Problem #1

Subject: Data Mining
Modules: re, pathlib
Filename: hw06_01.py

WORK

To achieve the task of data mining from a zipped file, the program begins by unzipping the provided file using the 'unzip_file' function. Following this, the number of subdirectories within the 'data' directory is counted using the 'count_subdirectories' function. Then, for each subdirectory, the program iterates through its files, counting the number of files in each subdirectory using the 'count_files_in_subdirectory' function. Within this iteration, the 'find_five_digit_numbers' function is used to locate and extract all five-digit numbers surrounded by dollar signs ('$') from each file. These numbers are accumulated in a list. Once all files have been processed, the total number of files and the total number of five-digit numbers found are printed. Finally, the sum of these five-digit numbers is calculated using the 'sum_five_digit_numbers' function, and the result is displayed.

OUTPUT

```
DATA MINING!!!

Number of subdirectories: 50

Number of files in data/rawiplzpi: 20
Number of files in data/soabjrhlol: 20
Number of files in data/kcqnhvlsx: 20
Number of files in data/ogrdqhgv: 20
Number of files in data/ewrdadpbfij: 20
Number of files in data/flwzhwqq: 20
Number of files in data/qzecubqtsg: 20
Number of files in data/qpcnzgxrntl: 20
Number of files in data/frhqlednfid: 20
Number of files in data/lqjtok: 20
Number of files in data/lrgtwkgnh: 20
Number of files in data/lldjdnefi: 20
Number of files in data/mjanmpa: 20
Number of files in data/kfwgsbupnnc: 20
Number of files in data/gulaksy: 20
Number of files in data/eukfmzthr: 20
Number of files in data/hbucfrcmpdtn: 20
Number of files in data/gaazbtxmzv: 20
Number of files in data/xddnovkveccz: 20
Number of files in data/rfhnwluk: 20
Number of files in data/yurfut: 20
Number of files in data/glugtlvqml: 20
Number of files in data/dpxrqorv: 20
Number of files in data/urivqszh: 20
Number of files in data/wolqauqeyr: 20
Number of files in data/malanmeydfse: 20
Number of files in data/qgjamvnqb: 20
Number of files in data/aywyboo: 20
```

```
Number of files in data/sctmwkqymfwr: 20
Number of files in data/rytkaanki: 20
Number of files in data/dsznbtqgg: 20
Number of files in data/hykpkfdqpnu: 20
Number of files in data/ddtuptgmm: 20
Number of files in data/cutshn: 20
Number of files in data/mqxmkemiolb: 20
Number of files in data/ryjztmtrdh: 20
Number of files in data/wzqyeac: 20
Number of files in data/ualhdniwsp: 20
Number of files in data/vireyzhfuoo: 20
Number of files in data/eersulmimg: 20
Number of files in data/tudquqly: 20
Number of files in data/zrfivzy: 20
Number of files in data/auiuyfsjj: 20
Number of files in data/zjfdypetesh: 20
Number of files in data/tteamqxpqns: 20
Number of files in data/cnyftz: 20
Number of files in data/bbncghlit: 20
Number of files in data/nibbhaohqibo: 20
Number of files in data/hcnkdfgktx: 20
Number of files in data/uyfjyeoixe: 20

Total number of files: 1000

Total number of five-digit numbers found: 4745

Sum of those numbers: 235762121
```

CODE

```
'''
PROGRAMMER: Jakob K. West
USERNAME: jwest21
PROGRAM: hw06_01.py

DESCRIPTION: DATA MINING!!!!!
Scan the contents of a zipped-up file to
find the five-digit numbers and sum them up
'''


import re
from pathlib import Path
import zipfile


# File name variables
zip_file_name = 'data.zip'
file_name = 'data'


# Function #1
def unzip_file(zip_file):

    # Unzip the file
    with zipfile.ZipFile(zip_file, 'r') as zip_ref:
        zip_ref.extractall()


# Function #2
def count_subdirectories(directory):

    # Count the number of subdirectories inside the given directory
    subdirectories = [subdir for subdir in Path(directory).iterdir() if
subdir.is_dir()]
    return len(subdirectories)


# Function #3
def count_files_in_subdirectory(subdirectory):

    # Count the number of files inside each subdirectory
    files = [file for file in Path(subdirectory).iterdir() if file.is_file()]
    return len(files)


# Function #4
def find_five_digit_numbers(file_path):

    # Find all five-digit numbers surrounded by dollar signs ($) in a file
    numbers = []
    with open(file_path, 'r') as file:
        for line in file:
```

```
            numbers.extend(re.findall(r'\$(\d{5})\$', line))

    return numbers


# Function #5
def sum_five_digit_numbers(numbers):

    # Sum up all the five-digit numbers
    return sum(int(number) for number in numbers)


# Main Function
def main(zip_file):

    # Unzip the file
    unzip_file(zip_file)

    # Count subdirectories inside the file_name directory
    num_subdirectories = count_subdirectories(file_name)
    print("Number of subdirectories:", num_subdirectories)
    print()

    # Loop through each subdirectory
    total_num_files = 0
    total_numbers = []
    for subdir in Path(file_name).iterdir():
        if subdir.is_dir():

            # Count the number of files in each subdirectory
            num_files = count_files_in_subdirectory(subdir)
            total_num_files += num_files
            print(f"Number of files in {subdir}: {num_files}")

            # Find and append five-digit numbers to the list
            for file_path in subdir.iterdir():
                if file_path.is_file():
                    total_numbers.extend(find_five_digit_numbers(file_path))

    print()

    # Sum up all the five-digit numbers
    total_sum = sum_five_digit_numbers(total_numbers)
    print("Total number of files:", total_num_files)
    print()
    print("Total number of five-digit numbers found:", len(total_numbers))
    print()
    print("Sum of those numbers:", total_sum)


if __name__ == "__main__":
    print("DATA MINING!!!", end='\n\n')
    main(zip_file_name)

# Jakob West
```