http://news.tf.co.kr/read/life/1820185.htm

파이썬으로 배우는 따릉이 데이터 분석과 시각화

11회차 수업

데이터 분석 & 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**
- 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 : 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

대여시간대별 x 대여구별 자전거 이용건수는?


```
bikes.head(100)
```

✓ 0.7s Python

	자전거 번호	대여일시	대여 대여소 번호	대여 대여소명	이용시간	이용거리	일자	대여시간대	요일	주말구분	대여구	대여점 위도	대여점 경도
0	SPB-22040	2019-06-03 08:49:00	646	장한평역 1번출구 (국민은행앞)	27	1330	3	8	월	평일	동대문구	37.56	127.06
1	SPB-20387	2019-06-05 08:27:00	646	장한평역 1번출구 (국민은행앞)	12	1930	5	8	수	평일	동대문구	37.56	127.06
2	SPB-16794	2019-06-05 08:46:00	646	장한평역 1번출구 (국민은행앞)	6	1340	5	8	수	평일	동대문구	37.56	127.06
3	SPB-13926	2019-06-11 08:29:00	646	장한평역 1번출구 (국민은행앞)	7	1360	11	8	화	평일	동대문구	37.56	127.06
4	SPB-14638	2019-06-12 08:29:00	646	장한평역 1번출구 (국민은행앞)	5	1340	12	8	수	평일	동대문구	37.56	127.06
5	SPB-18588	2019-06-17 08:34:00	646	장한평역 1번출구 (국민은행앞)	8	1360	17	8	월	평일	동대문구	37.56	127.06
6	SPB-21148	2019-06-17 08:47:00	646	장한평역 1번출구 (국민은행앞)	22	1330	17	8	월	평일	동대문구	37.56	127.06
7	SPB-24533	2019-06-18 08:36:00	646	장한평역 1번출구 (국민은행앞)	6	1230	18	8	화	평일	동대문구	37.56	127.06
	SPB-	2019-06-18	646	장한평역 1번출구 (국민	11	1200	18	8	화	평일	동대문	37.56	127.06

1. bikes 데이터프레임에서 '대여시간대'와 '대여구' 컬럼을 살펴본다.
2. 예를 들어 bikes['대여시간대']의 값이 8이고 bikes['대여구']가 '동대문구' 해당하는 bikes['자전거번호']를 count한다.
3. 옆의 표에서 보면 이에 해당하는 자전거번호는 9개이다.

대여시간대 x 대여구별 피봇테이블

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

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

1 2 3 4

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

✓ 0.4s

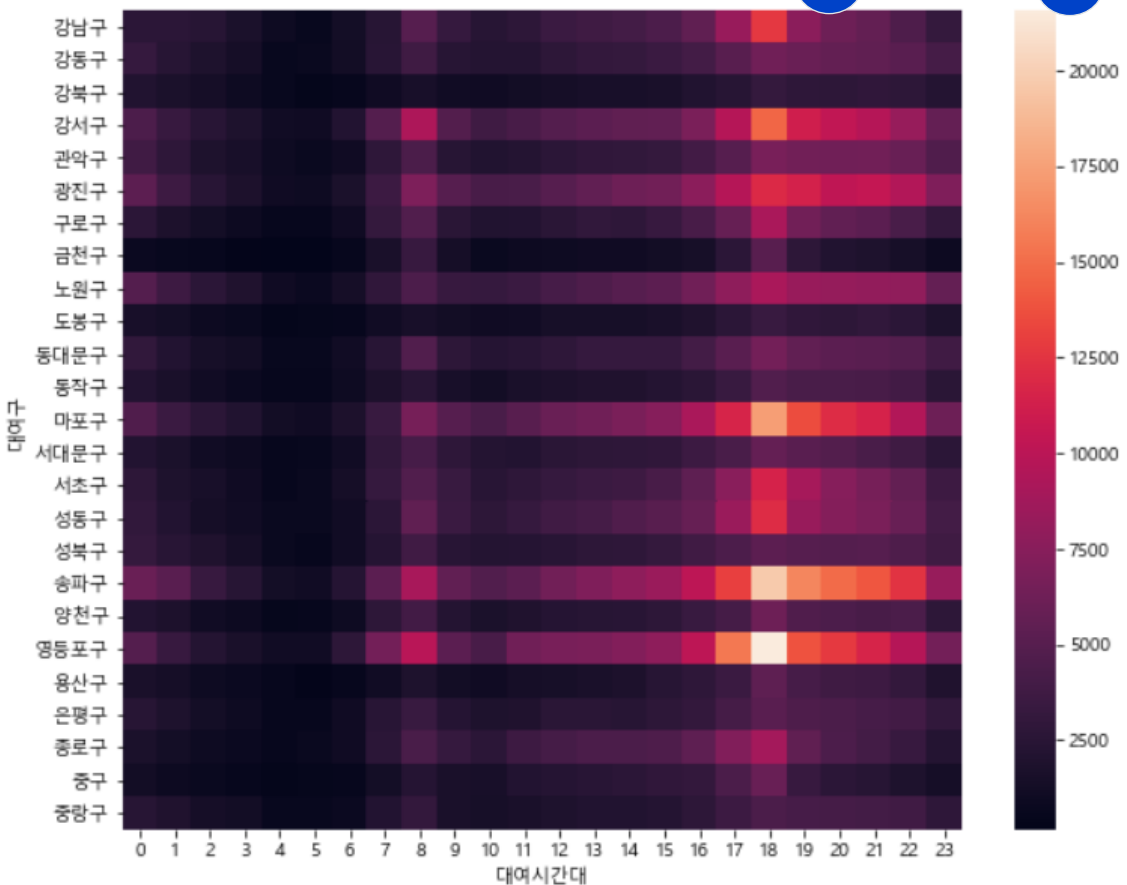
대여구	강남구	강동구	강북구	강서구	관악구	광진구	구로구	금천구	노원구	도봉구	...
대여시간대											
0	2643	3230	2066	4459	3769	5314	2650	790	4855	1529	...
1	2660	2342	1706	3306	2830	3621	1797	650	3632	1297	...
2	2360	1852	1378	2442	1899	2469	1291	496	2610	853	...
3	1689	1398	932	1909	1550	1830	853	328	1932	668	...

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

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');



5 6

```
# heatmap으로 표현한다. 흐린색일수록 이용건수가 많다.

plt.figure(figsize=(10, 8))
sns.heatmap(hourly_gu_use.T);
```

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

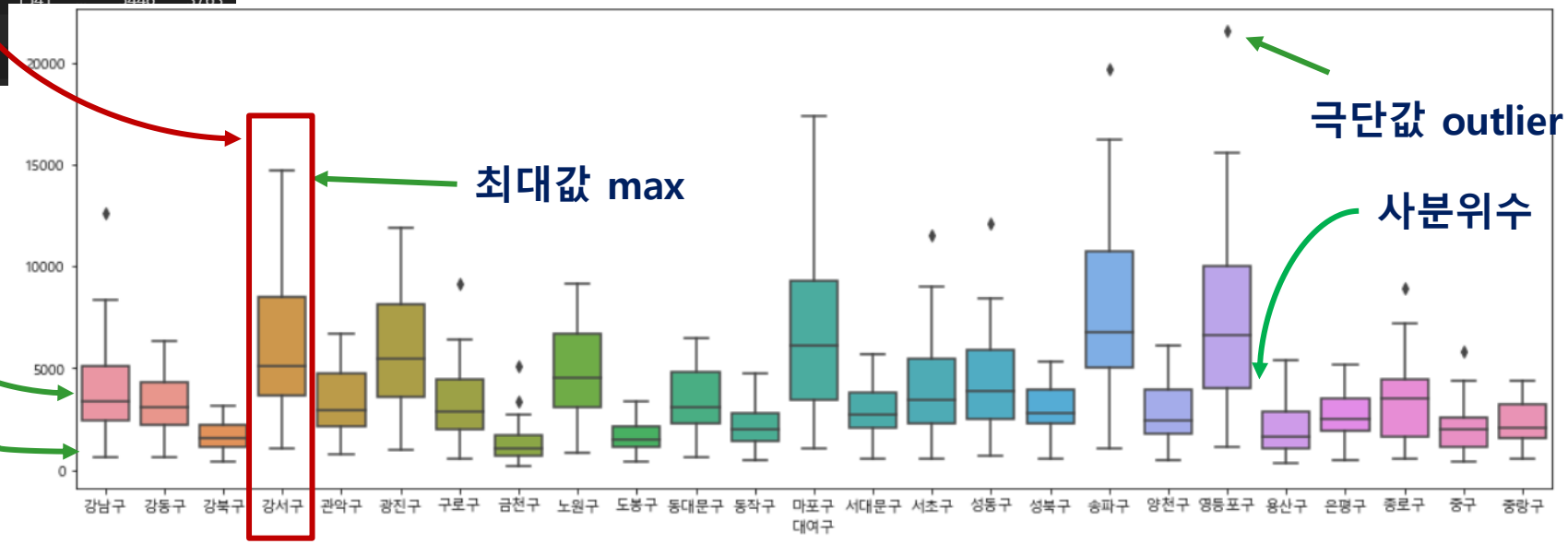
데이터 시각화 : 상자 그래프Box plot > sns.boxplot

상자 그래프는 다양한 통계량을 사용하여 데이터 분포를 파악하는 데 유용합니다.

대여구 대여시간대	강남구	강동구	강북구	강서구	관악구	광진구	구로구	금천구	노원구	도봉구	...	성동구	성북구
0	2643	3230	2066	4459	3769	5314	2650	790	4855	1529	...	3003	3165
1	2660	2342	1706	3306	2830	3621	1797	650	3632	1297	...	2096	2501
2	2360	1852	1378	2442	1899	2469	1291	496	2610	853	...	1394	1941
3	1689	1398	932	1909	1550	1830	853	328	1932	668	...	1103	1337
4	920	620	595	1152	990	998	572	173	1052	403	...	743	701
5	648	686	392	1072	745	992	543	161	829	416	...	698	522
6	1316	1173	499	2144	1042	1687	1081	511	1413	498	...	1092	1089
7	2404	2420	1047	4854	2800	3512	3091	1623	2841	1120	...	2588	2297
8	4951	3728	1533	9357	4350	6988	4691	3337	4491	1541		5446	3763
9	3202	2413	1160	4766	2414	4939	2550	1350	3259				
10	2376	2247	1078	3722	2037	4208	2071	678	3176				

```
# 박스 그래프로 표시하기
# seaborn 라이브러리를 사용하여 시각화

plt.figure(figsize=(18, 6))
sns.boxplot(data=hourly_gu_use);
```



나 지금 어느 단계를 공부하는 거지?

1.문제정의

2.데이터수집

3.데이터 가공

4.데이터 모델링

5.시각화 및 탐색

단계 3 : 시간 개념 x 장소적 특징 연관 분석 후 시각화

데이터 집계 -> df.pivot_table()

시본 라이브러리 명령어를 사용한 시각화

퀴즈

퀴즈를
풀어봅시다

1. 대여시간대와 대여구별 따릉이 이용건수를 나타내는 데이터프레임을 만들기 위해 여기서 사용한 명령어는 ?

2. 대여시간대와 대여구별 따릉이 이용건수를 2차원의 표에서 색깔로 표현해 주는 시각화 명령어는 ?

3. 대여시간대와 대여구별 따릉이 이용건수의 최대값, 최소값, 평균 등의 통계량을 박스로 표현해 주는 시각화 명령어는 ?

4. 데이터 분포에서 아주 멀리 떨어져 있는 값을 무엇이라고 하나요 ?

실습 순서

시간 x 장소 특징에 따른 데이터 분석

대여일자 x 대여구 별 이용건수 분석

대여시간 x 대여구 별 이용건수 분석

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 11회차 수업내용을 배워 보았습니다.

다음 시간에는 12회차 수업내용으로 주말과 평일에 이용건수가 많은 대여소 TOP 50을 찾고 시각화해 보겠습니다.

수고 많으셨어요. 다음 시간에 만나요.