FAQs

1. How and why did you use XGBoost?
   XGBoost was used to predict high spenders because it works well with imbalanced datasets. High spenders are fewer but essential for revenue, so we needed a model to identify them effectively. XGBoost achieved a recall of 0.98, meaning it captured most high spenders accurately.

2. Why did you choose the Black Friday dataset?
   The dataset offered a mix of demographic, product, and purchase information, making it ideal for studying customer behavior during a key shopping event. Its variety allowed us to explore multiple factors influencing spending.

3. Why use only one dataset?
   We used one dataset because it already contained the necessary demographic, product, and purchase details. Integration within the dataset involved combining these data types to analyze relationships, such as how demographics influence spending on specific products. This approach avoided unnecessary complexity while ensuring a comprehensive analysis.

4. How did you handle missing data in product categories?
   Missing values in Product_Category_2 and Product_Category_3 were replaced with -1 to retain records while marking gaps for later analysis. This ensured no data was lost unnecessarily.

5. Why normalize purchase amounts?
   Normalizing purchase values ensured fair comparisons between customers, especially since purchase amounts had a wide range. It also reduced the influence of outliers on analysis.

6. Why create features like Age × City?
   Interaction features like Age × City capture relationships between variables. For example, they help explore whether older customers in rural areas spend differently than younger ones in urban areas. These insights are valuable for targeted strategies.

7. Why split customers into high and low spenders?
   Splitting customers into High (\$15,000+) and Low (<\$5,000) spending segments helped identify distinct customer behaviors. High spenders are critical for revenue, while low spenders represent opportunities for growth.

8. Why did rural spending come out higher than urban spending?
   Rural customers may make larger or bulk purchases during Black Friday, treating it as a significant shopping event. Urban customers might shop more frequently but spend less per transaction.

9. What was the data integration process?

We integrated customer demographics, product details, and purchase data into a single structured dataset. This allowed us to explore relationships, such as how age or gender influences spending across product categories.

10. How did visualizations help the analysis?
Visualizations like bar plots and heatmaps made trends and relationships easier to interpret. For instance, bar plots showed spending by age and city, while heatmaps highlighted correlations among product categories.

11. What insights did the correlation matrix provide?
It revealed strong relationships between product categories, meaning customers often buy across multiple types. Demographics like Gender and Age showed weaker correlations, suggesting product preferences drive purchases more.

12. Why focus on high spenders?
High spenders are fewer in number but contribute disproportionately to revenue. Understanding their behavior allows businesses to design loyalty programs and personalized marketing to retain these valuable customers.

13. Why did you encode variables like gender and age?
Encoding variables like Gender (Female = 0, Male = 1) and Age groups into numeric values allowed us to use them in visualizations and machine learning models without losing their meaning.

14. What were the limitations of the dataset?
The dataset lacked time-series information, which would have helped study seasonal trends. It also had limited product descriptions, restricting deeper analysis at the product level.

15. What would you do differently with more time or resources?
We would integrate external datasets, such as economic indicators or regional trends, to enrich the analysis. Additionally, we would build time-series models to explore how spending patterns evolve over time.