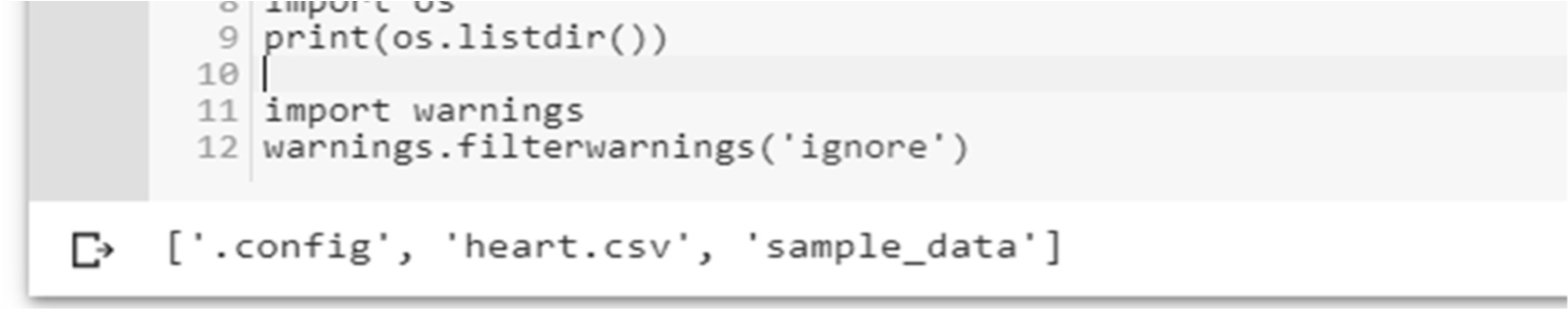
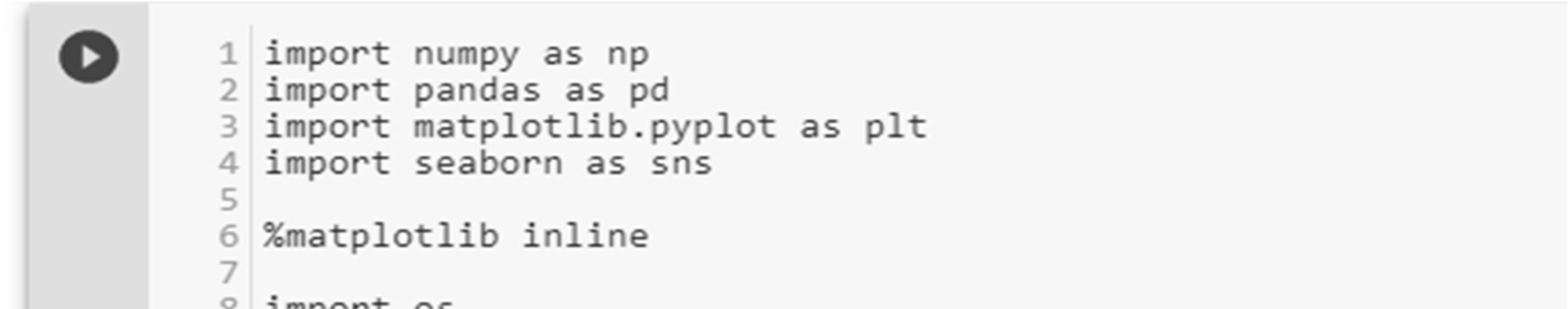
# Dataset Structure & Description

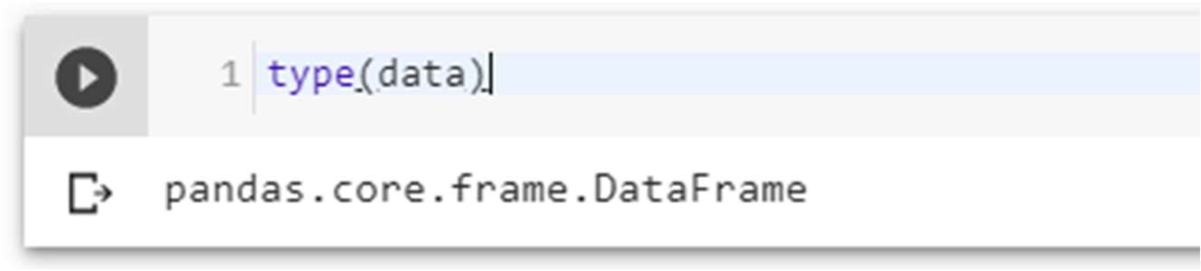
***Importing libraries***



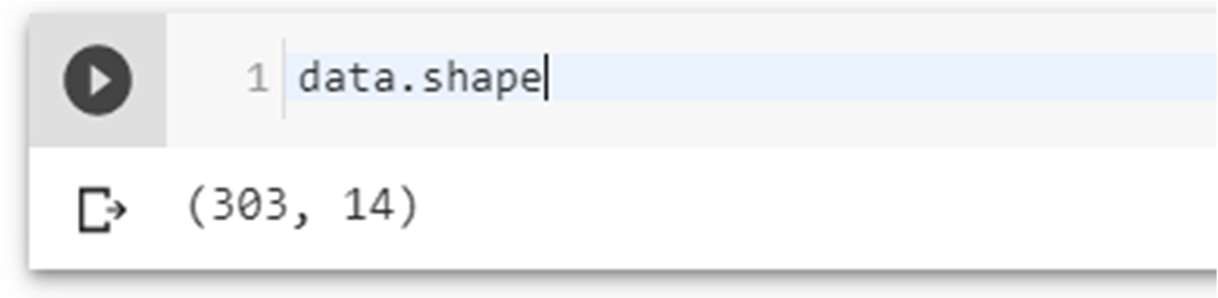
***Load data***



***Check the type of the dataset***



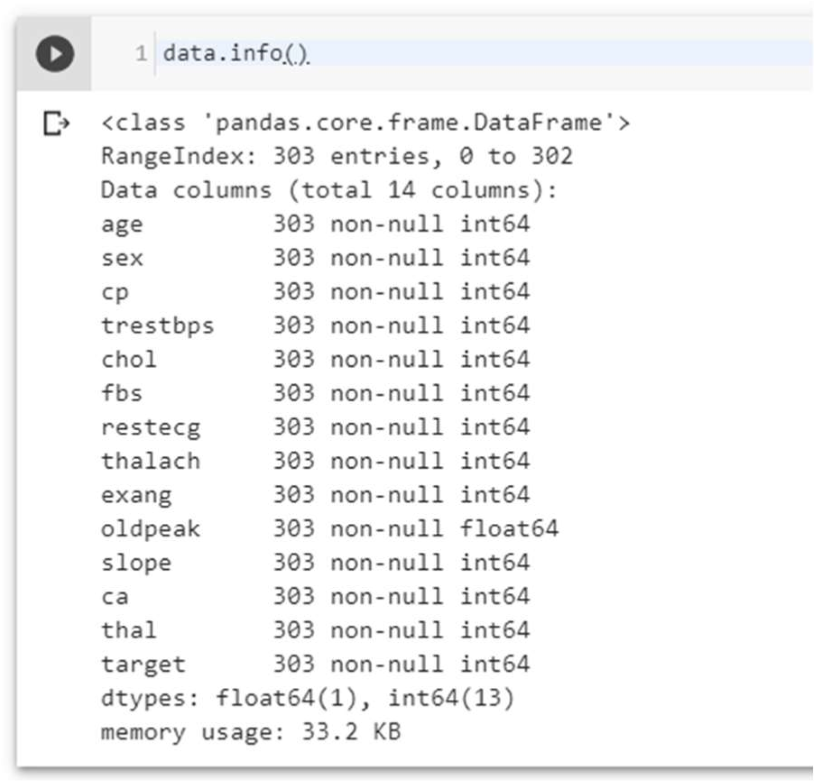
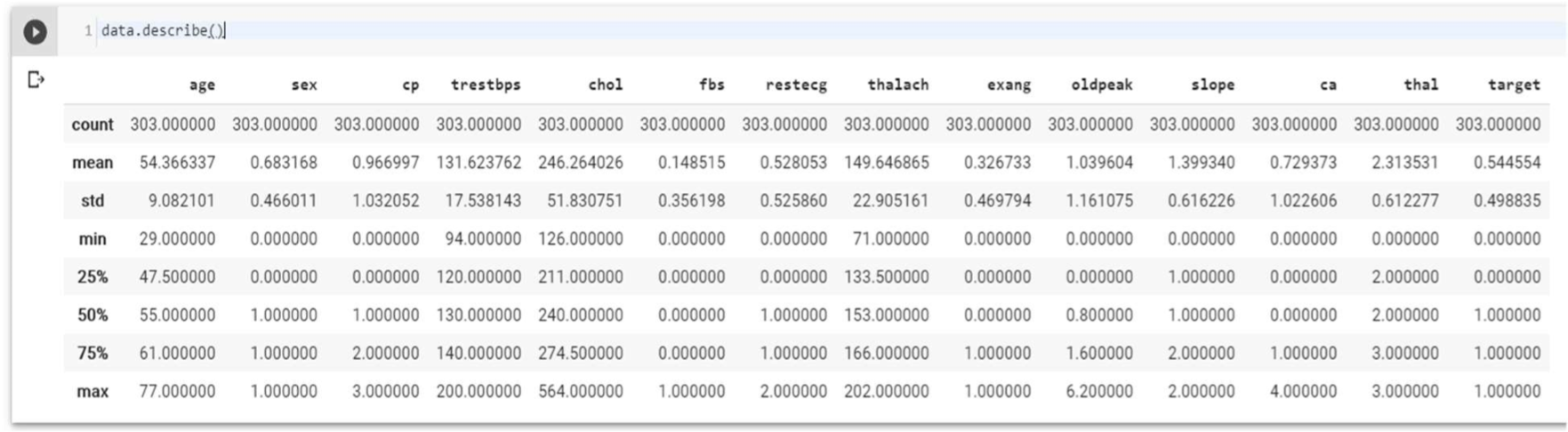
***Check the Shape of the data***

******

***Check the top four columns of the dataset***



***Dataset description***



The dataset used in this project contains 14 variables. The independent variable that needs to be predicted, 'diagnosis', determines whether a person is healthy or suffer from heart disease. Experiments with the Cleveland database have concentrated on endeavors to distinguish disease presence (values 1, 2, 3, 4) from absence (value 0). There are several missing attribute values, distinguished with symbol '?'. The header row is missing in this dataset, so the column names have to be inserted manually.[124] Features information:

* age - age in years
* sex - sex (1 = male; 0 = female)
* chest pain - chest pain type (1 = typical angina; 2 = atypical angina; 3 = nonanginal pain; 4 = asymptomatic)
* blood pressure - resting blood pressure (in mm Hg on admission to the hospital)
* serum cholesterol - serum cholesterol in mg/dl
* fasting blood sugar - fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
* electrocardiographic - resting electrocardiographic results (0 = normal; 1 = having

ST-T; 2 = hypertrophy)

* max heart rate - maximum heart rate achieved
* induced angina - exercise induced angina (1 = yes; 0 = no)
* ST depression - ST depression induced by exercise relative to rest
* slope - the slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = down sloping)
* no of vessels - number of major vessels (0-3) colored by fluoroscopy
* thalassemia - 3 = normal; 6 = fixed defect; 7 = reversable defect
* diagnosis - the predicted attribute - diagnosis of heart disease (angiographic disease status) (Value 0 = < 50% diameter narrowing; Value 1 = > 50% diameter narrowing)

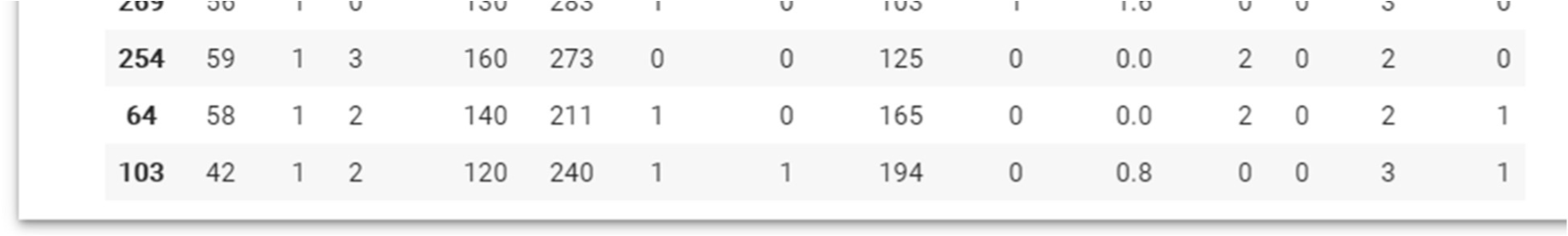
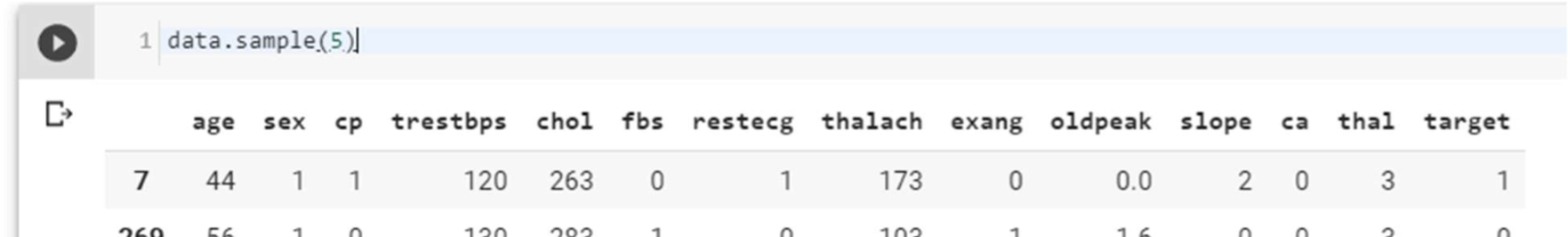
***Types of features***

Categorical features (Has two or more categories and each value in that feature can be categorized by them): sex, chest pain

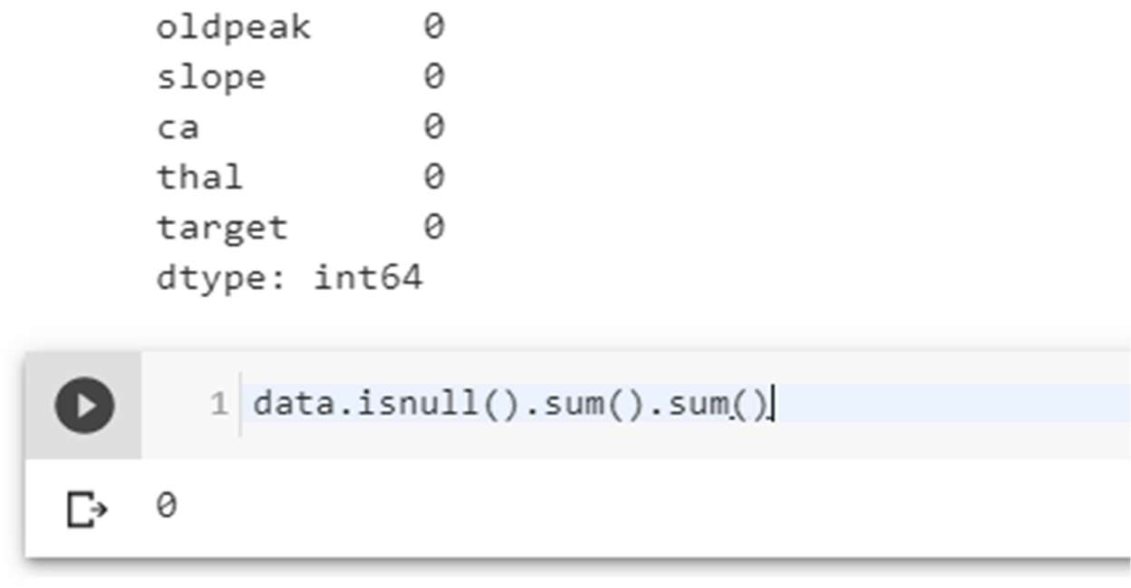
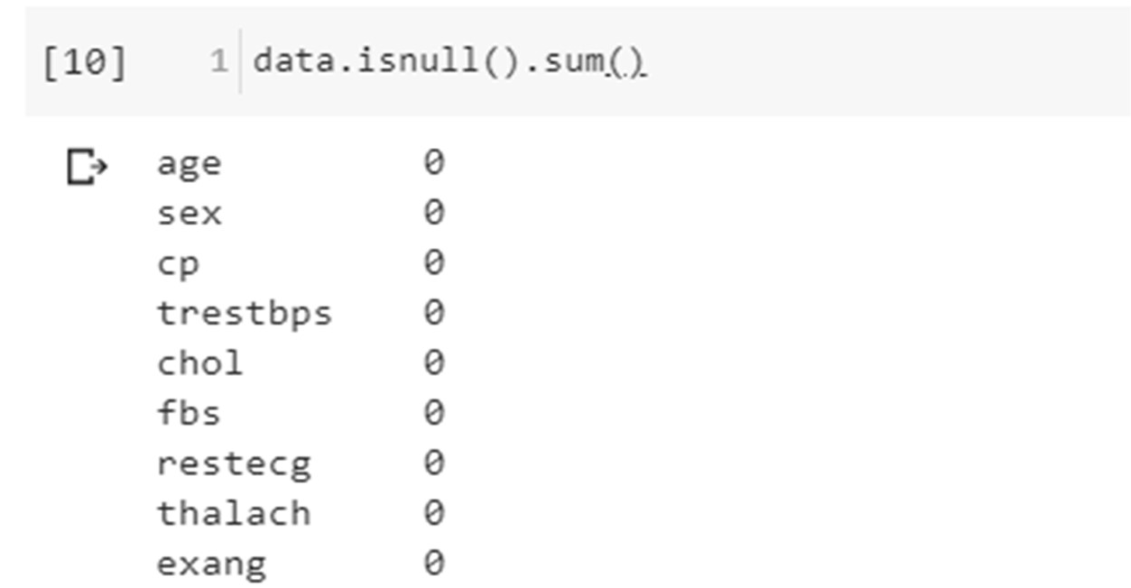
Ordinal features (Variable having relative ordering or sorting between the values): fasting blood sugar, electrocardiographic, induced angina, slope, no of vessels, thalassemia, diagnosis

Continuous features (Variable taking values between any two points or between the minimum or maximum values in the feature column): age, blood pressure, serum cholesterol, max heart rate, ST depression

***Some Random data columns***



***Check for missing Data***



No Data is missing, which is good.

***Check the correlation with target data***

