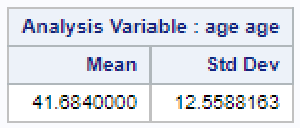
Introduction

This analysis is going to go through data that has been given by a telecommunications company that wants to increase its insight into the customers that it already has, and how it can reduce churn. The first objective is finding out what correlations there may be in the data relating to customers who churn, according to Ahmad, Jafar, and Aljoumaa (2019), being able to predict the what customers will churn, especially in the early stages, can be a large revenue source in it of itself. Finding these correlations will help the organization better predict what customers will churn, and how they might be able to counteract churn in the future. The second objective is to analyze data that has been boughten from an outside source, regarding the income of its customers. The ultimate goals for this, is to have the ability to predict the income of its customers and no longer need to purchase this data from an outside source. For this problem, a regression model will need to be made. Preferably with multiple independent variables that the organization already has accumulated data for, so that the dependent variable, income, will be easily predictable.

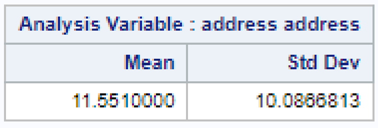
This analysis is going to go through nine different variables given from the telecommunications company, and look for possible correlations and useable information that can be pulled from the data. In the first section, the nine variables will be analyzed and set alongside graphs to get a visual representation of that data that is being worked with. The second section is going to go over the potential correlations with churn, trying to find a significant predictor for the variable churn. This will be done by attempting to make a logistic regression model with a different variable as the independent variable, and churn as the dependent variable, then validating the model. The third section will be similar to the second, but with multiple regression model, and using income as the dependent variable. This again will use the same technique of trying to make the model and then validate it. If the models are validated, then they are ready to make predictions, if they are not validated, then new independent variables must be found that can be validated as significant predictors.

Part 1 – Variable Analysis

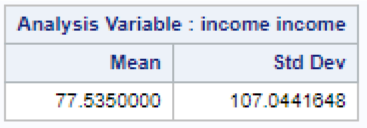
Here are the three continuous variables' statistics, with their interpretations and graphs following:



Analysis Variable : age age	
Mean	Std Dev
41.6840000	12.5588163

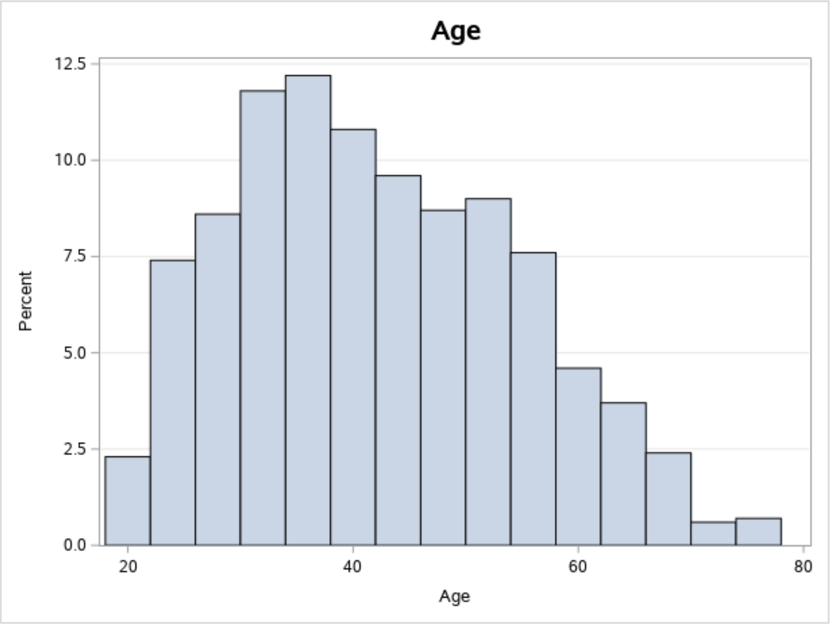


Analysis Variable : address address	
Mean	Std Dev
11.5510000	10.0866813

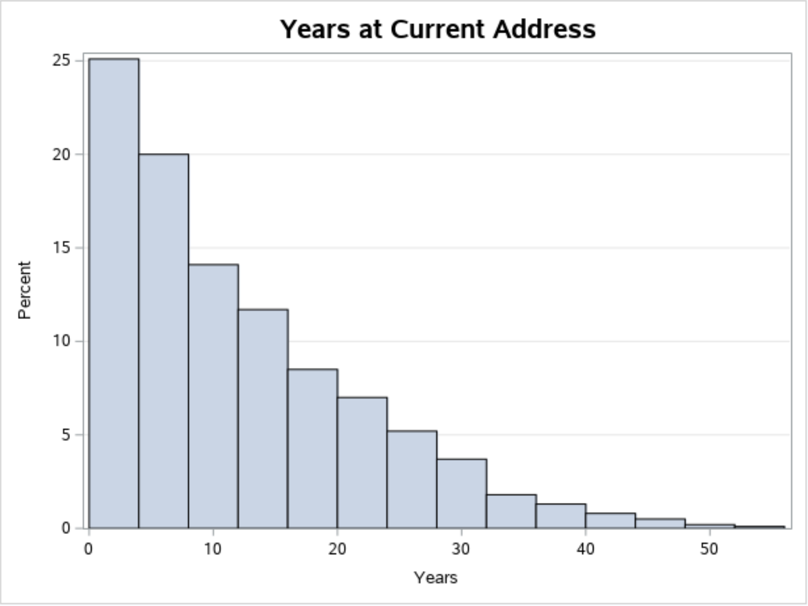


Analysis Variable : income income	
Mean	Std Dev
77.5350000	107.0441648

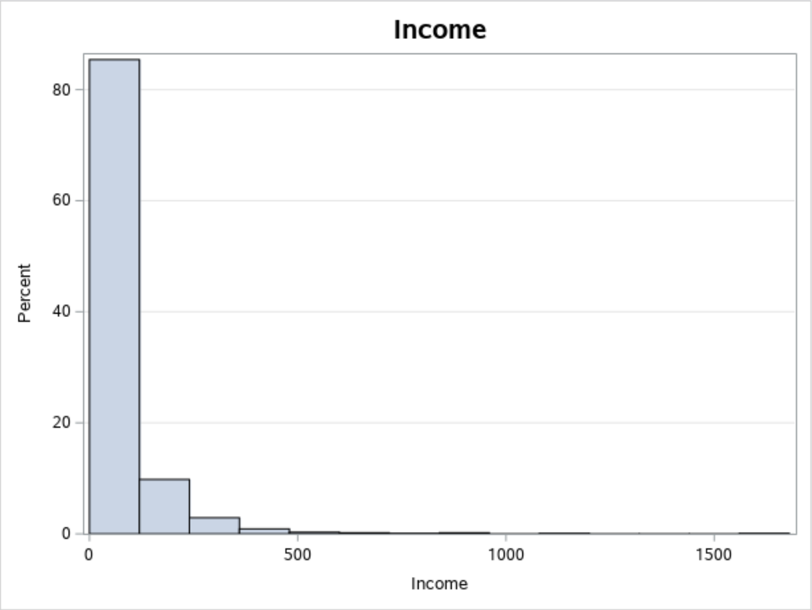
Age: On average, a telecommunications customer is about 41.7 years old within the given sample from TELCO. On average, a person's age deviates from 41.7 years old (mean) by about 12.6 years. Here is a graph to represent this data:



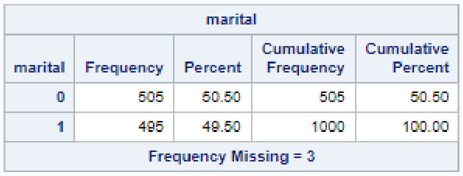
Years at Current Address: On average, each customer lives at their current address for about 11.55 years. On average, each customer's time spent at their current address deviates from 11.55 years (mean) by about 10.08 years. Here is a graph to represent this data:



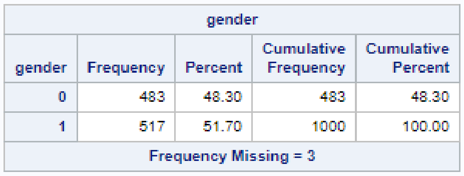
Income: On average, each customer makes about $77,535 annually. On average, each customer's income deviates from the mean annual income of $77,535, by $107,000 annually. Here is a graph to represent this data:



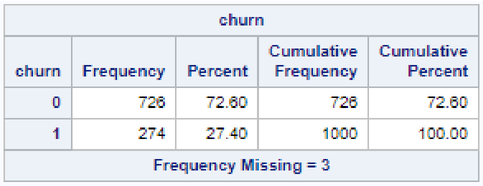
Here are three categorical variables' statistics that only have two categories, with their interpretations and graphs following:



marital				
marital	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	505	50.50	505	50.50
1	495	49.50	1000	100.00
Frequency Missing = 3				

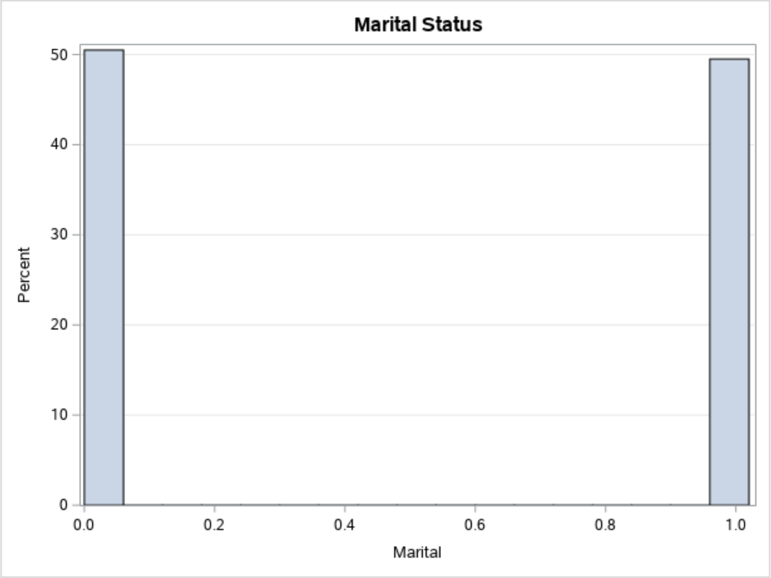


gender				
gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	483	48.30	483	48.30
1	517	51.70	1000	100.00
Frequency Missing = 3				

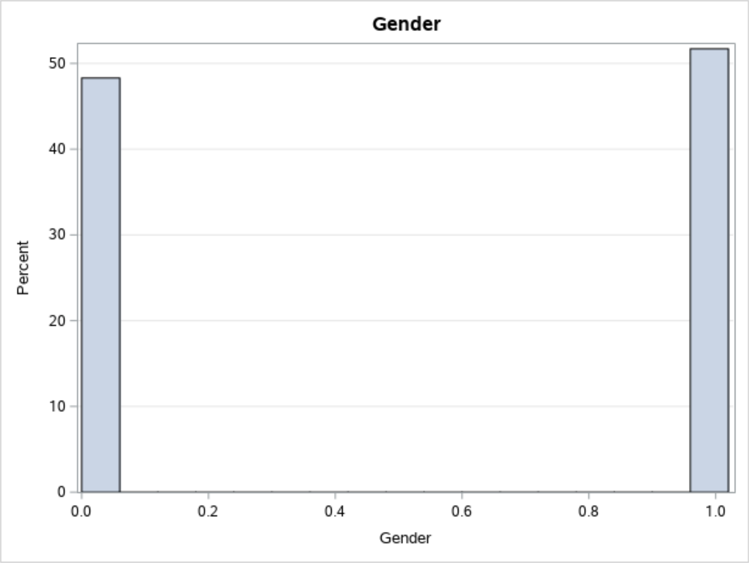


churn				
churn	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	726	72.60	726	72.60
1	274	27.40	1000	100.00
Frequency Missing = 3				

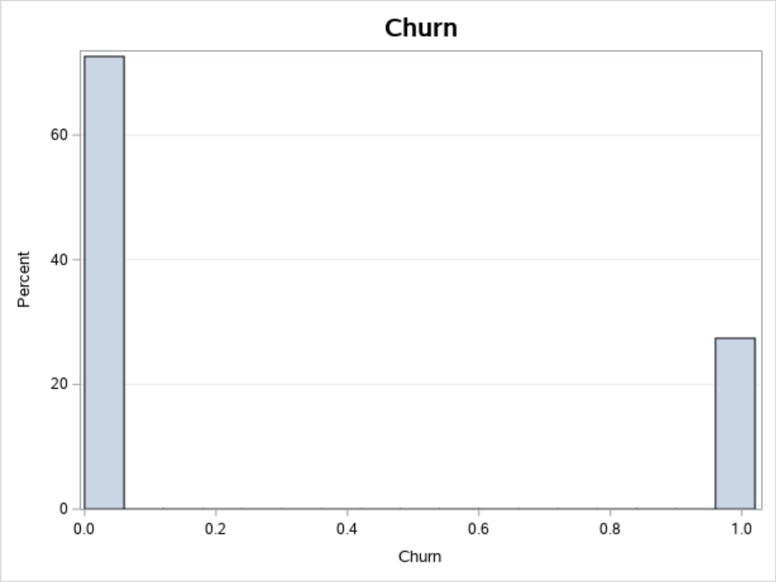
Marital: 50.5% of the customers are married, while 49.5% are unmarried. Here is a graph to represent this data:



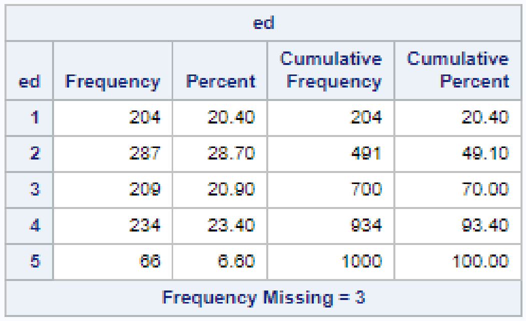
Gender: 48.3% of the customers are males, while 51.7% are females. Here is a graph to represent this data:



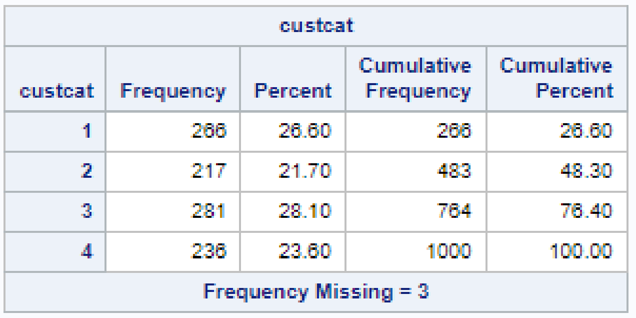
Churn: 72.6% of the customers are not going to churn, while 27.4% are going to churn. Here is a graph to represent this data:



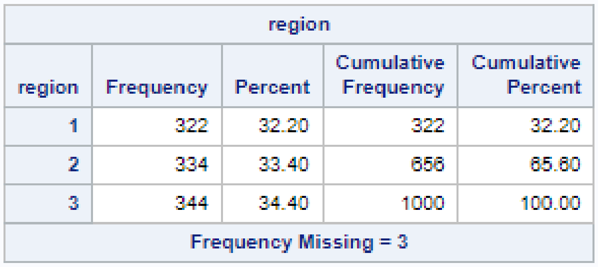
Here are the final three categorical variables' statistics that only have multiple categories, with their interpretations and graphs following:



ed				
ed	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	204	20.40	204	20.40
2	287	28.70	491	49.10
3	209	20.90	700	70.00
4	234	23.40	934	93.40
5	66	6.60	1000	100.00
Frequency Missing = 3				

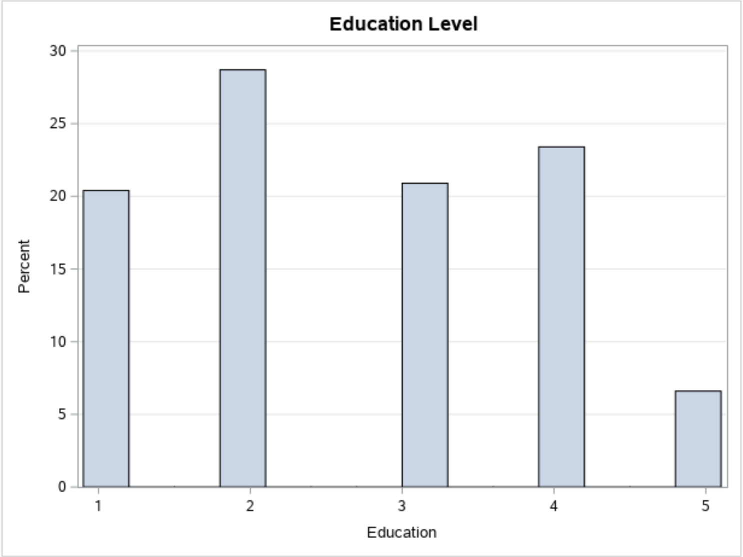


custcat				
custcat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	266	26.60	266	26.60
2	217	21.70	483	48.30
3	281	28.10	764	76.40
4	236	23.60	1000	100.00
Frequency Missing = 3				

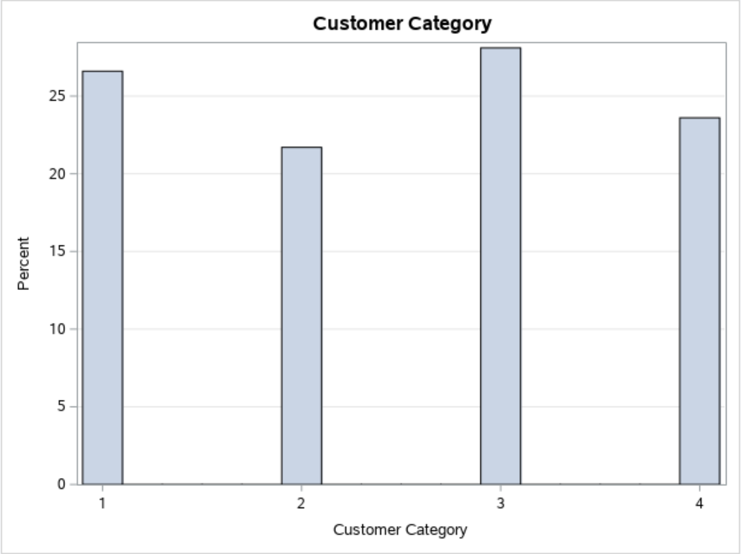


region				
region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	322	32.20	322	32.20
2	334	33.40	656	65.60
3	344	34.40	1000	100.00
Frequency Missing = 3				

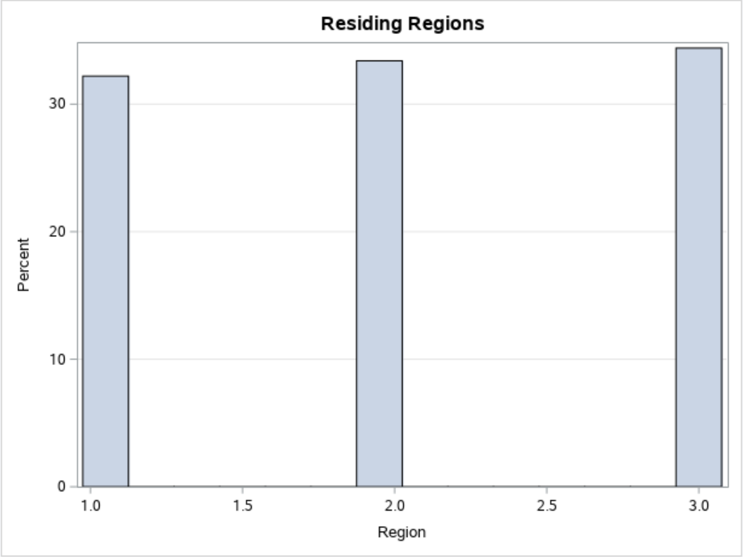
Education: 20.4% of customers did not complete high school, 28.7% of customers have a high school diploma, 20.9% have some college experience, 23.45 have a college degree, and 6.6% have an undergrad degree. Here is a graph to represent this data:



Customer Category: 26.6% of customers have the basic service, 21.7% have the E-service, 28.1% have the Plus service, 23.6% have the Total service. Here is a graph to represent this data:



Region: 32.3% of customers live in Zone 1, 33.4% live in Zone 2, 34.4% live in Zone 3. Here is a graph to represent this data:



Part 2 - Logistic Regression Model using Churn

In this model I will be building a model with Churn as my dependent variable, and will be trying to use Gender as the independent variable. A "logistic regression model allows the analysis of dichotomous or binary outcomes with 2 mutually exclusive levels" and "permits the use of continuous or categorical predictors"(LaValley, 2008), making it the perfect model to analyze the odds of churn with most of the nine variables that have been analyzed in part 1. After the model is built, it needs to be determined whether or not gender is a significant predictor of Churn. If this is validated, then it can be used as the predictor for the model. On the other hand, if it is not validated as a significant predictor, then a new variable will need to be found. The regression equation for this would be:

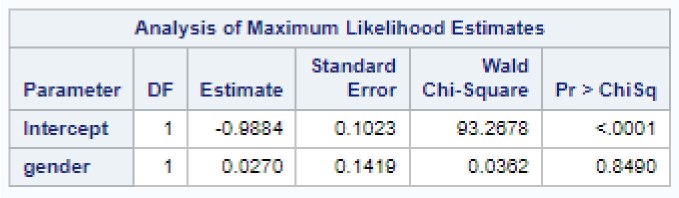
Log (odds to churn) = B0 + B1*Gender (a modified slope-point equation to find the slope of the correlation between the two variables)

The hypotheses are:

H0: B1 = 0, there is no correlation between the two variables and gender is not a significant predictor.

H1: B1 != 0, there is a correlation between the two variables and gender is a significant predictor.

We can then look at the Analysis of Maximum Likelihood Estimates Table to see if the p-value is less than the significance level of 0.05. Here is the table given from SAS:



Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.9884	0.1023	93.2678	<.0001
gender	1	0.0270	0.1419	0.0362	0.8490

As we can see from the Pr > ChiSq column, the p-value is .849 which is greater than 0.05, meaning that there is not a significant enough correlation between the two variables, and gender is not a significant predictor of Churn. This means that it is necessary to start over again and find a variable that is a significant predictor, and this time we can try Years at Address. This is a continuous variable rather than categorical dummy variable that is either one answer or another. The process needs to be started again from the beginning, so here is the regression equation for Years at Address and Churn:

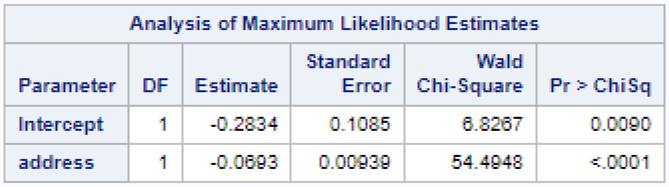
Log (odds to churn) = B0 + B1 * Address

The hypotheses are:

H0: B1 = 0, there is no correlation between the two variables and address is not a significant predictor.

H1: B1 != 0, there is a correlation between the two variables and address is a significant predictor.

We can again look at the Analysis of Maximum Likelihood Estimates Table to see if the p-value is less than the significance level of 0.05, and here is the table given from SAS:



Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.2834	0.1085	6.8267	0.0090
address	1	-0.0693	0.00939	54.4948	<.0001

We can again look at the Pr > ChiSq column to find the p-value and compare it to the significance level of 0.05. As we can see from the table, the p-value is equal to <.0001 which is less than 0.05. This means that we can then go ahead and reject the null hypothesis, and can conclude that Years at Address is a significant predictor of Churn.

Since we now know that we can use address within our logistic model, we can rewrite our regression equation as:

Log (odds to churn) = b0 + b1 * address

It is important to signify the change from B to b in these equations as these are all just estimations that SAS is providing, as there is no way that SAS can retrieve the actual slope and intercept from the data given. We can now pull the estimates from the table above and plug them in to make our equation:

Log (odds to churn) = -.2834 + -.0693(address)

Odds to churn = *e* ^(-.2834 + -.0693(address)) (simplified)

Now we can do two different predictions to make sure our equation makes sense. From what the equation says, the estimated intercept is at -.2834, with a slope of -.0693. We can look back at our Years at Address table in section 2 and see that the mean for Address was 11.55, so we can use that as our first number.

Odds to churn = *e* ^(-.2834 + -.0693(11.55))

Odds to churn = .3383

This means the estimated probability that somebody will churn, that has also lived at their current address for 11.55 years is equal to 33.8%

For the second number we can go up to 15 as the standard deviation is 10.08, so we are still well within the given data. Since we saw that the slope for the odds to churn go down when the Years at Address go up, then we should get a percentage that is lower than the one that we found at 11.55.

Odds to churn = *e* ^(-.2834 + -.0693(15))

Odds to churn = .2663

This means the estimated probability that somebody will churn, that has also lived at their current address for 15 years has gone down to 26.6%. This is what we had predicted when looking at the equation and the previous answer from the mean we initially input into the equation.

Part 3 - Multiple Regression Analysis using Income

Next is the multiple regression model using income. A multiple linear regression model is a model that can show what the dependent variable does when the independent variable(s) change (Zybooks, 2018). Since we had good luck using a continuous variable on the previous regression equation, we can again go for the two available continuous variables, age and address (since income is the other, but unavailable because it is our dependent variable). The regression equation for this model would be:

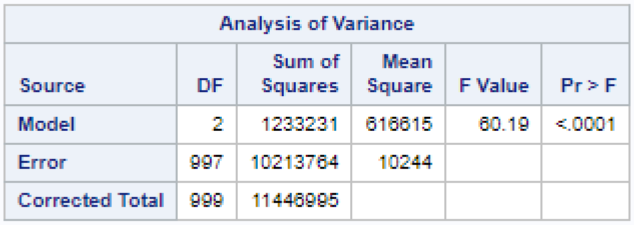
Income = B0 + B1 * Address + B2 * Age

The first test to be done to validate the significance would be to test the overall significance, and here are the hypotheses:

H0: B0 = B1 = B2, or none of the predictors used are significant.

H1: B0 != B1 and/or B1 != B2 and/or B0 != B2, or there is a significant predictor in the group.

Here is the Analysis of Variance table that SAS has produced:



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1233231	616615	60.19	<.0001
Error	997	10213764	10244		
Corrected Total	999	11446995			

We then compare the p-value to our significance level of 0.05, and again if it is less than that, we get to reject the null hypothesis and move on to the next test. As you can see in Pr > F column, the p-value is <.0001 which is less than 0.05, and the null hypothesis is rejected, meaning at least one of the predictors is significant. The next step is to test for which variable is significant, here is the hypotheses for the address test, as well as the age test:

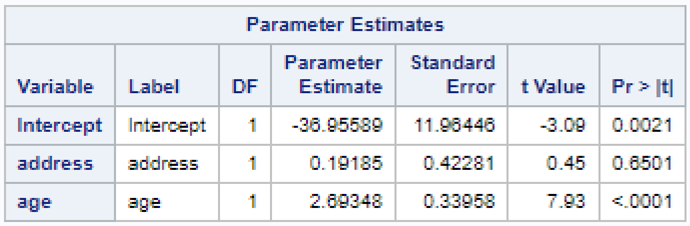
H0: B1 = 0, or address is not a significant predictor of income.

H1: B1 != 0, or address is a significant predictor of income.

H0: B2 = 0, or age is not a significant predictor of income.

H1: B2 != 0, or age is a significant predictor of income.

To test these hypotheses, we must look at the parameters estimates table:



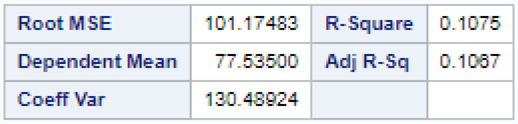
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-36.95589	11.96446	-3.09	0.0021
address	address	1	0.19185	0.42281	0.45	0.6501
age	age	1	2.69348	0.33958	7.93	<.0001

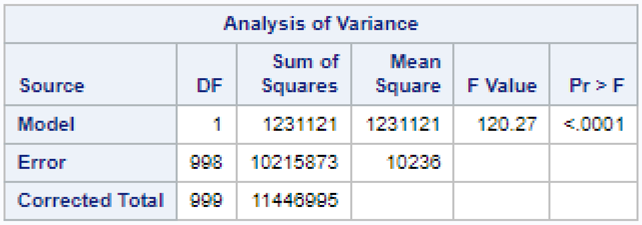
For the first test, or address test, we can compare the value found in the Pr > |t| column and address row to the significance value. The p-value found is .65, which is greater than 0.05, meaning that the null hypothesis is not rejected, and this is not a significant predictor of income.

For the second test, or age test, we can again compare the value found in the Pr > |t| column and age row to the significance value. The p-value found is <.0001 which is less than 0.05 meaning that the null hypothesis is rejected, and that age is a significant predictor of income.

Since we have now found one predictor that has been validated, we can move forward with the creation of the model, by dropping the insignificant predictor and keeping the significant one. This effectively changes the equation from a multiple linear regression model to a single linear regression model, that still needs to be validated. For this we can change the code in SAS to only include the age as an independent variable, and then pull the table containing r^2, and analysis of variance tables.



Root MSE	101.17483	R-Square	0.1075
Dependent Mean	77.53500	Adj R-Sq	0.1067
Coeff Var	130.48924		



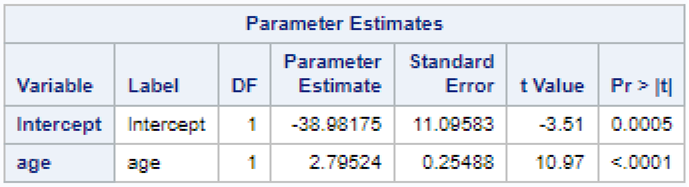
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1231121	1231121	120.27	<.0001
Error	998	10215873	10236		
Corrected Total	999	11446995			

We can start analyzing the model by looking at the r^2 value, or the coefficient of determination. This value tells how well the model fits the data, and if it is above .07, then it is a good fit. Luckily, the r^2 value is .1075, which is greater than .07, meaning that there should be a strong chance the regression model will be validated. The next step is to create the hypotheses and compare the F-statistic to the critical value of 5.039, and if the F-statistic is greater than the critical value, we can reject the null hypothesis.

H0: B1 = 0, or the slope of the regression equation is 0.

H1: B1 != 0, or the slope of the regression equation is not 0.

The F-value found on the analysis of variance table is 120.27, and this is greater than the critical value of 5.039, meaning we can reject the null hypothesis and validate that there is a slope to this regression equation. Since this has been validated, we can now read the parameter estimates table and write out our equation:



Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-38.98175	11.09583	-3.51	0.0005
age	age	1	2.79524	0.25488	10.97	<.0001

Income = b0 + b1 * age, with b0 = -38.98 and b1 = 2.79 therefor the validated equation is:

Income = -38.98 + 2.79(age)

We will again make two predictions to further validate our equation and make sure that the equation is doing what we think. Again, we can revert to the mean of the age, which is 41.7 and put it into the equation:

Income = -38.98 + 2.79(41.7)

Income = 77.363

This means that a customer that is 41.7 years old is estimated to be making around $77,363 annually.

Since we can look at our equation again and see a positive slope this time, we can assume that if we predict income for a customer that is older than 41.7, then there should be a growth in their estimated income. The standard deviation for age is 12.6, so we can go up one standard deviation from the mean for our next prediction, to 54.3 and put this into the equation:

Income = -38.98 + 2.79(54.3)

Income = 112.517

This means that a customer that is 54.3 years old is estimated to be making around $112,517 annually, and this has gone up, just like we had predicted based on the equation.

Conclusion

The first objective of finding more insight to correlation as to what makes customers churn has been found through the use of the first regression model. The first option of gender was chosen to see if there may be a strange correlation between churn and gender that was unlikely, but possible. This turned out to be a no or at least insignificant correlation between the two. The next option was Years at Address, which seemed to me as if there would again be little to no correlation, but as it turns out, the longer customers have lived at their current address can be related to how likely they are to churn. This surprised me, but could be seen as useful information to give the organization as it moves forward, to make smart business decisions around.

The second objective has also been achieved through the second regression model. The hopes of getting a multiple linear regression model didn't work out, but a strong single linear model was created. The two independent variables chosen were years at address and age. I was hoping to be surprised again by years at address being significant but it failed to validate, leaving the model with only age as a validated predictor. This will certainly help out the organization predict income of customers in the future without needing to purchase this information from an outside source.

References

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019, March 20). *Customer churn prediction in telecom using machine learning in big data platform*. Journal of Big Data. https://link.springer.com/article/10.1186/s40537-019-0191-6.

LaValley, M. P. (2008, May 6). Logistic Regression. Circulation. https://www.ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.106.682658.

ZyBooks (2018). Statistics in Business Analytics. ZyBooks.