Week 1

In [1]:
```
#Import necessary libraries
import pandas as pd
import numpy as np
import plotly.express as px
import matplotlib.pyplot as plt
import seaborn as sns
import opendatasets as od
```

In [2]:
```
#Download the diabtes dataset from kaggle
od.download("https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database?select
```

```
Skipping, found downloaded files in ".\pima-indians-diabetes-database" (use force=True t
o force download)
```

In [3]:
```
#Read csv into python dataframe
diabetes_df = pd.read_csv("pima-indians-diabetes-database/diabetes.csv")
diabetes_df.head(5)
```

Out[3]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

1. Write a summary of your data and identify at least two questions to explore visually with your data.

The diabetes data in this dataset is from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to understand the presence of significant association between diabetes and other individual health conditions.

Following are the features in the data -

- Pregnancies: Number of times pregnant.
- Glucose: Plasma glucose concentration
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/(height in m)^2)
- DiabetesPedigreeFunction: Diabetes pedigree function
- Age: Age (years)
- Outcome: Class variable (0 or 1) for diabetes

Questions: Following are a few questions to explore with the diabetes dataset.

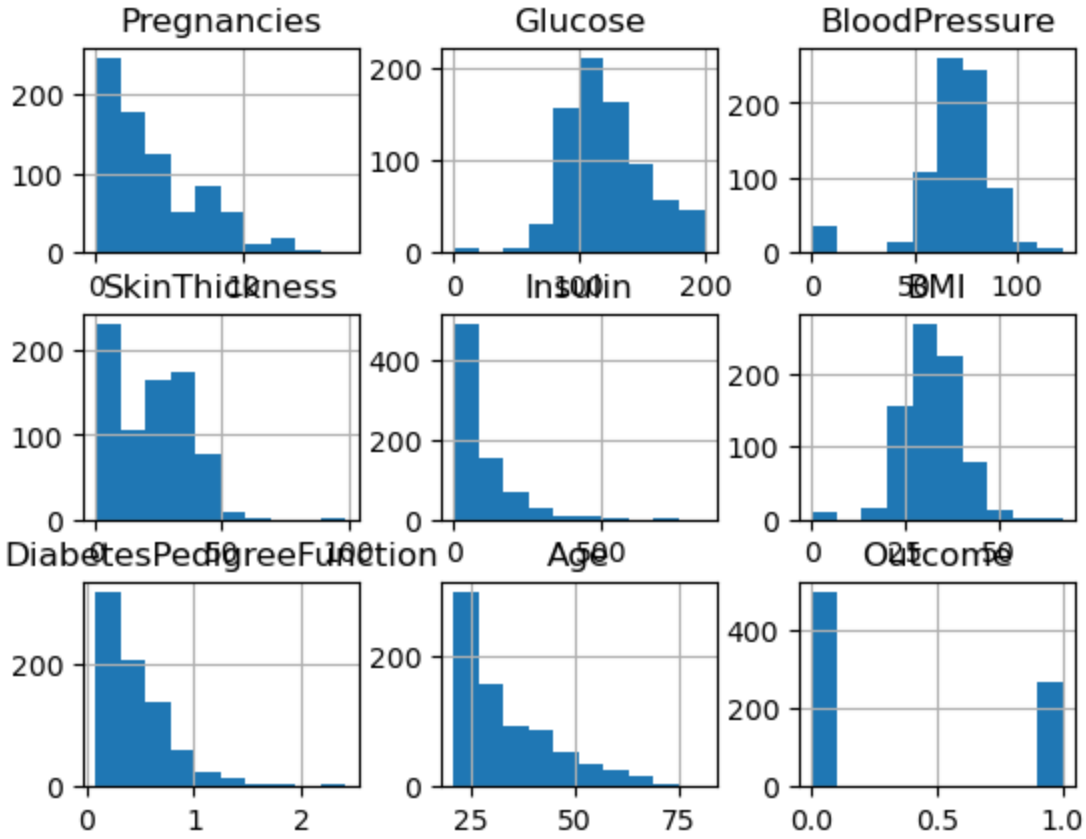
1. Impact of BMI on Diabtese
2. Impact of Age on Diabetes

3. Impact of Blood Pressure on Diabetes
4. Impact of Skin Thickness on Diabetes

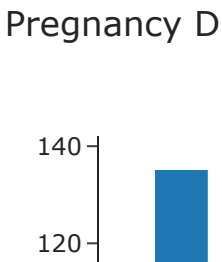
2. Create a histogram or bar graph from your data.

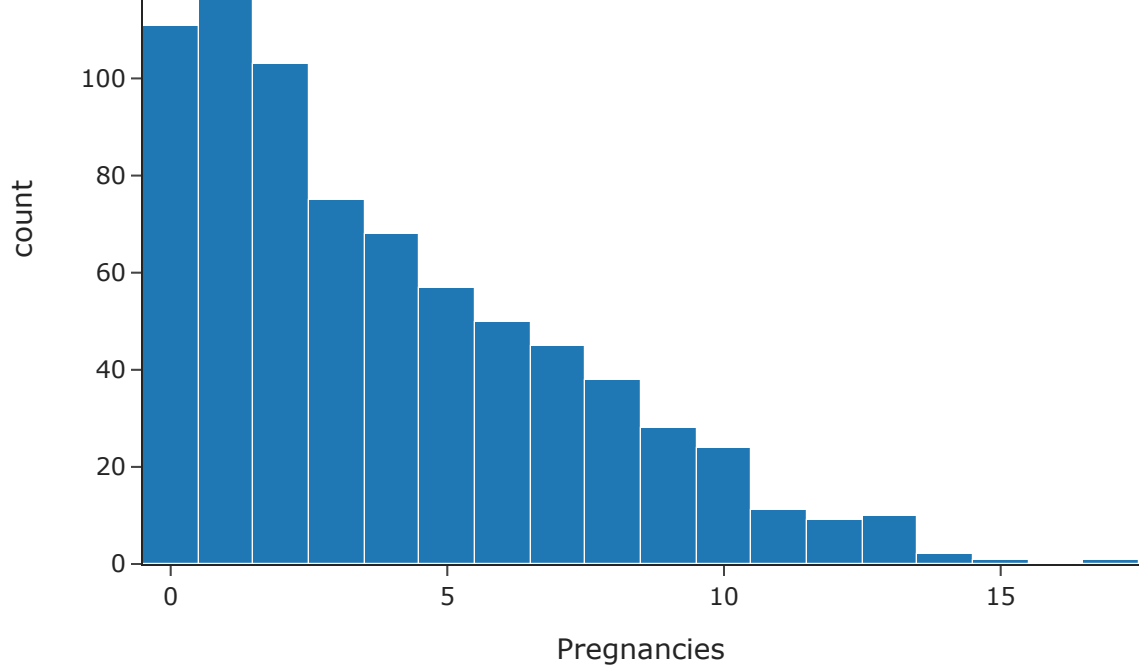
In [4]:
```
# Histogram for all  features in the dataframe.
plt.figure().set_figwidth(15)
diabetes_df.hist()
```

Out[4]:
```
array([[<Axes: title={'center': 'Pregnancies'}>,
        <Axes: title={'center': 'Glucose'}>,
        <Axes: title={'center': 'BloodPressure'}>],
       [<Axes: title={'center': 'SkinThickness'}>,
        <Axes: title={'center': 'Insulin'}>,
        <Axes: title={'center': 'BMI'}>],
       [<Axes: title={'center': 'DiabetesPedigreeFunction'}>,
        <Axes: title={'center': 'Age'}>,
        <Axes: title={'center': 'Outcome'}>]], dtype=object)
<Figure size 1500x480 with 0 Axes>
```



In [5]:
```
#Histogram on Pregnancies
fig = px.histogram(diabetes_df, x="Pregnancies", template="simple_white",
                   title="Pregnancy Distribution")
#fig.show('notebook')
fig.show()
```

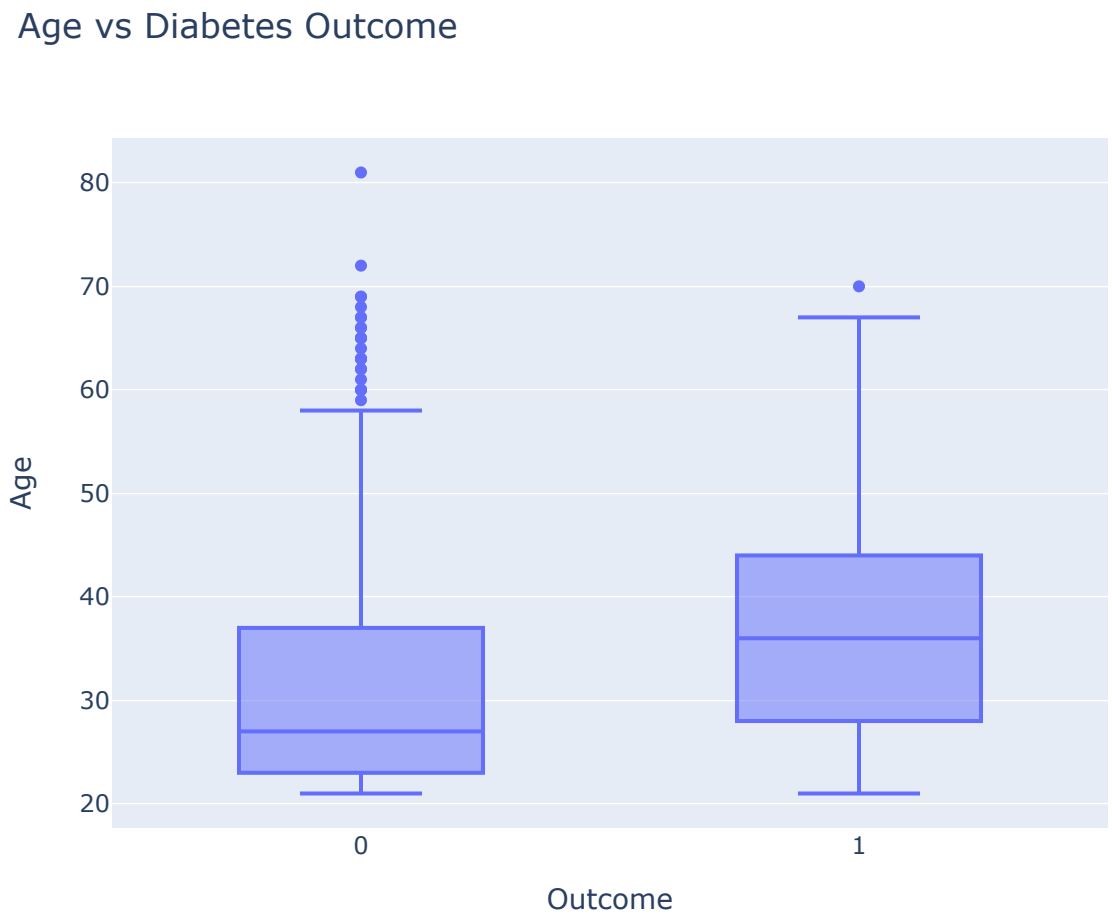




> The above chart shows the total count by pregnancy number

3. Create a boxplot from your data.

In [6]:
```
#Boxplot on Age and Outcome
fig = px.box(diabetes_df, y="Age", x="Outcome", title="Age vs Diabetes Outcome")
fig.show()
```

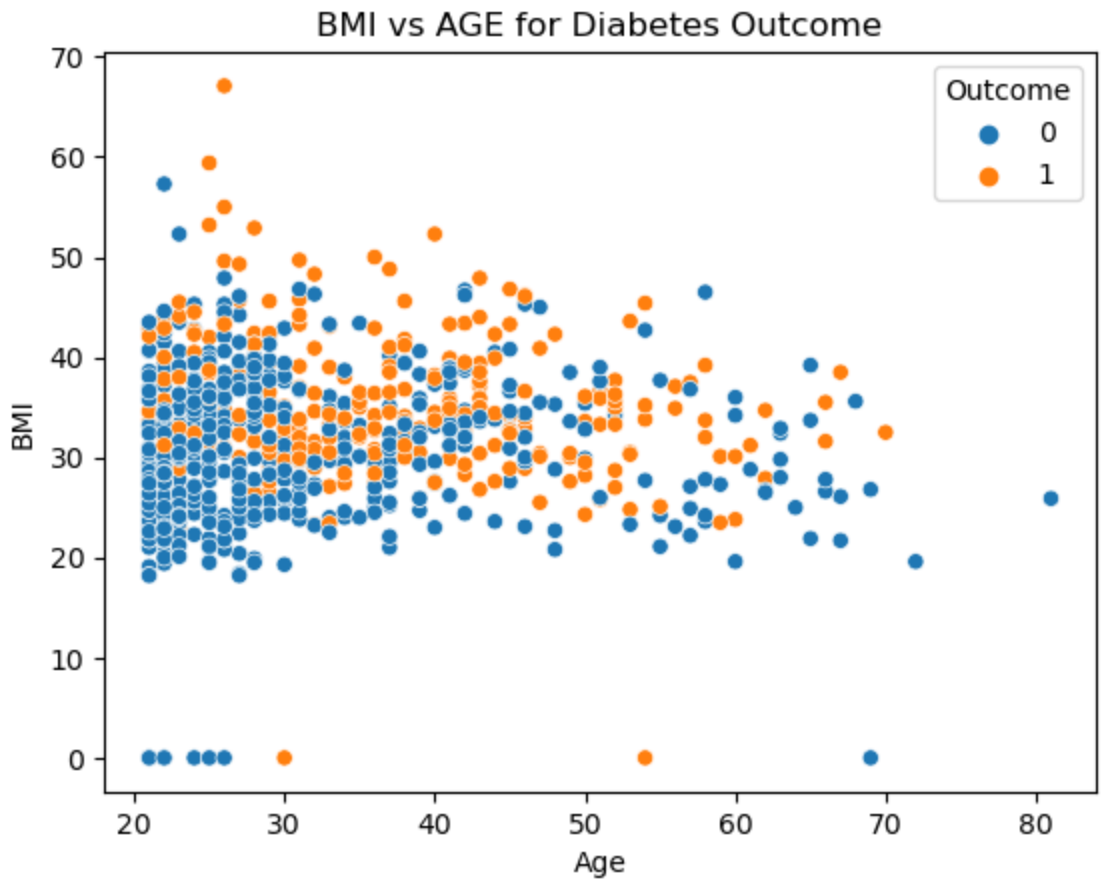


> From the above chart we can see the data distribution for diabetic and non-diabetics by age.

4. Create a bivariate plot from your data.

In [7]:
```
#Bivariate Scatter plot on Age and BMI
sns.scatterplot(data=diabetes_df, x="Age", y="BMI", hue="Outcome")
plt.title("BMI vs AGE for Diabetes Outcome")
```

Out[7]:
```
Text(0.5, 1.0, 'BMI vs AGE for Diabetes Outcome')
```

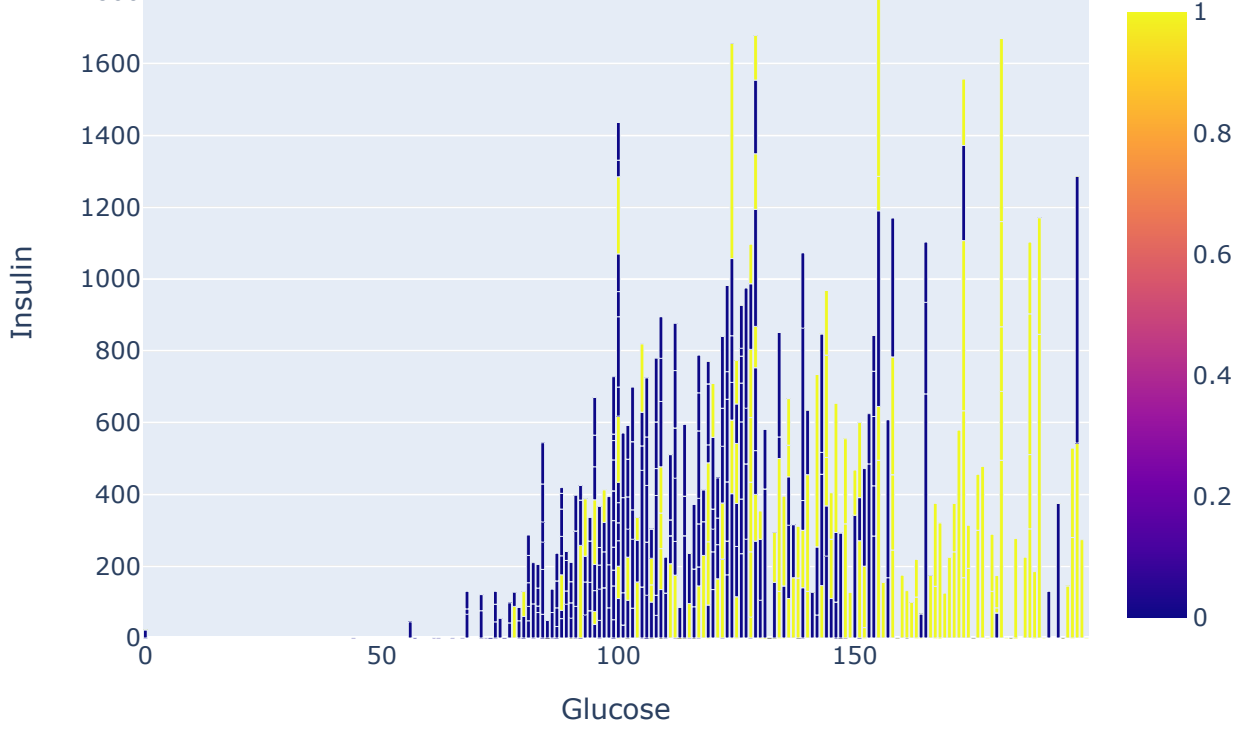


> From the above chart we can see that most of the data lies between the age 20 to 50.
> Diabetes is more common across people with a higher BMI (>30) and age between 20 to 50.

5. Create any additional visualizations that will help to answer the question(s) you want to answer.

In [16]:
```
#BAR chart on Glucose and Insulin levels
fig = px.bar(diabetes_df, x='Glucose', y='Insulin',color="Outcome",title="Glucose vs Ins
fig.show()
```

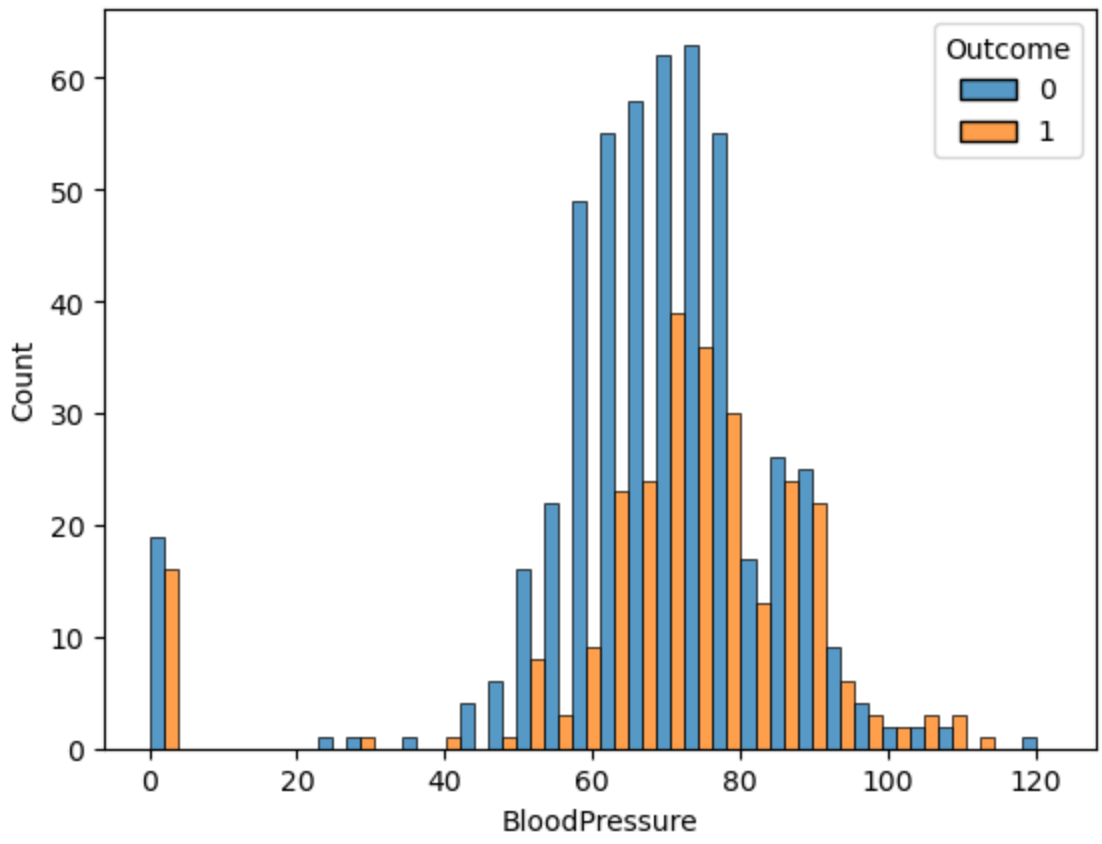
Glucose vs Insulin levels



> Ideally, for a glucose level less than 100, a patient is said to have normal blood sugar levels. From this dataset we can see that people with normal glucose levels are also diabetic, implying the normal levels could be a result of insulin usage.

In [9]:
```
#Histogram on BloodPressure and Outcome
sns.histplot(data=diabetes_df, x="BloodPressure", hue="Outcome", multiple="dodge")
```

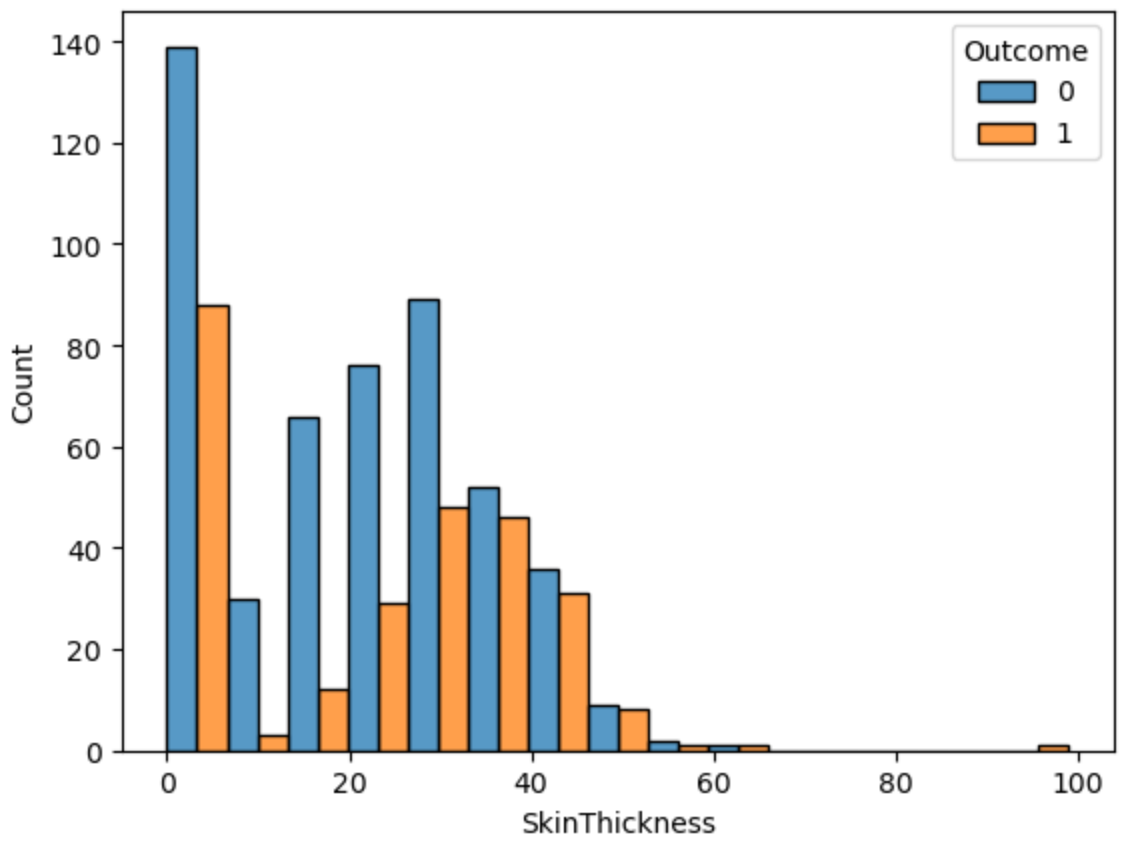
Out[9]:
```
<Axes: xlabel='BloodPressure', ylabel='Count'>
```



> From the dataset, Diastolic blood pressure levels between 60 to 90 have more cases of diabetes

In [10]:
```
#Histogram on SkinThickness and Outcome
sns.histplot(data=diabetes_df, x="SkinThickness", hue="Outcome", multiple="dodge")
```

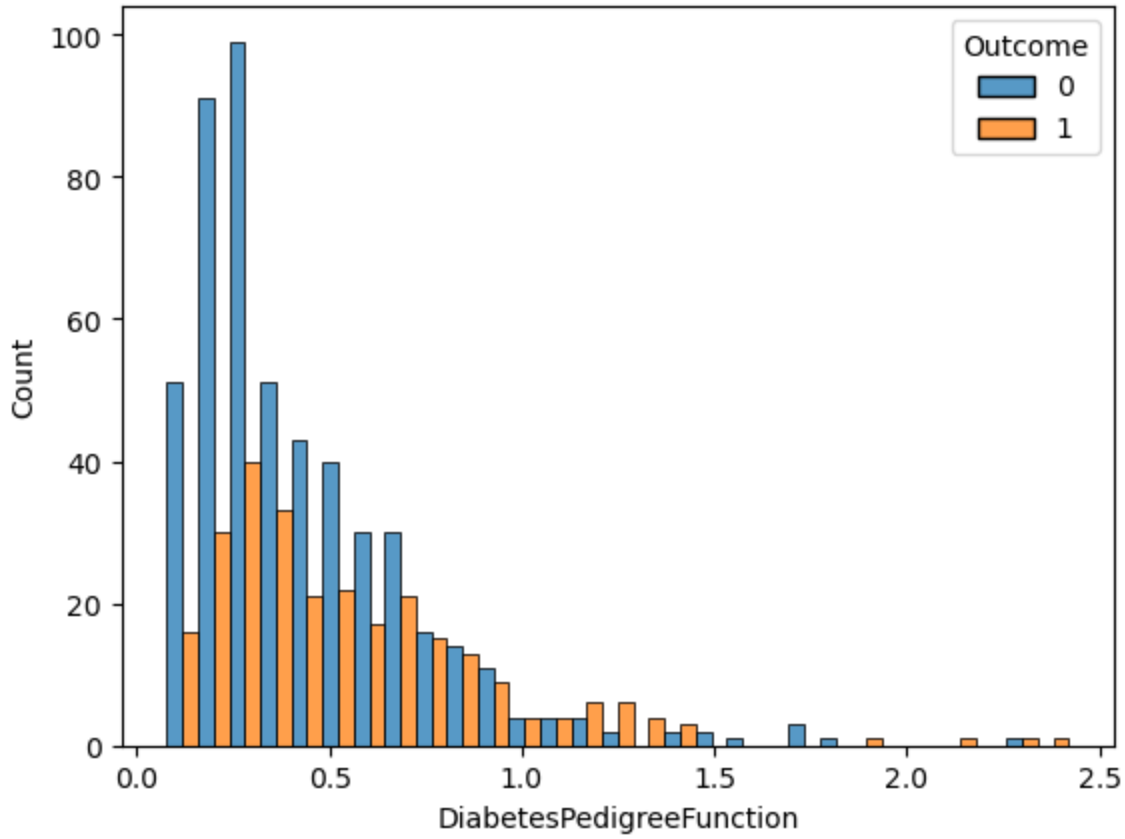
Out[10]:
```
<Axes: xlabel='SkinThickness', ylabel='Count'>
```



> People with diabetes develop tight, thick, waxy skin on the backs of their hands. Triceps skin fold thickness (mm) is said to have a significance in diabetes, but this is not clear in the dataset.

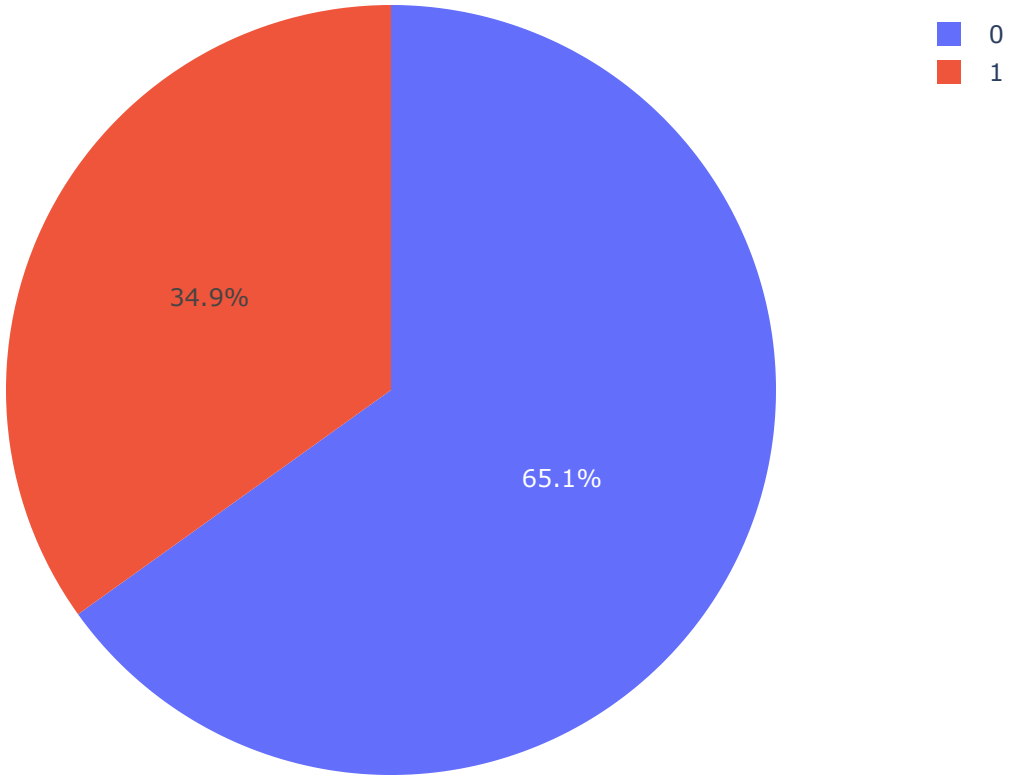
In [11]:
```
#Histogram on DiabetesPedigreeFunction and Outcome
sns.histplot(data=diabetes_df, x="DiabetesPedigreeFunction", hue="Outcome", multiple="do
```

Out[11]:
```
<Axes: xlabel='DiabetesPedigreeFunction', ylabel='Count'>
```



> Hereditary diabetes is another common cause for diabetes. From the above chart, we can see that most of the diabetic cases fall between 0 to 0.7 DiabetesPedigreeFunction score.

In [14]:
```
#Pie chart for outcome distribution
px.pie(diabetes_df, names='Outcome')
```



> The dataset has around 65.1% non-diabetic and 34.9% diabetic data.

6. Summarize your results and make a conclusion. Explain how you arrived at this conclusion and how your visualizations support your conclusion.

Data visualizations for the pima indians diabetes dataset from kaggle are as follows:

1. Histogram on Pregnancies - This chart depicts the total count by pregranancy numbers in the dataset
2. Boxplot on Age and Outcome - The boxplot shows that people aged 35 and older are at a higher risk of diabtes.
3. Bivariate Scatter plot - In this chart we can see that people with BMI over 30 are at a higher risk of diabetes.
4. Bar chart on Glucose and Insulin - From this bar chart, it is unclear if a higher glucose results in higher units of insulin.
5. Histogram on BloodPressure and Outcome - This chart shows that the number of diabetic outcomes is higher for a lower Diastolic blood pressure.
6. Histogram on SkinThickness and Outcome - Skinthickness does not have a glaring impact on diabetes.
7. Histogram on DiabetesPedigreeFunction and Outcome - Although it is known that diabetes has a stronger link to family history and lineage, I am not seeing this association clearly in the dataset.
8. Pie Chart - The dataset has around 65.1% non-diabetic and 34.9% diabetic data.

CONCLUSION

Based on the charts, we can say that BMI has a significant effect on the possibility of getting diabetes. Age also has some significance on the possibilities of getting diabetes.

There is not any glaring impact of skinthickness and diabetes pedigree function on diabetes.

Diabetes seems to be more of a lifestyle and age related disease.