21. Pandas – DataFrame - Concatenate Multiple csv files

Contents

21. Pandas – DataFrame - Concatenate Multiple csv files

1. Real time requirement

- ✓ Generally total data will be stored with multiple files
  - Yearly data: Jan sales, Feb sales, ..., Dec sales
- ✓ So, we need to concatenate all these monthly sales to bring year sales right

2. Loading csv files from all files

- ✓ The very first step is we need to access all files from specific folder.
- ✓ From that folder we need to capture only csv files.

3. os module

- ✓ os is a predefined module in python.
- ✓ By using this module we can load all files from the folder.

4. listdir(p) function

- ✓ listdir(p) is a predefined function in os module
- ✓ This function we should access with os module name.
- ✓ By using this function we can get all file names from folder.
- ✓ This function returns all file names in list.

Program Name: Accessing all files from the folder
demo1.py
Input file: **daniel**/jan_sales.csv,....,dec_sales.csv [15 files]

```
import os

path = "./daniel"

all_files = os.listdir(path)

print(all_files)
```

Output

```
['apr_sales.csv', 'aug_sales.csv', 'dec_sales.csv', 'feb_sales.csv',
'fun.jpg', 'jan_sales.csv', 'jul_sales.csv', 'jun_sales.csv',
'mar_sales.csv', 'may_sales.csv', 'nov_sales.csv', 'oct_sales.csv',
'progress.txt', 'sep_sales.csv', 'status.xlsx']
```

5. filter(p1, p2) function

- ✓ filter(p1, p2) is a predefined function in python
- ✓ We can access this function directly.
- ✓ By using this function we can apply Boolean logic and get results accordingly.

Program Name: Accessing only csv files from folder
demo2.py
Input file: **daniel**/jan_sales.csv,....,dec_sales.csv

```
import os

path = "./daniel"

all_files = os.listdir(path)

f = filter(lambda name: name.endswith('.csv'), all_files)
csv_files = list(f)

print(all_files)
print()
print(csv_files)
```

Output

```
['apr_sales.csv', 'aug_sales.csv', 'dec_sales.csv', 'feb_sales.csv',
'fun.jpg', 'jan_sales.csv', 'jul_sales.csv', 'jun_sales.csv',
'mar_sales.csv', 'may_sales.csv', 'nov_sales.csv', 'oct_sales.csv',
'progress.txt', 'sep_sales.csv', 'status.xlsx']

['apr_sales.csv', 'aug_sales.csv', 'dec_sales.csv', 'feb_sales.csv',
'jan_sales.csv', 'jul_sales.csv', 'jun_sales.csv', 'mar_sales.csv',
'may_sales.csv', 'nov_sales.csv', 'oct_sales.csv', 'sep_sales.csv']
```

6. Concatenating all csv file

- ✓ Once we loaded all csv file then we can concatenate all csv file.
- ✓ Based on requirement by using pandas we can concatenate all csv files into one csv file

Program Name: Concatenating all csv files
Input file: demo3.py
daniel/jan_sales.csv,....,dec_sales.csv [12 files]

```
import os
import glob
import pandas as pd

p = '.\daniel'

files = os.path.join(p, "*.csv")
csv_files = glob.glob(files)

result = (pd.read_csv(every)      for every in csv_files)

df = pd.concat(result, ignore_index = True)

print(df)
df.to_csv("year.csv", index = False)
```

Output

```
     Order_Id Customer_Name  Customer_Id               Product_Name  Product_Cost            Date
0        1320          Venu           23    LG ThinQ Refrigerator         55000   4/11/2019 9:00
1        1321  Jaya Chandra           21  Bose SoundSport Headphones      63000  4/13/2019 10:00
2        1322   Mallikarjun           13   Apple Airpods Headphones       69999  4/13/2019 11:00
3        1323        Siddhu           18     Samsung Galaxy S9 Plus       55000  4/14/2019 12:00
4        1324        Daniel            6                    iPhone 8      55000  4/15/2019 13:00
..        ...           ...          ...                        ...        ...             ...
103      1819        Daniel            6         LG Washing Machine       50000    9/3/2019 4:00
104      1820       Neelima           19               20in Monitor       55000    9/3/2019 5:00
105      1821       Karteek            4               20in Monitor       50000    9/3/2019 6:00
106      1822  Jaya Chandra           21    LG ThinQ Refrigerator         75999    9/3/2019 7:00
107      1823    Chaithanya           14           27in FHD Monitor       55000    9/3/2019 8:00

[108 rows x 6 columns]
```

Program Name: Loading year.csv files
Input file: demo4.py
year.csv

```
import pandas as pd

df = pd.read_csv("year.csv", parse_dates = ["Date"])

print(df)
```

Output

```
   Order_Id Customer_Name  Customer_Id               Product_Name  Product_Cost                Date
0      1320          Venu           23     LG ThinQ Refrigerator         55000 2019-04-11 09:00:00
1      1321  Jaya Chandra           21  Bose SoundSport Headphones         63000 2019-04-13 10:00:00
2      1322   Mallikarjun           13   Apple Airpods Headphones         69999 2019-04-13 11:00:00
3      1323        Siddhu           18     Samsung Galaxy S9 Plus         55000 2019-04-14 12:00:00
4      1324        Daniel            6                   iPhone 8         55000 2019-04-15 13:00:00
```