Unit 4 | Assignment - Pandas, Pandas, Pandas

Background

The data dive continues!

Now, it's time to take what you've learned about Python Pandas and apply it to new situations. For this assignment, you'll need to complete **one of two** (not both) Data Challenges. Once again, which challenge you take on is your choice. Just be sure to give it your all -- as the skills you hone will become powerful tools in your data analytics tool belt.

Option 1: Heroes of Pymoli

Congratulations! After a lot of hard work in the data munging mines, you've landed a job as Lead Analyst for an independent gaming company. You've been assigned the task of analyzing the data for their most recent fantasy game Heroes of Pymoli.

Like many others in its genre, the game is free-to-play, but players are encouraged to purchase optional items that enhance their playing experience. As a first task, the company would like you to generate a report that breaks down the game's purchasing data into meaningful insights.

Your final report should include each of the following:

Player Count

- Total Number of Players

Purchasing Analysis (Total)

- Number of Unique Items
- Average Purchase Price
- Total Number of Purchases
- Total Revenue

Gender Demographics

- Percentage and Count of Male Players
- Percentage and Count of Female Players
- Percentage and Count of Other / Non-Disclosed

Purchasing Analysis (Gender)

- The below each broken by gender
  - Purchase Count
  - Average Purchase Price
  - Total Purchase Value
  - Average Purchase Total per Person by Gender

Age Demographics

- The below each broken into bins of 4 years (i.e. <10, 10-14, 15-19, etc.)
  - Purchase Count
  - Average Purchase Price
  - Total Purchase Value
  - Average Purchase Total per Person by Age Group

Top Spenders

- Identify the the top 5 spenders in the game by total purchase value, then list (in a table):
  - SN
  - Purchase Count
  - Average Purchase Price
  - Total Purchase Value

Most Popular Items

- Identify the 5 most popular items by purchase count, then list (in a table):
  - Item ID
  - Item Name
  - Purchase Count
  - Item Price
  - Total Purchase Value

Most Profitable Items

- Identify the 5 most profitable items by total purchase value, then list (in a table):

- Item ID
- Item Name
- Purchase Count
- Item Price
- Total Purchase Value

As final considerations:

- You must use the Pandas Library and the Jupyter Notebook.
- You must submit a link to your Jupyter Notebook with the viewable Data Frames.
- You must include a written description of three observable trends based on the data.
- See Example Solution (HeroesOfPymoli/HeroesOfPymoli_starter.ipynb) for a reference on expected format.

Option 2: Academy of Py

Well done! Having spent years analyzing financial records for big banks, you've finally scratched your idealistic itch and joined the education sector. In your latest role, you've become the 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.

As a first task, you've been asked to analyze the district-wide standardized test results. You'll be given access to every student's math and reading scores, as well as various information on the schools they attend. Your responsibility is to aggregate the data to and showcase obvious trends in school performance.

Your final report should include each of the following:

District Summary

- Create a high level snapshot (in table form) of the district's key metrics, including:
  - Total Schools
  - Total Students
  - Total Budget
  - Average Math Score
  - Average Reading Score
  - % Passing Math
  - % Passing Reading
  - Overall Passing Rate (Average of the above two)

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)

Top Performing Schools (By Passing Rate)

- Create a table that highlights the top 5 performing schools based on Overall Passing Rate. Include:
  - 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)

Bottom Performing Schools (By Passing Rate)

- Create a table that highlights the bottom 5 performing schools based on Overall Passing Rate. Include all of the same metrics as above.

Math Scores by Grade**

- Create a table 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 table 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 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)

Scores by School Size

- Repeat the above breakdown, but this time group schools based on a reasonable approximation of school size (Small, Medium, Large).

Scores by School Type

- Repeat the above breakdown, but this time group schools based on school type (Charter vs. District).

As final considerations:

- Use the pandas library and Jupyter Notebook.
- You must submit a link to your Jupyter Notebook with the viewable Data Frames.
- You must include a written description of at least two observable trends based on the data.
- See Example Solution (PyCitySchools/PyCitySchools_starter.ipynb) for a reference on the expected format.

Hints and Considerations

- These are challenging activities for a number of reasons. For one, these activities will require you to analyze thousands of records. Hacking through the data to look for obvious trends in Excel is just not a feasible option. The size of the data may seem daunting, but pandas will allow you to efficiently parse through it.

- Second, these activities will also challenge you by requiring you to learn on your feet. Don't fool yourself into thinking: "I need to study pandas more closely before diving in." Get the basic gist of the library and then *immediately* get to work. When facing a daunting task, it's easy to think: "I'm just not ready to tackle it yet." But that's the surest way to never succeed. Learning to program requires one to constantly tinker, experiment, and learn on the fly. You are doing exactly the *right* thing, if you find yourself constantly practicing Google-Fu and diving into documentation. There is just no way (or reason) to try and memorize it all. Online references are available for you to use when you need them. So use them!

- Take each of these tasks one at a time. Begin your work, answering the basic questions: "How do I import the data?" "How do I convert the data into a DataFrame?" "How do I build the first table?" Don't get intimidated by the number of asks. Many of them are repetitive in nature with just a few tweaks. Be persistent and creative!

- Expect these exercises to take time! Don't get discouraged if you find yourself spending hours initially with little progress. Force yourself to deal with the discomfort of not knowing and forge ahead. Consider these hours an investment in your future!

- As always, feel encouraged to work in groups and get help from your TAs and Instructor. Just remember, true success comes from mastery and *not* a completed homework assignment. So challenge yourself to truly succeed!

Copyright