

데이터분석입문

Lecture 09. 대중교통 데이터와 그래프를 이미지로 저장하기

동양미래대학교 인공지능소프트웨어학과 강 환수

목차



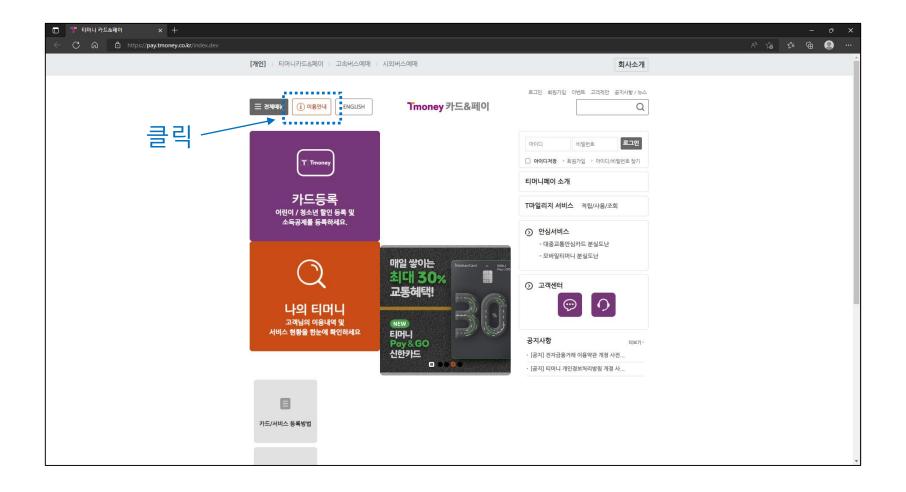
❖ 01. 대중교통 데이터 준비하기

❖ 02. savefig 함수로 그래프를 이미지로 저장하기

02. savefig 함수로 그래프를 이미지로 저장하기

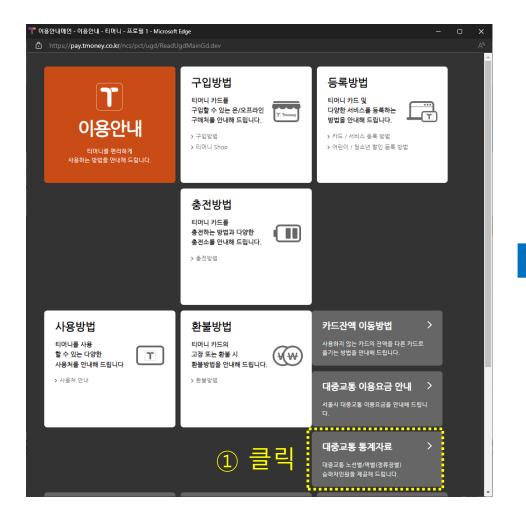


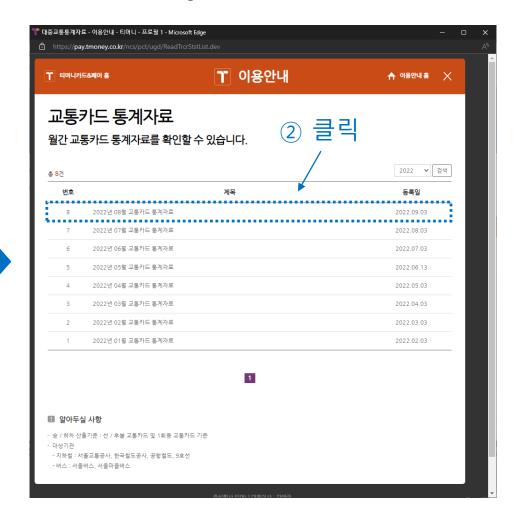
- ❖ ① 대중교통 데이터 내려받기 (1/4)
 - 대중교통 통계자료 수집 → 티머니 홈페이지(https://pay.tmoney.co.kr/index.dev) [①이용안내]





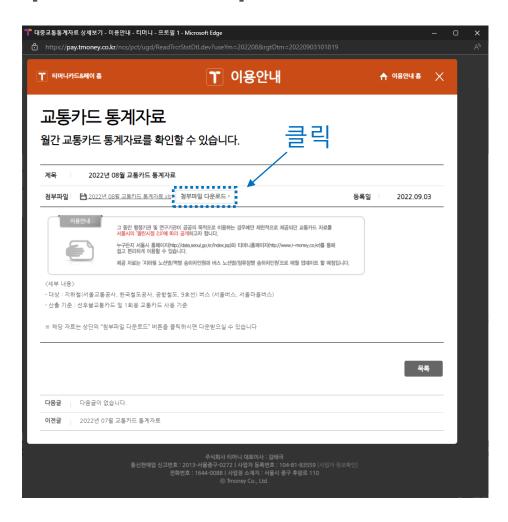
- ❖ ① 대중교통 데이터 내려받기 (2/4)
 - [대중교통 통계자료] 버튼 클릭 → [2022년 08월 교통카드 통계자료] 버튼 클릭







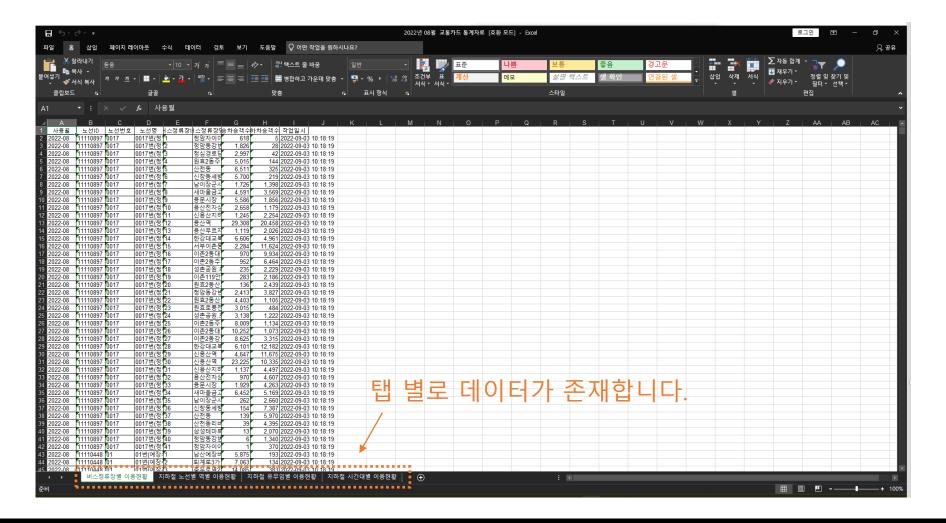
- ❖ ① 대중교통 데이터 내려받기 (3/4)
 - [첨부파일 다운로드] 버튼 클릭



다운로드(Downloads) 폴더에 "2022년 08월 교통카드 통계자료.xls" 엑셀 파일이 저장됩니다.

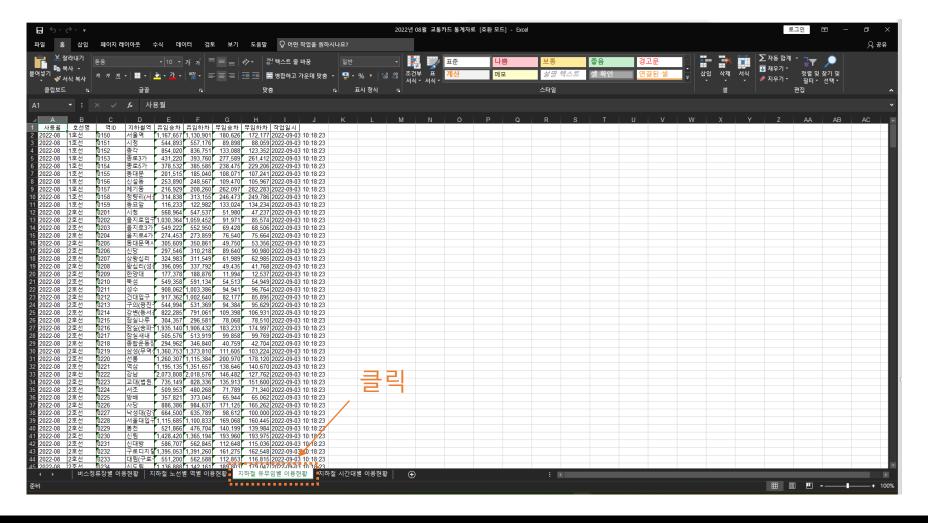


- ❖ ① 대중교통 데이터 내려받기 (4/4)
 - 버스정류장별, 지하철 노선별 역별, 지하철 유무임별, 지하철 시간대별 이용현황 데이터가 존재합니다.



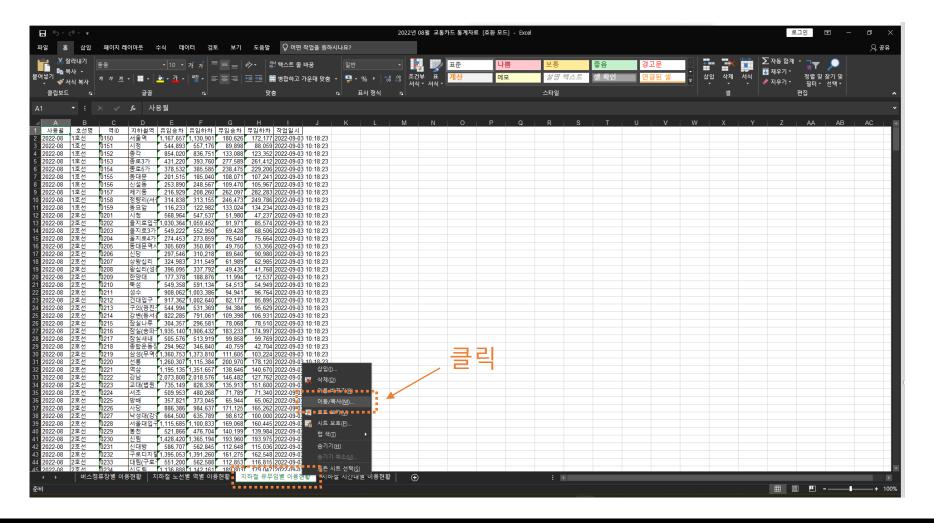


- ❖ ② 지하철 유무임별 이용현황 데이터 정제하기 (1/6)
 - 4개의 탭 중 "지하철 유무임별 이용현황" 탭을 클릭합니다.



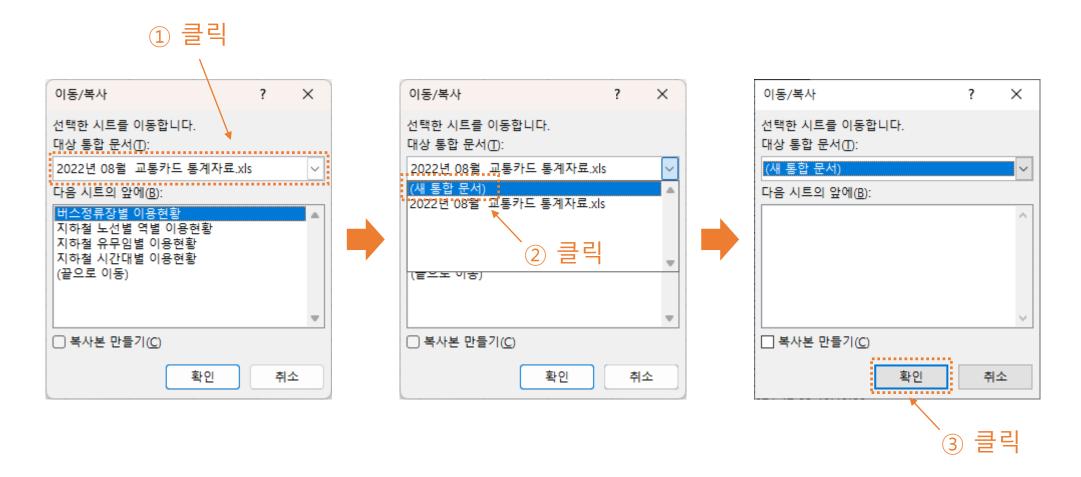


- ❖ ② 지하철 유무임별 이용현황 데이터 정제하기 (2/6)
 - 워크시트 이름 탭을 마우스 오른쪽 단추로 클릭하고, [이동/복사(M)] 버튼을 클릭합니다.



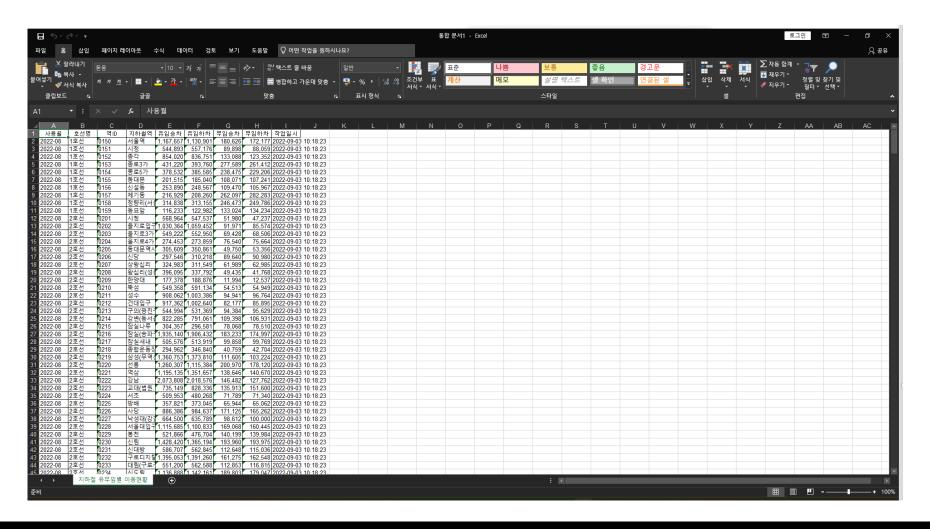


- ❖ ② 지하철 유무임별 이용현황 데이터 정제하기 (3/6)
 - [대상 통합 문서(T)] [(새 통합 문서)] [확인] 버튼을 클릭합니다.



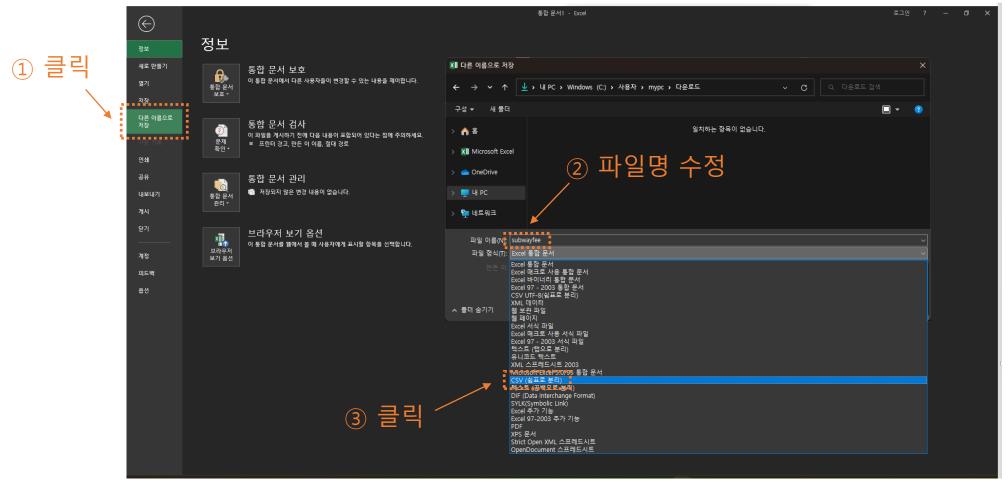


- ❖ ② 지하철 유무임별 이용현황 데이터 정제하기 (4/6)
 - 새 통합 문서가 열립니다.



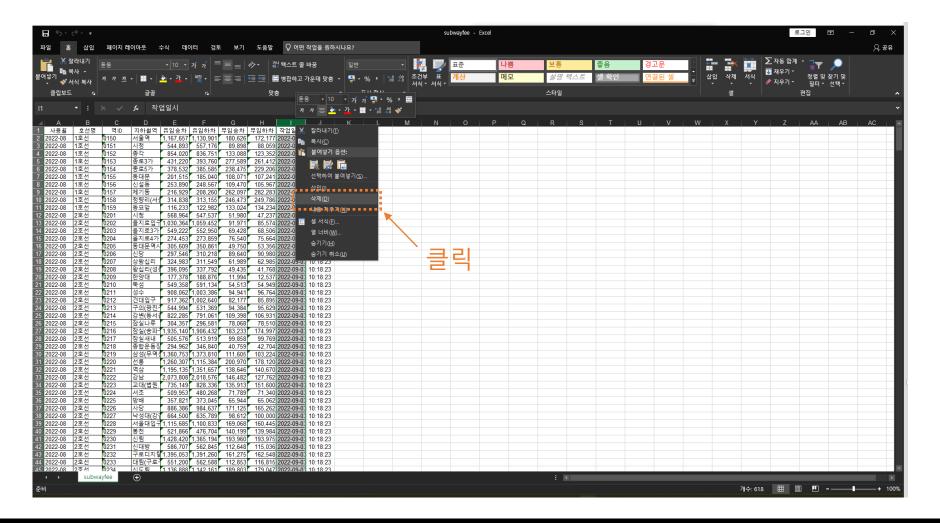


- ❖ ② 지하철 유무임별 이용현황 데이터 정제하기 (5/6)
 - [파일] [다른 이름으로 저장] 버튼을 클릭하고, 파일 이름을 subwayfee로 수정합니다.
 - 파일 형식을 "CSV (쉼표로 분리)"로 선택한 후 저장 버튼을 클릭합니다. (저장 경로는 다운로드 폴더)





- ❖ ② 지하철 유무임별 이용현황 데이터 정제하기 (6/6)
 - subwayfee.csv 파일을 열고, I열(Column)을 삭제 및 저장하고 파일을 닫습니다.





❖ ③ 생각해 보기 (1/2)

● 유임 승차 비율이 가장 높은 역은 어디일까요?

알고리즘(Algorithm)으로 생각하기

- ✓ Step 1) 데이터를 읽어온다.
- ✓ Step 2) 모든 역의 데이터를 바탕으로 각 역의 유임 승차 비율(Rate)을 계산한다.
- ✓ Step 3) 비율이 가장 높은 역을 찾는다.
- ✓ Step 4) 비율이 가장 높은 역이 어디인지, 그 비율이 얼마인지 출력한다.

유임(有賃): 값을 치름

■ 무임(無賃): 값을 치르지 않음



❖ ③ 생각해 보기 (2/2)

- 이번에는 유무임 승하차 인원이 가장 많은 역이 어디인지 찾아보겠습니다.
 - ◆ 유임 승차 인원이 가장 많은 역은?
 - ◆유임 하차 인원이 가장 많은 역은?
 - ◆ 무임 승차 인원이 가장 많은 역은?
 - ◆ 무임 하차 인원이 가장 많은 역은?

알고리즘(Algorithm)으로 생각하기

- ✓ Step 1) 데이터를 읽어온다.
- ✓ Step 2) 모든 역의 데이터를 바탕으로 유임 승차, 유임 하차, 무임 승차, 무임 하차 인원이가장 많은 역을 각각 찾는다.
- ✓ Step 3) 각각의 인원이 가장 많은 역을 출력한다.

■ 유임(有賃): 값을 치름

■ 무임(無賃): 값을 치르지 않음



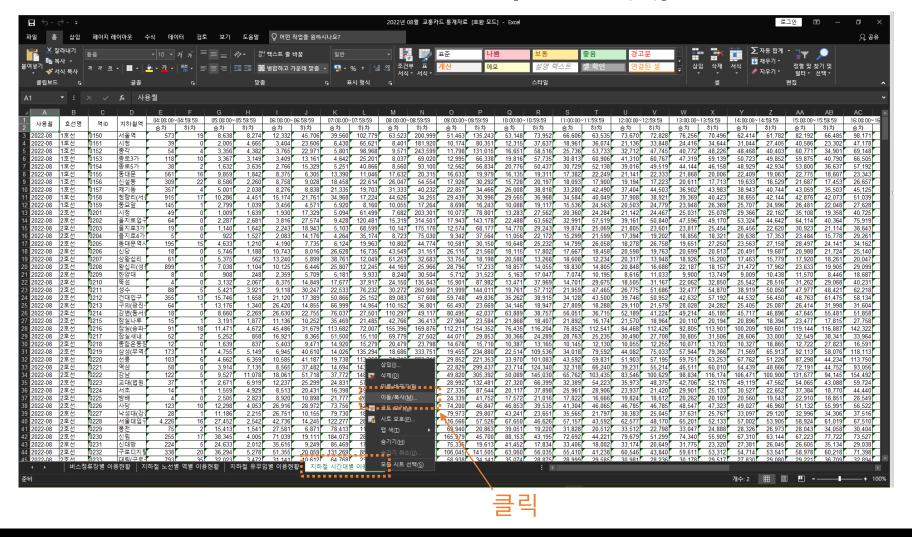
- ❖ ① 지하철 시간대별 이용 현황 데이터 정제하기 (1/6)
 - 4개의 탭 중 "지하철 시간대별 이용현황" 탭을 선택합니다.







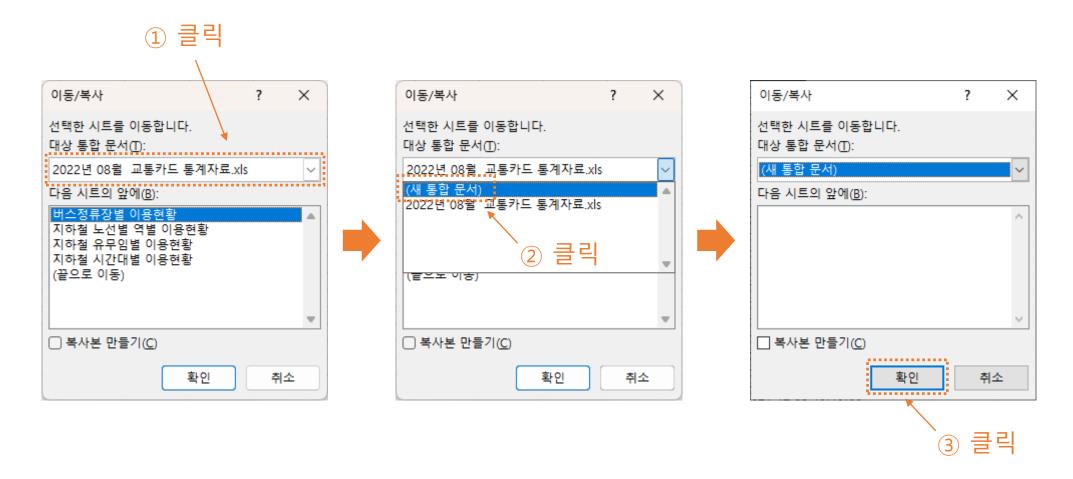
- ❖ ① 지하철 시간대별 이용 현황 데이터 정제하기 (2/6)
 - 워크시트 이름 탭을 마우스 오른쪽 단추로 클릭하고, [이동/복사(M)] 버튼을 클릭합니다.





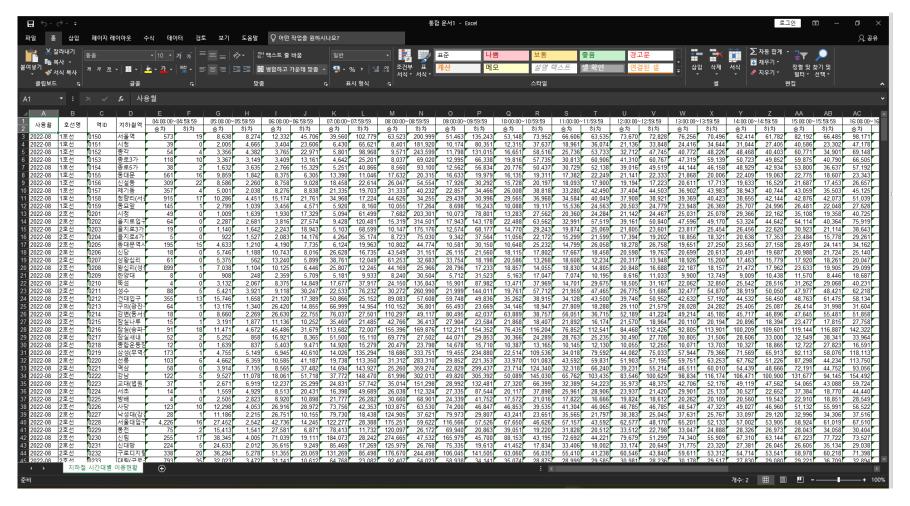


- ❖ ① 지하철 시간대별 이용 현황 데이터 정제하기 (3/6)
 - [대상 통합 문서(T)] [(새 통합 문서)] [확인] 버튼을 클릭합니다.



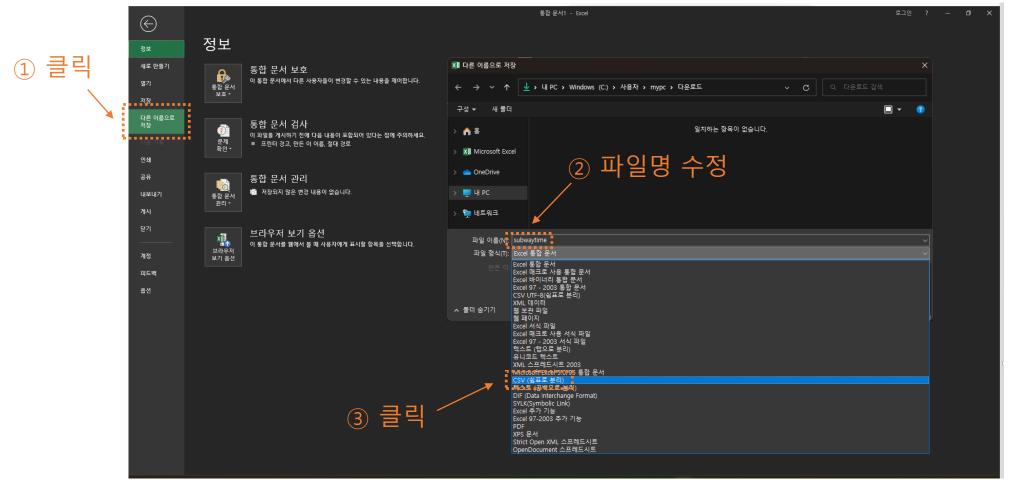


- ❖ ① 지하철 시간대별 이용 현황 데이터 정제하기 (4/6)
 - 새 통합 문서가 열립니다.



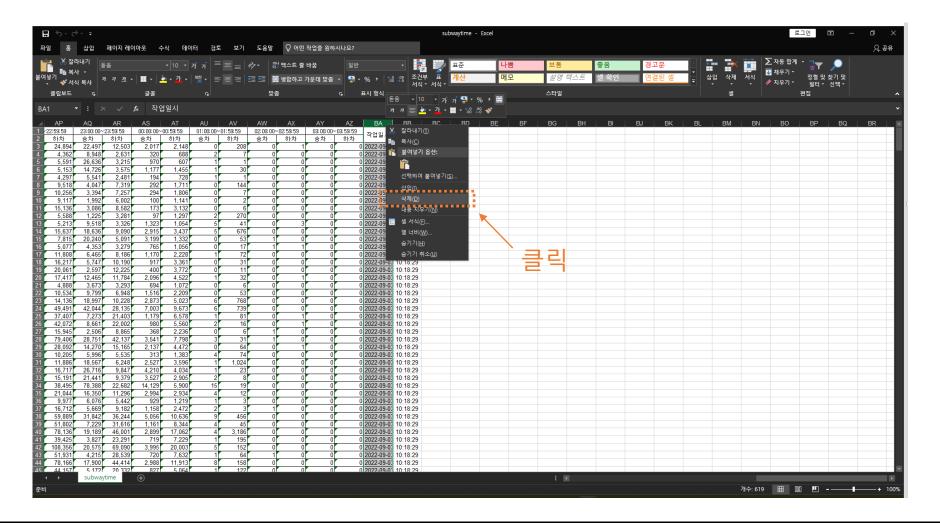


- ❖ ① 지하철 시간대별 이용 현황 데이터 정제하기 (5/6)
 - [파일] [다른 이름으로 저장] 버튼을 클릭하고, 파일 이름을 subwaytime으로 수정합니다.
 - 파일 형식을 "CSV (쉼표로 분리)"로 선택한 후 저장 버튼을 클릭합니다. (저장 경로는 다운로드 폴더)





- ❖ ① 지하철 시간대별 이용 현황 데이터 정제하기 (6/6)
 - subwaytime.csv 파일을 열고, BA열(Column)을 삭제 및 저장하고 파일을 닫습니다.





❖ ② 생각해 보기 (1/4)

● 출근 시간대 사람들이 가장 많이 타고 내리는 역은 어디일까요?

오전 7시 승차 데이터의 위치를 확인하니 10번 인덱스(Index)에 저장되어 있습니다.

Index 0 1 2 3 4 5 6 7 8 9 10 11 12 ...

4	Α	В	С	D	E	F	G	н	1	J	К	L	М	N
1	사용월	호선명	역ID	지하철역	04:00:00~0)4:59:59	05:00:00~0	05:59:59	06:00:00~0	06:59:59	07:00:00~0	7:59:59	08:00:00~0	8:59:59
2					승차	하차	승차	하차	승차	하차	승차	하차	승차	하차 :
3	Aug-22	1호선	150	서울역	573	19	8,638	8,274	12,332	45,706	39,560	102,779	63,523	200,999
4	Aug-22	1호선	151	시청	39	0	2,005	4,665	3,404	23,606	6,430	65,621	8,401	181,920
5	Aug-22	1호선	152	종각	54	4	3,356	4,382	3,765	22,971	5,801	98,968	9,571	243,599
6	Aug-22	1호선	153	종로3가	118	10	3,367	3,149	3,409	13,161	4,642	25,201	8,037	69,020
7	Aug-22	1호선	154	종로5가	38	2	1,632	3,635	2,766	15,329	5,251	40,866	8,560	93,100
8	Aug-22	1호선	155	동대문	561	16	9,859	1,842	8,375	6,305	13,390	11,046	17,632	20,315
9	Aug-22	1호선	156	신설동	309	22	8,586	2,260	8,758	9,028	18,458	22,614	26,047	54,554
10	Aug-22	1호선	157	제기동	357	4	5,001	2,038	8,276	8,838	21,335	19,703	31,333	40,232
11	Aug-22	1호선	158	청량리(서·	915	17	10,286	4,451	15,174	21,761	34,968	17,224	44,626	34,255
12	Aug-22	1호선	159	동묘앞	145	1	2,799	1,039	3,456	4,571	5,920	8,160	10,055	17,264
13	Aug-22	2호선	201	시청	49	0	1,009	1,639	1,930	17,329	5,094	61,499	7,682	203,301
14	Aug-22	2호선	202	을지로입	64	0	2,287	2,681	3,816	27,574	9,428	120,481	15,319	314,501
15	Aug-22	2호선	203	을지로3가	19	0	1,140	1,642	2,243	18,943	5,103	68,599	10,147	175,176
16	Aug-22	2호선	204	을지로4가	5	0	922	1,527	2,083	14,176	4,264	35,174	8,723	75,030
17	Aug-22	2호선	205	동대문역시	195	15	4,633	1,210	4,190	7,735	6,124	19,963	10,802	44,774
18	Διια-22	2호선	206	시단	18	n	5 746	1 122	10 743	2 N16	26.628	16 735	4 3 5 4 9	21 151



- ❖ ② 생각해 보기 (2/4)
 - 밤 11시에 사람들이 가장 많이 타는 역은 어디일까요?

승차 시간 t와 인덱스 i 사이의 관계식 찾기

Index	0	1	2	3	4	5	6	7	8	9	10	11	12	• • •	
	А	В	С	D	E	F	G	Н	1	J	К	L	М	N	
1	사용월	호선명	역ID	지하철역	04:00:00~()4:59:59	05:00:00~0	5:59:59	06:00:00~0	6:59:59	07:00:00~0	7:59:59	08:00:00~(8:59:59	C
2					승차	하차	승차	하차	승차	하차	승차	하차	승차	하차	ŧ
3	Aug-22	1호선	150	서울역	573	19	8,638	8,274	12,332	45,706	39,560	102,779	63,523	200,999	
4	Aug-22	1호선	151	시청	39	0	2,005	4,665	3,404	23,606	6,430	65,621	8,401	181,920	
5	Aug-22	1호선	152	종각	54	4	3,356	4,382	3,765	22,971	5,801	98,968	9,571	243,599	

승차 시간 t	인덱스 i	관계식
4	4	
5	6	
6	8	
7	10	i = 4 + (t - 4) * 2 = 2t - 4
8	12	
•••	•••	_
23	?	



- ❖ ② 생각해 보기 (3/4)
 - 시간대별로 사람들이 가장 많이 타고 내리는 역은 어디일까요?

시간대별 최대 승차 역 이름 및 승차 인원 출력하기

- ✓ 새벽 4시~새벽 3시 → 24시간을 1시간 단위로 구분 → 24개 → for 반복문 사용
- ✓ 변수 j와 인덱스 i 사이의 관계식 찾기

naex	Ü	I	2	3	4	5	6	/	8	9	10	11	12	• • •	
	A	В	С	D	E	F	G	н	I	J	K	L	М	N	
1	사용월	호선명	역ID	지하철역	04:00:00~0	4:59:59	05:00:00~()5:59:59	06:00:00~(6:59:59	07:00:00~0	7:59:59	08:00:00~(8:59:59	C
2					승차	하차	승차	하차	승차	하차	승차	하차	승차	하차	É
3	Aug-22	1호선	150	서울역	573	19	8,638	8,274	12,332	45,706	39,560	102,779	63,523	200,999	

변수 j	인덱스 i	관계식
0	4	
1	6	
2	8	: 2 * : . 4
•••	•••	i = 2 * j + 4
22	48	
23	50	



- ❖ ② 생각해 보기 (4/4)
 - 모든 지하철역에서 시간대별 승하차 인원의 총합은 어떻게 될까요?

알고리즘(Algorithm) 설계하기

- ✓ Step 1) 데이터를 읽어온다.
- ✓ Step 2) 모든 역에 대해 시간대별 승차 인원과 하차 인원을 누적해서 더한다.
- ✓ Step 3) 시간대별 승차 인원과 하차 인원을 그래프로 표현한다.

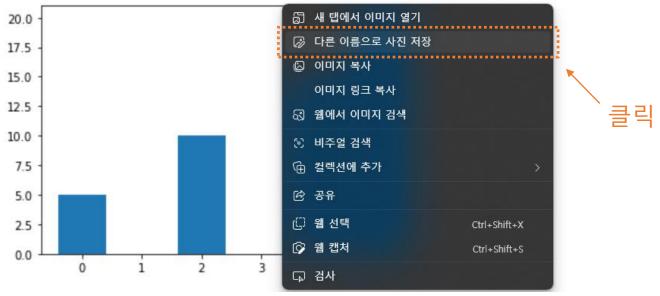


- ❖ 그래프를 이미지로 저장하는 방법 #1 (1/2)
 - 그래프 위에서 마우스 오른쪽 버튼 클릭 → [다른 이름으로 사진 저장] 클릭

```
import matplotlib.pyplot as plt

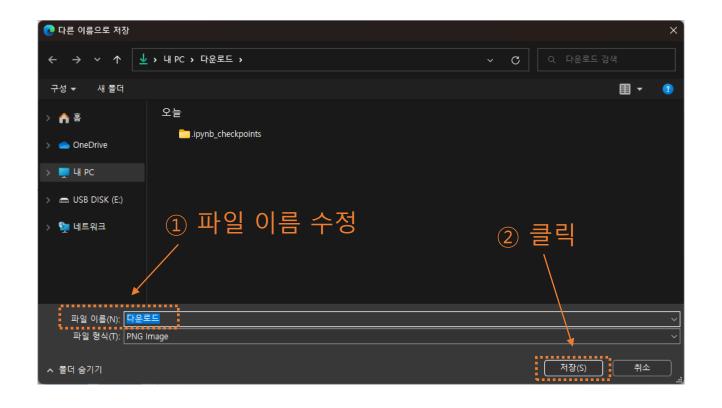
x = [0, 2, 4, 5]
y = [5, 10, 15, 20]

plt.bar(x, y)
plt.show()
```





- ❖ 그래프를 이미지로 저장하는 방법 #1 (2/2)
 - 파일 이름 수정 후 저장 버튼 클릭



파일 형식은 PNG Image가 기본입니다. PNG(Portable Network Graphics)는 비손실 그래픽 파일 포맷의 하나입니다.



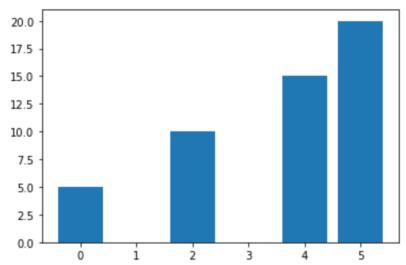
❖ 그래프를 이미지로 저장하는 방법 #2 (1/4)

● matplotlib 라이브러리의 pyplot 모듈에서 제공하는 savefig 함수를 이용

```
import matplotlib.pyplot as plt

x = [0, 2, 4, 5]
y = [5, 10, 15, 20]

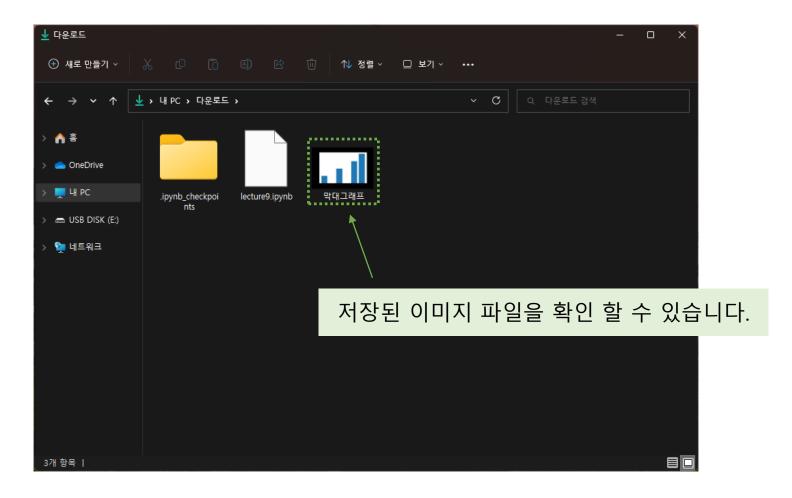
plt.bar(x, y)
plt.savefig('막대그래프.png')
plt.show()
```



savefig 함수의 입력 값으로는 저장할 파일의 이름을 적어주면 됩니다.



- ❖ 그래프를 이미지로 저장하는 방법 #2 (2/4)
 - 셀(Cell) 실행 후, 다운로드 폴더에 저장된 이미지 파일 확인하기





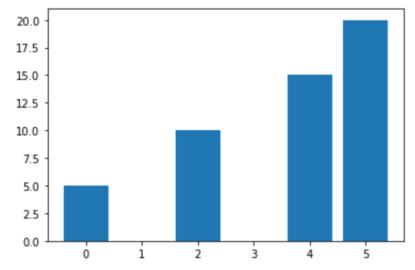
❖ 그래프를 이미지로 저장하는 방법 #2 (3/4)

● dpi를 설정하여 이미지를 원하는 해상도로 저장하기

```
import matplotlib.pyplot as plt

x = [0, 2, 4, 5]
y = [5, 10, 15, 20]

plt.bar(x, y)
plt.savefig('막대그래프_기본.png')
plt.savefig('막대그래프_dpi_50.png', dpi=50)
plt.savefig('막대그래프_dpi_200.png', dpi=200)
plt.show()
```



DPI(Dots per inch)는 디스플레이 해상도의 측정 단위이며, 1 제곱 인치 공간 안에 배치되는 픽셀(Pixel)의 수를 의미합니다.

DPI 값을 크게 지정하면 이미지가 선명해지는 대신 이미지 파일 크기가 커지게 됩니다.



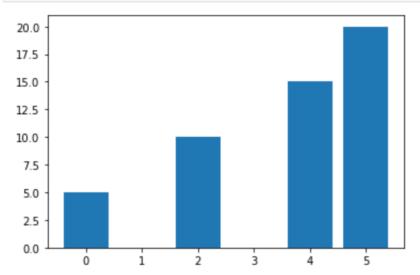
❖ 그래프를 이미지로 저장하는 방법 #2 (4/4)

● facecolor를 설정하여 저장할 이미지의 배경색 설정하기

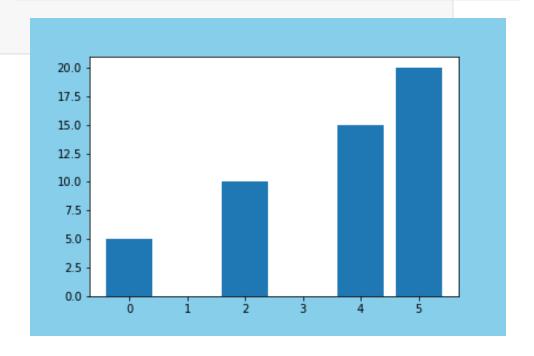
```
import matplotlib.pyplot as plt

x = [0, 2, 4, 5]
y = [5, 10, 15, 20]

plt.bar(x, y)
plt.savefig('막대그래프_facecolor.png', facecolor='skyblue')
plt.show()
```



지정된 색상으로 이미지의 배경색이 설정됩니다.



끝맺음



❖ 01. 대중교통 데이터 준비하기

❖ 02. savefig 함수로 그래프를 이미지로 저장하기



THANK YOU! Q & A

■ Name: 강환수

■ Office: 동양미래대학교 2호관 706호 (02-2610-1941)

■ E-mail: <u>hsknag@dongyang.ac.kr</u>

Homepage: https://github.com/ai7dnn/2023-DA