Homework 2

Due: Friday Sep 27, 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.

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

In [9]:

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

plt.style.use('default')
```

In [10]:

```python
```

Out[10]:

	Gender	marital status	age	country	size	type
78	Male	Married	44	Japanese	Small	Sporty
279	Male	Single	30	Japanese	Small	Sporty
35	Male	Single	24	Japanese	Medium	Sporty
136	Male	Married	33	American	Large	Family
126	Female	Married	28	American	Small	Family

In [11]:

```python
```

Out[11]:

```
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 [12]:

```python
```

Out[12]:

```
Married    64.686469
Single     35.313531
Name: marital status, dtype: float64
```

In [ ]:

```
```

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 [14]:

```
```

Out[14]:

```
Below 30    159
Above 30    144
dtype: int64
```

In [ ]:

```
```

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)](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?

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

In [16]:

```
```

Out[16]:

```
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)](https://pandas.pydata.org/docs/reference/api/pandas.qcut.html) to learn more about the `cut` method

In [17]:

```
```

Out[17]:

```
(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 [18]:

Out[18]:

```
Married    64.686469
Single     35.313531
Name: marital status, dtype: float64
```

In [19]:

Out[19]:

```
Male      54.455446
Female    45.544554
Name: Gender, dtype: float64
```

In [20]:

Out[20]:

```
count    303.000000
mean      30.719472
std        5.984294
min       18.000000
25%       26.000000
50%       30.000000
75%       34.500000
max       60.000000
Name: age, dtype: float64
```

6. Using `pandas`, summarize the data on the cars sold: `country`, `size`, and `type` (using relative frequency tables).

In [21]:

Out[21]:

```
Japanese    48.844884
American    37.953795
European    13.201320
Name: country, dtype: float64
```

In [22]:

Out[22]:

```
Small     45.214521
Medium    40.924092
Large     13.861386
Name: size, dtype: float64
```

In [23]:

Out[23]:

```
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
```

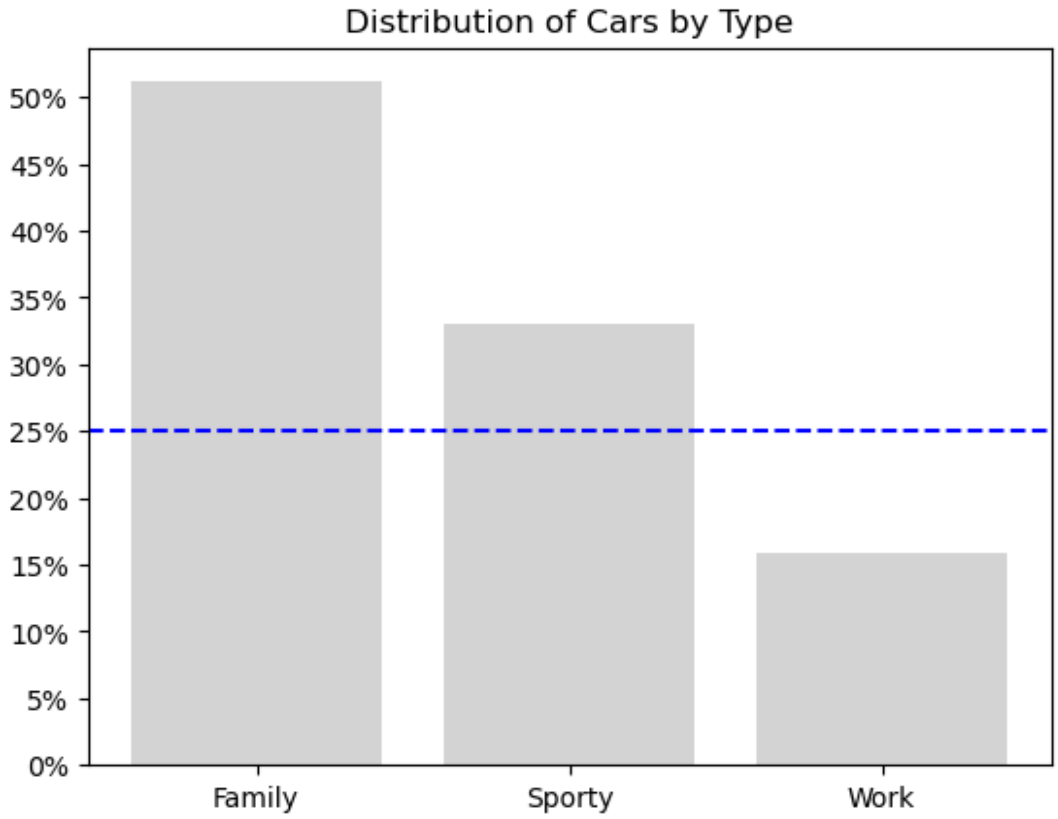
8. Create a bargraph that shows the distribution of car `type`. Your bargraph should be similar to the attached bargraph picture on blackboard ('CarsTypeDistribution.png'). In particular, make sure to:

- 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 [24]:

```

```



9. The dataset productioncost.xlsx, shows the various manufacturing costs of fertilizer production for a major producer in 4 of its plants. For this exercise, we are focusuing primarily on Plant (the name of the production Plant), Production Costs (which is overall production costs), Month (the month given from 1 to 12 of production).

aa. (4 points) Generate a Treemap for Total Production costs by Plants. Your graph should look as be shown below

In [1]:

```

```

In [25]:

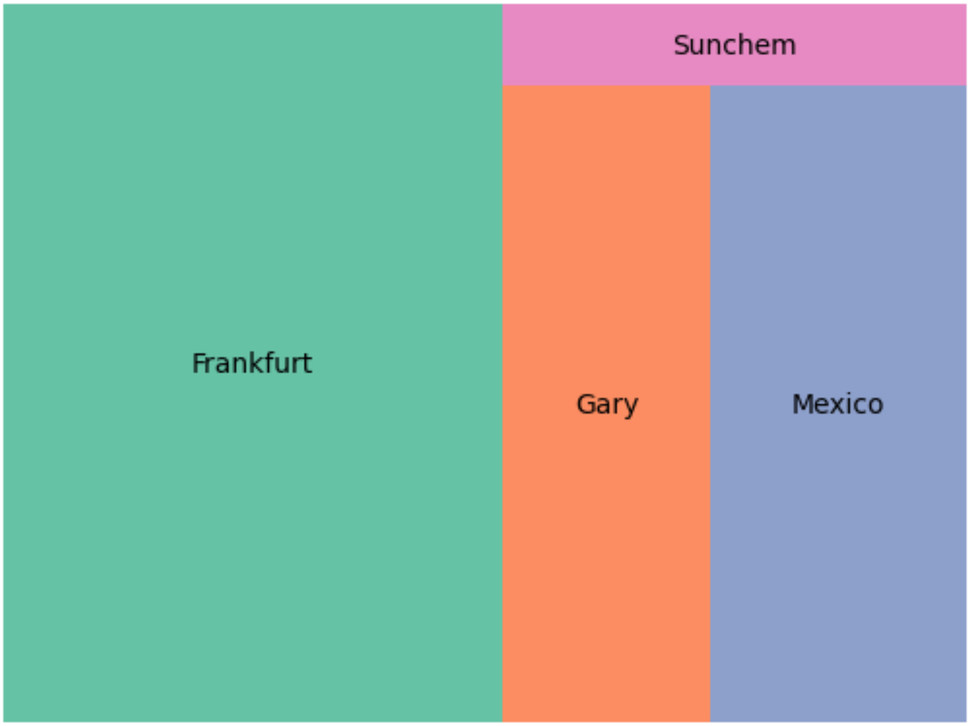
Out[25]:

Month Name	Apr	Aug	Dec	Feb	Jan	Jul	
Plant							
Frankfurt	104316.905567	104102.258968	104244.215355	104798.493893	104176.790839	103607.397194	104278.69
Gary	38247.200800	38211.712710	38321.345903	38176.851607	38516.631839	38129.793290	38439.61
Mexico	47189.629167	46952.890452	46930.365355	47535.256000	47224.978000	46989.544161	47190.66
Sunchem	11055.936400	10956.014581	10926.662806	10745.401393	11026.694581	10880.775968	10922.83

In [26]:

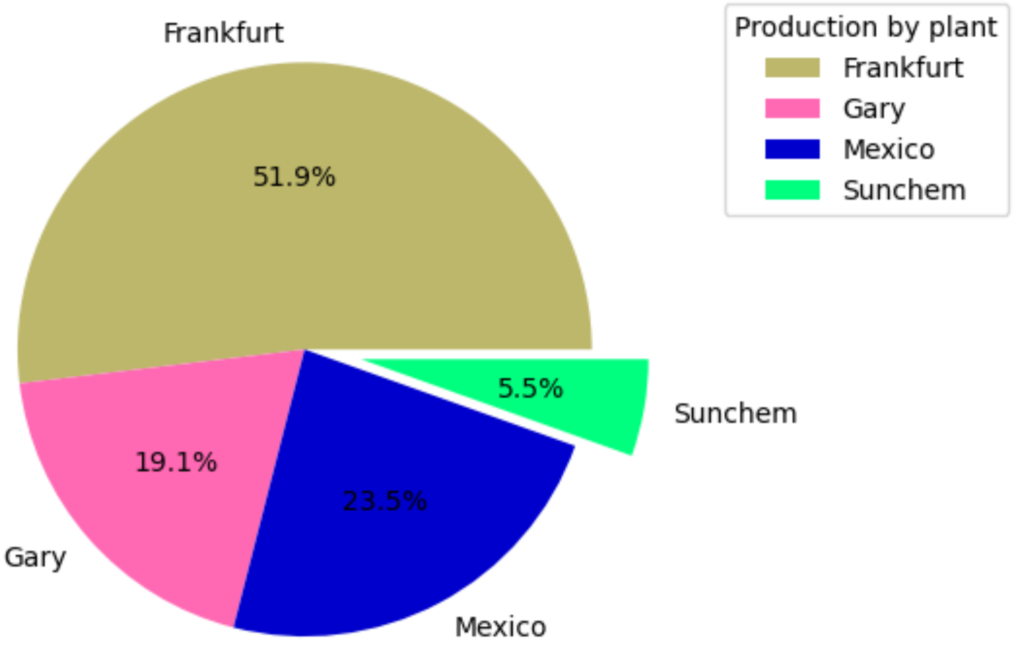
Out[26]:

```
(0.0, 300.0, 0.0, 150.0)
```



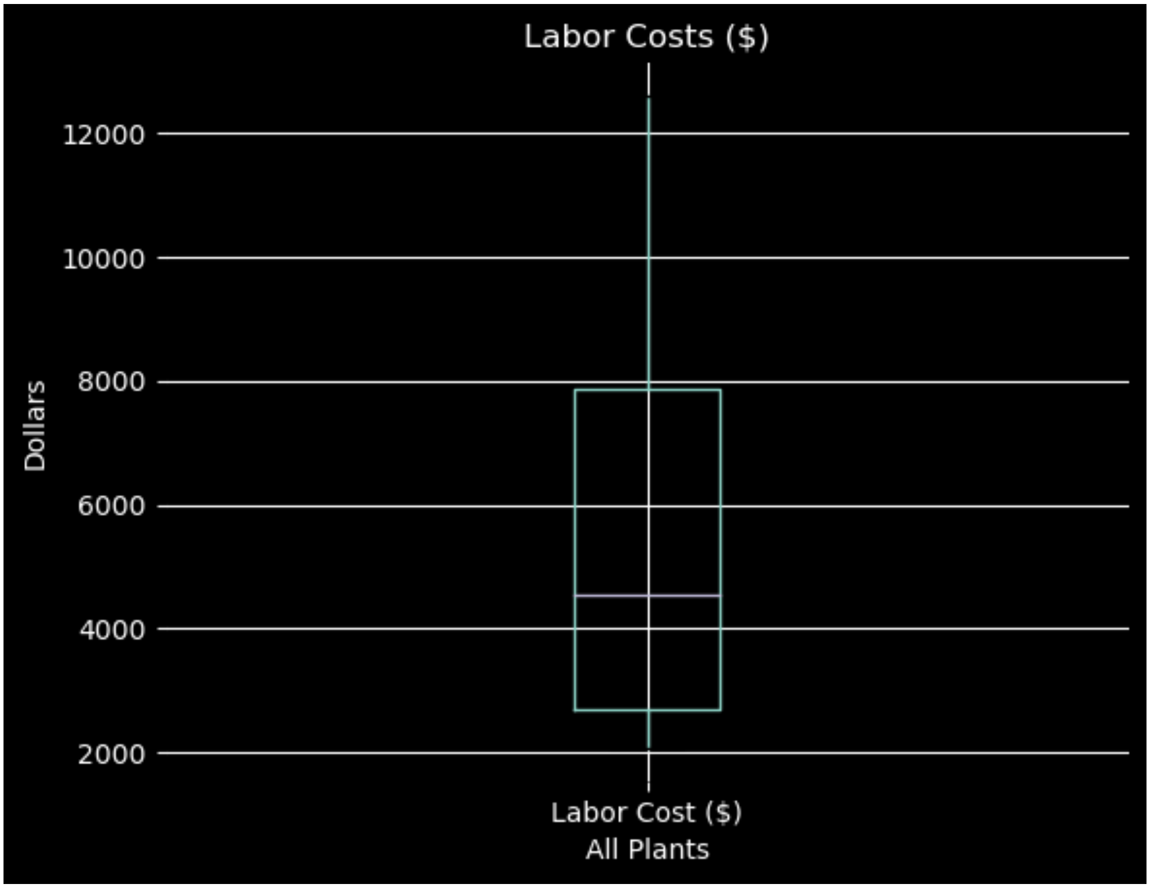
b. (4 points) Generate a pie chart to show Total Production Costs by Plant, 'exploding' out Sumchem's segment. Use 'darkkhaki','hotpink','mediumblue','springgreen'in your color palette, and show values to 1 decimal place. Your pie-chart should look as shown below:

```
In [27]:
```



c. (6 points) Generate a box-plot to show the overall Labor cost. Use the dark-background palette, and set the whiskers to the 5th and 95 percentile, and exclude outliers

```
In [28]:
```



Based on the boxplot, which of the followign are True

i. 50% of labor costs are approximately between 2.5K and 7.9K

ii. 75% of labor costs are higher than $2.5K

iii. 25% of labor costs are higher than $7.9K

iv. the distribution of production costs is skewed left

v. 50% of labor costs are below $4.3K

d. (4 points) Generate pie-charts to show the Total Production costs for each plant for months 1,4,7 and 10. Your chart titles should show the corresponding months of January, April,July and October, respectively, with values shown in percentages to 1 decimal place. Use 'hotpink', drakkhaki','blue and 'springgreen' for the colors. Your graphs should be look as shown below:

In [7]:

```

```

Out[7]:

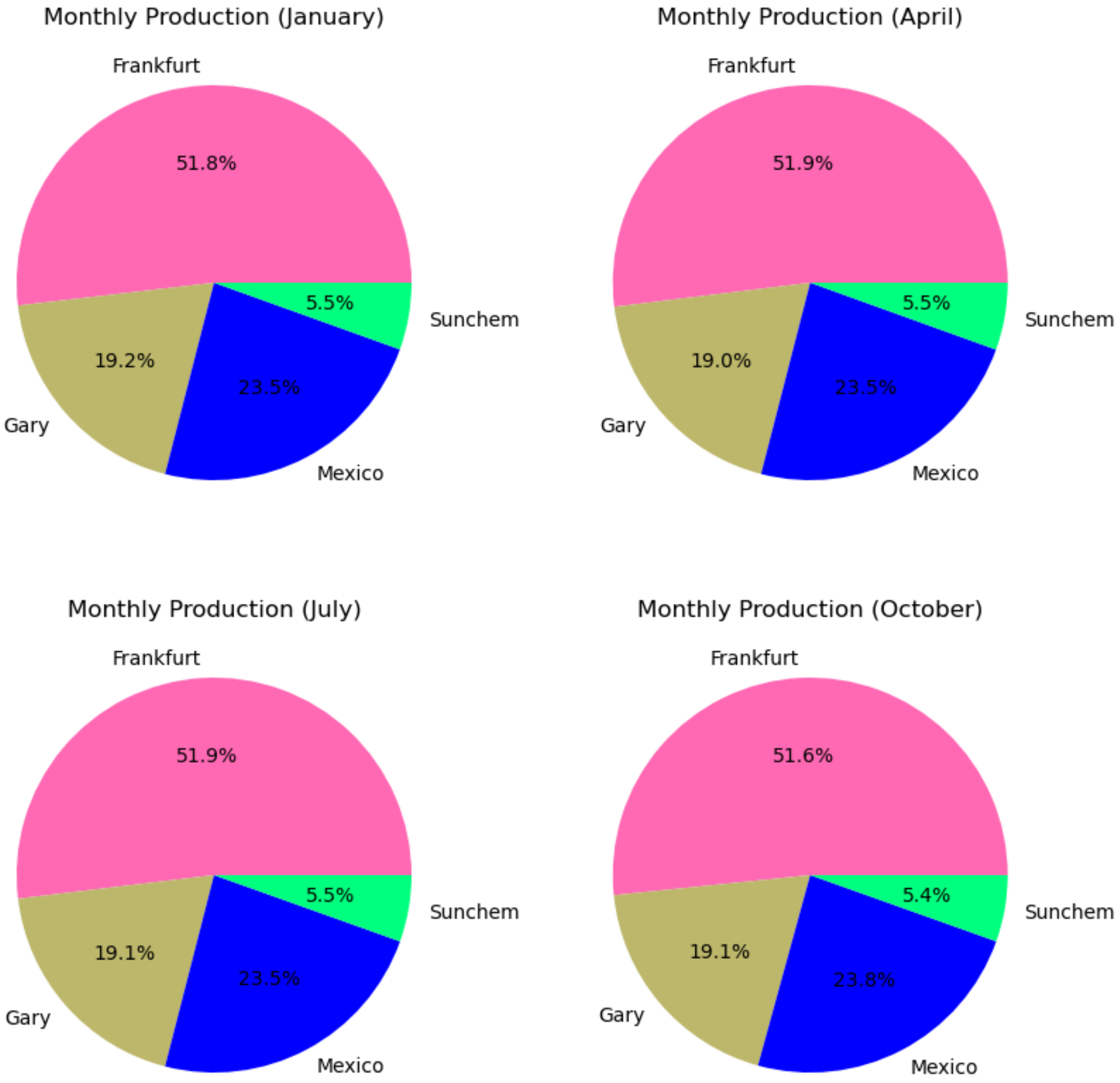
Month	1	2	3	4	5	6	7	
Plant								
Frankfurt	3229480.516	2934357.829	3221364.975	3129507.167	3223034.048	3128360.749	3211829.313	3227170.0
Gary	1194015.587	1068951.845	1187926.698	1147416.024	1199046.972	1153188.397	1182023.592	1184563.0
Mexico	1463974.318	1330987.168	1467613.665	1415688.875	1477013.397	1415719.919	1456675.869	1455539.6
Sunchem	341827.532	300871.239	342995.928	331678.092	341015.772	327685.145	337304.055	339636.4

In [8]:

```
```

Out[8]:

```
Text(0.5, 1.0, 'Monthly Production (October)')
```



Monthly Production (January)

Monthly Production (April)

Monthly Production (July)

Monthly Production (October)

In [ ]:

```
```