Python Libraries - Exam Practice

you are a data analyst in a well known college and you got the following datasets (according to 2023 year).
Note: All The provided courses have ended so all students should have final grades in the courses they registered to.

1. **teachers.csv file** → Contain data about all the active teachers.
   The csv has the following fields:
   a. id → Unique id for each teacher
   b. first_name
   c. last_name
   d. gender → Teacher gender (Male or Female)
   e. years_of_expirence
   f. salary → The teacher monthly salary in shekels

2. **students.csv file** → Contain data about all active students.
   The csv has the following fields:
   a. id → Unique id for each student
   b. first_name
   c. last_name
   d. gender → Student gender (Male or Female)
   e. age
   f. city → Student city
   g. education → The student education level
   h. email

3. **courses.csv file** → Contain data about all courses been opened during 2023 year

   The csv has the following fields:

   a. course_id
   b. course_name
   c. course_category → Each course should be associated to one category
   d. teacher_id → A pointer to the teacher that has been teaching the course. Each course should have only one teacher but a teacher can teach multiple courses
   e. course_start_date → The starting date of the course (all courses in the dataset are only from the year 2023)

4. **course_student.csv file** → Contain data about the association between student and course that appears in the students.csv and courses.csv files.

   The csv has the following fields:

   a. id
   b. student_id → Pointing to the associated student id
   c. course_id → Pointing to the associated course id
   d. final_grade → The final grade the student got in this specific course.

Data Preparation:

Make sure to use copy() of the original datasets in all data preparation actions.

1. Handle duplicate data in student dataset:
    a. In student dataset duplicate data is if you have:
        i. Different student rows with the same email
    b. In case you find duplicate data, remove the student with less information. Make sure you change associations in other data sets to the duplicated student id so your data will still be accurate.
    For example → If you choose to remove student_id 3 because it's duplicated with student_id 5, make sure to change any association in other data sets from student_id 3 to student_id 5.

2. Handle missing data in all datasets:
    a. In case the missing data in a specific column is above 5% of all provided data rows fill the missing data with a default valid value of your choice.
    b. In case the missing data is below 5% remove the row from your calculations and make sure to adjust your other datasets accordingly.
    c. In case you found a row with missing data with a mandatory column (like id, name, etc…) remove this row from your dataset.

Data analysis:

Use the copy() datasets from your data preparation answer and answer the following questions:

1. Explore the **students.csv** dataset and answer the following questions, base your answers with data calculations and visualizations if needed:
    a. Count the number of students by gender, show a dataframe with each gender type and how many students we have from that type.
    Plot bar chart to visualize your result.
    b. Plot the student ages distribution with histogram chart.
    c. Find what is the city with the highest number of registered students and what is the city with the lowest number of registered students.

2. Explore the **coursres.csv** dataset and answer the following questions, base your answers with data calculations and visualizations if needed:
    a. Count the number of courses by category, show a dataframe with each category and how many courses are associated with that category.
    Plot bar chart to visualize your result.
    b. Create a dataframe that shows for each category what are the unique courses names that are associated with that category.
    c. Create a dataframe that shows how many different courses have been started in each month during 2023 year.
    d. Use the dataframe from the previous exercise (2.c) and create a line chart that represent the trend of the number of

opened courses during each month (x-axis should be month date and y-axis should be number of courses in each month)

3. Explore the **teachers.csv** dataset and answer the following questions, base your answers with data calculations and visualizations if needed:
    a. Calculate the mean salary for a teacher in the college.
    b. Plot a scatter plot to find out the correlation between the teacher years of experience and its salary.
       Determine according to the plot if there is any correlation between them.

4. Use **all datasets** and answer the following questions:
    a. Find the id, name and salary of the teacher that is teaching the most courses during 2023 year.
       In case there are multiple teachers, include them all in your answer.
    b. Investigate if there is a linear trend between student education level and student grade in the course.
       Use a scatter plot to support your answer.
    c. Find for each course name (during the entire year) what was the mean grade.
       Plot a bar chart that represent your answer

d. Create a new dataframe that shows for each student what is its final_college_grade.
   final_college_grade is the mean grade that a specific student got in all of his courses together.
   If a student performs the same course more than once, take in your calculation the higher grade for that course.

ECOM SCHOOL
המכללה למקצועות הדיגיטל וההייטק