# NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

# 

# BIG DATA - LA - 2 - PROGRAMMING ASSIGNMENT

# BY -

## **KSHITIZ KUMAR - 1NT19IS073 PIYUSH GEETESH - 1NT19IS109 RAGHUNATHAN R - 1NT19IS122 MUSTAQEEM AHMED - 1NT19IS083** [ **SEMESTER - 6 - B - SECTION ]**

## **TO -**

## **Dr. K ADITYA SHASTRY**

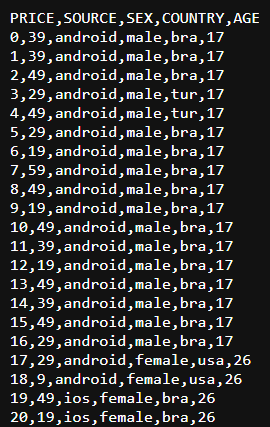
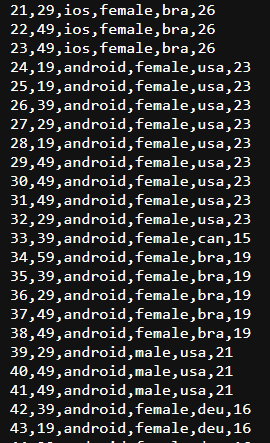
# Q1. Implement a Rule based classifier for performing classification

## **Explanation of Algorithm**

One of game company would like to learn how much money its new customers will potentially spend by observing the purchasing movements within the application

The variables used in our dataset are:

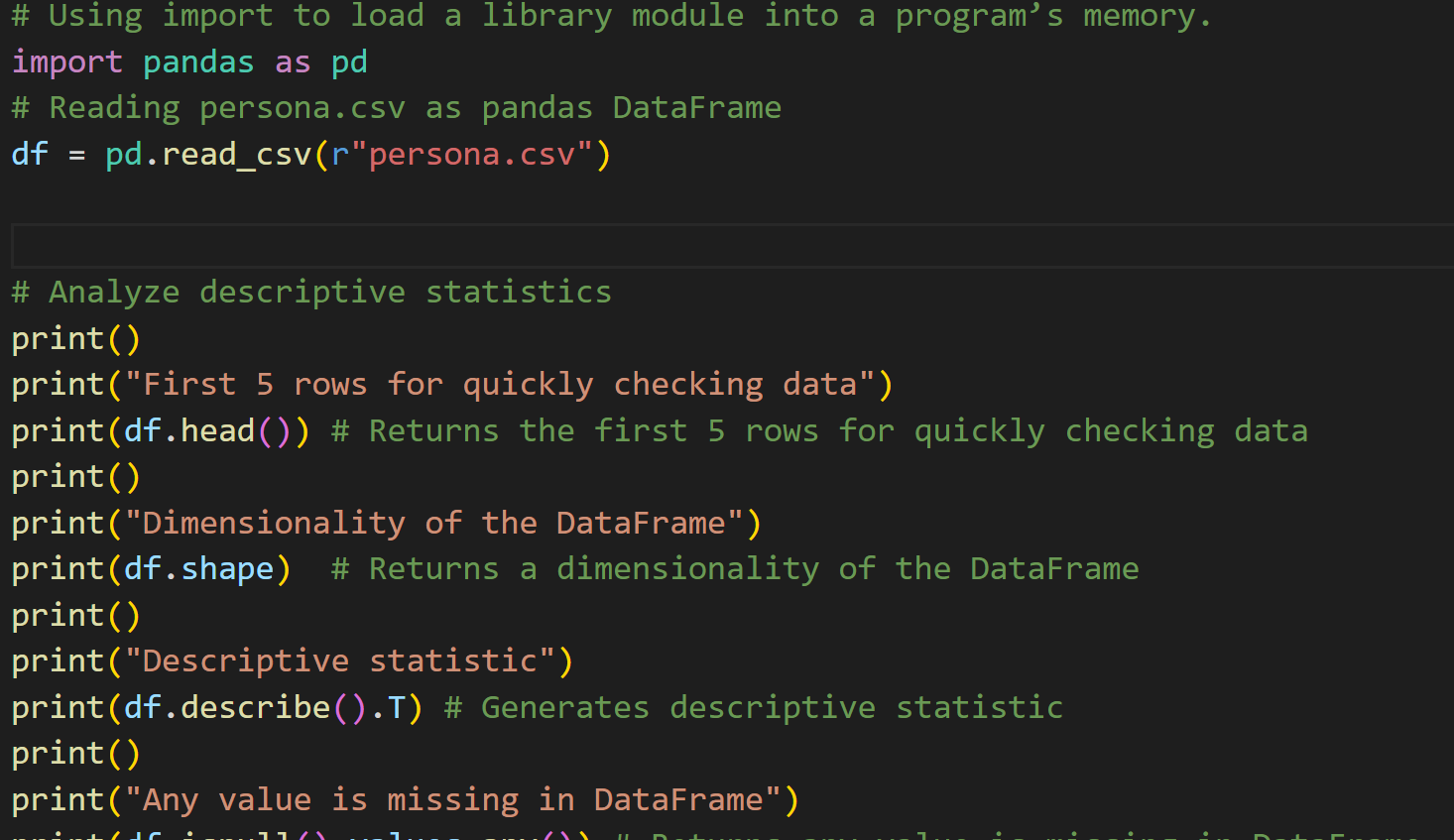
1. Price: Payments made by costumers
2. Source: The operating system used, including IOS and Android
3. Sex: Gender of users, Female and Male
4. Country : Information from which country the costumers are from
5. Age: Age of costumers

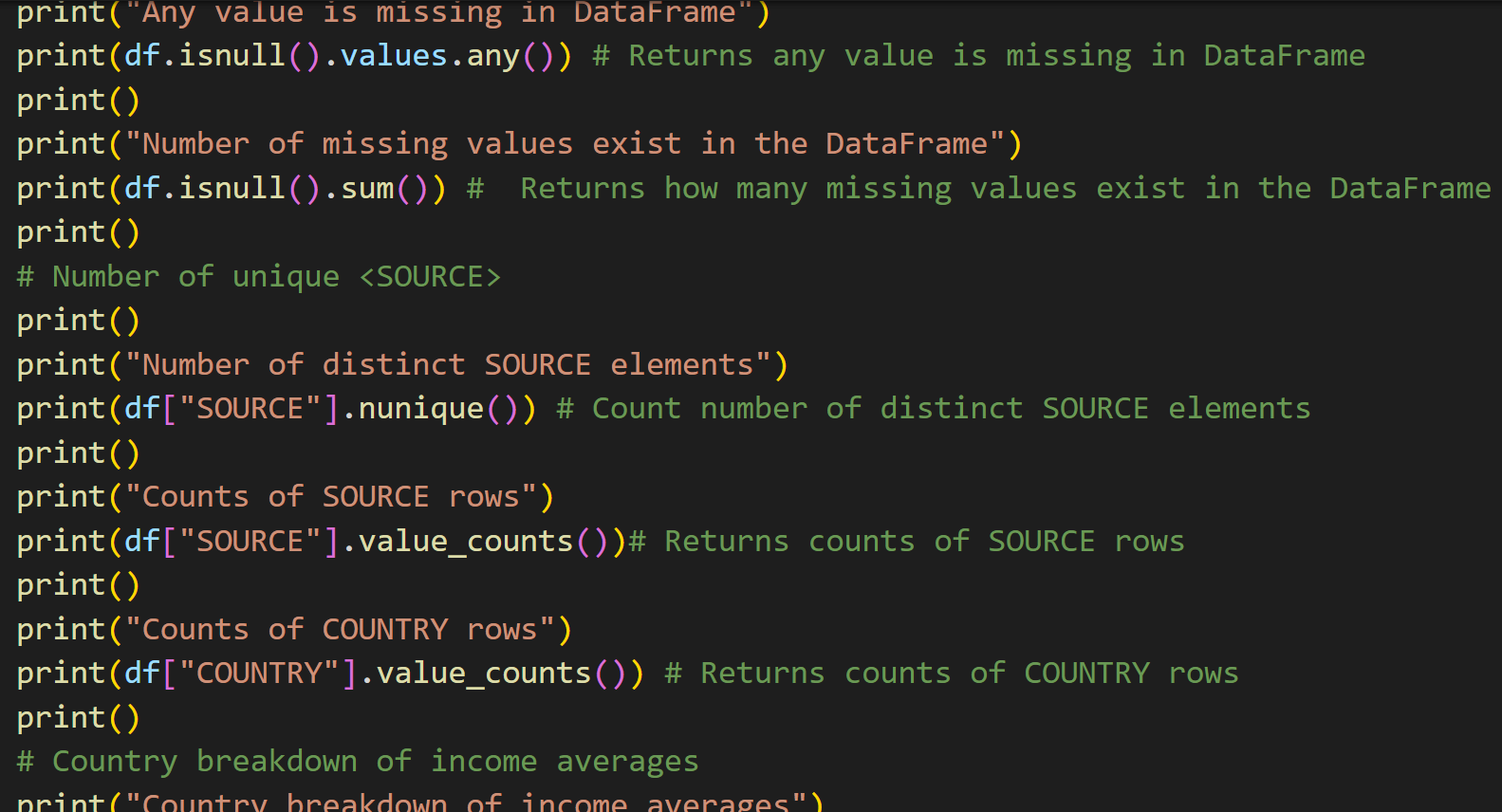
 

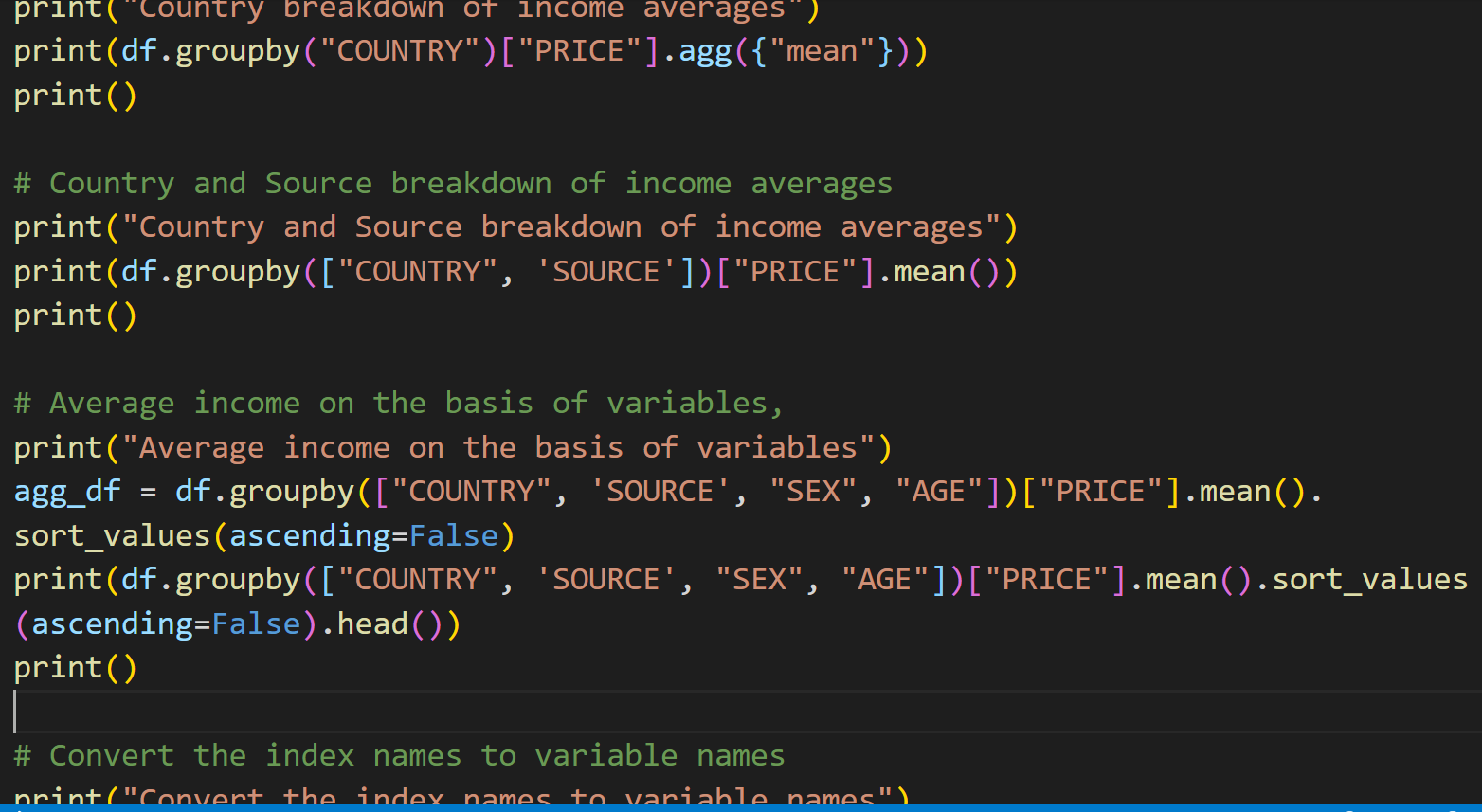
Algorithm:

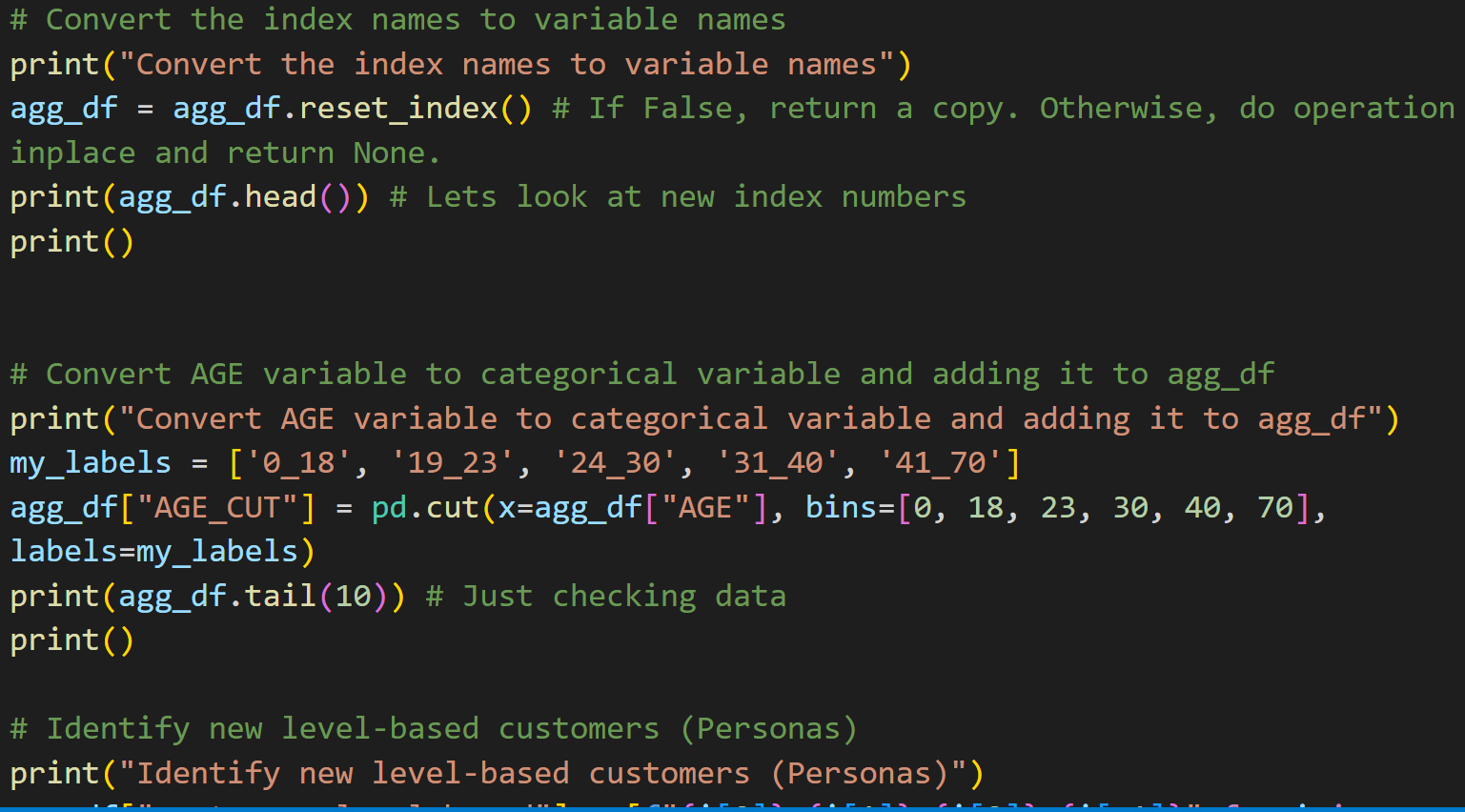
1. Import panda libray.
2. Read persona.csv file as pandas DataFrame.
3. The average gains within the combinations into our table in a DataFrame type named agg\_df is taken by grouping many variables in our data.
4. Convert AGE variable to categorical variable and adding it to agg\_df.
5. The personas are divided i into 4 groups from the least profitable customer group to the most profitable customer group. “D, C, B, A”
6. Based on the price of the new record into the segmented data, its group is determined.

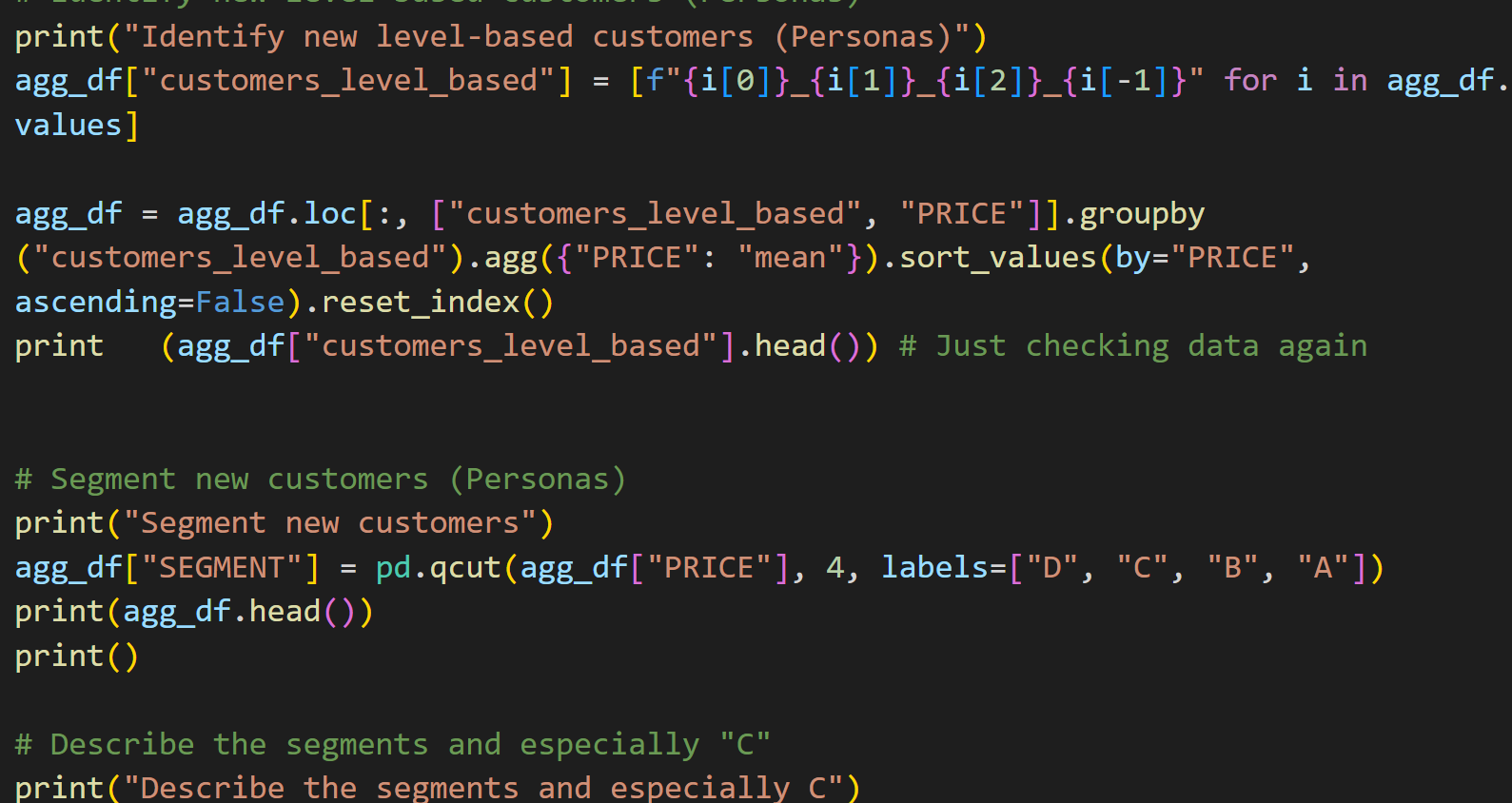
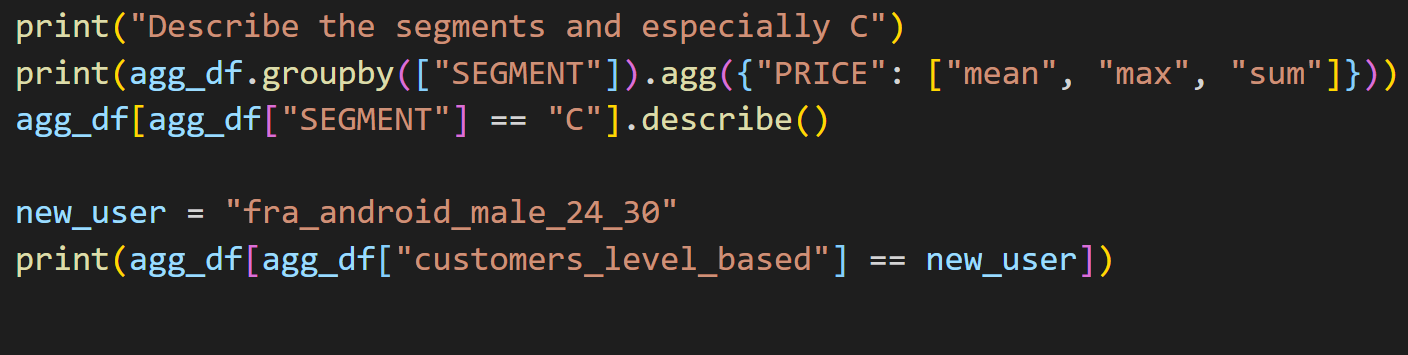
## **Source code**







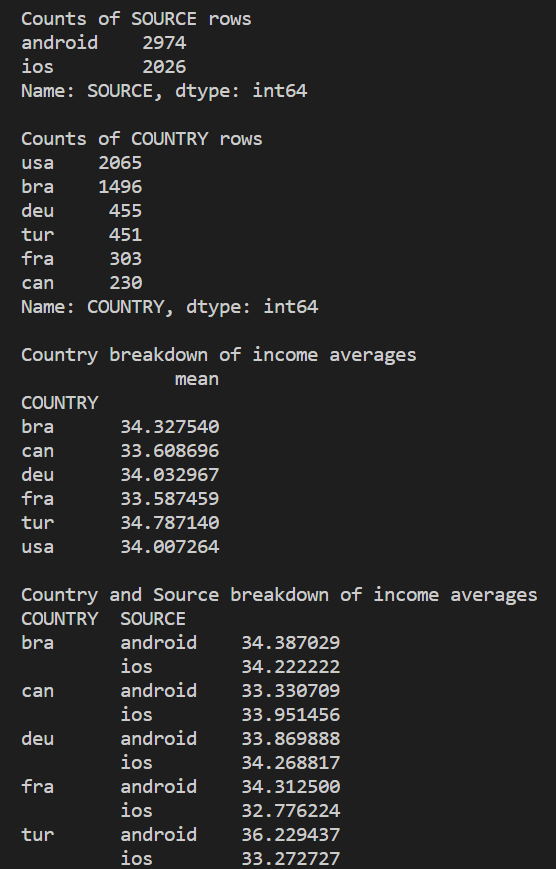


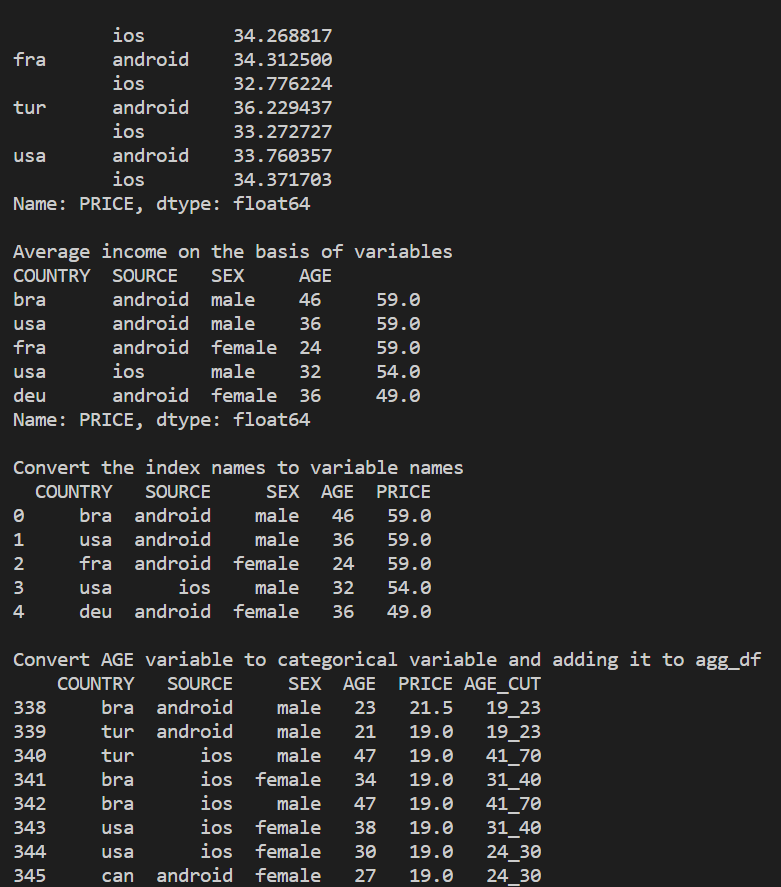
 

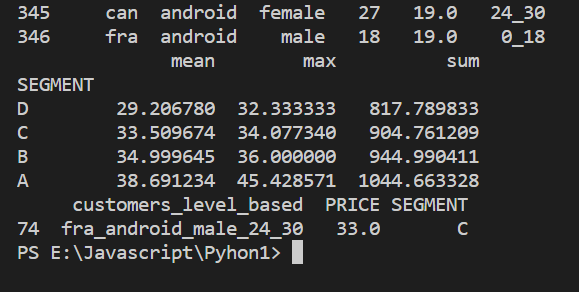
## 

## **Screenshots of Output**

## 







# Q2. Develop a program to implement Binarization

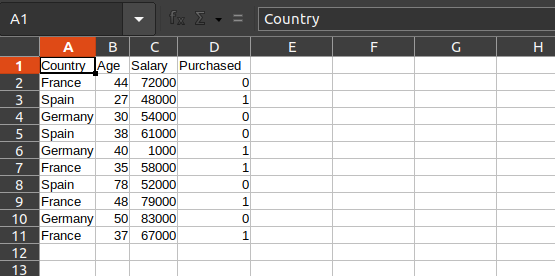
## **Explanation of Algorithm**

### What is the method used to perform the process ?

* sklearn.preprocessing.Binarizer() is a method which belongs to the processing module.
* It is used in the decentralization of continuous values.

|  |
| --- |
| sklearn.preprocessing.Binarizer(threshold, copy) |

* Parameters used
  + threshold : [float, optional] Values less than or equal to threshold is mapped to 0, else to 1. By default the threshold value is 0.0.
  + copy : [boolean, optional] If set to False, it avoids a copy. By default it is True.
* Method returns
  + Binarized Feature values



### What is the algorithm being used ?

1. Import all the required libraries
   1. Numpy -
   2. Matplotlib -
   3. Pandas -
   4. Sklearn -
2. Extracting values
   1. The dataset is read
   2. The rows and columns values are read and stored in respectives variables
3. Binarize the extracted data
   1. We use the Binarizer method from the Sklearn library to perform this method.
   2. We reshape the array of values obtained from the data set into a one dimensional array to be used during the binarization
   3. We create a binarizer with the required thresholds and store the values
4. Display the binarized features

## **Source code**

# Program to Binarize values from the given dataset.

# importing the required libraries

# numpy -> for reading the csv files and reshaping arrays

# matplotlib.pyplot -> for

# pandas -> for extracting rows using 'iloc'

# make sure you have numpy installed -> pip3 install numpy | pip3 install matplotlib | pip install pandas | pip install sklearn

import numpy as np

import matplotlib.pyplot as plt

import pandas as pd

# importing the Sklearn library

from sklearn import preprocessing

# importing the dataset from a .csv file

file\_path = './Sample\_Salaries\_Data\_Set.csv'

data\_set = pd.read\_csv(file\_path)

# .head() is used to select a number of rows. (n) means n rows will be selected.

data\_set.head()

print("\nThe data set being used is :\n\n", data\_set, "\n\n")

# extracting the values of different columns using slicing

# .iloc or .loc is used to extract the data from rows. () will extract from all rows. (n) will extract from the 'n' row.

# .iloc[:, n] selects the column until the nth column but excluding the nth column.

age = data\_set.iloc[:, 1].values

salary = data\_set.iloc[:, 2].values

# printing all the extracted values to check if it's right

print("\nThe extracted age values : \n", age)

print("\nThe extracted salary values : \n", salary)

# now binarizing the values

from sklearn.preprocessing import Binarizer

# reshape from numpy is used to convert a 2D array into a 1D array or vice versa

# Here reshape(1, -1) is used as '1' -> means a single feature is there in date. '-1' -> to reshape the 2D array into a 1D array

# reference : https://stackoverflow.com/questions/18691084/what-does-1-mean-in-numpy-reshape

x = age

x = x.reshape(1, -1)

y = salary

y = y.reshape(1, -1)

# https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.Binarizer.html

# Values greater than the threshold map to 1, while values less than or equal to the threshold map to 0. With the default threshold of 0, only positive values map to 1.

# Binarizer(35) means it will map all values greater than 35 to 1

binarizer\_1 = Binarizer(threshold = 35)

binarizer\_2 = Binarizer(threshold = 61000)

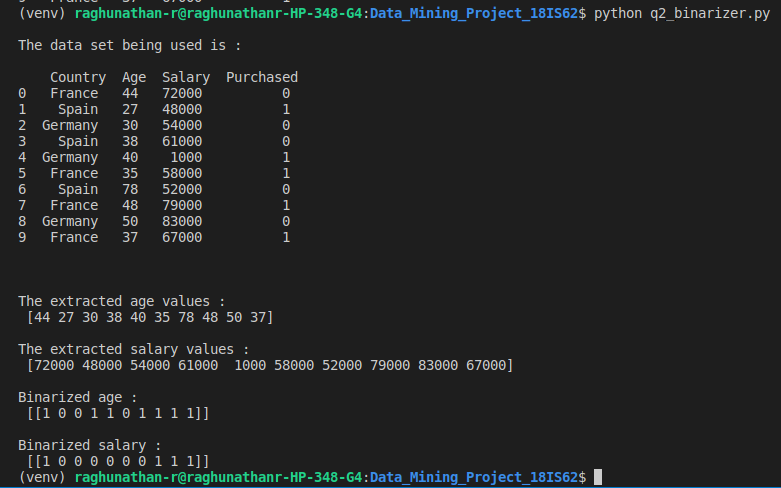
# transforming the features / columns

# fit\_transform(x) fits and transforms the date and returns the transformed version of x

print("\nBinarized age : \n", binarizer\_1.fit\_transform(x))

print("\nBinarized salary : \n", binarizer\_2.fit\_transform(y))

## **Screenshots of Output**



## 