```
#!/usr/bin/env python
# coding: utf-8
# Import required Python libraries
import requests
from bs4 import BeautifulSoup, re, Comment
import pandas as pd
import xlsxwriter

# Step 1: Scraping the primary "MEPS data file website",
 # finding the data file names that are within the "option"
 # comment tags, and saving them in a csv file

 def extractOptions(inputData):
     sub1 = str(re.escape('<option value="All">All data files</option>'))
     sub2 = str(re.escape('</select>'))
     result = re.findall(sub1+"(.*)"+sub2, inputData, flags=re.S)
     if len(result) > 0:
         return result[0]

 def extractData(inputData):
     sub1 = str(re.escape('>'))
     sub2 = str(re.escape('</option>'))
     result =  re.findall(sub1+"(.*)"+sub2, inputData, flags=re.S)
     if len(result) > 0:
         return result[0]
     return ''

 def main(base_url):
     response = requests.get(base_url)
     soup = BeautifulSoup(response.text, "html.parser")
     comments = soup.find_all(string=lambda text: isinstance(text, Comment))

     for c in comments:
         if '<select id="pufnumber" size=1 name="cboPufNumber">' in c:
             options = extractOptions(c)
             ops = options.splitlines() #split text into lines
             fp = open(r'C:/Data/MEPS_fn.csv', 'w')
             for op in ops:
                 data = extractData(op)
                 if data != '': #check if the data found
                     fp.write(data +'\n')
             fp.close()

             with open(r'C:/Data/MEPS_fn.csv', 'r') as buff:
                 for i, line in enumerate(buff, 1):
                     pass
                 print(f"{(i)}", 'file names listed in the MEPS website')

 main('https://meps.ahrq.gov/data_stats/download_data_files.jsp')

 # Step 2: Creating a Pandas DataFrame from the csv file 12/31/2022

 colname = ['file_name']
 df1 = pd.read_csv(r'C:/Data/MEPS_fn.csv',  sep='\t', names = colname)

 df1.drop(df1[df1['file_name'].str.contains('replaced|CD-ROM|NHC|NHEA|NHIS Link|HC-IC Linked|
1996 Parent IDs')].index, inplace=True)

 df1["file_id"] = df1["file_name"].str.extract(r"([A-Z])[A-Z]+-(\d+[A-
Z]*)").sum(axis=1).str.lower()
 df1['file_id'] = df1['file_id'].str.replace('h0', 'h').str.replace('h36', 'h036')
.str.replace('h36brr', 'h036brr')

 df1["url1"] =
"https://meps.ahrq.gov/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-" +
df1["file_name"].str.extract(r"(\d+[A-Z]*)").sum(axis=1).astype(str)

 df1.reset_index(drop = True, inplace = True)

 print("{:,}".format(len(df1)), 'MEPS public-use file names')
```

```
# Step 3: Scraping all MEPS data file-specific websites
# and saving format-spcific file names from each of
# those sites in a DataFrame 12/31/2022

url2_str_list = []
for item in df1.index:
    url1_str = df1['url1'][item]
    response = requests.get(url1_str)
    soup = BeautifulSoup(response.text, "html.parser")

    for link in soup.find_all('a'):
        if link.text.endswith('ZIP'):
            url2_str = 'https://meps.ahrq.gov' + link.get('href').strip('..')
            url2_str_list.append(url2_str)

df2  = pd.DataFrame(url2_str_list, columns=['url2'])
df2['file_id'] = df2['url2'].str.extract(r"([h]\d+[abcdefghir]*(?!\d))").sum(axis=1)
df2['file_id'] = df2['file_id'].str.replace('da', '')

df1 = df1.drop('url1',axis=1)
merged_df = pd.merge(df1, df2, on='file_id', validate ="one_to_many")
print("{:,}".format(len(merged_df)), 'URLs that are specific to data file formats')
with pd.ExcelWriter('merged_df.xlsx') as writer:
    merged_df.to_excel(writer, sheet_name='data_urls', index=False)
    writer.sheets['data_urls'].set_column(45, 3, 45)

```