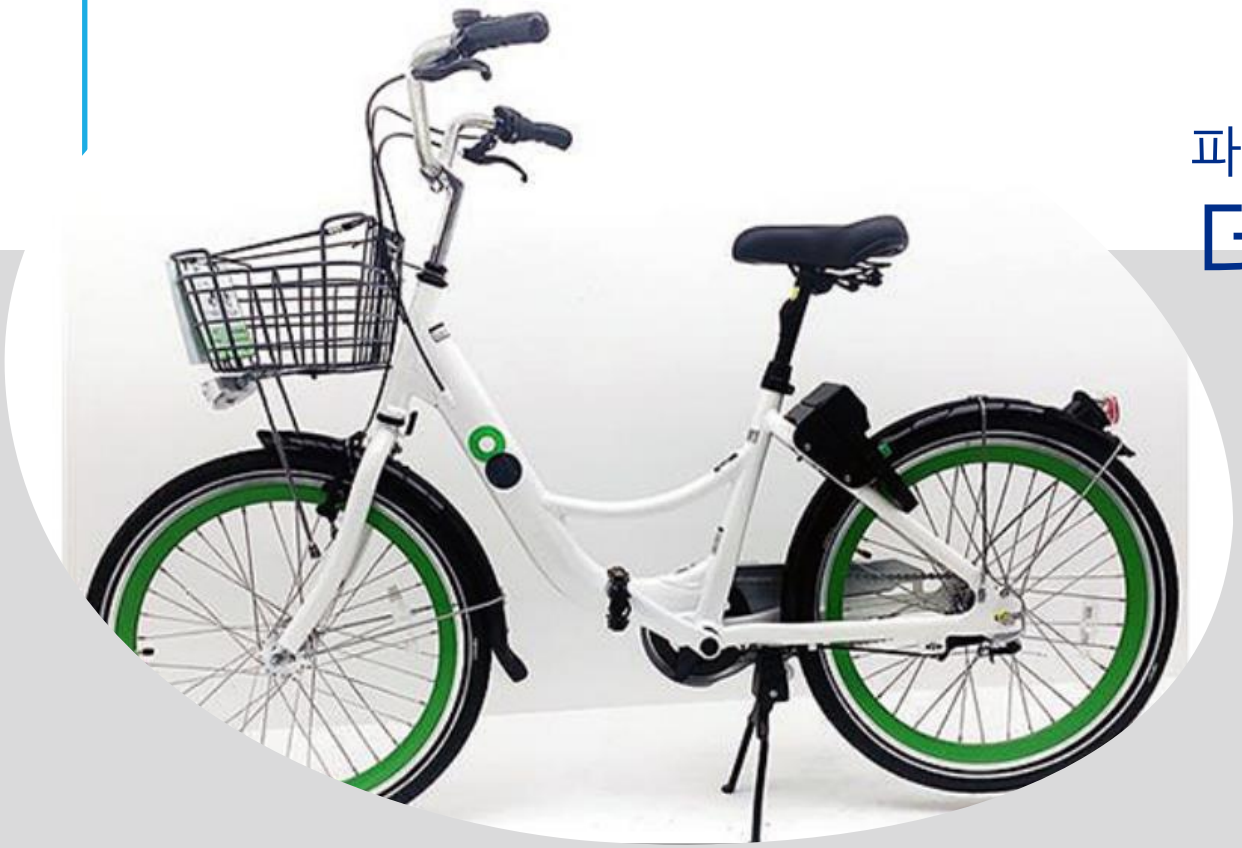
Elixirr



파이썬으로 배우는 따릉이 데이터 분석과 시각화

4회차 수업
데이터 가공

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

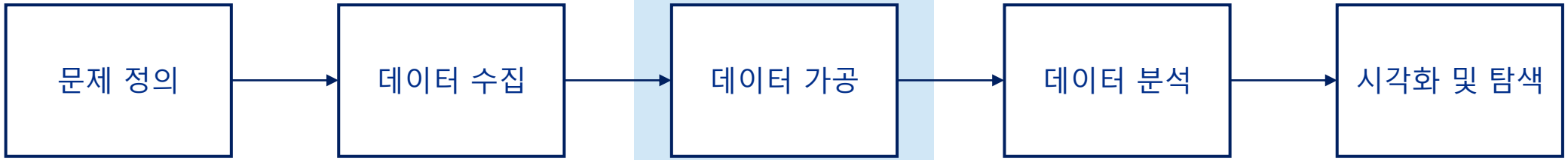
- 문제 정의에 맞춘 해결방안 도출
- 총정리

1. 문제정의
2. 데이터 수집
3. **데이터 가공**
4. 데이터 분석
5. 시각화 및 탐색

데이터 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

여기서 배울 내용은 ?

1.문제정의

2.데이터수집

3.데이터 가공

4.데이터 분석

5.시각화 및 탐색

단계 1 : 분석할 데이터프레임 만들기

단계 2 : 데이터프레임 관찰하기

단계 3 : 분석주제에 맞는 새로운 컬럼 추가하기

데이터프레임 관찰하기 : df.info()

bikes.info()

```
bikes.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2215632 entries, 0 to 2215631
Data columns (total 6 columns):
 #   Column    Dtype
---  ------    -----
 0   자전거번호     object
 1   대여일시      datetime64[ns]
 2   대여 대여소번호  int64
 3   대여 대여소명   object
 4   이용시간      int64
 5   이용거리      int64
dtypes: datetime64[ns](1), int64(3), object(2)
memory usage: 101.4+ MB
```

1. 데이터프레임 **전체 행수**를 알려준다 : **bikes.shape[0]**
2. 데이터프레임 **index**를 알려준다 : **bikes.index**
3. 데이터프레임 **컬럼 개수**를 알려준다 : **bikes.shape[1]**
4. 데이터프레임 **컬럼**을 알려준다 : **bikes.columns**
5. 데이터프레임의 **각 컬럼의 데이터타입**을 알려준다 : **bikes.dtypes**
6. 데이터프레임의 **메모리 사용량을 알려준다.**
7. 데이터프레임의 **전체 행과 컬럼의 수를 알려준다 : bikes.shape : (2215632, 6)**
   bikes.shape[0] : 2215632
   bikes.shape[1] : 6

데이터의 자료형

개별 컬럼은 정수, 실수, 문자, 날짜, 부울값 등의 자료형 (데이터 타입datatype)을 가집니다.

object

- 문자 또는 문자열 (작은 따옴표로 구분) 예 : '월', '화', '중랑센터', '7', '0.5'

int64

- 정수 예 : 0, 1, 5, -14, 21504

float64

- 실수 예 : 0.23, 0.00024, -0.125

datetime64

- 날짜 (작은 따옴표로 구분) 예 : '2019-06-03 08:33:22', '2019-06-26'

boolean

- 참(True) / 거짓(False)을 나타냄 예 : True(1), False(0)

데이터프레임 요약 통계 : df.describe()

bikes.describe()

```
bikes.describe()
```

[7]

	대여 대여소번호	이용시간	이용거리
count	2215632.00	2215632.00	2215632.00
mean	1226.15	27.17	3912.92
std	849.92	29.08	5145.47
min	3.00	1.00	0.00
25%	505.00	8.00	1220.00
50%	1159.00	17.00	2270.00
75%	1846.00	37.00	4670.00
max	9998.00	3133.00	232310.00

1. 데이터타입이 숫자인 컬럼의 **전체 갯수**
2. 데이터타입이 숫자인 컬럼의 **평균**
3. 데이터타입이 숫자인 컬럼의 **표준편차**
4. 데이터타입이 숫자인 컬럼의 **최소값**
5. 데이터타입이 숫자인 컬럼의 **하위 25% 값**
6. 데이터타입이 숫자인 컬럼의 **하위 50% 값**
7. 데이터타입이 숫자인 컬럼의 **하위 75% 값**
8. 데이터타입이 숫자인 컬럼의 **최대값**

데이터프레임 둘러보기 : 요약 통계 df.describe()

bikes.describe(include=['object', 'datetime64'])
1 **2**

```
bikes.describe(include=['object', 'datetime64'])
                        1          2
```

	자전거번호	대여일시	대여 대여소명
3 count	2215632	2215632	2215632
4 unique	19029	42853	1543
5 top	SPB-17237	2019-06-03 18:07:00	뚝섬유원지역 1번출구 앞
6 freq	292	266	12617
7 first	NaN	2019-06-01 00:00:00	NaN
8 last	NaN	2019-06-30 23:59:00	NaN

1. 데이터타입이 **'object'인 경우 – 문자열 등**
2. 데이터타입이 **날짜 데이터인 경우**
3. 각 컬럼의 **데이터 총 개수**
4. 각 컬럼의 **중복되지 않은 고유값의 갯수**
5. 각 컬럼의 **최대 빈도수를 가진 값**
6. **최대 빈도수**
7. 날짜 데이터인 경우 **최초 값**
8. 날짜 데이터인 경우 **마지막 값**

데이터프레임 관찰 : 각 컬럼 누락값 검사

bikes.isnull().sum()

bike_ride 데이터프레임의 각 컬럼의 갖는 누락값들의 합을 보여준다.

bikes.isnull()

- 해당 값이 null이면 True
- 해당 값이 null이 아니면 False

bikes.isnull().sum()

- True는 1, False는 0
- sum()을 수행해서 0보다 큰 수가 나오면 이 수가 해당 컬럼에서 null값의 개수가 된다.

```
bikes.isnull()
```

	자전거번호	대여일시	대여 대여소번호	대여 대여소명	이용시간	이용거리
0	False	False	False	False	False	False
1	False	False	False	False	False	False
2	False	False	False	False	False	False
3	False	False	False	False	False	False
4	False	False	False	False	False	False
...	...	...	...	...	...	...
2215627	False	False	False	False	False	False
2215628	False	False	False	False	False	False
2215629	False	False	False	False	False	False
2215630	False	False	False	False	False	False
2215631	False	False	False	False	False	False

2215632 rows × 6 columns

```
bikes.isnull().sum()
```

```
자전거번호          0
대여일시            0
대여 대여소번호     0
대여 대여소명       0
이용시간            0
이용거리            0
dtype: int64
```

데이터프레임의 처음과 마지막 5개 행 보여주기

bikes.head()

데이터프레임의 처음 5개 행을 보여준다.

bikes.tail()

데이터프레임의 끝의 5개 행을 보여준다.

```
bikes.head()
```

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

```
bikes.tail()
```

	자전거번호	대여일시	대여 대여소번호	대여 대여소명	이용시간	이용거리
2215627	SPB-22438	2019-06-24 07:40:00	240	문래역 4번출구 앞	13	900
2215628	SPB-24455	2019-06-25 07:34:00	240	문래역 4번출구 앞	6	880
2215629	SPB-24557	2019-06-26 08:19:00	240	문래역 4번출구 앞	7	930
2215630	SPB-00649	2019-06-27 07:38:00	240	문래역 4번출구 앞	11	1030
2215631	SPB-14209	2019-06-28 07:37:00	240	문래역 4번출구 앞	6	1010

나 지금 어느 단계를 공부하는 거지?

1.문제정의

2.데이터수집

3.데이터 가공

4.데이터 모델링

5.시각화 및 탐색

단계 2 : 데이터프레임 관찰하기

데이터프레임 정보얻기 -> **bikes.info()**

데이터프레임 요약통계 -> **bikes.describe()**

누락값 검사 : **bikes.isnull().sum()**

데이터프레임 처음 보여주기 -> **bikes.head()**
데이터프레임 마지막 보여주기 -> **bikes.tail()**

퀴즈

퀴즈를
풀어봅시다

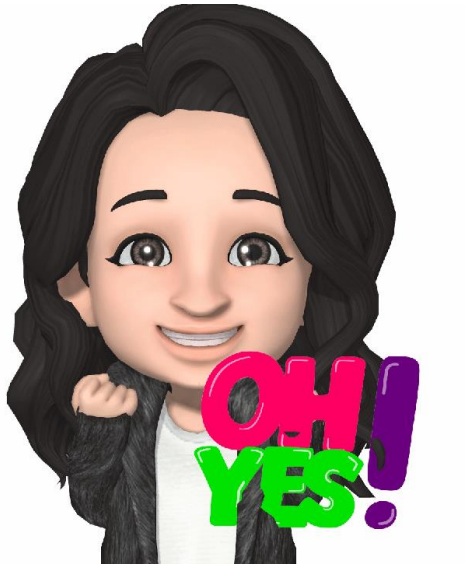
1. 데이터프레임의 정보를 알려주는 명령어는 ?

2. 데이터프레임에서 수치데이터의 요약 통계를 보여주는 명령어는 ?

3. bikes 데이터프레임에서 누락값을 검사하는 명령어 구문을 쓰세요.

4. 판다스 자료형 중에 문자 또는 문자열을 나타내는 것은?

실습 시작

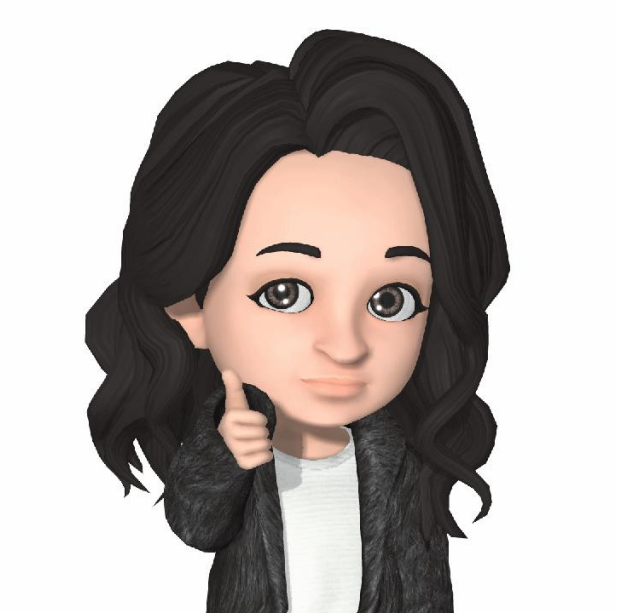


GD쌤

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 4회차 수업내용을 배워 보았습니다.

다음 시간에는 5회차 수업내용으로 시간 개념에 따른 데이터 분석을 위한 컬럼들을 추가해 보겠습니다.

수고 많으셨어요. 다음 시간에 만나요.