DS5500 HW3

Surui Yang

In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import csv
import warnings
warnings.filterwarnings('ignore')
```

Problem 1

Rank and visualize the states that take in the most federal funding (revenue).

In [2]:
```
district_level_fiscal= pd.read_csv('Sdf16_1a.csv',sep='\t')
```

In [3]:
```
federal_fund_state = district_level_fiscal[district_level_fiscal['TFEDREV']>0].groupby("STNA
federal_fund_state
```

Out[3]:

	FIPST	YEAR	CCDNF	CENFILE	V33	MEMBERSCH	TOTALREV	TFEDR
STNAME								
California	6108	16288	1018	1017	6203559	6187037	89110947000	7709275(
Texas	58368	19456	1216	1044	5296442	5296378	60768409000	6194317(
New York	24732	10992	687	677	2591958	2572154	67051220000	3374794(
Florida	804	1072	67	67	2776933	2776067	28125598000	3147329(
Illinois	16524	15552	972	970	2029830	2007587	32884195000	2334945(
Pennsylvania	31836	12128	758	592	1700375	1701253	32814988000	2037315(
Ohio	41340	16960	1059	702	1711138	1709658	24870176000	1837963(
Georgia	2782	3424	214	196	1741838	1741990	19532968000	1815242(
Michigan	23010	14160	885	595	1481694	1479649	20826612000	1731034(
North Carolina	9953	4304	269	115	1543632	1543375	14119703000	1587976(
Arizona	2416	9664	604	232	1078838	1073661	9830650000	1302010(

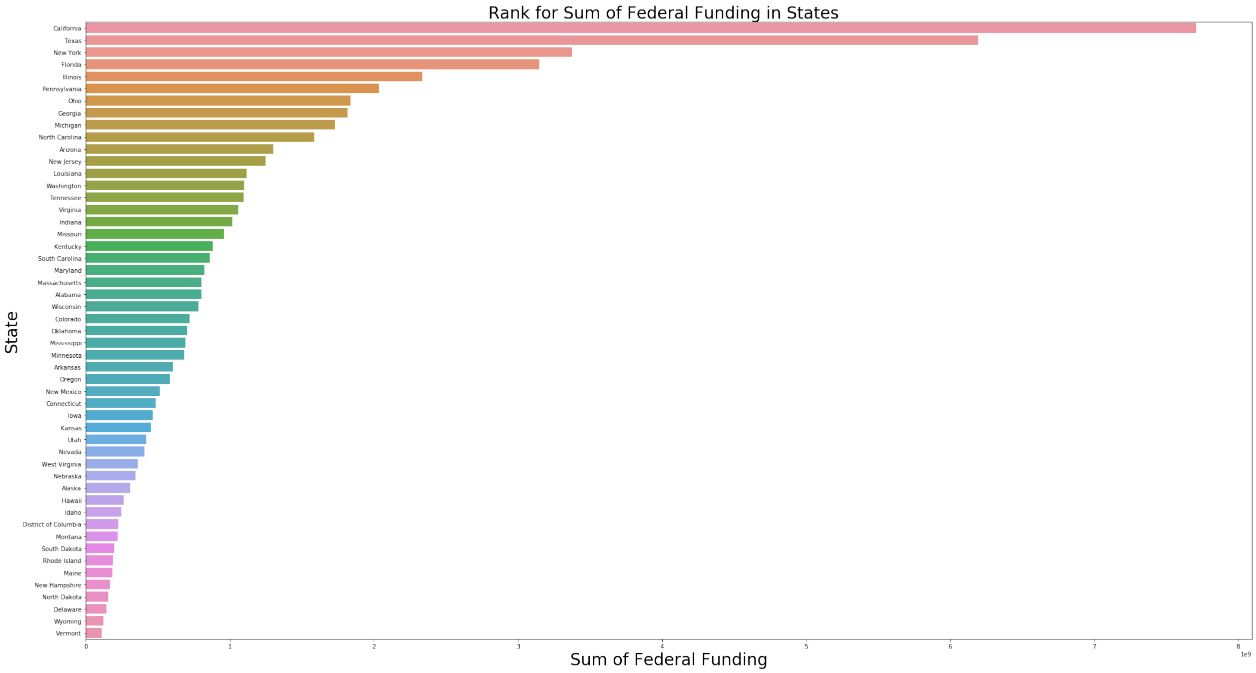
New Jersey	22950	10800	675	580	1407891	1368021	30773538000	1249741
Louisiana	3014	2192	137	69	716650	684744	9066633000	1115619
Washington	16218	4896	306	303	1084022	1084684	14972096000	1098332
Tennessee	6674	2272	142	142	999260	983814	9586098000	1096182
Virginia	6732	2112	132	132	1283493	1283451	16208682000	1058146
Indiana	7146	6352	397	309	1042311	1042215	13107646000	1015476
Missouri	16153	8912	557	518	917389	916511	11199359000	959978
Kentucky	3633	2768	173	173	686440	686094	7745928000	880296
South Carolina	4185	1488	93	89	763011	762977	9436787000	860867
Maryland	576	384	24	24	879196	879182	14409321000	823599
Massachusetts	9925	6352	397	320	961272	950659	18176359000	804595
Alabama	137	2192	137	135	743789	698300	7607098000	803907
Wisconsin	23430	6816	426	425	857796	857615	11697023000	782647
Colorado	1568	3136	196	195	895709	898376	10256106000	721719
Oklahoma	22080	8832	552	520	692605	692467	6233064000	703225
Mississippi	4032	2304	144	144	486245	371149	4755399000	690724
Minnesota	14715	8720	545	374	861784	863166	12869206000	685055
Arkansas	1350	4320	270	249	491251	491242	5513815000	606946
Oregon	8815	3440	215	215	574225	566665	7418000000	582560
New Mexico	5250	2400	150	89	334960	333323	3924863000	516289
Connecticut	1791	3184	198	173	511441	525555	11552177000	484186
Iowa	6555	5520	345	345	507996	500076	6919477000	464852
Kansas	5720	4576	286	286	495545	488229	6069563000	453922
Utah	7105	2320	145	41	647613	647613	5414412000	419642
Nevada	576	288	18	17	467371	467371	4668171000	405789
West Virginia	3510	1040	57	63	277436	277436	3420589000	360283
Nebraska	8122	4192	262	262	315520	315513	4398811000	346826
Alaska	108	864	54	54	132477	132477	2494691000	307320
Hawaii	15	16	1	1	181995	181995	3030519000	261131
Idaho	2448	2448	153	116	291737	291847	2378743000	248546
District of Columbia	682	992	62	1	82962	82144	2138284000	226202
Montana	12660	6752	422	422	145019	145039	1797849000	220340

South Dakota	6900	2400	150	150	134045	133668	1455737000	196644
Rhode Island	2728	992	62	40	141733	140353	2549378000	188204
Maine	5451	3792	236	231	180137	176148	2848422000	186523
New Hampshire	5610	2720	170	170	179664	173966	3146745000	169166
North Dakota	7676	3232	202	202	108285	106412	1787486000	155453
Delaware	470	752	47	19	134845	134306	2222987000	144707
Wyoming	2688	768	48	48	94511	94511	2044669000	123012
Vermont	9650	3088	193	193	59806	59305	1627661000	111891

51 rows × 132 columns

In [4]:
```python
f, ax = plt.subplots(figsize=(28, 15))
plt.tight_layout()
sns.barplot(x='TFEDREV',y='STNAME',data =federal_fund_state.reset_index())
plt.ylabel('State',fontsize=28)
plt.xlabel('Sum of Federal Funding',fontsize=28)
plt.title('Rank for Sum of Federal Funding in States  ',fontsize=28)
```

Out[4]:
```
Text(0.5, 1, 'Rank for Sum of Federal Funding in States  ')
```



Which states spend the most federal funding per student?

In [5]:
```python
federal_fund_state['spend_fd_pstu'] = federal_fund_state['TFEDREV']/federal_fund_state['V
federal_fund_state.sort_values(by='spend_fd_pstu',ascending= False)['spend_fd_pstu']
```

Out[5]:
```
STNAME
```

```
Out[5]: STNAME
        District of Columbia    2726.573612
        Alaska                  2319.798908
        Vermont                 1870.899241
        Louisiana               1556.713877
        New Mexico              1541.345235
        Montana                 1519.387115
        South Dakota            1466.999888
        North Dakota            1435.591264
        Hawaii                  1434.825133
        Mississippi             1420.526689
        Rhode Island            1327.877065
        New York                1302.024956
        Wyoming                 1301.562781
        West Virginia           1298.616618
        Kentucky                1282.407785
        California              1242.718091
        Arkansas                1235.510971
        Arizona                 1206.863310
        Pennsylvania            1198.156289
        Texas                   1169.524182
        Michigan                1168.280360
        Illinois                1150.315544
        Florida                 1133.383124
        South Carolina          1128.249789
        Nebraska                1099.220335
        Tennessee               1096.993775
        Alabama                 1080.826686
        Ohio                    1074.117342
        Delaware                1073.135823
        Missouri                1046.424145
        Georgia                 1042.141692
        Maine                   1035.450796
        North Carolina          1028.727054
        Oklahoma                1015.333415
        Oregon                  1014.515216
        Washington              1013.200839
        Indiana                  974.254325
        Connecticut              946.709396
        New Hampshire            941.568706
        Maryland                 936.763816
        Kansas                   916.005610
        Iowa                     915.070197
        Wisconsin                912.392923
        New Jersey               887.668861
        Nevada                   868.237439
        Idaho                    851.952272
        Massachusetts            837.010752
        Virginia                 824.426779
        Colorado                 805.751645
```

```
Minnesota            794.926571
Utah                 647.982669
Name: spend_fd_pstu, dtype: float64
```

Based on the above data, we konw that District of Columbia spent the most federal funding per student

Problem 2

Visualize the relationship between school districts’ total revenue and expenditures.

In [6]:

```
school_districts_state = district_level_fiscal[district_level_fiscal['TOTALREV']>0].groupby("S
school_districts_state
```

Out[6]:

	FIPST	YEAR	CCDNF	CENFILE	V33	MEMBERSCH	TOTALREV	TFEDR
STNAME								
Alabama	137	2192	137	135	743789	698300	7607098000	803907(
Alaska	108	864	54	54	132477	132477	2494691000	307320(
Arizona	2632	10528	658	236	1096992	1091693	9980177000	1302010(
Arkansas	1355	4336	271	249	491603	491594	5517204000	606946(
California	6300	16800	1050	1049	6203499	6186951	89224004000	7709275(
Colorado	1584	3168	198	197	895704	898770	10260558000	721719(
Connecticut	1809	3216	199	174	512461	526575	11552645000	484186(
Delaware	500	800	50	19	134841	134302	2223576000	144707(
District of Columbia	715	1040	65	1	82955	82175	2170632000	226202(
Florida	804	1072	67	67	2776933	2776067	28125598000	3147329(
Georgia	2808	3456	216	196	1755985	1756137	19610778000	1815242(
Hawaii	15	16	1	1	181995	181995	3030519000	261131(
Idaho	2480	2480	155	116	292082	292192	2382012000	248546(
Illinois	16779	15792	986	984	2029801	2007855	32918922000	2334945(
Indiana	7326	6512	407	313	1045066	1044966	13143063000	1015476(
Iowa	6555	5520	345	345	507996	500076	6919477000	464852(
Kansas	5720	4576	286	286	495545	488229	6069563000	453922(
Kentucky	3633	2768	173	173	686440	686094	7745928000	880296(

Louisiana	3058	2224	139	69	717223	685317	9082765000	1115619
Maine	6118	4256	258	259	180278	176223	2861888000	186523
Maryland	576	384	24	24	879196	879182	14409321000	823599
Massachusetts	9950	6368	398	320	961325	950712	18177253000	804595
Michigan	23322	14352	897	598	1486090	1484043	20863260000	1731034
Minnesota	15066	8928	558	386	861866	863429	12889641000	685055
Mississippi	4032	2304	144	144	486245	371149	4755399000	690724
Missouri	16153	8912	557	518	917389	916511	11199359000	959978
Montana	12870	6864	429	429	145149	145169	1800909000	220340
Nebraska	8122	4192	262	262	315520	315513	4398811000	346826
Nevada	608	304	19	17	467527	467527	4670678000	405789
New Hampshire	5808	2816	176	176	179652	173954	3150473000	169166
New Jersey	23494	11056	691	584	1408142	1368280	30820854000	1249741
New Mexico	5285	2416	151	89	335316	333679	3928180000	516289
New York	24912	11072	691	682	2591962	2572171	67055472000	3374794
North Carolina	10101	4368	273	115	1544648	1544391	14128774000	1587976
North Dakota	7866	3312	207	207	108320	106437	1788749000	155453
Ohio	42354	17376	1085	712	1712745	1711257	24929540000	1837963
Oklahoma	22160	8864	554	520	692658	692519	6237553000	703225
Oregon	8856	3456	216	216	574223	566663	7418055000	582560
Pennsylvania	32214	12272	766	595	1701007	1701886	32857966000	2037315
Rhode Island	2728	992	62	40	141733	140353	2549378000	188204
South Carolina	4185	1488	93	89	763011	762977	9436787000	860867
South Dakota	6900	2400	150	150	134045	133668	1455737000	196644
Tennessee	6674	2272	142	142	999260	983814	9586098000	1096182
Texas	59088	19696	1231	1046	5299681	5299615	60776728000	6194317
Utah	7448	2432	152	41	647599	647599	5415569000	419642
Vermont	15900	5088	318	316	86335	84124	2133017000	111891
Virginia	6783	2128	132	133	1283491	1283449	16259274000	1058146
Washington	16271	4912	307	304	1084025	1084687	14972163000	1098332
West Virginia	3510	1040	57	63	277436	277436	3420589000	360283
Wisconsin	23485	6832	427	426	857794	857612	11698660000	782647

```
Wyoming   2688   768   48   48   94511   94511   2044669000   123012(
```

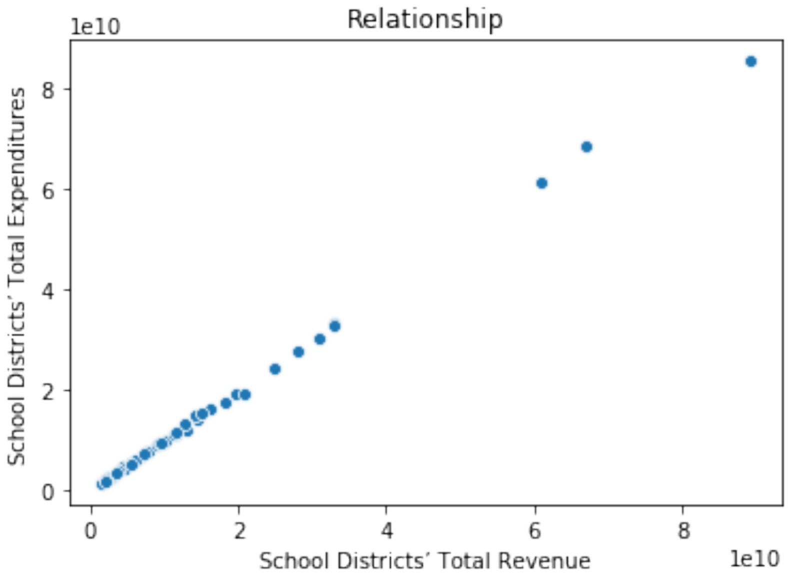
51 rows × 132 columns

In [7]:

```
plt.tight_layout()
sns.scatterplot(x='TOTALREV',y='TOTALEXP',data=school_districts_state)
plt.ylabel('School Districts’ Total Expenditures')
plt.xlabel('School Districts’ Total Revenue')
plt.title('Relationship')
```

Out[7]:

```
Text(0.5, 1.0, 'Relationship')
```



Based on the above plot, we could konw that there is a obvious positive linear relationship between school districts’ total revenue and expenditures

Which states have the most debt per student?

In [8]:

```
school_districts_state['debt_pstu'] = (school_districts_state['_41F'] + school_districts_state[
```

In [9]:

```
school_districts_state.sort_values(by='debt_pstu',ascending=False)['debt_pstu']
```

Out[9]:

```
STNAME
South Carolina    18375.689210
Minnesota         15607.783577
Texas             14836.154289
Pennsylvania      14489.936255
Michigan          12354.998688
Oregon            12017.822344
New York          11487.475897
Kansas            11238.315390
```

```
Washington              10514.506584
Alaska                  10360.492765
Illinois                10277.136527
California              10176.304050
Indiana                 10076.635351
Nebraska                 8746.675330
Arkansas                 8672.959685
Kentucky                 8544.569081
District of Columbia     8260.418299
Nevada                   8160.241013
Alabama                  8040.434855
Colorado                 7992.523200
Missouri                 7777.783470
Iowa                     7423.564359
Ohio                     7309.893767
South Dakota             6914.260137
North Dakota             6779.579025
Virginia                 6549.746745
New Mexico               6473.475766
Rhode Island             6383.432228
Wisconsin                6198.518525
Tennessee                5817.083642
Montana                  5729.326416
Louisiana                5636.000240
Massachusetts            5585.114295
Florida                  5277.146046
Maryland                 5039.002680
Maine                    5024.046195
New Jersey               4996.889518
New Hampshire            4944.804400
Idaho                    4935.572887
Utah                     4870.878429
North Carolina           4860.249066
Connecticut              4843.035860
Delaware                 4639.123115
Arizona                  4595.944182
Mississippi              3342.946457
Oklahoma                 3147.769895
Vermont                  2987.733828
Georgia                  2587.274379
West Virginia            1246.413587
Wyoming                   664.578726
Hawaii                      0.000000
Name: debt_pstu, dtype: float64
```

Based on the above data, we konw that South Carolina had the most debt per student

Problem 3

In [10]:
```
district_level_performance= pd.read_csv('math-achievement-lea-sy2015-16.csv')
```

ALL_MTH00PCTPROF_1516 is a blurred metric and I will use this as my performance matric

In [11]:
```
district_level_performance.groupby('ALL_MTH00PCTPROF_1516').count().reset_index()['ALL
```

Out[11]:
```
0         10
1      10-14
2         11
3      11-19
4         12
       ...
132     LE10
133     LE20
134      LE5
135     LT50
136       PS
Name: ALL_MTH00PCTPROF_1516, Length: 137, dtype: object
```

In [12]:
```
type(district_level_performance['ALL_MTH00PCTPROF_1516'][0])
```

Out[12]:
```
str
```

In [13]:

```
def process_blur(data):
    for i in range(len(data)):
        if 'LT' in data[i]:
            temp = int(data[i][2:])
            data[i] = np.mean([0,temp-1]) # if it is LT (less than), take mean of 0 and that num
        elif 'LE' in data[i]:
            temp = int(data[i][2:])
            data[i] = np.mean([0,temp]) # if it is LE (less than or equal to), take mean of 0 and
        elif 'GT' in data[i]:
            temp = int(data[i][2:])
            data[i] = np.mean([temp+1,100]) # if it is GT (greater than), take mean of 100 and
        elif 'GE' in data[i]:
            temp = int(data[i][2:])
            data[i] = np.mean([temp,100]) # if it is GE (less than), take mean of 100 and that n
        elif 'PS' in data[i]:
            data[i] = 'temp' # if it is suppressed to protect student privacy, wait for input
        elif '-' in data[i]:
            temp = data[i].split('-')
            data[i] = np.mean([int(temp[0]),int(temp[1])]) # if it is a range, take mean of start n
        else:
            data[i] = int(data[i])
    input_avg =np.mean(data[data!='temp'])
    for j in range(len(data)):
        if data[j]=='temp':
            data[j] = input_avg
    return data
```

In [14]:

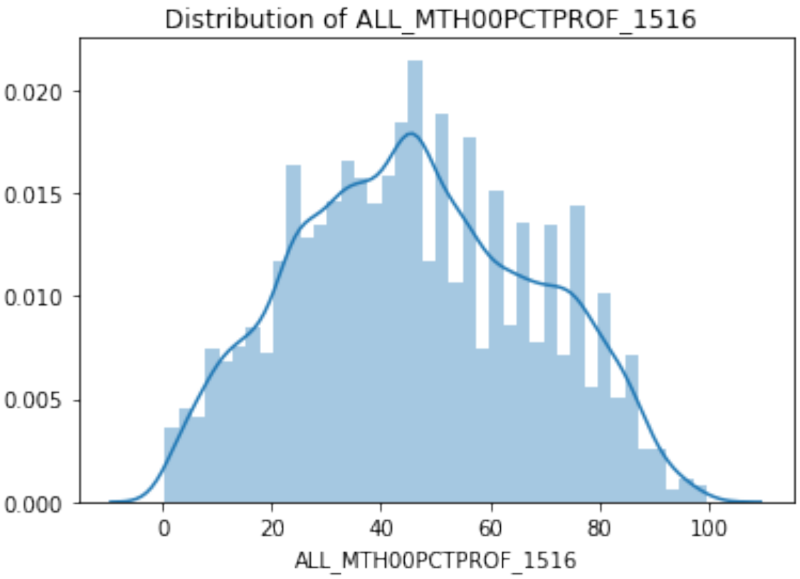
```
district_level_performance['ALL_MTH00PCTPROF_1516'] = process_blur(district_level_perfo
```

In [15]:

```
sns.distplot(district_level_performance['ALL_MTH00PCTPROF_1516'])
plt.title('Distribution of ALL_MTH00PCTPROF_1516 ')
```

Out[15]:

```
Text(0.5, 1.0, 'Distribution of ALL_MTH00PCTPROF_1516 ')
```



Problem 4

You are tasked with cutting 15% of the U.S. federal budget currently being spent on funding school districts. How much money is this?

In [16]:

```
cut_money = 0.15*federal_fund_state['TFEDREV'].sum()
cut_money
```

Out[16]:

```
8340411300.0
```

First Round(National Wide 10%)

In [17]:

```
school =district_level_fiscal[district_level_fiscal['TFEDREV']>0]
first_round_cut= 0.1*school['TFEDREV'].sum()
first_round_cut_list = school
first_round_cut_list['1_TFEDREV'] = 0.1*first_round_cut_list['TFEDREV']
first_round_cut_list['TFEDREV_after1'] = 0.9*first_round_cut_list['TFEDREV']
```

Second Round(Performance Focused)

In [18]:
```
threshold = np.percentile(district_level_performance['ALL_MTH00PCTPROF_1516'],75)
threshold
```

Out[18]:
```
62.0
```

In [19]:
```
cut_off_list = district_level_performance[district_level_performance['ALL_MTH00PCTPROF_
```

In [20]:
```
second_round_cut_list = pd.merge(cut_off_list,first_round_cut_list, on='LEAID',how='left')
total_have =second_round_cut_list['TFEDREV_after1'].sum()
percent = 100*(cut_money-first_round_cut)/total_have
percent # cut off percent of target school
```

Out[20]:
```
17.488760817082923
```

In [21]:
```
second_round_cut_list['2_TFEDREV'] = round((percent/100)*second_round_cut_list['TFEDR
```

In [22]:
```
final_cut_off = pd.merge(first_round_cut_list,second_round_cut_list,on='LEAID',how='left')
final_cut_off
```

Out[22]:

	LEAID	CENSUSID_x	FIPST	CONUM_x	CSA_x	CBSA_x	NAME_x	STNAME_x	S
0	100005	01504840100000	1	01095	290	10700	Albertville City	Alabama	
1	100006	01504800100000	1	01095	290	10700	Marshall County	Alabama	
2	100007	01503740100000	1	01073	142	13820	Hoover City	Alabama	
3	100008	01504530100000	1	01089	290	26620	Madison City	Alabama	
4	100011	01503710100000	1	01073	142	13820	Leeds City	Alabama	
...	...	...	...	...	...	...	...	...	
16534	5605762	51501900200000	56	56037	N	40540	Sweetwater County School District #2	Wyoming	
16535	5605820	51502200300000	56	56043	N	N	Washakie County School District #2	Wyoming	
16536	5605830	51502000200000	56	56039	N	27220	Teton County School District #1	Wyoming	
16537	5606090	51502300200000	56	56045	N	N	Weston County School District #7	Wyoming	
16538	5606240	51502200400000	56	56043	N	N	Washakie County School District #1	Wyoming	

16539 rows × 752 columns

In [23]:
```
final_cut_off.fillna({'1_TFEDREV_x':0,'2_TFEDREV':0},inplace=True)
final_cut_off['Amount of Funding Cut'] = final_cut_off['1_TFEDREV_x']+final_cut_off['2_TFE
```

In [24]:
```
final_cut_off['Amount of Funding Cut'].sum()
```

Out[24]:
```
8340411301.700001
```

In [25]:
```
final_cut_off[['LEAID','Amount of Funding Cut']]
```

Out[25]:
```
         LEAID   Amount of Funding Cut
    0   100005               1873091.4
    1   100006               1994068.9
    2   100007                608800.0
    3   100008                500700.0
    4   100011                391761.0
  ...      ...                     ...
16534  5605762                533073.0
16535  5605820                 66408.9
16536  5605830                476960.1
16537  5606090                 56370.3
16538  5606240                410293.8

16539 rows × 2 columns
```

Problem 5

If we are required to cut 15% of the U.S. federal budget currently being spent on funding school districts, I suggest conduct this by two rounds cutting off. The first round will be a national wide one, which cut off 10% of federal budget currently being spent on funding of every school districts. This will not make sure this plan highly biased, which is fair for the school districts in the second round list. The second round will focus on school districts that have undesirable performance. I use ALL_MTH00PCTPROF_1516 as metric and cut off the founds for those whose performance is below the 75 percentage. This is reasonable, because punishment mechanism should be considered and it make this plan fair to the the school districts that perform well. Then, I calculate the average cuf off percentage for the second round, which is around 17.5%. The last step is adding the amount of cutting in two rounds.