Diabetes statistics

- **Introduction**

Diabetes is a chronic health condition that affects how your body turns food into energy. There are two main types of diabetes: type 1 and type 2. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

- **Data Source**

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases / Kaggle

Columns are shown:

- Pregnancies: Number of pregnancies
- Glucose: 2-hour plasma glucose concentration in oral glucose tolerance test
- Blood Pressure: Blood Pressure (small blood pressure) (mm Hg)
- SkinThickness: Skin Thickness
- Insulin: 2-hour serum insulin (mu U/ml)
- DiabetesPedigreeFunction: Function (2-hour plasma glucose concentration in oral glucose tolerance test)
- BMI: Body mass index
- Age: Age (years)
- Outcome: Have the disease (1) or not (0)

- **Data Assessment for credibility & integrity**

1. Reliable — **HIGH** — no sample bias, the sample size is high .
2. Original — **HIGH —** National Institute of Diabetes and Digestive and Kidney Diseases
3. Comprehensive — **HIGH** —Data is within the parameters are clear and good.
4. Current — **medium** — data was sourced and put online 2022
5. Cited — **HIGH** — the data can be found on Kaggle.

- **PREPARE**

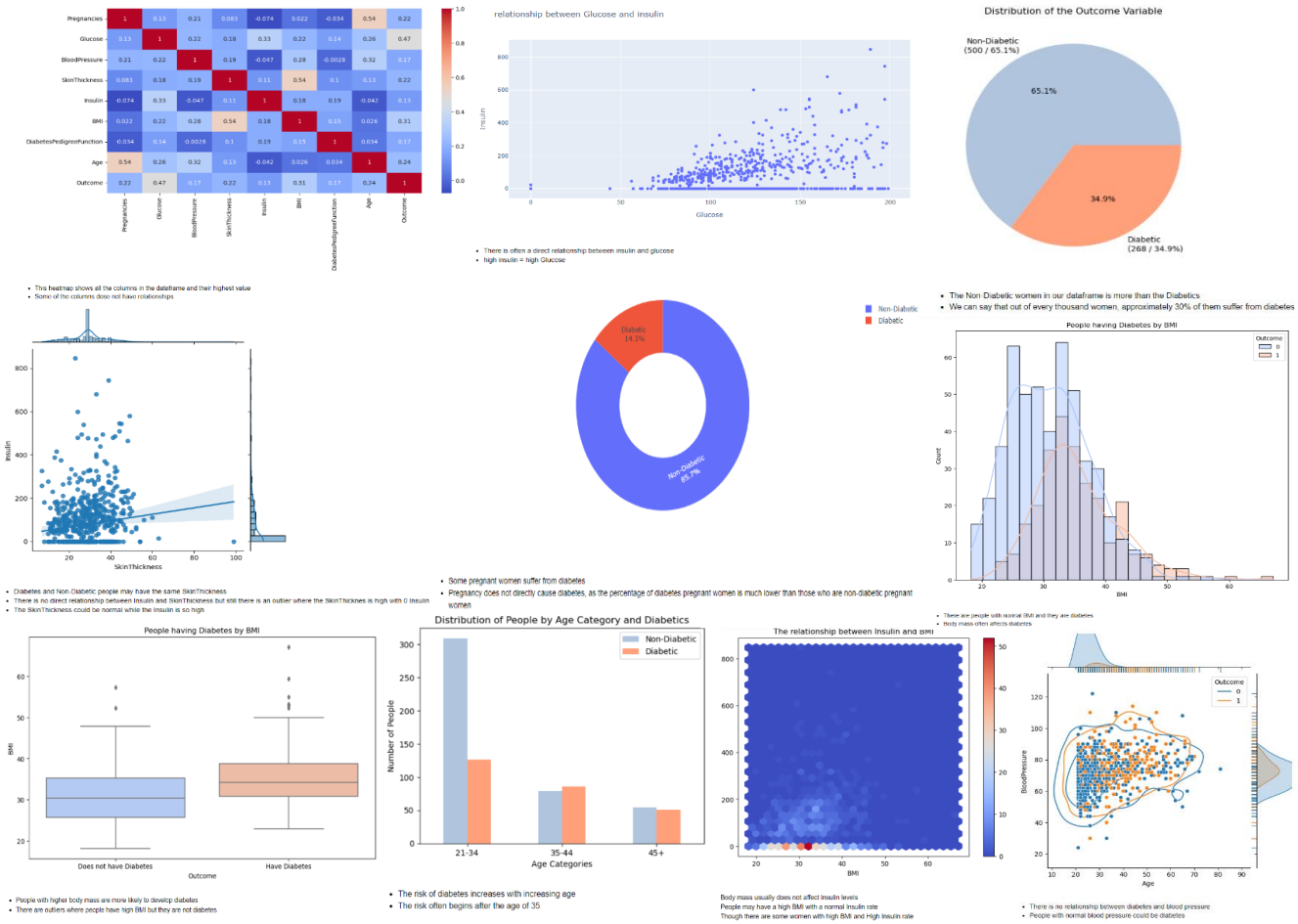
We will be using Python and some of its popular data science related packages. First of all, we will import pandas to read our data from a CSV file and manipulate it for further use. We will also use numpy to convert out data into a format more suitable, we used data profiling for the last report. We will use seaborn and matplotlib and plotly for visualizations.

- **PROCESS**

  Here, we will perform data cleaning operations to ensure the dataset is correct, complete and error free.

- **FINAL TEN INSIGHTS**

1. All variables except insulin seem to have some degree of normal distribution
2. All variables have outliers
3. There are no missing values



GROUP MEMBERS:

Zayed Alharbi

Hesham Alsadan

Refal Alboqami