Data-Driven Approaches to Understand the Factors Causing Heart Disease

Helena Hu, Estelle Hu, David Tolentino.

Author contributions

Helena Hu contributed in the discription of the dataset, the summary, results, and findings of visualization (barchart), and format the final report.

Estelle Hu contributed in abstract and background (introducing the topic and cocluding the indicators for heart disease).

David Tolentino contributed in the aims of the project, description of the variables, and the summary, results, and findings of PCA.

All three authors contributed in the discussion part.

Abstract

Thousands of patients with diverse attributes such as varying ages, genders, heights, serum cholesterol levels, and other characteristics have undergone testing to determine the presence or absence of heart disease. By using the dataset, we plotted a lot of graphs to show the direct correlation and eventually find out if there's a link between personal characteristics and heart disease risk. We first provided a summary of the basic properties of the dataset and found that there is no existing missing value. Then we change the column name to elaborate the column more precisely, such as changing 'chol' to 'cholesterol(mg/dl)'. Moreover, we defined the numerical code to its corresponding value. For example, in the sex code, we change 0 to male and 1 to female, to represent the data more visually and directly. Furthermore, we did several bar charts to separately show the relationship between different characteristics of patients and the count of records of getting heart disease. From the plot of age vs count of records, we see that people who are between 40 and 55 years old are more likely to get heart diseases based on the sample. From the plot of sex vs count of records, we see that the percentage of females who have heart disease is much larger than that of males who have heart disease. From the plot of chest pain type vs count of records, we see that people with atypical angina, non-anginal, and asymptomatic chest pain have the greatest risk of developing heart disease, while people who have heart disease have typical angina chest pain get a relatively smaller chance of getting heart disease. In the plot of resting blood pressure vs count of records, we observed that there's no significant relation between resting blooding pressure and the chance of developing heart disease. In the plot of cholesterol vs count of record, we found that people having heart disease mostly have serum cholesterol between 200 mg/dl to 300 mg/dl. With the graph of sugar in blood vs counts of records, we found that there isn't much relations between fasting blood sugar and the presence of heart disease. With the graph of electrical activity in the heart vs count of records, we can see that compared to other resting ECG types, people having ST-T wave abnormality are more like to have heart disease present. We see from the max heart rate vs counts of records, people with higher maximum heart rates tend to have a much higher possibility to develop heart diseases. In the plot of ST depression induced by exercise vs a number of records, we can see that people with st_slope of 2 obviously have a higher chance of developing heart disease. From the plot of a number of vessels vs count of records, we can see that people with the least major vessels colored by fluoroscopy have a significant chance of getting heart disease. From the plot of blood disorder thalassemia vs a number of records, we can tell that the fixed effect form of thalassemia has most significant influence on heart disease, and the reversible effect form of thalassemia has the least influence on heart disease. Overall, we can conclude that the factors age, sex, chest pain type, cholesterol, electrical activity in the heart, max heart rate, ST depression, number of vessels, and blood disorder thalassemia do contribute to the presence of heart disease risk. However, resting blood pressure, and sugar in the blood do not impact whether patients have the risk of getting heart disease. We also have to admit that there are some flaws in the conclusion as the sample is not randomly selected as they are just patients from selected clinics.

Introduction

Background

Heart disease is responsible for the majority of deaths in the United States. It is an umbrella term that encompasses various heart conditions. Coronary artery disease is the most common type of heart disease, which occurs when the arteries that supply blood to the heart become narrowed or blocked. This can lead to chest pain (angina), heart attack, or stroke. Making lifestyle changes and taking medication can significantly decrease the likelihood of developing heart disease. The dataset "heart.csv" was created in 1988 and comprises four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It consists of 76 characteristics, including the predicted attribute, although studies typically utilize only a subset of 14 of them. The "target" field refers to whether the patient has heart disease, with a value of 0 indicating no disease and a value of 1 indicating disease. Thousands of patients of different ages, sexes, treetops, serum cholesterol, and other attributes have been tested to determine whether or not they have heart disease. We're interested in finding out if there's a link between personal characteristics and heart disease risk. After tidying the dataset, making the plots, and drawing conclusions, we can say that factors age, sex, chest pain type, cholesterol, electrical activity in the heart, max heart rate, ST depression, number of vessels, blood disorder thalassemia have an impact on the presence of heart disease risk. However, resting blood pressure, and sugar in the blood do not influence whether patients have the risk of getting heart disease. So we can ask patients with potential risk factors, liken from 40 to 55 years old, or female, or with atypical angina, or with non-anginal, or with asymptomatic chest pain, or whose serum cholesterol between 200 mg/dl to 300 mg/dl, or whose ST-T wave abnormality, or with higher maximum heart rate, or with st_slope of 2, or with least major vessels colored by fluoroscopy, or with thalassemia to change their lifestyles. These patients with potential factors of getting heart diseases can maintain a healthy weight, exercise regularly, eat healthily, manage stress, or monitor blood pressure and cholesterol levels.

Aims

For our project we wanted to investigate how certain characteristics of a patient are associated with the prescence of heart disease. One of the research questions we posed involved identifying what variables in our dataset related to the prescence of heart disease. In order to answer this question we decided to create a correlation matrix, which allows us to examine the strength of relationship between our variables and our target variable of heart disease. This approach will help us to indetify which variables were most strongly correlated to heart disease and allow for further investigation. Other methods we will be going through include Principal Component Analysis (PCA) and visualizations. By using PCA we will be able to determine what variables in our data set are contributing the most to key patterns in our data. We will be plotting the loadings of each variable in order to see how they are contributing to our principal components. By calculating and visualizing our PCAs we will have a better undertstanding of what variables are most important in our data set.

Materials and methods

Datasets

The dimension of the data is (1025, 14). The dataset has 1025 rows(observations) and 14 columns(variables). There's no missing value in the dataset. The descriptions of variables in the dataset are listed as follows:

Variable name	Description	Type	Units of Measurement
age	age of patient	numeric	29-77 years
sex	patient gender	numeric	values from 0:1 (0: female, 1: male)
chest_pain_type	type of chest pain	numeric	values from 0:3 (0: typical angina, 1: atypical angina, 2: non-anginal pain, 3: asymptomatic)
rest_bp(mm Hg)	resting blood pressure	numeric	mm Hg
cholesterol(mg/dl)	amount of total cholestoral in blood	numeric	mg/dl
fbs(mg/dl)	if sugar in blood is greater than 120 mg	numeric	values from 0:1 (0: <= 120, 1: > 120)
rest_ecg	measure of electrical activity in heart	numeric	values from 0:1 (0: normal, 1: having ST-T wave abnormality, 2: showing probable or definite left ventricular hypertrophy by Estes' criteria)
max_heart_rate	max heart rate achieved	numeric	bpm
angina_exercise	exercise induced angina(chest pain)	numeric	values from 0:1 (0: no, 1: yes)
st_depression	ST depression induced by exercise	numeric	'ST' relates to positions on the ECG plot
st_slope	slope of peak exercise ST segment	numeric	values from 0:2
num_vessels	number of major vessels	numeric	values from 0:3 (1: normal, 2: fixed effect, 3:reversable effect)
thal	blood disorder thalassemia	numeric	values from 1:3 (1: normal, 2: fixed effect, 3:reversable effect)
heart_disease	if patient has heart disease	numeric	values from 0:1 (0: absence, 1: presence)

Here is the first 5 rows of the tided dataset:

index	age	sex	chest_pain_type	rest_bp(mm Hg)	cholesterol(mg/dl)	fbs(mg/dl)	rest_ecg	max_heart_rate	angina_exercise	st_depression	st_slope	num_vessels	thal	heart_disease
0	52	male	typical angina	125	212	\<= 120	ST-T wave abnormality	168	no	1.0	2	2	reversable effect	Absence
1	53	male	typical angina	140	203	>120	normal	155	yes	3.1	0	0	reversable effect	Absence
2	70	male	typical angina	145	174	\<= 120	ST-T wave abnormality	125	yes	2.6	0	0	reversable effect	Absence
3	61	male	typical angina	148	203	\<= 120	ST-T wave abnormality	161	no	0.0	2	1	reversable effect	Absence
4	62	female	typical angina	138	294	>120	ST-T wave abnormality	106	no	1.9	1	3	fixed effect	Absence

And here is the summary of numeric variables in the dataset:

index	age	rest_bp(mm Hg)	cholesterol(mg/dl)	max_heart_rate	st_depression	st_slope	num_vessels
count	1025.0	1025.0	1025.0	1025.0	1025.0	1025.0	1025.0
mean	54.43414634146342	131.61170731707318	246.0	149.11414634146342	1.0715121951219515	1.3853658536585365	0.7541463414634146
std	9.072290233244278	17.516718005376408	51.59251020618206	23.005723745977207	1.175053255150176	0.6177552671745918	1.0307976650242823
min	29.0	94.0	126.0	71.0	0.0	0.0	0.0
25%	48.0	120.0	211.0	132.0	0.0	1.0	0.0
50%	56.0	130.0	240.0	152.0	0.8	1.0	0.0
75%	61.0	140.0	275.0	166.0	1.8	2.0	1.0
max	77.0	200.0	564.0	202.0	6.2	2.0	4.0

Methods

For exploratory analysis, the technique we use to explore the relationship between the presence or absence of heart disease of patient and their health conditions is visualization. I plot barcharts for each variable separately and group the data by whether the patients have heart disease or not by contrast color so that we can see which health factor has the most significant influence on developing heart disease. Among all those factors in the dataset, we found that there are 7 factors, including age, chest pain type, max heart rate, angina exercise, st depression, number of major vessels, and form of thalassemia, have noticable effect on having heart disease. There is also certain degree of pattern between sex and heart disease, but total number of female patients is much less than that of male patients, so there might be some bias when applying the relationship from the dataset to general group, so we did not include sex factor when considering the most significant factor.

For the 7 factors listed above, for numeric variables, there's potive relationship between max heart rate and chance of getting heart disease and negative relationship between st depression, number of major vessels and chance of getting heart disease. For categorical variables, we can see that patients with typical angina types of chest pain or having angina exercise or reversable effect of thalassemia have lower chance of developing heart disease.

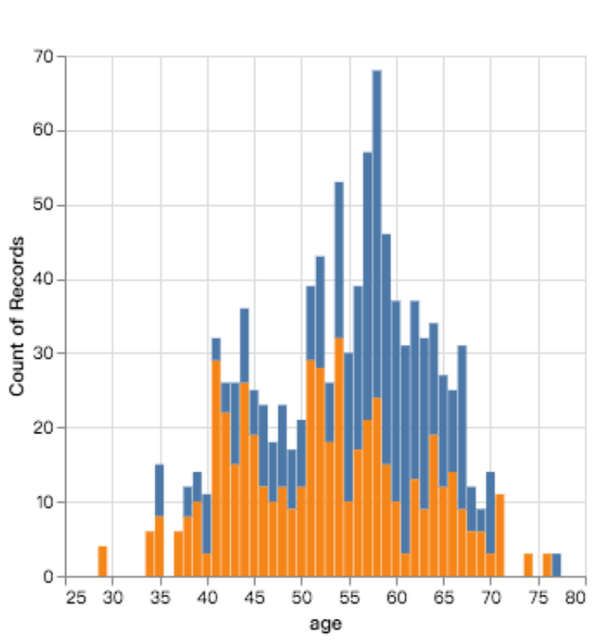
Except barchat, we also used a correlation matrix in order to visualize the strength of the relationship between variables in our data set. All categorical variables were dummy coded and then their correlation to each was then computed. Our target variable of prescence of heart disease was sliced from our correlation matrix in order to investigate which variables had the strongest correlation to heart disease. A PCA plot of varaibles was then created in order to identify which variables contributed the most to any key underlying patterns within our data.

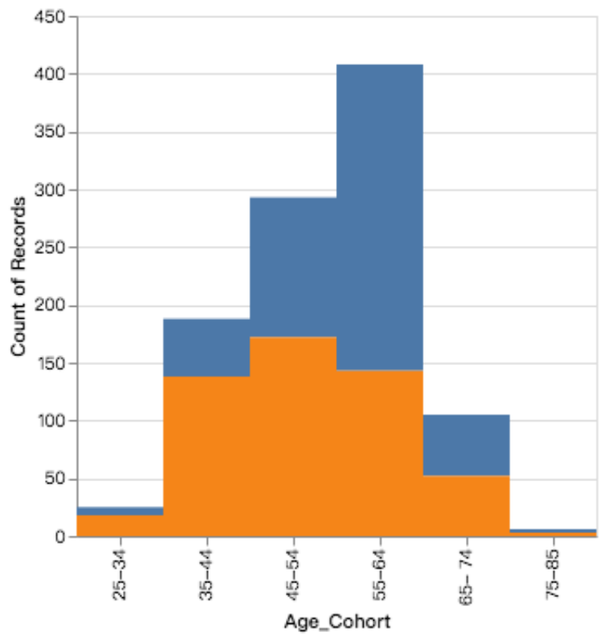
Results

Visualization (Barchart)

The following barchart shows the relationship between the proportions of getting heart disease and each health-related variables.

Age & Age Cohort

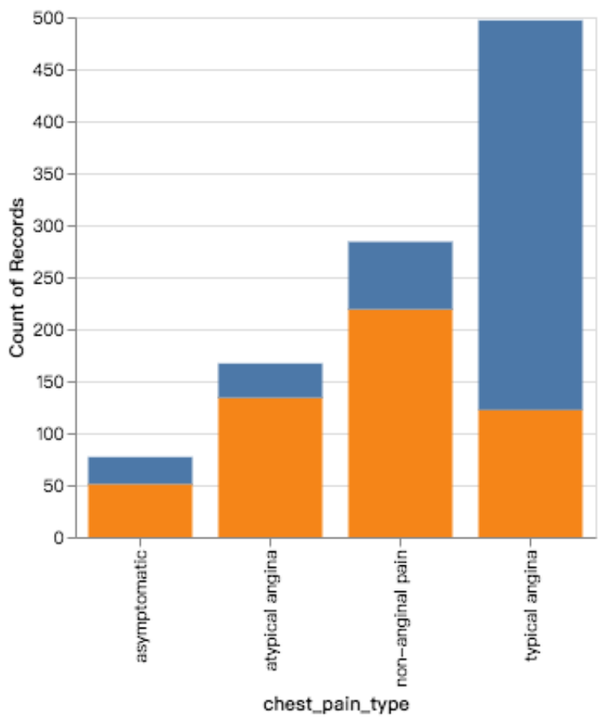


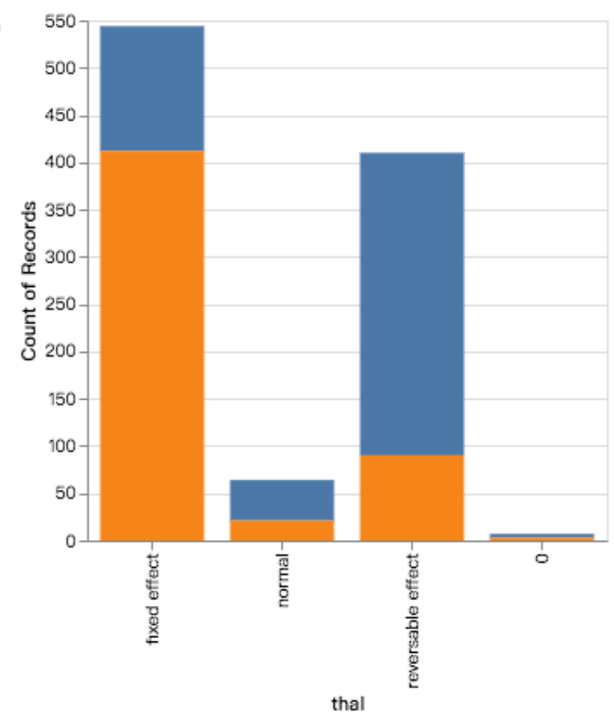


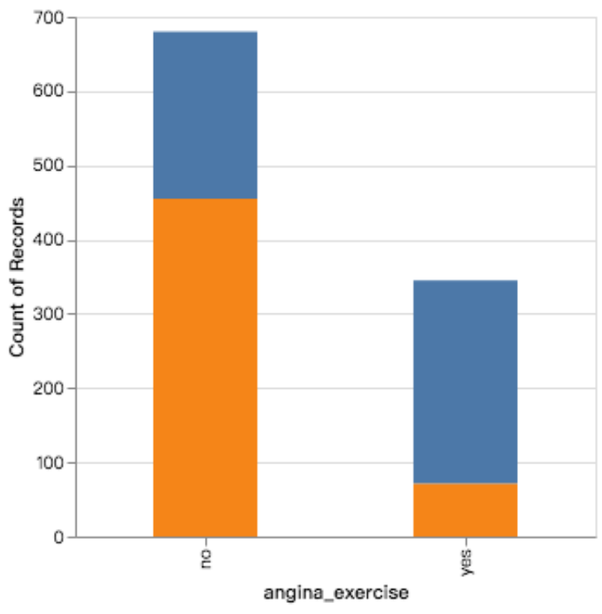
From the age plot we can see people who are 40 years old to 55 years old are more likely to have heart disease. It is kind of surprising that younger people have higher risk of developing heart disease. That's why we plan to focus on the age in later steps.

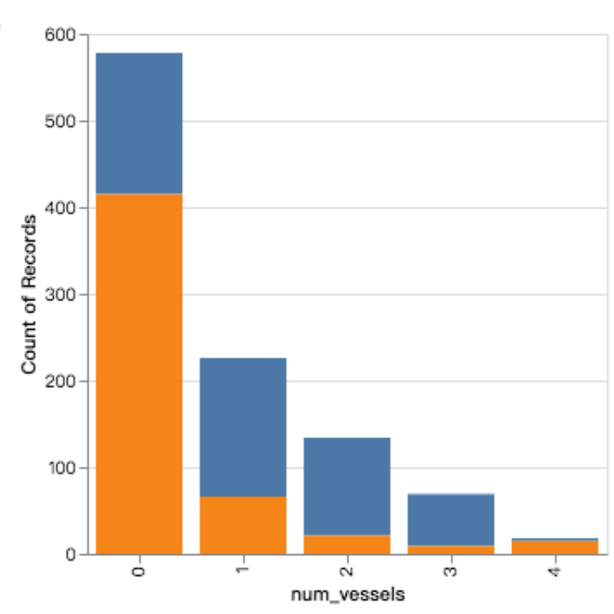
Histogram of age cohort displaying the distribution of age in the study population, stratified into cohorts of 10 years. The bars represent the frequency of participants in each age cohort, with the height of each bar indicating the number of individuals in the cohort. The histogram is divided into two subgroups, representing the counts of individuals with and without heart disease within each age cohort.

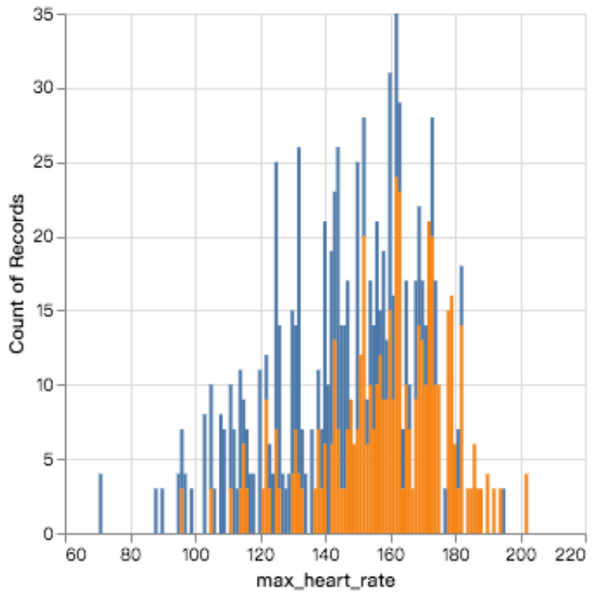
Other Significant Variables

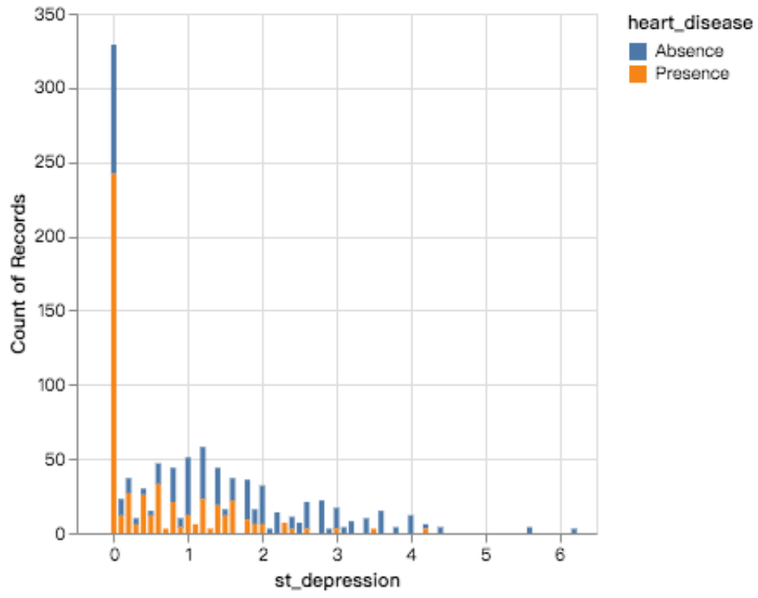












Chest Pain Type(`chest_pain_type`):We can see from the plot that people with atypical angina, non-anginal, and asymptomatic chest pain has greatest risk of developing heart disease. Though there are people who have heart disease has typical angina chest pain, the effect of it is much less significant compare to other types of chest pain.

Form of Thalassemia(`thal`): Fixed effect form of thalassemia have most significant influence on heart disease and reversable effect form of thalassemia have the least influence on heart disease. This make sense since fixed effect of thalassemia is more severe than reversable as explained by the name of the forms that fixed effect cannot be altered.

Angina Exercise(`angina_exercise`): Exercise with angina can help reduce the risk of developing heart disease, as there's greater portion of people who exercise without inducing angina have heart disease.

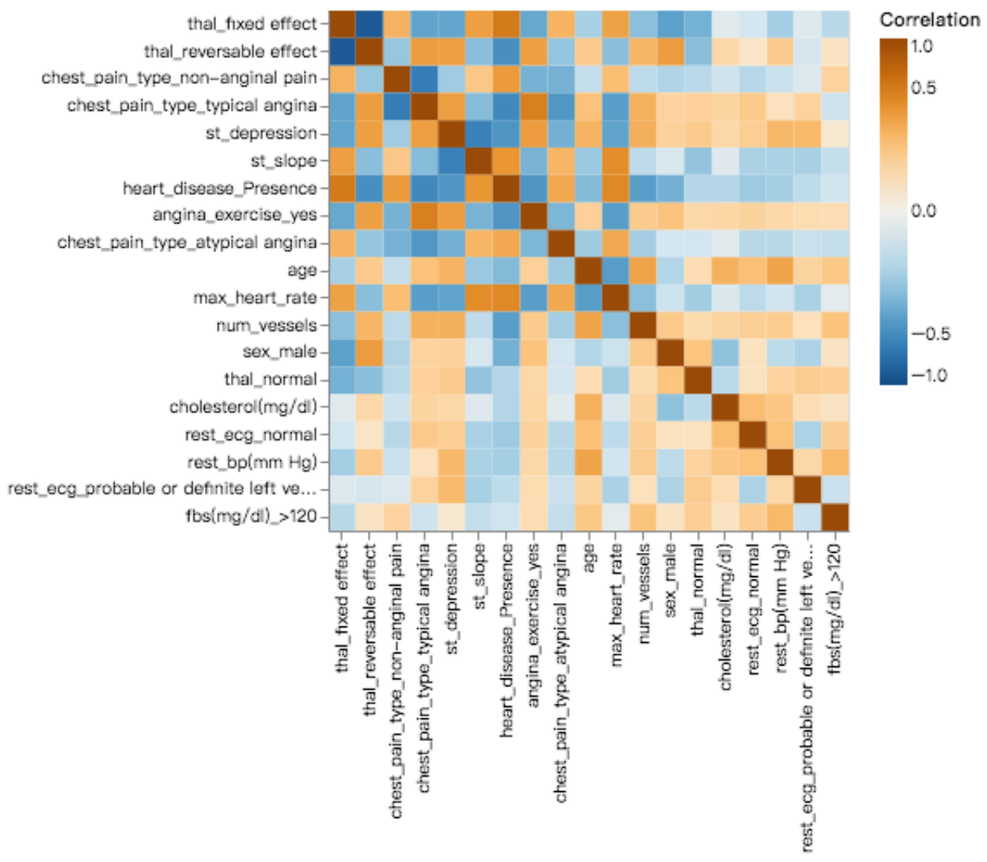
Number of Major Vessels(`num_vessels`): People with least major vessels colored by flourosopy have significant chance gettig heart disease. It might be explained by that only healthier vessels can be seen by flourosopy, so people with 0 vessels seen by flourosopy means that they have weaker vessels than others, which lead to a higher risk of developing heart disease.

Max Heart Rate(`max_heart_rate`): We can see a strong pattern in the plot. People with higher maximum heart rate tend to have a much higher possibility to develop heart disease. We can conclude that maximum heart rate is a major factor for the presence of heart disease.

ST Depression(`st_depression`): There's more observation with st_depression of 0 and we can look further into that in later steps. The chance of getting heart disease increase as `st_depression` decrease.

Principal Component Analysis (PCA)

Correlation and Correlation Matrix




```
chest_pain_type_typical angina                          -0.519621
thal_reversable effect                                  -0.479709
st_depression                                           -0.438441
angina_exercise_yes                                     -0.438029
num_vessels                                             -0.382085
sex_male                                                -0.279501
age                                                     -0.229324
rest_ecg_normal                                         -0.160308
rest_bp(mm Hg)                                          -0.138772
cholesterol(mg/dl)                                      -0.099966
thal_normal                                             -0.095541
rest_ecg_probable or definite left ventricular hypertrophy  -0.076357
fbs(mg/dl)_>120                                         -0.041164
chest_pain_type_atypical angina                          0.255288
chest_pain_type_non-anginal pain                         0.319504
st_slope                                                 0.345512
max_heart_rate                                           0.422895
thal_fixed effect                                        0.519543
heart_disease_Presence                                   1.000000
Name: heart_disease_Presence, dtype: float64
```

Correlation matrix in the form of a heat map plot. Shows the correlation between all variables within the data set. Boxes with warmer colors show a positive correlation, while boxes with cooler colors represent a negative correlation.

Correlation table shows a list of all variables in the data set and their correlation to the prescence of heart disease.

Proportion of Variance explained & Cumulative Variance Explained

index	Proportion of variance explained	Component	Cumulative variance explained
0	0.23275997892672742	1	0.23275997892672742
1	0.09361348519526133	2	0.32637346412198875
2	0.07644390945627728	3	0.402817373578266
3	0.0731322376361446	4	0.4759496112144106
4	0.06593263346137175	5	0.5418822446757824
5	0.0589493068030239	6	0.6008315514788063
6	0.05374656899224572	7	0.654578120471052
7	0.048781198445352024	8	0.703359318916404
8	0.04459016371120048	9	0.7479494826276045
9	0.042895050326415905	10	0.7908445329540205
10	0.04190641697895566	11	0.8327509499329762
11	0.03658315327318052	12	0.8693341032061567
12	0.03355395861848885	13	0.9028880618246455
13	0.02899378654973667	14	0.9318818483743823
14	0.02307136878638746	15	0.9549532171607698
15	0.01983595425260856	16	0.9747891714133783
16	0.01904185091164263	17	0.993831022325021
17	0.005536968170952468	18	0.9993679904959735
18	0.0006320095040269337	19	1.0000000000000004

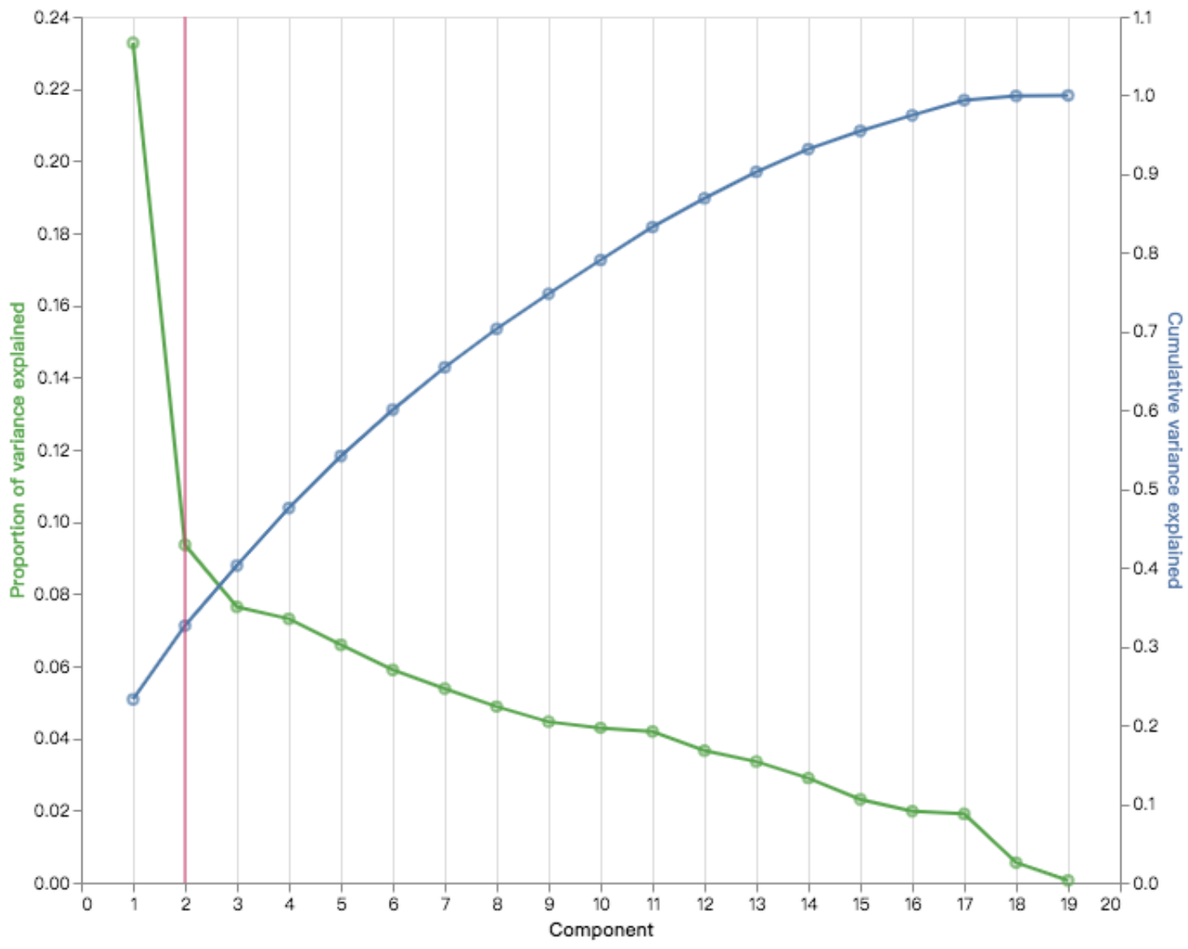


Table displaying the amount of variance explained by a principal component. Records the value of variance for each individual component and the cumulative variance at each added component.

The Point-Line plot displaying the loading values of each variable in the first principal component (PC1) of the dataset. The x-axis represents the variable names, and the y-axis represents the loading values. The loading values indicate the contribution of each variable to PC1, with larger loading values indicating stronger contribution

Principal Component

index	PC1	Variable
0	0.18001513705789243	age
1	0.11406880136432554	rest_bp(mm Hg)
2	0.05892630673114266	cholesterol(mg/dl)
3	-0.29281076456813904	max_heart_rate
4	0.3067127908366767	st_depression
5	-0.2641438762713849	st_slope
6	0.19606797938172318	num_vessels
7	0.13424004653094432	sex_male
8	-0.1991631885319484	chest_pain_type_atypical angina
9	-0.19223426134697255	chest_pain_type_non-anginal pain
10	0.3274019605529025	chest_pain_type_typical angina
11	0.04805254971401268	fbs(mg/dl)_>120
12	0.0984483165627004	rest_ecg_normal
13	0.06275767649626887	rest_ecg_probable or definite left ventricular hypertrophy
14	0.28888130531449546	angina_exercise_yes
15	-0.34422773086612285	thal_fixed effect
16	0.09319924761954368	thal_normal
17	0.3042527361090654	thal_reversable effect
18	-0.3769774499891303	heart_disease_Presence

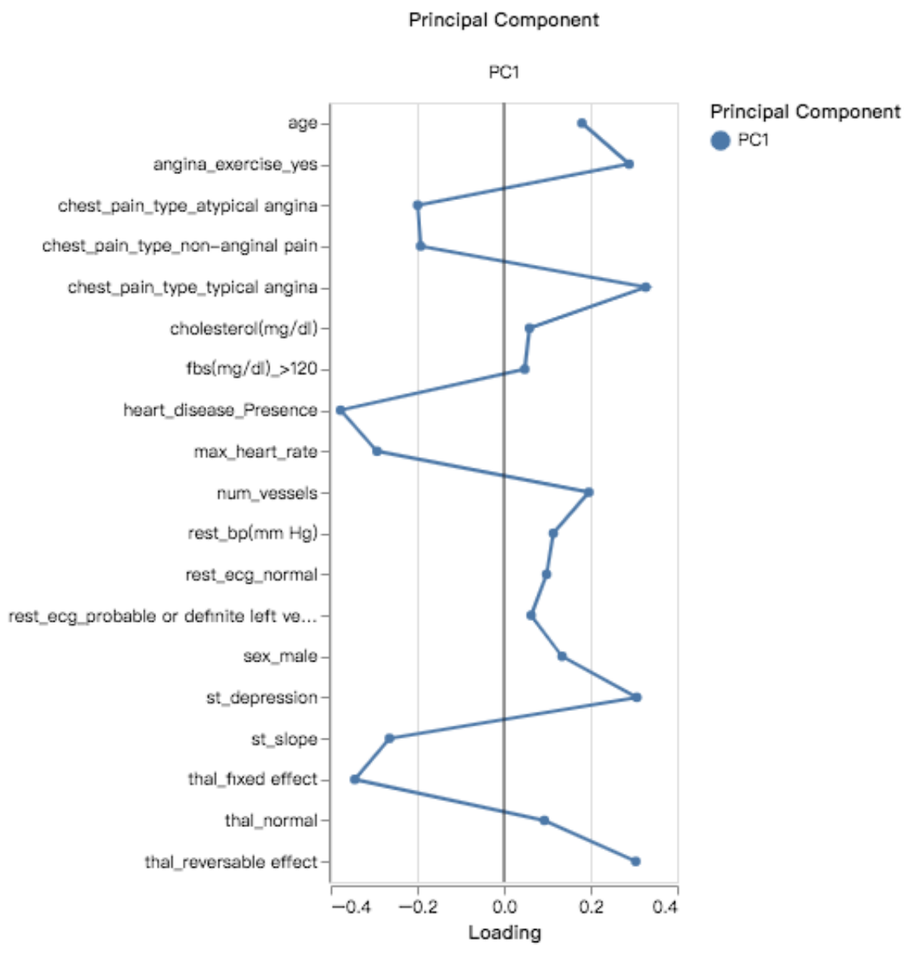


Table displaying the loading values of each variable in the first principal component (PC1) of the dataset. The table provides a comprehensive overview of the relative contributions of each variable to PC1, with larger loading values indicating stronger contributions.

Line plot displaying the loading values of each variable in the first principal component (PC1) of the dataset. The x-axis represents the variable names, and the y-axis represents the loading values. The loading values indicate the contribution of each variable to PC1, with larger loading values indicating stronger contribution

---

Discussion

By generating a correlation plot, we were able to identify the variables that are most strongly associated with the presence of heart disease. We found that thal_fixed effect, max_heart_rate, and st_slope exhibited high positive correlations with the target variable. This observation is further supported by our barchat plots, which show a higher incidence of heart disease for these variables. Conversely, we observed negative correlations between heart disease and variables such as chest_pain_type_typical angina, thal_reversible effect, and st_depression. These variables exhibited a lower incidence of heart disease in our barchart plots.

One way to confirm the presence of underlying patterns in our data is to apply a PCA (Principal Component Analysis) approach. By computing the principal components, we identified which variables explained the majority of the variance in the data. We focused on the first principal component, as it accounted for the greatest proportion of variation. Examining the loadings of the variables on this component reveals that several variables, including the type of chest pain, maximum heart rate, form of thalassemia, ST positions, and different forms of thalassemia, are strongly associated with the patterns observed in the data.

We noticed that there's a surprising finding in our analysis that younger people are more likely to develop heart disease. It might because of the unhealthy lifestyle younger people have due to study or work. Younger people have no time to eat or sleep due to pressure and huge work load, so they choose to eat fast food and do not have enough sleep, which end up with high rates of obesity and high heart rate, and eventually increase their risk of getting heart disease. We can see from the correlation matrix that there's a strong negative correlation between age and max heart rate, which support our speculation of the cause.

This time, our group explores data for patients with potential heart diseases and found indicators of heart diseases. Next time, we would also explore something related to health, as this is most important for human beings. I would like to analyze data like Healthcare utilization analysis, Patient health monitoring, or Disease outbreak modeling in the future.