Dataset Recap and Problem Scope:

I’m continuing with the Diabetes Health Indicators Dataset from Kaggle, which is derived from the 2015 CDC BRFSS survey. This dataset includes over 253,000 self-reported responses covering health behaviors, chronic conditions, and lifestyle factors. The business problem I’m addressing is: “How can we predict an individual’s diabetes status using behavioral, demographic, and health-related variables?” Since Module 1, I’ve refined the scope by removing 24,206 duplicate records and flagging domain-specific placeholders (e.g., 88, 77, 99), which will be addressed during preprocessing.

Data Dictionary (Selected Variables):

Variable	Data Type	Description
Diabetes_binary	Binary	1 = Diagnosed with diabetes; 0 = No
HighBP	Binary	High blood pressure (1 = Yes, 0 = No)
HighChol	Binary	High cholesterol (1 = Yes, 0 = No)
CholCheck	Binary	Had cholesterol check in past 5 years
BMI	Numeric	Body Mass Index (12–98 typical range)
MentHlth	Ordinal	Days mental health not good in past 30 days (0–30)
PhysHlth	Ordinal	Days physical health not good in past 30 days (0–30)
GenHlth	Ordinal	Self-rated general health (1 = Excellent to 5 = Poor)
Age	Categorical	Age group (1 = 18–24, …, 13 = 80+)
Income	Categorical	Income level (1 = <$10K to 8 = $75K+)

Summary Statistics:

- Diabetes_binary: Only 14% of respondents are diabetic, highlighting a class imbalance.
- BMI: Mean = 28.38, Max = 98; the upper tail indicates potential outliers.
- MentHlth & PhysHlth: Both are right-skewed with high standard deviations (7.41 and 8.72 respectively), suggesting uneven distributions across respondents.
- No missing values in standard format, but placeholder values (e.g., 88, 99) require contextual interpretation.

Data Cleaning:

- Removed 24,206 duplicate records.
- Flagged placeholder values (88, 77, 99) for potential recoding or transformation.
- All variables are either binary, ordinal, or numeric—no string-based categories.
- MentHlth and PhysHlth retained for exploratory purposes despite subjective nature.

Visualizations and Initial Insights:

- **Histogram of BMI:** Showed right-skewed distribution, confirming BMI as a key risk factor.
- **Boxplot of Age by Diabetes Status:** Diabetic individuals skew older; strong age-risk correlation observed.
- **Correlation Heatmap:** Revealed weak to moderate correlation across numeric predictors, suggesting opportunity for feature engineering or interaction terms.

AI Feedback:

ChatGPT helped identify that class imbalance was not yet addressed and suggested resampling or model weighting. It also pointed out overlooked transformations (e.g., binning BMI, recoding ordinal features), recommended exploring interactions (e.g., HighBP × BMI), and emphasized the need for clearer labeling. Additionally, potential multicollinearity among predictors (e.g., HighBP, Stroke) and logical inconsistencies in responses (e.g., 0 unhealthy days with walking difficulty) were flagged.

What I'd Like Feedback On:

Would it be advisable to group variables like Age or BMI into categorical bins to improve model interpretability and performance? Also, should I exclude subjective features like MentHlth and PhysHlth from predictive modeling altogether or retain them with transformations?