Dataset link: Healthcare_Lab3

Dataset Name: Healthcare Diabetes Dataset
https://www.kaggle.com/datasets/nanditapore/healthcare-diabetes

This dataset contains a diverse range of health-related attributes, meticulously collected to aid in the development of predictive models for identifying individuals at risk of diabetes.

```
> glimpse(data)
Rows: 2,768
Columns: 10
$ Id                       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21…
$ Pregnancies              <dbl> 6, 1, 8, 1, 0, 5, 3, 10, 2, 8, 4, 10, 10, 1, 5, 7, 0, 7, 1, 1, 3, 8, 7, 9…
$ Glucose                  <dbl> 148, 85, 183, 89, 137, 116, 78, 115, 197, 125, 110, 168, 139, 189, 166, 1…
$ BloodPressure            <dbl> 72, 66, 64, 66, 40, 74, 50, 0, 70, 96, 92, 74, 80, 60, 72, 0, 84, 74, 30,…
$ SkinThickness            <dbl> 35, 29, 0, 23, 35, 0, 32, 0, 45, 0, 0, 0, 0, 23, 19, 0, 47, 0, 38, 30, 41…
$ Insulin                  <dbl> 0, 0, 0, 94, 168, 0, 88, 0, 543, 0, 0, 0, 0, 846, 175, 0, 230, 0, 83, 96,…
$ BMI                      <dbl> 33.6, 26.6, 23.3, 28.1, 43.1, 25.6, 31.0, 35.3, 30.5, 0.0, 37.6, 38.0, 27…
$ DiabetesPedigreeFunction <dbl> 0.627, 0.351, 0.672, 0.167, 2.288, 0.201, 0.248, 0.134, 0.158, 0.232, 0.1…
$ Age                      <dbl> 50, 31, 32, 21, 33, 30, 26, 29, 53, 54, 30, 34, 57, 59, 51, 32, 31, 31, 3…
$ Outcome                  <dbl> 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1…
> 
```

1. Id: Unique identifier for each data entry.
2. Pregnancies: Number of times pregnant.
3. Glucose: Plasma glucose concentration over 2 hours in an oral glucose tolerance test.
4. BloodPressure: Diastolic blood pressure (mm Hg).
5. SkinThickness: Triceps skinfold thickness (mm).
6. Insulin: 2-Hour serum insulin (mu U/ml).
7. BMI: Body mass index (weight in kg / height in m^2).
8. DiabetesPedigreeFunction: Diabetes pedigree function, a genetic score of diabetes.
9. Age: Age in years.
10. Outcome: Binary classification indicating the presence (1) or absence (0) of diabetes.

Task 1:

Dependent variable: Skin Thickness
Independent Variable: Age
Pearson correlation coefficient: -0.11

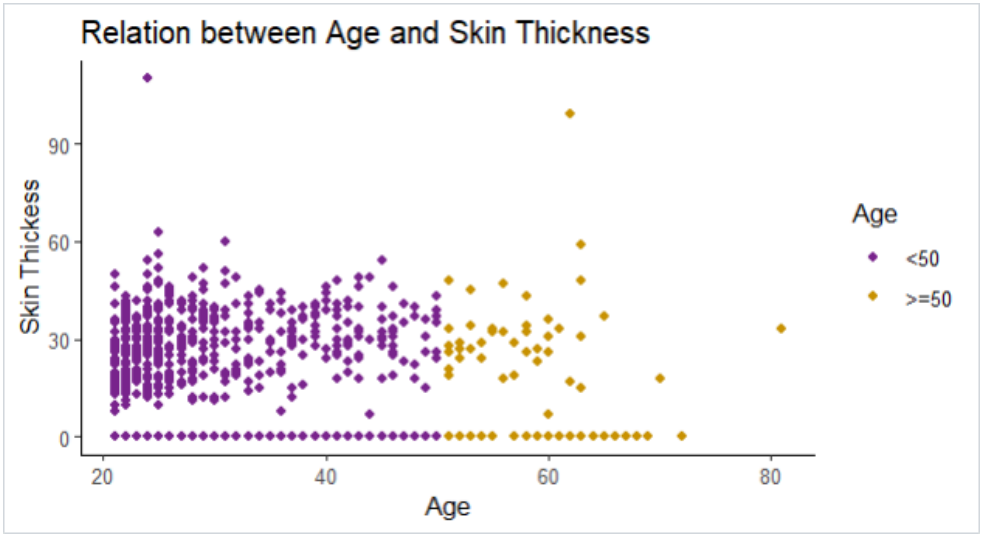
```
> cor(df_clean$Age,df_clean$SkinThickness)
[1] -0.1118954
```

```
> cor.test(df_clean$Age,df_clean$SkinThickness)

        Pearson's product-moment correlation

data:  df_clean$Age and df_clean$SkinThickness
t = -5.9221, df = 2766, p-value = 3.572e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.14853247 -0.07495157
sample estimates:
       cor
-0.1118954
```

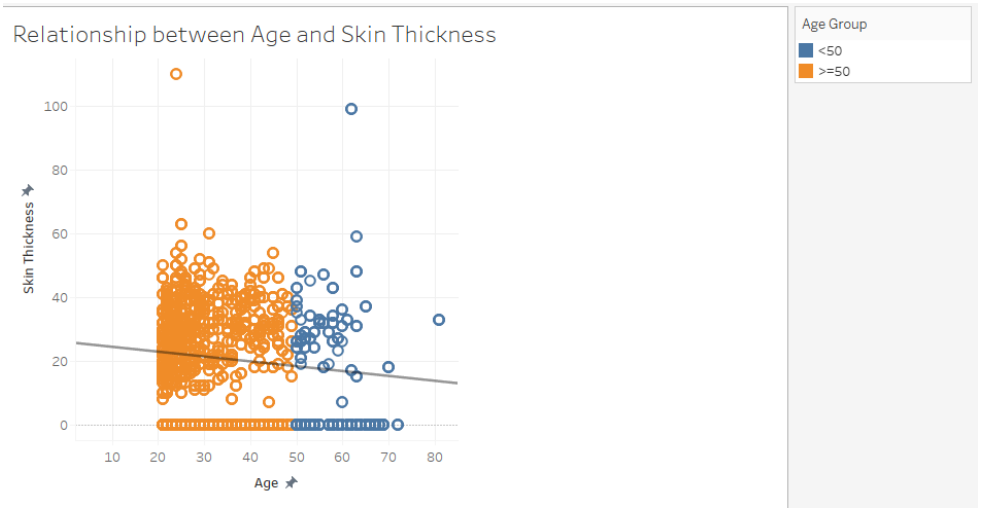
A hypothesis test was performed to verify the existence of a correlation between Age and Skin Thickness. The results of the test were the following: There is a negative, Weak
correlation between Age and Skin Thickness (r=-0.11), and it is statistically significant (p<=0.05).

Task 2:

Scatter Plot using R:



Scatter Plot using Tableau:

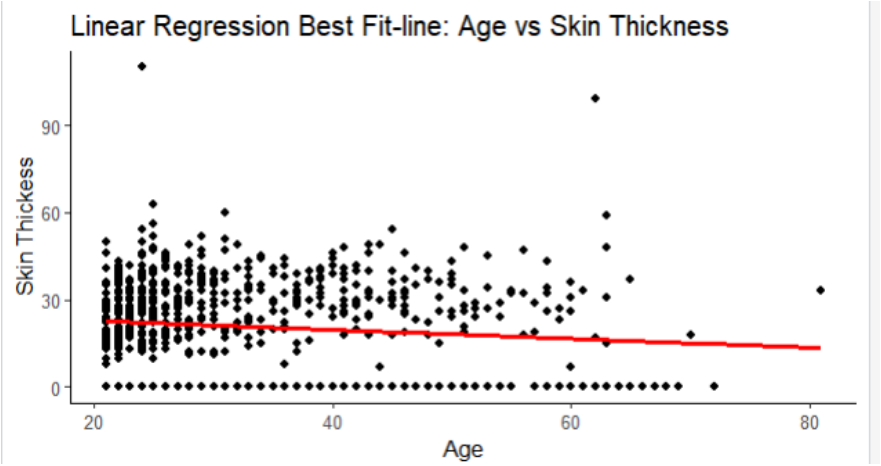


Task 3:

1. Perform Linear Regression Analysis using R or Python:

```
> summary(lm( df_clean$SkinThickness ~ df_clean$Age ))

Call:
lm(formula = df_clean$SkinThickness ~ df_clean$Age)

Residuals:
    Min      1Q  Median      3Q     Max
-22.676 -17.640   1.613  12.071  87.782

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  25.87981    0.90596  28.566  < 2e-16 ***
df_clean$Age -0.15258    0.02576  -5.922 3.57e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.96 on 2766 degrees of freedom
Multiple R-squared:  0.01252,   Adjusted R-squared:  0.01216
F-statistic: 35.07 on 1 and 2766 DF,  p-value: 3.572e-09
```

Best Fit-lineequation: Y = 25.87 - 0.15X



2. Perform Multiple Linear Regression Analysis using R or Python

```
> summary(lm(df_clean$SkinThickness ~ df_clean$Age + df_clean$Glucose
+            + df_clean$Insulin + df_clean$BloodPressure))

Call:
lm(formula = df_clean$SkinThickness ~ df_clean$Age + df_clean$Glucose +
    df_clean$Insulin + df_clean$BloodPressure)

Residuals:
    Min      1Q  Median      3Q     Max
-38.757 -12.902  -0.846  10.122  92.467

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           14.107104   1.370270  10.295  < 2e-16 ***
df_clean$Age          -0.142625   0.024268  -5.877 4.67e-09 ***
df_clean$Glucose      -0.042703   0.009179  -4.652 3.44e-06 ***
df_clean$Insulin       0.064029   0.002543  25.178  < 2e-16 ***
df_clean$BloodPressure 0.166107   0.014310  11.608  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.96 on 2763 degrees of freedom
Multiple R-squared:  0.2459,    Adjusted R-squared:  0.2448
F-statistic: 225.3 on 4 and 2763 DF,  p-value: < 2.2e-16

>
```

The independent variables (Age, Glucose, Insulin, and BloodPressure) are statistically significant predictors of the dependent variable (SkinThickness).
The R-squared value suggests that these variables collectively explain a substantial portion of the variance in SkinThickness (about 24.59%).
The model is statistically significant as indicated by the low p-value

Task 4:

Relationship between Age and Skin Thickness