ANLT5030 – Unit 9 Assignment 1 Tutorial

SAS Studio

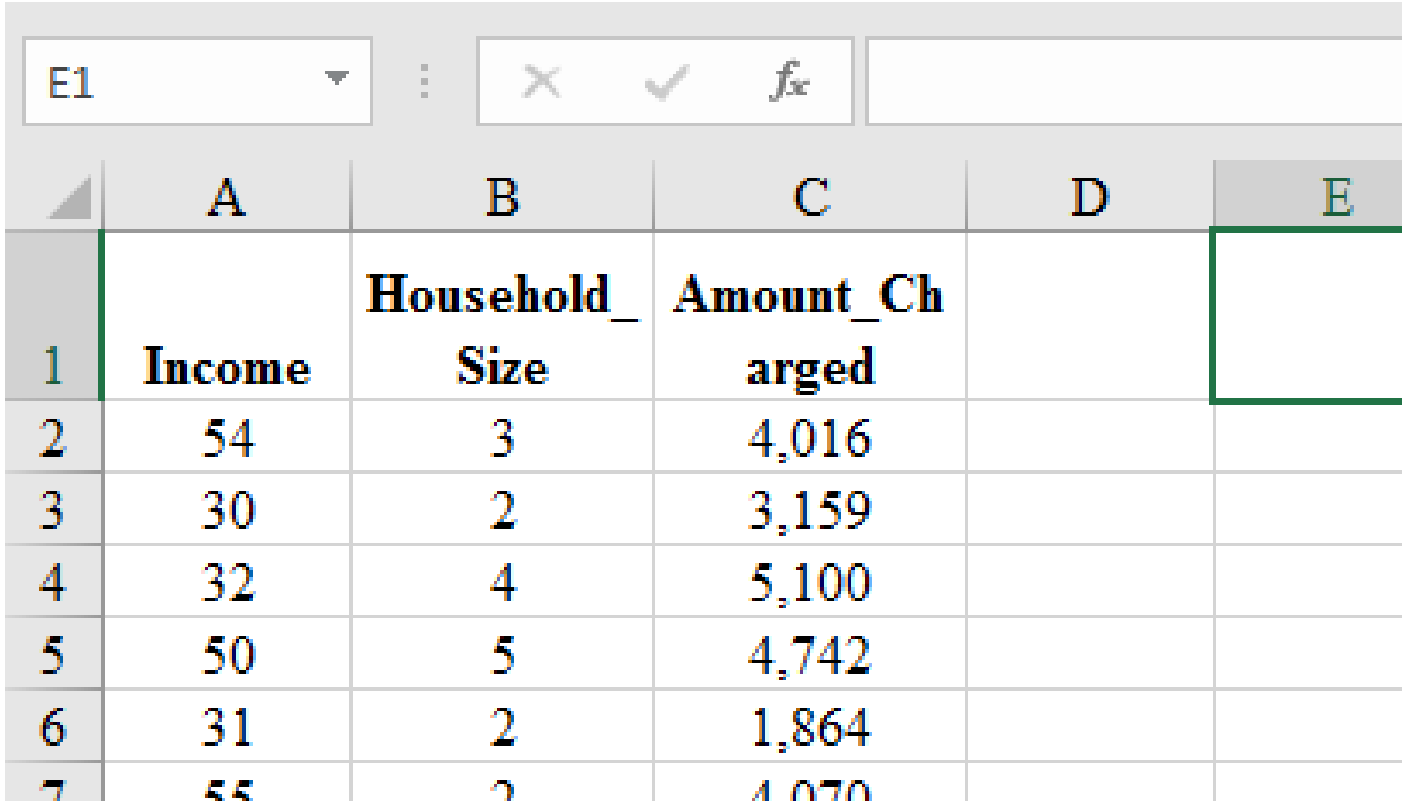
Instructions

- 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

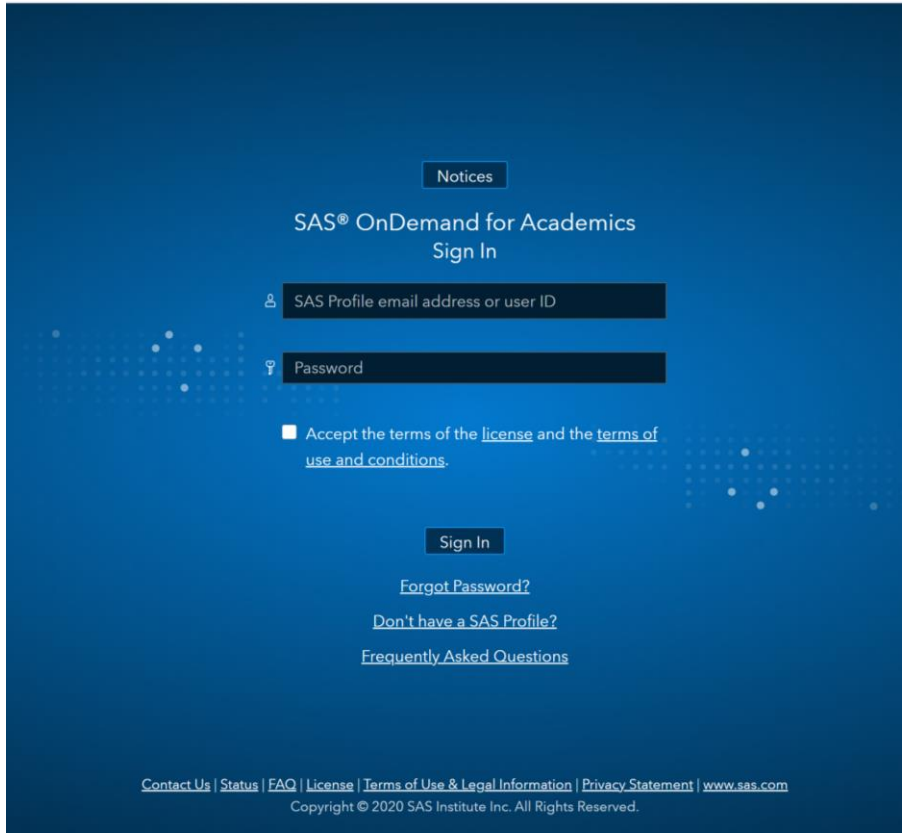
- Download the Consumer.xlsx data file from the Cengage datasets zip file, and also posted in the Unit 9 Welcome announcement in the course announcements.

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

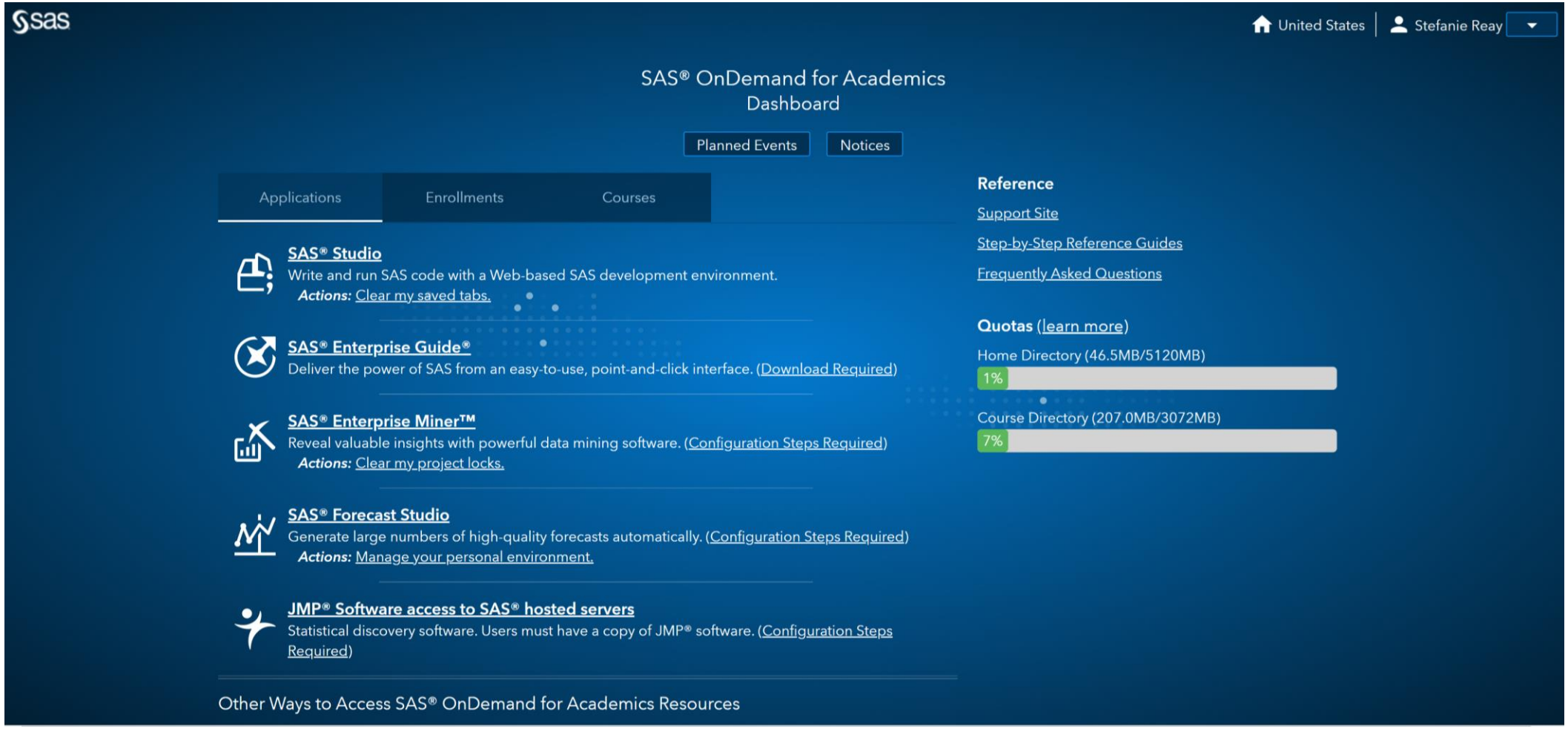


Access the SAS OnDemand for Academics Control Center

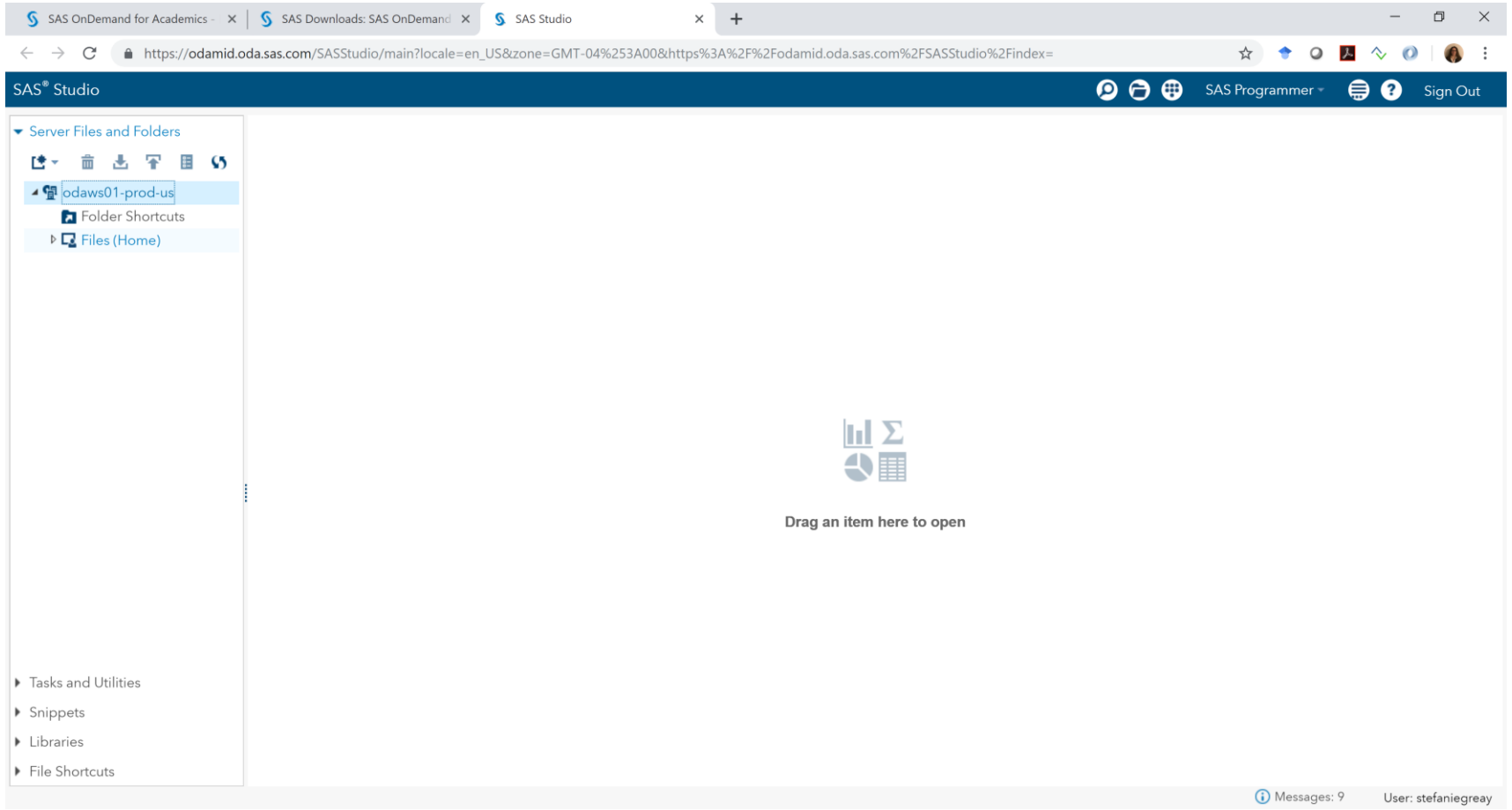
https://odamid.oda.sas.com/SASODAControlCenter



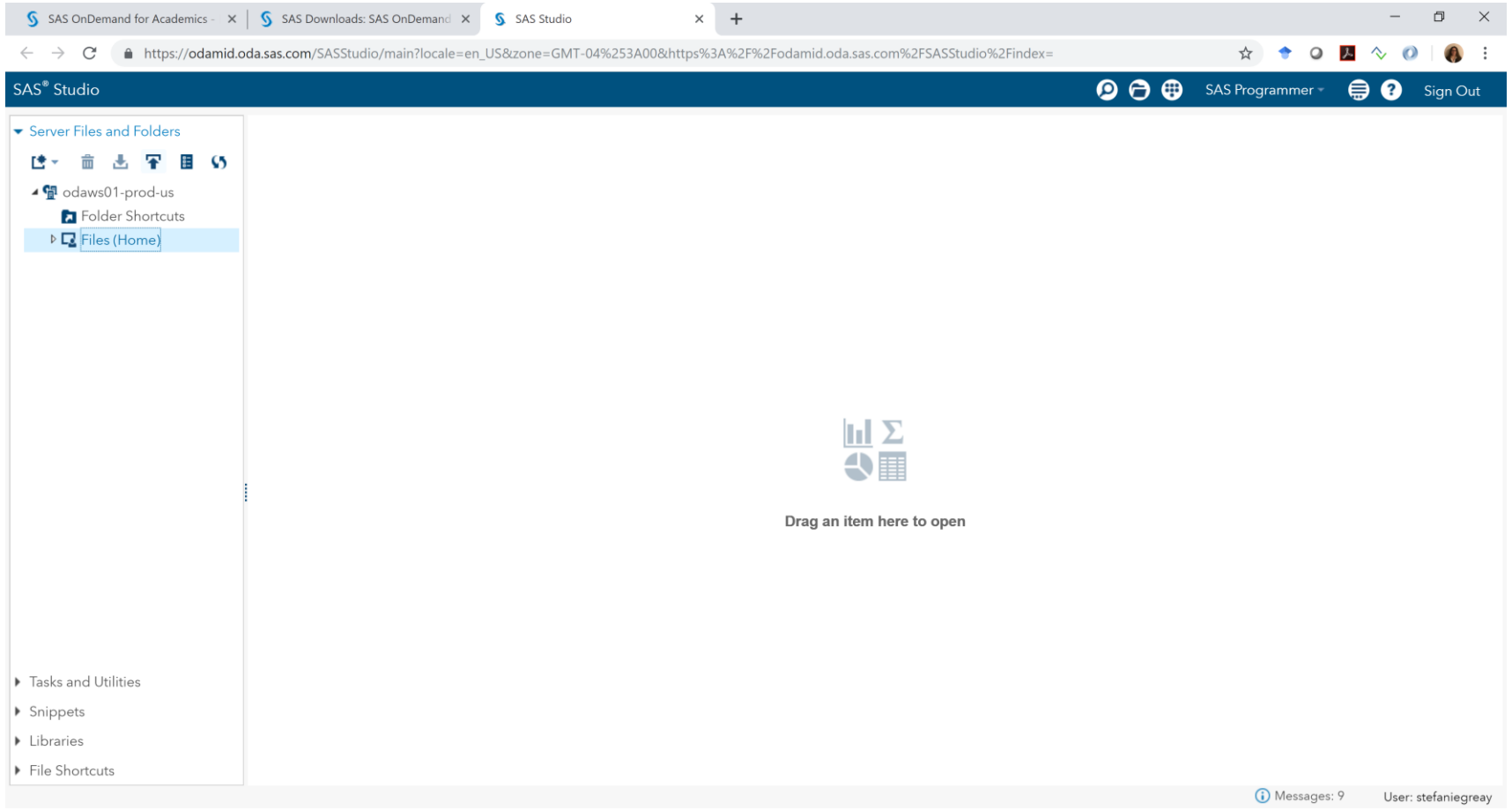
SAS OnDemand for Academics (SODA) Control Center



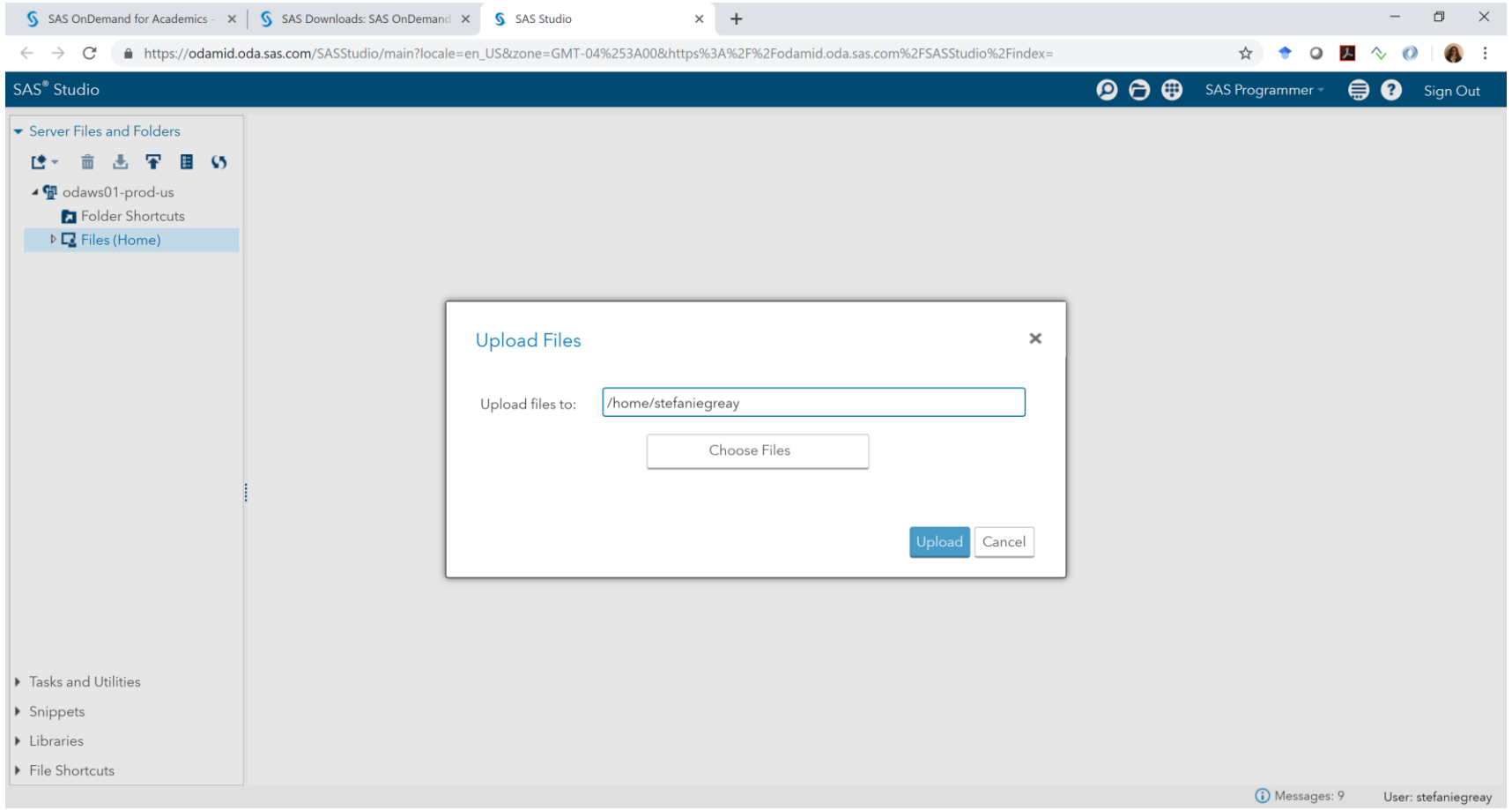
Click on Files(Home)



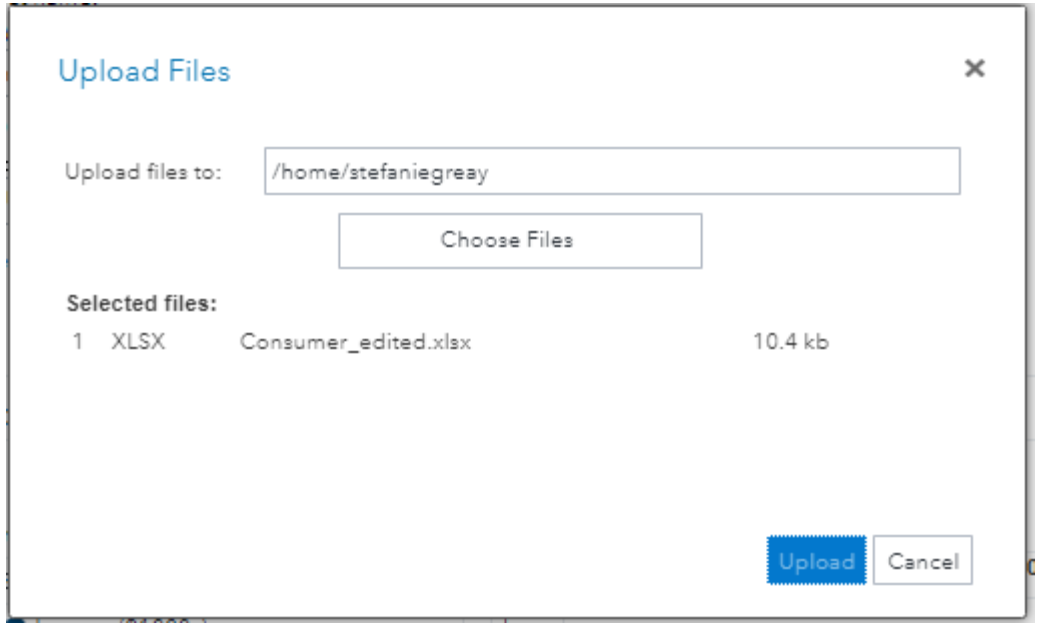
The Upload button will display in dark blue



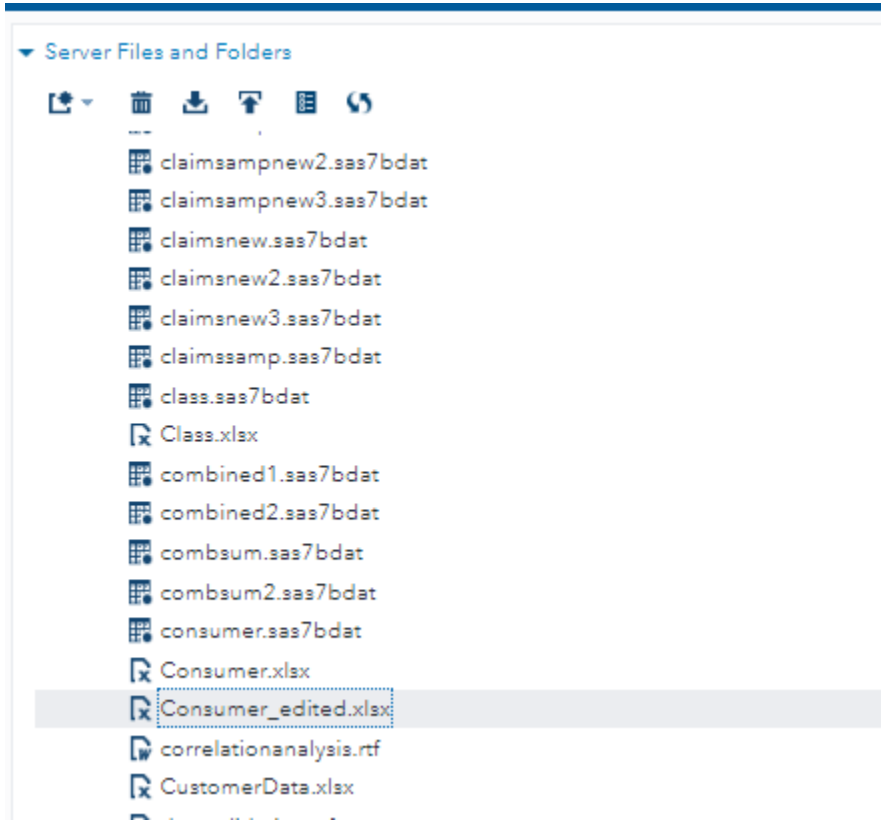
You can create a folder at this point, if you wish, or simply upload to your home directory.



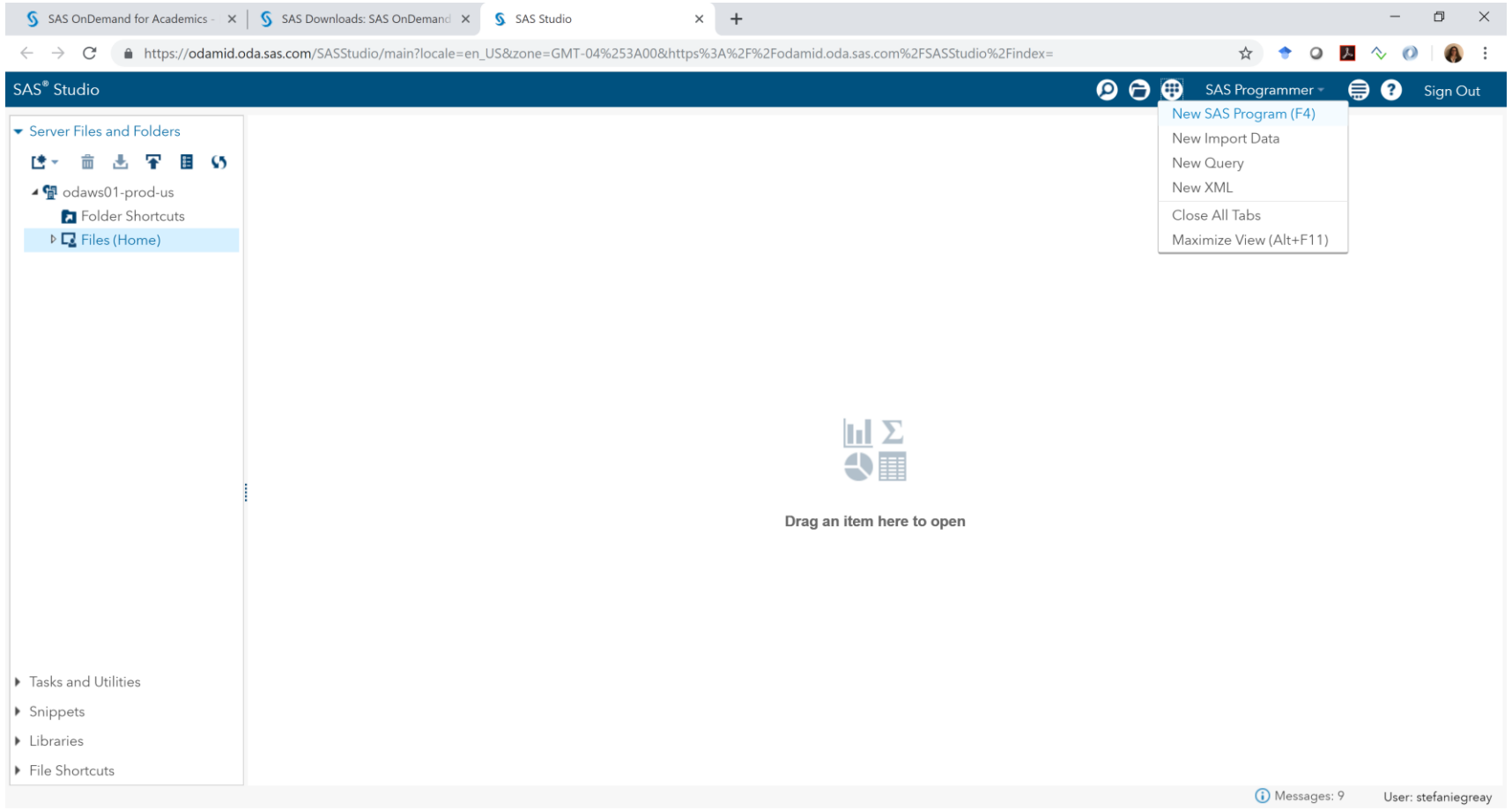
Select “Choose Files” to browse your computer for the dataset you want to upload. Once the dataset has been selected, click “Upload.”



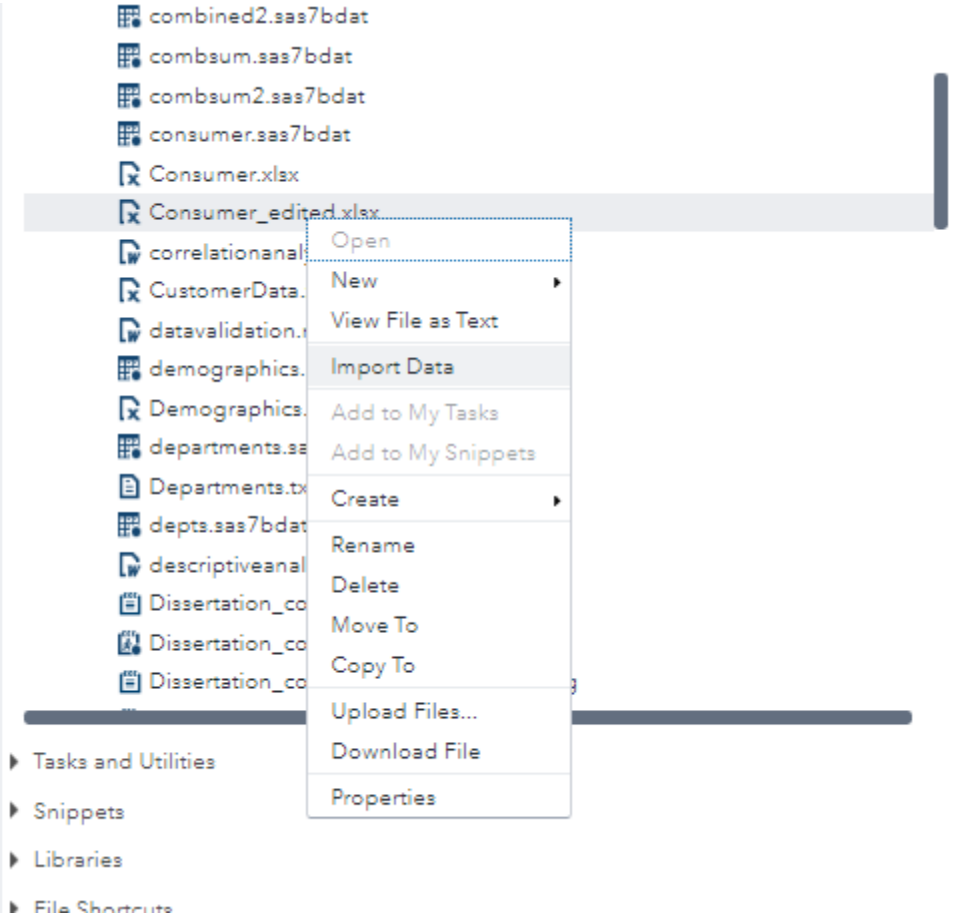
You will be able to view your files by clicking on “Files(Home)” to verify that your file successfully uploaded.



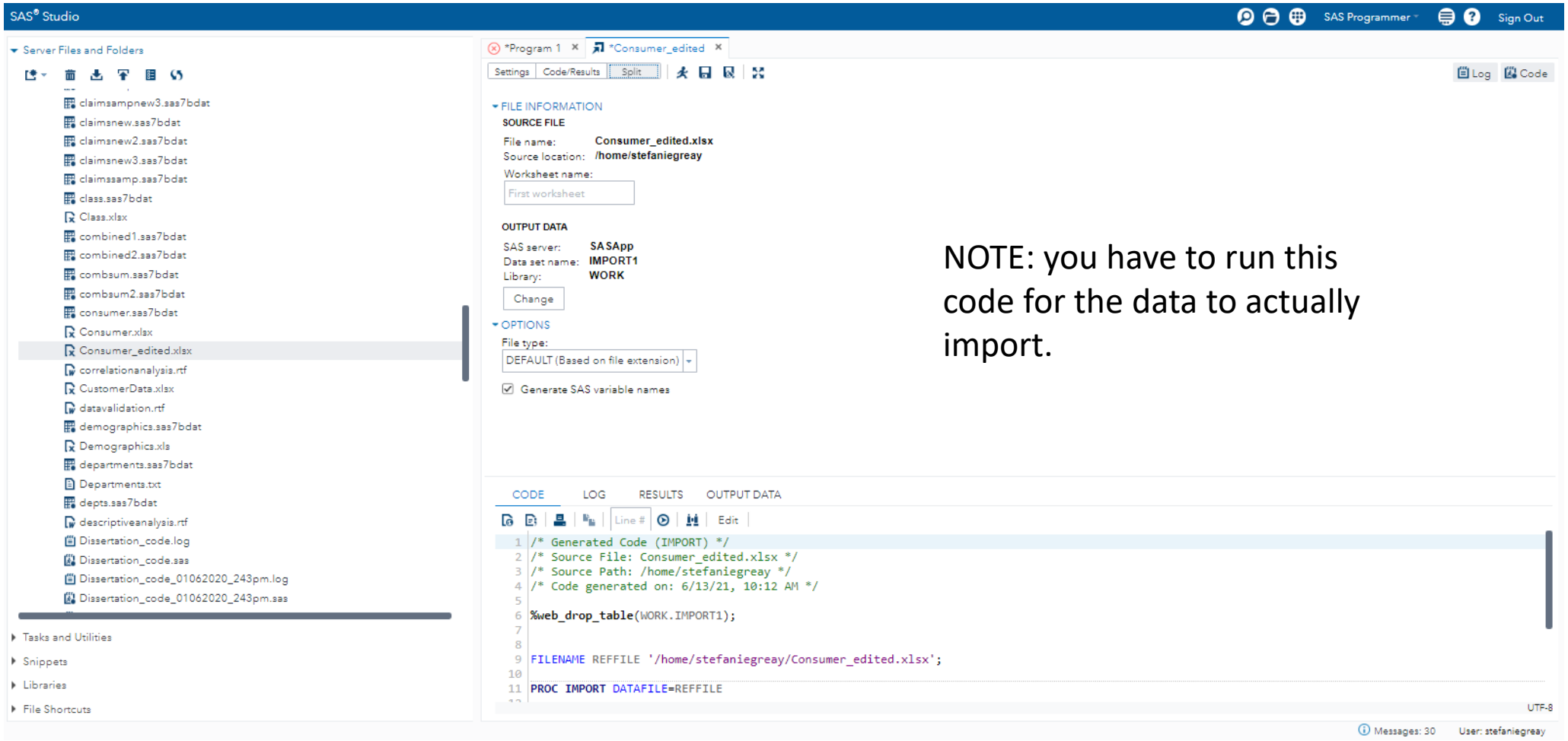
To get started with the SAS portion of the assignment, start a new SAS program.



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

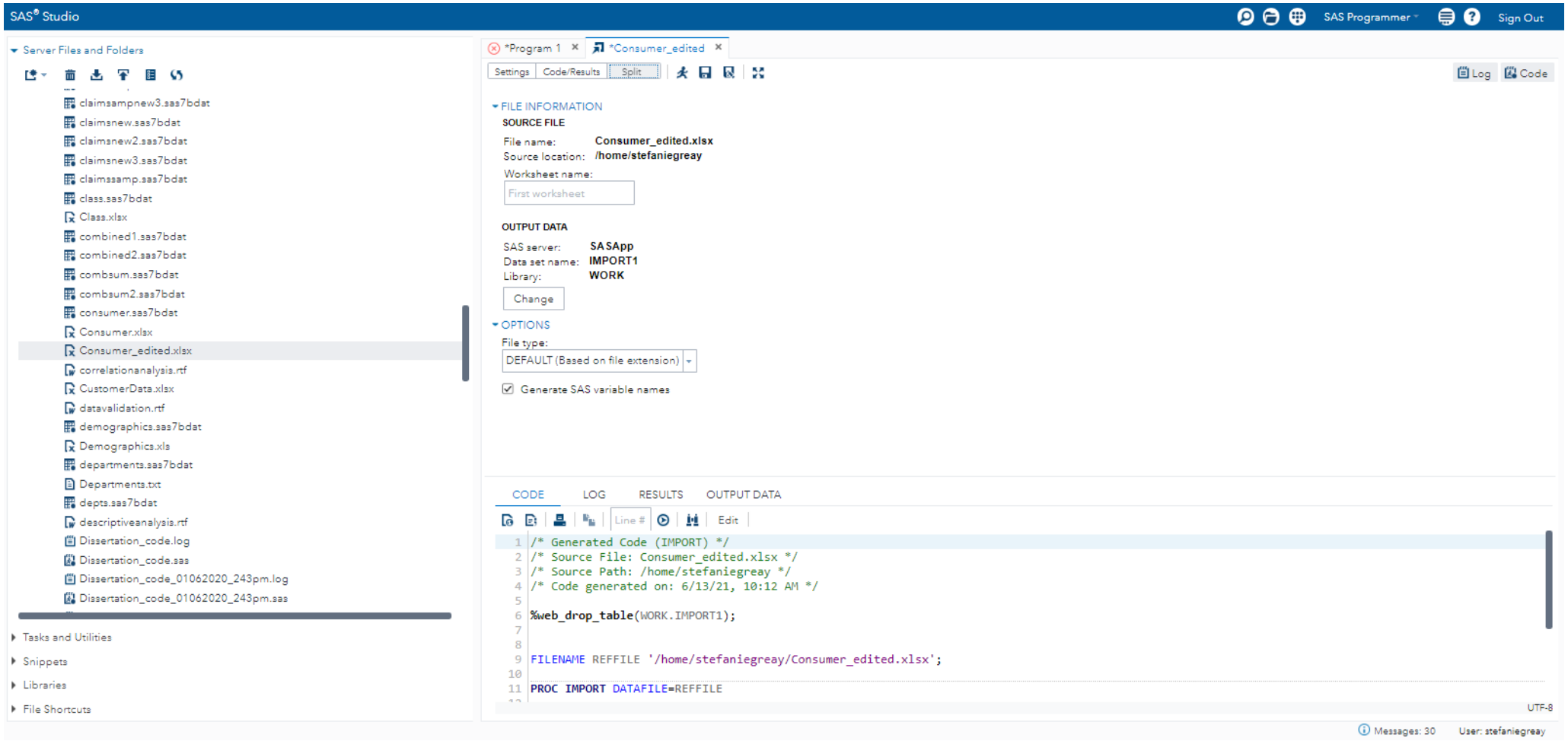


The Proc Import code will be written for you (save this as a template to use for future imports!)

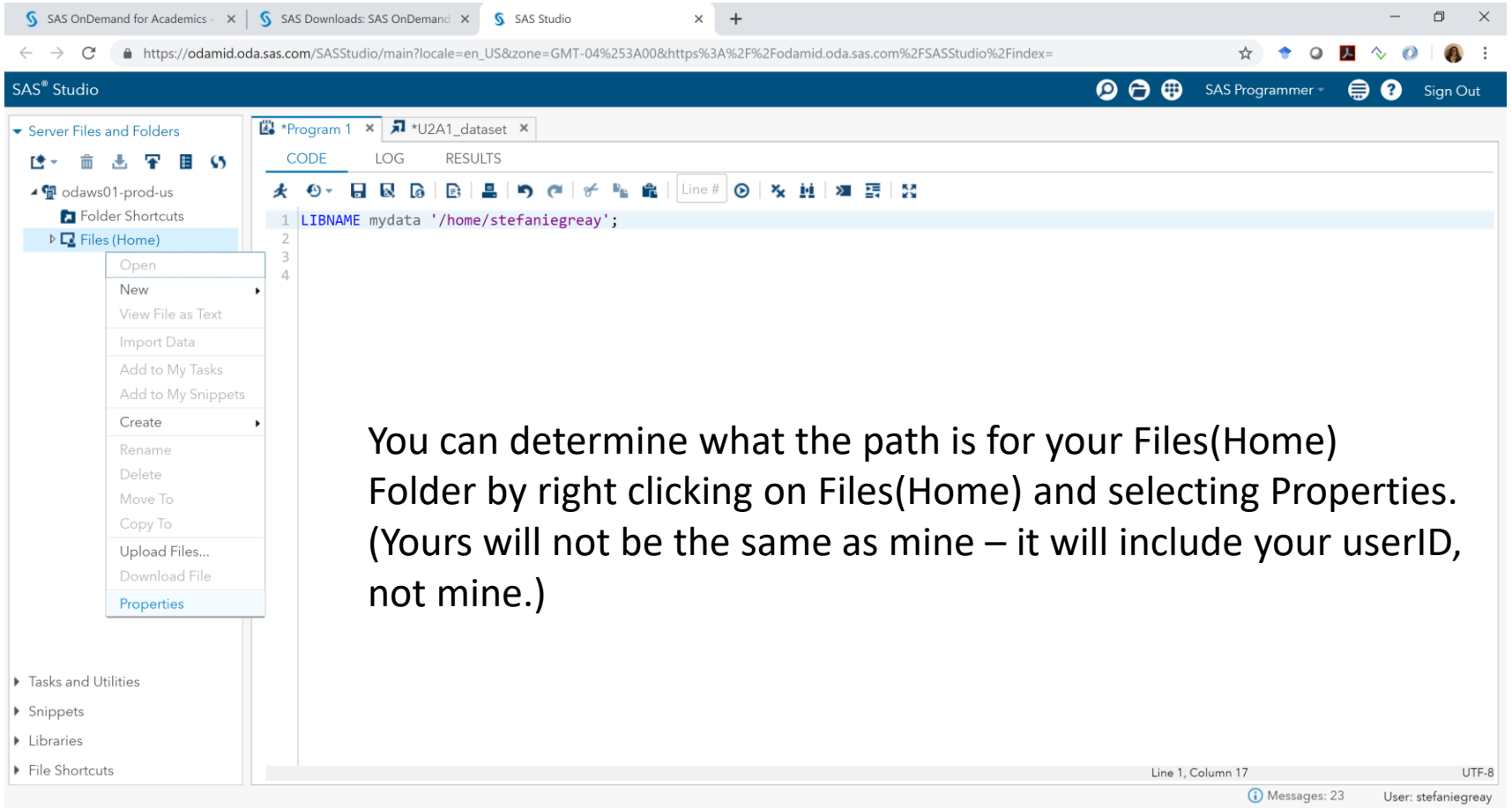


NOTE: you have to run this code for the data to actually import.

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

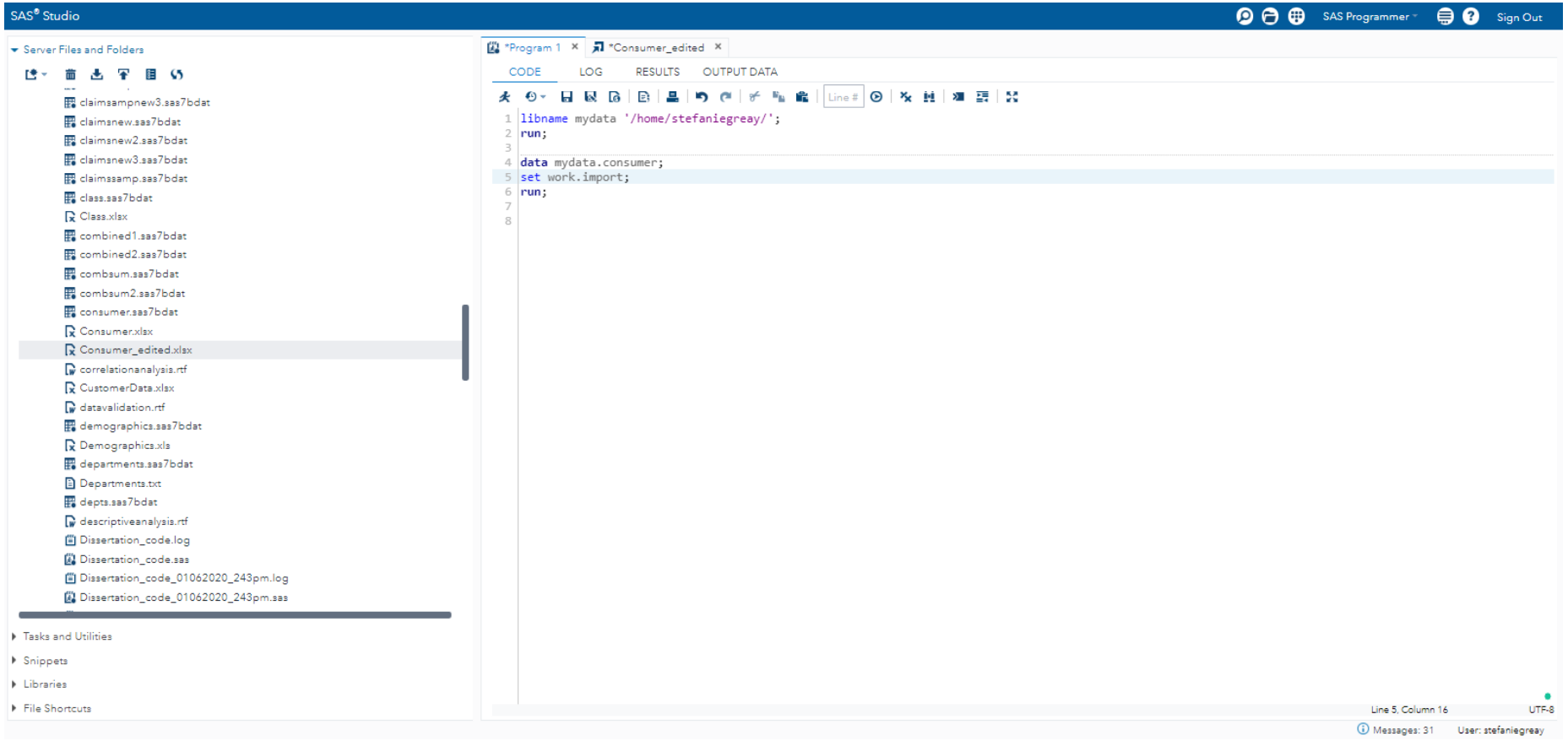


To create a SAS Library for your Files(Home) folder, you need to use a libname statement

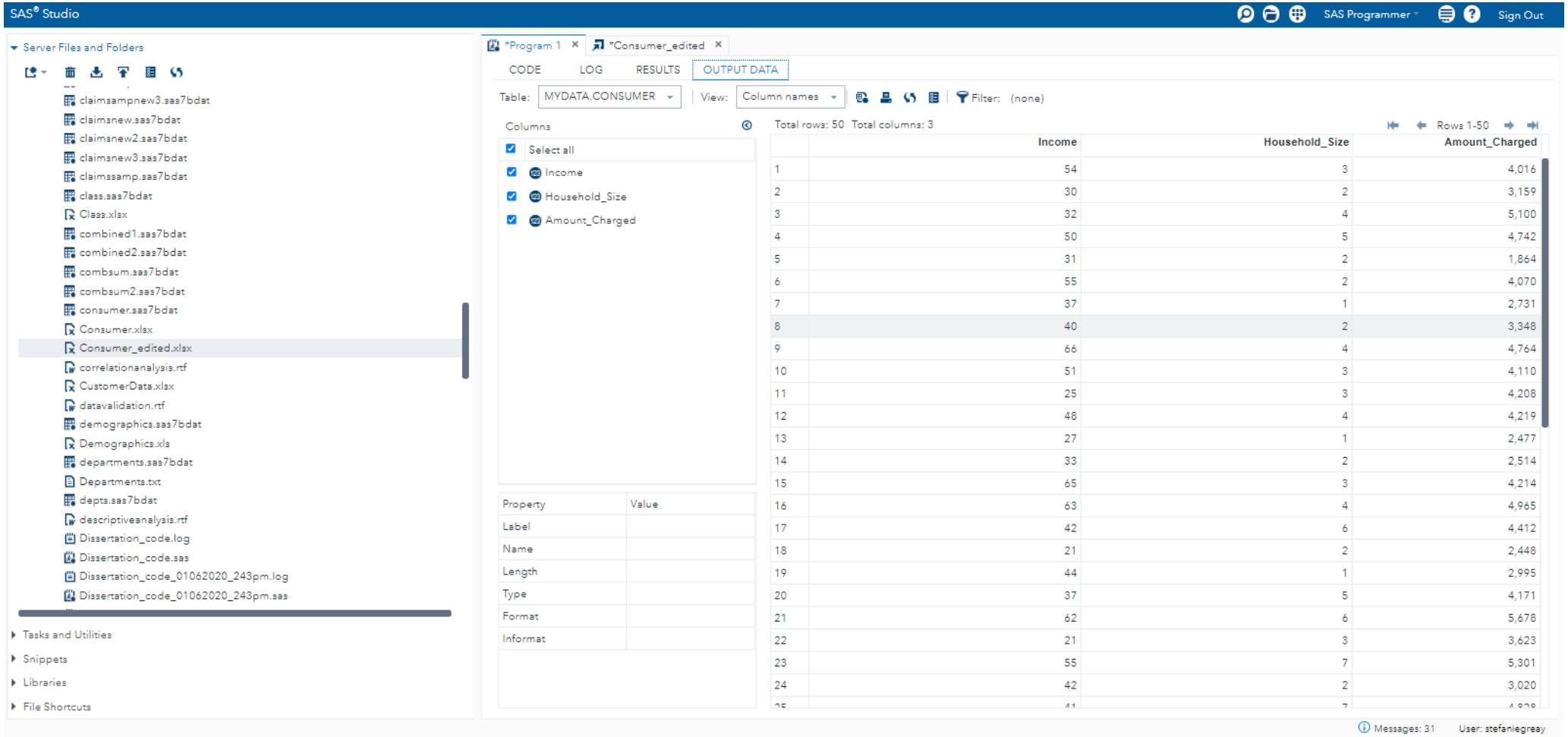


You can determine what the path is for your Files(Home) Folder by right clicking on Files(Home) and selecting Properties. (Yours will not be the same as mine – it will include your userID, not mine.)

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 ouput data window and can verify its success.



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

*Program 1 | *Consumer_edited

CODE | LOG | RESULTS | OUTPUT DATA

```
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;
run;

/*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;
```

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)