Covid-19_UK

September 24, 2020

#

Coronavirus Pandemic (COVID-19)

##

Country Profile: United Kingdom

###

by Noaman Mangera

0.1 Table of Contents

0.2 Introduction

This document explores the development of an infectious disease caused by a type of coronavirus, known as SARS-CoV-2.

The dataset is a collection of the COVID-19 data maintained by Our World in Data. It is updated daily and includes metrics on confirmed cases, deaths, and testing, as well as other variables of potential interest. A description of each variable is made available within the same repository in the csv labelled 'codebook.csv', along with the data source for each variable in the dataset.

[1]:

```python
#import necessary modules
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from urllib.request import urlretrieve
```

```
[2]: #plot withinin notebook environment
     %matplotlib inline
```

```
[3]: #prepare visualisations in notebook by setting a theme, a default plot size,␣
      ↪font and color
     sns.set_style('darkgrid')
     plt.rcParams['font.size'] = 14
     plt.rcParams['figure.figsize'] = (9,5)
     plt.rcParams['figure.facecolor'] = '#00000000'
```

0.3 Gather Data

```
[4]: #download data from owid and save file locally
     urlretrieve('https://covid.ourworldindata.org/data/owid-covid-data.csv',
                 'covid-daywise.csv')
```

```
[4]: ('covid-daywise.csv', <http.client.HTTPMessage at 0x12540f78a08>)
```

```
[5]: #read in locally saved csv into dataframe
     covid_df = pd.read_csv('covid-daywise.csv', index_col='date')
```

0.4 Assess & Clean

```
[6]: #visually inspect first five rows
     covid_df.head()
```

```
[6]:            iso_code continent     location  total_cases  new_cases  \
     date
     2019-12-31      AFG      Asia  Afghanistan          0.0        0.0
     2020-01-01      AFG      Asia  Afghanistan          0.0        0.0
     2020-01-02      AFG      Asia  Afghanistan          0.0        0.0
     2020-01-03      AFG      Asia  Afghanistan          0.0        0.0
     2020-01-04      AFG      Asia  Afghanistan          0.0        0.0

                 new_cases_smoothed  total_deaths  new_deaths  new_deaths_smoothed  \
     date
     2019-12-31                 NaN           0.0         0.0                  NaN
     2020-01-01                 NaN           0.0         0.0                  NaN
     2020-01-02                 NaN           0.0         0.0                  NaN
     2020-01-03                 NaN           0.0         0.0                  NaN
     2020-01-04                 NaN           0.0         0.0                  NaN

                 total_cases_per_million  …  gdp_per_capita  extreme_poverty  \
     date                                 …
     2019-12-31                      0.0  …        1803.987              NaN
     2020-01-01                      0.0  …        1803.987              NaN
```

```
2020-01-02                      0.0  …       1803.987                NaN
2020-01-03                      0.0  …       1803.987                NaN
2020-01-04                      0.0  …       1803.987                NaN

            cardiovasc_death_rate  diabetes_prevalence  female_smokers  \
date
2019-12-31                597.029                 9.59             NaN
2020-01-01                597.029                 9.59             NaN
2020-01-02                597.029                 9.59             NaN
2020-01-03                597.029                 9.59             NaN
2020-01-04                597.029                 9.59             NaN

            male_smokers  handwashing_facilities  hospital_beds_per_thousand  \
date
2019-12-31           NaN                  37.746                         0.5
2020-01-01           NaN                  37.746                         0.5
2020-01-02           NaN                  37.746                         0.5
2020-01-03           NaN                  37.746                         0.5
2020-01-04           NaN                  37.746                         0.5

            life_expectancy  human_development_index
date
2019-12-31            64.83                    0.498
2020-01-01            64.83                    0.498
2020-01-02            64.83                    0.498
2020-01-03            64.83                    0.498
2020-01-04            64.83                    0.498

[5 rows x 40 columns]
```

[7]:
```
#number of columns and rows
covid_df.shape
print('This dataset contains {} rows and {} columns.'.format(covid_df.shape[0],␣
 ↪covid_df.shape[1]))
```

```
This dataset contains 45639 rows and 40 columns.
```

[8]:
```
#column names and data types
covid_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 45639 entries, 2019-12-31 to 2020-09-23
Data columns (total 40 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   iso_code                         45371 non-null  object
 1   continent                        45103 non-null  object
 2   location                         45639 non-null  object
 3   total_cases                      45025 non-null  float64
 4   new_cases                        44821 non-null  float64
 5   new_cases_smoothed               44039 non-null  float64
 6   total_deaths                     45025 non-null  float64
 7   new_deaths                       44821 non-null  float64
 8   new_deaths_smoothed              44039 non-null  float64
 9   total_cases_per_million          44757 non-null  float64
 10  new_cases_per_million            44757 non-null  float64
 11  new_cases_smoothed_per_million   43974 non-null  float64
 12  total_deaths_per_million         44757 non-null  float64
 13  new_deaths_per_million           44757 non-null  float64
 14  new_deaths_smoothed_per_million  43974 non-null  float64
 15  new_tests                        16212 non-null  float64
 16  total_tests                      16608 non-null  float64
 17  total_tests_per_thousand         16608 non-null  float64
 18  new_tests_per_thousand           16212 non-null  float64
 19  new_tests_smoothed               18184 non-null  float64
 20  new_tests_smoothed_per_thousand  18184 non-null  float64
 21  tests_per_case                   16683 non-null  float64
 22  positive_rate                    17111 non-null  float64
 23  tests_units                      18997 non-null  object
 24  stringency_index                 37847 non-null  float64
 25  population                       45371 non-null  float64
 26  population_density               43308 non-null  float64
 27  median_age                       40706 non-null  float64
 28  aged_65_older                    40102 non-null  float64
 29  aged_70_older                    40495 non-null  float64
 30  gdp_per_capita                   40184 non-null  float64
 31  extreme_poverty                  26813 non-null  float64
 32  cardiovasc_death_rate            40714 non-null  float64
 33  diabetes_prevalence              42148 non-null  float64
 34  female_smokers                   31925 non-null  float64
 35  male_smokers                     31522 non-null  float64
 36  handwashing_facilities           19030 non-null  float64
 37  hospital_beds_per_thousand       36799 non-null  float64
 38  life_expectancy                  44801 non-null  float64
 39  human_development_index          39284 non-null  float64
dtypes: float64(36), object(4)
memory usage: 14.3+ MB
```

Observations:

The entire dataset contains approximately 45,000 recorded observations (this number will continue to increase as data is added daily) and 40 features (variables). The focus for this analysis will be a subset of this data, namely the headline figures cases, deaths and tests for the UK.

[9]:

```
covid_df.columns
```

```
[9]: Index(['iso_code', 'continent', 'location', 'total_cases', 'new_cases',
       'new_cases_smoothed', 'total_deaths', 'new_deaths',
       'new_deaths_smoothed', 'total_cases_per_million',
       'new_cases_per_million', 'new_cases_smoothed_per_million',
       'total_deaths_per_million', 'new_deaths_per_million',
       'new_deaths_smoothed_per_million', 'new_tests', 'total_tests',
       'total_tests_per_thousand', 'new_tests_per_thousand',
       'new_tests_smoothed', 'new_tests_smoothed_per_thousand',
       'tests_per_case', 'positive_rate', 'tests_units', 'stringency_index',
       'population', 'population_density', 'median_age', 'aged_65_older',
       'aged_70_older', 'gdp_per_capita', 'extreme_poverty',
       'cardiovasc_death_rate', 'diabetes_prevalence', 'female_smokers',
       'male_smokers', 'handwashing_facilities', 'hospital_beds_per_thousand',
       'life_expectancy', 'human_development_index'],
      dtype='object')
```

```
[51]: #subset data for UK
covid_uk_df = covid_df.loc[covid_df['location'] == 'United Kingdom',␣
 ↪['new_cases','new_cases_smoothed', 'total_cases', 'new_tests',␣
 ↪'new_deaths','new_deaths_smoothed', 'total_deaths',
                                                                  ␣
 ↪'new_tests_smoothed', 'total_tests', 'positive_rate']].copy()
covid_uk_df.head()
```

```
[51]:             new_cases  new_cases_smoothed  total_cases  new_tests  new_deaths  \
date
2019-12-31        0.0                 NaN          0.0        NaN         0.0
2020-01-01        0.0                 NaN          0.0        NaN         0.0
2020-01-02        0.0                 NaN          0.0        NaN         0.0
2020-01-03        0.0                 NaN          0.0        NaN         0.0
2020-01-04        0.0                 NaN          0.0        NaN         0.0

            new_deaths_smoothed  total_deaths  new_tests_smoothed  \
date
2019-12-31                  NaN           0.0                 NaN
2020-01-01                  NaN           0.0                 NaN
2020-01-02                  NaN           0.0                 NaN
2020-01-03                  NaN           0.0                 NaN
2020-01-04                  NaN           0.0                 NaN

            total_tests  positive_rate
date
2019-12-31          NaN            NaN
2020-01-01          NaN            NaN
2020-01-02          NaN            NaN
2020-01-03          NaN            NaN
2020-01-04          NaN            NaN
```

Observations:

Data is recorded from the 31/12/2019 onwards.

[52]:

```
#user defined function to calculate missing values
def missing_values_table(df):
        mis_val = df.isnull().sum()
        mis_val_percent = 100 * (df.isnull().sum() / len(df))
        mis_val_table = pd.concat([mis_val, mis_val_percent], axis=1)
        mis_val_table_ren_columns = mis_val_table.rename(
        columns = {0 : 'Missing Values', 1 : '% of Total Values'})
        mis_val_table_ren_columns = mis_val_table_ren_columns[
            mis_val_table_ren_columns.iloc[:,1] != 0].sort_values(
        '% of Total Values', ascending=False).round(1)
        print ("Your selected dataframe has " + str(df.shape[1]) + " columns.\n"
            "There are " + str(mis_val_table_ren_columns.shape[0]) +
              " columns that have missing values.")
        return mis_val_table_ren_columns
```

[53]:

```
#apply user defined function over subset of data
missing_values_table(covid_uk_df)
```

```
Your selected dataframe has 10 columns.
There are 6 columns that have missing values.
```

[53]:

```
                     Missing Values  % of Total Values
new_tests_smoothed              100               37.3
positive_rate                   100               37.3
new_tests                        93               34.7
total_tests                      93               34.7
new_cases_smoothed                6                2.2
new_deaths_smoothed               6                2.2
```

Observations:

There is less data available for the number of new tests recorded (contains more null values) than the other variables.

The distinction between 0 and null values is subtle but important. In this dataset, it represents daily test numbers that were not reported on specific dates.

[54]:

```
#first reported day of testing
covid_uk_df.new_tests.first_valid_index()
```

[54]: `'2020-04-01'`

Observations:

The UK only started publishing daily tests numbers on the 01/04/2020.

0.5 Exploratory Data Analysis

0.5.1 Univariate Exploration

[55]:

```
#summary statistics of numerical variables
covid_uk_df.describe().T
```

[55]:

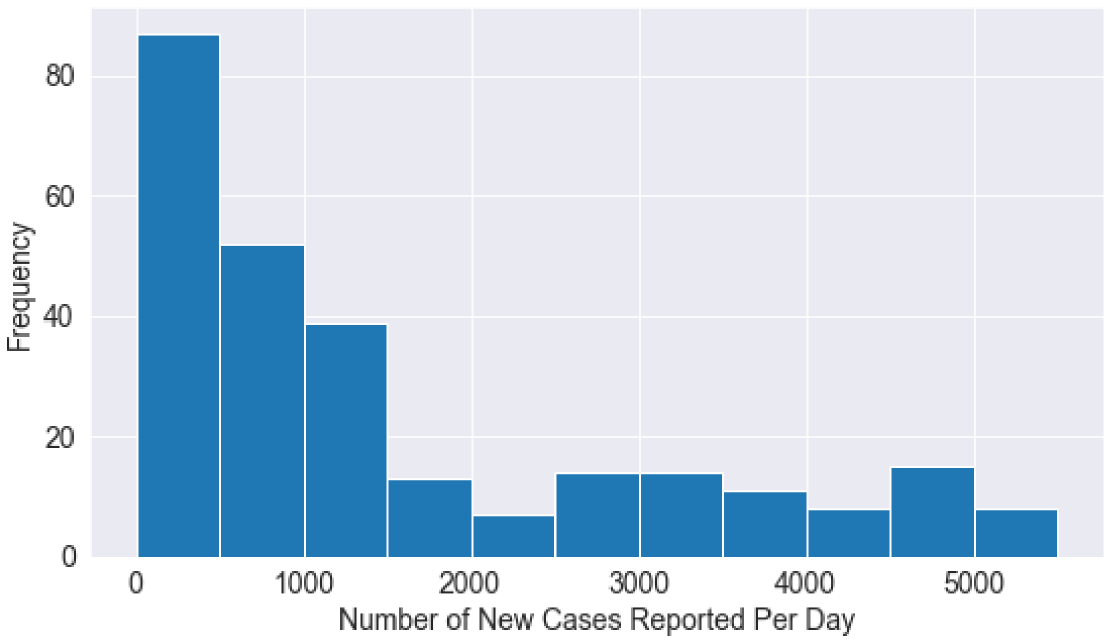
```
                    count          mean           std         min  \
new_cases           268.0  1.505787e+03  1.600387e+03       0.000
new_cases_smoothed  262.0  1.489945e+03  1.540290e+03       0.000
total_cases         268.0  1.697490e+05  1.397355e+05       0.000
new_tests           175.0  1.051608e+05  5.903500e+04   11896.000
new_deaths          268.0  1.560634e+02  2.775328e+02       0.000
new_deaths_smoothed 262.0  1.593620e+02  2.629501e+02       0.000
total_deaths        268.0  2.292306e+04  1.852375e+04       0.000
new_tests_smoothed  168.0  1.072603e+05  5.519494e+04   15713.000
total_tests         175.0  6.754594e+06  5.437085e+06  155174.000
positive_rate       168.0  4.520238e-02  7.650475e-02       0.004

                             25%           50%           75%           max
new_cases           5.575000e+01  9.120000e+02  2.597000e+03  5.487000e+03
new_cases_smoothed  7.985675e+01  9.623575e+02  2.521822e+03  4.846143e+03
total_cases         2.582500e+02  2.125130e+05  2.947822e+05  4.035510e+05
new_tests           6.694250e+04  9.317300e+04  1.528825e+05  2.525090e+05
new_deaths          0.000000e+00  1.800000e+01  1.565000e+02  1.224000e+03
new_deaths_smoothed 1.000000e+00  1.721450e+01  2.017142e+02  9.424290e+02
total_deaths        7.500000e-01  3.188600e+04  4.097725e+04  4.182500e+04
new_tests_smoothed  7.727400e+04  9.444700e+04  1.504452e+05  2.312570e+05
total_tests         1.905207e+06  5.604093e+06  1.078025e+07  1.889735e+07
positive_rate       6.000000e-03  1.200000e-02  3.350000e-02  3.020000e-01
```

Observations:

The standard deviation for the number of new cases, new deaths and new tests is signifiacnt, suggesting the mean is not an accurate measure of central tendency. This chimes with the appreciation that the disease has progressed at wildly different rates over the months.

[56]:

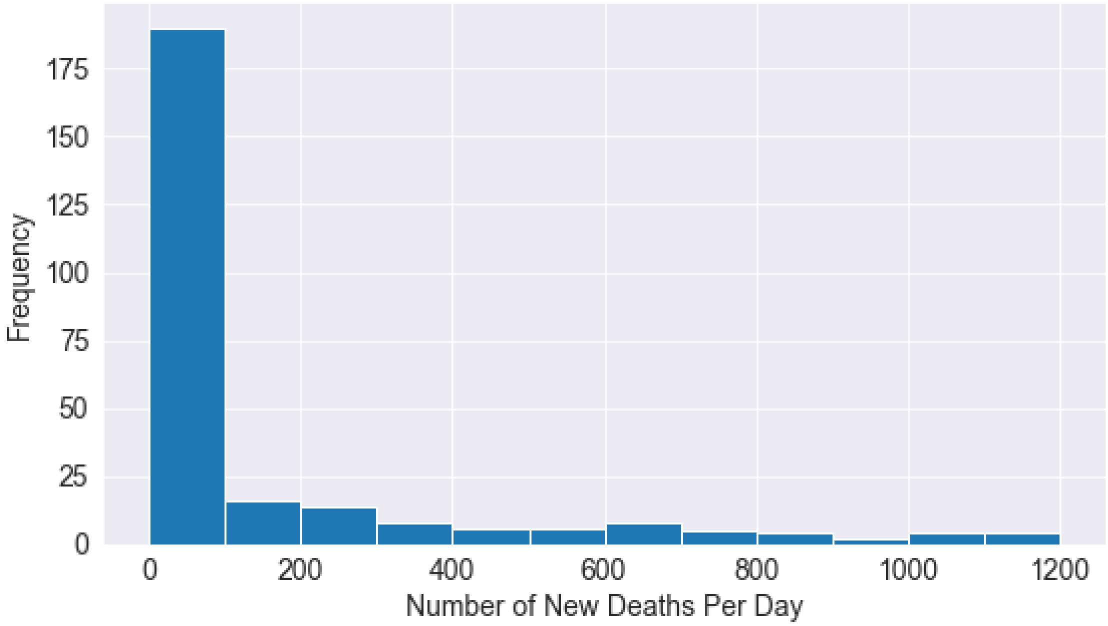
```
#plot histogram of number of new reported cases per day
plt.hist(covid_uk_df.new_cases, bins=np.arange(0, 6000, 500))

#set axis labels
plt.xlabel('Number of New Cases Reported Per Day');
plt.ylabel('Frequency');
```



Observations:

-Right skew, with the majority of days reporting less than 1000 new cases per day.

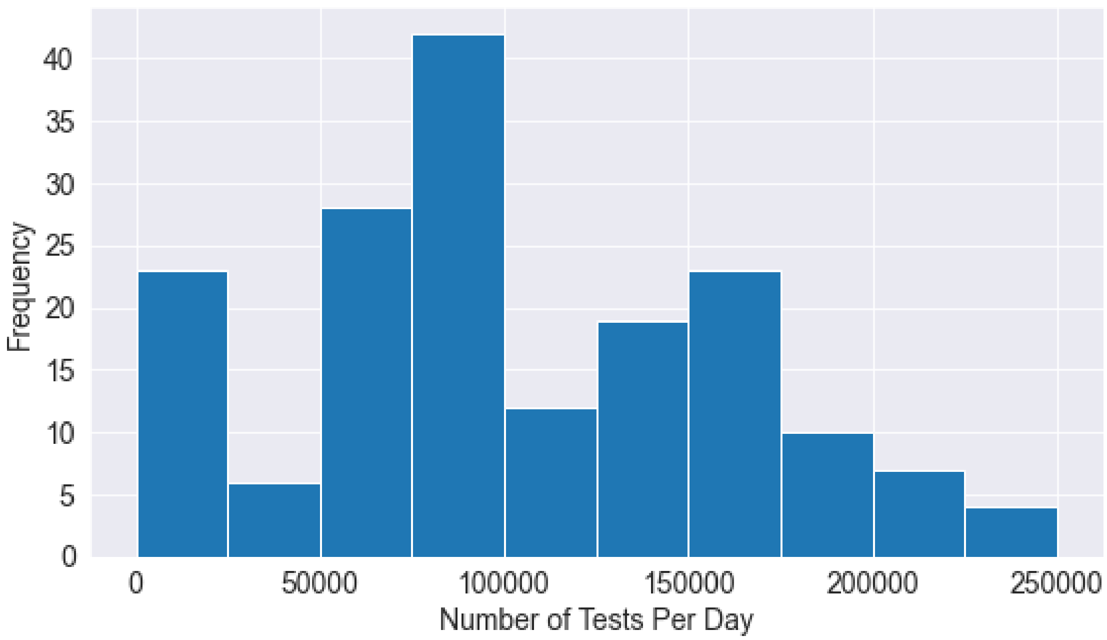
[57]:
```
#plot histgram with number of new deaths per day
plt.hist(covid_uk_df.new_deaths, bins=np.arange(0, 1300, 100))

#set axis labels
plt.xlabel('Number of New Deaths Per Day');
plt.ylabel('Frequency');
```



Observations:

-Right skew, with the majority of days reporting less than 100 new deaths per day.

[58]:
```
#plot histgram with number of new tests per day
plt.hist(covid_uk_df.new_tests, bins=np.arange(0, 275000,25000))

#set axis labels
plt.xlabel('Number of Tests Per Day');
plt.ylabel('Frequency');
```



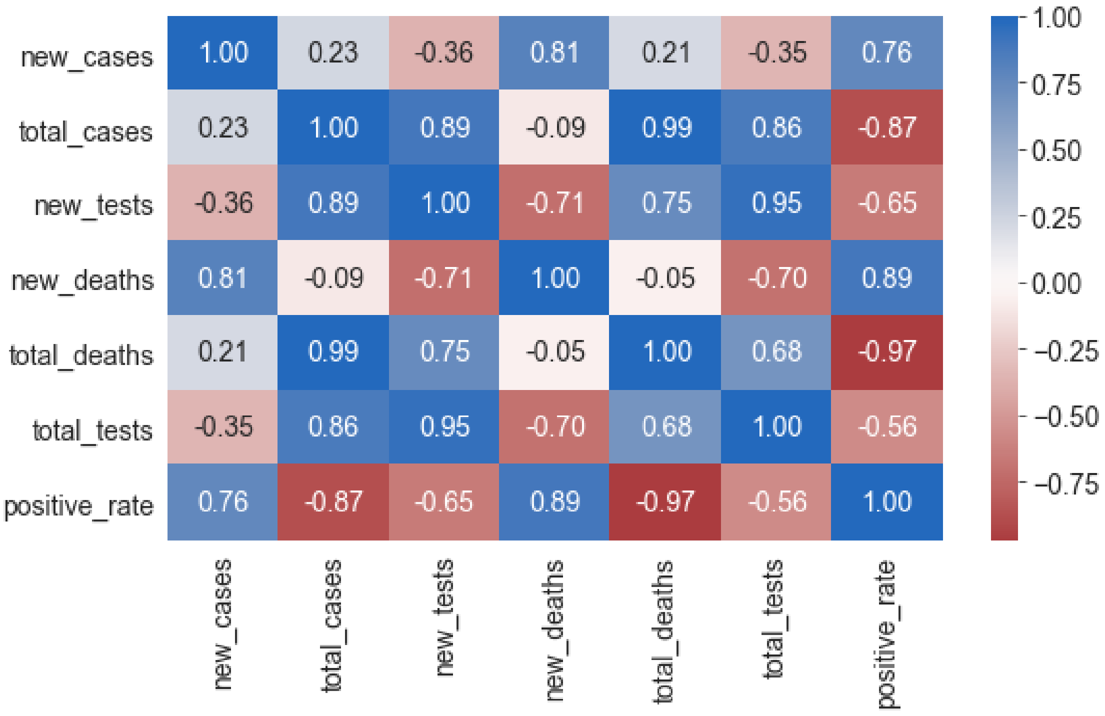
0.5.2 Bivariate Exploration

[59]:

```
#isolate for variables of interest
focus_vars = ['new_cases', 'total_cases', 'new_tests', 'new_deaths',␣
 ↪'total_deaths', 'total_tests', 'positive_rate']
```

[60]:

```
# correlation plot of numeric variables
sns.heatmap(covid_uk_df[focus_vars].corr(), annot = True, fmt = '.2f',
            cmap = 'vlag_r', center = 0);
```



	new_cases	total_cases	new_tests	new_deaths	total_deaths	total_tests	positive_rate
new_cases	1.00	0.23	-0.36	0.81	0.21	-0.35	0.76
total_cases	0.23	1.00	0.89	-0.09	0.99	0.86	-0.87
new_tests	-0.36	0.89	1.00	-0.71	0.75	0.95	-0.65
new_deaths	0.81	-0.09	-0.71	1.00	-0.05	-0.70	0.89
total_deaths	0.21	0.99	0.75	-0.05	1.00	0.68	-0.97
total_tests	-0.35	0.86	0.95	-0.70	0.68	1.00	-0.56
positive_rate	0.76	-0.87	-0.65	0.89	-0.97	-0.56	1.00

Observations:

- The number of reported cases and the number of deaths attributed to Covid-19 are highly correlated.
- The number of tests and the positive rate are inversely correlated. More people being tested means fewer people are actually diagnosed with Covid-19.

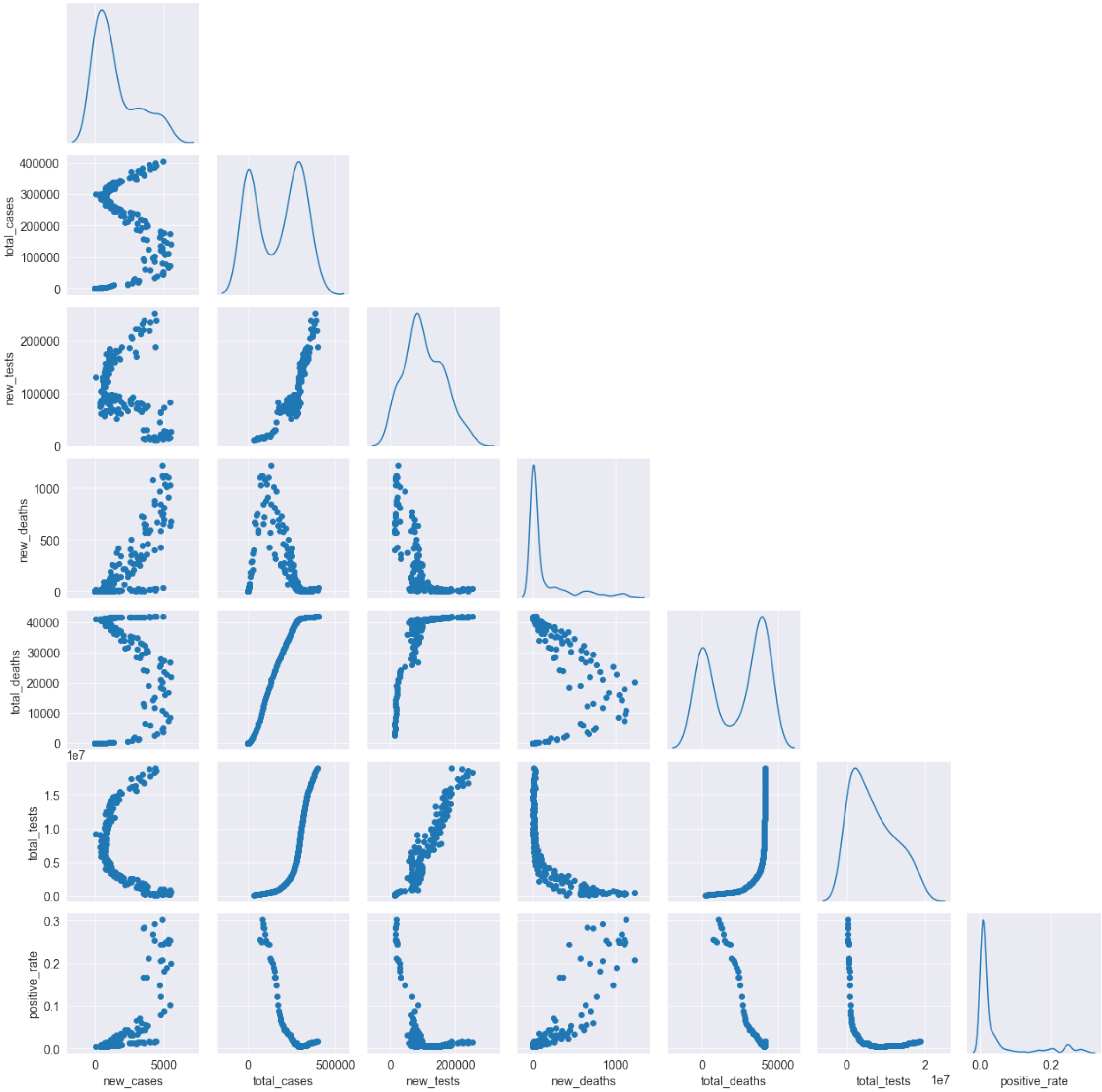
A word of caution:

1. Correlation does not imply causation. This means that although tests and deaths are inversely correlated, more testing does not necessarily lead to fewer fatalities.

2. Confounding variables are likely behind the correlations noted. For example, the positive rate is a composite measure of cases and tests, and therefore likely to influence the near perfect correlation between total deaths and the positive rate.

The heatmap above measures linear relationship. Scatter plots can be drawn to understand the presence of non-linear relationships.

[61]:
```
#pairwise plots of variables
g = sns.PairGrid(data = covid_uk_df, vars = focus_vars, diag_sharey=False, corner=True)
g.map_lower(plt.scatter)
g.map_diag(sns.kdeplot);
```



Observations:

The presence of higher order correlations between most variables.

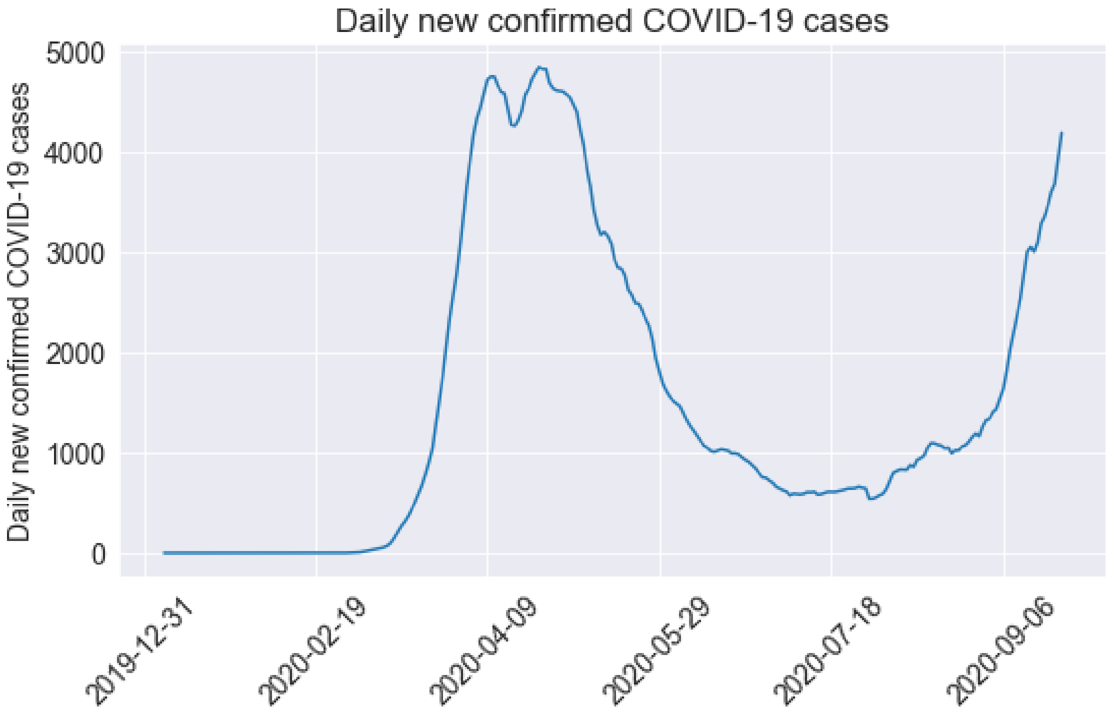
0.6 Question & Answers

For all data sources on the pandemic, daily data does not necessarily refer to the number of new confirmed cases on that day – but to the cases reported on that day. Since reporting can vary from day to day – irrespectively of any actual variation of cases – it is therefore helpful to look at a longer time span, which is less affected by the daily variation in reporting. This provides a clearer picture of where the pandemic is accelerating, staying the same, or reducing. A rolling average (7-day window) is therefore used to smooth short term variations.

Q: What is the daily number of confirmed cases?

[62]:

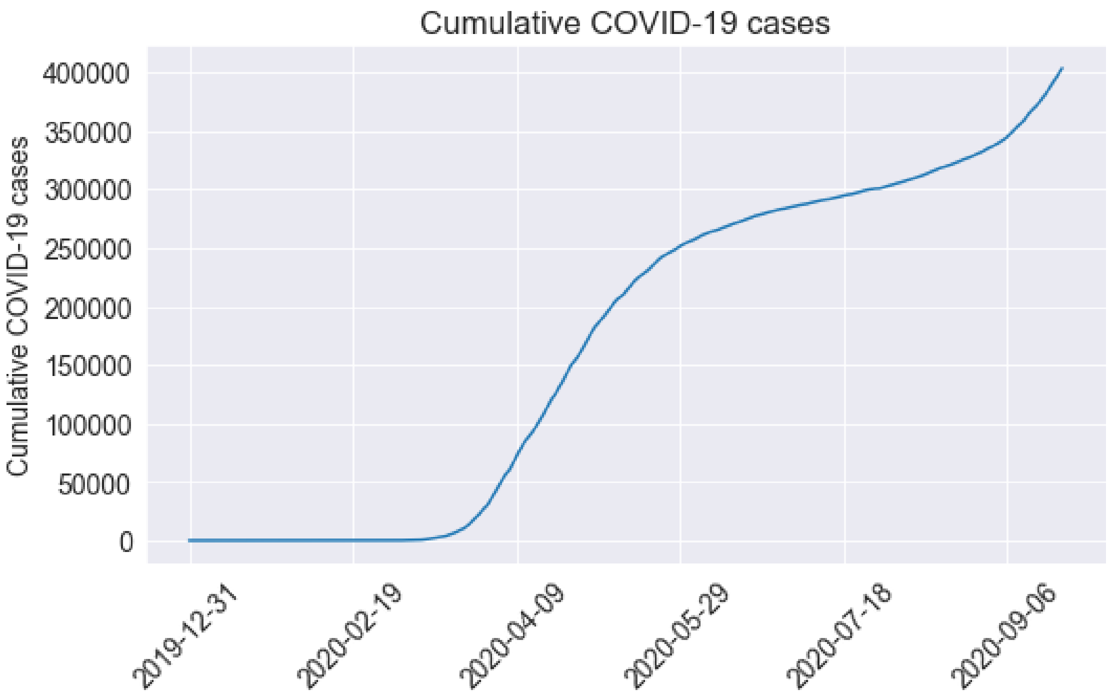
```
#plot line chart of number of new reported cases per day
covid_uk_df.new_cases_smoothed.plot()

#set title and axis labels
plt.title('Daily new confirmed COVID-19 cases')
plt.xticks(rotation=45)
plt.xlabel('')
plt.ylabel('Daily new confirmed COVID-19 cases');
```



Q: What is the total number of reported cases related to Covid-19 in the UK?

[63]:

```
#plot line chart of cumulative cases
covid_uk_df.total_cases.plot()

#set title and axis labels
plt.title('Cumulative COVID-19 cases')
plt.xticks(rotation=45)
plt.xlabel('')
plt.ylabel('Cumulative COVID-19 cases');
```

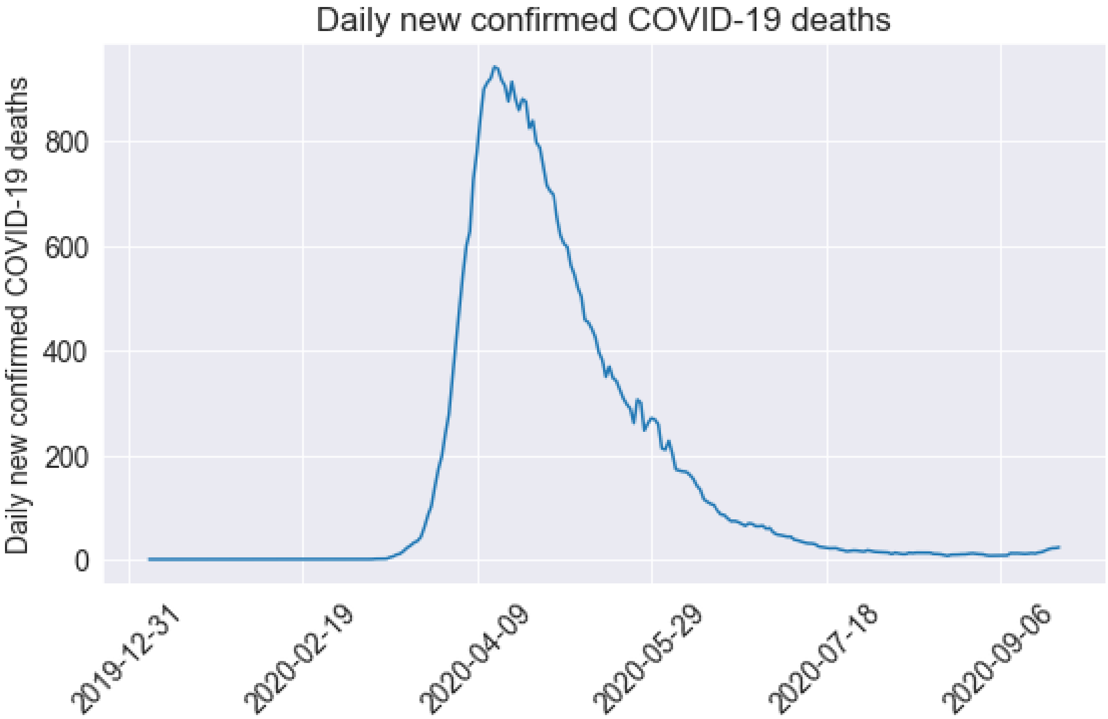


Observations:

The number of reported cases peaked at approx 4,900 on the 10th of April 2020. Since the 18th of July the number of daily reported cases has once again begun to grow. Is the UK prepared for a second wave?

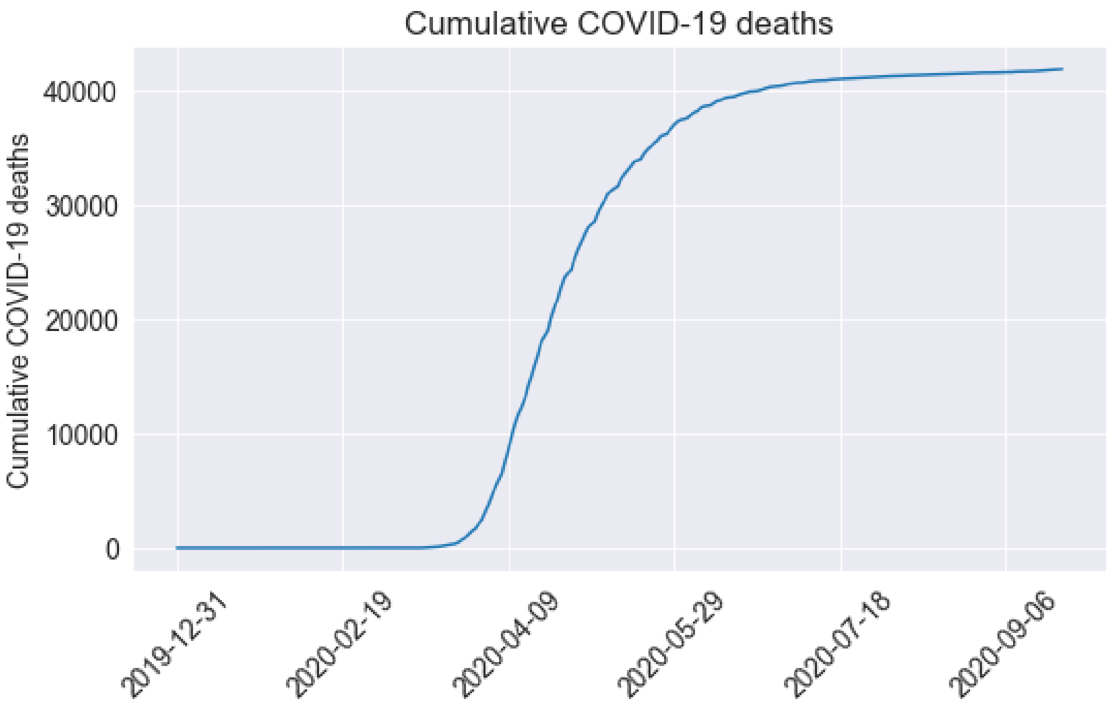
Q: What is the daily number of confirmed deaths?

[64]:
```
#plot line chart of new deaths per day
covid_uk_df.new_deaths_smoothed.plot()

#set title and axis labels
plt.title('Daily new confirmed COVID-19 deaths')
plt.xticks(rotation=45)
plt.xlabel('')
plt.ylabel('Daily new confirmed COVID-19 deaths');
```



[65]:

```
#plot line chart of cumulative deaths
covid_uk_df.total_deaths.plot()

#set title and axis labels
plt.title('Cumulative COVID-19 deaths')
plt.xticks(rotation=45)
plt.xlabel('')
plt.ylabel('Cumulative COVID-19 deaths');
```



Observations:

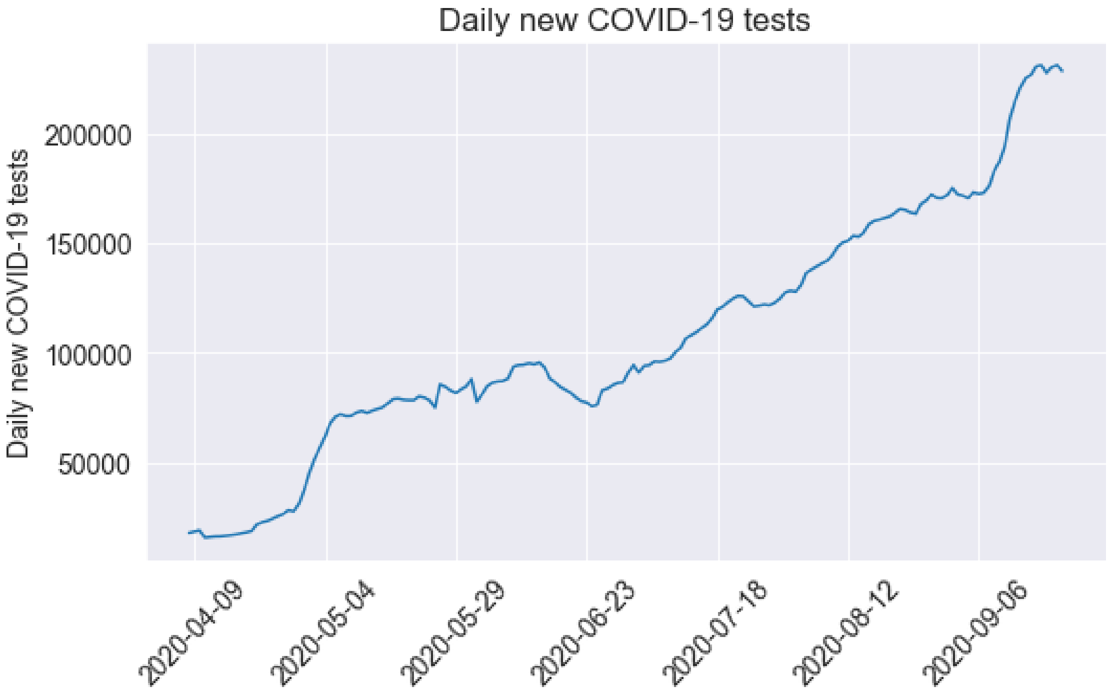
- Similar to the number of reported cases, the number of deaths peaked around the 10th of April 2020. Domain knowledge indicates the number of deaths should lag the number of cases by around 14 days. This is not clear from the data, raising questions about data consistency. A closer look at the literature reveals a retrospective revision in the number of deaths attributed to Covid-19.
- Given the rise in the number of reported daily cases, the number of daily confirmed deaths is expected to follow.

The widely available data on confirmed cases only becomes meaningful when it can be interpreted in light of how much a country is testing. Are countries testing enough to monitor the outbreak?

Q: What is the daily number of new tests?

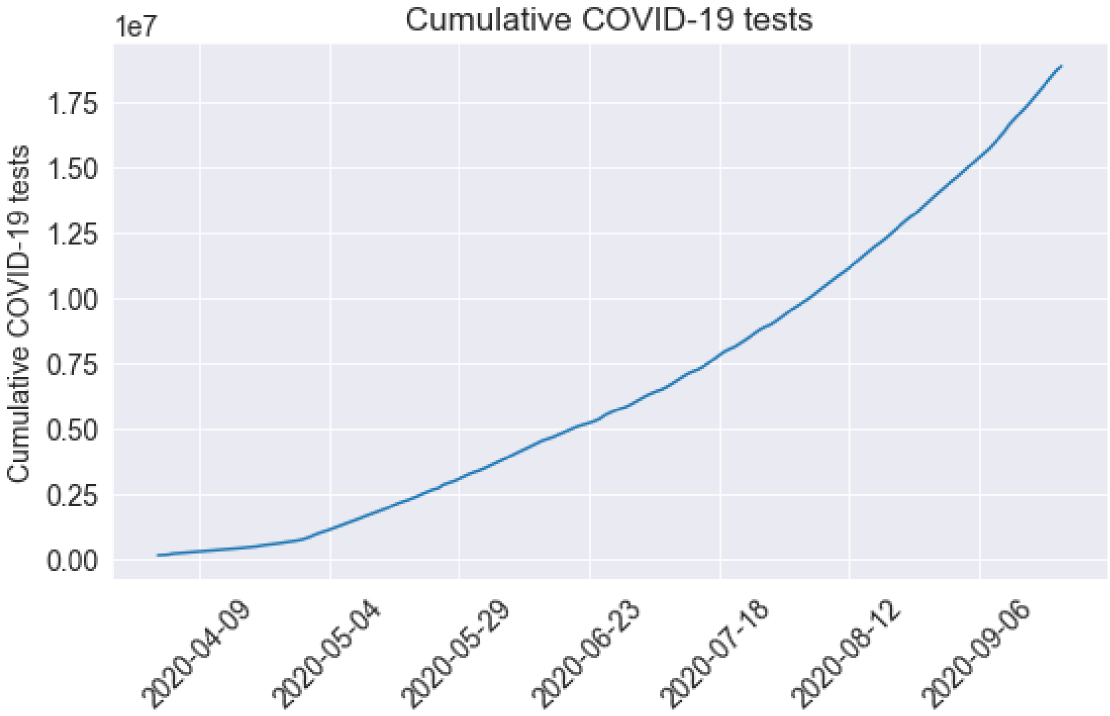
[66]:

```
#plot line chart of new tests per day
covid_uk_df.new_tests_smoothed.plot()

#set title, position of tick marks, and axis labels
plt.title('Daily new COVID-19 tests')
plt.xticks(rotation=45)
plt.xlabel('')
plt.ylabel('Daily new COVID-19 tests');
```



[67]:

```
#plot line chart of cumulative tests
covid_uk_df.total_tests.plot()

#set title, position of tick marks, and axis labels
plt.title('Cumulative COVID-19 tests')
plt.xticks(rotation=45)
plt.xlabel('')
plt.ylabel('Cumulative COVID-19 tests');
```



Observations:

As capacity is built the number of daily tests continues to rise.

Q: What is the death rate (ratio of confirmed deaths to reported cases)?

```
[68]: #create new variable by dividing exisiting variables
death_rate = covid_uk_df.total_deaths[-1] / covid_uk_df.total_cases[-1]

#print result
print("The latest reported 'death' rate in the UK is {}%.".
 ↪format(round(death_rate*100, 2)))
```

```
The latest reported 'death' rate in the UK is 10.36%.
```

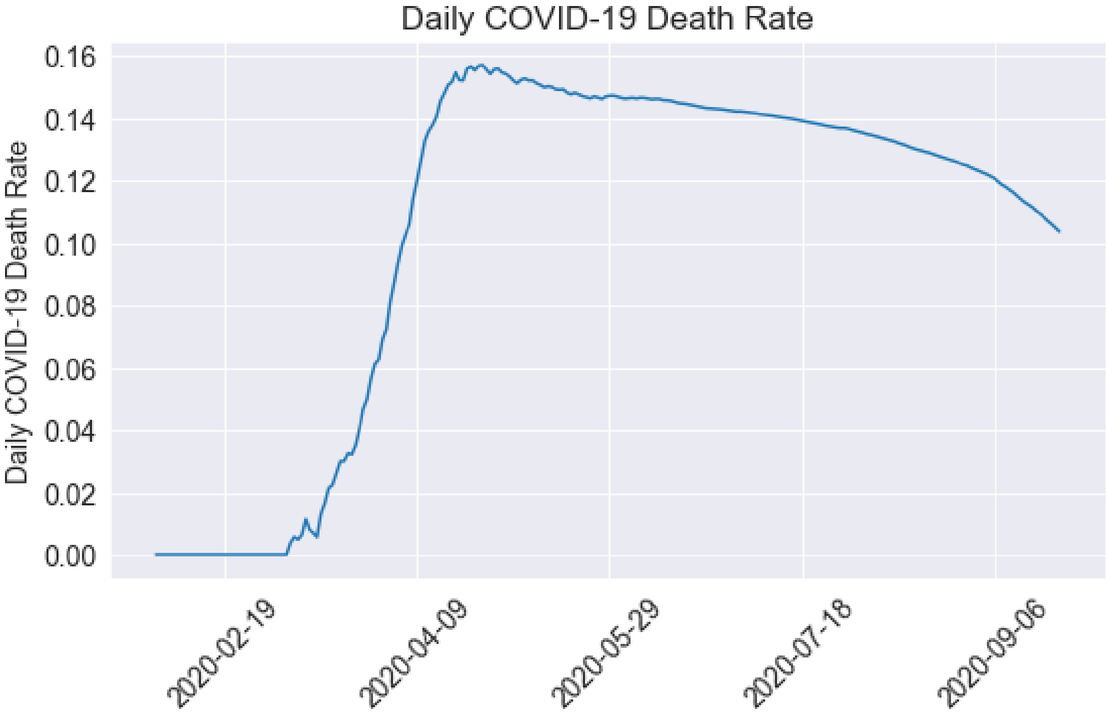
A word of caution:

This does not mean that 11% of people who contract the virus will suffer a fatality. The true number is likely to lower given many cases are asymptomatic, and yet many more cases are never diagnosed. To see this in play, consider the “death” rate as a function of time.

```
[69]: #create new variable by dividing exisiting variables element wise
covid_uk_df['death_rate_t'] = covid_uk_df.total_deaths / covid_uk_df.total_cases
```

```
[70]: #plot line chart of death rate
covid_uk_df.death_rate_t.plot()
```

```
#set title, position of tick marks, and axis labels
plt.title('Daily COVID-19 Death Rate')
plt.xticks(rotation=45)
plt.xlabel('')
plt.ylabel('Daily COVID-19 Death Rate');
```

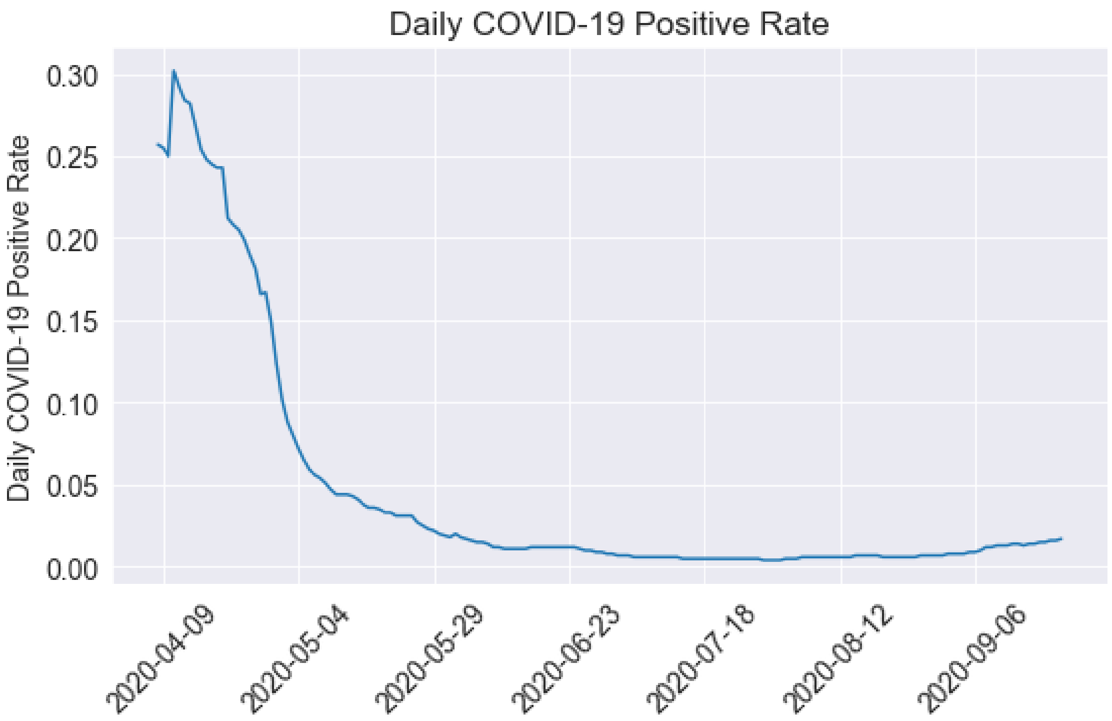


Observations:

At the height of the pandamic when testing was limited, the 'death rate' hovered around 16%. This number has steadily decreased as testing capacity is built. One important way to understand if countries are testing sufficiently is to ask: What share of the tests confirm a case? What is the positive rate?

Q: What fraction of test returned a positive result?

[71]:
```
#plot line chart of positive rate
covid_uk_df.positive_rate.plot()

#set title, position of tick marks, and axis labels
plt.title('Daily COVID-19 Positive Rate')
plt.xticks(rotation=45)
plt.xlabel('')
plt.ylabel('Daily COVID-19 Positive Rate');
```



A country is not testing adequately when it is finding a case for every few tests they perform. Here it is likely that the true number of new cases is much higher than the number of cases that were confirmed by tests. The WHO has suggested a positive rate of between 3% and 10% as a general benchmark of adequate testing.

Q: Where is the number of reported daily cases today in relation to the peak of the pandemic?

[72]:
```
#compute summary statistics for number of new cases
covid_uk_df.new_cases.describe()
```

[72]:
```
count      268.000000
mean      1505.787313
std       1600.386726
min          0.000000
25%         55.750000
50%        912.000000
75%       2597.000000
max       5487.000000
Name: new_cases, dtype: float64
```

[73]:
```
#subset dataframe for days when number of reported cases exceeds 75th percentile
high_cases_df = covid_uk_df[covid_uk_df.new_cases > 2600]
high_cases_df
```

```
[73]:             new_cases  new_cases_smoothed  total_cases  new_tests  new_deaths  \
      date
      2020-03-27     2692.0            1755.857      17717.0        NaN       181.0
      2020-03-28     3087.0            2046.143      20804.0        NaN       288.0
      2020-03-29     3197.0            2323.571      24001.0        NaN       292.0
      2020-03-30     2822.0            2555.571      26823.0        NaN       212.0
      2020-03-31     2858.0            2767.000      29681.0        NaN       374.0
      …               …                 …          …        …         …
      2020-09-19     4322.0            3465.571     385936.0   252509.0        27.0
      2020-09-20     4422.0            3597.714     390358.0   239885.0        27.0
      2020-09-21     3899.0            3679.000     394257.0   219723.0        18.0
      2020-09-22     4368.0            3928.571     398625.0   188865.0        11.0
      2020-09-23     4926.0            4189.000     403551.0        NaN        37.0

                  new_deaths_smoothed  total_deaths  new_tests_smoothed  \
      date
      2020-03-27              103.143         884.0                 NaN
      2020-03-28              139.714        1172.0                 NaN
      2020-03-29              173.143        1464.0                 NaN
      2020-03-30              198.286        1676.0                 NaN
      2020-03-31              240.857        2050.0                 NaN
      …                         …           …                 …
      2020-09-19               16.857       41732.0            227647.0
      2020-09-20               19.429       41759.0            230321.0
      2020-09-21               21.286       41777.0            231257.0
      2020-09-22               21.571       41788.0            228564.0
      2020-09-23               23.000       41825.0                 NaN

                  total_tests  positive_rate  death_rate_t
      date
      2020-03-27          NaN            NaN      0.049896
      2020-03-28          NaN            NaN      0.056335
      2020-03-29          NaN            NaN      0.060997
      2020-03-30          NaN            NaN      0.062484
      2020-03-31          NaN            NaN      0.069068
      …                 …            …           …
      2020-09-19   18248877.0          0.015      0.108132
      2020-09-20   18488762.0          0.016      0.106976
      2020-09-21   18708484.0          0.016      0.105964
      2020-09-22   18897349.0          0.017      0.104830
      2020-09-23          NaN            NaN      0.103642

      [67 rows x 11 columns]
```

Observations:

The number of daily reported new cases has recently reached levels last witnessed during the height of the pandemic in early April.

Q: How many cases, deaths and tests were recorded for each day of the month?

[74]:
```
#return date index to columns
covid_uk_df.reset_index(inplace=True)
```

[75]:
```
#convert data column to datetime object
covid_uk_df['date'] = pd.to_datetime(covid_uk_df.date)
```

[76]:
```
#extract year, month, day, and weekend from date and create new column for each
covid_uk_df['year'] = pd.DatetimeIndex(covid_uk_df.date).year
covid_uk_df['month'] = pd.DatetimeIndex(covid_uk_df.date).month
covid_uk_df['day'] = pd.DatetimeIndex(covid_uk_df.date).day
covid_uk_df['weekday'] = pd.DatetimeIndex(covid_uk_df.date).weekday
covid_uk_df
```

[76]:
```
          date  new_cases  new_cases_smoothed  total_cases  new_tests  \
0   2019-12-31        0.0                 NaN          0.0        NaN
1   2020-01-01        0.0                 NaN          0.0        NaN
2   2020-01-02        0.0                 NaN          0.0        NaN
3   2020-01-03        0.0                 NaN          0.0        NaN
4   2020-01-04        0.0                 NaN          0.0        NaN
..         …          …                   …            …          …
263 2020-09-19     4322.0            3465.571     385936.0   252509.0
264 2020-09-20     4422.0            3597.714     390358.0   239885.0
265 2020-09-21     3899.0            3679.000     394257.0   219723.0
266 2020-09-22     4368.0            3928.571     398625.0   188865.0
267 2020-09-23     4926.0            4189.000     403551.0        NaN

     new_deaths  new_deaths_smoothed  total_deaths  new_tests_smoothed  \
0           0.0                  NaN           0.0                 NaN
1           0.0                  NaN           0.0                 NaN
2           0.0                  NaN           0.0                 NaN
3           0.0                  NaN           0.0                 NaN
4           0.0                  NaN           0.0                 NaN
..          …                    …             …                   …
263        27.0               16.857       41732.0            227647.0
264        27.0               19.429       41759.0            230321.0
265        18.0               21.286       41777.0            231257.0
266        11.0               21.571       41788.0            228564.0
267        37.0               23.000       41825.0                 NaN

     total_tests  positive_rate  death_rate_t  year  month  day  weekday
0            NaN            NaN           NaN  2019     12   31        1
1            NaN            NaN           NaN  2020      1    1        2
2            NaN            NaN           NaN  2020      1    2        3
3            NaN            NaN           NaN  2020      1    3        4
4            NaN            NaN           NaN  2020      1    4        5
```

```
..           …            …           …   …    …   …      …
263   18248877.0           0.015      0.108132  2020      9   19        5
264   18488762.0           0.016      0.106976  2020      9   20        6
265   18708484.0           0.016      0.105964  2020      9   21        0
266   18897349.0           0.017      0.104830  2020      9   22        1
267          NaN             NaN      0.103642  2020      9   23        2

[268 rows x 16 columns]
```

[109]:

```
#exclude incomplete months, i.e September
covid_uk_exsep = covid_uk_df[covid_uk_df.month != 9].copy()
covid_uk_exsep
```

[109]:

```
          date  new_cases  new_cases_smoothed  total_cases  new_tests  \
0   2019-12-31        0.0                 NaN          0.0        NaN
1   2020-01-01        0.0                 NaN          0.0        NaN
2   2020-01-02        0.0                 NaN          0.0        NaN
3   2020-01-03        0.0                 NaN          0.0        NaN
4   2020-01-04        0.0                 NaN          0.0        NaN
..         …        …                 …          …        …
240 2020-08-27     1048.0            1106.857     328846.0   184461.0
241 2020-08-28     1522.0            1155.429     330368.0   178203.0
242 2020-08-29     1276.0            1190.143     331644.0   168684.0
243 2020-08-30     1108.0            1164.429     332752.0   170574.0
244 2020-08-31     1715.0            1260.714     334467.0   166871.0

     new_deaths  new_deaths_smoothed  total_deaths  new_tests_smoothed  \
0           0.0                  NaN           0.0                 NaN
1           0.0                  NaN           0.0                 NaN
2           0.0                  NaN           0.0                 NaN
3           0.0                  NaN           0.0                 NaN
4           0.0                  NaN           0.0                 NaN
..          …                  …           …                  …
240        16.0                9.714       41465.0            169546.0
241        12.0               10.571       41477.0            172228.0
242         9.0               11.571       41486.0            170658.0
243        12.0               10.714       41498.0            170542.0
244         1.0               10.000       41499.0            172026.0

     total_tests   positive_rate  death_rate_t  year  month  day  weekday
0            NaN             NaN           NaN  2019     12   31        1
1            NaN             NaN           NaN  2020      1    1        2
2            NaN             NaN           NaN  2020      1    2        3
3            NaN             NaN           NaN  2020      1    3        4
4            NaN             NaN           NaN  2020      1    4        5
..           …            …           …   …    …   …      …
240   13633416.0           0.007      0.126092  2020      8   27        3
```

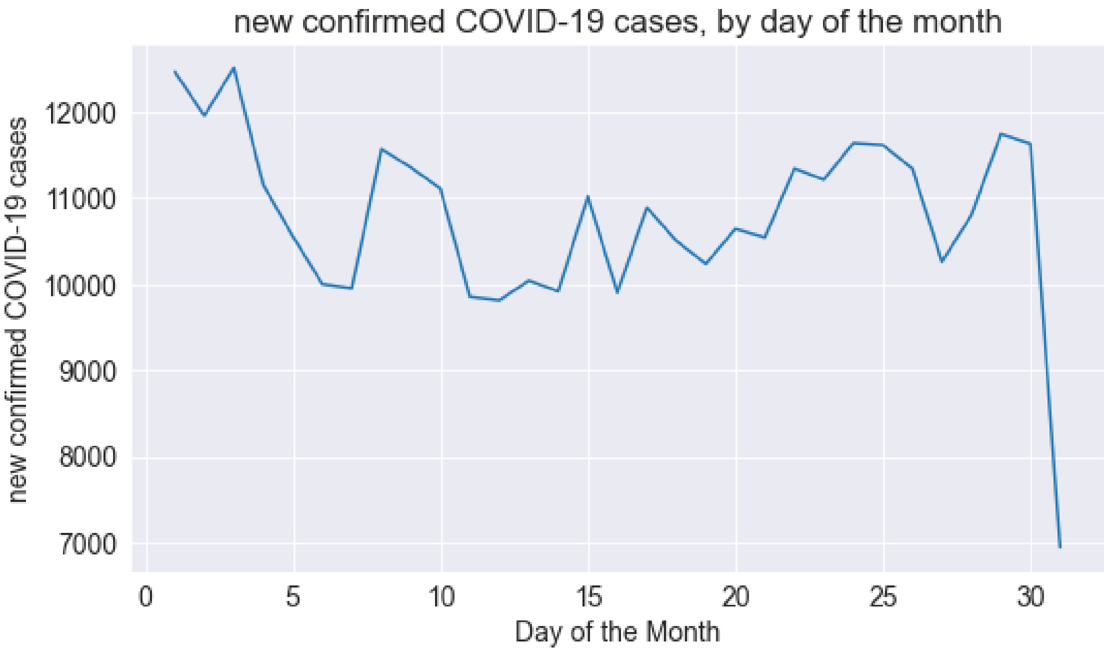
```
241  13823629.0         0.007     0.125548  2020       8   28         4
242  13992972.0         0.007     0.125092  2020       8   29         5
243  14163546.0         0.007     0.124711  2020       8   30         6
244  14330417.0         0.007     0.124075  2020       8   31         0

[245 rows x 16 columns]
```

[110]:

```
#sum cases, deaths and tests by day of the month
covid_uk_exsep = covid_uk_exsep.groupby('day')[['new_cases', 'new_deaths',␣
 ↪'new_tests']].sum()
```

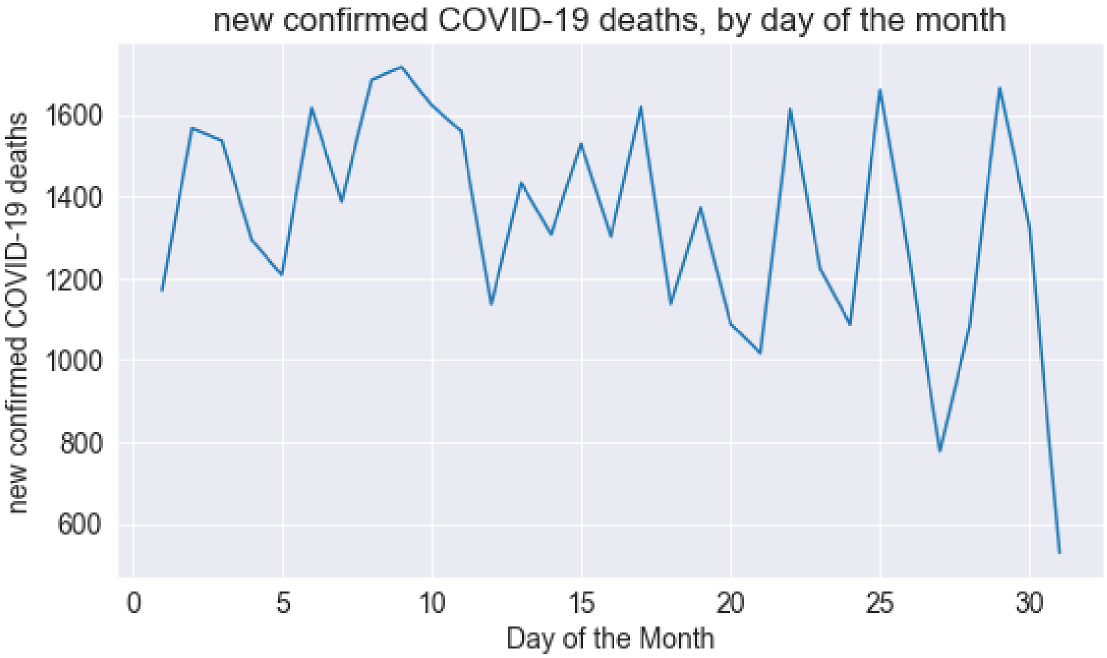
[129]:

```
#plot line chart of new cases by day of the month
covid_uk_exsep.new_cases.plot()

#set title, position of tick marks, and axis labels
plt.title('new confirmed COVID-19 cases, by day of the month')
plt.xlabel('Day of the Month')
plt.ylabel('new confirmed COVID-19 cases');
```



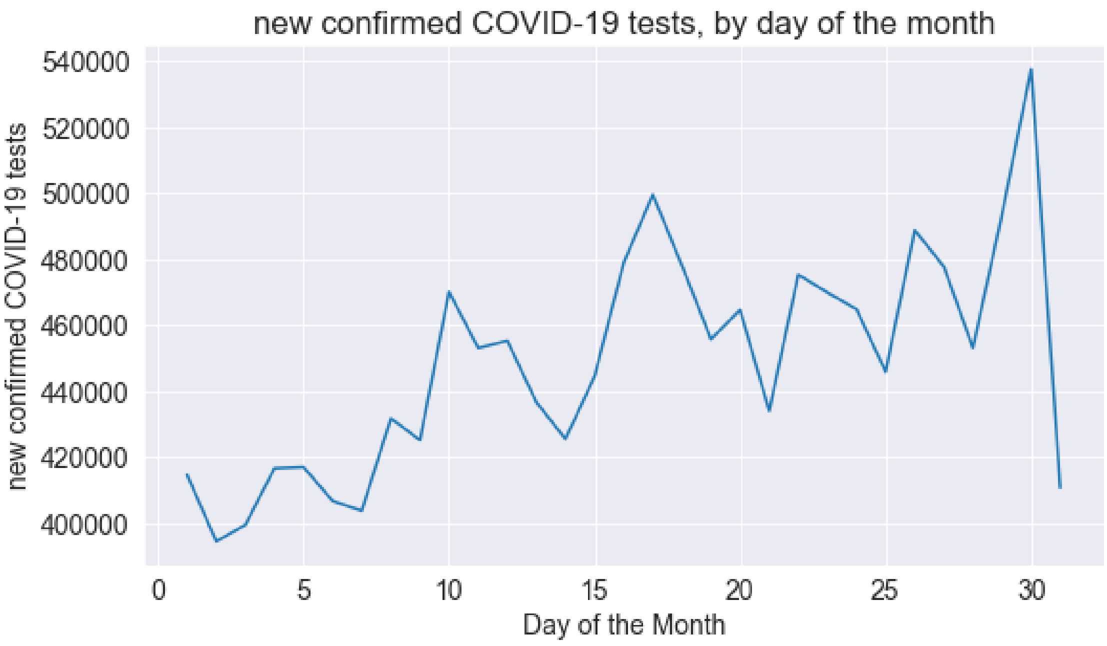
[130]:

```
#plot line chart of new cases by day of the month
covid_uk_exsep.new_deaths.plot()

#set title, position of tick marks, and axis labels
plt.title('new confirmed COVID-19 deaths, by day of the month')
plt.xlabel('Day of the Month')
plt.ylabel('new confirmed COVID-19 deaths');
```



[131]:

```
#plot line chart of new cases by day of the month
covid_uk_exsep.new_tests.plot()

#set title, position of tick marks, and axis labels
plt.title('new confirmed COVID-19 tests, by day of the month')
plt.xlabel('Day of the Month')
plt.ylabel('new confirmed COVID-19 tests');
```



Observations:

Variation in the number of deaths attributed to Covid-19 increases in the last 10 days of each month. Perhaps, this is linked to the increased number of tests conducted during during the same period. Whether these findings are statistically and/or practically significant would require further investigation.

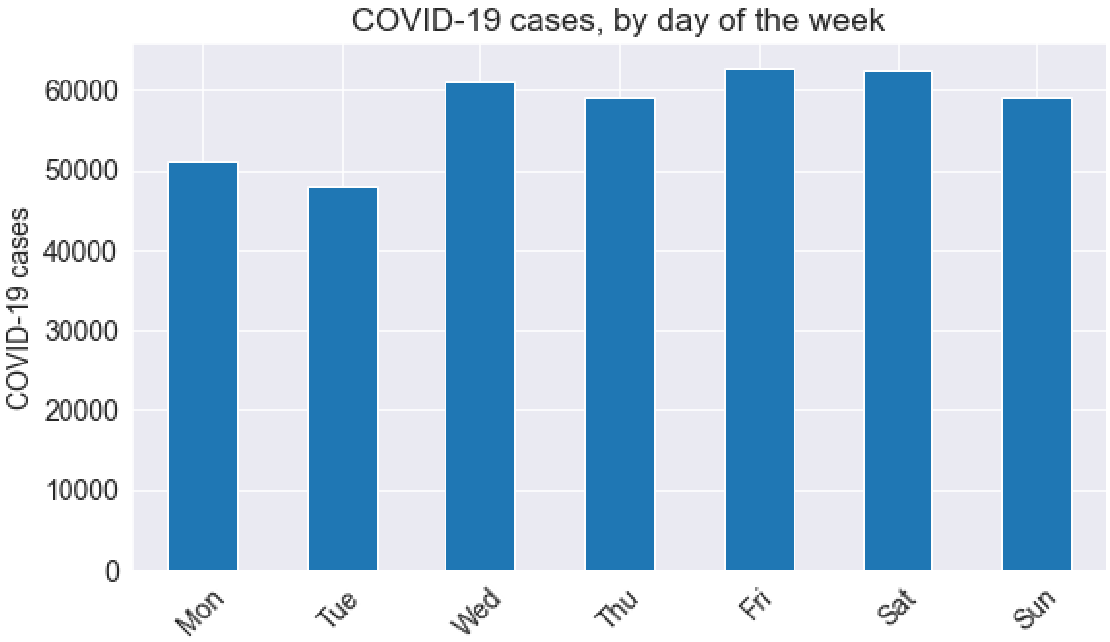
Q: How many cases, deaths and tests were recorded for each day of the week?

```
[120]: # sum cases, deaths and tests by day of the month (monday is 0)
       covid_weekday_df = covid_uk_df.groupby('weekday')[['new_cases', 'new_deaths',␣
        ↪'new_tests']].sum()
       covid_weekday_df
```

```
[120]:          new_cases  new_deaths  new_tests
       weekday
       0          51237.0      3527.0  2430179.0
       1          47857.0      3645.0  2245991.0
       2          60998.0      7930.0  2457385.0
       3          59012.0      7217.0  2749636.0
       4          62796.0      6359.0  2864353.0
       5          62619.0      7083.0  2912626.0
       6          59032.0      6064.0  2742976.0
```

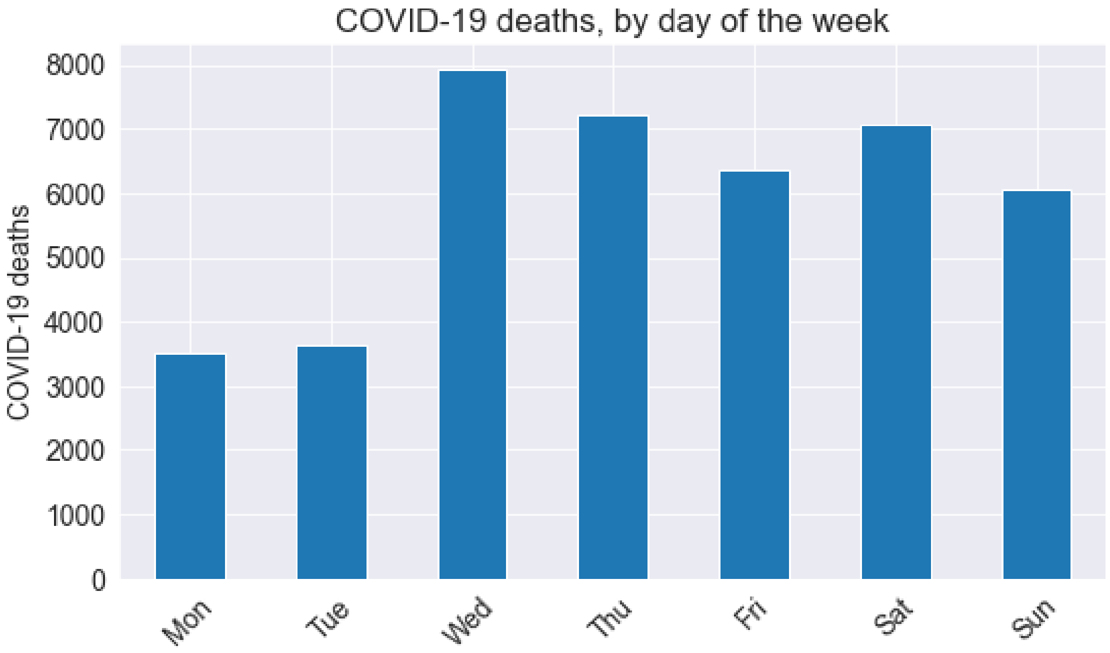
```
[137]: #plot line chart of new cases by day of the month
       covid_weekday_df.new_cases.plot(kind='bar')
```

```
#set title, position of tick marks, and axis labels
plt.title('COVID-19 cases, by day of the week')
day = ['Mon', 'Tue', 'Wed', 'Thu', 'Fri', 'Sat', 'Sun']
plt.xticks(np.arange(0, 7), day, rotation=45)
plt.xlabel('')
plt.ylabel('COVID-19 cases');
```



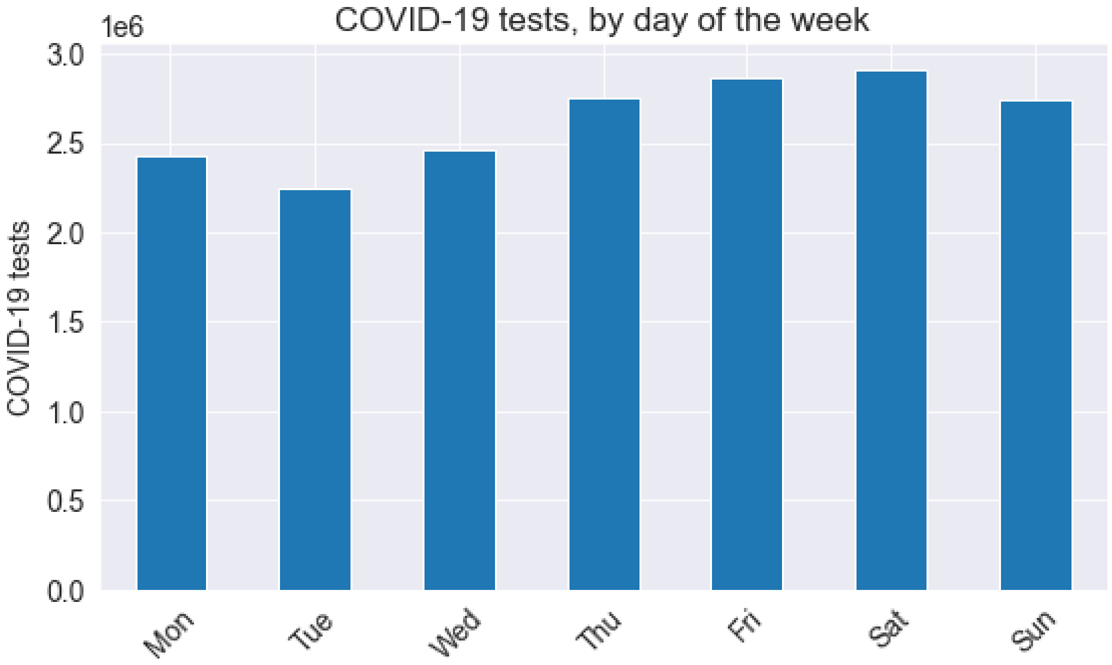
[138]:

```
#plot line chart of new cases by day of the month
covid_weekday_df.new_deaths.plot(kind='bar')

#set title, position of tick marks, and axis labels
plt.title('COVID-19 deaths, by day of the week')
day = ['Mon', 'Tue', 'Wed', 'Thu', 'Fri', 'Sat', 'Sun']
plt.xticks(np.arange(0, 7), day, rotation=45)
plt.xlabel('')
plt.ylabel('COVID-19 deaths');
```



[139]:

```
#plot line chart of new cases by day of the month
covid_weekday_df.new_tests.plot(kind='bar')

#set title, position of tick marks, and axis labels
plt.title('COVID-19 tests, by day of the week')
day = ['Mon', 'Tue', 'Wed', 'Thu', 'Fri', 'Sat', 'Sun']
plt.xticks(np.arange(0, 7), day, rotation=45)
plt.xlabel('')
plt.ylabel('COVID-19 tests');
```



Observations:

The number of deaths attributed to Covid-19 reach a lull on Monday & Tuesday. This may be due to beauracratic idiosyncracies rather than an accurate model of reality.

```
[140]: #save output to csv file
       covid_uk_df.to_csv('results.csv', index=False)
```

0.7 Conclusion

0.7.1 Summary

1. The number of reported cases peaked at approx 4,900 on the 10th of April 2020. Since the 18th of July the number of daily reported cases has once again begun to grow. Is the UK prepared for a second wave? The number of daily reported new cases has recently reached levels last witnessed during the height of the pandemic in early April.
2. Given the rise in the number of reported daily cases, the number of daily confirmed deaths is expected to follow.
3. As capacity is built the number of daily tests continues to rise.

0.7.2 Limitations:

What is important to note about these case figures? - The reported case figures on a given date does not necessarily show the number of new cases on that day: this is due to delays in reporting. - Keep in mind these are offically reported numbers, and the actual number of cases and deaths may be higher, as not all cases are diagnosed. - The actual number of cases is also likely to be much higher than the number of confirmed cases – this is due to limited testing. - Comorbidiy. Covid-19 may be a contributing factor but perhaps not the only cause of death.

0.7.3 Directions for Further Research

1. Statistical & Practical significance of day of the month/week differences