Task : 5 Author:- Sagar Bhoi

1. Write a python program for merge two excel file data to one file

In [1]:
```
import pandas as pd
```

In [2]:
```
#loading datasets
df_1 = pd.read_excel("Kalyani_Balance_Sheet_November_2022.xlsx")
df_2 = pd.read_excel("Kalyani_Balance_Sheet_December_2022.xlsx")

#merge datasets
df_combine = pd.concat([df_1, df_2])
```

In [3]:
```
#df_1
#df_2
df_combine
```

Out[3]:

	Name	Branch Team member	Department	Fees	Chennai	Aurangabad	Unnamed: 6	Student Count	Chennai.1	Aurangabad.1	Total	Month
0	Thirumurugan	Kalyani	Data Science	499	100	399	NaN	103	15092	36305.0	51397.0	NaT
1	Harsh Dodiya	Kalyani	Data Science	499	100	399	NaN	Month	2022-11-01 00:00:00	NaN	NaN	NaT
2	Kunal pahuja	Kalyani	Data Science	499	200	299	NaN	NaN	NaN	NaN	NaN	NaT
3	Arya Singh	Kalyani	Data Science	499	100	399	NaN	NaN	NaN	NaN	NaN	NaT
4	Prashansa Shree	Kalyani	Data Science	499	100	399	NaN	NaN	NaN	NaN	NaN	NaT
...	...	...	...	...	...	...	...	...	...	...	...	...
118	Sujeet Singh Rajpoot-k	Kalyani	Data Science	499	100	399	NaN	NaN	NaN	NaN	NaN	NaT
119	Kaligatla Sree Samhitha-K	Kalyani	Data Science	499	100	399	NaN	NaN	NaN	NaN	NaN	NaT
120	Manuru Sai Suhas-K	Kalyani	Data Science	499	100	399	NaN	NaN	NaN	NaN	NaN	NaT
121	Harsh Prasad-K	Kalyani	Data Science	499	100	399	NaN	NaN	NaN	NaN	NaN	NaT
122	Princy Gurnani-K	Kalyani	Data Science	499	100	399	NaN	NaN	NaN	NaN	NaN	NaT

226 rows × 12 columns

In [4]:
```
df_combine = df_combine.iloc[:, 0:6]
df_combine
```

Out[4]:

	Name	Branch Team member	Department	Fees	Chennai	Aurangabad
0	Thirumurugan	Kalyani	Data Science	499	100	399
1	Harsh Dodiya	Kalyani	Data Science	499	100	399
2	Kunal pahuja	Kalyani	Data Science	499	200	299
3	Arya Singh	Kalyani	Data Science	499	100	399
4	Prashansa Shree	Kalyani	Data Science	499	100	399
...	...	...	...	...	...	...
118	Sujeet Singh Rajpoot-k	Kalyani	Data Science	499	100	399
119	Kaligatla Sree Samhitha-K	Kalyani	Data Science	499	100	399
120	Manuru Sai Suhas-K	Kalyani	Data Science	499	100	399
121	Harsh Prasad-K	Kalyani	Data Science	499	100	399
122	Princy Gurnani-K	Kalyani	Data Science	499	100	399

226 rows × 6 columns

In [5]:
```
df_combine.to_excel('YoshopsBalanceSheet_NovDec_2022.xlsx', index = False)
```

In [6]:
```
new_df = pd.read_excel('YoshopsBalanceSheet_NovDec_2022.xlsx')
new_df
```

Out[6]:

	Name	Branch Team member	Department	Fees	Chennai	Aurangabad
0	Thirumurugan	Kalyani	Data Science	499	100	399
1	Harsh Dodiya	Kalyani	Data Science	499	100	399
2	Kunal pahuja	Kalyani	Data Science	499	200	299
3	Arya Singh	Kalyani	Data Science	499	100	399
4	Prashansa Shree	Kalyani	Data Science	499	100	399
...	...	...	...	...	...	...
221	Sujeet Singh Rajpoot-k	Kalyani	Data Science	499	100	399
222	Kaligatla Sree Samhitha-K	Kalyani	Data Science	499	100	399
223	Manuru Sai Suhas-K	Kalyani	Data Science	499	100	399
224	Harsh Prasad-K	Kalyani	Data Science	499	100	399
225	Princy Gurnani-K	Kalyani	Data Science	499	100	399

226 rows × 6 columns

In [7]:
```
#import os
#import shutil
```

2. Write a python program to shorting file in different folder means main folder containing 40 word file. Now after shorting create 4 child folder and store 10 file each folder.

In [8]:
```
import os
import shutil

target_folder = r'/Users/Sagar/Yoshops Data Science Intern/Task_5_week_5/Task 5'+'\\'
source_folder = r'/Users/Sagar/Yoshops Data Science Intern/Task_5_week_5'+'\\'

for path,dir,files in os.walk(source_folder):
    print(path)
    print(files)
```

```
/Users/Sagar/Yoshops Data Science Intern/Task_5_week_5\
['Kalyani_Balance_Sheet_December_2022.xlsx', 'Kalyani_Balance_Sheet_November_2022.xlsx', 'YoshopsBalanceSheet_NovDec_2022.xlsx', 'Yoshops_Task_5_Sagar_Bhoi.ipynb']
/Users/Sagar/Yoshops Data Science Intern/Task_5_week_5\.ipynb_checkpoints
['Yoshops_Task_5_Sagar_Bhoi-checkpoint.ipynb']
/Users/Sagar/Yoshops Data Science Intern/Task_5_week_5\Task 5
[]
```

In [9]:
```
for path,dir,files in os.walk(source_folder):
    if files:
        for file in files:
            if not os.path.isfile(target_folder + file):
                os.rename(path + '\\' + file, target_folder + file)
```

In [10]:
```
for path,dir,files in os.walk(target_folder):
    print(path)
    print(files)
```

```
/Users/Sagar/Yoshops Data Science Intern/Task_5_week_5/Task 5\
['Kalyani_Balance_Sheet_December_2022.xlsx', 'Kalyani_Balance_Sheet_November_2022.xlsx', 'YoshopsBalanceSheet_NovDec_2022.xlsx', 'Yoshops_Task_5_Sagar_Bhoi-checkpoint.ipynb', 'Yoshops_Task_5_Sagar_Bhoi.ipynb']
```

In [11]:
```
import os
dir_name = '/Users/Sagar/Yoshops Data Science Intern/Task_5_week_5/Task 5'
# Get list of all files in a given directory sorted by name
list_of_files = sorted( filter( lambda x: os.path.isfile(os.path.join(dir_name, x)),
                        os.listdir(dir_name) ) )
for file_name in list_of_files:
    print(file_name)
```

```
Kalyani_Balance_Sheet_December_2022.xlsx
Kalyani_Balance_Sheet_November_2022.xlsx
YoshopsBalanceSheet_NovDec_2022.xlsx
Yoshops_Task_5_Sagar_Bhoi-checkpoint.ipynb
Yoshops_Task_5_Sagar_Bhoi.ipynb
```

In [12]:
```
import os
path = '/Users/Sagar/Yoshops Data Science Intern/Task_5_week_5/Task 5'
for i in range(0,4):
    os.chdir(path)
    Newfolders = 'SubFolder'+str(i)
    os.makedirs(Newfolders)
```

In [13]:
```
i=0
j=0
k=0
```

```
for i in range(0,4):
    for j in range(0,10):
        src = '/Users/Sagar/Yoshops Data Science Intern/Task_5_week_5/Task 5/'+list_of_files[k]
        dst = '/Users/Sagar/Yoshops Data Science Intern/Task_5_week_5'+str(i)
        shutil.copy2(src,dst)
        k=k+1
```

```
---------------------------------------------------------------------------
IndexError                                Traceback (most recent call last)
Cell In[13], line 6
      4 for i in range(0,4):
      5     for j in range(0,10):
----> 6         src = '/Users/Sagar/Yoshops Data Science Intern/Task_5_week_5/Task 5/'+list_of_files[k]
      7         dst = '/Users/Sagar/Yoshops Data Science Intern/Task_5_week_5'+str(i)
      8         shutil.copy2(src,dst)

IndexError: list index out of range
```

3. Write a python programm separate duplicate file

In [ ]:
```
# Importing Libraries
import os
from pathlib import Path
from filecmp import cmp


# list of all documents
DATA_DIR = Path('/Users/Sagar/Yoshops Data Science Intern/Task_5_week_5/Task 5/')
files = sorted(os.listdir(DATA_DIR))

# List having the classes of documents
# with the same content
duplicateFiles = []

# comparison of the documents
for file_x in files:

    if_dupl = False

    for class_ in duplicateFiles:
        # Comparing files having same content using cmp()
        # class_[0] represents a class having same content
        if_dupl = cmp(
            DATA_DIR / file_x,
            DATA_DIR / class_[0],
```

```
            shallow=False
        )
        if if_dupl:
            class_.append(file_x)
            break

    if not if_dupl:
        duplicateFiles.append([file_x])

# Print results
print(duplicateFiles)
```

Check excel file and create sperate file and store duplicate data

In [ ]:
```
import pandas as pd
df_master = pd.read_excel('/Users/Sagar/Yoshops Data Science Intern/Task_5_week_5/Task 5/Kalyani_Balance_Sheet_December_2022.xlsx')
print(df_master)
```

In [ ]:
```
# Selecting duplicate rows except first
# occurrence based on all columns
duplicate = df_master[df_master.duplicated()]

print("Duplicate Rows :")

# Print the resultant Dataframe
duplicate
```

Sort files in a folder based on their size

In [ ]:
```
import os
import shutil
# The folder containing files.
directory = '/Users/Sagar/Yoshops Data Science Intern/Task_5_week_5/Task 5'

# Get all files.
list = os.listdir(directory)

# Loop and add files to list.
pairs = []
for file in list:

    # Use join to get full file path.
    location = os.path.join(directory, file)
```

```
    # Get size and add to list of tuples.
    size = os.path.getsize(location)
    pairs.append((size,file))

# Sort list of tuples by the first element, size.
pairs.sort(key=lambda s: s[0])

i=0
# Display pairs.
for pair in pairs:
    #src = 'F:/Yoshops/task_5/task_5/source/'+pair[1]
    #dst = 'F:/Yoshops/task_5/task_5/destination/size'+str(i)
    #shutil.copy2(src,dst)

    print(pair)
    i=i+1

path = r'/Users/Sagar/Yoshops Data Science Intern/Task_5_week_5/Task 5'
for j in range(0,i):
    os.chdir(path)
    Newfolders = 'size'+str(j)
    os.makedirs(Newfolders)
```

In [ ]:
```
i=0
# Display pairs.
for pair in pairs:
    src = r'/Users/Sagar/Yoshops Data Science Intern/Task_5_week_5'+pair[1]
    dst = r'/Users/Sagar/Yoshops Data Science Intern/Task_5_week_5/Task 5'+str(i)
    shutil.copy2(src,dst)
    i=i+1
```

In [ ]:
```

```