Practical Exam - Fitness Class

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

GoalZone is a fitness club chain in Canada.

GoalZone offers a range of fitness classes in two capacities - 25 and 15.

Some classes are always fully booked. Fully booked classes often have a low attendance rate.

GoalZone wants to increase the number of spaces available for classes.

They want to do this by predicting whether the member will attend the class or not.

If they can predict a member will not attend the class, they can make another space available.

Data

The dataset contains each record when a member registered for a fitness class.

The dataset can be downloaded from here.

Column Name	Criteria
booking_id	Nominal. The unique identifier of the booking. Missing values are not possible due to the database structure.
months_as_member	Discrete. The number of months as this fitness club member, minimum 1 month. Replace missing values with the overall average month.
weight	Continuous. The member's weight in kg, rounded to 2 decimal places. The minimum possible value is 40.00 kg. Replace missing values with the overall average weight.
days_before	Discrete. The number of days before the class the member registered, minimum 1 day. Replace missing values with 0.
day_of_week	Nominal. The day of the week of the class. One of “Mon”, “Tue”, “Wed”, “Thu”, “Fri”, “Sat” or “Sun”. Replace missing values with “unknown”.
time	Ordinal. The time of day of the class. Either “AM” or “PM”. Replace missing values with “unknown”.
category	Nominal. The category of the fitness class. One of “Yoga”, “Aqua”, “Strength”, “HIIT”, or “Cycling”. Replace missing values with “unknown”.
attended	Nominal. Whether the member attended the class (1) or not (0). Missing values should be removed.

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. Create a visualization that shows how many bookings attended the class. Use the visualization to:
    a. State which category of the variable attended has the most observations
    b. Explain whether the observations are balanced across categories of the variable attended
3. Describe the distribution of the number of months as a member. Your answer must include a visualization that shows the distribution.
4. Describe the relationship between attendance and number of months as a member. Your answer must include a visualization to demonstrate the relationship.
5. The business wants to predict whether members will attend using the data provided. State the type of machine learning problem that this is (regression/ classification/ clustering).
6. Fit a baseline model to predict whether members will attend using the data provided. You must include your code.
7. Fit a comparison model to predict whether members will attend 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.