FINAL PROJECT REFLECTIONS AND RESULTS

I started my project by creating a new conda environment named data-science-final. I RAND the mkdir command on my terminal (PowerShell).
Next, I formulated questions and answered them through various codes. I used google colab as my VS code was not working. If it were, then I would install the kernel and any extension required.

First, I downloaded the necessary libraries and loaded the file. Then, I found out the top 5 and bottom 5 entries in my data. Then I found out the number of NA values in each column and dropped the rows containing NA values. After doing that, I found the correlation between different columns and plotted it using a heatmap.
Furthermore, I divided my data into 2 parts- test and split. I didn't use the sk learn library to do so as I found the other method more easy.
I mainly focused on plotting scatterplots to find the relation between the outcome and the other features.

The inferences I made from the scatterplots I plotted

1. Glucose is a very strong factor and highly affects the outcome of the diabetes test. People with diabetes (outcome was 1) had higher glucose levels. People with low glucose levels were not diabetic.
2. According to the scatterplot, there was no clear trend between the pregnancies and the outcome.
3. According to the scatterplot, there was no clear trend between the blood pressure and the outcome.
4. According to the scatterplot, there was no significant relation between the skin thickness and the outcome.
5. According to the scatterplot, insulin also cannot be the only factor of the outcome, other factors are also needed as there isn’t any specific trend between the insulin levels and the diabetes outcome.
6. According to the scatterplot, there was no clear trend between the age and the outcome.
7. According to the scatterplot, there was no clear trend between the BMP and the outcome.
8. According to the scatterplot, there was no clear trend between the diabetes predigree function and the outcome.

Thank you