Elixirr

파이썬으로 배우는 따릉이 데이터 분석과 시각화

8회차

시각화 및 고찰

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

- **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 : 시간 개념에 따른 따릉이 이용패턴 분석 및 시각화

단계 2 : 장소적 특징에 따른 따릉이 이용패턴 분석 및 지도 시각화

단계 3 : 시간 개념 x 장소적 특징 연관 분석 후 시각화

단계 3 : 주말과 평일에 인기 있는 대여소 상위 50개 지도에 표시해보기

대여시간대 X 요일 따릉이 이용건수

피봇테이블에서 인덱스와 컬럼이 모두 필요한 경우로서 피봇테이블 수행 후 결과는 데이터프레임이다.

bikes.pivot_table(index='대여시간대', columns='요일', values='자전거번호', aggfunc='count')

1 2 3 4

```
bikes.pivot_table(\
                  index = '대여시간대', \
                  columns = '요일', \
                  values = '자전거번호', \
                  aggfunc = 'count'
                 )
```

✓ 0.3s

요일	금	목	수	월	일	토	화
대여시간대							
0	8353	9568	8624	7461	17025	13438	9907
1	6748	7508	7226	4846	11731	11232	7529
2	4762	5857	5328	3305	8804	8348	4497
3	3045	3992	3430	2102	8748	6221	3157

1. pivot_table의 인덱스로 정할 컬럼명 : '대여시간대'
2. pivot_table의 컬럼으로 정할 컬럼명 : '요일'
3. pivot_table의 값으로 정할 컬럼명 : '자전거번호'
4. 집계함수 : count()

꺾은선 그래프와 막대그래프의 문제점

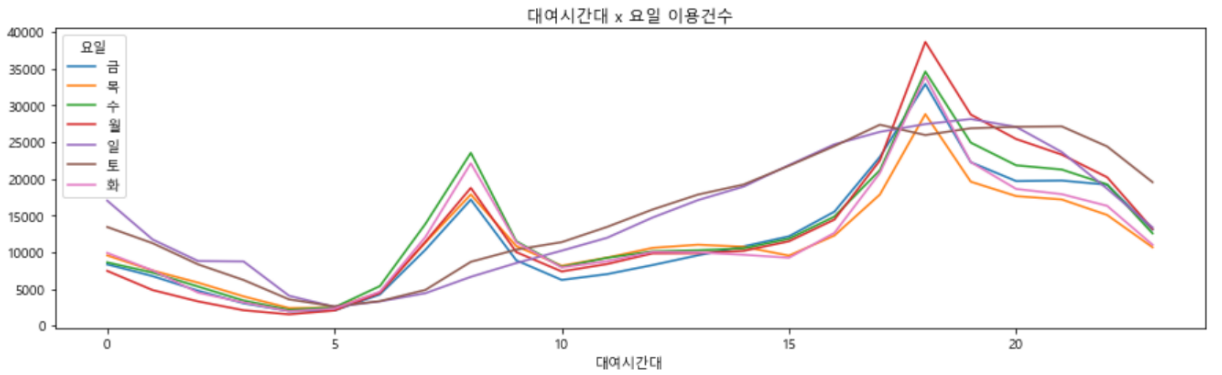
데이터프레임의 컬럼수가 너무 많으면 꺾은선이나 막대의 종류가 많아져서 그래프를 분석하기 어렵다. 다른 시각화 방안을 고려해보자.

```
hourly_dayofweek_ride
```
✓ 0.4s

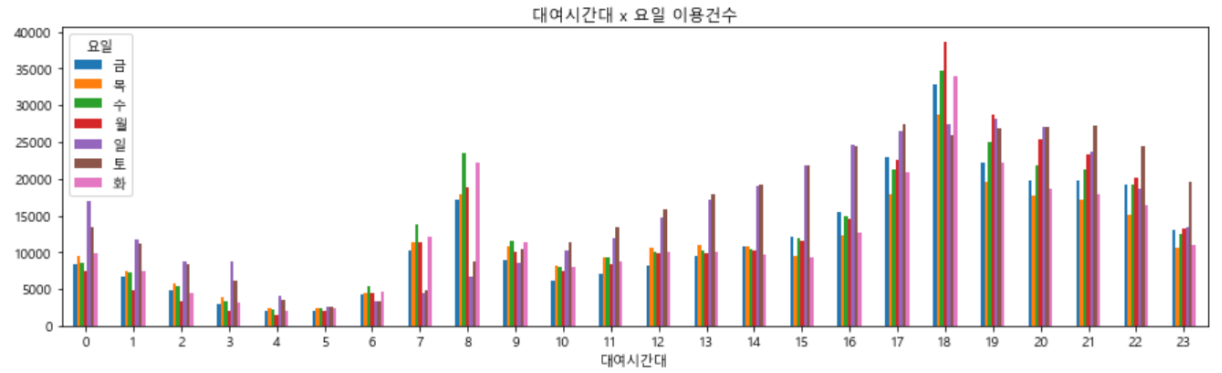
요일	금	목	수	월	일	토	화
대여시간대							
0	8353	9568	8624	7461	17025	13438	9907
1	6748	7508	7226	4846	11731	11232	7529
2	4762	5857	5328	3305	8804	8348	4497
3	3045	3992	3430	2102	8748	6221	3157
4	1979	2402	2160	1544	4076	3584	1971
5	2070	2451	2495	2069	2567	2619	2316
6	4252	4534	5393	4455	3311	3345	4633
7	10337	11324	13853	11418	4420	4869	12070
8	17164	17867	23542	18762	6651	8696	22119
9	8901	10833	11515	10017	8530	10357	11369
10	6232	8164	8006	7384	10215	11377	7924

컬럼수가 많다

```
hourly_dayofweek_ride.plot(kind='line', title = '대여시간대 x 요일 이용건수', figsize=(15, 4));
```
✓ 0.3s



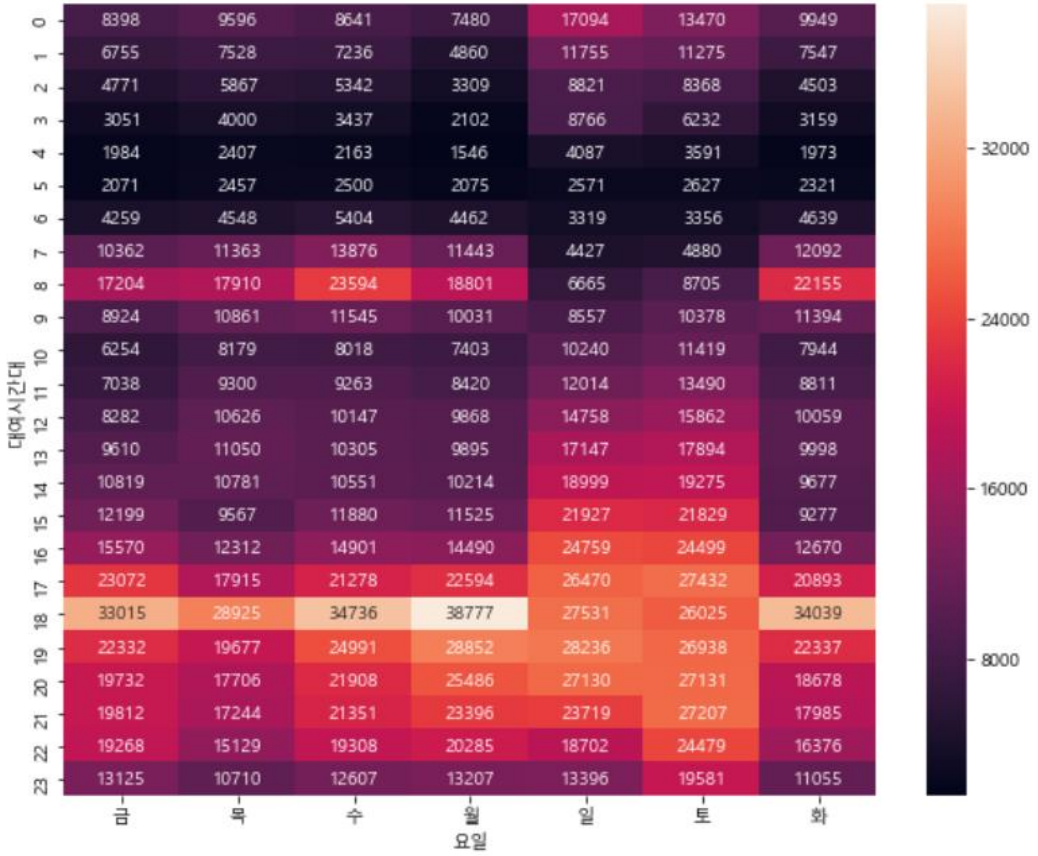

```
hourly_dayofweek_ride.plot(kind='bar', title = '대여시간대 x 요일 이용건수', \
                           figsize=(15, 4), rot=0);
```
✓ 0.6s



데이터 시각화 : 히트맵 Heatmap > sns.heatmap

X축과 Y축에 2개의 범주형 자료가 있을 때, 이들에 해당하는 값을 집계하고 집계한 값에 비례하여 색깔을 다르게 해서 2차원적으로 자료를 시각화 한다.

1 plt.figure(figsize=(10, 8))2
sns.heatmap(data=데이터프레임, annot=True, fmt='d');
3 4 5 6




```
plt.figure(figsize=(10, 8))
sns.heatmap(data=hourly_dayofweek_ride, annot=True, fmt='d');
```

1. plt.figure : 그래프 사이즈를 조절할 수 있는 명령어
2. figsize=(가로길이, 세로길이)
3. sns.heatmap : 집계한 수량을 색깔로 표시하는 그래프
4. data=데이터프레임 : 표시할 데이터프레임
5. annot=True : 색깔을 나타내는 칸에 해당 집계 수량도 표시
6. fmt='d' : 정수로 나타낸다는 표시

대여시간대 X 주말구분 따릉이 이용건수

피봇테이블에서 인덱스와 컬럼이 모두 필요한 경우로서 피봇테이블 수행 후 결과는 데이터프레임이다.

bikes.pivot_table(index='대여시간대', columns='주말구분', values='자전거번호', aggfunc='count')

1 2 3 4

```
weekdays_hourly_ride = bikes.pivot_table(\
                       index = '대여시간대', \
                       columns = '주말구분', \
                       values = '자전거번호', \
                       aggfunc = 'count')
weekdays_hourly_ride
```

✓ 0.4s

대여시간대	주말구분 주말	평일
0	30463	43913
1	22963	33857
2	17152	23749
3	14969	15726
4	7660	10056

1 pivot_table의 인덱스로 정할 컬럼명 : '대여시간대'

2 pivot_table의 컬럼으로 정할 컬럼명 : '주말구분'

3 pivot_table의 값으로 정할 컬럼명 : '자전거번호'

4 집계함수 : count()

꺾은선 그래프와 막대그래프

데이터프레임의 컬럼수가 너무 많으면 꺾은선이나 막대의 종류가 많아져서 그래프를 분석하기 어렵다. 다른 시각화 방안을 고려해보자.

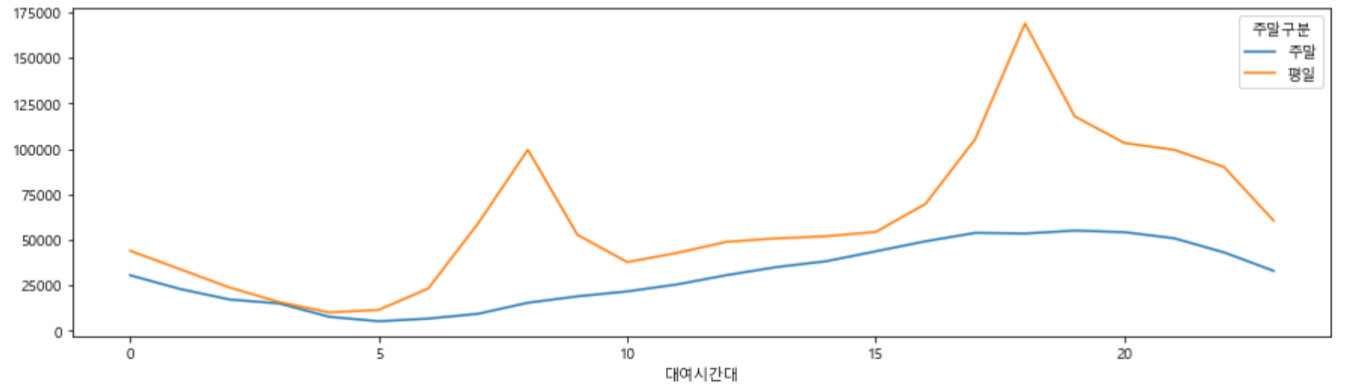
```
weekdays_hourly_ride = bikes.pivot_table(\
                       index = '대여시간대', \
                       columns = '주말구분', \
                       values = '자전거번호', \
                       aggfunc = 'count')
weekdays_hourly_ride
```

✓ 0.4s

주말구분	주말	평일
대여시간대		
0	30463	43913
1	22963	33857
2	17152	23749
3	14969	15726
4	7660	10056

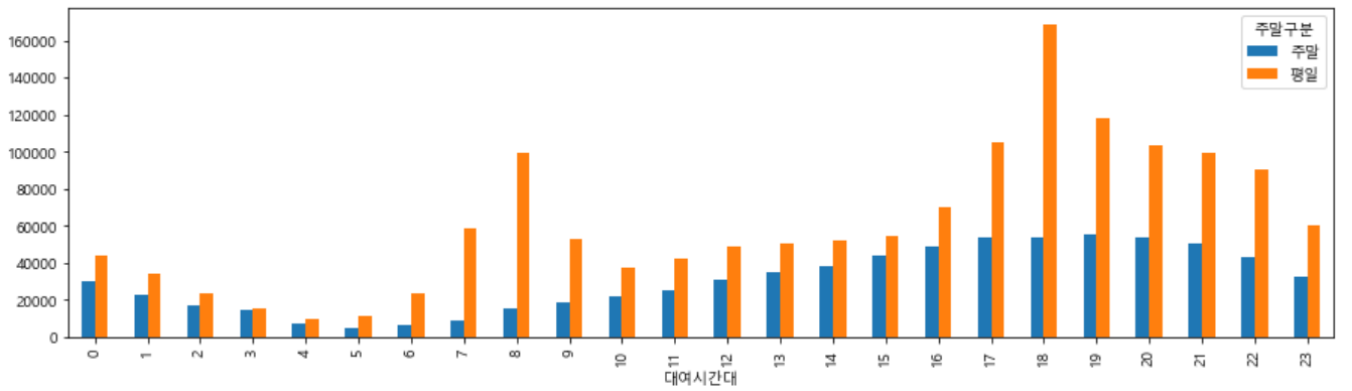
```
weekdays_hourly_ride.plot(kind='line', figsize = (15,4));
```

✓ 0.2s




```
weekdays_hourly_ride.plot(kind='bar', figsize = (15,4));
```

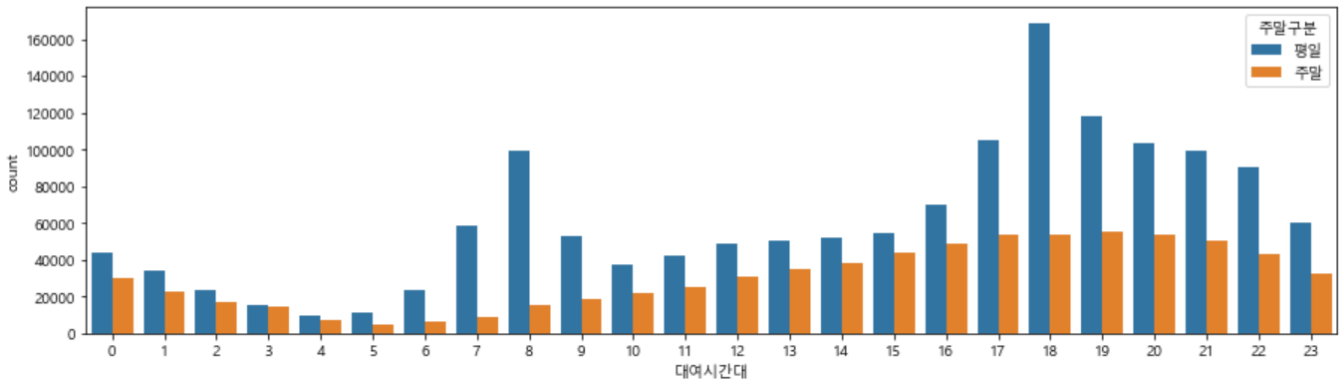
✓ 0.3s



Countplot & Heatmap


```
plt.figure(figsize=(15, 4))
sns.countplot(x='대여시간대', hue='주말구분', data=bikes);
```

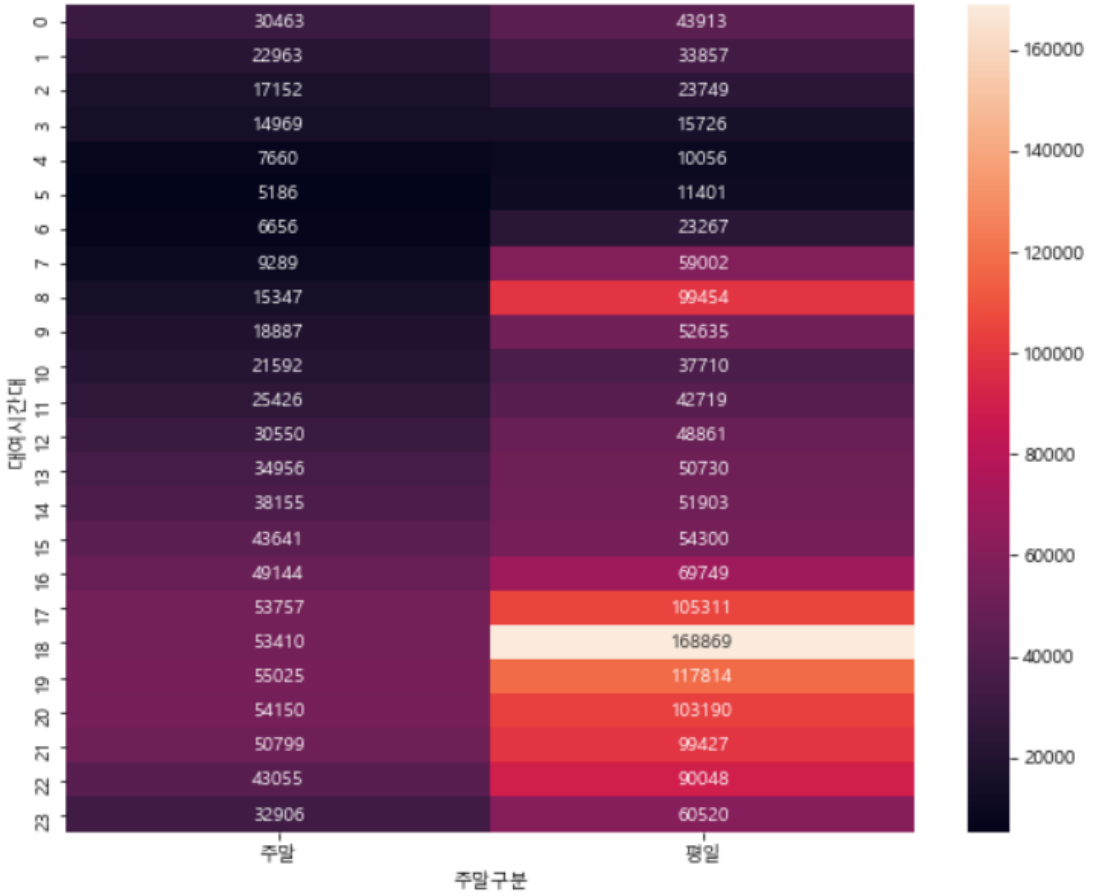
✓ 1.1s



- 주말과 평일은 두 그래프 모두 확연하게 차이를 보인다.
- 평일은 출퇴근 시간인 오전 8시와 오후 6시를 기준으로 이용건수가 많고, 주말은 오전 10시부터 이용건수가 증가한다.

```
plt.figure(figsize=(10, 8))
sns.heatmap(data=weekdays_hourly_ride, annot=True, fmt='d');
```

✓ 0.7s



나 지금 어느 단계를 공부하는 거지?

1.문제정의

2.데이터수집

3.데이터 가공

4.데이터 분석

5.시각화 및 탐색

단계 1 : 다양한 분석 명령어를 사용해서 시간 개념에 따른 따릉이 이용패턴 분석

내부적으로 집계해서 시각화 -> sns.countplot()

특정 컬럼들을 재구조화 -> bike_ride.pivot_table()

꺾은선 그래프, 막대 그래프, 히트맵 시각화

퀴즈

퀴즈를
풀어봅시다

1. 집계를 포함하는 다양한 시각화 명령어를 가지고 있고 high-level interface를 제공하는 라이브러리는?

2. 정돈된 데이터프레임에서 분석에 필요한 컬럼을 지정하면, 지정된 컬럼에 따라 내부적으로 데이터를 count해서 막대 그래프로 보여주는 명령어는 ?

3. 데이터분석을 위해 특정 컬럼을 인덱스와 컬럼으로 재구조화해서 집계함수를사용하는 명령어는 ?

4. 2개의 범주형 자료가 있을 때, 이들에 해당하는 값을 집계하고 집계한 값에 비례하여 색깔을 다르게 해서 2차원적으로 자료를 시각화 하는 명령어는 ?

실습 순서

시간 개념에 따른 데이터 분석

1. 일자별 따릉이 이용건수

2. 요일별 따릉이 이용건수

3. 대여시간대별 따릉이 이용건수

4. 대여시간대 x 요일 따릉이 이용건수

5. 대여시간대 x 주말구분 따릉이 이용건수

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 8회차 수업내용을 배워 보았습니다.

다음 시간에는 9회차 수업내용으로 장소적 특징에 따른 데이터 분석 및 시각화를 진행해 보겠습니다.

수고 많으셨어요. 다음 시간에 만나요.