UNIVERSITY OF EDINBURGH
Business School

2023-24

CMSE11615 DATA ANALYSIS AND STATISTICS FOR BUSINESS

ANALYSING PIMA INDIAN DIABETES DATA

Lahiru Panduka Alahakoon

Word count: 1895

TABLE OF CONTENTS

List of Tables

List of Figures

1. INTRODUCTION

1.1 Diabetes

Diabetes is a metabolic disease that occurs when the Glucose level in blood exceeds the desirable limit. Glucose is an essential source of energy for the body's cells to make up the muscles, tissues, and also, the main source of energy for the brain. It mainly comes from the Carbohydrates in foods and drinks. Glucose is absorbed into cells to be used for energy with the help of Insulin produced by the Pancreas. When the body doesn't produce or use Insulin effectively, glucose accumulates in bloodstreams and leads to High Blood Sugar. Consistent high blood sugar over time can damage the Kidneys, Nerves, Eyes, and Heart.

There are several types of diabetes, the most common types are Type 1, Type 2 and Gestational Diabetes.

- Type 1 Diabetes – Attacks and destroys cells which produce insulin in the pancreas. Diabetes type 1 is commonly diagnosed in children and young adults, but it can develop at any age.
- Type 2 diabetes – The body doesn't make enough insulin. The cells in pancreas may be making insulin but is not enough to respond to blood glucose level in the normal range. People with obesity and family history with the disease are more likely to develop type 2 diabetes.
- Gestational diabetes – More likely to develop during pregnancy. This type of diabetes mostly goes away after the baby is born. However, there is a high chance of developing type 2 diabetes later in life.

The major symptoms of diabetes include frequent urination, increased thirst and dry mouth, fatigue, blurred vision, unexplained weight loss and frequent skin infection

1.2 Previous Research

Several studies have been conducted on Pima Indians to analyze diabetes. Victor Chang, Jozeene Bailey, Qianwen Ariel Xu, and Zhili Sun developed a machine learning-based diabetes classification system for Pima Indians. This system is designed to diagnose type 2 diabetes mellitus. (Chang et al., 2022)

1.3 Objective of the Analysis

The objective of this analysis is to investigate and compare the mean Plasma Glucose Levels between Diabetic and Non-Diabetic patients, and to assess the potential association between pregnancy and diabetes.

2. DESCRIPTION OF THE DATASET

The Pima Indian Diabetes Dataset is a publicly available dataset that contains information on 768 Pima Indian women who were diagnosed with diabetes mellitus between 1991 and 1995. The dataset includes a variety of demographic and medical information, including:

- Personal characteristics: Age, Body Mass Index (BMI), Number of times been Pregnant, Triceps skin fold Thickness (mm)
- Diagnostic measurements: Diastolic Blood Pressure, Plasma Glucose Concentration, Serum Insulin Concentration

The data was collected by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) as part of a study to investigate the risk factors for diabetes mellitus in the Pima Indian population. The study aimed to identify factors that could be used to predict and prevent diabetes in this population. Dataset was later uploaded to Kaggle by UCI Machine Learning. Link: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. (Learning, 2016)

The dataset is important because it provides information on a population that is at high risk for diabetes. The Pima Indians have a prevalence of diabetes that is three times higher than the US national average. This makes the dataset valuable for studying the factors that contribute to diabetes in high-risk populations.

The dataset consists of several medical predictor variables and one target variable, ‘Outcome’. Predictor variables includes the number of Pregnancies the patient has had, their BMI, insulin level, age, etc.

Dataset includes a total of 768 entries and 9 variables. Metadata are as follows.

Column Name	Description	Data Type
Pregnancies	Number of times pregnant	int64
Glucose	Plasma glucose concentration	int64
Blood Pressure	Diastolic blood pressure (mm Hg)	int64
Skin Thickness	Triceps skin fold thickness (mm)	int64
Insulin	2-Hour serum insulin (mu U/ml)	int64
BMI	Body mass index (weight in kg/(height in m)^2)	float64
Diabetes Pedigree Function	Probability of diabetes based on family history	Float64
Age	Age (years)	Int64
Outcome	Whether a person is positive or negative for Diabetes; Positive: 1, Negative: 0	Int64

Table 1 - Metadata

2.1 Handling Missing Values

There are zero values present in the columns - Glucose, Blood Pressure, Skin Thickness, Insulin and BMI. Zero is not attainable for a living human being. Hence these will be treated as missing values.

Column Name	No. of Missing Values
Glucose	5
Blood Pressure	35
Skin Thickness	227
Insulin	374
BMI	11

Table 2 - Missing Values

Since there are only 768 entries, removing missing rows is not ideal. Missing values will be handled based on the nature of their outliers. Following Box plots show the maximum, minimum, range and outliers and their values. Missing values will be replaced with either mean or median of each variable.

Glucose

Figure 1 - Box Plot of Glucose

No extreme outliers. Missing values will be replaced with the mean.

Blood Pressure

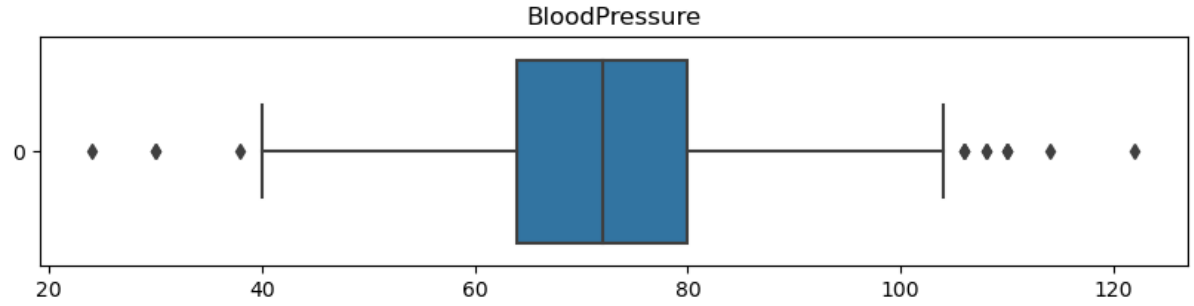

Figure 2 - Box Plot of Blood Pressure

Few Outliers are present. Mean Imputation will be used.

Skin Thickness

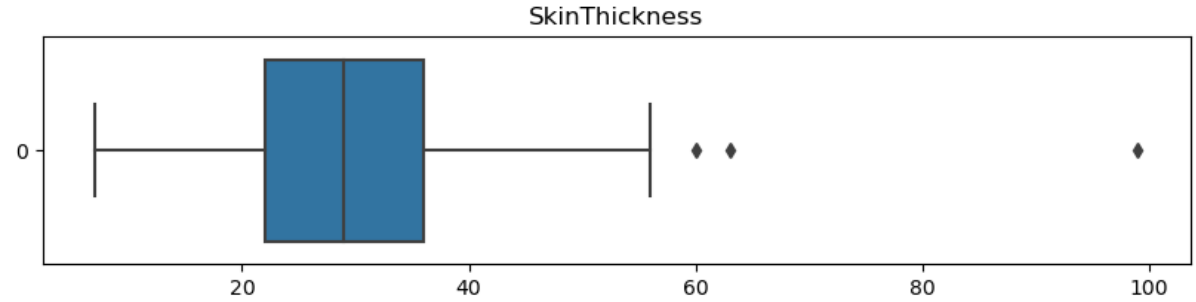

Figure 3 - Box Plot of Skin Thickness

Extreme outliers are present. Missing values will be replaced with the median value.

Insulin

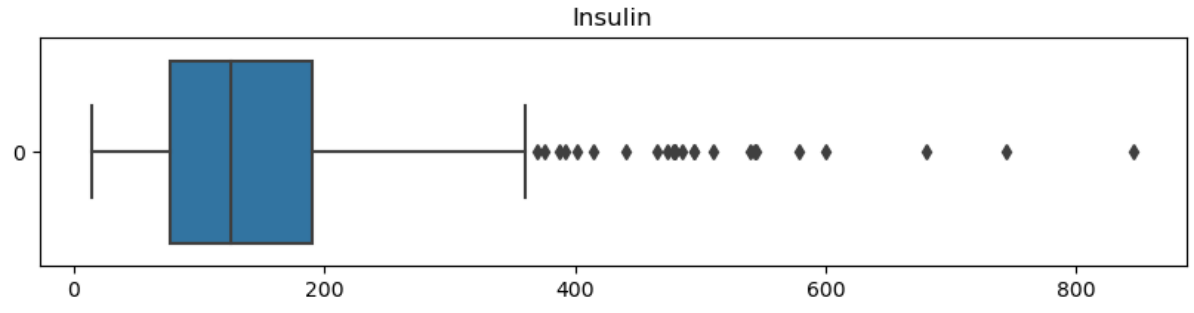

Figure 4 - Box Plot of Insulin

Extreme outliers are present. Missing values will be replaced with the median value.

BMI

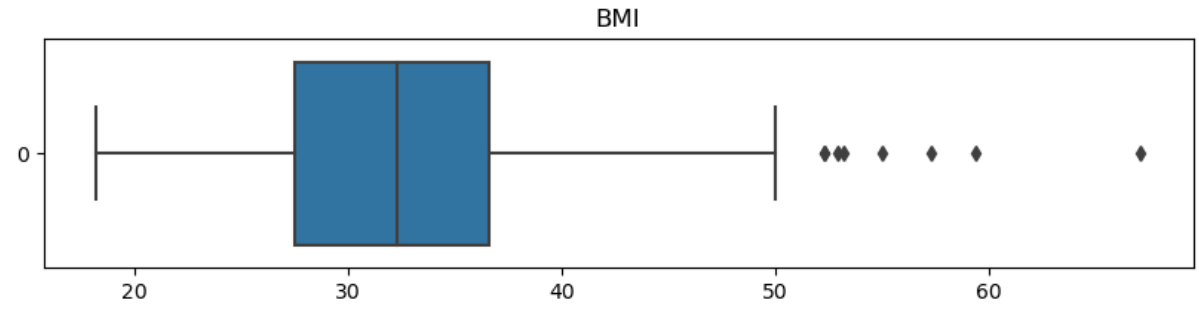

Figure 5 - Box Plot of BMI

Numerous outliers are present. Median Imputation will be used.

3. DESCRIPTIVE AND SUMMARY STATISTICS

3.1 Independent Variables

Dataset contains 8 independent quantitative variables and one dependent categorical variable.

Histograms and QQ Plots below illustrates the distribution of data for each column of the Pima Indians Diabetes dataset. The mean, mode and median values for each variable are also displayed on histograms.

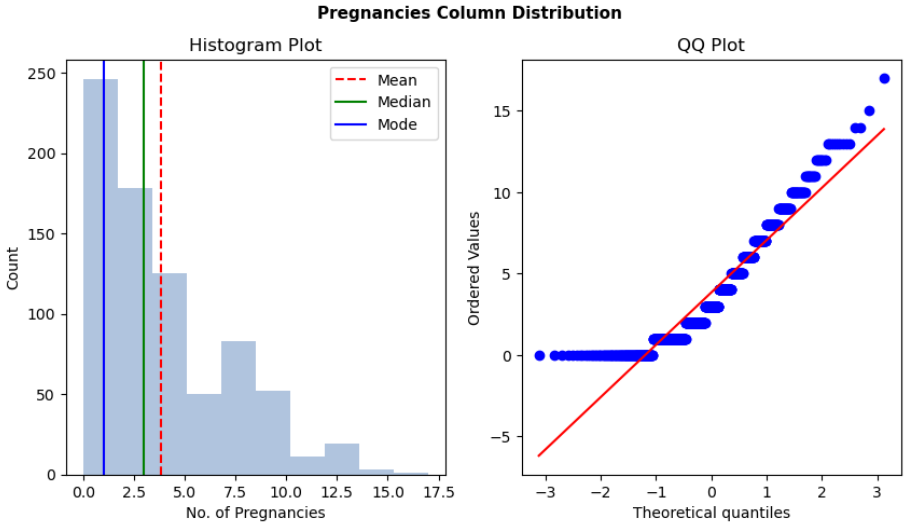

Figure 6 - Pregnancies Column Distribution

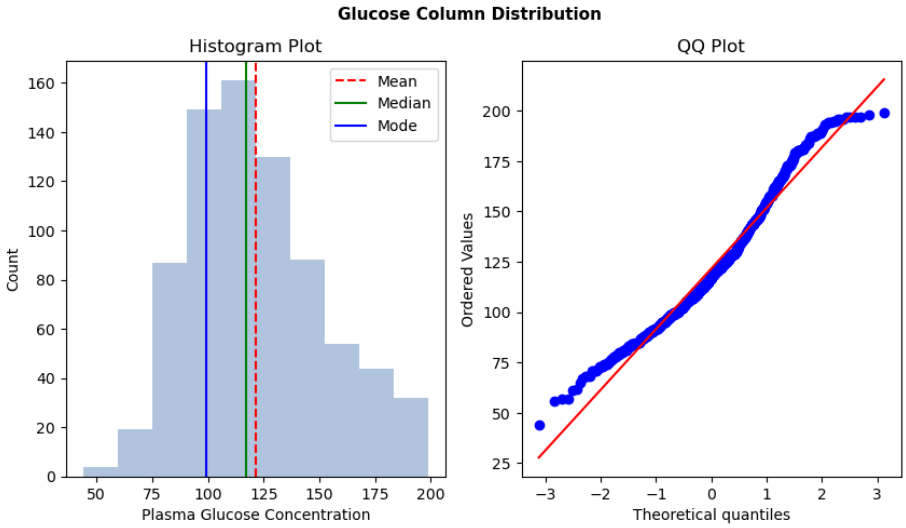

Figure 7 - Glucose Column Distribution

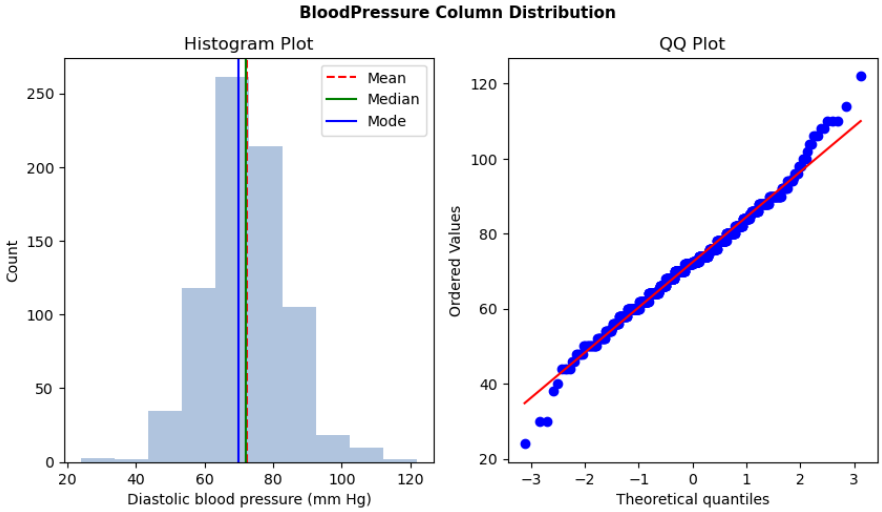

Figure 10 - Blood Pressure Column Distribution

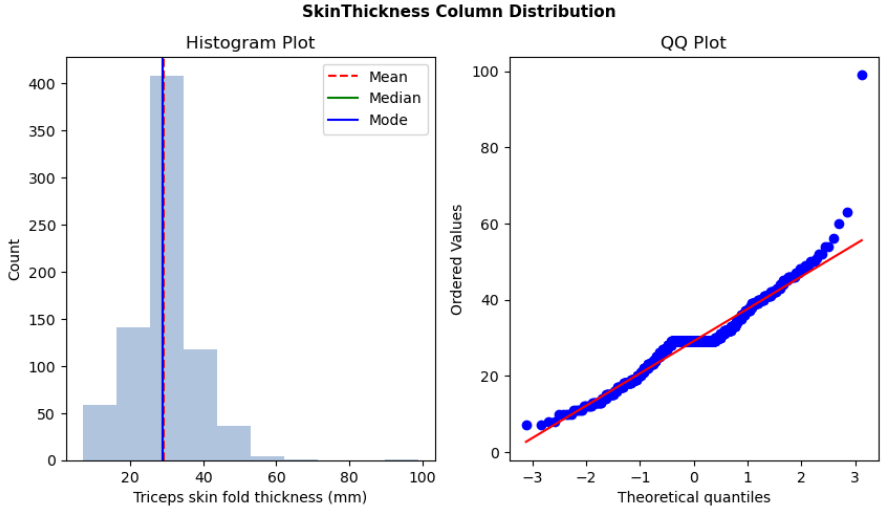

Figure 9 - Skin Thickness Column Distribution

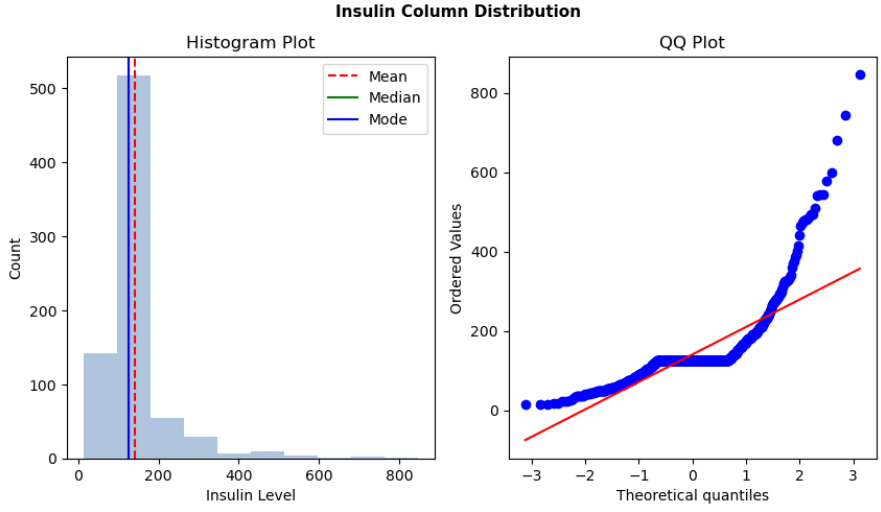

Figure 8 - Insulin Column Distribution

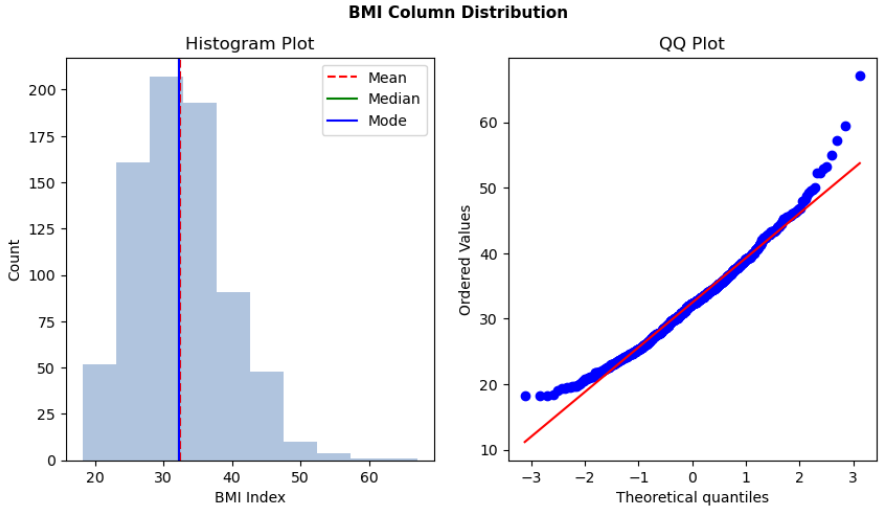

Figure 13 - BMI Column Distribution

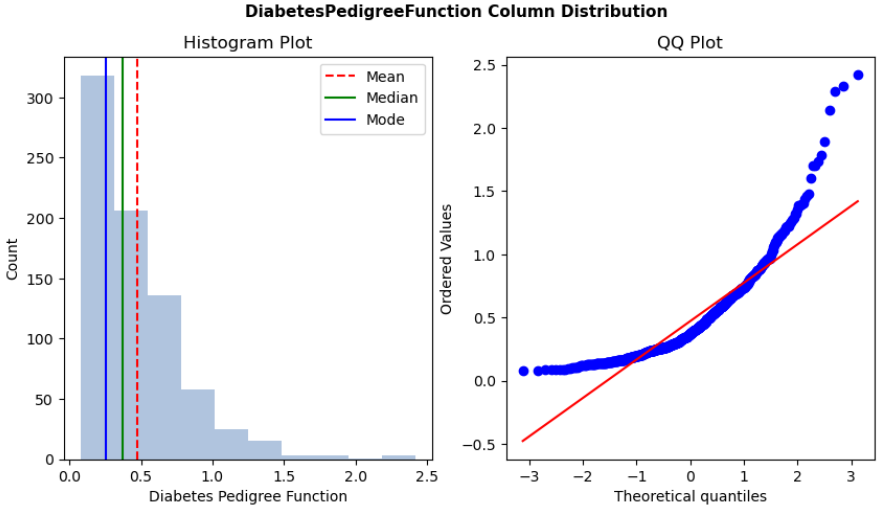

Figure 12 - Diabetes Pedigree Function Column Distribution

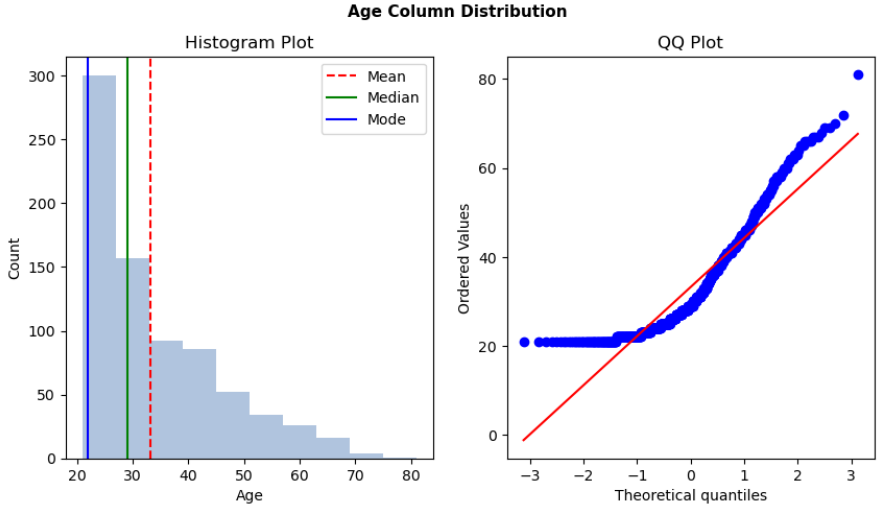

Figure 11 - Age Column Distribution

The above histograms and QQ plots show that Glucose, Blood Pressure and BMI follow a normal distribution whereas Pregnancies, Insulin, Diabetes Pedigree Function, and Age follow positively skewed distribution. The histogram of Skin Thickness is too peaked in the middle. QQ plots for all variables are also interpreted to compare the distribution of the dataset to a theoretical distribution.

3.1.1 Descriptive Statistical Methods

Following equations are being used to create the summary table.

Parameter	Equation/ Description
Mean	$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$
Mode	The mode is the most frequent value in the dataset.
Median	The median is the middle value in the dataset when the data is ordered from least to greatest.
Standard Deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$
Inter Quartile Range	$IQR = Q_3 - Q_1$ ($Q_3 -$ 75th percentile, $Q_1 -$ 25th percentile)
Range	The difference between the largest and smallest values in the dataset.

Table 3 - Descriptive Statistics Formulas

3.1.2 Summary Statistics

Summary of the calculation is shown in the following table.

Column Name	Mean	Standard Deviation	Min	Q1	Median (Q2)	Q3	Max	Range	IQR
Pregnancies	3.85	3.3696	0	1	3	6	17	17	5
Glucose	121.69	30.4359	44	99.75	117	140.25	199	155	40.50
Blood Pressure	72.41	12.0963	24	64	72.203	80	122	98	16
Skin Thickness	29.11	8.7912	7	25	29	32	99	92	7
Insulin	140.67	86.3831	14	121.50	125	127.25	846	832	5.75
BMI	32.46	6.8752	18	27.50	32.300	36.60	67	49	9.10
Diabetes Pedigree Function	0.47	0.3313	0.078	0.244	0.373	0.63	2.42	2.342	0.38
Age	33.24	11.7602	21	24	29	41	81	60	17

Table 4 - Summary Statistics

3.2 Dependent Variable

The dependent variable in the dataset is the Outcome. This variable represents whether a Patient has diabetes (Positive) or not (Negative).

3.2.1 Outcome Distribution

65.1% of the patients in the dataset tested negative for diabetes and 34.9% of the patients tested positive for diabetes. The bar chart provides numerical values of patients. Out of 768 patients, 500 were negative for diabetes and only 268 were positive for diabetes.

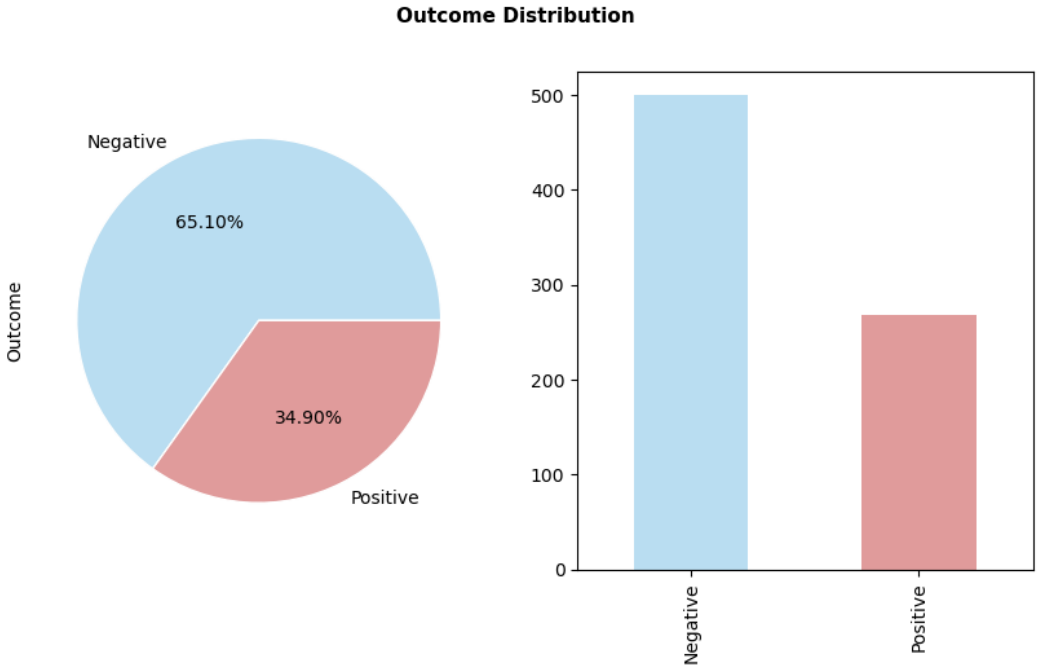

Figure 14 - Outcome Distribution

3.2.2 Box Plots

Below boxplots reflect the means, range, and distribution of two groups (diabetic positive and negative) against each independent variable.

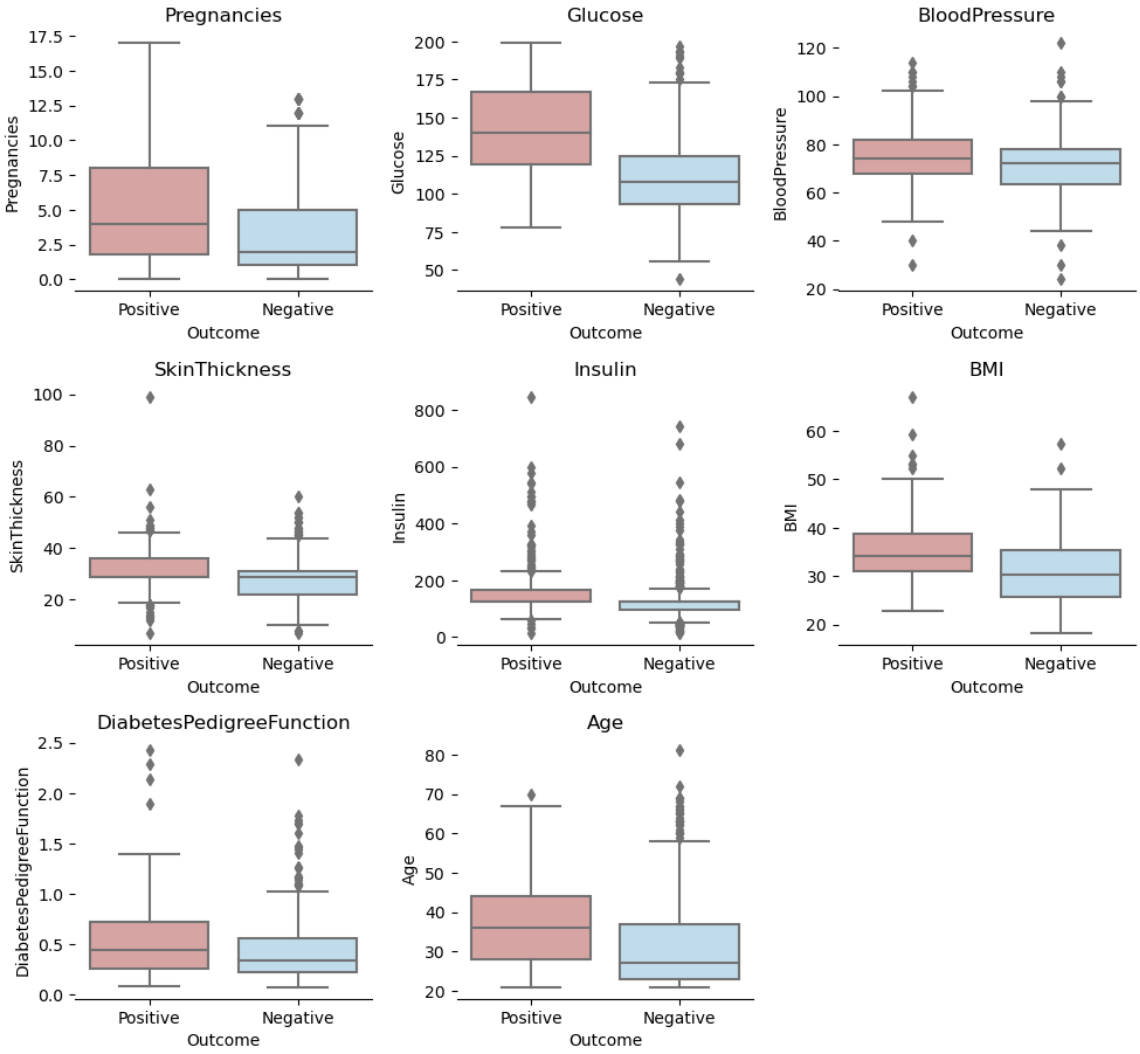

Figure 15 - Box Plots

3.2.3 Pair Plot

Pair plots graphically represent the relationship between every pair of variables in the dataset. The following plots demonstrate that Skin Thickness and BMI have nearly a linear relationship.

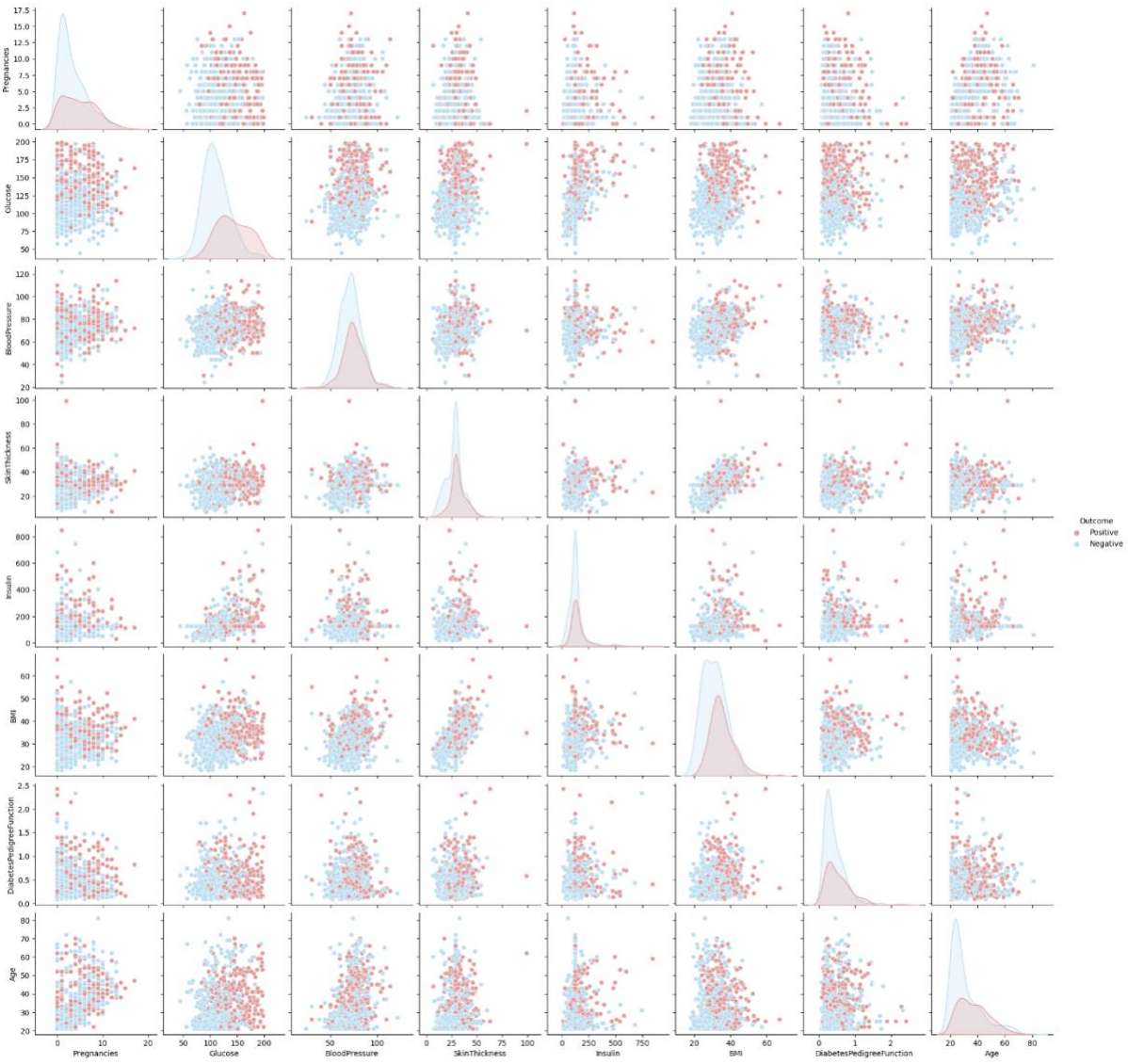

Figure 16 - Pair Plot

3.2.4 Heat Map

The heat map provides the overview of strength and direction of the linear relationship between two variables. Below heat map illustrates that the strongest positive correlation is between Pregnancies & Age, and Skin Thickness & BMI. Statistically, both correlations exhibit the same strength. Moreover, the analysis demonstrates the strongest negative correlation is between the Pregnancies and Diabetes Pedigree Function indicating there is an inverse relationship.

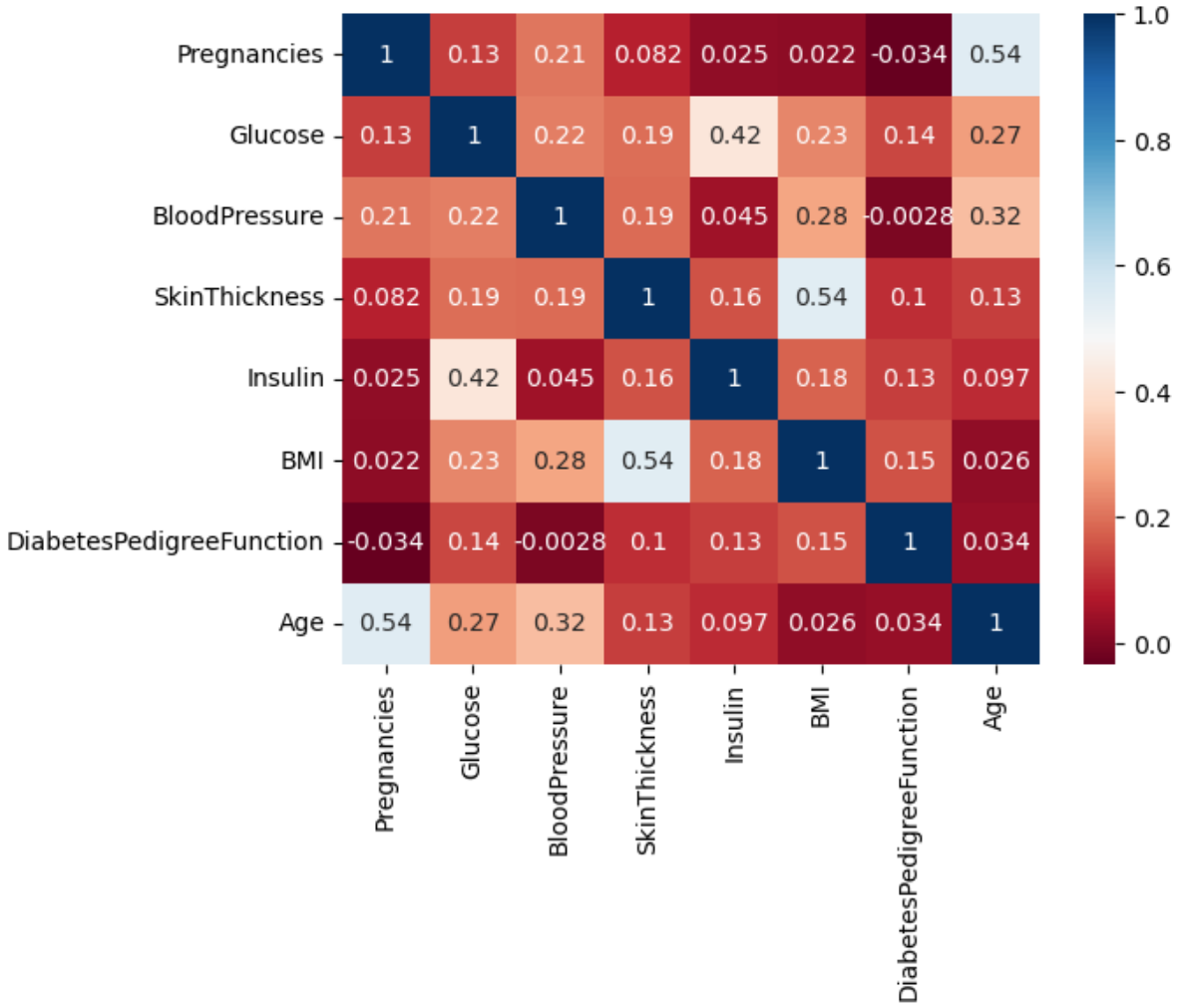

Figure 17 - Heat Map

4. INFERENTIAL STATISTICS

Inferential statistics are used to draw conclusions about the characteristics of a population based on sample data. Hypothesis testing is conducted using statistical methods to determine whether evidence supports or reject the hypothesis.

Inferential Statistic Method	Requirement
Z test	Z test can be used when the population variance is known.
T test	T test can be used when the population variance is unknown.
Chi Square	Chi square can be used when the data is categorical.
ANOVA	Dependent variable must be continuous and independent variables must be categorical.
Regression Analysis	Dependent variable must be continuous.

Table 5 - Inferential Statistical Methods

4.1 Is there an Association between Pregnancy and Diabetes?

To test the above research question, a Chi-Square test of independence will be used. The Chi – Square test is widely used to determine whether there is a significant association between two categorical variables.

In order to carry out a Chi-Squared test, the 'Pregnancies' column will be categorized. A new categorical variable ('Pregnancy History') will be introduced to the data frame, and it will contain 'yes' or 'no' based on their pregnancy history. Patients who have ever been pregnant will be marked as 'yes' and patients who have never been pregnant will be marked as 'no'.

Two categorical variables considered for the test are:

1. Outcome: ('Positive', 'Negative')
2. Pregnancy History: ('Yes', 'No')

Constructing the Hypotheses

- Null hypothesis (H_0): There is no association between pregnancy and diabetes.
- Alternative hypothesis (H_1): There is a significant association between pregnancy and diabetes.

Assumptions

- Observations should be independent of each other.
- The expected frequency for each cell in the contingency table should be at least five.
- The levels of categories and variables are mutually exclusive.

Contingency Table

		Pregnancy History		
		No	Yes	**Total**
Outcome	Negative	73	427	500
	Positive	38	230	268
	Total	111	657	768

Table 6 - Contingency Table

Parameter	**Equation**	**Calculated Value**
Degree of freedom	$dof = (\#rows - 1) \times (\#columns - 1)$	1
Pearson Chi Square Statics	$x^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - \tilde{e}_{ij})^2}{\tilde{e}_{ij}}$	0.002546320599888377
Critical Value	$x^2{}_{(r-1)\times(c-1),\alpha}$	3.841458820694124
Confidence Level	$1 - \alpha = 0.95$	95%
P Value	Probability of observed results occurring by chance	0.9597549634913316

Table 7 - Chi Square Test Summary

Using python packages Chi Square Statistic, P Value and Critical value are calculated.

$x^2 < 3.8414588$

$p\ value > 0.05$

No evidence to reject null hypothesis. **There is no association between pregnancy and diabetes**.

4.2 Is there a significant difference in mean Plasma Glucose Levels between Diabetic and Non-Diabetic patients?

Independent two-sample t-test is a statistical test which is used to compare the means of two independent groups.

Constructing the Hypotheses

- Null hypothesis (H_0): There is no significant difference in mean plasma glucose levels between diabetic and non-diabetic patients
- Alternative hypothesis (H_1): There is a significant difference in mean plasma glucose levels between diabetic and non-diabetic patients

Assumptions

- The data is continuous.
- The sample was randomly sampled from the population.
- The distribution is nearly normal.
- The observations in each group are independent of each other.
- The variances of the two groups are equal.

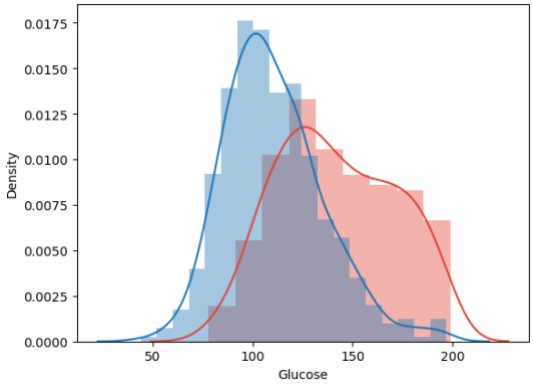

Figure 18 - Glucose Distribution Based on Outcome

Parameter	Equation	Value
T statistic	$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$	15.67989823120835
Confidence Level	$1 - \alpha = 0.95$	95%
P Value	Probability of observed results occurring by chance	2.909251656846331e-48

Table 8 - T test Summary

Using python packages t-Statistic and P Value are calculated.

$P\ value\ <\ 0.05$

The p-value is the probability of obtaining a test statistic as extreme or more extreme than the observed one, assuming the null hypothesis (no difference between the means) is true.

Since p-value is less than 0.05, we reject the null hypothesis and conclude that **there is a statistically significant difference between the means of the two groups**.

4.3 Reason Behind Selecting Chi Square and T test

This dataset and its variables have several limitations. Considering the nature of variables, two possible statistical tests were chi square test and t test.

Reason for not selecting below tests:

Z test	Population parameters are unknown
ANOVA	Dependent variable (‘Outcome’) is not continuous
Regression Analysis	Dependent variable (‘Outcome’) is not numerical

Table 9 - Statistical Tests and Dataset Limitations

4.4 Conclusion

There is no evidence to prove that there is an association between diabetes and pregnancy.

There is a statistical difference of mean glucose levels between diabetic and non-diabetic patients. It is safe to say that plasma glucose level is a critical factor in detecting diabetes while pregnancy is not.

5. FURTHER ANALYSIS POSSIBILITIES

In the previous sections I analysed the association and mean comparison of variables in the data. Evolving from the current analysis, we can explore further analysis possibilities.

- Predictive Models – A predictive model can be developed to identify individuals at high risk of diabetes based on their clinical characteristics. Machine Learning models for diabetes risk assessment have the potential to improve early identification and intervention. From small medical practices to state hospitals, medical practitioners will be able to identify diabetes patients at early stages and provide them with necessary treatments.

- Longitudinal Analysis – Change of a variable over time can be tracked individually and assess the effectiveness of treatment using longitudinal analysis. we can track each patient from the dataset and assess the impact of treatment or lifestyle modification over the time. This will identify the progression of diabetes and effectiveness of treatment strategies.

- Subgroup Analysis – Subgroup the analysis by factors such as age, BMI, etc. to examine potential relationships with diabetes across different subgroups. This could identify specific subgroups that are prone to diabetes and risk factors associated with these groups.

6. DATASET LIMITATION

The dataset only includes female patients who aged between 21 and 81. Hence, the results of this analysis or predictive models built based on this dataset cannot be generalized to other populations such as people of other ethnicities, men, or children.

The conducted survey was based on one hospital. There could be potential for selection bias, meaning that hospital's patient population is not representative of the general Pima Indian Population.

This data was collected during each patient's visit to the hospital. This only includes patient's health data at that time. It is not possible to track changes in risk factors over the time.

Also sample size of 768 is considered as small when assessing health factors that affect mass populations.

7. REFERENCES

- Learning, U.M. (2016) *Pima Indians Diabetes Database*, *Kaggle*. Available at: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database (Accessed: 05 December 2023).
- Chang, V. *et al.* (2022) *Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms*, *Neural computing & applications*. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8943493/ (Accessed: 05 December 2023).
- (No date) *NHS choices*. Available at: https://www.nhs.uk/conditions/diabetes/ (Accessed: 05 December 2023).
- *What is diabetes? - niddk* (no date) *National Institute of Diabetes and Digestive and Kidney Diseases*. Available at: https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes (Accessed: 05 December 2023).