Analysis of diabetes dataset

48165 Dominik Karaś
48173 Julia Giszczak

Laboratory of Data Analysis
Nome do Professor: Fátima Leal

Data: 15/04/2023

DEPARTAMENTO CIÊNCIA
E TECNOLOGIA

CONTENT

It is missing the Introduction.

Description of dataset’s characteristics

The analysis concerns the phenomenon of diabetes and factors that may cause it.

The domain of our project is healthcare. Data comes from Kaggle website. The study included women from 21 years of age of Pima Indian heritage. Dataset contains 9 variables and each of them has 768 entries. All variables are quantitative

caption

Variables	Description	Variable type
Pregnancies	Number of times pregnant	Discrete (int)
Glucose	Glucose level in blood	Discrete (int)
Blood pressure	Blood pressure measurement (mm Hg)	Discrete (int)
Skin thickness	Thickness of the skin (mm)	Discrete (int)
Insulin	Insulin level in blood (mu U/ml)	Discrete (int)
BMI	Body mass index (weight in kg / (height in m)^2)	Continuous (float)
DiabetesPedigreeFunction	Likelihood of diabetes based on family history	Continuous (float)
Age	Age (years)	Discrete (int)
Outcome	Final result (0 if no, 1 if yes)	Discrete (int)

The cleaning of data was not necessary, because there are no missing values in any of the columns. However, some of the variables such as: glucose, blood pressure, skin thickness, insulin, BMI, age and diabetes pedigree function contained values equal to zero, which is biologically impossible. These values have been replaced with the average of its column. Values equal to zero for such columns as pregnancies and outcome remained unchanged.

Statistical analysis

Data description

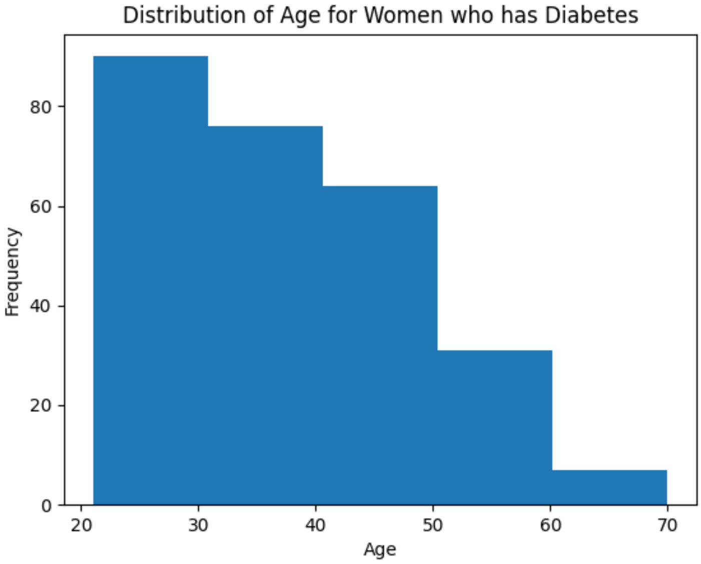
caption

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
Count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
Mean	3.845052	121.686763	72.405184	29.153420	155.548223	32.457464	0.471876	33.240885	0.348958
Std	3.369578	30.435949	12.096346	8.790942	85.021108	6.875151	0.331329	11.760232	0.476951
Min	0.000000	44.000000	24.000000	7.000000	14.000000	18.200000	0.078000	21.000000	0.000000
25%	1.000000	99.750000	64.000000	25.000000	121.500000	27.500000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.202592	29.153420	155.548223	32.400000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	155.548223	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

The dataset description confirms that each variable contains 768 values. As we can see, women included in the study were between 21 and 81 years old and the average age was 33.

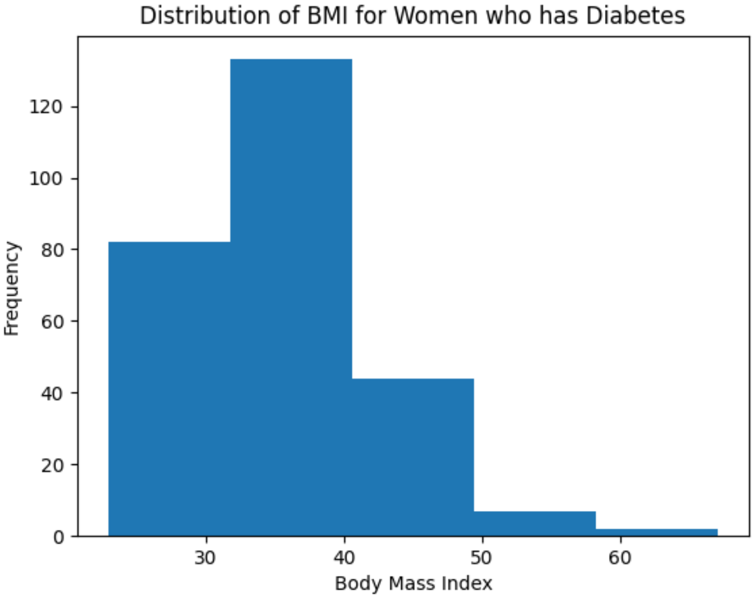
It is other section: Graph analysis

Using histogram, we can check the distribution of Age for women who has diabetes.



caption

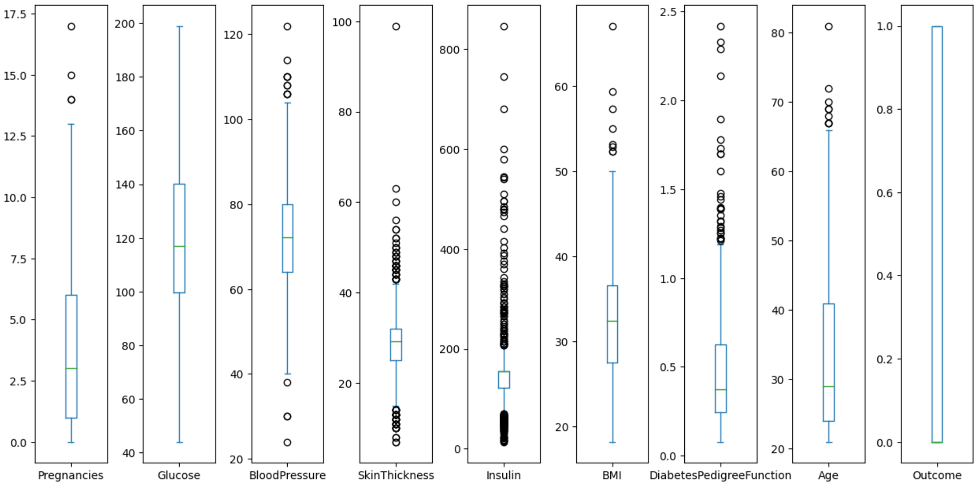
As we can see from the histogram, the largest number of women with diabetes were women aged from 21 to 30. What is interesting, with increasing age, the number of women with diabetes decreased.



Among the respondents, the largest number of women suffering from diabetes had a BMI in the range of 30-40, which means obesity of the 1st degree.

Boxplots for variables

Boxplot displays the five-number summary of each variable (minimum, first quartile, median, third quartile and maximum). We can also see if there are any outliers.



caption

Covariances

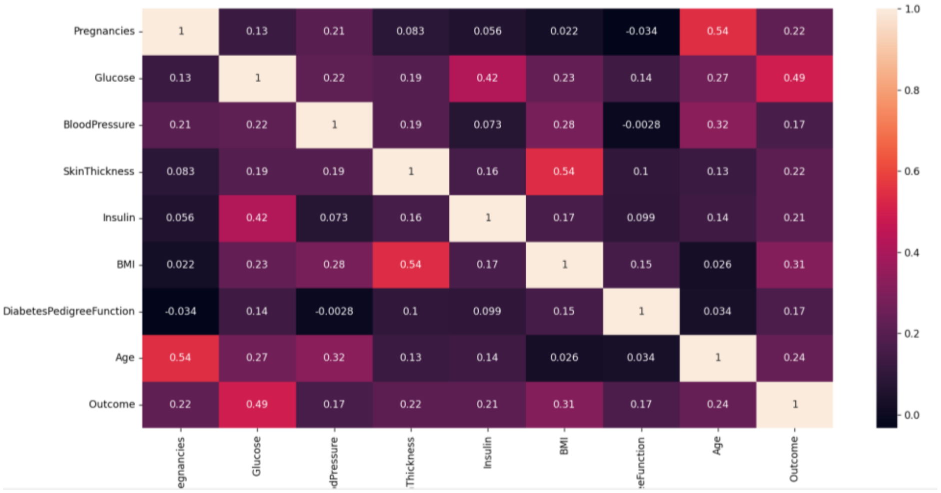
caption

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
Pregnancies	11.354056	13.118128	8.499282	2.458283	16.050914	0.499584	-0.037426	21.570620	0.356618
Glucose	13.118128	926.346983	80.394788	51.636823	1087.239699	48.324859	1.382151	95.401356	7.155569
Blood Pressure	8.499282	80.394788	146.321591	20.503705	74.579607	23.391407	-0.011075	46.175523	0.958140
Skin Thickness	2.458283	51.636823	20.503705	77.280660	118.195534	32.782007	0.294084	13.219905	0.902718
Insulin	16.050914	1087.239699	74.579607	118.195534	7228.588766	97.375072	2.778511	136.715802	8.694564
BMI	0.499584	48.324859	23.391407	32.782007	97.375072	47.267706	0.349435	2.063312	1.022835
Diabetes Pedigree Function	-0.037426	1.382151	-0.011075	0.294084	2.778511	0.349435	0.109779	0.130772	0.027472
Age	21.570620	95.401356	46.175523	13.219905	136.715802	2.063312	0.130772	138.303046	1.336953
Outcome	0.356618	7.155569	0.958140	0.902718	8.694564	1.022835	0.027472	1.336953	0.227483

Covariance is a statistical measure that shows whether two variables are related. Positive covariance means that both variables either increase or decrease, while negative value of this measure means that values of the variables change in opposite directions. In our dataset, positive covariance occurrs between for example blood pressure and glucose, which means that as the level of glucose in blood rises, the blood pressure also rises.

Correlations

Next step is correlation to determine the strength of a relationship between variables.



According to the heatmap generated for all variables, the highest, positive correlation occurred between:

- Age and pregnancies (0,54)
- BMI and skin thickness (0,54)
- Glucose and outcome (0,49)

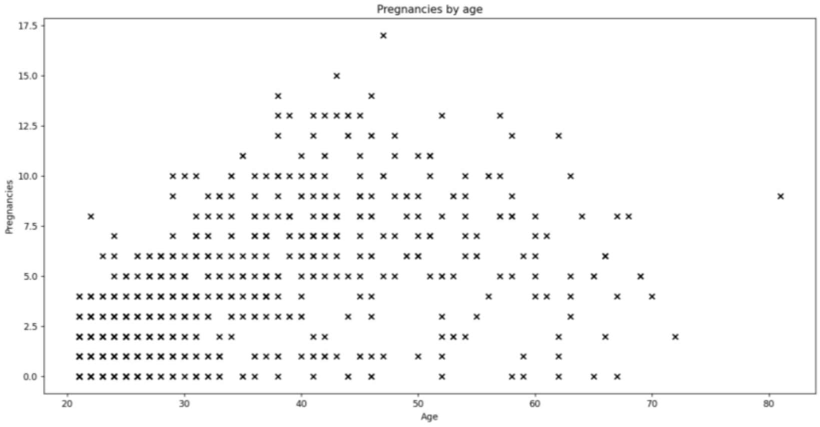
The correlation between age and pregnancies is logic. As women get older, their fertility declines and the likelihood of having a successful pregnancy decreases. Therefore, women who are older may have a higher number of pregnancies because they have been trying to conceive for a longer period of time.

The positive correlation between BMI and skin thickness means that people who have higher BMI tend to have thicker skin and people who weigh less are more likely to have thinner skin.

The positive correlation between Glucose and outcome shows that people with higher glucose level in blood are more likely to be diagnosed with diabetes. However it doesn't mean that increase in glucose level cause diabetes and the other way round.

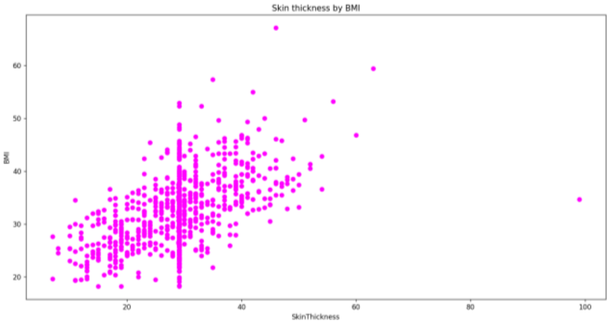
GRAPHS

1. PREGNANCIES BY AGE



As we can see from the graph shown above, relationship between pregnancies and age is linear. It is also positive, because the number of pregnancies rise along with rise of age.

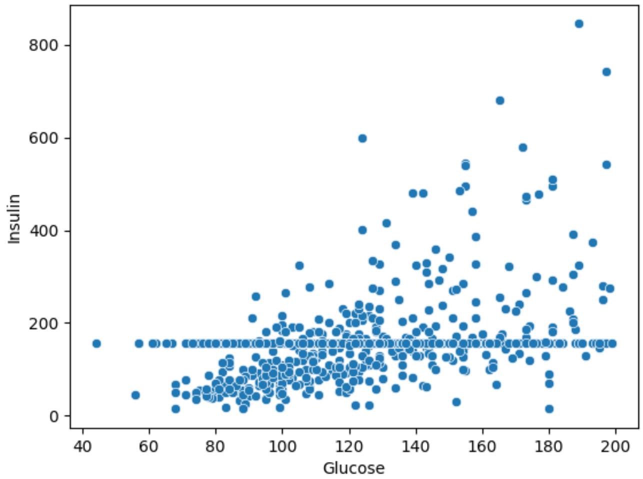
2. SKIN THICKNESS BY BMI



The shape of the scatterplot demonstrates that the relationship between skin thickness and BMI also assumes a positive linear relationship. We can also see outliers, one of which shows that skin thickness equals 99 mm with relatively low BMI.

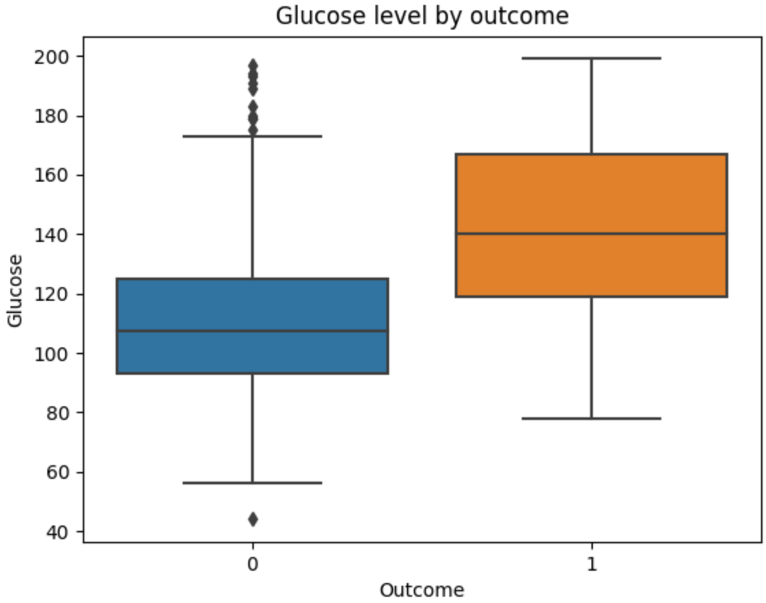
The sequence of points on the average level visible on the chart is caused by the replacement of zero values with the average value

3. GLUCOSE BY INSULIN



A positive relationship can be seen from the chart. However, it is much lower than relation between skin thickness and BMI which means that increase in glucose does relatively little change in insulin levels. The scatterplot contains some outliers, which may distort the result.

4. GLUCOSE LEVEL BY OUTCOME



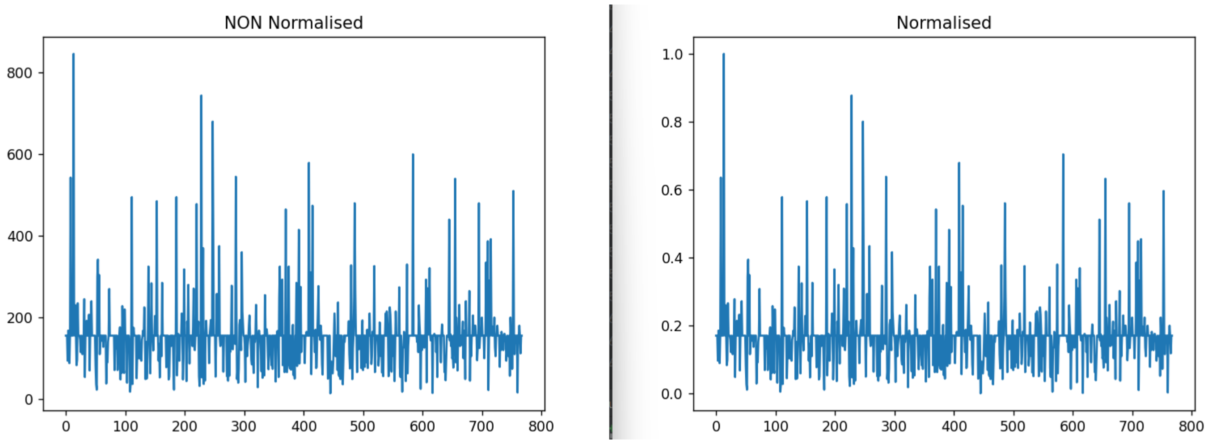
The median of glucose level for people with diabetes was higher than for healthy people. However we can see that in the case of healthy people there is significant number of outliers.

DATA TRANSFORMATION

5. NORMALIZATION

Normalization scales features between 0 and 1, retaining their proportional range to each other.

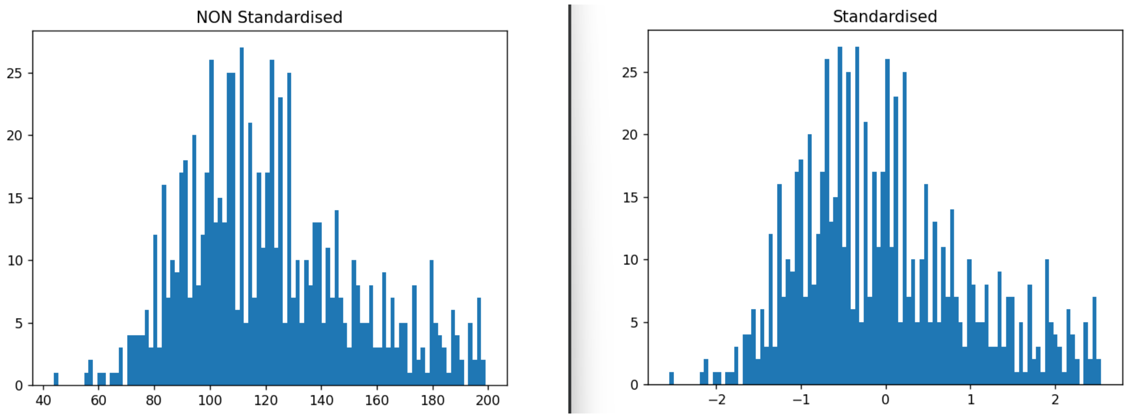
Example of normalization for insulin variable:



6. STANDARDIZATION

Standardization scales features to have a mean of 0 and standard deviation of 1.

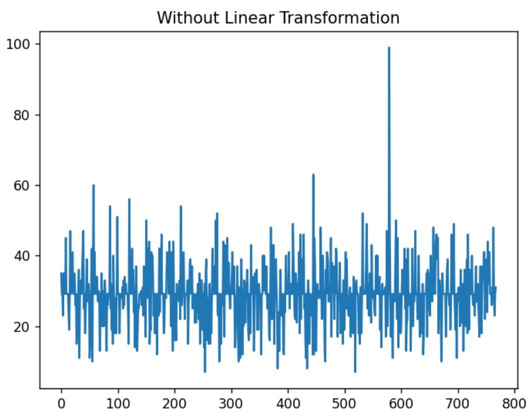
Example of standardization for glucose variable:

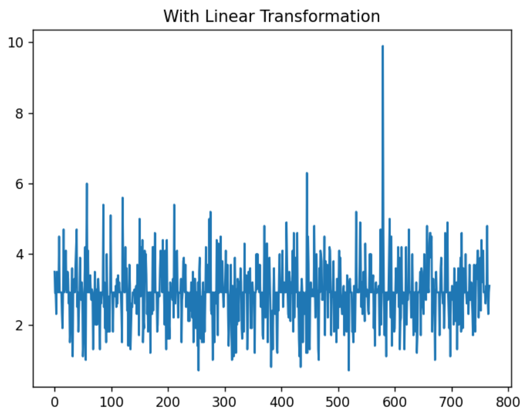


7. LINEAR TRANSFORMATION

Linear transformation is a function which changes the original value into new variable.

For example, linear transformation can be used to convert skin thickness unit from mm to cm.





In general, the analysis are good. You need to improve the reports. All figures and table should have a caption which need to be mentionated in the text.

The reports start with an Introduction which introduces the domain and the work to be done. The conclusion should summurise the results as well as the future work.

UNIVERSIDADE PORTUCALENSE
DO CONHECIMENTO À PRÁTICA