**A. Research Question**

**Research Question:** Does the consumption rate of various food products influence health outcomes?

**Hypotheses**: **H0 :** Food consumption has no impact on health outcomes in developing nations, **H1:** in nations with lower consumption of high-fat and high-caloric foods, there are more favorable health outcomes, **H2:** nations with greater consumption of high-fat and high-caloric foods result in better health outcomes.

This research is designed to explore the intersection of food consumption and health outcomes in developing countries. This research analyzes foods divided into 32 categories: rice (in all forms except flour), bread (fresh bread and special bread), cheese (cheese and curd), and pasta products to name a few. Consumption rates from each category were gathered from The World Bank. The data for nutritional value was then gathered for each type of item in the category and the median of the individual food items was used to represent the category. Finally, the health outcomes for each developing country were gathered from the World Health Organization. The Health Outcomes and consumption rate datasets are from 2010 data, but the nutritional data, gathered from nutritionvalue.org, is 2020 data. This is a constraint on the accuracy of the results attained.

This research aims to achieve a predictive model for health outcomes in developing nations based on their current consumption rates. This will provide health and government officials with the information necessary to focus resources and maximize care.

**B. Data Collection**

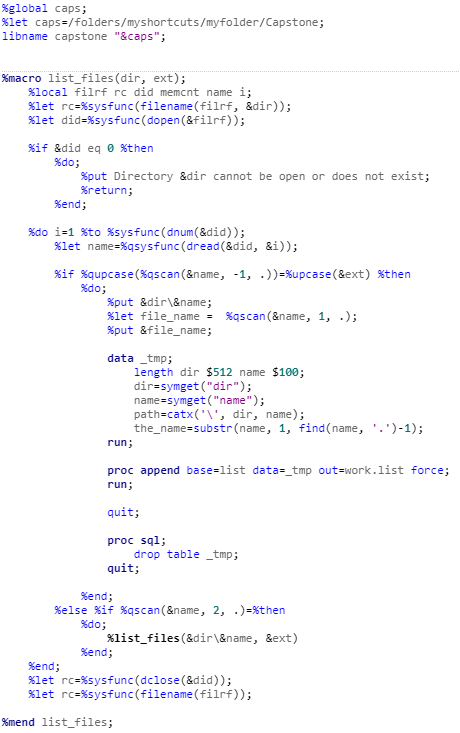
The World Bank and World Health Organizations had easily accessible datasets that downloaded into comma-separated values. Initially, this project was designed to collect data from Food Data Central for nutritional values. However, Food Data Central did not have records for a large portion of the food items and would have resulted in skewed and inaccurate results. The project then shifted to utilizing nutritionvalue.org. This provided enough data for each subcategory listed within the category. Not every single item and its contribution to the total consumption were able to be included due to data unavailability, untraceable local products, and underground markets.

While reviewing the data, higher rates of health outcomes and higher consumption rates were skewing some results. For this reason, Data Hub’s Population Figures by Country data set was used in order to create per capita figures that aided the practicability of the results.

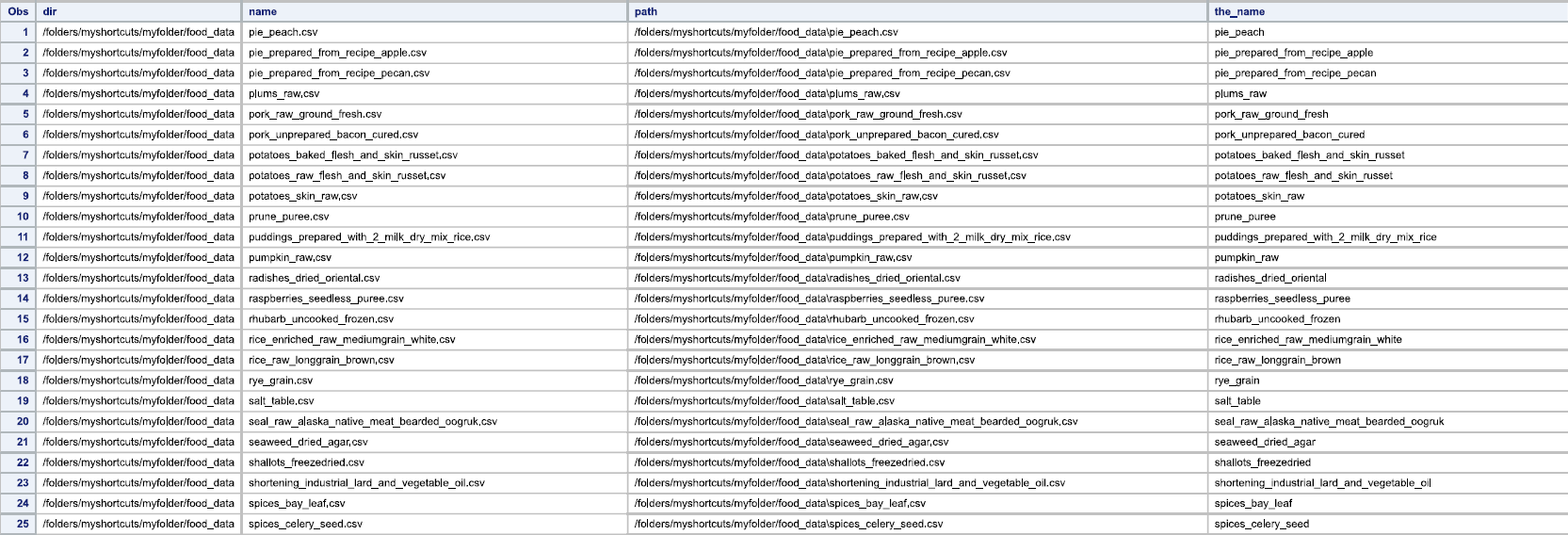
An advantage to the data gathering process is that the generalized approach gives a broad view of the relationship at hand without getting into product specifics. This is a useful approach for focusing nutrition information and health resources on a large scale.

**C. Data Extraction and Preparation**

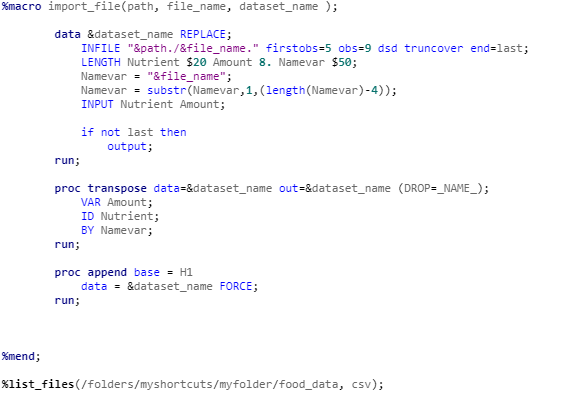
The first step necessary was to create a macro in SAS that would facilitate an iterative process through each csv file in the Food Data folder. There is a separate csv for each food item with the nutritional data in a column and the nutritional values in a separate column. The macro would first read through the folder of csv files and create a list of the filenames in a temporary file called “list.” This is preceded by creating a library to reference throughout for the folder “Capstone.” The macro method was selected in order to reduce redundancy in the code and increase efficiency, while an aftereffect is that the coding became increasingly obscure.



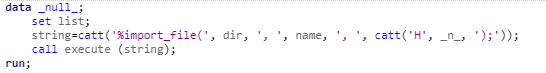
The following steps will then call on the “list” file variables to iterate through each file within the folder. This file is shown below.



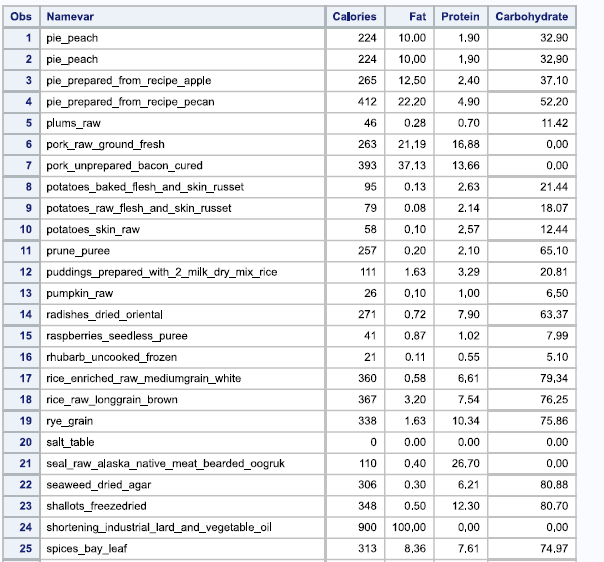
As the macro iterates through each folder, it is creating a variable with the name of each food and is reading the data for the four observations which include information on Calories, Protein, Carbohydrates, and Fat. The next procedure transposes each file from narrow to wide, making the observations Calories, Protein, Carbohydrates, and Fat, into variables and transforming each file to have a single observation with multiple variables. Then, the PROC APPEND procedure is used to create a single dataset with all observations compiled.

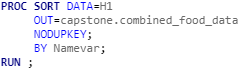


The data step here serves to trigger the iteration of the macro through each file in the “list” file created earlier.

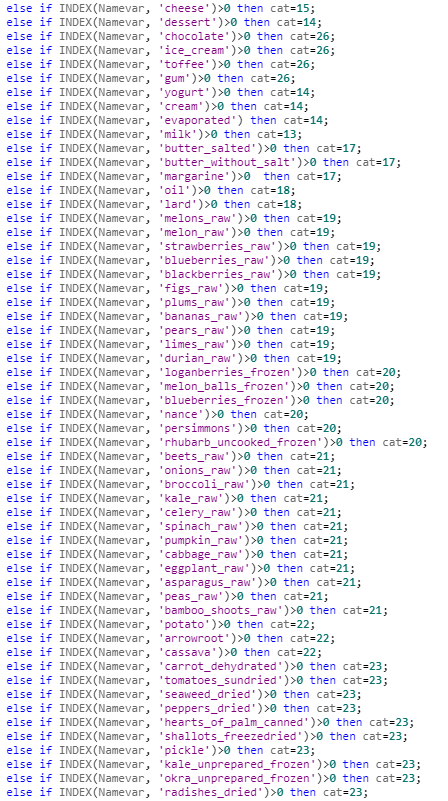
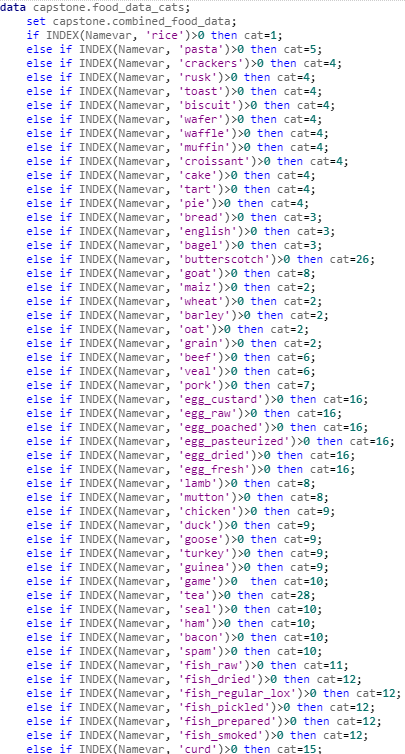


Once run, this final piece of the macro creates the following dataset, which has a few duplicate files for food items. This is a flaw in the selection of this method, but it was then resolved by the NODUPKEY function in the PROC SORT procedure.





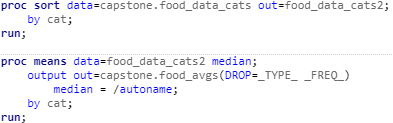
The food dataset is cleaned. Now, A category variable must be added to each food item in order to determine the median calorie, fat, protein, and carbohydrate levels for each category to facilitate category analysis. To do so, the function INDEX() is used to search through the name of each food item for a key word and categorize the food item based on that. The categories are strategically ordered in the IF statements so that items are not misclassified. The key words for each category were determined through the Global Consumption Database, Description of Sectors through the World Bank website. ­The main disadvantage with this technique is that ordering the food items for category classification was time intensive and resulted in a long procedure. However, this method avoided the creation of web scraping the food types from The World Bank website.

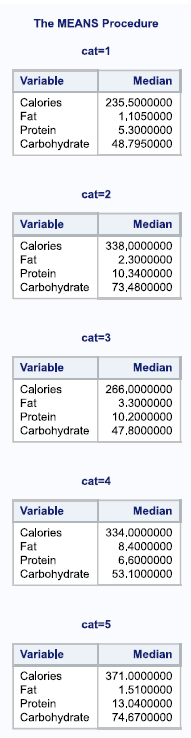


Above: Item 1

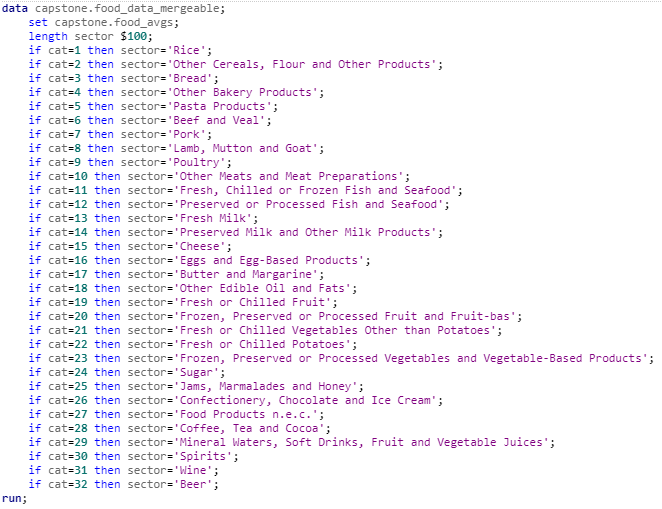
Above: Item 2

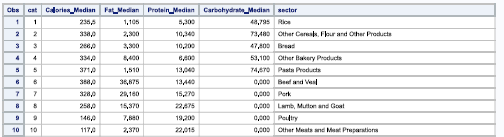
Each item has been categorized and is then sorted and PROC MEANS is used to find the median of each category and place these results into an output file named “food\_avgs.” The median was selected as a method to avoid outliers in the nutrition information, due to the fact that the food categories were diverse and the nutrition information varied greatly within a category. While this method avoided issues caused by the skewed distribution, the mean would consider the entirety of the data and perhaps be equally representative of the data.



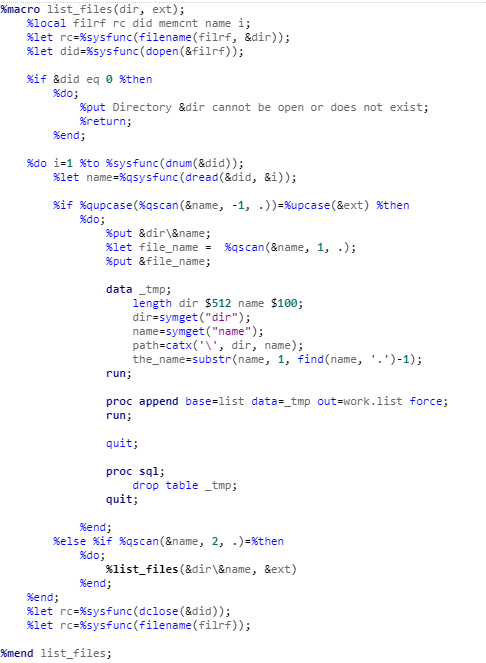


Now that the medians for each category have been identified, the file is given an additional variable named Sector and then given the proper name of the category instead of a number. This step was to facilitate a merge of this data onto the food consumption data in a later step.

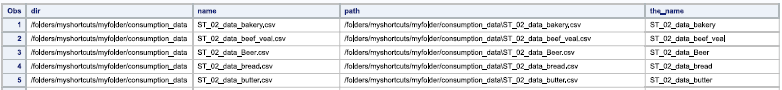




Moving on to the preparation of the Food Consumption Rate Dataset. The Food Consumption rate csv files were available on The World Bank website with each file separated by category. For this reason, the same methodology was used to process the data. First a macro was created, each file name was read into the file “list,” and then an iterative process transferred the csv files into SAS files, keeping only the US currency variable (out of the various choices for currency) and the National variable (out of the Rural, Suburban, and Urban additional designators). The data was then sorted by country and transposed, creating a wide observation for each country. The macro finishes out by merging all of the datasets into one.

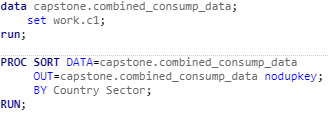


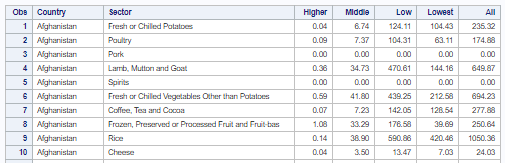




Above: List Dataset created to iterate through the Food Consumption categories

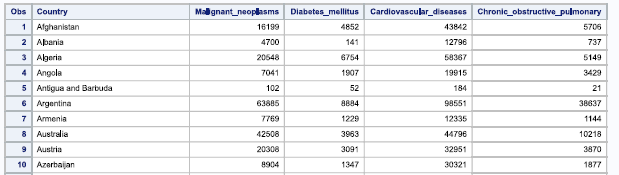
The next code turns the temporary file into a permanent one saved in the Capstone folder for use later on in the analysis, then sorts the file to remove duplicates.





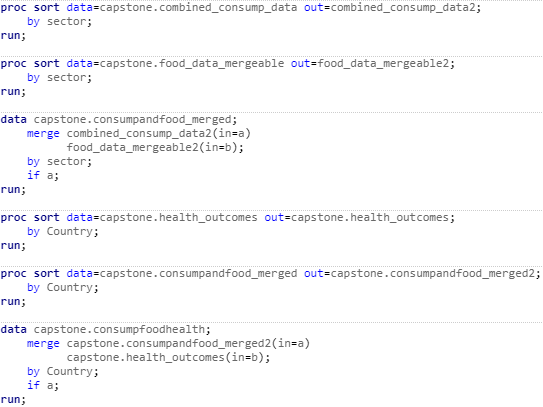
The final dataset, health outcomes, also needs to be transferred from a csv file to a SAS file. The input variables are reduced to four variables relevant to the research. Then, the file is reduced to only the observations for 2010 health outcomes. This is to align the data with the World Bank data, 2010 Food Consumption Rates. The dataset is sorted by country and then transposed to create a wide dataset with variables for each health outcome.

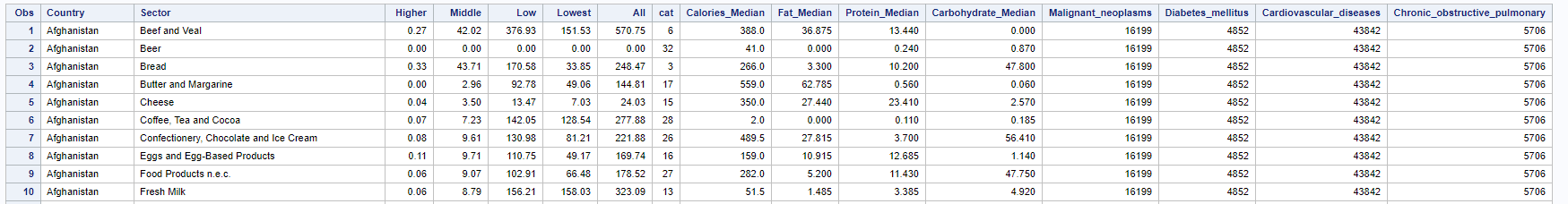




Above, several of the naming conventions for the Health Outcomes data set had to be adjusted to match the Food Consumption data set in order to prepare the set for a successful merge. This was executed with several if statements to rename countries. While this was a tedious and lengthy approach, it avoided the removal of the entire columns removal of additional characters and spaces, which would’ve resulted in lengthier procedures and muddled observations in that column.

The datasets were then ready to begin merging. The DATA step was selected for this procedure due to the fact that it was a simple merge. Contrastingly, SQL could have been used which would have reduced some code, but would have reduced processing time. The Food Consumption dataset and Food dataset are then sorted by Sector and merged. After merging, the new dataset is sorted by country alongside the Health Outcomes dataset in order to prepare for a second merge.





In the merge statements, the IN= paired with the IF clause was used to manufacture a “left join,” where only the observations that matched with the Food Consumption dataset were kept. This practice reduced missing values in the end product.

Then the population dataset needed to be read into SAS, sorted and merged with this dataset. This dataset was already in a csv file and was imported into SAS with a data step similar to the one used with the health outcomes data set.

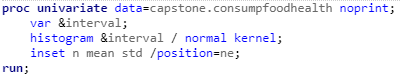


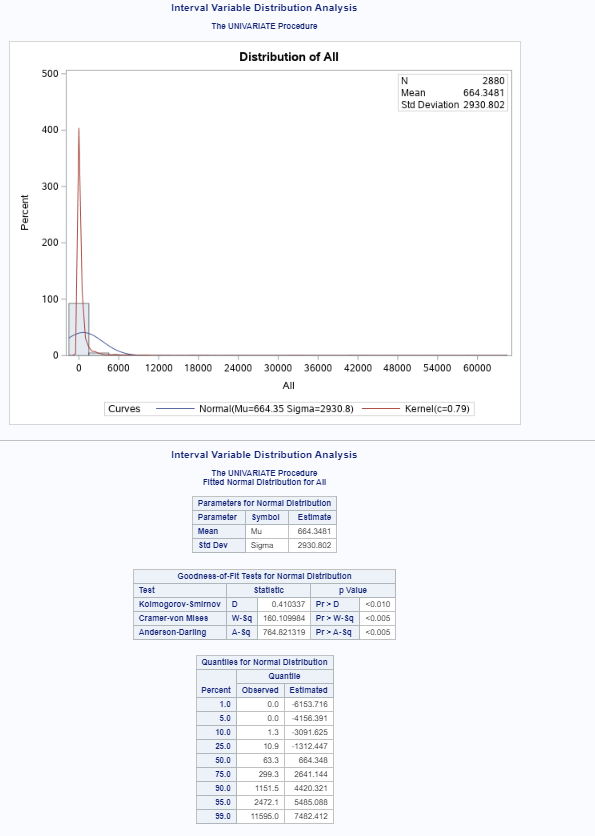
In the above procedures, the data set was paired down to essential variables and several of the Country observations were changed to prepare for the subsequent merge. After the merge, several variables were created. The “percapita\_all” variable showed the per capita consumption rates for each nation, adjusted for the “all” variable, which was in millions. Each health outcome was then adjusted to a per capita rate. This would facilitate more accurate results that were not inflated by countries with larger populations.

The final data set is clean, prepared, merged, and ready for exploratory data analysis.

**D. Exploratory Data Analysis**

**I. Distribution Analysis**

Upon reviewing the data, there are no categorical variables, only continuous variables. To analyze the distribution of these variables, the research uses the UNIVARIATE procedure with the histogram function. This procedure was selected in order to provide both graphs and statistics on the distribution of the numeric data. While the MEANS procedure might offer several of the same statistics, the UNIVARIATE procedure has more features for analyzing distribution and additionally produces graphs. Analyzing the distribution was key in determining whether this data would provide accurate models, or whether the distribution needed to be modified in order to improve the normalcy. From this analysis, it is shown that none of the variables have a strictly normal distribution. The only variables with a semi-normal distribution are protein median, fat median, carbohydrate median, and calories median. Though removing outliers would potentially resolve the issue, the data points were both verified and legitimate and the truth could be misrepresented without these outliers.



The above distribution analysis is representative of the analysis results for consumption rates amongst the lowest-income, low-income, middle-income, high-income, and all-income areas of developing countries. The standard deviation for each consumption rate indicated there is a large spread of the data which is shown in the graph. The Kolmogorov-Smirnov test for all five analyses is above .05, which indicated that the difference between the data and the normal curve was significant. The more powerful Anderson-Darling and Cramer-von Mises test show similar results to the KS test. This is also seen in the differences between observed and estimated quantiles for the distribution. Due to the tails of the data carrying much of the abnormalities, the Anderson-Darling Test is the most reliable for these variables. The Kolmogorov-Smirnov and Cramer von Mises tests are preferable when the center of the data contains the majority of the deviation (Yap and Sim, 2011).



Above is the distribution analysis for median calorie rates for each food group. The standard deviation indicated that the spread of the data was relatively small, which is also shown in the graph. The Kolmogorov-Smirnov test is above .05, which, though the distribution was much closer to normal than the consumption variables, indicated that the difference between the data and the normal curve was significant. The more powerful Anderson-Darling and Cramer-von Mises test show similar results to the KS test. In this case, the more centered skew in the data makes the KS and Cramer-von Mises statistics more reliable. This is also seen in the differences between observed and estimated quantiles for the distribution.

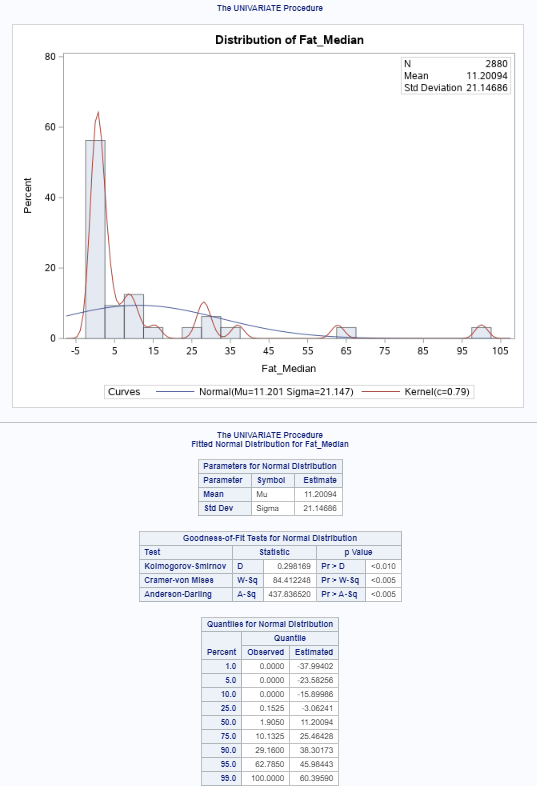
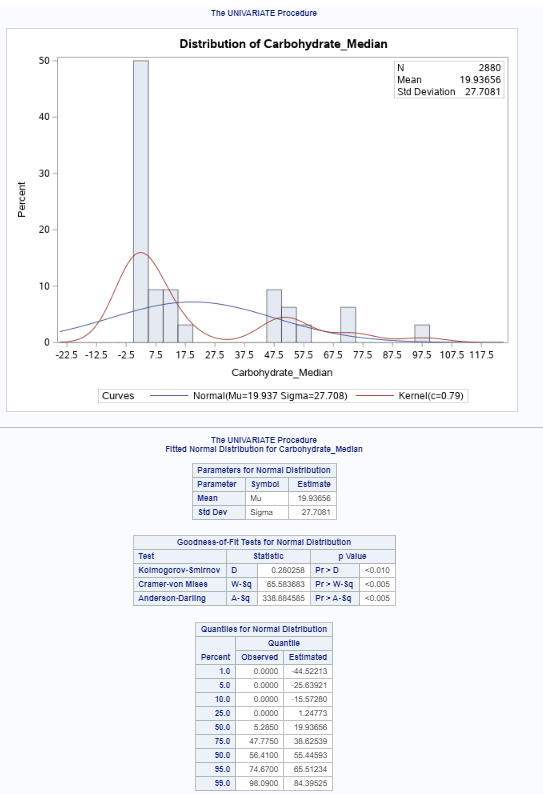
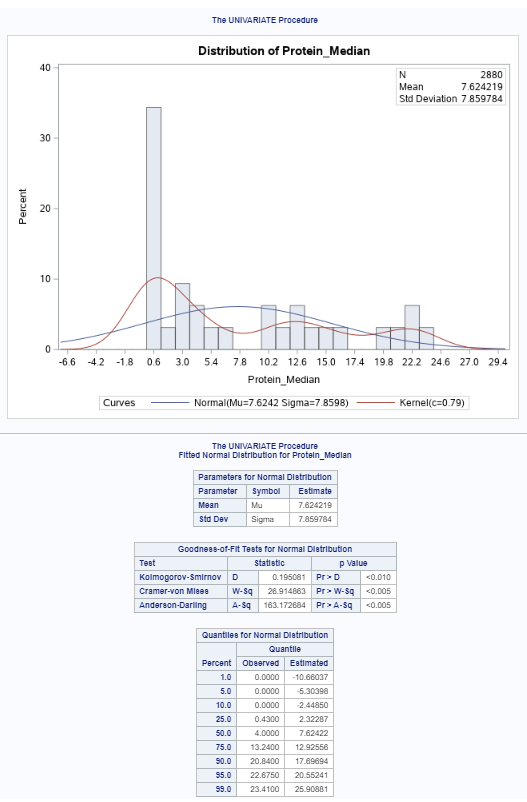


Image two: Protein\_Median Distribution

Image one: Fat\_Median Distribution

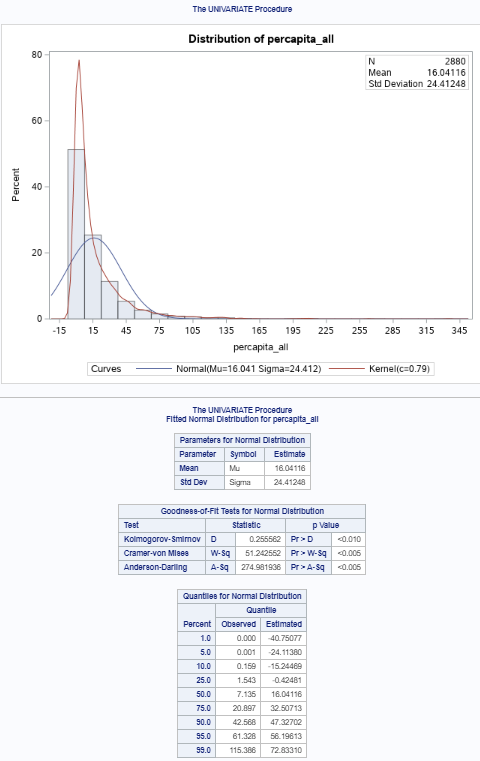


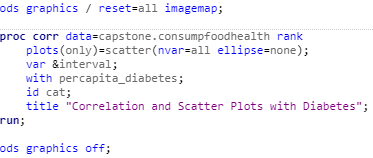
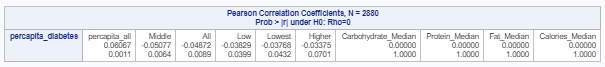
Image four: Percapita\_All Distribution

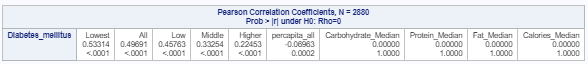
Image three: Carbohydrate\_Median Distribution

Above is the distribution analysis for median fat, protein, and carbohydrate levels for each food group, and also the per capita adjustment for the consumption rate. Interestingly, this per capita adjustment has drastically improved the normalcy of the distribution. The standard deviation indicated that the spread of the data was relatively small, which is also shown in each of the graphs. The Kolmogorov-Smirnov test is between .19 and .3 for each of the graphs, which again indicated that the difference between the data and the normal curve was significant. The more powerful Anderson-Darling and Cramer-von Mises test show similar results to the KS test. The above data shows irregularities around the center of the distribution, indicating that the KS and Cramer-von Mises tests are better representations of the distribution. These irregularities were also seen in the differences between observed and estimated quantiles for the distribution. Overall, each of the nutrient variables demonstrate a much closer-to-normal distribution than the consumption variables. Adjusting the nutrient variables with a logarithmic transformation proved unnecessary.

**II. Correlation Analysis**

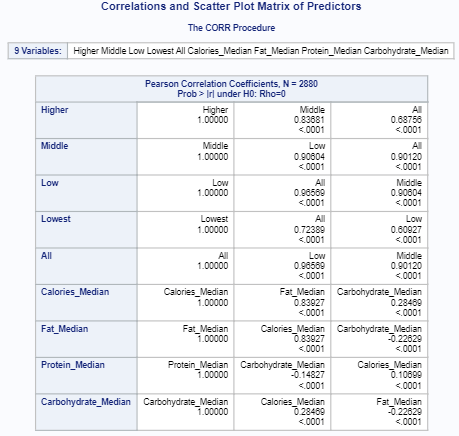
PROC CORR was used to produce correlation statistics and scatter plots in order to determine which variables are linearly associated with diabetes mellitus, chronic obstructive pulmonary disease, cardiovascular disease, and malignant neoplasms respectively. This analysis was designed to analyze relationships for further analysis and model building. The CORR procedure was utilized to calculate the Pearson Correlation Coefficients for the continuous variables. This method was selected to measure the strength of any linear relationship between the variables. Linear associations are a critical component of building a predictive model with the LOGISTIC procedure, which is why the Pearson Correlation Coefficient was selected as the methodology of choice. While many assumptions of this method were met, including that both variables were continuous, there was independence of cases, and the cases were paired, but one violation was the normality of the data (Laerd, 2020). Pearson’s Correlation was selected over Spearman’s correlation because the variables weren’t ordinal or ranked and was thus better represented by Pearson’s.





Above are the results of the correlation analysis for diabetes mellitus. These results show that, especially once there is an adjustment for population, the consumption rate is positively correlated to the rate of diabetes. When compared to the consumption rates that have been adjusted for income level, but not for population levels, there is a negative relationship. More than anything, this depicts the need for population adjustment levels in these areas. When diabetes is analyzed without a per capita adjustment, higher consumption is correlated with lower rates of diabetes mellitus. This is true of all of the consumption rates, but less so amongst the middle and higher-income classes. These results are representative of the correlation analysis for chronic obstructive pulmonary disease, cardiovascular disease, and malignant neoplasms. This analysis aided the model that would be built based on this research. The correlation analysis was concluded with a correlation table to investigate relationships amongst the predictors.

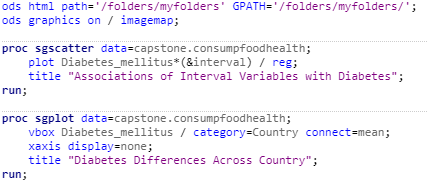




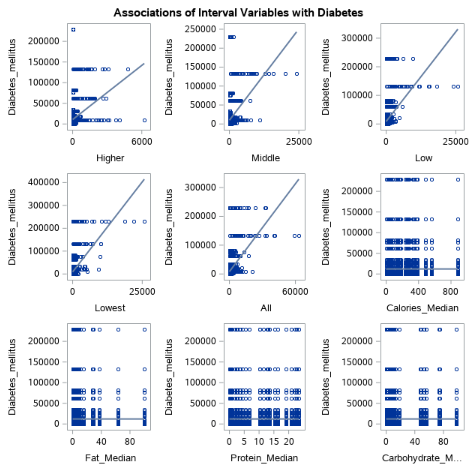
None of the variables had an unusual or surprising relationship with one another. Consumption rates had a significant correlation amongst themselves, as would be expected. Additionally, fat and calories were highly correlated, protein and carbohydrates had a negative relationship, and carbohydrates and fat had a negative relationship. These relationships provided a sound basis upon which to develop a model.

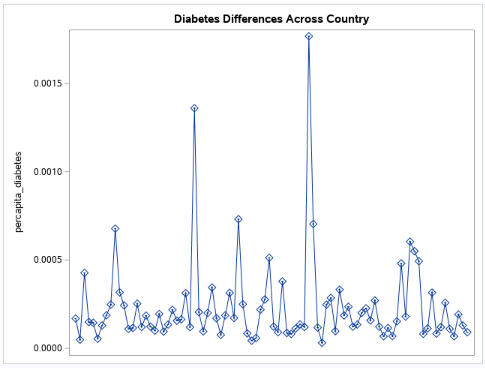
**III. Plotting the Data to Explore Associations**

After analyzing the correlations, it was necessary to investigate associations between the health outcome variables and the predictor variables. PROC SGSCATTER was used to develop scatter plots that would demonstrate the shape of the associations. The CORR procedure could have been selected to produce scatter plot matrices, but the matrix would not have trained the variables against diabetes specifically. For this reason, SGSCATTER was chosen. Then PROC SGPLOT was used to produce vertical bar charts. Diabetes mellitus was examined for its distribution by country. SGPLOT was the prime candidate for this graphical display because of its unique customization features and versatility.

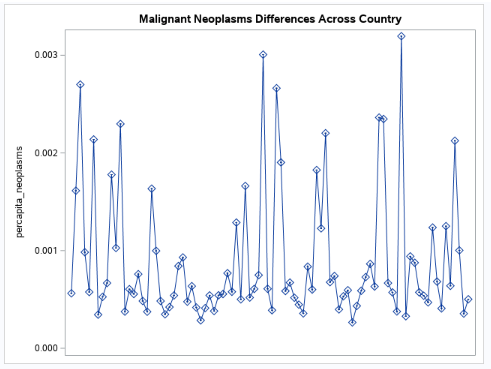


Above, the ODS options deliver x and y coordinate information in a data box as the user scrolled over a grid point. For this reason, the x axis is not displayed in the SGPLOT procedure.

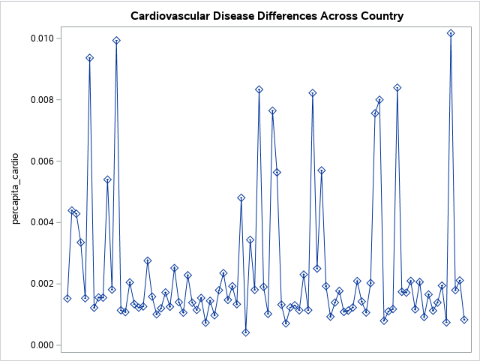




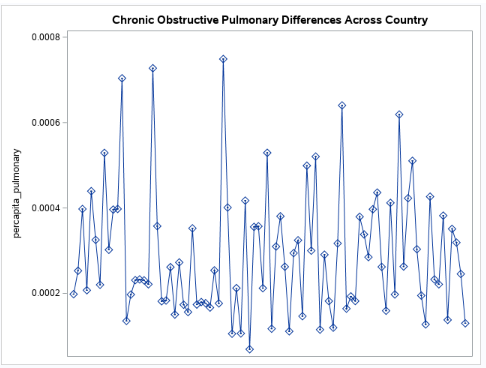
The associations in the scatter plots demonstrate the same results found in the correlation analysis, confirming the associations between consumption rates and diabetes, as well as the decreased strength in the associations amongst high-income and middle-income areas. Interestingly, the diabetes rates vs. country plot demonstrates that there is a cause-effect relationship worth investigating. There are countries whose per capita diabetes rate far exceeds that of other nations. In the above graph, Mauritius has the highest rate, followed in descending order by Fiji, Jamaica, Mexico, and Bosnia and Herzegovina. The model will attempt to explain possible predictors for these higher rates.



Shown above: per capita Malignant Neoplasm rates are led, in order, by Serbia, Latvia, Armenia. Lithuania, and Bulgaria



Shown above: per capita cardiovascular disease rates are led, in order, by Ukraine, Bulgaria, Belarus, Serbia, Latvia, Moldova

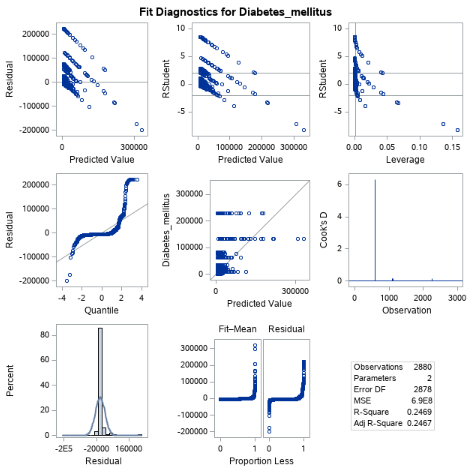
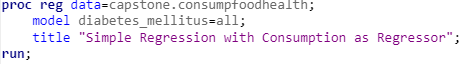


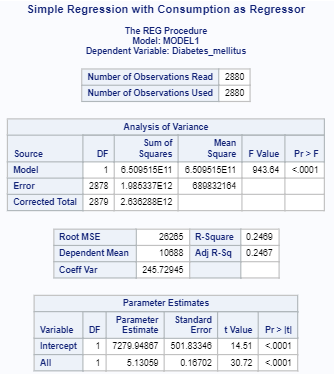
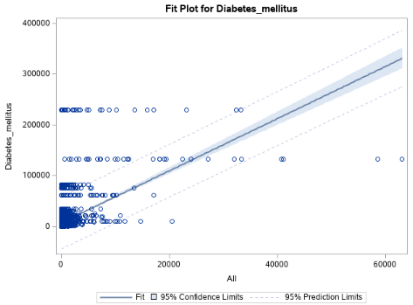
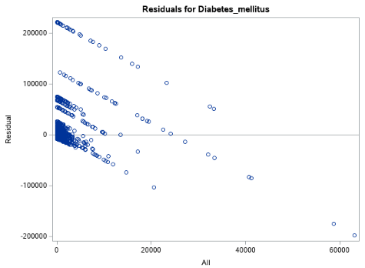
Shown above: per capita chronic obstructive pulmonary disease rates are led, in order, by India, China, Bulgaria, Nepal, Serbia

The differences in leading nations for each noncommunicable disease category opened a new possibility that different food category consumption rates could lead to higher rates of some noncommunicable diseases and not others. This was explored in the model building process.

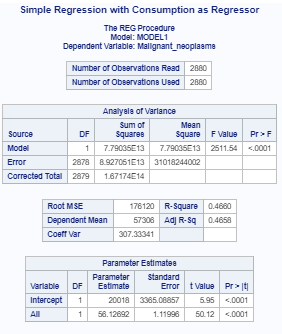
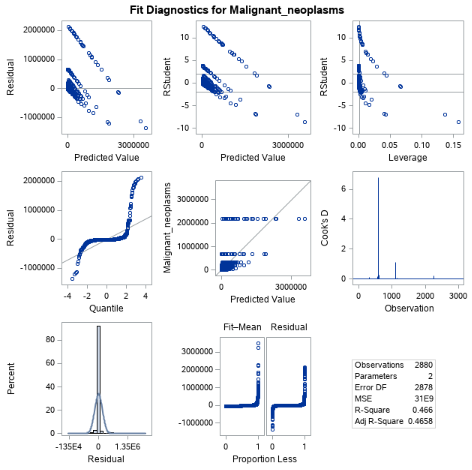
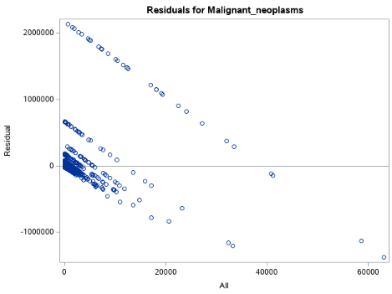
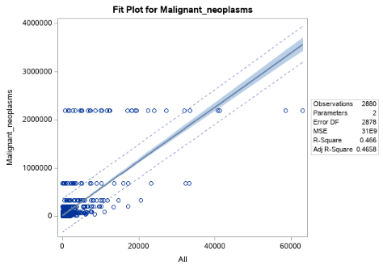
**IV. Simple Linear Regression**

In order to determine whether the rate of diabetes differs across various consumption rates, a linear regression was performed. The assumptions of the model were examined with Levene’s Test of Homogeneity, Q-Q plots, and residual plots. Levene’s Test was chosen over the Bartlett Test because it is less sensitive “to departures from normality” (Croarkin and Tobias, 2013). This is a critical aspect as the data does not follow the normal curve, though the downfall is that Levene’s Test is less sensitive. The Q-Q plot was selected to demonstrate whether the distribution of the residuals is normal and a residual plot to examine the dispersion of the data. These metrics are to determine whether the relationship between the variables meets basic assumptions prior to building a model. The REG procedure was selected because it maximizes the code to statistical results ratio. While the UNIVARIATE procedure will also generate the Q-Q Plot, the REG procedure offers more plots and statistics for in-depth analysis with a small amount of code.

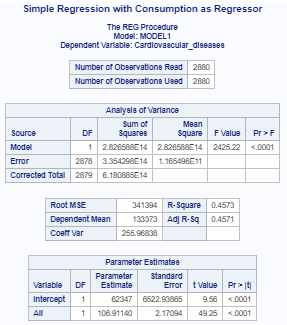
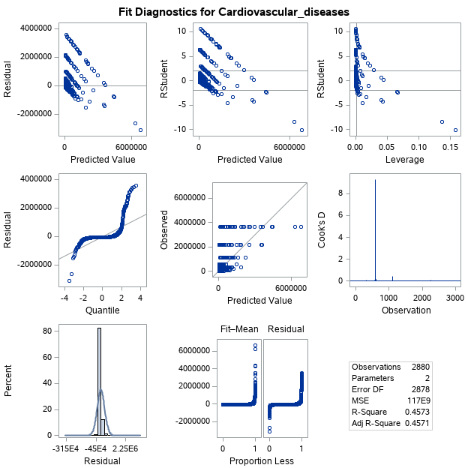
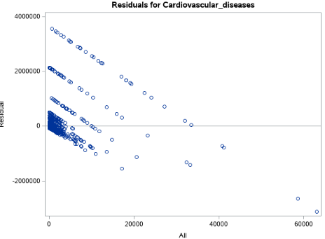


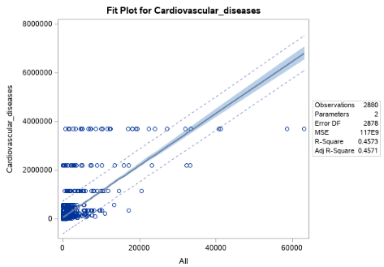


The p-value above is less than .05, so the regression fits the model better than a baseline. From the analysis, there is a significant linear relationship between diabetes and consumption rates. The coefficient of determination indicates that consumption rates explain 24.6% of the variation in diabetes. The Q-Q Plot and histogram of residuals fall relatively close to a straight line and bell-shaped curve, meaning that the deviations deviate from normality insignificantly.

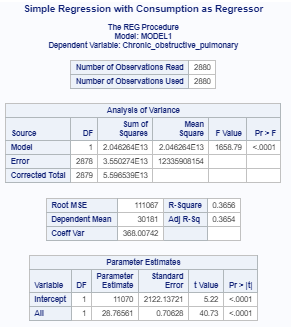
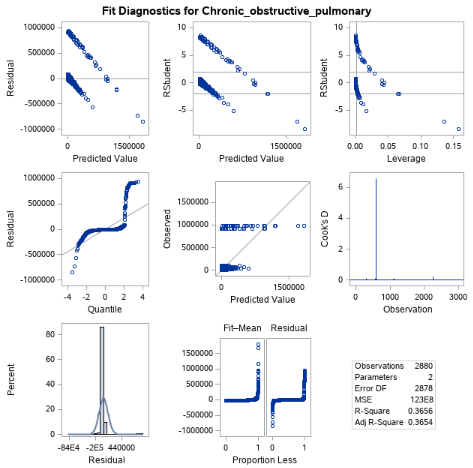


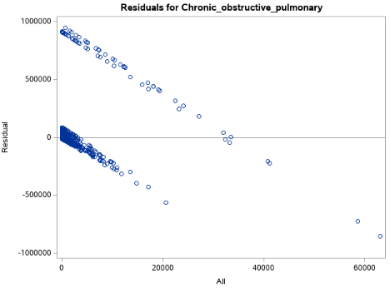
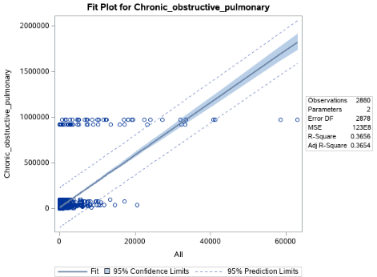












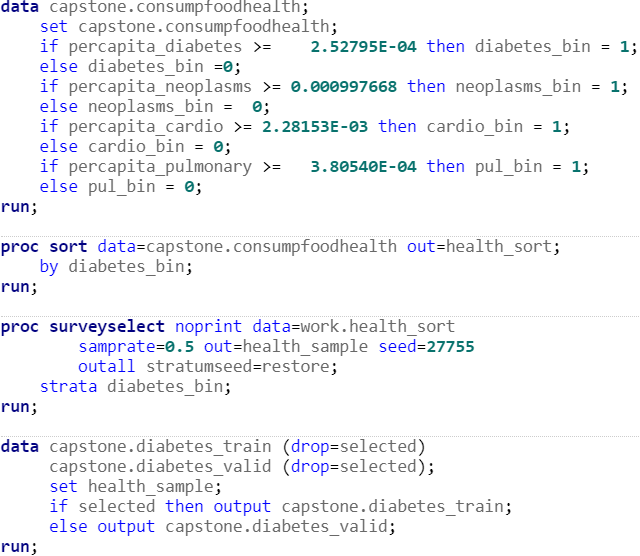
The above plots are very similar to the diabetes vs consumption rates. They are all related linearly and the models are better than the baseline. Conversely, the normality of each is not exact. These results indicated this relationship between consumption and health outcomes was strong.

**E. Predictive Modeling Diabetes, Malignant Neoplasms, Cardiovascular Disease, Chronic Obstructive Pulmonary Disease**

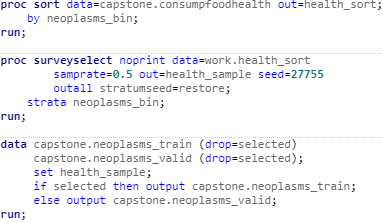
**I. Splitting the Data**

The model aims to identify predictors for high rates of NCDs. This was defined by the upper quantile or 25% of the data, in order to identify the upper quantile of the data as the target, a binary variable was created for each NCD. The data was then sorted by the diabetes mellitus binary variable and followed with the SURVEYSELECT procedure in order to split the data into training and validation data sets. SURVEYSELECT was chosen over the DATA step, which would have required more code and ended with a messier result. SURVEYSELECT was the more efficient method. Below are the data splitting procedures for diabetes, malignant neoplasms, cardiovascular disease and chronic obstructive pulmonary disease respectively.

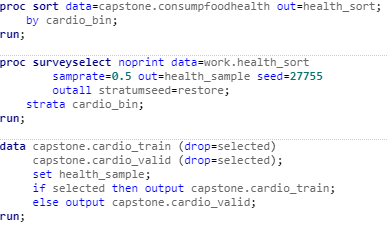
Diabetes:



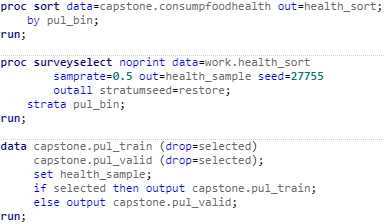
Malignant Neoplasms:



Cardiovascular Disease:



Chronic Obstructive Pulmonary Disease:



In the above code, the data is split 50/50 into training and validation datasets. Additionally, per capita rates of the NCDs are used in order to highlight the differences in health outcomes as opposed to differences in population rates.

**II. Handling Missing Values**

Prior to creating a model, it is necessary to remove missing values. The following step uses the MEANS procedure to identify any missing values in the training dataset. This method proved the simplest and shortest method for calculating the number of missing values across the dataset. By using the original dataset and not the training and validation datasets to calculate missing values, this eliminated the need to repeat this step on the validation dataset.

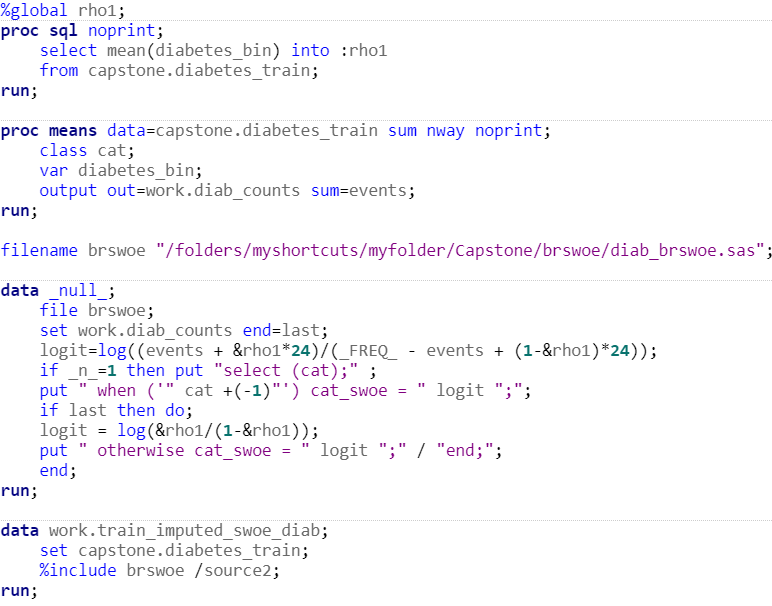


This procedure demonstrated that there are no missing values in the dataset, and thus, no further adjustments are necessary.

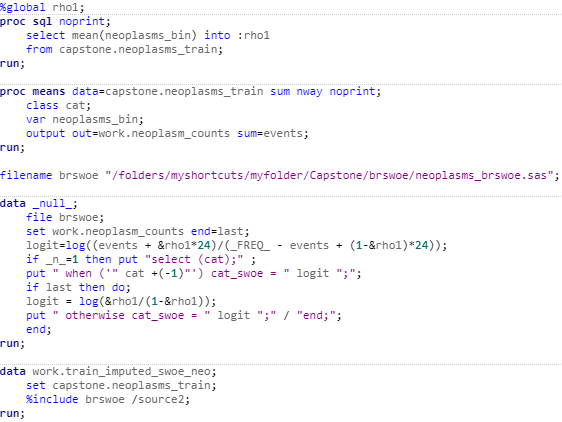
**III. Computing Smoothed Weight of Evidence**

When dealing with a categorical predictor, computing a smoothed weight of evidence assists with creating a predictive model. Categorical variables cause problems if left alone due to high dimensionality and quasi-complete separation. In this case, the food category variable necessitated smoothed weight of evidence to turn it into a continuous variable. Smooth Weight of Evidence reduced the risk of dimensionality that the alternative method, dummy coding, would have caused. Using this statistical method insured that sampling variability was accounted for and overfitting was avoided. Below is the code for computing smoothed weight of evidence for diabetes, malignant neoplasms, cardiovascular disease and chronic obstructive pulmonary disease.

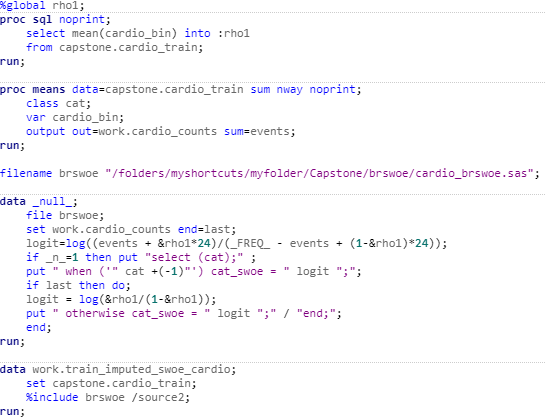
Diabetes:



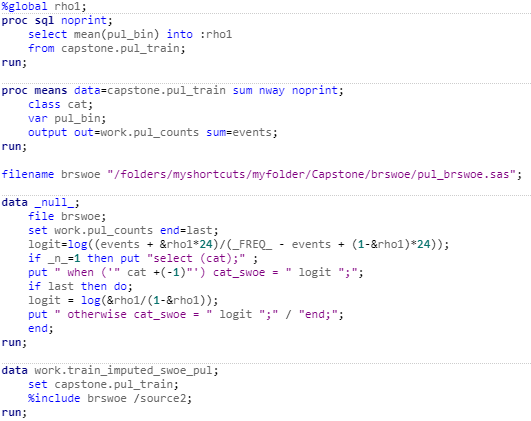
Malignant Neoplasms:



Cardiovascular Disease:



Chronic Obstructive Pulmonary:

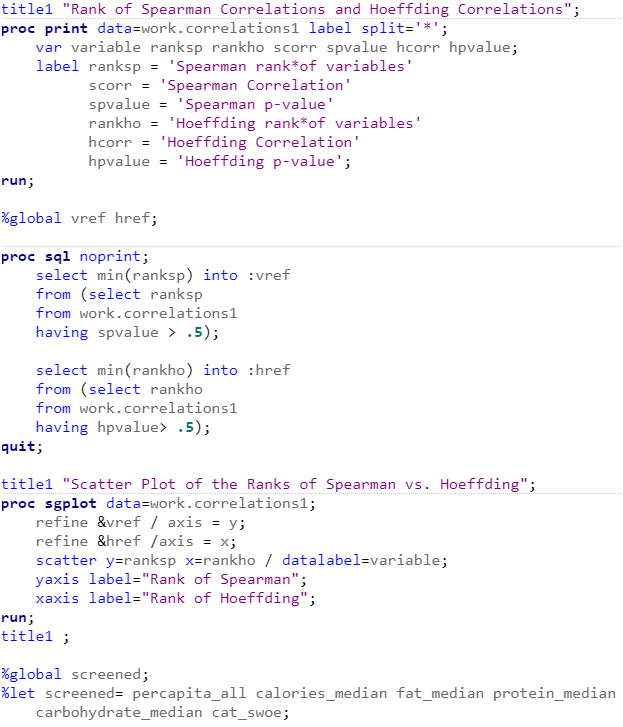


The above data takes the frequency of the target variable in each category and takes the logit in order to create a continuous variable that will be utilized in the model building phase.

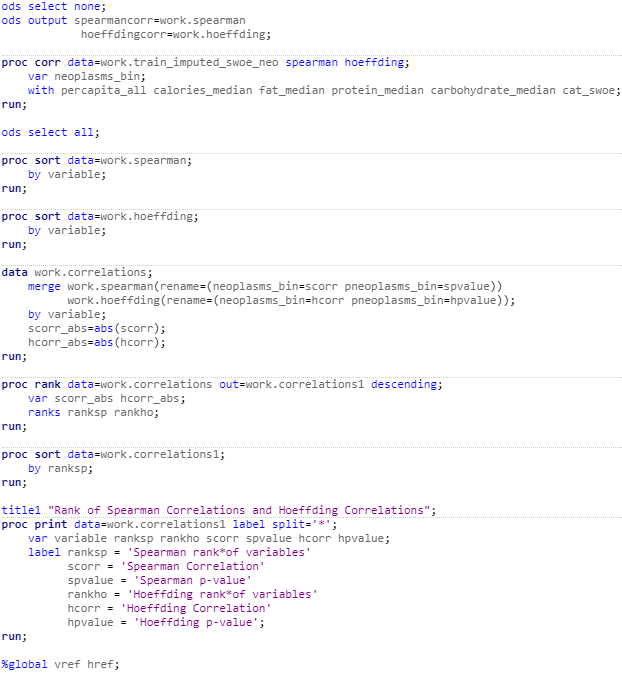
**IV. Detecting Nonlinear Relationships**

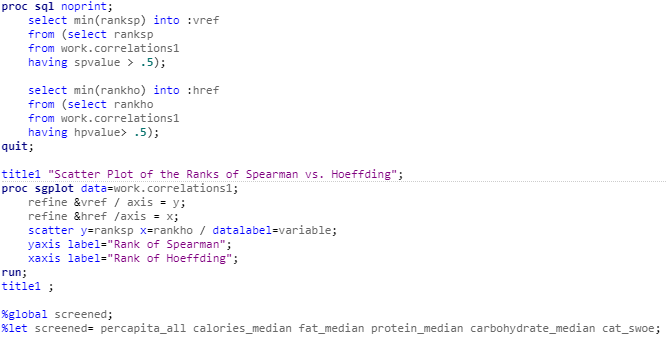
Due to the fact that logistic models are designed to best represent linear relationships, if nonlinear relationships exist in the data, adjustments must occur in order to proceed to model building. This study used Spearman’s Correlation and Hoeffding’s D statistics to evaluate whether the variable relationships were nonmonotonic with a high association, or nonlinear. Spearman’s Correlation is used here as opposed to Pearson’s for its lower sensitivity to outliers. Alternatively, TRANSREG can be used to model nonlinear relationships visually, but it does not illustrate the strength of these nonlinear relationships numerically nor is it possible to compare them easily across variables. Thus, Spearman’s and Hoeffding’s were selected. Below is the code for the detection of nonlinear relationships for diabetes:



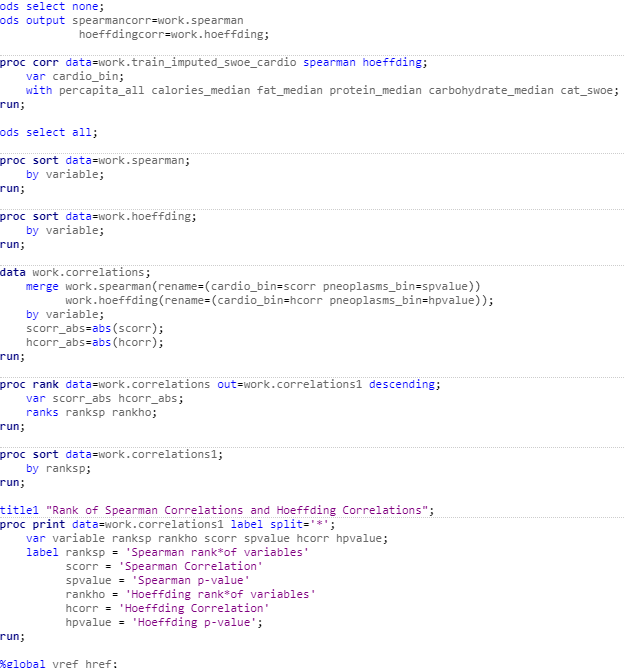


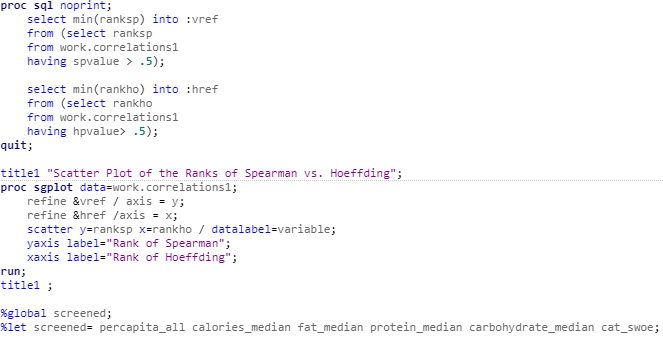
Malignant neoplasms:



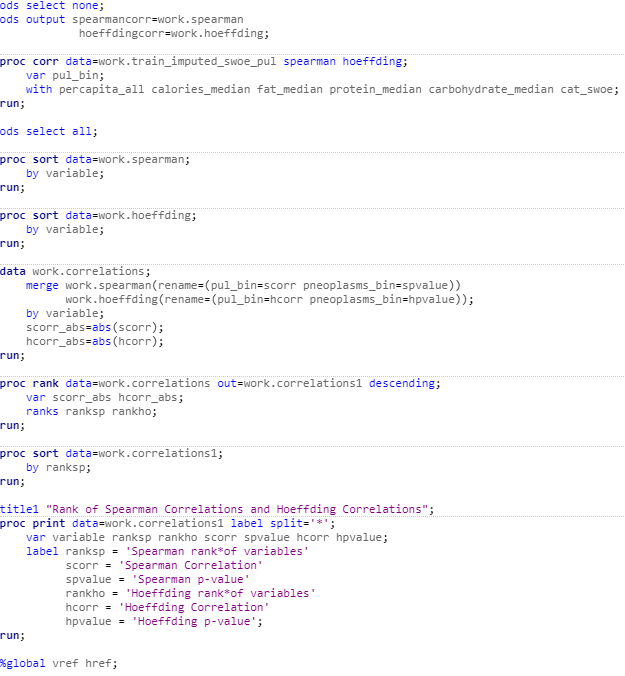


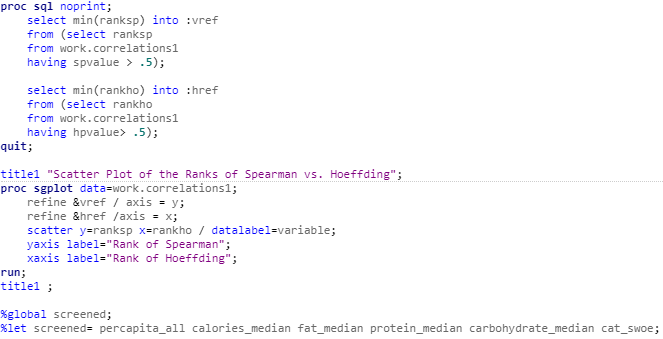
Cardiovascular disease:



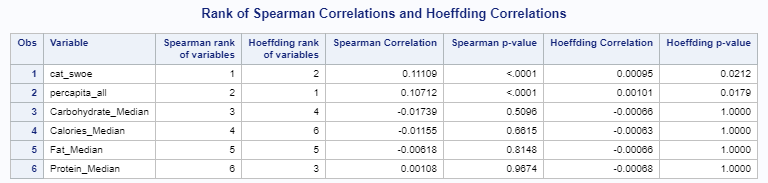


Chronic obstructive pulmonary disease:





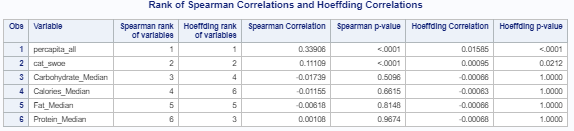
In the above code, the CORR procedure was used to create two files with the Spearman and Hoeffding statistics respectively. The two files were sorted by the variables and then merged. Two new variables were added that take the absolute value of the Spearman and Hoeffding statistics in order to sort the relationships between variables. The RANK procedure was then used to rank the statistics prioritizing the smallest statistics with the DESENDING option. This procedure output the following results for diabetes:



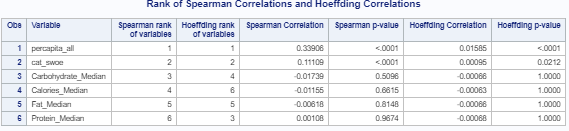
Malignant neoplasms:



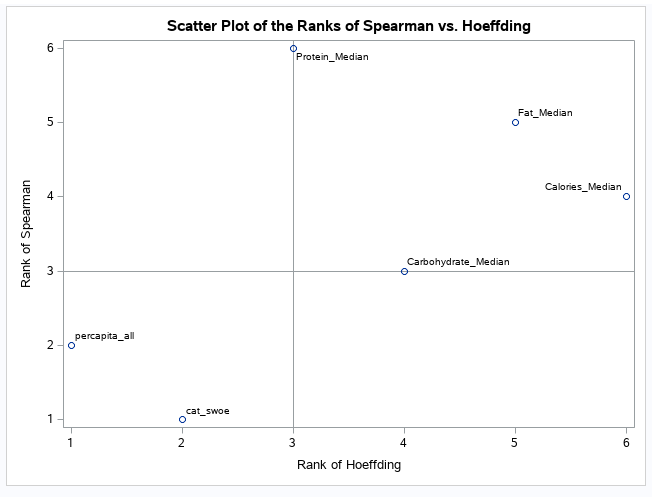
Cardiovascular disease:



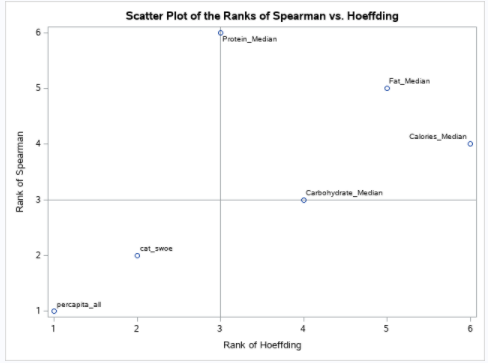
Chronic obstructive pulmonary:



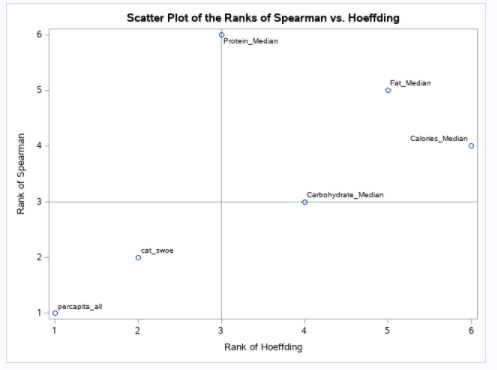
The ranks show that a nonlinear relationship likely did not exist. In order to highlight this further, a SQL procedure was used that highlighted the smallest Spearman and Hoeffding rank with a p-value greater than .5. Then SGPLOT was used to visually illustrate the statistics. The following graph was the output for diabetes:



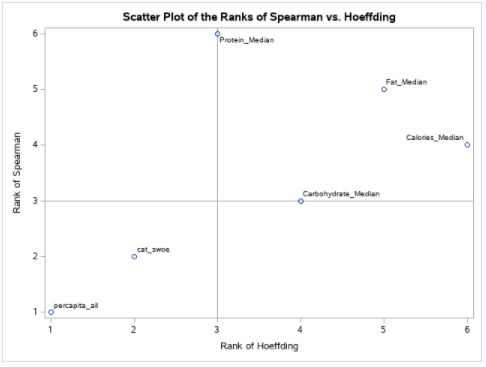
Malignant neoplasms:



Cardiovascular disease:



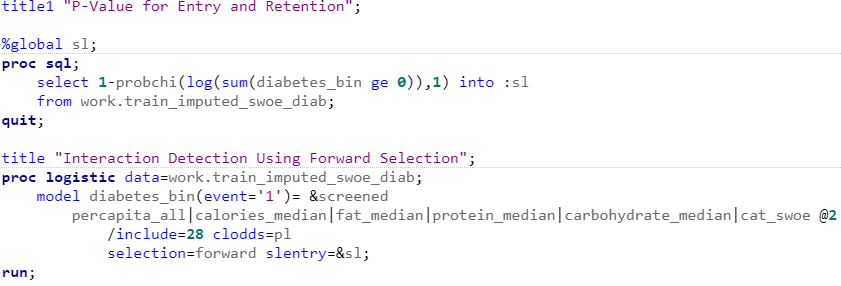
Chronic obstructive pulmonary:



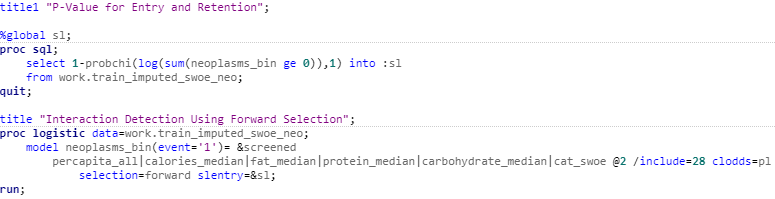
Due to the fact that there are no variables in the upper left quadrant, indicating a high Hoeffding and low Spearman statistic. This meant there were no linear relationships that necessitated adjustment prior to model building.

**V. Interaction Detection**

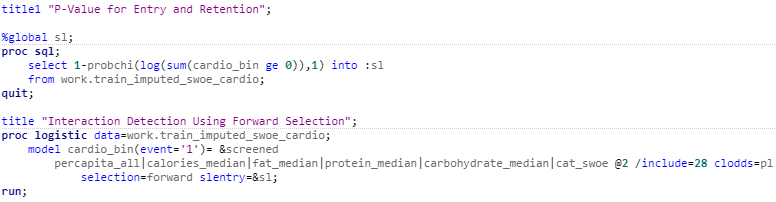
The data preparation then necessitated the discovery of interactions between model building. Detecting these interactions would assist in the most comprehensive model possible. The following code was used to determine interactions across the data for diabetes:



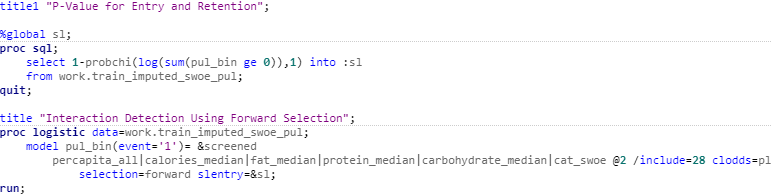
Malignant neoplasms:



Cardiovascular disease:

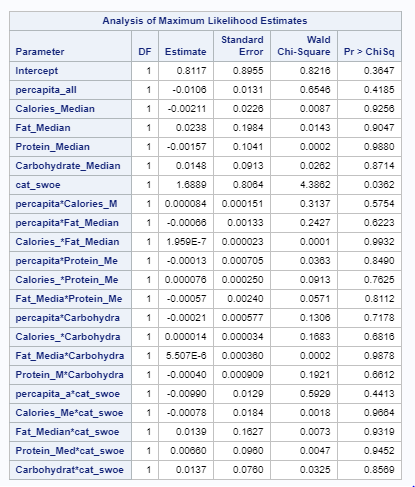


Chronic obstructive pulmonary:



The above code first uses SQL to determine the log of the number of observations. This figure would be the p-value utilized for entry and retention. The next procedure used forward selection in order to detect important two-factor interactions. This method was selected over the EFFECTPLOT procedure, which would have required a macro to iterate through the procedure for as many interaction plots necessitating presentation. Forward selection was chosen above backward selection because it is more efficient and does not have to consider the full model. Backward selection would have been selected if collinearity were present, in order to keep those variables in the model, but it was not. Finally, Stepwise selection was also considered, but due to the biased coefficients and p-values it produces, was not chosen in the end (Choueiry, 2020). The purpose of this method was to identify key interactions in the dataset that could be verified and selected with further analysis. The following tables show the results of interaction detection for diabetes, malignant neoplasms, cardiovascular disease, and chronic obstructive pulmonary.

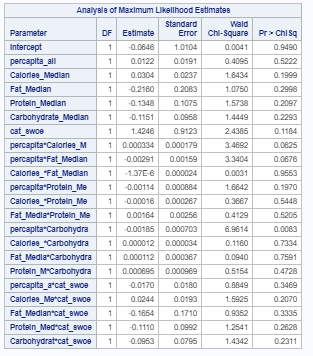
Diabetes:



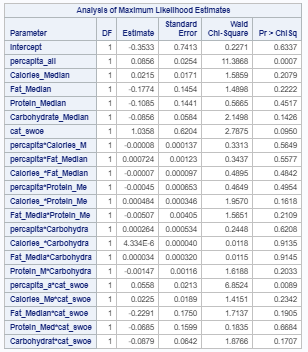
Malignant neoplasms:



Cardiovascular disease:



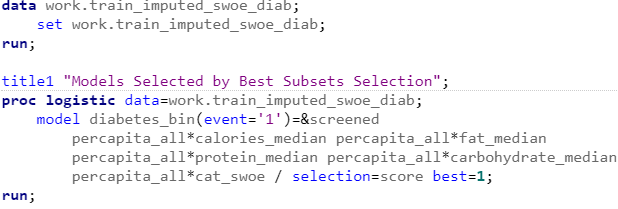
Chronic obstructive pulmonary:

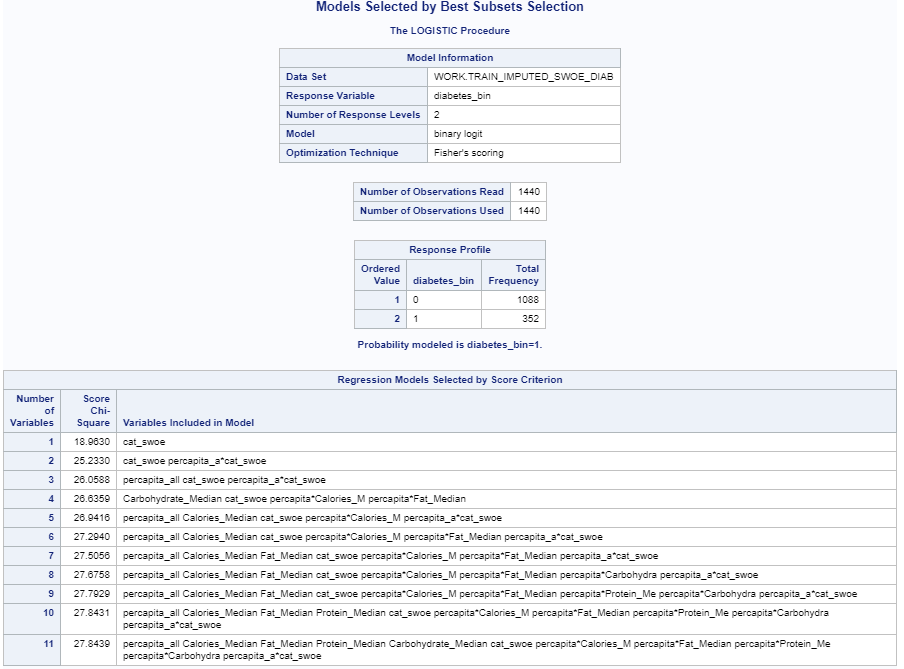


The above are the results of the interaction detection procedures. This illustrated that 21 interactions were detected. Once interactions were identified, this necessitated further exploration.

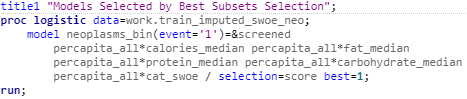
**VI. Best Subsets Selection Method**

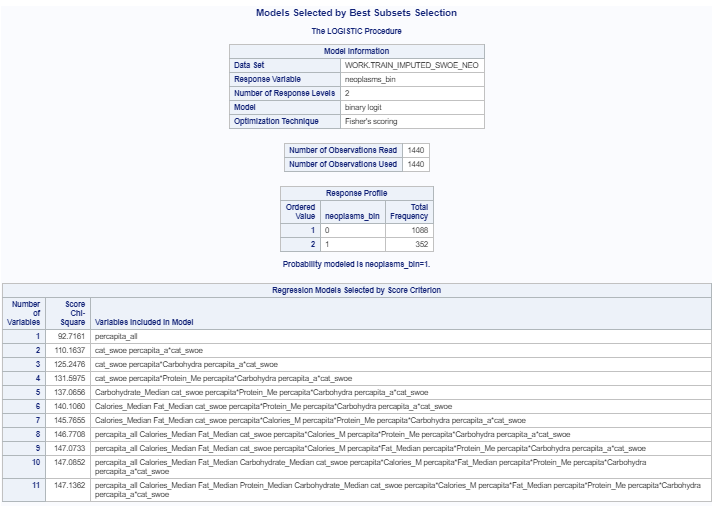
In order to determine the best model for each size, the best subset selection was used. Best subsets selection was chosen over stepwise because of the more in-depth information it presents. Stepwise selection would have arrived at a single best model, whereas best subset selection presents the best models for each number of variables, allowing the analyst to balance knowledge of the data, complexity, and best fit. This method selected the best model for each variable size based on the chi-square statistic. The chi-square statistic compares the actual data to the model predictions, causing it to be a useful tool in model selection. The following depicts the code for diabetes:



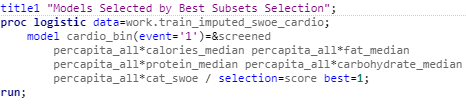


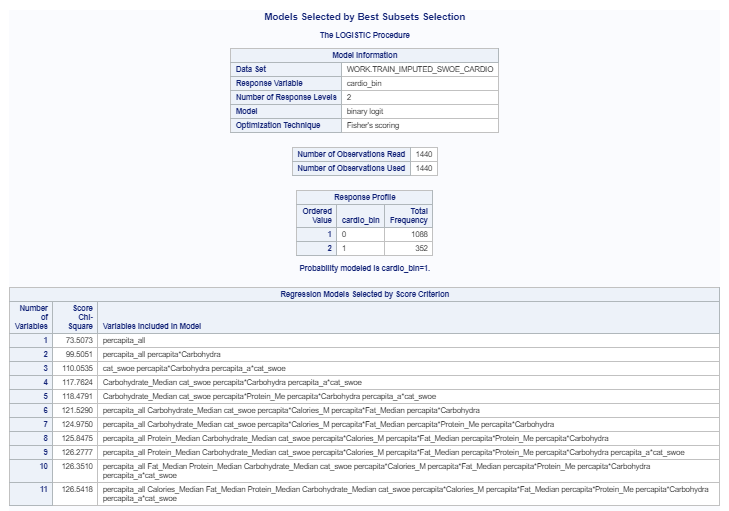
Malignant neoplasms:



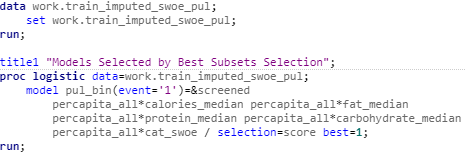


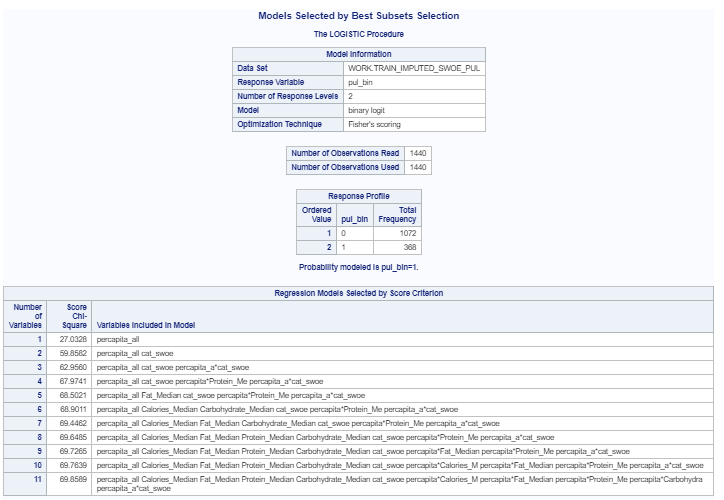
Cardiovascular disease:





Chronic obstructive pulmonary:

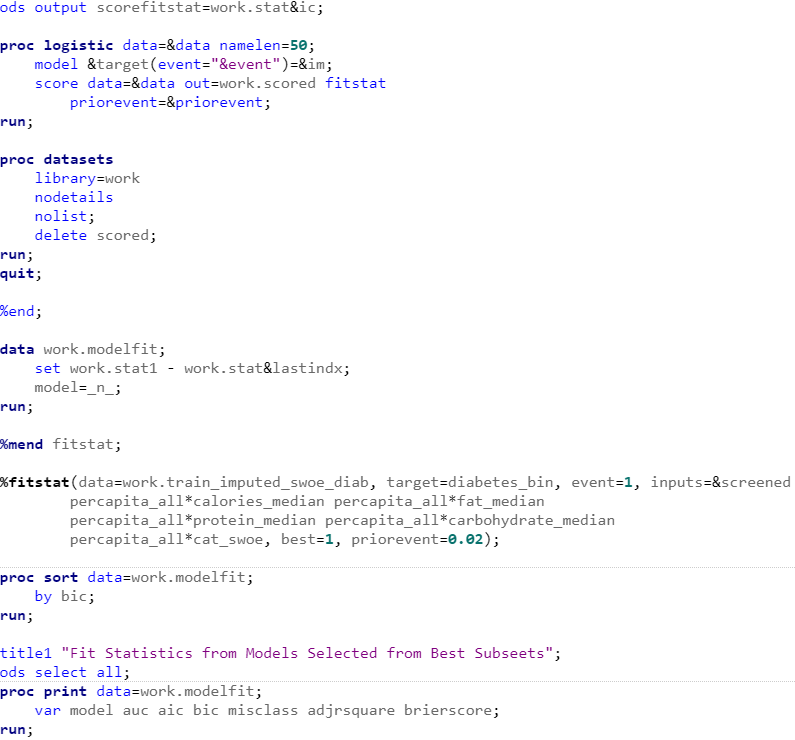
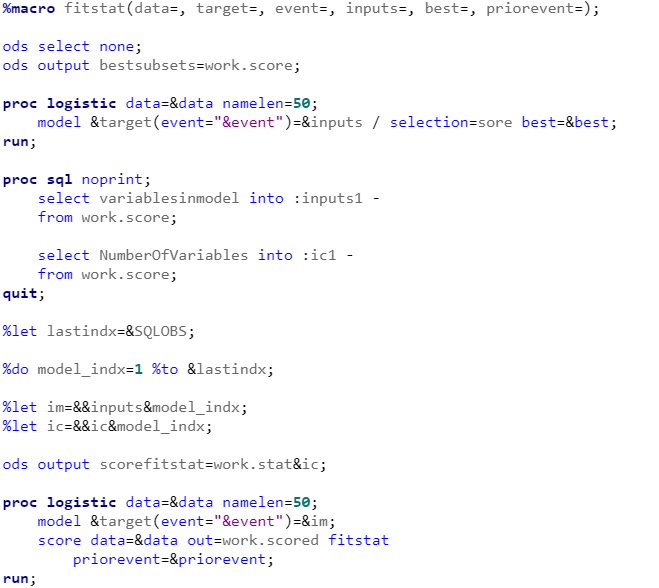




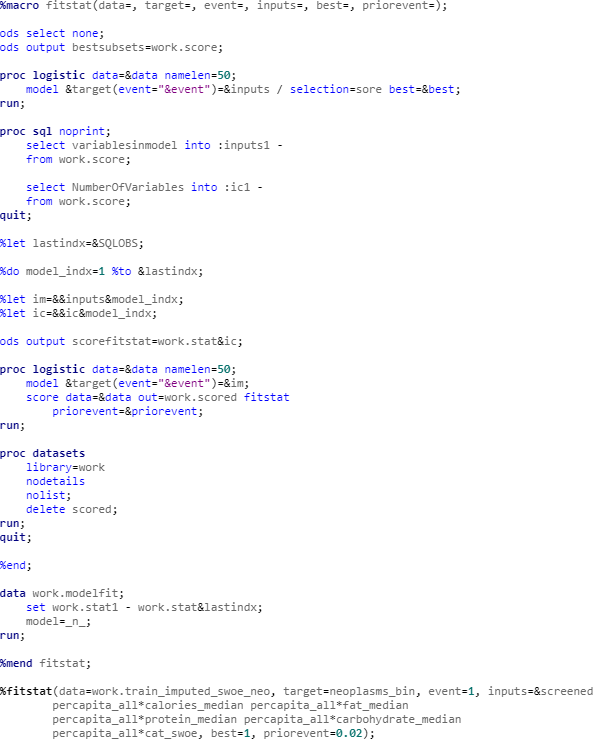
The above code uses the LOGISTIC procedure to conduct best subsets selection. Interestingly, cat\_swoe and percapita\_all, as well as their interaction are front runners in the model selection. This confirms the initial hypothesis that food category consumption rates may be a predictor for noncommunicable disease. The scoring technique used is not the best method for determining final model as it is skewed by model size. Therefore, the study continues with fit statistics to determine the best model.

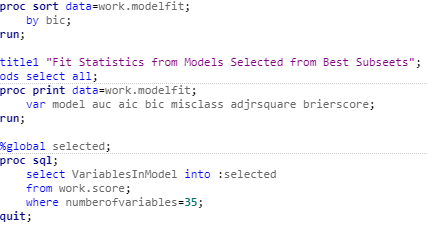
**VII. Fit Statistics for Model Selection**

Fit Statistics were selected as the most comprehensive method in model selection. The following code develops a FITSTAT macro to generate the fit statistics. FITSTAT in the LOGISTIC procedure was selected above the PHREG procedure, which would have depicted the Akaike Information Criterion, Schwarz Bayesian Criterion, and -2 log likelihood statistic. This occurred because the PHREG procedure did not offer enough statistics. An iterative loop is used to score each of the models generated and gather the statistics in a single file. Below is the loop code for diabetes:

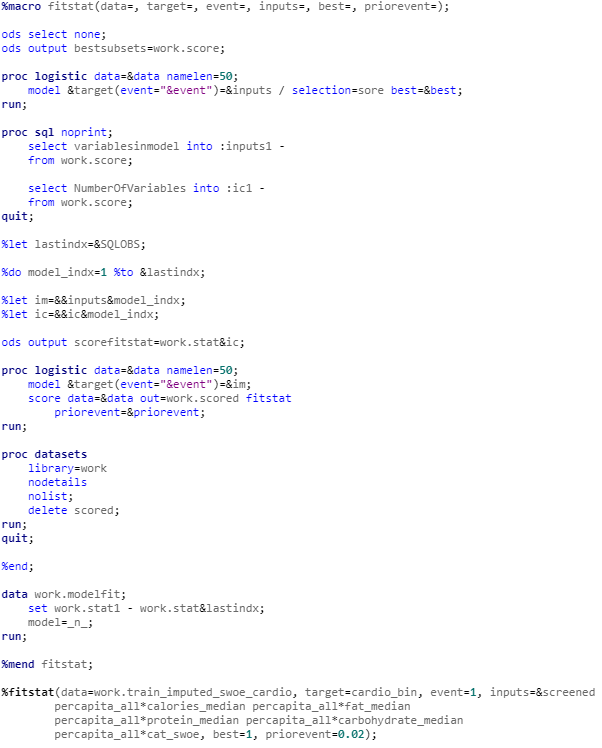


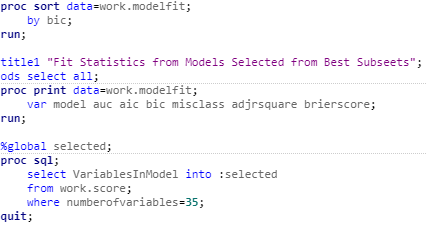
Malignant neoplasms:



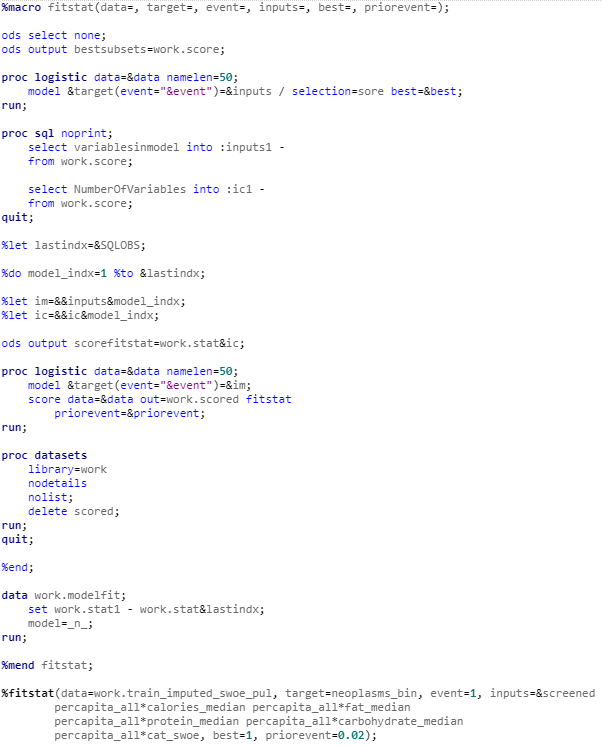


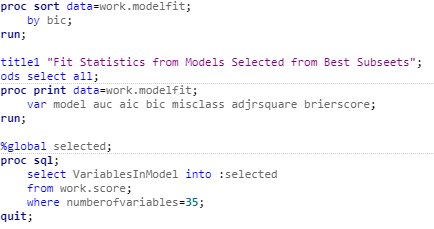
Cardiovascular disease:





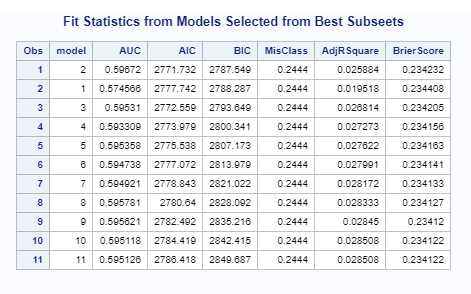
Chronic obstructive pulmonary:





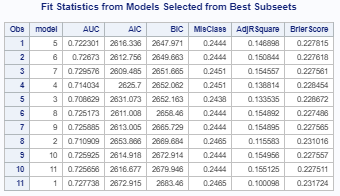
The above code generated all possible models with fit statistics for each and then saved the results into a file called modelfit. The file was then sorted by the Bayesian Information Criterion and the results were printed.

Fit statistics for diabetes:



As seen from the above diabetes table, the model with a low Bayesian Information Criterion combined with a sufficiently high adjusted r-square value is the model with three variables: percapita\_all, cat\_swoe, percapita\_all\*cat\_swoe. When analyzing the fit statistics, the adjusted r-square values for diabetes decrease at the second model, then increase only marginally after the third. For this reason, the third model was selected.

Malignant neoplasms:



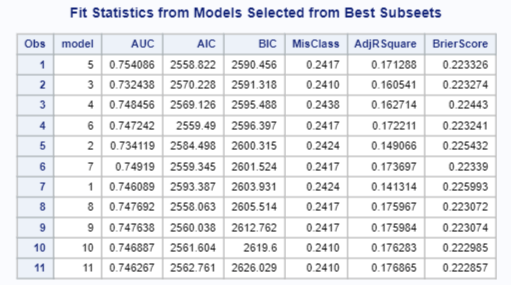
In the above malignant neoplasms table, the model with a low Bayesian Information Criterion and high adjusted r-square value is the model with three variables: cat\_swoe, percapita\_all\*Carbohydrate\_Median, and percapita\_all\*cat\_swoe. For this NCD’s fit statistics, the adjusted r-square values begin to decrease after the third model, then rise again only slightly after the fifth model. The third model was selected here again.

Cardiovascular disease:



For cardiovascular disease, the model with a low Bayesian Information Criterion combined and high adjusted r-square value is the model with four variables: Carbohydrate\_Median, cat\_swoe, percapita\_all\*Carbohydrate\_Median, and percapita\_all\*cat\_swoe. Looking at the fit statistics, the adjusted r-square values begin to decrease at the third model, rise for the fourth, then decrease again before a marginal improvement with more complex models. For this case, the fourth model was selected.

Chronic obstructive pulmonary:



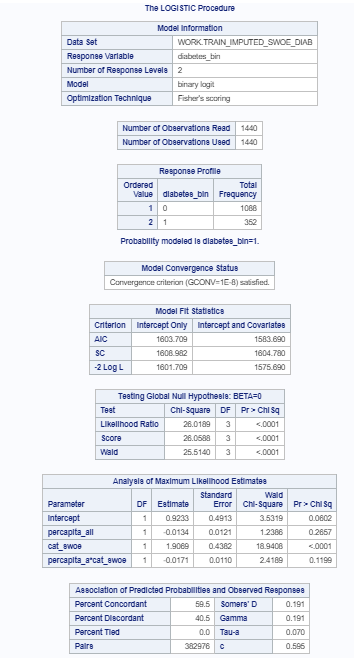
In the case of chronic obstructive pulmonary disease, the best model based on the outlined criteria was the model with four variables: percapita\_all, percapita\_all\*Fat\_Median, percapita\_all\*Protein\_Median, and percapita\_all\*cat\_swoe. When considering the fit statistics, the adjusted r-square value rises until the fourth variable, then decreases, with more complex models maintaining only a marginal improvement to the four variable model r-square value. In the end, these models were chosen because they balanced lower complexity, representation of the data, and the Bayesian Information Criterion.

**VIII. Model Selection**

Each selected model was then trained with the LOGISTIC procedure to examine the model further.

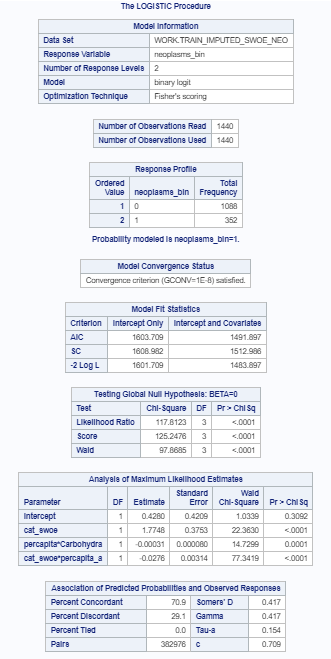
Diabetes:





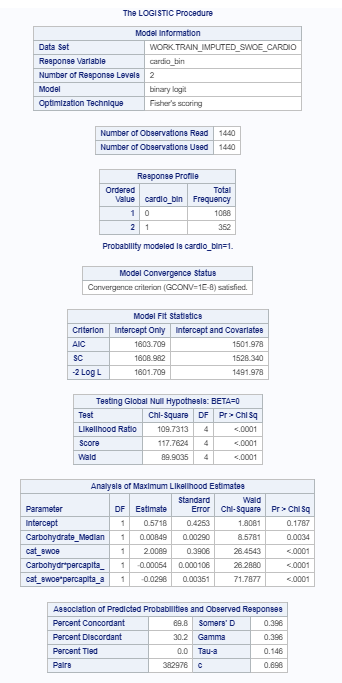
Malignant neoplasms:





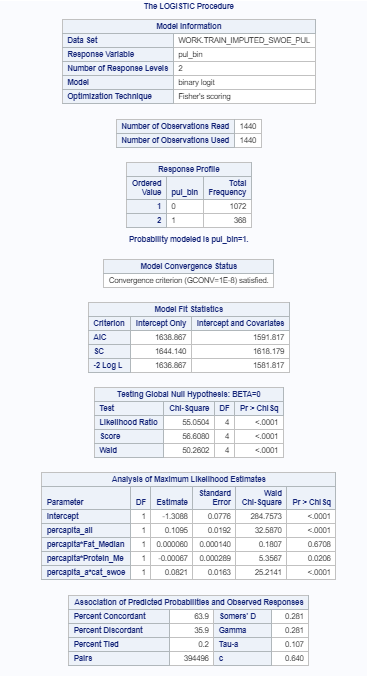
Cardiovascular disease

:



Chronic obstructive pulmonary:





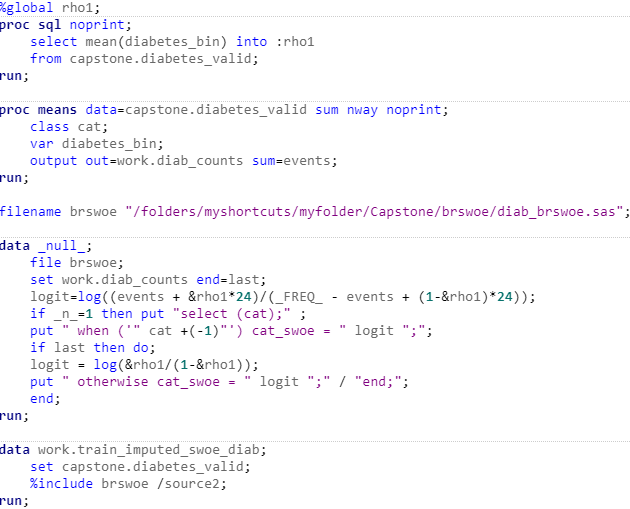
As is clear from the results of each procedure, the concordant pairs greatly outnumber the tied and discordant pairs, indicating the models are representative of the data. Of note, the training data c statistic is .595 for diabetes, .709 for malignant neoplasms, .696 for cardiovascular disease, and .640 for chronic obstructive pulmonary disease. This would prove useful for measuring the model performance later.

**F. Measuring Model Performance**

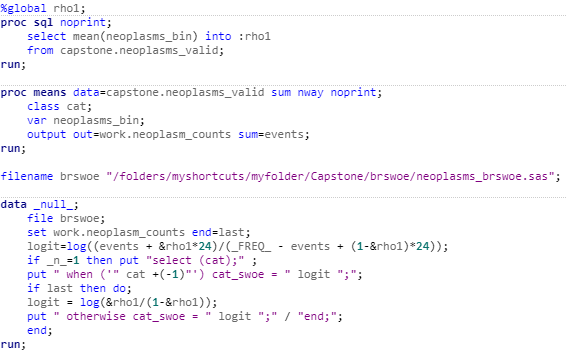
**I. Preparing the Validation Data**

Prior to scoring the validation data, the validation data set must be adjusted for the new variables that were added to the training dataset. This is done to most accurately measure the model’s ability to generalize to new data. In this case, only a smooth weight of evidence variable necessitates addition. The same procedure was conducted for each NCD.

Diabetes:



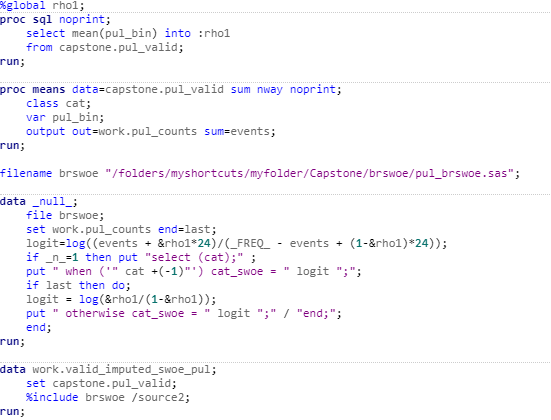
Malignant neoplasms:



Cardiovascular disease:

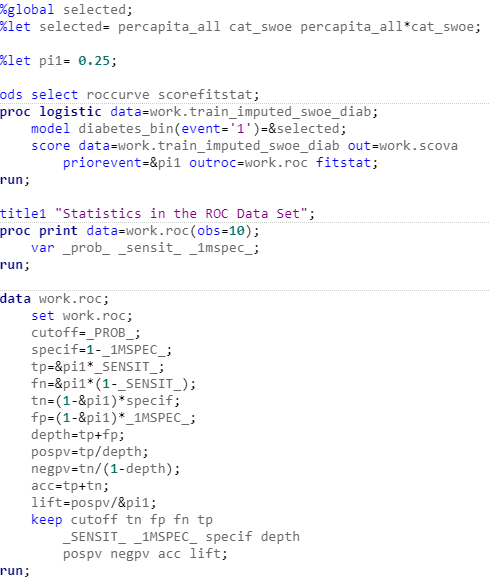


Chronic obstructive pulmonary:

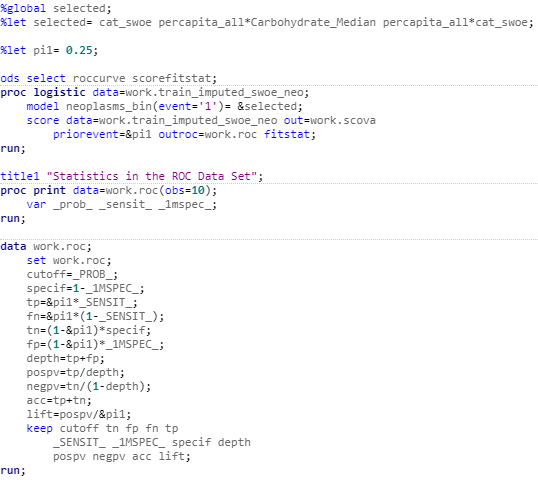


The cat\_swoe variable was added to the validation dataset using the same methodology and the data was then prepared for subsequent scoring and analysis.

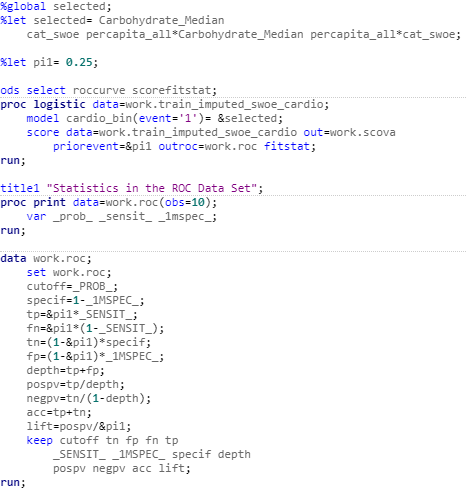
**B. Measuring Model Performance**

A receiver-operator curve (ROC) was utilized to illustrate the predictive power of the model. The receiver-operator curve is a “performance measurement for [the] classification problem at various threshold settings.” The area under the curve/c statistic is the ability the model has to predict the class to which the target will belong (Narkhede, 2018). This method was selected because it is widely used and easily interpreted. The gains and lift charts could also depict the same information, but may necessitate a more extensive explanation to the audience. Below is the ROC curve code for diabetes:

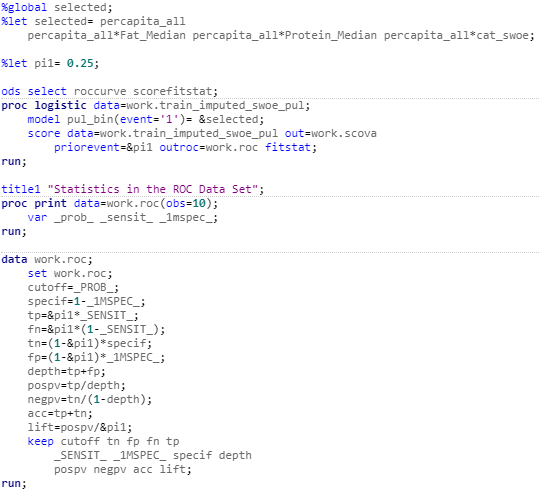
Malignant neoplasms:



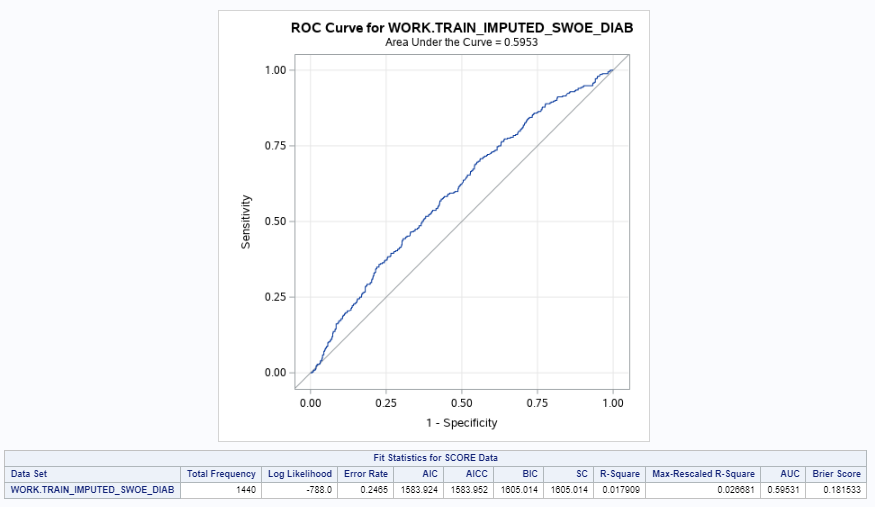
Cardiovascular disease:



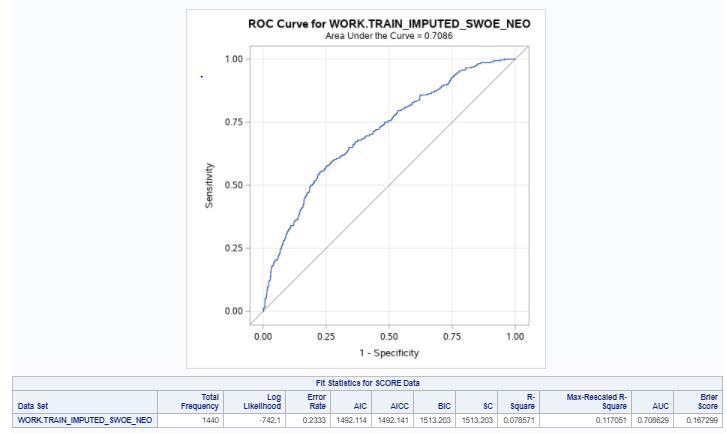
Chronic obstructive pulmonary:



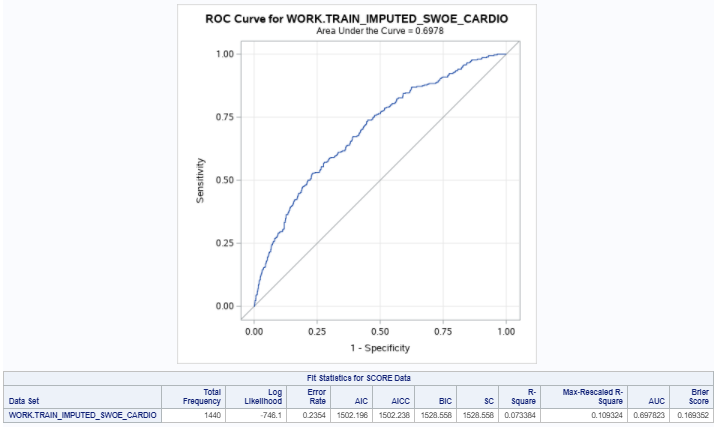
The above code depicts the fit statistics of the model on the new dataset and a receiver-operator curve to illustrates the sensitivity against one minus specificity. The curve is shown below for diabetes:



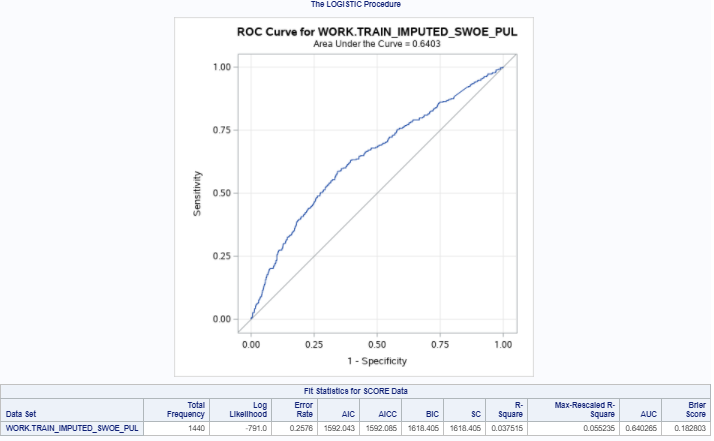
Malignant neoplasms:



Cardiovascular disease:



Chronic obstructive pulmonary:



As seen, the area under the diabetes curve was .5953, which was extremely close to the training data set c statistic of .595. The area under the malignant neoplasms curve was .7086 which was also close to the training data set c statistic of .709. The cardiovascular disease area under the curve was .6978, close to the training c statistic of .696, and finally the area under the curve for chronic obstructive pulmonary disease was .6403, which was again similar to the training c statistic of .640. The similarities between the areas under the curve and the training data set c statics indicate the model generalized well to new data.

**G. Data Summary and Implications**

In the end, the model that best fit the data was indeed the consumption rate, food category, and food category \* consumption rate interaction for diabetes. The best fit for malignant neoplasms was the food category, consumption rate \* carbohydrate interaction, and food category \* consumption rate interaction. The best model for cardiovascular disease was carbohydrate, food category, the food consumption \* carbohydrate interaction, and the food consumption \* food category interaction. Finally, the best model for chronic obstructive pulmonary disease involved food consumption rate, the food consumption \* fat interaction, the food consumption \* protein interaction, and the food consumption \* food category interaction. Ultimately nutrition facts, food categories, and consumption rates are excellent predictors for NCDs. Each of the models provides an accurate method for predicting rates of NCDs based on certain criteria.

However, the data is limited by the confusion caused by using a smoothed weight of evidence to represent the food category variable. As a result of this fact, the data is difficult to interpret. Though, if the data was interpreted with this in mind, the model results can serve as an indicator for emerging nations to shift the focus of their markets to certain food groups that are not correlated with higher rates of NCDs.

A main limitation of this analysis is the cumulative effect of limitations within each dataset. The food consumption rates are based on the consumption as best estimated by The World Bank. Accurate consumption rates were not available due to government entities failing to disclose, as well as the unavailability of rates due to informal markets. This combined with the fact that the median nutrition rates for each food category may not be the most accurate representation of the food category is a major constraint.

In summary, there is clearly a significant relationship between food consumption rates, the categories of food being consumed, the nutrition facts of those foods, and the subsequent health outcomes. This relationship should be further analyzed with more data that analyzes specific food groups combined with finely tuned nutrition information. This would provide a better model that would assist the resourcing of developing nations. A more in-depth analysis of the specific areas within these countries that are consuming each food would be more useful for these purposes as well.

**H. Resources**

Alwan, A., Dr. (Ed.). (2011). *Global status report on noncommunicable diseases 2010* (p. 33, Rep.). Italy: World Health Organization. Retrieved May 8, 2020, from https://apps.who.int/iris/bitstream/handle/10665/44579/9789240686458\_eng.pdf;jsessionid=B563EF98D7EC35CD759F3FC2408B9F85?sequence=1.

Choueiry, G. (2020). Understand Forward and Backward Stepwise Regression. Quantifying Health. https://quantifyinghealth.com/stepwise-selection/.

Croarkin, C., &amp; Tobias, P. (Eds.). (2013, October 10). 1.3.5.10. Levene Test for Equality of Variances. Engineering Statistics Handbook. https://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm.

Data Hub. (2018). In *Population Figures by Country*. John Snow Labs. Retrieved July 10, 2020, from https://datahub.io/JohnSnowLabs/population-figures-by-country#readme.

John Snow Labs. (2018). Data Hub. In Population Figures by Country (pp. 1–1). https://datahub.io/JohnSnowLabs/population-figures-by-country#readme.

Laerd Statistics. (2020). Pearson Product-Moment Correlation. Statistical tutorials and software guides. https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php.

Narkhede, S. (2018, June 26). Understanding AUC - ROC Curve. Medium. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

SAS Studio: Help Center. (2016, June 21). Retrieved May 08, 2020, from https://support.sas.com/software/products/sas-studio/faq/SAS\_whatis.htm.

SAS Institute. (2018). SAS University EditionVersion (3.8). S.l.

SAS Institute. (2020). *Predictive Modeling Using Logistic Regression (15.1)*. https://vle.sas.com/course/view.php?id=3472.

SAS Institute. (2020). *Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression*. https://vle.sas.com/course/view.php?id=2113.

U.S. Department of Agriculture. (2020, April). Download FoodData Central Data. Retrieved May 08, 2020, from https://fdc.nal.usda.gov/download-datasets.html.

World Bank Group. (2010). Global Consumption Database. Retrieved May 08, 2020, from http://datatopics.worldbank.org/consumption/product/Rice.

World Health Organization. (2018, April 5). NCD Deaths by Cause and Sex - Data by Country. Retrieved May 08, 2020, from https://apps.who.int/gho/data/view.main.NCDDEATHCAUSESNUMBERv?lang=en.

Yap, B. W., &amp; Sim, C. H. (2011). Comparisons of various types of normality tests. Journal of Statistical Computation and Simulation, 81(12), 2143–2155. https://doi.org/10.1080/00949655.2010.520163.