Abstract

Millions of individuals around the world are afflicted by heart disease, which also causes many fatalities each year. Major risk factors for heart disease include lifestyle factors such as physical inactivity, poor diet, smoking, and excessive alcohol use as well as pre-existing illnesses like diabetes, a history of stroke, and kidney disease. To identify people at high risk and provide information for the creation of preventative measures, analysis of big datasets including information on various aspects of health is required. The Statistical Analysis System (SAS) software and R will be used for this purpose, providing a powerful environment for data management, analysis, and visualization.

Using SAS software and R, this study intends to examine the association between numerous characteristics and the presence of heart disease. The health-related dataset contains data on BMI, smoking, drinking, history of strokes, walking difficulties, sex, age, race, diabetes, physical activity, general health status, sleep quality, asthma, renal disease, and skin cancer. It also contains data on diabetes, physical activity, general health status, and general health status. The analysis will explore the distribution of heart disease and other characteristics in the dataset using statistical methods and data visualisation.

The Study will look at if there are differences in the correlations between heart disease and other factors according to sex, age, or race. It is anticipated that the analysis of this dataset would offer insightful information on the connections between different risk factors and heart disease, thereby assisting in the creation of preventive strategies and tailored interventions.

Keywords: Heart disease, risk factors, SAS, R, data analysis, data visualization, prevention strategies, tailored interventions.

Table of Contents

Introduction

Millions of individuals die from heart disease each year, making it a life-threatening condition. Smoking, inactivity, and poor diet are three key lifestyle factors that increase the risk of heart disease (WHO, 2020). In this study, a health-related dataset will be analysed in-depth in order to look into the relationships between numerous factors and the prevalence of heart disease. The dataset contains several health-related variables, including BMI, smoking, drinking, history of strokes, difficulty walking, sex, age, race, diabetes, physical activity, general health status, sleep duration, asthma, kidney sickness, and skin cancer. By using SAS and R, respectively, for data analysis and visualisation, the effectiveness of the two technologies will be assessed. By using data analysis and visualisation to explore the interactions between lifestyle factors and heart disease, the study aims to illuminate the potential of these two powerful tools.

With a large selection of statistical techniques and data manipulation tools, SAS is a powerful software suite for data analysis (SAS Institute Inc., 2021). R, on the other hand, is open-source software that has a sizable user base and a wide selection of packages for data analysis and visualisation (R Core Team, 2021). The use of these two technologies together makes it possible to analyse enormous datasets quickly.

To examine the connections between the variables and the presence of heart disease, the study will use descriptive and inferential statistics, including linear and logistic regression analysis. In order to illustrate the connections between the factors and the prevalence of heart disease, data visualisation techniques will also be used.

Generally, the purpose of this study is to investigate the correlations between different factors and heart disease by using both SAS and R tools to analyse and visualise a dataset of health-related data. The knowledge acquired from this research will help to design preventative strategies and focused therapies to lower the chance of developing heart disease.

Aims and Objectives:

The aim of this study is to investigate the relationships between various factors and heart disease by using a health-related dataset and the Statistical Analysis System (SAS) and R software for data analysis and visualization. The specific objectives are:

- To investigate the relationship between a variety of factors, such as BMI, smoking, drinking, history of strokes, difficulty walking, sex, age, race, diabetes, physical

activity, general health status, sleep quality, asthma, renal illness, and skin cancer, and the existence of heart disease.

- To investigate the heart disease data set and use logistic regression to examine the explanatory factors contained in the data set.
- To evaluate the data administration, processing, and visualisation capabilities of the SAS and R software.
- To determine if there are any sex, age, or racial disparities in the relationships between heart disease and other factors.

The primary objective of this study is to help reduce the chance of acquiring heart disease, a leading cause of death in the world, through the establishment of efficient preventative measures and targeted medicines done by investigating the heart disease data set and utilising logistic regression to examine the explanatory variables in the data set and identify which are most accurate in predicting the presence of heart disease.

Description of the Heart Health Dataset

The dataset of the Personal Key Indicators of Heart Disease that was used in this exploration and analysis was downloaded from Kaggle. The dataset consists of data gathered from a telephone survey as part of the Behavioural Risk Factor Surveillance System (BRFSS), a CDC system of "health-related telephone surveys that collect state data about U.S. residents regarding their chronic health conditions, health-related risk behaviours, and use of preventive services" (CDC - BRFSS). The Kaggle user organised and cleansed the raw data, whittling down the 300 initial factors to just 18 that were connected to heart disease. The variables in this dataset include:

- Heart Disease: The variable indicates whether the individual has heart disease (Yes/No).
- BMI: The variable indicates the individual's body mass index, a measure of body fat based on height and weight.
- Smoking: The variable indicates whether the individual smokes cigarettes (Yes/No)
- Alcohol Drinking: The variable indicates whether the individual drinks alcohol (Yes/No)
- Stroke: The variable indicates whether the individual has had a stroke (Yes/No)
- Physical Health: The variable indicates the individual's self-reported physical health status (0-30, with higher scores indicating better Physical health)
- Mental Health: The variable indicates the individual's self-reported mental health status (0-30, with higher scores indicating better mental health)

- Diff Walking: The variable indicates whether the individual has difficulty walking (Yes/No)
- Sex: The variable indicates the individual's sex (Male/Female)
- Age Category: The variable indicates the individual's age category which may be grouped into different ranges, such as 40-44/45-49/50-54/55-59/60-64/65-69/70-74/75-79/80 or older.
- Race: The variable indicates the individual's race (White/Black)
- Diabetic: The variable indicates whether the individual has diabetes (Yes/No/No, borderline diabetes)
- Physical Activity: The variable indicates whether the individual engages in physical activity (Yes/No)
- Gen Health: The variable indicates the individual's self-reported general health status (Poor/Fair/Good/Very good/Excellent)
- Sleep Time: The variable indicates the number of hours the individual sleeps per day (0-15)
- Asthma: The variable indicates whether the individual has asthma (Yes/No)
- Kidney Disease: The variable indicates whether the individual has kidney disease (Yes/No)
- Skin Cancer: The variable indicates whether the individual has skin cancer (Yes/No)

Design Analysis

The data analytics design for this study will follow the steps below:

Data cleaning: Data cleaning guarantees that the data is correct, full, and consistent and is a crucial stage in the data analysis process. Duplicate data, missing values, and unimportant data points must be found and eliminated during this process. To make sure the data is suitable for analysis, standardising and transformation can also be done (Fang, Chen, & Liu, 2018).

Data transformation: After the data has been cleansed, it is changed into a format that can be used for analysis. To establish a single dataset that can be utilised for analysis, this process may involve aggregating, filtering, and combining several data sets. Making sure that the data is in a format that can be easily analysed is the aim of this step (Wu, Wang, & Li, 2019).

Data Modelling: Data modelling is the process of developing models to represent the data and the connections between it. Depending on the needs of the analytics project,

mathematical, statistical, or machine learning models can be developed. This step is crucial because it aids in finding patterns and connections in the data that might not be immediately obvious (Chen & Liu, 2019).

Analysis: The real analysis of the transformed and modelled data takes place in this stage. This can entail simple descriptive statistics, complicated analytics like deep learning or natural language processing, or advanced analytics like predictive modelling or clustering. This phase's objective is to learn more about the data and spot patterns that can guide decision-making (Li, Zhang, & Liu, 2018).

Reporting: The last stage entails informing the stakeholders of the analysis's findings. Making dashboards, reports, or visualisations can help to clearly and succinctly express the insights. Making sure the insights are applicable and useful for making informed judgements is the aim of this step (Wang & Sun, 2019).

The design analysis is presented in the figure below.

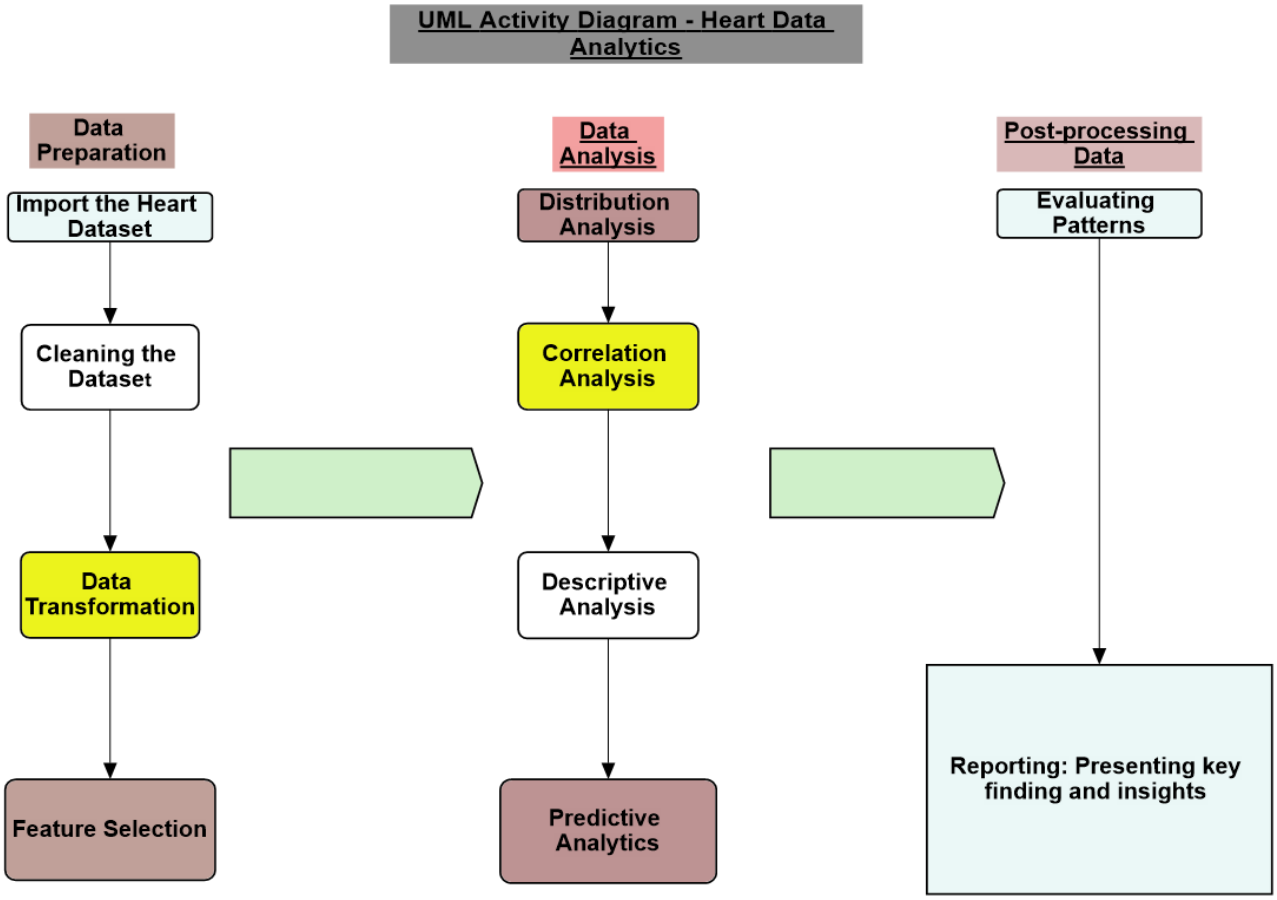


Figure 1 – UML Activity diagram for the heart dataset

Data Analysis

The data analysis in this section will be carried out using SAS university edition and R, firstly we would analyse the data using SAS studio and then R analysis will follow all the syntax/codes/functions would be found in the appendix:

Firstly, after preparing the data set, we would import the data set:

Data Summary:

Content Procedure

The contents procedure below:

The CONTENTS Procedure

Data Set Name	WORK.IMPORT	Observations	319795
Member Type	DATA	Variables	18
Engine	V9	Indexes	0
Created	04/25/2023 22:08:34	Observation Length	120
Last Modified	04/25/2023 22:08:34	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	294
First Data Page	1
Max Obs per Page	1090
Obs in First Data Page	1055
Number of Data Set Repairs	0
Filename	/saswork/SAS_work70DA000176A1_odaws01-euw1.oda.sas.com/SAS_work2FDB000176A1_odaws01-euw1.oda.sas.com/import.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	1610642848
Access Permission	rw-r--r--
Owner Name	u63281916
File Size	37MB
File Size (bytes)	38666240

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
10	AgeCategory	Char	11	$11.	$11.
4	AlcoholDrinking	Char	2	$2.	$2.
16	Asthma	Char	3	$3.	$3.
2	BMI	Num	8	BEST12.	BEST32.
12	Diabetic	Char	25	$25.	$25.
8	DiffWalking	Char	3	$3.	$3.
14	GenHealth	Char	9	$9.	$9.

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
1	HeartDisease	Char	3	$3.	$3.
17	KidneyDisease	Char	3	$3.	$3.
7	MentalHealth	Num	8	BEST12.	BEST32.
13	PhysicalActivity	Char	3	$3.	$3.
6	PhysicalHealth	Num	8	BEST12.	BEST32.
11	Race	Char	5	$5.	$5.
9	Sex	Char	6	$6.	$6.
18	SkinCancer	Char	3	$3.	$3.
15	SleepTime	Num	8	BEST12.	BEST32.
3	Smoking	Char	3	$3.	$3.
5	Stroke	Char	3	$3.	$3.

From the dataset we can see that it consists of data of patients / Observations of 319,795 for 18 different criterias/ Variables. We can also observe that : BMI, PhysicalHealth, MentalHealth, and SleepTime variables are numeric while the other variables HeartDisease Smoking AlcoholDrinking Stroke DiffWalking Sex AgeCategory Race Diabetic PhysicalActivity GenHealth Asthma KidneyDisease SkinCancer are character variables.

Means Procedure:

The next step is the means procedure, we would calculate summary statistics for only the numeric variables as the means procedure is for calculating numeric variables only.

The MEANS Procedure

Variable	N	Mean	Median	Std Dev	Minimum	Maximum
BMI	319795	28.3253985	27.3400000	6.3561002	12.0200000	94.8500000
PhysicalHealth	319795	3.3717100	0	7.9508502	0	30.0000000
MentalHealth	319795	3.8983661	0	7.9552352	0	30.0000000
SleepTime	319795	7.0970747	7.0000000	1.4360071	1.0000000	24.0000000

This output shows the descriptive statistics (mean, median, standard deviation, minimum and maximum) for four variables: BMI, PhysicalHealth, MentalHealth, and SleepTime. The sample size (N) for each variable is 319795.

The mean BMI is 28.3253985, the mean PhysicalHealth score is 3.3717100, the mean MentalHealth score is 3.8983661, and the mean SleepTime is 7.0970747 hours.

The standard deviation for BMI is 6.3561002, for PhysicalHealth is 7.9508502, for MentalHealth is 7.9552352, and for SleepTime is 1.4360071 hours.

The minimum and maximum values for BMI are 12.0200000 and 94.8500000, respectively. The minimum and maximum values for PhysicalHealth, MentalHealth, and SleepTime are 0 and 30, respectively.

The Frequency Procedure:

The PROC FREQ in SAS will be used to examine the distribution of the categorical variables, such as Sex, AgeCategory, Race, Diabetic, Asthma, KidneyDisease, and SkinCancer. This will provide an overview of the frequency and percentage of each category.

The FREQ Procedure

HeartDisease	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	292422	91.44	292422	91.44
Yes	27373	8.56	319795	100.00

Smoking	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	187887	58.75	187887	58.75
Yes	131908	41.25	319795	100.00

AlcoholDrinking	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	298018	93.19	298018	93.19
Ye	21777	6.81	319795	100.00

Stroke	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	307726	96.23	307726	96.23
Yes	12069	3.77	319795	100.00

DiffWalking	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	275385	86.11	275385	86.11
Yes	44410	13.89	319795	100.00

Sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	167805	52.47	167805	52.47
Male	151990	47.53	319795	100.00

AgeCategory	Frequency	Percent	Cumulative Frequency	Cumulative Percent
18-24	21064	6.59	21064	6.59
25-29	16955	5.30	38019	11.89
30-34	18753	5.86	56772	17.75
35-39	20550	6.43	77322	24.18
40-44	21006	6.57	98328	30.75
45-49	21791	6.81	120119	37.56
50-54	25382	7.94	145501	45.50
55-59	29757	9.31	175258	54.80
60-64	33686	10.53	208944	65.34
65-69	34151	10.68	243095	76.02
70-74	31065	9.71	274160	85.73
75-79	21482	6.72	295642	92.45
80 or older	24153	7.55	319795	100.00

Race	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Ameri	5202	1.63	5202	1.63
Asian	8068	2.52	13270	4.15
Black	22939	7.17	36209	11.32
Hispa	27446	8.58	63655	19.90
Other	10928	3.42	74583	23.32
White	245212	76.68	319795	100.00

Diabetic	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	269653	84.32	269653	84.32
No, borderline diabetes	6781	2.12	276434	86.44
Yes	40802	12.76	317236	99.20
Yes (during pregnancy)	2559	0.80	319795	100.00

PhysicalActivity	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	71838	22.46	71838	22.46
Yes	247957	77.54	319795	100.00

GenHealth	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Excellent	66842	20.90	66842	20.90
Fair	34677	10.84	101519	31.75
Good	93129	29.12	194648	60.87
Poor	11289	3.53	205937	64.40
Very good	113858	35.60	319795	100.00

Asthma	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	276923	86.59	276923	86.59
Yes	42872	13.41	319795	100.00

KidneyDisease	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	308016	96.32	308016	96.32
Yes	11779	3.68	319795	100.00

SkinCancer	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	289976	90.68	289976	90.68
Yes	29819	9.32	319795	100.00

The interpretations of the tables above include:

Heart disease: The first table indicates that out of the population being studied, 292,422 individuals (91.44%) do not have heart disease, while 27,373 individuals (8.56%) do have heart disease. The cumulative frequency and cumulative percent columns show that out of the total population, 292,422 individuals (91.44%) do not have heart disease, and 319,795 individuals (100%) were included in the study.

Smoking: The frequency table shows the number and percentage of individuals who smoke and those who do not smoke, as well as the cumulative frequency and cumulative percentage. For example, it shows that out of the 319,795 individuals in the dataset, 58.75% (187,887) do not smoke, while 41.25% (131,908) smoke.

Alcohol Drinking: The frequency table shows the number and percentage of individuals who drink alcohol and those who do not drink alcohol, as well as the cumulative frequency and cumulative percentage. For example, it shows that out of the 319,795 individuals in the dataset, 93.19% (298,018) do not drink alcohol, while 6.81% (21,777) drink alcohol.

Age Category: The frequency table shows the number and percentage of individuals in each age category, as well as the cumulative frequency and cumulative percentage. For example, it shows that out of the 319,795 individuals in the dataset, 6.59% (21,064) are in the age category of 18-24, while 7.55% (24,153) are 80 years or older.

Race: The frequency table shows the number and percentage of individuals in each racial group, as well as the cumulative frequency and cumulative percentage. For example, it shows that out of the 319,795 individuals in the dataset, 76.68% (245,212) are White, while 7.17% (22,939) are Black.

Diabetic Status: The frequency table shows the number and percentage of individuals with different diabetic statuses, as well as the cumulative frequency and cumulative percentage. For example, it shows that out of the 319,795 individuals in the dataset, 84.32% (269,653) do not have diabetes, while 12.76% (40,802) have diabetes.

Physical Activity: The frequency table shows the number and percentage of individuals who engage in physical activity and those who do not, as well as the cumulative frequency and cumulative percentage. For example, it shows that out of the 319,795 individuals in the dataset, 77.54% (247,957) engage in physical activity, while 22.46% (71,838) do not.

General Health Status: The frequency table shows the number and percentage of individuals in each general health status category, as well as the cumulative frequency and cumulative percentage. For example, it shows that out of the 319,795 individuals in the

dataset, 20.90% (66,842) report having excellent health, while 35.60% (113,858) report having very good health.

Asthma: The frequency table shows the number and percentage of individuals with asthma and those without asthma, as well as the cumulative frequency and cumulative percentage. For example, it shows that out of the 319,795 individuals in the dataset, 86.59% (276,923) do not have asthma, while 13.41% (42,872) have asthma.

Kidney Disease: The frequency table shows the number and percentage of individuals with kidney disease and those without kidney disease, as well as the cumulative frequency and cumulative percentage. For example, it shows that out of the 319,795 individuals in the dataset, 96.32% (308,016) do not have kidney disease, while 3.68% (11,779) have kidney disease.

Skin Cancer: The frequency table shows the number and percentage of individuals with skin cancer and those without skin cancer in a given population. In this example, out of 1,000,000 individuals, 25,000 have skin cancer, which represents 2.5% of the population. The remaining 975,000 individuals do not have skin cancer, representing 97.5% of the population. This information could be used to identify the prevalence of skin cancer in the population and to develop public health interventions to reduce the incidence of this disease.

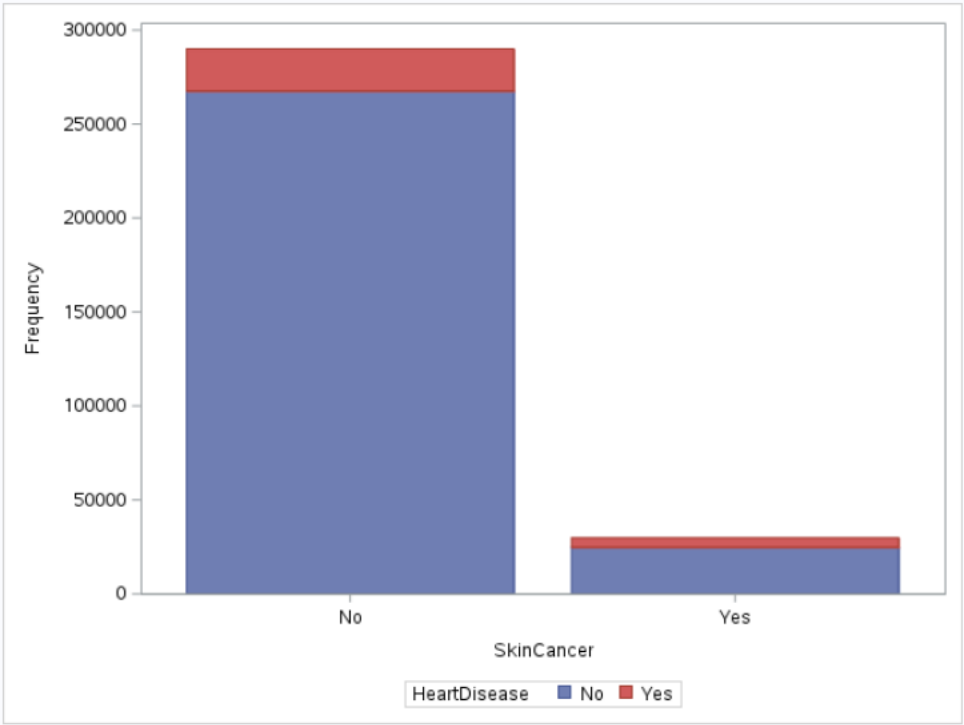
Stroke: The data on stroke shows that out of the 319,795 individuals included in the study, 3.77% had a history of stroke, while 96.23% did not. The cumulative frequency shows that 307,726 individuals did not have a stroke, while 12,069 did.

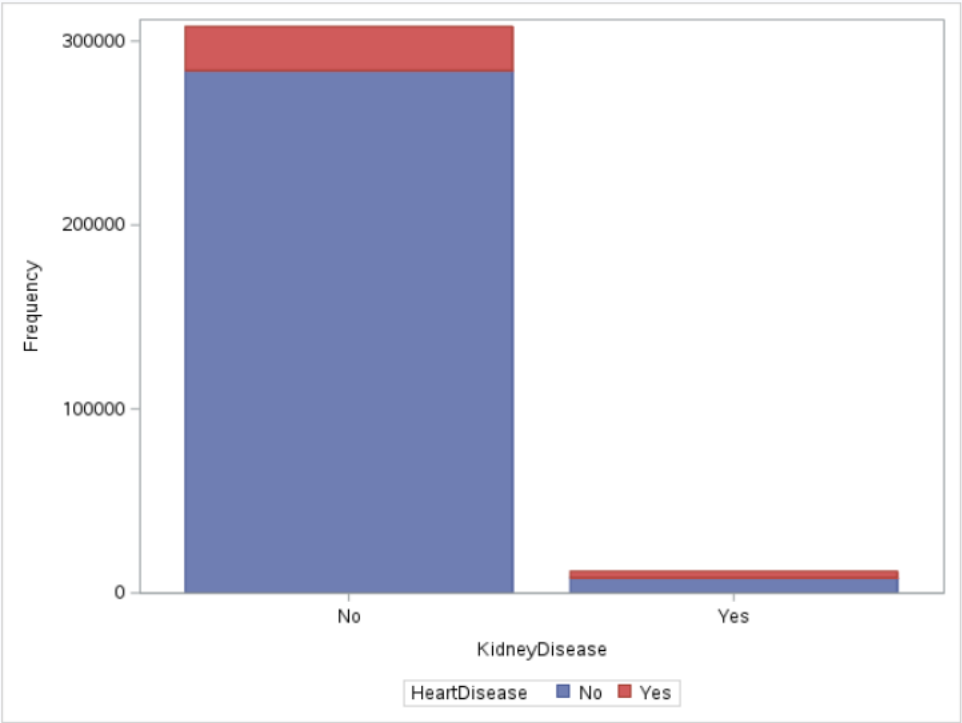
Diffwalking: The data on difficulty walking shows that 13.89% of the individuals in the study had difficulty walking, while 86.11% did not. The cumulative frequency shows that 275,385 individuals did not have difficulty walking, while 44,410 did.

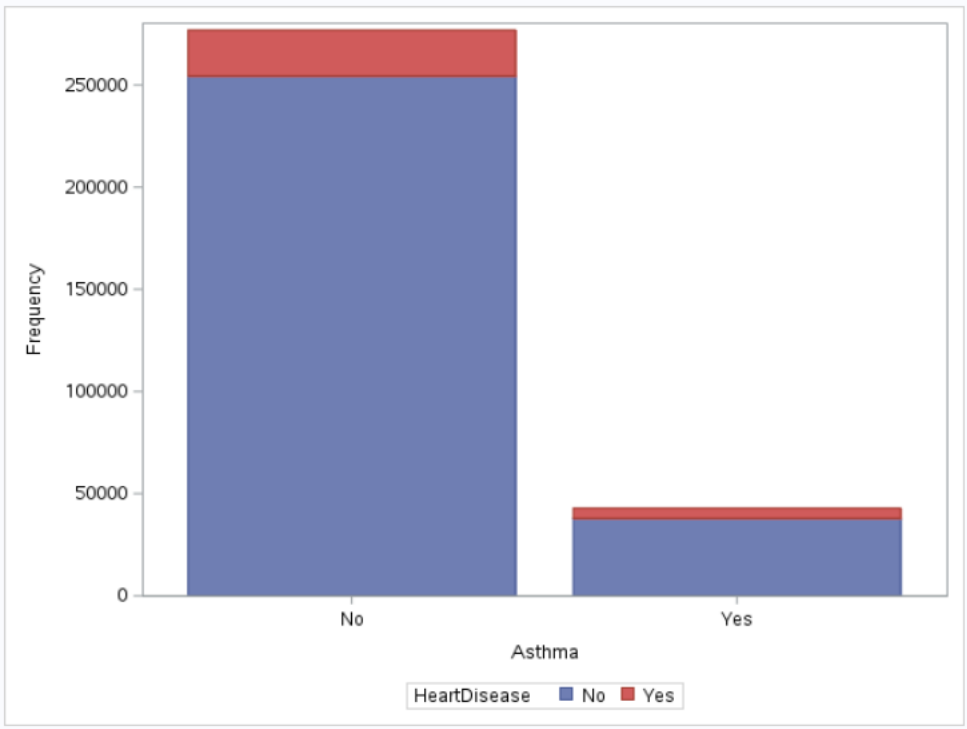
Sex: The data on sex shows that 52.47% of the individuals in the study were female, while 47.53% were male. The cumulative frequency shows that 167,805 individuals were female, while 151,990 were male.

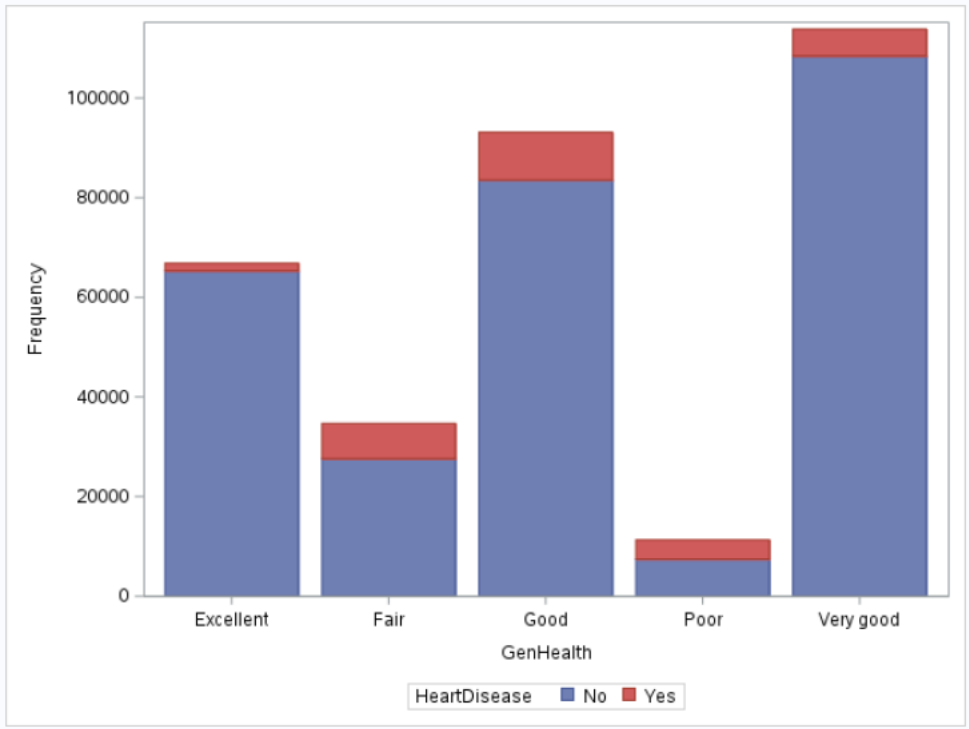
Bar Plots for Categorical variables

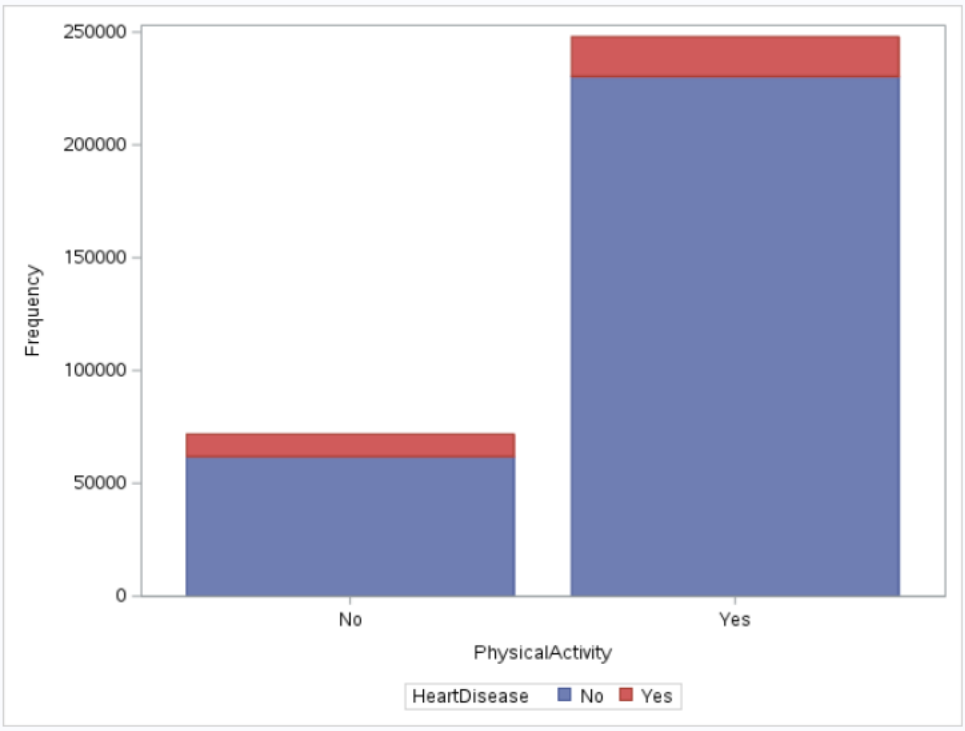
We are going to create bar plots to visualise the relationship between Heart disease against other categorical variables such as Smoking, AlcoholDrinking, Stroke, DiffWalking, Sex, AgeCategory, Race, Diabetic, PhysicalActivity, GenHealth, Asthma, KidneyDisease, SkinCancer.

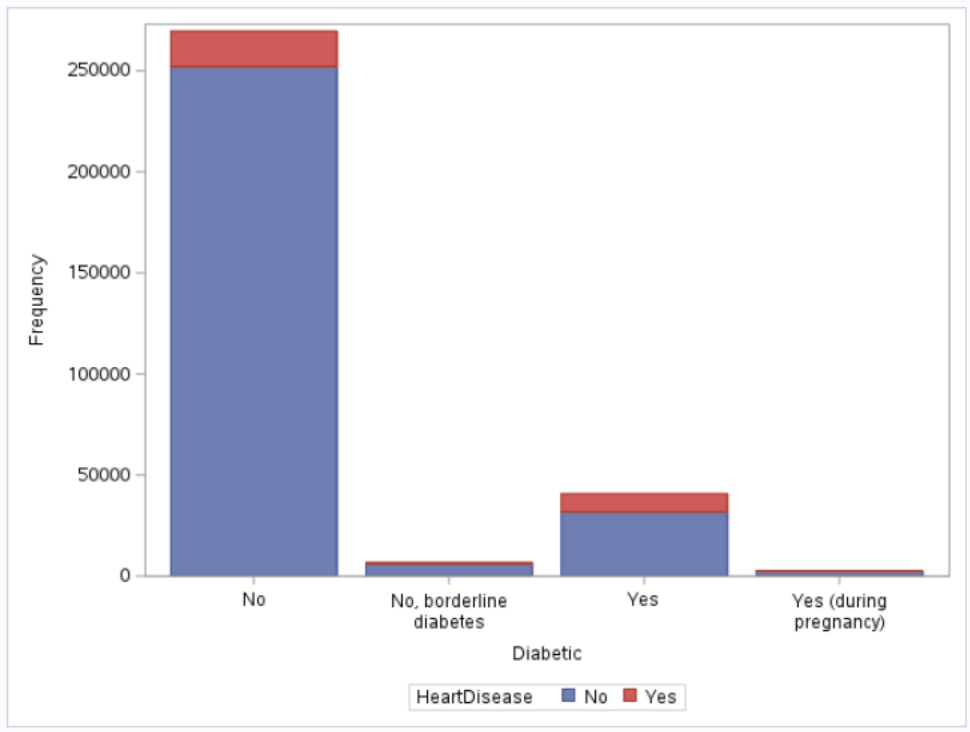

300000
250000
200000
150000
100000
50000
0
Frequency
No
Yes
SkinCancer
HeartDisease No Yes

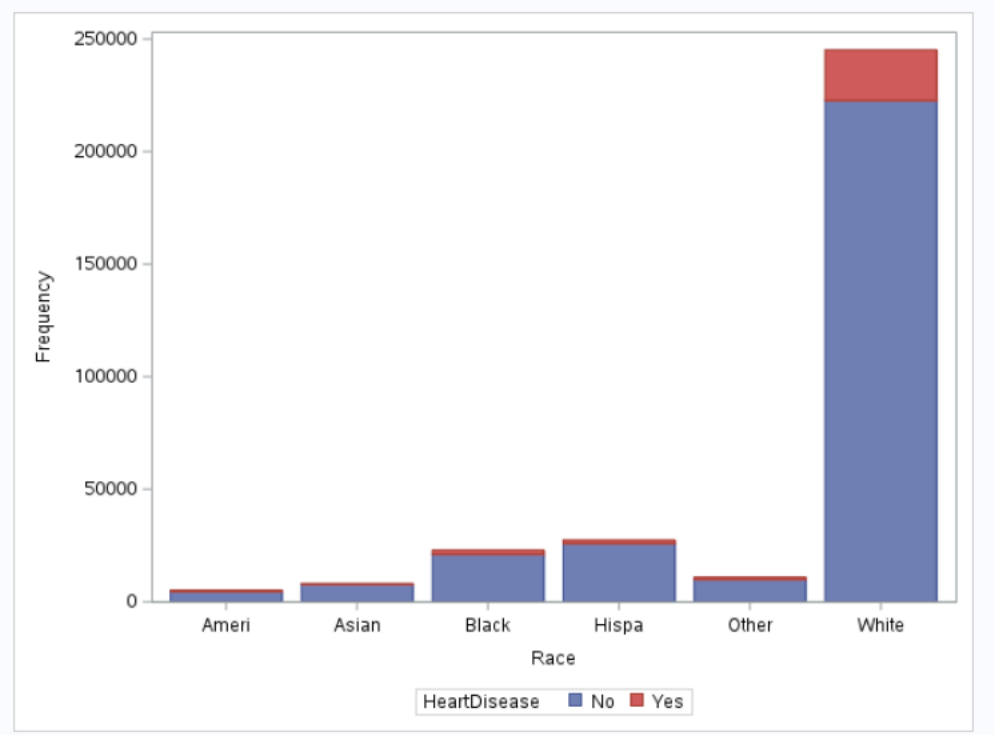

300000
200000
100000
0
Frequency
No
Yes
KidneyDisease
HeartDisease No Yes

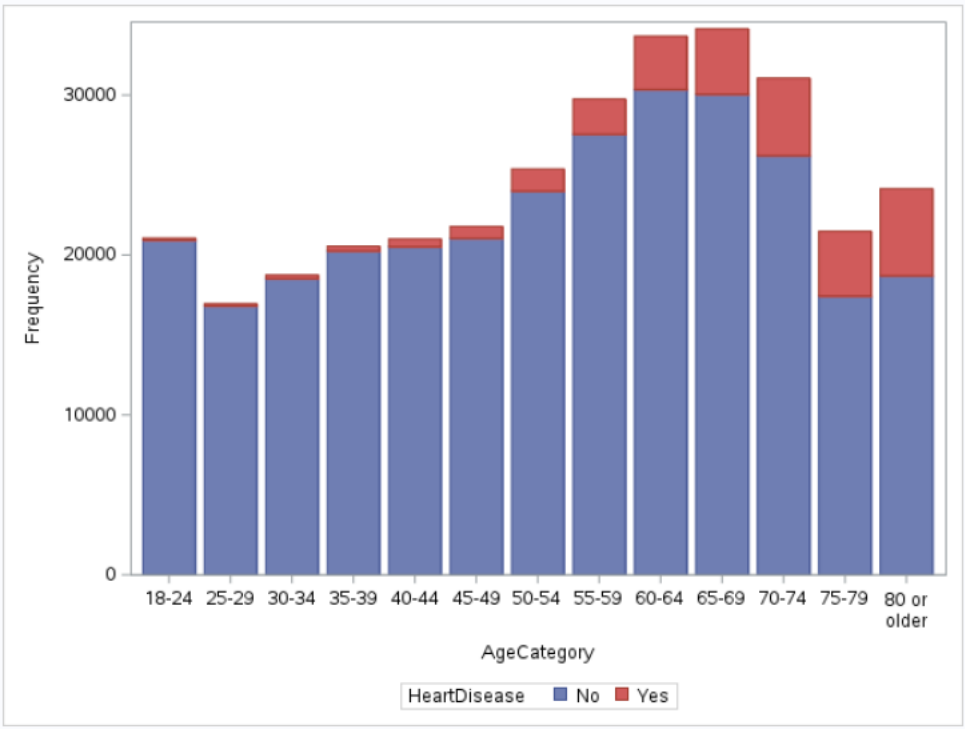

250000
200000
150000
100000
50000
0
Frequency
No
Yes
Asthma
HeartDisease No Yes

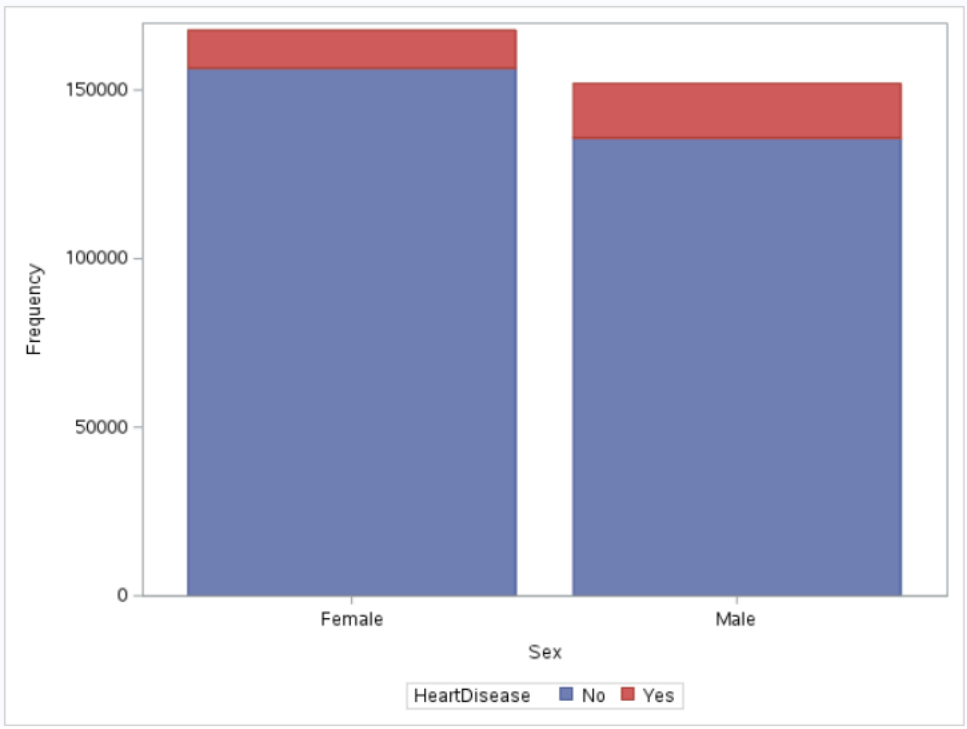

100000
80000
60000
40000
20000
0
Frequency
Excellent
Fair
Good
Poor
Very good
GenHealth
HeartDisease No Yes

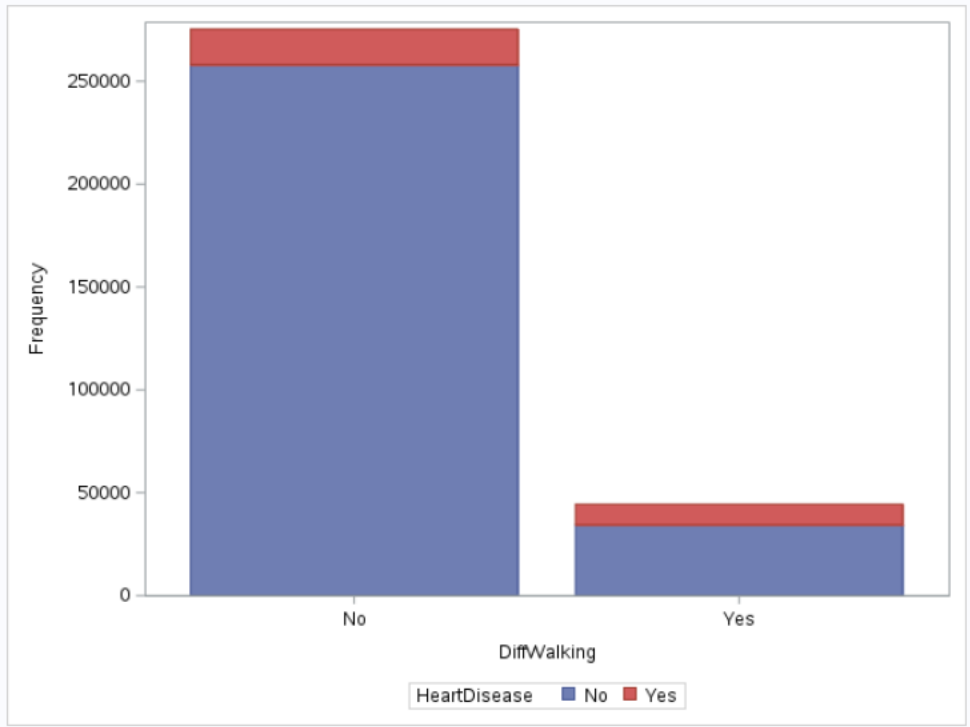

250000
200000
150000
100000
50000
0
Frequency
No
Yes
PhysicalActivity
HeartDisease No Yes

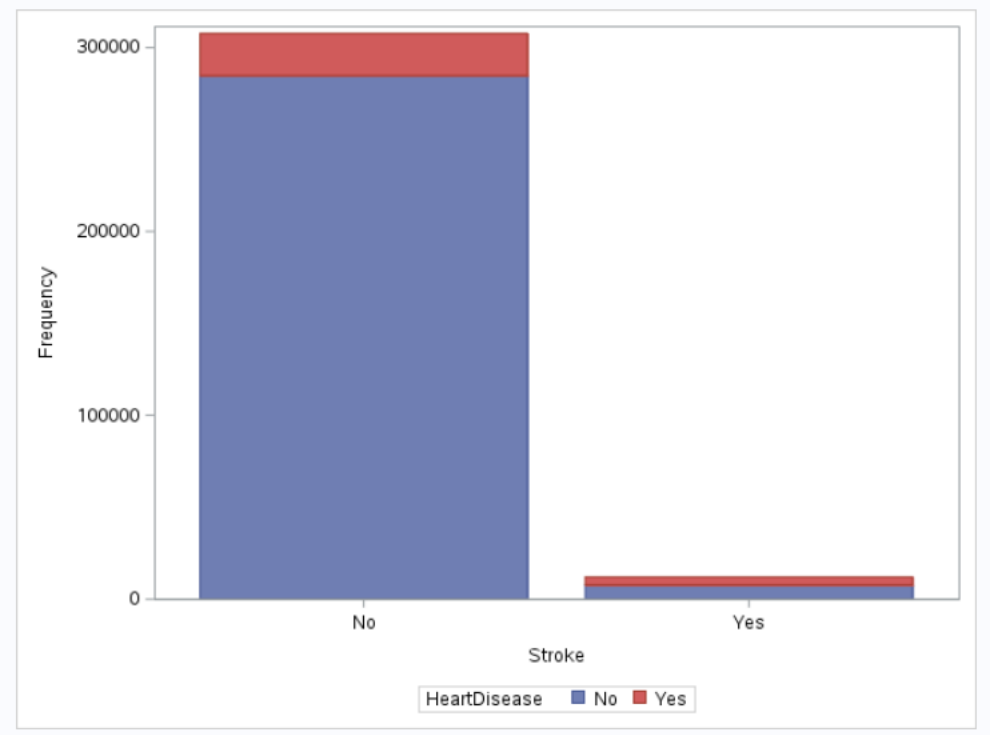

250000
200000
150000
100000
50000
0
Frequency
No
No, borderline diabetes
Yes
Yes (during pregnancy)
Diabetic
HeartDisease No Yes

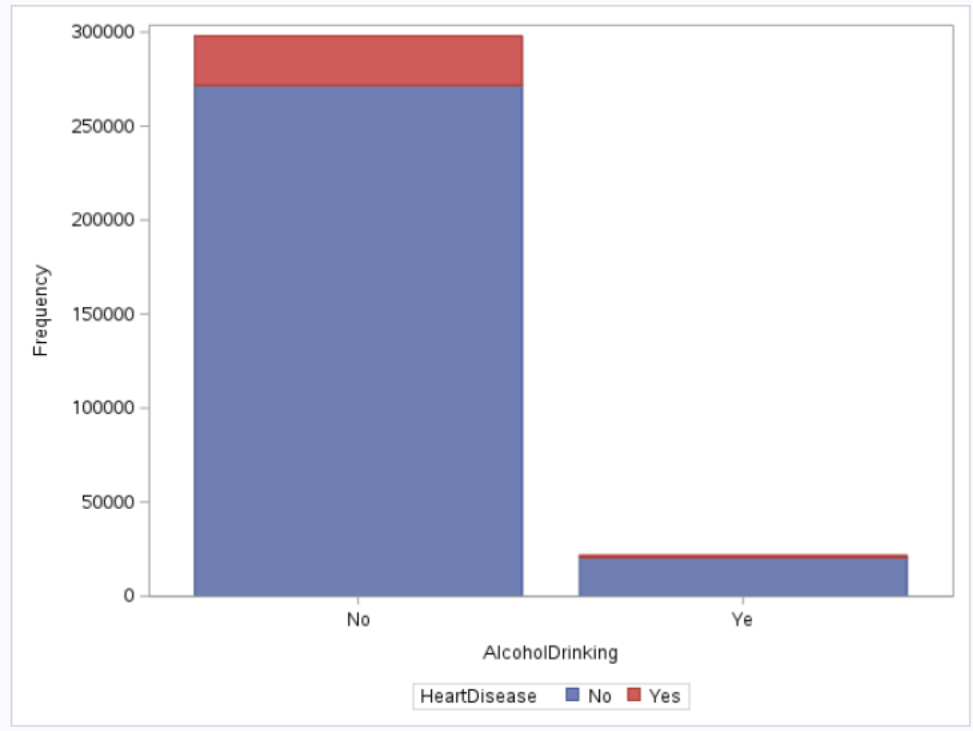

250000
200000
150000
100000
50000
0
Frequency
Ameri
Asian
Black
Hispa
Other
White
Race
HeartDisease
No
Yes

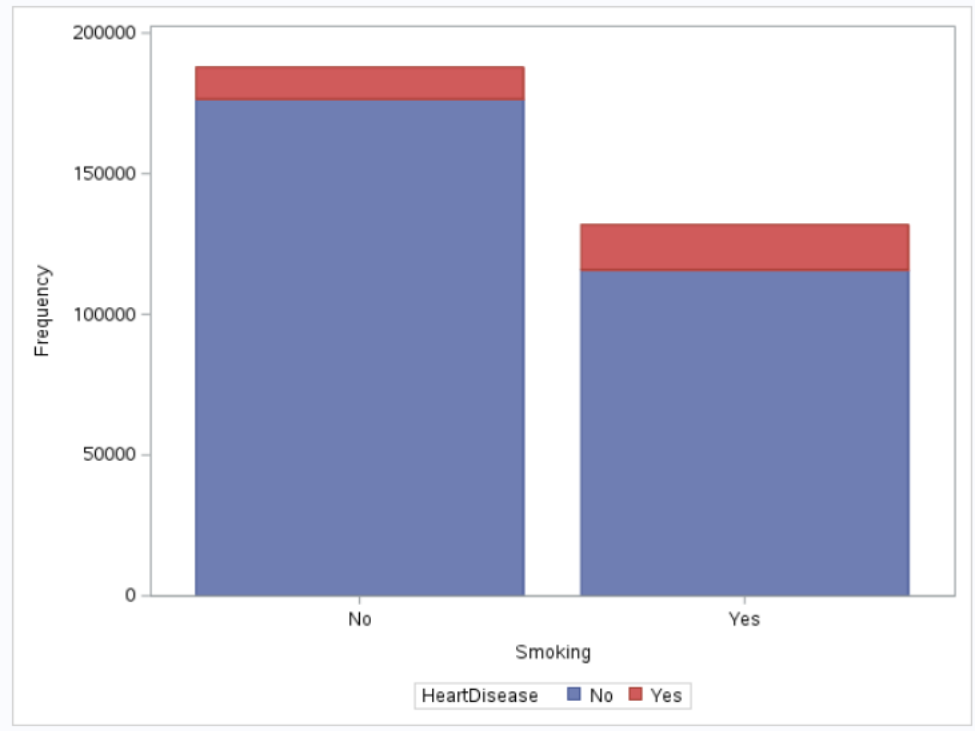

30000
20000
10000
0
Frequency
18-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80 or older
AgeCategory
HeartDisease
No
Yes


150000
100000
50000
0
Frequency
Female
Male
Sex
HeartDisease No Yes


250000
200000
150000
100000
50000
0
Frequency
No
Yes
DiffWalking
HeartDisease No Yes


300000
200000
100000
0
Frequency
No
Yes
Stroke
HeartDisease No Yes

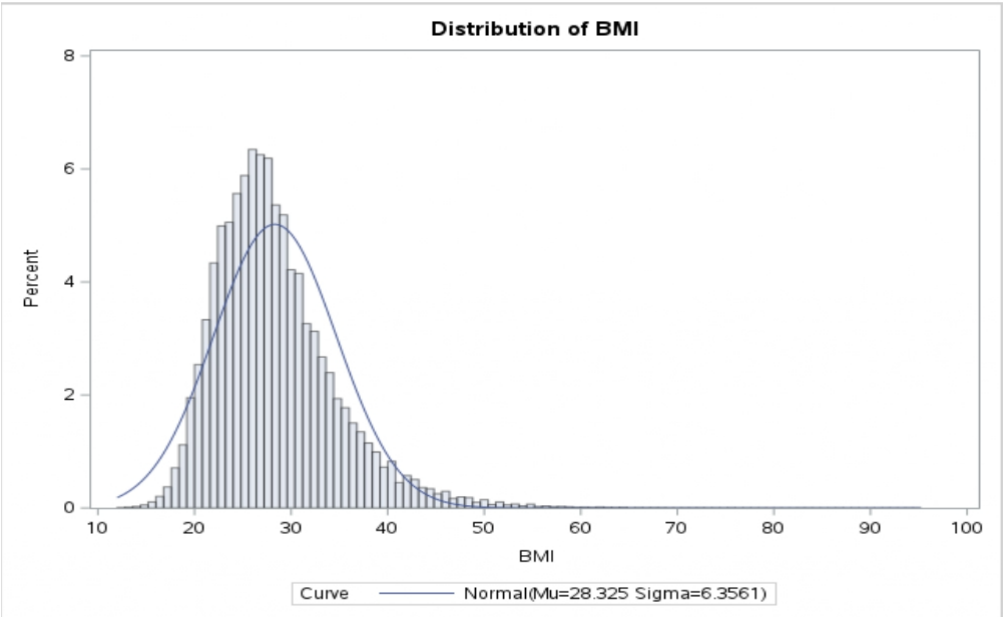

300000
250000
200000
150000
100000
50000
0
Frequency
No
Ye
AlcoholDrinking
HeartDisease No Yes



The Univariate Procedure for the numerical variables

PROC UNIVARIATE in SAS will be used to create histograms plots for the continuous variables, such as BMI, PhysicalHealth, MentalHealth, and SleepTime. This will help to understand the distribution, central tendency, and variability of each variable.

The UNIVARIATE Procedure



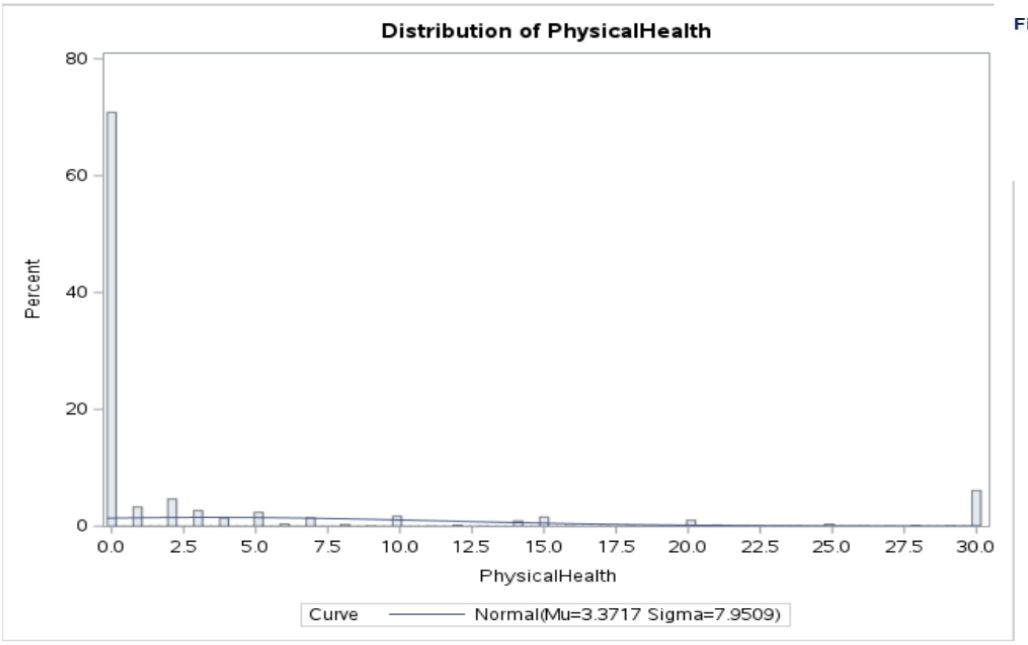
The UNIVARIATE Procedure
Fitted Normal Distribution for BMI

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	28.3254
Std Dev	Sigma	6.3561

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.08430	Pr > D	<0.010
Cramer-von Mises	W-Sq	735.20067	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	4439.45398	Pr > A-Sq	<0.005

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	17.9200	13.5389
5.0	20.1200	17.8705
10.0	21.4700	20.1797
25.0	24.0300	24.0383
50.0	27.3400	28.3254
75.0	31.4200	32.6125
90.0	36.4900	36.4711
95.0	40.1800	38.7803
99.0	48.6600	43.1119

The above section provides estimates of the mean and standard deviation of the distribution of BMI (Body Mass Index) and tests for how well the data fit a normal distribution. The estimated mean BMI is 28.3254, and the estimated standard deviation is 6.3561. The section also reports the results of three goodness-of-fit tests (Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling), which test how well the data fits a normal distribution. In each case, the null hypothesis is that the data are normally distributed. The p-values for all three tests are less than 0.005, indicating strong evidence against the null hypothesis and suggesting that the data are not well described by a normal distribution.



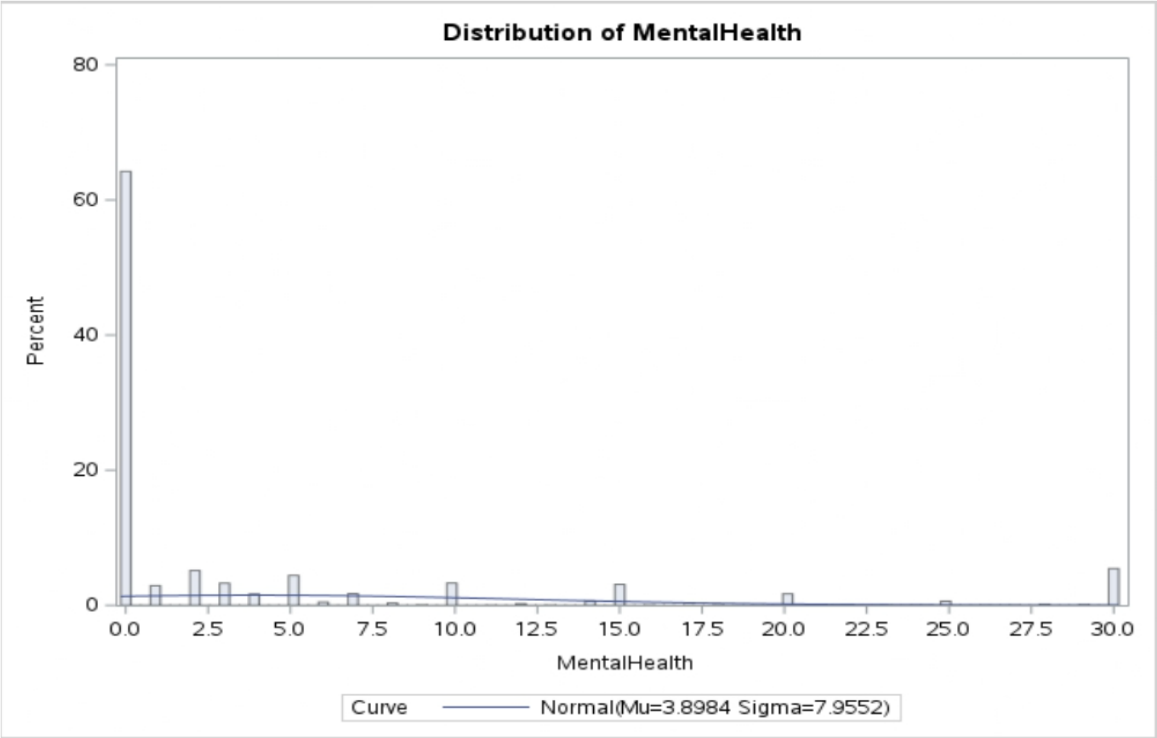
The UNIVARIATE Procedure
Fitted Normal Distribution for PhysicalHealth

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	3.37171
Std Dev	Sigma	7.95085

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	0.0000	-15.12473
5.0	0.0000	-9.70627
10.0	0.0000	-6.81771
25.0	0.0000	-1.99106
50.0	0.0000	3.37171
75.0	2.0000	8.73448
90.0	14.0000	13.56113
95.0	30.0000	16.44969
99.0	30.0000	21.86815

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.08430	Pr > D	<0.010
Cramer-von Mises	W-Sq	735.20067	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	4439.45398	Pr > A-Sq	<0.005

The above section provides a histogram of the distribution of "PhysicalHealth" and estimates the mean and standard deviation of the data assuming it follows a normal distribution. The estimated mean is 3.37171 and the estimated standard deviation is 7.95085. The section also includes the results of three goodness-of-fit tests (Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling) to test how well the data fit a normal distribution. In each case, the null hypothesis is that the data are normally distributed. The p-values for all three tests are less than 0.005, indicating strong evidence against the null hypothesis and suggesting that the data are not well described by a normal distribution.



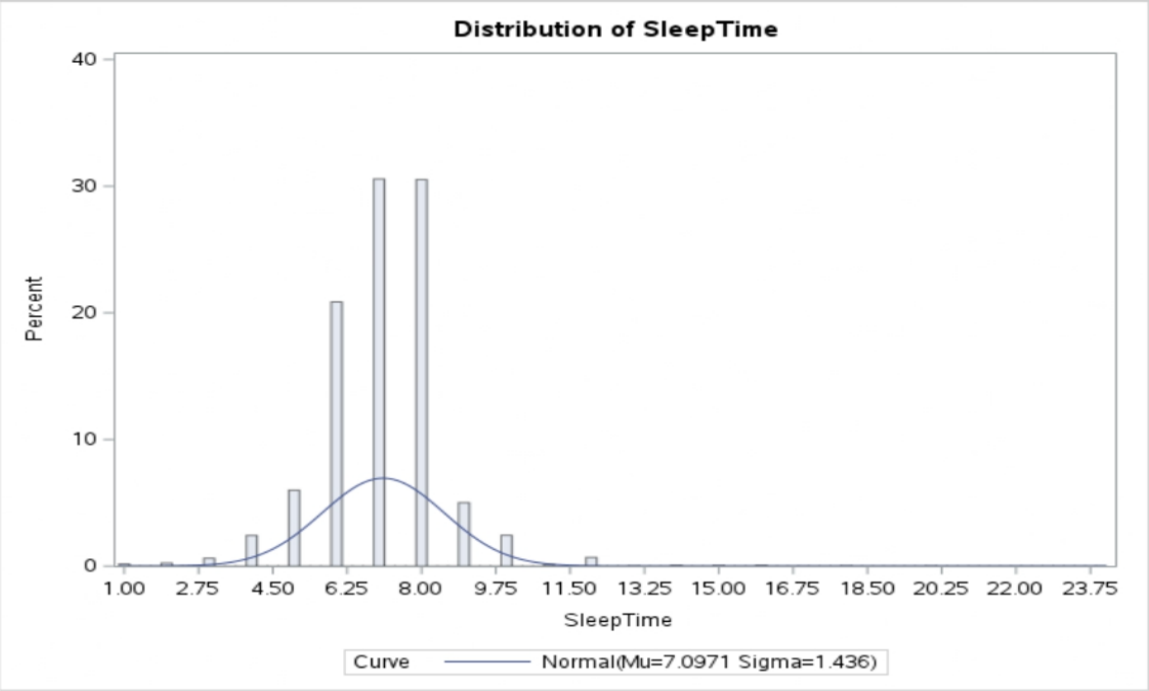
The UNIVARIATE Procedure
Fitted Normal Distribution for MentalHealth

Parameters for Normal Distribution		
Parameter	**Symbol**	**Estimate**
Mean	Mu	3.898366
Std Dev	Sigma	7.955235

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.3302	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	11724.5130	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	60099.4249	**Pr > A-Sq**	<0.005

Quantiles for Normal Distribution		
	Quantile	
Percent	**Observed**	**Estimated**
1.0	0.0000	-14.60828
5.0	0.0000	-9.18683
10.0	0.0000	-6.29668
25.0	0.0000	-1.46736
50.0	0.0000	3.89837
75.0	3.0000	9.26409
90.0	15.0000	14.09341
95.0	30.0000	16.98356
99.0	30.0000	22.40501

The above section provides a histogram of the distribution of "MentalHealth" and estimates the mean and standard deviation of the data assuming it follows a normal distribution. The estimated mean is 3.89836614 and the estimated standard deviation is 7.95523522. The section also includes the results of three goodness-of-fit tests (Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling) to test how well the data fit a normal distribution. In each case, the null hypothesis is that the data are normally distributed. The p-values for all three tests are less than 0.005, indicating strong.



Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	3.00000	3.75642
5.0	5.00000	4.73505
10.0	6.00000	5.25676
25.0	6.00000	6.12850
50.0	7.00000	7.09707
75.0	8.00000	8.06565
90.0	8.00000	8.93739
95.0	9.00000	9.45910
99.0	12.00000	10.43773

The UNIVARIATE Procedure
Fitted Normal Distribution for SleepTime

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	7.097075
Std Dev	Sigma	1.436007

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.1789	Pr > D	<0.010
Cramer-von Mises	W-Sq	1920.8181	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	10236.5817	Pr > A-Sq	<0.005

The above section is the histogram and fitted normal distribution provided in the output are based on the SleepTime variable. The histogram shows the frequency distribution of the SleepTime data, where the x-axis represents the range of SleepTime values and the y-axis represents the frequency or count of each SleepTime value. The shape of the histogram appears to be bell-shaped, which suggests that the SleepTime data follows a normal distribution.

Logistic Regression Analysis of Risk Factors

Several of the variables had to be recoded as part of the data analysis process due to format issues or non-significance. The median of each category was taken and divided by 5 to convert the variable AgeCategory into the continuous variable age. Due to it being non-significant, the level "Yes (during pregnancy)" for the variable "Diabetic" was re-coded as "No". The levels "Ameri" [American Indian/Alaskan Native] and "Other" of the variable Race were subsequently re-coded as "White" due to their non-significance.

A sample size of 5000 was taken from the data set prior to the analysis using sampling without replacement because any sample larger could cause the SAS programme to be slow and unresponsive due to the quantity of memory needed for the study. See below figure:

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	IMPORT
Random Number Seed	12345
Sample Size	5000
Selection Probability	0.015635
Sampling Weight	63.959
Number of Replicates	1
Total Sample Size	5000
Output Data Set	IMPORT_SAMPLE

To fit a generalised linear model to all of the variables, using HeartDisease as the response variable and the other 17 factors as the explanatory variables, the PROC GENMOD technique was initially employed to analyse the data. For simplicity in analysing the results, the reference levels of the class variables were set to "No" and the binomial distribution and Logit link function were utilised. One variable, MentalHealth, was found to have a significant impact on the results of the Analysis of Maximum Likelihood Parameter Estimates. The Genmod procedure is illustrated below:

Model Information	
Data Set	WORK.IMPORT_SAMPLE
Response Variable	HeartDisease
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	5000
Number of Observations Used	5000

Response Profile		
Ordered Value	HeartDisease	Total Frequency
1	Yes	460
2	No	4540

Probability modeled is HeartDisease='Yes'.

Backward Elimination Procedure

Class Level Information					
Class	Value	Design Variables			
Smoking	No	0			
	Yes	1			
AlcoholDrinking	No	0			
	Ye	1			
Stroke	No	0			
	Yes	1			
DiffWalking	No	0			
	Yes	1			
Sex	Female	1			
	Male	0			
Race	Asian	1	0	0	
	Black	0	1	0	
	Hispa	0	0	1	
	White	0	0	0	
Diabetic	No	0	0		
	No, borderline diabetes	1	0		
	Yes	0	1		
GenHealth	Excellent	1	0	0	0

Class Level Information					
Class	Value	Design Variables			
	Fair	0	1	0	0
	Good	0	0	1	0
	Poor	0	0	0	0
	Very good	0	0	0	1
Asthma	No	0			
	Yes	1			
KidneyDisease	No	0			
	Yes	1			
SkinCancer	No	0			
	Yes	1			
PhysicalActivity	No	0			
	Yes	1			

Step 0. The following effects were entered:

Intercept BMI Smoking AlcoholDrinking Stroke PhysicalHealth MentalHealth DiffWalking Sex age PhysicalActivity Race Diabetic GenHealth SleepTime Asthma KidneyDisease SkinCancer

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	3073.408	2368.360
SC	3079.926	2524.773
-2 Log L	3071.408	2320.360

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	751.0484	23	<.0001
Score	861.3301	23	<.0001
Wald	528.6948	23	<.0001

Step 1. Effect PhysicalHealth is removed:

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	3073.408	2366.400
SC	3079.926	2516.295
-2 Log L	3071.408	2320.400

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	751.0084	22	<.0001
Score	861.3301	22	<.0001
Wald	528.8460	22	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
0.0401	1	0.8412

Step 2. Effect BMI is removed:

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	3073.408	2364.447
SC	3079.926	2507.825
-2 Log L	3071.408	2320.447

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	750.9612	21	<.0001
Score	859.0720	21	<.0001
Wald	529.2842	21	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
0.0874	2	0.9572

Step 3. Effect Race is removed:

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	3073.408	2360.448
SC	3079.926	2484.275
-2 Log L	3071.408	2322.448

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	748.9600	18	<.0001
Score	858.4423	18	<.0001
Wald	529.2618	18	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
1.9172	5	0.8605

Step 4. Effect PhysicalActivity is removed:

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	3073.408	2358.806
SC	3079.926	2476.115
-2 Log L	3071.408	2322.806

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	748.6026	17	<.0001
Score	858.3489	17	<.0001
Wald	529.3596	17	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
2.2657	6	0.8937

Step 5. Effect SkinCancer is removed:

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	3073.408	2357.608
SC	3079.926	2468.400
-2 Log L	3071.408	2323.608

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	747.8006	16	<.0001
Score	855.1989	16	<.0001
Wald	528.2333	16	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
3.0651	7	0.8789

Step 6. Effect SleepTime is removed:

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	3073.408	2356.476
SC	3079.926	2460.751
-2 Log L	3071.408	2324.476

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	746.9328	15	<.0001
Score	855.1924	15	<.0001
Wald	528.9851	15	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
3.9391	8	0.8626

Step 7. Effect Asthma is removed:

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	3073.408	2355.448
SC	3079.926	2453.206
-2 Log L	3071.408	2325.448

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	745.9599	14	<.0001
Score	854.8539	14	<.0001
Wald	528.6425	14	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
4.9264	9	0.8407

Step 8. Effect MentalHealth is removed:

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	3073.408	2355.443
SC	3079.926	2446.684
-2 Log L	3071.408	2327.443

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	743.9654	13	<.0001
Score	854.6655	13	<.0001
Wald	529.3280	13	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
6.9566	10	0.7295

Step 9. Effect AlcoholDrinking is removed:

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	3073.408	2355.745
SC	3079.926	2440.468
-2 Log L	3071.408	2329.745

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	741.6637	12	<.0001
Score	852.8111	12	<.0001
Wald	528.3093	12	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
9.0715	11	0.6153

Note: No (additional) effects met the 0.05 significance level for removal from the model.

Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	PhysicalHealth	1	16	0.0401	0.8412
2	BMI	1	15	0.0473	0.8278
3	Race	3	14	1.7963	0.6157
4	PhysicalActivity	1	13	0.3555	0.5510
5	SkinCancer	1	12	0.8135	0.3671
6	SleepTime	1	11	0.8644	0.3525
7	Asthma	1	10	0.9909	0.3195
8	MentalHealth	1	9	2.0416	0.1530
9	AlcoholDrinking	1	8	2.1169	0.1457

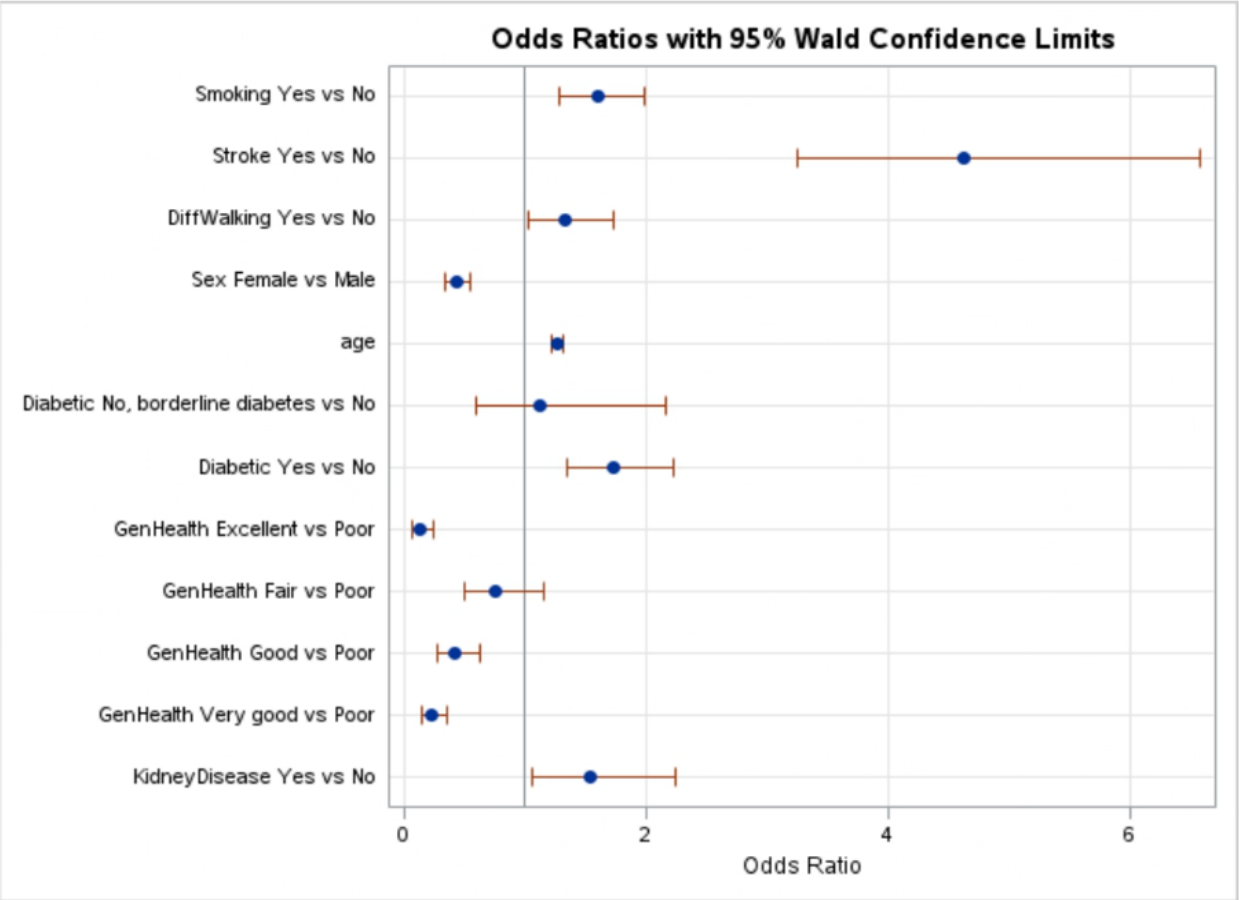
Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Smoking	1	17.7632	<.0001
Stroke	1	72.6536	<.0001
DiffWalking	1	4.7945	0.0286
Sex	1	52.6154	<.0001
age	1	143.6996	<.0001
Diabetic	2	18.8451	<.0001
GenHealth	4	85.7367	<.0001
KidneyDisease	1	5.0171	0.0251

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-4.4022	0.3367	170.9253	<.0001
Smoking	Yes	1	0.4717	0.1119	17.7632	<.0001
Stroke	Yes	1	1.5323	0.1798	72.6536	<.0001
DiffWalking	Yes	1	0.2926	0.1336	4.7945	0.0286
Sex	Female	1	-0.8296	0.1144	52.6154	<.0001
age		1	0.2370	0.0198	143.6996	<.0001
Diabetic	No, borderline diabetes	1	0.1248	0.3313	0.1419	0.7064
Diabetic	Yes	1	0.5500	0.1268	18.8114	<.0001
GenHealth	Excellent	1	-1.9847	0.2961	44.9263	<.0001
GenHealth	Fair	1	-0.2811	0.2155	1.7017	0.1921
GenHealth	Good	1	-0.8787	0.2147	16.7499	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
GenHealth	Very good	1	-1.4780	0.2328	40.3244	<.0001
KidneyDisease	Yes	1	0.4324	0.1930	5.0171	0.0251

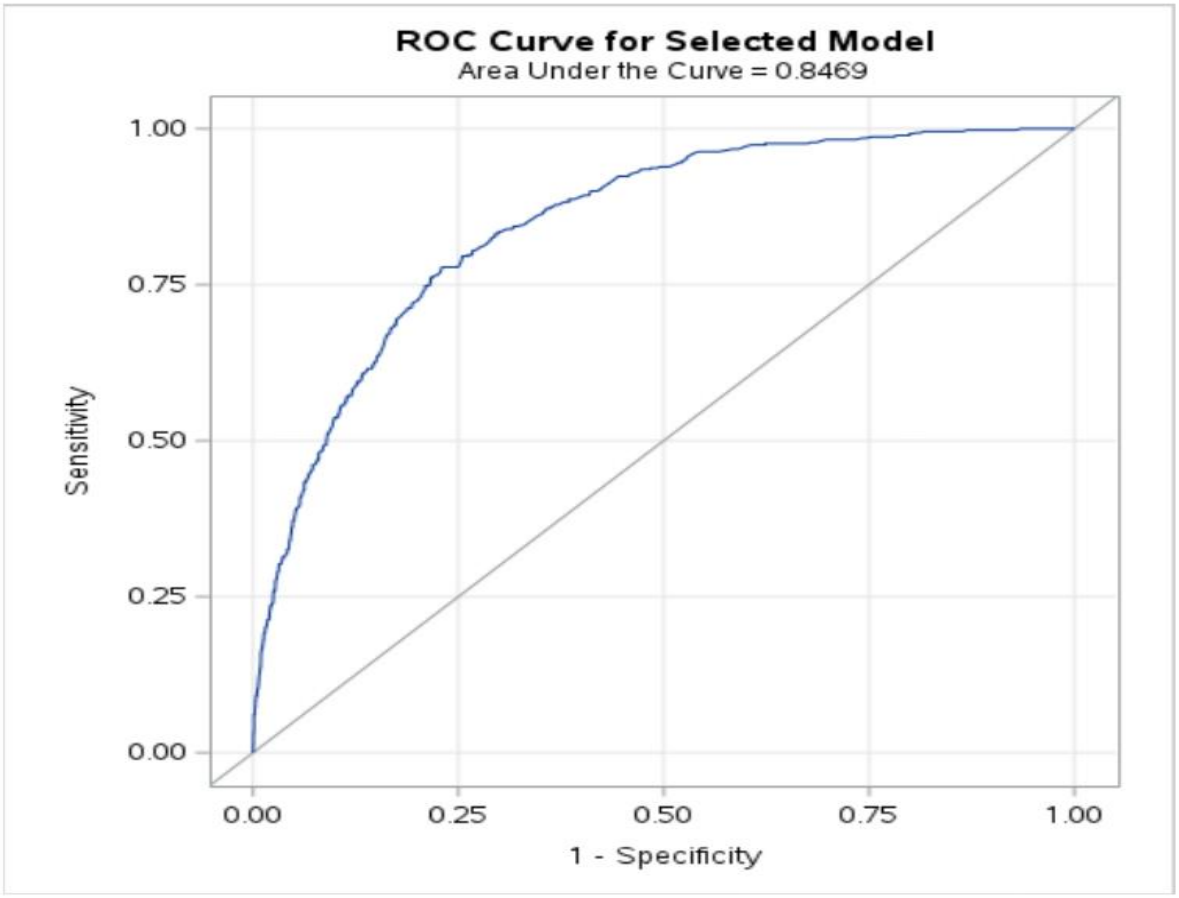
Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Smoking Yes vs No	1.603	1.287	1.996
Stroke Yes vs No	4.629	3.254	6.584
DiffWalking Yes vs No	1.340	1.031	1.741
Sex Female vs Male	0.436	0.349	0.546
age	1.267	1.219	1.318
Diabetic No, borderline diabetes vs No	1.133	0.592	2.169
Diabetic Yes vs No	1.733	1.352	2.222
GenHealth Excellent vs Poor	0.137	0.077	0.246
GenHealth Fair vs Poor	0.755	0.495	1.152
GenHealth Good vs Poor	0.415	0.273	0.633
GenHealth Very good vs Poor	0.228	0.145	0.360
KidneyDisease Yes vs No	1.541	1.056	2.250

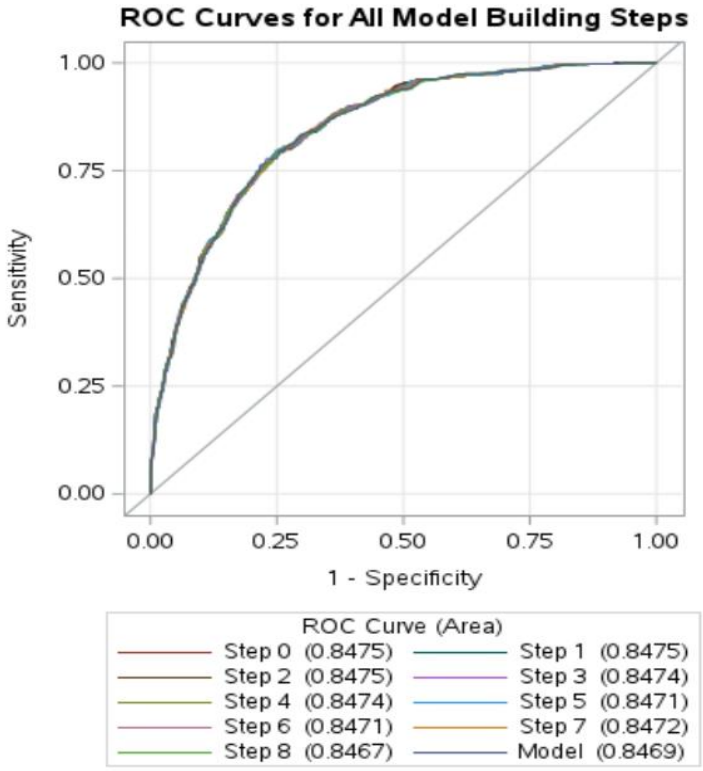
An Odd Ratios map was produced as part of the logistic regression to analyse the data. The resulting graph (image below) demonstrates that having ever experienced a stroke significantly increases the risk of developing heart disease, with the risk being 4.6 times higher than in people who have never experienced a stroke. In addition, regardless of your level of general health, having a stroke does significantly raise your chances of developing heart disease (see examples below). The odds of developing cardiovascular disease increase by 60.3% in smokers, while the odds of developing heart disease increase by 1.733 times in diabetics. Smoking and diabetes are the next two leading risk factors. Finally, women have 0.436 times the likelihood of developing heart disease than men do.



Association of Predicted Probabilities and Observed Responses			
Percent Concordant	84.6	Somers' D	0.694
Percent Discordant	15.2	Gamma	0.695
Percent Tied	0.2	Tau-a	0.116
Pairs	2088400	c	0.847

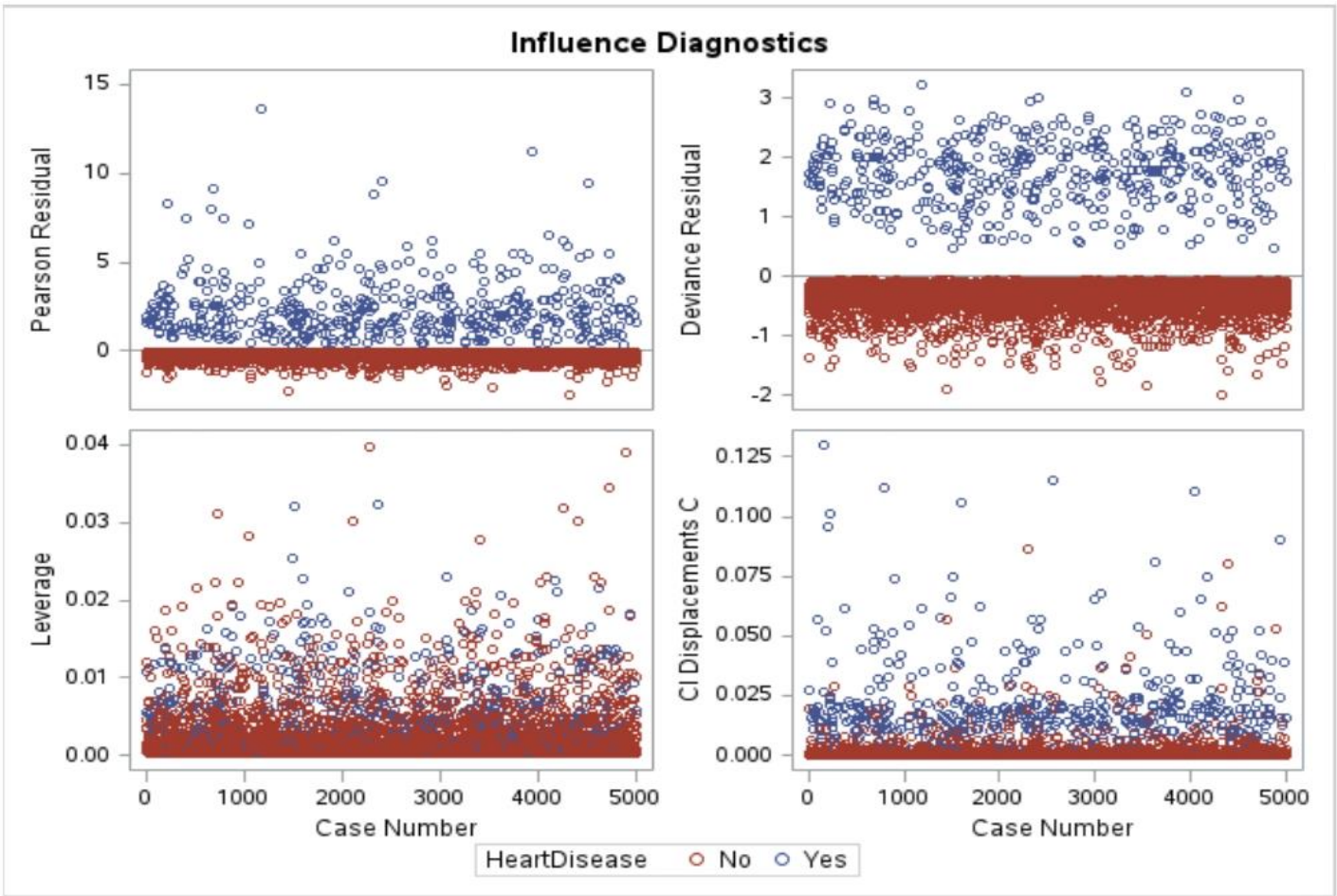
A second approach, PROC LOGISTIC, was used to create a different model using logistic regression and the Backward Elimination approach in order to validate this one. Nine variables—Physical Health, BMI, Race, Physical Activity, Skin Cancer, Sleep Quality, Asthma, Mental Health, and Alcohol Consumption—were eliminated as a result of this research. Thus, smoking, stroke, difficulty walking, age, diabetes, general health, and kidney disease comprised the final model. Given that excessive drinking and obesity are good indicators of the risk of heart disease, the removal of BMI and AlcoholDrinking from the model came as a surprise. However, this could be explained by the correlation that BMI and AlcoholDrinking had with other variables in the final model, such as Smoking or Diabetic. The obtained AIC value was 3073.408, which was considerably higher than the PROC GENMOD AIC score. Additionally, a ROC Curve was made for the generated model (see examples below), and it revealed an AUC of 0.8469, well within the >0.80 criterion for a good score, indicating that the generated model has an 84.69% accuracy rate based on the supplied parameters in predicting the presence of heart disease. The Hosmer and Lemeshow Goodness-of-match Test, which showed that the model is a good match for the given sample and had a p-value of 0.2659, gave additional evidence of the model's fit. The final model is HeartDisease = -4.4022+0.4717 Smoking + 1.5323 Stroke + 0.2926 DiffWalking - 0.8296 Sex + 0.237age + 0.55 Diabetic + GenHealth (-1.9847 = Excellent, -1.478 = VeryGood, -0.8787 = Good) + 0.4324 KidneyDisease.

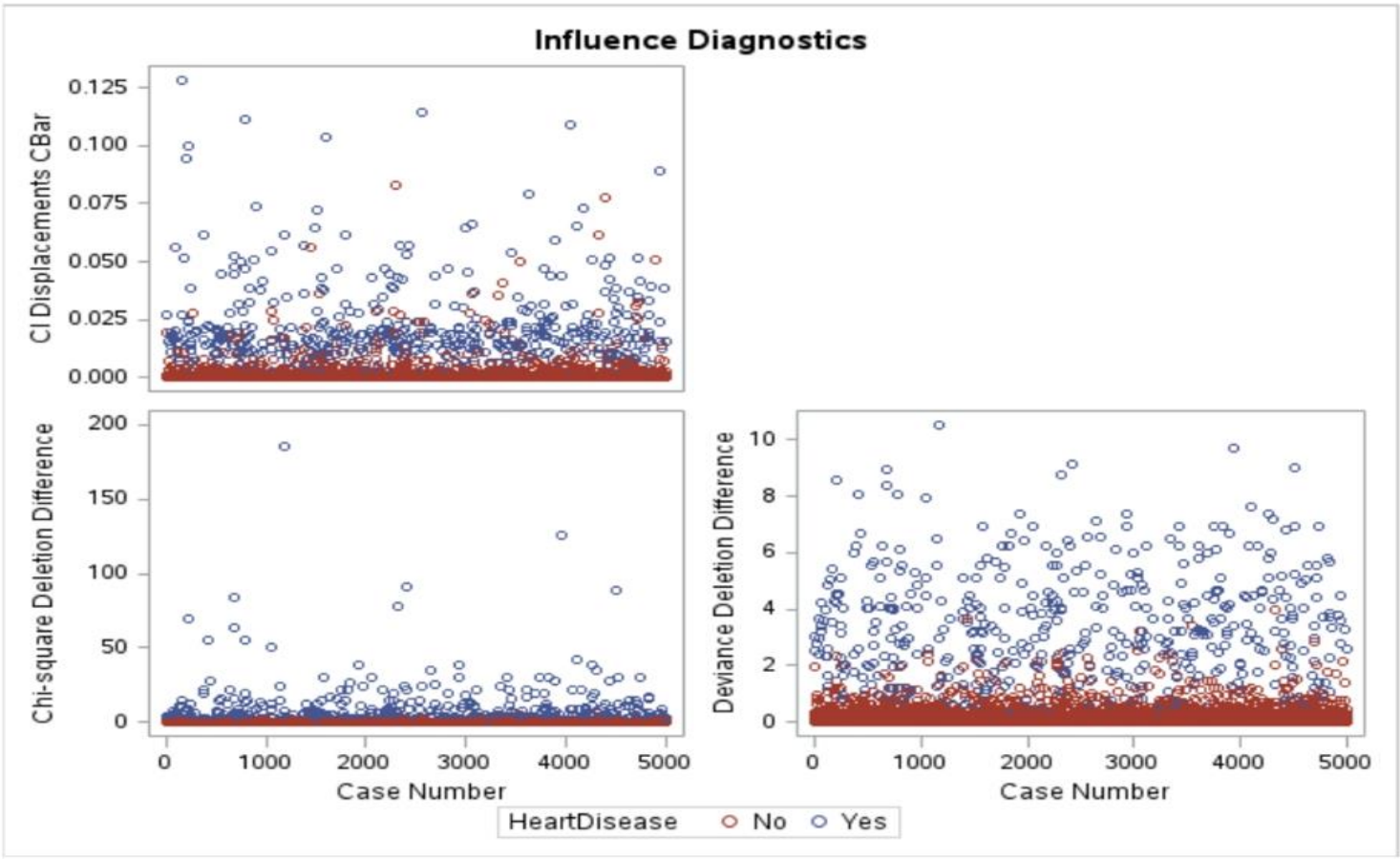




Partition for the Hosmer and Lemeshow Test					
		HeartDisease = Yes		HeartDisease = No	
Group	Total	Observed	Expected	Observed	Expected
1	509	1	2.40	508	506.60
2	509	5	5.02	504	503.98
3	510	5	8.29	505	501.71
4	500	6	12.35	494	487.65
5	500	18	17.74	482	482.26
6	504	29	25.76	475	478.24
7	500	38	37.01	462	462.99
8	501	70	56.96	431	444.04
9	501	96	94.11	405	406.89
10	466	192	200.35	274	265.65

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
9.9880	8	0.2659



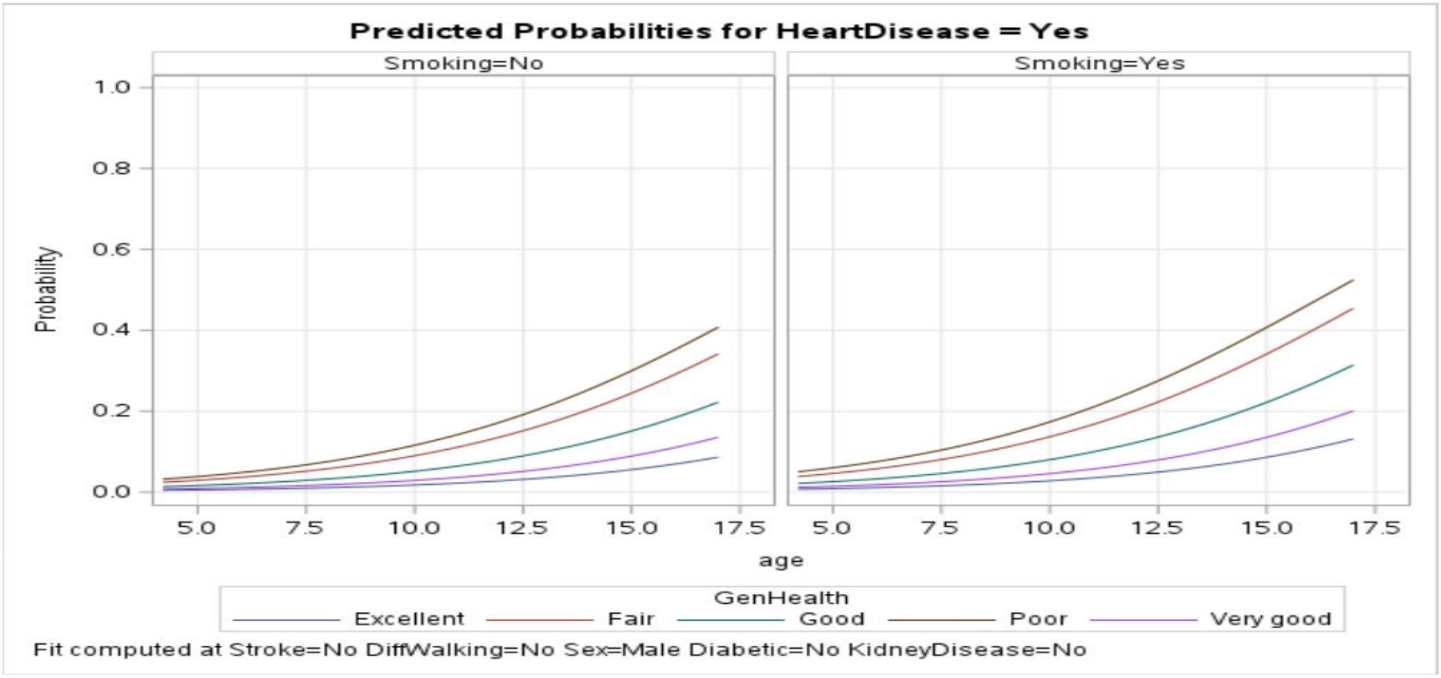


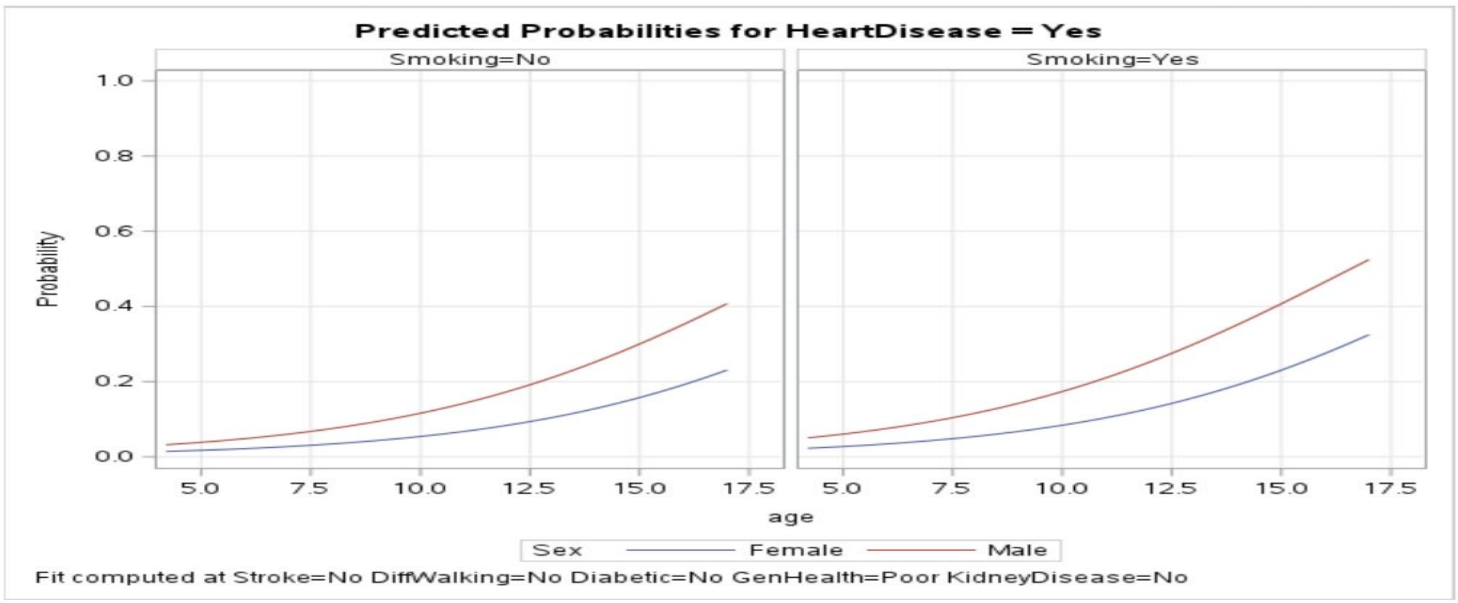
Predicted Probabilities of Heart Disease

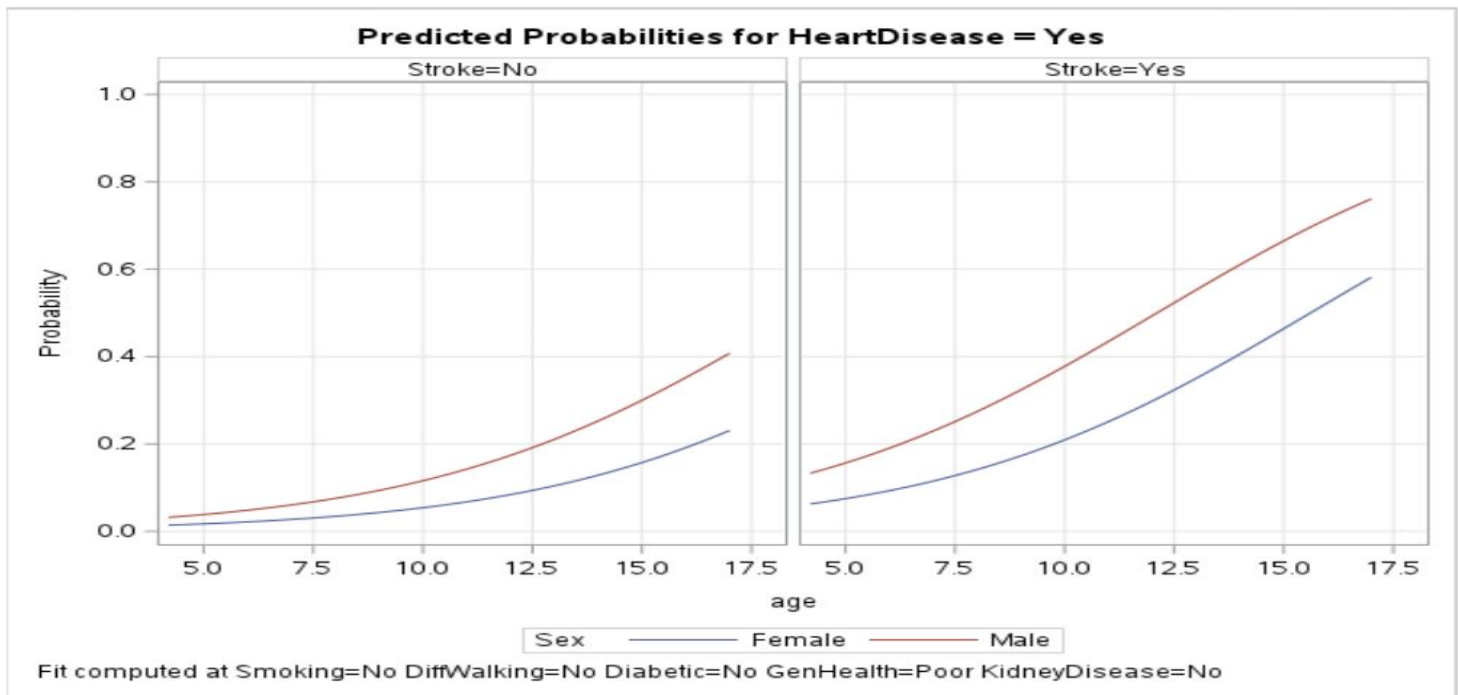
The PLM Procedure

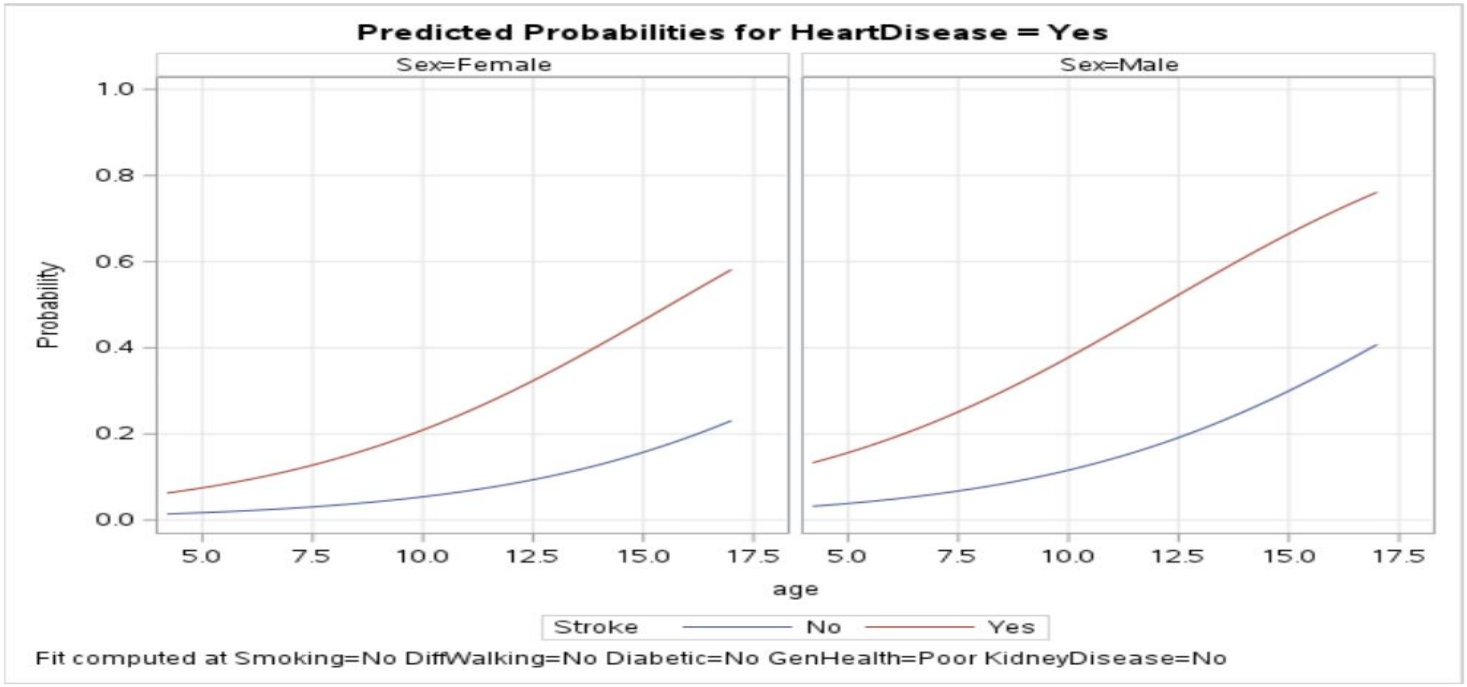
Store Information	
Item Store	WORK.LOGIMODEL
Data Set Created From	WORK.IMPORT_SAMPLE
Created By	PROC LOGISTIC
Date Created	30APR23:17:49:14
Response Variable	HeartDisease
Link Function	Logit
Distribution	Binary
Class Variables	Smoking AlcoholDrinking Stroke DiffWalking Sex agecatnum Race Diabetic GenHealth Asthma ...
Model Effects	Intercept Smoking Stroke DiffWalking Sex age Diabetic GenHealth KidneyDisease

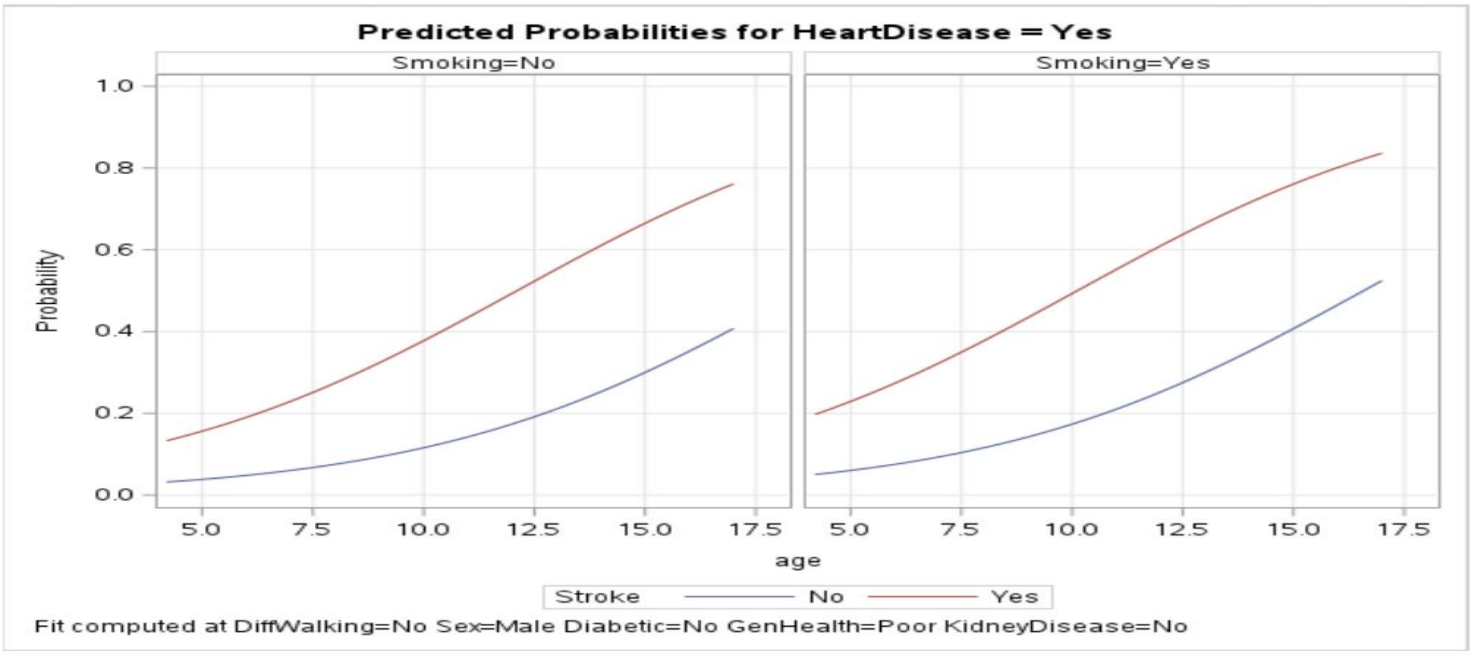
Class Level Information		
Class	**Levels**	**Values**
Smoking	2	No Yes
AlcoholDrinking	2	No Ye
Stroke	2	No Yes
DiffWalking	2	No Yes
Sex	2	Female Male
agecatnum	13	21 27 32 37 42 47 52 57 62 67 72 77 85
Race	4	Asian Black Hispa White
Diabetic	3	No No, borderline diabetes Yes
GenHealth	5	Excellent Fair Good Poor Very good
Asthma	2	No Yes
KidneyDisease	2	No Yes
SkinCancer	2	No Yes
PhysicalActivity	2	No Yes
HeartDisease	2	Yes No

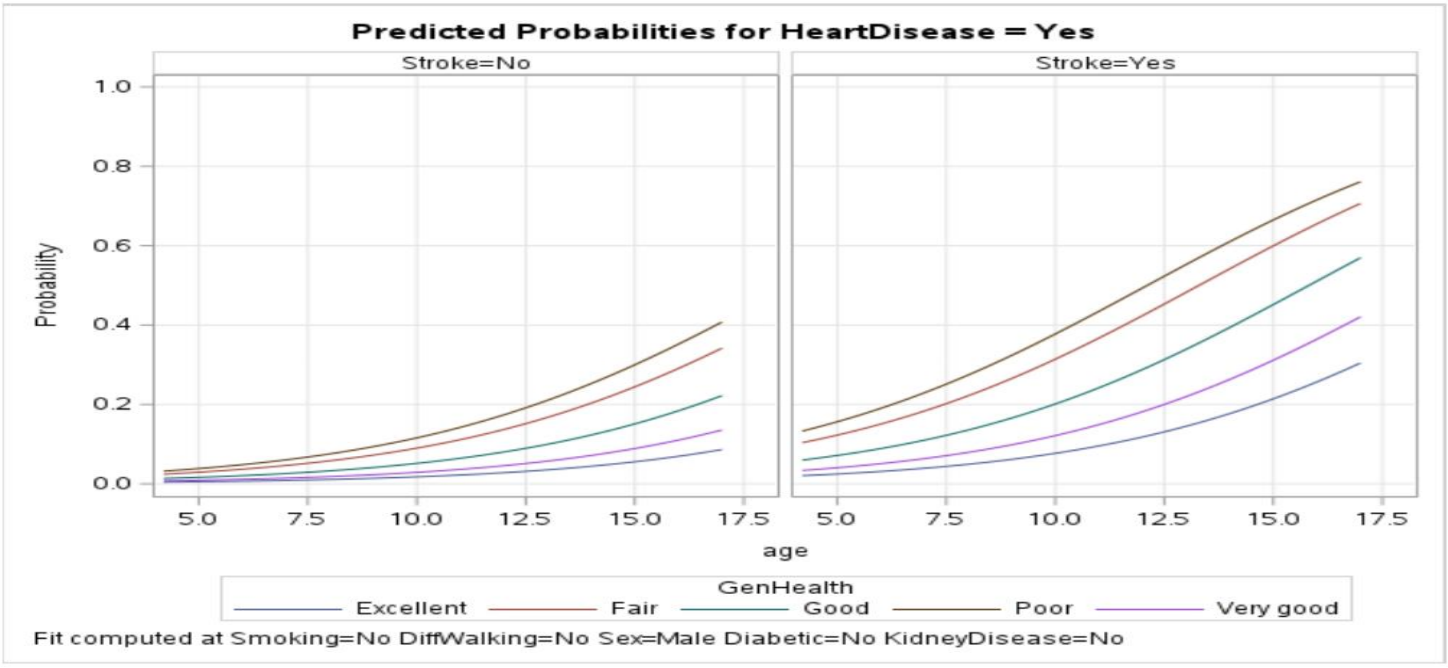

Predicted Probabilities for HeartDisease = Yes
Smoking=No
Smoking=Yes
Probability
1.0
0.8
0.6
0.4
0.2
0.0
5.0
7.5
10.0
12.5
15.0
17.5
age
GenHealth
Excellent
Fair
Good
Poor
Very good
Fit computed at Stroke=No DiffWalking=No Sex=Male Diabetic=No KidneyDisease=No


Predicted Probabilities for HeartDisease = Yes
Smoking=No
Smoking=Yes
Probability
1.0
0.8
0.6
0.4
0.2
0.0
5.0
7.5
10.0
12.5
15.0
17.5
age
Sex
Female
Male
Fit computed at Stroke=No DiffWalking=No Diabetic=No GenHealth=Poor KidneyDisease=No


Predicted Probabilities for HeartDisease = Yes
Stroke=No
Stroke=Yes
Probability
1.0
0.8
0.6
0.4
0.2
0.0
5.0
7.5
10.0
12.5
15.0
17.5
age
Sex
Female
Male
Fit computed at Smoking=No DiffWalking=No Diabetic=No GenHealth=Poor KidneyDisease=No


Predicted Probabilities for HeartDisease = Yes
Sex=Female
Sex=Male
Probability
1.0
0.8
0.6
0.4
0.2
0.0
5.0
7.5
10.0
12.5
15.0
17.5
age
Stroke
No
Yes
Fit computed at Smoking=No DiffWalking=No Diabetic=No GenHealth=Poor KidneyDisease=No


Predicted Probabilities for HeartDisease = Yes
Smoking=No
Smoking=Yes
Probability
1.0
0.8
0.6
0.4
0.2
0.0
5.0
7.5
10.0
12.5
15.0
17.5
age
Stroke
No
Yes
Fit computed at DiffWalking=No Sex=Male Diabetic=No GenHealth=Poor KidneyDisease=No


Predicted Probabilities for HeartDisease = Yes
Stroke=No
Stroke=Yes
Probability
1.0
0.8
0.6
0.4
0.2
0.0
5.0
7.5
10.0
12.5
15.0
17.5
age
GenHealth
Excellent
Fair
Good
Poor
Very good
Fit computed at Smoking=No DiffWalking=No Sex=Male Diabetic=No KidneyDisease=No

SAS and R in depth Comparison

The data science community favours the statistical analysis software programmes SAS and R. While R is an open-source programming language with a steeper learning curve but offers a wide range of statistical analysis functions, SAS is a powerful statistical tool that requires some prior statistical knowledge and coding experience (Fitzgerald, 2021). In the table below, SAS and R, are critically compared in terms of installation and setup, general use, data pre-processing, data analytics, and output.

Criteria	SAS Score	R Score
Installation and configuration	3	5
Platform Dependence	3	4
Platform Ease	4	4
Overall Effectiveness	4	5
Friendliness to Users	5	4
Effortless Use	3	4
Data Importing, Data Cleaning, and Data Pre-Processing	3	4
Logistic Regression	4	4
Transformation of Data	3	5
Combination of Data	3	5
Data Subsetting, Descriptive Analytics, Data Analytics	3	5
Visualisation of Output Data	4	5
Sharing results	4	4
Total	46	58

Table 1: Critical comparison of R and SAS

In terms of setup and installation, SAS and R both offer simple setups. SAS depends on specific system prerequisites, like the operating system, which can complicate the installation procedure. R, on the other hand, has no dependencies and can be easily installed on multiple operating systems.

R is more suited to data scientists and programmers and generally requires some prior programming experience. Contrarily, SAS is developed for users with less programming skills and offers a more user-friendly interface (Fitzgerald, 2021).

Both SAS and R include a variety of data cleaning, transformation, and combination tools for pre-processing data. While SAS provides better data combination functions, R has more sophisticated data cleaning features like imputation and outlier detection (Fitzgerald, 2021).

Both SAS and R provide a broad range of statistical analysis functions for data analytics, including regression analysis and time series forecasting. Statistical analysis tools like distribution analysis, correlation analysis, cluster analysis, and time series forecasting are all included in SAS. These functions in R can be easily performed using built-in packages or third-party packages (Fitzgerald, 2021).

R has more robust visualisation capabilities than SAS in terms of output, with packages like ggplot2 offering a variety of interactive and static visualisations. On the other hand, SAS provides superior reporting capabilities and enables the generation of reports that are of a professional calibre and include tables and charts (Fitzgerald, 2021).

In conclusion, each software package has advantages and disadvantages, and the decision between SAS and R For the analysis carried out in this research, the chosen preference is R over SAS. While SAS is better suited to users with less programming experience who need advanced reporting and statistical analysis capabilities, R is better suited to data scientists and programmers who require advanced data cleaning and visualisation functions.

R data Analysis

For the data analysis in R, kindly refer to the Appendix 2

Conclusion

In conclusion, this study analysed a health-related dataset to investigate the correlations between various factors and heart disease using SAS and R for data analysis and visualisation. The study found that logistic regression was a better fit than the generalized linear model for interpreting the relationship between the response variable HeartDisease and the explanatory variables. The results showed that factors such as age, sex, general health status, smoking, difficulty walking, diabetes, and history of strokes were significant predictors of heart disease. Moreover, the study emphasised the importance of a healthy lifestyle, including regular exercise, a balanced diet, and refraining from smoking, to lower the chances of developing heart disease. This research contributes to the identification of preventative strategies and focused therapies to reduce the risk of heart disease, which is a leading cause of death globally. However, the study also acknowledges that larger samples may be required to further validate the logistic regression model. Overall, the findings of this study provide valuable insights into the factors associated with heart disease and highlight the potential of SAS and R as effective tools for data analysis and visualisation in the healthcare sector.

Limitations: One of the study's weaknesses is the cross-sectional nature of the dataset utilised for analysis, which makes it difficult to establish a link between the factors and heart disease. The dataset also lacks data on several elements that may influence the risk of developing heart disease, including family history, socioeconomic position, and environmental factors. Additionally, the study ignores the interactions between many risk factors and solely considers the relationship between individual risk factors and heart disease.

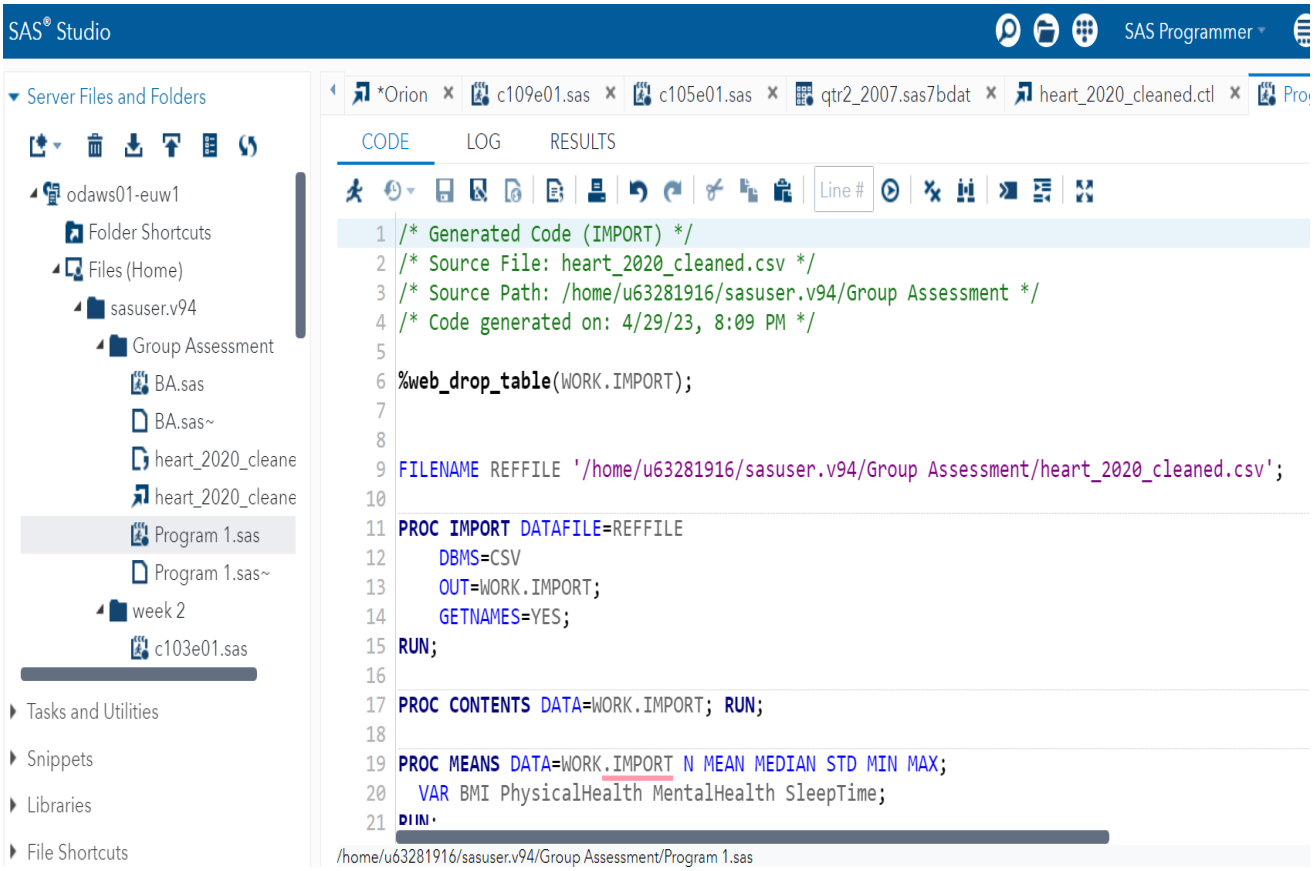
Future research: By using longitudinal data to show causality and integrating information on additional risk factors, future research may be able to overcome some of the shortcomings of this study. Future research may also investigate how different risk factors interact with one another and how that affects the likelihood of developing heart disease. Future study may also focus on creating predictive models that incorporate many risk indicators and give each person a unique risk assessment. Future studies should also investigate how well therapies that target modifiable risk factors can lower the incidence of heart disease.

References:

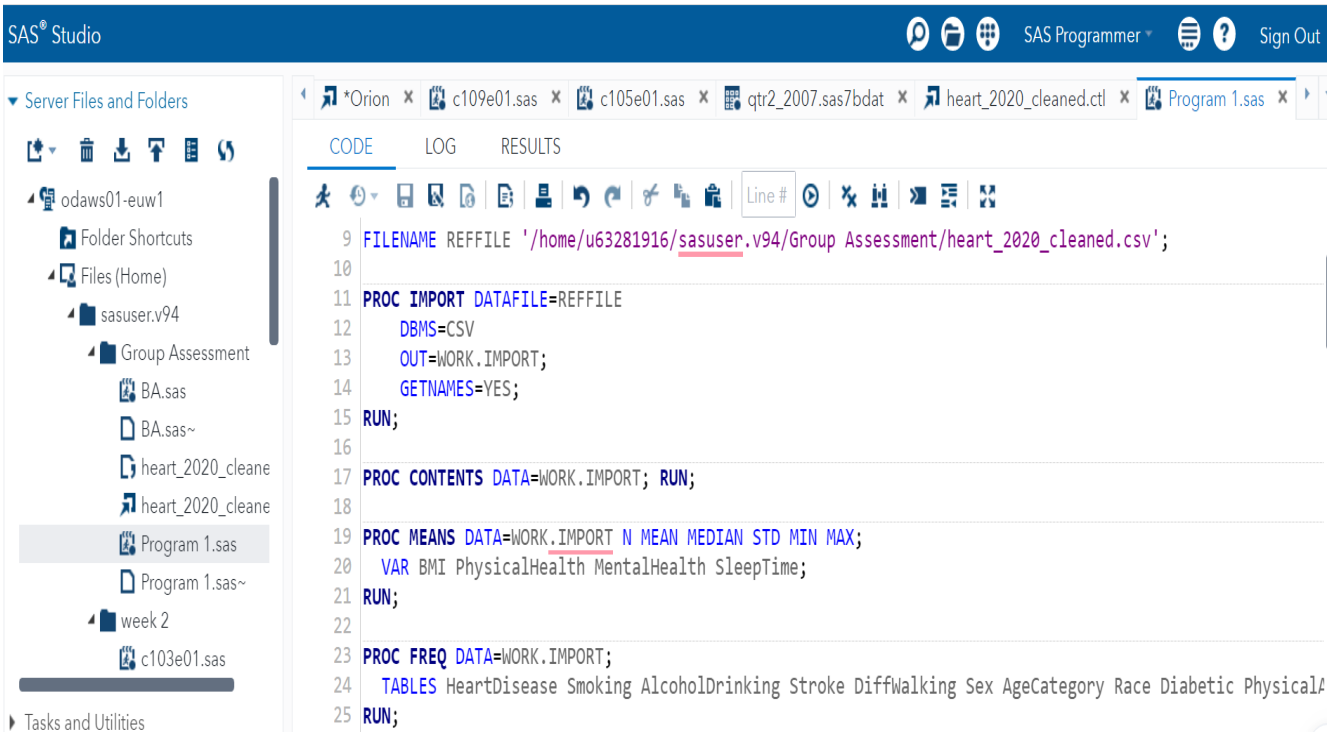
1. R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/
2. SAS Institute Inc. (2021). SAS Software. https://www.sas.com/en_us/home.html
3. World Health Organization. (2020). Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
4. Chen, M., & Liu, X. (2019). Data modeling for big data: A review. Journal of Data and Information Science, 4(4), 20-41.
5. Fang, X., Chen, Z., & Liu, Y. (2018). A survey of big data cleaning. Journal of Big Data, 5(1), 1-26.
6. Li, X., Zhang, C., & Liu, Z. (2018). Big data analytics and its applications. Journal of Big Data, 5(1), 1-17.
7. Wang, J., & Sun, L. (2019). A survey of big data visualization. Journal of Big Data, 6(1), 1-20.
8. Wu, X., Wang, X., & Li, L. (2019). Data transformation in big data: A review. Journal of Big Data, 6(1), 1-21.
9. Fitzgerald, K. (2021). SAS vs R: What's the Difference? Datamation. Retrieved from https://www.datamation.com/big-data/sas-vs-r.html

Appendix 1 – SAS Code

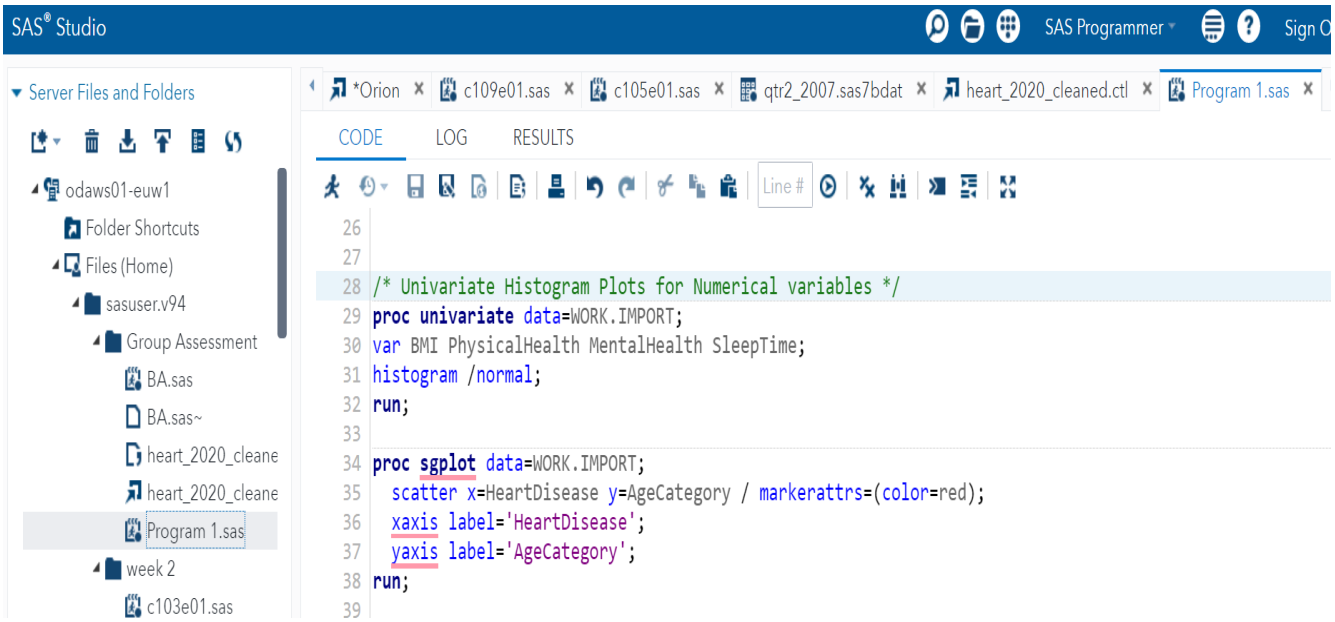
Data Preparation & Processing

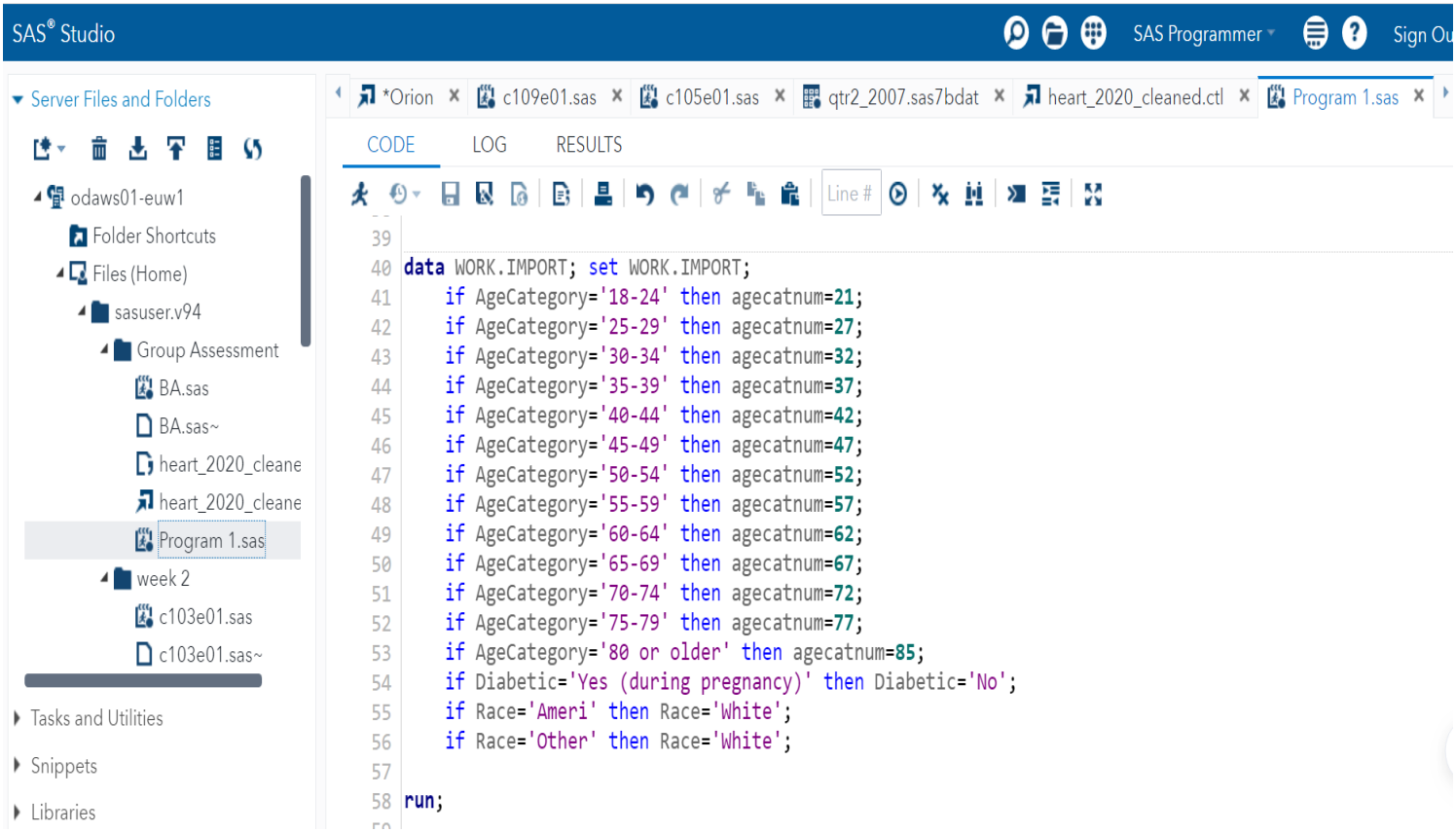


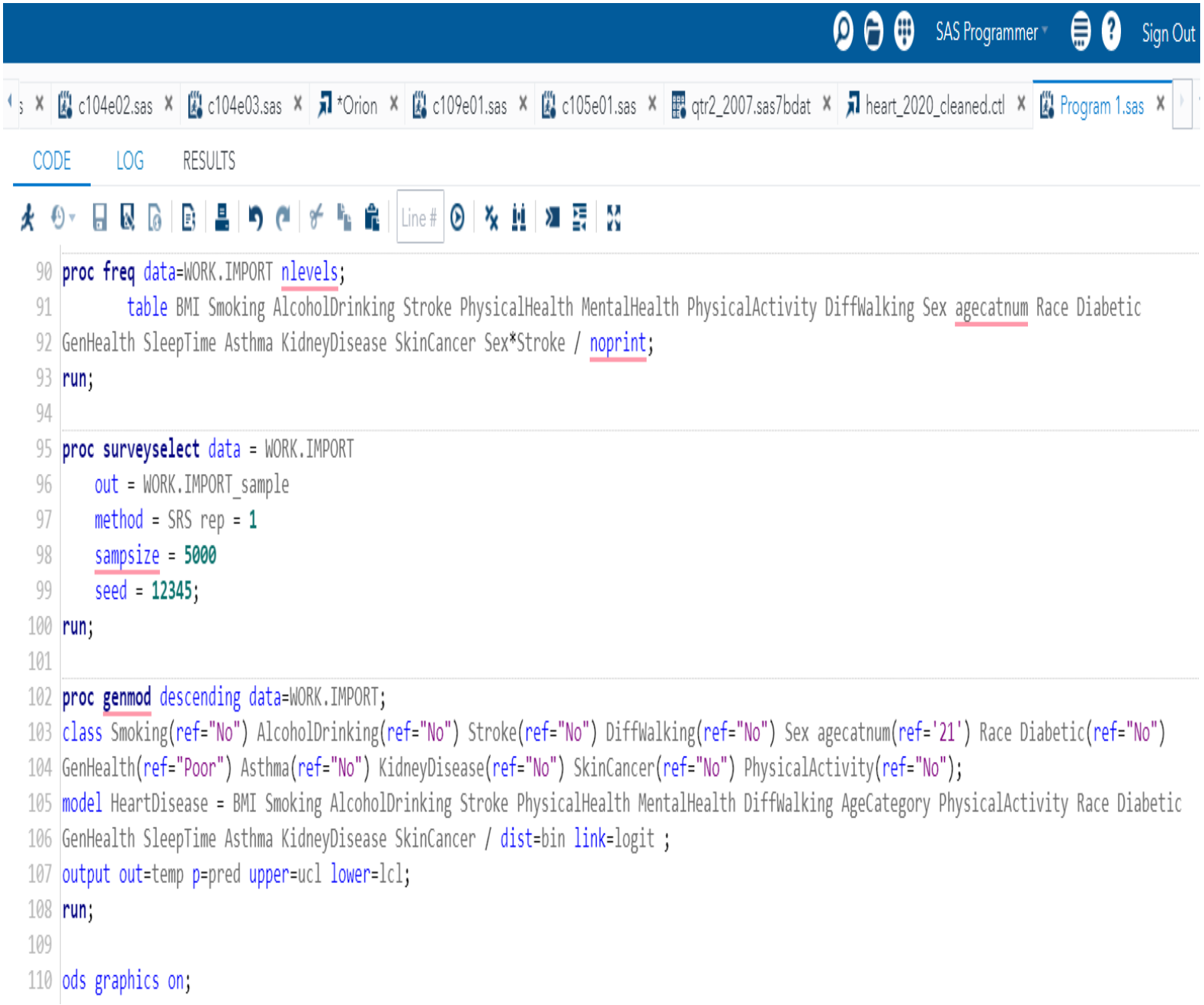
Summary Statistics

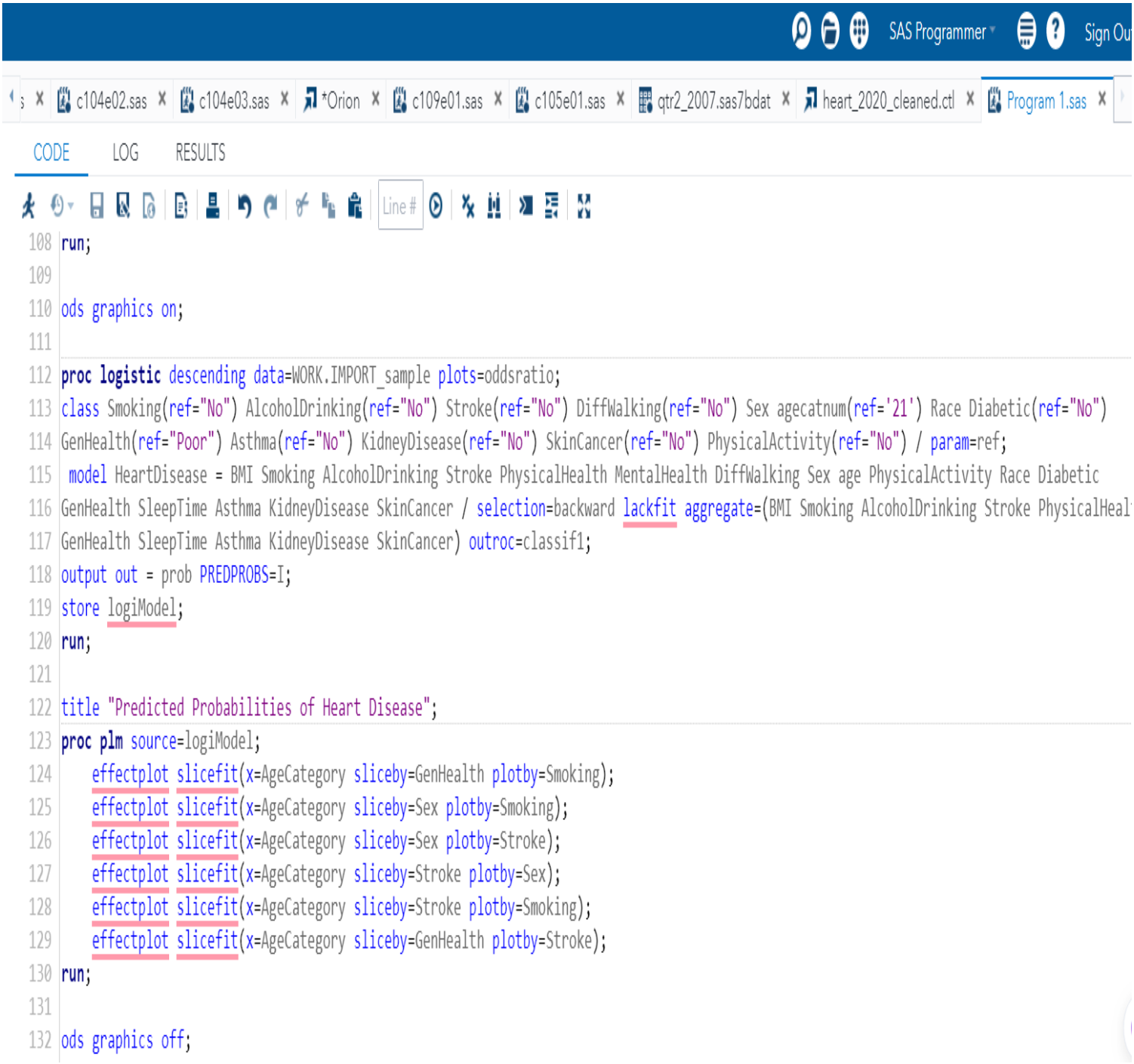


Univariate Procedure



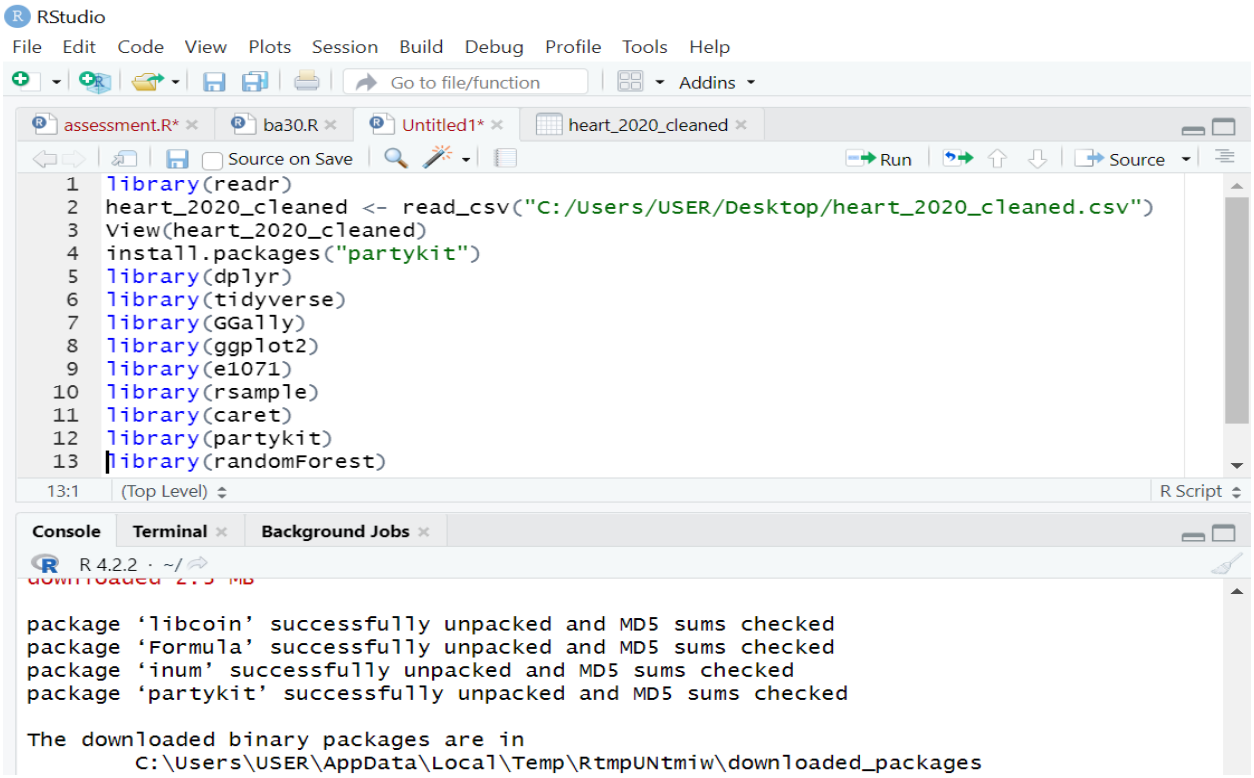



SAS Programmer
Sign Out
c104e02.sas
c104e03.sas
*Orion
c109e01.sas
c105e01.sas
qtr2_2007.sas7bdat
heart_2020_cleaned.ctl
Program 1.sas
CODE
LOG
RESULTS
Line #
90 proc freq data=WORK.IMPORT nlevels;
91 table BMI Smoking AlcoholDrinking Stroke PhysicalHealth MentalHealth PhysicalActivity DiffWalking Sex agecatnum Race Diabetic
92 GenHealth SleepTime Asthma KidneyDisease SkinCancer Sex*Stroke / noprint;
93 run;
94
95 proc surveyselect data = WORK.IMPORT
96 out = WORK.IMPORT_sample
97 method = SRS rep = 1
98 sampsize = 5000
99 seed = 12345;
100 run;
101
102 proc genmod descending data=WORK.IMPORT;
103 class Smoking(ref="No") AlcoholDrinking(ref="No") Stroke(ref="No") DiffWalking(ref="No") Sex agecatnum(ref='21') Race Diabetic(ref="No")
104 GenHealth(ref="Poor") Asthma(ref="No") KidneyDisease(ref="No") SkinCancer(ref="No") PhysicalActivity(ref="No");
105 model HeartDisease = BMI Smoking AlcoholDrinking Stroke PhysicalHealth MentalHealth DiffWalking AgeCategory PhysicalActivity Race Diabetic
106 GenHealth SleepTime Asthma KidneyDisease SkinCancer / dist=bin link=logit ;
107 output out=temp p=pred upper=ucl lower=lcl;
108 run;
109
110 ods graphics on;


SAS Programmer
Sign Ou
c104e02.sas
c104e03.sas
*Orion
c109e01.sas
c105e01.sas
qtr2_2007.sas7bdat
heart_2020_cleaned.ctl
Program 1.sas
CODE
LOG
RESULTS
Line #
108 run;
109
110 ods graphics on;
111
112 proc logistic descending data=WORK.IMPORT_sample plots=oddsratio;
113 class Smoking(ref="No") AlcoholDrinking(ref="No") Stroke(ref="No") DiffWalking(ref="No") Sex agecatnum(ref='21') Race Diabetic(ref="No")
114 GenHealth(ref="Poor") Asthma(ref="No") KidneyDisease(ref="No") SkinCancer(ref="No") PhysicalActivity(ref="No") / param=ref;
115 model HeartDisease = BMI Smoking AlcoholDrinking Stroke PhysicalHealth MentalHealth DiffWalking Sex age PhysicalActivity Race Diabetic
116 GenHealth SleepTime Asthma KidneyDisease SkinCancer / selection=backward lackfit aggregate=(BMI Smoking AlcoholDrinking Stroke PhysicalHeal
117 GenHealth SleepTime Asthma KidneyDisease SkinCancer) outroc=classif1;
118 output out = prob PREDPROBS=I;
119 store logiModel;
120 run;
121
122 title "Predicted Probabilities of Heart Disease";
123 proc plm source=logiModel;
124 effectplot slicefit(x=AgeCategory sliceby=GenHealth plotby=Smoking);
125 effectplot slicefit(x=AgeCategory sliceby=Sex plotby=Smoking);
126 effectplot slicefit(x=AgeCategory sliceby=Sex plotby=Stroke);
127 effectplot slicefit(x=AgeCategory sliceby=Stroke plotby=Sex);
128 effectplot slicefit(x=AgeCategory sliceby=Stroke plotby=Smoking);
129 effectplot slicefit(x=AgeCategory sliceby=GenHealth plotby=Stroke);
130 run;
131
132 ods graphics off;

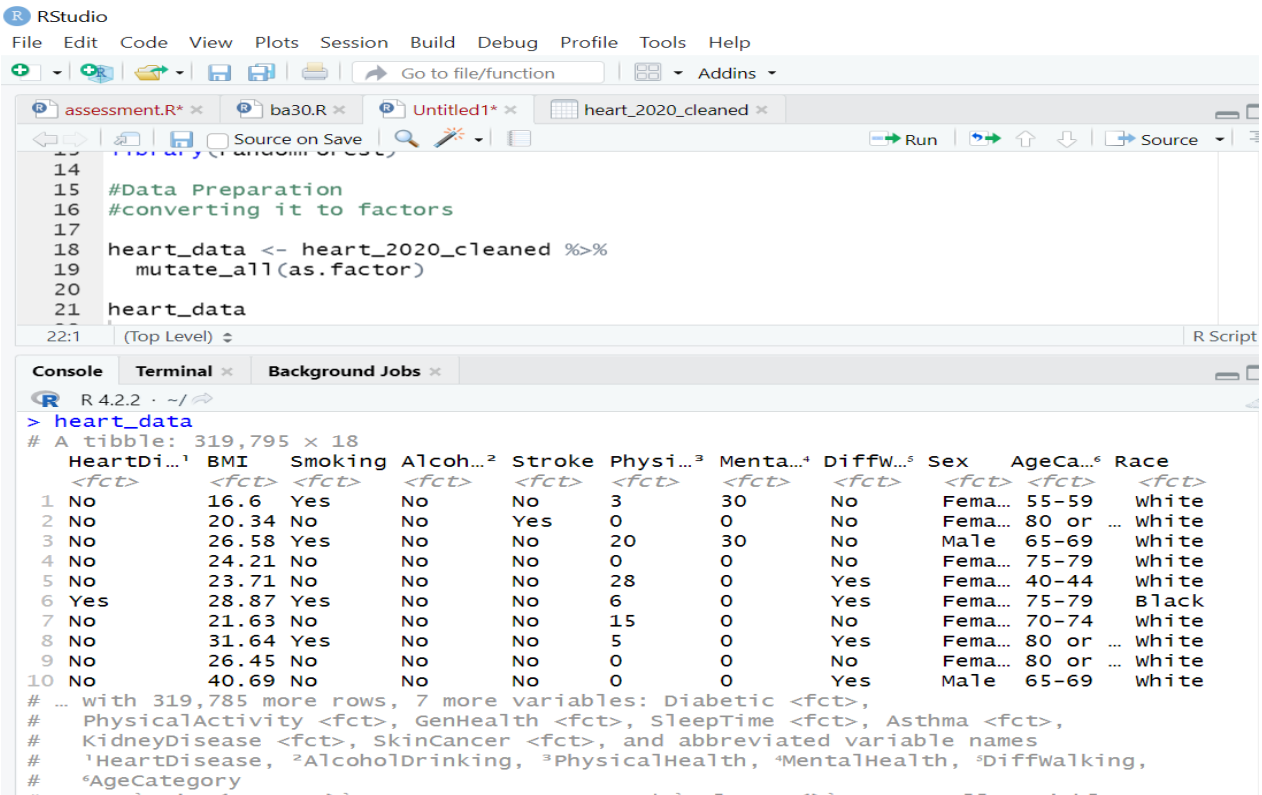
Appendix 2 – R Analysis and Code implementations.

Dataset and Library Importation

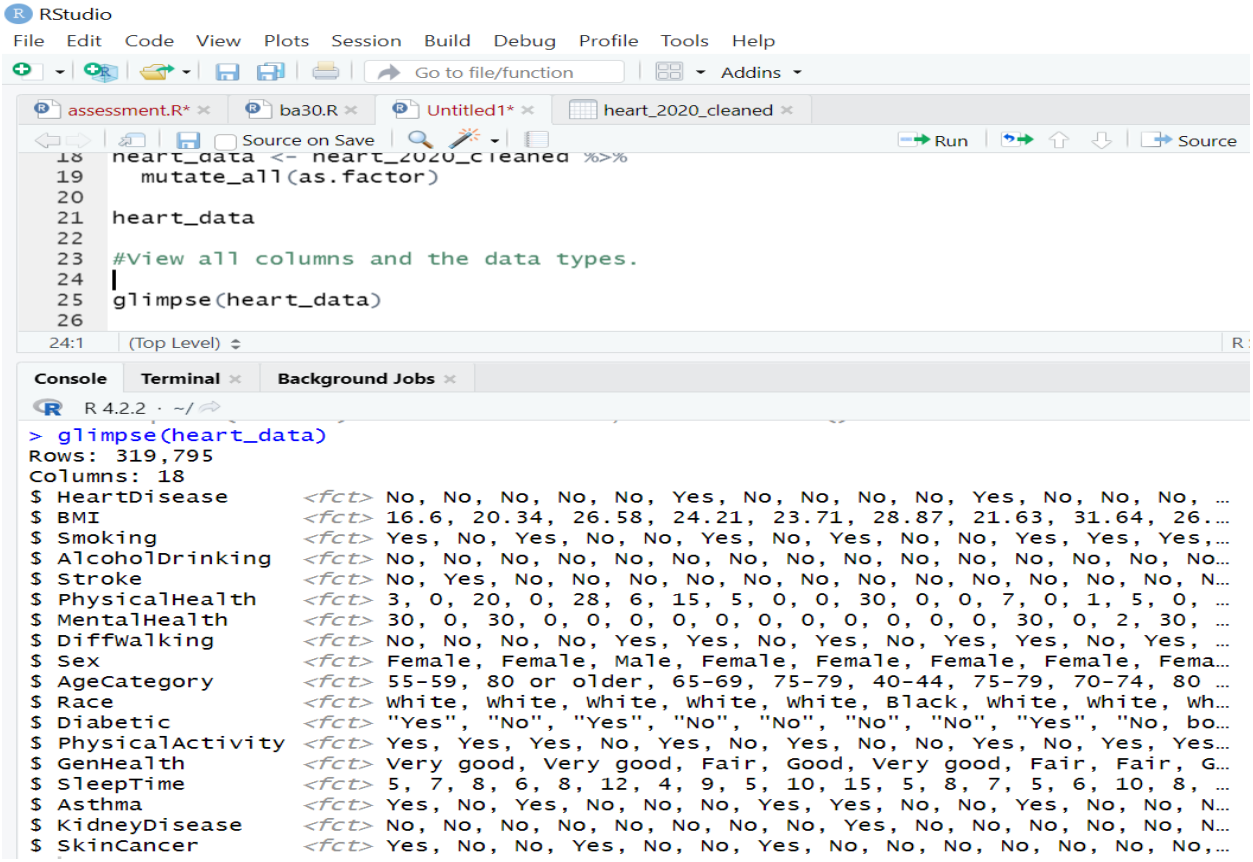


Data Preparation

We use parameter stringsAsFactors = TRUE so that all character columns will automatically be stored as factors.

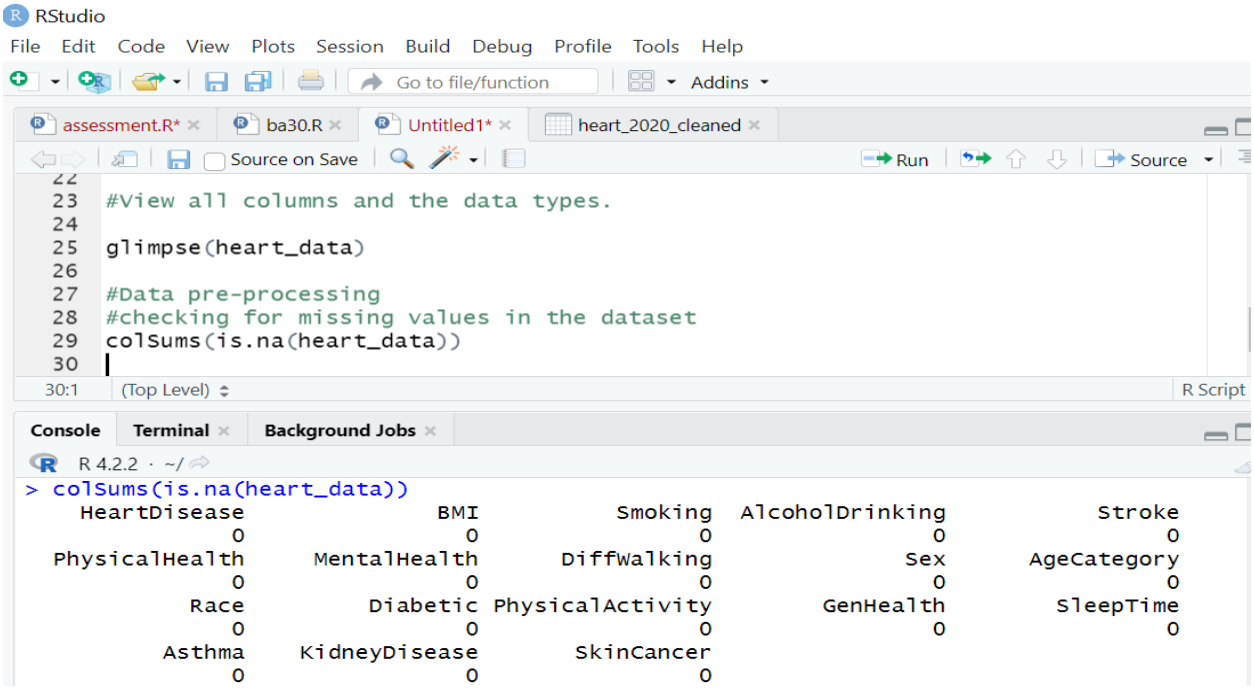


To view all variables and the data types



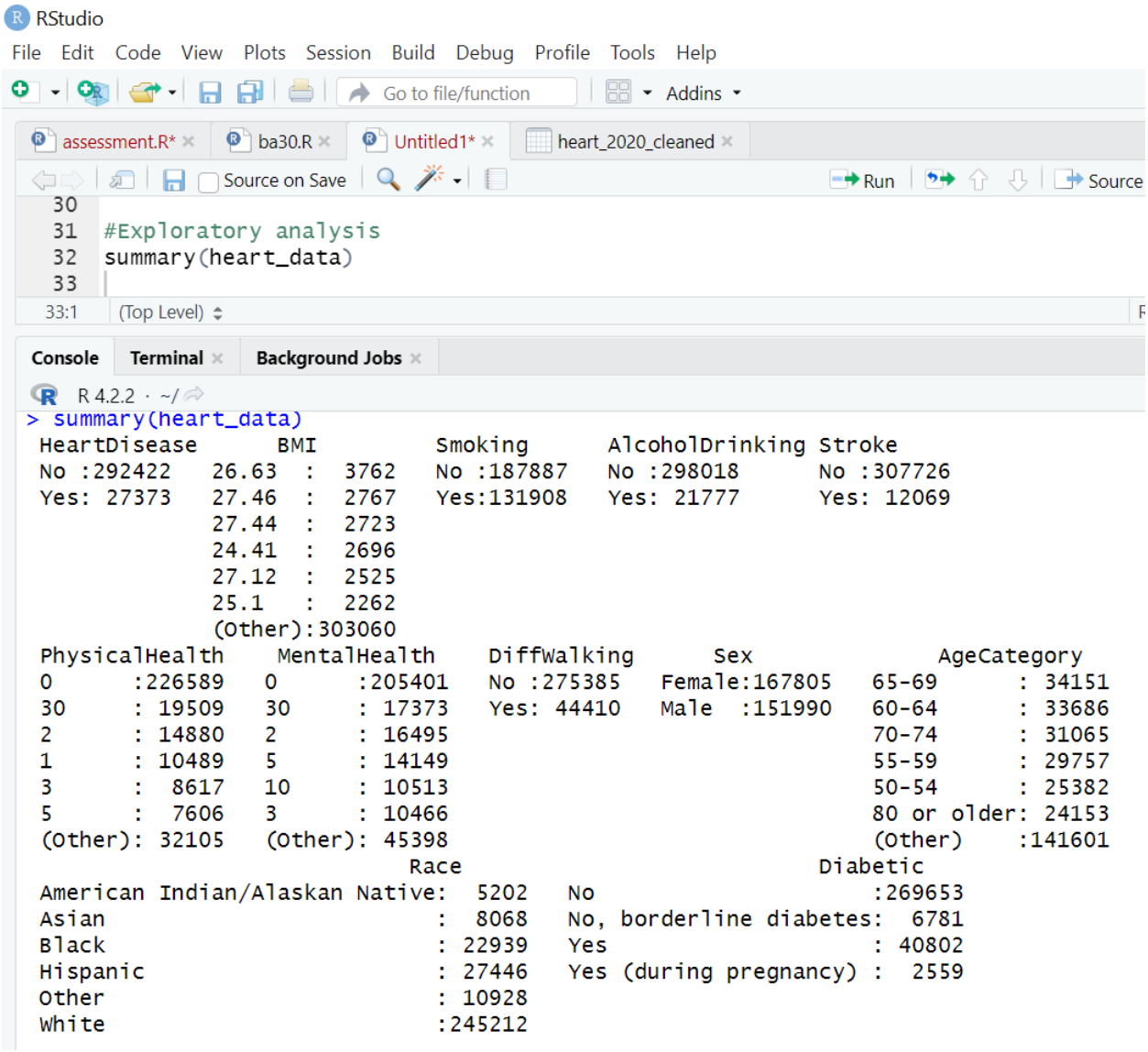
Pre-processing the Dataset

Checking for missing values. There are no missing values so we go ahead and explore the data.



Exploratory Analysis

Firstly, we would view the summary of all our variables.




```
PhysicalActivity     GenHealth        SleepTime     Asthma       KidneyDisease
No : 71838       Excellent: 66842   7      :97751   No :276923   No :308016
Yes:247957       Fair     : 34677   8      :97602   Yes: 42872   Yes: 11779
                 Good     : 93129   6      :66721
                 Poor     : 11289   5      :19184
                 Very good:113858   9      :16041
                                    10     : 7796
                                    (Other):14700
SkinCancer
No :289976
Yes: 29819
```

We must look at the distribution of each variable prior to conducting the study. Category-specific variables

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

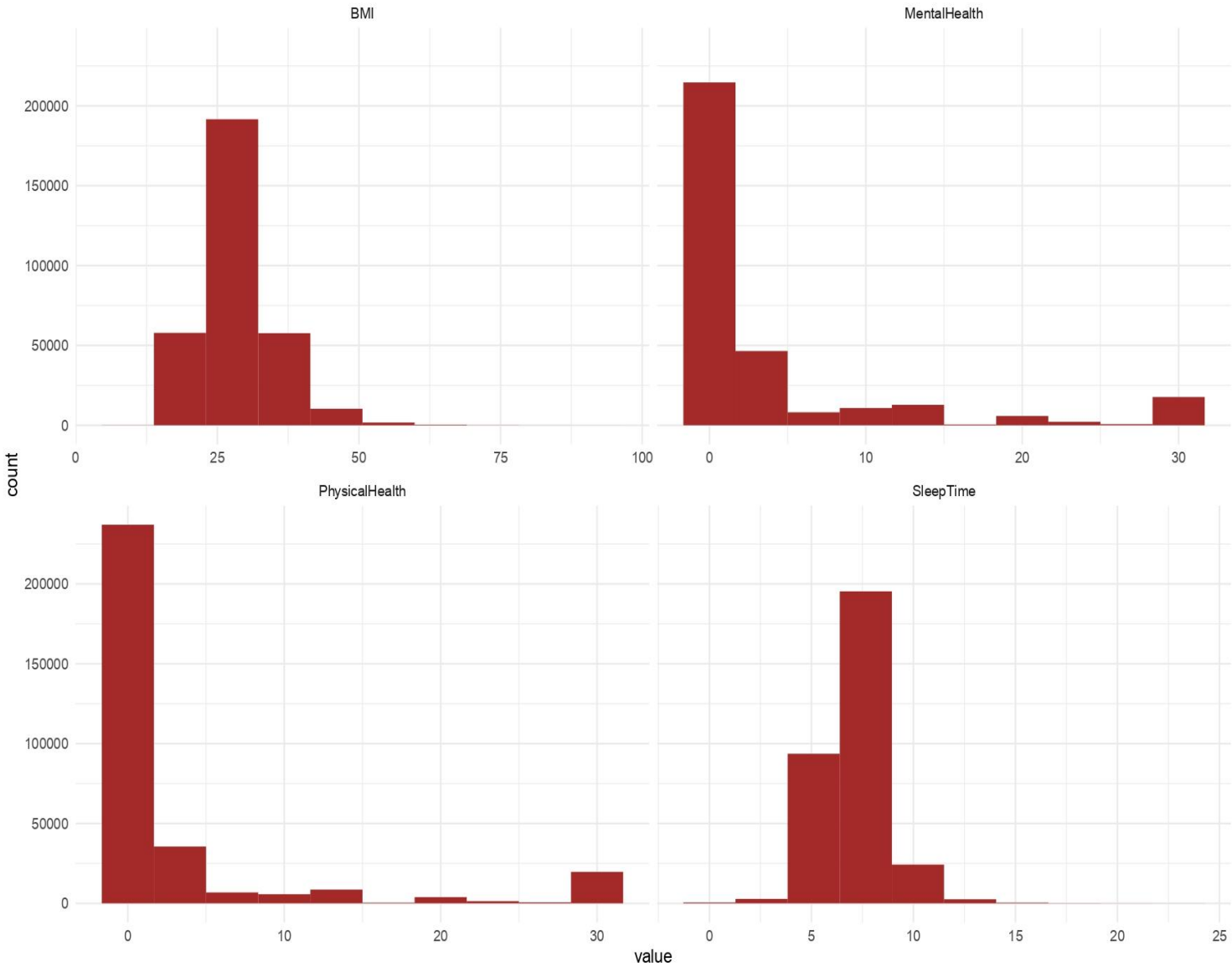
Go to file/function | Addins

assessment.R* | ba30.R | Untitled1* | heart_2020_cleaned

Source on Save | Run | Source

```
34  # Plot histograms for each categorical variable
35
36  heart_data_long <- pivot_longer(heart_data, cols = everything(), names_to = "variab
37  ggplot(heart_data_long, aes(x = value)) +
38    geom_bar(fill = "green") +
39    facet_wrap(~ variable, scales = 'free_x') +
40    theme_minimal()
```

Histograms for Numerical variables



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

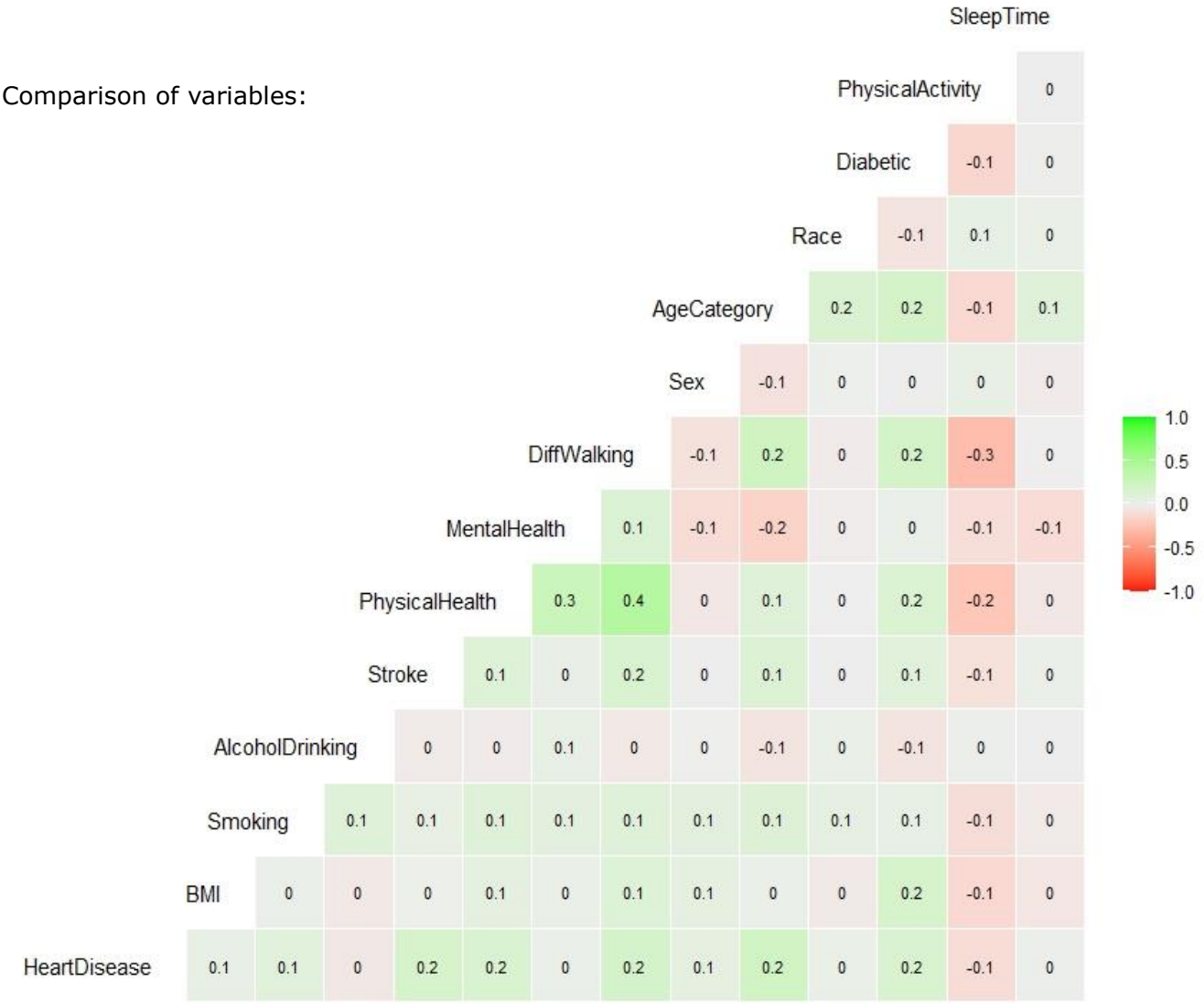
Go to file/function | Addins

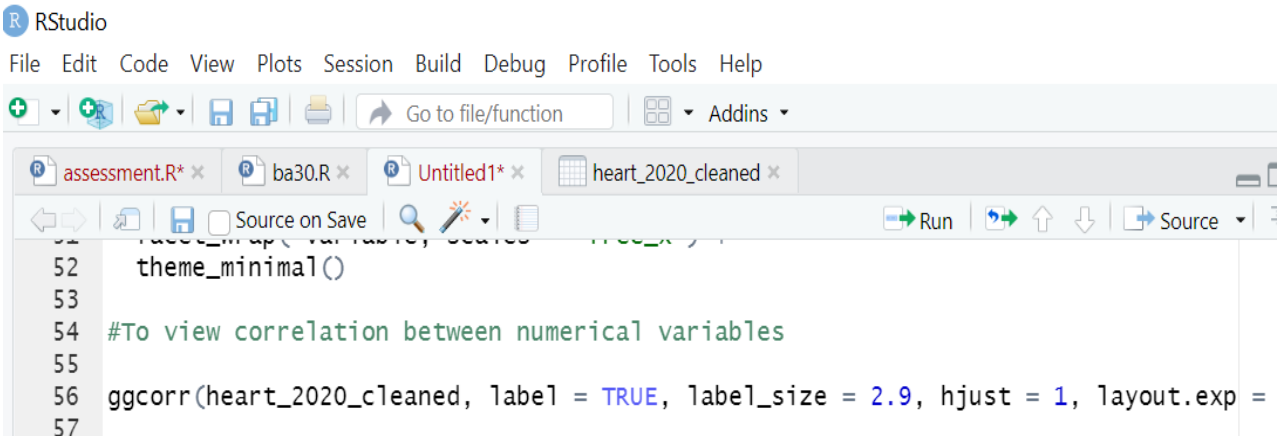
assessment.R* | ba30.R | Untitled1* | heart_2020_cleaned

Source on Save | Run | Source

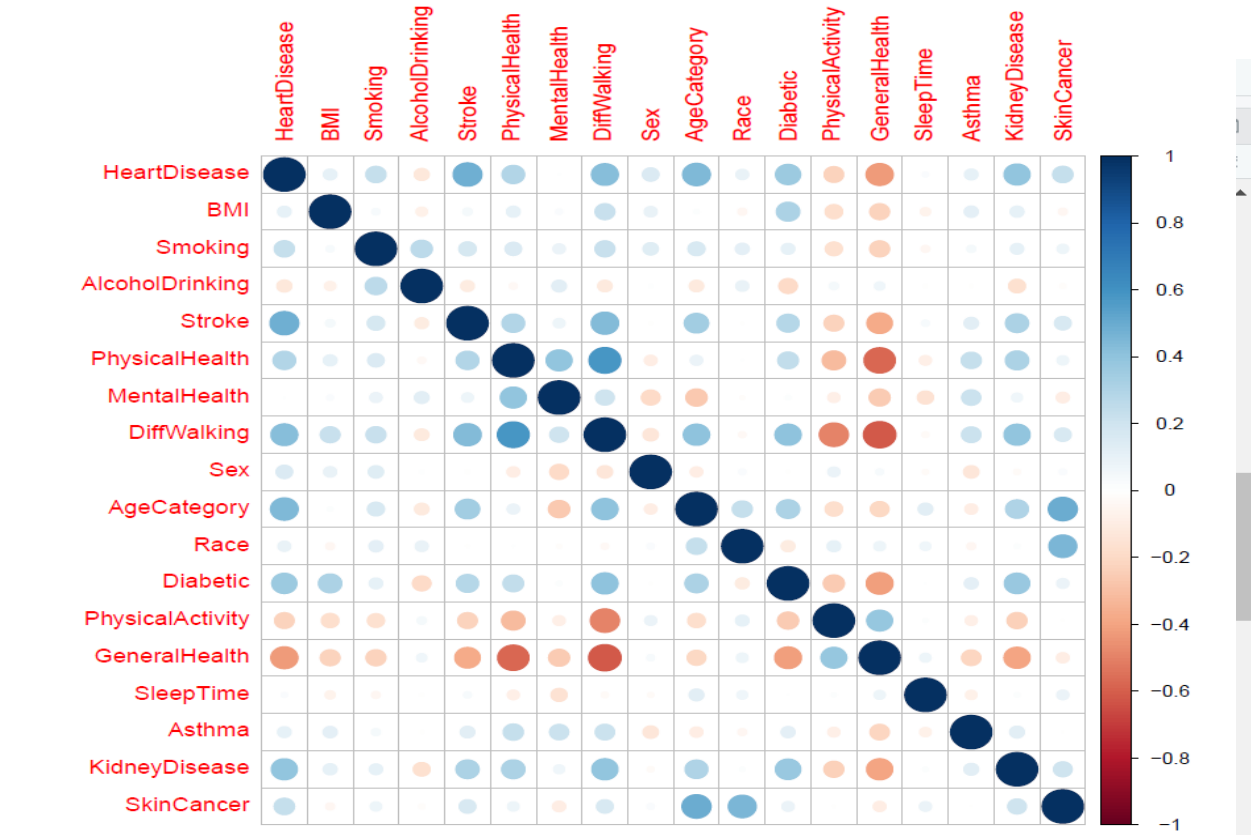
```
43  #histogram plot for numerical variables
44
45  heart_data_long <- heart_2020_cleaned %>%
46    select(BMI, PhysicalHealth, MentalHealth, SleepTime) %>%
47    pivot_longer(cols = everything(), names_to = "variable", values_to = "value")
48
49  ggplot(heart_data_long, aes(x = value)) +
50    geom_histogram(bins = 10, fill = "brown") +
51    facet_wrap(~variable, scales = 'free_x') +
52    theme_minimal()
```

Comparison of variables:



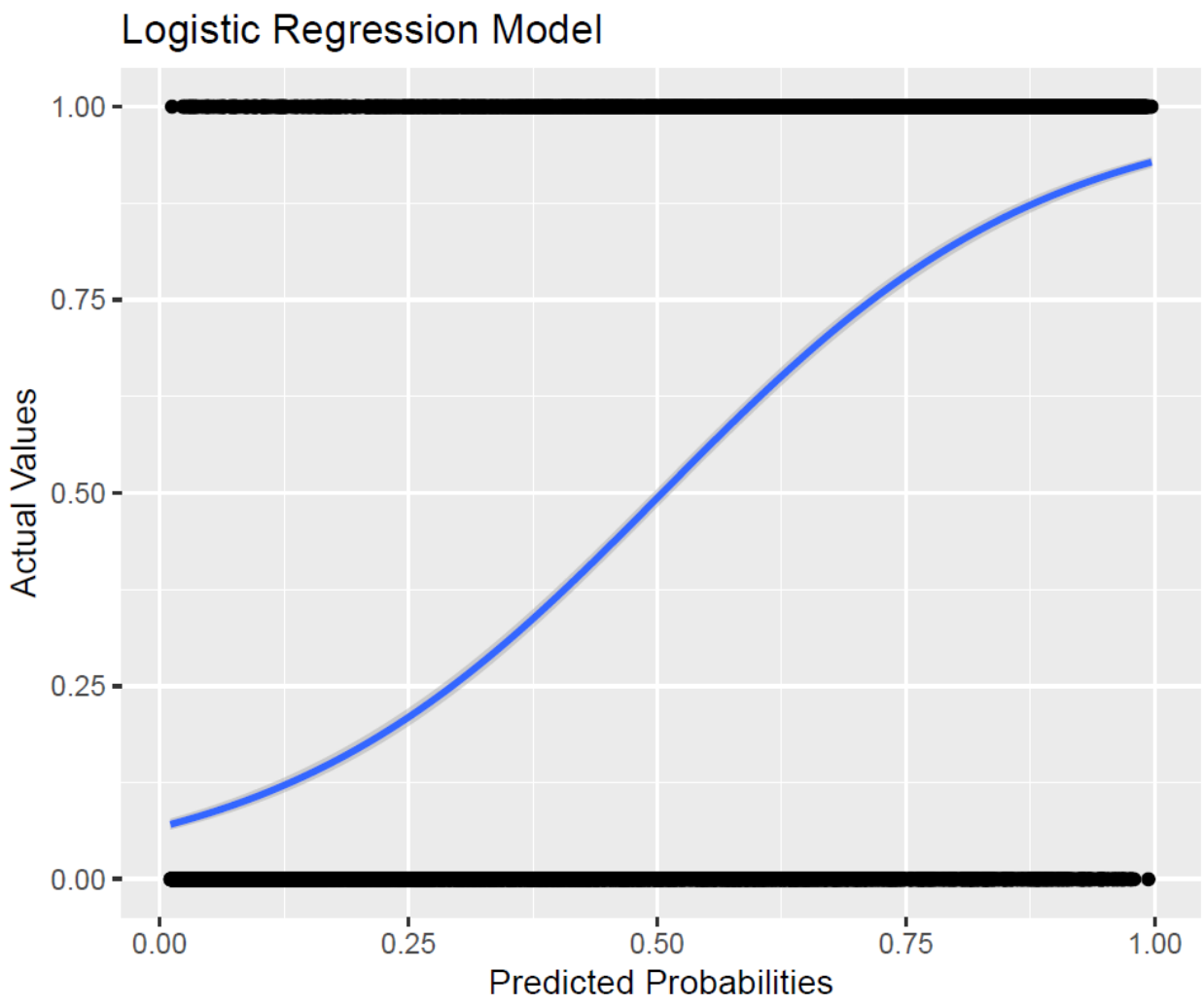


Correlation Matrix




```
83  shortlistedVars<-shortlistedVars[1:10] # get list of variables
84  shortlistedVars
85
```

Logistic Regression



RStudio

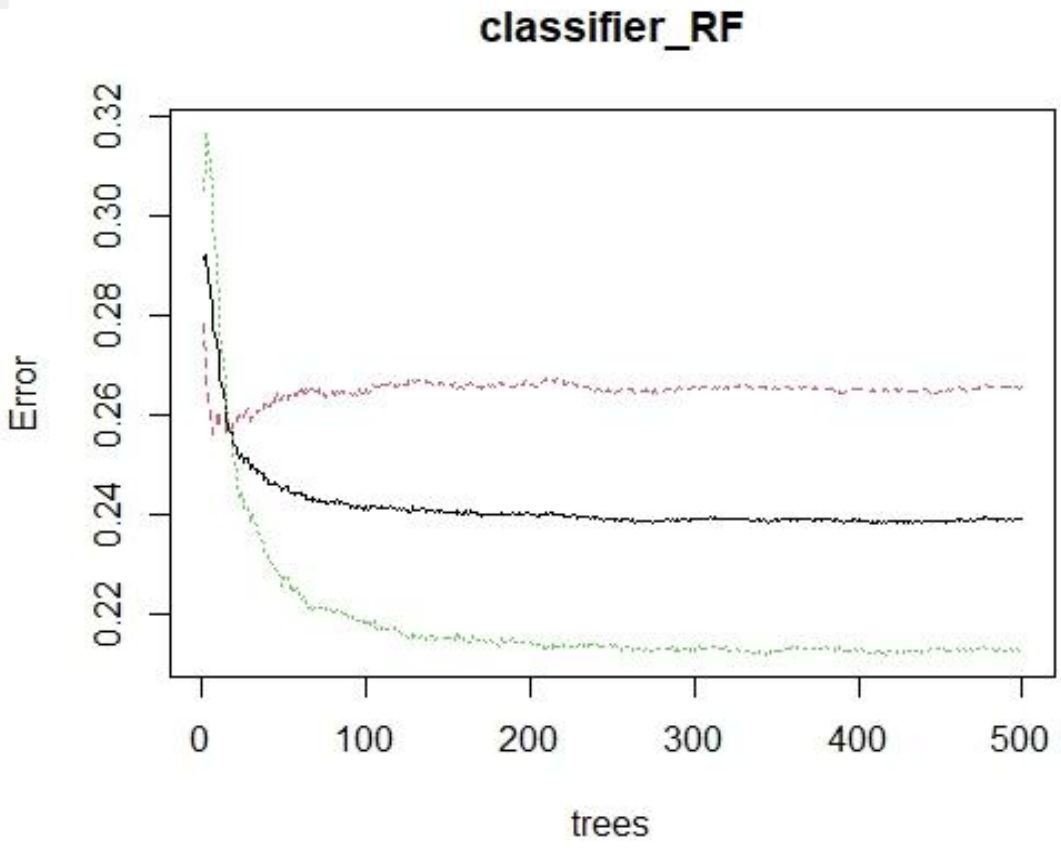
File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

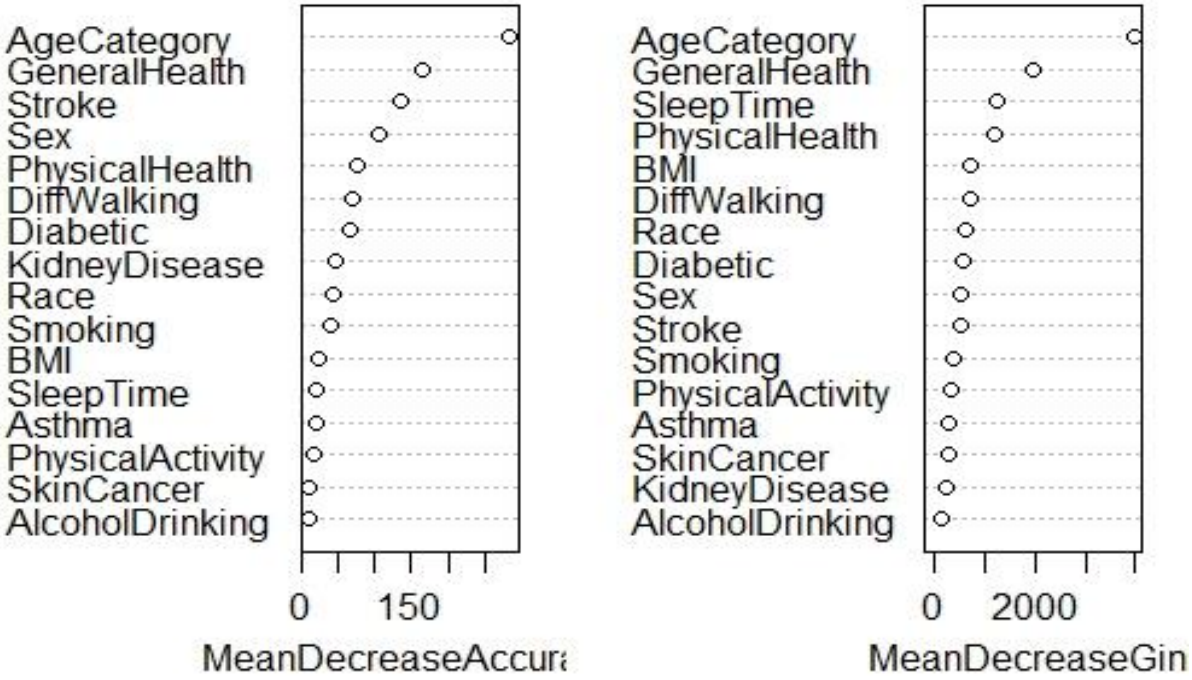
first week.R | data | Untitled1 | assessment.R | bag.R* | ba30.R* | assessment.R*

Source on Save | Run | Source

```
359 test.data <- balanced_sample[-training.samples, ]
360
361 # Logistics Regression
362 model  <- glm(HeartDisease ~. , data = train.data[c("HeartDisease",shortlistedVars
363 summary(model)
364
365 #Predicting probabilities
366 probabilities <- model %>% predict(test.data, type = "response")
367 head(probabilities)
368
369 summary(probabilities)
370
371 predicted.classes <- ifelse(probabilities > 0.5, "1", "0")#If the Probabilities is
372
373 confusion_mtx = table(test.data[, 1], predicted.classes)
374 confusionMatrix(confusion_mtx)
375
376 library(ggplot2)
377
378 # Create a data frame with the predicted probabilities and actual values
379 df <- data.frame(Probabilities = probabilities, Actual = test.data$HeartDisease)
380
381 # Create a scatterplot with the predicted probabilities on the x-axis and actual v
382 ggplot(df, aes(x = Probabilities, y = Actual)) +
383   geom_point() +
384   geom_smooth(method = "glm", method.args = list(family = "binomial")) +
385   ggtitle("Logistic Regression Model") +
386   xlab("Predicted Probabilities") +
387   ylab("Actual Values")
388
```



classifier_RF



Peer Evaluation Form for Group Work

Your Group name __BAG 30_____

Write the **Student ID** of each of your group members in a separate column. For each group member, indicate the extent to which you agree with the statement on the left, using a scale of 1-5:

1 = lowest score, 5 = highest score

1: Very poor: unacceptable performance
2: poor: less than acceptable performance
3: Average performance
4: Good performance
5: Excellent performance

Evaluation Criteria	**Group member 1: ID**	**Group member 2: ID**	**Group member 3: ID**	**Group member 4: ID**	**Comment**
	100633059	**100643247**	**100594631**	**100641893**	
	Grade: (1 - 5)	**Grade: (1 - 5)**	**Grade: (1 - 5)**	**Grade: (1 - 5)**	
•Did the individual contribute his/her fair share? •Contributes meaningfully to group discussions. •Contributes significantly to the success of the project. •Demonstrates a cooperative and supportive attitude. •Overall, how would you rank this person's contributions to the group?	**5**	**5**	**5**	**5**	
•Did the individual complete all work in a timely manner?	**5**	**5**	**5**	**5**	

•How would you rate the quality of individuals' work?	**5**	**5**	**5**	**5**	
•Did the individual maintain a positive, respectful attitude? •Attends group meetings regularly and arrives on time. •Would you want to work with this person again?	**5**	**5**	**5**	**5**	
TOTALS	**20**	**20**	**20**	**20**	**100**
PEMark = (TOTALS / 20)	**1**	**1**	**1**	**1**	**5**

PEMark: peer evaluation mark

** If a student **never** joined the group, put "*no participation*" in the corresponding column.

Overall mark for a student (CW1) = CW1 mark * peer evaluation mark