Bahria University,

Karachi Campus



COURSE: CEN-221

DATA MINING LAB

TERM: FALL 2024, CLASS: BSE- 6(B)

Submitted By:

AIMAN ZIA SATTI (02-131212-028)

(Name) (Reg. No.)

Submitted To:

Dr.Hina/Engr Hamza

Signed Remarks: Score:\_\_\_\_\_

INDEX

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SNO | DATE | LAB NO | LAB OBJECTIVE | SIGN |
| 01 |  |  |  |  |
| 02 |  |  |  |  |
| 03 |  |  |  |  |
| 04 |  |  |  |  |
| 05 |  |  |  |  |
| 06 |  |  |  |  |
| 07 |  |  |  |  |
| 08 |  |  |  |  |
| 09 |  |  |  |  |
| 10 |  |  |  |  |
| 11 |  |  |  |  |
| 12 |  |  |  |  |
| 13 |  |  |  |  |
| 14 |  |  |  |  |

Bahria University,

Karachi Campus



LAB EXPERIMENT NO.

**\_02\_**

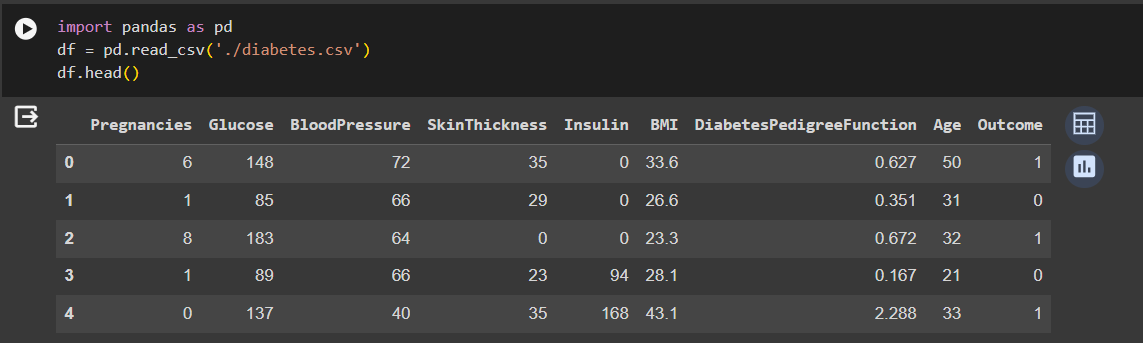
LIST OF TASKS

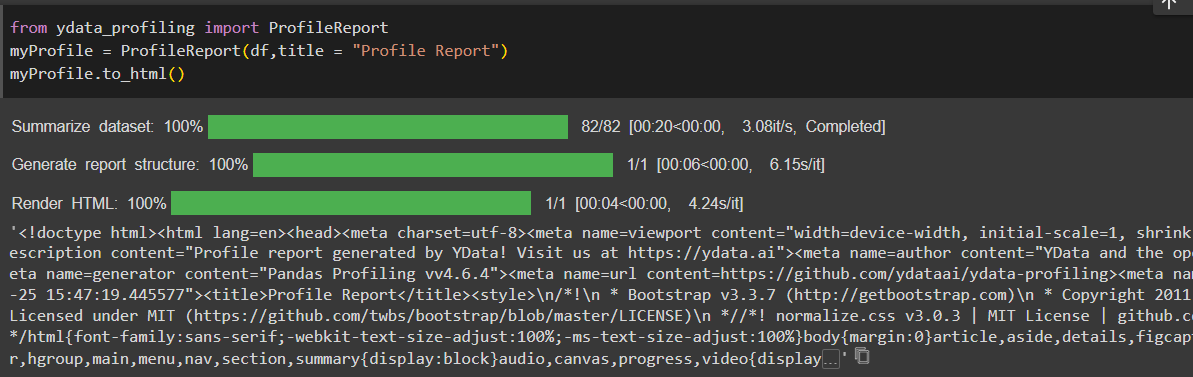
|  |  |
| --- | --- |
| TASK NO | OBJECTIVE |
| 1 | * 1. Perform a basic data profiling to understand the structure of the dataset, including the number of rows, columns, and data types.   2. Identify the target variable and the predictor variables.   3. Compute summary statistics (mean, median, standard deviation, etc.) for each numerical variable.   4. Identify missing values and their distribution across variables |
| 2 | * 1. Calculate pairwise correlation coefficients between all numerical variables.   2. Create a correlation matrix and visualize it using a heatmap.   3. Identify highly correlated variables and discuss their potential impact on model performance. |
| 3 | * 1. Identify potential outliers in the dataset using appropriate techniques, such as box plots, scatter plots.   2. Visualize the distribution of each numerical variable to identify any extreme values.   3. Discuss the potential impact of outliers on the analysis and modeling process. |
| 4 | * 1. Analyze the distribution of the target variable (diabetes or non-diabetes).   2. Visualize the target variable distribution using a histogram or a bar chart.   3. Identify any potential imbalance in the target variable and discuss its impact on model performance. |

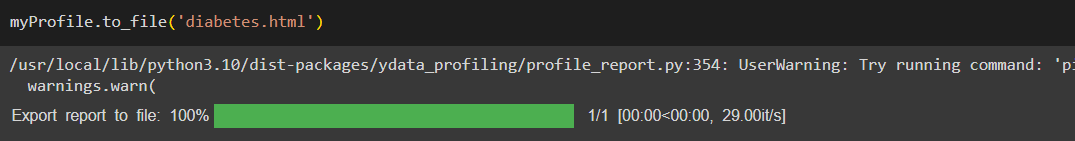
Submitted On:

**Date: \_\_\_\_\_22/02/2024\_\_\_**

**TASK 01 : Data Profiling**:







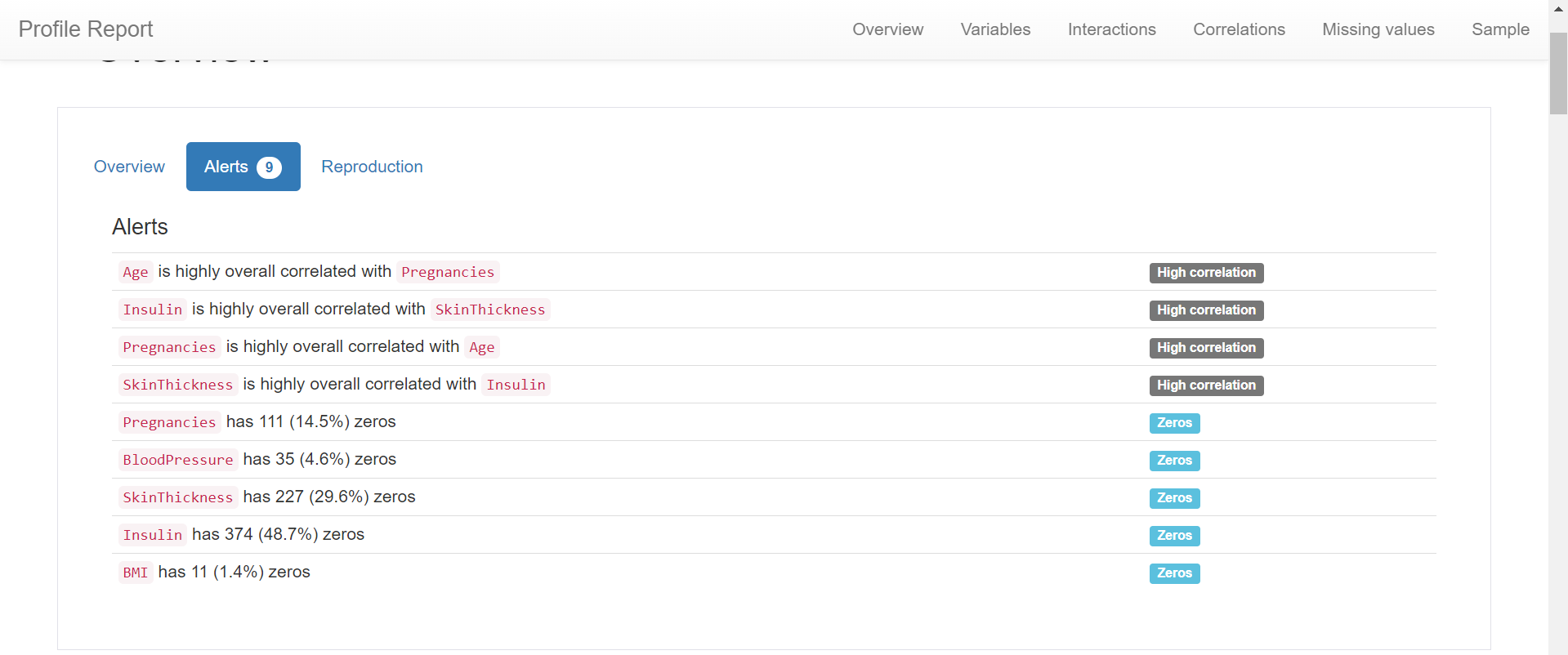
* 1. Perform a basic data profiling to understand the structure of the dataset, including the number of rows, columns, and data types.

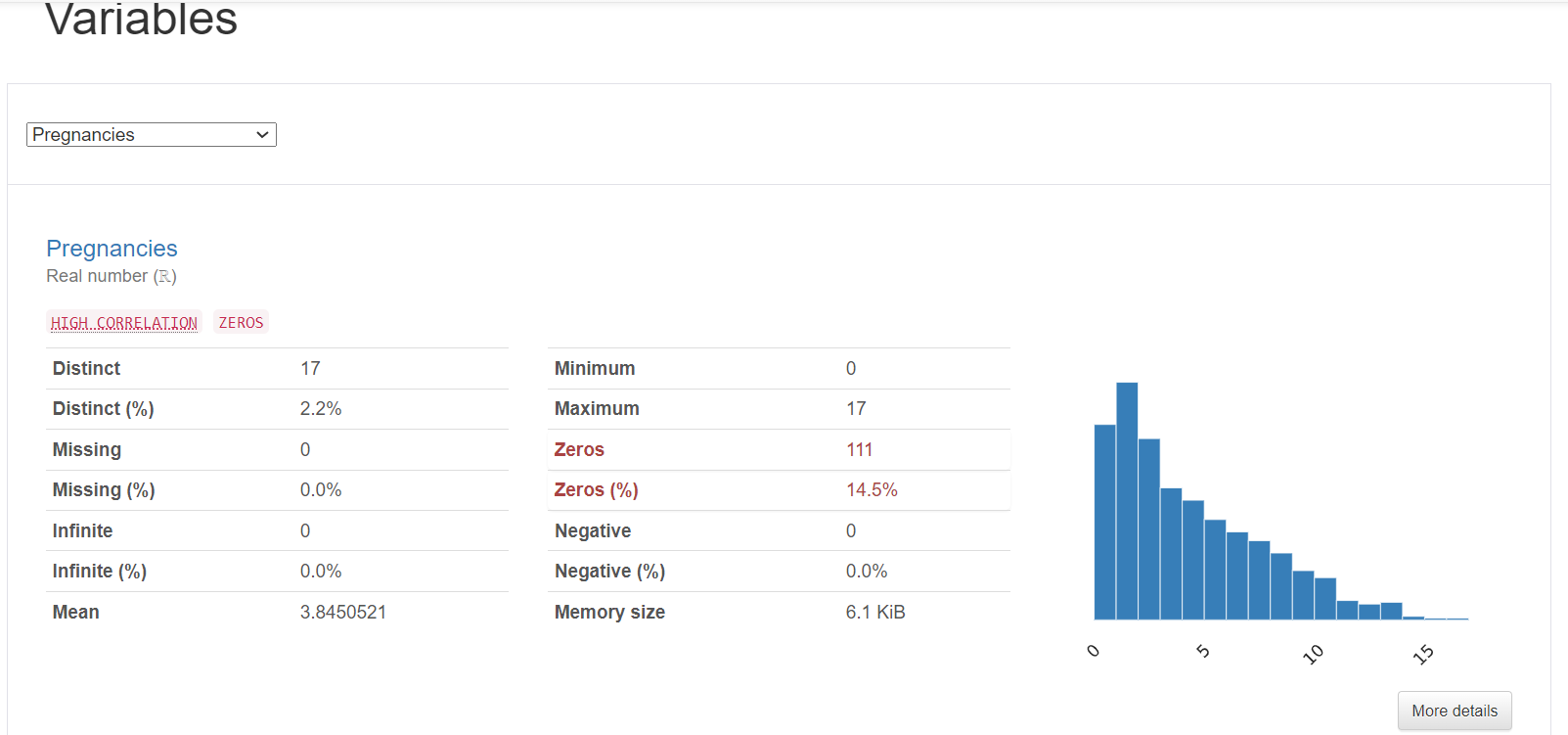
A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

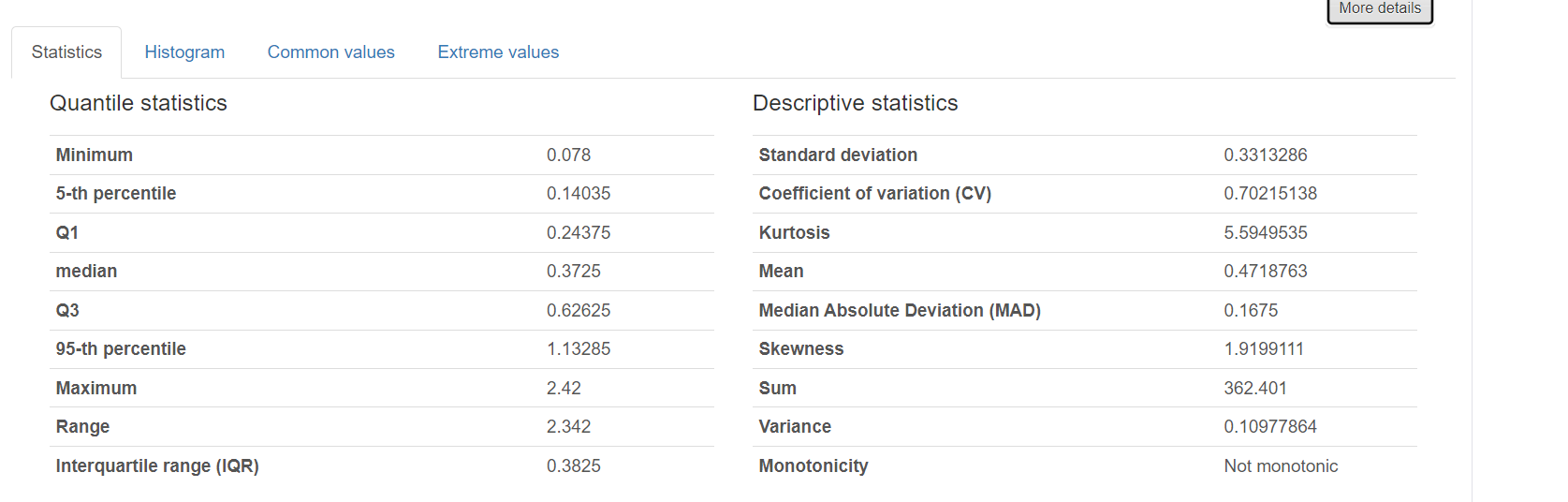




A screenshot of a graph

Description automatically generated

* 1. Compute summary statistics (mean, median, standard deviation, etc.) for each numerical variable.



A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Identify missing values and their distribution across variables.

A graph with blue lines

Description automatically generated

**TASK 02 : Feature Correlation**:

A screen shot of a graph

Description automatically generated

* 1. Create a correlation matrix and visualize it using a heatmap.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

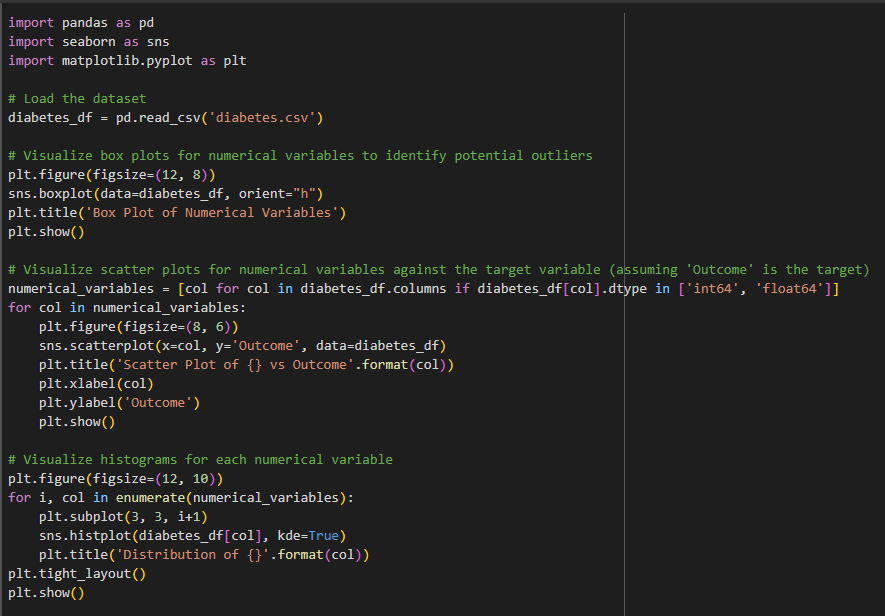
Description automatically generated

* 1. Identify highly correlated variables and discuss their potential impact on model performance.



**TASK 03 : Outlier Detection**:

* 1. Identify potential outliers in the dataset using appropriate techniques, such as box plots, scatter plots.



A graph showing a box plot of numerical variables

Description automatically generated

A graph with blue dots

Description automatically generatedA graph of blood pressure

Description automatically generated

* 1. Visualize the distribution of each numerical variable to identify any extreme values.

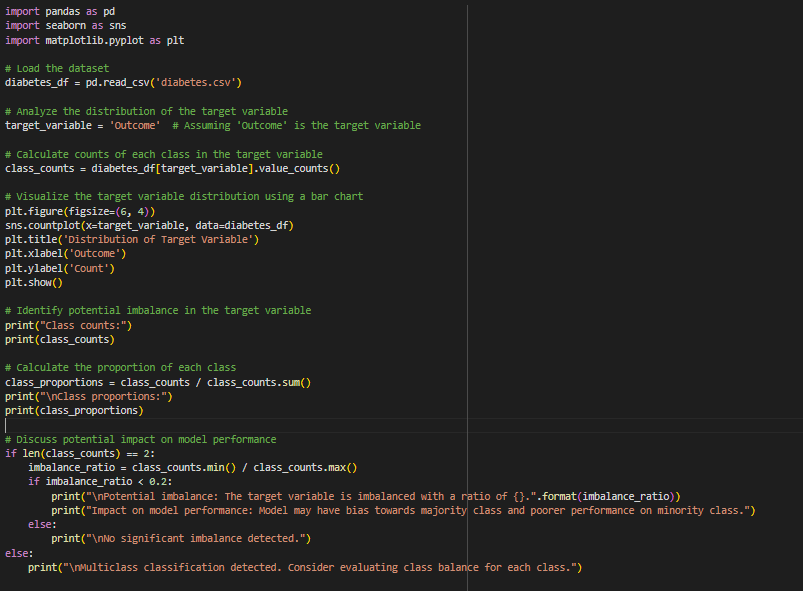
A group of graphs showing different sizes of blood

Description automatically generated with medium confidence

* 1. Discuss the potential impact of outliers on the analysis and modeling process.

Outliers can significantly impact the analysis and modeling process by skewing statistical measures, affecting model performance, and distorting the interpretation of results. They can lead to biased estimates of parameters, influence decision boundaries in classification models, and reduce the generalization ability of the model. Handling outliers appropriately, such as by removing them, transforming variables, or using robust statistical methods, is essential to ensure the accuracy and reliability of the analysis and model predictions.

**TASK 04 : Target Variable Analysis**:



A graph of a distribution of target variable

Description automatically generated A screenshot of a computer

Description automatically generated