

LearningSpoons Online

입문자를 위한 파이썬 데이터분석 & 시각화



- 01 데이터분석을위한파이썬기초
- 02 데이터분석을위한클래스-pandas
- 03 데이터시각화를위한클래스-seaborn, folium
- 04 Project1) "아직이트레이드에대해반대하는분계십니까?"
- 05 Project2) "이사람들다어디로가는거지??"
- 06 Project3) "나만의데이터지도만들기"

Section 01

데이터 분석을 위한 파이썬 기초

Unit 1 - 1

파이썬 준비하기

파이썬 설치하기(아나콘다)

아나콘다(Anaconda)? 파이썬 + 주요 라이브러리 + 입/출력 프로그램

Step1) 설치 환경 확인하기

- 내 컴퓨터 OS 확인하기: Mac / Windows / Linux.
- 운영체제 BIT 확인하기 : 32 bit / 64 bit

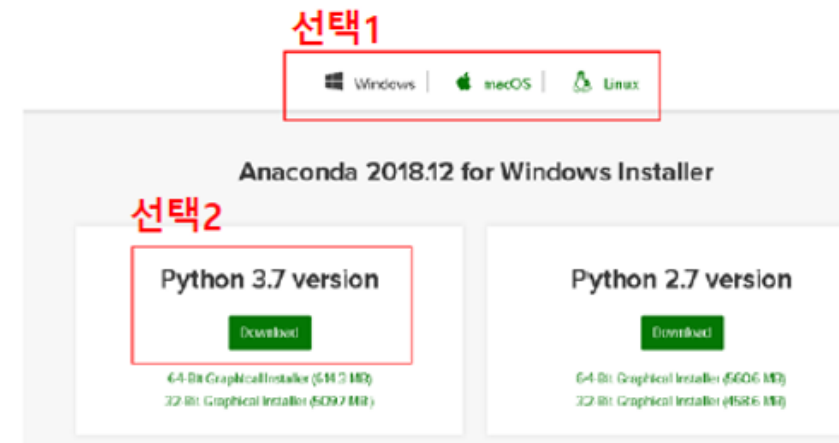


[운영체제 BIT 확인 방법] 윈도우 탐색기 → 내PC 우클릭 → 속성

파이썬 설치하기(아나콘다)

Step2) 아나콘다 다운로드

- <https://www.anaconda.com/download/>
- 파이썬 3.x 버전

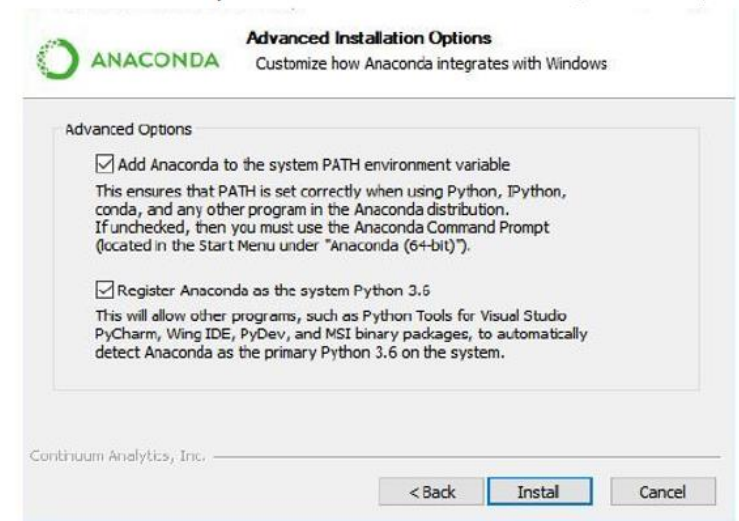
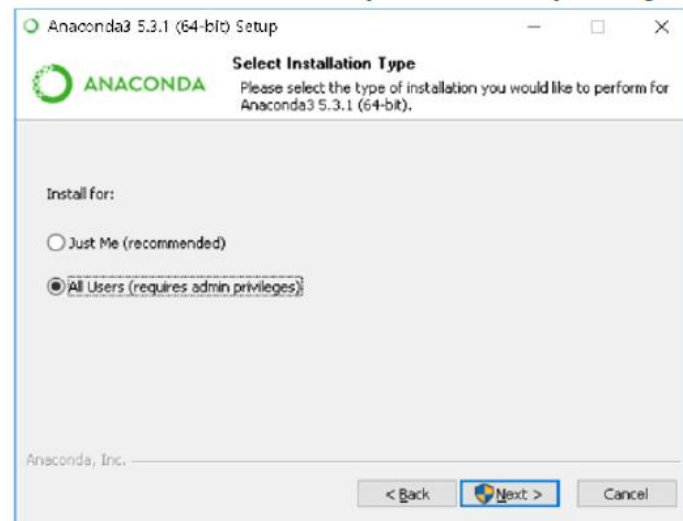


<Step 1에서 확인한 버전으로 다운로드>

파이썬 설치하기(아나콘다)

Step3) 아나콘다 설치

- Install for : All User(requires admin privileges)
- Advanced Options : 두 가지 모두 체크(윈도우 환경설정)



쥬피터 노트북 실행하기

실행방법

- (Mac / Linux) 커맨드 창에서 `jupyter notebook` 실행
- (windows) 시작 메뉴에서 `jupyter notebook`



<인터넷 브라우저가 실행되면서 localhost:8888/tree 접속 >

쥬피터 노트북 실행하기

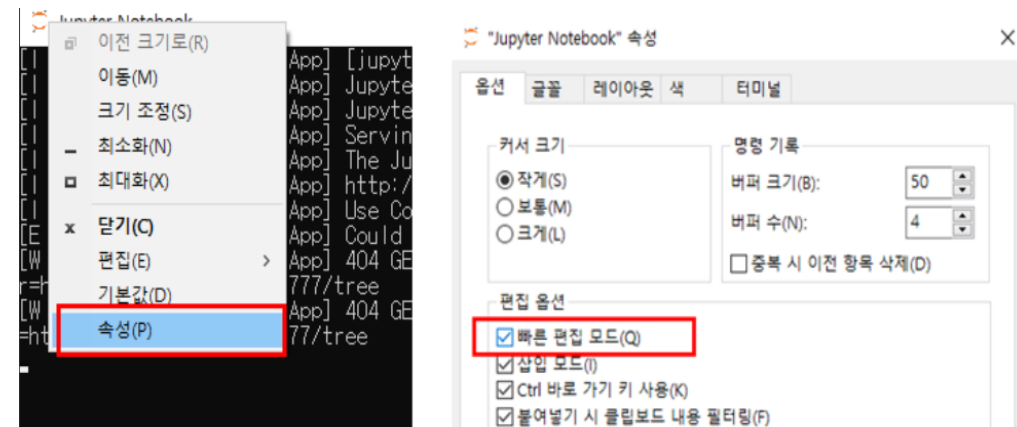
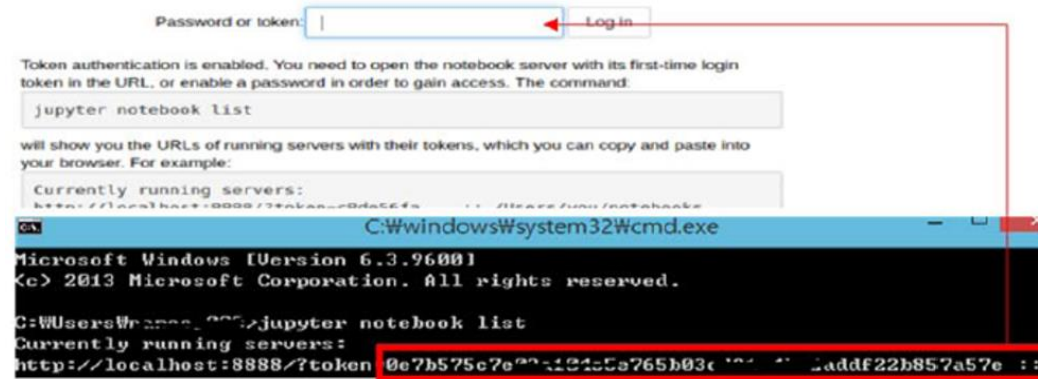
Q) 만약 토큰 값을 입력하라고 나온다면?

`http://localhost:8888/?token=토큰값`

토큰값 ← 복사&붙여넣기

Q) 만약 토큰값 복사하기가 되지 않는다면?

커맨드창 메뉴바 마우스 우클릭 → “속성” → “빠른 편집 모드” 체크



Unit 1-2

데이터 분석을 위한 파이썬 자료형

변수명 = 데이터

데이터 타입	내용	예시
int	정수	1, 2, 3, 4, 5
float	실수(소수점 존재)	3.14, 9.99
str	문자 나열	"파이썬", '대한민국'
list	순서대로 나열된 그룹	[1, 2, 3, 4, 5], ["가", "나", "다"]
dict	이름(key)별로 정의된 그룹	{ '이름' : '홍길동', '전화번호' : '010-0000-0000', '주소' : '대한민국 서울' }

Unit 1-3

데이터 분석을 위한 기본 문법

파이썬 기본 문법

- 반복문
- 조건문
- 문자열포매팅
- 함수

Section 02

데이터 분석을 위한 클래스 - pandas

Unit 2-1

판다스 기본 사용법

판다스 데이터 구조

DataFrame

		columns		
		국어	영어	수학
index	1번	70	80	75
	2번	68	95	55
	3번	90	100	95

columns = ['국어', '영어', '수학']

index = ['1번', '2번', '3번']

Series

	국어
1번	70
2번	68
3번	90

	영어
1번	80
2번	95
3번	100

	수학
1번	75
2번	55
3번	95

판다스 기본 사용법

- 데이터 파일 읽기 : `read_excel()`, `read_csv()`
- 데이터 선택하기 : `df.loc()`, `df.iloc()`
- 인덱스 / 컬럼 변경하기 : `columns/index`, `reset_index()`

Unit 2-2

데이터 병합하기

데이터 추가/병합하기

- 컬럼 데이터 추가하기
- 두 데이터 병합하기: `pd.merge()`

데이터 병합하기 pd.merge()

- "left": 왼쪽 데이터(A)에 속한 데이터 기준
- "right": 오른쪽 데이터(B)에 속한 데이터 기준
- "inner": 양쪽 (A와 B)에 모두 속한 데이터 기준
- "outer": 모든 데이터 기준

```
pd.merge( A , B , how = "left" , left_on = "A컬럼명" , right_on = "B컬럼명" ,  
         left_index = True , right_index = True )
```

A

	국어	영어	수학
1번	70	80	75
2번	68	95	55
3번	90	100	95

B

	과학	사회
1번	70	80
2번	80	85
4번	95	100
5번	90	70

왼쪽(A) 테이블
기준값 지정

오른쪽(B) 테이블
기준값 지정

Unit 2-3

정리/집계하기

데이터 정리/집계 하기

- 조건에 만족하는 데이터 살펴보기: `df [조건]`
- 특정 기준으로 테이블 변환하기: `df.pivot_table()`
- 정렬하기 : `df.sort_values()`

Unit 2-4

실습) 영어 이름 트렌드 살펴보기

Q) 남자는 James, 여자는 Mary 어때?

- 최근 영어 이름 트렌드 살펴보기

Section 03

데이터 시각화를 위한 클래스 - seaborn, folium

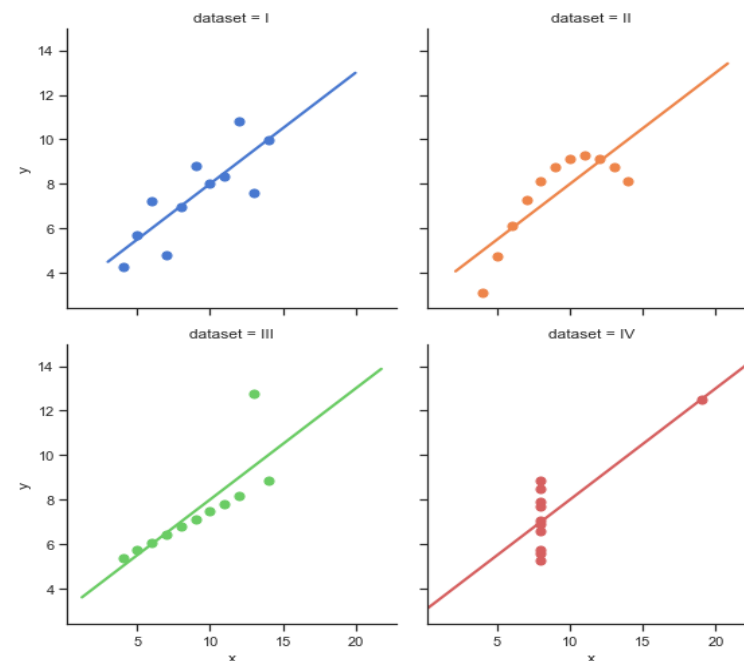
Unit 3-1

seaborn으로 시각화 하기

데이터 시각화의 중요성

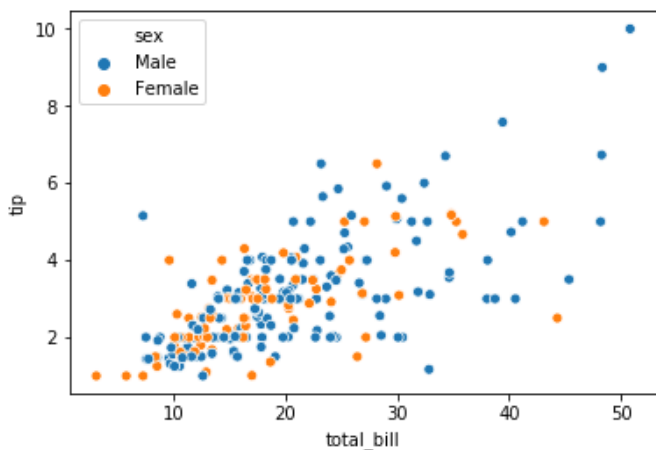
통계량이 모두 동일한 데이터 set → ~~똑같다??~~

Property	Value
Mean of x	9
Sample variance of x	11
Mean of y	7.5
Sample variance of y	4.125
Correlation between x and y	0.816
Linear regression line	$y = 3.00 + 0.500x$

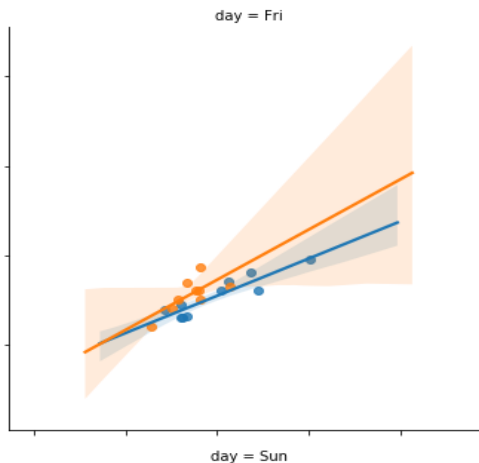


데이터 타입별 시각화: 수치형 x 수치형

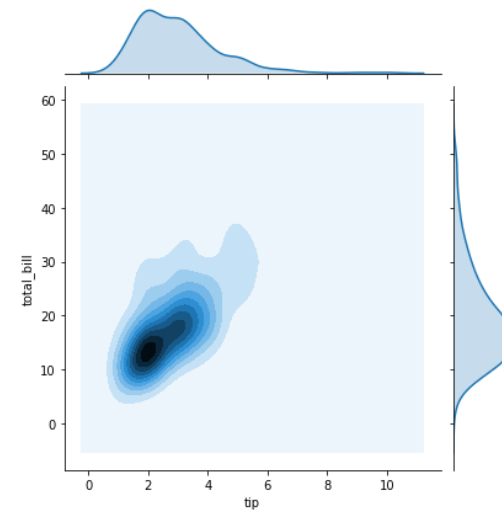
scatterplot



Implot

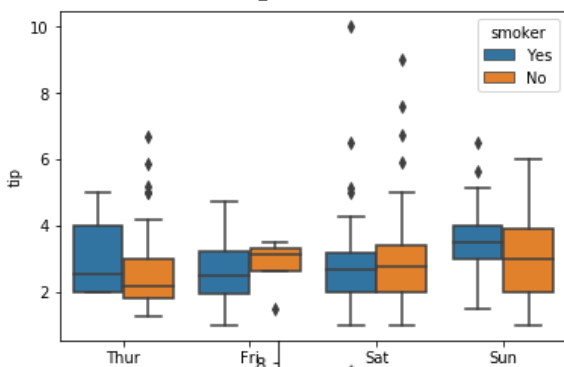


jointplot

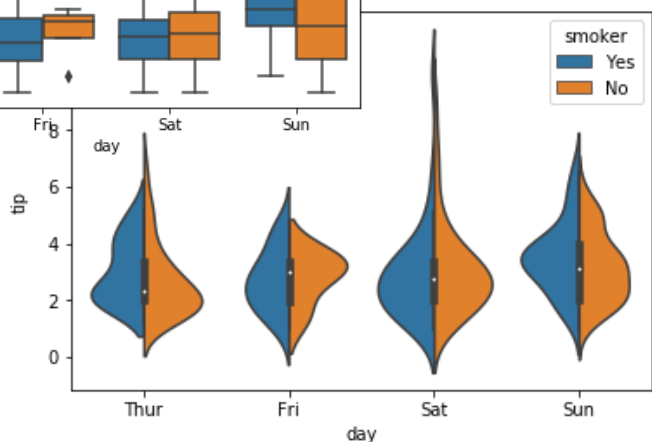


데이터 타입별 시각화: 수치형 x 카테고리형

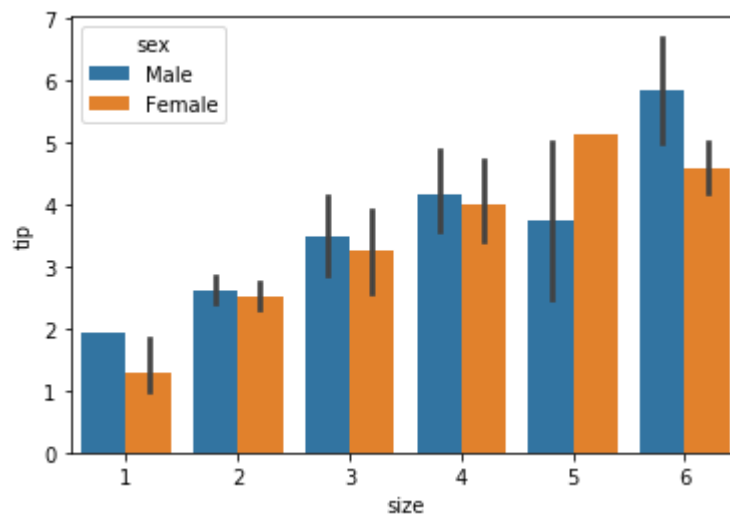
boxplot



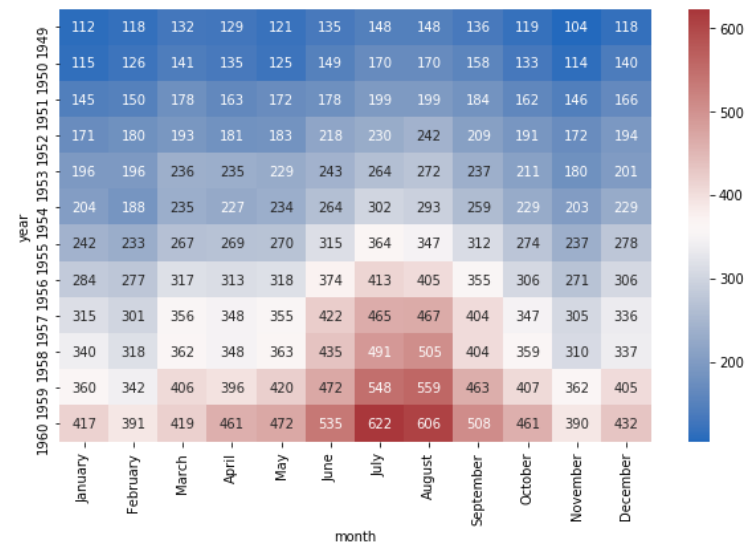
violinplot



barplot

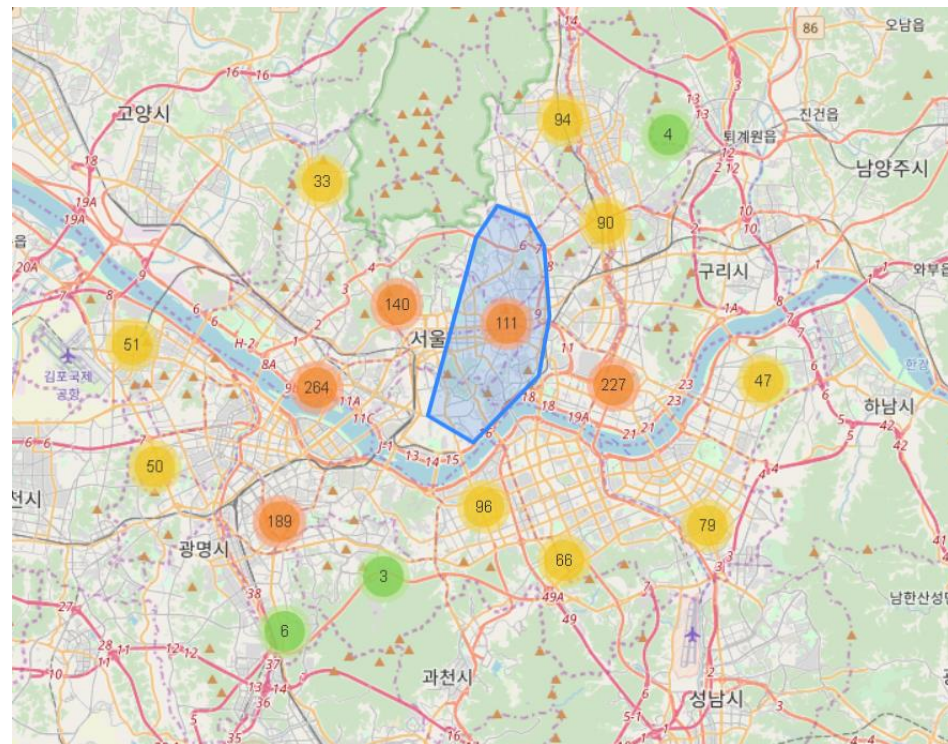


heatmap



데이터 타입별 시각화: 수치형 x 위치정보

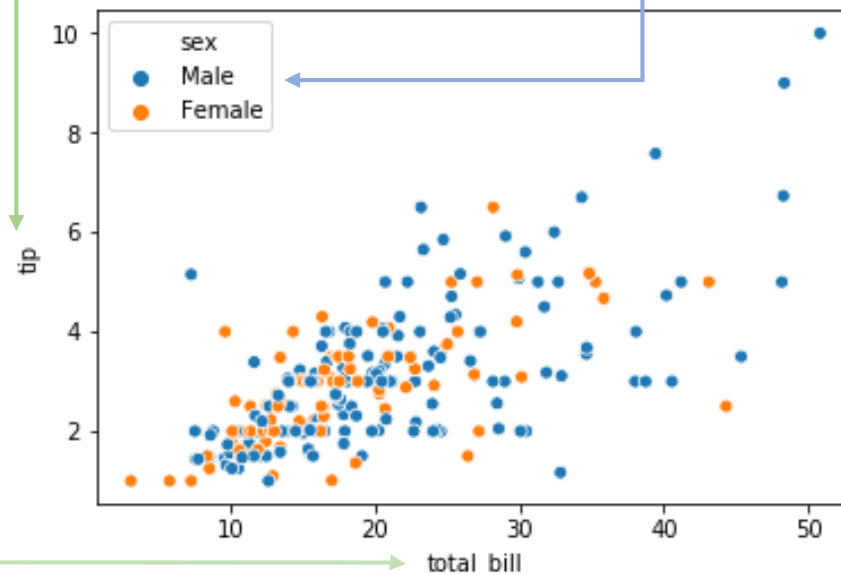
folium 라이브러리 활용



seaborn 함수 기본 형태

`sns.scatterplot` (`data = 데이터프레임`, `x = 'total_bill'`, `y = 'tip'`, `hue = 'sex'`)
그래프 종류

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4



heatmap 사용법

```
sns.heatmap( data = 데이터프레임,  
             annot = True, fmt = '.1f',  
             cmap = 'RdBu_r'  
            )
```

히트맵에서 실제 숫자 표현,
fmt 옵션을 조정하여 표시 형태 정리

fmt = '.1f' → 소수점첫째자리까지 표시
fmt = '.2f' → 소수점 둘째자리까지 표시
fmt = '.0f' → 정수로 표시

히트맵 색상 컬러 조정
추천색상: Reds, Blues, vlag, Pastel1

Unit 3-2

folium으로 지도시각화 하기

지도시각화 by folium

- 지도 생성하기
- 마커 추가하기
- 원 추가하기
- 툴팁/팝업 정보 추가하기

Section 04

“이 트레이드에 대해 반대하는 분 계십니까?”

– 야구 데이터 분석을 통한 KBO 타자 비교분석 –

Unit 4-1

Who is the Best player?

Q) 최고의 타자는?

지표	의미	계산식
타율	타격에 성공해 살아나는 정도	$= \text{타격 성공 횟수} / \text{타격 기회 수}$ $= \text{안타 수} / \text{타수}$
출루율	살아서 나가는 정도	$= \text{진루 성공 횟수} / \text{진루 기회 수}$ $= (\text{안타} + \text{볼넷} + \text{몸에맞는볼}) / (\text{타수} + \text{볼넷} + \text{몸에맞는볼} + \text{희생플라이})$
장타율	타격에 성공해 멀리 살아나는 정도	$= \text{진루한 베이스 수} / \text{타격 기회 수}$ $= \text{루타 수} / \text{타수}$ <p>* 타율에 거리 개념 추가, 2루타 = 1루타 x 2</p>
OPS	살아서 멀리 나가는 정도	$= \text{출루율} + \text{장타율}$

Unit 4-2

여름에 힘 떨어지는 타자? vs 꾸준한 타자

Q) 여름에 힘떨어지는 타자?? 꾸준한 타자?

- 월별 실적 변화 비교 분석하기

Unit 4-3

왜 우리 팀만 만나면 잘하는거야?

Q) 우리팀을 상대로 가장 강한 타자는?

- 팀별 실적 비교 분석하기

Section 05

“이 사람들 다 어디로 가는거지??”

- 지하철 이용현황 공공데이터 분석 -

Unit 5-1

여러 개의 엑셀 파일 통합 정리하기

엑셀 파일 통합하기

- 내 컴퓨터 파일 조회하기(os)
- 반복문을 통한 파일 열기 / 병합하기
- 저장하기

Unit 5-2

일자 별 승객 수 살펴보기

Q) 언제가 가장 승객수가 많을까?

- 월별 / 일자별 / 요일별 승객수 비교

Unit 5-3

지하철역별 승객 수 살펴보기

Q) 사람들은 어디서 탈까?? 어디로 갈까??

- 역별 승하차 인원 비교

Section 06

“나만의 데이터 지도 만들기”

– “따릉이” 자전거 조회 API 를 활용한 지도 만들기 –

Unit 6-1

오픈API 신청하기

데이터 수집 방법

- 이미 가지고 있는 데이터 준비/정리
- 웹에 있는 자료 Copy&Paste → 크롤링
- 회사/기관/개인에게 자료 요청하고 받기 → API

온라인 데이터 수집 방법

- 웹에 있는 자료 Copy&Paste → 크롤링
- 회사/기관/개인에게 자료 요청하고 받기 → API

Q) API 는 어떻게 사용할 수 있나요?

- 회사/기관/개인이 만든 규칙 활용

 - 주소 / 명령어 등의 기준에 따라 요청

- 일반적으로 API 사용량 등을 관리하기 위해 사용승인 필요

 - API Key 를 획득 한 후, 함께 전달 필요

공공데이터 API 사용 신청하기

1. 서울 열린 데이터광장(<http://data.seoul.go.kr/>) 회원가입/로그인
2. 서울특별시 공공자전거 실시간 대여정보

(<http://data.seoul.go.kr/dataList/OA-15493/A/1/datasetView.do>)

3. 인증키 신청

미리보기

닫힘 -

Open API

샘플 URL

Open API 이용안내 인증키 신청 명세서 다운로드

샘플 URL

서울특별시 공공자전거 실시간 대여정보
<http://openapi.seoul.go.kr/8068/인증키/json/bikeList/1/5/>

예제

```
[{"rentBikeStatus":{"list_total_count":5,"RESULT":{"INFO-000":{"MESSAGE":"정상 처리되었습니다."},"row":[{"rackTotCnt":"22","stationName":"102. 망원역 1번출구 앞","parkingBikeTotCnt":"14","shared":"0","stationLatitude":"37.55564880","stationLongitude":"126.91062927","stationId":"ST-4","rackTotCnt":"16","stationName":"103. 망원역 2번출구 앞","parkingBikeTotCnt":"3","shared":"0","stationLatitude":"37.55495071","stationLongitude":"126.91083527","stationId":"ST-5","rackTotCnt":"15","stationName":"104. 합정역 1번출구 앞","parkingBikeTotCnt":"10","shared":"0","stationLatitude":"37.55062866","stationLongitude":"126.91498566","stationId":"ST-6","rackTotCnt":"7","stationName":"105. 합정역 5번출구 앞","parkingBikeTotCnt":"1","shared":"0","stationLatitude":"37.55000687","stationLongitude":"126.91482544","stationId":"ST-7","rackTotCnt":"12","stationName":"106. 합정역 7번출구 앞","parkingBikeTotCnt":"8","shared":"0","stationLatitude":"37.54864502","stationLongitude":"126.91282654","stationId":"ST-8"}]}]}
```

요청인자

변수명	타입	변수설명	값설명
KEY	String(필수)	인증키	OpenAPI에서 발급된 인증키
TYPE	String(필수)	요청파일타입	xml:xml,xml파일 :xml, 예제파일 :xls,json파일:json
SERVICE	String(필수)	서비스명	bikeList
START_INDEX	INTEGER(필수)	요청시작위치	정수 입력(페이징 시작번호입니다:데이터 행 시작번호)
END_INDEX	INTEGER(필수)	요청종료위치	정수 입력(페이징 끝번호입니다:데이터 행 끝번호)

공공데이터 API 사용 신청하기

4. 가입 신청서 작성

- 사용URL: localhost
- 이메일: 본인 이메일 작성
- 활용용도 및 내용 등록

* 필수 입력항목

* 서비스(사용)환경	<input type="radio"/> 웹사이트개발 <input type="radio"/> 앱개발 (모바일 솔루션 등) <input checked="" type="radio"/> 연구 (논문 등) <input type="radio"/> 기타참고자료
* 사용URL (150자 이내)	<input type="text" value="localhost"/>
* 관리용 대표 이메일 (단체/기업/기관)	<input type="text" value="datago0ba0"/> @ <input type="text" value="gmail.com"/> <input type="button" value="선택"/>
* 활용용도	<input type="text" value="데이터분석"/>
* 내용 (200자 이내)	<div><input type="text" value="데이터분석"/></div> <div>5/200자</div>

공공데이터 API 사용 신청하기

5. 인증키 확인

- 사용URL: localhost
- 이메일: 본인 이메일 작성
- 활용용도 및 내용 등록

서울 열린데이터 광장

공공데이터 통계 소식&참여 **나의화면** 이용약관

로그아웃 사이트맵

나의화면 > 인증키신청 **인증키관리** 문의 관리 갤러리 관리

모든 서울시민을 위한 공공데이터

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

데이터셋	서비스	OpenAPI
5,326	12,093	4,560

6. 인증키 복사하기

일반 인증키(2)		실시간 지하철 인증키 발급대기현황(0)	예약 인증키 발급 대기현황(0)
정상	42- [redacted] 64c (2020/03/09)		인증키 복사 이용약관 활용갤러리 등록
정상	6c4 [redacted] 4767 (2019/10/03)		인증키 복사 이용약관 활용갤러리 등록

Unit 6-2

API를 통한 데이터 받기

API 활용하여 데이터 수집하기

- 데이터 요청하기: `requests.get(url)`
- json 데이터 정리하기
- 엑셀파일에 저장하기

Unit 6-3

따릉이 지도 만들기

따릉이 지도 만들기

- 데이터 불러오기
- 지도 시각화