Model Building and EDA to Predict a Possible Heart Attack

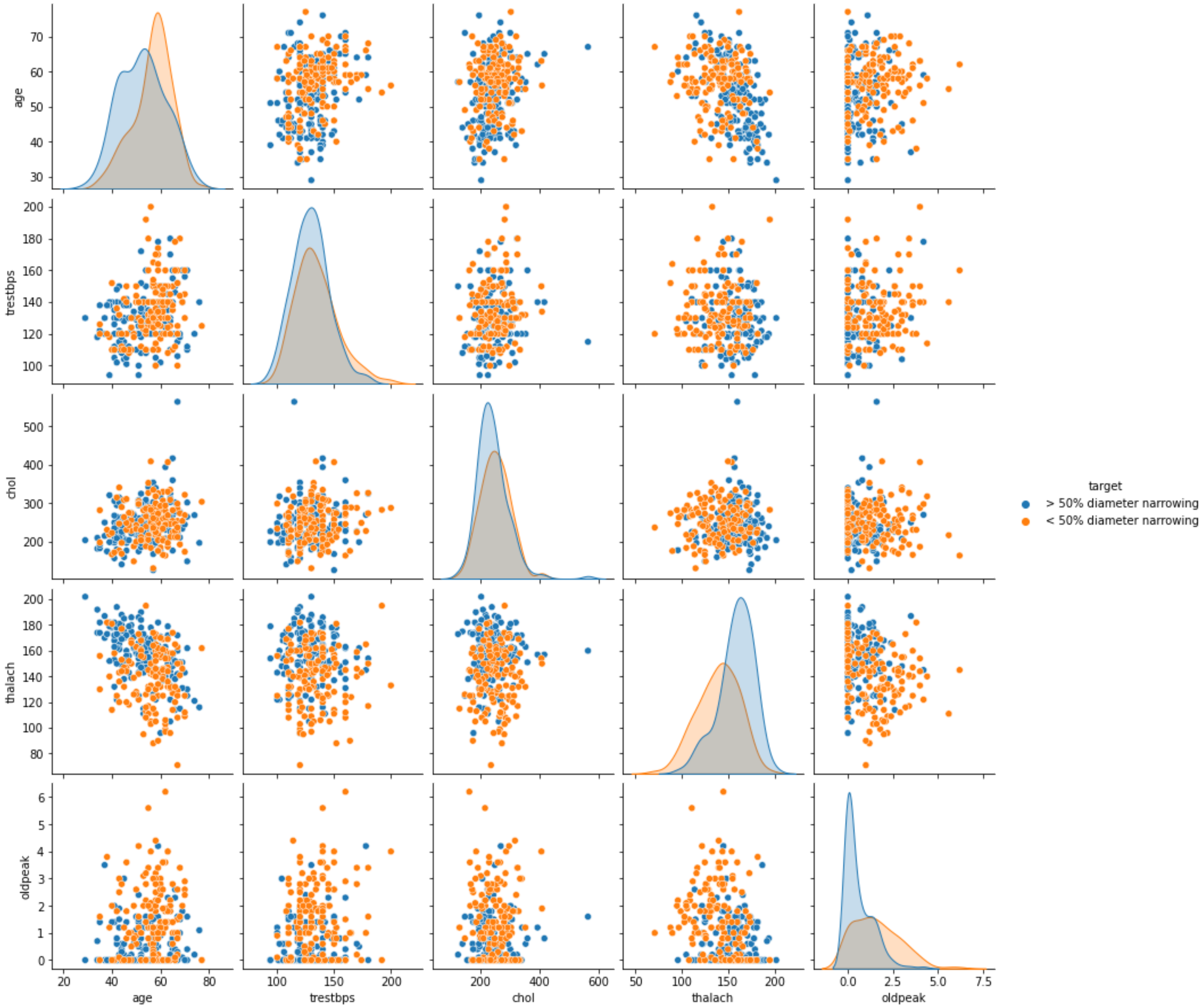
Description of Columns

- **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/d**
- **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: 3 = normal; 6 = fixed defect; 7 = reversable defect**
- **target: diagnosis of heart disease (angiographic disease status)**
- **-- Value 0: < 50% diameter narrowing**
- **-- Value 1: > 50% diameter narrowing**

Visualising Numeric Variables

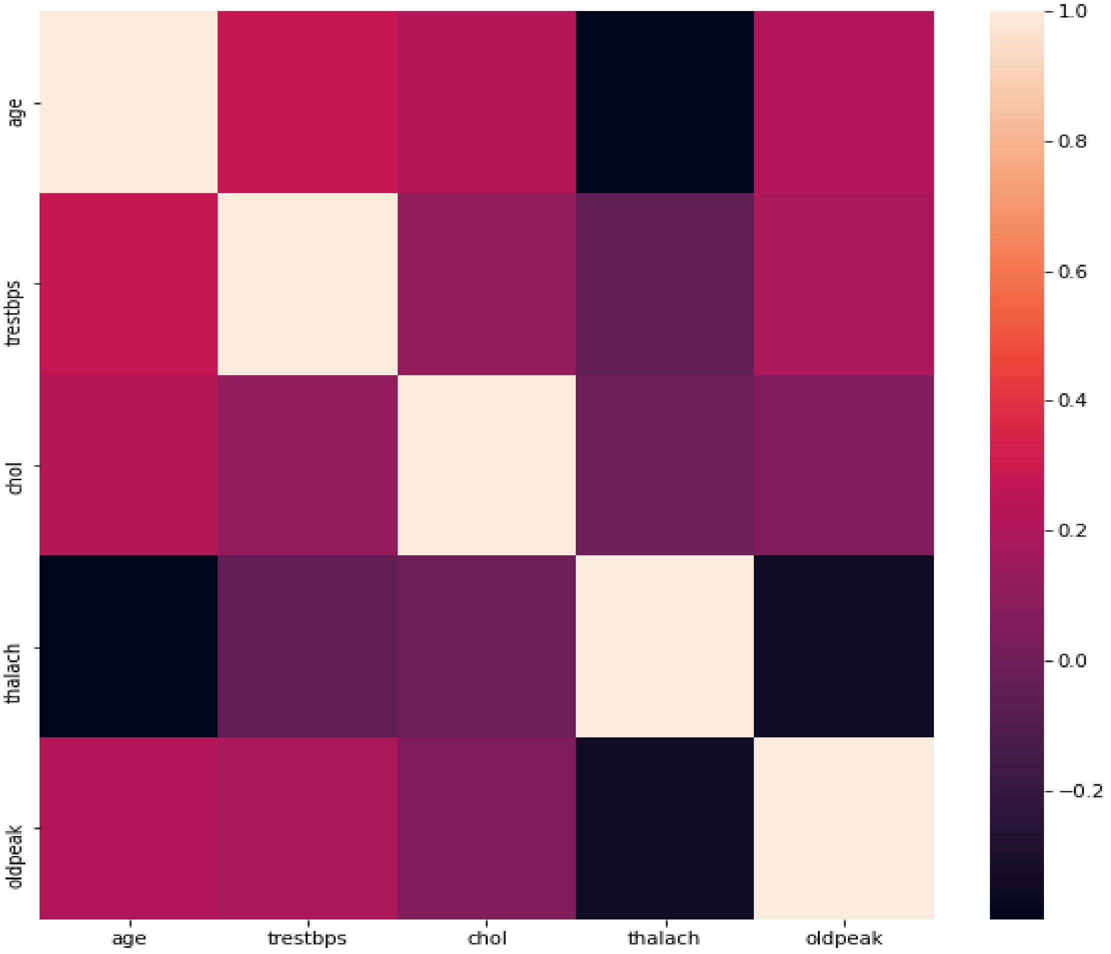
Let's make a pairplot of all the numeric variables

- **Pair plot is used to understand the best set of features to explain a relationship between two variables or to form the most separated clusters.**



Let's make a heatmap of all the numeric variables

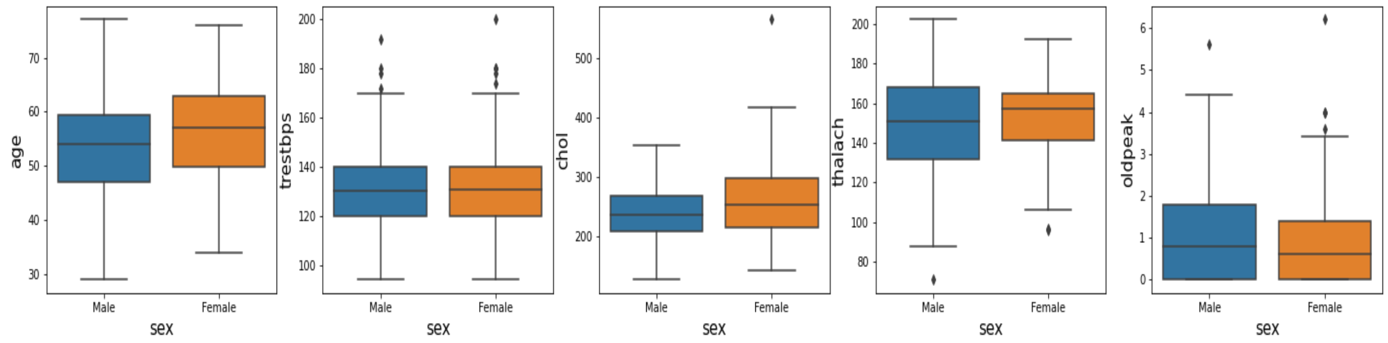
- Heatmap is also used to understand the best set of features to explain a relationship between two variables .
- It is usually not used that much in drawing any final conclusions but a great way to visualize the relation between features by looking at the different color representation of them.
- The heat map shows the relative intensity of values captured by your eye tracker by assigning each value a color representation.

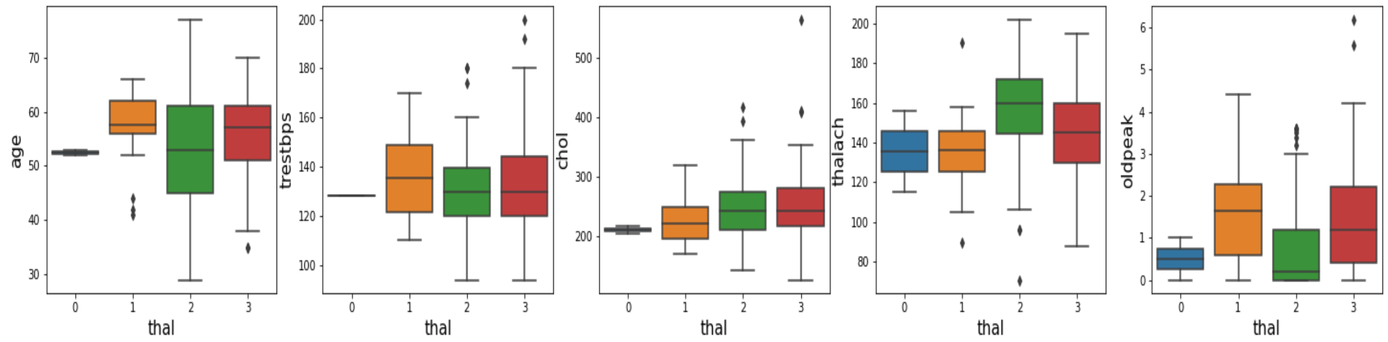


Visualising Categorical Variables

As you might have noticed, there are a few categorical variables as well. Let's make a boxplot for some of these variables.

- Outliers(for a normal distribution).
- “Minimum” and “Maximum” of different categories in a boxplot.





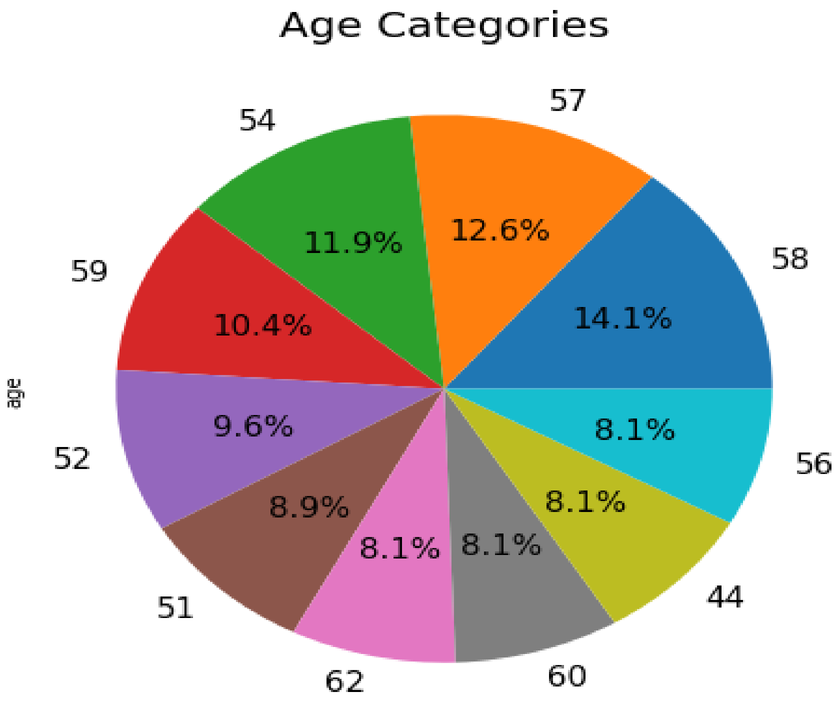
Observations:

By observing boxplots of sex against different features we get,

- Maximum and minimum of age in male is greater than and less than maximum and minimum of age in female respectively.It means that this dataset contains more number of younger and older males than females of respective category.
- Resting Blood Pressure is same in both sexes but also both male and female data of resting blood pressure contains outliers.
- **Cholestrol level** is slightly **more in female** compared to male.
- Inter Quartile Range of Maximum heart rate achieved for **male is between 140 and 170** whereas for **female it is between 140 and 160**.

By observing boxplots of thal(a blood disorder called thalassemia) against different features we get to know that

- Thal **value of 2**(**normal blood flow**) has lowest minimum and highest maximum which denotes its relationship against age.It has biggest IQR when compared to other thal values quartiles. Most common Thal value for all age groups is thal value 2(normal blood flow).
- Thal **value of 1**(**fixed defect** - no blood flow in some part of the heart) against age is between **50 and 70 years old.**
- Thal **value of 3**(**reversible defect** - a blood flow is observed but it is not normal) against age have its median at **58 years** old approximately.
- In **thal vs chol** plot we can see that the medians of cholestrol levels are increasing as thal values are increasing which shows that cholestrol level and thalessemia are positively correlated.
- In **thal vs thallach (Maximum heart rate achieved)** plot we can see that **median of thal value 2 is greatest** which denotes that even after having normal blood flow the range of maximum heart reate achieved is greater than other thal values.

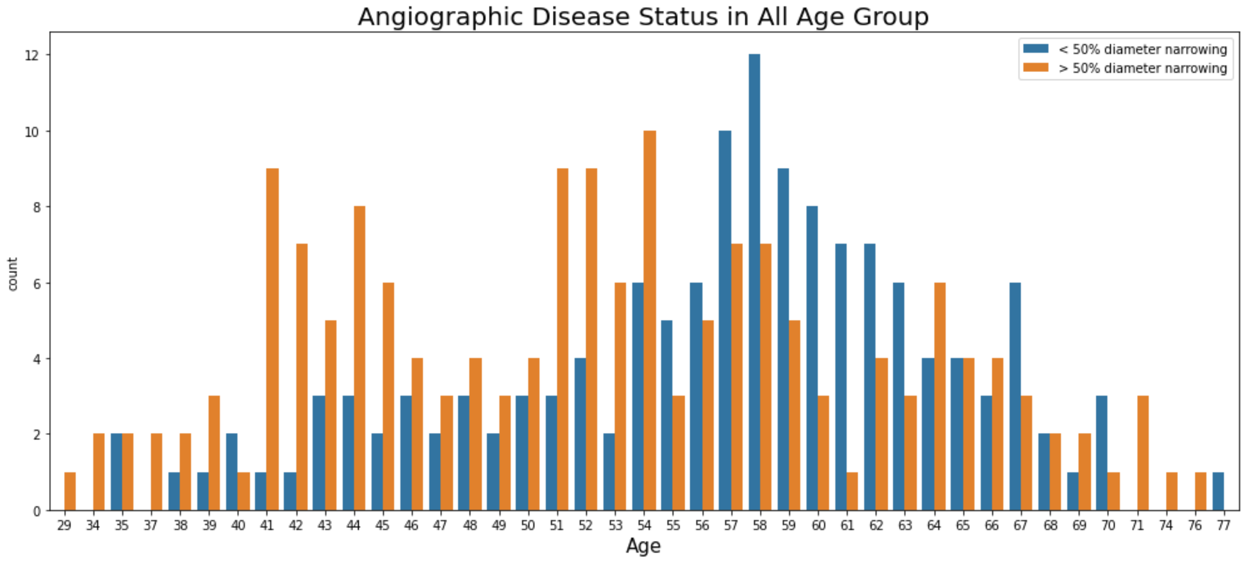


Observations:

- Maximum records belong to the age of **58 years old** in dataset with **14.1%.**
- We can see from the pie plot that the person present in the dataset mostly belongs to age group between **50 - 60 years old.**
- **However it is a very small dataset with very limited amount of observations so we can't rely on age as a deciding factor.**

Let's visualize more about some important relation between features and target variable

Histogram plot

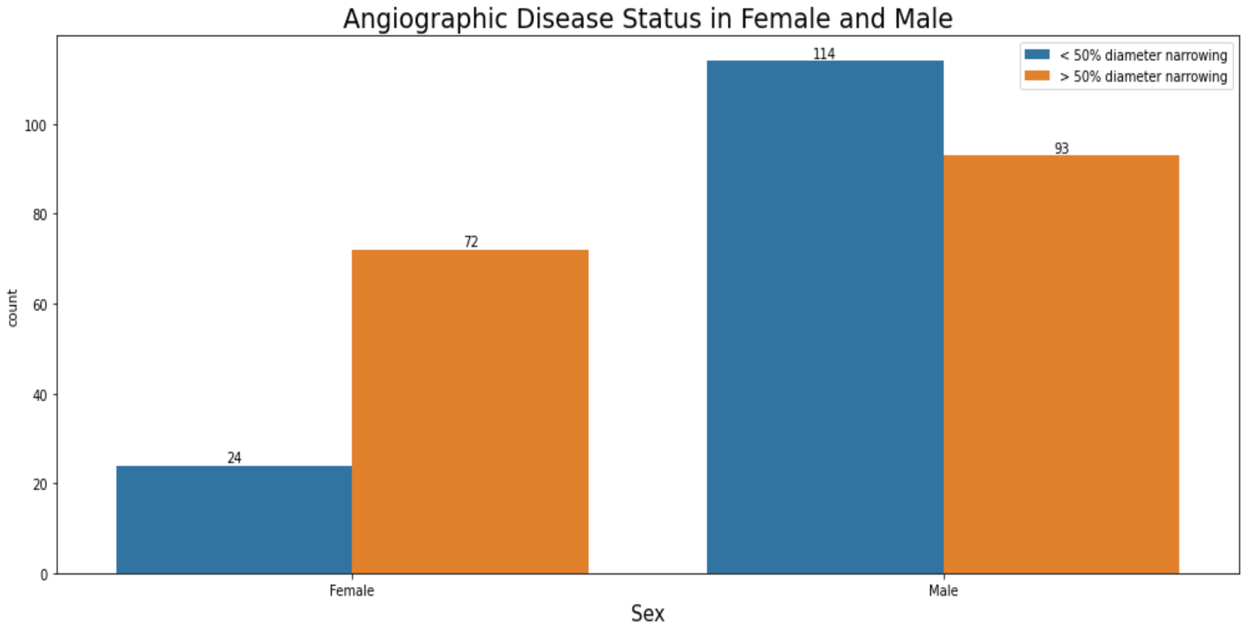


Observations:¶

According to dataset if we try to draw some conclusions on the basis of this plot then we conclude that,

- The orange bars shows that the person is at higher risk getting a heart attack and the blue bars shows that person is at less risk.
- The **age group 51-60 years old** is at more risk.

- Due to lack of more observations and small dataset we can see that according to chart, young people are at more risk than old people which is not true.

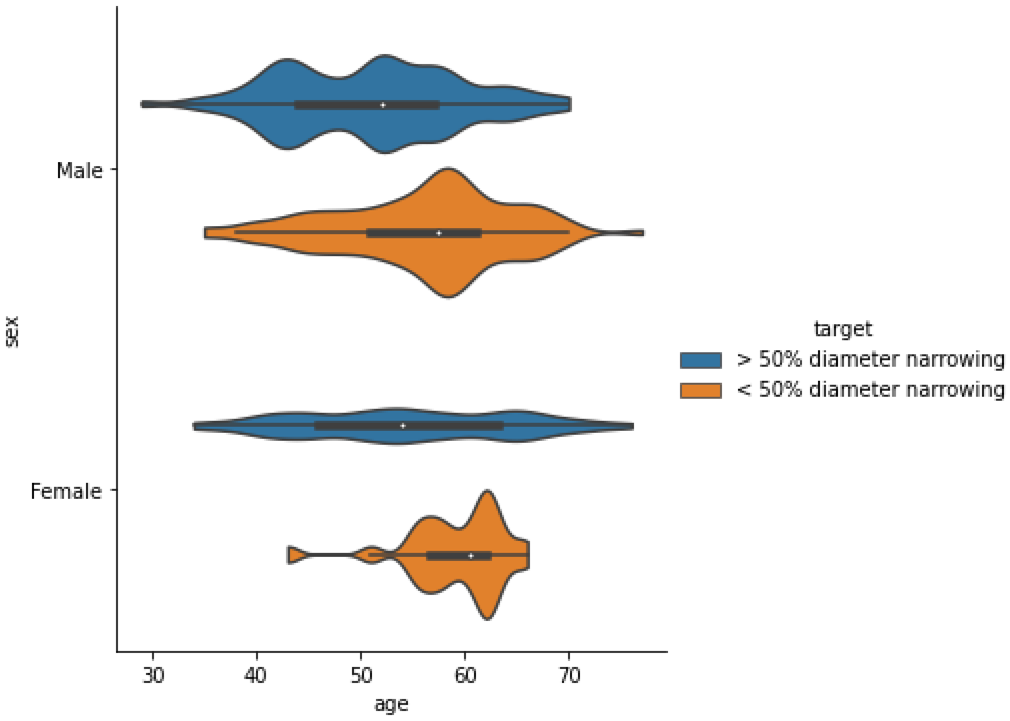


Observations

- *Such difference in male and female plot is due to less records of female present in dataset.*
- Ratio of more susceptible to a possible heart attack to less susceptible in **female** is **3:1**.
- Ratio of more susceptible to a possible heart attack to less susceptible in **male** is almost **3:4**.

Violin plot used for categorical data

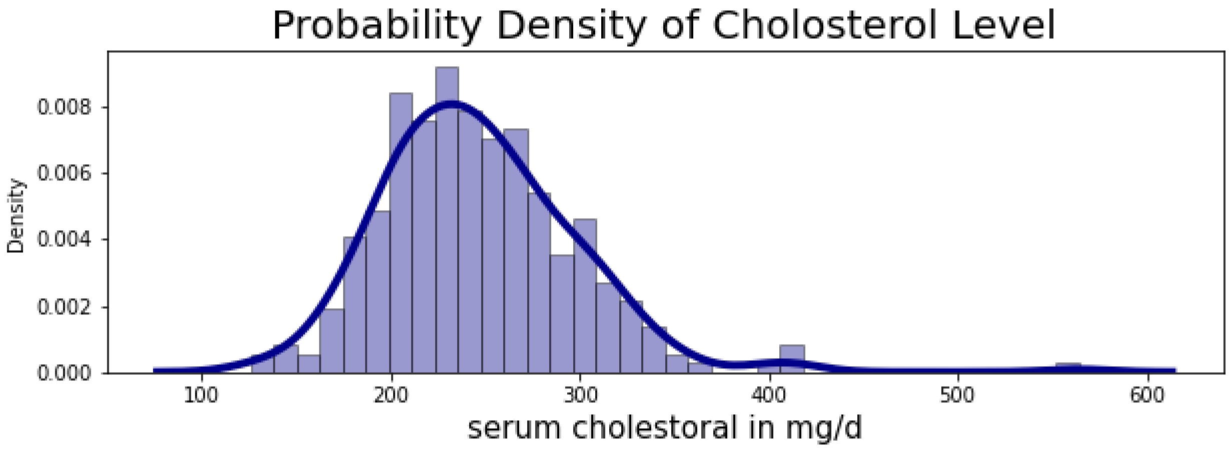
Analyzing relationship between age and sex with target variable in single plot

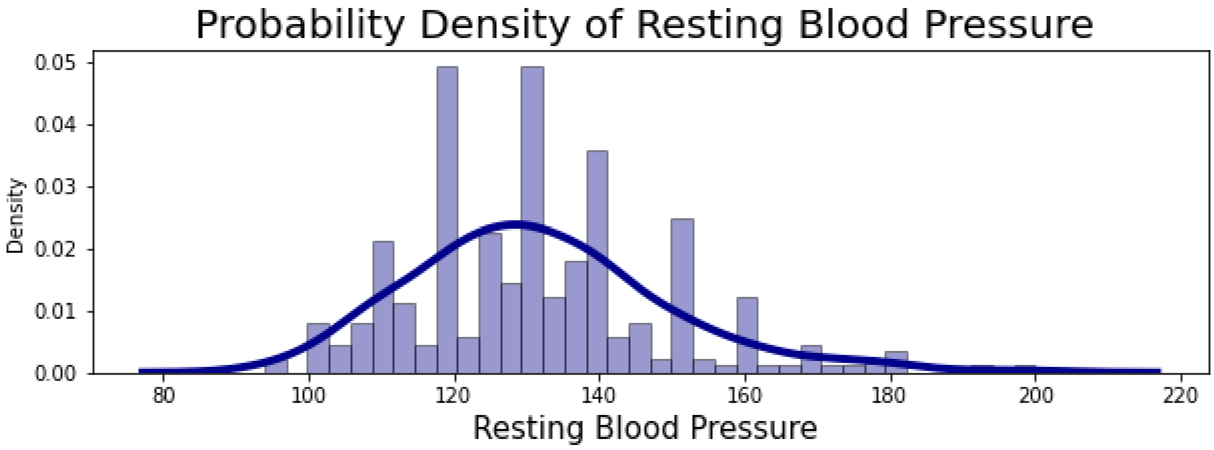


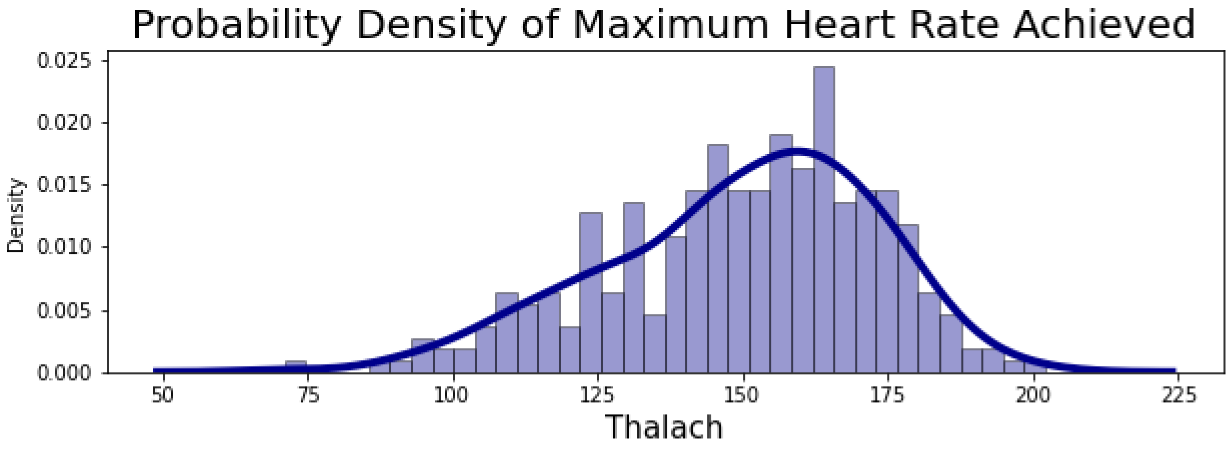
Observations:¶

- In plot of female, where target: > 50% diameter narrowing (more susceptible to a heart attack), the range of age is greater and evenly distributed with flatter peaks.
- Whereas in plot of female, where target: < 50% diameter narrowing (less susceptible to a heart attack), the range of age is shorter and has higher peaks at certain age groups near 55 years and 62 years.
- We can see there are very less outliers in dataset.

Probablity Density Plot of Some More Features


Probability Density of Cholosterol Level


Probability Density of Resting Blood Pressure

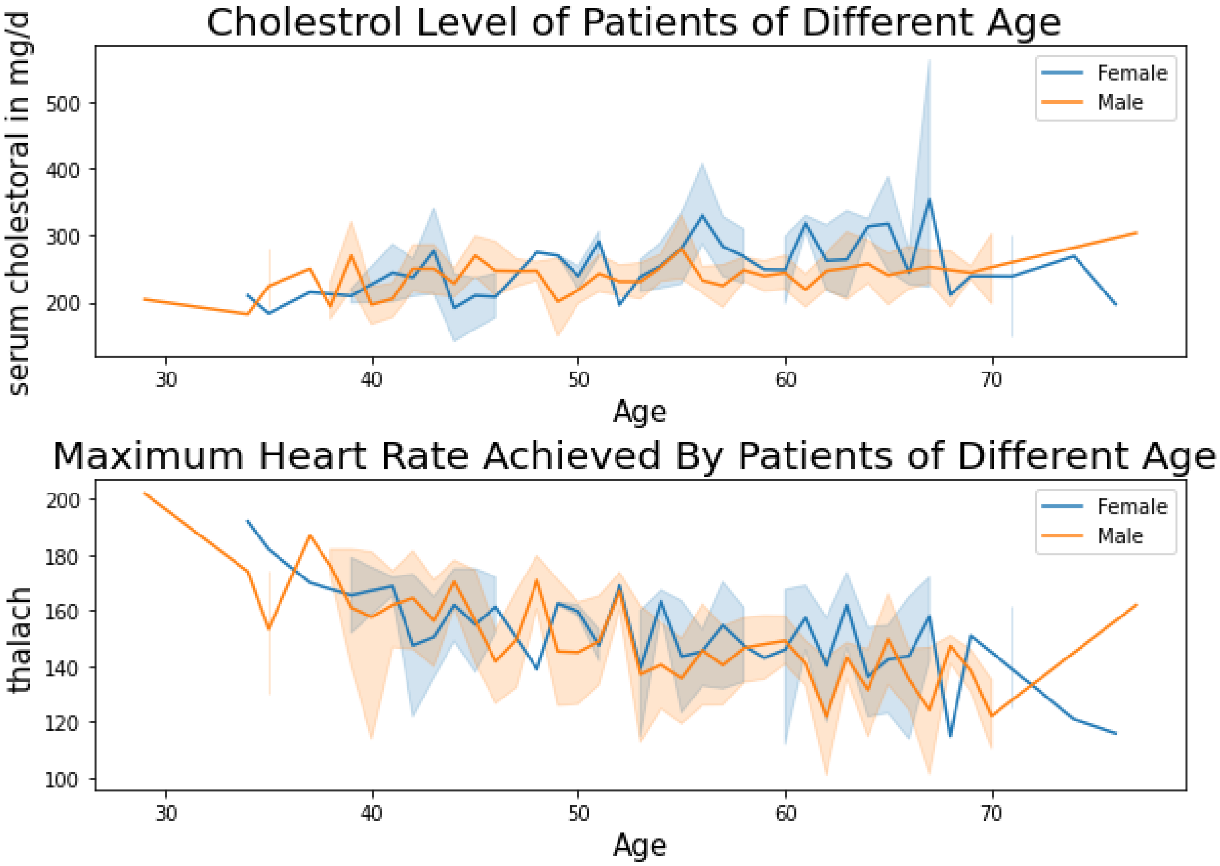

Probability Density of Maximum Heart Rate Achieved

Observations:

- Most common cholosterol level present in dataset is between **200 mg/d to 240 mg/d** with maximum value of **230 mg/d** which is considered **borderline to moderately elevated**.
- Few records of Thalach are at near **400 mg/d** and **550mg/d** which is very high and risky.
- Resting blood pressure at **120** and **125** have highest probablity density of **0.05**.
- Probablity density of **Thalach(maximum heart rate achieved)** is mostly evenly distributed between **140** and **175**.
- Highest probablity of **Thalach(maximum heart rate achieved)** is nearly at **165**.

Line plot of cholosterol level and Thalach in male and female with their ages

- Serum cholesterol levels helps in figuring out risk for developing heart disease.
- The maximum heart rate greater than MHR(= 220 - present age in years) is considerd hazardous



Observations:

- Serum cholestrol level in Female is higher than Male between the age of 52 years to 68 years.
- From the thalach plot we can see that as the age is increasing the chart is showing downward trend directing towards the fact that maximum heart rate achieved decreases as the person grows older.