Abstract and Dataset details:

**Abstract:**

Introduction:

Data analysis is the process of organizing, collecting raw data and analyzing it to better comprehend trends and patterns in a specific data set. These insights benefit individuals and businesses in decision making or solving business problems. On the other hand, data visualization is the process of gathering raw data and visualizing it using charts, tables, and graphs.

Problem statement:

The dataset used originally comes from the National Institute of Diabetes and Digestive Kidney Diseases. The main objective of this dataset is to predict and determine based on diagnostic measurements whether or not a patient has diabetes. However, several constraints were placed for the selection of data in this large database, for example, all patients used to collect the data from are females at least 21 years old and of Pima Indian heritage. Visualizing and understanding the data better will give us a clearer idea as to which variables have a direct correlation with a patient having diabetes. Pima Indians are a Native American group that are deemed to have a high incidence rate of diabetes, thus, the research involving them is thought to be significant and representative of global health.

Process/Approach:

To approach and investigate the independent variables that have a direct positive relationship with patients having diabetes. We can first create scatterplots between the variables that display whether or not an individual has diabetes or not (dependent variable) with all other independent variables. A pie chart can be created to display the distribution of different categorical variables. Moreover, histograms can be used to illustrate the distribution of data over a continuous interval or a defined period of time. Lastly, a regression output can also be created which shows in detail, the relationship between all independent variables and the dependent variable.

Conclusion:

After analyzing and visualizing the dataset, we should be able to conclude that variable such as number of pregnancies, age & BMI would have a direct, strong & positive relationship with the dependent variable (patient does/doesn’t have diabetes).

**Dataset Details:**

Dependent variable:

* Disease: outcome, 1 if diabetes, 0 if not

Independent variables:

* num\_preg: number of times pregnant.
* Plasma: plasma glucose concentration, 2 hours in an oral glucose tolerance test
* dias\_dp: Diastolic blood pressure (mm Hg)
* triceps: triceps skin fold thickness (mm)
* insu\_lev: 2-hour serum insulin (mu U/ml)
* bmi: body mass index (weight in kg/(height in m)^2)
* age: age in years

Dataset link: <https://www.kaggle.com/datasets/mathchi/diabetes-data-set>

Dataset:

