1. Description of dataset’s characteristics

The analysis concerns the phenomenon of diabetes and its causes. The goal of the analysis is to check which factors have the greatest impact on the risk of developing diabetes.

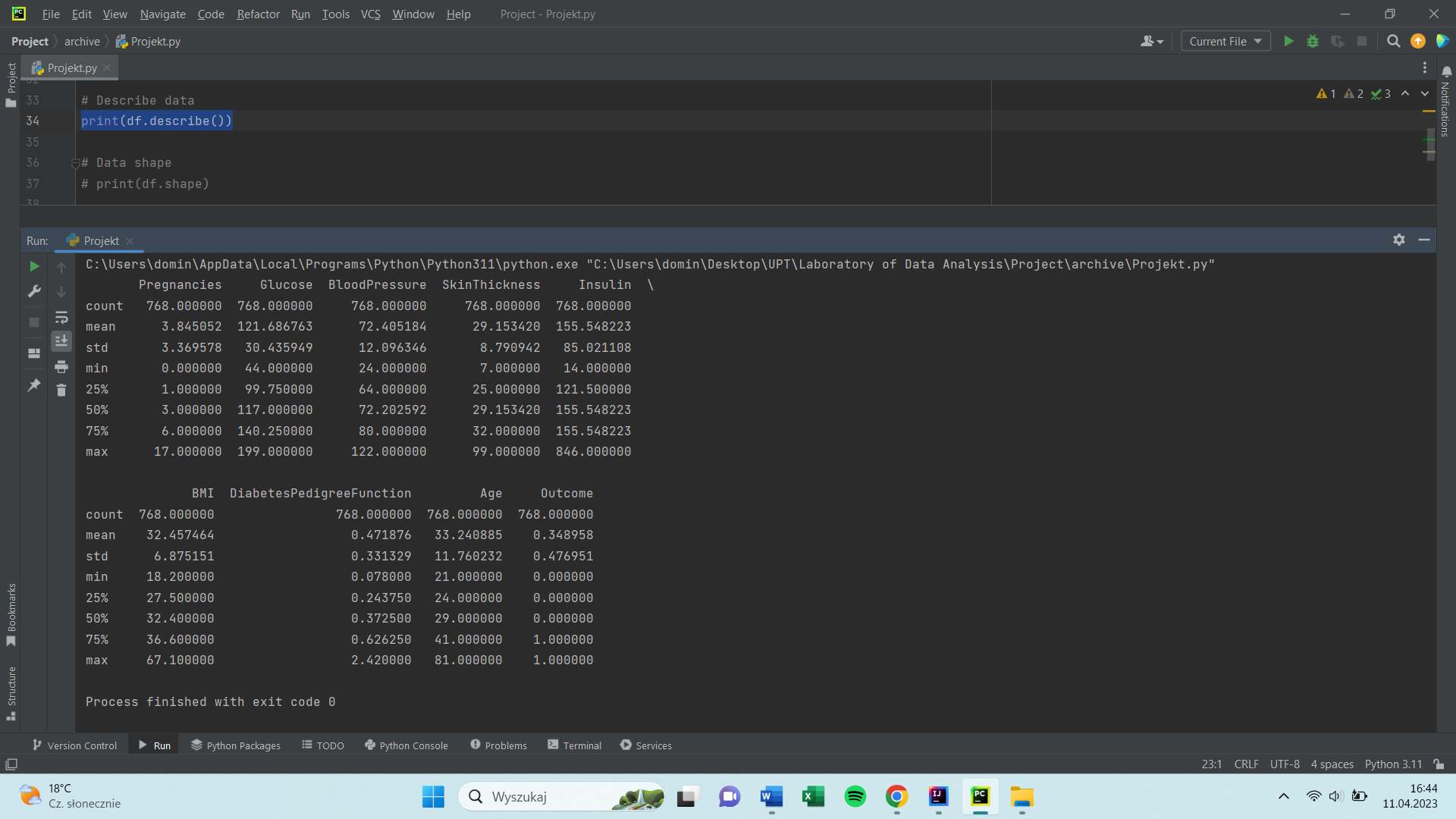
The domain of our project is healthcare. The source of the dataset is Kaggle. Dataset contains 9 variables and each of them has 768 entries. All variables are quantitative.

|  |  |  |
| --- | --- | --- |
| 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 values missing 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.

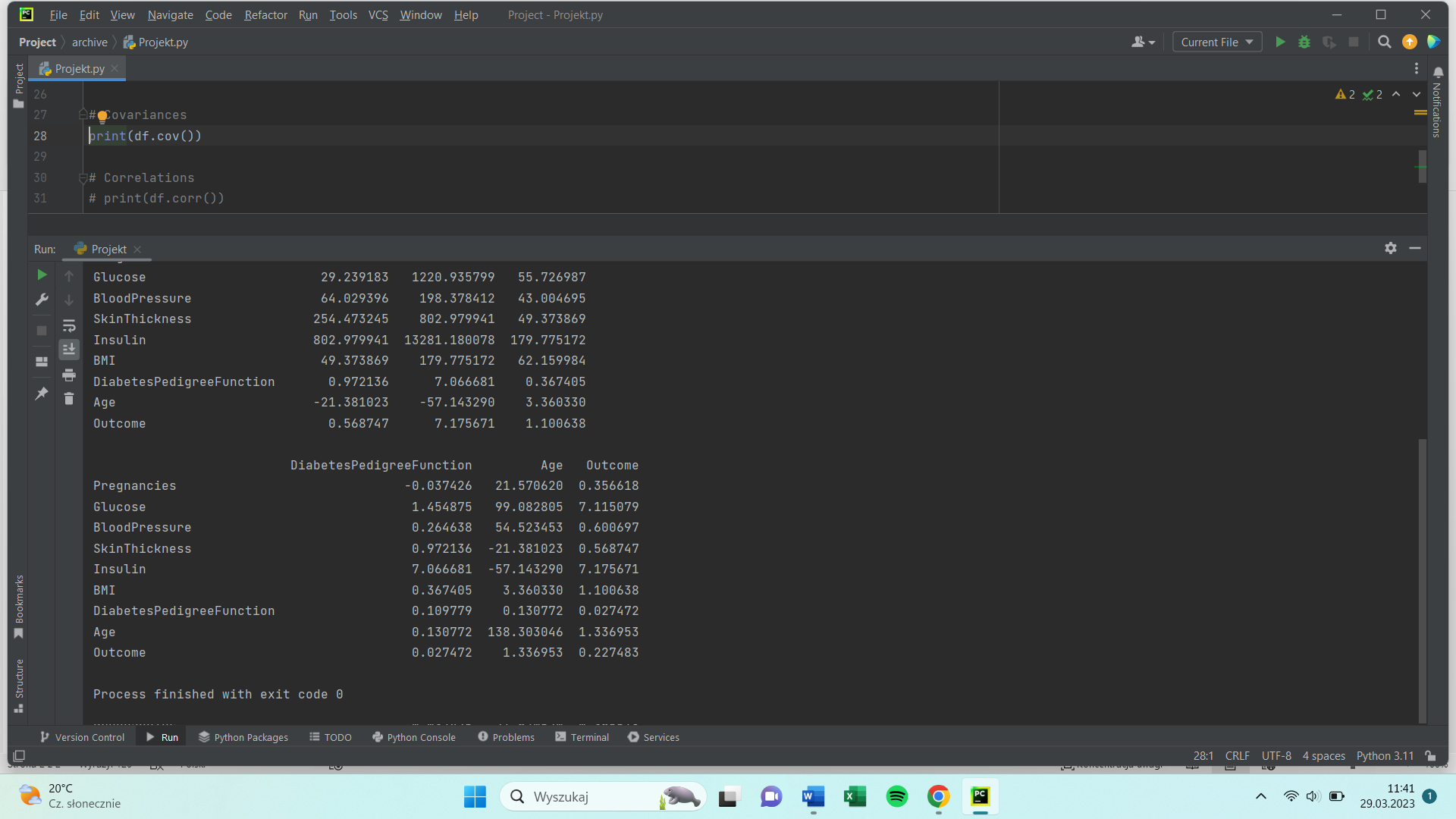
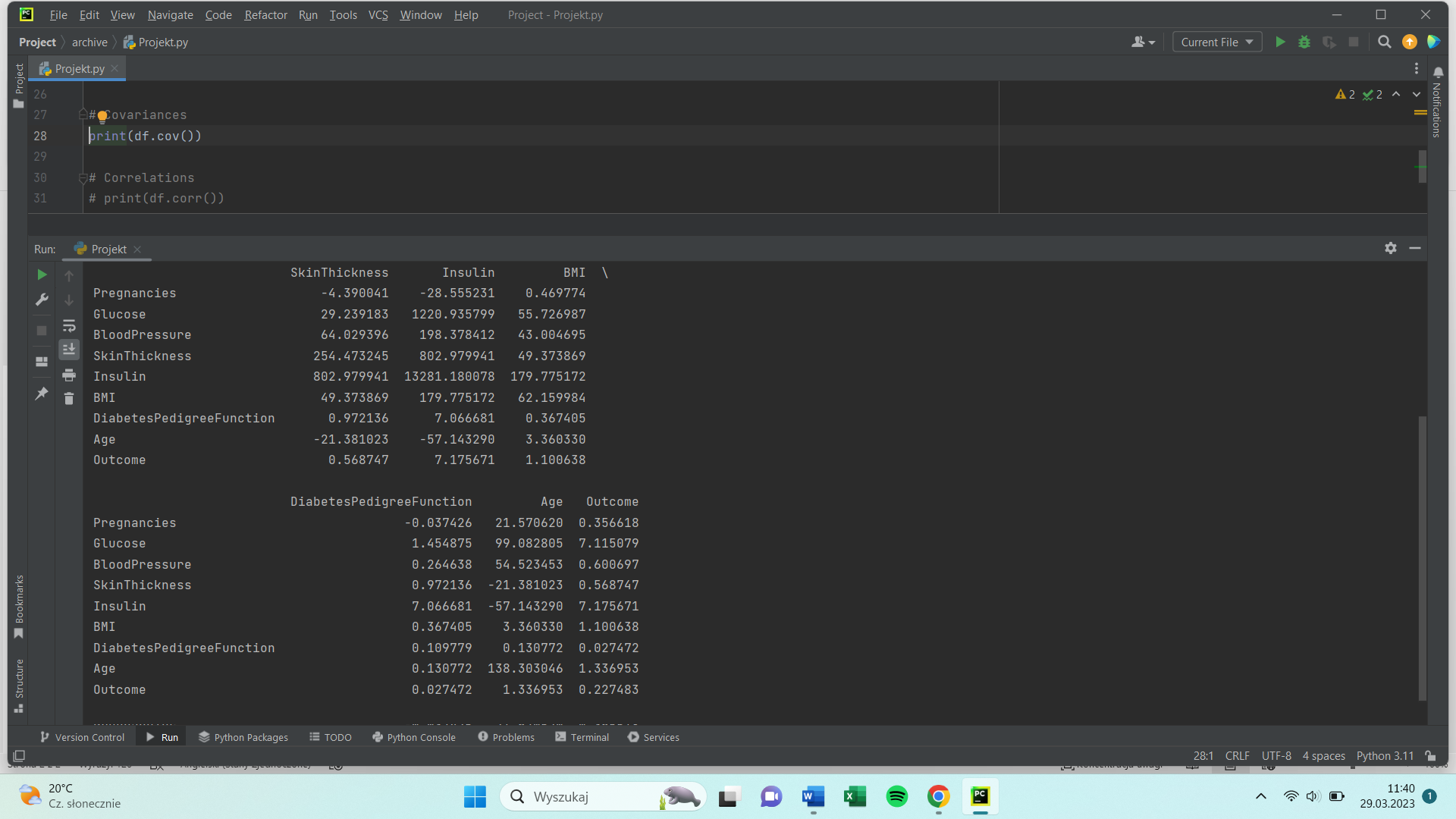
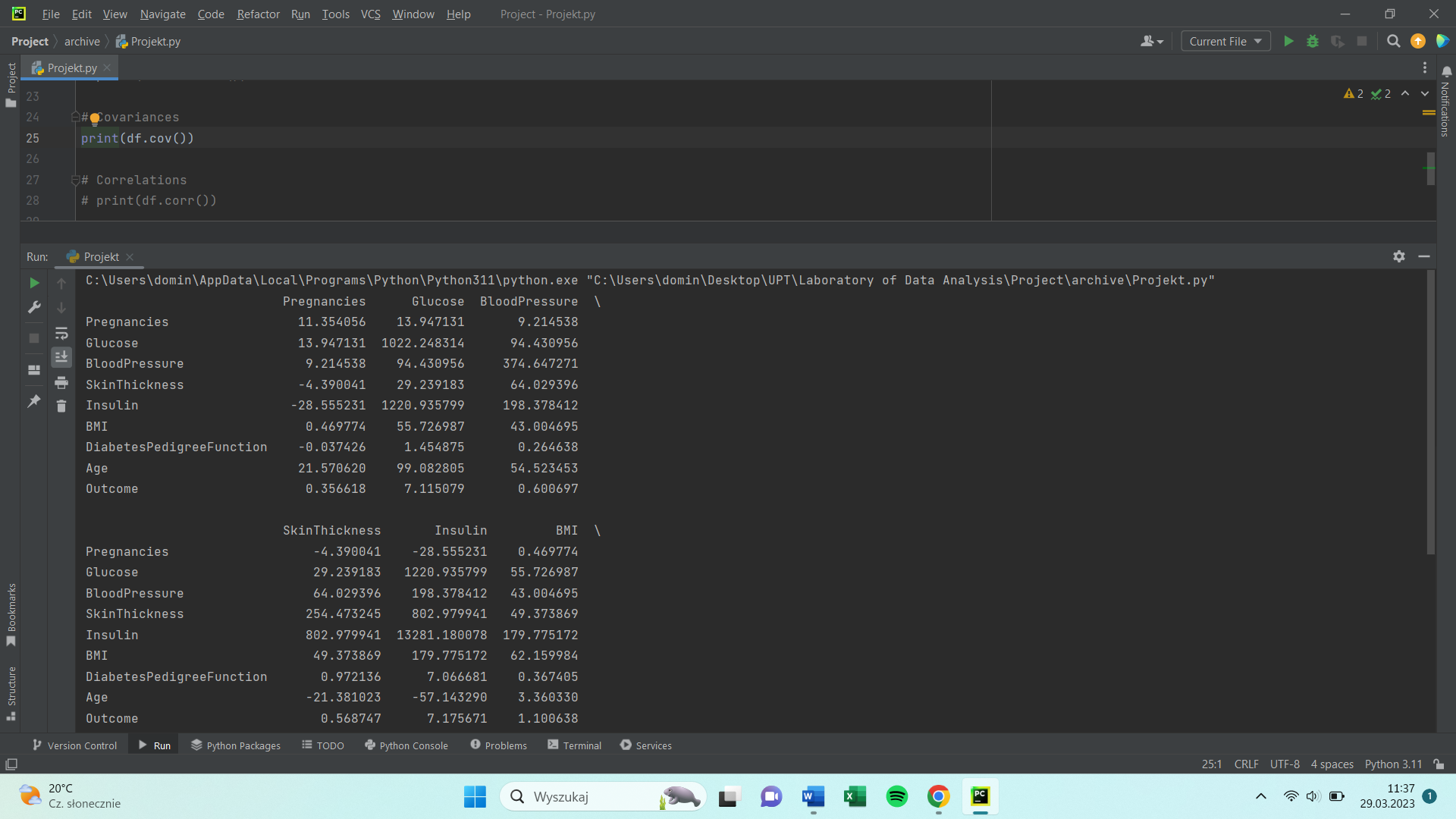
**Statistical analysis**

Data description



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.

Covariances



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.

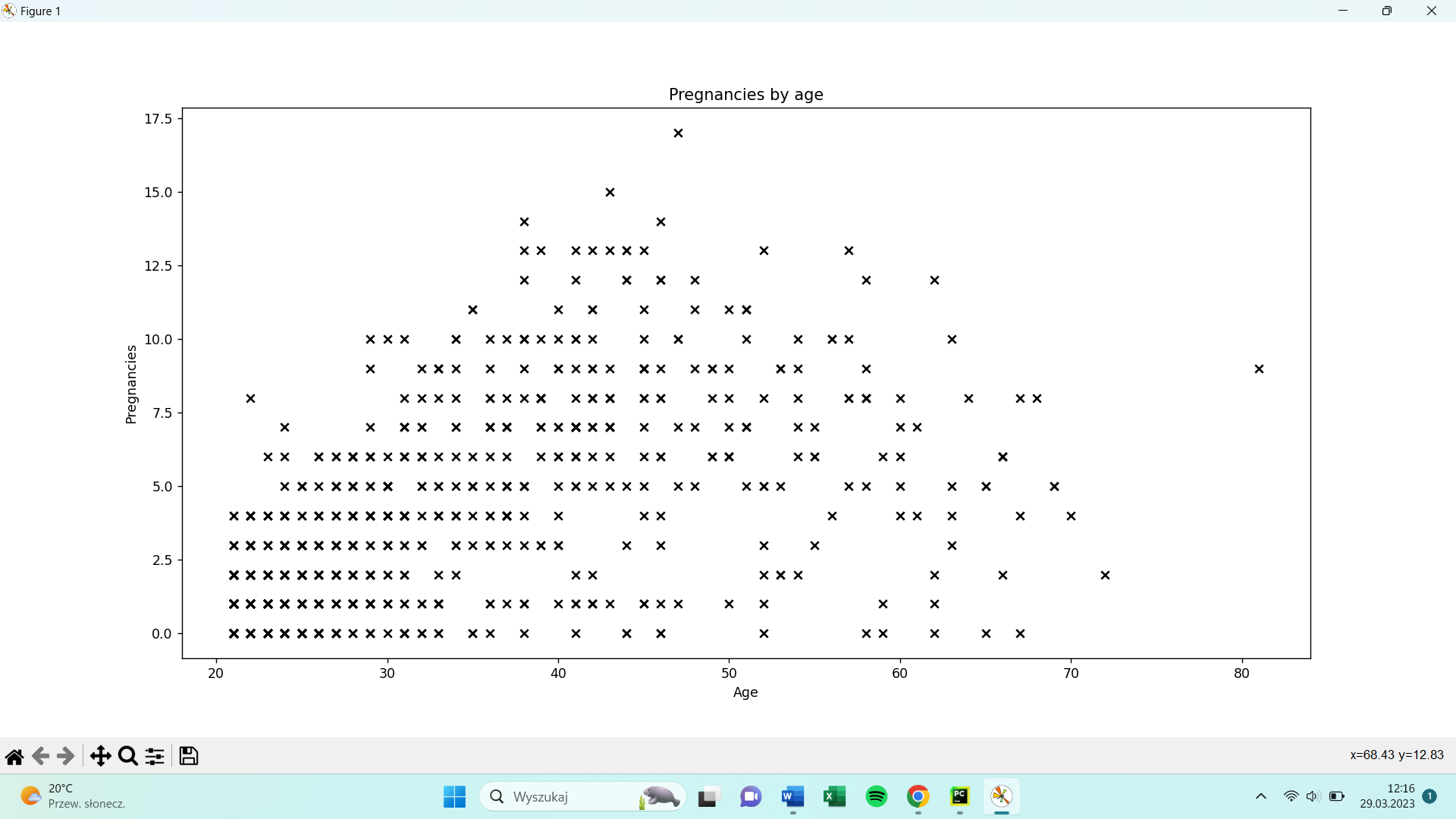


According to the heat map generated for all variables, the highest correlation occurred between:

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

Graphs

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.

Skin thickness by BMI

Obraz zawierający wykres

Opis wygenerowany automatycznie

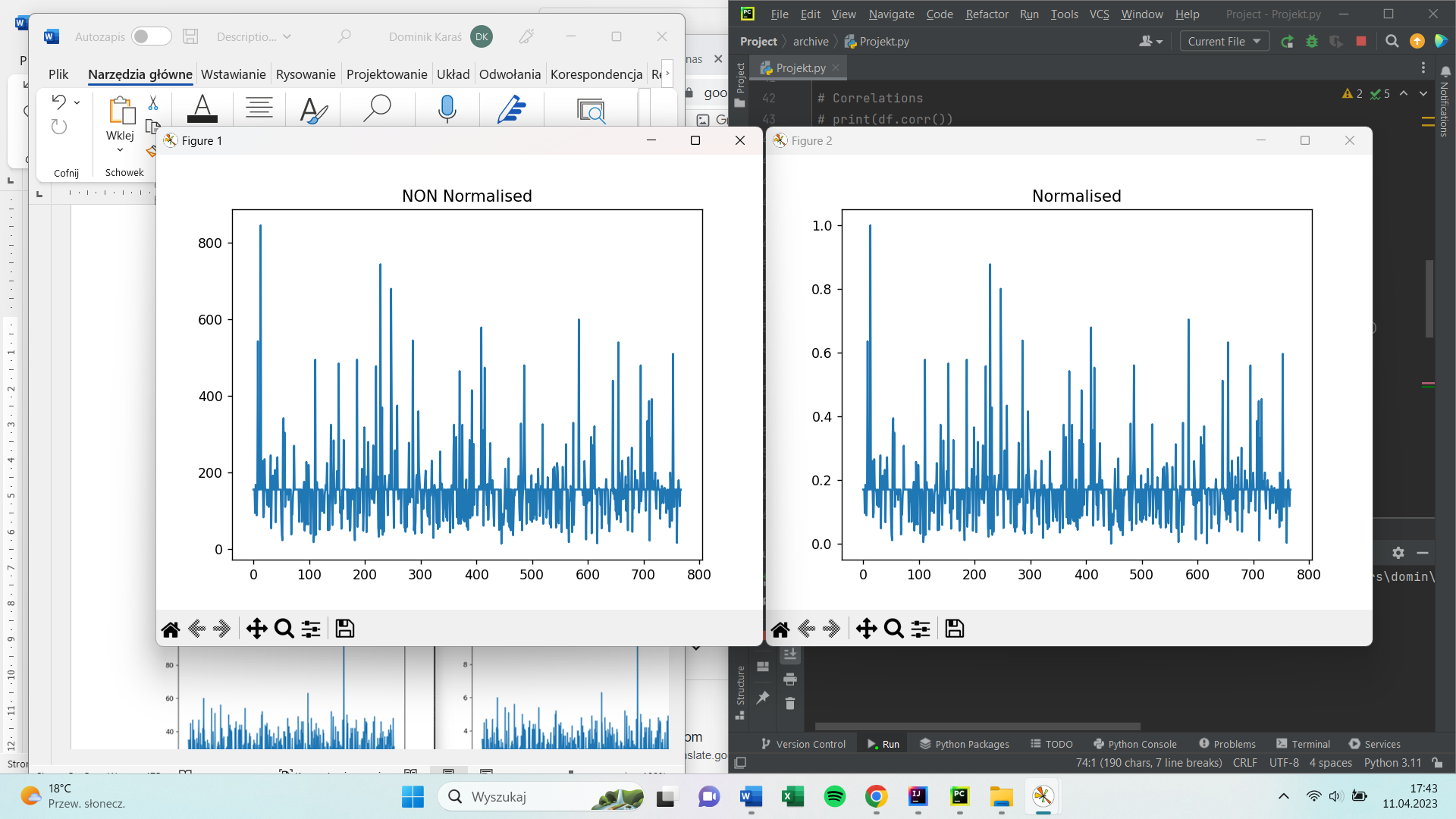
The relationship between skin thickness and BMI is the same as the relationship between pregnancies and age. We can also see one outlier, where skin thickness equals 99 mm.

Data transformation

Normalisation

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

Example of normalisation for insulin variable:



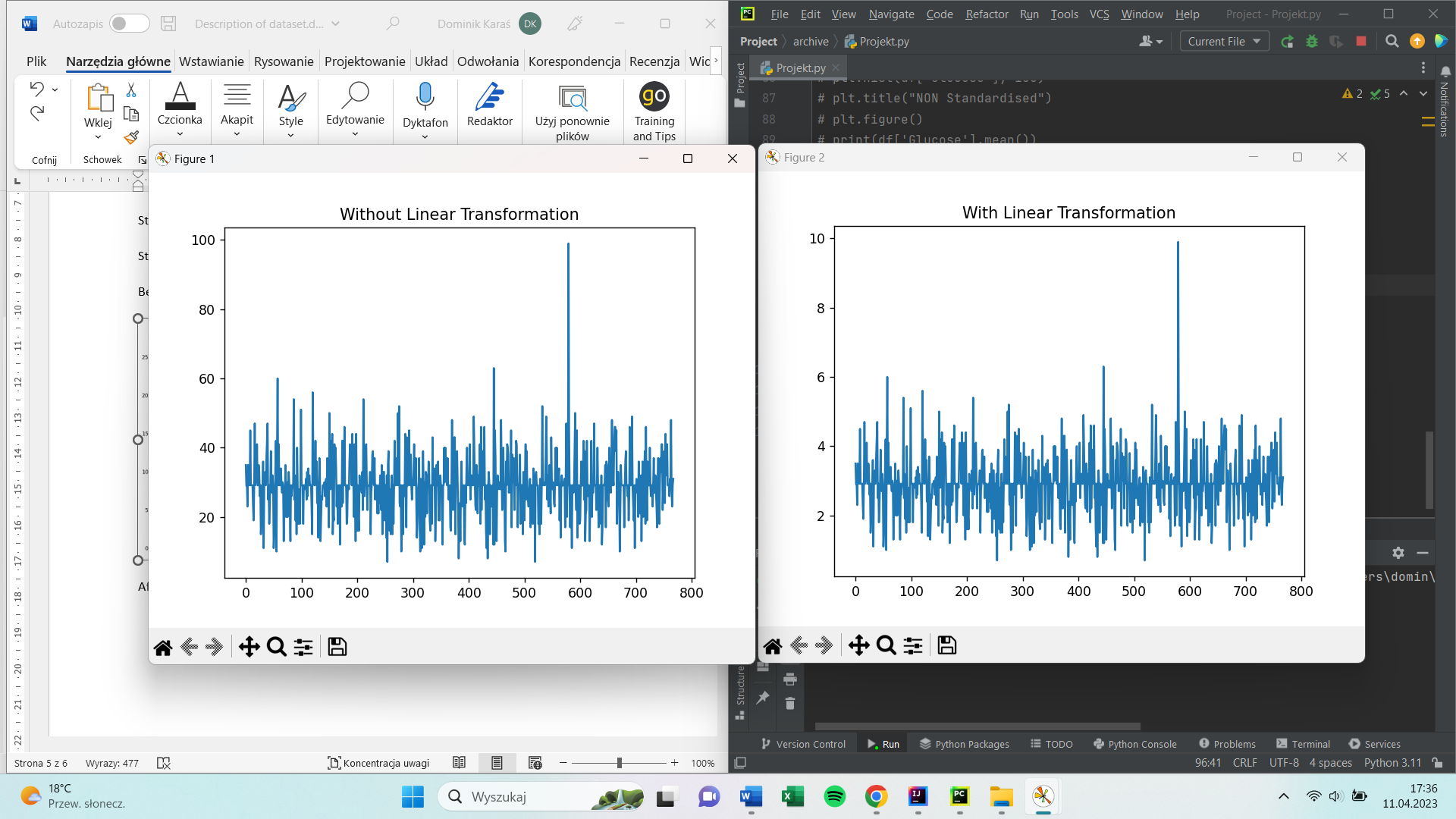
Standardization

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

Obraz zawierający wykres

Opis wygenerowany automatycznie

Linear Transformation



// does number of pregnancies influence blood pressure or other variables

Bmi and age (histogram)

Boxplot for all variables