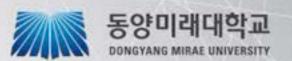
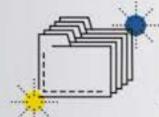


시계열 데이터

૭ 코로나 바이러스 시계열 데이터 분석



시계열 데이터

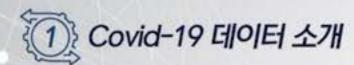


학습개요

- 1/ Covid-19 데이터 소개
- 2/ Covid-19 시계열 데이터 처리 및 분석







Covid-19 데이터 소개





① Covid-19 데이터 소개

✓ 2019년 12월, 중국 우한에서 발병

✓ 치사율 : 낮음

✓ 전염력 : 매우 강함







(1) Covid-19 데이터 소개

フ코로나 바이러스 (Covid-19) 데이터 셋

√ https://github.com/owid/covid-19-data/tree/master/public/data

Our World In Data https://ourworldindata.org

- ✓ 전세계에서 공지되는 데이터를 수집하여 매일매일 업데이트
- ✓ 세상에 산적해 있는 거대한 문제들을 해결하고 혁신하기 위한 연구를 진행하고 관련 데이터들을 제공
- ✓ 현재 약 300개의 주제에 3,000개 이상의 데이터 셋을 제공
- ✓ 누구나 접근 가능하며, 누구나 사용 가능
- ✓ 전세계 연구 기관, 언론 매체, 주요 대학 등에서 후원 받아 사용되고 있음





(1) Covid-19 데이터 소개

フ코로나 바이러스 (Covid-19) 데이터

- √ https://github.com/owid/covid-19-data/tree/master/public/data
- ✓ 전세계 200개 이상 국가의 코로나 현황(총 59개 컬럼)
 - ◎ 누적, 신규 확진자 수
 - ◎ 누적, 신규 사망자 수
 - ◎ 누적, 신규 검사자 수
 - ⊙ 나라별 인구수
 - ⊙ 인구 천명당 확진자 수 등
- ✓ 2020년 1월 1일부터 데이터 제공
- ✓ 2021년 2월 23일 기준, 약 8만 건의 데이터 제공



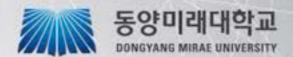
02 Covid-19 데이터 기반 시계열데이터처리및분석 **X** O





② Covid-19 데이터 기반 시계열 데이터 처리 및 분석

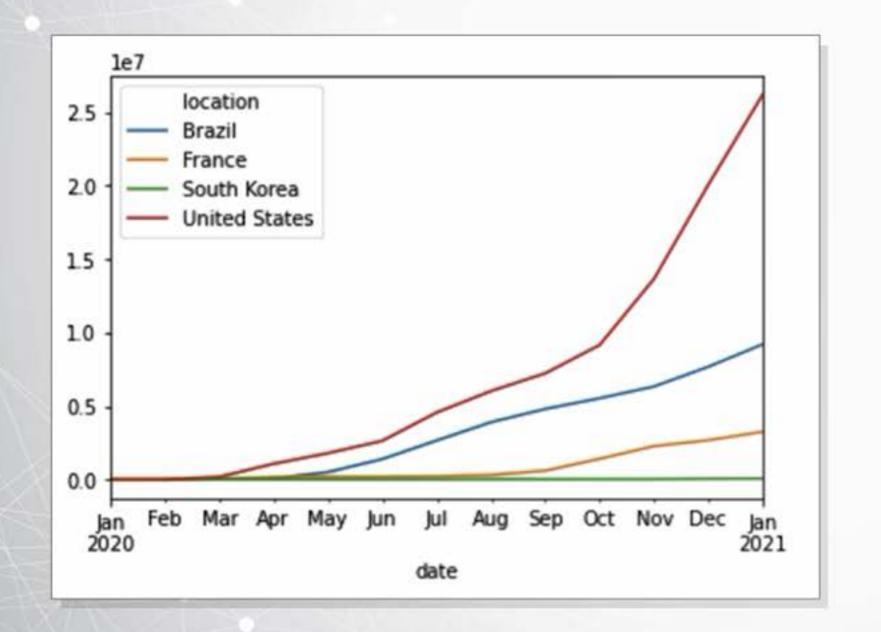
Covid-19 시계열 데이터 처리및분석

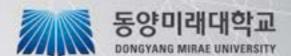




② Covid-19 데이터 기반 시계열 데이터 처리 및 분석

' 구요 나라들의 월별 누적 확진자 수 추이

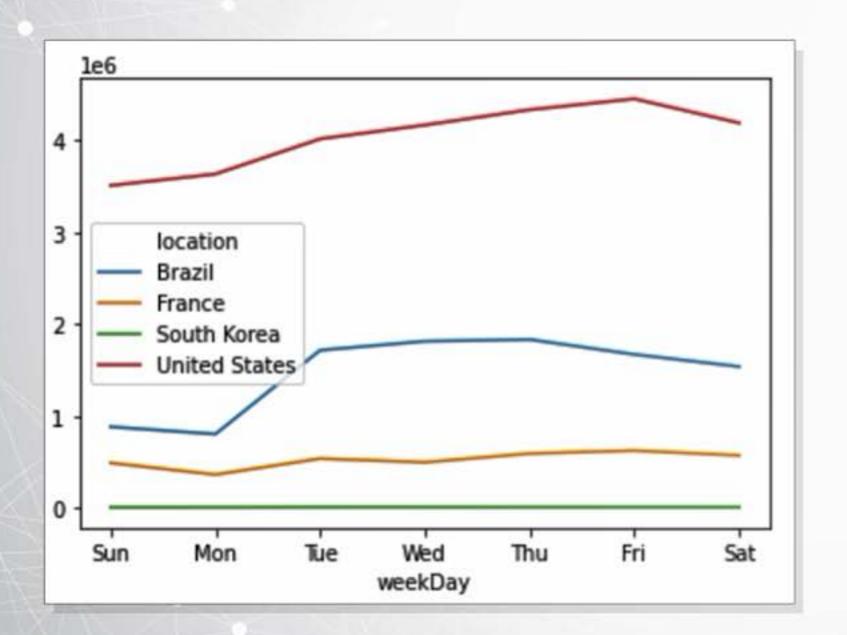


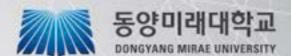




② Covid-19 데이터 기반 시계열 데이터 처리 및 분석

' 구요 나라들의 요일별 누적 확진자 수 추이

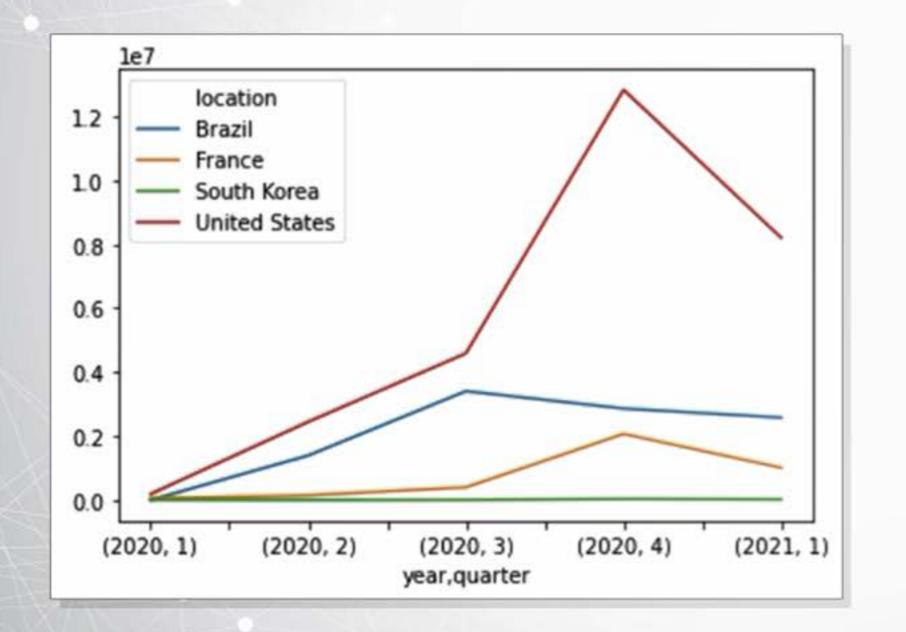






Covid-19 데이터 기반 시계열 데이터 처리 및 분석

· 주요 나라들의 분기별 누적 확진자 수 추이



◎ Covid-19 데이터 가져오기



https://github.com/owid/covid-19-data/tree/master/public/data

Owid-covid-data.xisx

Automated JHU update

5 minutes ago

README.md

Data on COVID-19 (coronavirus) by Our World in Data

Our complete COVID-19 dataset is a collection of the COVID-19 data maintained by Our World in Data. It is updated daily and includes data on confirmed cases, deaths, hospitalizations, testing, and vaccinations as well as other variables of potential interest.

Download our complete COVID-19 dataset : CSV | XLSX | JSON

We will continue to publish up-to-date data on confirmed cases, deaths, hospitalizations, testing, and vaccinations, throughout the duration of the COVID-19 pandemic.

The data you find here and our data sources

- Confirmed cases and deaths: our data comes from the COVID-19 Data Repository by the Center for Systems Science and Engineering
 (CSSE) at Johns Hopkins University (JHU). We discuss how and when JHU collects and publishes this data here. The cases & deaths dataset
 is updated daily. Note: the number of cases or deaths reported by any institution—including JHU, the WHO, the ECDC and others—on a given
 day does not necessarily represent the actual number on that date. This is because of the long reporting chain that exists between a new
 case/death and its inclusion in statistics. This also means that negative values in cases and deaths can sometimes appear when a country
 corrects historical data, because it had previously overestimated the number of cases/deaths. Alternatively, large changes can
 sometimes (although rarely) be made to a country's entire time series if JHU decides (and has access to the necessary data) to correct
 values retrospectively.
- Hospitalizations and intensive care unit (ICU) admissions: our data comes from the European Centre for Disease Prevention and Control
 (ECDC) for a select number of European countries; the government of the United Kingdom; the COVID Tracking Project for the United
 States; the COVID-19 Tracker for Canada. Unfortunately, we are unable to provide data on hospitalizations for other countries: there is
 currently no global, aggregated database on COVID-19 hospitalization, and our team at Our World in Data does not have the capacity to
 build such a dataset.
- . Testing for COVID-19: this data is collected by the Our World in Data team from official reports; you can find further details in our post on

https://covid.ourworldindata.org/data/owid-covid-dataxisx

◎ Covid-19 데이터 가져오기

In [227]:



- 2. 시계열 데이터 응용 Covid-19 현황 분석
- 1. Covid-19 데이터 가져오기
 - https://github.com/owid/covid-19-data/tree/master/public/data : 매일 업데이트된 파일을 제공함

```
In []: covid = pd.read_excel('data/owid-covid-data.xlsx')
In []: 2. date 컬럼을 datetime으로 변경하기
```

◎ Covid-19 데이터 가져오기



2. 시계열 데이터 응용 - Covid-19 현황 분석

- 1. Covid-19 데이터 가져오기
 - https://github.com/owid/covid-19-data/tree/master/public/data : 매일 업데이트된 파일을 제공함

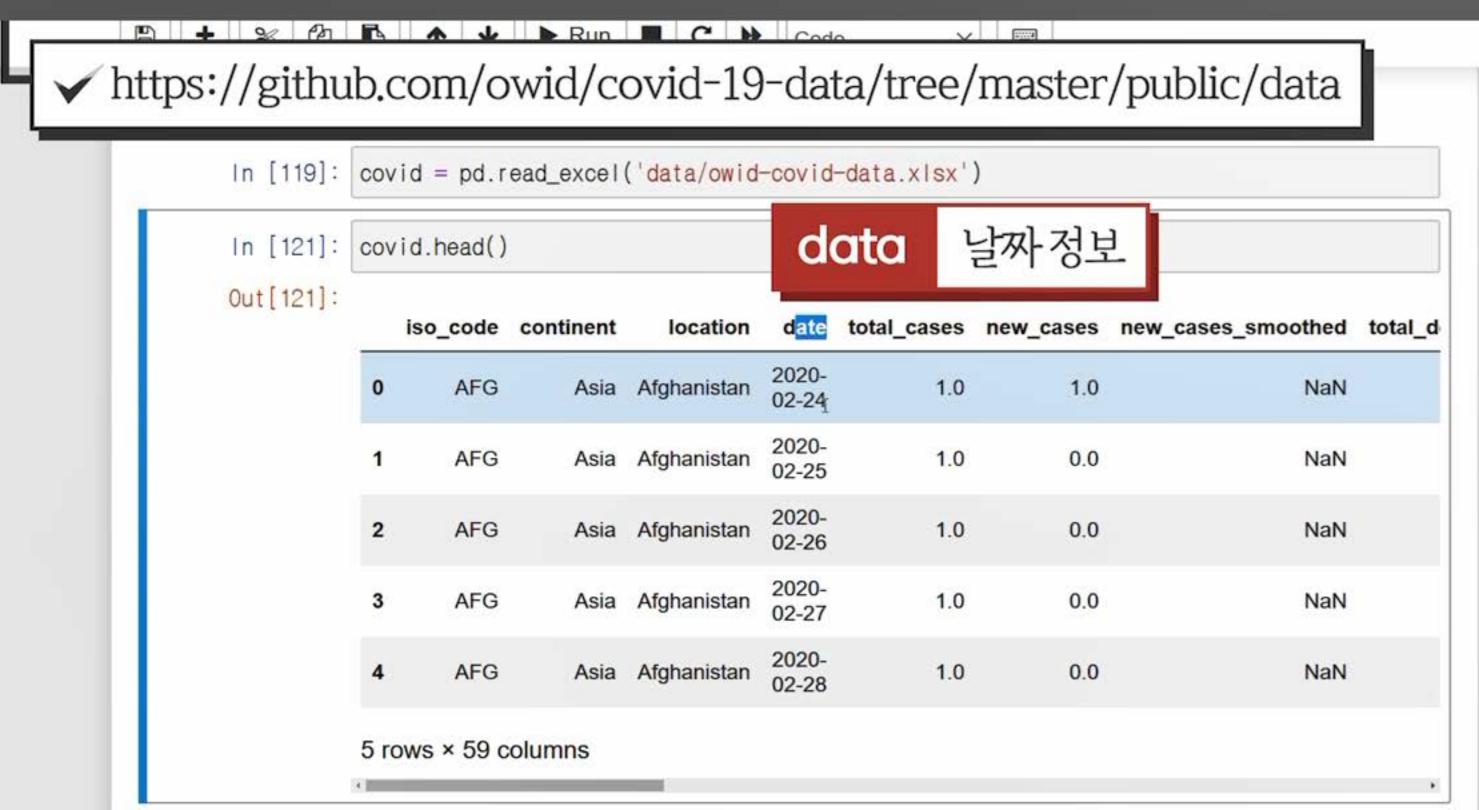
```
In [*]: covid = pd.read_excel('data/owid-covid-data.xlsx')

In [*]: covid.head()
```

2 data 커러은 datatima 0 2 변겨치기

실행 번호가 표시되는 곳에 [*]가 표시되면 '실행 중'이라는 의미입니다.

◎ Covid-19 데이터 가져오기



◎ Covid-19 데이터 가져오기

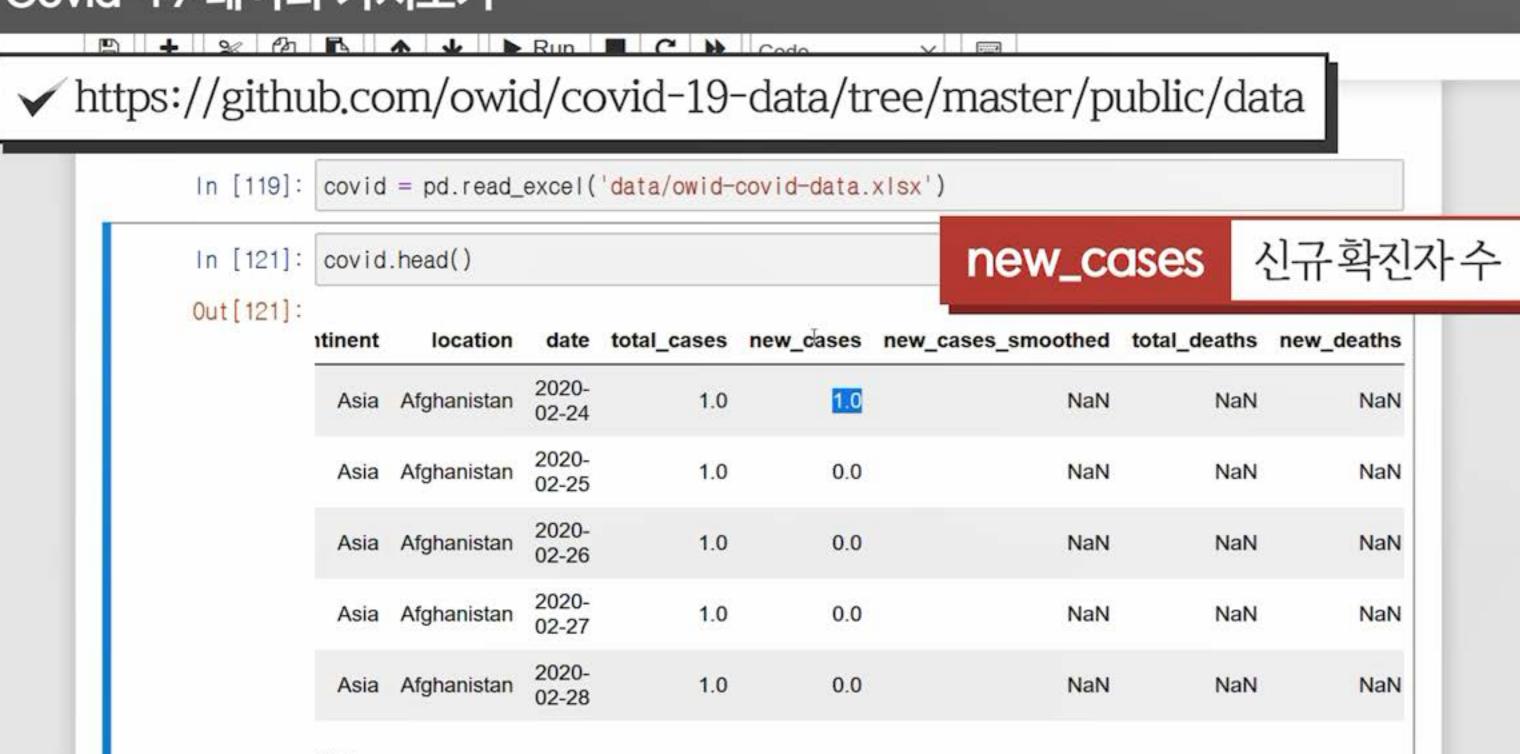


https://github.com/owid/covid-19-data/tree/master/public/data

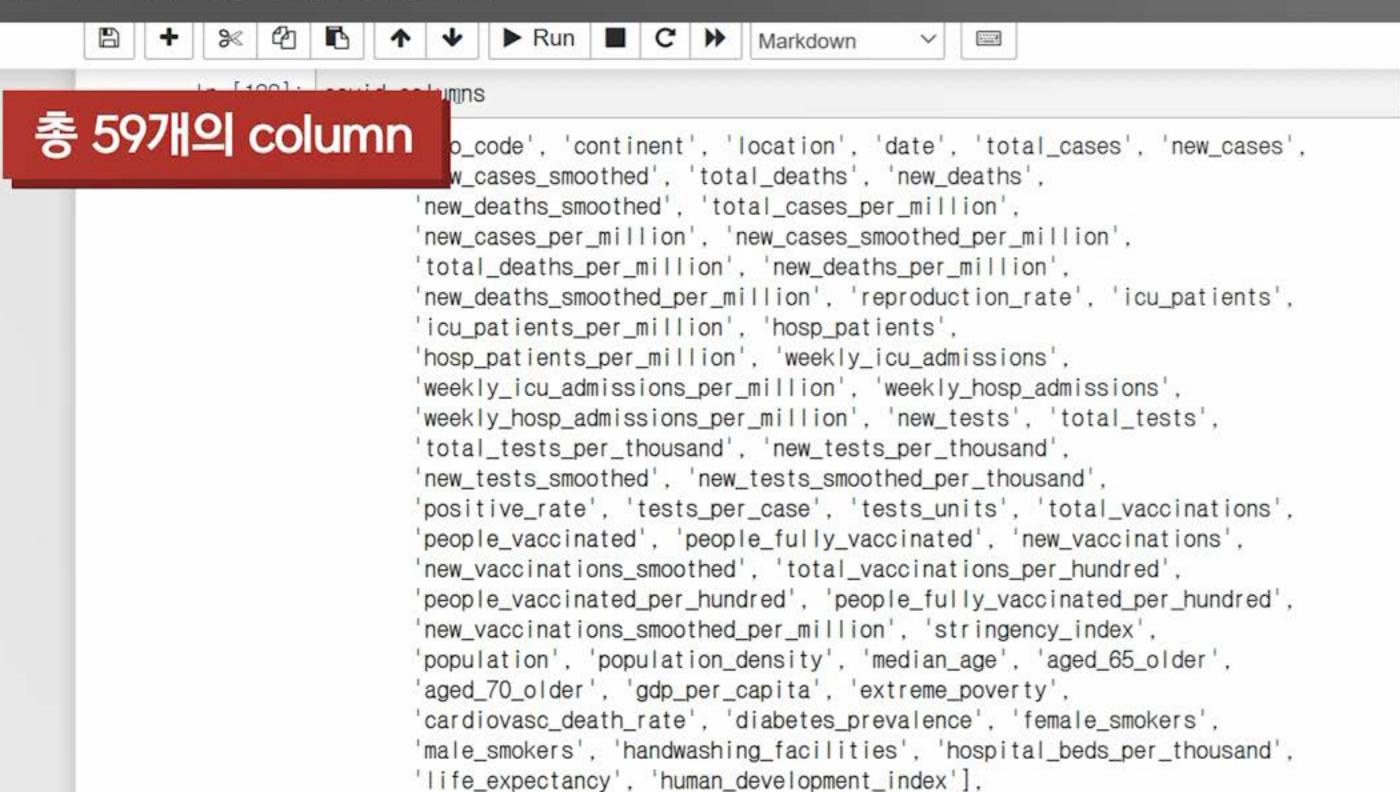


nns

◎ Covid-19 데이터 가져오기

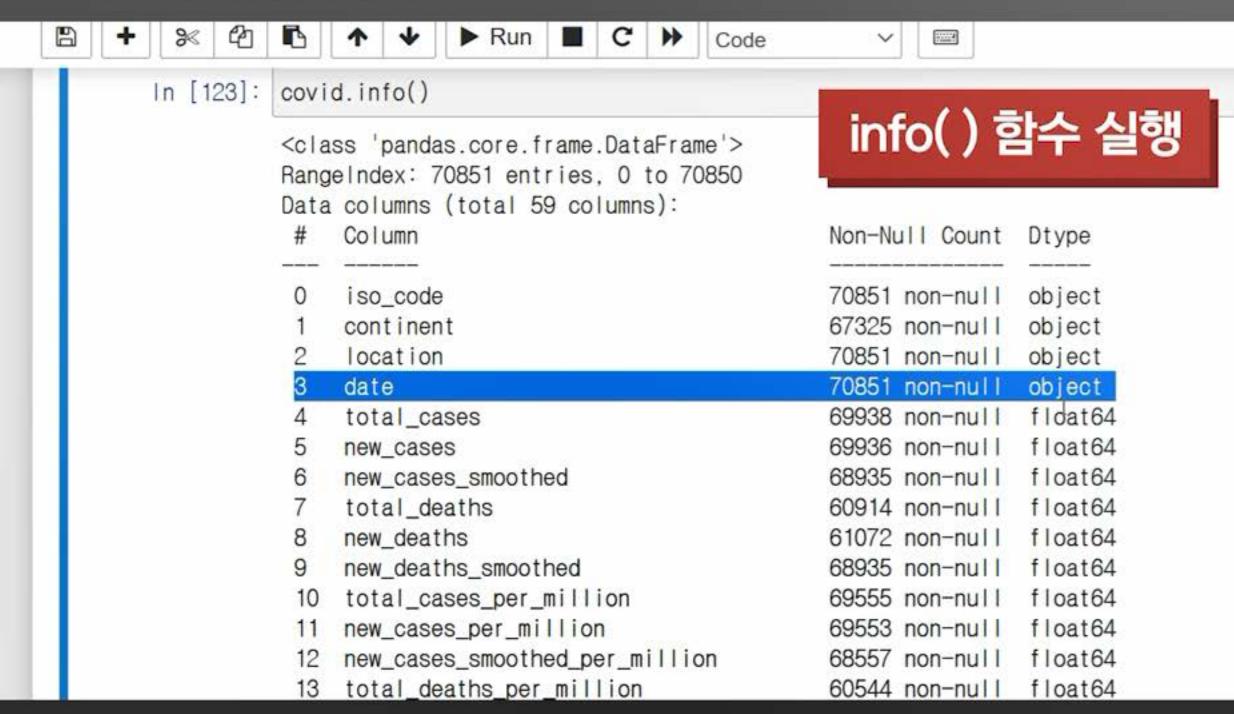


◎ Covid-19 데이터 가져오기



dtype- object)

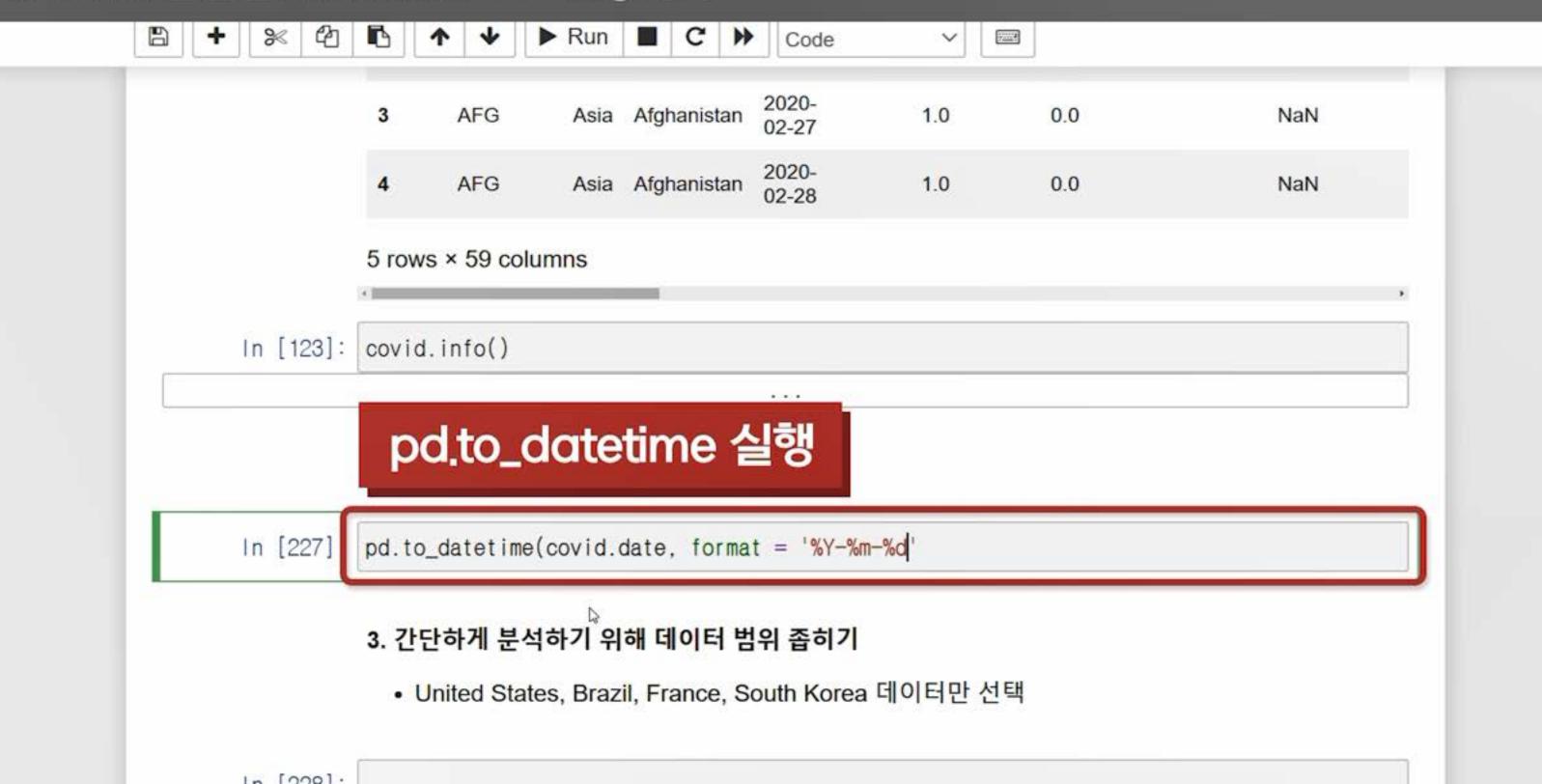
◎ Covid-19 데이터 가져오기



가장 먼저 date 컬럼의 데이터타입이 datetime인지 확인해야 합니다.

17 icu patients 55/38 non-null float64

◎ date 컬럼을 datetime으로 변경하기



◎ 간단한 분석을 위해 데이터 범위 좁히기



- 3. 간단하게 분석하기 위해 데이터 범위 좁히기
 - United States, Brazil, France, South Korea 데이터만 선택



◎ 간단한 분석을 위해 데이터 범위 좁히기



```
In [137]: covid = covid[covid.location.isin(['United States', 'Brazil', 'France', 'South Korea'])
```

• location, date, total cases, new cases 컬럼만 선택

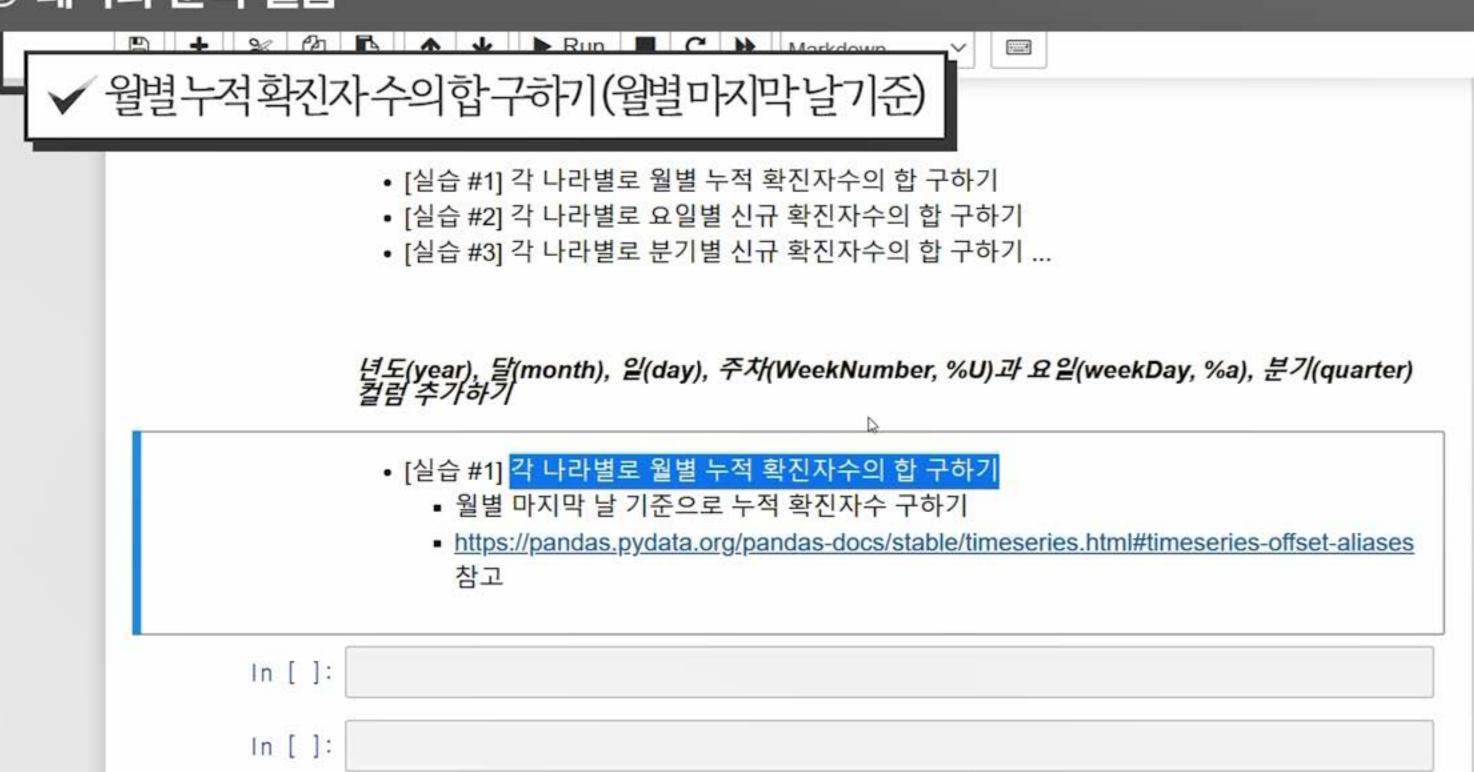
```
In [ ]: covid[[location', 'date', 'total_cases', 'new_cases']]
```

4. 데이터 분석 실습

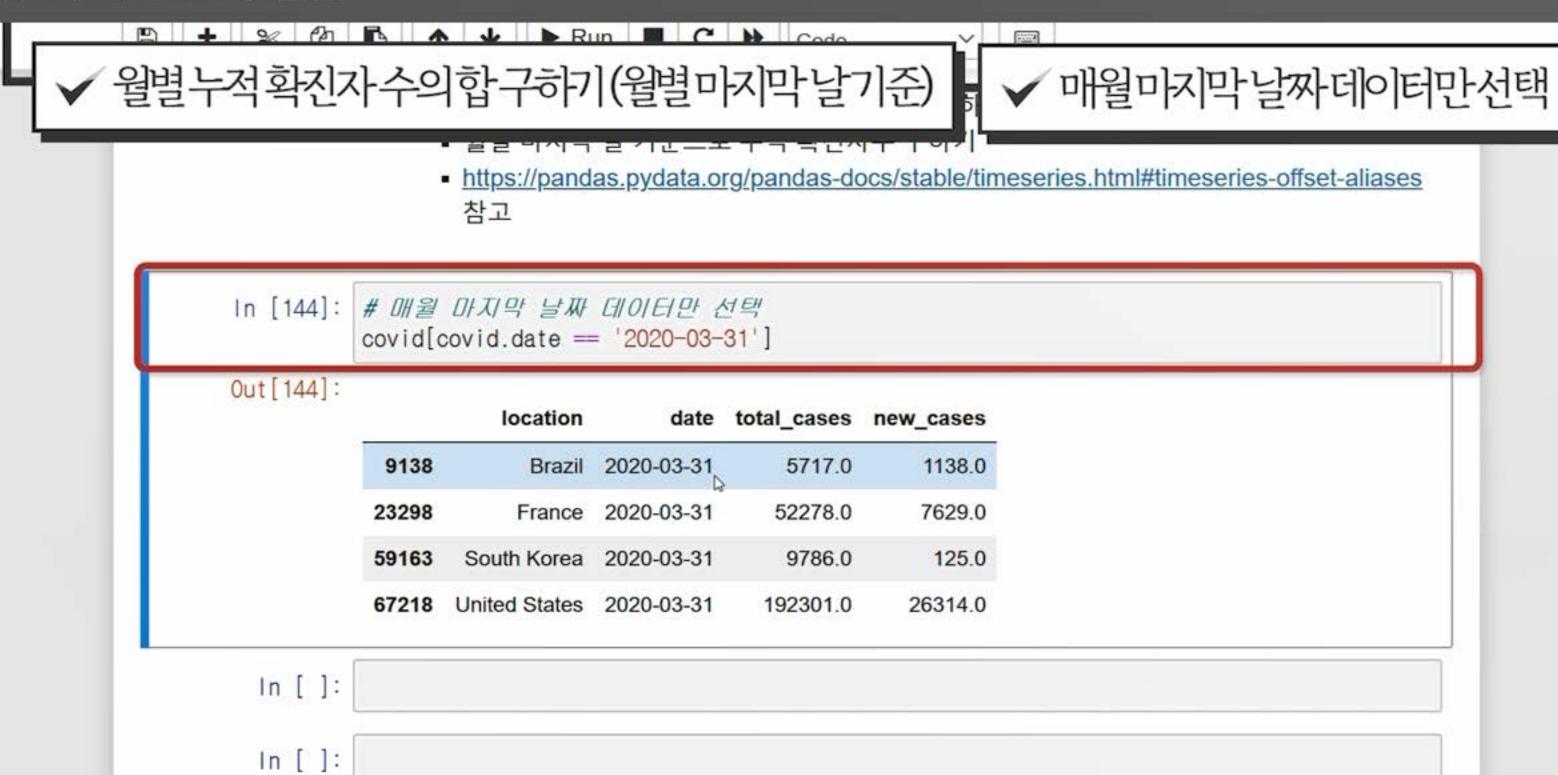
• [실습 #1] 각 나라별로 월별 누적 확진자수의 합 구하기

In [] .

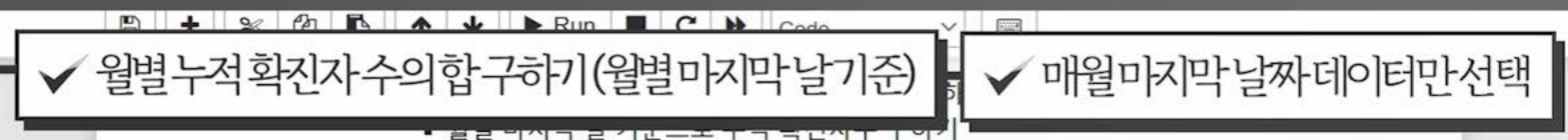
◎ 데이터 분석 실습



◎ 데이터 분석 실습



◎ 데이터 분석 실습



https://pandas.pydata.org/pandas-docs/stable/timeseries.html#timeseries-offset-aliases
 참고

```
In [145]: # 매월 마지막 날짜 데이터만 선택 covid[covid.date = '2020-05-31']

Out [145]:

| location | location
```

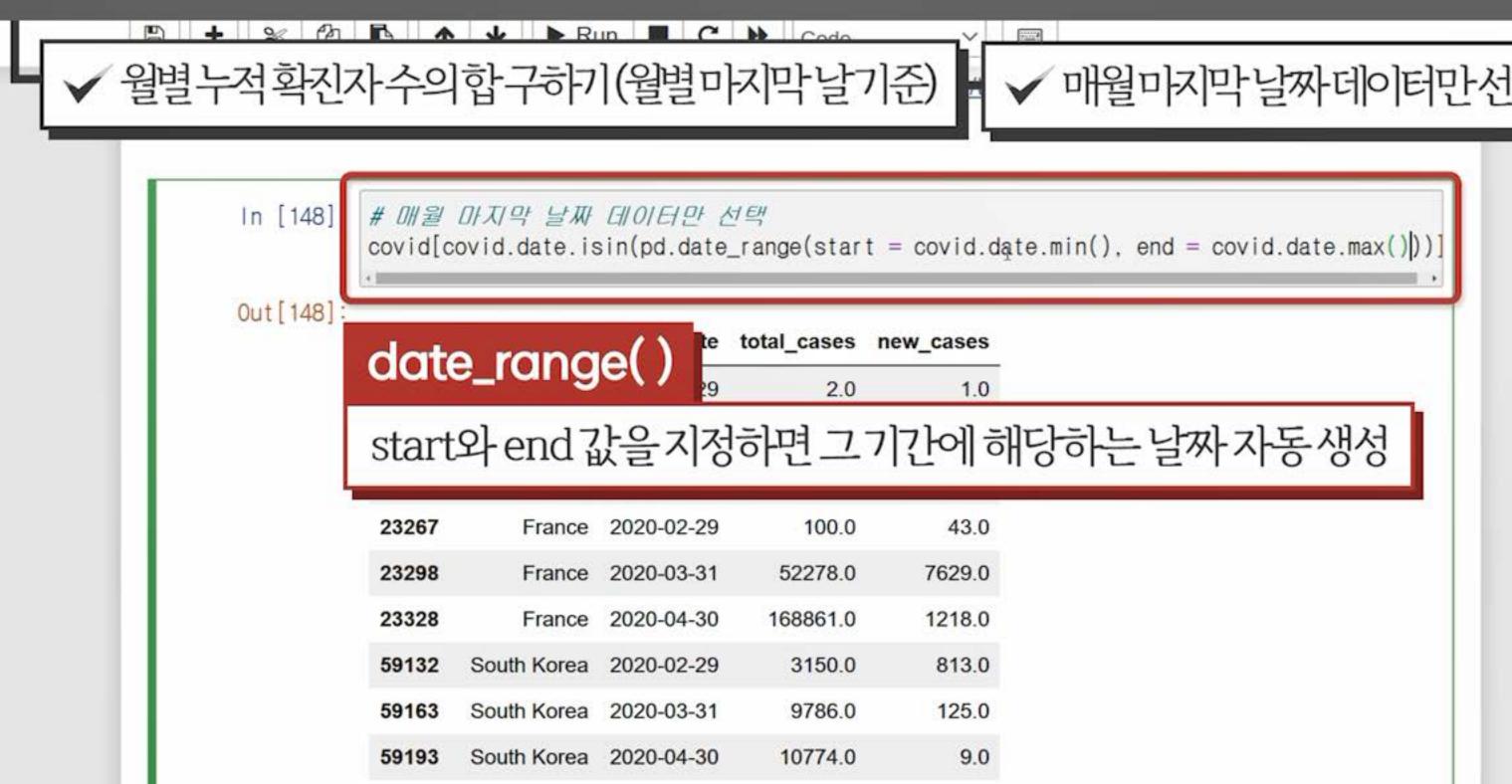
	location	uate	total_cases	new_cases
9199	Brazil	2020-05-31	514849.0	16409.0
23359	France	2020-05-31	190975.0	231.0
59224	South Korea	2020-05-31	11503.0	35.0
67279	United States	2020-05-31	1798713.0	18987.0

In []:

◎ 데이터 분석 실습



◎ 데이터 분석 실습

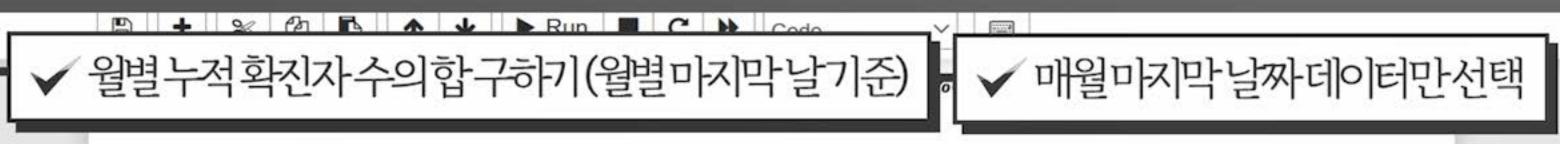


25.0

20

67197 United States 2020-02-29

◎ 데이터 분석 실습



- [실습 #1] 각 나라별로 월별 누적 확진자수의 합 구하기
 - 월별 마지막 날 기준으로 누적 확진자수 구하기
 - https://pandas.pydata.org/pandas-docs/stable/timeseries.html#timeseries-offset-aliases
 참고

freq

In [148]: # 배월 마지막 날짜 데이터만 선택
covid[covid.date.isin(pd.date_range(start = covid.date.min(),
end = covid.date.max(),
freq =))]

Out[148]:

	location	date	total_cases	new_cases
9107	Brazil	2020-02-29	2.0	1.0
9138	Brazil	2020-03-31	5717.0	1138.0
9168	Brazil	2020-04-30	87187.0	7502.0
23267	France	2020-02-29	100.0	43.0
23298	France	2020-03-31	52278.0	7629.0

◎ 데이터 분석 실습

In [145]: ts + pd.Timedelta(days=1)



월별누적확진자수의합구하기(월별마지막날기준)



매월마지막날짜데이터만선틱

minoto uara structures

Essential basic functionality

IO tools (text, CSV, HDF5, ...)

Indexing and selecting data

MultiIndex / advanced indexing

Merge, join, concatenate and compare

Reshaping and pivot tables

Working with text data

Working with missing data

Duplicate Labels

Categorical data

Nullable integer data type

Nullable Boolean data type

Visualization

Computational tools

Group by: split-apply-combine

Windowing Operations

Time series / date functionality

Time deltas

Styling

Options and settings

Enhancing performance

```
In [148]: friday.day_name()
Out[148]: 'Friday'

# Add 2 business days (Friday --> Tuesday)
In [149]: two_business_days = 2 * pd.offsets.8Day()

In [150]: two_business_days.apply(friday)
Out[150]: Timestamp('2018-01-09 00:00:00')

In [151]: friday + two_business_days
Out[151]: Timestamp('2018-01-09 00:00:00')

In [152]: (friday + two_business_days).day_name()
Out[152]: 'Tuesday'
```

Most pateoffsets have associated frequencies strings, or offset aliases, that can be passed into freq keyword arguments. The available date offsets and associated frequency strings can be found below:

	Frequency		
Date Offset	String	Description	
DateOffset	None	Generic offset class, defaults to absolute 24 hours	
BDay Of BusinessDay	'в'	business day (weekday)	
CDay OF CustomBusinessDay	,ε,	custom business day	
ieek	'м'	one week, optionally anchored on a day of the week	
weekOfMonth	'WOH'	the x-th day of the y-th week of each month	

Converting to timestamps

Generating ranges of

timestam

Timestamp limitations

Indexing

Time/date components

DateOffset objects

Parametric offsets

Using offsets with Series /

DatetimeIndex

Custom business days

Business hour

Custom business hour

Offset aliases

Combining aliases

Anchored offsets

Anchored offset semantics

Holidays / holiday calendars

Time series-related instance

Resampling

methods

Time span representation

Converting between

representations

Representing out-of-bounds

spans

Time zone handling

offset aliases라고 검색하면 freq 인자에 대한 정보를 확인할 수 있습니다.

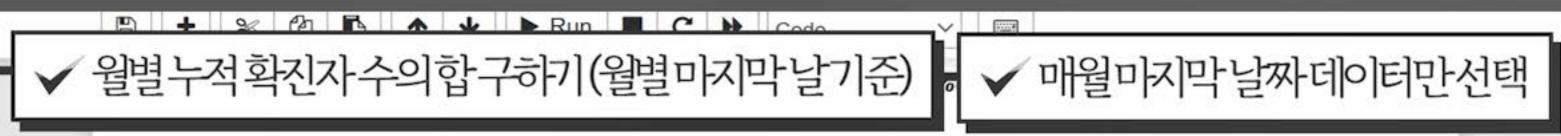
Cookbook

MonthRe

140

calendar month begin

◎ 데이터 분석 실습



- [실습 #1] 각 나라별로 월별 누적 확진자수의 합 구하기
 - 월별 마지막 날 기준으로 누적 확진자수 구하기
 - https://pandas.pydata.org/pandas-docs/stable/timeseries.html#timeseries-offset-aliases 참고

```
In [148]:
         # 매월 마지막 날짜 데이터만 선택
         covid[covid.date.isin(pd.date_range(start = covid.date.min(),
                                           end = covid.date.max(),
                                          freq = '0'))]
Out[148]:
```

11-	location	date	total_cases	new_cases
9107	Brazil	2020-02-29	2.0	1.0
9138	Brazil	2020-03-31	5717.0	1138.0

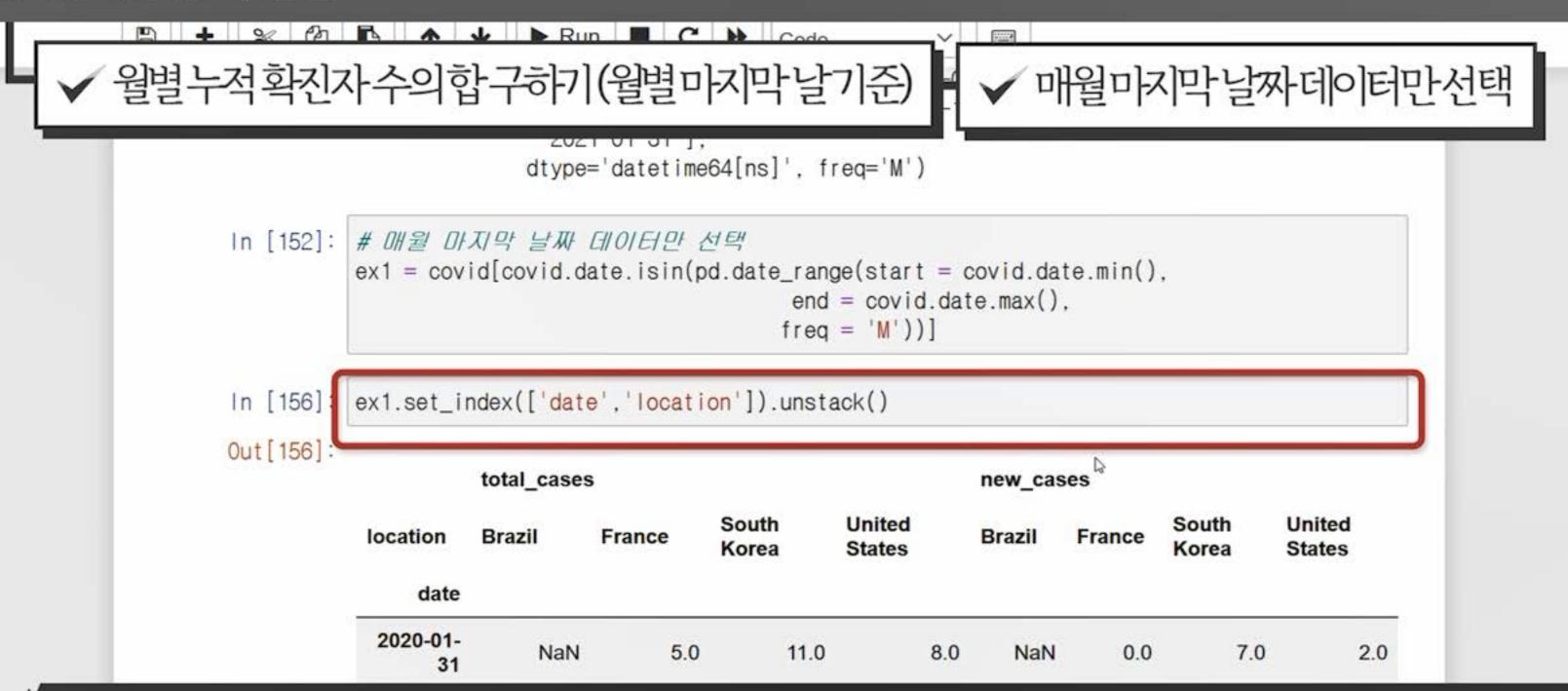
freq의 기본값은 day라서 하루 단위로 데이터를 생성합니다.

23298 France 2020-03-31

52278.0

7629.0

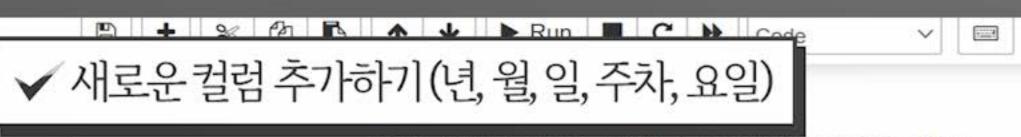
◎ 데이터 분석 실습



DataFrame.plot() 함수 실행을 위해 로우 인덱스와 컬럼 인덱스를 적절히 지정합니다.

2020-03- 5717 0 52279 0 0796 0 102201 0 1129 0 7620 0 125 0 2621

◎ 데이터 분석 실습



- [실습 #1] 각 나라별로 월별 누적 확진자수의 합 구하기
- [실습 #2] 각 나라별로 요일별 신규 확진자수의 합 구하기
- [실습 #3] 각 나라별로 분기별 신규 확진자수의 합 구하기 ...

apply() 함수를 활용하여 strftime() 함수 적용하기

```
covid.date.apply(lambda X: X.strftime(''))
In [160]
Out [160]:
          9104
                   2020-02-26
                   2020-02-27
           9105
                   2020-02-28
           9106
                   2020-02-29
           9107
                   2020-03-01
           9108
           67543
                   2021-02-19
                   2021-02-20
           67544
                   2021-02-21
           67545
                   2021-02-22
           67546
                   2021-02-23
           67547
```

◎ 데이터 분석 실습

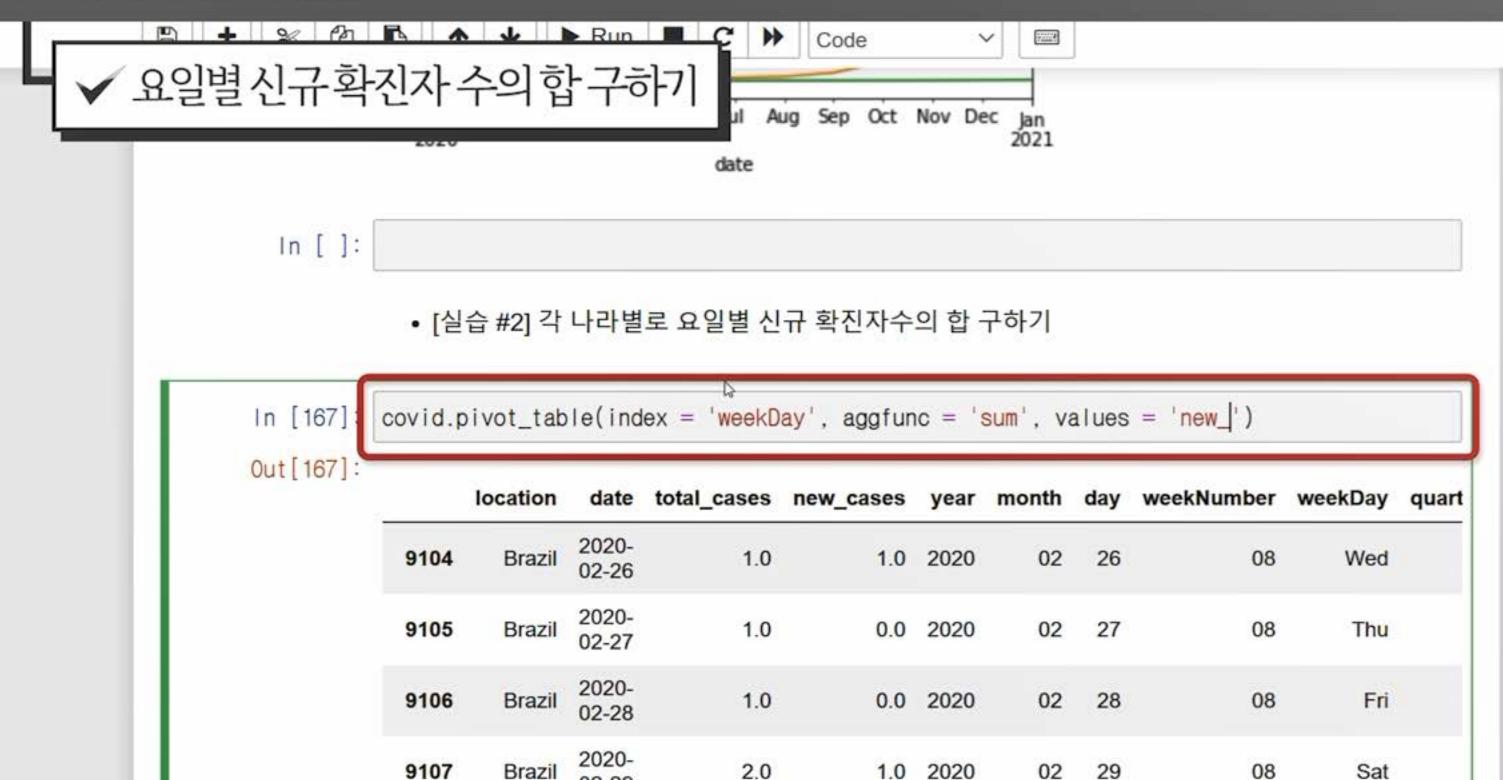
```
+ 90 Pm Run C >>
                                                                7222
                                                 Code
· 쿼터(분기) 정보추가하기
                                 bvid.date.apply(lambda X: X.strftime('%d'))
                                   = covid.date.apply(lambda X: X.strftime('%U'))
                   covid['weekDay'] = covid.date.apply(lambda X: X.strftime('%a'))
          In [163]: covid.date.apply(lambda X: X.strftime('%q'))
                   ValueFrror
                                                       Traceback (most recent call last)
                   <ipython-input-163-23db5bdb460b> in <module>
                   ----> 1 covid.date.apply(lambda X: X.strftime('%q'))
                   C:WProgramDataWAnaconda3WlibWsite-packagesWpandasWcoreWseries.py in apply(self, fun
                   c, convert_dtype, args, **kwds)
                     4198
                                    else:
                                       values = self.astype(object)._values
                     4199
                                        mapped = lib.map_infer(values, f, convert=convert_dtype)
                   -> 4200
                     4201
                                 if lan(manned) and is instance (manned[0] Series).
                     분기 정보는 strftime( ) 함수에서 제공하지 않기 때문에
                 Pandas에서 제공하는 다른 함수를 사용해서 추가해야 합니다.
```

---> 1 covid.date.apply(lambda X: X.strftime('%q'))

◎ 데이터 분석 실습

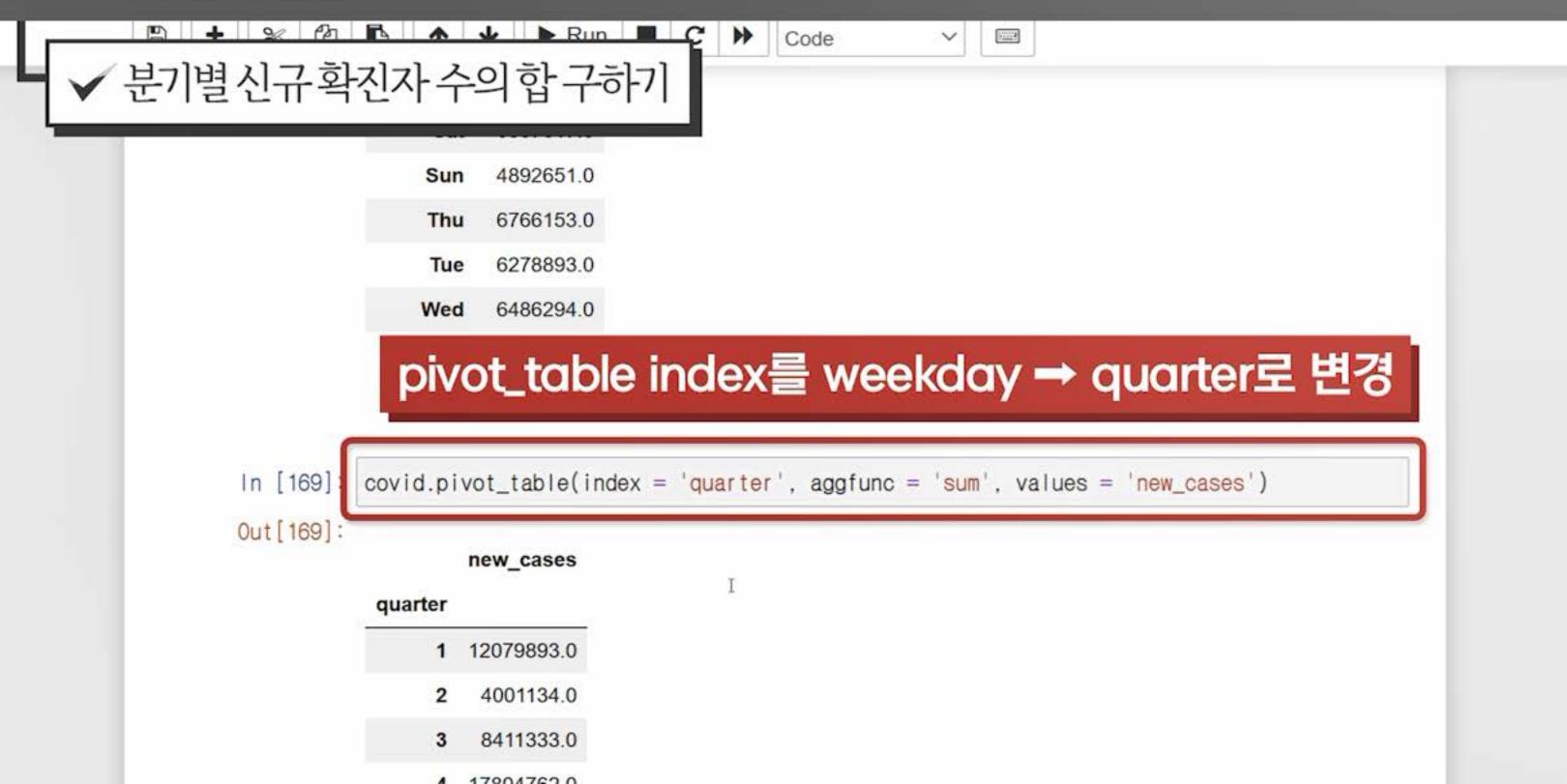
```
Widgets
                                                               Help
                                    Cell
                                           Kernel
             Edit
                    View
                            Insert
✔ 쿼터(분기) 정보추가하기
                                      Run C >>
                                                                       -----
                                                       Code
                      covid['day'] = covid.date.apply(lambda X: X.strftime('%d'))
                      covid['weekNumber'] = covid.date.apply(lambda X: X.strftime('%U'))
                      covid['weekDay'] = covid.date.apply(lambda X: X.strftime('%a'))
             In [163]
                      # 쿼터(분기) 정보 추가
                      covid.date.apply(lambda X: pd.Timestatmp(X).quart)
                                                      pd.Timestamp(X).quarter 실행
                      ValueError
                      <ipython-input-163-23db5bdb460l
                      ----> 1 covid.date.apply(lambda x. x.strrtime( xq ))
                      C:WProgramDataWAnaconda3WlibWsite-packagesWpandasWcoreWseries.py in apply(self, fun
                      c. convert_dtype, args, **kwds)
                         4198
                                         else:
                                             values = self.astype(object)._values
                         4199
                      -> 4200
                                             mapped = lib.map_infer(values, f, convert=convert_dtype)
                         4201
                                     if len(mapped) and isinstance(mapped[0], Series):
                         4202
                      pandasW_libsWlib.pyx in pandas._libs.lib.map_infer()
```

◎ 데이터 분석 실습



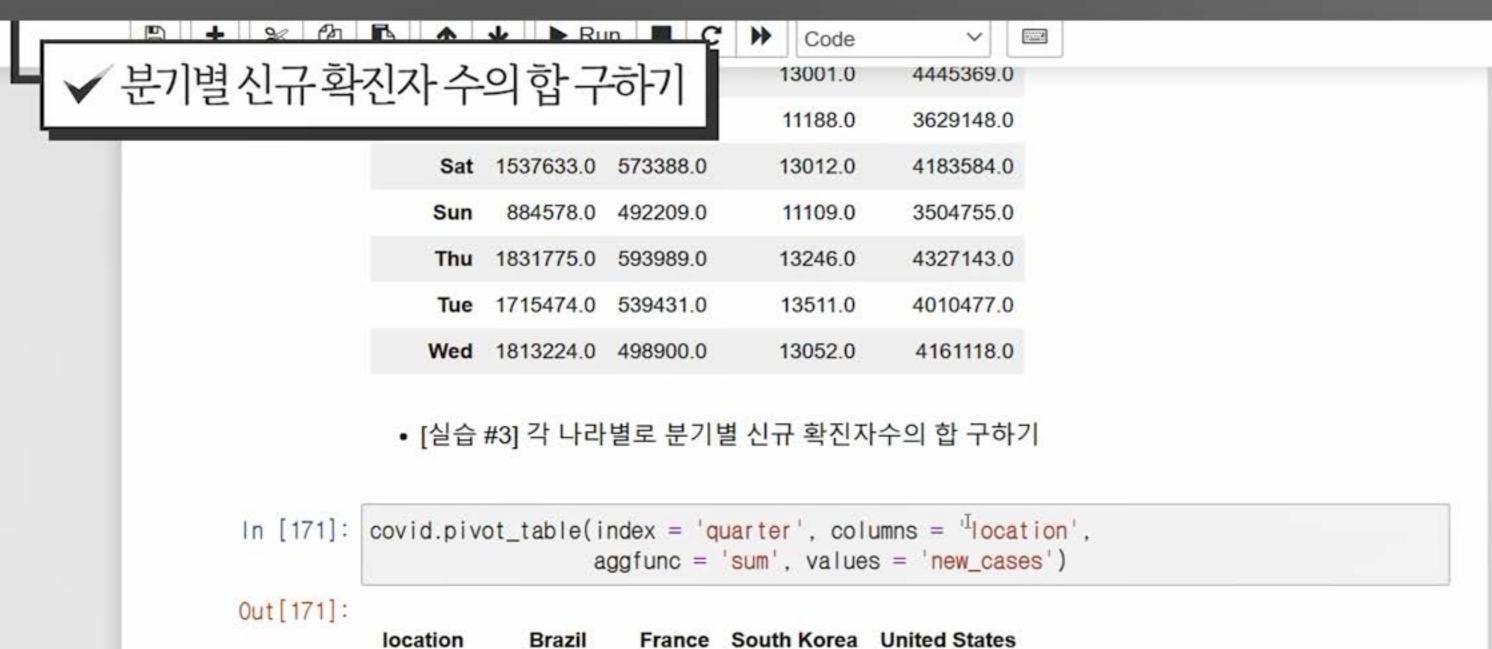
02-29

◎ 데이터 분석 실습



quarter

◎ 데이터 분석 실습



columns에 location을 추가하면 각 나라별, 분기별 확진자 수를 구할 수 있습니다.

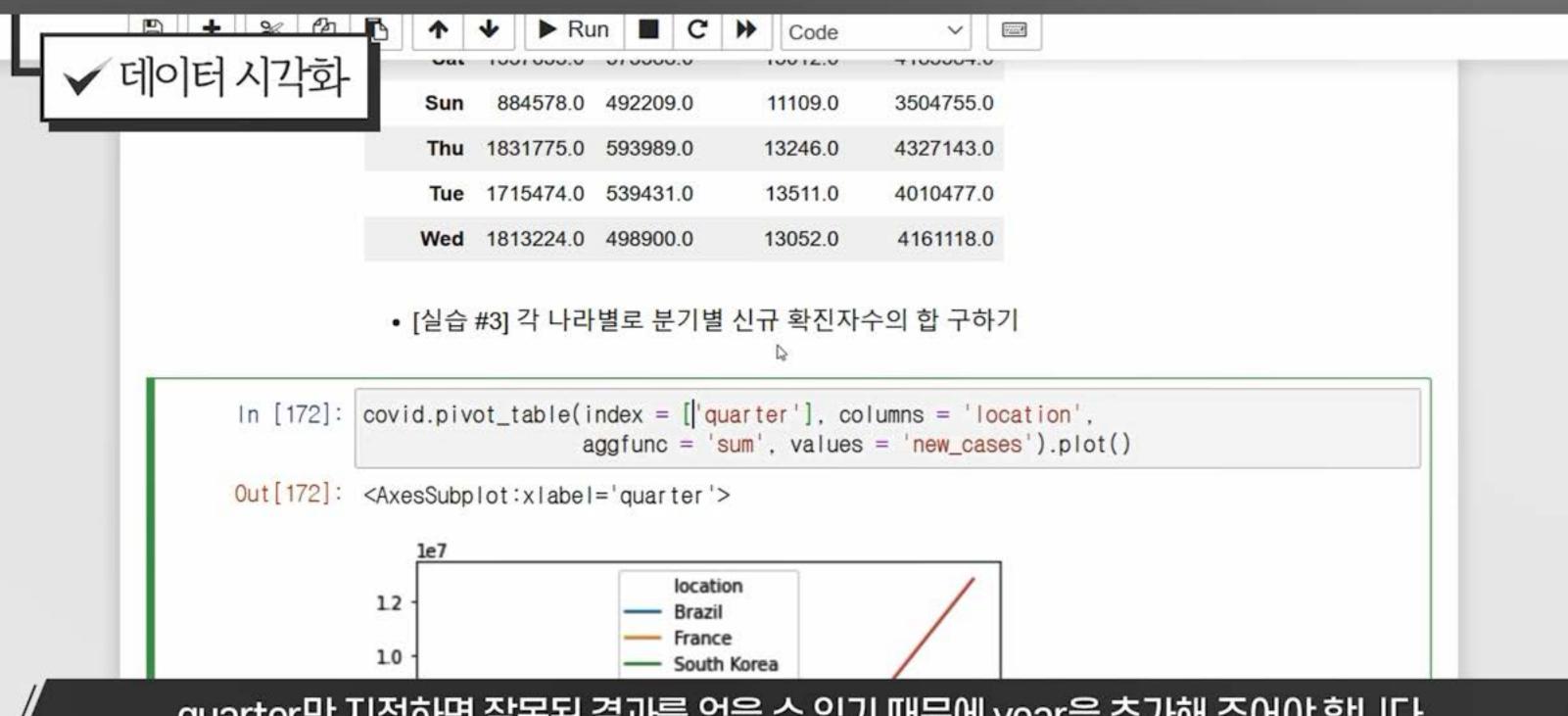
151966.0 1396324.0

3064.0

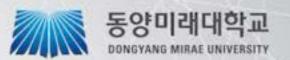
2449780.0

0.6

◎ 데이터 분석 실습



quarter만 지정하면 잘못된 결과를 얻을 수 있기 때문에 year을 추가해 주어야 합니다.



시계열 데이터



학습개요

- 1/ Covid-19 데이터 소개
- 2/ Covid-19 시계열 데이터 처리 및 분석



수고하셨습니다!