Data Exploration Pandas College Major

March 13, 2022

[1]:
```
# imports
import pandas as pd
```

[2]:
```
# number formats in the output
pd.options.display.float_format = '{:,.2f}'.format
```

[3]:
```
data_frame = pd.read_csv("salaries_by_college_major.csv")
```

1 Preliminary Data Exploration and Data Cleaning with Pandas

[4]:
```
# preview first 5 rows of our dataset
data_frame.head()
```

[4]:

	Undergraduate Major	Starting Median Salary	Mid-Career Median Salary	Mid-Career 10th Percentile Salary	Mid-Career 90th Percentile Salary	Group
0	Accounting	46,000.00	77,100.00	42,200.00	152,000.00	Business
1	Aerospace Engineering	57,700.00	101,000.00	64,300.00	161,000.00	STEM
2	Agriculture	42,600.00	71,900.00	36,300.00	150,000.00	Business
3	Anthropology	36,800.00	61,500.00	33,800.00	138,000.00	HASS
4	Architecture	41,600.00	76,800.00	50,600.00	136,000.00	Business

[5]:
```
# preview last 5 rows of our dataset
data_frame.tail()
```

[5]:
```
      Undergraduate Major  Starting Median Salary  Mid-Career Median Salary  \
46            Psychology                35,900.00                 60,400.00
47              Religion                34,100.00                 52,000.00
48             Sociology                36,500.00                 58,200.00
49               Spanish                34,000.00                 53,100.00
50  Source: PayScale Inc.                     NaN                       NaN

    Mid-Career 10th Percentile Salary  Mid-Career 90th Percentile Salary Group
46                          31,600.00                         127,000.00  HASS
47                          29,700.00                          96,400.00  HASS
48                          30,700.00                         118,000.00  HASS
49                          31,000.00                          96,400.00  HASS
50                                NaN                                NaN   NaN
```

[6]:
```
# check the number of rows and columns of our dataset
data_frame.shape
```

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

[7]:
```
# check the columns
data_frame.columns
```

[7]:
```
Index(['Undergraduate Major', 'Starting Median Salary',
       'Mid-Career Median Salary', 'Mid-Career 10th Percentile Salary',
       'Mid-Career 90th Percentile Salary', 'Group'],
      dtype='object')
```

[8]:
```
# check for missing values and chunk data
# the isna() checks if a cell is a NaN.
# NaN values are blank cells or cells that contain strings instead of numbers
data_frame.isna()
```

[8]:
```
    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
```

[9]:
```
# deleting the row that contains junk data / or not needed
# in our case we want to delete the last row
# we use dropna() on a new data_frame
clean_df = data_frame.dropna()
clean_df.tail()
```

[9]:

	Undergraduate Major	Starting Median Salary	Mid-Career Median Salary	Mid-Career 10th Percentile Salary	Mid-Career 90th Percentile Salary	Group
45	Political Science	40,800.00	78,200.00	41,200.00	168,000.00	HASS
46	Psychology	35,900.00	60,400.00	31,600.00	127,000.00	HASS
47	Religion	34,100.00	52,000.00	29,700.00	96,400.00	HASS
48	Sociology	36,500.00	58,200.00	30,700.00	118,000.00	HASS
49	Spanish	34,000.00	53,100.00	31,000.00	96,400.00	HASS

2 Accessing Columns and Individual Cells in a Dataframe.

[10]:
```
# find college major with the highest starting salary
starting_salaries = clean_df["Starting Median Salary"]
clean_df[starting_salaries == starting_salaries.max()]
```

[10]:

	Undergraduate Major	Starting Median Salary	Mid-Career Median Salary	Mid-Career 10th Percentile Salary	Mid-Career 90th Percentile Salary	Group
43	Physician Assistant	74,300.00	91,700.00	66,400.00	124,000.00	STEM

[11]:
```
# method 2 | getting the row that has the largest starting salary on average
starting_salaries.max()
```

[11]:
```
74300.0
```

[12]:
```
# method 3 | using the idxmax() to get the index
starting_salaries.idxmax()
```

[12]:
```
43
```

[13]:
```
clean_df["Undergraduate Major"].loc[43]
```

[13]:
```
'Physician Assistant'
```

[14]:
```
clean_df["Undergraduate Major"][43] # another way of accessing a particular␣
↪value
```

[14]:
```
'Physician Assistant'
```

[15]:
```
clean_df.loc[43] # retrieves data of the entire row
```

[15]:
```
Undergraduate Major                 Physician Assistant
Starting Median Salary                        74,300.00
Mid-Career Median Salary                      91,700.00
Mid-Career 10th Percentile Salary             66,400.00
Mid-Career 90th Percentile Salary            124,000.00
Group                                              STEM
Name: 43, dtype: object
```

[16]:
```
"""
Challenge

1. What college major has the highest mid-career salary?
   How much do graduates with this major earn? (Mid-career is defined as having␣
↪10+ years of experience).

2. Which college major has the lowest starting salary and how much do graduates␣
↪earn after university?

3. Which college major has the lowest mid-career salary and how much can people␣
↪expect to earn with this degree?
"""
```

[16]:
```
'\nChallenge\n\n1. What college major has the highest mid-career salary? \n
How much do graduates with this major earn? (Mid-career is defined as having 10+
years of experience).\n\n2. Which college major has the lowest starting salary
and how much do graduates earn after university?\n\n3. Which college major has
the lowest mid-career salary and how much can people expect to earn with this
degree? \n'
```

[17]:
```
# college major having the highest mid-career salary
mid_career_salaries = clean_df["Mid-Career 10th Percentile Salary"]
mid_career_salaries.max()
```

[17]:
```
71900.0
```

[18]:
```
# find the index of the row containing the highest mid-career salary
mid_career_salaries.idxmax()
```

[18]: 8

[19]:
```
# locate the major
clean_df["Undergraduate Major"][8]
```

[19]: 'Chemical Engineering'

[20]:
```
# college major having the lowest starting salary
lowest_starting_salaries = clean_df["Starting Median Salary"]
lowest_starting_salaries.min()
```

[20]: 34000.0

[21]:
```
# find the index of the row containing the minimum starting salary
lowest_starting_salaries.idxmin()
```

[21]: 49

[22]:
```
# locate the major
clean_df["Undergraduate Major"][49]
```

[22]: 'Spanish'

[23]:
```
# Spanish graduates earn after university
clean_df["Mid-Career Median Salary"][49]
```

[23]: 53100.0

[24]:
```
# college major having the lowest-mid career salary
lowest_mid_career = clean_df["Mid-Career 10th Percentile Salary"]
lowest_mid_career.min()
```

[24]: 26700.0

[25]:
```
# find the index of the row containing the minimum starting salary
lowest_mid_career.idxmin()
```

[25]: 39

[26]:
```
# locate the major
clean_df["Undergraduate Major"][39]
```

[26]: 'Music'

[27]:
```
# Music major expected salary to earn
clean_df["Mid-Career 90th Percentile Salary"][39]
```

[27]:
```
134000.0
```

3 Sorting Values & Adding Columns: Majors with the Most Potential vs Lowest Risk

[28]:
```
# calculate the difference between the earnings of the 10th and 90th percentile
# method 1
print( (clean_df["Mid-Career 90th Percentile Salary"] - clean_df["Mid-Career␣
 ↪10th Percentile Salary"]).head() )
```

```
0   109,800.00
1    96,700.00
2   113,700.00
3   104,200.00
4    85,400.00
dtype: float64
```

[29]:
```
# method 2
difference_in_salaries = clean_df["Mid-Career 90th Percentile Salary"].
 ↪subtract(clean_df["Mid-Career 10th Percentile Salary"])
```

[30]:
```
# print the first 5 rows
difference_in_salaries.head()
```

[30]:
```
0   109,800.00
1    96,700.00
2   113,700.00
3   104,200.00
4    85,400.00
dtype: float64
```

[31]:
```
# add difference_in_salaries Series to our existing DataFrame
clean_df.insert(1, "spread", difference_in_salaries)
```

[32]:
```
clean_df.head()
```

[32]:

	Undergraduate Major	spread	Starting Median Salary	\
0	Accounting	109,800.00	46,000.00	
1	Aerospace Engineering	96,700.00	57,700.00	
2	Agriculture	113,700.00	42,600.00	
3	Anthropology	104,200.00	36,800.00	
4	Architecture	85,400.00	41,600.00	

	Mid-Career Median Salary	Mid-Career 10th Percentile Salary	\
0	77,100.00	42,200.00	
1	101,000.00	64,300.00	

```
2                 71,900.00                          36,300.00
3                 61,500.00                          33,800.00
4                 76,800.00                          50,600.00

   Mid-Career 90th Percentile Salary     Group
0                        152,000.00  Business
1                        161,000.00      STEM
2                        150,000.00  Business
3                        138,000.00      HASS
4                        136,000.00  Business
```

[33]:
```
# Sorting by the Lowest Spread
lowest_risk = clean_df.sort_values("spread", ascending=False)
```

[34]:
```
type(lowest_risk)
```

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

[35]:
```
# display the head() of the Undergraduate Major and the Spread
# to do this we pass a list of columns to the DataFrame

print( "Majors having the Lowest Spread:\n" )
print( lowest_risk[ ["Undergraduate Major", "spread"] ].tail() )

print( "\nMajors having the Greatest Spread:\n" )
print( lowest_risk[ ["Undergraduate Major", "spread"] ].head() )
```

```
Majors having the Lowest Spread:

           Undergraduate Major    spread
27  Health Care Administration 66,400.00
49                     Spanish 65,400.00
41                   Nutrition 65,300.00
43         Physician Assistant 57,600.00
40                     Nursing 50,700.00

Majors having the Greatest Spread:

   Undergraduate Major     spread
17           Economics 159,400.00
22             Finance 147,800.00
37                Math 137,800.00
36           Marketing 132,900.00
42          Philosophy 132,500.00
```

[36]:
```
# Sorting by the Highest values in the 90th percentile
```

```
highest_values = clean_df.sort_values("Mid-Career 90th Percentile Salary",␣
↪ascending=False)
```

[37]:
```
highest_values[ ["Undergraduate Major", "Mid-Career 90th Percentile Salary"] ].
↪head()
```

[37]:

	Undergraduate Major	Mid-Career 90th Percentile Salary
17	Economics	210,000.00
22	Finance	195,000.00
8	Chemical Engineering	194,000.00
37	Math	183,000.00
44	Physics	178,000.00

[38]:
```
# Sorting by the Highest values in the Mid-Career Median Salary
mid_career_salary = clean_df.sort_values("Mid-Career Median Salary",␣
↪ascending=False)
```

[39]:
```
# highest mid-career median salary
mid_career_salary[ ["Undergraduate Major", "Mid-Career Median Salary"] ].head()
```

[39]:

	Undergraduate Major	Mid-Career Median Salary
8	Chemical Engineering	107,000.00
12	Computer Engineering	105,000.00
19	Electrical Engineering	103,000.00
1	Aerospace Engineering	101,000.00
17	Economics	98,600.00

[40]:
```
# Lowest mid-career median salary
mid_career_salary[ ["Undergraduate Major", "Mid-Career Median Salary"] ].tail()
```

[40]:

	Undergraduate Major	Mid-Career Median Salary
39	Music	55,000.00
32	Interior Design	53,200.00
49	Spanish	53,100.00
18	Education	52,000.00
47	Religion	52,000.00

[41]:
```
# difference between the Largest and Lowest mid-career median salaries
highest_mid_career_salary = mid_career_salary["Mid-Career Median Salary"].head()
lowest_mid_career_salary = mid_career_salary["Mid-Career Median Salary"].tail()
```

[42]:
```
highest_mid_career_salary
```

[42]:
```
8    107,000.00
12   105,000.00
19   103,000.00
1    101,000.00
```

```
17    98,600.00
Name: Mid-Career Median Salary, dtype: float64
```

[43]:
```
lowest_mid_career_salary
```

[43]:
```
39   55,000.00
32   53,200.00
49   53,100.00
18   52,000.00
47   52,000.00
Name: Mid-Career Median Salary, dtype: float64
```

[44]:
```
type(highest_mid_career_salary)
```

[44]:
```
pandas.core.series.Series
```

4 Grouping and Pivoting Data with Pandas

[45]:
```
"""
We have three categories in the 'Group' column:
STEM, HASS and Business.
Let's count how many majors we have in each category:
.groupby() method. This allows us to manipulate data similar to a Microsoft␣
 ↪Excel Pivot Table.
"""
clean_df.groupby("Group").count()
```

[45]:
```
          Undergraduate Major  spread  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
```

[46]:
```
type(clean_df.groupby("Group"))
```

[46]:
```
pandas.core.groupby.generic.DataFrameGroupBy
```

[47]:
```
# finding the average salary by group?
clean_df.groupby("Group").mean()
```

[47]:

Group	spread	Starting Median Salary	Mid-Career Median Salary	\
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	

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