NPTEL PYTHON FOR DATA SCIENCE - ASSIGNMENT 3 - SOLUTION

1. **Which one of the following syntaxes is used to import a csv file with all the special characters as NaN?**

   Solution: b)

   pandas.read_csv(file_name.csv, na_values =[ ])

2. **What type of exception will be raised for the code given below?**

   Solution: c) TypeError

```
In [1]:
```

```
num_1 = 546
num_2 = '786'
print (num_1 + num_2)
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
<ipython-input-1-d765a3870512> in <module>
      1 num_1 = 546
      2 num_2 = '786'
----> 3 print (num_1 + num_2)

TypeError: unsupported operand type(s) for +: 'int' and 'str'
```

3. **What will be output of the code given below?**

   Solution: c) 60

```
In [2]:
```

```
x = 10
def func(num):
    x = 5
    for i in num:
        x *= i
    return x

print(func((-2, -1, 1, 2, 3)))
```

```
60
```

4. **By default, the crosstab() function computes a __**

Solution: d) Frequency table

Read the comma-separated values file hotel_bookings.csv as a dataframe 'data' and answer the questions from 5 to 7

5. **Which of the following command is used to replace the column, is_canceled values’ 0 to ‘No’ and 1 to ‘Yes’?**

In [3]:

```
import pandas as pd
data = pd.read_csv('hotel_bookings.csv')
```

In [4]:

```
data.head()
data['is_canceled'].replace([0,1],['No', 'Yes'], inplace = True)

data['is_canceled'].replace({0:'No', 1:'Yes'}, inplace = True)
data.head()
```

Out[4]:

	hotel	is_canceled	arrival_date_year	arrival_date_month	arrival_date_day_of_month	stays_i
0	Resort Hotel	No	2015	July	1	
1	Resort Hotel	No	2015	July	1	
2	Resort Hotel	No	2015	July	1	
3	Resort Hotel	No	2015	July	1	
4	Resort Hotel	No	2015	July	1	

6. **From the bar plot given below find the day with maximum number of reservations.**

   Solution: c) 12

In [5]:

```
index = data['arrival_date_day_of_month'].value_counts().index.tolist()
day_list = data['arrival_date_day_of_month'].value_counts().tolist()
from matplotlib import pyplot as plt
plt.figure(figsize=(10,6))
plt.bar(index,day_list)
plt.xlabel('Days')
plt.ylabel('Number of reservation')
plt.show()
```

```
<Figure size 1000x600 with 1 Axes>
```

7. **Identify the correct statements**.

   I. Scatter plot is used to convey the relationship between two numerical variables

   II. Histogram is used to depict the shape and spread of a continuous variable

   III. Bar plot is used to depict the visual representation of statistical five-number summary of a variable

   Solution: a) I and II only

   Box plot is used to depict the visual representation of statistical five-number summary of a variable

8. **Which of the following parameters is an alias for ‘sep’ for the read_csv and read_table functions from Pandas? ****

   ****Solution: d)** delimiter

9. **While importing data using Pandas dataframes, by default the empty cells will be interpreted as: -**

   Solution: c) nan/NaN

Read the 'flavors_of_cocoa.csv' as a dataframe, 'data_csv' and answer Q10 & Q11

In [6]:

```
data_csv=pd.read_csv("flavors_of_cocoa.csv",delimiter=",")
```

10. **Which of the following commands will return the number of unique values in the column ‘Company Location’?**

    Solution: d)

In [7]:

```
import pandas as pd
import numpy as np
```

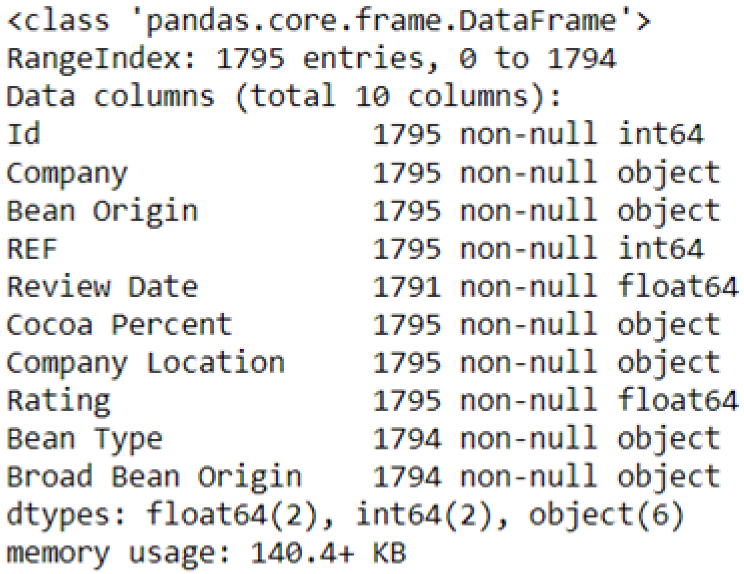
In [8]:

```
len(np.unique(data_csv['Company Location']))
```

Out[8]:

```
60
```

11. **According to the given data description for flavors_of_cocoa.csv, the column ‘Review Date’ denotes the year in which the chocolates were rated. Why is the column ‘Review Date’ read as float64?**



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1795 entries, 0 to 1794
Data columns (total 10 columns):
Id                   1795 non-null int64
Company              1795 non-null object
Bean Origin          1795 non-null object
REF                  1795 non-null int64
Review Date          1791 non-null float64
Cocoa Percent        1795 non-null object
Company Location     1795 non-null object
Rating               1795 non-null float64
Bean Type            1794 non-null object
Broad Bean Origin    1794 non-null object
dtypes: float64(2), int64(2), object(6)
memory usage: 140.4+ KB
```

Solution: b Because of missing values

In [9]:

```
data_csv.info()
data_csv.isnull().sum()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1795 entries, 0 to 1794
Data columns (total 7 columns):
Id                  1795 non-null int64
Company             1795 non-null object
Bean Origin         1795 non-null object
Review Date         1791 non-null float64
Cocoa Percent       1795 non-null object
Company Location    1795 non-null object
Rating              1795 non-null float64
dtypes: float64(2), int64(1), object(4)
memory usage: 98.3+ KB
```

Out[9]:

```
Id                  0
Company             0
Bean Origin         0
Review Date         4
Cocoa Percent       0
Company Location    0
Rating              0
dtype: int64
```

12. **What is the statistical measure related to the box plot?**

    Solution: c) Median

13. **The iris flower dataset containing 4 attributes Sepal length, Sepal width, Petal length, Petal width and a categorical feature 'Species' is loaded using**

In [10]:

```
import seaborn as sns
iris= sns.load_dataset("iris")
```

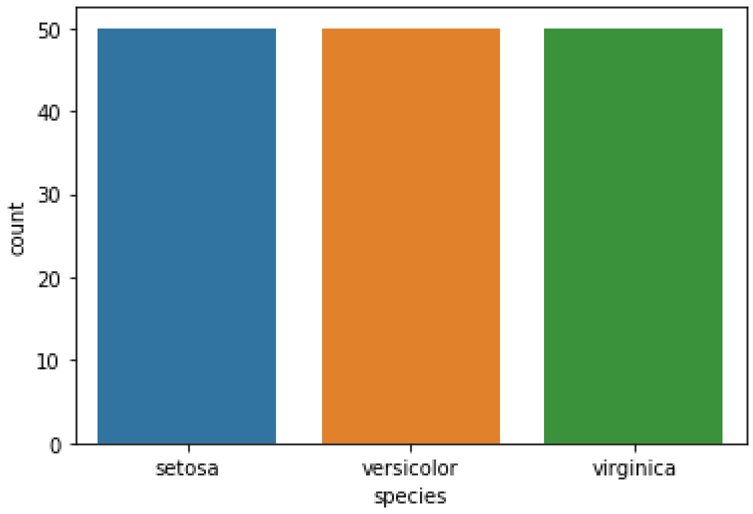
Which of the following code is used to plot the frequency distribution of the 'Species'?

In [11]:

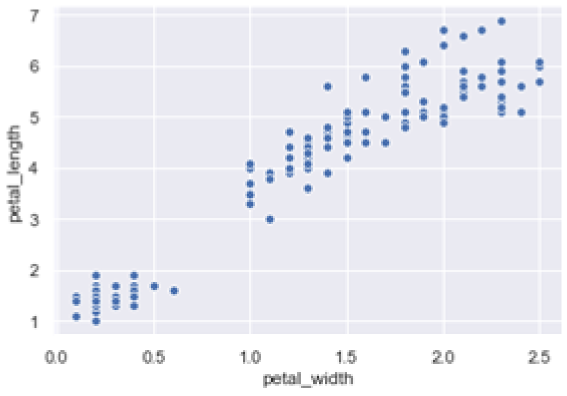
```
sns.countplot(x=iris['species'], data = iris)
```

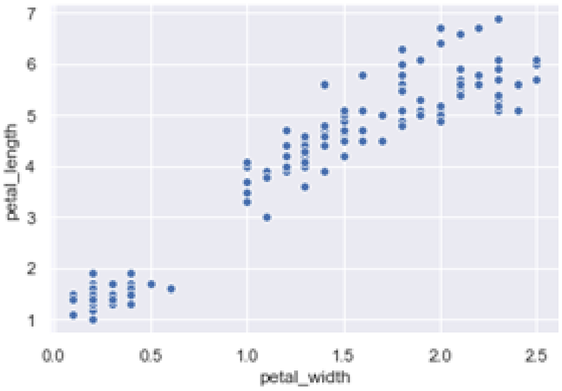
Out[11]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1bbe6ead7c8>
```



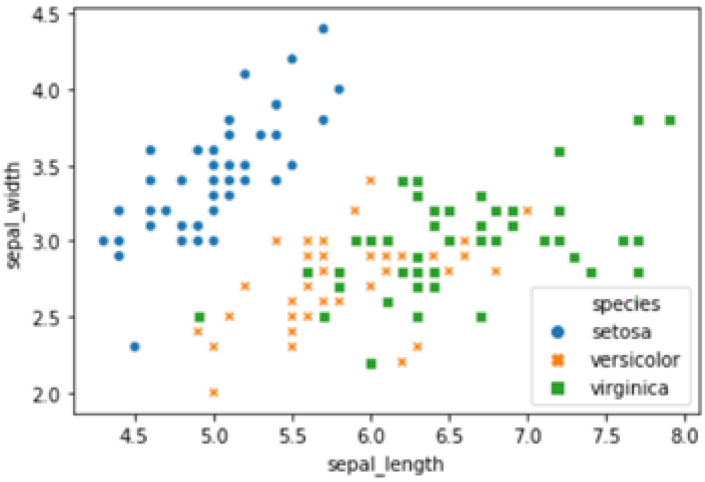
14. **What inferences can be made from the scatter plots shown below?**





Solution: d) Petal length & width are more linearly correlated than sepal length & width

15. **Fill in the blanks corresponding to the seaborn plot shown below:**



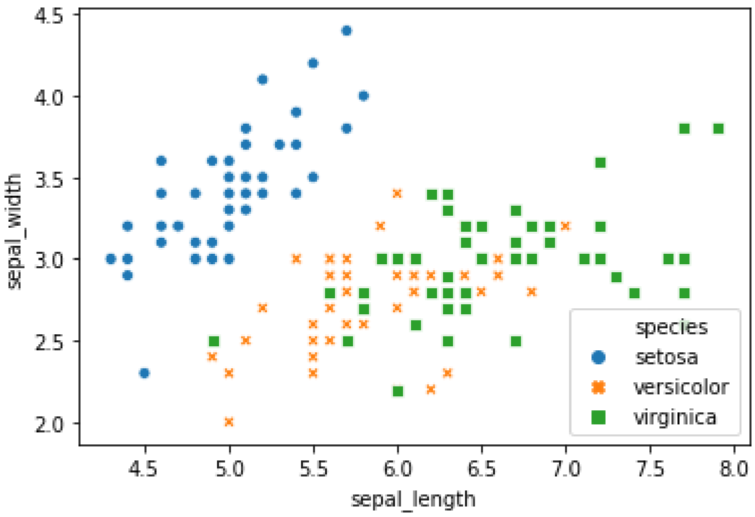
Solution: c)

In [12]:

```
sns.scatterplot(x=iris.sepal_length, y=iris.sepal_width,hue=iris.species, style=iris.specie
```

Out[12]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1bbe6f49588>
```


sepal_width
sepal_length
species
● setosa
✖ versicolor
■ virginica