



Future-Proofing Hearts: A Decade Ahead

Predicting Ten Year Risk of Coronary Heart Disease (CHD)

Introduction

Confronting Coronary Heart Disease

Background and Context:

- Heart disease is a prevalent health concern worldwide.
- It imposes a significant burden on public health and healthcare systems.
- The development of predictive models is crucial to address this issue effectively.

Significance of Predicting CHD Risk:

- Predicting Coronary Heart Disease (CHD) risk is vital for early intervention.
- It enables healthcare providers to offer personalized care and preventive strategies.
- Ultimately, it can reduce the incidence of CHD and improve patient outcomes.

Overview of the Project:

- Our project aims to predict CHD risk using data analytics.
- We leverage demographic and medical data to develop predictive models.
- Our goal is to provide actionable insights for healthcare professionals and individuals to mitigate CHD risk.

Objectives

Primary Goals:

- Our main objective is to develop a robust predictive model for CHD risk.
- This model will aid in estimating an individual's likelihood of developing coronary heart disease within the next 10 years.

Secondary Goals:

- In addition to prediction, we aim to identify and analyze the key risk factors associated with CHD.
- We will explore the dataset to uncover insights into the behavioral, medical, and demographic factors contributing to CHD risk.

Business Problem

The CHD Challenge

Healthcare Challenge:

- Coronary Heart Disease (CHD) is a pressing healthcare challenge.
- It's a leading cause of morbidity and mortality globally.
- The economic and human costs of CHD are substantial, affecting individuals and healthcare systems.

Project's Role:

- Our project is dedicated to addressing the CHD challenge through data analytics and machine learning.
- By leveraging data, we aim to:
 - Predict CHD risk for early intervention.
 - Provide insights for targeted healthcare strategies.
 - Contribute to reducing CHD incidence and improving public health.

Assumptions

- The dataset is representative of the population under study.
- The provided variables are sufficient to make accurate predictions about CHD risk.
- The data is free from significant errors or biases.
- The relationship between variables and CHD risk is stable over time.

Data Description

Our Data Treasure Trove

Dataset Overview:

- Our dataset is sourced from an ongoing cardiovascular study in Framingham, Massachusetts.
- It comprises of 3390 records, structured into 17 columns, providing rich insights into CHD risk factors.

Feature Descriptions:

- Features encompass demographic, behavioral, and medical attributes.

Target Variable: TenYearCHD

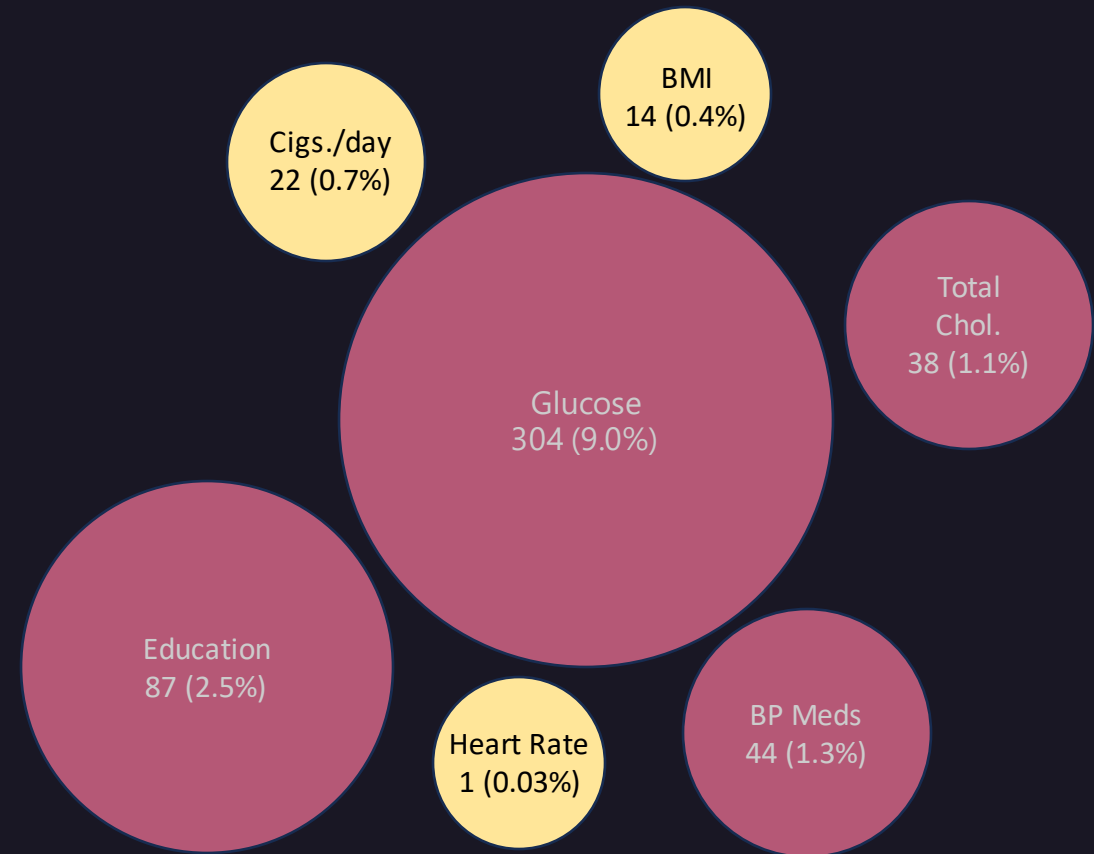
- The focal point of our analysis is the "TenYearCHD" variable.
- It signifies whether an individual is at risk of developing CHD within the next decade.


| Variable Name | Type | Description |
|---------------------|-------------|---|
| id | Unique | Unique identifier for each individual |
| age | Numeric | Age of the individual |
| education | Categorical | Education level of the individual |
| gender | Categorical | Gender of the individual |
| is_smoking | Categorical | Smoking status (Smoker/Non-Smoker) |
| cigsPerDay | Numeric | Average number of cigarettes smoked per day for smokers |
| BPMeds | Categorical | Blood Pressure Medication usage (Yes/No) |
| prevalentStroke | Categorical | History of prevalent stroke (Yes/No) |
| prevalentHyp | Categorical | Hypertensive condition (Yes/No) |
| diabetes | Categorical | Diabetic condition (Yes/No) |
| totChol | Numeric | Total cholesterol level |
| sysBP | Numeric | Systolic Blood Pressure |
| diaBP | Numeric | Diastolic Blood Pressure |
| BMI | Numeric | Body Mass Index |
| heartRate | Numeric | Heart rate |
| glucose | Numeric | Glucose level |
| TenYearCHD | Categorical | The target variable indicating 10-year risk of CHD (Yes/No) |
| Demographic Factors | | Behavioural Factors Medical Risk Factors |

Data Cleaning

Brewing the Data

- The features have already been encoded.
- Columns having missing values are Education, Cigs./day, BP Meds, Total Chol. and Glucose. Except the feature **Glucose**, all other missing values are less than 2.5% of data.
- We shall deal with the missing data in the model creation part - because we want to avoid information leak.
- We shall drop redundant column 'id', as it is a unique identifier.





Exploratory Data Analysis (EDA)

EDA – Summary Statistics

Unearthing Insights

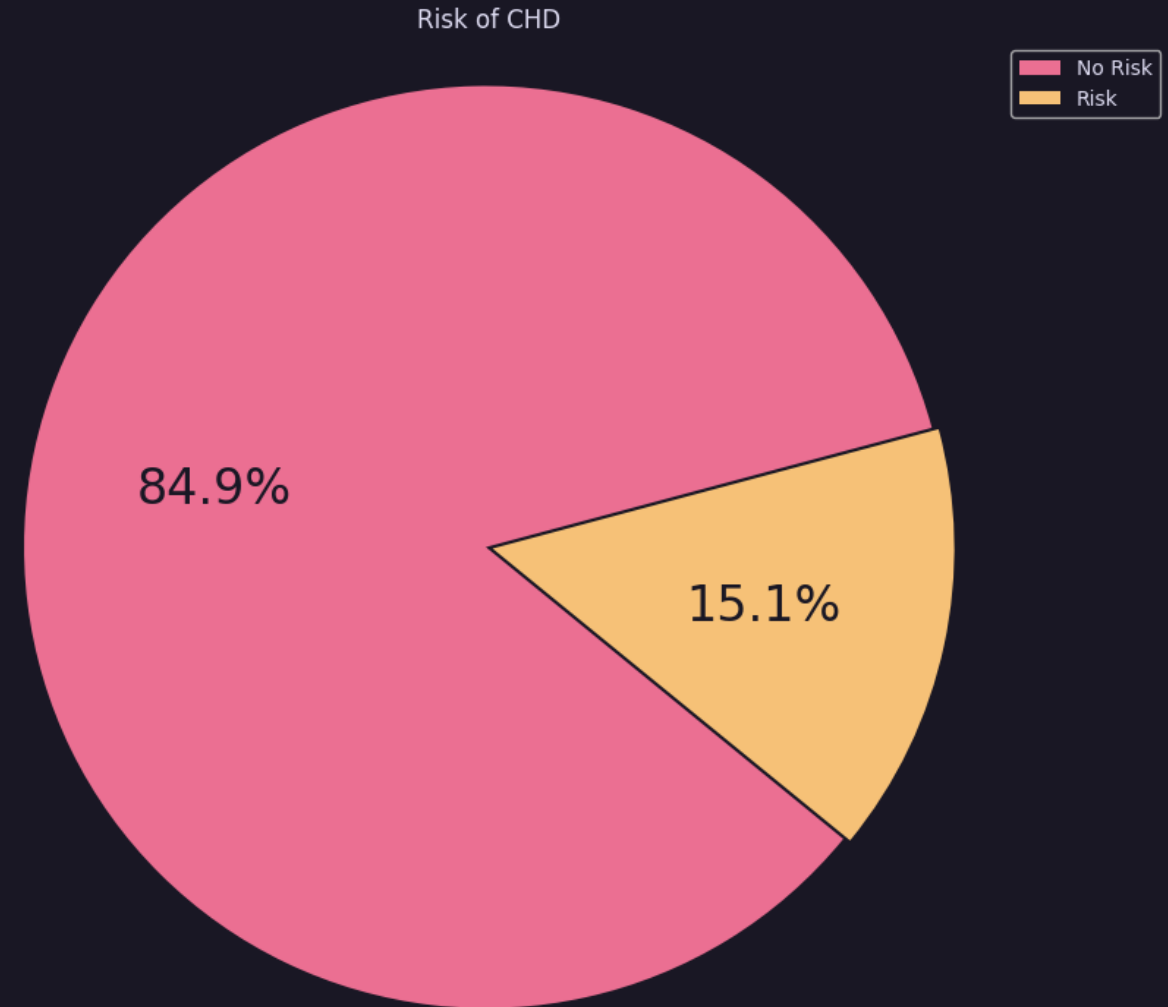
- All numeric variables (mean) are higher in the group who have a 10-year risk of coronary heart disease.

| Variables (Mean) | | |
|------------------------|--------|--------|
| TenYearCHD | 0 | 1 |
| Age | 48.73 | 54.13 |
| Education | 1.99 | 1.84 |
| Gender | 0.58 | 0.47 |
| Smoking | 0.49 | 0.54 |
| Cigarettes/Day | 8.73 | 10.95 |
| BP Medication | 0.02 | 0.07 |
| Prevalent Stroke | 0 | 0.02 |
| Prevalent Hypertension | 0.28 | 0.5 |
| Prevalent Diabetes | 0.02 | 0.06 |
| Total Cholestrol | 235.28 | 247.22 |
| Systolic BP | 130.6 | 143.85 |
| Diatolic BP | 82.19 | 86.76 |
| BMI | 25.68 | 26.45 |
| Heart Rate | 75.88 | 76.55 |
| Glucose | 80.66 | 89.97 |

EDA – Distribution Analysis

Unearthing Insights

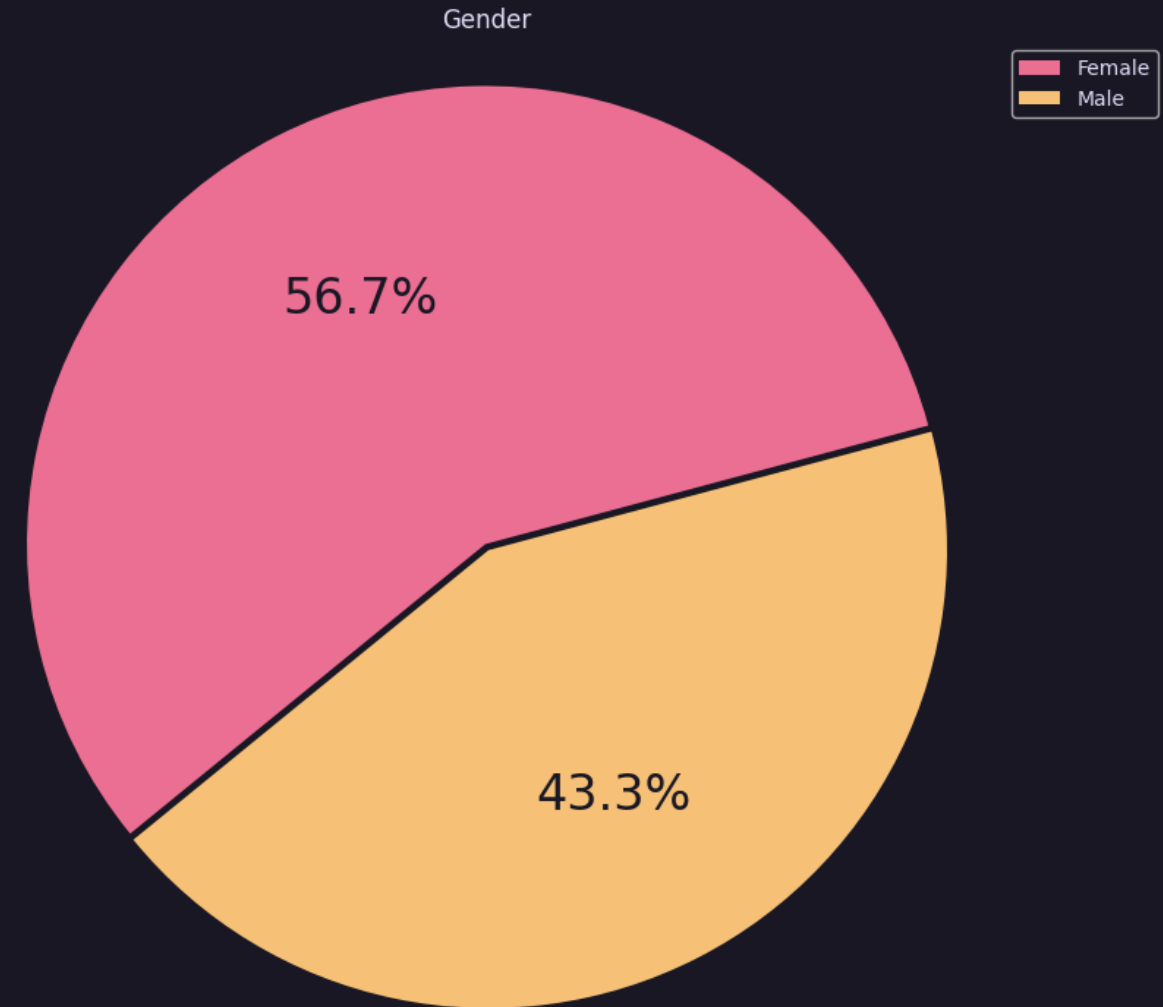
- Our target variable is imbalanced with 2879 (84.9%) records belonging to 'No Risk' class while 511 (15.1%) belonging to 'Risk' class.
- The number of 'No Risk' cases outweighs the number of 'Risk' cases.



EDA – Distribution Analysis

Unearthing Insights

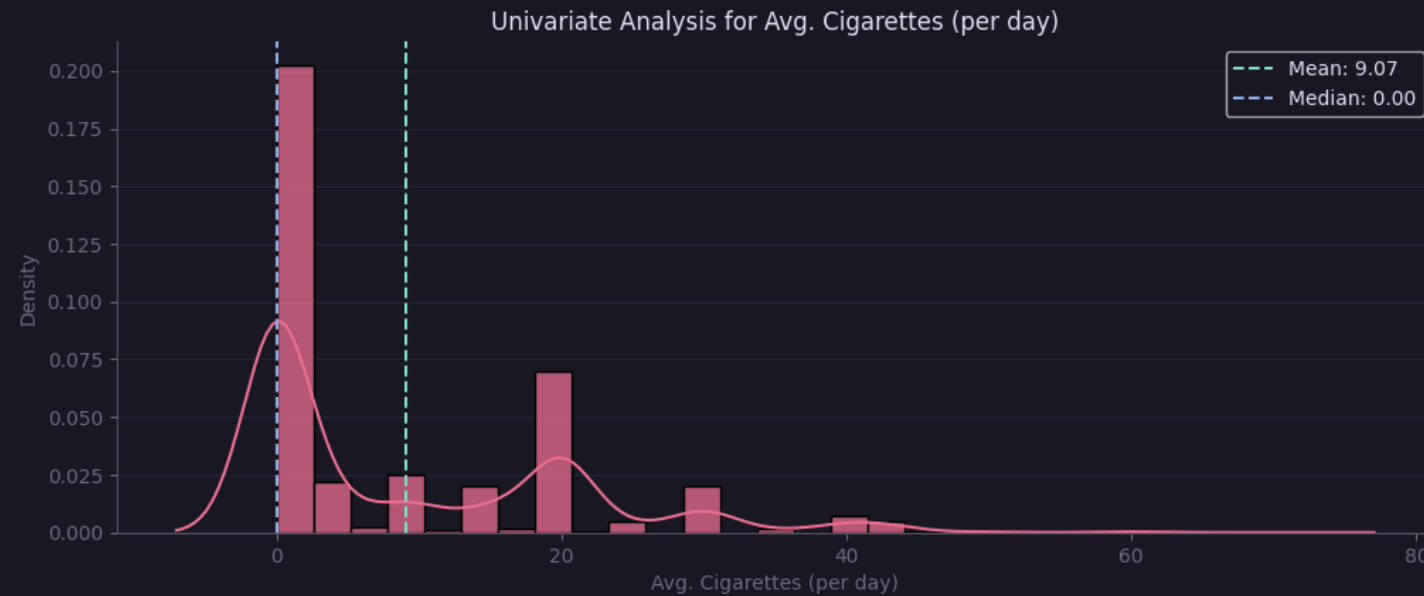
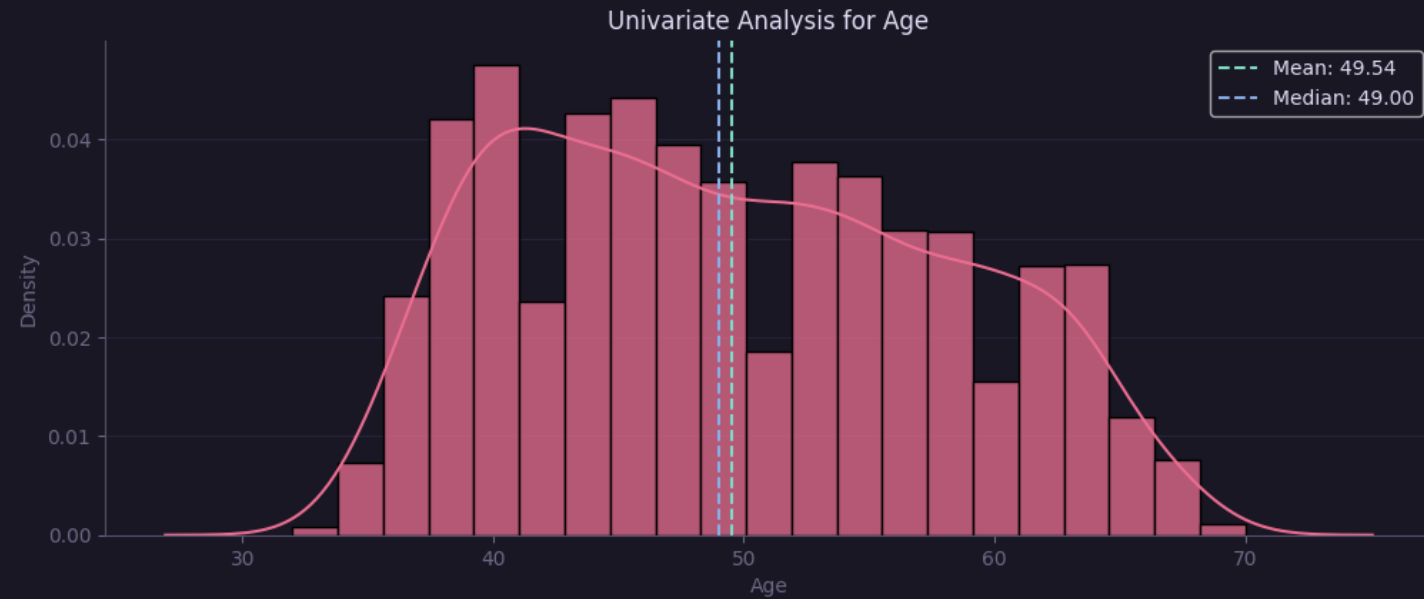
- Gender ratio in the dataset is slightly imbalanced with 1923 (56.7%) female and 1467 (43.3%) male.



EDA – Distribution Analysis

Unearthing Insights

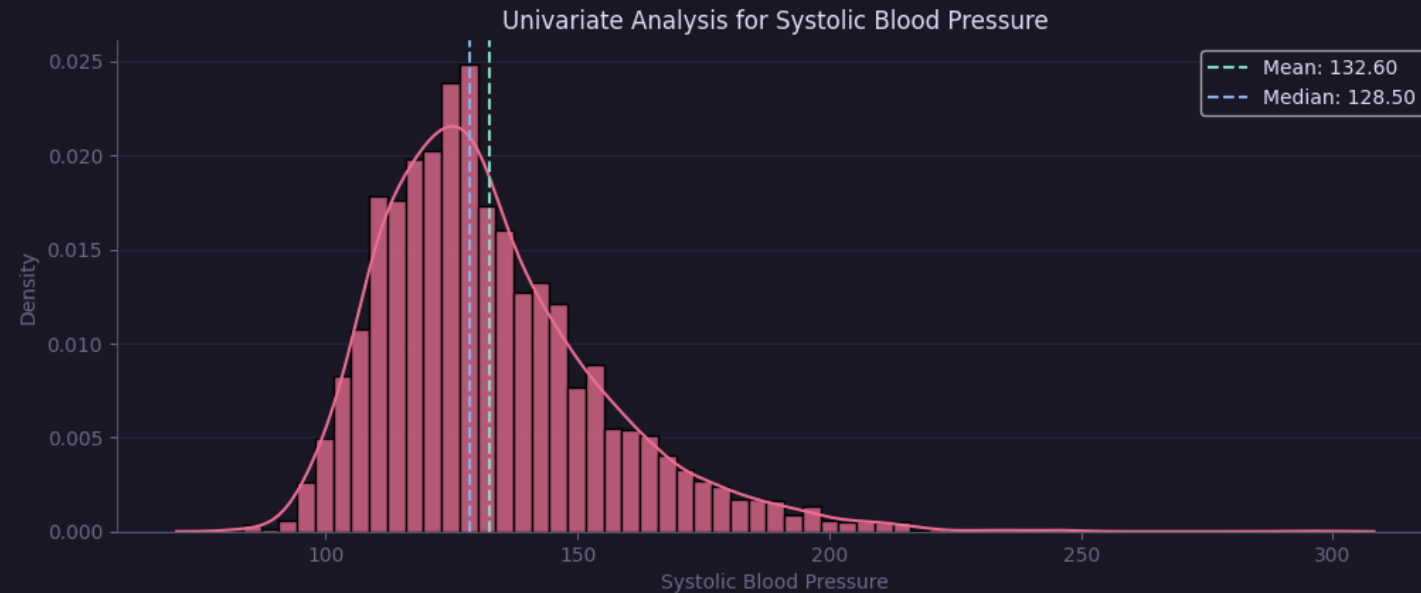
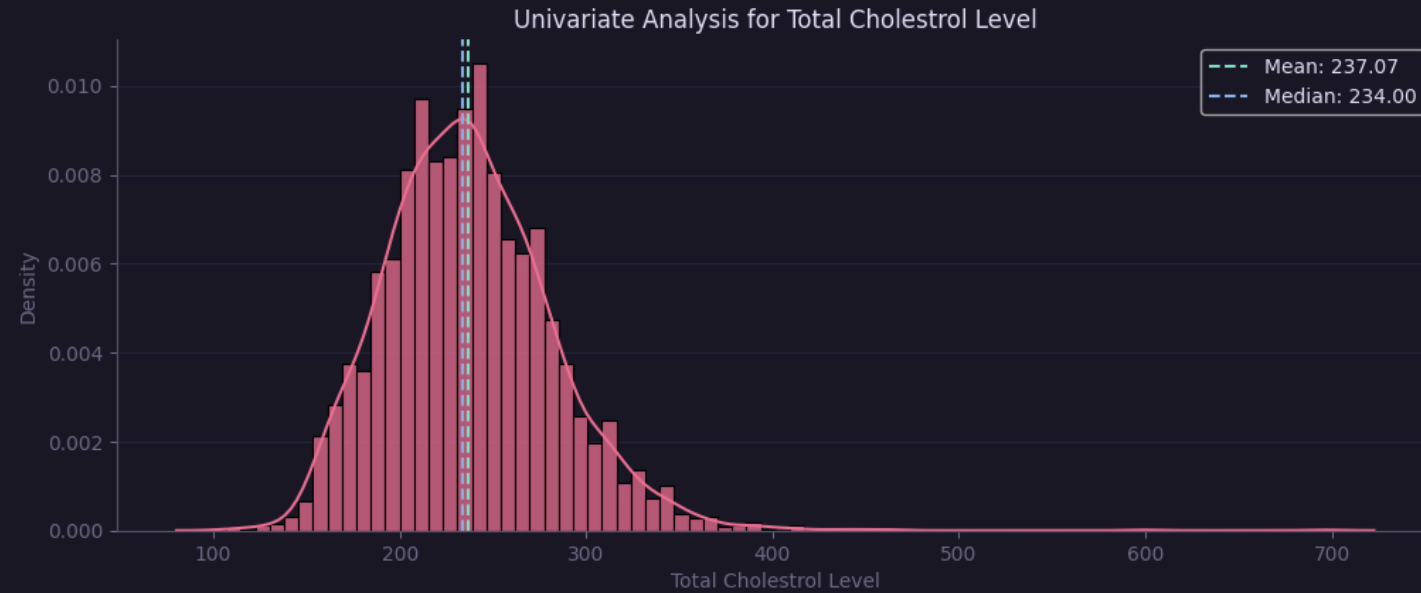
- Avg. Cigarettes (per day) has a highly uneven distribution with the most data present in 0.
- Avg. Cigarettes (per day) shows right skewness.



EDA – Distribution Analysis

Unearthing Insights

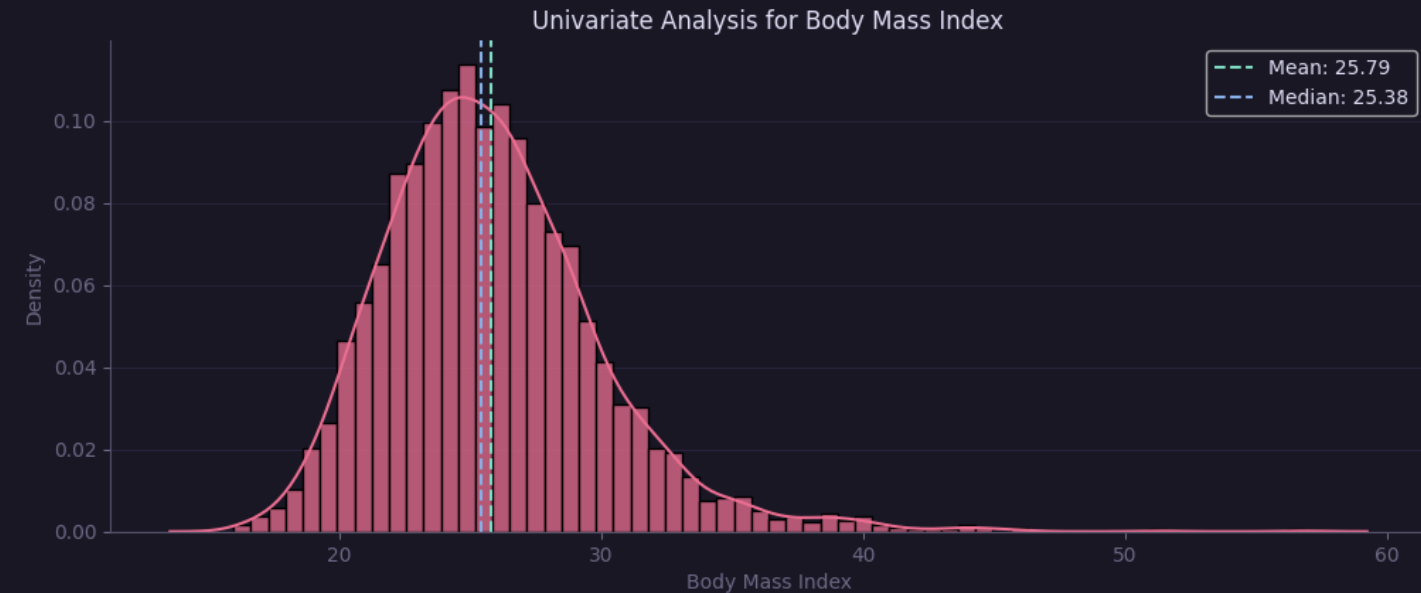
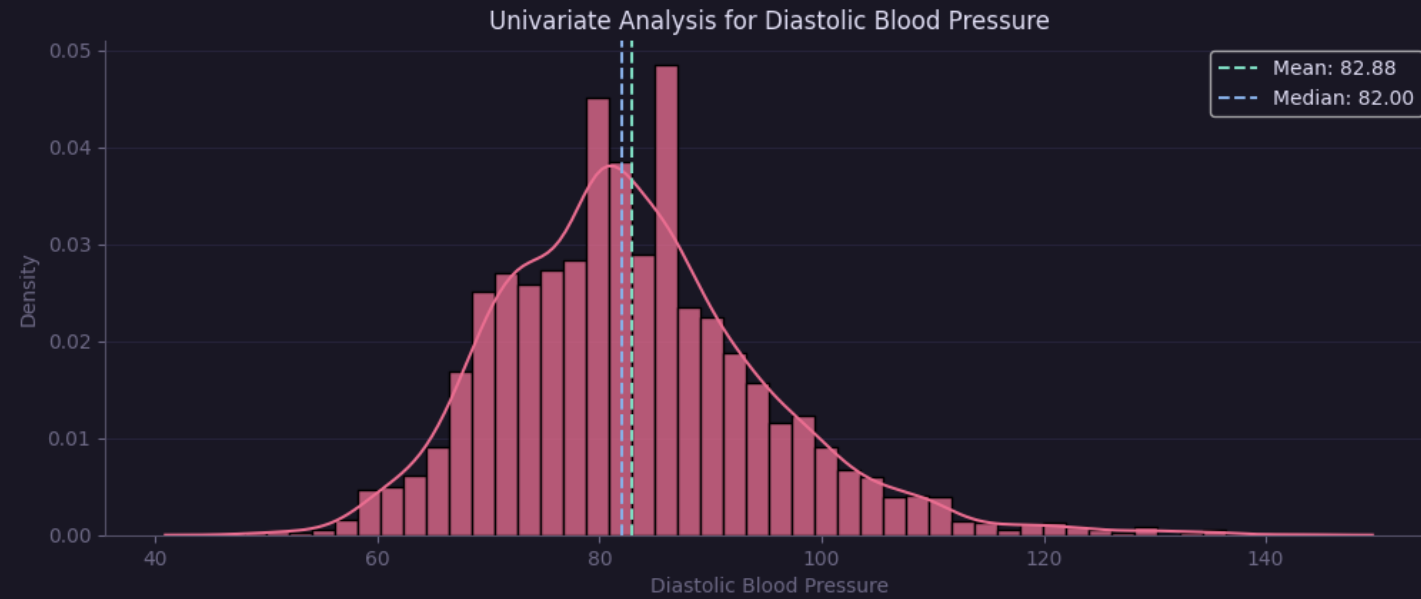
- Avg. Cigarettes (per day) has a highly uneven distribution with the most data present in 0.
- Avg. Cigarettes (per day) shows right skewness.
- Total Cholesterol has uniform distribution.
- Systolic BP shows slight right skewness.



EDA – Distribution Analysis

Unearthing Insights

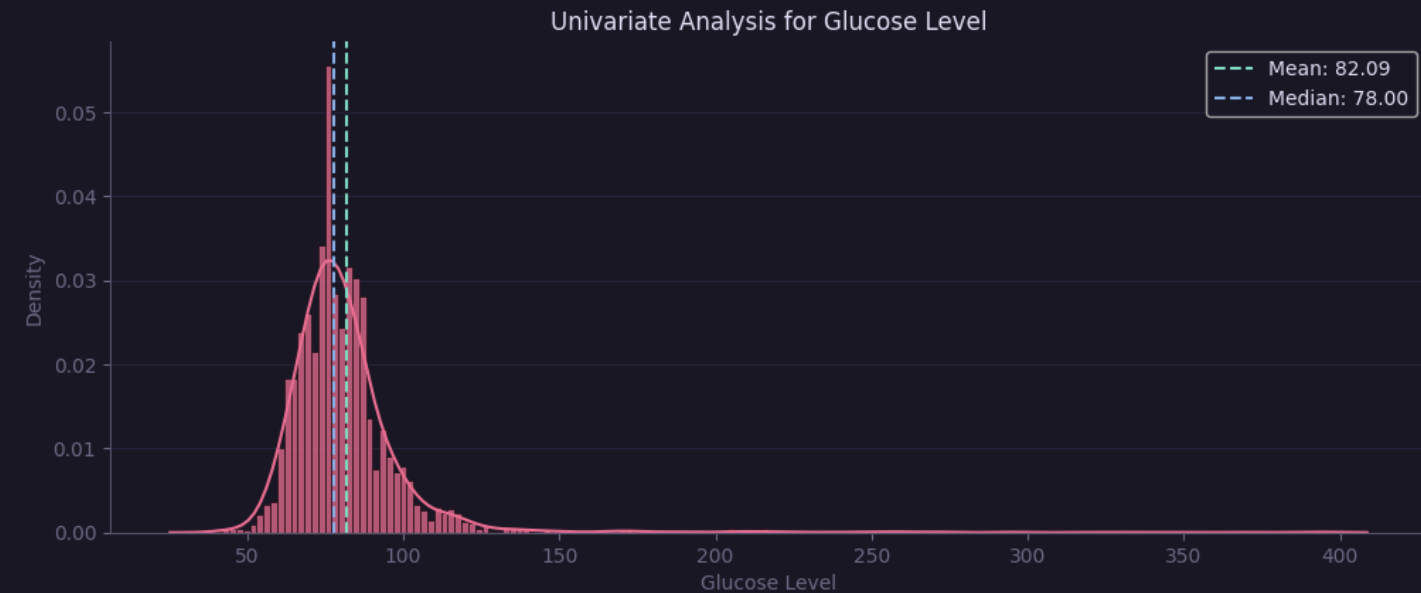
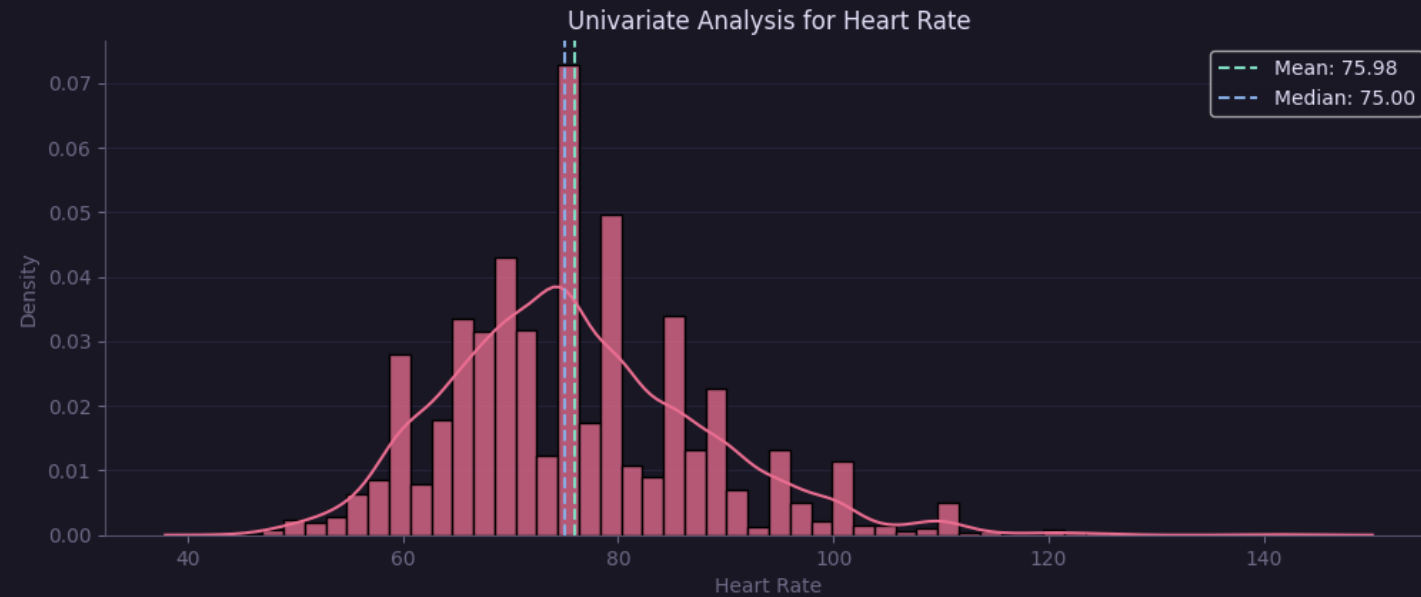
- Avg. Cigarettes (per day) has a highly uneven distribution with the most data present in 0.
- Avg. Cigarettes (per day) shows right skewness.
- Total Cholesterol has uniform distribution.
- Systolic BP shows slight right skewness.
- Diastolic BP and BMI has a uniform distribution.



EDA – Distribution Analysis

Unearthing Insights

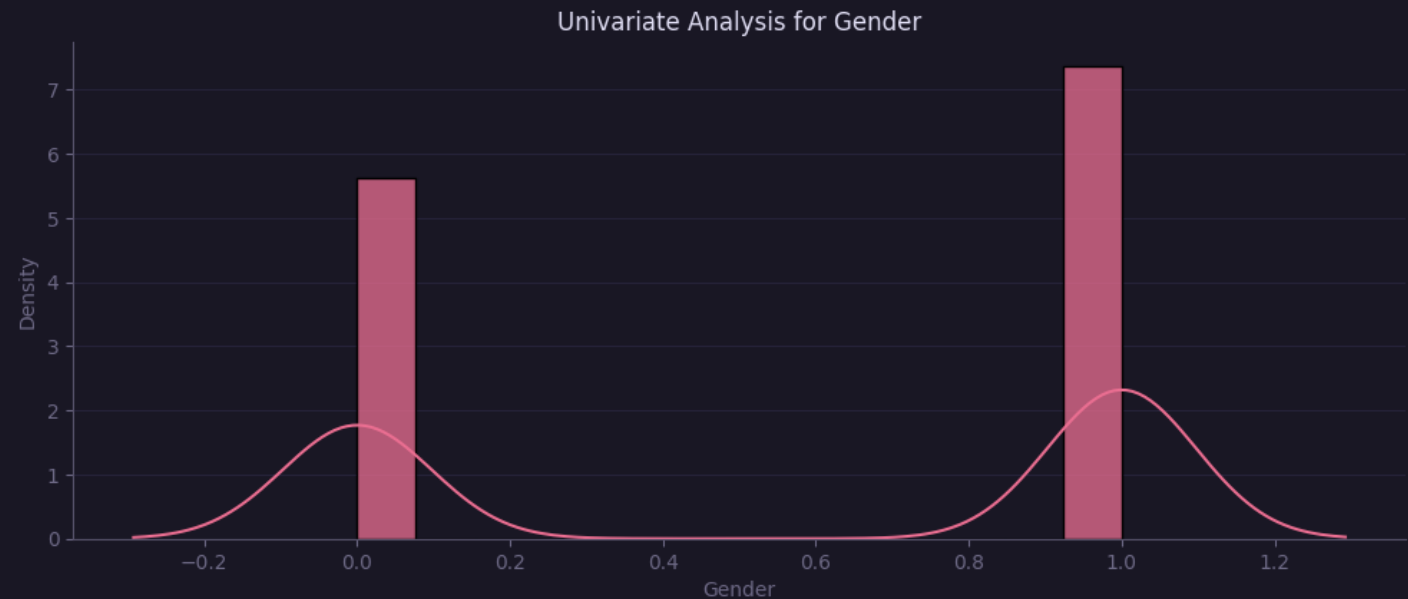
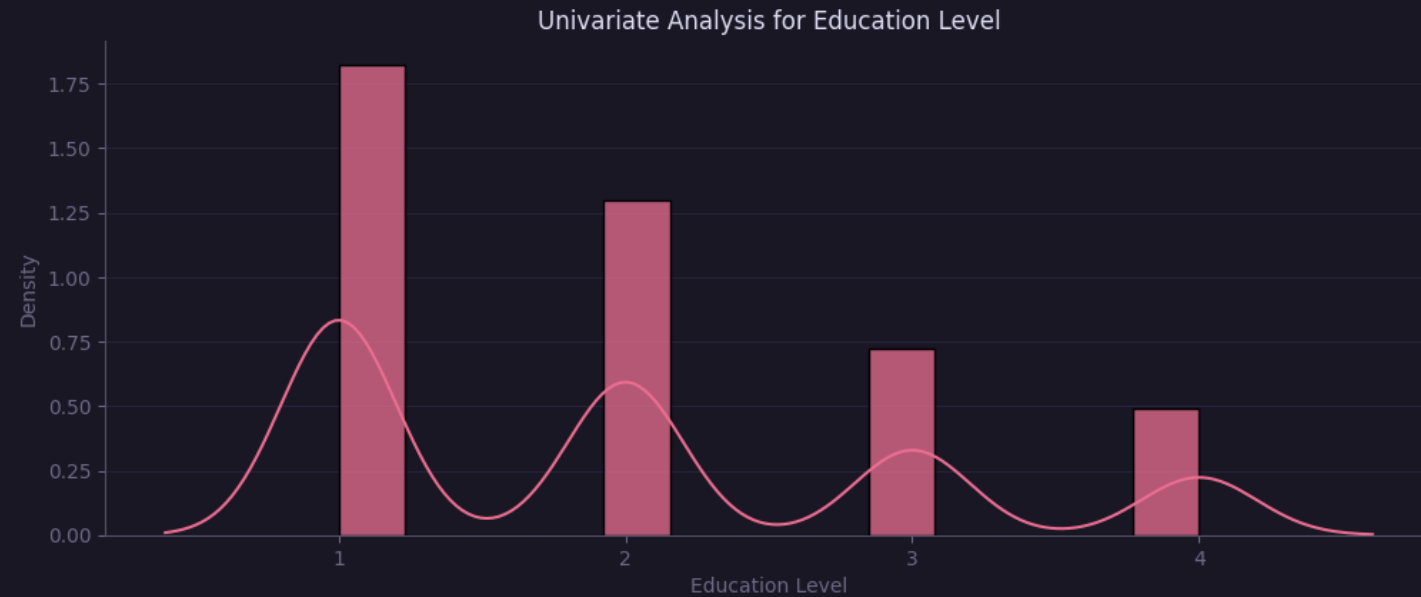
- Avg. Cigarettes (per day) has a highly uneven distribution with the most data present in 0.
- Avg. Cigarettes (per day) shows right skewness.
- Total Cholesterol has uniform distribution.
- Systolic BP shows slight right skewness.
- Diastolic BP and BMI has a uniform distribution while rest of the variables are unevenly distributed



EDA – Distribution Analysis

Unearthing Insights

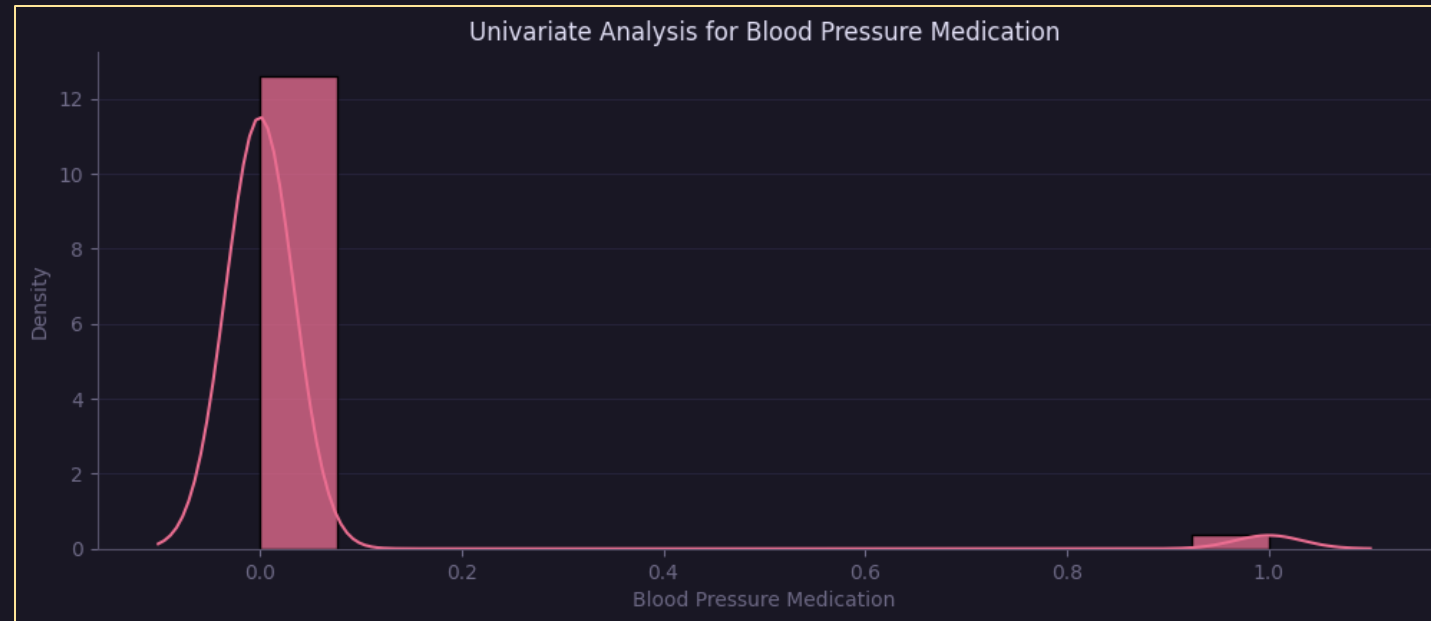
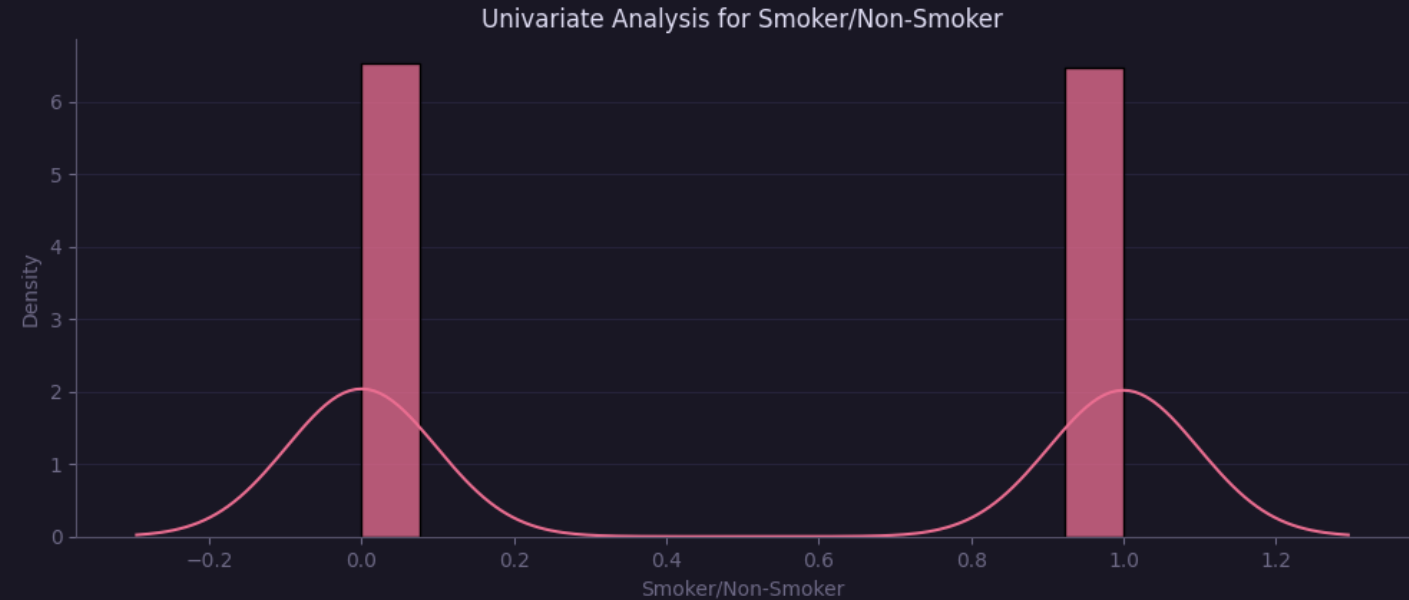
- There are four levels of education whereas the rest categorical features are all binary.



EDA – Distribution Analysis

Unearthing Insights

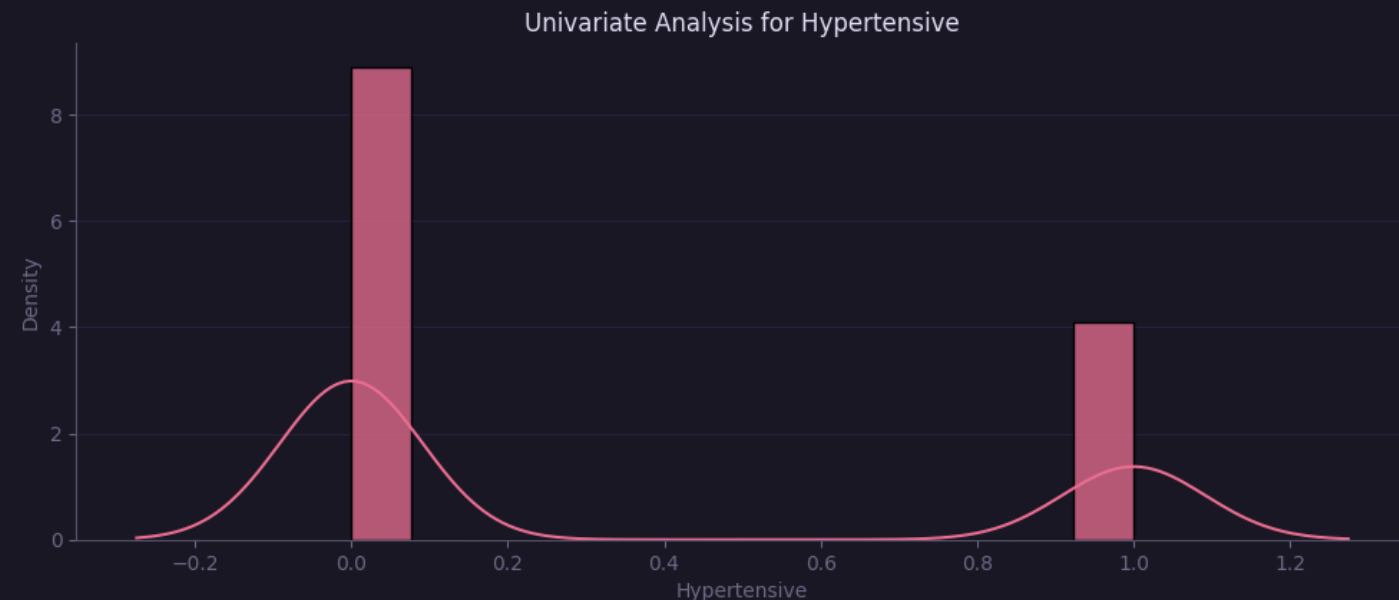
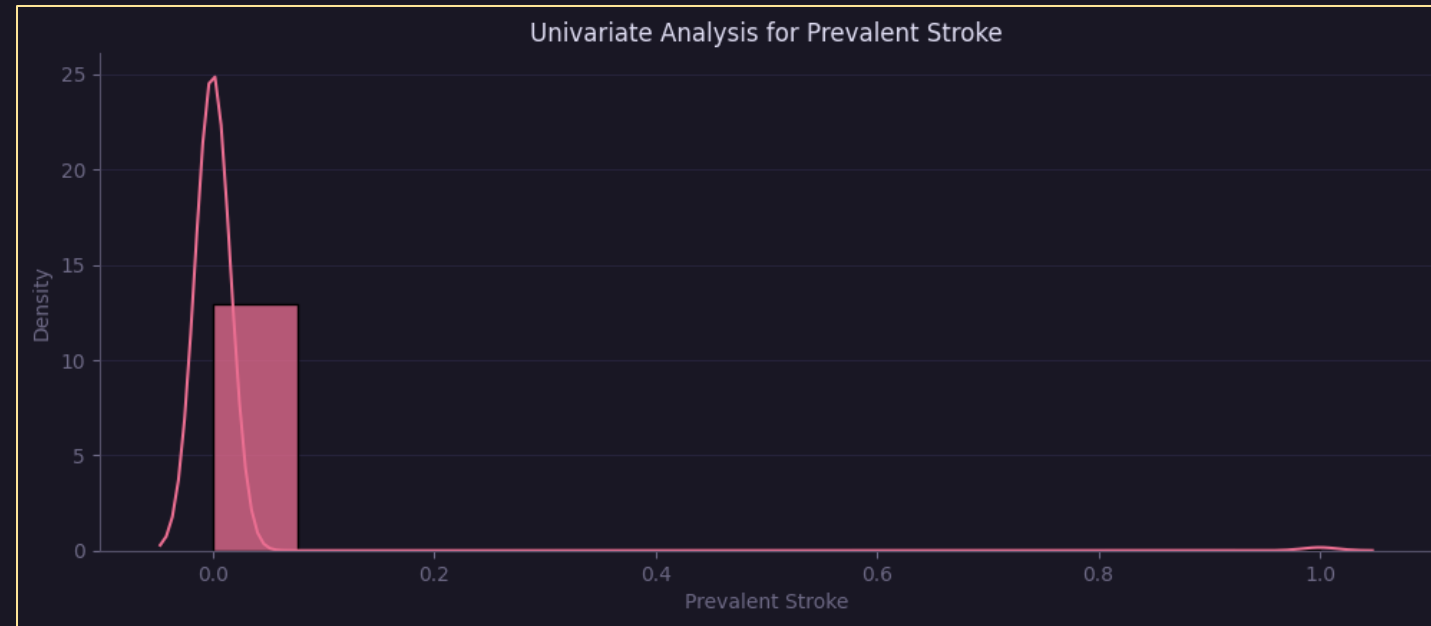
- There are four levels of education whereas the rest categorical features are all binary.
- The number of Smokers and Non-Smokers are almost the same.
- The dataset reveals a significant imbalance in variables, prevalent stroke, diabetes, and blood pressure medication.
 - **Blood Pressure Medications (BP Meds):** The dataset indicates that a relatively small proportion of individuals are taking blood pressure medications.



EDA – Distribution Analysis

Unearthing Insights

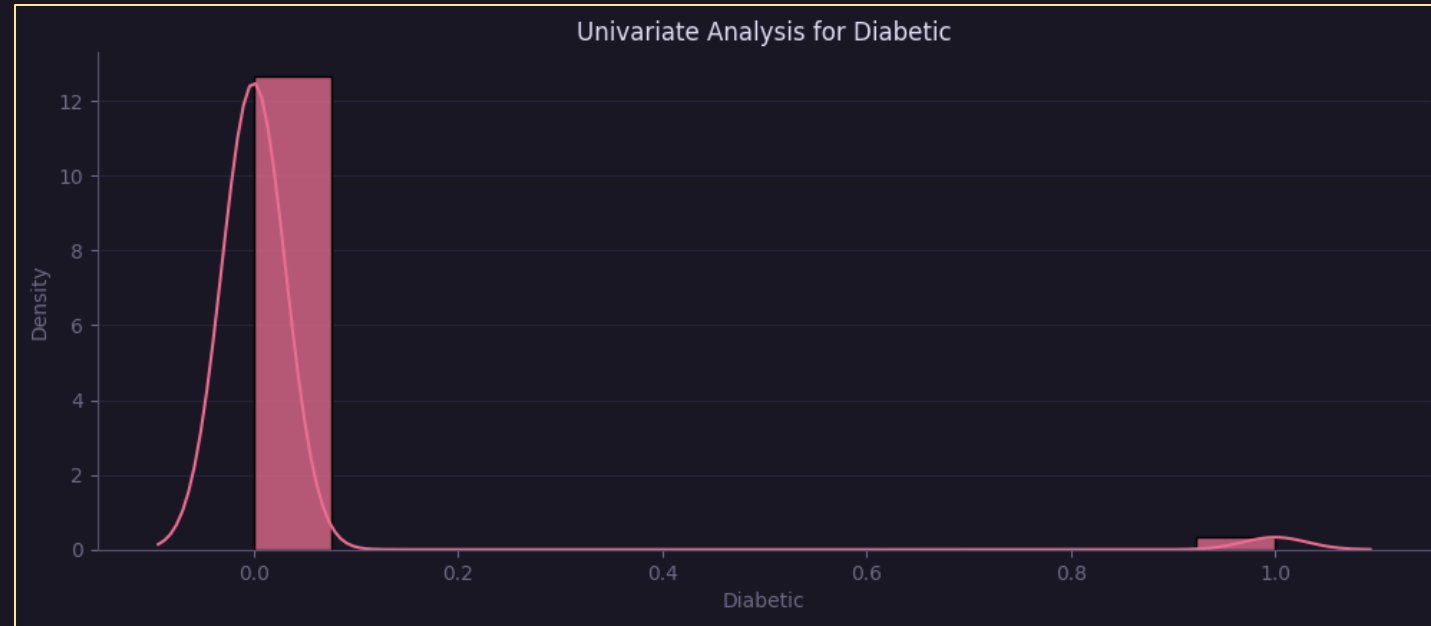
- There are four levels of education whereas the rest categorical features are all binary.
- The number of Smokers and Non-Smokers are almost the same.
- The dataset reveals a significant imbalance in variables, prevalent stroke, diabetes, and blood pressure medication.
 - **Blood Pressure Medications (BP Meds):** The dataset indicates that a relatively small proportion of individuals are taking blood pressure medications.
 - **Prevalent Stroke:** The dataset indicates that there are significantly more individuals without a history of prevalent stroke.



EDA – Distribution Analysis

Unearthing Insights

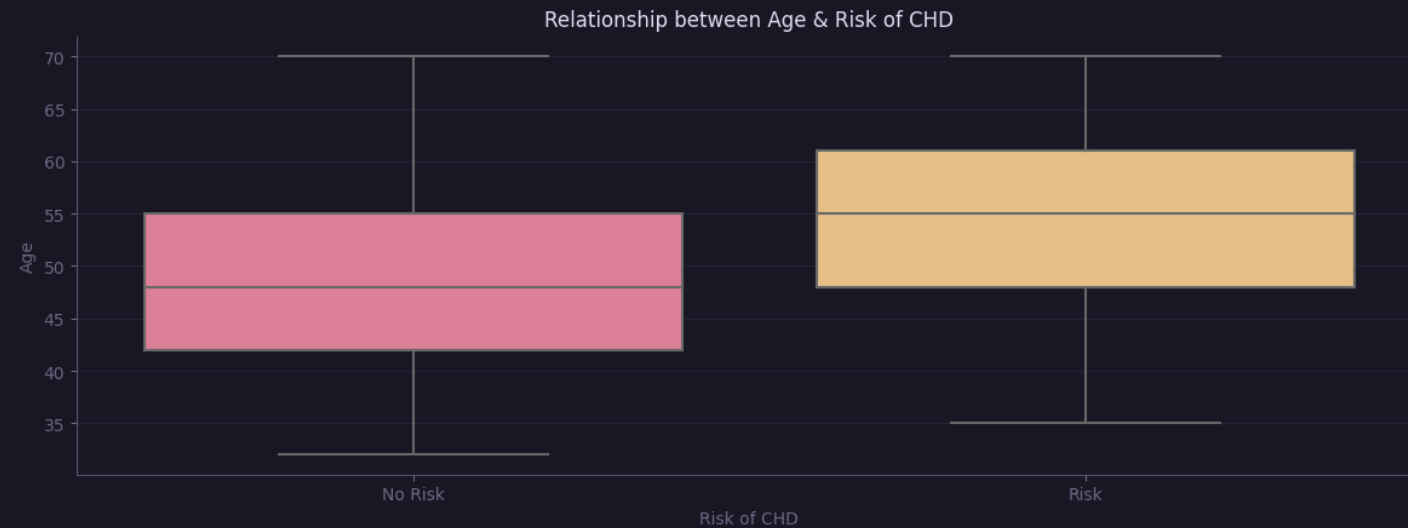
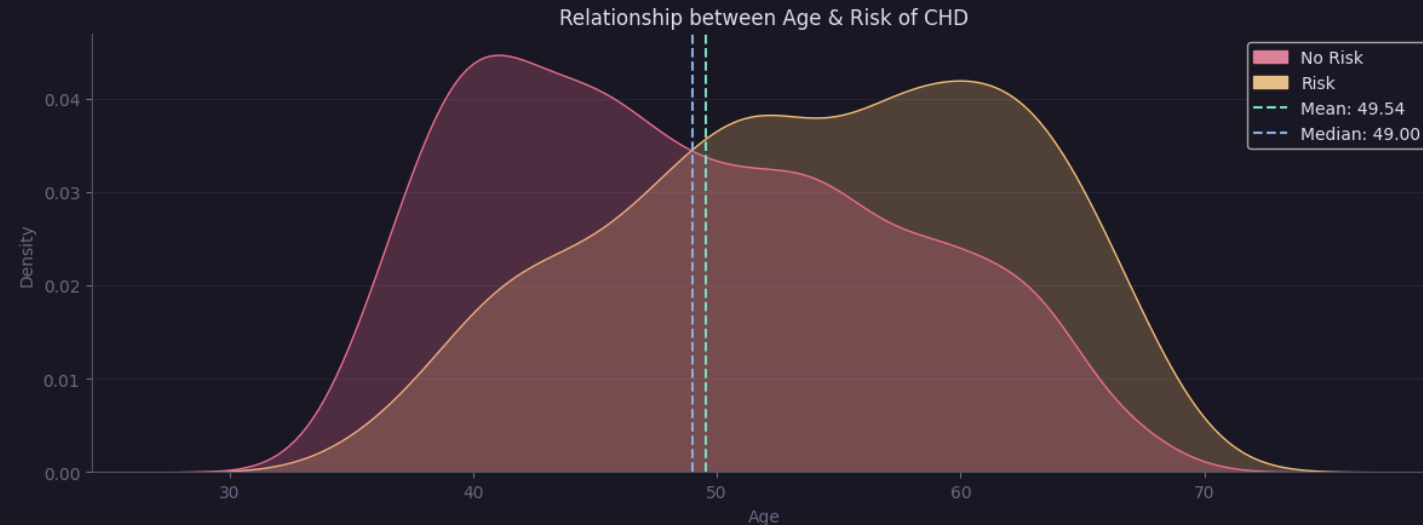
- There are four levels of education whereas the rest categorical features are all binary.
- The number of Smokers and Non-Smokers are almost the same.
- The dataset reveals a significant imbalance in variables, prevalent stroke, diabetes, and blood pressure medication.
 - **Blood Pressure Medications (BP Meds):** The dataset indicates that a relatively small proportion of individuals are taking blood pressure medications.
 - **Prevalent Stroke:** The dataset indicates that there are significantly more individuals without a history of prevalent stroke.
 - **Diabetes:** Similarly, the data suggests that the presence of diabetes is not very common among the individuals in the dataset.



EDA – Relationship Analysis

Unearthing Insights

- Patients between the ages of 50-65 are more likely to develop CHD, while those in the age range of 35-45 predominantly do not experience CHD.



EDA – Relationship Analysis

Unearthing Insights

- Patients between the ages of 50-65 are more likely to develop CHD, while those in the age range of 35-45 predominantly do not experience CHD.
- Patients who didn't smoke experiencing CHD is surprising, while a higher number of cigarettes smoked is associated with an increased risk of CHD.



EDA – Relationship Analysis

Unearthing Insights

- Patients between the ages of 50-65 are more likely to develop CHD, while those in the age range of 35-45 predominantly do not experience CHD.
- Patients who didn't smoke experiencing CHD is surprising, while a higher number of cigarettes smoked is associated with an increased risk of CHD.
- Cholesterol Levels doesn't show significant effect on CHD.



EDA – Relationship Analysis

Unearthing Insights

- Patients between the ages of 50-65 are more likely to develop CHD, while those in the age range of 35-45 predominantly do not experience CHD.
- Patients who didn't smoke experiencing CHD is surprising, while a higher number of cigarettes smoked is associated with an increased risk of CHD.
- Cholesterol Levels doesn't show significant effect on CHD.
- Higher systolic blood pressure (sysBP) is associated with an increased risk of CHD, while individuals with systolic blood pressure around 120 typically have a lower risk of CHD.



EDA – Relationship Analysis

Unearthing Insights

- Patients between the ages of 50-65 are more likely to develop CHD, while those in the age range of 35-45 predominantly do not experience CHD.
- Patients who didn't smoke experiencing CHD is surprising, while a higher number of cigarettes smoked is associated with an increased risk of CHD.
- Cholesterol Levels doesn't show significant effect on CHD.
- Higher systolic blood pressure (sysBP) is associated with an increased risk of CHD, while individuals with systolic blood pressure around 120 typically have a lower risk of CHD.
- Individuals with higher diastolic blood pressure (BP) tend to have an elevated risk of developing CHD, while those with diastolic BP levels within the range of 75-80 generally have a lower risk of CHD.



EDA – Relationship Analysis

Unearthing Insights

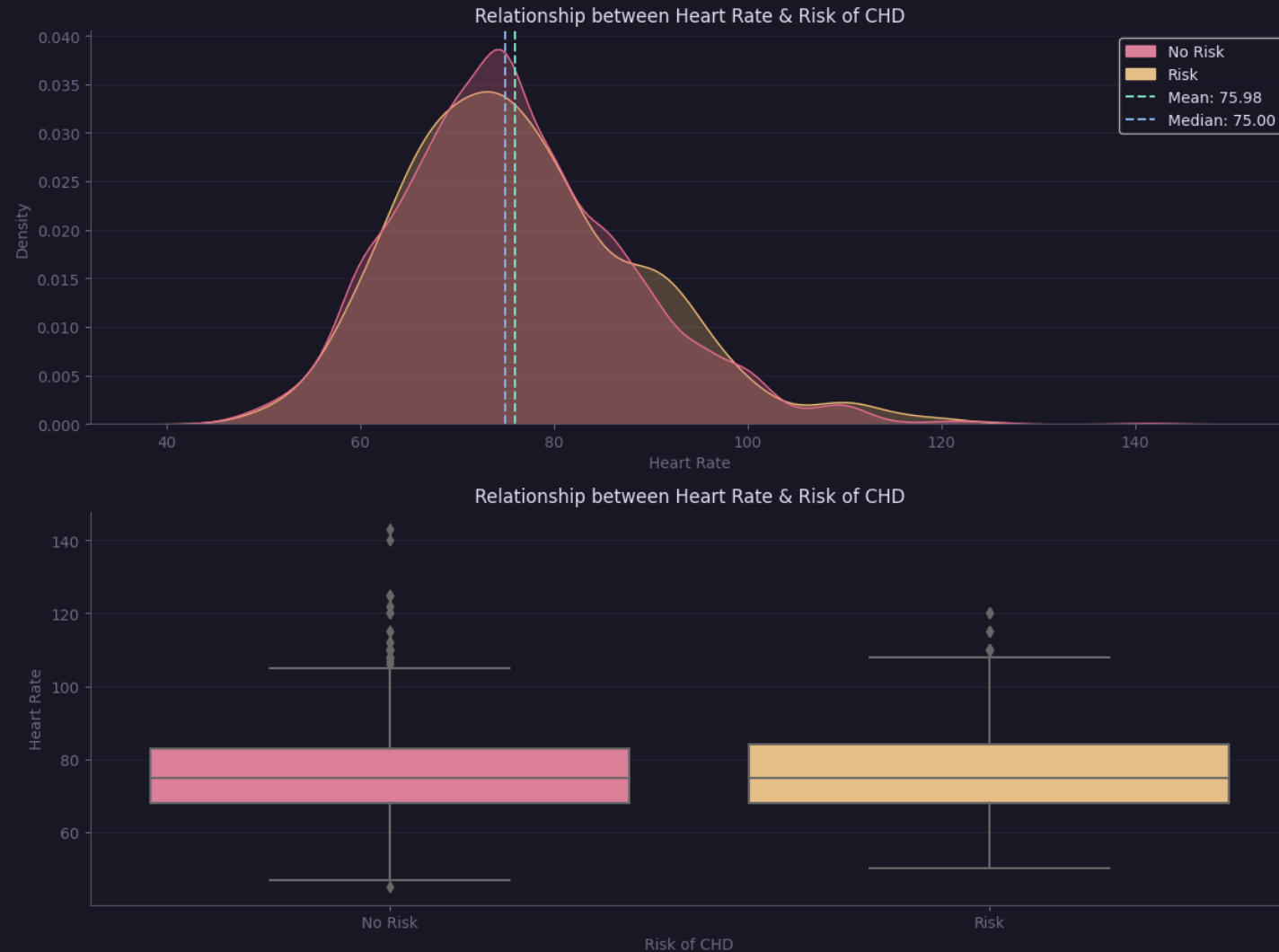
- Patients between the ages of 50-65 are more likely to develop CHD, while those in the age range of 35-45 predominantly do not experience CHD.
- Patients who didn't smoke experiencing CHD is surprising, while a higher number of cigarettes smoked is associated with an increased risk of CHD.
- Cholesterol Levels doesn't show significant effect on CHD.
- Higher systolic blood pressure (sysBP) is associated with an increased risk of CHD, while individuals with systolic blood pressure around 120 typically have a lower risk of CHD.
- Individuals with higher diastolic blood pressure (BP) tend to have an elevated risk of developing CHD, while those with diastolic BP levels within the range of 75-80 generally have a lower risk of CHD.
- It seems BMI doesn't affect chance of getting CHD.



EDA – Relationship Analysis

Unearthing Insights

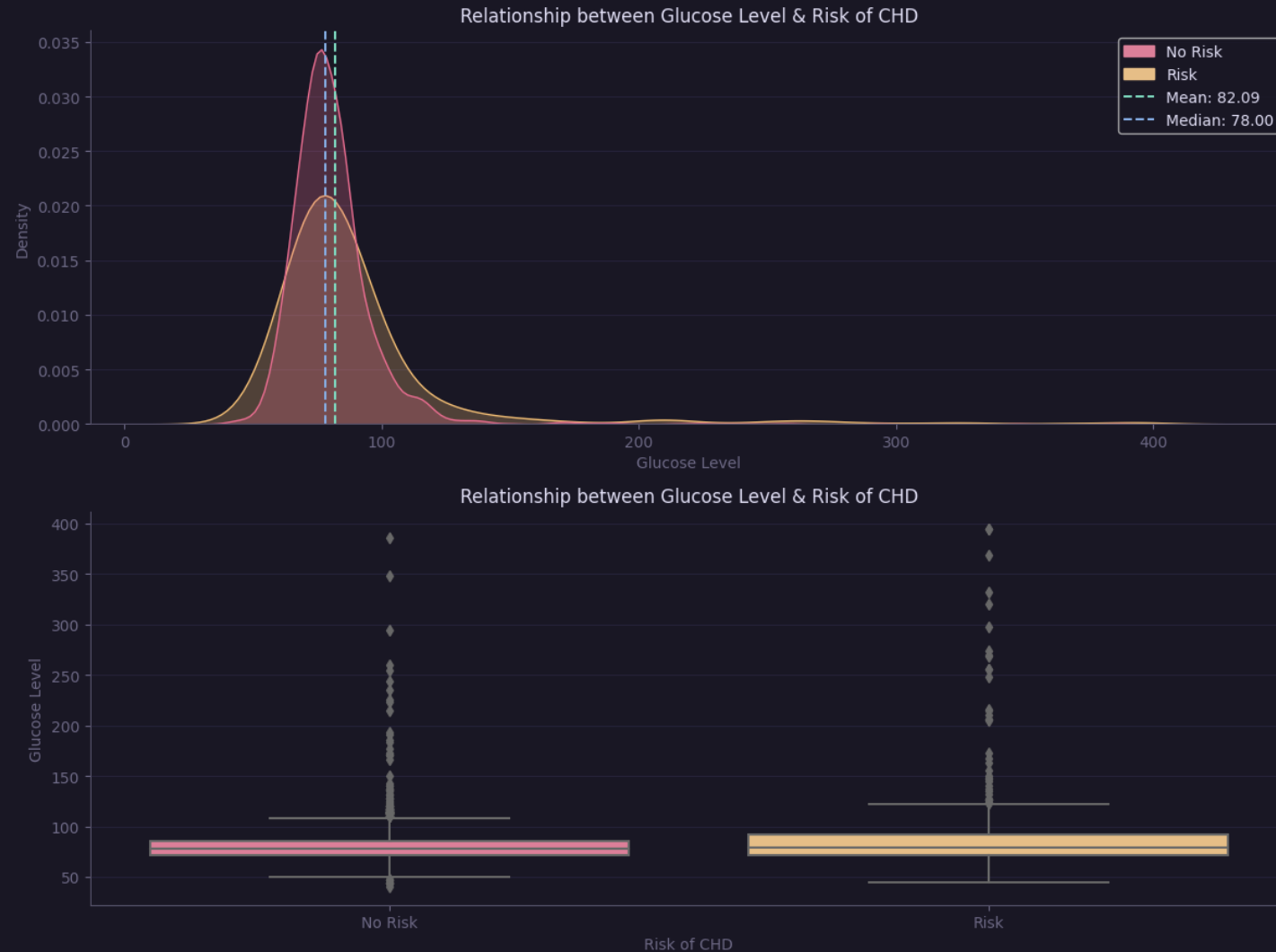
- Patients between the ages of 50-65 are more likely to develop CHD, while those in the age range of 35-45 predominantly do not experience CHD.
- Patients who didn't smoke experiencing CHD is surprising, while a higher number of cigarettes smoked is associated with an increased risk of CHD.
- Cholesterol Levels doesn't show significant effect on CHD.
- Higher systolic blood pressure (sysBP) is associated with an increased risk of CHD, while individuals with systolic blood pressure around 120 typically have a lower risk of CHD.
- Individuals with higher diastolic blood pressure (BP) tend to have an elevated risk of developing CHD, while those with diastolic BP levels within the range of 75-80 generally have a lower risk of CHD.
- It seems BMI doesn't affect chance of getting CHD.
- Maintaining a heart rate within the range of 70-80 beats per minute is generally considered safe; however, deviations either above or below this range may be associated with an increased risk of Coronary Heart Disease (CHD).



EDA – Relationship Analysis

Unearthing Insights

- Patients between the ages of 50-65 are more likely to develop CHD, while those in the age range of 35-45 predominantly do not experience CHD.
- Patients who didn't smoke experiencing CHD is surprising, while a higher number of cigarettes smoked is associated with an increased risk of CHD.
- Cholesterol Levels doesn't show significant effect on CHD.
- Higher systolic blood pressure (sysBP) is associated with an increased risk of CHD, while individuals with systolic blood pressure around 120 typically have a lower risk of CHD.
- Individuals with higher diastolic blood pressure (BP) tend to have an elevated risk of developing CHD, while those with diastolic BP levels within the range of 75-80 generally have a lower risk of CHD.
- It seems BMI doesn't affect chance of getting CHD.
- Maintaining a heart rate within the range of 70-80 beats per minute is generally considered safe; however, deviations either above or below this range may be associated with an increased risk of Coronary Heart Disease (CHD).
- The impact of glucose level on CHD appears to be limited.

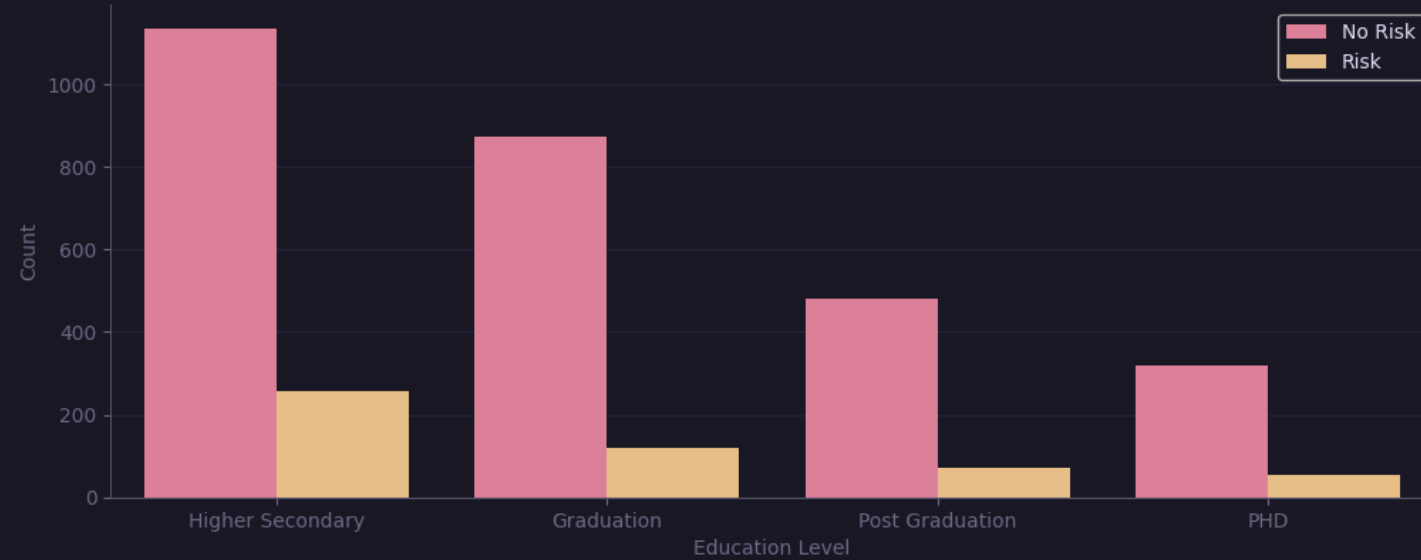


EDA – Relationship Analysis

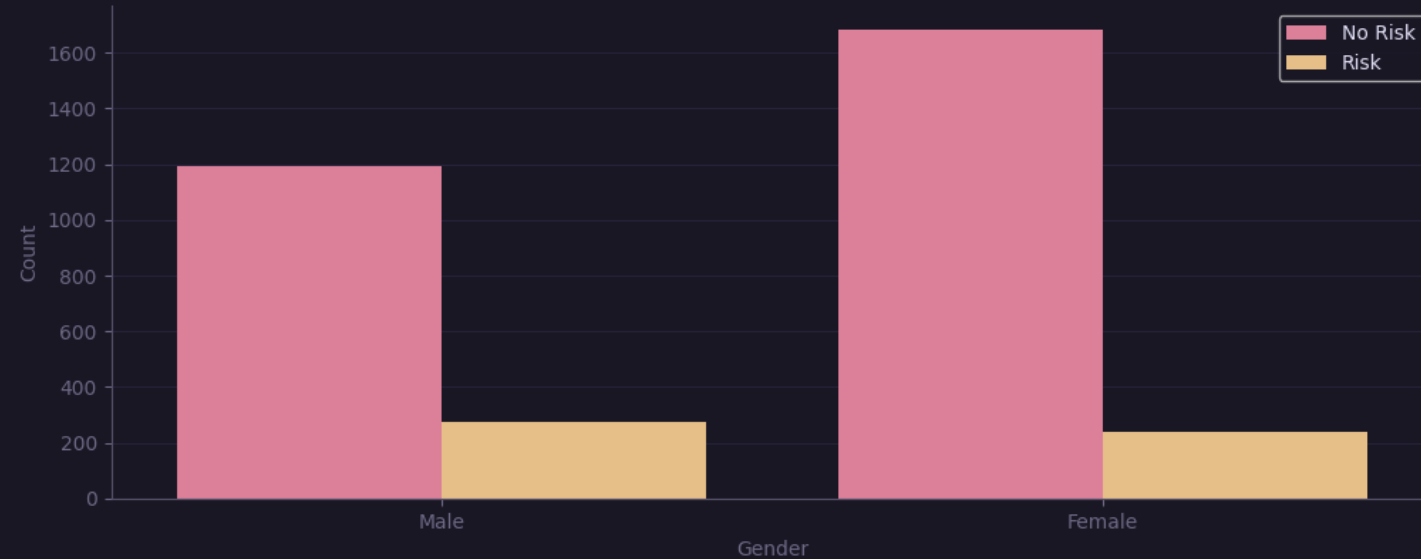
Unearthing Insights

- Gender disparity in CHD risk: Males > Females.

Relationship between Education Level & Risk of CHD



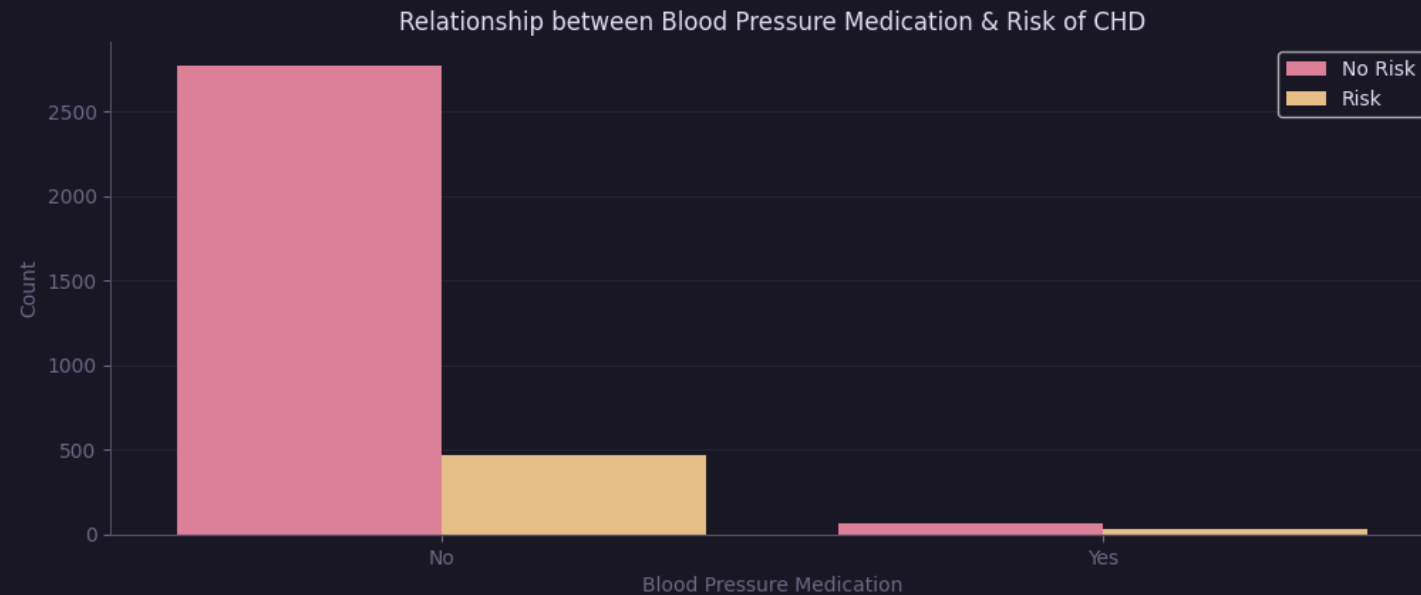
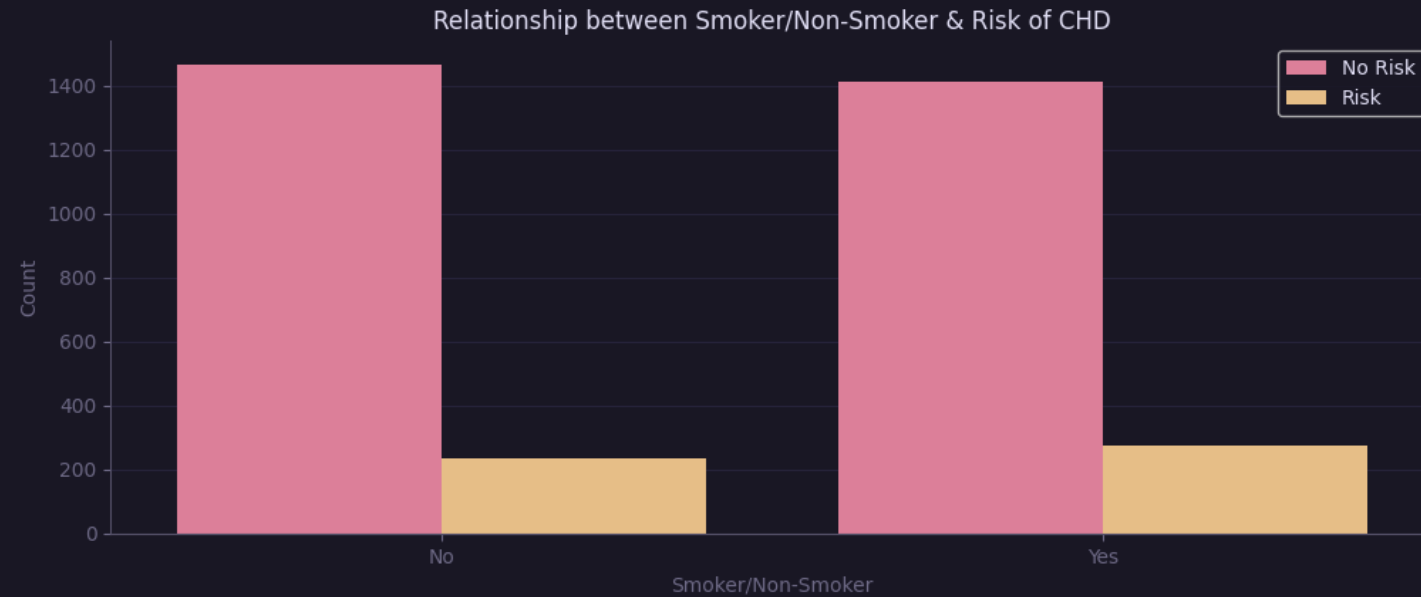
Relationship between Gender & Risk of CHD



EDA – Relationship Analysis

Unearthing Insights

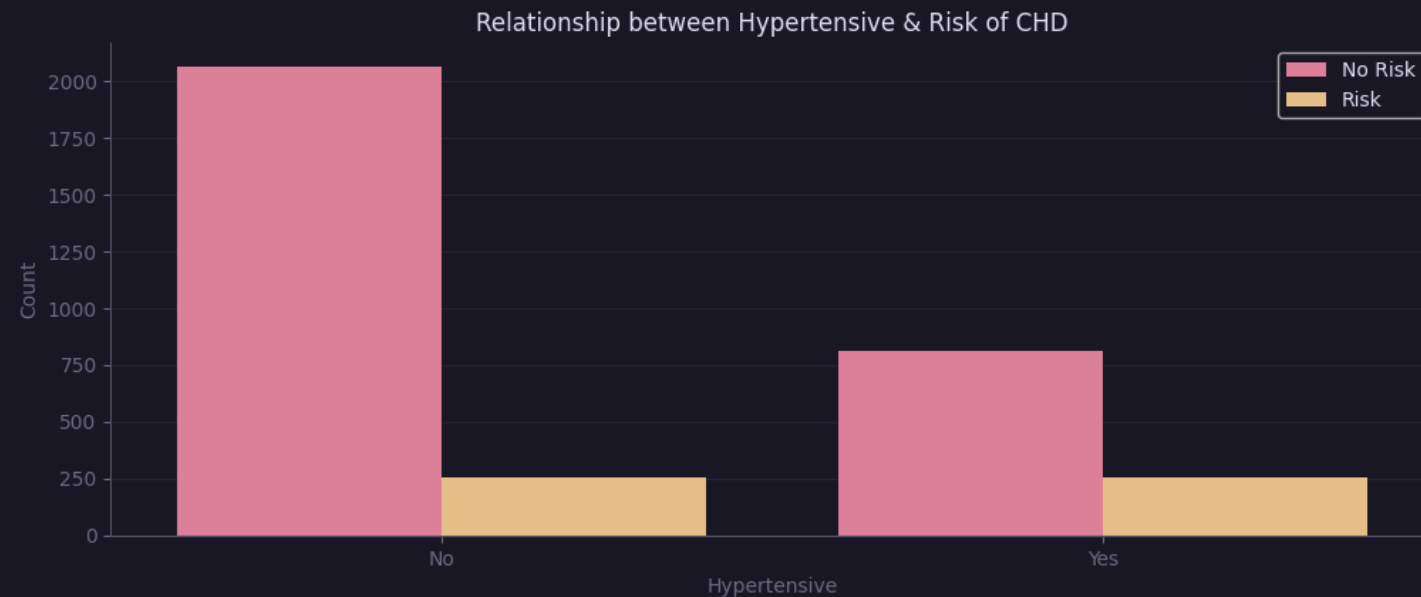
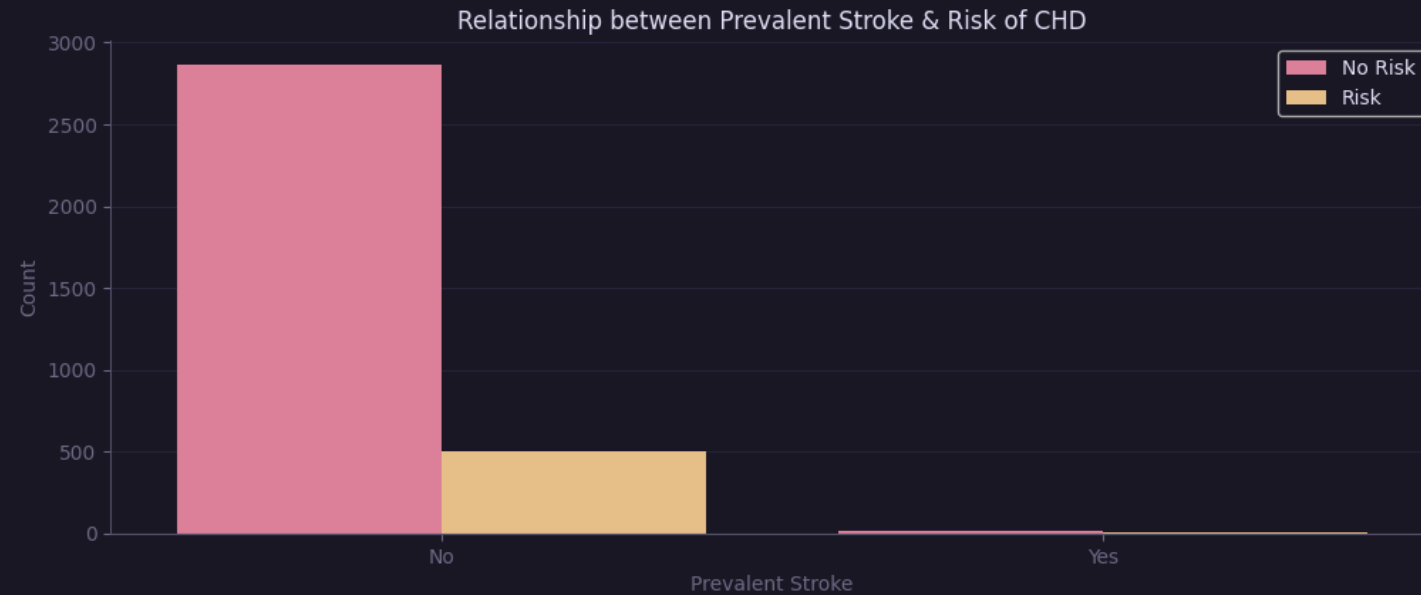
- Gender disparity in CHD risk: Males > Females.
- Smokers at risk of CHD tend to be around 50 years old, while non-smokers with risk are typically 65-70 years old, and most smokers with no risk are approximately 40 years old.
- Around 50-60% of individuals on BP medication appear to develop CHD.



EDA – Relationship Analysis

Unearthing Insights

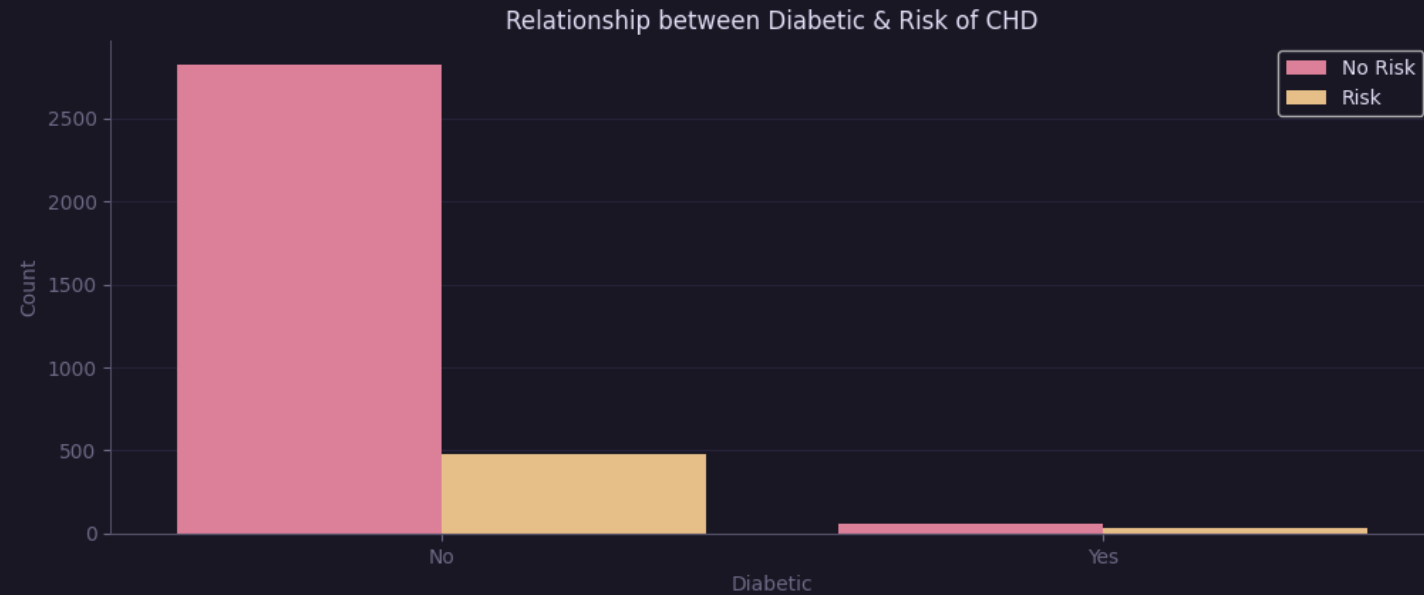
- Gender disparity in CHD risk: Males > Females.
- Smokers at risk of CHD tend to be around 50 years old, while non-smokers with risk are typically 65-70 years old, and most smokers with no risk are approximately 40 years old.
- Around 50-60% of individuals on BP medication appear to develop CHD.
- It appears that 90% of stroke patients also develop CHD.
- Elevated hypertension rates correlate with increased incidence of CHD.



EDA – Relationship Analysis

Unearthing Insights

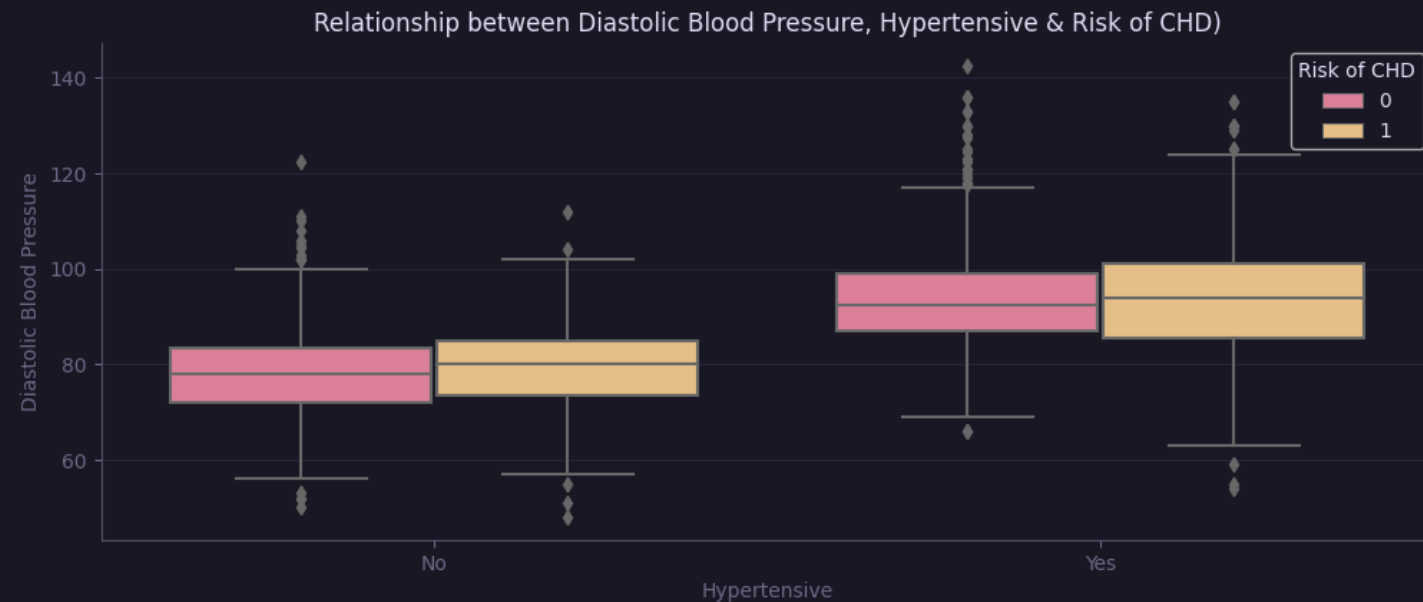
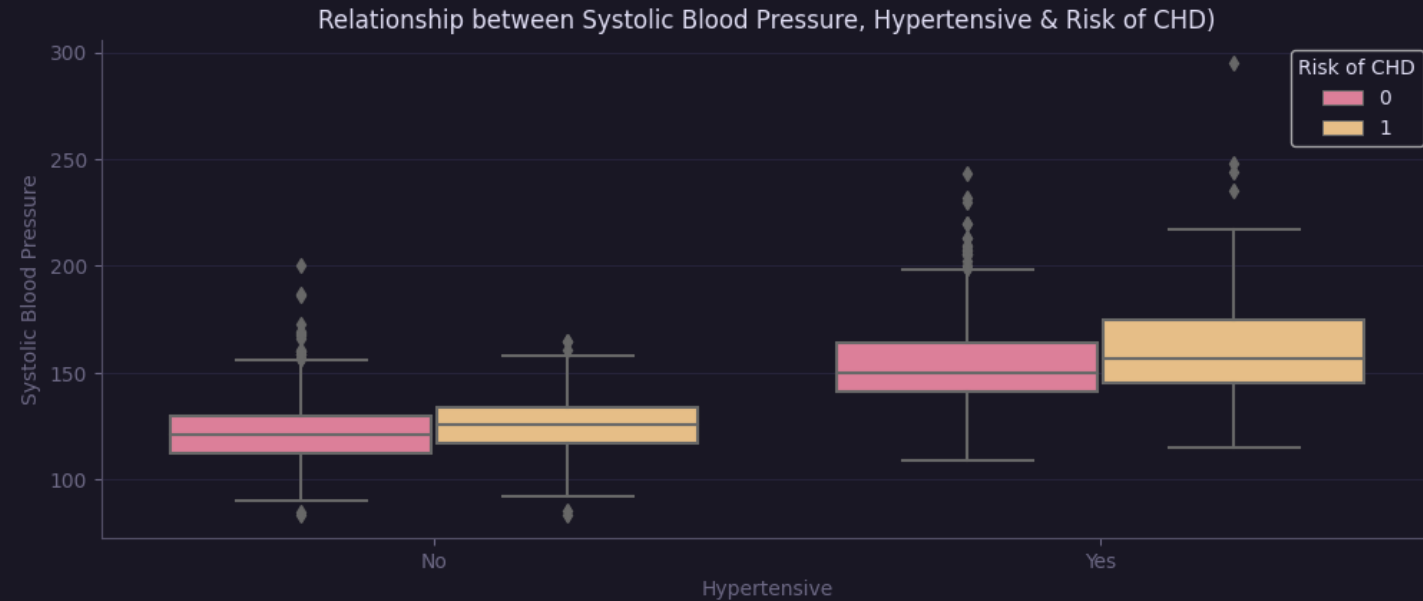
- Gender disparity in CHD risk: Males > Females.
- Smokers at risk of CHD tend to be around 50 years old, while non-smokers with risk are typically 65-70 years old, and most smokers with no risk are approximately 40 years old.
- Around 50-60% of individuals on BP medication appear to develop CHD.
- It appears that 90% of stroke patients also develop CHD.
- Elevated hypertension rates correlate with increased incidence of CHD.
- Approximately 60-80% of diabetic patients develop CHD.



EDA – Relationship Analysis

Unearthing Insights

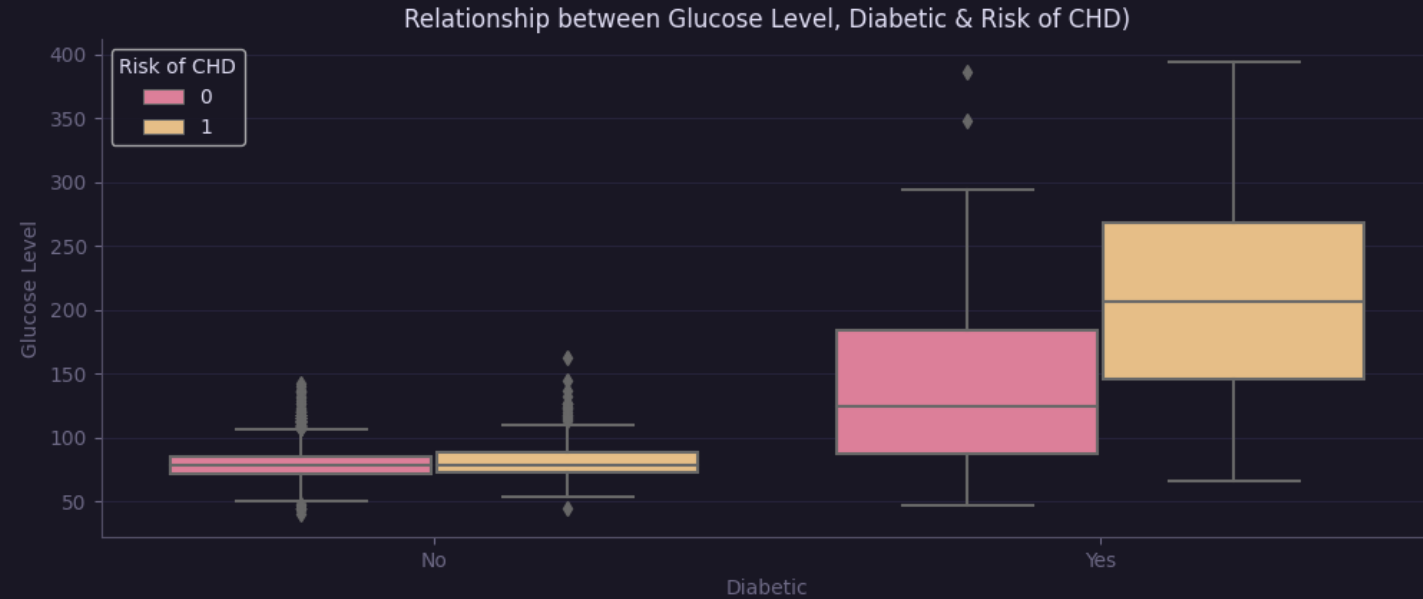
- Elevated systolic and diastolic blood pressure increase the risk of hypertension and consequently, the risk of coronary heart disease (CHD).



EDA – Relationship Analysis

Unearthing Insights

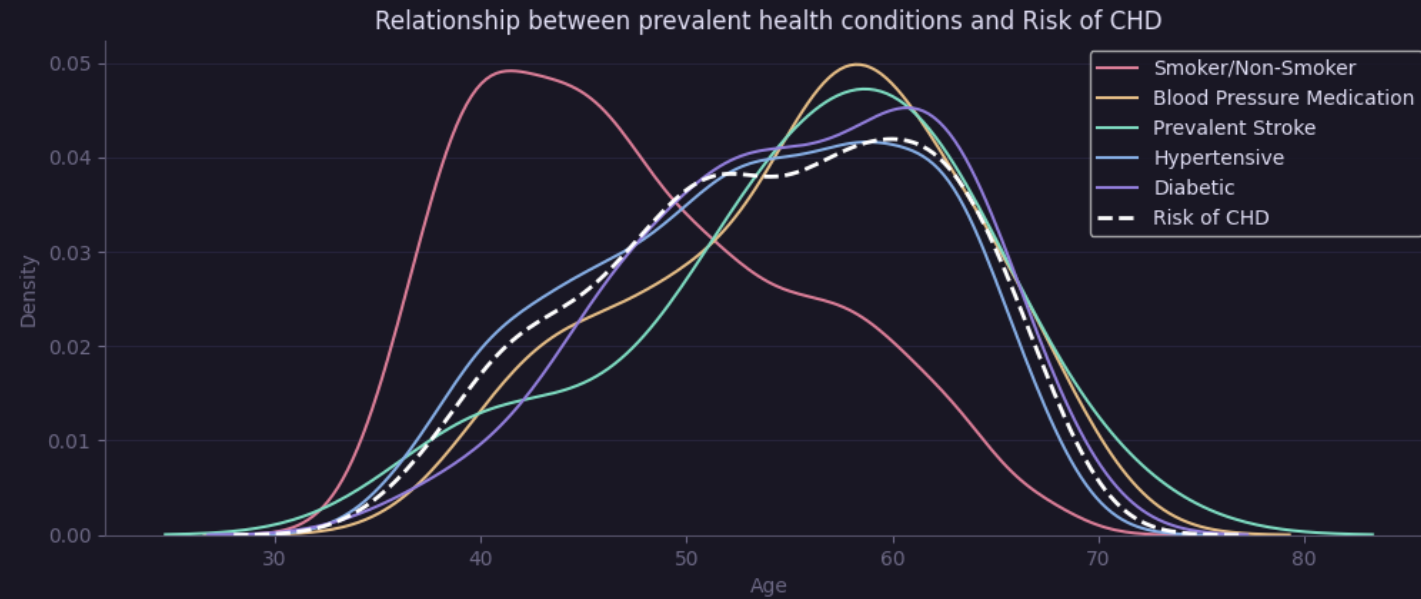
- Elevated systolic and diastolic blood pressure increase the risk of hypertension and consequently, the risk of coronary heart disease (CHD).
- Diabetic patients with glucose levels between 200-400 have an elevated risk of developing CHD.



EDA – Relationship Analysis

Unearthing Insights

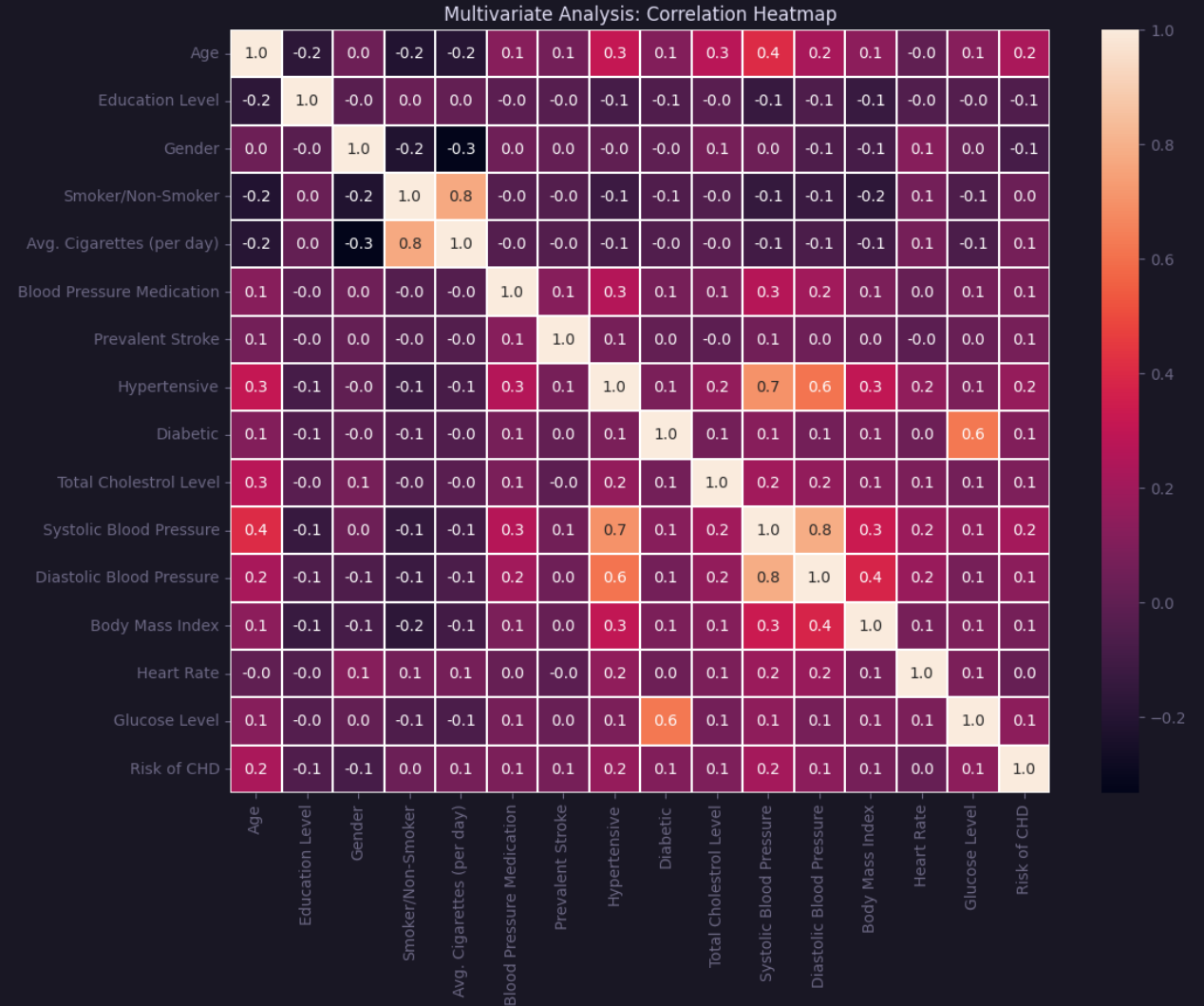
- As individuals grow older, they often choose to quit smoking, but they also face a **higher likelihood of experiencing prevalent health issues** such as stroke, hypertension, the need for blood pressure medication, diabetes, and as a result, an increased risk of coronary heart disease (CHD).



EDA – Correlation Analysis

Unearthing Insights

- In contrast to all the different data points, the correlation coefficient between Education and the target variable is notably low and, in fact, negative.
- Systolic BP and Diastolic BP, as well as Smoking and Cigs./day, exhibit a strong positive correlation, with coefficients close to 0.8.
- Additionally, Systolic BP, Diastolic BP, with Hypertension, and Glucose Level and Diabetes, show a moderate positive correlation, with coefficients around 0.6.



Preprocessing & Baseline Modelling

Missing Value Handling:

- The dataset initially contained 510 null values. Null values do not necessarily indicate non-existence but rather signify unknown or uncollected data.
- In the medical domain, null values often occur when certain observations are not collected due to low clinical relevance or unavailability of specific tests.
- Imputing missing medical data can be challenging as the efficacy of such imputations for disease detection is unclear. Therefore, in our approach, observations with missing values were removed.
- After removing null values, our dataset size stands at (2927, 16), ensuring data quality and reliability.

Baseline Modeling with Logistic Regression:

- We chose Logistic Regression as our initial model for its unique advantages:
 - Transparency: Logistic Regression offers transparency by allowing us to examine feature coefficients. This insight helps us understand how each feature influences predictions, crucial for domain experts' interpretability.
 - Efficiency: Its computational efficiency is paramount, especially when dealing with large datasets. This efficiency accelerates our data analysis and modeling processes.

| Metric | Value |
|-----------|-------|
| Accuracy | 0.85 |
| Precision | 0.7 |
| Recall | 0.08 |
| F1 Score | 0.14 |
| ROC AUC | 0.67 |

Second-Phase Baseline Modeling

Exploring Algorithmic Diversity for Enhanced Classification

To enhance classification performance, we addressed the class imbalance by employing Synthetic Minority Over-sampling Technique (SMOTE).

Algorithm Selection:

Logistic Regression (Interpretability)

Logistic Regression was retained for its interpretability and computational efficiency. It seamlessly aligned with our initial baseline and benefited from the enhanced data balance achieved through SMOTE.

Random Forest (Ensemble Approach)

Random Forest emerged as a robust choice, excelling with the balanced dataset post-SMOTE. Its ensemble nature significantly boosted performance metrics, making it a compelling candidate.

K-Nearest Neighbors (Non-Linearity)

KNN introduced non-linearity into our modeling, capturing intricate data relationships while capitalizing on the balanced dataset produced by SMOTE.

By incorporating SMOTE oversampling to address class imbalance, we leveled the playing field for our algorithms.

| Algorithm | Accuracy | Precision | Recall | F1 Score | ROC AUC |
|---------------------|----------|-----------|--------|----------|---------|
| Logistic Regression | 0.74 | 0.76 | 0.72 | 0.74 | 0.8 |
| Random Forest | 0.87 | 0.88 | 0.87 | 0.88 | 0.94 |
| K-Nearest Neighbors | 0.82 | 0.76 | 0.96 | 0.85 | 0.9 |

Model Building and Evaluation

Scaling:

To ensure uniformity in feature scales, we applied `StandardScaler` to both the features (X) and the target variable (y). This preprocessing step is vital for algorithms sensitive to feature magnitudes.

Classifier Initialization:

We initialized several classifiers, each with hyperparameters tuning and five-fold cross validation, to explore a variety of algorithmic approaches for our classification task. The chosen algorithms include XGBoost, Random Forest, Gradient Boosting, Logistic Regression, K-Nearest Neighbors, Naïve Bayes, Decision Tree, LightGBM, and Support Vector Classifier (SVC).

| Classifier | Accuracy | Precision | Recall | F1 Score | ROC AUC |
|---------------------|----------|-----------|--------|----------|---------|
| XGBoost | 0.88 | 0.88 | 0.9 | 0.89 | 0.95 |
| Random Forest | 0.88 | 0.89 | 0.88 | 0.88 | 0.95 |
| Gradient Boosting | 0.87 | 0.88 | 0.88 | 0.88 | 0.94 |
| Logistic Regression | 0.73 | 0.75 | 0.74 | 0.74 | 0.82 |
| K-Nearest Neighbors | 0.82 | 0.79 | 0.89 | 0.84 | 0.91 |
| Naïve Bayes | 0.71 | 0.73 | 0.71 | 0.72 | 0.76 |
| Decision Tree | 0.79 | 0.8 | 0.79 | 0.8 | 0.79 |
| LightGBM | 0.88 | 0.89 | 0.89 | 0.89 | 0.95 |
| SVC | 0.8 | 0.82 | 0.79 | 0.8 | 0.87 |

Model Building and Evaluation

Selected Model (XGBoost):

Based on the evaluation results, XGBoost emerged as the most promising model for our classification task, achieving the highest Recall, F1 Score, and ROC AUC among the classifiers.

Performance Metrics for Selected Model (XGBoost):

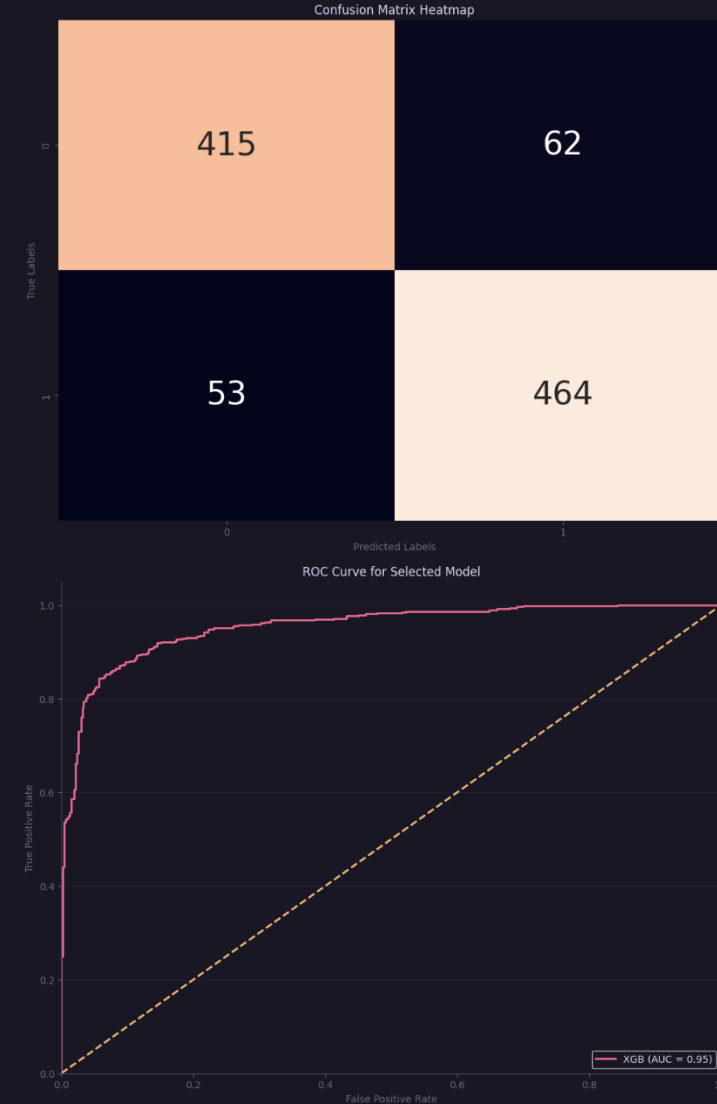
- Recall: 0.90
- F1 Score: 0.89
- ROC AUC: 0.95

Comparison with Baseline (XGBoost vs. Baseline Model (Random Forest)):

- Recall Improvement: 3%
- F1 Score Improvement: 1%
- ROC AUC Improvement: 1%

Justification:

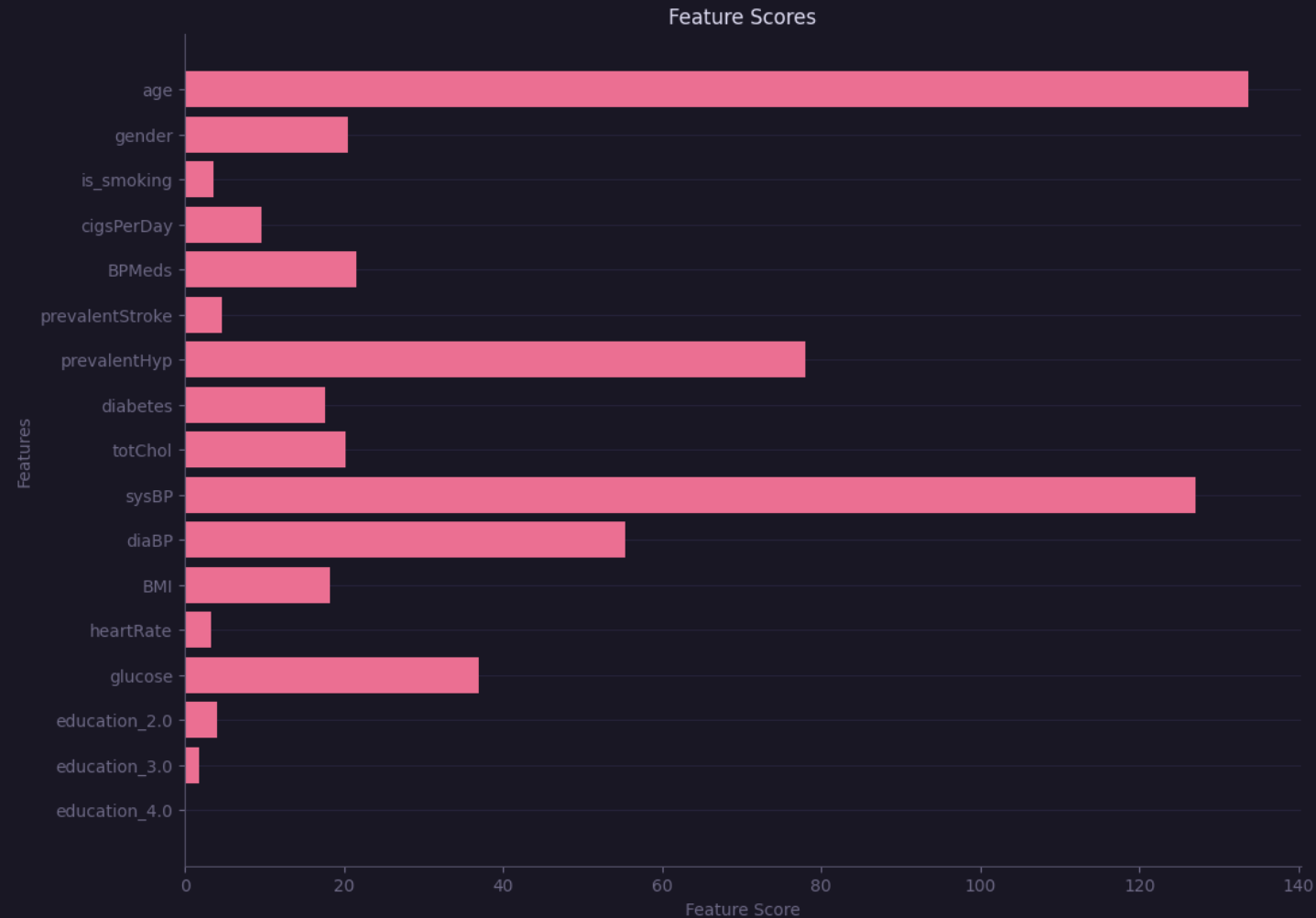
- Our choice of XGBoost as the selected model is validated by its superior performance in Recall, F1 Score, and ROC AUC, highlighting its potential to excel in medical classification tasks.
- The comparison with the baseline model (Random Forest) underscores the improvements achieved in the metrics most relevant to the medical domain, reinforcing our decision to proceed with XGBoost for further optimization and fine-tuning.



Model Building and Evaluation

Unveiling the Path to Enhanced Predictive Power

- We employed 'SelectKBest' algorithm with to identify the top features with the highest F-statistic scores, for enhancing model predictive power by focusing on the most relevant features.



Model Enhancement

Performance Metrics for Final Model (XGBoost with Top 10 Feature Selection):

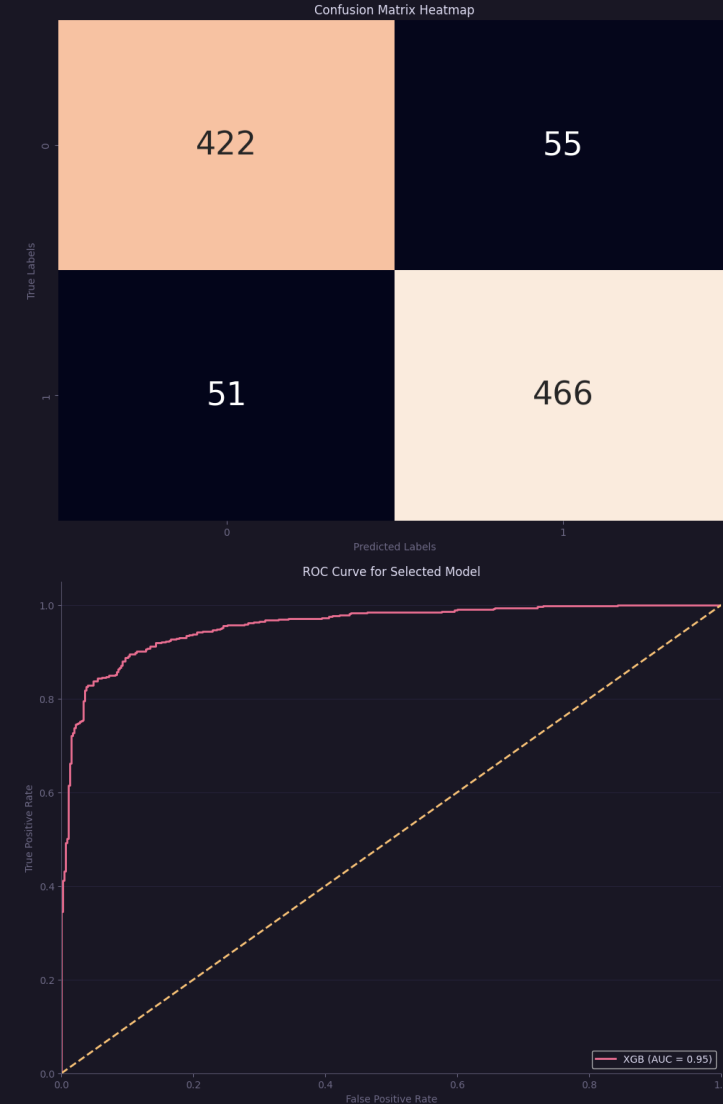
- Accuracy: 0.89
- Precision: 0.89
- Recall: 0.90
- F1 Score: 0.90
- ROC AUC: 0.96

Comparison:

- The final model, incorporating top 10 feature selection with XGBoost, has achieved outstanding results across all key performance metrics, showcasing its superiority in terms of Accuracy, Precision, Recall, F1 Score, and ROC AUC.
- This represents a significant advancement compared to our initial model, reaffirming the effectiveness of feature selection and algorithmic refinement in enhancing model performance for our critical classification task.

Justification:

Our pursuit of feature selection and algorithm optimization has culminated in a model that excels in both precision and recall, aligning with the stringent requirements of the medical domain.



"Thank you for joining us on this journey towards a heart-healthy future. Together, we can make a difference."