Running head: ANALYZING CUSTOMER CHURN RATES 1

# Option #1: Portfolio Project – Analyzing Customer Churn Rates

# Kasey Aldridge

# MIS445 – Statistics in Business Analytics

# Colorado State University = Global Campus

# Professor Alin Tomoiaga

# January 12, 2020

# Analyzing Customer Churn Rates

In this report I will take the data from the Telco Extra telecommunications company and try to find out why they are having such a high churn rate. Business analysts can make many strides in reducing the business losses by knowing exactly what is causing the problem. By locating the information in a timely manner this information can prove invaluable to the solvency of the company (Jamal & Bucklin, 2006).

**Purpose**

There are very good benefits to collecting churn data on a company’s customers. Not only can the company pinpoint where the biggest churn issues are coming from but they can try to come up with ways to stem the losses by focusing away from the biggest causes. So for example a company may change its marketing campaigns in order to draw in more customers that are predicted to have the lowest level of churn rates (Babu & Ananthanarayanan, 2016).

**Sample Population Characteristics**

The sample population of 1,000 customers will be analyzed, hypothesized and then interpreted. There are quite a few multiples but the only variables that we will be concerned with are the age of the customers, how long they have lived at their current address, their gender, what level of education do they possess, their household income, marital status, their region of the companies territories, what category or level of service they are contracted for, and churn rate. Not all of these variables will be able to be utilized as it is expected that some will have little to no bearing on the cause of the churn. This report is to decide which ones those will and will not be and to come up with the most accurate and useful information (Tomoiaga, 2019).

**Multiple Regression Models**

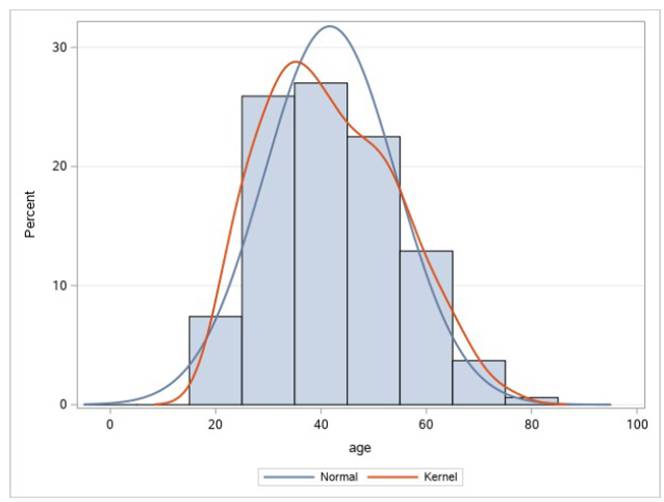
It is recommended to try to keep this as simple as possible. One of the ways to do that is to make the dependent variable one from the continuous category such as age, years at address, or income. Any of the independent variables used can be from either the previous category or it can be from the categorical ones. Avoiding categorical variables that have at least three categories such as the three regions the company is in charge of or the five different levels of education is usually the best advice for achieving simplicity. Gender or marriage are good dummy variables as there is only two answers, 0 and 1 (Tomoiaga, 2019), (Jamal & Bucklin, 2006).

**Continuous Variables – Age, Year at Current Address and Income**

**Sample Characteristics Age**

****

**Figure 1 above**

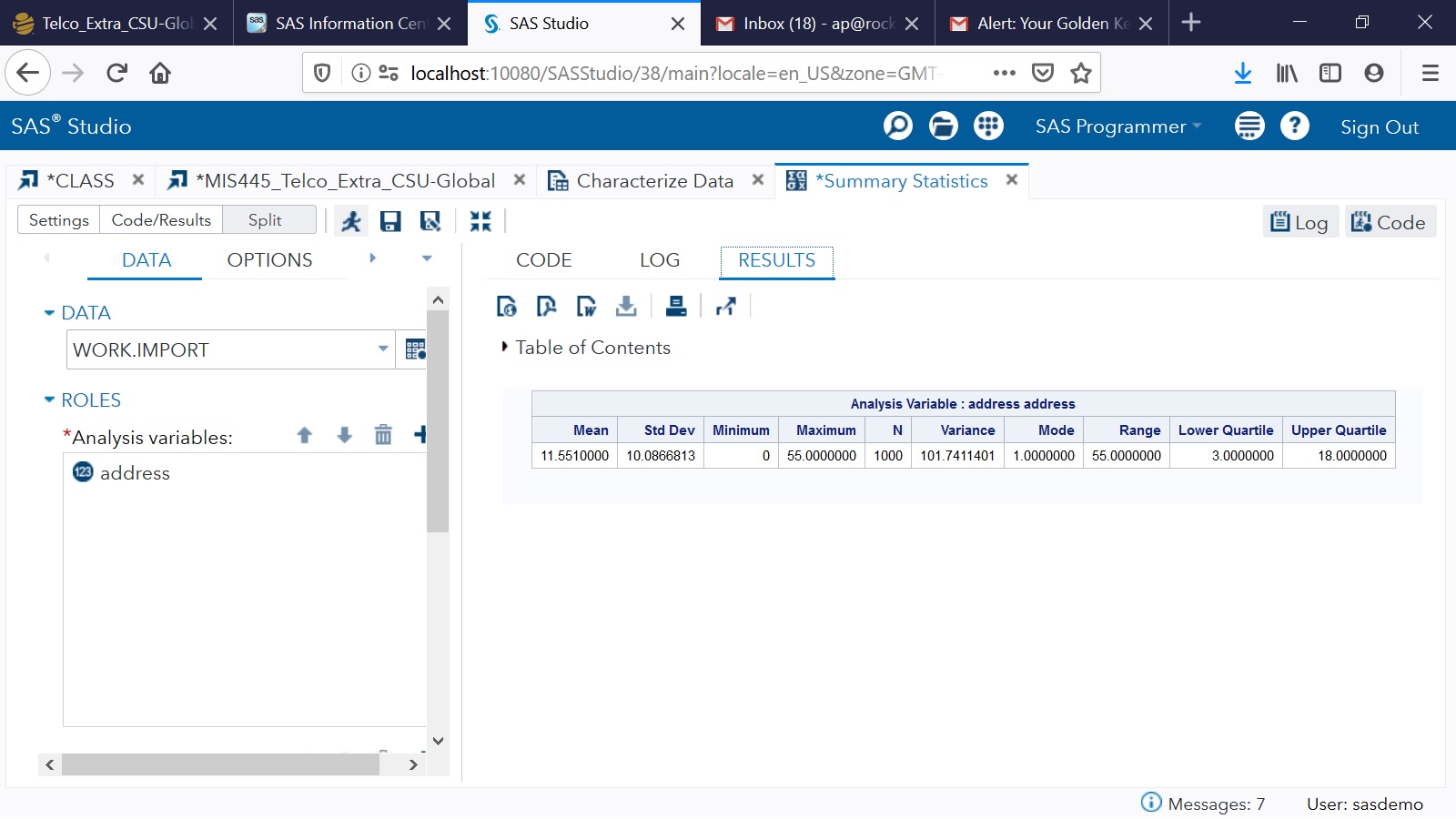
****

**Figure 2 above**

(SAS, n.d.)

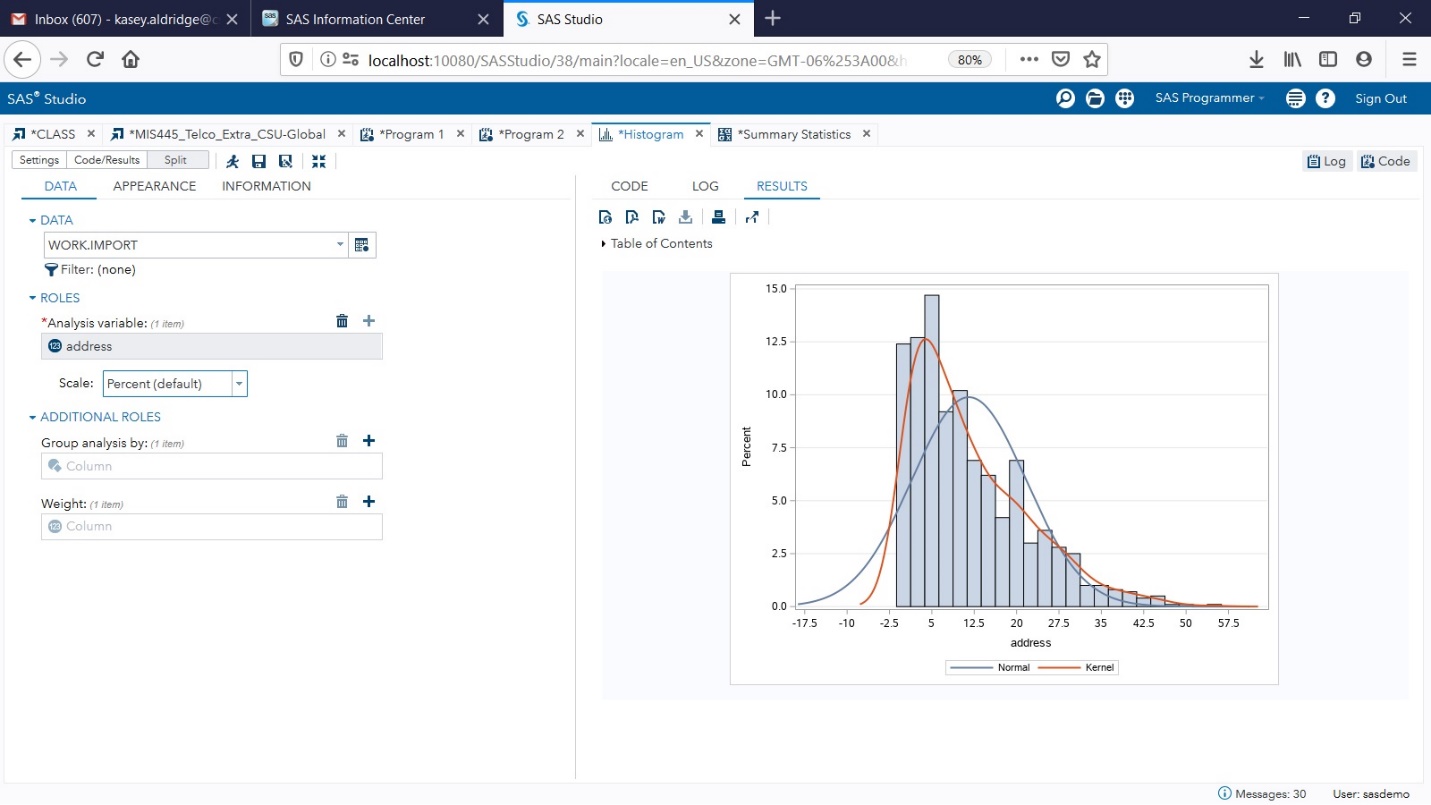
In the above table (figure 1) it shows that the mean age is 41.68 years old and the standard deviation 12.559. This shows that the company’s average customer age is 41.68 and that the majority are within the range of 12.559 older or younger. Figure 2 shows a visual example of this. This variable is measured on a number and continuous scale and is a quantitative variable (Tomoiaga, 2019).

**Sample Characteristics Years at Current Address**

****

**Figure 3 above**

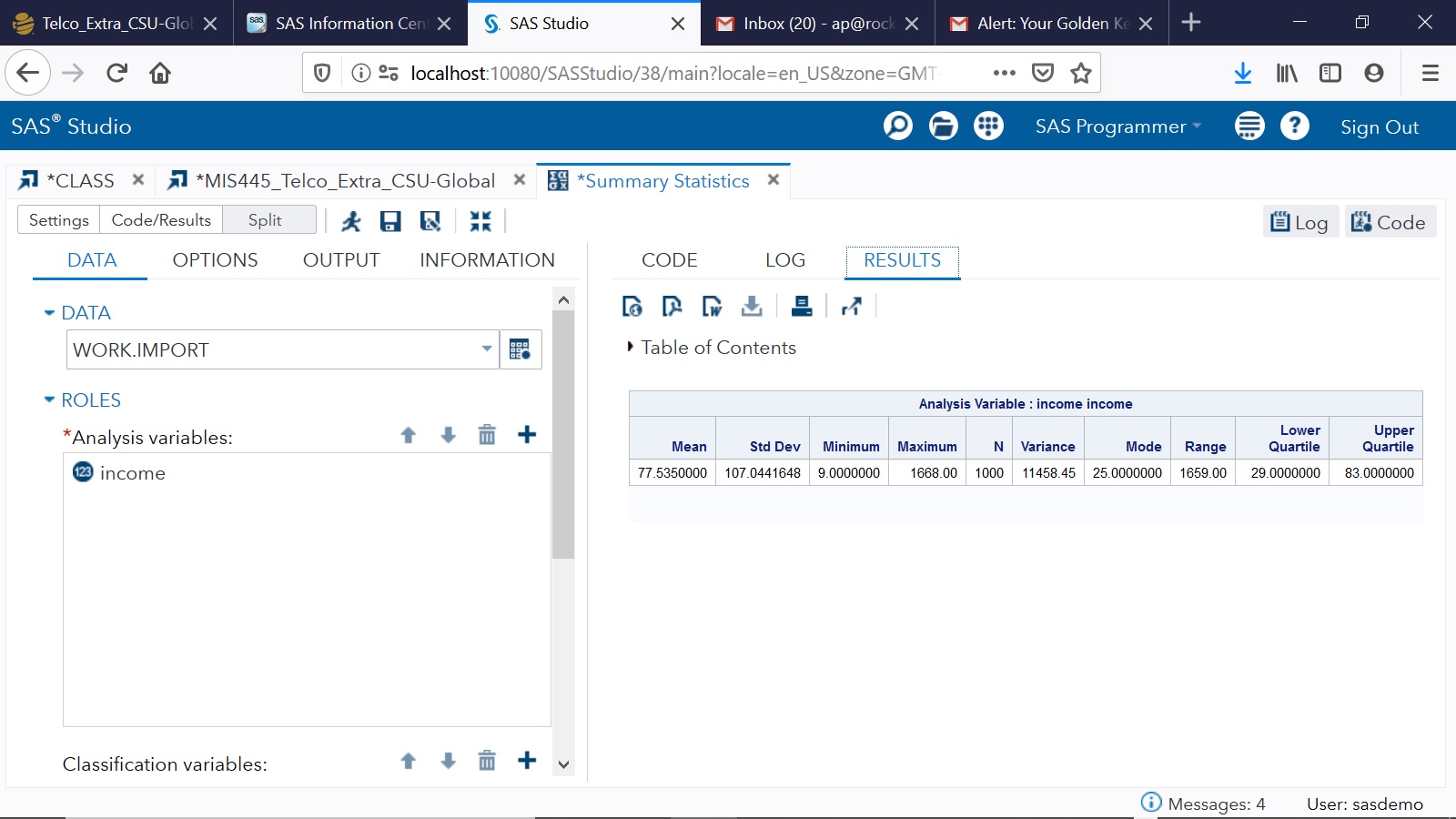
**Address at Current Address Graph**

****

**Figure 4 above**

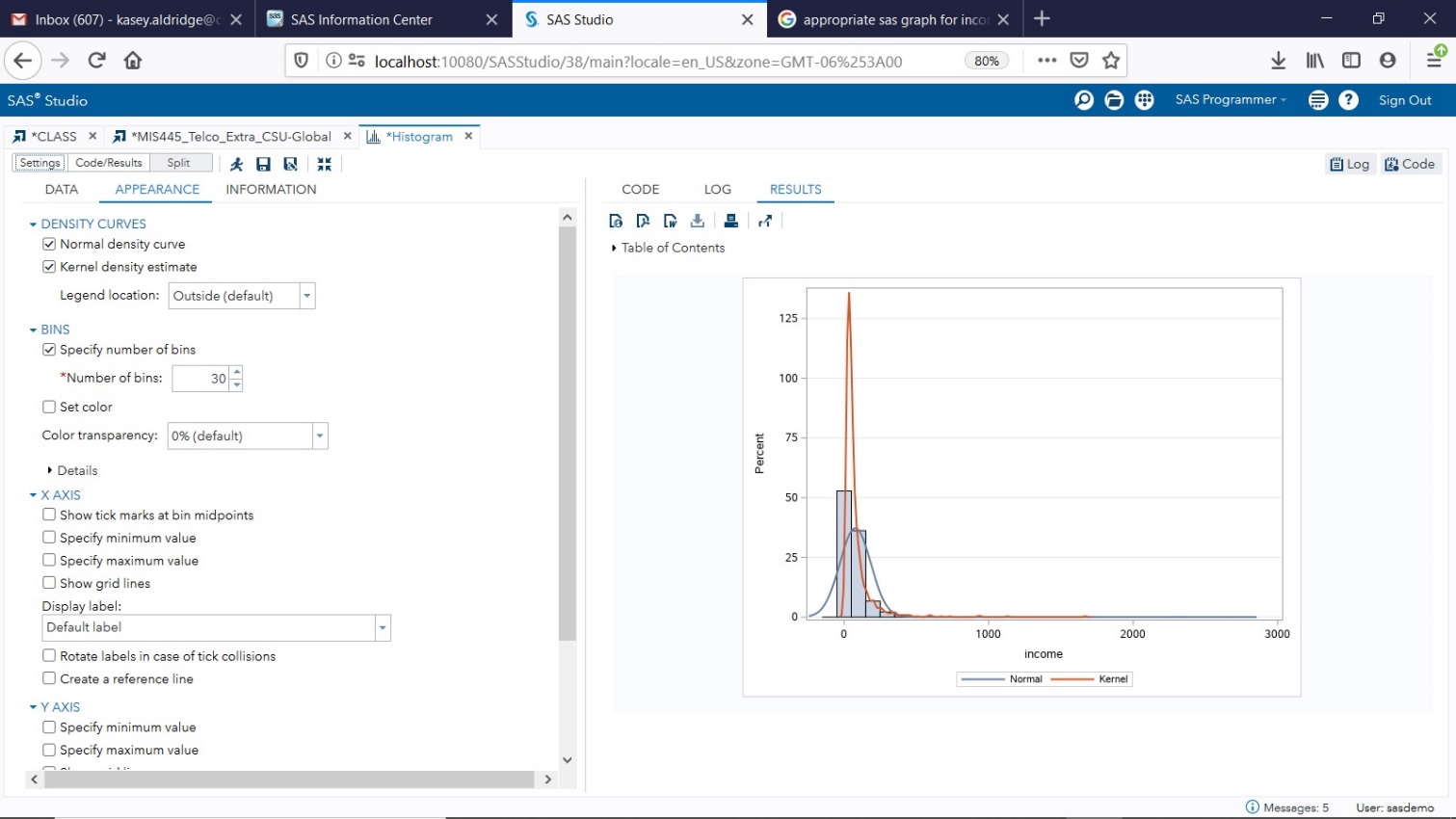
Above is the data for approximately how many years their customers have lived at the same address. Figure 3 above shows the mean to be 11.55 years and the standard deviation to be 10.08 years. Again, this is a quantitative variable and measured on a continuous scale. Looking at the right skewed graph above (figure 4) it looks like most of their customers rarely stay in one place for over 5 years (Tomoiaga, 2019).

**Sample Characteristics Income**

****

**Figure 5 above**

**Income Graph**

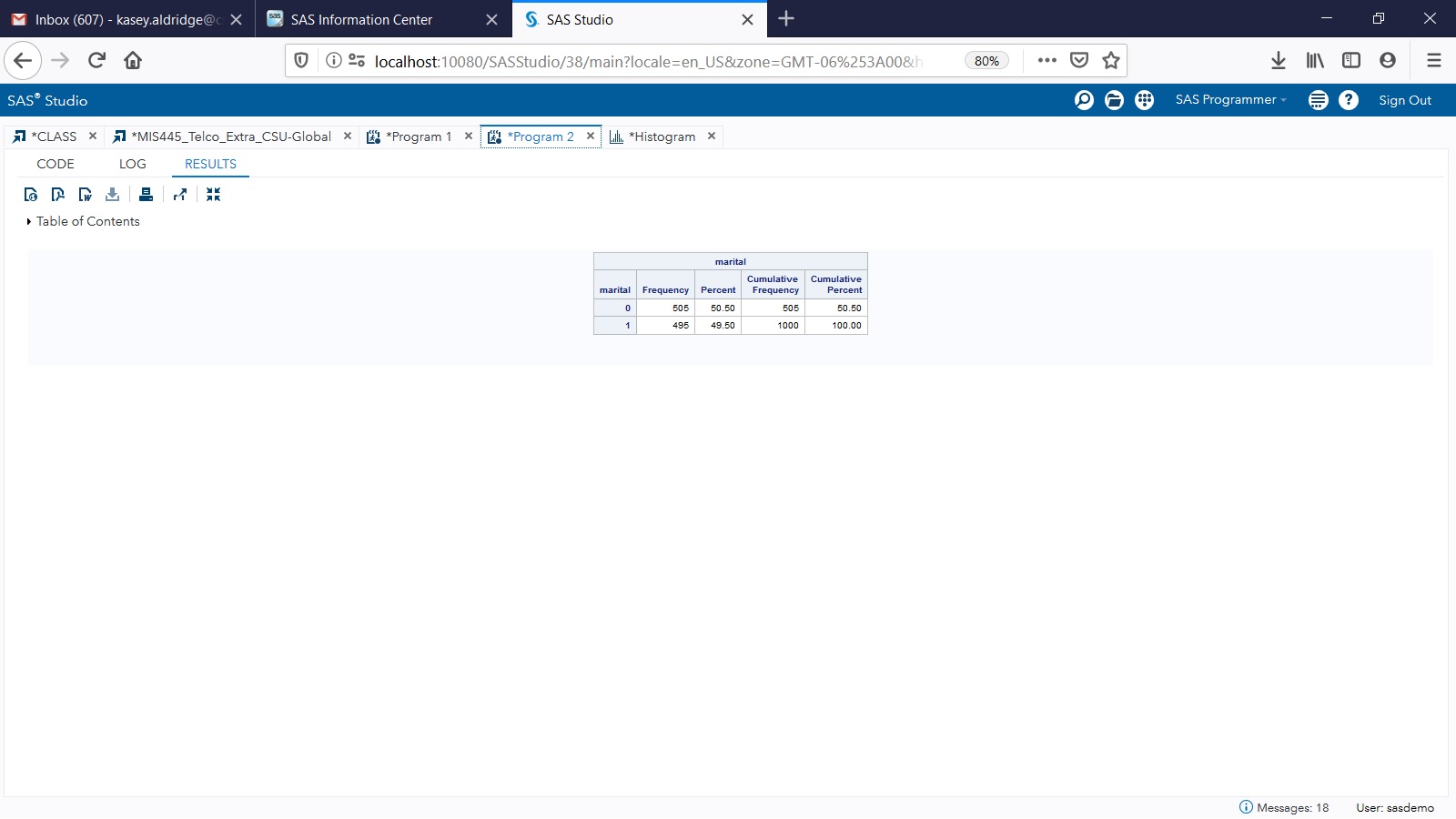
****

**Figure 6 above**

Income is the last category that is a continuous scale and a quantitative variable. This category is measured in thousands so figure 5 above shows that the average income is $77,535 with a standard deviation of 107.04 (Tomoiaga, 2019).

**Categorical Variables with Two Categories/Dummy Variables – Marital Status, Gender, and Churn**

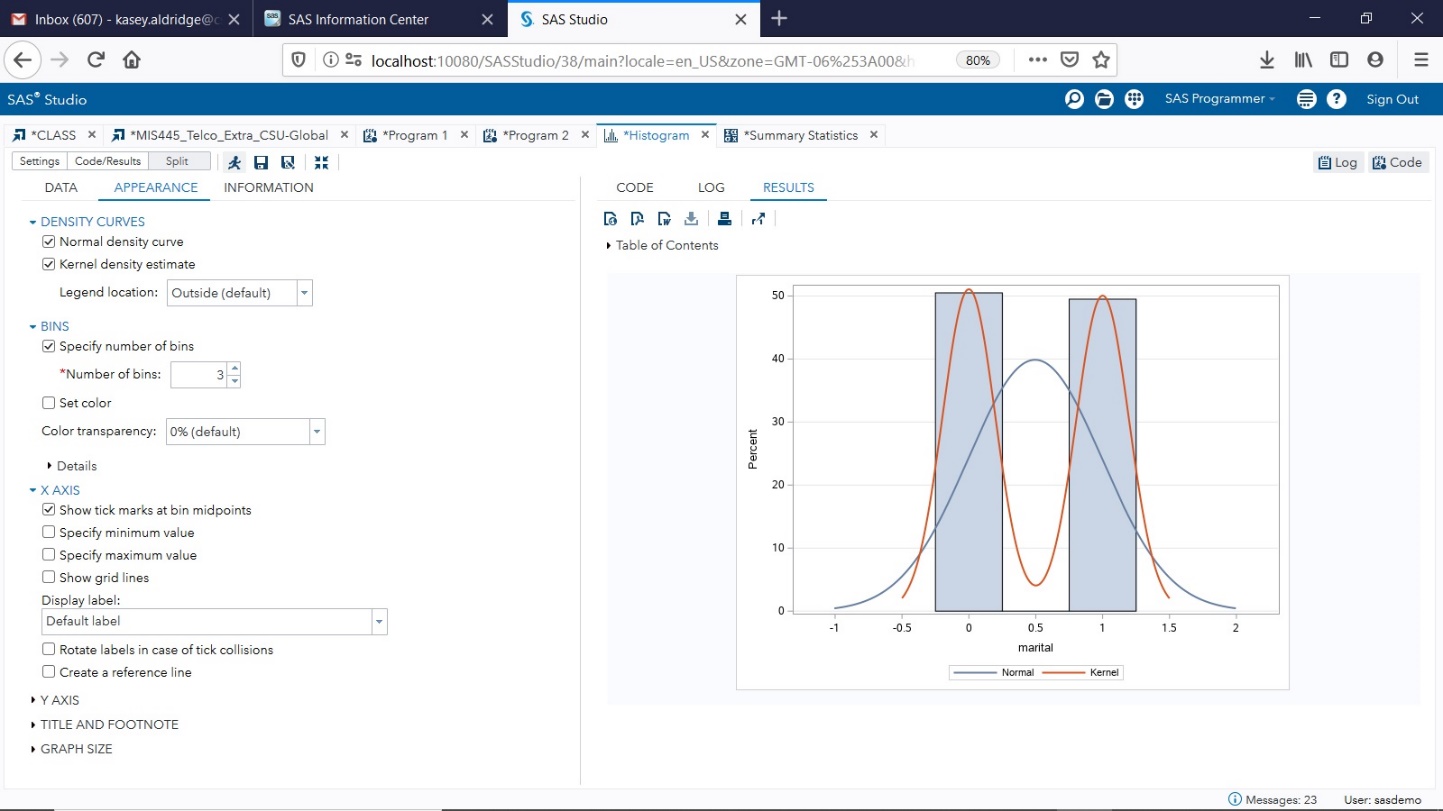
**Sample Characteristic Marital**

****

**Figure 7 above**

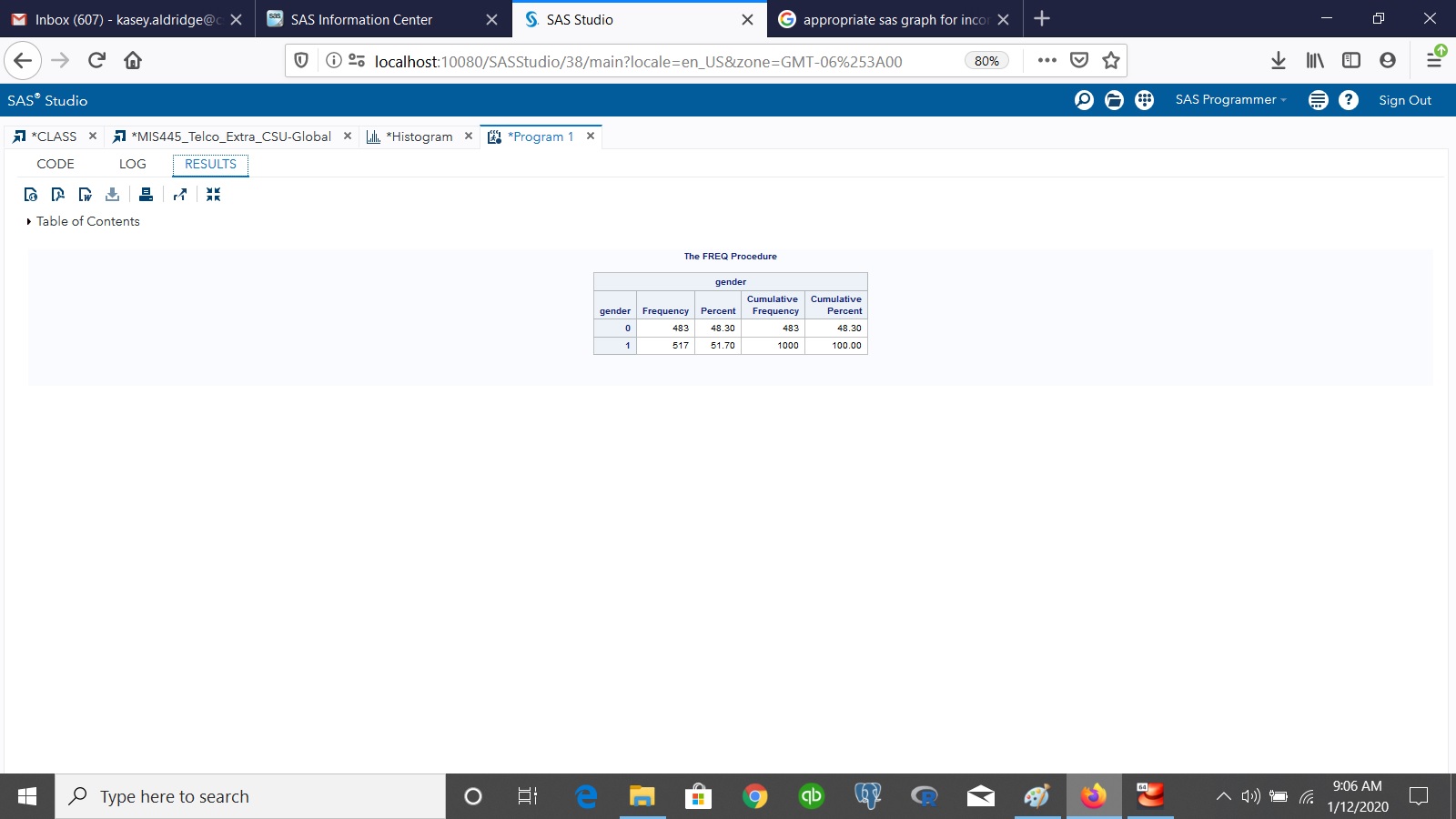
This categorical variable shows that 50.5% of the customers are married, while 49.5% remain unmarried with 0 = Married and 1 = Unmarried in figure 7 above (Tomoiaga, 2019).

**Marital Graph**

****

**Figure 8 above**

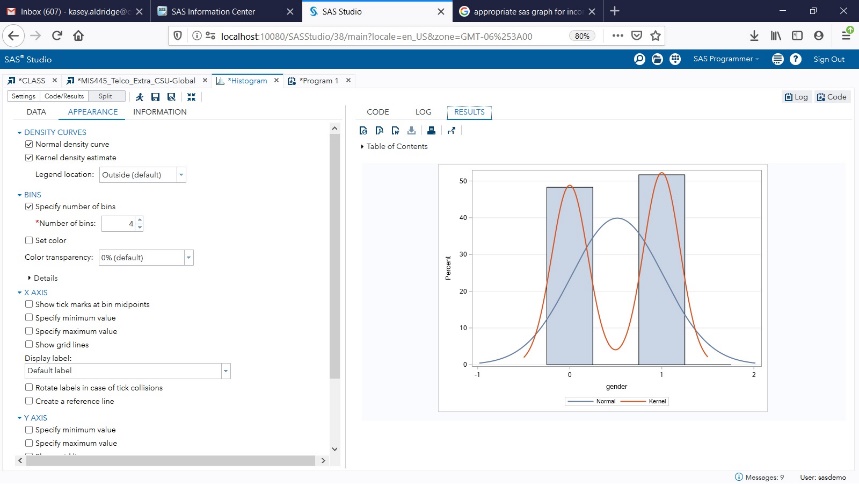
**Sample Characteristics Gender**

****

**Figure 9 above**

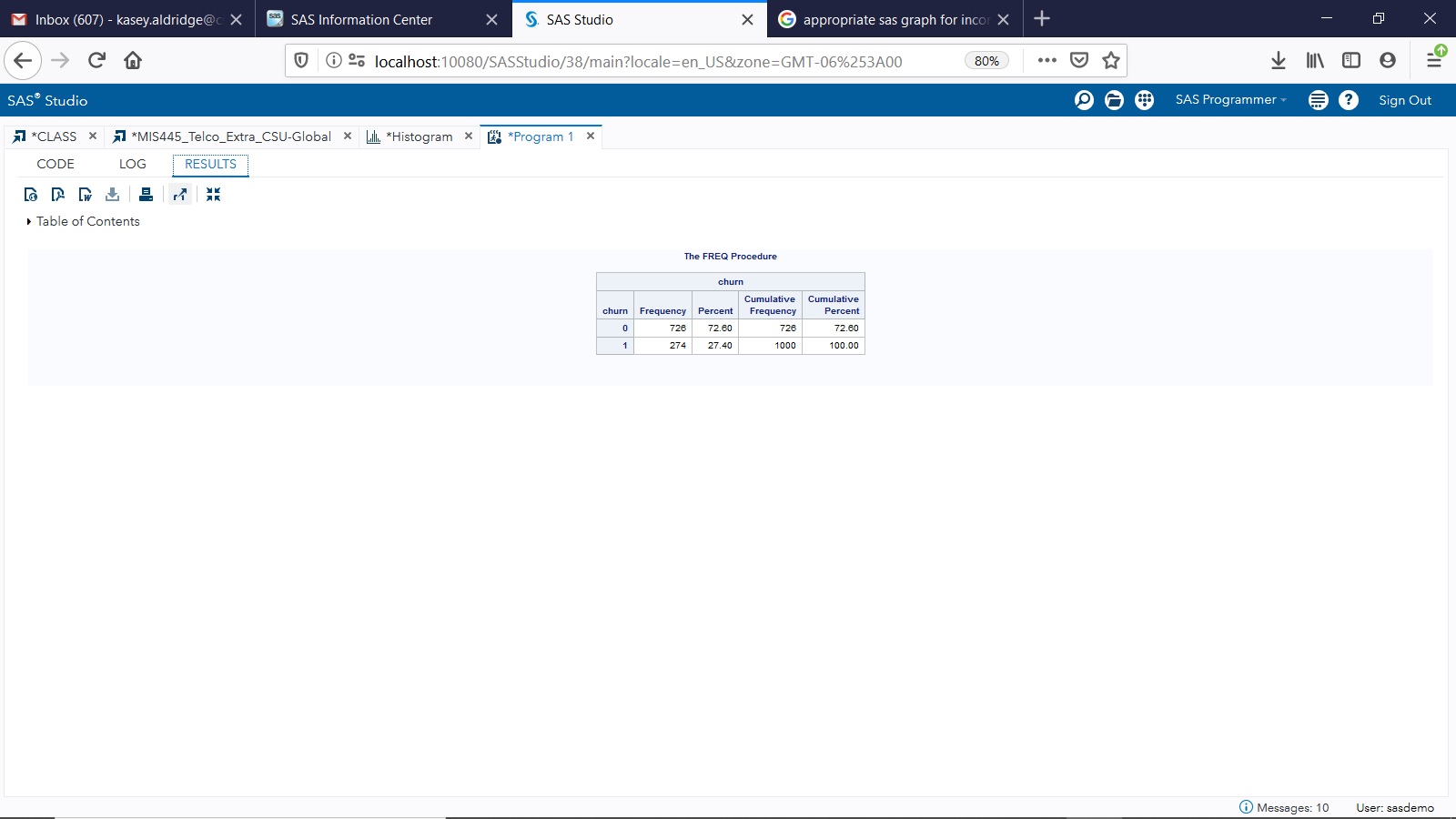
48.3% of the customers are male, while 51.7% are female with figure 9 above having 0 = Male and 1 = Female. This variable is a categorical variable (Tomoiaga, 2019).

**Gender Graph**

****

**Figure 10 above**

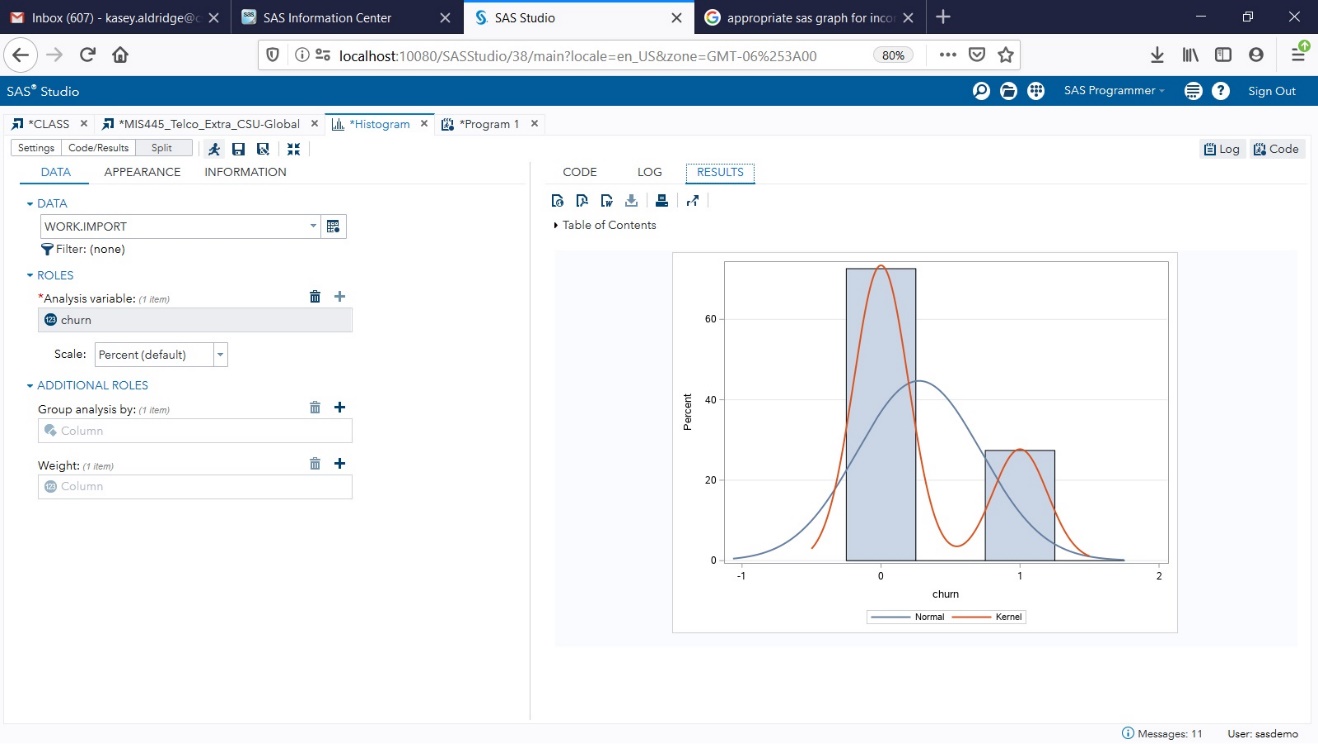
**Sample Characteristics Churn**

****

**Figure 11 above**

72.6% of the customers will not churn, while 27.4% will churn. Figure 11 above shows the churn rate where 0 = Not Churn and 1 = Churn. This variable is categorical that is measured on a nominal scale. This result shows that 27.4% of their customers have cancelled their services in the past month (Tomoiaga, 2019).

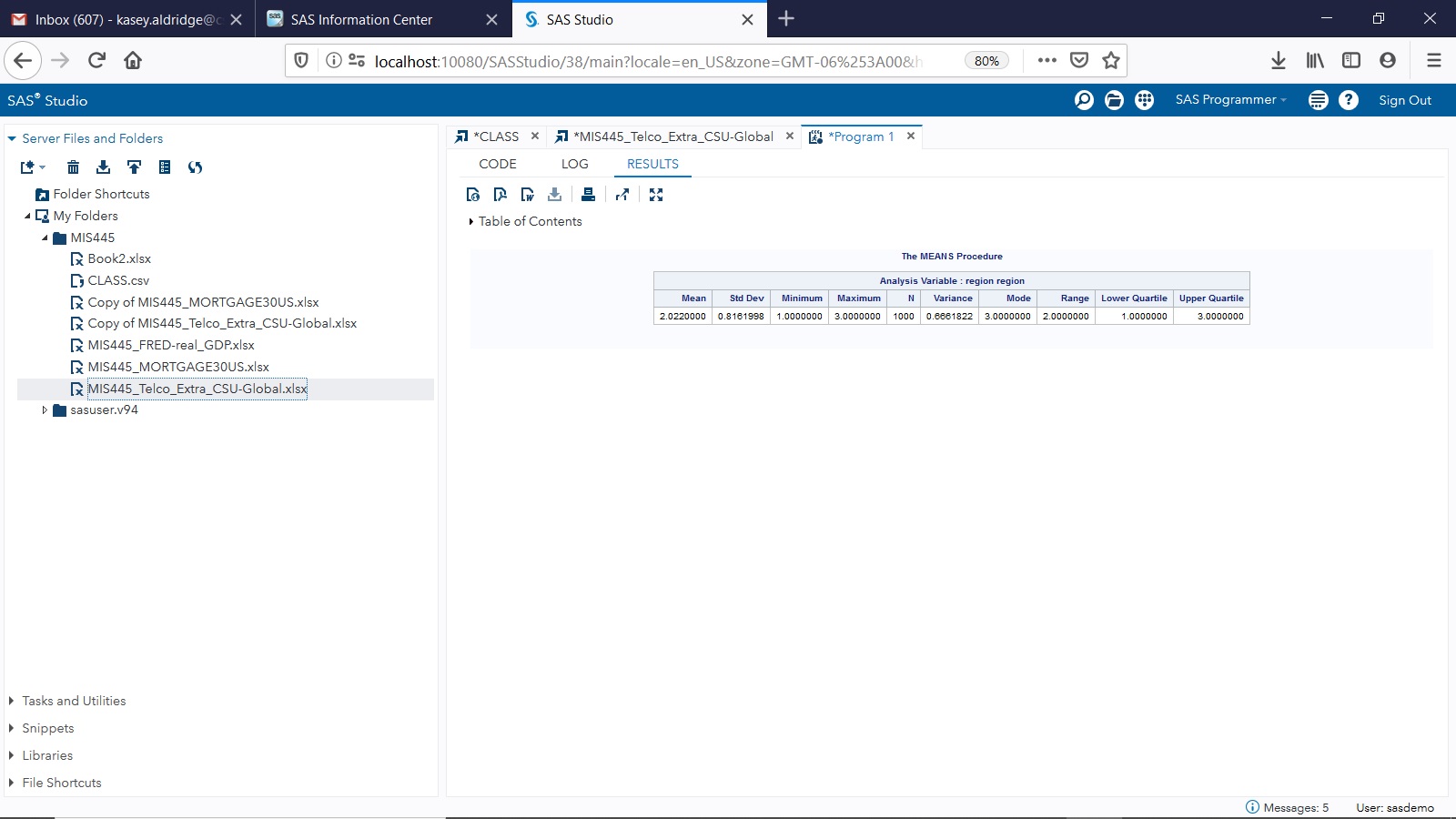
**Churn Chart**

****

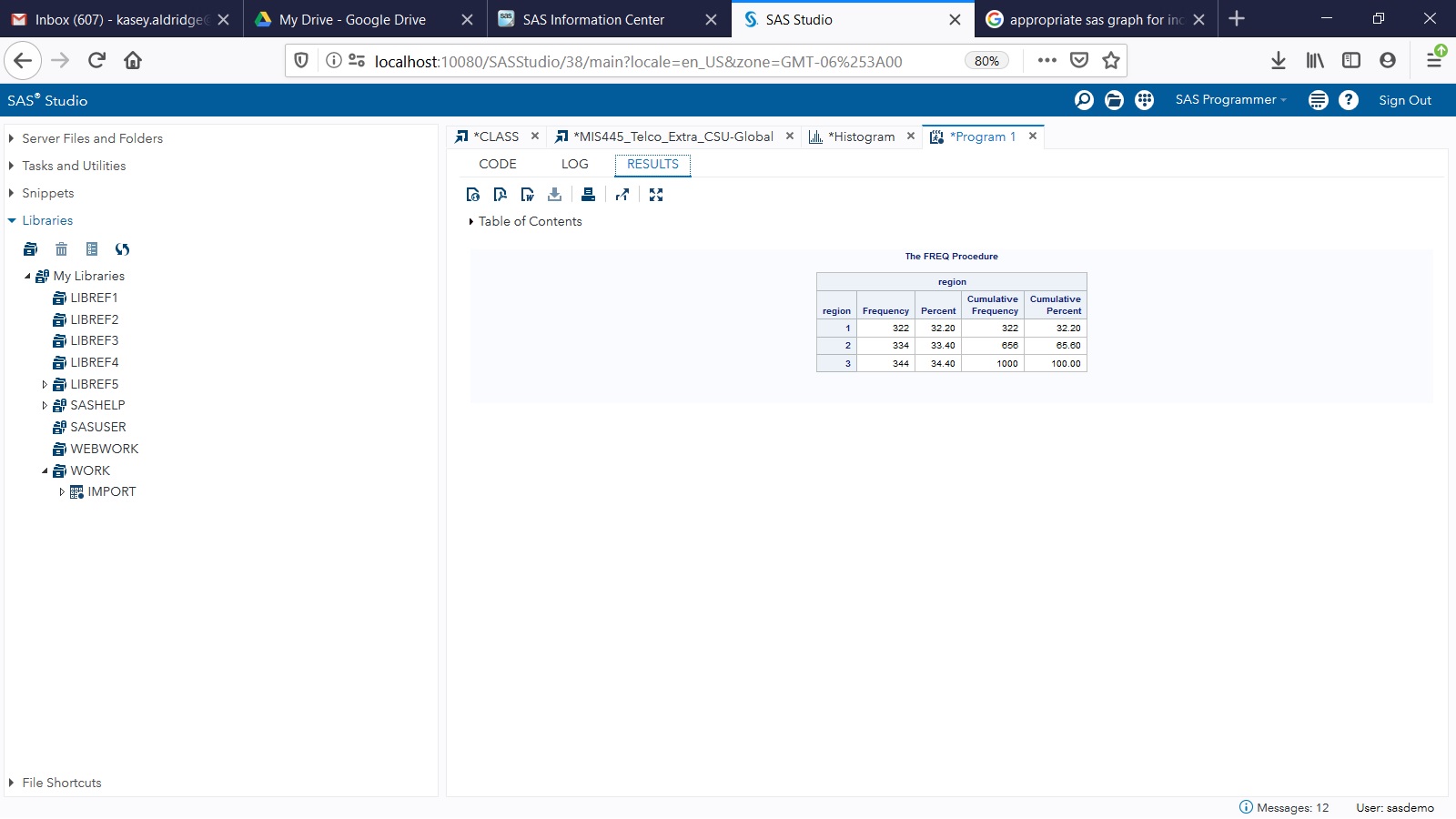
**Figure 12 above**

**Categorical Variables with more than Two Categories -Region, Level of Education, and Customer Categories**

**Sample Characteristics Region**

****

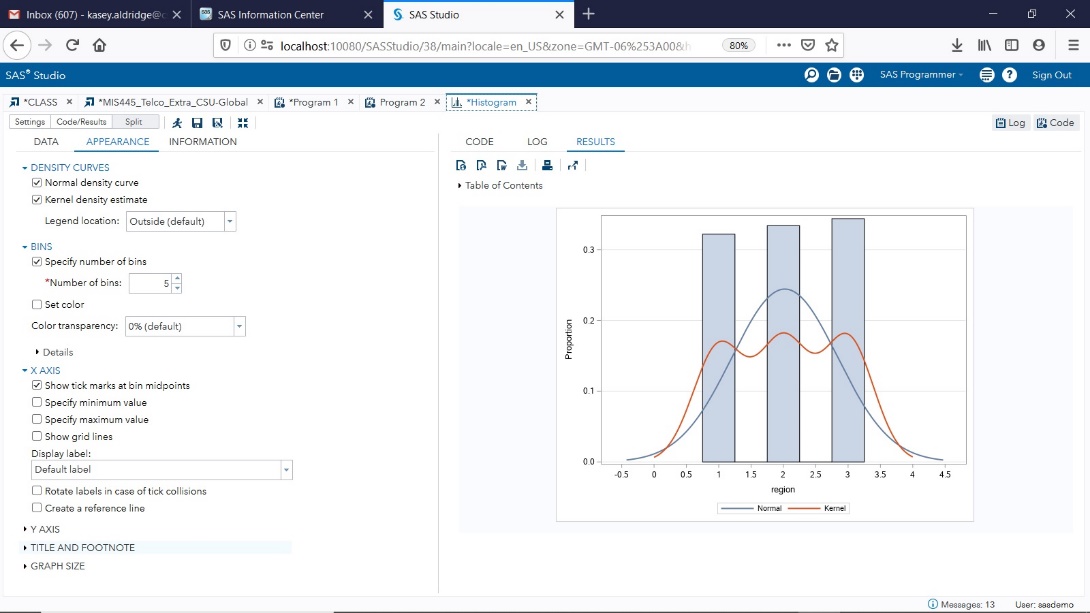
**Figure 13 above**

****

**Figure 14 above**

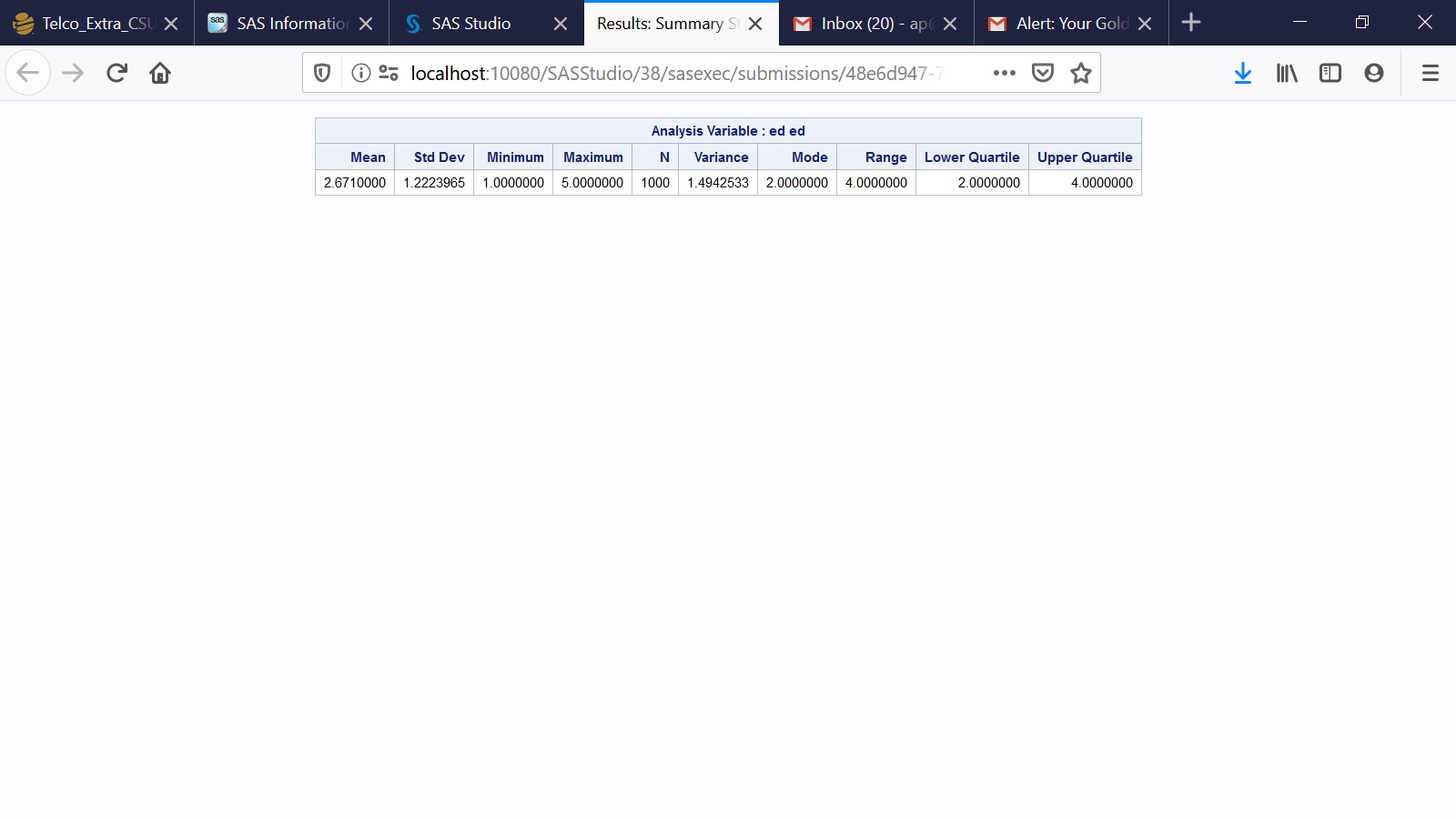
The region zones are listed above in figure 13 above. It shows that this categorical variable divvies up their customers in three ways. 32.2% of the customers are in Zone 1 1, 33.4% are in Zone 2, and 34.4% are in Zone 3. Although Zone 3 had the highest amount of customers the difference is negligible (Tomoiaga, 2019).

**Region graph**

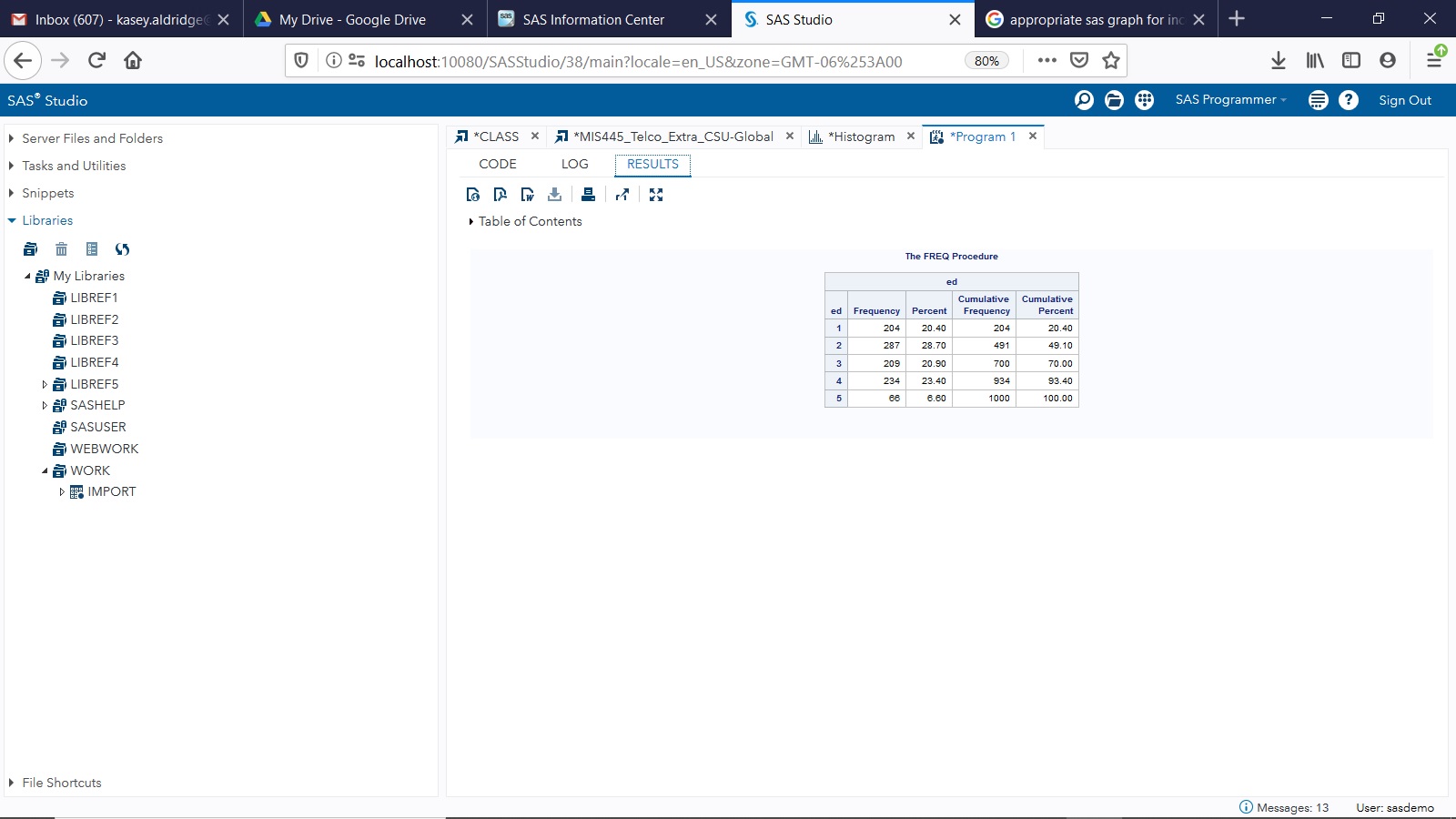
****

**Figure 15 above**

**Sample Characteristics Level of Education**

****

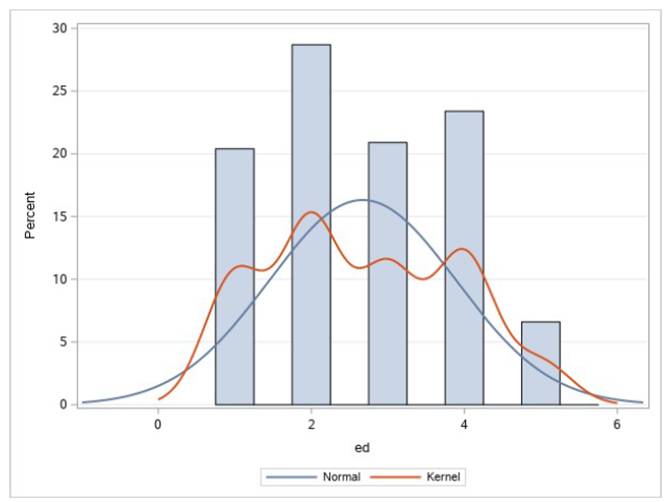
**Figure 16 above**

****

**Figure 17 above**

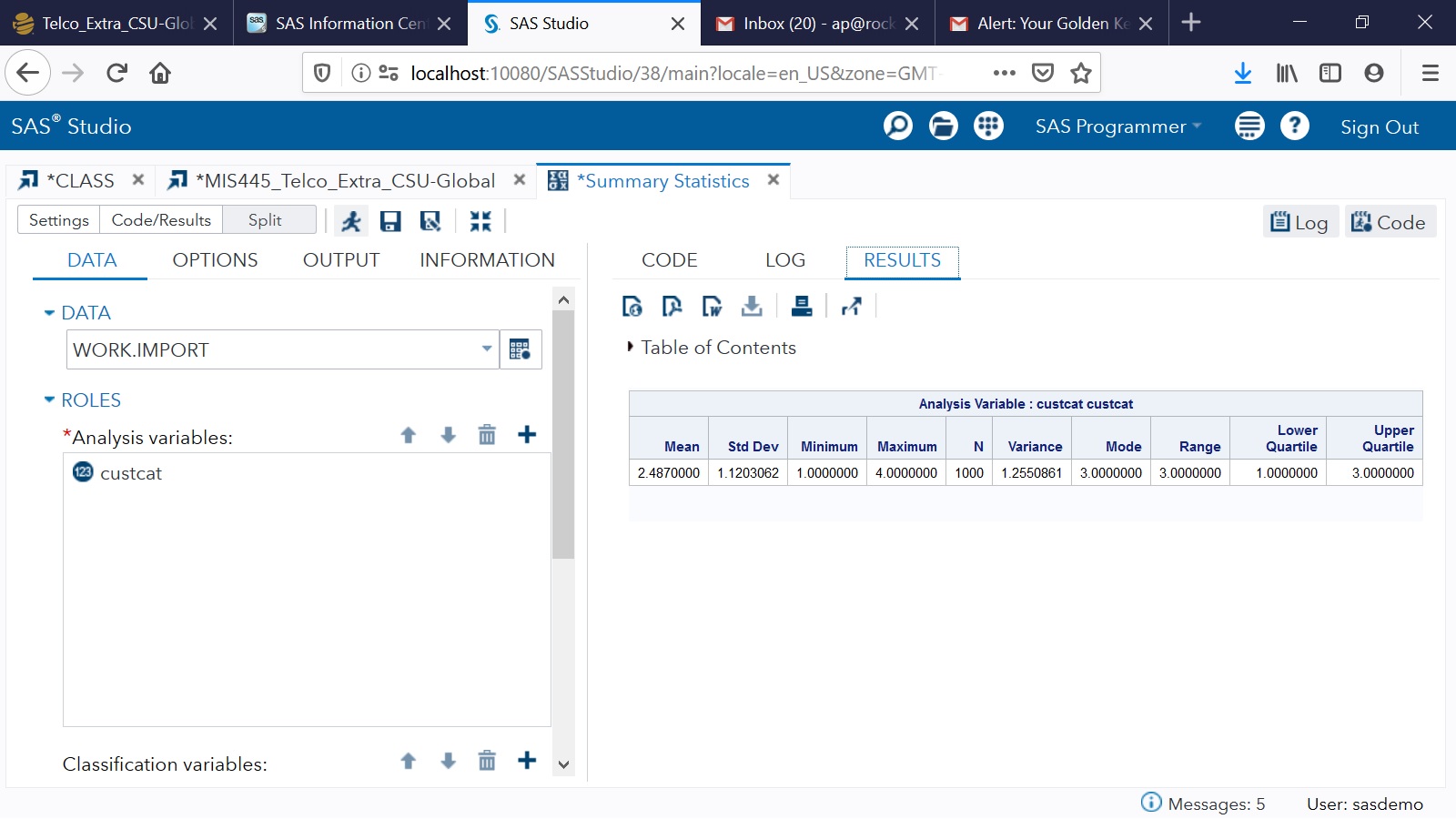
In figure 17 above, the five educational options and its corresponding percentage rates are: 20.4% of the customers did not complete high school, 28.7% have a high school degree, 20.9% have some college, 23.4% have a college degree, and 6.6% have a post undergraduate degree. This variable is categorical and measured on an ordinal scale (Tomoiaga, 2019).

**Level of Education Graph**

****

**Figure 18 above**

**Sample Characteristics Customer Category**

****

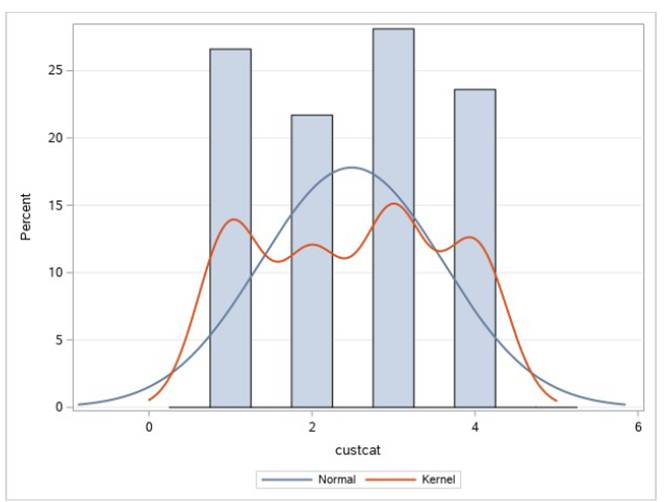
**Figure 19 above**

****

**Figure 20 above**

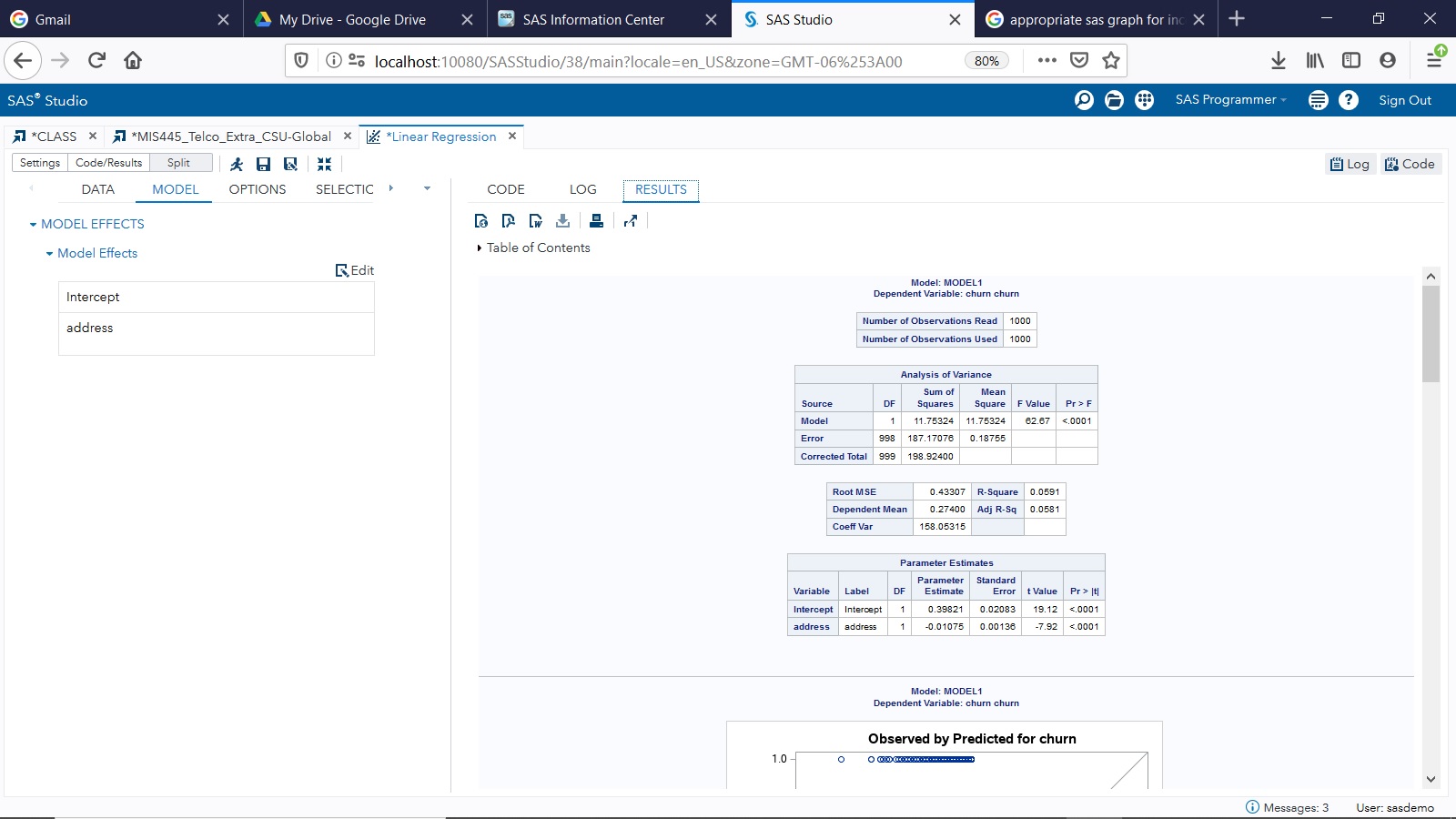
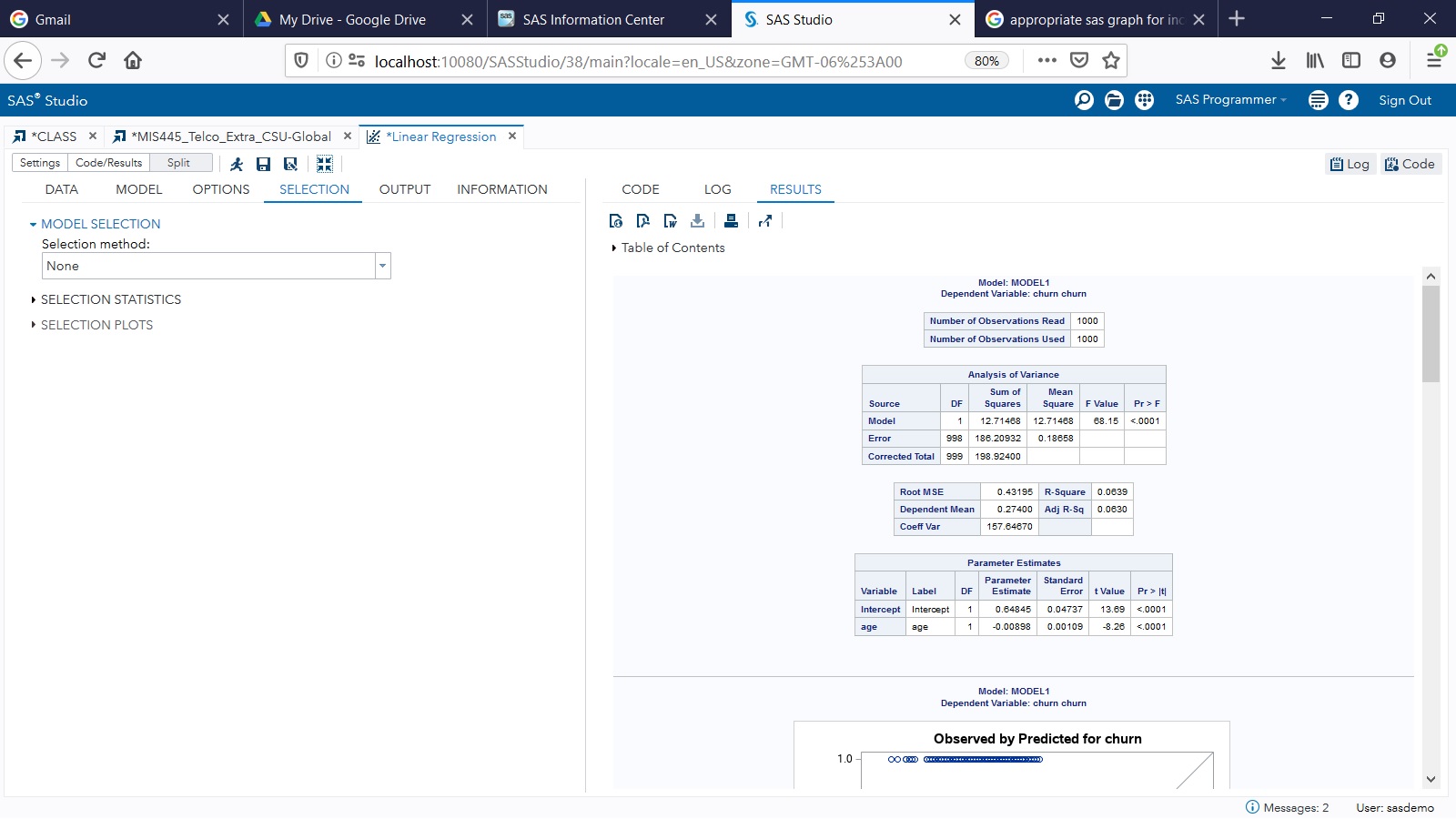
Figure 20 above shows the cuscat, also known as Customer Category. This categorical variable is measured on a ordinal scale. It shows that 26.6% of the customers prefer basic service, 21.7% desire E-service, 28.1% want Plus service, and 23.6% have purchased the total service package (Tomoiaga, 2019).

**Customer Category Graph**

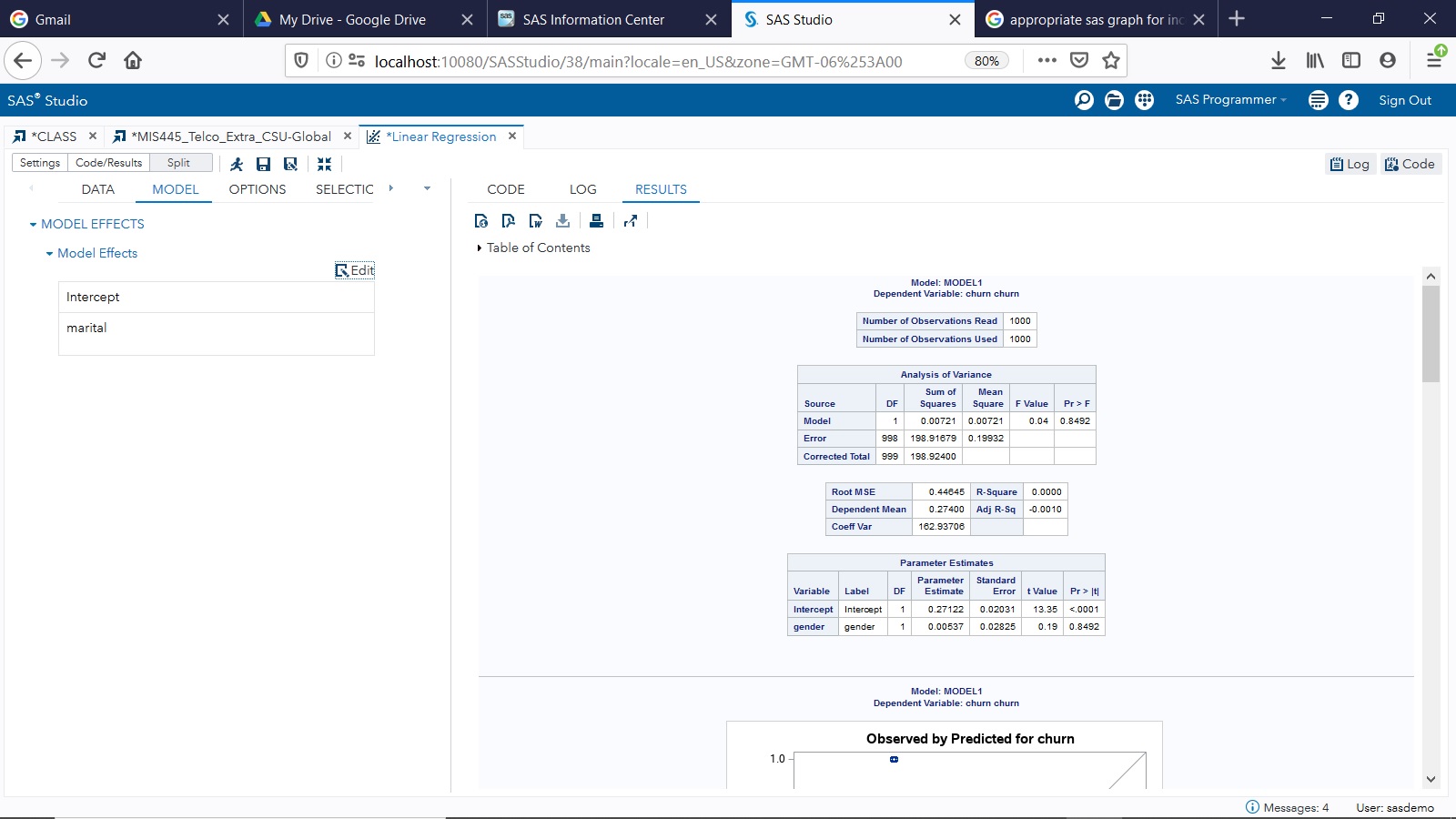
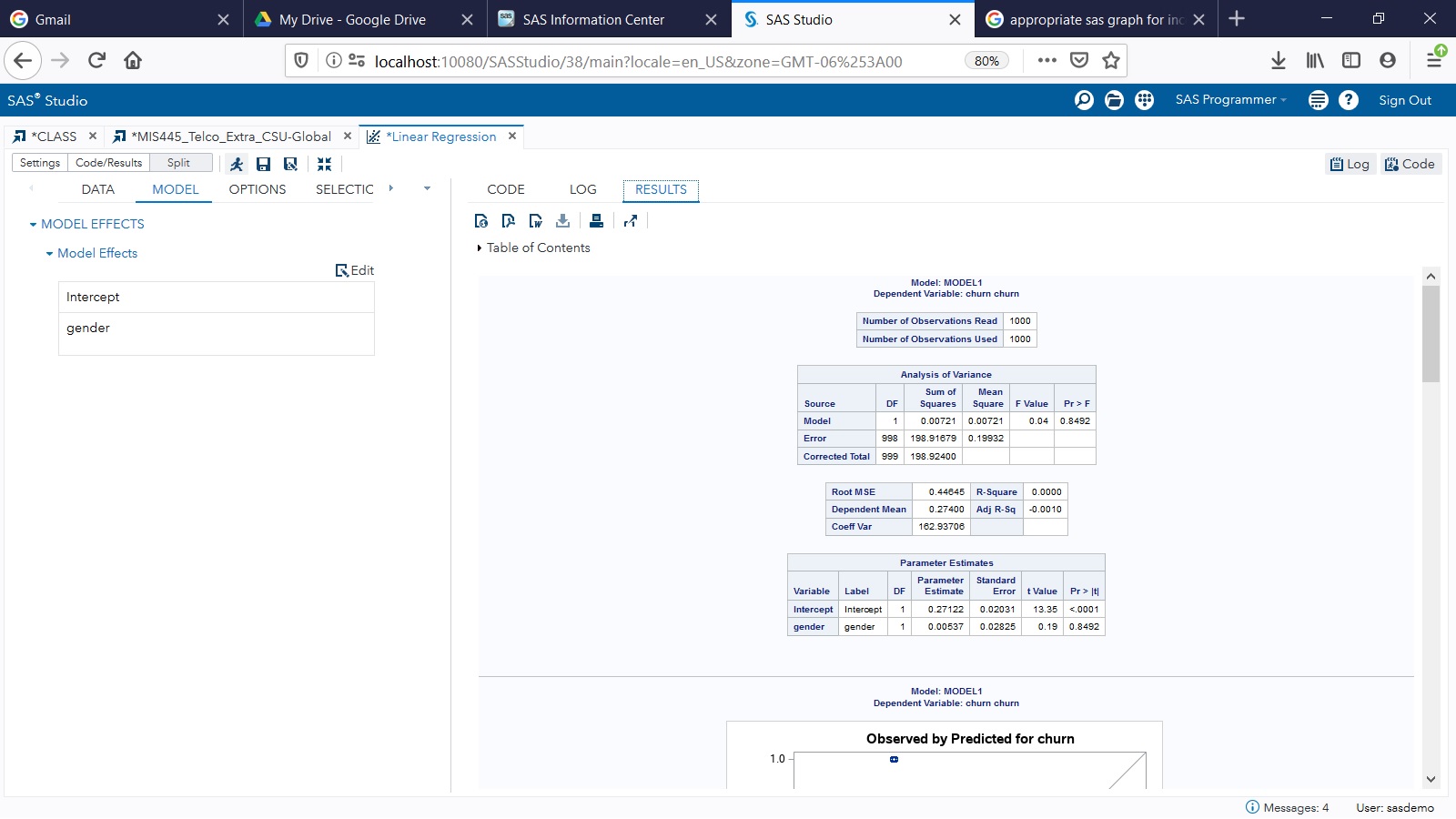
****

**Figure 21 above**

**Linear Regression Analysis**

****

**Figure 22 and 23 above**

****

**Figure 24 and 25 above**

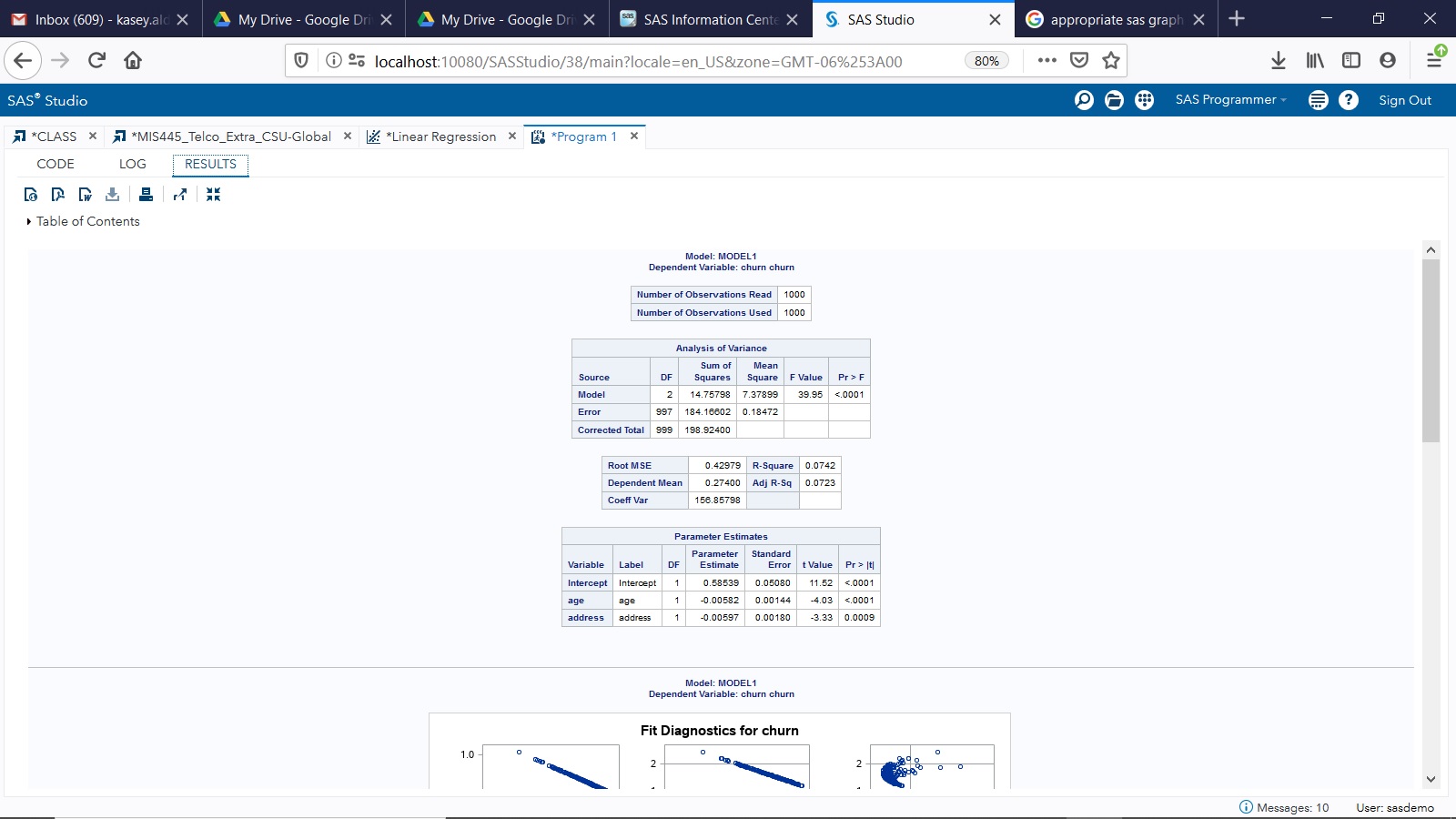
**Multiple Linear Regression Analysis – Dependent Variable: Churn**

Due to the churn rate being fairly high it is important to accurately predict what is the most likely cause. First, we will test churn as a dependent variable against the independent variable ‘age’. The second one that was chosen was years at an address. Both tests will check to see if any relationship exists or can be predicted (Tomoiaga, 2019).

**Hypothesis**

H0: ; In other words, the variable is not a significant predictor for our model;

H1: In other words, the variable is significant for our model.

****

**Figure 26 above**

**Interpretation**

The interpretation for is , then *7% of the variability of the variability of the dependent variable is explained by the variability of the independent variables* OR *7% of the variability of the dependent variable is explained by the regression model.* This small number seems to represent a weak positive linear relationship (Tomoiaga, 2019).

The hypothesis test is concluded by comparing the *p*-value (which is .0001 from figure 26 above) with the significance level α=0.05. Since the p-value is less than the significance level, then we reject the null hypothesis. In other words the variables chosen are significant for our model (Tomoiaga, 2019).

Hypothesis test for the significance of Age:

H0:

H1:

Hypothesis test for the significance of Address:

H0:

H1:

In both cases the p-value corresponds to each predictor is less than the significance level a= 0.05, so each predictor is significant for our model, so we validate the model and we may write the estimate of the regression equation as: churn = b0 + b1 \* Age + b2 \* Address, where are the estimates of and their values are given in the column **Parameter Estimate** (Tomoiaga, 2019)**.**

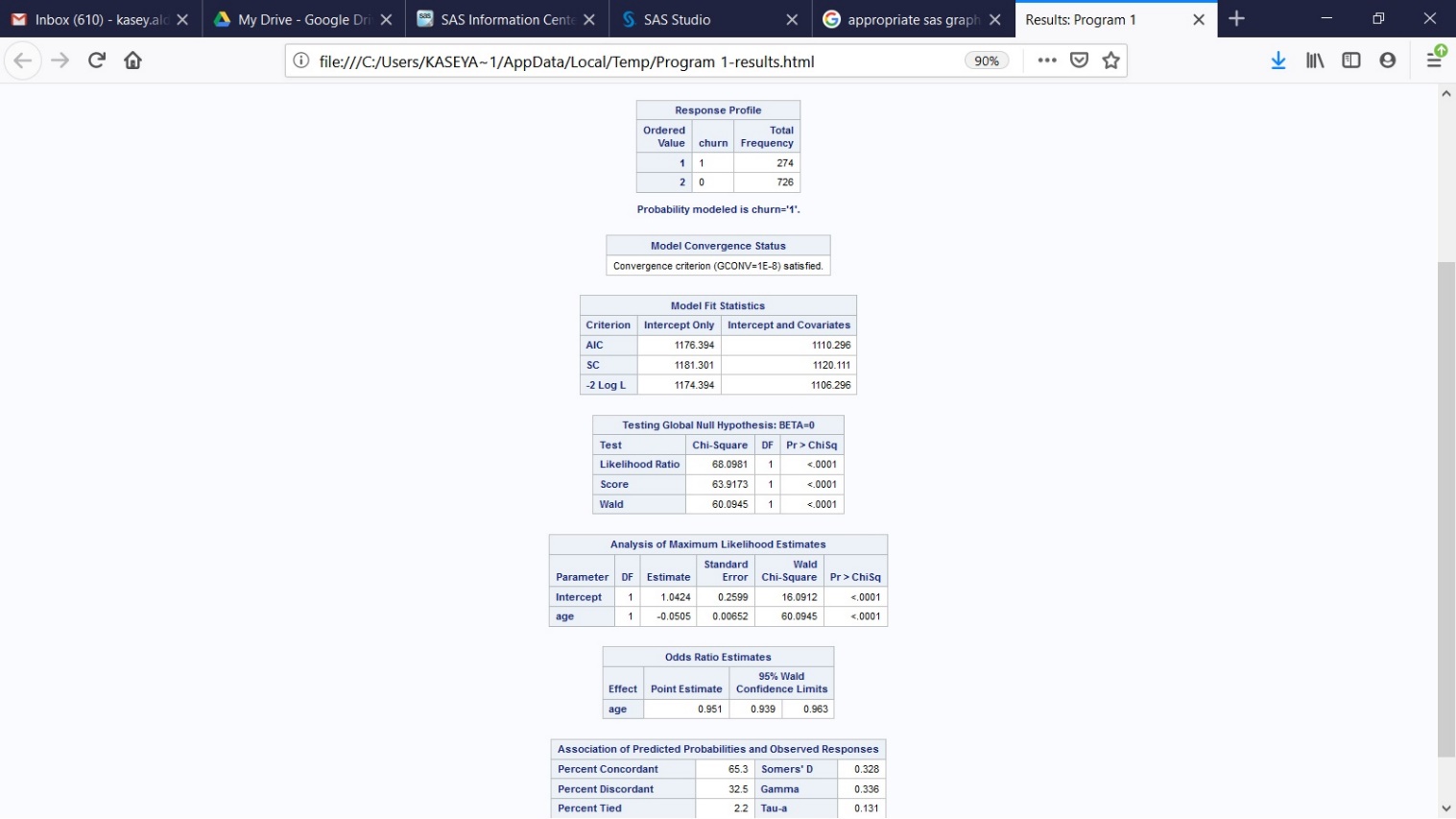


Figure 26a above

After running the regression code, the result is Probability modeled is churn=’1’ (figure 26a above). So, the regression equation is: After this we examined the Analysis of Maximum Likelihood Estimates table (figure 26a above) and using this table, we performed the following test on the significance of our predictor Age:

H0: =0; in other words, Age is not a significant predictor.

H1: ; in other words, Age is a significant predictor.

Then we analyzed the p-value indicated in the column called Pr > ChiSq (figure 26a above) and these results show that the value is less than the significant predictor so we reject the null hypothisis. This shows the logistic regression equation to be:

After this we studied the Analysis of Maximum Likelihood Estimates (figure 26a above) which shows that the estimate of the regression equation is: . So from there we get: odds to churn = *e*1.0424-0.0505\*age. So, since , let’s say 0.0505, then the correct interpretation would be: **for a one-unit increase in Age, we expect to see about 1.0424-1=0.0424, i.e. 4% increase in the odds of churning.** This tells us that once a customer gets older, the odds to churn are increasing (Tomoiaga, 2019).

Finally, **let’s make a prediction.** So, assume that we have a 27-years old customer. Then, the estimate of odds to churn for this person is:

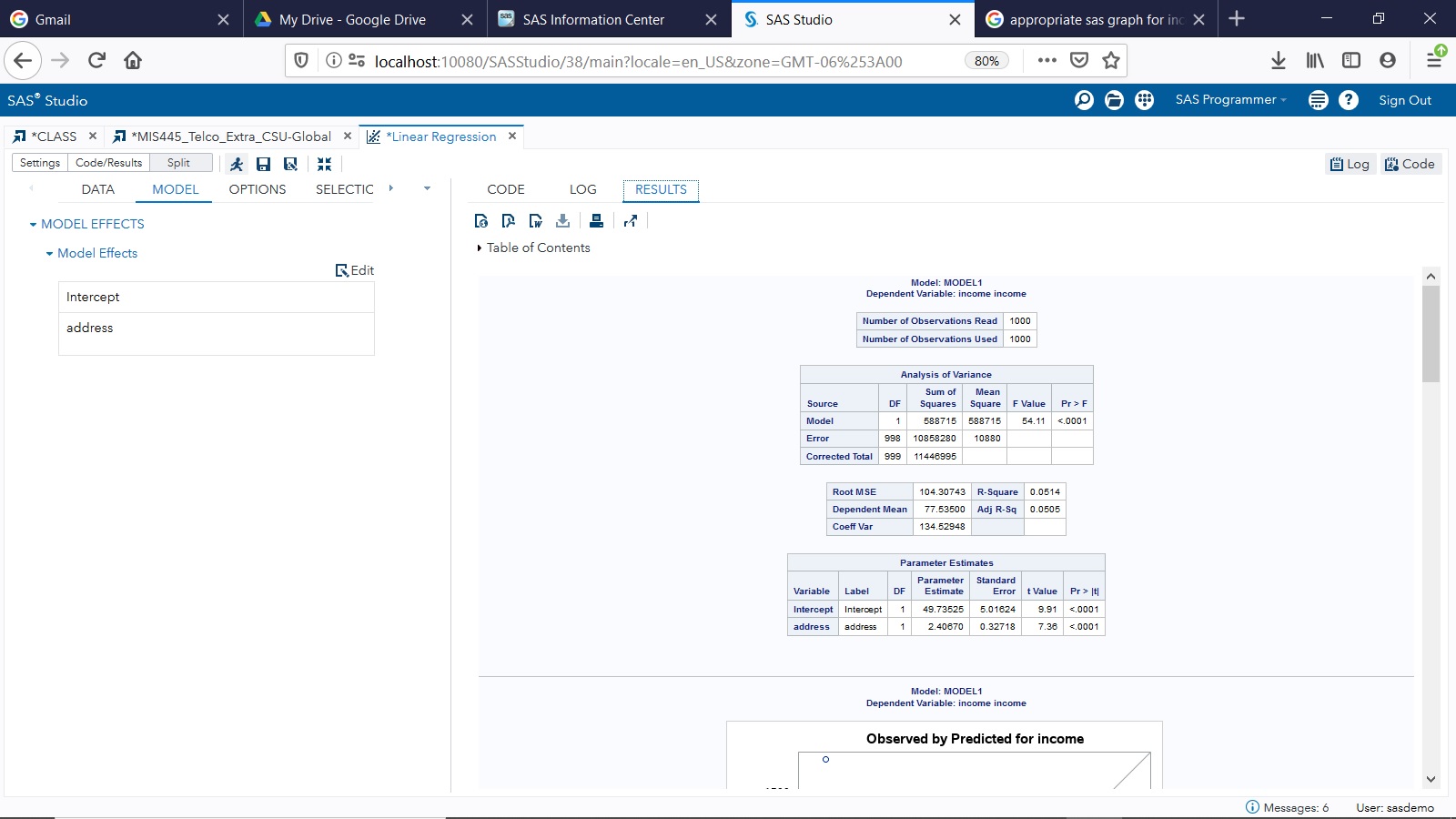
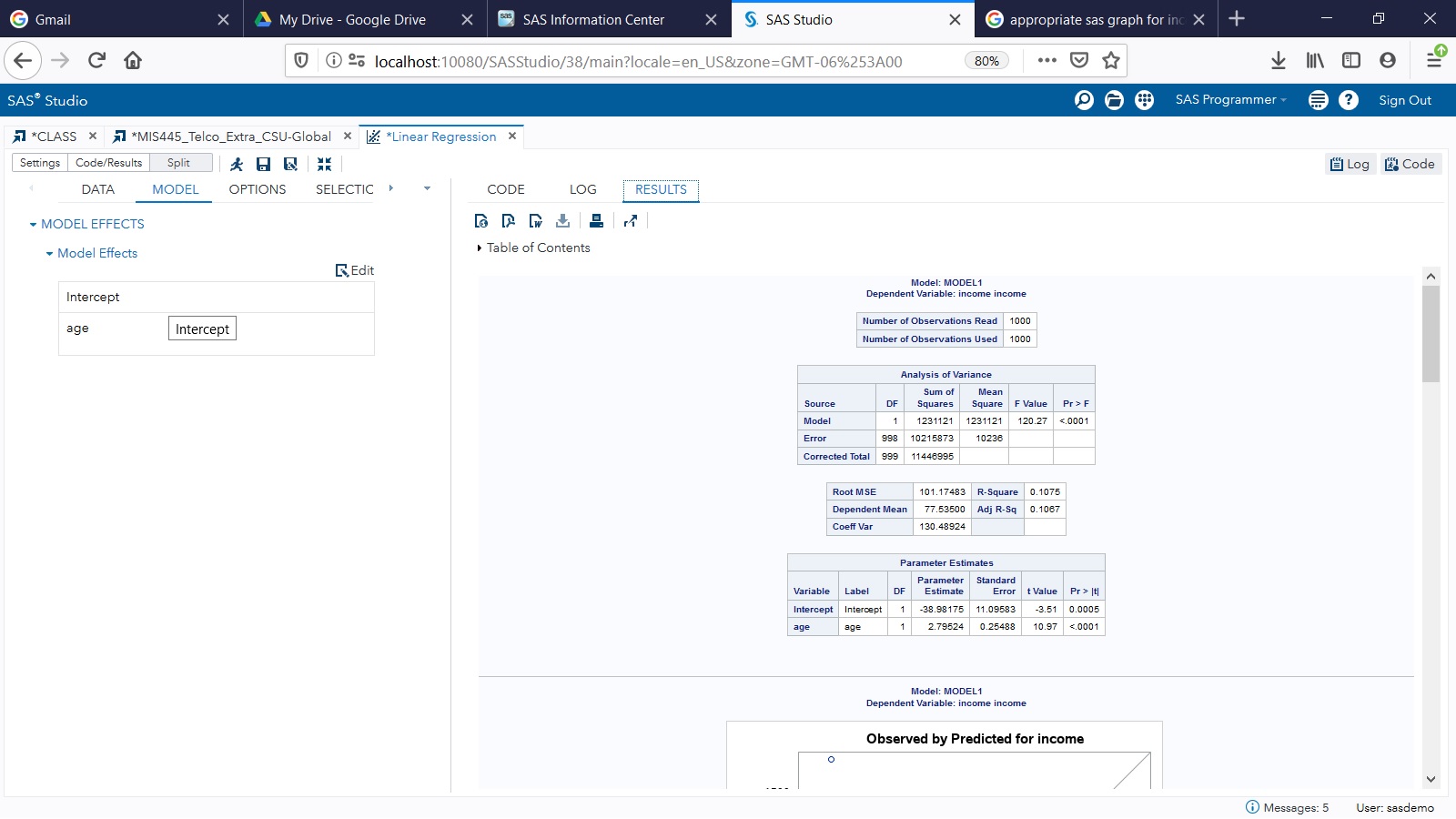
14.7.

Then, the estimated probability that a 27-years old person is going to churn is:

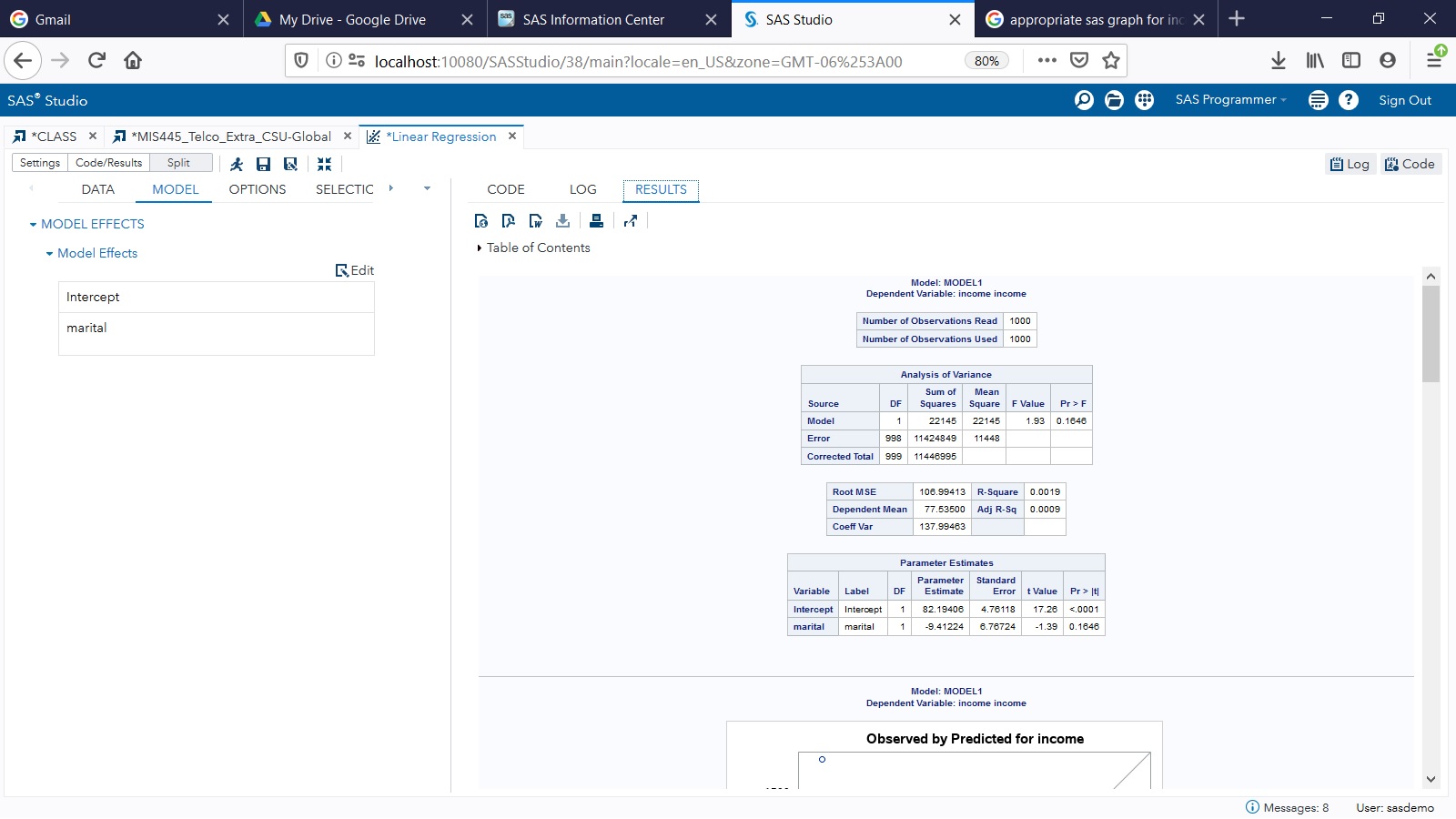
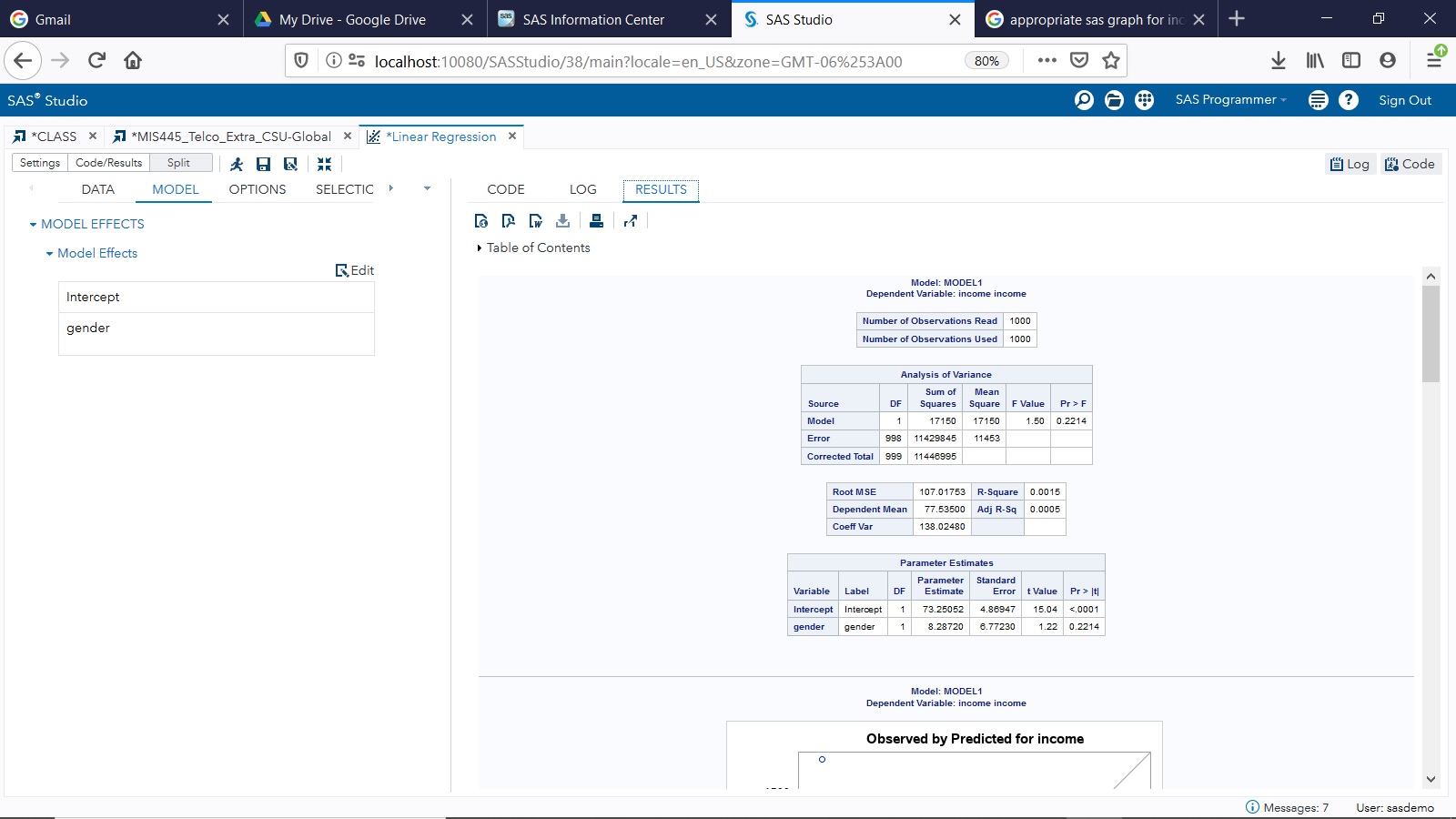
0.128.

So, the estimated probability that a 27-year old person is going to churn is 12.8% (Calculator.net, n.d.).

**Linear Regression Analysis**

****

**Figure 27 and 28 above**

****

**Figure 29 and 30 above**

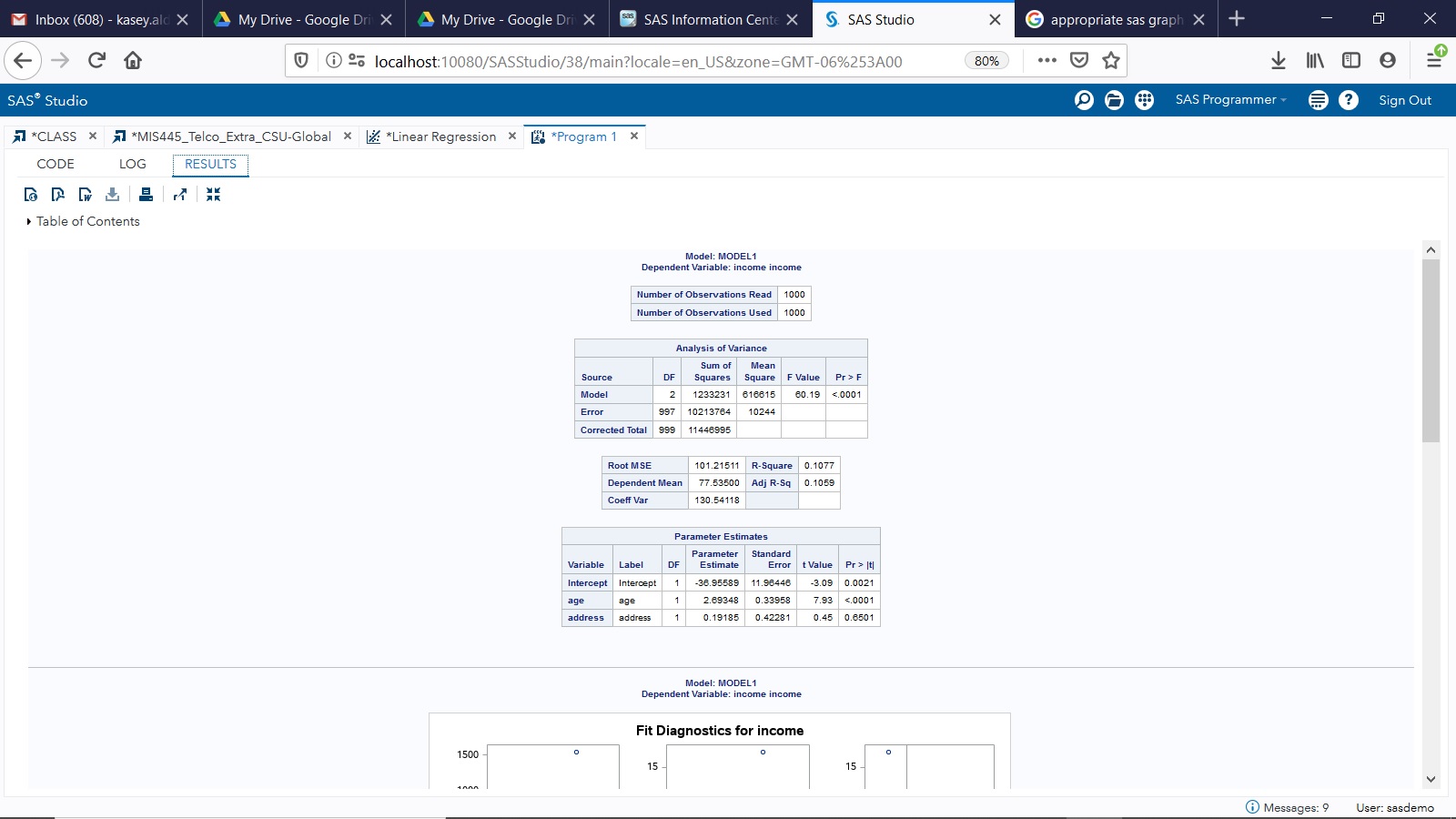
**Multiple Linear Regression Analysis – Dependent Variable: Income**

The income variable is the next dependent variable that will be tested. First, we will test churn as a dependent variable against the independent variable ‘age’. The second one that was chosen was years at an address. Both tests will check to see if any relationship exists or can be predicted (Tomoiaga, 2019).

**Hypothesis**

H0: ; In other words, the variable is not a significant predictor for our model;

H1: In other words, the variable is significant for our model.

****

**Figure 31 above**

**Interpretation**

The interpretation for is , then *11% of the variability of the variability of the dependent variable is explained by the variability of the independent variables* OR *11% of the variability of the dependent variable is explained by the regression model.* This small number seems to represent a weak positive linear relationship (Tomoiaga, 2019).

The hypothesis test is concluded by comparing the *p*-value (which is .0001 from figure 31 above) with the significance level α=0.05. Since the p-value is less than the significance level, then we reject the null hypothesis. In other words the variables chosen are significant for our model (Tomoiaga, 2019).

Hypothesis test for the significance of Age:

H0:

H1:

Hypothesis test for the significance of Address:

H0:

H1:

In both cases the p-value corresponds to each predictor is less than the significance level a= 0.05, so each predictor is significant for our model, so we validate the model and we may write the estimate of the regression equation as: income = b0 + b1 \* Age + b2 \* Address, where are the estimates of and their values are given in the column **Parameter Estimate.**

Unfortunately, I was unable to run the regression code in SAS due to the computations being terminated because the number of response levels, 218, exceeds the MAXRESPONSELEVELS=100. It seemed to be based on the income since both variables had the same result so I was unsure where to go from there.

Conclusion

This was a very interesting assignment and I hope that I can utilize this later on in my career. Knowing why people are leaving your company can be an extremely useful thing to know. Looking at the vast amount of information that was available in the file I can see how many can not make heads or tails out of these results, many of which are much higher than the amount in the Telco file.

References

Babu, S., & Ananthanarayanan, N. A. (2016). A Review on Customer Churn Prediction in Telecommunication Using Data Mining Techniques. *International Journal of Scientific Engineering and Research (IJSER)*, *4*(1), 35-40. Retrieved from https://pdfs.semanticscholar.org/ce77/0127176e52b10c7619eb81386cf1711f701c.pdf

Calculator.net. (n.d.). Exponent Calculator. Retrieved from https://www.calculator.net/exponent-calculator.html

Hood, K., & Green, J. (2006). Introduction to regression and data analysis Statlab Workshop. Retrieved from Yale University website: http://statlab.stat.yale.edu/help/workshops/IntroRegression/StatLabIntroRegression2006.pdf

Jamal, Z., & Bucklin, R. E. (2006). Improving the diagnosis and prediction of customer churn: A heterogeneous hazard modeling approach. *Journal of Interactive Marketing*, *20*(3-4), 16-29. Retrieved from http://eds.b.ebscohost.com.csuglobal.idm.oclc.org/eds/pdfviewer/pdfviewer?vid=1&sid=70e0536f-b023-48a4-8320-3f04b570131a%40sessionmgr103

SAS. (n.d.). Creating a Histogram in SAS Studio. Retrieved from https://video.sas.com/detail/videos/how-to-tutorials/video/4573023402001/creating-a-histogram-in-sas-studio?autoStart=true&page=2

Sturivant, R. X., Pardoe, I., Berrier, J., Watts, K., Vahid, F., Chan, C., & Nestler, S. (2016). Linear regression. In *Statistics for business analytics* [RedShelf]. Retrieved from https://csuglobal.instructure.com/courses/15535/external\_tools/21445

Tomoiaga, A. (2019). *Option 1 A theoretical approach over correlation and regression analysis* [Word Document]. Retrieved from CTA6\_template.docx