Case Study

This case study examines both modeling and basic data engineering strengths. Along with this document, you've received a dataset: Silver_Data.csv.

I. Load the Silver dataset provided (Silver_Data.csv). Write Python to:

Conduct Exploratory Data Analysis (EDA): Perform initial exploration of the Silver-level dataset to assess data quality, understand variable distributions, identify inconsistencies, and spot trends across shows and seasons.

- Summarize findings in a one-page document.

Conduct Data Cleaning: Transform the Silver-level dataset into the Gold-level format by:

- Handling duplicates
- Imputing or addressing missing values
- Sorting/unordering issues
- Aggregating daily entries into weekly metrics

1. Aggregate total impressions and $s by campaign by channel
2. Produce a report with campaign metrics
3. Answer scenario-based questions:
    - Briefly explain the purpose and main challenges of each medallion layer in 2-3 sentences
    - Given a sudden schema change upstream, describe your process for ensuring the downstream silver and gold layers still function without interruption
    - Describe how you would implement data lineage and monitoring across layers to quickly identify and resolve data quality issues

II. Load your cleaned (Gold) dataset. Write Python to:

Modeling: Fit a marketing mix model on the cleaned (Gold) data.

- Fit a marketing mix model using multiple linear regression, or Bayesian statistics, to estimate the effects of each spend channel
- Interpret the significant variables
- Try 1-3 approaches to adstock and/or saturation. Identify insights, challenges, recommend next steps.

Simulation (bonus): using the fitted model, simulate and plot the impact on viewers if digital spend increases by 20% for a given period, holding other spends constant. Provide:

- Python code for the simulation
- A plot visualizing predicted vs. actual viewers for the period of increased digital spend

Interpretation: Explain modeling assumptions, interpret the coefficients, and identify which channels are driving impact.

Optimization: Using the fitted model, simulate or recommend optimized media spend allocation to maximize revenue.

III. Submission Requirements:

1. Jupyter notebook (ipynb, html format) with code and short explanations
2. All code should be commented for clarity
3. Explain your assumptions and methodological choices

Dataset details

The dataset contains marketing spend and other details associated with a certain episode of a show. Assume the marketing spend channels are independent of each other. We have 3 channels.
Columns -

Show - Show Name - There are 8 shows
Season - Season Number of a show
Air Date - Date on which an episode aired.
Week Number - Linked to a specific show, marking the week during which a show was active - either through episode airs or prelaunch marketing
Episode Number - Episode Number of a specific show and season. One episode per week per show.
Episode Type - Prelaunch, Premiere, Regular or Finale. During Prelaunch, Episode number is 0. Prelaunch refers to the marketing done for a show before launch.
Network_TV_Spend - In Dollars, the amount of money spent on the Network TV marketing channel for a specific episode.
Cable_TV_Spend - In Dollars, the amount of money spent on the Cable TV marketing channel for a specific episode.
Digital_Spend - In Dollars, the amount of money spent on the Digital marketing channel for a specific episode.
Impressions - The amount of impressions received for the ads we aired marketing our shows. It is the sum of impressions received by each of the channels for their marketing spend.
Note - This DOES NOT refer to impressions received by the airing of an episode of our show. It refers to the impressions received by ads we aired marketing our shows, and hence is linked to marketing spend, not viewership or revenue.
True_Viewership - The verified number of viewers for a certain episode of a show.
Revenue - Revenue generated by the airing of an episode.
Note- This is entirely linked to the number of viewers. For the purposes of this case study, assume we get paid a fixed amount for each viewer.
Holiday - Binary Variable that marks whether the airing of an episode was on a Holiday.
Note - This is mocked up for the purposes of the case study and may not represent actual holidays.
LeadIn_Bonus - Binary variable that marks whether an episode received a lead-in, for example, by airing immediately after an NFL game.

Final note
For the purposes of the case study, the data does not include genre, audience demographics, or competition from other shows. All unobserved effects not present in the dataset should be implicitly captured in the base performance of each show.