Forecasting
Group 8
US Energy Consumption

Introduction:

Purpose: Energy consumption is essential for our modern lifestyle. This project analyzes historical U.S. energy data from the EIA to identify trends and patterns in commercial, and industrial sectors. By examining key variables, we aim to provide valuable insights for policymakers and energy providers. These insights will help optimize energy distribution, reduce waste, and promote sustainable energy practices.

Research Question: Key factors driving energy consumption in the commercial and industrial sectors, and which forecasting methods are most suitable for predicting future trends.

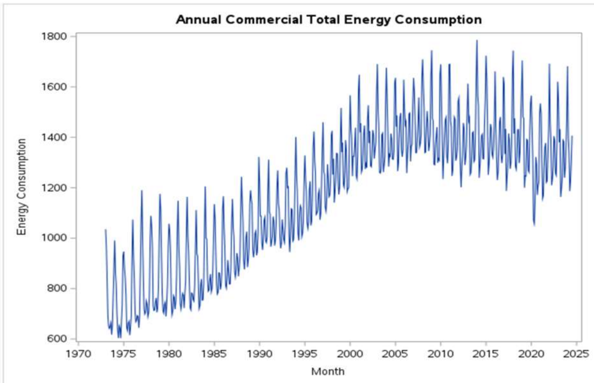
Data Collection: The dataset is derived from the U.S. Energy Information Administration (EIA), covering monthly records from January 1973 to July 2024. The dataset includes a wide range of variables, such as energy sales, primary consumption, end-use energy consumption, and total electricity losses ensuring adequate data for reliable analysis.

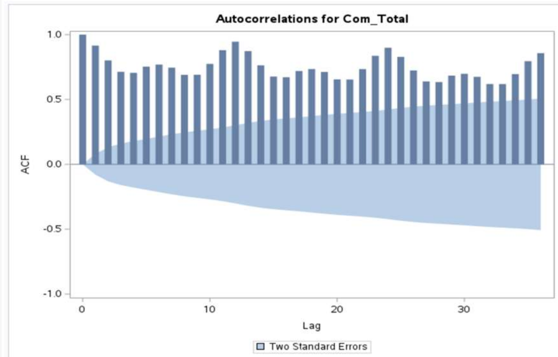
Original Name	Shortened Name
Date	Date
Primary Energy Consumed by the Commercial Sector	Com_Primary
Electricity Sales to Ultimate Customers in the Commercial Sector	Com_Elec_Sales
End-Use Energy Consumed by the Commercial Sector	Com_End_Use
Commercial Sector Electrical System Energy Losses	Com_Elec_Losses
Total Energy Consumed by the Commercial Sector	Com_Total
Primary Energy Consumed by the Industrial Sector	Ind_Primary
Electricity Sales to Ultimate Customers in the Industrial Sector	Ind_Elec_Sales
End-Use Energy Consumed by the Industrial Sector	Ind_End_Use
Industrial Sector Electrical System Energy Losses	Ind_Elec_Losses
Total Energy Consumed by the Industrial Sector	Ind_Total

Data Preparation: After importing the dataset, it underwent cleaning with Standardizing the "Date" column into SAS-compatible formats (MONYY7) and the dataset is partitioned into an 80% training set and a 20% validation set.

Data Partitioning: The dataset is partitioned into an 80% training set and a 20% validation set, consisting of 619 records for model building. The dependent variable considered was the Total energy consumption for the respective sectors, with a focus on predicting.

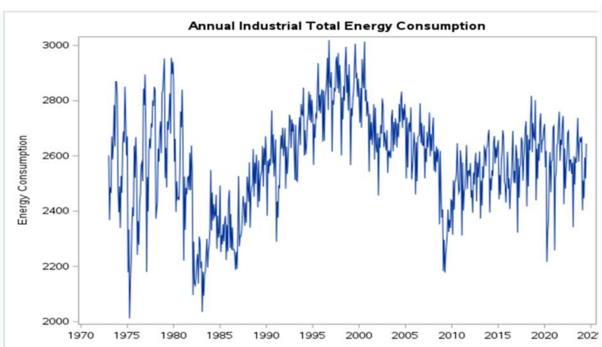
Graphs and Summary Statistics:
Commercial Sector:

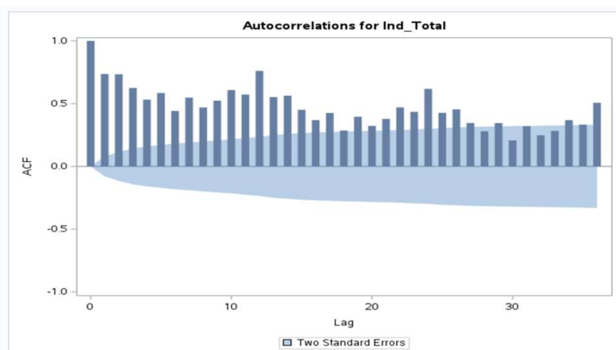




- The time series plot shows an overall positive trend and can recognize the seasonal pattern in energy consumption for the commercial sector.
- According to the ACF plot, the autocorrelations are not declining quickly towards zero which indicates a trend component. In addition, the autocorrelations are higher at lags 1, 12, 24, and 36 which indicates the presence of a seasonal component.

Industrial Sector:





The time series plot shows a non-linear trend and can recognize the seasonal pattern in energy consumption for the Industrial sector.

According to the ACF plot, the autocorrelations are not declining quickly towards zero which indicates a trend component. In addition, the autocorrelations are higher at lags 1, 12, 24, and 36, which indicates the presence of a seasonal component.

Models & Methods used for both the sectors:

1. Holt Winter's exponential Smoothing

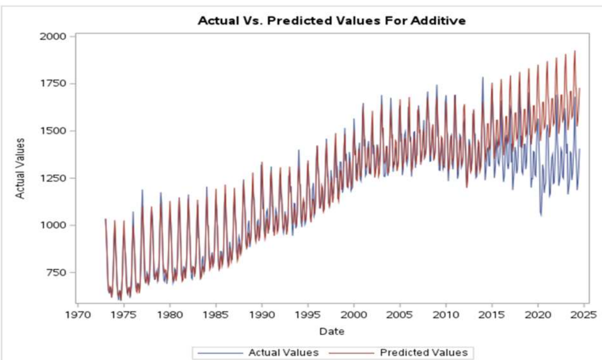
2. Multiple Linear Regression & Non linear Regression

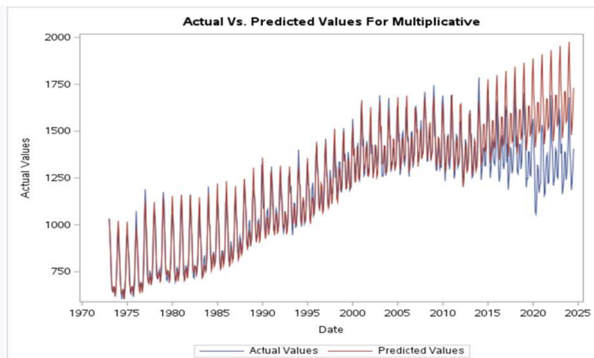
    - using dummy variables
    - using deseasonalizing and Reseasonalising

3. ARIMA

Commercial Sector:

Model 1: Holt's-Winter's Exponential Model:





	Additive	Multiplicative
MAPE fit	2.74	2.81
MAE fit	30.86	31.70
MSE fit	1637.98	1750.09
MAPE Acc	16.58	16.10
MAE Acc	220.48	215.56
MSE Acc	58329.23	56000.01

Winters Method (Additive) Parameter Estimates

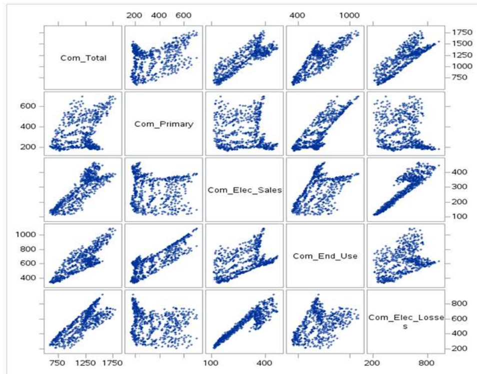
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|
Level Weight	0.25596	0.01999	12.81	<.0001
Trend Weight	0.0010000	0.0039067	0.26	0.7981
Seasonal Weight	0.31197	0.02643	11.80	<.0001

Winters Method (Multiplicative) Parameter Estimates

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|
Level Weight	0.24793	0.01951	12.71	<.0001
Trend Weight	0.0010000	0.0044528	0.22	0.8224
Seasonal Weight	0.42842	0.03047	14.06	<.0001

- Level weight for both additive and multiplicative models is closer to 0, which means less weight is assigned to the most recent observation.
- Trend weight shows the slope is hardly changing for both the models.
- The seasonal component is not changing drastically.
- Error values for Multiplicative accuracy are less, suggesting a better model.

Correlation Matrix:



- Except the Com_Primary, com_Elec_Losses, all other variables look relatively linear.

Model 2: Multiple Regression Actual vs Predicted plot: Linear vs Nonlinear using dummy variables:

Linear Model:

Root MSE	70.09413	R-Square	0.9384
Dependent Mean	1124.94488	Adj R-Sq	0.9368
Coeff Var	6.23089		

Parameter Estimates								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation	95% Confidence Limits	
Intercept	1	853.19311	12.27645	69.50	<.0001	0	829.07114	877.31509
t	1	1.71370	0.02205	77.72	<.0001	1.00025	1.67037	1.75703
jan	1	90.04782	15.38917	5.85	<.0001	1.85272	59.80967	120.28597
feb	1	-49.23326	15.38903	-3.20	0.0015	1.85268	-79.47113	-18.99539
mar	1	-97.03646	15.38892	-6.31	<.0001	1.85266	-127.27411	-66.79881
apr	1	-250.86494	15.48221	-16.20	<.0001	1.83458	-281.28589	-220.44398
may	1	-266.55493	15.48197	-17.22	<.0001	1.83453	-296.97542	-236.13443
jun	1	-233.96256	15.48177	-15.11	<.0001	1.83448	-264.38265	-203.54246
jul	1	-162.14221	15.48160	-10.47	<.0001	1.83444	-192.56196	-131.72245
aug	1	-162.81696	15.48145	-10.52	<.0001	1.83440	-193.23643	-132.39748
sep	1	-267.05017	15.48134	-17.25	<.0001	1.83438	-297.46943	-236.63091
oct	1	-252.99177	15.48127	-16.34	<.0001	1.83436	-283.41088	-222.57266
nov	1	-196.18664	15.48122	-12.67	<.0001	1.83435	-226.60565	-165.76763

The MEANS Procedure

Variable	Mean
mape_fit	4.894
mae_fit	54.418
mse_fit	4784.153
mape_acc	136824.428
mae_acc	282.661
mse_acc	70560.753

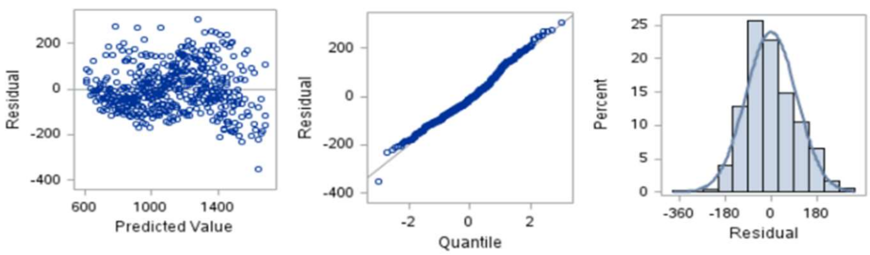
Equation of line fit: y=853.19311+1.71370t

The slope of **1.71370** indicates that for each one-unit increase in t, the total consumption increases by **1.71370** on average.

Model Evaluation:

- Multiple linear regression was applied after removing variables with a VIF greater than 10 to mitigate multicollinearity, leading to only t variable with positive slope making the model logical.
- The slope coefficients were statistically significant, with p-values less than alpha.
- The overall model demonstrated statistical significance.
- However, the adjusted R^2 of 93.68% indicates a good model fit.
- The nonlinear regression model is logical as the sign of the slope aligns with expectations. The slope coefficients are statistically significant, with p-values less than alpha. The model itself is statistically significant. An adjusted R^2 value of 93.68% suggests a strong model fit, and there is no evidence of multicollinearity.

Model Assumption:



Durbin-Watson D	0.915
Pr < DW	<.0001
Pr > DW	1.0000
Number of Observations	495
1st Order Autocorrelation	0.536

- For normality assumptions, the histogram looks bell shaped symmetric, so the assumption is true.
- For the constant variance assumption, the scatter plot does not show a pattern, so the assumption is true
- For the independence assumption, p-values of the DW test is less than alpha so there is serial correlation. The assumption is not true.

Non-Linear:

Root MSE	63.17788	R-Square	0.9500
Dependent Mean	1124.94488	Adj R-Sq	0.9487
Coeff Var	5.61609		

Parameter Estimates								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation	95% Confidence Limits	
Intercept	1	813.14171	11.69274	69.54	<.0001	0	790.16655	836.11686
t	1	2.01652	0.03481	57.93	<.0001	3.06789	1.94813	2.08492
t5	1	-7.0224E-12	6.62648E-13	-10.60	<.0001	3.06872	-8.3245E-12	-5.7204E-12
jan	1	91.12732	13.87108	6.57	<.0001	1.85282	63.87192	118.38273
feb	1	-48.02354	13.87105	-3.46	0.0006	1.85281	-75.27888	-20.76820
mar	1	-95.69218	13.87106	-6.90	<.0001	1.85281	-122.94754	-68.43681
apr	1	-251.80071	13.95484	-18.04	<.0001	1.83466	-279.22070	-224.38073
may	1	-267.38849	13.95457	-19.16	<.0001	1.83458	-294.80794	-239.96903
jun	1	-234.68975	13.95434	-16.82	<.0001	1.83452	-262.10874	-207.27076
jul	1	-162.75885	13.95413	-11.66	<.0001	1.83447	-190.17744	-135.34026
aug	1	-163.31884	13.95397	-11.70	<.0001	1.83442	-190.73710	-135.90058
sep	1	-267.43304	13.95383	-19.17	<.0001	1.83439	-294.85104	-240.01504
oct	1	-253.25135	13.95374	-18.15	<.0001	1.83436	-280.66916	-225.83354
nov	1	-196.31861	13.95368	-14.07	<.0001	1.83435	-223.73630	-168.90091

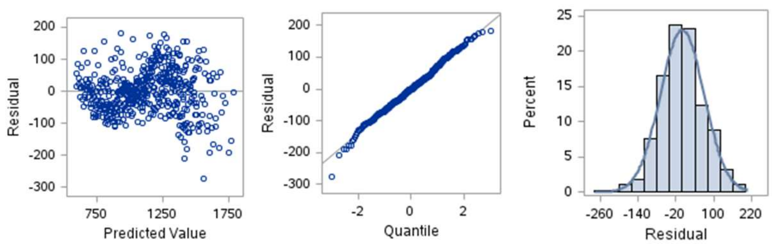
The MEANS Procedure

Variable	Mean
mape_fit	4.738
mae_fit	50.597
mse_fit	3878.555
mape_acc	136844.127
mae_acc	56.733
mse_acc	5514.759

Model Evaluation:

● Multiple linear regression was applied after removing variables with a VIF greater than 10 to mitigate multicollinearity, leading to only t variable with positive slope making the model logical.
● The slope coefficients were statistically significant, with p-values less than alpha.
● The overall model demonstrated statistical significance.
● However, the adjusted R^2 of 94.87% indicates a good model fit.
● The nonlinear regression model is logical as the sign of the slope aligns with expectations. The slope coefficients are statistically significant, with p-values less than alpha. The model itself is statistically significant. An adjusted R^2 value of 93.68% suggests a strong model fit, and there is no evidence of multicollinearity.

Model Assumption:



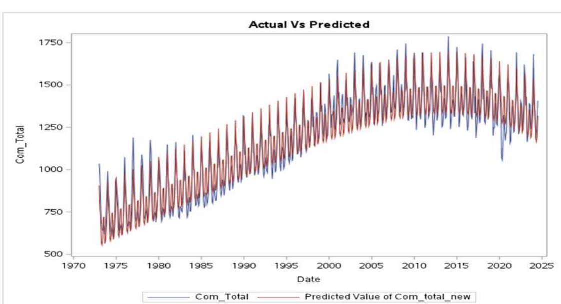
The REG Procedure
Model: MODEL1
Dependent Variable: Com_total_new

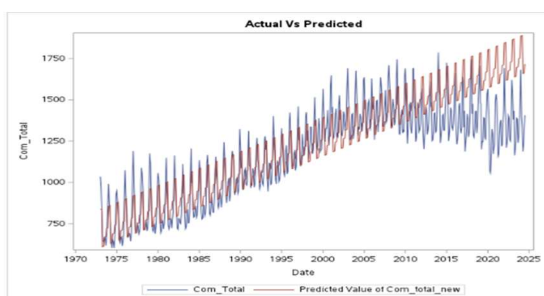
Durbin-Watson D	0.428
Pr < DW	<.0001
Pr > DW	1.0000
Number of Observations	495
1st Order Autocorrelation	0.783

Normality: True; the histogram displays a bell-shaped curve, and the QQ plot shows data points aligning closely with the diagonal, indicating that the residuals are normally distributed.

Equal Variance: True; the residuals versus predicted values plot shows no discernible pattern, confirming homoscedasticity.

Independence: False; the p-value from the Durbin-Watson test is less than the significance level (alpha), suggesting serial or positive correlation and a violation of the independence assumption.

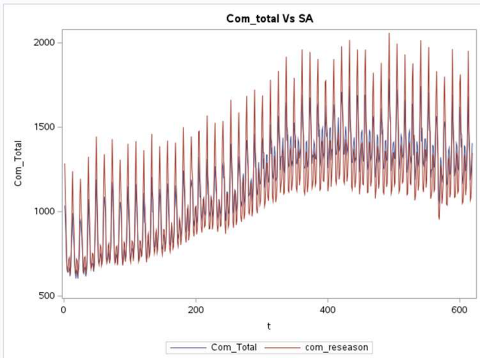


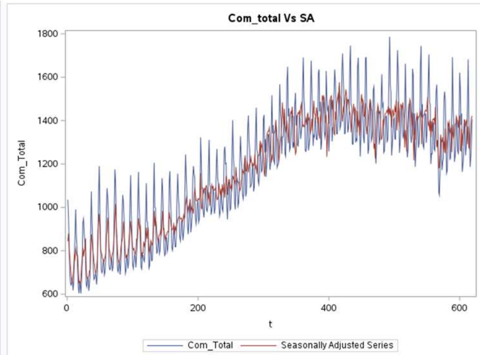


Linear Model Vs Non-Linear Model:

● The non-linear model provides better forecasting accuracy and fits the data more effectively, as reflected by lower error metrics and visually closer alignment between actual and predicted values.

Model 3: Deseasonalized & Reseasonalized Method:



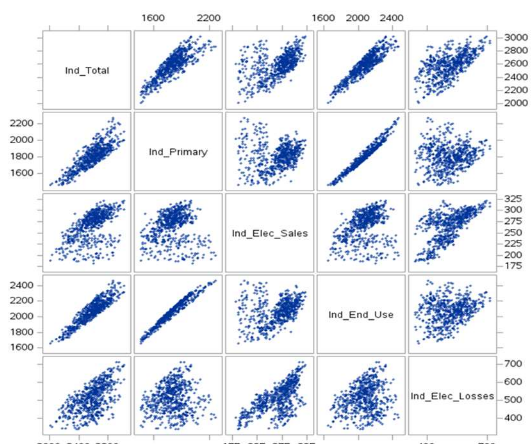


Variable	N	Mean
mape_fit	495	7.8105718
mape_acc	124	8.4383558
mae_fit	495	89.7277023
mae_acc	124	115.9162690
mse_fit	495	12042.96
mse_acc	124	16417.56

- The original time series exhibiting a clear upward trend and pronounced seasonal patterns.
- The deseasonalized series matches the original series closely, indicating that the seasonal component was correctly identified and removed.
- However, the error metrics did a somewhat good job in giving accuracy.

Industrial Sector:

Correlation Matrix



The variables *Ind_Primary*, *Ind_Elec_Sales*, and *Ind_End_Use* are likely significant contributors to *Ind_Total* (total energy consumption). While the correlation matrix reveals linear relationships among these variables, the time series trends suggest the presence of nonlinear patterns. A combination of linear and nonlinear models can be explored to identify the most suitable approach for understanding energy consumption in the industrial sector.

Model Evaluation: Multiple linear Regression

Multiple Non- linear Regression

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	3935491	327958	10.19	<.0001
Error	482	15518593	32196		
Corrected Total	494	19454085			

Root MSE	179.43314	R-Square	0.2023
Dependent Mean	2582.49336	Adj R-Sq	0.1824
Coeff Var	6.94806		

Parameter Estimates								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation	95% Confidence Limits	
Intercept	1	2595.30888	31.42634	82.58	<.0001	0	2533.55932	2657.05843
t	1	0.27317	0.05645	4.84	<.0001	1.00025	0.16226	0.38408
jan	1	-8.56652	39.39456	-0.22	0.8279	1.85272	-85.97281	68.83977
feb	1	-248.91200	39.39420	-6.32	<.0001	1.85268	-326.31758	-171.50642
mar	1	-100.20705	39.39391	-2.54	0.0113	1.85266	-177.61207	-22.80203
apr	1	-204.45058	39.63272	-5.16	<.0001	1.83458	-282.32484	-126.57633
may	1	-98.09997	39.63212	-2.48	0.0137	1.83453	-175.97304	-20.22690
jun	1	-107.73190	39.63160	-2.72	0.0068	1.83448	-185.60394	-29.85985
jul	1	-46.39014	39.63116	-1.17	0.2424	1.83444	-124.26131	31.48104
aug	1	10.83679	39.63079	0.27	0.7846	1.83440	-67.03367	88.70726
sep	1	-113.91394	39.63051	-2.87	0.0042	1.83438	-191.78385	-36.04403
oct	1	9.98875	39.63031	0.25	0.8011	1.83436	-67.88076	87.85827
nov	1	-56.45776	39.63019	-1.42	0.1549	1.83435	-134.32704	21.41152

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	16048619	1234509	174.37	<.0001
Error	481	3405466	7079.97068		
Corrected Total	494	19454085			

Root MSE	84.14256	R-Square	0.8249
Dependent Mean	2582.49336	Adj R-Sq	0.8202
Coeff Var	3.25819		

Parameter Estimates									
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	431.26645	54.01738	7.98	<.0001	0	325.12725	537.40565
Ind_Primary	Ind_Primary	1	1.16801	0.02751	42.45	<.0001	1.24187	1.11395	1.22207
t3		1	0.00000129	1.115614E-7	11.60	<.0001	1.03510	0.00000108	0.00000151
jan		1	-27.63903	18.47794	-1.50	0.1354	1.85361	-63.94648	8.66842
feb		1	-49.33251	19.06761	-2.59	0.0100	1.97380	-86.79861	-11.86641
mar		1	-13.76115	18.58705	-0.74	0.4594	1.87556	-50.28298	22.76069
apr		1	-13.35622	19.13007	-0.70	0.4854	1.94373	-50.94505	24.23261
may		1	48.09225	18.90636	2.54	0.0113	1.89854	10.94299	85.24152
jun		1	69.58149	19.05346	3.65	0.0003	1.92819	32.14318	107.01979
jul		1	80.23700	18.82546	4.26	<.0001	1.88232	43.24670	117.22729
aug		1	79.83477	18.65662	4.28	<.0001	1.84871	43.17624	116.49331
sep		1	6.34832	18.80061	0.34	0.7358	1.87736	-30.59316	43.28980
oct		1	9.97457	18.58407	0.54	0.5917	1.83436	-26.54141	46.49056
nov		1	6.24206	18.64291	0.33	0.7379	1.84599	-30.38954	42.87367

Model Equation of Line Fit: Linear regression

y=2595.30888+0.27317t

The slope of **0.27317** indicates that for each one-unit increase in t, the total consumption increases by **0.27317** on average.

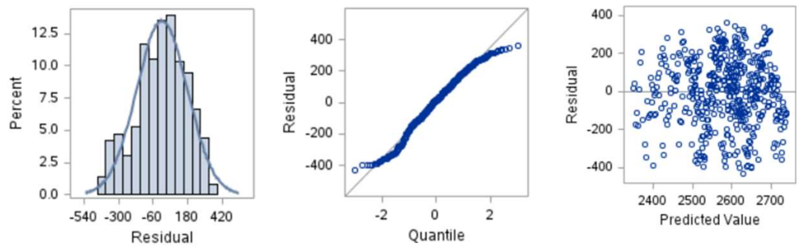
Model Equation of Line Fit: Non-Linear regression

y=431.26945+1.16801(Ind_Primary)+0.00000129(t3)

The slope of 0.00000129 indicates that for each one-unit increase in t3, the total consumption increases by 0.00000129 on average. For **Ind_Primary** indicates that for each one-unit increase in **Ind_Primary**, the total consumption increases by **1.16801** on average.

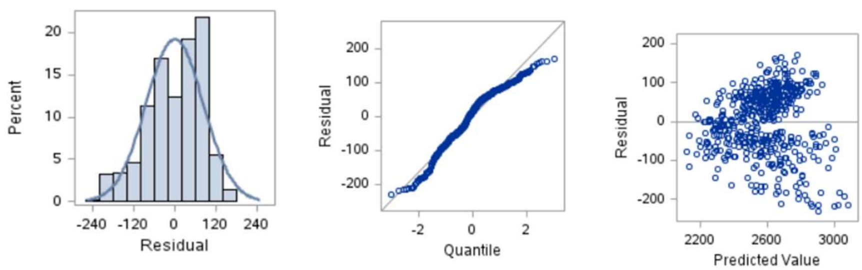
- Multiple linear regression was applied after removing variables with a VIF greater than 10 to mitigate multicollinearity, leading to only t variable with positive slope making the model logical.
- The slope coefficients were statistically significant, with p-values less than alpha.
- The overall model demonstrated statistical significance.
- However, the adjusted R^2 of 0.18% indicates a poor fit for the model.
- The nonlinear regression model is logical as the sign of the slope aligns with expectations. The slope coefficients are statistically significant, with p-values less than alpha. The model itself is statistically significant. An adjusted R^2 value of 82.02% suggests a strong model fit, and there is no evidence of multicollinearity.

Model Assumption: Linear regression



Durbin-Watson D	0.170
Pr < DW	<.0001
Pr > DW	1.0000
Number of Observations	495
1st Order Autocorrelation	0.915

Model assumption: Non-linear regression



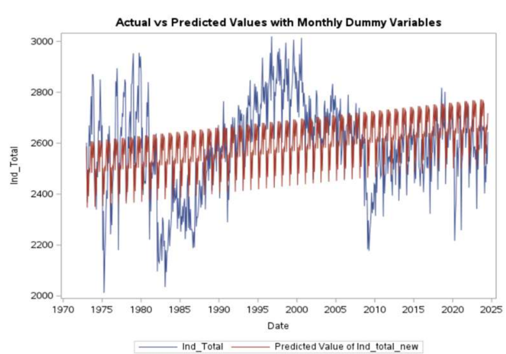
Durbin-Watson D	0.074
Pr < DW	<.0001
Pr > DW	1.0000
Number of Observations	495
1st Order Autocorrelation	0.957

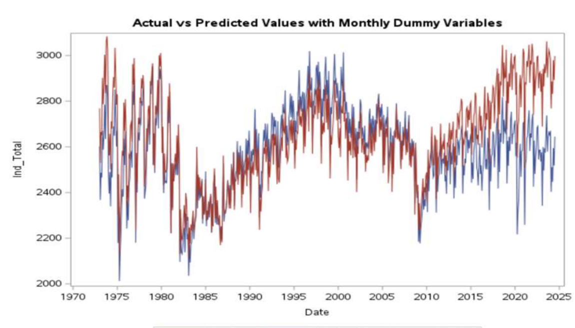
Normality: True; the histogram displays a bell-shaped curve, and the QQ plot shows data points aligning closely with the diagonal, indicating that the residuals are normally distributed.

Equal Variance: True; the residuals versus predicted values plot shows no discernible pattern, confirming homoscedasticity.

Independence: False; the p-value from the Durbin-Watson test is less than the significance level (alpha), suggesting serial or positive correlation and a violation of the independence assumption.

Multiple Regression Actual vs Predicted plot: Linear vs Nonlinear using dummy variables



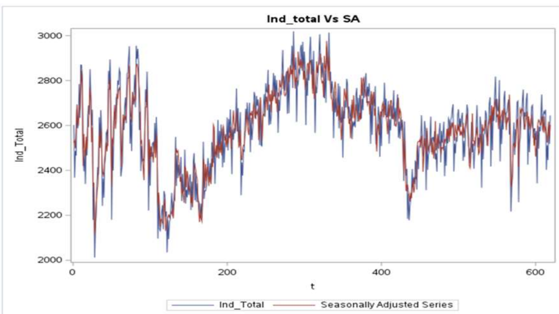


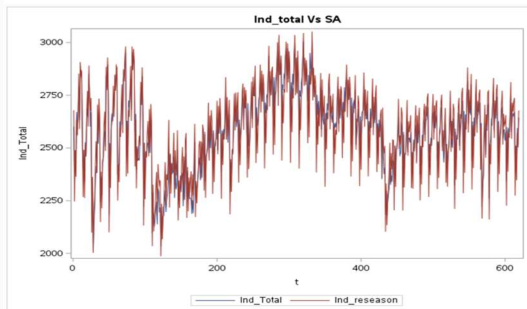
Variable	Mean
mape_fit	5.667
mae_fit	143.768
mse_fit	31350.693
mape_acc	258290.344
mae_acc	86.991
mse_acc	12062.022

Variable	Mean
mape_fit	2.666
mae_fit	69.365
mse_fit	6879.729
mape_acc	258283.676
mae_acc	256.224
mse_acc	71533.803

The nonlinear model effectively predicts values with only minor discrepancies observed toward the end. However, the presence of multicollinearity indicates that the model may be overfitting looking at the error metrics for nonlinear regression.

Deseasonalized & Reseasonalized Method:





The MEANS Procedure

Variable	N	Mean
mape_fit	495	2.3242671
mape_acc	124	2.1239893
mae_fit	495	59.5757882
mae_acc	124	54.8492350
mse_fit	495	5390.05
mse_acc	124	4523.37

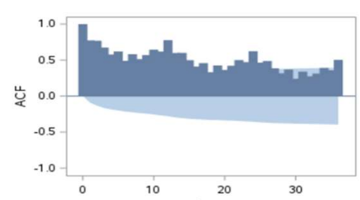
This method fails to simplify or filter out noise in the data. The residual plot clearly reflects even minor variations or noise. However, the **error metrics demonstrate strong performance in terms of accuracy.**

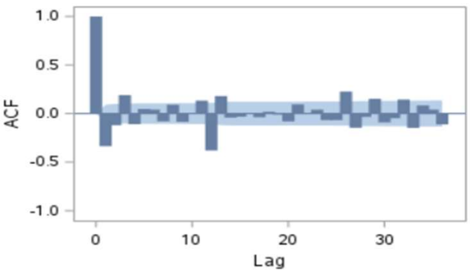
ARIMA Model: 2 times differencing

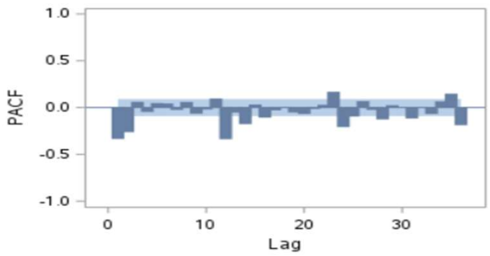
Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	1301.77	6	<.0001	0.774	0.769	0.675	0.577	0.620	0.489
12	2483.55	12	<.0001	0.582	0.513	0.571	0.644	0.619	0.777
18	3229.11	18	<.0001	0.599	0.600	0.499	0.411	0.458	0.330
24	3933.03	24	<.0001	0.426	0.364	0.422	0.500	0.472	0.624
30	4404.58	30	<.0001	0.466	0.487	0.386	0.316	0.369	0.241
36	4843.97	36	<.0001	0.339	0.274	0.314	0.391	0.359	0.503

Name of Variable = arima_New	
Mean of Working Series	2582.493
Standard Deviation	198.2453
Number of Observations	495







Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	.	0	.	-0.002	0.002	0.012	-0.025	0.012	0.068
12	.	0	.	-0.053	-0.026	-0.039	-0.003	-0.010	0.000
18	.	0	.	0.000	-0.000	-0.041	-0.000	-0.070	-0.021
24	16.56	1	<.0001	-0.078	-0.081	0.036	0.041	0.000	0.002
30	16.72	7	0.0193	-0.006	0.003	-0.002	-0.001	-0.001	-0.016
36	20.42	13	0.0853	-0.001	-0.001	0.000	0.084	0.002	-0.006
42	27.00	19	0.1047	0.032	-0.059	0.033	0.067	-0.047	0.011
48	40.46	25	0.0262	0.033	-0.102	-0.108	-0.000	0.046	0.005

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	.	0	.	-0.001	-0.000	0.011	-0.030	0.027	0.056
12	.	0	.	-0.051	-0.021	-0.048	-0.004	0.062	0.000
18	11.84	2	0.0027	0.000	-0.000	-0.031	-0.077	-0.051	-0.020
24	20.45	8	0.0088	-0.074	-0.073	0.031	0.073	-0.001	0.003
30	20.89	14	0.1044	0.012	0.003	-0.004	-0.001	-0.003	-0.026
36	31.93	20	0.0441	-0.095	0.036	-0.086	0.001	0.004	-0.059
42	38.24	26	0.0575	0.038	-0.055	0.027	0.070	-0.042	0.012
48	55.11	32	0.0068	0.050	-0.096	-0.121	0.013	0.045	0.056

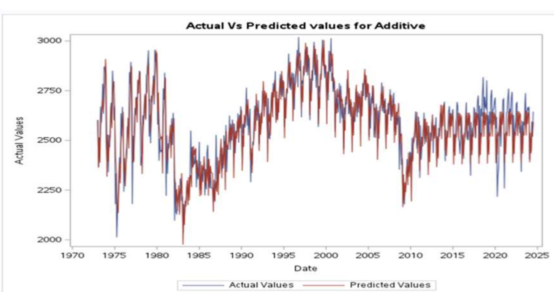
Models tried:

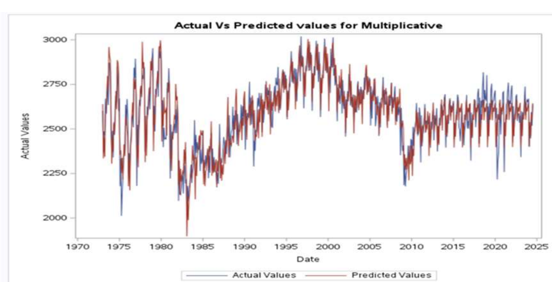
ARIMA(9,1,10)(3,1,1)

ARIMA(7,1,6)(2,1,1)

The residual analysis for both ARIMA(9,1,10)(3,1,1) and ARIMA(7,1,6)(2,1,1) reveals that the models are not performing optimally. This is evident from the presence of non-white noise residuals at multiple lags, indicating unaccounted patterns in the dataset.

Holt Winter's Exponential Model





Winters Method (Additive) Parameter Estimates				
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|
Level Weight	0.80590	0.03156	25.53	<.0001
Trend Weight	0.0010000	0.01054	0.09	0.9244
Seasonal Weight	0.0010000	0.03007	0.03	0.9735

Winters Method (Multiplicative) Parameter Estimates				
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|
Level Weight	0.35119	0.01783	19.69	<.0001
Trend Weight	0.0010000	0.0044048	0.23	0.8205
Seasonal Weight	0.58885	0.03592	16.39	<.0001

- Less weight is assigned to the most recent observation for multiplicative and more weight assigned for additive.
- Trend weight shows the slope is hardly changing for both the models.
- The seasonality component is moderately changing for multiplicative model.
- Error values for multiplicative accuracy are less, suggesting a better model.

	Additive	Multiplicative
MAPE fit	2.13	2.3
MAE fit	54.37	58.05
MSE fit	5167.9	5643.57
MAPE Acc	2.52	2.17
MAE Acc	65	55.6
MSE Acc	6382.22	5205.43

Model Fit Statistics:

- The additive model performs better on the training data (fit statistics) with lower MAE and MSE. However, its performance decreases when applied to new data (accuracy statistics), as seen from the higher MAPE, MAE, and MSE values.
- The multiplicative model performs slightly worse in the fit statistics but outperforms the additive model in accuracy metrics, particularly with a lower MAPE (2.17% vs. 2.52%) and significantly lower MAE (55.6 vs. 65). This indicates the multiplicative model generalizes better to test data.

Given these results, the multiplicative model is a better choice.

Conclusion:

Commercial Sector

	MAPE Fit	MAE Fit	MSE Fit	MAPE Acc	MAE Acc	MSE Acc
Holt Winter’s Exponential Model (Multiplicative)	2.81	31.7	1750.09	16.1	215.56	56000.01
Regression using Dummy variables Non-Linear	4.738	50.597	3878.555	136844.127	56.733	5514.759
Regression (linear) using De-Reseasonalization	7.81	89.727	12042.96	8.438	115.916	16417.56

- The Holt Winter’s exponential (Multiplicative) model performs well on the training set.

- The Multiple Linear Regression using deseasonalize method outperforms others on the validation set making it the most suitable Regression model for Forecasting the total Energy Consumption in Commercial Sector.

Industrial sector

	MAPE Fit	MAE Fit	MSE Fit	MAPE Acc	MAE Acc	MSE Acc
Holt Winter's Exponential Model (Multiplicative)	2.30	58.05	5643.57	2.17	55.60	5205.43
Regression using Dummy variables Non-Linear	2.67	69.37	6879.73	258283.68	256.22	71533.80
Regression(linear) using De-Reseasonalization	2.32	59.58	5390.05	2.12	54.85	4523.37

- The Holt Winters Exponential Model using dummy variables method performs well on the training set.
- The Multiple Linear Regression using deseasonalize method outperforms other on the validation set making it the most suitable Regression model for Forecasting the total Energy Consumption in Industrial Sector.

Key Findings

- Clear energy consumption trends with seasonal variations.
- Multiple Regression using deseasonalization method, is more suitable.
- Key drivers include energy sales, end-use, and primary consumption.

Shortcomings

- Potential biases.
- Serial correlation issues affect model independence.
- ARIMA models failed to produce reliable forecasts.

Suggestions

- Include additional variables for better analysis.
- Explorations for multicollinearity.
- Incorporate additional Weather-related variables like Temperature or Demographic & Economic variables like population growth, GDP growth rate.

References: https://www.eia.gov/totalenergy/data/monthly/

/* CODE For Commercial Sector*/

```
/* Importing Excel file into SAS - Monthly Data */
proc import out=energy_monthly
datafile="/home/u63735896/sasuser.v94/FORECASTING/Forecasting
Project/Energy_final.xlsx"dbms=xlsx replace;
run;
data energy_monthly;
   set energy_monthly;
   /* Convert character month and year to numeric if necessary */
   month_num = input(month, best12.);
   year_num = input(year, best12.);
   /* Convert year and month to a standard SAS date format */
   Date = mdy(month_num, 1, year_num);
   /* Format the Date column as "MONYY7." (e.g., "MAY23") */
   format Date monyy7.;
   *drop com_primary;
run;

/* Monthly data timeseries and acf plot */
proc sgplot data=energy_monthly;
   series x=date y=com_Total;
   title "Annual Commercial Total Energy Consumption";
   xaxis label="Month";
   yaxis label="Energy Consumption";
run;
proc timeseries data=energy_monthly plots=acf out=_null_;
        var com_total;
        corr acf/nlag=36;
run;

/* Holt-Winter's Exponential Smoothing */
/* Forecast accuracy using 124 observations as test set */
proc esm data=energy_monthly lead=124 back=124 outfor=energyout1 plot=forecasts
out=_null_ print=all;
        id Date interval=month;
        forecast com_Total/model=addwinters;
run;

proc esm data=energy_monthly lead=124 back=124 outfor=energyout2 plot=forecasts
out=_null_ print=all;
        id Date interval=month;
```

```
        forecast com_Total/model=winters;
run;
proc sgplot data=energyout1;
        series x=date y=actual;
        series x=date y=predict;
        title "Actual Vs. Predicted Values For Additive";
run;

proc sgplot data=energyout2;
        series x=date y=actual;
        series x=date y=predict;
        title "Actual Vs. Predicted Values For Multiplicative";
run;
```

/*Multiple Regression*/

```
data energy_monthly;
        set energy_monthly;
        t=_n_;
        Com_total_new=Com_total;
        if t>495 then Com_total_new=.;/* 124 observations for test set */
        month_num = month(date);
  jan = (month_num = 1);
  feb = (month_num = 2);
  mar = (month_num = 3);
  apr = (month_num = 4);
  may = (month_num = 5);
  jun = (month_num = 6);
  jul = (month_num = 7);
  aug = (month_num = 8);
  sep = (month_num = 9);
  oct = (month_num = 10);
  nov = (month_num = 11);

  /* December is the reference group, so no dummy variable for it */
run;

proc freq data=energy_monthly;
  table month_num jan feb mar apr may jun jul aug sep oct nov;
run;

proc reg data=energy_monthly;
  model Com_total_new = t jan feb mar apr may jun jul aug sep oct nov/ clb vif dwprob aic bic;
  output out=energy_monthly1 p=Com_total_predict r=Com_total_resid;
```

```
run;


proc sgplot data=energy_monthly1;
	series x=date y=Com_total;
	series x=date y=Com_total_predict;
run;

data energy_monthly1;
	set energy_monthly1;
	if t<=495 then
		do;
			mape_fit=abs(Com_total_resid/Com_Total_new)*100;
			mae_fit=abs(Com_total_resid);
			mse_fit=Com_total_resid**2;
		end;
	else if t>495 then
		do;
			mape_acc=abs(Com_Total-Com_total_predict/Com_Total)*100;
			mae_acc=abs(Com_Total-Com_Total_predict);
			mse_acc=(Res_Total-Com_total_predict)**2;
		end;
run;

proc means data=energy_monthly1 mean maxdec=3;
	var mape_fit mae_fit mse_fit mape_acc mae_acc mse_acc;
run;

/*___Non-Linear*/
data energy_monthly;
	set energy_monthly;
	t=_n_;
	Com_total_new=Com_total;
	if t>495 then Com_total_new=.;/* 124 observations for test set */
	month_num = month(date);
   jan = (month_num = 1);
   feb = (month_num = 2);
   mar = (month_num = 3);
   apr = (month_num = 4);
   may = (month_num = 5);
   jun = (month_num = 6);
   jul = (month_num = 7);
   aug = (month_num = 8);
   sep = (month_num = 9);
```

```
   oct = (month_num = 10);
   nov = (month_num = 11);

   /* December is the reference group, so no dummy variable for it */


t5=t*t*t*t*t;
run;

proc reg data=energy_monthly;
	model Com_total_new= t t5  jan feb mar apr may jun jul aug sep oct nov /dwprob vif
clb;
	output out=energy_monthly2 p=Com_total_predict r=Com_total_resid;
run;

proc sgplot data=energy_monthly2;
	series x=date y=Com_total;
	series x=date y=Com_total_predict;
	Title "Actual Vs Predicted";
run;

data energy_monthly2;
	set energy_monthly2;
	if t<=495 then
		do;
			mape_fit=abs(Com_total_resid/Com_total_new)*100;
			mae_fit=abs(Com_total_resid);
			mse_fit=Com_total_resid**2;
		end;
	else if t>495 then
		do;
			mape_acc=abs(Com_total-Com_total_predict/Com_total)*100;
			mae_acc=abs(Com_total-Com_total_predict);
			mse_acc=(Com_total-Com_total_predict)**2;
		end;
run;

proc means data=energy_monthly2 mean maxdec=3;
	var mape_fit mae_fit mse_fit mape_acc mae_acc mse_acc;
run;
```

/* Deseasonalize */

```
proc timeseries data=energy_monthly outdecomp=sa_com out=_null_;
```

```
	decomp sa;
	id date interval=month;
	var Com_total;
run;

data energy_monthly1;
	merge energy_Monthly sa_com;
	t=_n_;
	sa1=sa;
	if t>495 then sa1=.;
	si=Com_total/sa;
run;

proc sgplot data=energy_monthly1;
	series x=t y=Com_total;
	series x=t y=sa;
	title "Com_total Vs SA";
run;

proc reg data=energy_monthly1;
	model sa1=com_total;
	output out=com_out r=sa_resid p=sa_predict;
run;

data com_out;
	set com_out;
	com_reseason=si*sa_predict;
	if t<=495 then
		do;
			mape_fit=abs(sa_resid/sa1)*100;
			mae_fit=abs(sa_resid);
			mse_fit=sa_resid**2;
		end;
	else if t>495 then
		do;
			mape_acc=abs((sa-sa_predict)/sa)*100;
			mae_acc=abs(sa-sa_predict);
			mse_acc=(sa-sa_predict)**2;
		end;
run;

proc means data=com_out n mean;
	var mape_fit mape_acc mae_fit mae_acc mse_fit mse_acc;
run;
```

```
proc sgplot data=com_out;
	series x=t y=Com_total;
	series x=t y=com_reseason;
run;
```

/*Industrial Sector*/

```
/* Importing Excel file into SAS - Monthly Data */
proc import out=energy_monthly
   datafile="/home/u64002214/sasuser.v94/Energy_final.xlsx"
   dbms=xlsx replace;
run;

data energy_monthly;
   set energy_monthly;

   /* Convert character month and year to numeric if necessary */
   month_num = input(month, best12.);
   year_num = input(year, best12.);

   /* Convert year and month to a standard SAS date format */
   Date = mdy(month_num, 1, year_num);

   /* Format the Date column as "MONYY7." (e.g., "MAY23") */
   format Date monyy7.;
run;
/* Timeseries and acf plot */
proc sgplot data=energy_monthly;
   series x=date y=Ind_Total;
   title "Annual Industrial Total Energy Consumption";
   xaxis label="Month";
   yaxis label="Energy Consumption";
run;
proc timeseries data=energy_monthly plots=acf out=_null_;
	var Ind_total;
	corr acf/nlag=36;
run;

/* Holt-Winters Exponential Smoothing */
/* Forecast accuracy using 124 observations as test set, Winter Model Final */

proc esm data=energy_monthly lead=124 back=124 outfor=energyout1 plot=forecasts
out=_null_ print=all;
```

```
	id Date interval=month;
	forecast Ind_Total/model=addwinters;
run;

proc esm data=energy_monthly lead=124 back=124 outfor=energyout2 plot=forecasts
out=_null_ print=all;
	id Date interval=month;
	forecast Ind_Total/model=winters;
run;

proc sgplot data=energyout1;
	series x=date y=actual;
	series x=date y=predict;
	title'Actual Vs Predicted values for Additive';
run;

proc sgplot data=energyout2;
	series x=date y=actual;
	series x=date y=predict;
	title'Actual Vs Predicted values for Multiplicative';
run;

/* Multiple linear Regression with Seasonality */

proc sgscatter data=energy_monthly;
	matrix Ind_Total Ind_primary Ind_elec_sales Ind_End_use Ind_elec_losses;
run;
data energy_monthly;
	set energy_monthly;
	t=_n_;
	Ind_total_new=Ind_total;
	if t>495 then Ind_total_new=.;/* 124 observations for test set */
	month_num = month(date);
   jan = (month_num = 1); feb = (month_num = 2); mar = (month_num = 3); apr = (month_num
= 4);
   may = (month_num = 5); jun = (month_num = 6);  jul = (month_num = 7);aug = (month_num
= 8);
   sep = (month_num = 9); oct = (month_num = 10); nov = (month_num = 11);

   /* December is the reference group, so no dummy variable for it */
run;

proc freq data=energy_monthly;
   table month_num jan feb mar apr may jun jul aug sep oct nov;
```

```
run;

proc reg data=energy_monthly;
   model Ind_total_new =t jan feb mar apr may jun jul aug sep oct nov / clb vif dwprob aic bic;
   output out=energy_monthly1 p=Ind_total_predict r=Ind_total_resid;
run;

proc sgplot data=energy_monthly1;
   series x=date y=Ind_Total;
   series x=date y=Ind_total_predict;
   title "Actual vs Predicted Values with Monthly Dummy Variables";
run;

data energy_monthly1;
	set energy_monthly1;
	if t<=495 then
		do;
			mape_fit=abs(Ind_total_resid/Ind_Total_new)*100;
			mae_fit=abs(Ind_total_resid);
			mse_fit=Ind_total_resid**2;
		end;
	else if t>495 then
		do;
			mape_acc=abs(Ind_Total-Ind_total_predict/Ind_Total)*100;
			mae_acc=abs(Ind_Total-Ind_Total_predict);
			mse_acc=(Ind_Total-Ind_total_predict)**2;
		end;
run;

proc means data=energy_monthly1 mean maxdec=3;
	var mape_fit mae_fit mse_fit mape_acc mae_acc mse_acc;
run;

/* Multiple Regression - Non linear */
data energy_monthly;
	set energy_monthly;
	t=_n_;
	Ind_total_new=Ind_total;
	if t>495 then Ind_total_new=.;/* 124 observations for test set */
	month_num = month(date);
   jan = (month_num = 1);  feb = (month_num = 2); mar = (month_num = 3);apr = (month_num
= 4);
   may = (month_num = 5); jun = (month_num = 6); jul = (month_num = 7); aug = (month_num
= 8);
```

```
   sep = (month_num = 9); oct = (month_num = 10);  nov = (month_num = 11);
   t3 = t*t*t;
run;

proc reg data=energy_monthly;
   model Ind_total_new =Ind_primary t3 jan feb mar apr may jun jul aug sep oct nov / clb vif
dwprob aic bic;
   output out=energy_monthly1 p=Ind_total_predict r=Ind_total_resid;
run;
proc sgplot data=energy_monthly1;
        series x=date y=Ind_total;
        series x=date y=Ind_total_predict;
        title "Actual vs Predicted Values with Monthly Dummy Variables";
run;

data energy_monthly1;
        set energy_monthly1;
        if t<=495 then
                do;
                        mape_fit=abs(Ind_total_resid/Ind_Total_new)*100;
                        mae_fit=abs(Ind_total_resid);
                        mse_fit=Ind_total_resid**2;
                end;
        else if t>495 then
                do;
                        mape_acc=abs(Ind_Total-Ind_total_predict/Ind_Total)*100;
                        mae_acc=abs(Ind_Total-Ind_Total_predict);
                        mse_acc=(Ind_Total-Ind_total_predict)**2;
                end;
run;

proc means data=energy_monthly1 mean maxdec=3;
        var mape_fit mae_fit mse_fit mape_acc mae_acc mse_acc;
run;


/* ARIMA */
/* none of the models gave white noise residuals */
data energy_monthly;
   set energy_monthly;
   t =_n_;
   arima_New =Ind_Total;
   if t>495 then arima_New=.;
run;
```

```
proc arima data=energy_monthly;
	identify var=arima_New(12,1) nlag=36 whitenoise=ignoremiss;
	*estimate p=(1)(2)(11)(14)(16)(23)(28)(31)(35)(12)(24)(36)
q=(1)(2)(3)(11)(13)(26)(27)(29)(32)(33)(12) whitenoise=ignoremiss; /*ARIMA(9,1,10)(3,1,1) */
	estimate p=(1)(2)(14)(23)(28)(34)(35)(12)(24) q=(1)(3)(13)(26)(27)(29)(12)
whitenoise=ignoremiss; /*ARIMA(7,1,6)(2,1,1) */
	*estimate p=(1)(2)(11)(14)(23)(28)(31)(35)(12)(24)(36) q=(1)(2)(3)(13)(26)(29)(32)(12)
whitenoise=ignoremiss; /*ARIMA(8,1,7)(3,1,1) */
	*forecast id=month interval=month lead=124 out=energy_monthlyout1;
run;

/* Deseasonalize */

proc timeseries data=energy_monthly outdecomp=sa_Ind out=null;
	decomp sa;
	id date interval=month;
	var Ind_total;
run;

data energy_monthly1;
	merge energy_Monthly sa_Ind;
	t=_n_;
	sa1=sa;
	if t>495 then sa1=.;
	si=Ind_total/sa;
run;

proc sgplot data=energy_monthly1;
	series x=t y=Ind_total;
	series x=t y=sa;
	title "Ind_total Vs SA";
run;

proc reg data=energy_monthly1;
	model sa1=Ind_total;
	output out=Ind_out r=sa_resid p=sa_predict;
run;

data Ind_out;
	set Ind_out;
	Ind_reseason=si*sa_predict;
	if t<=495 then
		do;
```

```
				mape_fit=abs(sa_resid/sa1)*100;
				mae_fit=abs(sa_resid);
				mse_fit=sa_resid**2;
			end;
		else if t>495 then
			do;
				mape_acc=abs((sa-sa_predict)/sa)*100;
				mae_acc=abs(sa-sa_predict);
				mse_acc=(sa-sa_predict)**2;
			end;
run;

proc means data=Ind_out n mean;
	var mape_fit mape_acc mae_fit mae_acc mse_fit mse_acc;
run;

proc sgplot data=Ind_out;
	series x=t y=Ind_total;
	series x=t y=Ind_reseason;
run;
```