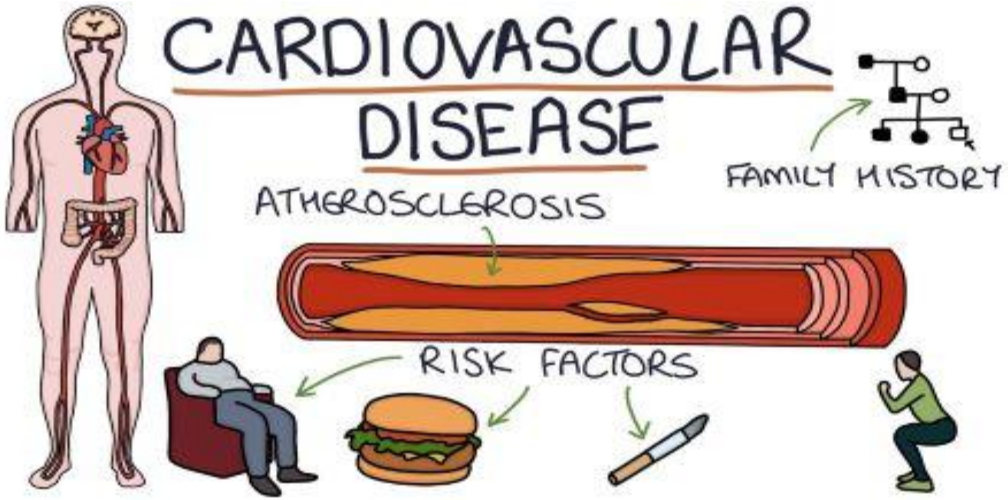
Capstone Project

Cardiovascular Risk Prediction

By- Samiksha Bandbuche

Problem Statement:

The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts.
The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD).
The dataset provides the patients' information. It includes over 4,000 records and 15 attributes.



Points to be discussed:

- Introduction
- Data Summary
- Exploratory Data Analysis
- Correlation Analysis
- Model Used
- Conclusion

Introduction:

Cardiovascular disease is a group of diseases affecting your heart and blood vessels. These diseases can affect one or many parts of your heart and/or blood vessels. A person may be symptomatic or asymptomatic .

Cardiovascular disease includes heart or blood vessel issues, including:

- Narrowing of the blood vessels in your heart, other organs or throughout your body.
- Heart and blood vessel problems present at birth.
- Heart valves that aren’t working right.
- Irregular heart rhythms.

Data Summary:

Data Set Name- data_cardiovascular_risk

Dataset -

Rows-3390

Columns-17

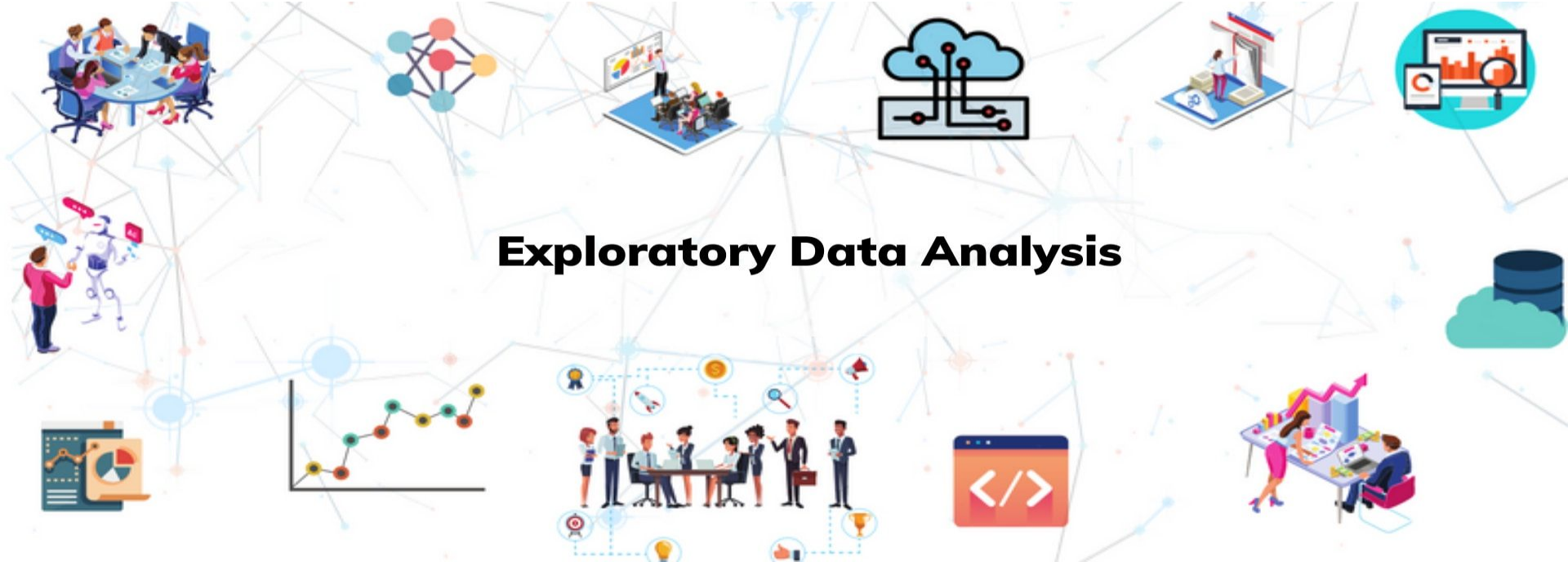
	id	age	education	sex	is_smoking	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	0	64	2.0	F	YES	3.0	0.0	0	0	0	221.0	148.0	85.0	NaN	90.0	80.0	1
1	1	36	4.0	M	NO	0.0	0.0	0	1	0	212.0	168.0	98.0	29.77	72.0	75.0	0
2	2	46	1.0	F	YES	10.0	0.0	0	0	0	250.0	116.0	71.0	20.35	88.0	94.0	0
3	3	50	1.0	M	YES	20.0	0.0	0	1	0	233.0	158.0	88.0	28.26	68.0	94.0	1
4	4	64	1.0	F	YES	30.0	0.0	0	0	0	241.0	136.5	85.0	26.42	70.0	77.0	0

Variables:

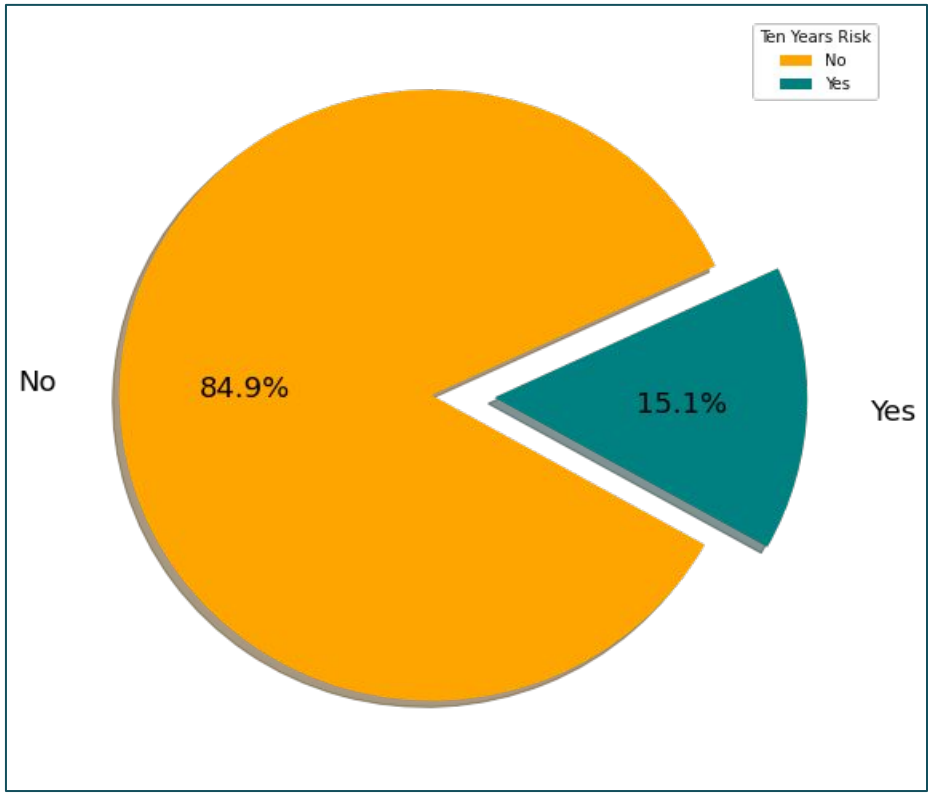
- **Sex:** male or female
- **Age:** Age of the patient
- **is_smoking:** whether or not the patient is a current smoker
- **Cigs Per Day:** the number of cigarettes that the person smoked on average in one day.
- **BP Meds:** whether or not the patient was on blood pressure medication
- **Prevalent Stroke:** whether or not the patient had previously had a stroke
- **Prevalent Hyp:** whether or not the patient was hypertensive
- **Diabetes:** whether or not the patient had diabetes
- **Tot Chol:** total cholesterol level
- **Sys BP:** systolic blood pressure
- **Dia BP:** diastolic blood pressure
- **BMI:** Body Mass Index
- **Heart Rate:** heart rate
- **Glucose:** glucose level
- **TenYearCHD:**10-year risk of coronary heart disease CHD(binary: “1”, means “Yes”, “0” means “No”) - DV

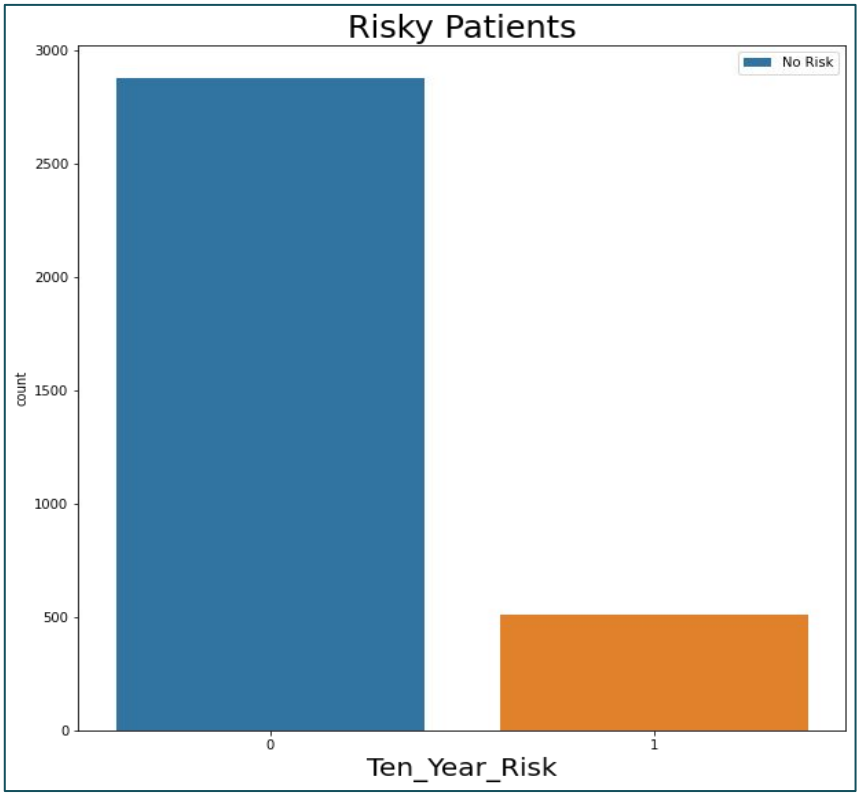
Exploratory Data Analysis:

Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics, often with visual methods.

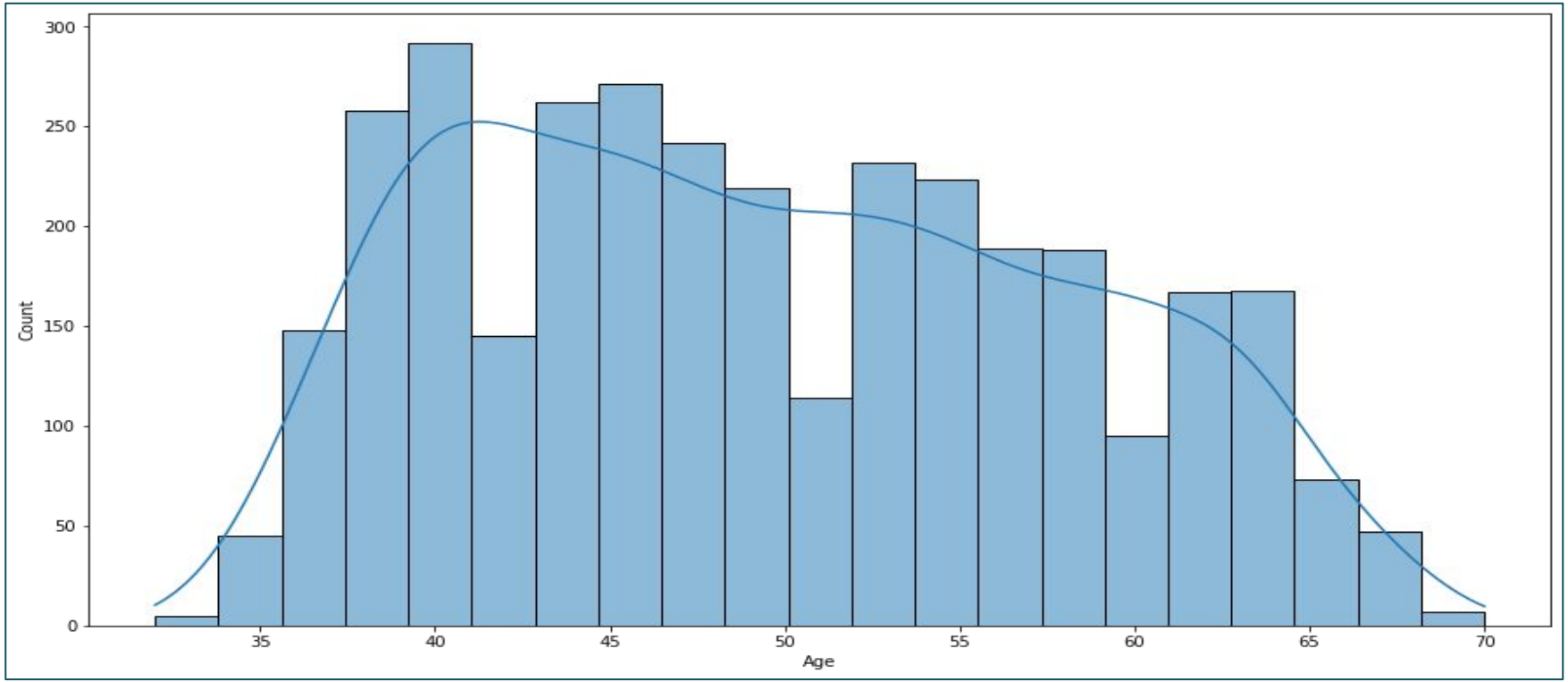


How many patients have risk of CHD?

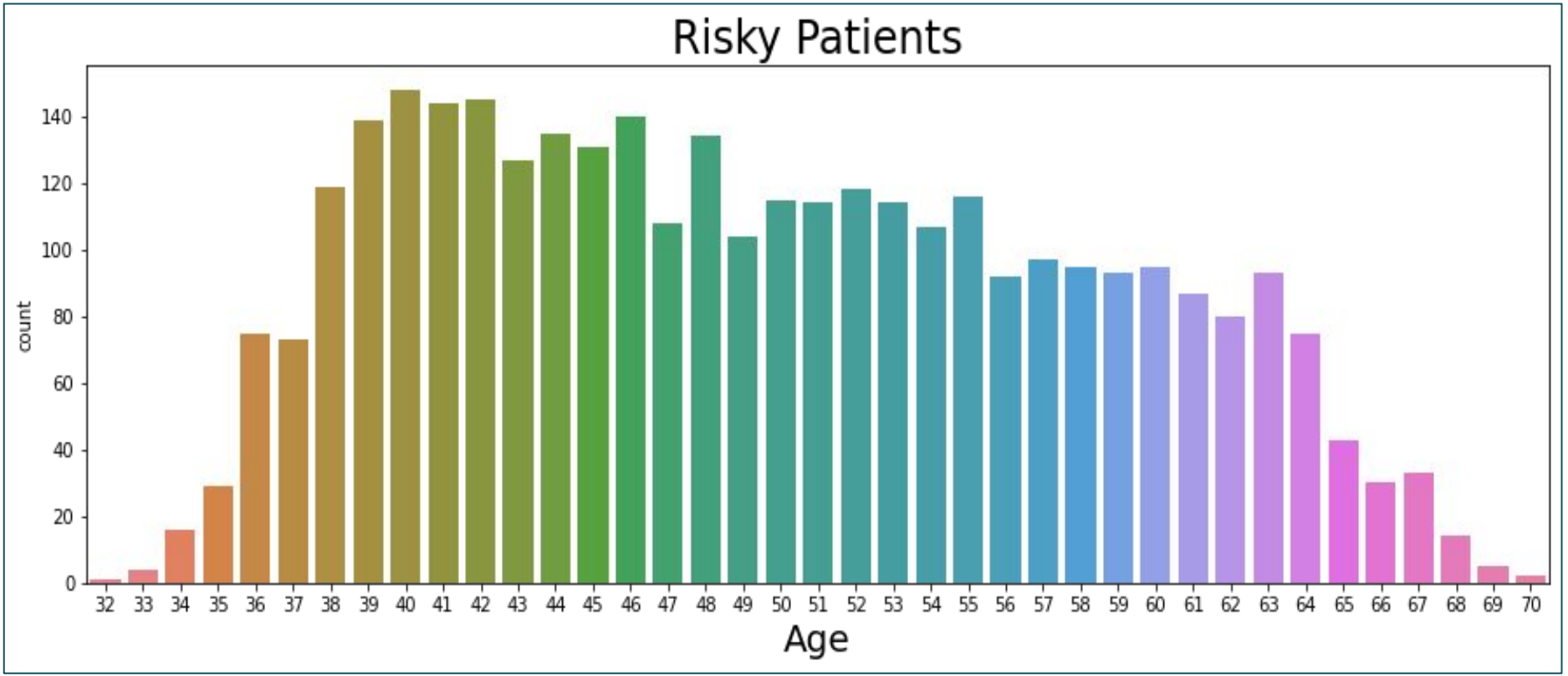




Which Age group is more vulnerable to CHD?

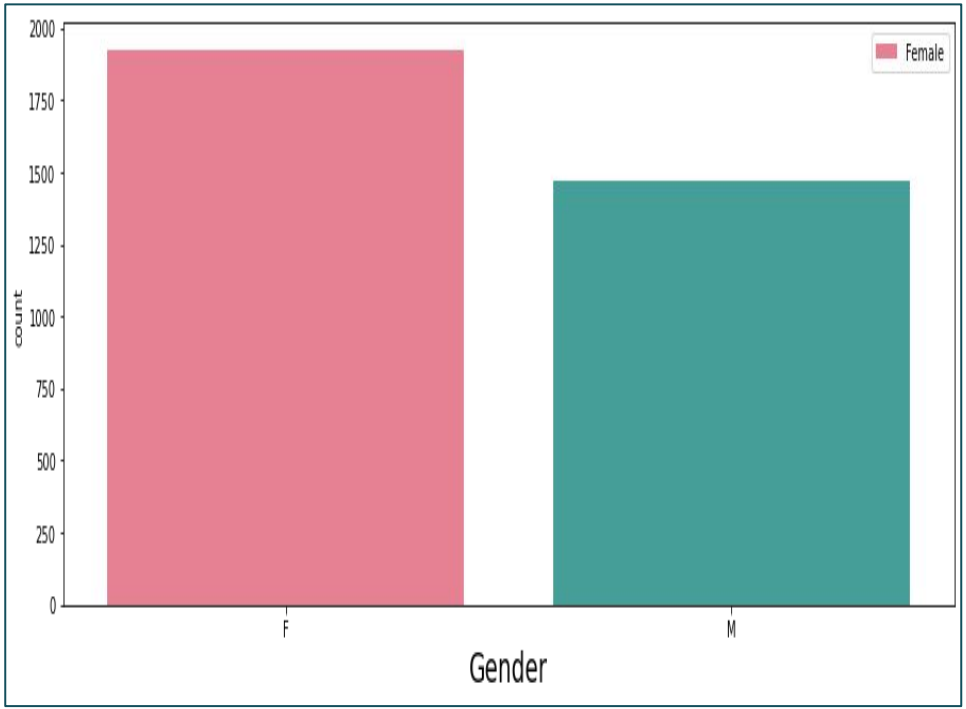


Patients at Risk With Respect To Age:

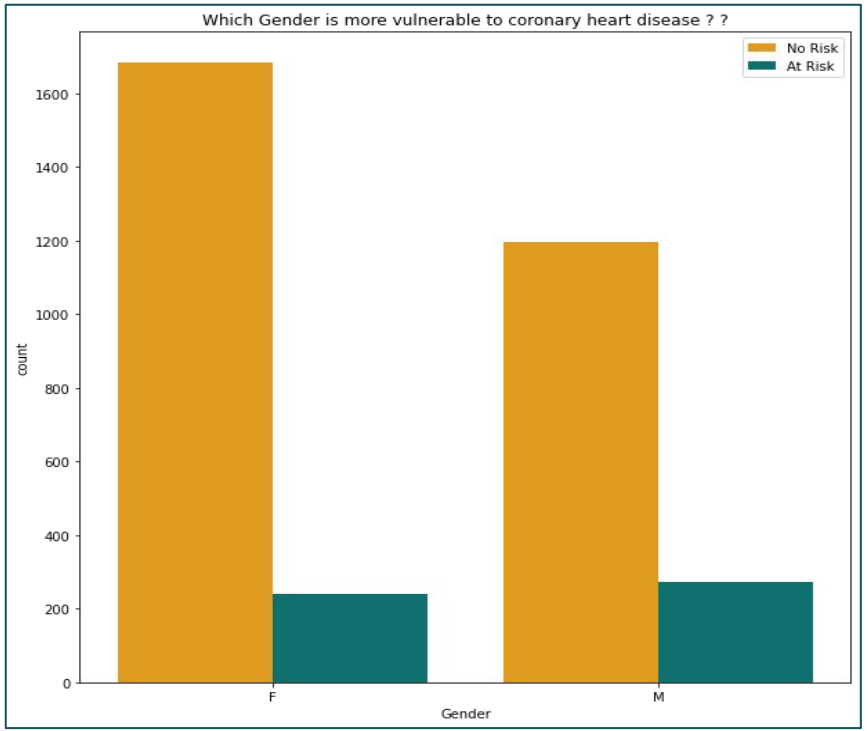


Which Gender is more vulnerable to coronary heart disease ?

Gender wise Distribution

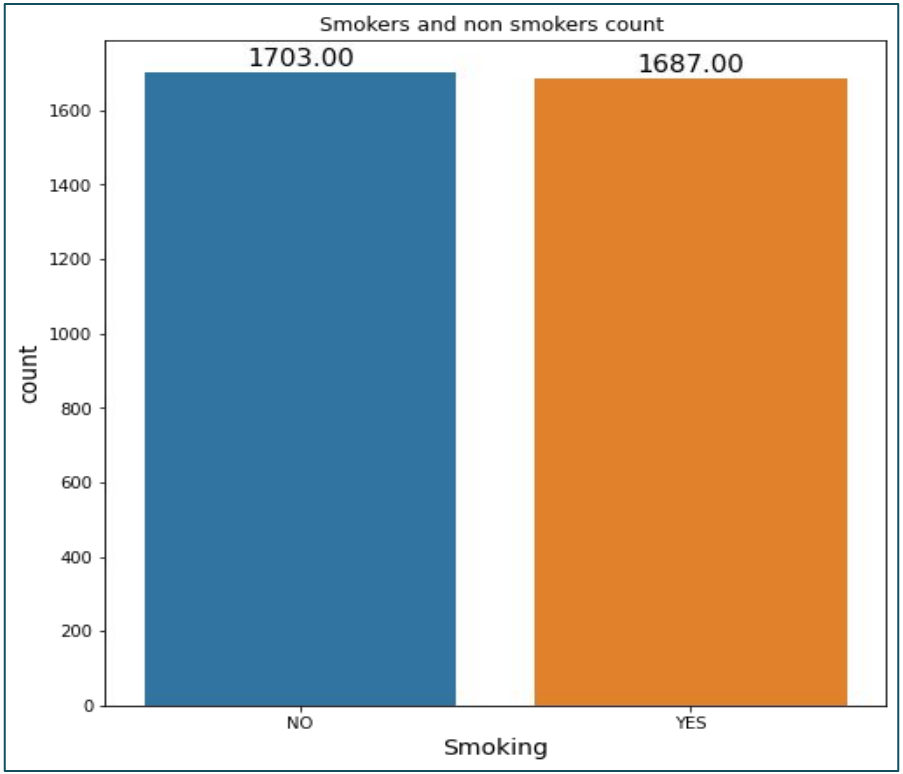


Patients at Risk With Respect To Gender

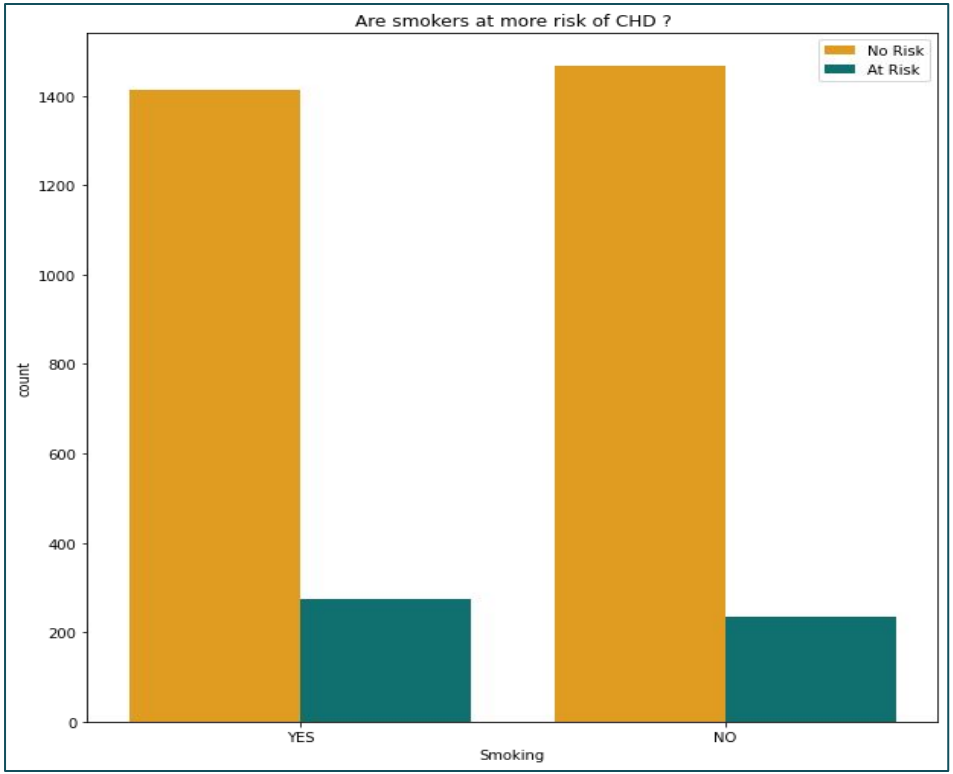


Are smokers at more risk of coronary heart disease ??

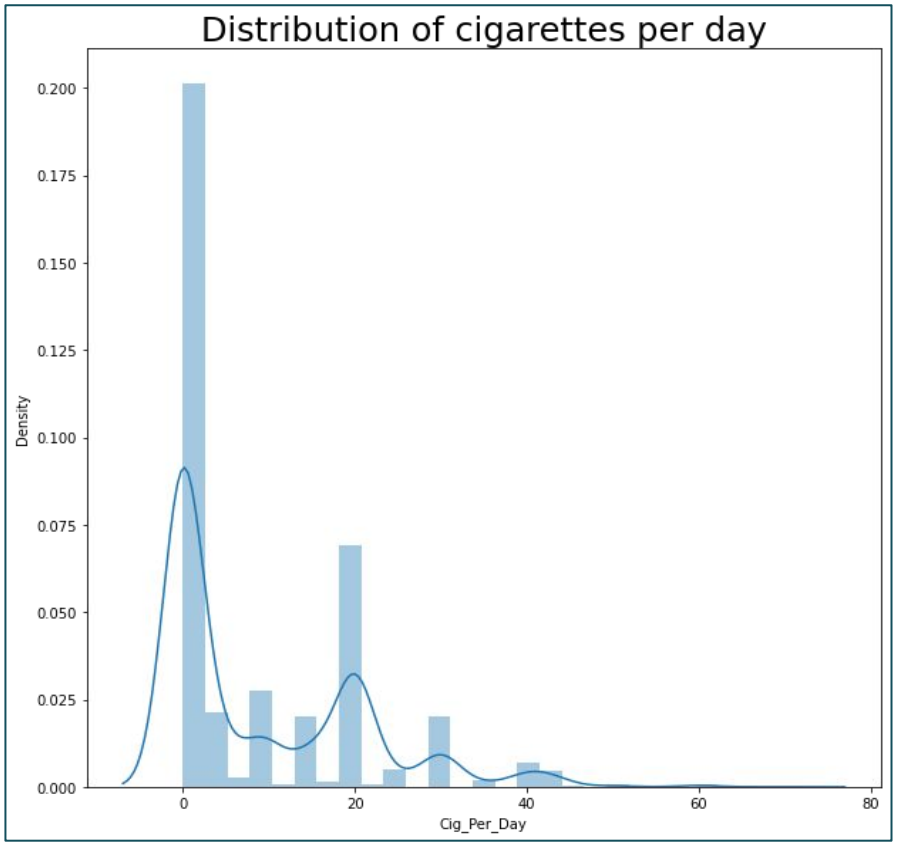
Distribution Of Smokers and Non-Smokers



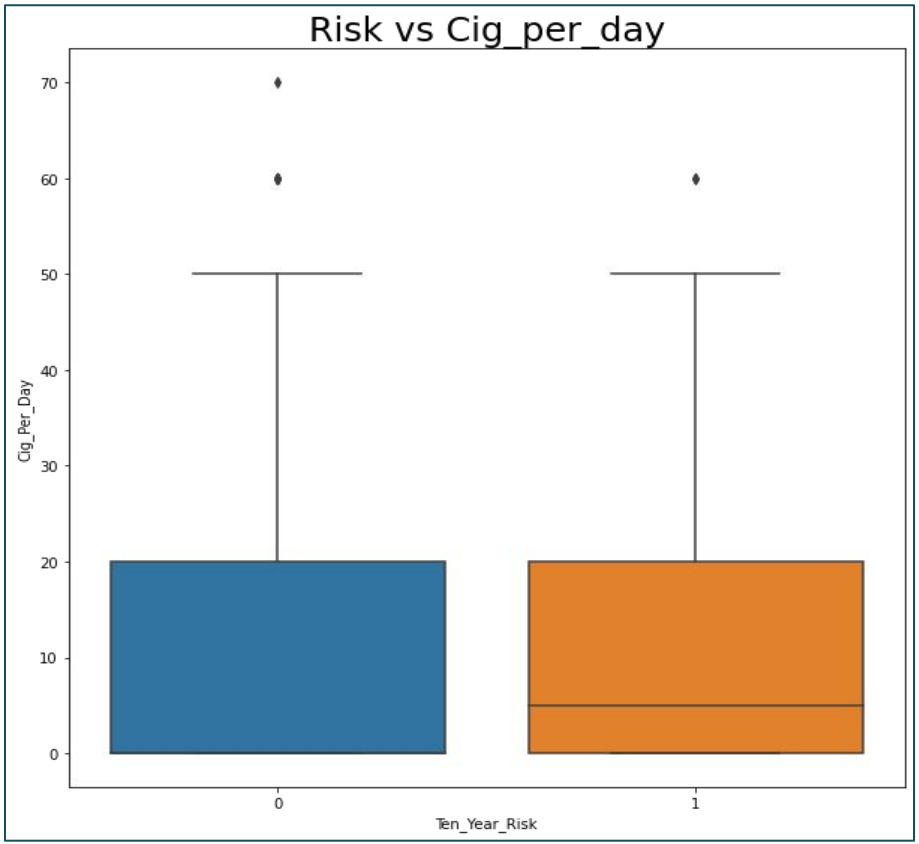
Smokers at Risk



Distribution of cigarettes per day:

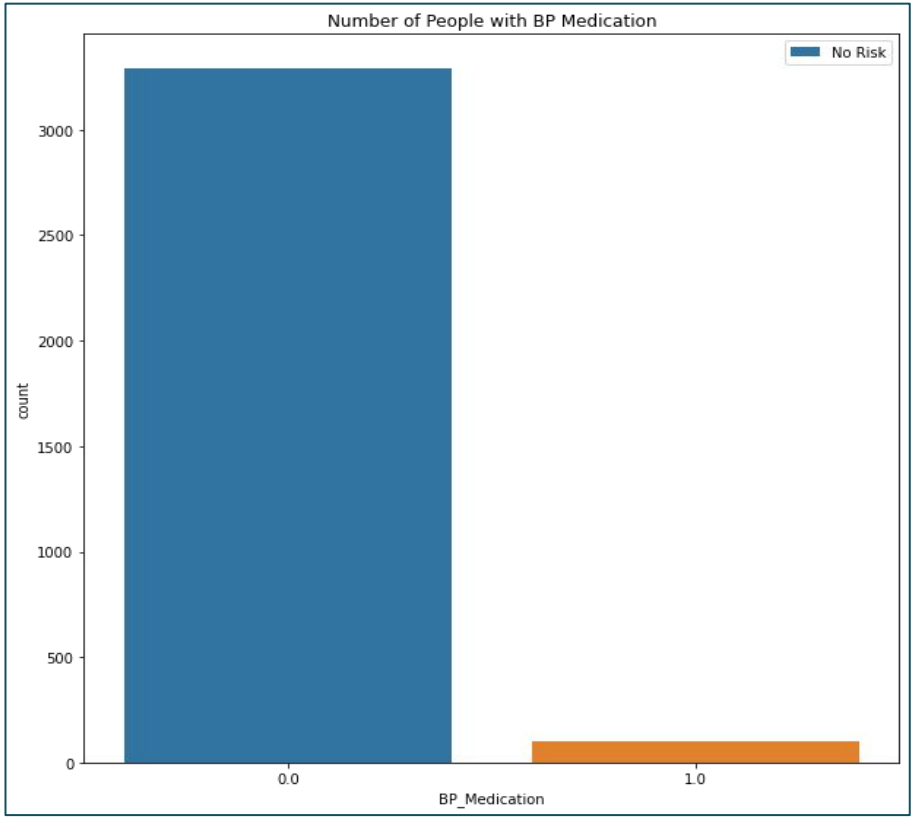


Risk vs Cigarettes Per Day

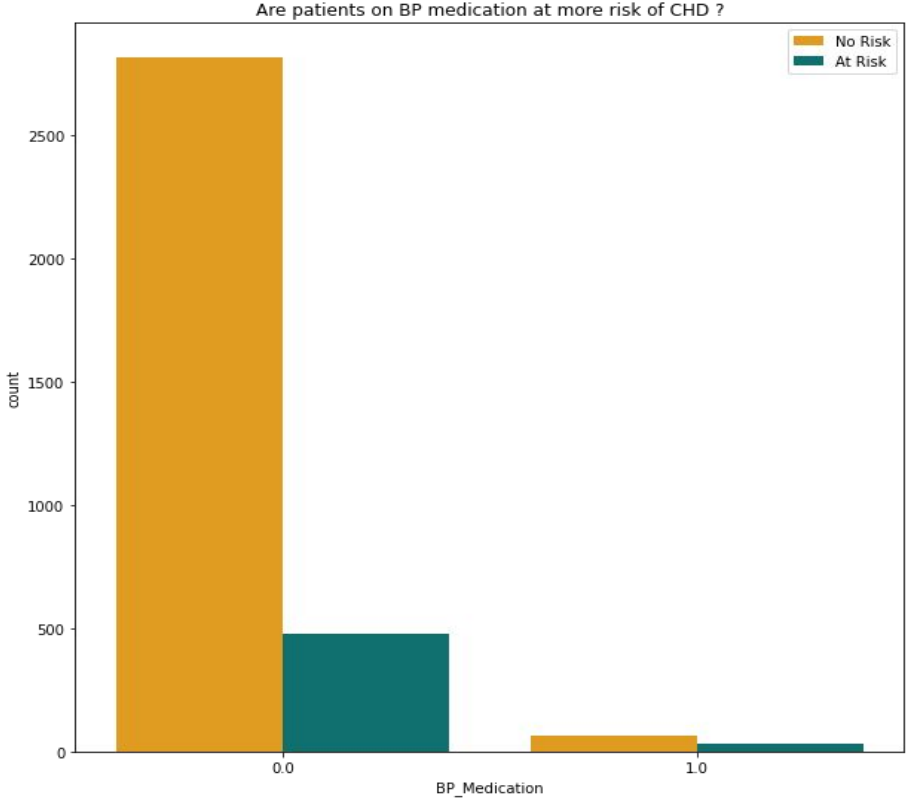


Are patients with blood pressure on medication at more risk of CHD ??

Distribution of people with BP Medication

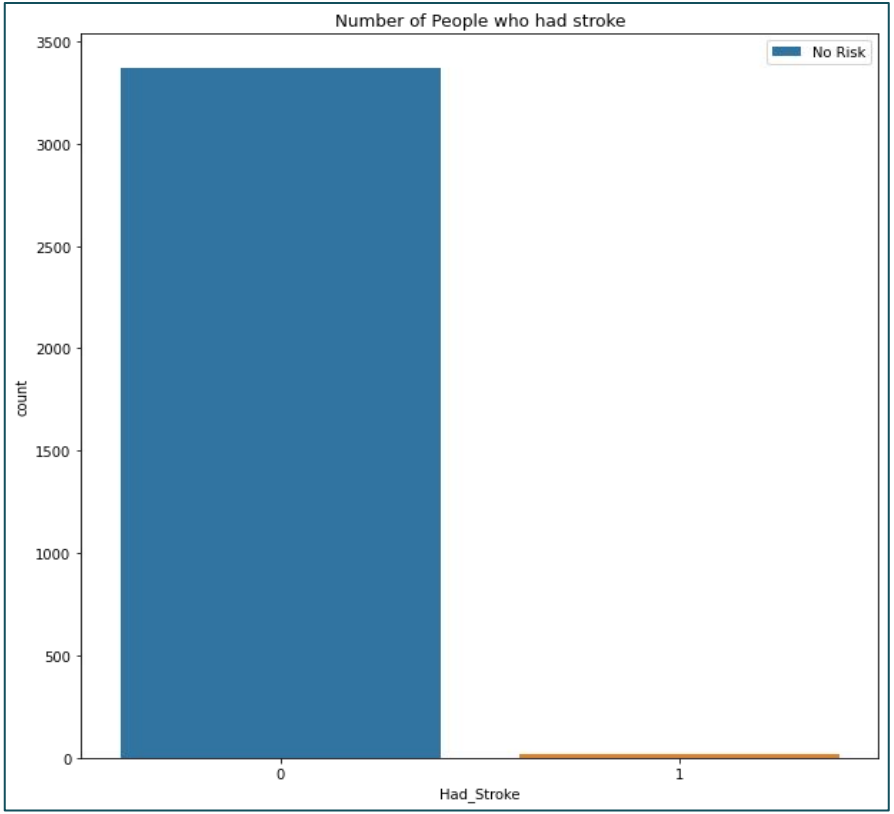


People at Risk with BP Medication

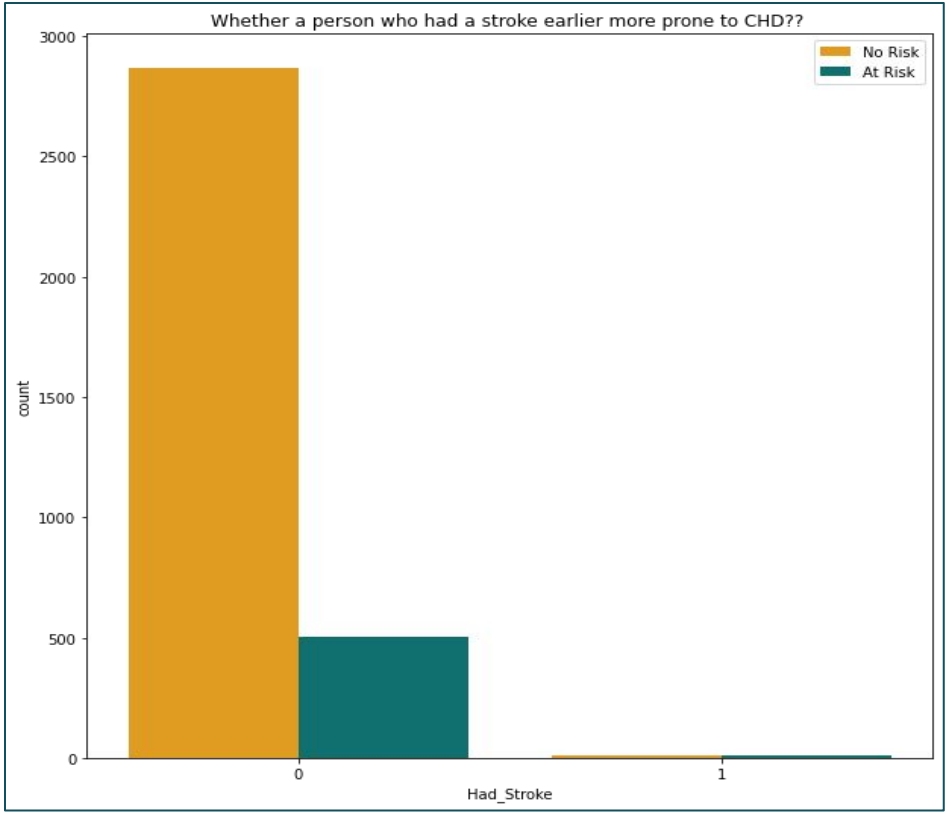


Whether a person who had a stroke earlier more prone to CHD?

Distribution of People who had stroke

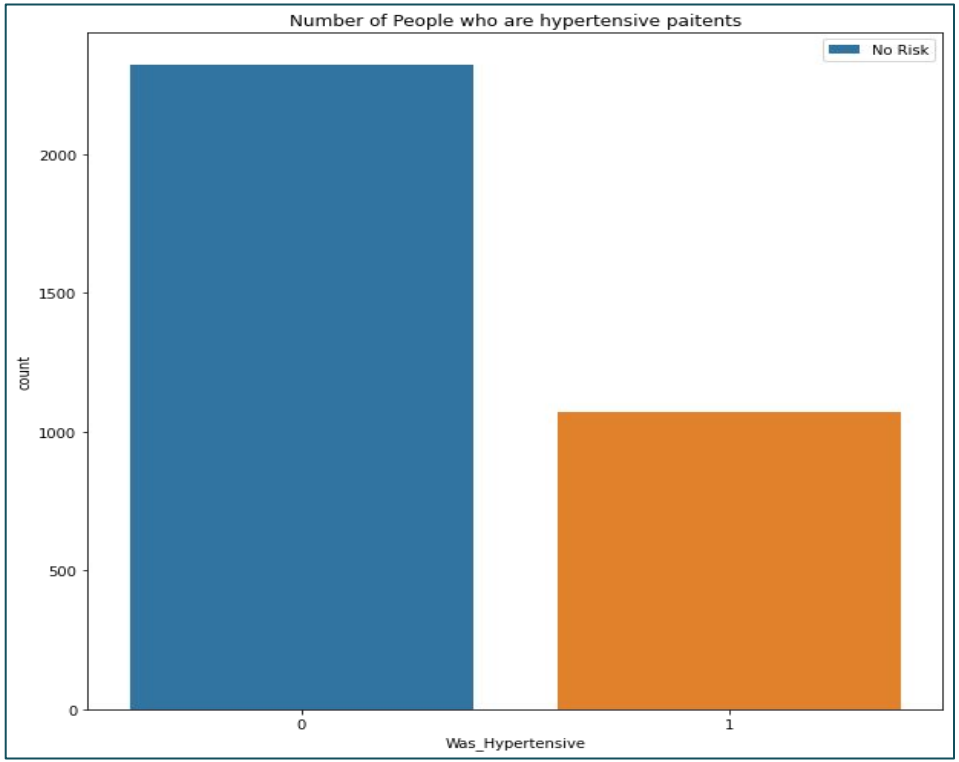


People who had Stroke with Risk

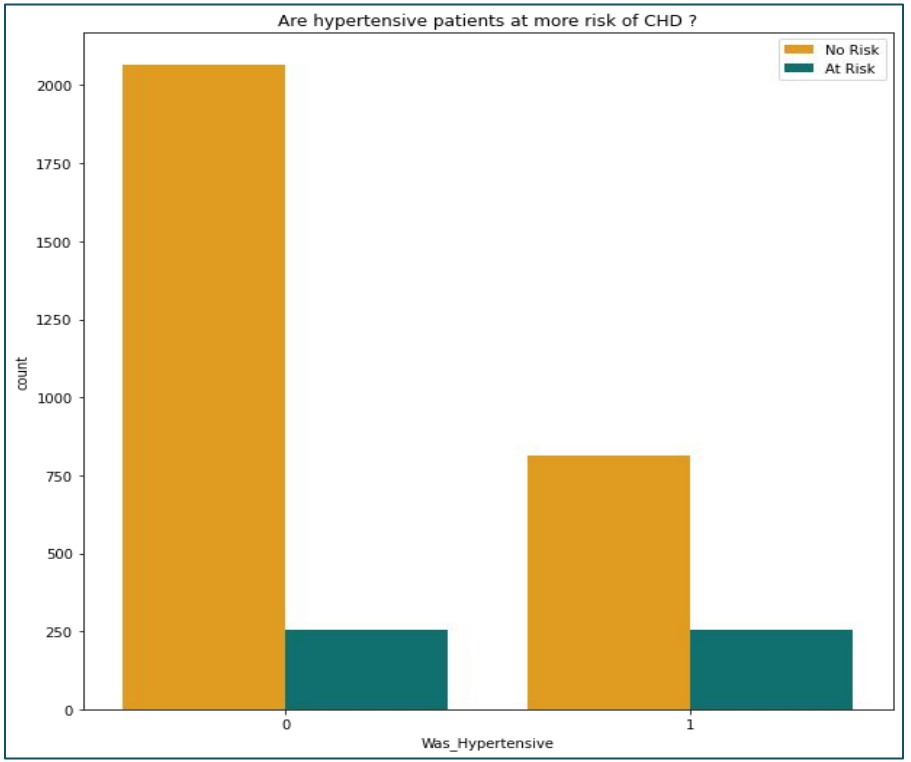


Are hypertensive patients at more risk of CHD ??

Distribution of Hypertensive Patients

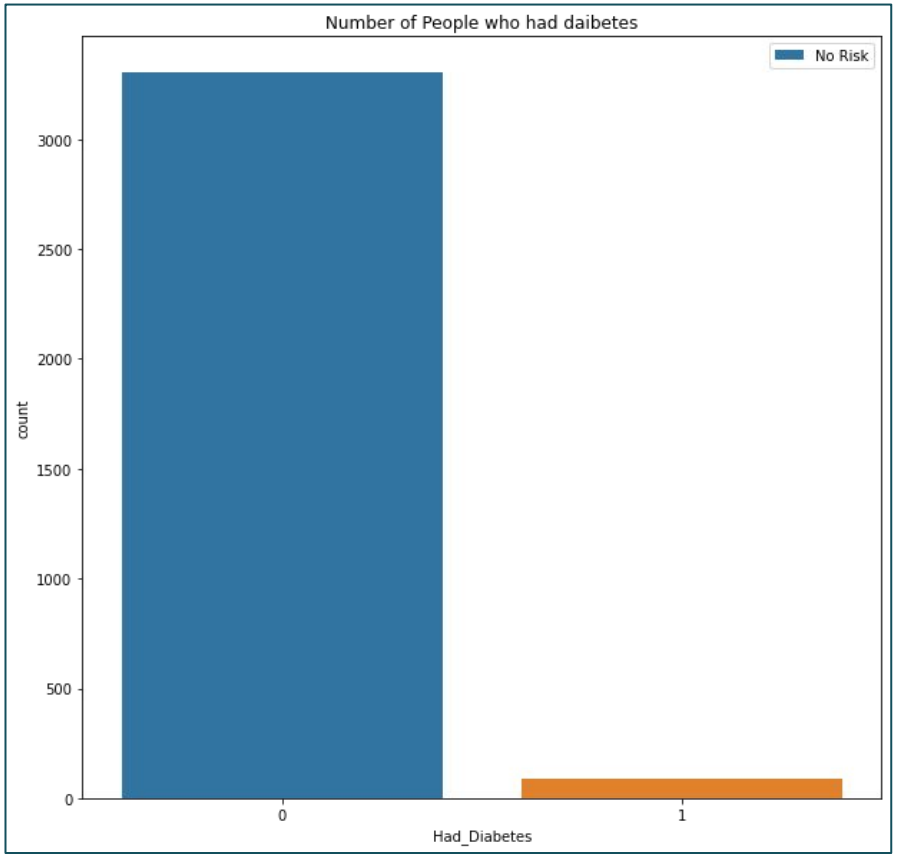


Hypertensive Patients At Risk

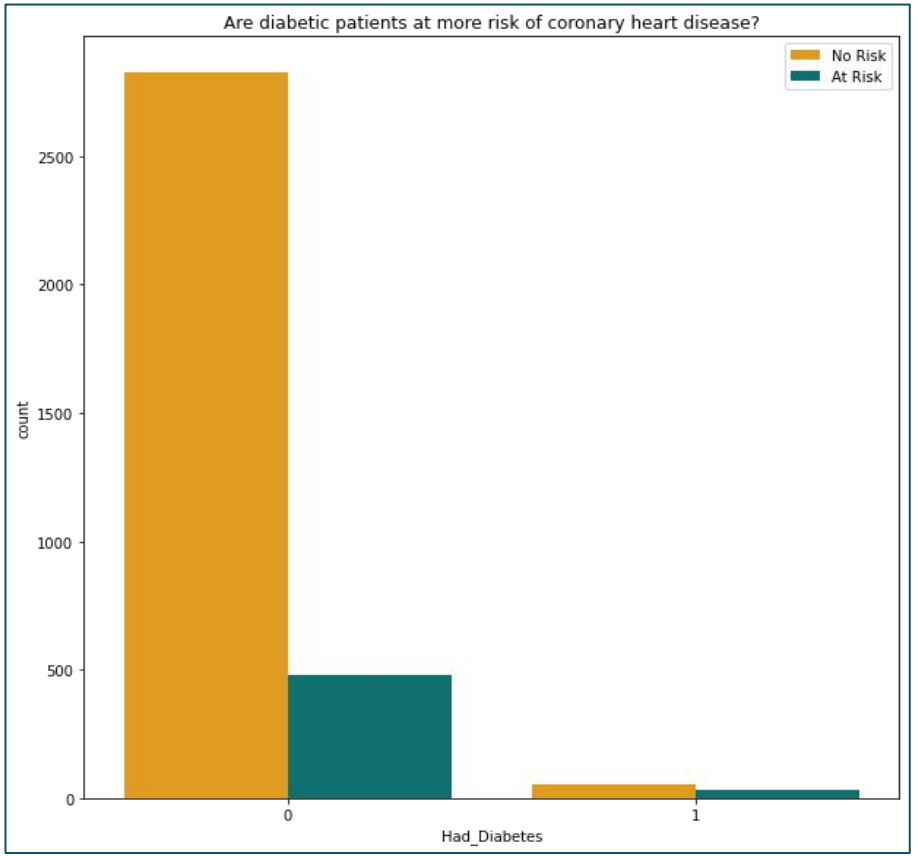


Are diabetic patients at more risk of CHD??

Distribution of Diabetes Patients

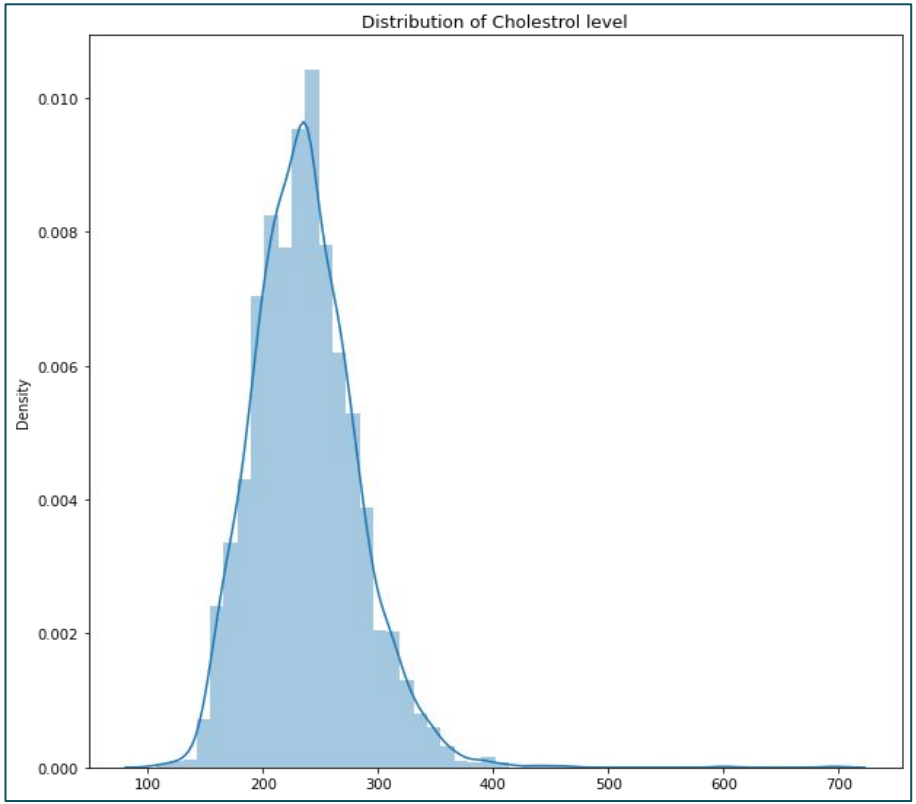


Diabetes Patients At Risk

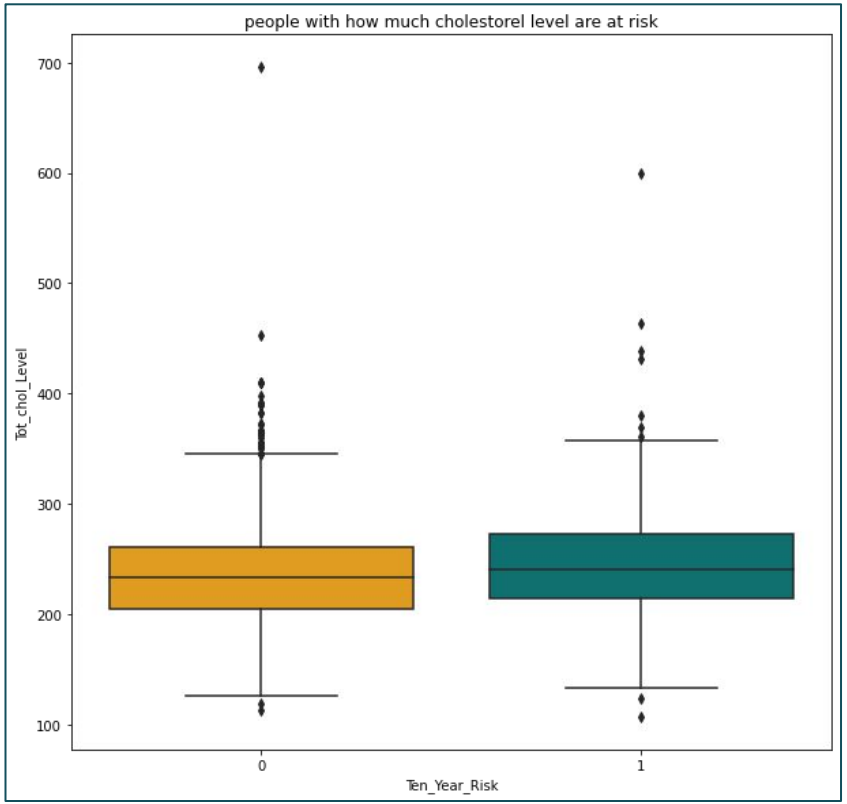


Lets check people with how much cholesterol level are at risk of CHD ??

Distribution Of Cholesterol Level

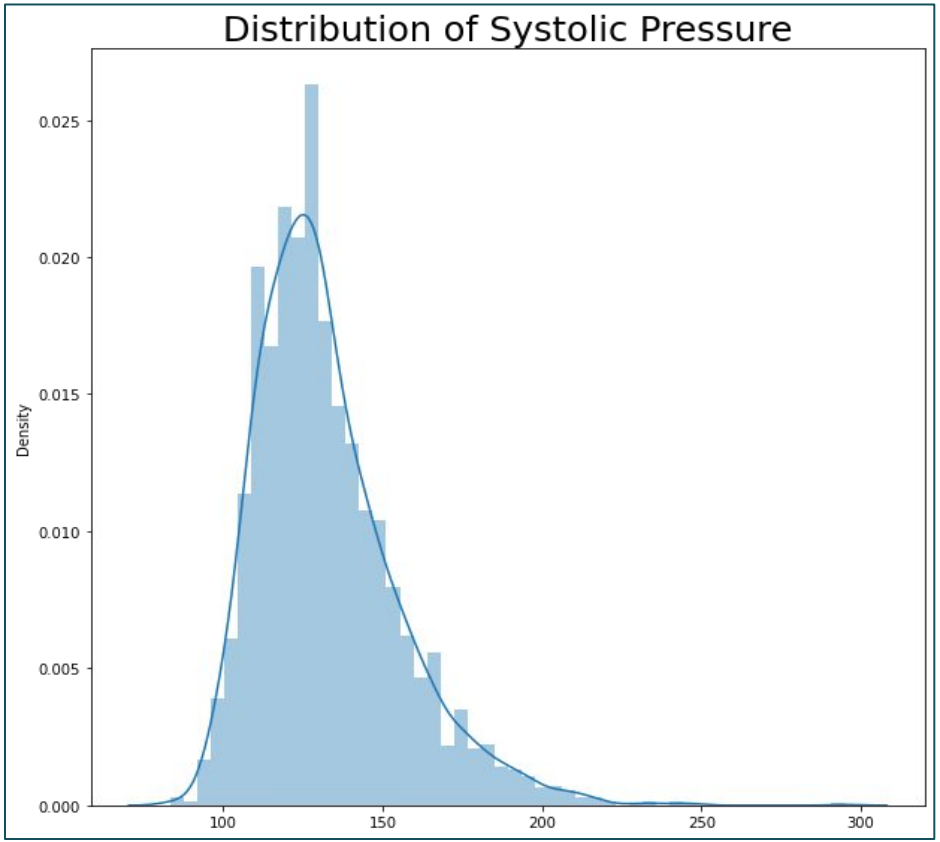


People with Cholesterol Level At Risk

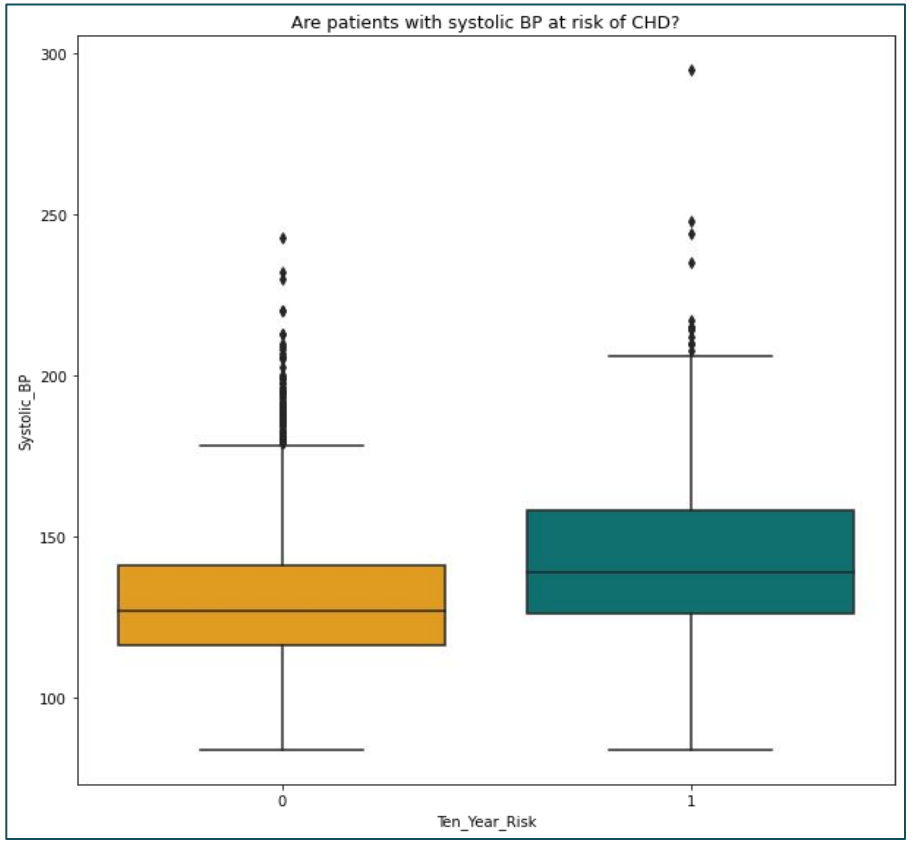


Are patients with systolic BP at risk of CHD?

Distribution Of Systolic BP

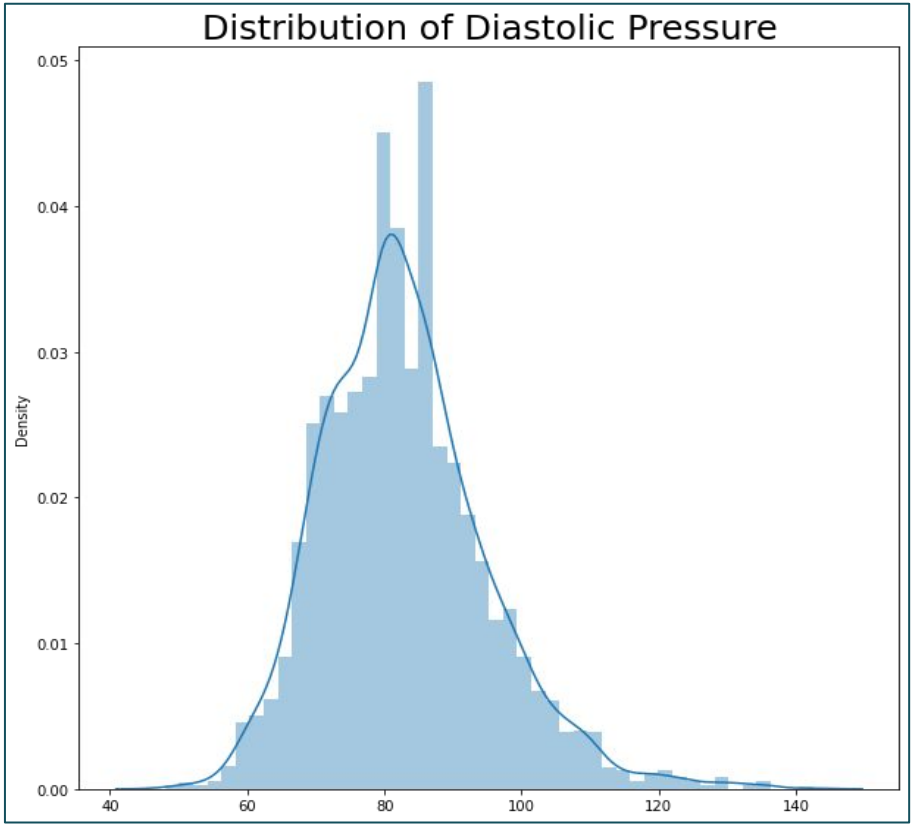


Patients With Systolic BP at Risk

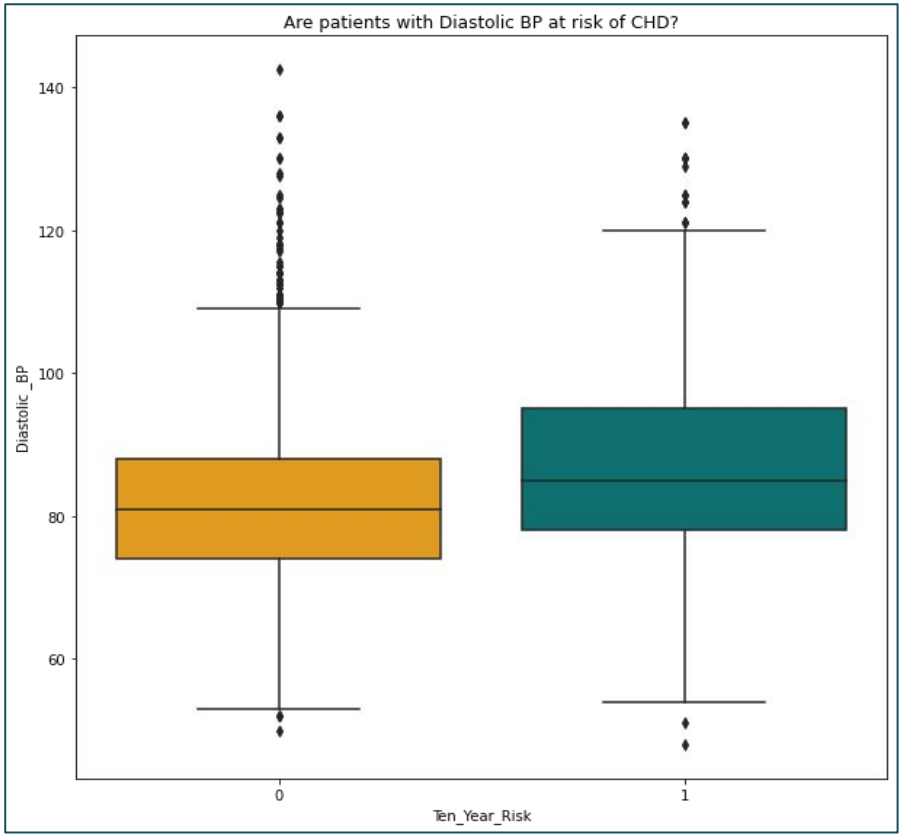


Are patients with Diastolic BP at risk of CHD?

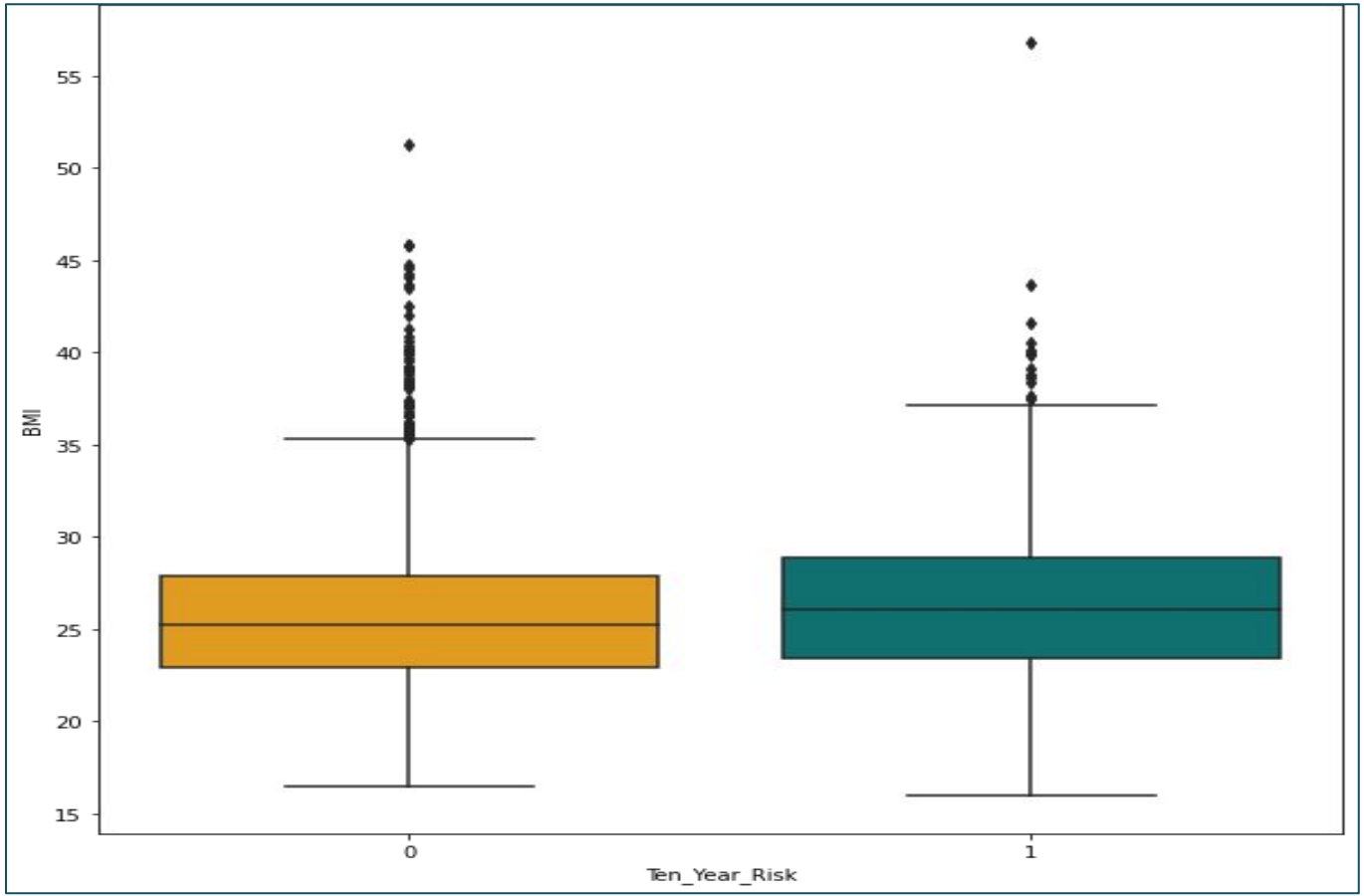
Distribution of Diastolic Pressure



Patients With Systolic BP at Risk

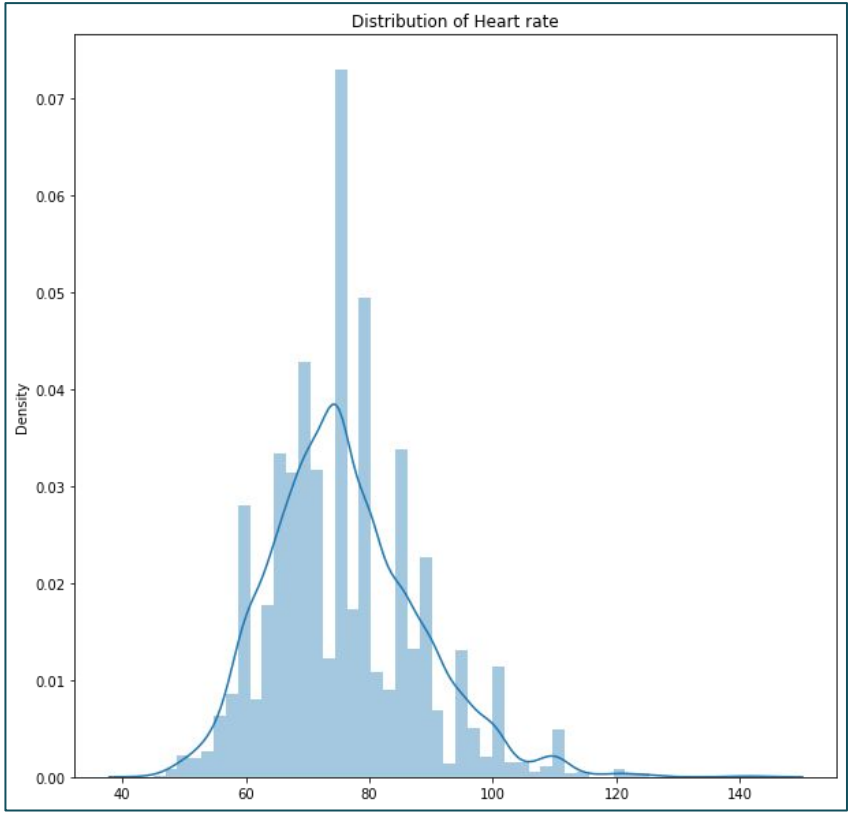


Are patients BMI important to show the risk of CHD?

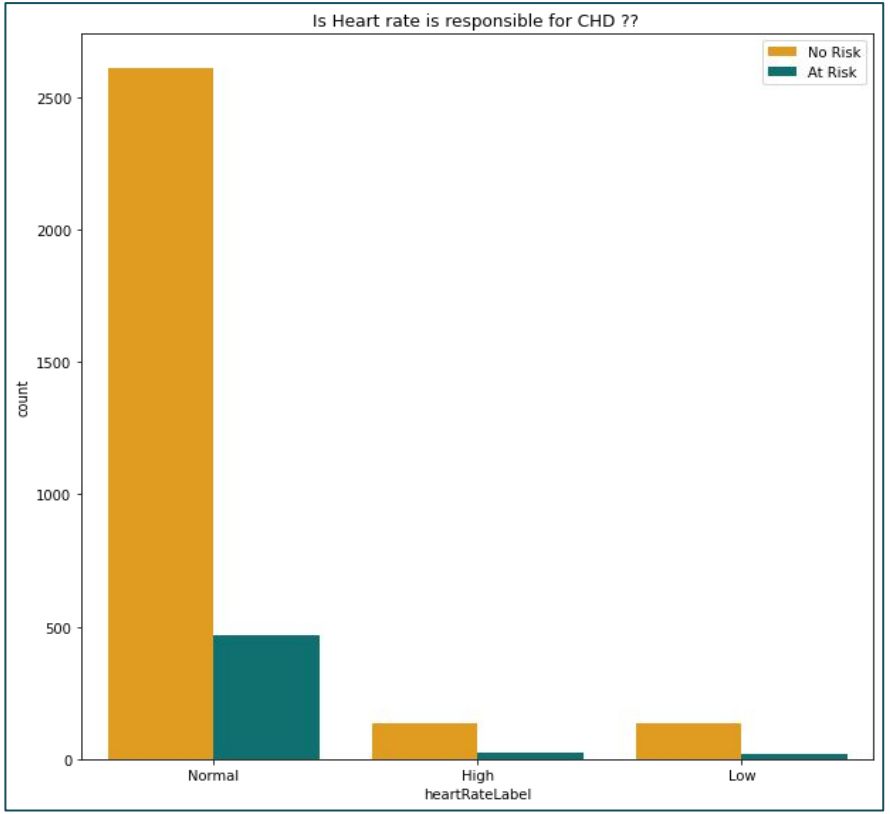


Is Heart rate responsible for CHD ??

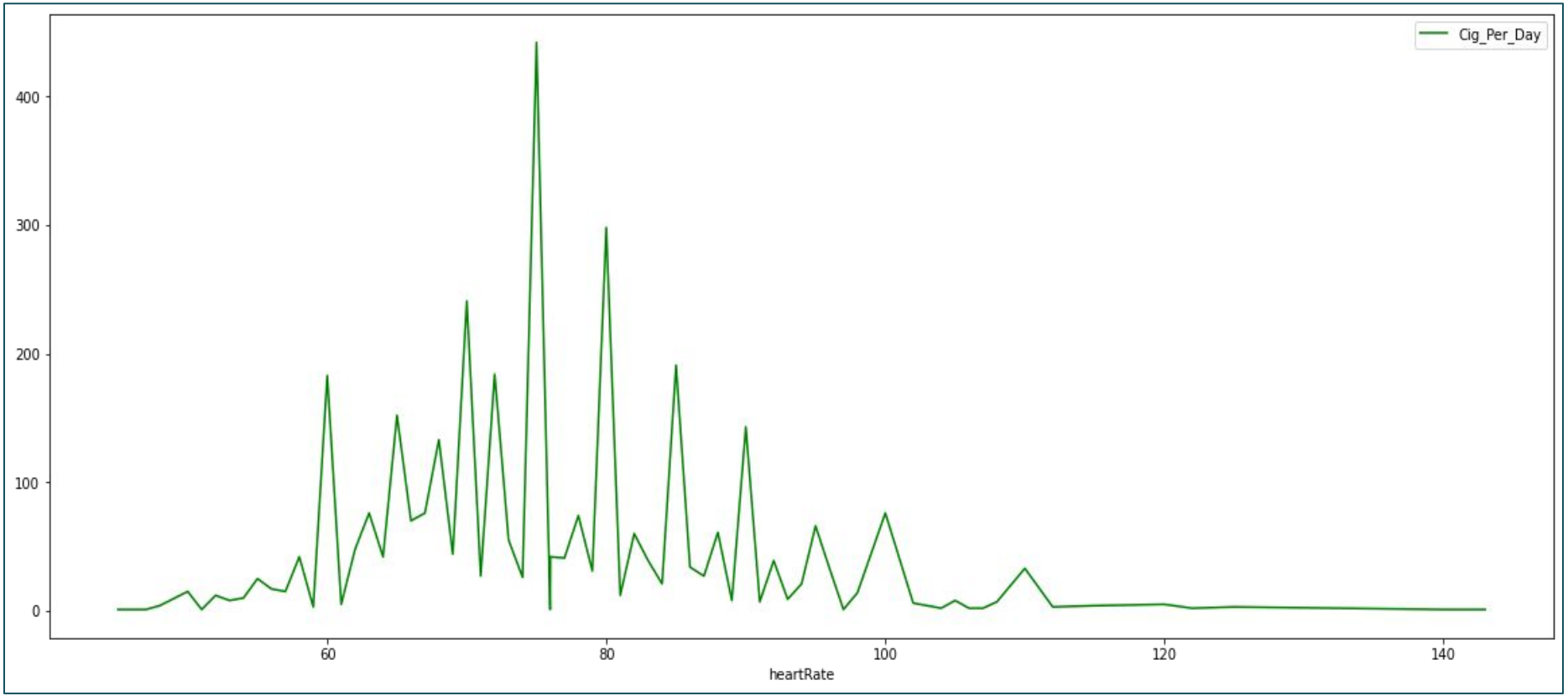
Distribution Of Heart Rate



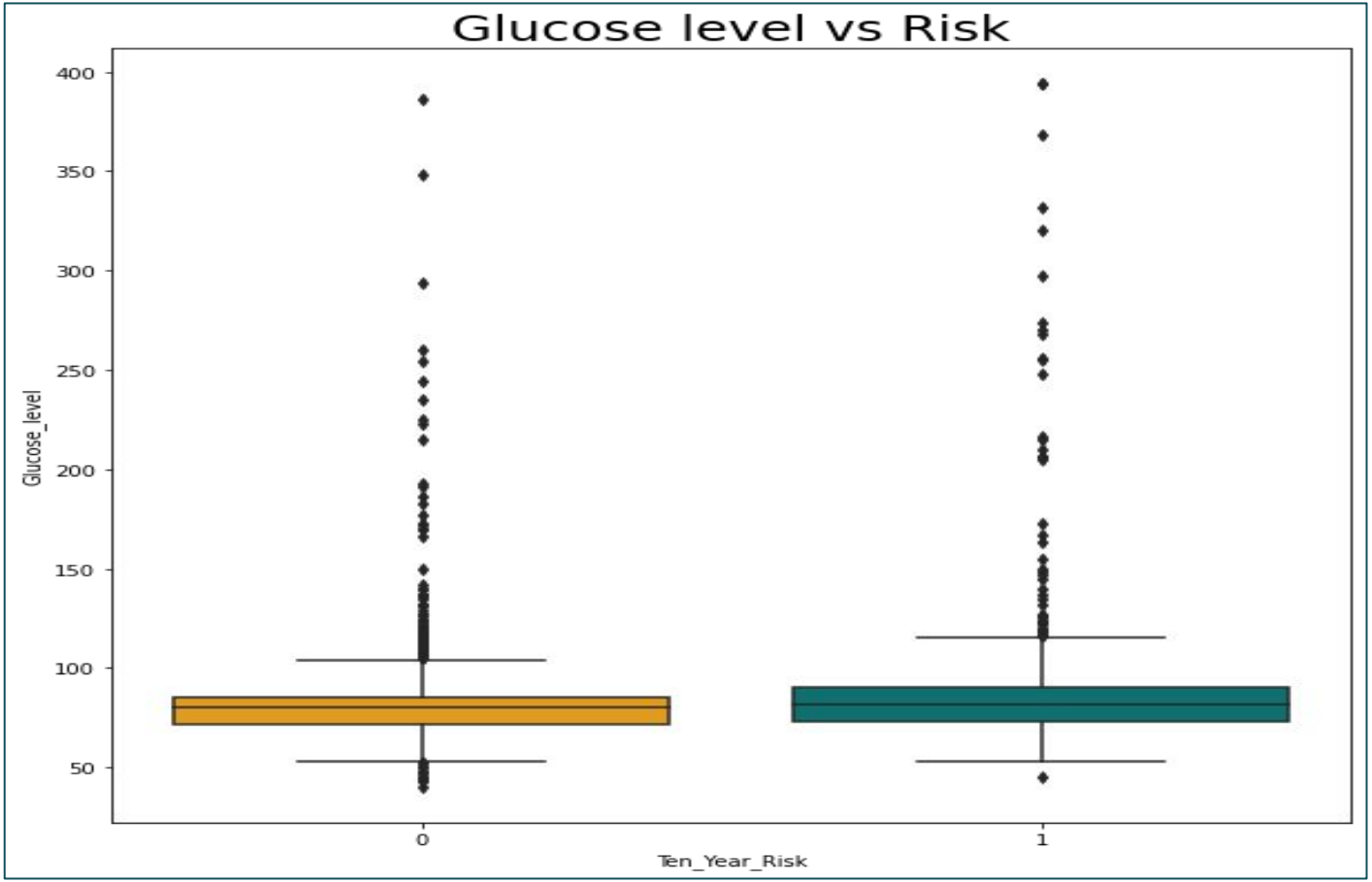
Heart Rate Vs Risk



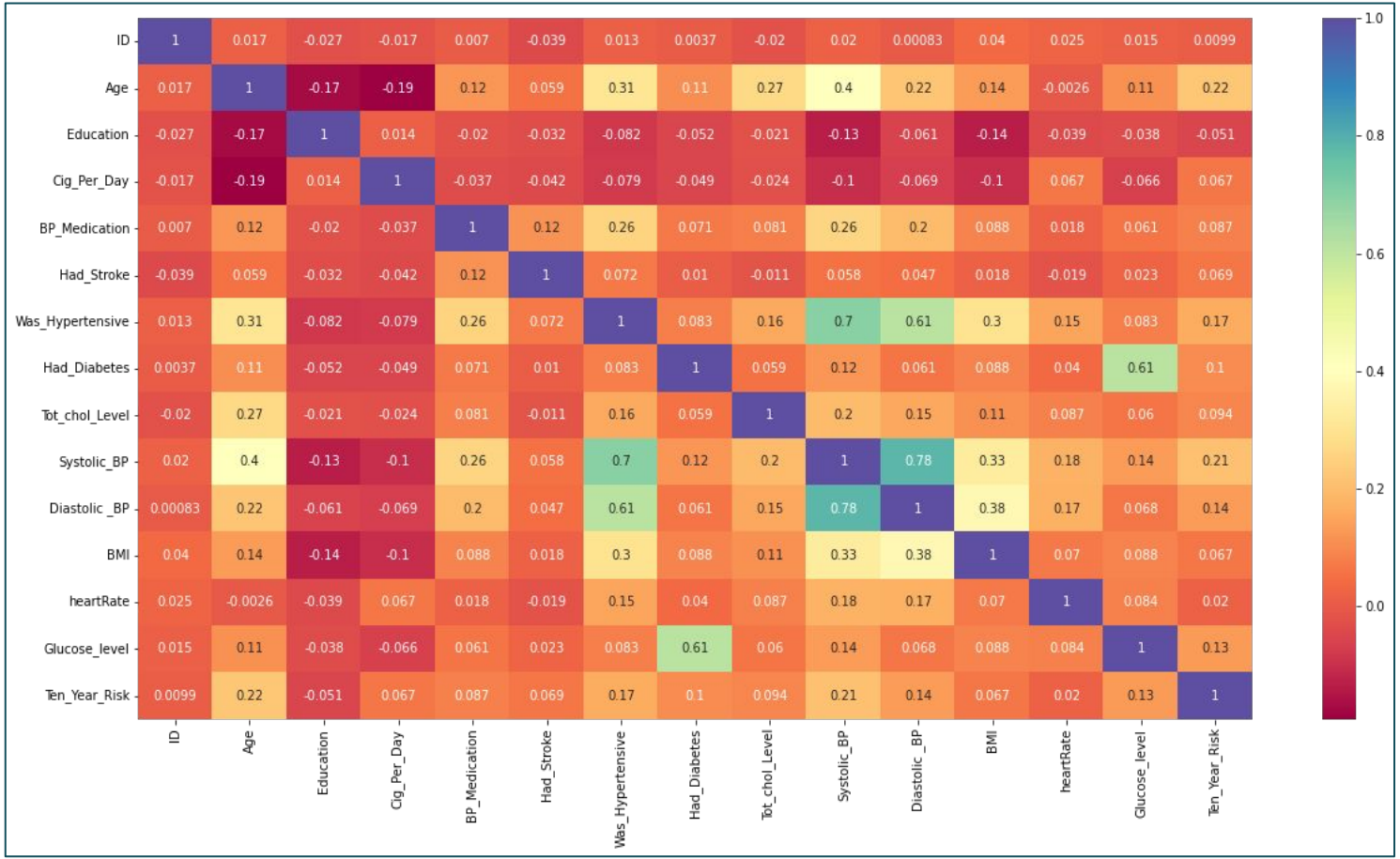
Heart Rate with respect to cigarettes per day:



Can patients Glucose levels show the risk of CHD?



Correlation Analysis:

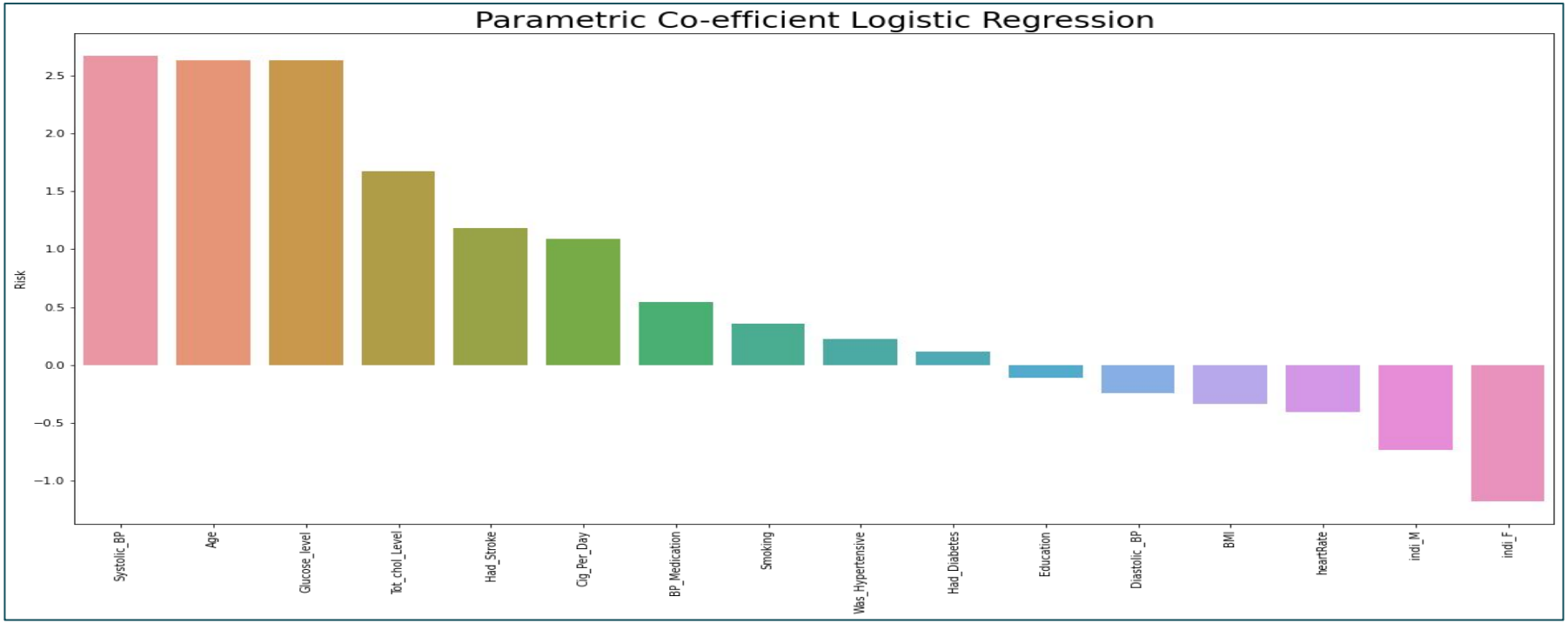


Models Used:

- Logistic Regression
- Support Vector Machine (SVM)
- Random Forest Classifier
- XGBoost Classifier
- Neural Networks

Logistic Regression:

Feature Importance For Logistic Regression

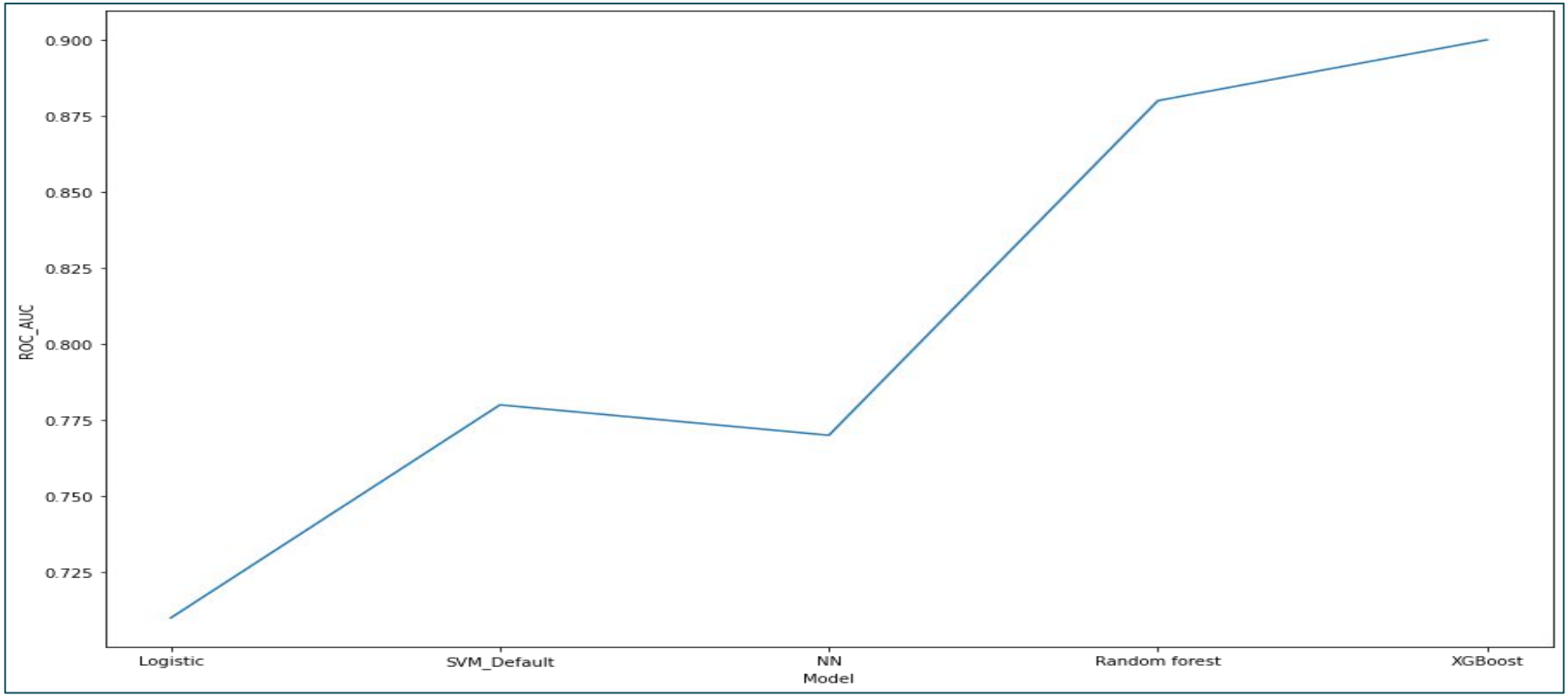


Model Evaluation Result:

Algorithm	Train Accuracy Score	Test Accuracy Score
Logistic Regression	0.68	0.66
Support Vector Machine	0.79	0.78
Neural Network	0.81	0.77
Random Forest Classifier	0.97	0.89
XGBoost Classifier	0.99	0.90

XGBoost Classifier have performed really well and got the best scores with XGBoost Classifier as compared to other Models, so I conclude XGBoost is my optimal model for use and we can use this model for further in predicting Cardiovascular risk.

ROC AUC Comparison:



Challenges:

- Execution takes time.
- Less amount of data available made it difficult to predict properly.
- Missing relevant/Important features in our dataset like Chest pain location, chest pain type, Family history of coronary artery, Exercise, etc.

Conclusion:

- There are **15.1 %** people in our dataset are are **at risk** for cardiovascular disease and **84.9 %** people **are safe** (ten year risk).
- There is more **risk** of cardiovascular disease in patients of **age between 51 to 63**.
- The **count of male and female are same in risk** which is around 200, though females are more than males in our dataset.
- Around **250 smokers** are in **risk** and around **210 non-smokers** are is **risk** for cardiovascular disease.
- We **can't evidentially state smoking will lead to heart disease**, as we seen from count plot there is no huge difference between these to commune and also our extreme smoker who smokes 70 cigarettes per day is not having ten year risk.
- There are very few people who are done with BP medication which are around **200** but many people have not taken any BP medication and they are around **3200**. We **cannot say that after taking medication person are safe**.
- Around **500 patients** who did not had stroke yet and are **at risk** and around **2800 patients are safe**.
- Around **250 people** with hypertensive are in **risk** and around **255 people** with no hypertensive are at **risk**.
- Here we can see people who did not had diabetes are more and around **500 people who did not had diabetes are at risk**.And there are very **few people who had diabetes are at risk**.

- Most of the people who are **not in risk** their Cholestrol level lies **between 210 to 280** and people who are **in risk** their cholestrol level lies **between 215 to 285** there in not huge difference it is quite normal.
- Most people who are not **in risk** their systolic BP lies **between 110 to 140** and people who are **at risk** their systolic BP lies **between 125 to 160**. We can say **people with high systolic BP are at risk**.
- Most people who are **not in risk** their diastolic BP lies **between 75 to 85** and people who are **at risk** their diastolic BP lies **between 89 to 90**. We can say there is a slight increase in diastolic BP of people who are in risk.
- Most people who are not in risk their BMI lies between **22 to 28** and people who are at risk their BMI lies between **23 to 29 approx**. WE cannot see any difference BMI is approx. same of risky and not risky people.
- Most people who are not in risk their heart rate lies between **68 to 83** and people who are at risk their heart rate lies between **68 to 84**. which is same for risky and not risky people.
- There is not that difference between the glucose level of risky and non risky patients. glucose level lies **between 70 to 80** for both risky and non risky patients.
- Most people smoke cigarettes between **1 to 10** approx. and there heart rate lies between **60 to 100**.

- With logistic regression we got the accuracy score of **0.68** on train data and **0.66** on test data.
- With Support vector machine we got the train accuracy score of **0.79** and test accuracy score of **0.78**.
- With Neural Networks we got the Train Accuracy score of **0.81** and test accuracy score of **0.77**.
- With Random forest classifier we got the train accuracy of **0.97** and test accuracy of **0.89**.
- With XGBoost classifier we got the train accuracy score of **0.99** and test accuracy of **0.90**.

XGBoost Classifier have performed really well and I got the best scores with XGBoost Classifier as compared to other Models, so I conclude XGBoost is my optimal model for use and we can use this model for further in predicting Cardiovascular risk.

Thank You!!