Exploratory Data Analysis of the Heart Disease Prediction Dataset

The Heart Disease Prediction dataset has been downloaded from the Kaggle website. It is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The dataset consists of more than 4000 observations and 16 attributes: sex, age, education, smoking status, number of cigarettes per day, BP meds (whether a patient is on blood pressure medication), Prevalent Stroke, Prevalent hypertension, diabetes, total cholesterol level, systolic blood pressure, diastolic blood pressure, BMI, heart rate, glucose level, and ten years risk of CHD (coronary heart disease).

First, we uploaded and cleaned the data. We used Proc Contents function to check the general information about our dataset: how many observations and variables are in the set, variables' names, and types. After, we used Proc Means function to observe the descriptive statistics for variables. We looked into such parameters as mean, standard deviations, minimum, maximum. We used Proc Gchart to create a pie chart to visualize the distribution of 10-year CHD. Based on the pie chart, we determined that 557 patients had a 10-year CHD.

Based on the article "Gender differences in cardiovascular disease" (https://www.sciencedirect.com/science/article/pii/S2590093519300256), the category of gender plays an important role in studying cardiovascular diseases. So, we continued exploring our dataset by determining how many males and females are in our dataset. We used Proc Gchart vbar function to create a vertical bar graph to see the total number of patients by gender. Based on the graph, there are 2034 females and 1622 males. We continued our exploratory data analysis by creating a vertical bar graph to demonstrate a 10-year CHD by gender. We see that the proportion of patients with a 10-year CHD is higher among male patients.

Using Proc Corr function we created a correlation table, and we found out that there is no significant correlation between variables age, cigsperday, totchol, BMI, heartrate, glucose, and for sysBP and diaBP there is some positive correlation.

Smoking has been known to contribute to cardiovascular diseases (https://www.hopkinsmedicine.org/health/conditions-and-diseases/smoking-and-cardiovascular-disease), so we proceeded with exploring the data to determine what is the proportion of smokers. We found out that almost 50% of patients are smoking (1868 non-smokers vs 1788 smokers), and among the smokers, there are 807 females and 981 males. To illustrate the proportion of smokers by gender, we used a horizontal bar graph.

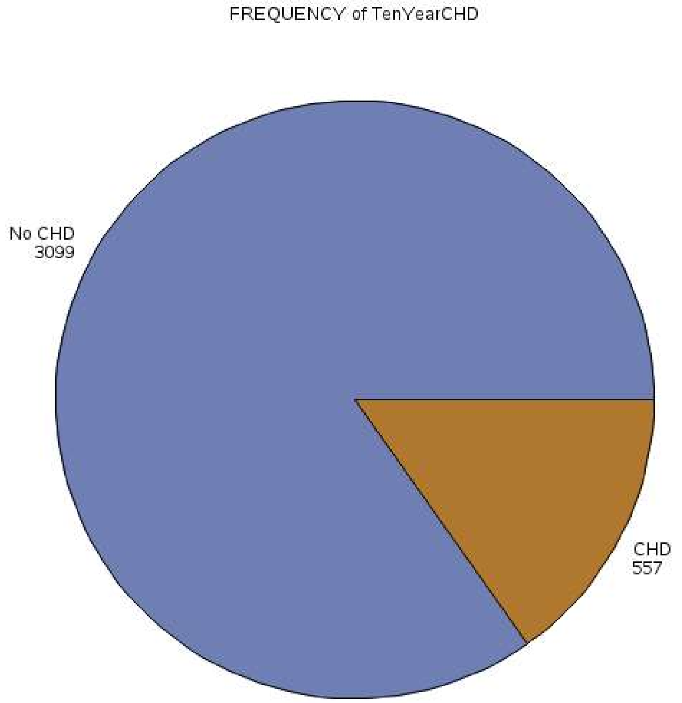
Another way to explore our dataset was to create a 2x2 frequency table. So, we looked at the relationship between prevalentHyp and TenYearCHD. Based on the table, if there is no prevalent hypertension, then the proportion of people not having CHD is approximately 89.15%; if there is prevalent hypertension, then the proportion of people having CHD is approximately 24.93%.

To find the relationship between BMI and 10-year CHD, we created a boxplot. And to find whether there is a difference between the mean BMI of patients with 10-year CHD and without CHD, we ran a t-test as well as Wilcoxon test. Both tests confirmed that the difference is significant; the mean BMI of patients with 10-year CHD is higher.

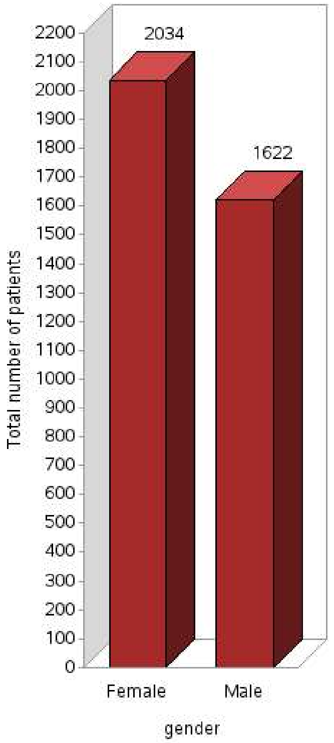
We wanted to explore the relationship between cigsPerDay of people with regard to their education levels. For this purpose, we created boxplots, and we observed that education level does not significantly affect smoking habits. Also, we created a scatter plot to find out the relationship between sysBP and diaBP pressure. There is a linear relationship between these two variables.

The last part of our exploratory data analysis was to create a histogram showing the distribution of age of people with respect to TenYearCHD. As we can see from the histogram, the mean age of patients without 10-year CHD is 48.71, and for those with 10-year CHD is 54.28.

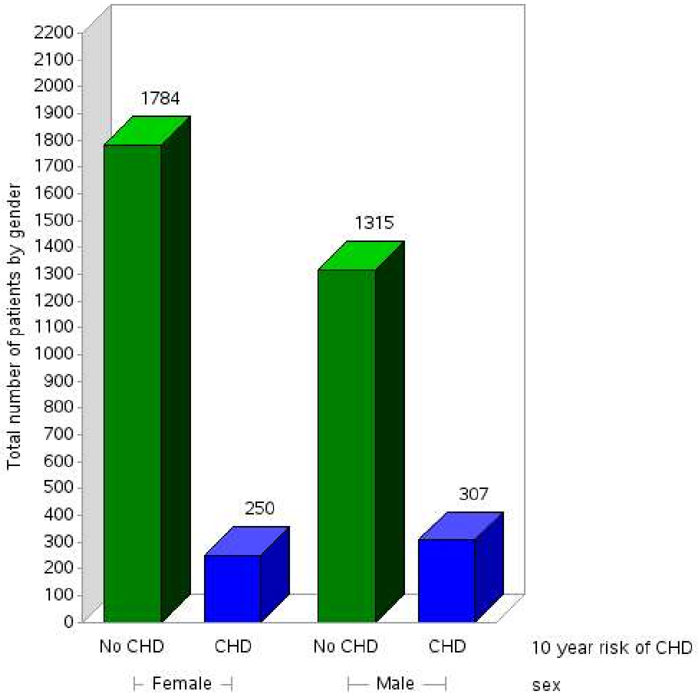
Pie chart of Distribution of TenYearCHD



Total number of patients by gender



10 year risk of coronary heart disease by gender



Correlations between age cigsperday totchol sysBP diaBP BMI heartRate glucose

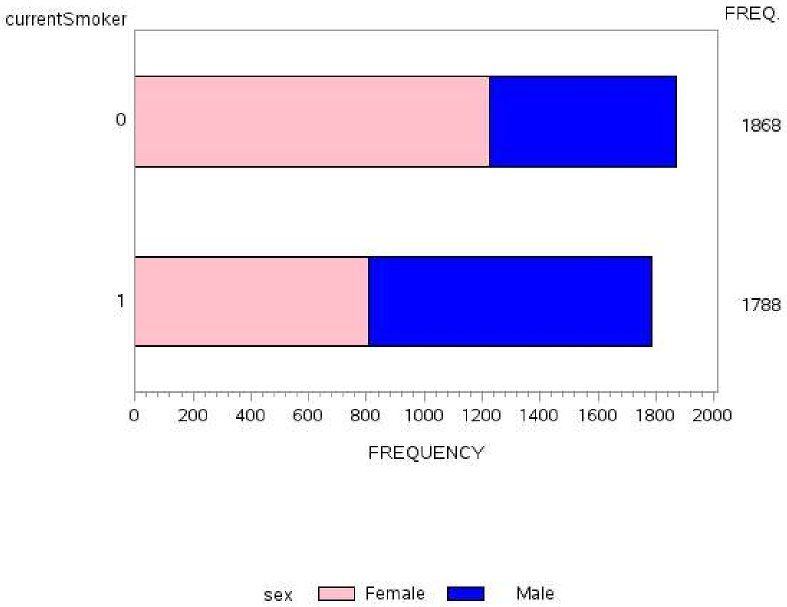
The CORR Procedure

8 Variables:	age cigsPerDay totChol sysBP diaBP BMI heartRate glucose

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
age	3656	49.55744	8.56113	181182	32.00000	70.00000
cigsPerDay	3656	9.02216	11.91887	32985	0	70.00000
totChol	3656	236.87309	44.09622	866008	113.00000	600.00000
sysBP	3656	132.36803	22.09244	483938	83.50000	295.00000
diaBP	3656	82.91206	11.97483	303127	48.00000	142.50000
BMI	3656	25.78418	4.06591	94267	15.54000	56.80000
heartRate	3656	75.73058	11.98295	276871	44.00000	143.00000
glucose	3656	81.85613	23.91013	299266	40.00000	394.00000

Pearson Correlation Coefficients, N = 3656 Prob > |r| under H0: Rho=0								
	age	cigsPerDay	totChol	sysBP	diaBP	BMI	heartRate	glucose
age	1.00000	-0.18910 <.0001	0.26776 <.0001	0.38855 <.0001	0.20888 <.0001	0.13717 <.0001	-0.00269 0.8711	0.11824 <.0001
cigsPerDay	-0.18910 <.0001	1.00000	-0.03022 0.0677	-0.09476 <.0001	-0.05665 0.0006	-0.08689 <.0001	0.06355 0.0001	-0.05380 0.0011
totChol	0.26776 <.0001	-0.03022 0.0677	1.00000	0.22013 <.0001	0.17499 <.0001	0.12080 <.0001	0.09306 <.0001	0.04975 0.0026
sysBP	0.38855 <.0001	-0.09476 <.0001	0.22013 <.0001	1.00000	0.78673 <.0001	0.33100 <.0001	0.18490 <.0001	0.13470 <.0001
diaBP	0.20888 <.0001	-0.05665 0.0006	0.17499 <.0001	0.78673 <.0001	1.00000	0.38561 <.0001	0.17901 <.0001	0.06370 0.0001
BMI	0.13717 <.0001	-0.08689 <.0001	0.12080 <.0001	0.33100 <.0001	0.38561 <.0001	1.00000	0.07440 <.0001	0.08367 <.0001
heartRate	-0.00269 0.8711	0.06355 0.0001	0.09306 <.0001	0.18490 <.0001	0.17901 <.0001	0.07440 <.0001	1.00000	0.09703 <.0001
glucose	0.11824 <.0001	-0.05380 0.0011	0.04975 0.0026	0.13470 <.0001	0.06370 0.0001	0.08367 <.0001	0.09703 <.0001	1.00000

Proportion of smokers by gender

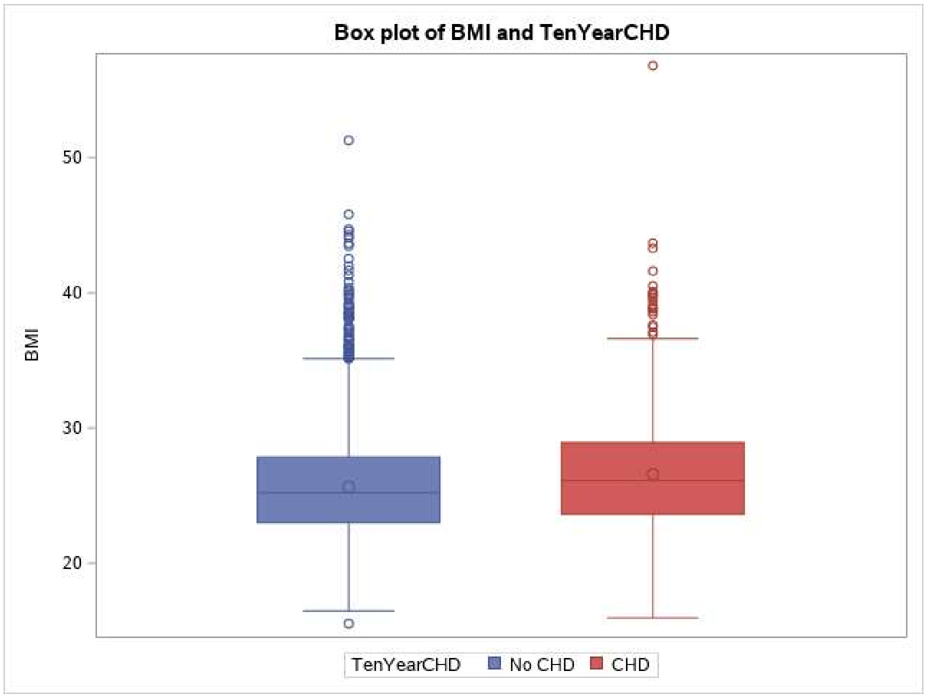


2X2 freq table of prevalentHyp vs TenYearCHD

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of prevalentHyp by TenYearCHD			
	TenYearCHD		
prevalentHyp	**0**	**1**	**Total**
0	2244 61.38 89.15 72.41	273 7.47 10.85 49.01	2517 68.85
1	855 23.39 75.07 27.59	284 7.77 24.93 50.99	1139 31.15
Total	3099 84.76	557 15.24	3656 100.00



T-test for BMI Run

The TTEST Procedure

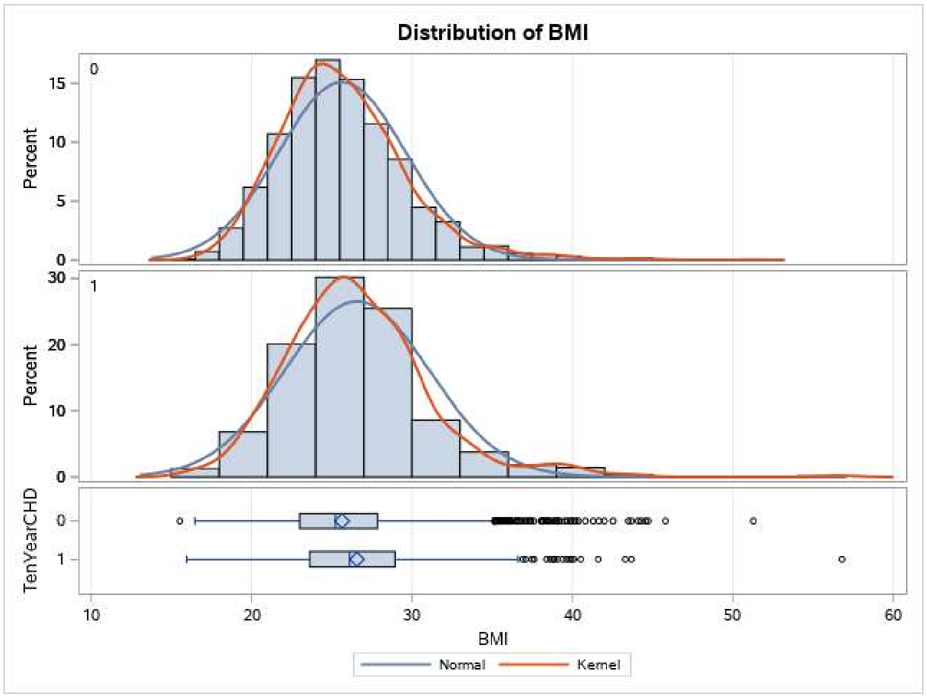
Variable: BMI

TenYearCHD	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		3099	25.6430	3.9653	0.0712	15.5400	51.2800
1		557	26.5698	4.5094	0.1911	15.9600	56.8000
Diff (1-2)	Pooled		-0.9269	4.0528	0.1865		
Diff (1-2)	Satterthwaite		-0.9269		0.2039		

TenYearCHD	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		25.6430	25.5033	25.7826	3.9653	3.8690	4.0666
1		26.5698	26.1945	26.9451	4.5094	4.2593	4.7911
Diff (1-2)	Pooled	-0.9269	-1.2926	-0.5612	4.0528	3.9620	4.1479
Diff (1-2)	Satterthwaite	-0.9269	-1.3272	-0.5265			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	3654	-4.97	<.0001
Satterthwaite	Unequal	718.79	-4.55	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	556	3098	1.29	<.0001



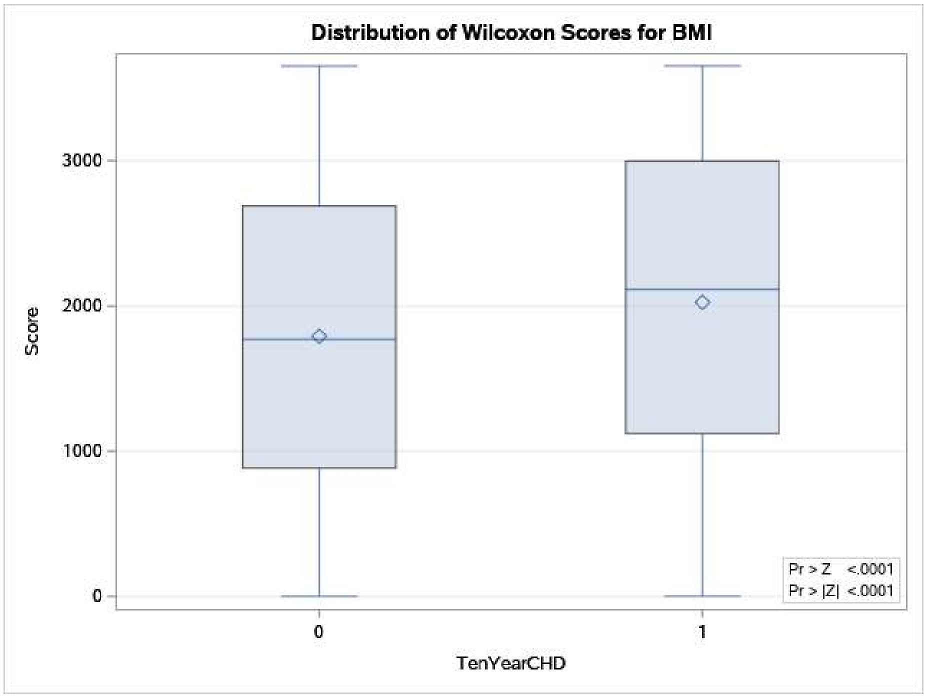
T-test for BMI Run

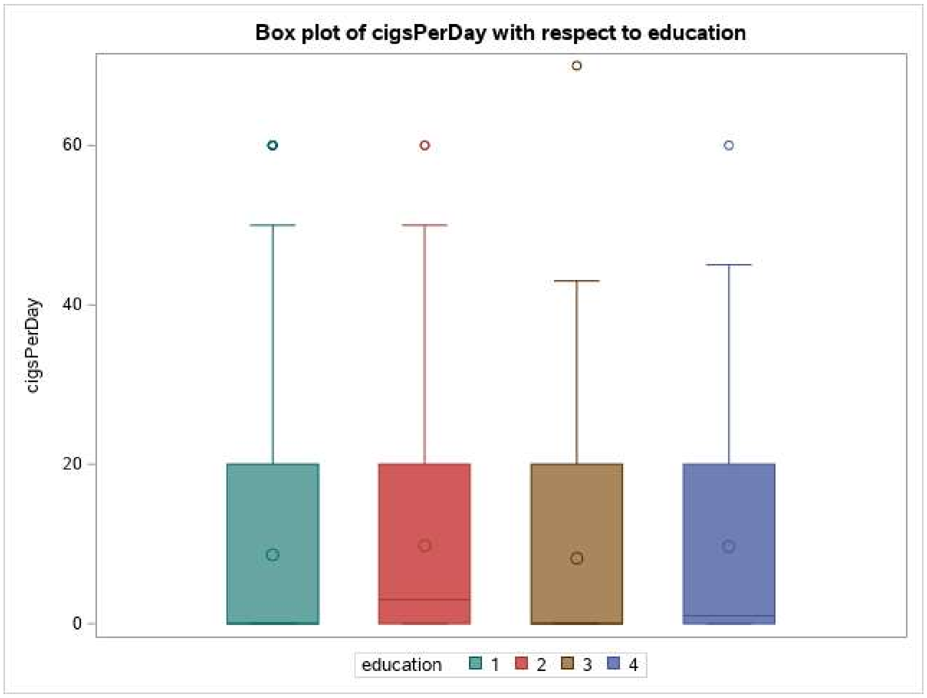
The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable BMI Classified by Variable TenYearCHD					
TenYearCHD	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	3099	5556423.50	5666521.50	22935.5800	1792.97306
1	557	1128572.50	1018474.50	22935.5800	2026.16248
Average scores were used for ties.					

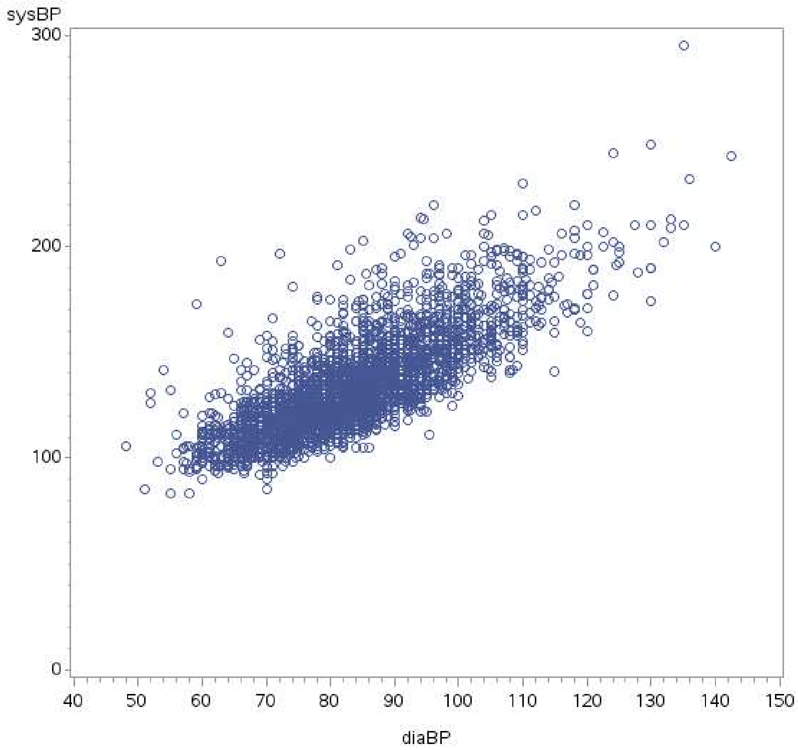
Wilcoxon Two-Sample Test					
				t Approximation	
Statistic	Z	Pr > Z	Pr > |Z|	Pr > Z	Pr > |Z|
1128573	4.8003	<.0001	<.0001	<.0001	<.0001
Z includes a continuity correction of 0.5.					

Kruskal-Wallis Test		
Chi-Square	DF	Pr > ChiSq
23.0430	1	<.0001


Distribution of Wilcoxon Scores for BMI
Score
3000
2000
1000
0
0
1
TenYearCHD
Pr > Z <.0001
Pr > |Z| <.0001

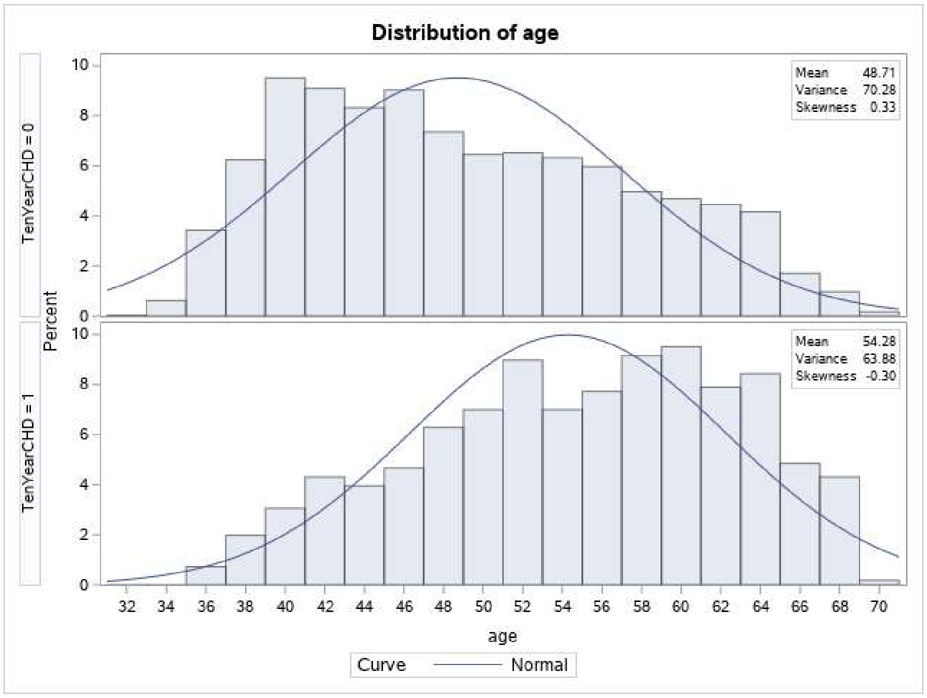

Box plot of cigsPerDay with respect to education
cigsPerDay
60
40
20
0
education 🟩 1 🟥 2 🟫 3 🟦 4

scatter plot of sysBP vs diaBP



Distribution of age

The UNIVARIATE Procedure



```
Proc Format;
Value sexft                1 = 'Male'
                            0 = 'Female'
                             Other   = 'Miscoded';
Value TenYearCHDft
                             1 = 'CHD'
                            0 = 'No CHD'
                             Other   = 'Miscoded';
run;
Proc Import datafile = "/home/u62112846/Homework/HeartDiseasePrediction.csv"
                                      out = HeartData
                DBMS = csv
               replace;
    getnames = yes;
    run;

    Data Heart_Data;
    set HeartData;
    glucose_new = input(glucose, BEST32.);
    drop glucose;
    rename glucose_new=glucose;
    run;

    Data Heart_clean1;
    set Heart_Data;
    if cmiss(of _all_) gt 0 then
     delete;
    run;

    *About Dataset variables;
    Proc Contents data = Heart_Data;
Run;

*Descriptive statistics of variables;
proc means data=heart_clean1;
run;

*Distribution of TenYearCHD;
option gstyle;
ods listing style=statistical;
goptions  hsize=15cm vsize=15cm;

Proc Gchart data = heart_clean1;
Title "Pie chart of Distribution of TenYearCHD";
pie TenYearCHD;
format TenYearCHD TenYearCHDft.;
run;

* total number of patients by gender;

pattern2 color=brown;
axis1 label=(a=90 "Total number of patients" ) minor=none order= (0 to 2200 by 100) offset=(0,0);
axis2 label=("gender");
Proc gchart data=heart_clean1;
Title "Total number of patients by gender";
vbar3d sex / discrete width=8 space=7 axis=axis1 maxis = axis2 outside=freq;
format sex sexft.;
run;

* plot gender against 10 year risk of CHD;

axis1 label=(a=90 "Total number of patients by gender" ) minor=none order= (0 to 2200 by 100) offset=(0,0);
axis2 label=("10 year risk of CHD");
proc gchart data=heart_clean1;
vbar3d TenYearCHD/discrete group=sex patternid=midpoint width=8 space=6 axis=axis1 maxis=axis2 outside=freq;
format sex sexft.;
format TenYearCHD TenYearCHDft.;
Title "10 year risk of coronary heart disease by gender ";
pattern1 c=green;
pattern2 c=blue;
run;
```

```
* correlation matrix;
proc corr data = Heart_clean1 plots=matrix;
title "Correlations between age cigsperday totchol sysBP diaBP BMI heartRate glucose ";
var age cigsperday totchol sysBP diaBP BMI heartRate glucose ;
run;

*smoking vs gender;
Proc Gchart data = heart_clean1;
hbar currentSmoker /discrete subgroup= sex width=6 space=6 outside=freq;
Title "Proportion of smokers by gender";
format sex sexft.;
 pattern1 c=pink;
 pattern2 c=blue;
run;

* freq table prevalentHyp & TenYearCHD;
proc freq data=heart_clean1;
Title "2X2 freq table of prevalentHyp vs TenYearCHD ";
table  prevalentHyp * TenYearCHD ;
run;

* box plot of BMI VS TenYearCHD;
proc sgplot data=heart_clean1;
Title "Box plot of BMI and TenYearCHD";
vbox BMI/group=TenYearCHD;
format TenYearCHD TenYearCHDft.;
run;

* T-test whether means of BMI of patients with CHD and without are equal or not   ;
Proc Ttest data=heart_clean1;
Class  TenYearCHD;
Var BMI;
Title " T-test for BMI"
Run;

* wilcoxon test;
Proc npar1way wilcoxon;
class TenYearCHD;
var BMI;
run;


* Distribution of cigsPerDay of people w.r.t their education levels;
proc sgplot data=heart_clean1;
Title "Box plot of cigsPerDay with respect to education";
vbox cigsPerDay/group=education grouporder=ascending;
run;



* scatter plot os sysBP vs diaBP;
 Symbol value = circle I = none;
Proc gplot data = heart_clean1;
Title "scatter plot of sysBP vs diaBP";
plot sysBP*diaBP;
run;

*Distribution of age of people w.r.t TenYearCHD;
proc univariate data=heart_clean1 noprint;
Title "Distribution of age";
class TenYearCHD;
histogram age/normal;
inset mean (6.2) var (7.2) skewness (6.2) / Pos = NE;
run;
```