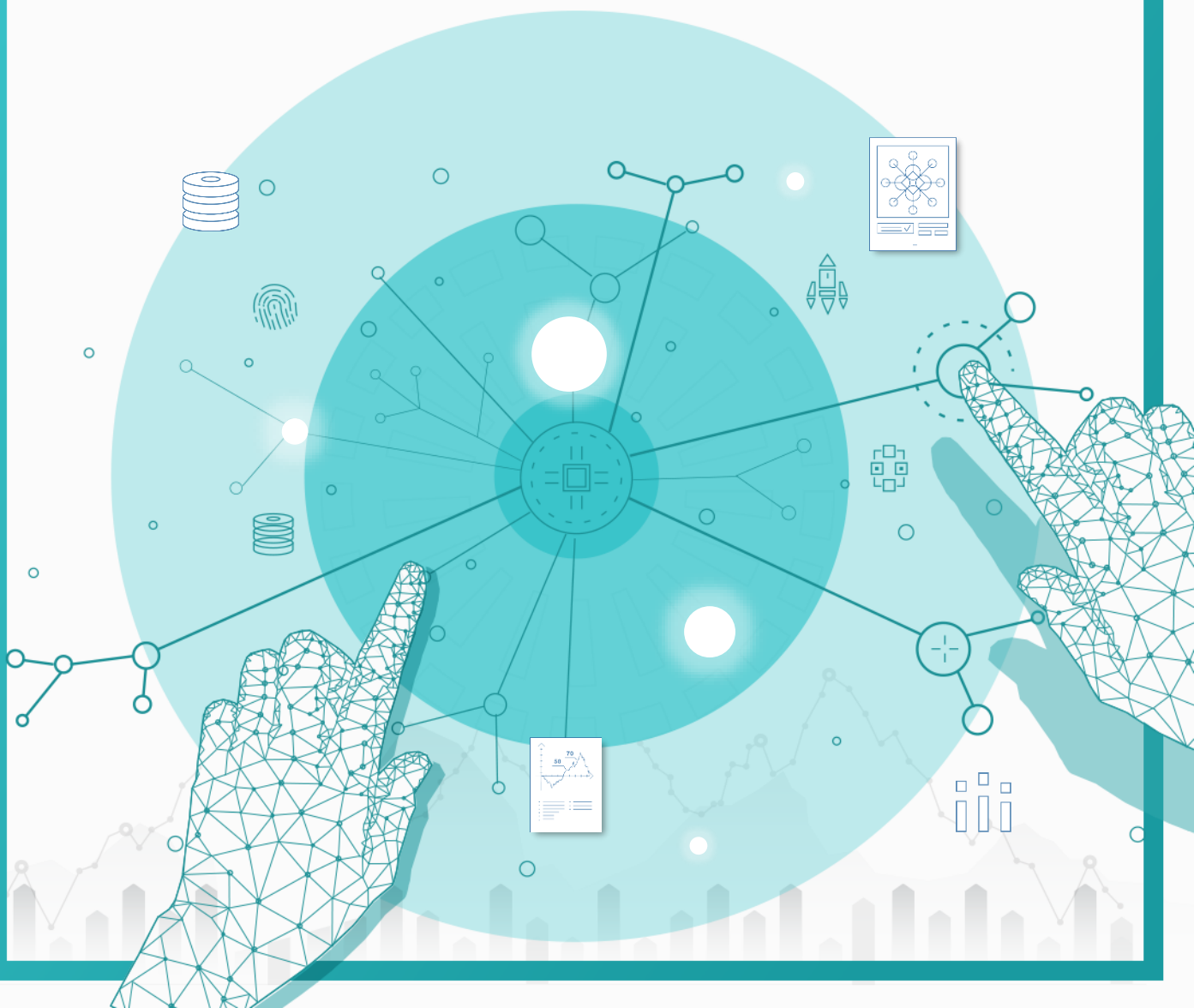




한국기술교육대학교  
온라인평생교육원

# 「파이썬 라이브러리로 하는 데이터 분석과 시각화」

파이썬 시작하기



# 파이썬 시작하기

## 학습 목표

1. 데이터 분석 및 시각화에 대한 필요성과 방법을 설명할 수 있다.
2. 파이썬을 설치하고, 직접 실행할 수 있다.
3. 데이터 분석의 개념을 설명할 수 있다.
4. 데이터 시각화의 개념을 설명할 수 있다.

## 학습 내용

1. 데이터 분석 및 시각화
2. 파이썬 개발 환경 구축
3. 데이터 분석의 개념과 이해
4. 데이터 시각화의 개념과 이해

### 1. 데이터 분석 및 시각화

#### 1) 데이터 분석 및 시각화의 필요성

##### (1) 데이터 수집과 분석의 필요성

- 빅데이터, 21세기의 석유?!
  - 잘 쓰이면 석유지만, 못 쓰다면 용량만 큰 쓰레기
- 데이터가 없는 곳이 없음
  - 인공지능, 사물인터넷, 스마트시티, 스마트 팩토리 등
- 4차 산업혁명의 시작과 끝은 데이터

➡ 다가올 미래를 주도하기 위해서 다양한 데이터를 수집, 분석하고 활용하는 능력이 필요!

##### (2) 실생활 속 데이터 분석, 시각화 예시

- 카드사별 AI 소비 컨설팅 서비스
  - 수많은 사용자의 카드 사용 내역(데이터)을 분석하여 패턴을 찾고 의미 있는 결과를 찾아냄
  - 고객별 맞춤형 서비스를 제공함

### 1. 데이터 분석 및 시각화

#### 1) 데이터 분석 및 시각화의 필요성

##### (1) 데이터 수집과 분석의 필요성

###### ▪ 맞춤형 선거 전략

(예) 2008년 미국 대선 당시 오바마 후보가 사용한 유권자 데이터 분석

- 사람들의 기본 인적 사항은 물론, 과거 투표 여부, SNS, 마시는 음료, 구독 잡지 등 수많은 데이터를 수집 및 분석, 시각화 함
- 유권자 맞춤형 선거 전략을 세워 흑인 대통령 당선에 결정적인 역할을 함

###### ▪ 의료, 유통, 제조 등 다양한 분야에서의 활용

- 데이터를 분석 및 시각화하여 유의미한 통계, 패턴 등을 도출하여 마케팅, 평가, 시스템 등에 적용함
- 예) 월마트의 맥주와 기저귀  
고객들의 구매 상품 리스트를 분석한 결과, 젊은 아버지가 퇴근길에 기저귀를 사면서 맥주도 함께 살 확률이 높다고 나옴



맥주 옆에 기저귀를 함께 배치하는 전략을 세워 매출 상승에 기여!

### 1. 데이터 분석 및 시각화

#### 2) 데이터 분석 및 시각화 방법

##### (1) 잘 알려진 데이터 분석 및 시각화 툴 활용하기

- 엑셀, 구글 애널리틱스, 구글 데이터 스튜디오, 파인리포트 등
  - 장점: 큰 진입장벽 없이 비교적 쉽게 접근, 사용할 수 있음
  - 단점: 커스터마이징이 어렵고, 일부 유료인 경우도 있음



Data Studio



FineReport

##### (2) 파이썬, 자바스크립트 등 프로그래밍 활용하기

- 파이썬의 Pandas, Matplotlib, 자바스크립트의 D3.js
  - 장점: 원하는 형태로 구현할 수 있으며, 확장성이 높음
  - 단점: 프로그래밍을 통한 구현으로, 상대적으로 구현이 어렵고 진입장벽이 높음



matplotlib



## 2. 파이썬 개발 환경 구축

### 1) 파이썬의 정의와 특징

#### (1) 파이썬이란?

- 사전적인 의미
  - 독이 없는 뱀으로 먹이를 몸으로 감아 압사시키는 큰 뱀
  - 그리스 신화에 나오는 악마
- 프로그래밍 언어
  - 1991년 프로그래머인 귀도 반 로섬(Guido van Rossum)이 발표한 고급 프로그래밍 언어
  - 플랫폼에 독립적이며 인터프리터식, 객체 지향적, 동적 타이핑 대화형 언어
  - 파이썬이라는 이름은 귀도가 좋아하는 코미디 <Monty Python's Flying Circus>에서 따옴

#### (2) 파이썬의 특징

- 초보자부터 전문가까지 사용자층이 다양함
- 쉽고, 단순하고, 직관적인 언어
- 스크립트 언어로 한 줄씩 실행하는 인터프리터(대화형) 언어
- 다양한 플랫폼에서 사용할 수 있음
- 다양한 라이브러리를 활용할 수 있음

## 2. 파이썬 개발 환경 구축

### 2) 파이썬 개발 환경 구축 방법

#### (1) 파이썬 설치 방법

- 파이썬 공식 홈페이지에서 다운로드
  - <https://www.python.org>
  - 가장 기본적인 파이썬만 설치
  - 데이터 분석, 시각화에 필요한 것은 직접 설치해야 함
- 아나콘다(Anaconda) 설치
  - <https://www.anaconda.com>
  - 파이썬과 파이썬의 패키지 관리, 라이브러리 관리, 다양한 수학, 과학 계산 등을 쉽고 편리하게 할 수 있는 파이썬 배포판

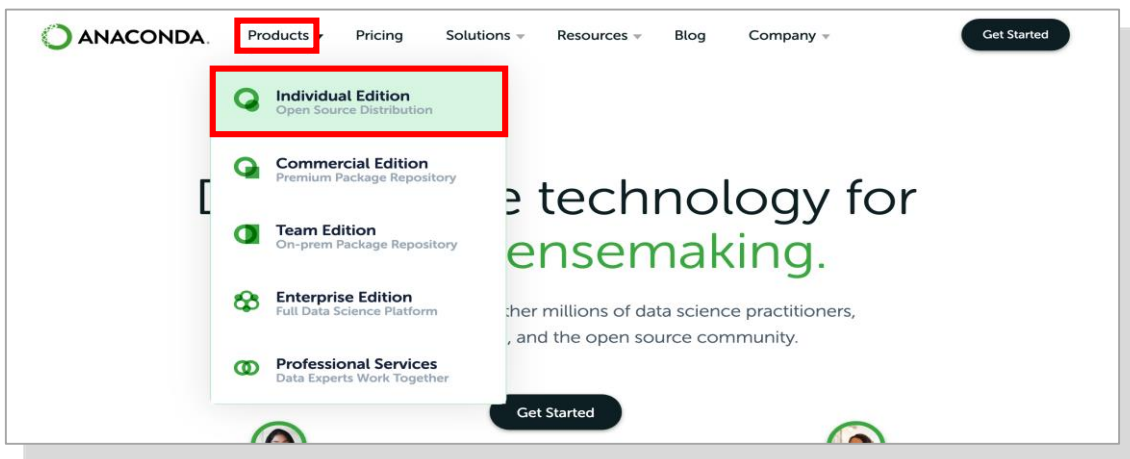
➡ 최근에는 거의 정석처럼 파이썬을 Anaconda로 설치 및 활용함

## 2. 파이썬 개발 환경 구축

### 2) 파이썬 개발 환경 구축 방법

#### (2) 아나콘다(Anaconda) 설치하기

- 아나콘다 홈페이지(<https://www.anaconda.com>)에 접속
  - Products → Individual Edition 클릭



- 페이지 하단에서 PC 환경에 맞는 버전 다운로드





## 2. 파이썬 개발 환경 구축

### 2) 파이썬 개발 환경 구축 방법

#### (2) 아나콘다(Anaconda) 설치하기

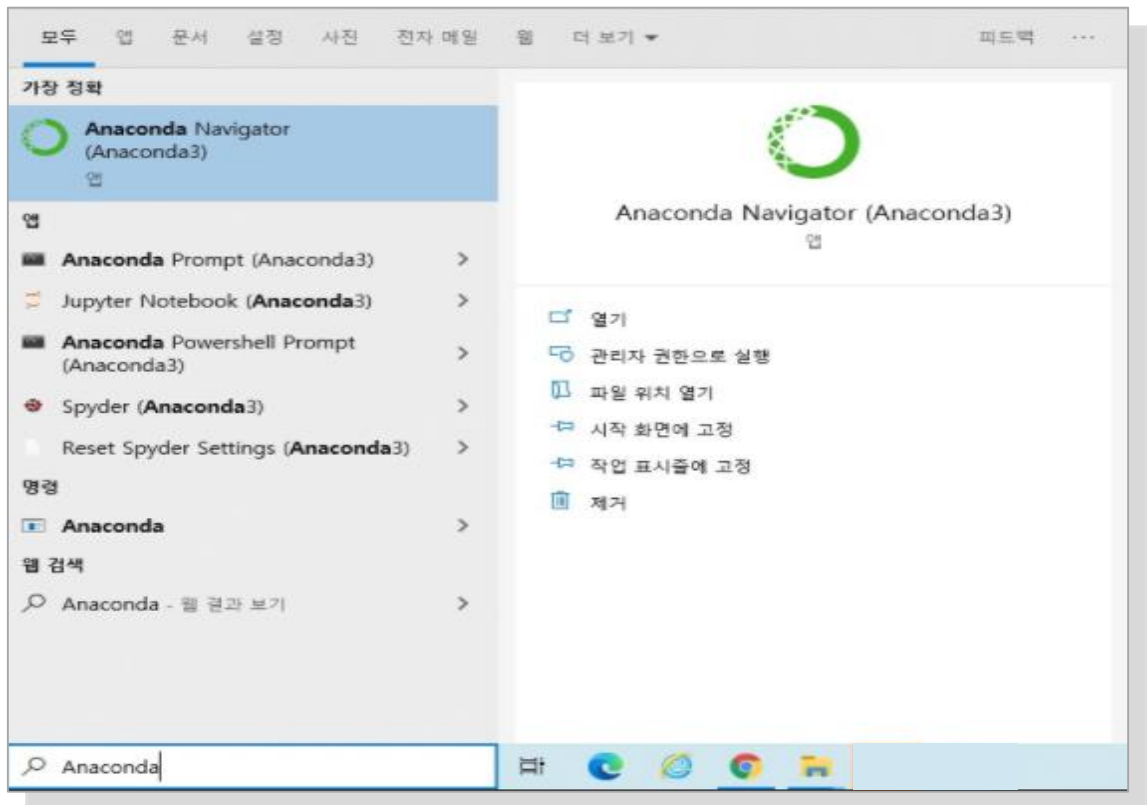
##### ■ 설치 시 주의사항

- Windows PC의 경우, 설치 경로에 한글이 없어야 함

(예) C:\Users\WW파이썬\Documents → X  
C:\Users\WWpython\Documents → O

##### ■ 파이썬 실행 방법(1) (Windows 기준)

- Anaconda Navigator 실행

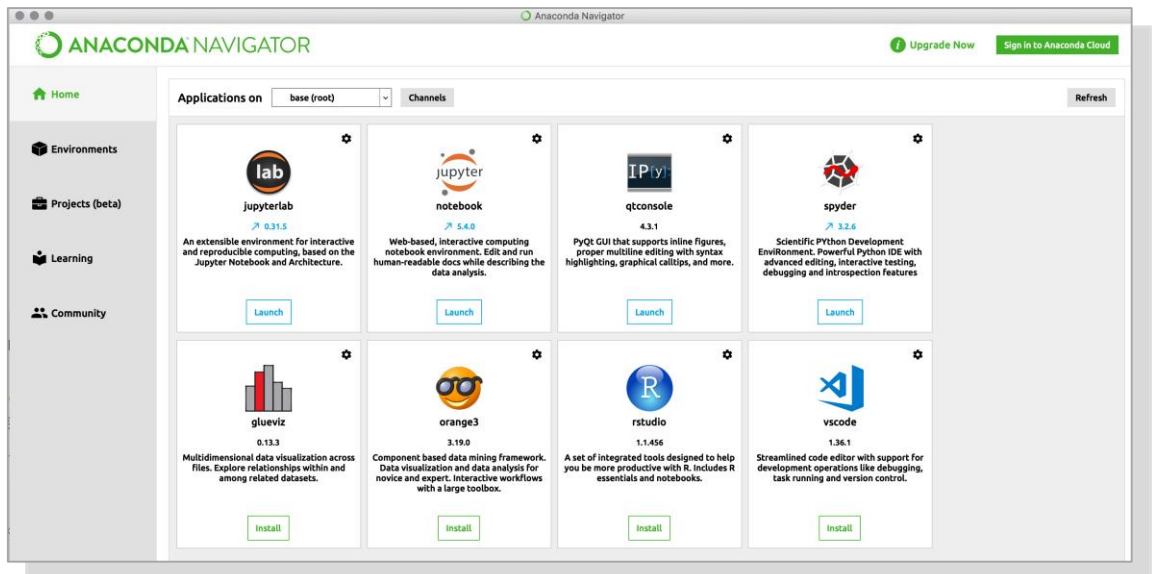


## 2. 파이썬 개발 환경 구축

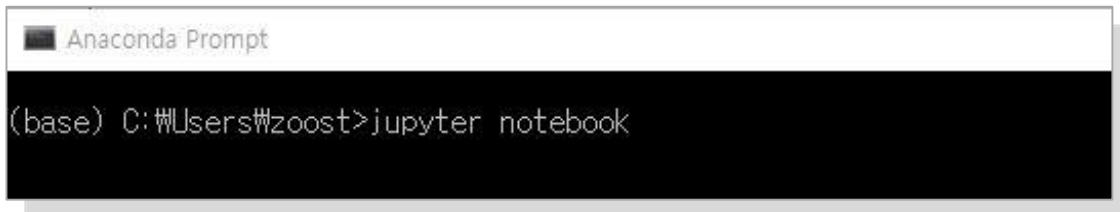
### 2) 파이썬 개발 환경 구축 방법

#### (2) 아나콘다(Anaconda) 설치하기

- 파이썬 실행 방법(1) (Windows 기준)
  - Jupyter Notebook 실행



- 파이썬 실행 방법(2) (Windows 기준)
  - 프롬프트 창에서 실행(Anaconda Prompt 실행)
  - 'jupyter notebook' 명령어 입력 후 엔터

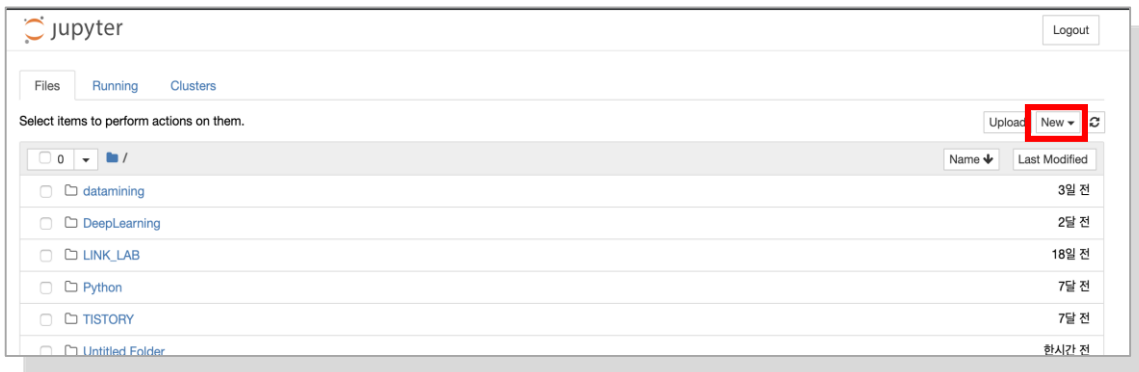


## 2. 파이썬 개발 환경 구축

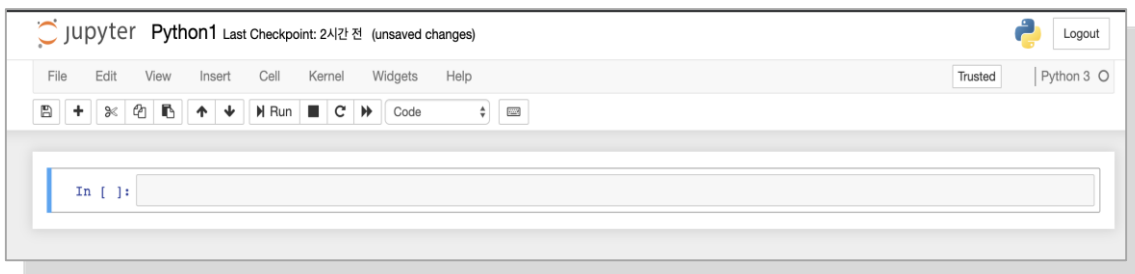
### 2) 파이썬 개발 환경 구축 방법

#### (2) 아나콘다(Anaconda) 설치하기

- Jupyter Notebook 사용하기
  - 우측 상단 New → Python3 클릭



- 코드 입력 후 Run 버튼을 사용하여 코드 실행



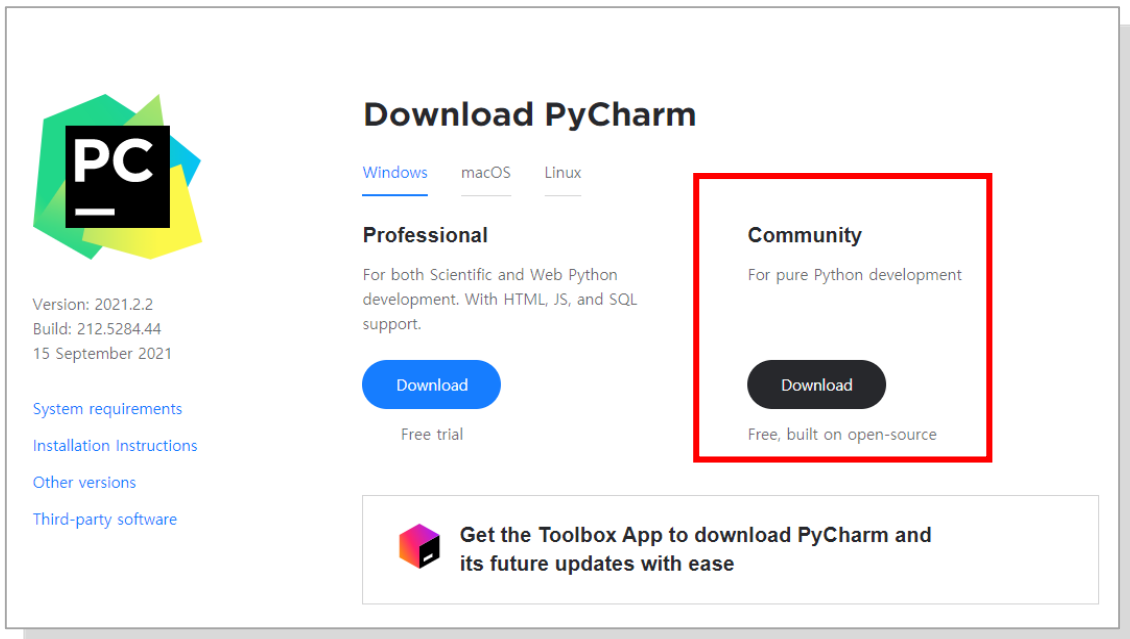
- 다양한 단축키 활용 가능
  - a, b 셀 추가
  - dd 셀 삭제
  - Shift + Enter 실행 등

## 2. 파이썬 개발 환경 구축

### 2) 파이썬 개발 환경 구축 방법

#### (3) 그 외 파이썬을 설치·활용하는 방법들

- PyCharm 설치
  - <https://www.jetbrains.com/pycharm/download>
  - 무료 버전 다운로드 및 설치(Community 버전)



- 무료 코드 에디터 설치
  - Sublime Text3 설치(<https://www.sublimetext.com/3>)
  - 아톰, 브라켓, VS Code 등
- 기타 개발 환경툴 사용
  - 비주얼 스튜디오, 이클립스, 캐노피 등

### 3. 데이터 분석의 개념과 이해

#### 1) 데이터 개념

##### (1) 데이터란?

- 문자, 숫자, 소리, 그림 등의 형태로 된 의미 단위로 보통 연구나 조사 등의 바탕이 되는 자료를 말하며, 이 자료를 의미 있게 정리하면 정보가 됨
- 흔히 우리가 알고 있는 데이터는 컴퓨터가 처리할 수 있는 자료를 말하며, 가공되지 않은 형태의 기초적인 자료를 말함

##### (2) 데이터의 구분

- 질적 자료(정성적 데이터)
  - 수치로 측정 불가능한 자료  
(예) 성별, 혈액형, 전화번호 등
- 양적 자료(정량적 데이터)
  - 수치로 측정 가능한 자료  
(예) 온도, 가격, 주가 등

### 3. 데이터 분석의 개념과 이해

#### 2) 데이터 분석의 개념

##### (1) 데이터 분석이란?

- 유용한 정보를 발굴하고, 결론 내용을 알리며 의사결정을 지원하는 것을 목표로 데이터를 정리, 변환, 모델링하는 과정
- 데이터에서 유용한 정보를 추출하고, 데이터 분석을 통해 결정을 내리거나 다양한 인사이트를 도출하기 위한 것

##### (2) 데이터 분석에 필요한 절차

- |     |                 |
|-----|-----------------|
| 1단계 | 데이터 분석 요구 사항 정리 |
| 2단계 | 데이터 수집 및 전처리    |
| 3단계 | 데이터 분석          |
| 4단계 | 데이터 평가 및 해석     |
| 5단계 | 데이터 시각화         |

- ➡ 데이터 분석'이란 용어는 데이터 수집, 가공, 분석, 시각화 등 다양한 뜻을 함께 포함하고 있음
- ➡ 데이터 분석은 데이터를 활용하기 위한 일종의 재생산 과정

### 3. 데이터 분석의 개념과 이해

#### 2) 데이터 분석의 개념

##### (3) 데이터 분석 기법

- 데이터 분석 기법은 다양함
- 우리가 아는 통계 또한 데이터 분석 방법 중 하나임

##### (4) 분석 목적에 따른 데이터 분석 분류

- 설명(요약, 통계)
  - 가장 기본적인 방법
  - 단순한 데이터의 계산 결과로 다양한 지표 등이 해당함
- 탐색(트렌드, 패턴)
  - 주로 시각화를 통해 데이터 사이에서 의미 있는 사실을 발견하고 패턴 등을 파악함
- 추론(샘플링)
  - 샘플과 모집단 간의 관계를 분석함
  - 한정된 데이터에서 문제를 해결할 때 주로 활용함
- 예측(기계학습 등)
  - 다양한 통계적 기법을 활용하여 데이터를 통해 아직 발생하지 않은 것들을 예측할 수 있음

### 3. 데이터 분석의 개념과 이해

#### 2) 데이터 분석의 개념

##### (5) 데이터 분석을 위한 고려사항

- 데이터 선정 시
  - 수집 가능한 데이터인가?
  - 데이터가 정확한가?
  - 데이터를 수집, 분석하기에 난이도가 적절한가?
- 데이터 수집 시
  - 분석할 데이터의 유형과 특징은 무엇인가?
  - 어떤 방법으로 데이터를 수집할 것인가?
- 데이터 전처리 및 저장 시
  - 수집한 데이터에서 분석에 필요한 주요 정보를 추려내고, 불필요하거나 이상한 데이터를 처리하는 일종의 클리닝 작업(전처리) 필요
  - 데이터를 쉽게 불러오고 분석할 수 있도록 저장을 위한 수단 필요



### 3. 데이터 분석의 개념과 이해

#### 3) 데이터 분석 예시

- 복잡하게 나열된 데이터에서 의미 있는 분석 결과를 찾아내 활용할 수 있음

2020년 12월 16일	21,352.20	19,434.70	21,525.30	19,299.70	199.81K	9.87%
2020년 12월 15일	19,434.90	19,273.90	19,556.30	19,076.20	98.39K	0.84%
2020년 12월 14일	19,273.80	19,176.40	19,346.50	19,007.00	67.11K	0.51%
2020년 12월 13일	19,176.80	18,808.10	19,403.60	18,716.80	82.45K	1.96%
2020년 12월 12일	18,808.90	18,026.70	18,936.20	18,025.10	71.37K	4.36%
2020년 12월 11일	18,023.60	18,250.50	18,288.30	17,600.10	111.53K	-1.23%
2020년 12월 10일	18,247.20	18,546.10	18,551.80	17,926.80	85.31K	-1.61%
2020년 12월 9일	18,546.00	18,322.40	18,637.90	17,658.80	129.72K	1.20%
2020년 12월 8일	18,326.60	19,170.80	19,288.60	18,229.20	101.41K	-4.40%
2020년 12월 7일	19,170.70	19,368.40	19,424.60	18,908.80	61.98K	-1.08%
2020년 12월 6일	19,379.90	19,146.00	19,417.10	18,873.00	54.56K	1.22%
2020년 12월 5일	19,146.50	18,657.50	19,172.70	18,507.10	59.44K	2.62%
2020년 12월 4일	18,658.10	19,431.10	19,529.50	18,610.80	110.63K	-3.99%
2020년 12월 3일	19,433.30	19,219.80	19,596.60	18,883.70	99.97K	1.12%
2020년 12월 2일	19,218.80	18,767.90	19,335.50	18,346.70	91.92K	2.39%
2020년 12월 1일	18,770.70	19,697.80	19,897.40	18,257.00	208.72K	-4.71%
2020년 11월 30일	19,698.10	18,186.00	19,831.20	18,186.00	185.59K	8.32%
2020년 11월 29일	18,185.50	17,729.70	18,336.30	17,530.40	79.12K	2.57%
2020년 11월 28일	17,730.70	17,142.60	17,874.40	16,874.40	95.20K	3.52%
2020년 11월 27일	17,127.10	17,160.90	17,450.30	16,481.60	138.80K	-0.20%
2020년 11월 26일	17,162.00	18,721.90	18,894.90	16,235.20	312.07K	-8.34%
2020년 11월 25일	18,722.00	19,151.50	19,486.70	18,527.70	146.12K	2.24%

- ➡ 시계열 데이터 A와 B가 있을 때, 어느 하나 일치하는 값이 없지만 상관관계가 높다면?

(예) A 주식 가격이 오를 때, B 주식도 오른다!

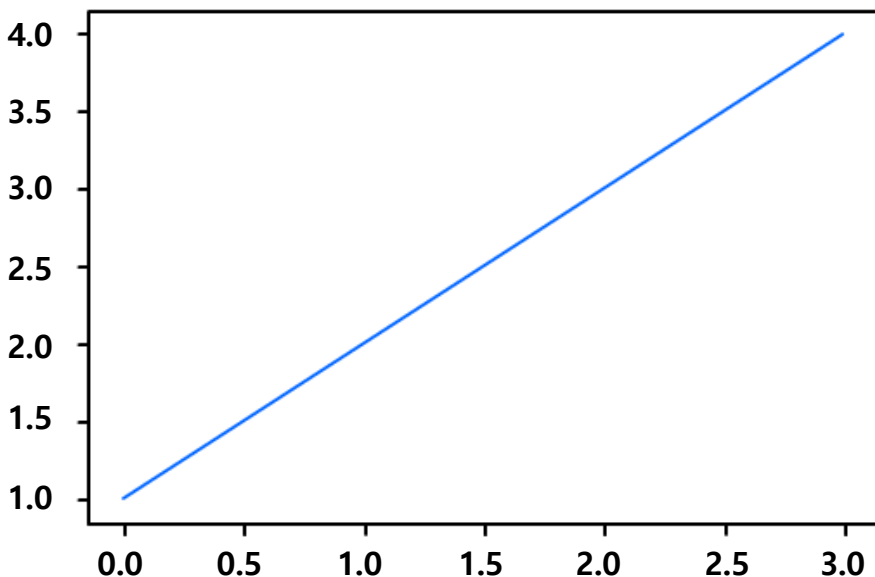
### 4. 데이터 시각화의 개념과 이해

#### 1) 데이터 시각화의 개념 및 필요성

##### (1) 데이터 시각화란?

- 데이터의 분석 결과를 쉽게 이해하고 판단할 수 있도록 데이터를 시각적으로 표현하고 전달하기 위한 과정
- 21세기 디지털 시대의 사람들은 수많은 미디어, SNS 등 사실상 무한정 데이터에 노출되어 있음
- 모두가 아는 정보는 정보가 아니듯, 단순히 데이터를 많이 보유하고 있다고 해서 큰 의미가 있지는 않음

➡ 즉, 데이터를 잘 활용하는 대표적인 방법이 바로 시각화!



### 4. 데이터 시각화의 개념과 이해

#### 1) 데이터 시각화의 개념 및 필요성

##### (2) 데이터 시각화의 필요성

- 많은 양의 데이터를 한눈에 볼 수 있음
  - 시각적 요소를 활용하여 데이터를 요약함
- 데이터 분석에 대한 전문 지식이 없더라도 누구나 해당 데이터를 인지하고 활용할 수 있음
  - 사람이 감각기관을 통해 획득하는 정보의 80% 이상이 시각을 통해 얻어짐
  - 시각적 입력은 다른 어떤 방법보다도 빠르고 쉽게 이해할 수 있음
  - 시각화 형태, 요소, 위치, 색 등을 기반으로 패턴이나 다양한 인사이트 도출 가능
- 단순한 데이터의 요약, 통계 분석 결과보다 정확한 데이터 분석 결과를 도출할 수 있음
  - 데이터의 패턴이나 추세 등 요약 통계, 표만 보서는 알 수 없던 결과들을 알 수 있음
- 다양한 분야에 활용할 수 있고, 데이터 분석 결과를 공유하기 편리함
  - 대시보드, 마케팅, 각종 지표 등 다양한 분야에 적용하여 활용할 수 있음

### 4. 데이터 시각화의 개념과 이해

#### 2) 파이썬에서의 데이터 분석 및 시각화

- 수많은 라이브러리를 제공하는 파이썬은 데이터 분석과 시각화에 필요한 여러 가지 라이브러리 또한 제공
  - 대표적인 라이브러리
    - 데이터 분석: Numpy, Pandas
    - 시각화: Matplotlib
  - 라이브러리 하나로 책이 출판될 만큼 강력하고 다양한 기능들로 편리하고 쉽게 데이터를 분석, 시각화할 수 있음
- ➡ 데이터를 잘 활용한다는 것은 수많은 데이터 속에서 원하는 정보를 찾아내고, 분석하고, 분석한 결과를 시각적으로 표현하는 것을 말함

## 21