

# 환경 설정하기

- 아나콘다 설치하기

아나콘다 파이썬 다운로드

전체 이미지 동영상 쇼핑 뉴스 웹 도서 : 더보기



anaconda.com

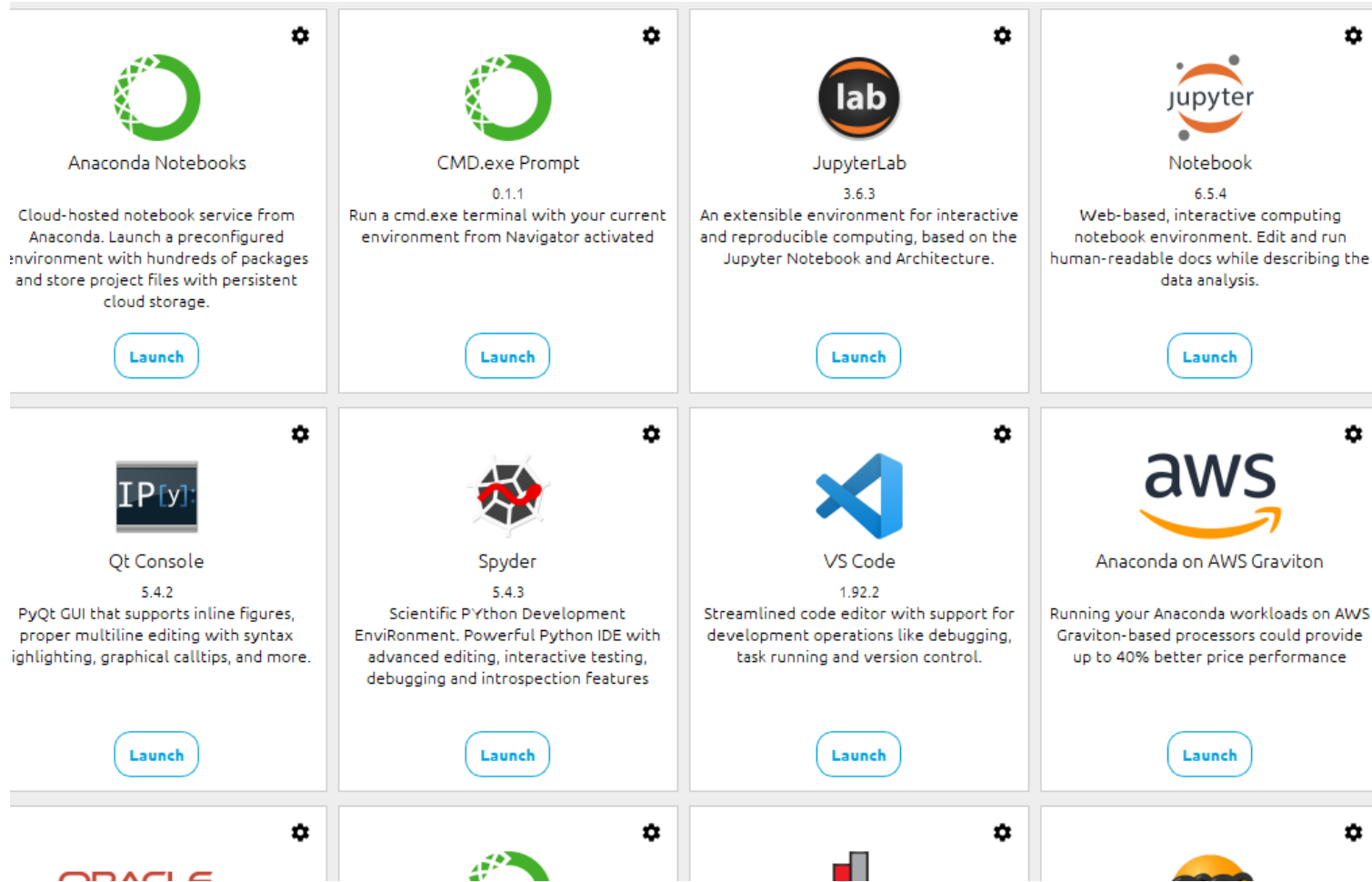
<https://www.anaconda.com> > download

## Download Anaconda Distribution

**Download Anaconda's open-source Distribution** today. Discover the easiest way to perform Python data science and machine learning on a single machine.

[Anaconda Navigator](#) · [Plans and Pricing](#) · [Toolbox for Notebooks](#) · [Open Source](#)

# 스파이더(spyder) 실행



# spyder 화면

The screenshot displays the Spyder Python IDE interface. The main window is divided into three panes: a code editor on the left, a variable explorer on the top right, and a console on the bottom right.

**Code Editor:** The code is written in Python and uses the `requests` library to fetch data from a URL. It then processes the JSON response into a pandas DataFrame and saves it as a JSON file.

```
5 @author: WD
6 """
7
8 import requests
9 url="http://data4library.kr/api/LoanItemSrc/
10 &startDt=2024-01-01&endDt=2024-08-18&age
11 r=requests.get(url)
12 data=r.json()
13 # print(data)
14
15 books=[]
16 for d in data['response']['docs']:
17     books.append(d['doc'])
18 import pandas as pd
19
20 books_df=pd.DataFrame(books)
21 print(books_df.columns)
22
23 books_df.to_json('best_book_20s2.json', force
24 #%%
25 books=books_df[['no', 'ranking', 'bookname',
26 'isbn13']]
27 print(books.head())
28 #%%
29 import requests
30 isbn=books.loc[0, 'isbn13']
31 url='https://www.yes24.com/product/search?q=
32 n=requests.get(url.format(isbn))
```

**Variable Explorer:** This pane shows the current state of variables in the workspace. It contains a table with columns: Name, Type, Size, and Value.

Name	Type	Size	Value
i	int	1	1
m	int	1	2
matrix	Array of int32	(2, 3)	$\begin{bmatrix} 1 & 2 & -1 \\ 0 & -2 & 3 \end{bmatrix}$
n	int	1	2
r	list	3	[3, 4, 0]

**Console:** The console shows the output of the code execution, including the file path for the JSON file and the resulting DataFrame structure.

```
In [98]: runfile('C:/Users/WD/OneDrive/2024/ 동아대/2 학기/
선행대/수학/source/ot(1).py', wdir='C:/Users/WD/OneDrive/2024/
동아대/2 학기/ 선행대/수학/source')
m x n 2 2
1 2 -1
3 4 0
[[-2 -4 -1]
 [ 3 4 0]]

In [99]: runfile('C:/Users/WD/OneDrive/2024/ 동아대/2 학기/
선행대/수학/source/ot(1).py', wdir='C:/Users/WD/OneDrive/2024/
동아대/2 학기/ 선행대/수학/source')
m x n 2 2
1 2 -1
3 4 0
[[-2 -4 -1]
 [-3 0 -2]]

In [100]: runfile('C:/Users/WD/OneDrive/2024/ 동아대/2 학기/
선행대/수학/source/ot(1).py', wdir='C:/Users/WD/OneDrive/2024/
동아대/2 학기/ 선행대/수학/source')
m x n 2 2
1 2 -1
3 4 0
t: [-3 -6 -1]
[[ 1 2 -1]
 [-3 0 -2]]

In [101]: runfile('C:/Users/WD/OneDrive/2024/ 동아대/2 학기/
선행대/수학/source/ot(1).py', wdir='C:/Users/WD/OneDrive/2024/
동아대/2 학기/ 선행대/수학/source')
m x n 2 2
```

# 데이터의 중요성

- 데이터 시대(the age of data)
  - ChatGPT, 제미나이, DALL·E, 소라(Sora)와 같은 생성형 AI 기술의 등장
  - 기존 기술을 뛰어넘는 생성형 AI 기술이 수많은 데이터를 학습한 결과
  - 컴퓨터가 개발되어 사용되기 시작한 이래 데이터는 꾸준히 주목을 받아 옴
  - '데이터 사이언스(data science)'라는 학문 분야가 새로 생길 정도로 데이터는 중요성을 인정받고 있음

# 데이터의 중요성

- 데이터가 중요한 구체적 이유
  - 의사결정
    - 데이터는 기업이 더 나은 결정을 내리고, 정부가 효과적인 공공 정책을 만들고, 개인들이 자신의 돈과 건강에 대해 정보를 바탕으로 의사결정을 내리는 데 도움
  - 문제 해결
    - 데이터는 현실에 존재하는 문제를 식별하고 정확한 이론을 개발하는 데 도움
    - Ex) 서울시는 데이터 분석을 기반으로 심야 교통 버스 노선 결정  
→ 시민들의 만족도 향상

# 데이터의 중요성

- 현실 이해
  - 검색 사이트들은 특정 사용자의 검색어를 분석하여 그 사용자가 관심 있어 할 만한 광고를 노출해 구매를 유도
- 기술 개발
  - 데이터는 머신러닝, 딥러닝 기술 개발을 위한 자원
  - 주가 예측, 질병 진단, 이미지 분석, 자동번역 등의 기술 개발을 위해 방대한 양의 데이터의 학습이 필요
  - 데이터를 기반으로 하는 디지털 전환(DX; Digital Transformation)은 오늘날 기업 생존을 위한 필수 요소로 인식

# 데이터의 중요성

- 새로운 비즈니스의 창출

- 데이터를 수집, 가공, 분석하는 분야에서 일하는 수많은 스타트업들

ex) 어떤 국내 스타트업은 빅데이터 상권분석 플랫폼을 운영한다. 매월 70만 개의 전국 상권 매출 및 주변 인구통계 데이터를 꾸준히 쌓아 현재까지 누적 1억 1000개 이상의 상권 데이터를 보유

# 빅데이터의 특징

- 크기(volume)
  - 빅데이터는 일반적으로 수십 테라바이트(terabyte) 혹은 수십 페타바이트(petabyte) 이상이 빅데이터의 범위에 해당
  - 1페타바이트는 6기가바이트 DVD 영화를 17만 4천 편 담을 수 있는 정도의 용량
- 속도(velocity)
  - 빅데이터는 빠른 증가 속도를 갖는 경우가 많다.
  - Ex) 지하철 승하차 정보, SNS상에서 오가는 메시지들
  - 데이터 중에는 자율주행 자동차와 같이 실시간 처리가 필요한 경우가 있는데 데이터의 빠른 증가는 이러한 실시간 처리를 어렵게 할 수 있다.



# 빅데이터의 특징

- 다양성(variety)
  - 데이터는 정형화 정도에 따라 정형(structured), 반정형(semi-Structured), 비정형(unstructured)으로 구분
  - 정형 데이터는 고정된 필드에 저장되는 데이터를 의미하며, 일정한 형식을 갖추고 있다. Ex) 엑셀 파일에 저장하는 테이블 형태의 데이터
  - 반정형 데이터는 고정된 필드로 저장되지는 않지만, XML이나 HTML같이 메타데이터나 구조 정보 등을 포함하며, 일정한 구조는 없으나 구조를 파악할 수 있다.
  - 비정형 데이터는 고정된 필드에 저장되지 않는 데이터를 의미
  - 사진, 동영상, 메신저로 주고받은 대화 내용, 스마트폰에 기록되는 위치 정보, 통화 내용 등

# 빅데이터의 활용 사례

## ■ 해외 사례: 타깃의 맞춤형 광고(계속)

- 타깃의 빅데이터 전문가들은 고객의 25가지 구매 행태를 분석하면 여성의 임신과 출산을 상당히 정확하게 예측할 수 있다는 사실을 확인했다. 예를 들어, 향이 나는 로션을 사용하던 여성이 무향의 로션으로 바꾸거나, 평소 사지 않던 미네랄 영양제를 갑자기 구매하는 경우다.

타깃은 이러한 패턴을 고객 데이터베이스에 적용했고, 전국적으로 수만 명의 임신 추정 여성들을 선별하여 관련 할인 쿠폰을 발송했다. 고객의 구매 행동을 유도할 수 있기 때문이다.

뉴욕 타임스가 보도한 이 사례는 빅데이터 분석의 강력한 영향력을 여실히 보여준다. 대형 마트는 고객이 가족에게도 알리지 않은 비밀을 파악하고, 앞으로 무엇이 필요할지 정확하게 예측해냈다.



# 빅데이터 분석 도구

## ■ 개요

- 효율적으로 데이터를 수집, 정리하고 분석하기 위해 중요한 요소가 바로 분석 도구
- 전통적으로 사용된 분석 도구는 바로 SAS, SPSS, Matlab과 같은 통계 도구
- 매우 고가이고 사용이 복잡
- 최근에는 프로그래밍 언어를 통해 분석하는 추세가 강하다
- 무료로 사용할 수 있는 오픈소스 도구들이 등장하여 인기를 얻고 있다. 그 중 대표적인 것이 R과 파이썬(Python) 언어



(a) R



(b) 파이썬

그림 1-8 R과 파이썬

# 데이터 분석 도구

## ■ 파이썬

- 1991년 네덜란드계 소프트웨어 엔지니어인 귀도 반 로섬이 발표한 고급 프로그래밍 언어
- '인터프리터를 사용하는 객체지향 언어'이자 '대화형 언어'
- 데이터를 처리하고 분석하는 데 필요한 numpy, pandas, scikit-learn과 같은 다양한 라이브러리(패키지)들을 제공
- 크로스 플랫폼 언어로, 윈도우, 맥, 리눅스 등 다양한 운영 체제에서 사용 가능

# 데이터 분석에 필요한 패키지 설치

표 1-2 데이터 분석에 유용한 라이브러리

라이브러리명	지원 기능
numpy	데이터 저장을 위한 배열과 다양한 수학 함수를 제공한다.
pandas	데이터 저장을 위한 데이터프레임과 데이터 가공을 위한 다양한 함수를 제공한다.
matplotlib	기본적인 그래프 작성 기능을 제공한다.
seaborn	matplotlib을 기반으로 한 고급 시각화 기능을 제공한다.
scikit-learn	다양한 머신러닝 분석 도구 및 예제 데이터셋을 제공한다.

- 설치된 라이브러리 목록은 `pip list` 명령으로 확인가능