Salaries_by_college_major

May 24, 2023

[ ]:
```
8+1
```

[2]:
```
import pandas as pd
```

[3]:
```
df = pd.read_csv('salaries_by_college_major.csv')
```

[ ]:
```
df.head()
```

[ ]:
```
     Undergraduate Major  Starting Median Salary  Mid-Career Median Salary  \
0             Accounting                 46000.0                   77100.0
1  Aerospace Engineering                 57700.0                  101000.0
2            Agriculture                 42600.0                   71900.0
3           Anthropology                 36800.0                   61500.0
4           Architecture                 41600.0                   76800.0

   Mid-Career 10th Percentile Salary  Mid-Career 90th Percentile Salary  \
0                            42200.0                           152000.0
1                            64300.0                           161000.0
2                            36300.0                           150000.0
3                            33800.0                           138000.0
4                            50600.0                           136000.0

      Group
0  Business
1      STEM
2  Business
3      HASS
4  Business
```

[ ]:
```
df.shape
```

[ ]:
```
(51, 6)
```

[ ]:
```
df.columns
```

[ ]:
```
Index(['Undergraduate Major', 'Starting Median Salary',
       'Mid-Career Median Salary', 'Mid-Career 10th Percentile Salary',
```

```
       'Mid-Career 90th Percentile Salary', 'Group'],
      dtype='object')
```

[ ]:
```
df.isna
```

[ ]:
```
df.tail
```

[ ]:
```
df.isna()
```

[ ]:

	Undergraduate Major	Starting Median Salary	Mid-Career Median Salary	\
0	False	False	False	
1	False	False	False	
2	False	False	False	
3	False	False	False	
4	False	False	False	
5	False	False	False	
6	False	False	False	
7	False	False	False	
8	False	False	False	
9	False	False	False	
10	False	False	False	
11	False	False	False	
12	False	False	False	
13	False	False	False	
14	False	False	False	
15	False	False	False	
16	False	False	False	
17	False	False	False	
18	False	False	False	
19	False	False	False	
20	False	False	False	
21	False	False	False	
22	False	False	False	
23	False	False	False	
24	False	False	False	
25	False	False	False	
26	False	False	False	
27	False	False	False	
28	False	False	False	
29	False	False	False	
30	False	False	False	
31	False	False	False	
32	False	False	False	
33	False	False	False	
34	False	False	False	
35	False	False	False	
36	False	False	False	

37	False	False	False
38	False	False	False
39	False	False	False
40	False	False	False
41	False	False	False
42	False	False	False
43	False	False	False
44	False	False	False
45	False	False	False
46	False	False	False
47	False	False	False
48	False	False	False
49	False	False	False
50	False	True	True

	Mid-Career 10th Percentile Salary	Mid-Career 90th Percentile Salary	\
0	False	False	
1	False	False	
2	False	False	
3	False	False	
4	False	False	
5	False	False	
6	False	False	
7	False	False	
8	False	False	
9	False	False	
10	False	False	
11	False	False	
12	False	False	
13	False	False	
14	False	False	
15	False	False	
16	False	False	
17	False	False	
18	False	False	
19	False	False	
20	False	False	
21	False	False	
22	False	False	
23	False	False	
24	False	False	
25	False	False	
26	False	False	
27	False	False	
28	False	False	
29	False	False	
30	False	False	

31	False	False
32	False	False
33	False	False
34	False	False
35	False	False
36	False	False
37	False	False
38	False	False
39	False	False
40	False	False
41	False	False
42	False	False
43	False	False
44	False	False
45	False	False
46	False	False
47	False	False
48	False	False
49	False	False
50	True	True

	Group
0	False
1	False
2	False
3	False
4	False
5	False
6	False
7	False
8	False
9	False
10	False
11	False
12	False
13	False
14	False
15	False
16	False
17	False
18	False
19	False
20	False
21	False
22	False
23	False
24	False

```
25  False
26  False
27  False
28  False
29  False
30  False
31  False
32  False
33  False
34  False
35  False
36  False
37  False
38  False
39  False
40  False
41  False
42  False
43  False
44  False
45  False
46  False
47  False
48  False
49  False
50   True
```

[ ]:
```
df.tail()
```

[ ]:
```
      Undergraduate Major  Starting Median Salary  Mid-Career Median Salary  \
46              Psychology                 35900.0                   60400.0
47                Religion                 34100.0                   52000.0
48              Sociology                  36500.0                   58200.0
49                Spanish                  34000.0                   53100.0
50  Source: PayScale Inc.                      NaN                       NaN

    Mid-Career 10th Percentile Salary  Mid-Career 90th Percentile Salary Group
46                            31600.0                           127000.0  HASS
47                            29700.0                            96400.0  HASS
48                            30700.0                           118000.0  HASS
49                            31000.0                            96400.0  HASS
50                                NaN                                NaN   NaN
```

[2]:
```
clean_df = df.dropna()
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
```

```
<ipython-input-2-dc972890cf10> in <cell line: 1>()
----> 1 clean_df = df.dropna()

NameError: name 'df' is not defined
```

[ ]:
```
clean_df.tail()
```

[ ]:
```
   Undergraduate Major  Starting Median Salary  Mid-Career Median Salary  \
45   Political Science                 40800.0                   78200.0
46          Psychology                 35900.0                   60400.0
47            Religion                 34100.0                   52000.0
48           Sociology                 36500.0                   58200.0
49             Spanish                 34000.0                   53100.0

    Mid-Career 10th Percentile Salary  Mid-Career 90th Percentile Salary Group
45                            41200.0                           168000.0  HASS
46                            31600.0                           127000.0  HASS
47                            29700.0                            96400.0  HASS
48                            30700.0                           118000.0  HASS
49                            31000.0                            96400.0  HASS
```

[ ]:
```
df['Starting Median Salary']
```

[ ]:
```
0     46000.0
1     57700.0
2     42600.0
3     36800.0
4     41600.0
5     35800.0
6     38800.0
7     43000.0
8     63200.0
9     42600.0
10    53900.0
11    38100.0
12    61400.0
13    55900.0
14    53700.0
15    35000.0
16    35900.0
17    50100.0
18    34900.0
19    60900.0
20    38000.0
21    37900.0
22    47900.0
23    39100.0
```

```
24    41200.0
25    43500.0
26    35700.0
27    38800.0
28    39200.0
29    37800.0
30    57700.0
31    49100.0
32    36100.0
33    40900.0
34    35600.0
35    49200.0
36    40800.0
37    45400.0
38    57900.0
39    35900.0
40    54200.0
41    39900.0
42    39900.0
43    74300.0
44    50300.0
45    40800.0
46    35900.0
47    34100.0
48    36500.0
49    34000.0
50        NaN
Name: Starting Median Salary, dtype: float64
```

[ ]:
```
clean_df['Starting Median Salary']
```

[ ]:
```
0     46000.0
1     57700.0
2     42600.0
3     36800.0
4     41600.0
5     35800.0
6     38800.0
7     43000.0
8     63200.0
9     42600.0
10    53900.0
11    38100.0
12    61400.0
13    55900.0
14    53700.0
15    35000.0
```

```
16    35900.0
17    50100.0
18    34900.0
19    60900.0
20    38000.0
21    37900.0
22    47900.0
23    39100.0
24    41200.0
25    43500.0
26    35700.0
27    38800.0
28    39200.0
29    37800.0
30    57700.0
31    49100.0
32    36100.0
33    40900.0
34    35600.0
35    49200.0
36    40800.0
37    45400.0
38    57900.0
39    35900.0
40    54200.0
41    39900.0
42    39900.0
43    74300.0
44    50300.0
45    40800.0
46    35900.0
47    34100.0
48    36500.0
49    34000.0
Name: Starting Median Salary, dtype: float64
```

[ ]:
```
clean_df['Starting Median Salary'].max()
```

[ ]:
```
74300.0
```

[ ]:
```
clean_df['Starting Median Salary'].idxmax()
```

[ ]:
```
43
```

[ ]:
```
clean_df['Undergraduate Major'][43]
```

[ ]:
```
'Physician Assistant'
```

```
[ ]: clean_df['Undergraduate Major'].loc[43]
```

```
[ ]: 'Physician Assistant'
```

```
[ ]: clean_df.loc[43]
```

```
[ ]: Undergraduate Major                 Physician Assistant
     Starting Median Salary                         74300.0
     Mid-Career Median Salary                       91700.0
     Mid-Career 10th Percentile Salary              66400.0
     Mid-Career 90th Percentile Salary             124000.0
     Group                                            STEM
     Name: 43, dtype: object
```

```
[ ]: clean_df['Mid-Career Median Salary']
```

```
[ ]: clean_df['Mid-Career Median Salary'].idxmax()
```

```
[ ]: clean_df['Mid-Career Median Salary'].loc[8]
```

```
[ ]: 107000.0
```

```
[ ]: clean_df.loc[8]
```

```
[ ]: Undergraduate Major                 Chemical Engineering
     Starting Median Salary                          63200.0
     Mid-Career Median Salary                       107000.0
     Mid-Career 10th Percentile Salary               71900.0
     Mid-Career 90th Percentile Salary              194000.0
     Group                                             STEM
     Name: 8, dtype: object
```

```
[ ]: clean_df['Starting Median Salary'].min()
```

```
[ ]: 34000.0
```

```
[ ]: clean_df['Undergraduate Major'].loc[clean_df['Starting Median Salary'].idxmin()]
```

```
[ ]: 'Spanish'
```

```
[ ]: clean_df.loc[clean_df['Starting Median Salary'].idxmin()]
```

```
[ ]: Undergraduate Major                 Spanish
     Starting Median Salary                34000.0
     Mid-Career Median Salary              53100.0
     Mid-Career 10th Percentile Salary     31000.0
     Mid-Career 90th Percentile Salary     96400.0
```

```
Group                                    HASS
Name: 49, dtype: object
```

[ ]:
```
clean_df.loc[clean_df['Mid-Career Median Salary'].idxmin()]
```

[ ]:
```
Undergraduate Major                 Education
Starting Median Salary                34900.0
Mid-Career Median Salary              52000.0
Mid-Career 10th Percentile Salary     29300.0
Mid-Career 90th Percentile Salary    102000.0
Group                                   HASS
Name: 18, dtype: object
```

[ ]:
```
clean_df.loc[clean_df['Mid-Career Median Salary'].idxmax()]
```

[ ]:
```
Undergraduate Major                 Chemical Engineering
Starting Median Salary                           63200.0
Mid-Career Median Salary                        107000.0
Mid-Career 10th Percentile Salary                71900.0
Mid-Career 90th Percentile Salary               194000.0
Group                                              STEM
Name: 8, dtype: object
```

[3]:
```
low_risk = clean_df.sort_values('Spread',ascending= False)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-3-ace4b19fa4bf> in <cell line: 1>()
----> 1 low_risk = clean_df.sort_values('Spread',ascending= False)

NameError: name 'clean_df' is not defined
```

[ ]:
```
clean_df.sort_values('Spread', ascending = False)
```

[ ]:
```
low_risk[['Undergraduate Major', 'Spread']].head()
```

[ ]:
```
   Undergraduate Major    Spread
17           Economics  159400.0
22             Finance  147800.0
37                Math  137800.0
36           Marketing  132900.0
42          Philosophy  132500.0
```

[ ]:
```
clean_df.sort_values('Mid-Career 90th Percentile Salary', ascending = False).
↪head()
```

[ ]:
```
      Spread   Undergraduate Major  Starting Median Salary  \
17  159400.0             Economics                 50100.0
22  147800.0               Finance                 47900.0
8   122100.0  Chemical Engineering                 63200.0
37  137800.0                  Math                 45400.0
44  122000.0               Physics                 50300.0

    Mid-Career Median Salary  Mid-Career 10th Percentile Salary  \
17                   98600.0                            50600.0
22                   88300.0                            47200.0
8                   107000.0                            71900.0
37                   92400.0                            45200.0
44                   97300.0                            56000.0

    Mid-Career 90th Percentile Salary     Group
17                           210000.0  Business
22                           195000.0  Business
8                            194000.0      STEM
37                           183000.0      STEM
44                           178000.0      STEM
```

[ ]:
```
highest_potential = clean_df.sort_values('Mid-Career 90th Percentile Salary',␣
↪ascending=False)
highest_potential[['Undergraduate Major', 'Mid-Career 90th Percentile Salary']].
↪head()
```

[ ]:
```
     Undergraduate Major  Mid-Career 90th Percentile Salary
17             Economics                           210000.0
22               Finance                           195000.0
8   Chemical Engineering                           194000.0
37                  Math                           183000.0
44               Physics                           178000.0
```

[ ]:
```
clean_df.groupby('Group').count()
```

[ ]:
```
          Spread  Undergraduate Major  Starting Median Salary  \
Group
Business      12                   12                      12
HASS          22                   22                      22
STEM          16                   16                      16

          Mid-Career Median Salary  Mid-Career 10th Percentile Salary  \
Group
Business                        12                                 12
HASS                            22                                 22
STEM                            16                                 16
```

```
          Mid-Career 90th Percentile Salary
Group
Business                                 12
HASS                                     22
STEM                                     16
```

[ ]:
```
clean_df.groupby('Group').mean()
```

```
<ipython-input-64-61aa9b47bd8a>:1: FutureWarning: The default value of
numeric_only in DataFrameGroupBy.mean is deprecated. In a future version,
numeric_only will default to False. Either specify numeric_only or select only
columns which should be valid for the function.
  clean_df.groupby('Group').mean()
```

[ ]:
```
                Spread  Starting Median Salary  Mid-Career Median Salary  \
Group
Business  103958.333333            44633.333333              75083.333333
HASS       95218.181818            37186.363636              62968.181818
STEM      101600.000000            53862.500000              90812.500000

          Mid-Career 10th Percentile Salary  Mid-Career 90th Percentile Salary
Group
Business                       43566.666667                     147525.000000
HASS                           34145.454545                     129363.636364
STEM                           56025.000000                     157625.000000
```

[ ]:
```
pd.options.display.float_format = '{:,.2f}'.format
```

[ ]:
```
clean_df.groupby('Group').mean()
```

```
<ipython-input-66-61aa9b47bd8a>:1: FutureWarning: The default value of
numeric_only in DataFrameGroupBy.mean is deprecated. In a future version,
numeric_only will default to False. Either specify numeric_only or select only
columns which should be valid for the function.
  clean_df.groupby('Group').mean()
```

[ ]:
```
             Spread  Starting Median Salary  Mid-Career Median Salary  \
Group
Business 103,958.33               44,633.33                 75,083.33
HASS      95,218.18               37,186.36                 62,968.18
STEM     101,600.00               53,862.50                 90,812.50

          Mid-Career 10th Percentile Salary  Mid-Career 90th Percentile Salary
Group
Business                          43,566.67                        147,525.00
HASS                              34,145.45                        129,363.64
STEM                              56,025.00                        157,625.00
```

[ ]: