LAB 1
Understanding Data through Statistical and Visualization Techniques

Data: Data are raw, unorganized facts, figures, or values. They lack inherent meaning until processed and interpreted. Data can exist in various forms, including numbers, text, images, and sounds. They serve as the foundational elements from which information is derived.

There are two main types:

- Quantitative: Numerical (e.g., temperature, age)
- Qualitative: Descriptive (e.g., survey responses).

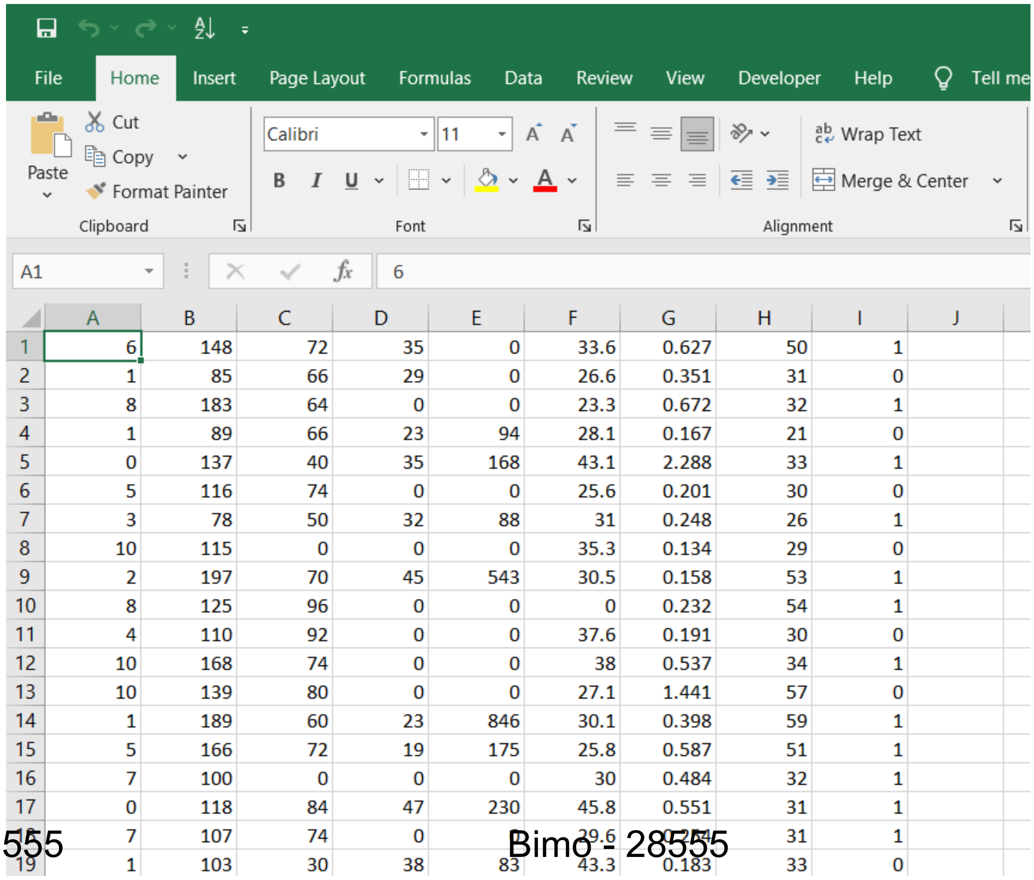
Data can be structured (organized) or unstructured (not organized).

Exploratory Data Analysis (EDA): EDA is an approach for analyzing data sets. The main purpose is to summarize their characteristics, often with visual methods. This process is used to discover patterns, spot anomalies, test hypotheses, and check assumptions before formal modeling or further statistical analysis.

The key benefits of EDA include:

- **Initial Data Understanding:** EDA provides an initial, objective look at the data. This helps analysts become familiar with the dataset's structure, including the number of variables, data types, and the overall distribution of values.
- **Identification of Issues:** Through EDA, errors, inconsistencies, and missing values in the data can be identified. These data quality issues could lead to flawed analysis and inaccurate models if left unaddressed.
- **Feature Identification:** EDA helps in discovering patterns and relationships between variables. This process assists in selecting the most relevant features for a predictive model, which can improve its performance and efficiency.
- **Outlier Detection:** Outliers can be easily spotted using EDA's visual methods. Understanding and handling these outliers is important, as they can significantly skew statistical results and model outcomes.
- **Hypothesis Generation:** The insights gained from exploring the data can lead to the formation of new hypotheses or research questions that were not initially considered. This can guide further, more focused analysis.
- **Assumption Checking:** Many statistical methods and machine learning models rely on specific assumptions about the data. EDA helps to visually and statistically check these assumptions, ensuring that the chosen methods are appropriate for the dataset.

Dataset Description: This PIMA dataset contains 768 observations and nine features. The features include medical predictor variables and one target variable, Diabetic. The Diabetic variable indicates whether a person has diabetes (1) or not (0).



	A	B	C	D	E	F	G	H	I	J
1	6	148	72	35	0	33.6	0.627	50	1	
2	1	85	66	29	0	26.6	0.351	31	0	
3	8	183	64	0	0	23.3	0.672	32	1	
4	1	89	66	23	94	28.1	0.167	21	0	
5	0	137	40	35	168	43.1	2.288	33	1	
6	5	116	74	0	0	25.6	0.201	30	0	
7	3	78	50	32	88	31	0.248	26	1	
8	10	115	0	0	0	35.3	0.134	29	0	
9	2	197	70	45	543	30.5	0.158	53	1	
10	8	125	96	0	0	0	0.232	54	1	
11	4	110	92	0	0	37.6	0.191	30	0	
12	10	168	74	0	0	38	0.537	34	1	
13	10	139	80	0	0	27.1	1.441	57	0	
14	1	189	60	23	846	30.1	0.398	59	1	
15	5	166	72	19	175	25.8	0.587	51	1	
16	7	100	0	0	0	30	0.484	32	1	
17	0	118	84	47	230	45.8	0.551	31	1	
18	7	107	74	0	0	29.6	0.254	31	1	
19	1	103	30	38	83	43.3	0.183	33	0	

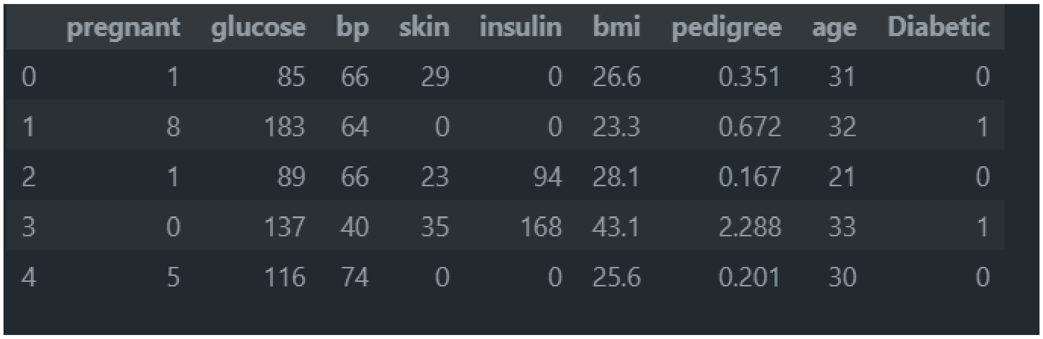
Python Codes

# Import

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
```

# Load dataset and attach corresponding label to each column of the raw data

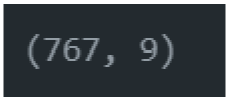
```
col_names = ['pregnant', 'glucose', 'bp', 'skin', 'insulin', 'bmi', 'pedigree', 'age', 'Diabetic']
dbts_ds= pd.read_csv('pima_diabetes.csv', header=0, names=col_names)
dbts_ds.head()
```



	pregnant	glucose	bp	skin	insulin	bmi	pedigree	age	Diabetic
0	1	85	66	29	0	26.6	0.351	31	0
1	8	183	64	0	0	23.3	0.672	32	1
2	1	89	66	23	94	28.1	0.167	21	0
3	0	137	40	35	168	43.1	2.288	33	1
4	5	116	74	0	0	25.6	0.201	30	0

# Check dimensions i.e. number of rows and columns in this dataset

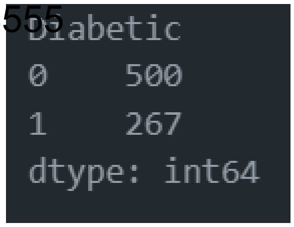
```
dbts_ds.shape
```



```
(767, 9)
```

# Determine how many rows are related to diabetic (1) class and non diabetic (0) in the total 768 rows of data

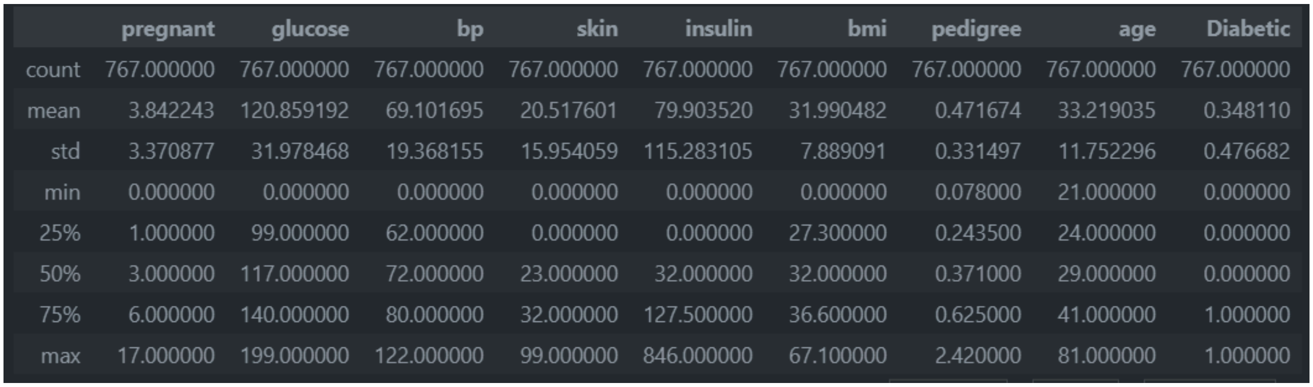
```
dbts_ds.groupby('Diabetic').size()
```



```
Diabetic
0    500
1    267
dtype: int64
```

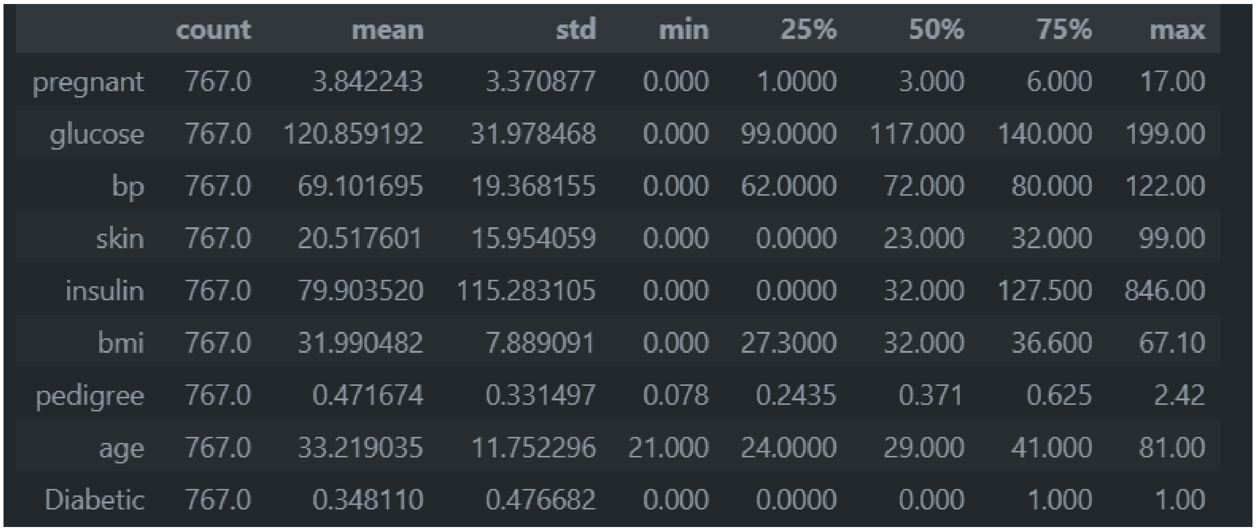
# View statistical details about the data

```
dbts_ds.describe()
```



	pregnant	glucose	bp	skin	insulin	bmi	pedigree	age	Diabetic
count	767.000000	767.000000	767.000000	767.000000	767.000000	767.000000	767.000000	767.000000	767.000000
mean	3.842243	120.859192	69.101695	20.517601	79.903520	31.990482	0.471674	33.219035	0.348110
std	3.370877	31.978468	19.368155	15.954059	115.283105	7.889091	0.331497	11.752296	0.476682
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243500	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	32.000000	32.000000	0.371000	29.000000	0.000000
75%	6.000000	140.000000	80.000000	32.000000	127.500000	36.600000	0.625000	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

```
dbts_ds.describe().T
```



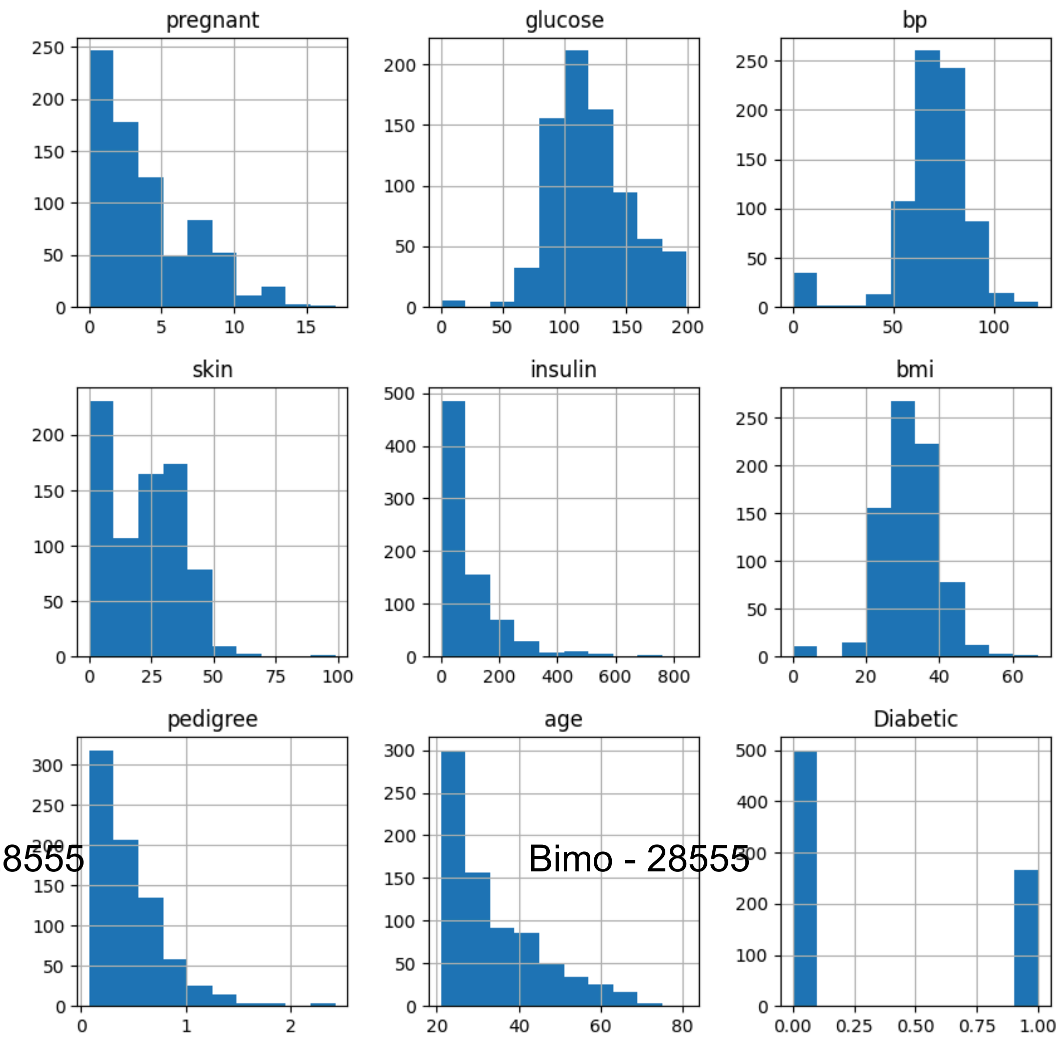
	count	mean	std	min	25%	50%	75%	max
pregnant	767.0	3.842243	3.370877	0.000	1.0000	3.000	6.000	17.00
glucose	767.0	120.859192	31.978468	0.000	99.0000	117.000	140.000	199.00
bp	767.0	69.101695	19.368155	0.000	62.0000	72.000	80.000	122.00
skin	767.0	20.517601	15.954059	0.000	0.0000	23.000	32.000	99.00
insulin	767.0	79.903520	115.283105	0.000	0.0000	32.000	127.500	846.00
bmi	767.0	31.990482	7.889091	0.000	27.3000	32.000	36.600	67.10
pedigree	767.0	0.471674	0.331497	0.078	0.2435	0.371	0.625	2.42
age	767.0	33.219035	11.752296	21.000	24.0000	29.000	41.000	81.00
Diabetic	767.0	0.348110	0.476682	0.000	0.0000	0.000	1.000	1.00

Exploratory Data Analysis (EDA)

Histogram: Here we look at the distribution of each attribute by discretizing the continuous values into buckets and count the frequency in each bucket as histograms.
Histogram is formed by counting the frequency of each value in the attribute and then plotting it as a bargraph.

This lets us note interesting properties of the attribute distributions such as the possible normal distribution of attributes associated with the features of the dataset or skewness of the data feature.

```
dbts_ds.hist(figsize = (10,10))
```



Histogram Data Interpretation of each attributes above:

Each picture above is a histogram that shows the distribution of each feature or attribute. Each figure represents a univariate distribution. i.e. involving only one variable or feature.

Certain insights from the histogram plot:

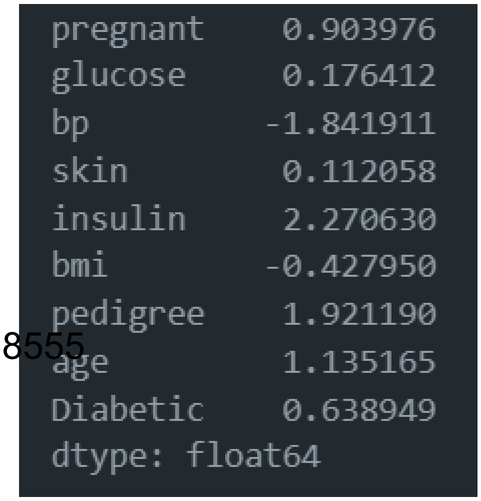
1. We have identified whether the data is normally distributed or skewed. i.e. "glucose" as normally distributed while "insulin" is right skewed.)
2. Once we identify the type of distribution we can apply a specific approach for missing values based on data distribution. i.e. We may replace the missing value of "glucose" as mean and missing value of "insulin" as median.
3. Most of the learning models are based on Normal/Gaussian distribution (the bell shaped curve). So if any data feature is skewed we can transform it into normal distribution then impute the missing values and then feed the learning model.Because such models make good predictions if data are distributed normally.

4. Algorithms are biased when the data distribution is skewed i.e. if the learning algorithm (model) gets trained on the above insulin distribution then the model gives more accurate results on people with insulin level less than 250 but gives incorrect prediction on people with insulin level greater then 350.

# Check the skewness degree of each attribute.

Values near zero are less skewed as compared to values away from it. Distributions that have skewness value less than -1 or greater than +1 are skewed.

```
skew_attrib_val = dbts_ds.skew()
print(skew_attrib_val)
```



```
pregnant    0.903976
glucose     0.176412
bp         -1.841911
skin        0.112058
insulin     2.270630
bmi        -0.427950
pedigree    1.921190
age         1.135165
Diabetic    0.638949
dtype: float64
```

Data Interpretation of above result:

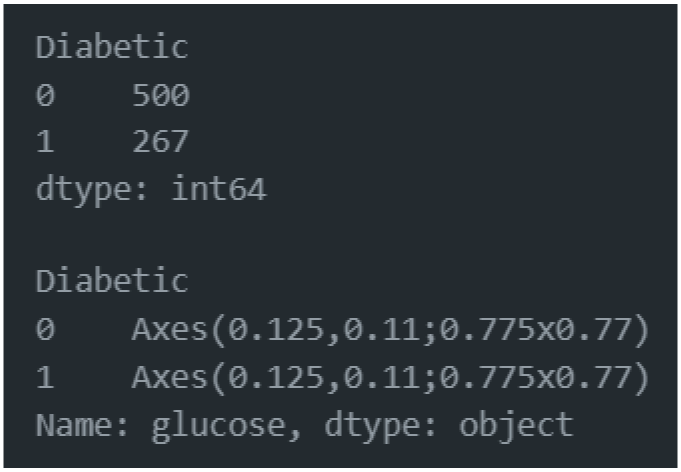
From the above skew value, We found that glucose, pregnancy are normally distributed while features like bp, insulin, pedigree etc are skewed.

Another Insight from histogram data: The imbalanced class in the classification problem.

Below figure shows there's a lot more observations for one class than another and may need special handling in the data preparation.

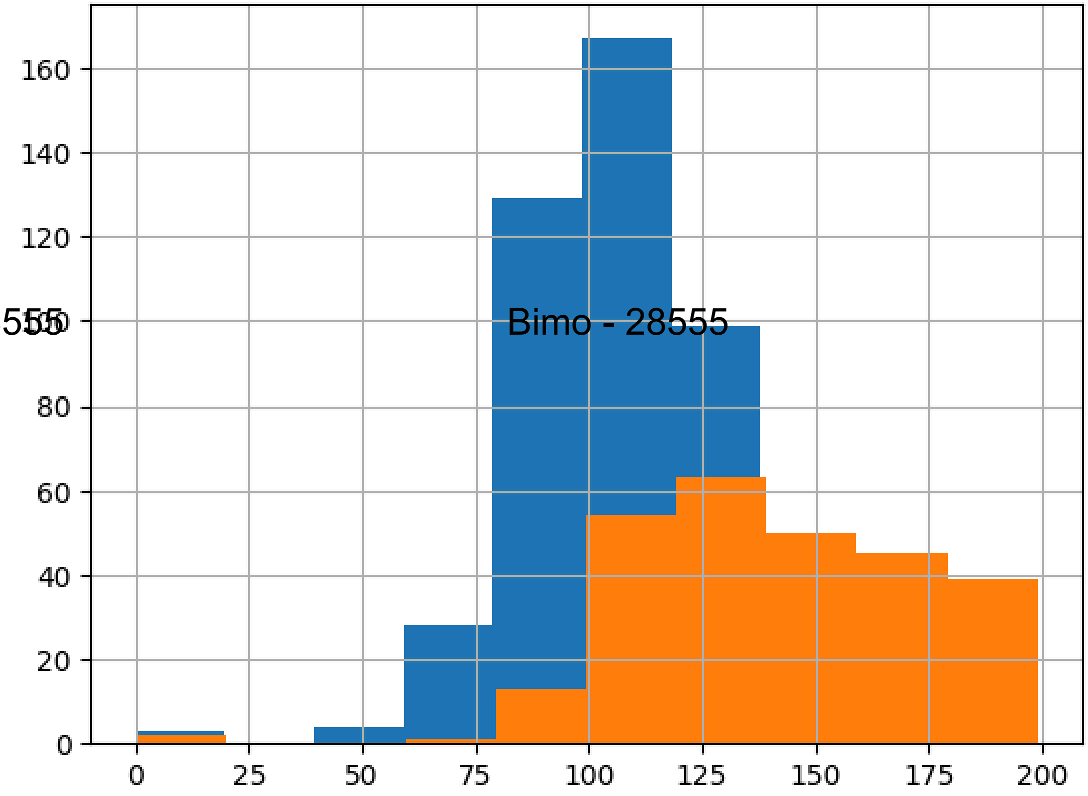
To know such imbalance: Either use a count function or visualize through a histogram. Below figure shows a lot more non-diabetic observations than diabetic which needs special handling.

```
output_grp = dbts_ds.groupby('Diabetic').size()
print(output_grp)
dbts_ds.groupby('Diabetic').glucose.hist(alpha=1)
```



```
Diabetic
0    500
1    267
dtype: int64

Diabetic
0    Axes(0.125,0.11;0.775x0.77)
1    Axes(0.125,0.11;0.775x0.77)
Name: glucose, dtype: object
```



The above figure shows that there are a lot more non diabetic patients than diabetes patients in the total 768 observations.

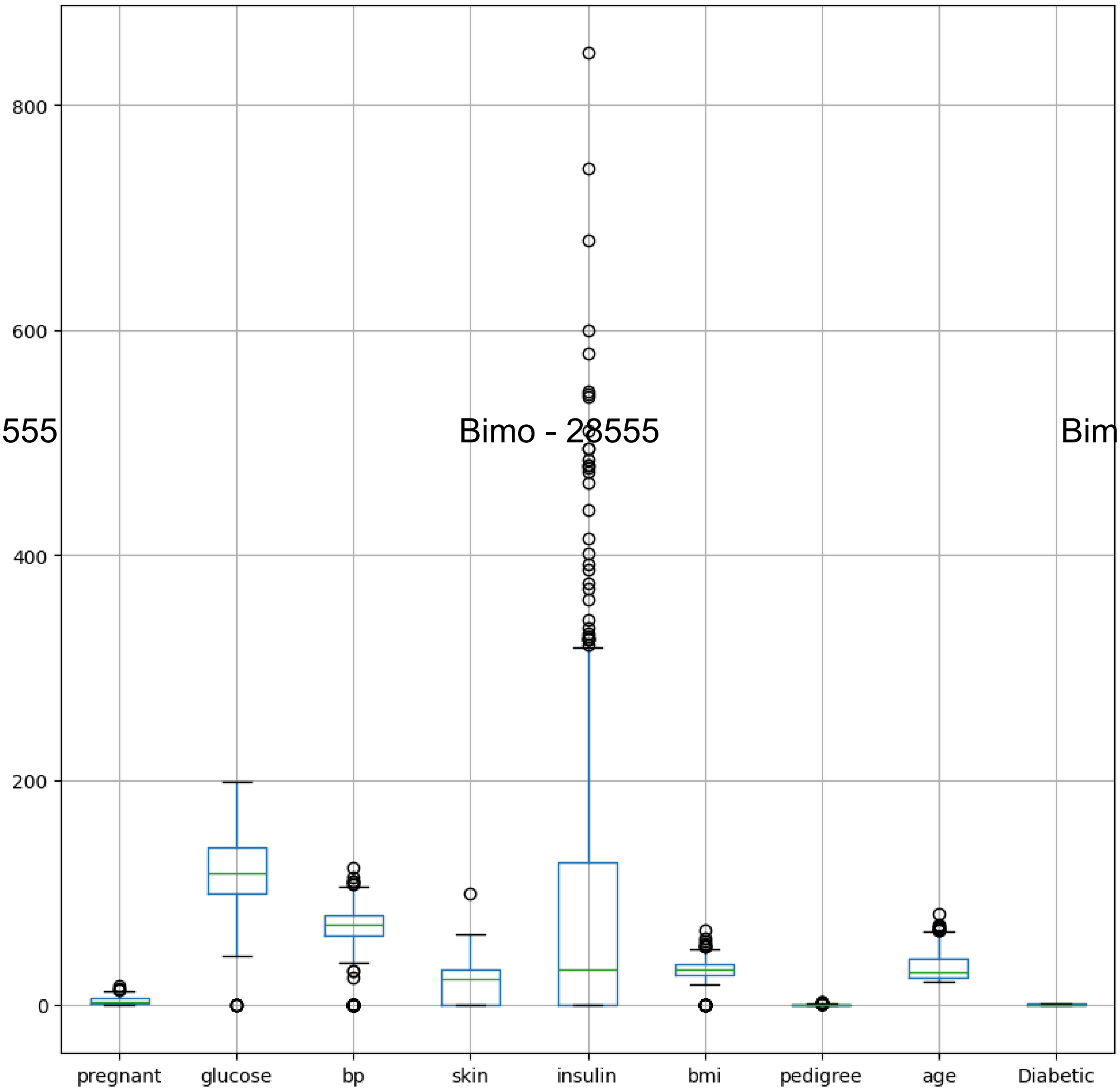
Training models on such imbalanced classes will likely predict the majority class (non diabetic) and hence the model will have low accuracy predicting diabetic population.

Box Plots

Visualize data distribution's median as a thick line within the box, interquartile range with 25th percentile as left most boundary and 75th percentile as right most boundary. Data outside of these interquartile ranges are considered outliers.

From the diagram below Insulin has the highest outliers while glucose has the lowest.

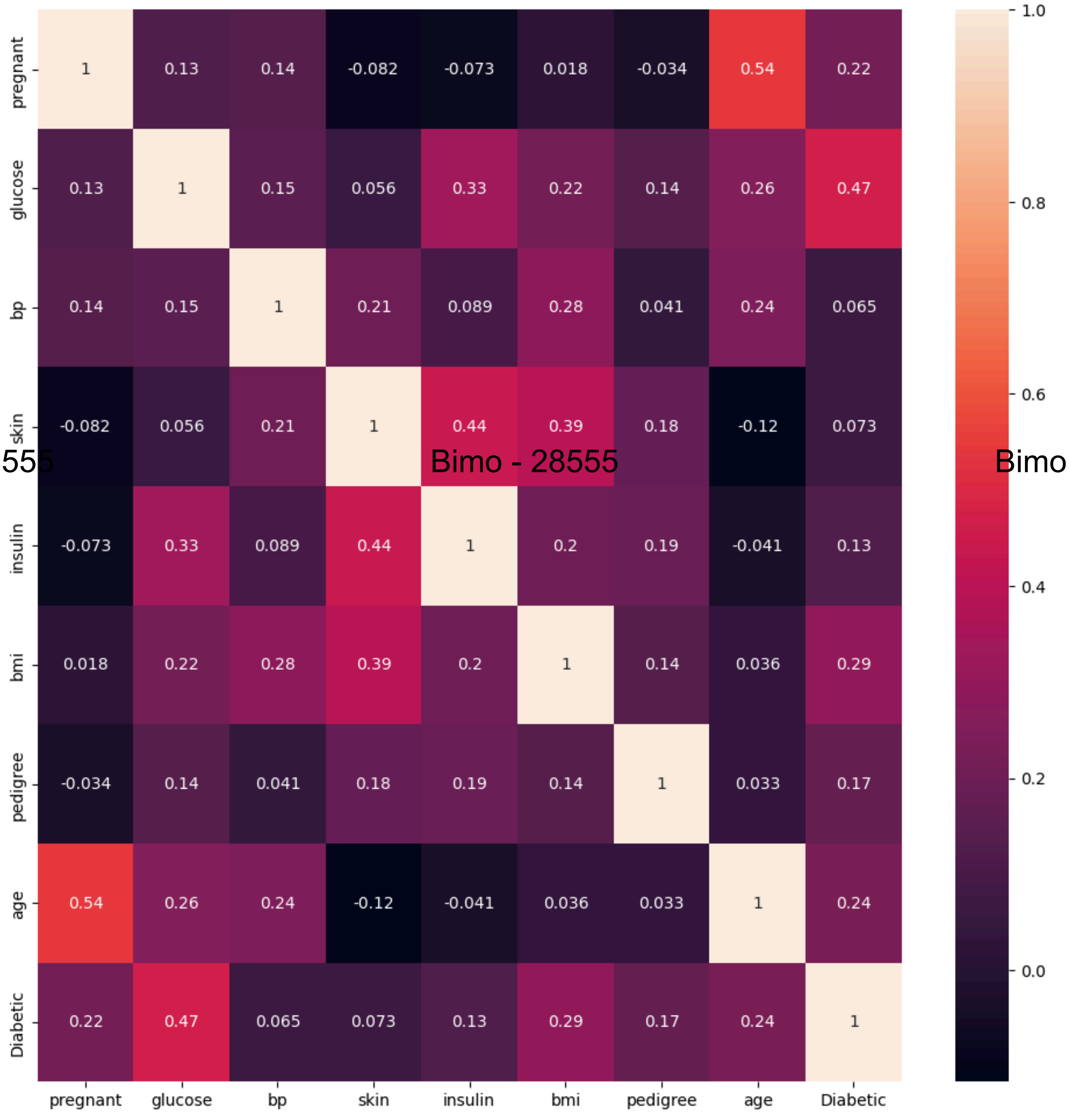
```
dbts_ds.boxplot(figsize =(10,10))
print("MODIFIED  PIMA_NEW DATASET")
```



Heatmap Plot

Identifying correlation between two attributes, values nearer to +1 represents stronger linear relationship between two variables while values nearer to -1 represents inverse relation.

```
plt.figure(figsize=(12,12))  # we set the size of the figure to 12 by 12.
dbts_ds=sns.heatmap(dbts_ds.corr(), annot=True)  # use seaborn library heatmap
```


pregnant 1 0.13 0.14 -0.082 -0.073 0.018 -0.034 0.54 0.22
glucose 0.13 1 0.15 0.056 0.33 0.22 0.14 0.26 0.47
bp 0.14 0.15 1 0.21 0.089 0.28 0.041 0.24 0.065
skin -0.082 0.056 0.21 1 0.44 0.39 0.18 -0.12 0.073
insulin -0.073 0.33 0.089 0.44 1 0.2 0.19 -0.041 0.13
bmi 0.018 0.22 0.28 0.39 0.2 1 0.14 0.036 0.29
pedigree -0.034 0.14 0.041 0.18 0.19 0.14 1 0.033 0.17
age 0.54 0.26 0.24 -0.12 -0.041 0.036 0.033 1 0.24
Diabetic 0.22 0.47 0.065 0.073 0.13 0.29 0.17 0.24 1
pregnant glucose bp skin insulin bmi pedigree age Diabetic