PROJECT REPORT (GROUP-3)

TITLE: -

Extracting data from an Excel file with the concept of field value.

INTRODUCTION: -

- The project requires extraction of data from a given excel file using the concept of field value.
- Using the concept of field value implies that the corresponding fields and their respective values will exist side by side in the excel file.
- Excel sheets are a very intuitive and user-friendly way to manipulate large data sets.
- The idea behind the project is to ease the accessibility of data while working with excel files. Programmatically accessing the data proves beneficial as it allows to bring data from multiple sources into a single program.
- The project extracts and reads data from the provided excel sheet.

SOFTWARE DEVELOPMENT CYCLE: -

Stage 1: - PLANNING AND ANALYSIS

This stage was dedicated to analysis of the requirements of the given project and understanding the ways in which the project could be approached. The expectations from the project and the desired outcomes were thought over to finalise the working methodology.

Stage 2: - DEFINING REQUIREMENTS

The requirements of the project were structured in a definite manner in this stage of project development. Based on the requirements, the working methodology was finalised. Furthermore, the complete structural requirement of the project was finalised.

Stage 3: - DESIGNING THE FUNCTIONS

The third stage of project development was dedicated to deciding the modules and framing the user defined functions to be used in the project. The modules to be used were researched upon and the methods were listed down. Finally, the user defined functions to be used in the project were structured and designed.

Stage 4: - BUILDING THE PROJECT

The fourth stage of project development was dedicated to combining all the user defined functions to be used and completing the source code of the project. The additional requirements of the source code were pondered upon and finalised in this stage. Finally, the conclusions from all the previous stages were taken into consideration and the final source code was written.

Stage 5: - TESTING THE PROJECT

The fifth and final stage of project development was dedicated to testing the project code with the test cases from the sample sheet. The outputs were crosschecked and verified to be in accordance with the requirements of project. The source code was re-verified for the final time and thereby the project development was finalised and concluded.

SYNTAX USED: -

- The modules used in the project are as follows: -

  - **Pandas**: - Pandas is a python library used to analyse data and to work with data sets. It has functions for analysing, cleaning, exploring and manipulating data. The name “Pandas” has a reference to both “Panel Data”, and “Python Data Analysis”.

  - **Sys**: - The sys module provides functions and variables used to manipulate different parts of the Python runtime environment. Furthermore, the sys module provides information about constants, functions and methods of the Python interpreter.

- The methods which have been used from the above-mentioned modules are as follows: -

  • **series( ) :-** It is used to create series. A Pandas Series is like a column in a table. It is a one-dimensional array holding data of any type.

  • **set_index() :-** It is a method to set a List, Series or Data frame as index of a Data Frame. Index column can be set while making a data frame too. But sometimes a data frame is made out of two or more data frames and hence later index can be changed using this method.

  • **read_excel( ) :-** The read_excel( ) function reads the excel sheet data into a DataFrame object.

  • **sys.exit ( ):-** It is a method used to exit the program or a particular block of code explicitly. Argument is optional in this case.

☆What is DataFrame ? :-

A pandas DataFrame is a 2-dimensional data structure, like a 2-dimensional array, or a table with rows and columns.

CODE: -

```
import sys
import pandas as pd

logs=open("C:\logFile\log.txt","a")
logs.write("Name of the user: ")
name=input("Enter your name: ")
logs.write(name)
logs.write("\n")

print("Hello, "+name+"! We help you to search and read information from an excel file")
file_name=input("Enter the path of your excel file: ")

logs=open("C:\logFile\log.txt","a")

try:
  xls=pd.ExcelFile(file_name)
except IOError:  #Exception used when file is not found
   print("File not found, Please enter the correct path of your file")
   sys.exit()

# Read information from a particular file
def read_OneSheet(sheet):
   logs=open("C:\logFile\log.txt","a")
   logs.write("Reading data from ")
   logs.write(sheet)
   logs.write("\n")
   df=pd.read_excel(xls,sheet,header=None)
   print()
   print(df)
   q=df.to_string()
   logs.write(q)
   logs.write("\n")



# Read information from a particular file
def read_allSheet():
   logs=open("C:\logFile\log.txt","a")
   logs.write("Reading data from all the sheets\n")
   logs.write("\n")

   sheet_name=xls.sheet_names
   for tab in sheet_name:
      print()
      print("****************** "+tab+" ******************")
      logs.write("****************** ")
      logs.write(tab)
      logs.write("****************** \n")
      df=pd.read_excel(xls,tab,header=None)

      print()
      print(df)
      q=df.to_string()
      logs.write(q)
      logs.write("\n")
```

```
60
61  # Find the value of the given field
62  def field_value(sheet,field):
63      #updating log file
64      logs=open("C:\logFile\log.txt","a")
65      df=pd.read_excel(xls,sheet,header=None)
66
67
68      df.set_index(0 ,inplace=True)
69      indices = pd.Series(df.index)
70      x=indices[indices== field].index[0]
71      li=list(df.iloc[x,0:])
72      print(field+": ",li)
73
74      logs.write(field)
75      logs.write(": ")
76      for i in li:
77
78          nan_value = float("NaN")
79          if i!=nan_value:
80              logs.write(str(i))
81              logs.write(" ")
82      logs.write("\n")
83
84
85
86  ch=1
87  while True:
88
89      print()
90      print("1.List the names of all the sheets")
91      print("2.Read any one sheet")
92      print("3.Read all the sheets")
93      print("4.Display the value to the field")
94      print("Press -1 to end the program")
95      ch=int(input("Choose: "))
96      logs=open("C:\logFile\log.txt","a")
97      if ch==1:
98          print(xls.sheet_names)
99          logs.write("Sheet names are ")
100         for i in xls.sheet_names:
101             logs.write(i)
102             logs.write(" ")
103         logs.write("\n")
104         print()
105     elif ch==2:
106         sheet=input("Enter the name of sheeet: ")
107         read_OneSheet(sheet)
108     elif ch==3:
109         read_allSheet()
110     elif ch==4:
111         sheet=input("Enter the name of sheeet: ")
112         field=input("Enter the field to obtain corresponding values: ")
113         print()
114         field_value(sheet,field)
115         print()
116
117     elif ch==-1:
118         logs.close()
119         break
120     else:
121         print("Wrong choice")
122
```

EXPLANATION: -

- The project has been made with the help of the modules Pandas and sys. The function of both the modules and the methods used have already been explained above.

```
try:
   xls=pd.ExcelFile(file_name)
except IOError:  #Exception used when file is not found
   print("File not found, Please enter the correct path of your file")
   sys.exit()
```

- We have used the try and except block to handle incorrect entries of the path of file.

```
def read_OneSheet(sheet):
    logs=open("C:\logFile\log.txt","a")
    logs.write("Reading data from ")
    logs.write(sheet)
    logs.write("\n")
    df=pd.read_excel(xls,sheet,header=None)
    print()
    print(df)
    q=df.to_string()
    logs.write(q)
    logs.write("\n")
```

- Next, we have used a user defined function read_OneSheet().

- The function uses the read_excel() method to read the sheet data into a DataFrame object and then prints it.

- The function also stores the data in a log file after extraction using the concept of file handling.

```
38  def read_allSheet():
39      logs=open("C:\logFile\log.txt","a")
40      logs.write("Reading data from all the sheets\n")
41      logs.write("\n")
42
43      sheet_name=xls.sheet_names
44      for tab in sheet_name:
45          print()
46          print("****************** "+tab+" ******************")
47          logs.write("****************** ")
48          logs.write(tab)
49          logs.write("****************** \n")
50          df=pd.read_excel(xls,tab,header=None)
51
52          print()
53          print(df)
54          q=df.to_string()
55          logs.write(q)
56          logs.write("\n")
```

- Then, we have a user defined function named read_allSheet().
- The function firstly extracts the names of all the sheets and then uses a for loop for printing the sheet data with the help of a DataFrame object using the read_excel() method.
- The function also stores the data in a log file after extraction using the concept of file handling.

```
62  def field_value(sheet,field):
63      #updating log file
64      logs=open("C:\logFile\log.txt","a")
65      df=pd.read_excel(xls,sheet,header=None)
66
67
68      df.set_index(0 ,inplace=True)
69      indices = pd.Series(df.index)
70      x=indices[indices== field].index[0]
71      li=list(df.iloc[x,0:])
72      print(field+": ",li)
73
74      logs.write(field)
75      logs.write(": ")
76      for i in li:
77
78          nan_value = float("NaN")
79          if i!=nan_value:
80              logs.write(str(i))
81              logs.write(" ")
82      logs.write("\n")
```

- Thereafter, we have used another user defined function named field_value().
- The function firstly reads the sheet data into a DataFrame object and then uses the set_index() and Series() methods to store and print the values of the field entered by the user.
- The function also stores the data in a log file after extraction using the concept of file handling.

```
ch=1
while True:

    print()
    print("1.List the names of all the sheets")
    print("2.Read any one sheet")
    print("3.Read all the sheets")
    print("4.Display the value to the field")
    print("Press -1 to end the program")
    ch=int(input("Choose: "))
    logs=open("C:\logFile\log.txt","a")
    if ch==1:
        print(xls.sheet_names)
        logs.write("Sheet names are ")
        for i in xls.sheet_names:
            logs.write(i)
            logs.write(" ")
        logs.write("\n")
        print()
    elif ch==2:
        sheet=input("Enter the name of sheeet: ")
        read_OneSheet(sheet)
    elif ch==3:
        read_allSheet()
    elif ch==4:
        sheet=input("Enter the name of sheeet: ")
        field=input("Enter the field to obtain corresponding values: ")
        print()
        field_value(sheet,field)
        print()

    elif ch==-1:
        logs.close()
        break
    else:
        print("Wrong choice")
```

- Finally, we have used an infinite while loop to provide choices to the users and based on that the different functions are called to give the required output.

EXECUTION: -

The screenshots of execution of all the user defined functions are attached hereafter: -

1) Following is the execution screenshot of the read_OneSheet() function

```
Choose: 2

Enter the name of sheeet: Sheet2

                0                          1          2          3
0  Patient's Name              Pratyush Chand        NaN        NaN
1             Age                          43        NaN        NaN
2        Phone No                  9123436477        NaN        NaN
3         Address  2a parkstreet,kolkata-54        NaN        NaN
4           Email  pratyush.chand@gmail.com        NaN        NaN
5             NaN                         NaN        NaN        NaN
6       Medicines                Paracetamol     Clavam  Trobicin
7            Time                    morning  afternoon     night
8           Price                         55         90        105
9           Doses                          2          1          2
```

2) Following is the execution screenshot of the read_allSheet() function

```
Choose: 3

****************** Sheet1 *************************

          0              1      2       3    4
0      Name  Anurag Adarsh    NaN     NaN  NaN
1   Roll no    13000219122    NaN     NaN  NaN
2   Section              B    NaN     NaN  NaN
3    Gender           Male    NaN     NaN  NaN
4       NaN            NaN    NaN     NaN  NaN
5       NaN            NaN    NaN     NaN  NaN
6  Subjects        English  Maths  Python  Dsa
7       CA1             22     25      20   14
8       CA2             23     25      20   14
9       CA3             21     12      17   24
10      CA4             19     16      19   22
11     PCA1             35     32      22   27
12     PCA2             21     25      24   34

****************** Sheet2 *************************

                0                          1          2          3
0  Patient's Name              Pratyush Chand        NaN        NaN
1             Age                          43        NaN        NaN
2        Phone No                  9123436477        NaN        NaN
3         Address  2a parkstreet,kolkata-54        NaN        NaN
4           Email  pratyush.chand@gmail.com        NaN        NaN
5             NaN                         NaN        NaN        NaN
6       Medicines                Paracetamol     Clavam  Trobicin
7            Time                    morning  afternoon     night
8           Price                         55         90        105
9           Doses                          2          1          2
```

3) Following are the execution screenshots of the field_value() function for Sheet1 and Sheet2 respectively

```
Choose: 4

Enter the name of sheeet: Sheet1

Enter the field to obtain corresponding values: Subjects

Subjects:  ['English', 'Maths', 'Python', 'Dsa']
```

```
Choose: 4

Enter the name of sheeet: Sheet2

Enter the field to obtain corresponding values: Patient's Name

Patient's Name:  ['Pratyush Chand', nan, nan]
```

4) Following are the screenshots of data being extracted and stored in the log file.

```
Name of the user: Vishal
Subjects: English Maths Python Dsa
CA1: 22 25 20 14
CA2: 23 25 20 14
CA3: 21 12 17 24
CA4: 19 16 19 22
PCA1: 35 32 22 27
PCA2: 21 25 24 34
Medicines: Paracetamol Clavam Trobicin
Time: morning afternoon night
Price: 55 90 105
Doses: 2 1 2
```

```
Name of the user: vishal
Sheet names are Sheet1 Sheet2
Reading data from Sheet1
           0             1      2       3    4
0       Name  Anurag Adarsh    NaN     NaN  NaN
1    Roll no    13000219122    NaN     NaN  NaN
2    Section              B    NaN     NaN  NaN
3     Gender           Male    NaN     NaN  NaN
4        NaN            NaN    NaN     NaN  NaN
5        NaN            NaN    NaN     NaN  NaN
6   Subjects        English  Maths  Python  Dsa
7        CA1             22     25      20   14
8        CA2             23     25      20   14
9        CA3             21     12      17   24
10       CA4             19     16      19   22
11      PCA1             35     32      22   27
12      PCA2             21     25      24   34
```

REFERENCES: -

1) https://wiki.python.org/moin/

2) https://www.w3schools.com/python/

3) https://www.geeksforgeeks.org/python-pandas-dataframe-set_index/

4)https://stackoverflow.com/questions/35818873/pandas-series-creation-using-dataframe-columns-returns-nan-data-entries/

MEMBER DETAILS AND CONTRIBUTION: -

GROUP 3

- Piyush Kumar(13000219124) :- Research for module used and help in code.
- Sejal Agarwal(13000219106) :- Writing the source code.
- Anurag Adarsh(13000219122) :- Framing the contents of Project Report.
- Pratyush Chand(13000219103):- Preparation of the Project Report.
- Vishal Kumar Jha(13000219136):- Preparation of the Project PPT.