BIS 634 Final Report

Wei Pang

GETTING STARTED

Describe dataset and why it is interesting

https://www.kaggle.com/datasets/nareshbhat/health-care-data-set-on-heart-attack-possibility

I found this dataset on Kaggle, which contains 303 samples with 13 independent variables and 1 dependent variable. The dependent variable in this dataset is the outcome of whether an individual has had a heart attack, which is recorded as a binary variable with a value of "1" indicating that the individual has had a heart attack and a value of "0" indicating that they have not. Heart attack is a major public health problem and understanding the risk factors associated with it is crucial for developing strategies to prevent and treat it. This dataset includes a wide range of potential risk factors and could be used to develop predictive models to estimate an individual's risk of heart attack based on their specific risk profile. Overall, this dataset would be useful for researchers and healthcare professionals interested in understanding and reducing the risk of heart attack.

Attribute	Description
age	age in years
sex	sex (1 = male; 0 = female)
cp	chest pain type
	-- Value 1: typical angina
	-- Value 2: atypical angina
	-- Value 3: non-anginal pain
	-- Value 4: asymptomatic
trestbps	resting blood pressure (in mm Hg on admission to the hospital)
chol	serum cholestoral in mg/dl
fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
restecg	resting electrocardiographic results
	-- Value 0: normal
	-- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
	-- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
thalach	maximum heart rate achieved
exang	exercise induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slope	the slope of the peak exercise ST segment
	-- Value 1: upsloping
	-- Value 2: flat
	-- Value 3: downsloping
ca	number of major vessels (0-3) colored by flourosopy
thal	0 = normal; 1 = fixed defect; 2 = reversable defect

Explain how acquired it

Since this dataset was found from Kaggle, it requires the use of Kaggle's library to access it via API. In addition, Kaggle account authentication is required for this to work. So even though this dataset is open source, access to the dataset requires platform authentication. So for convenience, I downloaded the dataset to the host.

Discuss FAIRness

- Findability: This dataset is open source and easy to find. And medical terminology standards are used to describe the variables.
- Accessibility: This dataset is easy to access and only requires a Kaggle account to download it. a Kaggle account is completely free. It is also possible to analyze the dataset directly on Kaggle.
- Interoperability: As a CSV document, it can easily be processed by various programming languages.
- Reusability: The usefulness of this dataset is relatively homogeneous and is limited to the prediction of the cause of heart attack and the description of patient status.

Describe data cleaning or preprocessing

The dataset is very clean with uniform data types. There are no null or residual values. No additional normalization operations are required. For ease of handling, I have marked

the category variables in it for easy differentiation from continuous variables.

```
heart_df.dtypes
```

```
age            int64
sex         category
cp          category
trestbps       int64
chol           int64
fbs         category
restecg     category
thalach        int64
exang       category
oldpeak      float64
slope       category
ca          category
thal        category
target         int64
dtype: object
```

Put data in standard format if necessary

Yes, as mentioned above, I changed the data type of the category variable from int64 to category for easier processing. Data is stored in a dataframe.

ANALYSIS

Any issues with summary statistics

No issues were found from the summary statistics

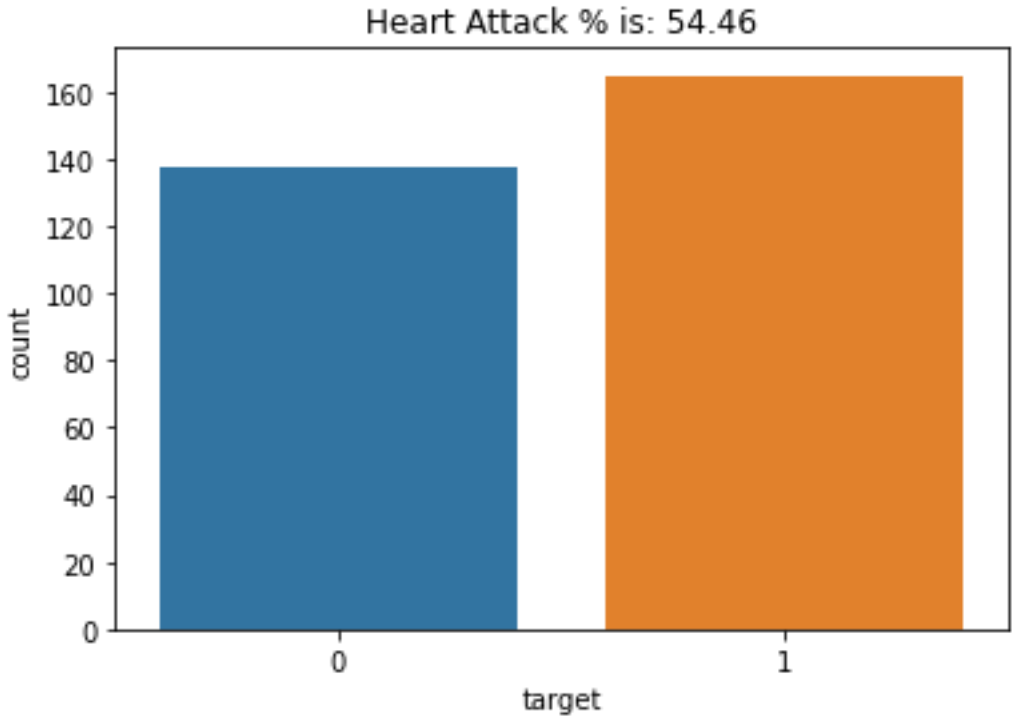
```
heart_df.describe()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.00
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.31
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.61
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.00
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.00
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.00
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.00
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.00

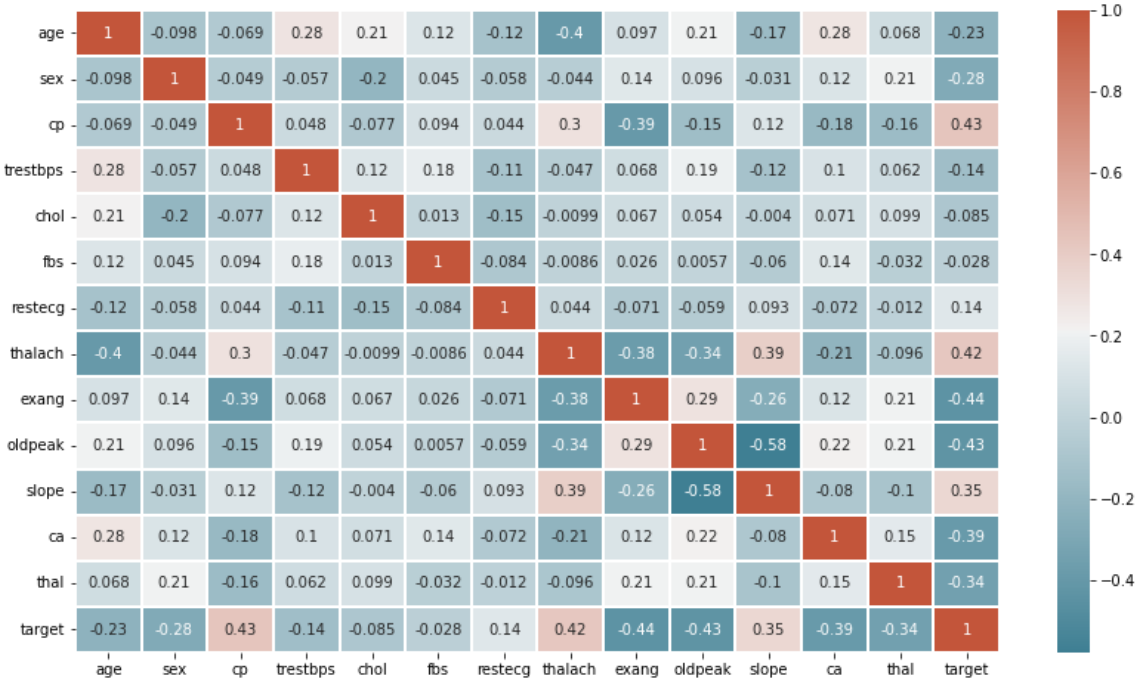
There were 303 patients, ranging in age from 29 to 77 years. The "count" showed no null values

Discuss the analyses ("interesting") you chose to run

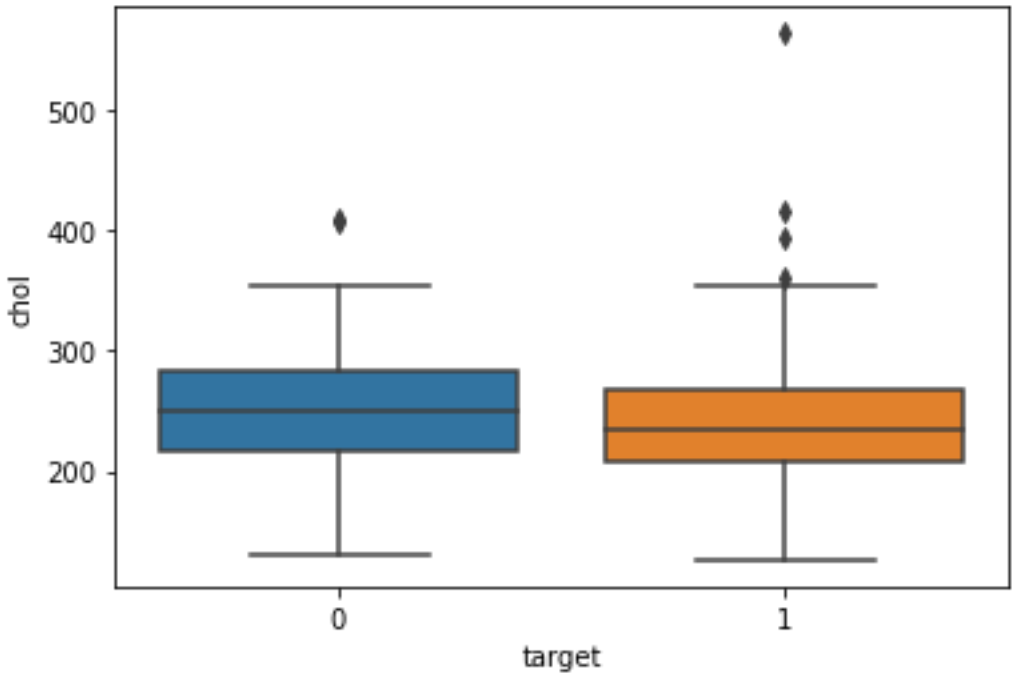
Bar graphs were plotted for Target, and 54.46% of the sample had heart attacks.

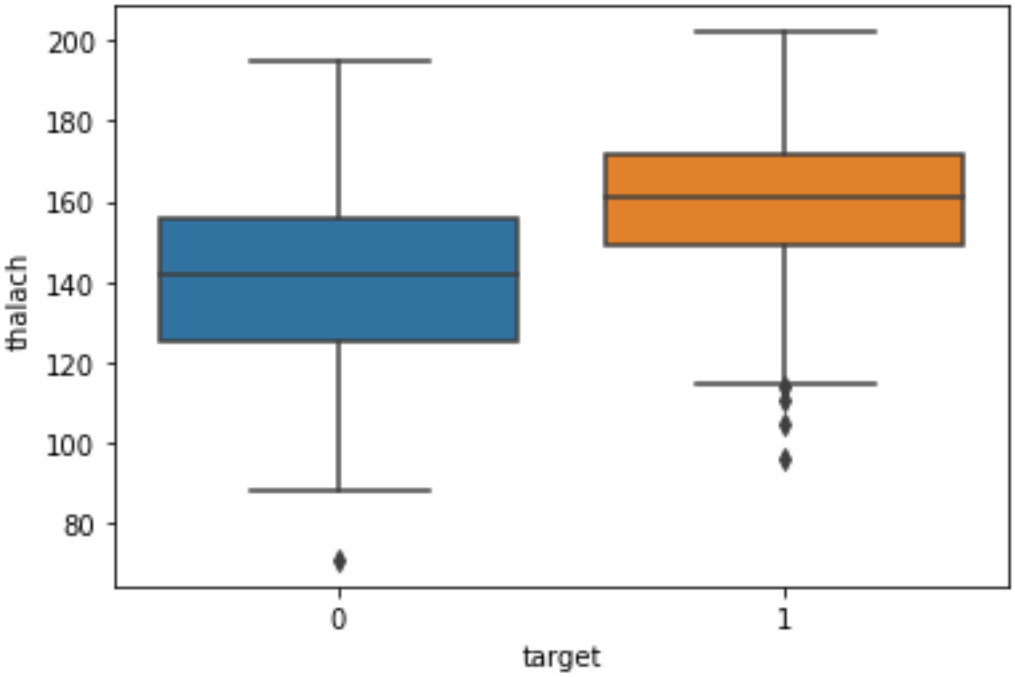


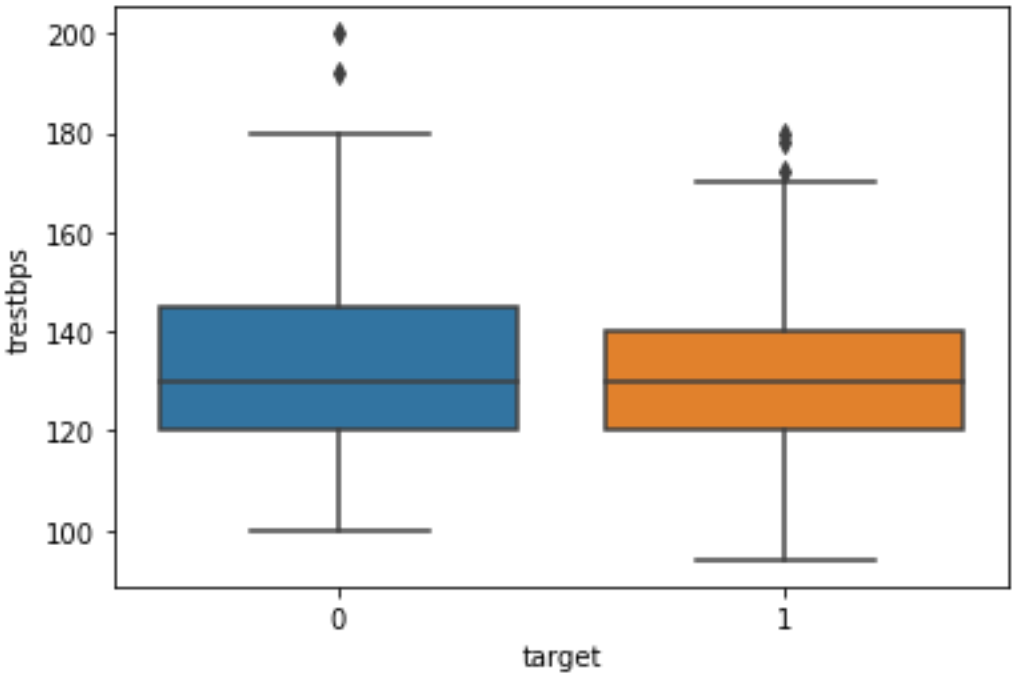
Pearson's correlation coefficient is a way to understand the correlation between variables. I use a heat map to present the results. It can be found that the variables cp, thalach and the result Target are significantly positively correlated. And exang, oldpeak and ca are significantly negatively correlated with the result Target.



I performed a box plot analysis for each of the three continuous variables thalach, chol, and trestbps. I found no significant differences in cholesterol levels and blood pressure between patients with heart attacks and healthy individuals, but their maximum heart rate was significantly higher than that of healthy individuals.

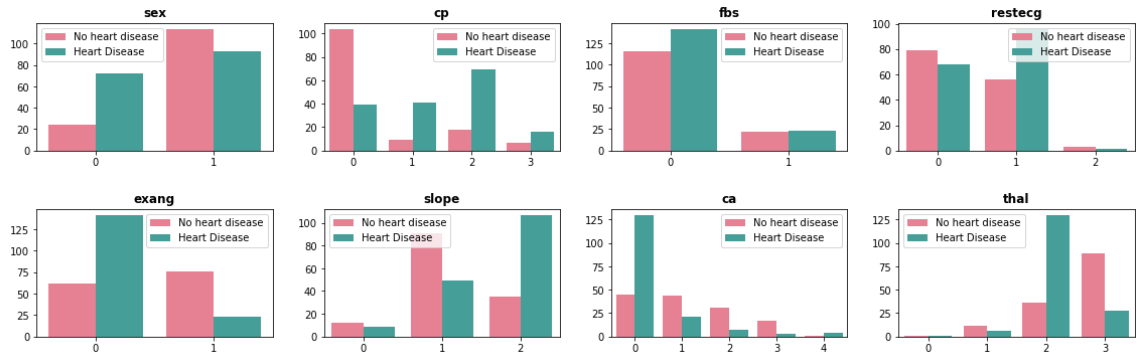

chol
500
400
300
200
0
1
target


thalach
200
180
160
140
120
100
80
0
1
target



A

Visual representation of the bar chart was performed for each of the eight category variables. Men were found to be more likely to suffer from heart attacks. Heart attack and the history of diagnosis of myocardial ischemia was correlated. Patients with diseases that cause irreversible myocardial damage (e.g., thalassemia) were more likely to have heart attack. Exercise and heart attack were negatively correlated. One of the valid diagnostic markers of heart attack is the electrocardiographic index.



- Why these questions?

What causes heart attacks and what can be done to avoid them? What is the pattern of heart attack?

- What were the results?

Heart attack is a common condition. It is not clear why men in the sample were more likely to develop it. Heart attacks are associated with myocardial ischemia and a history of disease that damages the heart muscle. A healthy lifestyle and timely diagnosis and

treatment can help avoid heart attacks. Studying the relationship between variables can identify ways to avoid heart attacks, which is relevant for the prevention of a disease with high morbidity.

- Any surprises?

Many people experience angina after exercise, even young people like me. At one time I would have thought that over-exercise had caused damage to my heart. But the data show a significant negative correlation between post-exercise angina and heart attack. After further research I found that post-exercise angina is normal and increases the oxygen supply to the heart, thus preventing heart attacks.

Men are more likely to have heart attacks. Probably because men are also more likely to engage in behaviors that increase the risk of heart disease, such as smoking, drinking alcohol in excess, and not exercising regularly. These behaviors can damage the heart and blood vessels, increasing the risk of a heart attack.

- validation of analyses?

I applied the insights derived from these analyses to a subsequent machine learning model and used it to analyze the probability of heart attack. The model achieved a 91% correct rate. Although my conclusions were not necessarily correct, there was a greater likelihood of successfully predicting the outcome.

I also reviewed several peer reviewed papers related to heart attack and found that the above statistical analysis results are indeed the main cause of heart attack.

- do more than just summary statistics

I performed machine learning on the dataset using the Free Forest classifier and GridSearchCV. First I used StratifiedShuffleSplit to split the dataset into two significantly different clusters. Then the model was trained. The accuracy for the two clusters were 0.91 and 0.76.

	precision	recall	f1-score	support
0	0.84	0.96	0.90	91
1	0.97	0.88	0.92	136
accuracy			0.91	227
macro avg	0.91	0.92	0.91	227
weighted avg	0.92	0.91	0.91	227

	precision	recall	f1-score	support
0	0.69	0.77	0.73	31
1	0.83	0.76	0.79	45
accuracy			0.76	76
macro avg	0.76	0.76	0.76	76
weighted avg	0.77	0.76	0.76	76

- two analyses that generate graphs

I generated a variety of graphical results such as heat maps, box plots, bar graphs, etc.

- at least one analysis that takes a parameter

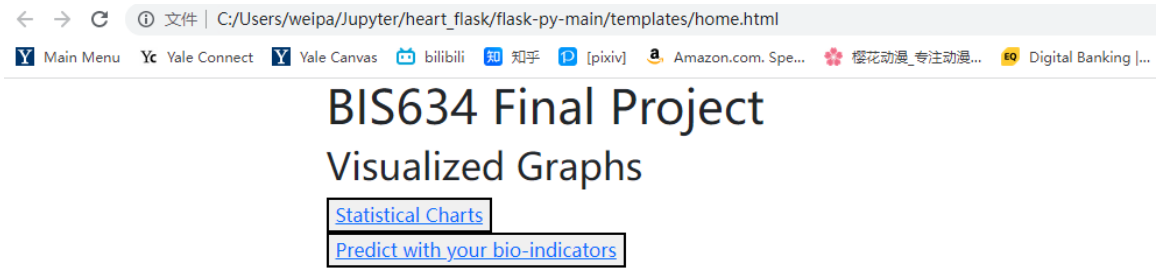
Machine learning trained predictive models use a variety of parameters. I used GridSearchCV to perform parameter tuning.

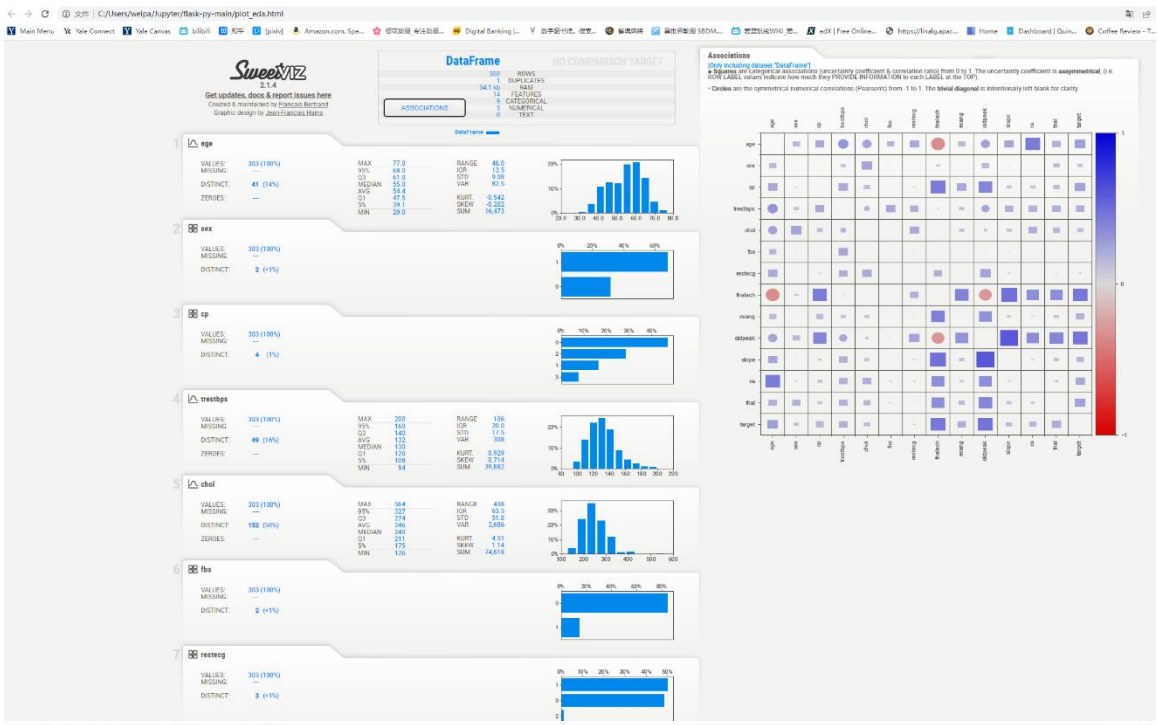
WEB BACKEND AND FRONTEND (?/25 POINTS TOTAL)

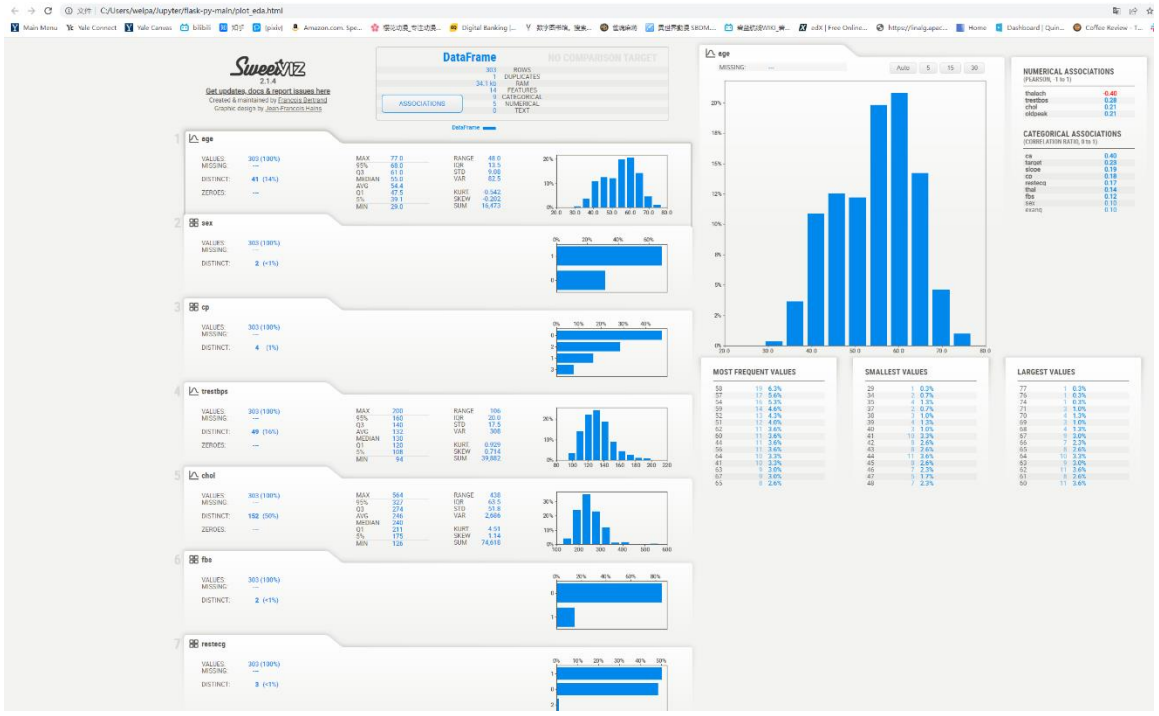
Describe your server API and the web front-end

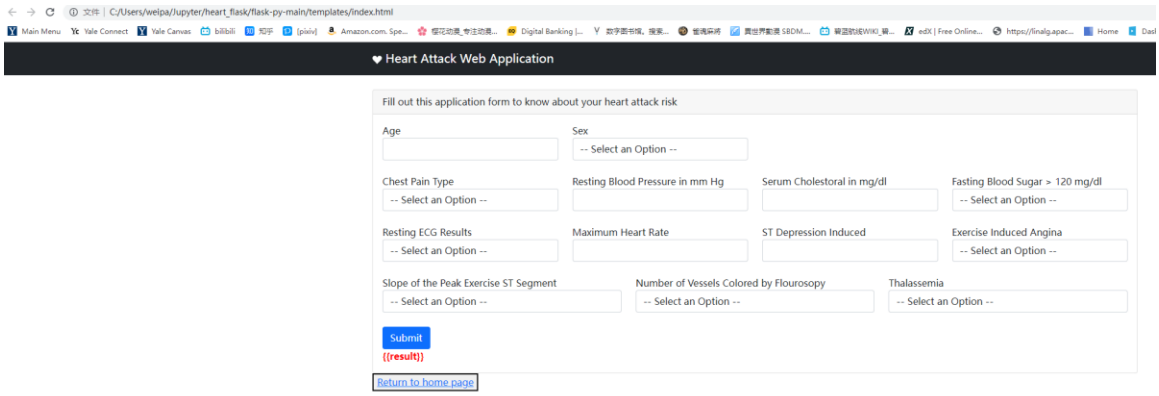
My web page consists of a homepage, a page presenting statistical charts, and an interactive page.

Has a web interface


文件 | C:/Users/weipa/Jupyter/heart_flask/flask-py-main/templates/home.html
Main Menu
Yale Connect
Yale Canvas
bilibili
知乎
[pixiv]
Amazon.com. Spe...
樱花动漫_专注动漫...
Digital Banking |...
BIS634 Final Project
Visualized Graphs
Statistical Charts
Predict with your bio-indicators


C:/Users/welpa/jupyter/flask-py-main/plot_eda.html
SweetViz
2.1.4
Get updates, docs & report issues here
DataFrame
NO COMPARISON TARGET
303 ROWS
1 DUPLICATES
34.1 kb RAM
14 FEATURES
9 CATEGORICAL
5 NUMERICAL
0 TEXT
ASSOCIATIONS
Associations
1 age
2 sex
3 cp
4 trestbps
5 chol
6 fbs
7 restecg


C:/Users/welpa/jupyter/flask-py-main/plot_eda.html
SweetViz
2.1.4
Get updates, docs & report issues here
DataFrame
NO COMPARISON TARGET
303 ROWS
1 DUPLICATES
34.1 kb RAM
14 FEATURES
9 CATEGORICAL
5 NUMERICAL
0 TEXT
ASSOCIATIONS
NUMERICAL ASSOCIATIONS
CATEGORICAL ASSOCIATIONS
MOST FREQUENT VALUES
SMALLEST VALUES
LARGEST VALUES



interface

Users can choose to access the graphical interface or access the interactive interface for parameter input and result query. The user must enter 13 parameters in the interactive interface. For different parameter inputs, the trained model will give different answers to assess the risk of heart attack.

Routing is established via flask to fetch the graphs and trained prediction models from local. The results are presented as a graph