PYTHON PROJECT

About the dataset

This dataset contains self-reported clothing-fit feedback from customers as well as other side information like reviews, ratings, product categories, catalog sizes, customers’ measurements (etc.) from website:

1. Kaggle

CODE :

import pandas as pd

df = pd.read\_csv(r"C:\Users\nitis\Downloads\diabetes-data.csv")

print(df.head())

print(df.describe())

print(df['Age'].value\_counts())

print(df.dtypes)

df['BMI'] = pd.to\_numeric(df['BMI'])

print(df.isnull().sum())

df.dropna(inplace=True)

import matplotlib.pyplot as plt

# Histogram

plt.hist(df['BloodPressure'], bins=5)

plt.xlabel('BloodPressure')

plt.ylabel('Frequency')

plt.show()

# Scatter plot

plt.scatter(df['Age'], df['Glucose'])

plt.xlabel('Age')

plt.ylabel('Glucose')

plt.show()

# correlation matrix

import seaborn as sns

correlation\_matrix = df.corr()

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm')

plt.show()

# Box plot

plt.boxplot(df['Pregnancies'])

plt.xlabel('Pregnancies')

plt.show()

# Group by a categorical column (outcome)(Age) and calculate mean

grouped = df.groupby('Outcome')['Age'].mean()

print(grouped)

# Random sampling

sampled\_df = df.sample(n=100, random\_state=42)

print(sampled\_df)

# Binning numeric data(outcome column) into categories

bins = [-1,0,1]

labels = ['negative','positive']

df['report\_outcome'] = pd.cut(df['Outcome'], bins=bins, labels=labels)

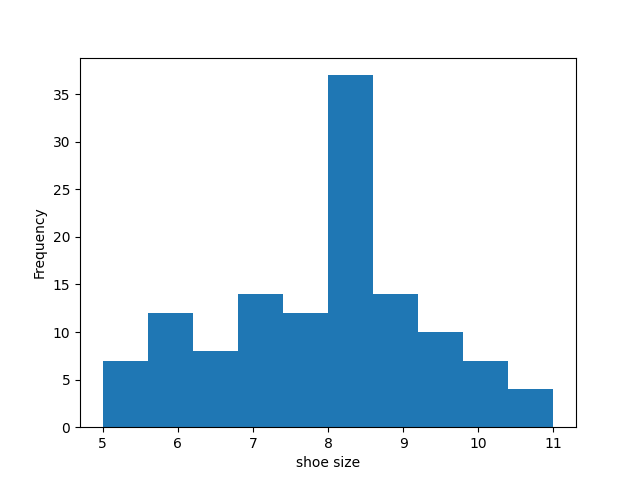
print(df.head())

# Export

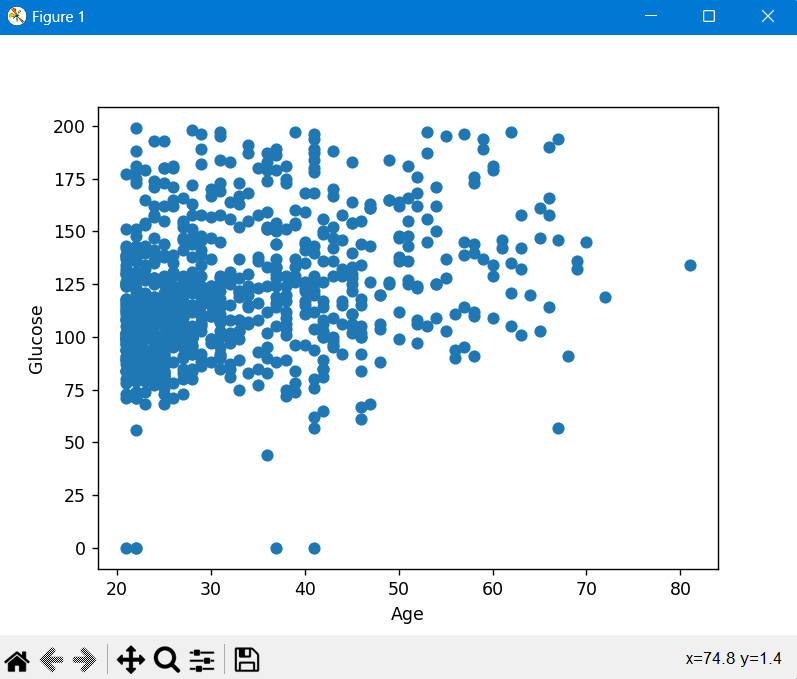
df.to\_csv(r"C:\Users\nitis\Downloads\updated-diabetes-data.csv", index=False)

OUTPUTS :

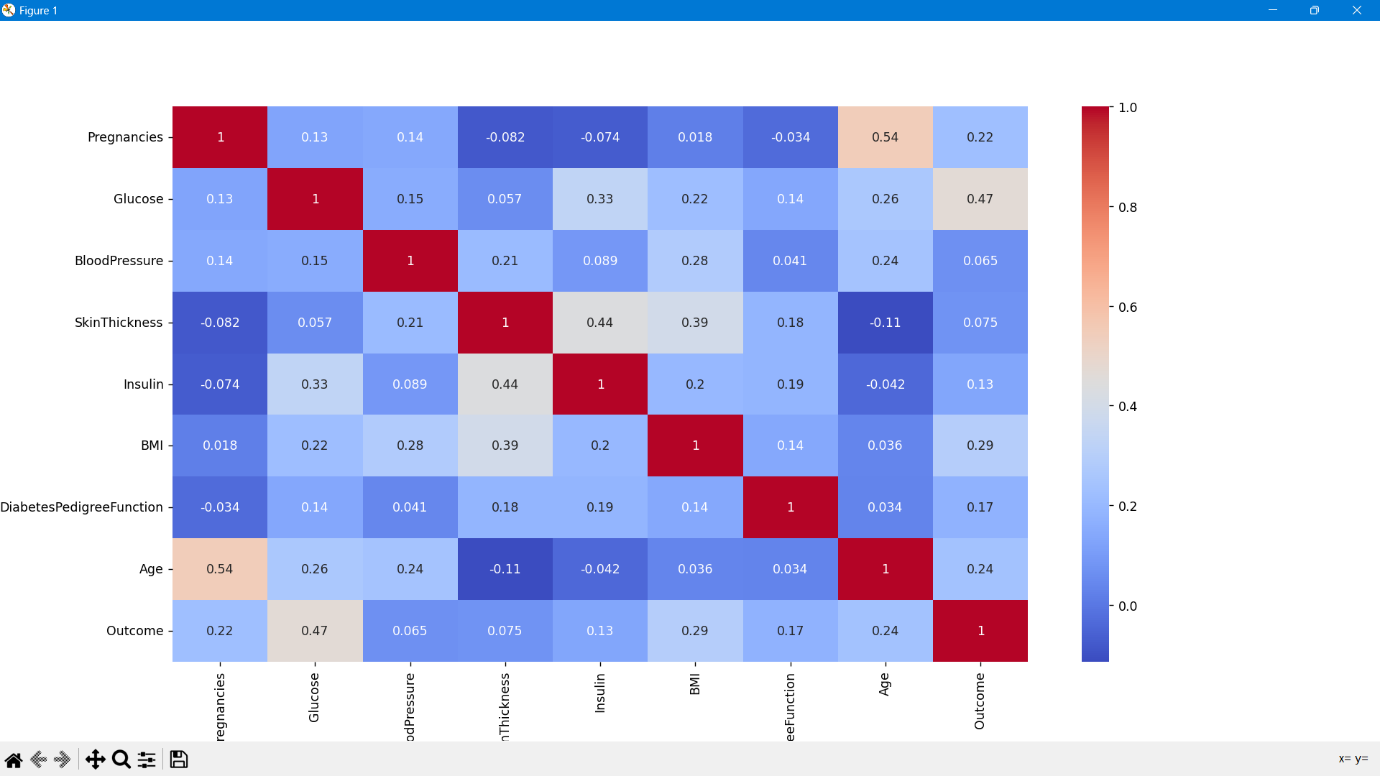
# HISTOGRAM



# SCATTER PLOT



# CORRELATION MATRIX



# BOX PLOT

A screenshot of a computer

Description automatically generated

# OUTPUT SNIPPETS :

