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

Analysis Variable : age age									
Mean	Std Dev	Minimum	Maximum	N	Variance	Mode	Range	Lower Quartile	Upper Quartile
41.6840000	12.5588163	18.0000000	77.0000000	1000	157.7238679	33.0000000	59.0000000	32.0000000	51.0000000

Figure 1 above

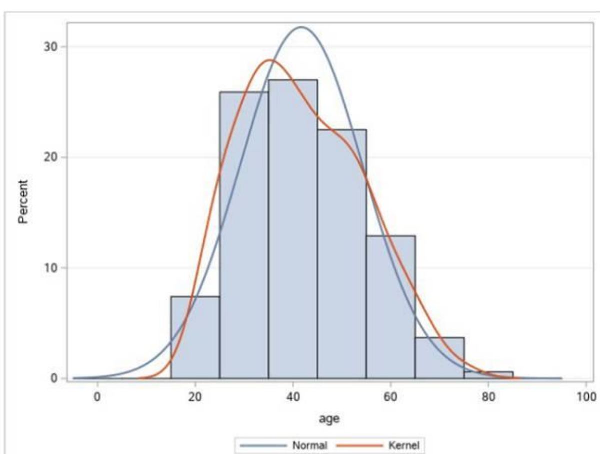

Figure 2 above

(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

Analysis Variable : address address									
Mean	Std Dev	Minimum	Maximum	N	Variance	Mode	Range	Lower Quartile	Upper Quartile
11.5510000	10.0866813	0	55.0000000	1000	101.7411401	1.0000000	55.0000000	3.0000000	18.0000000

Figure 3 above

Address at Current Address Graph

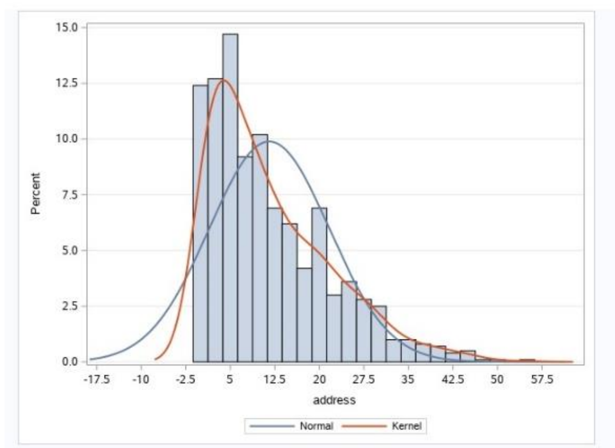

Figure 4 above

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

Analysis Variable : income income									
Mean	Std Dev	Minimum	Maximum	N	Variance	Mode	Range	Lower Quartile	Upper Quartile
77.5350000	107.0441648	9.0000000	1668.00	1000	11458.45	25.0000000	1659.00	29.0000000	83.0000000

Figure 5 above

Income Graph

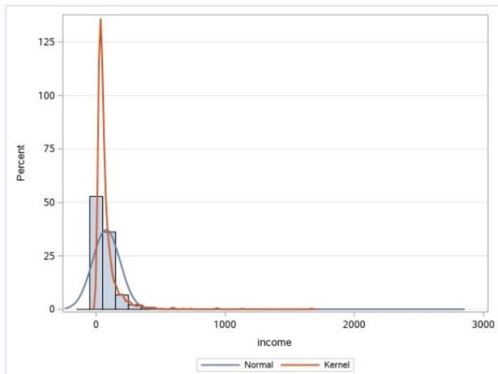

Figure 6 above

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

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

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

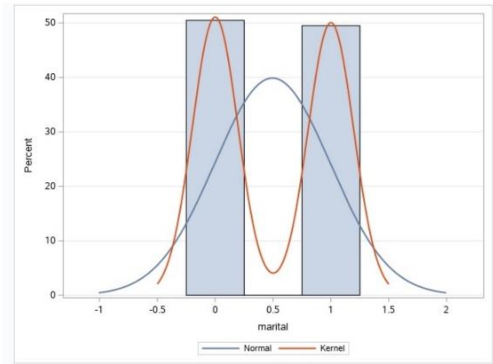

Figure 8 above

Sample Characteristics Gender

The FREQ Procedure

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

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

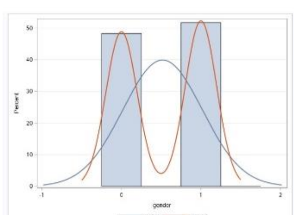

Figure 10 above

Sample Characteristics Churn

The FREQ Procedure

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

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

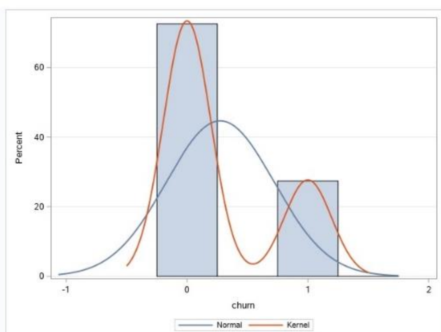

Figure 12 above

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

Sample Characteristics Region

The MEANS Procedure

Analysis Variable : region region									
Mean	Std Dev	Minimum	Maximum	N	Variance	Mode	Range	Lower Quartile	Upper Quartile
2.0220000	0.8161998	1.0000000	3.0000000	1000	0.6661822	3.0000000	2.0000000	1.0000000	3.0000000

Figure 13 above

The FREQ Procedure

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

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

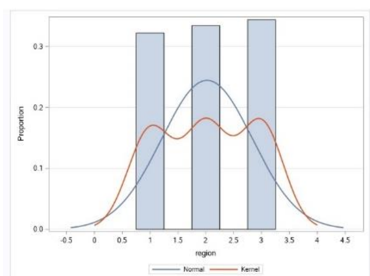

Figure 15 above

Sample Characteristics Level of Education

Analysis Variable : ed ed									
Mean	Std Dev	Minimum	Maximum	N	Variance	Mode	Range	Lower Quartile	Upper Quartile
2.6710000	1.2223965	1.0000000	5.0000000	1000	1.4942533	2.0000000	4.0000000	2.0000000	4.0000000

Figure 16 above

The FREQ Procedure

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

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

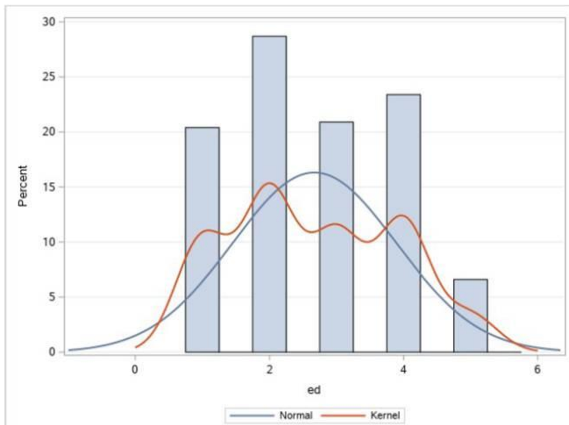

Figure 18 above

Sample Characteristics Customer Category

Analysis Variable : custcat custcat									
Mean	Std Dev	Minimum	Maximum	N	Variance	Mode	Range	Lower Quartile	Upper Quartile
2.4870000	1.1203062	1.0000000	4.0000000	1000	1.2550861	3.0000000	3.0000000	1.0000000	3.0000000

Figure 19 above

The FREQ Procedure

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

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

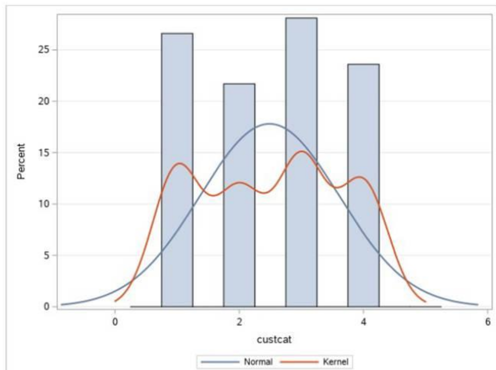

Figure 21 above

Linear Regression Analysis

Model: MODEL1
Dependent Variable: churn churn

Number of Observations Read	1000
Number of Observations Used	1000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	12.71468	12.71468	68.15	<.0001
Error	998	186.20932	0.18658		
Corrected Total	999	198.92400			

Root MSE	0.43195	R-Square	0.0639
Dependent Mean	0.27400	Adj R-Sq	0.0630
Coeff Var	157.64670		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.64845	0.04737	13.69	<.0001
age	age	1	-0.00898	0.00109	-8.26	<.0001

Model: MODEL1
Dependent Variable: churn churn

Number of Observations Read	1000
Number of Observations Used	1000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	11.75324	11.75324	62.67	<.0001
Error	998	187.17076	0.18755		
Corrected Total	999	198.92400			

Root MSE	0.43307	R-Square	0.0591
Dependent Mean	0.27400	Adj R-Sq	0.0581
Coeff Var	158.05315		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.39821	0.02083	19.12	<.0001
address	address	1	-0.01075	0.00136	-7.92	<.0001

Figure 22 and 23 above

Model: MODEL1
Dependent Variable: churn churn

Number of Observations Read	1000
Number of Observations Used	1000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.00721	0.00721	0.04	0.8492
Error	998	198.91679	0.19932		
Corrected Total	999	198.92400			

Root MSE	0.44645	R-Square	0.0000
Dependent Mean	0.27400	Adj R-Sq	-0.0010
Coeff Var	162.93706		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.27122	0.02031	13.35	<.0001
gender	gender	1	0.00537	0.02825	0.19	0.8492

Model: MODEL1
Dependent Variable: churn churn

Number of Observations Read	1000
Number of Observations Used	1000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.00721	0.00721	0.04	0.8492
Error	998	198.91679	0.19932		
Corrected Total	999	198.92400			

Root MSE	0.44645	R-Square	0.0000
Dependent Mean	0.27400	Adj R-Sq	-0.0010
Coeff Var	162.93706		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.27122	0.02031	13.35	<.0001
gender	gender	1	0.00537	0.02825	0.19	0.8492

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: $\beta_1 = 0$; In other words, the variable X_1 is not a significant predictor for our model;

H1: $\beta_1 \neq 0$; In other words, the variable X_1 is significant for our model.

Model: MODEL1
Dependent Variable: churn churn

Number of Observations Read	1000
Number of Observations Used	1000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	14.75798	7.37899	39.95	<.0001
Error	997	184.16602	0.18472		
Corrected Total	999	198.92400			

Root MSE	0.42979	R-Square	0.0742
Dependent Mean	0.27400	Adj R-Sq	0.0723
Coeff Var	156.85798		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.58539	0.05080	11.52	<.0001
age	age	1	-0.00582	0.00144	-4.03	<.0001
address	address	1	-0.00597	0.00180	-3.33	0.0009

Figure 26 above

Interpretation

The interpretation for r^2 is $r^2 = 0.07$, 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: $\beta_{1=}0; in\ other\ words\ Age\ is\ not\ a\ significant\ predictor\ for\ our\ model.$

H1: $\beta_1\ \neq 0; in\ other\ words\ Age\ is\ \ a\ significant\ predictor\ for\ our\ model.$

Hypothesis test for the significance of Address:

H0: $\beta_{2=}0; in\ other\ words\ Address\ is\ not\ a\ significant\ predictor\ for\ our\ model.$

H1: $\beta_2\ \neq 0; in\ other\ words\ Address\ is\ a\ significant\ predictor\ for\ our\ model.$

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 = b_0 + b_1 * Age + b_2 * Address, where b_0, b_1, b_2 are the estimates of $\beta_{0,}\beta_1, \beta_2$ and their values are given in the column **Parameter Estimate** (Tomoiaga, 2019)**.**

Response Profile		
Ordered Value	churn	Total Frequency
1	1	274
2	0	726

Probability modeled is churn='1'.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1176.394	1110.296
SC	1181.301	1120.111
-2 Log L	1174.394	1106.296

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	68.0981	1	<.0001
Score	63.9173	1	<.0001
Wald	60.0945	1	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.0424	0.2599	16.0912	<.0001
age	1	-0.0505	0.00652	60.0945	<.0001

Figure 26a above

After running the regression code, the result is Probability modeled is churn='1' (figure 26a above). So, the regression equation is: $\log(odds\ to\ churn) = \beta_0 + \beta_1 * age$. 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: β_1=0; in other words, Age is not a significant predictor.

H1: $\beta_1 \neq 0$; 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: $\log(odds\ to\ churn) = \beta_0 + \beta_1 * age$.

After this we studied the Analysis of Maximum Likelihood Estimates (figure 26a above) which shows that the estimate of the regression equation is: $\log(odds\ to\ churn) = 1.0424 - 0.0505 * age$. So from there we get: odds to churn = $e^{1.0424\text{-}0.0505\text{*}age}$. So, since $e^{b_1} > 1.0424$, let's say e^{b_1} =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:

$$odds\ to\ churn = e^{1.0424-0.0505*27} = \quad 14.7.$$

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

$$p = \frac{odds\ to\ churn}{1+odds\ to\ churn} = \frac{0.147}{1.147} = 0.128.$$

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

Linear Regression Analysis

Model: MODEL1
Dependent Variable: income income

Number of Observations Read	1000
Number of Observations Used	1000

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			

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

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

Model: MODEL1
Dependent Variable: income income

Number of Observations Read	1000
Number of Observations Used	1000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	588715	588715	54.11	<.0001
Error	998	10858280	10880		
Corrected Total	999	11446995			

Root MSE	104.30743	R-Square	0.0514
Dependent Mean	77.53500	Adj R-Sq	0.0505
Coeff Var	134.52948		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	49.73525	5.01624	9.91	<.0001
address	address	1	2.40670	0.32718	7.36	<.0001

Figure 27 and 28 above

Model: MODEL1
Dependent Variable: income income

Number of Observations Read	1000
Number of Observations Used	1000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	17150	17150	1.50	0.2214
Error	998	11429845	11453		
Corrected Total	999	11446995			

Root MSE	107.01753	R-Square	0.0015
Dependent Mean	77.53500	Adj R-Sq	0.0005
Coeff Var	138.02480		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	73.25052	4.86947	15.04	<.0001
gender	gender	1	8.28720	6.77230	1.22	0.2214

Model: MODEL1
Dependent Variable: income income

Number of Observations Read	1000
Number of Observations Used	1000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	22145	22145	1.93	0.1646
Error	998	11424849	11448		
Corrected Total	999	11446995			

Root MSE	106.99413	R-Square	0.0019
Dependent Mean	77.53500	Adj R-Sq	0.0009
Coeff Var	137.99463		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	82.19406	4.76118	17.26	<.0001
marital	marital	1	-9.41224	6.76724	-1.39	0.1646

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: $\beta_1 = 0$; In other words, the variable X_1 is not a significant predictor for our model;

H1: $\beta_1 \neq 0$; In other words, the variable X_1 is significant for our model.

Model: MODEL1
Dependent Variable: income income

Number of Observations Read	1000
Number of Observations Used	1000

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			

Root MSE	101.21511	R-Square	0.1077
Dependent Mean	77.53500	Adj R-Sq	0.1059
Coeff Var	130.54118		

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

Figure 31 above

Interpretation

The interpretation for r^2 is $r^2 = 0.11$, 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: $\beta_{1=}0; in\ other\ words\ Age\ is\ not\ a\ significant\ predictor\ for\ our\ model.$

H1: $\beta_1 \neq 0; in\ other\ words\ Age\ is\ \ a\ significant\ predictor\ for\ our\ model.$

Hypothesis test for the significance of Address:

H0: $\beta_{2=}0; in\ other\ words\ Address\ is\ not\ a\ significant\ predictor\ for\ our\ model.$

H1: $\beta_2 \neq 0; in\ other\ words\ Address\ is\ a\ significant\ predictor\ for\ our\ model.$

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 = b_0 + b_1 * Age + b_2 * Address, where b_0, b_1, b_2 are the estimates of $\beta_{0,}\beta_1, \beta_2$ 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