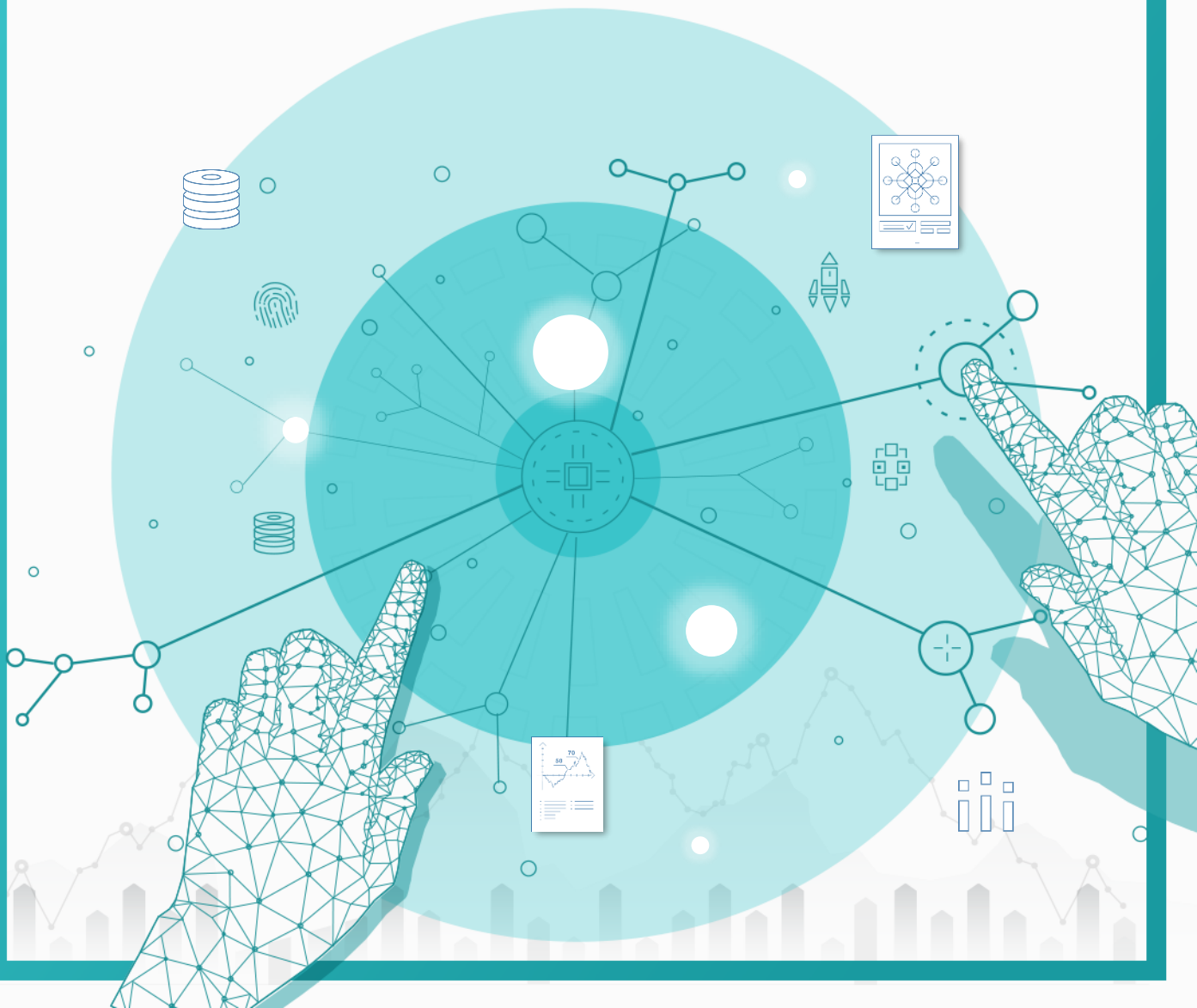
한국기술교육대학교
온라인평생교육원

파이썬 라이브러리로 하는 데이터 분석과 시각화

데이터 분석 기초



데이터 분석 기초

학습 목표

1. 데이터 수집의 개념을 설명할 수 있다.
2. 데이터 전처리의 개념을 설명할 수 있다.
3. 데이터 분석의 개념을 설명할 수 있다.

학습 내용

1. 데이터 수집의 이해
2. 데이터 전처리의 이해
3. 데이터 분석의 이해

1. 데이터 수집의 이해

1) 데이터 선정

(1) 수집할 데이터 선정

(예) 축구 승리에 영향을 미치는 요인이 있을까?

- 축구와 직접 관련된 데이터
  - 축구 기록 관련: 포메이션, 볼 점유율, 패스 성공률, 슈팅 시도, 태클 등
  - 축구 선수 관련: 연봉, 최근 경기 성적, 평균 나이 등
  - 그 외: 홈 원정 여부, 경기 시작 시간(낮·밤), 직전 경기의 승패 여부 등

- 축구와 직접 관련이 없는 데이터
  - 경기 당일의 날씨
  - 관중의 수, 응원단 유무
  - 유니폼의 색깔
  - 그 외 관련이 있을 것 같은 데이터

2) 데이터 수집 방법 선정

(2) 데이터 수집 방법 선정

▪ 데이터를 어디서 구할 수 있는가?

➡ 검색을 통해 수집할 데이터를 어디서 구할 수 있는지 확인함

- 어디서 제공해 주는가?
- 제공한다면 무료인가?
- 다운로드하는 파일 형식은 어떠한가?
- 수집한 데이터를 사용해도 법적으로 문제가 없는가?

▪ 데이터를 쉽게 구할 수 없다면?

➡ API나 CSV 파일 등 쉽게 다운로드할 수 있도록 데이터를 제공한다면 좋지만, 만약 그렇지 않다면?

- 웹 크롤링 등으로 필요한 데이터를 수집할 수 있는가?
- 다른 데이터를 활용하여 필요한 데이터를 생성할 수 있는가?

(예) 일자별 최고 기온 데이터가 없는 대신 시간대별 기온 데이터가 있다면, 직접 일자별 최고 기온 데이터를 생성할 수 있음

2) 데이터 수집 방법 선정

(2) 데이터 수집 방법 선정

- 수집할 데이터의 양은 충분한가?
  - 의미 있는 결과를 얻으려면 수집할 데이터는 한쪽에 편향되지 않아야 하며, 충분히 양이 많아야 함
  - 그렇지 않으면 잘못된 분석 결과가 나올 수 있음

(예) 평균 이용량이 가장 많은 지하철 호선은?

➡ 1호선은 7월, 2호선은 10월, 3호선은 6월, 학생들의 방학, 휴가철 등 여러 가지 변수가 있기 때문에 균형 있고 충분한 데이터를 수집한 뒤 분석해야 함

- 신뢰할 만한 데이터인가?
  - 4차 산업혁명, 인터넷 발달 등으로 무수히 많은 데이터 중 원하는 데이터를 수집하기가 쉬워졌음
  - 쉬워진 만큼 수집한 데이터가 신뢰할 만한 데이터인지 구별하는 것이 중요해짐

(예) 출처가 불분명한 데이터: 지식IN, 카페 글 등

3) 데이터 수집 도구 선정

(3) 데이터 수집 도구 선정

- 일회성 데이터인가? 주기적으로 데이터를 수집해야 하는가?

➡ 이에 따라 수집 방법 및 도구가 달라질 수 있음

- 일회성 데이터: 수집한 데이터를 CSV, TXT 등의 파일 형식으로 저장하여 활용하거나 필요할 때마다 데이터를 수집하여 그때그때 활용해도 됨

(예) 변하지 않는 데이터: 국가, 도시 정보 등

- 주기적으로 수집이 필요한 데이터: 필요할 때마다 데이터를 처음부터 끝까지 수집하기에 어려움이 있으므로 데이터베이스 활용, 자동화 시스템 구축 등을 통해 주기적으로 데이터를 수집·저장하는 과정이 필요함

(예) 변하는 데이터: 시계열 데이터 등

3) 데이터 수집 도구 선정

(3) 데이터 수집 도구 선정

- 수집한 데이터의 형식에 따른 도구 선정

➡ 데이터의 형식에 따라 수집 방법이 달라져야 함

- TXT 파일 형식의 로그 데이터: TXT 파일을 불러오기 위한 파이썬의 open 함수와 정형화된 데이터의 패턴을 기준으로 원하는 데이터만 추출하기 위해 파이썬 자료형, 정규표현식 등의 문법을 활용

- 홈페이지 내 특정 데이터: 홈페이지 소스를 가져오기 위한 파이썬의 Requests와 같은 HTTP 관련 모듈과 홈페이지 소스에서 특정 데이터를 추출하기 위한 BS4, Selenium과 같은 모듈 활용

```
<!doctype html><html lang="ko" dir="ltr" itemscope itemtype="http://schema.org/Search

 Copyright The Closure Library Authors.
 SPDX-License-Identifier: Apache-2.0
*/
'use strict';var a=!1,e=window,l=e.performance,n=m();e.cc_latency_start_time=l&&l.now
e.onaft=function(c){p("aft");if(!a&&e.ebp){var f=e.ebp(c);if(null!==f){c=f.source;var
e.l=function(c){function f(b){var d={};d[b]=m();e.cc_latency.push(d)}function h(b){va
!k))return!1;if(!d.getBoundingClientRect)return!0;g=d.getBoundingClientRect();d=g.lef
e.aft_counter.indexOf(b),-1!==b&&(b=1===e.aft_counter.splice(b,1).length,0===e.aft_cc
l('TvbFac')</script><script nonce="MgyEXsT54JaMQr1kApukkw">var _F_cssRowKey = 'boq-se
/*# sourceURL=/_/mss/boq-search/_/ss/k=boq-search.VisualFrontendUi.By7pLIWUURI.L.B1.C
```

3) 데이터 수집 도구 선정

(3) 데이터 수집 도구 선정

- 수집한 데이터의 형식에 따른 도구 선정
  - 항상 똑같은 형식, 개수의 데이터: 관계형 데이터베이스 등을 통해 데이터를 수집하고 저장하면 추후에 수집한 데이터를 다시 불러와서 사용하기에 편리함
  - 언제든 새로운 항목이 추가될 수 있는 데이터: JSON과 같은 파일 포맷의 데이터를 저장하여 언제든 새로운 항목을 추가, 수정, 삭제할 수 있음

```
{
    "이름": "홍길동",
    "나이": 25,
    "성별": "여",
    "주소": "서울특별시 양천구 목동",
    "특기": ["농구", "도술"],
    "가족관계": {"#": 2, "아버지": "홍판서", "어머니": "춘섬"},
    "회사": "경기 수원시 팔달구 우만동"
}
```

3) 데이터 수집 도구 선정

(3) 데이터 수집 도구 선정

- 무조건 코딩으로 데이터를 수집해야 하는가?

➡ 모든 데이터를 파이썬의 모듈을 활용해 수집할 필요는 없음

쉽게 수집 가능한 데이터라도 코딩이 필요한 경우

① 하루에 한 번씩, 한 시간에 한 번씩 클릭해야 할 때
② 정확한 시간에 수집해야 할 때
③ 하루 이틀이 아닌 한 달, 6개월, 1년처럼 장기적으로 수집이 필요할 때

2. 데이터 전처리의 이해

1) 데이터 전처리의 개념

- 데이터 전처리란?
  - 데이터 정제, 데이터 전처리, 데이터 클렌징, 데이터 가공, 데이터 핸들링 등으로 다양하게 표현함
  - 수집한 데이터를 분석에 적합하게 만들기 위해(정제, 클렌징, 가공 등) 별도 과정을 거친다는 것을 의미함
  - 수집한 데이터와 분석하려는 목적에 따라 해당 데이터를 분석에 가장 알맞도록 정제하는 과정을 말함

➡ 데이터 전처리 역시 데이터 분석 과정의 일부임

2) 데이터 전처리의 필요성

- '콩 심은 데 콩 나고 팥 심은 데 팥 난다'처럼 아무 데이터나 넣고 분석하면 좋은 결과를 기대하기 어려움
- 데이터 분석에 좋다는 최신 기술을 사용하더라도 충분히 전처리 되지 않은 데이터를 입력한다면 좋은 결과를 얻을 수 없을 뿐만 아니라 왜곡된 분석 결과를 얻을 수도 있음
- 빅데이터 시대에 수집된 데이터가 정형화되고 신뢰할 만한 데이터만 있을 수는 없기 때문에 데이터 전처리 과정은 필수로 이루어져야 함
- 정형화된 데이터도 경우에 따라 데이터의 전처리가 필요할 수 있음

(예) 홍 길동, 길동 홍 등 같은 사람이 다르게 인식될 수도 있음

2) 데이터 전처리의 필요성

- 수집한 데이터 내에 이상 값들이 있다면 전처리 과정을 통해 데이터를 정제해야 함
  - 이상 값이란?
    - 결측 값: 특정 데이터가 없는 경우
    - 입력 오류: 입력하면서 오류가 발생한 경우
    - 데이터 처리 오류: 데이터를 수집하는 과정에서 잘못 가져온 경우
    - 그 외 정상적이지 않은 데이터들: 측정, 실험 등의 오류 값 등
  - 이상 값들을 처리하지 않으면?
    - 편향되거나 오차 분산이 커지는 등 데이터 분석 결과가 크게 달라질 수 있음

3) 데이터 전처리의 종류

- 전처리를 하는 방법은 수집된 데이터의 특징과 분석 목적에 따라 여러 가지로 나뉘며, 파이썬의 문법, 모듈 등을 활용할 수 있음
  - 이상 데이터 제거
  - 데이터 통합
  - 데이터 변환(정규화, 요약, 이산화 등)
  - 데이터 축소, 특징 추출

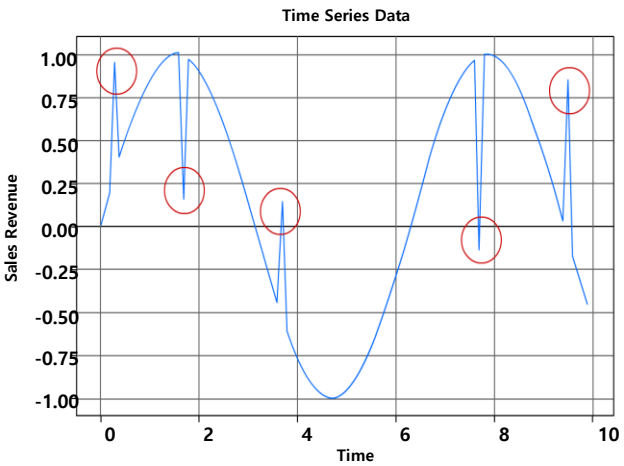
(1) 이상 데이터의 예시

- 오류, 실수 등으로 인해 값이 없거나 이상한 값들이 들어가 있는 경우

(예) 시계열 데이터에서의 이상 데이터

번호	이름	나이
1	홍길동	24
2	이길동	63
3	김길동	9142
4	최길동	-5.2
5	박길동	-

[이상 데이터 예시]



* 출처: Aditya Bhattacharya, 시계열 이상 탐지를위한 효과적인 접근 방법, https://vo.la/WkSgl

3) 데이터 전처리의 종류

(2) 데이터 통합의 예시

- 관련 있는 데이터들을 통합하여 일관성 있는 데이터 형태로 변환하는 과정

번호	이름	나이
1	홍길동	22
2	이길동	19
3	김길동	34
4	최길동	56
5	박길동	42

이름	성별	혈액형
김길동	남	A
홍길동	여	A
최길동	여	O
박길동	여	AB
이길동	남	B

➡ 두 테이블의 데이터를 하나로 통합하여 활용할 수 있음

(3) 데이터 변환의 예시

- 여러 데이터 간의 비교를 위해 정규화하는 과정 등 기존 데이터를 변환하여 활용

이름	타율	홈런
홍길동	0.245	34
이길동	0.321	20
김길동	0.333	11
최길동	0.297	5
박길동	0.258	40

➡

이름	타율	홈런
홍길동	0.245	0.828
이길동	0.321	0.714
김길동	0.333	0.171
최길동	0.297	0.0
박길동	0.258	1.0

3) 데이터 전처리의 종류

(4) 데이터 축소, 특징 추출의 예시

- 분석 목적에 필요 없는 데이터를 줄이거나 데이터에서 특징을 추출하고, 데이터의 차원을 축소하는 등

(예) 개와 고양이를 구분하기 위한 데이터

분류	꼬리 길이	눈동자 모양	다리 개수	유전자 형태
개				
고양이			분석 목적에 맞지 않는 데이터	
고양이				
개				
개				

3. 데이터 분석의 이해

1) 데이터 분석 방법

- 데이터를 분석하는 방법은 데이터의 종류, 데이터 분석의 목적, 분야, 분석 도구 등에 따라 다양함
- 수학, 통계학, 기계학습, 데이터 시각화 등 다양한 분석 방법이 있고, 파이썬, 엑셀, R, Matlab 등 다양한 도구를 활용해 분석할 수 있음

➡ 데이터 분석의 목적은 수많은 데이터 중에서 의사 결정 등에 도움이 되는 정보를 발견하고 이를 활용하여 가치를 창출하는 데 있음



*출처: FineReport '데이터 분석 방법 13가지 모음! 더 이상 망설일 필요가 없다!'

3. 데이터 분석의 이해

1) 데이터 분석 방법

(1) 통계 분석

- 통계를 기반으로 분석하는 방법으로 다양한 분석 기법이 있음
  - 회귀분석, 상관분석, 군집분석, 주성분분석 등

(2) 기계학습 분석

- 컴퓨터에 스스로 학습하고 문제를 해결하는 능력을 줌
  - 분류, 예측, 군집 등

(3) 시각화

- 분석 결과를 좀 더 자세하게 파악하기 위해 시각화하기도 하지만 여러 데이터 요소 간의 관계나 단순 나열된 데이터에서 알지 못한 인사이트를 시각화를 통해 발견할 수 있음

3. 데이터 분석의 이해

2) 데이터 종류에 따른 분석 방법

(1) 정량적 데이터 분석

- 데이터가 수치화된 형태일 때 분석하는 방법
- 객관적으로 데이터를 분석, 평가할 수 있음

➡ 주로 정형 데이터로 통계 분석 등을 적용할 수 있음

(2) 정성적 데이터 분석

- 숫자가 아닌 질적으로 평가되는 데이터를 분석하는 방법
- 서술 형태로 표현되는 범주형 데이터를 분석함

➡ 주로 비정형 데이터로 텍스트 내 빈도 분석, SNS 데이터 분석 등을 할 수 있음

3. 데이터 분석의 이해

3) 파이썬에서 데이터 수집 및 분석 방법

(1) 파이썬의 기본 문법 활용

- 파이썬의 기본 자료형, 기본 모듈을 활용하여 데이터를 수집, 전처리, 분석할 수 있음

(2) 파이썬의 외부 모듈 활용

- 잘 알려진 외부 라이브러리를 설치해 데이터를 수집 및 분석할 수 있음
  - 웹 크롤링: BS4, Selenium, Requests 등
  - 기계학습: Scipy, Tensorflow, Keras, Pytorch 등
  - 통계, 빅테이터 분석: Numpy, Pandas
  - 시각화: Matplotlib, Seaborn