C745 Capstone

Jonathan DiQuattro

TCM2 Data Analytics Report

Daniel Smith, PhD

October 16, 2021

Abstract

This paper is being submitted for the C772 Data Analytics Capstone course and demonstrates the skills learned over the course of the Data Analytics Program to include the ability to gather, analyze and report on large data sets for the purpose of making business-oriented decisions.

C745 Proficiency Assessment

**Report on the results of the research identified in TCM1 (appendix 1).**

**A. Research Question**

According to Pew research, minorities may face substantial bias when applying for mortgage loans. For instance, Black and Hispanic applicants are disapproved at higher rates than their Asian and White counterparts (DeSilver & Bialik, 2017). This national trend can be analyzed further and subjected to various statistical tests to validate the trends reported by pew research. This study was conducted on Arkansas mortgage application data and consists of an ample amount of records (over 125,000 entries). The data for this study was pulled from the Consumer Financial Protection Bureau and includes geographical information that was entered by the customer when applying for a mortgage loan. The question proposed by this research is “To what extent do the independent variables predict approved loans?” To understand the justification for researching this question, it is important to understand the federal protections in the mortgage industry, which have been put into place to prevent racial discrimination amongst applicants. According to the Department of Housing and Urban Development website, lenders are prohibited from discriminating against any applicant based on race, yet there is still a disparity in disapproval rates according to Pew research. For that reason, it is important to investigate the trends in the data and to apply statistical tests to determine if there is statistical significant evidence to support the unsettling data being reported by Pew research, and determine how and if race and other geographical data affects approval rates.

There is a very real and sad history of racial discrimination that has taken place in this country. To understand the context of this research, one must understand first the well documented racial discriminations tactics that have been used to deny minorities access to homeownership, some of which are documented by (Steil et al., The Social Structure of Mortgage Discrimination 2017) in their article for the HHS Public Access Journal. The historic and documented racial discrimination of minorities seeking housing is the basis for the Fair Housing and Equal Credit Opportunity Acts, but does not answer the question of how and if race impacts the probability of being approved or disapproved for a mortgage loan on a broader population. Explicit discrimination has been documented, but the aim of this research is to understand how racial bias impacts loan approval and disapproval probabilities. Due to the Fair Housing and Equal Credit Opportunity Acts, the applicants are prompted to enter geographical data, which in a fair and equal world should not influence the decision to approve or disapprove their loan requests.

The hypothesis of this research is that “There is no statistically significant relationship between the independent variables and approved loans.” That is to say, the geographical data entered by the applicants has no effect on the response variable of approved loans. This hypothesis would be rejected after running statistical tests if there is evidence that would suggest one or more of the variables affects the dependent variable. If after running various statistical tests it is determined that there is no significant relationship between the predictor variables and dependent variable, then the null hypothesis would be accepted.

**B. Data Collection**

The data used for this report was downloaded from the Consumer Financial Protection Bureau as a comma separated file and consisted of 70+ categorical and numerical variables. The data was very organized and included numerous .pdf files that explained the encoding of each categorical variable. From my experience combing the data.gov website, I was pleasantly surprised by how organized and easy to work with the data was, however it was not without faults or challenges and contained several columns of missing data which were promptly removed. While researching data for this project, I found it difficult to land on a dataset with the variables I wanted to investigate. Another issue I faced was that the numerical variables were all string types and were being imported as such into SAS. This made getting the descriptive statistics impossible because they were being seen as 125,000+ thousand levels. I simply wrote a SAS program to convert the string types to numeric types.

The SAS Studio online software was the methodology chosen to process and analyze the data, but presented advantages and disadvantages. I like SAS because I am certified on the technology and because it has all the tools needed for this project. I also felt that I was familiar enough with SAS to run all the procedures needed to complete this research. Some of the more difficult procedures, like PROC LOGISTIC, are made incredibly easy in SAS and would have required substantial research on other platforms. There are disadvantages to using SAS Studio online, which included not being able to upload files greater than 100MB. This really impacted the data chosen because using another technology, like Python or R to complete the research, would have added significant time to completing the project.

**C. Data Extraction and Preparation**

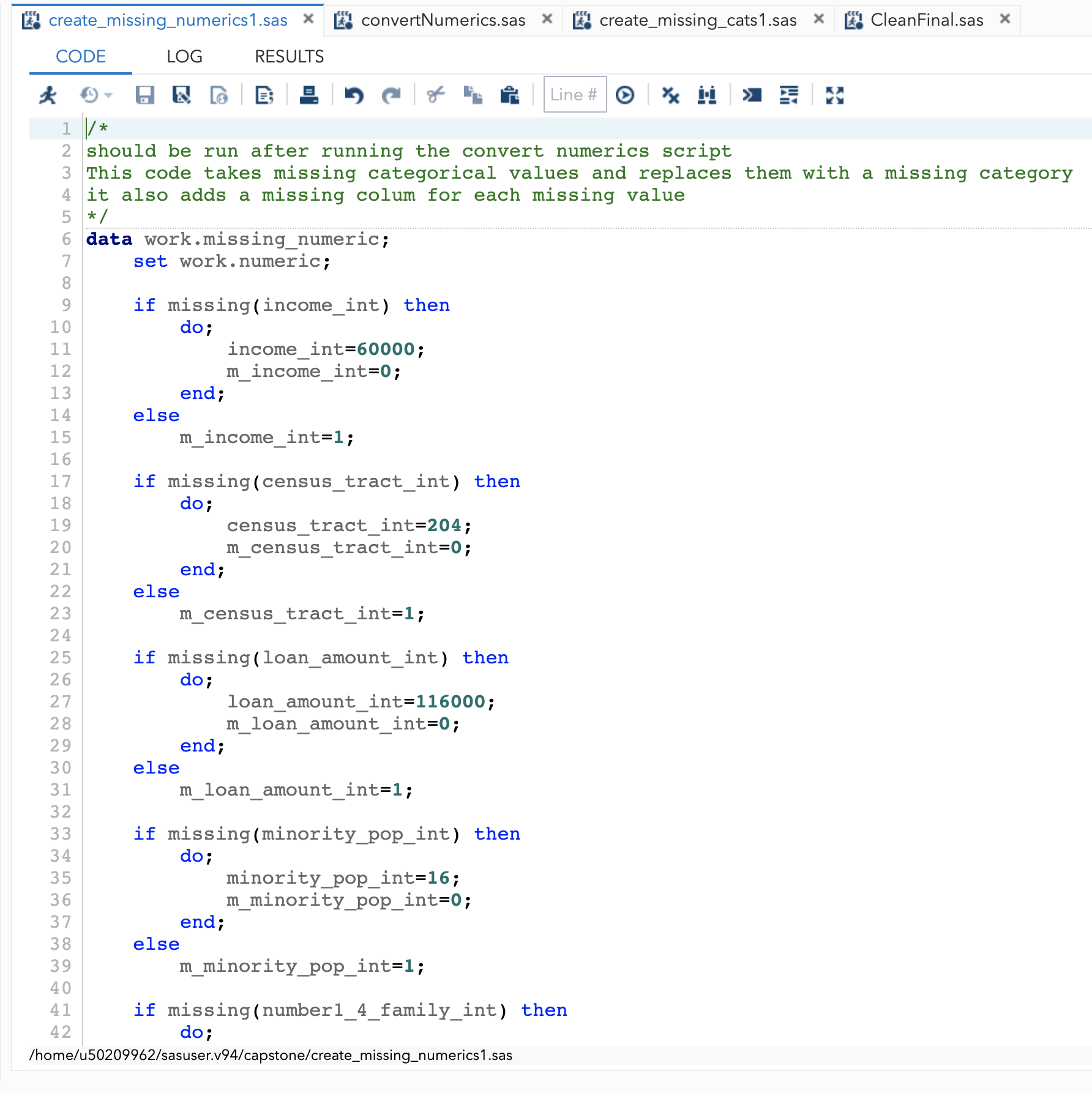
While the data was quite organized, the majority of this project was spent cleaning the data, which contained 75 variables, many of them being ordinal categorical variables. There were 7 numeric variables, which included the dependent variable “action\_taken\_name”. All the numeric variables were encoded as character types and had to be converted into numeric variables. Several variables were removed from the data for being empty for every entry. These were different from the random missing values that were accounted for in the other variables, and consisted of entire columns of empty data.

To account for the other randomly missing variables, a missing category was imputed for categorical variables, and median values imputed for each missing numerics.

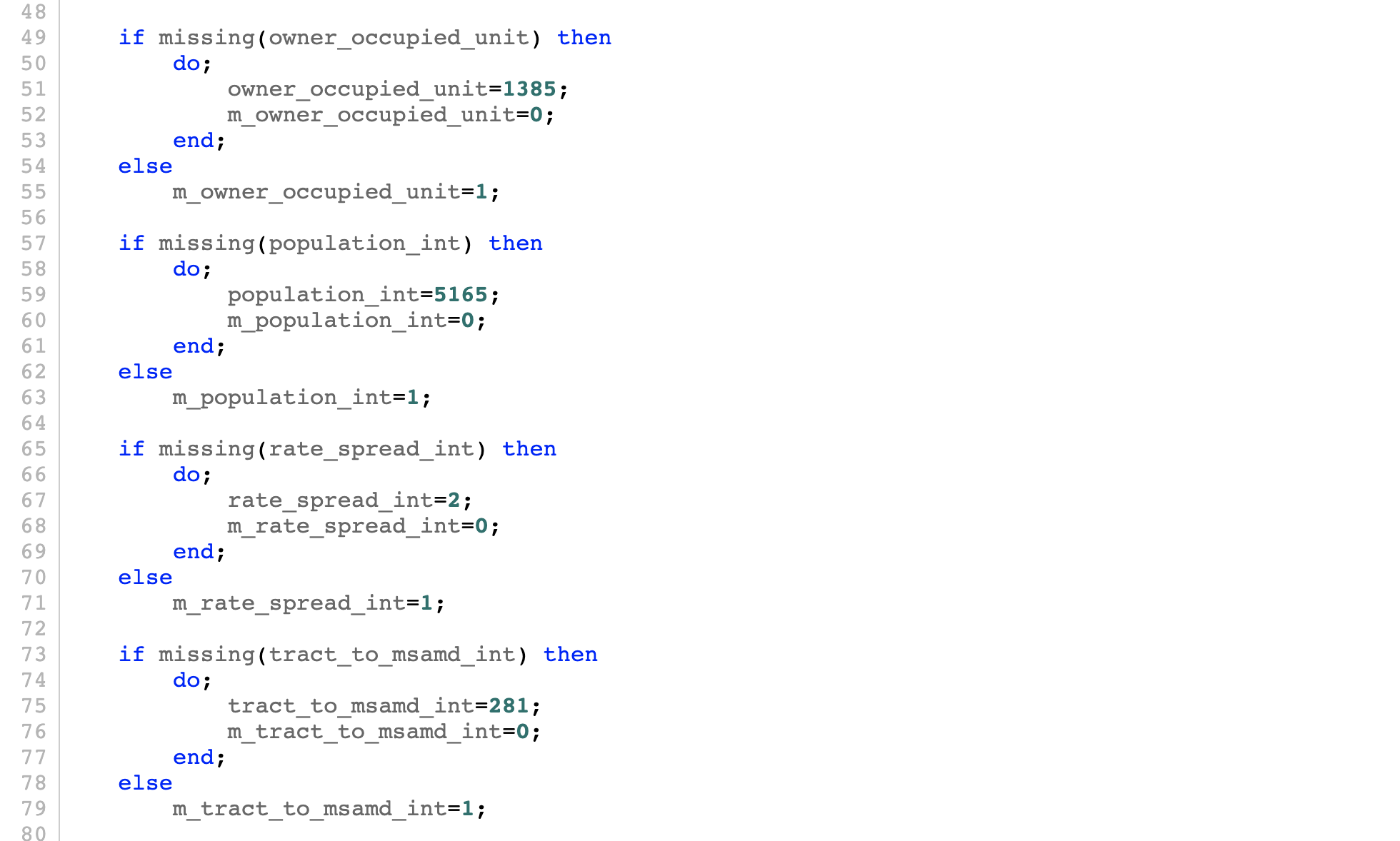
Along with the missing numeric medians, the missing-ness was captured in a missing column for each numeric variable that contained missing values (Image C.1-C.3).

Creating missing variables for each missing categorical variables allows the relationship for the missing entry to be accounted for in the response variable, and while the Fair Housing and Equal Credit Opportunity Acts require geographical data to be collected, it does not require the applicants to actually enter the data. For some data like income\_int the applicant may not want to answer the question for fears of discrimination. For that reason it’s important to account for missing-ness, which could be involved in an interaction with the dependent variable or a secondary relationship that could affect the dependent variable.

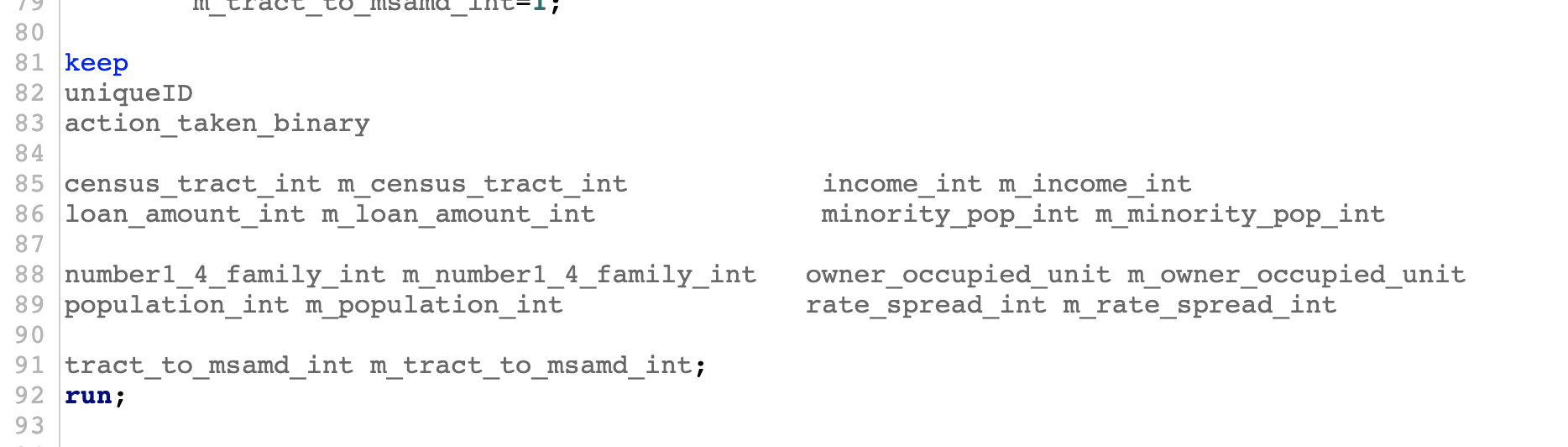
**C.1**



**C.2**

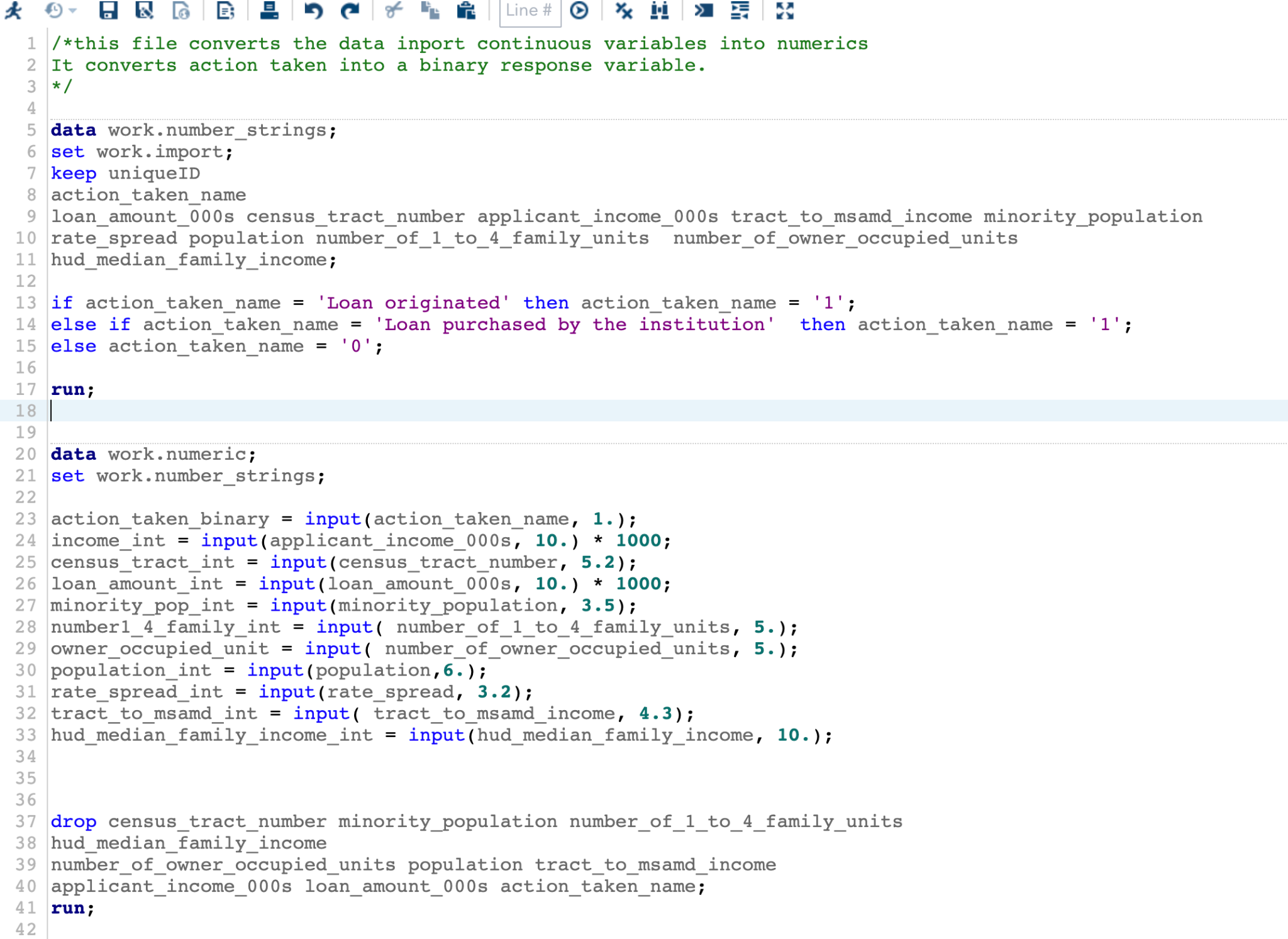
****

**C.3**

****

Before the missing numerics could be imputed with median values, they had to be converted from character types into numeric types for computing the medians (C.4). For this task, a procedure was written to convert the actions\_taken\_name categorical variable into a binary variable. This was done in conjunction with the encoding schema that was downloaded with the data set. If the loan was “originated” or “purchased” by an institution,” it represents an approved loan, and any of the other levels represent a denied loan.

**C.4**

****

The other numerics were converted to their corresponding values. Income and loan\_amount were converted into decimal values from strings. For instance, “10k” would be converted into the value 10000. This was done to preserve the relationship between the numeric value and the binary response variable to keep a more accurate representation of the data.

The missing categorical variables were resolved by creating a missing category and imputing that category for each missing value.

**C.5**

****

The advantage of using a missing category is that during the analysis the missing category can be used like any other category. A missing value could be statistically significant to the dependent variable, depending on why it is missing. An alternative would be to remove the records with missing values, but that would drastically reduce the usable records and would not account for the potential relationship between missing-ness and the dependent variable.

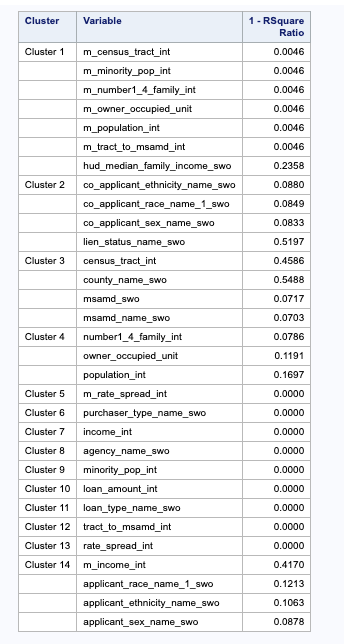
**D. Analysis**

There were multiple phases used to analyze the data, with the ultimate goal of using a binary logistic regression algorithm to quantify any potential relationships between variables in the dataset and the dependent variable actions\_taken\_name.

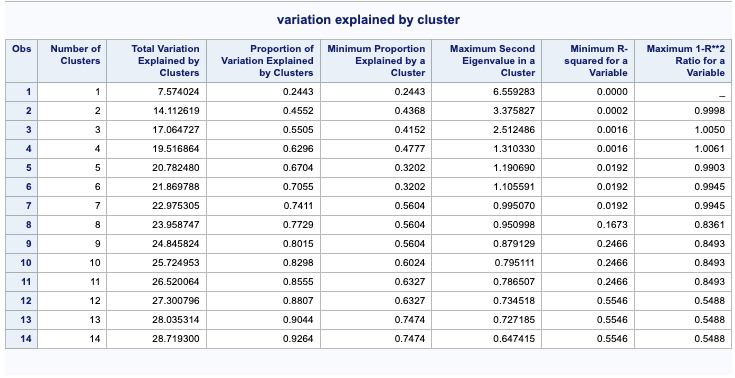
The benefit of using binary regression is that the parameter estimates can be used to quantify the relationship between any numeric variable in the model and the dependent variable. As the parameter increases or decreases, there is a relative response in the dependent variable, where the probability of the event increases or decreases for every unit of change. The same parameter estimates will also be calculated for each level of the categorical variables compared to a control category.

The first phase consisted of exploring the variables. For the categorical variables, frequency tables were made. For the numeric variables, descriptive statistics were generated using the PROC MEANS procedure. Both frequency tables and descriptive statistics were used to verify the data at each phase of the process. The next step, after accounting for missing variables detected in the first phase, was to convert categorical variables into numerics using a Smooth Weight Of Evidence conversion. Since the majority of the variables were categorical, this was a lengthy process. This conversion made clustering the variables in a PROC VARCLUS possible, since PROC VARCLUS only works with numeric variables. The advantage of clustering in this manner is that it greatly reduces the amount of variables and dimensionality of the logistic model. This reduction of dimensions makes decisions regarding which variables to use in the model abundantly clear. This strategy resulted in 14 clusters and highlights redundancy in the variables. For instance, in image D.1 hud\_median\_family\_income\_swo is redundant to multiple missing variables, and has a larger impact on the dependent variable as indicated by the R Square Ratio. Using these clusters, combined with data obtained about the data set from the source, allowed for the systematic inclusion into the final model. After selecting the variables from the clusters, further analysis of the relationships can be conducted to include chi square and Spearman correlations. These tests will not only detect and validate relationships, but give a quantification of the relationship on a scale of 0 to 1, where anything over a .2 Spearman correlation indicates a weak to moderate relationship.

**D.1**

****

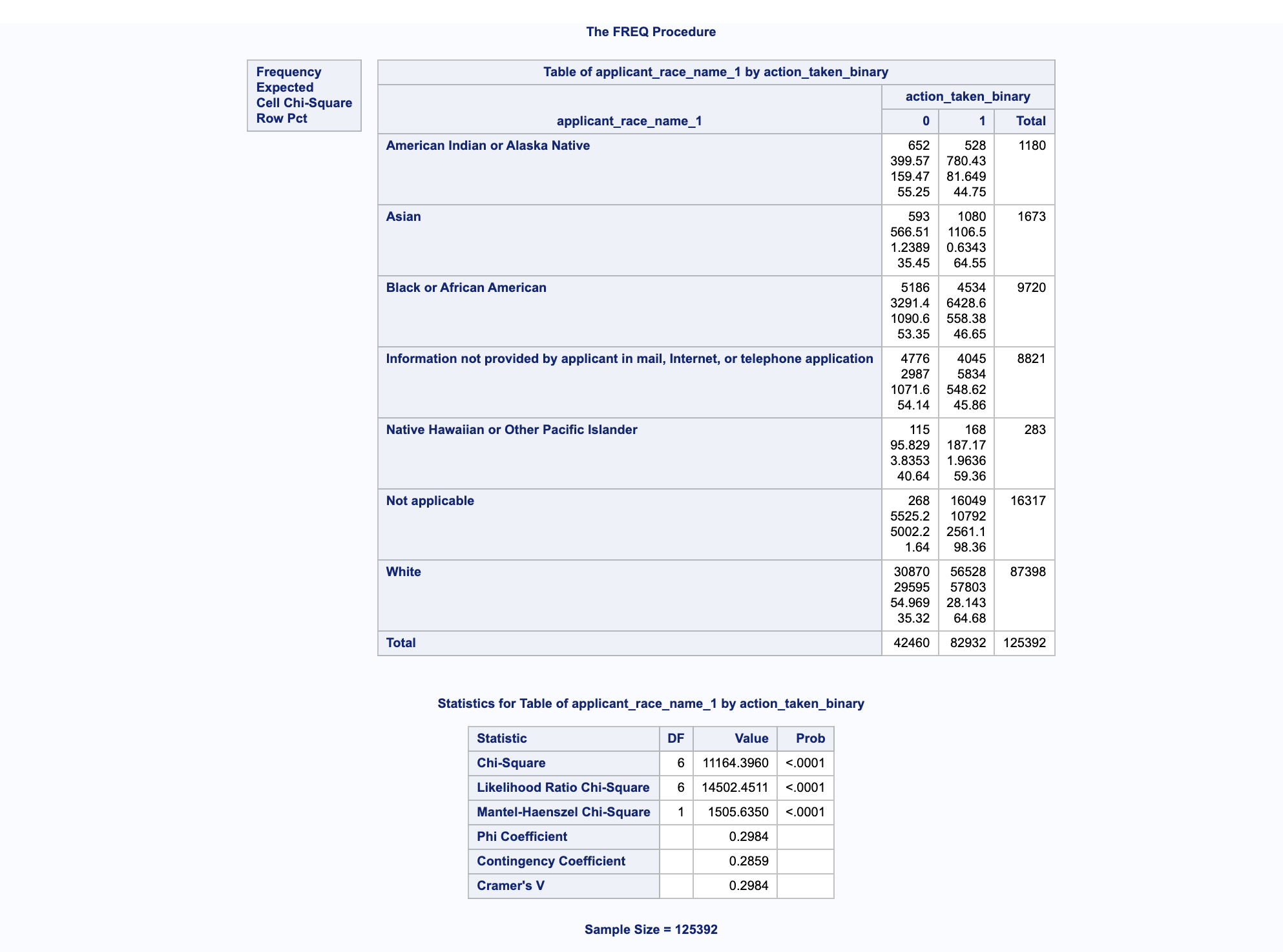
**D.2**

****

The advantage of using Smooth Weight of Evidence is that there is no need to keep the original variables included in the dataset. Other clustering methods, like Multiple Component Analysis, results in eigien values being substituted for the categorical values. This approach requires the original variables to be maintained in the data set. The end goal of this research is not necessarily to create a model for predicting new values, but to explain an existing relationship. This process creates a model that could be trained on a subset of the overall data, and tweaked using post hoc procedures to create a model that fits well to new data.

A PROC FREQ procedure was run to analyze the Pearson chi square of applicant\_race\_name\_1 and other categorical variables selected from the clusters. The results show, at the .05 level, that there is statistical evidence to suggest that applicant\_race\_name\_1 is related to the actions\_taken\_name dependent variable.

**D. 2**



What is even more compelling is that the category “White” has almost twice the expected amount of approved loans, while the category “Black or African American” is about 13% lower than expected. Upon further analysis of the “Cramer’sV” statistic, it is clear that there is a weak to moderate relationship between applicant\_race\_name\_1 and the action\_taken\_binary dependent variable.

Once the model variables were selected from the clusters, they were added to the logistic regression model, which gives a parameter estimate based on each level. Since this study is most concerned with how minorities are effected when applying for a loan, the decision was made to use the original applicant\_race\_name\_1 categorical variable. This will allow us to see how each race compares to the control level of “White.” The logistic regression algorithm outputs the “Type 3 Analysis of Effects” table, which displays the Chi Squared and Wald statistics to determine the statistical significance of the relationship between each variable and the dependent variable. The null hypothesis being that there is no significant evidence of a relationship, and the alternate hypothesis being there is a statistically significant relationship (D.2).

**D.4**

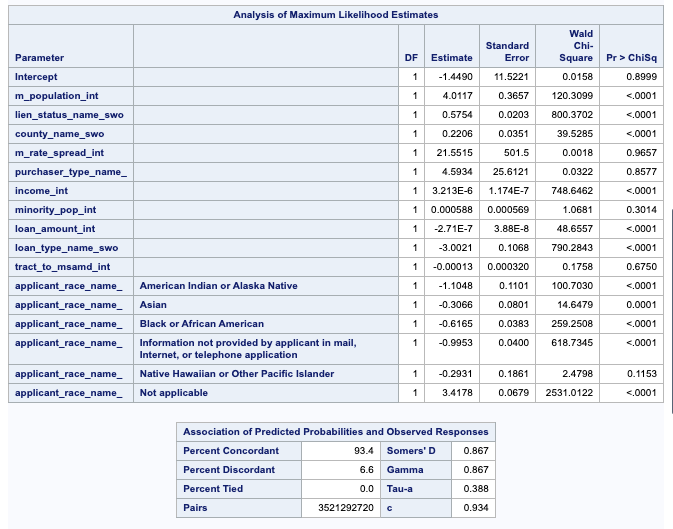
****

The results of Chi Sqare and Wald show that each of the variables m\_population\_int, lien\_status\_name\_swo, county\_name\_swo, income\_int, loan\_amount\_int, loan\_type\_name\_swo, and applicant\_race\_name1 imputed into the model are statistically significant at the .05 level, and thus we reject the null hypothesis for those variables.

The “Analysis of Maximum Likelihood Estimates” shows an alarming result. At the .05 level of significance for each level of the applicant\_race\_name\_1 (minus Native Hawaiian or Pacific Islander), there is a decrease in the likelihood of being approved for a loan compared to the control. Most unsettling is that the “Black or African American” category has a negative coefficient of -.6165. Which can be converted to a percentage by exponentiation and subtracting 1 to get to get ~46%. This interprets to the “Black or African American” category being ~46% less likely to be approved for a loan vs the control.

Binary logistic regression is a good tool for this analysis because it outputs chi square and Wald scores for each variable. Since SAS has many other statistical test options built into the PROC LOGISTIC procedure, it is a very powerful tool. To get the same output from an open-source technology like R or Python would require researching multiple packages, and could be overly time-consuming.

**D.5**

****

**E. Data Summary and Implications**

This research was conducted to answer the question of “To what extent do the independent variables predict approved loans?” for the purpose of testing the claims by Pew research that minorities face bias when applying for a home mortgage. This was accomplished by applying a logistic regression algorithm to data collected from the Consumer Financial Protection Bureau, which contains over 125,000 loan applicants from the state of Arkansas during 2017.

The logistic regression outputs a Pearson Chi square table, which indicates at the .5 level of significance that there is a relationship between the variables included in the model and the dependent variable actions\_taken\_name\_1.

The Coefficient or Estimate column listed in the Analysis of Maximum Likelihood Estimates table shows that at the .05 level of significance that there is a decrease in likelihood of being approved for members of the levels “African or African American”, “American Indian or Alaskan Native”, and “Asian”, moreover there is statistically significant evidence that applicants who responded “African or African American” are ~46% less likely to be approved for a home loan than the control group “White”.

In conclusion, the null hypothesis that “there is no statistically significant relationship between the independent variables and approved loans” is rejected.

The limited financial data included in the data set, which does not include a credit score or overall assets leaves the totality of the research hindered, and intuitively credit score and asset wealth would be larger drivers of the decision to approve or disapprove loans. Due to these limitations, further research could be conducted to determine how those data points affect the dependent variable. Another limitation to this study is that there are not just people making risk decisions about who gets approved for loans (Martinez, 2021), but there are algorithms that financial institutions use to determine risk for each applicant, and including a risk score would pose an interesting question of how bias affects not just the human element but the computer element as well.

In conclusion, I would recommend financial institutions investigate internal procedures and data collection processes to determine why race is playing such a large factor in deciding who is approved for a mortgage loan. I would also suggest moving more of the application process to an online format, where the user is shielded from any naturally occurring biases during the loan application process. An interesting study could be conducted on loan officers to determine how race affects approval rates when the loan officer knows the race of the applicant versus not knowing.

**F. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.**

The citations for this paper can be found on the reference page below.

References

U.S. Consumer Financial Protection Bureau. (2017). Retrieved September 19, 2021, from <https://www.consumerfinance.gov/data-research/hmda/historic-data/?geo=ar&records=all-records&field_descriptions=labels>

History of fair housing - HUD: Hud.gov / U.S. Department of Housing and Urban Development (HUD). HUD. (n.d.). Retrieved October 21, 2021, from https://www.hud.gov/program\_offices/fair\_housing\_equal\_opp/aboutfheo/history.

DeSilver, D., &; Bialik, K. (2017, January 10). Blacks, hispanics face mortgage challenges. Pew Research Center. Retrieved September 19, 2021, from [https://www.pewresearch.org/fact-tank/2017/01/10/blacks-and-hispanics-face-extra-challenges-in-getting-home-loans/](https://www.pewresearch.org/fact-tank/2017/01/10/blacks-and-hispanics-face-extra-challenges-in-getting-home-loans/.)

Steil, J. P., Albright, L., Rugh, J. S., &amp; Massey, D. S. (2017). The Social Structure of Mortgage Discrimination. HHS Public Access. https://doi.org/10.1080/02673037.2017.1390076

Martinez, E. (2021, August 25). The secret bias hidden in mortgage-approval algorithms. AP NEWS. Retrieved October 22, 2021, from https://apnews.com/article/lifestyle-technology-business-race-and-ethnicity-mortgages-2d3d40d5751f933a88c1e17063657586.

SAS. (n.d.). The VARCLUS Procedure. SAS help center. Retrieved September 19, 2021, from <https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.4/statug/statug_varclus_syntax01.htm>