BigData

실습은 소스 코드를 참고 바랍니다.

<CSV 파일>
CSV (Comma-Separated Values) 파일은 텍스트 파일 형식 중 하나로, 데이터를 쉼표(콤마) 또는 다른 구분자를 사용하여 각 열의 데이터를 분리하는 형식입니다.

CSV를 활용해서 DB에 넣어서 DB를 시각화 할 수 있음.

<CSV로 생성할 표 데이터 예제>
아이디,이름,인원,주소,국번,전화 번호,평균 키,데뷔 일자
TWC,트와이스,9,서울,02,11111111,167,2015.10.19
BLK,블랙핑크,4,경남,055,22222222,163,2016.08.08
WMN,여자친구,6,경기,031,33333333,166,2015.01.15
OMY,오마이걸,7,서울,,,160,2015.04.21
GRL,소녀시대,8,서울,02,44444444,168,2007.08.02
ITZ,있지,5,경남,,,167,2019.02.12
RED,레드벨벳,4,경북,054,55555555,161,2014.08.01
APN,에이핑크,6,경기,031,77777777,164,2011.02.10
SPC,우주소녀,13,서울,02,88888888,162,2016.02.25
MMU,마마무,4,전남,061,99999999,165,2014.06.19
이와 같이 사용을 함

대게 보통은 콤마로 구별하는데 스페이스(공백)로 구분해서 사용하기도 함.

◎ 정부에서 제공하는 공공데이터포털 파일 포맷이 CSV이다.

이와 같이 데이터가 읽어짐
Python 3.11.5 (tags/v3.11.5:cce6ba9, Aug 24 2023, 14:38:34) [MSC v.1936 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.

= RESTART: C:\Users\admin\Desktop\Code BigData\01.py
아이디,이름,인원,주소,국번,전화 번호,평균 키,데뷔 일자
TWC,트와이스,9,서울,02,11111111,167,2015.10.19.

```
with open("C:/Users/admin/Desktop/Python, BigData/source/CSV/singer1.csv", "r") as inFp :
    with open("C:/Users/admin/Desktop/Python, BigData/source/CSV/new_singer2.csv", "w") as outFp:
        header = inFp.readline()
        header = header.strip()
        header_list= header.split(',')
        idx1 = header_list.index('아이디')
        idx2 = header_list.index('이름')
        idx3 = header_list.index('평균 키')
        header_list = [header_list[idx1], header_list[idx2], header_list[idx3]]
        header_str = ','.join(map(str, header_list))
        outFp.write(header_str + '\n')
        for inStr in inFp:
```

```
            inStr = inStr.strip()
            row_list = inStr.split(',')
            if int(row_list[idx3]) >= 165 :
                row_list = [row_list[idx1], row_list[idx2], row_list[idx3]]
                row_str = ','.join(map(str, row_list))
                outFp.write(row_str + '\n')

print('Save. OK~')
```

아이디	이름	평균 키
TWC	트와이스	167
WMN	여자친구	166
GRL	소녀시대	168
ITZ	있지	167
MMU	마마무	165

와 같이 나오게 됨

```
<CSV Library>
import csv

# with를 쓰면 Close를 사용하지 않아도 된다.
with open("C:/Users/admin/Desktop/Python, BigData/source/CSV/singer2.csv", "r") as inFp :
    csvReader = csv.reader(inFp)
    header_list = next(csvReader)
    print(header_list[1],header_list[6])
    for row_list in csvReader:
        youtube = int(row_list[6].replace(',',''))
        youtube = int(youtube/10000)
        print(row_list[1], str(youtube)+"만")
```

[결과값]

```
이름 유튜브 조회수
트와이스 333만
블랙핑크 44만
여자친구 0만
오마이걸 0만
소녀시대 111만
있지 2만
레드벨벳 4만
에이핑크 0만
우주소녀 0만
마마무 0만
```

<Windows Teminal 사용>
pip, pip3 활용

MS Office 97~2003 Version으로 저장 - 공용 오픈 버전

```
import xlrd

workbook = xlrd.open_workbook('C:/Users/admin/Desktop/Python,
BigData/source/Excel/singer.xls')
sheetCount = workbook.nsheets
print('워크시트는 %d개 입니다' % (sheetCount))

wsheetList = workbook.sheets()
for worksheet in wsheetList :
    print('** 워크시트의 이름 : %s' % (worksheet.name) )
    print(" 행 수는 %d, 열 개수는 %d 입니다." % (worksheet.nrows, worksheet.ncols))
```

```
워크시트는 2개 입니다
** 워크시트의 이름 : senior
 행 수는 6, 열 개수는 7 입니다.
** 워크시트의 이름 : junior
 행 수는 6, 열 개수는 7입니다.

** 워크시트의 이름 : senior
아이디 이름   인원   주소   평균 키       데뷔 일자     유튜브 조회수
WMN  여자친구      6.0    경기   166.0  2015.01.15    800.0
GRL  소녀시대      8.0    서울   168.0  2007.08.02    1114600.0
RED  레드벨벳      4.0    경북   161.0  2014.08.01    44500.0
APN  에이핑크      6.0    경기   164.0  2011.02.10    2900.0
MMU  마마무 4.0    전남   165.0  2014.06.19    6900.0

** 워크시트의 이름 : junior
아이디 이름   인원   주소   평균 키       데뷔 일자     유튜브 조회수
TWC  트와이스      9.0    서울   167.0  2015.10.19    3334500.0
BLK  블랙핑크      4.0    경남   163.0  2016.08.08    443700.0
OMY  오마이걸      7.0    서울   160.0  2015.04.21    3500.0
ITZ    있지   5.0    경남   167.0  2019.02.12    21300.0
SPC  우주소녀      13.0   서울   162.0  2016.02.25    350.0
```

```
import xlrd

workbook = xlrd.open_workbook('C:/Users/admin/Desktop/Python,
BigData/source/Excel/singer.xls')
sheetCount = workbook.nsheets

personNum = 0
personIdx = 2
rowCount = 0
wsheetList = workbook.sheets()
for worksheet in wsheetList :
    rowCount += worksheet.nrows-1
    for row in range(1, worksheet.nrows) :
        personNum +=  int(worksheet.cell_value(row, personIdx))
```

```
print("전체 가수그룹 인원 합계 : ", personNum)
print("가수그룹 인원 평균 : ", personNum/rowCount)
전체 가수그룹 인원 합계 :  66
가수그룹 인원 평균 :  6.6


import xlrd
import xlwt

workbook = xlrd.open_workbook('C:/Users/admin/Desktop/Python,
BigData/source/Excel/singer.xls')
outWorkbook = xlwt.Workbook()

wsheetList = workbook.sheets()
for worksheet in wsheetList :
    outSheet = outWorkbook.add_sheet(worksheet.name)
    for row in range(worksheet.nrows) :
        for col in range(worksheet.ncols) :
            outSheet.write(row, col, worksheet.cell_value(row, col))

outWorkbook.save('C:/Users/admin/Desktop/Python, BigData/source/Excel/outSinger1.xls')
print("Save. OK~")

Save. OK~
```

아이디	이름	인원	주소	평균 키	데뷔 일자	유튜브 조회수
WMN	여자친구	6	경기	166	2015.01.15	800
GRL	소녀시대	8	서울	168	2007.08.02	1114600
RED	레드벨벳	4	경북	161	2014.08.01	44500
APN	에이핑크	6	경기	164	2011.02.10	2900
MMU	마마무	4	전남	165	2014.06.19	6900

```
import xlrd
import xlwt

workbook = xlrd.open_workbook('C:/Users/admin/Desktop/Python,
BigData/source/Excel/singer.xls')
outWorkbook = xlwt.Workbook()
idx = 4 # 평균 키의 인덱스

wsheetList = workbook.sheets()
outSheet = outWorkbook.add_sheet("singer")
worksheet = wsheetList[0]
for col in range(worksheet.ncols):
    outSheet.write(0, col, worksheet.cell_value(0, col))

totalRow = 0
for worksheet in wsheetList :
    for row in range(1, worksheet.nrows) :
        height = worksheet.cell_value(row, idx)
        if int(height) >= 165 :
```

```
            totalRow += 1
            for col in range(worksheet.ncols) :
                outSheet.write(totalRow, col, worksheet.cell_value(row, col))

outWorkbook.save('C:/Users/admin/Desktop/Python, BigData/source/Excel/outSinger2.xls')
print("Save. OK~")
```

아이디	이름	인원	주소	평균 키	데뷔 일자	유튜브 조회수
WMN	여자친구	6	경기	166	2015.01.15	800
GRL	소녀시대	8	서울	168	2007.08.02	1114600
MMU	마마무	4	전남	165	2014.06.19	6900
TWC	트와이스	9	서울	167	2015.10.19	3334500
ITZ	있지	5	경남	167	2019.02.12	21300

<openpyxl Library>
MS Office 2007 Excel 이후 버전의 라이브러리 xlsx의 확장자

엑셀 이미지 출력