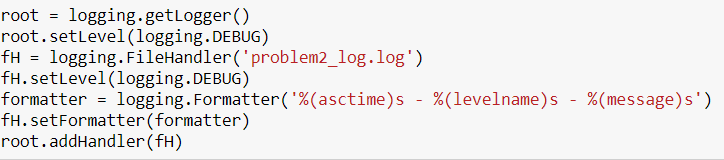
REPORT: Problem 2

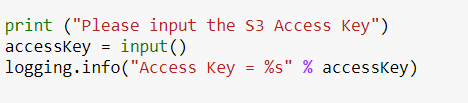
Missing Data Analysis

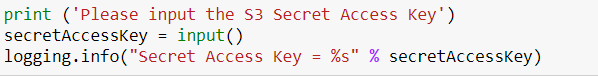
Problem Statement: To analyze the EDGAR Log File Data Set [https://www.sec.gov/data/edgar-log-filedata-set.html]. The page lists the meta data for the datasets and you are expected to develop a pipeline which does the following. Given a year, your program should get data for the first day of the month for every month in the year and process the file for handling missing data.

1. First, Initialize the log file. Use the following code to do that.



1. The program takes the parameter from the user. Here the year of the data would be needed, and it must be ranged from 2003 to 2017. If an invalid year is provided, the program will exit.

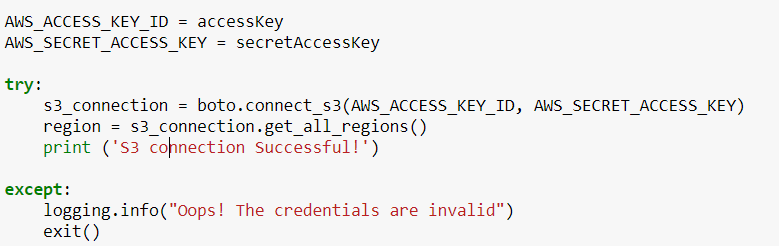








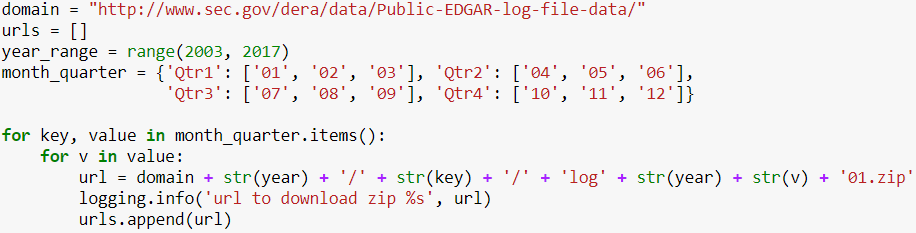
1. Validate the AWS account



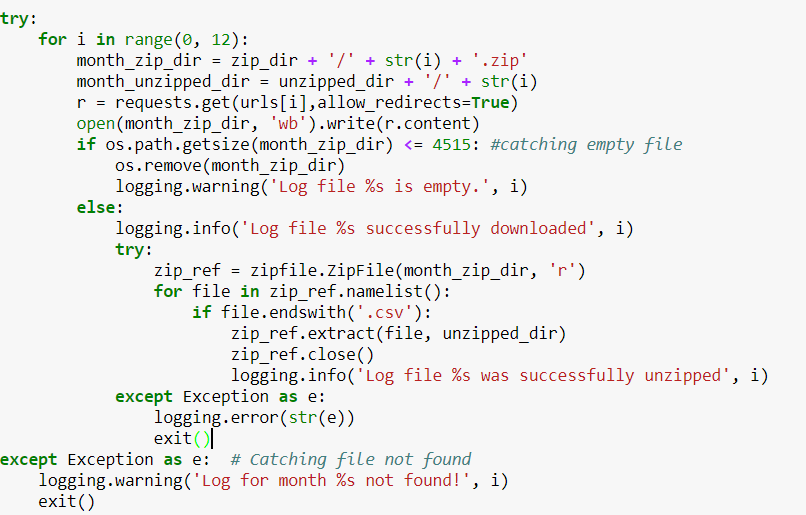
1. Cleaning up the required directory through the following code.



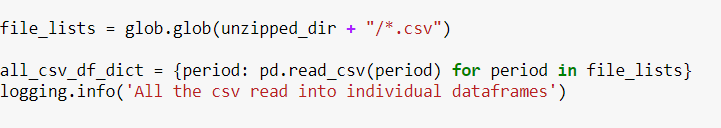
1. Generate the URL to download the zip file. The URL is formed with domain and key-value pairs in month quarter dictionary



1. Download the zip file and unzip it. If the file does not exist, the program will log a warning.



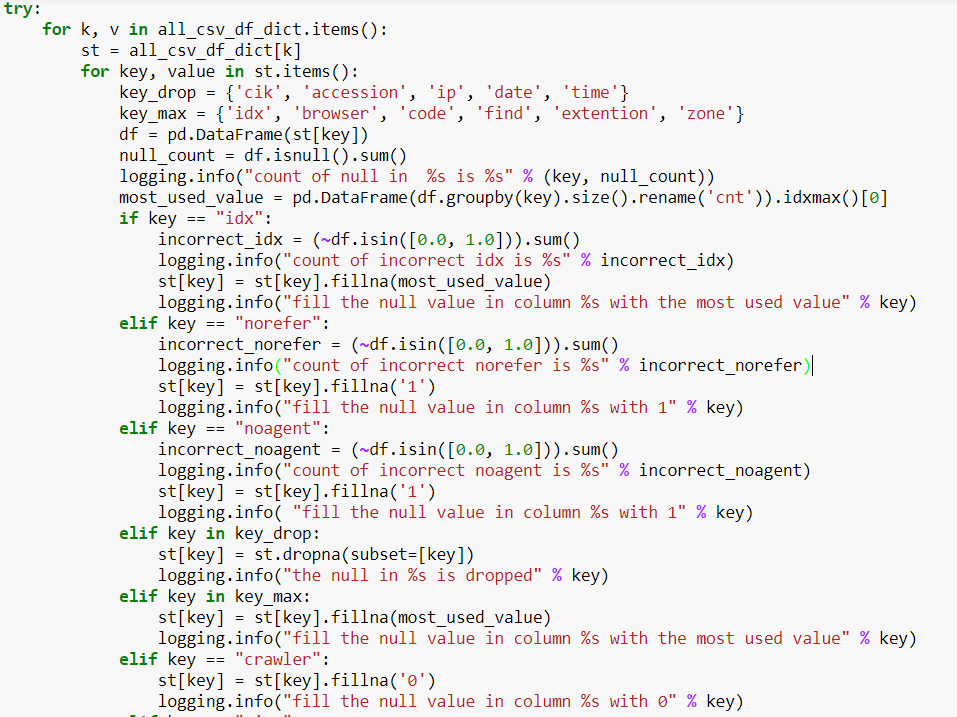
1. Read all unzipped csv files and put them in a Data Frame structure.

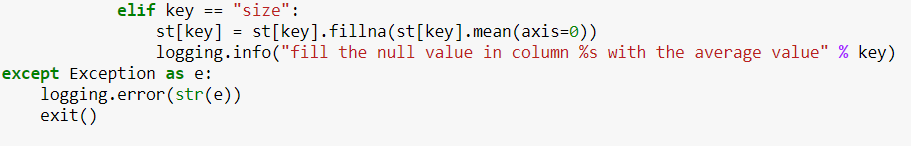


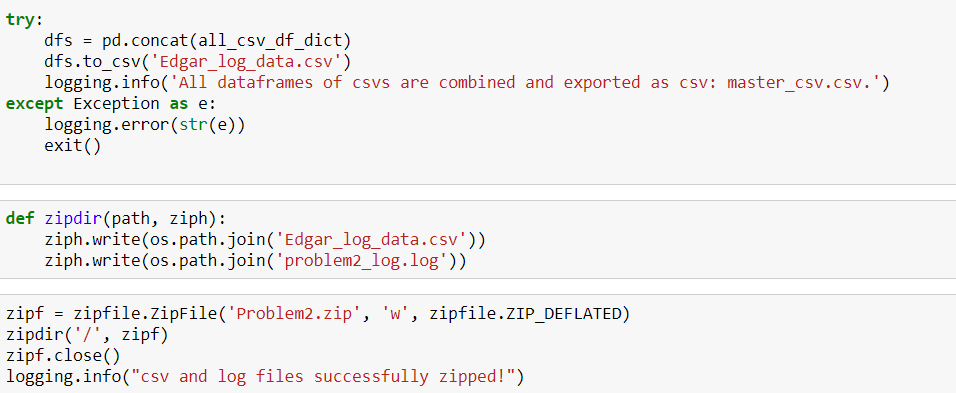
1. Detect the anomalies and clean the data set. By doing so, the program counts the null values, checks if there is any incorrect value instead of 0 and 1 in idx, norefer and noagent fields.

Handle the missing data. The program will delete the rows which have missing values in CIK, accession number, ip, date and time columns. For the columns of idx, browser, code, find, extension, zone, the program will fill the missing values with the most used data of that column; For the columns of norefer, noagent, replace missing values with 1; For the column of crawler, replace the missing values with 0; For the column of size, replace the missing value with average value.

For each month log file, compute the summary matrix





1. Combine all the individual Data Frames into one Data Frame and export a csv file.
2. Zip the csv file
3. Create a new Amazon S3 bucket and upload all files.

