**EC 9560 – DATA MINING**

**LAB 03**

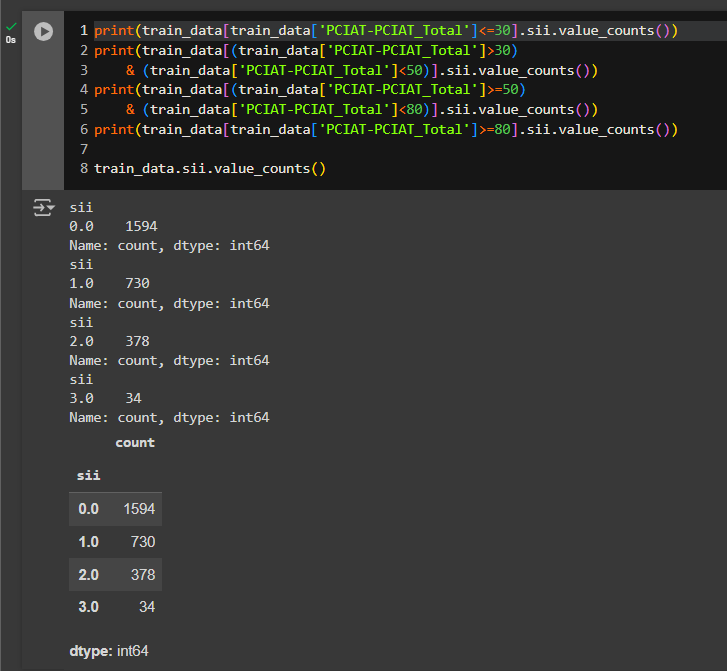
**DARMILA.T**

**2020/E/027**

**SEMESTER 07**

**30th OCTOBER 2024**

**Data Preprocessing**



This sii values are correctly distributed within the PCIAT-PCIAT\_Total score intervals.

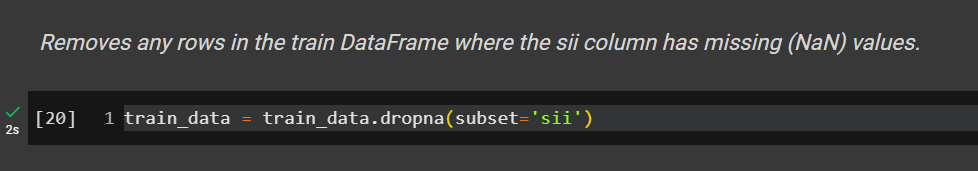
**A screen shot of a computer program

Description automatically generated**

To ensure the sii target variable distribution aligns with the expected categories of PCIAT-PCIAT\_Total score ranges, verify that each range of scores corresponds correctly to its designated sii value.

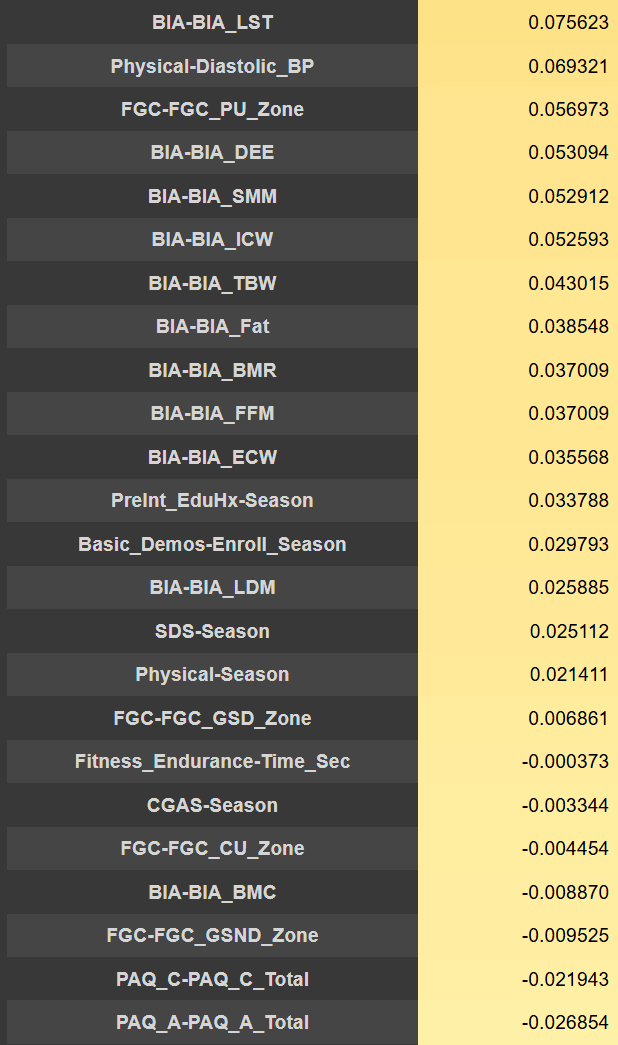
**Define the Expected Ranges**: Based on the original criteria:

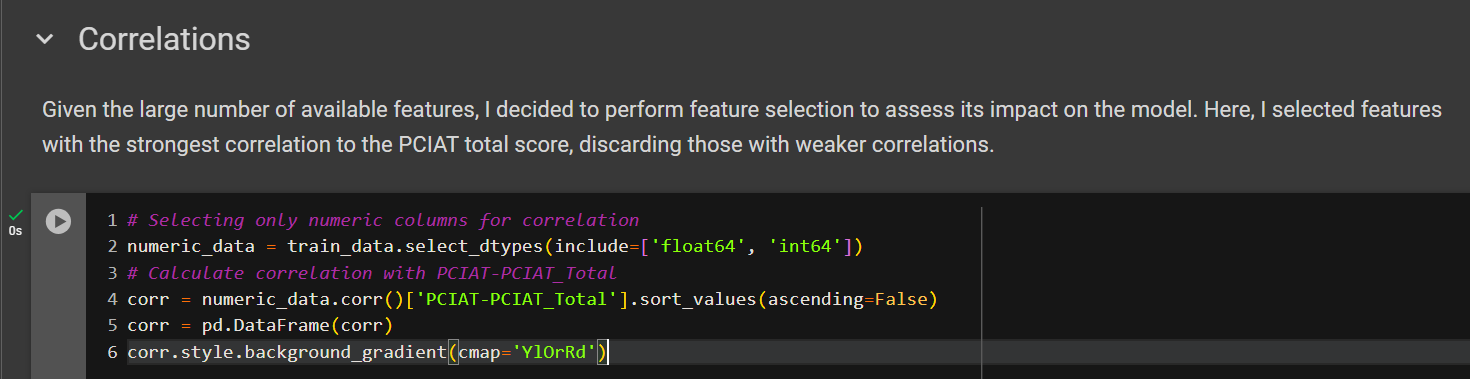
* PCIAT-PCIAT\_Total between 0–30 should have sii = 0.
* PCIAT-PCIAT\_Total between 31–49 should have sii = 1.
* PCIAT-PCIAT\_Total between 50–79 should have sii = 2.
* PCIAT-PCIAT\_Total between 80–100 should have sii = 3.

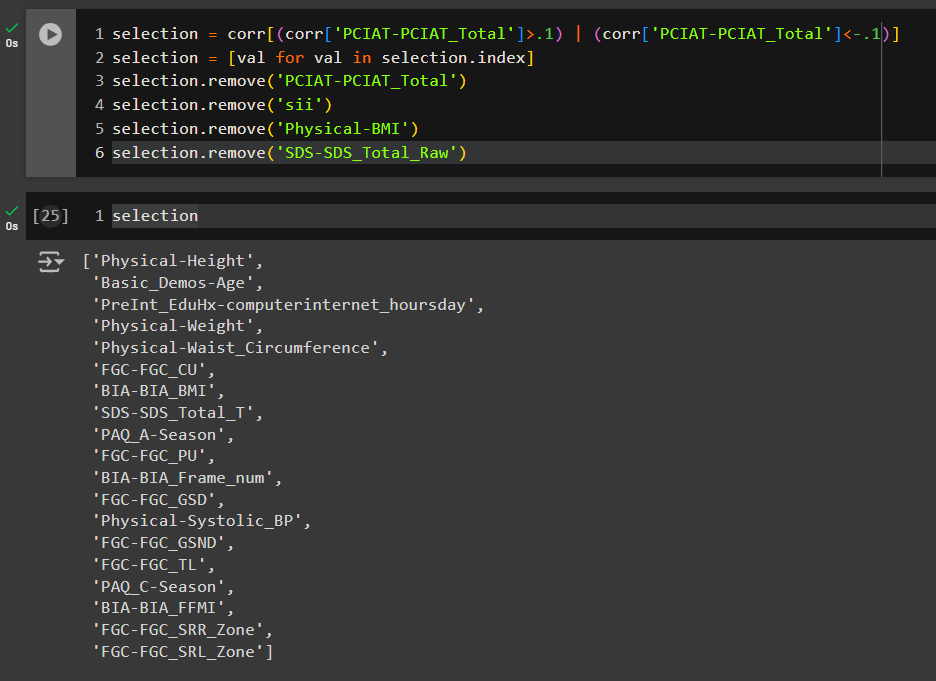


Here, I clean the data by ensuring all rows in train have valid values in the sii column, which is likely important if sii is used as a target variable or feature in further analysis.

A screenshot of a graph

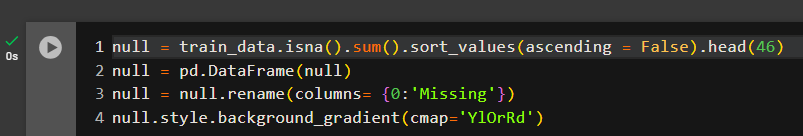
Description automatically generatedA screenshot of a computer

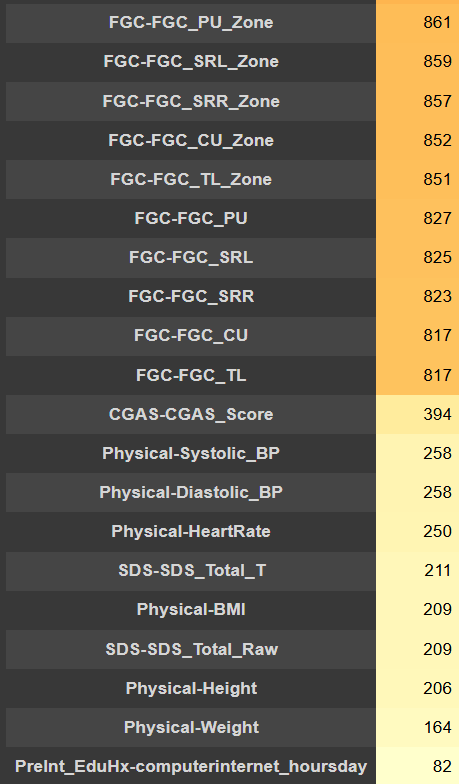
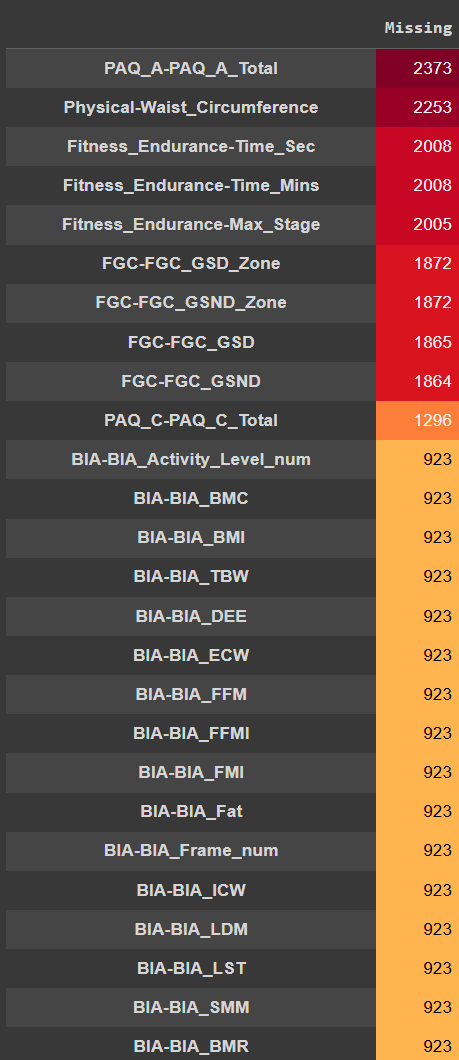
Description automatically generated

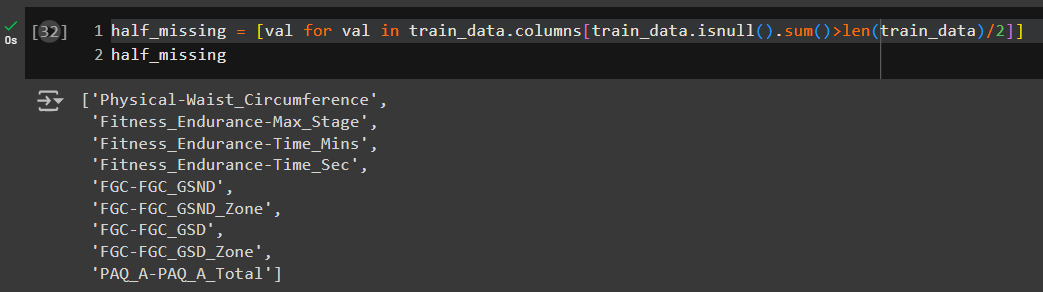


Here, I select a subset of features that have a moderate correlation (either positive or negative) with PCIAT-PCIAT\_Total, while excluding certain specified columns. The goal is to narrow down the list of features to those most relevant for analysis or modeling, based on their correlation with PCIAT-PCIAT\_Total, and to exclude target or potentially redundant features.

**Find the missing values in the train\_data**

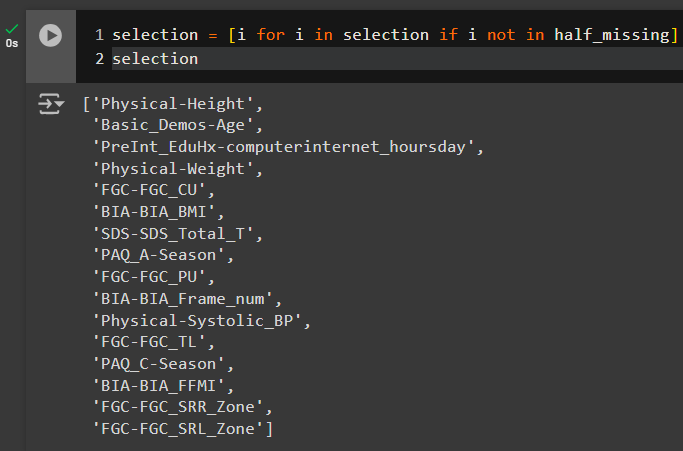






Here, Identifies the column that more than one value is missing in the data frame.

Then, I check is any selected features have more than half value is missing, If I Found remove that in from the selection list features.



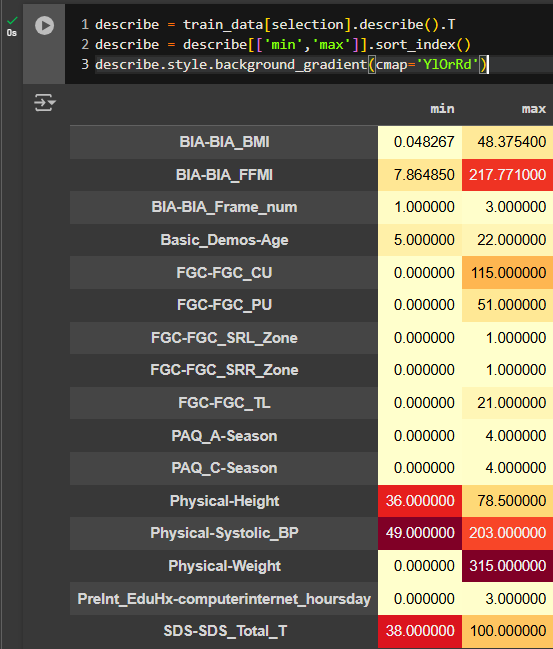
I have now identified 16 selected features based on two criteria:

a) their correlation with the target variable and

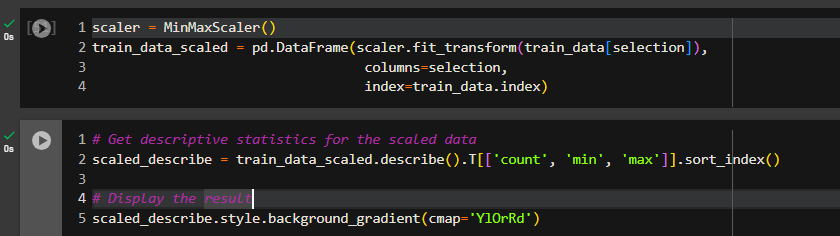
b) the presence of relatively few missing values.

Here, I want to create a model that makes accurate predictions by using only the most useful information while avoiding distractions from irrelevant data.

In the following section, I display the minimum and maximum values of the selected features in the training data.



Now, I scaled scale the selected features in your train\_data DataFrame using MinMaxScaler.



A screenshot of a computer

Description automatically generated

Split the dataset

