**DOCUMENTATION**

**Heart Disease Prediction**

**World Health Organization has estimated 12 million deaths occur worldwide, every year due to Heart diseases**

1. **The goal of gathering this data**

* Scientists from around the world can communicate with each other thanks to the collecting of this information.
* In order for the WHO to monitor cardiovascular illnesses.
* For scientists to be able to have access to one document that shows statistics on the heart disease risk factors.

1. **What problem/s is this dataset intended to solve**

* This dataset aims to eliminate disarray
* Give precise details on the most important heart disease risk factors.

1. **Why do I need to use this dataset, and what issue am I trying to resolve?**

* The goal of this study is to identify the most important heart disease risk variables and to use logistic regression to estimate the overall risk.

**List of columns**

linkcode

sex: male or female

age: age of the patient

currentSmoker: whether or not the patient is a current smoker

cigsPerDay: the number of cigarettes that the person smoked on average in one day

BPMeds: whether or not the patient was on blood pressure medication

prevalentStroke: whether or not the patient had previously had a stroke

prevalentHyp: whether or not the patient was hypertensive

diabetes: whether or not the patient had diabetes

totChol: total cholesterol level

sysBP: systolic blood pressure

diaBP: diastolic blood pressure

BMI: Body Mass Index

heartRate: heart rate

glucose: glucose level

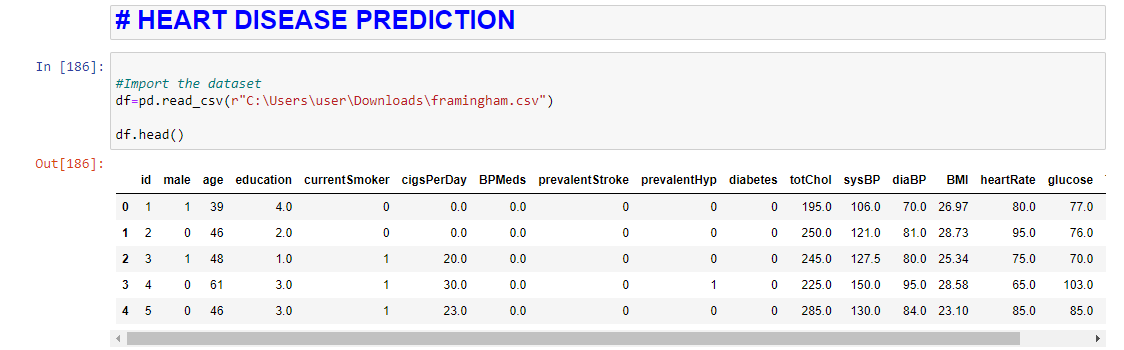
TARGET COLUMN:

10 year risk of coronary heart disease CHD

**Data Understanding**

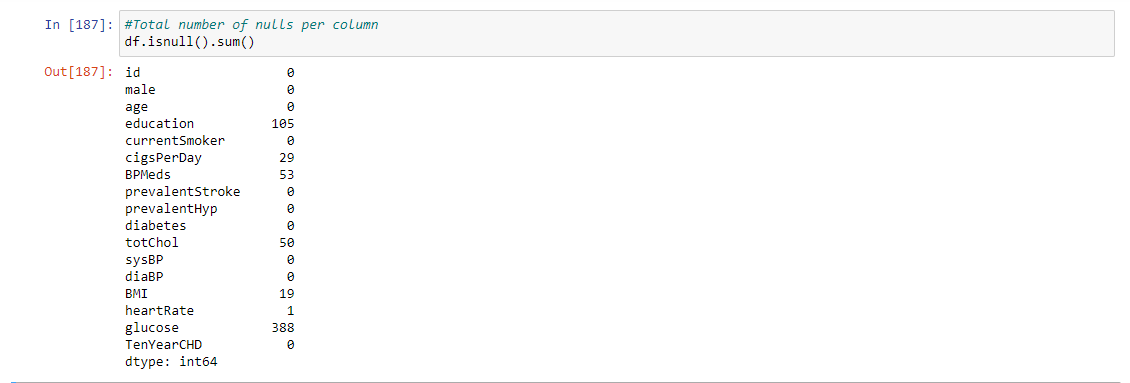
* Gathering Data

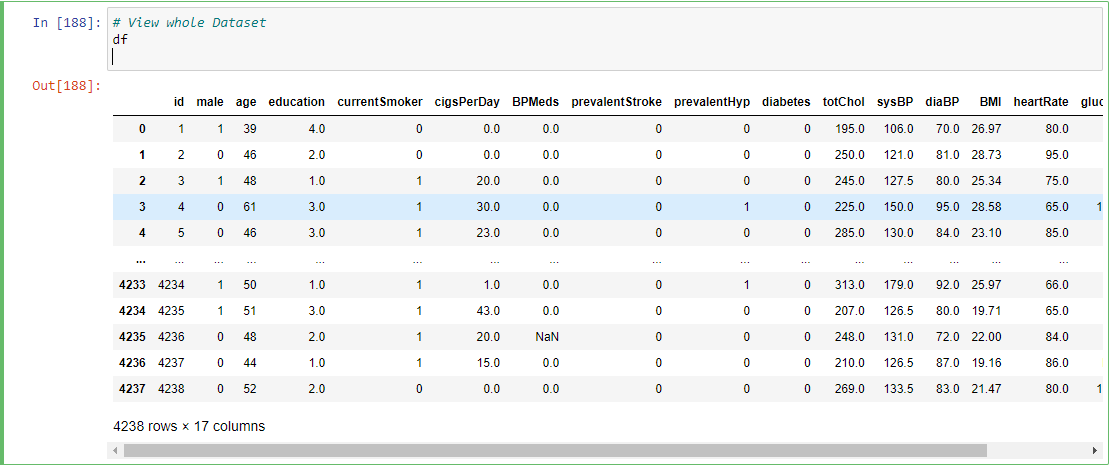




* Exploring data

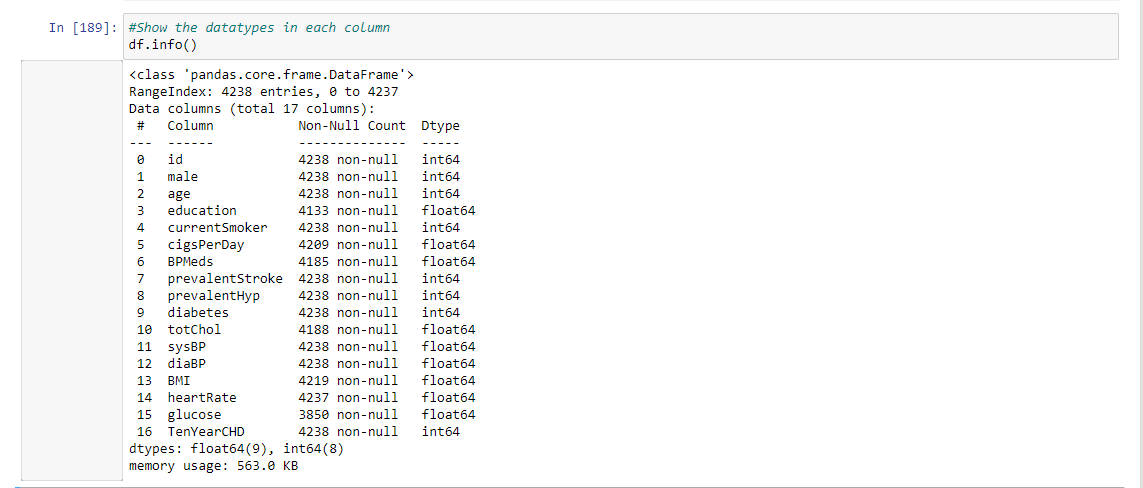
There are 17 columns with 7 of them having missing data

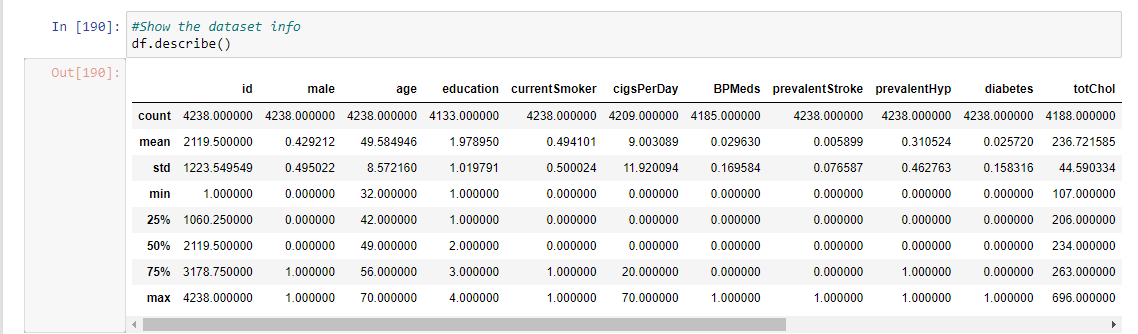


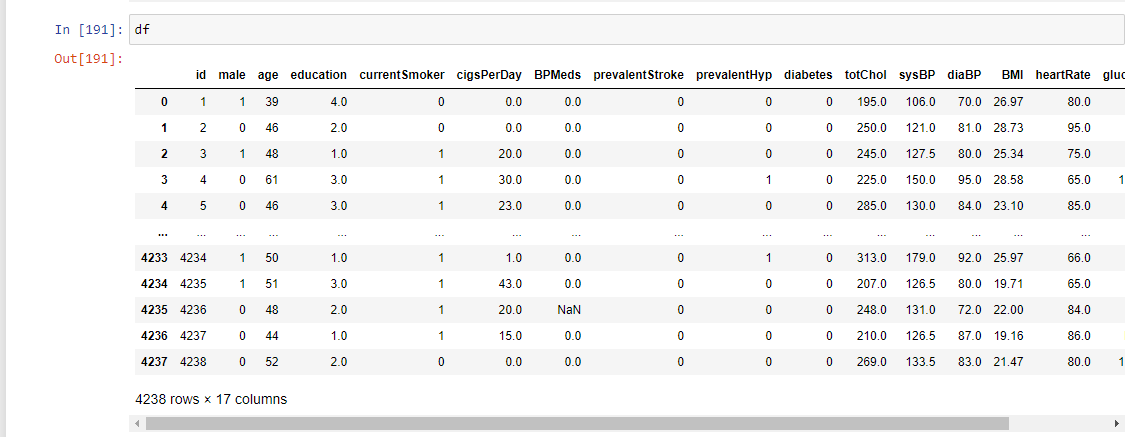


* Describing Data

With describing the data and finding more info, it is found that there are no object types in the dataset, only floats and integers.







**Data Cleaning**

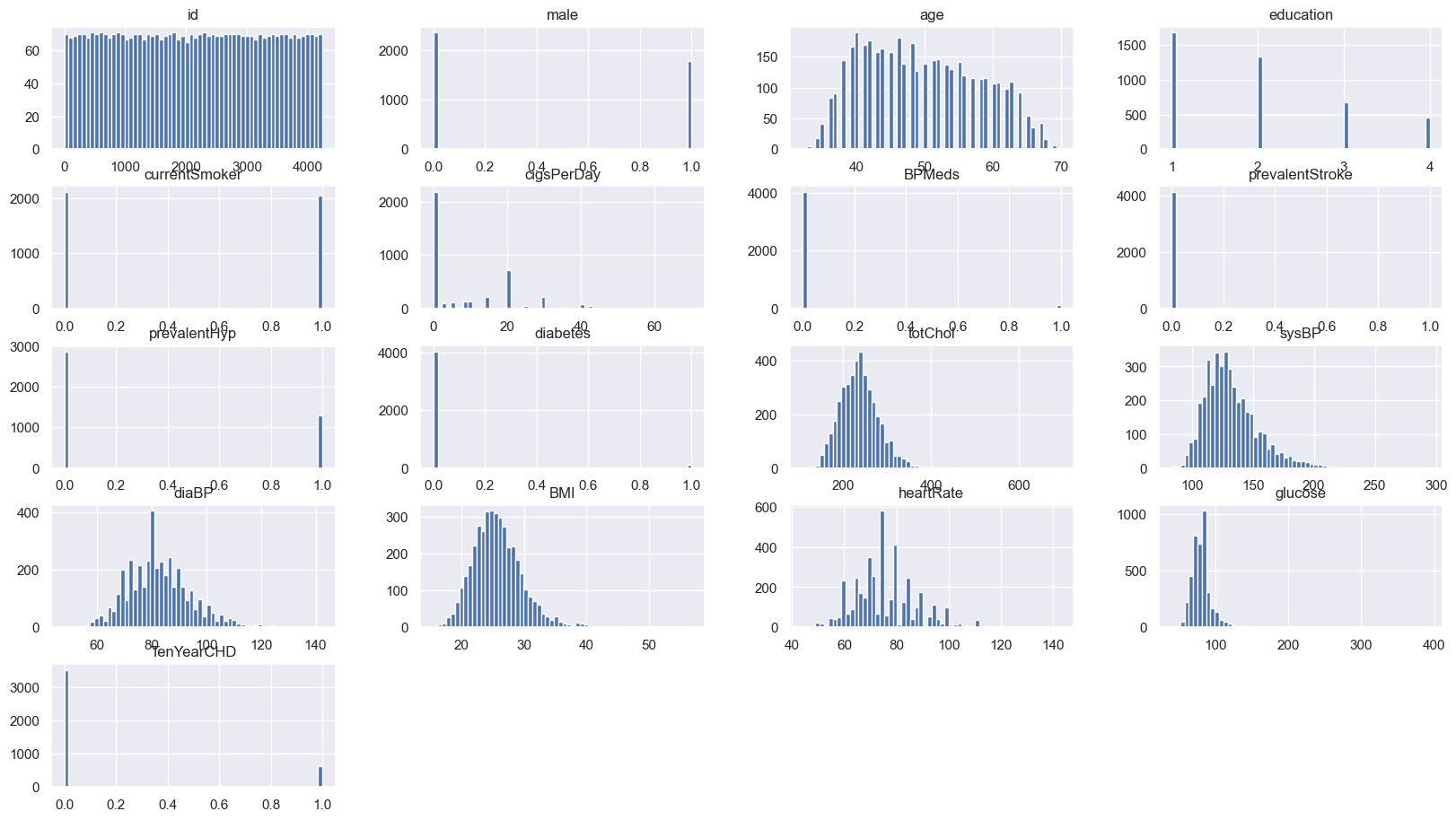
Handling Missing Values

* Nulls in the column education were filled with the median value of education
* Nulls in the cigsPerDay column were filled with the mode value of cigsPerDay column
* Nulls in the BPMeds column were filled with the mode value of BPMeds column
* Nulls in totChol column were filled with the mean value of totChols column
* Nulls in BMI column were filled with the mean value of BMI column
* Nulls in Glucose were filled with the mean value of the Glucose column
* Drop all Nulls which were not filled with the median, mean and mode

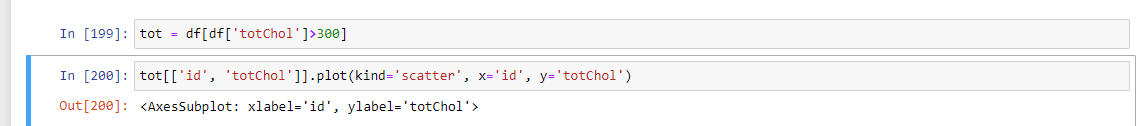


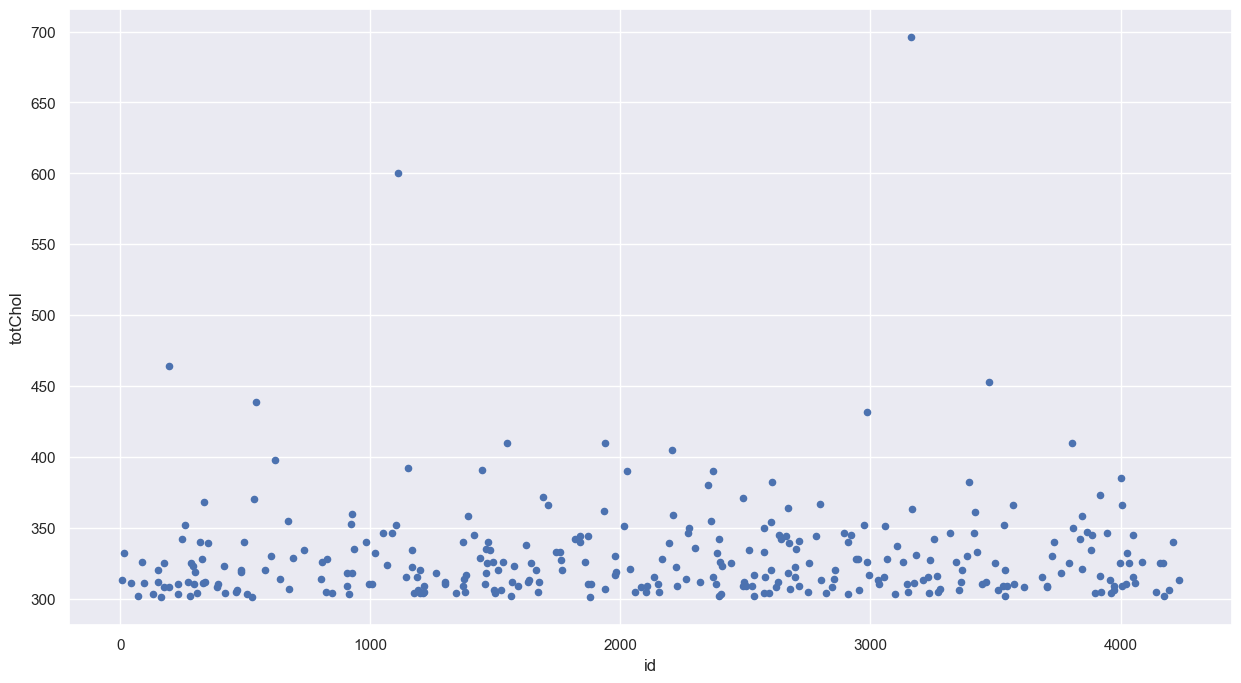
**Plotting**





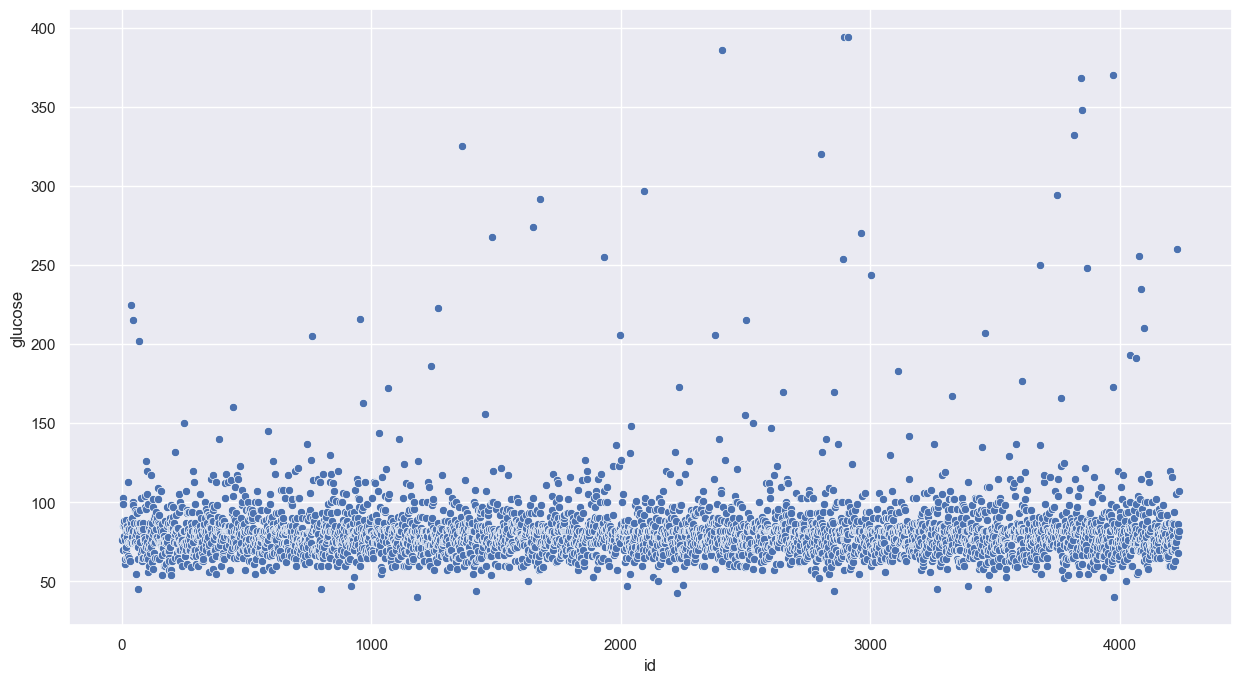
* Scatterplot for totChol



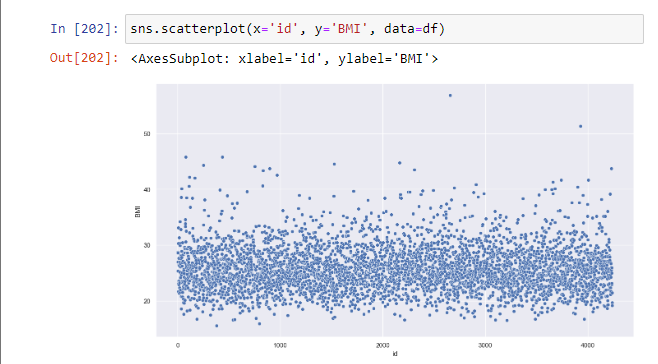


* Scatterplot for glucose

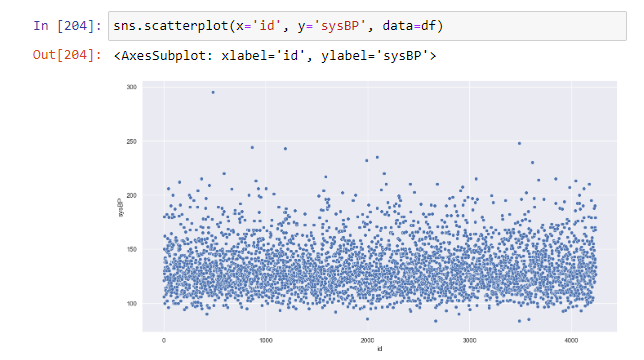




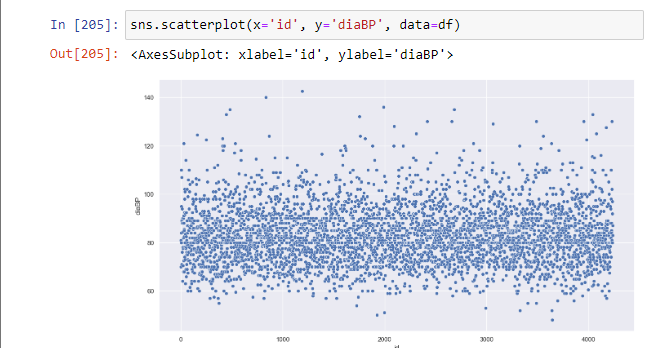
* Scatterplot for BMI



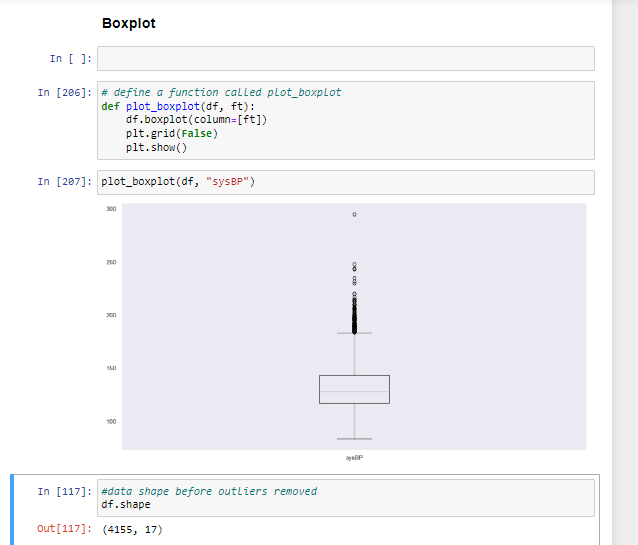
* Scatterplot for sysBP

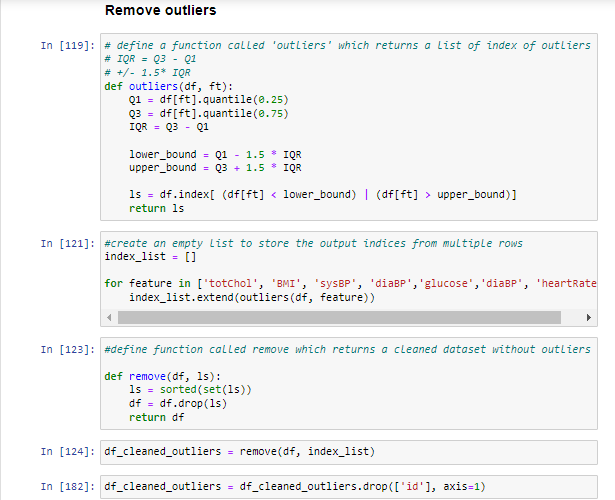


* Scatterplot for diaBP

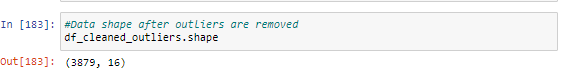


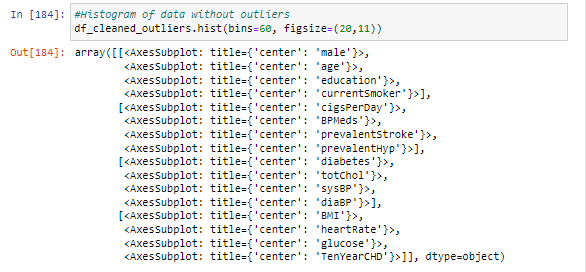
**Dealing with Outliers**

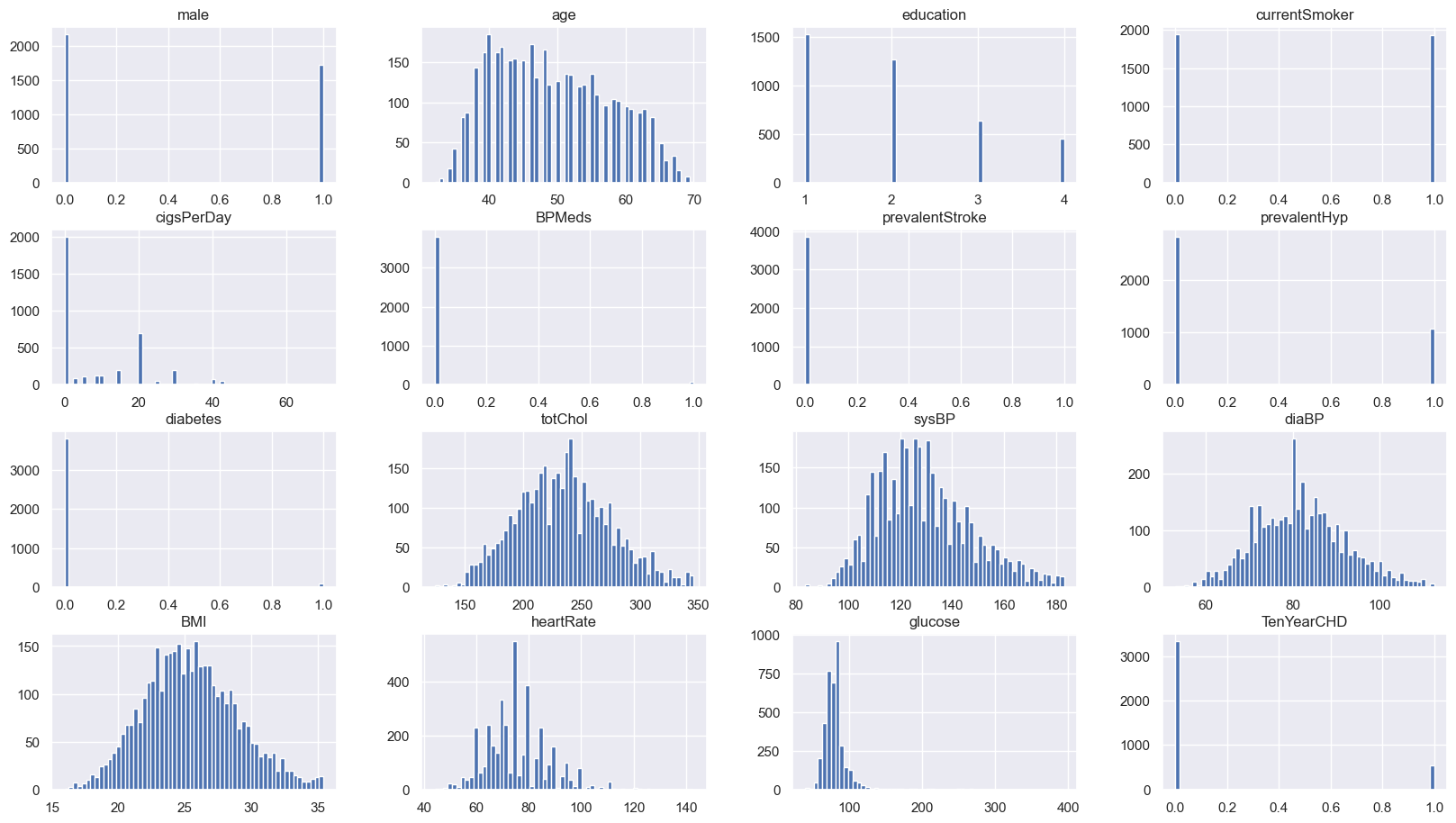


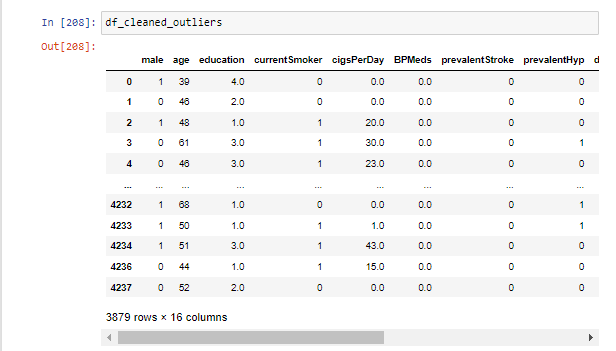


Plot the Histogram of the cleaned dataset with removed outliers

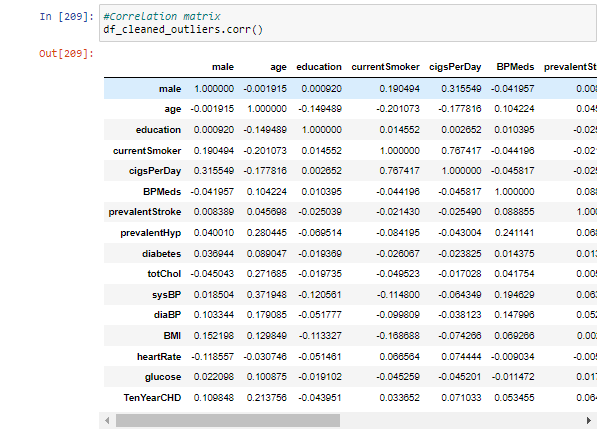








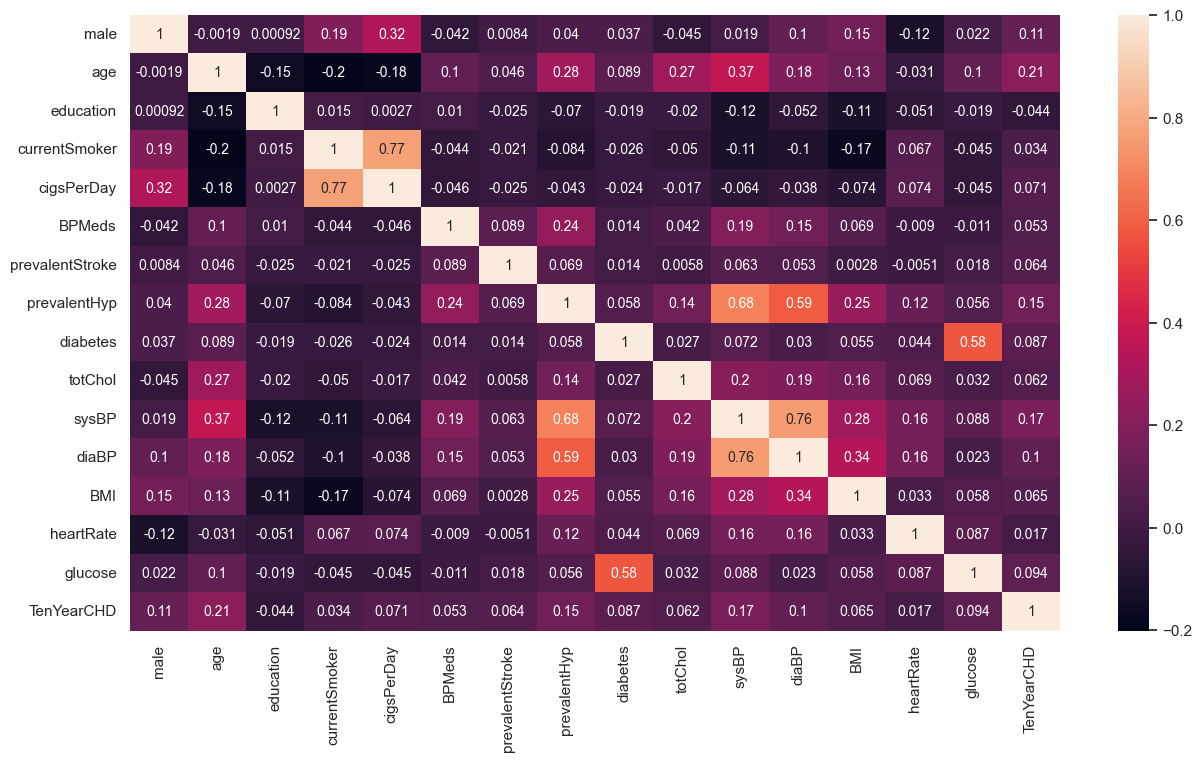
**Correlation Matrix**



**Correlation matrix heatmap**

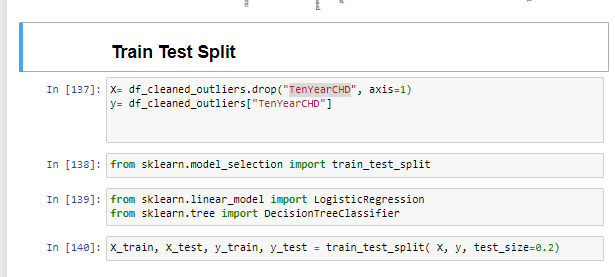
The matrix depicts the correlation between all the possible pairs of the values in a table.

The lighter the color the more it correlates and the darker the color the less it correlates.



**Data Modelling**

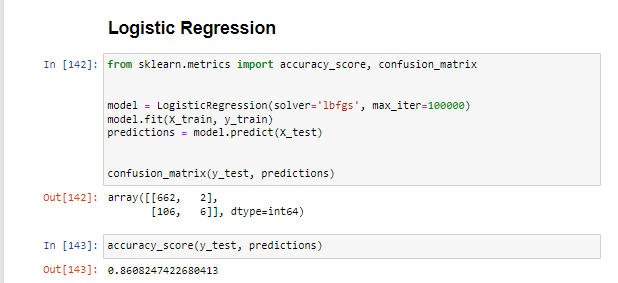
* Save the column Target column TenYearCHD in the variable y
* Save the Columns besides TenYearCHD in the variable X
* Split the data, 20% for testing and 80% for training.



**Logistic Regression**

With the logistic regression algorithm, 86% accuracy was scored.

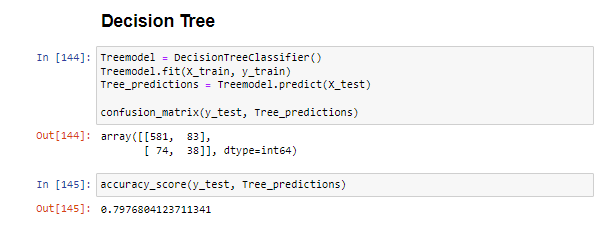
* With the confusion matrix displaying, 662 True positives; 106 False Negatives; 2 False Positives and 6 True Negatives.



**Decision Tree**

With the Decision Tree algorithm, 79% accuracy was scored.

* With the confusion matrix displaying, 581 True positives; 74 False Negatives; 83 False Positives and 38 True Negatives.



**Random Forest**

With the Random Forest algorithm, 85% accuracy was scored.

* With the confusion matrix displaying, 661 True positives; 106 False Negatives; 3 False Positives and 6 True Negatives.

