

**TASK**

**Exploratory Data Analysis on the diabetes Dataset**



**Introduction**

Diabetes dataset shows correlation between medical characteristics and the risk of having diabetes. The dataset is stored in a comma separated values file. This file has 9 columns and 768 rows.

Column names and meaning:

* Pregnancies- [continuous] number of pregnancies
* Glucose- [continuous] level of blood glucose
* BloodPressure- [continuous] blood pressure
* SkinThickness- [continuous] thickness of the skin
* Insulin- [continuous] level of insulin in blood
* BMI- [continuous] calculated BMI
* DiabetesPedigreeFunction- [continuous] likelyhood of diabetes based on family history
* Age- [continuous] age of patient
* Outcome- [categorical] 0 for not being diabetic and 1 for being diabetic

**DATA CLEANING**

* removing duplicates
* no need to change types of columns
* dealing with missing values

**MISSING DATA**

In this dataset missing data is stored with value 0. We are not able to determin if the values 0 in pregnancies and outcome are true values or a missing data, that's why these columns will remain unchanged.

Looking at the other columns the value can't be 0 and here we can safely replace '0' value with numpy not a number value.

Turned out that 7 rows were lacking 4 out of 7 values. These rows were removed.

The rest of the missing values were replaced with median values of each column.

**DATA STORIES AND VISUALISATIONS**

Percent of people with diabetes in the dataset



Almost 35% of entries are for patients with diabetes. In the UK around 7% which means that certain type of people attend this kind of screening.

Correlation between columns



* The highest impact on being diabetic has glucose level, BMI and age.
* There is a high correlation between glucose level and insulin level.
* Age has high correlation with number of pregnancies.
* Skin thickness depends on BMI.

The most significant impact on being diagnosed with diabetes has sugar level. Let's see how the statistics form for entiries with diabetes and without diabetes.



We can see that the average and the min values differ but the max values don't. This is exactly what we would expect from description of the disease itself. The blood sugar stays high for longer because of insulin resistance.

Lets see how glucose and insuline depend on each other. For this purpose glucose level is binned into three ranges of values.



We can see that insulin level raises with sugar level. Also as the glucose level rises the insulin level has wider range. This again explain the insulin resistance. People with diabetes have their insulin level high most of the time but it doesn't lower glucose level as it should.

Let's also compare more significant factors in developing diabetes.



As we can see, people with normal BMI can be diabetic if their glucose level is high. Being overweigh and having glucose level highier than 100 almost guarantees disease. For patients with obesity even when their blood sugar is low they can be ill. Usually as people age they gain weight, the thickness of the skin grows and sugar level can be higher. This does't necessarily mean that they will develop diabetes.

**THIS REPORT WAS WRITTEN BY : MALGORZATA KOLKA**

