**Portfolio Milestone: Income Prediction**

Kelsey Thompson

Colorado State University Global

MIS 445: Statistics in Business Analytics

Professor Mark Bateh

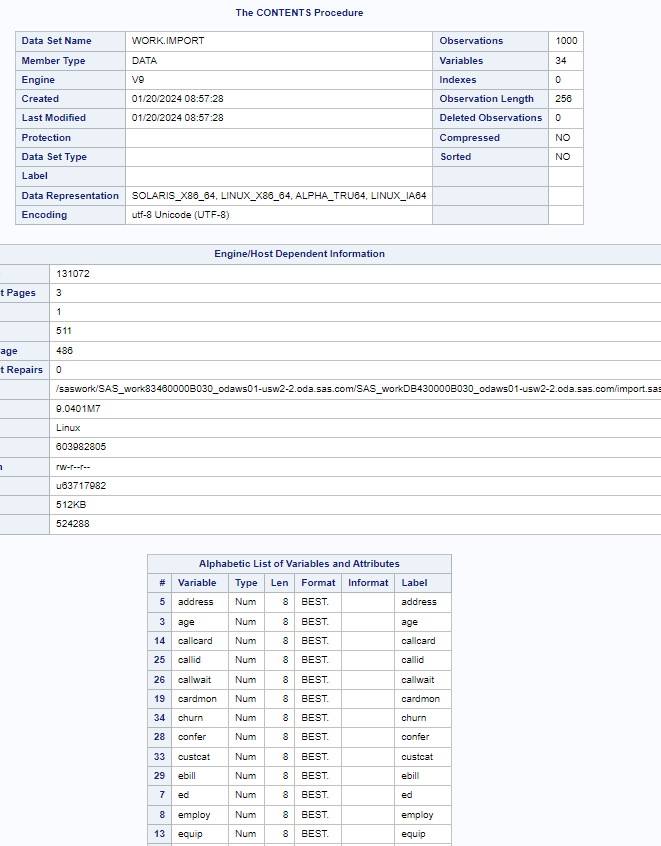
February 11, 2024

**Portfolio Milestone: Income Prediction**

1. An introduction for the purpose of my paper:

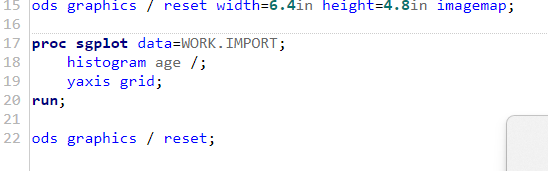
This project is centered around the dataset Telco Extra CSUGlobal.xls, containing customer information for a telecommunications company. The objective is to predict customer income, with current income data sourced externally.

To forecast customer income, I will utilize descriptive/summary statistics for each variable, conduct multiple linear regression analysis, establish hypotheses, and evaluate the strength and direction of the relationship between predictors and *churn* followed by *income*. The aim of this analysis is to assess the variables' significance on income.

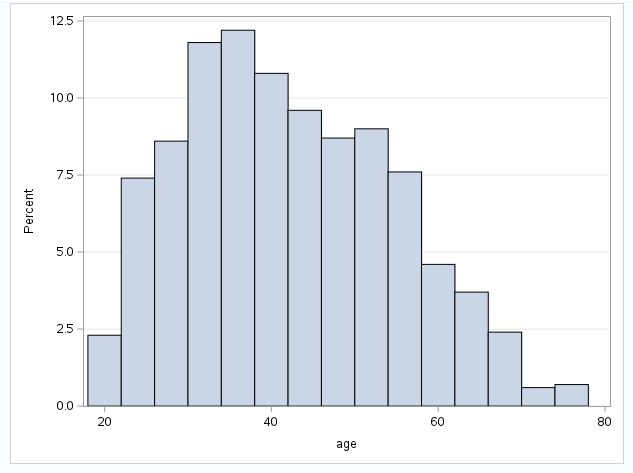
Commencing with the utilization of SAS, I accessed the file to initiate the analysis:

I focused on analyzing the variables *age, years* at current address, *gender*, level of *ed*ucation, *income, marital* status, *region, custcat*, and *churn*. The subsequent section presents the descriptive statistics results, including the appropriate graph for each variable, which I conducted using SAS. I created histograms, one-way frequencies, and/or summary statistics for these variables.

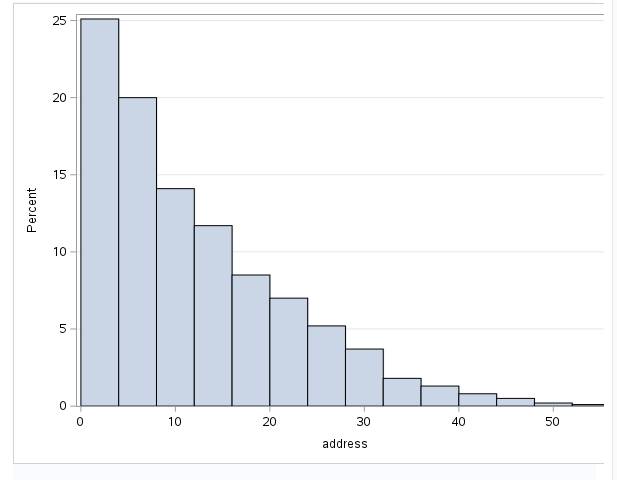
The provided code below is for *age*, which I systematically tailored to each of the remaining quantitative variables, generating histograms for each.



1. The *Age* data, displayed below, demonstrates a rightward skew, suggesting that the majority of customers are older, with an average age of around 40 years.



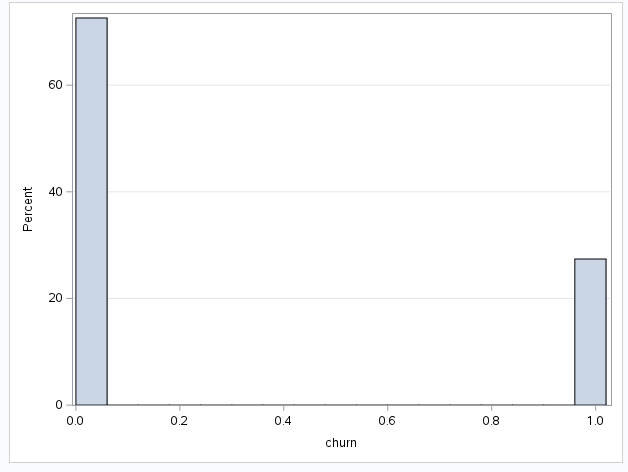
1. *Address*: The data indicates that most customers have lived at their current address for fewer than ten years, as evidenced by a rightward skew, displayed below.



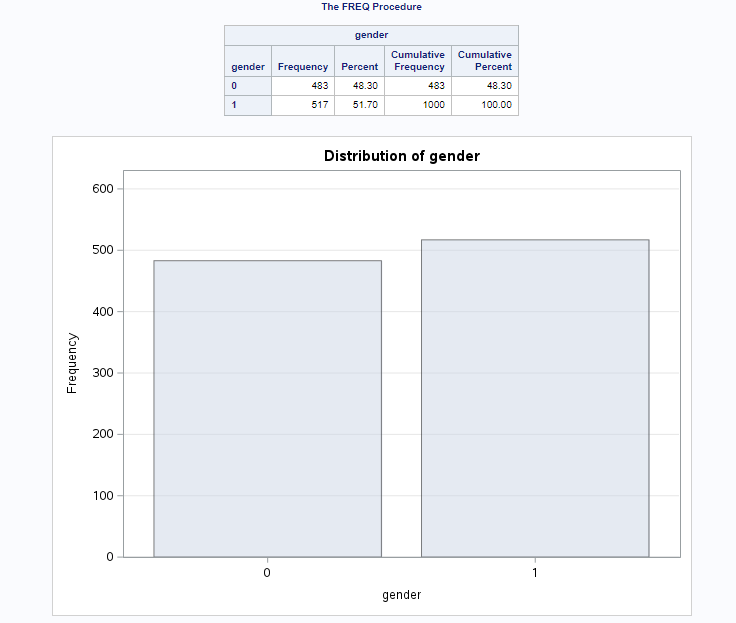
c) *Income* (with added mid-tick points for clarity): The dataset demonstrates a right-skewed tail. The histogram illustrates that the majority of clients earn between $9,000 (minimum from Summary Statistics) and $119,000 annually. Moreover, none report an income exceeding $1.26 million (maximum from Summary Statistics). To ensure clarity, I added Summary Statistics, confirming a maximum yearly income of $1.668 million and a minimum income of $9,000 per year, as mentioned earlier.



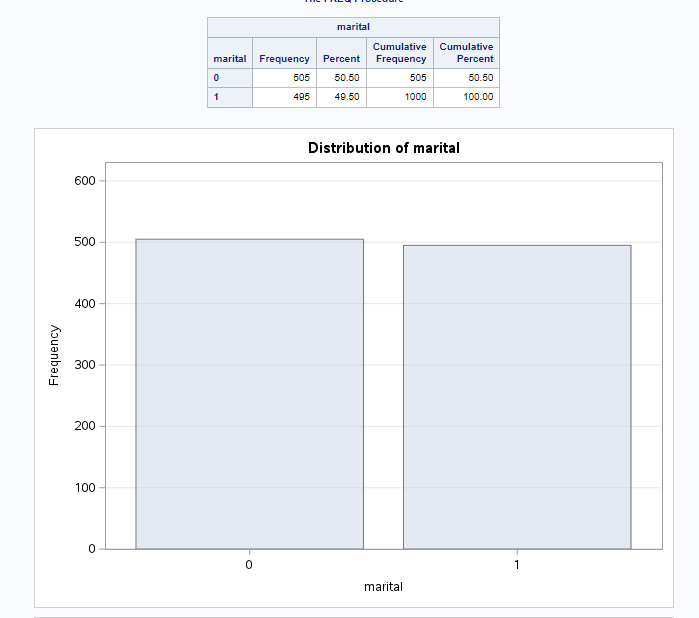
d) *Churn* is denoted by a binary value of 1 or 0, indicating whether a customer has opted to either discontinue or continue using the services of the telecommunications company. It's apparent that there are more clients with a value of 0 than those are, with a value of 1.



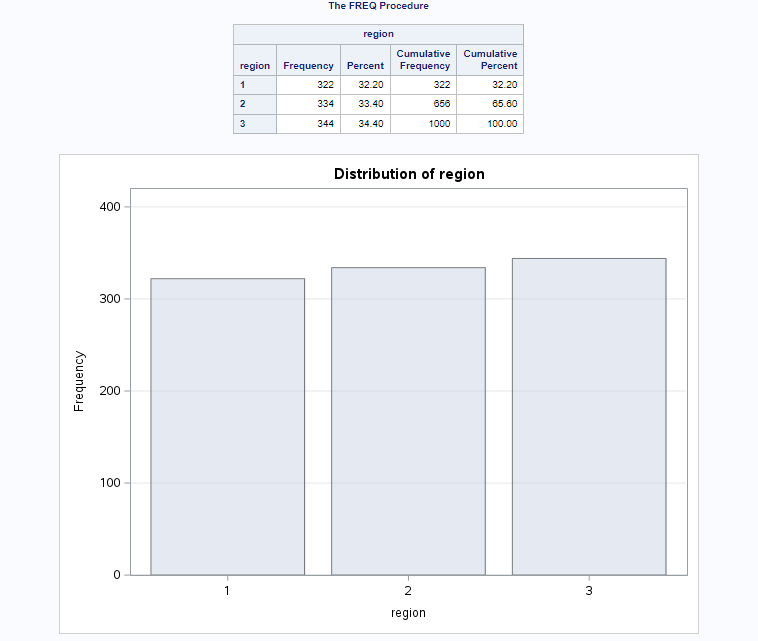
e) For *gender* I utilized a One-Way Frequency analysis. There are 483 instances for gender 0 and 517 instances for gender 1:



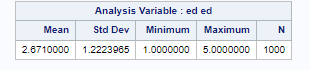
f) For *marital* status, I conducted a One-Way Frequency analysis. There are 505 individuals categorized as not married (factor 0). The second frequency is 495, representing those who are married (factor 1):

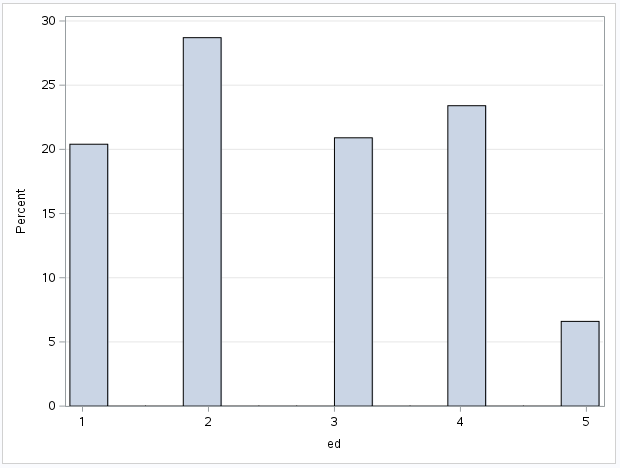


g) For the variable *region* of address, I also conducted a One-Way Frequency analysis. There are 322 instances for region 1, 334 for region 2, and 344 for region 3.

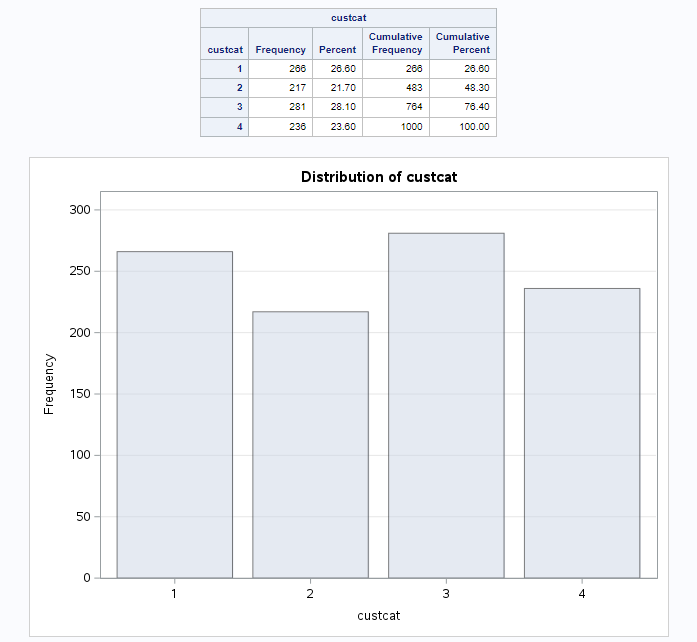


h) For the level of *ed*ucation, I utilized Summary Statistics and generated a histogram. The maximum level is 5, the minimum is 1, with the majority falling under level 2.



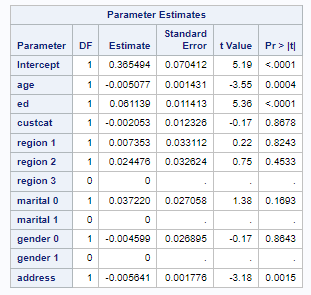


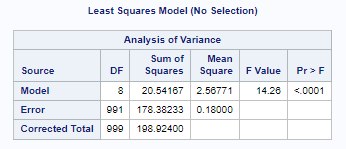
1. I conducted a One-Way Frequency analysis for *custcat*. All categories (1-4) have frequencies ranging between 225 and 275.



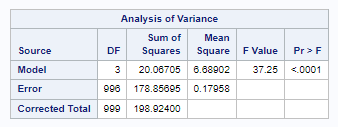
3) As instructed, I selected *churn* as the dependent variable and omitted *income* for this analysis. I then performed a multiple linear regression analysis with the remaining six variables from part 2.

a) Upon reviewing the Parameter Estimate P-values, it is evident that most are significant (p-values less than 0.05). However, *custcat, region*, and *gender* are not significant, indicating that these variables do not have a significant impact on *churn* and are not useful predictors.





b) I isolated the non-significant variables, and the resulting analysis revealed a comparably larger F-value of 37.25. This corresponds to a smaller p-value compared to the previous value of 14.26. Below is the table illustrating the analysis of variance.



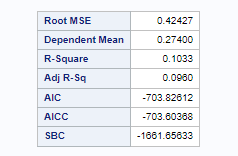
4) Listed, is the Hypotheses for part 3?

• The Null Hypothesis states that all of the coefficients are equal to zero.

• The Alternative Hypothesis suggests that at least one of the coefficients will be significantly different from zero.

• Since the P-value is very small, we conclude that at least one of the coefficients is significantly different from zero. The Null Hypothesis is rejected.

1. Below, is my Interpretation of the SAS output in terms of strength and direction of the relationship.
2. Here is the original (all variables):



1. Here are the numbers that I am focused on:



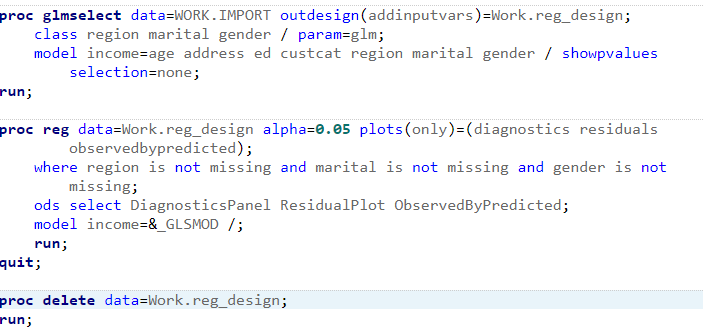
The variation in the x-variables explains only 10.33% variation of the *churn*

1. Here is the revised model (after extracting insignificant variables) along with the strength and direction of the relationships:

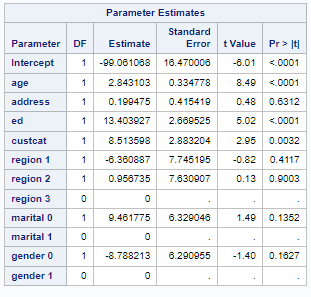


The variance in the x-variables accounts for only 10.27% of the variation in *churn*. The adjusted value improved slightly to 9.91%.

1. Following the instructions, I designated *income* as the dependent variable and utilized the other six variables as predictor variables (as mentioned in part 2), omitting *churn*. I conducted a Multiple Linear Regression Analysis in SAS. The code, table specifics, and output are outlined below:







From the Parameter Estimates (Multiple Linear Regression Analysis), the following are significant: *age, ed*, and *cuscut*. The following are not significant (and could be excluded): *address, region, marital* and *gender*.

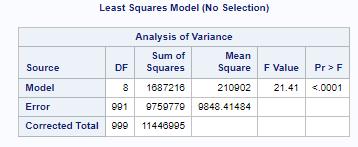
1. The hypotheses in part 6 address the question of whether the model is significant as a whole.

a) The answer suggests that while the model is better than having no model at all, it is significant but not as accurate as desired.

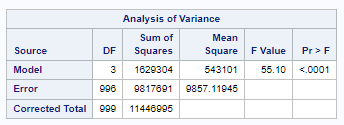
• Null Hypothesis: All estimates are equal to zero.

• Alternative Hypothesis: At least one estimate is significantly different from zero and serves as a predictor.

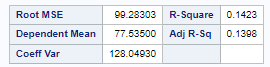
b) The result of the F-test is less than .0001. The Alternative Hypothesis is chosen, as the Null Hypothesis is rejected. The result of the f-test (the p-value) is seen below:



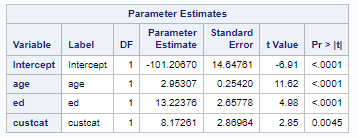
1. Lastly, I interpreted the SAS output in terms of strength and direction of the relationship.
2. Since the variables *address, region, marital* and *gender* are not significant, I extracted them. The model is more significant as a whole now. Here is the output:



1. Concerning the strength and direction of the relationship, the R-square stands at 14.23%, indicating the percentage change in income explained by the change in the x-variables. The strength is relatively low, scoring at 14.23% on a scale from 1 to 100%.



c) The direction of the relationships is apparent in the Parameter Estimates, all of which are positive. This suggests that as the variables increase, income also increases.



Summary:

a) The statistical results presented in the tables offer support for my hypotheses concerning both strength and direction, as indicated by the low P-values observed for the significant variables.

b) There were some unexpected findings during my analysis:

Notably, variables such as *address, region, marital* status, and *gender* were found to be non-significant and consequently excluded. This outcome was unexpected, particularly given societal norms where gender and marital status often have an effect on one’s income.

For example, married white men typically earn the highest incomes in America, as stated by the Federal Reserve Bank of St. Louis (2020), a trend not reflected in the dataset. However, I acknowledge that this dataset, Telco Extra CSUGlobal.xls, specifically pertains to the telecommunications sector. Additionally, it's worth mentioning that age often exhibits a polynomial relationship with income, with income typically increasing with age and experience until old age, after which it may decline.

References

Elliot, A.C. & Woodward, W.A. (2023). SAS essentials: Mastering SAS for analytics (3rd ed.). John Wiley and Sons. ISBN-13: 978-1-119-90161-7

Federal Reserve Bank of St. Louis. (2020, September 21). *Taking a Closer Look at Marital Status and the Earnings Gap*. Federal Reserve Bank of St. Louis. [https://www.stlouisfed.org/on-the-economy/2020/september/taking-closer- look-marital-status-earnings-gap](https://www.stlouisfed.org/on-the-economy/2020/september/taking-closer-%09look-marital-status-earnings-gap)

Kantardzic, M. (2020). Data mining (3rd ed.). John Wiley and Sons, Inc. ISBN 13: 978-1- 119-51604-0

SAS® Help Center (2012). <https://documentation.sas.com/doc/en/helpcenterwlcm/> 1.0/home.htm#/?softwareId=STUDIOMID&softwareVersion=3.7&requestor=ina pp&locale=en\_US

zyBooks. (2022). Statistics for data analytics.