**SD LAB3: DATA PRE-PROCESSING**

**NAME: KEMGOU LEKANE ALLICSON D**

**DEPARTMENT: SOFTWARE ENGINEERING**

**Campus B**

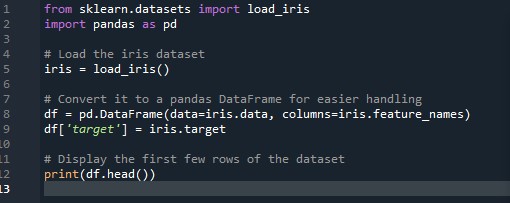
**Introduction**

Data preprocessing is the process of transforming raw data into a clean, structured, and usable format before feeding it into a machine learning model. This step is essential in data science and machine learning because real-world data is often messy, incomplete, and inconsistent, which can affect the accuracy and performance of models. In summary, data preprocessing is a critical step to ensure the data is prepared for modeling, ultimately leading to more reliable and efficient machine learning models. This different models can be seen below:

**Step 1. Data Collection**

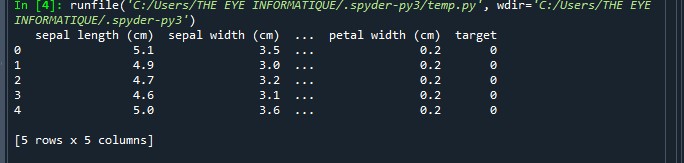
Gathering data from various sources, such as databases, sensors, surveys, or web scraping.

Ensuring the data is collected in a structured and usable format, often involving different formats or types (e.g., text, numeric, images).



***1 .This figure shows us how data will be collected***

***OUTPUT:***

******

***2.This OUTPUT shows us how data is been collected and been displaced in rows and columns.***

**Step 2. Data Cleaning**

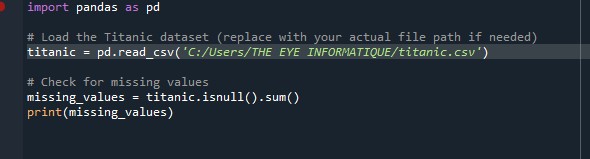
Removing or correcting any errors, inconsistencies, or irrelevant data to improve quality.

Handling missing values, removing duplicates, and correcting inaccuracies in the data.

Converting data to a standard format to ensure consistency.

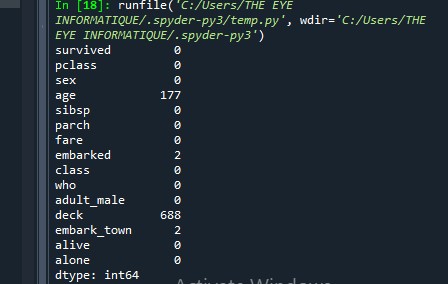
1. **Inspection of Missing Values**

This will detect and show us all the missing values in the dataset

****

***Figure 3. This step will show us how to find missing values***

***OUTPUT:***

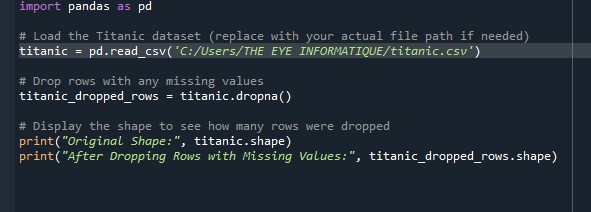
******

***Figure 4. The missing values can be seen from the table above***

**b. Handle Missing Values by:**

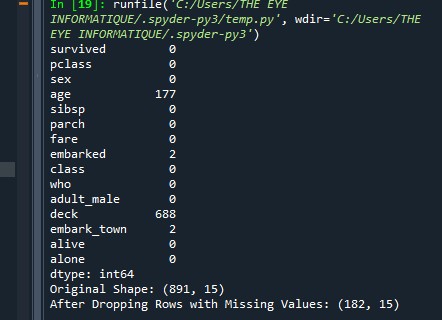
**1.Dropping Rows**

**It will show the missing values in the rows and columns of he datasets**

****

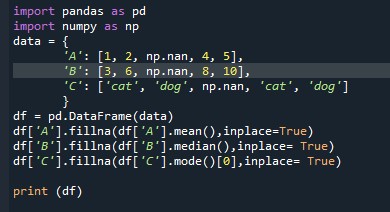
***Figure 5. This figure show us how to find missing values by dropping rows***

***OUTPUT:***

******

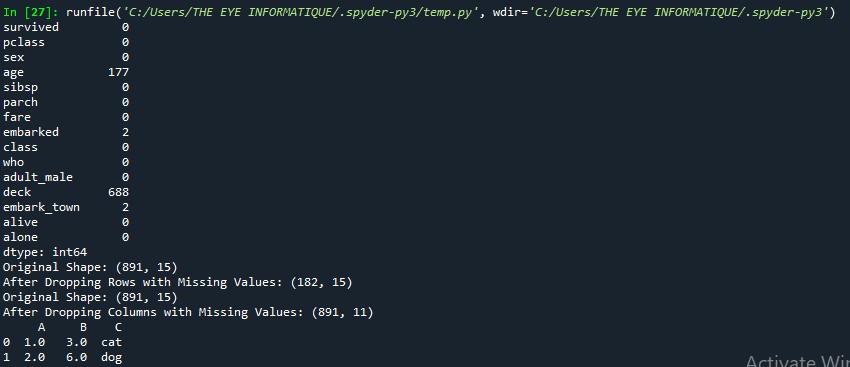
***Figure 6. this figure show us the missing value by dropping rows***

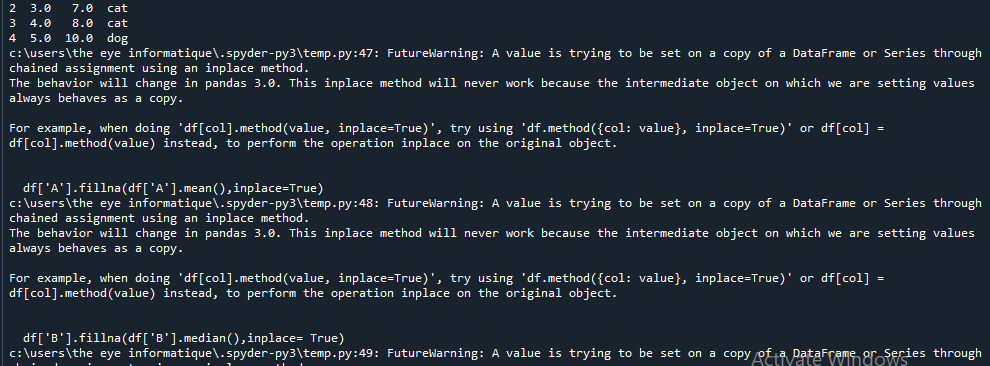
**2. Imputing Missing Values with Mean, Median, Mode**

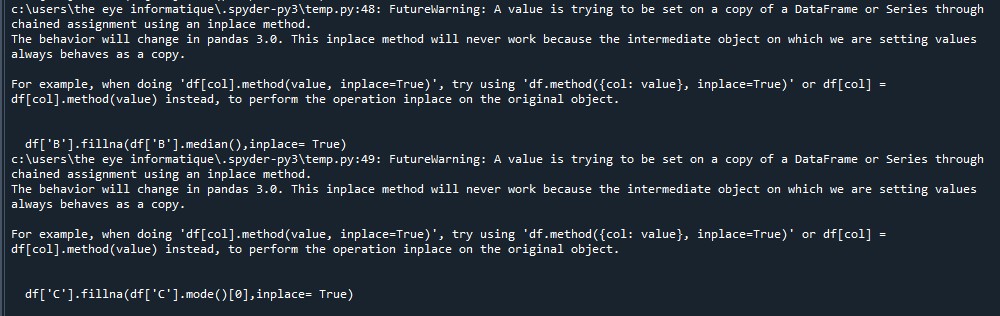
****

***Figure 7. This Code is to Compile the Mean, Median and Mode***

***OUTPUT:***

****

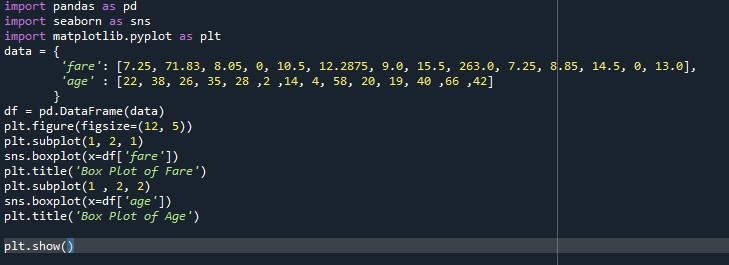
****

****

***Figure 8. from the above Diagram, the Data is been Collected, Compile and Giving us the result***

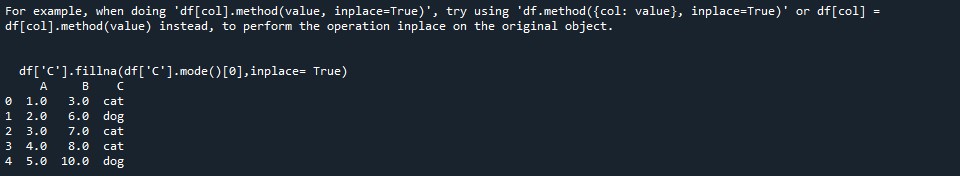
**STEP 3. Handling Outliers**

1.Using ‘fare’ and ‘age’ column(box plot)



***This figure 9 is going to show the result below concerning fare, age and boxplot.***

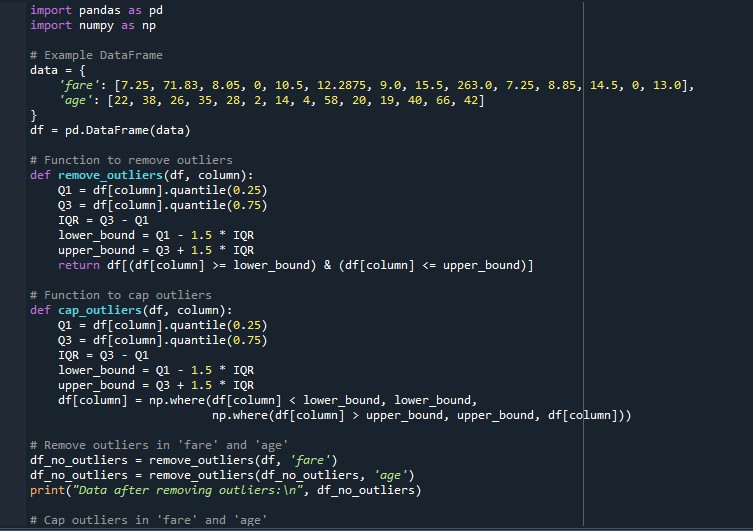
***OUTPUT:***



***Figure10 . This is the result of the above compilation***

**2. Capping and Removing Outliers**

This are two common techniques for handling outliers in data preprocessing.



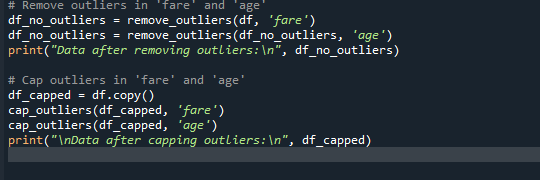
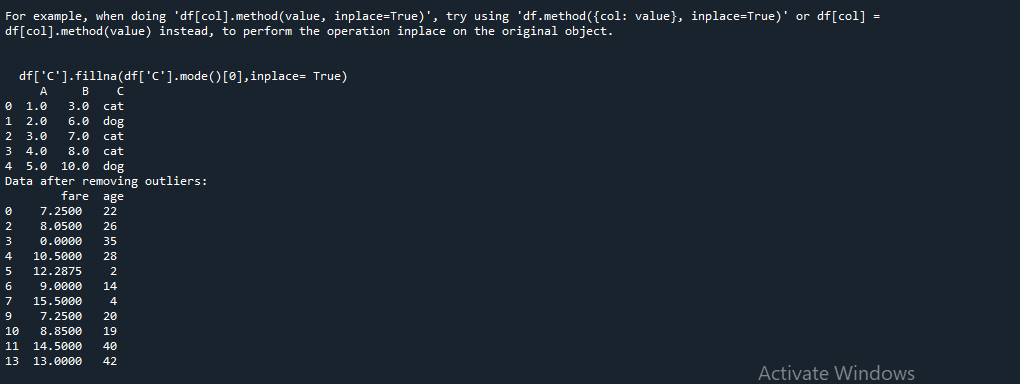
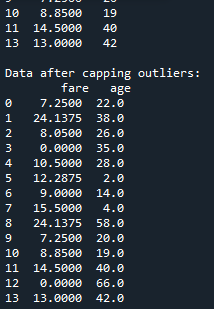


Figure 11: ***The above techniques help in improving models performance by reducing impact of outliers***

***OUTPUT:***

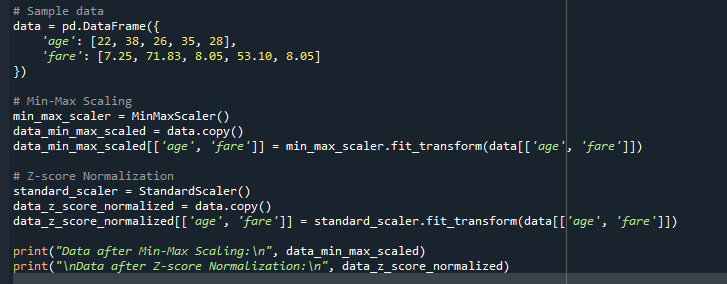
******

******

***Figure12: This figure shows how capping and removing outliers*** ***are been done in a dataset***

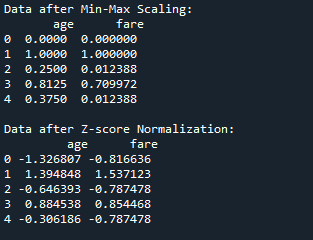
**STEP 4: Data Normalization**

*To normalize numerical features like age and fare, we can use either Min-max Scaling or Z-score normalization.*

******

***Figure13 .The above code is going to print the original values along with the normalized values for each method***

***OUTPUT :***

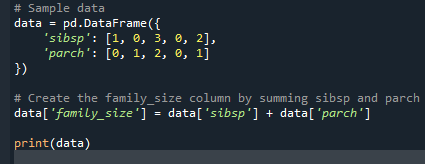
******

***.The above transform data by scaling feature to a fixed range, usually between 0 and 1***

**STEP 14: Feature Engineering**

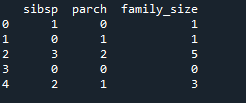
Feature engineering is the process of transforming raw data into meaningful features (variables or attributes) that can improve the performance of machine learning models. It involves creating, modifying, or selecting variables from the raw data that help the model capture patterns, relationships, and insights. Well-designed features can make a significant difference in the accuracy and effectiveness of a model.

a) Family size



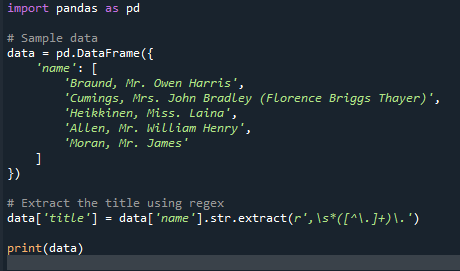
***Figure15: This will transform raw data in to meaningful features, well-designed features can make a significant difference in the accuracy and effectiveness of a model*.**

***OUTPUT:***

****

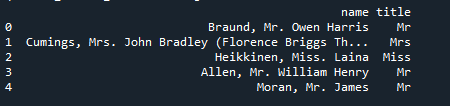
***Figure 16: This step is often time-consuming but critical as well-designed features improve model accuracy interpretability.***

***b) the extraction of columns***

******

***Figure17: This often look to pull out any prefixes from a person full name***

***OUTPUT:***

******

***Figure18: This approach isolates the title, allowing for cleaner data and easier sorting or filtering based on titles.***

**STEP 18: Features Selection**

I choose the correlation analysis technique to identify the most relevant features.

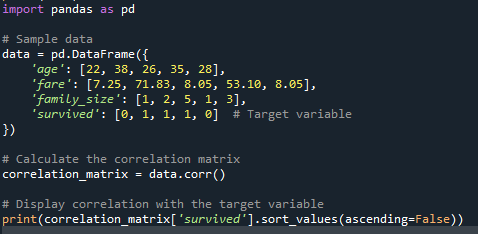
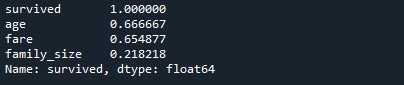


Figure19***: the above data will help in calculating the correlation coefficients.***

***OUTPUT:***

******

***Figure 20:This figure shows us the most relevant features find by the correlation analysis technique.***