# 54r4,1 A description of the dataset

## 1.1 Problem of interest

The dataset I have chosen to work with is called “South African Heart Disease”, containing data of a sample of males in a heart-disease high-risk region in South Africa. There are roughly two controls per case of Congenital heart defect (CHD). Most of these CHD-positive patients have undergone different programs and treatments to reduce their risk factors, and some of these measurements were made after those treatments. The attributes refer to systolic blood pressure, cumulative tobacco (kg), low density lipoprotein cholesterol, adiposity, family history heart disease, type A behaviors, obesity, current alcohol consumption, age and coronary heart disease (Trevor Hastie, u.d.).

With this dataset it is possible to apply various statistical tools in order to describe and analyze it. Furthermore, various machine learning models can be used to make predictions on certain attributes.

## 1.2 Data source

The data was gathered from the Stanford.edu website called The Elements of Statistical Learning, and the data is taken from a larger dataset described in (J E Rossouw, 1983).

## 1.3 Previous work on the data

As mentioned earlier, the data was taken from a larger dataset described in the South African Medical Journal. This article describes the coronary risk factor screening in three rural communities, and revealed that the significant reversible danger factors hypercholesterolaemia, hypertension and smoking, just as 'minor' factors, for example, inertia, corpulence, hyperuricemia, coronary-inclined conduct and the irreversible danger elements of chest torment, ischaemic changes on the ECG and a family background of ischaemic coronary illness (IHD), were incredibly common. Separately or combined, the significant danger factors were available in the dominant part of the examination populace after the age of 44 years. The association of elevated levels of way of life prompted hazard factors with constitutional inclination could sufficiently clarify any abundance danger of IHD in the African speaking community. The practically general danger predominating this examination has significant implications for any preventive technique (J E Rossouw, 1983). The data is analysed using different methods, such as histograms to depict the data.

# 1.4 Machine learning modeling aim

The large datasets used for this study is drawn from Rossouw et al. considering the survey sample of males collected from risk prone region i.e. Western Cape of South Africa to detect the heart disease of the individuals indicating negative (0) or positive (1) . The positive result holders were asked to undergo remedial steps including blood pressure reduction to bring down the level of risk during post treatment.

Each high-risk patient has been monitored and the following patient attributes were obtained: systolic blood pressure (sbp), cumulative tobacco (tobacco), low density lipoprotein cholesterol (ldl), adiposity, family history of heart disease (famhist), type-A behavior (typea), obesity, alcohol, and age. A total of 462 samples are included in this data set. Adiposity is a measure of % body fat, whereas obesity measures weight-to-height ratios (body-mass-index, bmi). Type-A behaviour pattern is characterised by an excessive competitive drive, impatience and anger/hostility.

Therefore, it allows for various types of machine learning analyses and is ideal to address different research questions. For each machine learning task, we identified an interesting analysis that we wish to perform, we specified what we aim to accomplish and we selected the relevant attributes for it.

In Supervised learning by using known dataset termed as labelled training dataset which includes both input data and response values and thereby generating a model to make predictions of the response values for a target data. As a result, validation of the above generated model is done through testing the dataset. Various applications of the supervised machine learning algorithms include biometric attendance or ATM, spam filters, weather prediction, predicting winning % between two teams, Face detection, Text and speech categorization, Signature recognition and Medicine. Supervised learning includes two categories of algorithm:

1. **Classification**: for categorical response values, where the data can be separated into specific ”classes”.

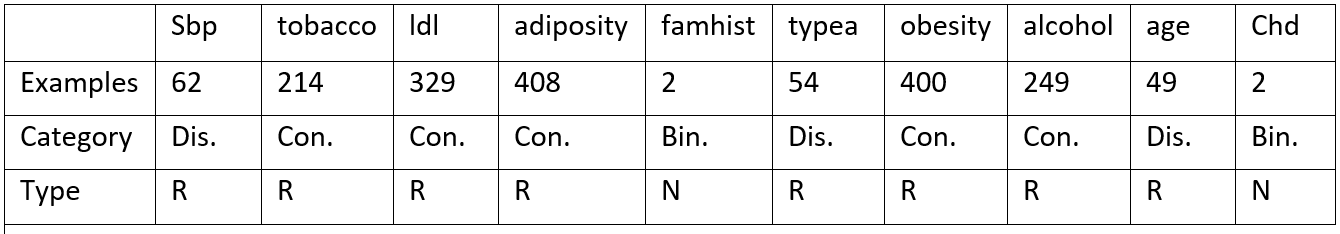
2. **Regression**: for continuous-response values.

With regard to **classification**, it would be interesting to predict whether a patient tests positive for coronary heart disease or not based on the measurements. This analysis aims to predict the ‘CHD’ attribute which will be our label. The (1) in the ‘CHD’ attribute denotes that the patient is tested positive for ‘Coronary Heart Disease’ and vice versa 0 represents that the patient is test negative.

In the case of **regression,**  it would be interesting if we try to predict low density lipoprotein cholesterol (ldl) from the other attributes. ldl is the ”bad” cholesterol, and we expect high ldl levels to be strongly associated with adiposity and obesity. However, we aim to build a predictive model for ldl using the most informative subset of variables.

# 2 A detailed explanation of the data attributes

## 2.1 The attributes of the data set

****

Attributes by category and type.

* Con. = Continuous, Dis. = Discrete, Bin. = Binary.
* R = Ratio, I = Interval, N = Nominal

## 2.2 Data issues and description

Fortunately, this dataset contains no missing values. However it contains a column “row.names” which is quite trivial, when working on that data, therefore we removed it while working on our data.

## 2.3 Basic statistics of attributes

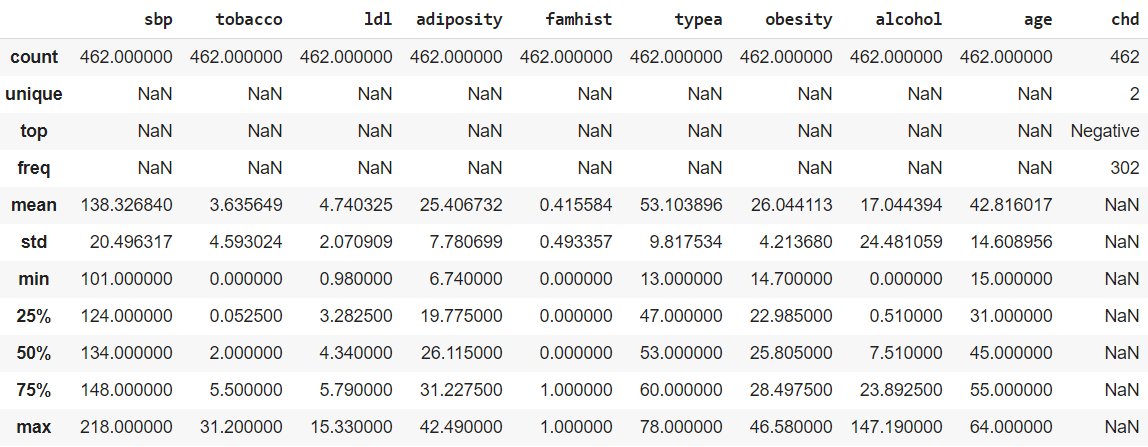


Table 2: Summary statistics

In table 2 we have expressed the general statistics describing the data.On interval and nominal data we have not described mean, variance, standard deviation, quartiles and the median, since these would be redundant information of no importance because it does not make sense to impose the average on an interval, nor which is best between Europe, Japan and America. We have converted the ‘famhist’ column 1s and 0s as 1 represents ‘Present’ and 0 represents ‘Absent’ respectively.

# 

# 3 Data visualization

## 3.1 Outliers and issues:

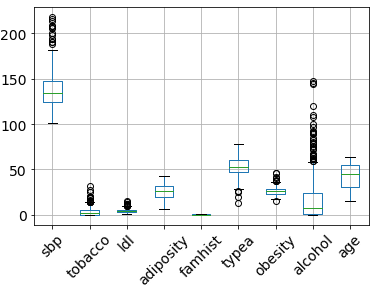


Figure 1: Boxplot

Using the method of identifying outliers through the quartiles we found a few samples that were outside of the expected area.

Outliers.above = Q75 + 1.5 ∗ (Q75 − Q25) (1)

Outliers.below = Q25 − 1.5 ∗ (Q75 − Q25) (2)

The following attributes contained outliers which are given below.

**sbp** = [206,194,208,194,208,188,190,206,190,198,218,208,200,216,214]

**tobacco** = [14.1,14,18.2,31.2,14,15.5, 27.4 , 15, 20,18,15, 14.4,19.45,19.6,25.01,19.2,20.,16.,13.8]

**ldl** = [15.33, 12.42, 9.65, 11.32, 10.58, 11.17, 10.53, 11.89, 11.41,10.49, 10.19, 9.78, 14.16, 11.61]

**typea** = [13,26,20,25]

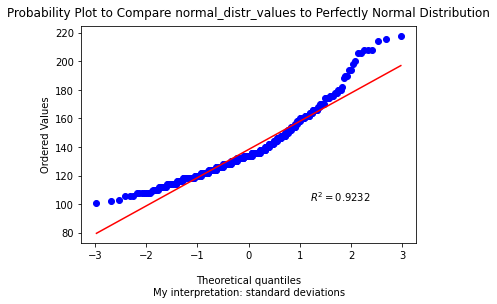
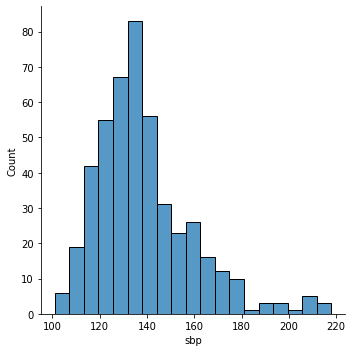
**ob** = [38.8 , 46.58, 45.72, 37.41, 40.34, 41.76, 37.24, 37.71, 14.7 ]

Even though these represent data points that are out of the expected area, we do not find any ”exceptional” good reason for removing these, since they all are within a reasonable spectrum, and do not appear to be invalid in any manner that would justify them to be removed.

## 3.2 The distribution of the attributes

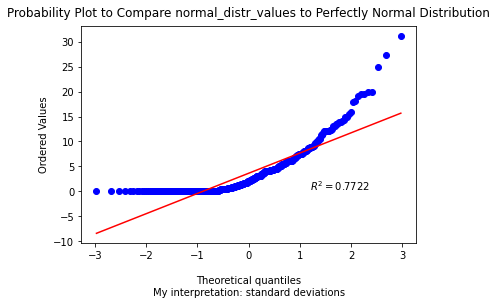
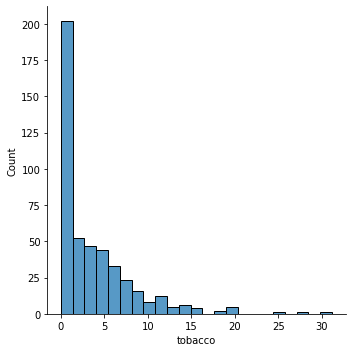
In this section, an analysis of the distributions of the attributes will be made on the data set.

#### Histogram and Normal Probability Plot for the attribute SBP



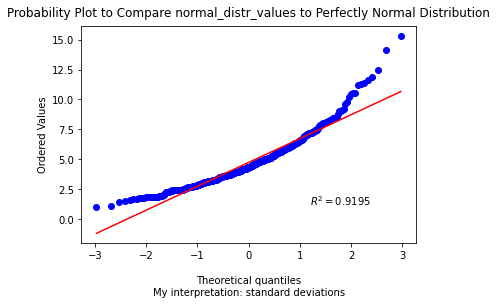
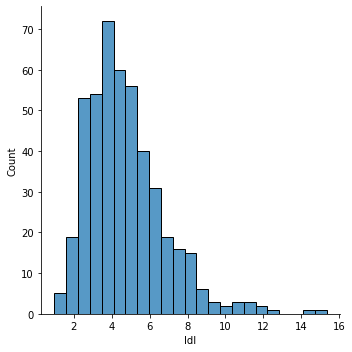
In this case, we see that the data for the attribute sbp resembles a normal distribution a bit, but it skews more to the right . This indicates that the mean is right to the median.

#### Histogram and Normal Probability Plot for the attribute tobacco



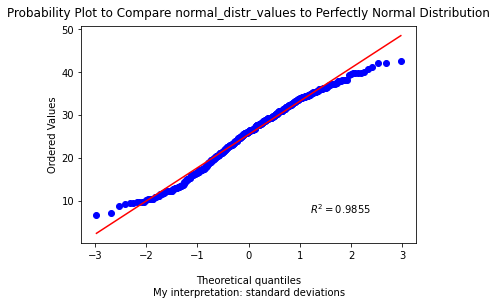
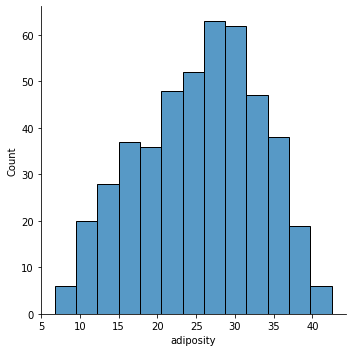
When considering tobacco attributes, the data is skewed more towards right according to the normal probability plot and histogram.

#### Histogram and Normal Probability Plot for the attribute ldl



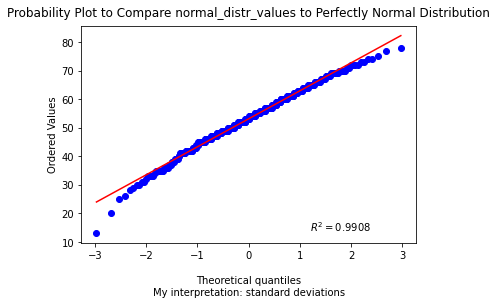
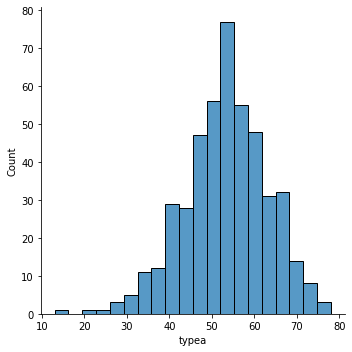
As for ldl, by looking at the histogram and the normal probability plot, we can see that it is skewed a little towards right.

#### Histogram and Normal Probability Plot for the attribute adiposity



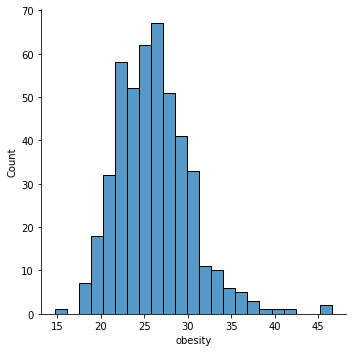
For Adiposity the graph is skewed a little towards the left as it is visible from histogram and probability plot.

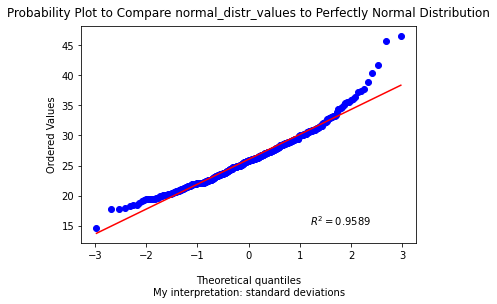
#### Histogram and Normal Probability Plot for the attribute typea



For Type-A attributes the data is skewed towards the left a little bit but it is in the acceptable limit for normal distribution.

#### Histogram and Normal Probability Plot for the attribute obesity

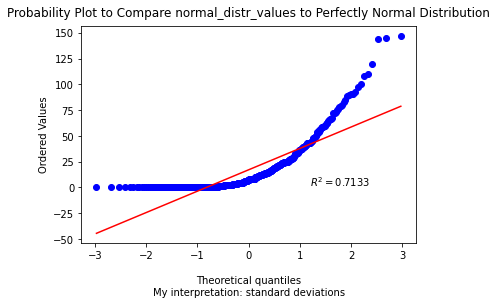
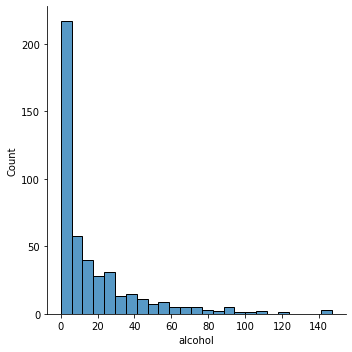




For obesity the data is skewed towards the right.

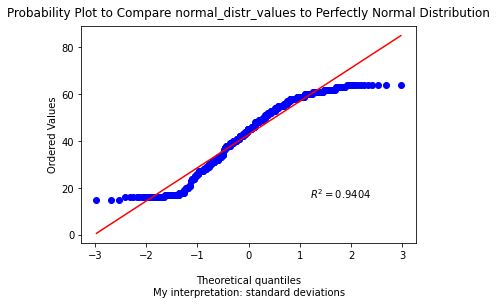
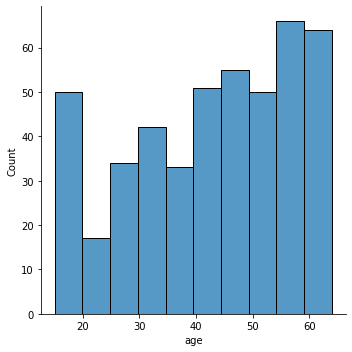
### 

#### Histogram and Normal Probability Plot for the attribute alcohol



For alcohol the data is skewed towards the right.

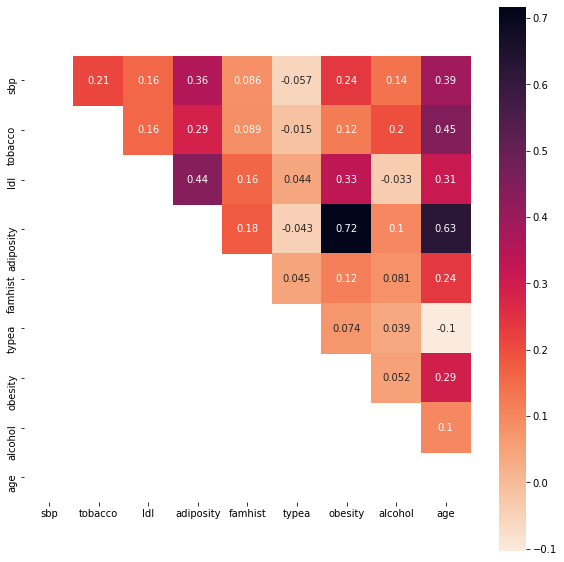
#### Histogram and Normal Probability Plot for the attribute age



For Age the data is skewed towards the left.

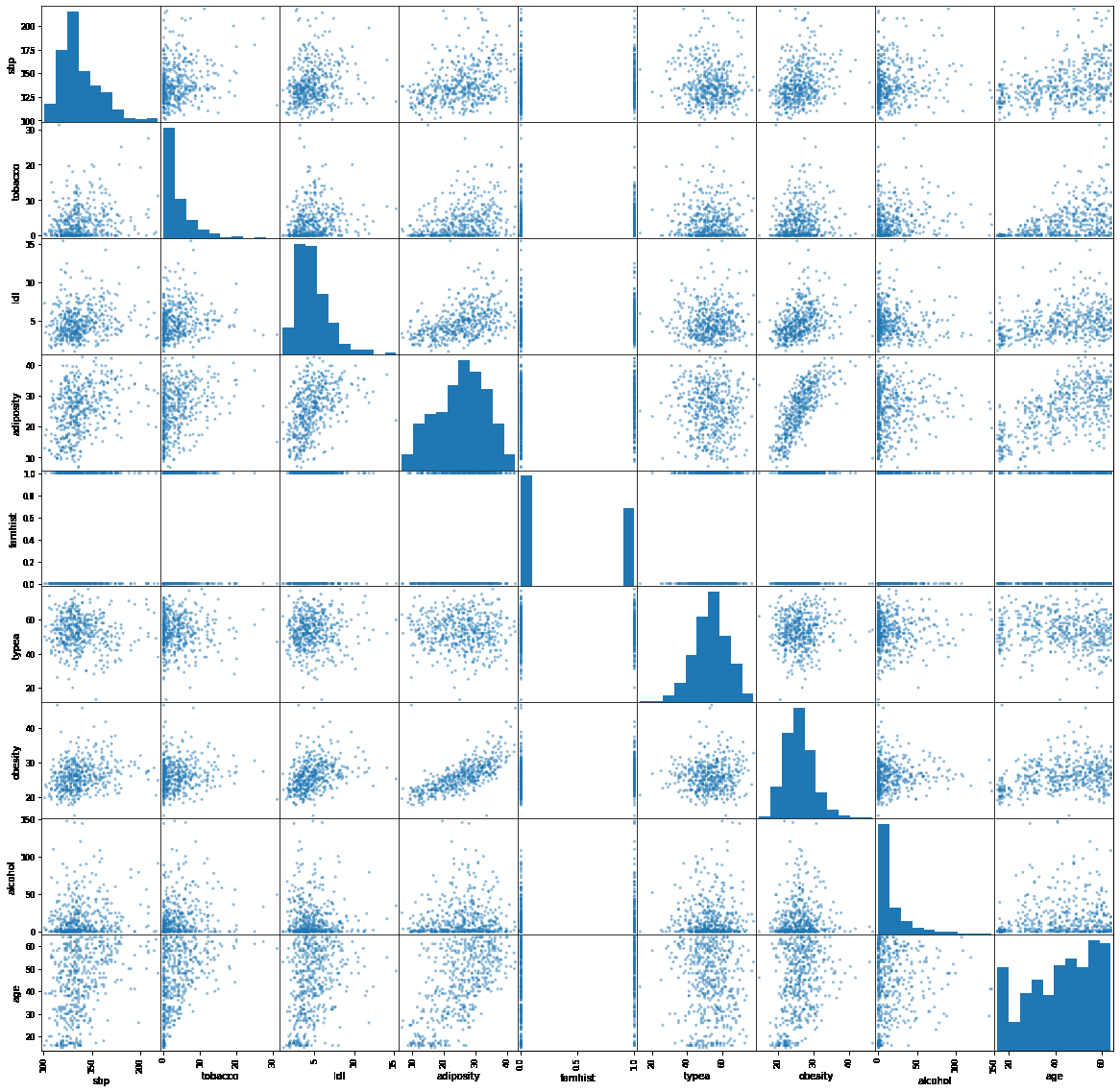
## 3.3 Correlation between the variables

#### HeatMap

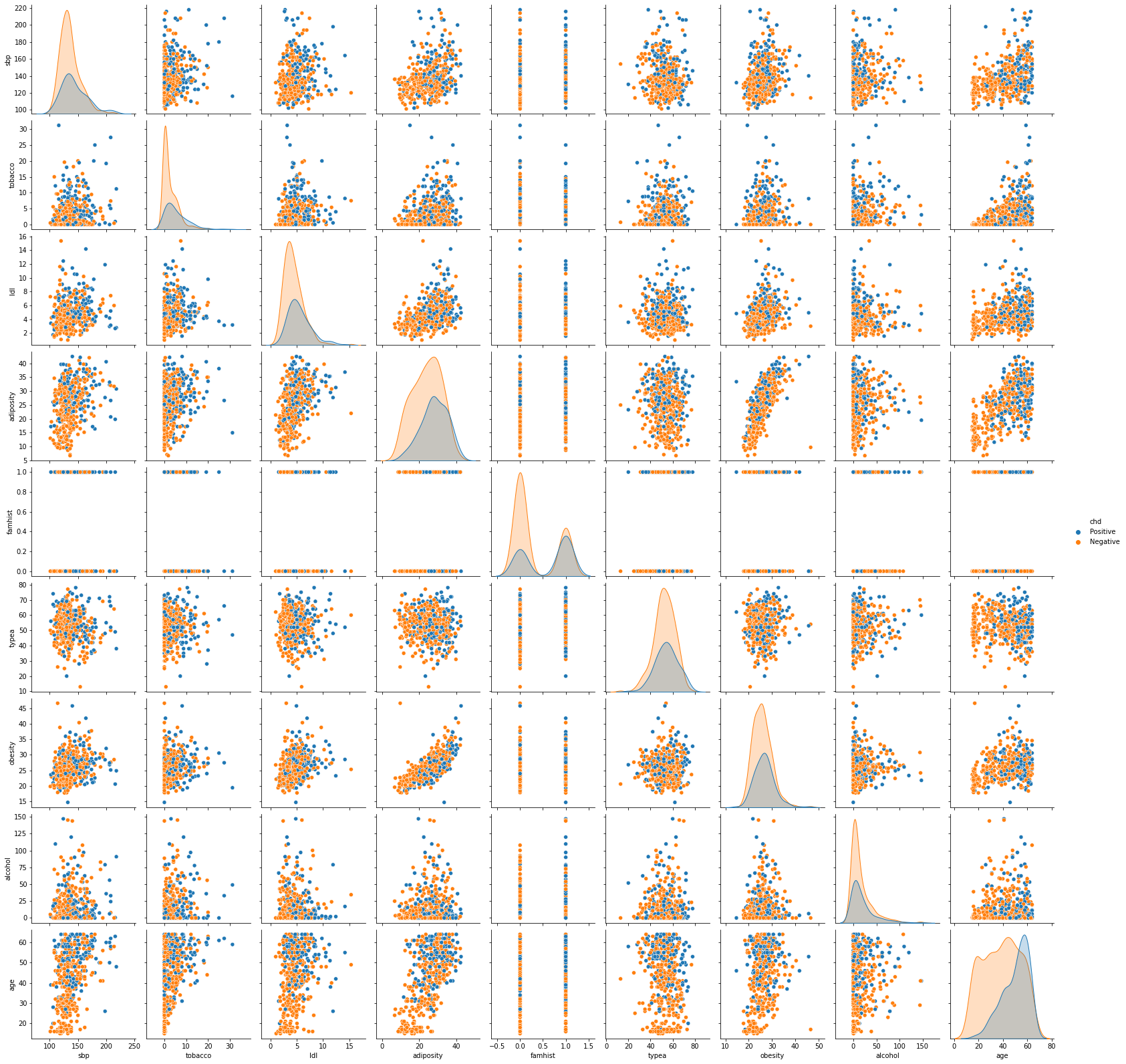


The attributes have both positive and negative correlations. It is interesting to notice that typea is negatively correlated with the amount of age, adiposity, sbp and tobacco. On the other hand, adiposity, obesity, and age are strongly correlated with each other. This sounds reasonable as adiposity and sbp tends to be higher in obese people.

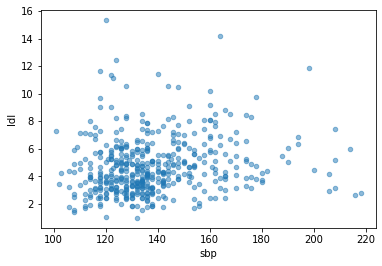
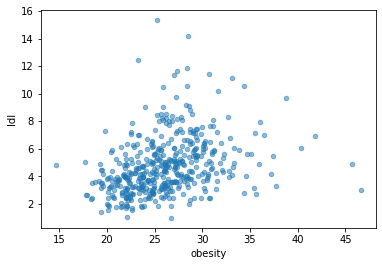
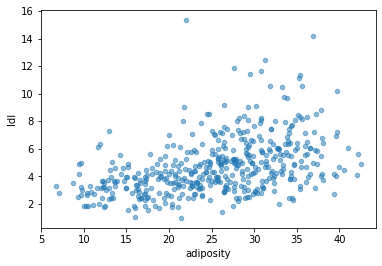
#### Scatter Matrix



#### PairPlot



#### Plotted Correlation of Specific Attributes

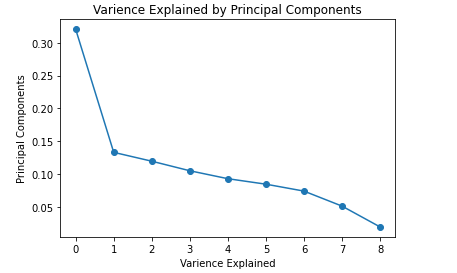


## 3.4 Primary machine learning applicability based on visualizations

Based on the initial visualizations there are good indications that the data set is capable of supporting the intended aim with the respective methods proposed. The data is very consistent and in several cases normally distributed and the continuous attributes hold regressional possibilities. There appear to be various correlations between the continuous attributes, that can be used as indicators. Comments on the PCA are given in the following sections.

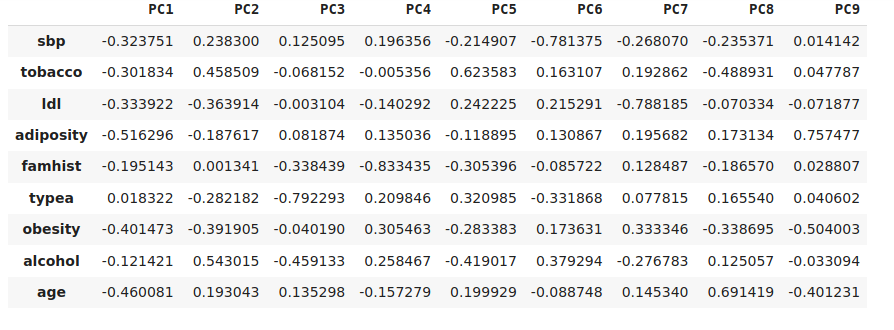
# 

## 3.5 The amount of variation explained by the PCA components



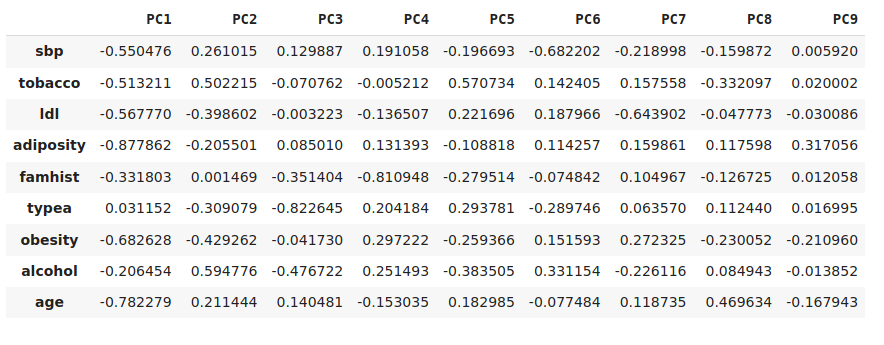
Above figure displays the proportion of variance explained by the 9 PCs. The variables have been standardized before performing the PCA. It is possible to notice that the first PC recovers around 32.05% of the variance of the original variables. Moreover, the first PC has an eigenvalue of 2.89, significantly higher than 1. This shows the great relevance of the component in summarizing the data, and, at a certain extent, this was expected, since many of the variables are correlated, therefore the component is able to explain the variance of these variables at the same time. The second component, instead, recovers around 13.30% of the variance and has an eigenvalue of 1.199. Although the percentage and the eigenvalue are not very high, we believe it could be interesting to analyse the second component, since it might recover the variance of a specific variable. We are not considering the remaining components relevant for the analysis. In conclusion, we will analyse the first 2 principal components, which recover a cumulative proportion of 45.35% of the variance.

## 3.6 Principal component analysis: directions and PC interpretation



### Eigenvectors of the PCA components

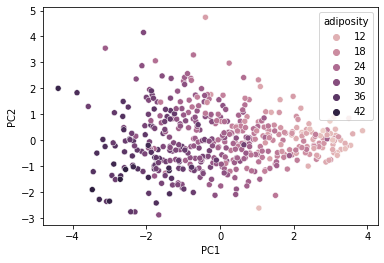
The above figure displays the eigenvectors, which define the weight of each variable in the definition of each principal component. **Adiposity,** **obesity** and **age** are important in the definition of the first principal component. On the other hand, **adiposity** and **age** are less important in defining the second component. However, the eigenvectors do not provide information about the relevance of the tendencies described by the components. Therefore, to better describe the relations among the variables and the information summarized by the PCs, we use the loadings, i.e. the correlations between the original variables and the PCs, which are based both on eigenvectors and on eigenvalues.

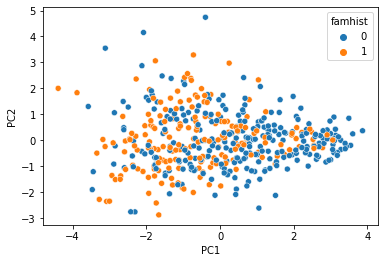
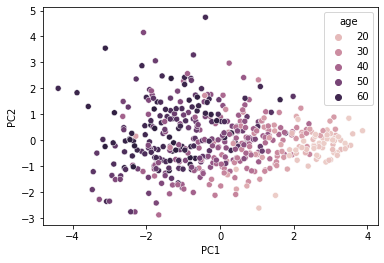


Loadings of the PCA components

The loadings of components clearly show that the first component has an extremely high proportion of variance explained, because it is highly correlated with all the variables except **typea** and **alcohol**, thus, very important in describing them. It opposes people with lower **obesity**, **age**, **sbd**, **tobacco**, **adiposity**, **ldl** and absent **famhist** to people with **high obesity**, **age**, **sbd**, **tobacco**, **adiposity**, **ldl** and present **famhist**. Inshort, the first component opposes people who do not have coronary heart disease.

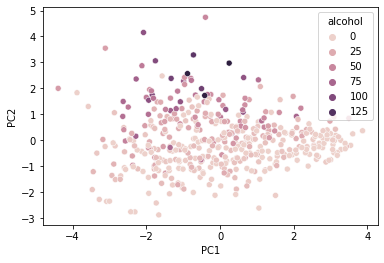
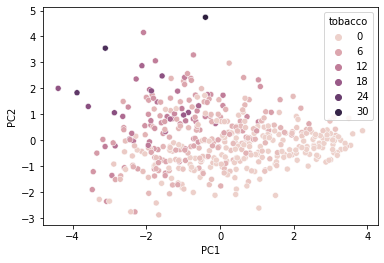
We can verify it by seeing these graphs.



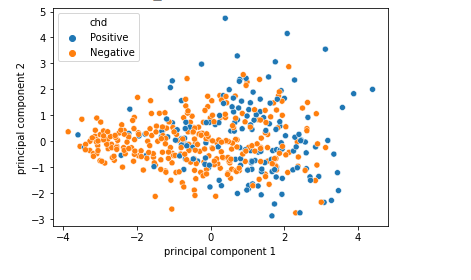


The second PC, instead of alcohol, which is less related to the other variables. The second PC mainly opposes people with high alcohol and tobacco intake. Inshort, the second component opposes people who have coronary heart disease.

We can verify it by these graphs.

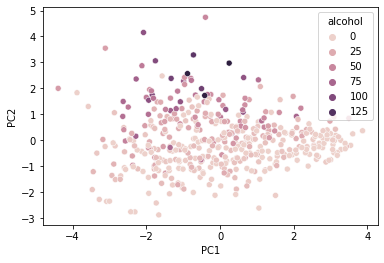


In these graphs we can see that people with higher intake of **Tobacco** and **Alcohol** have higher **PC2** than others.



### Directions of the PCA components

Looking at the figure, it is possible to evaluate if chd positive and negative cases are mapped on the graph with PC1 and PC2 as x-axis and y-axis respectively. It seems that points with negative PC1 are showing people with coronary heart disease. It implies that the people with higher adiposity, obesity, age and people having coronary heart disease in their family members are likely to have coronary heart disease than other people.



As for the tobacco and alcohol intake. People with greater intake of alcohol and tobacco generally have higher PC2 and have a higher chance of having coronary heart disease.

# 4 Discussion

## 4.1 Summary

In this study, analysis of heart disease is applied using a South African heart disease dataset. The analysis of heart disease is done using nine features, and the role of analysis is to gain insight about the relationship of the disease with these features. The analysis could improve the treatment quality and treatment process.

All in all, we can conclude the following points about what we have learned from the data:

* There seems to be a rather strong correlation between some of the attributes. These are: the adiposity and ldl, tobacco and age, adiposity and obesity, and adiposity and age. This makes sense as a person is obese, he is more likely to have higher **adiposity** and **ldl**, and the probability of him being aged is greater. As for the negative correlations, there seems to be weak ones between **typea** and the other continuous attributes which are not enough to argue for causality.
* As for the PCA, we decided to analyze the first two PCs as they cover a cumulative proportion of 45.35% of the variance. Since the first PC is highly negatively correlated with the **adiposity**, **obesity** and **age** it is used to describe those attributes. The first component opposes people who do not have coronary heart disease. With regard to the second PC, it is described by its major influence by **alcohol** and **tobacco**. This means that the second PC also opposes cardiac heart disease patients but the second PC has higher acceptance for cardiac heart patients than the first PC.
* When it comes to the PCA directions, it shows those people who are obese, with family history of heart disease and higher age are more prone to coronary heart diseases.
* Similarly, people with higher intake of tobacco and alcohol are more prone to coronary heart disease.

## 4.2 Feasibility of the primary modeling tasks

Thanks to the detailed analysis of the attributes and to the PCA performed, we have gained awareness of the feasibility of the machine learning methods initially suggested. The strong relations between the variables and the normal distribution of the attributes seem to make the application of machine learning techniques possible. In particular, the prediction of the type of heart disease (classification method) seems to be feasible, since there appears to be a relation between the variable origin and the other ones. As for regression, the analysis of low density lipoprotein cholesterol is affected by other variables proves to be an interesting task. In that regard, it might be necessary to control confounding variables.