

GITAM

(DEEMED TO BE UNIVERSITY)

**Analysis on Pima using Python**

**Program: BBA -BA**

**Course: DATA ANLYSIS WITH PYTHON**

**Instructor: Dr. M. Kamakshaiah**

**Academic Year:2021-2024**

**GROUP MEMEBERS:**

**PRASAD**

**VEDAHITHA**

**TARUN**

**REETHU AKANKSHA**

**LEELA ADITYA**

**1.INTRODUCTION TO DATA SET**

This Pima dataset is collection of Indian diabetes insights and data variables. This dataset consists of several medical predictor (independent) variables and one target (dependent) variable, Outcome. Independent variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on. Here, we are analysing different aspects of diabetes in the Pima diabetes analysis by doing exploratory data analysis.

The dataset has the following information:

1. Pregnancies — Number of times pregnant
2. Glucose Plasma — glucose concentration 2 hours in an oral glucose tolerance test
3. Blood Pressure — Diastolic blood pressure (mm Hg)
4. Skin Thickness — Triceps skin-fold thickness (mm)
5. Insulin — Two hours of serum insulin (mu U/ml)
6. BMI — Body mass index (weight in kg/(height in m)²)
7. Diabetes Pedigree Function — Diabetes pedigree function
8. Age — Age in years
9. Outcome — Class variable (0 or 1)

The first eight columns represent the independent variables, and the last column denotes the binary dependent variable. There are a total of 768 entries in the dataset. The outcome variable is set to 1 for 268 entries, and the rest are set to 0.

**2.QUESTIONS/PROBLEMS**

**ANALYSIS TASKS**

1. Is there a correlation between BMI and glucose levels, and how does it differ between patients with and without diabetes?
2. What is the distribution of age in the dataset, and how does it differ between patients with and without diabetes?
3. What is the distribution of glucose levels in the dataset, and how does it differ between patients with and without diabetes?
4. Are there any differences in the average number of pregnancies, BMI, and glucose levels between patients with and without diabetes?
5. Correlation matrix plot between all variables

**3.Statistical analysis of the Dataset**

Here we imported Pima India diabetes dataset and named it as Pima. We first import all the required libraries pandas which is for data analysis, NumPy for linear algebra and arrays, matplotlib for plotting and visualization. This data set consists of several medical terms of the human body that related to the diabetes.

We imported the libraries

**CODE**

import numpy as np

import pandas as pd

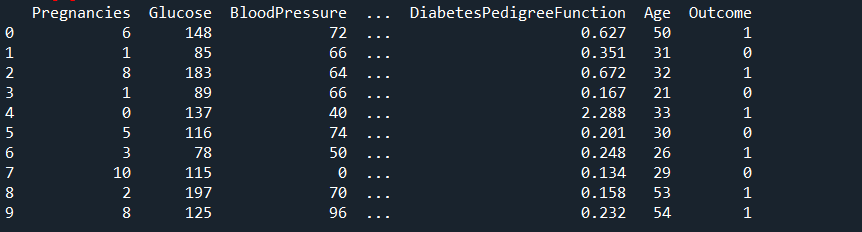
#we import the data

pima = pd.read\_csv("C:/Users/prasa/Downloads/diabetes.csv")

**UNDERSTANDING THE DATA:**

**HEAD OF THE DATA**

pima.head(10)



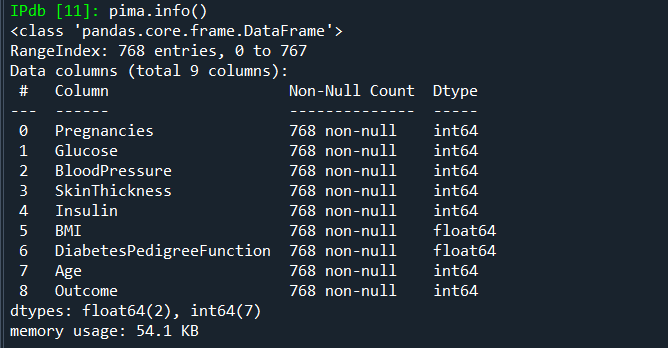
**SHAPE OF THE DATA**

pima.shape

Out [10]: (768, 9)

**INFO OF THE DATA**

**pima.info()**

****

**INPUT:**

**col\_idx = pima.columns**

**col\_idx**

**Out [12]:**

**Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',**

**'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],**

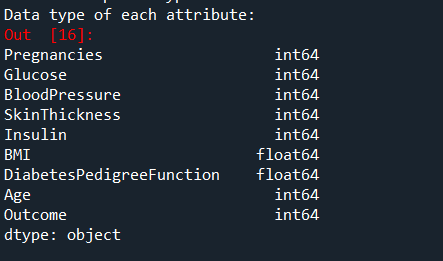
**dtype='object')**

**DATA TYPE OF THE ATTRIBUTES:**

**# Find data type for each attribute**

**print("Data type of each attribute:")**

**pima.dtypes**

****

**Summarisation of Data**

Exploratory Data Analysis of Pima India diabetes analysis. This data is about different medical terms which related to diabetes and with attributes related to human body terms

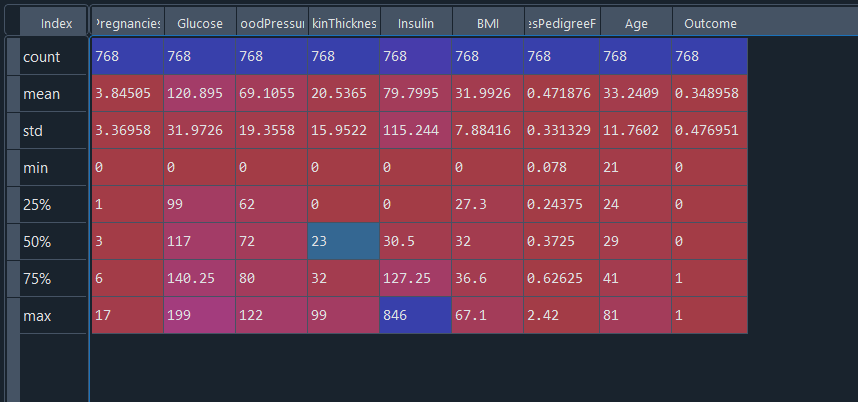
Industry: Medical

Dimensions: 768 observations, 9 variables.

Missing Data / Null Values: No

Duplicated Values: No

Structure: Data frame



**4.Analysis and Insights**

**1. Correlation matrix plot between all pairs of attributes**

**CODE:**

#correlation matrix

# import required package

import numpy as np

# plot correlation matrix

fig = pyplot.figure()

ax = fig.add\_subplot(111)

cax = ax.matshow(correlations, vmin=-1, vmax=1)

fig.colorbar(cax)

ticks = np.arange(0,9,1)

ax.set\_xticks(ticks)

ax.set\_yticks(ticks)

names = pima.columns

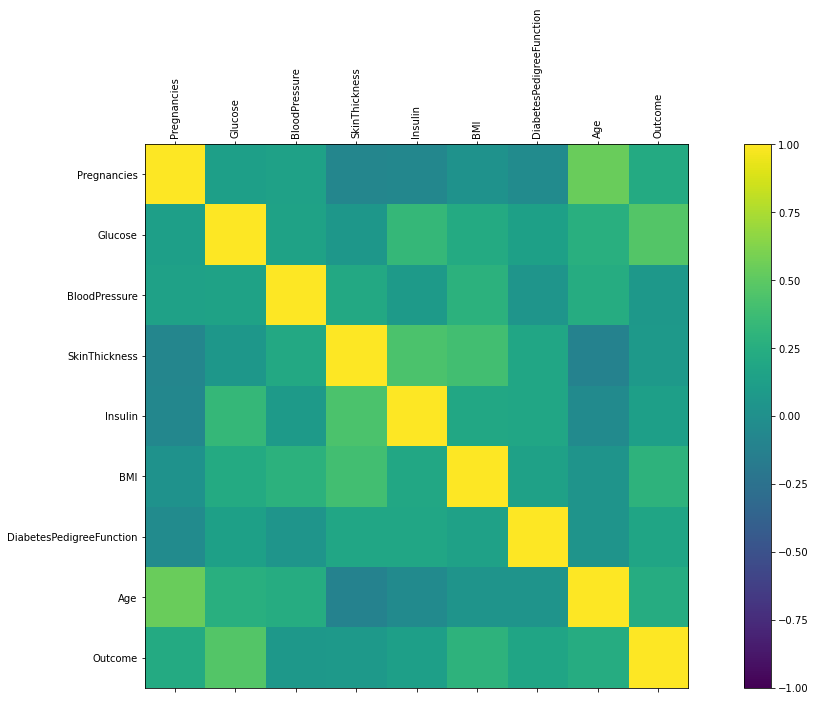
# Rotate x-tick labels by 90 degrees

ax.set\_xticklabels(names,rotation=90)

ax.set\_yticklabels(names)

pyplot.show()

**OUTPUT:**



**INSIGHTS:**

* The diagonal of the correlation matrix shows the correlation between each attribute and itself, which is always equal to 1.0. Therefore, the diagonal has a uniform colour throughout the plot.
* The darkest colours on the plot represent the highest positive correlations, while the lightest colours represent the highest negative correlations. For example, we can see that glucose and age have a relatively strong positive correlation, while skin thickness and BMI have a weak positive correlation.
* The plot can help identify multicollinearity between different attributes. For example, we can see that age and pregnancies have a strong positive correlation, which could indicate that they are measuring similar aspects of a patient's health. This could lead to overfitting if both attributes are included in a predictive model. Therefore, it may be necessary to remove one of these attributes to avoid redundancy.

**2. Is there a correlation between BMI and glucose levels, and how does it differ between patients with and without diabetes?**

**CODE:**

# Correlation between BMI and glucose levels

sns.scatterplot(data=pima, x='BMI', y='Glucose', hue='Outcome')

pyplot.show()

# Correlation between BMI and glucose levels for patients with and without diabetes

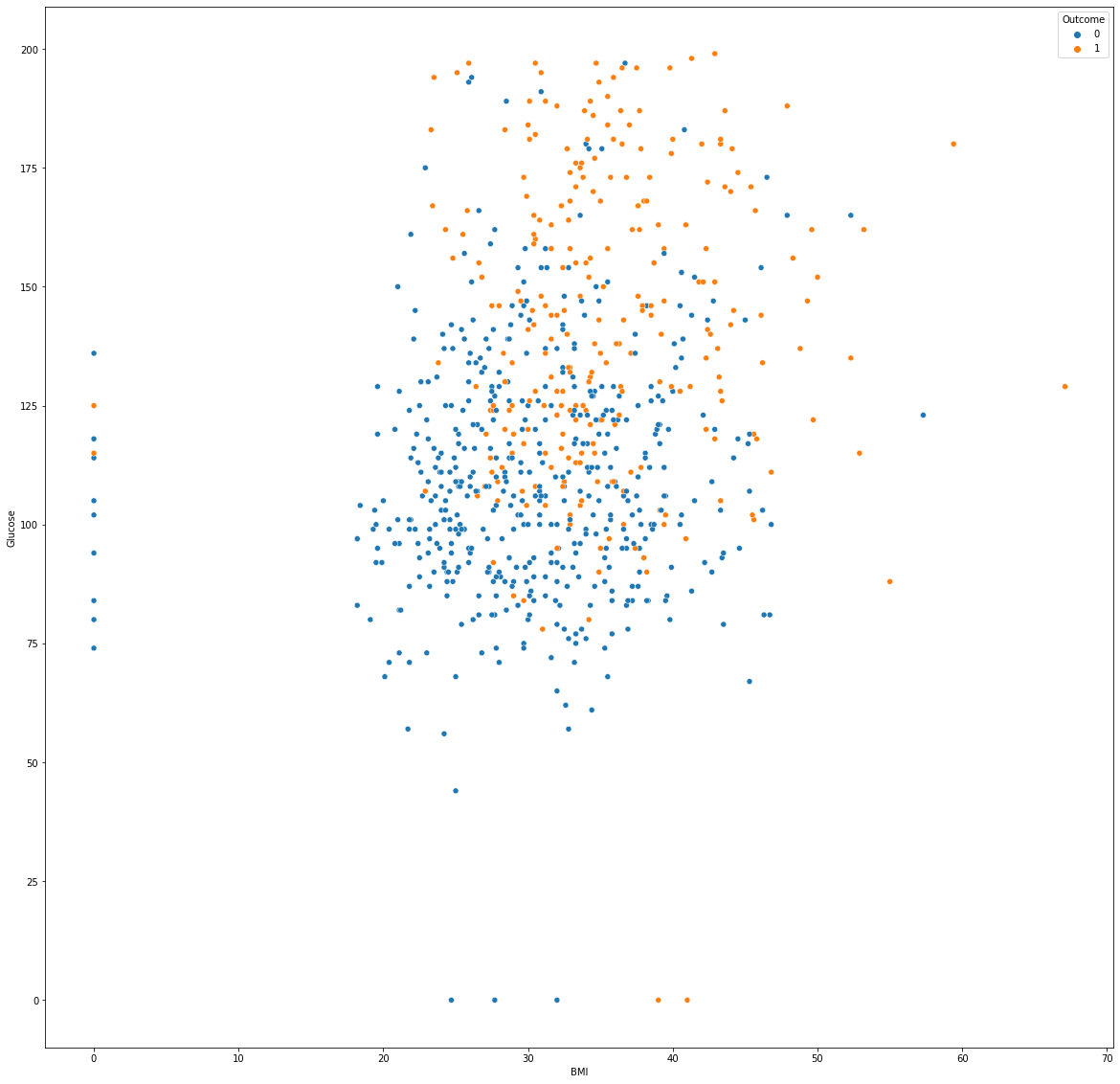
sns.scatterplot(data=pima, x='BMI', y='Glucose', hue='Outcome')

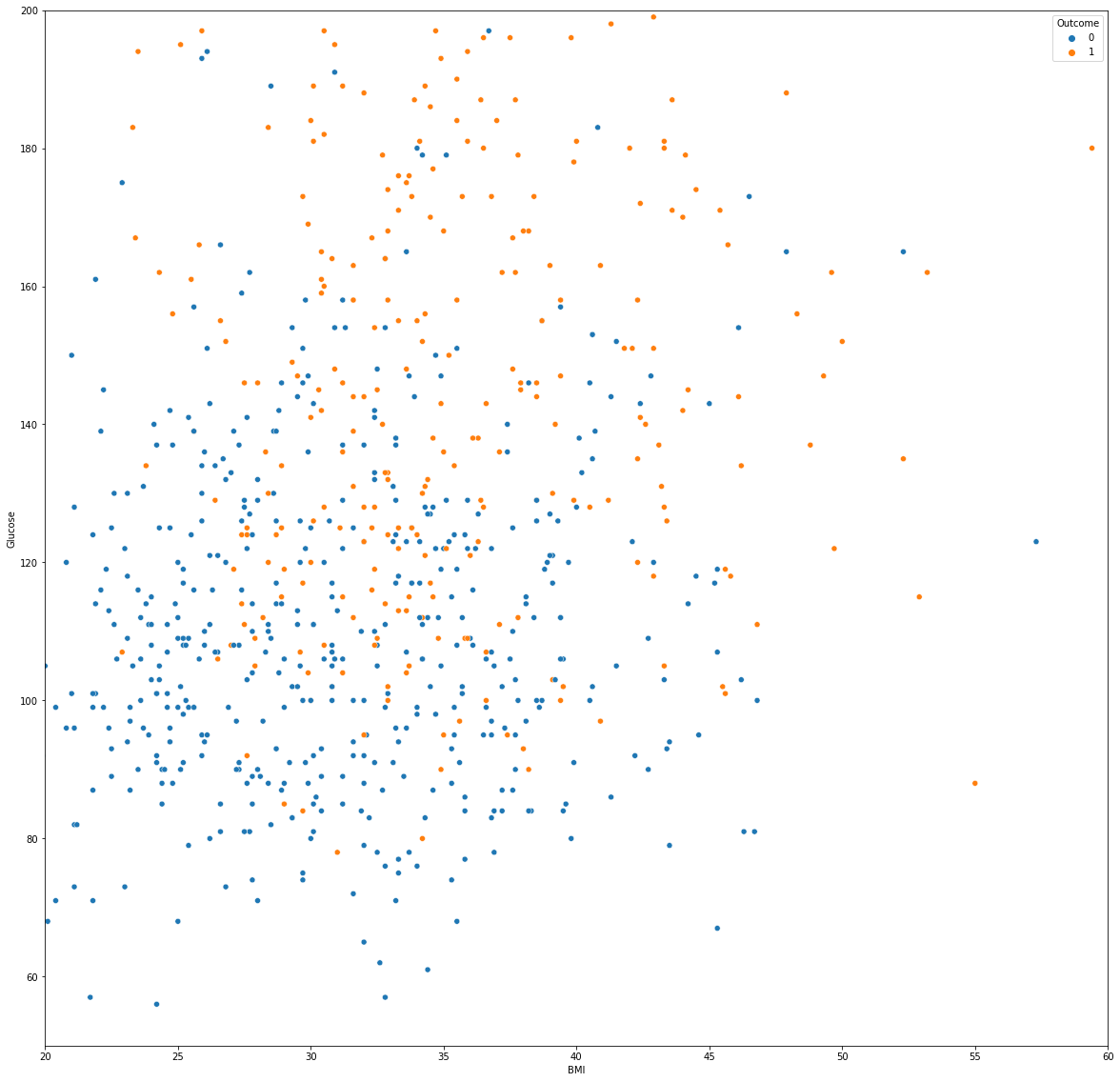
pyplot.xlim(20, 60)

pyplot.ylim(50, 200)

pyplot.show()

OUTPUT:





**INSIGHTS:**

* There is a moderate positive correlation between BMI and glucose levels in the dataset.
* The correlation between BMI and glucose levels is stronger for patients with diabetes than for patients without diabetes.
* Patients with diabetes tend to have higher BMI and glucose levels than patients without diabetes, and the correlation between BMI and glucose levels is stronger for these patients.

**3.** **What is the distribution of glucose levels in the dataset, and how does it differ between patients with and without diabetes?**

**CODE:**# Distribution of glucose levels

sns.histplot(data=pima, x='Glucose', kde=True)

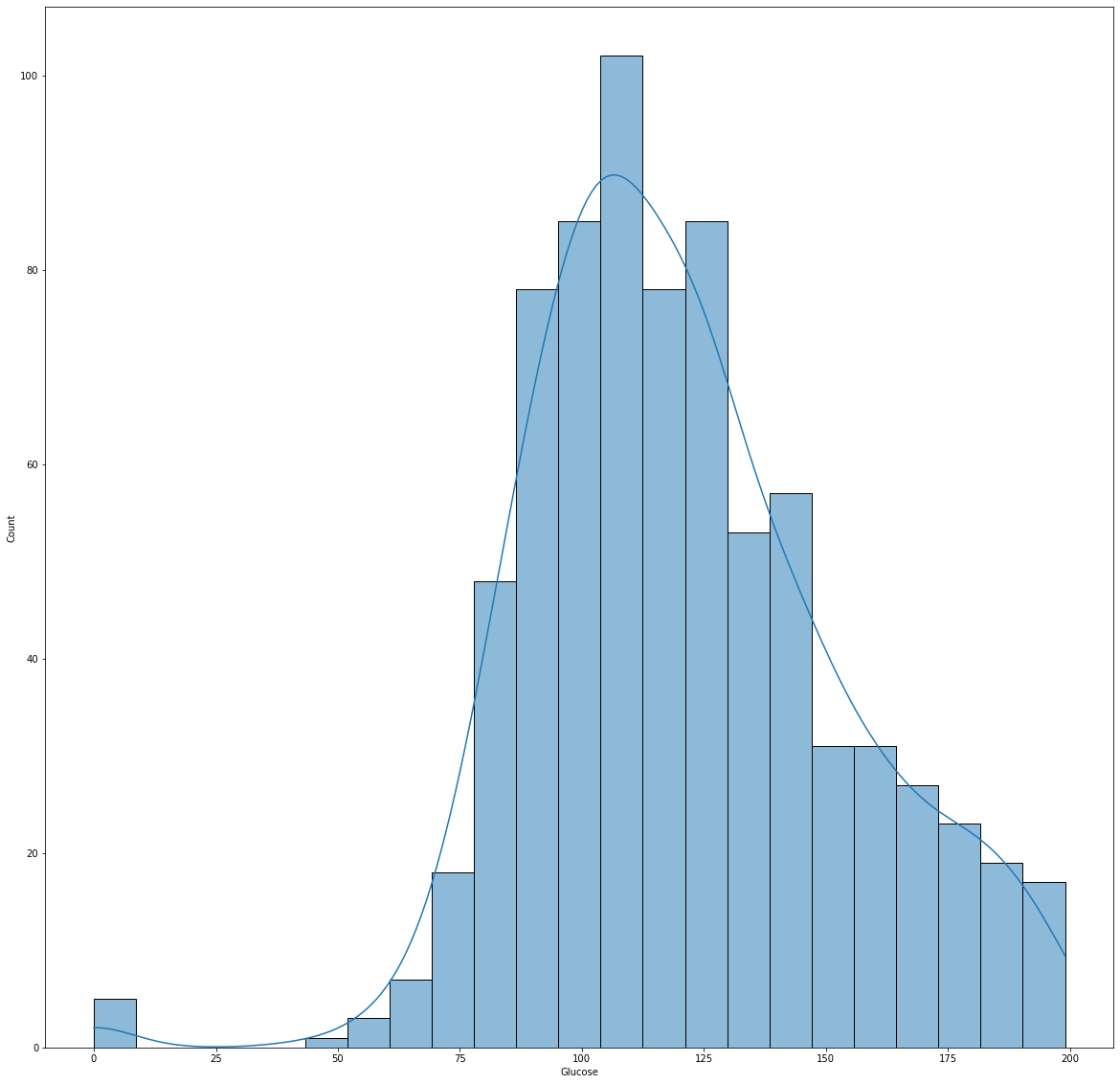
pyplot.show()

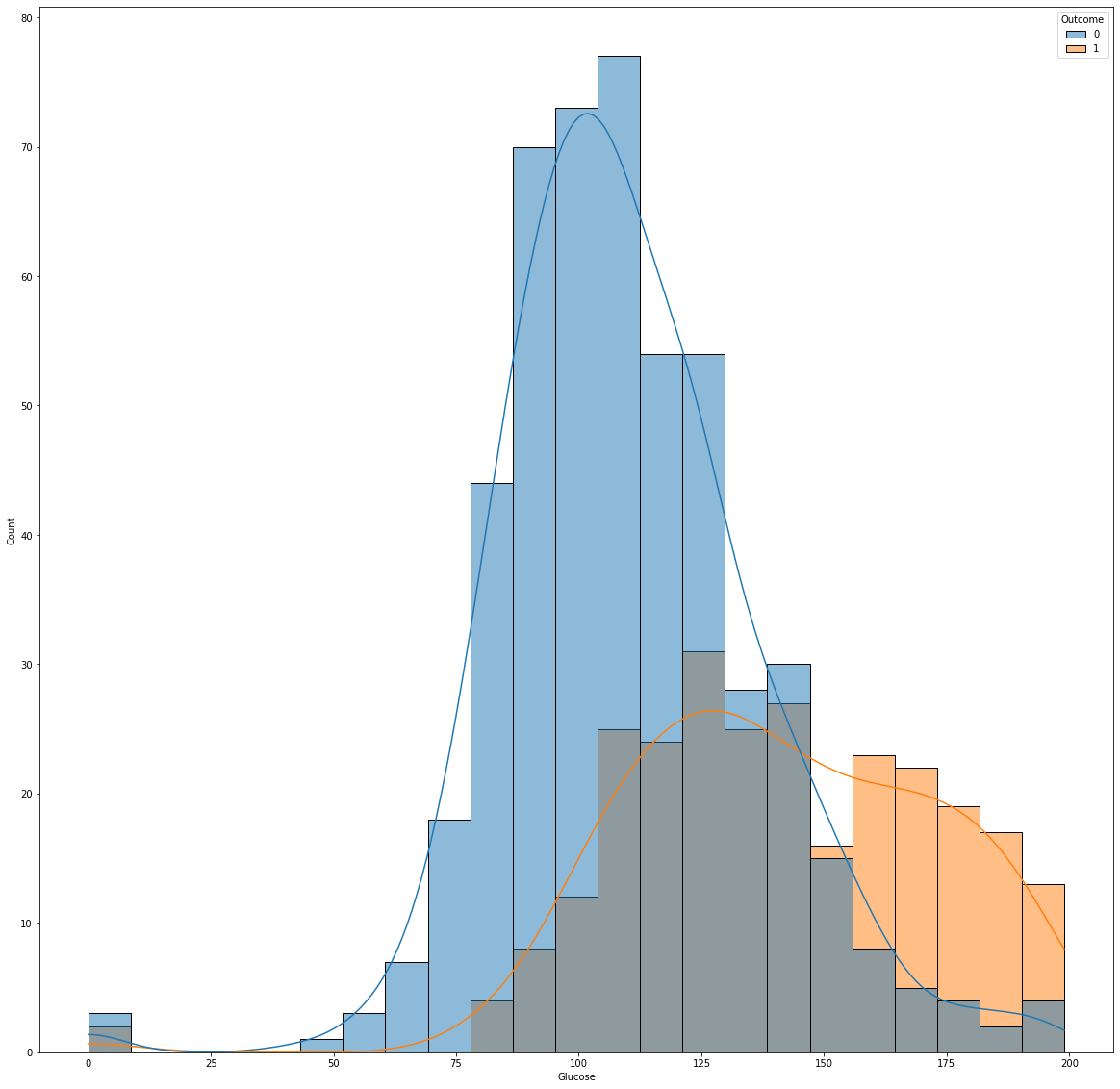
# Distribution of glucose levels by diabetes status

sns.histplot(data=pima, x='Glucose', hue='Outcome', kde=True)

pyplot.show()

**OUTPUT:**





**INSIGHTS:**

* The distribution of glucose levels in the dataset is right-skewed, with most patients having glucose levels between 80 and 140 mg/dL.
* Patients with diabetes tend to have higher glucose levels than patients without diabetes, with a larger proportion of patients having glucose levels above 140 mg/dL.
* The distribution of glucose levels for patients with diabetes is more spread out than that of patients without diabetes.

**4. Are there any differences in the average number of pregnancies, BMI, and glucose levels between patients with and without diabetes?**

**CODE:**

# Differences in average number of pregnancies, BMI, and glucose levels by diabetes status

sns.barplot(data=pima, x='Outcome', y='Pregnancies')

plt.show()

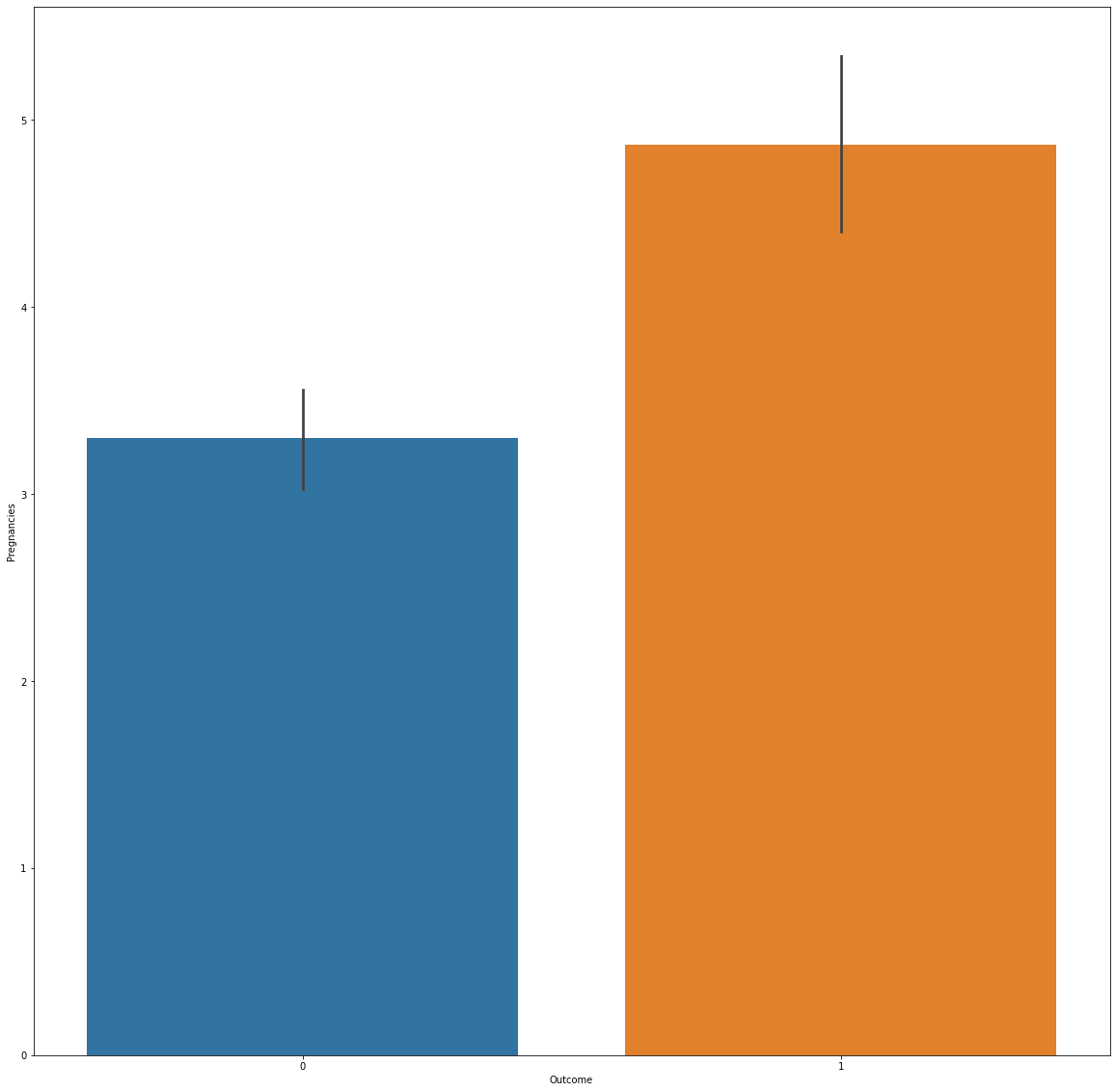
sns.barplot(data=pima, x='Outcome', y='BMI')

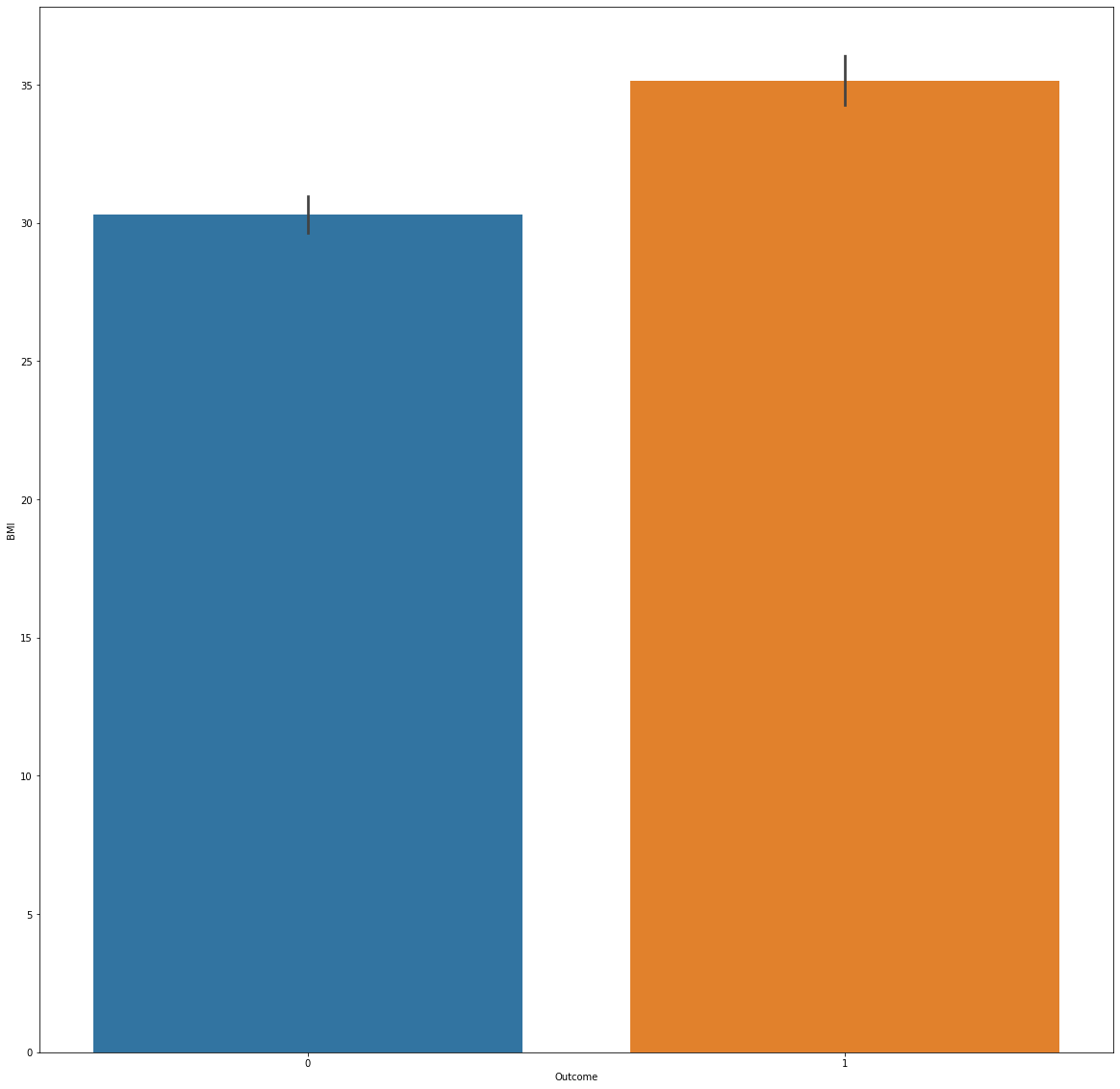
pyplot.show()

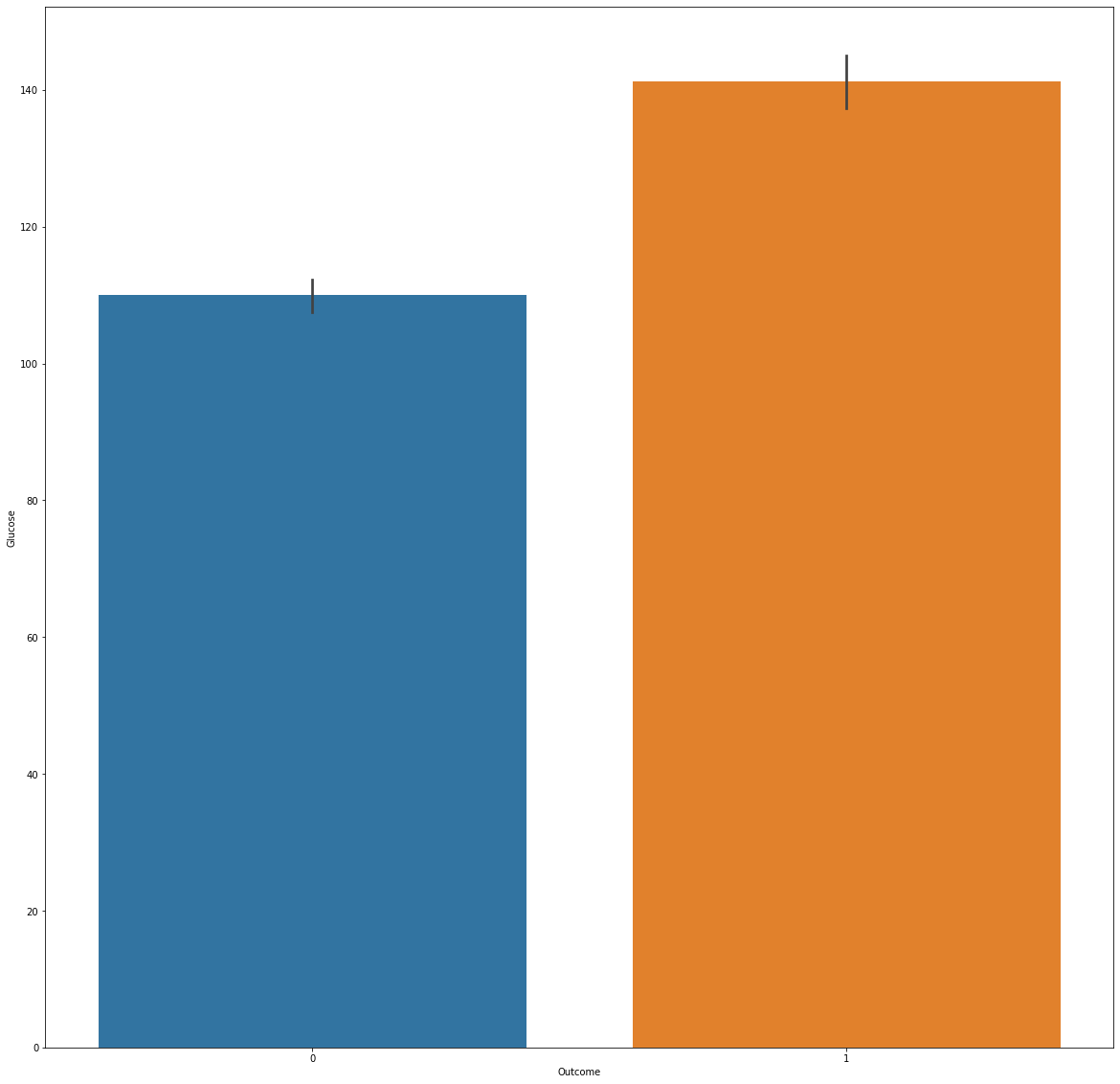
sns.barplot(data=pima, x='Outcome', y='Glucose')

pyplot.show()

**OUTPUT:**







**INSIGHTS:**

* Patients with diabetes tend to have a higher average number of pregnancies than patients without diabetes.
* Patients with diabetes tend to have a higher average BMI than patients without diabetes.
* Patients with diabetes tend to have a higher average glucose level than patients without diabetes.

**FINDINGS:**

1. The dataset contains information about female patients of Pima Indian heritage, with various health measures including glucose levels, BMI, and insulin levels.
2. A significant proportion of the patients in the dataset have been diagnosed with diabetes.
3. The dataset includes missing values, which need to be handled carefully during analysis and modeling.
4. Some health measures, such as glucose levels and BMI, are correlated with the likelihood of diabetes diagnosis.
5. Age and number of pregnancies are also correlated with diabetes diagnosis, but to a lesser extent.
6. Overall, the dataset can be used to develop predictive models for diabetes diagnosis based on various health measures.