Task3.1:

Data preprocessing is nothing but transforming/ converting raw data into suitable format for a machine learning model.
The dataset in our code is usually a CSV **file and may contain** incorrect, incomplete, inaccurate, missing data.

Two ways to handle missing data:

By deleting the particular row: The first way is used to commonly deal with null values. In this way, we just delete the specific row or column which consists of null values. But this way is not so efficient and removing data may lead to loss of information which will not give the accurate output.

By calculating the mean: In this way, we will calculate the mean of that column or row which contains any missing value and will put it on the place of missing value. This strategy is useful for the features which have numeric data such as age, salary, year, etc. Here, we will use this approach.

For removing specific symbol from the row or column:
Syntax: dataframe[columns].replace({symbol:},regex=True)
First, select the columns which have a symbol that needs to be removed. And inside the method replace() insert the symbol example replace("h":"")
Regular expressions (regex or regexp) are powerful tools used for pattern matching and searching within text. They are widely used in various programming languages and tools to perform tasks like validation, data extraction, and text manipulation.

4 Steps of data preprocessing:

1. Data Cleaning
2. Data Integration
3. Data Reduction
4. Data Transformation

Data Cleaning:
Process of removal of incorrect, incomplete, inaccurate data and also replacing missing data.

Missing data:
We can find missing value with "NA" and replace with mean values, or replace by median value.
We can also use most probable values for replacing.

We can remove these values from our dataframe too. Additionally some index values get affected and are not in order.

Error Data:

1. Binning: Sort data first.The sorted data is stored in bins.
   We can replace the error data using the mean value, median or boundary(min or max value).
2. Regression: Numerical Prediction of data
   The data to be predicted is a dependent variable and from which data is going to be calculated is an independent variable.
   Outliers are data which give wrong direction to the expected data. Outliers are very min and max data.
3. Clustering: Similar error items are grouped and removed all together

Task 3.2:

Diffusion Models work by destroying training data through the successive addition of Gaussian noise, and then learning to recover the data by reversing this noising process. Diffusion models encode images step by step adding more and more noise. Similarly decode by removing the noise in multiple steps to generate the image.
After training, we can use the Diffusion Model to generate data by simply passing randomly sampled noise through the learned denoising process.
The goal of training a diffusion model is to learn the reverse process, to generate new data

Prior matching
Reconstruction
Denoising Matching