Black Friday - (Assignment 3 - Question 5)

Subject Machine Learning / AI

DESCRIPTION

Black Friday falls on the Friday following the 'Thanksgiving Day' and is used as an occasion by many stores to offer highly promoted Sales.

You have the Black Friday dataset, which is an input data file **blackfriday.csv** present at the location **/data/training/blackfriday.csv**

This dataset contains information about purchases made in a retail store on Black Friday sale. Here's a brief description of the columns in the sample dataset:

- **USER_ID**: ID of the user
- **Gender**: F or M
- **Age**: Age group to which the customer belongs
- **Occupation**: ID of occupation of the customer
- **City_Category**: A or B or C
- **Stay_In_Current_City_Years**: 0 to 4+
- **Marital_Status**: 0: Unmarried, 1: Married
- **Purchase**: Purchase amount in dollars

This is a preview of the data under consideration:

User_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Purchase
1000001	F	0-17	10	A	2	0	8370
1000001	F	0-17	10	A	2	0	15200
1000001	F	0-17	10	A	2	0	1422
1000001	F	0-17	10	A	2	0	1057
1000003	M	26-35	15	A	3	0	15227
1000004	M	46-50	7	B	2	1	19215
1000004	M	46-50	7	B	2	1	15854
1000004	M	46-50	7	B	2	1	15686

The retailer wants to analyse this data and improve its future sales based on the analysis. In all the questions of this Assignment, we have to perform analysis on this data.

- Purchases made by customers on Black Friday sale are stored in the column named **Purchase**
- **Age** represents the age group the customer belongs to out of 0-17, 18-25, 26-35, 36-45, 46-50, 51-55 and 55+
- **Gender** represents the gender of the customer as F or M
- **City_Category** represents the category of city the customer belongs to as A, B or C

In this question, we have to perform probability calculations on the above data as explained below.

Questio

1. Calculate which **Gender** has a higher probability of making a purchase in the specified **Age** group.

******NOTE: You will find that many duplicate customer id are present here, please do not remove these. Here we will not filter the data for checking the unique user data.***

2. Calculate the probability that a customer belongs to the specified **City_Category**

******NOTE: Do not consider the customer category or gender while checking the probability.***

Input Format

- The first file to be read will be **blackfriday.csv**, which contains the data as mentioned above. This file is in **.csv** format and is present at the location **(/data/training/blackfriday.csv)**
- The second file to be read is **testcaseprobability.txt** which is present at **(/data/training/testcaseprobability.txt)**
- **testcaseprobability.txt** has the following lines:
- The first line contains the number of test cases T
- From the second line, every two lines are a pair wherein:
- the first line contains the value for **Age** for which the **Gender** with higher probability of making a purchase needs to be calculated
- the second line contains the value for **City_Category** for which the probability of any customer belonging to it needs to be calculated

******NOTE: Do not consider the customer category or gender while checking the probability for City_Category.***

Output Format

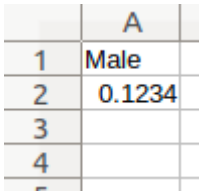
- For each test case **T**, create an output file, **output1.csv, output2.csv, ..., outputn.csv** where **n** represents the test case number
- **outputn.csv** should be present at the location (**/code/output/outputn.csv**) . This file should consist of the values for the **Gender (**Male or Female) & the **probability** (value rounded to **4** decimal places) for the value of Age & City_Category given in the corresponding test case
- **outputn.csv** should consist of only the values on two **separate rows**
- Do not write any headers or additional labels in the **outputn.csv** file

Sample Input

Read the input file **/data/training/blackfriday.csv**

Sample Output

Example: output1.csv will contain data looking like:



	A
1	Male
2	0.1234
3	
4	

DATASETS

- Training dataset

EXECUTION TIME LIMIT

Default.