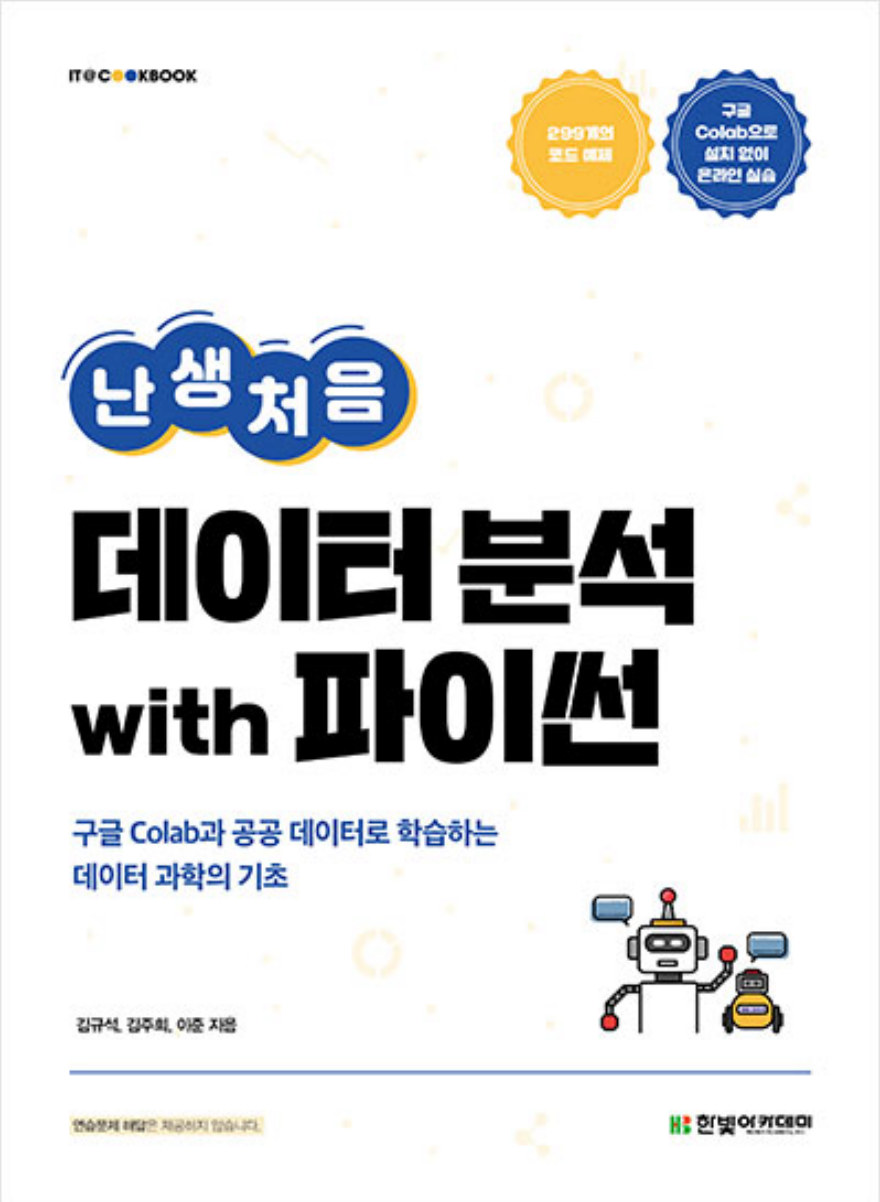
[IT@CookBook, 난생처음 데이터 분석 with 파이썬]
연습문제 해답



Chapter 01. 데이터 과학 이해하기

01 인공지능에 대한 설명 중 틀린 것을 고르시오.
① 인공지능은 인간처럼 생각하고 행동할 수 있는 기계이다.
② 딥러닝은 인공지능을 개발하기 위한 기술이다.
③ 인공지능 연구는 1990년대에 하드웨어 기술 발전 이후에 시작되었다.
④ 분야를 막론한 여러 기업에서 인공지능 기술을 도입 중이다.

02 빅데이터 기술에 대한 설명 중 틀린 것을 고르시오.
① 빅데이터 기술은 빅데이터 자체와 이것을 분석하기 위한 기술을 통칭한다.
② 빅데이터 기술은 인공지능을 개발하기 위한 기반 기술로만 사용한다.
③ 빅데이터 기술로 이전에는 발견하지 못했던 지식이나 가치를 발굴한다.
④ 빅데이터 기술을 활용하여 정형 데이터 및 비정형 데이터를 분석할 수 있다.

03 다음 중 데이터 과학자에게 필요한 기술과 거리가 먼 것을 고르시오.
① 통계 기반 데이터 분석 기술
② 인공지능 활용 기술
③ 비정형 데이터 수집 및 가공 기술
④ GPU 가속 기술

04 인공지능과 빅데이터 기술의 관계에 대한 설명으로 옳은 것을 고르시오.
① 인공지능과 빅데이터 기술은 별개의 기술이라서 서로 상관이 없다.
② 빅데이터 분석을 위하여 인공지능 기술이나 모형을 활용하지 않는다.
③ 인공지능 학습 데이터를 만들기 위해 빅데이터 기술을 활용한다.
④ 데이터 과학자는 빅데이터 기술만을 활용한다.

05 인간 뇌의 신경을 모사한 초기 인공신경 모형을 고르시오.
① 클로바
② 퍼셉트론
③ 랜덤 포레스트
④ 베이즈 이론

06 다음 중 파이썬의 특징이 아닌 것을 고르시오.
① 문법이 아주 쉽다.
② 컴퓨터 중심 언어이다.
③ 다른 언어와 쉽게 결합 가능하다.
④ 다양한 공개 라이브러리가 있다.

07 파이썬에 대한 내용 중 틀린 것을 고르시오.

① 파이썬은 고급 언어에 속한다.

② 파이썬으로만 인공지능을 개발할 수 있다.

③ 파이썬은 대화형 인터프리터를 제공한다.

④ 파이썬 프로그래밍을 위한 다양한 IDE가 있다.

08 이진법에 대한 설명 중에 옳은 것을 고르시오.

① 이진법은 한 자리에 두 가지 상태를 표현할 수 있다.

② 이진법은 숫자 1, 2만 사용한다.

③ 이진법 수를 십진법으로 변환할 수 없다.

④ 이진법은 두 자리 이상인 수의 사칙연산이 불가능하다.

09 다음 중 구글 Colab에 대한 내용으로 틀린 것을 고르시오.

① 구글 Colab은 유료 서비스이기 때문에 비용을 지불해야 사용할 수 있다.

② 구글 Colab에서 파이썬 프로그래밍이 가능하다.

③ 구글 Colab의 파일 단위는 노트북이다.

④ 구글 Colab에서는 파이썬 코드와 일반 텍스트를 모두 입력할 수 있다.

10 다음 중 구글 Colab을 사용할 때 필요한 것으로 거리가 가장 먼 것을 고르시오.

① 인터넷이 가능한 PC　　② 구글 계정

③ 웹 브라우저　　④ GPU

Chapter 02. 데이터 분석을 위한 파이썬 기초

01 다음 중 파이썬에서 지원하지 않는 자료형을 고르시오.

① 숫자 ② 문자열

③ 리스트 ④ 구조체

02 다음 문장을 누락 없이 모두 출력하는 코드를 작성하시오.

```
Teacher said that "Python is 'very' easy."
```

```
print('Teacher said that "Python is \'very\' easy."')
print("Teacher said that \"Python is 'very' easy.\"")
print('''Teacher said that "Python is 'very' easy."''')
```

03 실행 결과가 다음과 같이 나타나도록 빈칸에 들어갈 연산자를 선택하시오.

```
a = 3
b = 8
print(a _______ b)
```

```
True
```

① = ② ==

③ != ④ !!

04 실행 결과가 다음과 같이 나타나도록 조건식을 완성하시오.

```
word = 'school'
if __________ :
  print('high school')
else :
  print('university')
```

```
high school
```

연산 결과가 True인 모든 식이 정답입니다.

```
True
1
word == 'school'
type(word) == str
len(word) <= 7
```

```
len(word) % 2 == 0
's' in word
'u' not in word
```

05 반복문의 실행 결과가 다음과 같이 나타나도록 코드를 완성하시오.

```
data = ['kim', 'lee', 'park']
for i in data :
  print(i)
```

```
kim
lee
park
```

06 for 반복문을 작성하여 구구단 6단의 결과를 출력하는 코드를 완성하시오.

```
a = 6
for i in ______ :
  print(a*i)
```

```
range(1,10)
```

07 실행 결과가 다음과 같이 나타나도록 리스트의 데이터 일부를 삭제하려 한다. 빈칸에 들어갈 명령어를 작성하시오.

```
data = ['kim', 'lee', 'park']
data.remove('lee')
print(data)
```

```
['kim', 'park']
```

08 if 조건문을 작성하여 리스트에서 가장 큰 값을 반환하는 함수를 정의하는 코드를 완성하시오.

```
def max_list(a):
  j = 0
  for i in a:
    if ___ⓐ___ :
      ___ⓑ___
    return j
```

```
ⓐ i > j ⓑ j = i
```

09 다음 딕셔너리에 Key = '가족' , Value = ['아빠', '엄마', '동생']을 추가하는 코드를 작성하시오.

```
dict_1 = {'name': '홍길동', 'age': 22}
```

```
dict_1['가족'] = ['아빠', '엄마', '동생']
```

10 리스트 요소의 합을 구하는 함수를 정의하려 한다. 빈칸에 들어갈 명령어를 작성하시오.

```
def sum_list(a):
  j = 0
  for i in a:
    j = j + i
  return j
```

Chapter 03. 파일 입출력

01 다음 중 CSV 데이터를 모두 고르시오.

① a / b / c / d ② a, b, c, d ③ a: b: c: d ④ 'a', 'b', 'c', 'd'

02 다음 중 파이썬 csv 라이브러리의 함수 open( )의 권한 옵션이 잘못된 것을 고르시오.

① 읽기 권한: r
② 수정 권한: w
③ 추가 권한: a
④ 저장 권한: s

03 다음 코드를 실행했을 때 변수 result에 저장되는 값을 고르시오.

```
result = sheet1['A1'].value
```

[sheet1의 데이터 구조]

	A	B	C	D
1	Banana	apple	Tomato	Cherry
2	3.0	11.7	4.2	5.6
3	Yellow	Red	Red	Purple

① Yellow ② Banana ③ Red ④ Purple

04 문제 03의 데이터 구조에서 C1 셀에 위치한 데이터의 값을 알고 싶다. 작성해야 하는 코드를 고르시오.

① `sheet1['C:1']`
② `Sheet1['C1']`
③ `sheet1['C1']`
④ `sheet1['C', '1']`

05 다음 중 파이썬 csv 라이브러리의 함수가 아닌 것을 고르시오.

① `writer` ② `writerow` ③ `adder` ④ `close`

06 다음 중 워크북(Workbook)과 워크시트(Worksheet)에 관한 설명으로 옳지 않은 것을 고르시오.

① 하나의 워크북에 워크시트는 1개 이상 존재해야 한다.
② 워크북은 엑셀 파일을 의미한다.
③ 파이썬 함수를 사용하여 워크시트의 이름을 변경할 수 있다.
④ 워크북을 생성할 때 워크시트 1개를 같이 생성해야 한다.

07 아래와 같은 코드가 기본 선언되어 있을 때, 정상적으로 수행되지 않는 코드를 고르시오.

```
import openpyxl
wb = openpyxl.Workbook( )
```

① ws2 = wb.create_sheet('Sheet2')

② del wb['Sheet2']

③ sheet1 = wb['Sheet1']

④ print(wb.sheetnames)

08 파이썬 csv 라이브러리를 사용하여 한글이 포함된 csv 파일을 읽을 때, 인코딩 옵션으로 적절한 것을 고르시오. 단, 파일은 윈도우 운영체제에서 만들어졌다.

① 'cp949'

② 'utf-8'

③ 'html'

④ 'hex'

09 CSV 파일에 대한 설명으로 틀린 것을 고르시오.

① CSV 파일은 메모장과 같은 텍스트 편집기로도 읽을 수 있다.

② CSV 파일은 탭(Tab)으로 구분된 일반 텍스트이다.

③ 셀에 쉼표(Comma)를 포함시키고 싶다면 해당 셀 전체를 큰따옴표(" ")로 묶어야 한다.

④ CSV 파일에는 다수의 워크시트 개념이 없다.

10 CSV 파일과 엑셀 파일에 대한 설명으로 옳은 것을 고르시오.

① 엑셀 파일은 파일 자체에서 수식 저장 등의 기능을 제공한다.

② CSV 파일은 파일 자체에서 수식 저장 등의 기능을 제공한다.

③ 엑셀 파일을 메모장과 같은 텍스트 편집기로도 읽을 수 있다.

④ CSV 파일을 엑셀 프로그램으로 읽을 수 없다.

Chapter 04. 웹 크롤링

01 다음 중 셀레니움(Selenium)에 관한 설명으로 틀린 것을 고르시오.

① 셀레니움은 코드를 작성하면 웹 브라우저에서 동작을 수행하는 자동화 도구이다.

② 셀레니움은 C#, Java 등의 프로그래밍 언어로는 사용할 수 없고 파이썬으로만 동작한다.

③ 셀레니움은 무료 프레임워크이다.

④ 셀레니움으로 웹에 있는 데이터를 수집할 수 있다.

02 다음 코드로 셀레니움 관련 모듈을 설치했다. 코드와 수행 동작이 바르게 연결되지 않은 것을 고르시오.

```
from selenium import webdriver
from selenium.webdriver.common.keys import Keys

options = webdriver.ChromeOptions( )
options.add_argument('--headless')                #Headless 설정하기
options.add_argument('--no-sandbox')
options.add_argument('--disable-dev-shm-usage')
driver = webdriver.Chrome('chromedriver', options=options)

driver.get('https://www.naver.com')
```

① print(driver.page_source) : 웹페이지의 소스 코드 출력하기

② driver.find_element('xpath', 'A').click( ) : A 위치의 객체 클릭하기

③ driver.find_element('xpath', 'B').send_keys('서울') : B 위치의 xpath 객체에 '서울' 입력하기

④ driver.find_element('xpath', 'B').send_keys(ENTER) : B 위치의 xpath 객체에 Enter 입력하기

03 셀레니움의 '--headless' 옵션에 대한 설명으로 옳은 것은?

① 웹페이지의 제목을 표시하지 않는다.

② 셀레니움에서 작업하는 웹 브라우저를 보이지 않게 한다.

③ 크롬 브라우저에서는 동작하지 않는 옵션이다.

④ 웹페이지에서 텍스트만 남기고 다른 객체들은 표시하지 않는다.

04~08 다음 코드를 실행하여 셀레니움과 관련 라이브러리를 불러온 상태일 때, 질병관리청 홈페이지에 접속하여 '감기예방'을 검색하는 코드를 차례로 작성하시오.

```
from selenium import webdriver
from selenium.webdriver.common.keys import Keys

options = webdriver.ChromeOptions( )
```

```
options.add_argument('--headless')                #Headless 설정하기
options.add_argument('--no-sandbox')
options.add_argument('--disable-dev-shm-usage')
driver = webdriver.Chrome('chromedriver', options=options)
```

04 질병관리청 홈페이지에 접속하는 코드를 작성하시오.

```
webdriver.get('www.kdca.go.kr')
```

05 검색어 입력란을 선택하는 코드를 작성하시오.

```
webdriver.find_element("xpath", '//*[@id="header"]/div/div[2]/a[1]').click()
```

06 검색어를 입력하는 코드를 작성하시오.

```
search_for = webdriver.find_element("xpath",'//*[@id="searchTermMobile123"]')
search_for.click()
search_for.send_keys('감기예방')
```

07 Enter를 눌러 입력한 검색어를 검색하는 코드를 작성하시오.

```
search_for.send_keys(Keys.ENTER)
```

08 페이지의 소스 코드 전체를 출력하는 코드를 작성하시오.

```
print(driver.page_source)
```

09 객체 여러 개에 공통 속성이 있을 때 find_element( ) 함수가 아닌 A 함수를 활용하면 코드의 양을 줄여 간단히 접근할 수 있다. A 함수로 적절한 것은?

① find_element_all( )
② find_all_elements( )
③ find_elements( )
④ find_all( )

10 다음 명령어에 대한 설명으로 잘못된 것은?

```
!pip install selenium
!apt-get update
!apt install chromium-chromedriver
!cp /usr/lib/chromium-browser/chromedriver /usr/bin
```

① 셀레니움을 구글 Colab에서 활용하기 위하여 관련 라이브러리를 설치하는 명령어이다.
② 위 명령어는 셀레니움을 통해 새로운 웹페이지에 접속하려면 매번 다시 실행해야 한다.
③ 셀레니움과 크롬 브라우저를 연동하기 위한 라이브러리를 설치하는 명령어가 포함되어 있다.
④ 구글 Colab에서는 위 명령어를 실행해야 라이브러리를 가져올 수 있다.

Chapter 05. 데이터베이스

01 SQL에 대한 설명을 읽고 맞으면 '○', 틀리면 '×'를 표시하시오.

① SQL은 구조화된 질의 언어이다. (O)

② NULL은 0 또는 공백을 의미한다. (X)

③ 테이블은 행과 열로 구성된다. (O)

④ 기본키로 NULL이나 중복되는 값을 입력할 수 있다. (X)

02 SQL 언어에서 테이블 두 개를 수직방향으로 합하기 위해 사용하는 명령어를 고르시오.

① UNION

② INTERSECT

③ MINUS

④ JOIN

03 다음 _____에 들어갈 알맞은 단어를 고르시오.

```
SELECT * FROM 학생 WHERE 학년 _____ (1,2,3);
```

① =

② IN

③ OR

④ LIKE

04~10은 다음 학생 테이블과 수강 테이블을 보고 답하시오.

학번	이름	주소	전화번호	생년월일
s1	홍길동	경기 파주	010-1111-1111	2001-01-15
s2	강감찬	전북 전주	010-2222-2222	2002-12-25
s3	을지문덕	서울 강남	010-3333-3333	2000-05-05
s4	이순신	대전 유성	010-4444-4444	2002-07-17
s5	김유신	강원도 원주	해당사항 없음	NULL

(a) 학생 테이블

학번	과목번호	학점
s1	c1	A
s1	c2	A
s2	c2	B
s2	c3	D
s4	c1	C
s4	c3	A
s5	c1	B

(b) 수강 테이블

04 수강 테이블에서 과목별 수강인원을 구하기 위한 명령이다. 빈칸에 들어갈 명령어를 고르시오.

```
SELECT 과목번호, _____ AS 수강인원 FROM 수강 GROUP BY 과목번호;
```

① MAX(*)

② SUM(*)

③ COUNT(*)

④ MIN(*)

05 수강 테이블에서 수강 신청이 된 과목번호를 한 번씩만 검색하는 명령이다. 빈칸에 들어갈 명령어를 고르시오.

```
SELECT ______ 과목번호 FROM 수강 ;
```

① ALL

② GROUP

③ UNION

④ DISTINCT

06 학생 테이블에서 학번이 's3'인 학생의 전화번호를 '010-1234-5678'로 변경하는 명령을 완성하시오.

```
UPDATE ___ⓐ___ SET ______ = ___ⓒ___ WHERE ___ⓓ___ = ___ⓔ___ ;
```

ⓐ 학생 ⓑ 전화번호 ⓒ '010-1234-5678' ⓓ 학번 ⓔ 's3'

07 학생 테이블에서 학번이 's3'인 학생의 데이터를 삭제하는 명령을 적으시오.

DELETE FROM 학생 WHERE 학번 = 's3';

08 학생 테이블에서 주소가 '대전' 또는 '서울'이면서 태어난 해가 '2000년'인 학생의 이름과 주소, 생년월일을 구하는 명령을 완성하시오.

```
SELECT 이름, 주소, 생년월일
FROM 학생
WHERE (주소 LIKE '대전%' ___ⓐ___ 주소 LIKE '___ⓑ___')
AND 생년월일 BETWEEN '2000-01-01' ___ⓒ___ '2000-12-31';
```

ⓐ OR ⓑ 서울% ⓒ AND

09 수강 테이블에서 c1 과목을 수강하는 학생 수를 구하는 명령이다. 별칭을 사용해서 제목이 '학생 수'라고 출력되도록 완성하시오.

```
SELECT ___ⓐ___ 학번 ___ⓑ___ "학생 수"
FROM 수강
WHERE 과목번호='c1';
```

ⓐ COUNT ⓑ AS

10 학생 테이블에서 이름에 '신'이 들어있는 학생의 학번과 이름을 검색하는 명령을 적으시오.

SELECT 학번, 이름 FROM 학생 WHERE 이름 LIKE '%신%';

Chapter 06. 넘파이와 판다스

01 다음은 0부터 11까지 정수를 차례로 채운 (3, 4) 넘파이 배열을 생성하는 코드이다. 빈칸을 채워 완성하시오.

```
import numpy as np
a = ___ⓐ___(12).___ⓑ___(3,4)
```

ⓐ np.arange ⓑ reshape

02 다음 코드의 실행 결과를 쓰시오.

```
import numpy as np
list1= [[1, 2, 3], [4, 5, 6], [7, 8, 9]]
a = np.array(list1)
b = a[0:2, 0:2]
print(b)
c = a[1:, 1:]
print(c)
```

```
[[1 2]
 [4 5]]
[[5 6]
 [8 9]]
```

03 다음은 넘파이 배열을 만들어 그 모양을 출력하는 코드이다. 실행 결과를 쓰시오.

```
import numpy as np
arr = np.array([[[1, 2, 3], [4, 5, 6]],
                [[7, 8, 9], [10, 11, 12]]])
print(arr.shape)
```

(2, 2, 3)

04 다음은 2행 3열의 넘파이 배열을 1차원 배열로 변경하고 출력하는 코드이다. 빈칸을 채우시오.

```
import numpy as np
a = np.array([[1, 2, 3], [4, 5, 6]])

___ⓐ___
___ⓑ___
```

ⓐ a = a.flatten()
ⓑ print(a)

05 다음은 넘파이 배열 a의 모양을 변경하는 코드이다. 빈칸에 들어갈 속성과 실행 결과를 쓰시오.

```
import numpy as np

a = np.arange(8)
a.shape = (4, 2)
print(a)
```

- 실행 결과:
[[0, 1],
[2, 3],
[4, 5],
[6, 7]]

06 다음 데이터프레임 df에서 나이가 25세 이상이면서 성별이 '여자'인 데이터를 조회하는 코드를 작성하시오.

	이름	성별	나이	키
0	허준호	남자	30	183
1	이가원	여자	24	162
2	배규민	남자	23	179
3	고고림	남자	21	182
4	이새봄	여자	28	160
5	이보람	여자	26	163
6	이루리	여자	24	157
7	오다현	여자	24	172

```
df[(df['나이'] >= 25) & (df['성별'] == '여자')]
```

07 문제 06의 데이터프레임 df에서 df.descibe( )의 결과를 완성하시오.

	이름	성별	나이	키
count	8	8	8	8
unique	8	2	ⓐ	8
top	허준호	ⓑ	ⓒ	183
freq	1	5	3	1

ⓐ 6 ⓑ 여자 ⓒ 24

08 문제 06의 데이터프레임 df에서 인덱스로 설정하기에 가장 적합한 열을 선택하고 그렇게 생각한 이유를 적으시오. 그리고 인덱스를 변경하는 코드를 작성하시오.

- 가장 적합한 열: '이름' 열
- 이유: 모든 행의 값이 서로 다른 것은 '이름' 열과 '키' 열 두 개이지만 이름은 각 행을 대표하는 의미가 있으므로 '이름' 열이 가장 적절합니다.
- 인덱스를 변경하는 코드: df.set_index('이름', inplace=True)

09 다음은 다차원 배열에서 짝수와 홀수를 따로 출력하는 코드와 실행 결과이다. 빈칸을 채우시오.

```
import numpy as np
list1= [[1, 2],
        [3, 4]]
arr = np.array(list1)
bool_index = (arr % 2 == 0)
print(bool_index)
print('짝수\n', arr[bool_index])
#배열 arr의 인덱스에 바로 조건식을 넣어 간단하게 표현할 수 있음
res = arr[___ⓐ___]
print('홀수\n', res)
```

```
[[___ⓑ___, ___ⓒ___],
 [___ⓓ___, ___ⓔ___]]
짝수
[2 4]
홀수
[___ⓕ___ ___ⓖ___ ]
```

ⓐ arr % 2 != 0 ⓑ False ⓒ True ⓓ False ⓔ True ⓕ 1 ⓖ 3

10 넘파이 배열 arr1을 복사하여 사본 arr2를 만들고 arr1의 데이터만 변경하고자 한다. 다음과 같이 코드를 작성했으나 의도와 다르게 동작한다. 그 이유를 설명하고 원래 의도대로 동작하도록 코드를 수정하시오.

```
#2행 4열 넘파이 배열
arr1 = np.array([[1,2,3,4], [5,6,7,8]])

#배열 복사하기
arr2 = arr1
print('arr1', '\n',arr1)
print('arr2', '\n',arr2)
arr1[0,0] = 10
print('='*20)
print('변경 후 -> arr1', '\n',arr1)
```

```
print('변경 후 -> arr2', '\n',arr2)
```

```
arr1
 [[1 2 3 4]
 [5 6 7 8]]
arr2
 [[1 2 3 4]
 [5 6 7 8]]
====================
변경 후 -> arr1
 [[10 2 3 4]
 [ 5 6 7 8]]
변경 후 -> arr2
 [[10 2 3 4]
 [ 5 6 7 8]]
```

- 이유: 얕은 복사를 하면 새로 만든 변수도 같은 배열을 참조하게 됩니다. 그래서 한 배열의 값을 변경할 때 다른 배열의 내용도 동일하게 변경됩니다. 그렇게 하지 않으려면 깊은 복사를 해야 합니다.
- 수정 내용: arr2 = arr1를 arr2 = arr1.copy()로

Chapter 07. 데이터 시각화

01 다음 파이플롯 함수가 생성하는 그래프 종류를 찾아서 짝지으시오.

① boxplot( )	a. 히스토그램
② scatter( )	b. 산점도
③ pie( )	c. 파이 차트
④ hist()	d. 상자 그래프

① d ② b ③ c ④ a

02 파이플롯의 plot( ) 함수 포맷스타일 인자 값과 표시되는 마커 모양이 잘못 짝지어진 것을 찾으시오.

① o: 원 ② +: 덧셈 기호 ③ s: 사각형 ④ ^: 다이아몬드

03 각 함수와 가장 관련 있는 기능을 보기에서 찾아 적으시오.

① plt.title( ): (i)	② plt.xlabel( ): (d)	③ plt.text( ) : (e)
④ plt.xlim( ): (c)	⑤ plt.ylim( ): (g)	⑥ plt.show( ): (f)
⑦ plt.ylabel( ): (b)	⑧ plt.subplot( ): (h)	⑨ plt.legend( ): (a)

a. 범례	b. y축의 제목	c. x축의 값의 범위
d. x축의 제목	e. 그래프에 문자열 작성	f. 그래프 출력
g. y축 값의 범위	h. 다중 그래프	i. 그래프의 제목

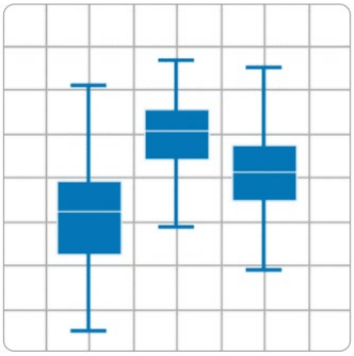
04 다음 함수의 기능을 적으시오.

① re.sub( ) : 정규표현식을 이용하여 특정 문자를 다른 문자열로 치환합니다.
② df.head(3) : 데이터프레임 df의 처음 세 행을 표시합니다.
③ G = nx.Graph( ): 네트워크 그래프 객체를 생성합니다.
④ nx.draw_networkx( ): 네트워크 그래프를 그립니다.

05 다음 중 워드클라우드에 관한 설명으로 옳지 않은 것을 고르시오.

① 워드클라우드는 텍스트 데이터의 단어 빈도수를 시각화하는 방법이다.
② 워드클라우드를 만들 때 단어의 빈도수를 이용하여 워드클라우드 객체를 생성하고, 이를 이미지로 저장하는 과정이 있다.
③ 워드클라우드는 자연어 처리와 관련이 있다.
④ 워드클라우드에서 크기가 큰 단어는 빈도수가 작은 단어이다.

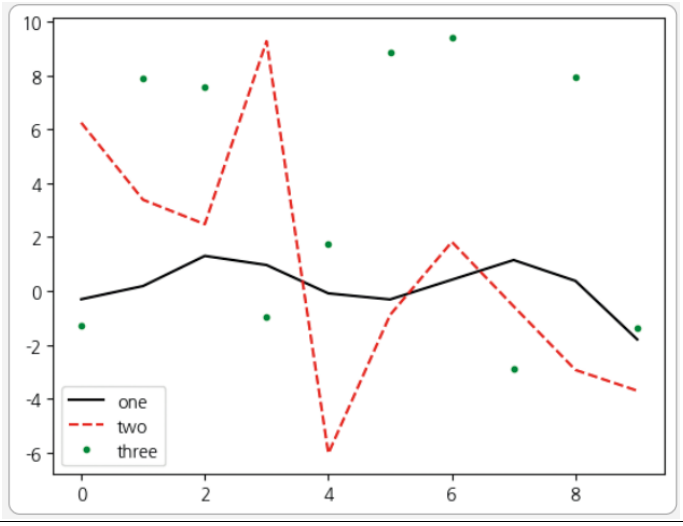
06 다음 중 상자 그래프의 각 요소에 대한 설명으로 옳지 않은 것을 고르시오.



① 상자 그래프에서 최솟값을 가장 아래쪽 선으로 표현한다.
② 상자 그래프에서 제1사분위수는 데이터의 하위 25%에 해당하는 값으로 상자 아래쪽 끝에서 시작하는 가로선으로 표현한다.
③ 상자 그래프에서 제4사분위수는 데이터의 상위 75%에 해당하는 값으로 상자 위쪽 끝 가로선으로 표현된다.
④ 상자 그래프에서 이상치는 상자 바깥에 점으로 표현할 수 있다. 이상치란 전체 데이터 분포에서 벗어나 극단적으로 크거나 작은 값을 의미한다.

07 다음 코드에서 범례를 표시하는 코드를 고르시오.

```
(1) import numpy as np
(2) plt.plot(np.random.randn(10), 'k', label='one')
(3) plt.plot(np.random.randn(10)*3, 'r--', label='two')
(4) plt.plot(np.random.randn(10)*10, 'g.', label='three')
(5) plt.legend()
(6) plt.show()
```



08 다중 그래프를 생성하여 다음과 일치하는 실행 결과가 나타나도록 빈칸에 들어갈 값을 적으시오.

```
import matplotlib.pyplot as plt
plt.subplot(2, __ⓐ__, __ⓑ__)
plt.plot(np.random.randn(10), 'b--')
plt.subplot(2, __ⓒ__, __ⓓ__)
plt.plot(np.random.randn(100), 'r', alpha=0.7)
plt.subplot(2, __ⓔ__, __ⓕ__)
plt.plot(np.random.randn(10), 'y^')
plt.subplot(2, __ⓖ__, __ⓗ__)
plt.plot(np.random.randn(10), 'g.')
plt.show( )
```

ⓐ 2 ⓑ 1 ⓒ 2 ⓓ 2 ⓔ 2 ⓕ 3 ⓖ 2 ⓗ 4

09 다음 중 네트워크X와 네트워크 그래프에 관한 설명으로 옳지 않은 것을 고르시오.

① 네트워크X는 파이썬으로 작성된 네트워크 분석 라이브러리이다.

② 네트워크X는 그래프 이론과 관련된 다양한 알고리즘을 지원한다.

③ 네트워크X는 무료로 사용할 수 있는 오픈소스 라이브러리이다.

④ 네트워크 그래프를 그리려면 맷플롯립 라이브러리를 반드시 불러와야 한다.

10 마트에서 일주일 동안 고객들의 구매 품목을 수집했다. 전체 구매자 500명 중에 맥주와 오징어를 함께 구매한 사람은 200명이었고 오징어만 구매한 사람은 100명이었으며 맥주만 구매한 사람은 없었다. 이때 지지도, 신뢰도, 향상도를 계산하시오. 그리고 장바구니에 맥주를 담은 고객에게 오징어를 추천할 만한지 판단하고 이유를 설명하시오.

맥주&오징어 지지도 = 200 ÷ 500 = 0.4

오징어 지지도 = 300 ÷ 500 = 0.6

맥주&오징어 신뢰도 = 200 ÷ 200 = 1

향상도 = 맥주&오징어 신뢰도 ÷ 오징어 지지도 = 1 ÷ 0.6 = 약 1.7

향상도가 1.7로 1보다 큽니다. 맥주와 관계없이 오징어를 구매할 확률(0.6)에 비해 맥주와 오징어를 함께 구매할 확률(1)이 높다는 의미입니다. 따라서 맥주를 구매하려는 고객에게 오징어를 추천하면 효과적일 것입니다.

Chapter 08. 데이터 전처리

01 설명이 맞으면 '○', 틀리면 '×'로 표시하시오.

① 이상치는 표본의 전체 패턴에서 크게 벗어나는 값이다. (O)

② 데이터 분포를 시각화하면 이상치가 있는지 확인할 수 있다. (O)

③ 이상치를 제거하는 데 IQR을 활용할 수 있다. (O)

④ 데이터 전처리 기법인 정규화는 데이터를 평균이 0이고 표준편차가 1이 되도록 변환하는 것이다. (X)

⑤ Z 점수는 어떤 값 X가 평균에서 얼마나 떨어져 있는지를 나타낸다. (O)

02 데이터가 X일 때, 데이터 전처리에 사용하는 Z 점수를 구하는 계산식을 적으시오.

Z 점수 = (X - 평균) ÷ 표준편차

$$Z\text{점수} = \frac{(X - \text{평균})}{\text{표준편차}}$$

03 다음은 A반 학생 네 명과 B반 학생 네 명의 성적 Z 점수를 구하는 코드와 실행 결과이다. 빈칸을 채우시오.

```
import pandas as pd
#데이터프레임 생성하기
df = pd.DataFrame({'A':[80,90,96,60],
                   'B':[70,70,90,58]})
#각 열 Z 점수 구하기
df_z = (df - ___ⓐ___ ) / ___ⓑ___

df_z
```

	A	B
0	-0.095059	-0.150756
1	0.538666	-0.150756
2	0.918900	1.356801
3	-1.362507	-1.055290

ⓐ df.mean() ⓑ df.std()

04 한빛대학교에 올해 입학한 1학년이 데이터 분석 과목 중간시험과 기말시험을 쳤다. 전체 학생의 중간시험 평균은 81점이고 표준편차는 15점이었다. 기말시험의 평균은 76점이고 표준편차는

13점이었다. 세일이의 1학기 중간시험 점수는 96점이고 기말시험 점수는 90점이다. Z 점수를 이용하여 세일이는 중간시험과 기말시험 중 어느 것을 더 잘했는지 구하시오.

중간시험 z1 = (96 -81) ÷ 15 = 1.0
기말시험 z2 = (90 – 76) ÷ 13 = 1.0769
기말시험의 Z 점수가 더 큽니다. 따라서 세일이는 기말시험을 더 잘 쳤습니다.

05 다음 계산식을 사용하는 데이터 전처리 기법은 무엇인가?

$$X' = \frac{(X - Xmin)}{(Xmax - Xmin)}$$

정규화

06 다음은 데이터프레임 df에서 price 열의 IQR을 구해서 iqr_ price에 저장하는 코드이다. 빈칸을 채우시오.

```
q3_price = df[___ⓐ___].___ⓑ___(q=___ⓒ___)
q1_price = df[___ⓓ___].___ⓔ___(q=___ⓕ___)
iqr_price = _______ⓖ_______
```

ⓐ 'price' ⓑ quantile ⓒ 0.75 ⓓ 'price' ⓔ quantile ⓕ 0.25 ⓖ q3_price - q1_price

07 데이터프레임 df에서 price 열의 상한값보다 큰 이상치 개수를 구하는 코드이다. 빈칸을 채우시오.

```
upper_price = q3_price + ___ⓐ___ * iqr_price
print((df['price'] > upper_price).___ⓑ___) #상한값을 벗어난 이상치 개수
```

ⓐ 1.5 ⓑ sum( )

08 데이터프레임 df의 결측치 개수를 확인하는 코드를 적으시오.

```
df.isnull( ).sum( )
```

09 데이터프레임 df의 결측치를 바로 위에 있는 행의 값으로 대체해서 변수 df_ffill에 저장하는 코드를 작성하시오.

```
df_ffill = df.fillna(method='ffill')
```

10 데이터프레임 df의 결측치를 각 열의 평균으로 대체해서 변수 df_mean에 저장하는 코드를 작성하시오.

```
df_mean = df.fillna(df.mean( ))
```

Chapter 09. 상관관계 분석

01 다음 중 상관관계에 대한 설명으로 틀린 것을 고르시오.

① 상관계수가 양수일 때 두 변수는 양의 상관관계를 갖는다.

② 상관계수가 클수록 두 변수의 선형 상관도가 낮다.

③ 상관분석은 인과관계를 의미하는 것은 아니다.

④ 정규분포를 따르지 않는 데이터도 스피어만 상관분석을 수행할 수 있다.

02 아래 코드의 빈칸에 피어슨 상관분석을 수행하는 명령을 작성하시오.

```
import pandas as pd
data = {'A' : [1, 2, 3], 'B' : [4, 5, 6]}
df = pd.DataFrame(data)
coef = df.corr(method = 'pearson')
print(coef)
```

03 다음 중 스피어만 상관분석을 수행하는 코드를 고르시오.

① df.corr(method='pearson')

② df.spearman( )

③ df.corr( )

④ df.corr(method='spearman')

04 다음 데이터의 변수 '직장 만족도'와 '월급'의 관계를 피어슨 상관분석으로 구하는 코드를 작성하시오.

```
import pandas as pd
data = {'직장 만족도': [6, 7, 9, 10, 3, 6, 4, 10, 8, 5, 7],
        '월급': [350, 360, 400, 390, 230, 310, 280, 500, 390, 230, 400]}
df = pd.DataFrame(data)
coef = df.corr(method='pearson')
print(coef)
```

05 04 코드의 실행 결과를 완성하시오.

```
             직장 만족도          월급
[__ⓐ__]       [__ⓑ__]       [__ⓒ__]
[__ⓓ__]       [__ⓔ__]       1.00000
```

ⓐ 직장 만족도 ⓑ 1.00000 ⓒ 0.89133 ⓓ 월급 ⓔ 0.89133

06 05의 분석 결과를 해석하는 문장을 완성하시오.

변수 '직장 만족도'와 '월급'은 __ⓐ__ 양의 선형 상관관계가 있다. 일반적으로 상관계수의 절댓값이 __ⓑ__ 보다 클 때 강한 선형 상관관계라고 판단할 수 있기 때문이다. 상관분석 결과에 따라 직장 만족도가 더 높은 사람은 월급을 더 __ⓒ__ 받고 있을 것이라고 추측할 수 있다.

ⓐ 강한 ⓑ 0.5 ⓒ 많이

07 최분석 사원이 사장님께 04~06의 피어슨 상관분석 결과와 해석을 제출하며 자신의 직장 만족도가 10이기 때문에 월급을 500만 원으로 인상해야 한다고 주장했다. 상관관계 분석의 성질을 근거로 하여 주장에 반박하시오.

두 변수에 선형 상관관계가 있더라도 인과관계가 없을 수 있습니다. 그러므로 높은 직장 만족도가 높은 월급의 원인이라고 말할 수 없습니다.

08 다음 데이터의 변수 '키'와 '몸무게'의 상관관계를 분석하는 코드를 작성하고 결과를 해석하시오.

키(cm)	153	176	183	173	177	166	158	163	190	150	155	175
체중(kg)	60	77	80	82	90	50	55	57	80	52	48	60

```
import pandas as pd
data = {'키': [153, 176,183,173,177,166,158,163,190,150,155,175],
        '몸무게': [60,77,80,82,90,50,55,57,80,52,48,60]}
df = pd.DataFrame(data)
coef = df.corr()
print(coef)
```

```
             키     몸무게
키       1.000000  0.787468
몸무게   0.787468  1.000000
```

상관계수가 0.787468이므로 키와 몸무게는 강한 양의 선형 상관관계가 있다.

09 다음 데이터의 변수 '과학 점수'와 '영어 점수', '과학 점수'와 '수학 점수'의 상관관계를 분석하는 코드를 작성하고 결과를 해석하시오.

영어 점수	89	92	99	66	70	90	80	100	80	70	100	60
과학 점수	80	90	77	80	50	80	50	98	90	30	40	55
수학 점수	98	99	77	86	49	50	33	100	96	30	40	40

```
import pandas as pd
```

```
data = {'Eng': [89, 92, 99, 66, 70, 90, 80, 100, 80, 70, 100, 60],
        'Sci': [80, 90, 77, 80, 50, 80, 50, 98, 90, 30, 40, 55],
        'Math': [98, 99, 77, 86, 49, 50, 33, 100, 96, 30, 40, 40]}
df = pd.DataFrame(data)
corr_matrix = df.corr()
print(corr_matrix)
```

```
          Eng       Sci      Math
Eng  1.000000  0.391075  0.372532
Sci  0.391075  1.000000  0.898752
Math 0.372532  0.898752  1.000000
```

영어 점수와 수학 점수는 상관계수가 0.37로 약한 양의 상관관계가 있으며, 과학 점수와 수학 점수는 상관계수가 0.89로 강한 양의 상관관계가 있다.

Chapter 10 회귀분석

01 다음 중 회귀분석에 대한 설명으로 틀린 것을 고르시오.

① 독립변수의 기울기가 되는 값을 계수라고 한다.

② 독립변수의 계수가 작을수록 다른 변수에 비해 영향력이 크다.

③ 종속변수는 단 한 개만 존재한다.

④ 독립변수의 유의수준이 0.05 미만일 때 유의한 변수로 판단한다.

02 다음 중 선형 회귀분석에 대한 설명으로 틀린 것을 고르시오.

① 원인인 x 변수가 한 개이면 단순 선형 회귀분석, 두 개 이상이면 다중 선형 회귀분석이다.

② 잔차는 실제 값과 예측 값의 차이이다.

③ R^2는 데이터 표본에 대한 회귀모형의 설명력이다.

④ R^2가 작을수록 데이터를 잘 설명하는 모형이다.

03 다음 중 통계적 가설검정에 대한 설명으로 틀린 것을 고르시오.

① 유의수준은 일반적으로 0.05로 정한다.

② 귀무가설은 대립가설과 반대되는 개념이다.

③ p-값이 클수록 대립가설이 받아들여지기 쉽다.

④ 유의수준이 0.05이면 귀무가설이 참인데도 받아들이지 않을 가능성이 5% 미만이라는 것이다.

04 다음 중 회귀분석과 통계적 가설검정에 대한 설명으로 틀린 것을 고르시오.

① 더미 변수는 '예' 또는 '아니오'로 변환하여 1 또는 0으로 표현할 수 있는 독립변수이다.

② 회귀분석에서 결과인 y 변수는 단 한 개 존재하며, 원인인 x 변수는 두 개 이상 존재할 수 있다.

③ 귀무가설은 연구자가 입증하고자 하는 가설이다.

④ 범주형 데이터를 표현하기 위해 더미 변수를 활용한다.

05 다음 코드의 빈칸을 채워 데이터프레임 df의 종속변수 y와 독립변수 x1, x2의 회귀분석을 완성하시오.

```
import pandas as pd
from statsmodels.formula.api import ols
df = pd.DataFrame(Data)
ols = (y ~ x1 + x2, data=df).fit( )
```

06 다음 중 선형 회귀분석의 결과를 출력하는 코드를 고르시오.

① output( )
② organize( )
③ results( )
④ summary( )

07 다음 내용을 읽고 선형 회귀분석의 종속변수와 독립변수를 설정하시오.

- 30 세 이상 60 세 미만의 성인을 대상으로 삶의 행복 점수와 그 요인을 조사하였다.
- 전 연령대의 성인에게 삶의 행복 점수에 가장 큰 영향을 미치는 요인은 소득이었으며, 가족 관계, 여가 시간, 친구 관계 순으로 나타났다.

- 종속변수: 행복 점수
- 독립변수: 소득, 가족 관계, 여가 시간, 친구 관계

08 다음 선형 회귀분석의 결과를 보고 회귀모형을 작성하시오. ©Rohit Gupta, C# Corner

OLS Regression Results

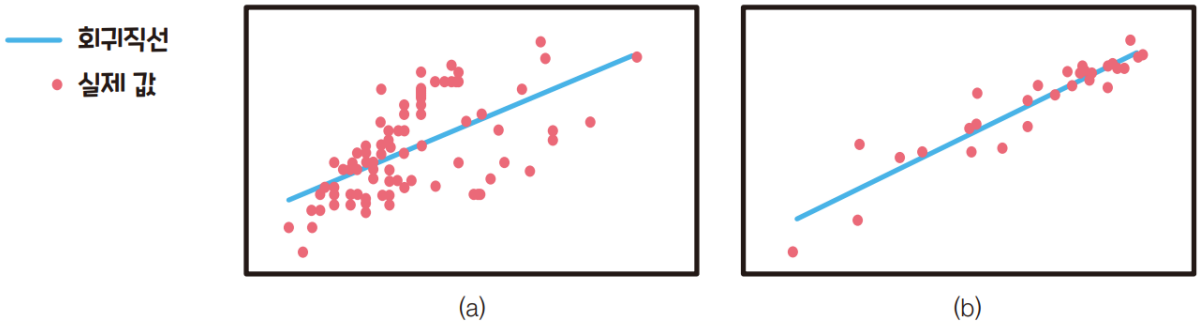
Dep. Variable:	Stock_Index_Price	R-sqared:	0.898
Model:	OLS	Adj. R-squared:	0.888
Method:	Least Squares	F-statistic:	92.07
Date:	Sun, 27 Oct 2019	Prob (F-statistic):	4.04e-11
Time:	11:55:09	Log-Likelihood:	-134.61
No. Observations:	24	AIC:	275.2
Df Residuals:	21	BIC:	278.8
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1798.4040	899.248	2.000	0.059	-71.685	3668.493
Interest_Rate	345.5401	111.367	3.103	0.005	113.940	577.140
Unemployment_Rate	-250.1466	117.950	-2.121	0.046	-495.437	-4.856

Omnibus:	2.691	Durbin-Watson:	0.530
Prob(Omnibus):	0.260	Jarque-Bera (JB):	1.551
Skew:	-0.612	Prob(JB):	0.461
Kurtosis:	3.226	Cond. No.	394.

y = 1798.4040 + 345.5401 × (Interest_Rate) + (-250.1466) × (Unemployment_Rate) + ε

09 두 선형 회귀분석의 그래프를 보고 결정계수 R^2 값이 더 클 것으로 예상되는 모형은 어느 것인지 설명하시오.



(b)의 결정계수 값이 더 클 것입니다. 데이터가 비교적 회귀직선과 가깝게 분포해 있어 잔차가 작기 때문입니다.

10 다음 데이터는 알코올 중독 여부와 심장질환 여부 조사 결과이다. 알코올 중독자인 사람이 심장질환이 있는 사건의 오즈비를 구하시오.

항목	심장질환 있음	심장질환 없음	합계
알코올 중독	8	4	12
알코올 중독 아님	42	96	138
합계	50	100	150

OR = (알코올 중독이면서 심장질환 있을 확률) ÷ (알코올 중독이면서 심장질환 없을 확률) = (8÷12) ÷ (4÷12) = 2

Chapter 11 인공지능 분석

01 다음 중 분류와 예측에 대한 설명으로 틀린 것을 고르시오.

① 분류 문제의 결과로 데이터를 그룹으로 묶을 수 있다.

② 분류 문제의 결과 클래스가 3개 이상일 때는 기계학습으로 해결할 수 없다.

③ 레이블이 연속적인 값을 가질 때 예측 문제의 해결 방식을 적용한다.

④ 범주형 자료형의 종류에는 명목형과 순서형이 있다.

02 웹에서 제공하는 CSV 파일을 읽어와서 콤마(,)를 구분자로 하여 데이터프레임 df_test에 저장하는 코드를 완성하시오.

```
df_test = pd.___ⓐ___('http://www.hanbit.co.kr/test.csv', ___ⓑ___',')
```

ⓐ read.csv ⓑ sep=

03 다음 중 특성 간의 상관관계를 계산하여 시각화하는 함수를 고르시오.

① output( )

② maps( )

③ heatmap( )

④ heats( )

04 원핫 인코딩을 수행하기 위해 사이킷런의 라이브러리를 로드하려고 한다. 빈칸에 알맞은 라이브러리를 입력하시오.

```
from sklearn.preprocessing import LabelEncoder
```

05 다음 중 인공지능에 대한 설명으로 틀린 것을 고르시오.

① 인공지능 모형이 학습을 하려면 학습 데이터가 필요하다.

② 하이퍼파라미터 튜닝으로 모형의 성능을 극대화할 수 있다.

③ 데이터의 품질에 관계없이 데이터가 많을수록 모형 성능이 강화된다.

④ 테스트 데이터를 통해 모형의 성능을 평가할 수 있다.

06 데이터프레임(df_test)에서 결측치(NaN)를 찾고 결측치가 존재하는 행을 모두 제거하는 코드를 완성하시오.

```
df_test.dropna()
```

07 다음 중 인공지능 모형 학습에 대한 설명으로 옳은 것을 고르시오.

① 모형 학습의 결과는 정확도로만 평가한다.

② 딥러닝 모형이 기계학습 모형보다 무조건 성능이 좋으므로 적극 활용해야 한다.

③ 레이블의 불균형에 따른 평가 오류는 발생하지 않는다.

④ 모형 성능 향상을 위하여 데이터 전처리는 매우 중요하다.

08 다음 중 인공신경망을 구성할 때 층을 추가하는 함수를 고르시오.

① plus( )

② add( )

③ surplus( )

④ deep( )

09 인공신경망의 출력층에 사용할 수 있는 활성화 함수를 모두 고르시오.

① relu( )

② sigmoid( )

③ softmax( )

④ lstm( )

10 다음 중 인공신경망 구성에 대한 설명으로 옳은 것을 고르시오.

① 은닉층에서는 데이터 특성 정보를 다시 입력받을 입력 은닉층이 필요하다.

② 은닉층에서 뉴런 수는 입력층보다는 작고 출력할 클래스 수보다는 크게 설정한다.

③ 출력층에서의 뉴런 개수는 분류 대상의 클래스 개수와 무관하다.

④ 입력층에서 활성화 함수는 필요 없다.

Chapter 12 시계열 예측

01 시계열 예측에 대한 설명 중 틀린 것을 고르시오.

① 시계열 예측을 위해서는 시간을 기준으로 수집 또는 생성된 데이터가 필요하다.

② ADF 검정은 p-value가 0.05 이상이면 귀무가설을 기각하여 정상성으로 판정한다.

③ 대표적인 시계열 데이터 패턴에는 추세, 주기성, 계절성 등이 있다.

④ AR, MA 등 통계 모형을 시계열 예측에 사용할 수 있다.

02 데이터 AirPassengers를 air라는 이름의 데이터프레임에 저장하는 코드를 완성하시오.

```
air = pd.DataFrame(data('AirPassengers'))
```

03 다음 중 ADF 검정을 수행하기 위한 stattools 라이브러리의 함수명을 고르시오.

① adf( )

② kpss( )

③ adfuller( )

④ tsa( )

04 계절성을 가지는 비정상성 시계열 데이터에 대하여 계절 차분을 수행하는 코드를 완성하시오.

```
df_log_air_diff_season = df_log_air_diff.diff(12)
```

05 시계열 데이터의 정상성과 비정상성에 대한 설명 중 틀린 것을 고르시오.

① ADF 검정은 시계열 데이터의 추세를 잘 검출할 수 있다.

② KPSS 검정은 계절성 파악에 활용할 수 있다.

③ 시계열 데이터에 추세가 존재하면 무조건 변환을 수행하여야 한다.

④ pmdarima 라이브러리는 정상성 검정을 위한 다양한 함수를 제공한다.

06 데이터에서 최댓값과 최솟값의 편차가 클 때 데이터가 적절한 범위에 속하도록 스케일링해야 한다. 데이터 스케일링에 필요한 사이킷런의 라이브러리를 로드하는 코드를 완성하시오.

```
from sklearn.preprocessing import MinMaxScaler
```

08 다음 중 시계열 데이터 변환에 대해 올바른 설명을 모두 고르시오.

① 시계열 데이터는 항상 비정상성을 가지고 있다.

② 시계열 데이터의 추세를 로그 변환을 통해 제거할 수 있다.

③ 시계열 데이터의 계절성을 차분 변환을 통해 제거할 수 있다.

④ 시계열 데이터 예측 문제에서 데이터 전처리는 필요하지 않다.

09 다음 중 RNN에서 장기간 순차적인 데이터에 대한 기억을 유지시키고 싶을 때 사용하는 활성화 함수를 고르시오.

① softmax

② lstm

③ relu

④ sigmoid

10 다음 중 학습이 완료된 인공신경망을 활용하여 시계열 예측을 수행할 때 호출할 함수를 고르시오.

① call( )

② predict( )

③ reshape( )

④ save( )

11 시계열 데이터의 특성을 파악하기 위해 그래프 시각화를 수행하는 코드를 완성하시오.

```
import matplotlib.___ⓐ___ as plt
plt.___ⓑ___(data)
plt.___ⓒ___( )
```

ⓐ pyplot ⓑ plot ⓒ show