Knowledge Streams

Pandas Mini Project 01

Read the submission instructions carefully. Any points deduction based on the below statements will not be entertained for queries. Hence, it is your responsibility to review each of them carefully.

1. Only 1 file is to be submitted.
2. All tasks are to be completed and any step missing will result in appropriate marks deduction.
3. The project requires advanced applications of all the concepts that you have studied in class and intermediate logic development.
4. There are some questions with constraints on them. Using any other command/method, even with correct logic and output will result in zero marks. Those constraints are in place to check your understanding of certain commands.
5. The instructors may not help you with this project during the timeline. **(Read next page for more)**

Objectives: Students will be able to perform EDA using Pandas in great depth.

- **Assignment Title:** PANDAS MINI PROJECT 01
- **Due Date:** 4 August, 2024 - 23:59 (PST)
- **Submission Method:** GCR and LMS
- **Weightage:** 20% of your overall grade
- **Instructions:** Next Page for detailed Instructions
- **Resources:** Pandas Documentation
- **Contact Information:** Rehman Sadaqat

Project Explanation

This project is a self-guided project.

What is a self-guided project?

A self-guided project is one where you can only use the resources available to you. Instructors may not help you even with the questions. The resource that you may use for your understanding is on next page.

Why do we have to do a self-guided project when instructors are there to help us?

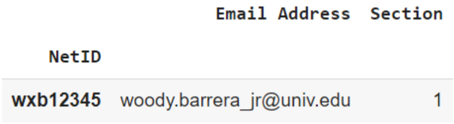
Suppose you are working as a data analyst at an XYZ company and your team lead is busy with another project and may not be able to help you. He gave you some tasks and what he expects. No logic and nothing else. How will you navigate that scenario? You will use the resources available to you. This project not only helps you with the concepts but also gives you an industrial scenario where you have to do everything on your own. Since this is your first project, I will keep the self-resources to a maximum but they will decrease with each Mini Project later in the cohort 😉

Rehman Sadaqat

Question Guidelines & Explanations

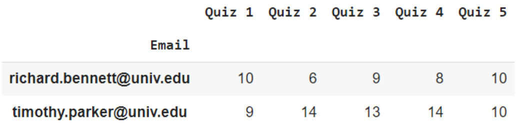
You are a data analyst at a university tasked with analyzing the grades of students in a course. Your goal is to provide insights into the performance of students, identify patterns, and make recommendations for improving student outcomes. You will use Python, Pandas, and Numpy to load, clean, and analyze the data where appropriate

1. Load the necessary libraries.
2. Certain EDA steps required on the roster df is already given in the file, expect your output to be something like this.



NetID	Email Address	Section
wxb12345	woody.barrera_jr@univ.edu	1

3. For the Homework grade df, the steps required are already given. No expected output will be shared for this, I already gave you one as a demo above. You must use a lambda filter for filtering in this code block or there will be appropriate points deduction.
4. For the Quiz df(s), the steps are as follows:

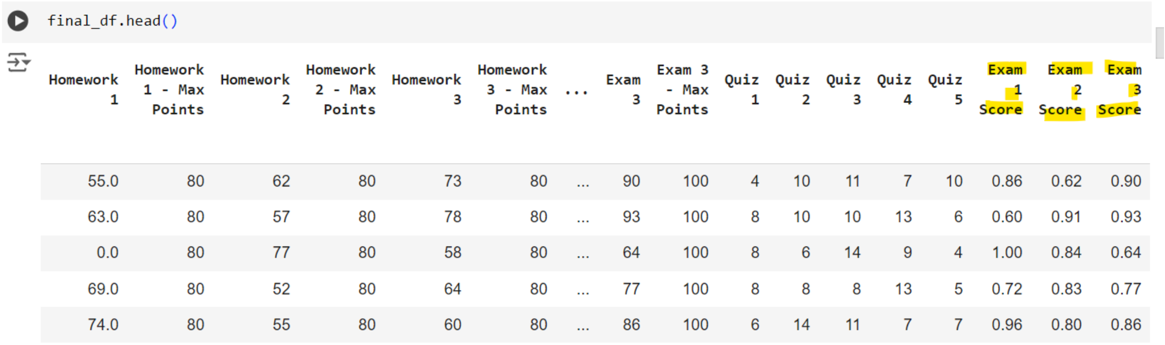


Email	Quiz 1	Quiz 2	Quiz 3	Quiz 4	Quiz 5
richard.bennett@univ.edu	10	6	9	8	10
timothy.parker@univ.edu	9	14	13	14	10

Since this requires some logic building, there are a few hints on one of the approaches that I used. Your approach may differ from this so these hints might not apply to everyone.

Hints: To tackle this task, start by creating a blank data structure to store multiple quiz grades. You'll be reading several files from a directory, each containing quiz results, so think about a method to loop through these files efficiently. For each file, extract a meaningful column name from the file's title, ensuring it's formatted neatly. As you read the data, focus on selecting only the necessary columns and standardizing any important fields like email addresses. Consider setting one of these columns as an index for easier data management. Finally, look into how you can merge this data with your existing structure, ensuring that any new columns align correctly with the data you've already processed. Handling cases where data might be missing for some entries will also be crucial. Use online resources to understand how to read files, manipulate strings, and combine data structures, and don't hesitate to experiment and debug your code as you go.

5. In Merging the dataframe, NetID should be your index.
6. To find percentage in Exam grades, the formula is obtained marks/total marks. You may use f-strings and loops for the scenario.



```
final_df.head()
```

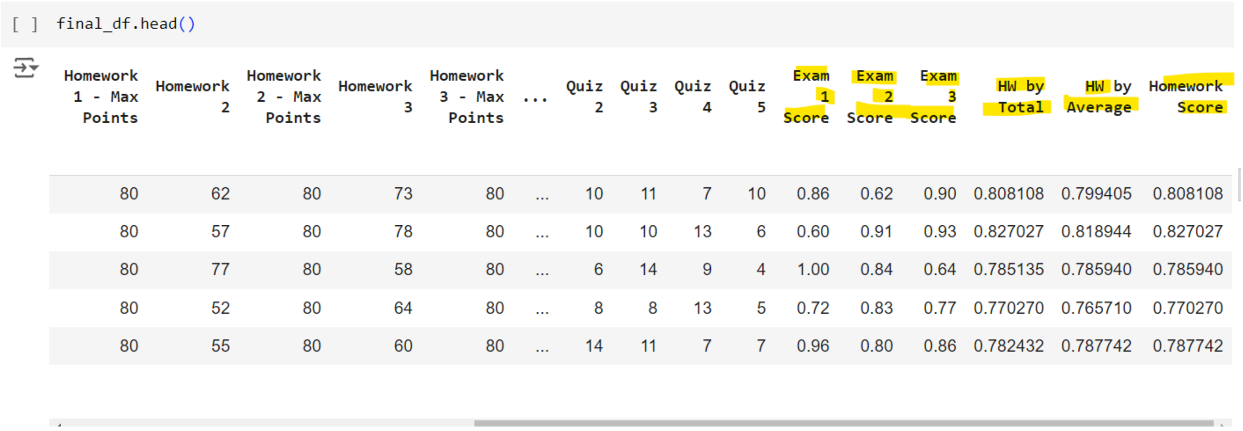
Homework 1	Homework 1 - Max Points	Homework 2	Homework 2 - Max Points	Homework 3	Homework 3 - Max Points	...	Exam 3	Exam 3 - Max Points	Quiz 1	Quiz 2	Quiz 3	Quiz 4	Quiz 5	Exam 1 Score	Exam 2 Score	Exam 3 Score
55.0	80	62	80	73	80	...	90	100	4	10	11	7	10	0.86	0.62	0.90
63.0	80	57	80	78	80	...	93	100	8	10	10	13	6	0.60	0.91	0.93
0.0	80	77	80	58	80	...	64	100	8	6	14	9	4	1.00	0.84	0.64
69.0	80	52	80	64	80	...	77	100	8	8	8	13	5	0.72	0.83	0.77
74.0	80	55	80	60	80	...	86	100	6	14	11	7	7	0.96	0.80	0.86

7. For Homework calculation, steps 1-3 are easy so I am skipping their resource. Step 4 might get you guys all riled up.

   Hint for Q4: hw_max_data to match the columns in hw_cols. The hw_max_data DataFrame initially contains the maximum possible points for each homework assignment, and hw_cols contains the actual scores. Renaming the columns ensures that both data sets align correctly by assignment when performing operations between them.

   PS: Check out what set_axis can do in this 😉

   Hint for Q5: To find the average, you can divide the actual homework score by the corresponding maximum points (element-wise division), sum these ratios for each student, and then divide by the total number of homework assignments. This gives the average proportion of points earned out of the maximum points across all assignments, reflecting the student's overall performance in homework.

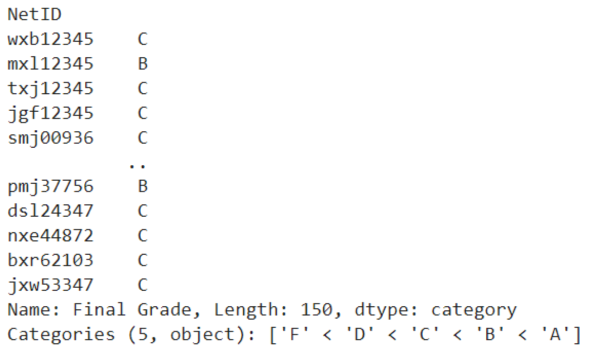


```
[ ] final_df.head()
```

Homework 1 - Max Points	Homework 2	Homework 2 - Max Points	Homework 3	Homework 3 - Max Points	...	Quiz 2	Quiz 3	Quiz 4	Quiz 5	Exam 1 Score	Exam 2 Score	Exam 3 Score	HW by Total	HW by Average	Homework Score
80	62	80	73	80	...	10	11	7	10	0.86	0.62	0.90	0.808108	0.799405	0.808108
80	57	80	78	80	...	10	10	13	6	0.60	0.91	0.93	0.827027	0.818944	0.827027
80	77	80	58	80	...	6	14	9	4	1.00	0.84	0.64	0.785135	0.785940	0.785940
80	52	80	64	80	...	8	8	13	5	0.72	0.83	0.77	0.770270	0.765710	0.770270
80	55	80	60	80	...	14	11	7	7	0.96	0.80	0.86	0.782432	0.787742	0.787742

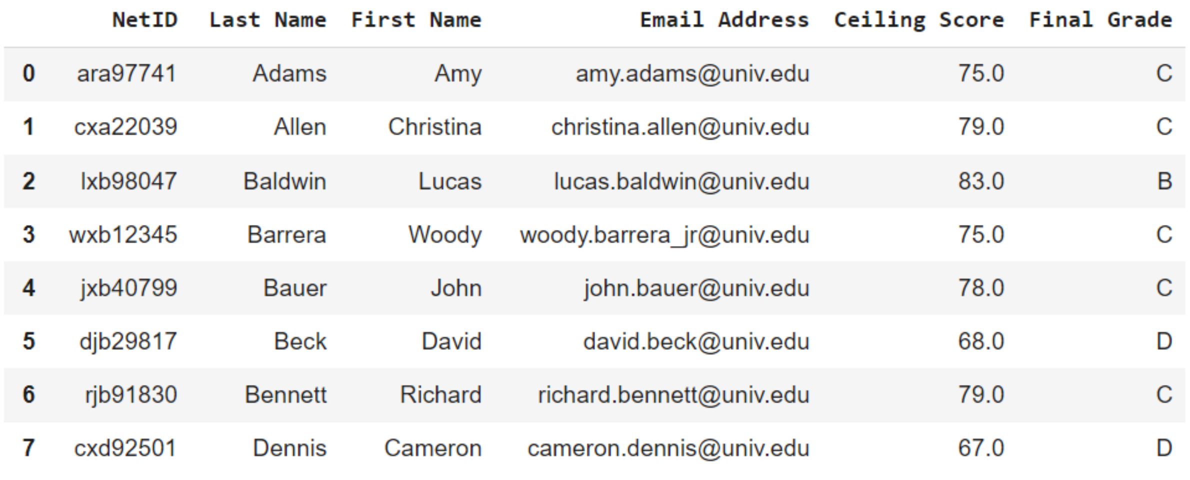
8. You MUST use regex to filter out the columns containing the word Quiz and then the appropriate number in front of it.

9. After grouping the data and performing all the steps written in the ipynb file, this is the expected output.



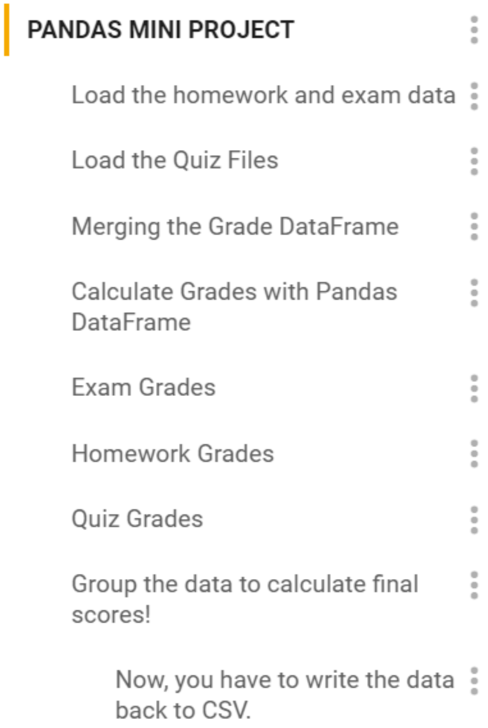
```
NetID
wxb12345    C
mxl12345    B
txj12345    C
jgf12345    C
smj00936    C
           ..
pmj37756    B
dsl24347    C
nxe44872    C
bxr62103    C
jxw53347    C
Name: Final Grade, Length: 150, dtype: category
Categories (5, object): ['F' < 'D' < 'C' < 'B' < 'A']
```

10. The last step is to write the data back into a CSV. But how? Figure it out yourself or through documentation. This is the final expected output.



	NetID	Last Name	First Name	Email Address	Ceiling Score	Final Grade
0	ara97741	Adams	Amy	amy.adams@univ.edu	75.0	C
1	cxa22039	Allen	Christina	christina.allen@univ.edu	79.0	C
2	lxb98047	Baldwin	Lucas	lucas.baldwin@univ.edu	83.0	B
3	wxb12345	Barrera	Woody	woody.barrera_jr@univ.edu	75.0	C
4	jxb40799	Bauer	John	john.bauer@univ.edu	78.0	C
5	djb29817	Beck	David	david.beck@univ.edu	68.0	D
6	rjb91830	Bennett	Richard	richard.bennett@univ.edu	79.0	C
7	cxd92501	Dennis	Cameron	cameron.dennis@univ.edu	67.0	D

Grading Scheme



PANDAS MINI PROJECT

Load the homework and exam data

Load the Quiz Files

Merging the Grade DataFrame

Calculate Grades with Pandas DataFrame

Exam Grades

Homework Grades

Quiz Grades

Group the data to calculate final scores!

Now, you have to write the data back to CSV.

- **Steps 1 - 4, 8 - 9 are 3 points each.**
- **Steps 5 - 7 are of 4 marks each.**

If there are any new guidelines or announcements regarding this, they will be shared in class OR the WA group.