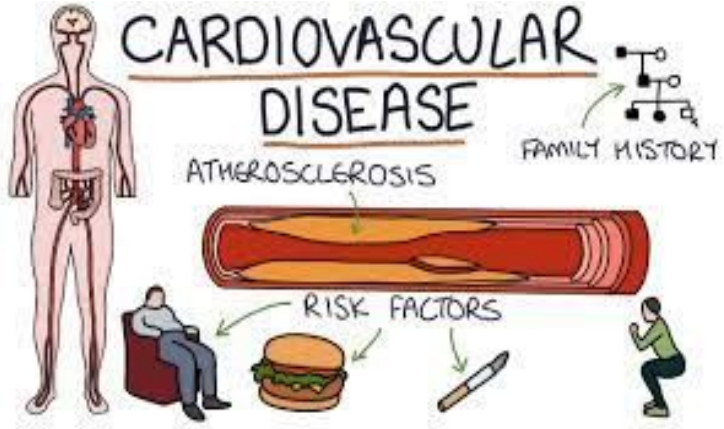
Capstone Project

Cardiovascular Risk Prediction

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.



Key Steps:

- Defining the problem statement
- Data Cleaning
- EDA and data visualization
- Data preprocessing
- Feature selection
- Preparing Dataset for model
- Applying model
- Model validation and selection

Why is predictive analytics useful for Cardiovascular risk ?

- To know which patients are in risk:
- To know which disease lead to Cardiovascular risk:
- To know what habit lead to Cardiovascular risk:
- To know what should be BMI, BP, Diabetes and Cholesterol level:

Handling Null Values:

1. KNN imputer
2. Simple imputer

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

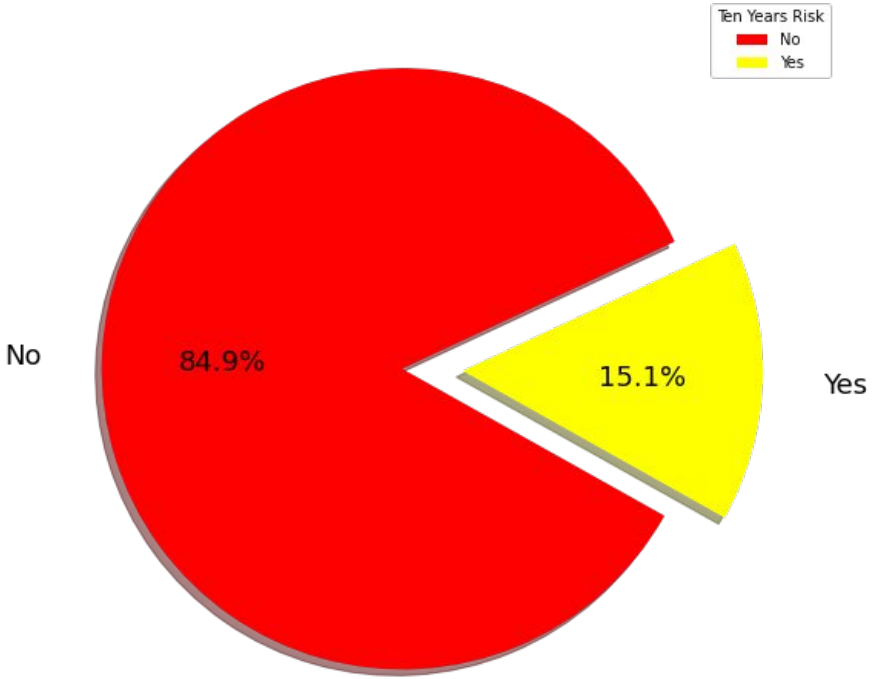
Variable Names:

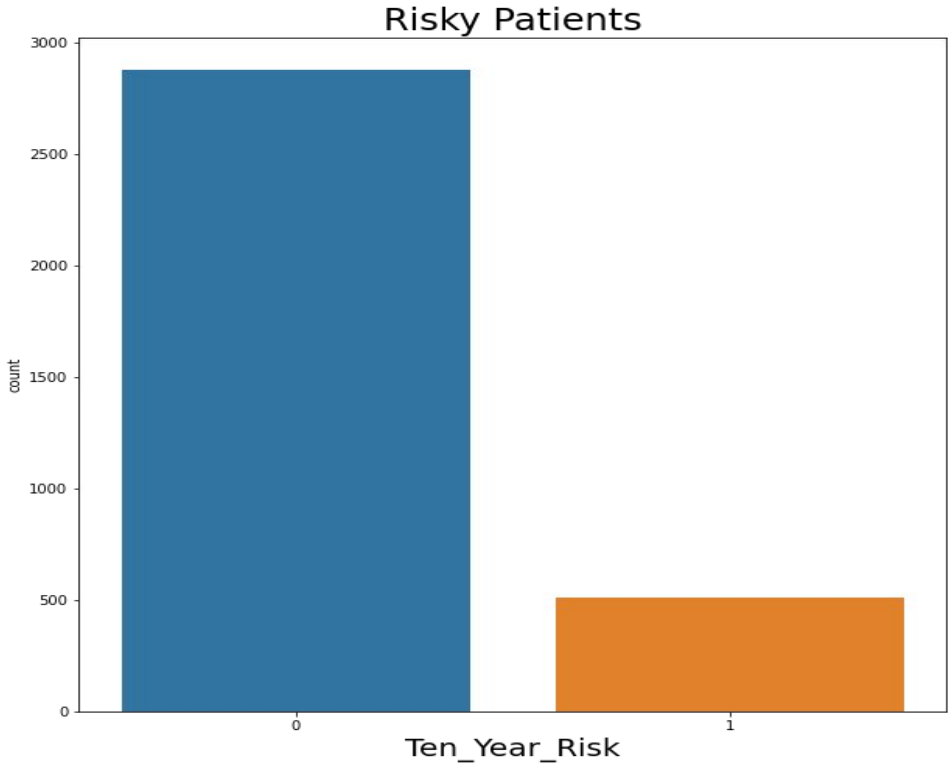
- Sex:
- Age:
- is_smoking:
- Cigs Per Day:
- BP Meds:
- Prevalent Stroke:
- Prevalent Hyp:
- Diabetes:
- Tot Chol:
- Sys BP:
- Dia BP:
- BMI:
- Heart Rate:
- Glucose:
- 10-year risk:

Exploratory Data Analysis:

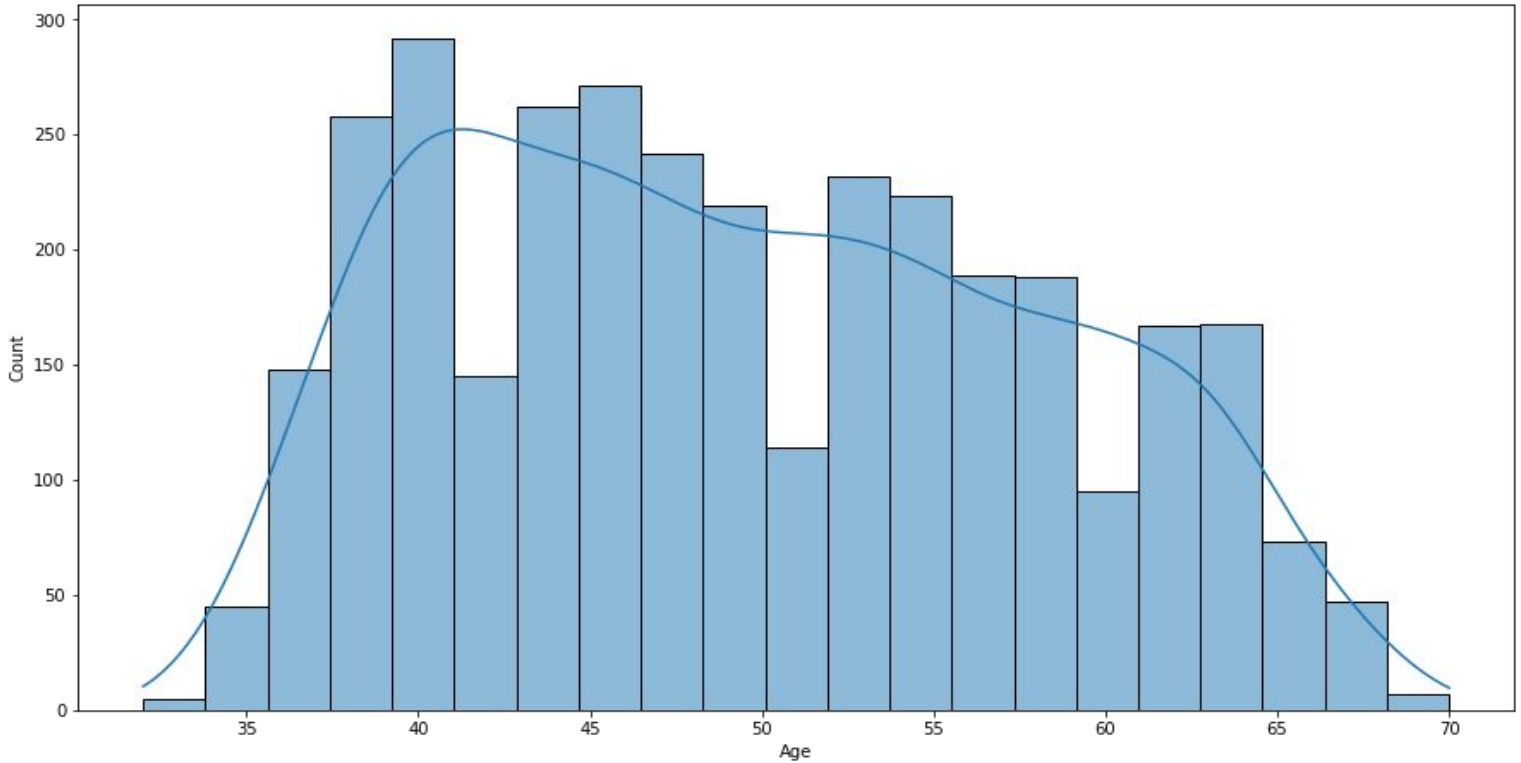
EDA is used for analyzing what the data can tell us before the modeling or by applying any set of instructions/code.

Risky Patients:

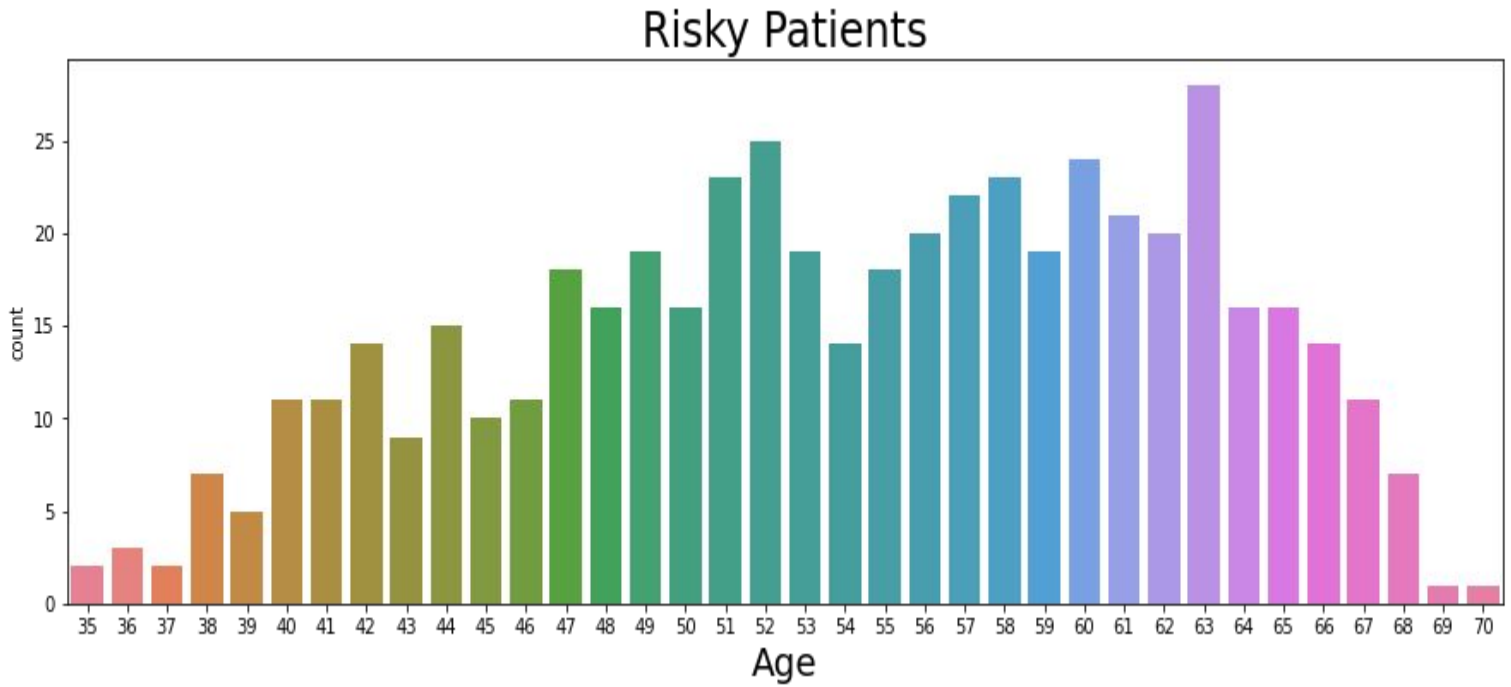




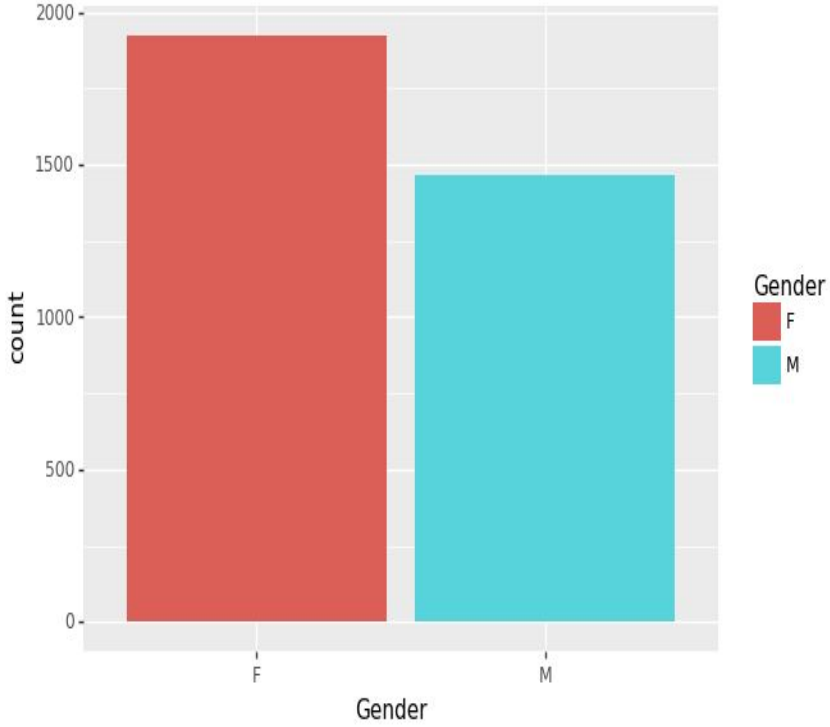
Distribution of age:

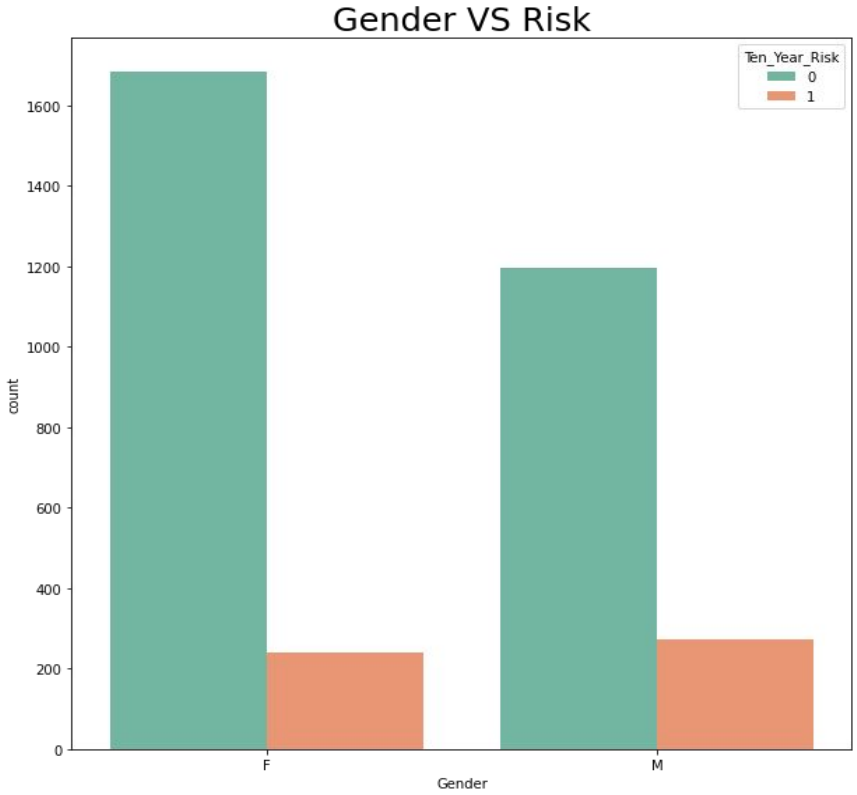


Risky Patients With Respect To Age:

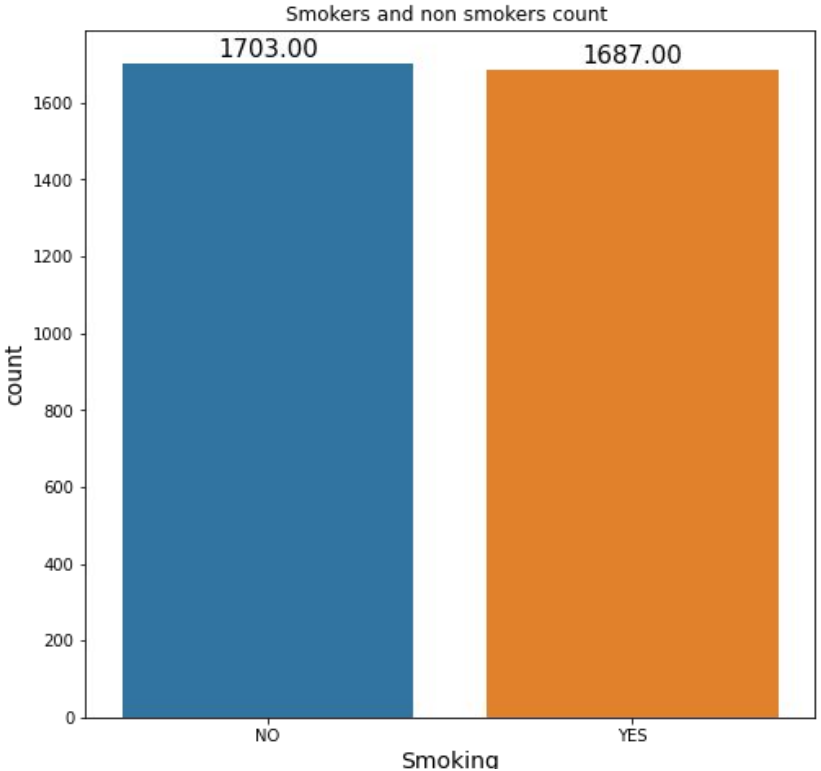


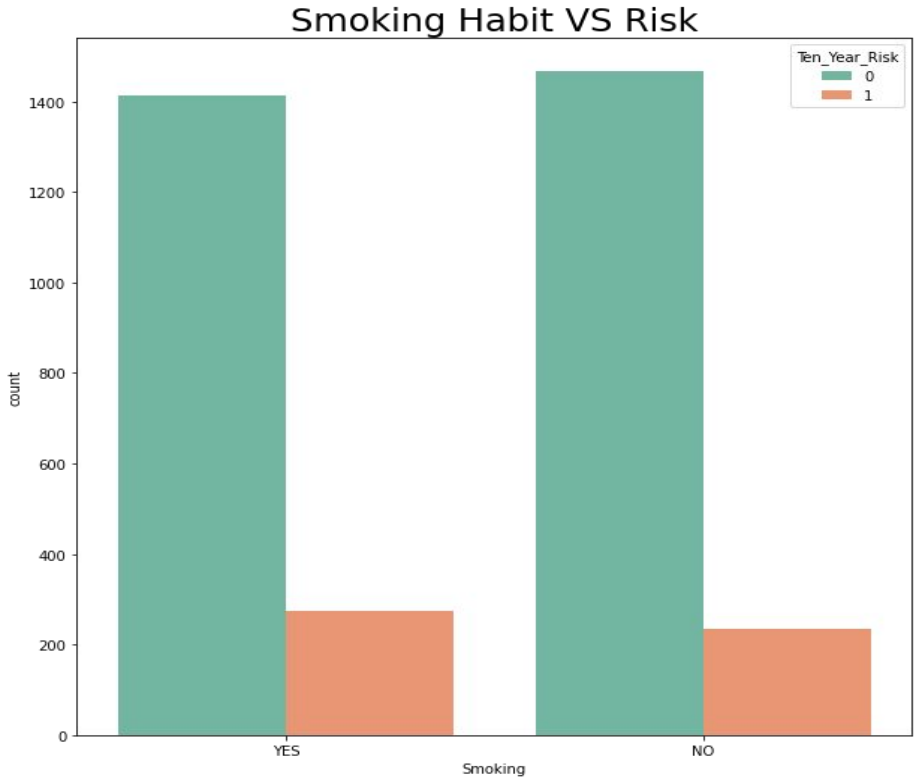
Risk with respect to gender:



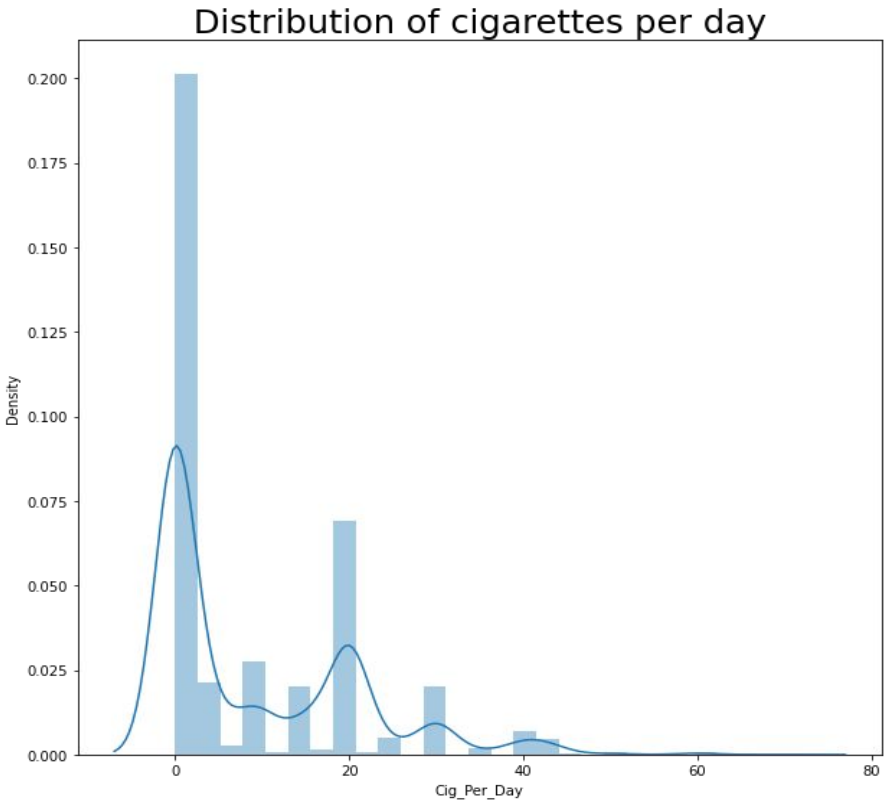


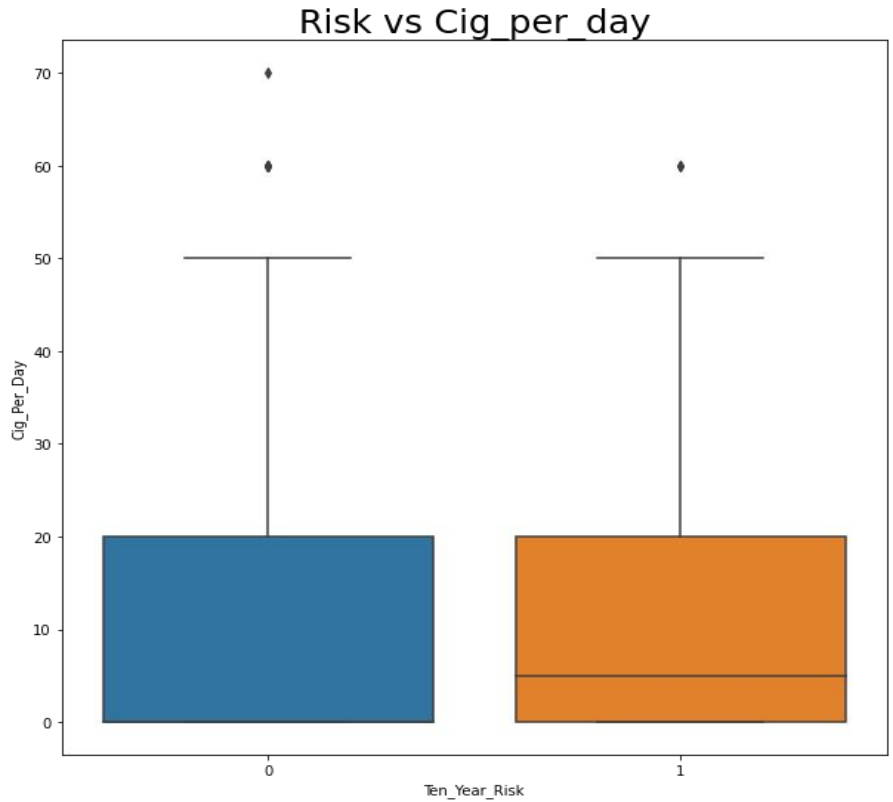
Risk With Respect To Smoking Habit:



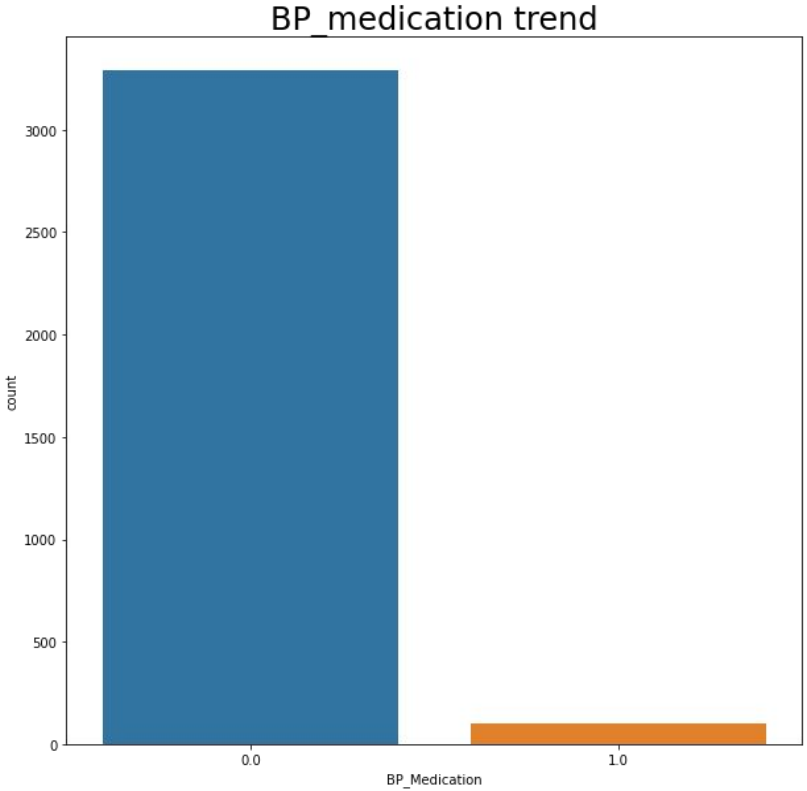


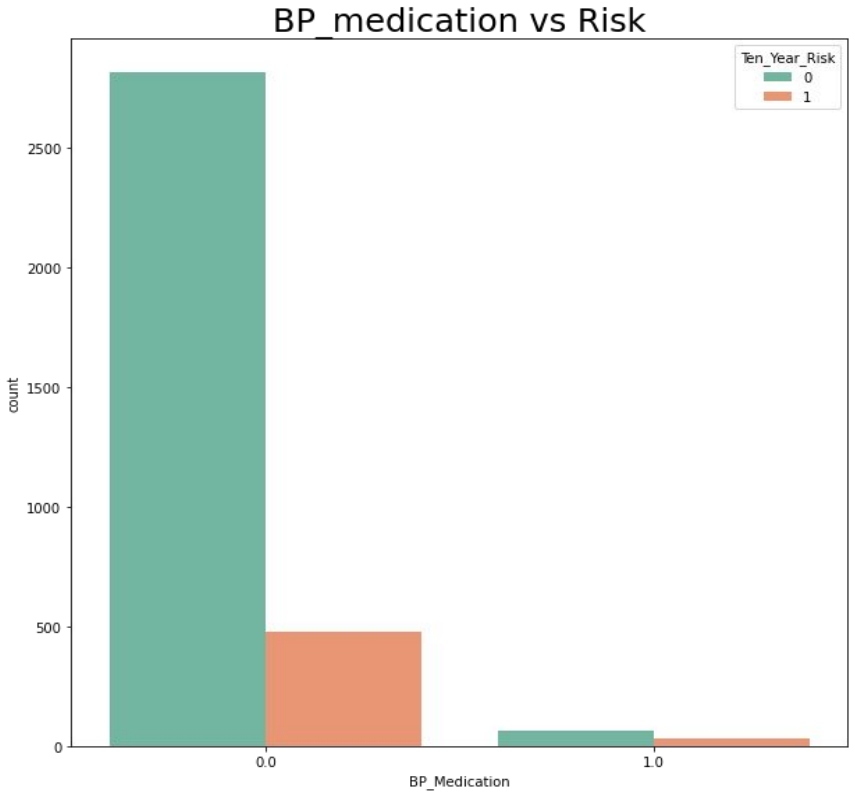
Risk With Respect To Cigarettes Per Day:



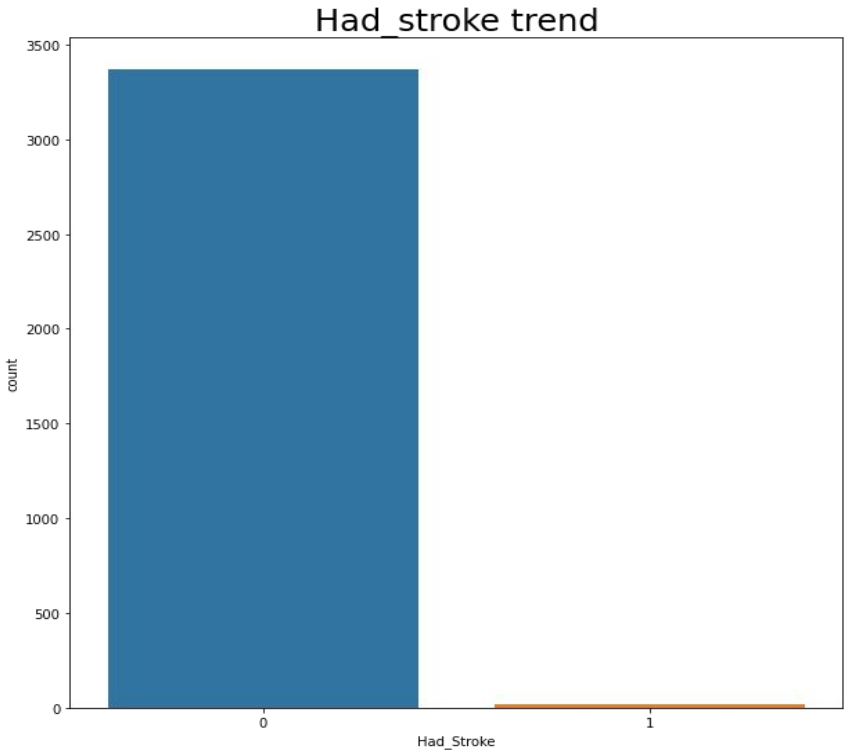


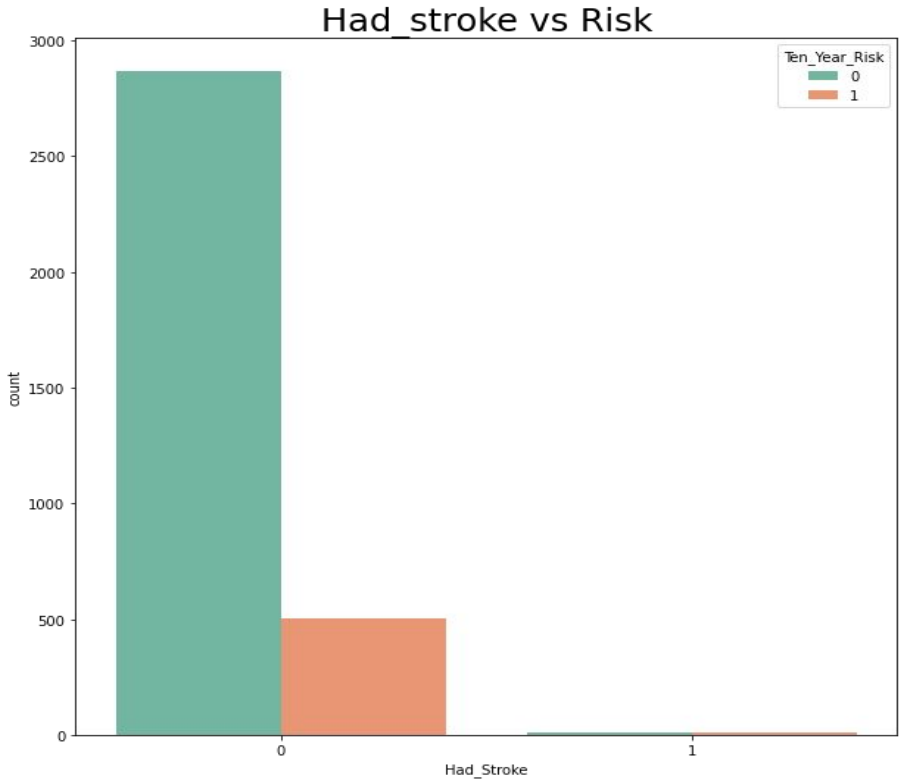
Risk With Respect To BP Medication:



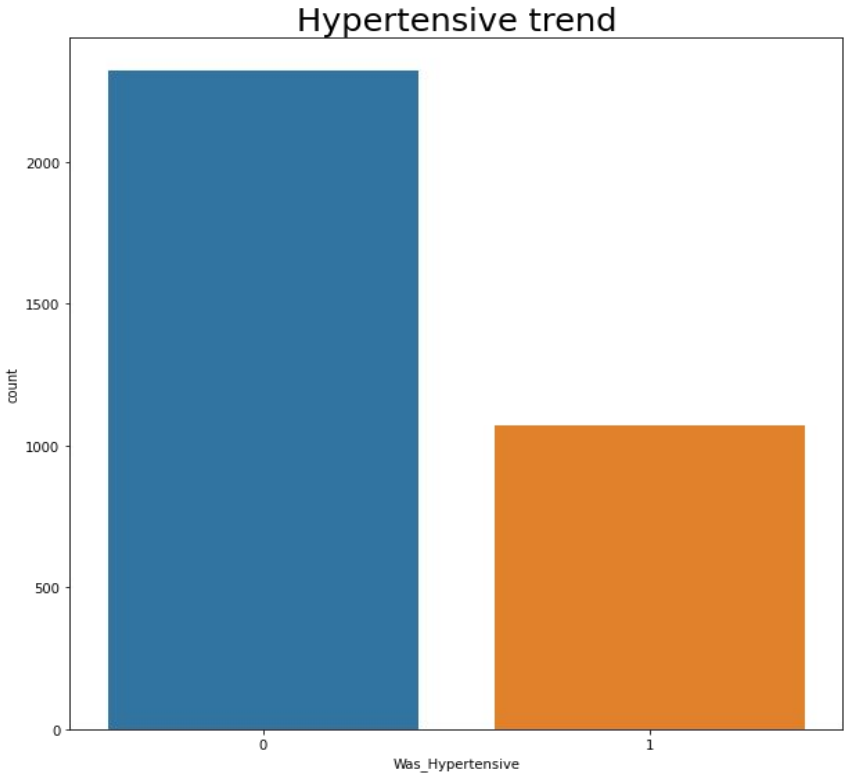


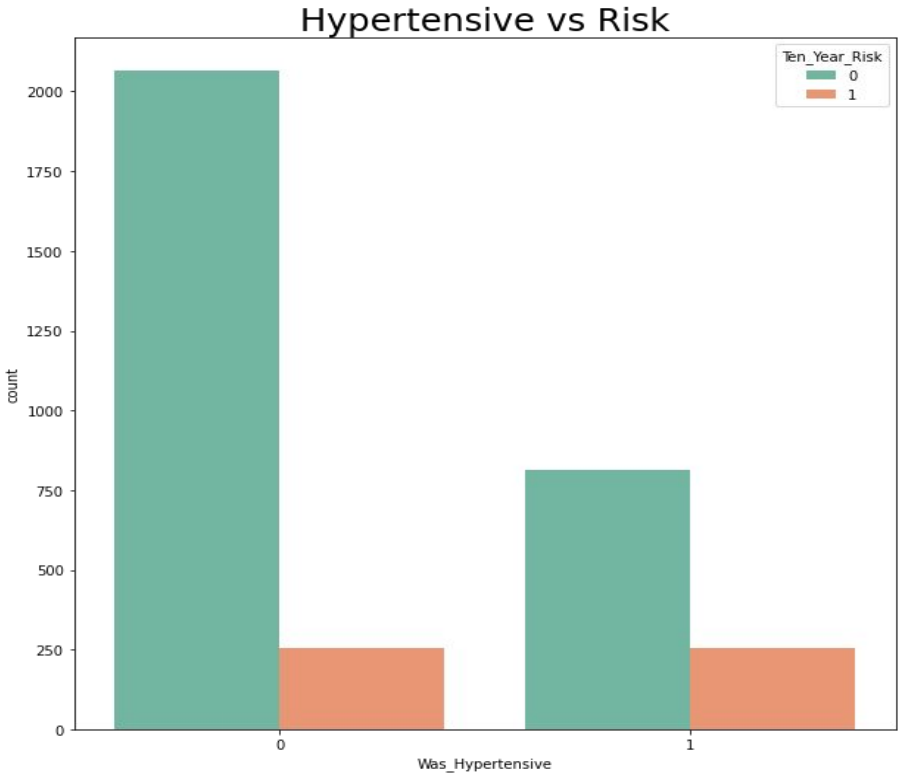
Risk With Respect To Stroke:



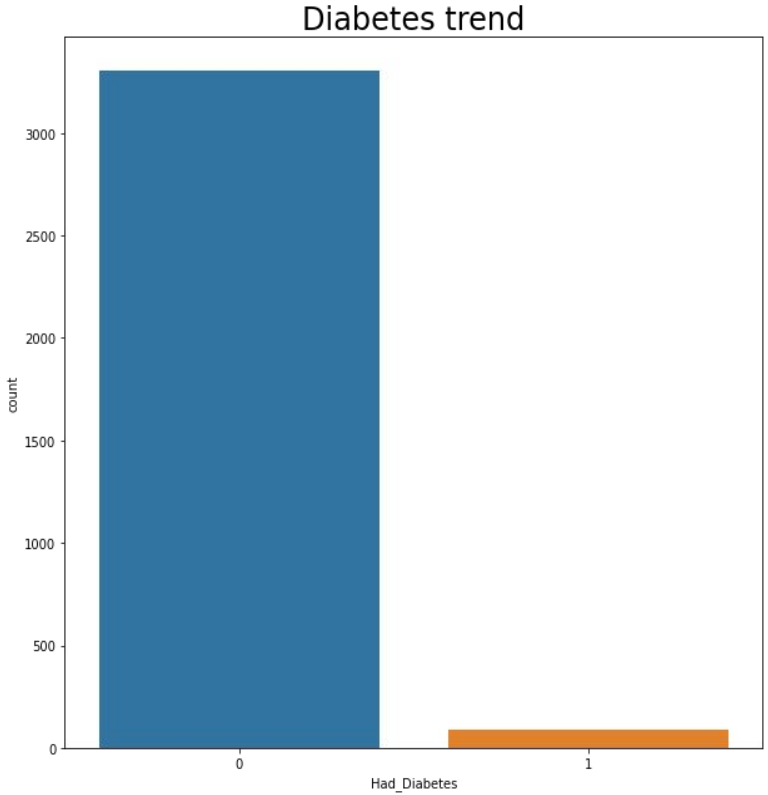


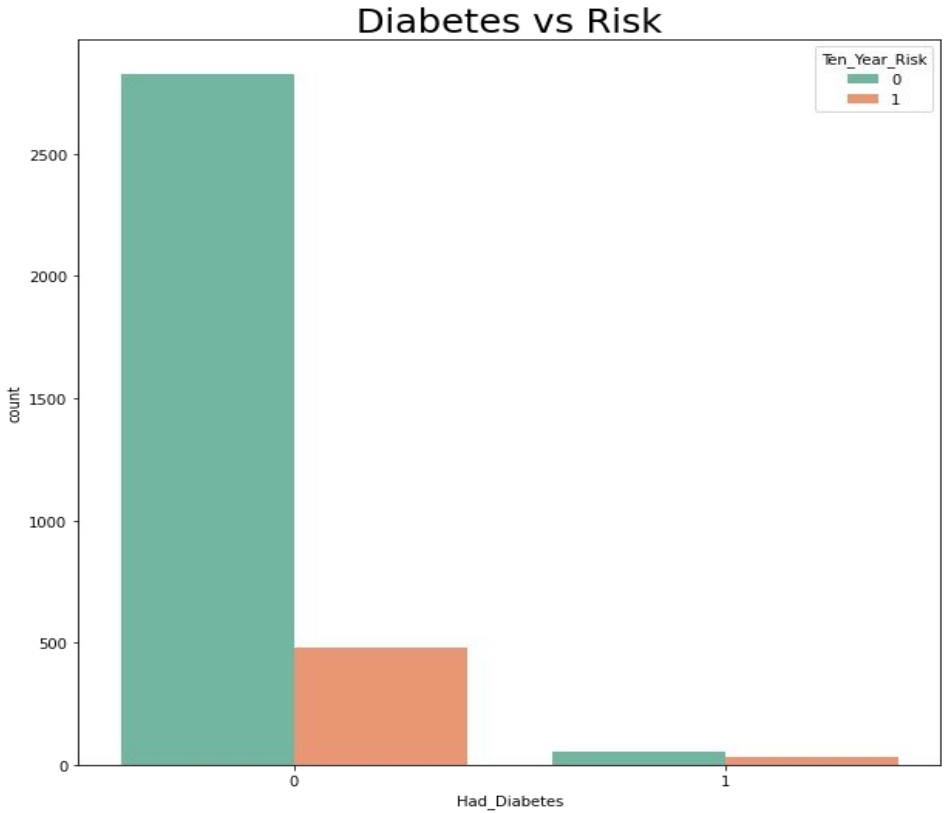
Risk With Respect To Hypertensive:



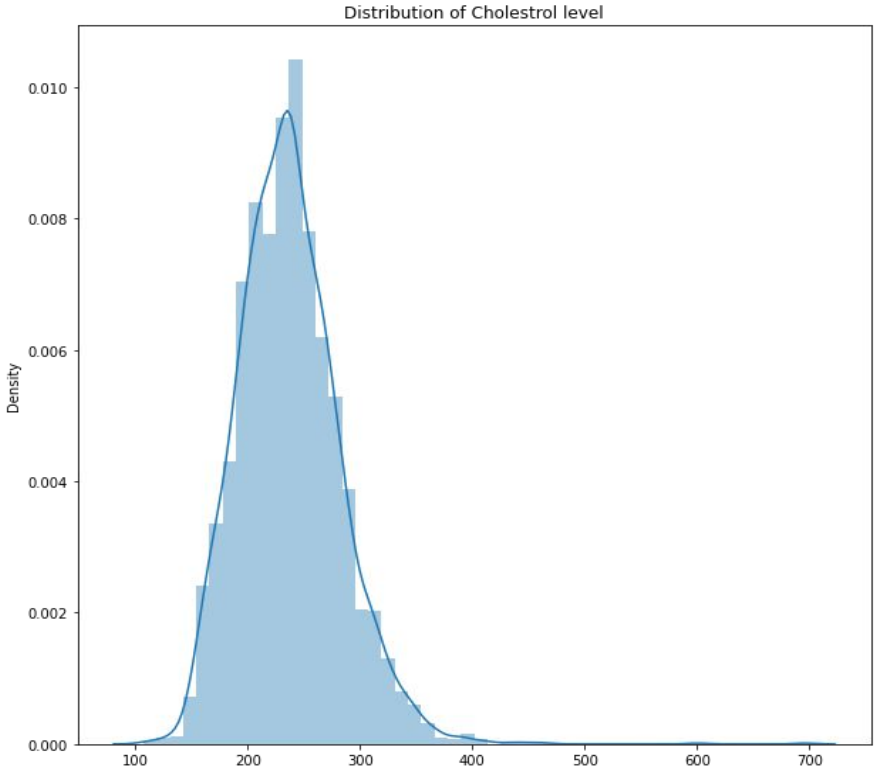


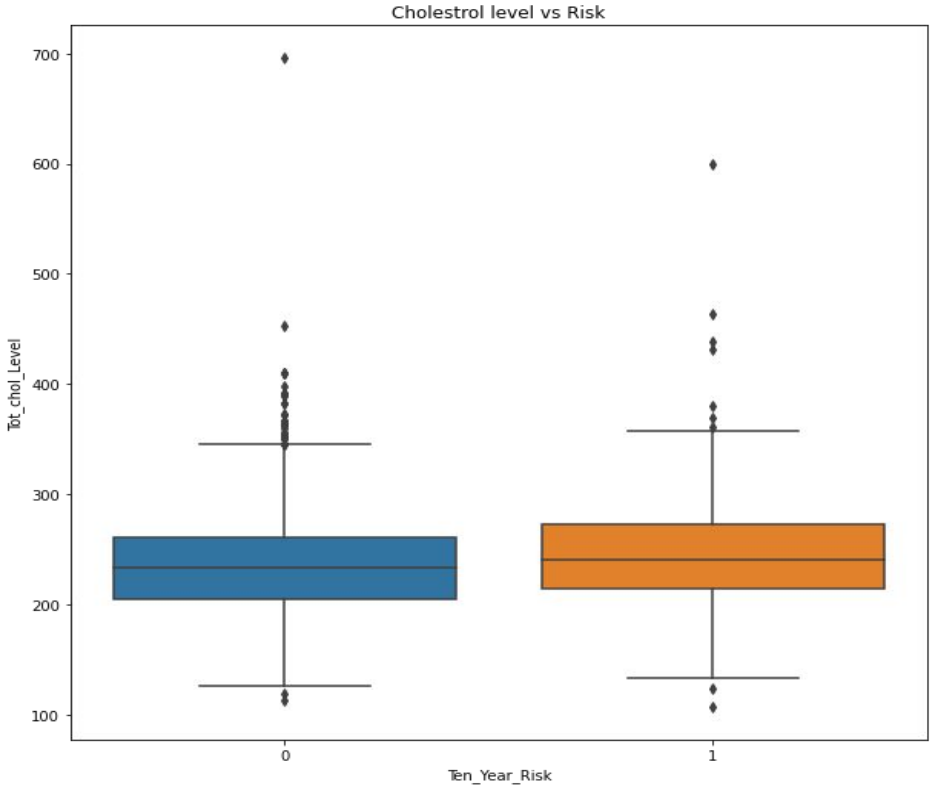
Risk With Respect To Diabetes:



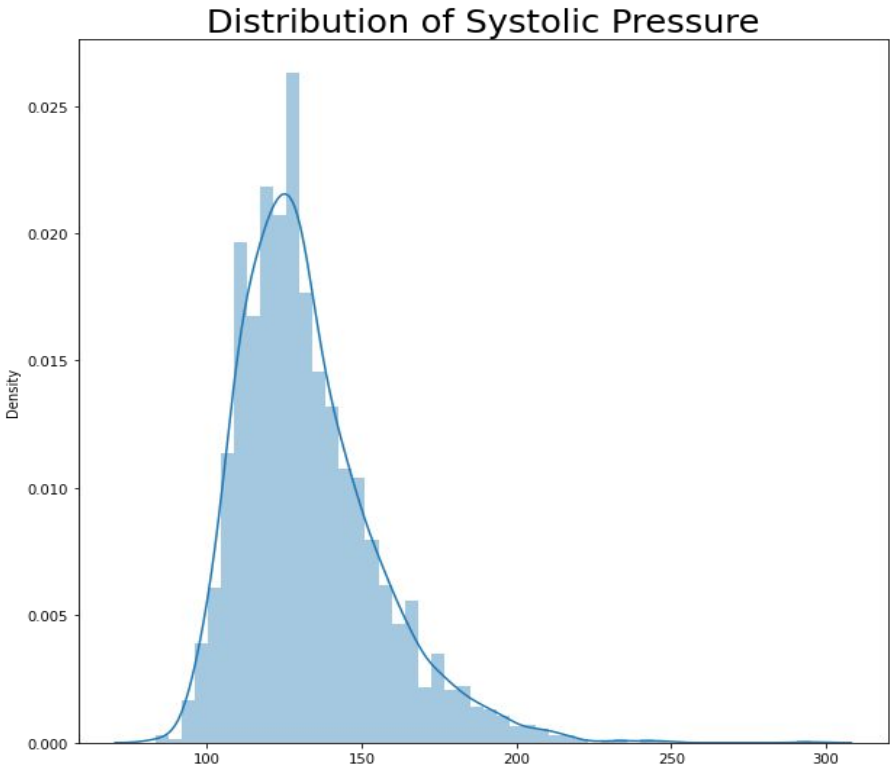


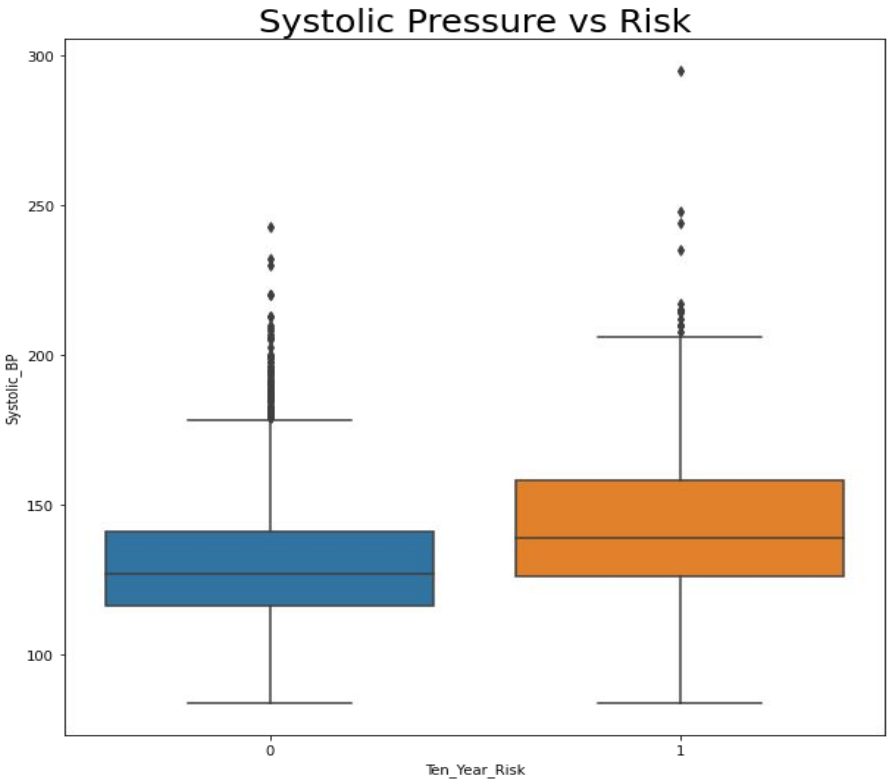
Risk With Respect To Cholesterol level:



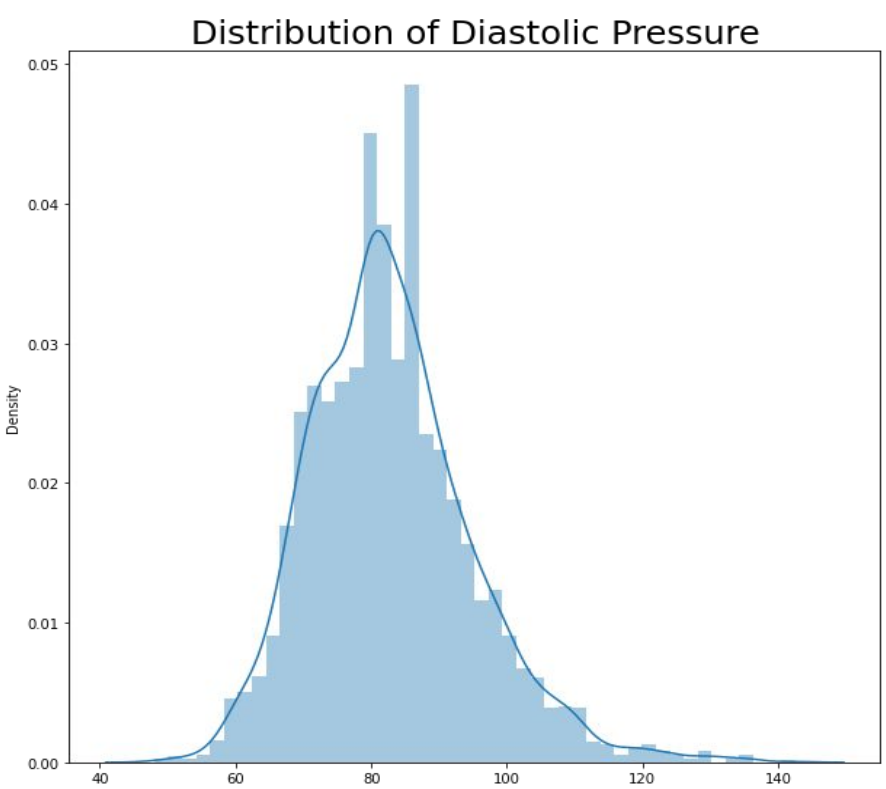


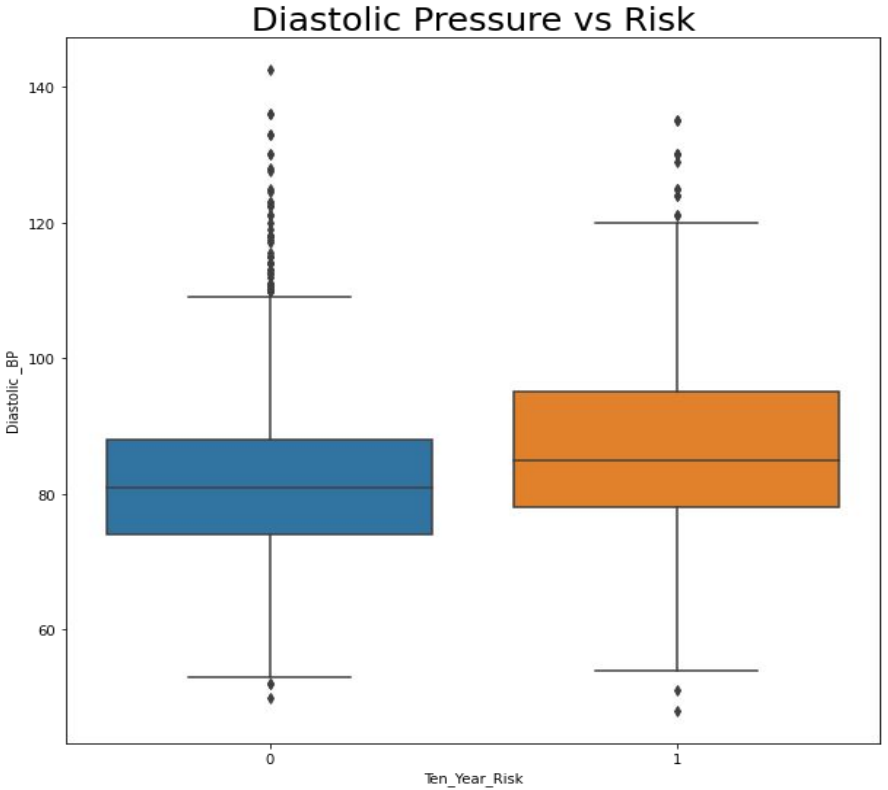
Risk With Respect To Systolic Pressure:



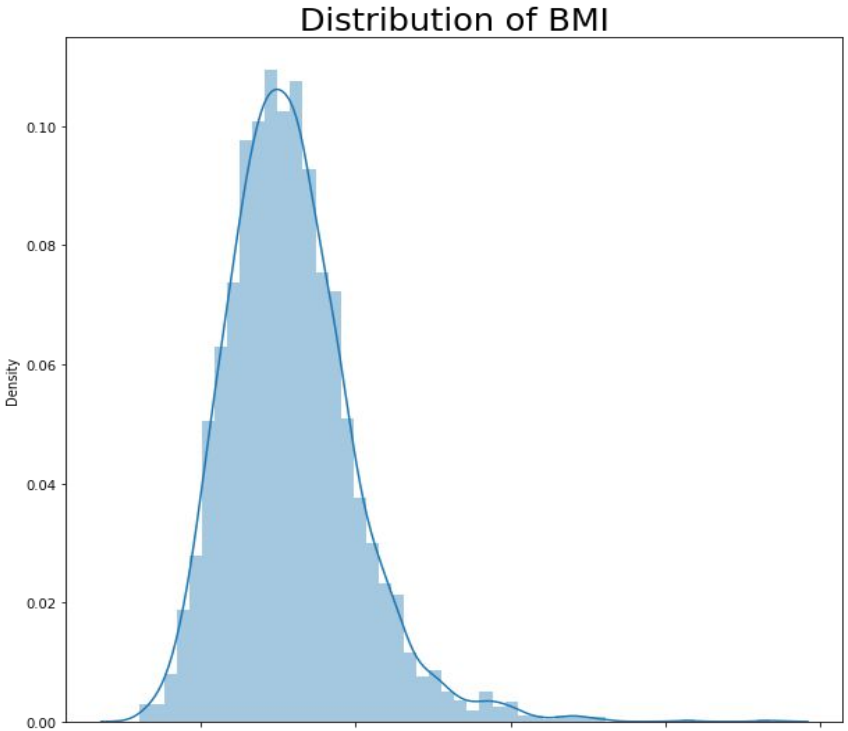


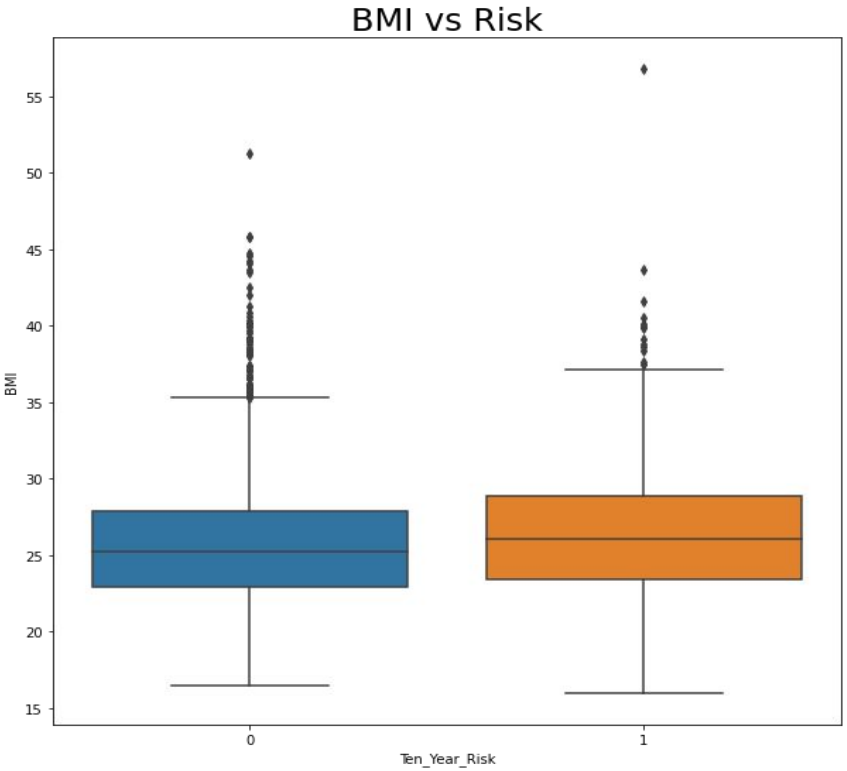
Risk With Respect To Diastolic Pressure:



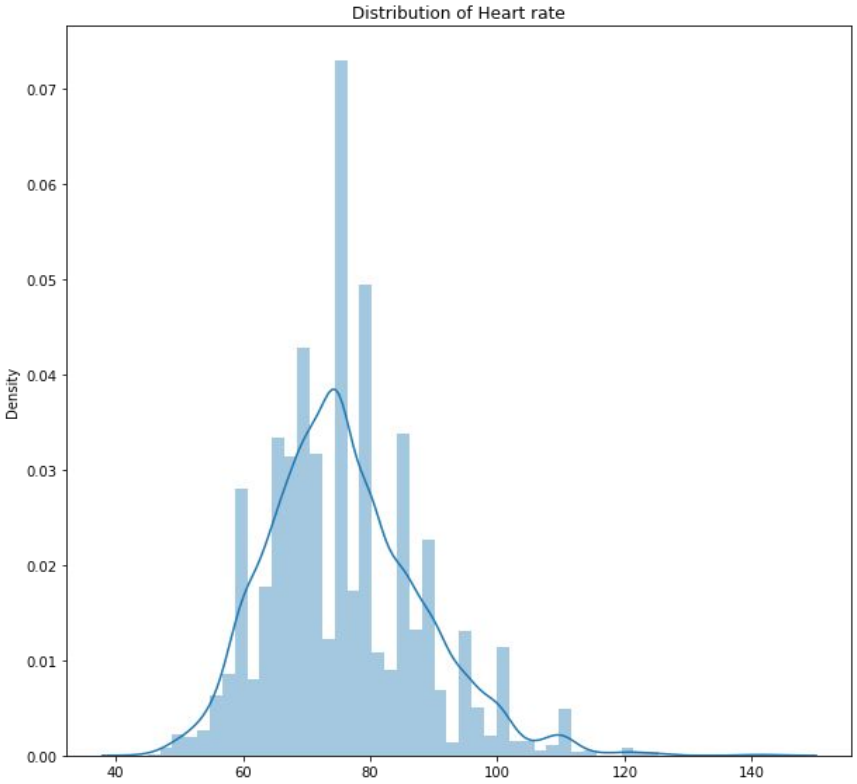


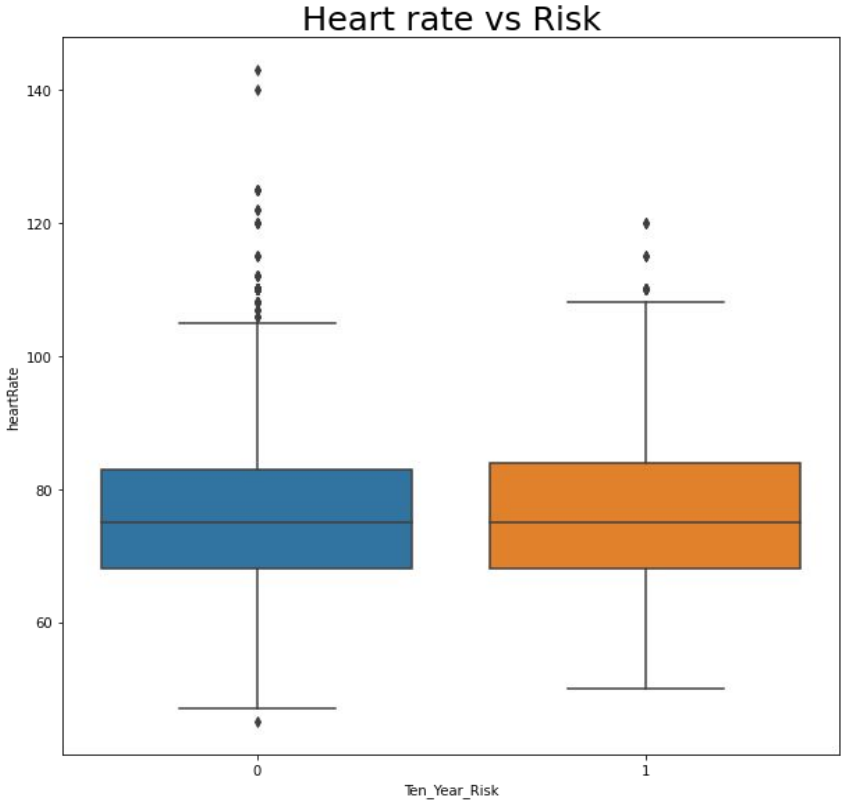
Risk With Respect To BMI:



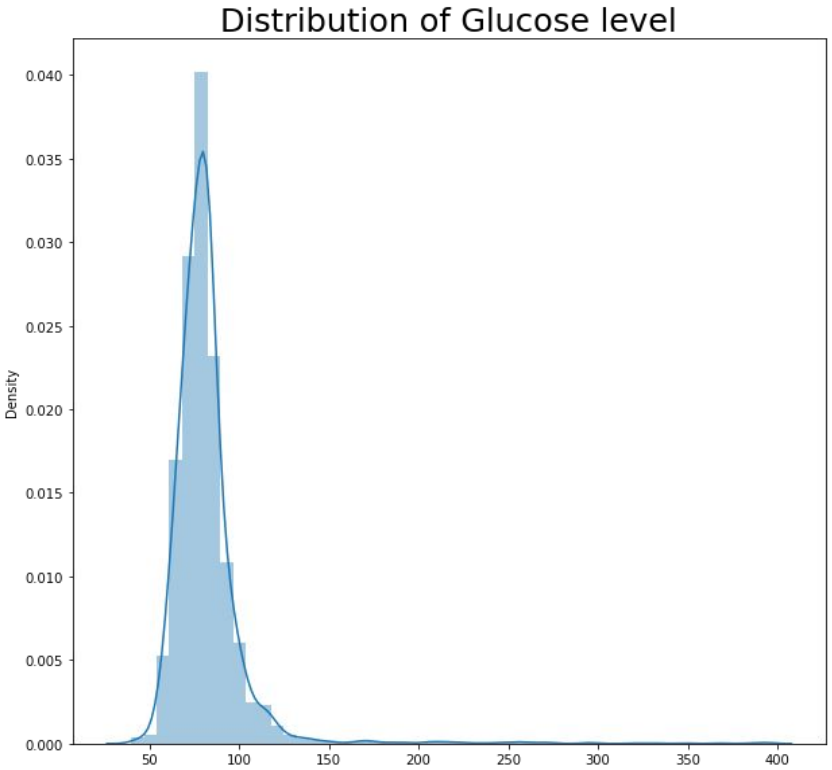


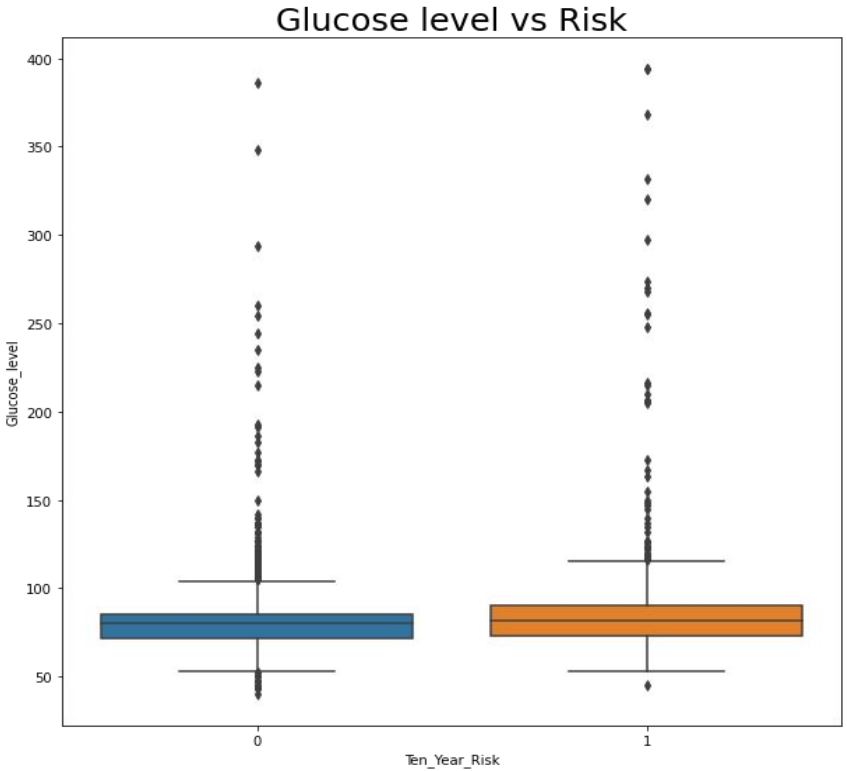
Risk With Respect To Heart Rate:



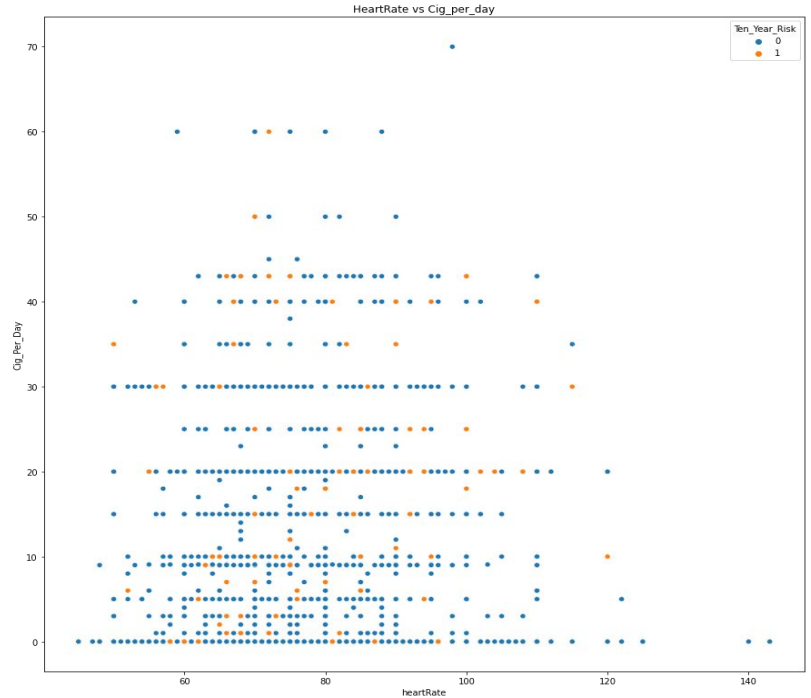


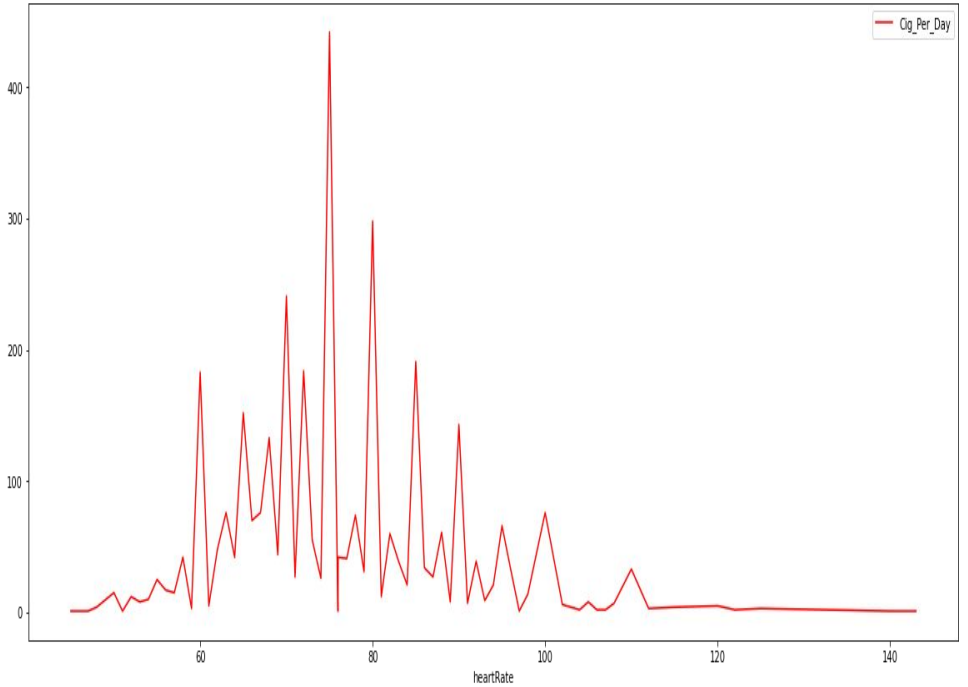
Risk With Respect To Glucose Level:



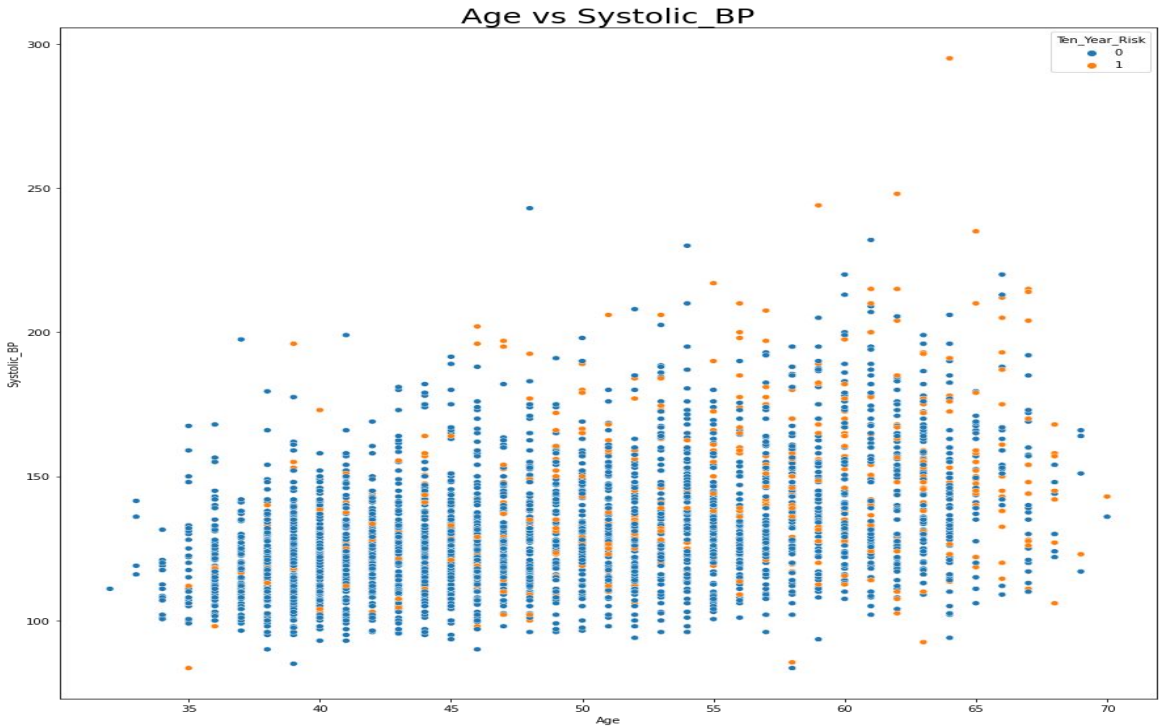


Heart Rate With Respect To Cigarettes Per Day:

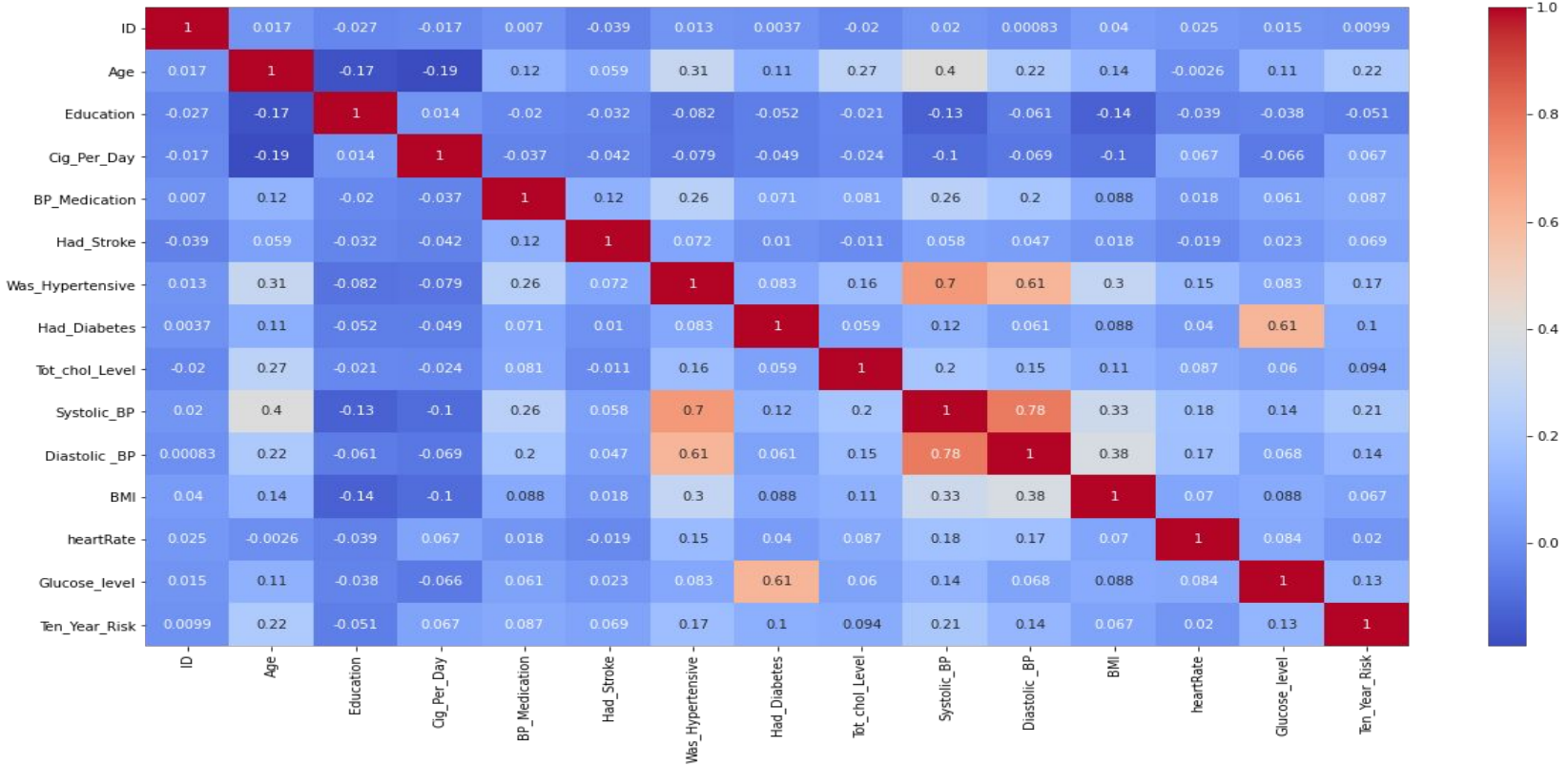




Systolic BP With Respect To Age:



Correlation Matrix:



Data Preprocessing:

Data After Converting to Numerical :

Rows: 3390

Columns: 18

Which columns we have converted ?

- Smoking
- Gender

Models Used:

1. Logistic Regression
2. Support Vector Machine (SVM)
3. SVM with Linear Kernel
4. SVM with Polynomial Kernel
5. SVM with Gaussian Radial Base Kernel
6. Neural Networks
7. Random Forest Classifier
8. XGBoost

Model Validation and Selection:

Observation 1:

At first I've tried Logistic Regression but scores were not that satisfying.

Observation 2:

Then I've tried using Support Vector Machine, i got good scores as compared to logistic regression. Then I used SVM with Kernel Tricks I have used Linear kernel, Polynomial Kernel and Gaussian Radial Base Kernel I got good scores with polynomial kernel and radial base kernel.

Observation 3:

Then I've used Neural networks with one hidden layer but we did not got good scores.

Observation 4:

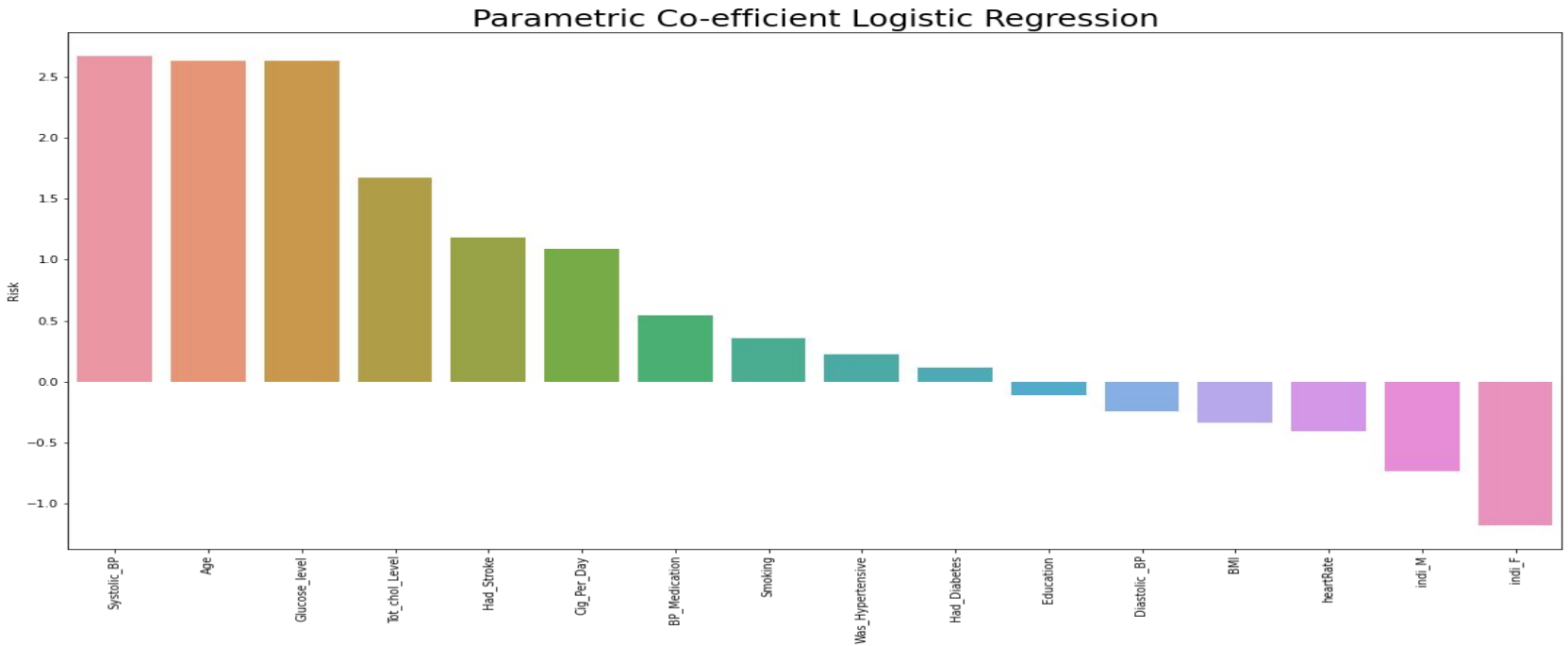
And lastly I've used Ensemble Learning models like Random Forest Classifier and XGBoost Classifier and both algorithms performed really well as compared to other models and with XGBoost I got the best Scores, so my optimal model is XGBoost Classifier.

Model Validation and Selection:

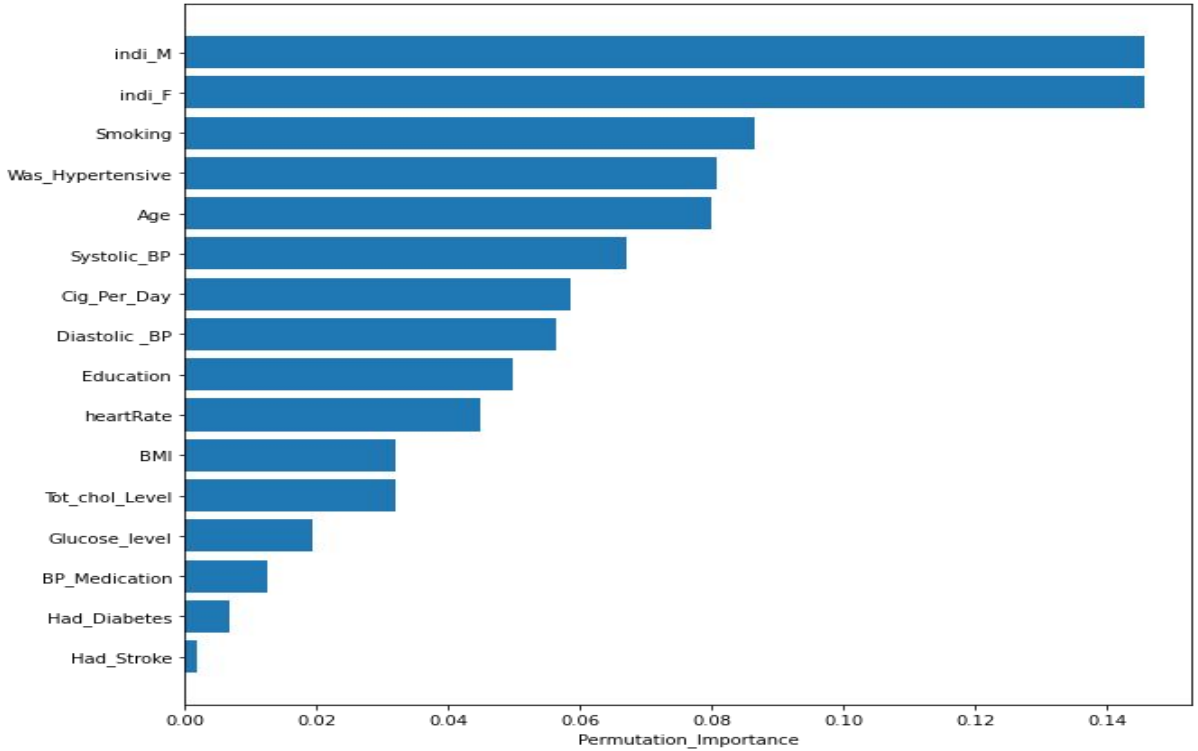
As XGBoost had performed really well so the best hyperparameters we got are:

```
'colsample_bytree': 0.7
'learning_rate': 0.03
'max_depth': 7
'min_child_weight': 4
'n_estimators': 500
'nthread': 4
'objective': 'reg:linear'
'silent': 1
'subsample': 0.7
```

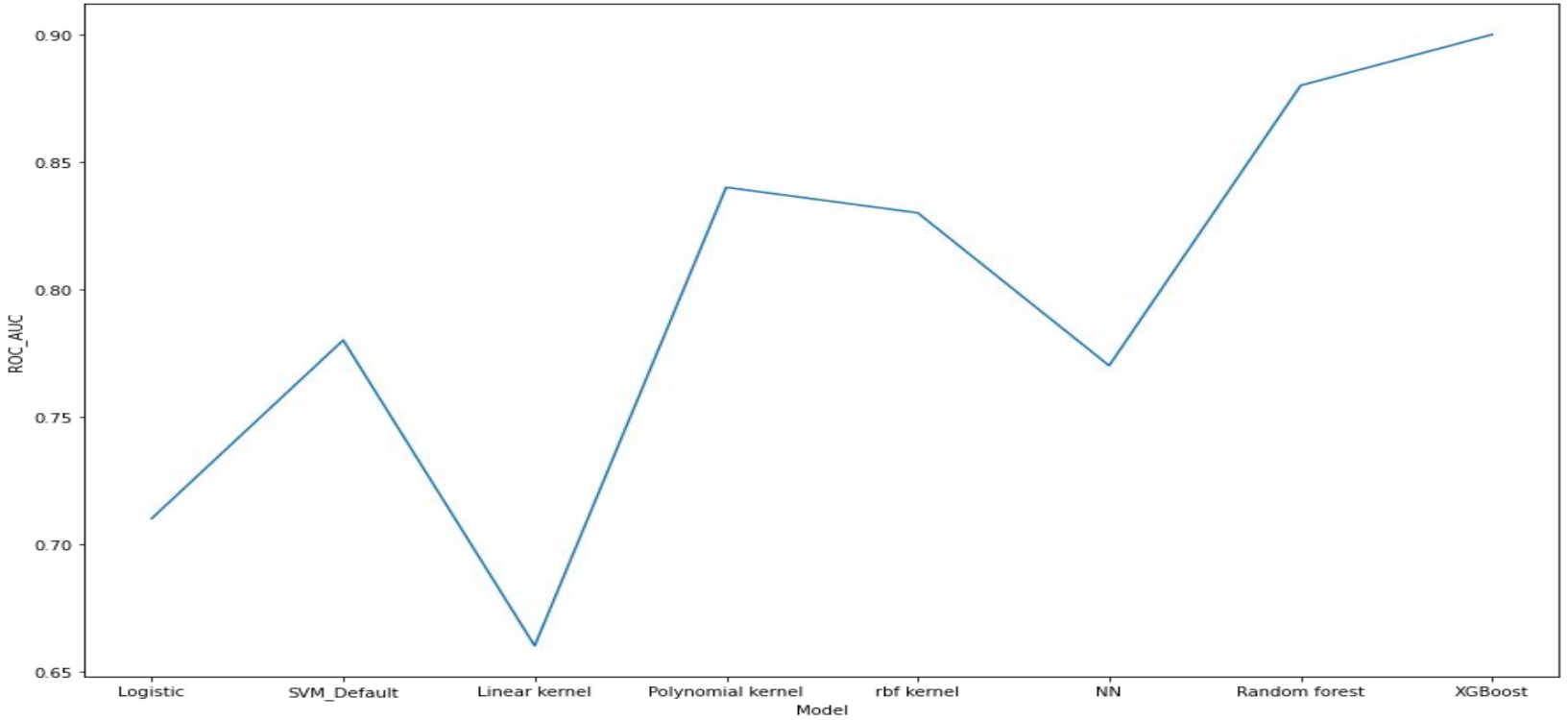
Feature Importance for Logistic Regression:



Feature Importance for Support Vector Machine:



ROC AUC Comparison:



Challenges:

- Execution takes time.
- As there were many null values present in data set it took time to clean the dataset.
- Difficulty in selecting the appropriate graph for trend.

Summary:

I started the study by handling missing values, our dataset was having many columns with null entities I have removed null values using KNN Imputer and Simple imputer I've seen how smoking, systolic BP, diastolic BP, BMI, Heart rate, glucose, hypertensive, cholesterol, diabetes, etc affects the person.

Factors like Blood Pressure, Glucose Level, Age had created a huge impact on a person's heart condition. We checked the correlations between the factors. Handled the class Imbalance using SMOTE and experimented with a combination of SMOTE + Tomek links. SMOTE gave a good result of 50-50 class balanced data.

Then I started building classification models. I started with Logistic regression with default parameters but I did not get a good score.

Then I used a Support Vector Machine with various Kernel tricks with respective hyperparameters. I have used linear kernel, polynomial kernel and gaussian radial based kernel. Polynomial and rbf kernel gave good scores though it took a large amount of time in rendering.

Then I tried a neural network with a single hidden layer there we got decent numbers but less as compared to the rbf poly kernel.

Then I tried the Random Forest Classifier and I got great Scores. Random Forest Classifier performed well.

And at last I've tried XGBoost Classifier and it had 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!