

파이썬으로 배우는 **따릉이** 데이터 분석과 시각화

6회차 데이터 가공

이 자료는 Elixir의 사전 서면 승인 없이 외부에 배포하기 위해
그 일부를 배포, 인용 또는 복제 할 수 없습니다.

© Copyright Elixir



수업 일정

전체 수업은 13회로 구성된다.



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



- 비주얼 스튜디오 코드 설치
- 따릉이 데이터 수집



- 파이썬 라이브러리
- 따릉이 데이터프레임 만들기



- 따릉이 데이터프레임 관찰하기



- 시간 개념에 따른 데이터 분석을 위한 컬럼 추가



- 장소적 특징에 따른 데이터 분석을 위한 컬럼 추가



- 시간 개념에 따른 데이터 분석 및 시각화-(1)



- 시간 개념에 따른 데이터 분석 및 시각화-(2)



- 장소 특징에 따른 데이터 분석 및 시각화-(1)



- 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

1. 문제정의

2. 데이터 수집

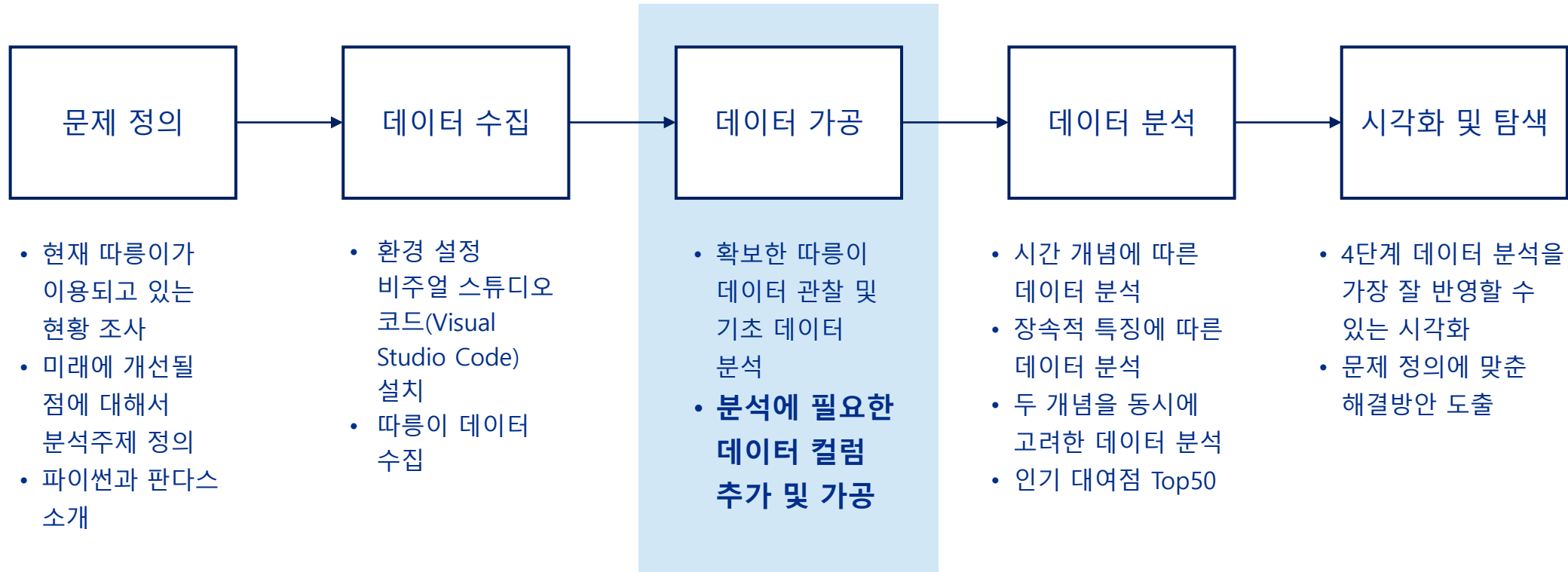
3. 데이터 가공

4. 데이터 분석

5. 시각화 및 탐색

데이터 분석 단계에 맞추어 따릉이 데이터 분석을 수행한다.

데이터 분석의 5단계





여기서 배울 내용은 ?

3.데이터 가공

1.문제정의

2.데이터수집

3.데이터 가공

4.데이터 모델링

5.시각화 및 탐색

단계 1 : 분석할 데이터프레임 만들기

단계 2 : 데이터프레임 관찰하기

단계 3 : 분석주제에 맞는 새로운 컬럼 추가하기

따릉이 데이터 : '서울 열린 데이터 광장'

3.데이터 가공

따릉이 데이터는 '서울 열린 데이터 광장' 사이트를 통해 공유 되어 있다.
사이트에 접속하여 '따릉이' 로 검색한다.

data.seoul.go.kr

1 data.seoul.go.kr

로그 주피터 Home ppticon redmine preswot Google Google 계정 따릉이 머신러닝 OverTheMoon

서울특별시 시인감사유부조위원회 출범 5주년 기념행사 온라인 참석 시민 모집 서울소식 응답소 정보공개 분야별정보

서울 열린데이터 광장 공공데이터 통계 소식&참여 이용안내 AI학습데이터 로그인 회원가입 사이트맵

모든 서울시민을 위한 공공데이터

열린데이터광장에서 서울시와 연계 기관이 공개한 공공데이터를 확인하실 수 있습니다.
서울시와 관련된 다양한 공공데이터를 확인해 보세요.

2

Q 찾고 싶은 데이터를 입력해 주세요. 검색하기

데이터셋 6,983 서비스 14,395 Open API 5,437

보건	일반행정	문화/관광	산업/경제	복지	환경
교통	도시관리	교육	안전	인구/가구	주택/건설

관심데이터 36종 생활이슈 자세히보기 >

뉴스레터 구독안내 Newsletter 열광뉴스레터를 신청하세요!

1 서울 열린데이터 광장
data.seoul.go.kr 접속

2 검색창에 '따릉이'를
입력하고 '검색하기' 클릭

서울특별시 공공자전거 대여소 정보'로 들어간다.

서울특별시 빠져 나갈 없는 사회... '서울시 청렴사회' 민관협의회'가 주도

서울소식 응답소 정보공개 분야별정보

서울 열린데이터 광장 공공데이터 통계 소식&참여 이용안내 로그인 회원가입 사이트맵

데이터셋 Home > 공공데이터 > 데이터셋

🔍 찾고 싶은 데이터를 입력해 주세요. 검색하기

교통

공공데이터

활용갤러리 등록 URL 복사 목록 이동

1

서울특별시 공공자전거 대여소 정보

서울특별시 공공자전거 대여소(따릉이) 현황정보입니다.
대여소의 이름, 관리번호, 위치정보와, 기차대수 정보를 제공합니다.

파일내려받기 * 파일에 이상이 있는 경우 '오류신고'를 통해 운영자에게 알려주세요. 오류신고

NO	항목	파일명	용량 (MB)	수정일	내려받기
1	데이터	공공자전거 대여소 정보(20.07.13 기준).xlsx	0.3	2020.10.22	
2	데이터	서울특별시 공공자전거 대여소 정보(19.12.9).xlsx	0.3	2019.12.10	
3	데이터	시 공공자전거 대여소 설치 일시 정보(20190517).xlsx	0.2	2019.08.28	
4	데이터	서울특별시 공공자전거 대여소 정보(배치정보)(2019.06.24).xlsx	0.1	2019.07.29	
5	데이터	공공자전거 대여소 정보 201905.xlsx	0.1	2019.06.20	

[전체 파일보기](#)

- **1** 을 클릭해서 해당 파일을 다운로드 받는다.
- 대여소 정보 파일은 2019년 5월기준이다.


수집된 데이터를 엑셀로 열어서 데이터를 확인한다.

bikes.head()

Python

	자전거번호	대여일시	대여 대여소번호	대여 대여소명	이용시간	이용거리	일자	대여시간대	요일	주말구분
0	SPB-22040	2019-06-03 08:49:00	646	장한평역 1번출구 (국민은행앞)	27	1330	3	8	월	평일
1	SPB-07446	2019-06-03 08:33:00	526	용답토속공원 앞	54	1180	3	8	월	평일
2	SPB-20387	2019-06-05 08:27:00	646	장한평역 1번출구 (국민은행앞)	12	1930	5	8	수	평일
3	SPB-16794	2019-06-05 08:46:00	646	장한평역 1번출구 (국민은행앞)	6	1340	5			
4	SPB-18266	2019-06-10 08:27:00	529	장한평역 8번 출구 앞	5	1230	10			

구분	위도	경도
마포구	37.549561	126.90575
마포구	37.556	126.91045
마포구	37.554951	126.91084
마포구	37.550629	126.91499
마포구	37.550007	126.91483
마포구	37.548645	126.91283
마포구	37.55751	126.9185


[교통]

서울특별시 공공자전거 대여소 정보

서울특별시 공공자전거 대여소(따릉이) 현황정보입니다. 대여소의 이름, 관리번호, 위치정보와, 거치대...
수정일자: 2020-10-22 제공기관: 서울특별시 제공부서: 도시교통실 보행친화기획관 자전거...

FILE

공공 데이터

	A	B	C	D	E	F
1	구분	대여소번호	대여소명	위도	경도	거치대수
2	마포구	101	101. (구)합정동 주민센터	37.549561	126.90575	5
3	마포구	102	102. 망원역 1번출구 앞	37.556	126.91045	20
4	마포구	103	103. 망원역 2번출구 앞	37.554951	126.91084	14
5	마포구	104	104. 합정역 1번출구 앞	37.550629	126.91499	13
6	마포구	105	105. 합정역 5번출구 앞	37.550007	126.91483	5
7	마포구	106	106. 합정역 7번출구 앞	37.548645	126.91283	10
8	마포구	107	107. 신한은행 서교동금융센터점 앞	37.55751	126.9185	5

데이터프레임으로 읽어 들이기 : pd.read_excel()

3.데이터 가공

판다스 라이브러리에 있는 pd.read_excel() 명령어를 사용해서 대여소 위치정보가 있는 엑셀 파일을 읽어들이어 데이터프레임 bike_shop을 생성한다.

```
bike_shop = pd.read_excel( 파일명 )
```

```
bike_shop = pd.read_excel('data/공공자전거 대여소 정보_201905.xlsx')
bike_shop.head()
```

	구분	대여소번호	대여소명	위도	경도	거치대수
0	마포구	101	101. (구)합정동 주민센터	37.55	126.91	5
1	마포구	102	102. 망원역 1번출구 앞	37.56	126.91	20
2	마포구	103	103. 망원역 2번출구 앞	37.55	126.91	14
3	마포구	104	104. 합정역 1번출구 앞	37.55	126.91	13
4	마포구	105	105. 합정역 5번출구 앞	37.55	126.91	5

```
# 필요한 컬럼들을 추출해서 bike_gu 라는 변수에 할당한다.
```

```
bike_gu = bike_shop[['구분', '대여소번호', '대여소명', '위도', '경도']]
bike_gu.head(1)
```

	구분	대여소번호	대여소명	위도	경도
0	마포구	101	101. (구)합정동 주민센터	37.55	126.91

- 작은 따옴표 ' ' 안에 디렉토리를 포함한 파일명
- 변수 = 숫자, 문자와 같은 값들을 저장하는 공간

- pd. : 판다스 (as pd) 라이브러리에 속한
- read_excel() : 명령어로 엑셀 파일을 읽고
- bike_shop 라는 공간 (= 변수) 에 저장함

데이터프레임 연결하기 : pd.merge()

3.데이터 가공

pd.merge(df1, df2, left_on='df1_컬럼명', right_on='df2_컬럼명')

①

②

③

④

- ① **pd.merge** : 두 개의 데이터프레임에서 공통된 열을 기준으로 동일한 값을 가지는 행을 각 데이터프레임에서 찾은 후, 이를 병합시킨다.
- ② **df1, df2** : 연결할 데이터프레임 두개
- ③ **left_on='df1_컬럼명'** : df1 데이터프레임에 있는 컬럼명
- ④ **right_on='df2_컬럼명'** : df2 데이터프레임에 있는 컬럼명

같은 내용의 컬럼이
두 개씩 존재한다.

```
bikes = pd.merge(bikes, bike_gu, left_on='대여 대여소번호', right_on='대여소번호')
bikes.head(3)
```

bikes : 대여 대여소번호,
대여 대여소명
bike_gu : 대여소번호,
대여소명

	자전 거번호	대여일시	대여 대 여소번호	대여 대여소명	이용 시간	이용 거리	일 자	대여 시간 대	요 일	주말 구분	구분	대여 소번호	대여소명	위도	경도
0	SPB- 22040	2019-06-03 08:49:00	646	장한평역 1번출구 (국민은행앞)	27	1330	3	8	월	평일	동대 문구	646	646. 장한평역 1번출 구 (국민은행앞)	37.56	127.06
1	SPB- 20387	2019-06-05 08:27:00	646	장한평역 1번출구 (국민은행앞)	12	1930	5	8	수	평일	동대 문구	646	646. 장한평역 1번출 구 (국민은행앞)	37.56	127.06
2	SPB- 16794	2019-06-05 08:46:00	646	장한평역 1번출구 (국민은행앞)	6	1340	5	8	수	평일	동대 문구	646	646. 장한평역 1번출 구 (국민은행앞)	37.56	127.06

데이터프레임내의 컬럼 삭제하기 : df.drop()

3.데이터 가공

`bikes.drop([삭제할 컬럼명], axis='columns', inplace=True)`

①

②

③

④

- ① `bikes.drop` : 데이터프레임에서 여러 개의 인덱스나 컬럼을 삭제하는 명령어
- ② `[삭제할 컬럼명]` : 삭제할 컬럼명을 리스트 형식으로 입력한다. -> `[]` 사용
- ③ `axis='columns'` : 인덱스를 삭제할 때는 `'index'`, 컬럼을 삭제할 때는 `'columns'`를 써준다.
- ④ `inplace=True` : 변경된 내용을 `bikes` 데이터프레임에 고정시킨다.

```
bikes = pd.merge(bikes, bike_gu, left_on='대여 대  
bikes.head(3)
```

✓ 1.7s Python

	자전 거번 호	대여일시	대여 대 여소번호	대여 대여소명	이용 시간	이용 거리	일 자	대여 시간 대	요 일	주말 구분	구분	대여 소번 호	대여소명	위도	경도
0	SPB- 22040	2019-06-03 08:49:00	646	장한평역 1번출구 (국민은행앞)	27	1330	3	8 월	평일	동대 문구	646	646. 장한평역 1번출 구 (국민은행앞)			

```
bikes.drop(['대여 소번호', '대여소명'], axis='columns', inplace=True)  
bikes.head(1)
```

Python

	자전거번 호	대여일시	대여 대여소 번호	대여 대여소명	이용시 간	이용거 리	일 자	대여시간 대	요 일	주말구 분	구분	위도	경도
0	SPB- 22040	2019-06-03 08:49:00	646	장한평역 1번출구 (국민 은행앞)	27	1330	3	8 월	평일	동대문 구		37.56	127.06

```
bikes.rename( columns={ 변경전 컬럼명 : 변경후 컬럼명 }, inplace=True )
```

1

2

3

- 1 **bikes.rename** : 데이터프레임의 컬럼이름을 바꾸는 명령어
- 2 **columns={ 변경전 컬럼명 : 변경후 컬럼명 }** : 컬럼명을 바꾸고 싶으면 columns를 써준다.
변경전 컬럼명과 변경후 컬럼명을 딕셔너리 형식으로 입력한다.
- 3 **inplace=True** : 변경된 내용을 bikes 데이터프레임에 고정시킨다.

```
bikes.rename(columns={'구분' : '대여구', \
                      '위도' : '대여점위도', \
                      '경도' : '대여점경도'}, \
              inplace=True)
bikes.head()
```



	자전거 번호	대여일시	대여 번호	대여소명	이용시 간	이용거 리	일 자	대여시 간대	요 일	주말구 분	대여구	대여점 위도	대여점 경도
0	SPB- 22040	2019-06-03 08:49:00	646	장한평역 1번출구 (국민 은행앞)	27	1330	3	8	월	평일	동대문 구	37.56	127.06
1	SPB- 20387	2019-06-05 08:27:00	646	장한평역 1번출구 (국민 은행앞)	12	1930	5	8	수	평일	동대문 구	37.56	127.06
2	SPB- 16794	2019-06-05 08:46:00	646	장한평역 1번출구 (국민 은행앞)	6	1340	5	8	수	평일	동대문 구	37.56	127.06



나 지금 어느 단계를 공부하는 거지?

3.데이터 가공

1.문제정의

2.데이터수집

3.데이터 가공

4.데이터 모델링

5.시각화 및 탐색

단계 3 : 분석주제에 맞는 새로운 컬럼 추가하기

데이터프레임에 장소 관련 컬럼 추가

엑셀 파일을 읽어들인다 -> `pd.read_excel()`

두개의 데이터프레임 연결 -> `pd.merge()`

여러 개의 컬럼 삭제 -> `bikes.drop()`

여러 개의 컬럼명 변경 -> `bikes.rename()`



퀴즈를
풀어봅시다

1. 대여소정보 파일은 엑셀 파일이다. 이 파일을 읽어들이 데이터프레임을 만드는 명령어는 ?

2. 두 개의 데이터프레임에서 공통된 열을 기준으로 동일한 값을 가지는 행을 각 데이터프레임에서 찾은 후, 이를 병합시키는 명령어는 ?

3. 데이터프레임에서 여러 개의 인덱스나 컬럼을 삭제하는 명령어는 ?

4. 데이터프레임의 컬럼이름을 바꾸는 명령어는 ?

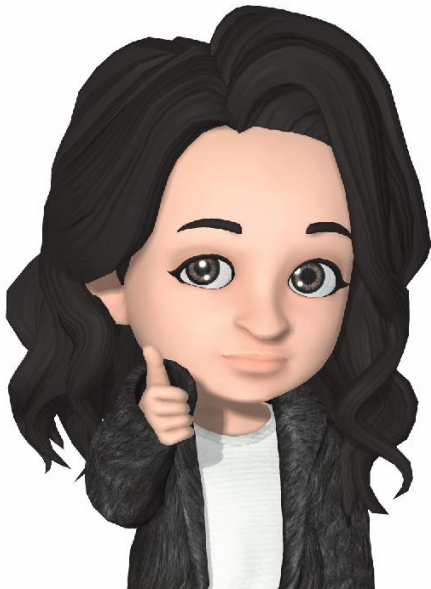


GD쌤

이제부터 Visual Studio Code 실습 환경에서 지금까지 배운 내용을 실습해 보겠습니다.

앞에서 배웠던 내용을 Visual Studio Code에서 직접 실습해보면 더욱 이해하기 편리할 것입니다.

수업 마무리



GD쌤

지금까지 6회차 수업내용을 배워 보았습니다.

다음 시간에는 7회차 수업내용으로 시간 개념에 따른 데이터 분석 및 시각화를
동시에 진행해 보겠습니다.

수고 많으셨어요. 다음 시간에 만나요.