Data Analytics Questions

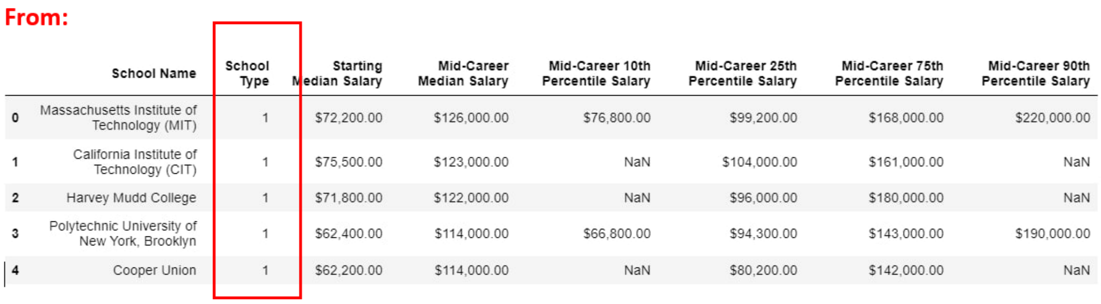
You were given a 'practice_dataset.csv' dataset, that contains data about average salary of some school graduates. Please read in this file here, and explore it.

In [ ]:

```
#please code here
```

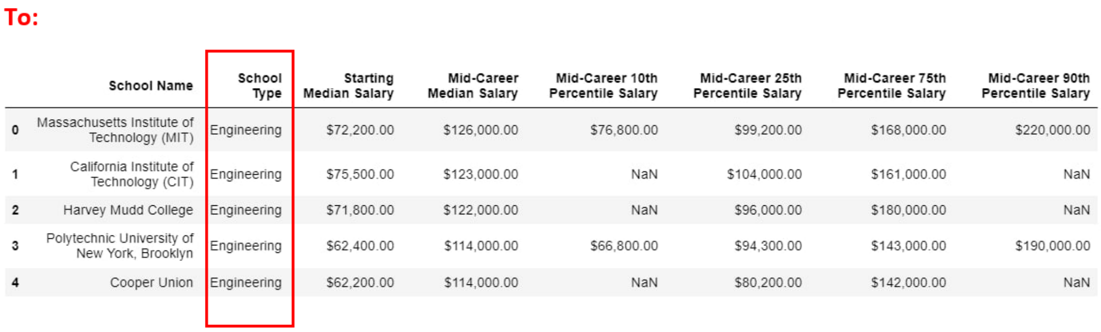
Task 1.

In the dataset there is a 'School Type' column that has numircal values: they are IDs for keys that are given in 'school_type.json' file. Please, map over these IDs to replace them with their keys. Here is an expected outcome:



From:

	School Name	School Type	Starting Median Salary	Mid-Career Median Salary	Mid-Career 10th Percentile Salary	Mid-Career 25th Percentile Salary	Mid-Career 75th Percentile Salary	Mid-Career 90th Percentile Salary
0	Massachusetts Institute of Technology (MIT)	1	$72,200.00	$126,000.00	$76,800.00	$99,200.00	$168,000.00	$220,000.00
1	California Institute of Technology (CIT)	1	$75,500.00	$123,000.00	NaN	$104,000.00	$161,000.00	NaN
2	Harvey Mudd College	1	$71,800.00	$122,000.00	NaN	$96,000.00	$180,000.00	NaN
3	Polytechnic University of New York, Brooklyn	1	$62,400.00	$114,000.00	$66,800.00	$94,300.00	$143,000.00	$190,000.00
4	Cooper Union	1	$62,200.00	$114,000.00	NaN	$80,200.00	$142,000.00	NaN



To:

	School Name	School Type	Starting Median Salary	Mid-Career Median Salary	Mid-Career 10th Percentile Salary	Mid-Career 25th Percentile Salary	Mid-Career 75th Percentile Salary	Mid-Career 90th Percentile Salary
0	Massachusetts Institute of Technology (MIT)	Engineering	$72,200.00	$126,000.00	$76,800.00	$99,200.00	$168,000.00	$220,000.00
1	California Institute of Technology (CIT)	Engineering	$75,500.00	$123,000.00	NaN	$104,000.00	$161,000.00	NaN
2	Harvey Mudd College	Engineering	$71,800.00	$122,000.00	NaN	$96,000.00	$180,000.00	NaN
3	Polytechnic University of New York, Brooklyn	Engineering	$62,400.00	$114,000.00	$66,800.00	$94,300.00	$143,000.00	$190,000.00
4	Cooper Union	Engineering	$62,200.00	$114,000.00	NaN	$80,200.00	$142,000.00	NaN

In [16]:

```
import json
import re
import pandas as pd

with open('school_type.json', 'r') as json_file:
    json_text = json_file.read()

#Replacing <'> with <"> in JSON, e.g. 'Engineering' to "Engineering"
json_text = re.sub(r"\'(.*?)\'", r'"\1"', json_text)

school_type_data = json.loads(json_text)

school_type_mapping = {entry['ID']: entry['VALUE'] for entry in school_type_data}

csv_file_path = 'practice_dataset.csv'

df = pd.read_csv(csv_file_path)

df['School Type'] = df['School Type'].map(school_type_mapping)
output_csv_file_path = 'updated_practice_dataset.csv'
df.to_csv(output_csv_file_path, index=False)
print(df)
```

```
                                      School Name  School Type  \
0     Massachusetts Institute of Technology (MIT)  Engineering   
1        California Institute of Technology (CIT)  Engineering   
2                             Harvey Mudd College  Engineering   
3    Polytechnic University of New York, Brooklyn  Engineering   
4                                    Cooper Union  Engineering   
..                                            ...          ...   
264                  Austin Peay State University        State   
265                    Pittsburg State University        State   
266                      Southern Utah University        State   
267           Montana State University - Billings        State   
268                  Black Hills State University        State   

    Starting Median Salary Mid-Career Median Salary  \
0              $72,200.00             $126,000.00   
1              $75,500.00             $123,000.00   
2              $71,800.00             $122,000.00   
3              $62,400.00             $114,000.00   
4              $62,200.00             $114,000.00   
..                    ...                     ...   
264            $37,700.00              $59,200.00   
265            $40,400.00              $58,200.00   
266            $41,900.00              $56,500.00   
267            $37,900.00              $50,600.00   
268            $35,300.00              $43,900.00   

    Mid-Career 10th Percentile Salary Mid-Career 25th Percentile Salary  \
0                          $76,800.00                        $99,200.00   
1                                  NaN                      $104,000.00   
2                                  NaN                       $96,000.00   
3                          $66,800.00                        $94,300.00   
4                                  NaN                       $80,200.00   
..                                 ...                              ...   
264                        $32,200.00                        $40,500.00   
265                        $25,600.00                        $46,000.00   
266                        $30,700.00                        $39,700.00   
267                        $22,600.00                        $31,800.00   
268                        $27,000.00                        $32,200.00   

    Mid-Career 75th Percentile Salary Mid-Career 90th Percentile Salary  
0                         $168,000.00                       $220,000.00  
1                         $161,000.00                               NaN  
2                         $180,000.00                               NaN  
3                         $143,000.00                       $190,000.00  
4                         $142,000.00                               NaN  
..                                ...                               ...  
264                        $73,900.00                        $96,200.00  
265                        $84,600.00                       $117,000.00  
266                        $78,400.00                       $116,000.00  
267                        $78,500.00                        $98,900.00  
268                        $60,900.00                        $87,600.00  

[269 rows x 8 columns]
```

Task 2

We defined a function that takes any 'School Type' value, and estimates rounded average 'Mid-Career Median Salary'for it. However, our funciton is not working. Please find an error and try to fix it.

In [1]:

```
import pandas as pd

csv_file_path = 'updated_practice_dataset.csv'
df = pd.read_csv(csv_file_path)

df['Mid-Career Median Salary'] = df['Mid-Career Median Salary'].replace('[\$,]', '', reg

def function_1(df, school_type):
    result = round(df[df['School Type'] == school_type]['Mid-Career Median Salary'].mean
    return result

df.to_csv(csv_file_path, index=False)
```

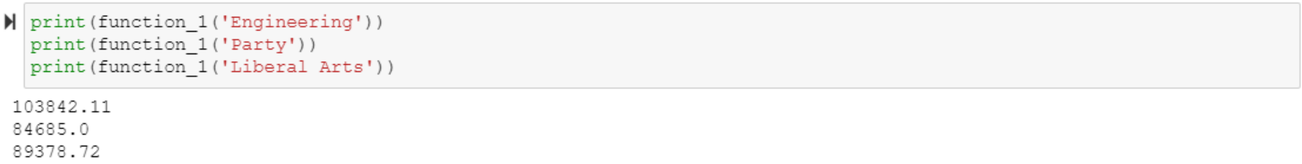
In [21]:

```
function_1(df,'Engineering')
```

Out[21]:

```
103842.11
```

If you fix an error, apply this function to values 'Engineering', 'Party', 'Liberal Arts' and print output of the function, the end result must looks like this:



```
print(function_1('Engineering'))
print(function_1('Party'))
print(function_1('Liberal Arts'))
```

```
103842.11
84685.0
89378.72
```

In [22]:

```
print(function_1(df, 'Engineering'))
print(function_1(df, 'Party'))
print(function_1(df, 'Liberal Arts'))
```

```
103842.11
84685.0
89378.72
```

Task 3

According to the National Occupational Employment and Wages Estimates, the average salary in the United States is 56,310 USD annually. Iterate over 'Starting Median Salary' column and assign value 'more than national average' if it is more than 56,310 USD, else 'less than national average'. The result is supposed to be as such:

	School Name	School Type	Starting Median Salary	Mid-Career Median Salary	Mid-Career 10th Percentile Salary	Mid-Career 25th Percentile Salary	Mid-Career 75th Percentile Salary	Mid-Career 90th Percentile Salary
0	Massachusetts Institute of Technology (MIT)	Engineering	more than national average	126000	$76,800.00	$99,200.00	$168,000.00	$220,000.00
1	California Institute of Technology (CIT)	Engineering	more than national average	123000	NaN	$104,000.00	$161,000.00	NaN
2	Harvey Mudd College	Engineering	more than national average	122000	NaN	$96,000.00	$180,000.00	NaN
3	Polytechnic University of New York, Brooklyn	Engineering	more than national average	114000	$66,800.00	$94,300.00	$143,000.00	$190,000.00
4	Cooper Union	Engineering	more than national average	114000	NaN	$80,200.00	$142,000.00	NaN

In [2]:

```
import pandas as pd

csv_file_path = 'updated_practice_dataset.csv'
df = pd.read_csv(csv_file_path)

national_average_salary = 56310

df['Starting Median Salary'] = df['Starting Median Salary'].str.replace('[$,]', '', rege

df['Starting Median Salary'] = df['Starting Median Salary'].apply(lambda x: 'more than n

df.to_csv(csv_file_path, index=False)
```

Now, display all state schools that have less than national average salary.

In [25]:

```
print(df)
```

```
                                    School Name  School Type  \
0     Massachusetts Institute of Technology (MIT)  Engineering
1        California Institute of Technology (CIT)  Engineering
2                             Harvey Mudd College  Engineering
3    Polytechnic University of New York, Brooklyn  Engineering
4                                    Cooper Union  Engineering
..                                            ...          ...
264                     Austin Peay State University        State
265                       Pittsburg State University        State
266                         Southern Utah University        State
267            Montana State University - Billings        State
268                     Black Hills State University        State

         Starting Median Salary Mid-Career Median Salary  \
0    more than national average              $126,000.00
1    more than national average              $123,000.00
2    more than national average              $122,000.00
3    more than national average              $114,000.00
4    more than national average              $114,000.00
```

Task4

You might have realised that some columns have missing values. Display all rows that has at least one missing value in any column. Then, consider how would you handle these missing values? Please, describe below your thoughts

In [ ]:

```
# We can use the isna() or isnull() method along with the any() function.
# Setting "axis=1" will help identify rows with missing values. Regarding how to handle
# Remove Rows with Missing Values, Imputation with Mean/Median, Imputation with Mode, Fo
```

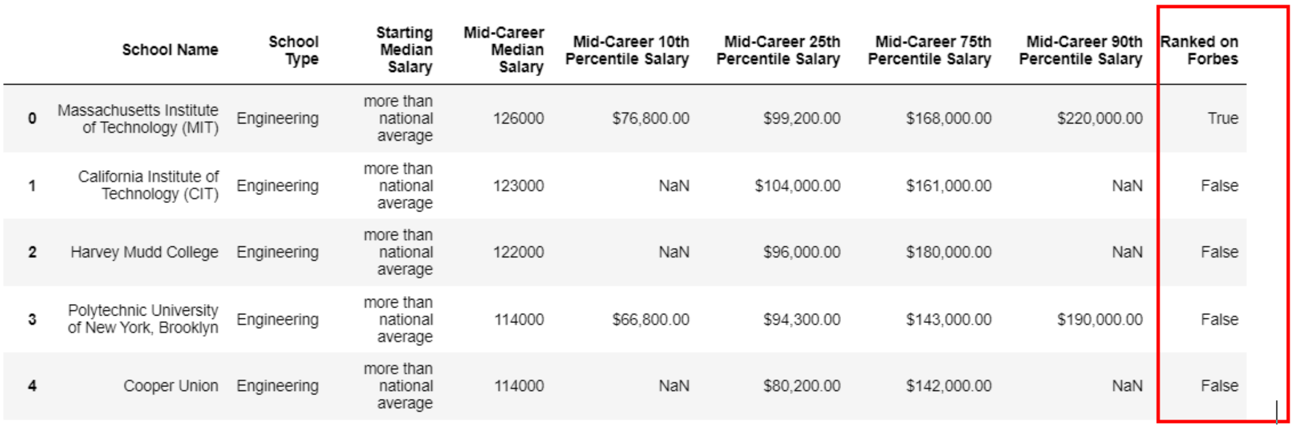
In [26]:

```
import pandas as pd

csv_file_path = 'updated_practice_dataset.csv'
df = pd.read_csv(csv_file_path)

rows_with_missing_values = df[df.isnull().any(axis=1)]

print(rows_with_missing_values)
```

```
58                       $131,000.00                          NaN
59                       $185,000.00                          NaN
62                       $129,000.00                          NaN
63                       $132,000.00                          NaN
66                       $123,000.00                          NaN
67                       $123,000.00                          NaN
68                       $125,000.00                          NaN
69                       $122,000.00                          NaN
70                       $128,000.00                          NaN
72                       $148,000.00                          NaN
73                       $101,000.00                          NaN
74                       $101,000.00                          NaN
75                       $111,000.00                          NaN
77                       $110,000.00                          NaN
78                       $131,000.00                          NaN
79                       $116,000.00                          NaN
80                       $147,000.00                          NaN
81                        $94,900.00                          NaN
83                        $94,000.00                          NaN
```

Task 5

Please visit this web page: https://www.upgradabroad.com/articles/forbes-unveils-americas-top-colleges-2022-list-news/ (https://www.upgradabroad.com/articles/forbes-unveils-americas-top-colleges-2022-list-news/) It contains several tables, that show some university rankings. Scroll till you reach "Forbes college rankings" field. First, scrape HTML table from this field and save as forbes_ranking. Then, write a function that creates new column "Ranked on Forbes" in practice_dataset and accepts Boolean values (True or False) based on the fact whether this university in forbes_ranking or not. Final result should look as follows:



	School Name	School Type	Starting Median Salary	Mid-Career Median Salary	Mid-Career 10th Percentile Salary	Mid-Career 25th Percentile Salary	Mid-Career 75th Percentile Salary	Mid-Career 90th Percentile Salary	Ranked on Forbes
0	Massachusetts Institute of Technology (MIT)	Engineering	more than national average	126000	$76,800.00	$99,200.00	$168,000.00	$220,000.00	True
1	California Institute of Technology (CIT)	Engineering	more than national average	123000	NaN	$104,000.00	$161,000.00	NaN	False
2	Harvey Mudd College	Engineering	more than national average	122000	NaN	$96,000.00	$180,000.00	NaN	False
3	Polytechnic University of New York, Brooklyn	Engineering	more than national average	114000	$66,800.00	$94,300.00	$143,000.00	$190,000.00	False
4	Cooper Union	Engineering	more than national average	114000	NaN	$80,200.00	$142,000.00	NaN	False

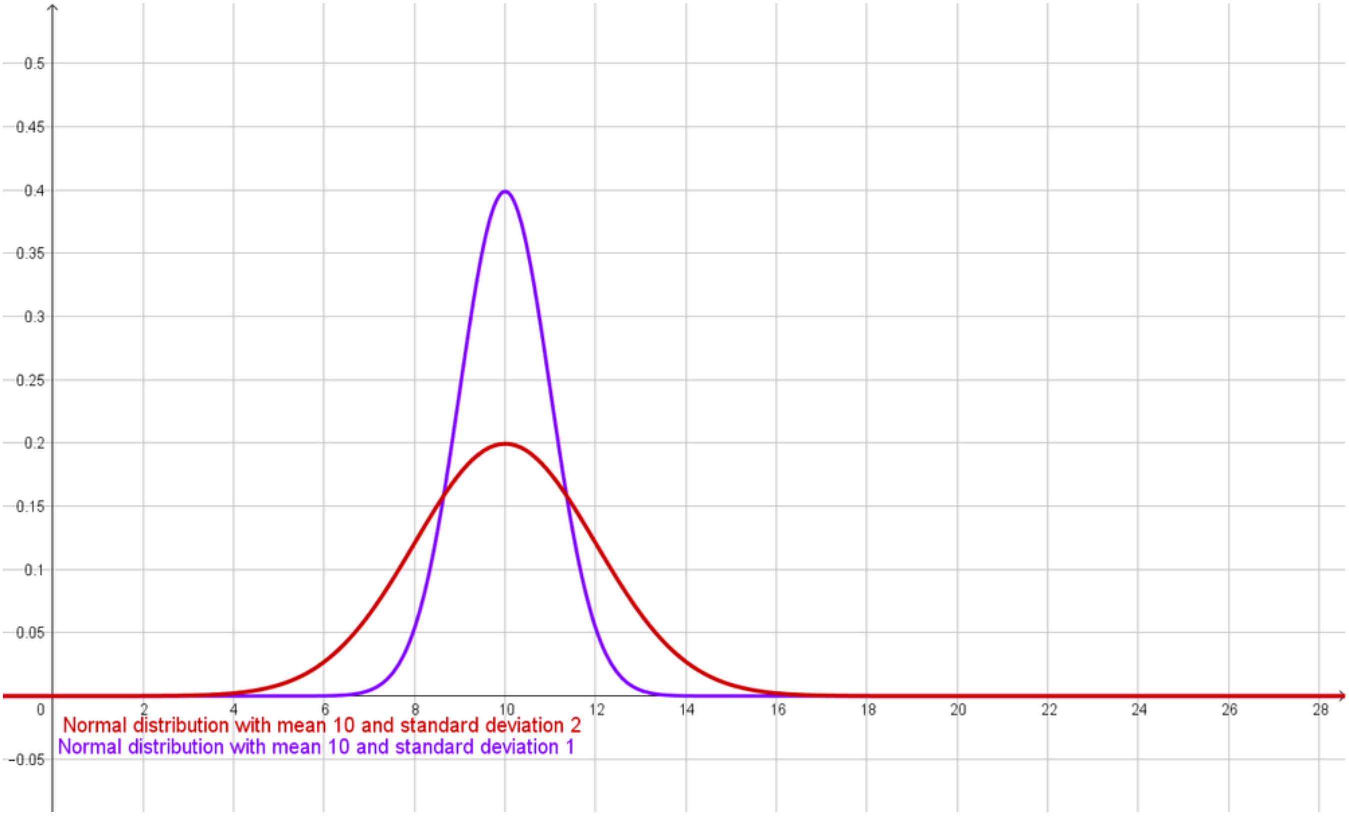
In [5]:

```
import pandas as pd

csv_file_path = 'updated_practice_dataset.csv'
df = pd.read_csv(csv_file_path)

url = 'https://www.upgradabroad.com/articles/forbes-unveils-americas-top-colleges-2022-l
all_uni = pd.read_html(url, header=0)

#Checking number of tables in the page
#print(len(all_uni))

#there is only one table
all_uni[0]

forbes_ranking = all_uni[0]
forbes_ranking.to_csv('forbes_ranking.csv', index=False)

#print(forbes_rankings)

forbes_colleges = set(forbes_ranking['Colleges'])

# Function to check if a school is ranked on Forbes
def is_ranked_on_forbes(school_name):
    return school_name in forbes_colleges

# Add a new column "Ranked on Forbes" based on the check
df['Ranked on Forbes'] = df['School Name'].apply(is_ranked_on_forbes)

df.to_csv(csv_file_path, index=False)

# Display the updated DataFrame
print(df)
```

```
                                      School Name  School Type  \
0     Massachusetts Institute of Technology (MIT)  Engineering   
1        California Institute of Technology (CIT)  Engineering   
2                             Harvey Mudd College  Engineering   
3    Polytechnic University of New York, Brooklyn  Engineering   
4                                    Cooper Union  Engineering   
..                                            ...          ...   
264                  Austin Peay State University        State   
265                    Pittsburg State University        State   
266                      Southern Utah University        State   
267           Montana State University - Billings        State   
268                  Black Hills State University        State   

         Starting Median Salary  Mid-Career Median Salary  \
0    more than national average                  126000.0   
1    more than national average                  123000.0   
2    more than national average                  122000.0   
3    more than national average                  114000.0   
4    more than national average                  114000.0   
```

Task 6

If time spent by website visitors on two difernet landing pages could be drawn as below, so that average



In [ ]:

```
# The purple landing page's time spent distribution is narrower and more concentrated ar
# This means that visitors' time spent on the purple landing page is more consistent
# and less variable compared to the red landing page.
```

Task 7

if a die is thrown 6 times, what is the probability of 3 of the numbers being even numbers?

In [15]:

```
from math import comb

n = 6 # Number of trials (throws of the die)

k = 3 # Number of successful outcomes (even numbers)

# Probability of getting an even number on a single throw
p = 1/2  # Since there are 3 even numbers out of 6 possible outcomes

probability = comb(n, k) * (p ** k) * ((1 - p) ** (n - k))

print(f"The probability of getting exactly 3 even numbers is: {probability:.4f}")
```

```
The probability of getting exactly 3 even numbers is: 0.3125
```

In [ ]:

```

```