

# DS Midterm I, November 7th, 2023

## Download the data

1. Consider the file [sciprogram-ds-2023-01-19-FIRSTNAME-LASTNAME-ID.zip](#) and extract it on your desktop.
2. Rename [sciprogram-ds-2023-01-19-FIRSTNAME-LASTNAME-ID](#) folder:

**put your name, lastname and id number**

like [sciprogram-ds-2023-01-19-luca-marchetti-432432](#)

From now on, you will be editing the files in that folder.

3. Edit the files following the instructions.
4. At the end of the exam, compress the folder in a zip file

[sciprogram-ds-2023-01-19-luca-marchetti-432432.zip](#)

and submit it. This is what will be evaluated. Please, include in the zip archive all the files required to execute your implementations!

## Exercise 1 [FIRST MODULE]

The Consumer Behavior and Shopping Habits Dataset provides a detailed overview of consumer preferences and purchasing behaviors. It includes demographic information, purchase history, product preferences, and preferred shopping channels (online or offline). This dataset is essential for businesses aiming to tailor their strategies to meet customer needs and enhance their shopping experience, ultimately driving sales and loyalty.

The dataset looks like the following:

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied	Promo Code Used
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express	Yes	Yes
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express	Yes	Yes
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shipping	Yes	Yes

- 1) Load the dataset
- 2) Create a function that **returns** the gender of the oldest individual whose location is specified as Montana.

```
def get_older(dataset):
```

```
    ...
```

```
    return res
```

Then print the output of the function. i.e.

```
>The gender of the oldest person in Montana is
```

```
>Male
```

- 3) What is the mean review rating customers provide based on their gender and clothing size?

Develop a function called "get\_avg\_rev" that calculates the average "Review Rating" for a given gender and size. The function should accept the input dataset, gender, and size as parameters. The default size should be set as "M".

```
def get_avg_rev(dataset,gender,size ....
```

```
    ...
```

```
    return res
```

Finally, test the function with:

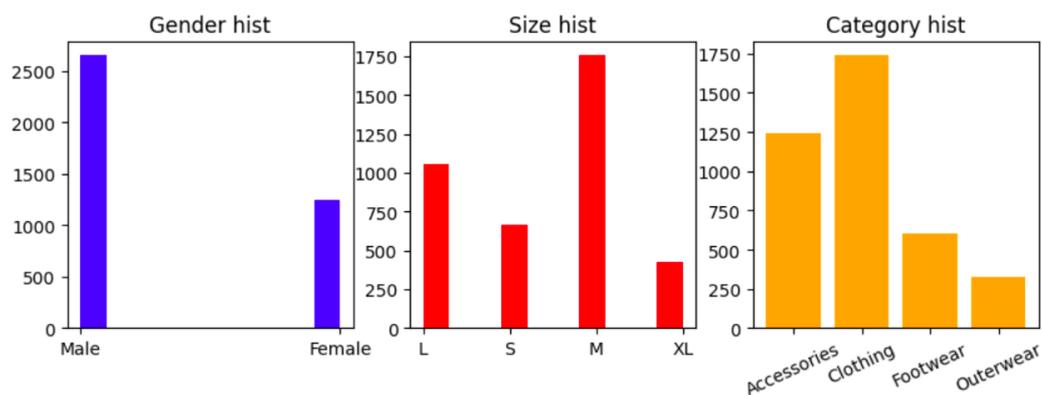
```
>get_avg_rev(df,"Male","S")
```

```
>get_avg_rev(df,"Male","M")
```

```
>get_avg_rev(df,"Male","L")
```

```
>get_avg_rev(df,"Male","XL")
```

- 4) Create a function named "my\_plot" that reproduces the following chart:



The initial plot displays a histogram representing gender, the second plot showcases the histogram for sizes, and the final plot illustrates the histogram for categories. It's important to specify titles for each plot, with colors set as blue, red, and orange correspondingly. Additionally, ensure that the labels on the Category histogram are rotated by 25 degrees.

Finally save the figure in pdf format using “plt.savefig("myplot.pdf", bbox\_inches = 'tight')”

- 5) Create a function called "build\_dictionary" that generates and provides a dictionary where the keys represent locations, and the corresponding values indicate the percentage of customers utilizing the discount. For example, if there were 74 customers in Oregon, with 38 of them using the discount and 36 not using it, the percentage of customers who used the discount would be calculated as  $38/74 = 0.513$  and stored in the dictionary under the "Oregon" key.

```
def build_dictionary(dataset):  
    ...  
    return dictionary
```

The resulting dictionary should encompass all the locations and be structured as follows:

```
{  
    'Kentucky': 0.43037974683544306,  
    'Maine': 0.35064935064935066,  
    'Massachusetts': 0.4861111111111111,  
    'Rhode Island': 0.3968253968253968,  
    'Oregon': 0.5135135135135135,  
    'Wyoming': 0.4225352112676056,  
    'Montana': 0.375,  
    ...  
}
```

- 6) Next, create a function that accepts the previous dictionary as input. For each location where the percentage of people using the discount falls below 0.35, print the most frequently purchased item. Name the function "most\_common\_item(dictionary)"

```
def most_common_item(dictionary):  
    ...  
    print ("The most common item in ",LOCATION," is ",ITEM)
```

output:

```
>The most common item in Kansas is Blouse  
>The most common item in Arizona is Backpack  
>The most common item in Connecticut is Coat
```