Pandas Data Frame

In [ ]:
```
'''
Q1: (1) Read "employee.csv" into a Pandas data frame (i.e., "employee_db").
    (2) Print a concise summary of "employee_db".
    (3) Print the first five rows of the data frame.
'''
```

In [33]:
```
import pandas as pd
import numpy as np

employees_db = pd.read_csv('employee.csv')
print(employees_db)
```

```
      UNIQUE_ID               POSITION_TITLE                      DEPARTMENT  \
0             0  ASSISTANT DIRECTOR (EX LVL)    Municipal Courts Department
1             1            LIBRARY ASSISTANT                         Library
2             2               POLICE OFFICER  Houston Police Department-HPD
3             3            ENGINEER/OPERATOR  Houston Fire Department (HFD)
4             4                  ELECTRICIAN    General Services Department
...         ...                          ...                             ...
1995       1995               POLICE OFFICER  Houston Police Department-HPD
1996       1996       COMMUNICATIONS CAPTAIN  Houston Fire Department (HFD)
1997       1997               POLICE OFFICER  Houston Police Department-HPD
1998       1998               POLICE OFFICER  Houston Police Department-HPD
1999       1999                 FIRE FIGHTER  Houston Fire Department (HFD)

      BASE_SALARY                        RACE EMPLOYMENT_TYPE  GENDER  \
0        121862.0            Hispanic/Latino       Full Time  Female
1         26125.0            Hispanic/Latino       Full Time  Female
2         45279.0                      White       Full Time    Male
3         63166.0                      White       Full Time    Male
4         56347.0                      White       Full Time    Male
...           ...                        ...             ...     ...
1995      43443.0                      White       Full Time    Male
1996      66523.0  Black or African American       Full Time    Male
1997      43443.0                      White       Full Time    Male
1998      55461.0     Asian/Pacific Islander       Full Time    Male
1999      51194.0            Hispanic/Latino       Full Time    Male

     EMPLOYMENT_STATUS   HIRE_DATE    JOB_DATE
0               Active  2006-06-12  2012-10-13
1               Active  2000-07-19  2010-09-18
2               Active  2015-02-03  2015-02-03
3               Active  1982-02-08  1991-05-25
4               Active  1989-06-19  1994-10-22
...                ...         ...         ...
1995            Active  2014-06-09  2015-06-09
1996            Active  2003-09-02  2013-10-06
1997            Active  2014-10-13  2015-10-13
1998            Active  2009-01-20  2011-07-02
1999            Active  2009-01-12  2010-07-12

[2000 rows x 10 columns]
```

In [34]:
```
type(employees_db)
```

Out[34]: pandas.core.frame.DataFrame

In [35]:
```
employees_db.shape
```

Out[35]: (2000, 10)

In [36]:
```
employees_db.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 10 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   UNIQUE_ID          2000 non-null   int64  
 1   POSITION_TITLE     2000 non-null   object 
 2   DEPARTMENT         2000 non-null   object 
 3   BASE_SALARY        1886 non-null   float64
 4   RACE               1965 non-null   object 
 5   EMPLOYMENT_TYPE    2000 non-null   object 
 6   GENDER             2000 non-null   object 
 7   EMPLOYMENT_STATUS  2000 non-null   object 
 8   HIRE_DATE          2000 non-null   object 
 9   JOB_DATE           1997 non-null   object 
dtypes: float64(1), int64(1), object(8)
memory usage: 156.4+ KB
```

In [37]:
```
employees_db.head()
```

Out[37]:

	UNIQUE_ID	POSITION_TITLE	DEPARTMENT	BASE_SALARY	RACE	EMPLOYMENT_TYPE
0	0	ASSISTANT DIRECTOR (EX LVL)	Municipal Courts Department	121862.0	Hispanic/Latino	Full Time
1	1	LIBRARY ASSISTANT	Library	26125.0	Hispanic/Latino	Full Time
2	2	POLICE OFFICER	Houston Police Department-HPD	45279.0	White	Full Time
3	3	ENGINEER/OPERATOR	Houston Fire Department (HFD)	63166.0	White	Full Time
4	4	ELECTRICIAN	General Services Department	56347.0	White	Full Time

In [38]:
```
employees_db.tail()
```

Out[38]:

	UNIQUE_ID	POSITION_TITLE	DEPARTMENT	BASE_SALARY	RACE	EMPLOYMENT_TYPE
1995	1995	POLICE OFFICER	Houston Police Department-HPD	43443.0	White	Full Time
1996	1996	COMMUNICATIONS CAPTAIN	Houston Fire Department (HFD)	66523.0	Black or African American	Full Time
1997	1997	POLICE OFFICER	Houston Police Department-HPD	43443.0	White	Full Time
1998	1998	POLICE OFFICER	Houston Police Department-HPD	55461.0	Asian/Pacific Islander	Full Time
1999	1999	FIRE FIGHTER	Houston Fire Department (HFD)	51194.0	Hispanic/Latino	Full Time

In [ ]:
```
'''
Q2: (1) Change the data type of "BASE_SALARY" into float32.
    (2) Print the data type of "BASE_SALARY" column.
'''
```

In [39]:
```
employees_db.dtypes
```

Out[39]:
```
UNIQUE_ID              int64
POSITION_TITLE        object
DEPARTMENT            object
BASE_SALARY          float64
RACE                  object
EMPLOYMENT_TYPE       object
GENDER                object
EMPLOYMENT_STATUS     object
HIRE_DATE             object
JOB_DATE              object
dtype: object
```

In [40]:
```
employees_db_BASE_SALARY = list(employees_db.select_dtypes(include = 'float64'))
```

In [41]:
```
employees_db[employees_db_BASE_SALARY] = employees_db[employees_db_BASE_SALARY].astype(
```

In [42]:
```
employees_db.dtypes
```

Out[42]:
```
UNIQUE_ID              int64
POSITION_TITLE        object
DEPARTMENT            object
BASE_SALARY          float32
```

```
RACE                 object
EMPLOYMENT_TYPE      object
GENDER               object
EMPLOYMENT_STATUS    object
HIRE_DATE            object
JOB_DATE             object
dtype: object
```

In [43]:
```
employees_db['BASE_SALARY'].dtypes
```

Out[43]:
```
dtype('float32')
```

In [ ]:
```
'''
Q3: (1) Add 'Clemson_University' and 'Clemson_Tigers' as new columns into the data frame
    (2) Rename the column 'Clemson_Tigers' to 'Clemson Tigers'
    (3) Print the column names of the data frame.
'''
```

In [44]:
```
employees_db.columns
```

Out[44]:
```
Index(['UNIQUE_ID', 'POSITION_TITLE', 'DEPARTMENT', 'BASE_SALARY', 'RACE',
       'EMPLOYMENT_TYPE', 'GENDER', 'EMPLOYMENT_STATUS', 'HIRE_DATE',
       'JOB_DATE'],
      dtype='object')
```

In [45]:
```
employees_db['CLEMSON_UNIVERSITY'] = True
employees_db['CLEMSON TIGERS'] = True
```

In [46]:
```
employees_db.columns
```

Out[46]:
```
Index(['UNIQUE_ID', 'POSITION_TITLE', 'DEPARTMENT', 'BASE_SALARY', 'RACE',
       'EMPLOYMENT_TYPE', 'GENDER', 'EMPLOYMENT_STATUS', 'HIRE_DATE',
       'JOB_DATE', 'CLEMSON_UNIVERSITY', 'CLEMSON TIGERS'],
      dtype='object')
```

In [47]:
```
employees_db.head()
```

Out[47]:

	UNIQUE_ID	POSITION_TITLE	DEPARTMENT	BASE_SALARY	RACE	EMPLOYMENT_TYPE
0	0	ASSISTANT DIRECTOR (EX LVL)	Municipal Courts Department	121862.0	Hispanic/Latino	Full Time
1	1	LIBRARY ASSISTANT	Library	26125.0	Hispanic/Latino	Full Time
2	2	POLICE OFFICER	Houston Police Department-HPD	45279.0	White	Full Time
3	3	ENGINEER/OPERATOR	Houston Fire Department (HFD)	63166.0	White	Full Time

	UNIQUE_ID	POSITION_TITLE	DEPARTMENT	BASE_SALARY	RACE	EMPLOYMENT_TYPE
4	4	ELECTRICIAN	General Services Department	56347.0	White	Full Time

In [48]:
```
columns_renamed = {'CLEMSON TIGERS': 'CLEMSON_TIGERS'}
employees_db = employees_db.rename(columns = columns_renamed)
employees_db.columns
```

Out[48]:
```
Index(['UNIQUE_ID', 'POSITION_TITLE', 'DEPARTMENT', 'BASE_SALARY', 'RACE',
       'EMPLOYMENT_TYPE', 'GENDER', 'EMPLOYMENT_STATUS', 'HIRE_DATE',
       'JOB_DATE', 'CLEMSON_UNIVERSITY', 'CLEMSON_TIGERS'],
      dtype='object')
```

In [ ]:
```
'''
Q4: (1) Add a new row to the data frame by following information.

        POSITION_TITLE: ASSISTANT PROFESSOR
        DEPARTMENT: MANAGEMENT
        BASE_SALARY: 100000
        RACE: ASIAN
        EMPLOYMENT_TYPE: Full Time
        GENDER: Male
        EMPLOYMENT_STATUS: Active
        HIRE_DATE: 2006-06-12
        JOB_DATE: 2016-06-12
        DSA_PROGRAM: True
        DSA_8640: True

    (2) Change the "GENDER" of the last row to 'Female'.
    (3) Print the last five rows of the data frame.
'''
```

In [49]:
```
employees_db = employees_db.append({'POSITION_TITLE': 'ASSISTANT PROFESSOR', 'DEPARTMEN
```

In [50]:
```
employees_db.tail()
```

Out[50]:

	UNIQUE_ID	POSITION_TITLE	DEPARTMENT	BASE_SALARY	RACE	EMPLOYMENT_TYPE
1996	1996.0	COMMUNICATIONS CAPTAIN	Houston Fire Department (HFD)	66523.0	Black or African American	Full Time
1997	1997.0	POLICE OFFICER	Houston Police Department-HPD	43443.0	White	Full Time
1998	1998.0	POLICE OFFICER	Houston Police Department-HPD	55461.0	Asian/Pacific Islander	Full Time

	UNIQUE_ID	POSITION_TITLE	DEPARTMENT	BASE_SALARY	RACE	EMPLOYMENT_TYPE
1999	1999.0	FIRE FIGHTER	Houston Fire Department (HFD)	51194.0	Hispanic/Latino	Full Time
2000	NaN	ASSISTANT PROFESSOR	MANAGEMENT	100000	ASIAN	Full Time

In [51]:
```
employees_db.at[2000, 'GENDER'] = 'Female'
```

In [52]:
```
employees_db.tail()
```

Out[52]:

	UNIQUE_ID	POSITION_TITLE	DEPARTMENT	BASE_SALARY	RACE	EMPLOYMENT_TYPE
1996	1996.0	COMMUNICATIONS CAPTAIN	Houston Fire Department (HFD)	66523.0	Black or African American	Full Time
1997	1997.0	POLICE OFFICER	Houston Police Department-HPD	43443.0	White	Full Time
1998	1998.0	POLICE OFFICER	Houston Police Department-HPD	55461.0	Asian/Pacific Islander	Full Time
1999	1999.0	FIRE FIGHTER	Houston Fire Department (HFD)	51194.0	Hispanic/Latino	Full Time
2000	NaN	ASSISTANT PROFESSOR	MANAGEMENT	100000	ASIAN	Full Time

In [ ]:
```
'''
Q5: (1) Drop the column DSA_8640.
    (2) Drop the first row of the data frame, "employee_db".
    (3) Reset the index of the data frame, "employee_db".
    (4) Print the first five rows of the data frame, "employee_db".
    (5) Create a new data frame, 'employee_date',
        which contains columns having 'DATE' keyword from "employee_db" data frame.
    (6) Print the last five rows of the data frame, "employee_date".
'''
```

In [53]:
```
employees_db = employees_db.drop('DSA_8640', axis = 'columns')
```

In [54]:
```
employees_db = employees_db.drop(0)
```

In [55]:
```
employees_db = employees_db.reset_index()
```

In [56]:
```
employees_db.head()
```

Out[56]:

	index	UNIQUE_ID	POSITION_TITLE	DEPARTMENT	BASE_SALARY	RACE	EMPLOYMENT_
0	1	1.0	LIBRARY ASSISTANT	Library	26125.0	Hispanic/Latino	Full
1	2	2.0	POLICE OFFICER	Houston Police Department-HPD	45279.0	White	Full
2	3	3.0	ENGINEER/OPERATOR	Houston Fire Department (HFD)	63166.0	White	Full
3	4	4.0	ELECTRICIAN	General Services Department	56347.0	White	Full
4	5	5.0	SENIOR POLICE OFFICER	Houston Police Department-HPD	66614.0	Black or African American	Full

In [57]:
```
employees_date = employees_db[['HIRE_DATE', 'JOB_DATE']]
```

In [58]:
```
employees_date.tail()
```

Out[58]:

	HIRE_DATE	JOB_DATE
1995	2003-09-02	2013-10-06
1996	2014-10-13	2015-10-13
1997	2009-01-20	2011-07-02
1998	2009-01-12	2010-07-12
1999	2006-06-12	2016-06-12

In [ ]:
```
'''
Q6: Create two data frames with following information:

(1) Create a data frame, "Fruit_1", by using a dictionary as an input of the data frame
(2) Create a data frame, "Fruit_2", by using a numpy array as an input of the data frame
(3) Print the data frames (i.e., Fruit_1 and Fruit_2).
'''
```

In [59]:
```
apples = {'apples':[3, 2, 1, 0]}
```

In [60]:
```
Fruit_1 = pd.DataFrame(apples)
```

In [61]:
```
oranges = np.array([[0], [3], [7], [2]])
```

In [62]:
```
Fruit_2 = pd.DataFrame(oranges, columns = ['oranges'])
```

In [63]:
```
Fruit_1
```

Out[63]:

	apples
0	3
1	2
2	1
3	0

In [64]:
```
Fruit_2
```

Out[64]:

	oranges
0	0
1	3
2	7
3	2

In [ ]:
```
'''
Q7: (1) Read "diamonds.csv" into a Pandas data frame (i.e., "diamonds_db").
    (2) Create another data frame, "new_diamonds_db",
        which includes the "carat", "cut", and "color" columns of the first one thousand
    (3) Sort the column names in alphabetical order.
    (4) Print the column names of the data frame.
    (5) Print a concise summary of "new_diamonds_db".
'''
```

In [65]:
```
diamonds_db = pd.read_csv('diamonds.csv')
```

In [66]:
```
new_diamonds_db = diamonds_db[['carat', 'cut', 'color']]
new_diamonds_db = new_diamonds_db.loc[0:999]
```

In [67]:
```
new_diamonds_db = new_diamonds_db.sort_index(axis = 'columns')
```

In [68]:
```
new_diamonds_db.columns
```

Out[68]:
```
Index(['carat', 'color', 'cut'], dtype='object')
```

In [69]:
```
type(new_diamonds_db)
```

Out[69]:
```
pandas.core.frame.DataFrame
```

In [70]:
```
new_diamonds_db.shape
```

Out[70]:
```
(1000, 3)
```

In [71]:
```
new_diamonds_db.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   carat   1000 non-null   float64
 1   color   1000 non-null   object
 2   cut     1000 non-null   object
dtypes: float64(1), object(2)
memory usage: 23.6+ KB
```

In [72]:
```
new_diamonds_db.head()
```

Out[72]:

	carat	color	cut
0	0.23	E	Ideal
1	0.21	E	Premium
2	0.23	E	Good
3	0.29	I	Premium
4	0.31	J	Good

In [73]:
```
new_diamonds_db.tail()
```

Out[73]:

	carat	color	cut
995	0.54	D	Ideal
996	0.72	E	Ideal
997	0.72	F	Good
998	0.74	D	Premium
999	1.12	J	Premium