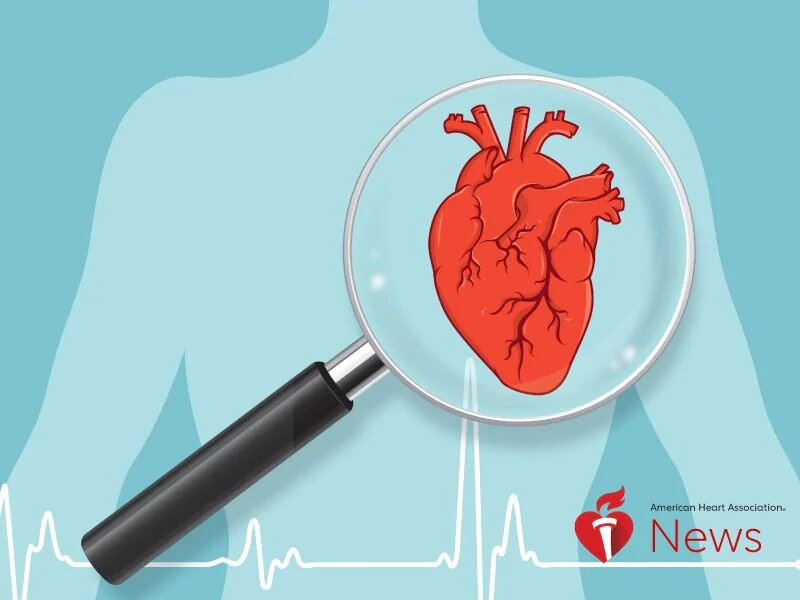
**Nuriye Merve TATLIDİL – 150212002**

Heart Disease Data Set Project



**1-Introduction**

I worked with Jupyter Notebook in my project. I found the dataset I was working on kaggle.com.This data set consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "**target**" field refers to the presence of heart disease in the patient. It is integer valued

0 = no disease

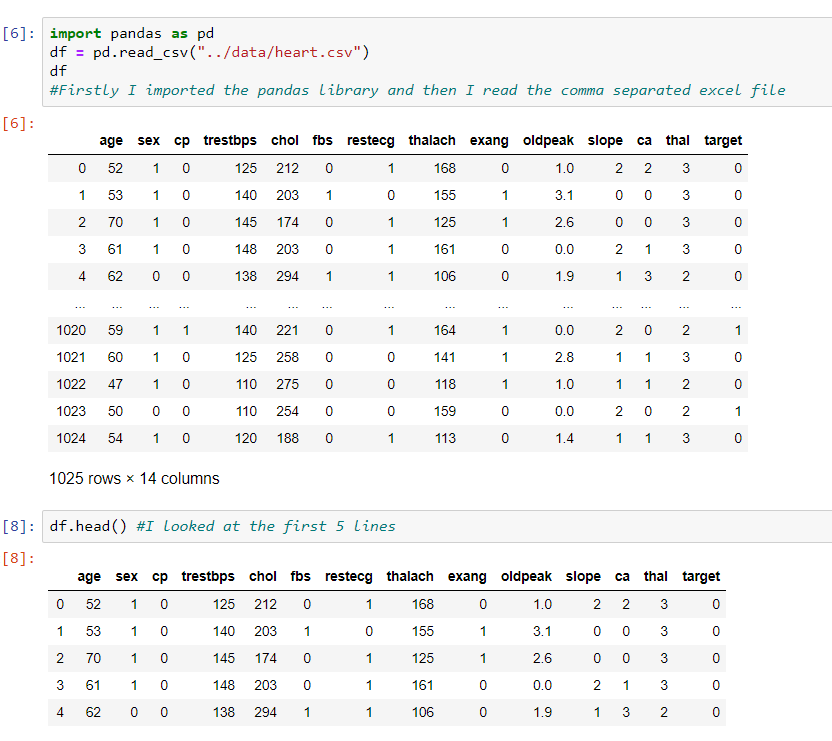
1 = disease.

### 2- General View of the Data

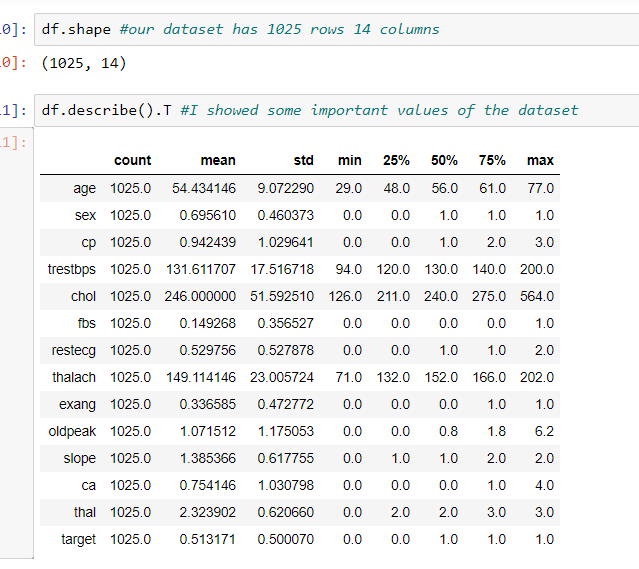
1. Age (In years)
2. Sex 1 - Male 0 - Female
3. CP (Chest Pain Type)  0 - Typical Angina (Heart related) 1 - Atypical Angina (Non-heart related) 2 - Non-Anginal pain (Non-heart related) 3 - Asymptomatic (No disease)
4. TRESTBPS (Resting Blood Pressure (in mm Hg on admission to the hospital))
5. CHOL (Serum Cholestoral in mg/dl) Healthy serum cholesterol is less than 200 mg/dL
6. FPS (Fasting blood sugar > 120 mg/dl) 1 - True 0 - False
7. RESTECH (Resting Electro Cardio Graphic results)
8. THALACH (Maximum heart rate achieved)
9. EXANG (Exercise induced Angina) 1 - Yes 0 - No
10. OLDPEAK (ST depression induced by exercise relative to rest)
11. SLOPE (Slope of the peak exercise ST segment)
12. CA (Number of major vessels (0-3) colored by Flouroscopy)
13. THAL 0 - Normal 1 - Fixed defect 2 - Reversible defect
14. TARGET 1 - Heart Problem 0 - No Heart Problem

**3-Importing Library**

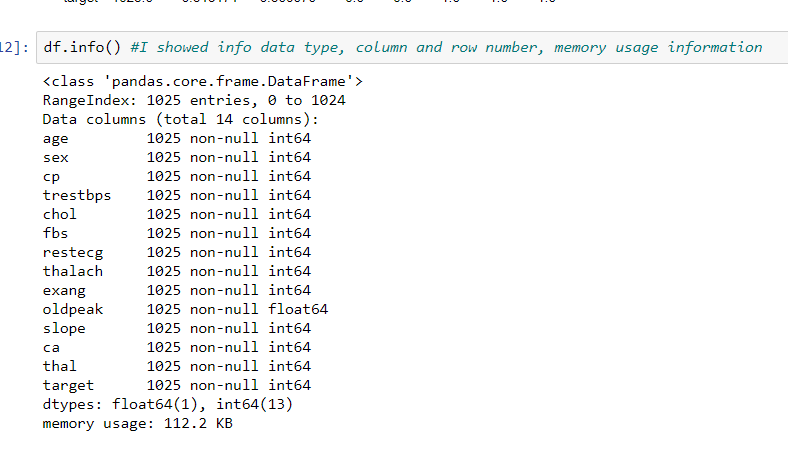
Since I will use the dataset as a dataframe, I first imported the pandas library. Then I added the dataset to my project.



I showed some important values of the attributes in the data set. I applied the transposition because I thought it looked more understandable when I received the transposition.



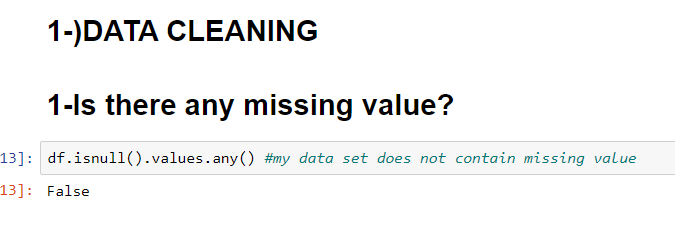
It is important for the application of algorithms that the types of our data are categorical or numerical.For this reason, I looked at how many numerical and categorical variables there are because we would convert categorical variables into numerical variables in the next steps.We can also see the size of the data set in this way.



**4-Pre-Processing Step**

There are many steps in the data preprocessing phase. One of them is data cleaning. In this phase, we check the missing, noisy and outlier data.

First, I check for missing data :

****

My data set does not contain missing data.

Secondly, I check for outlier data :

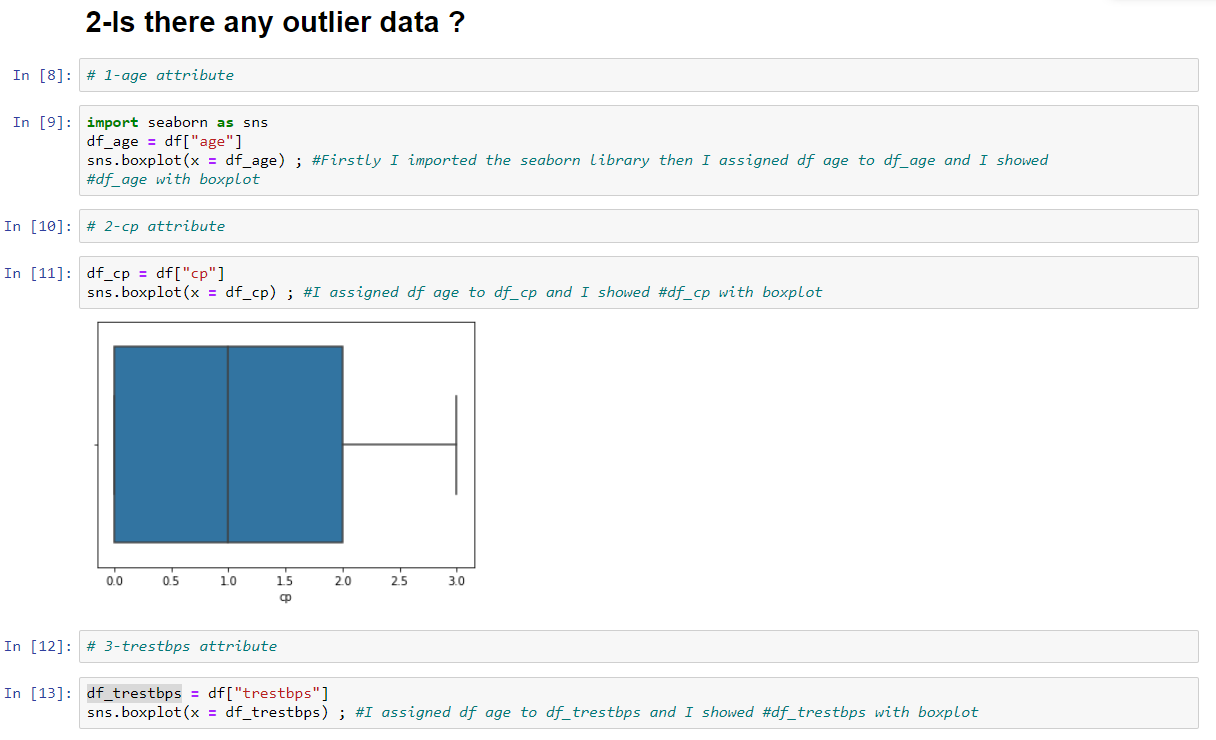
I know 2 methods to find outlier values:

1- Showing with Boxplot graphic

2- Extreme Value Determination

Firsly I show with the age attribute boxplot :

Age and cp attribute don't have outlier values.



The trestbps attribute has outlier data.in this way, the data with Extreme Value Determination to look at which rows are outlier data.

For Outlier data

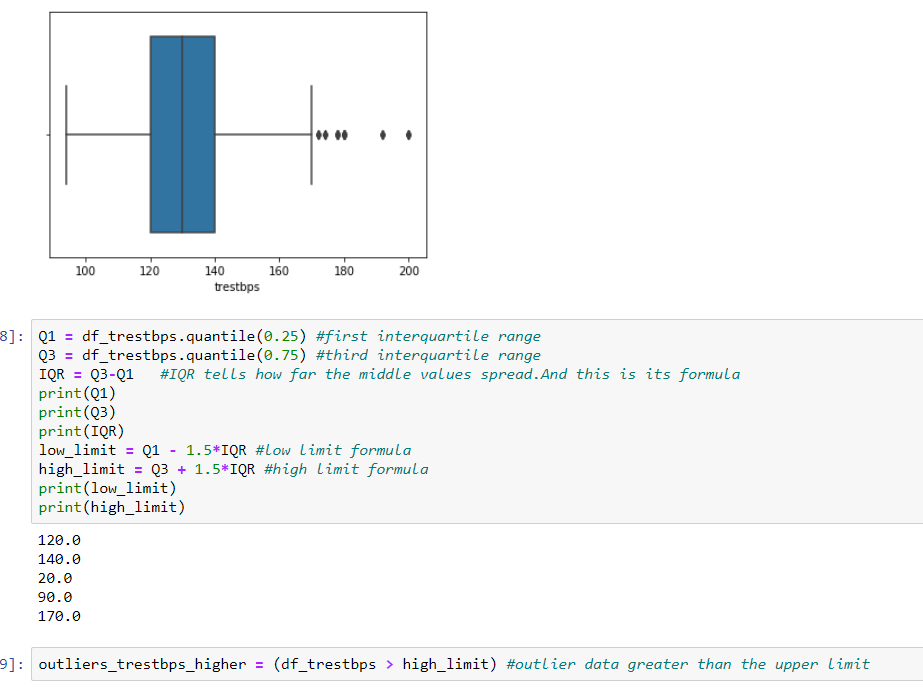
1-deletion

2-filling with mean

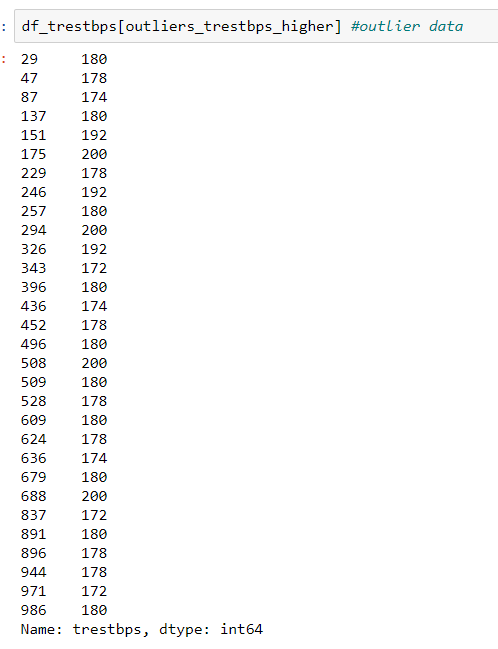
3- suppression.

we can do it. Outlier data I chose to apply the printing process in all the attributes I found.Because there are many characteristics of outlier data.I could have lost a lot of information if I deleted it all.when it comes to filling with mean,

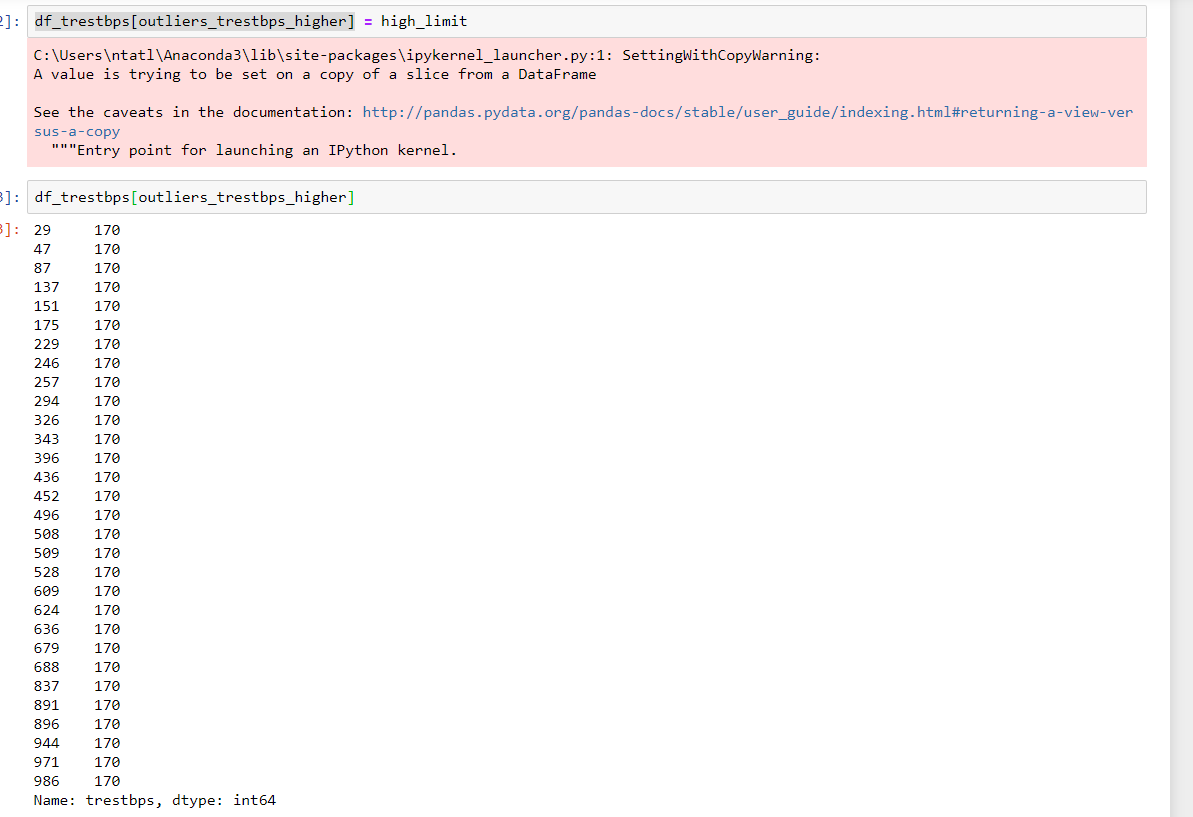
For example the outlier value is 170.let the mean value of the attribute be 80.I thought I might get a wrong result because the difference would be too much. that's why I applied suppression.

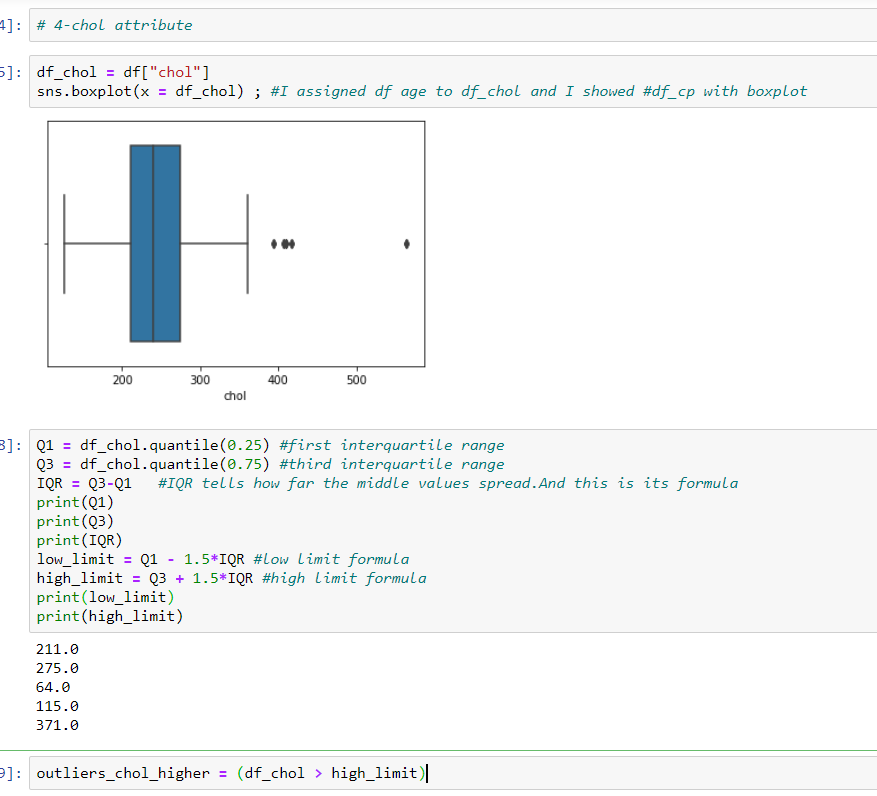


Outlier data rows :

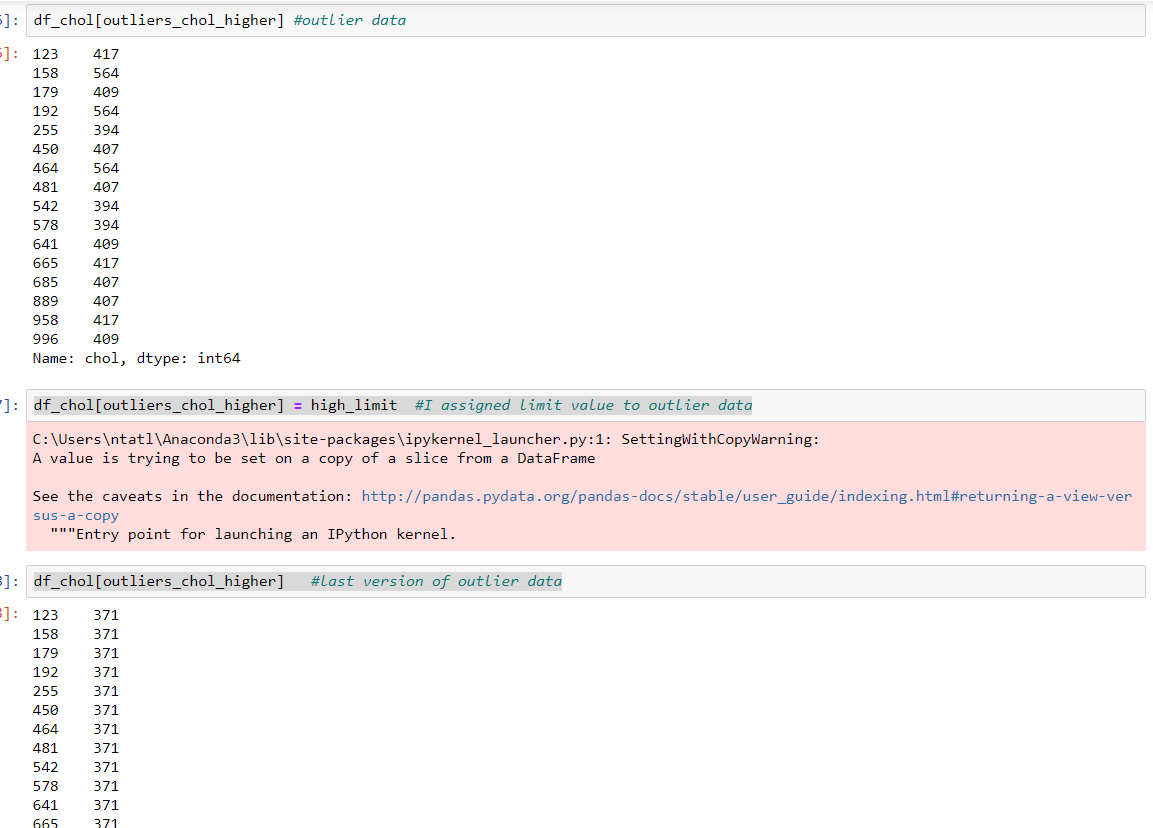


Trestbps outlier data is above the upper limit.So I filled out the outlier data with the upper limit value.So they were in the range I had set, but they had the number closest to them in the range.





Chol outlier data is above the upper limit.So I filled out the outlier data with the upper limit value.So they were in the range I had set, but they had the number closest to them in the range.

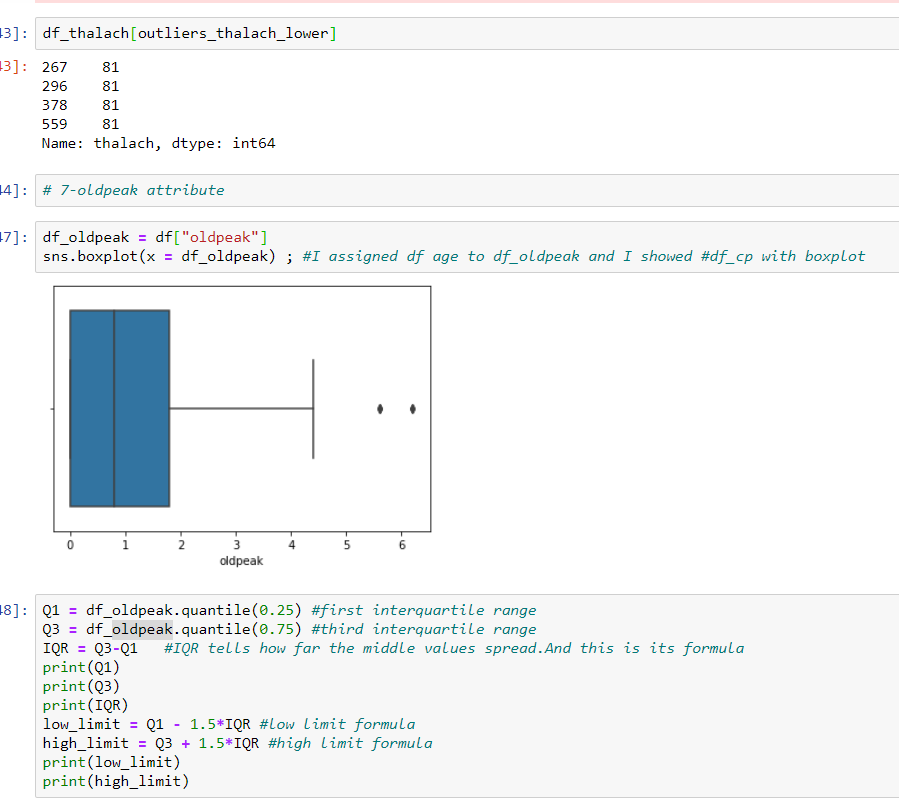


Restecg attribute has not outlier datas.



Thalach outlier data is above the lower limit.So I filled out the outlier data with the lower limit value.So they were in the range I had set, but they had the number closest to them in the range.





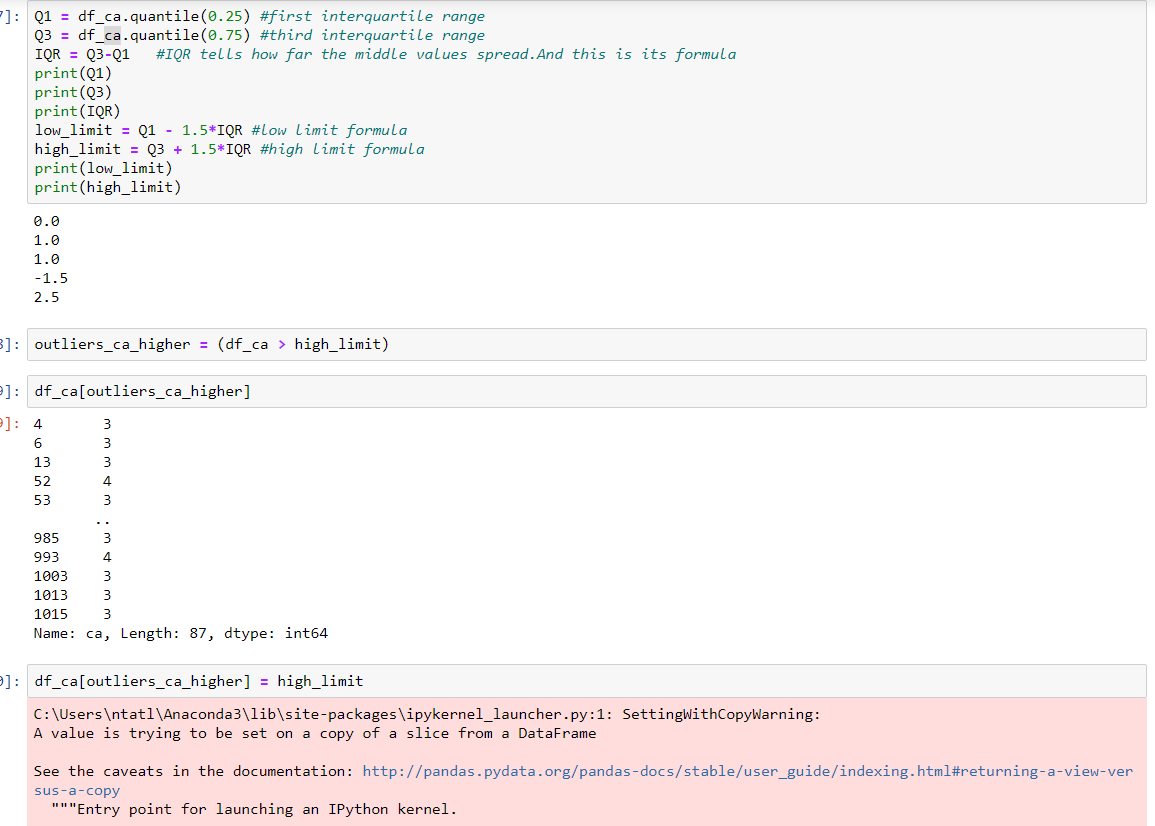
Oldpeak outlier data is above the upper limit.So I filled out the outlier data with the upper limit value.So they were in the range I had set, but they had the number closest to them in the range.



Slope attribute has not outlier datas.



Ca outlier data is above the upper limit.So I filled out the outlier data with the upper limit value.So they were in the range I had set, but they had the number closest to them in the range.

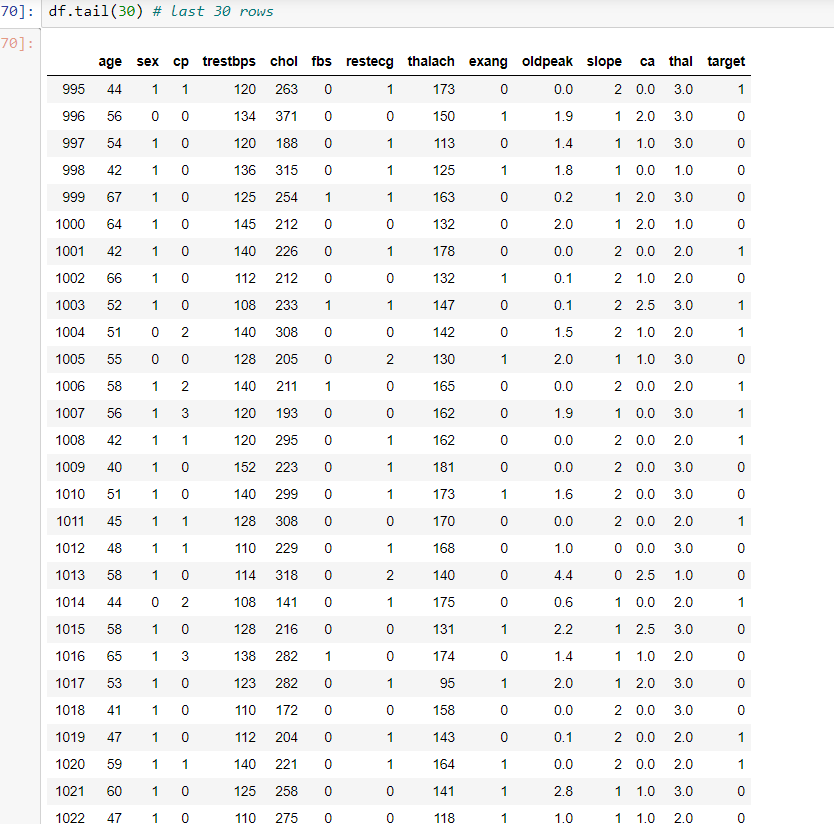






Thal outlier data is above the lower limit.So I filled out the outlier data with the lower limit value.So they were in the range I had set, but they had the number closest to them in the range.

If I look at the final attributes :

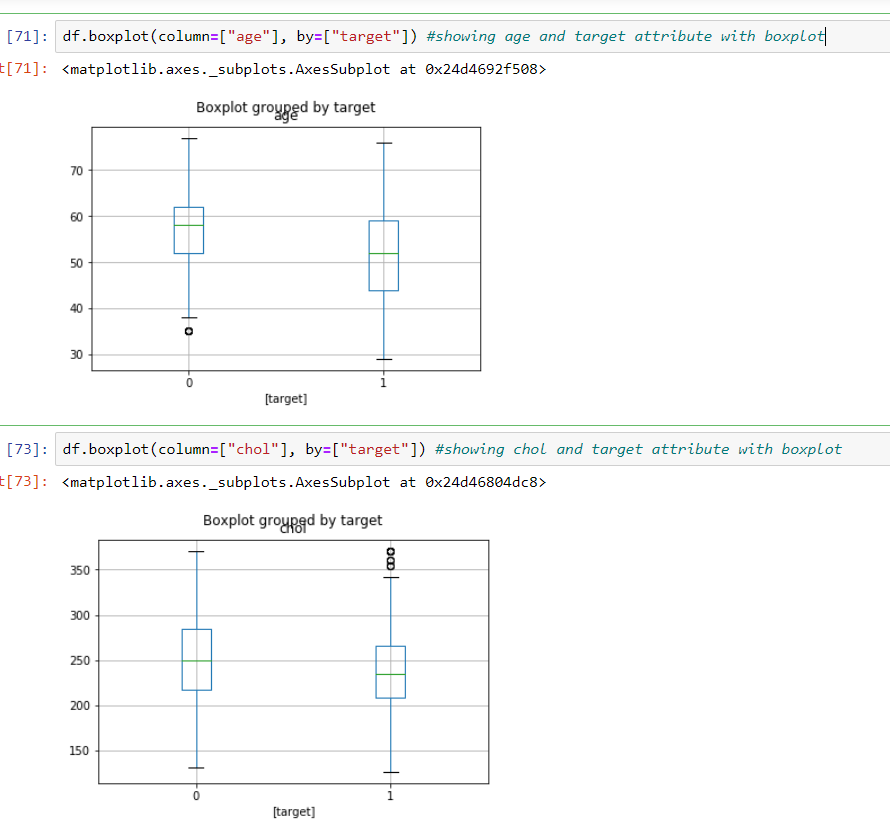


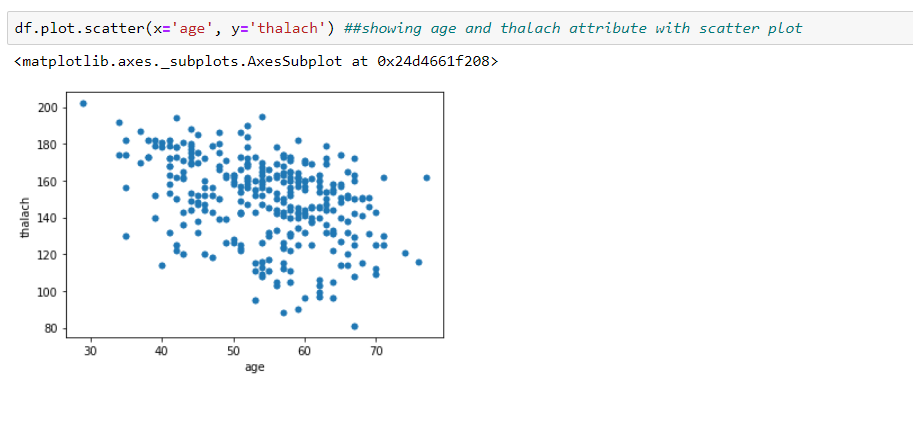
I have looked at these lines as the changed lines are in the last 30 lines.

One of the steps of data preprocessing is to perform data transformation.Since there are no categorical variables in my dataset, I don't need to do data transformation.

**4-Data Visualization**

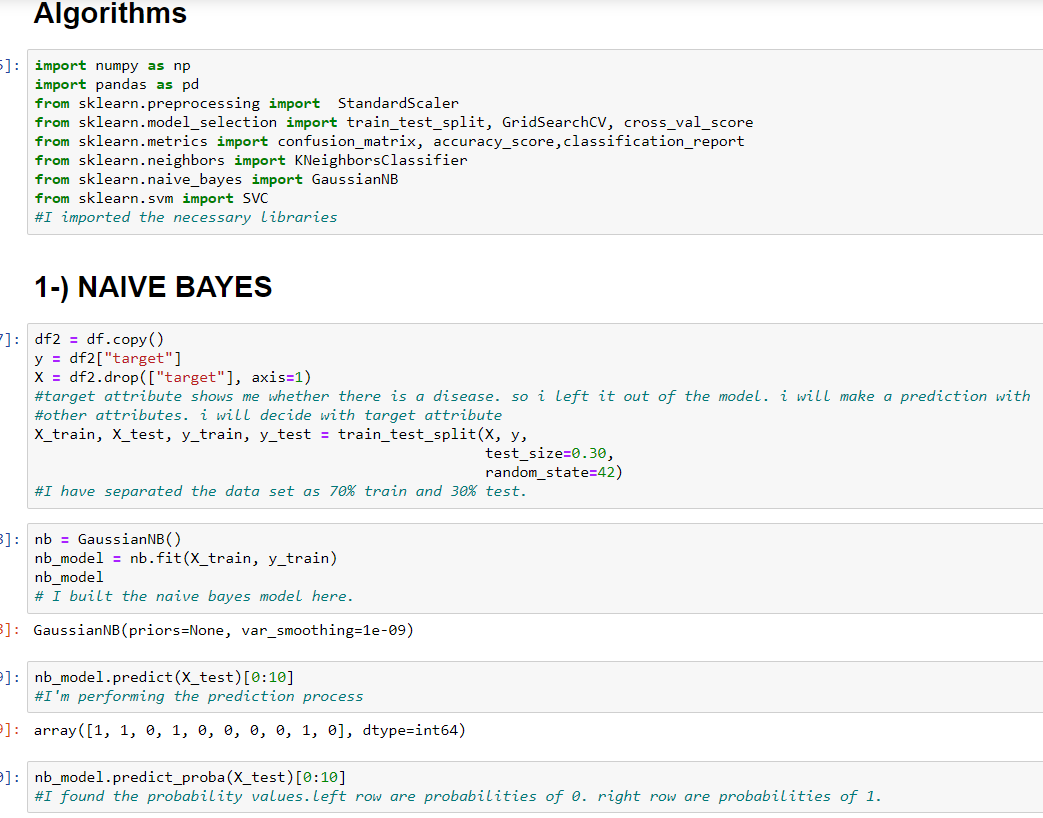
Displaying data with graphics is important for understanding the status of data in data science, so I have visualized the data with some types of graphics we learned in the advanced topics lesson.

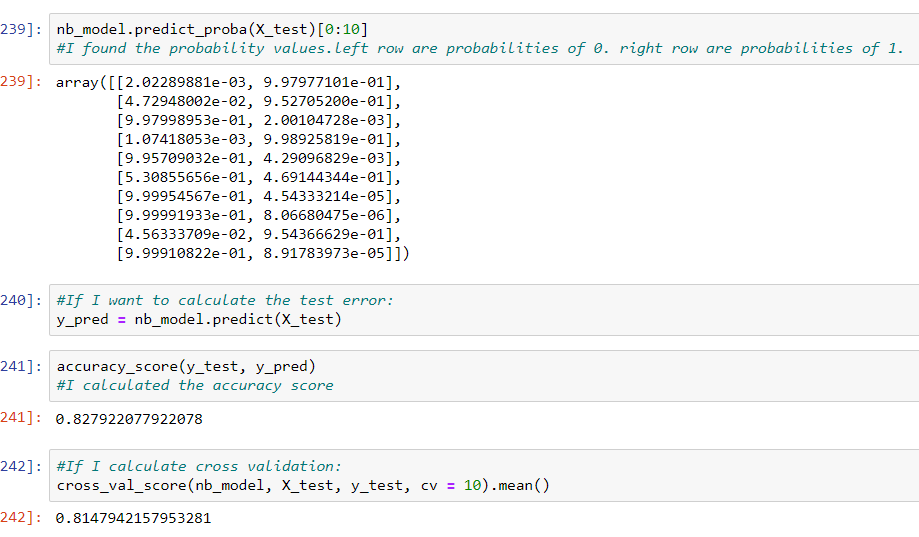
****

****

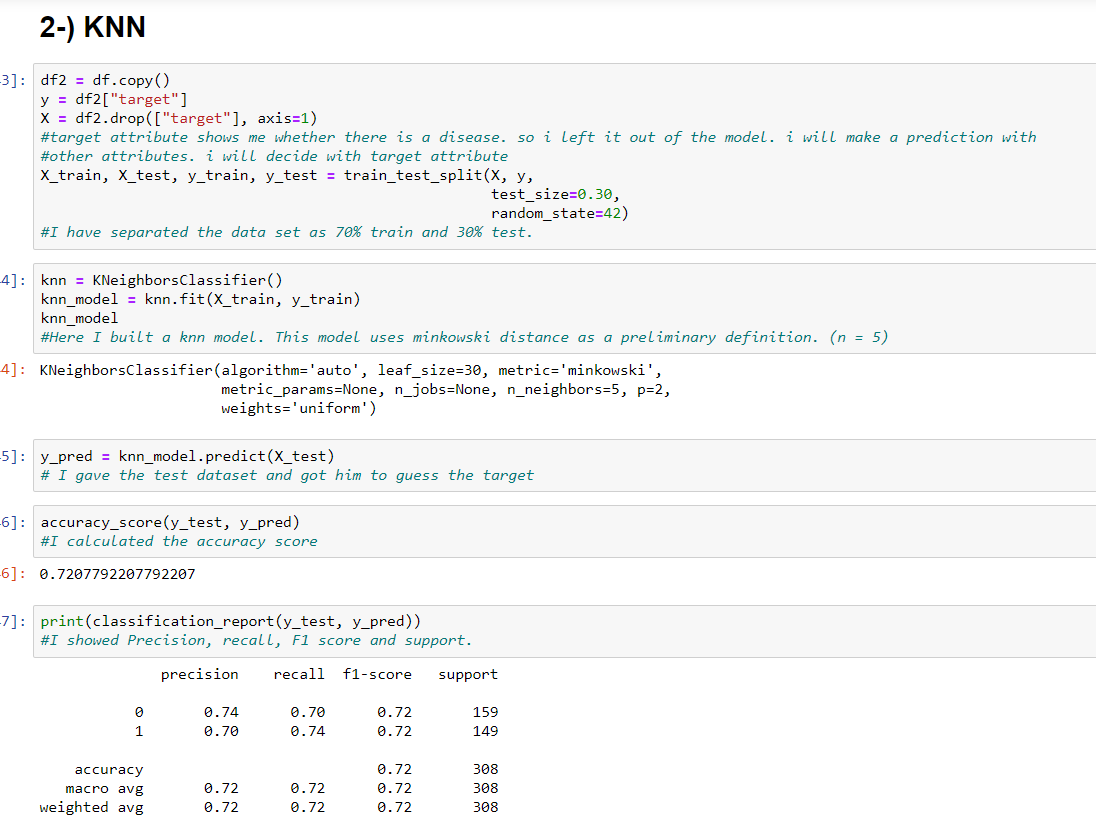
**5-Algorithms Step**

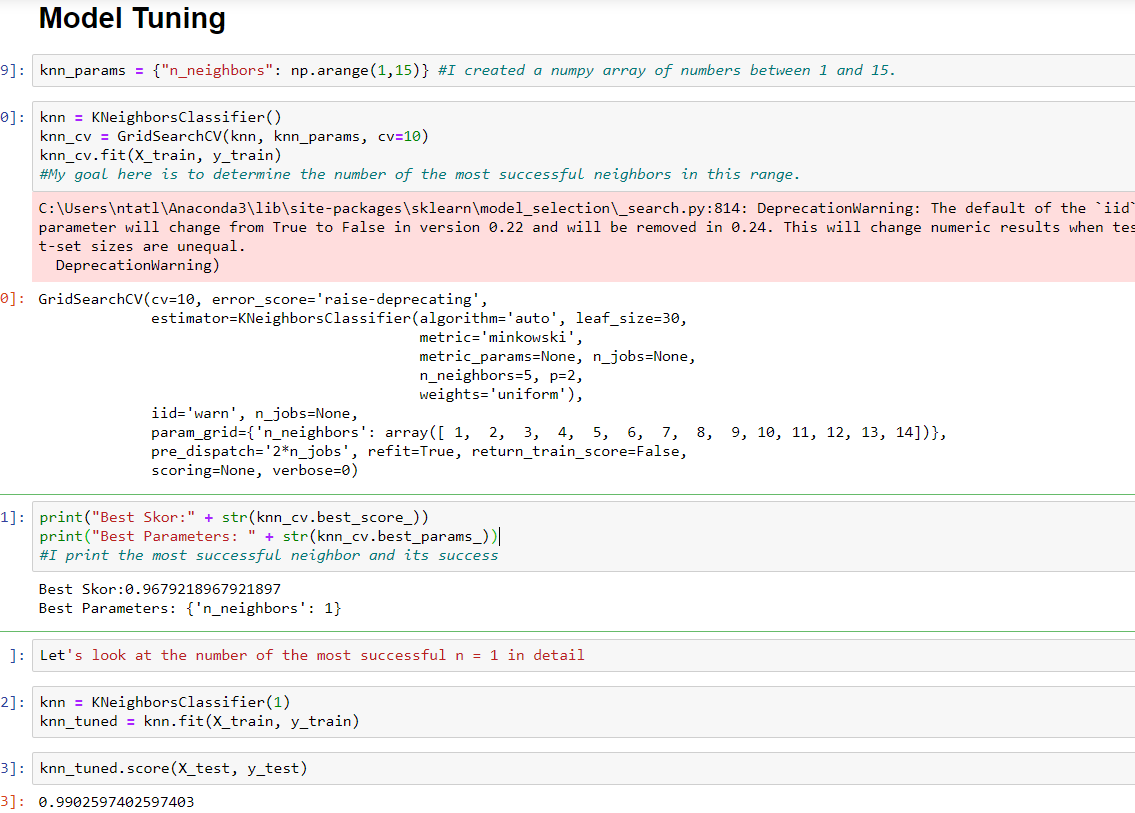
1. **Naive Bayes Algorithm**

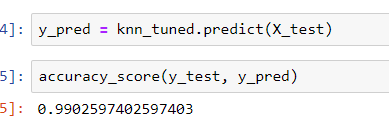
****

****

1. **KNN (K-Nearest Neighbor) Algorithm**

****

****

****

1. **SVC (Support Vector Machine) Algorithm**

