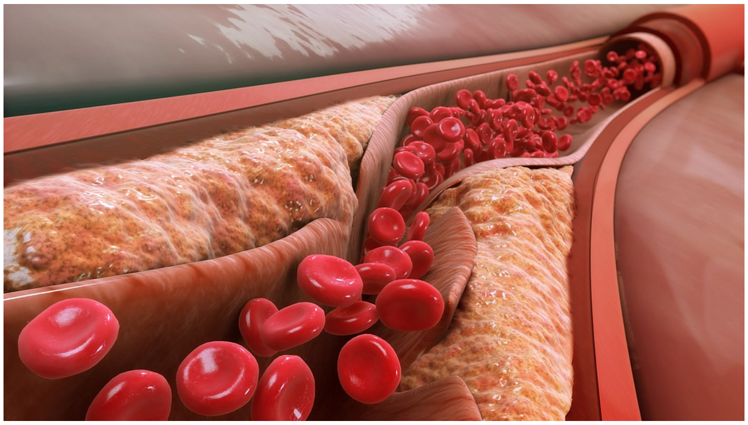
HEARTH DISEASE PREDICTION



Siphu Langeni, MS

February 27, 2020

TABLE OF CONTENTS

Introduction

According to the World Health Organization (WHO), cardiovascular disease is the number one killer worldwide.[1] It has maintained this position continuously for the last 15 years. More than half of the 17.9 million deaths that occur from it can be attributed to coronary artery disease (CAD).

CAD causes a huge burden to individuals, families and societies. The damage can range from decrease in quality of life to loss of life. These devastating effects are accompanied by a tremendous financial burden to healthcare systems. Globally, the cost of CAD was estimated to be US$863 billion in 2010 and projected to reach US$1.044 trillion by 2030[2]. Implementing strategies to prevent the disease before it begins may be instrumental in decreasing the health and economic costs of CAD.

The Study

The data is a subset of a larger dataset from a longitudinal study for coronary risk factors[3]. The dataset has 462 participants from three rural communities in the Western Cape of South Africa with known high risk for CAD. All participants are white, Afrikaaner males aged 15-64.

The Variables

The study has nine predictor variables and one target variable. The predictor variables are systolic blood pressure (mmHg), tobacco use (kg), low density lipoprotein cholesterol (mmol/L), adiposity (% bodyfat), family history of CAD, type A traits, obesity (kg/m^2), alcohol consumption and age at onset of symptoms. The target variable is whether or not a participant has CAD.

[1] https://www.who.int/health-topics/cardiovascular-diseases/

[2] http://www.championadvocates.org/en/champion-advocates-programme/the-costs-of-cvd

[3] Rossouw, J., Du Plessis, J., Benadé, A., Jordaan, P., Kotzé, J., Jooste, P. & Ferreira, J. (1983). Coronary Risk Factor Screening in Three Rural Communities. The CORIS Baseline Study. *S Afr Med J*, 64(12), 430-6.

The Study Framework

Of the nine predictor variables, six are modifiable while three are non-modifiable. I have chosen to study four of these modifiable risk factors, namely systolic blood pressure, tobacco use, low density lipoprotein cholesterol and obesity.

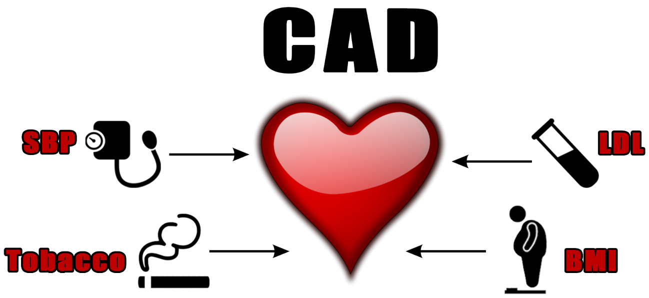


Figure 1

Hypotheses

Using this dataset, I hope to answer several questions about CAD:

1. Will patients with systolic hypertension experience more CAD than those who are normotensive?
2. Are patients with elevated LDL levels at higher risk for CAD?
3. Do patients who consume more tobacco products have higher incidence of CAD?
4. Do patients that are overweight or obese have more CAD than those with normal and low BMI?

Methodology

SAS 9.4 (TSM15) was used in reading, management, analysis and modelling of the data. A total of 12 PROC statements were used in the analysis of the data (see Appendix).

An infile statement was used to import the dataset for maximum control of the final output. Data types were established at import with two categorical variables and eight numerical variables. Custom formats were created for improved analysis and readability.

A total of 462 records were imported. There were no missing values in the dataset. Analysis will be on only four of the selected modifiable risk factors. Modeling will include all variables determined to contribute during backward feature elimination.

Descriptive Analysis

There were 302 cases where there was no CAD and 160 cases where there was CAD, 65.37% and 34.63%, respectively. Notably, this is much higher than the prevalence in the general population. This is due to the fact that inclusion criteria in the original study was symptoms of CAD in an already high risk subset of the population.

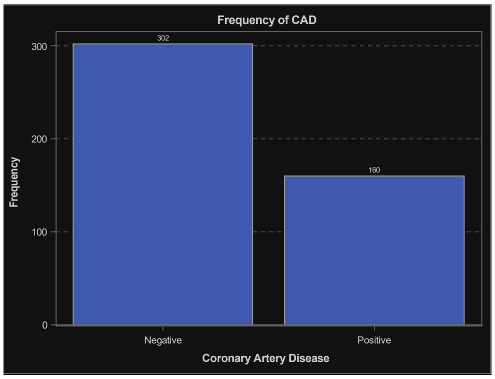


Figure 2

Univariate Analysis

Univariate analysis can give a better understanding of the distribution of each variable in the framework. We can observe that each of the variables have a near normal distribution. This will allow us to use parametric statistics later that require normality as a prerequisite. Tobacco use was given a cubic transformation in order to see a more normal like distribution in this particular variable.

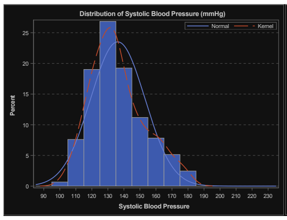


Figure 3.1

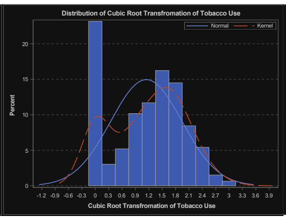


Figure 3.2

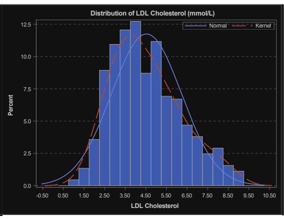


Figure 3.3

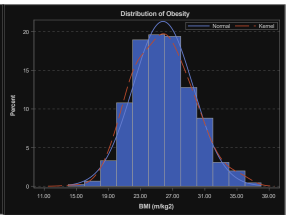


Figure 3.4

Outlier Detection

Outliers are datapoints that do not seem to fit in the normal distribution of the data. These can be problematic as they can cause a skewness in the data, pulling the mean away from the median in the direction of the outlier. PROC univariate is useful in compiling the five number summary: minimum, Q1, median, Q2, max. Interquartile range can also be reported in this procedures. The outliers have been defined by the following two equations:

$$\boldsymbol{outlier} > \ Q3 + 1.5IQR \text{ (the upper bound)}$$

$$\boldsymbol{outlier} < Q1 - 1.5IQR \text{ (the lower bound)}$$

The outliers were removed from the dataset to do the bivariate analysis.

Bivariate analysis

It is important to compare two variables to each other in the analysis. Boxplots can show the distribution of a variable based on a group that it is a part of. In this case, a comparison of each distribution is seen when CAD is present and when it is not.

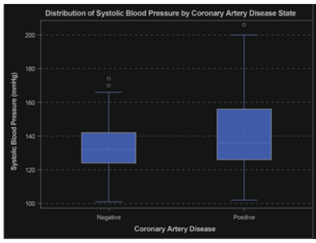


Figure 4.1

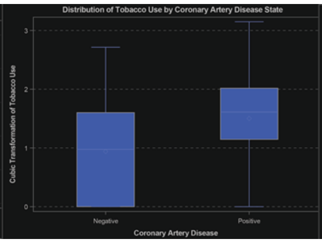


Figure 4.2

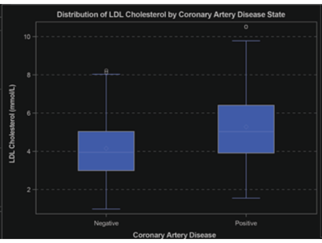


Figure 4.3

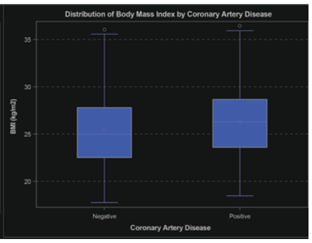


Figure 4.4

Inferential Analysis

Visually, it appears that there is a difference between the means of the SBP, Tobacco use, LDL and BMI based on presence of CAD or not. To be absolutely sure, two sample t-test must be performed to show if the means are from the same or different distributions.

Two-Sample T-test

Variable	Mean (with CAD) Mean (with no CAD)	Statistically Significant	p-value
Systolic Blood Pressure	143.7 mmHg 135.5 mmHg	Yes	< 0.0001
Tobacco Use	1.5022 kg^{-3} 0.9398 kg^{-3}	Yes	< 0.0001
LDL Cholesterol	5.2800 mmol/L 4.1505 mmol/L	Yes	< 0.0001
Body Mass Index	26.3017 kg/m^2 25.3975 kg/m^2	Yes	0.0088

Table 1 Two sample t-test between study variables in presence and absence of CAD

Chi-square Analysis

A chi-square analysis allows comparison between categorical variables and determine if the relationship is significant by rejection of the null hypothesis. Using the contingency table from PROC FREQ and the Chi-square statistic, we are able to determine a relationship even though we are not able to quantify the magnitude of that relationship. CAD seems to have a relationship with overweight to obese weight categories. CAD does not show any statistical relationship between systolic hypertension or normotension.

Scatterplot Matrix

A scatterplot is a good tool to show correlation between continuous variables. In the event of a continuous target variable, we may get a good sense of how much each variable will likely contribute to the model. This type of plot is less meaningful where the target is categorical, as is the case in this study. Before modelling, this is a good way to identify any variables which may have multicollinearity which can cause problems for explanatory > predictive models. A scatterplot matrix places multiple variables to see one by one correlation between each and every variable against all others.

For example, there is a strong positive linear relationship between Adiposity and Obesity. Both of these measure distribution of weight, making it logical that they would be highly correlated.

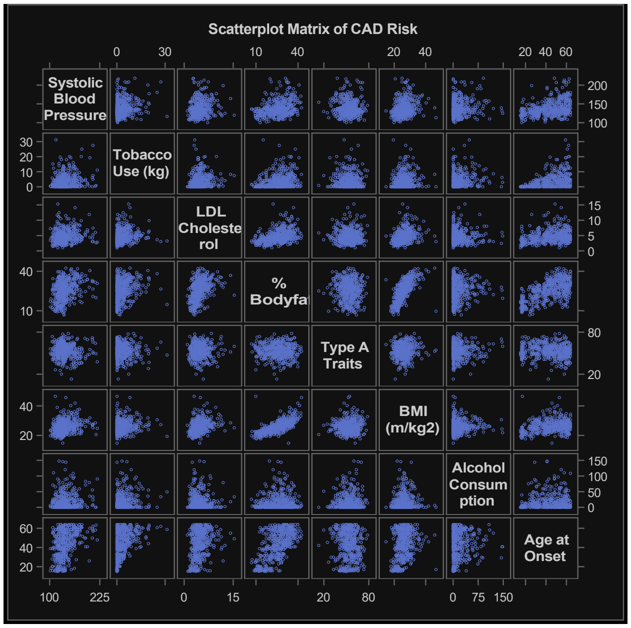


Figure 5

Logistic Regression

All the variables in the study are put into the logistic regression model, including those that were not a part of the study framework. The variables in the study framework focus on modifiable risk factors. There is still interplay between all of the risk factors so they have to be taken into consideration. By minimizing the negative loss likelihood, logistic regression aims to linearly separate the data between the two outcomes of CAD or not.

Backward feature elimination is a technique of dimensionality reduction that eliminates features that do not contribute meaningfully to the model. Sequentially, Alcohol, Adiposity, Systolic Blood Pressure and Obesity were removed. The model has high concordance at 79.4%. Odds ratios are calculated and provide a method of describing the strength of the relationship between a predictor and the target by comparing an exposure to an outcome. The following equation is a mathematical representation of the odds ratio of CAD given Tobacco Use, LDL Cholesterol, Family History, Type A and Age of Onset:

$$\boldsymbol{Odds\,ratio} \;=\; e \;\hat{}\; (-6.2288 \;+\; 0.5349(Tobacco) \;+\; 0.1508(LDL) \;-0.4403(FH\ Absent) + 0.0367(Type\ A) \;+\; 0.0496(Age))$$

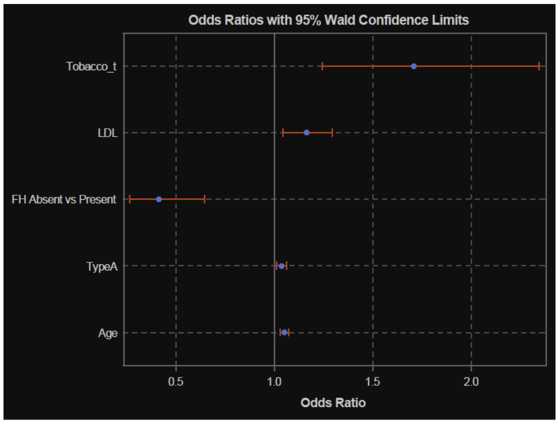


Figure 6

Summary of Results

1. Patients with systolic HTN **DO NOT** experience more CAD than those who are normotensive.
   Twenty-five percent of normotensive participants have CAD while 35.9% of hypertensive participants have CAD. Using chi-square analysis, we fail to reject the null hypothesis with a p-value of 0.0689.

2. Patients with elevated LDL levels **ARE** at higher risk for CAD.
   The odds ratio indicates that higher levels of LDL indicate a higher risk for CAD. At an odds ratio of 1.163 with a 95% confidence interval of 1.044 – 1.295 that does not span 1.0 (the null hypothesis). This shows high precision and statistical significance with a p-value of 0.0061.

3. Patients who consume more tobacco products **WILL** have higher incidence of CAD.
   Performing a two-sample t-test shows that study participants with CAD consume more tobacco than those that do not (p <0.0001)

4. Patients that are overweight or obese **WILL** have more CAD than those with normal and low BMI.
   While obesity was not significant in the logistic regression model, there is a relationship between obesity and CAD when we segment the weight groups as low-normal and overweight-obese. Chi-square analysis shows a statistically significant relationship (p = 0.0044).

Logistic regression model indicates the most contributory variables in the study are: **Tobacco Use, LDL,** FH**,** Type A Traits**,** Age of Onset. Only two of the four variables from the study framework have shown to be great predictors for CAD. These will be the focus of our recommendations.

Recommendations

It has been established that the outcomes of CAD can be devastating. We also know that there are some risk factors that are linked with modifiable behaviors and lifestyles. The best treatment for CAD should involve preventative measures.

My recommendation is to create a community program aimed at early identification of modifiable risk factors. Prevalence of CAD is much higher in low to middle income demographies[4]. These communities may have difficulty with financial, geographical or other forms of access. By pinpointing high risk areas, early identification can be established.

A comprehensive community education program should include:

- Informing the community about CAD and its effects.
- A nutrition program to promote a heart healthy diet.
- Behavior modification programs centered around not smoking or smoking cessation, whichever is relevant.
- And finally, prescription of medication for management of LDL, if already outside of normal parameters

[4] Schultz, W., Kelli, H., Lisco, J., Varghese, Shen, J., Sandesara, P., Quyyumi, A., Taylor, H., Gulati, M., Harold, J., Mieres, J., Ferdinand, K., Mensah, G. & Sperling, L. (2018). Socioeconomic Status and Cardiovascular Outcomes. *Circulation*, 137(20), 2166-2178.

APPENDIX

PROC Statements Used[5]

PROC FORMAT enables you to define your own informats and formats for variables

PROC PRINT prints the observations in a SAS data set

PROC CONTENTS shows the contents of a SAS data set and prints the directory of the SAS library

PROC MEANS provides data summarization tools to compute descriptive statistics for variables across all observations and within groups of observations

PROC FREQ produces one-way to n-way frequency and contingency (crosstabulation) tables

PROC SORT orders SAS data set observations by the values of one or more character or numeric variables

PROC UNIVARIATE provides a variety of descriptive measures, graphical displays, and statistical methods, which you can use to summarize, visualize, analyze, and model the statistical distributions of numeric variables

PROC SGPLOT creates one or more plots and overlays them on a single set of axes

PROC SGPANEL creates a panel of graph cells for the values of one or more classification variables

PROC SGSCATTER creates a paneled graph of scatter plots for multiple combinations of variables

[5] https://documentation.sas.com/

PROC TTEST performs *t* tests and computes confidence limits for one sample, paired observations, two independent samples, and the AB/BA crossover design

PROC LOGISTIC investigate the relationship between these discrete responses and a set of explanatory variables

SAS Scripts

```
%let path = D:\SASProject;
options validvarname = v7;

*Create a library;
libname SL "&path";

*Custom formats;
proc format;
        value $FHFMT
                '0' = 'Absent'
                '1' = 'Present'
                ;

        value $CADFMT
                '0' = 'Negative'
                '1' = 'Positive'
                ;
run;

*Import data, change cat data display, format, label;
data SL.HD;
        infile "&path\HeartDisease.csv" dlm = ',' firstobs = 2;
        input
                SBP
                Tobacco
                LDL
                Adiposity
                FH $
                TypeA
                Obesity
                EtOH
                Age
                CAD $
                ;

                if FH = '2' then FH = '0';

                if CAD = '1' then CAD = '0';
                else CAD = '1';

        format
                Tobacco 5.2
                LDL 5.2
                Adiposity 5.2
                Obesity 5.2
                EtOH 6.2
                ;

        label
                SBP = 'Systolic Blood Pressure (mmHg)'
                Tobacco = 'Tobacco Use (kg)'
                LDL = 'LDL Cholesterol (mmol/L)'
                Adiposity = '% Bodyfat'
                FH = 'Family History'
                TypeA = 'Type A Traits'
                Obesity = 'BMI (kg/m2)'
                EtOH = 'Alcohol Consumption'
                Age = 'Age at Onset'
                CAD = 'Coronary Artery Disease'
                ;
run;

*Global options;
options nodate nonumber;
```

```
ods graphics on;
ods noproctitle;

/*View contents of data*/
ods pdf file = "&path\Contents.pdf" style = MoonFlower notoc;
proc contents data = SL.HD order = varnum;
		title 'Contents of Heart Disease Prediction Datatset';
run;
title;
ods pdf close;

*Missing values;
*Numeric;
ods pdf file = "&path\Missing Cont.pdf" style = MoonFlower notoc;
proc means data = SL.HD nmiss;
		title 'Missing Continuous Values';
run;
title;
ods pdf close;

*Catgeorical;
ods pdf file = "&path\Missing Cat.pdf" style = MoonFlower notoc;
proc freq data = SL.HD;
		table FH / missing;
		format FH $FHFMT.;
		title 'Missing Categorical Values';
run;
title;
ods pdf close;

*How many have CAD or not;
proc freq data = SL.HD;
		table CAD / nocum;
		format CAD $CADFMT.;
		title 'Frequency of Coronary Artery Disease';
run;
title;

*Barchart for CAD status;
ods pdf file = "&path\CAD Frequency.pdf" style = MoonFlower notoc;
proc sgplot data = SL.HD;
		vbar CAD / datalabel;
		yaxis grid;
		title 'Frequency of Coronary Artery Disease';
		format CAD $CADFMT.;
run;
quit;
ods pdf close;
title;

*Sort the data by CAD class to use in a by statement;
proc sort data = SL.HD out = SL.HD_sort_CAD;
		by descending CAD;
run;

*Framework variables 5 num sum;
proc means data = SL.HD_sort_CAD min q1 median q3 max qrange;
		var SBP Tobacco LDL Obesity;
		by descending CAD;
		format CAD $CADFMT.;
		title 'Five Number Summary of the Framework Variables';
run;
title;

*Univariate Analysis - original distribution;
*SBP;
ods pdf file = "&path\UnivarSBP.pdf" style = MoonFlower notoc;
```

```
proc sgplot data = SL.NoLie;
        histogram SBP / showbins nbins = 15;
        density SBP;
        density SBP / type = kernel;
        yaxis grid;
        keylegend / location = inside
                                        position = topright;
        title 'Distribution of Systolic Blood Pressure (mmHg)';
run;
quit;
ods pdf close;
title;

*SBP;
ods pdf file = "&path\UnivarSBPBefore.pdf" style = MoonFlower notoc;
proc sgpanel data = SL.HD_sort_CAD;
        panelby CAD / columns = 1;
        histogram SBP;
        density SBP;
        density SBP / type = kernel;
        colaxis label = 'Systolic Blood Pressure in mmHg';
        format CAD $CADFMT.;
        title 'Distribution of Systolic Blood Pressure by Coronary Artery Disease State';
run;
quit;
ods pdf close;
title;

*Tobacco;
ods pdf file = "&path\UnivarTobaccoBefore.pdf" style = MoonFlower notoc;
proc sgpanel data = SL.HD_sort_CAD;
        panelby CAD / columns = 1;
        histogram Tobacco;
        density Tobacco;
        density Tobacco / type = kernel;
        colaxis label = 'Tobacco Use in kg';
        format CAD $CADFMT.;
        title 'Distribution of Tobacco Use by Coronary Artery Disease State';
run;
quit;
ods pdf close;
title;


data SL.HD_transform;
        set SL.HD_sort_CAD;
        Tobacco_t = Tobacco ** (1/3);
run;

*Tobacco transformed;
ods pdf file = "&path\UnivarTobacco.pdf" style = MoonFlower notoc;
proc sgplot data = SL.HD_transform;
        histogram Tobacco_t / showbins;
        density Tobacco_t;
        density Tobacco_t / type = kernel;
        yaxis grid;
        xaxis label = 'Cubic Root Transfromation of Tobacco Use';
        keylegend / location = inside
                                        position = topright;
        title 'Distribution of Cubic Root Transfromation of Tobacco Use';
run;
quit;
ods pdf close;
title;

*Tobacco transformed;
ods pdf file = "&path\UnivarTobaccoTransBefore.pdf" style = MoonFlower notoc;
```

```
proc sgpanel data = SL.HD_transform;
        panelby CAD / columns = 1;
        histogram Tobacco_t;
        density Tobacco_t;
        density Tobacco_t / type = kernel;
        colaxis label = "Cube Root Transformation of Tobacco Use in 1/kg(*ESC*){Unicode '00b3'x}";
        format CAD $CADFMT.;
        title 'Distribution of Cube Root Transformation of Tobacco Use by Coronary Artery Disease State';
run;
quit;
ods pdf close;
title;

*LDL;
ods pdf file = "&path\UnivarLDL.pdf" style = MoonFlower notoc;
proc sgplot data = SL.NoLie3;
        histogram LDL / showbins nbins = 20;
        density LDL;
        density LDL / type = kernel;
        yaxis grid;
        keylegend / location = inside
                                        position = topright;
        title 'Distribution of LDL Cholesterol (mmol/L)';
run;
quit;
ods pdf close;
title;

*LDL;
ods pdf file = "&path\UnivarLDLBefore.pdf" style = MoonFlower notoc;
proc sgpanel data = SL.HD;
        panelby CAD / columns = 1;
        histogram LDL;
        density LDL;
        density LDL / type = kernel;
        colaxis label = 'LDL Cholesterol in mmol/L';
        format CAD $CADFMT.;
        title 'Distribution of Low Density Lipoprotein Cholesterol by Coronary Artery Disease State';
run;
quit;
ods pdf close;
title;

*Obesity;
ods pdf file = "&path\UnivarBMI.pdf" style = MoonFlower notoc;
proc sgplot data = SL.NoLie4;
        histogram Obesity / showbins;
        density Obesity;
        density Obesity / type = kernel;
        yaxis grid;
        keylegend / location = inside
                                        position = topright;
        title 'Distribution of Obesity';
run;
quit;
ods pdf close;
title;

*Obesity;
ods pdf file = "&path\UnivarObesityBefore.pdf" style = MoonFlower notoc;
proc sgpanel data = SL.HD;
        panelby CAD / columns = 1;
        histogram Obesity;
        density Obesity;
        density Obesity / type = kernel;
        colaxis label = 'Distribution of Obesity';
        format CAD $CADFMT.;
```

```
            title 'Distribution of Obesity by Coronary Artery Disease State';
run;
quit;
ods pdf close;
title;

*Create CAD, noCAD version of the dataset;
data SL.CAD SL.noCAD;
            set SL.HD_transform;
            drop Adiposity FH TypeA EtOH Age;
            if CAD = '1' then output SL.CAD;
            else output SL.noCAD;
run;

*Create Q1 Q3 IQR for each variable with CAD;
*SBP;
proc univariate data = SL.CAD noprint;
            var SBP;
            output out = SL.FindLiersSBPCAD q1 = q1 q3 = q3 qrange = iqr;
run;

*Tobacco transformed;
proc univariate data = SL.CAD noprint;
            var Tobacco_t;
            output out = SL.FindLiersTobacco_tCAD q1 = q1 q3 = q3 qrange = iqr;
run;

*LDL;
proc univariate data = SL.CAD noprint;
            var LDL;
            output out = SL.FindLiersLDLCAD q1 = q1 q3 = q3 qrange = iqr;
run;

*Obesity;
proc univariate data = SL.CAD noprint;
            var Obesity;
            output out = SL.FindLiersObesityCAD q1 = q1 q3 = q3 qrange = iqr;
run;

*Remove outliers from SL.CAD;
*4 outliers removed from SBP with CAD;
data SL.NoLiersCAD (drop = q1 q3 iqr);
            if _n_ = 1 then set SL.FindLiersSBPCAD;
            set SL.CAD;
            if SBP < q1 - 1.5 * iqr or SBP > q3 + 1.5 * iqr then delete;
run;

*0 outliers removed from Tobacco with CAD;
data SL.NoLiersCAD (drop = q1 q3 iqr);
            if _n_ = 1 then set SL.FindLiersTobacco_tCAD;
            set SL.NoLiersCAD;
            if Tobacco_t < q1 - 1.5 * iqr or Tobacco_t > q3 + 1.5 * iqr then delete;
run;

*6 outliers removed from LDL with CAD;
data SL.NoLiersCAD (drop = q1 q3 iqr);
            if _n_ = 1 then set SL.FindLiersLDLCAD;
            set SL.NoLiersCAD;
            if LDL < q1 - 1.5 * iqr or LDL > q3 + 1.5 * iqr then delete;
run;

*5 outliers removed from Obesity with CAD;
data SL.NoLiersCAD (drop = q1 q3 iqr);
            if _n_ = 1 then set SL.FindLiersObesityCAD;
            set SL.NoLiersCAD;
            if Obesity < q1 - 1.5 * iqr or Obesity > q3 + 1.5 * iqr then delete;
run;
```

```
*Create Q1 Q3 IQR for each variable with noCAD;
*SBP;
proc univariate data = SL.noCAD noprint;
		var SBP;
		output out = SL.FindLiersSBPnoCAD q1 = q1 q3 = q3 qrange = iqr;
run;

*Tobacco;
proc univariate data = SL.noCAD noprint;
		var Tobacco_t;
		output out = SL.FindLiersTobacco_tnoCAD q1 = q1 q3 = q3 qrange = iqr;
run;

*LDL;
proc univariate data = SL.noCAD noprint;
		var LDL;
		output out = SL.FindLiersLDLnoCAD q1 = q1 q3 = q3 qrange = iqr;
run;

*Obesity;
proc univariate data = SL.noCAD noprint;
		var Obesity;
		output out = SL.FindLiersObesitynoCAD q1 = q1 q3 = q3 qrange = iqr;
run;

*Remove outliers from SL.noCAD;
*10 outliers removed from SBP with no CAD;
data SL.NoLiersnoCAD (drop = q1 q3 iqr);
		if _n_ = 1 then set SL.FindLiersSBPnoCAD;
		set SL.noCAD;
		if SBP < q1 - 1.5 * iqr or SBP > q3 + 1.5 * iqr then delete;
run;

*0 outliers removed from Tobacco with no CAD;
data SL.NoLiersnoCAD (drop = q1 q3 iqr);
		if _n_ = 1 then set SL.FindLiersTobacco_tnoCAD;
		set SL.NoLiersnoCAD;
		if Tobacco_t < q1 - 1.5 * iqr or Tobacco_t > q3 + 1.5 * iqr then delete;
run;

*7 outliers removed from LDL with no CAD;
data SL.NoLiersnoCAD (drop = q1 q3 iqr);
		if _n_ = 1 then set SL.FindLiersLDLnoCAD;
		set SL.NoLiersnoCAD;
		if LDL < q1 - 1.5 * iqr or LDL > q3 + 1.5 * iqr then delete;
run;

*3 outliers removed from Obesity with no CAD;
data SL.NoLiersnoCAD (drop = q1 q3 iqr);
		if _n_ = 1 then set SL.FindLiersObesitynoCAD;
		set SL.NoLiersnoCAD;
		if Obesity < q1 - 1.5 * iqr or Obesity > q3 + 1.5 * iqr then delete;
run;

*Join the oulier-free datasets;
data SL.NoLiers;
		set SL.NoLiersCAD SL.NoLiersnoCAD;
run;

*Bivariate analysis;
*SBP boxplot by CAD;
ods pdf file = "&path\BoxSBPAfter.pdf" style = MoonFlower notoc;
proc sgplot data = SL.NoLiers;
		vbox SBP / category = CAD;
		yaxis grid;
		format CAD $CADFMT.;
```

```
          title 'Distribution of Systolic Blood Pressure by Coronary Artery Disease State';
          title2 '(After removal of outliers)';
run;
quit;
title;
title2;
ods pdf close;

*Tobacco transposed boxplot by CAD;
ods pdf file = "&path\BoxTobacco_tAfter.pdf" style = MoonFlower notoc;
proc sgplot data = SL.NoLiers;
          vbox Tobacco_t / category = CAD;
          yaxis grid;
          format CAD $CADFMT.;
          label Tobacco_t = 'Cubic Transformation of Tobacco Use';
          title 'Distribution of Cubic Transformation of Tobacco Use by Coronary Artery Disease State';
          title2 '(After removal of outliers)';
run;
quit;
title;
title2;
ods pdf close;

*LDL boxplot by CAD;
ods pdf file = "&path\BoxLDLAfter.pdf" style = MoonFlower notoc;
proc sgplot data = SL.NoLiers;
          vbox LDL / category = CAD;
          yaxis grid;
          format CAD $CADFMT.;
          title 'Distribution of Low Density Lipoprotein Cholesterol by Coronary Artery Disease State';
          title2 '(After removal of outliers)';
run;
quit;
title;
title2;
ods pdf close;

*Obesity boxplot by CAD;
ods pdf file = "&path\BoxObesityAfter.pdf" style = MoonFlower notoc;
proc sgplot data = SL.NoLiers;
          vbox Obesity / category = CAD;
          yaxis grid;
          format CAD $CADFMT.;
          title 'Distribution of Body Mass Index by Coronary Artery Disease State';
          title2 '(After removal of outliers)';
run;
quit;
title;
title2;
ods pdf close;

*
F-test for pooled variance
Must be determined before using two sample t-test
H0: pooled variance
Ha: non-pooled variance
;

*
Two-way t-test
H0: mean SBP in CAD(0) = mean SBP in CAD(1)
Ha: mean SBP in CAD(0) > mean SBP in CAD(1)
;
*T-test SBP;
ods pdf file = "&path\TTestSBP.pdf" style = MoonFlower notoc;
proc ttest data = SL.HD H0 = 0 sides = l plots = none;
          class CAD;
```

```
        var SBP;
        format CAD $CADFMT.;
        title 'Two-Sample T-Test';
run;
quit;
title;
ods pdf close;

*T-test Tobacco_t;
ods pdf file = "&path\TTestTobacco_t.pdf" style = MoonFlower notoc;
proc ttest data = SL.NoLiers H0 = 0 sides = l plots = none;
        class CAD;
        var Tobacco_t;
        format CAD $CADFMT.;
        label Tobacco_t = 'Cubic Transformation of Tobacco Use';
        title 'Two-Sample T-Test';
run;
quit;
title;
ods pdf close;

*T-test LDL;
ods pdf file = "&path\TTestLDL.pdf" style = MoonFlower notoc;
proc ttest data = SL.NoLiers H0 = 0 sides = l plots = none;
        class CAD;
        var LDL;
        format CAD $CADFMT.;
        title 'Two-Sample T-Test';
run;
quit;
title;
ods pdf close;

*T-test Obesity;
ods pdf file = "&path\TTestObesity.pdf" style = MoonFlower notoc;
proc ttest data = SL.NoLiers H0 = 0 sides = l plots = none;
        class CAD;
        var Obesity;
        format CAD $CADFMT.;
        title 'Two-Sample T-Test';
run;
quit;
title;
ods pdf close;

*Custom format normo- vs hypertensive;
proc format;
        value HTN
                low - 120 = 'Normotensive'
                120 - high = 'Hypertensive'
                ;
run;

*Chi square to assess relationship between Hypertension and CAD;
ods pdf file = "&path\ChiHTN.pdf" style = MoonFlower notoc;
proc freq data = SL.NoLiers;
        tables SBP * CAD / expected chisq;
        format SBP HTN. CAD $CADFMT.;
        title 'Contingency Table for Systolic Hypertension vs. Coronary Artery Disease State';
run;
ods pdf close;
title;


*Custom format BMI levels;
proc format;
        value BMI
```

```
                low - <25 = 'Under - Normal'
                25 - high = 'Overweight - Obese'
                ;
run;

*Chi square to assess relationship between BMI and CAD;
ods pdf file = "&path\ChiObesity.pdf" style = MoonFlower notoc;
proc freq data = SL.NoLiers;
        tables Obesity * CAD / expected chisq;
        format Obesity BMI. CAD $CADFMT.;
        title 'Contingency Table for Body Mass Index vs. Coronary Artery Disease State';
run;
ods pdf close;
title;

data SL.HD_transform;
        set SL.HD;
        Tobacco_t = Tobacco ** (1/3);
run;

*Bi-variate analysis;
ods pdf file = "&path\BivarScatterMatrix.pdf" style = MoonFlower notoc;
proc sgscatter data = SL.HD_transform;
        matrix SBP Tobacco_t LDL Adiposity TypeA Obesity EtOH Age;
        label Tobacco_t = 'Tobacco (transformed)';
        title 'Scatterplot Matrix of Coronary Artery Disease Risk Factors';
run;
quit;
title;
ods pdf close;

*Logisic Regression model for prediction, odds ratio;
ods pdf file = "&path\LogReg.pdf" style = MoonFlower notoc;
proc logistic data = SL.HD_transform desc plots(only) = oddsratio plots(only) = roc;
        model CAD = SBP Tobacco_t LDL Adiposity FH TypeA Obesity EtOH Age / expb selection = backward;
        format FH $FHFMT.;
        output out = outdata p = pred_prob lower = low upper = up;
        title 'Logistic Regression for Coronary Artery Disease Prediction';
run;
quit;
title;
ods pdf close;

*Odds ratio
P /(1 - P) = e ^ (-6.2288+ 0.5349(Tobacco_t) + 0.1508(LDL) - 0.4403(FH Absent) + 0.0367(TypeA) + 0.0496(Age))
```