

ChatGPT와 함께 하는 데이터 분석

강사 소개

김영우

데이터저널 대표

前 히든그레이스 데이터분석팀장

데이터 분석 교육(Python, R, SPSS, AMOS, 데이터분석방법론)

youtube.com/dataholic4

facebook.com/groups/datacommunity





ChatGPT랑 친해지기

chat.openai.com

ChatGPT 똑똑하게 만들기

ChatGPT 똑똑하게 만들기 = 질문 잘하기

ChatGPT 똑똑하게 만들기 = 질문 잘하기

- 명확한 표현 사용하기

ChatGPT 똑똑하게 만들기 = 질문 잘하기

- 명확한 표현 사용하기
- 맥락 설명하기

ChatGPT 똑똑하게 만들기 = 질문 잘하기

- 명확한 표현 사용하기
- 맥락 설명하기
- 역할 부여하기, 어조 설정하기

특징 & Tip

특징 & Tip

- 같은 질문에도 매번 답변이 바뀐다
 - 같은 질문 여러 번 반복하기

특징 & Tip

- 같은 질문에도 매번 답변이 바뀐다
 - 같은 질문 여러 번 반복하기
- 대화 내용 기억하지만 대화 길어지면 잊어버린다
 - 새 대화창 열기

할루시네이션 조심하기!

할루시네이션(Hallucination)

할루시네이션(Hallucination)

- ChatGPT의 답변을 그대로 믿지 않기
- 오류가 없는지 검토하기
- 검색 도구가 아니라 창작 도구로 사용하기
- 코드 작성할 때도 주의하기

코드 만들기

ChatGPT의 다양한 기능들

파이썬이랑 친해지기

안녕, 파이썬?

킹왕짱 범용 프로그래밍 언어

- 오픈소스, 공짜
- 다양한 데이터 분석 기법

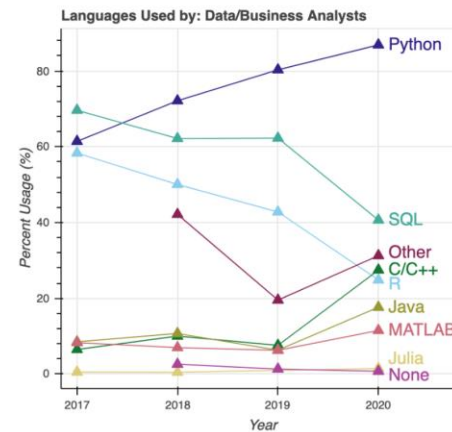
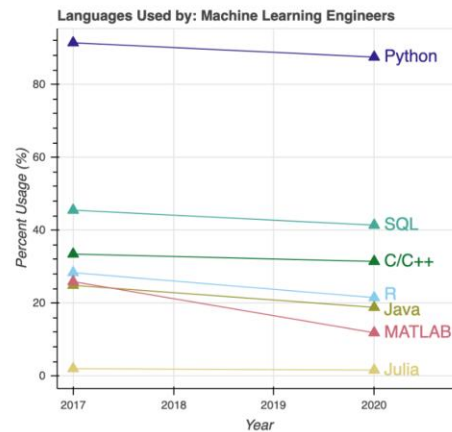
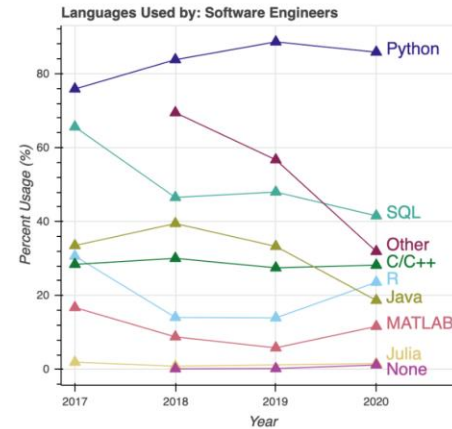
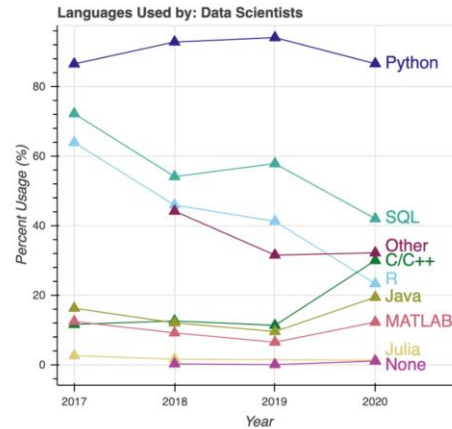
만든이

- 귀도 반 로섬(Guido van Rossum), 1991년
- 네덜란드의 프로그래머

쓰는 곳

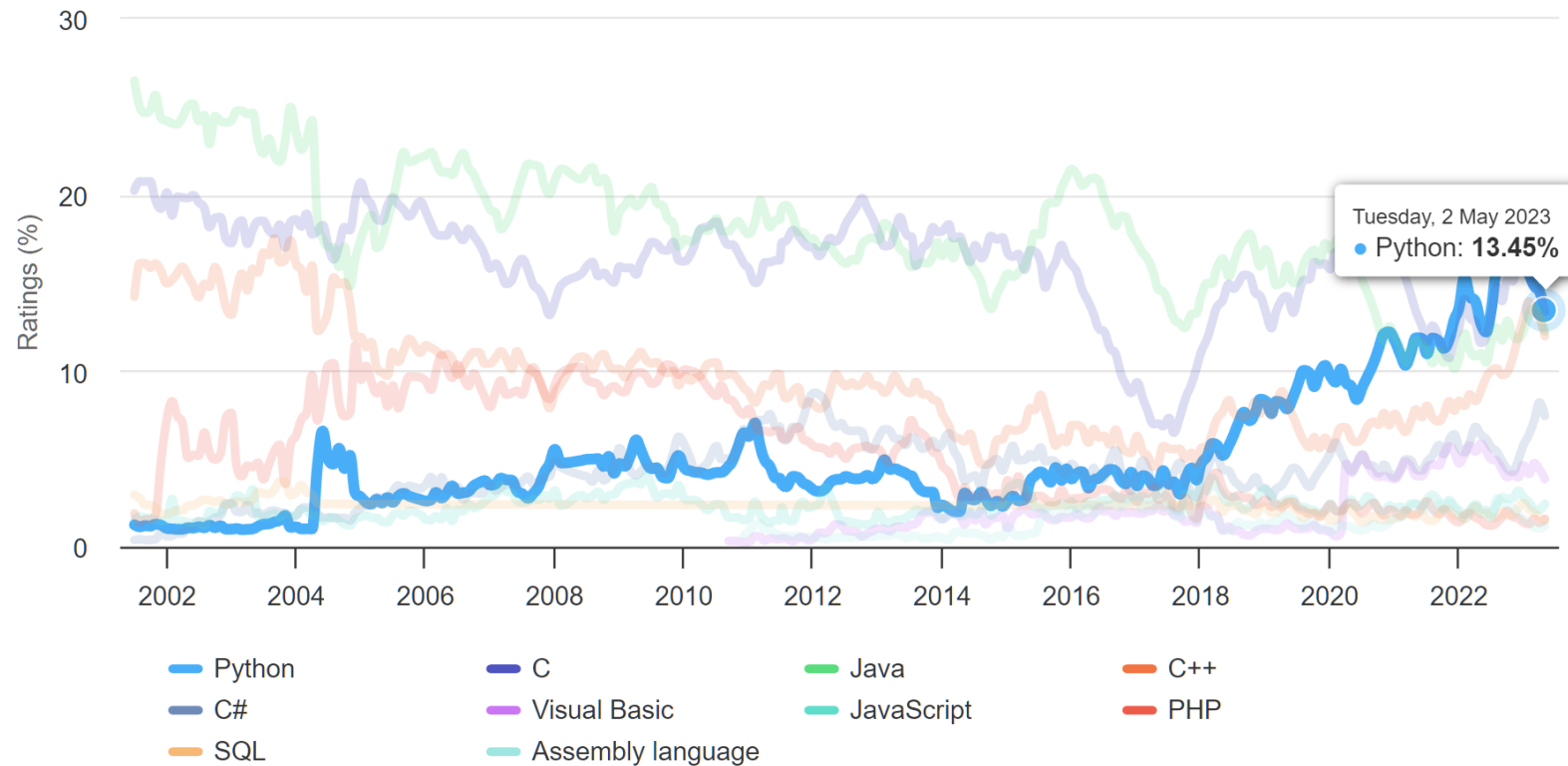
- Google, Facebook, Instagram, Netflix, Dropbox, Spotify...

kaggle



<https://towardsdatascience.com/data-science-trends-based-on-4-years-of-kaggle-surveys-60878d68551f>

TIOBE Index



<https://tiobe.com/tiobe-index>


장점

- 최신 분석 기법
- 무료 → 사용자가 많다 → 공부하기 좋다

Getting Started Prediction Competition


Titanic - Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

 Kaggle · 21,617 teams · Ongoing

OverviewDataCodeDiscussionLeaderboardRules


New Notebook

 Search notebooks

Filters

AllYour WorkShared With YouFavorites

Most Votes




Titanic Data Science Solutions

Updated 2y ago

1605 comments · Titanic - Machine Learning from Disaster

7080

Gold




Introduction to Ensembling/Stacking in Python

Updated 3y ago

977 comments · Titanic - Machine Learning from Disaster

5248

Gold




A Data Science Framework: To Achieve 99% Accuracy

Updated 3y ago

Score: 0.88516 · 571 comments · Titanic - Machine Learning from Disaster

4354

Gold




Exploring Survival on the Titanic

Updated 3y ago

Score: 0.80382 · 1012 comments · Titanic - Machine Learning from Disaster

3530

Gold




Titanic Tutorial

Updated 17d ago

943 comments · Titanic - Machine Learning from Disaster

3432

Gold

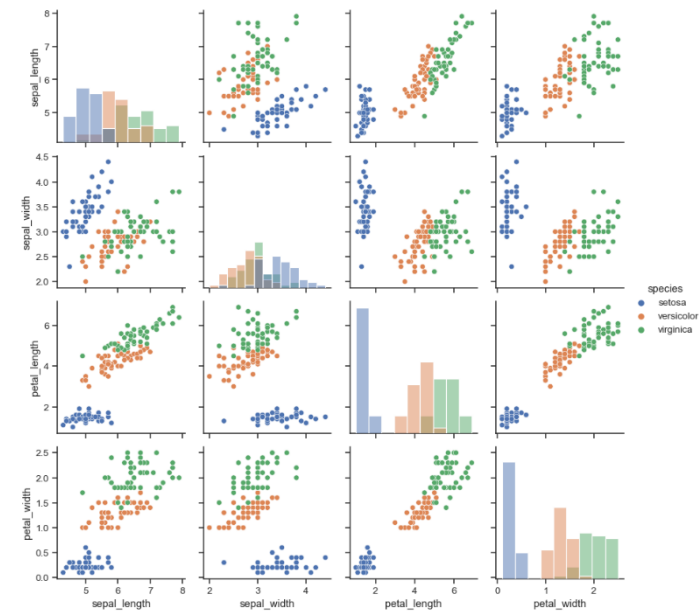
 Fast campus

Python을 어떻게?

Exploratory Data Analysis(EDA)

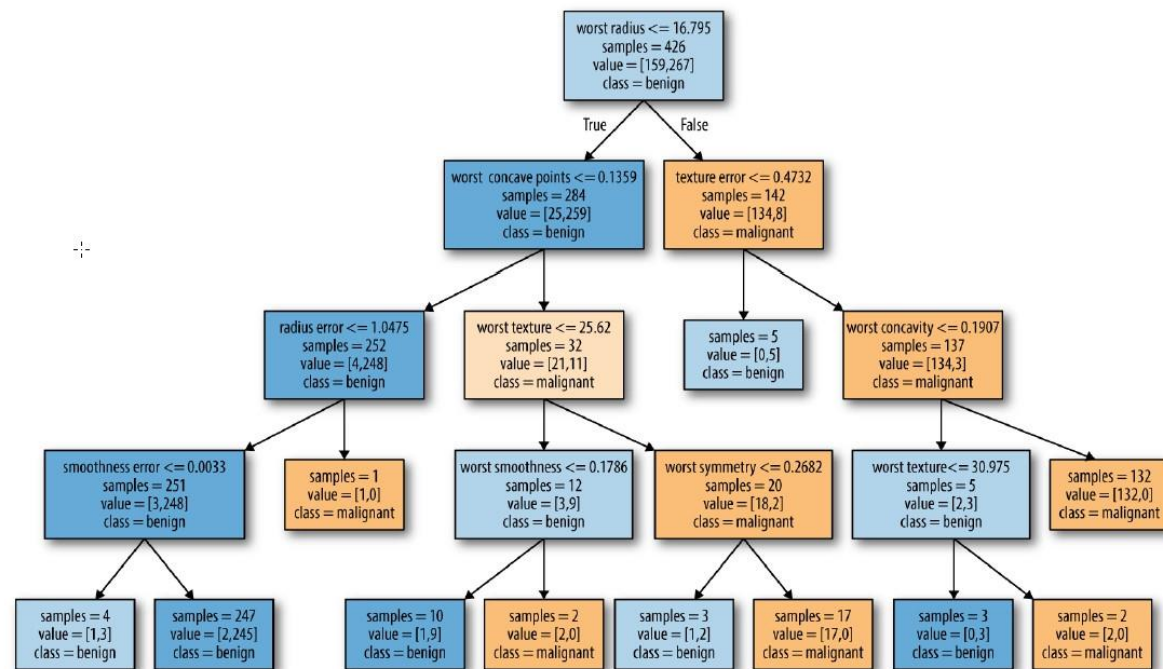
	Area Code	Item Code	Element Code	latitude	longitude	Y1961	Y1962
count	21477.000000	21477.000000	21477.000000	21477.000000	21477.000000	17938.000000	17938.000000
mean	125.449411	2694.211529	5211.687154	20.450613	15.794445	195.262069	200.782250
std	72.868149	148.973406	146.820079	24.628336	66.012104	1864.124336	1884.265591
min	1.000000	2511.000000	5142.000000	-40.900000	-172.100000	0.000000	0.000000
25%	63.000000	2561.000000	5142.000000	6.430000	-11.780000	0.000000	0.000000
50%	120.000000	2640.000000	5142.000000	20.590000	19.150000	1.000000	1.000000
75%	188.000000	2782.000000	5142.000000	41.150000	46.870000	21.000000	22.000000
max	276.000000	2961.000000	5521.000000	64.960000	179.410000	112227.000000	109130.000000

<https://www.shanelynn.ie/using-pandas-dataframe-creating-editing-viewing-data-in-python/>



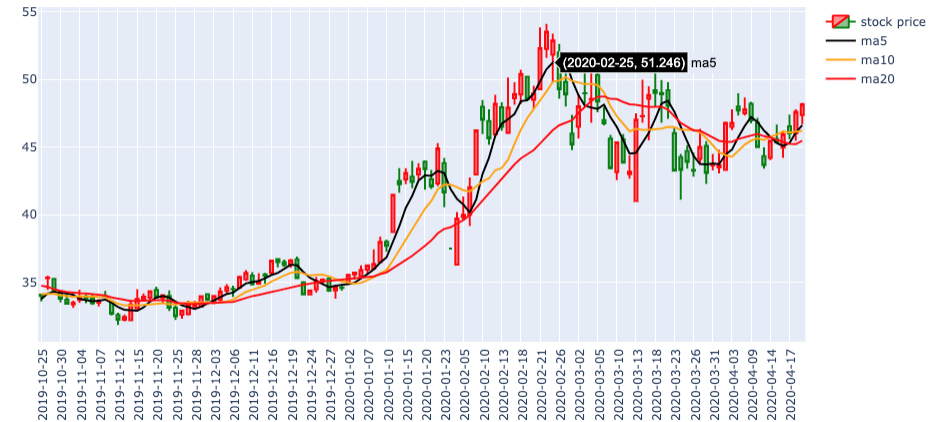
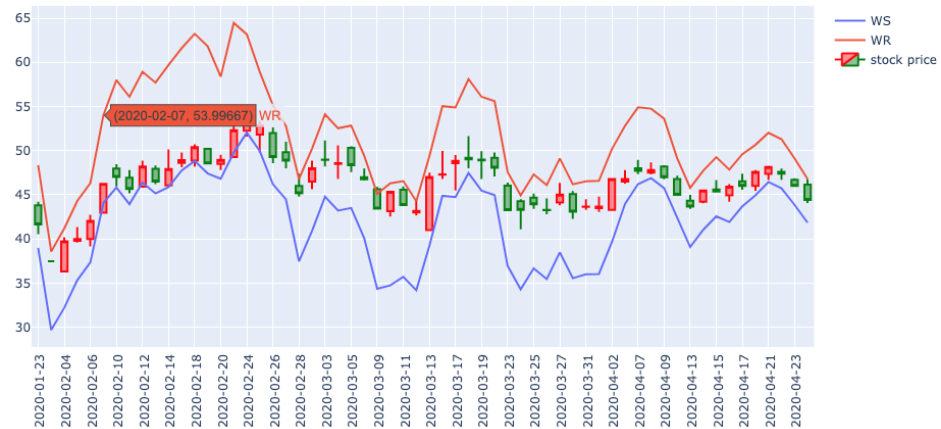
https://seaborn.pydata.org/tutorial/axis_grids.html

Predictive Analysis



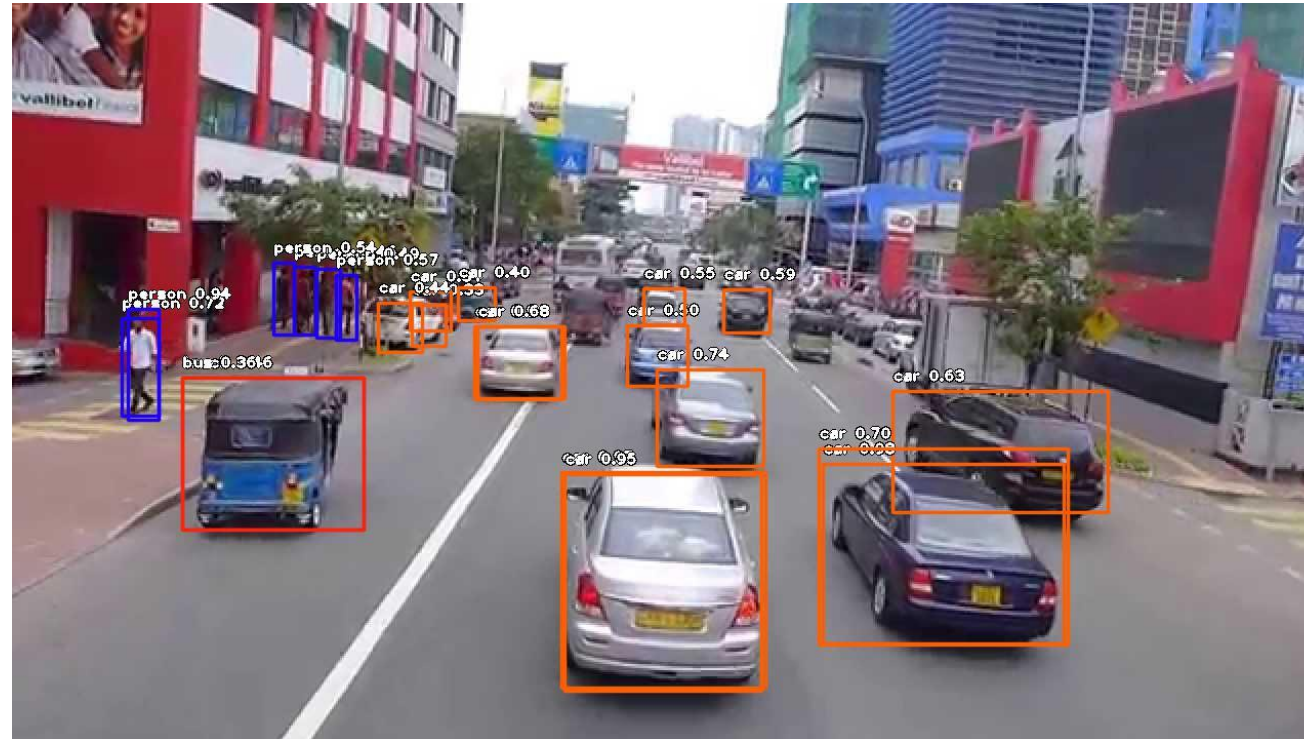
<https://www.thetopsites.net/article/58561179.shtml>

주식



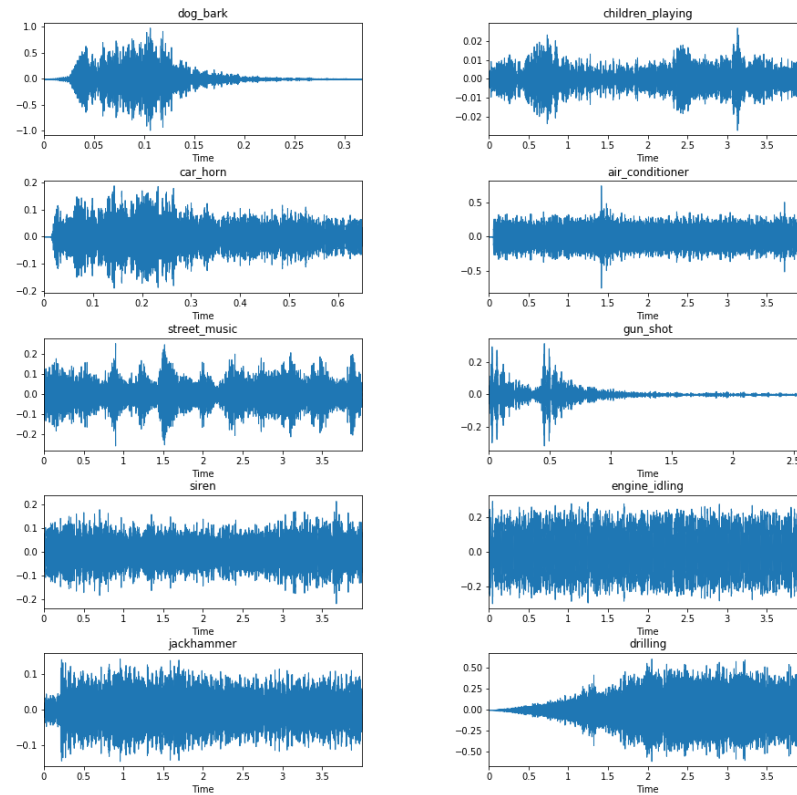
<https://github.com/charlesdong1991/StockInsider>

이미지



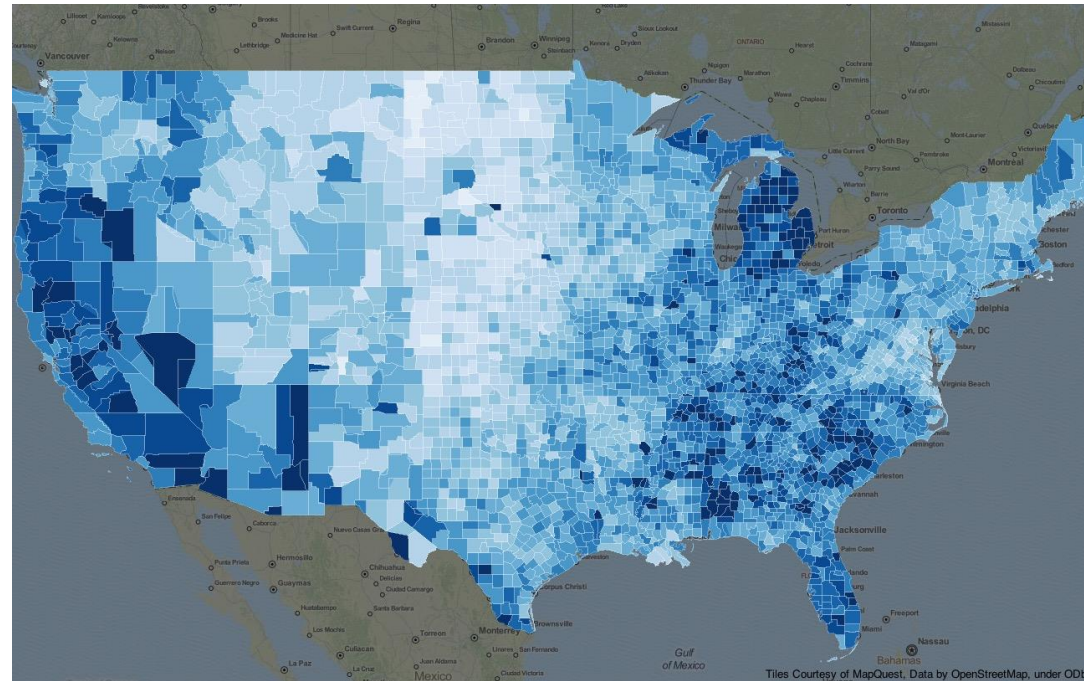
<https://stackabuse.com/object-detection-with-imageai-in-python/>

사운드



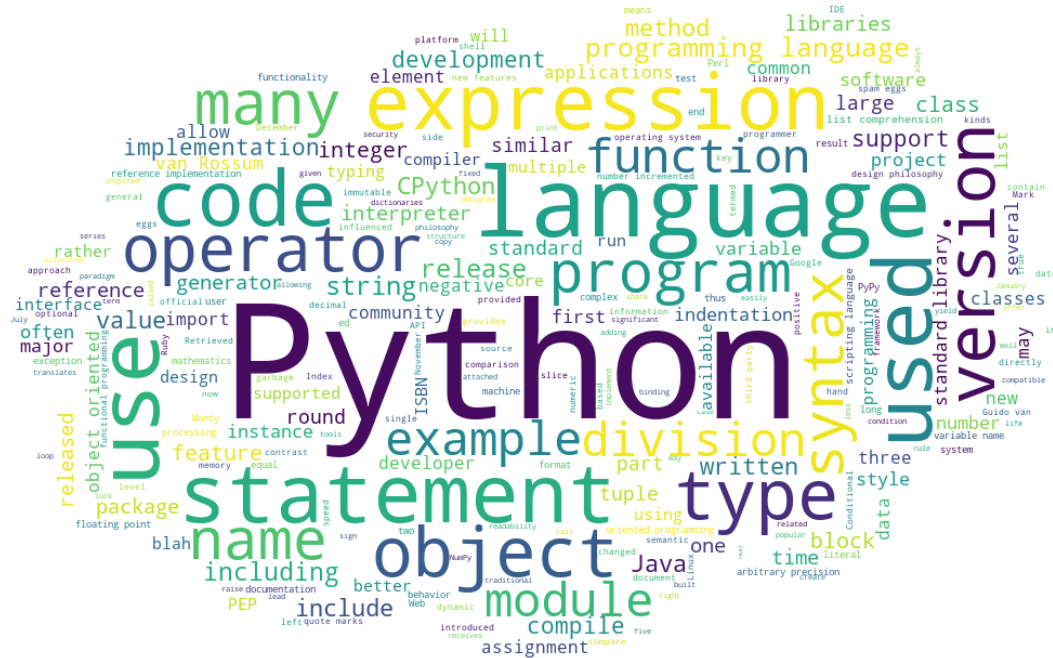
<https://towardsdatascience.com/how-to-apply-machine-learning-and-deep-learning-methods-to-audio-analysis-615e286fcbbc>

Map Data Visualization



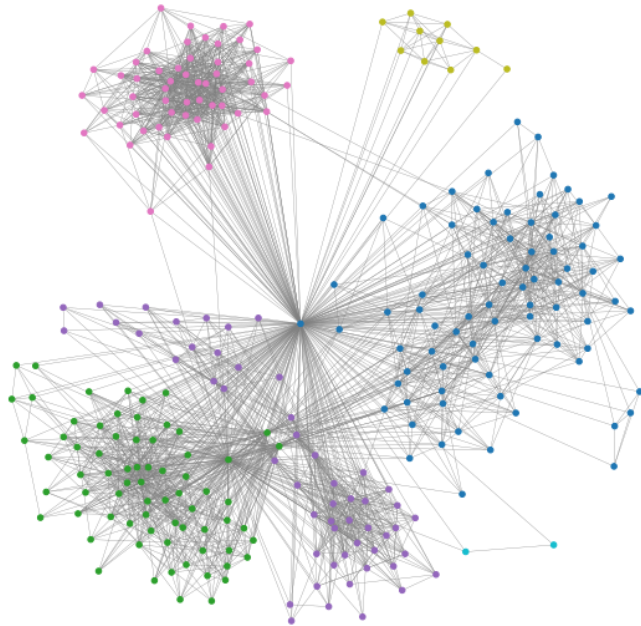
<https://deepai.org/publication/geoplotlib-a-python-toolbox-for-visualizing-geographical-data>

텍스트 마이닝

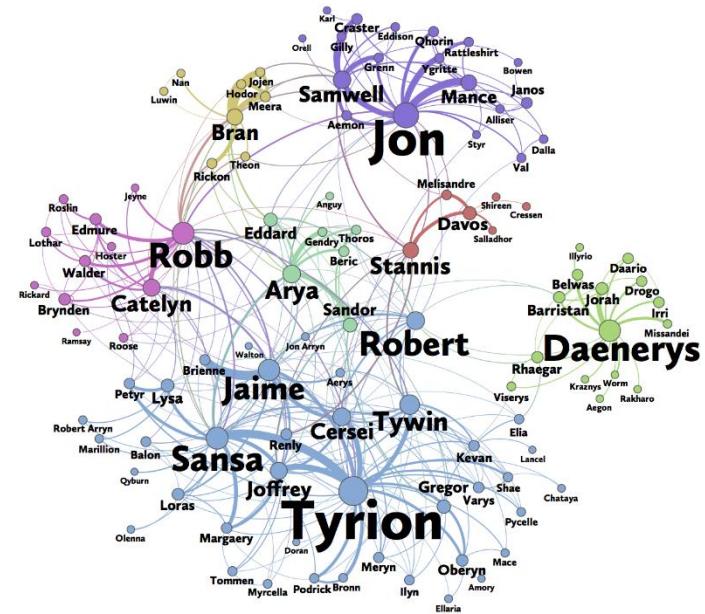


<https://cppsecrets.com/users/165711297114971154610611710811764103109971051084699111109/Using-WordCloud-lib-to-create-a-python-wordcloud.php>

Network Analysis



<https://www.databentobox.com/2019/07/28/facebook-friend-graph/>



<https://pub.towardsai.net/social-network-analysis-of-game-of-thrones-in-networkx-ff21ef65dc11>

Dashboard



<https://dash.gallery/dash-world-cell-towers/>

파이썬 설치하기

anaconda.com/download

JupyterLab 다루기

-
- **JupyterLab 바로가기 만들기**
 - **워킹 디렉터리 설정하기**

C:\Users\USER\anaconda3\Scripts

무작정 따라해보기!

파이썬 기초 문법

변하는 수, '변수' 이해하기

소득	성별	학점	국적
1000만원	남자	3.8	대한민국
2000만원	남자	4.2	대한민국
3000만원	여자	2.6	대한민국
5000만원	여자	4.5	대한민국

소득	성별	학점	국적
1000만원	남자	3.8	대한민국
2000만원	남자	4.2	대한민국
3000만원	여자	2.6	대한민국
5000만원	여자	4.5	대한민국

소득	성별	학점	국적
1000만원	남자	3.8	대한민국
2000만원	남자	4.2	대한민국
3000만원	여자	2.6	대한민국
5000만원	여자	4.5	대한민국

변수(Variable) = 변하는 수

소득	성별	학점	국적
1000만원	남자	3.8	대한민국
2000만원	남자	4.2	대한민국
3000만원	여자	2.6	대한민국
5000만원	여자	4.5	대한민국

- 하나의 속성, 다양한 값
- 분석의 대상
 - 남녀 중 누가 소득이 높은가?
 - 성별에 따라 학점이 다른가?
 - 학점이 증가할 수록 소득이 증가하는가?

변수(Variable) = 변하는 수

소득	성별	학점	국적
1000만원	남자	3.8	대한민국
2000만원	남자	4.2	대한민국
3000만원	여자	2.6	대한민국
5000만원	여자	4.5	대한민국

- Data : 변수들의 덩어리
- 분석 = 변수 간 관계 파악
- 분석 기법 = 변수간의 관계를 파악하는 방법

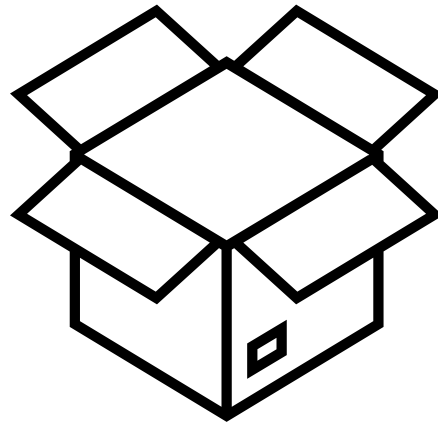
상수(Constant) = 안 변하는 수

소득	성별	학점	국적
1000만원	남자	3.8	대한민국
2000만원	남자	4.2	대한민국
3000만원	여자	2.6	대한민국
5000만원	여자	4.5	대한민국

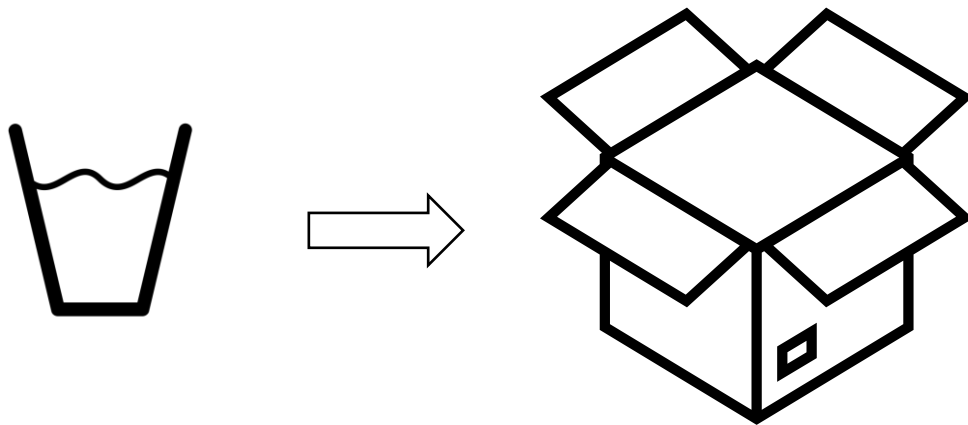
- 하나의 속성, 고정된 값
- 분석의 대상 X
 - 국적에 따른 소득 차이???
 - 국적에 따른 여성 비율???

마술 상자 같은 '함수' 이해하기

함수 = Magic Box

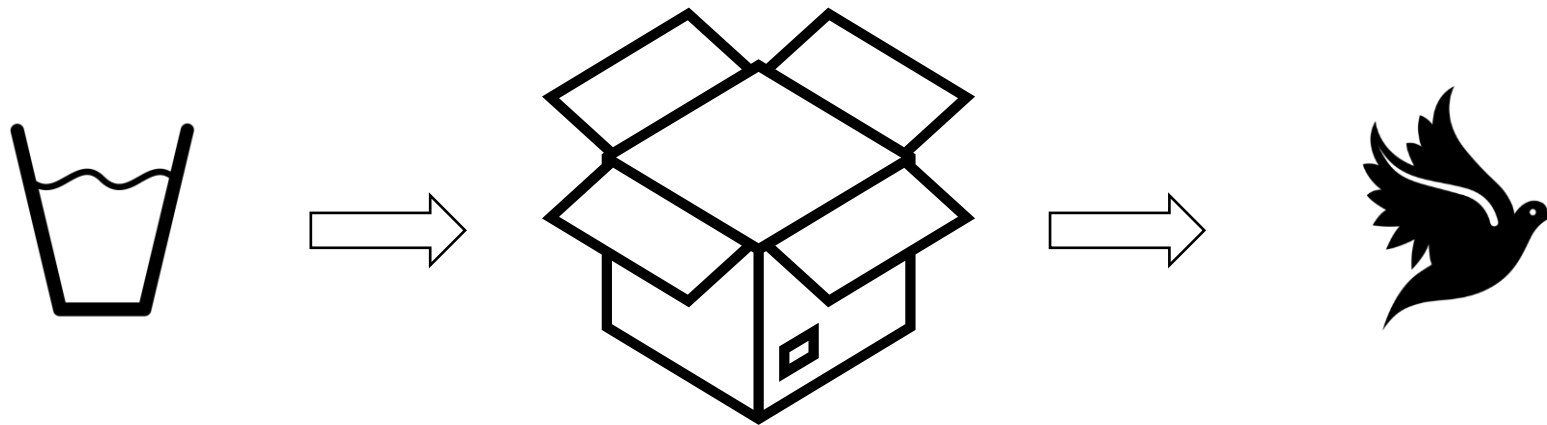


함수 = Magic Box



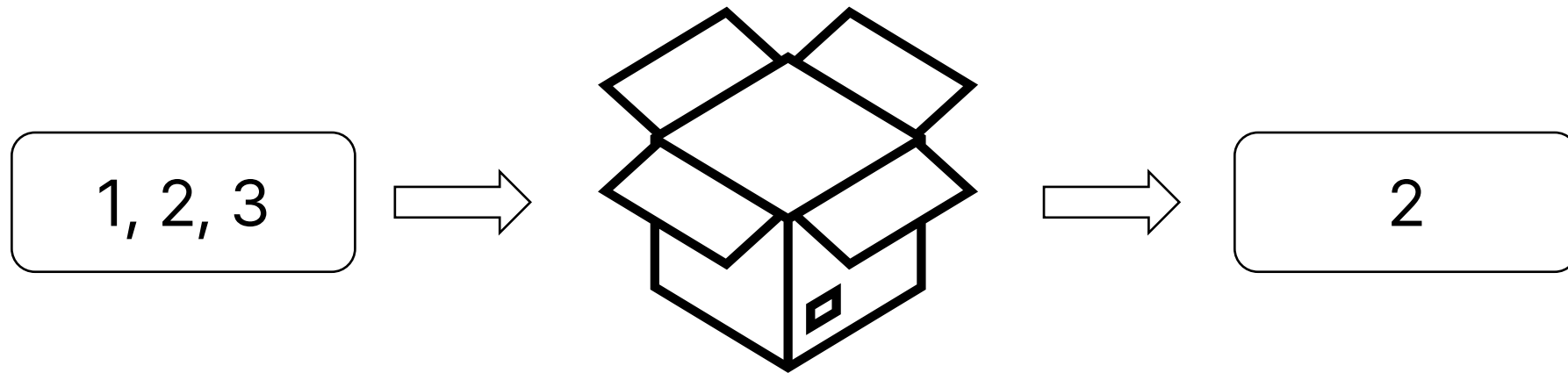
함수 = Magic Box

- 기능이 들어있음



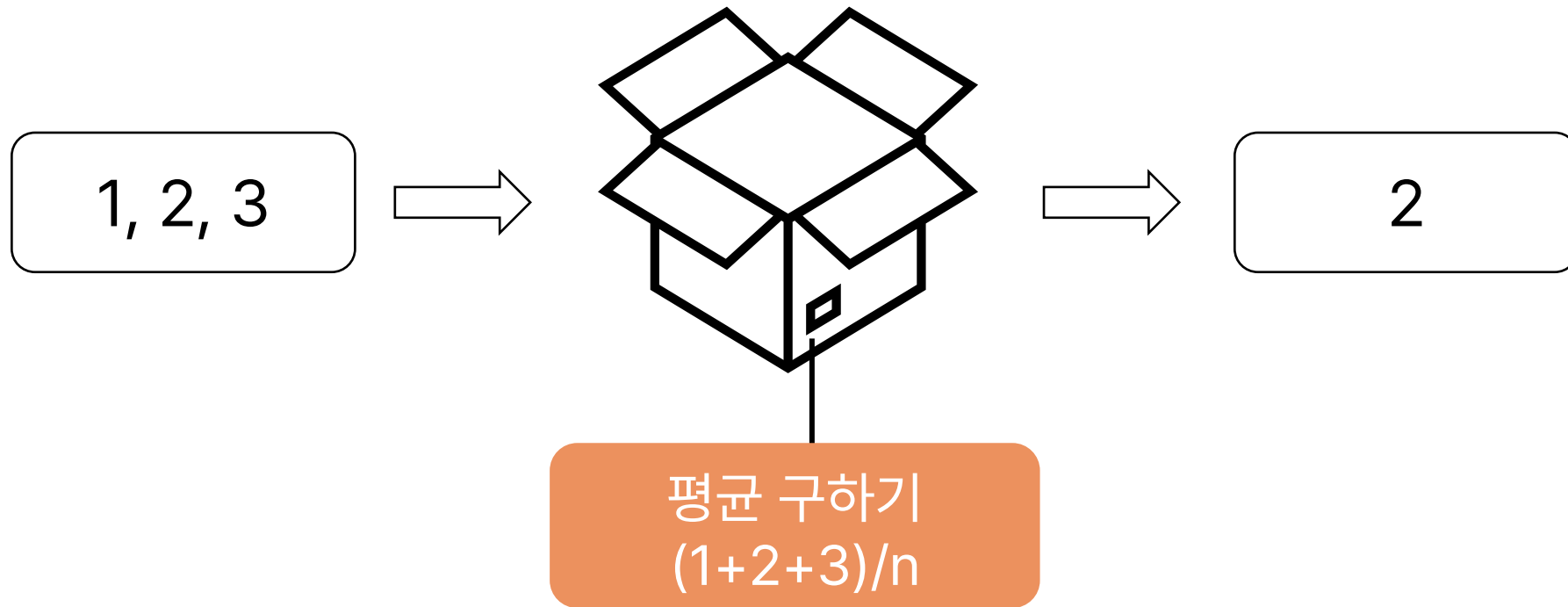
함수 = Magic Box

- 기능이 들어있음



함수 = Magic Box

- 기능이 들어있음
- 변수 넣으면 새로운 변수 탄생
- 분석 = 함수를 이용해서 변수를 조작하는 일



함수 꾸러미, '패키지' 이해하기

패키지(package) = 함수 꾸러미



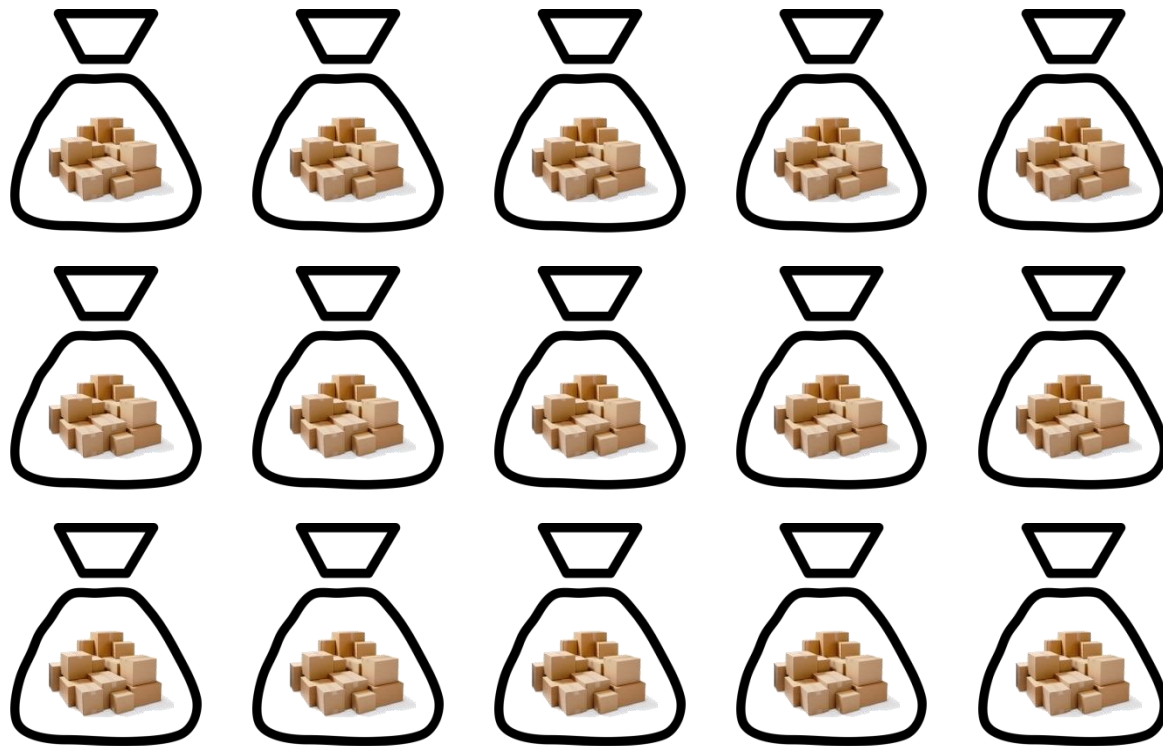
패키지(package) = 함수 꾸러미

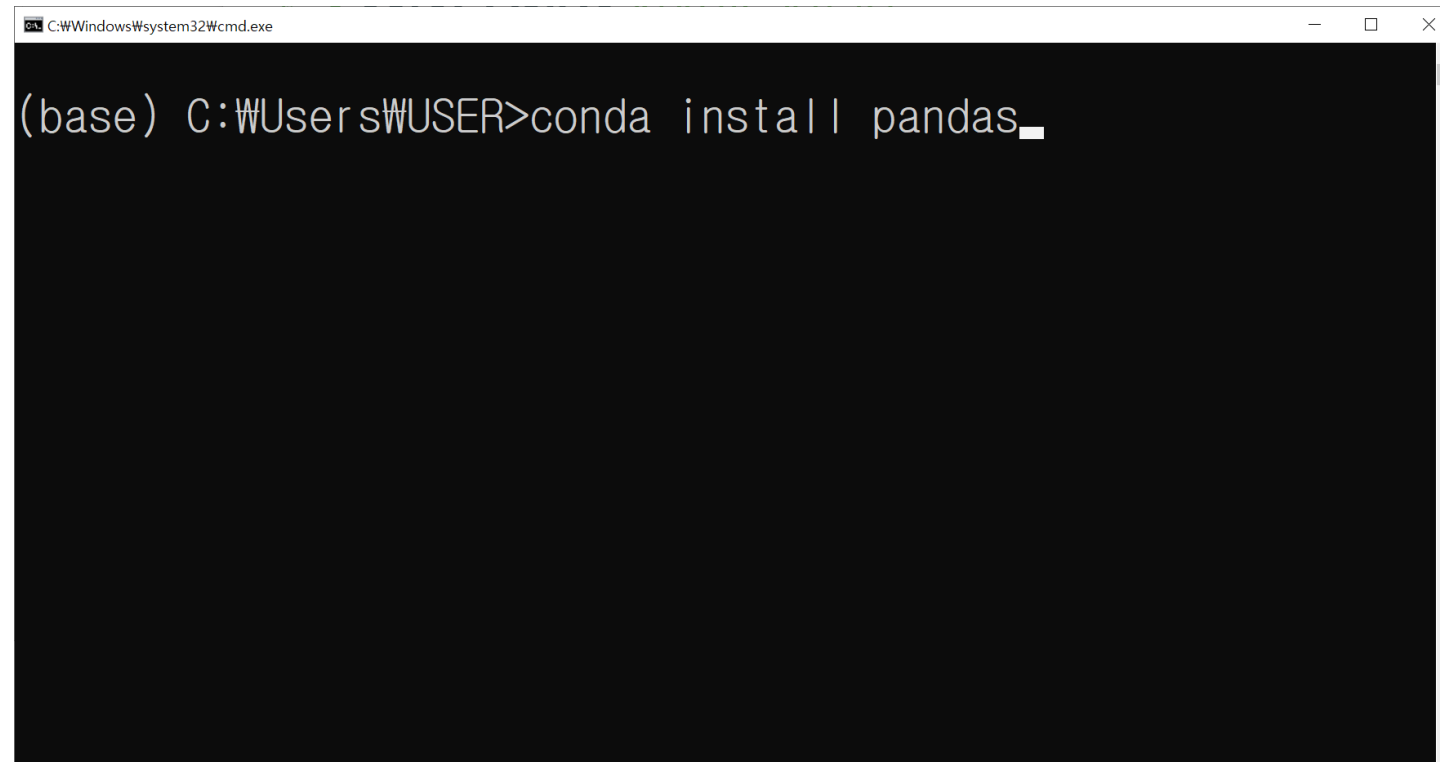
- 함수 쓰려면 반드시!
 - 설치 & 로드
 - 노트북 열 때마다
 - 내장 함수는 X
- 어플깰듯 입맛대로 설치
 - 필요한 기능에 따라
 - 최신 분석 기법 수시로 업로드
- ex) pandas
 - 데이터 핸들링 패키지
 - describe(), groupby(), agg(), sort_values(), merge()...



패키지(package) = 함수 꾸러미

- pypi 등록 프로젝트 > 20만개





```
C:\Windows\system32\cmd.exe
(base) C:\Users\WUSER>conda install pandas_
```

ChatGPT와 파이썬으로 데이터 갖고 놀기

데이터 프레임의 세계로

데이터 프레임(Data Frame)

소득	성별	학점	국적
1000만 원	남자	3.8	대한민국
2000만 원	남자	4.2	대한민국
3000만 원	여자	2.6	대한민국

데이터 프레임(Data Frame)

소득	성별	학점	국적
1000만원	남자	3.8	대한민국
2000만원	남자	4.2	대한민국
3000만원	여자	2.6	대한민국

- 가장 일반적인 데이터 형태
- 행(Row)과 열(Column)의 조합

데이터 프레임(Data Frame)

3개의 Row

1	2	3	4	5
소득	성별	학점	국적	
1000만원	남자	3.8	대한민국	
2000만원	남자	4.2	대한민국	
3000만원	여자	2.6	대한민국	

- 가장 일반적인 데이터 형태
- 행(Row)과 열(Column)의 조합

데이터 프레임(Data Frame)

		A	B	C	D	
		↓	↓	↓	↓	4개의 Column
		소득	성별	학점	국적	
3개의 Row	1	→	1000만원	남자	3.8	대한민국
	2	→	2000만원	남자	4.2	대한민국
	3	→	3000만원	여자	2.6	대한민국

- 가장 일반적인 데이터 형태
- 행(Row)과 열(Column)의 조합

데이터 프레임(Data Frame)

		A ↓	B ↓	C ↓	D ↓	4개의 Column
		소득	성별	학점	국적	
3개의 Row	1 →	1000만원	남자	3.8	대한민국	
	2 →	2000만원	남자	4.2	대한민국	
	3 →	3000만원	여자	2.6	대한민국	

- Row = 행, 한 사람의 데이터
 - 100명이면 100 Row
 - 활용예) Row가 몇이예요? Case가 몇 개예요?

데이터 프레임(Data Frame)

		A ↓	B ↓	C ↓	D ↓	4개의 Column
		소득	성별	학점	국적	
3개의 Row	1 →	1000만원	남자	3.8	대한민국	
	2 →	2000만원	남자	4.2	대한민국	
	3 →	3000만원	여자	2.6	대한민국	

- Column = 열, 변수
 - 변수가 100개면 100 Column
 - 활용예) Column이 몇 개예요? 변수가 몇 개예요?

데이터 프레임(Data Frame)

		A ↓	B ↓	C ↓	D ↓	4개의 Column
		소득	성별	학점	국적	
3개의 Row	1 →	1000만 원	남자	3.8	대한민국	
	2 →	2000만 원	남자	4.2	대한민국	
	3 →	3000만 원	여자	2.6	대한민국	

“데이터가 크다”

1. Row가 많다
2. Column이 많다

데이터 프레임(Data Frame)

		A	B	C	D
		↓	↓	↓	↓
		소득	성별	학점	국적
3개의 Row	1	1000만원	남자	3.8	대한민국
	2	2000만원	남자	4.2	대한민국
	3	3000만원	여자	2.6	대한민국

“데이터가 크다”

1. Row가 많다 → 컴퓨터가 버벅 → 고사양 장비 구축(upgrade)
2. Column이 많다

데이터 프레임(Data Frame)

		A	B	C	D	4개의 Column
		↓	↓	↓	↓	
		소득	성별	학점	국적	
3개의 Row	1	→	1000만 원	남자	3.8	대한민국
	2	→	2000만 원	남자	4.2	대한민국
	3	→	3000만 원	여자	2.6	대한민국

“데이터가 크다”

- 1. Row가 많다 → 컴퓨터가 버벅 → 고사양 장비 구축(upgrade)
- 2. Column이 많다 → 분석방법의 한계 → 고급 분석방법(machine learning)

데이터 분석 기초

자유자재로 데이터 가공하기

그래프 만들기

데이터 합치기

데이터 정제하기

실전! 데이터 분석 프로젝트

한국복지패널 데이터

GPT-4 날개 달기

GPT-4

- 고성능 모델, GPT-3.5 업그레이드 버전
- 모든 면에서 더 뛰어남
 - 응답 품질, 길이
 - 프롬프트 길이
 - 맥락 추론
 - 대화 내용 기억

부가 기능

- Advanced Data Analysis
- Browsing
- DALL·E
- GPTs
- Plugins

한계점

- 요청 횟수 제한
- 느린 속도
- 할루시네이션

Advanced Data Analysis

Advanced Data Analysis

- 파이썬이 내장된 ChatGPT
- 파일 업로드, 다운로드 가능
- 분석 요청하면 파이썬 코드 생성, 실행하여 결과물 보여줌
- 고급 분석

Advanced Data Analysis

- 한계점
 - 요청 횟수 제한
 - 파일 업로드 크기 제한
 - 시간 제한: 코드 실행, 세션
 - 패키지 사용 제약, 파이썬 버전 제약
 - 보안

할루시네이션

할루시네이션

데이터에 오류 있는지 모른 채 분석 요청

- ChatGPT는 오류 가능성 염두 안 하고 분석함

할루시네이션

데이터에 오류 있는지 모른 채 분석 요청

- ChatGPT는 오류 가능성 염두 안 하고 분석함

없는 데이터 분석 요청

- ChatGPT가 그럴듯한 값을 생성해서 분석함

할루시네이션

가장 위험한 사용법: 데이터와 분석 목표만 주기

ex) "데이터를 분석해서 인구와 GDP의 관계를 설명한 보고서를 만들어줘"

1. 단계별로 분석, 각 단계 오류 확인

1. 단계별로 분석, 각 단계 오류 확인

- 1) 분석 기획 - 목표, 방법
- 2) 데이터 탐색 - 빈도 분석, 요약 통계량
- 3) 오류 검토, 수정 - 결측치 처리, 이상치 처리
- 4) 요약표 만들기
- 5) 그래프 만들기
- 6) 추론 통계 분석, 머신러닝 모델링
- 7) 결과 해석

1. 단계별로 분석, 각 단계 오류 확인

- [View analysis] 코드 검토, 재분석 요청
- 결과 해석 검토, 재해석 요청

2. 분석 내용을 프롬프트에 입력, 오류 점검 요청

2. 분석 내용을 프롬프트에 입력, 오류 점검 요청

“너는 경력 10년차 데이터 분석가이자 데이터 분석 과정 검수 전문가다.

너는 데이터 분석 대회 심사 경력이 많다.

내가 입력한 [분석 내용]에 오류가 있는지 점검해서 알려줘.

특히 다음 요소를 평가해서 알려줘.

- 일반적인 데이터 분석 절차에 맞게 수행했는가?
- 분석 과정에 누락한 작업이 없는가?
- 존재하지 않는 데이터를 인위적으로 생성하지 않았는가?
- 분석 결과를 해석할 때 과도하게 일반화하거나 무리한 결론을 내리지 않았는가?

[분석 내용]

”

한계점

- 메시지 횟수 제한 금방 넘김
- 프롬프트 떠올리기 어려움, 오류 완벽히 잡아내지 못함
- 전체 분석 내용을 프롬프트에 입력하기 어려움

한계점

- 메시지 횟수 제한 금방 넘김
- 프롬프트 떠올리기 어려움, 오류 완벽히 잡아내지 못함
- 전체 분석 내용을 프롬프트에 입력하기 어려움
- 번거로움, 한 번에 최종 결과물까지 원하게 됨
- 점점 무비판적으로 ChatGPT에 의지하게 됨

가장 좋은 방법: 분석 과정을 ChatGPT가 주도하기

가장 좋은 방법: 분석 과정을 ChatGPT분석가**가 주도하기**

가장 좋은 방법: 분석 과정을 ChatGPT**분석가**가 주도하기

- 분석 목적, 절차 설계
- 파이썬 코드 생성, IDE에서 실행
- 분석 단계별 오류 파악

가장 좋은 방법: 분석 과정을 ChatGPT**분석가**가 주도하기

안전함

- 할루시네이션, 보안

가장 좋은 방법: 분석 과정을 ChatGPT**분석가**가 주도하기

안전함

- 할루시네이션, 보안

제약 없음

- 파일 용량, 분석 횟수, 분석 시간, 최신 패키지

데이터 분석에 활용하기

코드 만들기

- 코드 설명 요청하기, 주석 달기
- 함수 매뉴얼, 패키지 매뉴얼 설명 요청하기
- 패키지 바꾸기
- 코드 개선하기
- 언어 바꾸기

데이터 분석 기법 학습하기

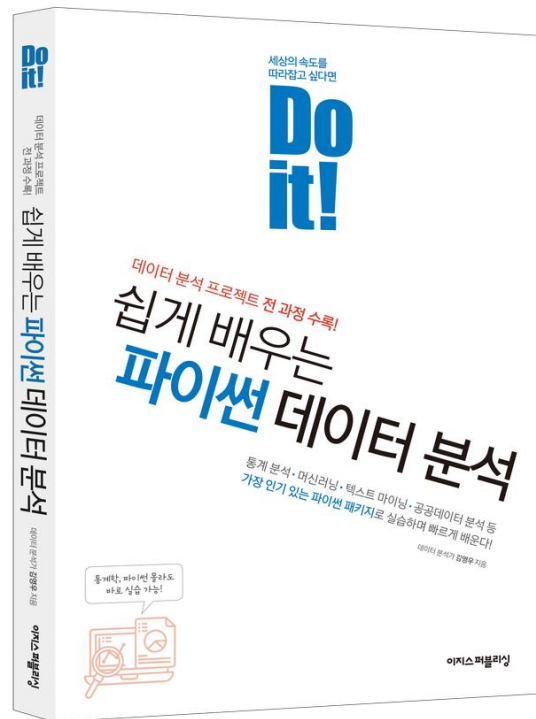
- 분석 기법 물어보기
- 튜터 역할 부여하기
- 데이터 분석 문제 출제하기
- 샘플 데이터 만들기
- 웹에 있는 데이터로 데이터 프레임 만들기

데이터 분석 보고서 작성하기

- 분석 아이디어 얻기
- 분석 결과 해석 문장 작성하기
- 분석 절차 검수하기
- 보고서 양식으로 정리하기

데이터 분석 기술 효율적으로 익히기

파이썬 활용법 익히기



네트워킹

데이터 분석 커뮤니티

<https://fb.com/groups/datacommunity>

지피터스

<https://gpters.org>

모두가 궁금해하는 데이터의 모든 것 데이터홀릭!

박박사 김팀장 조과장

데이터홀릭! dataholic4
구독자 1.98천명

구독중

홈 동영상 재생목록 커뮤니티 채널 정보

업로드한 동영상 ▾ 모두 재생 ▹ 정렬 기준

<p>데이터를 봐야 분석을 하지! - 데이터 공유 문화 1:30:08</p> <p>데이터를 봐야 분석을 하지! - 데이터 공유 문화 조회수 131회 · 3일 전</p>	<p>보안팀 이야기(2) 55:32</p> <p>Ep(76) 보안팀이야기(2) IT가 늘어나는 만큼 늘어나는 보... 조회수 51회 · 1주 전</p>	<p>보안팀 이야기(1) 59:17</p> <p>Ep(75) 보안팀이야기(1) 제발 내가 너를 도울 수 있게 도와... 조회수 112회 · 2주 전</p>	<p>소카의 본격채용특집! 데이터공채용 (일군을) 1:03:19</p> <p>Ep(73) 차가 필요한 모든 순간 소카의 본격 채용 특집! 데이... 조회수 315회 · 3주 전</p>
<p>소카의 본격채용특집! 데이터공채용 (2)직무면 42:16</p> <p>Ep(74) 차가 필요한 모든 순간 소카의 본격 채용 특집! 데이... 조회수 174회 · 3주 전</p>	<p>데이터가 돈이 되는 퀸트의 세계(2) 48:18</p> <p>Ep(72) 데이터가 곧 돈이 되는 퀸트의 세계(2) - 이현열님 조회수 186회 · 1개월 전</p>	<p>데이터가 돈이 되는 퀸트의 세계(1) 52:59</p> <p>Ep(71) 데이터가 곧 돈이 되는 퀸트의 세계(1) - 이현열님 조회수 410회 · 1개월 전</p>	<p>스트리머의 친구 트립, 트게더의 회사 EUN의 본격 채용특집 (2) 직무소개편 1:08:03</p> <p>Ep(70) 스트리머의 친구 트립, 트게더의 회사 EUN의 본격 ... 조회수 213회 · 1개월 전</p>
<p>스트리머의 친구 트립, 트게더의 회사 EUN의 본격 채용특집 (1) 회사소개편 53:04</p> <p>Ep(69) 스트리머의 친구 트립, 트게더의 회사 EUN의 본격 ... 조회수 340회 · 1개월 전</p>	<p>데이터로 발굴하는 인류의 역사, 아니, 만사 58:53</p> <p>Ep(68) 데이터로 발굴하는 인류의 역사, 아니, 만사. - 예맨... 조회수 144회 · 1개월 전</p>	<p>데이터가 소중하다면 고개를 들어 고고학을 보라 1:06:56</p> <p>Ep(67) 데이터가 소중하다면 고개를 들어 고고학을 보라... 조회수 157회 · 1개월 전</p>	<p>기업교육에서는 이런 데이터를 어떻게 분석합니까? 1:17:34</p> <p>Ep(66) 기업교육에서는 이런 데이터를 어떻게 분석합니까? 조회수 260회 · 2개월 전</p>

실전에 활용할 시간!