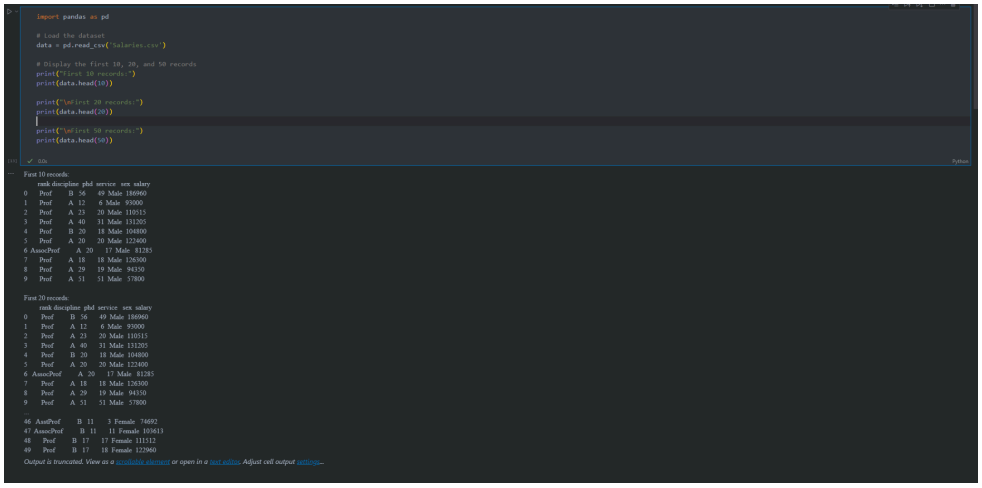
Assignment 3 - Data Analysis

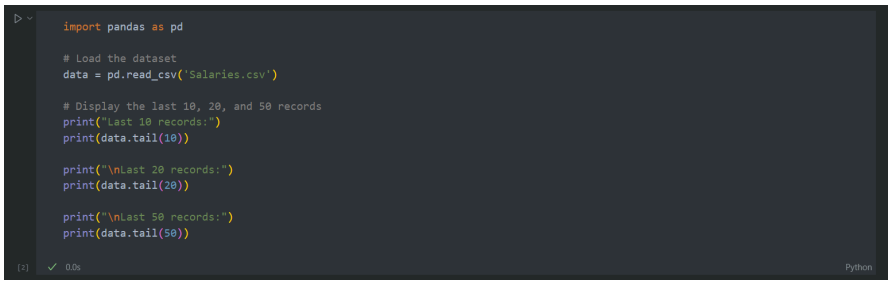
1) Try to read the first 10, 20, 50 records;
   Ans: Use head(n) to retrieve first n records



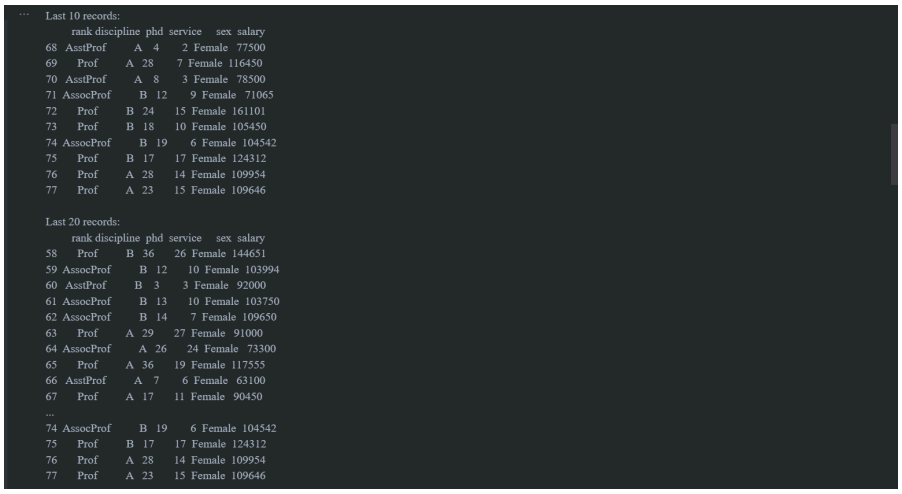
```
import pandas as pd

# Load the dataset
data = pd.read_csv('Salaries.csv')

# Display the first 10, 20, and 50 records
print("First 10 records:")
print(data.head(10))

print("\nFirst 20 records:")
print(data.head(20))

print("\nFirst 50 records:")
print(data.head(50))
```

```
First 10 records:
   rank discipline  phd  service   sex  salary
0  Prof          B   56       49  Male  186960
1  Prof          A   12        6  Male   93000
2  Prof          A   23       20  Male  110515
3  Prof          A   40       31  Male  131205
4  Prof          B   20       18  Male  104800
5  Prof          A   20       20  Male  122400
6  AssocProf     A   20       17  Male   81285
7  Prof          A   18       18  Male  126300
8  Prof          A   29       19  Male   94150
9  Prof          A   51       51  Male   57800

First 20 records:
   rank discipline  phd  service   sex  salary
0  Prof          B   56       49  Male  186960
1  Prof          A   12        6  Male   93000
2  Prof          A   23       20  Male  110515
3  Prof          A   40       31  Male  131205
4  Prof          B   20       18  Male  104800
5  Prof          A   20       20  Male  122400
6  AssocProf     A   20       17  Male   81285
7  Prof          A   18       18  Male  126300
8  Prof          A   29       19  Male   94150
9  Prof          A   51       51  Male   57800

46  AsstProf     B   11        3  Female  74692
47  AssocProf    B   11       11  Female  103613
48  Prof         B   17       17  Female  111512
49  Prof         B   17       18  Female  122960
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

2) Can you guess how to view the last few records;
   Ans: Instead of head, in this case we use tail(n) to retrieve last n records

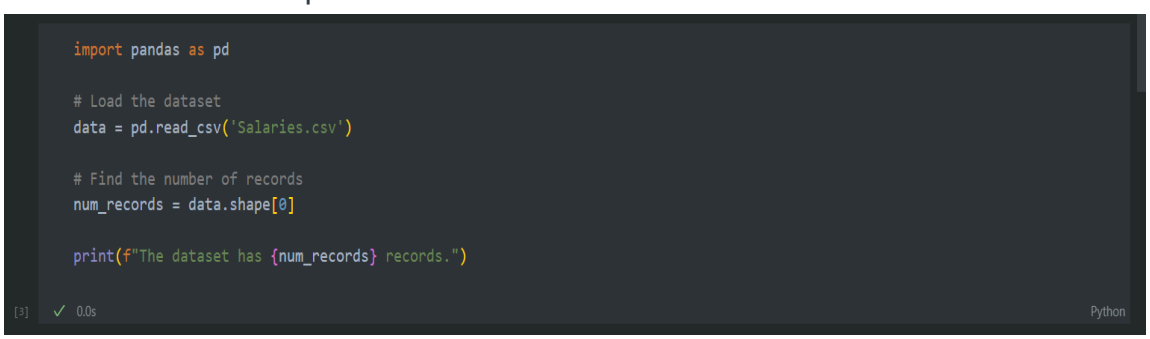


```
import pandas as pd

# Load the dataset
data = pd.read_csv('Salaries.csv')

# Display the last 10, 20, and 50 records
print("Last 10 records:")
print(data.tail(10))

print("\nLast 20 records:")
print(data.tail(20))

print("\nLast 50 records:")
print(data.tail(50))
```



```
Last 10 records:
    rank discipline  phd  service   sex  salary
68  AsstProf     A    4        2  Female   77500
69  Prof         A   28        7  Female  116450
70  AsstProf     A    8        3  Female   78500
71  AssocProf    B   12        9  Female   71065
72  Prof         B   24       15  Female  161101
73  Prof         B   18       10  Female  105450
74  AssocProf    B   19        6  Female  104542
75  Prof         B   17       17  Female  124312
76  Prof         A   28       14  Female  109954
77  Prof         A   23       15  Female  109646

Last 20 records:
    rank discipline  phd  service   sex  salary
58  Prof         B   36       26  Female  144651
59  AssocProf    B   12       10  Female  103994
60  AsstProf     B    3        3  Female   92000
61  AssocProf    B   13       10  Female  103750
62  AssocProf    B   14        7  Female  109650
63  Prof         A   29       27  Female   91000
64  AssocProf    A   26       24  Female   73300
65  Prof         A   36       19  Female  117555
66  AsstProf     A    7        6  Female   63100
67  Prof         A   17       11  Female   90450
...
74  AssocProf    B   19        6  Female  104542
75  Prof         B   17       17  Female  124312
76  Prof         A   28       14  Female  109954
77  Prof         A   23       15  Female  109646
```

3) Find how many records this data frame has;
   Ans: we can use .shape attribute to to find the total number of record in the dataframe



```
import pandas as pd

# Load the dataset
data = pd.read_csv('Salaries.csv')

# Find the number of records
num_records = data.shape[0]

print(f"The dataset has {num_records} records.")
```

```
The dataset has 78 records.
```

4) How many elements are there?
   Ans: we can use .size attribute to find the total number of elements in the dataframe

```python
import pandas as pd

# Load the dataset
data = pd.read_csv('Salaries.csv')

# Find the total number of elements
num_elements = data.size

print(f"The dataset has {num_elements} elements.")
```

```
[4] ✓ 0.0s Python
The dataset has 468 elements.
```

5) What are the column names?
   Ans: we can use .columns attribute to get the column names of the data frame

```python
import pandas as pd

# Load the dataset
data = pd.read_csv('Salaries.csv')

# Get the column names
column_names = data.columns

print("Column names:")
print(column_names)
```

```
[5] ✓ 0.0s Python
Column names:
Index(['rank', 'discipline', 'phd', 'service', 'sex', 'salary'], dtype='object')
```

6) What types of columns we have in this data frame?
   Ans. we can use dtypes attribute to find out the data types of each column in the data frame

```python
import pandas as pd

# Load the dataset
data = pd.read_csv('Salaries.csv')

# Get the data types of the columns
column_types = data.dtypes

print("Column types:")
print(column_types)
```

```
[6] ✓ 0.0s Python
Column types:
rank        object
discipline  object
phd          int64
service      int64
sex         object
salary       int64
dtype: object
```

7) Calculate the basic statistics for the salary column;
   Ans. .describe methods basically provides a summary statistics for the numerical columns and in this case ‘salary’.

```python
import pandas as pd

# Load the dataset
data = pd.read_csv('Salaries.csv')

# Calculate basic statistics for the 'salary' column
salary_stats = data['salary'].describe()

print("Basic statistics for the 'salary' column:")
print(salary_stats)
```

[7] ✓ 0.0s Python

```
Basic statistics for the 'salary' column:
count       78.000000
mean    108023.782051
std      28293.661022
min      57800.000000
25%      88612.500000
50%     104671.000000
75%     126774.750000
max     186960.000000
Name: salary, dtype: float64
```

8) Find how many values in the salary column (use count method);
Ans. inorder to find out how many values are there in the salary column first we'll exclude the non null columns and then count them out using the .count() method. [.count will count only non null values]

```python
import pandas as pd

# Load the dataset
data = pd.read_csv('Salaries.csv')

# Count the number of non-null values in the 'salary' column
salary_count = data['salary'].count()

print(f"There are {salary_count} non-null values in the 'salary' column.")
```

[8] ✓ 0.0s Python

```
There are 78 non-null values in the 'salary' column.
```

9) Calculate the average salary;
Ans: Calculating average is simple, the .mean() method will give the average salary value.

```python
import pandas as pd

# Load the dataset
data = pd.read_csv('Salaries.csv')

# Calculate the average salary
average_salary = data['salary'].mean()

print(f"The average salary is {average_salary:.2f}.")
```

[12] ✓ 0.0s Python

```
The average salary is 108023.78.
```