Datasets for Final Project

José Ignacio Díez Ruiz – 100487766
Carlos Roldán Piñero – 100484904
Pablo Vidal Fernández – 100483912

2022-11-14

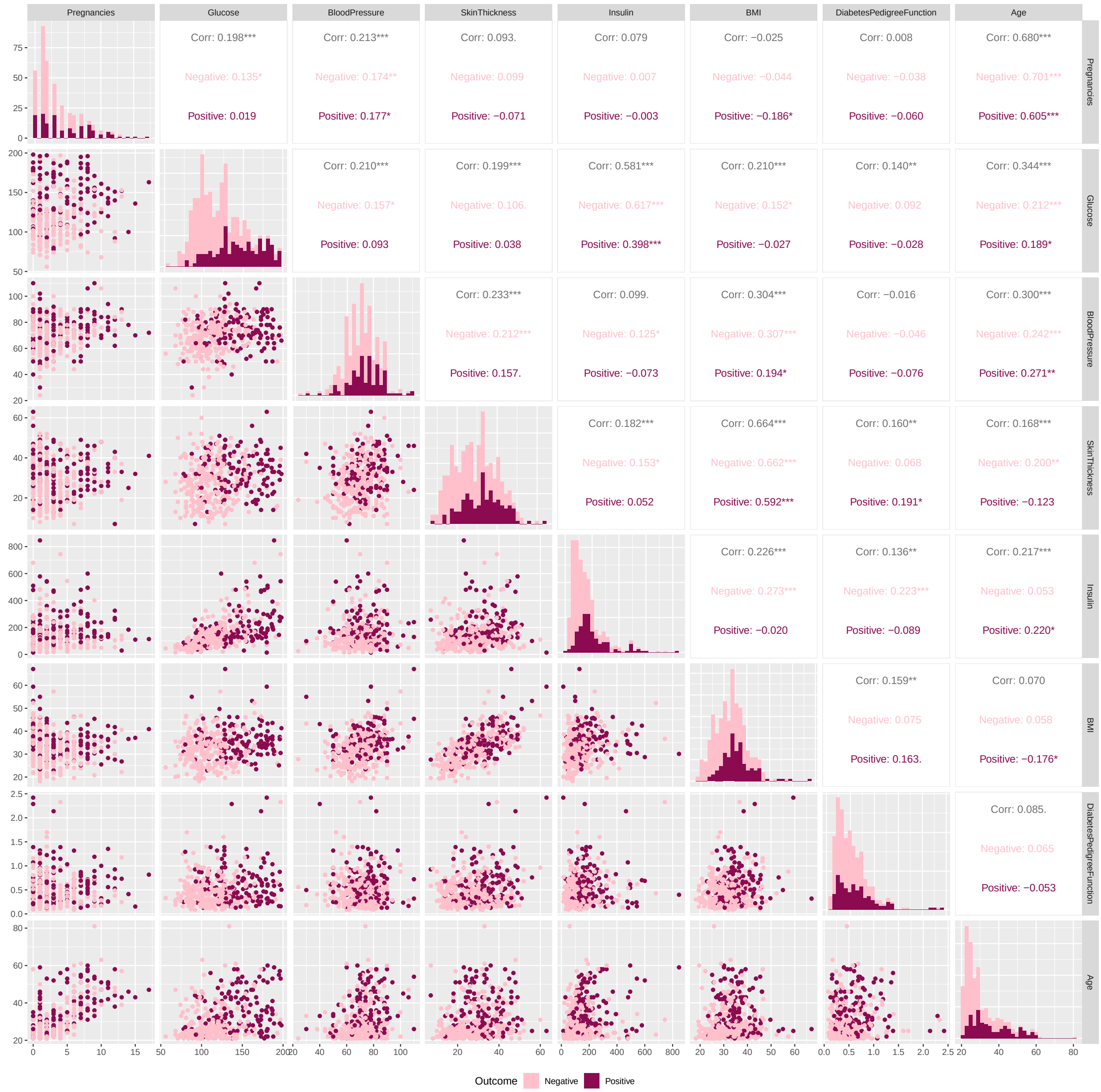
We have found two datasets of interest. Here we place a little description of the variables in each of them, as well as a plot of the distribution for each of the quantitative variables among them.

Dataset 1: Diabetes

This dataset contains information regarding diagnostic measurements of females over 21 years old of Pima Indian Heritage. It is comprised of 768 individuals and 9 variables. The dataset can be found in https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

On the particulatities of the dataset it is worth to comment that although there are not explicit NA values, they are present as 0 values in characteristics or tests that do not make sense to have a 0 result. In order to properly define the present variables, we did clean those fake measurements.

Before we start adressing the individual variables, we do a quick grid plot to see if any of our variables is strongly correlated with another so that we may reduce the number of them.

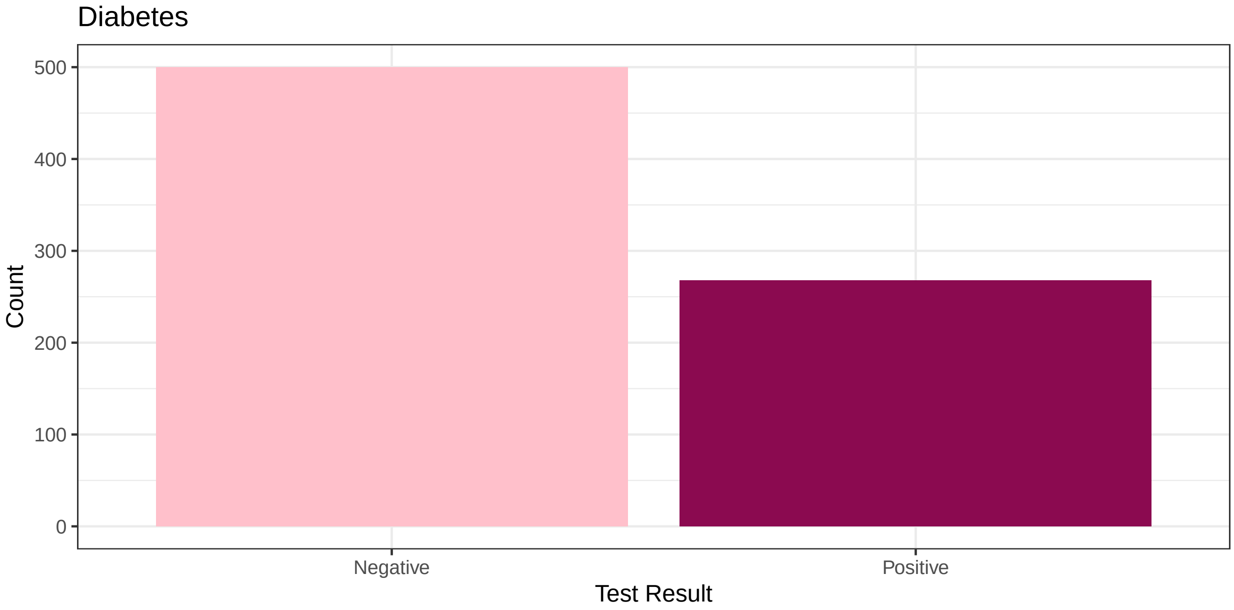


We do observe some correlations between them, although not at a level close enough to consider a variable reduction at this point. We may hence start reviewing each of the variables alone.

Outcome

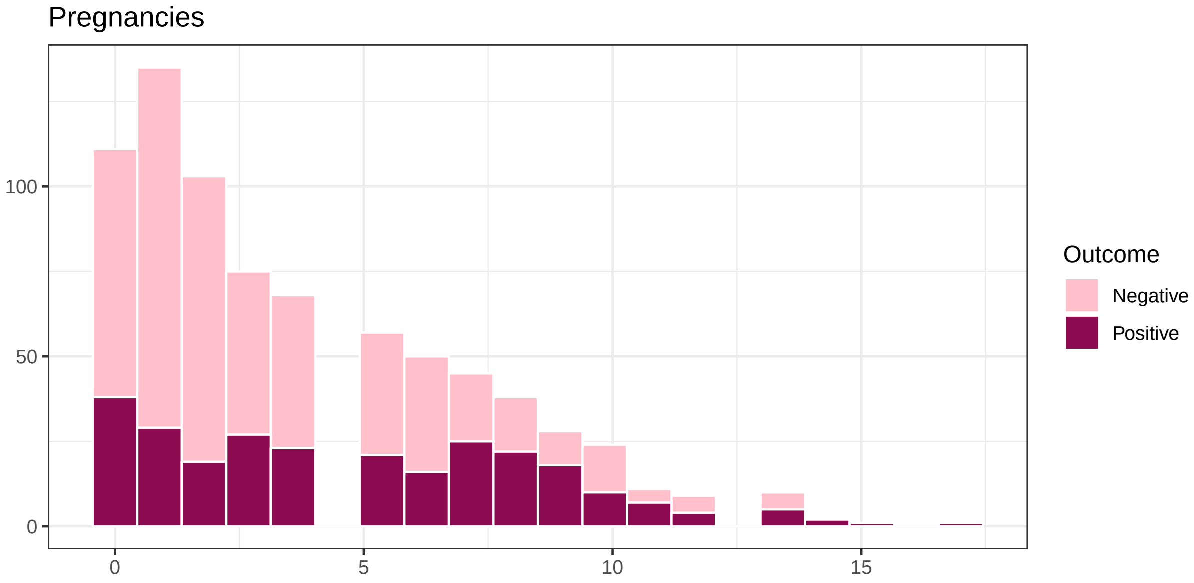
Qualitative nominal variable that indicates if the individual has diabetes or not.

Its values has been factorized as *Positive* for positive diabetes tested individuals and *Negative* for those without diabetes.



Pregnancies

Quantitative discrete variable detailing the number of times the individual has been pregnant. Its distribution is the following:

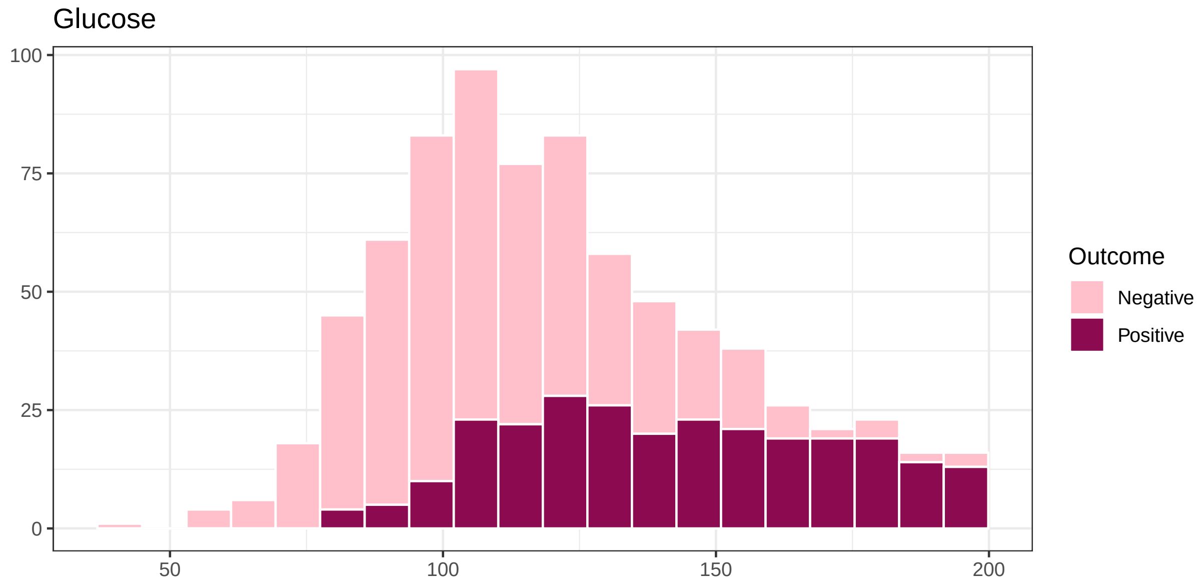


As observed, the majority of the individuals had either none or a low number of pregnancies. However there is a non-negligible presence of outliers.

At first glance we appreciate that, for high number of pregnancies, the ratio between positive and negative diagnostized individuals converges to a stalemate. This could indicate a positive dependency such that, as the number of pregnancies increase, the risk of developing diabetes increases too.

Glucose

Quantitative continuous variable that reflects the plasma glucose concentration at 2 hours in an oral glucose tolerance test in mg/dl. The following distribution resembles a normal distribution with a right tail. As expected, the majority of positive diagnostized cases fall within the high glucose regime.

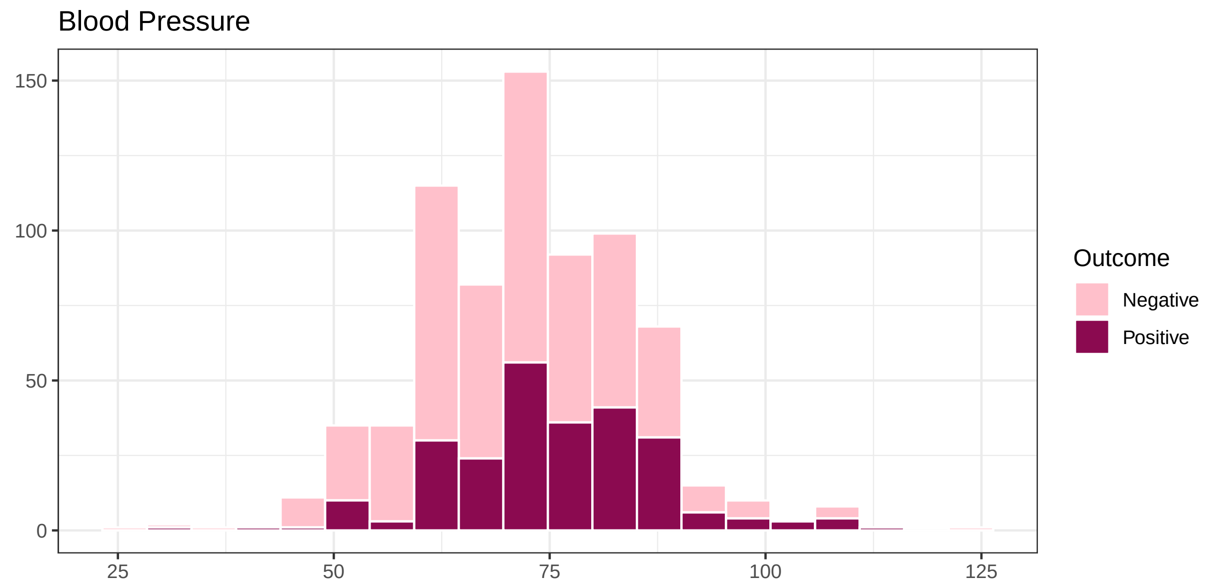


About the distributions, it is interesting to note that the positive tested individuals seem to also exhibit a, somewhat normal, distribution for glucose, although with a higher mode. For reference, we include a table with the mean value of glucose in mg/dl for tested individuals.

Group	Glucose (mg/dl)
All	121.7
Negative	110.6
Positive	142.3

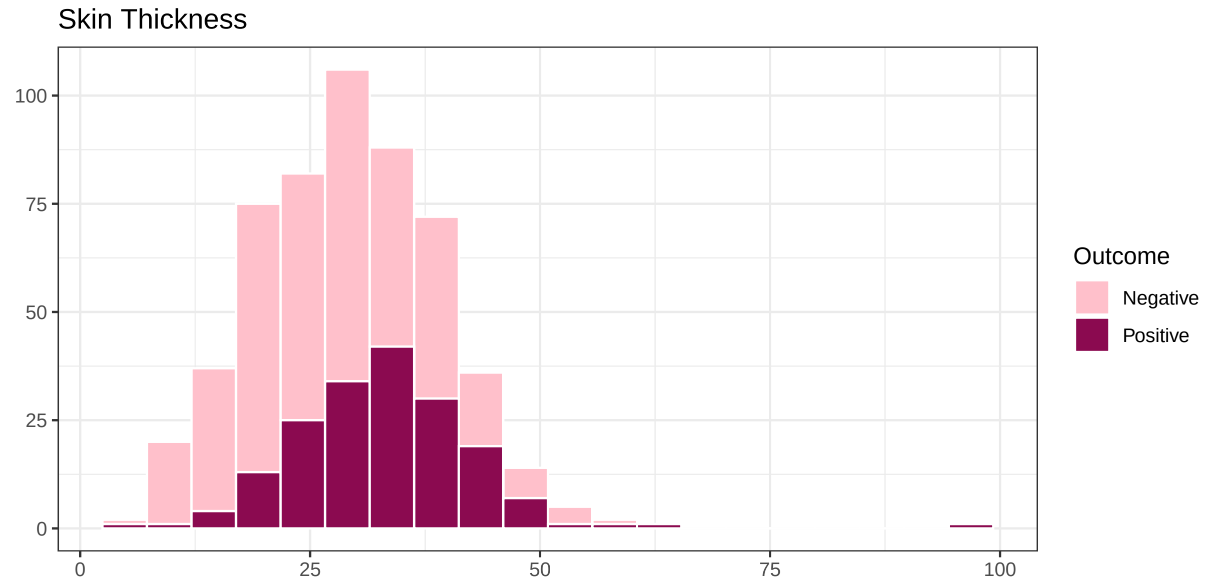
Blood Pressure

Quantitative continuous variable that measures diastolic blood pressure, in mm Hg. Judging with the histogram, we do not observe any significant difference between the two groups.



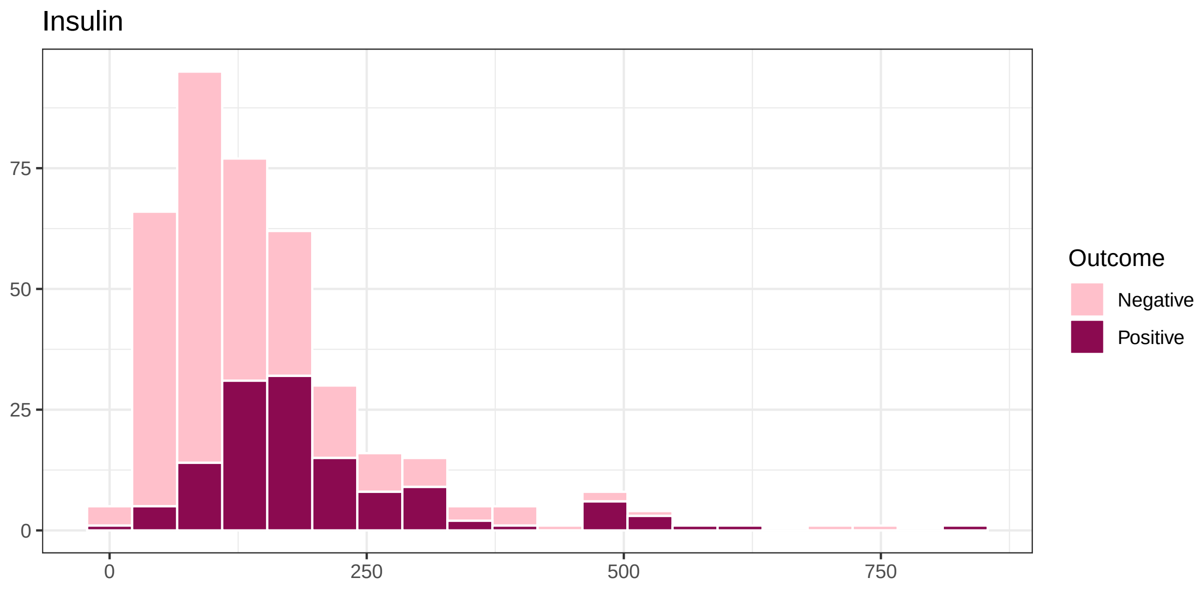
Skin thickness

Quantitative continuous variable that measures the triceps skin fold thickness in mm. For this case we do observe a small shift towards thicker skins for positive tested individuals. However, this shift appears to be small enough so that eyeball techniques are not enough.



Insulin

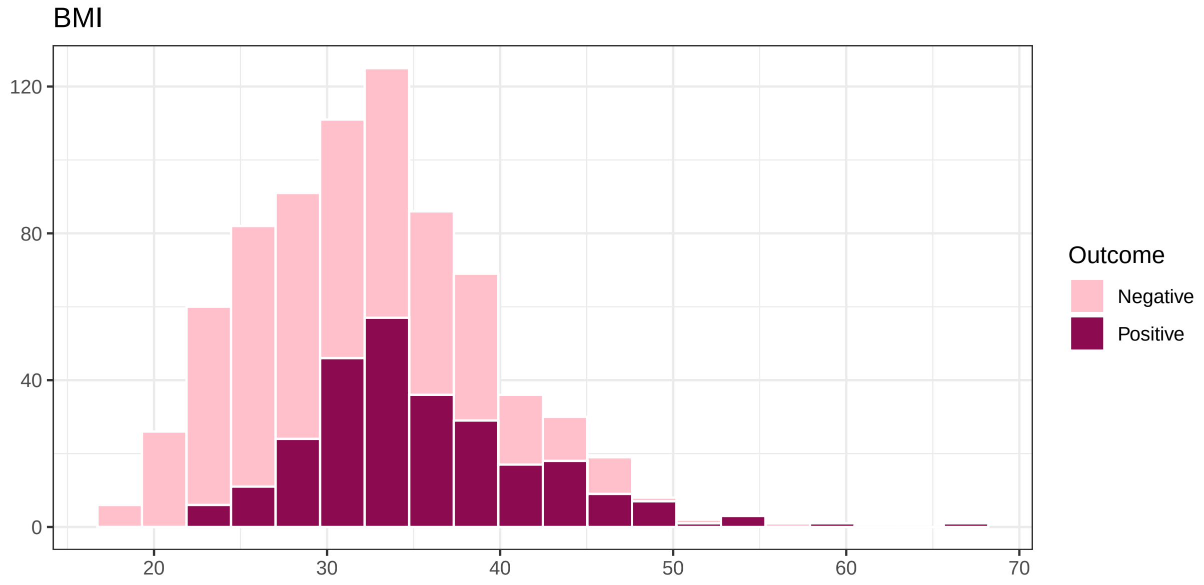
Quantitative continuous variable that measures the 2-hour serum insulin in μU/ml. This units are quantity of insulin matter over volume of the disolution (in the International System of Units it would be measured in mol/m^3), a density of insulin.



The observed distribution has a noticeable right-skewness with a more positive skewness, and more right shifted, histogram for the diabetes groupd We do hereby anticipate a dependency between the 2-Hour serum insulin and whether or not the individual had diabetes.

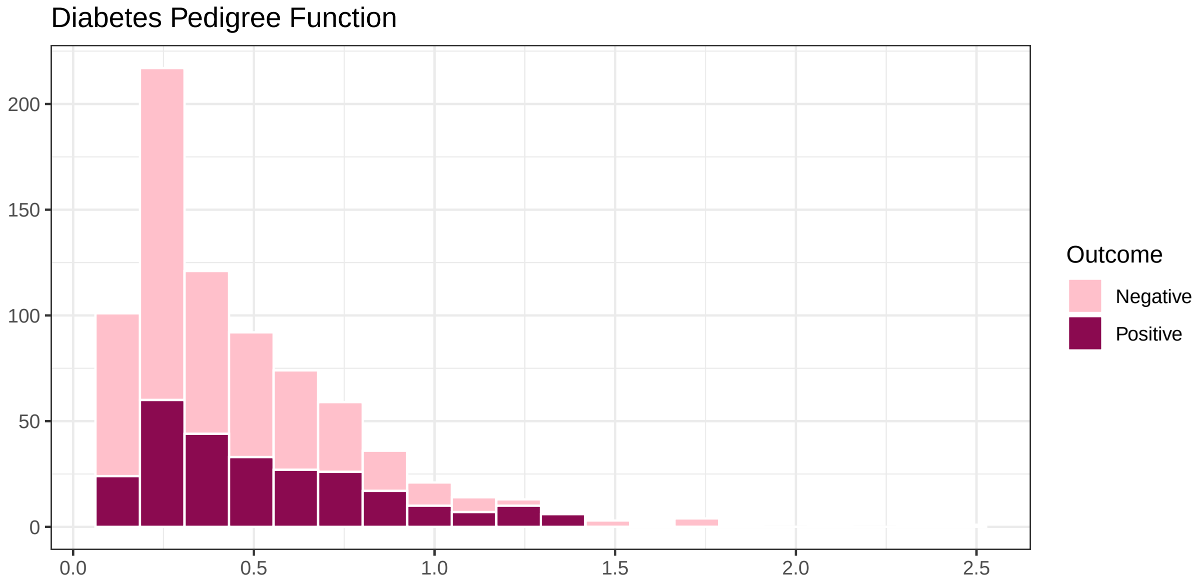
BMI

Quantitative continuous variable that measures the body mass index (BMI), defined as (weight in kg) / (height in m)^2. We do not observe any aparent relation between the body mass index (BMI) and diabetes testing.



Diabetes pedigree function

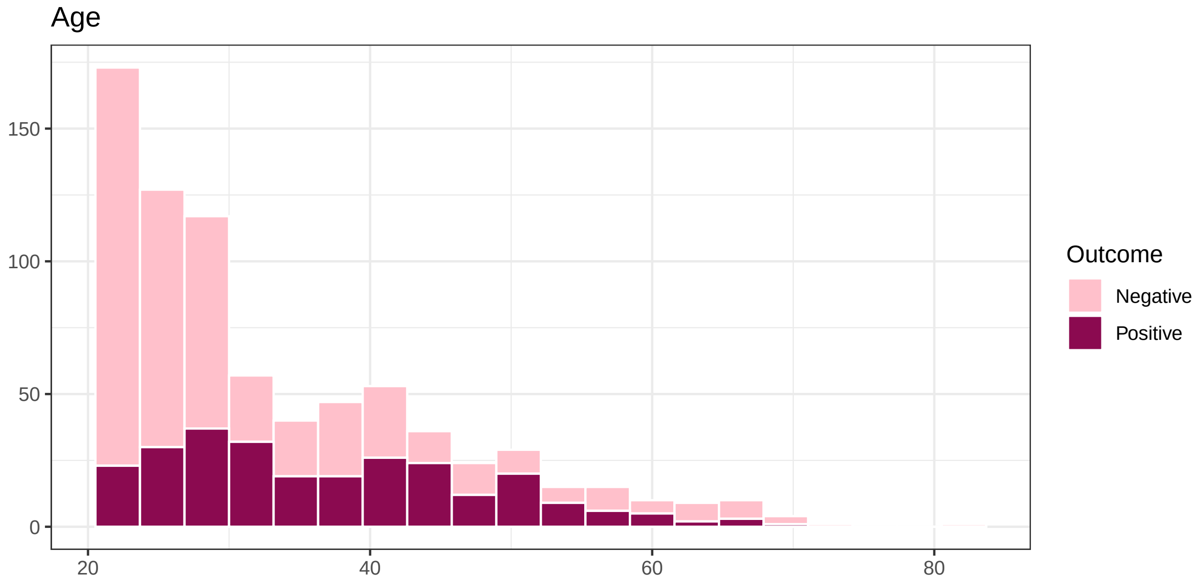
Quantitative continuous variable that indicates the function which scores likelihood of diabetes based on family history. This is an interesting case as although we do not observe *a priori* a dependency between the value of this function and the result of the diabetes test on the individual, it is expected to have relevance as this function was especifically designed to measure the likelihood of having diabetes based on family history.



Age

Quantitative continuous variable that indicates the age of the individual in years. Like with many medical conditions, the older the individual the more likely they are to test positive.

We also note that the majority of tested individuals were young people.



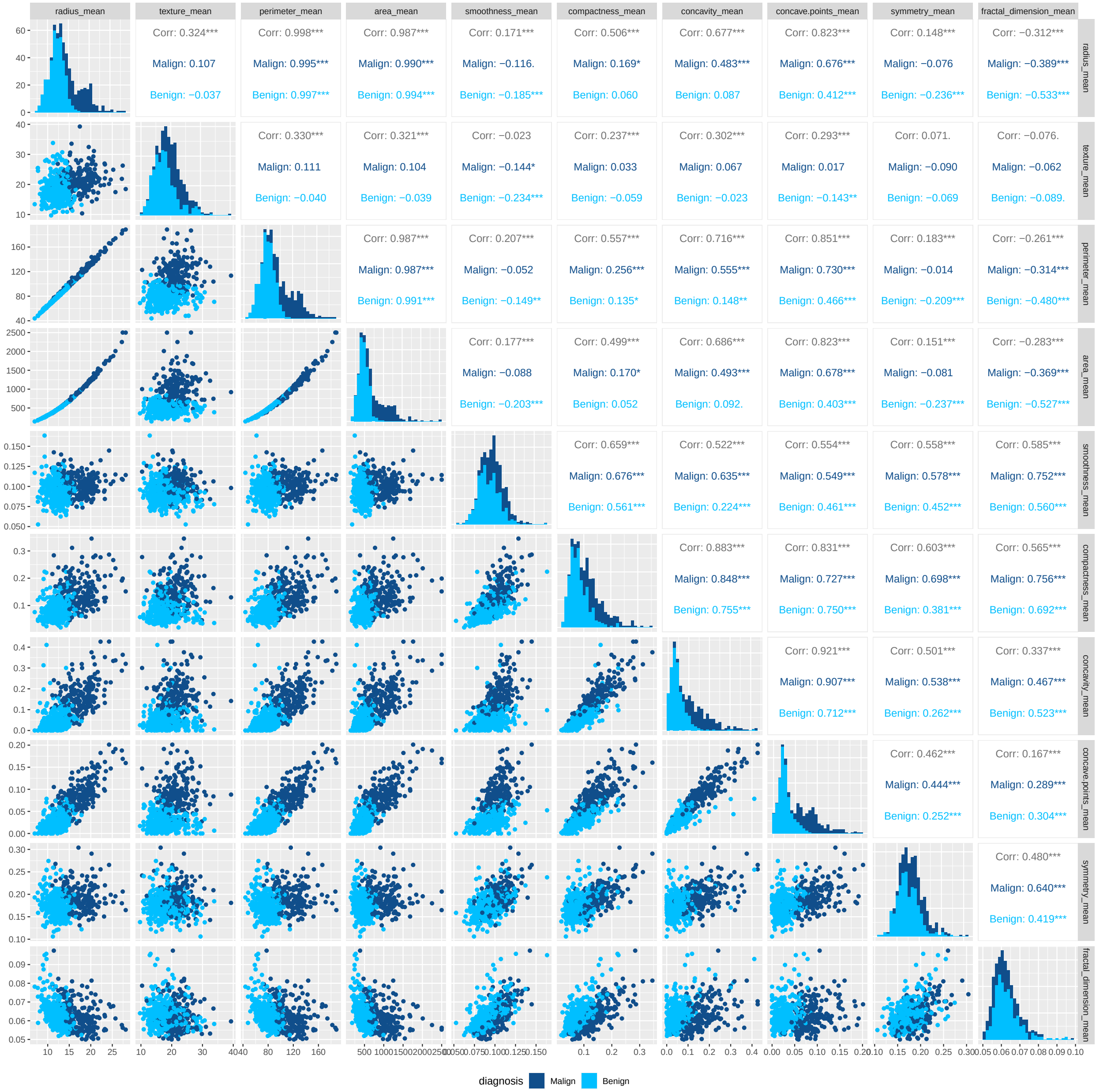
Dataset 2: Breast Cancer

This dataset contains information regarding diagnostic measurements of a digitalized image of a fine needle aspirate of a breast mass. It is comprised of 569 samples and 32 variables. These are an id, a diagnosis variable of the tissue and 10 sets of 3 variables. Each of these sets represent a variable describing the mass and has fields:

- Mean value.
- Standard error.
- Mean of the 3 most negative cases.

Each of them are qualitative continous.

The dataset can be found in https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data. As before we do a scatter plot matrix to check for correlations.



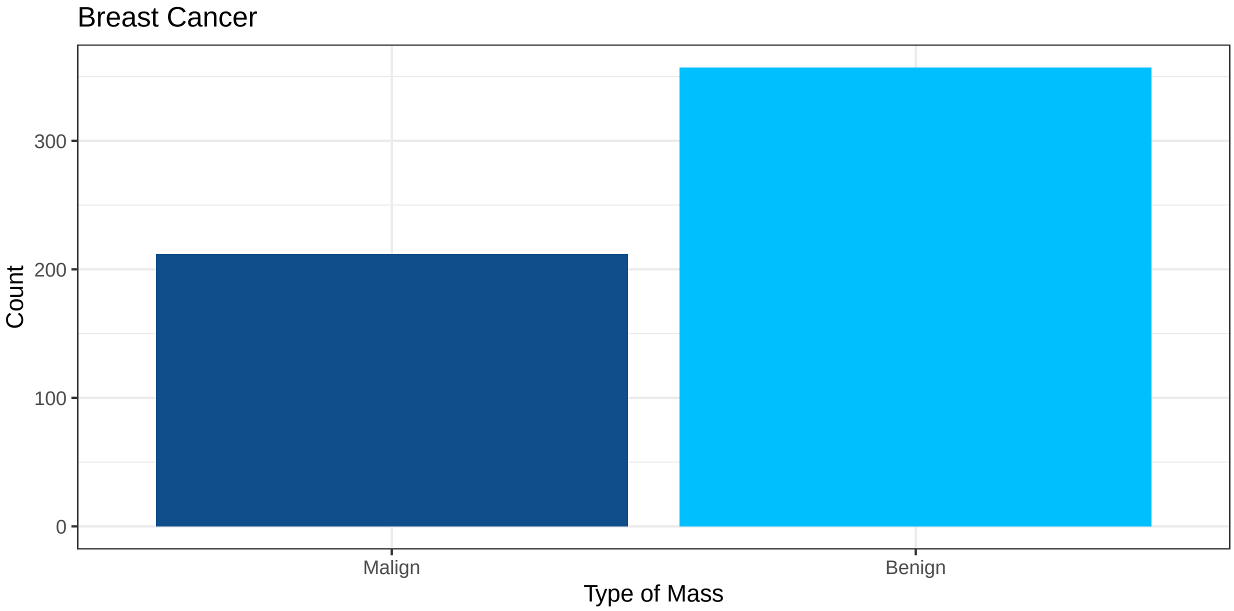
We observe a perfect correlation for the radius, perimeter and area. This is expected as from geometry

$$P = 2\pi R\ , \qquad A = \pi R^2\ .$$

Hence we may ignore the perimeter and the area as the information will be carried by the radius variable. Note that the compactness although defined from the perimeter and the area, it is done in a way such that the radius dependency is cancelled and is independent again.

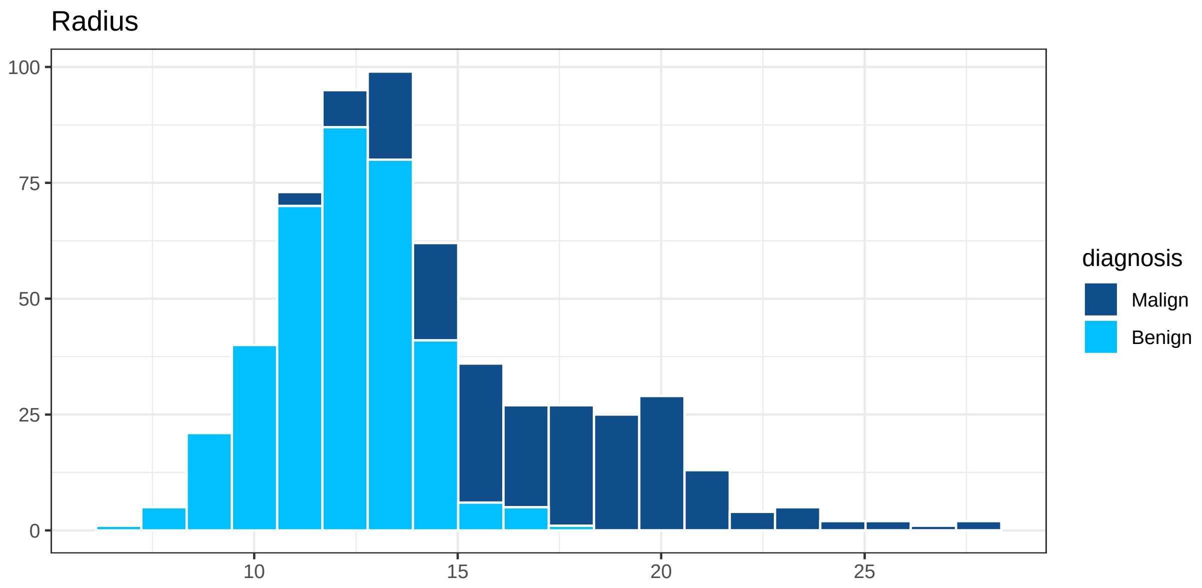
diagnosis

The diagnosis of breast tissues. This is a qualitative nominal variable with values *Benign* and *Malign*.



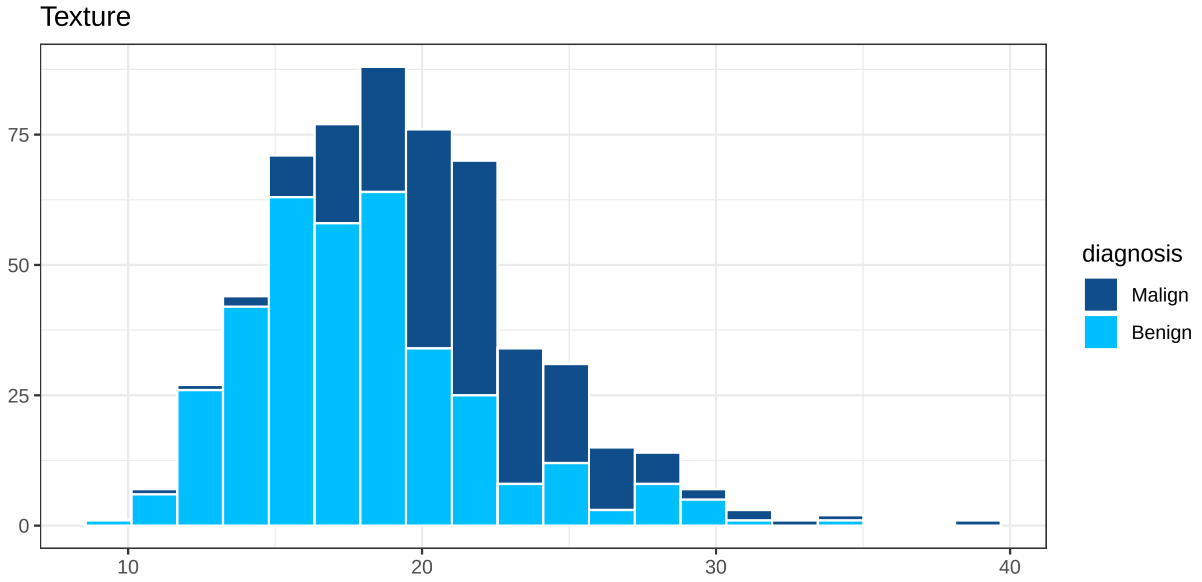
radius

Mean of the distances from center to points on the perimeter in mm. We observe that people with malign masses have a higher probability of having bigger radius.



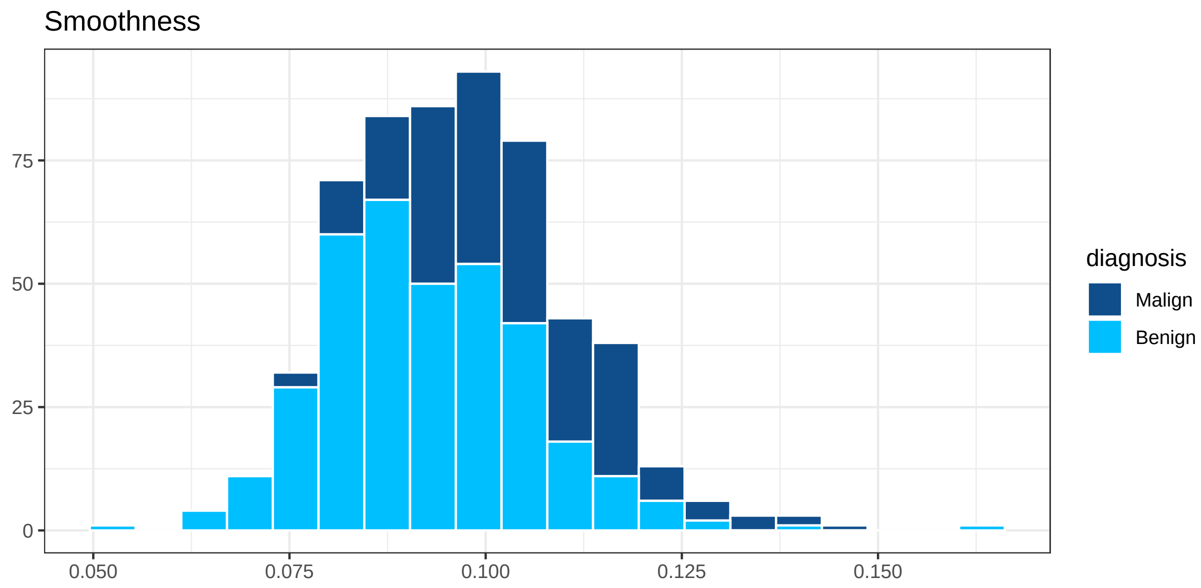
texture

Standard deviation of gray-scale values. There seems to be a shift to the right for people with malign masses.



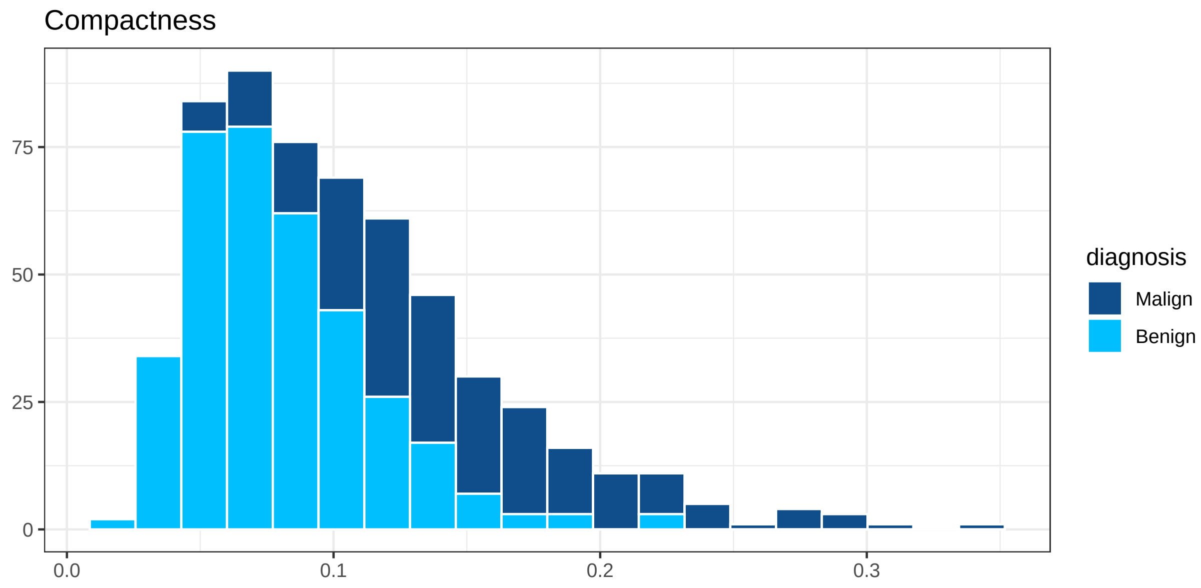
smoothness

Local variation in radius lengths. This represents how abruptly the mass is noted in the image. From eyeballing the histogram we cannot infer a clear relation.


Smoothness

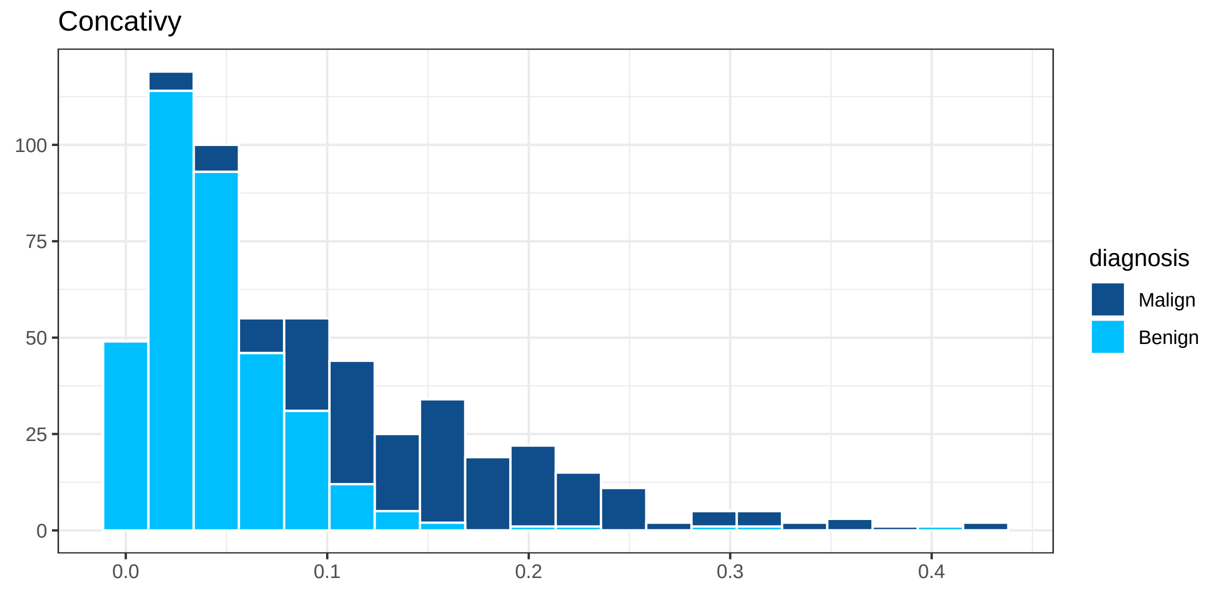
compactness

Measures how compact the mass is. It is defined as a ratio of the perimeter squared over the area. This way it represents the amount of border per surface. Like with other variables, for people with malign masses it exhibits a more right-skewed distribution.


Compactness

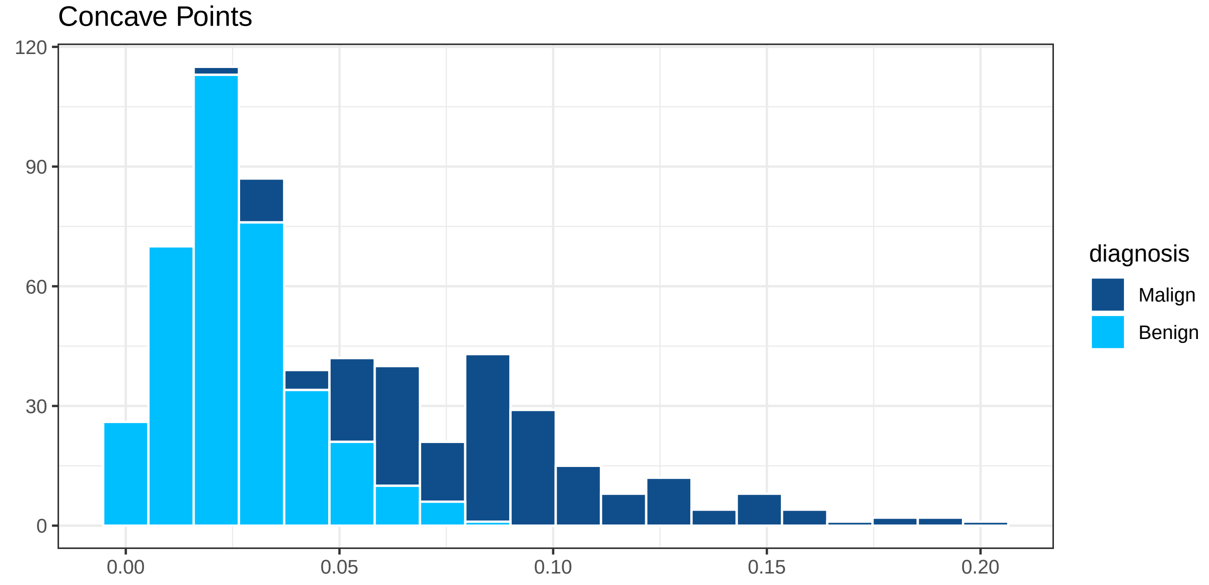
concavity

Severity of concave portions of the contour. One again we observe the right-skewness effect.



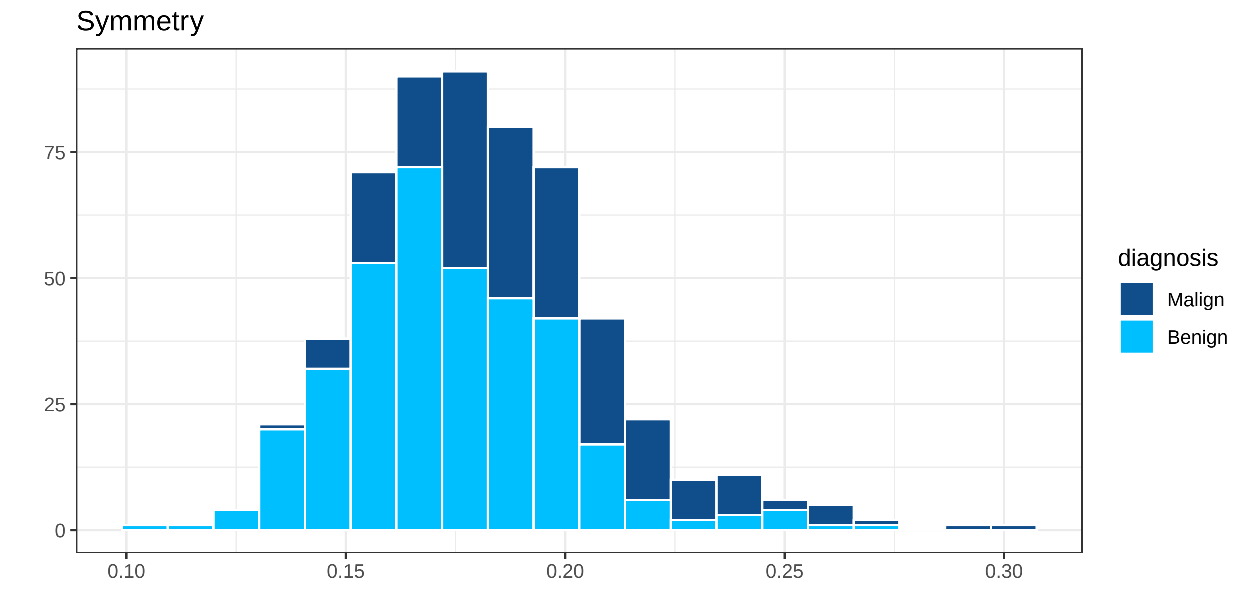
concave_points

Number of concave portions of the contour. One again we observe the right-skewness effect.



symmetry

Eccentricity of the mass. Measures how similar it is to a circle with 0 an exact circle. Always less than 1. From the histogram it does not seem to have a big effect onto the nature of the mass.

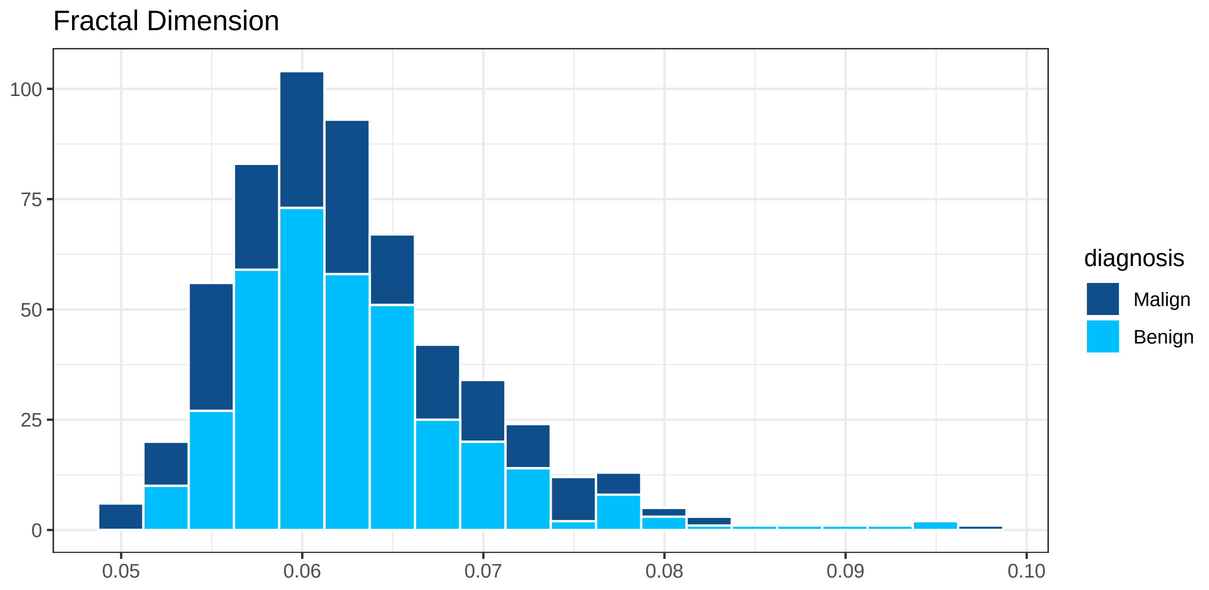


fractal_dimension

Measure of the complexity of a pattern. Mathematically,

$$D = -\frac{\log N}{\log \varepsilon} \ ,$$

where if we imagine a lattice of points N is the number of them inside the mass and ε the distance between them in lattice units. It is again related to the geometry although this time the relation with the kind of mass is not visible at plain sight.


Fractal Dimension
diagnosis
Malign
Benign