Universidad Politécnica de Yucatán

Data Engineering

H3

Exploratory Data Analysis

10/16/19

Brenda Martín Várguez.

The Heart Disease Data

**Description of the data base:**

To start describing this data base, it is needed first, to describe what a heart disease is. Heart disease describes a range of diseases that affect the heart. Diseases classified as heart disease are diseases of the blood vessels, such as coronary artery diseases, heart rhythm problems (arrhythmias) and heart defects with which it was born (congenital heart defects), among others.

For this project, the Heart Disease Data Set was used, which is available on the UC Irvine Machine Learning Repository. This database contains another four databases concerning heart disease diagnosis from the Cleveland Clinic Foundation, the Hungarian Institute of Cardiology (Budapest), the V.A. Medical Center (Long Beach, CA) and the University Hospital (Zurich, Switzerland), although for this project only the information provided by the Cleveland Clinic Foundation was analyzed. It is important to mention that, according to the UC Irvine Machine Learning Repository, all the information provided by all the clinical centers dates from 1988. The authors of this data set are: from the Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D., from the University Hospital, Zurich, Switzerland: William Steinbrunn, M.D., from the University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D. and from the V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Each data base has the same instance format. According with the description contained on the UC Irvine Machine Learning Repository, “This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date.” And for this project in particular, as it has been mentioned before, also only the Cleveland database provided by the Cleveland Clinic Foundation has been used, and from the 76 attributes, only a subset of 14 of them were selected to work with.

**Attribute information:**

Only 14 used:

1. Age
2. sex
3. sp
4. trestbps
5. chol
6. fbs
7. restecg
8. thalach
9. exang
10. oldpeak
11. slope
12. ca
13. thal
14. num

**Complete attribute documentation:**

1. **age**: age of patients in years
2. **sex**: 1 = male; 0 = female
3. **cp**: chest pain type
   1. Value 1: typical angina
   2. Value 2: atypical angina
   3. Value 3: non-anginal pain
   4. Value 4: asymptomatic
4. **trestbps**: resting blood preasure (in mm Hg on admission to the hospital)
5. **chol**: serum cholestoral in mg/dl
6. **fbs**: fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7. **restecg**: resting electrocardiographic results.
   1. Value 0: normal
   2. Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
   3. Value 2: Showing probable or definite left ventricular hypertrophy by Estes’ criteria.
8. **thalach**: maximum heart rate achieved
9. **exang**: exercise induced angine (1 = yes; 0 = no)
10. **oldpeak**: ST depression induced by exercise relative to rest
11. **slope**: the slope of the peak exercise ST segment
    1. Value 1: upsloping
    2. Value 2: flat
    3. Value 3: downsloping
12. **ca**: number of major vessels (0-3) colored by flourosopy
13. **thal**: **thalassemia**- 3 = normal; 6 = fixed defect; 7 = reversable defect
14. **target:** 0(no presence) 1(presence)

**Important concepts**

Before start analyzing the dataframe, it is important to understand some basic concepts that were used on this EDA.

According to the Engineering Statistics Handbook, exploratory data analysis (EDA) is a very important step to analyze data due to here, the researches try to make some sense of it, and it’s often the first step in data analysis. Using EDA can be really helpful for error detection like missing data and other mistakes, also to check assumptions associated with any hypothesis as well as performing a preliminary selection of appropriate models, among others.

Now, let’s talk about methods of analyzing data in exploratory data analysis. First, we have the graphical and non-graphical method. And on second place, a method can be univariable or multivariable.

According to information provided by the Engineering Statistics Handbook: “Non-graphical methods generally involve calculation of summary statistics, while graphical methods obviously summarize the data in a diagrammatic or pictorial way. Univariate methods look at one variable (data column) at a time, while multivariate methods look at two or more variables at a time to explore relationships […]”

**Data type**

**Categorical Data**

Categorical variables contain a finite number of different categories or groups. Categorical data may not have a logical order; for example, sex of a person. The characteristics of interest for a categorical variable are simply the range of values and the frequency of occurrence for each value (Engineering Statistics Handbook), while quantitative variables are numbers that usually represent a control or a measurement.

**Discrete variable**

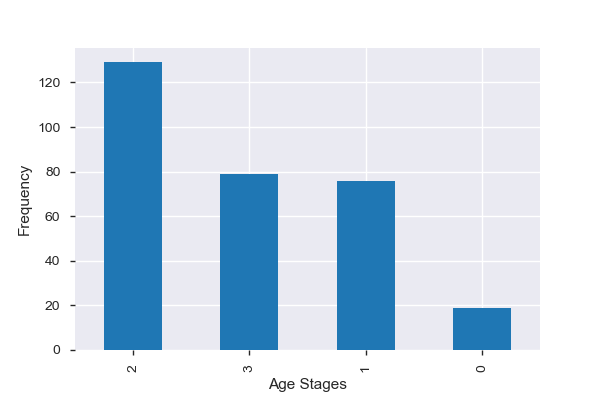
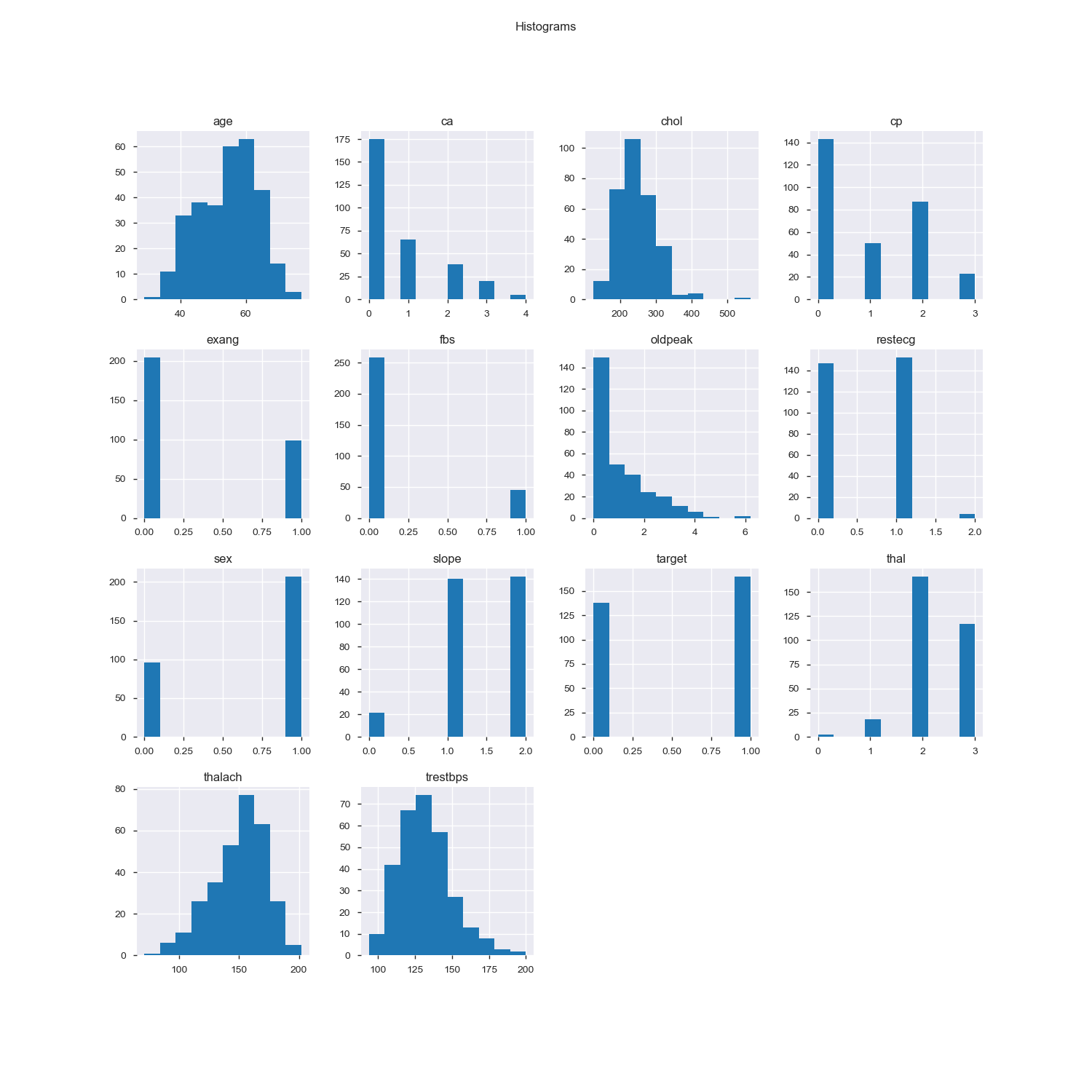
Discrete variables are numerical variables that have a countable number of values between any two values. A discrete variable is always numerical. For example, the number of failures or defects.

**Continuous variable**

Continuous variables are numerical variables that have an infinite number of values between any two values. A continuous variable can be numeric or date / time. For example, the age of a person.

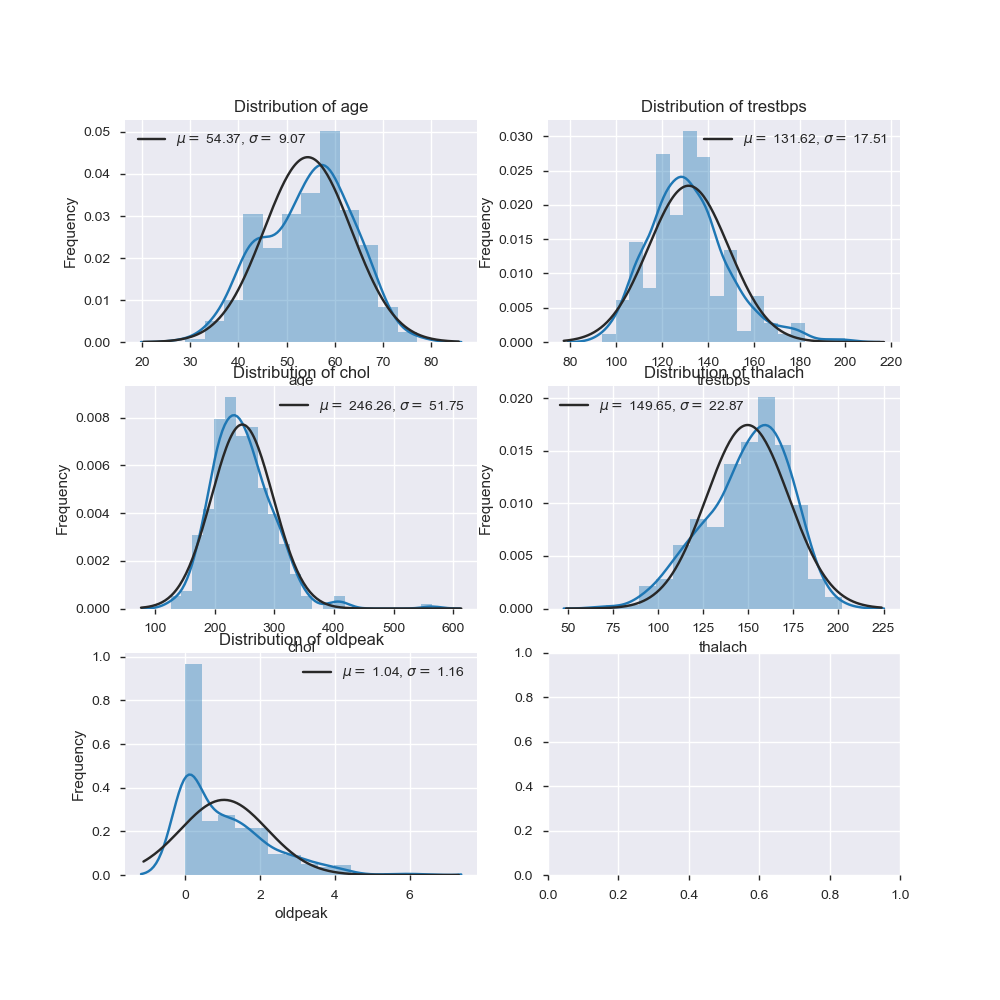
**Data analysis for some variables**

A histogram is a graphical representation of a variable in the form of bars, where the surface of each bar is proportional to the frequency of the values represented. They serve to obtain an overview of the distribution of the population/sample, with respect to a characteristic, quantitative and continuous. First, let’s start with the age graphs.



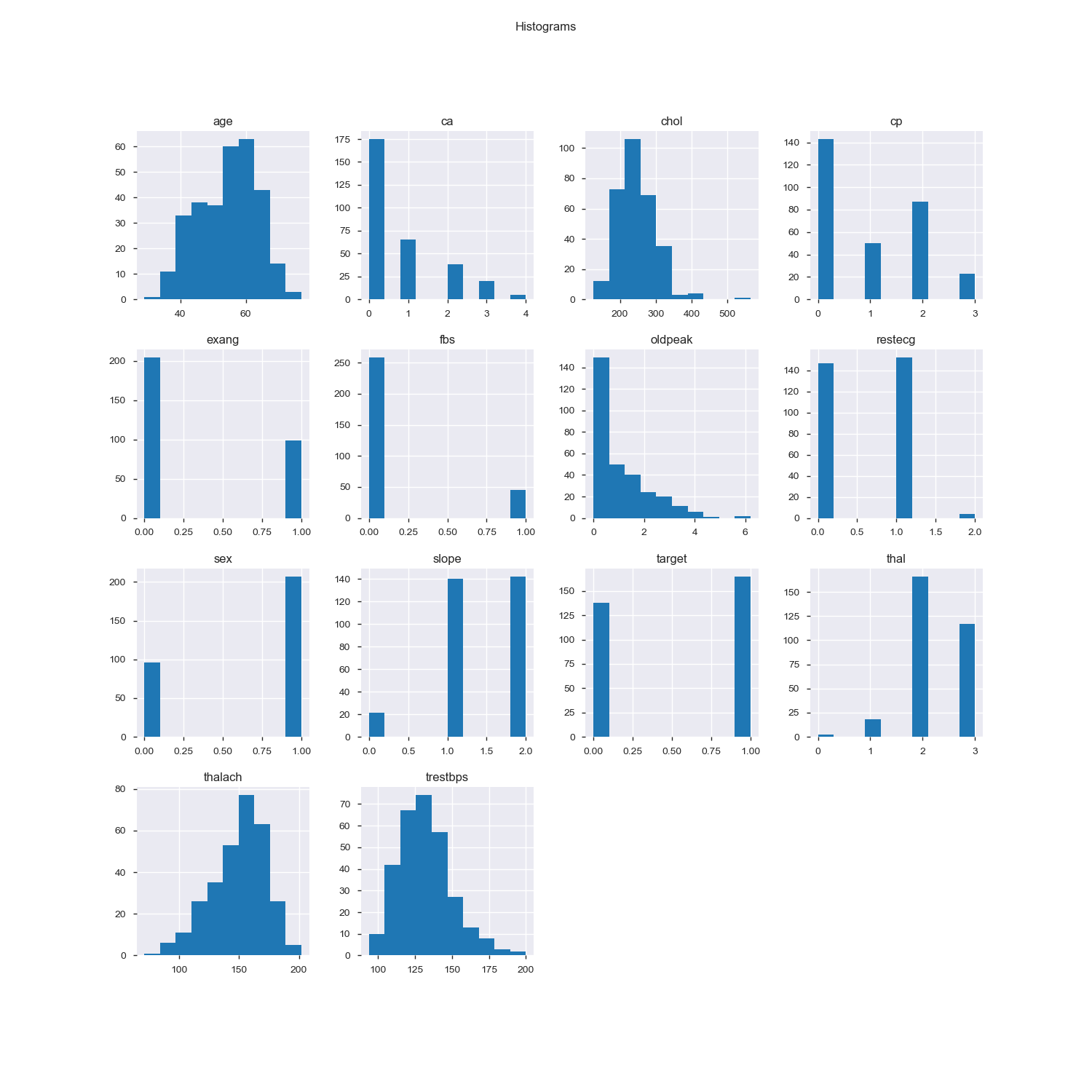
As expected, from the first picture it can be seen that we have a bigger presence of older people rather than young people for heart disease according by age. And from the second picture, as the age of patients was grouped into stages, it shows the frequency of appearance of stages for peoples’ age and it confirms that, for patients with heart diseases, the frequency of young people is much less than the frequency of mature/old people. According with Mayo Clinic (2019), one of the main risk factors to have a heart disease is the age due to aging increases the risk of arteries becoming damaged and narrowing, and of the weakening or thickening of the heart muscle.

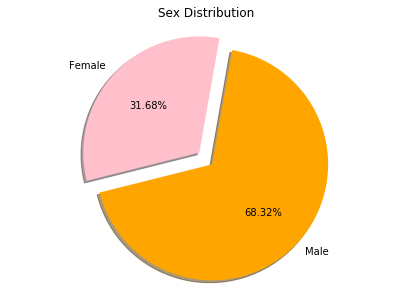
Now we are going the analyze non-graphical techniques for age variable. We have a mode = 58, median = 55, mean = 54. 366337 and a standard deviation of 9.082101. It is important to notice that the mode and median are greater than the mean. If we decided to see the distribution that the age variable follows, it would look like this:



As you can appreciate, the distribution is very close to a normal distribution but is not symmetric, why? Because by definition, for a normal distribution to be completely symmetric, the mean, median and mode values must be equals, this is**: μ = median = mode. Since median > μ,** without seeing the graph we can assume that our values will be skew to the left.

Now, we will continue with sex variable. From the first histogram, we can clearly see that the presence of heart disease is much more common in men than in women, in this case, we can said that, in the case of this dataset, the presence of men is approximately the double of the women population. From the second graph we can see that, from the total population of patients, which are 303, just the 31.68% of that population are women, while the other 68.32% are men, which is not rare because, according to information given by Mayo Clinic, in general, men are at greater risk of heart disease. However, the risk for women increases after menopause.





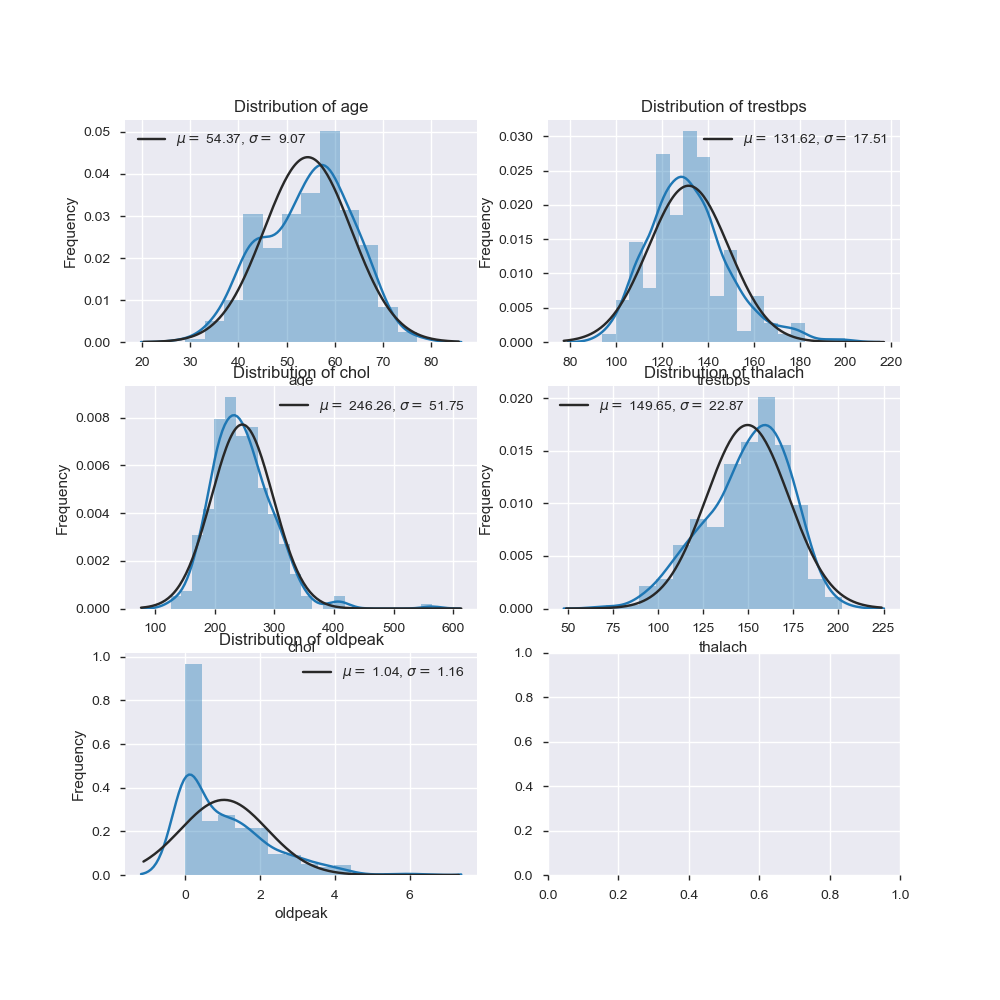
Now, let’s analyze the chol variable. On the chol variable the cholesterol in mg/dl of patients is shown. Before jump into the graphs, let’s take a look to non-graphical data. The central tendency measures are the following:

Mean = 246.264026

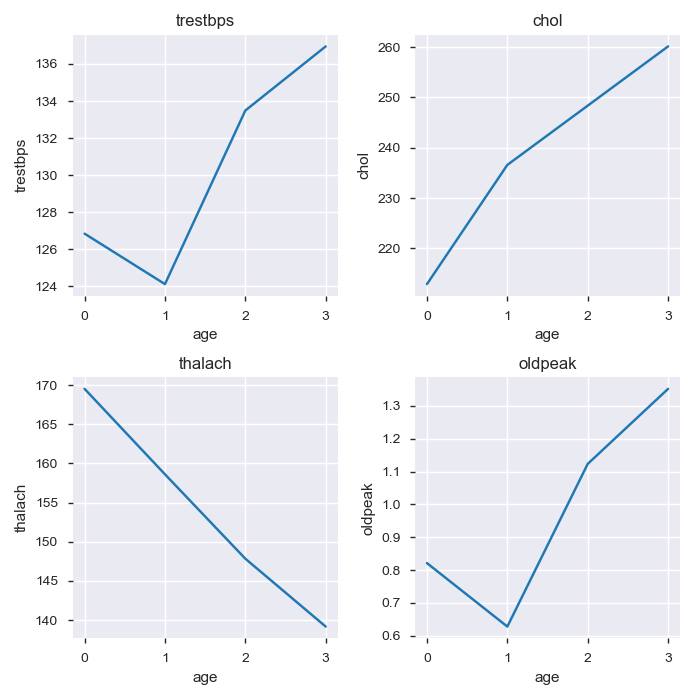
Mode = 197

Median = 240

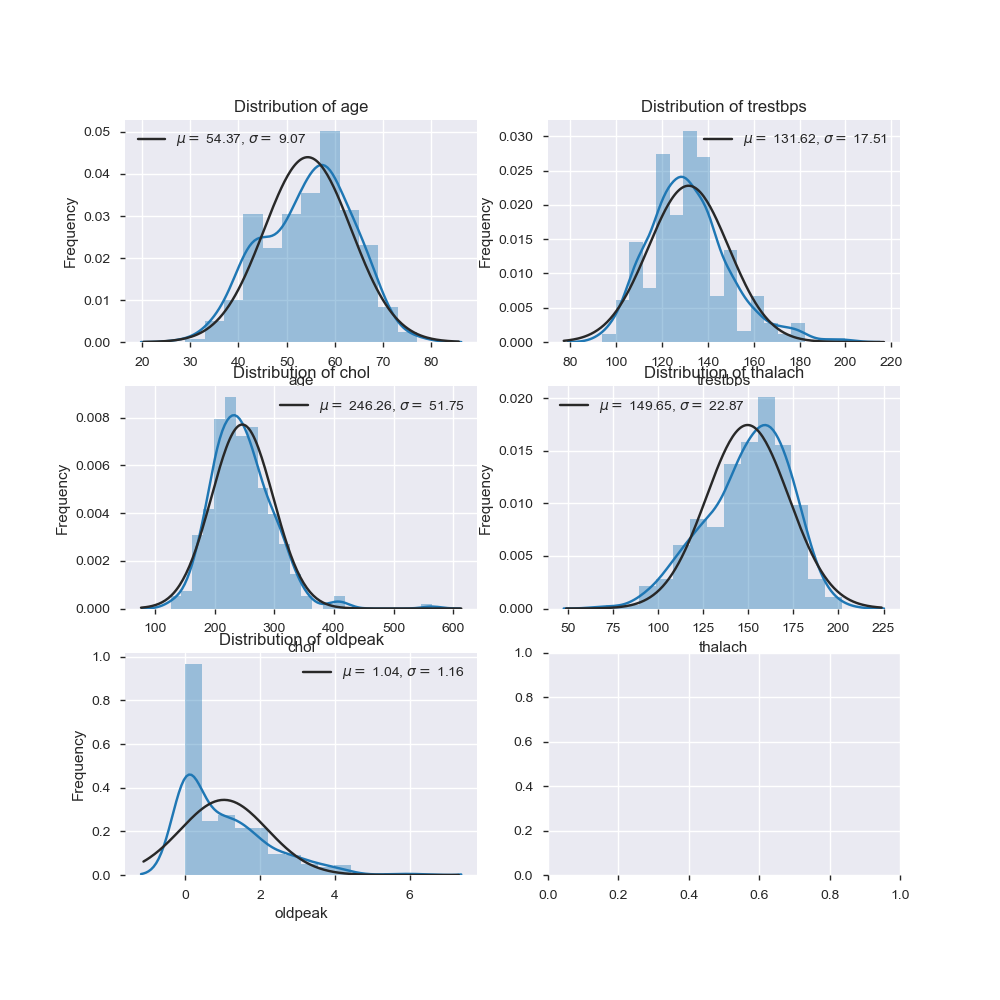
Here we can notice that the mode and median are less than the mean, thus, since **μ > median,** we can assume that our distribution is softly skew to the right. If we decided to see the distribution that the chol variable follows, it would look like this:



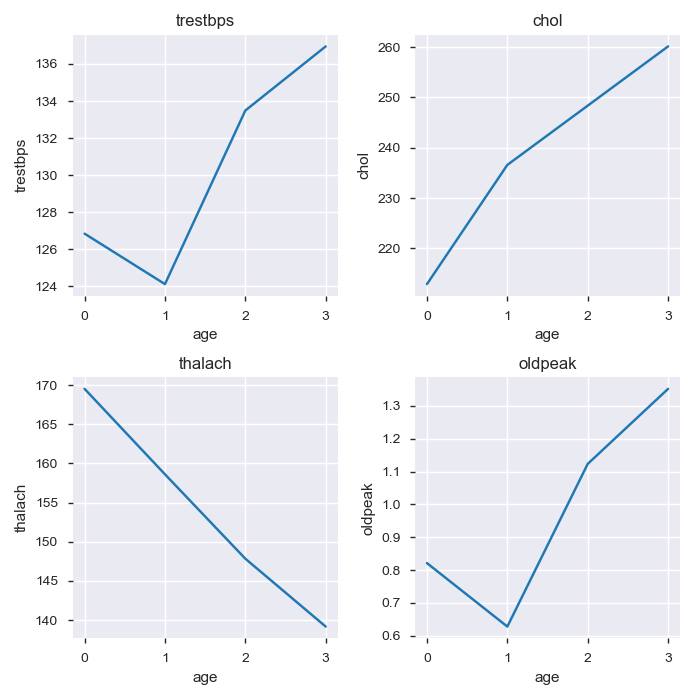
But just with the central tendency measures, it can be noticed that cholesterol values are greater than the normal amount. According to the American Heart Association, normal levels of cholesterol for a person older than 20 years old, a healthy level of cholesterol should lay in a range from 125 to 200 mg/dL, but in this case, the mode is equal to 197 mg/dL, we can say that the mode is almost touching the upper limit to be consider a healthy level, and the maximum value = 564 mg/dL, which is far away from the range of healthy values and it’s attributed to a patient who is 67 years old. If we pay close attention, it is possible to “accept” this data because as shows on the following graph, the cholesterol levels tend to increase with age, and non-controlled high levels of cholesterol in blood may increase the risk of plaque formation and **atherosclosis**.

Here, the x-axis are the stages of ages that, if you remember, were defined on the analysis of the age variable. And the y-axis is the cholesterol levels of patients. As expected, the colesterol level increases with age and on this graph, the stage 3, where the oldest ages of patients were grouped, has a major level of cholesterol.

Now, let’s take a look on the **trestbps** variable**.** The **trestbps** variable stands for resting blood pressure (in mm Hg on admission to the hospital). The mean value = 131.623762, median = 130 and mode = 120. Since **μ > median,** without looking at the graph distribution it can be assumed that the values population are skewed to the right. According to the World Health Organization (WHO), the normal blood pressure in adults is contained in a range of 120 mm Hg – 140 mm Hg/80 mm Hg – 90 mm Hg with an open interval. When the systolic tension is equal to or greater than 140 mm Hg and / or the diastolic tension is equal to or greater than 90 mm Hg, the blood pressure is considered high. By considering the central tendency measures, it can be observed that mean and mode values are quite a bit away from the healthy values. Uncontrolled high blood pressure can cause hardening and thickening of the arteries, which narrows the vessels through which blood circulates, also known as **atherosclosis.**



As espected, the distribution of blood preasure is skewed to the rigth and the levels got increase with age, as shown in both grahics.



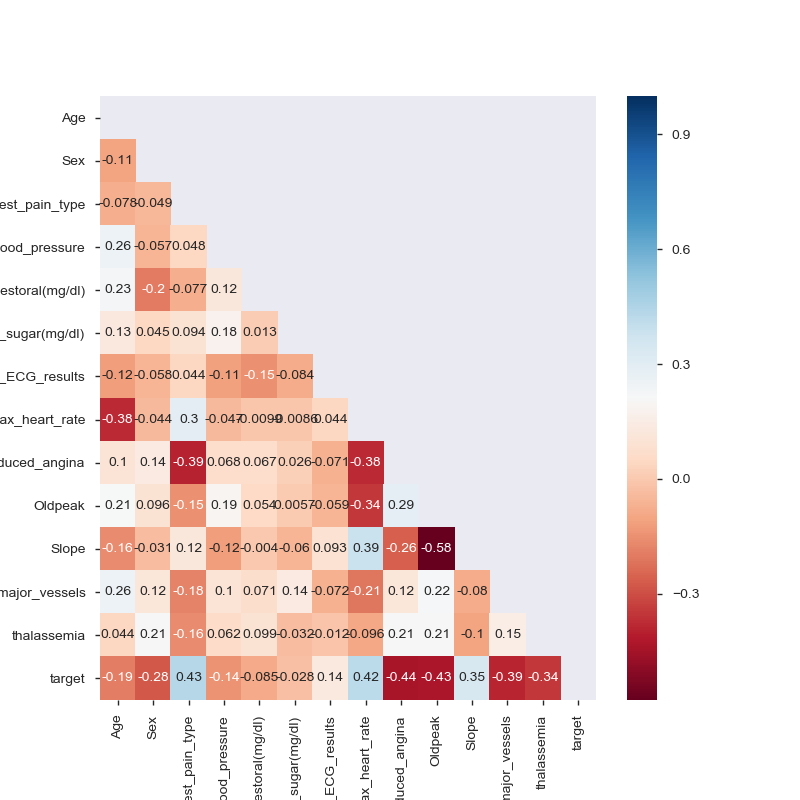
**Correlation between variables**

Here we have the correlation matrix. The correlation is a statistical measure that indicates how strongly two variables are related.

0 ≤ value ≤ 1: means that as one variable increase the other does the same: possitive correlation, thus, it is said that there is a possitive correlation between variables.

-1 ≤ value ≤ 0: means that if one increases, the other one dicreases. This property is know as negative correlation.

value = 0: means there is no relationship between the variables and it’s calles zero correlation.



From the heatmap it can be observed that there is not a strong correlation between variables, neither positive nor negative correlation, and most of them have a zero correlation. At the beginning, this I thought that there was an error on the dataset, but after analyzing the heatmap, what I think is happening here is that what the heatmap is showing is correct, maybe between variables there is no correlation or the correlation is quite low, almost zero; however, what may be happening is that the variables alone or compared between each other are not strongly relevant, but its relevance may come when you find the presence of many of them in an individual, this is, the “sum” of all these variables in a human being can be dangerous for someone’s health, it is possible that variables become relevant when they are found all together, the presence of many of them can be think as a set, it is possible that when you find a dangerous set of this variables on a patient, it is because the patient has a heart disease, and by dangerous I mean huge or quite low values. But as I have mentioned before, these are my thoughts.

References

* Mayo Clinic. (2019, January 15th). Enfermedad cardíaca - Síntomas y causas - Mayo Clinic. Recovered on October 17th, 2019, from <https://www.mayoclinic.org/es-es/diseases-conditions/heart-disease/symptoms-causes/syc-20353118>
* K2datascience. (s.f.). Heart Disease Study. Recovered on October 17th, 2019, from <https://github.com/k2datascience/advanced-classification-example>
* American Heart Association. (s.f.). Recovered on October 17th, 2019, from <http://www.heart.org/idc/groups/heart-public/@wcm/@gsa/documents/downloadable/ucm_487867.pdf>
* World Health Organization. (April 26th, 2016). Preguntas y respuestas sobre la hipertensión. Recovered on October 17th, 2019, from <https://www.who.int/features/qa/82/es/>
* Statistcs class 2. Recovered on October 17th, 2019, from <http://www.mat.uda.cl/hsalinas/cursos/2011/2do/clase2.pdf>