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.

Note

- Instructions have been included for each segment. You do not have to follow them exactly, but they are included to help you think through the steps.

In [1]:
```
# Dependencies and Setup
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# File to Load (Remember to Change These)
school_data_to_load = "Resources/schools_complete.csv"
student_data_to_load = "Resources/students_complete.csv"

# Read School and Student Data File and store into Pandas Data Frames
school_data = pd.read_csv(school_data_to_load)
student_data = pd.read_csv(student_data_to_load)

# Combine the data into a single dataset
school_data_complete = pd.merge(student_data, school_data, how="left", on=["school_name", "school_name"])
```

District Summary

- Calculate the total number of schools
- Calculate the total number of students
- Calculate the total budget
- Calculate the average math score
- Calculate the average reading score
- Calculate the overall passing rate (overall average score), i.e. (avg. math score + avg. reading score)/2
- Calculate the percentage of students with a passing math score (70 or greater)
- Calculate the percentage of students with a passing reading score (70 or greater)
- Create a dataframe to hold the above results
- Optional: give the displayed data cleaner formatting

In [2]:
```
s_df = school_data_complete.copy()
tot_sch = s_df['School ID'].nunique()
s_df.head()
#  print(tot_sch)
tot_stu = s_df['Student ID'].nunique()
tot_stu
#  print('{:,}'.format(tot_stu))
tot_stu_str = '{:,}'.format(tot_stu)
total_budget = school_data['budget'].sum()
#print('{:,}'.format(total_budget))
total_budget = '{:,}'.format(total_budget)
#  print(total_budget)
```

In [3]:
```
# '{:,}'.format(school_data['budget'].sum())
# print('{:,}'.format(school_data['budget'].sum()))
# '{:,.6f}'.format(student_data['math_score'].mean())

avg_math = student_data['math_score'].mean()
avg_reading = student_data['reading_score'].mean()
#  print('{:,.6f}'.format(avg_math))
#  print('{:,.6f}'.format(avg_reading))

pass_math_df = student_data[student_data['math_score'] >= 70]
pct_math = (pass_math_df['Student ID'].count() * 100)/tot_stu
#  print('{:,.6f}%'.format(pct_math))

pass_read_df = student_data[student_data['reading_score'] >= 70]
pct_read = (pass_read_df['Student ID'].count() * 100)/tot_stu
#  print('{:,.6f}%'.format(pct_read))

overall_pass_pct = (avg_math + avg_reading)/2
#  print('{:,.6f}%'.format(overall_pass_pct))

summary_df = pd.DataFrame(
    {'Total Schools': [tot_sch],
    'Total Students': [tot_stu_str],
    'Total Budget': [total_budget],
    'Average Math Score': [avg_math],
    'Average Reading Score': [avg_reading],
    '% Passing Math': [pct_math],
    '% Passing Reading': [pct_read],
    '% Overall Passing Rate': [overall_pass_pct]})
summary_df
```

Out[3]:

	Total Schools	Total Students	Total Budget	Average Math Score	Average Reading Score	% Passing Math	% Passing Reading	% Overall Passing Rate
0	15	39,170	24,649,428	78.985371	81.87784	74.980853	85.805463	80.431606

School Summary

- Create an overview table that summarizes key metrics about each school, including:
  - School Name
  - School Type
  - Total Students
  - Total School Budget
  - Per Student Budget
  - Average Math Score
  - Average Reading Score
  - % Passing Math
  - % Passing Reading
  - Overall Passing Rate (Average of the above two)
- Create a dataframe to hold the above results

Top Performing Schools (By Passing Rate)

- Sort and display the top five schools in overall passing rate

In [4]:
```
summary_df = pd.DataFrame()
sch_type = s_df.groupby(['school_name'])['type'].max()
summary_df['School Type']= sch_type
stu_cnt = s_df.groupby(['school_name'])['Student ID'].count()
summary_df['Total Students']= stu_cnt
sch_budget = s_df.groupby(['school_name'])['budget'].max()
# Alternative way
#summary_df['Total School Budget']= sch_budget.apply(lambda x: "${:,.2f}".format(x))
summary_df['Total School Budget']= sch_budget.map("${:,.2f}".format)
summary_df['Per Student Budget']= (sch_budget/stu_cnt).apply(lambda x: "${:,.2f}".format(x))
avg_math = s_df.groupby(['school_name'])['math_score'].mean()
summary_df['Average Math Score']= avg_math
avg_reading = s_df.groupby(['school_name'])['reading_score'].mean()
summary_df['Average Reading Score']= avg_reading
pass_math = s_df[s_df['math_score'] >= 70]
pct_math = (pass_math.groupby(['school_name'])['Student ID'].count() * 100)/stu_cnt
summary_df['% Passing Math']= pct_math
# Alternate way to drop all indices that have math score < 70
#pct_math = s_df.groupby(['school_name', 'type', 'Student ID']).filter(lambda x: x['math_score'] >= 70)
pass_reading = s_df[s_df['reading_score'] >= 70]
pct_reading = (pass_reading.groupby(['school_name'])['Student ID'].count() * 100)/stu_cnt
summary_df['% Passing Reading']= pct_reading
overall_pass_pct = (pct_math + pct_reading)/2
summary_df['% Overall Passing Rate']= overall_pass_pct
summary_df.reset_index()
del summary_df.index.name
perf_df = summary_df.copy()
summary_df.sort_values('% Overall Passing Rate', ascending = False).head(5)
```

Out[4]:

	School Type	Total Students	Total School Budget	Per Student Budget	Average Math Score	Average Reading Score	% Passing Math	% Passing Reading	% Overall Passing Rate
Cabrera High School	Charter	1858	$1,081,356.00	$582.00	83.061895	83.975780	94.133477	97.039828	95.586652
Thomas High School	Charter	1635	$1,043,130.00	$638.00	83.418349	83.848930	93.272171	97.308869	95.290520
Pena High School	Charter	962	$585,858.00	$609.00	83.839917	84.044699	94.594595	95.945946	95.270270
Griffin High School	Charter	1468	$917,500.00	$625.00	83.351499	83.816757	93.392371	97.138965	95.265668
Wilson High School	Charter	2283	$1,319,574.00	$578.00	83.274201	83.989488	93.867718	96.539641	95.203679

Bottom Performing Schools (By Passing Rate)

- Sort and display the five worst-performing schools

In [5]:
```
summary_df.sort_values('% Overall Passing Rate', ascending = True).head(5)
```

Out[5]:

	School Type	Total Students	Total School Budget	Per Student Budget	Average Math Score	Average Reading Score	% Passing Math	% Passing Reading	% Overall Passing Rate
Rodriguez High School	District	3999	$2,547,363.00	$637.00	76.842711	80.744686	66.366592	80.220055	73.293323
Figueroa High School	District	2949	$1,884,411.00	$639.00	76.711767	81.158020	65.988471	80.739234	73.363852
Huang High School	District	2917	$1,910,635.00	$655.00	76.629414	81.182722	65.683922	81.316421	73.500171
Johnson High School	District	4761	$3,094,650.00	$650.00	77.072464	80.966394	66.057551	81.222432	73.639992
Ford High School	District	2739	$1,763,916.00	$644.00	77.102592	80.746258	68.309602	79.299014	73.804308

Math Scores by Grade

- Create a table that lists the average Reading Score for students of each grade level (9th, 10th, 11th, 12th) at each school.
  - Create a pandas series for each grade. Hint: use a conditional statement.
  - Group each series by school
  - Combine the series into a dataframe
  - Optional: give the displayed data cleaner formatting

In [6]:
```
summary_df = pd.DataFrame()
math_9_grade = s_df[s_df['grade'] == '9th']
math_9_avg = math_9_grade.groupby('school_name')['math_score'].mean()
math_9_avg.head()
summary_df['9th'] = math_9_avg
math_10_grade = s_df[s_df['grade'] == '10th']
math_10_avg = math_10_grade.groupby('school_name')['math_score'].mean()
math_10_avg.head()
summary_df['10th'] = math_10_avg
math_11_grade = s_df[s_df['grade'] == '11th']
math_11_avg = math_11_grade.groupby('school_name')['math_score'].mean()
math_11_avg.head()
summary_df['11th'] = math_11_avg
math_12_grade = s_df[s_df['grade'] == '12th']
math_12_avg = math_12_grade.groupby('school_name')['math_score'].mean()
math_12_avg.head()
summary_df['12th'] = math_12_avg
del summary_df.index.name
summary_df
```

Out[6]:

	9th	10th	11th	12th
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

Reading Score by Grade

- Perform the same operations as above for reading scores

In [7]:
```
summary_df = pd.DataFrame()
reading_9_grade = s_df[s_df['grade'] == '9th']
reading_9_avg = reading_9_grade.groupby('school_name')['reading_score'].mean()
reading_9_avg.head()
summary_df['9th'] = reading_9_avg
reading_10_grade = s_df[s_df['grade'] == '10th']
reading_10_avg = reading_10_grade.groupby('school_name')['reading_score'].mean()
reading_10_avg.head()
summary_df['10th'] = reading_10_avg
reading_11_grade = s_df[s_df['grade'] == '11th']
reading_11_avg = reading_11_grade.groupby('school_name')['reading_score'].mean()
reading_11_avg.head()
summary_df['11th'] = reading_11_avg
reading_12_grade = s_df[s_df['grade'] == '12th']
reading_12_avg = reading_12_grade.groupby('school_name')['reading_score'].mean()
reading_12_avg.head()
summary_df['12th'] = reading_12_avg
del summary_df.index.name
summary_df
```

Out[7]:

	9th	10th	11th	12th
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

Scores by School Spending

- Create a table that breaks down school performances based on average Spending Ranges (Per Student). Use 4 reasonable bins to group school spending. Include in the table each of the following:
  - Average Math Score
  - Average Reading Score
  - % Passing Math
  - % Passing Reading
  - Overall Passing Rate (Average of the above two)

In [8]:
```
# Sample bins. Feel free to create your own bins.
spending_bins = [0, 585, 615, 645, 675]
group_names = ["<$585", "$585-615", "$615-645", "$645-675"]
summary_df = pd.DataFrame()
perf_df['Per Student Budget'] = perf_df['Per Student Budget'].replace('\$|\.00', '', regex=True).astype('int32')
perf_df['Spending Ranges (Per Student)'] = pd.cut(perf_df['Per Student Budget'], spending_bins, labels=group_names)
summary_df['Average Math Score'] = perf_df.groupby(['Spending Ranges (Per Student)'])['Average Math Score'].mean()
summary_df['Average Reading Score'] = perf_df.groupby(['Spending Ranges (Per Student)'])['Average Reading Score'].mean()
summary_df['% Passing Math'] = perf_df.groupby(['Spending Ranges (Per Student)'])['% Passing Math'].mean()
summary_df['% Passing Reading'] = perf_df.groupby(['Spending Ranges (Per Student)'])['% Passing Reading'].mean()
summary_df['% Overall Passing Rate'] = perf_df.groupby(['Spending Ranges (Per Student)'])['% Overall Passing Rate'].mean()
summary_df.head()
```

Out[8]:

	Average Math Score	Average Reading Score	% Passing Math	% Passing Reading	% Overall Passing Rate
Spending Ranges (Per Student)					
<$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

Scores by School Size

- Perform the same operations as above, based on school size.

In [9]:
```
# Sample bins. Feel free to create your own bins.
size_bins = [0, 1000, 2000, 5000]
group_names = ["Small (<1000)", "Medium (1000-2000)", "Large (2000-5000)"]
summary_df = pd.DataFrame()
perf_df['School Size'] = pd.cut(perf_df['Total Students'], size_bins, labels=group_names)
summary_df['Average Math Score'] = perf_df.groupby(['School Size'])['Average Math Score'].mean()
summary_df['Average Reading Score'] = perf_df.groupby(['School Size'])['Average Reading Score'].mean()
summary_df['% Passing Math'] = perf_df.groupby(['School Size'])['% Passing Math'].mean()
summary_df['% Passing Reading'] = perf_df.groupby(['School Size'])['% Passing Reading'].mean()
summary_df['% Overall Passing Rate'] = perf_df.groupby(['School Size'])['% Overall Passing Rate'].mean()
summary_df.head()
```

Out[9]:

	Average Math Score	Average Reading Score	% Passing Math	% Passing Reading	% Overall Passing Rate
School Size					
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	82.766634	76.364998

Scores by School Type

- Perform the same operations as above, based on school type.

In [10]:
```
summary_df = pd.DataFrame()
summary_df['Average Math Score'] = perf_df.groupby(['School Type'])['Average Math Score'].mean()
summary_df['Average Reading Score'] = perf_df.groupby(['School Type'])['Average Reading Score'].mean()
summary_df['% Passing Math'] = perf_df.groupby(['School Type'])['% Passing Math'].mean()
summary_df['% Passing Reading'] = perf_df.groupby(['School Type'])['% Passing Reading'].mean()
summary_df['% Overall Passing Rate'] = perf_df.groupby(['School Type'])['% Overall Passing Rate'].mean()
summary_df.head()
```

Out[10]:

	Average Math Score	Average Reading Score	% Passing Math	% Passing Reading	% Overall Passing Rate
School Type					
Charter	83.473852	83.896421	93.620830	96.586489	95.103660
District	76.956733	80.966636	66.548453	80.799062	73.673757