Systematic Sampling Exercise

One commonly used sampling method is systematic sampling, which is implemented with a simple two-step process:

1. Place each member of a population in some order.

2. Choose a random starting point and select every nth member to be in the sample.

Let's learn how to perform systematic sampling on a pandas DataFrame in Python.

Suppose a teacher wants to obtain a sample of 100 students from a school with 500 total students. She chooses to use systematic sampling. She places each student in alphabetical order according to their last name, randomly chooses a starting point, and picks every 5th student in the sample.

The following code shows how to create a fake data frame to work within Python:

```
import pandas as pd
import numpy as np
import string
import random

#make this example reproducible
np.random.seed(0)

#create simple function to generate random last names
def randomNames(size=6, chars=string.ascii_uppercase):
    return ''.join(random.choice(chars) for _ in range(size))

#create DataFrame
df = pd.DataFrame({'last_name': [randomNames() for _ in range(500)],
                               'GPA': np.random.normal(loc=85, scale=3,
size=500)})

#view the first six rows of DataFrame
df.head()
```

```
last_name  GPA
0     PXGPIV     86.667888
1     JKRRQI     87.677422
2     TRIZTC     83.733056
3     YHUGIN     85.314142
4     ZVUNVK     85.684160
```

And the following code shows how to obtain a sample of 100 students through systematic sampling:

```
#obtain a systematic sample by selecting every 5th row
sys_sample_df = df.iloc[::5]

#view first six rows of DataFrame
sys_sample_df.head()
```

```
  last_name      gpa
3      ORJFW 88.78065
8      RWPSB 81.96988
13     RACZU 79.21433
18     ZOHKA 80.47246
23     QJETK 87.09991
28     JTHWB 83.87300
```

```
#view dimensions of the data frame
sys_sample_df.shape
```

```
(100, 2)
```

Notice that the first member included in the sample was in the first row of the original data frame. Each subsequent member in the sample is located 5 rows after the previous member.

And from using shape(), we can see that the systematic sample we obtained is a data frame with 100 rows and 2 columns.