**Predictive Analytics in Cardiology: Machine Learning Approaches to Forecasting Heart Failure**

**Introduction:**

Heart failure stands as a formidable challenge in the realm of public health, affecting millions worldwide and serving as a leading cause of hospitalization and morbidity. The ability to predict heart failure with accuracy is not merely an academic exercise but a pressing necessity that could revolutionize patient outcomes and healthcare strategies. This research paper delves into the application of predictive analytics within cardiology, focusing on the deployment of machine learning algorithms to forecast the risk of heart failure.

The advent of big data and advanced computational techniques has opened new horizons in medical diagnostics and prognostics. By utilising the power of machine learning, researchers and clinicians can go through through large amounts of patient data to identify patterns and predictors of heart failure. This paper aims to explore the efficacy of various machine learning models, assess their performance in clinical settings, and discuss the implications of their use in early intervention and personalized medicine.

Through a comprehensive analysis of existing datasets and a meticulous construction of predictive models, this study seeks to contribute to the burgeoning field of digital health. It endeavours to bridge the gap between data science and clinical expertise, offering insights that could lead to more timely and targeted treatments for those at risk of heart failure. As we stand on the cusp of a new era in healthcare, the fusion of data science and medicine holds the promise of enhancing the quality of life for patients across the globe, making the quest for accurate prediction models not only relevant but imperative.

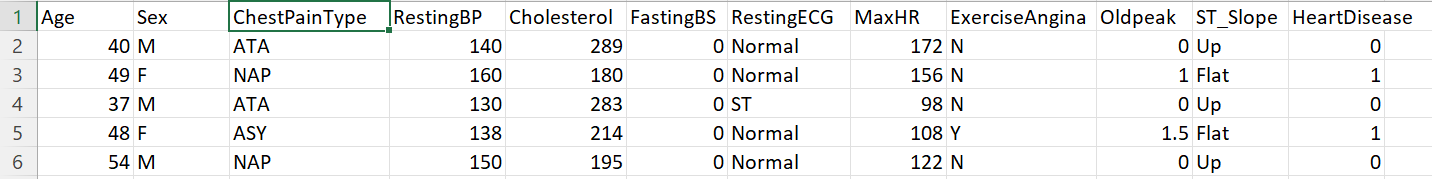
**What is our data?**

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Four out of 5CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs and this dataset contains 11 features that can be used to predict a possible heart disease.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

**Attribute Information -**

* Age: age of the patient [years]
* Sex: sex of the patient [M: Male, F: Female]
* ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
* RestingBP: resting blood pressure [mm Hg]
* Cholesterol: serum cholesterol [mm/dl]
* FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
* RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
* MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
* ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
* Oldpeak: oldpeak = ST [Numeric value measured in depression]
* ST\_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
* HeartDisease: output class [1: heart disease, 0: Normal]



This dataset was created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes. The five datasets used for its curation are:

* Cleveland: 303 observations
* Hungarian: 294 observations
* Switzerland: 123 observations
* Long Beach VA: 200 observations
* Stalog (Heart) Data Set: 270 observations

Total: 1190 observations  
Duplicated: 272 observations  
Final dataset: 918 observations

**Identifying Patterns in Heart Failure:**

Before we begin, we must find out which variables in our dataset show correlations so that we can narrow down upon which factors are the most important in developing our model. Upon running the program, we can see this plot:

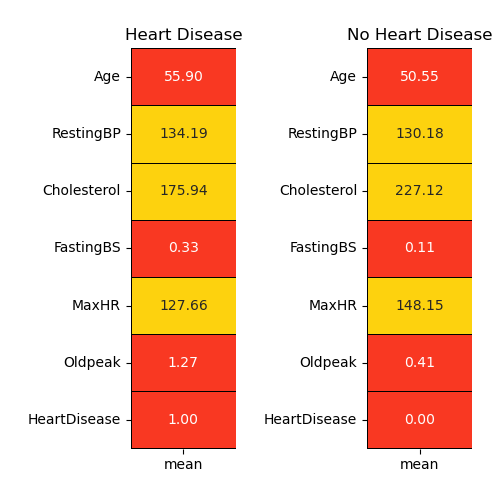
A screenshot of a graph

AI-generated content may be incorrect.

From the above diagram, we can make 3 conclusions:

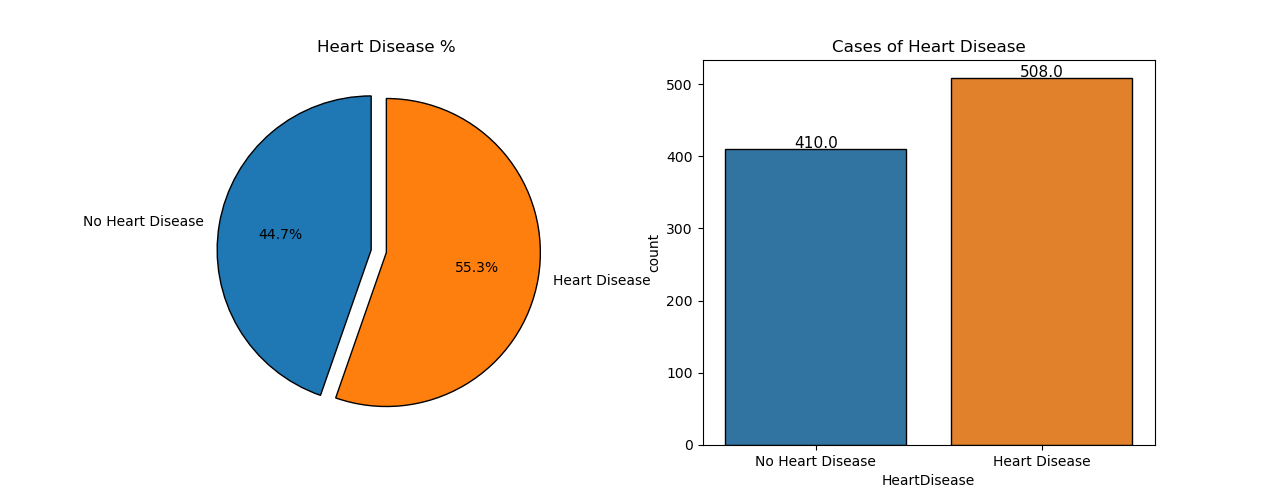
1. Age, Fasting Blood Sugar, and Old age are 3 factors that positively correlate to Heart Disease
2. Serum Cholesterol, and Maximum Heart Rate are 2 factors that negatively correlate to Heart Disease.
3. Resting Blood Pressure shows little to no correlation at all.

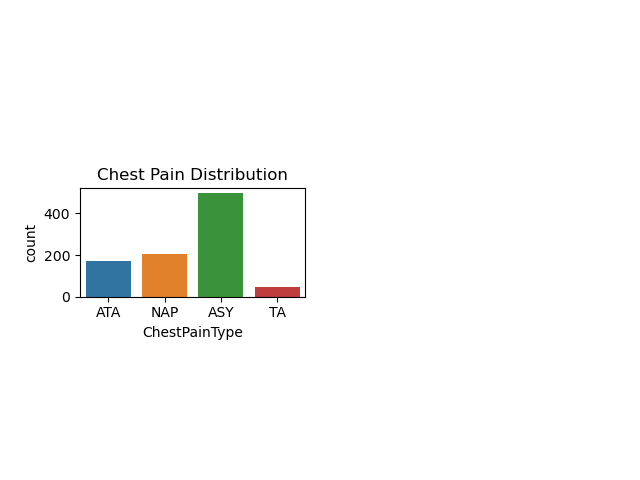
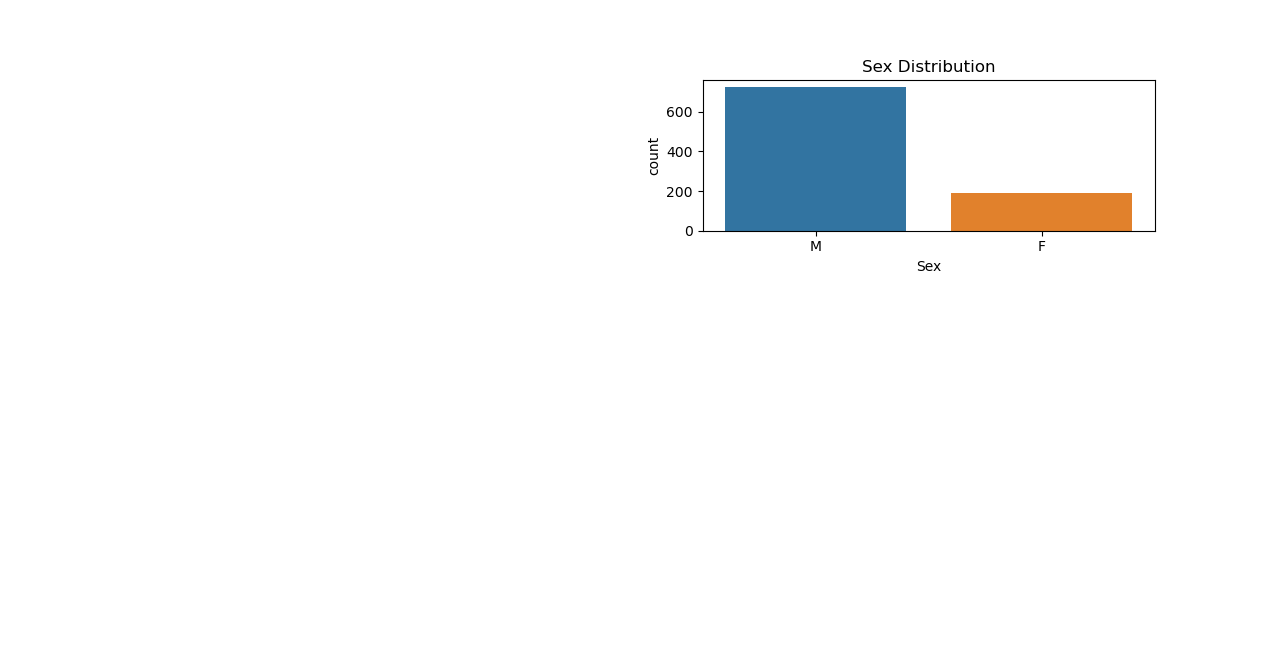
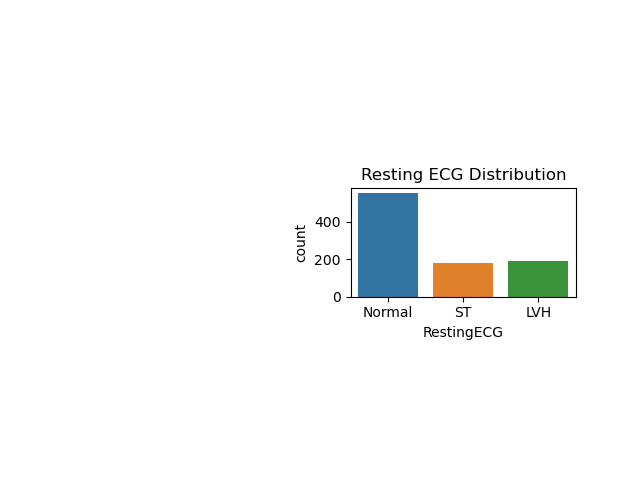
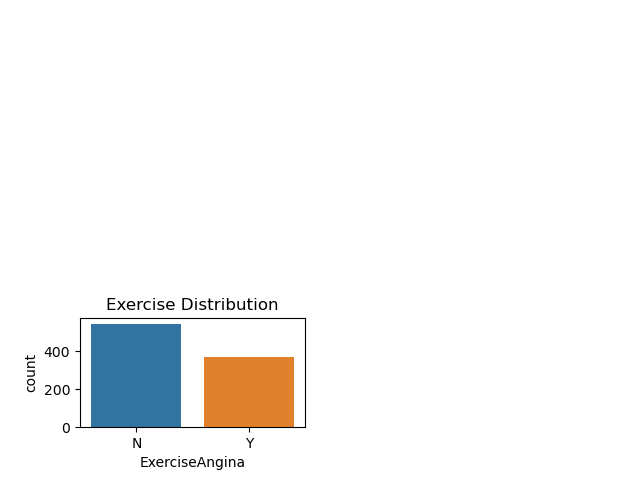
Now, let us take a look at the mean of each of these factors in order to verify these conclusions:



As expected, we can see that those with heart disease versus those without are elder, have higher fasting blood sugar and higher old peak. On the other hand, maximum heart rate and cholesterol are significantly lower for those with heart disease. Resting blood sugar can be ignored in this case as there is minimal difference between the 2 groups.

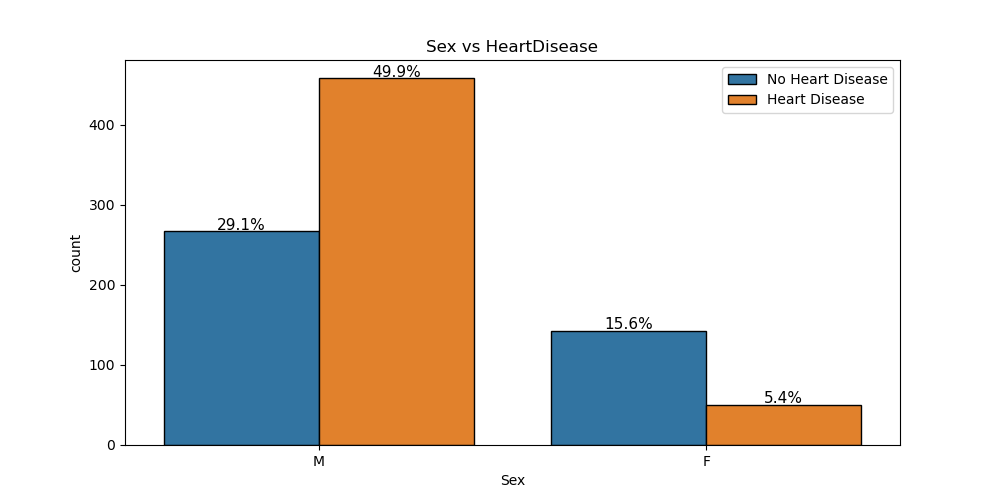
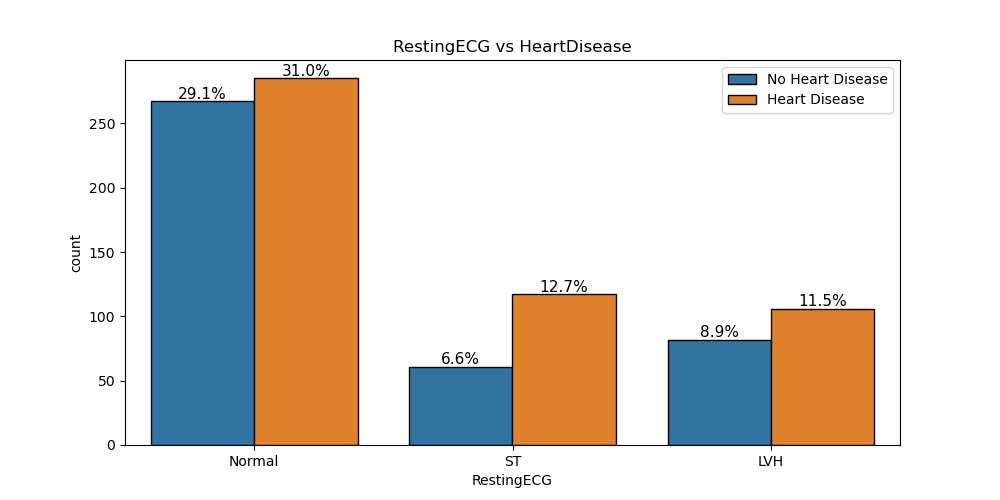
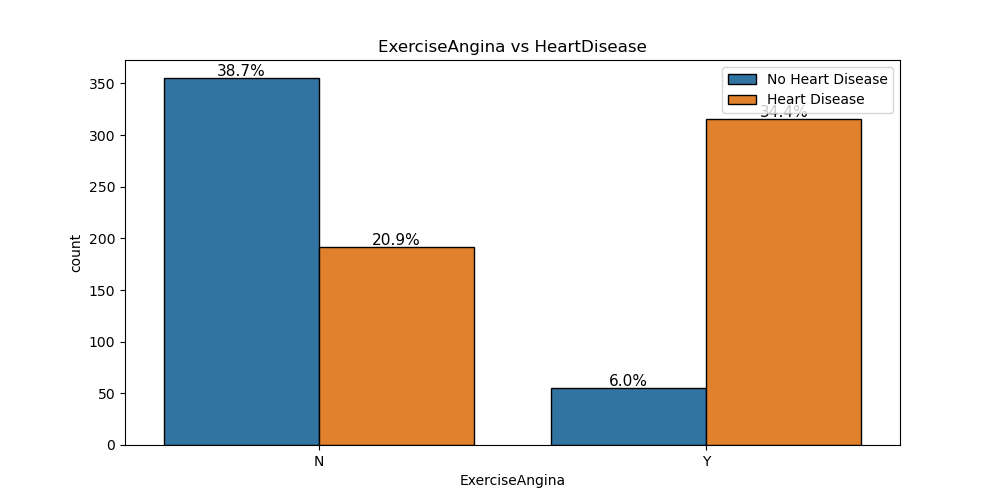
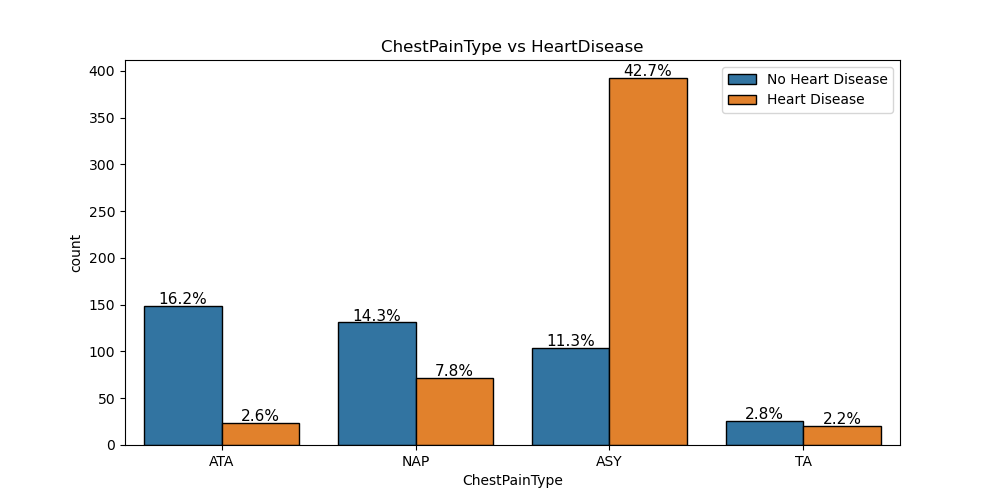
We have now managed to find correlations in the numerical data available in our dataset. However, how do we find correlations in the categorical data? We can do this by first determining the distribution of people with heart disease and without in our dataset as shown below:



As we can see above, there is nearly equal distribution of people with and without heart disease in our dataset. Based on this information, we can take a look at the data distribution of each of our categories and make a few assumptions:

1. Having a chest pain type of ASY can be an indicator of heart disease
2. Heart disease if more likely among males compared to females
3. Any kind of resting electrocardiograph that is not normal may suggest heart disease.
4. Having no exercise angina can be a predictor of heart disease as well

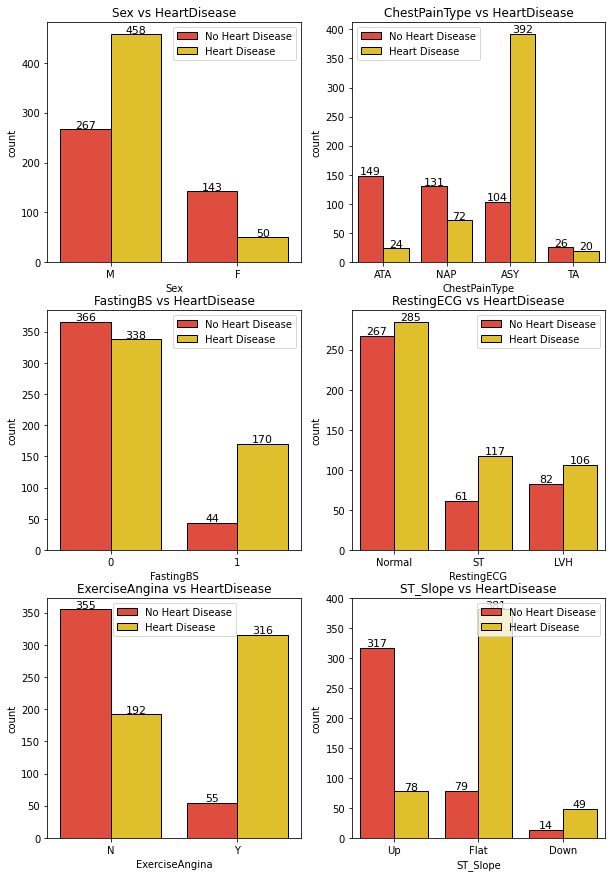
However, we cannot simply proceed based on these assumptions. Therefore, to confirm these hypotheses, we will take a look at the categorical data distribution as well:



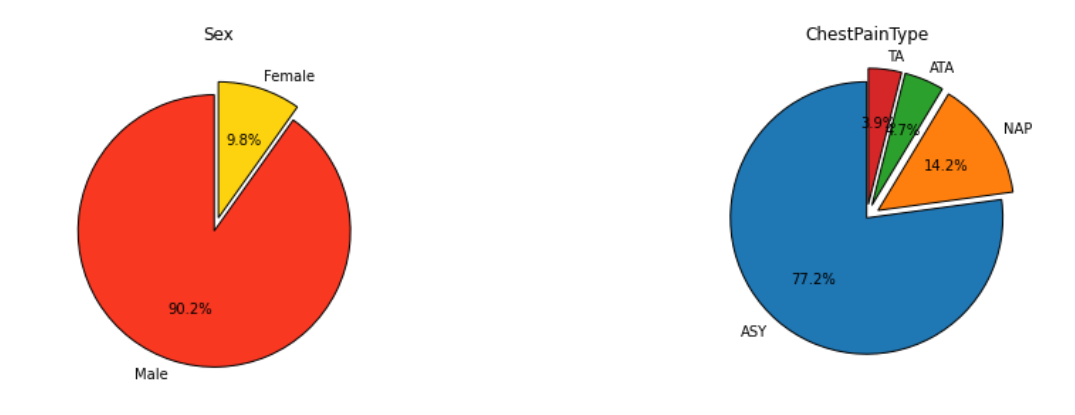
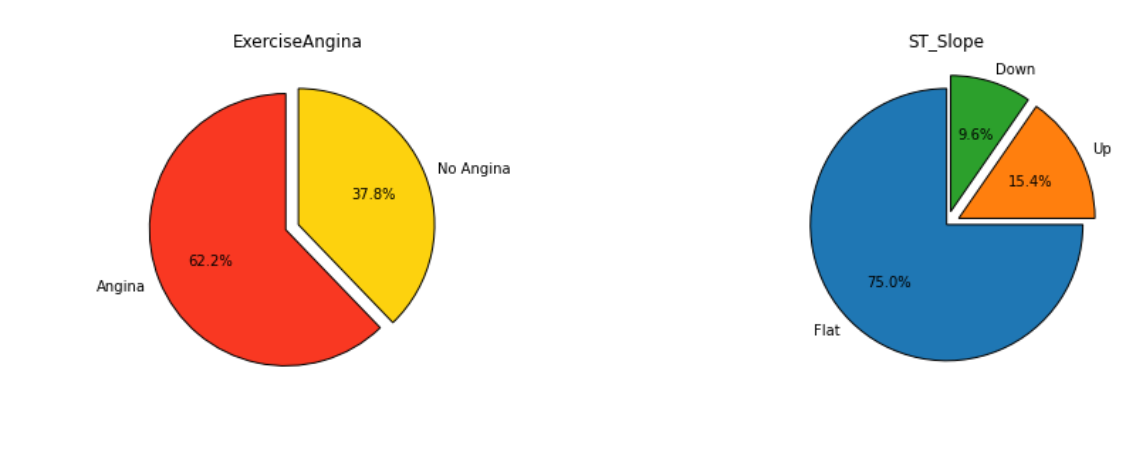
Now, based on the above distributions of people with and without heart disease in each of the categories, including Fasting Blood Sugar, we can compare them with our previous assumptions to reach some conclusions:

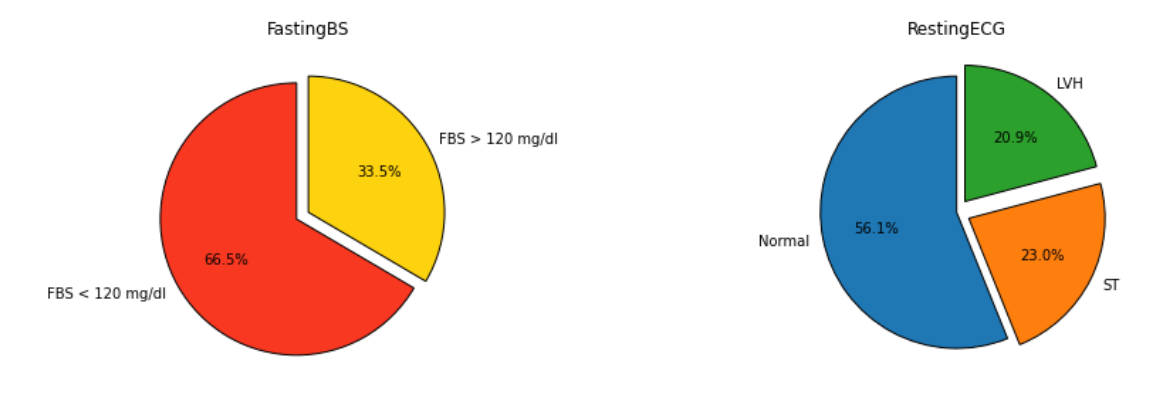
* As predicted previously, heart is much more likely among males than females.
* ASY type of chest pain boldly points towards major chances of heart disease.
* Resting ECG does not present with a clear-cut category that highlights heart disease patients. All the 3 values consist of high number of heart disease patients. However, we can notice that the number of heart disease patients for ST is nearly double compared to the number of patients without heart disease. On the other hand, the percentages for the other 2 categories are nearly similar.
* Exercise Induced Engina definitely bumps the probability of being diagnosed with heart disease as shown by the graph.

Let’s finally map out the distribution for all of our categories in terms of those that have heart disease and those who don’t to verify all the conclusions made about our data so far:



As we can see above, all of our conclusions map out appropriately with the data distribution of the given dataset.

However, for further confirmation, let us see the comparison of each factor with the positive heart disease cases:

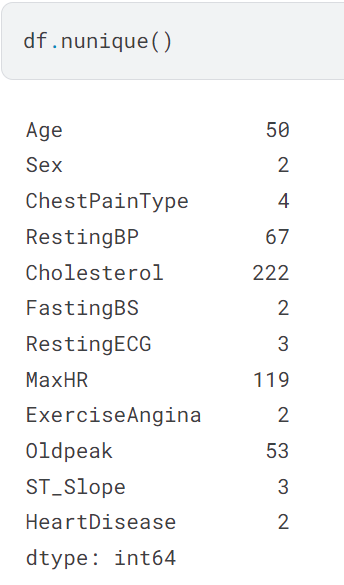
We can observe the following results:

* Heart disease is more common in males – 90% of the patient are males
* ASY Chest pain type is a strong indicator – 77% of the patients had ASY chest pain
* Resting ECG of ST may be likely for heart disease – 56% of patients were normal
* Yes for Exercise induced Angina is common and another indicator – 62% of patients said yes

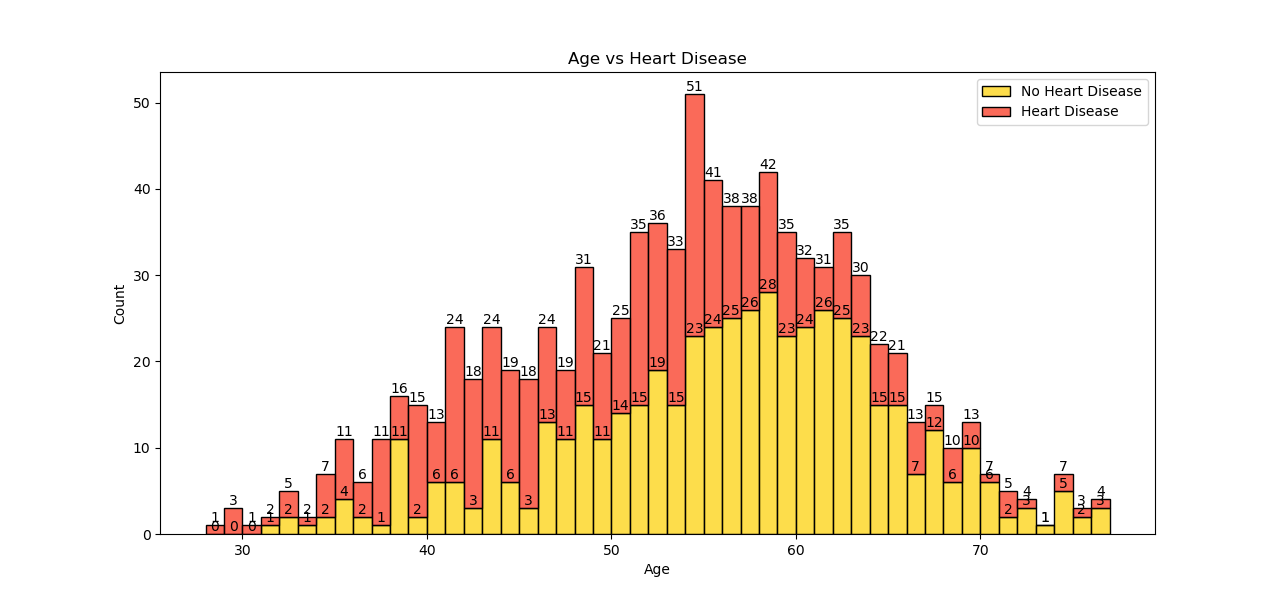
Additional observations we have made from this data are:

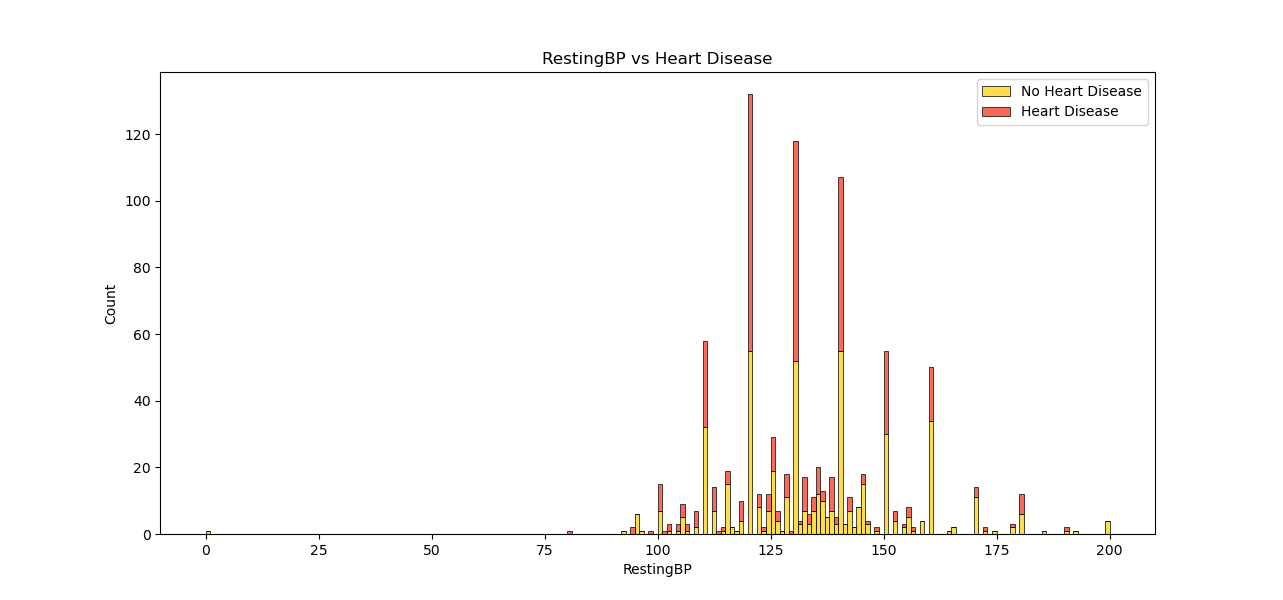
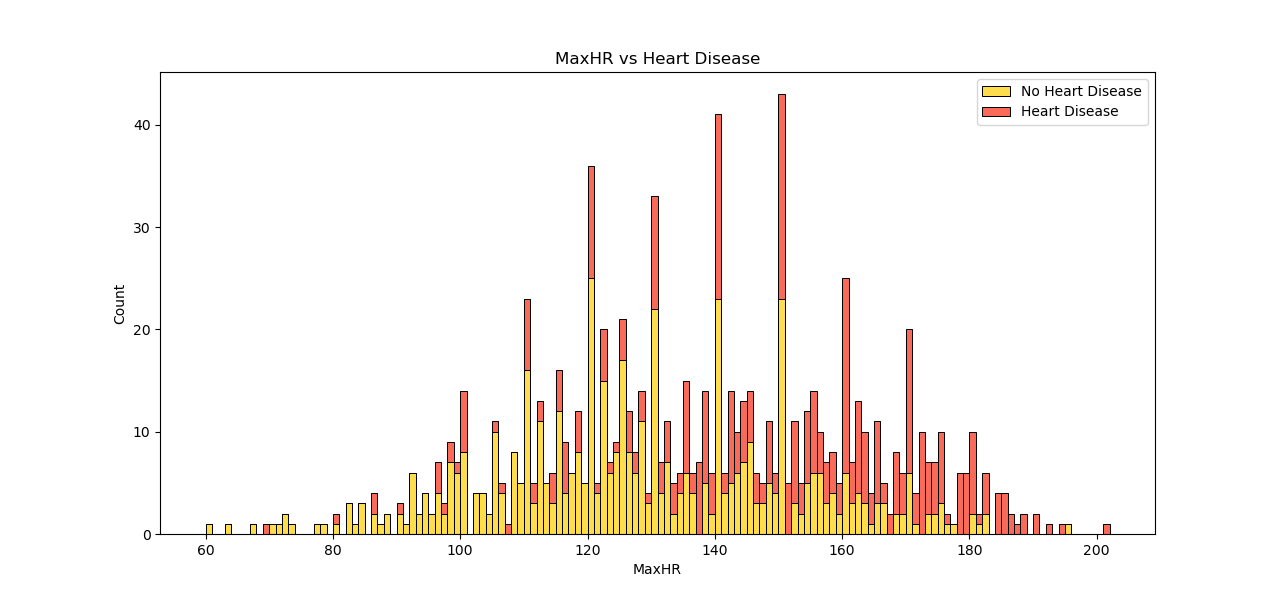
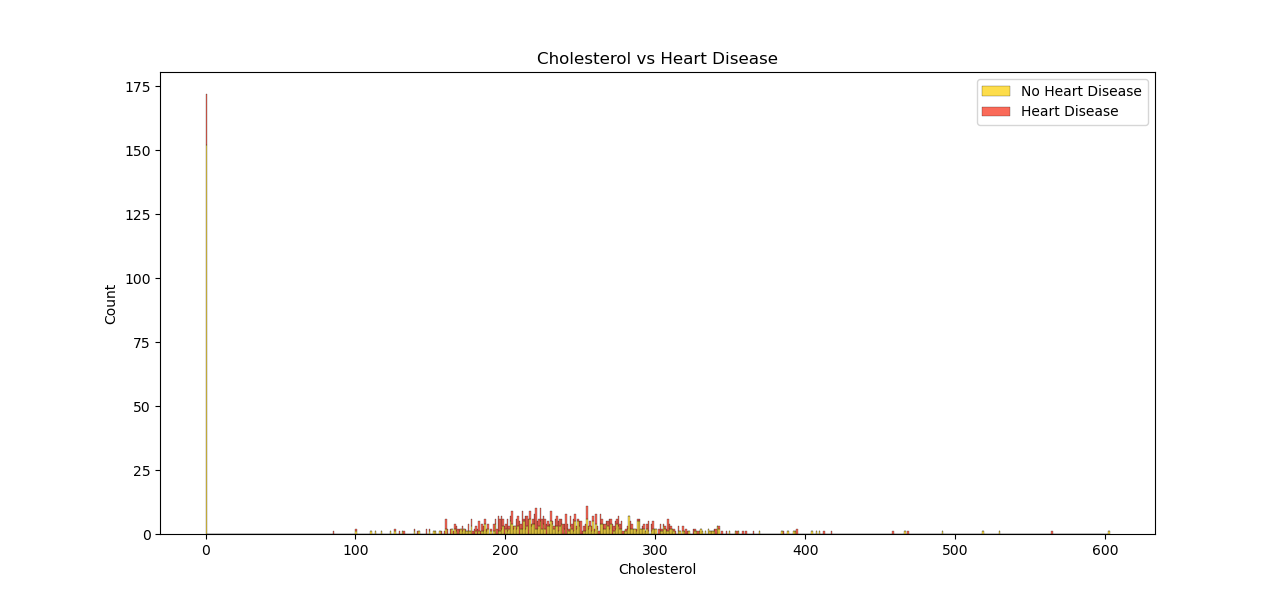
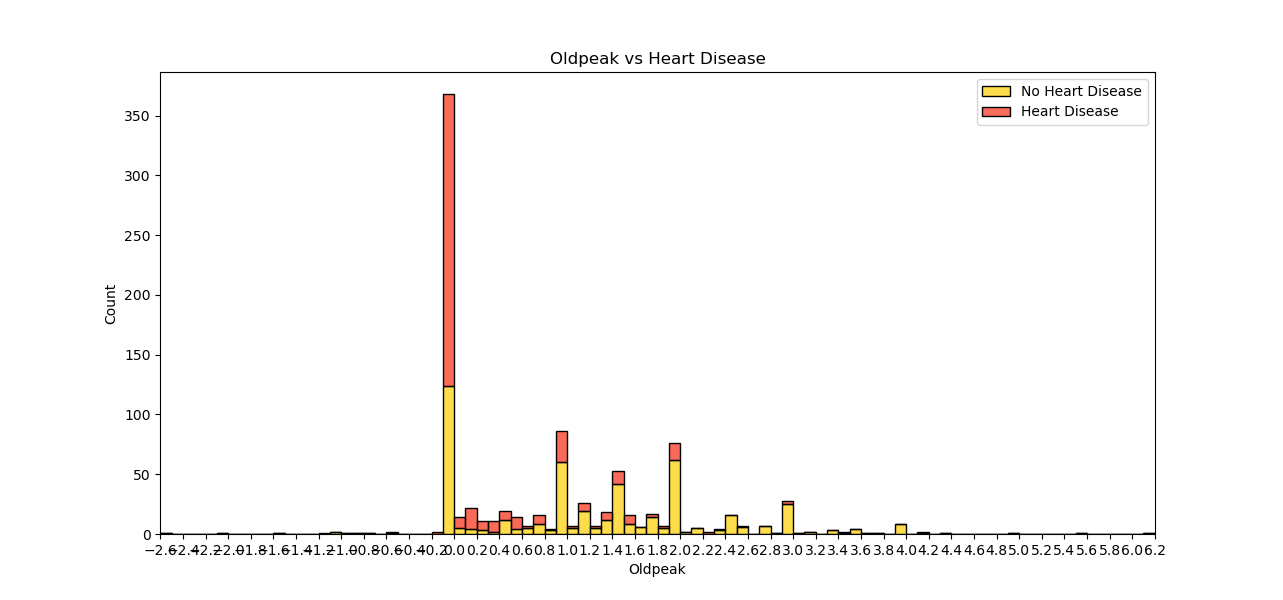
* Fasting Blood Sugar below 120 mg/dl is an indicator as shown by nearly 2/3rd of patients
* ST-slope of flat is an indicator as shown by 3/4th of the patients.

These observations give us a better idea of what we’re specifically looking for in order to successfully diagnose a patient.

**Numerical Features vs Target Variable (Heart Disease):**

This data shows us the number of unique values in each category. For numerical categories such as Age, RestingBP, etc; the more unique values we have the better. This helps provide us a general overview of how heart disease might vary over a range. Thus, more unique values means a larger range allowing for greater accuracy in our data.

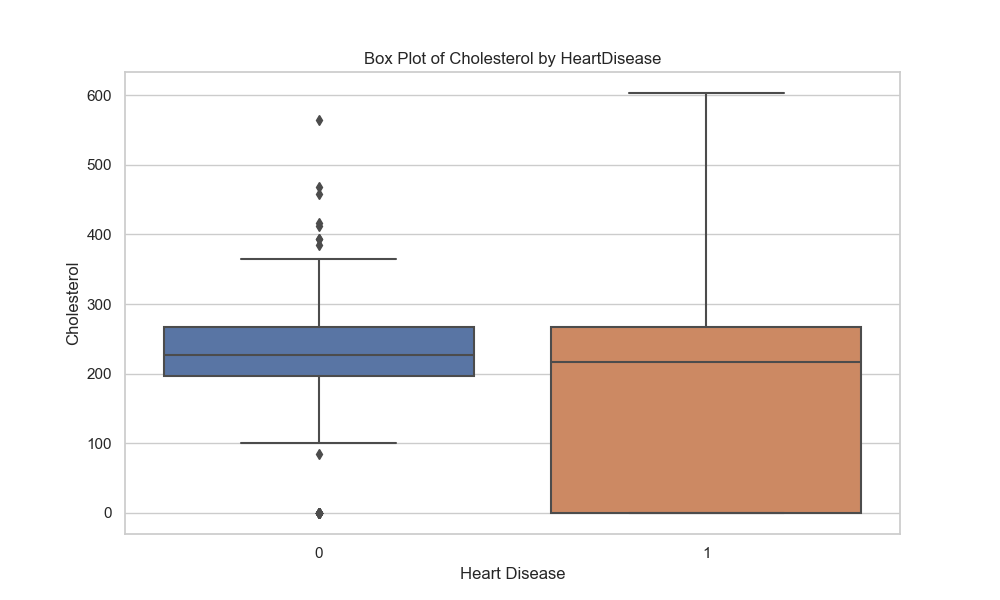
**-**

****

A chart with a red and yellow box

AI-generated content may be incorrect.We mentioned before that having several unique values allows us to explore a range of data for a better overview. However, we have encountered a unique problem here. Due to the presence of too many data values, our bar graphs are spread too far apart. This makes it difficult to derive any sort of valid information. Thus, we can represent these values in boxplots instead to understand where most of our data lies.

A chart of a box with a red box

AI-generated content may be incorrect.A chart with a red and yellow box

AI-generated content may be incorrect.A graph of a diagram

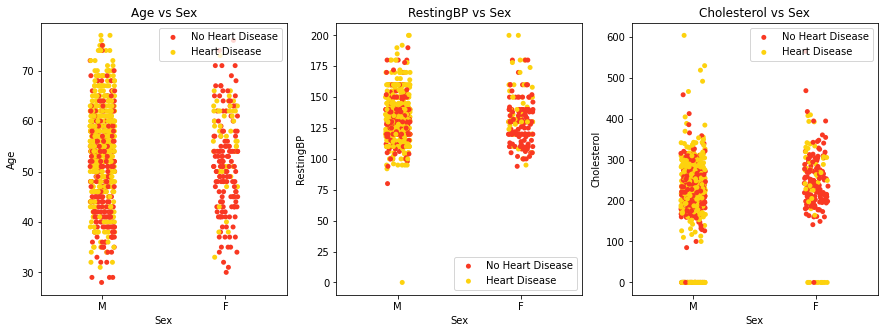
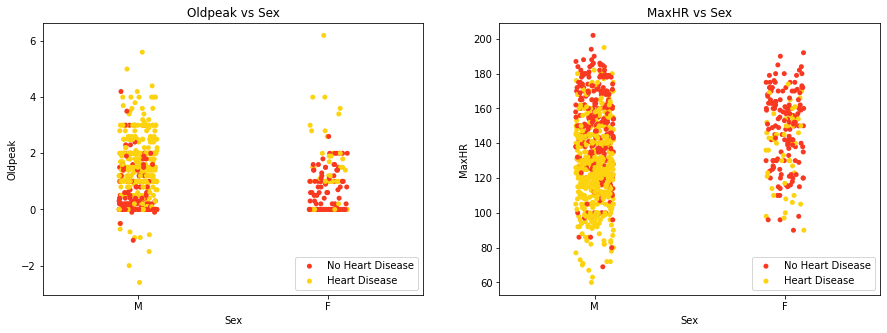
AI-generated content may be incorrect.A diagram of a box with a red and yellow line

AI-generated content may be incorrect.

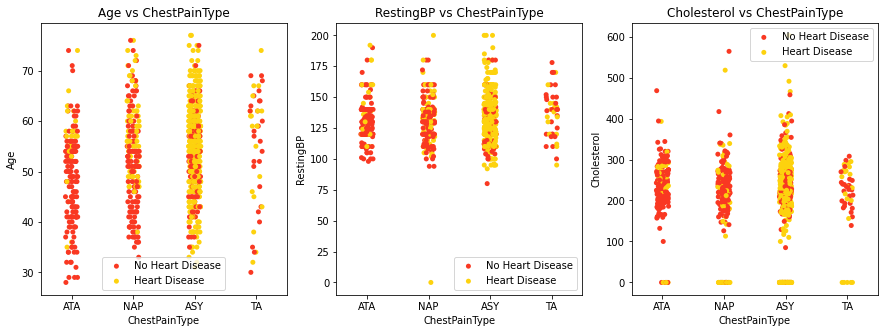
As visible from the boxplots, we can tell that our data lies in the following regions with heart disease:

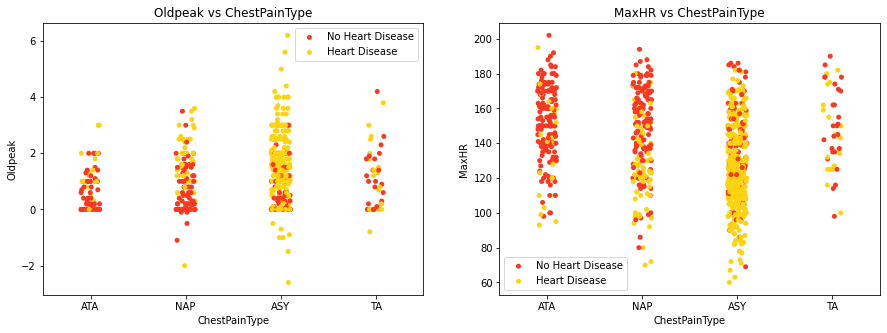
* Age: 50 – 65 years
* Cholesterol: 0 - 350
* MaxHR: 110 – 150
* Oldpeak: 0 - 8
* RestingBP: 110 – 150

You may notice that these figures are not the exact values represented on the boxplots, and that’s right. This is for a valid reason. While most of our data lies within the upper and lower quartiles as represented by the boxplots, there is some essential data that we may have missed out outside these regions. Thus, the ranges taken into consideration are a little higher and lower than the upper and lower ranges of the boxplots respectively. Since Oldpeak is a small range, we will take the entire range into consideration in this case.  
Now, let us take the categorical features into consideration and compare them against the numerical features to find correlations and make our assumptions of where the data lies even more accurate.

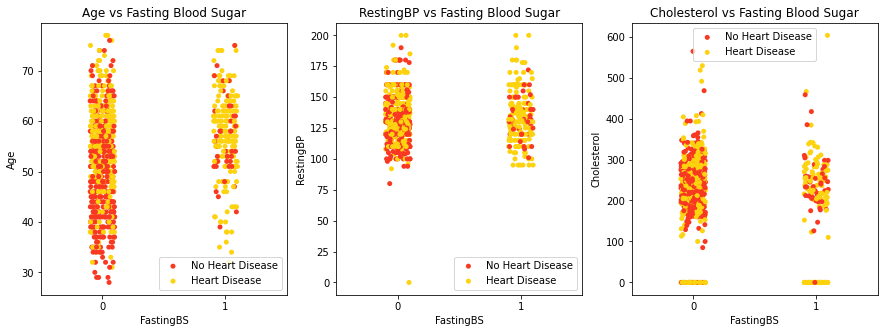
1. Numerical Features vs Sex:

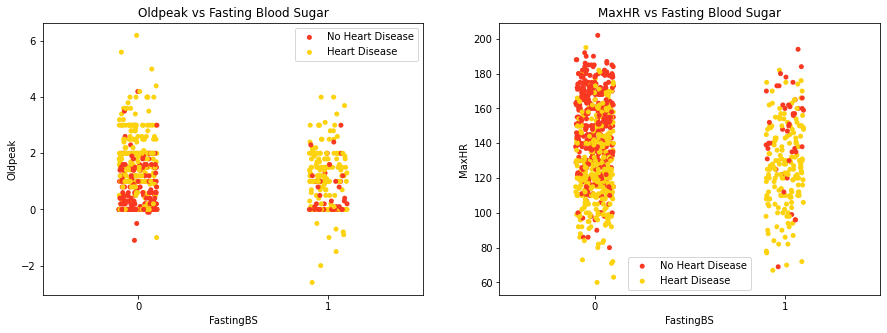
For males, the likelihood of heart disease is highest between ages 40 and 70, with oldpeak and maximum heart rate between 90 and 160. For females, there is much lesser data values for heart disease to make any valid observations from these graphs visually.

1. Numerical Features vs ChestPainType:

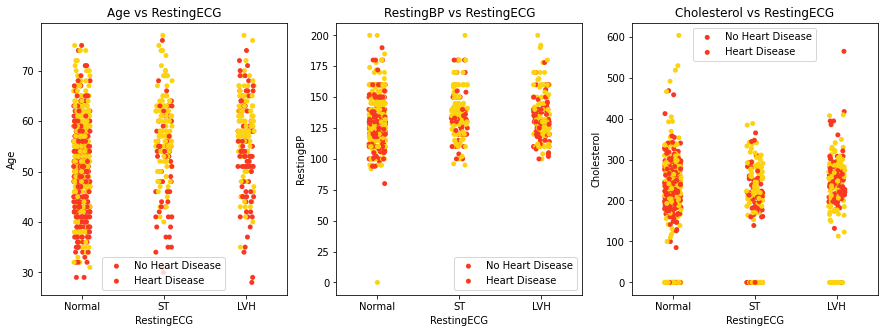


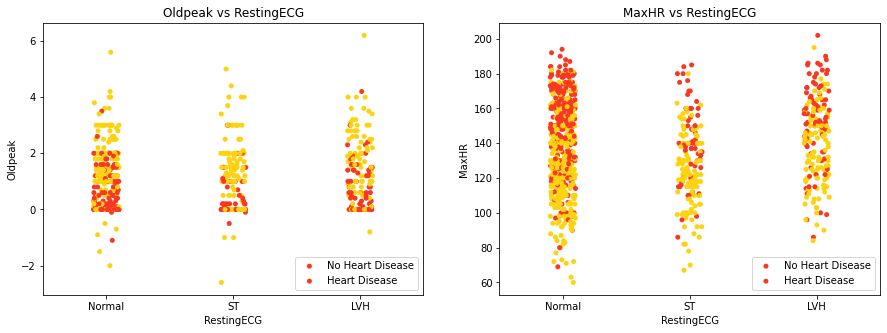
ChestPainType of ASY clearly dominates in all numerical features for the entire range of data values. So, there is not much to infer from this scenario.

1. Numerical Features vs FastingBS:



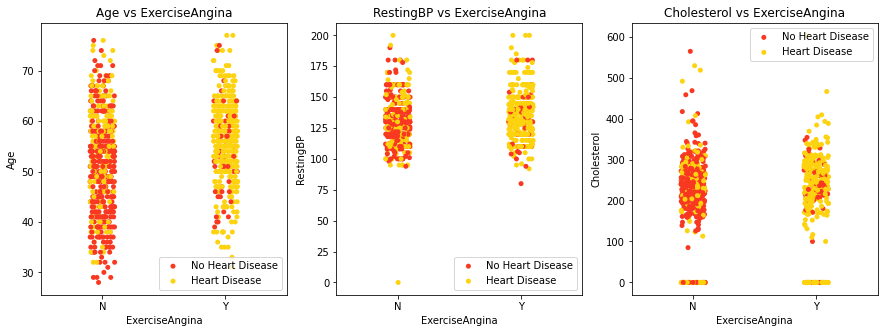
Irrespective of whether the patient has FastingBS, any patient above the age of 50 seems to be extremely likely to have heart disease. While most cases of heart disease in RestingBP were above 100, the chances are significantly higher for those with FastingBS. In the case of Cholesterol, it is difficult to make a reasonable assumption based on the data. This seems to be the same case for Oldpeak where most heart disease cases are above 0. In the case of MaxHR, most heart disease cases appear below 140.

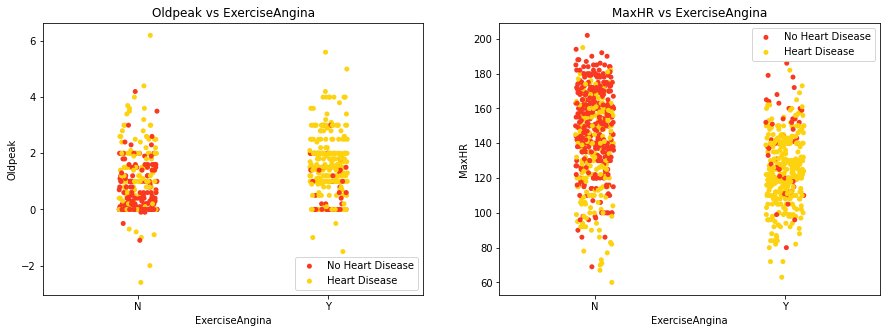
1. Numerical Features vs Resting ECG:



Heart diseases are detected based on RestingECG values, with Normal, ST, and LVH values indicating risk from ages 30, 40, and 40 respectively. Patients over 50 are at higher risk regardless of RestingECG values. Heart diseases are consistently observed across all RestingBP and RestingECG values. Patients with cholesterol levels between 200 and 300 and an ST RestingECG value show a specific pattern of heart disease. For maximum heart rate values, diseases are often detected in patients with values below 140 points and a Normal RestingECG, while ST and LVH values across maximum heart rate levels also show cases of heart disease.

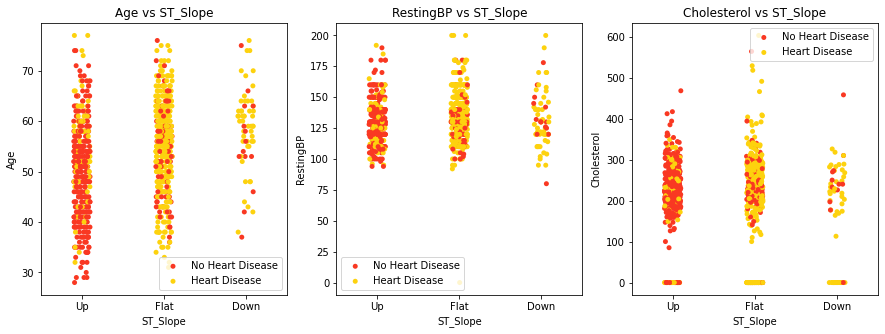
1. Numerical Features vs Exercise Angina:

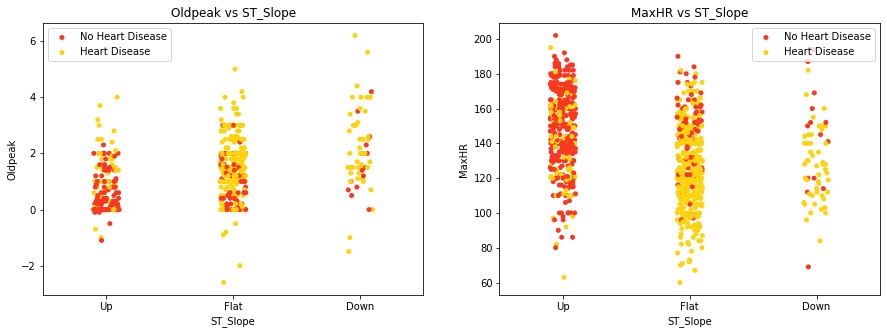




A clear observation reveals a positive correlation between heart disease cases and Exercise Induced Angina across all numerical features.

1. Numerical Features vs ST slope:





Another clear observation is the positive correlation between ST\_Slope values and heart disease cases. Specifically, a Flat ST\_Slope is associated with a high probability of heart disease, a Down ST\_Slope with a moderate probability, and an Up ST\_Slope with a low probability.

**Summary of our observations:**

Categorical Features (Ranked):

* Sex: Male > Female
* Chest Pain Type: ASY > NAP > ATA > TA
* Fasting Blood Sugar (FBS): (FBS < 120 mg/dl) > (FBS > 120 mg/dl)
* Resting ECG: Normal > ST > LVH
* Exercise Angina: Angina > No Angina
* ST Slope: Flat > Up > Down

Numerical Features (Ranges):

* Age: 50+
* Resting Blood Pressure (BP): 95 - 170
* Cholesterol: 160 - 340
* Maximum Heart Rate (MaxHR): 70 - 180
* Oldpeak: 0 – 4

With the typical values of these features understood, we can now proceed to select the appropriate features for modelling.

**Build Machine Learning Models using our data:**

Let us now proceed to build various machine learning models to find the most accurate predictor of heart disease using patients’ physical attributes and the symptoms that we’ve identified to be contributors of this medical issue.

**Precautions:**

* **Underfitting** – A model of extremely low accuracy is not ideal as it may make wrong predictions
* **Overfitting** – A model of extremely high accuracy is not ideal as well since it may perform specific to the given dataset which contains some outliers. Thus, in a case of general usage it may not make the right decisions as well.
* **Data bias** – We have identified factors that contribute towards heart disease and we will give them considerable importance in our models. However, we must make sure not to give them disproportionate weight resulting in other essential factors going unnoticed leading to a wrong prediction.
* **Non-representative training data** – Our data only contains 1190 observations compared to the nearly 500 million people affected by cardiovascular diseases. This dataset may not include several unique conditions and unique diseases.

To identify the most effective model for predicting heart disease based on patient attributes, we trained and evaluated six distinct machine learning classification algorithms using the processed dataset. These models were selected for their varied learning strategies and suitability for structured, tabular medical data.

The selected algorithms include:

* **Logistic Regression** – a linear model used for binary classification problems
* **Decision Tree Classifier** – a rule-based tree structure for interpretable classification
* **Random Forest Classifier** – an ensemble of decision trees that improves generalization
* **Gradient Boosting Classifier** – a powerful boosting technique that reduces bias
* **K-Nearest Neighbors (KNN)** – a distance-based lazy learning algorithm
* **Support Vector Machine (SVM)** – a hyperplane-based classifier optimized for margin maximization

**Data Preprocessing:**

The features were standardized using StandardScaler to ensure uniform scaling across numerical variables. This step is essential for algorithms sensitive to feature magnitude, such as KNN and SVM. The dataset was then split into training and testing sets with an 80:20 ratio, maintaining stratified class distribution.

Although our earlier data analysis identified certain features such as chest pain type, exercise-induced angina, ST segment slope, and age to be more strongly associated with heart disease, we did not manually assign them greater importance when training the machine learning models. Instead, all models were trained on the full feature set, allowing the algorithms themselves to determine which variables were most predictive. In particular, tree-based models like Random Forest and Gradient Boosting are known to automatically evaluate and weigh feature importance internally. This approach prevents human bias from overriding the learning process, while still capturing the natural influence of more significant variables.

**Model Evaluation:**

Each model was trained on the training set and tested on unseen data. We evaluated the models using multiple performance metrics:

* **Accuracy** – overall correctness of predictions
* **Precision** – indicates how many predicted positives were actually correct (important in reducing false alarms).
* **Recall** – sensitivity to capture how many actual positives were identified correctly (crucial for medical diagnosis).
* **F1-Score** – harmonic mean of precision and recall. It is helpful when false positives and false negatives are both critical.
* **AUC-ROC** – ability to distinguish between classes at various threshold levels. It gives a broader measure of model discrimination ability across all thresholds where higher score is better.

These metrics allow for a well-rounded understanding of each model’s strengths, not just in terms of how often it is correct (accuracy), but also in how well it distinguishes between heart disease positive and negative cases, especially under imbalanced or borderline scenarios.

**Model Performance Overview and Comparison:**

All six models showed promising results, with some outperforming others in specific metrics. The table below summarizes the classification results, with each model revealing different insights:

A screenshot of a computer screen

AI-generated content may be incorrect.

Let’s break this down model by model:

**Logistic Regression -** Performs well, with high precision of 0.91 and good recall score of 0.82, resulting in a solid F1-score of 0.86. Since it’s a linear model, it may struggle with complex nonlinear patterns, hence slightly lower recall compared to ensemble models. It is still a very interpretable and dependable baseline.

**Decision Tree -** Has the lowest AUC score of 0.794 and recall score of 0.757 among all the models.This model tends to overfit easily on small datasets without pruning or tuning, which likely explains its subpar generalization.Performance can improve with techniques like pruning, depth limitation, or transitioning to ensemble methods (e.g. Random Forest).

**Random Forest -** Delivers excellent balance across all metrics with both precision and recall at 0.897, and AUC of 0.942. Its strong performance comes from using multiple decision trees to reduce overfitting and improve generalization as we just mentioned above. It is suitable when interpretability is less important than predictive power.

**Gradient Boosting -** Very high precision of 0.92 with solid recall score at 0.86, leading to an F1-score of 0.889 and AUC of 0.936.It builds trees sequentially and corrects earlier mistakes, thus often leading to better performance on structured or tabular data like this.Slightly lower recall than Random Forest suggests that it may miss a few true positives but avoids false alarms better.

**K-Nearest Neighbours –** KNN matches Logistic Regression in most metrics, though with slightly higher AUC at 0.922.It is likely limited by the curse of dimensionality and sensitivity to feature scaling, especially with continuous features.KNN can sometimes struggle when there are many numerical features because it relies heavily on calculating distances between points. This makes it sensitive to how the data is scaled or distributed. This can be improved by tuning the k value (number of neighbours) and trying different distance metrics.

**Support Vector Machine -** Best overall AUC score at 0.949 and high F1-score of 0.882, making it the top performer. SVM tends to work well when the dataset has many features (multiple dimensions). It also includes built-in techniques to avoid overfitting, making it more reliable on complex data.It has slightly lower precision than Gradient Boosting, but better recall, which is often more critical in medical settings (fewer missed heart disease cases).

In addition to overall performance metrics, confusion matrices were examined for each model to better understand the types of errors they made. These matrices reveal the number of true positives (correct heart disease predictions), true negatives (correct non-heart disease predictions), false positives (incorrectly predicting heart disease), and false negatives (failing to detect actual heart disease).

Among all models, those with higher recall such as Support Vector Machine, Random Forest, and Gradient Boosting produced fewer false negatives, which is especially important in medical contexts where missing a heart disease case could lead to serious consequences. Conversely, models with lower recall, such as the Decision Tree, had more false negatives and thus higher clinical risk. This reinforces the importance of prioritizing recall and AUC over just accuracy in healthcare prediction tasks.  
  
In the context of heart disease prediction, recall and AUC are arguably the most important because high recall value ensures that we catch as many true heart disease cases as possible, while also avoiding false negatives. Alongside this, AUC shows how well a model can distinguish between positive and negative classes across all thresholds, not just at 0.5. Thus, SVM, Gradient Boosting, and Random Forest offer the strongest blend of sensitivity and reliability and are so far the best models as per our results.

To visually interpret the above metrics, bar plots were generated showing each model’s performance across the four key classification metrics.

A graph of different colored bars

AI-generated content may be incorrect.

As the differences between models are often small at this high level of performance, a zoomed-in version of the bar plot was created with a y-axis limited to 0.70–1.00, making it easier to distinguish their relative effectiveness.

A graph of different colored bars

AI-generated content may be incorrect.

This zoomed graph makes it easier to spot subtle but significant differences, especially when all models perform within the 0.80–0.95 range.

Key Observations:

* Random Forest shows consistent height across all metrics — a clear indicator of well-rounded performance.
* Gradient Boosting stands out in precision, suggesting it avoids false positives better.
* SVM excels in recall and AUC, implying it identifies actual heart disease patients more reliably.
* Decision Tree, by contrast, has visibly lower bars — reinforcing its weaker generalization.\

In addition to metric-based evaluation, ROC Curves were plotted for all models, showing the trade-off between the True Positive Rate (Sensitivity) and the False Positive Rate. A model with a ROC curve closer to the top-left corner is considered more effective. The ROC curves below reveal that SVM, Gradient Boosting, and Random Forest exhibit steep and high-performing curves, confirming their robustness across threshold variations.

A graph of a number of different colored lines

AI-generated content may be incorrect.

Analysis of Curves:

* Support Vector Machine achieves the most steep and well-separated ROC curve, with AUC = 0.949, indicating high discriminatory power.
* Random Forest and Gradient Boosting also have strong, consistent ROC curves (AUCs above 0.93).
* Decision Tree’s curve is noticeably flatter, reflecting more false positives or false negatives depending on the threshold — which aligns with its lower recall and AUC.

ROC (Receiver Operating Characteristic) curves were generated for all models to evaluate performance across all classification thresholds. The AUC (Area Under Curve) quantifies each model’s ability to distinguish between heart disease and non-disease cases. In medical prediction, it’s not always best to use the default threshold (e.g., 0.5). The AUC gives us a holistic view across all possible thresholds, helping clinicians adjust the sensitivity/specificity trade-off as needed.

**Ranking of Machine Learning Models:**

Based on the combination of evaluation metrics — particularly **recall**, **F1-score**, and **AUC**, which carry higher weight in medical prediction — the models can be ranked from most to least effective as follows:

**1) Support Vector Machine (SVM) -** Highest AUC (0.949) and strong F1-score (0.882), with excellent recall (0.869), making it the most reliable in identifying heart disease.

**2) Gradient Boosting** - Outstanding precision (0.920) and balanced performance overall, with an AUC of 0.936.

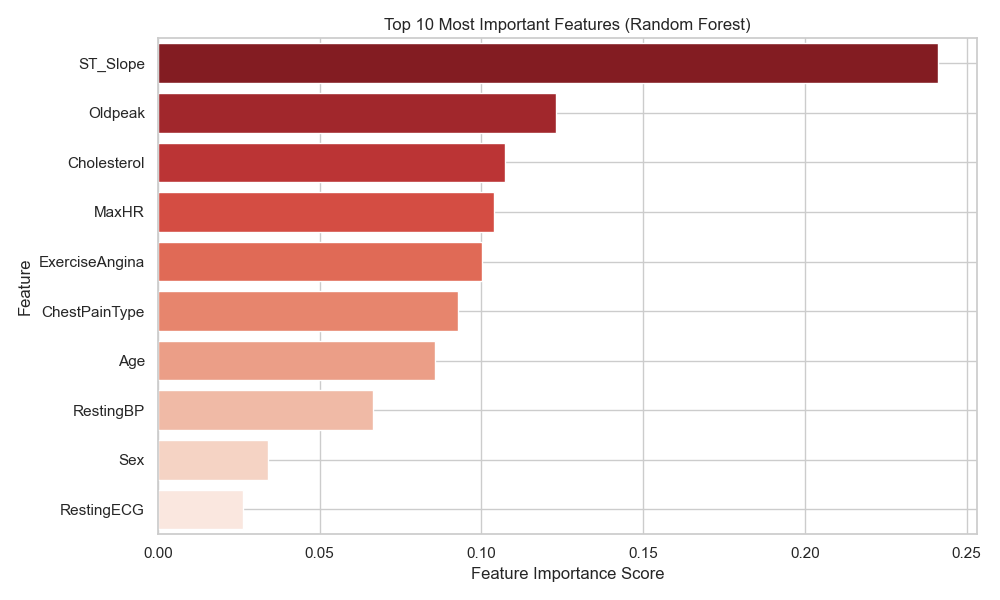
**3) Random Forest** - Very stable across all metrics with AUC of 0.942, though slightly lower F1-score than SVM.

**4) Logistic Regression & K-Nearest Neighbors (Tied)** - Both had solid performance with F1-scores of 0.863 and identical accuracy/recall values.

**5) Decision Tree** - Simpler and more interpretable, but lowest AUC (0.794) and weaker generalization due to overfitting tendencies.

**Feature Importance Interpretation:**

To better understand which attributes most influenced the model's decision-making, we analyzed feature importance scores derived from the trained Random Forest classifier. The chart below displays the top 10 most important predictors based on how frequently they were used to split decision trees in the ensemble.



The results closely align with earlier data analysis. ST\_Slope, a measure of ST segment depression during exercise, emerged as the most influential feature. It was followed by Oldpeak (also ST-related), Cholesterol, MaxHR (maximum heart rate), and ExerciseAngina, all of which had shown strong relationships with heart disease in our exploratory phase. Features like Age, ChestPainType, and RestingBP also held considerable predictive value.

This alignment confirms that the Random Forest model internally prioritized features that were already identified as clinically and statistically significant. It also supports the interpretability of ensemble methods, which not only perform well but provide transparency in how predictions are made

**How to Further Improve These Models:**

1. Feature Engineering

We can create new features from the existing data that may better capture the patterns behind heart disease. For example: Combine features like cholesterol and age into a new one called "cholesterol per age". We could also use a method called PCA (Principal Component Analysis) to remove noise and focus only on the most important parts of the data.

2. Hyperparameter Tuning

Every model has certain internal settings that affect how it learns.  
Using tools like GridSearchCV or RandomizedSearchCV, we can test different combinations of settings to find the best version of each model — especially for SVM, Random Forest, and Gradient Boosting.

3. Handling Imbalanced Data

If one class (e.g., “No Heart Disease”) has more examples than the other, models might become biased.  
To fix this, we can use SMOTE, a technique that creates more examples of the smaller class. We could also adjust model settings to give more importance to the minority class, so the model learns to detect it better.

4. Combining Models (Ensembling)

Instead of relying on just one model, we can combine the predictions of the top three (SVM, Random Forest, Gradient Boosting).  
This is called a soft voting classifier, and it helps reduce individual model weaknesses by averaging their outputs.

5. Using Cross-Validation

Rather than training and testing the model on one fixed split of the data, we can test it across multiple different splits using something called k-fold cross-validation.  
This gives us a more accurate and reliable idea of how the model will perform on new, unseen data.

**Conclusion:**

This study demonstrates the effectiveness of machine learning in predicting heart disease using patient attributes and symptoms. Among the six tested models, Support Vector Machine achieved the highest overall performance, with Gradient Boosting and Random Forest close behind. These findings highlight the promise of data-driven approaches in improving early diagnosis and clinical decision-making in cardiology.

The models developed in this study demonstrate the potential of machine learning to serve as a supportive tool in clinical decision-making. For example, these models could be integrated into hospital systems to assist physicians in early risk screening, helping identify patients more likely to have heart disease based on readily available attributes like chest pain type, ST slope, cholesterol levels, and blood pressure. Such tools can improve speed and consistency in diagnosis, especially in resource-limited settings where access to specialists may be delayed.

While the models showed strong predictive performance, it’s important to recognize that the dataset may not fully represent all population subgroups. Factors such as age, gender, ethnicity, and pre-existing conditions could influence heart disease risk but may be underrepresented or unbalanced in this data. To ensure fairness and avoid bias in real-world deployment, future versions of this model should be trained and validated on diverse, demographically representative datasets, and tested for fairness across subgroups. Additionally, these tools should be used as aiding systems, not replacements for medical judgment.