Homework 2

Due: Tuesday Sep 19, at 11:59pm via Blackboard

A car dealership wants to understand their customers and their buying habbits. The data (cardealership.csv) represents a randsome sample of their sales.

DESCRIPTION	VARIABLE
gender for customer	Gender
is the customer 'Married' or 'Single'?	marital status
age of the customer	age
country make of the car	country
the size of the car they bought ('Small', 'Medium', 'Large')	size
the type of the car they bought ('Family', 'Sporty', 'work')	type

```
In [1]: #import the necessary libraries
```

In [2]:

Out[2]:

	Gender	marital status	age	country	size	type
55	Male	Married	27	American	Small	Family
2	Male	Married	23	Japanese	Small	Family
107	Female	Married	24	American	Medium	Sporty
273	Male	Married	32	American	Large	Family
162	Male	Married	34	Japanese	Small	Work

```
In [16]: #What is the Shape
```

Out[16]: 6

1. Select all the married customers in the given dataset, and save it in a variable (married_customers). What is the percentage of married customers in the sample?

```
In [3]:
```

Out[3]: Married 64.686469 Single 35.313531

Name: marital status, dtype: float64

```
In [10]: #OR
```

Out[10]: 0.6468646864686468

2. Use a list comprehension to create a list with two age categories. The category is Below or equal to 30 if age <= 30, otherwise the category is Above 30. Use the result from this question to compute the number of customers in each category.

```
In [3]:
Out[3]: Relow 30 159
```

Out[3]: Below 30 159 Above 30 144 dtype: int64

- 3. The current version of Pandas has 142 methods including (DataFrame(), Series(), value_counts(), etc.). In this question, you are expected to learn about the cut() method which allows you to categorize a numerical vector into user-defined categories. Click here (https://pandas.pydata.org/docs/reference/api/pandas.cut.html">Click here (https://pandas.pydata.org/docs/reference/api/pandas.cut.html) to learn more about the cut method.
 - Use the cut() method to categorize the age variable into three buckets: (0,30], (30, 34], and (34,60]. (For this exercise, you don't have to add the new column to the original dataframe. You can save it in a seperate variable instead)
 - Rename the labels of the buckets to the ones shown in the table below.
 - How many element are there in each category?

label	bucket
Below 30	(0,30]
Between 30 and 34	(30, 34]
Above 34	(34,60]

In [6]:

Out[6]: Below 30 159
Above 34 76
Between 30 and 34 68
Name: age, dtype: int64

4. Pandas has another method called qcut, which allows you to categorize a numerical variable into equal-sized buckets based on quantiles. Use the qcut() method to categorize age into quartiles (4 buckets). Click here (https://pandas.pydata.org/docs/reference/api/pandas.qcut.html) to learn more about the cut method

```
In [48]:
Out[48]:
          (17.999, 26.0]
                             85
          (34.5, 60.0]
                             76
          (26.0, 30.0]
                             74
          (30.0, 34.5]
                             68
          Name: age, dtype: int64
            5. Using pandas, summarize the customer characteristics: Gender, marital status
              (using relative frequency tables) and age (using the describe() method).
In [11]:
Out[11]: Married
                      64.686469
          Single
                      35.313531
          Name: marital status, dtype: float64
In [12]:
Out[12]: Male
                     54.455446
          Female
                     45.544554
          Name: Gender, dtype: float64
In [13]:
Out[13]:
          count
                    303.000000
                     30.719472
          mean
          std
                      5.984294
          min
                     18.000000
          25%
                     26.000000
          50%
                     30.000000
          75%
                     34.500000
                     60.000000
          max
          Name: age, dtype: float64
            6. Using pandas, summarize the data on the cars sold: country, size, and type (using
              relative frequency tables).
 In [6]:
 Out[6]:
          Japanese
                       48.844884
                       37.953795
          American
          European
                       13.201320
          Name: country, dtype: float64
 In [7]: d
 Out[7]: Small
                     45.214521
          Medium
                     40.924092
          Large
                     13.861386
          Name: size, dtype: float64
```

In [8]:

Out[8]: Family 51.155116

Sporty 33.003300 Work 15.841584

Name: type, dtype: float64

7. Write a summary paragraph describing the customers and cars sold data. Round all numbers in this paragraph to nearest integers.

Customers

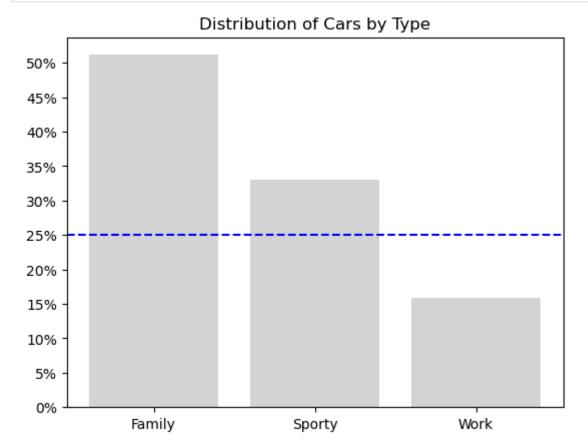
_

Cars sold

-

- 8. Create a bargraph that shows the distribution of car type . Your bargraph should be similar to the attached bargraph picture on blackboard ('CarsTypeDistribution.png'). In particular, make sure to:
- · Use default matplotlib plot style
- Use % for the labels of the y-axis ticks
- Use lightgrey for the bars color
- Overlay a horizontal line (y=25). The line's style is "dashed", and the color is "blue"

In [55]:



5 of 5