Capstone 2 Milestone Report #1

TITLE:	Predicting Cardiovascular Heart Disease
AUTHOR:	Caitlin Jansson
DATASET :	https://www.kaggle.com/christofel04/cardiovascular-study-dataset-predict-heart-disea
GIT REPOSITORY:	https://github.com/CJEJansson/Springboard_Projects
SLIDES:	Presentation Slides

PROBLEM STATEMENT

Cardiovascular heart disease (CHD) is the leading cause of death annually worldwide. Cardiovascular diseases can, however, be managed if caught early and simple lifestyle changes are made. This project explores a set of data for patients measuring known factors for heart disease to develop a machine learning model to predict risk of developing heart disease within the next ten years.

GENERAL OVERVIEW OF THE DATASET:

The data is provided from the Kaggle.com HME Workshop on Oct 3, 2020 targeted at increasing prediction of risk of developing CHD within the next ten years. This dataset is a subset of the Framingham, MA heart study data set. This data consists of a large group of initially "healthy" patients between the ages of 30-59 who were then tracked for 20 years to determine if they developed CHD [1]. The subset of data utilized in this project is divided into a test (80%) and train (20%) dataset and contains information on over 4,200 patients. The majority of the feature data has been converted to ordinal format (typically numeric), and there is no descriptive data included.

Overall the data was relatively clean, after a brief exploration. The ratio of male to female patients is unbalanced, with the dataset being approximately 43% male. The ratio of smokers to non-smokers is also unbalanced, with the dataset being 50.5% nonsmokers. When reviewing the feature for the number of cigarettes smoked per day, it was found that a null value was reported if a smoker did not divulge the number of cigarettes each day. All non-smokers were assigned zero values. Of those patients that did smoke, the majority reported smoking one pack a day (assuming 20 cigarettes per pack).

During exploratory analysis it was determined that the ages of patients do in fact range from 32-70, which is consistent with 20 years of observation for patients starting at approximately age 30. Patients ranged in education level from some high school to a higher education degree, while some did not report their education level.

To simplify later analysis, both gender and smoking status were converted to ordinal values, consistent with the rest of the data set. Values were assigned to Male/Female as 0/1, and Non-smoker/Smoker as 0/1. This will allow for analysis as though the data is continuous, rather than discrete. Converting all continuous data to discrete values (eg. blood pressure to Low, Normal, High, etc) was considered, but it was decided to take the more expedient route. The use of actual continuous values for health measurement data will also allow for potentially more accurate models during the model development phase.

A summary of observations for minimum and maximum values can be seen in Appendix A, in the table "Summary of Dataset by Feature".

INITIAL FINDINGS:

After cleaning, the dataset was explored via Exploratory Data Analysis (EDA) and hypothesis testing was used by leveraging inferential statistics. This analysis was started by using a series of

As expected, demographic data was not of much interest, with the exception of age. Age had the highest correlation with risk of developing Cardiovascular Heart Disease in the next ten years. Other significant factors were blood pressure, body weight, and other preexisting health conditions. This was expected as it is consistent with the results of the Framingham Heart Study's researcher identified milestones [2]. Those milestones were reviewed after completing data exploration and statistical analysis, and were consistent with the findings of this project.

Three methods were utilized to study the correlation between each feature and the risk of developing CHD in the next decade. First a correlation barplot was created to quickly visualize the magnitude of each relationship, seen below in Figure 1. Then a heatmap was created and modified to return only the bottom half of the results - to eliminate duplication. It is not necessary to report the entire heatmap as these variables are symmetric. If the dataset were asymmetric (the value of x is known but y cant be determined, and if the value of y is known the x is guaranteed), it would be necessary to return the entire heat map. This map can be seen in appendix A. The full heat map can be seen in the attached code.

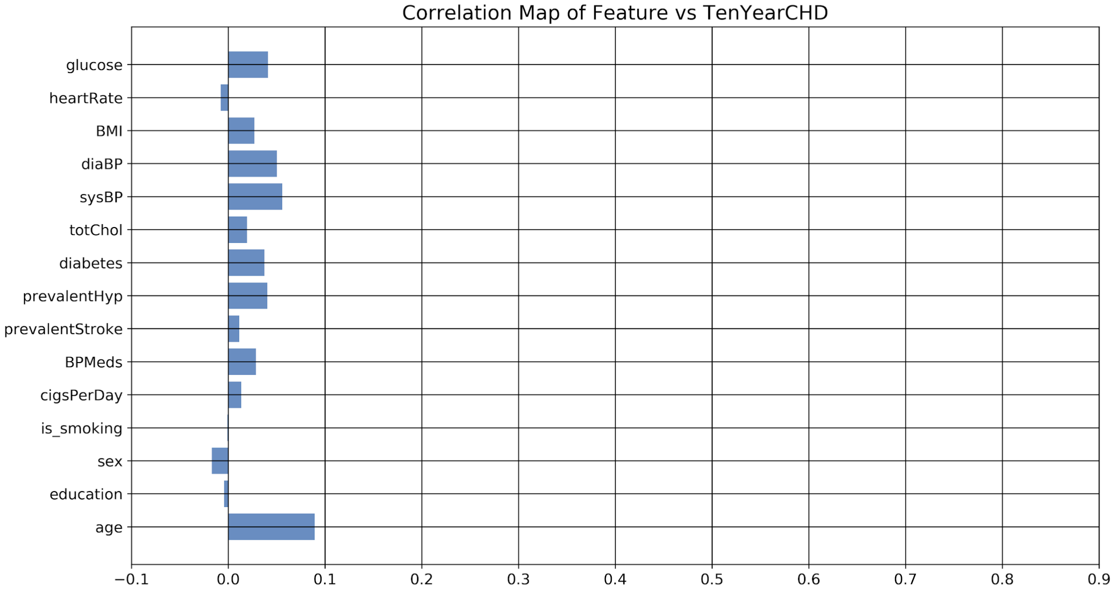


FIGURE 1: Barplot of correlation between each feature and the risk of developing CHD in the next ten years.

Finally, the correlation was visualized by plotting distribution plots to show the influence of each feature on the predicted value (risk of developing cardiovascular heart disease in the next 10 years). To do this, each feature will be plotted using a Kernel Density Estimation (KDE). The KDE shows the non-parametric probability density function of each data feature. This plot is created for each scenario - one for cases where CHD is observed, and a second for cases in which CHD is not observed. The amount of overlap of each density function is then compared to determine correlation. If there is a high overlap of both density functions for the feature, it is implied that there is little to no correlation between the dependent(CHD) and independent(each feature) variables. The less overlap between plots, the more significant the correlations.

For those variables that are discrete, each category will have 2 plots. For example, when looking at the KDE for gender, there are 4 curves. To determine correlation one must compare the CHD curves for male, and then consider the CHD curves for female patients. The comparison for gender vs heart disease must be visualized using a different method. The completed KDE can be seen in Appendix A for all features in the dataset. This showed consistent results with the other methods.

Those variables found to have a significant correlation were unsurprisingly, age, previous medical conditions (stroke, hypertension- and treatment of hypertension with blood pressure medication, diabetes), number of cigarettes smoked per day, and body mass (BMI). Of note, it was interesting that there was stronger correlation between the number of cigarettes smoked per day and the risk of developing CHD than when only studying whether or not a patient smoked. As expected, there is an increased risk of CHD with smoking, and this risk increases with the number of cigarettes smoked per day, as seen in Figure 2.

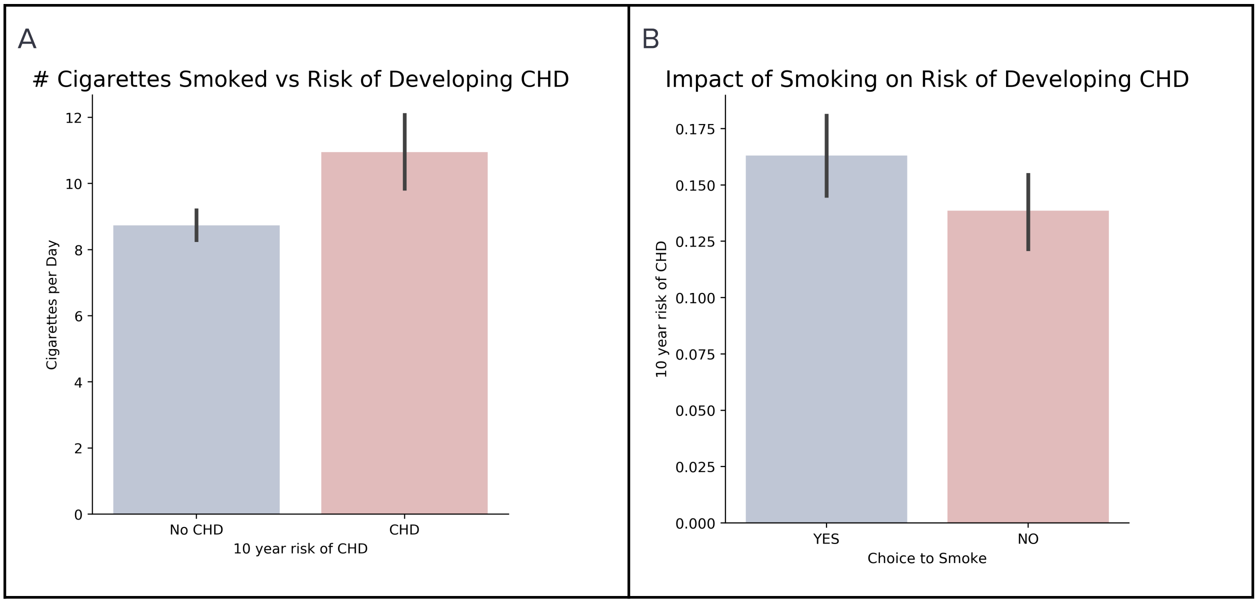
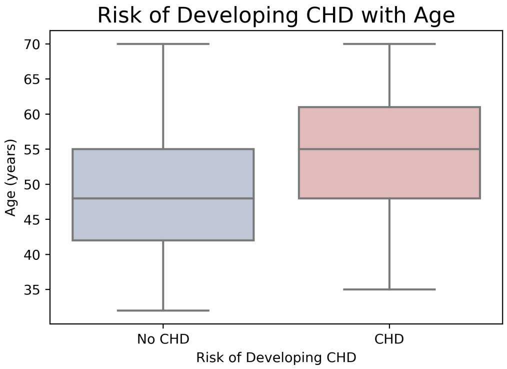


FIGURE 2 A) Risk of developing CHD as compared to # of cigarettes smoked per day; B) Risk of developing CHD with smoking

The risk of developing CHD was then investigated as patients age, since this was the most significant feature shown in the correlation plot. There was a distinct relationship between age and risk of developing CHD, showing that, on average, as patients age the risk increases. This relationship can be seen in the figure, to the right. This relationship was further explored to study the effect of age on prevalence of stroke, hypertension, and diabetes, all of which increased with age. Also of note, risk of developing CHD increased with prevalence of diabetes, with prevalence of stroke, and with prevalence of hypertension. These plots can be seen in Appendix A.



As blood pressure is known to have an effect on heart health and risk of stroke as well as other health problems, the effect of blood pressure was also investigated with regards to risk of CHD. It was found, as expected, that on average the risk of developing CHD increases with increased blood pressure. This can be seen in Figure 3 below:

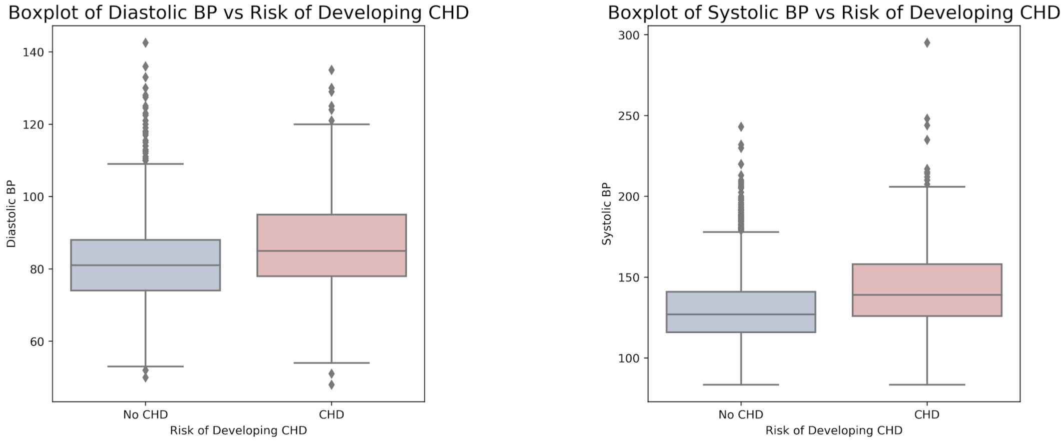


FIGURE 3 Effect of blood pressure, both systolic and diastolic vs risk of developing cardiovascular heart disease.

Gender was investigated, and it was found that, on average, males have a higher risk of developing heart disease, and the difference in this risk increases with age. Body weight also has an effect on risk of developing CHD, with risk increasing as BMi increases. These results can both be seen below in Figures 4, below, and Figure 5 on the next page.

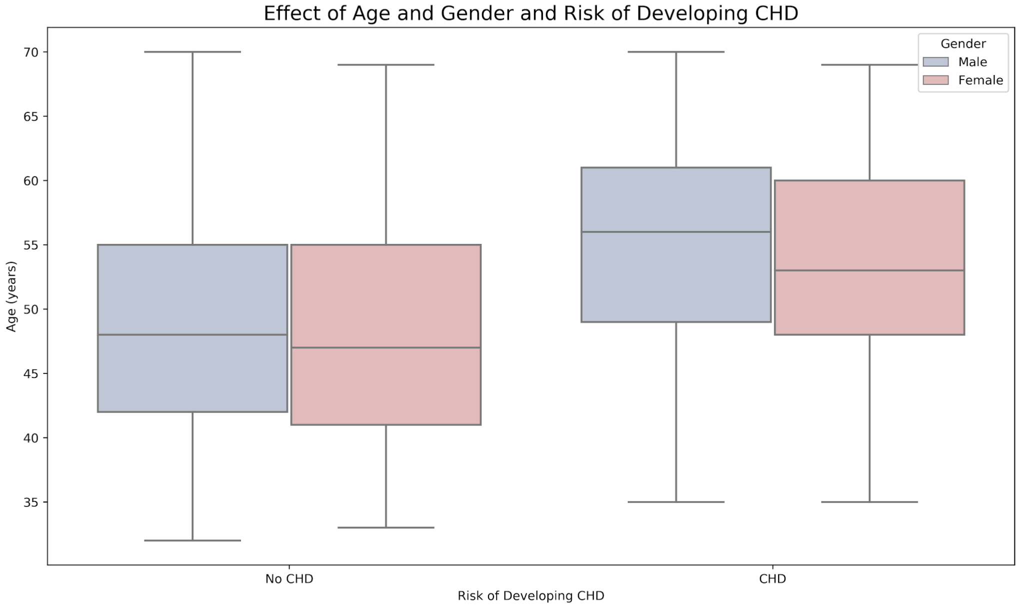


FIGURE 4 Effect of blood age and gender vs risk of developing cardiovascular heart disease.

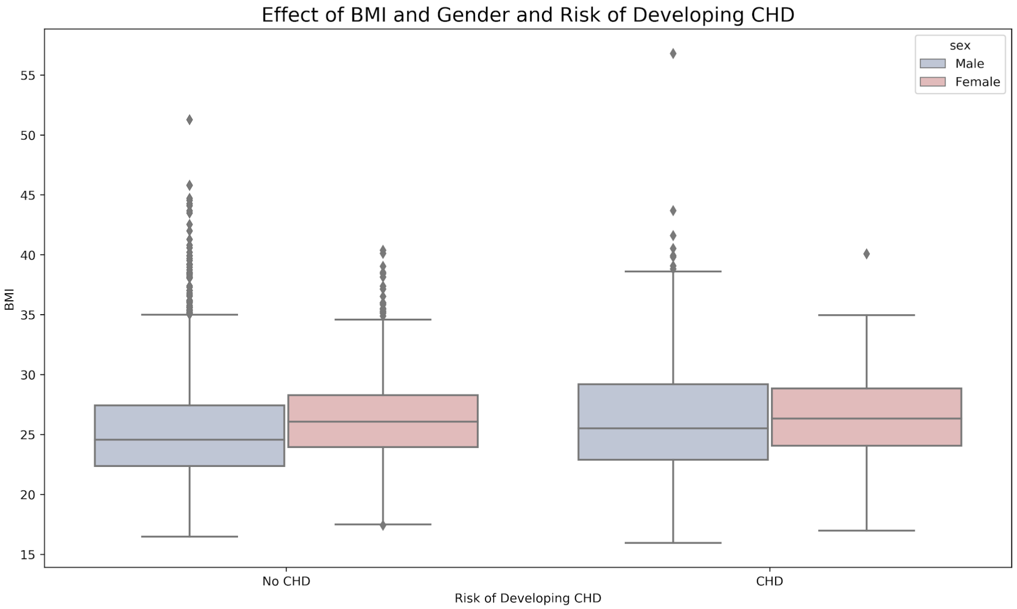


FIGURE 5 Effect of BMI and gender vs risk of developing cardiovascular heart disease.

The results of EDA were consistent with those found during the correlation analysis, and consistent with those observations made by other researchers who have explored this dataset. Now it's time to verify these findings via hypothesis testing, the details of which are discussed below.

STATISTICAL ANALYSIS :

After completing exploratory data analysis, Statistical analysis was conducted to perform high level hypothesis testing on each of the variables in the dataset. This was done to verify the correlation relationships revealed graphically, and will also serve to validate/verify findings from the Framingham Heart Study reported by researchers. Two types of statistical testing will be conducted, Chi-Squared testing and T-testing, based upon variable type.

Chi-Squared testing is used in situations when testing statistical independence or association between categorical variables. The following categorical variables will be tested using this method: education level, gender, choice to smoke, use of blood pressure medication, prevalence of stroke, prevalence of hypertension. Hypotheses tested were set up to determine whether there was or was not an association between the feature and risk of developing CHD.

T-testing was used to compare the means between groups of continuous data. This method was chosen rather than ANOVA because only two groups of data are being compared to determine if they are different. The response variables that will be tested using this method are continuous.

These variables included age, number of cigarettes smoked per day, total cholesterol, blood pressure, BMI, heart rate, and blood glucose. Hypotheses were set up to determine if patients in each group (with and without CHD) have the same risk of developing heart disease given the value of each feature.

During hypothesis testing, the majority of results were consistent with those found during the correlation analysis. Relationships that were interesting and of note were found. For example, hypothesis testing showed the association between smoking and CHD was not significant. However, when testing the number of cigarettes smoked per day and CHD three was a significant relationship found. While this is consistent with correlation values, the difference in significance was interesting. Another interesting observation was that a patient's resting heart rate had no significant association with risk of developing CHD. This is interesting because an elevated resting heart rate can sometimes be indicative of heart disease or overall wellness. It is entirely possible there is still some small relationship, but that it is not significant given the value of alpha utilized, which was 5%.

Overall statistical results can be seen in Appendix A and are shown as compared to the correlation values from the heat map.

SUMMARY:

Overall, the data showed results expected, and the results found here were consistent with those results found by the researchers doing the study. The risk of developing CHD is increased with smoking, and the more a patient smokes per day the higher the risk. High cholesterol and high blood pressure (looking at both systolic and diastolic) increase the risk of developing CHD. Increase of developing CHD increases with age. Patients with a prevalence of stroke, hypertension, and diabetes all also increase the risk of developing CHD with age.

Knowing from the full study that risk of CHD can be decreased, it would have been interesting to study this as part of the project, but this data was not included in the provided dataset for analysis. These insights can be leveraged by the client to target specific patient groups and ultimately lower the risk of developing CHD. It is hoped that these insights will lend themselves to the development of an accurate risk prediction engine as the project continues.

CITATIONS

[1] https://www.thelancet.com/journals/lancet/article/PIIS0140-6736%2813%2961752-3/fulltext

[2] https://web.archive.org/web/20170710152157/https://www.framinghamheartstudy.org/index.php

Appendix A

SUMMARY OF DATASET BY FEATURE:

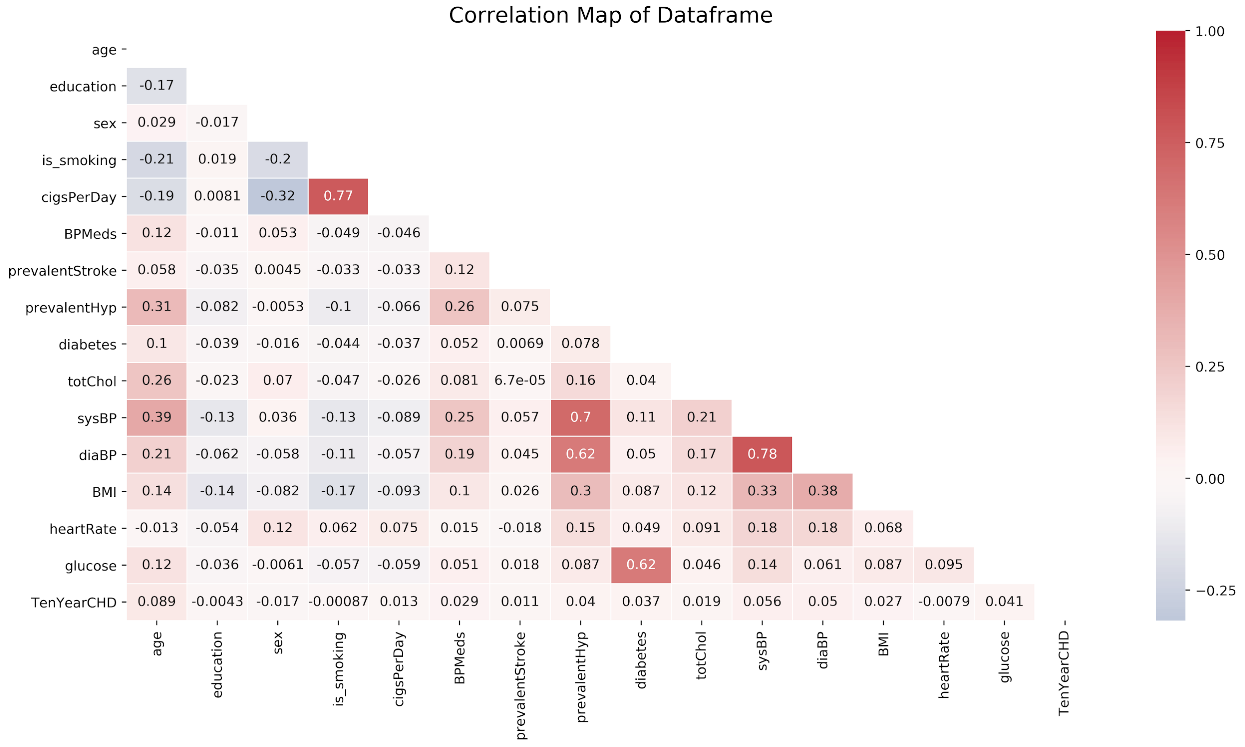
Feature	Values	Data Type	# Null	Description
Age	32-70	Continuous	None	Patient age in years, only whole numbers
Education	1, 2, 3, 4	Discrete	None	Education Level: 1-Some High School, 2-High School Diploma/GED, 3-College, 4-Degree
Sex	M, F	Discrete	None	Patient gender (M/F or 0=M, 1=F)
is_smoking	Yes, No	Discrete	None	If the patient is a current smoker (Yes/No or 1=yes, 0=no)
Cigs per Day	0-70	Continuous	29	Number of Cigarettes smoked per day (null = unknown)
BP Meds	0, 1	Discrete	53	Whether the patient is taking Blood Pressure Medications (0=no, 1=yes, null=unknown
prevalentStroke	0, 1	Discrete	None	Prevalence of stroke (0=none, 1=has had occurences of stroke)
prevalentHyp	0, 1	Discrete	None	Prevalence of hypertension (0=none, 1= has prevalence hypertension
diabetes	0, 1	Discrete	None	If the patient has diabetes (0=no, 1=yes)
totChol	107-696	Continuous	50	Total Cholesterol
sysBP	83.5-295	Continuous	None	Systolic Blood Pressure
diaBP	48-142.5	Continuous	None	Diastolic Blood Pressure
BMI	15-54-56.8	Continuous	19	Body Mass Index
heartRate	44-143	Continuous	1	Resting heart rate in beats per minute (bpm)
glucose	40-394	Continuous	388	Blood glucose level. (mg/dL)
TenYearCHD	0,1	Discrete (calculated)	848 (test data)	Risk of developing CHD in next decade (0=no risk, 1=risk)

STATISTICAL RESULTS OF FEATURE SPECIFIC INVESTIGATION:

*Correlation value reported vs risk of developing CHD in the next 10 years

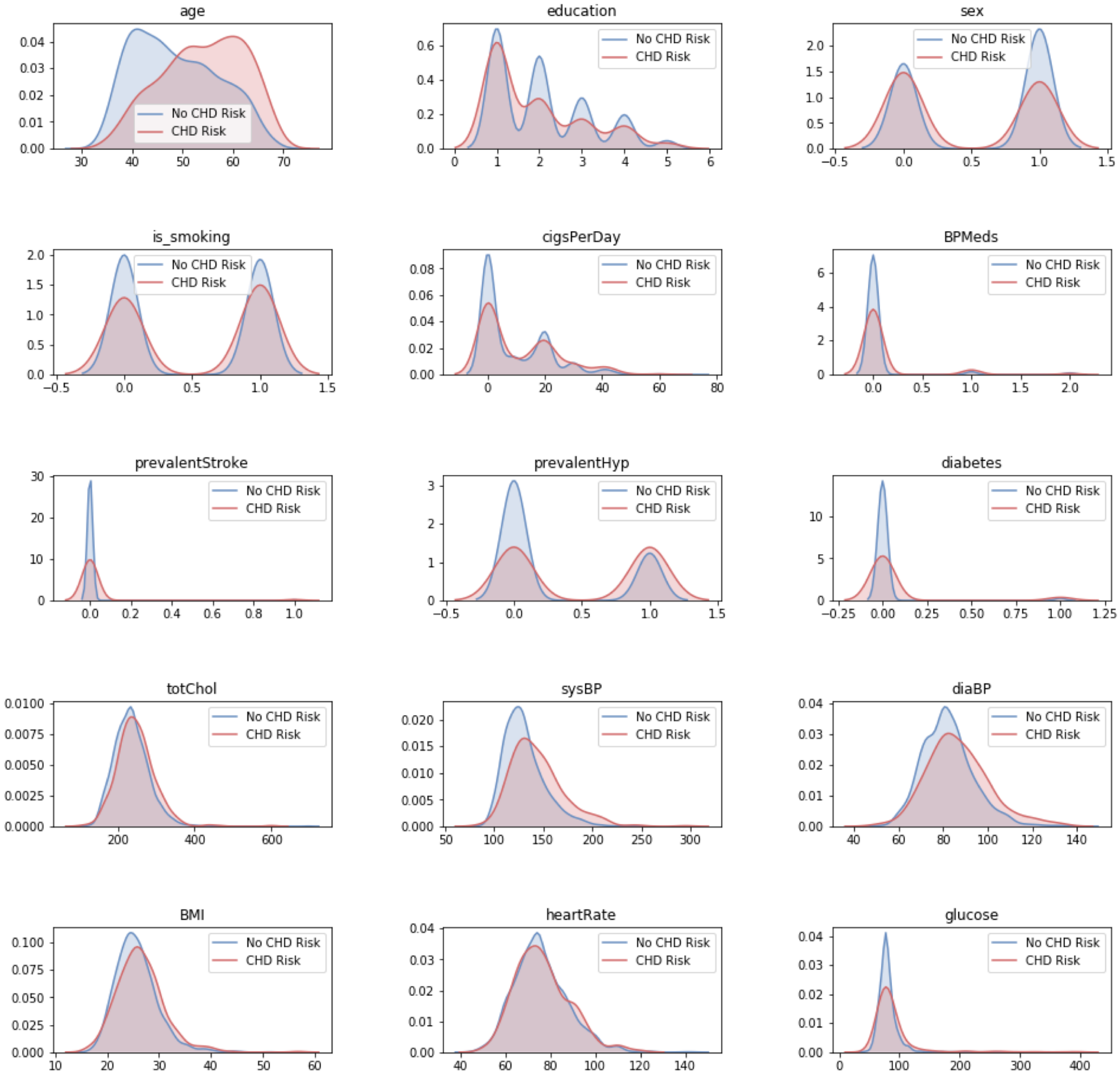
Feature	Correlation	Hypothesis Testing		
		Test	***Result***	***P-Value***
Age	0.0890	T-Test	Reject Ho - Correlated	1.85E-38
Education Level	-0.0043	Chi-Square	Reject Ho - Correlated	4.87E-04
Gender	-0.0170	Chi-Square	Reject Ho - Correlated	3.37E-06
Smoker	-0.0009	Chi-Square	Fail to Reject	9.07E-02
Number of Cigarettes per Day	0.0130	T-Test	Reject Ho - Correlated	5.37E-04
Use of BP Meds	0.0290	Chi-Square	Reject Ho - Correlated	2.07E-06
Prevalence of Stroke	0.0110	Chi-Square	Reject Ho - Correlated	9.32E-05
Prevalence of Hypertension	0.0400	Chi-Square	Reject Ho - Correlated	1.03E-21
Diabetic	0.0370	Chi-Square	Reject Ho - Correlated	1.27E-08
Total Cholesterol	0.0190	T-Test	Reject Ho - Correlated	5.31E-07
Systolic BP	0.0560	T-Test	Reject Ho - Correlated	5.52E-24
Diastolic BP	0.0500	T-Test	Reject Ho - Correlated	8.90E-12
BMI	0.0270	T-Test	Reject Ho - Correlated	4.45E-04
Heart Rate	-0.0079	T-Test	Fail to Reject	2.47E-01
Blood Sugar	0.0410	T-Test	Reject Ho - Correlated	3.41E-06

Correlation Heatmap showing correlation between features:

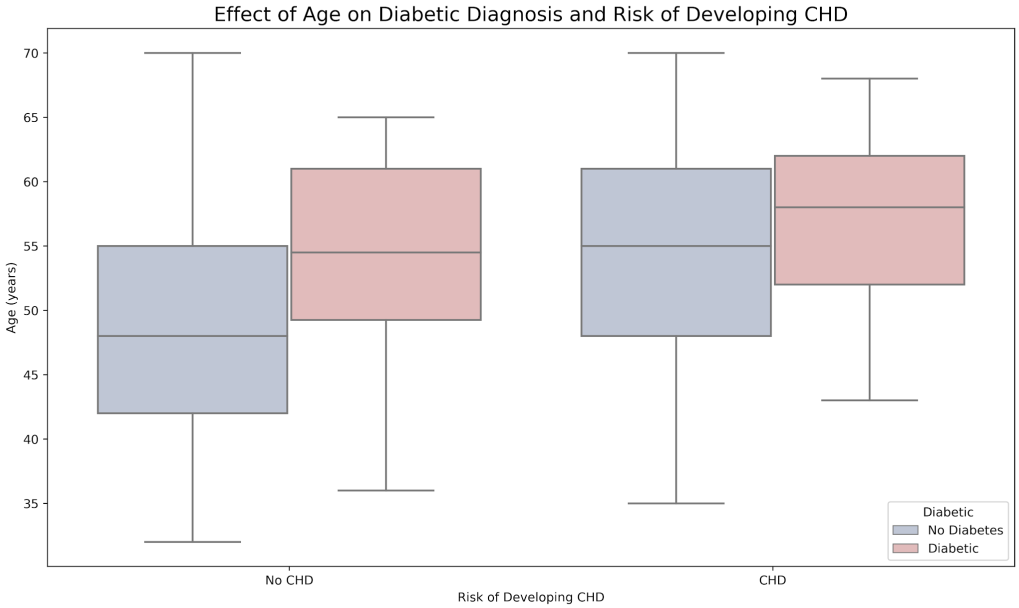


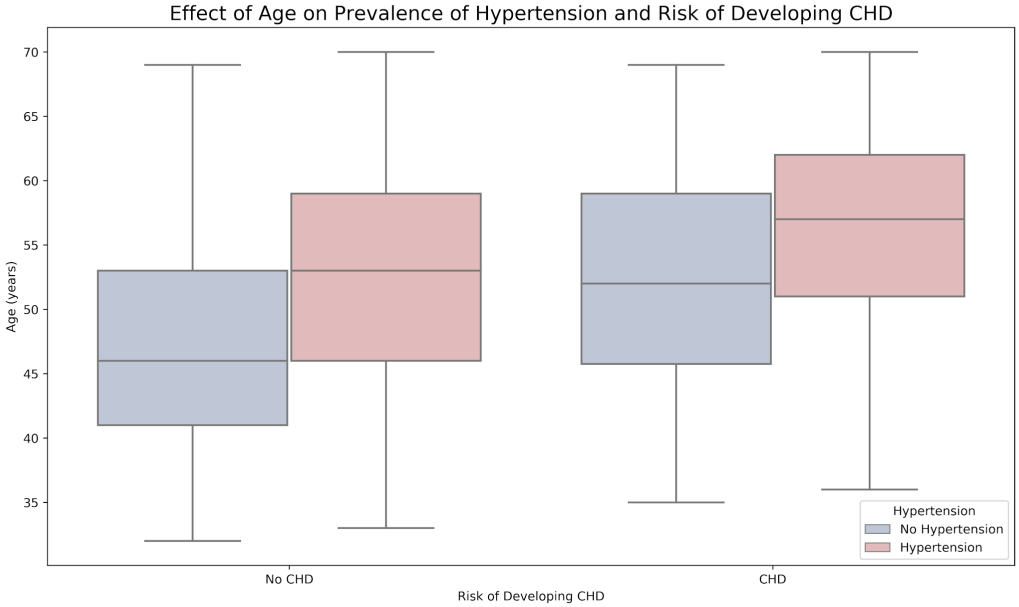
	age	education	sex	is_smoking	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
age																
education	-0.17															
sex	0.029	-0.017														
is_smoking	-0.21	0.019	-0.2													
cigsPerDay	-0.19	0.0081	-0.32	0.77												
BPMeds	0.12	-0.011	0.053	-0.049	-0.046											
prevalentStroke	0.058	-0.035	0.0045	-0.033	-0.033	0.12										
prevalentHyp	0.31	-0.082	-0.0053	-0.1	-0.066	0.26	0.075									
diabetes	0.1	-0.039	-0.016	-0.044	-0.037	0.052	0.0069	0.078								
totChol	0.26	-0.023	0.07	-0.047	-0.026	0.081	6.7e-05	0.16	0.04							
sysBP	0.39	-0.13	0.036	-0.13	-0.089	0.25	0.057	0.7	0.11	0.21						
diaBP	0.21	-0.062	-0.058	-0.11	-0.057	0.19	0.045	0.62	0.05	0.17	0.78					
BMI	0.14	-0.14	-0.082	-0.17	-0.093	0.1	0.026	0.3	0.087	0.12	0.33	0.38				
heartRate	-0.013	-0.054	0.12	0.062	0.075	0.015	-0.018	0.15	0.049	0.091	0.18	0.18	0.068			
glucose	0.12	-0.036	-0.0061	-0.057	-0.059	0.051	0.018	0.087	0.62	0.046	0.14	0.061	0.087	0.095		
TenYearCHD	0.089	-0.0043	-0.017	-0.00087	0.013	0.029	0.011	0.04	0.037	0.019	0.056	0.05	0.027	-0.0079	0.041	

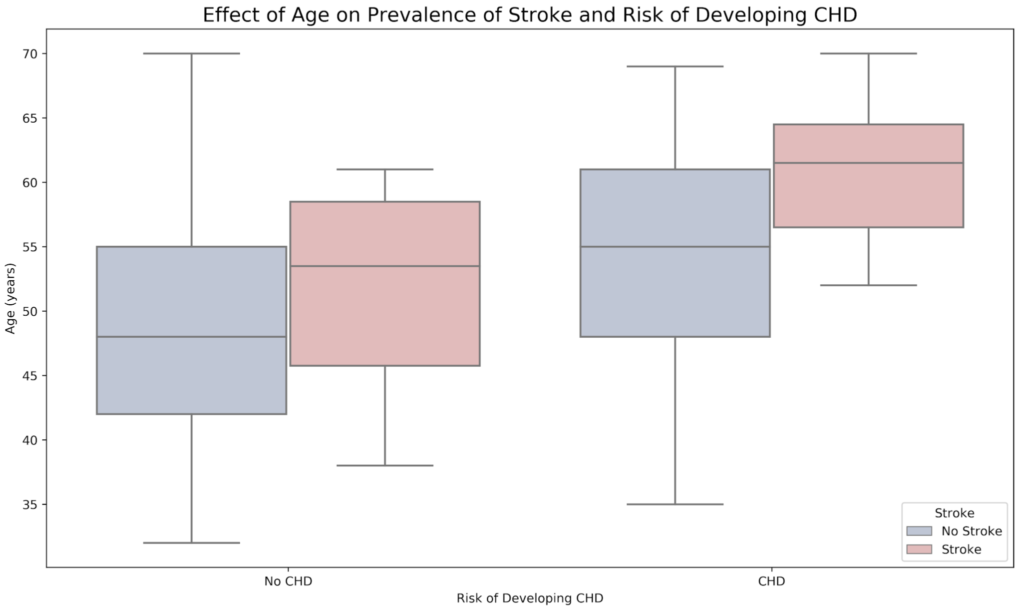
Kernel Density Estimation Distribution For Each Feature vs TenYearCHD showing correlation between features:



Boxplots of Features vs Risk of Developing CHD (by Feature)






Effect of Age on Prevalence of Stroke and Risk of Developing CHD
Age (years)
70
65
60
55
50
45
40
35
No CHD
CHD
Risk of Developing CHD
Stroke
No Stroke
Stroke