**SVKM’s NMIMS**

**Mukesh Patel School of Technology Management & Engineering**

Program: B.Tech\MBA.Tech

**Course: Machine Learning**

**Experiment No.01**

PART A

(PART A : TO BE REFFERED BY STUDENTS)

**A.1 Aim: To understand and implement data exploration techniques using Pandas Library.**

**Task 1:** **Perform Exploratory data analysis on Car dataset and write the inferences for each question.**

1. Read the Toyota.csv file into a DataFrame.
2. Explore size, shape, data types of each column in the dataset.
3. List down the columns of dataset
4. Find out ‘Fuel Type’ for the 4th row.
5. Find out value for second column for the 4th row.
6. Select all rows for column “Fuel Type”
7. Select all rows for columns “KM”, “HP” and “Automatic”
8. Display 1 to 5 rows for columns 2 to 4 (excluding row 5 and column 4)
9. Display the info of dataset and state your observations
10. Identify unique values for columns “KM”, “HP” and “Doors”
11. Create a new data frame, by replacing “??” with NAN
12. Replace the categorical values in the “Doors” column with its corresponding numeric value
13. Convert data types of columns “Doors”, “MetColor” and “Automatic” to int, and object
14. Identify the total number of null values in each column of the data set
15. Drop rows with null values
16. Identify total number of cars that runs on “Petrol”, “Diesel” or “CNG”
17. Identify mean of “KM” for the cars that runs on “Diesel”

**Task 2:**

**Perform one hot and label encoding on relationship column of “adults” dataset**

**A.2 Prerequisite:**

Python Programming, Pandas library

**A.3 Outcome:**

**After successful completion of this experiment students will be able to:**

* 1. Read different types of data files(csv, excel, text file etc.)
  2. Obtain metadata of given dataset
  3. Understand finding of null values and replacing null values
  4. Understand and implement class label encoding
  5. Understand and implement one hot encoding

**A.4 Theory:**

**Exploratory Data Analysis:**

Exploratory Data Analysis (EDA) is an open-ended process where we calculate statistics and make figures to find trends, anomalies, patterns, or relationships within the data. The goal of EDA is to learn what our data can tell us. It generally starts out with a high level overview, then narrows in to specific areas as we find intriguing areas of the data. The findings may be interesting in their own right, or they can be used to inform our modeling choices, such as by helping us decide which features to use.

**Pandas Library:**

**DataFrame** is a 2-dimensional labeled data structure with columns of potentially different types. You can think of it like a spreadsheet or SQL table, or a dict of Series objects. It is generally the most commonly used pandas object. Like Series, DataFrame accepts many different kinds of input:

* Dict of 1D ndarrays, lists, dicts, or Series
* 2-D numpy.ndarray
* [Structured or record](http://docs.scipy.org/doc/numpy/user/basics.rec.html) ndarray
* A Series
* Another DataFrame

**Encoding:**

**One hot encoding:**

 One-hot encoding converts the categorical data into numeric data by splitting the column into multiple columns. The numbers are replaced by 1s and 0s, depending on which column has what value

**Label encoding:**

This approach is very simple and it involves converting each value in a column into a number.

PART B

(PART B : TO BE COMPLETED BY STUDENTS)

***(Students must submit the soft copy as per following segments within two hours of the practical.)***

|  |  |
| --- | --- |
| Roll No. N052 | Name: Pratyush Kumar |
| Class: MBA Tech CE | Batch: B2 |
| Date of Experiment: 14-12-2023 | Date of Submission: 21-12-2023 |
| Grade : |  |

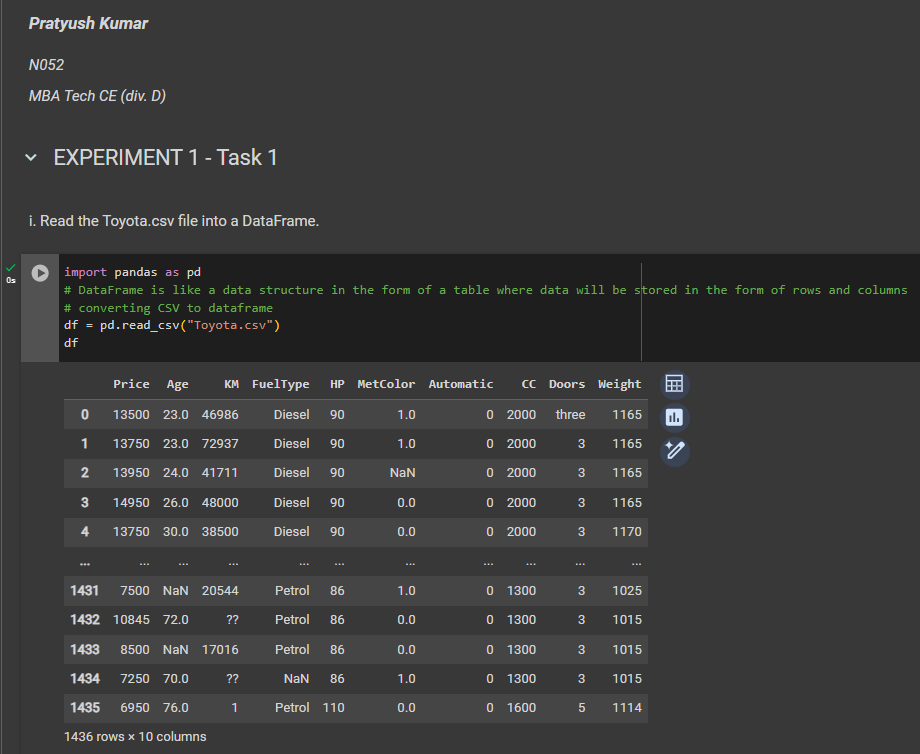
**B.1 Task1**

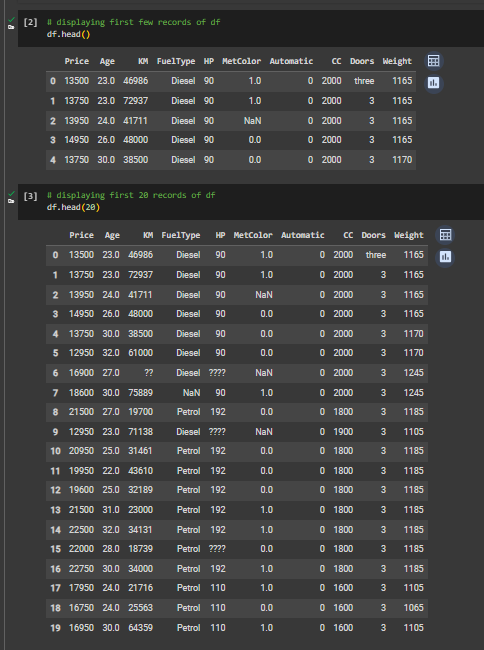
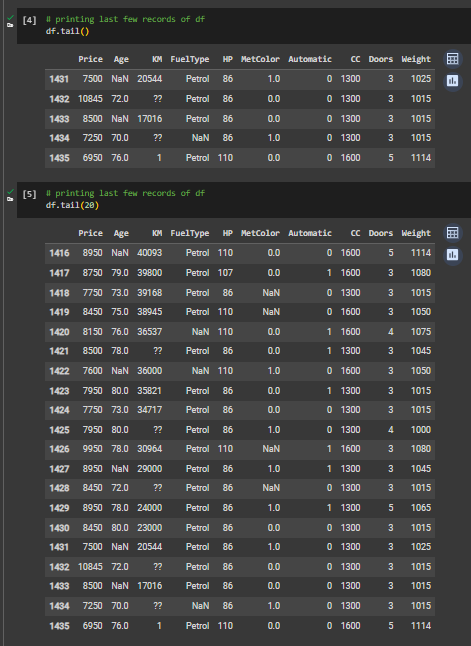
**Colab link:** [**https://colab.research.google.com/drive/12saxmansdQRF95O6XYGPfj0sK1hrlFOr?usp=sharing**](https://colab.research.google.com/drive/12saxmansdQRF95O6XYGPfj0sK1hrlFOr?usp=sharing)

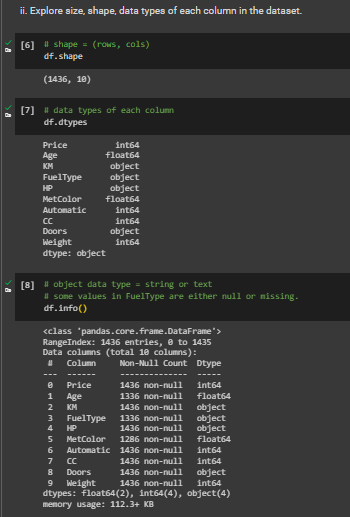
* **Source Code**

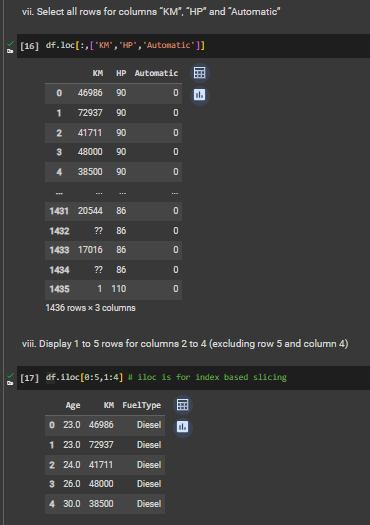
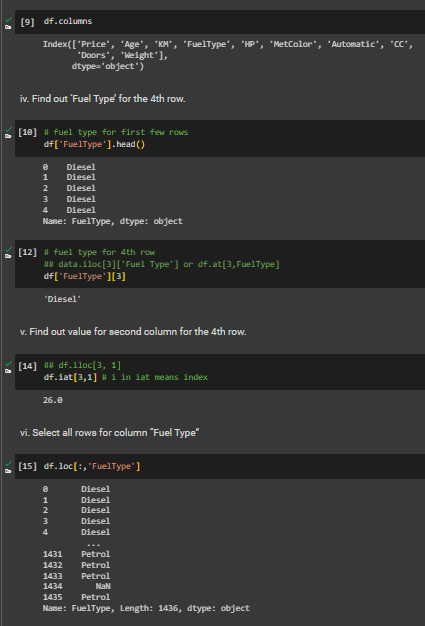
*"""  
 \* This file contains code snippets for performing exploratory data analysis on car dataset  
 \* ML-E1-Task1  
 \*   
 \* Original file is located at: https://colab.research.google.com/drive/12saxmansdQRF95O6XYGPfj0sK1hrlFOr  
 \* @author Pratyush Kumar (github.com/pratyushgta)  
"""*"i. Read the Toyota.csv file into a DataFrame."  
import pandas as pd  
  
# DataFrame is like a data structure in the form of a table where data will be stored in the form of rows and columns  
# converting CSV to dataframe  
df = pd.read\_csv("Toyota.csv")  
df  
# displaying first few records of df  
df.head()  
# displaying first 20 records of df  
df.head(20)  
# printing last few records of df  
df.tail()  
# printing last few records of df  
df.tail(20)  
  
"ii. Explore size, shape, data types of each column in the dataset."  
# shape = (rows, cols)  
df.shape  
# data types of each column  
df.dtypes  
# object data type = string or text  
# some values in FuelType are either null or missing.  
df.info()  
  
"iii. List down the columns of dataset"  
df.columns  
  
"iv. Find out ‘Fuel Type’ for the 4th row."  
# fuel type for first few rows  
df['FuelType'].head()  
# fuel type for 4th row  
## data.iloc[3]['Fuel Type'] or df.at[3,FuelType]  
df['FuelType'][3]  
  
"v. Find out value for second column for the 4th row."  
## df.iloc[3, 1]  
df.iat[3, 1] # i in iat means index  
  
"vi. Select all rows for column “Fuel Type”"  
df.loc[:, 'FuelType']  
  
"vii. Select all rows for columns “KM”, “HP” and “Automatic”"  
df.loc[:, ['KM', 'HP', 'Automatic']]  
  
"viii. Display 1 to 5 rows for columns 2 to 4 (excluding row 5 and column 4)"  
df.iloc[0:5, 1:4] # iloc is for index based slicing  
  
"ix. Display the info of dataset and state your observations"  
df.info()  
df.dtypes  
"""  
Observations:  
1. Dataset contains integer, object (string) and float datatypes  
2. Some values in FuelType are either null or missing.  
3. Doors column appears to have integer values which are represented as string in dataset (first value is string and rest are numeric values with datatype object)  
  
x. Identify unique values for columns “KM”, “HP” and “Doors”  
"""  
  
df['KM'].unique()  
  
df['KM'].head(50).unique()  
  
df['HP'].head(50).unique()  
  
df['Doors'].head(50).unique()  
  
"xi. Create a new data frame, by replacing “??” with NAN"  
  
# df\_new = df.replace('??',"NAN", inplace=True)  
df\_new = pd.read\_csv('Toyota.csv', na\_values=["??", "????"])  
  
df\_new  
  
df\_new.info()  
  
# comparing null value representation in original and modified dataset  
print("Unique values of HP in df:", df['HP'].unique())  
print("Unique values of HP in df\_new:", df\_new['HP'].unique())  
  
"""xii. Replace the categorical values in the “Doors” column with its corresponding numeric value"""  
  
# replacing string in Doors with its corresponding numerical values  
# door\_mapping = {'two': 2, 'three': 3, 'four': 4, 'five': 5}  
# data['Doors'] = data['Doors'].replace(door\_mapping) or #df3=df\_new['Doors]  
df\_new['Doors'].replace('three', '3', inplace=True)  
df\_new['Doors'].replace('four', '4', inplace=True)  
df\_new['Doors'].replace('five', '5', inplace=True)  
  
df\_new.head()  
  
"""xiii. Convert data types of columns “Doors”, “MetColor” and “Automatic” to int, and object"""  
df\_new['Doors'].dtypes  
# before conversion  
df\_new.info()  
  
df\_new['Doors'] = df\_new['Doors'].astype('int64')  
df\_new['MetColor'] = df\_new['MetColor'].astype('object')  
df\_new['Automatic'] = df\_new['Automatic'].astype('object')  
  
# after conversion  
df\_new.info()  
df\_new['Doors'].dtypes  
df\_new.head()  
  
"xiv. Identify the total number of null values in each column of the data set"  
  
# isnull returns true if null is present in each column  
df\_new.isnull()  
  
df\_new.isnull().sum()  
  
"xv. Drop rows with null values"  
  
df\_new.shape  
df\_new = df\_new.dropna()  
df\_new.shape  
  
"xvi. Identify total number of cars that runs on “Petrol”, “Diesel” or “CNG”"  
df\_new['FuelType'].unique()  
df\_new['FuelType'].value\_counts()  
  
"xvii. Identify mean of “KM” for the cars that runs on “Diesel”"  
# df\_new[df\_new['FuelType']=='Diesel'].KM.mean()  
df\_new[df\_new.FuelType == 'Diesel'].KM.mean()

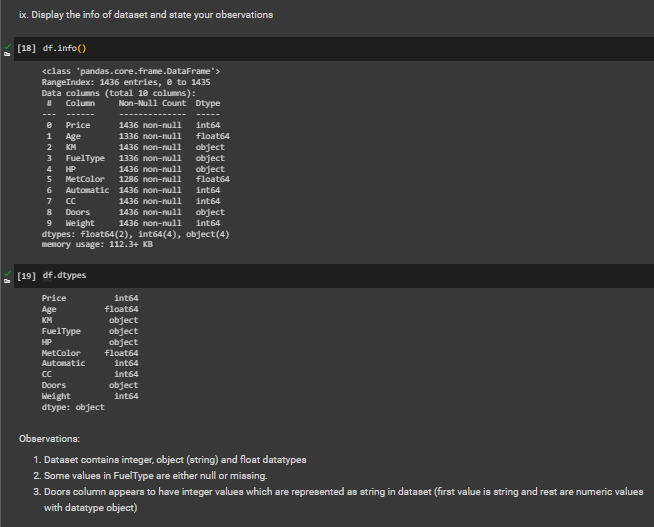
* **Input/ Output**

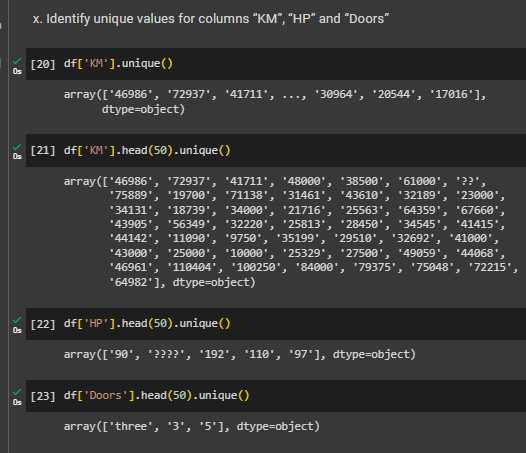
****

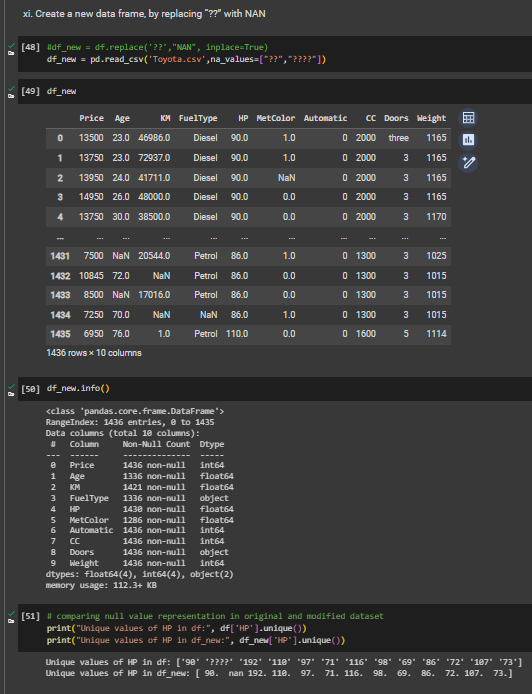
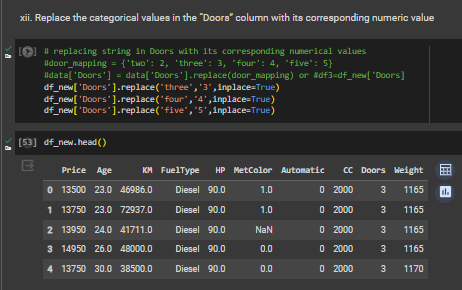
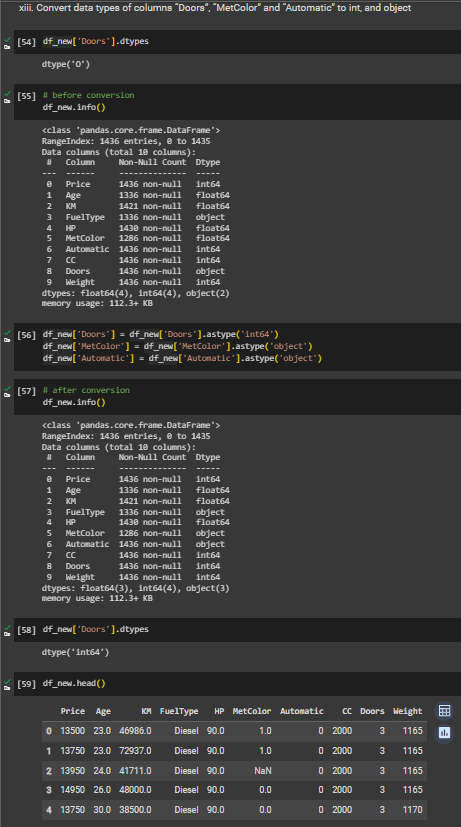
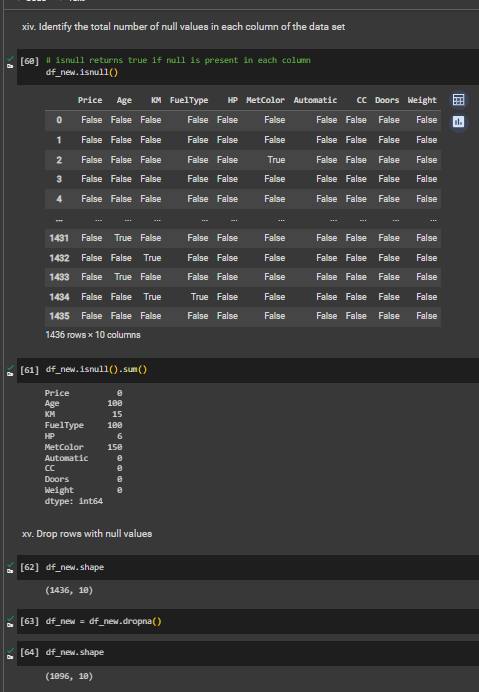
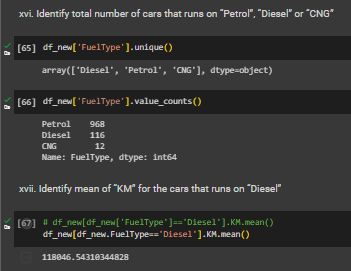
****

****

****

****

****

****

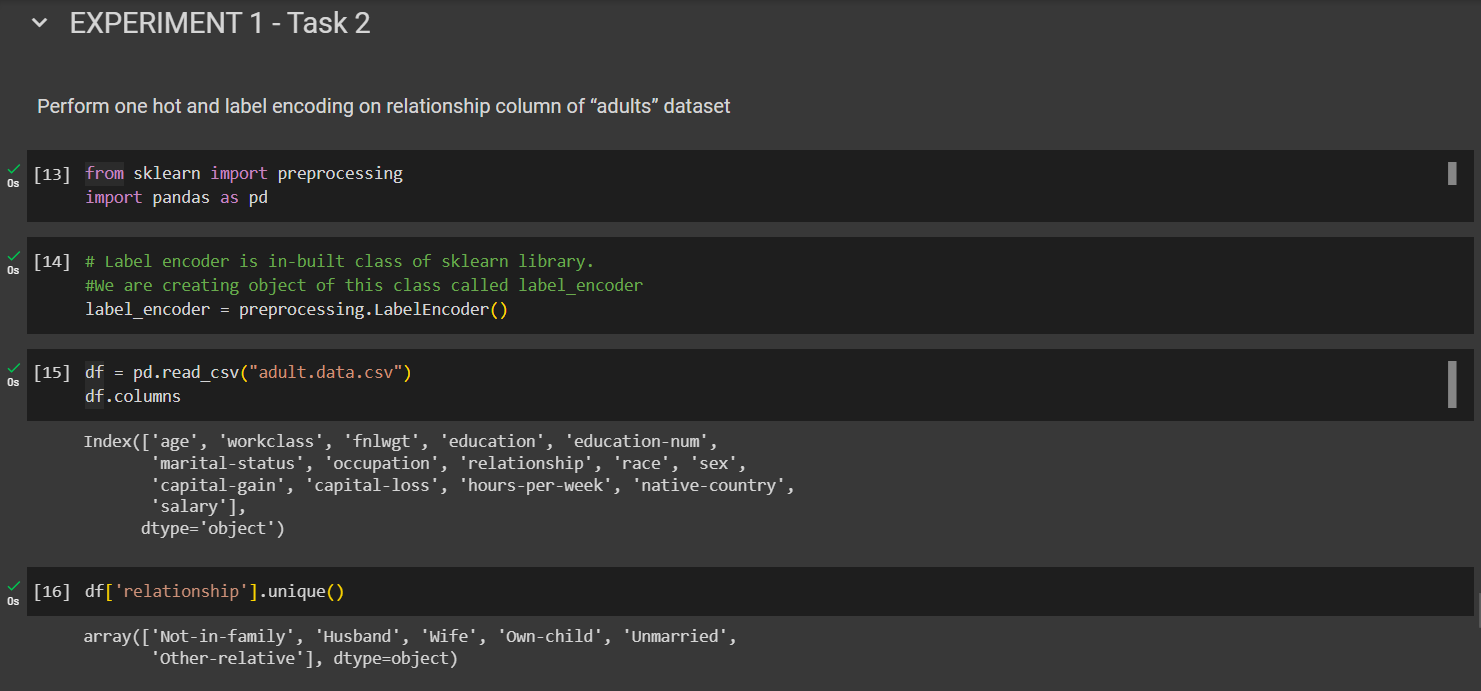
**B.2 Task 2**

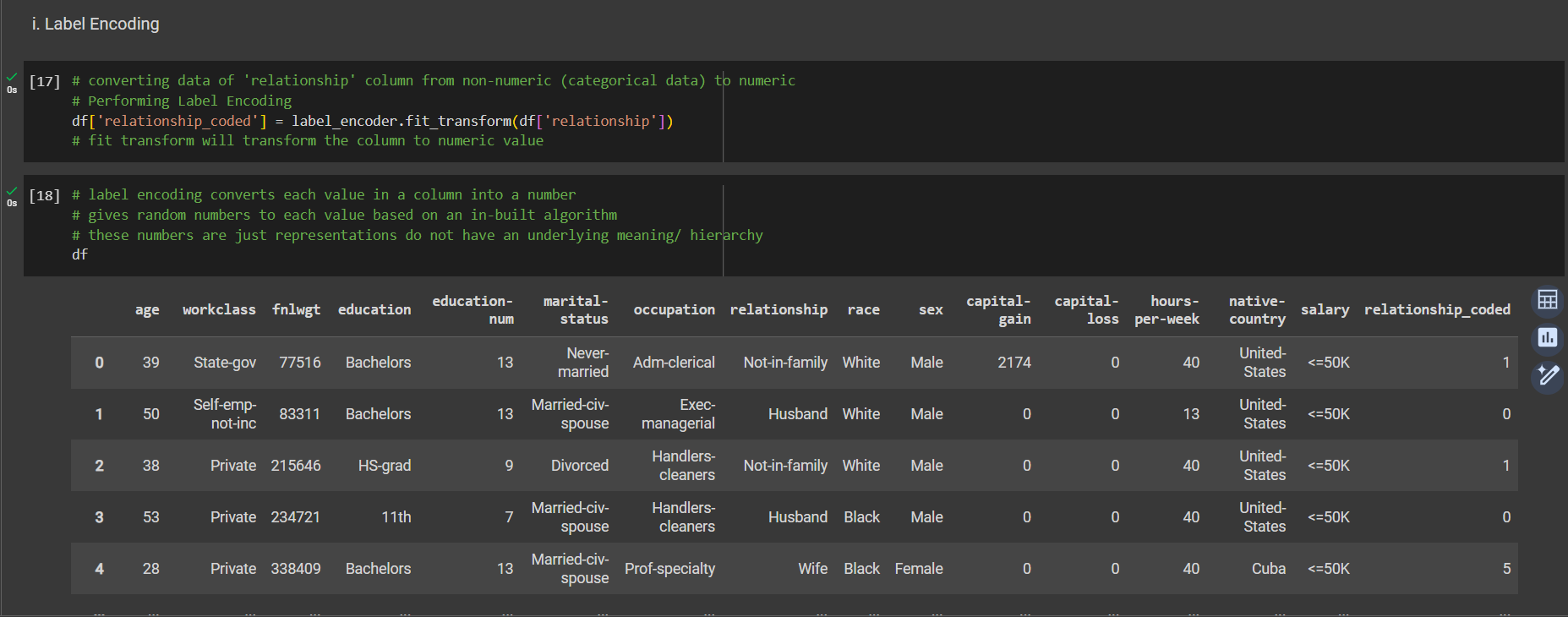
**Colab link:** [**https://colab.research.google.com/drive/12saxmansdQRF95O6XYGPfj0sK1hrlFOr?usp=sharing**](https://colab.research.google.com/drive/12saxmansdQRF95O6XYGPfj0sK1hrlFOr?usp=sharing)

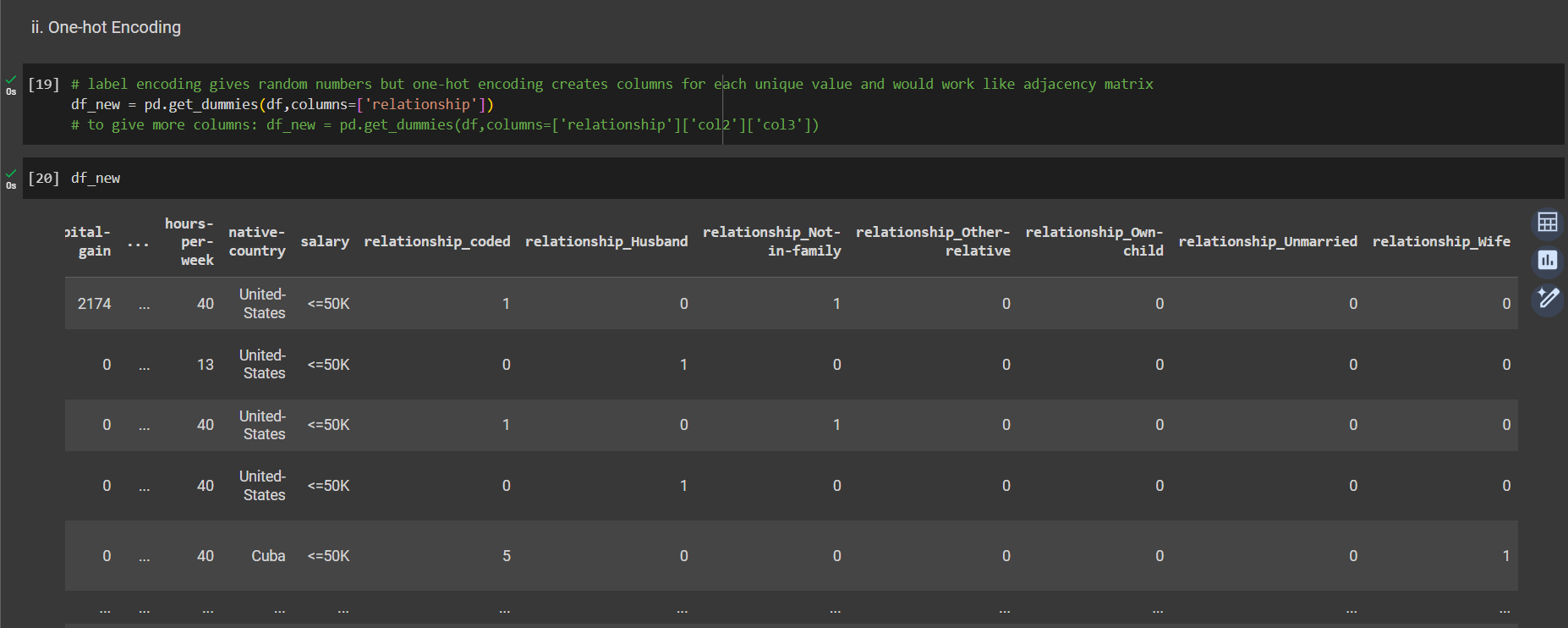
* **Source Code**

*"""  
 \* This file contains code snippets for performing exploratory data analysis on adult dataset  
 \* ML-E1-Task2  
 \*  
 \* Original file is located at: https://colab.research.google.com/drive/12saxmansdQRF95O6XYGPfj0sK1hrlFOr  
 \* @author Pratyush Kumar (github.com/pratyushgta)  
"""*"Perform one hot and label encoding on relationship column of “adults” dataset"  
from sklearn import preprocessing  
import pandas as pd  
  
# Label encoder is in-built class of sklearn library.  
# We are creating object of this class called label\_encoder  
label\_encoder = preprocessing.LabelEncoder()  
  
df = pd.read\_csv("adult.data.csv")  
df.columns  
  
df['relationship'].unique()  
  
"i. Label Encoding"  
  
# converting data of 'relationship' column from non-numeric (categorical data) to numeric  
# Performing Label Encoding  
df['relationship\_coded'] = label\_encoder.fit\_transform(df['relationship'])  
# fit transform will transform the column to numeric value  
  
# label encoding converts each value in a column into a number  
# gives random numbers to each value based on an in-built algorithm  
# these numbers are just representations do not have an underlying meaning/ hierarchy  
df  
  
"ii. One-hot Encoding"  
  
# label encoding gives random numbers but one-hot encoding creates columns for each unique value and would work like adjacency matrix  
df\_new = pd.get\_dummies(df,columns=['relationship'])  
# to give more columns: df\_new = pd.get\_dummies(df,columns=['relationship']['col2']['col3'])  
  
df\_new

* **Input/ Output**

****

****

****

**B.4 Conclusion:**

*(Students must write the conclusion in their own words.)*

Explored and analyzed the ‘Toyota’ car dataset using Python and Pandas library. Conducted exploratory data analysis on the dataset to gain insights and draw inferences from the data. The analysis included examining the dataset's size, shape, and data types of each column, exploring specific rows and columns, replacing categorical values with numeric equivalents, converting data types of columns, and identifying the count of null values in each column.

Performed one-hot and label encoding on the ‘adults’ dataset relationship column to transform categorical data into numerical representations suitable for ML algorithms.