Elixirr

파이썬으로 배우는 따릉이 데이터 분석과 시각화

1회차

문제 정의

수업 일정

전체 수업은 13회로 구성된다.

1
- 따릉이 이용현황 파악
- 문제 정의
- 파이썬 및 사용할 라이브러리 소개

2
- 비주얼 스튜디오 코드 설치
- 따릉이 데이터 수집

3
- 파이썬 라이브러리
- 따릉이 데이터프레임 만들기

4
- 따릉이 데이터프레임 관찰하기

5
- 시간 개념에 따른 데이터 분석을 위한 컬럼 추가

6
- 장소적 특징에 따른 데이터 분석을 위한 컬럼 추가

7
- 시간 개념에 따른 데이터 분석 및 시각화-(1)

8
- 시간 개념에 따른 데이터 분석 및 시각화-(2)

9
- 장소 특징에 따른 데이터 분석 및 시각화-(1)

10
- 장소 특징에 따른 데이터 분석 및 시각화-(2)

수업 일정

전체 수업은 13회로 구성된다.

- 시간 개념 X 장소 특징에 따른 데이터 분석 및 시각화

- 주말과 평일에 이용건수가 많은 대여소 데이터 분석 및 시각화

- 문제 정의에 맞춘 해결방안 도출
- 총정리

데이터 분석의 5 단계

데이터 분석은 아래와 같이 5단계로 수행된다. 시각화 과정이 생략될 수 있다.

데이터 분석의 5단계

- 현재 상황에 따른 문제점 파악
- 미래 개선될 점에 대해서 분석주제 정의

- 문제 정의에 따라서 분석에 필요한 데이터 선택
- 데이터 수집

- 확보한 데이터 관찰 및 기초 데이터 분석
- 분석에 필요한 데이터 컬럼 추가 및 가공

- 문제 정의를 반영한 데이터 분석
- 연관될 수 있는 다른 데이터의 수집, 연결 및 분석

- 데이터 분석을 반영할 수 있는 시각화
- 문제 정의에 맞춘 해결방안 도출

1. 문제 정의

2. 데이터 수집

3. 데이터 가공

4. 데이터 분석

5. 시각화 및 탐색

데이터 분석의 5 단계

데이터 분석 단계에 맞추어 따릉이 데이터 분석을 수행한다.

데이터 분석의 5단계

- **현재 따릉이가 이용되고 있는 현황 조사**
- **미래에 개선될 점에 대해서 분석주제 정의**
- **파이썬과 판다스 소개**

- 환경 설정
  비주얼 스튜디오 코드(Visual Studio Code) 설치
- 따릉이 데이터 수집

- 확보한 따릉이 데이터 관찰 및 기초 데이터 분석
- 분석에 필요한 데이터 컬럼 추가 및 가공

- 시간 개념에 따른 데이터 분석
- 장속적 특징에 따른 데이터 분석
- 두 개념을 동시에 고려한 데이터 분석
- 인기 대여점 Top50

- 4단계 데이터 분석을 가장 잘 반영할 수 있는 시각화
- 문제 정의에 맞춘 해결방안 도출

따릉이 : 따릉이는 어떻게 이용할까?

따릉이 : 서울시 자전거 대여 서비스

서울시, 따릉이 5년...

56% '출퇴근시간대' 집중, 20분 이내 57%

대중교통 이용 전후 '퍼스트-라스트 마일' 문제 해결

강남-여의도가 도심보다 원거리 이동 많아...

도심지 자전거도로 등 인프라 확충

인기 대여소는 여의도 한강, 뚝섬

따릉이 : 서울시 자전거 대여 서비스

GD쌤

2015년 9월 2000대로 출발한 따릉이는 2020년 12월 기준 3만7500대로 20배가 늘었고, 대여소도 150개에서 2175개로 급증했어요.

2019년 따릉이 만족도 설문조사에서 36.3%가 출퇴근용, 26.8%가 여가 및 취미, 17%가 운동 목적으로 따릉이를 이용한다고 응답했어요.

여러분, 이제부터 우리는 2019년 6월 따릉이 데이터를 수집해서 분석할거예요.

주제1. 시간 개념에 따른 이용 패턴

미디어에 알려진 따릉이 이용 패턴에 대해서 궁금한 것이 생겼어요. 언제 많이 타는지(이용건수)와 한 번 대여 하면 어느 정도의 시간을 대여하는지(이용시간) 에 대한 분석이 필요해 보인다.

출퇴근 목적으로 사용하는 것이 가장 많고 이용시간도 20분 이내가 가장 많네.

해리

그럼 출퇴근용이 아닌 경우에는 더 오래 탈까?

제니

주말과 평일에 대여 건수의 차이가 있을까?

론

주제2. 장소적 특징에 따른 이용패턴

대여소가 위치한 장소적 특징에 따라서 사용자의 이용패턴이 달라진다.

우리 집 앞은 언덕이 심해서 자전거를 타기가 힘들어.

해리

지역구區 별로 자전거를 이용하기 쉬운 장소가 인기가 많지 않을까?

제니

자전거 전용도로가 있는 곳이 사람들이 많이 타지 않을까?

론

주제3. 시간 개념 X 장소적 특징에 따른 이용패턴

시간 개념과 대여소가 위치한 장소적 특징을 모두 고려해서 사용자의 이용패턴을 분석해 본다.

서울시의 모든 지역구별로 따릉이 이용시간의 평균을 계산해서 가장 많이 사용하는 지역구를 알아볼까?

해리

6월 한달 일자에 따라 또는 하루 24시간 시간대에 따라 지역구별로 따릉이 이용건수를 세어볼까?

제니

주말과 평일에 각각 인기있는 대여점 50개를 한번 알아보자.

론

수업 환경 – 파이썬 (Python)

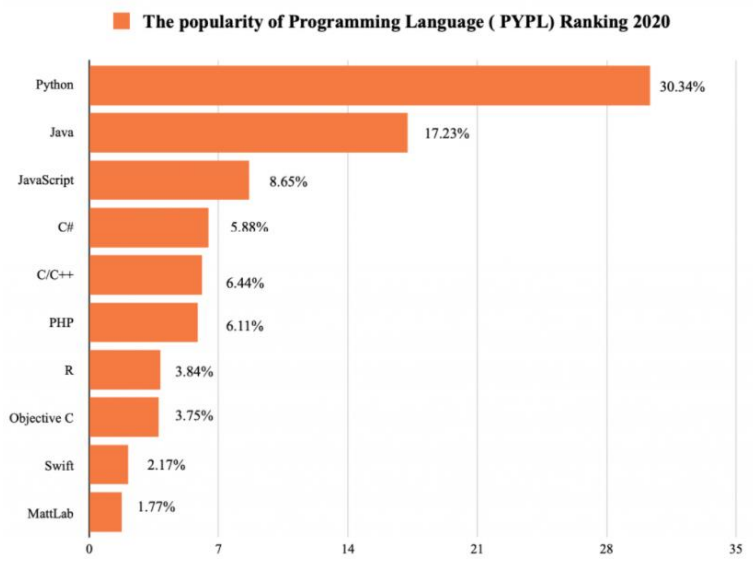
이번 수업은 파이썬을 기본 언어로 사용한다.

파이썬 언어

자료 : Guido Van Rossum, at 2006 O'Reilly Open Source Convention (OSCON), Doc Searls, Wikepedia

- 창시자 : 네덜란드 개발자 귀도 반 로섬
- 인터프리터 프로그래밍 언어
- 문법이 간결하고 표현 구조가 인간의 사고 체계와 유사
- 초보자부터 전문가까지 다양한 사용층
- 풍부한 라이브러리 → 대학을 비롯한 연구기관 및 산업계에서 이용이 활발
- 파이썬은 웹 개발 뿐 아니라 데이터 분석, 머신 러닝, 딥러닝 등 다양한 분야에서 활용

프로그래밍 언어 인기 순위 (PYPL)



자료 : https://darly.solutions/the-most-popular-programming-languages-in-2021

- 2020년 기준 사람들이 선호하는 인기 프로그래밍 언어 1위
- The popularity of Programming Language (PYPL) : 구글에서 가장 많이 검색된 프로그래밍 언어 교재 순

수업 환경 – 판다스 (Pandas)

파이썬은 데이터 과학 관련하여 다양한 라이브러리를 보유하고 있다. 이번 수업에서는 데이터 분석 기본 라이브러리인 판다스를 활용한다.

파이썬의 데이터 과학 관련 라이브러리

TensorFlow

Keras

NumPy

pandas

seaborn

PyTorch

matplotlib

scikits learn

- 파이썬은 데이터 분석, 머신 러닝, 딥러닝 등 데이터 과학 분야에 다양하고 풍부한 라이브러리를 보유
- 판다스는 데이터 분석에 특히 많이 사용되는 라이브러리임

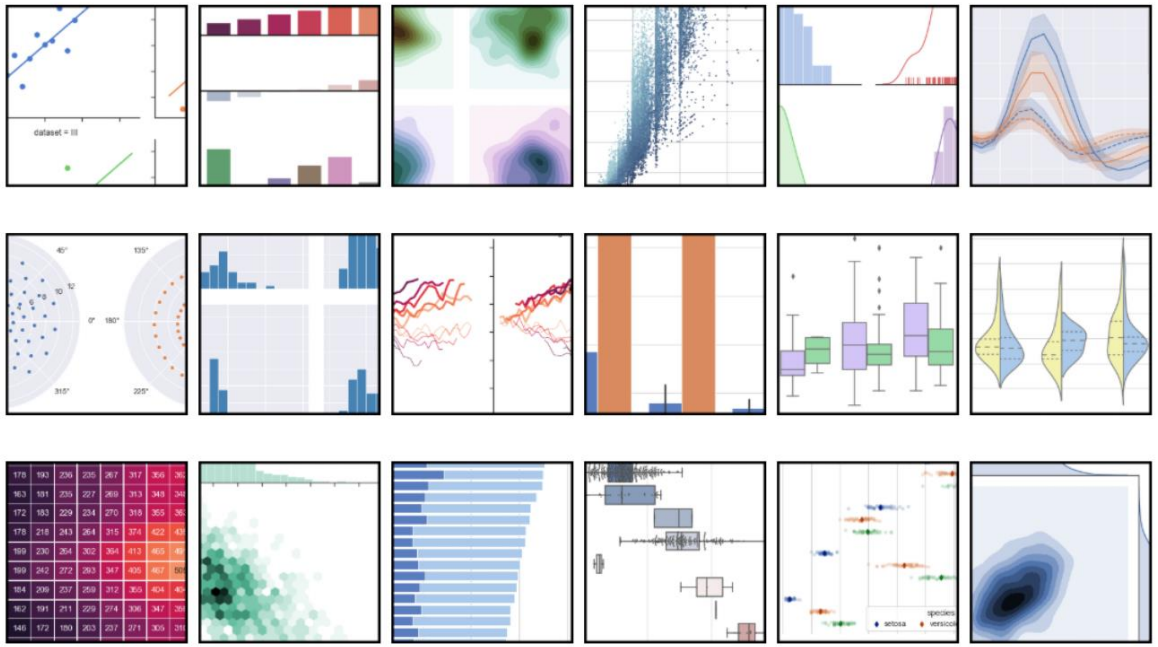
수업 환경 – 맷플랏립 (Matplotlib) & 시본 (Seaborn)

파이썬은 다양한 그래프를 그려주는 라이브러리를 보유하고 있다. 이번 수업에서는 맷플랏립과 시본 라이브러리를 활용한다.

matplotlib

Matplotlib : 파이썬으로 기본적인 차트들을 쉽게 그릴 수 있도록 도와주는 시각화 라이브러리

seaborn



- Seaborn : matplotlib 기반으로 만들어진 통계 데이터 시각화 라이브러리로 high-level interface를 제공

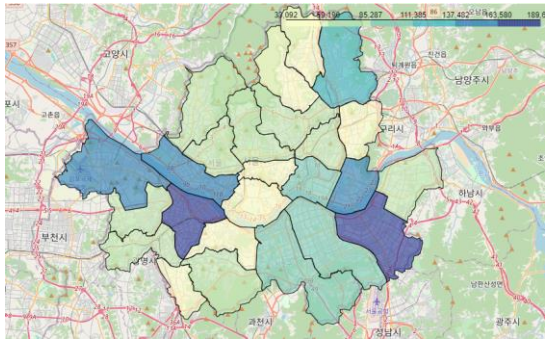
수업 환경 – 폴리움 (Folium)

파이썬은 지도 시각화를 해주는 다양한 함수를 가지고 있는 폴리움 라이브러리를 보유하고 있다.

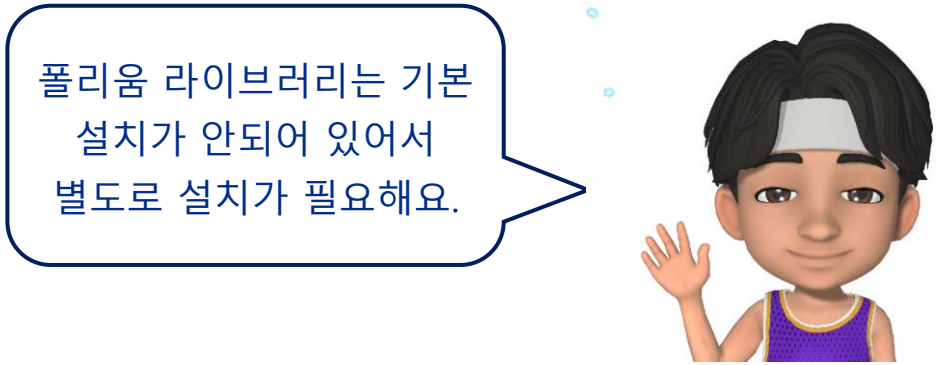
이번 수업에서는 폴리움 라이브러리를 활용한다.

Folium

- 파이썬에서 제공해주는 라이브러리로서 지도를 다루는 대표적인 라이브러리
- 폴리움을 사용하기 위해서는 위도와 경도 데이터를 알아야 한다.



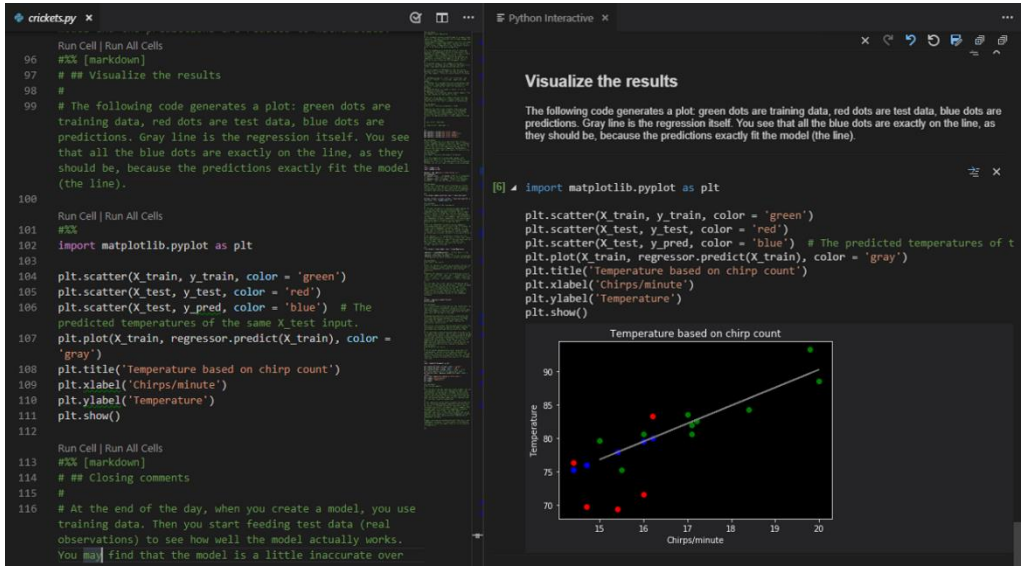
- 대여소 위도, 경도 데이터를 활용하여 지도에 분석한 내용을 시각화



수업 환경 – 비주얼 스튜디오 코드 (Visual Studio Code)

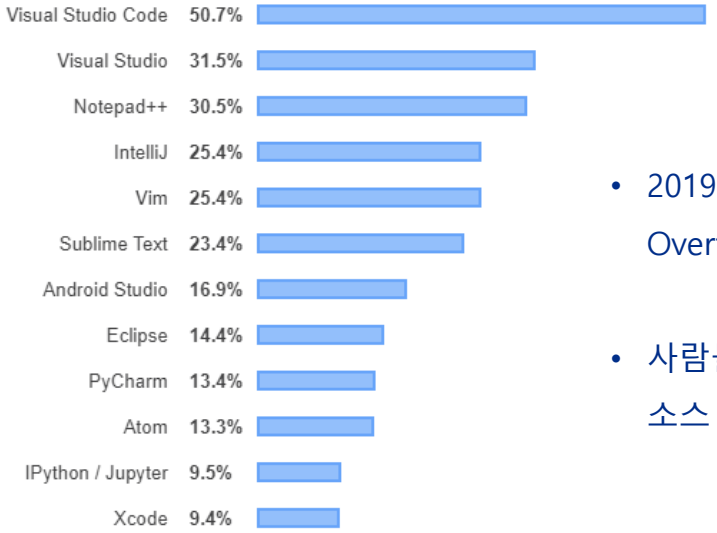
이번 수업에서는 파이썬 코드 작성 시 비주얼 스튜디오 코드 편집기를 사용한다.

비주얼 스튜디오 코드



- 마이크로소프트에서 윈도우, macOS, 리눅스 용으로 개발한 '소스 코드 편집기'
- 다양한 언어 지원, 편리한 기능, 확장 팩 등 제공

코드 편집기 인기 순위



- 2019년 Stack Overflow Survey 조사
- 사람들이 선호하는 소스 코드 편집기 1위

자료 : Most Popular Development Environments, All Respondents, https://insights.stackoverflow.com/survey/2019

따릉이 수업 과정

데이터 수집

	A	B	C	D	E	F	G	H	I	J	K
1	자전거번호	대여일시	대여 대여소번호	대여 대여소명	대여거치대	반납일시	반납대여소번호	반납대여소명	반납거치대	이용시간	이용거리
2	SPB-22040	2019-06-03 8:49	646	장한평역 1번출구 (국민은행앞)	1	2019-06-03 9:17	3	중랑센터	14	27	1330
3	SPB-07446	2019-06-03 8:33	526	용답토속공원 앞	8	2019-06-03 9:27	3	중랑센터	14	54	1180
4	SPB-20387	2019-06-05 8:27	646	장한평역 1번출구 (국민은행앞)	1	2019-06-05 8:41	3	중랑센터	2	12	1930
5	SPB-16794	2019-06-05 8:46	646	장한평역 1번출구 (국민은행앞)	6	2019-06-05 8:53	3	중랑센터	14	6	1340
6	SPB-18266	2019-06-10 8:27	529	장한평역 8번 출구 앞	10	2019-06-10 8:33	3	중랑센터	2	5	1230
7	SPB-13926	2019-06-11 8:29	646	장한평역 1번출구 (국민은행앞)	4	2019-06-11 8:37	3	중랑센터	2	7	1360
8	SPB-14638	2019-06-12 8:29	646	장한평역 1번출구 (국민은행앞)	4	2019-06-12 8:35	3	중랑센터	2	5	1340
9	SPB-18588	2019-06-17 8:34	646	장한평역 1번출구 (국민은행앞)	5	2019-06-17 8:44	3	중랑센터	2	8	1360
10	SPB-21148	2019-06-17 8:47	646	장한평역 1번출구 (국민은행앞)	6	2019-06-17 9:10	3	중랑센터	14	22	1330

- 분석주제에 맞는 파일 다운로드
- 파일 읽어 들이기

데이터 가공

```
# 일자 비교를 위해서 요일 컬럼 추가
dayofweek = ['월','화','수','목','금','토','일']
bike_ride['요일'] = bike_ride['대여일시'].dt.dayofweek.apply(lambda x: dayofweek[x])


# 주중/주말공휴일 구분
bike_ride['주말구분'] = bike_ride['요일'].apply(\
                        lambda x: '평일' if x not in (['토', '일']) else '주말')

# 일자 컬럼 추가
bike_ride['일자'] = bike_ride['대여일시'].dt.day

# 시간대 컬럼 추가
bike_ride['대여시간대'] =bike_ride['대여일시'].dt.hour
bike_ride['반납시간대'] =  bike_ride['반납일시'].dt.hour

bike_ride.head(2)
```

	자전거번호	대여일시	대여 대여소번호	대여 대여소명	대여거치대	반납일시	반납대여소번호	반납대여소명	반납거치대	이용시간	이용거리	요일	주말구분	일자	대여시간대	반납시간대
0	SPB-22040	2019-06-03 08:49:27	646	장한평역 1번출구 (국민은행앞)	1	2019-06-03 09:17:10	3	중랑센터	14	27	1330.00	월	평일	3	8	9
1	SPB-07446	2019-06-03 08:33:22	526	용답토속공원 앞	8	2019-06-03 09:27:16	3	중랑센터	14	54	1180.00	월	평일	3	8	9

- 판다스 데이터프레임 생성
- 필요한 컬럼 생성 및 추가

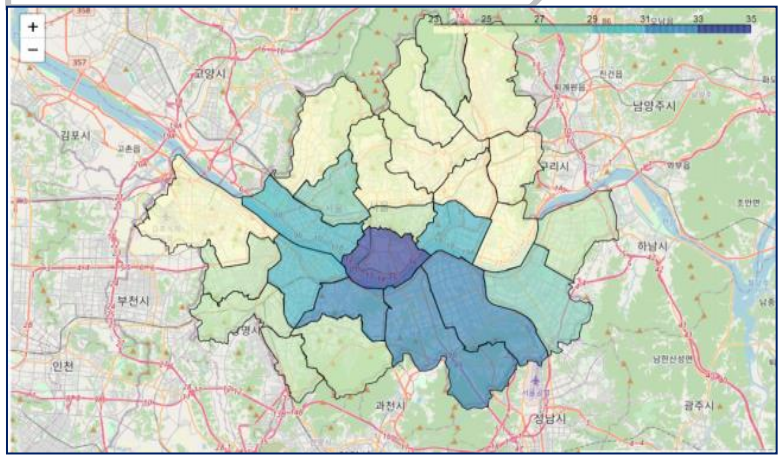
데이터 모델링

```
day_gu_use = bike_ride.pivot_table( \
                    index = '일자', \
                    columns = '대여구', \
                    values = '자전거번호', \
                    aggfunc = 'count')
day_gu_use
```

대여구	강남구	강동구	강북구	강서구	관악구	광진구	구로구	금천구	노원구	도봉구	...	성동구	성북구	송파구	양천구	영등포구	용산구	은평구	종로구	중구	중랑구
일자																					
1	3217	2766	1458	4456	3023	5198	2571	933	4007	1366	...	3493	2510	7189	2200	6285	1869	2280	2683	1559	1976
2	3047	2579	1515	4235	3182	5088	2484	907	3930	1347	...	3620	2441	6522	2022	6242	1876	2181	2461	1384	1801
3	3864	2729	1397	4973	3022	5289	2851	1327	4084	1411	...	3667	2609	6227	2164	6791	1657	2165	3083	1834	2044
4	3589	2818	1413	5100	3012	5209	2949	1263	4439	1439	...	3993	2720	6579	2346	6456	1538	2175	2892	1833	2080
5	3296	2733	1440	5266	3028	4868	2795	1271	4327	1383	...	3661	2566	6496	2250	6272	1592	2168	3113	1935	2108
6	1446	1366	739	2118	1440	2296	1265	451	2097	727	...	1652	1353	3042	999	2551	734	1145	1209	668	1092
7	2451	1874	896	3311	1957	3195	1843	669	2737	879	...	2473	1727	4671	1528	4370	1143	1531	1993	1090	1216
8	3323	2970	1539	4685	3323	5427	2729	940	4143	1493	...	3615	2616	7329	2347	6816	1917	2360	2570	1470	2076
9	2369	2191	1215	3343	2547	3768	1889	653	3129	1118	...	2715	2082	5230	1686	4526	1313	1664	2054	1121	1511
10	2834	1942	1018	3746	2307	3937	2144	943	3269	1100	...	2813	2067	5078	1558	4647	1262	1694	2325	1408	1499

- 다양한 집계와 분석 명령어를 사용해서 모델링

데이터 시각화 및 탐색



- 지도로 시각화
- 꺾은선 그래프, 막대 그래프로, 박스 그래프로 표현

나 지금 어느 단계를 공부하는 거지?

1.문제정의

2.데이터수집

3.데이터 가공

4.데이터 분석

5.시각화 및 탐색

시간적 개념 : 일자별, 요일별, 시간대별로 따릉이 이용건수 및 이용시간 분석

장소적 특징 : 지역구별로 따릉이 이용건수 및 이용시간 분석

시간적 개념 x 장소적 특징 : 두 개념을 동시에 고려해 분석

퀴즈

퀴즈를
풀어봅시다

1. 소스 코드 에디터로서 코드 자동 완성 등 편리한 기능을 포함하고 파이썬, 주피터 노트북 등 다양한 확장 팩 (Extension)을 가진 것은 ?

2. 파이썬으로 기본적인 차트들을 쉽게 그릴 수 있도록 도와주는 시각화 라이브러리는 ?

3. matplotlib 기반으로 만들어진 통계 데이터 시각화 라이브러리로 high-level interface를 제공하는 것은?

4. 파이썬에서 제공해주는 라이브러리로서 지도를 다루는 대표적인 라이브러리는 ?

수업 마무리

GD쌤

지금까지 1회차 수업내용을 배워 보았습니다.

다음 시간에는 2-1회차 수업내용으로 비주얼 스튜디오 코드, 아나콘다와 폴리움 라이브러리를 설치하는 환경설정을 하겠습니다.

수고 많으셨어요. 다음 시간에 만나요.