PyCity School Analysis

Module 4 - Challenge

**OVERVIEW**

The PyCitySchools project is an in-depth analysis of school district data using the Pandas library in Python and Jupyter Notebook. Data from separate CSV files are read into Pandas and merged to create multiple different dataframes and analyzed as per instructions provided in the Module 4. The sample solution called PyCitySchools\_starter.ipynb located in the .zip file is used to review the desired format for this assignment.

**BACKGROUND**

You are the new Chief Data Scientist for your city's school district. In this capacity, you'll be helping the school board and mayor make strategic decisions regarding future school budgets and priorities.

**PURPOSE**

The purpose of this analysis is to utilize Pandas to analyze school data provided, analyze the district-wide standardized test results and to aggregate the data to showcase obvious trends in school performance.

**ANALYSIS**

* The budget is not responsible for better performance or higher budgets, did not yield better test results. Take for example schools with higher spending $645-675 per student actually underperformed compared to schools with smaller budgets $585 per student performed better.
* All top 5 performing schools were Charter schools while bottom performing were District schools
* As a whole, smaller and medium sized schools dramatically out-performed large sized schools on passing math performances (90-91% overall passing vs 52-53%).
* Except one exception, per student spending is higher in bottom performing schools than top performing. Therefore, spending per student is not related to school performance
* Small (<1000) and medium (1000-2000) size schools have much higher passing rate than Large (2000-5000) size school

**District Summary**

Perform the necessary calculations and then create a high-level snapshot of the district's key metrics in a DataFrame.

Include the following:

* Total number of unique schools
* Total students
* Total budget
* Average math score
* Average reading score
* % passing math (the percentage of students who passed math)
* % passing reading (the percentage of students who passed reading)
* % overall passing (the percentage of students who passed math AND reading)

**School Summary**

Perform the necessary calculations and then create a DataFrame that summarizes key metrics about each school.

Include the following:

* School name
* School type
* Total students
* Total school budget
* Per student budget
* Average math score
* Average reading score
* % passing math (the percentage of students who passed math)
* % passing reading (the percentage of students who passed reading)
* % overall passing (the percentage of students who passed math AND reading)

**Highest-Performing Schools (by % Overall Passing)**

Sort the schools by % Overall Passing in descending order and display the top 5 rows.

Save the results in a DataFrame called "top\_schools".

**Lowest-Performing Schools (by % Overall Passing)**

Sort the schools by % Overall Passing in ascending order and display the top 5 rows.

Save the results in a DataFrame called "bottom\_schools".

**Math Scores by Grade**

Perform the necessary calculations to create a DataFrame that lists the average math score for students of each grade level (9th, 10th, 11th, 12th) at each school.

**Reading Scores by Grade**

Create a DataFrame that lists the average reading score for students of each grade level (9th, 10th, 11th, 12th) at each school.

**Scores by School Spending**

Create a table that breaks down school performance based on average spending ranges (per student).

Use the code provided below to create four bins with reasonable cutoff values to group school spending.

spending\_bins = [0, 585, 630, 645, 680]

labels = ["<$585", "$585-630", "$630-645", "$645-680"]

Use pd.cut to categorize spending based on the bins.

Use the following code to then calculate mean scores per spending range.

spending\_math\_scores = school\_spending\_df.groupby(["Spending Ranges (Per Student)"])["Average Math Score"].mean()

spending\_reading\_scores = school\_spending\_df.groupby(["Spending Ranges (Per Student)"])["Average Reading Score"].mean()

spending\_passing\_math = school\_spending\_df.groupby(["Spending Ranges (Per Student)"])["% Passing Math"].mean()

spending\_passing\_reading = school\_spending\_df.groupby(["Spending Ranges (Per Student)"])["% Passing Reading"].mean()

overall\_passing\_spending = school\_spending\_df.groupby(["Spending Ranges (Per Student)"])["% Overall Passing"].mean()

Use the scores above to create a DataFrame called spending\_summary.

Include the following metrics in the table:

* Average math score
* Average reading score
* % passing math (the percentage of students who passed math)
* % passing reading (the percentage of students who passed reading)
* % overall passing (the percentage of students who passed math AND reading)

**Scores by School Size**

Use the following code to bin the per\_school\_summary.

size\_bins = [0, 1000, 2000, 5000]

labels = ["Small (<1000)", "Medium (1000-2000)", "Large (2000-5000)"]

Use pd.cut on the "Total Students" column of the per\_school\_summary DataFrame.

Create a DataFrame called size\_summary that breaks down school performance based on school size (small, medium, or large).

**Scores by School Type**

Use the per\_school\_summary DataFrame from the previous step to create a new DataFrame called type\_summary.

This new DataFrame should show school performance based on the "School Type".