**Batch:B4**

**RollNo.: 16010122221**

**Experiment02**

**Grade:AA/ AB / BB / BC / CC/ CD /DD**

**Title:**Datasetpreparing/pre-processing

# Objective:

1. **To learn how to prepare the dataset**

# To learn various steps in Data -Preprocessing

**Course Outcome:**

# CO1: Learn how to locate and download datasets, extract insights from that data and present their findings in a variety of different formats.

**Books/ Journals/ Websites referred:**

* 1. Data Visualization made simple New York: Routledge - Kristen Sosulski, First edition, 2019
  2. Sosulski, K. Data Visualization Made Simple: Insights into Becoming Visual, First edition, 2018
  3. <https://www.kaggle.com/uciml/adult-census-income>
  4. <https://archive.ics.uci.edu/ml/datasets/adult>
  5. [https://ori.hhs.gov/education/products/n\_illinois\_u/datamanagement/dctopic.ht](https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/dctopic.html) [ml](https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/dctopic.html)
  6. A review of research process, data collection and analysis - Surya Raj Niraula
  7. <https://www.jigsawacademy.com/blogs/data-science/what-is-data-processing/>

# Resources used:

1. <https://www.kaggle.com/uciml/adult-census-income>
2. <https://archive.ics.uci.edu/ml/datasets/adult>

# Theory (About Data Preprocessing):

Concept of Data processing is collecting and manipulating data into a usable and appropriate form. The automatic processing of data in a predetermined sequence of operations is the manipulation of data.

Data Preprocessing is a Data Mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends and is likely to contain many errors. Data Preprocessing is a proven method of resolving such issues. we need to transform or organize it to make it into a proper format by using Data Preprocessing.

Preprocessing of data is mainly to check the data quality. The quality can be checked by the following:

* Accuracy: To check whether the data entered is correct or not.
* Completeness: To check whether the data is available or not recorded.
* Consistency: To check whether the same data is kept in all the places that do or do not match.
* Timeliness: The data should be updated correctly.
* Believability: The data should be trustable.
* Interpretability: The understandability of the data.

*Major Tasks in Data Preprocessing:*

1. Data cleaning: process to remove incorrect data, incomplete data and inaccurate data from the datasets, including removing missing values.
2. Data integration: process of combining multiple sources into a single dataset.
3. Data reduction: process helps in the reduction of the volume of the data which makes the analysis easier.
4. Data transformation: process in which change is made in the format or the structure of the data.

# WHY WE SHOULD USE DATA PROCESSING

In the modern era, most of the work relies on data, therefore collection of large amounts of data for different purposes. The processing of this data collected is essential so that the data goes through all the above-stated steps and gets sorted, stored, filtered, presented in the required format and analyzed.

# IMPLEMENTATION:

*Working (Put the code and Output for each Data Preprocessing task):*

Different steps in Data Preprocessing:

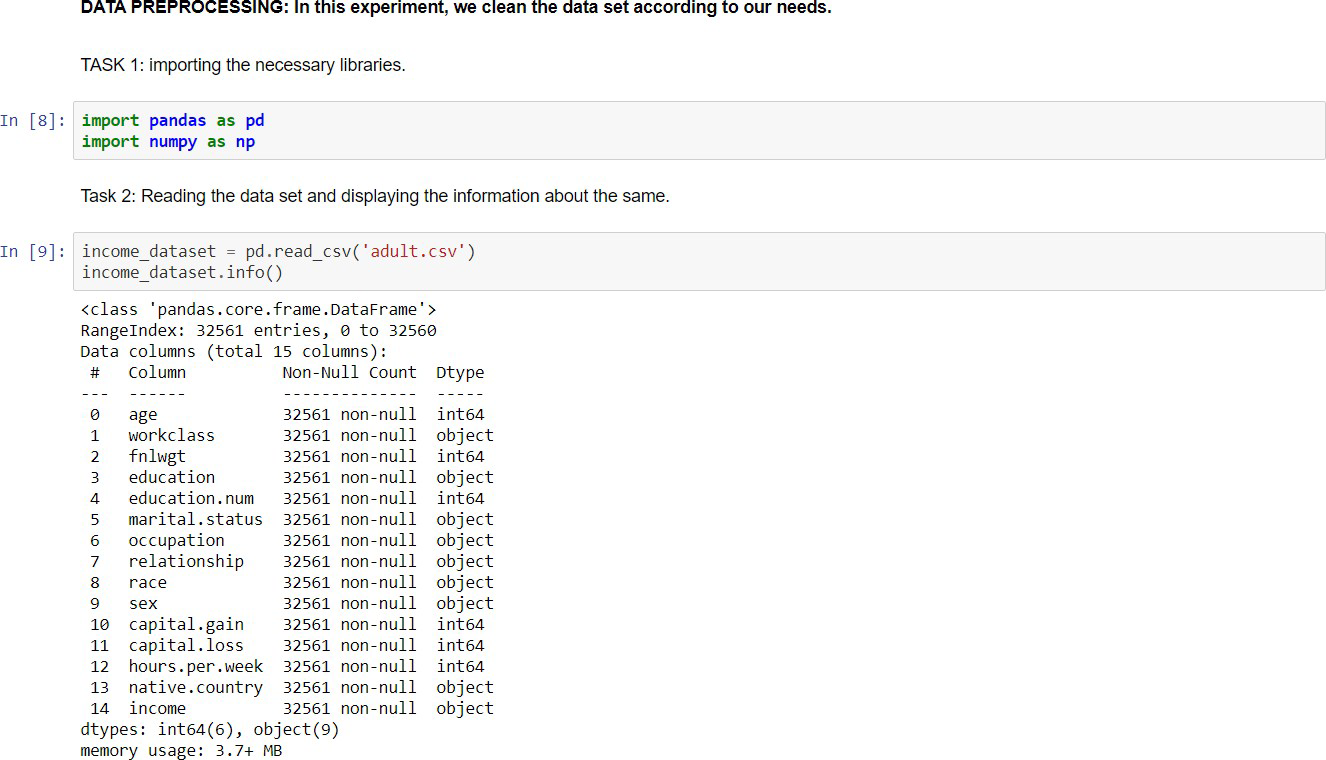
# Finding missing, null values etc.

* + **Replacing missing, null values with statistical parameters.**

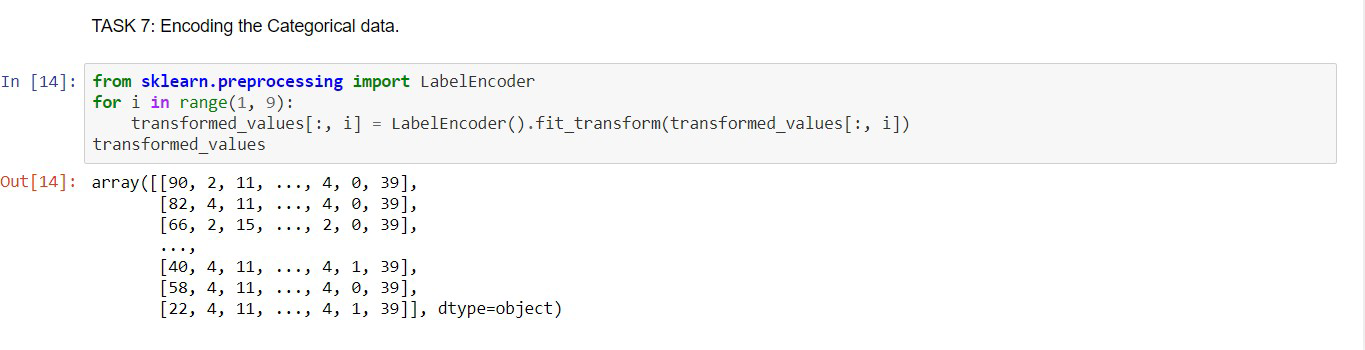
# Encoding categorical data

* + **Normalization**

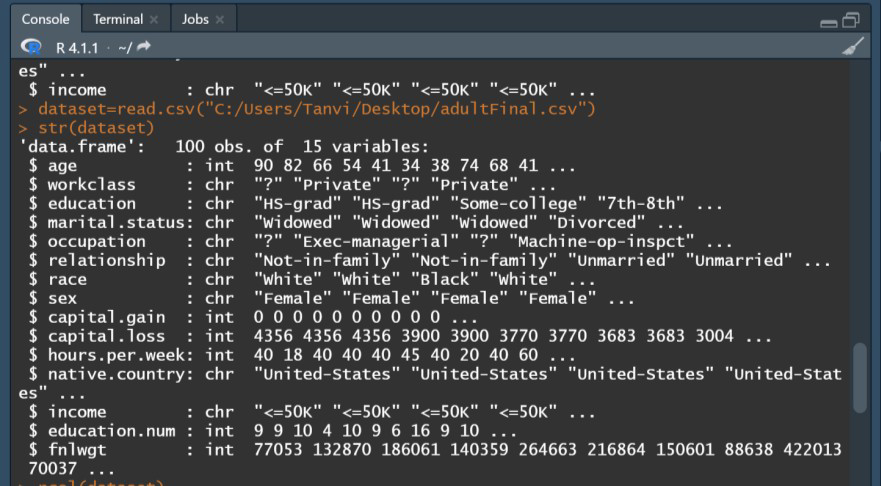
# PYTHON



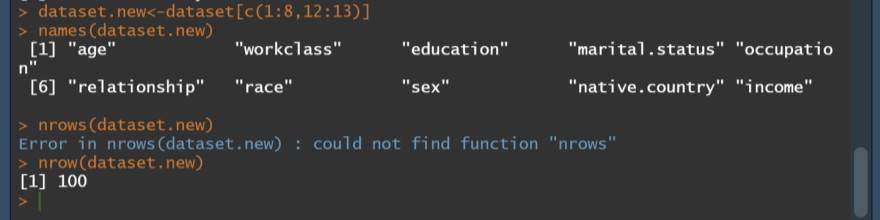




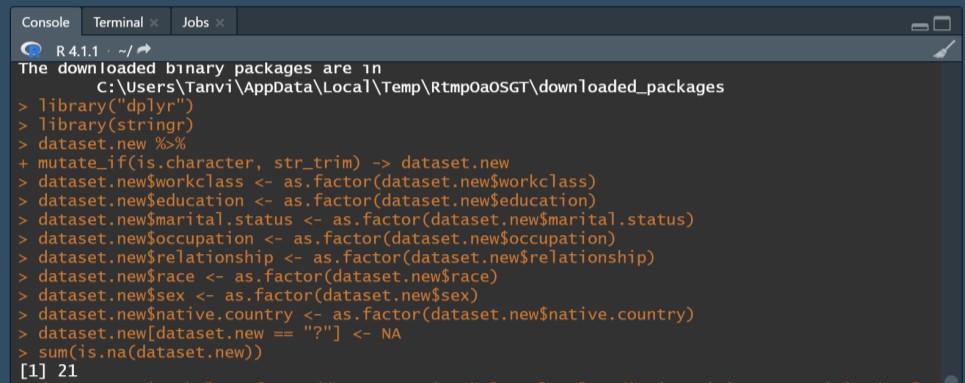
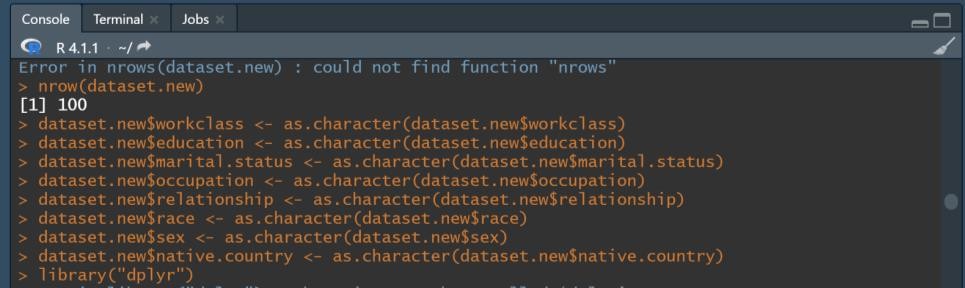
1. **RStudio**

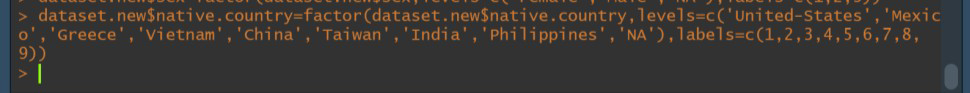
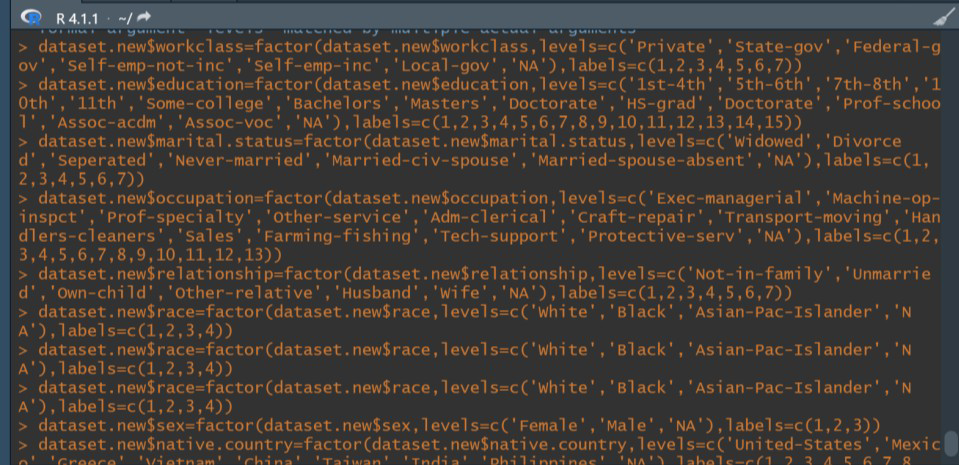
Task 1: Reading the dataset and displaying it.

Task 2: Removing the columns that are not required (Creating a new dataset and storing the required columns)



Task 3: Taking care of missing data be replacing “?” with NA

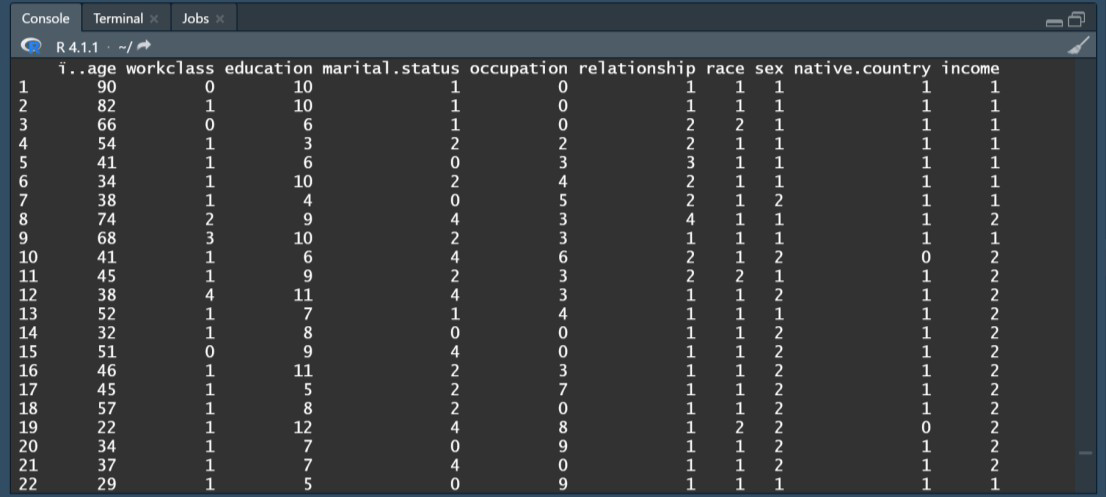


Task 4: Encoding the Categorical Data

Task 5: NA to 0



Final: (Result from 1 to 20 observations)



# Conclusion (Students should write in their own words):

Through this experiment, we were working on a chosen dataset. We learnt about data processing methods i.e., cleaning, integrations, reduction, transformations; and steps and implemented the same on our “Adult Income” dataset.

# Post Lab Question:

**1. Write the importance of Data Preprocessing in Software System Designing**

When exploring this wealth of information data pre-processing cleans and prepares the data before predictive models are developed. Predictions from incorrect data can be difficult to debug, or worse, can lead to inaccurate or misleading results that impact system performance and reliability. The goal here is to find the most predictive features

of the data and filter it so it will enhance the predictive power of the analytics model. Some common techniques include feature selection to reduce high-dimension data, feature extraction and transformation for dimensionality reduction, and domain analysis such as signal, image, and video processing.

The information gathered from data pre-processing is then taken and implemented across a number of analytics-driven embedded systems. An example of this is the innovation in using Big Data and analytics to make cars smarter. Automotive OEMs are collecting enormous amounts of data from real-world driving situations (think millions of miles of driving), recording data such as engine performance, video, radar, and other signals. This data is used to generate important metrics such as fuel economy and performance at the fleet level. Engineering teams are also using this real-world data to design, develop, and test new types of automotive systems, such as advanced driver assistance systems (ADAS).

.