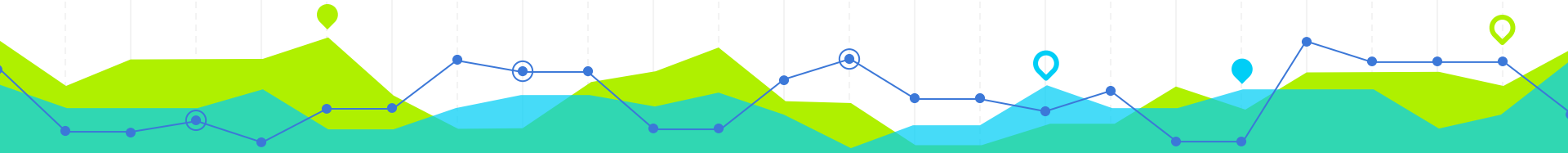
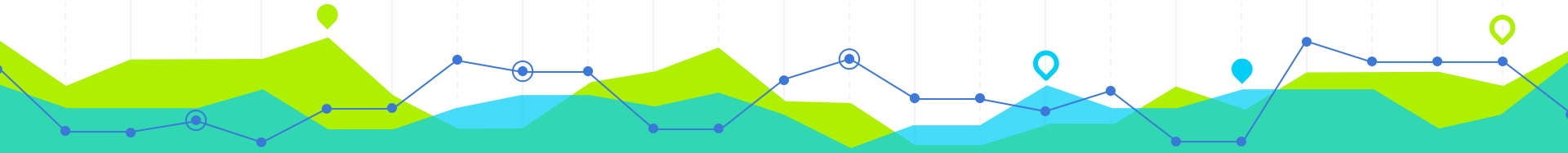


Data Analysis Group A

Contents

- Dataset description
- Questions
- Data Cleaning (5 steps)
- Descriptive Analysis
  - Frequency
  - Central Tendency & Dispersion
- Crosstabs
  - Question 1 & 2
  - Question 3
  - Question 4



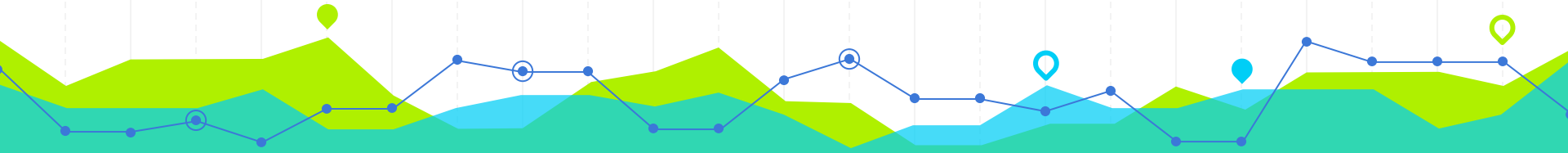
Data Set Description

1

Heart Disease Prediction

Heart Disease Prediction Dataset

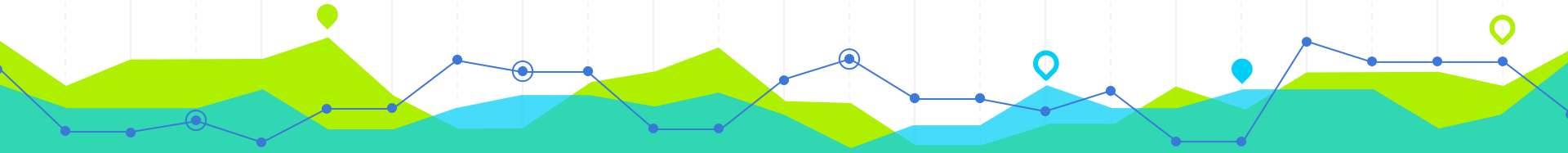
- 14 variables including Age, Sex, BP, etc…
- 270 rows of data
- The variables consist of a mix of numerical and categorical variables



Questions 2

Questions

- Question 1: Highest prevalence of heart disease between sexes?
- Question 2: Is there an association between heart disease and chest pain type?
- Question 3: Does the presence of exercise angina affect that of heart disease?
- Question 4: What's the distribution between the EKG results and heart disease



Data Cleaning

3

5 Steps of Data Cleaning

Data Cleaning Steps

Step one: Remove irrelevant data

- None

Step two: Deduplicate the data

- There were no duplicates

Step three: Fix structural errors

- None

Step 4: Filter out data outliers

- Out-of-range values:
  a. Gender has two anomalies, 11 and 10, while the values should be 1 or 0 (male or female). 11 is clearly an entry mistake that was meant to be 1, but we wouldn't be able to know what 10 is, so it was removed from the dataset.
  b. Chest pain type is supposed to have 4 categories, but there is one anomaly (44) that was changed to 4.
  c. Max HR has two anomalies (1154 and 1380), since it is impossible for a human to have such heart rates. These values were removed from the data.

Step 5: Deal with missing data

1. **The univariate statistics analysis** output using SPSS shows the number and percentage of missing values for each variable:

Univariate Statistics

	N	Mean	Std. Deviation	Missing Count	Missing Percent	No. of Extremes[a] Low	No. of Extremes[a] High
BP	255	131.09	18.116	15	5.6	0	11
Cholesterol	253	248.76	51.406	17	6.3	0	4
MaxHR	259	149.32	22.995	11	4.1	1	0
STdepression	266	1.042	1.1416	4	1.5	0	6
Age	270	54.43	9.109	0	.0	0	0
Sex	270			0	.0		
Chestpaintype	270			0	.0		
FBSover120	270			0	.0		
EKGresults	270			0	.0		
Exerciseangina	270			0	.0		
SlopeofST	270			0	.0		
Numberofvesselsfluro	270			0	.0		
Thallium	270			0	.0		
HeartDisease	270			0	.0		

a. Number of cases outside the range (Q1 - 1.5*IQR, Q3 + 1.5*IQR).

Step 5: Deal with missing data

The number of nonmissing values for each variable appears in the N column, and the number of missing values appears in the Missing Count column. The Missing Percent column displays the percentage of cases with missing values and provides a good measure for comparing the extent of missing data among variables.

The recommended percentage of missing values is 5 percent or less. Therefore the variables **‘Cholestrol’ and ‘BP’** will require further analysis.

Univariate Statistics

	N	Mean	Std. Deviation	Missing Count	Missing Percent	No. of Extremes[a] Low	No. of Extremes[a] High
BP	255	131.09	18.116	15	5.6	0	11
Cholesterol	253	248.76	51.406	17	6.3	0	4
MaxHR	259	149.32	22.995	11	4.1	1	0
STdepression	266	1.042	1.1416	4	1.5	0	6
Age	270	54.43	9.109	0	.0	0	0
Sex	270			0	.0		
Chestpaintype	270			0	.0		
FBSover120	270			0	.0		
EKGresults	270			0	.0		
Exerciseangina	270			0	.0		
SlopeofST	270			0	.0		
Numberofvesselsfluro	270			0	.0		
Thallium	270			0	.0		
HeartDisease	270			0	.0		

a. Number of cases outside the range (Q1 - 1.5*IQR, Q3 + 1.5*IQR).

Step 5: Deal with missing data

1. **Separate variance t-tests:**
   The separate-variance t tests table can help to identify variables whose pattern of missing values may be influencing the quantitative (scale) variables. The t test is computed using an indicator variable that specifies whether a variable is present or missing for an individual case. The subgroup means for the indicator variable are also tabulated. The second output gives the results of the t test conducted for all the quantitative variables using indicator variables with more than 5% missing values.

Separate Variance t Tests[a]

		BP	Cholesterol	MaxHR	STdepression	Age
BP	t	.	-.3	1.2	-1.9	-.9
	df	.	13.6	15.5	14.9	15.3
	# Present	255	240	244	251	255
	# Missing	0	13	15	15	15
	Mean(Present)	131.09	248.58	149.76	1.000	54.30
	Mean(Missing)	.	252.08	142.20	1.740	56.67
Cholesterol	t	-.5	.	.0	-.3	-.3
	df	15.2	.	16.8	16.7	19.0
	# Present	240	253	243	250	253
	# Missing	15	0	16	16	17
	Mean(Present)	130.92	248.76	149.33	1.036	54.39
	Mean(Missing)	133.87	.	149.19	1.138	55.06

For each quantitative variable, pairs of groups are formed by indicator variables (present, missing).

a. Indicator variables with less than 5% missing are not displayed.

Step 5: Deal with missing data

Interpretation:

- When BP is missing, the mean of Cholesterol is 252.08, as opposed to 248.58 when BP is present. This could be considered as a minimal difference that is likely due to chance, which also applies to the remaining variables. This indicates that BP may be **missing completely at random (MCAR).**
- When Cholesterol is missing, all the numerical variables undergo an insignificant change with their respective means. This indicates that Cholestrol may also be **missing completely at random (MCAR).**

Separate Variance t Tests[a]

		BP	Cholesterol	MaxHR	STdepression	Age
BP	t	.	-.3	1.2	-1.9	-.9
	df	.	13.6	15.5	14.9	15.3
	# Present	255	240	244	251	255
	# Missing	0	13	15	15	15
	Mean(Present)	131.09	248.58	149.76	1.000	54.30
	Mean(Missing)	.	252.08	142.20	1.740	56.67
Cholesterol	t	-.5	.	.0	-.3	-.3
	df	15.2	.	16.8	16.7	19.0
	# Present	240	253	243	250	253
	# Missing	15	0	16	16	17
	Mean(Present)	130.92	248.76	149.33	1.036	54.39
	Mean(Missing)	133.87	.	149.19	1.138	55.06

For each quantitative variable, pairs of groups are formed by indicator variables (present, missing).

a. Indicator variables with less than 5% missing are not displayed.

Step 5: Deal with missing data

3. Crosstabulations of Categorical Versus Indicator Variables

The crosstabulations of categorical variables versus indicator variables show information similar to that found in the separate-variance t test table. Indicator variables are once again created, except this time they are used to calculate frequencies in every category for each categorical variable. The values can help you determine whether there are differences in missing values among categories.

Sex

			Total	Female	Male
BP	Present	Count	255	81	174
		Percent	94.4	94.2	94.6
	Missing	% SysMis	5.6	5.8	5.4
Cholesterol	Present	Count	253	82	171
		Percent	93.7	95.3	92.9
	Missing	% SysMis	6.3	4.7	7.1

Indicator variables with less than 5% missing are not displayed.

Step 5: Deal with missing data

3. Crosstabulations of Categorical Versus Indicator Variables

The crosstabulations of categorical variables versus indicator variables show information similar to that found in the separate-variance t test table. Indicator variables are once again created, except this time they are used to calculate frequencies in every category for each categorical variable. The values can help you determine whether there are differences in missing values among categories.

Chestpaintype

			Total	1	2	3	4
BP	Present	Count	255	19	39	78	119
		Percent	94.4	95.0	92.9	98.7	92.2
	Missing	% SysMis	5.6	5.0	7.1	1.3	7.8
Cholesterol	Present	Count	253	19	38	75	121
		Percent	93.7	95.0	90.5	94.9	93.8
	Missing	% SysMis	6.3	5.0	9.5	5.1	6.2

Indicator variables with less than 5% missing are not displayed.

Step 5: Deal with missing data

3. Crosstabulations of Categorical Versus Indicator Variables

The crosstabulations of categorical variables versus indicator variables show information similar to that found in the separate-variance t test table. Indicator variables are once again created, except this time they are used to calculate frequencies in every category for each categorical variable. The values can help you determine whether there are differences in missing values among categories.

Chestpaintype

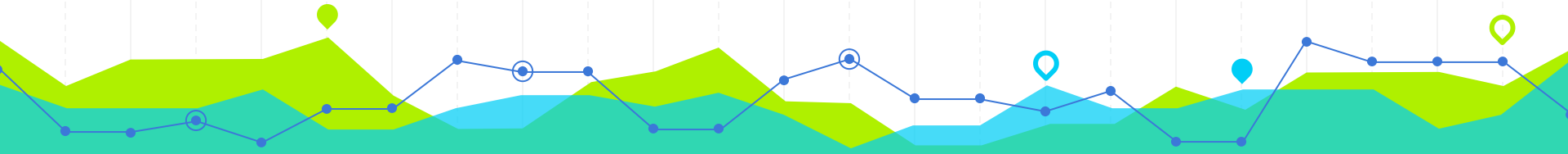
			Total	1	2	3	4
BP	Present	Count	255	19	39	78	119
		Percent	94.4	95.0	92.9	98.7	92.2
	Missing	% SysMis	5.6	5.0	7.1	1.3	7.8
Cholesterol	Present	Count	253	19	38	75	121
		Percent	93.7	95.0	90.5	94.9	93.8
	Missing	% SysMis	6.3	5.0	9.5	5.1	6.2

Indicator variables with less than 5% missing are not displayed.

Step 5: Deal with missing data

The results of the analysis show that the discrepancies between variables when the indicator variables are missing or present are likely due to chance. The absence of some values is therefore **independent of both the observed and unobserved data (MCAR).**

Considering the results of the missing value analysis, pairwise deletion will be implemented, which deletes cases only if the data missing are required for the analysis being conducted.

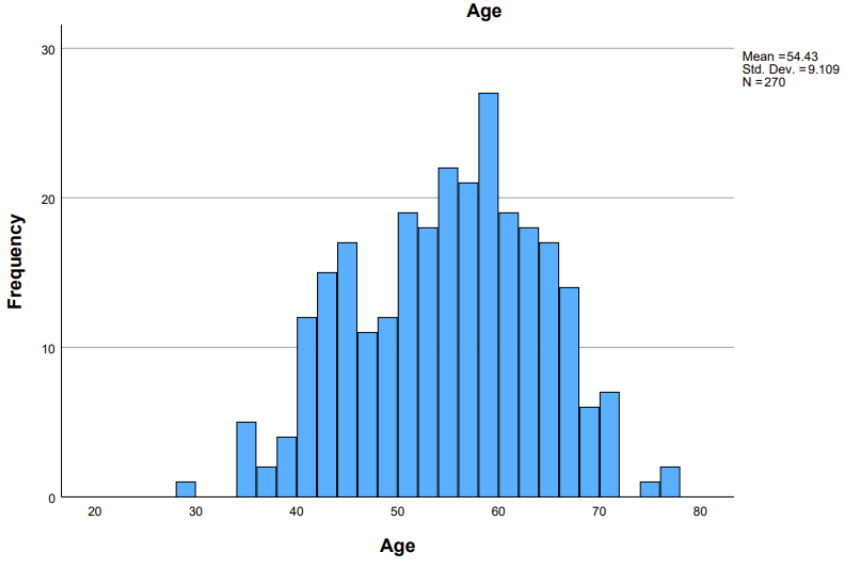


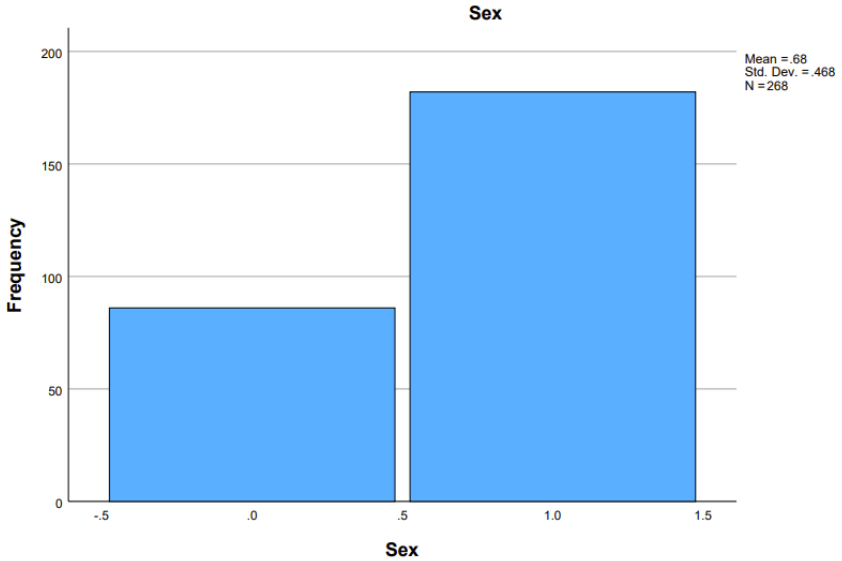
Descriptive Analysis

4

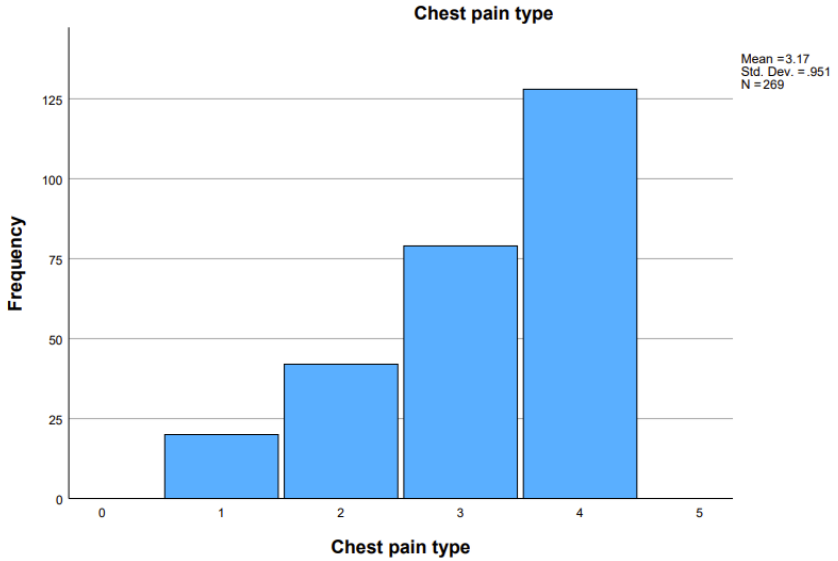
Frequency, Central Tendency, and Dispersion

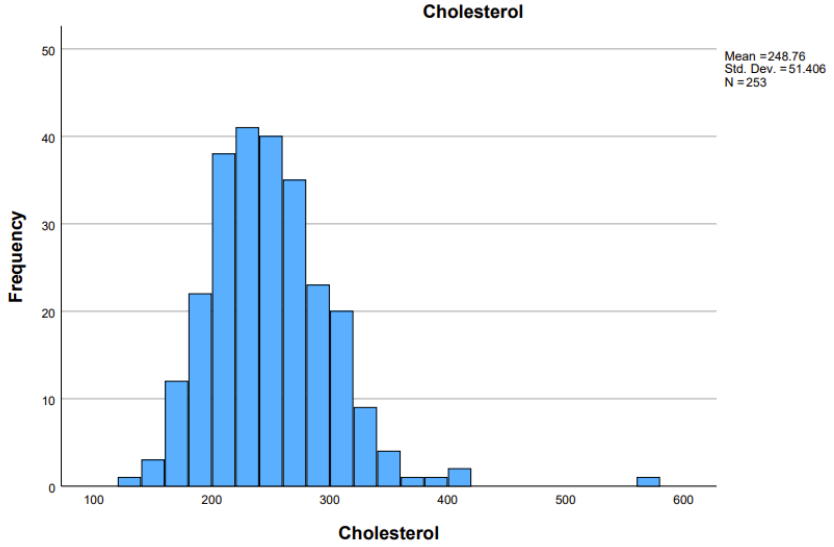
Frequency


Age
Mean =54.43
Std. Dev. =9.109
N =270
Frequency
30
20
10
0
20
30
40
50
60
70
80
Age

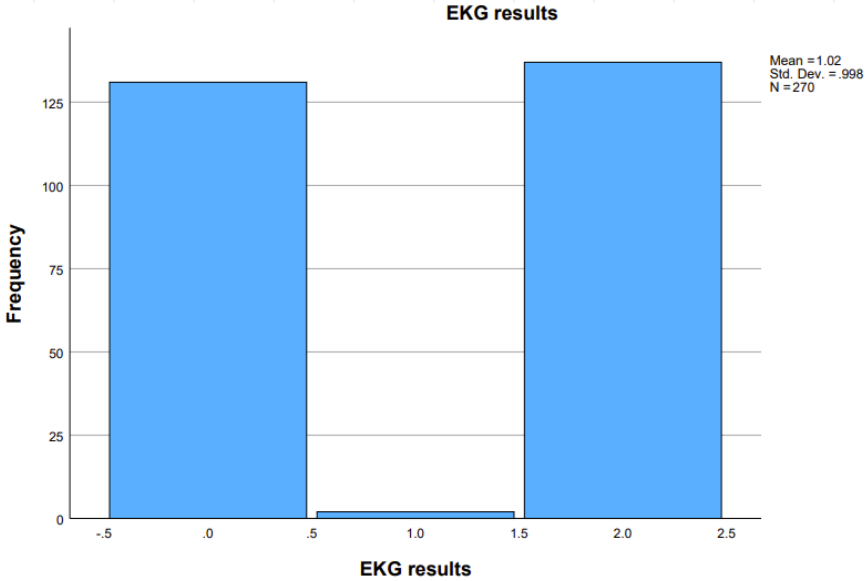

Sex
Mean =.68
Std. Dev. =.468
N =268
Frequency
200
150
100
50
0
-.5
.0
.5
1.0
1.5
Sex

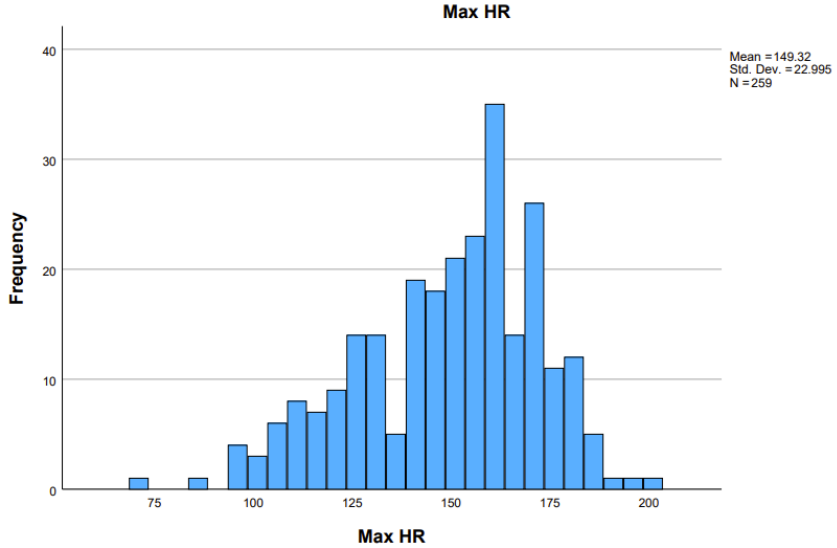
Frequency


Chest pain type
Mean =3.17
Std. Dev. =.951
N =269
Frequency
125
100
75
50
25
0
0
1
2
3
4
5
Chest pain type

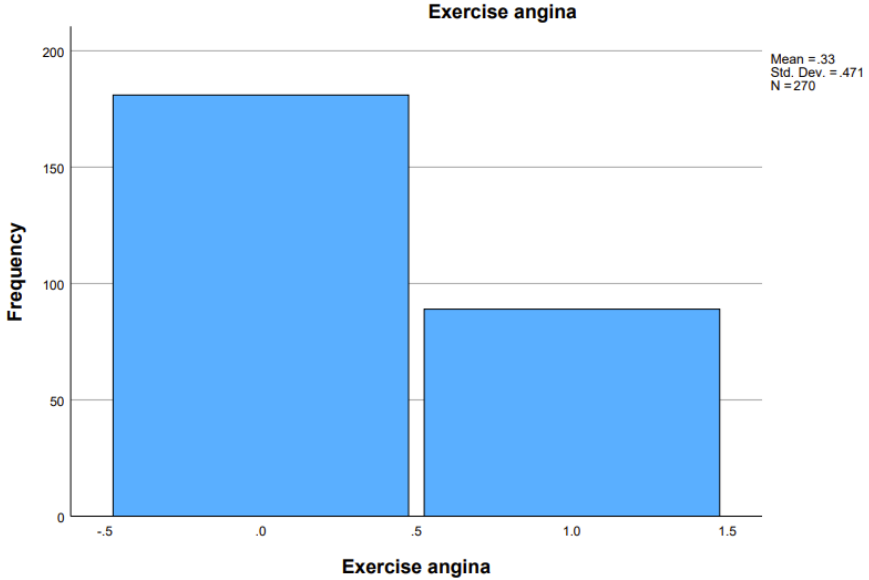

Cholesterol
Mean =248.76
Std. Dev. =51.406
N =253
Frequency
50
40
30
20
10
0
100
200
300
400
500
600
Cholesterol

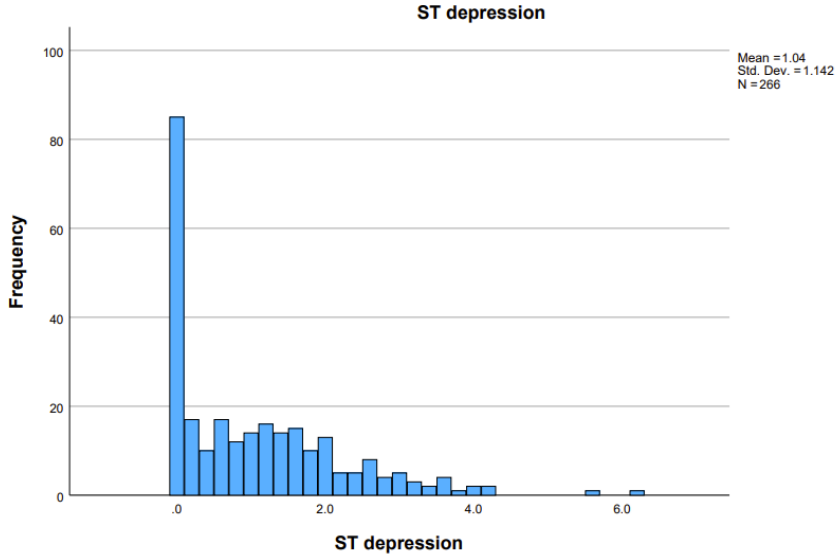
Frequency


EKG results
Mean = 1.02
Std. Dev. = .998
N = 270
Frequency
125
100
75
50
25
0
-.5
.0
.5
1.0
1.5
2.0
2.5
EKG results

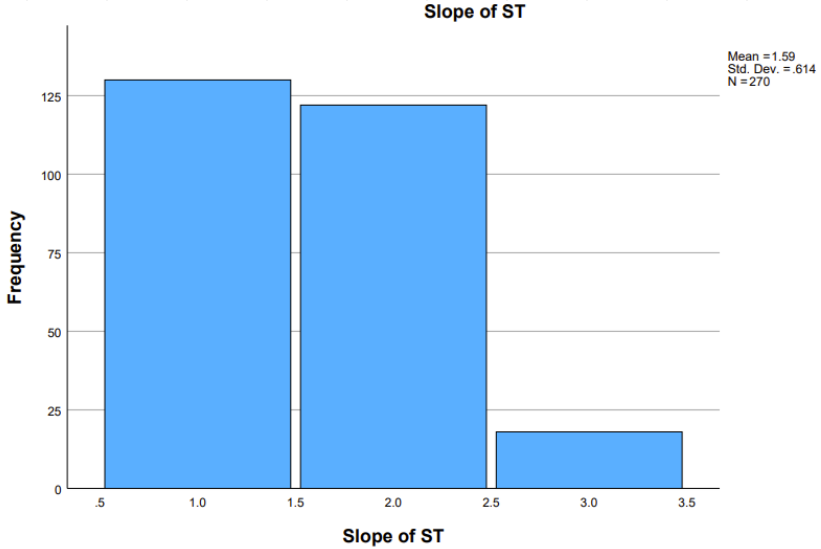

Max HR
Mean = 149.32
Std. Dev. = 22.995
N = 259
Frequency
40
30
20
10
0
75
100
125
150
175
200
Max HR

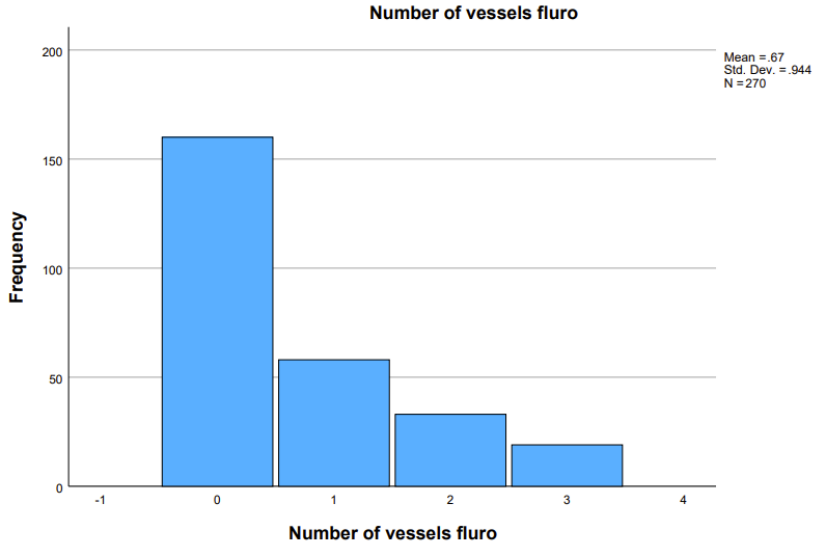
Frequency


Exercise angina
Mean = .33
Std. Dev. = .471
N = 270
Frequency
200
150
100
50
0
-.5
.0
.5
1.0
1.5
Exercise angina

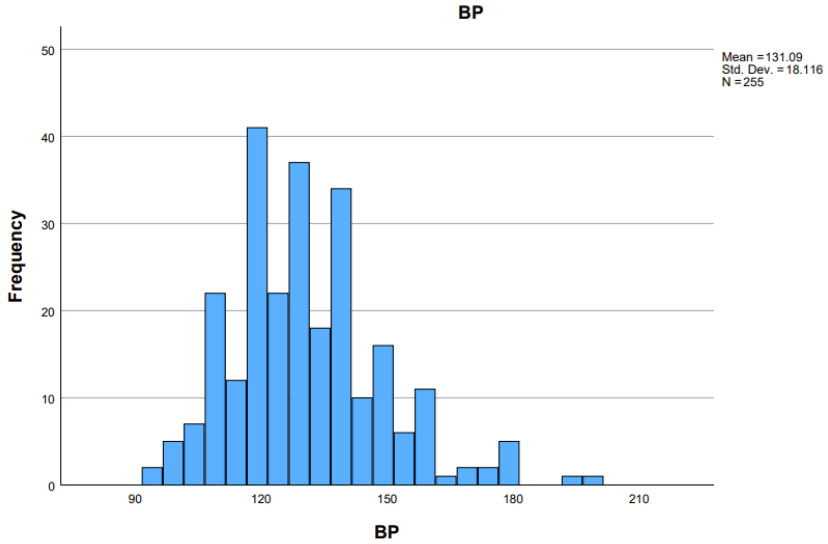

ST depression
Mean = 1.04
Std. Dev. = 1.142
N = 266
Frequency
100
80
60
40
20
0
.0
2.0
4.0
6.0
ST depression

Frequency


Slope of ST
Mean = 1.59
Std. Dev. = .614
N = 270
Frequency
125
100
75
50
25
0
.5
1.0
1.5
2.0
2.5
3.0
3.5
Slope of ST


Number of vessels fluro
Mean = .67
Std. Dev. = .944
N = 270
Frequency
200
150
100
50
0
-1
0
1
2
3
4
Number of vessels fluro

Frequency

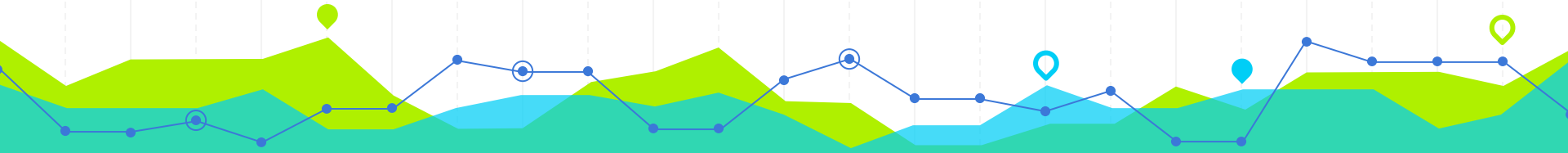

BP
Mean = 131.09
Std. Dev. = 18.116
N = 255
Frequency
50
40
30
20
10
0
90
120
150
180
210
BP

Central Tendency and Dispersion

A summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
Age	255	48	29	77	54,20	9,226	85,121
SMEAN(BP)	255	98,0	94,0	192,0	129,896	16,4039	269,088
SMEAN(Cholesterol)	255	234,0	126,0	360,0	246,349	43,0355	1852,055
SMEAN(MaxHR)	255	131,0	71,0	202,0	149,346	22,9499	526,699
Valid N (listwise)	255						



Cross Tabulation

Answers to questions

5

1. Highest Prevalence of Heart-disease amongst the sexes?

Significantly more of sex 1 (making up nearly 70% of the entire study) were recorded in the study than sex 0 (whom we both cannot determine without making assumptions). Male or female?

More than 80% within sex 0 cases had no heart disease. Whereas more than 80% of sex 1 make up our heart disease cases.

Sex * Heart Disease Crosstabulation

			Heart Disease		
			Absence	Presence	Total
Sex	0	Count	63	14	77
		% within Sex	81,8%	18,2%	100,0%
		% within Heart Disease	44,1%	12,5%	30,2%
		% of Total	24,7%	5,5%	30,2%
	1	Count	80	98	178
		% within Sex	44,9%	55,1%	100,0%
		% within Heart Disease	55,9%	87,5%	69,8%
		% of Total	31,4%	38,4%	69,8%
Total		Count	143	112	255
		% within Sex	56,1%	43,9%	100,0%
		% within Heart Disease	100,0%	100,0%	100,0%
		% of Total	56,1%	43,9%	100,0%

2. Is there any connection between heart-disease and chest-pain types?

Chest pain type * Heart Disease Crosstabulation

			Heart Disease		
			Absence	Presence	Total
Chest pain type	1	Count	14	4	18
		% within Chest pain type	77,8%	22,2%	100,0%
		% within Heart Disease	9,8%	3,6%	7,1%
		% of Total	5,5%	1,6%	7,1%
	2	Count	33	7	40
		% within Chest pain type	82,5%	17,5%	100,0%
		% within Heart Disease	23,1%	6,3%	15,7%
		% of Total	12,9%	2,7%	15,7%
	3	Count	60	17	77
		% within Chest pain type	77,9%	22,1%	100,0%
		% within Heart Disease	42,0%	15,2%	30,2%
		% of Total	23,5%	6,7%	30,2%
	4	Count	36	84	120
		% within Chest pain type	30,0%	70,0%	100,0%
		% within Heart Disease	25,2%	75,0%	47,1%
		% of Total	14,1%	32,9%	47,1%
Total		Count	143	112	255
		% within Chest pain type	56,1%	43,9%	100,0%
		% within Heart Disease	100,0%	100,0%	100,0%
		% of Total	56,1%	43,9%	100,0%

Nearly half of the study frequency were classified as chest-pain type 4. 70% cases in this study, who were type 4 for chest-pain type categorization, made up 75% of heart-disease cases.

Whilst, nearly 78% of those who were type 1 had no presence of heart-disease. But type 1 were only 7% of the total.

It can be observed that there are higher chances of being found to possess heart-disease when you are a chest-pain type 4.

However, it can be observed that more factors than just your chest-pain type category, contribute to one having a heart-disease diagnosis.

3. Does the presence of exercise angina affect the probability of heart disease?

The presence of exercise angina shows greater probability to have heart diseases (73.5%) opposed to those without it (29.7%).

Exercise angina * Heart Disease Crosstabulation

			Heart Disease		
			Absence	Presence	Total
Exercise angina	0	Count	121	51	172
		% within Exercise angina	70,3%	29,7%	100,0%
		% within Heart Disease	84,6%	45,5%	67,5%
		% of Total	47,5%	20,0%	67,5%
	1	Count	22	61	83
		% within Exercise angina	26,5%	73,5%	100,0%
		% within Heart Disease	15,4%	54,5%	32,5%
		% of Total	8,6%	23,9%	32,5%
Total		Count	143	112	255
		% within Exercise angina	56,1%	43,9%	100,0%
		% within Heart Disease	100,0%	100,0%	100,0%
		% of Total	56,1%	43,9%	100,0%

4. What's the distribution between the EKG results and heart-disease?

- The likelihood of having heart disease is greater if the result of the electrocardiogram is 2. If the result is 0, the probability of having heart disease is 33.9%.

EKG results * Heart Disease Crosstabulation

			Heart Disease		
			Absence	Presence	Total
EKG results	0	Count	82	42	124
		% within EKG results	66,1%	33,9%	100,0%
		% within Heart Disease	57,3%	37,5%	48,6%
		% of Total	32,2%	16,5%	48,6%
	1	Count	1	1	2
		% within EKG results	50,0%	50,0%	100,0%
		% within Heart Disease	0,7%	0,9%	0,8%
		% of Total	0,4%	0,4%	0,8%
	2	Count	60	69	129
		% within EKG results	46,5%	53,5%	100,0%
		% within Heart Disease	42,0%	61,6%	50,6%
		% of Total	23,5%	27,1%	50,6%
Total		Count	143	112	255
		% within EKG results	56,1%	43,9%	100,0%
		% within Heart Disease	100,0%	100,0%	100,0%
		% of Total	56,1%	43,9%	100,0%

Thank you