Pandas Project

November 21, 2021

[1]:
```
import pandas as pd
import numpy as np

school_data_to_load = "schools_complete.csv"
student_data_to_load = "students_complete.csv"

school_data = pd.read_csv(school_data_to_load)
student_data = pd.read_csv(student_data_to_load)

school_data_complete = pd.merge(student_data, school_data, how="left",␣
↪on=["school_name", "school_name"])
```

[2]:
```
schools = school_data_complete["school_name"].unique()
len(schools)
```

[2]: 15

[3]:
```
students = len(student_data["student_name"])
students
```

[3]: 39170

[4]:
```
totalbudget = school_data["budget"].sum()
totalbudget
```

[4]: 24649428

[5]:
```
mathave = student_data["math_score"].mean()
mathave
```

[5]: 78.98537145774827

[6]:
```
readdave = student_data["reading_score"].mean()
readdave
```

[6]: 81.87784018381414

[7]:
```
overallave = (mathave + readdave)/2
overallave
```

[7]:
```
80.43160582078121
```

[8]:
```
passmath = len(student_data[student_data["math_score"] >= 70])
percpassmath = (passmath/ students)*100
percpassmath
```

[8]:
```
74.9808526933878
```

[9]:
```
passread = len(student_data[student_data["reading_score"] >=70])
percpassread = (passread/students)*100
percpassread
```

[9]:
```
85.80546336482001
```

[10]:
```
summary = [{'Total Schools': len(schools),  'Total Students': students, 'Total␣
 ↪Budget': totalbudget, 'Avg. Math Score': mathave, 'Avg. Reading Score':␣
 ↪readdave, '% Passing Math': percpassmath, '% Passing Reading': percpassread,␣
 ↪'% Overall Passing Rate': overallave}]
summary = pd.DataFrame(summary)
summary = summary[['Total Schools', 'Total Students', 'Total Budget', 'Avg.␣
 ↪Math Score', 'Avg. Reading Score', '% Passing Math', '% Passing Reading', '%␣
 ↪Overall Passing Rate']]
summary
```

[10]:

	Total Schools	Total Students	Total Budget	Avg. Math Score	Avg. Reading Score	% Passing Math	% Passing Reading	% Overall Passing Rate
0	15	39170	24649428	78.985371	81.87784	74.980853	85.805463	80.431606

[11]:
```
SchoolTypes = school_data.set_index(["school_name"])['type']
```

[12]:
```
studentsgroup = school_data_complete["school_name"].value_counts()
studentsgroup
```

[12]:
```
Bailey High School       4976
Johnson High School      4761
Hernandez High School    4635
Rodriguez High School    3999
Figueroa High School     2949
Huang High School        2917
Ford High School         2739
```

```
Wilson High School       2283
Cabrera High School      1858
Wright High School       1800
Shelton High School      1761
Thomas High School       1635
Griffin High School      1468
Pena High School          962
Holden High School        427
Name: school_name, dtype: int64
```

[13]:
```
budgetgroup = school_data_complete.groupby(["school_name"]).mean()["budget"]
budgetgroup
```

[13]:
```
school_name
Bailey High School       3124928.0
Cabrera High School      1081356.0
Figueroa High School     1884411.0
Ford High School         1763916.0
Griffin High School       917500.0
Hernandez High School    3022020.0
Holden High School        248087.0
Huang High School        1910635.0
Johnson High School      3094650.0
Pena High School          585858.0
Rodriguez High School    2547363.0
Shelton High School      1056600.0
Thomas High School       1043130.0
Wilson High School       1319574.0
Wright High School       1049400.0
Name: budget, dtype: float64
```

[14]:
```
perstubud = budgetgroup / studentsgroup
perstubud
```

[14]:
```
Bailey High School       628.0
Cabrera High School      582.0
Figueroa High School     639.0
Ford High School         644.0
Griffin High School      625.0
Hernandez High School    652.0
Holden High School       581.0
Huang High School        655.0
Johnson High School      650.0
Pena High School         609.0
Rodriguez High School    637.0
Shelton High School      600.0
Thomas High School       638.0
```

```
Wilson High School       578.0
Wright High School       583.0
dtype: float64
```

[15]:
```
avgmathgroup = school_data_complete.groupby(["school_name"]).
↪mean()["math_score"]
avgmathgroup
```

[15]:
```
school_name
Bailey High School       77.048432
Cabrera High School      83.061895
Figueroa High School     76.711767
Ford High School         77.102592
Griffin High School      83.351499
Hernandez High School    77.289752
Holden High School       83.803279
Huang High School        76.629414
Johnson High School      77.072464
Pena High School         83.839917
Rodriguez High School    76.842711
Shelton High School      83.359455
Thomas High School       83.418349
Wilson High School       83.274201
Wright High School       83.682222
Name: math_score, dtype: float64
```

[16]:
```
avgreadgroup = school_data_complete.groupby(["school_name"]).
↪mean()["reading_score"]
avgreadgroup
```

[16]:
```
school_name
Bailey High School       81.033963
Cabrera High School      83.975780
Figueroa High School     81.158020
Ford High School         80.746258
Griffin High School      83.816757
Hernandez High School    80.934412
Holden High School       83.814988
Huang High School        81.182722
Johnson High School      80.966394
Pena High School         84.044699
Rodriguez High School    80.744686
Shelton High School      83.725724
Thomas High School       83.848930
Wilson High School       83.989488
Wright High School       83.955000
Name: reading_score, dtype: float64
```

[17]:
```
mathpassgroup = school_data_complete[(school_data_complete["math_score"] >= 70)]
percmath = mathpassgroup.groupby(["school_name"]).count()["student_name"]/
 ↪studentsgroup*100
percmath
```

[17]:
```
Bailey High School       66.680064
Cabrera High School      94.133477
Figueroa High School     65.988471
Ford High School         68.309602
Griffin High School      93.392371
Hernandez High School    66.752967
Holden High School       92.505855
Huang High School        65.683922
Johnson High School      66.057551
Pena High School         94.594595
Rodriguez High School    66.366592
Shelton High School      93.867121
Thomas High School       93.272171
Wilson High School       93.867718
Wright High School       93.333333
dtype: float64
```

[18]:
```
readpassgroup = school_data_complete[(school_data_complete["reading_score"] >=␣
 ↪70)]
percread = readpassgroup.groupby(["school_name"]).count()["student_name"]/
 ↪studentsgroup*100
percread
```

[18]:
```
Bailey High School       81.933280
Cabrera High School      97.039828
Figueroa High School     80.739234
Ford High School         79.299014
Griffin High School      97.138965
Hernandez High School    80.862999
Holden High School       96.252927
Huang High School        81.316421
Johnson High School      81.222432
Pena High School         95.945946
Rodriguez High School    80.220055
Shelton High School      95.854628
Thomas High School       97.308869
Wilson High School       96.539641
Wright High School       96.611111
dtype: float64
```

[19]:
```
overallpassgroup = ((percmath + percread)/2)
overallpassgroup
```

[19]:
```
Bailey High School       74.306672
Cabrera High School      95.586652
Figueroa High School     73.363852
Ford High School         73.804308
Griffin High School      95.265668
Hernandez High School    73.807983
Holden High School       94.379391
Huang High School        73.500171
Johnson High School      73.639992
Pena High School         95.270270
Rodriguez High School    73.293323
Shelton High School      94.860875
Thomas High School       95.290520
Wilson High School       95.203679
Wright High School       94.972222
dtype: float64
```

[20]:
```
summary = pd.DataFrame({"Total Students": studentsgroup,
                        "Type": SchoolTypes,
                        "Total School Budget": budgetgroup,
                        "Per Student Budget": perstubud,
                        "Average Math Score": avgmathgroup,
                        "Average Reading Score": avgreadgroup,
                        "% Passing Math": percmath,
                        "% Passing Reading": percread,
                        "Overall Passing Rate": overallpassgroup })
summary = summary[["Total Students", "Type", "Total School Budget", "Per␣
↪Student Budget", "Average Math Score", "Average Reading Score", "% Passing␣
↪Math", "% Passing Reading", "Overall Passing Rate"]]
summary
```

[20]:

	Total Students	Type	Total School Budget	\
Bailey High School	4976	District	3124928.0	
Cabrera High School	1858	Charter	1081356.0	
Figueroa High School	2949	District	1884411.0	
Ford High School	2739	District	1763916.0	
Griffin High School	1468	Charter	917500.0	
Hernandez High School	4635	District	3022020.0	
Holden High School	427	Charter	248087.0	
Huang High School	2917	District	1910635.0	
Johnson High School	4761	District	3094650.0	
Pena High School	962	Charter	585858.0	
Rodriguez High School	3999	District	2547363.0	
Shelton High School	1761	Charter	1056600.0	
Thomas High School	1635	Charter	1043130.0	
Wilson High School	2283	Charter	1319574.0	
Wright High School	1800	Charter	1049400.0	

	Per Student Budget	Average Math Score	\
Bailey High School	628.0	77.048432	
Cabrera High School	582.0	83.061895	
Figueroa High School	639.0	76.711767	
Ford High School	644.0	77.102592	
Griffin High School	625.0	83.351499	
Hernandez High School	652.0	77.289752	
Holden High School	581.0	83.803279	
Huang High School	655.0	76.629414	
Johnson High School	650.0	77.072464	
Pena High School	609.0	83.839917	
Rodriguez High School	637.0	76.842711	
Shelton High School	600.0	83.359455	
Thomas High School	638.0	83.418349	
Wilson High School	578.0	83.274201	
Wright High School	583.0	83.682222	

	Average Reading Score	% Passing Math	\
Bailey High School	81.033963	66.680064	
Cabrera High School	83.975780	94.133477	
Figueroa High School	81.158020	65.988471	
Ford High School	80.746258	68.309602	
Griffin High School	83.816757	93.392371	
Hernandez High School	80.934412	66.752967	
Holden High School	83.814988	92.505855	
Huang High School	81.182722	65.683922	
Johnson High School	80.966394	66.057551	
Pena High School	84.044699	94.594595	
Rodriguez High School	80.744686	66.366592	
Shelton High School	83.725724	93.867121	
Thomas High School	83.848930	93.272171	
Wilson High School	83.989488	93.867718	
Wright High School	83.955000	93.333333	

	% Passing Reading	Overall Passing Rate
Bailey High School	81.933280	74.306672
Cabrera High School	97.039828	95.586652
Figueroa High School	80.739234	73.363852
Ford High School	79.299014	73.804308
Griffin High School	97.138965	95.265668
Hernandez High School	80.862999	73.807983
Holden High School	96.252927	94.379391
Huang High School	81.316421	73.500171
Johnson High School	81.222432	73.639992
Pena High School	95.945946	95.270270
Rodriguez High School	80.220055	73.293323

```
Shelton High School            95.854628             94.860875
Thomas High School             97.308869             95.290520
Wilson High School             96.539641             95.203679
Wright High School             96.611111             94.972222
```

[21]:
```
topschools = summary.sort_values("Overall Passing Rate", ascending=False)
topschools.head()
```

[21]:
```
                     Total Students     Type  Total School Budget  \
Cabrera High School            1858  Charter            1081356.0
Thomas High School             1635  Charter            1043130.0
Pena High School                962  Charter             585858.0
Griffin High School            1468  Charter             917500.0
Wilson High School             2283  Charter            1319574.0

                     Per Student Budget  Average Math Score  \
Cabrera High School                582.0           83.061895
Thomas High School                 638.0           83.418349
Pena High School                   609.0           83.839917
Griffin High School                625.0           83.351499
Wilson High School                 578.0           83.274201

                     Average Reading Score  % Passing Math  % Passing Reading  \
Cabrera High School              83.975780       94.133477          97.039828
Thomas High School               83.848930       93.272171          97.308869
Pena High School                 84.044699       94.594595          95.945946
Griffin High School              83.816757       93.392371          97.138965
Wilson High School               83.989488       93.867718          96.539641

                     Overall Passing Rate
Cabrera High School             95.586652
Thomas High School              95.290520
Pena High School                95.270270
Griffin High School             95.265668
Wilson High School              95.203679
```

[22]:
```
bottomschools = summary.sort_values("Overall Passing Rate")
bottomschools.head()
```

[22]:
```
                       Total Students      Type  Total School Budget  \
Rodriguez High School            3999  District            2547363.0
Figueroa High School             2949  District            1884411.0
Huang High School                2917  District            1910635.0
Johnson High School              4761  District            3094650.0
Ford High School                 2739  District            1763916.0

                       Per Student Budget  Average Math Score  \
```

```
Rodriguez High School                637.0           76.842711
Figueroa High School                 639.0           76.711767
Huang High School                    655.0           76.629414
Johnson High School                  650.0           77.072464
Ford High School                     644.0           77.102592

                      Average Reading Score  % Passing Math  \
Rodriguez High School              80.744686       66.366592
Figueroa High School               81.158020       65.988471
Huang High School                  81.182722       65.683922
Johnson High School                80.966394       66.057551
Ford High School                   80.746258       68.309602

                      % Passing Reading  Overall Passing Rate
Rodriguez High School          80.220055             73.293323
Figueroa High School           80.739234             73.363852
Huang High School              81.316421             73.500171
Johnson High School            81.222432             73.639992
Ford High School               79.299014             73.804308
```

[23]:
```
fresh2 = school_data_complete[school_data_complete["grade"] == "9th"].
 ↪groupby("school_name").mean()["math_score"]
soph2 = school_data_complete[school_data_complete["grade"] == "10th"].
 ↪groupby("school_name").mean()["math_score"]
jun2 = school_data_complete[school_data_complete["grade"] == "11th"].
 ↪groupby("school_name").mean()["math_score"]
sen2 = school_data_complete[school_data_complete["grade"] == "12th"].
 ↪groupby("school_name").mean()["math_score"]

SummMath = pd.DataFrame({"9th Grade": fresh2, "10th Grade": soph2, "11th Grade":
 ↪ jun2, "12th Grade": sen2})
SummMath
```

[23]:
```
                     9th Grade  10th Grade  11th Grade  12th Grade
school_name
Bailey High School     77.083676   76.996772   77.515588   76.492218
Cabrera High School    83.094697   83.154506   82.765560   83.277487
Figueroa High School   76.403037   76.539974   76.884344   77.151369
Ford High School       77.361345   77.672316   76.918058   76.179963
Griffin High School    82.044010   84.229064   83.842105   83.356164
Hernandez High School  77.438495   77.337408   77.136029   77.186567
Holden High School     83.787402   83.429825   85.000000   82.855422
Huang High School      77.027251   75.908735   76.446602   77.225641
Johnson High School    77.187857   76.691117   77.491653   76.863248
Pena High School       83.625455   83.372000   84.328125   84.121547
Rodriguez High School  76.859966   76.612500   76.395626   77.690748
Shelton High School    83.420755   82.917411   83.383495   83.778976
```

```
Thomas High School     83.590022   83.087886   83.498795   83.497041
Wilson High School     83.085578   83.724422   83.195326   83.035794
Wright High School     83.264706   84.010288   83.836782   83.644986
```

[24]:
```
fresh = school_data_complete[school_data_complete["grade"] == "9th"].
 ↪groupby("school_name").mean()["reading_score"]
soph = school_data_complete[school_data_complete["grade"] == "10th"].
 ↪groupby("school_name").mean()["reading_score"]
jun = school_data_complete[school_data_complete["grade"] == "11th"].
 ↪groupby("school_name").mean()["reading_score"]
sen = school_data_complete[school_data_complete["grade"] == "12th"].
 ↪groupby("school_name").mean()["reading_score"]

SummRead = pd.DataFrame({"9th Grade": fresh, "10th Grade": soph, "11th Grade":␣
 ↪jun, "12th Grade": sen})
SummRead
```

[24]:
```
                     9th Grade  10th Grade  11th Grade  12th Grade
school_name
Bailey High School     81.303155   80.907183   80.945643   80.912451
Cabrera High School    83.676136   84.253219   83.788382   84.287958
Figueroa High School   81.198598   81.408912   80.640339   81.384863
Ford High School       80.632653   81.262712   80.403642   80.662338
Griffin High School    83.369193   83.706897   84.288089   84.013699
Hernandez High School  80.866860   80.660147   81.396140   80.857143
Holden High School     83.677165   83.324561   83.815534   84.698795
Huang High School      81.290284   81.512386   81.417476   80.305983
Johnson High School    81.260714   80.773431   80.616027   81.227564
Pena High School       83.807273   83.612000   84.335938   84.591160
Rodriguez High School  80.993127   80.629808   80.864811   80.376426
Shelton High School    84.122642   83.441964   84.373786   82.781671
Thomas High School     83.728850   84.254157   83.585542   83.831361
Wilson High School     83.939778   84.021452   83.764608   84.317673
Wright High School     83.833333   83.812757   84.156322   84.073171
```

[25]:
```
spending_bins = [0, 585, 615, 645, 675]
group_names = ["<$585", "$585-615", "$615-645", "$645-675"]
summary["Spending Ranges"] = pd.cut(perstubud, spending_bins,␣
 ↪labels=group_names)
mathspend = summary.groupby(["Spending Ranges"]).mean()["Average Math Score"]
readspend = summary.groupby(["Spending Ranges"]).mean()["Average Reading Score"]
passmathspend = summary.groupby(["Spending Ranges"]).mean()["% Passing Math"]
passreadspend = summary.groupby(["Spending Ranges"]).mean()["% Passing Reading"]
passoverallspend = (passmathspend + passreadspend)/2
spendsumm = summary[["Average Math Score","Average Reading Score","% Passing␣
 ↪Math","% Passing Reading", "Overall Passing Rate"]]
spendsumm = pd.DataFrame({"Average Math Score"     :mathspend,
```

```
                          "Average Reading Score"     :readspend,
                          "% Passing Math"     :passmathspend,
                          "% Passing Reading"     :passreadspend,
                          "Overall Passing Rate"     :passoverallspend})
spendsumm.groupby("Spending Ranges").head(15)
```

[25]:

Spending Ranges	Average Math Score	Average Reading Score	% Passing Math	% Passing Reading	Overall Passing Rate
<$585	83.455399	83.933814	93.460096	96.610877	95.035486
$585-615	83.599686	83.885211	94.230858	95.900287	95.065572
$615-645	79.079225	81.891436	75.668212	86.106569	80.887391
$645-675	76.997210	81.027843	66.164813	81.133951	73.649382

[26]:
```
size_bins = [0, 1000, 2000, 5000]
group_names2 = ["Small (<1000)", "Medium (1000-2000)", "Large (2000-5000)"]
summary["Size Ranges"] = pd.cut(studentsgroup, size_bins, labels=group_names2)
sizemath = summary.groupby(["Size Ranges"]).mean()["Average Math Score"]
sizeread = summary.groupby(["Size Ranges"]).mean()["Average Reading Score"]
sizepassmath =summary.groupby(["Size Ranges"]).mean()["% Passing Math"]
sizepassread =summary.groupby(["Size Ranges"]).mean()["% Passing Reading"]
sizeoverallpass = (sizepassmath + sizepassread)/2
sizesumm = summary[["Average Math Score","Average Reading Score","% Passing␣
↪Math","% Passing Reading", "Overall Passing Rate"]]
sizesumm = pd.DataFrame({"Average Math Score"     :sizemath,
                          "Average Reading Score"     :sizeread,
                          "% Passing Math"     :sizepassmath,
                          "% Passing Reading"     :sizepassread,
                          "Overall Passing Rate"     :sizeoverallpass})
sizesumm.groupby("Size Ranges").head(15)
```

[26]:

Size Ranges	Average Math Score	Average Reading Score	% Passing Math	% Passing Reading	Overall Passing Rate
Small (<1000)	83.821598	83.929843	93.550225	96.099437	94.824831
Medium (1000-2000)	83.374684	83.864438	93.599695	96.790680	95.195187
Large (2000-5000)	77.746417	81.344493	69.963361		

```
Large (2000-5000)          82.766634             76.364998
```

[27]:
```
typemath = summary.groupby(["Type"]).mean()["Average Math Score"]
typeread = summary.groupby(["Type"]).mean()["Average Reading Score"]
typepassmath =summary.groupby(["Type"]).mean()["% Passing Math"]
typepassread =summary.groupby(["Type"]).mean()["% Passing Reading"]
typoverallpass = (typepassmath + typepassread)/2
typesumm = pd.DataFrame({"Average Math Score"     :typemath,
                         "Average Reading Score"     :typeread,
                         "% Passing Math"          :typepassmath,
                         "% Passing Reading"           :typepassread,
                         "Overall Passing Rate"          :typoverallpass})
typesumm = typesumm[["Average Math Score", "Average Reading Score", "% Passing␣
↪Math", "% Passing Reading", "Overall Passing Rate"]]
typesumm.groupby("Type").head()
```

[27]:
```
          Average Math Score  Average Reading Score  % Passing Math  \
Type
Charter            83.473852              83.896421       93.620830
District           76.956733              80.966636       66.548453

          % Passing Reading  Overall Passing Rate
Type
Charter           96.586489             95.103660
District          80.799062             73.673757
```

1 PyCity Schools Analysis

- As a whole, schools with higher budgets, did not yield better test results. By contrast, schools with higher spending per student actually (\$645 - \$675) underperformed compared to schools with smaller budgets (\$585 per student).
- As a whole, smaller and medium sized schools dramatically out-performed large sized schools on passing math performances (89-91% passing vs 67%).
- As a whole, charter schools out-performed the public district schools across all metrics. However, more analysis will be required to glean if the effect is due to school practices or the fact that charter schools tend to serve smaller student populations per school.