Lifeblood data scientist take-home assignment: instructions

Time required: ~2 hours

Task:
The synthetic dataset provided contains information about songs that were released over a five-year period. Your task is to perform exploratory data analysis to inform the development of two models (a and b):

a) A model to predict which songs will sell more than two million copies.
b) A statistical model that can be used to evaluate the popularity of a genre and/or band over sequential time periods (year or quarter).

Please note that the task is not to build the models, but to explore the dataset, identify any potential problems or complexities, identify any necessary preprocessing and feature engineering steps, and decide how to approach the modelling tasks.

You can assume that inflation is zero. If you make any additional assumptions (either about the dataset or how a model would be used), please state those clearly.

Dataset details:
The dataset includes the following columns: song_id, band_id, genre, date, energy, loudness, length, and copies_sold. Each song has a unique ID (song_id) and there's one row for each song released from 2018-2022. Bands are identified by band_id and each band can have multiple songs. Date is the date a song was released. Loudness (arbitrary units), length (minutes), and energy (arbitrary units) describe characteristics of the song. The copies_sold column is the number of copies of the song sold (in millions).

Format:
The exploratory data analysis should be done in a notebook format (e.g. RMarkdown, Quarto, Jupyter notebook) and the code provided to us prior to the interview. You'll be asked to take us (data scientists) through the analysis during the interview, so it doesn't need to be a polished standalone document. A combination of plots and brief commentary in the form of dot points with some headings for structure is fine.

Interview:
In the interview, we will ask you to take us through your analysis and ask some questions to help us understand how you approached the task, what you found, and how that would inform the development of the two models.

Below are some questions you might like to consider during your analysis and while preparing for the interview. Since this is designed to be a short task, there's no expectation that your analysis be exhaustive or that you answer all these questions.

- Is there anything missing in the data? If so, how would you handle this?
- Are there any outliers? If so, how would you handle this?
- Are there any temporal patterns?
- Are there any features that could be good predictors?
- Are there any features you would add to include as predictors in a model?
- Is there anything about the dataset that might make the task challenging?
- Are there any assumptions you're making (either about the dataset or how a model would be used)?
- What kind of models would be suitable for this task?

- Are there any conditions where you would expect a model would not perform well in production?
- For model b) what is an appropriate response variable and why?