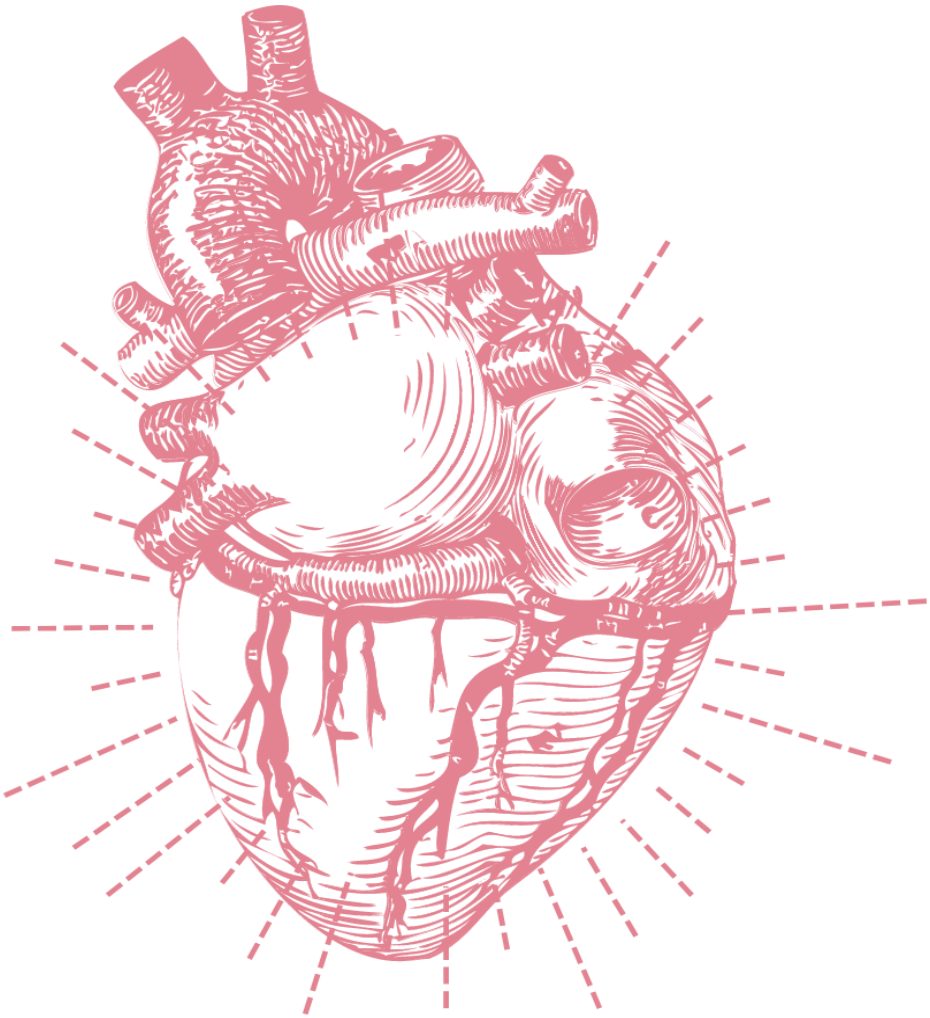
Heart Diseases Prediction



Prepared by :

Mohamed Samy AbdulKareem (Sec. 24)

Youssef AbdulHameed Farag (Sec. 35)

Nourhan Ayman Mohamed (Sec. 32)

Haneen Ibraheem Emam (Sec. 9)

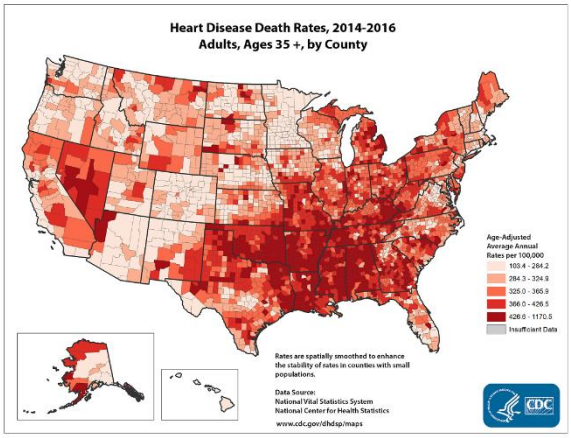
"All team members are level 2 students"

Introduction:

This project aims to study the likelihood of having heart diseases,

Which is a quite interesting subject to study where:

- Heart disease is the leading cause of death in the United States and globally.
- One person dies every 36 seconds in the United States from cardiovascular disease.
- About 659,000 people in the United States die from heart disease each year, that's 1 in every 4 deaths.



Source: www.cdc.gov/heartdisease

So, to help People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidemia 'excessive fats' or already established disease) who need early detection and management.

We decided to analyze this subject statistically using graphical and numerical methods and train a machine learning model that can make good use of the dataset.

Summary of research:

Dataset:

Our dataset is based on a combination of 5 datasets with 11 common features, it contains 918 observations after removing duplicates

Used datasets are:

- Cleveland: 303 observations
- Hungarian: 294 observations
- Switzerland: 123 observations
- Long Beach VA: 200 observations
- Stalog (Heart) Data Set: 270 observations

Dataset source: fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved [December 2021] from https://www.kaggle.com/fedesoriano/heart-failure-prediction.

Dataset details: the dataset comes in form of a ‘‘.CSV’’ file, it will be analyzed using “Pandas” and “Matplotlib” libraries on python 3.9 using “Jupyter Notebooks” with “Conda” environment and it contains 12 attributes divided into columns which are:

- Age: age of the patient [years]
- Sex: sex of the patient [M: Male, F: Female]
- ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
- RestingBP: resting blood pressure [mm Hg]
- Cholesterol: serum cholesterol [mm/dl]
- FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
- RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
- MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
- ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
- Oldpeak: oldpeak = ST [Numeric value measured in depression]
- ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
- HeartDisease: output class [1: heart disease, 0: Normal]

LEVEL 1:

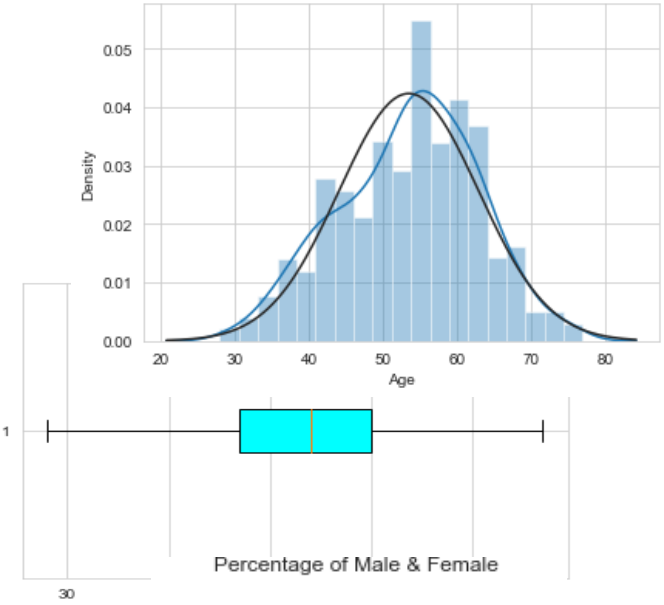
Exploratory data Analysis:

By studying and analyzing this data, we could deduce the following:

	RestingBP	Cholesterol	MaxHR	Oldpeak	Age
count	918.000000	918.000000	918.000000	918.000000	918.000000
mean	132.396514	244.628758	136.809368	0.887364	53.510893
std	18.514154	53.318031	25.460334	1.066570	9.432617
min	0.000000	85.000000	60.000000	-2.600000	28.000000
25%	120.000000	214.000000	120.000000	0.000000	47.000000
50%	130.000000	244.600000	138.000000	0.600000	54.000000
75%	140.000000	267.000000	156.000000	1.500000	60.000000
max	200.000000	603.000000	202.000000	6.200000	77.000000

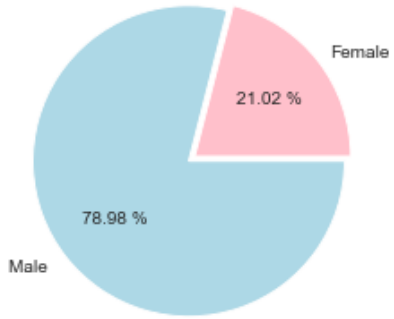
Age:

- Mean of Ages: 53.510893246187365
- Median of Ages: 54.0
- Mode of Ages: 54 repeated 51 times
- Variance of Ages: 88.9742
- Standard deviation of Ages: 9.4326
- Q1= 47.0, Q2 = 54.0, Q3 = 60.0
- Interquartile range (IQR) =1 3.0
- Range: minimum: 27.5, maximum 79.5
- The distribution is normal and slightly skewed to the left



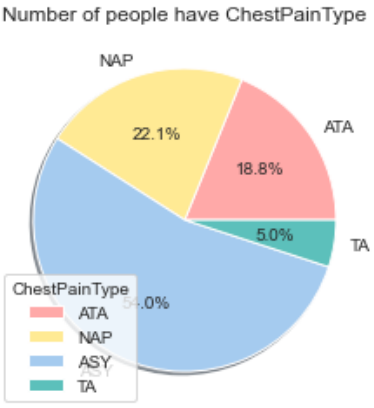
Sex:

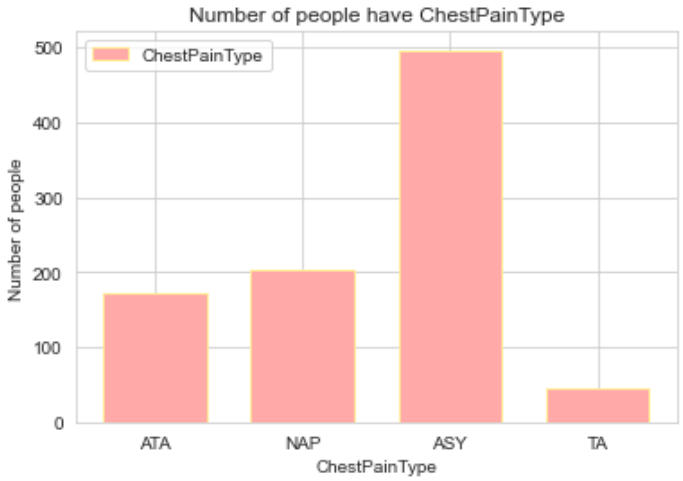
- Males are: 78.98% of the sample, while
- Females are: 21.01% of it.



Chest pain type:

- 54% of observations had Asymptomatic pain (ASY)
- 22.1% had Non-Anginal Pain (NAP)
- 18.8% had Atypical Angina (ATA)
- 5% Typical Angina (TA)



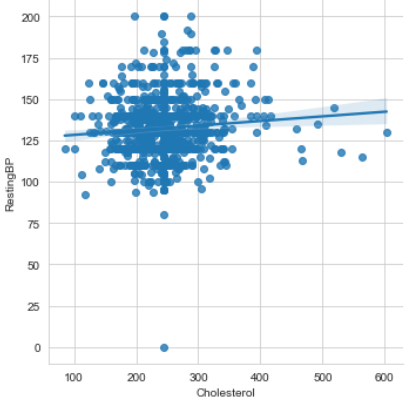


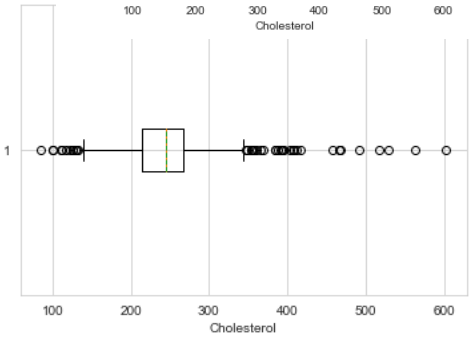
RestingBP:

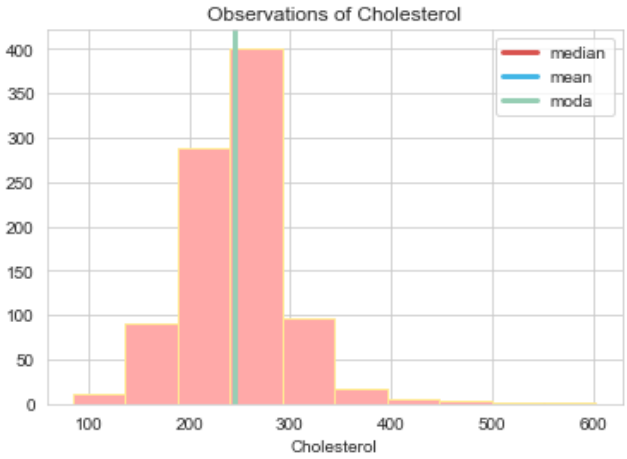
- Mean of RestingBP: 132.39651416122004
- Median of RestingBP: 130.0
- Mode of RestingBP: 120 repeated 132 times
- Variance of RestingBP: 342.7739
- Standard deviation of RestingBP: 18.5141

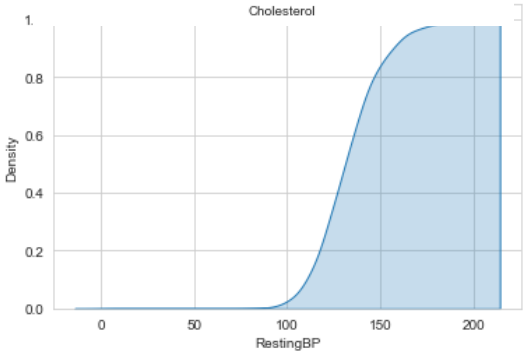
Cholesterol:

- Mean of Cholesterol is 244.62875816993574
- Median of Cholesterol is 244.6
- Mode of Cholesterol is 244.6
- Distribution of shape is NORMAL DISTRIBUTION
- Variance of Cholesterol is 2842.8124
- Standard deviation of Cholesterol is 53.3180
- Range of data:
- Minimum Value = 85.0
- Maximum Value = 603.0
- IQR of Cholesterol is 53.0



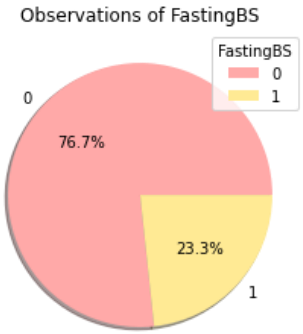


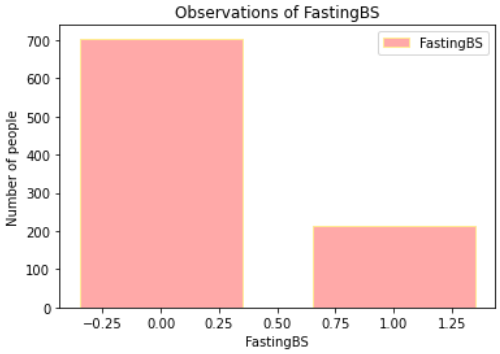




FastingBs:

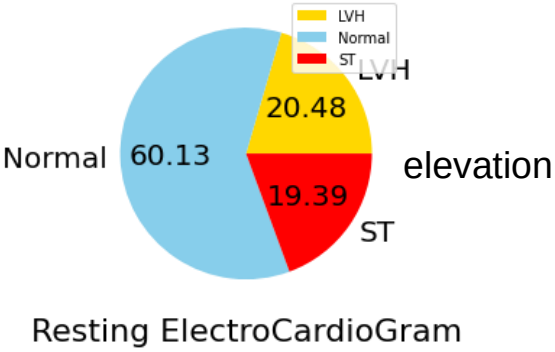
- 76.7% of observations had fasting blood sugar greater than 120 mg/dl
- 23.3% of observations had fasting blood sugar less than 120 mg/dl





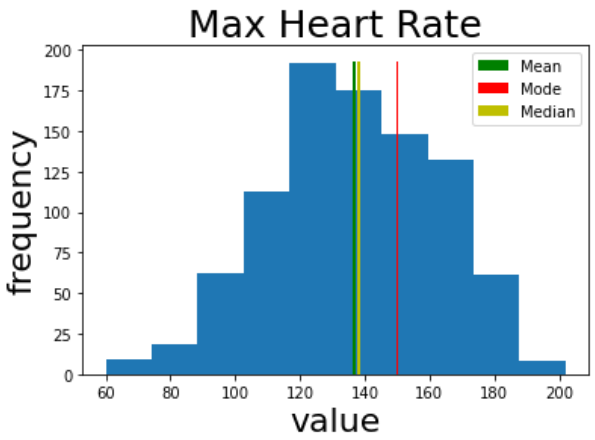
RestingECG:

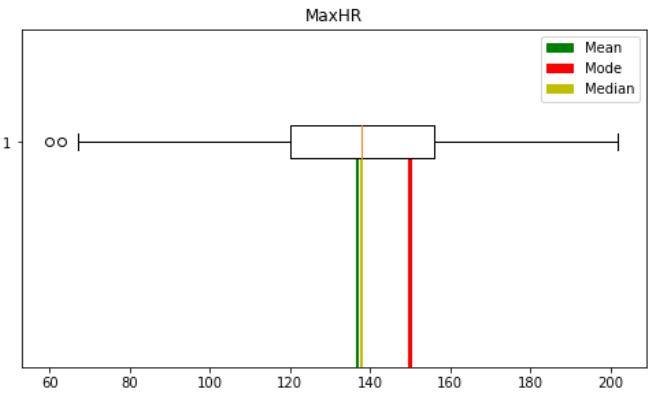
- 60.13% of observation are **Normal**
- 19.39% of observation having **ST**-T wave abnormality (T wave inversions and/or ST or depression of > 0.05 mV)
- 20.48% **LVH**: showing probable or definite left ventricular hypertrophy by Estes' criteria



MaxHR:

- Mean: 136.8093
- Median: 138
- Mode: 150
- Standard deviation: 25.4603
- Variance: 648.2268
- Q1: 120
- Q2: 138
- Q3: 156
- IQR: 36
- Minimum: 60
- Maximum: 202
- Range: 142

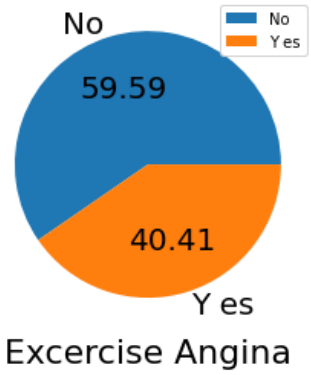




Distribution is normal, Skewness: -0.1443
Slightly skewed to the left

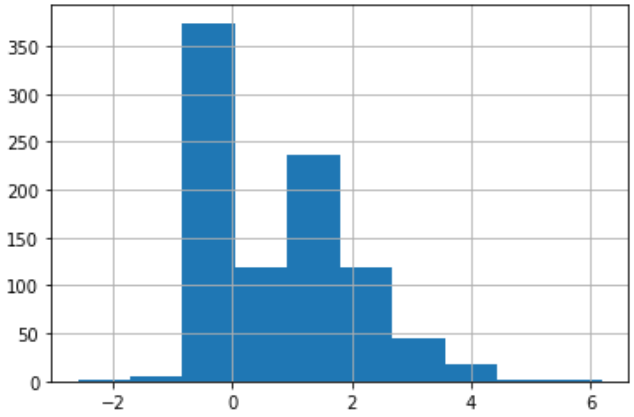
ExerciseAngina:

- 59.59% of observation didn't have Exercise Angina
- 40.41% of observation had Exercise Angina



Oldpeak:

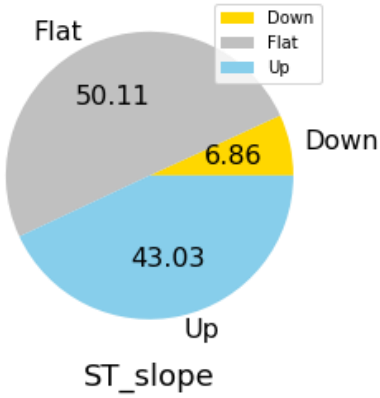
- Mean of Oldpeak is 0.8873
- Median of Oldpeak is 0.6
- Mode of Oldpeak is [0.0]
- distribution shape is **positively skewed**
- Variance of Oldpeak is 1.1375
- Standard deviation of Oldpeak is 1.0665
- Minimum Value = -2.6
- Maximum Value = 6.2
- IQR of Oldpeak is 1.5



ST_Slope:

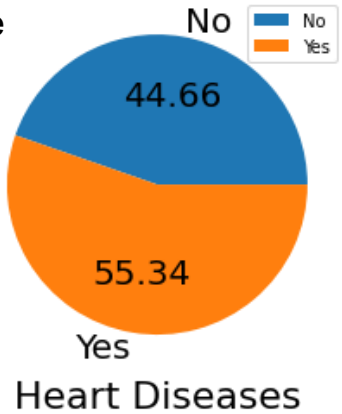
the slope of the peak exercise ST segment

- 50.11% have **FLAT** slope
- 43.03% have an **UP** slope
- 6.86% have a **DOWN** slope



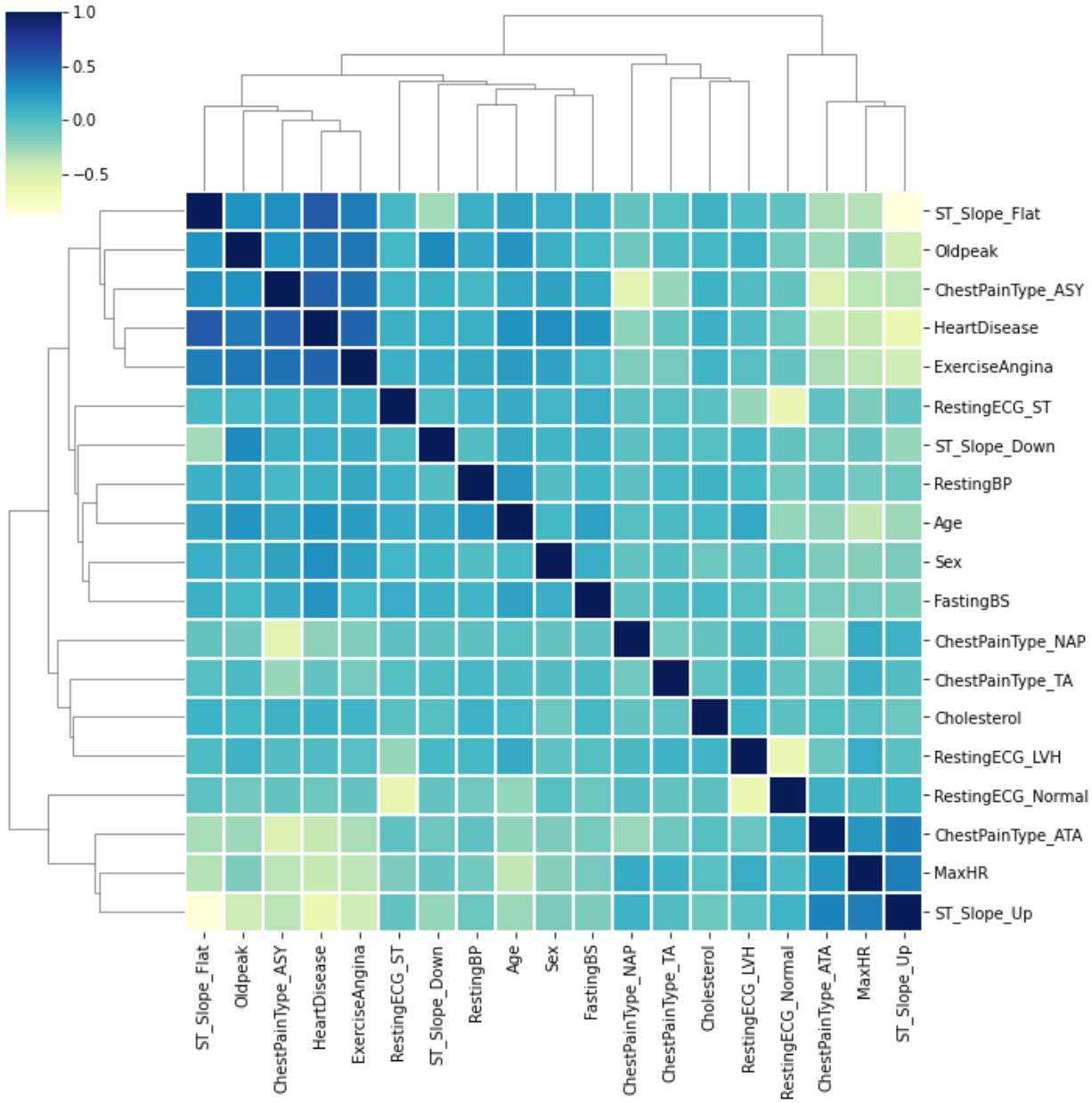
HeartDisease:

- 44.66% of the observations didn't have heart disease
- 55.34% of the observations had heart disease



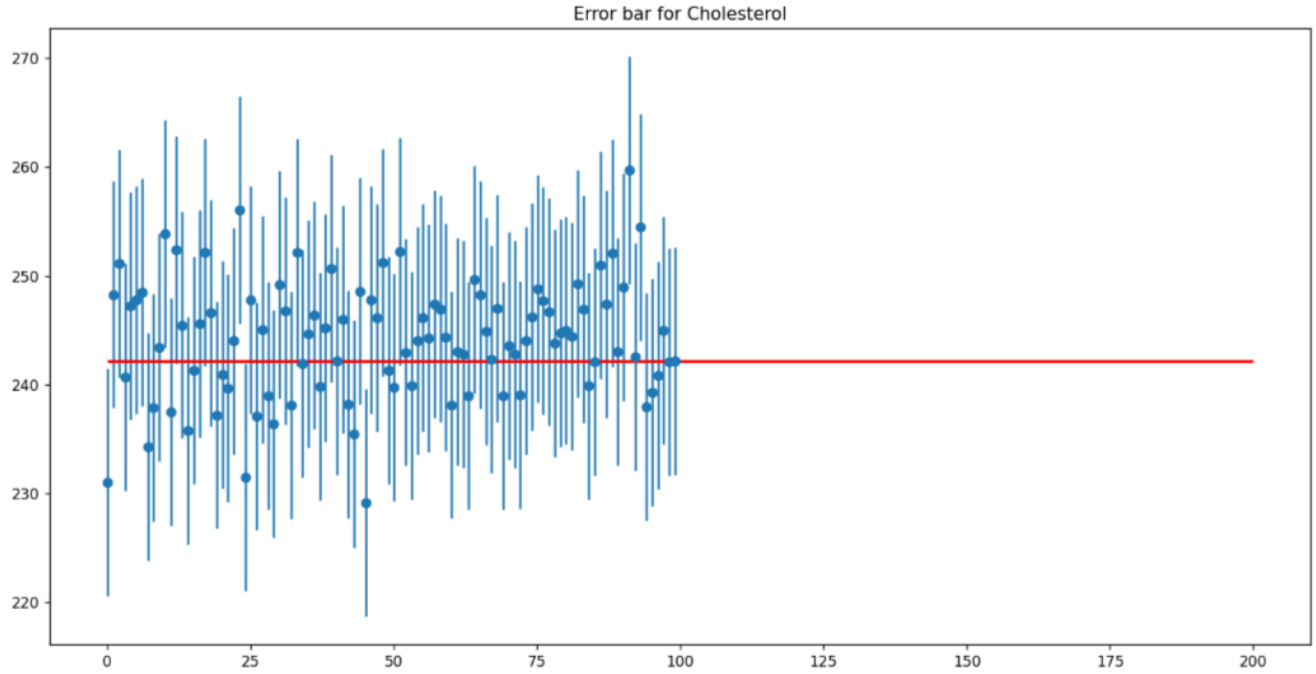
Correlation of 2D data:

	HeartDisease	Sex	RestingBP	Cholesterol	FastingBS	MaxHR	ExerciseAngina	Oldpeak	Age	ChestPainType_ASY	ChestPainType_ATA	ChestPainType_NAP	ChestPainType_TA	RestingECG_LVH	RestingECG_Normal	RestingECG_ST	ST_Slope_Down	ST_Slope_Flat	ST_Slope_Up
HeartDisease	1.000000	0.305445	0.107589	0.093989	0.267291	-0.400421	0.494282	0.403951	0.282039	0.516716	-0.401924	-0.212964	-0.054790	0.010670	-0.091580	0.102527	0.122527	0.554134	-0.622164
Sex	0.305445	1.000000	0.005133	-0.101750	0.120076	-0.189186	0.190664	0.105734	0.055750	0.183876	-0.161522	-0.066486	-0.004031	-0.049518	-0.010634	0.063715	0.066036	0.116077	-0.150942
RestingBP	0.107589	0.005133	1.000000	0.080742	0.070193	-0.112135	0.155101	0.164803	0.254399	0.048824	-0.046153	-0.041348	0.049855	0.053166	-0.116851	0.090447	-0.007912	0.099207	-0.096146
Cholesterol	0.093989	-0.101750	0.080742	1.000000	0.042925	-0.017166	0.077528	0.053036	0.053337	0.084485	-0.015241	-0.062222	-0.047310	0.075526	-0.042412	-0.024566	-0.008944	0.093598	-0.089954
FastingBS	0.267291	0.120076	0.070193	0.042925	1.000000	-0.131438	0.060451	0.052698	0.198039	0.131176	-0.140514	-0.039249	0.026885	-0.011656	-0.093028	0.127110	0.105102	0.107006	-0.161730
MaxHR	-0.400421	-0.189186	-0.112135	-0.017166	-0.131438	1.000000	-0.370425	-0.160691	-0.382045	-0.354963	0.253735	0.134580	0.100025	0.125793	0.023801	-0.157879	-0.073316	-0.342581	0.383397
ExerciseAngina	0.494282	0.190664	0.155101	0.077528	0.060451	-0.370425	1.000000	0.408752	0.215793	0.430034	-0.300365	-0.166030	-0.128105	-0.016382	-0.072924	0.107036	0.136439	0.382237	-0.455676
Oldpeak	0.403951	0.105734	0.164803	0.053036	0.052698	-0.160691	0.408752	1.000000	0.258612	0.280026	-0.262124	-0.106212	0.032231	0.086794	-0.116719	0.059958	0.322130	0.283295	-0.450577
Age	0.282039	0.055750	0.254399	0.053337	0.198039	-0.382045	0.215793	0.258612	1.000000	0.166607	-0.218165	-0.011335	0.032042	0.145727	-0.230566	0.136798	0.138397	0.185568	-0.258067
ChestPainType_ASY	0.516716	0.183876	0.048824	0.084485	0.131176	-0.354963	0.430034	0.280026	0.166607	1.000000	-0.522432	-0.577670	-0.249003	0.002289	-0.063606	0.076438	0.103407	0.303645	-0.359443
ChestPainType_ATA	-0.401924	-0.161522	-0.046153	-0.015241	-0.140514	0.253735	-0.300365	-0.262124	-0.218165	-0.522432	1.000000	-0.256767	-0.110679	-0.085791	0.107941	-0.046111	-0.097754	-0.304667	0.357588
ChestPainType_NAP	-0.212964	-0.066486	-0.041348	-0.062222	-0.039249	0.134580	-0.166030	-0.106212	-0.011335	-0.577670	-0.256767	1.000000	-0.122381	0.035299	0.005010	-0.042236	-0.040816	-0.072031	0.093583
ChestPainType_TA	-0.054790	-0.004031	0.049855	-0.047310	0.026885	0.100025	-0.128105	0.032231	0.032042	-0.249003	-0.110679	-0.122381	1.000000	0.081407	-0.057719	-0.011611	0.016651	-0.010486	0.002087
RestingECG_LVH	0.010670	-0.049518	0.053166	0.075526	-0.011656	0.125793	-0.016382	0.086794	0.145727	0.002289	-0.085791	0.035299	0.081407	1.000000	-0.623227	-0.248892	0.043755	0.015091	-0.037582
RestingECG_Normal	-0.091580	-0.010634	-0.116851	-0.042412	-0.093028	0.023801	-0.072924	-0.116719	-0.230566	-0.063606	0.107941	0.005010	-0.057719	-0.623227	1.000000	-0.602314	-0.060564	-0.047172	0.078563
RestingECG_ST	0.102527	0.063715	0.090447	-0.024566	0.127110	-0.157879	0.107036	0.059958	0.136798	0.076438	-0.046111	-0.042236	-0.011611	-0.248892	-0.602314	1.000000	0.030345	0.043017	-0.058936
ST_Slope_Down	0.122527	0.066036	-0.007912	-0.008944	0.105102	-0.073316	0.136439	0.322130	0.138397	0.103407	-0.097754	-0.040816	0.016651	0.043755	-0.060564	0.030345	1.000000	-0.272040	-0.235904
ST_Slope_Flat	0.554134	0.116077	0.099207	0.093598	0.107006	-0.342581	0.382237	0.283295	0.185568	0.303645	-0.304667	-0.072031	-0.010486	0.015091	-0.047172	0.043017	-0.272040	1.000000	-0.870951
ST_Slope_Up	-0.622164	-0.150942	-0.096146	-0.089954	-0.161730	0.383397	-0.455676	-0.450577	-0.258067	-0.359443	0.357588	0.093583	0.002087	-0.037582	0.078563	-0.058936	-0.235904	-0.870951	1.000000

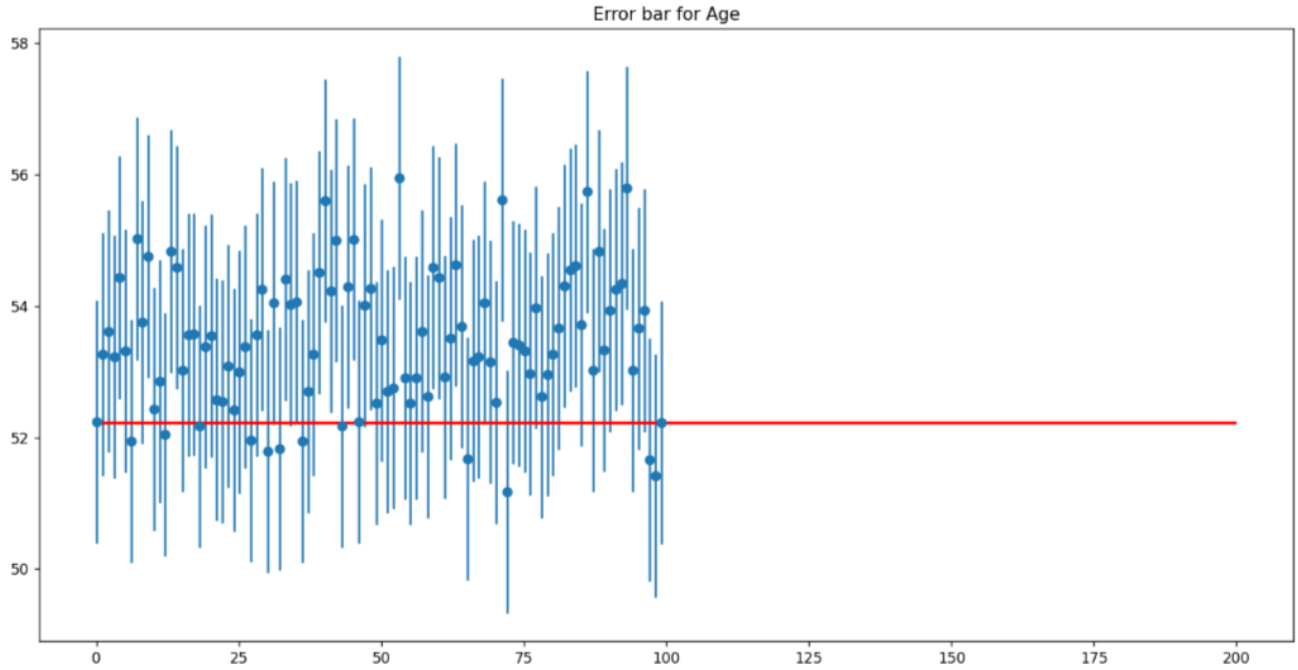


LEVEL 2:

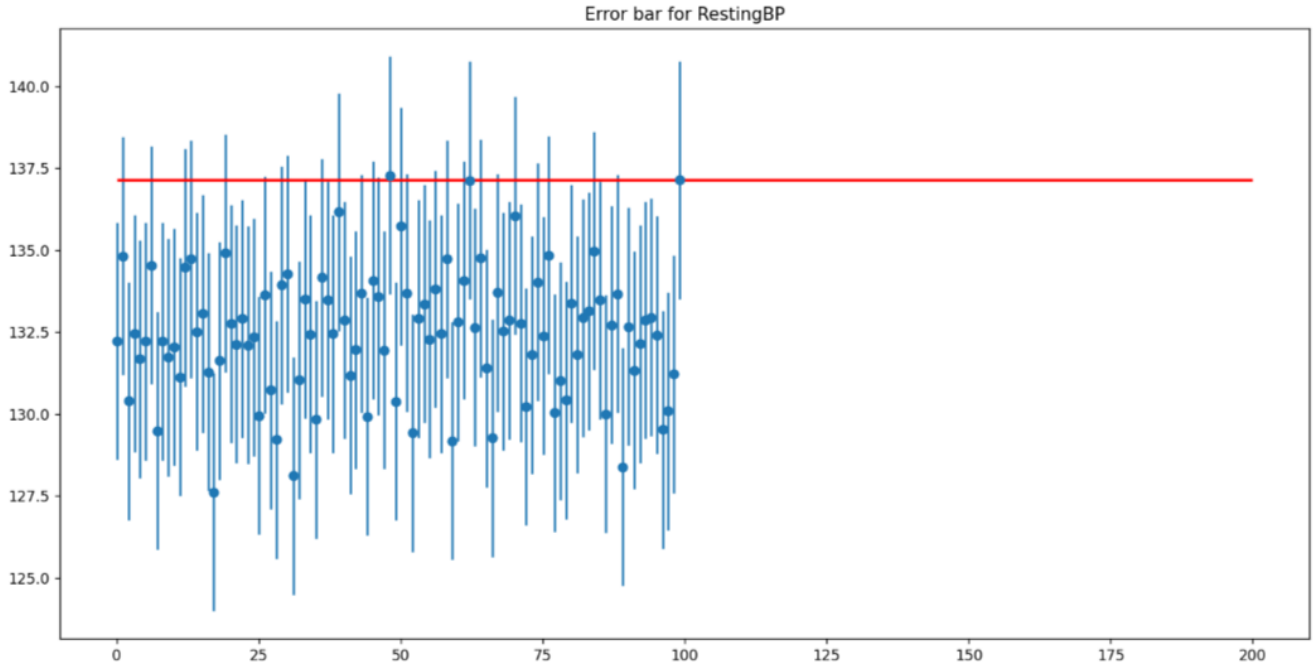
- The confidence interval using critical-z for Cholesterol = [234.9915, 255.8804]
- The confident interval using critical-t for Cholesterol = [235.6967, 256.9592]



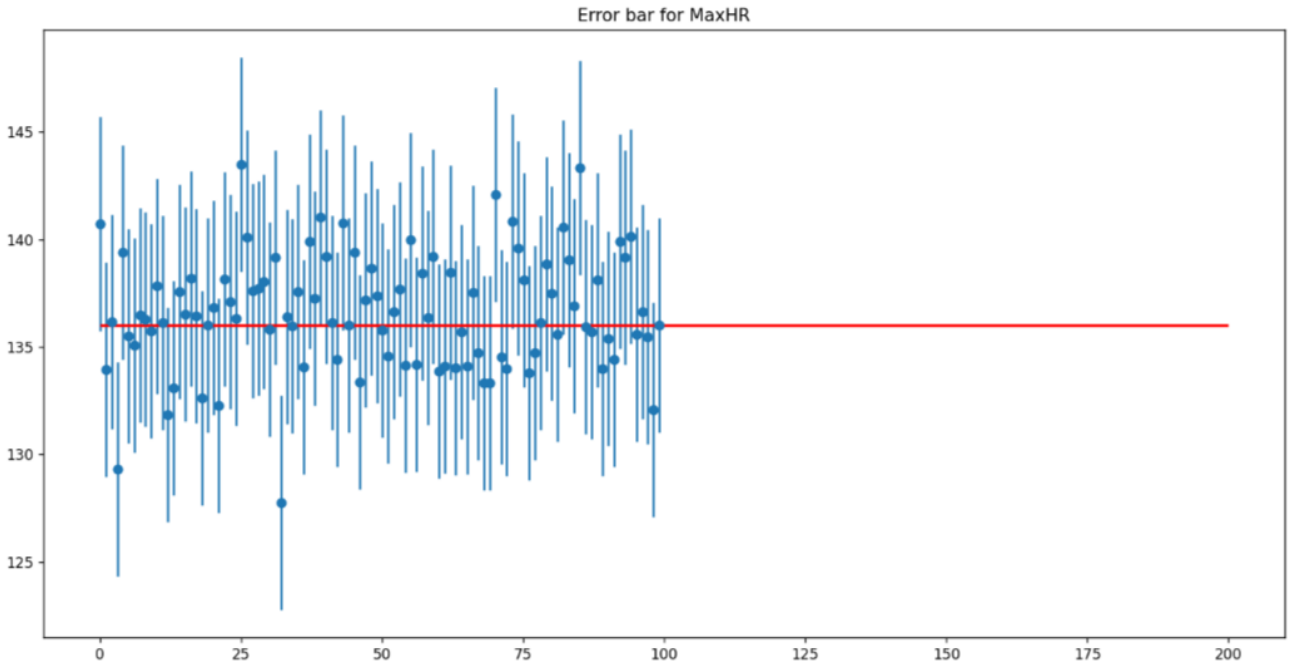
- The confidence interval using critical-z for Age = [50.3622, 54.0577]
- The confident interval using critical-t for Age = (51.6281, 55.4918)



- The confidence interval using critical-z for RestingBP = [131.6432, 138.8967]
- The confident interval using critical-t for RestingBP = (128.7058, 135.3941)



- The confidence interval using critical-z for MaxHR = [129.6925, 139.6674]
- The confident interval using critical-t for MaxHR = (127.5521, 137.0278)



- Haven't fasting blood sugar proportion estimate = 79.0%
- Haven't fasting blood sugar proportion estimate = 21.0%
- Haven't heart disease proportion estimate = 47.0%
- Haven't heart disease proportion estimate = 53.0%

- don't make exercise angina proportion estimate = 59.0%
- don't make exercise angina proportion estimate = 41.0%

Machine Learning approach:

- To make good use of our dataset we decided to train 3 different machine learning models such that each model depends on different algorithms than the others.
- Before being able to train the models, dataset had to be prepared first by removing duplicates, null or missing values, outliers and using dummies for categorial data (like chestPainType)
- All models were trained using a part of the data set and tested using the other part
    - The first model was trained using multiple linear regression, it had poor accuracy of only 55%
    - Second model was trained using decision tree, and it was a great leap forward as the accuracy has improved to about 80% to 85%, depending on the random portion of dataset used during the training.
    - Third model was trained using K-Nearest-Neighbors (KNN) algorithm and had accuracy of 85%
- “scikit-learn” library from “Conda” environment with python 3.9 was used to train the models and “joblib” to save them as binaries for later use
- A Jupyter notebook was used to divide code into cells to be easy for reading and documenting along the whole study
- To test for an observation, the required attributes need to be passed in form of (.CSV) file to the model
- Each model produces output of:
  either (1: “have heart disease”) or (0: “Doesn’t have heart disease”)