

파이썬 데이터 분석 실무과정

이 선 순

4차 산업혁명과 데이터 과학

IT 기술의 발전

■ IT 기술의 발전과 일상생활의 변화

• 사례1. 대학생 A는 늦잠을 자는 바람에 학교에 늦어 카카오톡 어플로 택시를 부르고 외출 준비를 한다. 예약 시간에 택시를 타고 택시 안에서 페이스북에 댓글을 달면서 친구의 SNS에도 방문한다. 학교에 도착하여 학생 카드로 출석 체크를 하고, 리포트는 웹 디스크에 올린다. 점심시간에는 식당에 있는 키오스크를 통해 돈가스를 주문하며, 공강 시간에 는 원하던 청바지의 사용 후기를 보면서 최저가 검색을 한 후 결제한다.

• 사례2. 직장인 B는 저녁 모임을 위해 맛집을 검색하여 카카오톡 단체 방에 장소와 시간을 공지한다. 모임에 참석하기 위해 지하철역으로 들어가 전광판을 보고 다음 전철이 5분 뒤에 도착한다는 것을 확인한다. 목적지 근처 역에 도착하여 길 찾기 어플을 확인하며 약속 장소로 향한다. 모임 장소에 도착하여 친구들과 수다를 떨며 주문한 음식과 단체 사진을 인스타그램에 올린다. 모임이 끝나고 집으로 돌아와 뉴스 기사를 검색하고 유튜브를 보다가 잠이 든다

IT 기술의 발전

■ 일상이 된 스마트폰



그림1. 스마트폰

IT 기술의 발전

■ 일상이 된 스마트폰

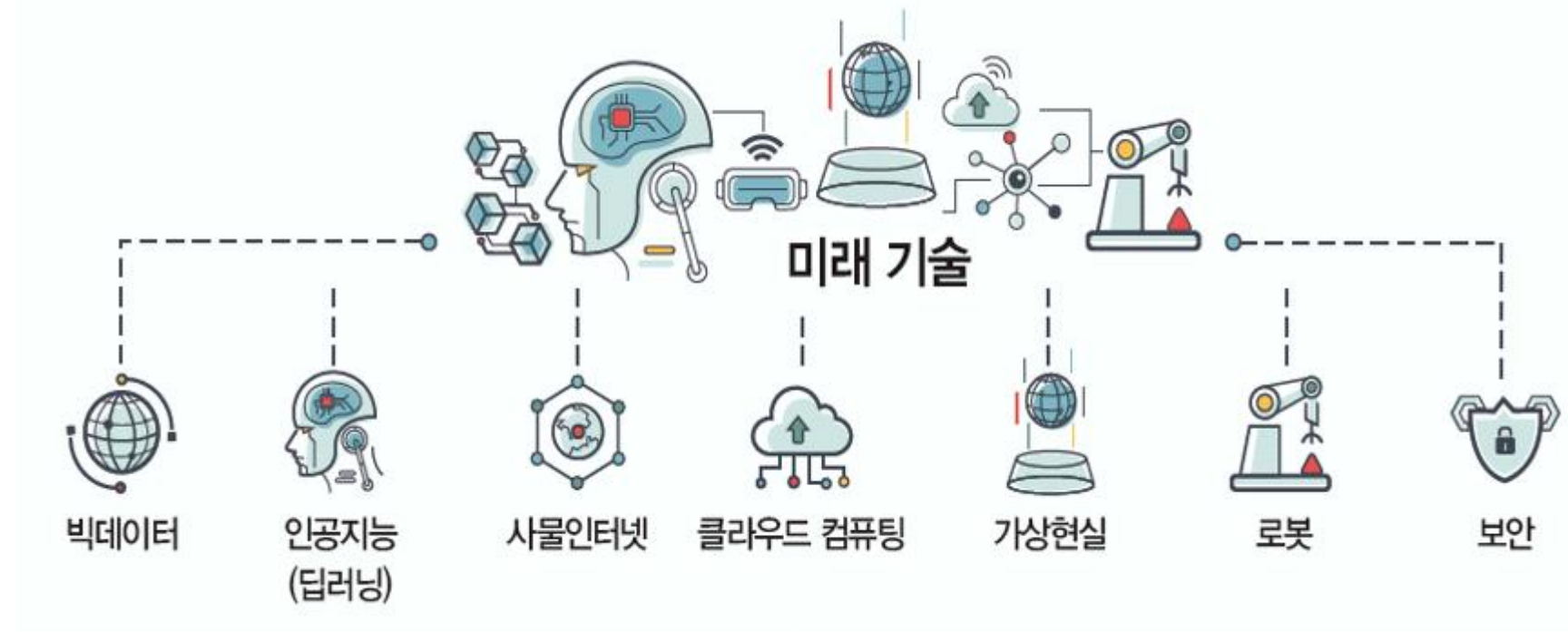


그림2. 미래기술

4차 산업혁명의 이해

- 18세기 : 증기기관을 기반으로 한 기계화 혁명의 1차 산업혁명 시대
- 19~20세기 초반 : 전기를 사용한 대량생산혁명의 2차 산업혁명 시대
- 20세기 후반 : 컴퓨터와 인터넷 보급, 지식정보혁명의 3차 산업혁명 시대
- 21세기 현재 : 4차 산업혁명 시대, 초연결, 초지능, 초융합이 특징

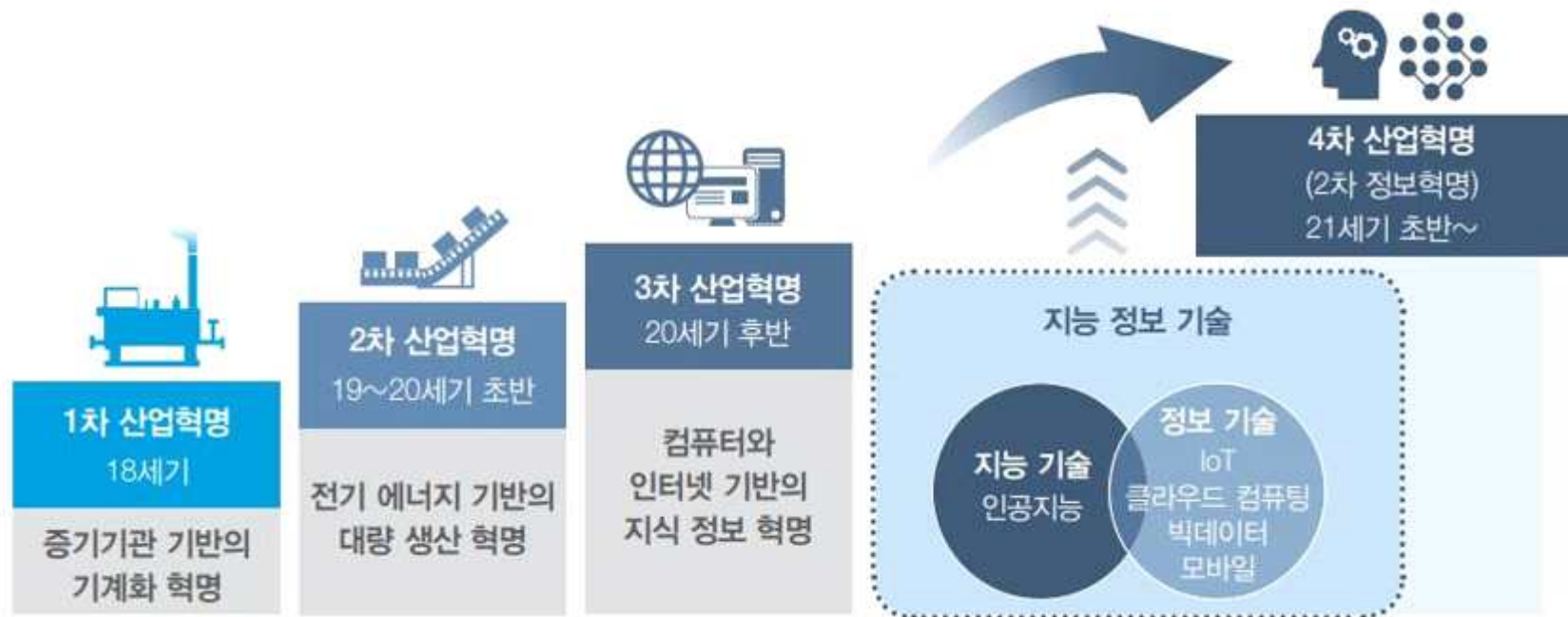


그림3. 산업혁명의 흐름(출처 : 미래창조과학부 블로그)

4차 산업혁명의 이해

■ 4차 산업혁명의 정의

- 융합 기술을 기반으로 하는 초연결사회(hyper-connected society)

4차 산업혁명

융합 기술을 기반으로 하는 초연결 사회

4차 산업혁명의 이해

■ 초연결(hyper-connectivity)

▶ 초연결(hyper-connectivity)

- 사물인터넷의 진화와 디지털화로 인해 사물과 공간, 인터넷의 상호의존성이 증폭, 제품과 서비스의 연결성이 무한 확장
- 사물인터넷, 클라우드컴퓨팅 등, 인간과 인간, 인간과 사물, 사물과 사물간의 연결 확대
- 초연결 사회: 세상의 모든 물건이 인터넷으로 연결되는 세상
- 사물인터넷과 5세대 통신(5G)은 초연결사회를 실현시킨 대표적 기술

▶ 사물인터넷(Internet of Things, IoT)

- IoT는 '언제나, 어디서나, 어느 것과도 연결될 수 있는 새로운 통신 환경, 사물들(things)이 연결된 세상
- 사물인터넷은 전자제품, 교통 전광판, 스마트 홈, 스마트워치등과 같이 다양한 기계들이 인터넷에 연결되어 부가가치를 생성하는 기술

4차 산업혁명의 이해

▶ 사물인터넷(Internet of Things, IoT)

- 사물인터넷은 전자제품, 교통 전광판, 스마트 홈, 스마트워치등과 같이 다양한 기계들이 인터넷에 연결되어 부가가치를 생성하는 기술



지하철 전광판



스마트 홈



무인 슈퍼마켓 아마존고의 매장 전경



스마트워치



전기와 IT기술을 접목한 지능형 전력망 스마트 그리드



사물인터넷 기술을 이용한 스마트 주차장

4차 산업혁명의 이해

▶ 사물인터넷(Internet of Things, IoT)

- 스마트시티 : 스마트 주차장, 스마트 자동차, 스마트 그리드, 스마트 건물과 같이 모든 물건들이 유기적으로 연결되는 도시



4차 산업혁명의 이해

▶ 사물인터넷(Internet of Things, IoT)

- 사람, 업무, 데이터까지 모든 것이 연결되어 상호 통신하는 만물인터넷(loE)으로 발전할 전망



그림4. 네트워크에 연결된 사물과 세계 인구수 비교(출처:CISCO,UBSG,HP)

4차 산업혁명의 이해

■ 초연결(hyper-connectivity)

▶ 5세대 통신 5G

- 5G의 특징은 초고속, 초연결, 초저지연으로 요약 가능

표1. 5G의 특징

특징	설명
초고속	초광대역 무선통신(eMBB)enhanced Mobile BroadBand • 유선과 무선의 차이가 없는 대용량 및 고속 의 데이터 이용 환경 제공 • 4G에 비해 최대 20배 더 빠른 20Gbps까지 구현 가능(일상적으로는 100Mbps 보장 목표) • 모바일로 8K 콘텐츠 송수신 가능
초연결	대규모 사물통신(mMTC)massive Machine Type Communication • 산업 또는 일반인에게 IoT 사용 환경 제공 • 현재보다 최대 500배 더 많은 기기와 고밀집 연결 가능 • 고에너지 효율 • 스마트폰의 인터넷, PC의 인터넷을 넘어 진정한 IoT 가능
초저지연	고신뢰/초저지연 통신(uRLLC)ultra-Reliable Low-Latency Communication • 최대 1ms까지의 낮은 지연성 과 고안정성 을 목표로 데이터 통신 서비스의 품질(QoS)을 제공

4차 산업혁명의 이해

■ 초지능

• 슈퍼 인텔리전스

- 인공지능의 지능이 인간을 넘어서는 특이점을 통과하여 거의 모든 영역에서 인간의 인지 능력을 크게 능가 하는 경우

• 하이퍼 인텔리전스

- 인간의 지능과 인공지능이 협력하여 더 스마트한 서비스를 제공하는 것
- 인공지능 기능을 추가하여 사물을 더 스마트하게 만드는 사물의 지능화를 의미

• 빅데이터

- 초지능의 원료
- 인공지능 : 초지능을제작

4차 산업혁명의 이해

■ 초지능

▶ 빅데이터

- 디지털 환경에서 발생하는 모든 데이터를 의미
- 4차 산업혁명의 서비스는 이전에는 사용하지 못했던 빅데이터를 사용하게 되면서 가능
- 빅데이터의 발생위치와 시점에 따라 위치 기반의 상황 인지 서비스가 가능해졌고 개인에 대한 빅데이터 사용이 가능해짐에 따라 개인별 맞춤 서비스가 제공

▶ 인공지능(AI)

- 인공지능(Artificial Intelligence, AI) : 인지·학습·추론·이해 능력과 같이 인간의 고차원적인 정보 처리 능력을 구현하는 기술
- 1950년 앨런 튜링의 튜링 기계와 이미테이션 게임에서 컴퓨터 기계와 지능에 대한 논의가 시작
- 1956년에는 다트머스대학교의하계 컨퍼런스에서 Artificial Intelligence라는 용어가 처음 사용(1차 전성기)
- 1970년에 들어서자 컴퓨터의 계산 기능과 논리 체계의 한계로 인공지능 이론 구현에 실패(1차 인공지능 겨울)
- 1980년대에는 신경망 다층 퍼셉트론 개발(2차 전성기)
- 신경망의 성능을 높이기 위해 필요한 데이터가 부족하고 프로세서의 계산 능력이 한계에 도달(2차 인공지능 겨울)
- 2000년대에 들어서면서 메모리, CPU, GPU 등의 하드웨어와 네트워크의 성능 향상으로 신경망 연구가 다시 활발해짐
- 빅데이터가 출현하면서 딥러닝의 성능 향상이 가속, 구글 딥마인드의 알파고가 바둑대회에서 우승(3차 전성기)

4차 산업혁명의 이해

■ 초지능

▶ 인공지능(AI)

- 인공지능의 주요 기술 분야

표2. AI의 주요 기술 분야

분야	설명
추론	인간의 사고 능력을 모방하는 기술
지식 표현 및 언어 지능	사람이 사용하는 자연어 이해를 기반으로 사람과 상호작용하는 기술
청각 지능	음성, 음향, 음악을 분석, 인식, 합성, 검색하는 기술
시각 지능	사물의 위치, 종류, 움직임, 주변과의 관계 등 시각 이해를 기반으로 지능화된 기능을 제공하는 기술
복합 지능	시공간, 촉각, 후각 등 주변의 상황을 인지 및 예측하고 상황에 적합한 대응을 제공하는 기술
지능형 에이전트	개인 비서, 챗봇 등 가상 환경에 위치하여 특별한 응용 프로그램을 다루는 사용자를 도울 목적으로 반복적인 작업을 자동화시켜 주는 기술
인간과 기계의 협업	인간의 감성이나 의도를 이해하기 위해 인간의 뇌 활동에 기계가 연동되어 작동하는 기술
AI 기반 하드웨어	초고속 지능 정보 처리를 구현하게 지원하는 하드웨어 기술

4차 산업혁명의 이해

■ 초융합(hyper-convergence)

- 초융합은 초연결 환경이 조성되면서 이전에는 생각할 수 없었던 여러 산업 및 기술이 결합되어 새로운 융합산업 또는 서비스가 촉진되는 것을 의미
- 4차 산업혁명의 지능 정보기술은 초연결성과 초지능화라는 기술적 특성을 통해 '모든 것이 상호 연결되고 보다 지능화된 사회로 변화'되는 초융합으로 진보

▶ 디지털 트랜스포메이션(digital transformation)

- 디지털 기술을 활용하여 기존 산업의 운영 및 생산의 효율성과 경쟁력을 높이는 프로세스의 변화를 의미
- 기업은 디지털 기술을 활용하여 다양한 산업 분야에서 지속적인 혁신을 추진, 특히 제조업에 주목 : 스마트 팩토리
- 기존 비즈니스 모델뿐만 아니라 고객의 경험을 변화시키고 추가 수익 흐름을 창출하여 새로운 방식으로 산업을 변화

▶ 로봇 프로세스 자동화(RPA : Robotic Process Automation)

- RPA는 업무과정에서 발생하는 데이터를 정형화하고 논리적으로 자동 수행하게 하는 기술
- RPA를 통해 단순하고 반복적인 업무를 감소시키고 오류를 줄여 업무 생산성 향상을 기대할 수 있음
- 금융 분야에서 비중이 가장 큼. 제조, 서비스 등의 다양한 분야로 확산 중
- RPA는 인간의 노동을 디지털 노동으로 전환하는 역할

4차 산업혁명과 데이터 과학

■ 데이터 과학

- 데이터 과학은 목적에 따라 데이터로부터 내재된 패턴을 발견하거나 분석하여 전략적 의미를 추론하는 방법론
- 빅데이터의 활용 중요성이 높아지면서 빅데이터와 관련 기술을 포괄하는 학문 분야로 중요하게 다루어짐
- 데이터 과학을 기반으로 혁신을 추구하는 과정에서 4차 산업혁명이 실현됨

▶ 데이터과학과 {IoT + 빅데이터 + AI}

- IoT와 5G의 확산으로 초연결성이 강화되어 더 많은 빅데이터가 생성, 전달, 저장
- AI가 이것들을 처리하고 분석
- 혁신적인 결과물(제품, 서비스 등)을 만들기 위해서는 목적을 정하고 그에 필요한 IoT, 빅데이터, AI 등의 기술을 적용하도록 프로세스를 관리하는 방법론, 데이터 과학이 필요함

▶ 데이터 과학과 21세기 핵심 역량

- 4차 산업혁명으로 인해 산업분야가 요구하는 주요 능력 및 역량에 변화 : 2020년에는 모든 산업 분야 가운데 3분의 1 이상이 복합 문제 해결 능력을 필요로 할 것(세계경제포럼)
- 기초 문해 6가지 (문해, 수해, 과학 문해, ICT 문해, 재정 문해, 문화 및 시민 문해)
- 복잡한 문제를 대하는 방법에 관한 역량 4가지(비판적 사고/문제 해결, 창의성, 의사소통, 협력)
- 변화하는 환경에 대한 대처 능력으로 인성 자질 6가지(호기심, 주도성, 일관성/끈기, 적응력, 리더십, 사회 및 문화 의식)

연습문제

- 4차 산업혁명의 3가지 키워드를 나열하고 각각의 의미를 설명하시오.
- 데이터 과학이란 무엇인가?
- IoT, 빅데이터, AI 란 무엇인가?
- IoT, 빅데이터, AI, 데이터 과학의 관계를 설명하시오.

빅데이터의 이해와 활용

빅데이터의 개념

■ 빅데이터의 정의

▶ 데이터의 규모 측면

- 인터넷의 확산으로 수많은 정보가 발생하고 소셜 미디어나 스마트 기기의 증가로 개인과 관련된 비정형 데이터가 대량으로 생성 및 축적되면서 빅데이터로 발전
- 디지털 환경에서 발생하는 대량의 모든 데이터 : PC와 인터넷, 모바일 기기, IoT 환경의 사물에서 발생하는 모든 데이터
- 다양한 기관에서 정의한 빅데이터

표1. 빅데이터의 정의

기관	정의
맥킨지	일반적인 데이터베이스 소프트웨어가 수집, 저장, 관리, 분석할 수 있는 범위를 초과하는 대규모의 데이터다.
가트너	향상된 시사점과 더 나은 의사결정을 위해 사용되는 것으로 비용 효율이 높고 혁신적이며 대용량 고속 및 다양성을 가지는 정보 자산이다.
위키피디아	기존 데이터베이스 관리 도구의 수집, 저장, 관리, 분석 역량을 넘어서는 대량의 정형 또는 비정형 데이터셋 및 이러한 데이터로부터 가치를 추출하고 결과를 분석하는 기술이다.
국가전략위원회	대용량 데이터를 활용 및 분석하여 가치 있는 정보를 추출하고 생성된 지식을 바탕으로 능동적으로 대응하거나 변화를 예측하기 위한 정보화 기술이다.
삼성경제연구소	기존의 관리 및 분석 체계로는 감당할 수 없을 정도의 거대한 데이터 집합으로 대규모 데이터와 관계된 기술 및 도구(수집, 저장, 검색, 공유, 분석, 시각화 등)를 모두 포함한다.
한국정보화진흥원	저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터와 이것을 저장, 관리, 분석할 수 있는 하드웨어 및 소프트웨어 기술, 데이터를 유통 및 활용하는 과정을 통틀어 나타낸다. 즉, 빅데이터를 구성하는 하드웨어, 소프트웨어 그리고 이를 포괄하는 모든 프로세스를 의미하는 거대 플랫폼이다.

빅데이터의 개념

■ 빅데이터의 정의

▶ 데이터의 활용 측면

- 데이터의 규모와 기술 측면에서 출발한 빅데이터 정의는 가치와 활용 효과 측면으로 확대
- 빅데이터란 대규모의 데이터만을 의미하는 것이 아니라, 대규모의 데이터를 저장 · 관리 · 분석할 수 있는 하드웨어 및 소프트웨어 기술, 데이터를 유통 · 활용하는 모든 프로세스를 포함하는 빅데이터 플랫폼으로 의미를 확장
- 빅데이터 플랫폼
 - 하드웨어와 소프트웨어 기술 그리고 애플리케이션을 이용하여 빅데이터를 다루는 데이터 과학은 빅데이터가 가치있게 활용될 수 있도록 해줌
 - 빅데이터 플랫폼을 구성하는 하드웨어, 소프트웨어, 애플리케이션 간의 유기적 순환에 의해 가치를 창출
 - 애플리케이션은 빅데이터를 분석하고 그 결과를 사용자에게 제공
- 빅데이터의 활용 중요성이 높아지면서 빅데이터와 관련 기술을 포괄하는 학문 분야로 중요하게 다루어짐
- 동의없이 개인정보를 어디까지 사용할 수 있는가의 문제는 여전히 남아 있음



그림1. 빅데이터 플랫폼

빅데이터의 개념

■ 빅데이터의 출현

- ▶ 기술의 발달과 비용 저하, 소셜 네트워크 서비스 발달, 그림자 정보와 사물 정보 증가 등의 ICT 패러다임의 변화 ⇒ 데이터의 양적 팽창
- ▶ 빅데이터에 전문 역량과 기술을 더하여 전략적으로 활용할 방법이 주목됨
- ▶ 경제적 가치 창출, 사회 문제 해결, 새로운 ICT 패러다임 견인이라는 신가치 창출

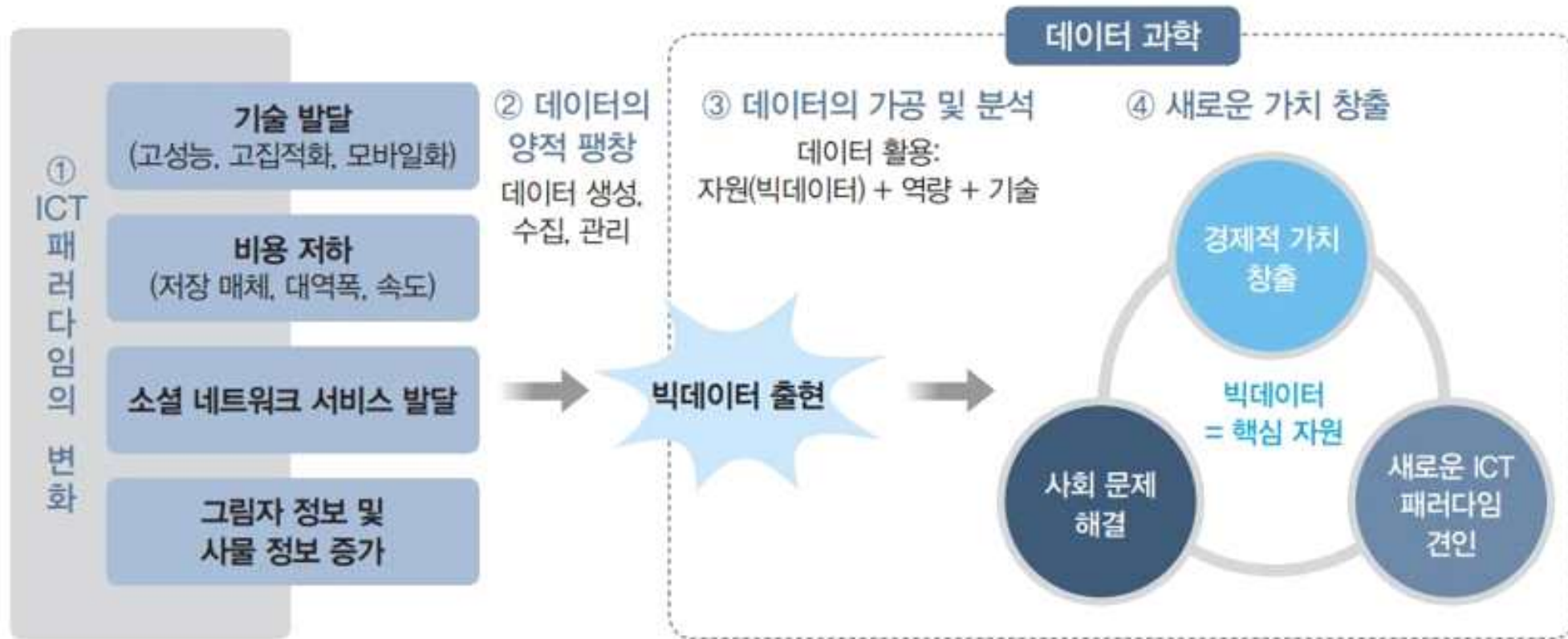


그림2. 빅데이터의 출현과 신가치 창출의 흐름

빅데이터의 개념

■ 빅데이터의 분류

▶ 정형데이터

- 일정한 규칙으로 체계적으로 정리된 것으로 그 자체로 해석이 가능하여 바로 활용할 수 있음

▶ 반정형데이터

- 고정된 필드에 저장되어 있지는 않지만 XML, HTML 등의 메타데이터와 스키마를 포함하는 것으로 파일 형태로 저장

▶ 비정형 데이터

- 고정된 필드나 스키마가 없는 것
- 스마트 기기에서 페이스북, 트위터, 유튜브 등으로 생성되는 소셜 데이터
- IoT환경에서 생성되는 위치 정보나 센서 데이터와 같은 사물 데이터 등에 해당

빅데이터의 개념

■ 빅데이터의 분류

표2. 정형화 정도에 따른 빅데이터의 분류

구분	설명	수집 및 처리 난이도
정형 데이터	<ul style="list-style-type: none">고정된 필드에 저장관계형 데이터베이스처럼 스키마 형식에 맞게 저장예: RDB, 스프레드시트	<ul style="list-style-type: none">내부 시스템에 의한 데이터라 수집하기 쉬움파일 형태의 스프레드시트는 형식을 가지고 있어 처리하기 쉬움처리 난이도: 하
반정형 데이터	<ul style="list-style-type: none">고정된 필드에 저장되어 있지는 않지만 메타 데이터나 스키마 등을 포함예: XML, HTML, JSON, 웹 문서, 웹 로그	<ul style="list-style-type: none">API 형태로 제공되므로 데이터 처리 기술이 필요함처리 난이도: 중
비정형 데이터	<ul style="list-style-type: none">데이터 구조가 일정하지 않음규격화된 데이터 필드에 저장되지 않음예: 소셜 데이터, 텍스트 문서, 이미지/동영상/음성 데이터, 문서 파일(PDF)	<ul style="list-style-type: none">파일을 데이터 형태로 파싱해야 하므로 처리하기 어려움처리 난이도: 상

빅데이터의 개념

■ 빅데이터의 특징

▶ 데이터 측면

- 초기에는 빅데이터의 특징을 3V로 일컬어지는 규모, 다양성, 속도로 나타냄
- 빅데이터를 통한 가치 창출이 중요해지면서 정확성과 가치를 추가한 5V로 나타냄

표3. 빅데이터의 특징 - 데이터 측면

구분	특징	설명
1차 특징	규모	• ICT 기술의 발전으로 디지털 정보량이 기하급수적으로 폭증하여 제타바이트 시대로 진입
	다양성	• 데이터의 종류 증가 : 로그 기록, 소셜/위치/소비/현실 데이터 등 • 데이터의 유형 다양화 : 텍스트 외에 멀티미디어 등의 비정형 데이터 증가
	속도	• 센서, 모니터링 등의 사물 정보와 스트리밍 등의 실시간 정보가 증가하면서 데이터의 생성 및 이동(유통) 속도 증가 • 대규모 데이터를 처리하고 가치 있는 정보를 활용하기 위한 데이터 처리 및 분석 속도 증가
추가 특징	정확성	• 방대한 데이터를 기반으로 분석을 수행하므로 정확성 향상
	가치	• 빅데이터 분석으로 도출된 최종 결과물이 문제 해결을 위한 통찰력을 제공하므로 새로운 가치 창출 가능

빅데이터의 개념

■ 빅데이터의 특징

▶ 데이터 분석 환경 측면

- 데이터 분석 시스템의 구성 요소인 데이터, 하드웨어, 소프트웨어 분석 방법은 분석 환경에 따라 다른 특징을 나타냄

표4. 빅데이터의 특징 - 데이터 분석 환경 측면

요소	과거의 데이터 분석 환경	현재의 데이터 분석 환경
데이터	<ul style="list-style-type: none">• 정형화된 수치 중심의 자료	<ul style="list-style-type: none">• 비정형의 다양한 데이터• 예 : 문자데이터(SMS, 검색어), 영상 데이터(CCTV, 동영상), 위치 데이터 등
하드 웨어	<ul style="list-style-type: none">• 고가의 저장 장치• 데이터베이스• 대규모 데이터웨어하우스	<ul style="list-style-type: none">• 클라우드 컴퓨팅 : 비용 대비 효율성 증대
소프트웨어 분석 방법	<ul style="list-style-type: none">• 관계형 데이터베이스 : RDBMS• 통계패키지 : SAS, SPSS• 데이터마이닝• 머신 러닝• 지식 발견	<ul style="list-style-type: none">• 오픈 소스 형태의 무료 소프트웨어• 오픈소스 통계 솔루션 : R• 텍스트 마이닝• 오피니언 마이닝• 감성 분석

빅데이터의 개념

■ 빅데이터의 특징

▶ 데이터 처리 방식 측면

- 빅데이터는 기존 데이터베이스 관리 시스템(DBMS)으로 처리하던 것에 비해 100배 이상 많은 정형, 비정형 데이터를 처리

표5. 빅데이터의 특징 - 데이터 처리 방식 측면

구분	이전의 데이터 처리 방식	빅데이터 처리 방식
데이터 트래픽	<ul style="list-style-type: none">• 테라바이트 수준	<ul style="list-style-type: none">• 페타바이트 수준 : 최소 100테라바이트 이상• 정보의 장기간 수집 및 분석• 방대한 처리량
데이터 유형	<ul style="list-style-type: none">• 정형 데이터 중심	<ul style="list-style-type: none">• 비정형 데이터 비중이 높음 : SNS데이터, 로그 파일, 클릭스트림 데이터, 콜센터 로그 통신 등• 처리 복잡성 증대
프로세스 및 기술	<ul style="list-style-type: none">• 단순한 프로세스 및 기술• 정형화된 처리 및 분석결과• 원인 및 결과 규명 중심	<ul style="list-style-type: none">• 다양한 데이터 소스와 복잡한 로직 처리• 처리 복잡도가 높아 분산 처리 기술 필요• 새롭고 다양한 처리 방법 필요 : 정의된 데이터 모델/상관관계/절차 등이 없음• 상관관계 규명 중심• 하둡, NoSQL 등 개방형 소프트웨어 사용

빅데이터의 활용

■ 빅데이터의 활용분야

- ▶ 상품 추천 시스템 : 쇼핑몰 사이트, 인터넷 사용 패턴을 분석한 광고, 유튜브 맞춤 동영상이나 넷플릭스 추천 동영상
- ▶ 의료 서비스 : 건강보험관리공단에 있는 질병 관련 데이터 활용, 나이와 질병 관계, 생활 방식과 질병 관계 등 다양한 데이터를 추출하고, 나이별로 필요한 예방접종을 실행하고, 어떤 건강검진을 받아야 하는지 결정
- ▶ 보험 설계 : 위험 요인과 확률을 추출하여 보험료를 결정, 탈세자나 사기 범죄자를 색출
- ▶ 광고판의 안면 인식 : 광고판이 내 성향을 파악하고 나에게 맞는 정보를 제공

■ 빅데이터의 역할

- ▶ 미래 사회의 특성은 불확실성, 리스크, 스마트, 융합으로 대변됨
- ▶ 빅데이터를 활용해 여러 가지 가능성에 대한 시나리오 시뮬레이션을 하면 불확실한 상황 변화에 유연하게 대처 가능
- ▶ 빅데이터에 기반한 정보 패턴 분석으로 리스크에 대응할 수 있음
- ▶ 개인화 및 지능화된 서비스를 제공하여 삶의 질을 향상시킴

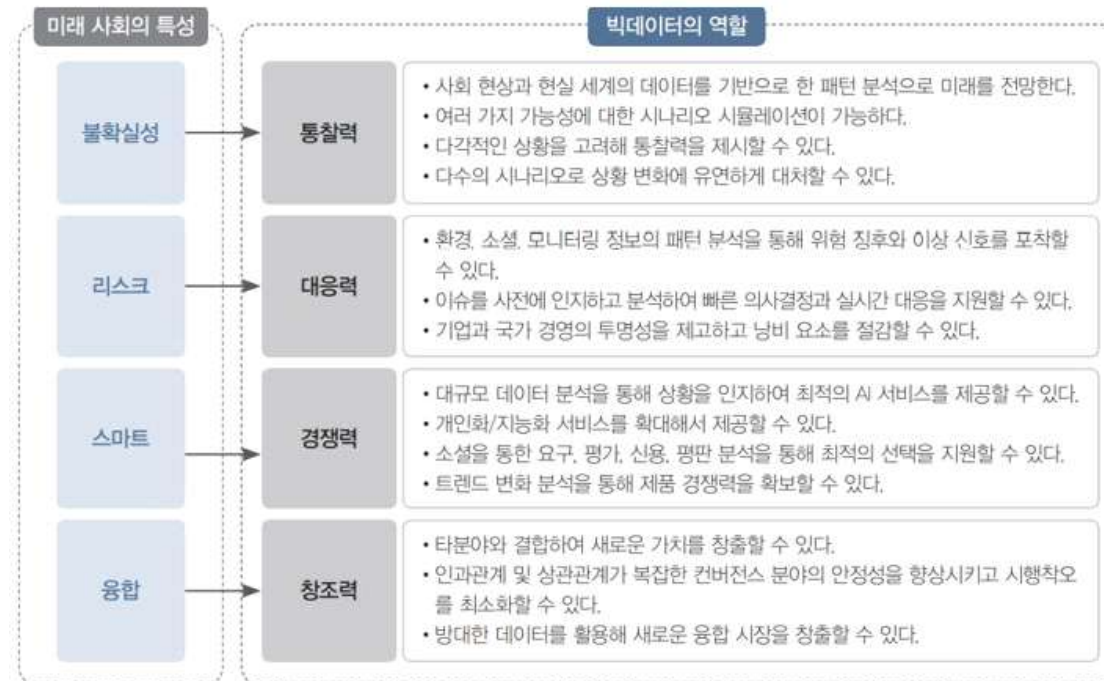


그림3. 미래사회의 특성과 대응되는 빅데이터의 역할

■ 성공적인 빅데이터 활용을 위한 3요소

- ▶ 자원, 기술, 인력의 3가지 요소에 대한 전략을 수립

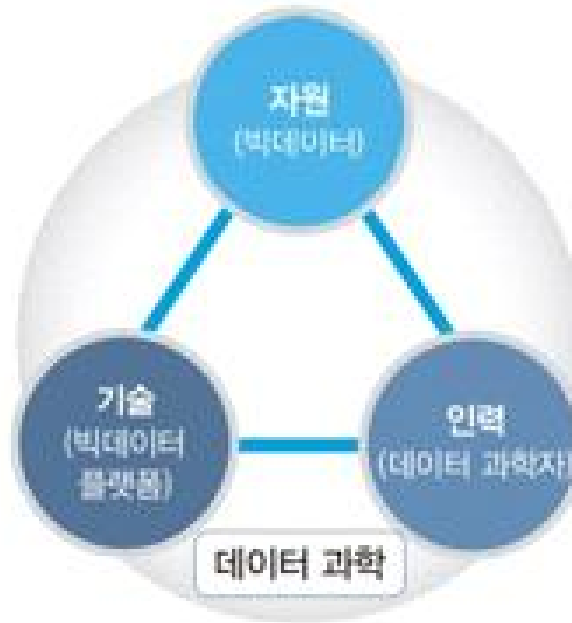


그림4. 성공적인 빅데이터 활용을 위한 3요소

■ 데이터 과학자 역량 강화

- ▶ 빅데이터 시대에는 데이터를 분석하고 관리할 수 있는 인력에 대한 중요성이 커짐
- ▶ 대규모 데이터 속에서 숨겨진 정보를 찾아내는 데이터 과학자는 '빅데이터 시대의 연금술사'
- ▶ 존 라우치 : 데이터 과학자에게 필요한 6가지 기본 자질
 - ① 수학 역량
 - ② 공학 역량
 - ③ 데이터를 분석할 때 필수적인 가설을 세우거나 검증할 때 필요한 비판적 시각
 - ④ 이를 잘 작성할 수 있는 글쓰기 역량
 - ⑤ 다른 사람에게 잘 전달할 수 있는 대화 능력
 - ⑥ 호기심과 개인의 행복
- ▶ 데이터 과학자는 외부보다는 내부 인력으로 내재화하여 활용

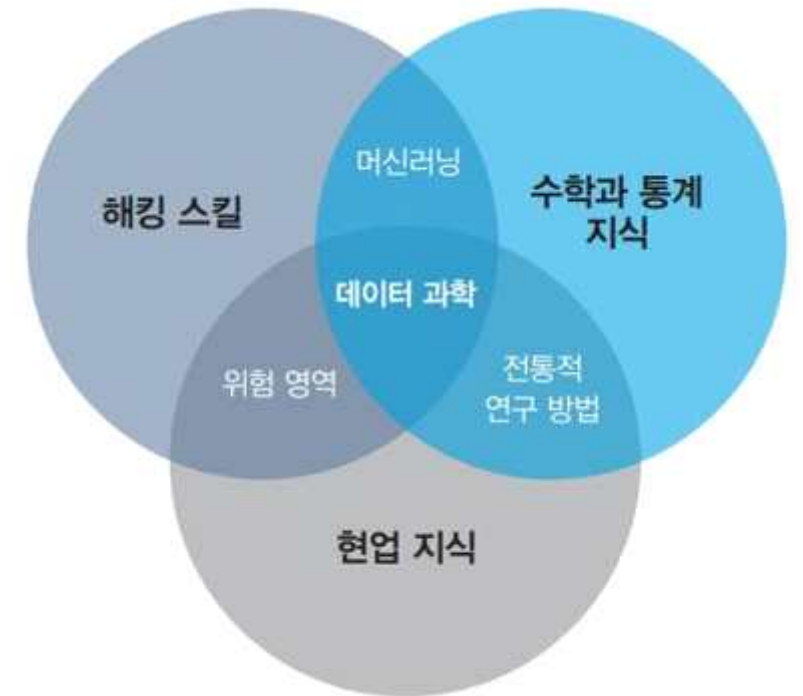


그림5. 데이터 과학자가 지녀야 할 역량 벤다이어그램

데이터 과학 방법론

■ 데이터 과학 방법론 6단계

- 데이터 과학 방법론은 6단계로 구성되며 필요에 따라 특정 단계를 반복해서 수행 가능



그림6. 데이터 과학 방법론의 6단계 구성

데이터 과학 방법론

■ 데이터 과학 방법론 6단계

▶ 1단계 : 연구 목표 설정

- 무엇을 분석할지, 결과가 어디에 필요한지, 어떤 데이터가 필요한지 등을 이해당사자와 논의하여 결정

▶ 2단계 : 데이터 수집

- 연구 또는 프로젝트에 필요한 데이터의 위치와 형태를 확인하고 원시 데이터(raw data) 수집
- 내부 데이터베이스 또는 외부 기관등에서 수집

▶ 3단계 : 데이터 준비

- 수집한 원시 데이터의 품질을 높이기 위해 정제 후 사용 가능한 형태로 가공하는 단계
- 수집한 데이터를 다음 단계에서 사용할 수 있게 오류를 여과(filtering)하거나 수정하여 정제하고 필요에 따라서는 데이터를 통합하거나 형태를 변환함

데이터 과학 방법론

■ 데이터 과학 방법론 6단계

▶ 4단계 : 데이터 탐색

- 데이터 탐색은 데이터와 변수 간의 관계나 상호 작용을 이해하기 위한 단계
- 변수 간의 관련성, 데이터의 분포, 편차, 패턴 존재 여부를 확인하는 탐색적 데이터 분석(EDA : Exploratory Data Analysis)
- 꺾은선 그래프, 히스토그램, 분포도 등과 같은 그래픽 기법을 많이 사용함

▶ 5단계 : 데이터 모델링

- 데이터 모델링은 이전 단계에서 얻은 데이터 탐색 결과로 프로젝트나 연구에 대한 답을 찾는 단계
- 변수를 선택하여 모델을 구성하고 실행 및 평가하는 과정을 반복 수행하여 문제 해결 모델을 완성함
- 분석하려는 데이터의 특성과 목적에 따라 모델 유형을 선택할 수 있음
- 기술통계, 상관분석, 회귀분석, 분산분석, 주성분분석, 데이터마이닝, 텍스트마이닝, 소셜네트워크 분석 등

▶ 6단계 : 결과 발표 및 분석 자동화

- 수행 결과가 연구 목표를 달성했는지를 이해 당사자, 특히 의사 결정자에게 이해시키고 가능하다면 이후의 유사 프로젝트 수행을 위해 분석 과정을 자동화하는 단계

연습문제

1. 다음 중 비정형 데이터에 대한 설명에 해당하지 않은 것을 고르시오.

- ① 스마트기기 등을 통하여 생성되는 데이터다.
- ② 빅데이터 분석은 비정형 데이터를 대상으로 한다
- ③ 규격화된 데이터 필드에 저장되지 않은 데이터다.
- ④ 대표적인 예로 소셜데이터, 사물 인터넷에서 생성되는 사물 데이터가 있다.

2. 빅데이터의 특징 중 데이터 측면에 속하지 않는 것을 고르시오.

- ① 규모의 증가 ② 속도의 증가 ③ 가치의 증가 ④ 다양성의 증가

3. 다음 중 빅데이터의 역할에 해당하지 않는 것을 고르시오.

- ① 시나리오 시뮬레이션을 통해 불확실한 상황 변화에 유연하게 대처할 수 있다.
- ② 이슈를 사전에 인지하고 분석하여 빠른 의사결정을 내릴 수 있으므로 리스크에 실시간에 대응할 수 있다.
- ③ 개인화 및 지능화된 서비스를 확대해서 제공하여 스마트 사회에서 삶의 질을 향상시킬 수 있다.
- ④ 타분야와의 결합보다는 빅데이터 활용에 맞는 새로운 분야를 개발할 수 있다.

4. 다음 중 성공적인 빅데이터 활용 전략의 3요소를 바르게 나타낸 것을 고르시오.

- ① 자원-기술-인력 ② 자원-기술-품질 ③ 기술-품질-인력 ④ 기술-투자-인력

5. 다음 중 존라우저가 선정한 데이터 과학자의 6가지 역량에 해당되지 않는 것을 고르시오.

- ① 자원-기술-인력 ② 자원-기술-품질 ③ 기술-품질-인력 ④ 기술-투자-인력

데이터 분석 프로그램

Spyder 소개

■ Spyder

▶ Anaconda에는 Jupyter Notebook과 같은 웹기반의 editor와 Spyder, PyCharm 등의 IDE등이 있음

- 통합개발환경(Integrated Development Environment, IDE)은 소스코드 작성 및 실행, 저장 등 개발의 전 과정을 지원하는 도구
- 기본 python 인터프리터보다 IDE나 editor가 긴 코드를 입력하거나 코드를 저장할 때 편리함
- 파이썬 인터프리터(IDLE) 셸모드 사용시 오타나 코드 수정이 직접 안되고 이점은 프로그래밍 작업 시에 불편함

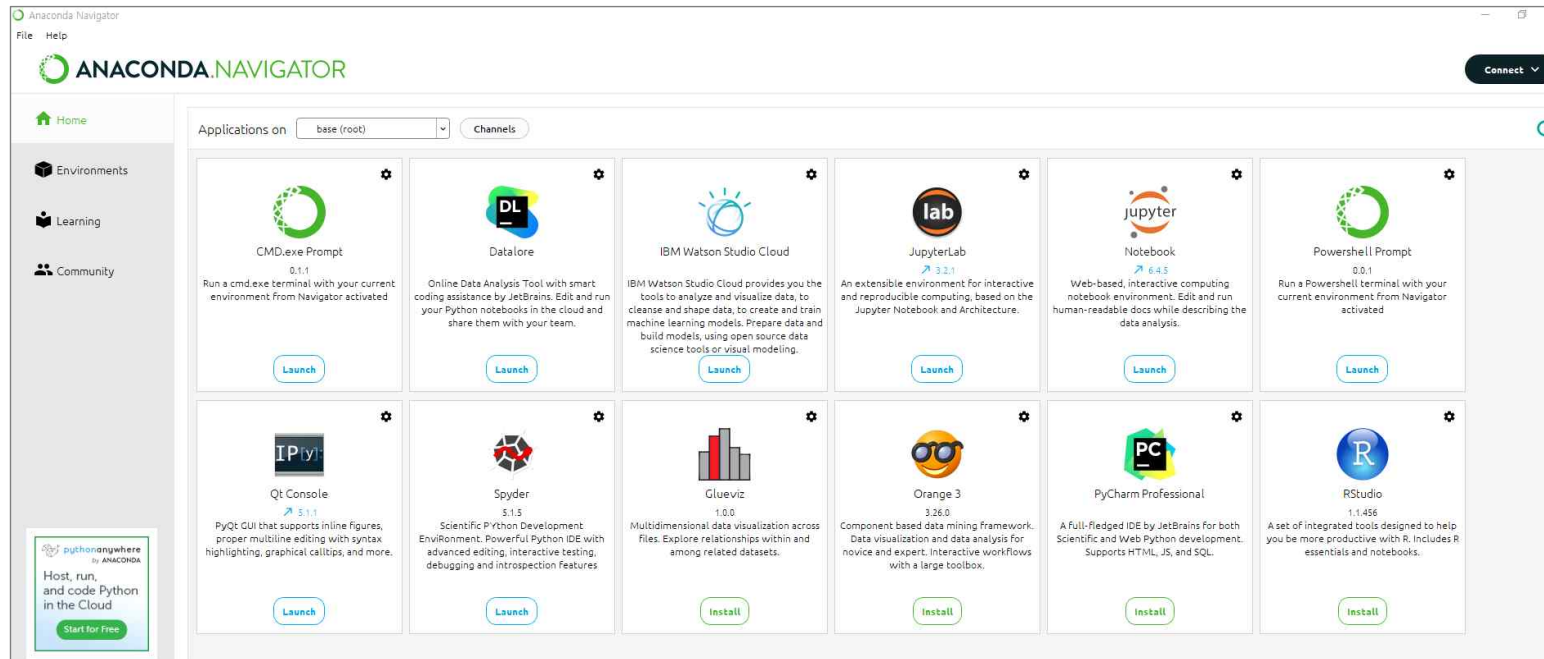


그림1. ANACONDA NAVIGATOR

Spyder 소개

■ Spyder 실행

- ▶ 본 수업에서는 Spyder를 사용하여 강의하려고 함(수강생들은 자유롭게 선택)
- ▶ 윈도우 시작 화면에서 Spyder 바로가기 클릭하거나 Anaconda Navigator를 실행하여 Spyder 바로가기 클릭하여 Spyder 실행



그림2. Spyder_윈도우시작화면

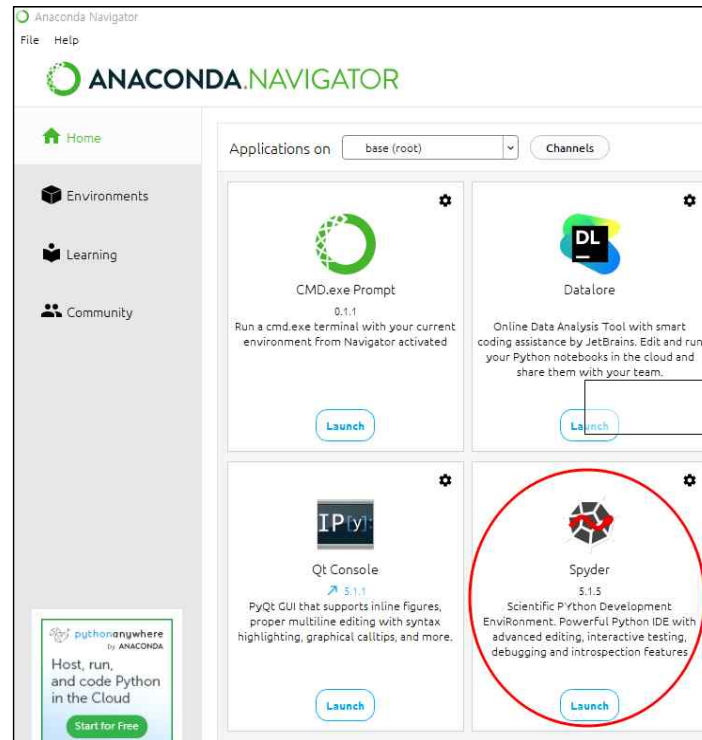


그림3. Spyder_ANACONDA NAVIGATOR

Spyder 소개

■ Spyder 기본창

▶ Spyder는 기본적으로 세 개의 창으로 구성

- ① script Editor : 파이썬 스크립트(명령어)를 작성하고 데이터를 볼 수 있고 스크립트의 수정, 변경이 가능함
- ② Variable Explorer : 파이썬 실행시 작성한 변수들이 기록
- ③ Ipython console : 스크립트를 작성하고 실행하면 그 실행결과를 반환

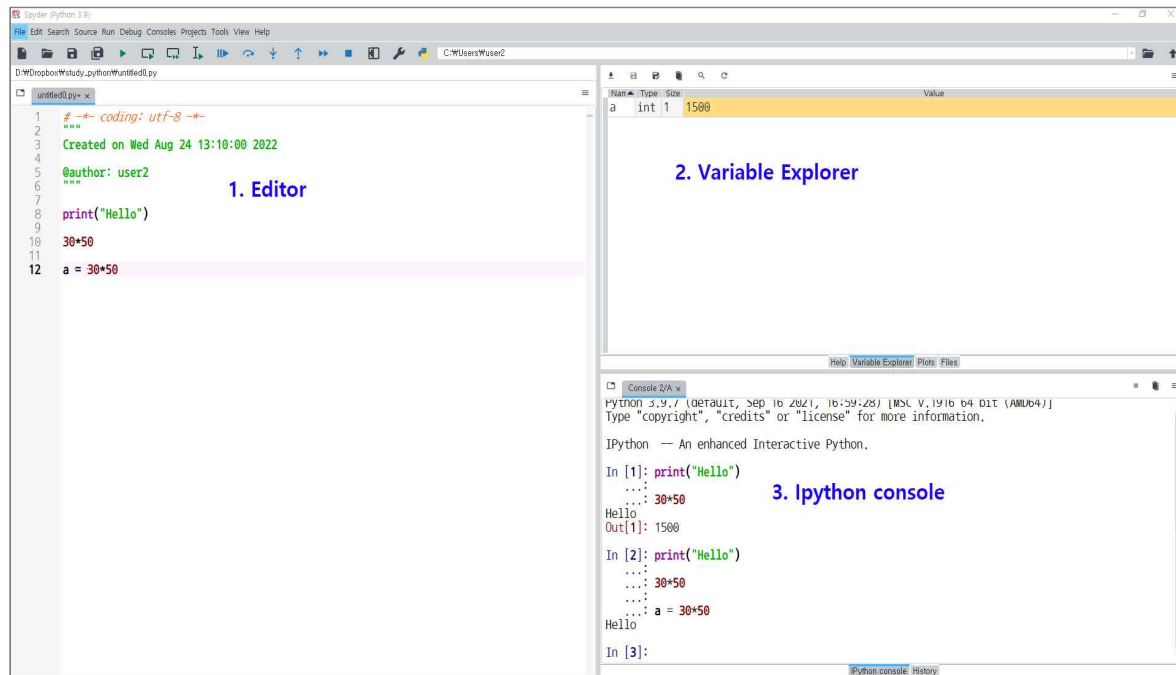




그림5. Spyder 기본창

Spyder 소개

■ Spyder에서 코드 입력, 실행, 저장

▶ 예 : `print("Hello")`, `30*50` 실행

- script Editor에 다음과 같이 수식을 입력한 뒤,  를 눌러 실행하거나  를 실행 ⇒ 결과는 Ipython console창에서 확인

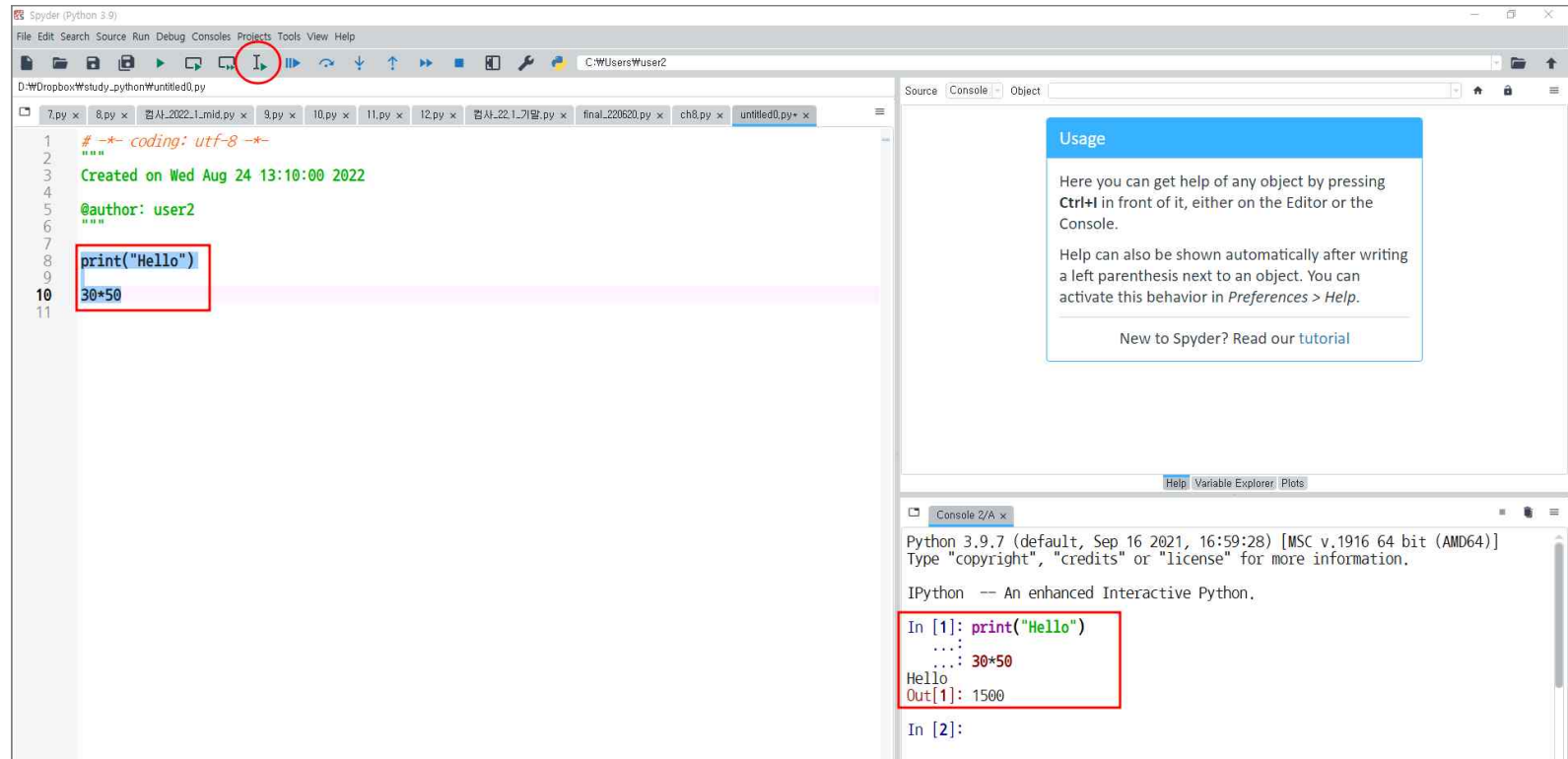


그림6. Spyder 코드 입력, 실행 결과

Spyder 소개

■ Spyder에서 코드 실행(Run) 방법

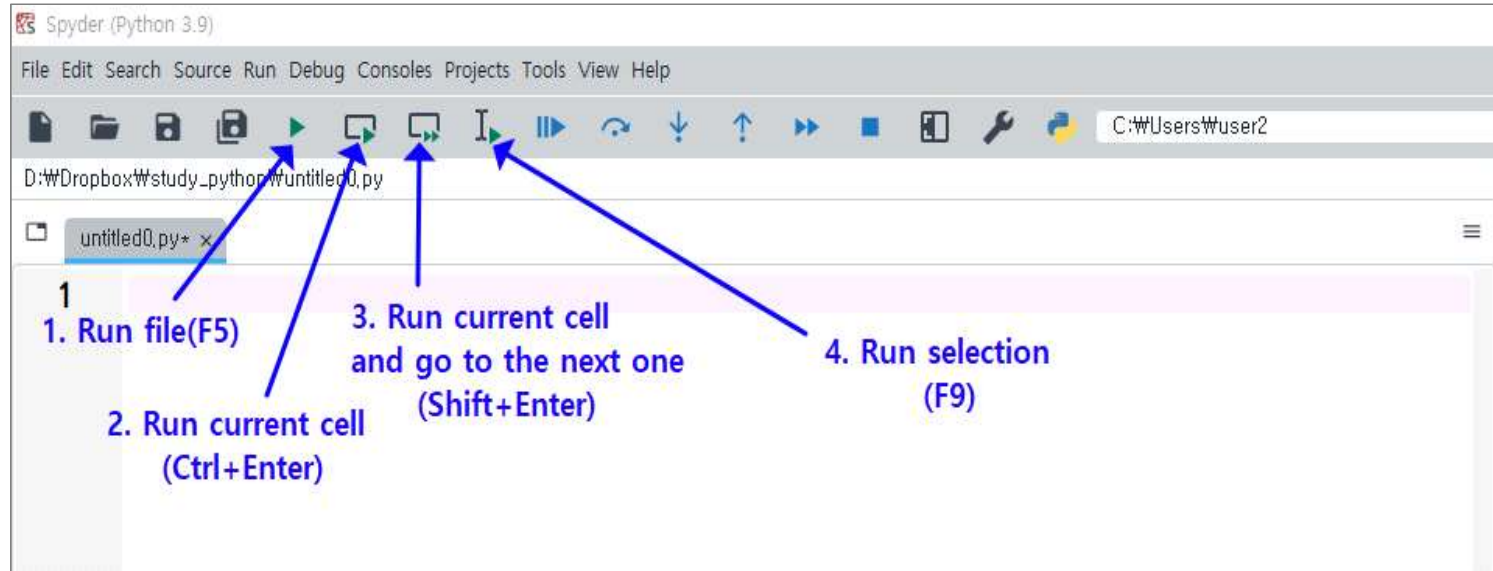


그림7. Spyder 실행방법

- ① 저장한 파일 전체를 처음부터 끝까지 실행 (Run file : F5)
- ② 현재 커서가 위치한 cell 전체를 실행 (Run current cell) 후 커서를 현 cell에 위치 : Ctrl+Enter
- ③ 현재 커서가 위치한 cell 전체를 실행 후 다음 번 cell로 커서를 이동(Run current cell and go to the next one) : Shift+Enter
- ④ 현재 커서가 위치한 행(row) 또는 선택한 행 전체를 실행(Run selection) : F9

Spyder 소개

- Spyder에서 파일 저장 : [File]-[Save as]-'sample1'로 저장(sample1.py)

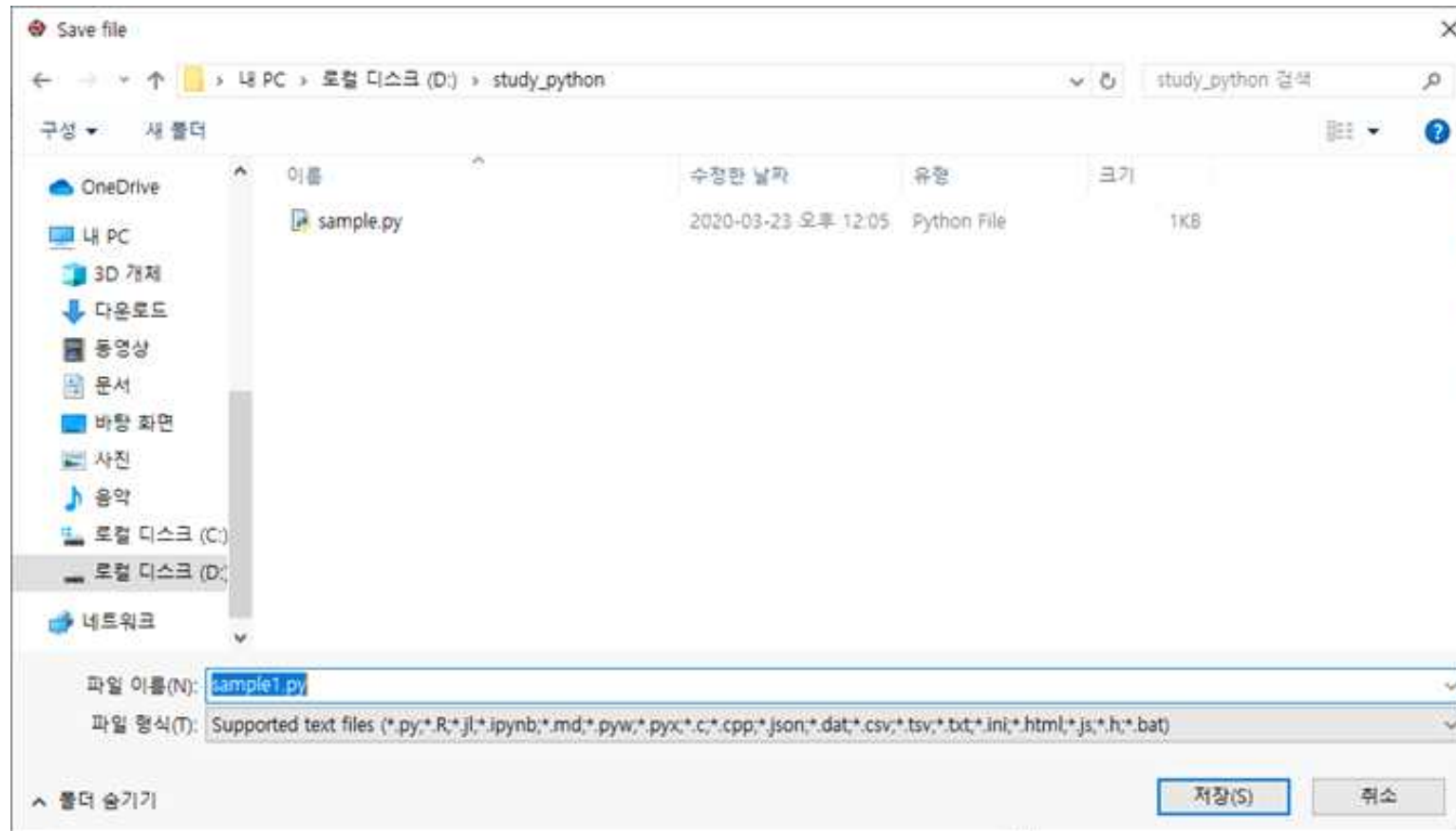
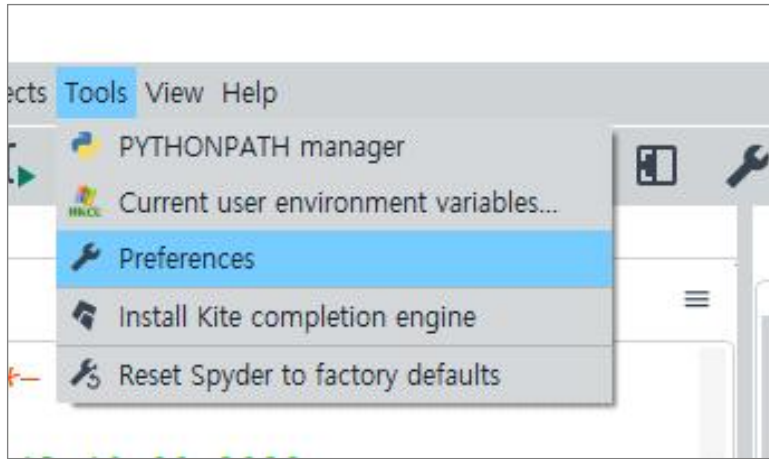


그림8. Spyder 실행결과 저장

Spyder 소개

■ Spyder의 환경 설정

- ▶ 메뉴에서 Tools>Preferences>Appearance - Syntex highlighting theme, Fonts 등 설정 변경



데이터의 분류와 데이터프레임

자료(Data)의 분류

▶ 양적자료(quantitative data) 또는 숫자형 자료(numerical data)

- 계측데이터(metric)
 - 길이, 무게 등과 같이 양적인 수치로 측정되거나 몇 개인가를 세어 측정하는 변수
 - 사칙연산 가능
- ① 연속형 자료(Continuous data) : 무게, 키, 전구수명등과 같이 셀 수 없는 소수점을 포함하는 형태의 자료
- ② 이산형 자료(Discrete data) : 자녀의 수, 교통사고건수, 불량품의 수등과 같이 셀 수 있는 정수 형태의 자료

▶ 질적자료(qualitative data) 또는 범주형 자료(categorical data)

- 비계측자료(nonmetric)
 - 학년, 성별, 등급 등
 - 조사대상을 특성에 따라 범주로 구분하여 측정된 변수
 - 사칙연산 불가능
- ① 명목형 자료(Nominal data) : 성별, 직업, 지역등과 같이 자료값의 크기나 순서에 대한 의미가 없고 자료값 자체의 이름만 의미를 부여한 자료
- ② 순서형 자료(Ordinal Data): 학년, 순위, 성적등급 등과 같이 어떤 기준에 따라 자료값들의 순서의 개념이 있는 자료

pandas 데이터프레임 생성

- 데이터 분석에서 자주 사용하는 테이블 형태를 다룰 수 있는 라이브러리
- pandas 설치 및 호출

<code>pip install pandas</code>	<ul style="list-style-type: none">• pandas 설치,• anaconda 플랫폼을 설치하여 IDE를 사용하는 경우, 추가 설치 없이 바로 사용할 수 있음
<code>import pandas as pd</code>	<ul style="list-style-type: none">• pandas를 사용하기 위해서는 pandas를 import

- 1차원 자료구조인 시리즈(Series), 2차원 자료구조인 데이터프레임(DataFrame), 3차원 자료구조인 패널(Panel) 지원
- 시리즈, 데이터프레임을 많이 사용함

- 예. 5명의 학생들에 대한 시험 성적 자료

번호	성별	수학	영어
1	M	66	70
2	M	64	68
3	F	48	46
4	F	46	48
5	M	78	84

■ 시리즈 자료형 생성

- ▶ pandas의 시리즈는 딕셔너리, 리스트, arange를 이용하여 생성할 수 있음

```
import pandas as pd
```

```
math = pd.Series([66,64,48,46,78])
```

```
english = pd.Series([70,68,46,48,84])
```

```
gender = pd.Series(['m','m','f','f','m'])
```

- ▶ pandas를 import한다.

- ▶ pd.Series([])를 이용하여 Series를 작성


```
In [36]: import pandas as pd
...:
...: math = pd.Series([66,64,48,46,78])
...: math
Out[36]:
0    66
1    64
2    48
3    46
4    78
dtype: int64

In [37]: gender = pd.Series(['m','m','f','f','m'])
...: gender
Out[37]:
0    m
1    m
2    f
3    f
4    m
dtype: object
```

■ 데이터프레임 자료형 생성

- ▶ 데이터프레임(=데이터셋) : 변수와 개체(관측치), 두 개의 차원에 의하여 직사각형으로 배열된 자료값들의 집합
- ▶ 변수(variable) : 데이터프레임의 한 열(column)로서 구체적인 속성을 나타내는 자료값들의 집합
 - 문자, 숫자, 특수문자 및 기호 등으로 이루어진 문자형(character) 변수와 숫자만으로 이루어진 숫자형(numeric) 변수로 분류됨
- ▶ 개체(observation) : 데이터프레임의 한 행(row)으로서 동일한 관찰단위에 대한 모든 변수들의 자료값들의 집합

	ID	Name	Gender	Weight	Height	← 변수(variable)
1	101	김철수	M	74	170	
2	102	이영희	F	68	166	
3	103	안수지	F	55	155	
4	104	박민호	M	72	167	
5	105		M	66	169	
6	106	강지영	F	52	.	

개체(observation) ← (row index 1-6)
결측값(missing value) ← (row 6, column 5)

- ▶ 파이썬에서는 pandas 패키지의 DataFrame()함수를 이용하여 데이터프레임을 생성하거나 내장되어 있는 데이터를 호출하여 데이터프레임으로 사용

■ 값을 직접 입력하여 데이터프레임 생성

```
from pandas import DataFrame

rawdata = {'ID': [1,2,3,4,5],
           'Gender': ["M","M","F","F","M"],
           'Math' : [66,64,48,46,78],
           'English' : [70,68,46,48,84]
           }

score = DataFrame(rawdata)
```

- ▶ 판다스 패키지로부터 DataFrame 함수 import하기
- ▶ 딕셔너리를 이용하여 각 컬럼별로 데이터를 작성한 후 rawdata 생성
 - ID, Gender, Math, English를 key로 하여 대응하는 값들을 리스트로 입력
- ▶ 딕셔너리 rawdata에 pandas 패키지의 DataFrame함수 적용하여 score 데이터프레임 생성
- ▶ 데이터프레임은 문자와 숫자가 같이 사용되어도 숫자가 문자로 바뀌지 않음

```
In [1]: from pandas import DataFrame
...:
...: rawdata = {'ID': [1, 2, 3, 4, 5],
...:            'Gender': ["M", "M", "F", "F", "M"],
...:            'Math': [66, 64, 48, 46, 78],
...:            'English': [70, 68, 46, 48, 84]}
...:
```

```
In [2]: rawdata
```

```
Out[2]:
```

```
{'ID': [1, 2, 3, 4, 5],
 'Gender': ['M', 'M', 'F', 'F', 'M'],
 'Math': [66, 64, 48, 46, 78],
 'English': [70, 68, 46, 48, 84]}
```

```
In [3]: score = DataFrame(rawdata)
```

```
In [4]: score
```

```
Out[4]:
```

	ID	Gender	Math	English
0	1	M	66	70
1	2	M	64	68
2	3	F	48	46
3	4	F	46	48
4	5	M	78	84

▶ score 데이터프레임에서 변수 불러오기

`score.Gender`

`score.Math`

`score['English']`

▶ score의 Gender 변수를 불러온다

▶ score의 Math 변수를 불러온다

```
In [13]: score.Gender
Out[13]:
0      M
1      M
2      F
3      F
4      M
Name: Gender, dtype: object

In [14]: score.Math
Out[14]:
0      66
1      64
2      48
3      46
4      78
Name: Math, dtype: int64
```

■ 외부 데이터 파일을 읽어 들여 데이터프레임 생성

- ▶ pandas 패키지는 다양한 형태의 외부 파일을 읽어와서 행과 열이 있는 데이터프레임으로 변환
- ▶ 텍스트 파일, 공백 또는 콤마(,)등으로 구분되어 있는 csv(comma-separated values)형식의 외부 파일, Excel 파일 등을 읽어들이 데이터프레임으로 생성
- ▶ 어떤 파일이든 pandas 패키지의 데이터프레임으로 변환되고 나면 pandas 패키지의 모든 함수와 기능을 자유롭게 사용할 수 있음

▶ 텍스트 파일로부터 데이터프레임 생성하기 : cdc.txt

```
pd.read_csv("d:/data/cdc.txt", sep=" ")  
파일경로:/파일명.확장자(txt, csv)
```

```
import pandas as pd  
cdc = pd.read_csv("d:/data/cdc.txt", sep=" ")
```

- ▶ 외부에 저장된 cdc.txt 파일을 읽어오기
 - cdc.txt 파일의 경로와 확장자를 정확히 지정
 - cdc.txt 파일을 pandas 데이터프레임으로 저장
- ▶ sep=" " 옵션 : 텍스트 데이터의 변수들이 공백(" ")으로 구분되어 있어 지정해줌.
 - 파일에 따라 쉼표(,) 또는 탭(wt)으로 구분되어 있을수도 있으므로 원시 데이터 확인 후 적절한 옵션을 지정

- spyder의 variable explorer 탭에서 cdc 데이터프레임 확인

```
In [53]: import pandas as pd
```

```
In [54]: cdc = pd.read_csv("d:/data/cdc.txt", sep=" ")
```

```
In [55]: cdc
```

```
Out[55]:
```

	genhlth	exerany	hlthplan	smoke100	...	weight	wt Desire	age	gender
1	good	0	1	0	...	175	175	77	m
2	good	0	1	1	...	125	115	33	f
3	good	1	1	1	...	105	105	49	f
4	good	1	1	0	...	132	124	42	f
5	very good	0	1	0	...	150	130	55	f
...
19996	good	1	1	0	...	215	140	23	f
19997	excellent	0	1	0	...	200	185	35	m
19998	poor	0	1	0	...	216	150	57	f
19999	good	1	1	0	...	165	165	81	f
20000	good	1	1	1	...	170	165	83	m

```
[20000 rows x 9 columns]
```

▶ csv(comma-separated values)형식의 외부 파일로부터 데이터프레임 생성하기 : cars93.csv

```
import pandas as pd
cars93 = pd.read_csv("d:/data/cars93.csv")
```

- ▶ 외부에 저장된 cars93.csv 파일을 읽어들이 데이터프레임 생성
 - cars93.csv 경로와 확장자를 정확히 지정
 - cars93.csv 파일을 pandas 데이터프레임으로 저장

```
In [60]: import pandas as pd
```

```
In [61]: cars93 = pd.read_csv("d:/data/cars93.csv")
```

```
In [62]: cars93
```

```
Out[62]:
```

	Unnamed: 0	Manufacturer	Model	...	Weight	Origin	Make
0	1	Acura	Integra	...	2705	non-USA	Acura Integra
1	2	Acura	Legend	...	3560	non-USA	Acura Legend
2	3	Audi	90	...	3375	non-USA	Audi 90
3	4	Audi	100	...	3405	non-USA	Audi 100
4	5	BMW	535i	...	3640	non-USA	BMW 535i
..
88	89	Volkswagen	Eurovan	...	3960	non-USA	Volkswagen Eurovan
89	90	Volkswagen	Passat	...	2985	non-USA	Volkswagen Passat
90	91	Volkswagen	Corrado	...	2810	non-USA	Volkswagen Corrado
91	92	Volvo	240	...	2985	non-USA	Volvo 240
92	93	Volvo	850	...	3245	non-USA	Volvo 850

```
[93 rows x 28 columns]
```


▶ Excel 파일로부터 데이터프레임 생성하기 : test.xlsx

```
pd.read_excel("d:/data/test.xlsx")
```

```
import pandas as pd  
test = pd.read_excel("d:/data/test.xlsx")
```

- ▶ 외부에 저장된 test.xlsx파일을 읽어들이 데이터프레임 생성
- 경로와 확장자를 정확히 지정

```
In [132]: import pandas as pd  
  
In [133]: test = pd.read_excel("d:/data/test.xlsx")  
In [134]: test  
Out[134]:
```

	id	class	gender	kor	math	eng
0	101	1	M	10	10	10
1	102	2	M	30	30	30
2	103	4	F	30	30	10
3	104	4	M	30	30	20
4	105	4	M	10	10	10
.....						

■ 내장된 데이터셋을 읽어 데이터프레임 생성

- 패키지에 내장된 데이터셋을 load 하여 데이터프레임을 이용할 수 있다.
- 파이썬의 seaborn 패키지에는 다양한 내장데이터가 있다.

```
In [66]: import seaborn as sns
```

```
In [67]: sns.get_dataset_names( )
```

```
Out[67]:
```

```
['anagrams',  
 'anscombe',  
 'attention',  
 'brain_networks',  
 'car_crashes',  
 'diamonds',  
 'dots',  
 'dowjones',  
 'exercise',  
 'flights',  
 'fmri',  
 'geyser',  
 'glue',  
 'healthexp',  
 'iris',  
 'mpg',  
 'penguins',
```

▶ seaborn 패키지 import하기

▶ sns.get_dataset_names() : seaborn 데이터셋 목록 확인

```
In [68]: iris = sns.load_dataset('iris')
```

```
In [69]: iris
```

```
Out[69]:
```

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
..
145	6.7	3.0	5.2	2.3	virginica
146	6.3	2.5	5.0	1.9	virginica
147	6.5	3.0	5.2	2.0	virginica
148	6.2	3.4	5.4	2.3	virginica
149	5.9	3.0	5.1	1.8	virginica

```
[150 rows x 5 columns]
```

- ▶ seaborn에 내장되어 있는 iris 데이터를 load하여 iris라고 저장한다

감사합니다