Title: electrical energy output Prediction using Linear Regression

Project overview

This project analyzes the Combined Cycle Power Plant from the UCI Machine Learning Repository to predict the net hourly electrical energy output (EP). The goal was to identify the most significant factors affecting the output using **Multiple and Simple Linear Regression** techniques.

Methodology

- **Linear Regression**

Tool used: IBM SPSS Statistics / Microsoft Excel.

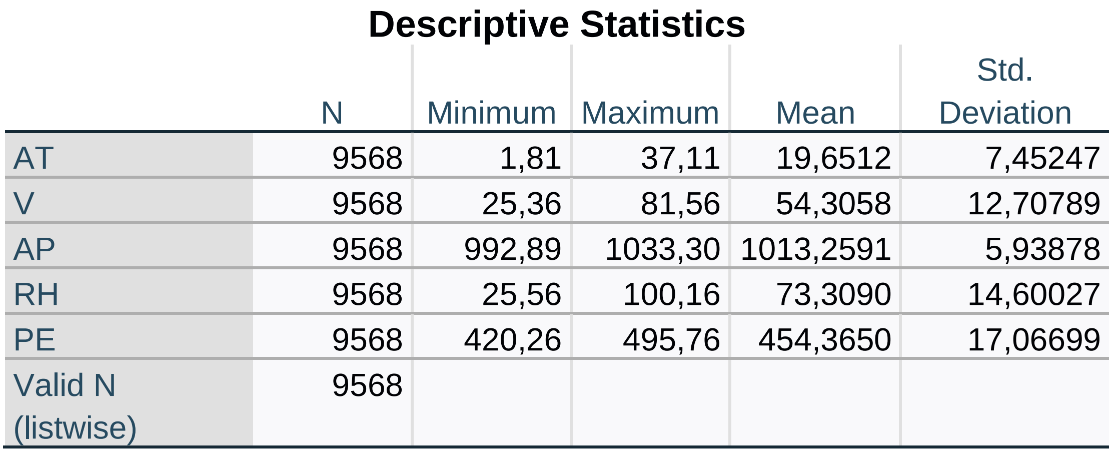
MAIN PART

First let’s think about the concept again. We have five variables:

- AT (Ambient Temperature - Θερμοκρασία)
- V (Exhaust Vacuum - Πίεση κενού)
- AP (Ambient Pressure)
- RH (Relative Humidity)
- PE (Power Energy)

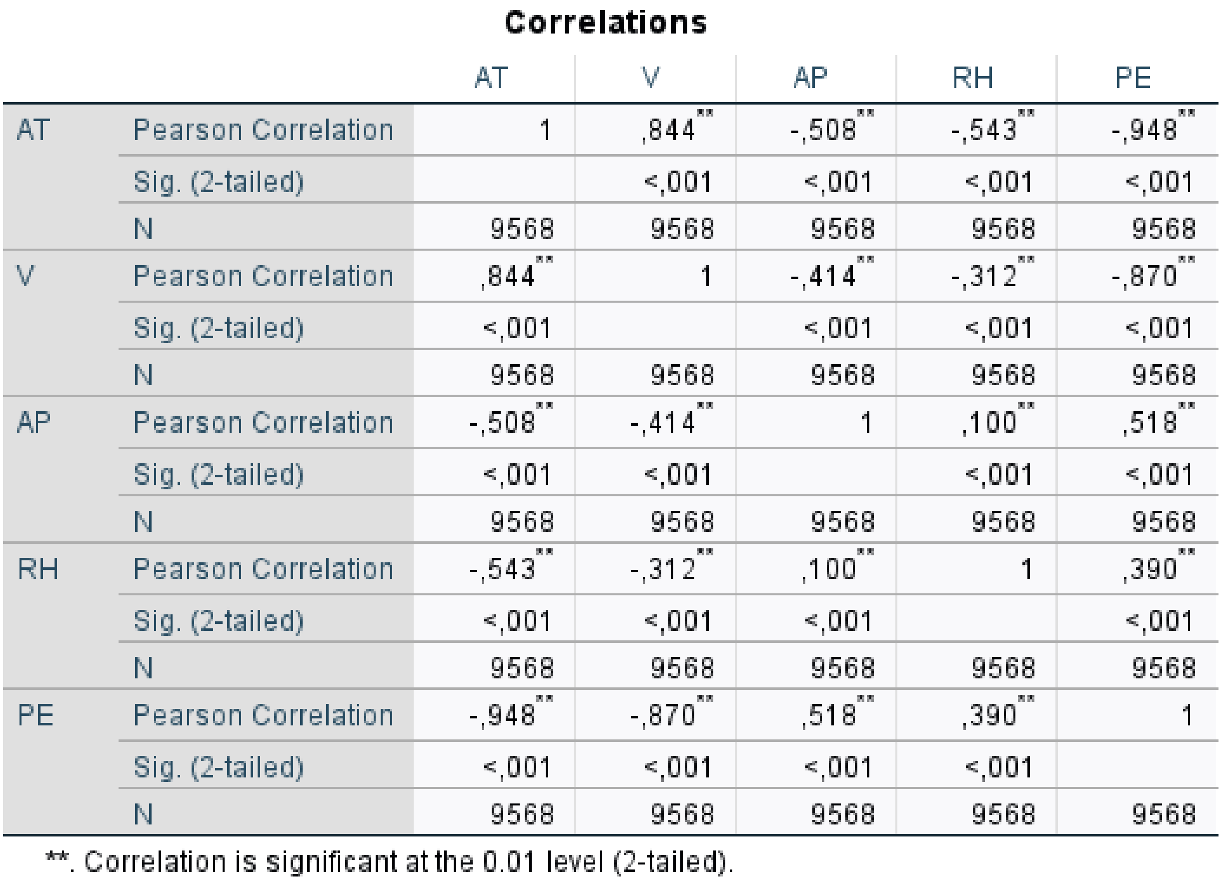
As the concept creator says “A combined cycle power plant (CCPP) is composed of gas turbines (GT), steam turbines (ST) and heat recovery steam generators. In a CCPP, the electricity is generated by gas and steam turbines, which are combined in one cycle, and is transferred from one turbine to another. While the Vacuum is collected from and has effect on the Steam Turbine, the other three of the ambient variables effect the GT performance.”

Lets begin with some descriptive statistics



Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
AT	9568	1,81	37,11	19,6512	7,45247
V	9568	25,36	81,56	54,3058	12,70789
AP	9568	992,89	1033,30	1013,2591	5,93878
RH	9568	25,56	100,16	73,3090	14,60027
PE	9568	420,26	495,76	454,3650	17,06699
Valid N (listwise)	9568				



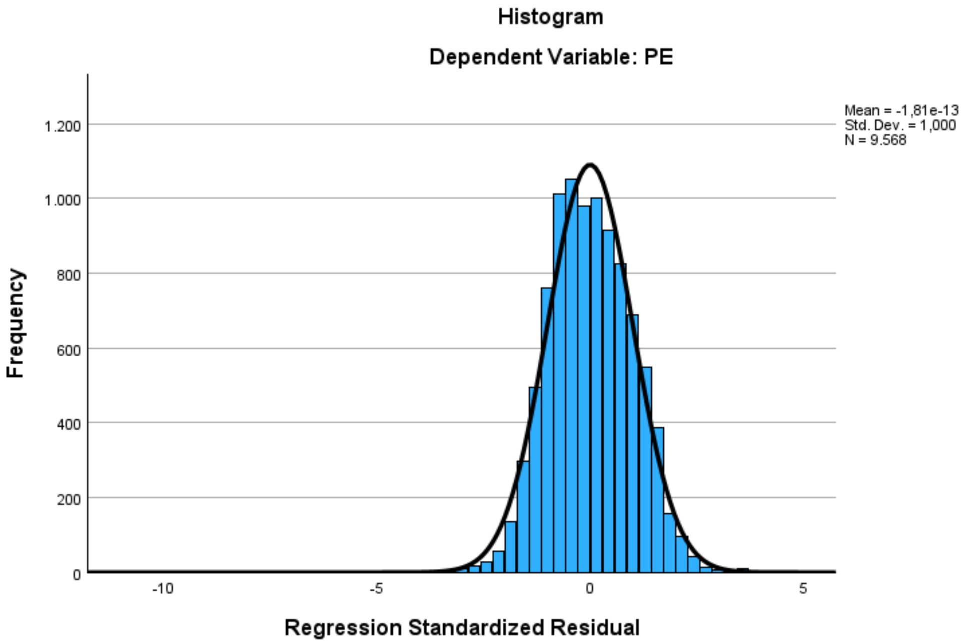
Correlations

		AT	V	AP	RH	PE
AT	Pearson Correlation	1	,844**	-,508**	-,543**	-,948**
	Sig. (2-tailed)		<,001	<,001	<,001	<,001
	N	9568	9568	9568	9568	9568
V	Pearson Correlation	,844**	1	-,414**	-,312**	-,870**
	Sig. (2-tailed)	<,001		<,001	<,001	<,001
	N	9568	9568	9568	9568	9568
AP	Pearson Correlation	-,508**	-,414**	1	,100**	,518**
	Sig. (2-tailed)	<,001	<,001		<,001	<,001
	N	9568	9568	9568	9568	9568
RH	Pearson Correlation	-,543**	-,312**	,100**	1	,390**
	Sig. (2-tailed)	<,001	<,001	<,001		<,001
	N	9568	9568	9568	9568	9568
PE	Pearson Correlation	-,948**	-,870**	,518**	,390**	1
	Sig. (2-tailed)	<,001	<,001	<,001	<,001	
	N	9568	9568	9568	9568	9568

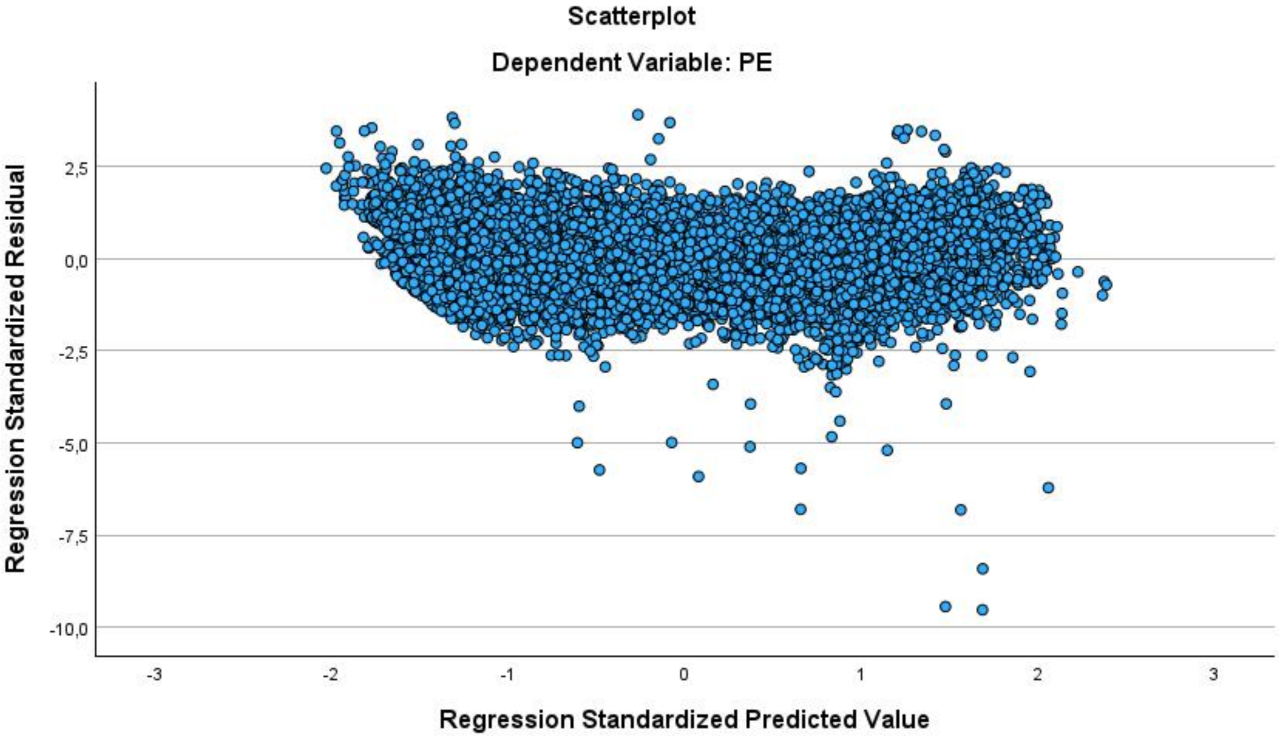
**. Correlation is significant at the 0.01 level (2-tailed).

We can easily see that all the variables are significant in relation to PE which is the variable of interest. It is worth to mention the very high correlation of ambient temperature (AT) and Exhaust Vacuum (V) with PE.

First, we need to check the assumption of our model which is that the residuals are normally distributed and homoscedastic, as we mention, we use SPSS for that purpose

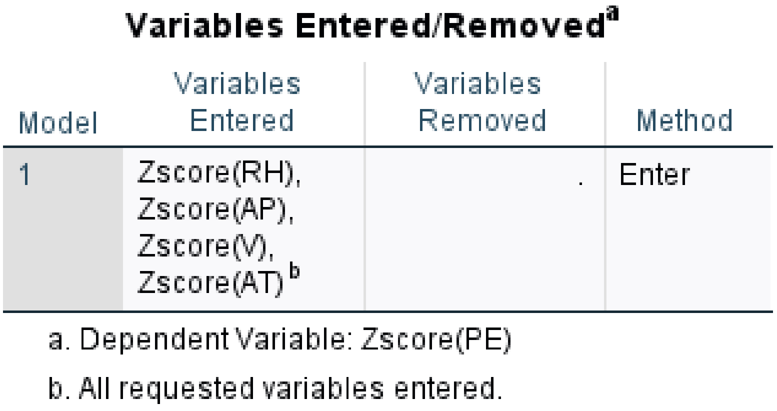


Close enough to a normal distribution. Also, the next plot shows that we have nothing to worry about-the residuals are also homoscedastic



Now, due to various and different units of measurement , we will first run our model with the standardized variables. It is important to see if some of the under consideration factors have anything to offer in our analysis. So, that can be achieve getting this factors to the same “scale”.

Regression with the standardized variables:

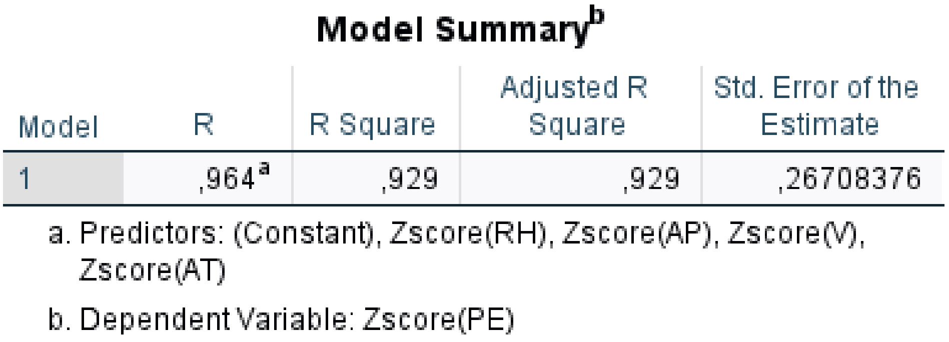


Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Zscore(RH), Zscore(AP), Zscore(V), Zscore(AT)[b]	.	Enter

a. Dependent Variable: Zscore(PE)

b. All requested variables entered.

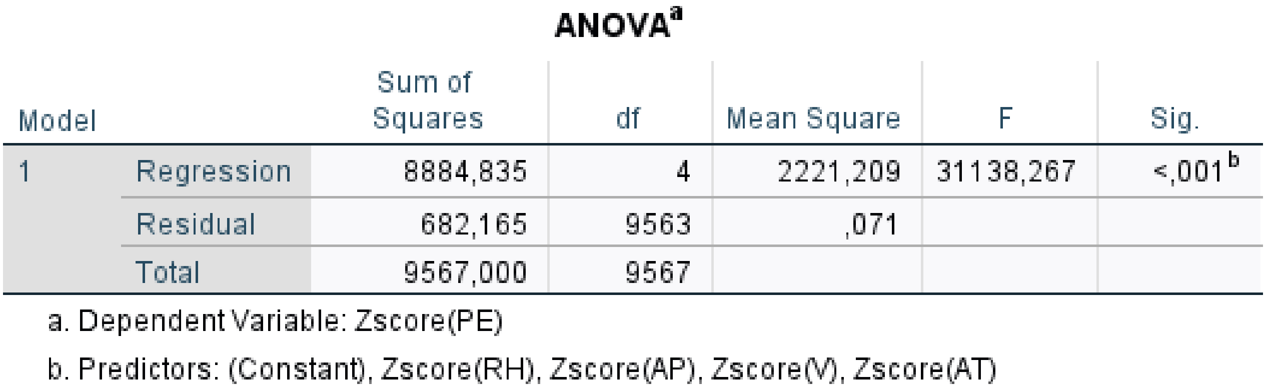


Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,964[a]	,929	,929	,26708376

a. Predictors: (Constant), Zscore(RH), Zscore(AP), Zscore(V), Zscore(AT)

b. Dependent Variable: Zscore(PE)



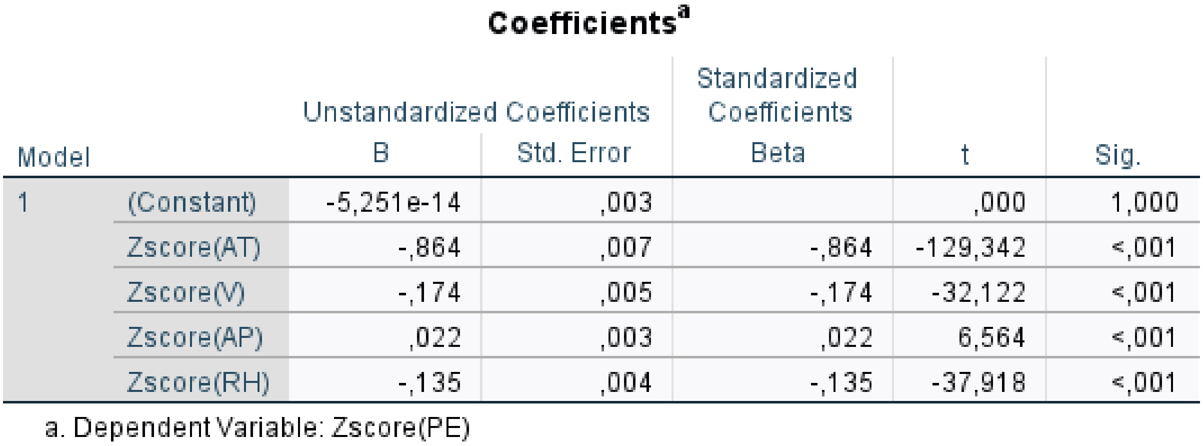
ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8884,835	4	2221,209	31138,267	<,001[b]
	Residual	682,165	9563	,071		
	Total	9567,000	9567			

a. Dependent Variable: Zscore(PE)

b. Predictors: (Constant), Zscore(RH), Zscore(AP), Zscore(V), Zscore(AT)

Obviously, as we can see from the high score of the R and from the Sig of the F-test in the ANOVA, a linear model can adjust very well in our data. Now we can surely continue our analysis.



Coefficients[a]

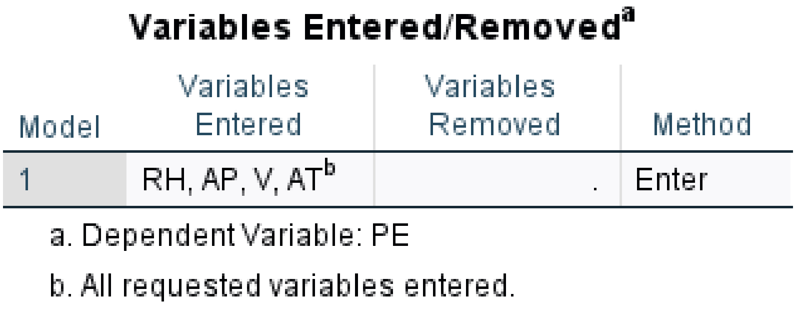
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-5,251e-14	,003		,000	1,000
	Zscore(AT)	-,864	,007	-,864	-129,342	<,001
	Zscore(V)	-,174	,005	-,174	-32,122	<,001
	Zscore(AP)	,022	,003	,022	6,564	<,001
	Zscore(RH)	-,135	,004	-,135	-37,918	<,001

a. Dependent Variable: Zscore(PE)

The last panel shows that all the coefficients of our independent variables are significant. However, it would be nice if we could simplify the things even more. Looking at the Standardized Beta coefficients, AT shows the highest absolute value, indicating it is the strongest predictor so the AT weighted more in our model. We will talk about this later, for completeness reasons lets run our model with the initial unstandardized variables (now we take on consideration and the units of measurement)

.

Regression with the unstandardized variables:

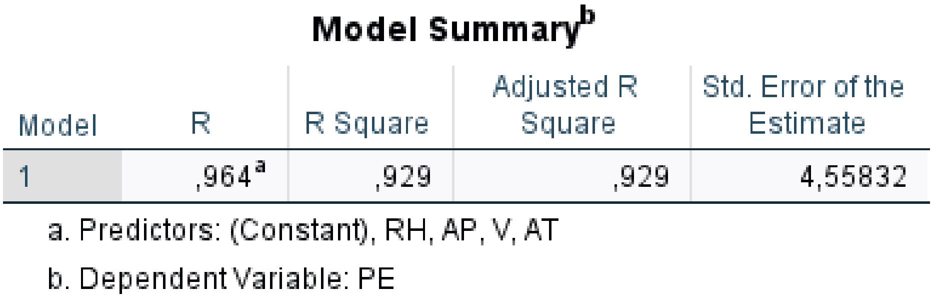


Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	RH, AP, V, AT[b]	.	Enter

a. Dependent Variable: PE

b. All requested variables entered.

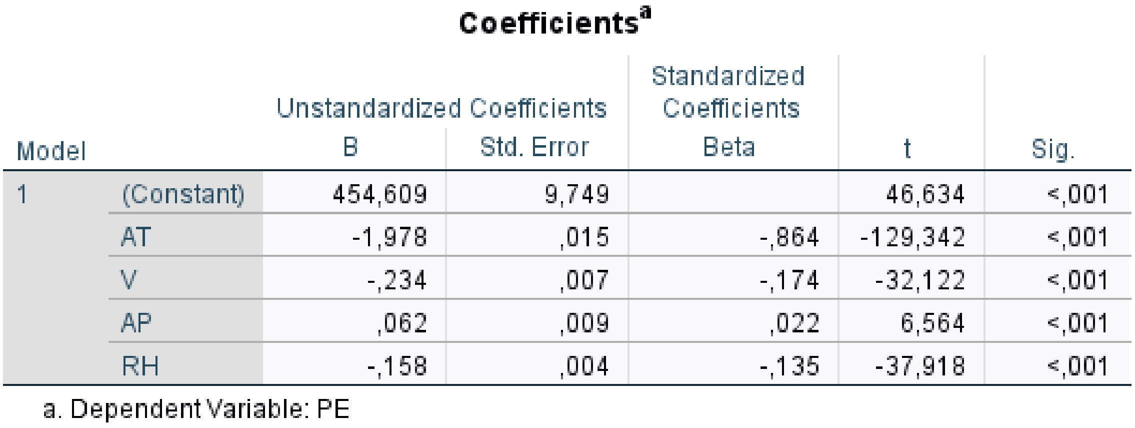


Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,964[a]	,929	,929	4,55832

a. Predictors: (Constant), RH, AP, V, AT

b. Dependent Variable: PE



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	454,609	9,749		46,634	<,001
	AT	-1,978	,015	-,864	-129,342	<,001
	V	-,234	,007	-,174	-32,122	<,001
	AP	,062	,009	,022	6,564	<,001
	RH	-,158	,004	-,135	-37,918	<,001

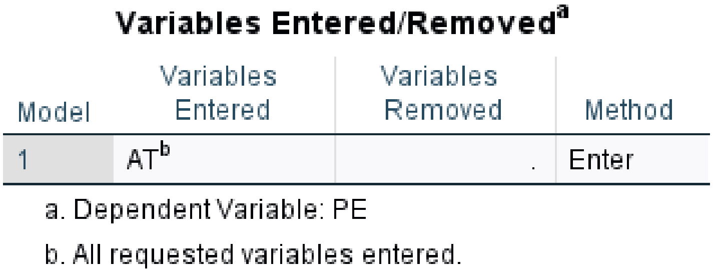
a. Dependent Variable: PE

Again, all the coefficients are significant, and the equation that our linear model produce is this

Y = 454.6 - 1.978*AT – 0.234*V + 0.62*AP – 0.158*RH

As we promised, lets simplify the things now and see what will happen. We run our model only with the AT as independent variable.

Regression only with the AT variable :

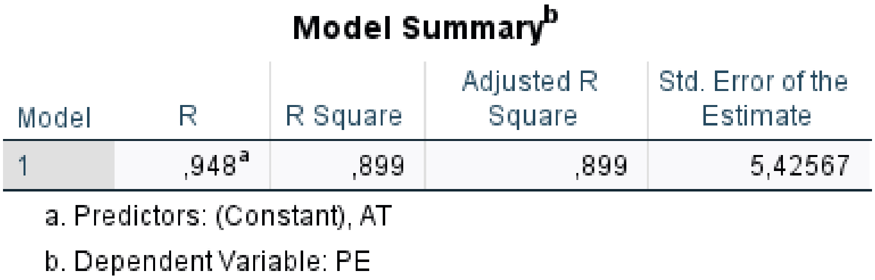


Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	AT[b]	.	Enter

a. Dependent Variable: PE

b. All requested variables entered.

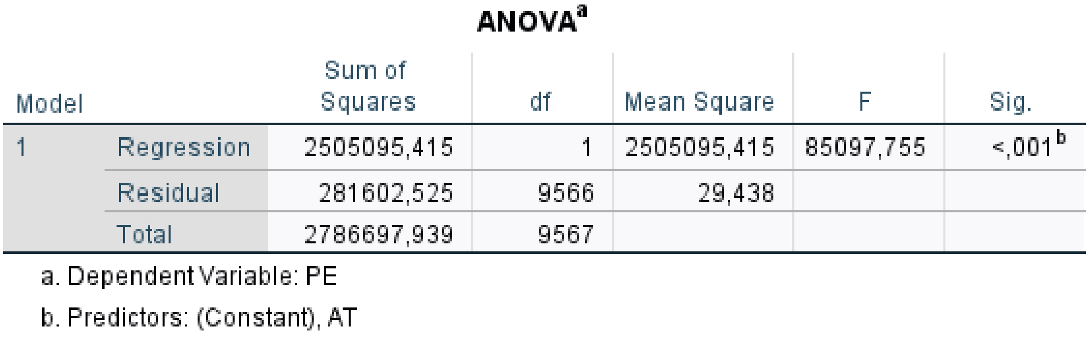


Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,948[a]	,899	,899	5,42567

a. Predictors: (Constant), AT

b. Dependent Variable: PE

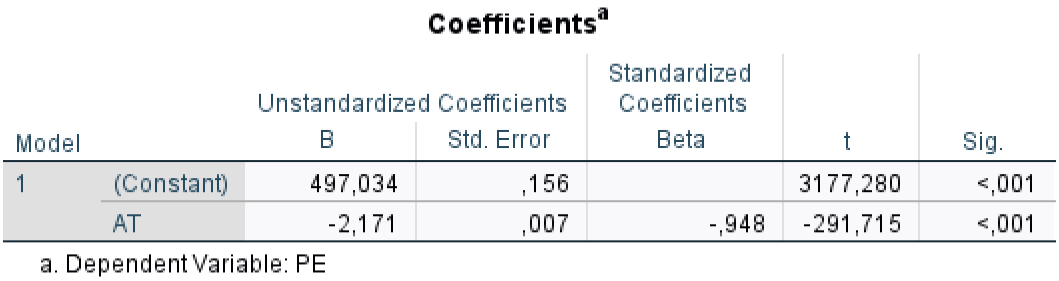


ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2505095,415	1	2505095,415	85097,755	<,001[b]
	Residual	281602,525	9566	29,438		
	Total	2786697,939	9567			

a. Dependent Variable: PE

b. Predictors: (Constant), AT



Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	497,034	,156		3177,280	<,001
	AT	-2,171	,007	-,948	-291,715	<,001

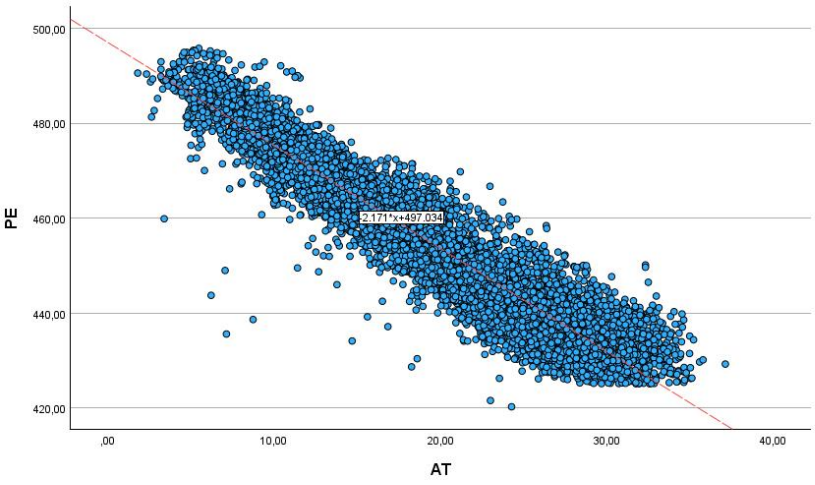
a. Dependent Variable: PE

And the produced equation is: **Y = 497.034 -2.171*AT**

The model fits the data almost perfectly proving that temperature alone is an excellent predictor, also its simpler and now we can have a very nice plot, but first let's see what this equation means more carefully.

CONCLUSIONS

The final simple linear regression model demonstrates a strong inverse relationship between ambient temperature and power output. Specifically, for every 1°C increase in temperature, the plant's electrical energy output is predicted to decrease by approximately 2.171 MW. This result is highly significant ($p < 0.001$), confirming that temperature is a critical driver of thermodynamic efficiency in combined cycle power plants. Also, if the temperature is zero the CCPP produce 497.034 MW.


PE
-2.171*x+497.034
AT