INDIVIDUAL ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

CT075-3-2 DTM

DATA MANAGEMENT (PART 1)

APD2F2211CS(DA)/APU2F2211CS(DA)

HAND OUT DATE: 20 DECEMBER 2022
HAND IN DATE: 6 MARCH 2023
WEIGHTAGE: 25%

---

1 **INSTRUCTIONS TO CANDIDATES:**
2 **Submit your assignment at the administrative counter**
3 **Student are advised to underpin their answer with the use of references (cited using the Harvard Name System of Referencing)**
4 **Late submission will be awarded zero (0) unless Extenuating Circumstances (EC) are upheld**
5 **Cases of plagiarism will be penalized**
6 **The assignment should be bound in an appropriate style (comb bound or stapled).**
7 **Where the assignment should be submitted in both hardcopy and softcopy, the softcopy of the written assignment and source code (where appropriate) should be on a CD in an envelope / CD cover and attached to the hardcopy.**
8 **You must obtain 50% overall to pass this module.**

Table of Contents

List of Tables

List of Figures

1 Introduction

The given dataset is provided by Framingham Heart Study, an organization dedicated to identifying the common factors that contribute to cardiovascular disease since 1948. It is stored in the SAS studio's SASHELP folder, SASHELP.HEART. This dataset has 5209 rows and 17 attributes. We will conduct Exploratory Data Analysis (EDA) on this dataset to have a further understanding of this dataset. EDA here includes a few different parts, including data category, descriptive statistics, data visualization, and analysis. Then, we will proceed to data pre-processing, removing outliers, missing data, incomplete data, and inconsistent data existing in the dataset.

2 Type of attributes

Alphabetic List of Variables and Attributes				
#	Variable	Type	Len	Label
12	AgeAtDeath	Num	8	Age at Death
5	AgeAtStart	Num	8	Age at Start
3	AgeCHDdiag	Num	8	Age CHD Diagnosed
15	BP_Status	Char	7	Blood Pressure Status
14	Chol_Status	Char	10	Cholesterol Status
13	Cholesterol	Num	8	
2	DeathCause	Char	26	Cause of Death
8	Diastolic	Num	8	
6	Height	Num	8	
10	MRW	Num	8	Metropolitan Relative Weight
4	Sex	Char	6	
11	Smoking	Num	8	
17	Smoking_Status	Char	17	Smoking Status
1	Status	Char	5	
9	Systolic	Num	8	
7	Weight	Num	8	
16	Weight_Status	Char	11	Weight Status

Table 1.0.2.1.1 dataset attributes

2.1 Nominal data

Nominal data is a type of qualitative data in which variables are labeled into categories. It cannot be measured or ordered, and it does not have equal spacing between values or a true zero value.

Attributes	Description	Remark
Status	Status of the individuals, alive or dead	With categories that could not be ranked
Death Cause	The reason for individual death.	
Sex	The gender of the individual	

Table 2.1.1 Nominal data table

2.2 Ordinal data

Ordinal data is a type of qualitative data that groups variables into ordered categories. Hence, the criterion for determining ordinal data is whether the features can be ordered, ranked, or have a hierarchical scale.

Attributes	Description	Remark
Chol_Status	Cholesterol status is classified using cholesterol content measured, Borderline, High or Desirable.	Can be ordered and ranked
BP_Status	Blood Pressure status of individuals, classified using high, normal, or optimal	
Weight_Status	The weight status of the individual, classified using normal, overweight, or underweight	
Smoking_Status	The addiction level of individuals, classified using light, moderate, non-smokers, or very heavy	

Table 2.2.1 Ordinal data table

2.3 Ratio data

Ratio data is the most intricate type of data which helps deal with the broadest array of analyses. It is a quantitative data type that measures variables as continuous values with equidistant adjacent values. In other words, for the ratio data, the relationship between data can be expressed as a fraction. Therefore, it has a "true zero" and does not accept negative values. The scale can have either an infinite range of values or a finite endpoint.

Attributes	Description	Remark
AgeCHDdiag	Age Coronary Heart Disease diagnosed is the age at which the disease is diagnosed.	Not allow negative values
AgeAtStart	The age of the respondent started having unknown conditions.	
AgeAtDeath	The age of passing away for the individual.	
Height	Body height of the individual	
Weight	Body weight of the individual	
Smoking	Smoking level of respondents. The higher the marks, the heavier the individual’s smoking habit	

Table 2.3.1 Ratio data table

2.4 Interval data

Interval data is a quantitative data type that is represented along a scale, with equal distance between any two points. It is usually expressed in numerical values, where the gap between any two points on the scale is consistent. For example, since the difference between any two points on the scale of temperature degrees is the same, the temperature degree is interval data.

Attributes	Description	Remark
MRW	Metropolitan Relative Weight is similar to BMI (Body Mass Index) and is used to observe body fatness in clinical studies. The higher marks symbolize that the fatter the individual is.	The gap between the two values is the same
Systolic	The blood pressure when an individual's heart beats are measured by mmHg. The higher values indicate that the individual is suffering from high blood pressure.	
Diastolic	The blood pressure data that is measured within the timeframe of two continuous heartbeats.	
Cholesterol	Cholesterol level in blood measured by mg/dL. The higher the value means the more cholesterol within the particular volume of an individual’s blood.	

Table 2.4.1 Interval data table

3 Exploratory Data Analysis

3.1 Descriptive Statistics

Variable	Label	N	N Miss	Minimum	Mean	Median	Maximum	Std Dev	Mode	Lower Quartile	Upper Quartile	Variance	Quartile Range	Skewness
AgeCHDdiag	Age CHD Diagnosed	1449	3760	32.0000000	63.3029676	63.0000000	90.0000000	10.0182154	59.0000000	57.0000000	70.0000000	100.3646390	13.0000000	-0.0099678
AgeAtStart	Age at Start	5209	0	28.0000000	44.0687272	43.0000000	62.0000000	8.5749541	36.0000000	37.0000000	51.0000000	73.5298379	14.0000000	0.1947589
Height		5203	6	51.5000000	64.8131847	64.5000000	76.5000000	3.5827074	62.5000000	62.2500000	67.5000000	12.8357921	5.2500000	0.1774198
Weight		5203	6	67.0000000	153.0866808	150.0000000	300.0000000	28.9154261	138.0000000	132.0000000	172.0000000	836.1018660	40.0000000	0.5559412
Diastolic		5209	0	50.0000000	85.3586101	84.0000000	160.0000000	12.9730913	80.0000000	76.0000000	92.0000000	168.3010976	16.0000000	0.8759436
Systolic		5209	0	82.0000000	136.9095796	132.0000000	300.0000000	23.7395964	120.0000000	120.0000000	148.0000000	563.5684355	28.0000000	1.4875765
MRW	Metropolitan Relative Weight	5203	6	67.0000000	119.9575245	118.0000000	268.0000000	19.9834015	113.0000000	106.0000000	131.0000000	399.3363347	25.0000000	1.1323756
Smoking		5173	36	0	9.3665185	1.0000000	60.0000000	12.0314511	0	0	20.0000000	144.7558161	20.0000000	1.2224041
AgeAtDeath	Age at Death	1991	3218	36.0000000	70.5364139	71.0000000	93.0000000	10.5594062	68.0000000	63.0000000	79.0000000	111.5010603	16.0000000	-0.3160539
Cholesterol		5057	152	96.0000000	227.4174412	223.0000000	568.0000000	44.9355238	200.0000000	196.0000000	255.0000000	2019.20	59.0000000	0.8163442

Table 3.1.1 descriptive statistics for numeric data

The data in the AgeCHDdiag column is the most missing with 72% of the data being unavailable, while only 6 rows are missing in the height, weight, and MRW columns. Smoking data is missing in 36 rows, and there is a 62% absence of data in the AgeAtDeath column, with 3218 missing rows. Additionally, 152 data are missing in the Cholesterol column.

Apart from that, the minimum height of the dataset is 51.5, but the minimum age of respondents in the dataset is 28 years old. The dataset given is from SAS, an American enterprise, hence the reasonable unit for height column should be an inch since 51.5 inches is around 130.81 cm. A similar situation is also applicable to the weight column, the mean weight is 153.1, so the unit should be in the pound as 153.1 pounds is about 69.4 KG which is more reasonable.

Besides, the mode of the smoking marks column is 0, which indicates that the samples in this data space are mostly not smokers.

Notably, the weight and cholesterol columns have the highest quartile range among all columns, recorded as 40 and 59 respectively, which means that the gap among data in the columns is considerably big.

Frequencies for Categorical Variables

Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Alive	3218	61.78	3218	61.78
Dead	1991	38.22	5209	100.00

Table 3.1.2 frequency of the status column

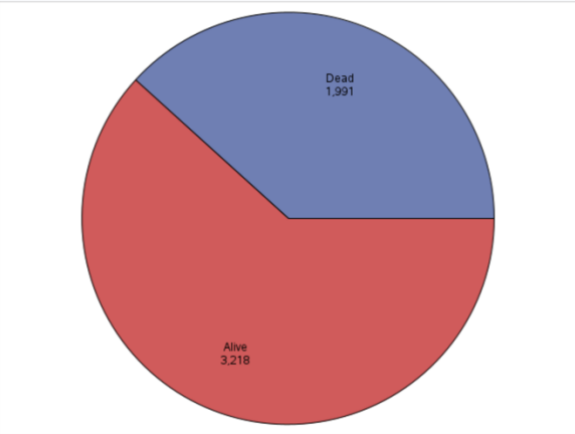

Figure 3.1.1 Status pie chart

This dataset contains information on 5209 individuals, with 3218 (62%) being recorded as alive and 1991 (38%) being recorded as dead.

Cause of Death				
DeathCause	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Cancer	539	27.07	539	27.07
Cerebral Vascular Disease	378	18.99	917	46.06
Coronary Heart Disease	605	30.39	1522	76.44
Other	357	17.93	1879	94.37
Unknown	112	5.63	1991	100.00
Frequency Missing = 3218				

Table 3.1.3 frequency of the cause of death

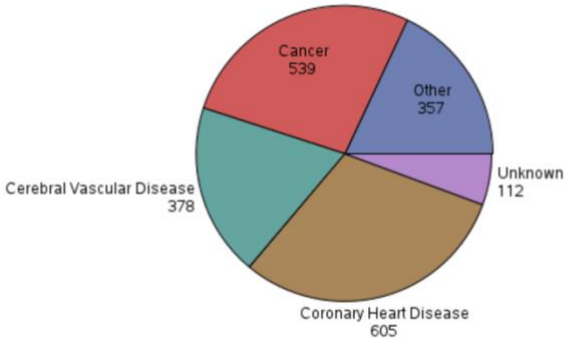

Figure 3.1.2 Death cause pie chart

The cause of death for the 38% of samples who have passed away. 27% or 539 of them died of cancer, and 19% or 378 of them died of Cerebral Vascular Diseases. Following that, coronary heart disease caused 30% or 605 of them to die. People died of other diseases and unknown diseases contain 18% (357 people) and 5.6% (112 people) respectively. The 3218 rows of missing data here mean that those people are still alive.

Sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	2873	55.15	2873	55.15
Male	2336	44.85	5209	100.00

Table 3.1.4 frequency of the sex column

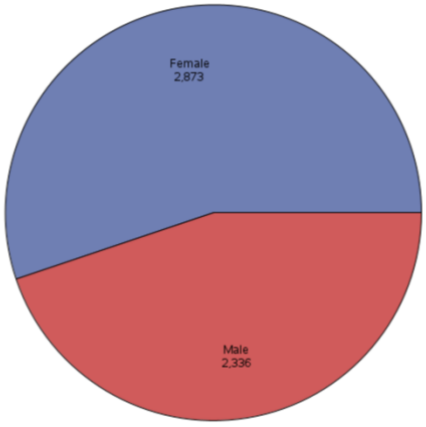

Figure 3.1.3 Sex pie chart

This dataset has information on 5209 people, with 2873 of them being female and 2336 being male. The proportion of female respondents is 55% while the proportion of male respondents is 45%.

Cholesterol Status				
Chol_Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Borderline	1861	36.80	1861	36.80
Desirable	1405	27.78	3266	64.58
High	1791	35.42	5057	100.00
Frequency Missing = 152				

Table 3.1.5 frequency of the cholesterol status column

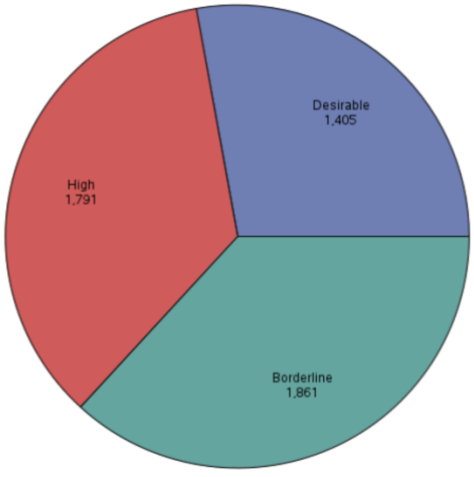

Figure 3.1.4 Cholesterol status pie chart

36.8% (1861) of them have considerably low cholesterol levels of the respondents in the dataset. 27.8% (1405) people accounted that they have a healthy cholesterol status, while 35.4% (1791) people in the given dataset are suffering from high cholesterol. For the missing data, only 152 rows are found empty in this column.

Blood Pressure Status				
BP_Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
High	2267	43.52	2267	43.52
Normal	2143	41.14	4410	84.66
Optimal	799	15.34	5209	100.00

Table 3.1.6 frequency of the blood pressure status column

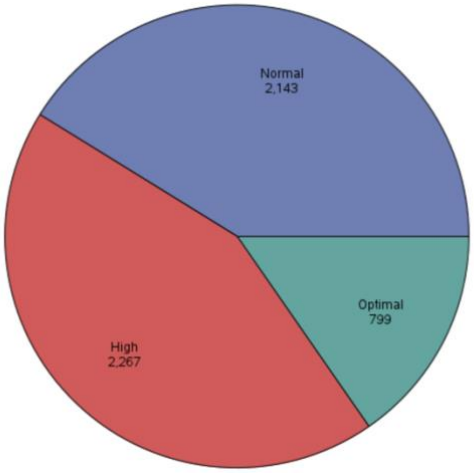

Figure 3.1.5 Blood Pressure status pie chart

The table above refers to the blood pressure status of individuals in this dataset. The results show that 43.5% of the people, or 2267 individuals, have high blood pressure. Meanwhile, 41.1% or 2143 people have a normal blood pressure status. Only 15.3% of the individuals, or 799 people, have optimal blood pressure levels.

Weight Status				
Weight_Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Normal	1472	28.29	1472	28.29
Overweight	3550	68.23	5022	96.52
Underweight	181	3.48	5203	100.00
Frequency Missing = 6				

Table 3.1.7 frequency of the weight status column

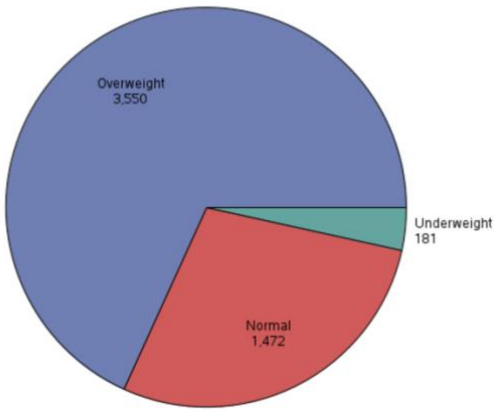

Figure 3.1.6 Weight Status pie chart

Aside from the 6-missing data, 28.3% (3550) of the individuals are considered to be of normal weight. On the other hand, the majority (68.2%, 3550) of the people in the dataset are classified as overweight. In comparison, only a small portion of the population, 3.5% (181 people), are considered underweight.

Smoking Status				
Smoking_Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Heavy (16-25)	1046	20.22	1046	20.22
Light (1-5)	579	11.19	1625	31.41
Moderate (6-15)	576	11.13	2201	42.55
Non-smoker	2501	48.35	4702	90.90
Very Heavy (> 25)	471	9.10	5173	100.00
Frequency Missing = 36				

Table 3.1.8 frequency of the smoking status column

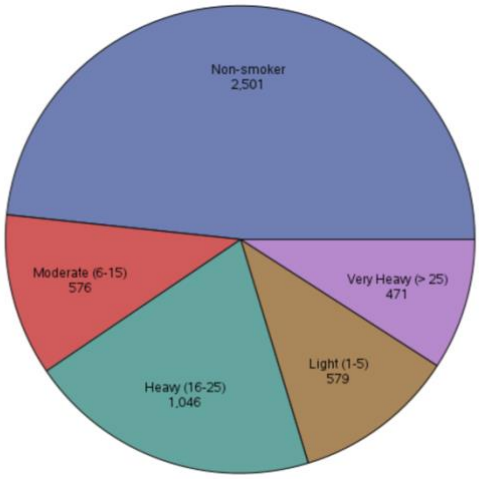

Figure 3.1.7 Smoking status pie chart

The smoking status of individuals in this dataset is classified using the smoking marks of individuals. Approximately one-fifth (1046) of them are heavy smoking while 579 of them are light smokers. 11% (576) of them get 6 to 15 marks, which means that they are moderate smokers. Non-smoker in this dataset constitutes a large portion of the dataset, recording 48.4% (2501) people. There are also 9% (471) of people who are heavy smokers. Missing data for this column recorded 36 rows.

3.2 Data Visualization

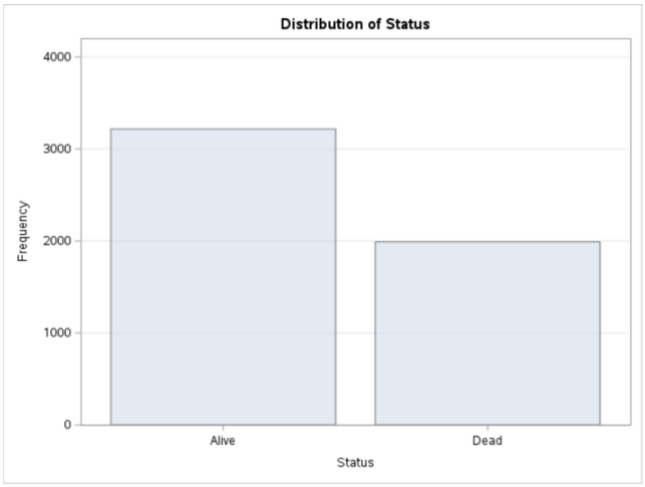

Figure 3.2.1 Distribution of Status

The sample encompassed in the dataset has more than 3000 alive individuals, 1000 more than the death cases.

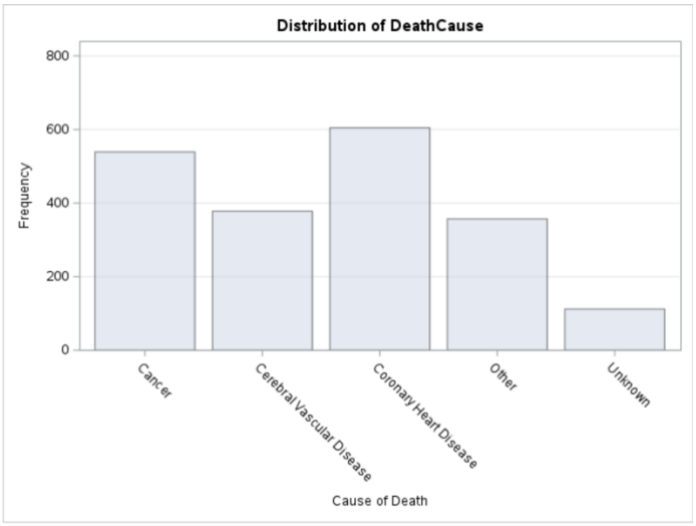

Figure 3.2.2 Distribution of cause of death

The dataset records the causes of death for the individuals in the sample, with coronary heart disease being the most common cause of death, accounting for 600 cases. Cancer and cerebral vascular disease were the second and third most common causes of death, with approximately 550 and 390 cases respectively. There were also around 400 cases of deaths due to other diseases, and 100 cases of deaths due to unknown causes.

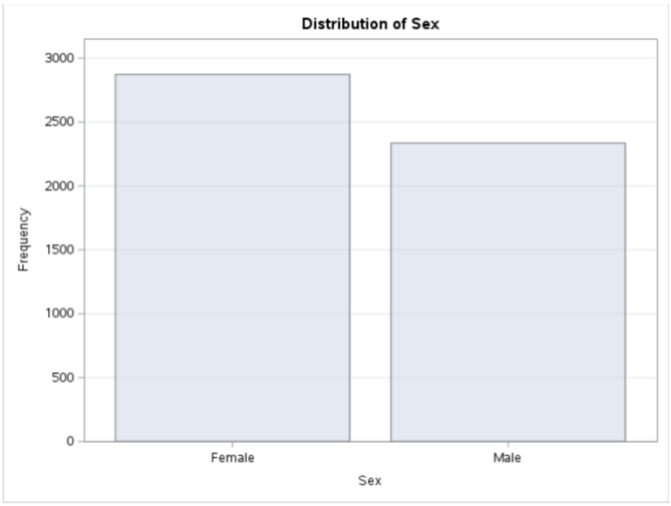

Figure 3.2.3 Distribution of sex

In this dataset, there is nearly an equal number of male and female cases, with around 2400 male cases and 2900 female cases.

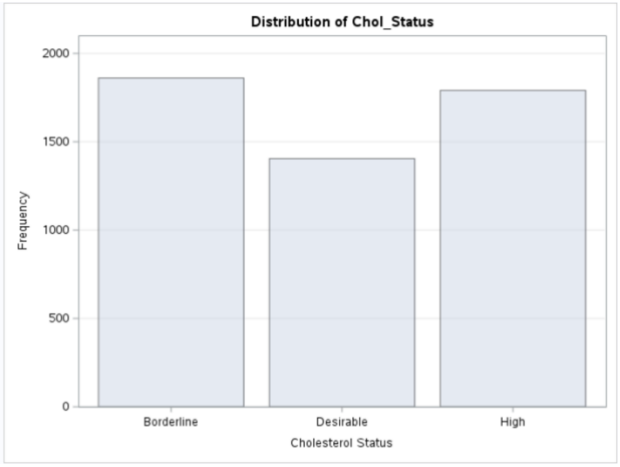

Figure 3.2.4 Distribution of cholesterol status

For the cholesterol status of individuals, most of them are either at the borderline or high cholesterol, accounting for around 1800 people in the given dataset. People with desirable cholesterol level is comparatively less and only has approximately 1400 cases.

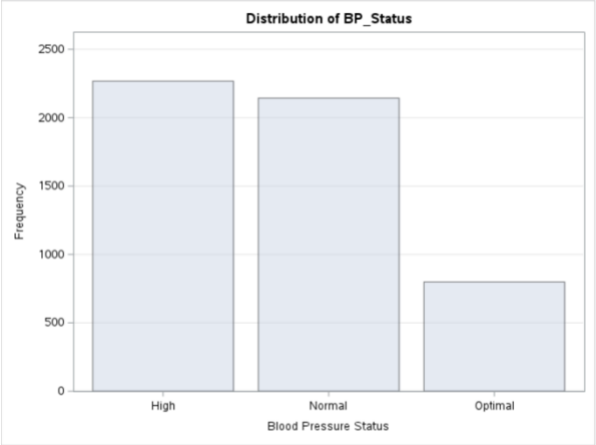

Figure 3.2.5 Distribution of blood pressure status

There are approximately 2300 individuals in this dataset who appear to be experiencing high blood pressure, while an almost equal number of individuals have normal blood pressure. Additionally, 900 individuals in the dataset have been recorded as having optimal blood pressure.

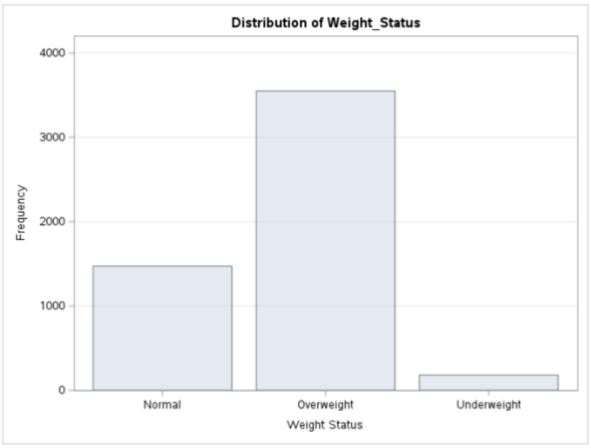

Figure 3.2.6 Distribution of weight status

Based on the distribution graph shown, it can be observed that the majority of the samples in the dataset are classified as overweight, accounting for around 3500 individuals. The second largest group comprises individuals with normal weight, numbering around 1500. In contrast, only a small number of individuals are categorized as underweight.

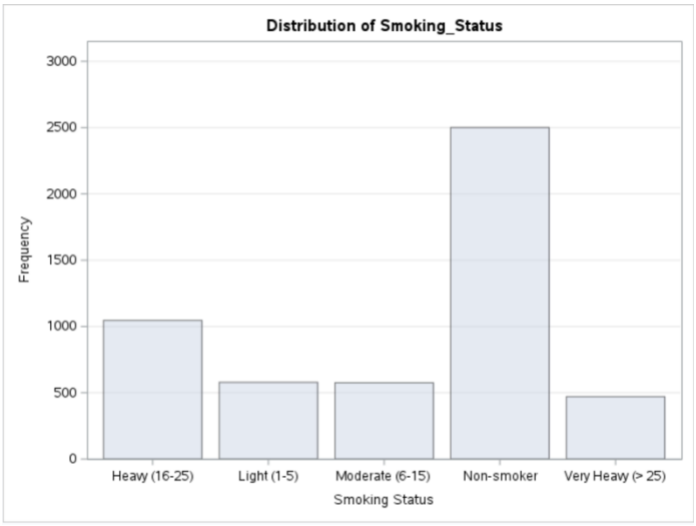

Figure 3.2.7 Distribution of smoking status

According to the descriptive statistics table, non-smokers make up the largest proportion of individuals in this dataset, with around 2500 people reporting that they do not smoke. The next largest group is heavy smokers, with approximately 1000 people reporting that they smoke a significant amount. The remaining individuals in the dataset are categorized as light smokers, moderate smokers, or very heavy smokers, with each group comprising around 500 people.

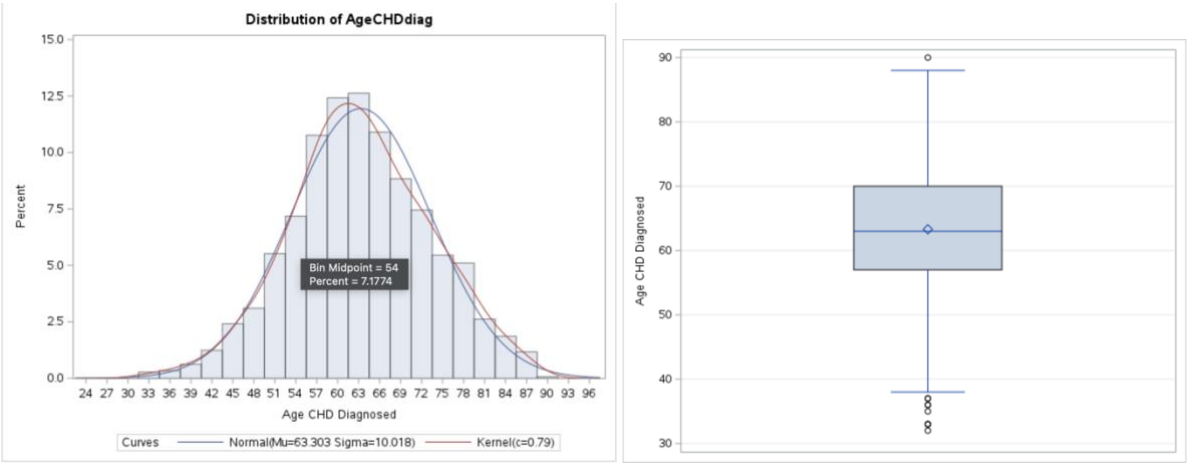

Figure 3.2.8 Distribution and boxplot of AgeCHDdiag

The mean of the age CHD diagnosed from the boxplot graph is around 65 years old and most of the people found are when 57 to 69 years old. The ideal distribution (blue curve) and the actual distribution (red curve) are fairly similar. Hence the distribution analysis will be more accurate.

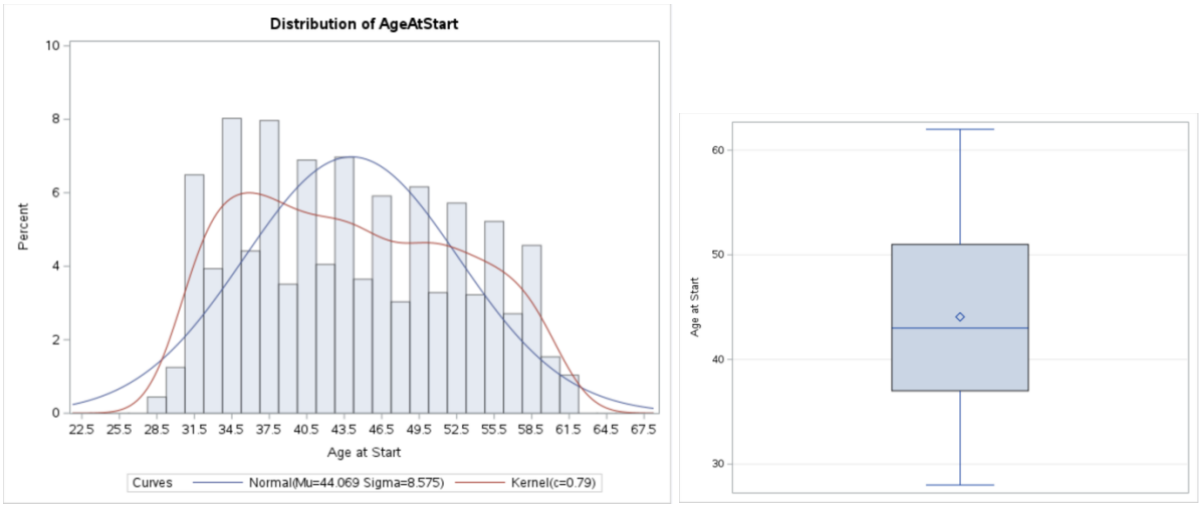

Figure 3.2.9 Distribution and boxplot of AgeAtStart

The frequency of age at the start found in the dataset fluctuated between 31.5 to 58.5 years old with an average of 44 years old. However, the peak of the theoretical distribution is higher than the actual distribution, which means that the dataset given is not enough representative and balanced for the actual situation, and potential biases might exist in the dataset.

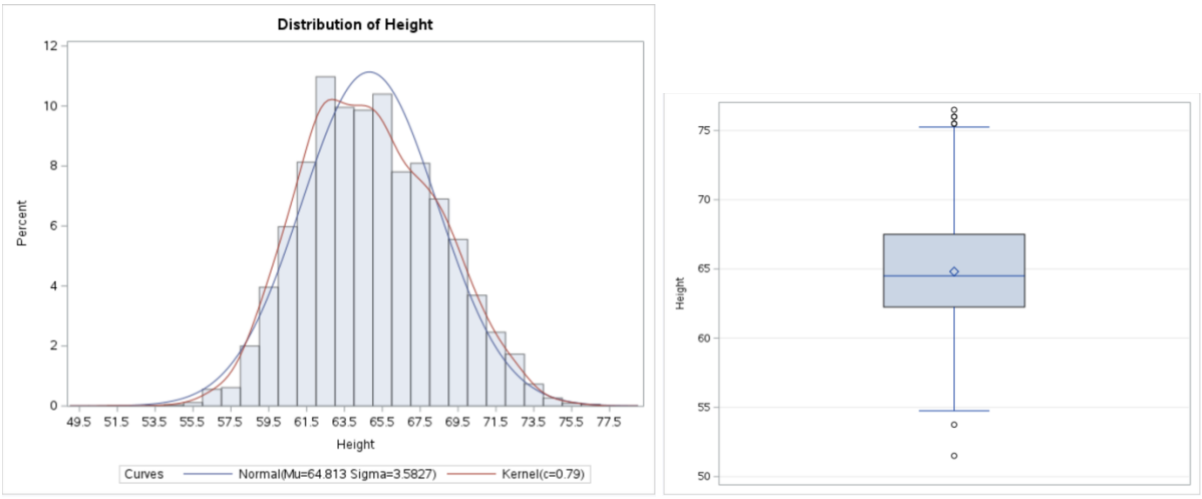

Figure 3.2.10 Distribution and boxplot of height

The average height in the dataset is 65 inches. Overall, the empirical distribution of height is nearly the same as the theoretical curve given except at the highest point. So, the analysis related to the height between 63.5 to 67.5 inches might be not that accurate and reliable.

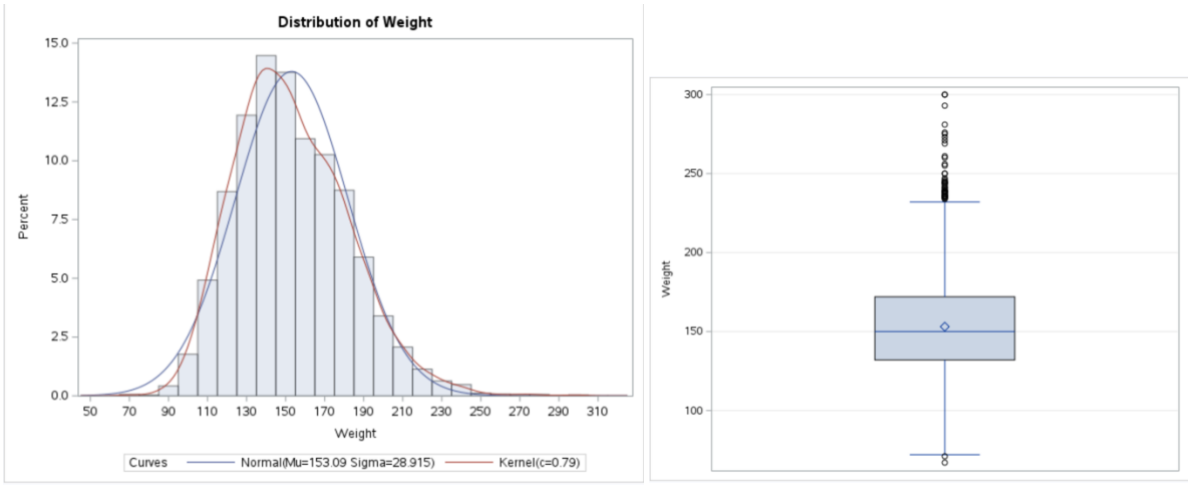

Figure 3.2.11Distribution and boxplot of weight

The midpoint of the weight values is about 150 pounds. Meanwhile, there are quite a lot of outliers that are above the 3rd quartile. The bell curve in the distribution graph demonstrates that the kurtosis between theoretical and actual distribution has little difference and their peak are staggered.

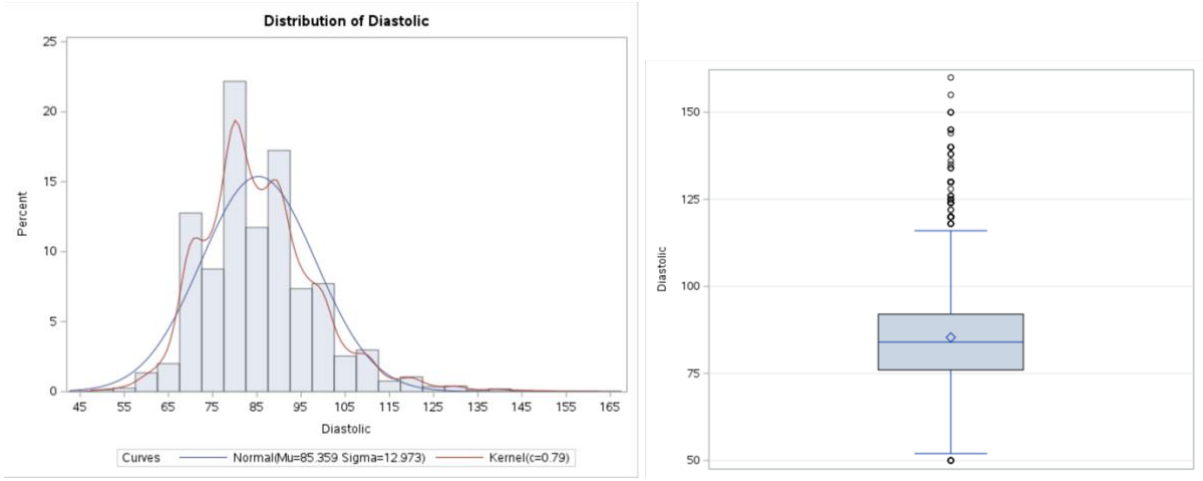

Figure 3.2.12 Distribution and boxplot of Diastolic

The distribution of an individual's diastolic fluctuated starting from 70, hence the density curve fluctuated as well. However, it fluctuated higher than the normal distribution expected and some result that comes up by using the diastolic data higher than 70 might not reliable.

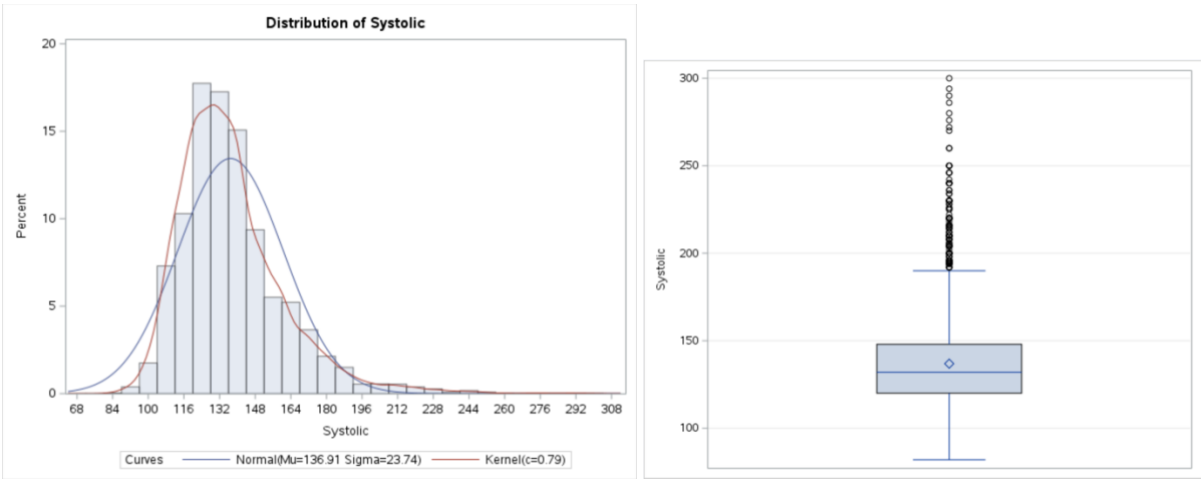

Figure 3.2.13 Distribution and boxplot of Systolic

In terms of systolic distribution, it has the same situation as the diastolic distribution as the peak of the density curve is higher than the expected distribution of systolic between 124 and 140. From the boxplot, the mean of systolic drops at 140 while a row of outliers occurs between values of 190 to 300.

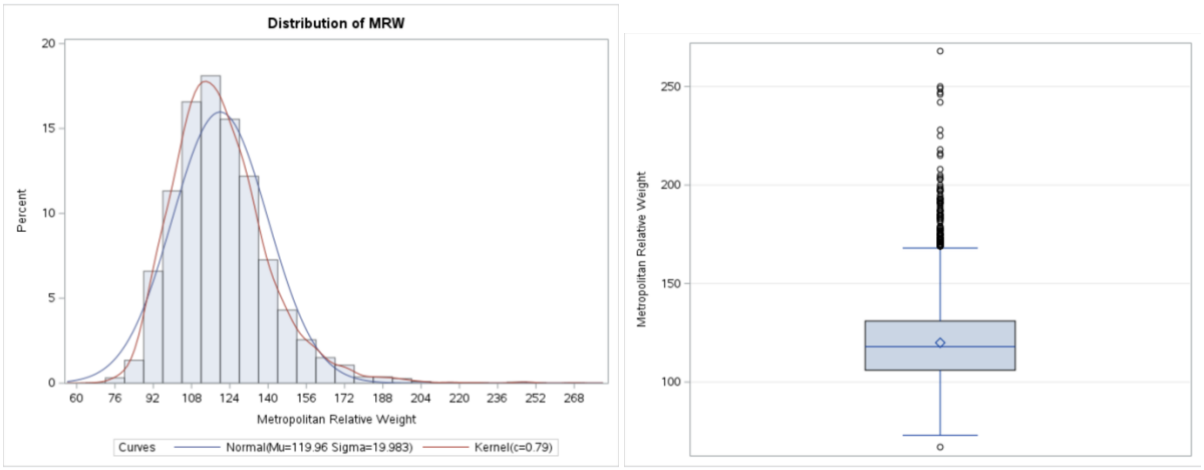

Figure 3.2.14 Distribution and boxplot of MRW

Based on the above graph, the distribution of the MRW data in this dataset is right-skewed. The peak of the theoretic curve is lower than the expectation. For most individuals, their MRW resides between 108 to 124. From the boxplot, there are quite a lot of outliers occurring above the upper fence. The mean of MRW falls at around 120.

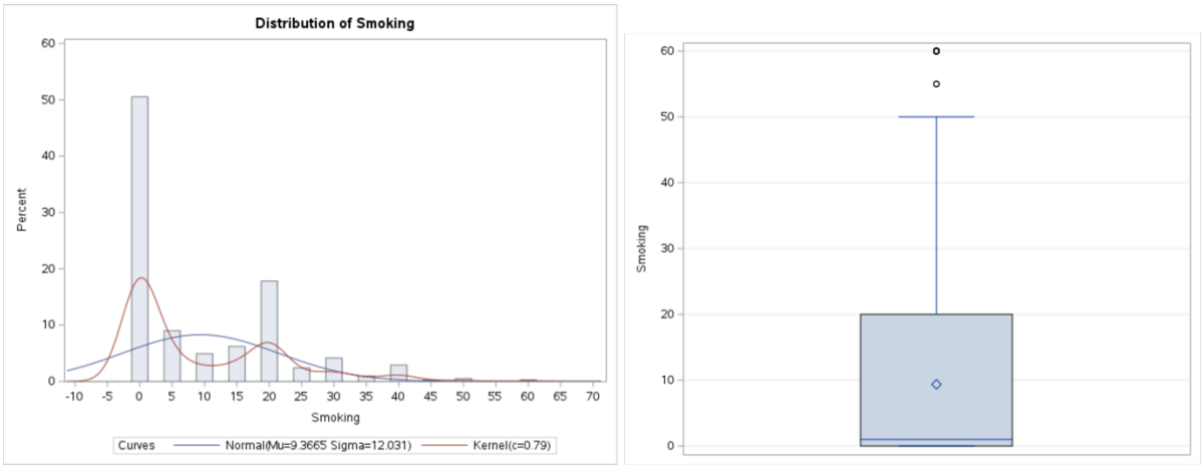

Figure 3.2.15 Distribution and boxplot of smoking

The distribution of smoking marks is not balanced and most of the samples in this dataset are non-smokers. Hence, the distribution graph is right-skewed, and the quartile of data resides at the bottom side.

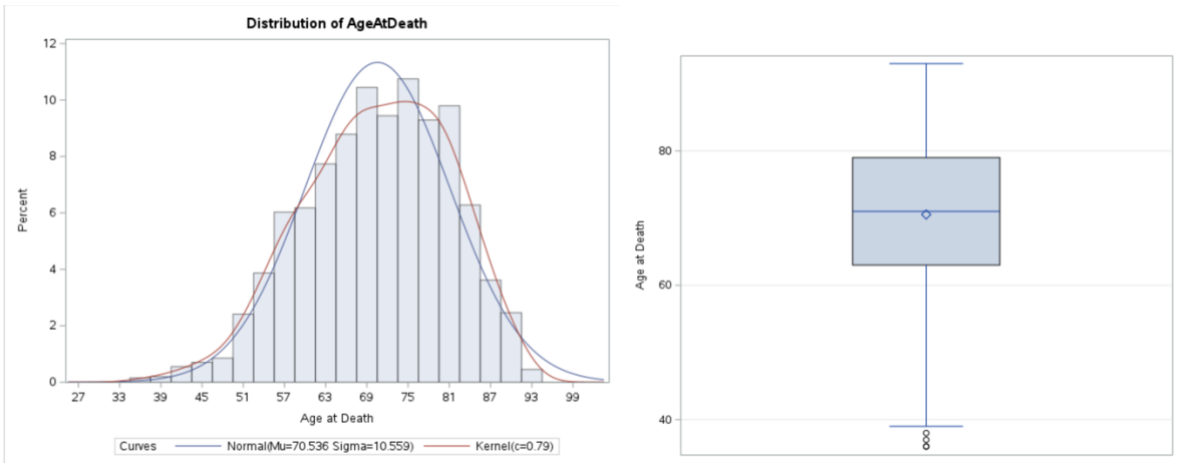

Figure 3.2.16 Distribution and boxplot of AgeAtDeath

For the distribution of AgeAtDeath data, it is slightly left-skewed, and the actual peak is lower than the theoretical peak. The average age of death is around 70. Meanwhile, the outlier of this column only occurs below the lower fence.

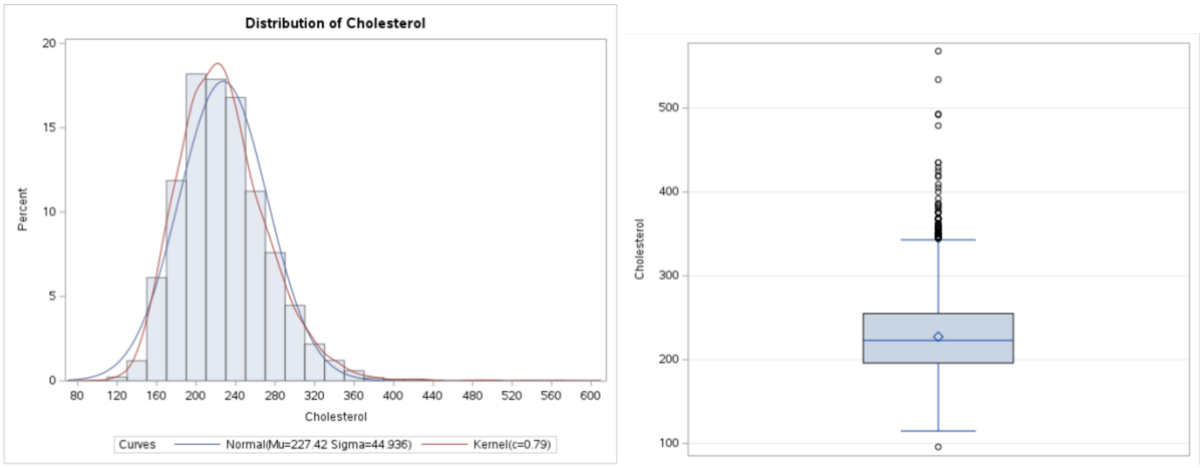

Figure 3.2.17 Distribution and boxplot of cholesterol

Cholesterol data in this dataset is highly right-skewed, and most of its outliers reside above the upper fence. However, its theoretical distribution is aligned with the actual distribution. Most of the cholesterol values are above 200, and their average is around 210 to 220.

3.2.1 Attributes Relationship Analysis

3.2.1.1 Data Transform

```
data WORK.NEW_Heart;
set WORK.NEW_Heart;
Height_Square_in_meter = (Height * 0.0254) ** 2;
bmi = Weight * 0.454 / (Height * 0.0254) ** 2;
if bmi < 19 then Weight_Status = 'Underweight';
else if bmi < 24 then Weight_Status = 'Normal';
else Weight_Status = 'Overweight';
run;
```

```
DATA WORK.NEW_Heart;
set WORK.NEW_Heart;
bp_status_mark = diastolic + (systolic - diastolic) / 3;
run;
```

```
data work.new_heart;
set work.new_heart;
lifespanAfterStarting = ageatdeath - ageatstart;
run;
```

According to the BMI formula, we have counted the BMI index of individuals in the previous section. (Myrna B. Schnur, MSN, RN, 2017) Then, according to the formula given by an article found in the United States National Library of Medicine, we calculate the blood pressure mark of individuals to check the status of their blood pressure by using diastolic and systolic data.

$$BP\_Status\_Mark = \frac{diastolic}{\frac{(systolic - diastolic)}{3}}$$

Figure 3.2.18 Formula Blood Pressure status mark (Saints et al., 2016)

Then, we also establish a new column which is counting the lifespan of one by using AgeAtDeath and AgeAtStart to check the relationship between lifespan and AgeAtStart, AgeAtDeath.

$$LifespanAfterStarting = AgeAtDeath - AgeAtStart$$

Figure 3.2.19 Formula Lifespan After Starting

3.2.1.2 Correlation analysis

```
*create a heat map graph with P-Values in the squares;
proc sgplot data=CorrLong noautolegend;
heatmap x=Variable y=CorrelationID / colorresponse=Correlation
name="nope1" discretex discretey x2axis colormodel=ThreeColorRamp;
*Colorresponse allows discrete squares for each correlation. x2axis bring the label to the top;
text x=Variable y=CorrelationID text=p_value  / textattrs=(size=10pt)
x2axis name='nope2'; /*To overlay significance, create a variable that contans that info and set text=VARIABLE */
label correlation='Pearson Correlation';
yaxis reverse display=(nolabel);
x2axis display=(nolabel);
gradlegend;

run;

proc format;
    value CorrSignif -0.0-<0.2 = "white"
                     0.2-<0.4, -0.4-<-0.2 = "orange"
                     0.4-<0.6, -0.6-<-0.4 = "very_light_orange"
                     0.6-<0.8, -0.8-<-0.6 = "very_light_orange"
                     0.8-<1.0, -1.0-<-0.8= "light_red"
                     1, -1 = "red";
run;

proc corr data=work.corrHeart outp=Corrout (where=(_type_='CORR')) noprint;
run;

proc print data=CorrOut (drop=_type_ rename=(_name_=Variable))
           style(column)={backgroundcolor= CorrSignif.} noobs;
run;
/*corr finish*/
```

Figure 3.2.20 HeatMap table code

Variable	AgeCHDdiag	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	AgeAtDeath	Cholesterol	Height_Square_in_meter	bmi	bp_status_mark	lifespanAfterStarting
AgeCHDdiag	1.00000	0.55091	-0.21035	-0.13751	-0.03856	0.07013	0.00628	-0.28108	0.74811	0.03331	-0.21016	-0.00801	0.01336	0.38685
AgeAtStart	0.55091	1.00000	-0.13166	0.09354	0.27585	0.38732	0.20445	-0.16698	0.68860	0.28303	-0.13215	0.20522	0.34713	-0.10148
Height	-0.21035	-0.13166	1.00000	0.51725	-0.00911	-0.05668	-0.13623	0.28450	-0.13620	-0.07724	0.99941	-0.08553	-0.03305	-0.02165
Weight	-0.13751	0.09354	0.51725	1.00000	0.33138	0.27487	0.76717	0.08800	0.00487	0.07724	0.51899	0.80329	0.32334	-0.01008
Diastolic	-0.03856	0.27585	-0.00911	0.33138	1.00000	0.78017	0.38353	-0.06430	0.01095	0.18954	-0.00868	0.39034	0.95148	-0.10370
Systolic	0.07013	0.38732	-0.05668	0.27487	0.78017	1.00000	0.36124	-0.08990	0.10528	0.20922	-0.05580	0.36324	0.93481	-0.13432
MRW	0.00628	0.20445	-0.13623	0.76717	0.38353	0.36124	1.00000	-0.12376	0.10300	0.13997	-0.13461	0.99393	0.39543	0.00914
Smoking	-0.28108	-0.16698	0.28450	0.08800	-0.06430	-0.08990	-0.12376	1.00000	-0.27713	-0.00937	0.28285	-0.09413	-0.08072	-0.07866
AgeAtDeath	0.74811	0.68860	-0.13620	0.00487	0.01095	0.10528	0.10300	-0.27713	1.00000	0.10085	-0.13580	0.09686	0.05895	0.65151
Cholesterol	0.03331	0.28303	-0.07724	0.07724	0.18954	0.20922	0.13997	-0.00937	0.10085	1.00000	-0.07610	0.13974	0.21052	0.00346
Height_Square_in_meter	-0.21016	-0.13215	0.99941	0.51899	-0.00868	-0.05580	-0.13461	0.28285	-0.13580	-0.07610	1.00000	-0.08347	-0.03238	-0.02049
bmi	-0.00801	0.20522	-0.08553	0.80329	0.39034	0.36324	0.99393	-0.09413	0.09686	0.13974	-0.08347	1.00000	0.40028	0.00139
bp_status_mark	0.01336	0.34713	-0.03305	0.32334	0.95148	0.93481	0.39543	-0.08072	0.05895	0.21052	-0.03238	0.40028	1.00000	-0.12555
lifespanAfterStarting	0.38685	-0.10148	-0.02165	-0.01008	-0.10370	-0.13432	0.00914	-0.07866	0.65151	0.00346	-0.02049	0.00139	-0.12555	1.00000

Table 3.2.1 HeatMap table

Through the calculation using coding, we make a heatmap table that shows the coefficient correlation score and highlights the cell when the relationship meets a certain level. Higher the saturation of the color, the more the relationship between the two attributes.

Through the table, we found attributes that are potential-highly related, and we will analysis on these relationships in the following sections.

3.2.1.3 Analysis on AgeAtStart and AgeAtDeath

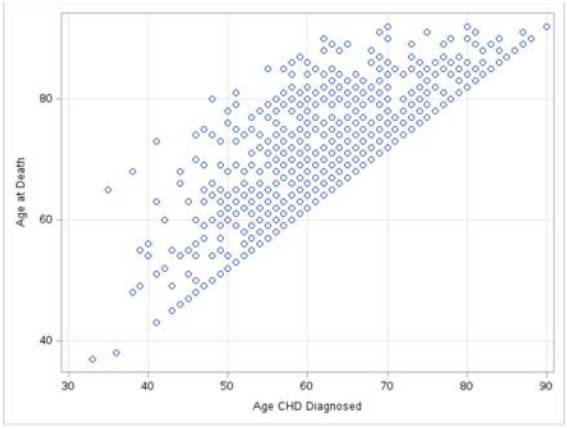

Figure 3.2.21Age at Death vs Age CHD diagnosed

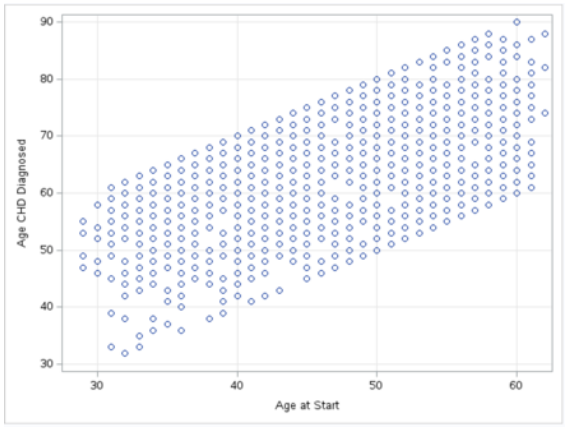

Figure 3.2.22 Age CHD Diagnosed vs Age at Start

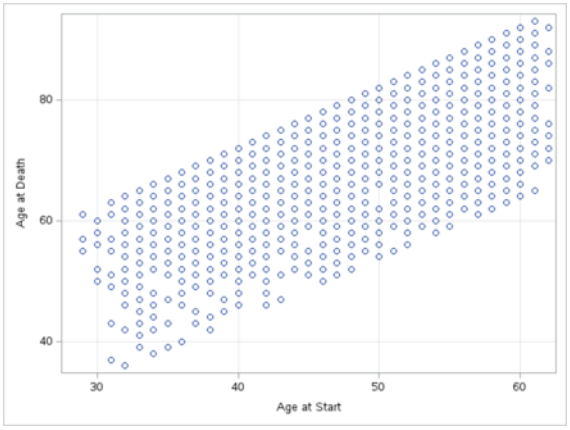

Figure 3.2.23 Age at Death vs Age at Start

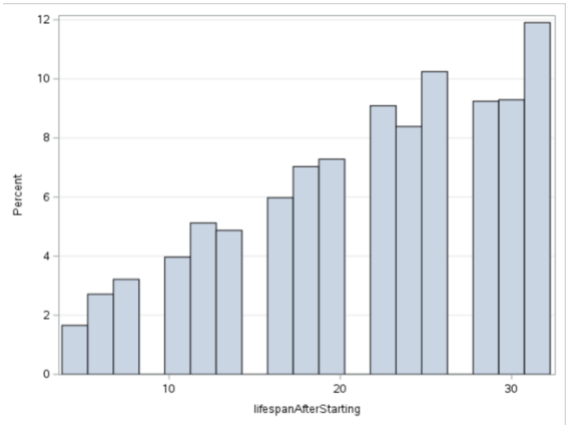

Figure 3.2.24 Lifespan After Starting histogram

Figure 3.2.1.3.1 the age at death and age of CHD diagnosed also have a direct proportion relationship. The later the CHD was diagnosed, the later the respondent died. After CHD is diagnosed, the respondents will die after 20 to 30 years.

Figure 3.2.1.3.2 the people who start to have the habits are likely to diagnose with CHD after around 30 years. The direct proportional relationships shown in the graph conclude that the later the behavior starts, the later the CHD is diagnosed.

Figure 3.2.1.3.3 Age at death is directly proportional to the age at start, the later the age at start, the later the age at death. Meanwhile, the time duration between age at start and age at

death is all similar. In other words, once the respondents start the behavior, they probably left around 20 to 30 years of lifespan. This observation same as the findings of the first graph.

Figure 3.2.1.3.4 the histogram of lifespan after the behavior starts demonstrates that most people have around 30 years of lifespan, which is the same as our observations in the previous figures.

In a summary, the timeframe between the age at start and the age of CHD diagnosed is around 20 to 30 years and after that, in most cases, CHD patients will die 30 years after CHD is diagnosed. However, the above analysis result requires further data collection and research as the observation summary for age at death and age at start are too intuitive. If this conclusion is true, the behavior pointed out by the age at start is not suggested to be cultivated as it will shorten the lifespan of one.

3.2.1.4 Analysis of Sex and Smoking Status

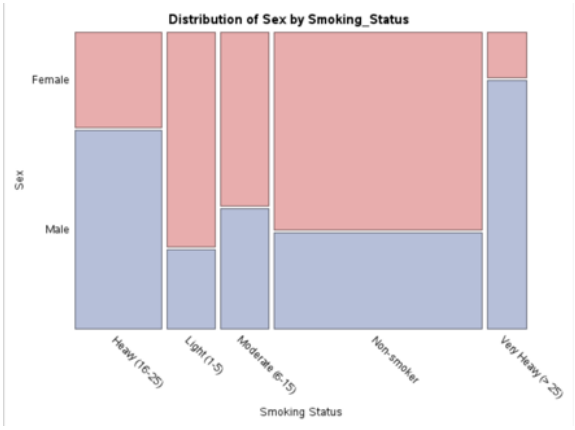

Figure 3.2.25 Distribution of sex and smoking status

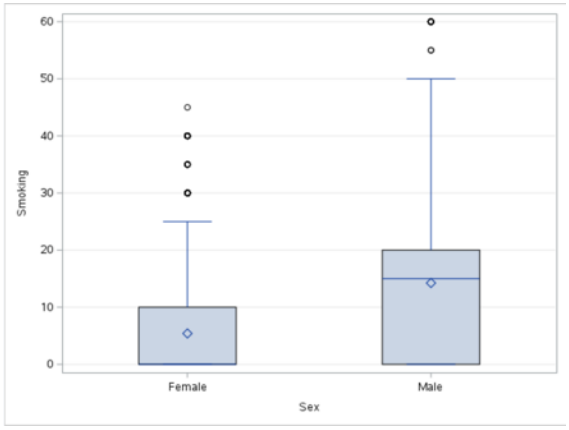

Figure 3.2.26 Smoking boxplot for sex

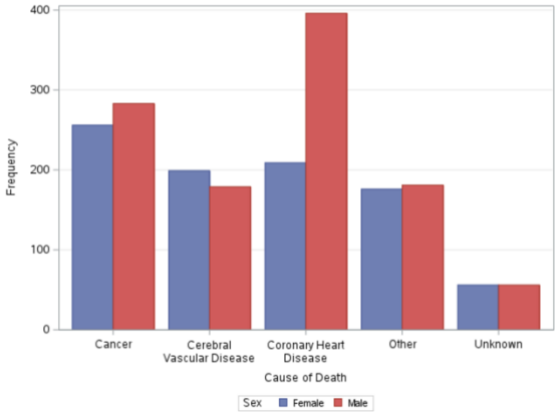

Figure 3.2.27 Distribution of death cause and sex

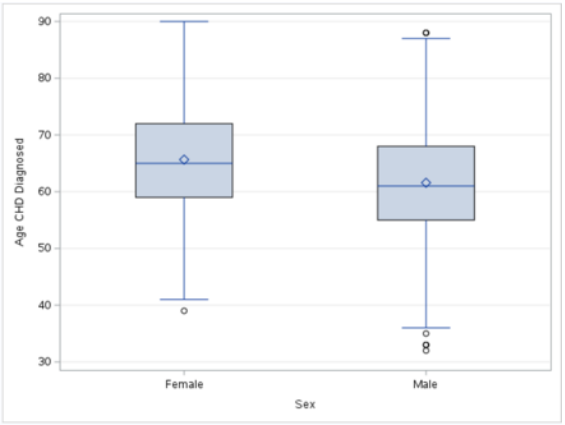

Figure 3.2.28 AgeCHDdiag boxplot for sex

Figure 3.2.1.4.1 we found that in this dataset, the non-smoker population forms the majority, and among the non-smokers, the females are more than the males.

Figure 3.2.1.4.2 the boxplot illustrates the same findings as in the last figures, the smoking average scores for the female are visibly lower than for the male.

Figure 3.2.1.4.3 we continue to find that the number one killer for males is coronary heart disease (CHD) while cancer is the number one killer for females.

Figure 3.2.1.4.4, graph implies that the average CHD-diagnosed age for females is later than for males.

From the observations above, we can conclude that heavy smoking might lead to CHD as those male heavy smokers in the dataset tend to suffer from CHD. These findings align with the conclusion from the European journal that also upholds the relationship between

cigarettes and coronary artery disease. (Inoue, 2004) Smoking might be the reason for the males in this dataset having a shorter lifespan than females. Henceforth, the result implies the setbacks of smoking as it leads to a shorter lifespan and several chronic diseases especially the CHD mentioned in this dataset.

3.2.1.5 Analysis of Height, Weight, MRW by Sex

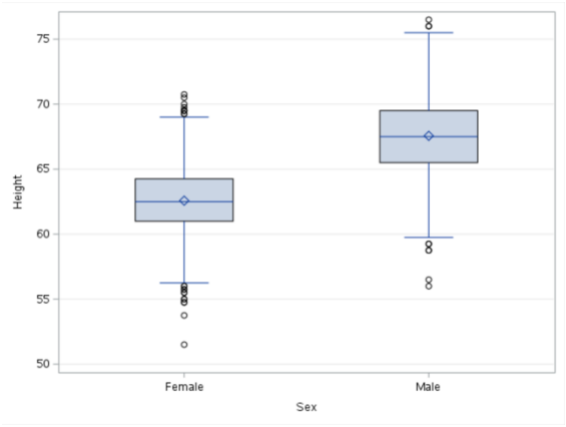

Figure 3.2.29 Height boxplot by sex

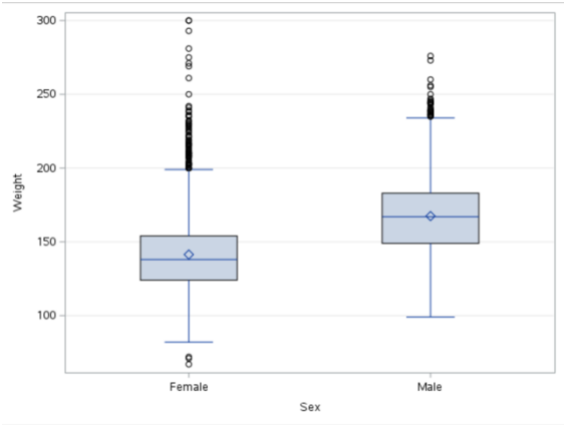

Figure 3.2.30 Weight boxplot by sex

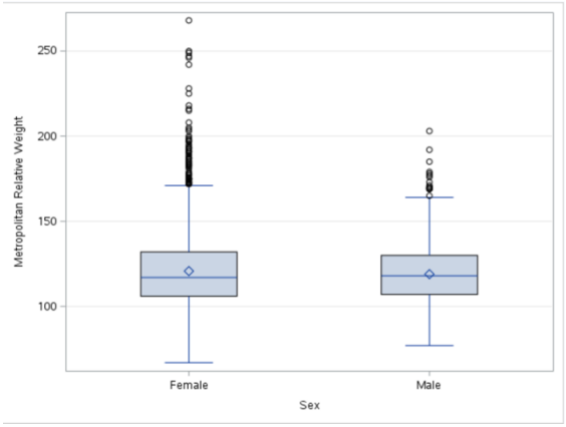

Figure 3.2.31 MRW boxplot by sex

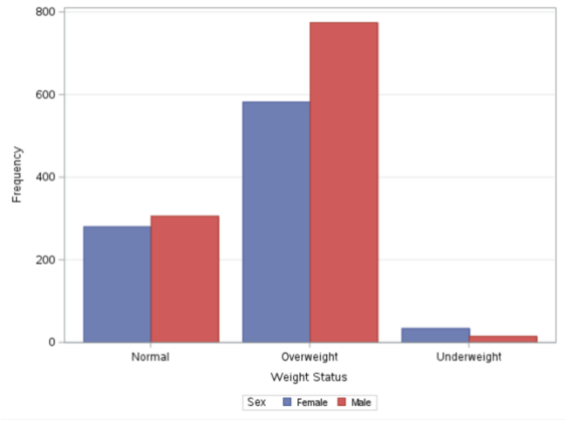

Figure 3.2.32 Weight Status bar chart by sex

Figure 3.2.1.5.1 we find that the males are 5 inches averagely higher than females in terms of body height.

Figure 3.2.1.5.2 indicates that the weight of males is also 30 pounds heavier than females on average.

Figure 3.2.1.5.3 indicates that the MRW of both females and males is almost the same.

Figure 3.2.1.5.4 points out that men in this dataset are overall having obesity problems.

Based on the research, MRW is an index used in Framingham Heart Study and is calculated using weight and height. (P. Simopoulos, 1986) Hence, the male's MRW that is calculated using these 2 variables should be higher than the female's. However, the observations from the first and second figures are against the last figures. To determine the relationship between height, weight, and MRW, further data collection and analysis should be conducted.

On the other hand, the weight status of males shows that there is a more obese male population in this dataset compared to females. In other words, males are not genetically or congenitally higher and heavier than females as the male samples in this dataset are highly obese.

3.2.1.6 Analysis of Diastolic, Systolic, and Blood Pressure Status by Sex

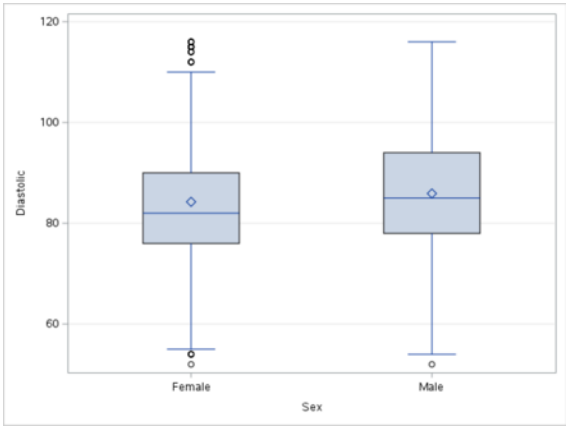

Figure 3.2.33 Diastolic boxplot by sex

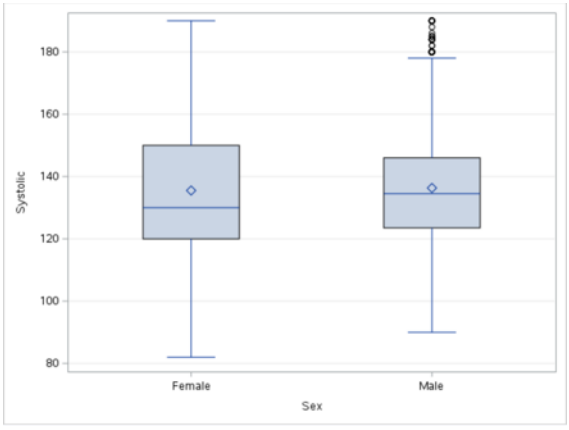

Figure 3.2.34 Systolic boxplot by sex

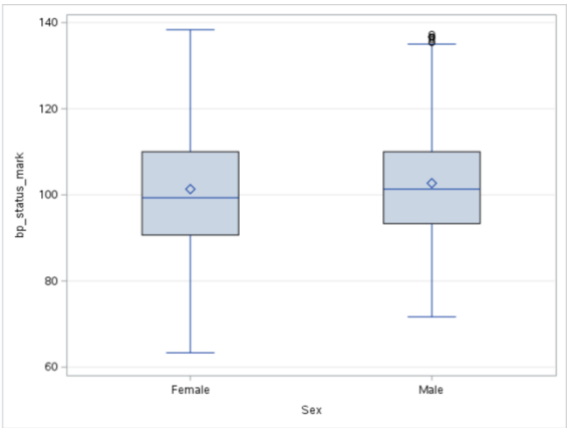

Figure 3.2.35 BP_Status_Mark boxplot by sex

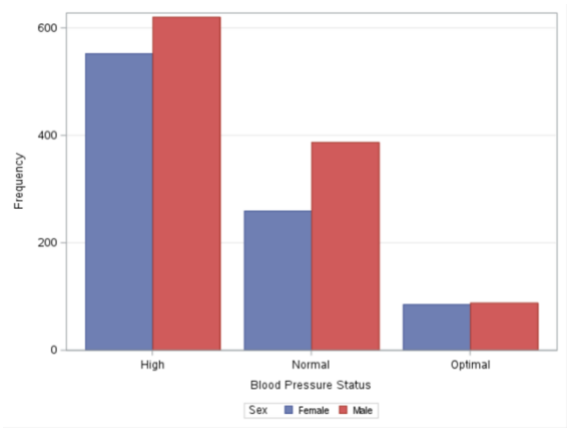

Figure 3.2.36 Blood Pressure status bar chart by sex

Figure 3.2.1.6.1 indicates that the diastolic measured in males is slightly higher than in females on average.

Figure 3.2.1.6.2 shows that the systolic value of males is also higher than females on average.

Figure 3.2.1.6.3 denotes that the males have slightly lower blood pressure than the males on average.

Figure 3.2.1.6.4 implies that the males in this dataset have higher blood pressure than females. But at the same time, men with normal blood pressure also have more than females.

It is interesting to find that the male's diastolic, systolic, and blood pressure is probably higher than the female's. However, when we check the frequency of blood pressure status in the dataset according to sex, we find that this finding might be due to the unbalanced sample residing in these three columns. The males with higher blood pressure have more than the females. Although the number of males with normal high blood pressure reduces the gap in the mean blood pressure mark between the two sexes, the average blood pressure of males is still slightly higher than the blood pressure of females. Nevertheless, male does not generally have higher blood pressure compared to female.

3.2.1.7 Analysis of the Cause of Death and Cholesterol Status

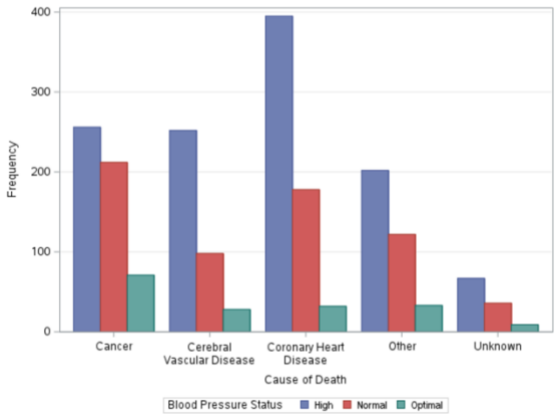

Figure 3.2.37 Cause of Death bar chart by Blood Pressure Status

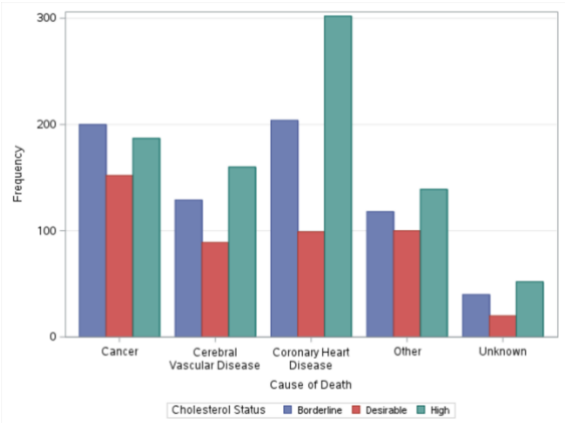

Figure 3.2.38 Cause of Death bar chart by Cholesterol Status

Figure 3.2.1.7.1 illustrates that people with high blood pressure usually die from CHD, cancer, and cerebral vascular disease.

Figure 3.2.1.7.2 indicates that people with high cholesterol status die from CHD, cancer, and cerebral vascular disease.

The figures above have shown that the top killer for people with high cholesterol and high blood pressure. In order, those diseases include CHD, cancer, and cerebral vascular disease. In short, hypertension and high cholesterol should not be neglected since they tend to bring serious diseases and death to one.

3.2.1.8 Analysis of Smoking Status, AgeAtDeath, and CHDdiag

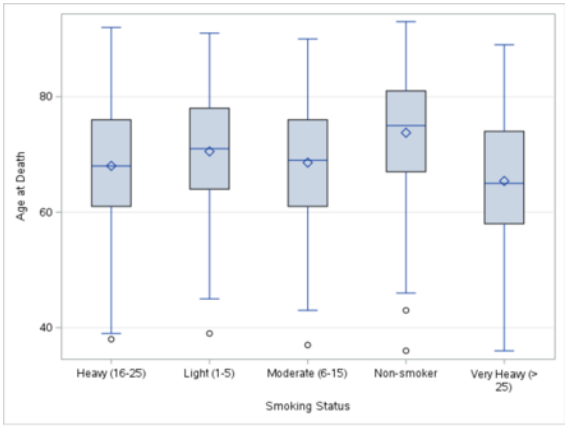

Figure 3.2.39 Age at Death boxplot by Smoking Status

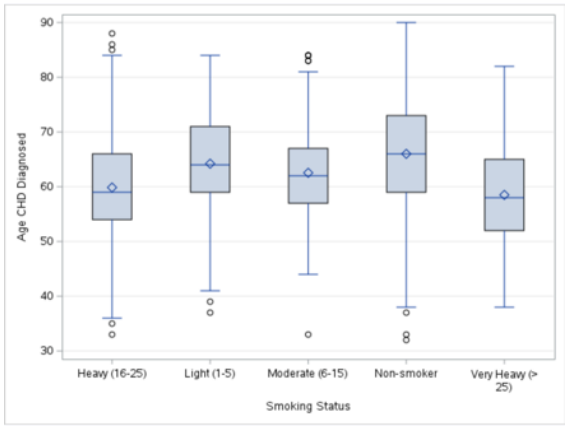

Figure 3.2.40 Age CHD diagnosed boxplot by Smoking Status

Figure 3.2.1.8.1 shows that the non-smokers have the latest age at death, followed by light-smokers, moderate-smokers, heavy-smokers, and very-heavy-smokers.

Figure 3.2.1.8.2 implies that non-smokers are the latest ones to suffer from CHD, followed by light-smokers, moderate-smokers, and heavy and very heavy smokers.

For this analysis, both figures point out that the smoking habit brings death and CHD diseases to smokers. People without smoking can enjoy longer lifespans and are unlikely to suffer from CHD disease at a young age.

4 Data Preprocessing

4.1 Incomplete data

Common methods for in filling missing data are using mean, median, and mode. The following section will further explain the reason for choosing methods for in filling missing data.

$$\text{mean} = \frac{\sum_{i=1}^{N} X_i}{N}$$

Figure 4.1.1.0.1 Mean formula (Luke, 2021)

$$m = L + \left(\frac{\frac{N}{2} - F}{f_m} \right) c$$

Figure 4.1.0.2 Median formula (Onlinetuition.com, 2013)

4.1.1 Filling in missing data with the median

Among the three-missing data-filling methods aforementioned, the mode is the method for missing categorical data. Hence, the mean and median will be the main consideration for the missing numeric data below. As the graph shows above, height has a distribution that is lower than the expected distribution. The weight column has a right-skewed distribution, which means its extreme data mostly reside on the right side, leading to a stagger between the red curve and the blue curve in the distribution graph. Besides, smoking, MRW, and cholesterol are highly right-skewed attributes. Filling missing data using mean value brings drawbacks of changing on standard deviation and quartile values of particular columns when the data distribution is not symmetric and highly skewed since mean calculation involved every value in the dataset. (Jim, 2019) Henceforth, filling these missing data with a median is a more ideal method compared to the mean. (Dario, 2019)

```
/*Filling in missing data with median*/
proc stdize data=SASHELP.Heart out=WORK.NEW_Heart reponly
        method=median;
        var Cholesterol Smoking MRW Weight Height;
run;
```

Figure 4.1.2 replacing value with the median

By using the standardizing function in SAS coding, we can replace missing data with a median. (Wicklin, 2020)

4.1.2 Filling in missing data by using calculation

Among the variables, the status of individuals in terms of cholesterol, weight status, and level of smoking is determined by using cholesterol rate in blood, weight, height, and smoking marks respectively. In other words, they are determined by using calculation. Hence, the column missing data should be filled by using the related standard. For cholesterol, we determine its status by using the standard given by the author from MedicalNewsToday and Cleveland clinic, and it is aligned with the standard in the original dataset.

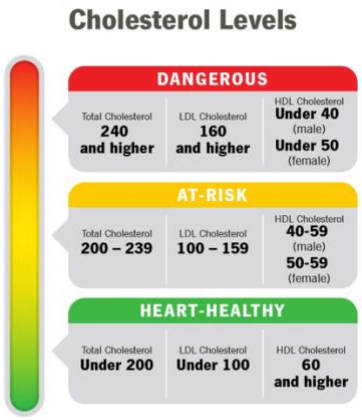

Figure 4.1.3 Cholesterol status standard (Brazier, 2021) (Cleveland Clinic, 2022)

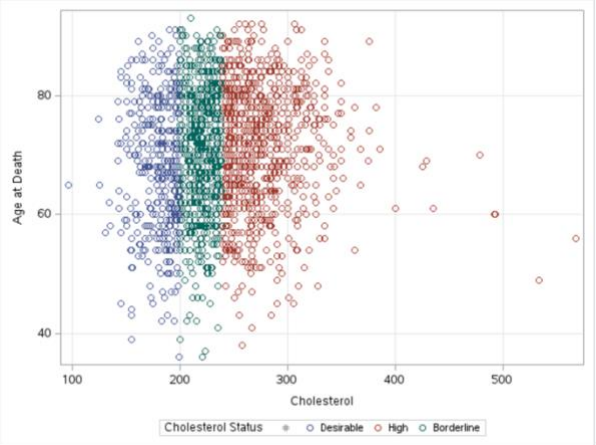

Figure 4.1.4 distribution of cholesterol status in the dataset

The standard of weight status, we found that it is determined by using the BMI (Body Mass Index) by comparing the distribution of weight status in the dataset and the measuring standard of BMI.

Height	Normal weight BMI 19–24	Overweight BMI 25–29	Obesity BMI 30–39	Severe obesity BMI 40+
4 ft 10 in (58 in)	91–115 lb	119–138 lb	143–186 lb	191–258 lb
4 ft 11 in (59 in)	94–119 lb	124–143 lb	148–193 lb	198–267 lb
5ft (60 in")	97–123 lb	128–148 lb	153–199 lb	204–276 lb
5 ft 1 in (61 in)	100–127 lb	132–153 lb	158–206 lb	211–285 lb
5 ft 2 in (62 in)	104–131 lb	136–158 lb	164–213 lb	218–295 lb
5 ft 3 in (63 in)	107–135 lb	141–163 lb	169–220 lb	225–304 lb
5 ft 4 in (64 in)	110–140 lb	145–169 lb	174–227 lb	232–314 lb
5 ft 5 in (65 in)	114–144 lb	150–174 lb	180–234 lb	240–324 lb
5 ft 6 in (66 in)	118–148 lb	155–179 lb	186–241 lb	247–334 lb
5 ft 7 in (67 in)	121–153 lb	159–185 lb	191–249 lb	255–344 lb
5 ft 8 in (68 in)	125–158 lb	164–190 lb	197–256 lb	262–354 lb
5 ft 9 in (69 in)	128–162 lb	169–196 lb	203–263 lb	270–365 lb
5 ft 10 in (70 in)	132–167 lb	174–202 lb	209–271 lb	278–376 lb
5 ft 11 in (71 in)	136–172 lb	179–208 lb	215–279 lb	286–386 lb
6 ft (72 in)	140–177 lb	184–213 lb	221–287 lb	294–397 lb
6 ft 1 in (73 in)	144–182 lb	189–219 lb	227–295 lb	302–408 lb
6 ft 2 in (74 in)	148–186 lb	194–225 lb	233–303 lb	311–420 lb
6 ft 3 in (75 in)	152–192 lb	200–232 lb	240–311 lb	319–431 lb
6 ft 4 in (76 in)	156–197 lb	205–238 lb	246–320 lb	328–443 lb

Figure 4.1.5 BMI standard given by Lippincott NursingCenter (Myrna B. Schnur, MSN, RN, 2017)

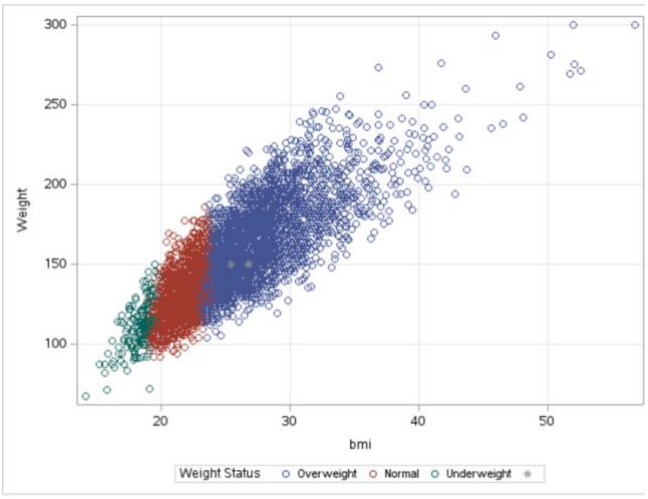

Figure 4.1.6 distribution of weight status in the dataset

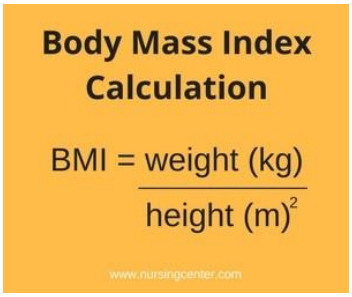

Figure 4.1.7 BMI formula (Myrna B. Schnur, MSN, RN, 2017)

```
/*Filling in missing data with calculation*/
data WORK.NEW_Heart;
set WORK.NEW_Heart;
if Cholesterol <= 200 then Chol_Status = 'Desirable';
else if Cholesterol <= 239 then Chol_Status = 'Borderline';
else Chol_Status = 'High';
run;

data WORK.NEW_Heart;
set WORK.NEW_Heart;
Height_Square_in_meter = (Height * 0.0254) ** 2;
bmi = Weight * 0.454 / (Height * 0.0254) ** 2;
if bmi < 19 then Weight_Status = 'Underweight';
else if bmi < 24 then Weight_Status = 'Normal';
else Weight_Status = 'Overweight';
run;

data WORK.NEW_Heart;
set WORK.NEW_Heart;
if Smoking = 0 then Smoking_Status = 'Non-smoker';
else if Smoking <= 5 then Smoking_Status = 'Light (1-5)';
else if Smoking <= 15 then Smoking_Status = 'Moderate (6-15)';
else if Smoking <= 25 then Smoking_Status = 'Heavy (16-25)';
else Smoking_Status = 'Very Heavy (>25)';
run;
```

Figure 4.1.8 Recalculation of missing data

By using SAS coding, we make recalculations on cholesterol status, weight status, and smoking status. The weight and height are converted to metric units (Kg and Meter) for BMI calculation. The smoking status is calculated based on the standard given by the dataset.

4.1.3 Meaningful missing data

There are a few meaningful attributes in the dataset, meaning the missing data in the columns are supposed to be deleted.

```
/*Meaningful missing data*/
DATA WORK.NEW_Heart;
    set WORK.NEW_Heart;
    /*Filling in missing data with mode*/
    if DeathCause=' ' and Status='Dead' then
        DeathCause="Coronary Heart Disease";
    /*Filling in missing data with median*/
    if AgeAtDeath='.' and Status='Dead' then AgeAtDeath=71;
run;
```

Figure 4.1.9 replacing missing data with mode

The missing data in the Death Cause column is due to the individuals are still alive and we have made sure that there are no individuals that are dead without Death Cause data in the dataset. If the situation exists, the missing data will be filled with mode (coronary heart disease) since Death Cause is categorical data.

Age At Death is missing due to the individual is still alive. So, we should not fill in the missing rows in Age At Death column.

Besides, the missing AgeCHDdiag data might be because the individual has never been confirmed for having coronary heart disease, whether alive or dead.

4.2 Noisy data

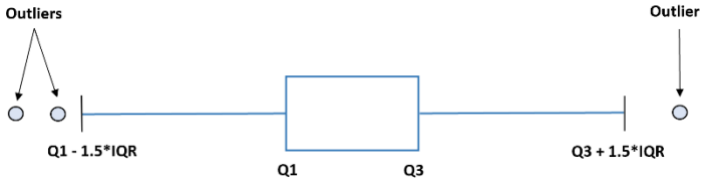

Figure 4.2.0.1 Determining of outliers (Zach, 2021)

Outliers mean the data values that are lesser than the lower fence (Q1 – 1.5 * IQR) or greater than the upper fence (Q3 + 1.5 * IQR). The treatments of outliers include winsorizing, trimming, and replacing values. (Magaga, 2021)

Winsorizing, also called flooring and capping, is a statistical technique in which outliers in a dataset are converted to a certain percentile value. By replacing the values at particular percentiles like the 5^{th} and 95^{th}, effect of outliers on statistical analysis will be minimized.

Trimming data has several ways, we can delete the data rows or filter the data column based on the percentiles. Trimming the data column helps data analyst to focus on the mainstream of the data when the outliers are the actual cases that exist in the real world, or it could be a data entry error as well. (Protobi, 2018) (Zach, 2021)

Replacing outliers with mean values is also considerable in dealing with noisy data. However, replacing data values triggered the problem of changing its data distribution and other statistical data indicators.

4.2.1 Winsorizing (value replacing)

Diastolic, Systolic, and cholesterol outliers will be winsorized. This is due to the outliers in these columns seems not like error data. In other words, they are actual data. Henceforth, to maintain the trend of data, these outliers should not be replaced with mean values or deleted. The winsorizing method can also make sure that those outliers stay at the edge of the distribution, including in the further data analysis result.

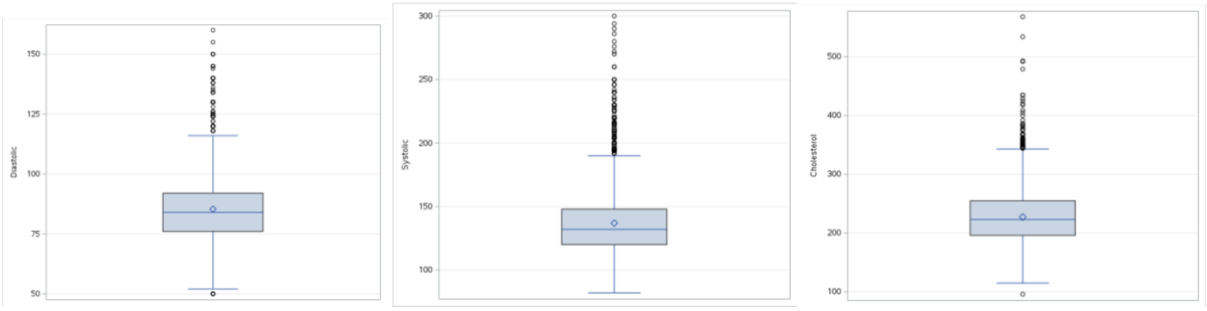

Figure 4.2.1 Initial boxplot graph of Diastolic, Systolic, and cholesterol

The graph above is the initial boxplot graph of the diastolic, systolic, and cholesterol columns.

```
/*Noisy data*/
/*Winsorizing data*/
DATA WORK.NEW_Heart;
    set WORK.NEW_Heart;
    minimum=76-(1.5*16);
    maximum=92+(1.5*16);
    /*5% to 95%*/
    if Diastolic > maximum then Diastolic=110;
    else if Diastolic < minimum then Diastolic=68;
run;

DATA WORK.NEW_Heart;
    set WORK.NEW_Heart;
    minimum=120-(1.5*28);
    maximum=148+(1.5*28);
    /*5% to 95%*/
    if Systolic > maximum then Systolic=180;
    else if Systolic < minimum then Systolic=108;
run;

DATA WORK.NEW_Heart;
    set WORK.NEW_Heart;
    minimum=196-(1.5*59);
    maximum=255+(1.5*59);
    if Cholesterol > maximum then Cholesterol=307;
    else if Cholesterol < minimum then Cholesterol=163;
run;
```

Figure 4.2.2 code of winsorizing data

Through SAS coding, we set the winsorizing percentile at 5% and 95%, so that we can minimize our modification towards the original dataset.

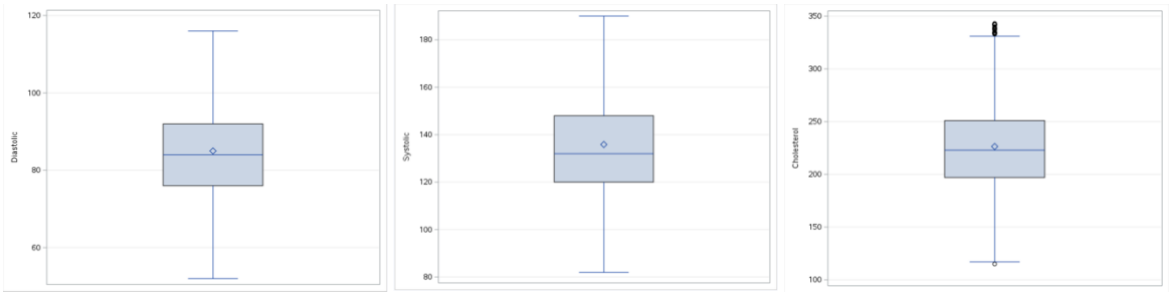

Figure 4.2.3 Boxplot graph for 1st round winsorizing

However, after the first time of winsorizing, cholesterol data's Q1, Q3, and quartile range changed, leading to the new outliers. Therefore, we have to make another round of winsorizing for cholesterol data.

The MEANS Procedure

Analysis Variable : Cholesterol

N	N Miss	Minimum	Median	Maximum	Std Dev	Lower Quartile	Upper Quartile	Quartile Range
5209	0	115.0000000	223.0000000	343.0000000	41.5027505	197.0000000	251.0000000	54.0000000

Table 4.2.1 Cholesterol Q1, Q3 and quartile range change

```
proc means data=WORK.NEW_HEART n nmiss min median median max std q1 q3 qrange;
    var Cholesterol;
run;

DATA WORK.NEW_Heart;
    set WORK.NEW_Heart;
    minimum=197-(1.5*54);
    maximum=251+(1.5*54);
    if Cholesterol > maximum then Cholesterol=307;
    else if Cholesterol < minimum then Cholesterol=163;
run;
```

Figure 4.2.4 check descriptive statistics and proceed to 2nd round of winsorizing

For the second round of winsorizing, we capped the percentile at 5% and 95% and we have successfully removed those outliers in cholesterol columns.

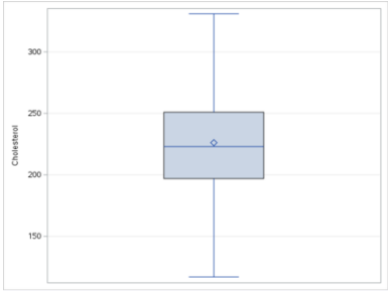

Figure 4.2.5 Cholesterol boxplot after 2nd winsorizing

4.2.2 Trimming (Quartile-based data filtering)

Other columns that contain outliers include AgeCHDdiag, height, weight, MRW, smoking, and AgeAtDeath. They are the actual measured data and there are no unreasonable data among them. Those outliers in weight, height, smoking, and AgeAtDeath seem reasonable. As a summary, we found that these outliers are mostly real data that should not be modified unless it is a data entry error. However, for some outliers like AgeCHDdiag and MRW is hard to determine whether there is a data entry error that occurs. As a result, for those real data or those data whose data credibility is undetermined, we can only remove them from the dataset.

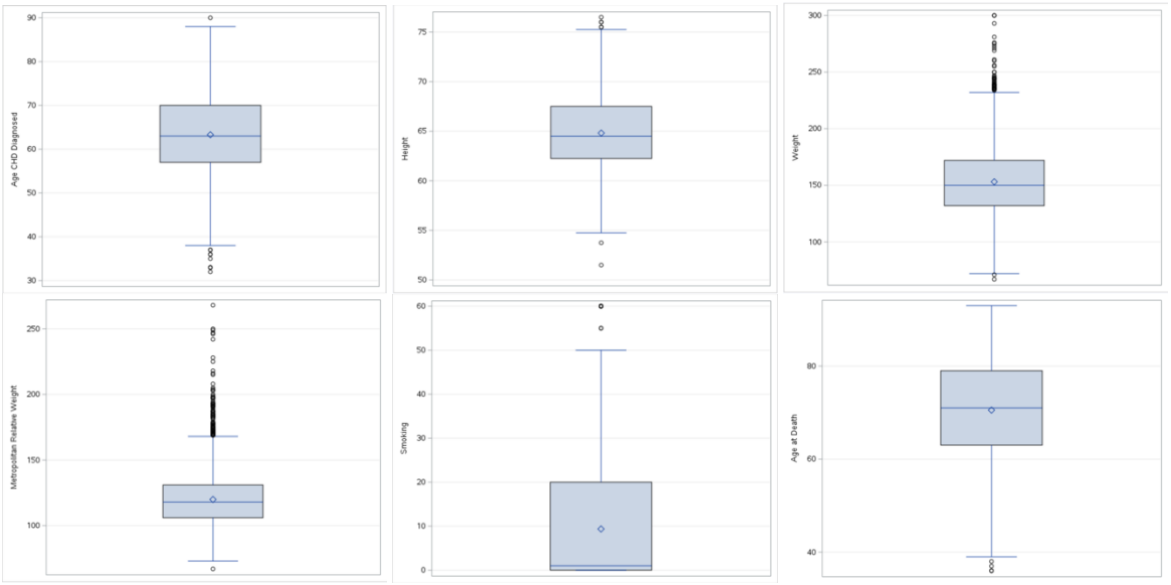

Figure 4.2.6 initial boxplot graph for AgeCHDdiag, Height, Weight, MRW, Smoking, and AgeAtDeath

The above boxplot graphs describe the situation of AgeCHDdiag, Height, Weight, MRW, Smoking, and AgeAtDeath.

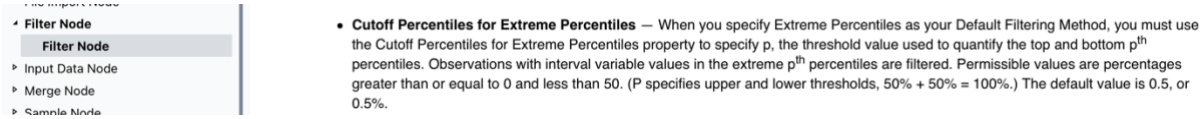

Figure 4.2.7 Using of filter node for winsorizing techniques (SAS, 2020)

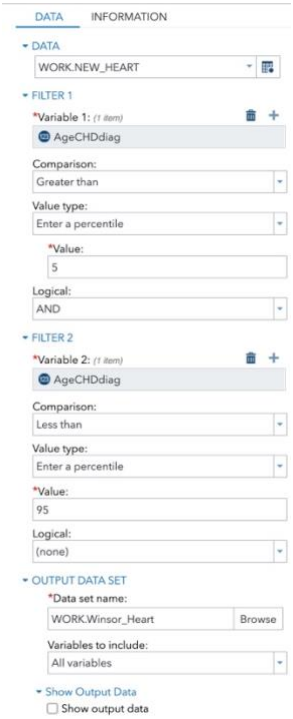

Figure 4.2.8 data trimming setting

To minimize the data portion trimmed, we set the percentile of trimming at 5% and 95%. The filter data function in SAS studio can filter the data from the 5th to 95th percentile (filter the data below 5% and above 95%). This function shown will leave the data greater than the 5th percentile and lesser than the 95th percentile. We have made the same filtering to AgeCHDdiag, Height, Weight, and MRW. For the Smoking and AgeAtDeath columns, since their outliers only occur either below the upper fence or above the lower fence, we only trim the data above 95% and below 5% respectively.

The following graph shows the boxplot after data trimming:

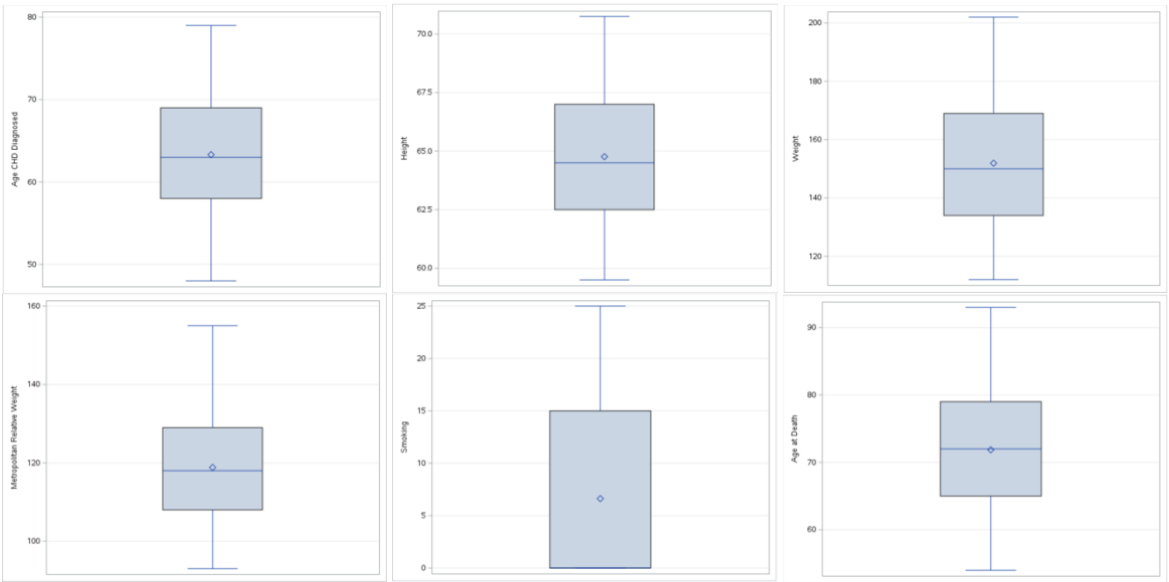

Figure 4.2.9 Boxplot graph after data trimming

4.3 Inconsistent data

Inconsistent data format exists in categorical data could be the difference in the uppercase, lowercase, and white space between the words. (Isdiarti, 2021). Furthermore, inconsistency also occurs when the same data exists in the dataset.

4.3.1 Data format Inconsistency checking

```
/*Inconsistent data*/
/*Categorical*/
proc sql;
select Weight_Status
from work.new_heart
where Status~='Dead' and Status~='Alive';
run;

proc sql;
select DeathCause
from work.new_heart
where DeathCause~='Cancer' and DeathCause ~='Cerebral Vascular Disease'
and DeathCause ~='Coronary Heart Disease' and DeathCause ~='Other' and DeathCause ~='Unknown';
run;

proc sql;
select Sex
from work.new_heart
where Sex~='Male' and Sex~='Female';
run;

proc sql;
select Chol_Status
from work.new_heart
where Chol_Status~='Borderline' and Chol_Status~='Desirable' and Chol_Status~='High';
run;

proc sql;
select BP_Status
from work.new_heart
where BP_Status~='High' and BP_Status~='Normal' and BP_Status~='Optimal';
run;

proc sql;
select Weight_Status
from work.new_heart
where Weight_Status~='Normal' and Weight_Status~='Overweight' and Weight_Status~='Underweight';
run;

proc sql;
select Smoking_Status
from work.new_heart
where Smoking_Status~='Heavy (16-25)' and Smoking_Status~='Very Heavy (>25)'
and Smoking_Status~='Light (1-5)' and Smoking_Status~='Moderate (6-15)'
and Smoking_Status~='Non-smoker';
run;
```

```
/*Filling in missing data with calculation*/
data WORK.NEW_Heart;
set WORK.NEW_Heart;
if Cholesterol <= 200 then Chol_Status = 'Desirable';
else if Cholesterol <= 239 then Chol_Status = 'Borderline';
else Chol_Status = 'High';
run;

data WORK.NEW_Heart;
set WORK.NEW_Heart;
bmi = Weight * 0.454 / Height * 2.54 ** 2;
if bmi < 19 then Weight_Status = 'Underweight';
else if bmi < 24 then Weight_Status = 'Normal';
else Weight_Status = 'Overweight';
run;

data WORK.NEW_Heart;
set WORK.NEW_Heart;
if Smoking = 0 then Smoking_Status = 'Non-smoker';
else if Smoking <= 5 then Smoking_Status = 'Light (1-5)';
else if Smoking <= 15 then Smoking_Status = 'Moderate (6-15)';
else if Smoking <= 25 then Smoking_Status = 'Heavy (16-25)';
else Smoking_Status = 'Very Heavy (>25)';
run;
```

Figure 4.3.1 data inconsistency checking using SQL

For data inconsistency, we utilize the integrated SQL in SAS coding to search for the data inconsistency. Furthermore, since we have recomputed the value of cholesterol, weight, and smoking status in the previous section, the categorical data found no inconsistency here.

4.3.2 Duplicate Data Checking

```
proc sql;
select COUNT(*) as row_count from work.new_heart;
run;

proc sort data=work.new_heart out=work.no_dups noduprecs;
by _all_;
run;

proc sql;
select COUNT(*) as row_count from work.no_dups;
run;
```

row_count
5209

row_count
5209

Figure 4.3.2 SQL code and row count after removing duplicate data

By using SQL coding integrated with SAS studio, we compare the data row number after applying the duplicate data removing code to the dataset. We found that there are no duplicate data in the dataset.

However, since this dataset records people with different health conditions including smoking status, CHD disease, age at death, etc. Even if duplicate data exists in this dataset, we might also consider not deleting them as the sample people could have similar or the same health conditions.

4.4 New dataset

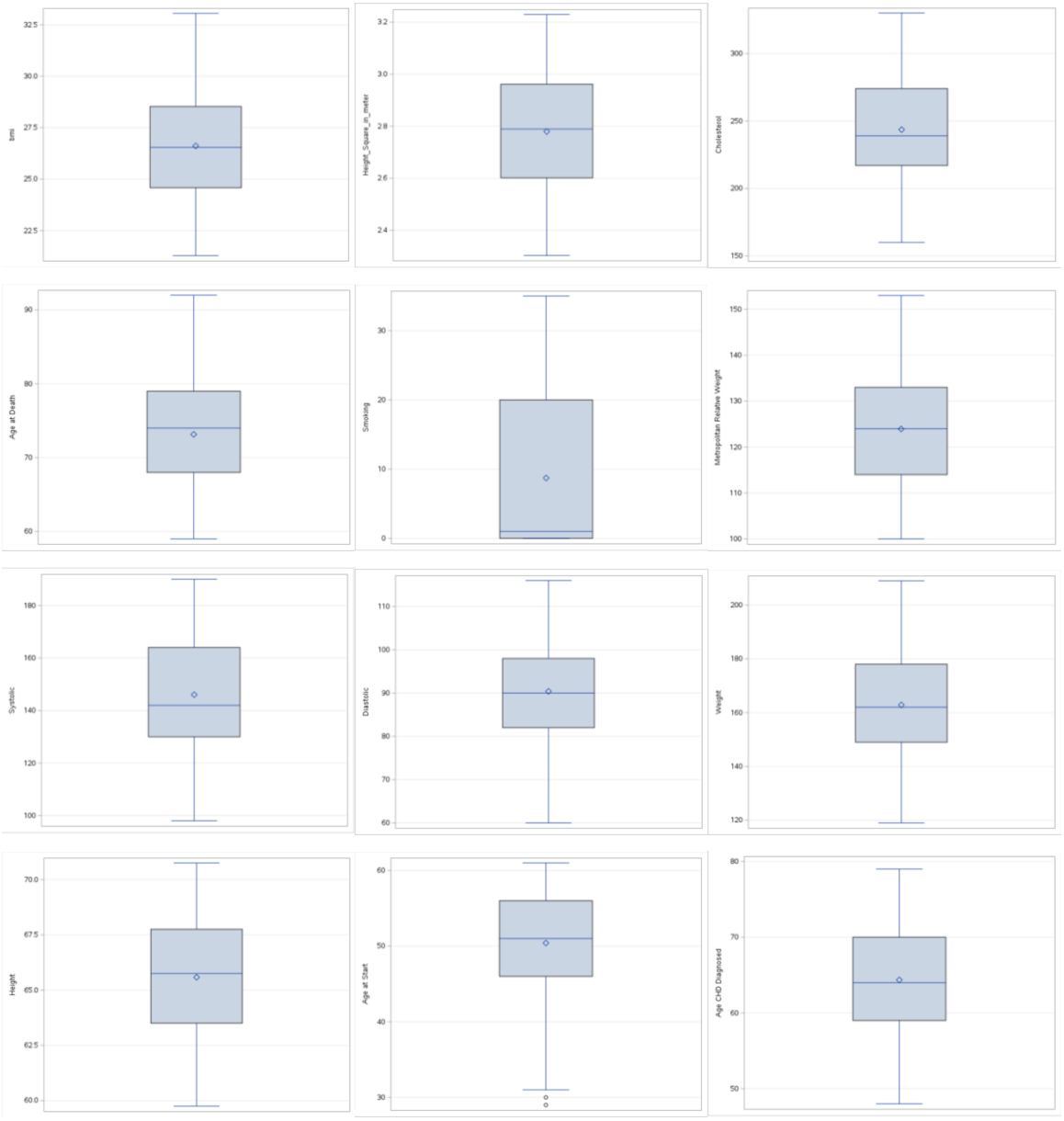

Figure 4.4.1 all attributes boxplot from new dataset

The exploratory data analysis and preprocessing process were conducted on the given dataset. Outliers were identified in several numerical columns, including diastolic, systolic, and cholesterol. Winsorizing was used to adjust these outliers, while trimming was applied to columns such as AgeCHDdiag, Height, Weight, and MRW to remove extreme values. No inconsistency was found in categorical data after using integrated SQL coding. Finally, the

dataset was checked for duplicate data, and none were found. Overall, the preprocessing steps were necessary to ensure the data's quality and prepare it for further analysis.

Smoking Status

Smoking_Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Heavy (16-25)	113	23.64	113	23.64
Light (1-5)	43	9.00	156	32.64
Moderate (6-15)	60	12.55	216	45.19
Non-smoker	233	48.74	449	93.93
Very Heavy (>25)	29	6.07	478	100.00

Blood Pressure Status

BP_Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
High	305	63.81	305	63.81
Normal	143	29.92	448	93.72
Optimal	30	6.28	478	100.00

Weight Status

Weight_Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Normal	85	17.78	85	17.78
Overweight	393	82.22	478	100.00

Cholesterol Status

Chol_Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Borderline	164	34.31	164	34.31
Desirable	76	15.90	240	50.21
High	238	49.79	478	100.00

Cause of Death

DeathCause	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Cancer	56	11.72	56	11.72
Cerebral Vascular Disease	59	12.34	115	24.06
Coronary Heart Disease	314	65.69	429	89.75
Other	40	8.37	469	98.12
Unknown	9	1.88	478	100.00

Sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	151	31.59	151	31.59
Male	327	68.41	478	100.00

Frequencies for Categorical Variables

Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Dead	478	100.00	478	100.00

Descriptive Statistics for Numeric Variables

Variable	Label	N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
AgeCHDdiag	Age CHD Diagnosed	478	0	48.0000000	64.3598326	64.0000000	79.0000000	7.2707606
AgeAtStart	Age at Start	478	0	29.0000000	50.4142259	51.0000000	61.0000000	6.7336234
Height		478	0	59.7500000	65.5794979	65.7500000	70.7500000	2.7599084
Weight		478	0	119.0000000	162.8682008	162.0000000	209.0000000	20.5354271
Diastolic		478	0	60.0000000	90.4037657	90.0000000	116.0000000	11.8054702
Systolic		478	0	98.0000000	146.0292887	142.0000000	190.0000000	21.7293042
MRW	Metropolitan Relative Weight	478	0	100.0000000	123.9435146	124.0000000	153.0000000	12.3432286
Smoking		478	0	0	8.7217573	1.0000000	35.0000000	10.2585030
AgeAtDeath	Age at Death	478	0	59.0000000	73.1548117	74.0000000	92.0000000	7.2679394
Cholesterol		478	0	160.0000000	243.5899582	239.0000000	330.0000000	39.5018677
Height_Square_in_meter		478	0	2.3032615	2.7795246	2.7890670	3.2293887	0.2327127
bmi		478	0	21.2839745	26.6156029	26.5437795	33.0565392	2.6601153
minimum		478	0	116.0000000	116.0000000	116.0000000	116.0000000	0
maximum		478	0	332.0000000	332.0000000	332.0000000	332.0000000	0

Table 4.4.1 all attributes in new dataset without noisy data

5 Conclusion

Exploratory Data Analysis (EDA) is to evaluate and study data sets by summarizing their key features and employing data visualization techniques. By applying EDA to the dataset, we can improve data comprehension by spotting patterns, anomalies, correlations, and other characteristics that may be used for further analysis and modeling. In this SASHELP.HEART dataset, we have been able to spot possible relationships between the attributes of the dataset including Height, Weight, MRW, Cause of Death, smoking status and the list goes on. These findings might be helpful in disease treatment and guiding people to not have bad habits like smoking. In this EDA, we reveal that:

1. Age at Death and Age at Start is highly related. The age of starting to have specific behavior potentially leads to lifespan decreases.
2. Smoking habits lead to lifespan reduction and various diseases.
3. Males are not genetically higher and heavier than females.
4. the top killer for people with high cholesterol and high blood pressure include CHD, cancer, and cerebral vascular disease.
5. People without smoking can enjoy longer lifespans and are unlikely to suffer from CHD disease at a young age.
6. The male does not generally have higher blood pressure compared to the female.

Following that, data pre-processing is carried out on the given dataset. Data preprocessing is a vital step in converting raw data into a clear and defined dataset that is useful for conducting data mining and analysis. It is necessary to preprocess data correctly as different input sources can affect data quality. Preprocessing can improve data accuracy, consistency, and reliability by removing missing or inconsistent data values. In this section, we have dealt with the incomplete data by filling in using the median and calculation. However, for those meaningful missing data, we have left them in the dataset. For noisy data, we apply winsorizing and trimming methods to remove outliers from the dataset. In the end, we try to find the inconsistent data by checking. However, there are no inconsistent data exists in this dataset.

6 References

American Heart Association. (2017, January 9). *Common Types of Heart Defects | American Heart Association*. Www.Heart.Org. https://www.heart.org/en/health-topics/congenital-heart-defects/about-congenital-heart-defects/common-types-of-heart-defects

Beaumont Street. (2023). *BP Calculator*. 28 Beaumont Street : NHS GP Surgery, Oxford. https://www.28beaumontstreet.co.uk/bp-calculator

Bhalla, D. (n.d.). *Complete Guide to PROC UNIVARIATE*. ListenData; ListenData. Retrieved February 17, 2023, from https://www.listendata.com/2016/07/complete-guide-to-proc-univariate.html

Bhandari, P. (2022, September 18). *What Is Interval Data? | Examples & Definition*. Scribbr. https://www.scribbr.co.uk/stats/interval-data-meaning/

Blog, F. (2019, October 23). *What is Interval Data? + [Examples, Variables & Analysis]*. Create Free Online Forms & Surveys in 2 Mins | Formplus; Formplus. https://www.formpl.us/blog/interval-data

Brazier, Y. (2021, November 16). *How much should I weigh for my height and age? BMI calculator & chart*. Medical and Health Information; Medical News Today. https://www.medicalnewstoday.com/articles/323446

CDC US. (2021, May 18). *High Blood Pressure Symptoms and Causes | cdc.gov*. Centers for Disease Control and Prevention. https://www.cdc.gov/bloodpressure/about.htm

Cleveland Clinic. (2022, July 28). *Understanding Cholesterol Levels and Numbers*. Cleveland Clinic. https://my.clevelandclinic.org/health/articles/11920-cholesterol-numbers-what-do-they-mean

Code, S. E. (2020, August 26). *How to Replace Missing Values in SAS - SAS Example Code*. SAS Example Code. https://sasexamplecode.com/replace-missing-values-in-sas/#Using_PROC_STDIZE

Code, S. E. (2021a, March 20). *3 Easy Ways to Calculate Percentiles in SAS (Examples)*. SAS Example Code. https://sasexamplecode.com/3-easy-ways-to-calculate-percentiles-in-sas-examples/

Code, S. E. (2021b, October 30). *3 Easy Ways to Calculate the Median in SAS - SAS Example Code*. SAS Example Code. https://sasexamplecode.com/3-easy-ways-to-calculate-the-median-in-sas/#Calculate_the_Median_of_Multiple_Variables-2

Dario, R. (2019, September 17). *Stop Using Mean to Fill Missing Data*. Towards Data Science. https://towardsdatascience.com/stop-using-mean-to-fill-missing-data-678c0d396e22

Inoue, T. (2004). Cigarette Smoking as a Risk Factor of Coronary Artery Disease and its Effects on Platelet Function. *Tobacco Induced Diseases*, *1*, 27. https://doi.org/10.1186/1617-9625-2-1-27

Isdiarti, N. (2021, June 24). *Data Cleaning: Inconsistent Data Entry | by Nindya Isdiarti | Nerd For Tech | Medium*. Medium; Nerd For Tech. https://medium.com/nerd-for-tech/data-cleaning-inconsistent-data-entry-7731ac3c52c7

IvyProSchool. (2022, October 31). *How to identify Inconsistent data using Excel? | Data Cleaning using Excel Ep 5 | IvyProSchool*. YouTube. https://www.youtube.com/watch?v=5EthgmXUY6c

Jim, F. (2019). *Mean, Median, and Mode: Measures of Central Tendency*. Statistic by Jim. https://statisticsbyjim.com/basics/measures-central-tendency-mean-median-mode/

Luke, G. (2021, March 17). *The Correct Way to Average the Globe*. Towards Data Science. https://towardsdatascience.com/the-correct-way-to-average-the-globe-92ceecd172b7

Magaga, A. M. (2021, April 5). *Identifying, Cleaning and replacing outliers | Titanic Dataset | by Alamin Musa Magaga | Analytics Vidhya | Medium*. Medium; Analytics Vidhya. https://medium.com/analytics-vidhya/identifying-cleaning-and-replacing-outliers-titanic-dataset-20182a062893

Myrna B. Schnur, MSN, RN. (2017, August 23). *BMI vs BSA Formulas: What's the Difference? | NursingCenter*. Lippincott NursingCenter | Professional Development for Nurses. https://www.nursingcenter.com/ncblog/august-2017/body-mass-index-and-body-surface-area-what-s-the-d

Onlinetuition.com. (2013, October 6). *7.1c Median - SPM Additional Mathematics*. SPM Additional Mathematics. https://spmaddmaths.blog.onlinetuition.com.my/2013/10/7-1c-median.html

P. Simopoulos, A. (1986). OBESITY AND BODY WEIGHT STANDARDS . *Annu. Rev. Public Health*. https://www.annualreviews.org/doi/pdf/10.1146/annurev.pu.07.050186.002405

Protobi. (2018, August 1). *Protobi*. Protobi. https://www.protobi.com/post/extreme-values-winsorize-trim-or-retain

Roy, B. (2019, September 4). *All About Missing Data Handling*. https://towardsdatascience.com/all-about-missing-data-handling-b94b8b5d2184

Saints, G., Millia, R., Palazzolo, G., Ibba, G., Marongiu, E., Roberto, S., Pinna, V., Ghiani, G., Tocco, F., & Crisafulli, A. (2016, August 5). *Mean Blood Pressure Assessment during Post-Exercise: Result from Two Different Methods of Calculation - PMC*. PubMed Central (PMC); National Library of Medicine National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4974855/

SAS. (2013). *SAS/STAT® 13.1User'sGuide The STDIZE Procedure*. SAS. https://support.sas.com/documentation/onlinedoc/stat/131/stdize.pdf

SAS. (2020, March 18). *SAS Help Center*. SAS Help Center. https://documentation.sas.com/doc/en/emref/15.1/p0hcukglth6mm2n104elqzvyztfz.htm

Wicklin, R. (2020, August 10). *4 ways to standardize data in SAS - The DO Loop*. The DO Loop. https://blogs.sas.com/content/iml/2020/08/10/standardize-data-sas.html

Zach. (2021, January 4). *How to Find Outliers Using the Interquartile Range*. Statology. https://www.statology.org/find-outliers-with-iqr/