Interactive exercise week #7c –Data wrangling2

Liping Wu 300-958-061

In this exercise we will do the following:

- Generate random numbers
- Merge datasets

Pre-requisites:

1- Install Anoconda

2- We will be using a lot of Public datasets these datasets are available at https://goo.gl/zjS4C6 under a folder named "Datasets for Predictive Modelling with Python", the datasets are organized in the order of the text book chapters: Python: Advanced Predictive Analytics, chapter # 3 files are required

1- Open your spider IDE

2- Import numpy

a. Generate a random number between 1 and 100

b. Generate a random number between 0 and 1

c. Define a function named "randint_range _firstname" to generate several random numbers in a range

d. Generate three random numbers between 0 and 100, which are all multiples of 5.

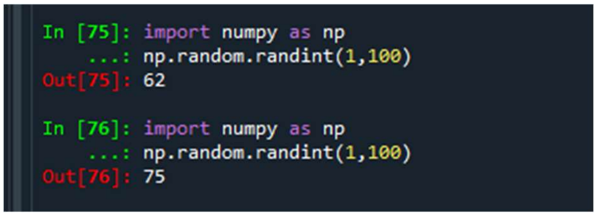
e. Select three numbers randomly from a list of numbers.

f. Generate a set of random numbers that retain their value, i.e. use the seed option

g. Shuffle a list of five numbers

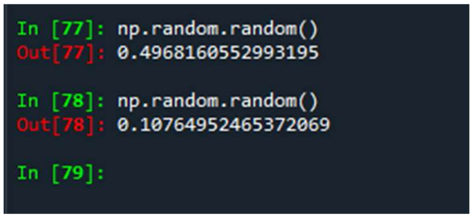
Following is the code, *make sure you update the function name correctly*:

```
#Generate one number between 1 and 100
import numpy as np
np.random.randint(1,100)
```



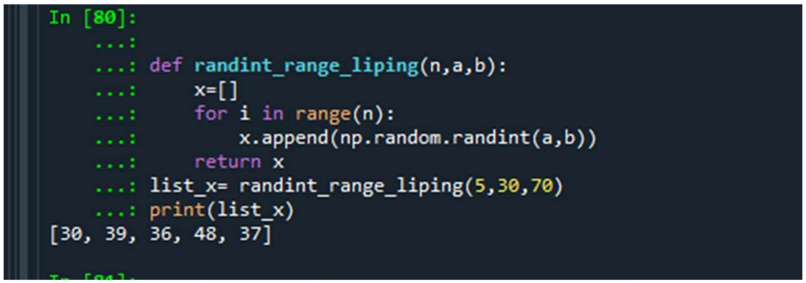
```
In [75]: import numpy as np
    ...: np.random.randint(1,100)
Out[75]: 62

In [76]: import numpy as np
    ...: np.random.randint(1,100)
Out[76]: 75
```

```
#Generate a random number between 0 and 1
import numpy as np
np.random.random()
```

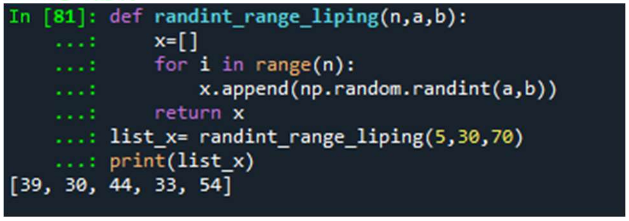


```
In [77]: np.random.random()
Out[77]: 0.4968160552993195

In [78]: np.random.random()
Out[78]: 0.10764952465372069

In [79]:
```

```
#Define a function to generate several random numbers in a range
def randint_range_liping(n,a,b):
   x=[]
   for i in range(n):
      x.append(np.random.randint(a,b))
   return x
list_x= randint_range_liping(5,30,70)
print(list_x)
```

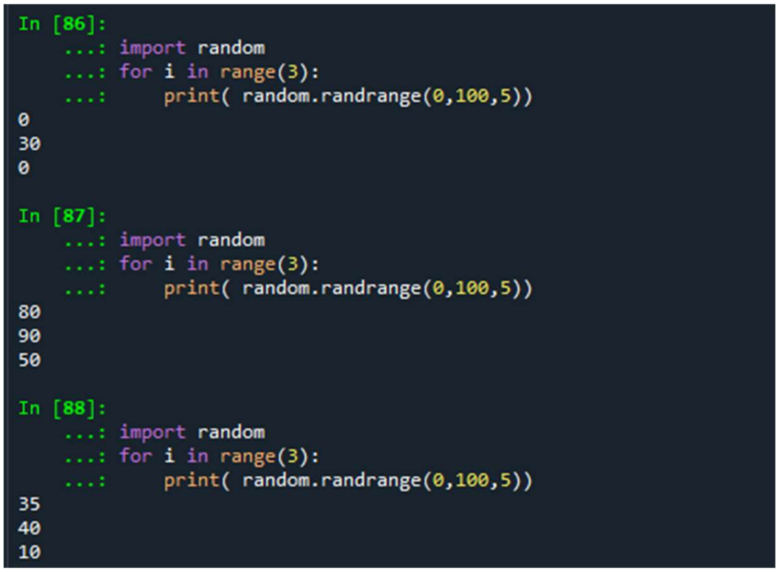


```
In [80]:
    ...:
    ...: def randint_range_liping(n,a,b):
    ...:     x=[]
    ...:     for i in range(n):
    ...:         x.append(np.random.randint(a,b))
    ...:     return x
    ...: list_x= randint_range_liping(5,30,70)
    ...: print(list_x)
[30, 39, 36, 48, 37]
```



```
In [81]: def randint_range_liping(n,a,b):
    ...:     x=[]
    ...:     for i in range(n):
    ...:         x.append(np.random.randint(a,b))
    ...:     return x
    ...: list_x= randint_range_liping(5,30,70)
    ...: print(list_x)
[39, 30, 44, 33, 54]
```

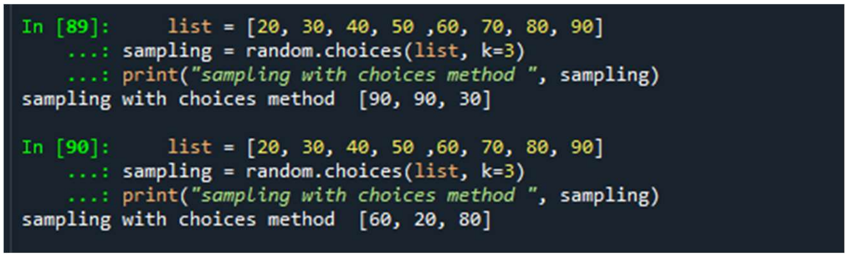
```
#d.  Generate three random numbers between 0 and 100, which are all multiples of 5
import random
for i in range(3):
   print( random.randrange(0,100,5))
```



```
In [86]:
    ...: import random
    ...: for i in range(3):
    ...:     print( random.randrange(0,100,5))
0
30
0

In [87]:
    ...: import random
    ...: for i in range(3):
    ...:     print( random.randrange(0,100,5))
80
90
50

In [88]:
    ...: import random
    ...: for i in range(3):
    ...:     print( random.randrange(0,100,5))
35
40
10
```

# Select three numbers randomly from a list of numbers

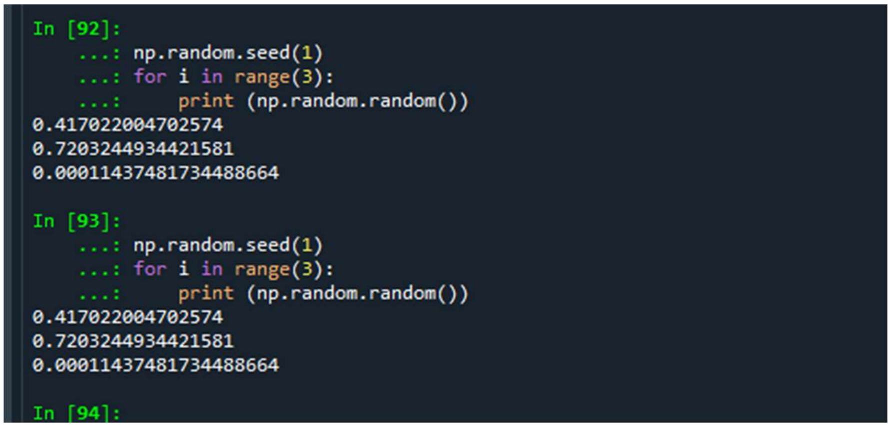
```
   list = [20, 30, 40, 50 ,60, 70, 80, 90]
sampling = random.choices(list, k=3)
print("sampling with choices method ", sampling)
```



```
In [89]:     list = [20, 30, 40, 50 ,60, 70, 80, 90]
    ...: sampling = random.choices(list, k=3)
    ...: print("sampling with choices method ", sampling)
sampling with choices method  [90, 90, 30]

In [90]:     list = [20, 30, 40, 50 ,60, 70, 80, 90]
    ...: sampling = random.choices(list, k=3)
    ...: print("sampling with choices method ", sampling)
sampling with choices method  [60, 20, 80]
```

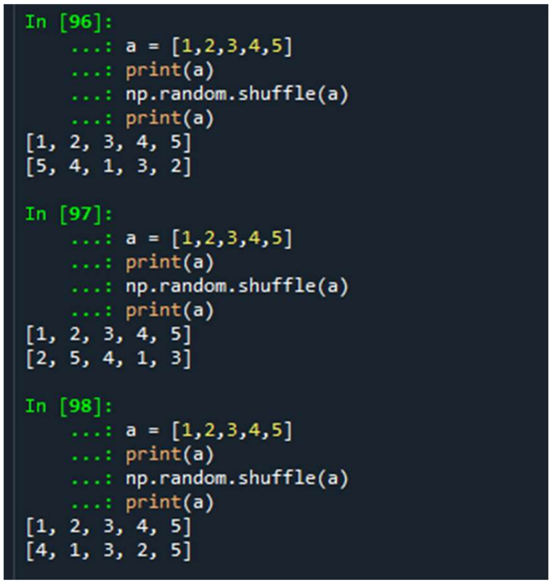
#Generate a set of random numbers that retain their value, i.e. use the seed option

```
np.random.seed(1)
for i in range(3):
   print (np.random.random())
```



```
In [92]:
    ...: np.random.seed(1)
    ...: for i in range(3):
    ...:     print (np.random.random())
0.417022004702574
0.7203244934421581
0.00011437481734488664

In [93]:
    ...: np.random.seed(1)
    ...: for i in range(3):
    ...:     print (np.random.random())
0.417022004702574
0.7203244934421581
0.00011437481734488664

In [94]:
```

```
#Shuffle a list of 5 numbers
a = [1,2,3,4,5]
print(a)
np.random.shuffle(a)
print(a)
```



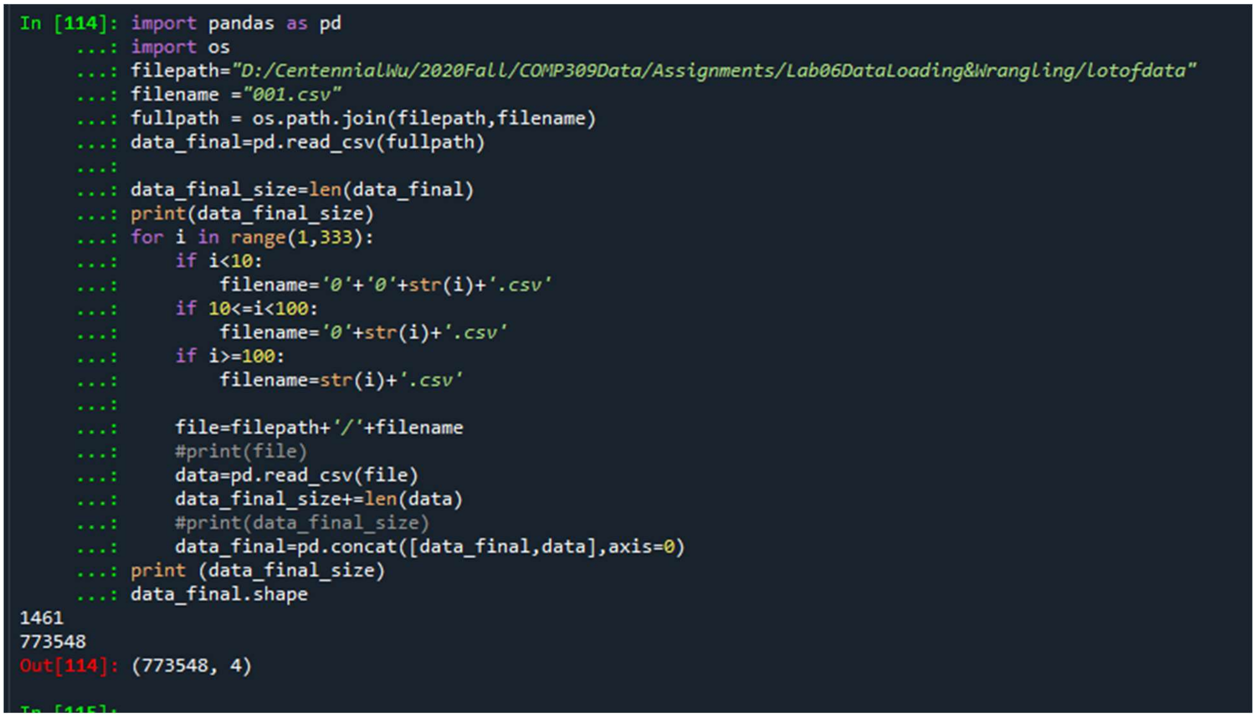
```
In [96]:
    ...: a = [1,2,3,4,5]
    ...: print(a)
    ...: np.random.shuffle(a)
    ...: print(a)
[1, 2, 3, 4, 5]
[5, 4, 1, 3, 2]

In [97]:
    ...: a = [1,2,3,4,5]
    ...: print(a)
    ...: np.random.shuffle(a)
    ...: print(a)
[1, 2, 3, 4, 5]
[2, 5, 4, 1, 3]

In [98]:
    ...: a = [1,2,3,4,5]
    ...: print(a)
    ...: np.random.shuffle(a)
    ...: print(a)
[1, 2, 3, 4, 5]
[4, 1, 3, 2, 5]
```

3- Dealing with several files containing daily collected data. You will need to:

a. Import the first file.
b. Loop through all the files.
c. Import them one by one.
d. Append them to the first file.
e. Repeat the loop.
f. Check the output

Following is the code, *make sure you update the path to the correct path where you placed the files:*

```
import pandas as pd
import os
filepath="D:/CentennialWu/2020Fall/COMP309Data/Assignments/Lab06DataLoading&Wrangling/lotofdata"
filename ="001.csv"
fullpath = os.path.join(filepath,filename)
data_final=pd.read_csv(fullpath)

data_final_size=len(data_final)
print(data_final_size)
for i in range(1,333):
    if i<10:
        filename='0'+'0'+str(i)+'.csv'
    if 10<=i<100:
        filename='0'+str(i)+'.csv'
    if i>=100:
        filename=str(i)+'.csv'

    file=filepath+'/'+filename
    #print(file)
    data=pd.read_csv(file)
    data_final_size+=len(data)
    #print(data_final_size)
    data_final=pd.concat([data_final,data],axis=0)
print (data_final_size)
data_final.shape
```



```
In [114]: import pandas as pd
     ...: import os
     ...: filepath="D:/CentennialWu/2020Fall/COMP309Data/Assignments/Lab06DataLoading&Wrangling/lotofdata"
     ...: filename ="001.csv"
     ...: fullpath = os.path.join(filepath,filename)
     ...: data_final=pd.read_csv(fullpath)
     ...:
     ...: data_final_size=len(data_final)
     ...: print(data_final_size)
     ...: for i in range(1,333):
     ...:     if i<10:
     ...:         filename='0'+'0'+str(i)+'.csv'
     ...:     if 10<=i<100:
     ...:         filename='0'+str(i)+'.csv'
     ...:     if i>=100:
     ...:         filename=str(i)+'.csv'
     ...:
     ...:     file=filepath+'/'+filename
     ...:     #print(file)
     ...:     data=pd.read_csv(file)
     ...:     data_final_size+=len(data)
     ...:     #print(data_final_size)
     ...:     data_final=pd.concat([data_final,data],axis=0)
     ...: print (data_final_size)
     ...: data_final.shape
1461
773548
Out[114]: (773548, 4)

In [115]:
```