SI618 FINAL PROJECT REPORT

TABASSUM NISHA (Uniquename- tabbie)

MOTIVATION:

The term “Heart Disease” refers to a wide range of conditions that affects the normal functioning of heart such as a number of blood vessel diseases like coronary artery disease, abruption in the rhythm of the heart (arrhythmias) and congenital heart defects.
The term "Heart Disease" and “cardiovascular Disease” are often used interchangeably. Cardiovascular disease generally involves conditions such as narrowing or blocking of blood vessels leading to heart attacks, chest pains(anginas) or stroke.
There are several factors that determine the heart health of an individual. Some of the major ones being:

- Total cholesterol: Should not be higher than 200 mg/dL. About 56% of the Americans suffer from high cholesterol levels.
- Blood sugar: About 9 percent of the adult population suffer from diabetes. With growing rates of diabetes around 35% of the Americans are at risk of diabetes and hence marked pre-diabetic.
- Blood pressure: About 33 percent of Americans have blood pressure higher than 140/90 mm Hg.

Therefore, I performed an analysis on the heart disease data from UCI to determine which factors lead to heart diseases.

DATASOURCE:

The dataset is from UCI machine learning repository and the creators of the dataset are as follows:

- Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
- University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
- University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
- V.A. Medical Center, Long Beach, and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

The data was created on 2018-06-26 and it was collected from the Cleaveland clinic database. This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to
this date.

The dataset is in the "csv" format and was collected from https://www.kaggle.com/ronitf/heart-disease-uci
Each row represents one individual whose various details were recorded. The data contains the following column variables which are of numerical type:

- age: age in years
- 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 cholesterol 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 0: upsloping
    - Value 1: flat

    - Value 2: downsloping
- ca: number of major vessels (0-3) colored by fluoroscopy
- thal: thallium stress test
    - Value 0: null
    - Value 1: 6 = fixed defect
    - Value 2: 3 = normal
    - Value 3: 7 = reversable defect

14. target: have disease or not (1 = yes, 0 = no)

DATA MANIPULATION AND CLEANING:

The dataframe contains 303 rows and 14 columns comprising of numerical and float values. To check for any missing values or noisy data, I performed the statistical analysis using describe function and implemented the drop NaN values function. I prepared the additional categorial columns from the existing numerical columns for the further analysis.
I created the age_group column from age column, heart_disease column from target column, gender from sex column, and fbs_greater_120 from fbs column. Looking at the maximum and minimum values in the age column, I generated 6 age groups. If a value in the age column falls in the following group, the relevant age-group is assigned to that row.
The gender column comprises of two categories, male and female which corresponds to values in the sex column(1=male, 0=female). The heart_disease column also comprises o two categories corresponding to the respective values in the target column(1=yes, 0=no). Similarly, the fbs_greater_120 has two categories corresponding to the fasting blood sugar column(1=true, 0=false).

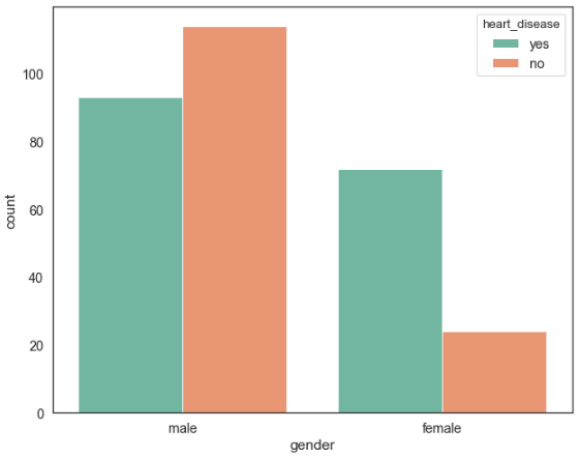
QUESTIONS:

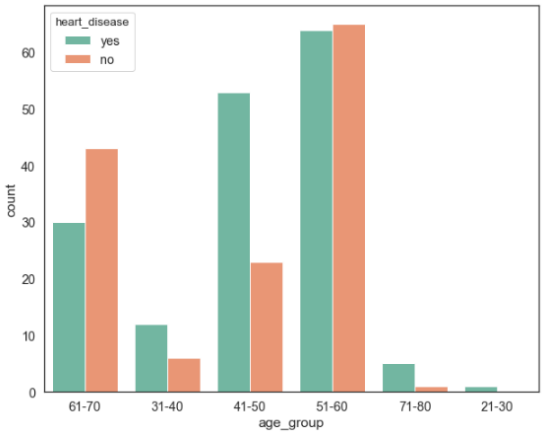
Are heart diseases gender specific and/or age specific? How many clusters do you think there are in the data frame for that obtained gender in the obtained age range?

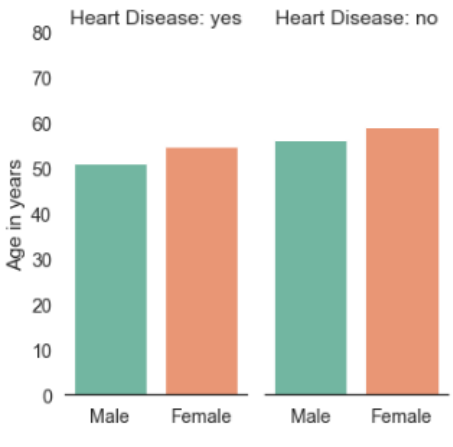
MOTIVATION AND METHODS:

In order to determine whether or not heart diseases are gender and or age specific, I realized I needed the age column in a grouped manner so as to have a less crowded countplot. Hence, I performed the desired data manipulation beforehand. I also felt it would be better to utilize the categorical columns to avoid labeling the countplots. I filtered out the columns age, age_group, gender and heart disease from the main dataframe(heart_data). For the catplot it was better to use the numerical age values rather than categories.
For the cluster analysis depending on the above results from visualistations, I first created a copy of the original dataframe. From the copy dataframe I filtered out all the rows from the relevant age group and then from this filtered dataframe I filtered out all the rows corresponding to females. I dropped all the categorical columns from this filtered dataframe. I used both K-mean method and Silhouette method for determining the cluster patterns in the filtered data. I also used dendogram for Hierarchical Clustering to visualize how the datapoints of a particular cluster show similarity.

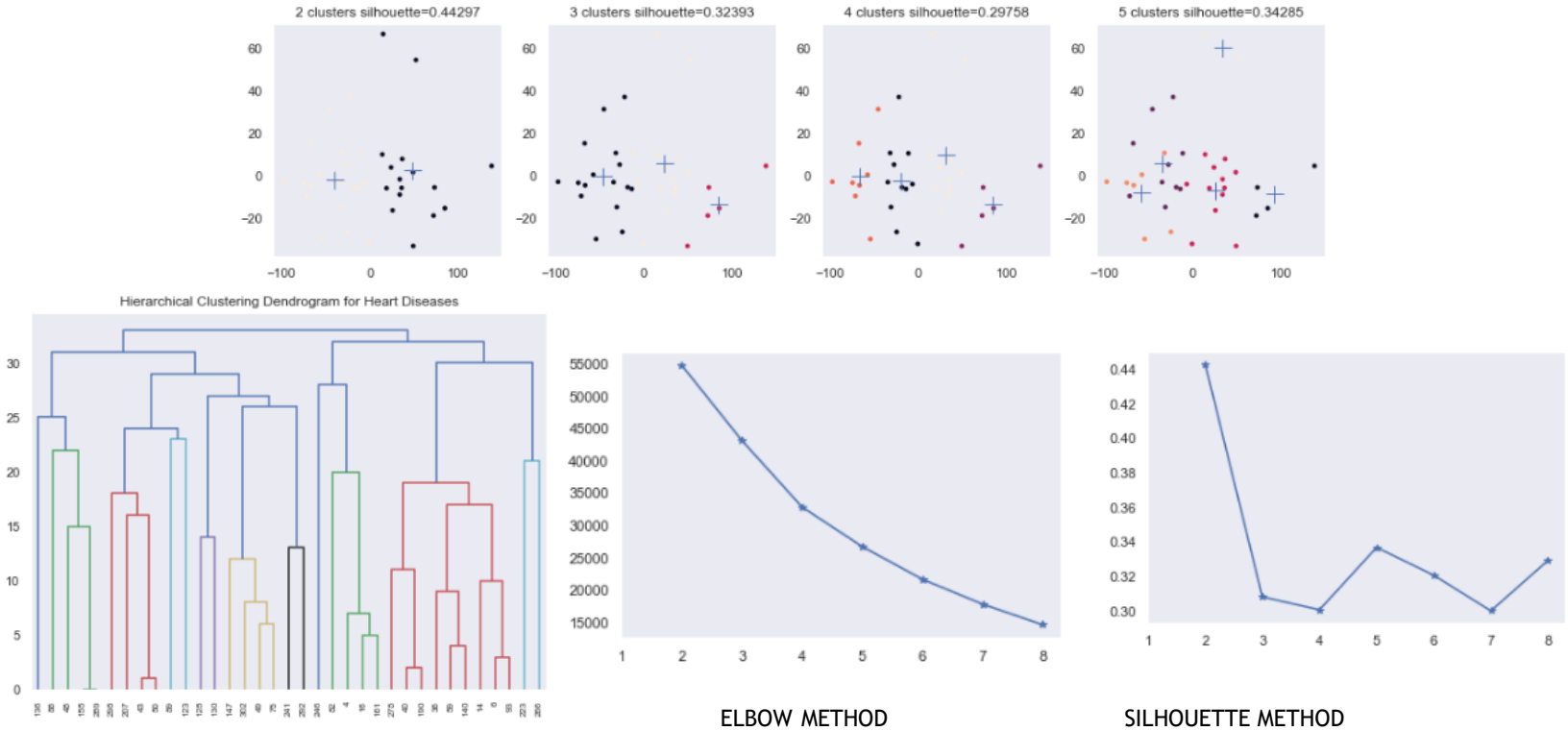
ANALYSIS:







From the first visualization we observe that males show greater frequency for having heart diseases. Hence, they have greater tendency for heart diseases as compare to females. From the second visualization we observe that the age group 51-60 years shows the highest frequency of having heart diseases followed by age group 41-50 years. Hence, they have greater tendency for heart diseases as compare to females. But upon aggregating the two factors in the third visualization, we observe that females of the age about 55 years have heart diseases. Therefore, I conclude that according to this data heart diseases tend to be more age specific rather than gender specific.



ELBOW METHOD SILHOUETTE METHOD

From the above countplots/graphs, I chose females of the age group 51-60 years for examining the clusters. With the help of rule of thumb we obtained the number of clusters which are 4. Elbow method gives an idea on what a good k number of clusters would be based on the sum of squared distance (SSE) between data points. The best k-value is at the spot where SSE starts to flatten out and forming an elbow. The obtained number of clusters from this method is also 4. The silhouette value/score from the scatterplot is 0.29 at number of clusters equal to 4 and hence it indicates that the object is well matched to neighboring clusters and poorly matched to it's own cluster. Hence, there is more seperation and less cohesion among the datapoints. We therefore use a line graph the average distance from all data points in the same cluster and the average distance from all data points in the closest cluster.The coefficient is closer to 0 and thus the sample is very close to the neighbouring clusters.

Which variable/variables show strong relationships to the having a heart disease?

MOTIVATION AND METHODS:

For determining the correlations between the variables, I filtered all the columns with numerical values from the original dataframe(heart_data). I generated a heat map of the filtered dataframe. From the results obtained from the heatmap, I implemented OLS and ANOVA regression model of the variables that displayed strong correlations. This gave me the desired statistical results for analysis. I also generated regression jointplots on the strongly correlated variables. As a supporting visualization, I generation scatterplots also between the strongly correlated variables. The challenges I faced here was understanding the OLS regression model on multiple variables and generating a heatmap of accurate size to be able to view the coefficient numbers in the boxes.

ANALYSIS:

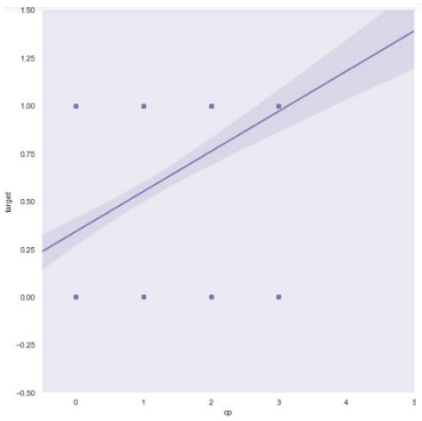
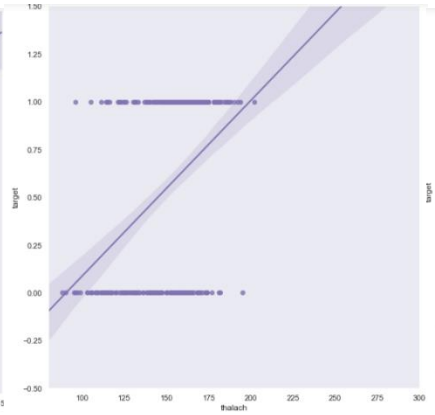
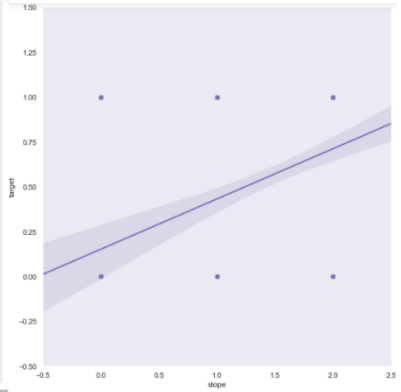
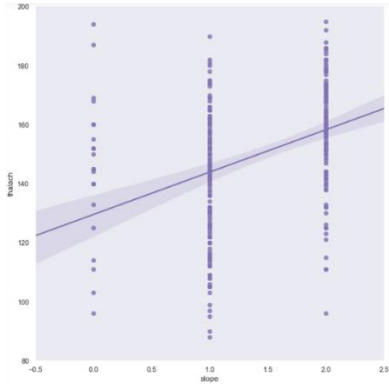
From the above heatmap, we can observe that the correlation coefficients of cp(chest pain) and target(heart disease), thalach(maximum heart rate achieved) and target(heart disease), thalach(maximum heart rate achieved) and slope(slope of the ST segment of electrocadiograph during exercise) and slope(slope of the ST segment of electrocadiograph during exercise) and target(hear disease) are 0.43, 0.42, 0.39 and 0.35. This means that these variables are strongly related as compared to other correlations between the rest of the variables. Therefore, we understand that if an individual experiences chest pains of any kind, has a high heart rate and the slope the ST segment in the ECG are abrupt they're positive for having a heart disease.

Dep. Variable:	target	R-squared:	0.321
Model:	OLS	Adj. R-squared:	0.314
Method:	Least Squares	F-statistic:	47.13
Date:	Mon, 20 Apr 2020	Prob (F-statistic):	5.72e-25
Time:	03:46:44	Log-Likelihood:	-160.05
No. Observations:	303	AIC:	328.1
Df Residuals:	299	BIC:	342.9
Df Model:	3		
Covariance Type:	nonrobust		

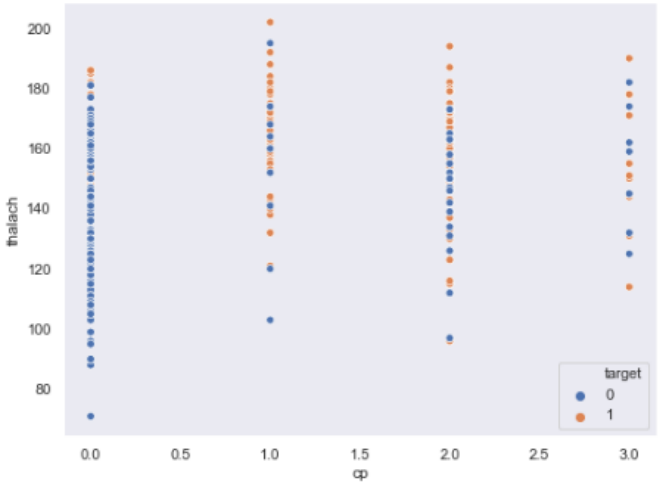
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.6352	0.159	-3.990	0.000	-0.948	-0.322
cp	0.1631	0.024	6.765	0.000	0.116	0.211
thalach	0.0052	0.001	4.462	0.000	0.003	0.008
slope	0.1723	0.042	4.118	0.000	0.090	0.255

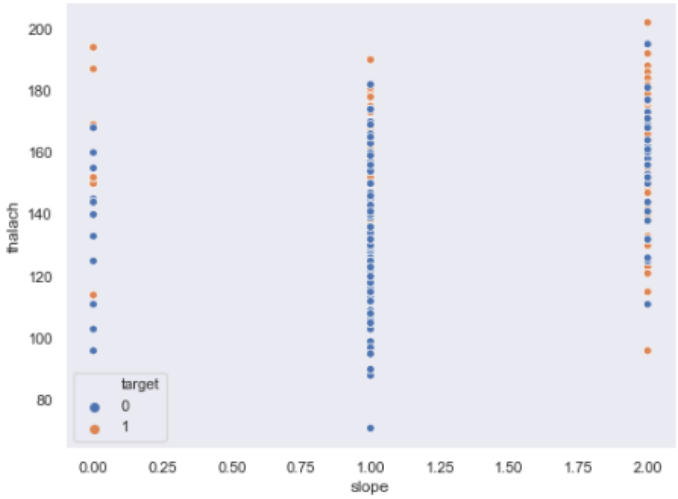
Dep. Variable:	thalach	R-squared:	0.150
Model:	OLS	Adj. R-squared:	0.147
Method:	Least Squares	F-statistic:	52.95
Date:	Mon, 20 Apr 2020	Prob (F-statistic):	2.99e-12
Time:	03:46:50	Log-Likelihood:	-1353.7
No. Observations:	303	AIC:	2711.
Df Residuals:	301	BIC:	2719.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	129.5288	3.020	42.889	0.000	123.586	135.472
slope	14.3768	1.976	7.277	0.000	10.489	18.265

target vs thalach | target vs cp | target vs slope | thalach vs slope





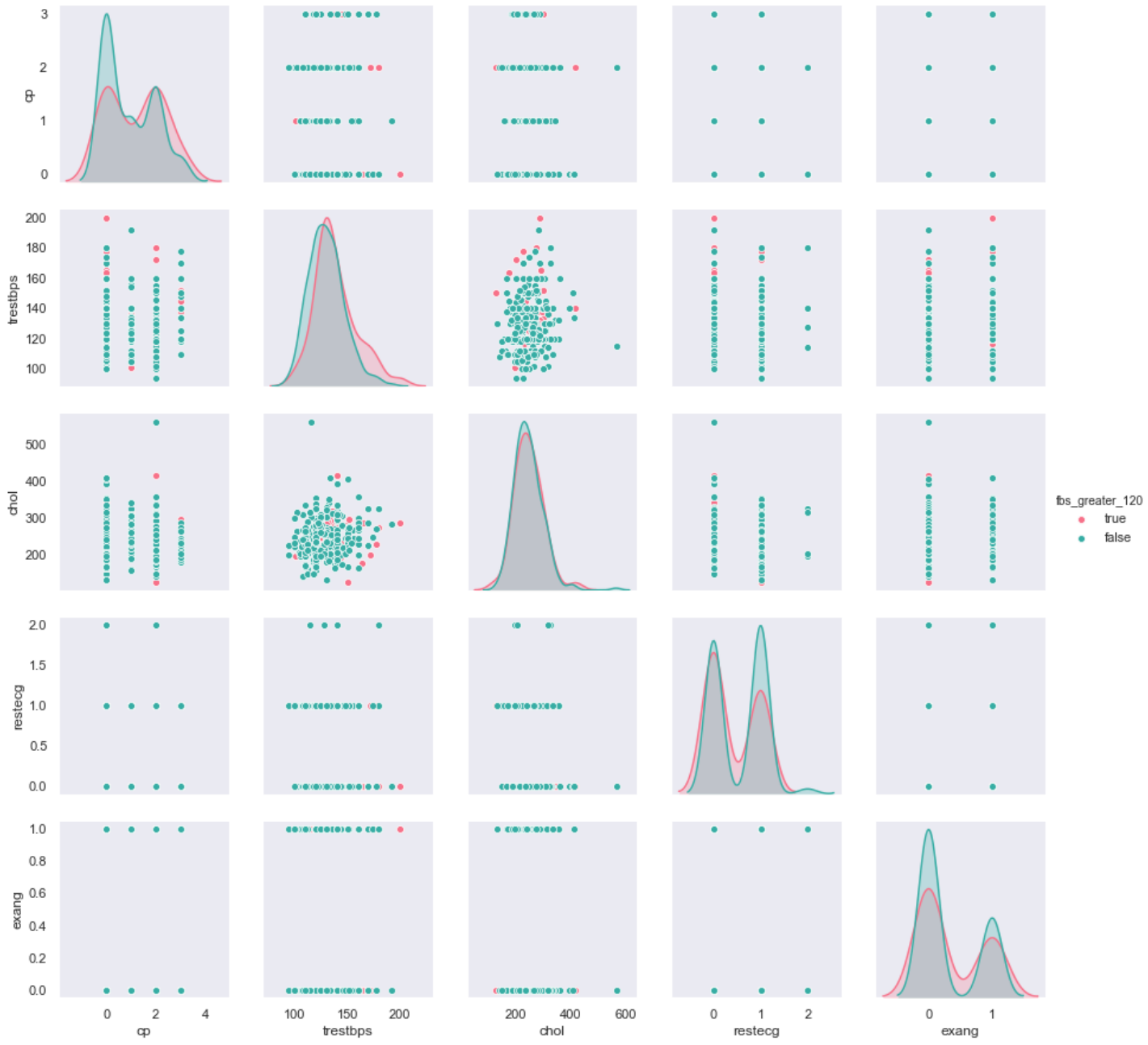
From the above multiple regression model we can observe that the R-squared value is 0.32 which is a satisfactory variation of the dependent variables chest pain, maximum heart rate and slope of ST segment to the target(heart disease). Since the probability of f-statistic and p-value are 0 which is less that 0.05 and that establishes that the two variables in the model are significantly different and the model is a good fit. Therefore we accept that there is variability in the model and hence the dependent variables do relate to the independent variable (target).Individuals with atypical anginal and non-anginal type of chest pain and those with high maximum heart rate and a flat ST slope segment strongly indicate presence of heart diseases. Hence, if an individual experiences chest pains of any kind, has a high heart rate and the slope the ST segment in the ECG are abrupt they're positive for having a heart disease.

How are the factors cp, trestbps, chol, restecg, exang related on the basis of categorical factors?

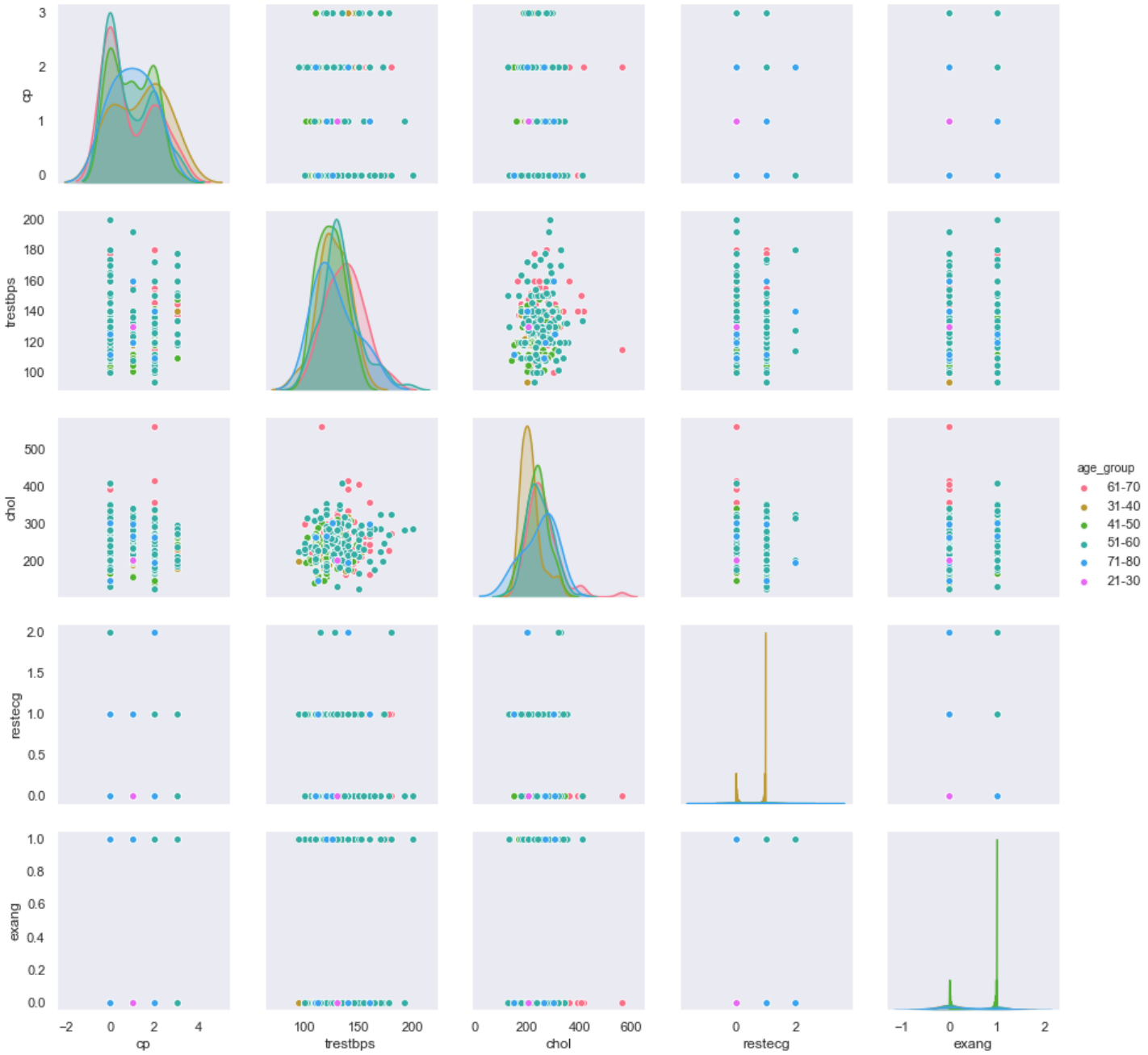
MOTIVATION AND METHODS:

For determining the effect of fasting blood sugar level, heart disease, age group and gender on chest pain, rest blood pressure, cholesterol, resting ECG and and exercise induced angina, I decided to use pairplots. For this approach, I created four separate dataframes from the main dataframe(heart_data) each containing one categorical column and the five numerical columns mentioned in the question. I then implemented pairplots on these four dataframes. The major challenge that I faced was with the pairplot having age group as the hue. But with the help of the some visualtsations from first question, it was easy to correlate my results on some variables.

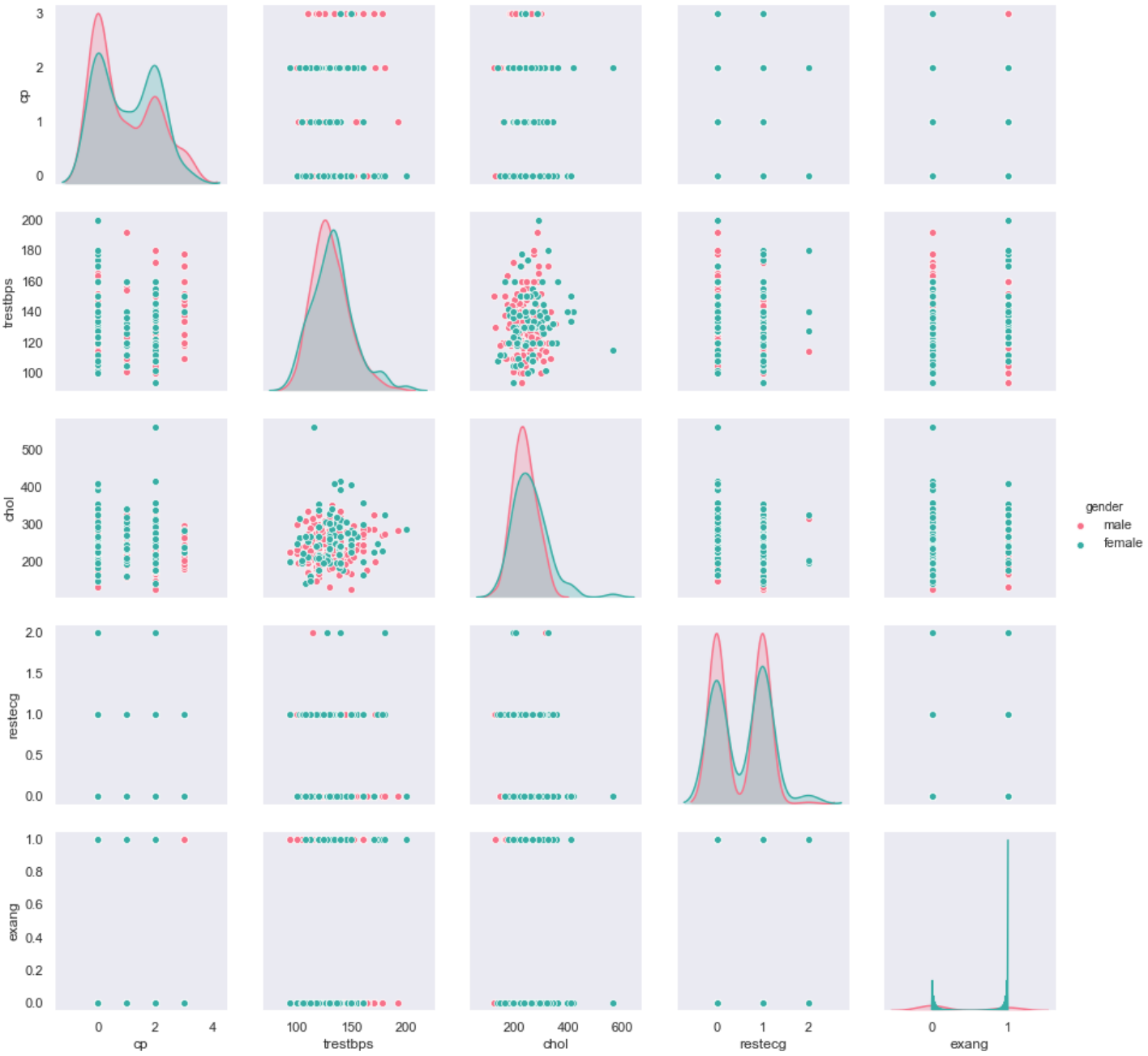
ANALYSIS:



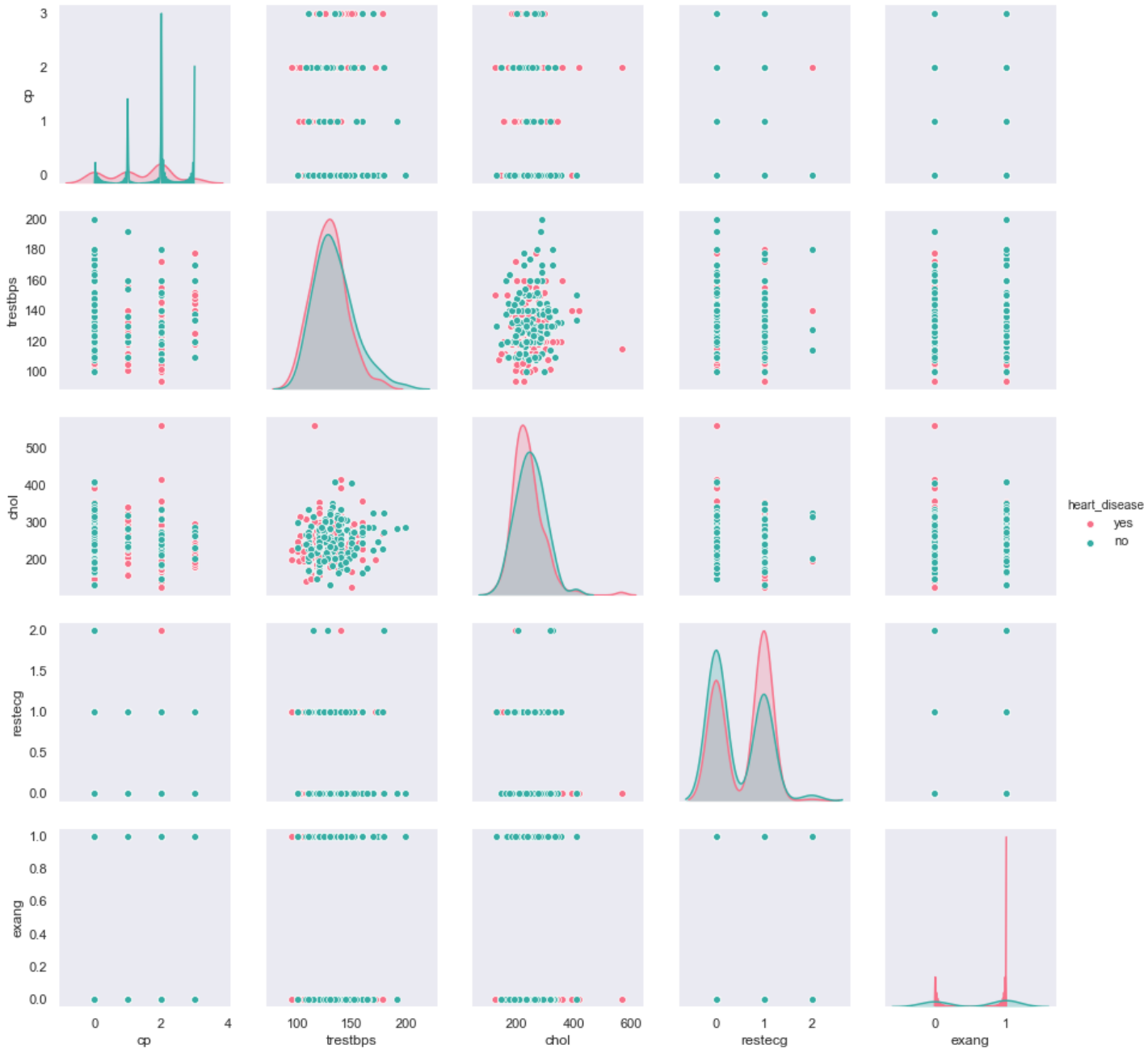
From the above pairplot we observe that majority of the individuals with fasting blood sugar level greater than 120mg/dl tend to have normal resting ECG, no exercise induced angina, and non-anginal type chest pain. On the contrary, majority of individuals with fasting blood sugar level below 120mg/dl tend to have abnormalities in their ST segment of ECG, exercise induced angina, and typicial angina type chest pain. Therefore, we can conclude that it is not necessary that individuals suffering from diabetes are at risk of developing or having heart diseases.



From the above pairplot, we observe that individuals in the age group of 51-60 years tend to have high resting blood pressure and typical angina type chest pain. Individuals in the age-group 31-40 years show highest cholesterol levels but relatively moderate to low resting blood pressure and fairly no chest pain or atypical type of anginal chest pain. It's also interesting to observe that the scatterplot of resting blood pressure indicates that majority of age group 31-40 years have resting blood pressure above 120 mmHg and cholesterol levels around 300 mg/dl. We can also observe from the scatterplots that age group 51-60 years show high abnormality in ST-T wave of the rest ECG and have high positives for exercise induced angina. Therefore, we can conclude that individuals in the age group 51-60 tend to be relatively more positive for heart diseases as they display more number of predisposing factors.



From the above pairplot we observe that males have higher positives for typical angina type chest pain in comparison to other type of chest pains where as females display relatively lower chances of experiencing typical angina type chest pain. Males also display high resting blood pressure and high cholesterol levels as compared to females. Males show high number of abnormalities in the ST-T wave of the resting ECG but surprisingly females shows highest positives for exercise induced angina. Therefore, given the results we can conclude that males display more number of factors that makes them prone to heart diseases as compared to females.



From the above pairplot we can observe that, individuals with atypical anginal chest pain do not display heart diseases. Individuals with resting blood pressure greater than 120 mmHg and cholesterol levels greater than 200 mg/dl are positive for heart diseases. Those positive for heart diseases also display abnormalities in the ST-T wave of resting ECG and present with exercise induced angina. Hence, we can conclude that for having heart diseases factors such as blood pressure, cholesterol, resting ECG and exercise induced angina play a major role.

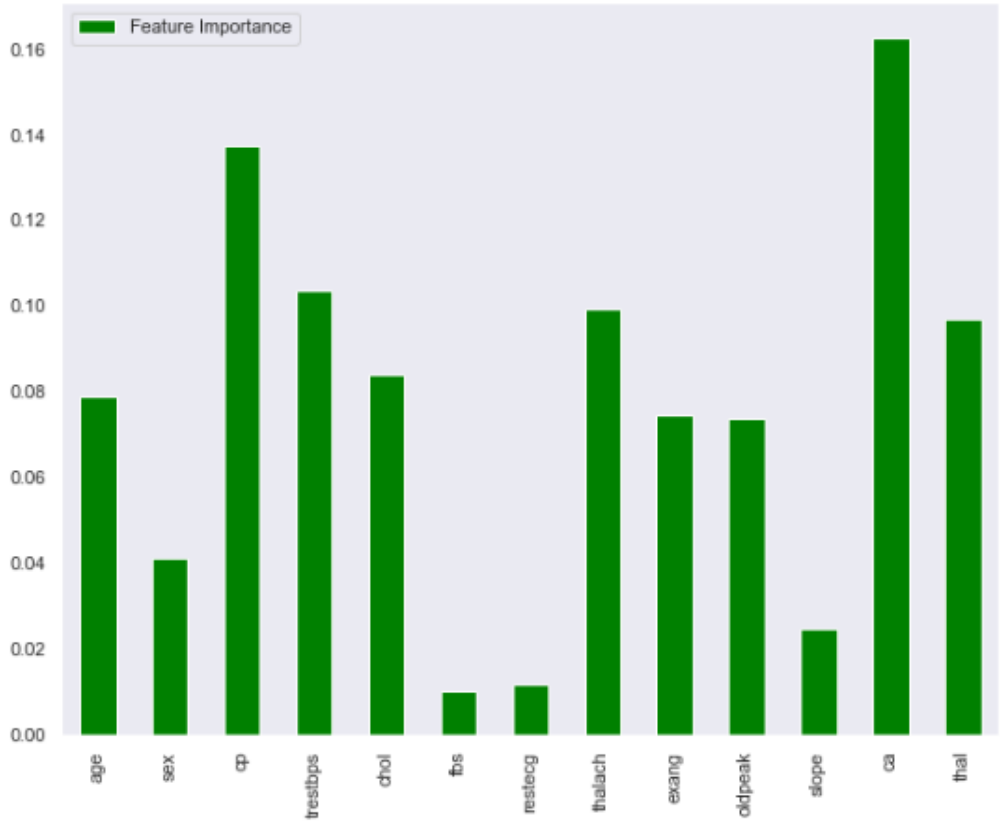
What causes heart disease?

A machine learning approach

MOTIVATION AND METHODS:

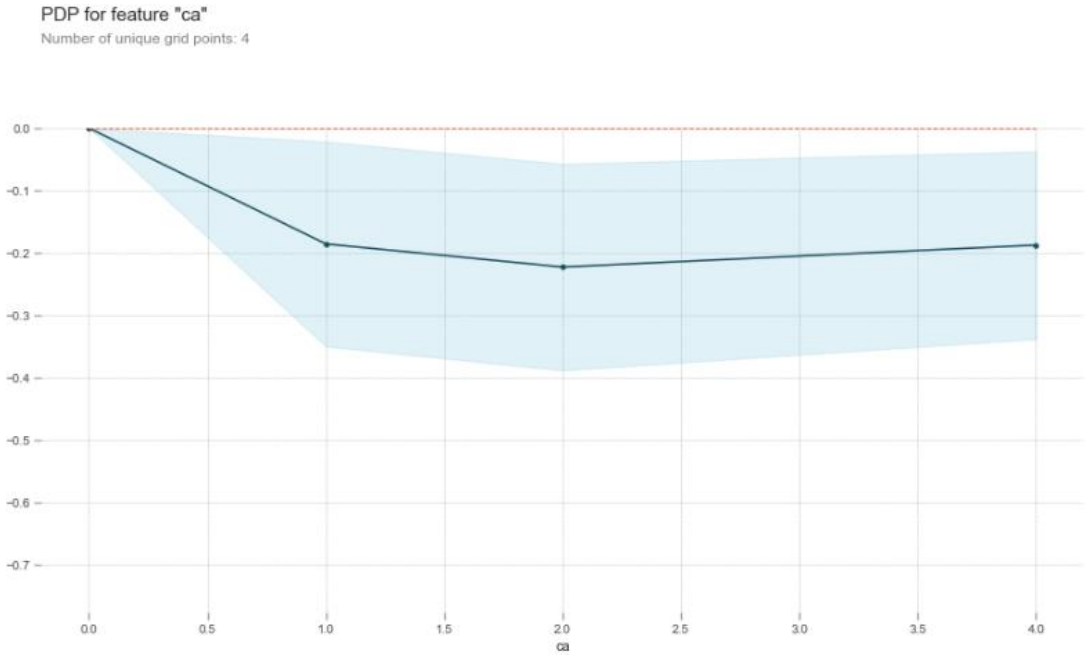
From all of the above analysis there were certain factors such as resting blood pressure, chest pain, abruptions in the ECG and exercise induced angina that surfaced out to be the predisposing factors for heart diseases. I wanted to train a machine learning model on this dataframe so as to predict the prevalence of heart dieases in an individual on the basis of these factors and that which factors play a major role in it. For this approach I isolated numerical data from the original dataframe(heart_data) as the training and test data for the model and I isolated the categorical column heart_disease as the label for the training and test data. After splitting the data and labels into training and test category I used a random forest classifier to create and model which I tuned using the training data and training label. I then generated the predicted labels and matched them with the test labels to analyse the accuracy of the model. Art this point, the major challenge was to achieve a good enough accuracy to generate a useful feature importance plot. I played with the parameters of the classifier while tuning it with the training data and label to obtain the desired accuracy. After generating the feature importance plot. I generated the partial dependence plot using the model, test data, basic features and dependence of the desired individual feature on them.

ANALYSIS:

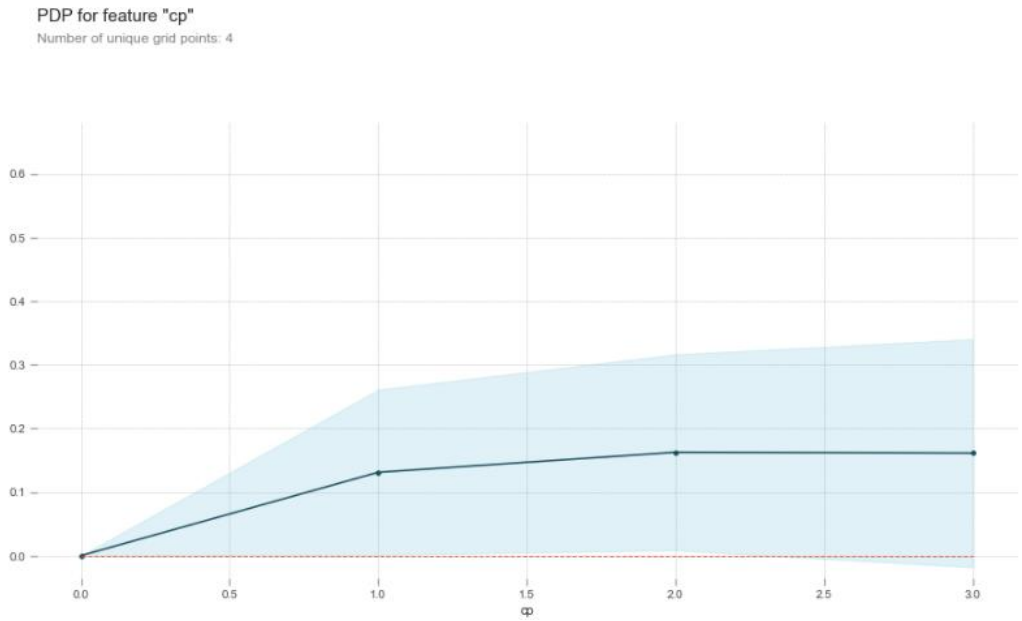


From the above feature importance graph generated bases on the results of our model (accuracy=89%) we can see that the features ca (number of major vessels recorded upon fluoroscopy), cp (type of chest pain), trestbps (resting blood pressure) and thalach (maximum heart rate achieved) are of greater importance in comparison to the others. Therefore, we may conclude that for an individual to either have or not have heart diseases depends greater on these factors as compared to the others.

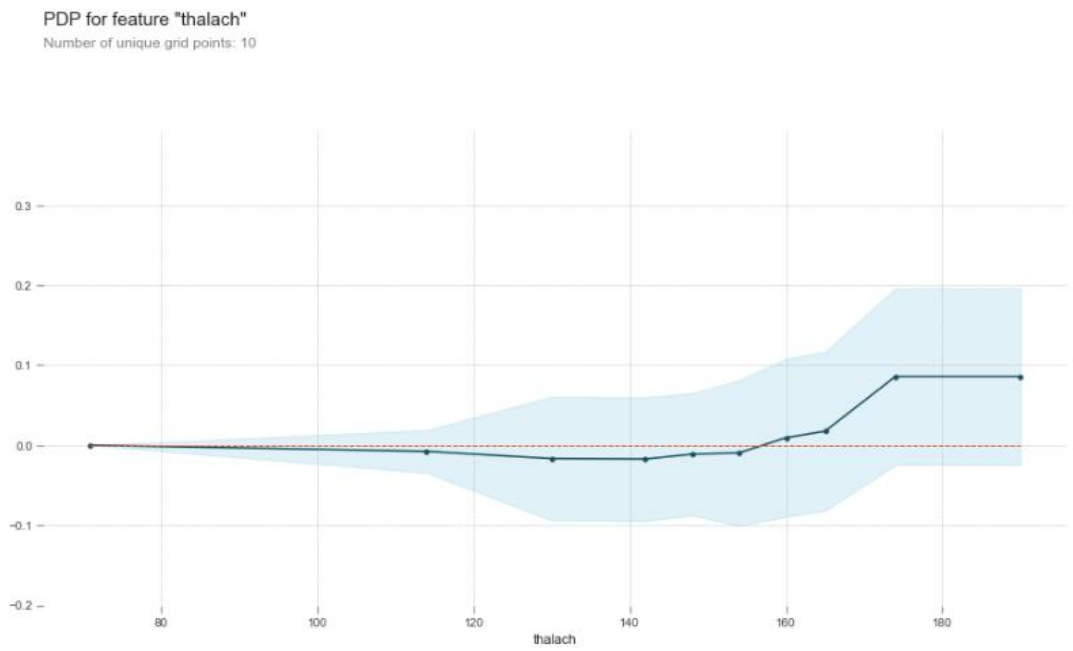
Reference for graphs: https://www.kaggle.com/tentotheminus9/what-causes-heart-disease-explaining-the-model



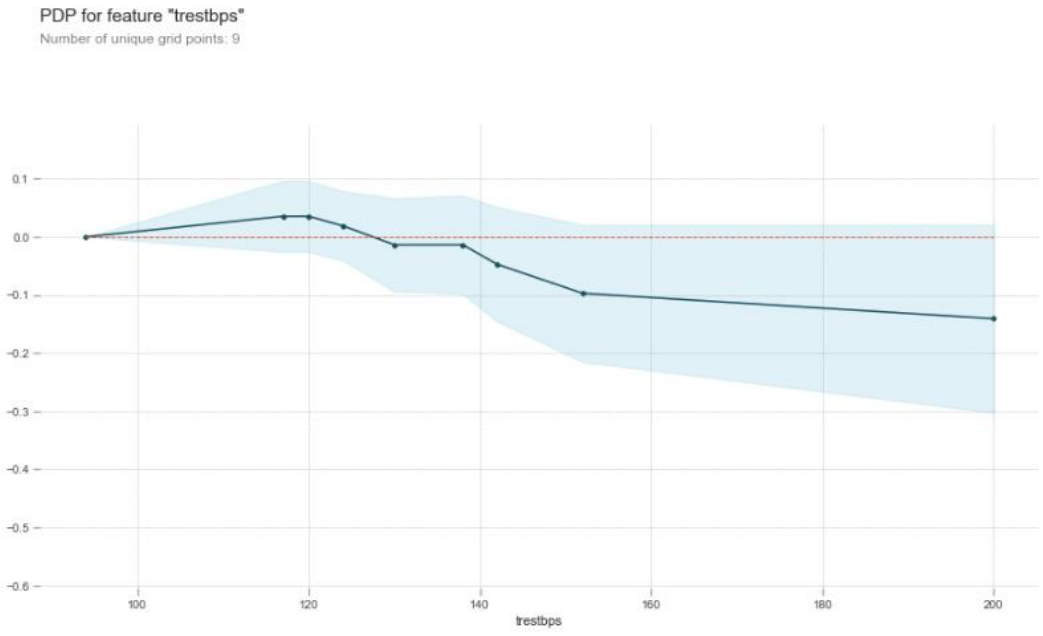
In order to achieve better insights of the results from feature importance graph we can look into the partial dependence plot. The above plot shows that the greater the number of major blood vessels identified during fluoroscopy the lesser are the chances of having a heart disease (negative y-axis value). This indicates absence blockage of the majority of major blood vessels and hence they can be spotted from fluoroscopy results. Prevalence of blockage in the major blood vessels means lack of major blood vessels being identified during fluoroscopy. Therefore, such blockages of blood vessels leads to heart disease. Hence, this factor is an important feature for the presence or absence of heart disease. The dotted red line on the graph represents absence of disease.



The above plot shows that, for a reading of cp zero and above the probability of having heart diseases are positive (y-axis values greater than null). The dotted red line on the graph represents absence of disease. Therefore, presence of any type of anginal chest pain is an indicator of heart diseases.



From the above plot, we can observe that if the maximum heart rate achieved during exercise (thalach) is more than 150 is a strong indication of having a heart diseases. For heart rates between 70-140 during exercise are considered normal and there is almost absence of heart diseases. Hence, we may conclude that maximum heart rate recorded during exercise does play an importance in deciding whether or not an individual has heart diseases.



From the above plot, it is surprising to notice that for higher readings of resting blood pressure is it not necessary for an individual to be having a heart diseases. Although, it can be seen that for resting blood pressure above 120 mmHg give a positive probability of having a heart disease the successive reading gives a negative probability. Anyhow, given that the model is 89% accurate we may conclude that having a high resting blood pressure may or may not predispose to heart dieases.