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 predict the factors likely to cause Diabetes.

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)

Questions: What are the factors most likely to cause Diabetes?

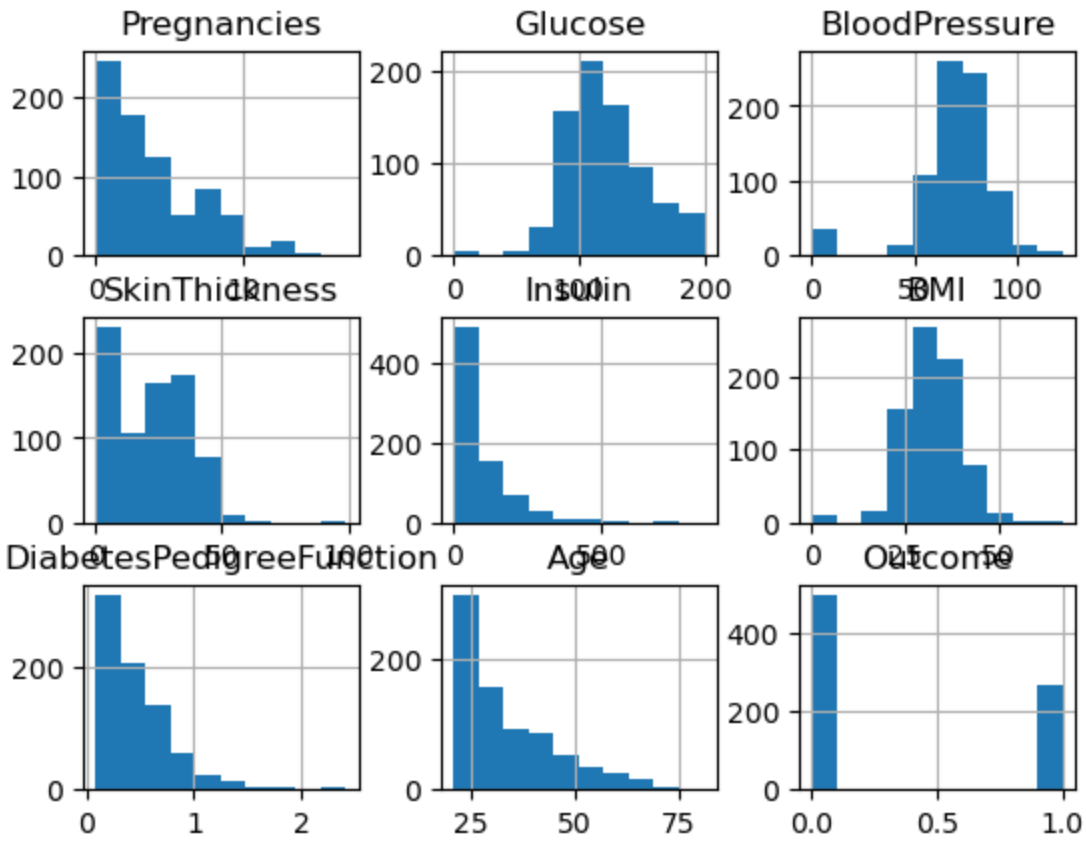
- Impact of BMI on Diabtese
- Impact of Age on Diabetes
- Impact of Blood Pressure on Diabetes

- Impact of Skin Thickness on Diabetes

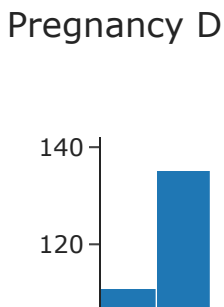
2. Create a histogram or bar graph from your data.

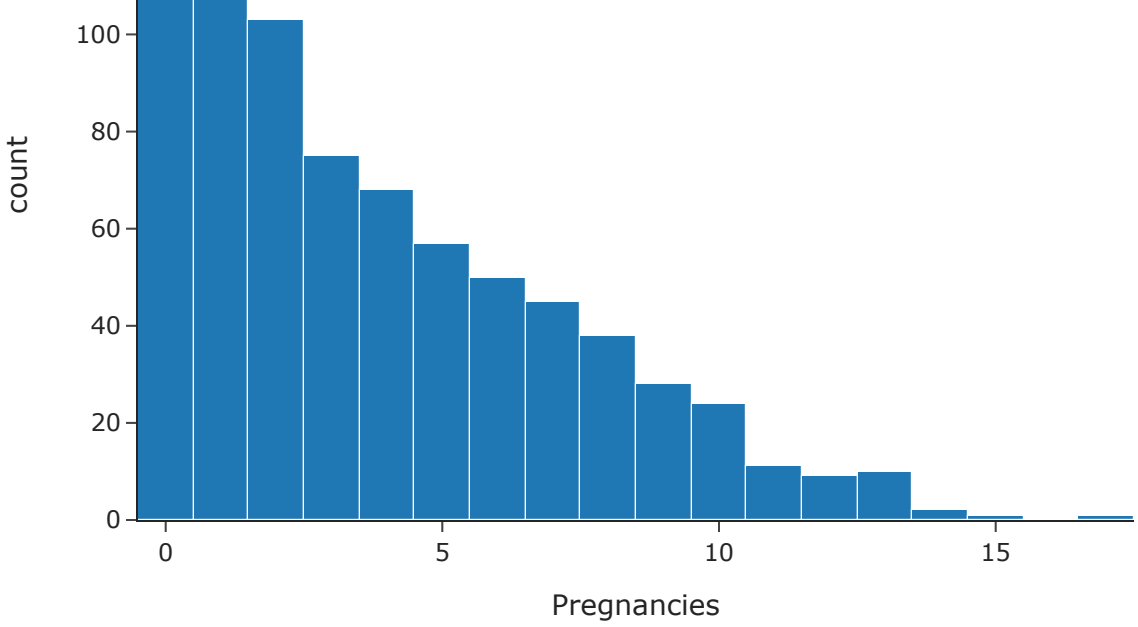
In [4]:
```
# Histogram for all  features in the dataframe.
plt.figure().set_figwidth(15)
diabetes_df.hist()
```

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>
```



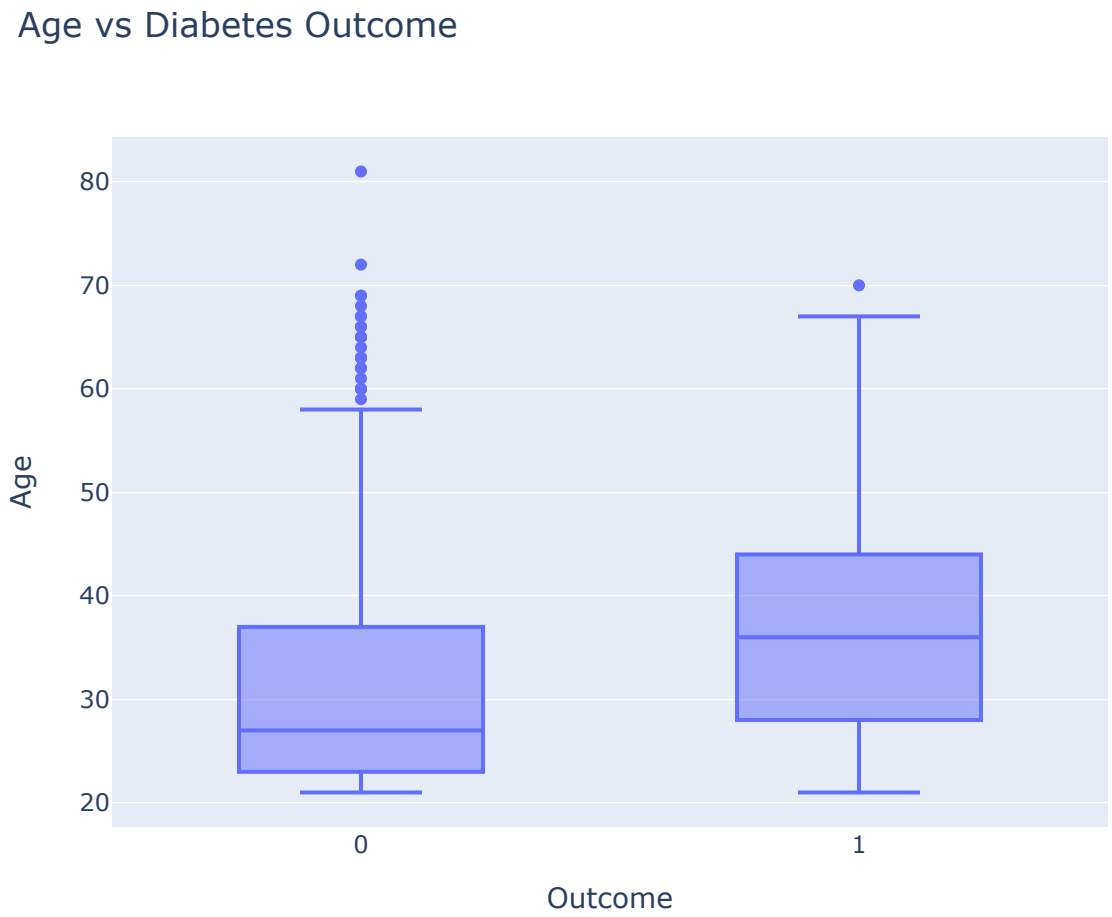
In [5]:
```
#Histogram on Pregnancies
fig = px.histogram(diabetes_df, x="Pregnancies", template="simple_white",
                   title="Pregnancy Distribution")
#fig.show('notebook')
fig.show()
```





3. Create a boxplot from your data.

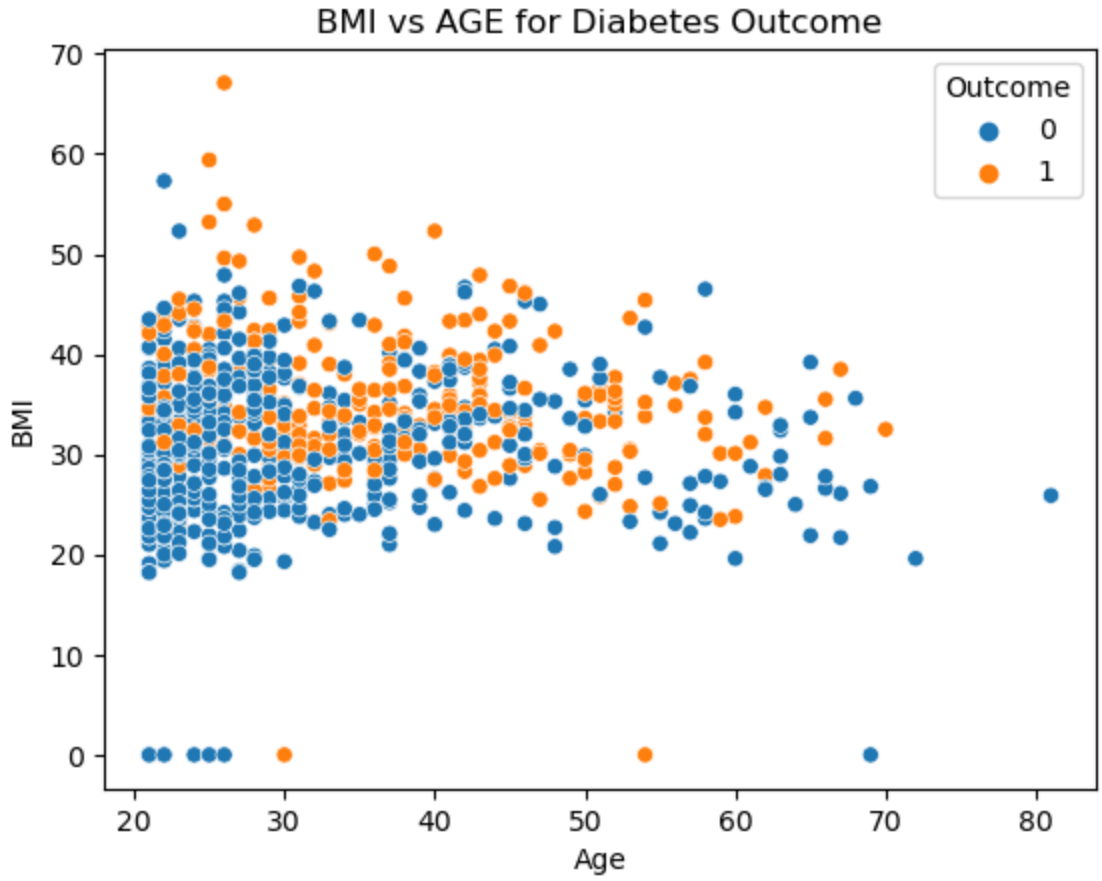
In [6]:
```
#Boxplot on Age and Outcome
fig = px.box(diabetes_df, y="Age", x="Outcome", title="Age vs Diabetes Outcome")
fig.show()
```



4. Create a bivariate plot from your data.

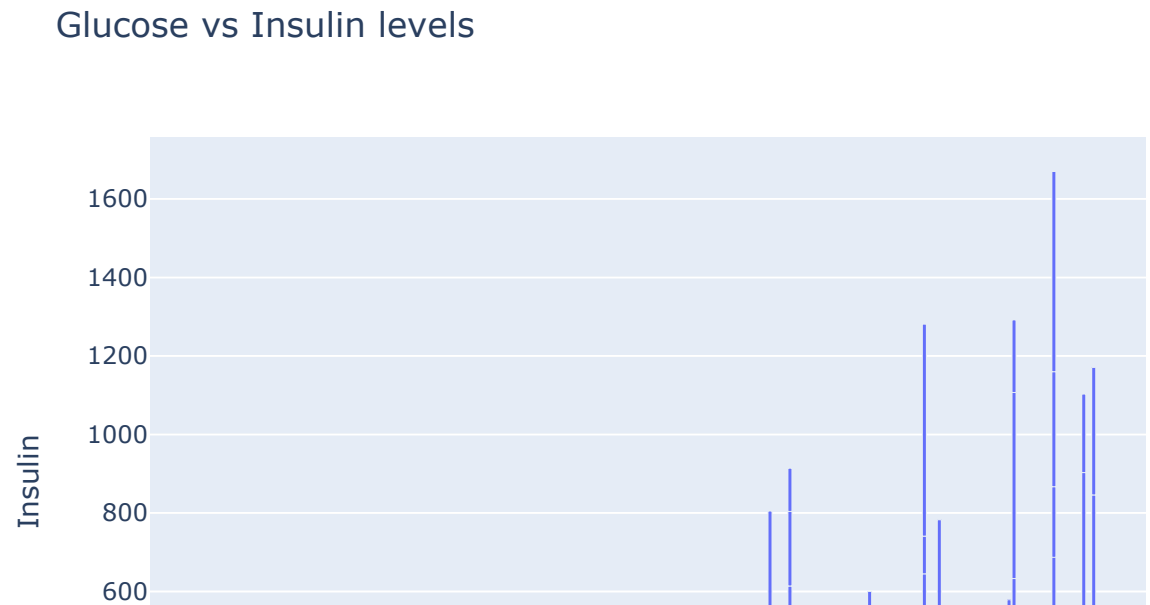
In [7]:
```
#Bivariate Scatter plot on Age and BMI
sns.scatterplot(data=diabetes_df, x="Age", y="BMI", hue="Outcome")
plt.title("BMI vs AGE for Diabetes Outcome")
```

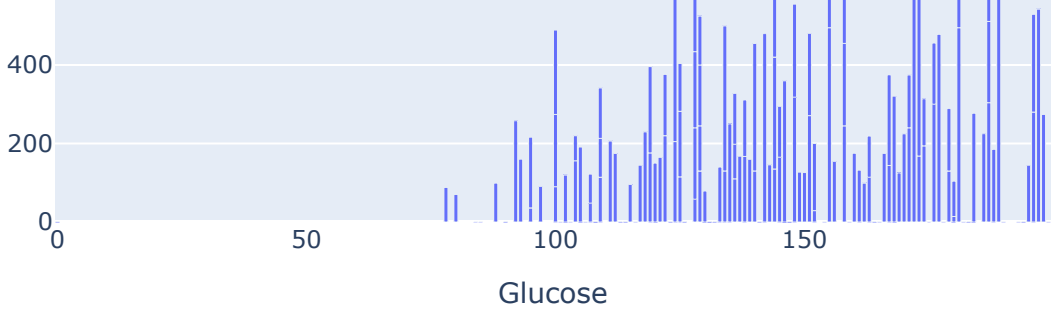
Out[7]:
```
Text(0.5, 1.0, 'BMI vs AGE for Diabetes Outcome')
```



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

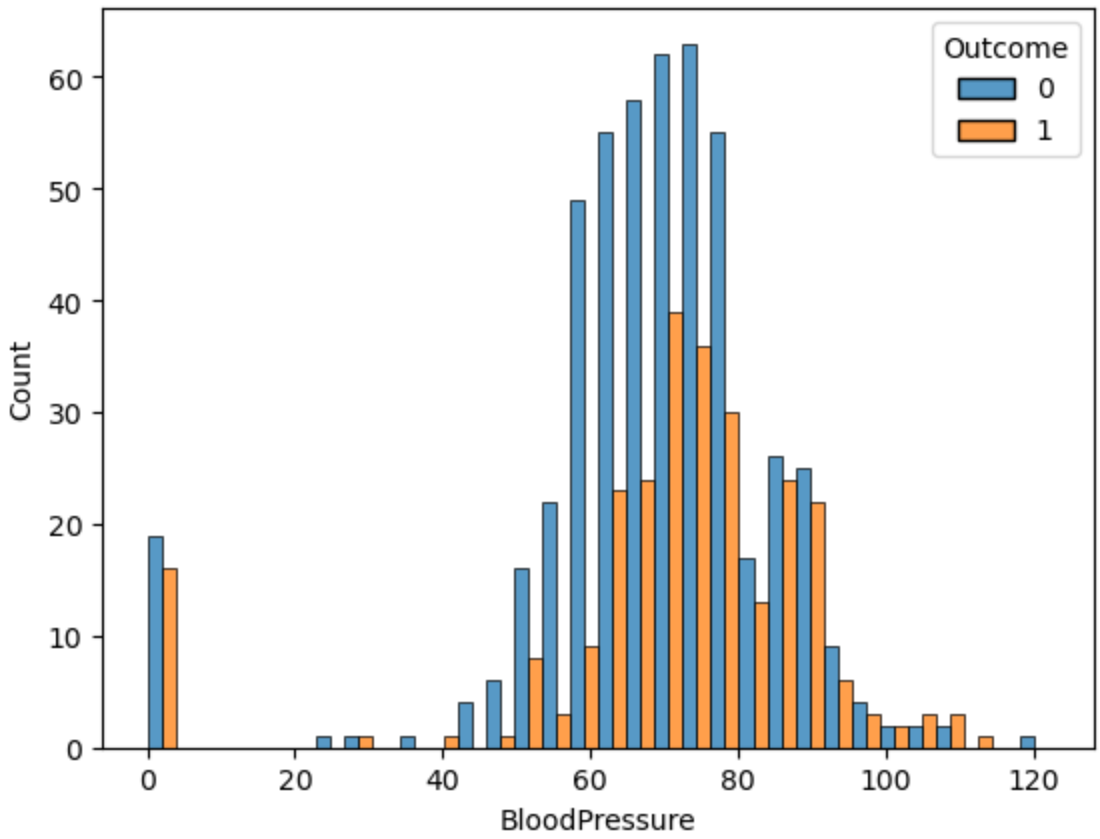
In [8]:
```
#BAR chart on Glucose and Insulin levels
diabetes_data = diabetes_df[diabetes_df.Outcome==1]
fig = px.bar(diabetes_data, x='Glucose', y='Insulin',title="Glucose vs Insulin levels")
fig.show()
```





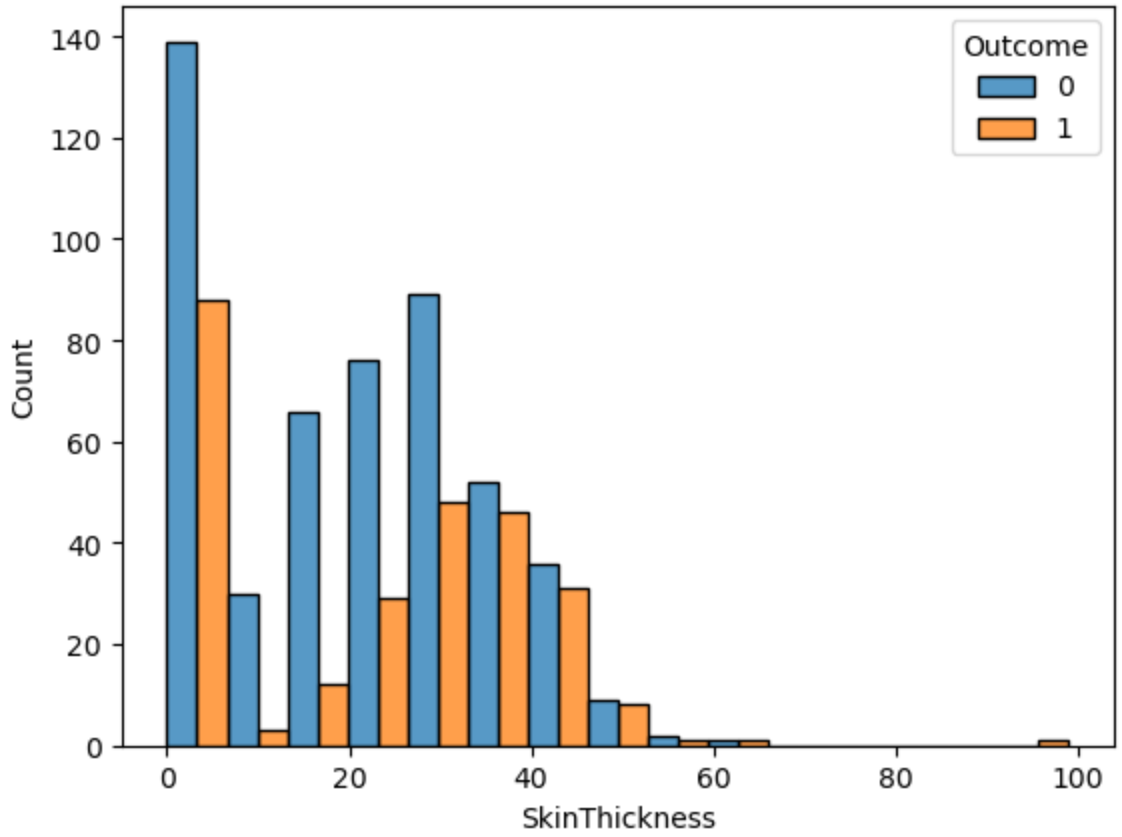
In [9]:
```
#Histogram on BloodPressure and Outcome
sns.histplot(data=diabetes_df, x="BloodPressure", hue="Outcome", multiple="dodge")
```

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



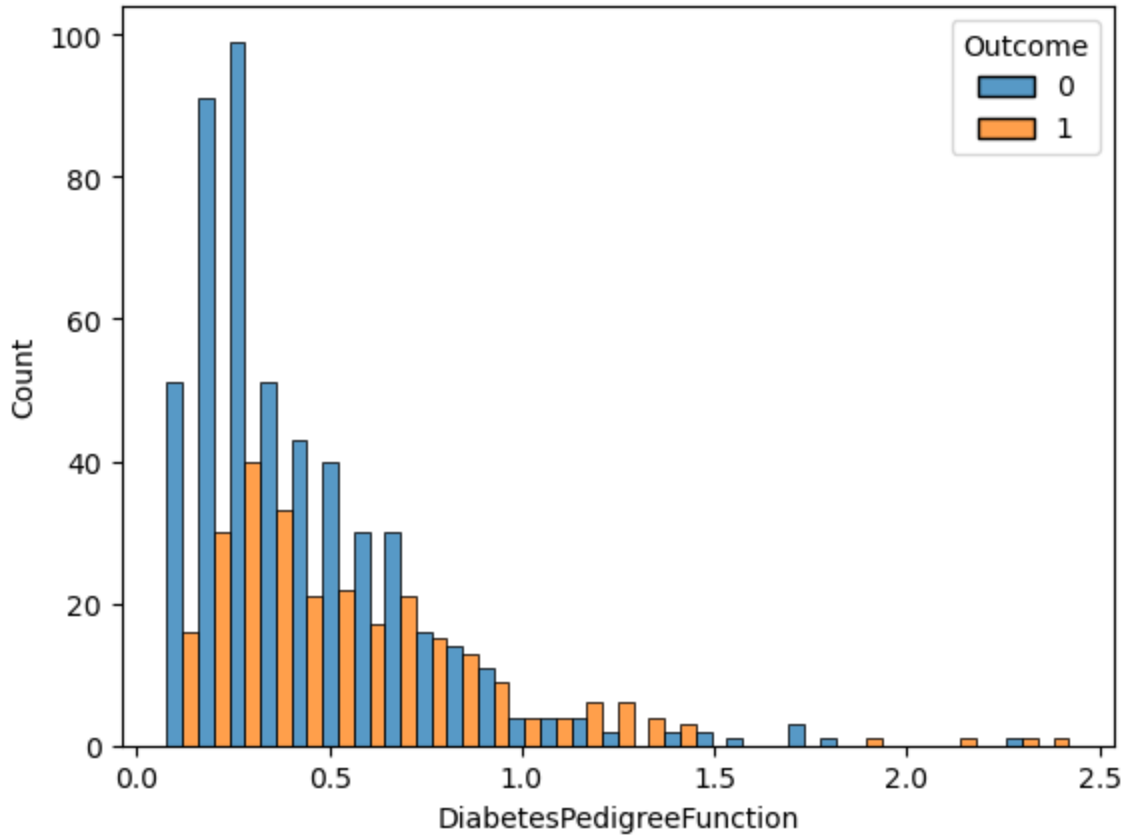
In [10]:
```
#Histogram on SkinThickness and Outcome
sns.histplot(data=diabetes_df, x="SkinThickness", hue="Outcome", multiple="dodge")
```

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



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'>
```



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