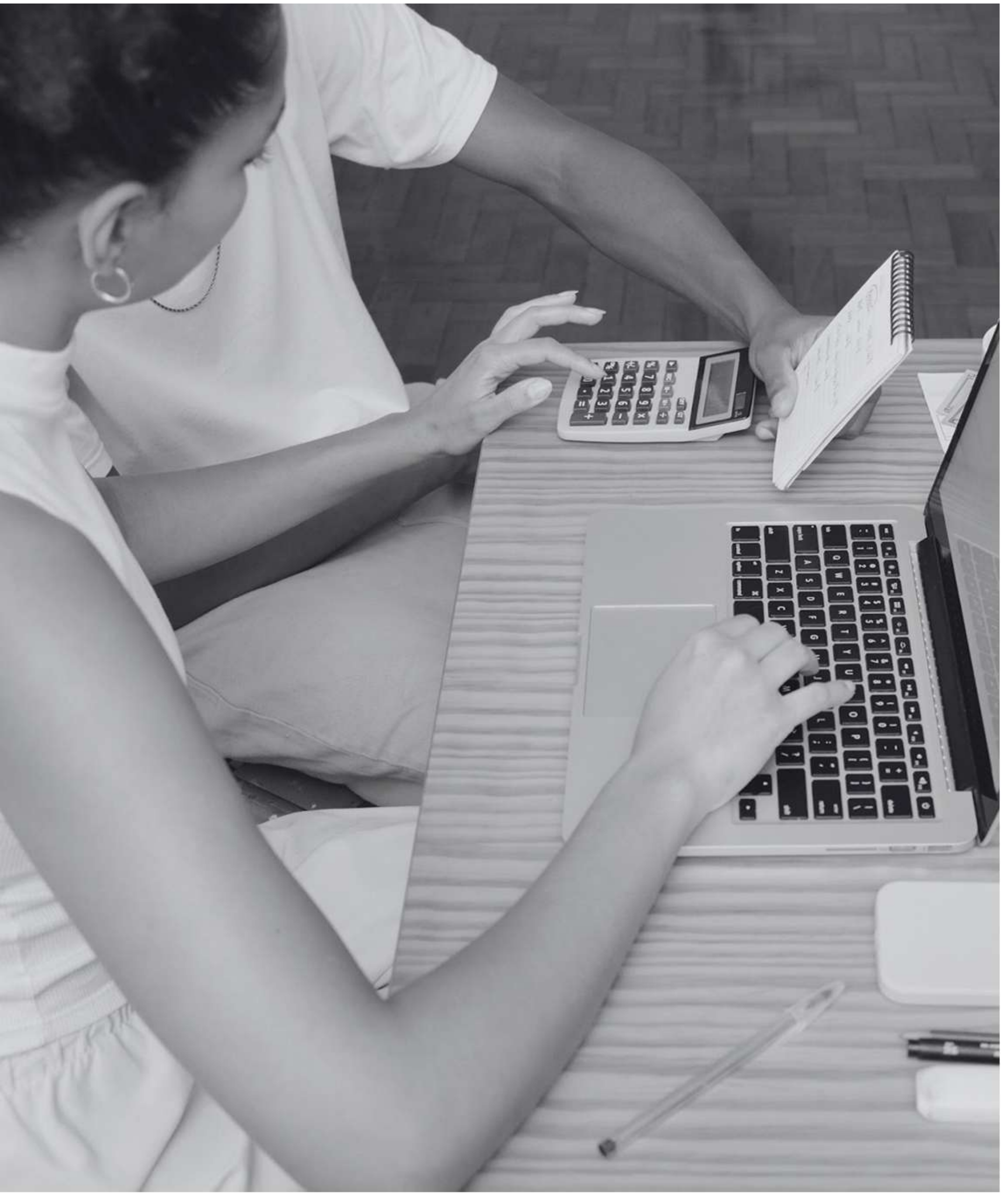
Group 8

Us Energy Consumption



What this report will cover

- Introduction
- Dataset
- Dataset Preparation & Cleaning
- EDA
- Model Used
- Comparison of Models
- Conclusion

Vision

Analyze the trends and patterns in energy consumption across the **Commercial** and **Industrial** sectors in the United States to forecast future energy consumption in these sectors to assist in planning.

Introduction

- 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.

Dataset

Dataset from U.S. Energy Information Administration:	EIA
Monthly data	1973 to 2024
Dependent Variables	Total Energy consumption
Independent Variables	Primary energy input, electricity sales, and system losses
Training set	January 1973 to February 2014
Test Set	March 2014 to July 2024

Key Variables in the Dataset

Com_Primary :	Total primary energy consumption in the commercial sector.
Com_Elec_Sales	Total electricity sales to the commercial sector.
Com_End_Use	Total end-use energy consumption in the commercial sector.
Com_Elec_Losses	Total electricity losses in the commercial sector, including transmission losses.
Ind_Primary	Total primary energy consumption in the Industrial sector.
Ind_Elec_Sales	Total electricity sales to the Industrial sector.
Ind_End_Use	Total end-use energy consumption in the Industrial sector.
Ind_Elec_Losses	Total electricity losses in theIndustrial sector, including transmission losses.

Dataset Preparation

Imported Dataset

Monthly (1973–2024) and annual records from the EIA.

.

Standardized & Format

Standardized the "Date" column into SAS-compatible formats (MONYY7.)

Data Cleaning

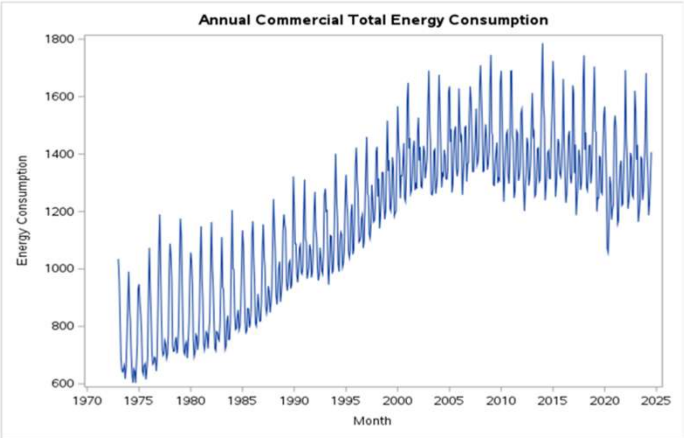
Cleaned missing values and resolved inconsistencies in energy metrics.

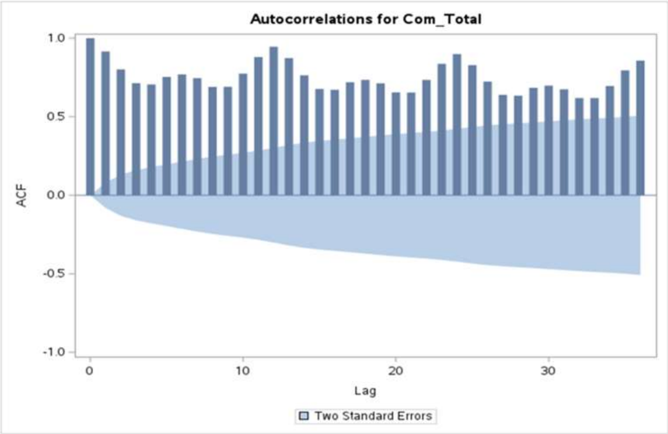
Data Partition

80-20 split.

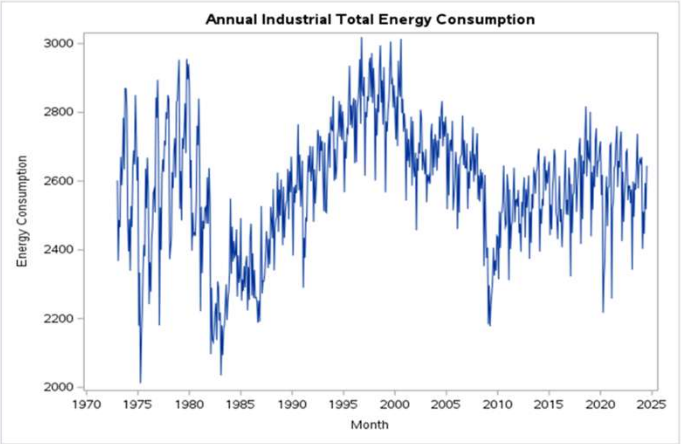
Time Series & ACF

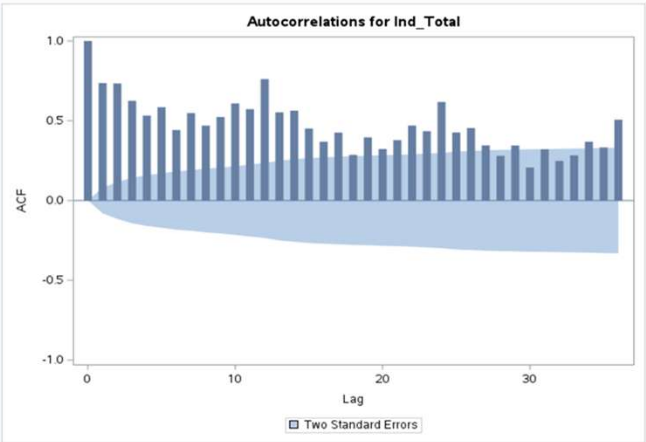
Commercial Sector: Trend & Seasonality





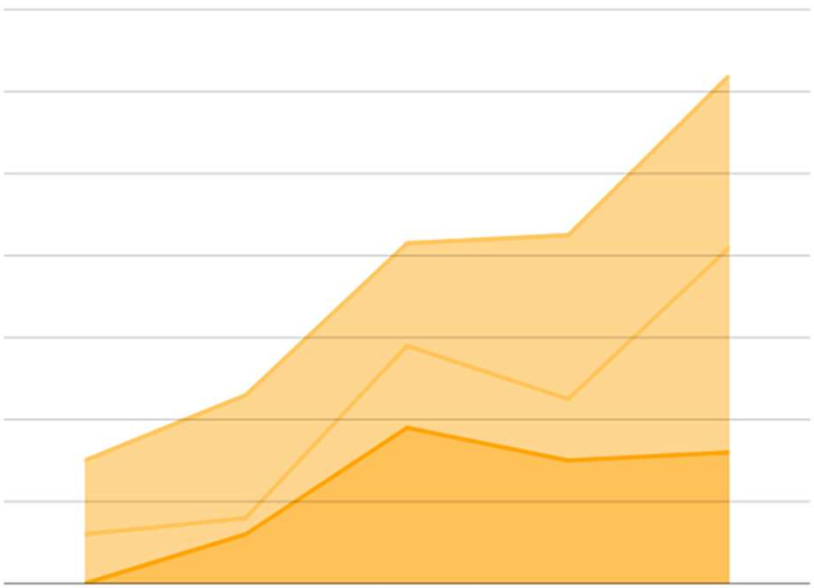
Industrial Sector: Non-linear Trend & Seasonality





Models & Methods used

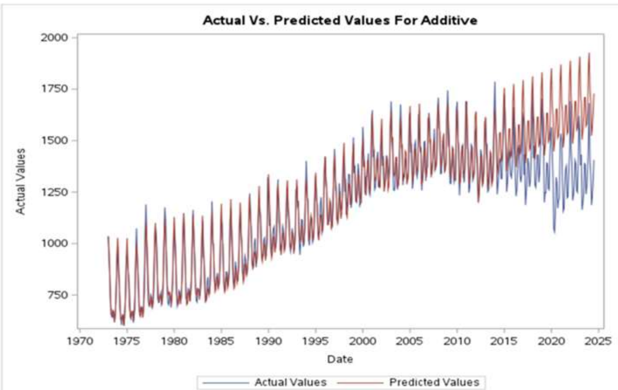
- Holt Winter's exponential Smoothing
- Multiple Linear Regression & Non linear Regression
  - using dummy variables
  - using deseasonalizing and Reseasonalising
- ARIMA
- Time series decomposition
- Classical Decomposition

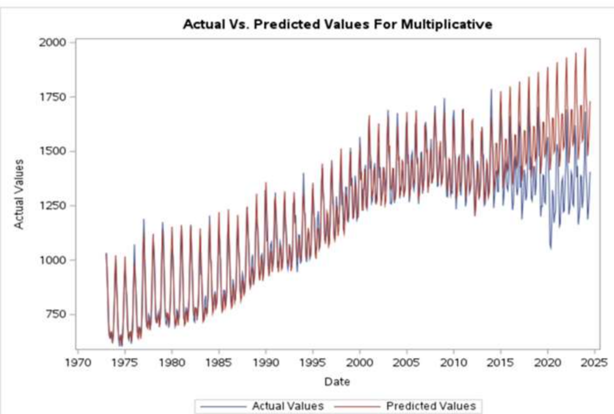


Commercial Sector

Holt’s Winter Exponential Model

Commercial Sector: Additive and Multiplicative:





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

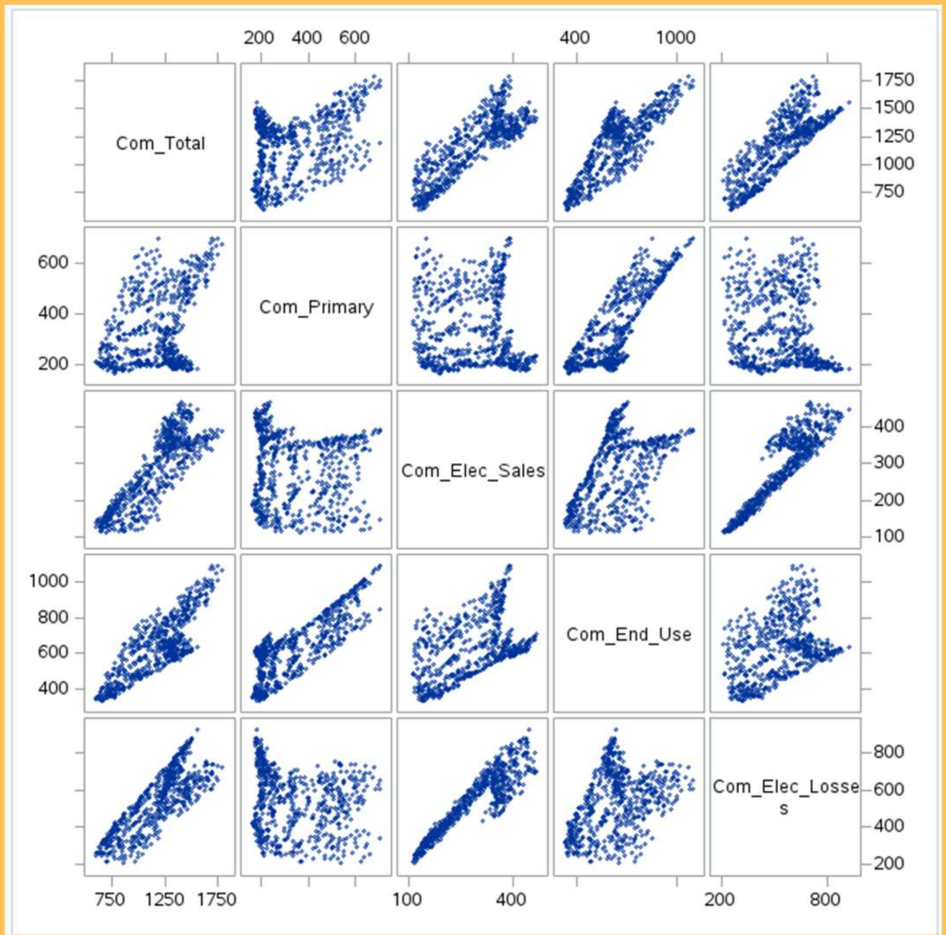
Winters Method (Multiplicative) Parameter Estimates

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

	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

- Magnitude of seasonal component changes overtime, suggests multiplicative.
- Error values for Multiplicative accuracy is less, suggesting a better model.
- The Multiplicative model consistently outperforms the Additive model across all metrics in terms of accuracy. It is the better model to use based on these results.

Correlation Matrix



Commercial Sector

Except the Com_Primary, com_Elec_Losses, other all variables looks relatively linear.

Multiple Regression: Model Evaluation

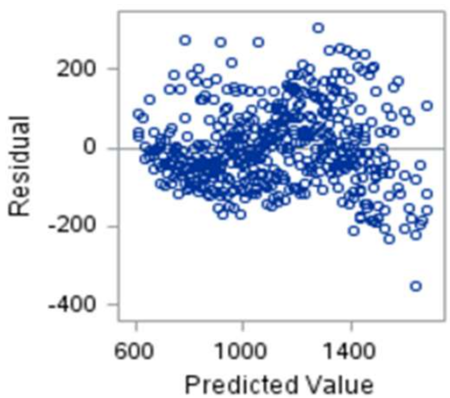
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	37202508	3382046	1345.94	<.0001
Error	483	1213670	2512.77437		
Corrected Total	494	38416178			

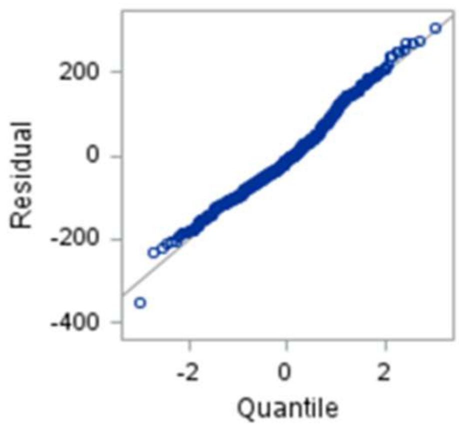
Root MSE	50.12758	R-Square	0.9684
Dependent Mean	1124.94488	Adj R-Sq	0.9677
Coeff Var	4.45600		

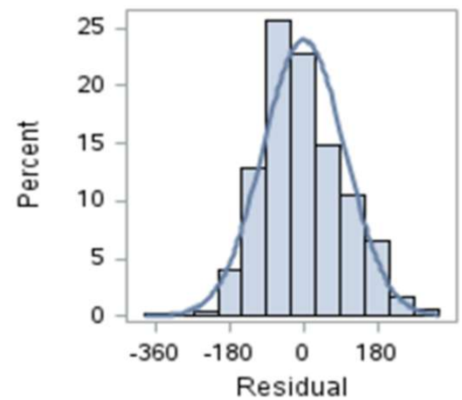
Parameter Estimates									
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	371.86748	14.88993	24.97	<.0001	0	342.61044	401.12453
t		1	1.20010	0.02459	48.80	<.0001	2.43230	1.15178	1.24842
jan		1	-35.23547	12.38354	-2.85	0.0046	2.34573	-59.56773	-10.90321
feb		1	-111.52227	10.91991	-10.21	<.0001	1.82401	-132.97867	-90.06586
mar		1	-94.01125	9.75227	-9.64	<.0001	1.45479	-113.17337	-74.84913
apr		1	-136.07753	9.04314	-15.05	<.0001	1.22382	-153.84628	-118.30879
may		1	-85.32189	9.42547	-9.05	<.0001	1.32949	-103.84187	-66.80191
jun		1	-35.80478	9.61910	-3.72	0.0002	1.38468	-54.70524	-16.90432
sep		1	-74.50389	9.55082	-7.80	<.0001	1.36509	-93.27017	-55.73761
oct		1	-88.88239	9.26745	-9.59	<.0001	1.28529	-107.09189	-70.67289
nov		1	-99.14941	9.04918	-10.96	<.0001	1.22546	-116.93003	-81.36879
Com_End_Use	Com_End_Use	1	0.86391	0.03162	27.32	<.0001	4.98510	0.80179	0.92604

- The model is logical because the sign of slope is intuitive.
- Slope terms statistically significant with P-value less than alpha.
- The model is statistically significant.
- Using adjusted R2 of **96.77%** indicates a good model fit.
- There is no indication of multicollinearity.

Multiple Regression:Model Assumptions







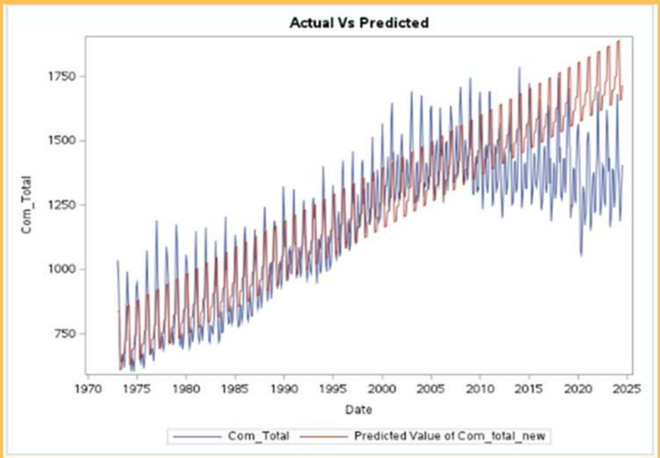
Durbin-Watson D	0.915
Pr < DW	<.0001
Pr > DW	1.0000
Number of Observations	495
1st Order Autocorrelation	0.536

- For the 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.**

Linear Vs Non-Linear

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.

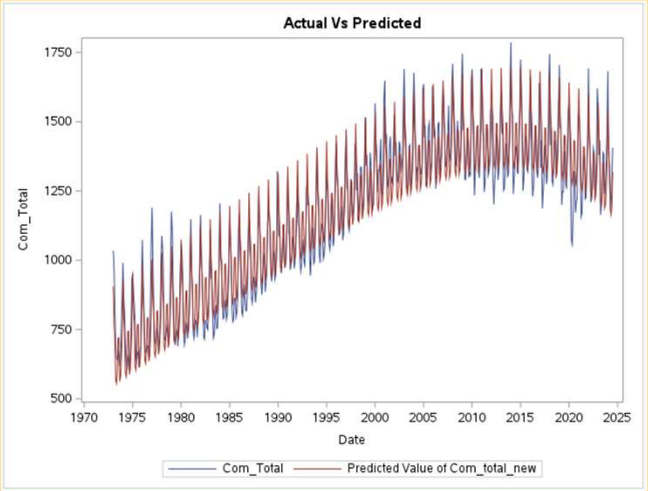
Linear



The MEANS Procedure

Variable	Mean
mape_fit	3.197
mae_fit	36.986
mse_fit	2451.859
mape_acc	136826.008
mae_acc	263.758
mse_acc	58486.373

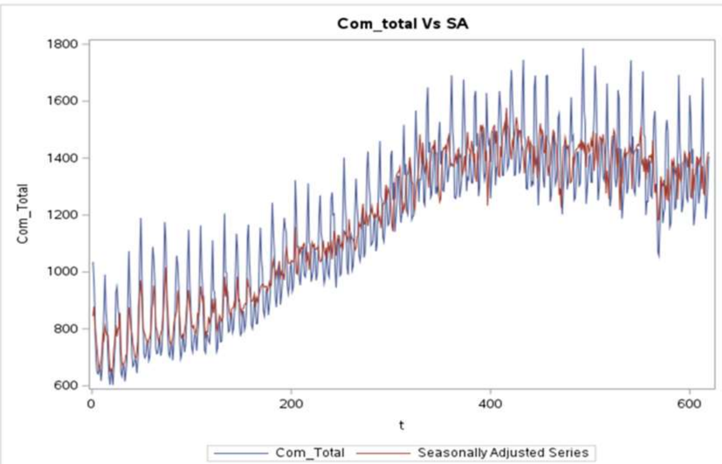
Non- Linear

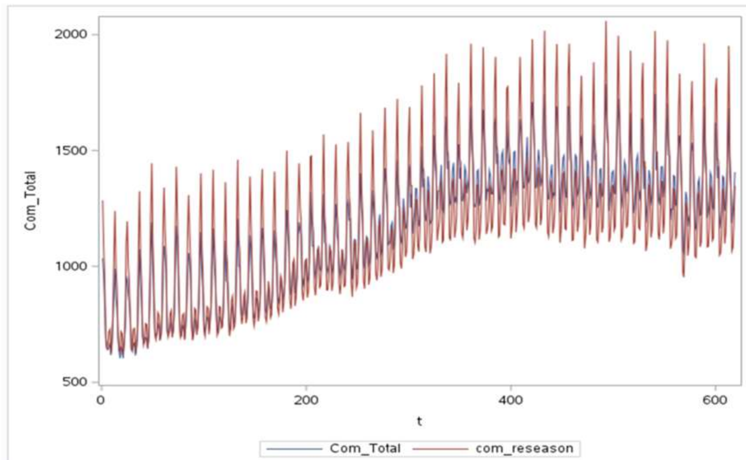


The MEANS Procedure

Variable	Mean
mape_fit	2.994
mae_fit	32.815
mse_fit	1673.209
mape_acc	136844.258
mae_acc	42.862
mse_acc	2927.747

Deseasonalize Method





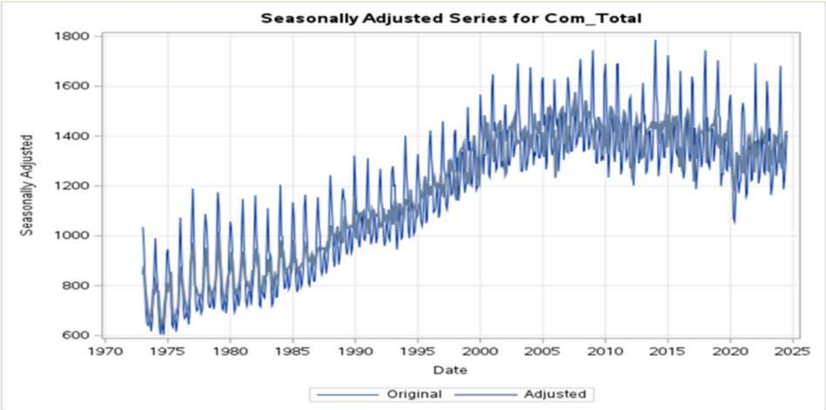
The MEANS Procedure

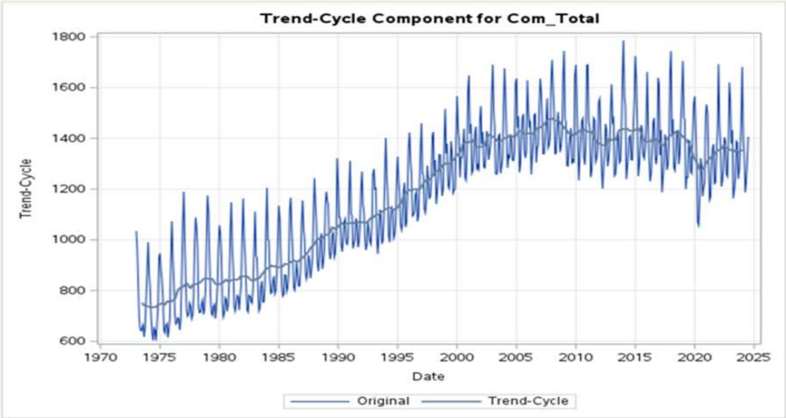
Variable	N	Mean
mape_fit	495	7.8105718
mape_acc	124	8.4383558
mae_fit	495	89.7277023
mae_acc	124	115.9162690
mse_fit	495	12042.96
mse_acc	124	16417.56

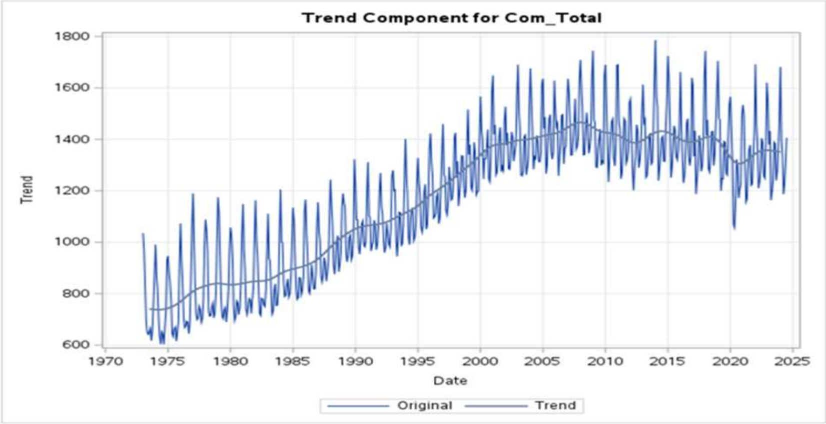
- The original time series, exhibiting a clear upward trend and pronounced seasonal patterns.
- The deseasonalized series matches the original series closely, indicating that the seasonal component was correctly identified and removed.
- However, the error metrics did a somewhat good job in giving accuracy.

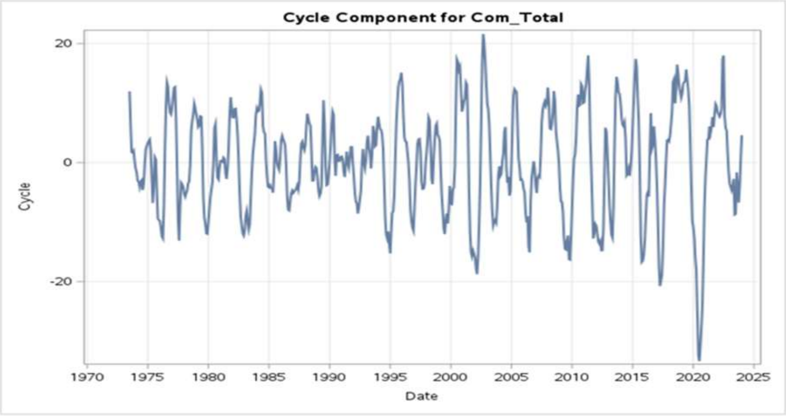
Classical Decomposition Model

Commercial Sector:



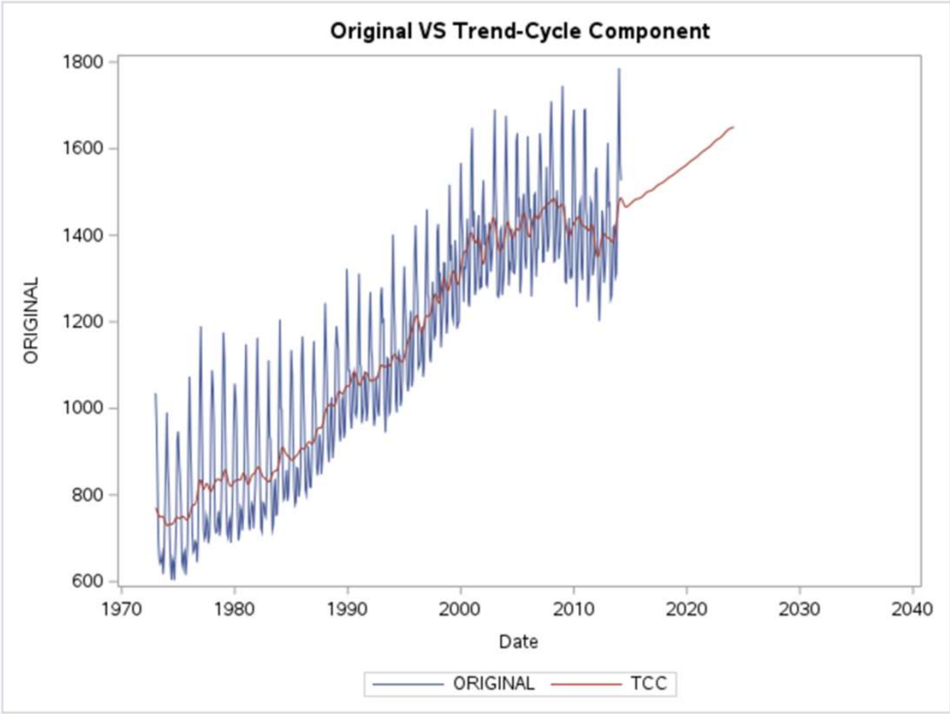


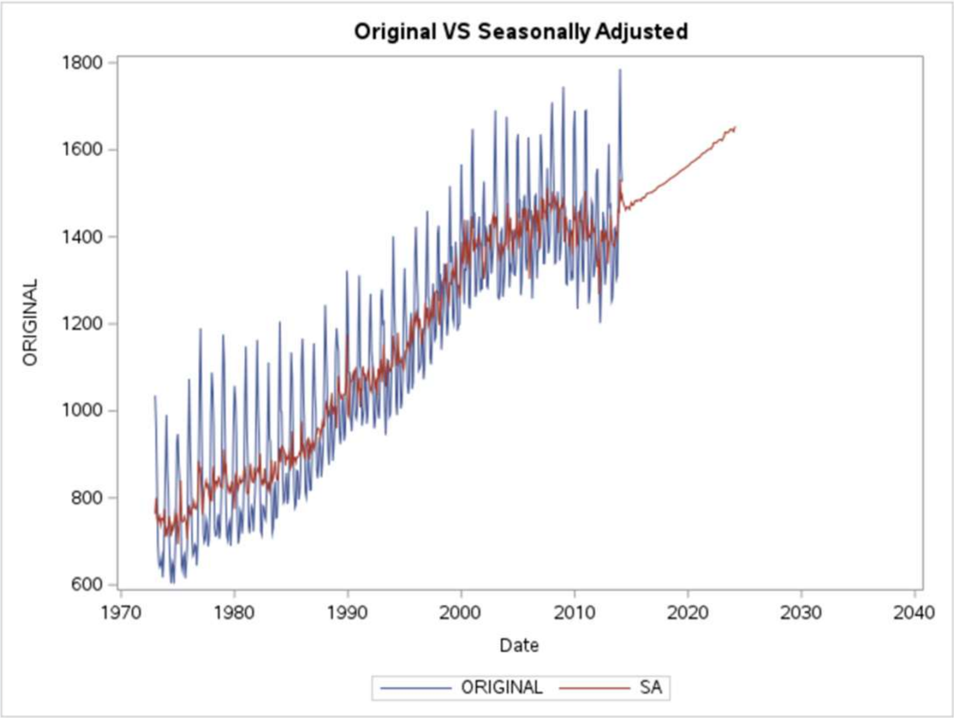




X11 Decomposition Model

Commercial Sector:





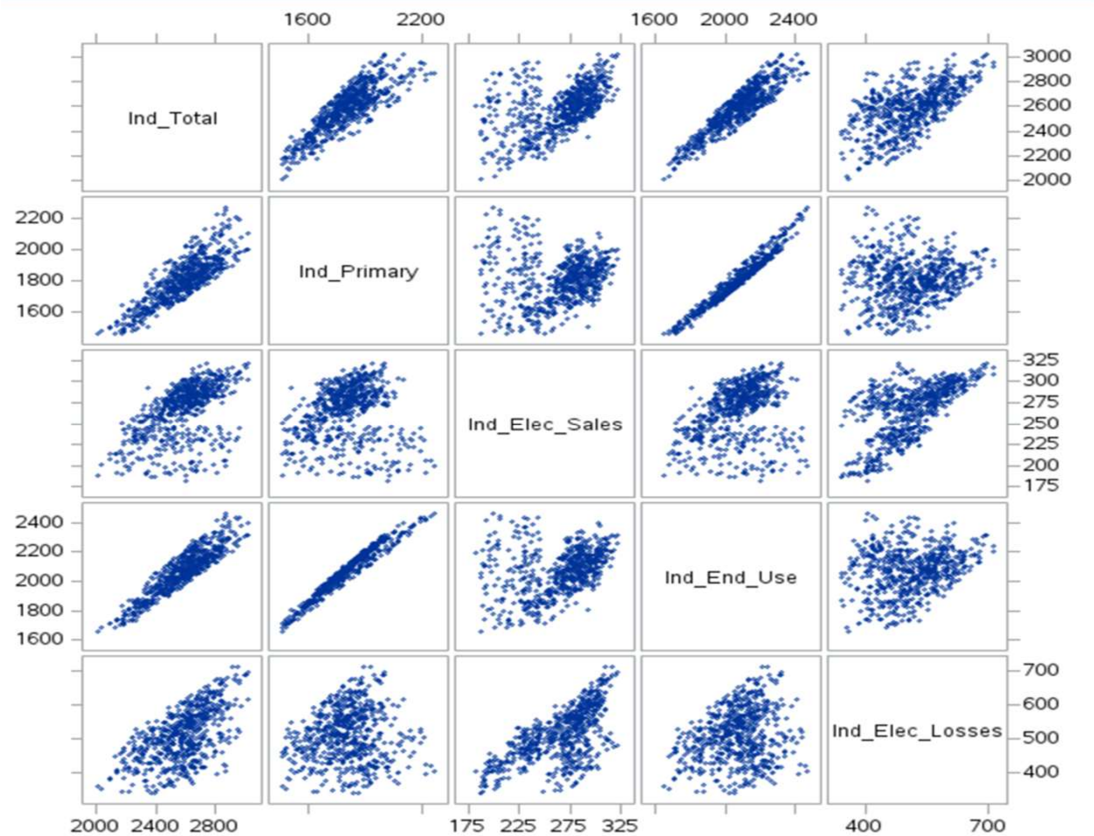
Industrial Sector

Correlation Matrix

Variables Ind_Primary, Ind_Elec_Sales,and Ind_End_Use are likely key drivers of total energy consumption (Ind_Total).

The correlation matrix indicates linear relationships between variables, although the time series trend suggests a nonlinear pattern.

Using a combination of linear and nonlinear models to understand what suits the best for industrial sector.



Multiple linear Regression: Model Evaluation

y=−91.86−2.58·Ind_Primary+3.59·Ind_End_Use+0.18·t−0.00054·t2

- Most of the independent variables have negative slope, indicating the model is not logical.
- Slope terms statistically significant with P-value less than alpha.
- The model is statistically significant.
- Using adjusted R2 of 99.24% indicates a good model fit.
- There is a clear indication of multicollinearity with VIF >10 of Ind_primary & Ind_end_user.

Analysis of Variance

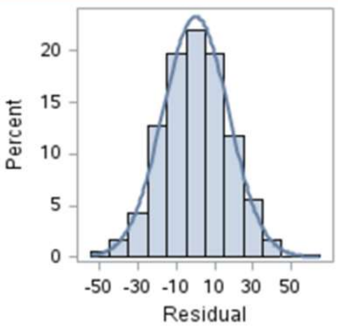
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	19309753	1379268	4586.99	<.0001
Error	480	144332	300.69126		
Corrected Total	494	19454085			

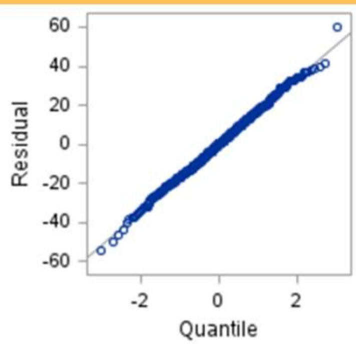
Root MSE	17.34045	R-Square	0.9926
Dependent Mean	2582.49336	Adj R-Sq	0.9924
Coeff Var	0.67146		

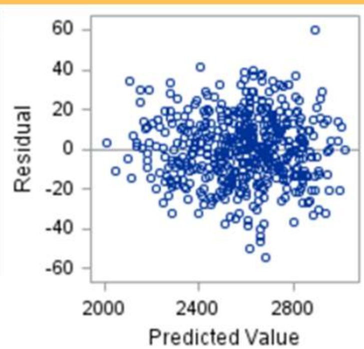
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	-91.86238	12.29102	-7.47	<.0001	0	-116.01324	-67.71153
Ind_Primary	Ind_Primary	1	-2.58364	0.05152	-50.15	<.0001	102.51411	-2.68487	-2.48242
Ind_End_Use	Ind_End_Use	1	3.59180	0.04840	74.21	<.0001	97.36994	3.49669	3.68690
t		1	-0.20046	0.01112	-18.03	<.0001	4.15745	-0.22232	-0.17861
jan		1	-18.05837	3.80970	-4.74	<.0001	1.85526	-25.54413	-10.57262
feb		1	-53.27506	3.93580	-13.54	<.0001	1.98011	-61.00859	-45.54153
mar		1	-31.37068	3.84329	-8.16	<.0001	1.88812	-38.92244	-23.81893
apr		1	-43.52210	3.97924	-10.94	<.0001	1.98022	-51.34098	-35.70323
may		1	-7.32719	3.98760	-1.84	0.0668	1.98855	-15.16249	0.50811
jun		1	-4.11746	4.08031	-1.01	0.3134	2.08209	-12.13494	3.90001
jul		1	8.49907	4.01946	2.11	0.0350	2.02046	0.60116	16.39699
aug		1	-7.83018	4.03681	-1.94	0.0530	2.03793	-15.76218	0.10181
sep		1	-61.22298	3.99798	-15.31	<.0001	1.99892	-69.07868	-53.36727
oct		1	-36.21965	3.88096	-9.33	<.0001	1.88362	-43.84542	-28.59387
nov		1	-16.00146	3.85684	-4.15	<.0001	1.86027	-23.57984	-8.42309

Multiple linear Regression: Model assumption







Durbin-Watson D	1.056
Pr < DW	<.0001
Pr > DW	1.0000
Number of Observations	495
1st Order Autocorrelation	0.471

- Normality : True; Histogram the plot exhibits bell shaped and the QQ plot data points are in line, thereby residuals are normally distributed.
- Equal Variance – True; the Residuals v/s predicted shows no pattern which indicates homoscedasticity.
- Independent – False; the p-value of d-w test is less than alpha showing serial/positive correlation therefore the independent assumption is false.

Multiple Non-linear Regression: Model Evaluation

- Most of the independent variables have negative slope, indicating the model is not logical.
- Slope terms statistically significant with P-value less than alpha.
- The model is statistically significant.
- Using adjusted R2 of 99.29% indicates a good model fit.
- There is a clear indication of multicollinearity with VIF >10 of Ind_primary & Ind_end_user.

Analysis of Variance

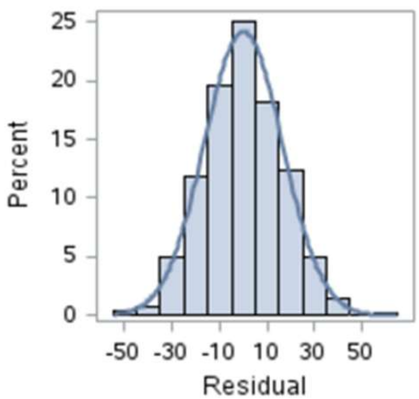
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	19319564	1287971	4586.18	<.0001
Error	479	134521	280.83724		
Corrected Total	494	19454085			

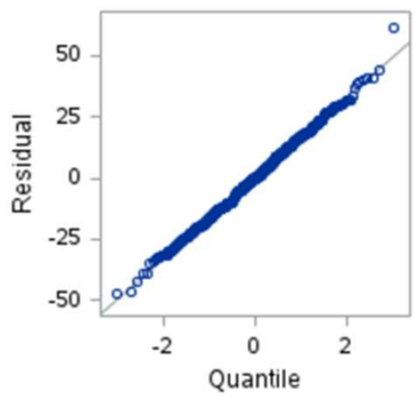
Root MSE	16.75820	R-Square	0.9931
Dependent Mean	2582.49336	Adj R-Sq	0.9929
Coeff Var	0.64892		

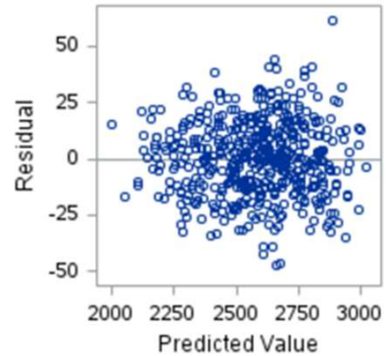
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	-74.63312	12.23077	-6.10	<.0001	0	-98.66571	-50.60052
Ind_Primary	Ind_Primary	1	-1.99913	0.11072	-18.06	<.0001	507.01050	-2.21669	-1.78158
Ind_End_Use	Ind_End_Use	1	3.04548	0.10359	29.40	<.0001	477.55093	2.84193	3.24904
t		1	0.17847	0.06501	2.75	0.0063	152.08513	0.05073	0.30620
t2		1	-0.00054229	0.00009175	-5.91	<.0001	79.48181	-0.00072257	-0.00036200
jan		1	-18.90246	3.68455	-5.13	<.0001	1.85805	-26.14235	-11.66258
feb		1	-49.49461	3.85705	-12.83	<.0001	2.03610	-57.07345	-41.91578
mar		1	-26.88530	3.79098	-7.09	<.0001	1.96694	-34.33430	-19.43629
apr		1	-36.41733	4.02911	-9.04	<.0001	2.17370	-44.33425	-28.50041
may		1	3.04947	4.23477	0.72	0.4718	2.40126	-5.27155	11.37049
jun		1	9.37517	4.55641	2.06	0.0402	2.77988	0.42214	18.32820
jul		1	21.05829	4.42770	4.76	<.0001	2.62504	12.35817	29.75840
aug		1	6.43530	4.58751	1.40	0.1613	2.81795	-2.57882	15.44942
sep		1	-49.43455	4.34816	-11.37	<.0001	2.53157	-57.97837	-40.89074
oct		1	-29.14064	3.93724	-7.40	<.0001	2.07569	-36.87703	-21.40424
nov		1	-11.84730	3.79302	-3.12	0.0019	1.92642	-19.30032	-4.39428

Multiple Non-linear Regression: Model assumption



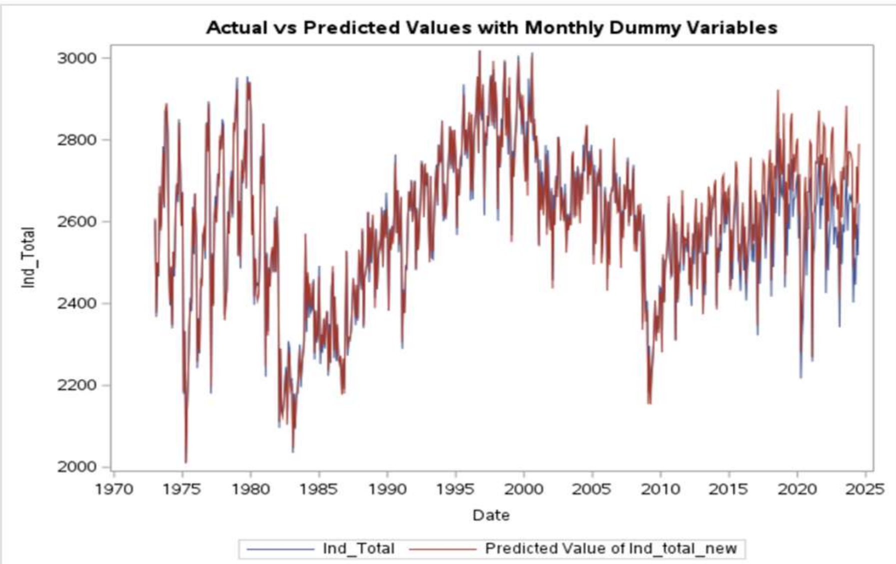


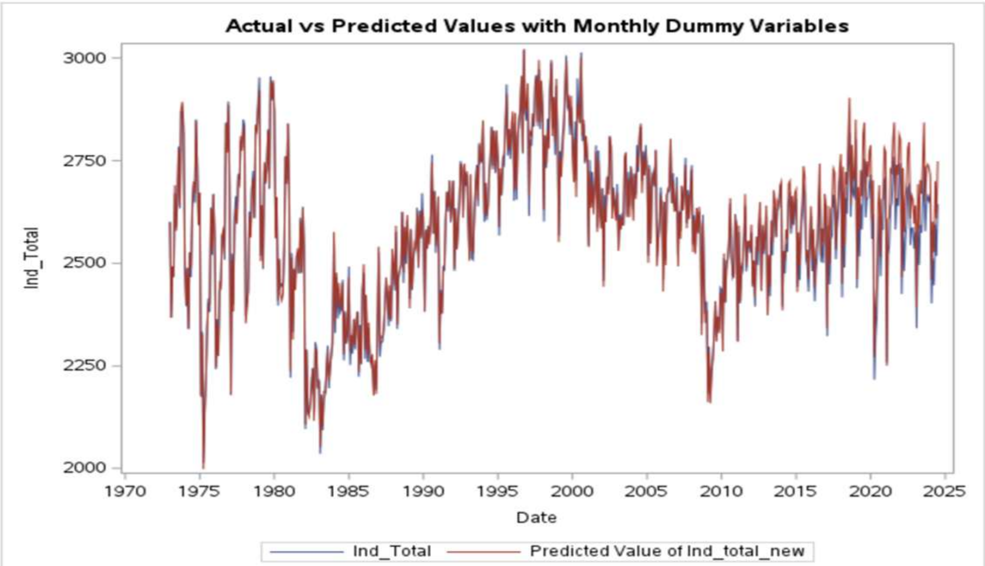


Durbin-Watson D	1.104
Pr < DW	<.0001
Pr > DW	1.0000
Number of Observations	495
1st Order Autocorrelation	0.448

- Normality : True; histogram plot exhibits a symmetrical bell-shaped pattern. The of the data points lie on the line of PQ plot. Both suggests the residuals are normally distributed.
- Equal Variance – True; the Residuals v/s predicted shows no pattern which indicates that the assumption of equal variance is true.
- Independent – False; the p-value of d-w test is less than alpha showing serial correlation therefore the independent assumption is false.

Multiple Regression: Linear vs Nonlinear using dummy variables





The MEANS Procedure

Variable	Mean
mape_fit	0.531
mae_fit	13.670
mse_fit	291.579
mape_acc	258290.648
mae_acc	77.218
mse_acc	7112.355

The model has better performance in Nonlinear multiple regression model.

However, since there is a presence of multicollinearity the model is showing overfitting.

The MEANS Procedure

Variable	Mean
mape_fit	0.513
mae_fit	13.225
mse_fit	271.760
mape_acc	258291.319
mae_acc	59.958
mse_acc	4265.748

Deseasonalized & Reseasonalized Method

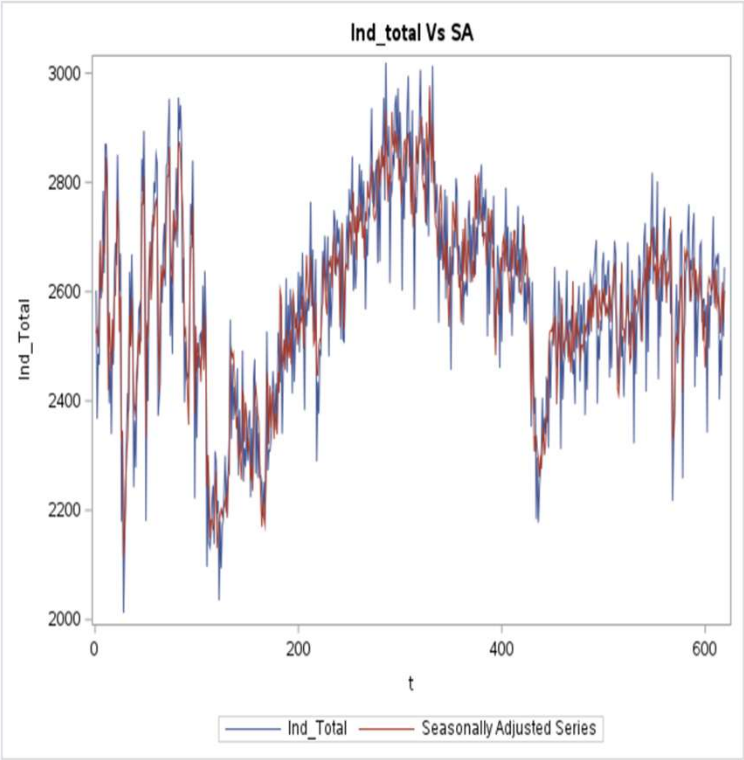
This method couldn't simplify and ignore noises in the data.

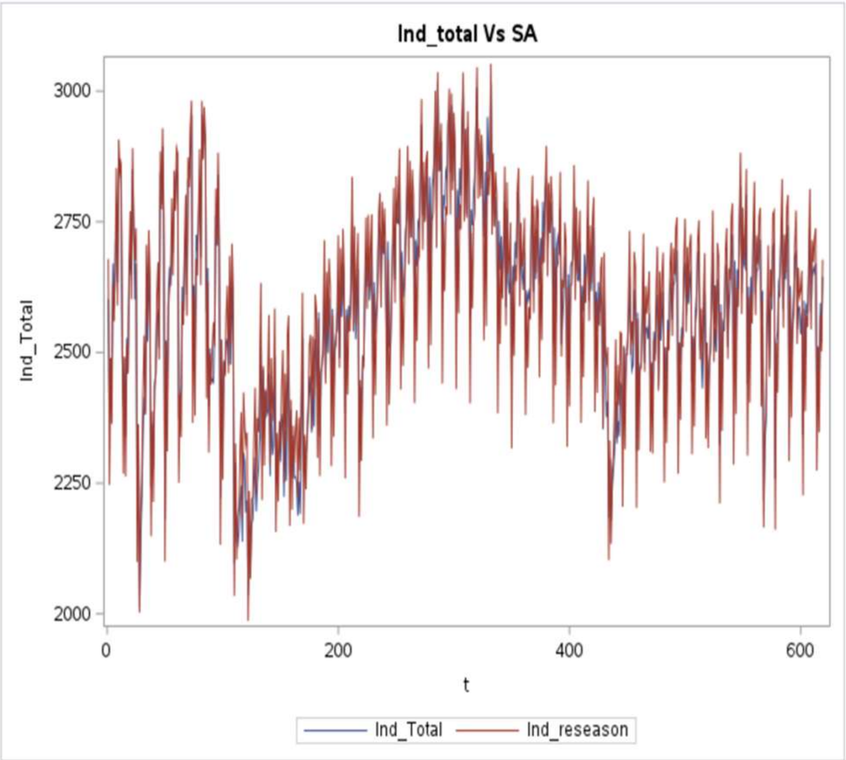
The reseasoned plot clearly captures even a small change/noises in data.

The error metrics does a good job in accuracy stats.

The MEANS Procedure

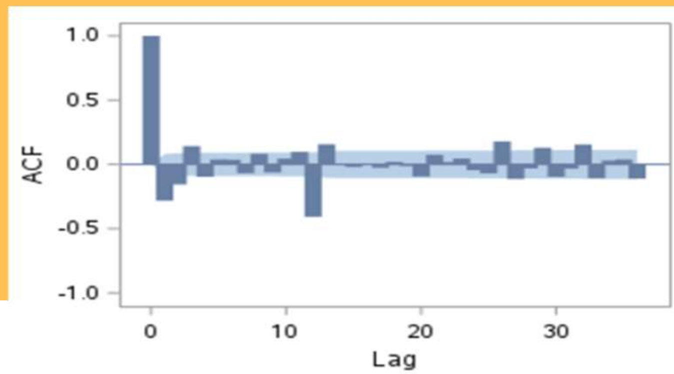
Variable	N	Mean
mape_fit	495	2.3242671
mape_acc	124	2.1239893
mae_fit	495	59.5757882
mae_acc	124	54.8492350
mse_fit	495	5390.05
mse_acc	124	4523.37

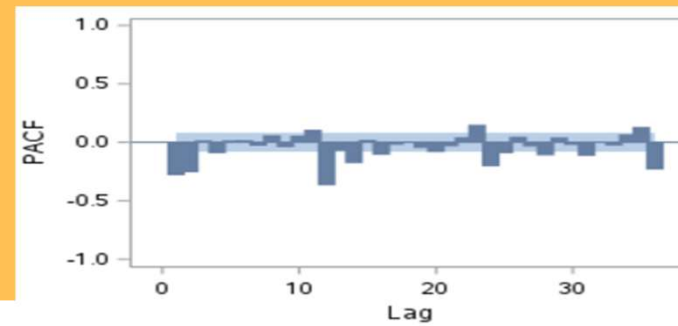




ARIMA

To attain stationarity in the acf plot: Differentiated the model twice - One to remove the seasonality and other to remove the trend.





None of the ARIMA models gave white noise residuals to proceed with forecasting, indicating it to not be a suitable choice for the data.

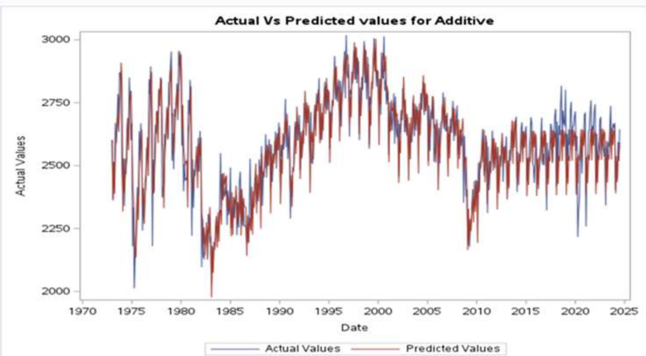
ARIMA(11,1,11)(3,1,1
)

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	.	0	.	-0.001	-0.000	0.004	-0.000	-0.007	0.042
12	.	0	.	-0.068	0.023	-0.043	0.033	0.022	0.000
18	.	0	.	0.000	-0.000	-0.036	0.000	-0.045	-0.028
24	.	0	.	-0.084	-0.061	0.020	0.052	-0.000	0.001
30	18.42	4	0.0010	0.000	0.002	-0.003	0.000	-0.001	-0.033
36	19.25	10	0.0372	-0.001	0.000	0.001	0.036	0.001	-0.003
42	26.96	16	0.0419	0.033	-0.048	0.038	0.069	-0.045	0.014
48	38.12	22	0.0178	0.025	-0.061	-0.106	0.004	0.036	-0.003

ARIMA(6,1,5)(3,1,1)

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	.	0	.	0.008	-0.025	0.008	-0.000	0.015	0.053
12	.	0	.	-0.045	-0.006	-0.047	0.040	0.033	0.000
18	11.72	2	0.0029	0.000	-0.000	-0.030	-0.078	-0.035	-0.005
24	24.89	8	0.0016	-0.080	-0.095	0.005	0.074	-0.001	0.000
30	30.60	14	0.0063	-0.010	0.001	-0.074	-0.000	0.041	-0.041
36	41.09	20	0.0036	-0.063	0.060	-0.080	0.048	0.003	-0.001
42	48.85	26	0.0043	0.033	-0.045	0.037	0.065	-0.050	0.026
48	61.32	32	0.0014	0.034	-0.061	-0.108	0.018	0.046	0.000

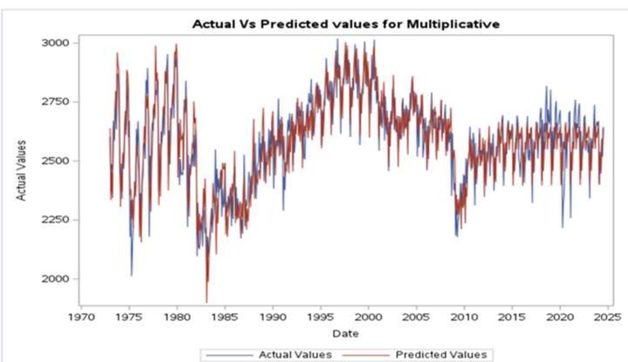
Holt Winter's Exponential Model



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

	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



- 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 is less, suggesting a better model.

Model Comparison: Commercial Sector

	MAPE Fit	MAE Fit	MSE Fit	MAPE Acc	MAE Acc	MSE Acc
Holt Winter's Exponential Model (Multiplicative)	2.81	31.70	1750.09	16.10	215.56	56000.01
Regression using Dummy variables Non-Linear	2.994	32.815	1673.209	136844.258	42.862	2927.747
Regression(linear) using De-Reseasonalization	7.810	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 other on the validation set making it the most suitable Regression model for Forecasting the total Energy Consumption in Commercial Sector.**

Model Comparison: Industrial Sector

	MAPE Fit	MAE Fit	MSE Fit	MAPE Acc	MAE Acc	MSE Acc
Holt Winter's Exponential Model (Multiplicative)	2.3	58.05	5643.57	2.17	55.6	5205.43
Regression using Dummy variables Non-Linear	0.513	13.225	271.760	258291.319	59.958	4265.748
Regression(linear) using De-Reseasonalization	2.3245	59.576	5390.05	2.123	54.849	4523.37

- **The Multiple Non-Linear Regression 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.**

Conclusion

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.

Thank You!