Heart Matters: exploring the science behind cardiovascular health

By: Shreya Jayakumar

Outline

- Introduction
- Research Questions
- Methodology (EDA & Hypothesis Tests)
- Results
- Conclusion

Introduction

- Heart disease is an umbrella term for various cardiovascular diseases
- Many factors influence occurrence of heart disease
- We analyzed a heart disease dataset with 303 rows, 14 columns to answer
- 9 Categorical and 5 Numerical Variables

	age	sex	chest_pain	resting_bp	cholesterol	fasting_blood_sugar	rest_ecg	thalachh	exng	oldpeak	slope	num_major_vessels	thalassem
0	63	1	3	145	233	1	0	150	0	2.3	0	0	
1	37	1	2	130	250	0	1	187	0	3.5	0	0	
2	41	0	1	130	204	0	0	172	0	1.4	2	0	
3	56	1	1	120	236	0	1	178	0	0.8	2	0	
4	57	0	0	120	354	0	1	163	1	0.6	2	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	

Glimpse of our dataset

What are we trying to solve?

We are trying to see whether the following variables in the data set influence the likelihood of having heart disease

Variables

- Age
- Sex
- Chest_pain
- resting_bp
- cholesterol
- fasting_blood_sugar
- rest_ecg
- Thalach
- exang
- oldpeak
- slope
- Thalassemia
- heart_disease

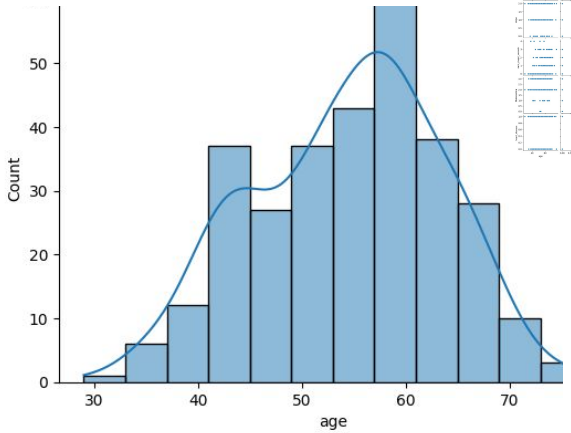
Exploratory Data Analysis

- EDA performed to generate hypothesis
- Looked at null values, summary statistics, unique values
- Univariate and bivariate distributions using matplotlib and seaborn

Pairplot

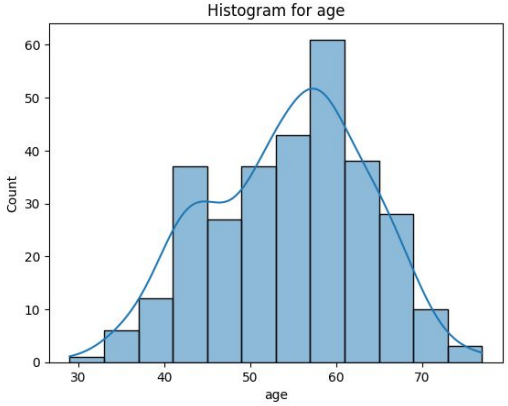
	age	resting_bp	cholesterol	thalachh	oldpeak
count	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	131.623762	246.264026	149.646865	1.039604
std	9.082101	17.538143	51.830751	22.905161	1.161075
min	29.000000	94.000000	126.000000	71.000000	0.000000
25%	47.500000	120.000000	211.000000	133.500000	0.000000
50%	55.000000	130.000000	240.000000	153.000000	0.800000
75%	61.000000	140.000000	274.500000	166.000000	1.600000
max	77.000000	200.000000	564.000000	202.000000	6.200000

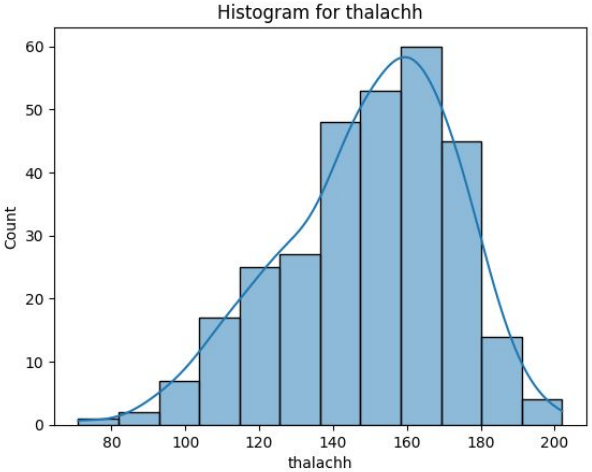
Summary Stats of Num. Variables

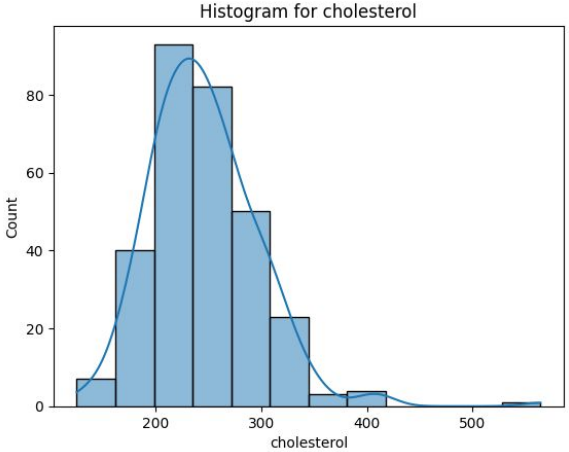


Histogram of Age

Histograms

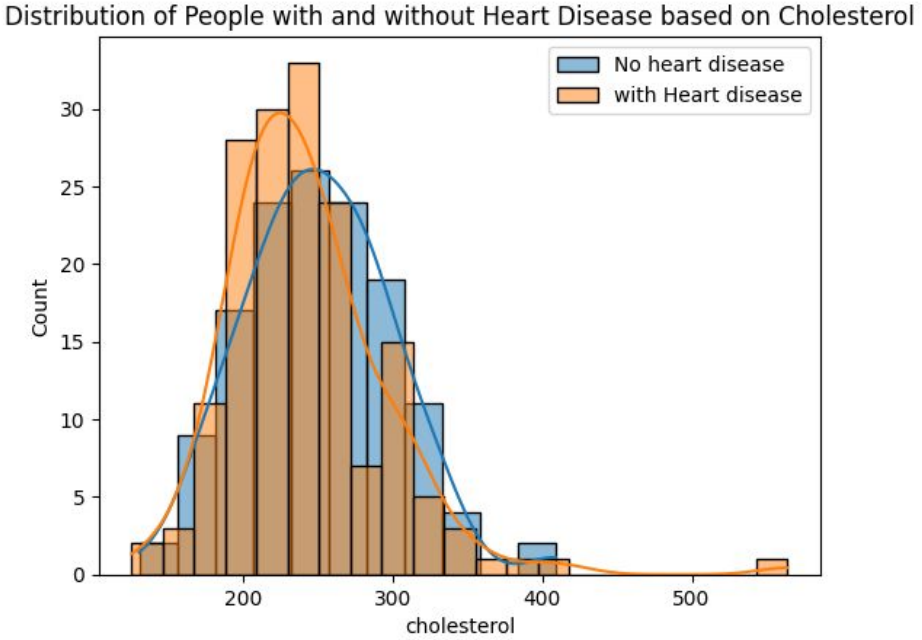

Histogram for age
Count
age
0
10
20
30
40
50
60
30
40
50
60
70


Histogram for thalachh
Count
thalachh
0
10
20
30
40
50
60
80
100
120
140
160
180
200


Histogram for cholesterol
Count
cholesterol
0
20
40
60
80
200
300
400
500

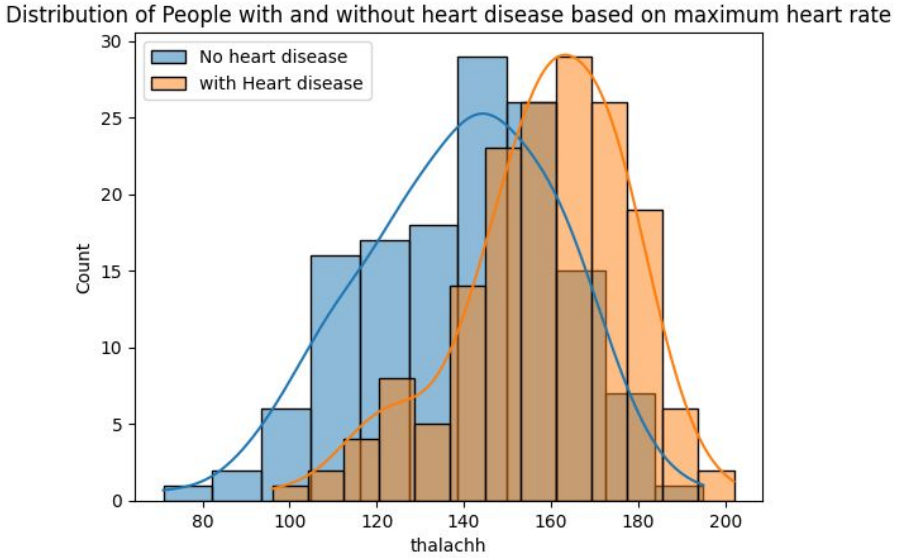
Hypothesis 1: Is there a relationship between the average Cholesterol level and heart disease?

- Used **One-Way ANOVA**
- Unequal sample size
- Used scipy, defined functions
- F-statistic = 2.2029, p-value = 0.13879
- Fail to reject the null hypothesis.
- Insufficient evidence to suggest that there is a significant difference between the average cholesterol levels of people with and without heart disease.



Hypothesis 2: Is there a relationship between maximum heart rate and heart disease?

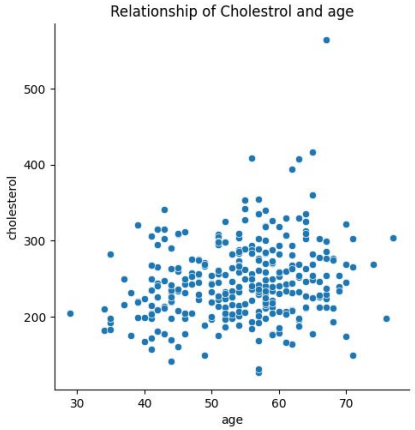
- Used **Welch Independent T Test**
- Data Visualization- graph
- P-Value: 5e-14
- T-Statistic: -7.95
- Since p-value is very small and < 0.05, we can reject the null hypothesis of Welch's t-test.
- Therefore, we can conclude that there is sufficient evidence to say that that having or not having heart disease lead to different mean maximum heart rate



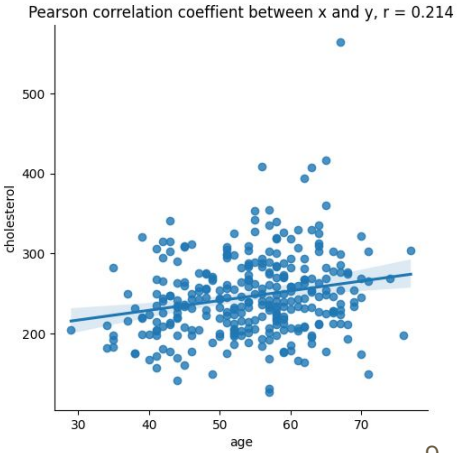
From the graph we can see that data is approximately normally distributed and that each subject belongs to one group. This means that a person can or can't have heart disease based on their maximum heart rate

Hypothesis 3: Are age and cholesterol level related?

- Used **Pearson Correlation**
- Data Visualization- scatterplot and correlation graph

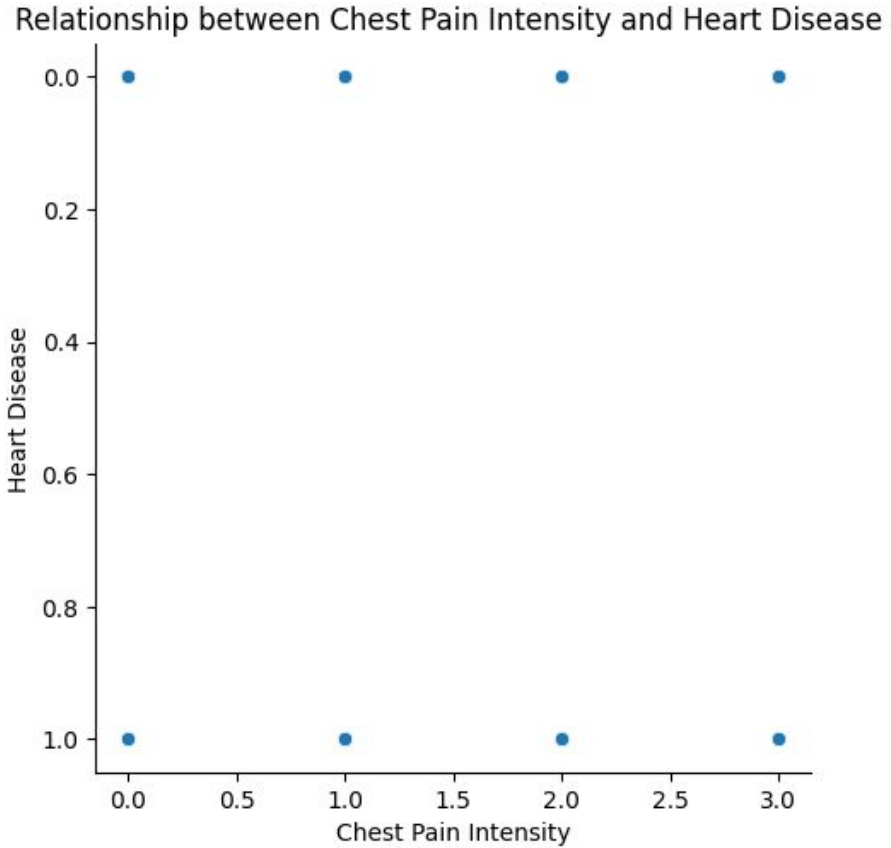


- Pearson Coefficient = 0.214 ,P-value = 0.0001786
- Weakly pos relationship
- p value 0.0001786 < a = 0.05
- pearson coefficient is statistically significant gives us strong evidence to reject the null



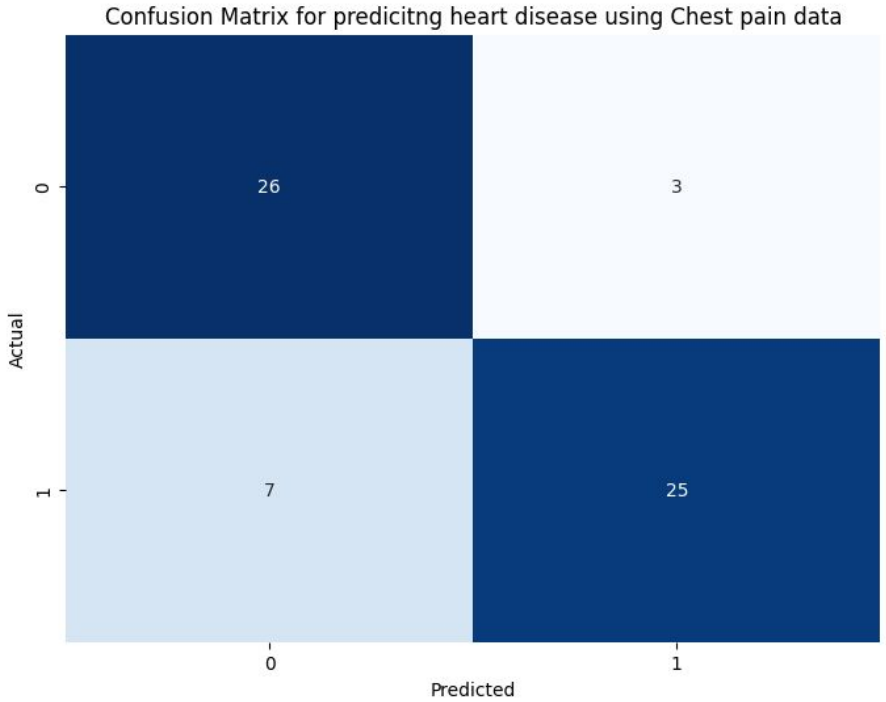
Data Visualization

You can see no useful information about the relationship between the two variables is observable



Hypothesis 4: Can we predict presence or absence of heart disease based on Chest Pain?

- Cannot perform Correlation
- Chest Pain: 0=typical angina, 1=atypical angina, 2=non-anginal pain, 3= asymptomatic
- Used **Logistic Regression**
- Used sklearn to train, test. Sklearn.metrics accuracy_score, statsmodels.api
- Chest pain coeff =0.984
  P-value = 1.45e-12
  Accuracy of model= 83.6%
- Chest pain is a significant predictor of heart disease in your logistic regression model.



Result Summary

Sr.No.	Hypothesis Question	Statistical Test	Result
1.	Is there a relationship between average cholesterol level and the presence/absence of heart disease?	One-Way ANOVA	F-statistic = 2.20298 , P-value = 0.13879 **no significant difference in the mean cholesterol levels between individuals with heart disease and those without**
2.	Is there a relationship between maximum heart rate and heart disease?	Welch Independent T Test	T-Statistic: -7.95 ,P-Value: 5e-14 **No significant difference between the average cholesterol levels and having or not having HD**
3.	Are age and cholesterol level related?	Pearson Correlation	Pearson Coefficient = 0.214 ,P-value = 0.0001786 **Weakly positive relationship**
4.	Is there a relationship between different kinds of chest pain and the likelihood of having heart disease?	Logistic Regression	Chest pain coeff =0.984 ,P-value = 1.45e-12 Accuracy of model= 83.6% **chest pain is a significant predictor of heart disease**

Conclusions

Insights and/or decisions you draw from the data analyses

From all the data we analyze through our hypotheses, we can conclude.

- each variable is independent and of no correlation to one another

When looking at the results from each of our tests, even though some had very little correlation, the majority of the data set is completely independent.

For example, when looking at Hypothesis 1, the age did not really affect whether a person has high cholesterol or not, as the data was scattered all over the place. As a whole, our data analysis shows a stable pattern of independence across the variables, with only weak correlations seen in certain tests.

Future directions and/or areas that can be improved

- Using a bigger dataset on a larger scare
- Using different hypothesis tests to with different CI for better statistical analysis
- Validate our observations for predictions to improve patient outcomes

Thank You