Heart Disease dataset

Project - Data Analysis Industrial Traineeship

Group A - ***Carla Mota Leal***

Brainnest Supervisor:
Elnaz Gholipour

Report outline

Part 1	• Data cleaning • Descriptive statistics
Part 2	• Normality of variables • Visual representation of Normality • Sampling
Part 3	• Hypothesis about the data • Parametric tests
Part 4	• Non-parametric tests • Correlation • Regression

Part 1

Objectives

1. Determine the types of data for each variables in the dataset.
2. Explore and clean the data.
3. Identify missing or noisy data.
4. Deal with missing and noisy data.
5. Perform simple descriptive statistics to explore the data before performing any further analysis.

Data Understanding

Target: Heart Disease

Features:

- **Age**: Age of the patient
- **Sex**: Sex/Gender of the patient
- **Chest pain type:** Chest pain type of the patient
  - Value 1: typical angina
  - Value 2: atypical angina
  - Value 3: non-anginal pain
  - Value 4: asymptomatic
- **BP**: Blood pressure of the patient (in mm Hg)
- **Cholesterol:** Cholesterol in mg/dl fetched via BMI sensor
- **FBS over 120**: fasting blood sugar > 120 mg/dl
  - (1 = true; 0 = false)
- **EKG results:** Electrocardiographic results
  - 0 = normal
  - 1 = having ST-T wave abnormality
  - 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria
- **Max HR:** The patient's maximum heart rate achieved
- **Exercise angina:** type of chest pain caused by reduced blood flow to the heart
- **ST depression:** Specific outcome that may appear in a person's ECG results
- **Slope of ST**: The slope of the peak exercise ST segment
  - Value 1: upsloping
  - Value 2: flat
  - Value 3: downsloping
- **Number of vessels fluro:** Number of major vessels (0-3)
- **Thallium**: A radioactive tracer to see how much blood is reaching different parts of the heart.

References:

Simmons II, B. (2021). *Investigating Heart Disease Datasets and Building Predictive Models* (Doctoral dissertation, Elizabeth City State University). https://libres.uncg.edu/ir/ecsu/f/Brandon_Simmons_Thesis-Final.pdf

V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D. https://archive.ics.uci.edu/ml/datasets/heart+disease

Features Type

Features
Age: Continuous/Scale
Sex: Binary/Nominal
Chest Pain Type: Ordinal
BP: Continuous/Scale
Cholesterol: Continuous/Scale
FBS over 12: Binary/Nominal
EKG result: Ordinal
Max HR: Continuous/Scale
Exercise angina: Binary/Nominal
ST depression: Continuous/Scale
Slop of ST: Ordinal
Number of vessels: Ordinal
Thallium: Ordinal

Data Preparation

Remove Irrelevant Feature✓

Deduplicate Check✓

Noise Treatment

- Some noise are not logical, due to human error, such as Sex 11. It should be either 0 or 1.
- Remove the noise and replace the value.

Outlier Treatment

- Check the type of distribution of each feature
- According to the type of distribution, we can find outliers data using A or B methods.

Missing Data Handling

- Understand the overall descriptive of the data distribution.
- Replace the missing value using the right method regarding the data distribution.

Extra Steps for Modeling:

- Qualitative data to Quantitative data
- Scaling
- Balancing

Descriptive Statistics

Statistics

		Age	Sex	Chest pain type	BP	Cholesterol	FBS over 120	EKG results	Max HR	Exercise angina	ST depression	Slope of ST	Number of vessels fluro	Thallium
N	Valid	270	270	270	255	253	270	270	261	270	266	270	270	270
	Missing	0	0	0	15	17	0	0	9	0	4	0	0	0
Mean		54.43	.75	3.32	131.09	248.76	.15	1.02	157.89	.33	1.042	1.59	.67	4.70
Median		55.00	1.00	3.00	130.00	245.00	.00	2.00	154.00	.00	.800	2.00	.00	3.00
Mode		54	1	4	120	234	0	2	162	0	.0	1	0	3
Std. Deviation		9.109	.965	2.660	18.116	51.406	.356	.998	100.789	.471	1.1416	.614	.944	1.941
Variance		82.975	.931	7.074	328.204	2642.612	.127	.996	10158.515	.222	1.303	.377	.891	3.766
Skewness		-.164	7.647	13.336	.757	1.200	1.992	-.045	10.630	.729	1.274	.543	1.210	.287
Std. Error of Skewness		.148	.148	.148	.153	.153	.148	.148	.151	.148	.149	.148	.148	.148
Range		48	11	43	106	438	1	2	1309	1	6.2	2	3	4
Minimum		29	0	1	94	126	0	0	71	0	.0	1	0	3
Maximum		77	11	44	200	564	1	2	1380	1	6.2	3	3	7

Noise Treatment

- Sex→ 2 Noise (11,10)
- Chest Pain Type→ 1 Noise (44)
- Max HR→ 2 Noise (1154,1380)

These numbers are impossible for their feature. We have several options to solve this issue in order of priority.

1. Refer to the expert of the organization and get the correct data.
2. Emptying the relevant cell and filling it through the machine learning algorithm.
3. Deleting data can be considered as the last and worst option.

Because we don't have access to experts in this case. So, we choose the second option.

Outlier Treatment

In this part, we must first check the type of distribution of each feature, and after that, according to the type of distribution, we can find outliers data using A or B methods.

A: if our data is normally distributed, we can use standard deviation from mean.

B: if our data is not normally distributed, we can use Box Plot.

Field	Storage	Status		Distribution
Age	Real	✓		Weibull
Chest pain type	Real	✓		Poisson
BP	Real	✓		Lognormal
Cholesterol	Real	✓		Lognormal
EKG results	Real	✓		NegativeBinomialF...
Max HR	Real	✓		Lognormal
ST depression	Real	✓		Normal
Slope of ST	Real	✓		Poisson
Number of vessels...	Real	✓		NegativeBinomialF...
Thallium	Real	✓		Weibull

ST depression: 2 Outliers

BP: 9 Outliers

Cholesterol: 3 Outliers and 1 extreme Outlier

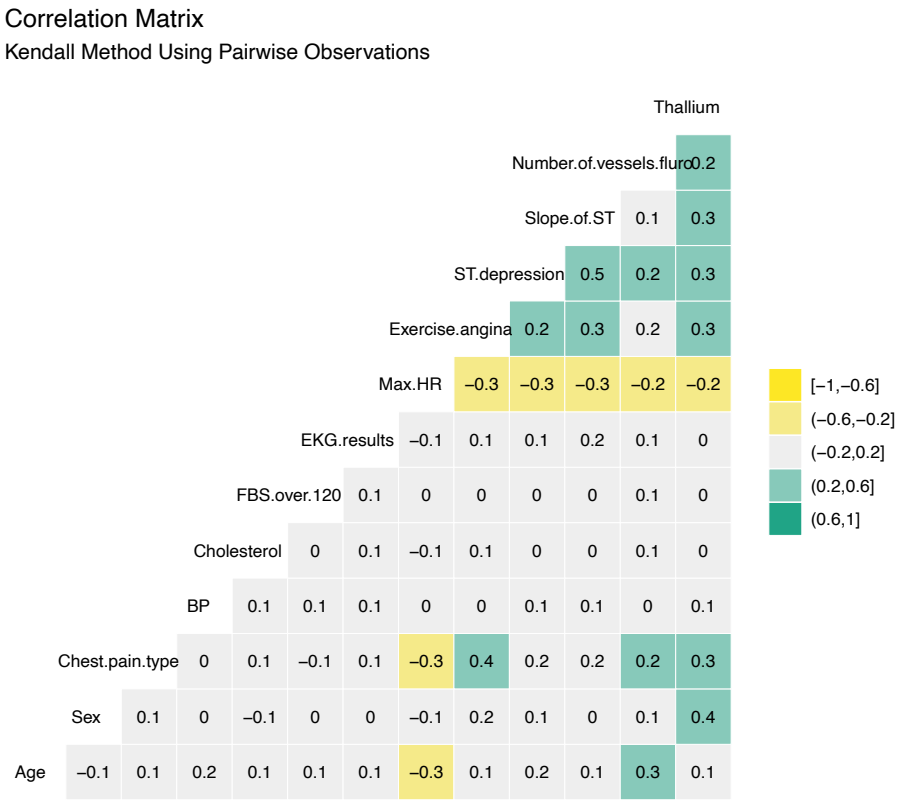
Max HR: 1 Outlier

Missing Data Handling

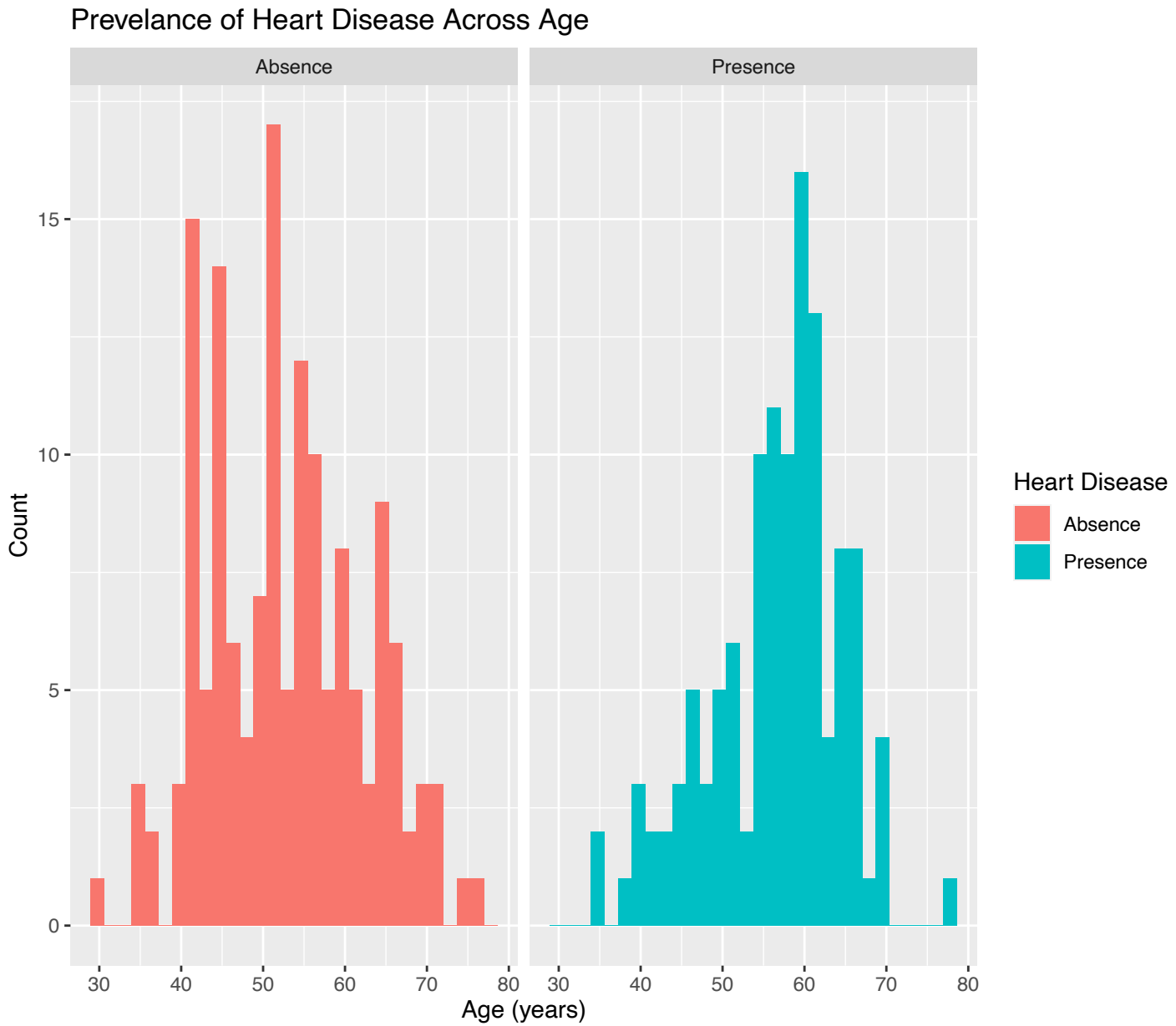
- Sex→ 2 Missing
- BP→ 15 Missing
- Cholesterol→ 18 Missing
- Max HR→ 11 Missing
- ST depression→ 4 Missing
- Chest Pain Type→ 1 Missing

Because the number of missing data is small, we use model CART (Classification And Regression Tree) for all of them

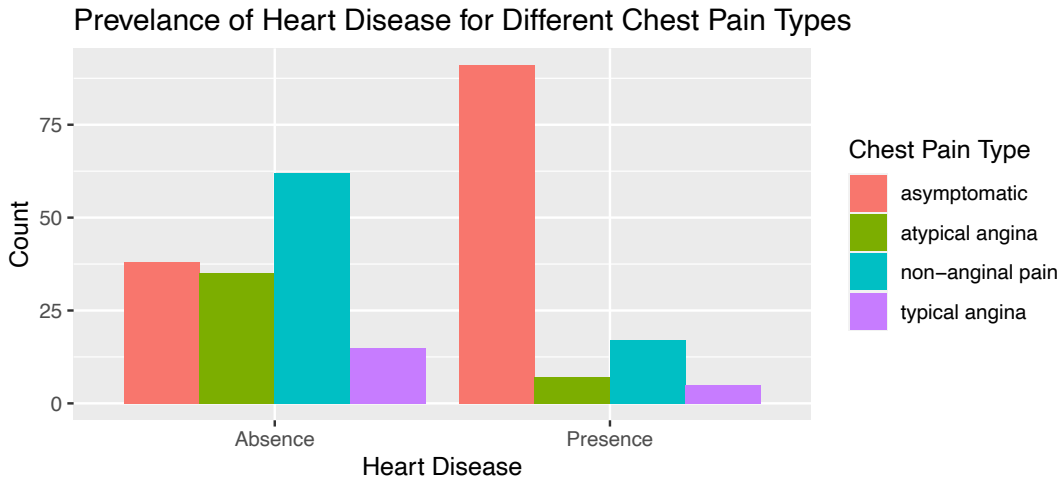
Exploratory analysis

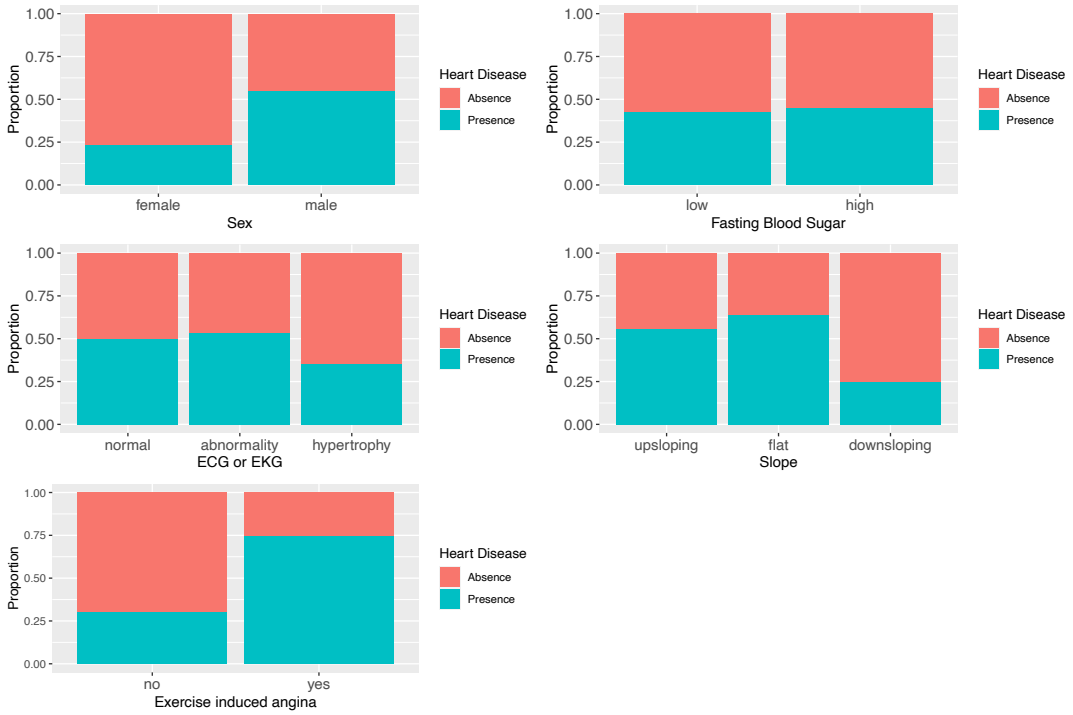


Exploratory analysis

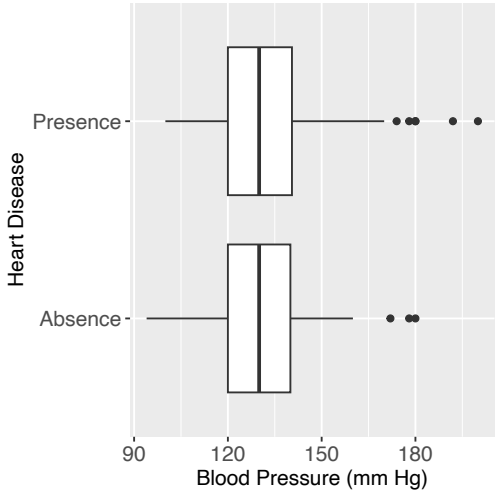


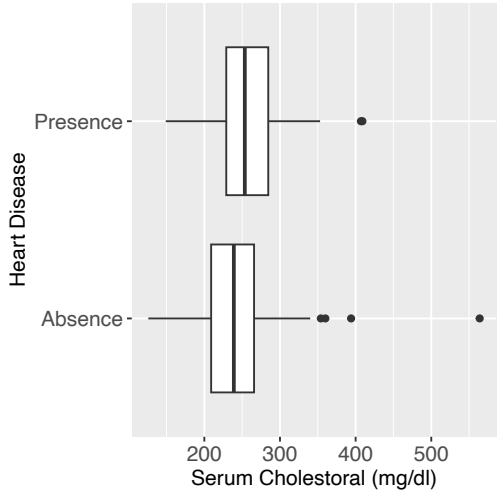
Exploratory analysis

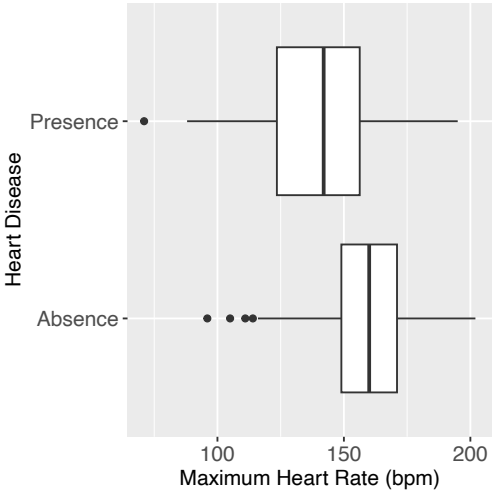


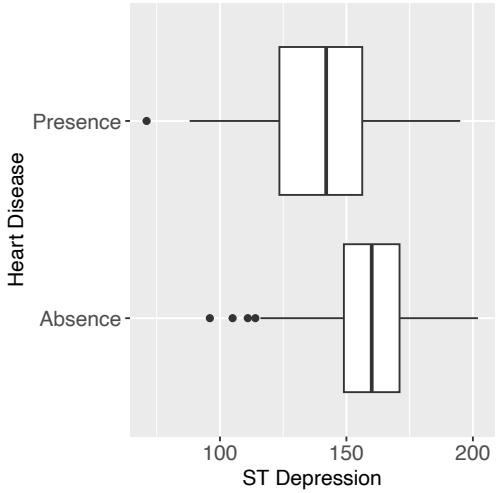


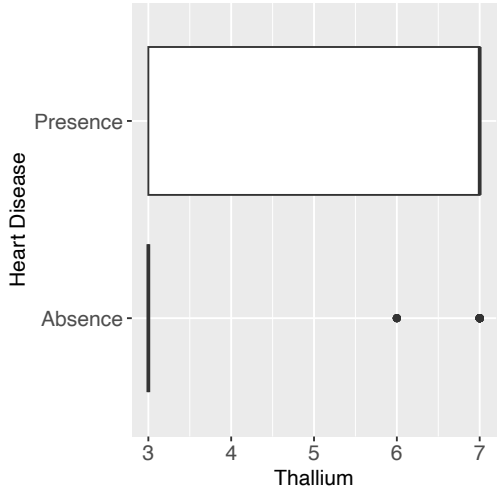
Exploratory analysis











Extra Steps for Modeling Qualitative data to Quantitative data

We should change the qualitative data to quantitative data because we want to use the model which works with distance.

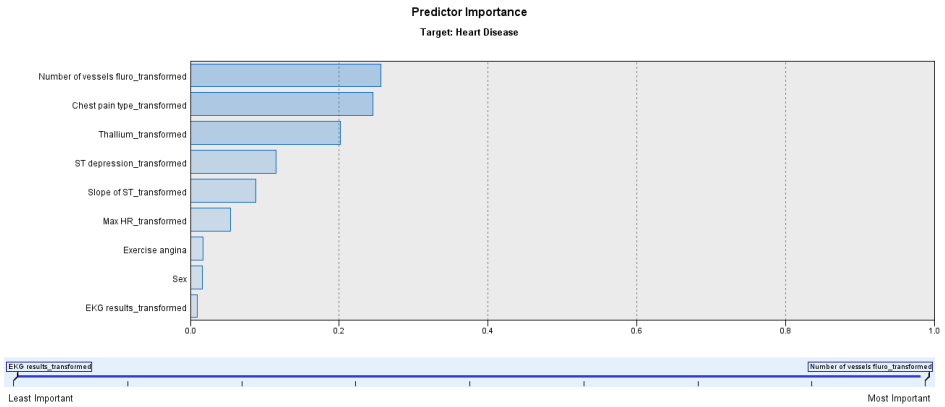
- Chest Pain Type
- EKG result: Ordinal
- Slop of ST: Ordinal
- Number of vessels: Ordinal

Extra Steps for Modeling Scaling

We need to do this because we want to check the distance between the data, so we need all the features to be on the same scale.

We use the min-max transformation for scaling the data. Now all data are between 0 and 100.

Modelling



Results for output field Heart Disease

Comparing $C-Heart Disease with Heart Disease

'Partition'	1_Training		2_Testing	
Correct	186	91.18%	59	89.39%
Wrong	18	8.82%	7	10.61%
Total	204		66	

We C5.0 classification model. The table on the right side measures the accuracy of the model. The table on the left shows the importance of each feature for this model.

Conclusion

- In total there are 14 different features in which six are ratio and ordinal, and another four are nominal type of data
  - **Scale** or **ratio**: age, blood pressure, cholesterol, maximum heart rate, ST depression.
  - **Ordinal**: Chest Pain Type, EKG, slope of peak exercise ST segment, number of major vessels, Thallium injection.
  - **Nominal**: Sex, FBS over 120, Exercise angina, and presence of heart disease.
- Our data showed to have different characters and missing values. It was used regression to replace missing values.
- Data cleaning, outlier handling, and missing value analysis are important steps before analyzing the data, as it might influence the means, central tendencies, and the overall results of the analysis.

Part 2

Normality tests

OBJECTIVE normality tests

- The specific objective of this session is to check the distribution of the data before we do any analysis.
- By knowing the right distribution, we can run the right analysis. For instance, if we want to see the correlation between two variables and the data is not following the normal distribution, we should use the Spearman-Correlation test instead of Pearson-Correlation test.

NORMALITY TEST

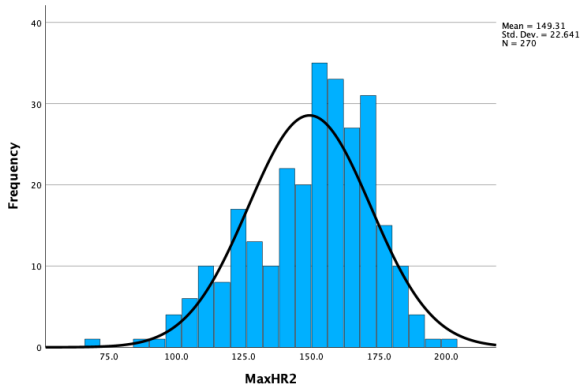
- Applied methods:
  - Visual:
    - Q-Q plot & P-P plots
    - Histogram, or frequency distribution
    - Box plots
  - Statistical test
    - Skewness & Kurtosis z-values
    - Shapiro-Wilk & Kolmogorov-Smirnov

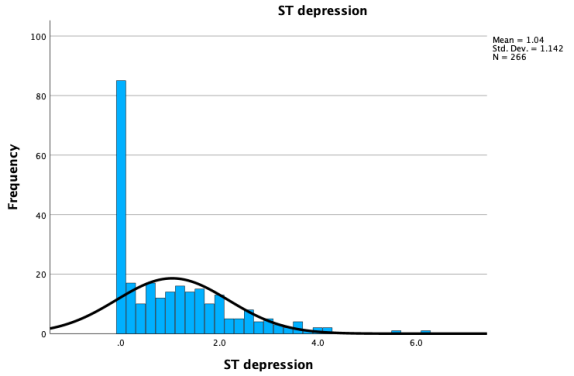
NORMALITY TEST

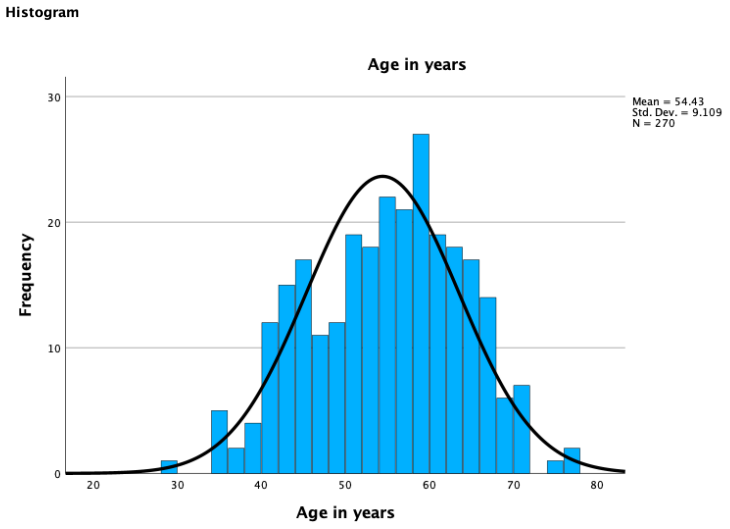
- Method:

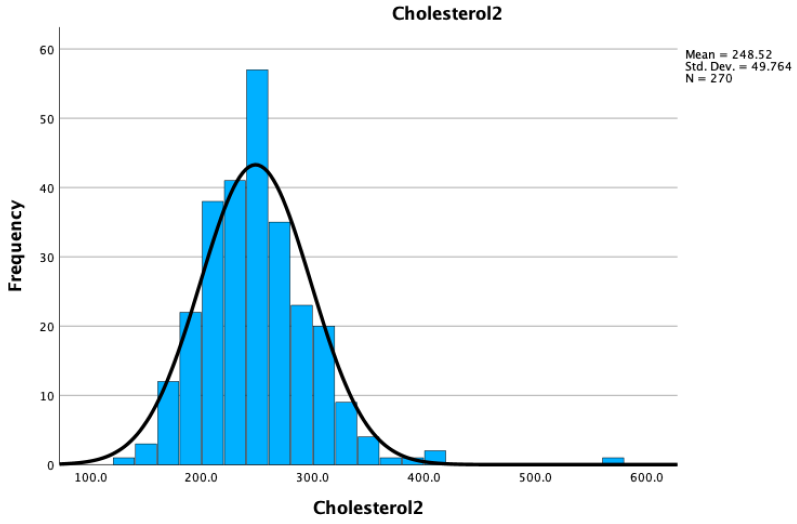
  Visual

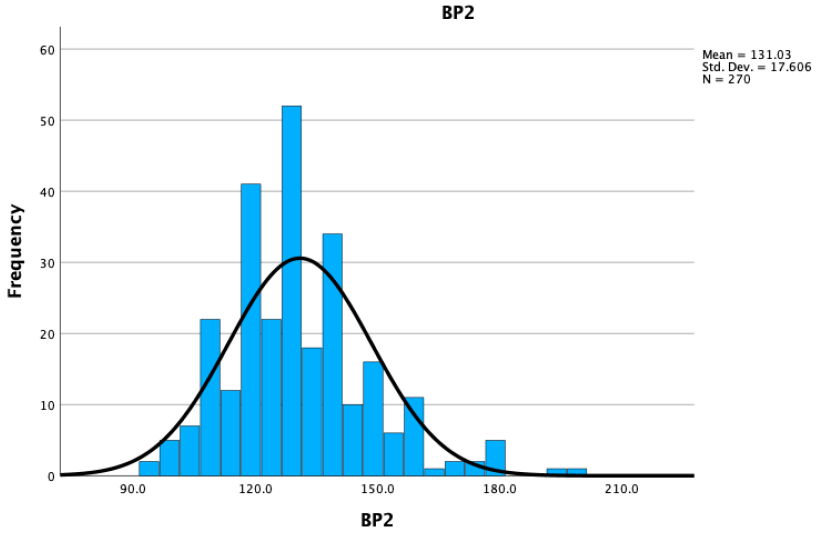
  - Histogram of cholesterol showed to be the closest to the normal base line.







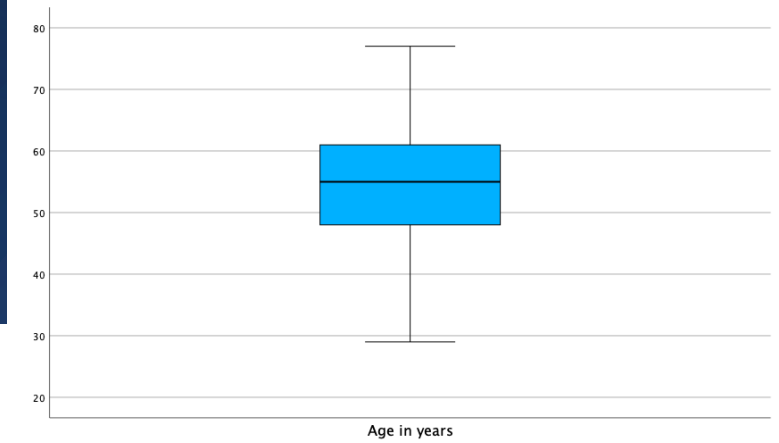


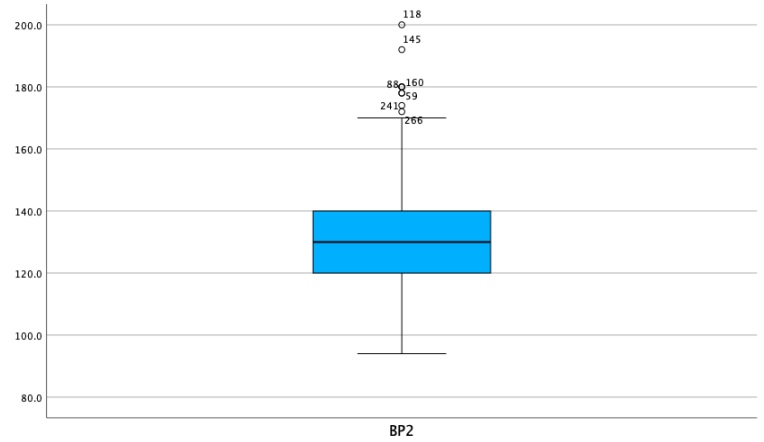


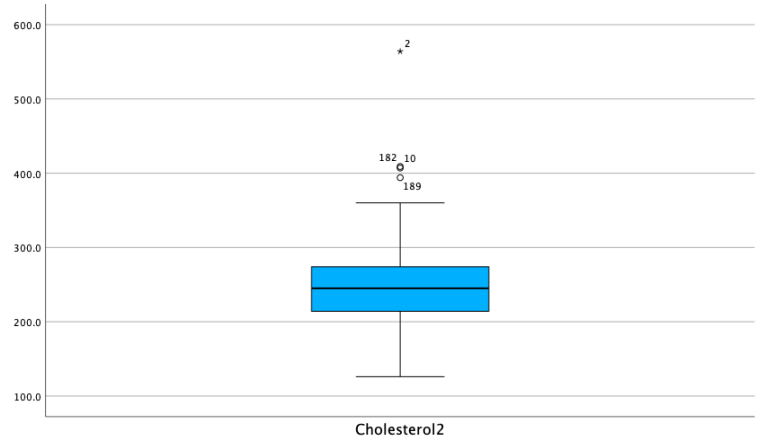
NORMALITY TEST

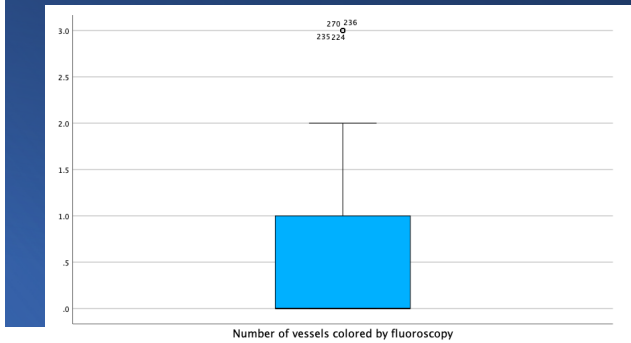
- Method:
  Visual
  - Box plots

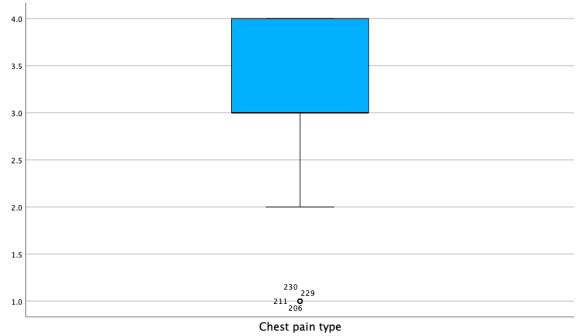
From all features, age, Blood Pressure, Cholesterol, and maximum heart rate are assumed symmetrical.

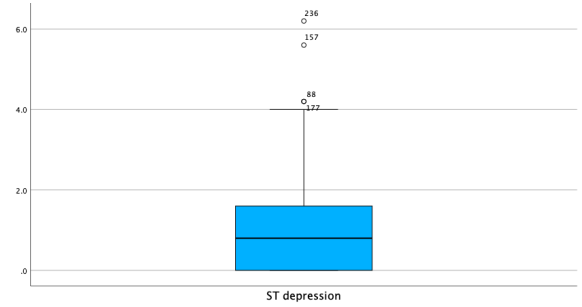


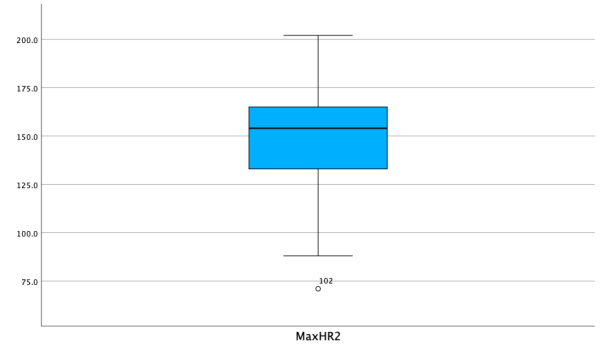












NORMALITY TEST

- Method:

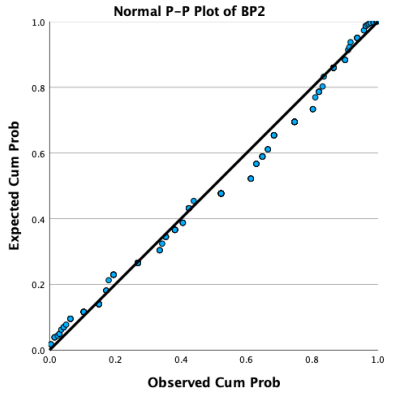
  Visual

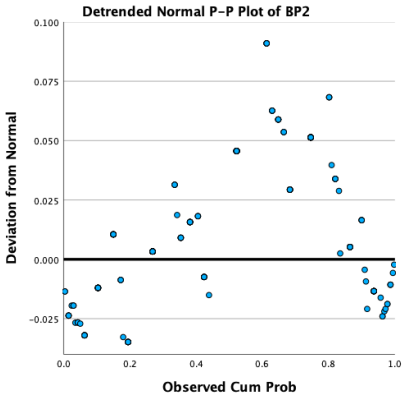
  - P-P plots

Deviations from straight line indicate deviatic
normality.

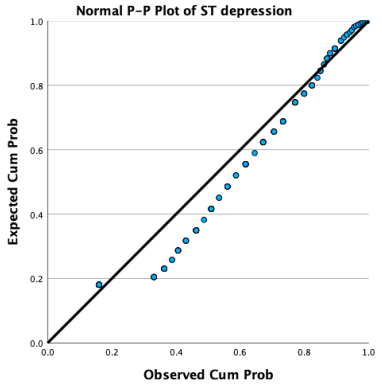
All features deviated from the base line. The le
cholesterol level.

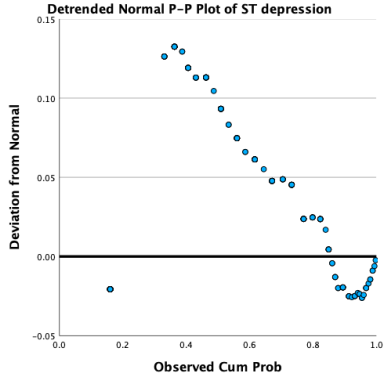
BP2

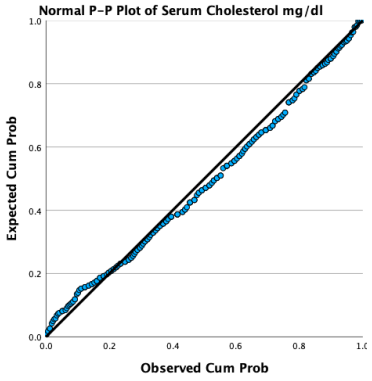


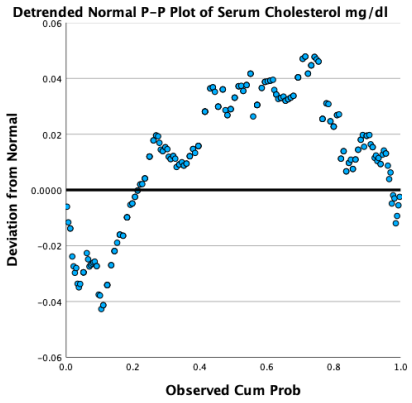


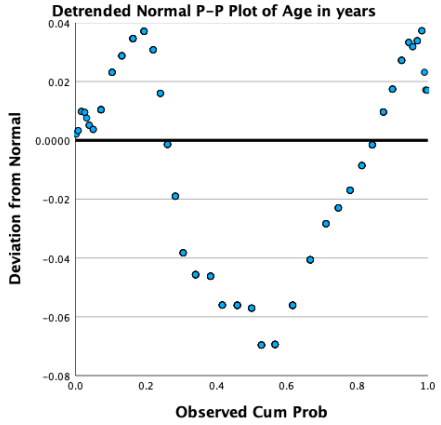
ST depression

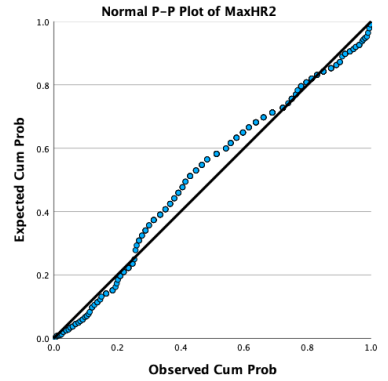


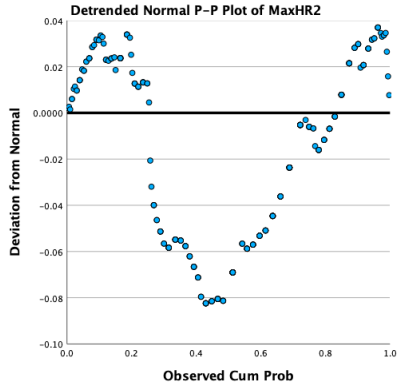












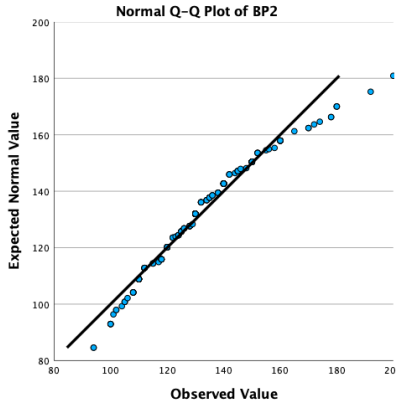
NORMALITY TEST

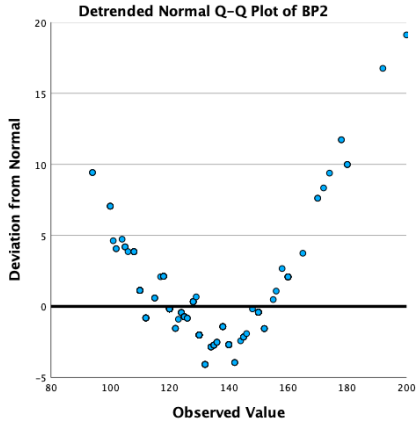
- Method:
  Visual
  - Q-Q plot

Deviations from straight line indicate deviations from normality.

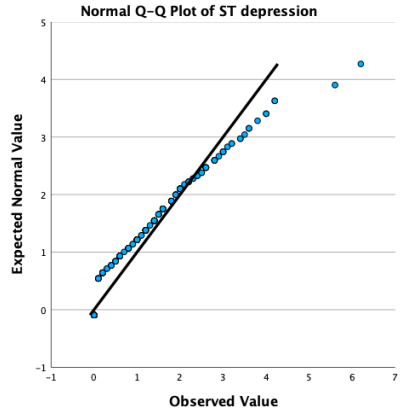
Similar outputs from P-P plots.

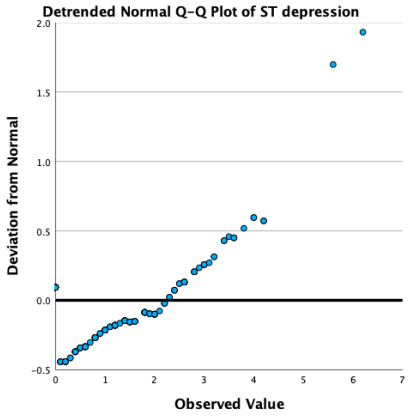
BP2

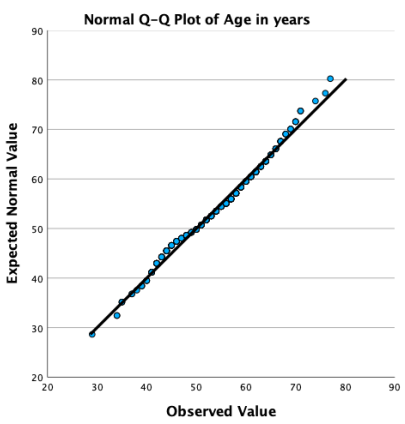


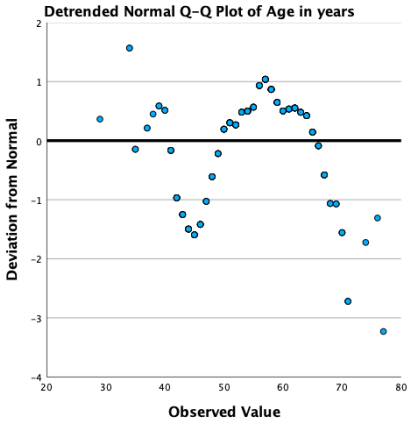


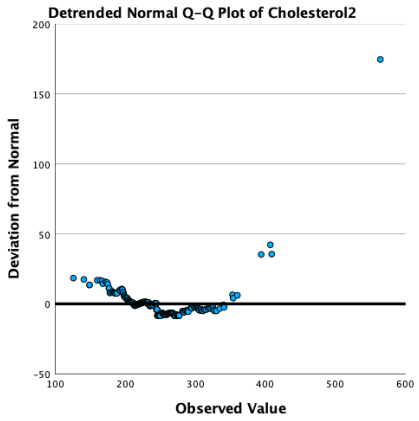
ST depression

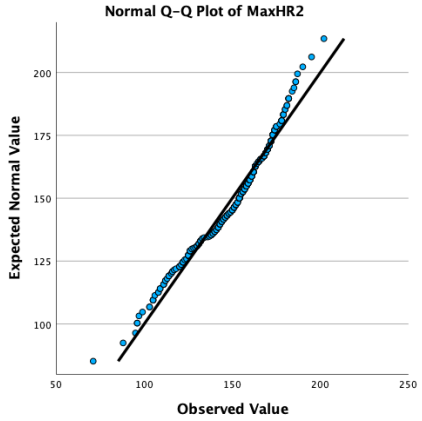


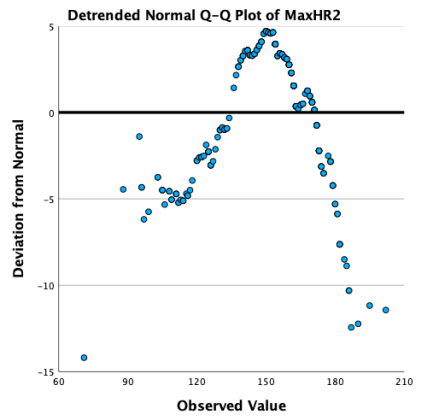












NORMALITY TEST

Statistical test

Skewness & Kurtosis z-values

- Z-value should be between -1.96 and +1.96.
  - Age skewness = -1.114
  - Age kurtosis = -1.731
  - Chest pain skewness = -5.832
  - Chest pain kurtosis = -1.097
- In FBS over 120, measure is too large compared to its standard error.
- Besides FBS and chest pain, age is somewhat skewed and kurtotic but we can assume they are approximately normal.

Descriptives

			Statistic	Std. Error
Age in years	Mean		54.53	.554
	95% Confidence Interval for Mean	Lower Bound	53.44	
		Upper Bound	55.62	
	5% Trimmed Mean		54.59	
	Median		55.00	
	Variance		81.518	
	Std. Deviation		9.029	
	Minimum		29	
	Maximum		77	
	Range		48	
	Interquartile Range		13	
	Skewness		-.166	.149
	Kurtosis		-.516	.298
Chest pain type	Mean		3.17	.059
	95% Confidence Interval for Mean	Lower Bound	3.05	
		Upper Bound	3.28	
	5% Trimmed Mean		3.24	
	Median		3.00	
	Variance		.911	
	Std. Deviation		.954	
	Minimum		1	
	Maximum		4	
	Range		3	
	Interquartile Range		1	
	Skewness		-.869	.149
	Kurtosis		-.327	.298
FBS over 120	Mean		.15	.022
	95% Confidence Interval for Mean	Lower Bound	.11	
		Upper Bound	.19	
	5% Trimmed Mean		.11	
	Median		.00	
	Variance		.128	
	Std. Deviation		.358	
	Minimum		0	
	Maximum		1	
	Range		1	
	Interquartile Range		0	
	Skewness		1.967	.149
	Kurtosis		1.885	.298

NORMALITY TEST

Statistical test

Skewness & Kurtosis z-values

- Z-value should be between -1.96 and +1.96.
  - ECG skewness = -0.409
- All measure is too large compared to its standard error (kurtosis)
- ECG is somewhat skewed

Descriptives

			Statistic	Std. Error
ECG readings	Mean		1.03	.061
	95% Confidence Interval for Mean	Lower Bound	.91	
		Upper Bound	1.15	
	5% Trimmed Mean		1.03	
	Median		2.00	
	Variance		.995	
	Std. Deviation		.998	
	Minimum		0	
	Maximum		2	
	Range		2	
	Interquartile Range		2	
	Skewness		-.061	.149
	Kurtosis		-2.004	.298
Exercise angina	Mean		.33	.029
	95% Confidence Interval for Mean	Lower Bound	.27	
		Upper Bound	.38	
	5% Trimmed Mean		.31	
	Median		.00	
	Variance		.221	
	Std. Deviation		.470	
	Minimum		0	
	Maximum		1	
	Range		1	
	Interquartile Range		1	
	Skewness		.741	.149
	Kurtosis		-1.461	.298
ST depression	Mean		1.042	.0700
	95% Confidence Interval for Mean	Lower Bound	.904	
		Upper Bound	1.180	
	5% Trimmed Mean		.932	
	Median		.800	
	Variance		1.303	
	Std. Deviation		1.1416	
	Minimum		.0	
	Maximum		6.2	
	Range		6.2	
	Interquartile Range		1.7	
	Skewness		1.274	.149
	Kurtosis		1.834	.298

NORMALITY TEST

Statistical test

Skewness & Kurtosis z-values

- Z-value should be between -1.96 and +1.96.
  - Slope skewness = 3.832
  - Slope kurtosis = -1.976
  - Thallium skewness = 1.852
  - N. v. fluo. kurtosis = 1.060

- Thallium and Number of vessels colored by fluo, measure is too large compared to its standard error.

- We can assume they are not normal.

Descriptives

			Statistic	Std. Error
Slope of ST	Mean		1.58	.038
	95% Confidence Interval for Mean	Lower Bound	1.50	
		Upper Bound	1.65	
	5% Trimmed Mean		1.53	
	Median		2.00	
	Variance		.381	
	Std. Deviation		.617	
	Minimum		1	
	Maximum		3	
	Range		2	
	Interquartile Range		1	
	Skewness		.571	.149
	Kurtosis		-.589	.298
Number of vessels colored by fluoroscopy	Mean		.67	.058
	95% Confidence Interval for Mean	Lower Bound	.56	
		Upper Bound	.78	
	5% Trimmed Mean		.58	
	Median		.00	
	Variance		.894	
	Std. Deviation		.945	
	Minimum		0	
	Maximum		3	
	Range		3	
	Interquartile Range		1	
	Skewness		1.218	.149
	Kurtosis		.316	.298
Thallium	Mean		4.71	.119
	95% Confidence Interval for Mean	Lower Bound	4.47	
		Upper Bound	4.94	
	5% Trimmed Mean		4.67	
	Median		3.00	
	Variance		3.770	
	Std. Deviation		1.942	
	Minimum		3	
	Maximum		7	
	Range		4	
	Interquartile Range		4	
	Skewness		.276	.149
	Kurtosis		-1.907	.298

NORMALITY TEST

Statistical test

Skewness & Kurtosis z-values

- Z-value should be between -1.96 and +1.96.
  - BP skewness =5.382
  - BP kurtosis = 4.030
  - MxHR skewness = -4.094
  - MxHR kurtosis = -0.007
- In Cholesterol, measure is too large compared to its standard error.
- We can assume they are not normal.

Descriptives

			Statistic	Std. Error
BP2	Mean		131.169	1.0770
	95% Confidence Interval for Mean	Lower Bound	129.049	
		Upper Bound	133.290	
	5% Trimmed Mean		130.299	
	Median		130.000	
	Variance		308.534	
	Std. Deviation		17.5651	
	Minimum		94.0	
	Maximum		200.0	
	Range		106.0	
	Interquartile Range		20.0	
	Skewness		.802	.149
	Kurtosis		1.201	.298
Cholesterol2	Mean		248.560	3.0622
	95% Confidence Interval for Mean	Lower Bound	242.531	
		Upper Bound	254.590	
	5% Trimmed Mean		246.546	
	Median		245.000	
	Variance		2494.368	
	Std. Deviation		49.9436	
	Minimum		126.0	
	Maximum		564.0	
	Range		438.0	
	Interquartile Range		60.3	
	Skewness		1.256	.149
	Kurtosis		5.761	.298
MaxHR2	Mean		149.486	1.3937
	95% Confidence Interval for Mean	Lower Bound	146.742	
		Upper Bound	152.231	
	5% Trimmed Mean		150.308	
	Median		154.000	
	Variance		516.683	
	Std. Deviation		22.7307	
	Minimum		71.0	
	Maximum		202.0	
	Range		131.0	
	Interquartile Range		31.3	
	Skewness		-.610	.149
	Kurtosis		.002	.298

NORMALITY TEST

- Statistical test
  - Shapiro-Wilk & Kolmogorov-Smirnov

Tests of Normality

	Kolmogorov–Smirnov[a]			Shapiro–Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Age in years	.066	266	.006	.989	266	.035
Sex	.432	266	<.001	.589	266	<.001
Chest pain type	.285	266	<.001	.789	266	<.001
FBS over 120	.512	266	<.001	.427	266	<.001
ECG readings	.346	266	<.001	.641	266	<.001
Exercise angina	.430	266	<.001	.591	266	<.001
Slope of ST	.315	266	<.001	.740	266	<.001
Number of vessels colored by fluoroscopy	.354	266	<.001	.713	266	<.001
Thallium	.370	266	<.001	.653	266	<.001
ST depression	.181	266	<.001	.850	266	<.001
BP2	.128	266	<.001	.958	266	<.001
Cholesterol2	.071	266	.002	.938	266	<.001
MaxHR2	.099	266	<.001	.968	266	<.001

All features shows to be <=0.05

We therefore have significant evidence to **reject** the null hypothesis, in this test,
that the variable follows a normal distribution.

Questions to apply z-score

- What is the probability of the patients whose blood pressure is more than 120?

N = 270, Mean = 131.033, Std = 17.6058

For 120 under ZBP new column, -.6267

P(x=120, z=-.6267) = 26.76%

P(x>120) = 1-26.76% = **73.24%**

<u>Alternative Method (The Distribution is Not Normal)</u>

We sort the data and count the patients with blood pressure above 120.

We found that 181 patients have blood pressure above 120.

Therefore, the probability of the patients with blood pressure above 120 is: 181/270*100% = **67.04%**

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic
BP2	270	106.0	94.0	200.0	131.033	1.0715	17.6058

Questions to apply z-score

- What is the probability of the patients whose cholesterol is more than 240 mg/dl? Considered high cholesterol level.

N = 270, Mean = 248.522, Std = 49.7639

For 240 under ZChol new column, 0.13362

P(x=240, z=.1336) = 55.17%

P(x>240) = 1-55.17% = **44.83%**

Alternative Method (The Distribution is Not Normal)

We sort the data and count the patients with cholesterol level above 240.

We found that 142 patients have cholesterol level above 240.

Therefore, the probability of the patients with cholesterol level above 240 is: 142/270 * 100% = **52.59%**

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic
BP2	270	106.0	94.0	200.0	131.033	1.0715	17.6058
Cholesterol2	270	438.0	126.0	564.0	248.522	3.0285	49.7639

Questions to apply z-score

- What is the probability of maximum heart rate be more than 170 (from 50 years age)?

N = 270, Mean =149.313, Std = 22.641

For 170 under ZmaxHR new column, 0.914

P(x=170, z=.914) = 81.86%

P(x>170) = 1-81.86% = **18.14%**

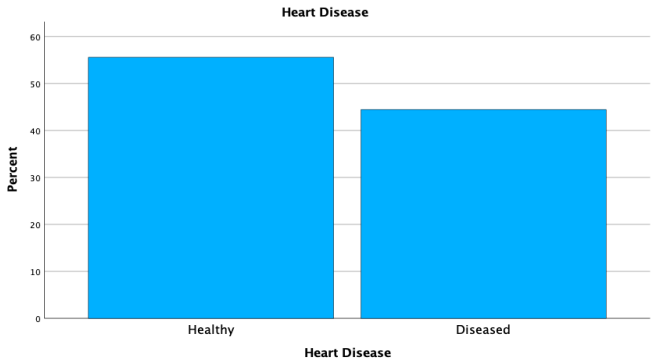
Alternative Method (The Distribution is Not Normal)

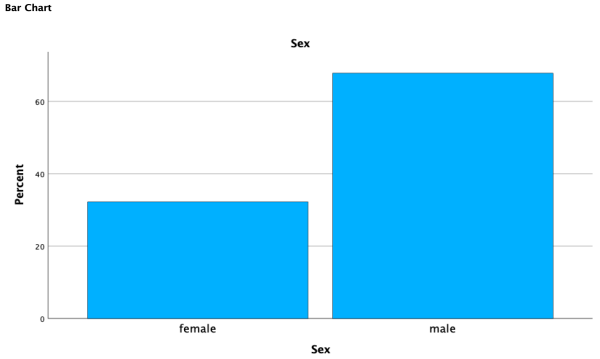
We sort the data and count the patients with maximum heart rate above 170. We found that 48 patients have maximum heart rate above 170. Therefore, the probability of the patients with maximum heart rate above 170 is: 48/270* 100% = **17.77%**

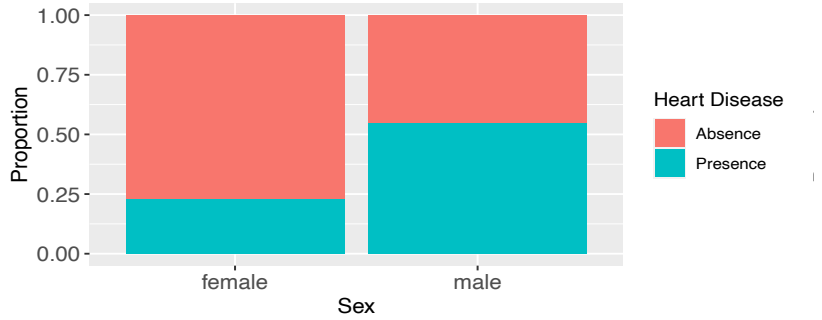
Descriptive Statistics

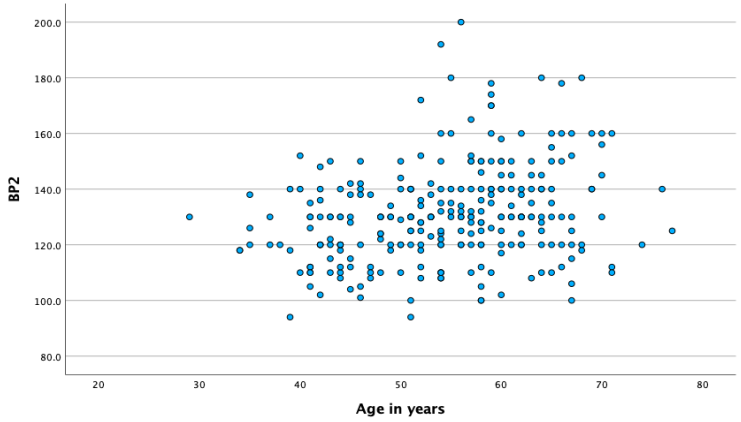
	N	Minimum	Maximum	Sum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
MaxHR2	270	71.0	202.0	40314.4	149.313	1.3779	22.6414	512.634

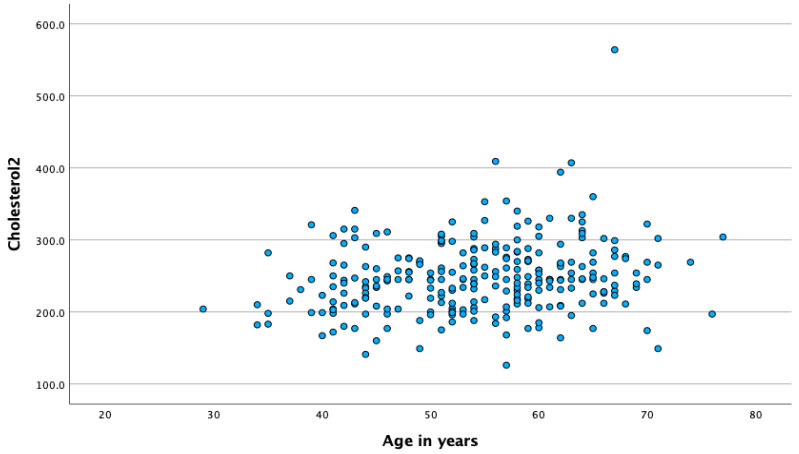
More visual representation of data

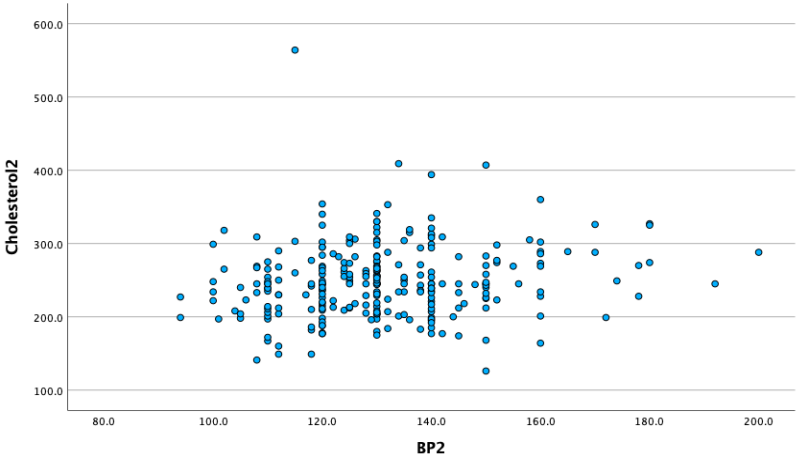












SAMPLING

- We can see that some variables in the current dataset are not normally distributed. This might be caused by the number of samples that are not representative to the population. The current number of dataset is 270.
- By changing (increasing or decreasing) the number of sample, the distribution of the data might be changed. The skewness and kurtosis might be changed.
- For instance, we reduce the amount of data to a half.

SAMPLING

Number of sample: 270

	Age	Sex	Chest pain type	BP	Cholesterol	FBS over 120	EKG results	Max HR	Exercise angina	ST depression	Slope of ST	Number of vessels fluro	Thallium
Valid	270	269	270	255	250	270	270	261	270	264	270	270	270
Missing	0	1	0	15	20	0	0	9	0	6	0	0	0
Mode	54.000	1.000	4.000	120.000	234.000	0.000	2.000	162.000	0.000	0.000	1.000	0.000	3.000
Median	55.000	1.000	3.000	130.000	244.000	0.000	2.000	153.000	0.000	0.800	2.000	0.000	3.000
Mean	54.433	0.680	3.174	131.094	246.224	0.148	1.022	149.149	0.330	1.005	1.585	0.670	4.696
Skewness	−0.164	−0.778	−0.879	0.757	0.228	1.992	−0.045	−0.563	0.729	0.953	0.543	1.210	0.287
Std. Error of Skewness	0.148	0.149	0.148	0.153	0.154	0.148	0.148	0.151	0.148	0.150	0.148	0.148	0.148
Kurtosis	−0.545	−1.406	−0.297	0.927	−0.115	1.983	−2.005	−0.116	−1.480	0.138	−0.607	0.298	−1.901
Std. Error of Kurtosis	0.295	0.296	0.295	0.304	0.307	0.295	0.295	0.300	0.295	0.299	0.295	0.295	0.295

Number of sample: 135

	Age	Sex	Chest pain type	BP	Cholesterol	FBS over 120	EKG results	Max HR	Exercise angina	ST depression	Slope of ST	Number of vessels fluro	Thallium
Valid	135	134	135	128	123	135	135	132	135	133	135	135	135
Missing	0	1	0	7	12	0	0	3	0	2	0	0	0
Mode	57.000	1.000	4.000	120.000	234.000	0.000	2.000	163.000	0.000	0.000	1.000	0.000	3.000
Median	56.000	1.000	4.000	130.000	235.000	0.000	2.000	150.000	0.000	0.600	2.000	0.000	3.000
Mean	54.815	0.664	3.207	130.898	242.374	0.148	1.022	145.856	0.348	0.951	1.578	0.659	4.630
Skewness	−0.072	−0.703	−0.898	0.790	0.178	2.003	−0.045	−0.513	0.645	1.131	0.571	1.218	0.355
Std. Error of Skewness	0.209	0.209	0.209	0.214	0.218	0.209	0.209	0.211	0.209	0.210	0.209	0.209	0.209
Kurtosis	−0.736	−1.529	−0.348	1.075	−0.067	2.043	−2.013	−0.346	−1.608	0.458	−0.579	0.467	−1.865
Std. Error of Kurtosis	0.414	0.416	0.414	0.425	0.433	0.414	0.414	0.419	0.414	0.417	0.414	0.414	0.414

SAMPLING

- In this case, we suggest to take a larger number of sample, to make the observation more representative to the population.
- We also suggest to balance the number of sample between male and female.
- Depending on the objective of the research, we can also take specific sample from our observation. For instance we need to observe the data of the people within the age of 60-70:

	Age	Sex	Chest pain type	BP	Cholesterol	FBS over 120	EKG results	Max HR	Exercise angina	ST depression	Slope of ST	Number of vessels fluro	Thallium
Valid	78	78	78	70	68	78	78	76	78	76	78	78	78
Missing	0	0	0	8	10	0	0	2	0	2	0	0	0
Mode [a]	60.000	1.000	4.000	140.000	254.000	0.000	2.000	132.000	0.000	0.000	2.000	0.000	3.000
Median	64.000	1.000	4.000	136.500	253.500	0.000	2.000	144.000	0.000	1.400	2.000	1.000	3.000
Mean	63.962	0.590	3.308	136.157	256.132	0.167	1.103	139.934	0.359	1.396	1.718	1.115	4.885
Skewness	0.373	−0.372	−1.329	0.291	0.457	1.824	−0.210	−0.654	0.600	0.336	0.201	0.420	0.091
Std. Error of Skewness	0.272	0.272	0.272	0.287	0.291	0.272	0.272	0.276	0.272	0.276	0.272	0.272	0.272
Kurtosis	−0.848	−1.911	0.648	−0.188	0.319	1.362	−2.008	0.074	−1.684	−0.680	−0.542	−1.139	−2.006
Std. Error of Kurtosis	0.538	0.538	0.538	0.566	0.574	0.538	0.538	0.545	0.538	0.545	0.538	0.538	0.538

Conclusion

- After visualization of all the different methods to check normality, we can confirm that our dataset, with its different features, are not normally distributed.

- Possible change in the number of sample size of the features, or use of lognormal in our dataset could be done to achieve normality standards.

Part 3

Parametric tests

INTRODUCTION

Variables that can be found in this dataset includes:

- age, blood pressure, serum cholesterol, fasting blood sugar level, maximum heart rate achieved, ST depression induced by exercise relative to rest, slope of peak exercise ST segment, number of major vessels, Thallium injection, and presence of heart disease.

- In total there are 14 different data in which six are ratio and ordinal, and another four are nominal type of data as shown in Data Description.

Scale or **ratio**: age, blood pressure, cholesterol, maximum heart rate, ST depression.

Ordinal: Chest Pain Type, EKG, slope of peak exercise ST segment, number of major vessels, Thallium injection.

Nominal: Sex, FBS over 120, Exercise angina, and presence of heart disease.

OBJECTIVE Parametric tests

- In general, the main goal is to be able to accurately classify as having or not the heart disease based on some given variables.
    - To get better insight by analyzing the influence of some variables that might lead to the presence of heart disease.

ASSUMPTION AND TEST

- Assumed our data is normally distributed.

- T-test:
  1) One sample t-test (metric scale)
  2) Independent sample t-test (two groups)
  3) Paired-samples t-test (group bet two points in time)

T-TEST

1) One sample t-test

H0 – The patients in the dataset follows the standard *normal value of 120 blood pressure (no matter your age).

H1 – The mean value of the patients is different than specified value.

T-Test

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
BP2	270	131.033	17.6058	1.0715

One-Sample Test

Test Value = 120

	t	df	Significance One-Sided p	Significance Two-Sided p	Mean Difference	95% Confidence Interval of the Difference Lower	95% Confidence Interval of the Difference Upper
BP2	10.298	269	<.001	<.001	11.0333	8.924	13.143

H1 -> There is significant difference between the mean of population (patients) and standard normal blood pressure sample (120)

*Whelton PK, Carey RM, Aronow, WS, Casey DE, Collins KJ, Himmelfarb CD, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *J Am Coll Cardiol.* 2018;71(19):e127–e248.

*National High Blood Pressure Education Program. *The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure* [PDF – 223K]. Bethesda, MD: National Heart, Lung, and Blood Institute; 2003.

POWER AND EFFECT SIZE ANNALYSIS

Dependent variable: cholesterol level

Independent variable: Sex

Univariate Analysis of Variance

Between-Subjects Factors

		N
Sex	0	87
	1	183

Tests of Between-Subjects Effects

Dependent Variable: Cholesterol

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	11661.229[a]	1	11661.229	5.918	.016	.022	5.918	.679
Intercept	14715390.54	1	14715390.54	7468.262	<.001	.965	7468.262	1.000
Sex	11661.229	1	11661.229	5.918	.016	.022	5.918	.679
Error	528064.553	268	1970.390					
Total	17049130.67	270						
Corrected Total	539725.782	269						

a. R Squared = .022 (Adjusted R Squared = .018)

b. Computed using alpha = .05

Effect size: %2.2 → There is a meaningful difference or correlation between groups

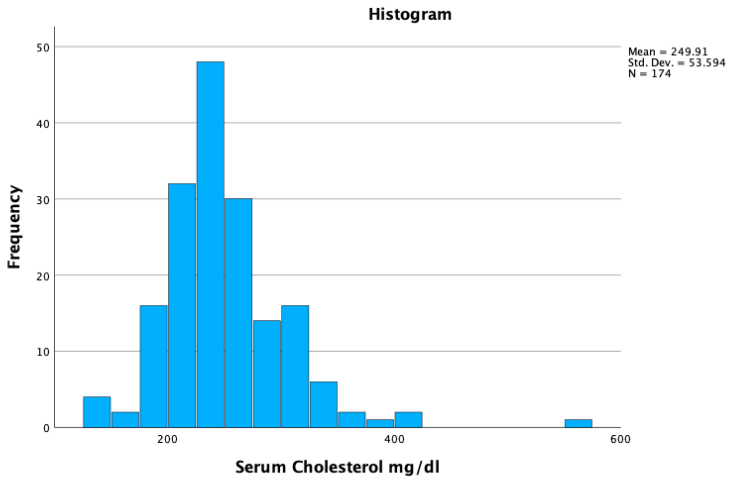
Power: 0.679 → Shows that the number of our sample is enough. So, we can figure out something from this data

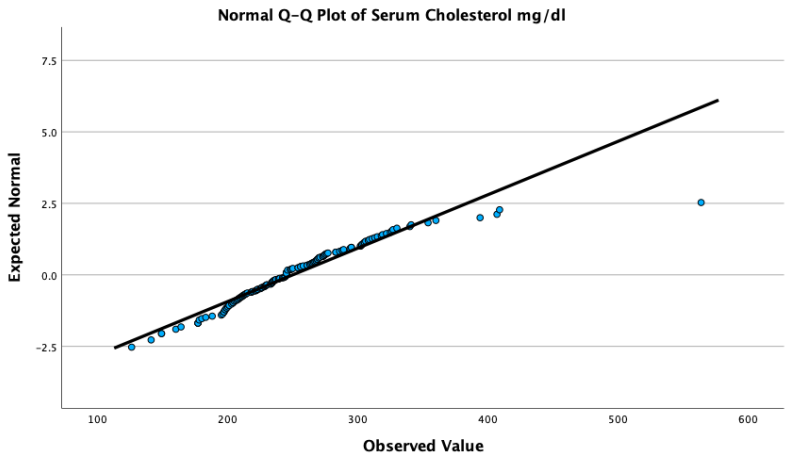
T-TEST

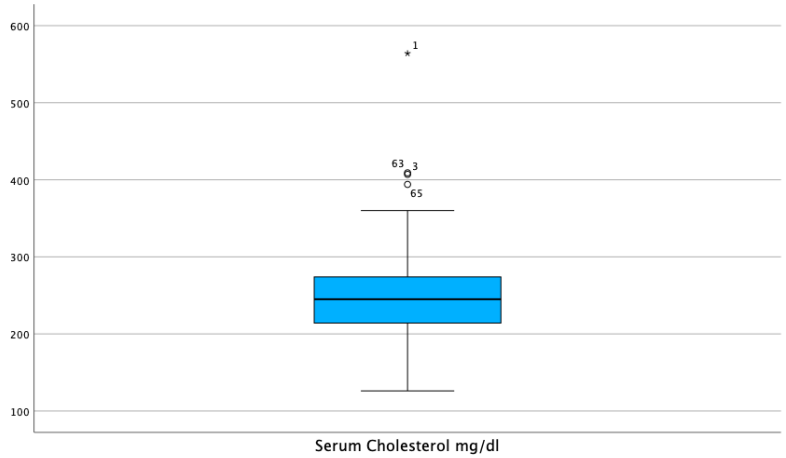
2) Independent sample t-test (two groups)

Subsampled the data between male and female (1 and 0). Using the same sample size for these two independent groups.

Assumed normal.







T-TEST

2) Independent sample t-test (two groups)

H0 -> Cholesterol level is the same for both independent variables (female and male patients).

H1 -> Difference between cholesterol level between females and males.

T-Test

Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
Serum Cholesterol mg/dl	Female	87	261.31	64.061	6.868
	Male	87	238.52	37.568	4.028

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance				95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
Serum Cholesterol mg/dl	Equal variances assumed	10.761	.001	2.863	172	.002	.005	22.793	7.962	7.077	38.509
	Equal variances not assumed			2.863	138.896	.002	.005	22.793	7.962	7.051	38.535

H1 -> Cholesterol level is significantly different between the male and female groups.

T-TEST

3) Paired-samples t-test (Same variable with different value)

T-Test

```
[DataSet1] C:\Users\suley\OneDrive\Documents\Job\Interviews\Brainnest\Data\Heart_Disease_3_MVA.sav
```

Warnings

The Paired Samples Correlations table is not produced.
The Paired Samples Test table is not produced.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Cholesterol	246.22[a]	250	45.445	2.874
	TREND(Cholesterol)	246.224[a]	250	45.4446	2.8742

a. The correlation and t cannot be computed because the standard error of the difference is 0.

Cholesterol= Variable with missing values.
TREND(Cholesterol)= Variable with replaced missing values.

The result shows that both variable has similar Mean.

ANOVA

1) One way (only one independent variable)
2) Two way (two independent variables)

ANOVA

1) One way (only one independent variable)

H0 -> Chest pain type has no influence on blood pressure.

H1 -> Chest pain type has an influence on blood pressure.

Oneway

Tests of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
BP	Based on Mean	1.436	3	266	.233
	Based on Median	1.381	3	266	.249
	Based on Median and with adjusted df	1.381	3	265.251	.249
	Based on trimmed mean	1.316	3	266	.270

→ Our data is homogeneous

ANOVA

BP

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2788.216	3	929.405	3.475	.017
Within Groups	71147.701	266	267.473		
Total	73935.917	269			

→ 0.017<0.05 → H0 reject

ANOVA

2) Two way (two independent variables)

Hypotheses

1) H0 -> Sex has no influence on blood pressure.

H1 -> Sex has an influence on blood pressure.

2) H0 -> EKG results has no influence on blood pressure.

H1 -> EKG results has an influence on blood pressure.

3) H0 -> Sex has no influence on EKG results .

H1 -> Sex has an influence on EKG results .

ANOVA

2) Two way (two independent variables)

Tests of Between-Subjects Effects

Dependent Variable: BP

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	2314.187[a]	4	578.547	2.141	.076	.031	8.562	.630
Intercept	533070.382	1	533070.382	1972.357	<.001	.882	1972.357	1.000
Sex	141.335	1	141.335	.523	.470	.002	.523	.111
EKGresults	1985.054	2	992.527	3.672	.027	.027	7.345	.673
Sex * EKGresults	1.489	1	1.489	.006	.941	.000	.006	.051
Error	71621.730	265	270.271					
Total	4691585.474	270						
Corrected Total	73935.917	269						

a. R Squared = .031 (Adjusted R Squared = .017)

b. Computed using alpha = .05

Accept H0

Reject H0

Accept H0

ANOVA

2) Two way (two independent variables) → Heart Disease

Hypotheses

1) H0 -> Chest Pain Type has no influence on Heart Disease.
   H1 -> Chest Pain Type has an influence on Heart Disease.
2) H0 -> Cholesterol has no influence on Heart Disease.
   H1 -> Cholesterol has an influence on Heart Disease.
3) H0 -> Chest Pain Type has no correlation with Cholesterol.
   H1 -> Chest Pain Type has correlation with Cholesterol.

Simmons II, B. (2021). *Investigating Heart Disease Datasets and Building Predictive Models* (Doctoral dissertation, Elizabeth City State University). https://libres.uncg.edu/ir/ecsu/f/Brandon_Simmons_Thesis-Final.pdf

V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D. https://archive.ics.uci.edu/ml/datasets/heart+disease

Chest Pain Type
1: typical angina
2: atypical angina
3: non-anginal pain
4: asymptomatic

Heart Disease (Transformed)
0: Absence
1: Presence

Cholesterol (Transformed)
1: Normal (<200)
2: Risk (200-239)
3: High Risk (>240)

ANOVA

2) Two way (two independent variables)

ANOVA - Heart_Disease

Cases	Sum of Squares	df	Mean Square	F	p
Chest_pain_type	12.301	3	4.100	22.048	< .001
Cholesterol	1.032	2	0.516	2.775	0.064
Chest_pain_type ✻ Cholesterol	0.737	6	0.123	0.660	0.682
Residuals	47.982	258	0.186		

Reject H0

Accept H0

Accept H0

Kruskal-Wallis Test

Kruskal-Wallis Test

Factor	Statistic	df	p
Chest_pain_type	68.334	3	< .001
Cholesterol	7.595	2	0.022

Conclusions and Suggestions

The Blood Pressure data for this sample has higher mean than the suggested value of 120 blood pressure

Cholesterol level is significantly different between the male and female groups.

Sex has no influence on blood pressure.

EKG results has an influence on blood pressure.

Chest Pain Type has an influence on Heart Disease.

Cholesterol has no direct influence on Heart Disease.

Suggestion

Since many reference said that Cholesterol actually has influence on the presence of Heart Disease, we suggest that more representative dataset is required.

Part 4

Non-parametric tests

INTRODUCTION

Variables that can be found in this dataset includes:

- age, blood pressure, serum cholesterol, fasting blood sugar level, maximum heart rate achieved, ST depression induced by exercise relative to rest, slope of peak exercise ST segment, number of major vessels, Thallium injection, and presence of heart disease.

- In total there are 14 different data in which six are ratio and ordinal, and another four are nominal type of data as shown in Data Description.

Scale or **ratio**: age, blood pressure, cholesterol, maximum heart rate, ST depression.

Ordinal: Chest Pain Type, EKG, slope of peak exercise ST segment, number of major vessels, Thallium injection.

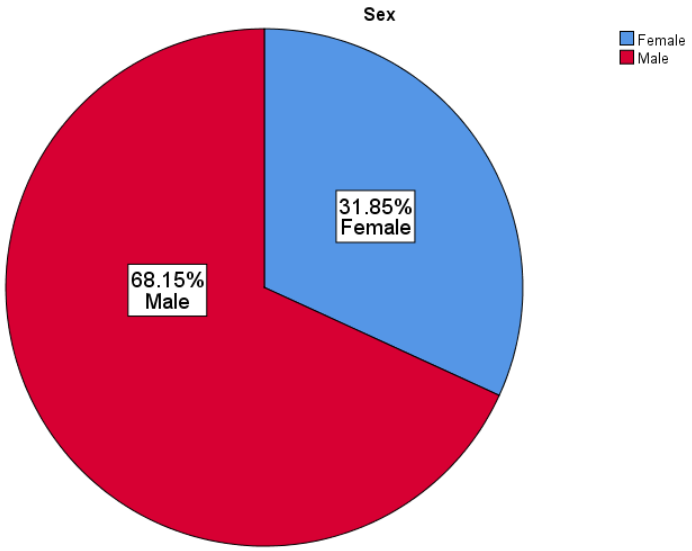
Nominal: Sex, FBS over 120, Exercise angina, and presence of heart disease.

OBJECTIVE

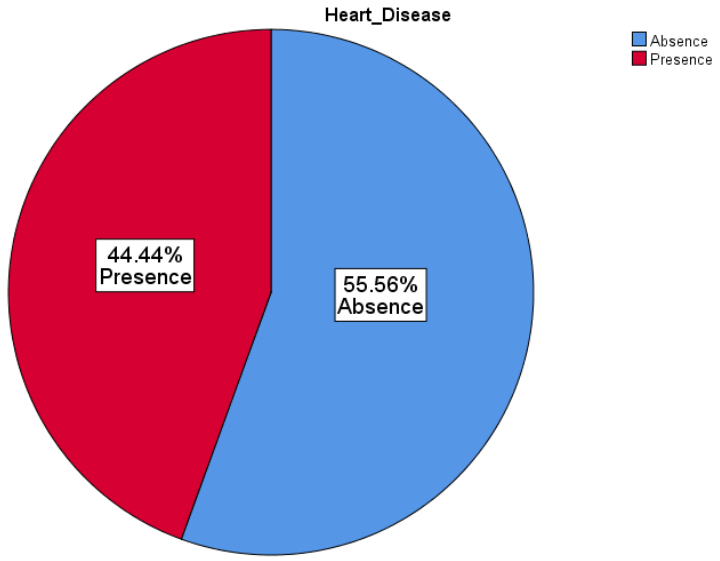
- In general, the main goal is to be able to accurately classify as having or not the heart disease based on some given variables.
  - To get better insight by analyzing the influence of some variables that might lead to the presence of heart disease with reviewing of non-parametric tests and correlation.
  - Using a linear regression model for features that may affect heart disease

Descriptive

- Heart disease and Gender



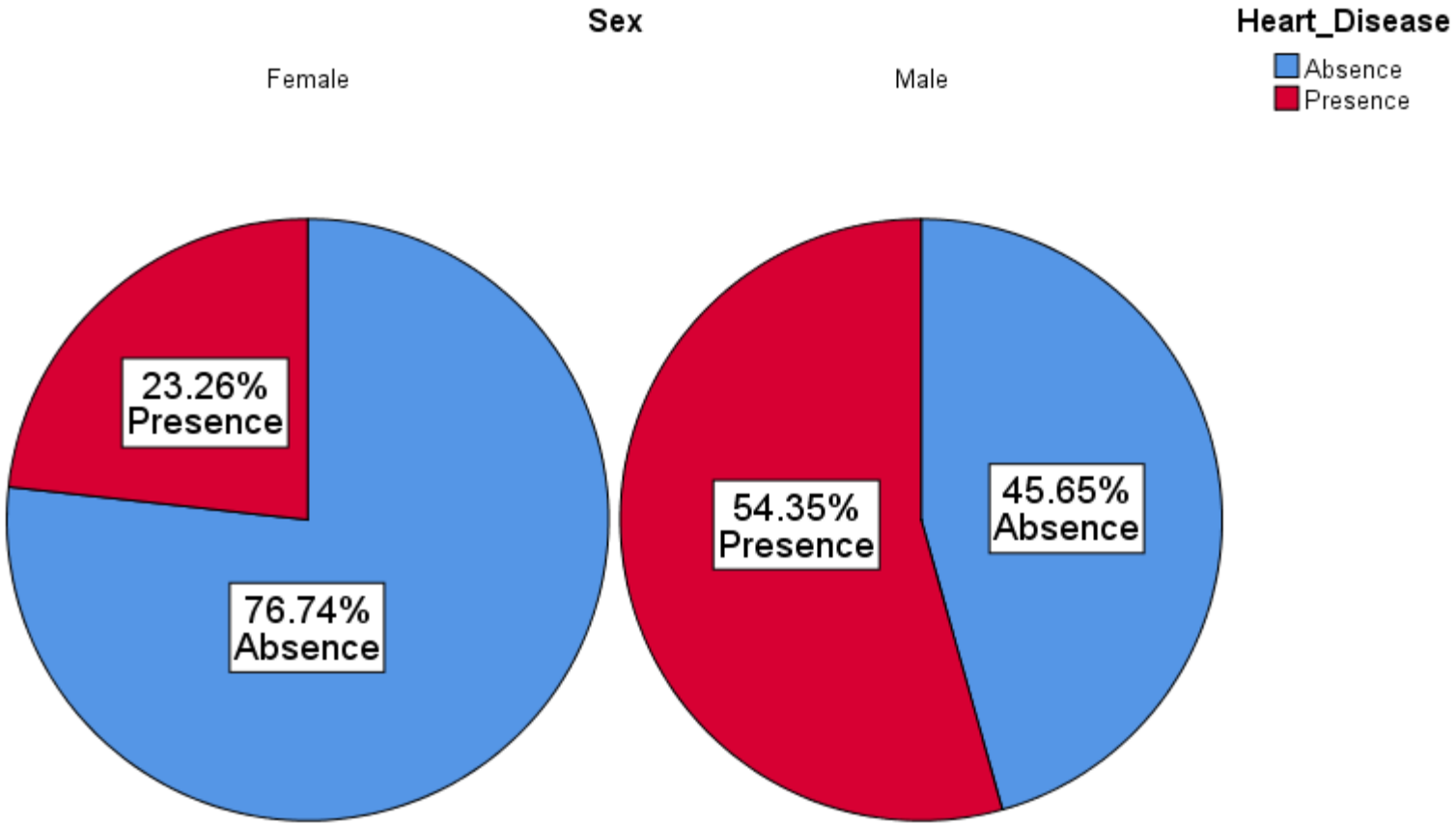
From 270 patients, 68.15% are Male (nominal symbol = 1), and 31.85% are Female (nominal symbol = 0).



Among 270 patients, 44.44% are having Heart Disease (nominal symbol = 1), and 55.56% are not (nominal symbol = 0).

Descriptive

- Heart disease and Gender



In the Female group, only 23.26% are having the Heart Disease. In the Male group, 54.35% are having the Heart Disease. **Why are the Male group having more Heart Disease compared to the Female group?**

Maybe the Male group are having a higher blood pressure, or higher cholesterol level, or higher maximum heart rate?

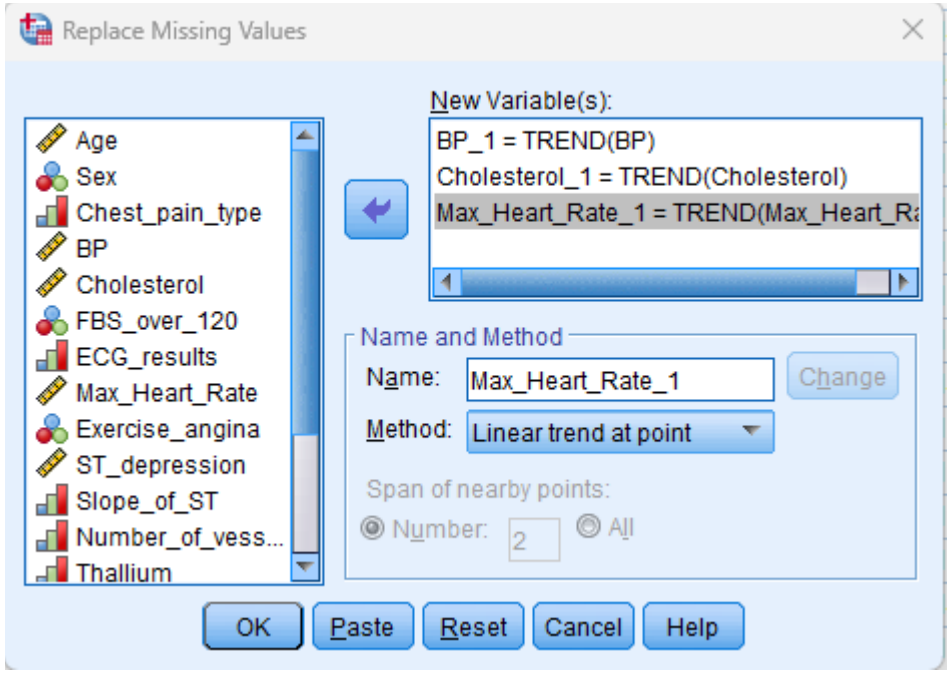
We will test these three variables using Mann-Whitney U-Test.

Tests

- We see in Part 2 that our data is not normally distributed. We will use Nonparametric test:

- Wilcoxon test:
  1) One sample (metric scale)
  2) Paired-samples/ two dependent samples (group between two points in time). We use this test to see if there is significant difference in some variables between before and after replacing the missing values. The missing values are replaced using **Linear trend at point**:

     Replaces missing values with the linear trend for that point. The existing series is regressed on an index variable scaled 1 to n. Missing values are replaced with their predicted values. (https://www.ibm.com/docs/sr/spss-statistics/beta?topic=values-estimation-methods-replacing-missing)



Tests

- We see in Part 2 that our data is not normally distributed. We will use Nonparametric test:

- Mann-Whitney U-Test. Two independent sample (two groups, Male and Female)
- Kruskal Wallis test (alternative to ANOVA)
- Correlation test (Spearman Correlation)

In addition, we also performed:

- Multiple Linear Regression
- Logistic Regression

Wilcoxon Test: One sample (metric scale)

- Hypothesis:
  - H0 – The patients' central tendencies in the dataset follows the standard *normal value of 120 mm Hg blood pressure (no matter your age).
  - H1 – The central tendencies of patients is different than specified value (120 mm Hg).

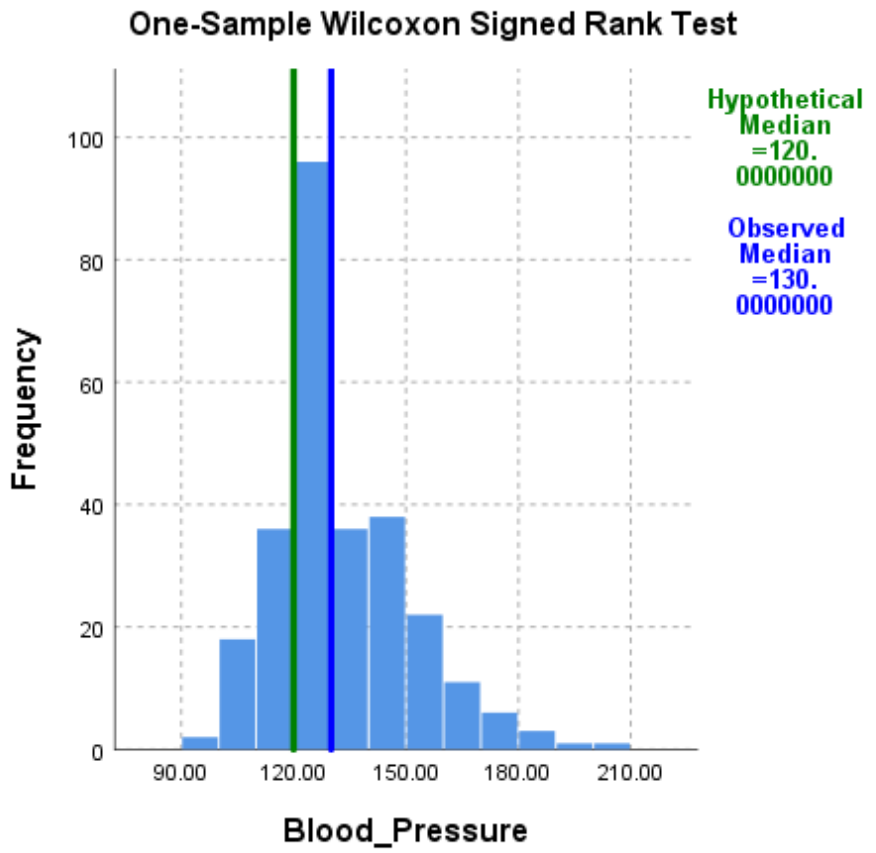
BLOOD PRESSURE CATEGORY	SYSTOLIC mm Hg (upper number)	and/or	DIASTOLIC mm Hg (lower number)
NORMAL	LESS THAN 120	and	LESS THAN 80
ELEVATED	120 – 129	and	LESS THAN 80
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 1	130 – 139	or	80 – 89
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 2	140 OR HIGHER	or	90 OR HIGHER
HYPERTENSIVE CRISIS (consult your doctor immediately)	HIGHER THAN 180	and/or	HIGHER THAN 120

Healthy and unhealthy blood pressure ranges
https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings

One-Sample Wilcoxon Signed Rank Test Summary

Total N	270
Test Statistic	23670.000
Standard Error	1055.244
Standardized Test Statistic	9.068
Asymptotic Sig.(2-sided test)	.000

H1 -> There is significant difference between the central tendencies of population (patients) and standard normal blood pressure sample (120)



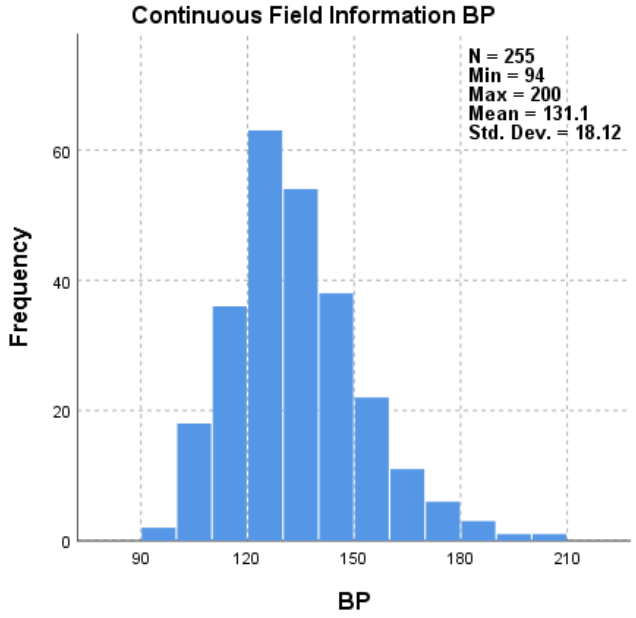
Wilcoxon Test: Paired-samples/ two dependent samples (group between two points in time) → Blood Pressure

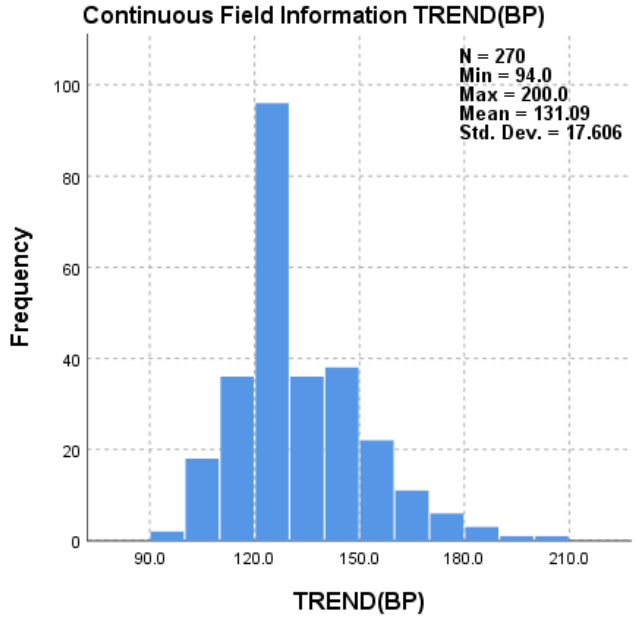
- Hypothesis:
  - H0 – The central tendencies of Blood Pressure from the two dependent samples are the same.
  - H1 – The central tendencies of Blood Pressure from two dependent samples are unequal in the population.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between BP and TREND(BP) equals 0.	Related-Samples Wilcoxon Signed Rank Test	1.000	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .050.





H0 – The central tendencies of Blood Pressure from the two dependent samples are the same.

There is no significant difference after replacing the missing values.

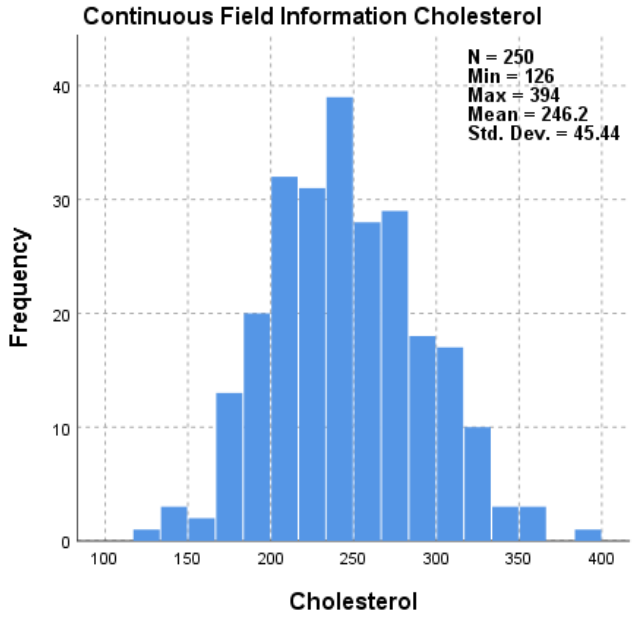
Wilcoxon Test: Paired-samples/ two dependent samples (group between two points in time) → Cholesterol

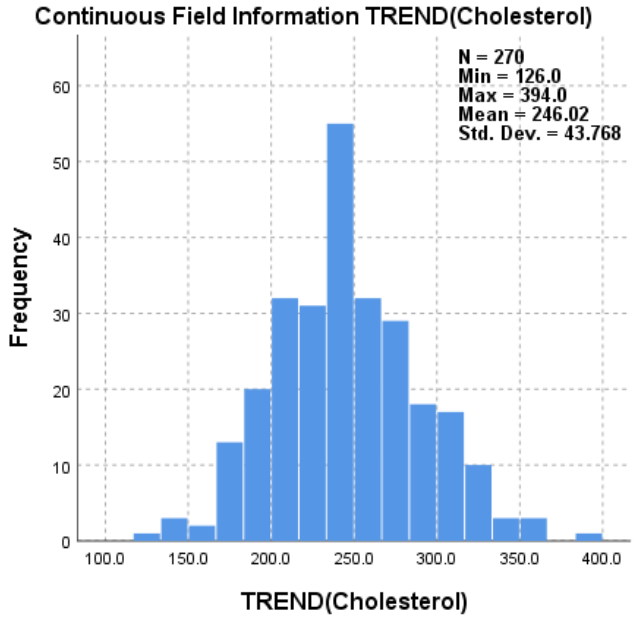
- Hypothesis:
  - H0 – The central tendencies of Cholesterol from the two dependent samples are the same.
  - H1 – The central tendencies of Cholesterol from two dependent samples are unequal in the population.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between Cholesterol and TREND(Cholesterol) equals 0.	Related-Samples Wilcoxon Signed Rank Test	1.000	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .050.





H0 – The central tendencies of Cholesterol from the two dependent samples are the same.

There is no significant difference after replacing the missing values.

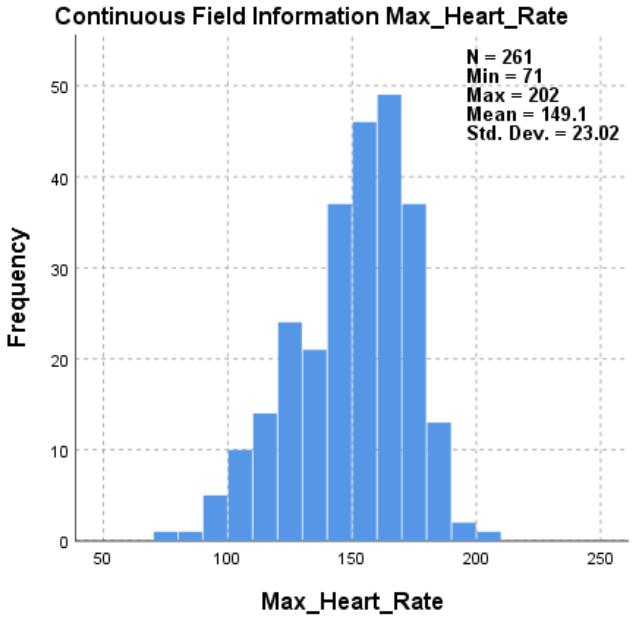
Wilcoxon Test: Paired-samples/ two dependent samples (group between two points in time) → Maximum Heart Rate

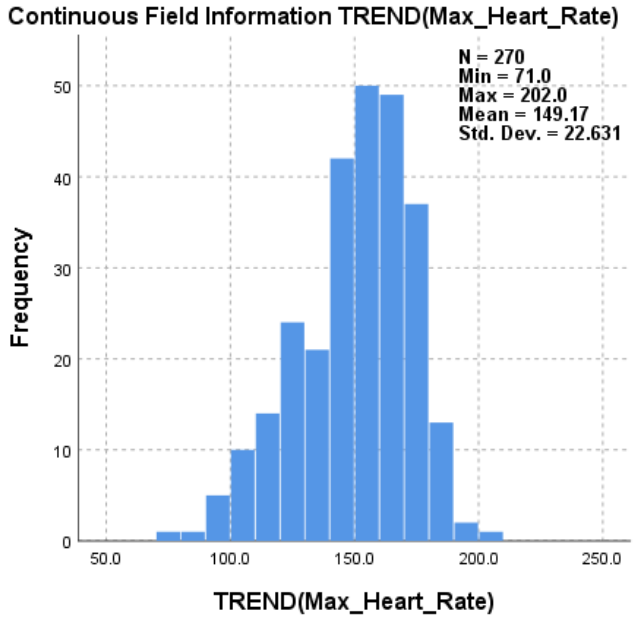
- Hypothesis:
  - H0 – The central tendencies of Maximum Heart Rate from the two dependent samples are the same.
  - H1 – The central tendencies of Maximum Heart Rate from two dependent samples are unequal in the population.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between Max_Heart_Rate and TREND(Max_Heart_Rate) equals 0.	Related-Samples Wilcoxon Signed Rank Test	1.000	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .050.





H0 – The central tendencies of Maximum Heart Rate from the two dependent samples are the same.

There is no significant difference after replacing the missing values.

Mann-Whitney U-Test: Two independent sample (two groups, Male and Female), **Same Sample Size**.

- Hypothesis:
  - H0 – The sum of the rankings of Blood Pressure/Cholesterol/Maximum Heart Rate from the two independent samples are the same.
  - H1 – The sum of the rankings of Blood Pressure/Cholesterol/Maximum Heart Rate from two independent samples are unequal in the population.

	W	df	p	Rank-Biserial Correlation	SE Rank-Biserial Correlation	95% CI for Rank-Biserial Correlation	
						Lower	Upper
Blood_Pressure	3985.500		0.378	0.078	0.088	−0.095	0.246
Cholesterol_1	4341.000		0.049	0.174	0.088	0.003	0.335
Max_Heart_Rate	4484.500		0.016	0.213	0.088	0.043	0.371

Note. For the Mann-Whitney test, effect size is given by the rank biserial correlation.
Note. Mann-Whitney U test.

Blood Pressure

H0 – The sum of the rankings of Blood Pressure from the two independent samples are the same.

There is no significant difference between Male and Female in Blood Pressure.

Cholesterol

H1 – The sum of the rankings of Cholesterol from two independent samples are unequal in the population.

There is significant difference between Male and Female in Cholesterol.

Maximum Heart Rate

H1 – The sum of the rankings of Maximum Heart Rate from two independent samples are unequal in the population.

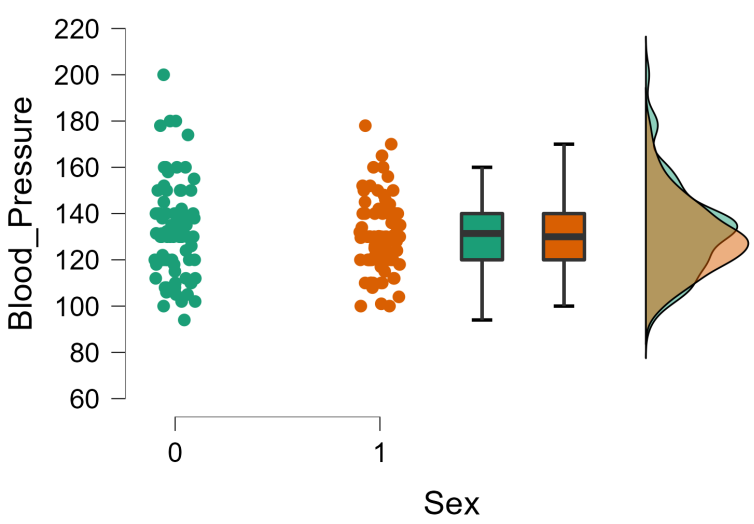
There is significant difference between Male and Female in Maximum Heart Rate.

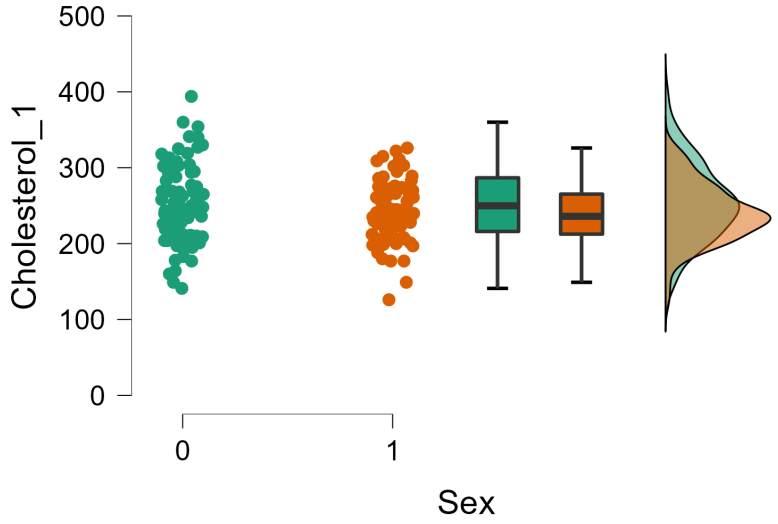
Effect size	Trivial	Small	Medium	Large
Rank -biserial (r_B)	<0.1	0.1	0.3	0.5

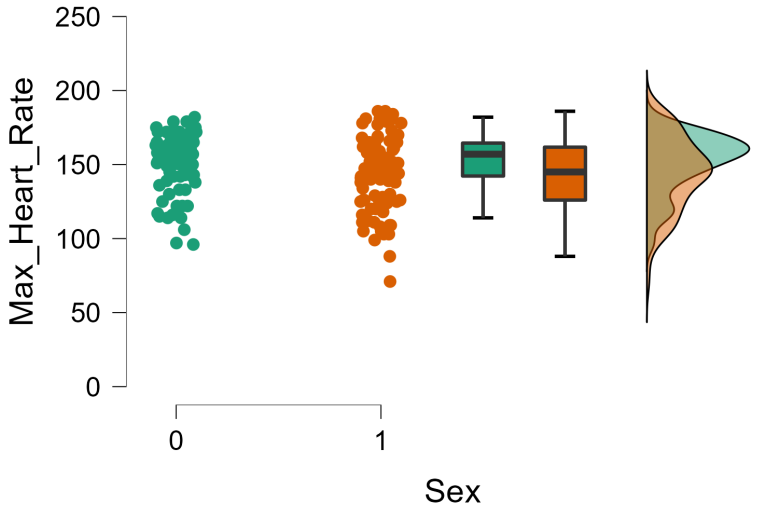
Group Descriptives

	Group	N	Mean	SD	SE	Coefficient of variation
Blood_Pressure	0	86	132.710	19.979	2.154	0.151
	1	86	129.811	15.851	1.709	0.122
Cholesterol_1	0	86	253.363	50.146	5.407	0.198
	1	86	238.998	37.987	4.096	0.159
Max_Heart_Rate	0	86	151.613	19.576	2.111	0.129
	1	86	143.542	24.843	2.679	0.173

Mann-Whitney U-Test: Two independent sample (two groups, Male and Female), Same Sample Size.







Mann-Whitney U-Test: Two independent sample (two groups, Male and Female), Different Sample Size.

- Hypothesis:
  - H0 – The sum of the rankings of Blood Pressure/Cholesterol/Maximum Heart Rate from the two independent samples are the same.
  - H1 – The sum of the rankings of Blood Pressure/Cholesterol/Maximum Heart Rate from two independent samples are unequal in the population.

	W	df	p	Rank-Biserial Correlation	SE Rank-Biserial Correlation	95% CI for Rank-Biserial Correlation	
						Lower	Upper
Blood_Pressure	8463.500		0.356	0.070	0.075	−0.078	0.215
Cholesterol_1	8912.500		0.094	0.126	0.075	−0.021	0.268
Max_Heart_Rate	8683.000		0.197	0.097	0.075	−0.050	0.241

Note. For the Mann-Whitney test, effect size is given by the rank biserial correlation.
Note. Mann-Whitney U test.

Effect size	Trivial	Small	Medium	Large
Rank -biserial (r_B)	<0.1	0.1	0.3	0.5

We tried to include the whole group in Mann-Whitney U-Test, since we can run this test with **unequal sample size**. In this test, we can see that the effect size is small.

Blood Pressure

H0 – The sum of the rankings of Blood Pressure from the two independent samples are equal.

There is no significant difference between Male and Female in Blood Pressure.

Cholesterol

H0 – The sum of the rankings of Cholesterol from two independent samples are equal.

There is no significant difference between Male and Female in Cholesterol.

Maximum Heart Rate

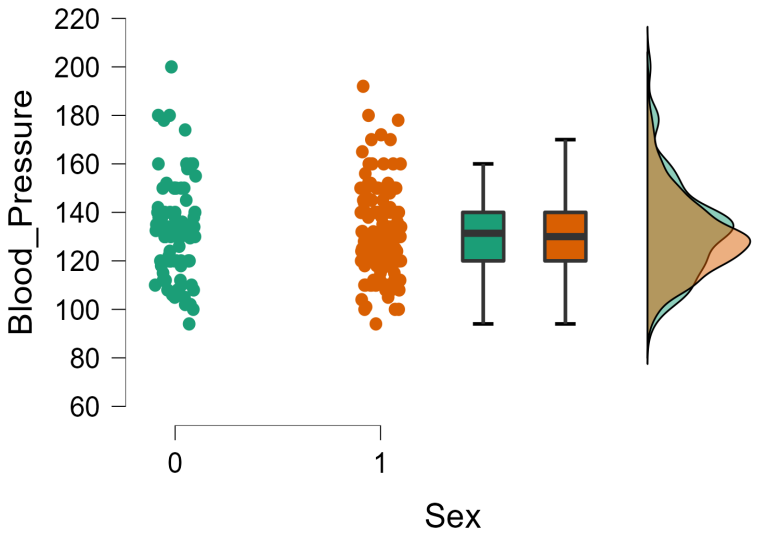
H0 – The sum of the rankings of Maximum Heart Rate from two independent samples are equal.

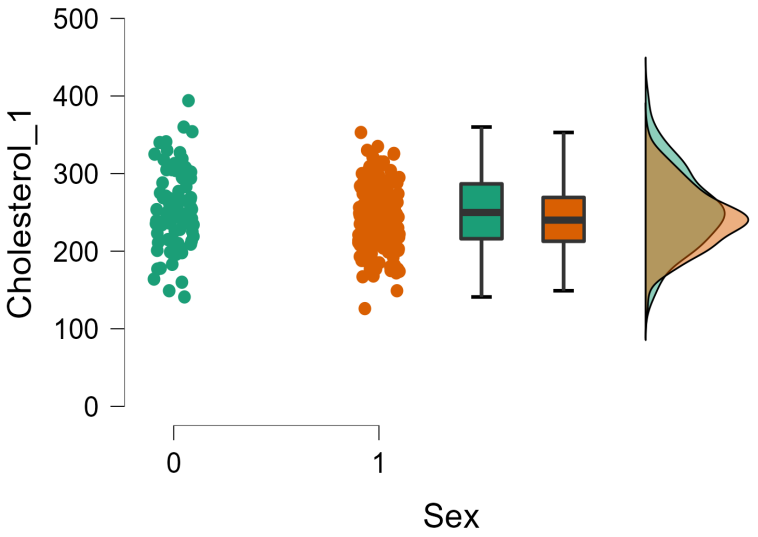
There is no significant difference between Male and Female in Maximum Heart Rate.

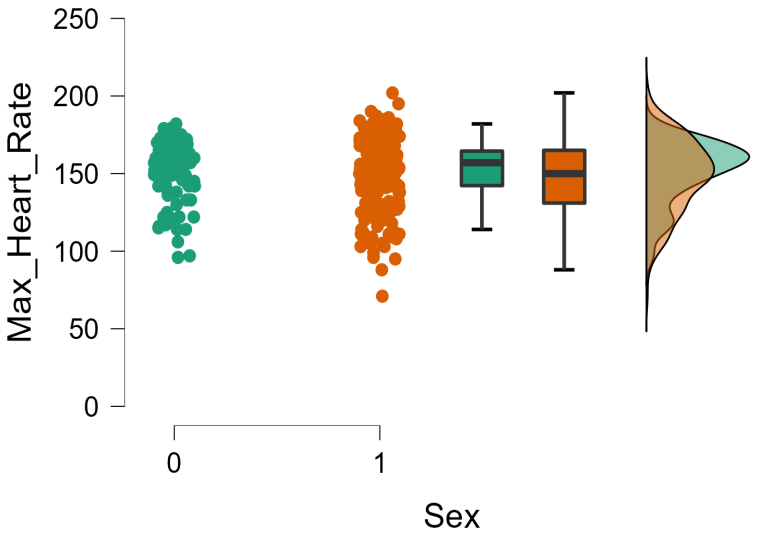
Group Descriptives

	Group	N	Mean	SD	SE	Coefficient of variation
Blood_Pressure	0	86	132.710	19.979	2.154	0.151
	1	184	130.330	16.384	1.208	0.126
Cholesterol_1	0	86	253.363	50.146	5.407	0.198
	1	184	242.592	40.134	2.959	0.165
Max_Heart_Rate	0	86	151.613	19.576	2.111	0.129
	1	184	148.027	23.890	1.761	0.161

Mann-Whitney U-Test: Two independent sample (two groups, Male and Female), Different Sample Size.







Mann-Whitney U-Test: Two independent sample (two groups, Male and Female), Different Sample Size.

- Hypothesis:
  - H0 – The sum of the rankings of Heart Disease from the two independent samples are the same.
  - H1 – The sum of the rankings of Heart Disease from two independent samples are unequal in the population.

	W	df	p	Rank-Biserial Correlation	SE Rank-Biserial Correlation	95% CI for Rank-Biserial Correlation	
						Lower	Upper
Heart_Disease	5452.000		< .001	−0.311	0.075	−0.438	−0.172

Note. For the Mann-Whitney test, effect size is given by the rank biserial correlation.
Note. Mann-Whitney U test.

Effect size	Trivial	Small	Medium	Large
Rank -biserial (r_B)	<0.1	0.1	0.3	0.5

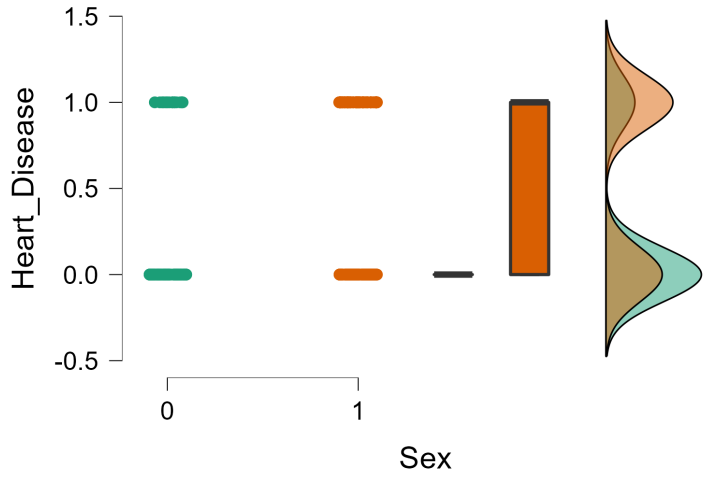
Group Descriptives

	Group	N	Mean	SD	SE	Coefficient of variation
Heart_Disease	0	86	0.233	0.425	0.046	1.827
	1	184	0.543	0.499	0.037	0.919

<u>Heart Disease</u>

H1 – The sum of the rankings of Heart Disease from the two independent samples are unequal in the population.

There is significant difference between Male and Female in Heart Disease.



Although we cannot find significant difference in Blood Pressure/Cholesterol/Maximum Heart Rate from the two group (Male & Female), the occurrence of the Heart Disease is significant. The negative value of Rank-Biserial Correlation indicates that all values of the second sample (Male, 1) are larger than all the values of the first sample (Female, 0).

Non-Parametric Test: Kruskal-Wallis

Practitioners want to know if there is a difference in the Blood pressure of patients who have different Chest Pain Types.

First should be checked Normality.

Descriptives

			Statistic	Std. Error
Blood_Pressure	Mean		131.0881666	1.071466676
	95% Confidence Interval for Mean	Lower Bound	128.9786394	
		Upper Bound	133.1976937	
	5% Trimmed Mean		130.2461110	
	Median		130.0000000	
	Variance		309.971	
	Std. Deviation		17.60599403	
	Minimum		94.0000000	
	Maximum		200.0000000	
	Range		106.0000000	
	Interquartile Range		20.0000000	
	Skewness		.779	.148
	Kurtosis		1.157	.295

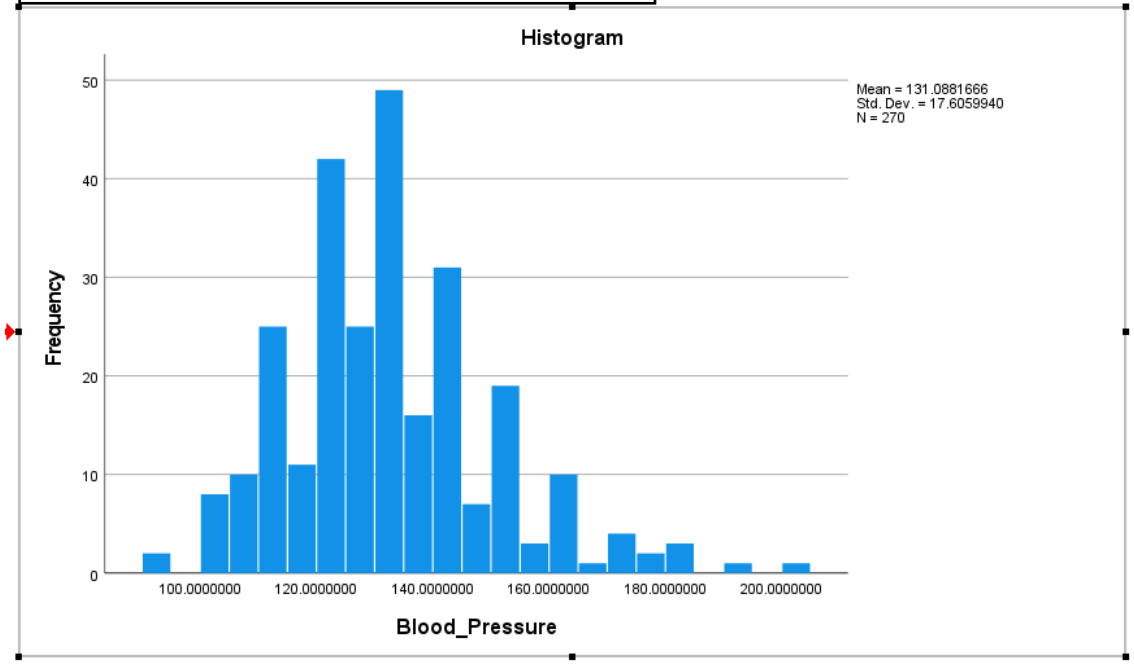
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Blood_Pressure	.103	270	<.001	.961	270	<.001

a. Lilliefors Significance Correction

Chest pain types:

- Value 1: typical angina
- Value 2: atypical angina
- Value 3: non-anginal pain
- Value 4: asymptomatic



Non-Parametric Test: Kruskal-Wallis

H0 -> There is no difference between Blood pressure of patients who have different categories of Chest Pain Types.

H1 -> There is a difference between Blood pressure of patients who have different categories of Chest Pain Types.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of BP is the same across categories of Chest pain type.	Independent-Samples Kruskal-Wallis Test	.093	Retain the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.

H0 -> There is not a difference between Blood pressure of patients who have different categories of Chest Pain Types.

Practitioners want to know if there is a difference in the Blood pressure of patients who have different Cholesterol Types.

H0 -> There is no difference between Blood pressure of patients who have different categories of Cholesterol Types.

H1 -> There is a difference between Blood pressure of patients who have different categories of Cholesterol Types.

Cholesterol Types:
- 1: Normal (<200)
- 2: Risk (200-239)
- 3: High Risk (>240)

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of Blood_Pressure is the same across categories of Cholesterol.	Independent-Samples Kruskal-Wallis Test	.059	Retain the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.

H0 -> There is not a difference between Blood pressure of patients who have different categories of Cholesterol Types.

Correlation test

Correlation

H0 -> There is no correlation between the ST depression and Heart Disease.

H1 -> In contrast, the alternative hypothesis assumes that there is a correlation between ST depression and Heart Disease.

Spearman Correlation test is performed because the data is not Normal.

Correlations

			ST_depression	Heart_Disease
Spearman's rho	ST_depression	Correlation Coefficient	1.000	.391**
		Sig. (2-tailed)	.	<.001
		N	270	270
	Heart_Disease	Correlation Coefficient	.391**	1.000
		Sig. (2-tailed)	<.001	.
		N	270	270

**. Correlation is significant at the 0.01 level (2-tailed).

H1 -> There is a correlation between ST depression and Heart Disease.

Regression tests

Multiple Linear Regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.337[a]	.114	.107	.9940846255

a. Predictors: (Constant), Exercise_angina, Blood_Pressure

Only 11.4 percentage a variability of model can be explained by independent variables

We don't have sufficient independent variables

ST depression= 0.01(BP) + 0.628(Exercise angina) -0.537

Independent variables have an impact on Cholesterol

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	33.806	2	16.903	17.105	<.001[b]
	Residual	263.851	267	.988		
	Total	297.657	269			

a. Dependent Variable: ST_depression

b. Predictors: (Constant), Exercise_angina, Blood_Pressure

ST depression can be predicted by BP, 17.1 percent

ST depression can be predicted by Exercise, 28.1 percent

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-.537	.456		-1.179	.239
	Blood_Pressure	.010	.003	.171	2.955	.003
	Exercise_angina	.628	.129	.281	4.875	<.001

a. Dependent Variable: ST_depression

Multiple Linear Regression

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.216[a]	.047	.040	42.89003839

a. Predictors: (Constant), Exercise_angina, Blood_Pressure

b. Dependent Variable: Cholesterol_1

Only 5 percentage a variability of model cand be explained by independent variables

We don't have sufficient independent variables

Cholesterol = 0.433(BP) + 11.06(Exercise angina) + 185.574

Independent variables have an impact on Cholesterol

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	24151.244	2	12075.622	6.564	.002[b]
	Residual	491161.290	267	1839.555		
	Total	515312.534	269			

a. Dependent Variable: Cholesterol_1

b. Predictors: (Constant), Exercise_angina, Blood_Pressure

Cholesterol can be predicted by BP, 17.4 percent

Cholesterol can be predicted by Exercise, 11.9 percent

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	185.574	19.659		9.440	<.001
	Blood_Pressure	.433	.149	.174	2.913	.004
	Exercise_angina	11.060	5.561	.119	1.989	.048

a. Dependent Variable: Cholesterol_1

Logistic Regression

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	164.464[a]	.535	.716

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

53.5 to 71.6 percentage a variability of model cand be explained by independent variables

if the probability of a case being classified into the "yes" category is greater than .500, then that particular case is classified into the "yes" category.

92 percentage of cases that can be correctly classified as "no"

Classification Table[a]

Observed		Predicted		
		Heart_Disease		Percentage Correct
		0	1	
Step 1	Heart_Disease 0	138	12	92.0
	1	21	99	82.5
	Overall Percentage			87.8

a. The cut value is .500

Logistic Regression

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Age	-.022	.026	.681	1	.409	.979	.929	1.030
	Sex(1)	1.728	.578	8.928	1	.003	5.631	1.812	17.495
	Chest_pain_type			16.438	3	<.001			
	Chest_pain_type(1)	1.341	.922	2.116	1	.146	3.824	.628	23.299
	Chest_pain_type(2)	.704	.790	.793	1	.373	2.021	.430	9.507
	Chest_pain_type(3)	2.516	.797	9.957	1	.002	12.374	2.594	59.032
	Blood_Pressure	.026	.013	4.315	1	.038	1.027	1.001	1.052
	Cholesterol_1	.010	.005	4.027	1	.045	1.010	1.000	1.021
	FBS_over_120(1)	-.549	.603	.831	1	.362	.577	.177	1.881
	ECG_results			2.499	2	.287			
	ECG_results(1)	.845	3.362	.063	1	.801	2.329	.003	1692.122
	ECG_results(2)	.655	.416	2.481	1	.115	1.925	.852	4.351
	Max_Heart_Rate	-.020	.012	2.795	1	.095	.980	.958	1.003
	Exercise_angina(1)	.488	.460	1.129	1	.288	1.630	.662	4.011
	ST_depression	.348	.258	1.821	1	.177	1.416	.854	2.348
	Slope_of_ST			5.592	2	.061			
	Slope_of_ST(1)	1.190	.510	5.441	1	.020	3.289	1.209	8.941
	Slope_of_ST(2)	.515	.931	.306	1	.580	1.674	.270	10.392
	Number_of_vessels_fluor			24.540	3	<.001			
	Number_of_vessels_fluor(1)	2.033	.545	13.898	1	<.001	7.634	2.622	22.226
	Number_of_vessels_fluor(2)	3.035	.798	14.464	1	<.001	20.800	4.353	99.386
	Number_of_vessels_fluor(3)	2.329	.910	6.545	1	.011	10.267	1.724	61.143
	Thallium			12.918	2	.002			
	Thallium(1)	-.026	.862	.001	1	.976	.975	.180	5.278
	Thallium(2)	1.593	.467	11.618	1	<.001	4.919	1.968	12.294
	Constant	-7.808	3.261	5.733	1	.017	.000		

The chance of having heart disease ("yes" category) is 5.631 times greater for males as opposed to females.

did not add significantly to the model

Significant features to the model:
7 feature
Age, Sex, Chest pain type, BP, Cholesterol, Number vessels, and Thallium

Conclusion

- Non-parametric tests shown to be the most suited to our dataset, since our data is not normally distributed.
- In Mann-Whitney U-Test, the difference in Blood Pressure/Cholesterol/Maximum Heart Rate from the two group (Male & Female) is not significant. However, the occurrence of the Heart Disease is significant, with the Male group having a higher occurrence of Heart Disease.
- There is a correlation between ST depression and Heart Disease.
- Regression model - Significant features to the model: Age, Sex, Chest pain type, BP, Cholesterol, Number vessels, and Thallium.
- Males are having 5.631 times greater risk of heart disease.

Thank you!

Group A