KENNESAW STATE UNIVERSITY
COLLEGE OF SCIENCE AND MATHEMATICS
Department of Statistics and Analytical Sciences

STAT 3010- Final Project

- Read the guidelines for all graded written and video assignments in the D2L documentation and syllabus.
- Use all template headings and subheadings (note the descending level styles) in the order provided for your written report. Remove the section descriptions and insert information pertaining to your data.
- Light data preparation such as recoding, dealing with missing values and making subsets of the original data set can be done in EXCEL. However, ALL project tasks in sections 2 and 3 MUST be done using SAS or SAS Studio. Format tables and figures in a scholarly format a software like MS Word.
- Use the APPENDIX template to insert your annotated code.

Introduction

This dataset is originally called Byar and Greene Prostate Cancer Data and it is obtained from the website of the Vanderbilt University Department of Biostatistics. This topic is chosen because of personal interest for biostatistics hoping to learn more about how data analysis can help the healthcare industry to improve its efficiency in saving lives and maintaining the public health of the community. Specifically, this dataset can potentially contribute more to cardiovascular disease knowledge and in what way some factors in our lives including weight, blood pressure, etc. can affect the possibility of getting prostate cancer or related disease.

Statistical Question 1: Bivariate Analysis of Two Qualitative Variables

How does a patient's history of cardiovascular disease (hx; 0=No, 1=Yes) affect the results of an electrocardiogram test (ekg)?

Statistical Question 2: Bivariate Analysis of a Quantitative Variable and a Qualitative Variable

There has been some research suggesting that the serum level of hemoglobin in bone metastatic patients was statistically lower than those with non-bone metastases conditions. Is hemoglobin measure (hg) associated with the patient's history of bone metastases (bm) according to this dataset?

Statistical Question 3: Bivariate Analysis of Two Continuous Variables

Is there a linear relationship between weight index (wt) and systolic blood pressure (sbp)?

Section 1: Data set overview and preparation

1.1: Data set overview

Filename: prostate Extension: .csv

1.2: Data set summary

There was no information found on how or when the data were collected. There are 502 observations and 18 variables in the original dataset. The only change made was renaming the history of cardiovascular disease and bone metastases variables observations from 0, 1 to "Yes" and "No".

1.3: Data preparation methods

All 502 observations were used in this analysis. The chosen categorical variables are History of Cardiovascular (hx), Electrocardiogram Results (ekg), and Bone Metastases History (bm). The reason for choosing these variables is that they are clearer and more direct comparing to other categorical variables. For instance, the variable status has more than 9 different categories comparing to ekg which only has 4 categories, the big number of categories can make the analysis harder and repetitive. The used quantitative variables are Systolic blood pressure (sbp) and Weight Index (wt). Given that these two variables are the most familiar in our daily lives, for example, in our diets or lifestyle, they are known to have various effects on many things, not only heart disease. An analysis between these variables can help us acknowledge more the risk or benefits of these factors contributing our lives' quality.

1.4: Table of variable definitions

Table 1. Variables Used of Prostate.csv

Variable	Descriptive/Values	General Type	Specific Type	Measurement Units
wt	Weight Index (=weight(kg)-height(cm)+200) Integer = 69 to 152	Quantitative	Discrete	N/A
hx	History of Cardiovascular Disease (0 = No, 1 = Yes)	Qualitative	Nominal	N/A
sbp	Systolic blood pressure (Integer = 80 to 300)	Quantitative	Discrete	Millimeters of mercury (mmHg)
ekg	Electrocardiography results (normal, benign, rhythmic disturb & electrolyte ch, heart block or conduction def, heart strain, old MI, recent MI)	Qualitative	Nominal	N/A
hg	Serum hemoglobin (Real number interval = 5.899414 to 21.19922)	Quantitative	Continuous	grams/100 mL
bm	Bone Metastases History (0 = No, 1 = Yes)	Qualitative	Nominal	N/A

Section 2: Univariate descriptive statistics and visualizations

2.1: Ordinal variable

For your ordinal variable: a) provide an ordered table containing the frequency, relative frequency, cumulative frequency, and totals, b) generate a pie chart and an ordered bar chart, and c) discuss the distribution of the variable and whether the order of the categories reveals something about the trend.

Table 2.1: Contingency Table of Patient's EKG Test Results:

ekg	Frequency	Percent	Cumulative Frequency	Cumulative Percent
rhythmic disturb & electrolyte ch	51	10.32	51	10.32
recent MI	1	0.20	52	10.53
old MI	75	15.18	127	25.71
normal	168	34.01	295	59.72
heart strain	150	30.36	445	90.08
heart block or conduction def	26	5.26	471	95.34
benign	23	4.66	494	100.00
Frequency Missing = 8				

Figure 2.1: Ordered Bar Chart of Patients' EKG Results

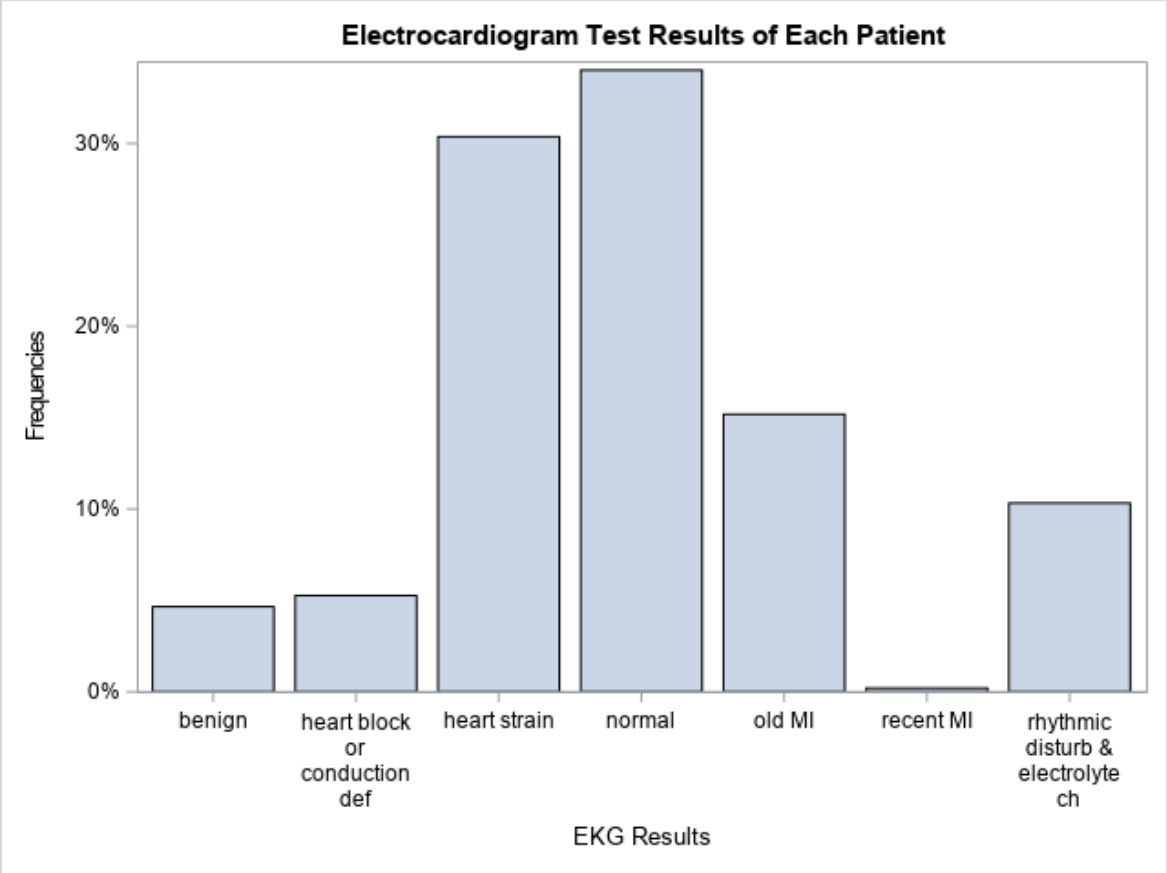
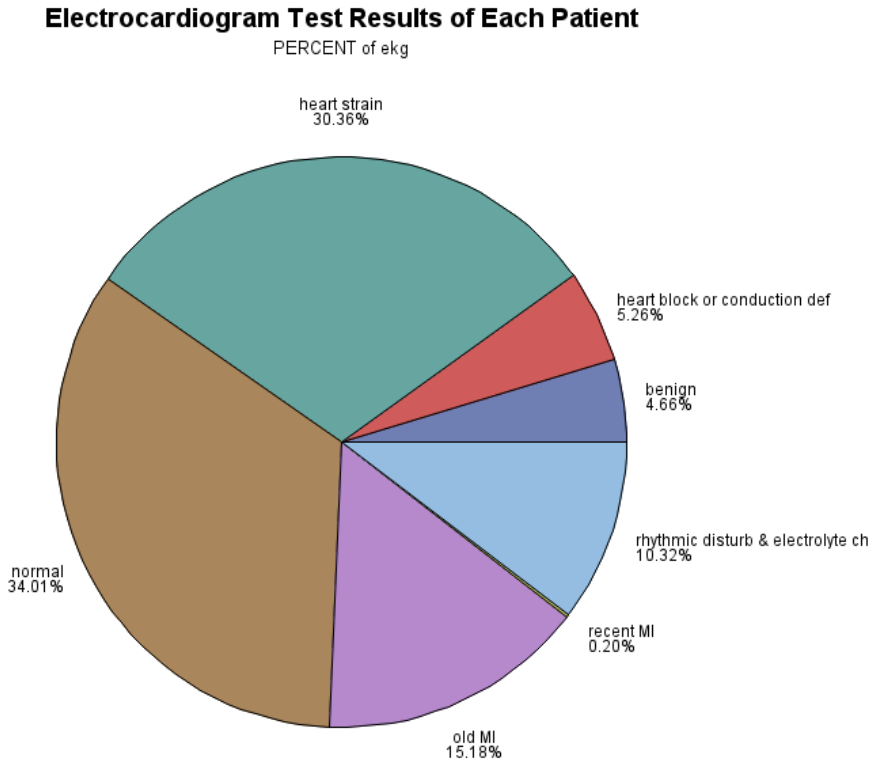


Figure 2.2: Pie Chart of Patients' EKG Results



Interpretation: Among the 502 observations, 168 patients, which contributes 34.01% of the total, have their EKG test results came back as "Normal", while 150 patients, which contributes 30.36% of the total, have "Heart strain" as the results. The lowest frequency is the number of patients having "recent myocardial infarction (MI)", which is 0.20%, as their test results.

Section 3: Multivariate descriptive statistics and visualizations

3.1: Bivariate Analysis of Two Qualitative Variables

Provide the following: a) a contingency table containing the frequency, relative frequency, row percent, and column percent of the groups b) a side by side bar chart, c) a 100% stacked bar chart of row percent, d) a response to the statistical question that uses evidence from the analysis as support.

Table 3.1.1: Frequencies of Patient's History of Cardiovascular Disease by Corresponding EKG:

	Corresponding EKG Results							
History of cardiovascular disease	Benign	Heart block or conduction def	Heart Strain	Normal	Old MI	Recent MI	Rhythmic Disturb and Electrolyte	Total
No	15	16	76	117	28	0	29	281
Yes	8	10	74	51	47	1	22	213
Total	23	26	150	168	76	1	51	494

Table 3.1.2: Relative Frequency of Patient's History of Cardiovascular Disease by Corresponding EKG:

	Corresponding EKG Results							
History of cardiovascular disease	Benign	Heart block or conduction def	Heart Strain	Normal	Old MI	Recent MI	Rhythmic Disturb and Electrolyte	Total
No	3.04	3.24	15.38	23.68	5.67	0	5.87	56.88
Yes	1.62	2.02	14.98	10.32	9.51	0.20	4.45	43.12
Total	4.66	5.26	30.36	34.01	15.18	0.20	10.32	100.00

Table 3.1.3: Row Percent of Patient's History of Cardiovascular Disease by Corresponding EKG:

	Corresponding EKG Results						
History of cardiovascular disease	Benign	Heart block or conduction def	Heart Strain	Normal	Old MI	Recent MI	Rhythmic Disturb and Electrolyte
No	5.34	5.69	27.05	41.64	9.96	0	10.32
Yes	3.76	4.69	34.74	23.94	22.07	0.47	10.33

Table 3.1.4: Column Percent of Patient's History of Cardiovascular Disease by Corresponding EKG:

	Corresponding EKG Results						
History of cardiovascular disease	Benign	Heart block or conduction def	Heart Strain	Normal	Old MI	Recent MI	Rhythmic Disturb and Electrolyte
No	65.22	61.54	50.67	69.64	37.33	0	56.86
Yes	34.78	38.46	49.33	30.36	62.67	100.00	43.14

Figure 3.1.1: Side-by-side Bar Charts of Patient's History of Cardiovascular Disease by Corresponding EKG:

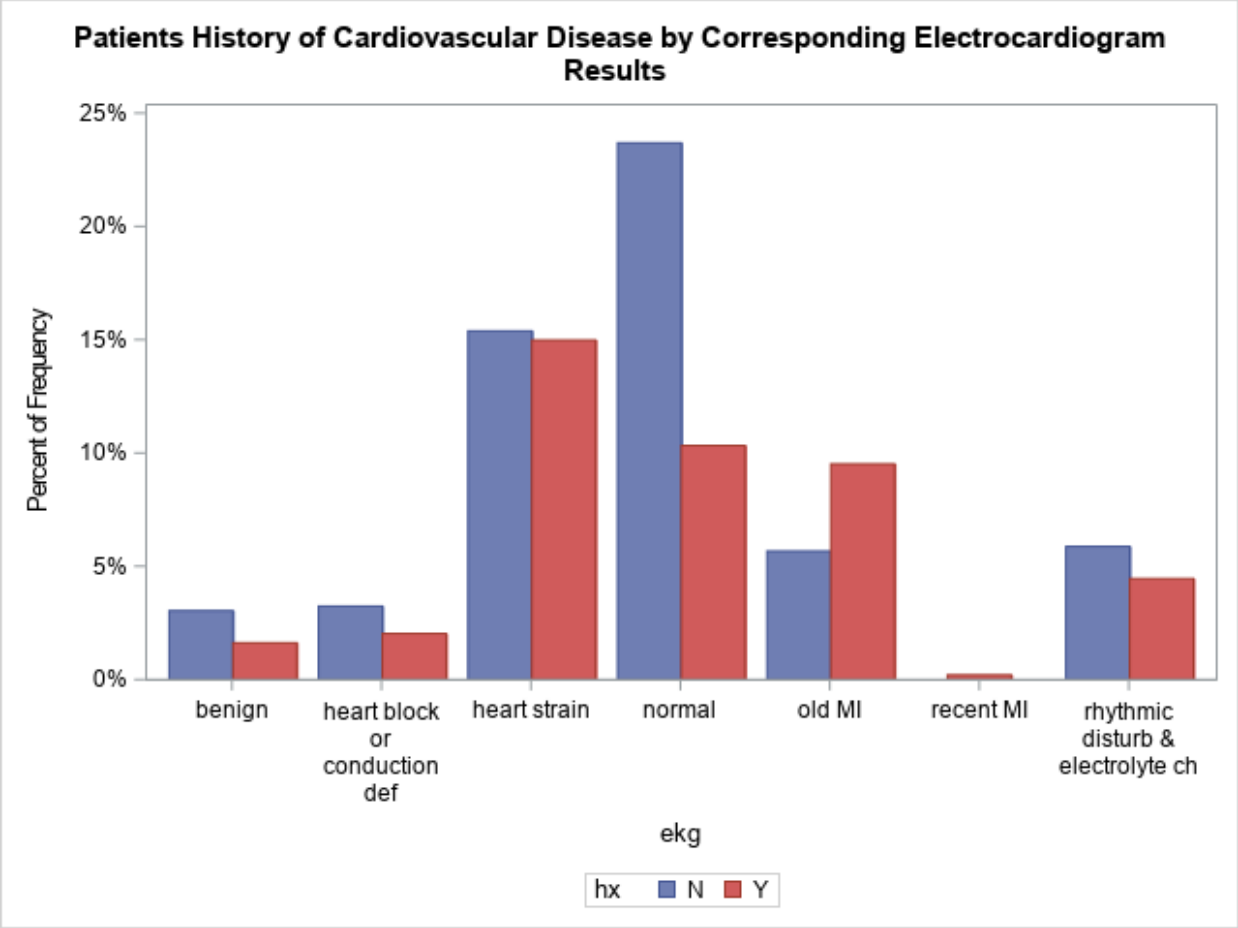
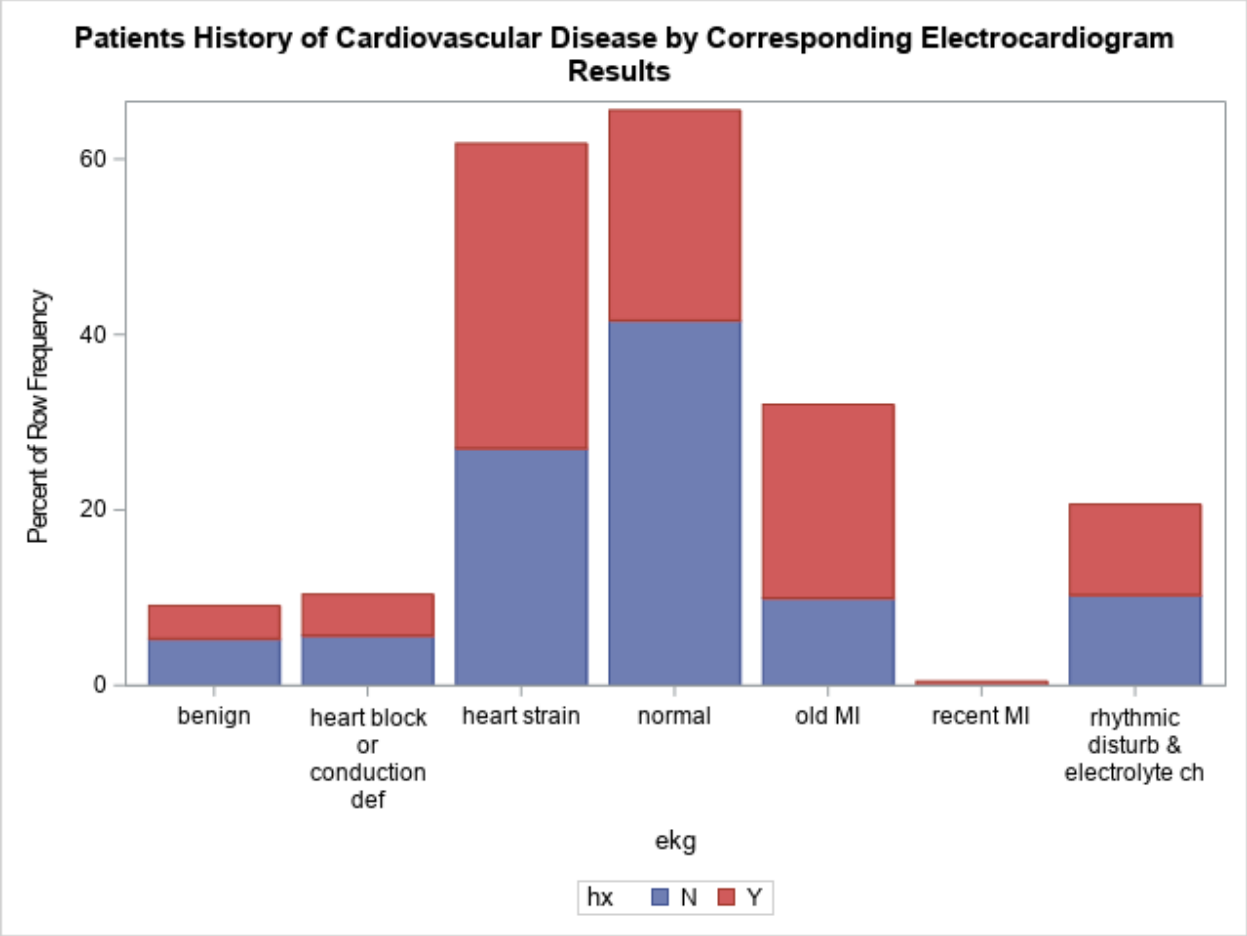


Figure 3.1.2: 100% Stacked Bar Chart Charts of Patient's History of Cardiovascular Disease by Corresponding EKG



Interpretation: According to the frequency table and generated bar charts, most of the patients indicated that they have a history of cardiovascular disease have more acute myocardial infarction as their results of the EKG while the ones with no history are likely to receive benign and normal as their results from the EKG. However, there was not much of a difference between the two groups when it comes to having a heart block, heart strain, and rhythmic disturbance. The frequency between the two only differed from 1 to 6. For example, there are 16 patients (3.24%) indicating they have not had any history of cardiovascular who received the EKG test results as having a heart block or cardiac conduction comparing to only 10 patients (2.02%) who have had history of cardiovascular disease received the same test results.

3.2: Bivariate Analysis of a Quantitative Variable and a Qualitative Variable

Provide the following: a) a table containing the descriptive statistics for each group, b) a side by side box-whiskers plot, c) the 95% confidence interval for the mean of each group and the interpretation of each CI in context, and e) a response to the statistical question that uses evidence from the analysis as support

Table 3.2.1: Descriptive Statistics of Patients' Serum Hemoglobin Measures by Corresponding History of Bone Metastases

	History of Bone Metastases	
	Yes	**No**
Count	82	420
Mean	12.258	13.679
Standard Deviation	2.331	1.781
Minimum	7.000	5.899
Q1	10.500	12.600
Median	12.649	13.799
Q3	13.799	14.799
Maximum	21.199	18.199
95% Confidence Intervals	(11.746, 12.770)	(13.508, 13.849)

Figure 3.2.1: Side-by-side boxplots and whisker of Patients' Serum Hemoglobin Measures by Corresponding History of Bone Metastases

Interpretation:

In Figure 3, the distribution for patients who answered Yes has more variance than the ones that answered No. However, the group with no history of bone metastases has more outliers compared to

the other. Both appears to be asymmetrical with some skewness to the left. Therefore, measure of central tendency for this data would be the median.

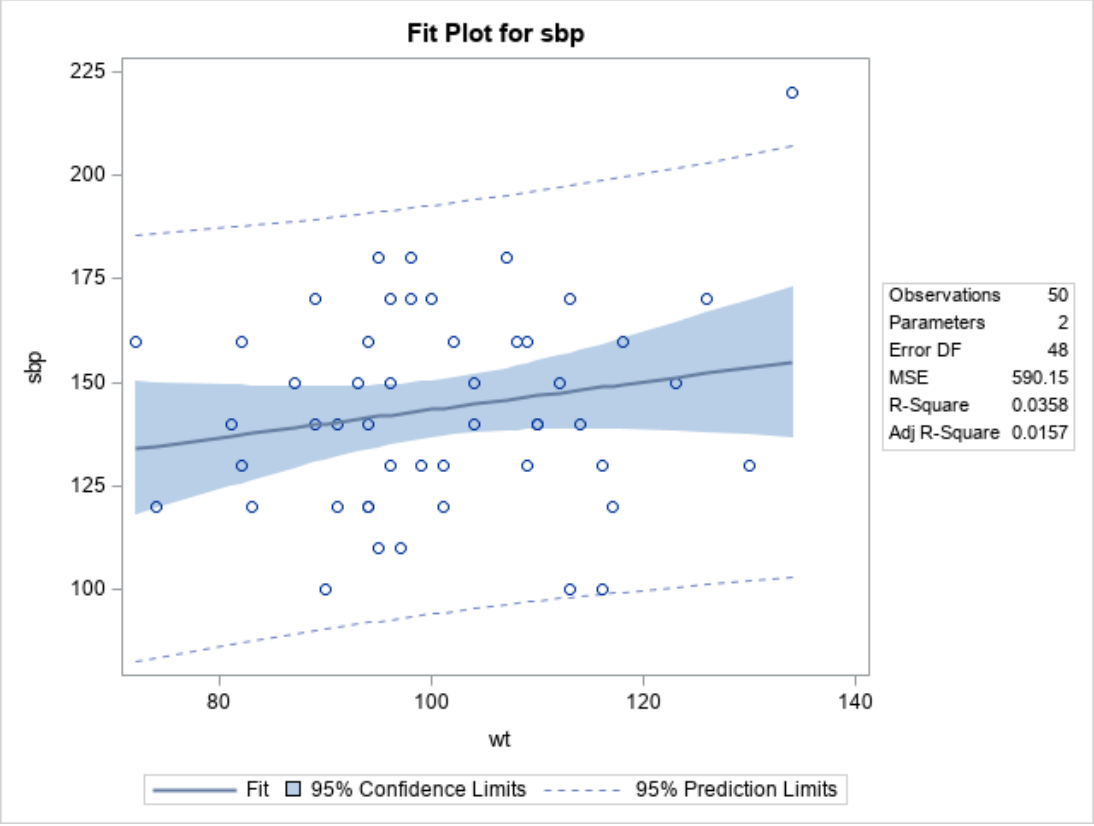
For the patients who answered “No”, the interquartile range has a breadth of 3.299 grams/100 mL. For the patients who answered “Yes”, the interquartile range has a breadth of 2.199 grams/ 100 mL. As seen from the descriptive statistics and figure, the median for the “No” group (13.799 grams/100 mL) is higher than the other group (12.649 grams/100 mL).

We are 95% confident that the true mean for the group that said “No” to having a history of bone metastases lies between 13.508 and 12.849 grams/100 mL while the true mean for the group that said “Yes” is between 11.746 to 12.770 grams/ 100 mL. Since the two confidence intervals do not overlap, there is a statistically significant difference between the two groups. To restate, there is a statistically significant difference in serum hemoglobin measures in each patient based on their indication on having any history of bone metastases.

3.3: Bivariate Analysis of Two Continuous Variables

Provide: a) a seed and draw an SRS where the sample size is equal to 10% of your data rows (round up if you get a decimal), b) a scatter plot with least squares regression line overlaid, c) a contextual interpretation of the correlation and of the slope, d) one prediction for the response variable using a value of the explanatory variable of your choosing and interpret the prediction in the context of the relationship, and e) a response to the statistical question that uses evidence from the analysis as support.

Figure 3.3.1: Scatter plot with least regression line of Systolic Blood Pressure by Weight Index Using SRS:



Interpretation:

After generating the scatterplot, the R-square value was only 0.0358 which indicates a very weak linear relationship with little correlation between the two variables. There were many outliers in the plot, and the data was widely spread. For instance, according to the scatter plot, a patient whose weight index is 90 will have a systolic blood pressure of 130. In conclusion, there is a small linear relationship between weight index and systolic blood pressure. The higher one's weight index, the higher their systolic blood pressure will be.

Section 4: Conclusion

To recap, in question 1, we learned that patients who do not have history of cardiovascular disease are more likely to receive the results from EKG Test indicating their tumor is benign, normal, or no myocardial infarction comparing to the ones that answered "Yes". From question 2, it is concluded that one's serum hemoglobin measures differ significantly when it comes to whether having a history of bone metastases. Lastly, unfortunately, according to the data in question 3, we cannot find any linear relationship between weight index and systolic blood pressure since the correlation and R squared value is too low. For future plans, I have not decided on how to make more use out of this topic. However, I think it would be interesting to use this dataset for further research on how to prevent prostate cancer and determine besides the variables recorded, would there be any other significant factor that can be added to.

APPENDIX

Annotated SAS Code

Section 1

```
/*Create a library connection*/

libname Ksu "C:\Users\Meggin\Desktop\SAS";

/*Import the data file and store it in the library*/

proc import datafile='C:\Users\Meggin\Desktop\STAT\prostate.csv'

out=Ksu.prostate dbms=csv replace;

guessingrows=20;

getnames=yes;

run;

/*View the data*/

proc contents data=Ksu.prostate;

run; *Variable attributes;

proc print data=Ksu.prostate (obs=25);

run; *Print first 25 rows;

/*Renaming the yes no variables*/

proc format;

value yn

0='N'

1='Y';

run;

data Ksu.prostate;

drop tmp_:;
```

```
set Ksu.prostate (rename=(hx=tmp_hx));

hx=put(tmp_hx,yn.);

run;
```

Section 2

```
/**Univariate Analysis-Qualitative***/

/*Table ordered as in data for ekg*/

proc sort data= Ksu.prostate;

by descending ekg;

run;

proc freq data=Ksu.prostate order=data; *Table in descending order;

tables ekg;

run;

proc sgplot data=Ksu.prostate; *Vertical bar chart-relative frequency;

title 'Electrocardiogram Test Results of Each Patient';

vbar ekg/stat=percent;

xaxis label='EKG Results';

yaxis label='Frequencies';

run;

proc gchart data=Ksu.prostate;

pie ekg/type=pct;

legend;

run;
```

Section 3

3.1: Bivariate Analysis of Two Qualitative Variables

```
/**Bivariate-Two Categorical Variables**/

*Contigency table of summary statistics;

proc freq data=Ksu.prostate nlevels; *Introduce NLEVELS option;

table hx*ekg; *Row by column;

run;
```

```
*Side by side bar chaty-relative frequency;
proc sgplot data=Ksu.prostate;
vbar ekg/group=hx groupdisplay=cluster stat=percent;
title 'Patients History of Cardiovascular Disease by Corresponding Electrocardiogram Results';
run;
*Method for 100% stacked bar chart;
*1-Write the frequencies to a new data set;
proc freq data=Ksu.prostate;
tables hx*ekg;
ods output crosstabfreqs= Ksu.categoricals; *Write output data set of counts;
run;
*2-Extract the row frequencies into a new data set;
data Ksu.categoricals2;
set Ksu.categoricals (drop=table);
if not missing (rowpercent);
run;
*3-Generate the plot;
proc sgplot data= Ksu.categoricals2;
vbar ekg/group=hx groupdisplay=stack response=rowpercent;
title "Patients History of Cardiovascular Disease by Corresponding Electrocardiogram Results";
run;
```

3.2: Bivariate Analysis of a Quantitative Variable and a Qualitative Variable

```
/**Bivariate-Quantitative by qualitative**/
*Numerical sumaries by group with CI;
data Ksu.prostate;
drop tmp_:;
set Ksu.prostate (rename=(bm=tmp_bm));
bm=put(tmp_bm,yn.);
run;
```

```
proc means data=Ksu.prostate n mean std min q1 median q3 max nmiss clm alpha=0.05 maxdec=3;
var hg;
class bm;
run;
*Side by side boxplots;
proc sgplot data=Ksu.prostate;
vbox hg/category=bm;
title"Serum homoglobin measures of each patient by their corresponding histories of bone metastases";
yaxis label="Serum Hemoglobin Measures (grams/100mL)";
run;
*Histograms by group;
proc sort data=Ksu.prostate;
by bm;
run;
proc sgplot data=Ksu.prostate;
histogram hg;
by bm;
xaxis label="Serum Hemoglobin Measures (grams/100mL)";
run;
```

3.3: Bivariate Analysis of Two Continuous Variables

```
/*Two quantitative variables*/
/*Simple Random Sample*/
proc surveyselect data=Ksu.prostate out=Ksu.srsprostate method=srs samplesize=50 seed=502;
run;
*Correlation;
proc corr data= Ksu.srsprostate nomiss outp=Ksu.CorrOutp; *outp=creates a data set of the output;
var sbp wt;
run;
proc print data=Ksu.CorrOutp noobs;
```

```
run;
*Scatterplot and regression line coefficients;
proc reg data=Ksu.srsprostate alpha=0.05 plots(only)=(residuals fitplot);
model sbp=wt; *Form y=x;
run;
quit;
```