

# ANLT5020 – Unit 3

## Assignment 1 Tutorial

SAS Studio



# Instructions

- Create a new SAS dataset using data from Teachers.xlsx.
- Complete the following using this dataset:
- For the character variables FirstName, LastName, Gender, and Subject, run an SAS procedure that will detect the invalid values within these variables. List the invalid values for each variable, if any.
- For the variable Salary, run an SAS procedure to find any salaries that seem to be unusual compared to the rest of the dataset. List the name of the employee with an invalid Salary value.
- For the variable LastName, note that there are extra blanks in some of the last names. Use a DATA step with a SET statement and a COMPRESS command to create a new dataset without these blanks. (Hint: You will need to create a new variable name for the compressed LastName.)
- Write another DATA step to fix other errors in the new dataset you created. (Hint: Make use of the SET statement.) There are three errors in the new dataset:
- Some teacher names are spelled incorrectly.
- All Genders should be capital letters.
- The actual salary of Employee\_ID=E030 is \$73,000.
- Use a PROC PRINT statement to print out the new dataset with the errors correcte.

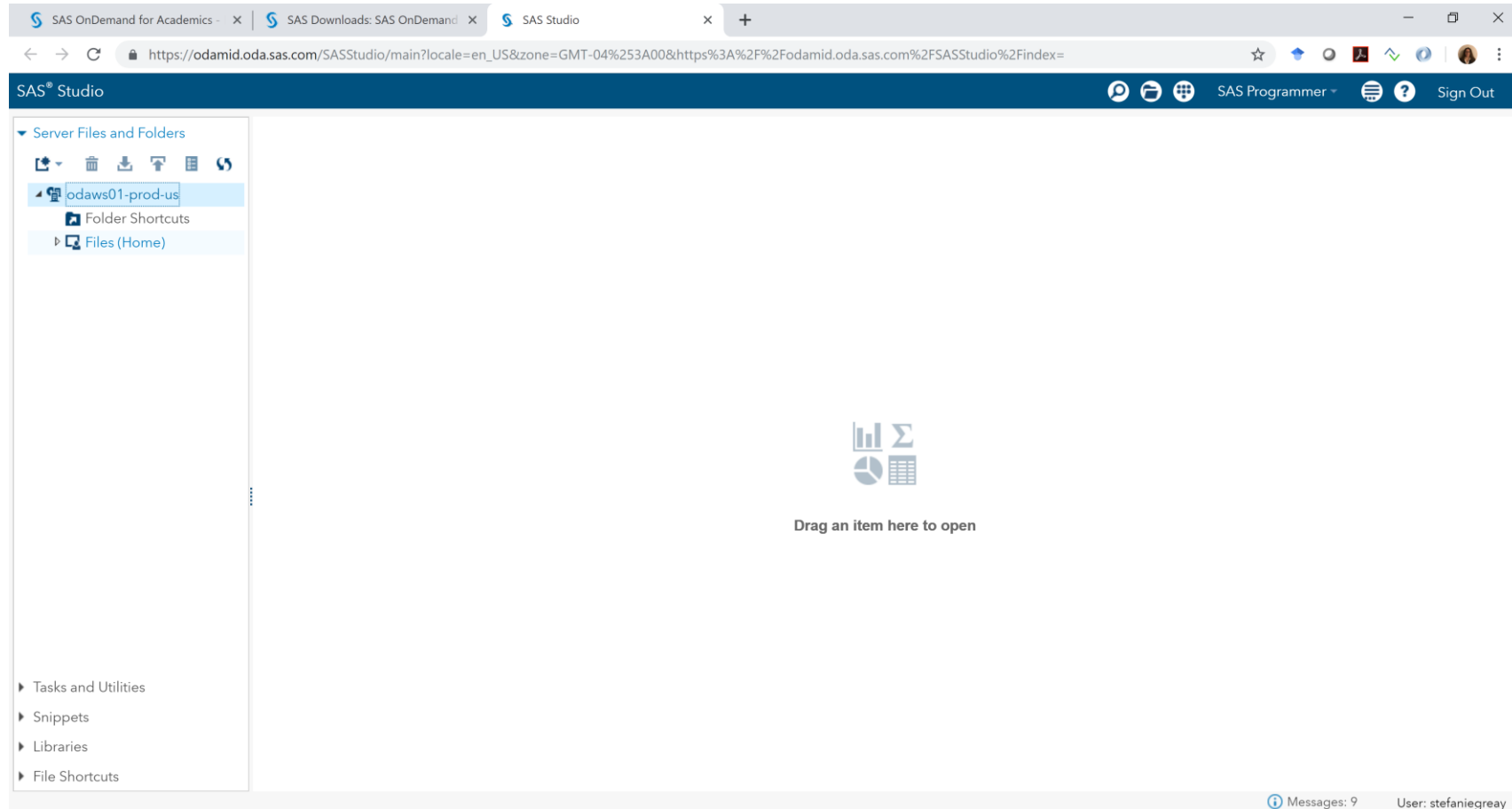


# Dataset

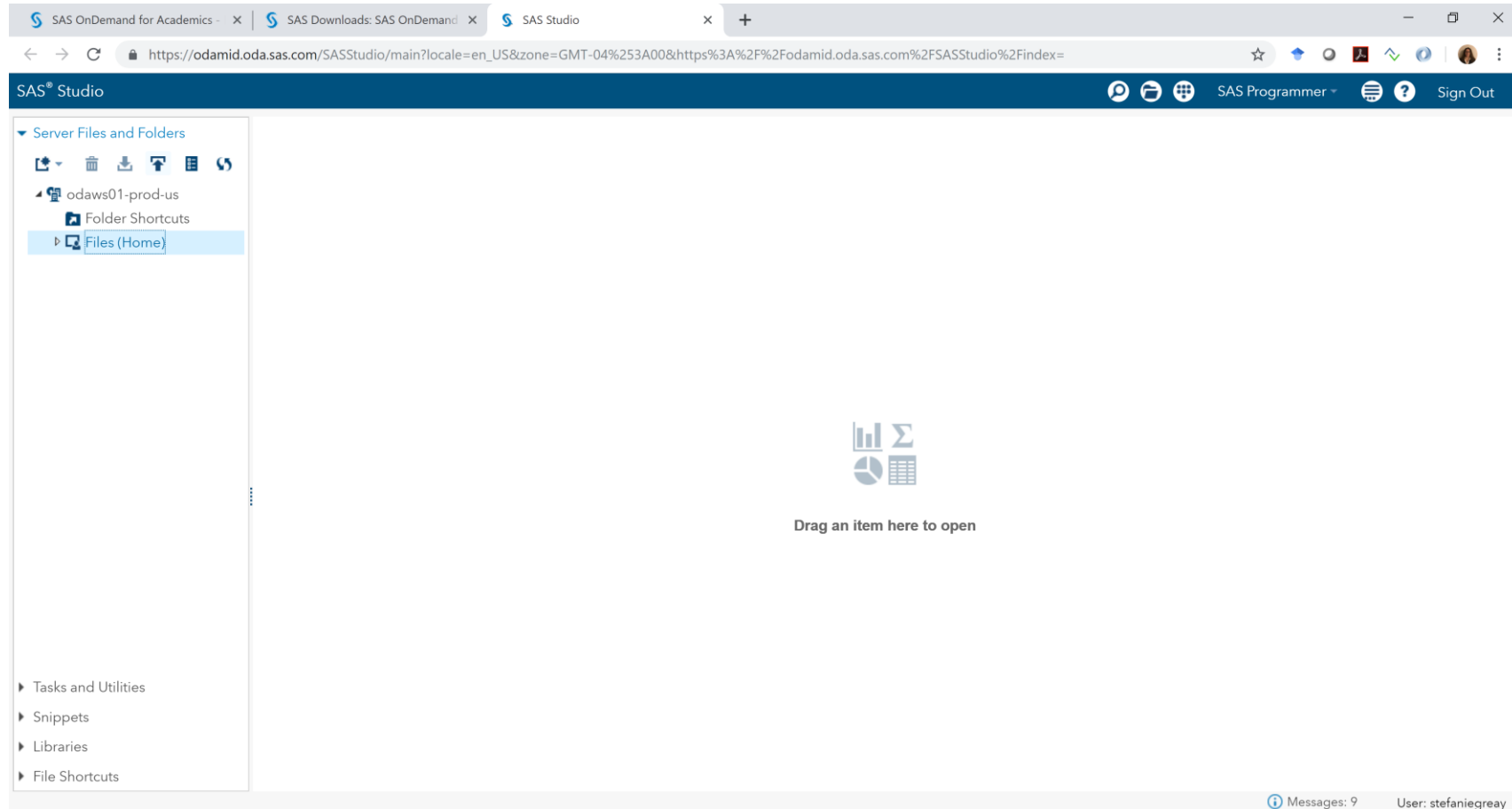
- Download the Teachers.xlsx file from the course datasets zip file or from the Unit 3 Welcome announcement in the course announcements.



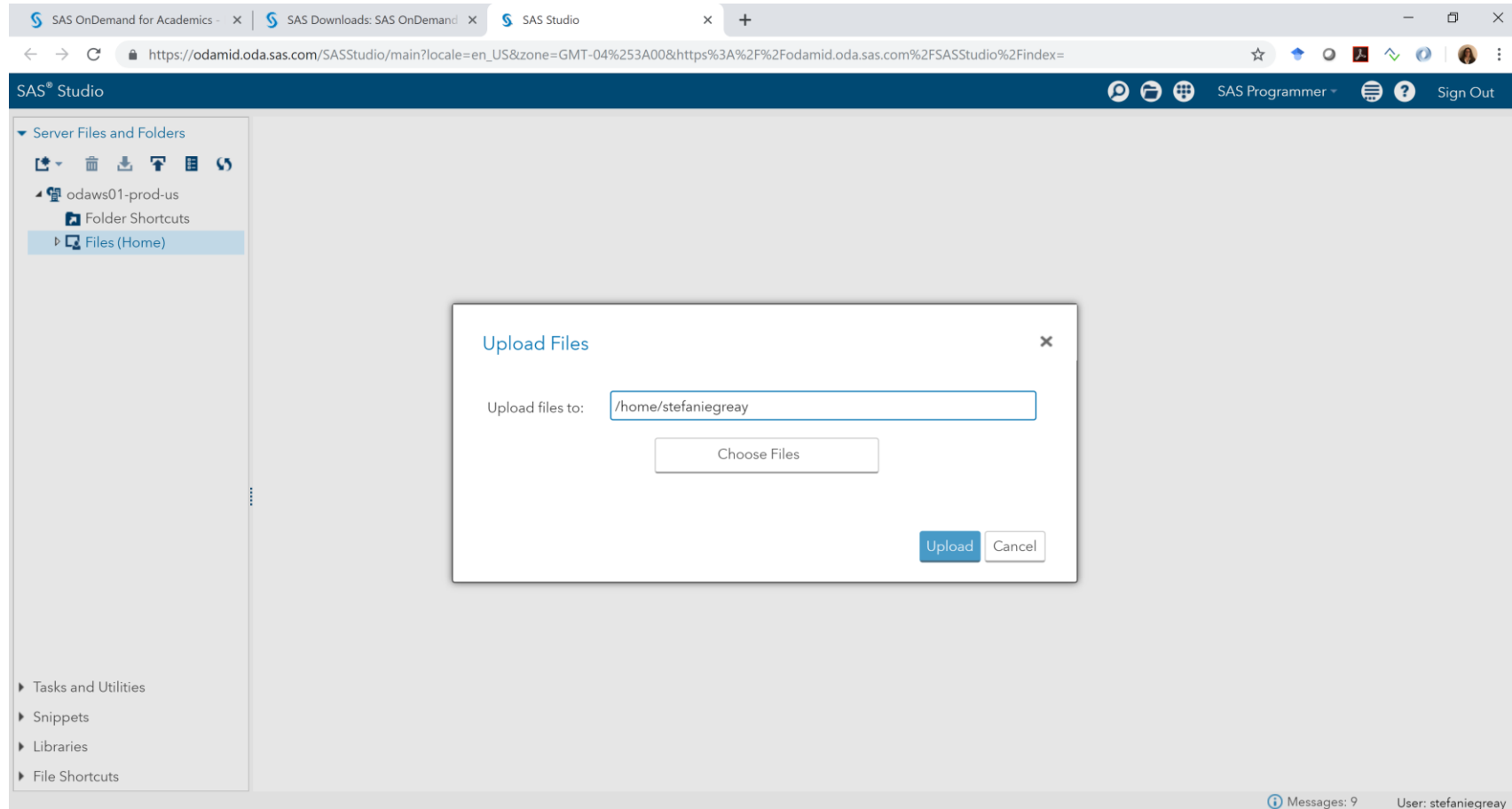
# Click on Files(Home)



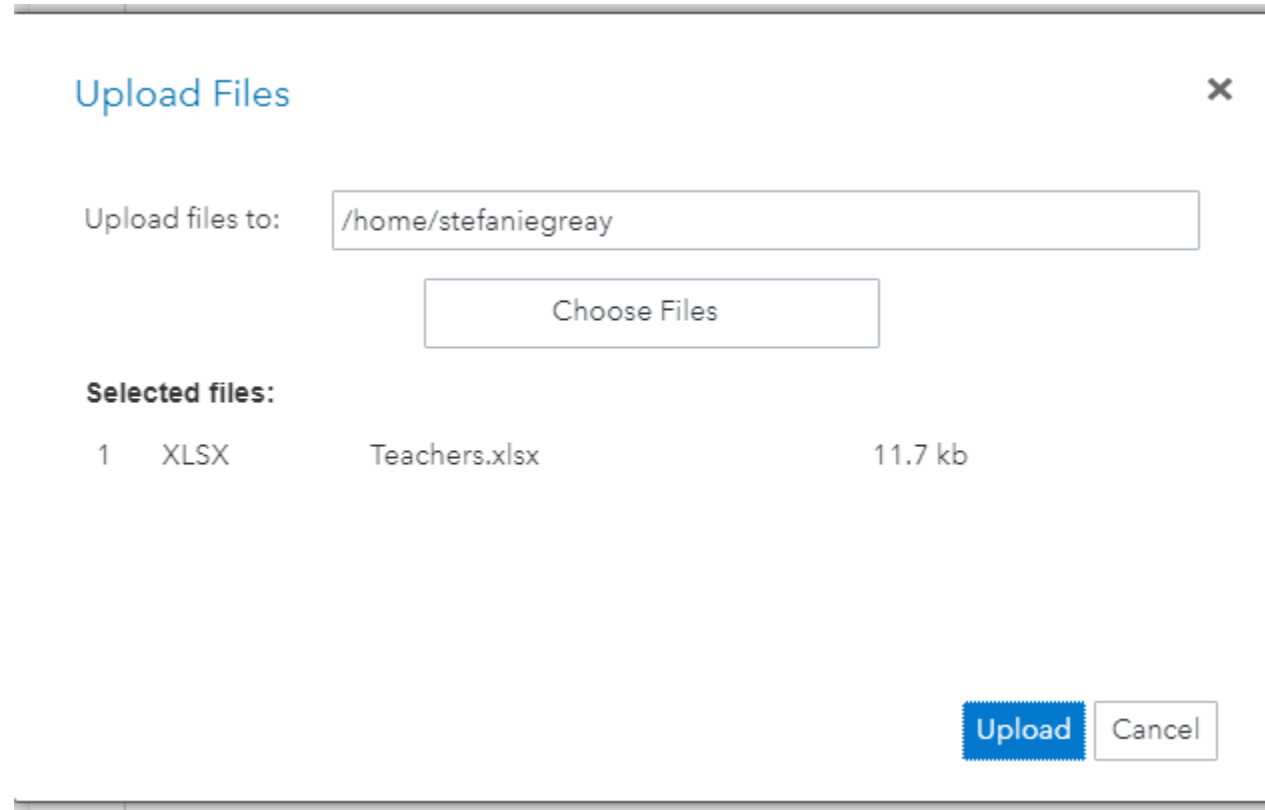
# The Upload button will display in dark blue



You can create a folder at this point, if you wish, or simply upload to your home directory.



Select “Choose Files” to browse your computer for the dataset you want to upload. Once the dataset has been selected, click “Upload.”

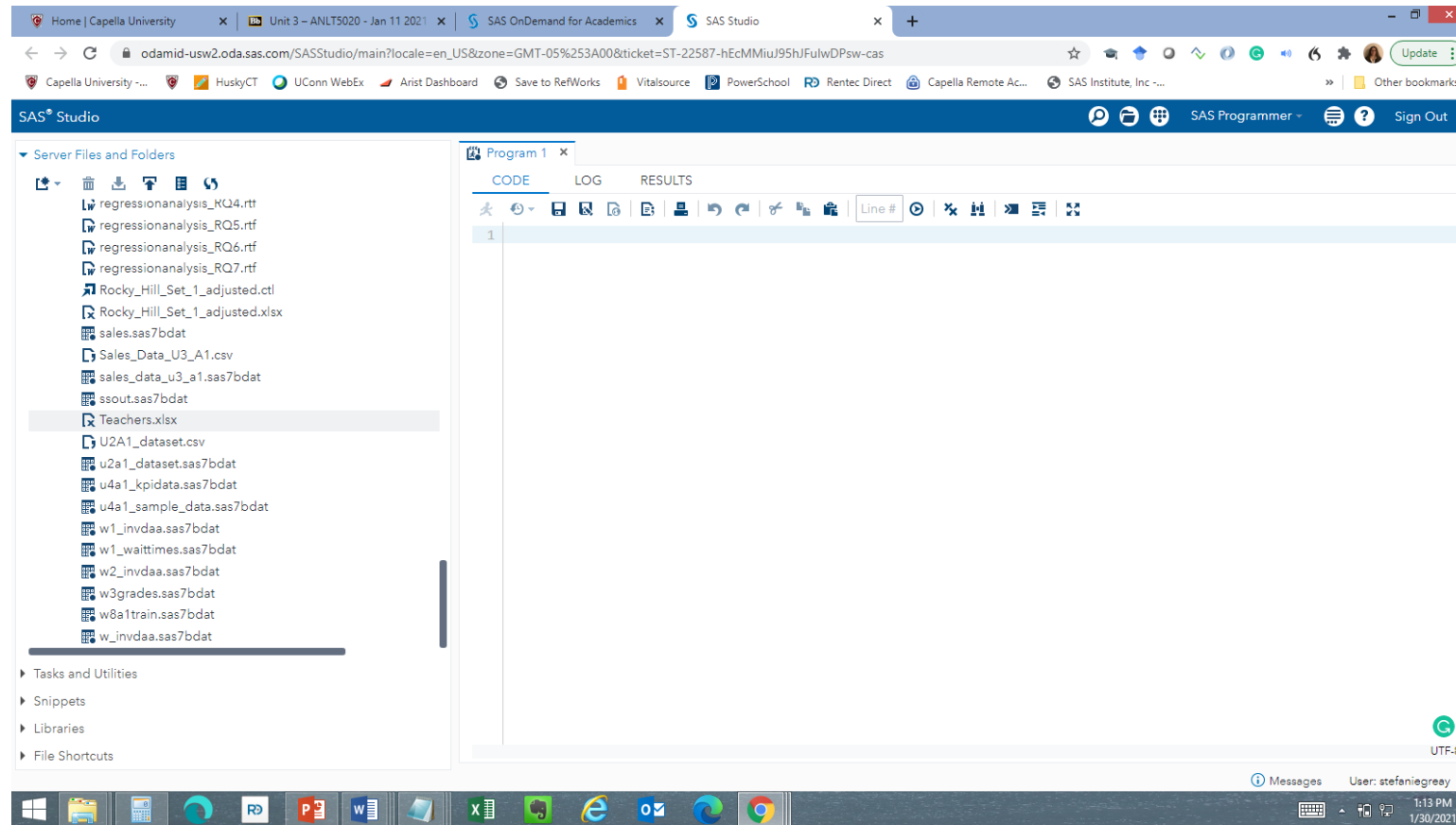


The screenshot shows a web-based 'Upload Files' dialog box. At the top, the title 'Upload Files' is in blue, with a close button (X) to its right. Below the title, there is a text input field labeled 'Upload files to:' containing the path '/home/stefaniegreay'. Underneath this field is a button labeled 'Choose Files'. Below the 'Choose Files' button, the section 'Selected files:' is displayed. It contains a single entry: '1 XLSX Teachers.xlsx 11.7 kb'. At the bottom right of the dialog, there are two buttons: a blue 'Upload' button and a white 'Cancel' button with a grey border.

Selected files:			
1	XLSX	Teachers.xlsx	11.7 kb

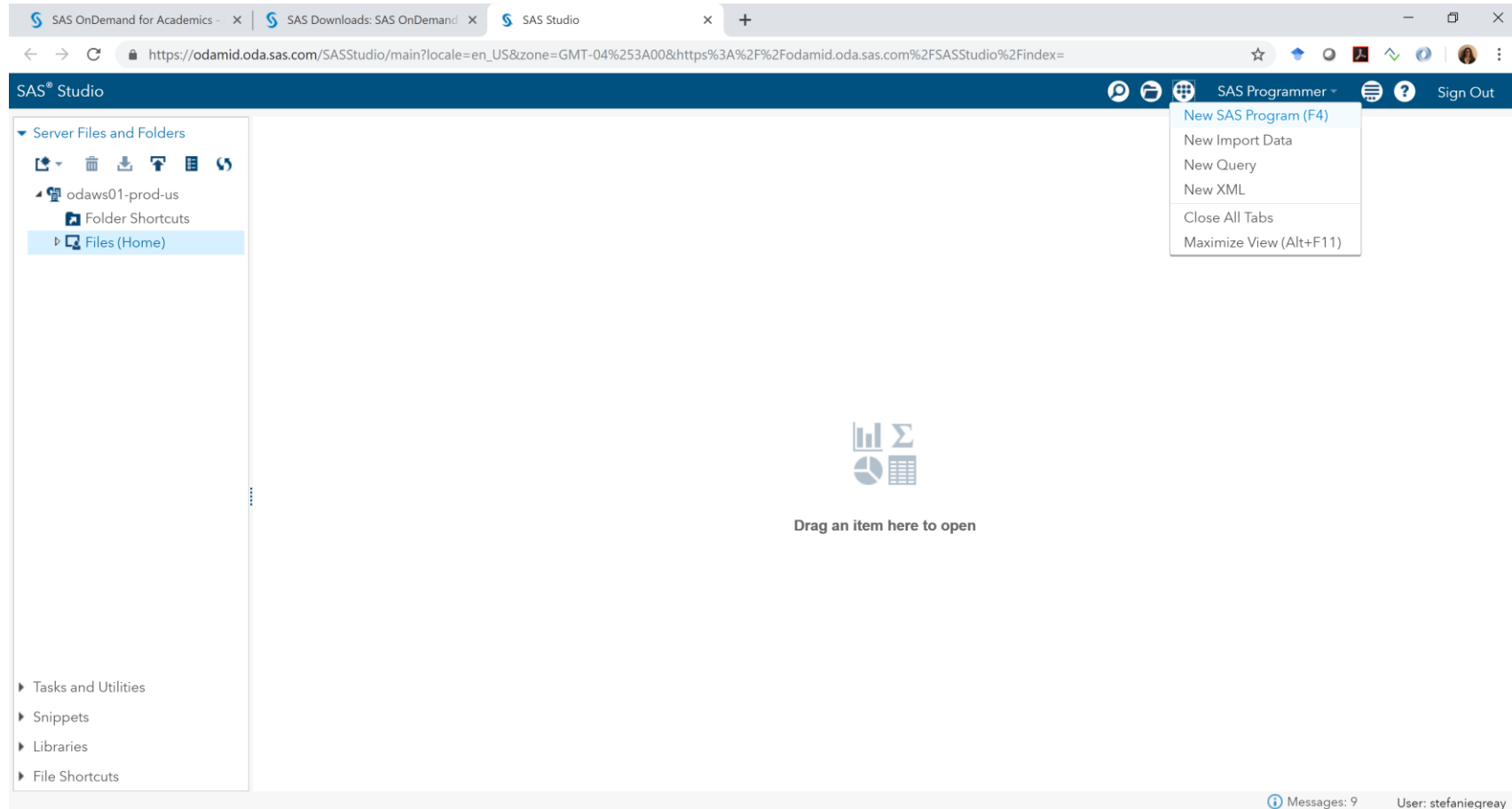


You will be able to view your files by clicking on “Files(Home)” to verify that your file successfully uploaded.

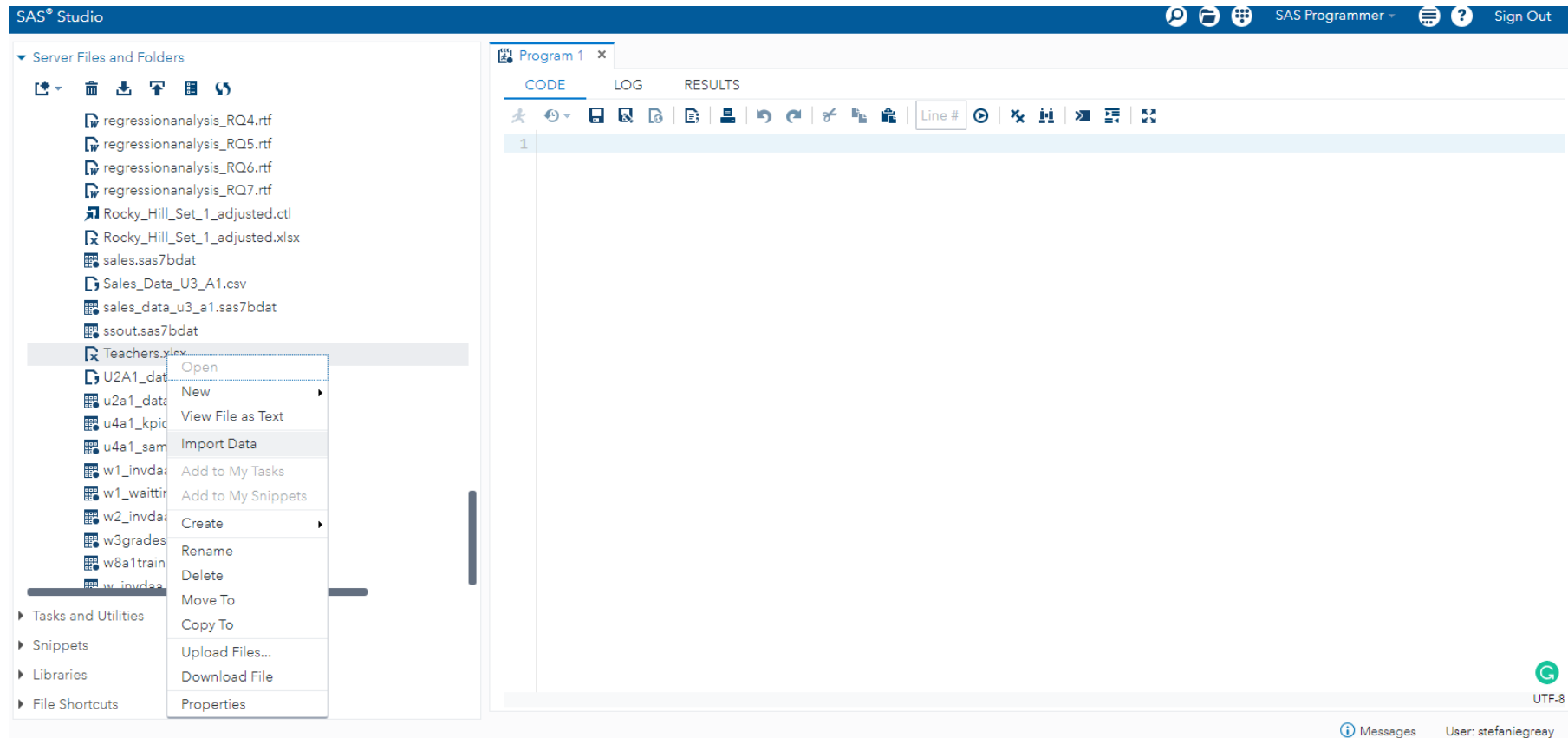




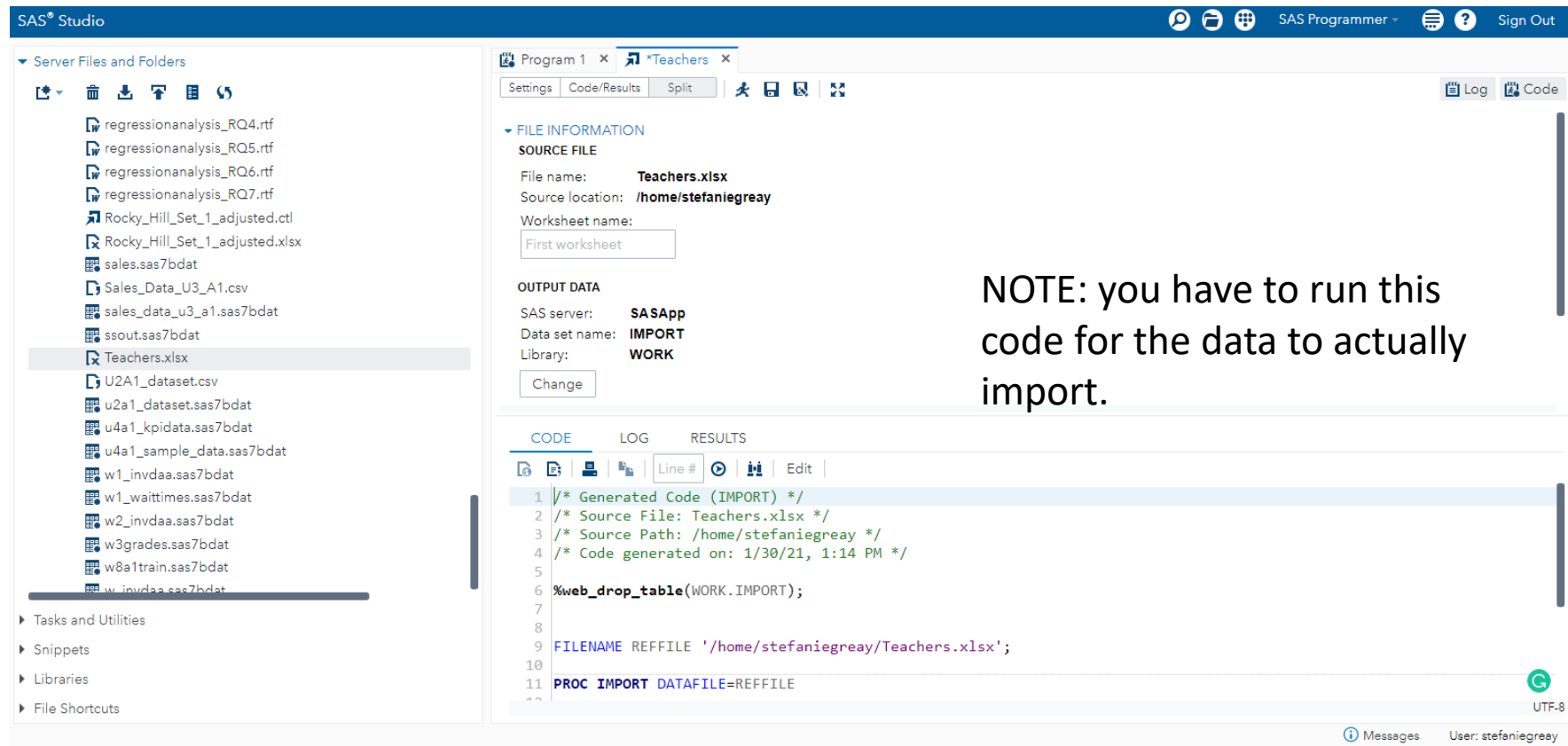
# To get started with the SAS portion of the Unit 3 Assignment 1 assignment, start a new SAS program.



# Import the dataset into a SAS dataset format (from the current xlsx format)



# The Proc Import code will be written for you (save this as a template to use for future imports!)



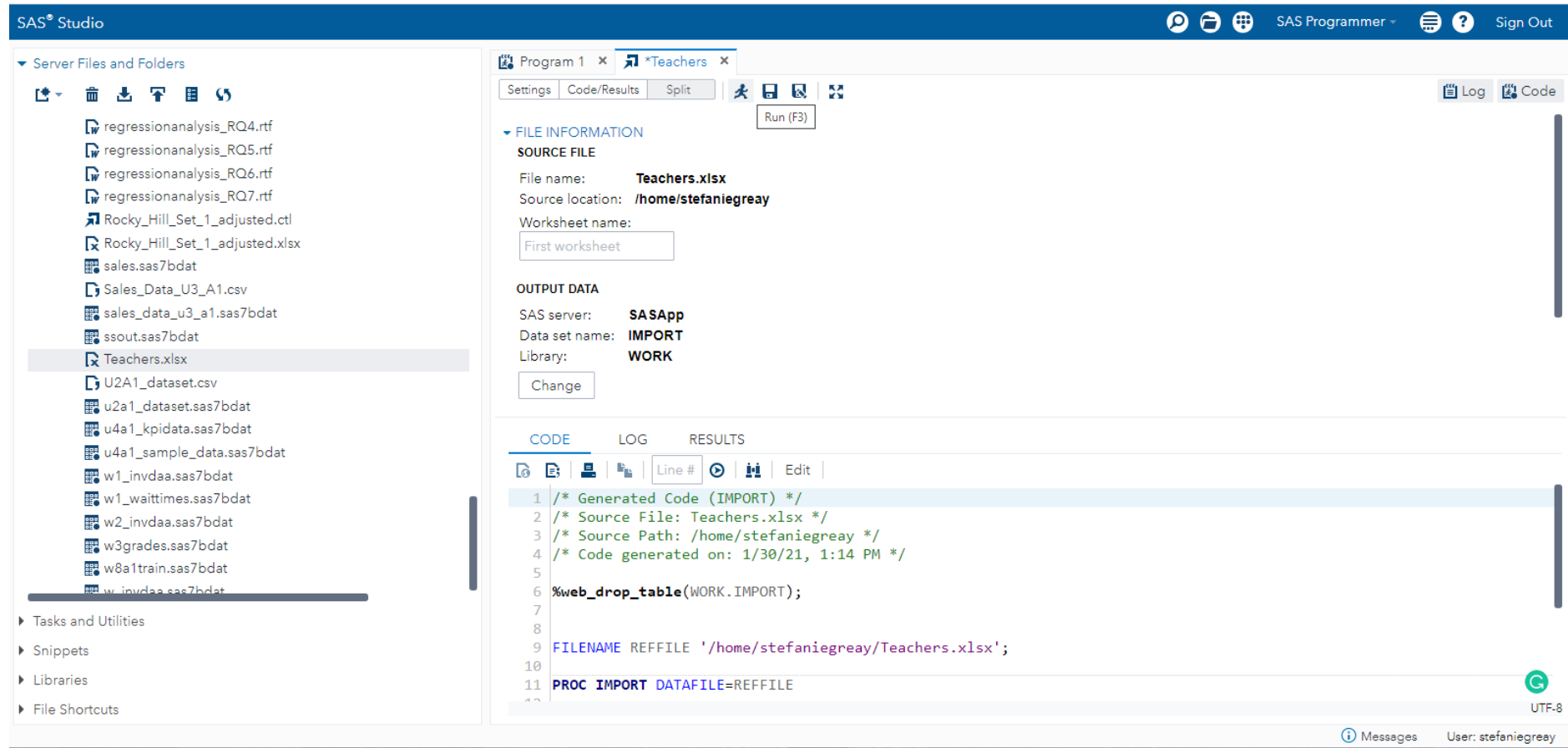
The screenshot shows the SAS Studio interface. On the left, the 'Server Files and Folders' pane lists various files, with 'Teachers.xlsx' selected. The main window is divided into three sections: 'FILE INFORMATION', 'OUTPUT DATA', and 'CODE'. The 'FILE INFORMATION' section shows the source file 'Teachers.xlsx' at the location '/home/stefaniegreay' with the worksheet 'First worksheet'. The 'OUTPUT DATA' section shows the SAS server 'SASApp', data set name 'IMPORT', and library 'WORK'. The 'CODE' section displays the generated SAS code for importing the data.

**NOTE: you have to run this code for the data to actually import.**

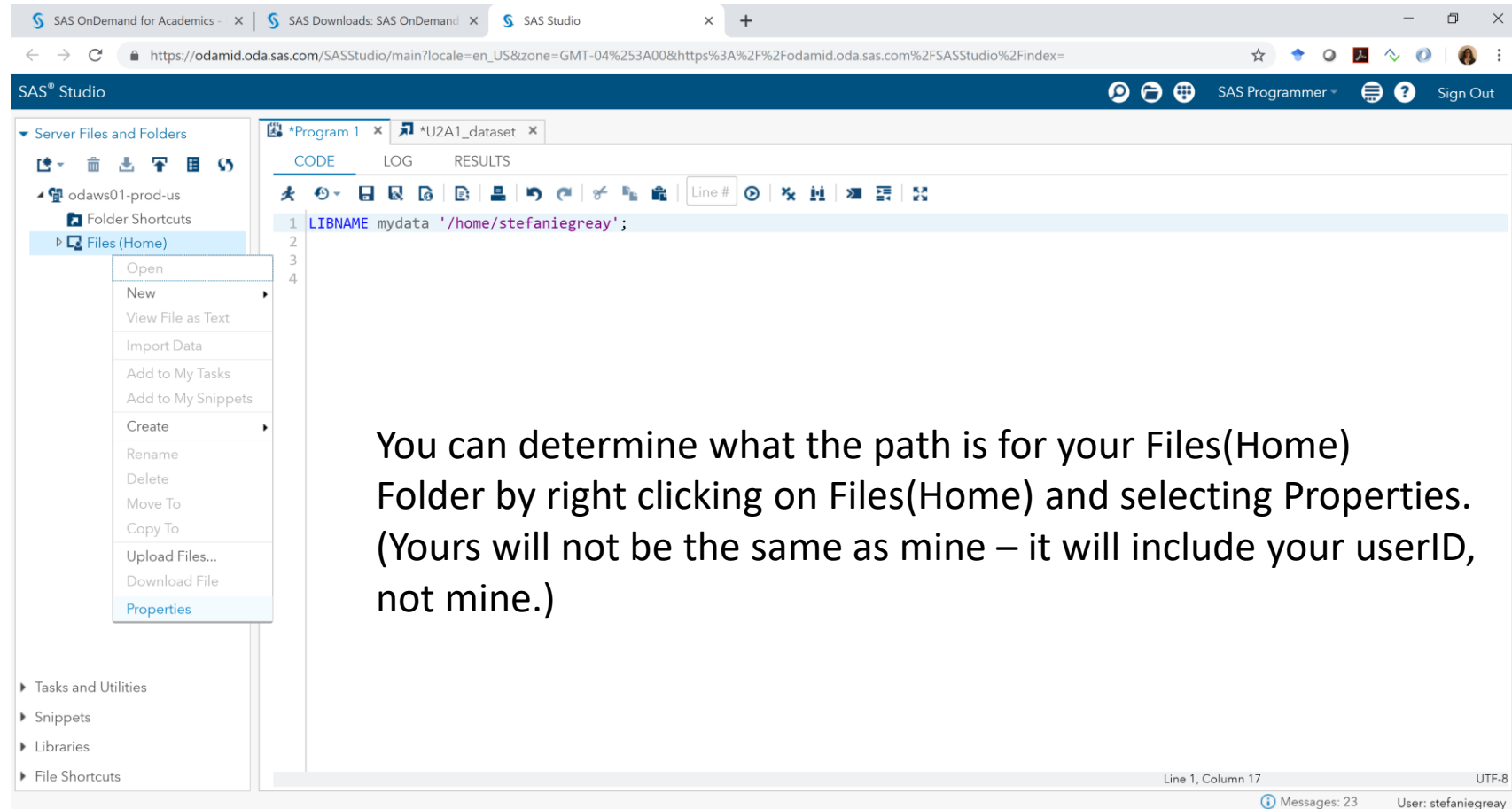
```
1 /* Generated Code (IMPORT) */  
2 /* Source File: Teachers.xlsx */  
3 /* Source Path: /home/stefaniegreay */  
4 /* Code generated on: 1/30/21, 1:14 PM */  
5  
6 %web_drop_table(WORK.IMPORT);  
7  
8  
9 FILENAME REFFILE '/home/stefaniegreay/Teachers.xlsx';  
10  
11 PROC IMPORT DATAFILE=REFFILE  
12
```



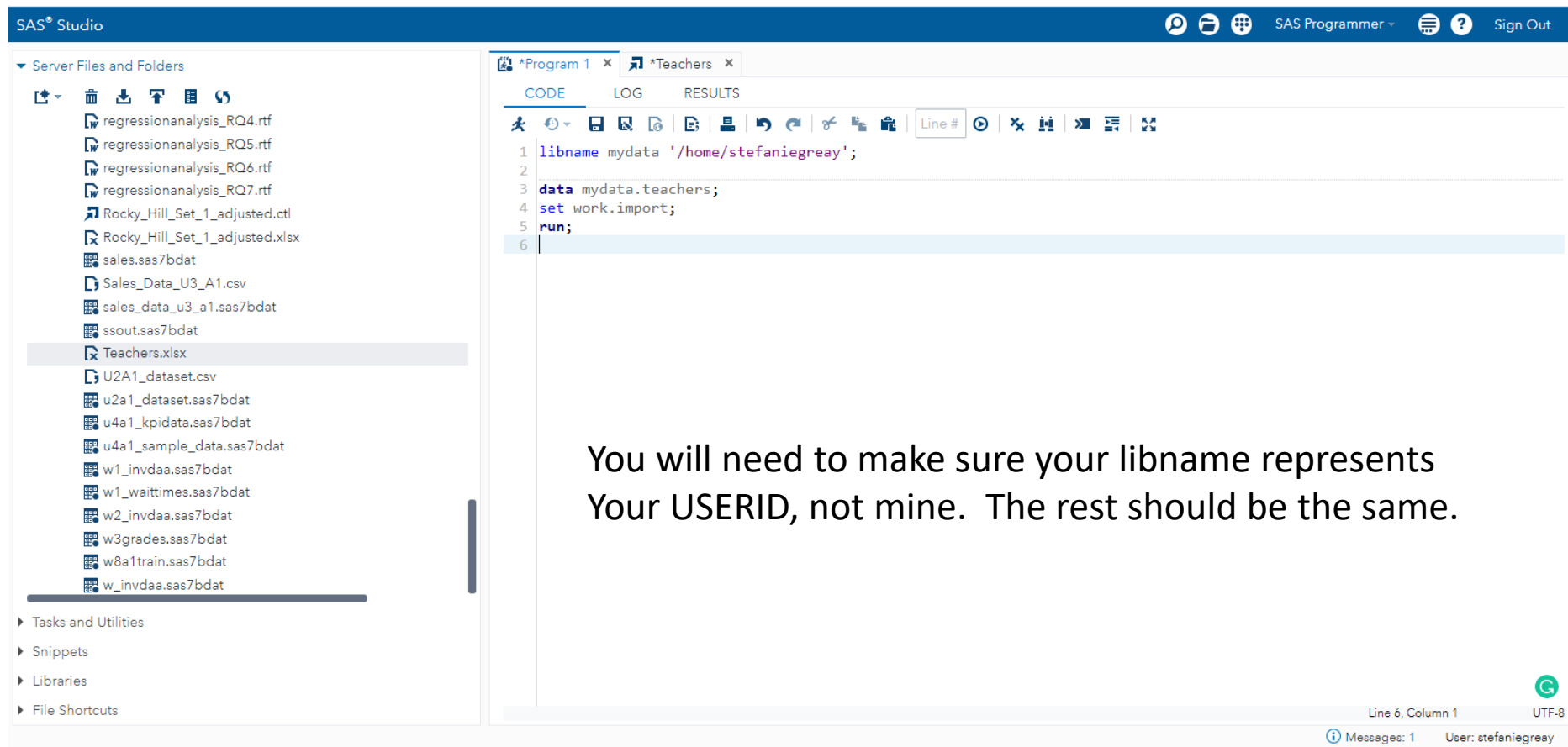
# To run the code, click the icon that looks like a guy running.



# To create a SAS Library for your Files(Home) folder, you need to use a libname statement



# Save the temporary SAS dataset created by the import to your library using the following sample code.



When you run the code, you will see the dataset in the output data window and can verify its success.

SAS® Studio

Server Files and Folders

- regressionanalysis\_RQ5.rtf
- regressionanalysis\_RQ6.rtf
- regressionanalysis\_RQ7.rtf
- Rocky\_Hill\_Set\_1\_adjusted.ctl
- Rocky\_Hill\_Set\_1\_adjusted.xlsx
- sales.sas7bdat
- Sales\_Data\_U3\_A1.csv
- sales\_data\_u3\_a1.sas7bdat
- ssout.sas7bdat
- teachers.sas7bdat
- Teachers.xlsx**
- U2A1\_dataset.csv
- u2a1\_dataset.sas7bdat
- u4a1\_kpidata.sas7bdat
- u4a1\_sample\_data.sas7bdat
- w1\_invdaa.sas7bdat
- w1\_waittimes.sas7bdat
- w2\_invdaa.sas7bdat
- w3grades.sas7bdat
- w8a1train.sas7bdat
- w\_invdaa.sas7bdat

Tasks and Utilities

Snippets

Libraries

File Shortcuts

\*Program 1 x \*Teachers x

CODE LOG RESULTS **OUTPUT DATA**

Table: MYDATA.TEACHERS View: Column names Filter: (none)

Columns

- ☒ Select all
- ☒ Employee\_ID
- ☒ FirstName
- ☒ LastName
- ☒ Gender
- ☒ DOH
- ☒ Years
- ☒ Salary
- ☒ Subject

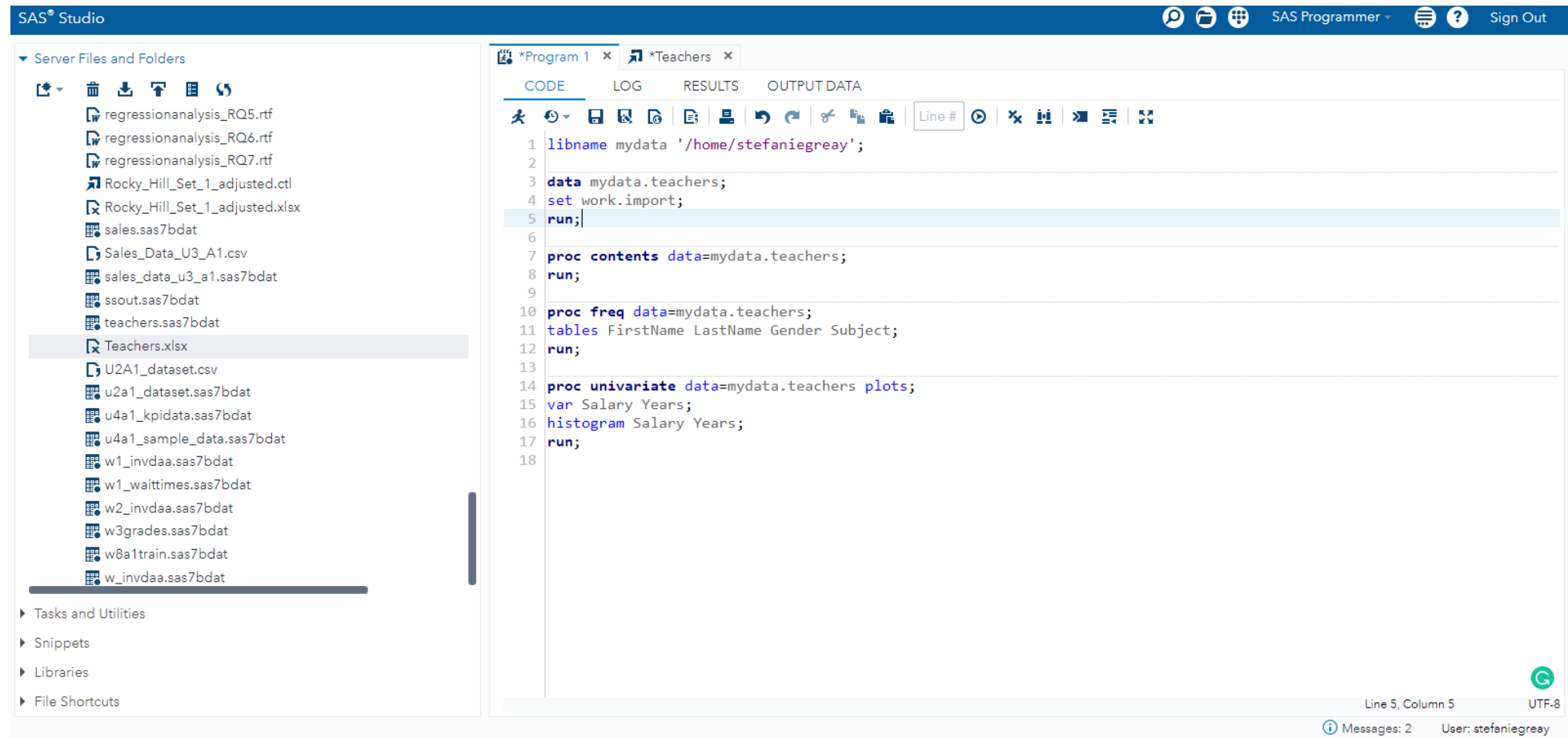
Property Value

Employee_ID	FirstName	LastName	Gender	DOH	Years	Salary	Subject
1 E001	Michael	Jones	M	12/31/2002	14	\$70,000	Math
2 E002	Jason	Myers	M	01/19/2008	8	\$55,000	Englis
3 E003	Michelle	Lewis	F	04/13/1996	20	\$80,000	Physic
4 E004	Timothy	Pal mer	M	06/04/1999	17	\$70,000	Math
5 E005	Rosa	Brown	F	07/12/2001	15	\$70,000	Chem
6 E006	Patricia	Butler	F	07/18/1998	18	\$75,000	Accou
7 E007	Anthony	Lawrence	M	05/10/2007	9	\$55,000	Comp
8 E008	Katherine	Stewart	F	07/26/1995	21	\$83,000	Accou
9 E009	Rebecca	Miller	F	10/28/2004	12	\$62,000	Math
10 E010	Morgan	Smith	f	07/25/1995	21	\$84,000	Biolog
11 E011	Barbara	Small	F	09/22/2005	11	\$59,000	Histor
12 E012	David	McDaniel	M	11/17/2003	13	\$65,000	Physic
13 E013	Mark	Frost	M	07/14/1996	20	\$81,000	Englis
14 E014	Robert	Wagner	M	11/25/2003	13	\$67,000	Mth
15 E015	Patricia	Keller	F	09/07/1998	18	\$78,000	Comp
16 E016	Lisa	Waters	F	03/26/2007	9	\$53,000	Techn
17 E017	Laura	Green	F	10/06/2004	12	\$63,000	Music
18 E018	Jason	Palmer	m	04/08/2002	14	\$71,000	Enalis

Messages: 2 User: stefaniegreay



# You can now run any procedures against that dataset via the code window.





# Sample Code for identifying issues

```
libname mydata '/home/stefaniegreay';

data mydata.teachers;
set work.import;
run;

proc contents data=mydata.teachers;
run;

proc freq data=mydata.teachers;
tables FirstName LastName Gender Subject;
run;

proc univariate data=mydata.teachers plots;
var Salary Years;
histogram Salary Years;
run;
```



Once you run the code, you can review the output to identify data issues, like those in the LastName variable.

The screenshot shows the SAS Studio interface. On the left, the 'Server Files and Folders' pane lists various files, with 'Teachers.xlsx' selected. The main window displays the 'RESULTS' tab for a program named '\*Program 1'. It shows a 'Table of Contents' and a data table with 5 columns. The data table contains 20 rows of names and numerical values. At the bottom right, it indicates 'Messages: 3' and 'User: stefaniegreay'.

Green	1	2.56	9	23.08
Henry	1	2.56	10	25.64
Holder	1	2.56	11	28.21
Jacob s	1	2.56	12	30.77
Jacobs	1	2.56	13	33.33
Jone s	1	2.56	14	35.90
Jones	1	2.56	15	38.46
Jordan	1	2.56	16	41.03
Keller	1	2.56	17	43.59
Lawrence	1	2.56	18	46.15
Lewis	1	2.56	19	48.72
McDaniel	1	2.56	20	51.28
Miller	1	2.56	21	53.85
Myers	1	2.56	22	56.41
Nelson	1	2.56	23	58.97
O'Connor	1	2.56	24	61.54
Pal mer	1	2.56	25	64.10
Palmer	1	2.56	26	66.67
Powell	1	2.56	27	69.23
Richards	1	2.56	28	71.79
Robbins	1	2.56	29	74.36
Rollins	1	2.56	30	76.92
Small	1	2.56	31	79.49
Smi th	1	2.56	32	82.05
Smith	3	7.69	35	89.74
Stewart	1	2.56	36	92.31
Tyler	1	2.56	37	94.87
Wagner	1	2.56	38	97.44
Waters	1	2.56	39	100.00



# Sample Code

```
libname mydata '/home/stefaniegreay';
```

```
data mydata.teachers;  
set work.import;  
run;
```

```
proc contents data=mydata.teachers;  
run;
```

```
proc freq data=mydata.teachers;  
tables FirstName LastName Gender Subject;  
run;
```

```
proc univariate data=mydata.teachers plots;  
var Salary Years;  
histogram Salary Years;  
run;
```

```
data mydata.teachers2;  
set mydata.teachers;  
LastName_c = Compress(LastName);  
run;
```

```
data mydata.teachersc;  
set mydata.teachers2;  
Gender_c = UPCASE(Gender);  
/* The teachers' names errors resolve by the  
compression in the last name (i.e. removing  
spaces)*/  
/* There are issues in the spelling of the subjects,  
though. */  
if Subject='Acconting' then Subject='Accounting';  
if Subject='Eglish' then Subject='English';  
if Subject='Mth' then Subject='Math';  
if Employee_ID='E030' then Salary = 73000;  
run;
```

