Practical Exam - University Enrollment

Instructions

- Use Python or R to perform the tasks required.
- Write your solutions in the workspace provided from your certification page.
- Include all of the visualizations you create to complete the tasks.
- Visualizations must be visible in the published version of the workspace. Links to external visualizations will not be accepted.
- You do not need to include code unless the question says you must.
- You must pass all criteria to pass this exam. The full criteria can be found here.

Background

You are working as a data scientist at a local University.

The university started offering online courses to reach a wider range of students.

The university wants you to help them understand enrollment trends.

They would like you to identify what contributes to higher enrollment. In particular, whether the course type (online or classroom) is a factor.

Data

The dataset contains records for each course offered over the last 5 years.

The dataset can be downloaded from here.

Column Name	Criteria
course_id	Nominal. The unique identifier of the course. Missing values are not possible due to the database structure.
course_type	Nominal. Whether the course is “online” or “classroom” based. Replace missing values with “classroom”.
year	Discrete. The year the course was offerered. Any year from 2011 to 2022. Replace missing values with 2011.
enrollment_count	Discrete. The number of students enrolled onto the course. Replace missing values with 0.
pre_score	Continuous. The average score of the enrolled students on the pre course exam. Replace missing values with 0.
post_score	Continuous. The average score of students who complete the course on the post course exam. Replace missing values with 0.
pre_requirement	Nominal. The previous course completion requirement for students to enroll. One of “None”, “Beginner”, “Intermediate”. Replace missing values with “None”.
department	Nominal. The department of the university offering the course. One of “Science”, “Technology”, “Engineering”, “Mathematics” Replace missing values with “unknown”.

Tasks

Write your answers in your workspace.

1. For every column in the data:
    a. State whether the values match the description given in the table above.
    b. State the number of missing values in the column.
    c. Describe what you did to make values match the description if they did not match.
2. Describe the distribution of the enrollment counts. Your answer must include a visualization that shows the distribution.
3. Create a visualization that shows how many courses were of each type. Use the visualization to:
    a. State which type of course has the most observations
    b. Explain whether the observations are balanced across the types.
4. Describe the relationship between course type and the enrollment count. Your answer must include a visualization to demonstrate the relationship.
5. The university wants to predict how many students will enroll in a course. State the type of machine learning problem that this is (regression/ classification/ clustering).
6. Fit a baseline model to predict how many students will enroll using the data provided. You must include your code.
7. Fit a comparison model to predict how many students will enroll using the data provided. You must include your code.
8. Explain why you chose the two models used in parts 6 and 7.
9. Compare the performance of the two models used in parts 6 and 7, using any method suitable. You must include your code.
10. Explain which model performs better and why.