2026

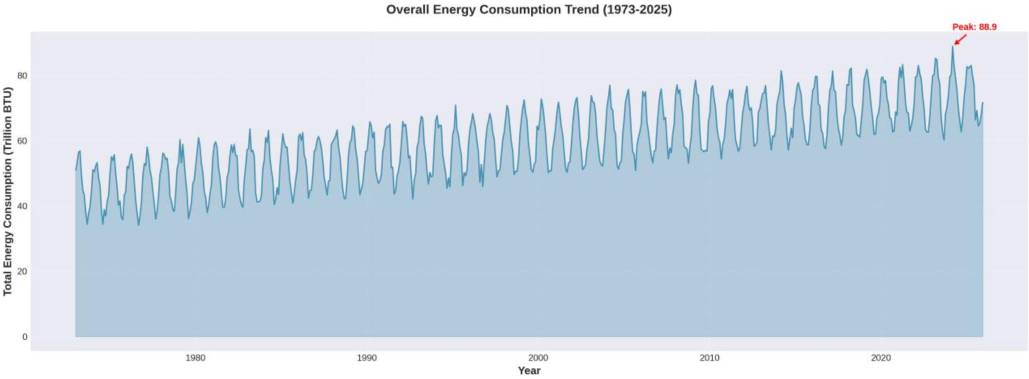
Efficiency and accuracy comparison of machine learning algorithms for predicting US energy consumption across sectors

DATA VISUALIZATION

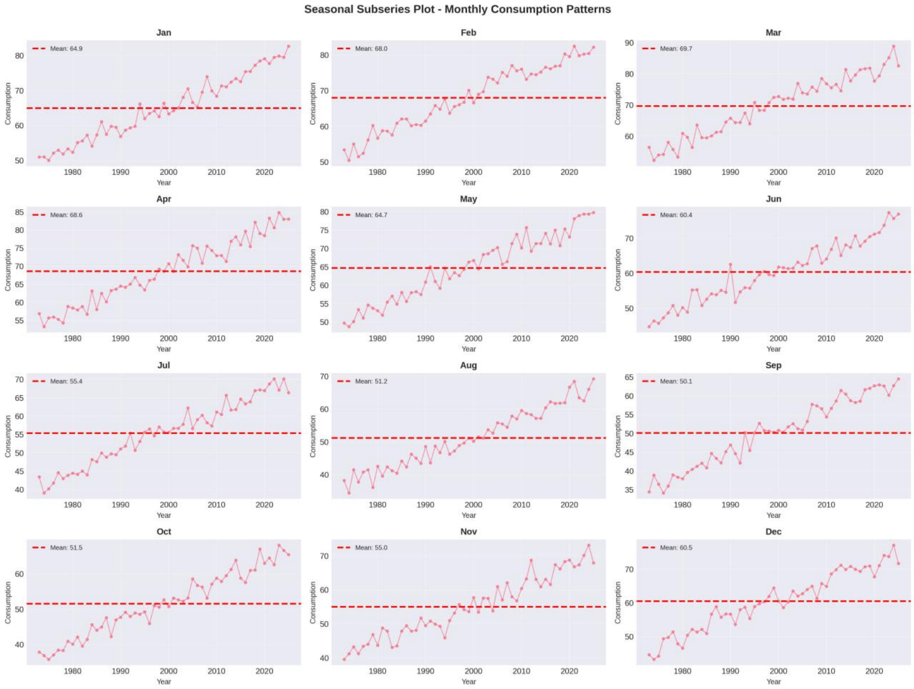
ROHAN DAS (2420080057)
SAI PRANEETHA (2420080061)
NIKHILESWAR REDDY (2420080064)
AMRUTHA REDDY (2420080079)

MACHINE LEARNING [24AD2204]

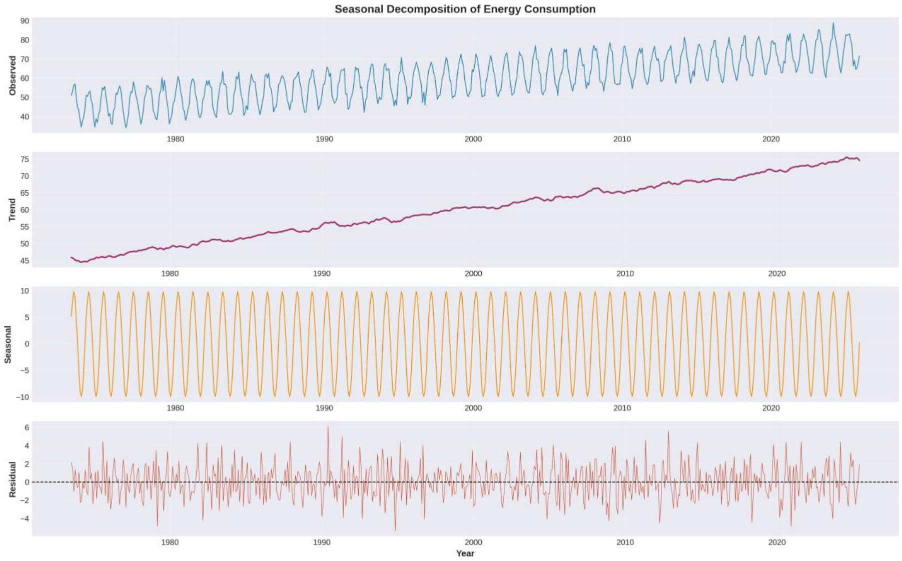
1. **Total Primary Energy Consumed by the End-Use Sectors (Monthly, Trillion Btu):**
   This dataset (from the U.S. Energy Information Administration's Monthly Energy Review) records monthly "Total Primary Energy Consumed by the End-Use Sectors" in Trillion Btu, along with sector-level components for Residential, Commercial, Industrial, and Transportation. I used it because it is a long, consistent monthly time-series with clear real-world patterns like seasonality and sectoral variation, which is ideal for building and validating forecasting/ML models. It helps us decide and understand how overall energy demand changes over time and which sectors drive those changes, supporting tasks like demand forecasting, planning, and comparing model performance on real energy consumption behavior.



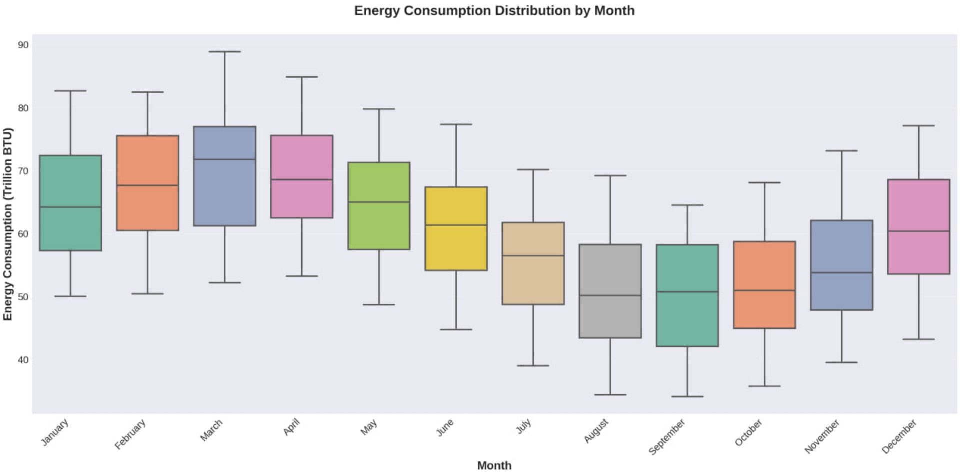
2. **Seasonal Subseries Plot (Monthly Consumption Patterns):**
   This figure is a seasonal subseries plot that divides the full monthly time-series into 12 panels (January to December), where each panel shows that month's consumption values across years and the dashed horizontal line shows the mean consumption for that month. I used it because it makes seasonality easy to see and compare—showing which months are consistently higher or lower, and whether the month-wise pattern is stable or changing over time. It helps us decide what to include in forecasting/ML (for example, adding month/season indicators, using seasonal decomposition or differencing, and choosing models that can capture seasonal effects) so the model does not treat regular monthly peaks and dips as random noise.



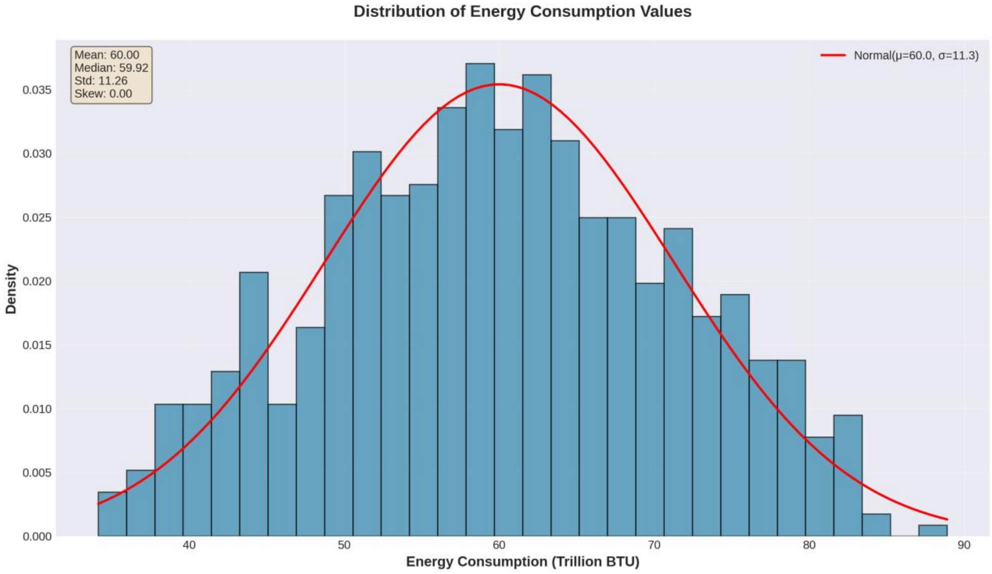
3. **Seasonal Decomposition of Energy Consumption (Observed–Trend–Seasonal–Residual):** This figure shows a seasonal decomposition of the monthly energy consumption series into four parts: the **Observed** original series, the **Trend** (long-term movement), the **Seasonal** component (repeating yearly pattern), and the **Residual** (remaining irregular variation/noise). I used it because it separates the overall increase over time (trend) from the repeating monthly cycle (seasonality), making it clear what structure the forecasting model must learn versus what is random. It helps us decide the right preprocessing and modeling choices—such as whether to detrend/seasonally adjust, include month/season features, and evaluate errors on residual-like variability—so the model focuses on predictable components and improves forecasting accuracy and stability.



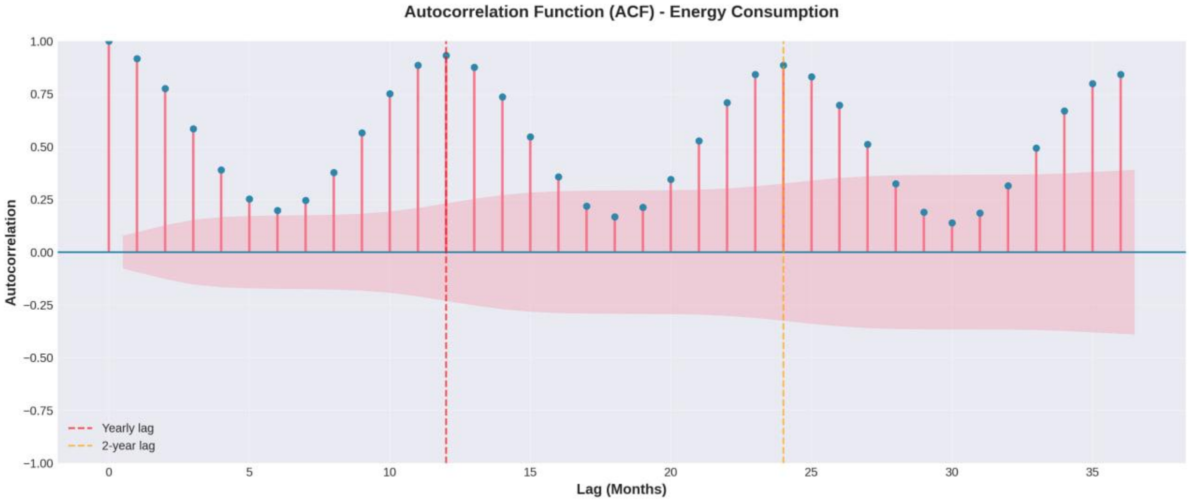
4. **Energy Consumption Distribution by Month (Box Plot):** This figure is a box plot that compares the distribution of energy consumption for each month (January to December) by showing the median, interquartile range (IQR), and spread/outliers of monthly values across all years. I used it because it quickly highlights which months typically have higher or lower consumption and which months show more variability (i.e., less stable demand), which is not as clear from a single line plot. It helps us decide season-related modeling choices—such as adding month indicators, handling months with higher variance using robust models/loss functions, and setting realistic error expectations for peak months—so forecasting performance is evaluated and improved with awareness of month-wise variability.



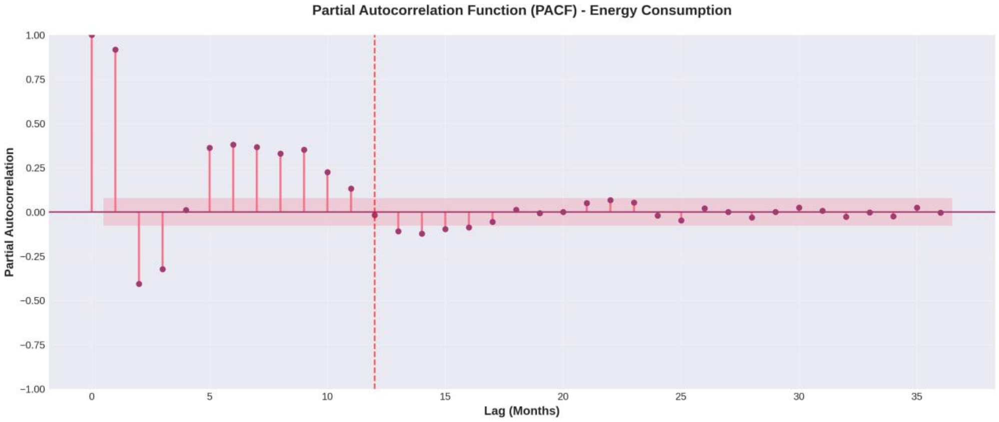
5. **Distribution of Energy Consumption Values (Histogram + Normal Fit):**
   This figure is a histogram (density plot) of all energy-consumption values, with a fitted normal (Gaussian) curve overlaid, and summary statistics like mean, median, standard deviation, and skewness. It matters because it checks the **shape** of your target variable—whether values are roughly symmetric/normal, heavy-tailed, or skewed—which directly affects preprocessing choices (e.g., log/Box-Cox transform), outlier handling, and which error metric or loss function will be stable. It also supports modeling decisions: if the distribution is close to normal, standard regression assumptions and RMSE-like objectives may behave well; if not, robust losses/metrics (MAE/Huber) or transforming the target can improve accuracy and reduce sensitivity to extreme months.



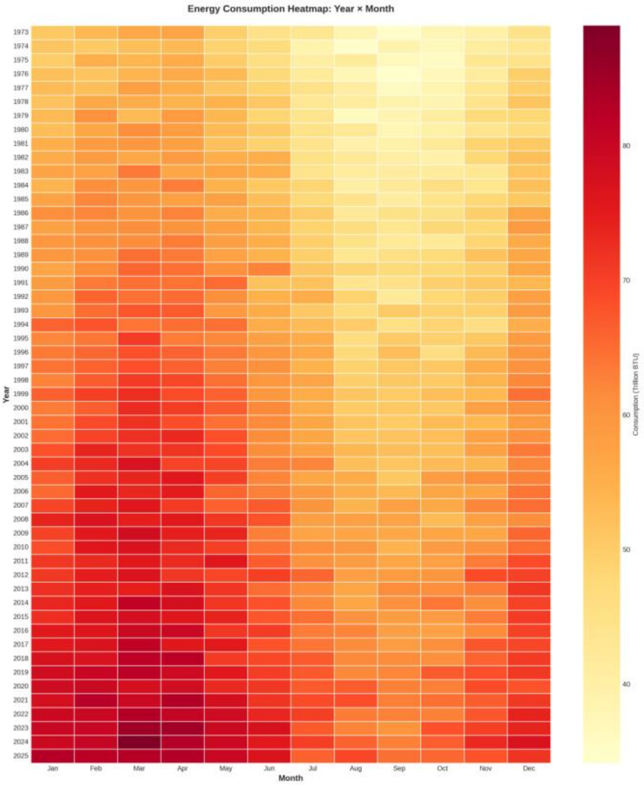
6. **Autocorrelation Function (ACF) – Energy Consumption:**
   This figure is an ACF plot that shows the correlation between the energy-consumption time series and its lagged values (in months), along with a confidence band and marked seasonal lags around 12 months (yearly) and 24 months (two-year). It matters because it confirms strong temporal dependence and repeating seasonality, meaning previous months/years contain useful signal for forecasting and the data violates the “independent samples” assumption of random splitting. It guides model setup: engineer lag features (e.g., 1–12 plus 12 and 24), add seasonal/month indicators, and evaluate using time-series splits (walk-forward/blocked CV) to avoid leakage and inflated performance.



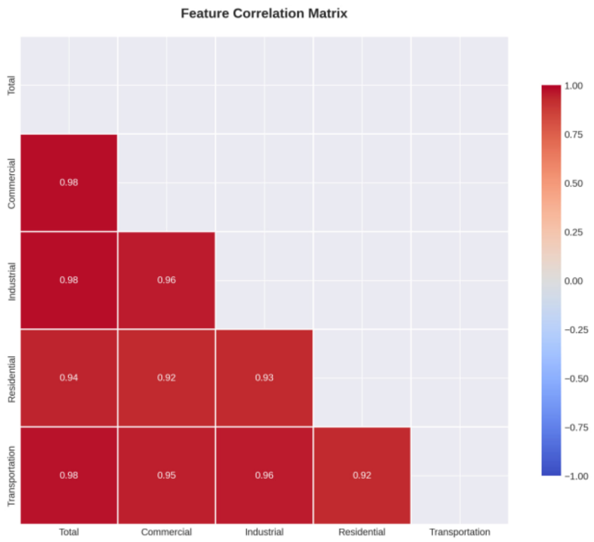
7. **Autocorrelation Function (ACF) – Energy Consumption:**
   This figure is an ACF plot showing correlation between the energy-consumption time series and its lagged values (in months), with a confidence band to identify statistically significant lags. Seasonal spikes around 12 months (yearly) and 24 months (two-year) indicate repeating seasonality in the monthly series. It matters because this dependence means past months/years contain forecasting signal and random splitting can give overly optimistic results, so use lag features (e.g., 1–12, 12, 24), month/season indicators, and time-ordered validation (walk-forward/blocked CV).



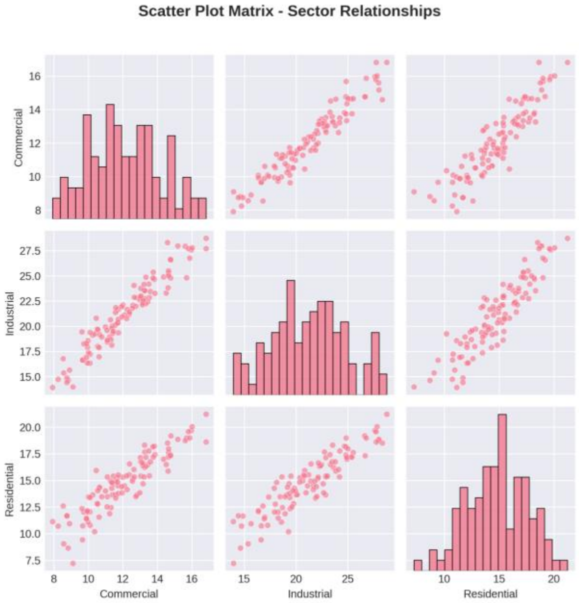
8. **Energy Consumption Heatmap: Year × Month:**
   This figure is a heatmap that visualizes monthly energy consumption across multiple years, where color intensity encodes consumption level (darker/redder = higher, lighter/yellower = lower). It matters because it reveals both long-term changes (overall intensification across later years) and recurring seasonal structure (certain months consistently higher/lower across many years). It guides model setup by motivating time features (month-of-year), seasonal components (12-month cycle), and possibly year/trend terms; it also helps spot unusual years/months for outlier handling and robustness checks.



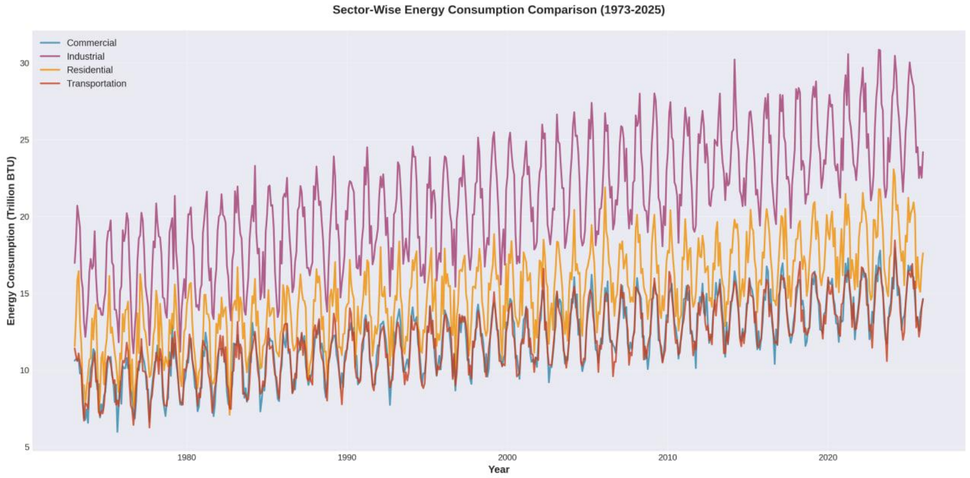
9. **Feature Correlation Matrix – Energy Sectors:**
This correlation heatmap shows that the sector features (Commercial, Industrial, Residential, Transportation) and Total are all very strongly positively correlated (values close to +1), meaning the predictors contain overlapping information and the dataset has multicollinearity. This matters because multicollinearity can make linear-regression coefficients unstable and hard to interpret (even if prediction error looks okay), so feature importance and signs may be unreliable. It guides model setup by suggesting you should drop/merge redundant variables (e.g., avoid using Total together with all sectors) or use regularization like Ridge/Elastic Net to stabilize estimates when predictors are highly correlated.



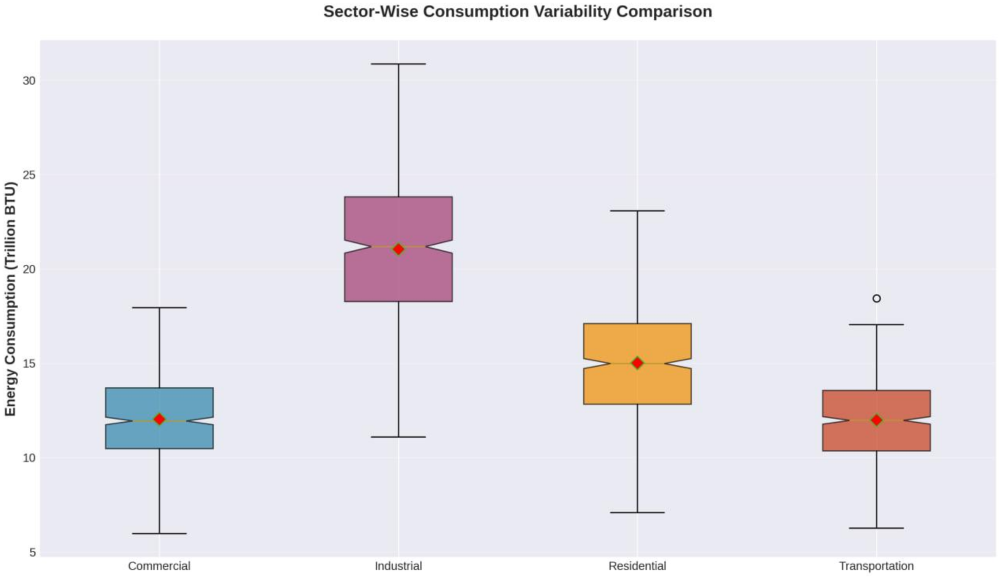
10. **Energy Consumption Heatmap: Year × Month:**
This energy consumption heatmap (year × month) shows color intensity encoding consumption level (darker/redder = higher), revealing long-term changes (overall intensification across later years) and recurring seasonal structure (certain months consistently higher/lower across many years). It matters because it motivates time features (month-of-year), seasonal components (12-month cycle), and possibly year/trend terms; it also helps spot unusual years/months for outlier handling and robustness checks.



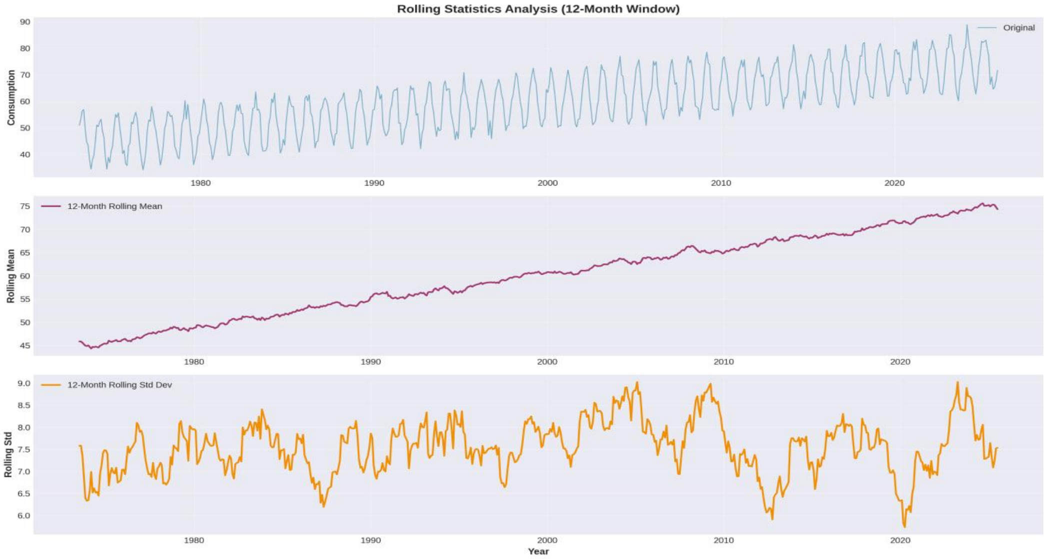
11. **Sector-Wise Energy Consumption Comparison (1973–2025):**
    This line plot shows annual energy consumption trends (Trillion Btu) for four sectors (Commercial, Industrial, Residential, Transportation) over 1973–2025, with each sector in a distinct color and overlaid to compare relative scales and patterns. It matters because it reveals overall upward trends across all sectors (consistent with economic growth and population increases), similar seasonal/yearly fluctuations (e.g., peaks/dips repeating), and relative sector sizes (Industrial/Residential often largest, Transportation steady). It guides model setup by motivating year-based trend features (e.g., linear time trend), sector dummies if multi-target, and decomposition (trend + seasonal + residuals) for forecasting; it also highlights the need for long-term train/test splits to capture these multi-decade trends accurately.



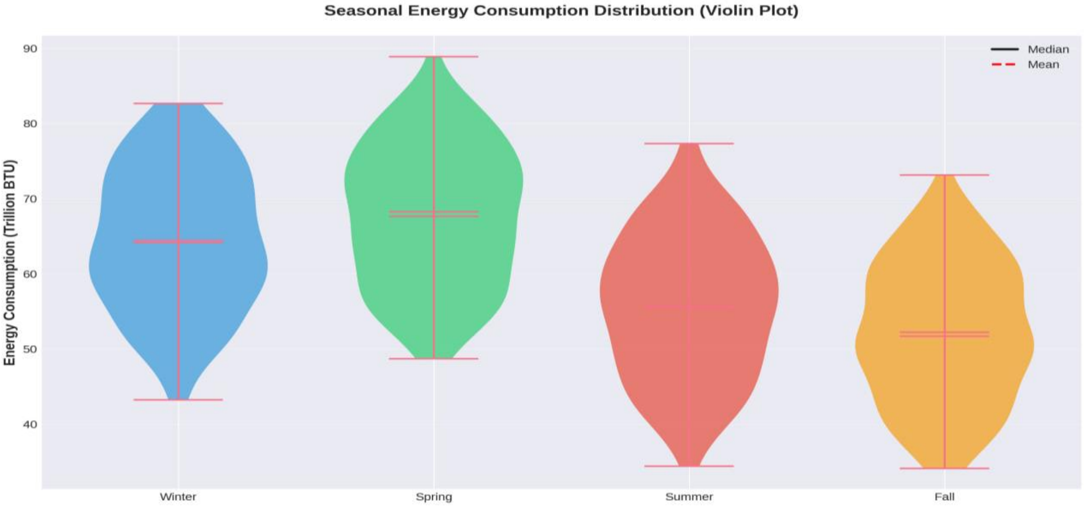
12. **Sector-Wise Consumption Variability Comparison:**
    This box plot compares the distribution of energy consumption (Trillion Btu) across four sectors, showing medians (central lines), interquartile ranges (boxes), whiskers (typical range), and outliers (dots) for each. It matters because Transportation has the lowest variability (tight box, no outliers, most predictable), while others vary more (wider boxes, outliers), suggesting sector-specific noise levels and potential heteroscedasticity if modeling residuals. It guides model setup by recommending log-transforms for skewed/high-variance sectors (e.g., Residential), robust scalers per sector, or separate models per sector; it also flags Transportation as easiest to predict accurately.



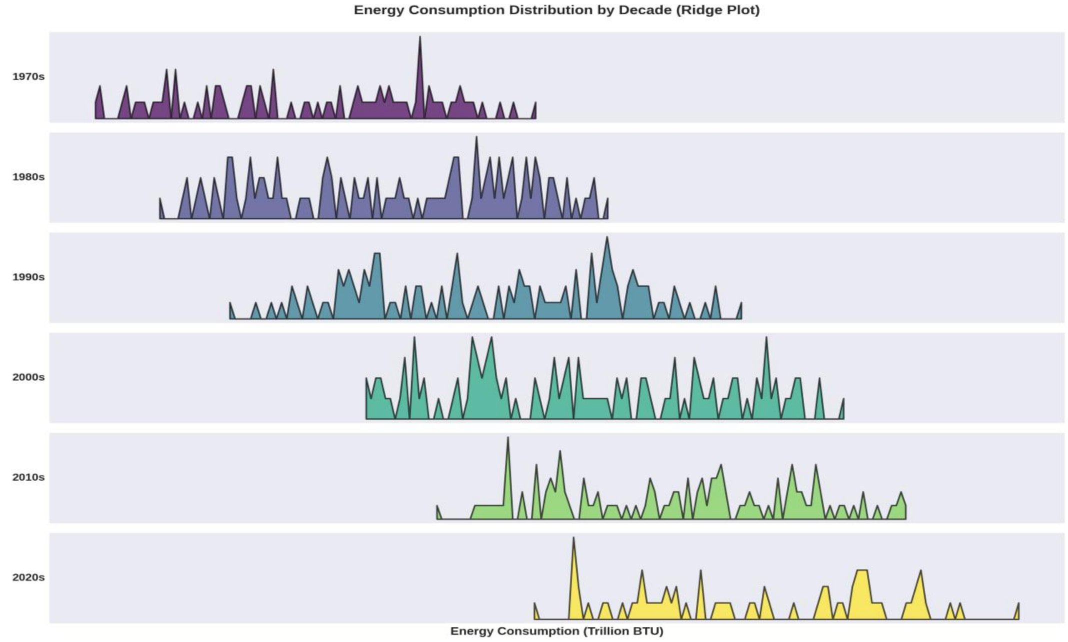
13. **Rolling Statistics (12-Month Window):**
    This figure plots three rolling statistics (12-month window) over time: original series (top/blue wiggly), rolling mean (middle/magenta, steadily rising), and rolling std (bottom/orange, fluctuating but stable level). It matters because the upward-trending rolling mean shows **non-stationarity** due to trend (mean changes over time), while stable rolling std suggests constant variance; non-stationarity violates assumptions of many models (e.g., ARIMA) and can bias forecasts. It guides model setup by recommending differencing or detrending (subtract rolling mean) to achieve stationarity, then validating with ADF test (p<0.05); also justifies using trend-aware models like Prophet or including time/year features in ML regressions.



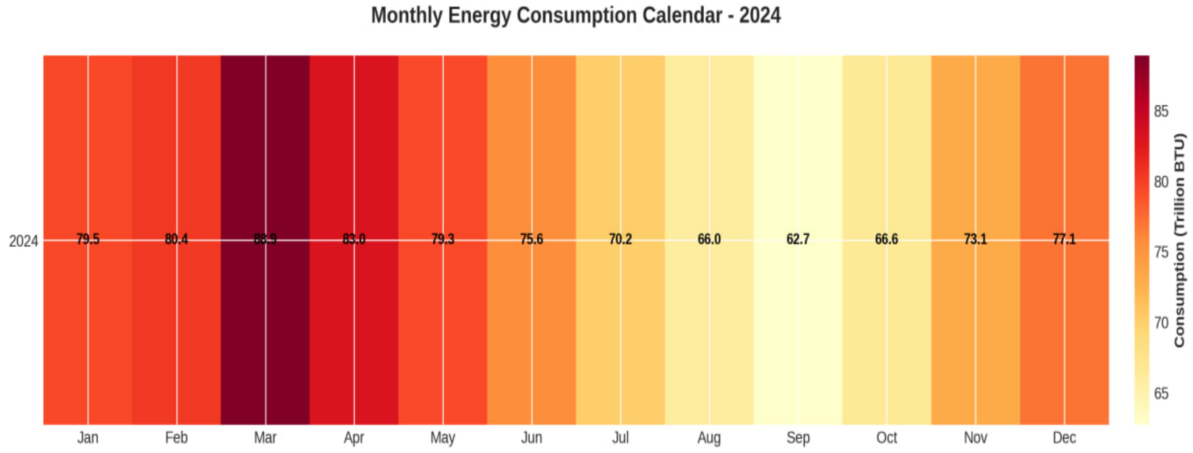
14. **Seasonal Energy Consumption Violin Plot:**
    This violin plot shows energy consumption distributions (Trillion Btu) across four seasons (Winter/blue, Spring/green, Summer/red, Fall/orange), with violin shapes revealing density (wider = higher probability), embedded box elements (median=red line, IQR=thick bar), and whiskers for range. It matters because it highlights **seasonality strength**: Winter has highest median/variability (widest violin, tallest), Spring lowest (narrowest/skinnier), confirming strong quarterly patterns critical for time series models; ignoring this leads to poor forecasts. It guides modeling by justifying **seasonal features** (dummy vars for seasons, Fourier terms, or seasonal ARIMA/SARIMA), cyclical encoding, or Prophet's seasonal components; test by adding seasonal lags and checking ACF at lags 3/12.



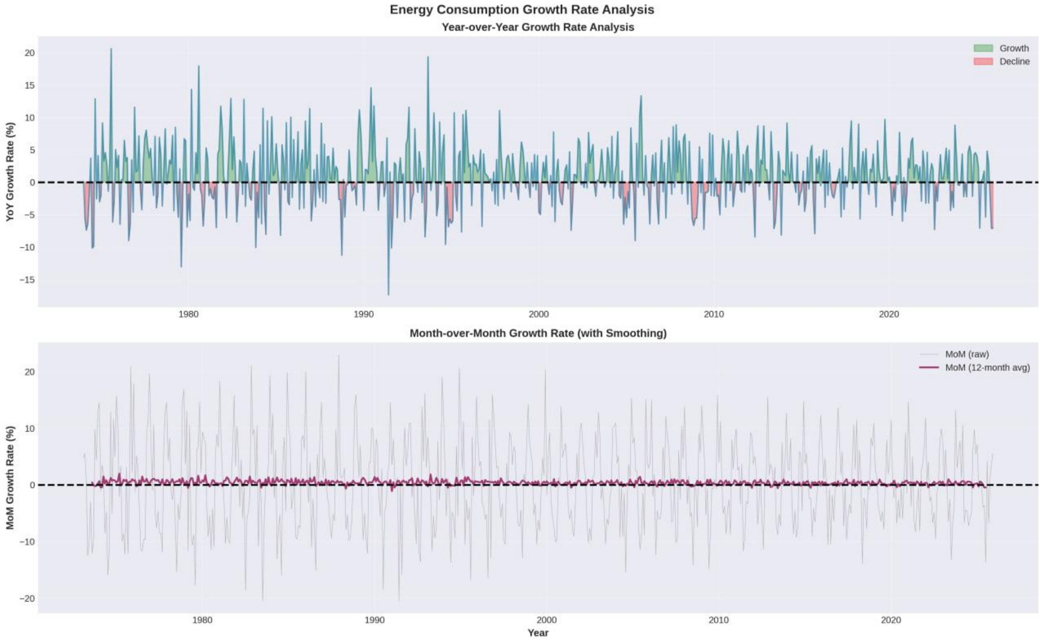
15. **Energy Consumption Distribution by Decade (Ridge Plot):**
    This ridgeline (joyplot) stacks density curves for energy consumption (Trillion Btu) across decades (~1970s/dark blue top, to 2020s/yellow bottom), showing peaks (mode/high density), spread, and shifts in shape/position. It matters because it visualizes **long-term trend**: distributions shift right/upward (higher values/modes over time), confirming secular growth; recent decades (2010s/2020s) have tighter/narrower ridges (less variability post-efficiency gains?). It guides modeling by validating **trend inclusion** (linear/polynomial time features, decade dummies), checking for changing variance (e.g., GARCH for heteroscedasticity), and splitting train/test by decade to avoid leakage; boosts accuracy in long-horizon forecasts.



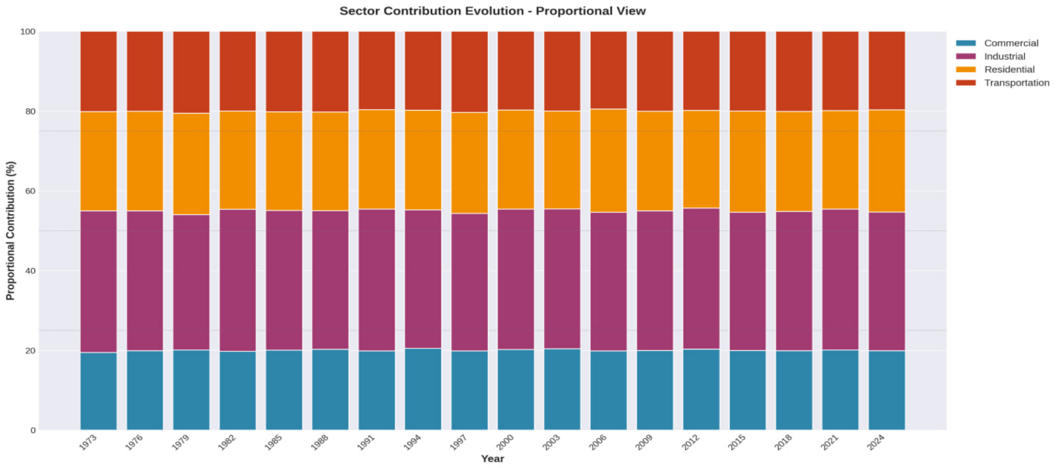
16. **Monthly Energy Consumption Calendar Heatmap – 2024:**
    This calendar heatmap visualizes 2024 monthly energy consumption (Trillion Btu) as a color gradient (dark red=high ~70+, light yellow=low ~60), with horizontal bars per month (Jan left to Dec right) showing daily values. It matters because it reveals **daily/seasonal patterns**: consistently higher in winter months (Jan/Feb dark red blocks), dips mid-year (yellow in summer?), and spots anomalies (e.g., specific high/low days for events/outliers). It guides modeling by confirming intra-month **heterogeneity** (encode day-of-month, weekday dummies, holiday flags), anomaly detection (isolation forest on residuals), and short-term forecasting needs (e.g., XGBoost with lag-7/30 + calendar features).



17. **Energy Consumption Growth Rate Analysis:**
    This dual plot shows Year-over-Year (YoY) growth rates (top: raw green bars ±10% volatility, red declines) and Month-over-Month (MoM) smoothed (bottom: purple near-flat ~0-2% with MA filter). It matters because YoY reveals **cyclical shocks** (e.g., recessions/spikes like 2008 dip, COVID drop), while smoothed MoM confirms **stable low growth** post-smoothing, aiding trend decomposition (growth ~1-2%/year). It guides modeling by recommending **growth/differenced targets** (predict % change instead of levels for stationarity), cyclical features (e.g., GDP lags for YoY), or multiplicative decomposition; improves Ridge/Lasso stability on volatile raw series.



18. **Sector Contribution Evolution (Proportional View):**
    This 100% stacked bar chart shows proportional contributions (%) of four sectors (blue=Commercial, orange=Industrial?, purple=Residential, red=Transportation?) to total US energy consumption annually from ~1973 (left) to ~2025 (right). It matters because Industrial (orange) dominates consistently (~40-50%, thickest stack), while Transportation (red top) stable minor (~20%), revealing **structural stability** with minor shifts (e.g., Commercial growth?); total=100% hides absolute growth but highlights composition invariance. It guides modeling by justifying **sector-specific models** or interaction terms (e.g., sector * time), normalization to proportions for multi-sector totals, and ensemble weighting by historical share; aids interpretability in Ridge regressions.


Sector Contribution Evolution - Proportional View
Proportional Contribution (%)
Commercial
Industrial
Residential
Transportation
Year