

데이터통계분석

김영주

wrapup@dongyang.ac.kr

교수 소개

- 김영주
- 회사: (주) 푸드테크
- 부서: Data Intelligence

데이터 과학자	데이터 엔지니어	데이터 분석가
AI/머신러닝 모델 개발/최적화	데이터 플랫폼, 파이프라인의 아키텍처 구성 및 개발/운영	데이터 기반 의사결정을 위해 통계적 데이터 분석 및 시각화

Data Scientist
also known as Data Managers, statisticians.



A data scientist will be able to take data science projects from end to end. They can help store large amounts of data, create predictive modelling processes and present the findings.

Skills: Mathematics, Programming, Communication

Will use programmes such as:
SQL, Python, R

Data Engineers
also known as database administrators and data architects.



They are versatile generalists who use computer science to help process large datasets. They typically focus on coding, cleaning up data sets, and implementing requests that come from data scientists.

Skills: Programming, Mathematics, Big data

Will use programmes such as:
Hadoop, NoSQL, and Python

Data Analysts
also known as business Analysts.

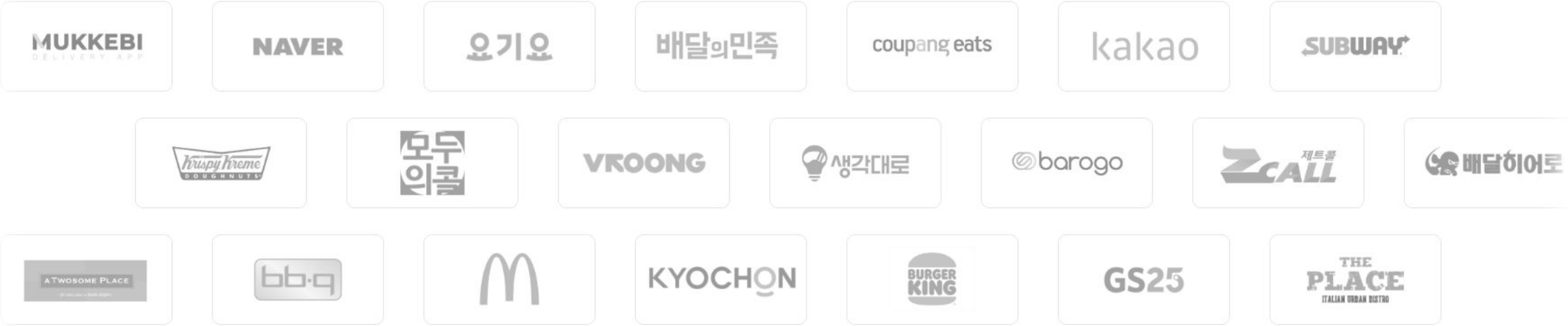


They typically help people from across the company understand specific queries with charts.

Skills: Statistics, Communication, Business knowledge

Will use programmes such as:
Excel, Tableau, SQL

<https://socrates-dissatisfied.tistory.com/5>



무엇을 배우나요? : 데이터 통계 분석

- 이론 → 수식 → 실습의 3단계로 데이터를 분석

기초 통계학에 대한 핵심적인 이론을 학습하고, 수식을 통해 개념이 의미하는 바를 파악합니다. 그런 다음, 파이썬으로 실습하면서 다양한 형태의 결과물(값, 표, 그림 등)을 살펴봅니다.

- 복잡한 수식 없이 데이터 분석의 기초를 학습

파이썬을 활용해 복잡한 수식이 없이도 통계분석을 학습할 수 있습니다.

- 파이썬 소스코드 제공

- 책에 수록된 파이썬 소스코드는 한빛 아카데미 홈페이지

(https://www.hanbit.co.kr/support/supplement_list.html)를 통해 다운받을 수 있습니다.

평가

평가 항목	평가 비율 (%)
중간 고사	30
기말 고사	40
과제	10
출결	20

강의 계획 (실습 언어 : Python)

1	교과목 소개 및 데이터의 개념과 변수	
2	[기본통계학] 1차원 데이터 - 지표 계산, 시각화 실습	
3	[기본통계학] 2차원 데이터 - 지표 계산, 시각화 실습	
4	추측통계의 기본	
5	이산형 확률변수와 대표적인 이산형 확률분포	과제 1 공지
6	연속형 확률변수 와 대표적인 연속형 확률 분포	과제 1 제출
7	중간고사 전 내용 정리 및 회고	
8	중간 고사	
9	독립동일분포	
10	통계적 추정	
11	통계적 가설검정	
12	회귀분석 - 단순회귀모형, 중회귀모형	과제 2 공지
13	회귀분석 - 모형의 선택, 모형의 타당성	과제 2 제출
14	기말고사 전 내용 정리 및 회고	
15	기말 고사	

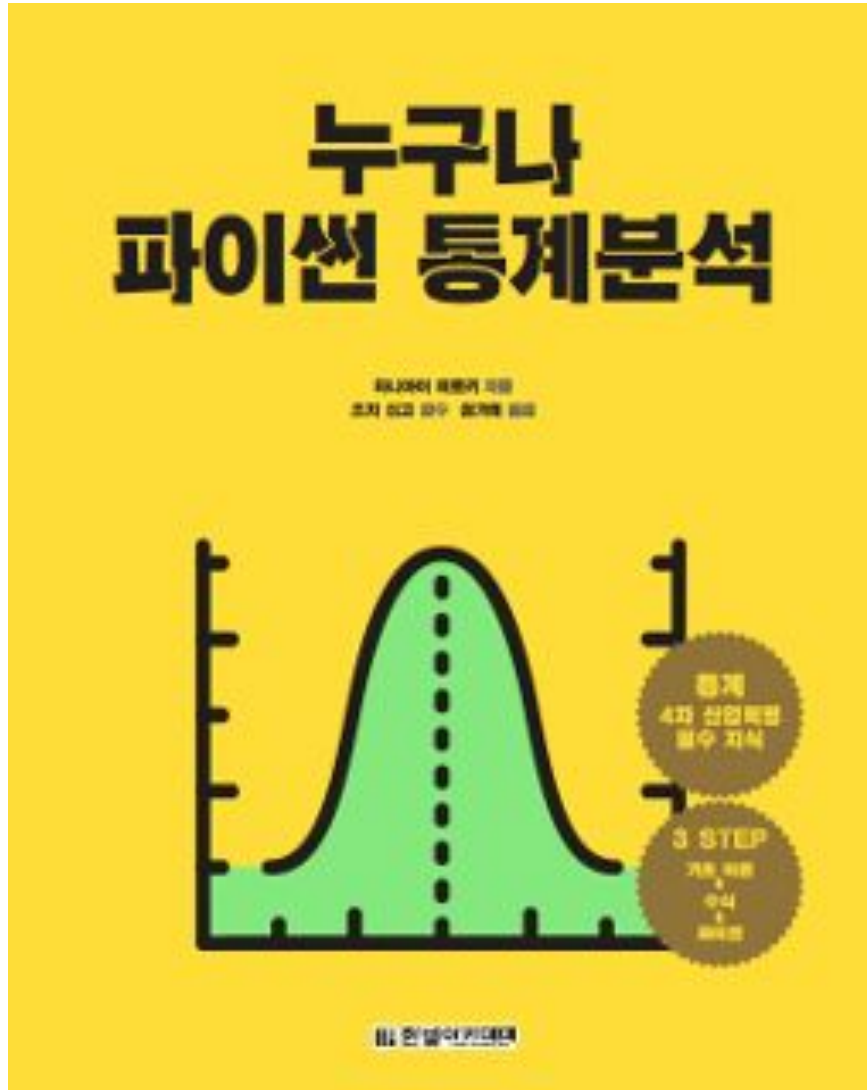
중간, 기말 고사

- **중간 고사, 기말 고사**

문제에 대한 답안 작성, 실기 시험(Jupyter notebook)

- **과제**

간단한 퀴즈 형식의 프로그래밍 과제



- **도서명:** 누구나 파이썬 통계분석
- **ISBN:** 979-11-5664-488-0 93310
- **저자:** 타니아이 히로키
- **감수자:** 츠지 신고
- **역자:** 권기태
- **출판사:** 한빛아카데미(주)
- **페이지:** 384p
- **정가:** 26,000원

CHAPTER 01



데이터

CONTENTS

- 1.1 데이터의 크기
- 1.2 변수의 종류
- 1.3 정리

데이터: 실습 데이터 살펴보기

표 데이터 처리에 특화된 통계분석 라이브러리 Pandas의 `read_csv` 함수를 사용하여 체력 테스트 결과를 담은 `csv` 파일을 읽어들이м

In [1]:

```
# Pandas를 pd라는 이름으로 임포트
import pandas as pd
```

In [2]:

```
# 학생번호를 인덱스로 하여 csv 파일을 읽어 들여, 변수 df에 저장
df = pd.read_csv( '../data/ch1_sport_test.csv',
                  index_col= '학생번호' )
```

```
# 변수 df를 표시
df
```

```
학생번호,학년,악력,윗몸일으키기,점수,순위
1,1,40.2,34,15,4
2,1,34.2,14,7,10
3,1,28.8,27,11,7
4,2,39.0,27,14,5
5,2,50.9,32,17,2
6,2,36.5,20,9,9
7,3,36.6,31,13,6
8,3,49.2,37,18,1
9,3,26.0,28,10,8
10,3,47.4,32,16,3
```

참고: 엑셀 포맷 - <https://xlsxwriter.readthedocs.io/worksheet.html>

데이터: 실습 데이터 살펴보기

체력 테스트 결과를 DataFrame으로 읽어들이는 결과

Out [2] :

학생번호	학년	악력	윗몸일으키기	점수	순위
1	1	40.2	34	15	4
2	1	34.2	14	7	10
3	1	28.8	27	11	7
4	2	39.0	27	14	5
5	2	50.9	32	17	2
6	2	36.5	20	9	9
7	3	36.6	31	13	6
8	3	49.2	37	18	1
9	3	26.0	28	10	8
10	3	47.4	32	16	3

학생번호,학년,악력,윗몸일으키기,점수,순위
1,1,40.2,34,15,4
2,1,34.2,14,7,10
3,1,28.8,27,11,7
4,2,39.0,27,14,5
5,2,50.9,32,17,2
6,2,36.5,20,9,9
7,3,36.6,31,13,6
8,3,49.2,37,18,1
9,3,26.0,28,10,8
10,3,47.4,32,16,3

데이터: 실습 데이터 살펴보기

DataFrame에서 악력에 대한 열을 추출하면 1차원 데이터 구조인 Series 반환

In [3] :

```
df[ ' 악력 ' ]
```

Out [3] :

학생번호

1	40.2
2	34.2
3	28.8
4	39.0
5	50.9
6	36.5
7	36.6
8	49.2
9	26.0
10	47.4

Name: 악력, dtype: float64

학생번호,학년,악력,윗몸일으키기,점수,순위

1	1	40.2	34	15	4
2	1	34.2	2	14	7
3	1	28.8	8	27	11
4	2	39.0	27	14	5
5	2	50.9	9	32	17
6	2	36.5	20	9	9
7	3	36.6	31	13	6
8	3	49.2	37	18	1
9	3	26.0	28	10	8
10	3	47.4	32	16	3

데이터: 데이터의 크기

DataFrame의 크기는 shape라는 인스턴스 변수 참조

In [4]:

```
df.shape
```

Out [4]:

```
(10, 5)
```

데이터 개수 10 (레코드 수), 변수 개수 5 (컬럼 수)

```
학생번호,학년,악력,윗몸일으키기,점수,순위
1,1,40,2,34,15,4
2,1,34,2,14,7,10
3,1,28,8,27,11,7
4,2,39,0,27,14,5
5,2,50,9,32,17,2
6,2,36,5,20,9,9
7,3,36,6,31,13,6
8,3,49,2,37,18,1
9,3,26,0,28,10,8
10,3,47,4,32,16,3
```

실습

데이터: 변수의 종류 - 질적변수와 양적변수

- 질적 변수 : 선택이 필요한 변수, 종류를 구별하기 위한 변수

1. 매우 좋음 2. 좋음 3. 보통 4. 나쁨 5. 매우 나쁨

A형 B형 O형 AB형

- 양적 변수 : 양을 표현하는 변수

데이터: 변수의 종류 - 척도수준

[표 1-1] 척도 수준

척도	예	대소 관계	차이	비
명의 척도	학생번호	x	x	x
순서 척도	성적 순위	o	x	x
간격 척도	온도	o	o	x
비례 척도	키	o	o	o

데이터: 변수의 종류 – 이산형변수와 연속형변수

- **이산형 변수**

- 하나하나의 값을 취하는 변수
- 서로 인접한 숫자 사이에 값이 존재하지 않음
- 주사위의 눈, 결석 횟수, 결석 학생 수

- **연속형 변수**

- 연속적인 값을 취할 수 있는 변수
- 어떤 두 숫자 사이에도 반드시 숫자가 존재
- 길이, 무게, 시간

Q&A

다음 강의

- 교재 : 02장 1차원 데이터 정리