Lecture 01: Introduction to Data Science & Python

Hyeryung Jang
(hyeryung.jang@dgu.ac.kr)
AI Department, Dongguk University

Instructor

- **Instructor: 장혜령**
  - 소속: 일반대학원 인공지능학과

장혜령 교수

최종학력	한국과학기술원 전기 및 전자공학부 공학박사		
전공분야	기계학습		
세부연구분야	그래프 기반 학습/추론, 확률/최적화 기반 기계학습, 지능형 네트워크		
E-mail	hyeryung.jang at dgu dot ac dot kr	**연락처**	공지예정
교수연구실	공지예정		

  - 약력:
    - 카이스트 전기및전자공학과 공학사/공학석사/공학박사
    - 영국 King’s College London Informatics 박사후연구원 (Post-doc)

Course Introduction

General Information

- Instructor: 장혜령
  - Email: **hyeryung dot jang at dgu dot ac dot kr**
  - Office: 만해관 1층 304-174호

- 수업 시간: 화 (실습), 목 (이론) 10:00 - 12:00
  - Office Hour: 화, 목 오후 (사전 이메일 연락)

- 수업 장소: (실습) 정보문화관 Q202 / (이론) 정보문화관 P404

- 수업 형태: e-class를 통한 비대면 실시간 Webex 수업
  - 실습, 이론 수업 모두 비대면 실시간 수업으로 진행
  - COVID 상황에 따라 대면-비대면 혼용, 혹은 대면수업 전환의 가능성이 있음

General Information

- Textbook:
  - Python Data Science Handbook: Essential Tools for Working With Data

  - ✓ Online page: https://jakevdp.github.io/PythonDataScienceHandbook/
  - ✓ Github code: https://github.com/jakevdp/PythonDataScienceHandbook

- Prerequisites:
  - Data Science 개론, Python programming 기초, Machine Learning and Data Science
- TA: **홍영준**
  - Email: yjayhong37@gmail.com

Syllabus

- Part 1: **Introduction to Data Science and Python (Week 1-3)**
  - Introduction to Data Science & Python, Environment Set-up & Python Basics
- Part 2: **Python for Data Analysis (Week 4-7)**
  - NumPy, Pandas & Web Crawling
- **Mid-term Exam (Week 8)**
- Part 3: **Python for Data Visualization (Week 9-10)**
  - Matplotlib, Seaborn & Advanced tools
- Part 4: **Machine Learning with Python (Week 11-13)**
  - Supervised, Unsupervised Learning & Recommender System
- **Project Presentation (Week 14)**
- **Final Exam (Week 15)**

Policy

- Grading Scheme
  - **출석 10%**: 출석 점수 총합 (보강 제외)
  - **중간고사 30%**: 4월 22일 목요일
  - **기말고사 40%**: 6월 10일 목요일
  - **과제 20 %**: 간단한 실습 숙제, 퀴즈, term project 발표 점수 총합

- Attendance
  - e-Class 출결 기준 출석 10점, 지각 5점, 결석 0점
  - 결석 혹은 지각의 사유가 있을 때에는 사전 이메일

- Mid-term and Final Exams
  - 시험 방식에 관하여는 추후 자세하게 공지

Policy

- Homework and/or Quiz
  - 매주 실습 수업 후 실습 내용 관련 숙제 혹은 퀴즈: 수업 후에 공지
  - 기한: 다음 실습 수업 전날 21:59분까지
  - e-class로 온라인 제출

- Term Project
  - **팀 프로젝트 (2인 1팀)**
  - Presentation 발표, report 제출 (week 14)
  - Peer-review 평가

- 강의 자료
  - 강의 후 다시보기 제공, 수업 자료는 e-Class로 제공

Overview of This Course

Python Programming for Data Science

- 데이터 사이언스를 위한 파이썬 프로그래밍

Python Programming for Data Science

- 데이터 사이언스를 위한 파이썬 프로그래밍 **환경 구축**
- 파이썬 라이브러리를 활용하여 **다양한 데이터를 효율적으로 가공, 변환, 처리, 분석하고 시각화하는 프로그래밍** 방법
- 대표적인 **머신러닝 알고리즘을 사용하여 데이터의 정보를 분석/표현하고 의미 있는 지식, 통찰력을 추출하여 의사결정하는 기술 및 구현** 능력

What is Data Science?

"The ability to take (usually large) data – to be able

- ✓ To **understand** it
- ✓ To **process** it
- ✓ To **extract value** from it
- ✓ To **visualize** it
- ✓ To **communicate** it

– that's going to be a hugely important skill in the next decades"

Hal Varian, chief economist at Google and UC Berkeley professor of information sciences, business, and economics

What is Data Science?

- “데이터” 를 다루는 여러 분야에서
- 새로운 **질문**을 하고,
- **데이터를 효과적으로 다루는 기술**을 활용하여
- **유용한 지식 (knowledge)이나 통찰력 (insights)을 추출**하여,
- 데이터 기반의 **대답 혹은 의사결정**을 가능케 하는 능력

annotation, cleaning, organizing, storing, analyzing and visualizing

What is Data Science?

- Example of Data Science in Banking: **Fraud Detection**

- Data science와 Machine learning을 활용하여 비정상적인 사기를 감지
  - 그동안의 금융 사기 관련 데이터를 분석하여
  - 비정상적인 금융 사기 패턴을 파악함으로써
  - 어떠한 금융 활동의 이상 여부를 정확하게 판단하는 모델을 학습하고
  - 잘 훈련된 모델을 사용해 실시간으로 금융 사기를 감지

What and Why Data Science?

- Data Science in Today

✓ Vast quantity of data, massively varied sources

  - 이메일, 소셜 미디어, 센서 데이터, 감시 카메라, 스마트 기기, … and many more
  - 다양한 경로로, 다양한 종류의, 방대한 양의 데이터가 생성

✓ Computing power

  - CPU, GPU 등 컴퓨팅 능력의 향상
  - 효율적이고 효과적인 데이터 처리, 분석이 가능

- 데이터를 다루는 다양한 분야에서 각광받고 있음

What and Why Data Science?

- 데이터 사이언스를 위해 갖추어야 할 역량들

✓Statistics
  - 방대한 데이터를 모델링하고 요약하는 능력

✓Computer science
  - **데이터를 효과적으로 저장, 처리, 시각화하는 알고리즘을 설계하고 구현하는 능력**

✓Computational engineering
  - 주어진 domain에 관한 새로운 질문을 만들고, 적합한 분석 방법 (예를 들어, **machine learning**) 을 사용하여 대답을 얻어내고 이해하는 능력

✓Domain knowledge
  - 생물학, 의학, 공학, 사회학, 인문과학 등 '데이터'를 다루는 모든 응용 분야

Learning Objectives

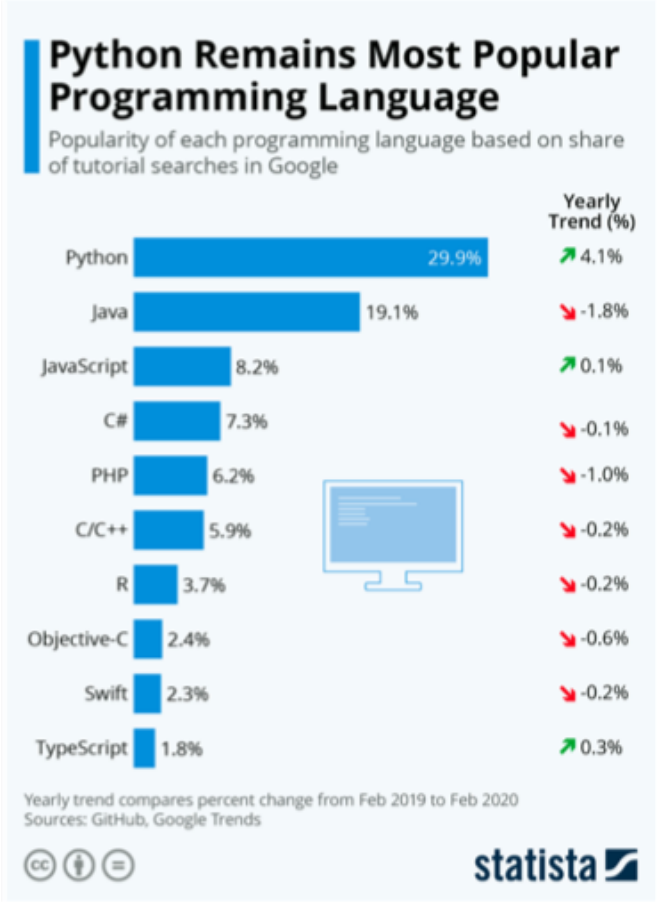
- You will be able to
  - Find useful datasets
  - Form research questions about the data
  - Perform basic data analysis to help answer your research questions
  - Present your findings

→ **Term project**

- Python을 이용해 데이터를 효과적으로 다루는 각종 라이브러리를 배우고 직접 사용할 수 있는 능력을 갖추는 것
  - 파이썬 프로그래밍 Basics에 익숙함을 가정
  - 데이터 사이언스가 무엇인지 알고 있음을 가정

Why Python?

- Python is the MOST popular programming language
  - Easy and Simple
  - Free and Open source
  - Interpreted
  - Large Standard Libraries
  - …



Why Python?

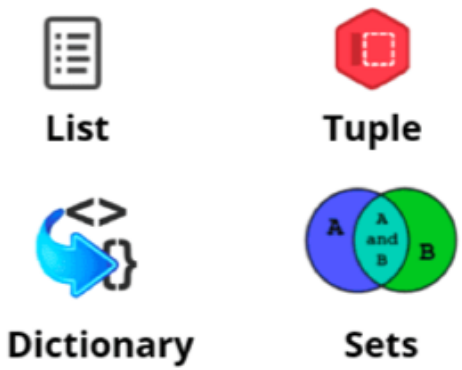
- 다양한 기능을 제공하는 다양한 패키지 구축
  - 배열 기반의 데이터를 처리하는 **NumPy**
  - 레이블이 붙은 데이터를 처리하는 **Pandas**
  - 시각화 도구를 제공하는 **Matplotlib**, **Seaborn**
  - 머신러닝을 위한 **Scikit-Learn**, **TensorFlow**, **PyTorch** 등
  - … and Much More!

Quick Overview

- Refresh basics of Python: Python Data Structures

```
[ ] def my_func(param1='default'):
        """
        Docstring goes here.
        """
        print(param1)

[ ] my_func()

    default

[ ] my_func(param1='new param')

    new param

[ ] def square(x):
        return x**2

[ ] out = square(2)

[ ] print(out)

    4
```

functions 함수



Quick Overview

- Learn about methodology, practices, and requirements to understand how to solve problem with data
  - How to **manipulate and analyze data** using Python libraries (NumPy and Pandas)

```
[4] import numpy as np
```

```
[5] np.random.randn(2,4)

    array([[-1.20725667,  0.96993926, -0.2696715 ,  0.13225716],
           [ 0.40032321, -0.3867656 , -0.25318173,  0.76155693]])
```

```
[7] arr = np.arange(0,5)
    arr

    array([0, 1, 2, 3, 4])
```

```
[9] np.sqrt(arr)

    array([0.        , 1.        , 1.41421356, 1.73205081, 2.        ])
```

```
[10] np.log(arr)

     /usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:1: Runtim
       """Entry point for launching an IPython kernel.
     array([      -inf, 0.        , 0.69314718, 1.09861229, 1.38629436])
```


```
[11] import pandas as pd
     import numpy as np
```

```
[12] from numpy.random import randn
     np.random.seed(101)
```

```
[13] df = pd.DataFrame(randn(5,4),index='A B C D E'.split(),columns='W X Y Z'.split()
```

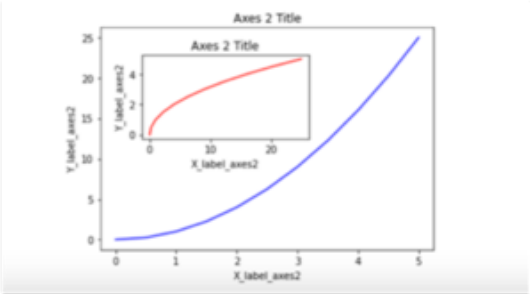
```
[14] df
```

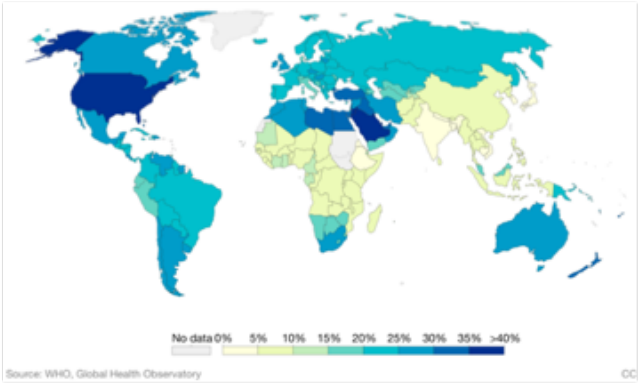
	W	X	Y	Z
A	2.706850	0.628133	0.907969	0.503826
B	0.651118	-0.319318	-0.848077	0.605965
C	-2.018168	0.740122	0.528813	-0.589001
D	0.188695	-0.758872	-0.933237	0.955057
E	0.190794	1.978757	2.605967	0.683509

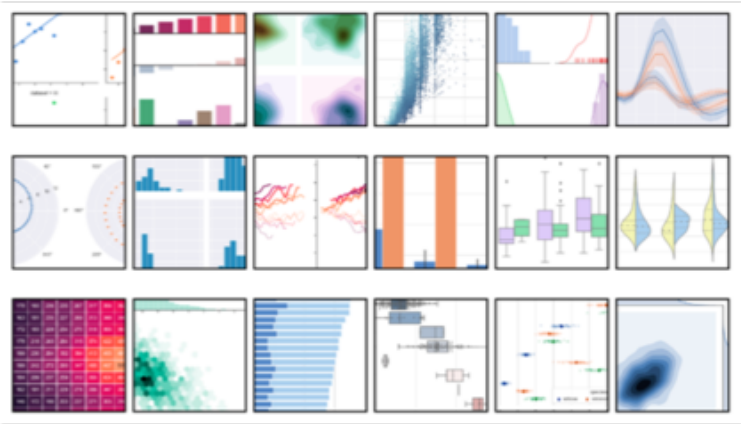
Quick Overview

- Learn about methodology, practices, and requirements to understand how to solve problem with data
  - How to **visualize data** using Python libraries (Matplotlib, Seaborn, Wordcloud, …)

Matplotlib





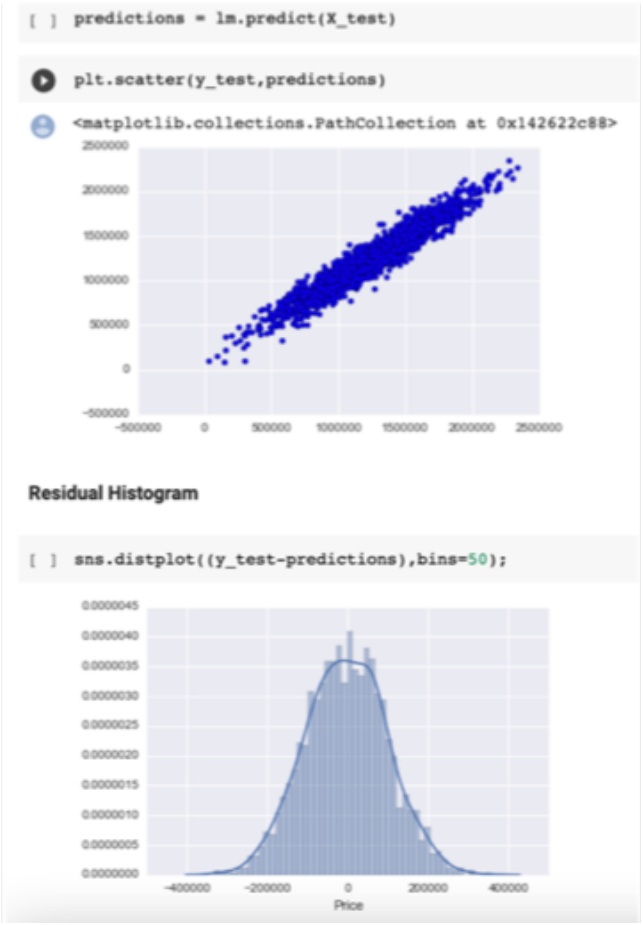


Seaborn

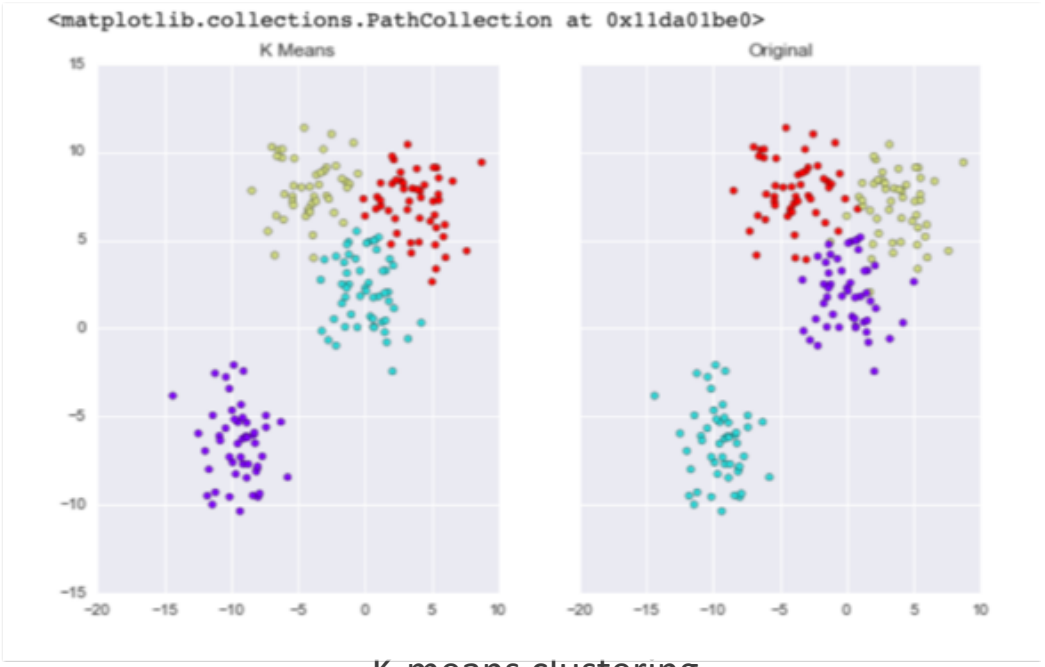
Wordcloud

Quick Overview

- Implement popular **Machine Learning (ML) algorithms** using Python



Regression



K-means clustering

In this lesson, you have learned:

- Data Science is the study of large quantities of data, which can reveal hidden insights that help us make strategic decisions and/or answers

- Data Science (usually) involves a little math, a little science, a little programming, and a lot of curiosity about data

- Quick overview of this course includes **“How to work with Data using Python?”**
  - NumPy, Pandas, Matplotlib, Seaborn, Scikit-Learn, ….

Thank you!

Any Questions?

hyeryung.jang@dgu.ac.kr