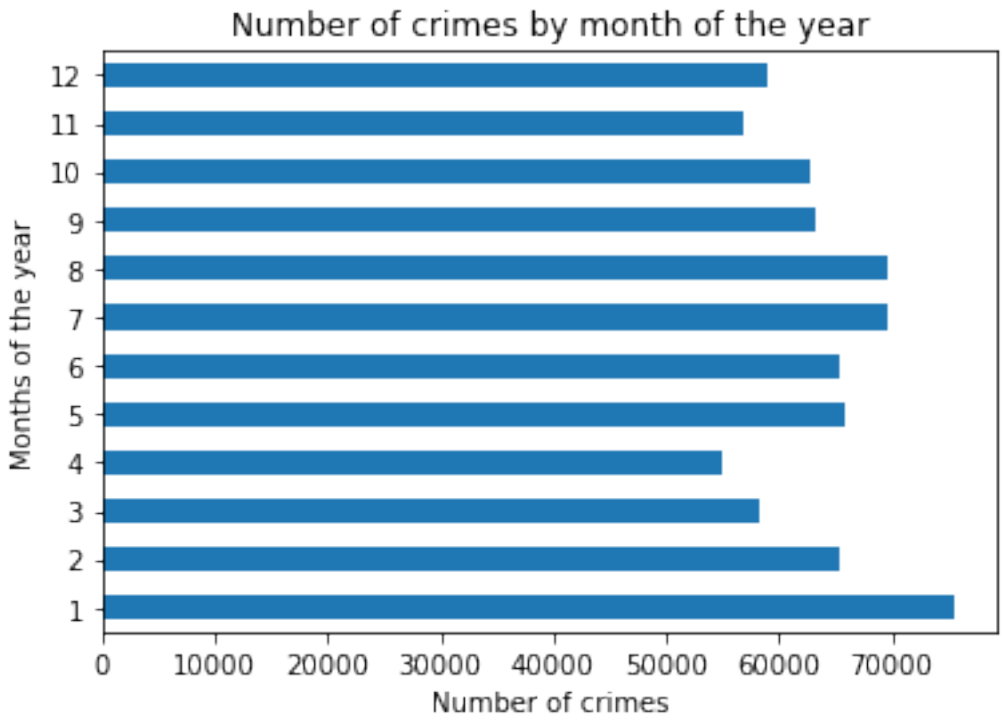
[54]:
```
days = ['Monday','Tuesday','Wednesday',  'Thursday', 'Friday', 'Saturday',␣
↪'Sunday']
```

```
crimes.groupby([crimes.index.dayofweek]).size().plot(kind='barh')
plt.ylabel('Days of the week')
plt.yticks(np.arange(7), days)
plt.xlabel('Number of crimes')
plt.title('Number of crimes by day of the week')
plt.show()
```



Now Let's look at crimes per month and see if certain months show more crimes than others.

[22]:
```
crimes.groupby([crimes.index.month]).size().plot(kind='barh')
plt.ylabel('Months of the year')
plt.xlabel('Number of crimes')
plt.title('Number of crimes by month of the year')
plt.show()
```



Crimes rates seem to peak at summer months!

3.1 Description of data analysis result 2

3.1.1 Number of crimes by day of the week - There is not much difference in number of crimes by days of the week except for friday which is a little higher

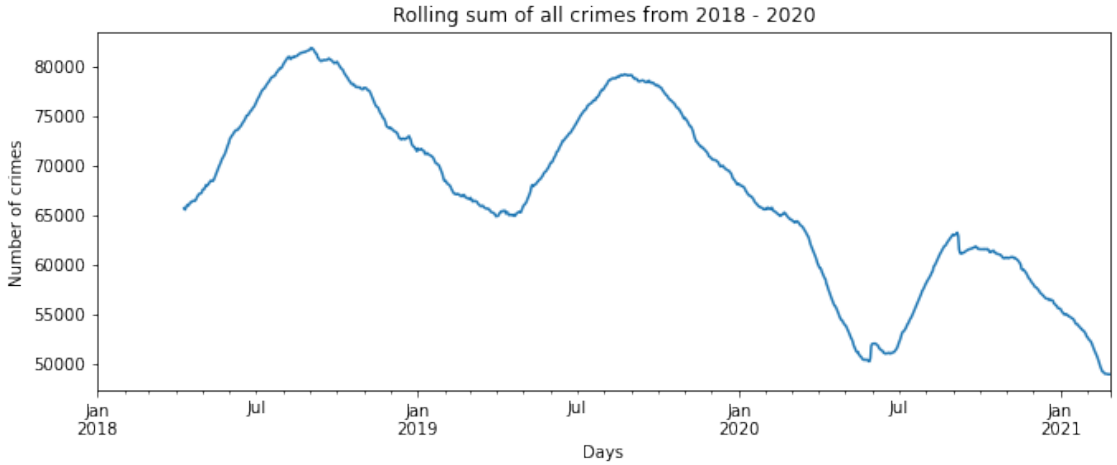
3.1.2 Number of crimes by month of the year - the 1st month has more crimes than any other month, while the 4th month has lowest crimes

4 Code for data analysis 3

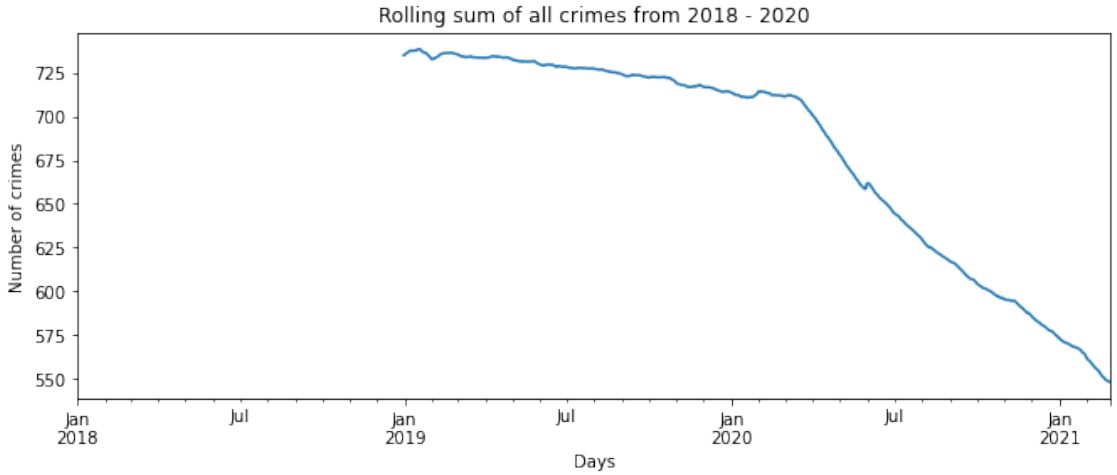
You can place the code for your third data analysis result in this section. Add as many code cells as you need.

[ ]:

[43]:
```
plt.figure(figsize=(11,4))
crimes.resample('D').size().rolling(100).sum().plot() ## rolling sum with step␣
↪size 100
plt.title('Rolling sum of all crimes from 2018 - 2020') ## title of the plot
plt.ylabel('Number of crimes')
plt.xlabel('Days')
plt.show()
```



[45]:
```
plt.figure(figsize=(11,4))
crimes.resample('D').size().rolling(365).mean().plot() ## rolling mean with␣
 ↪step size 365
plt.title('Rolling sum of all crimes from 2018 - 2020')
plt.ylabel('Number of crimes')
plt.xlabel('Days')
plt.show()
```



4.1 Description of data analysis result 3

4.1.1 Rolling Sum (Step size 100) - the rolling sum of the crimes have decreased over the years from 2018 to 2021 from a peak of 80000 to a lowest of 50000

4.1.2 Rolling Mean (Step Size 365) - the rolling mean of the crimes have decreased from a peak 725 to 550