

ANLT5030 – Unit 9

Assignment 1 Tutorial

SAS Studio



Instructions

- Use methods of descriptive statistics to summarize the data and describe the findings.
- Develop an estimated regression equation using annual income as the independent variable.
- Develop an estimated regression equation using household size as the independent variable.
- Analyze whether annual income or household size is the better predictor of annual credit card charges and provide a rationale.
- Develop an estimated regression equation with annual income and household size as the independent variables and analyze your findings.
- Determine the predicted annual credit card charge for a three-person household with an annual income of \$40,000.
- Assess the need for additional independent variables that could be added to an estimated regression model and the value they would add.

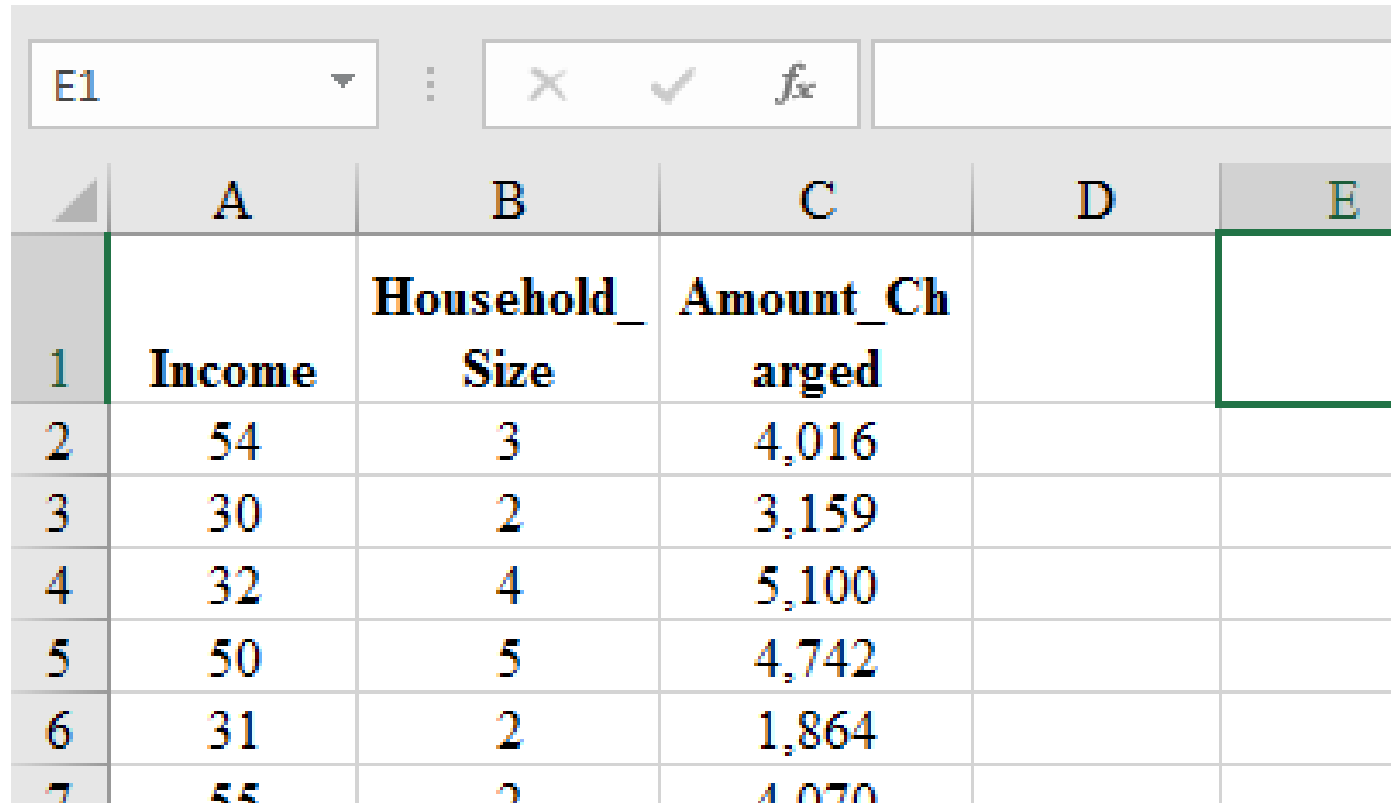


Dataset

- Download the Consumer.xlsx data file from the Cengage datasets zip file, and also posted in the Unit 9 Welcome announcement in the course announcements.



Since SAS cannot accept special characters in the variable names, we have to edit the column/variable names in the spreadsheet first.



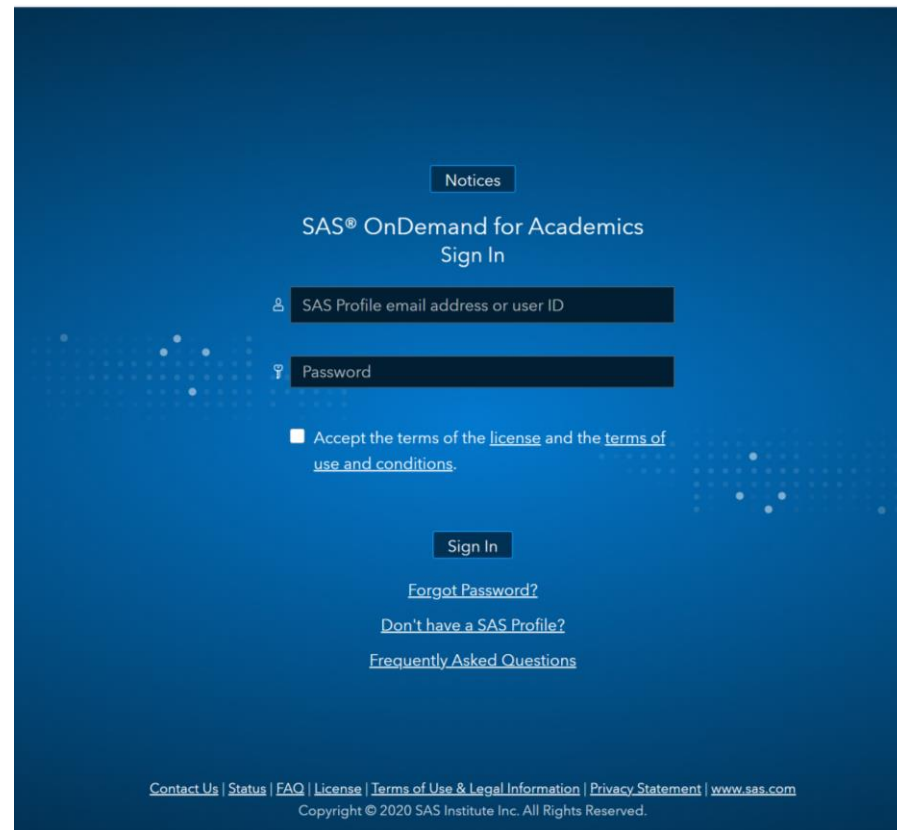
The image shows a spreadsheet interface with a formula bar at the top displaying 'E1'. Below the formula bar is a grid of cells. The first row contains headers: 'Income' in column A, 'Household Size' in column B, and 'Amount Charged' in column C. The subsequent rows contain numerical data. The variable names 'Household Size' and 'Amount Charged' contain spaces, which are not allowed in SAS variable names, necessitating an edit before importing into SAS.

	A	B	C	D	E
1	Income	Household Size	Amount Charged		
2	54	3	4,016		
3	30	2	3,159		
4	32	4	5,100		
5	50	5	4,742		
6	31	2	1,864		
7	55	2	4,070		



Access the SAS OnDemand for Academics Control Center

<https://odamid.oda.sas.com/SASODAControlCenter>



The screenshot shows the SAS OnDemand for Academics Sign In page. The background is a dark blue gradient with a subtle pattern of white dots. At the top, there is a "Notices" button. Below it, the text "SAS® OnDemand for Academics" and "Sign In" are displayed. The sign-in form includes two input fields: "SAS Profile email address or user ID" and "Password". Below the password field, there is a checkbox labeled "Accept the terms of the [license](#) and the [terms of use and conditions](#)." A "Sign In" button is positioned below the checkbox. At the bottom of the form, there are three links: "Forgot Password?", "Don't have a SAS Profile?", and "Frequently Asked Questions". The footer contains a row of links: "Contact Us", "Status", "FAQ", "License", "Terms of Use & Legal Information", "Privacy Statement", and "www.sas.com", followed by the copyright notice "Copyright © 2020 SAS Institute Inc. All Rights Reserved."

Notices

SAS® OnDemand for Academics
Sign In

SAS Profile email address or user ID

Password

☐ Accept the terms of the [license](#) and the [terms of use and conditions](#).

Sign In

[Forgot Password?](#)
[Don't have a SAS Profile?](#)
[Frequently Asked Questions](#)

[Contact Us](#) | [Status](#) | [FAQ](#) | [License](#) | [Terms of Use & Legal Information](#) | [Privacy Statement](#) | [www.sas.com](#)
Copyright © 2020 SAS Institute Inc. All Rights Reserved.



SAS OnDemand for Academics (SODA) Control Center

The screenshot displays the SAS OnDemand for Academics (SODA) Control Center dashboard. The interface features a dark blue header with the SAS logo on the left, and user information (United States, Stefanie Reay) on the right. The main content area is titled "SAS® OnDemand for Academics Dashboard" and includes tabs for "Planned Events" and "Notices". Below these are three main sections: "Applications", "Enrollments", and "Courses". The "Applications" section lists five items: SAS® Studio, SAS® Enterprise Guide®, SAS® Enterprise Miner™, SAS® Forecast Studio, and JMP® Software access to SAS® hosted servers. Each item includes a description and a link to "Actions". The "Reference" section on the right provides links to the Support Site, Step-by-Step Reference Guides, and Frequently Asked Questions. The "Quotas" section shows progress bars for Home Directory (1% of 46.5MB/5120MB) and Course Directory (7% of 207.0MB/3072MB). The footer contains a link to "Other Ways to Access SAS® OnDemand for Academics Resources".

SAS

United States | Stefanie Reay

SAS® OnDemand for Academics Dashboard

Planned Events Notices

Applications Enrollments Courses

SAS® Studio
Write and run SAS code with a Web-based SAS development environment.
Actions: [Clear my saved tabs.](#)

SAS® Enterprise Guide®
Deliver the power of SAS from an easy-to-use, point-and-click interface. ([Download Required](#))

SAS® Enterprise Miner™
Reveal valuable insights with powerful data mining software. ([Configuration Steps Required](#))
Actions: [Clear my project locks.](#)

SAS® Forecast Studio
Generate large numbers of high-quality forecasts automatically. ([Configuration Steps Required](#))
Actions: [Manage your personal environment.](#)

JMP® Software access to SAS® hosted servers
Statistical discovery software. Users must have a copy of JMP® software. ([Configuration Steps Required](#))

Reference

[Support Site](#)
[Step-by-Step Reference Guides](#)
[Frequently Asked Questions](#)

Quotas ([learn more](#))

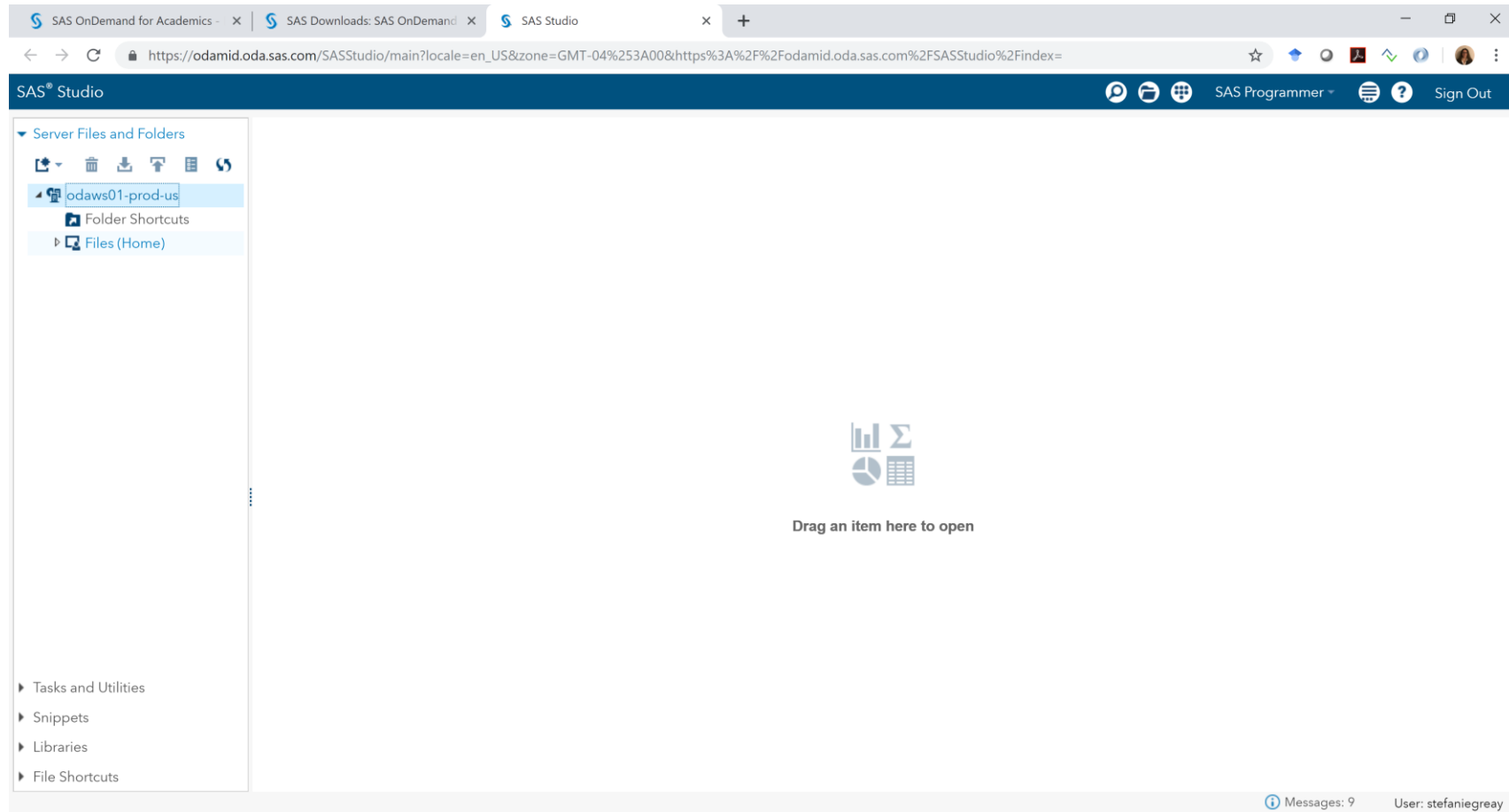
Home Directory (46.5MB/5120MB)
1%

Course Directory (207.0MB/3072MB)
7%

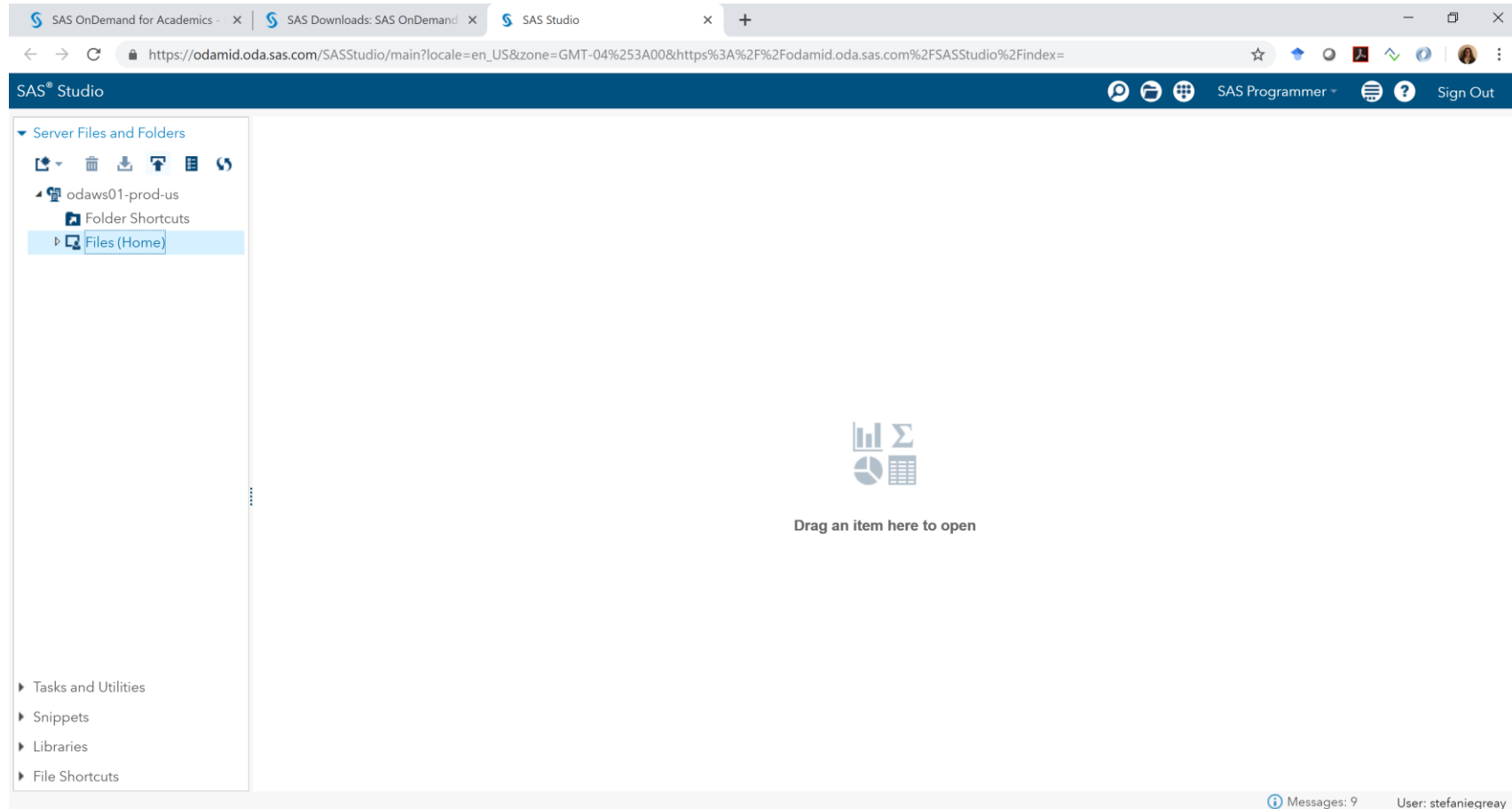
[Other Ways to Access SAS® OnDemand for Academics Resources](#)



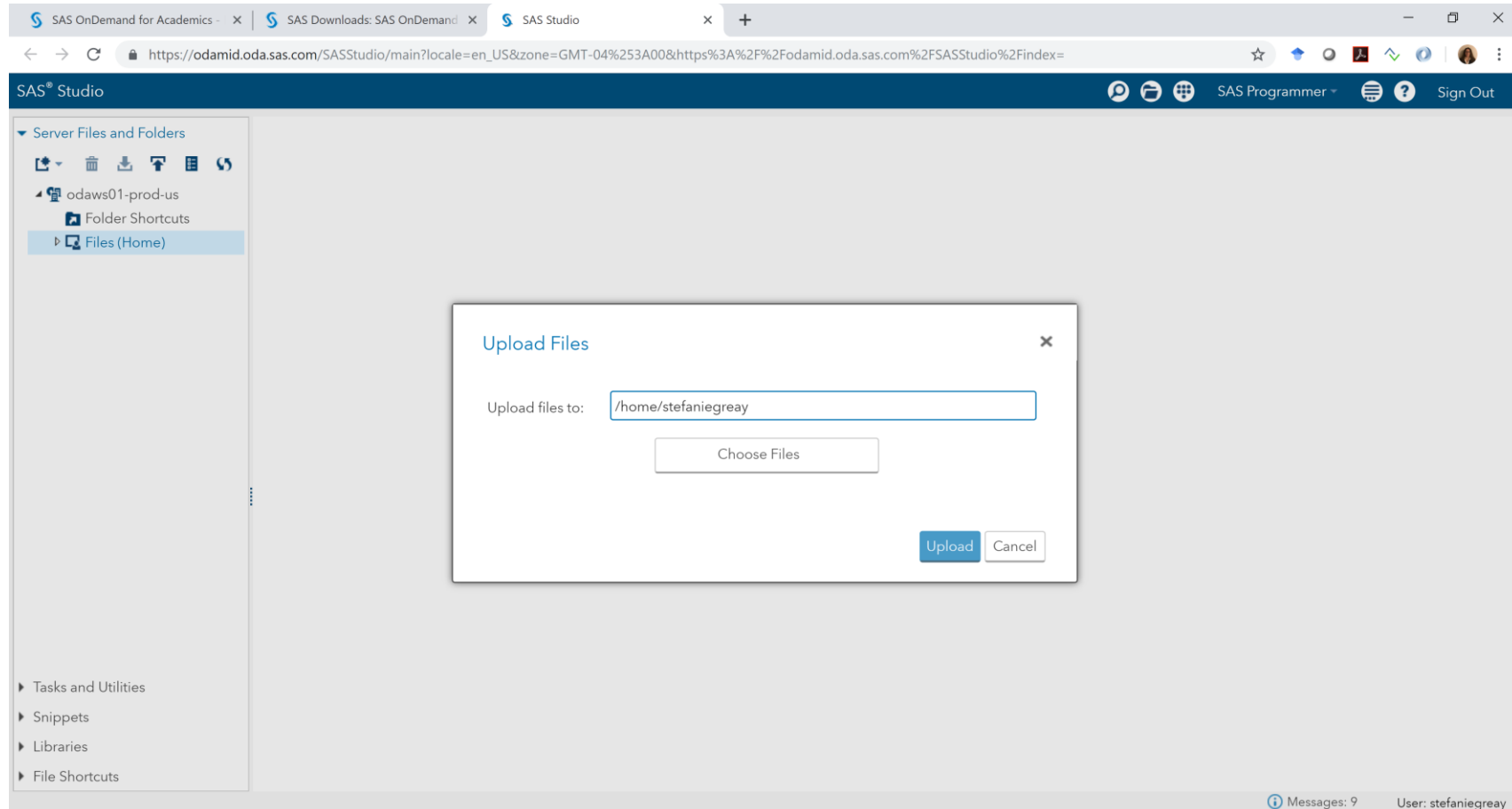
Click on Files(Home)



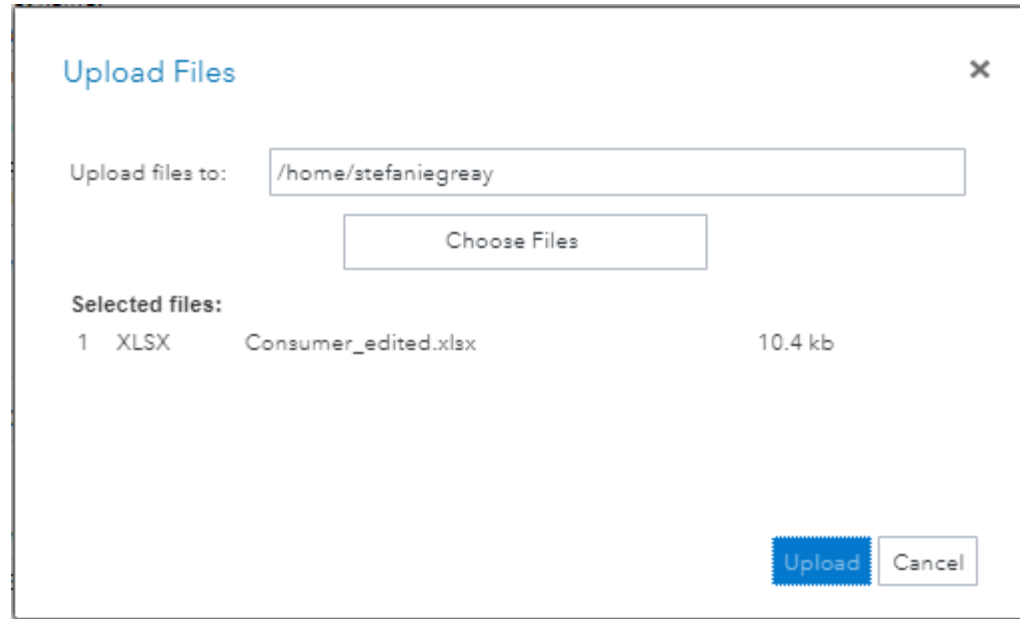
The Upload button will display in dark blue



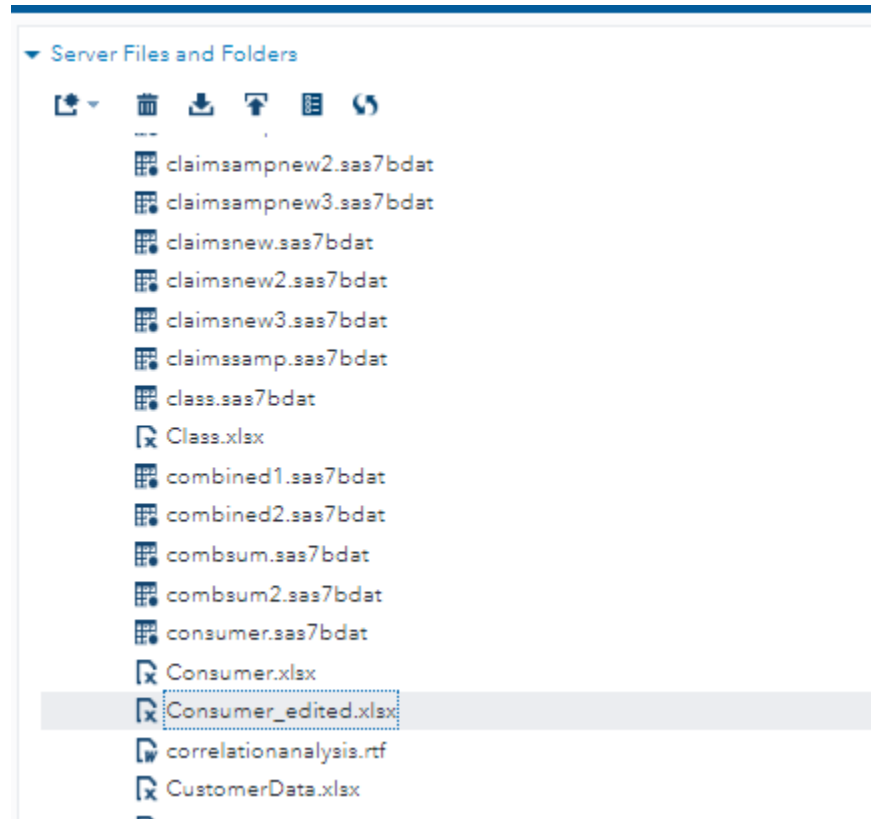
You can create a folder at this point, if you wish, or simply upload to your home directory.



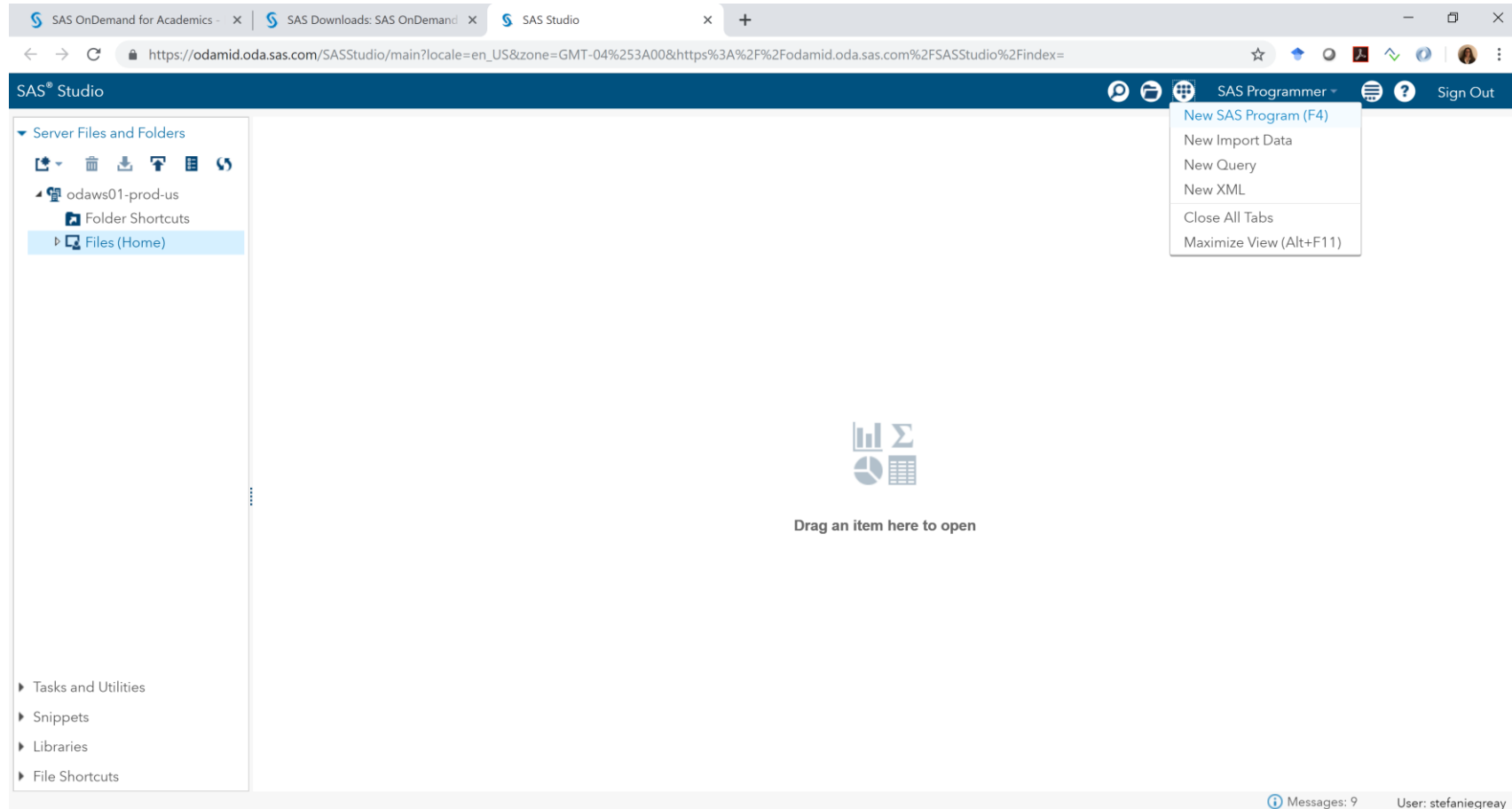
Select “Choose Files” to browse your computer for the dataset you want to upload. Once the dataset has been selected, click “Upload.”



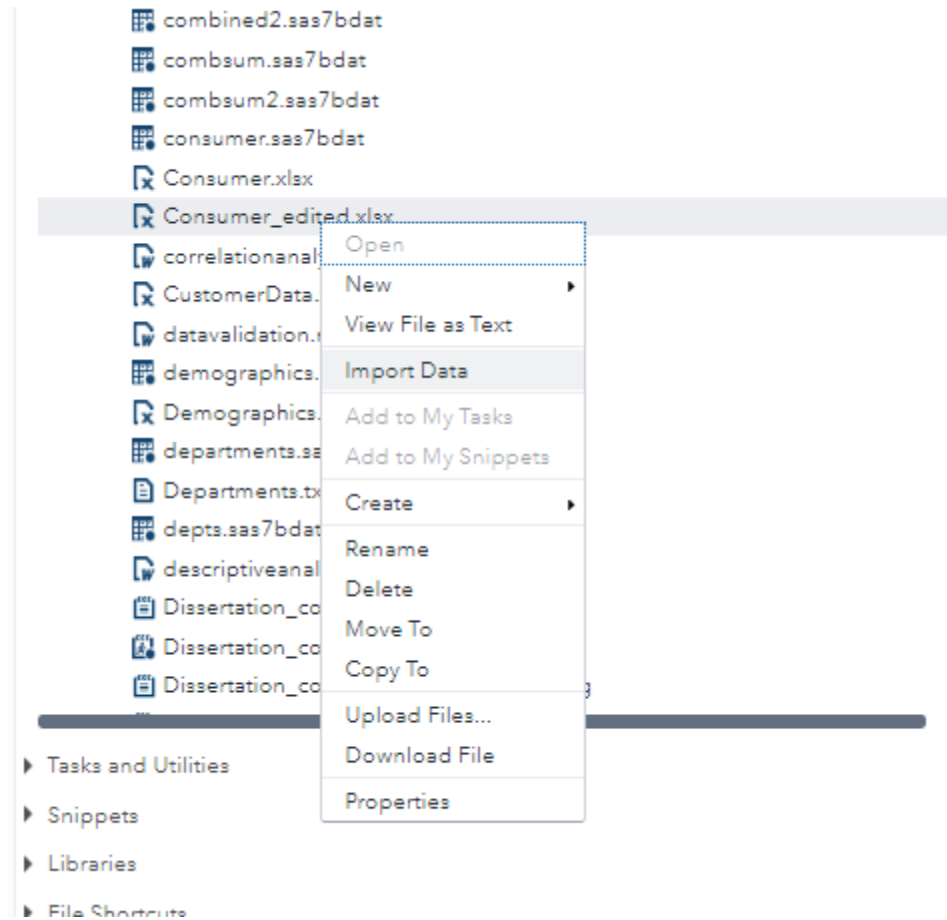
You will be able to view your files by clicking on “Files(Home)” to verify that your file successfully uploaded.



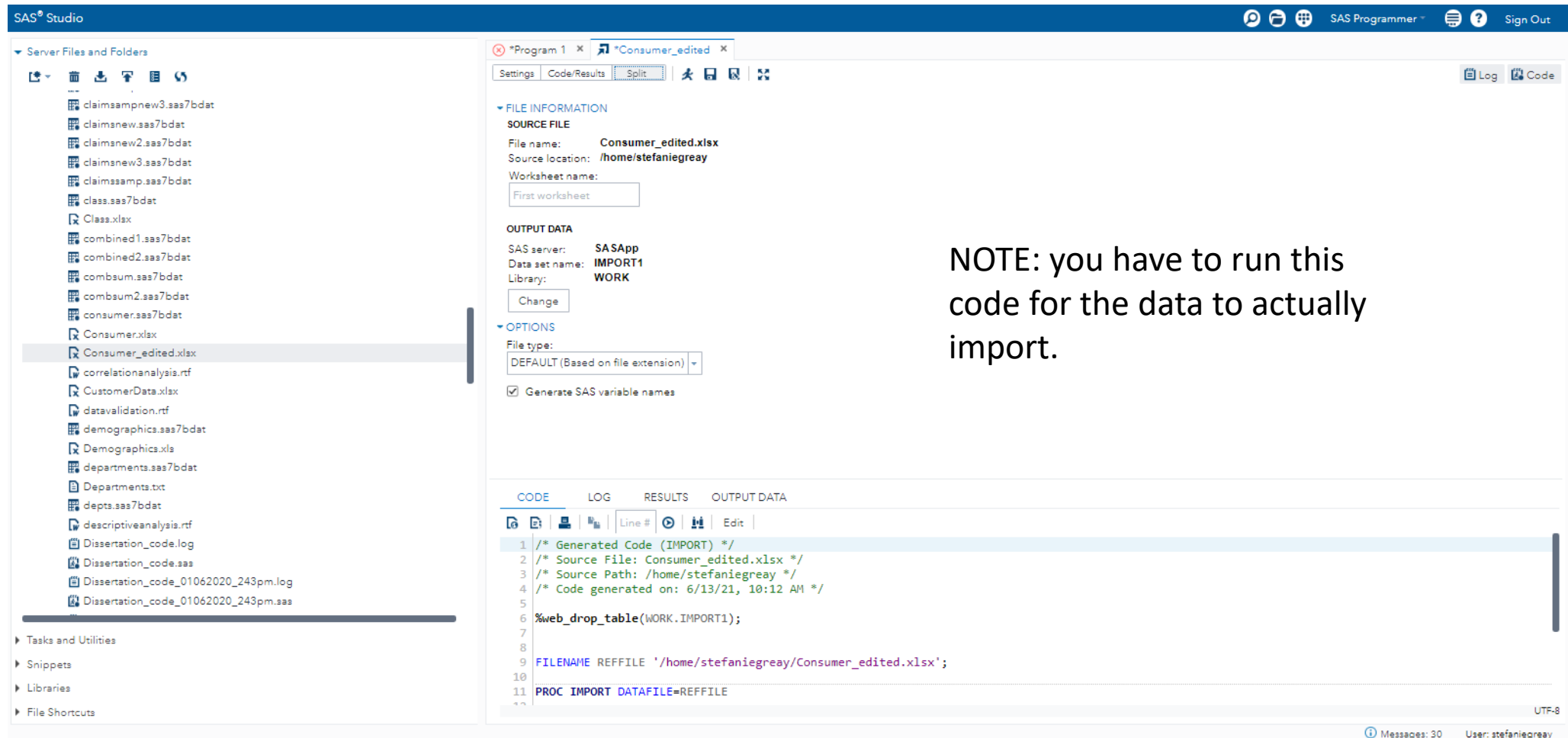
To get started with the SAS portion of the assignment, start a new SAS program.



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!)



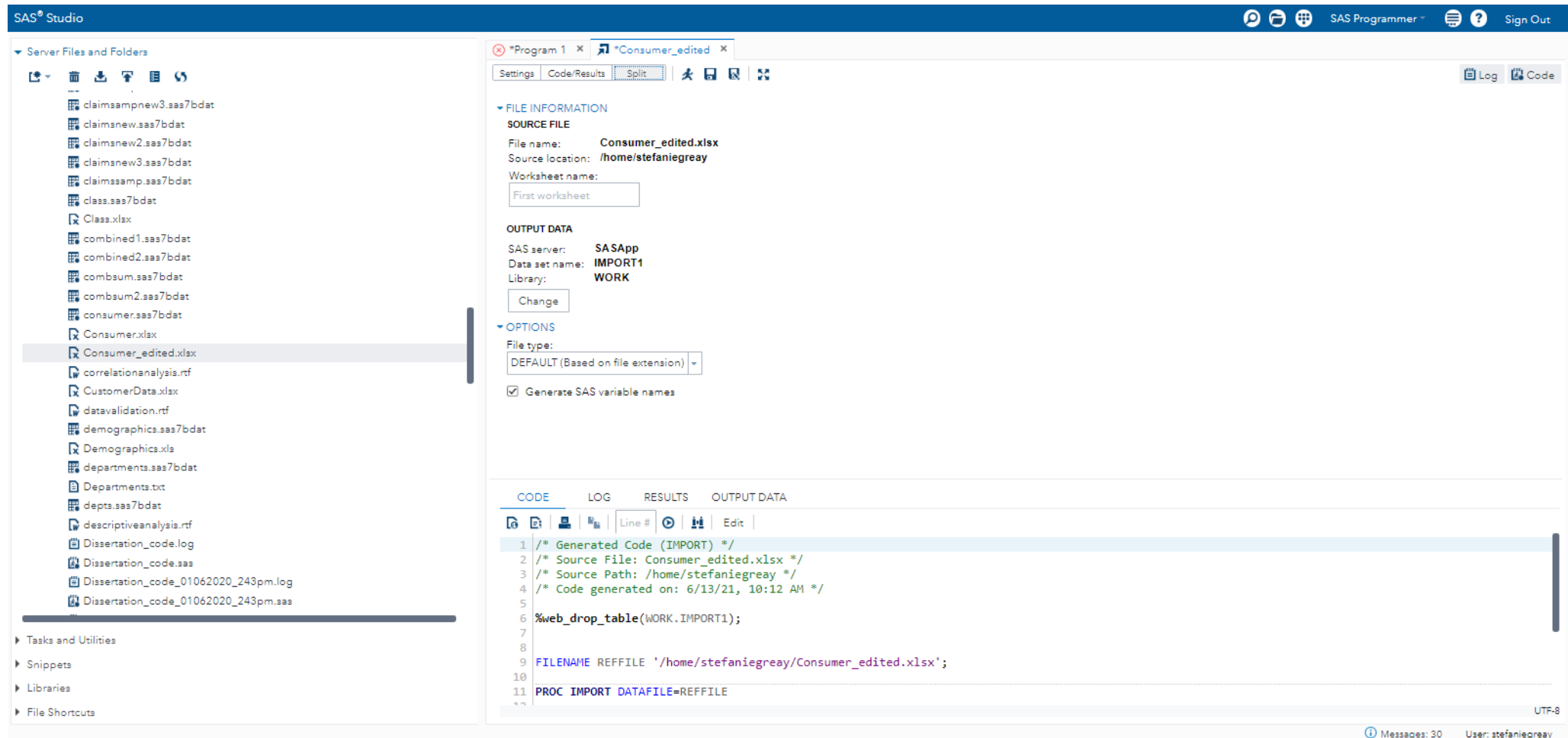
The screenshot displays the SAS Studio interface. On the left, the 'Server Files and Folders' pane shows a list of files, with 'Consumer_edited.xlsx' selected. The main window is divided into two panes. The top pane, titled 'FILE INFORMATION', shows the source file 'Consumer_edited.xlsx' and the output data 'IMPORT1'. The bottom pane, titled 'CODE', shows the generated SAS code for importing the file. The code includes comments about the source file and path, and the actual SAS code for the import process.

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

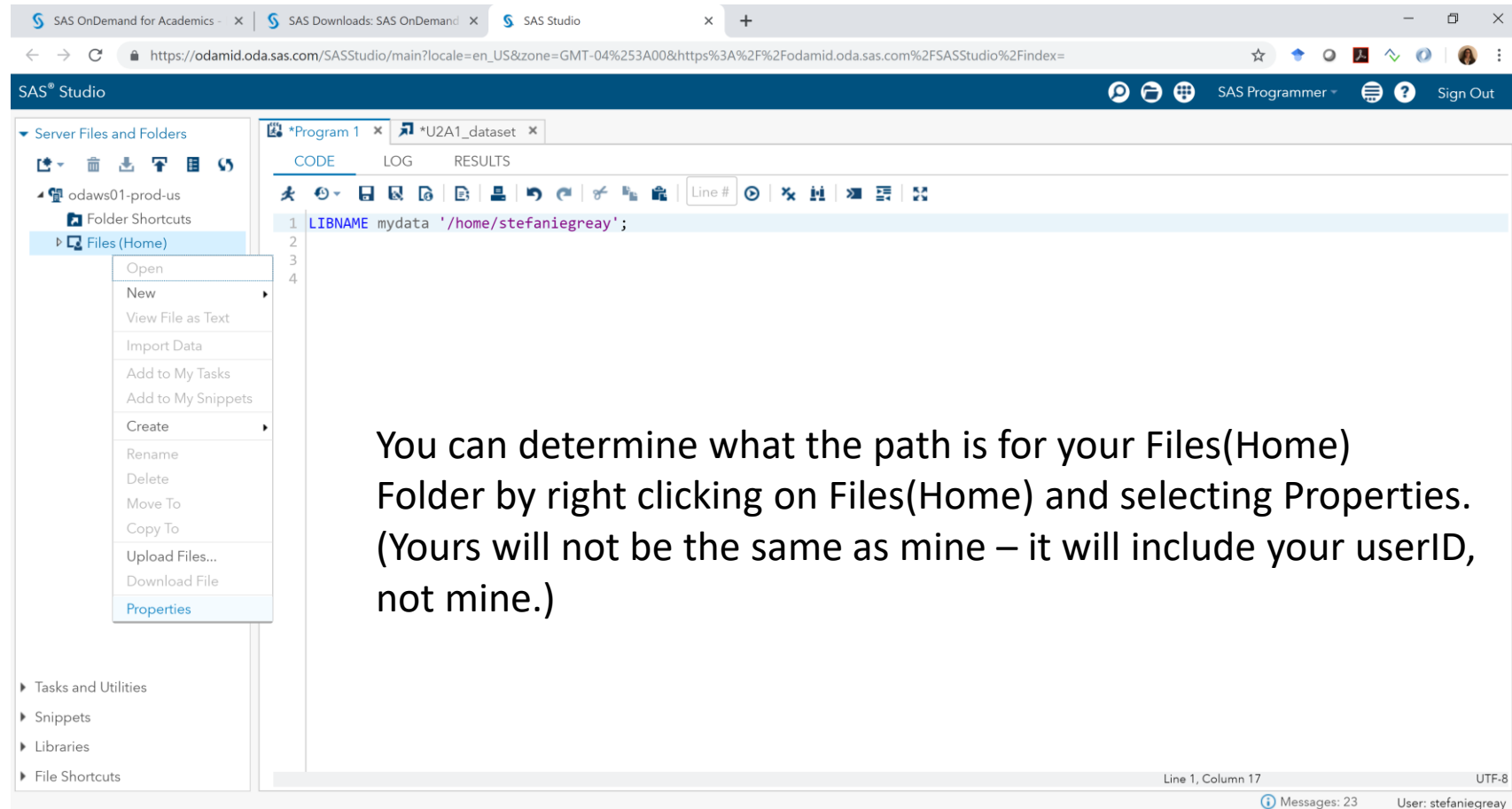
```
1 /* Generated Code (IMPORT) */
2 /* Source File: Consumer_edited.xlsx */
3 /* Source Path: /home/stefaniegreay */
4 /* Code generated on: 6/13/21, 10:12 AM */
5
6 %web_drop_table(WORK.IMPORT1);
7
8
9 FILENAME REFFILE '/home/stefaniegreay/Consumer_edited.xlsx';
10
11 PROC IMPORT DATAFILE=REFFILE
```



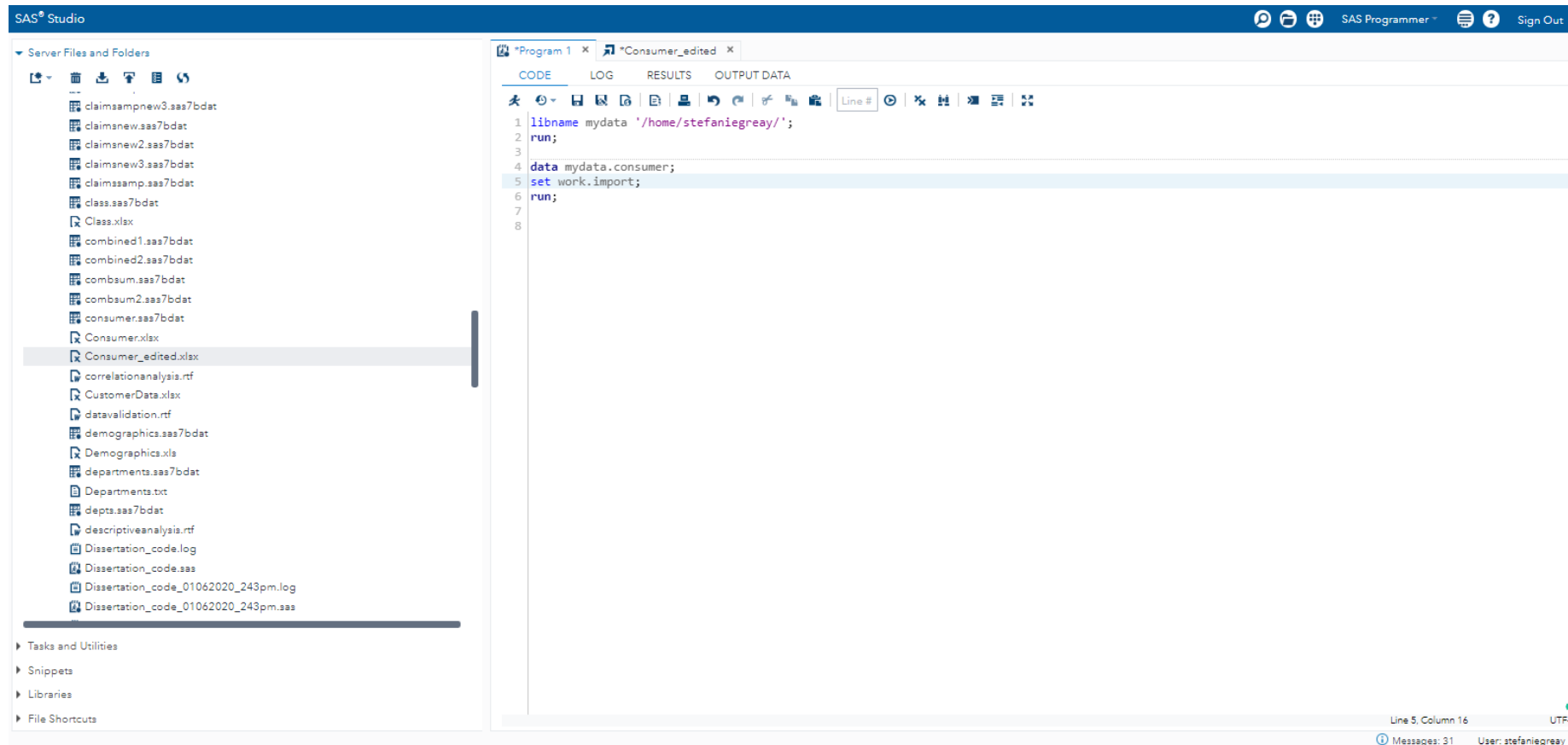
To run the code, click the icon that looks like a guy running.



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



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



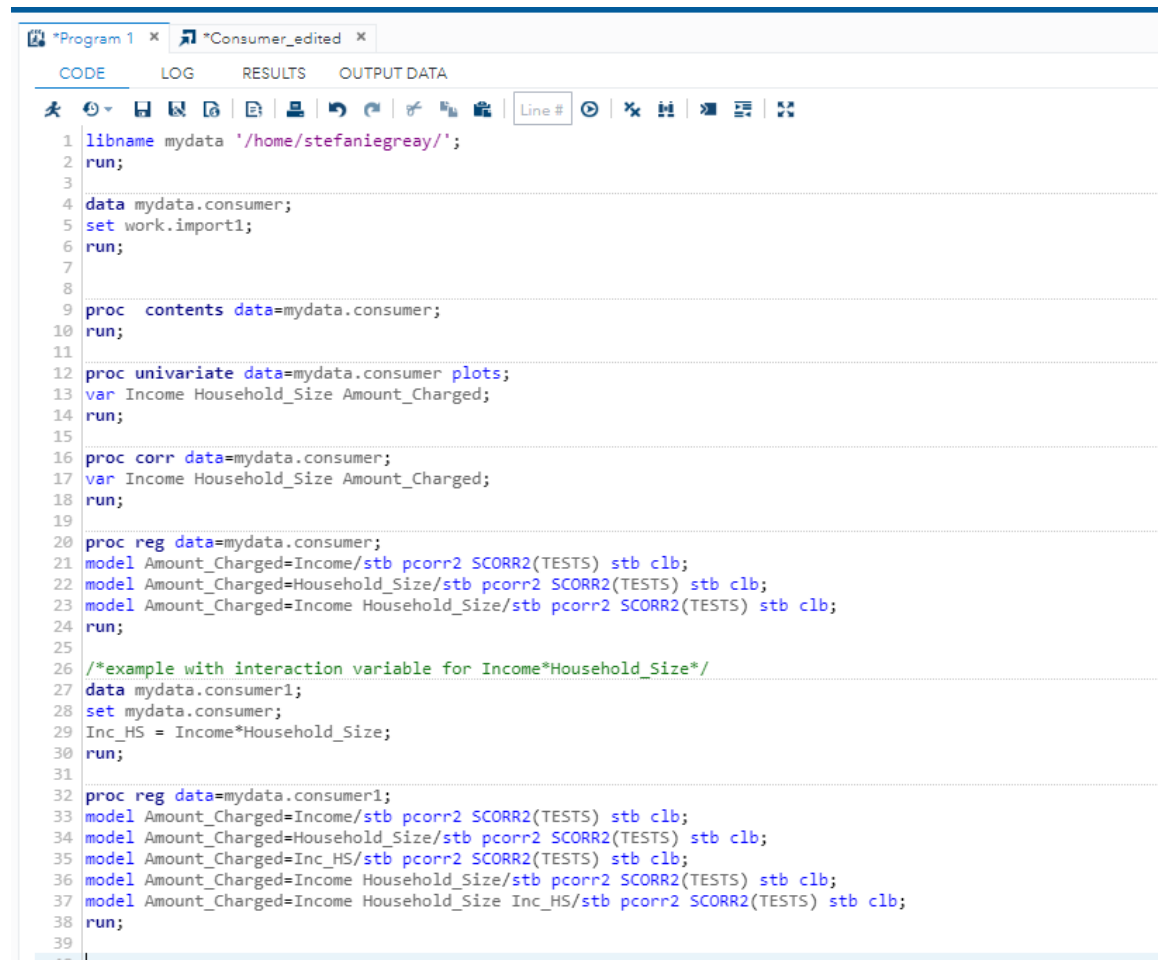
When you run the code, you will see the dataset in the output data window and can verify its success.

The screenshot displays the SAS Studio interface. On the left, the 'Server Files and Folders' pane lists various files, with 'Consumer_edited.xlsx' highlighted. The main window is divided into four tabs: 'CODE', 'LOG', 'RESULTS', and 'OUTPUT DATA'. The 'OUTPUT DATA' tab is active, showing a table named 'MYDATA.CONSUMER'. The table has three columns: 'Income', 'Household_Size', and 'Amount_Charged'. The table contains 25 rows of data. Below the table, there is a 'Property Value' section with fields for Label, Name, Length, Type, Format, and Informat.

	Income	Household_Size	Amount_Charged
1	54	3	4,016
2	30	2	3,159
3	32	4	5,100
4	50	5	4,742
5	31	2	1,864
6	55	2	4,070
7	37	1	2,731
8	40	2	3,348
9	66	4	4,764
10	51	3	4,110
11	25	3	4,208
12	48	4	4,219
13	27	1	2,477
14	33	2	2,514
15	65	3	4,214
16	63	4	4,965
17	42	6	4,412
18	21	2	2,448
19	44	1	2,995
20	37	5	4,171
21	62	6	5,678
22	21	3	3,623
23	55	7	5,301
24	42	2	3,020
25	41	7	4,698



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



```
*Program 1 x *Consumer_edited x
CODE LOG RESULTS OUTPUT DATA
1 libname mydata '/home/stefaniegreay/';
2 run;
3
4 data mydata.consumer;
5 set work.import1;
6 run;
7
8
9 proc contents data=mydata.consumer;
10 run;
11
12 proc univariate data=mydata.consumer plots;
13 var Income Household_Size Amount_Charged;
14 run;
15
16 proc corr data=mydata.consumer;
17 var Income Household_Size Amount_Charged;
18 run;
19
20 proc reg data=mydata.consumer;
21 model Amount_Charged=Income/stb pcorr2 SCORR2(TESTS) stb clb;
22 model Amount_Charged=Household_Size/stb pcorr2 SCORR2(TESTS) stb clb;
23 model Amount_Charged=Income Household_Size/stb pcorr2 SCORR2(TESTS) stb clb;
24 run;
25
26 /*example with interaction variable for Income*Household_Size*/
27 data mydata.consumer1;
28 set mydata.consumer;
29 Inc_HS = Income*Household_Size;
30 run;
31
32 proc reg data=mydata.consumer1;
33 model Amount_Charged=Income/stb pcorr2 SCORR2(TESTS) stb clb;
34 model Amount_Charged=Household_Size/stb pcorr2 SCORR2(TESTS) stb clb;
35 model Amount_Charged=Inc_HS/stb pcorr2 SCORR2(TESTS) stb clb;
36 model Amount_Charged=Income Household_Size/stb pcorr2 SCORR2(TESTS) stb clb;
37 model Amount_Charged=Income Household_Size Inc_HS/stb pcorr2 SCORR2(TESTS) stb clb;
38 run;
39
40
```



Sample Code for the assignment

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

```
data mydata.consumer;  
set work.import1;  
run;
```

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

```
proc univariate data=mydata.consumer plots;  
var Income Household_Size Amount_Charged;  
run;
```

```
proc corr data=mydata.consumer;  
var Income Household_Size Amount_Charged;  
run;
```

```
proc reg data=mydata.consumer;  
model Amount_Charged=Income/stb pcorr2 SCORR2(TESTS) stb clb;  
model Amount_Charged=Household_Size/stb pcorr2 SCORR2(TESTS) stb clb;  
model Amount_Charged=Income Household_Size/stb pcorr2 SCORR2(TESTS) stb clb;
```

Note that is just sample code...you will need to check the assumptions and significance of contributions, and select the “best” model.



Sample Code with Interaction Variable

```
/*example with interaction variable for Income*Household_Size*/  
data mydata.consumer1;  
set mydata.consumer;  
Inc_HS = Income*Household_Size;  
run;  
  
proc reg data=mydata.consumer1;  
model Amount_Charged=Income/stb pcorr2 SCORR2(TESTS) stb clb;  
model Amount_Charged=Household_Size/stb pcorr2 SCORR2(TESTS) stb clb;  
model Amount_Charged=Inc_HS/stb pcorr2 SCORR2(TESTS) stb clb;  
model Amount_Charged=Income Household_Size/stb pcorr2 SCORR2(TESTS) stb clb;  
model Amount_Charged=Income Household_Size Inc_HS/stb pcorr2 SCORR2(TESTS) stb clb;  
run;
```

An interaction variable captures the contribution to predicting the response variable that may be due to the interaction between the selected explanatory variables. These are sometimes significant and sometimes not, but are helpful in understanding the contribution of each variable individually vs the combination of the variables, especially in a simple case with only two independent/explanatory variables.



Additional Resources for Multiple Linear Regression and SAS's Proc Reg

- SAS's Proc Reg documentation:

https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.3/statug/statug_reg_toc.htm

- Institute for Digital Research and Education Statistical Consulting Proc Reg annotated output:

<https://stats.idre.ucla.edu/sas/output/regression-analysis/>



Assumptions for Multiple Linear Regression

Assumptions

- 1) Linear, additive relationship between the explanatory/independent variables and the response/dependent variable.
- 2) Multivariate normality (the error terms are assumed to be normally distributed)
- 3) No multicollinearity (no linear relationships between the explanatory/independent variables)
- 4) Heteroscedasticity or homogeneity of variance (variance of errors are distributed equally across the values of each of the independent variables and dependent variables)

