1. Create a new conda environment named: data-science-final:
    a. Input: `conda create --name data-science-final`
    b. Activate conda: `conda activate data-science-final`

2. Load this environment in VS Code and install the ”kernel”. Install any extension it requires.
    a. Load the environment in VS using the command: `code .`
    b. Download pandas package using `%pip install pandas`

WRITE-UP:

"What are the correlations between predictor variables and the occurrence of diabetes within the Pima Indian population?"

I chose to work on this question because I wanted to know which predictor variable is the most credible to refer to when gauging for a probability of diabetes. To explore this, I first created histograms for each predictor variable. But due to the data being widely spread out, it didn't show any obvious correlations.After that I plotted scatter plots for each variable. Some scatter plots showed a relation but others were quite vague and hence incredible to account for a conclusion.
Hence I plotted a heatmap to show the correlation for each variable with the outcome. Since I wanted the relationship of each variable with outcome, heatmaps were the most suitable option as they depict the correlation on a scale.

The Conclusions:

The heatmap shows the correlation coefficients between various predictor variables and the outcome variable (likely indicating diabetes status). Here are some conclusions based on the correlation coefficients:

Glucose:

Has the highest positive correlation with the outcome (0.47). This suggests that higher glucose levels are strongly associated with the presence of diabetes.

BMI:

Has a moderately positive correlation with the outcome (0.30). This indicates that higher BMI is associated with a higher likelihood of diabetes.

Age:

Shows a moderate positive correlation with the outcome (0.24). Older age is associated with a higher likelihood of diabetes.

Pregnancies:

Has a positive correlation with the outcome (0.22). This suggests that the number of pregnancies is somewhat associated with a higher risk of diabetes.

The other variables have a weak relationship with the outcome and hence can't be relied on.

General Observations: Among all the predictor variables, glucose level has the strongest association with the diabetes outcome. Hence, while predicting or gauging the probability of diabetes, the glucose levels should be prioritized first.

The heatmap also provided some additional insights into the correlation of the predictor variables amongst themselves. For instance:

1. Pregnancies and Age show the positive correlation implying that older people have a higher number of pregnancies.
2. Insulin levels and Skin thickness: There is a moderate positive correlation between insulin levels and skin thickness, indicating that higher insulin levels might be associated with higher skin thickness measurements.
3. Skin thickness and Age: There is a negative correlation between these two suggesting that aging reduces the skin thickness.
4. BMI and Skin Thickness: Skin thickness and BMI have a moderate positive correlation, implying that individuals with higher BMI tend to have higher skin thickness measurements.

These are some of the conclusions made from the heatmap. The Heatmap was the best method to test this as it provided an insight into the relationship of each variable with the outcome and other predictor variables.