Anova Regression Correlation analysis. A portfolio of work in Statistical Techniques with SPSS

Article · January 2022

CITATION
1

READS
2,155

1 author:

Ilias Kalemis
University of Derby
5 PUBLICATIONS **1** CITATION

SEE PROFILE

Anova Regression Correlation analysis. A portfolio of work in Statistical Techniques with SPSS.

Ilias Kalemis [1]

Derby University, Kedleston Rd, Derby DE2, GB.
i.kalemis@mc-class.gr

Contents

Abstract. The specific report will focus on statistical technics that may be applied in any multivariate dataset. Nowadays the data can give us as clear answers for queries and hypothesis that may observed as it is easier than ever to collect data and analyze them. As we do know a standard statistical procedure involves the collection from dataset where the model deployment such as mean, standard deviation, regression bivariate analysis, anova, chi square test, correlation' can either give us a pattern recognition or a range within a decision maker can have the best results from it. These techniques are helpful in providing insights about data however, the final decision should remain into human favor where we must have or give to others our structural opinion based on statistical analysis. The subject report will analyze two different datasets and try to implement some of the structural techniques in order to give to the readers results based on the analysis charts and tables where it may lead him into valuable decisions. The report will extract the results from a statistical software which is one of the best tools for these techniques SPSS [2].

Keywords: SPSS, Statistical analysis, multivariate analysis, regression analysis, bivariate analysis, anova, chi square test, correlation, t – tests analysis.

1 Introduction

The subject report will distinct the main vulnerabilities that statistical techniques may have on the implementation in two datasets. In the following sections we will have one multivariate analysis, regression analysis and five bivariate analysis such as anova, chi square test, correlation, and 2 t – tests. The first datasets that we will use on the subject report is Gas Turbine CO and NOx Emission. The dataset contains 36733 instances of 11 sensor measures aggregated over one hour, from a gas turbine located in Turkey for the purpose of studying flue gas emissions, namely CO and NOx [3]. Second dataset is the AI4I 2020 Predictive Maintenance Dataset is a synthetic dataset that reflects real predictive maintenance data encountered in industry [4].

2 Regression and correlation analysis techniques

In this section will start with implementation of regression correlation techniques on gas turbine Co and NOx emissions dataset. In statistical techniques we use correlation to denote association between two quantitative variables. We also assume that the association is linear, that one variable increase or decreases a fixed amount for a unit increase or decrease in the other metrics.

2.1 Regression analysis techniques

The other technique that is often used in these circumstances is regression, which involves estimating the best straight line to summarize the association. a regression analysis model is presented with dependent variable the CO emissions from gas turbines and independent variables the ambient temperature (AT), pressure (AP) mbar, the humidity (AH) (%), the air filter difference pressure (AFDP) mbar, the gas turbine exhaust pressure (GTEP) mbar, the turbine inlet temperature (TIT), the turbine after temperature (TAT) C, the compressor discharge pressure (CDP) mbar and the turbine energy yield (TEY) MWH (Table 1. Descriptive statistics). The data were part from the UCI Machine Learning Repository in regards of Gas Turbine CO and NOx Emission Data Set Data Set. The data concerned the article were written from Heysem, PÄ±nar & ErdinÃ§ Uzun (2019) and aimed to predict CO and NOx emissions from gas turbines[3]. They were collected in an operating range of the turbine between partial load (75%) and full load (100%). The initial data set contained 36733 instances and concerned five years, 2011 – 2015. In this study we analyzed the data from year 2015 which contained 7384 observations.

	N	Minimum	Maximum	Mean	Std. Deviation
Ambient temperature	7384	-62,35	99,97	25,5812	22,96355

Ambient pressure	7384	989,4	1036,6	1014,509	6,8954
Ambient humidity	7384	,03	96,67	61,8800	24,17223
Air filter difference pressure	7384	,00	5,24	3,2558	1,16020
Gas turbine exhaust pressure	7384	,02	40,72	23,5340	8,92576
Turbine inlet temperature	7384	1016,0	1100,4	1078,975	19,7624
Turbine after temperature	7384	516,04	550,59	546,6425	5,48907
Compressor discharge pressure	7384	,01	100,00	11,0304	5,84373
Carbon monoxide	7384	,00	10,00	2,5968	1,74456
Nitrogen oxides	7384	,00	10,00	5,3126	2,05560
Turbine energy yield	7384	100,02	179,50	133,9934	16,17921

Table 1. Descriptive statistics

In the next set of charts graphs, we saw how each independent variable is related to the dependent variable. Thus we took each variable and tried to scale it, based on the graph charts it can be easily determined that there is a relationship between the dependent variable and each of the independent variables or in the case there is a relationship it does not seem that it has a linear form.

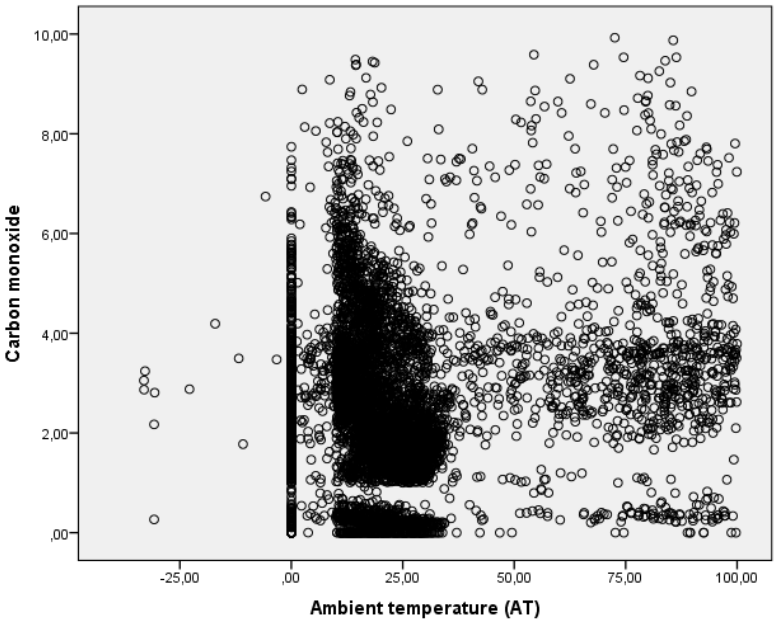

Figure 1. Scatterplot between CO and temperature

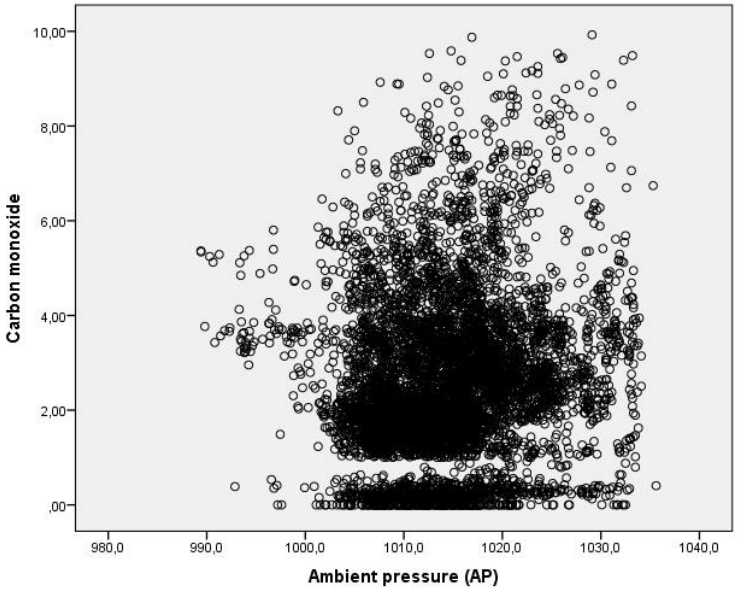

Figure 2. Scatterplot between CO and pressure

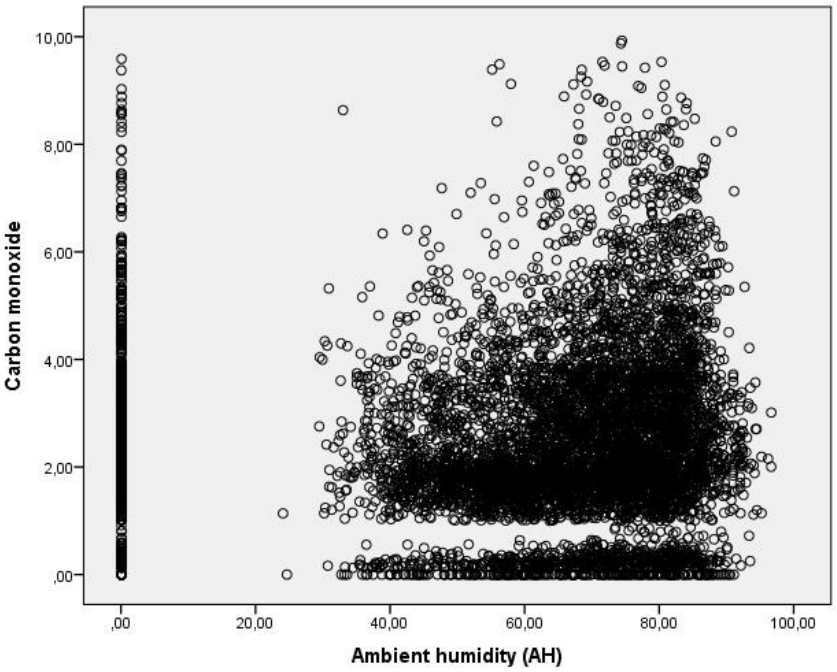

Figure 3. Scatterplot between CO and humidity

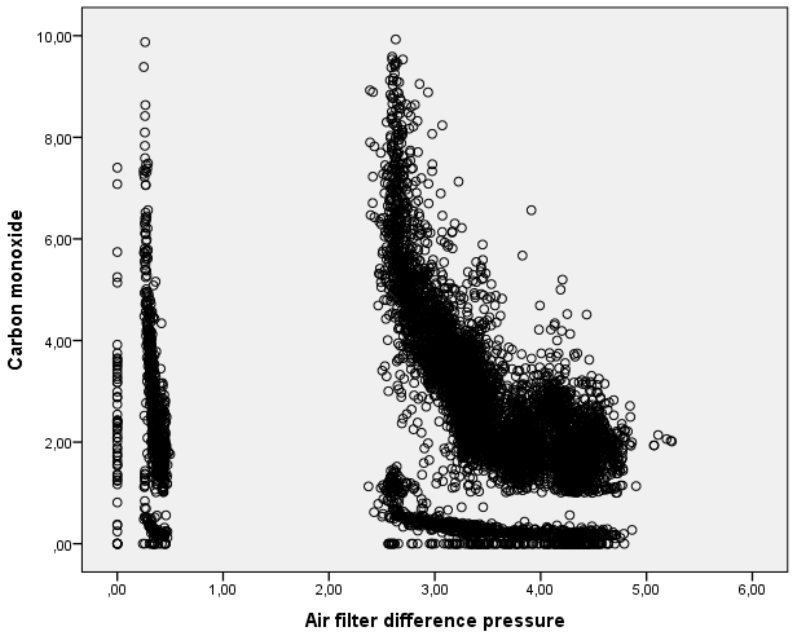

Figure 4. Scatterplot between CO and air filter pressure

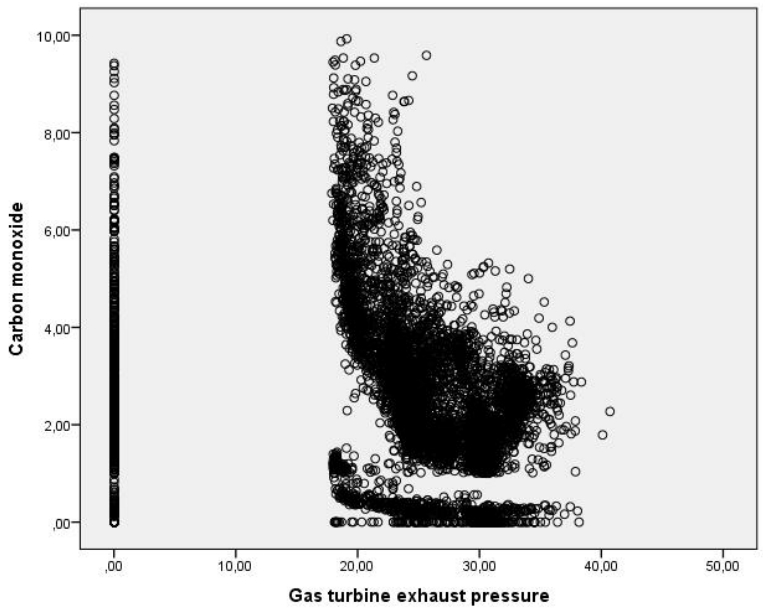

Figure 5. Scatterplot between CO and gas turbine pressure

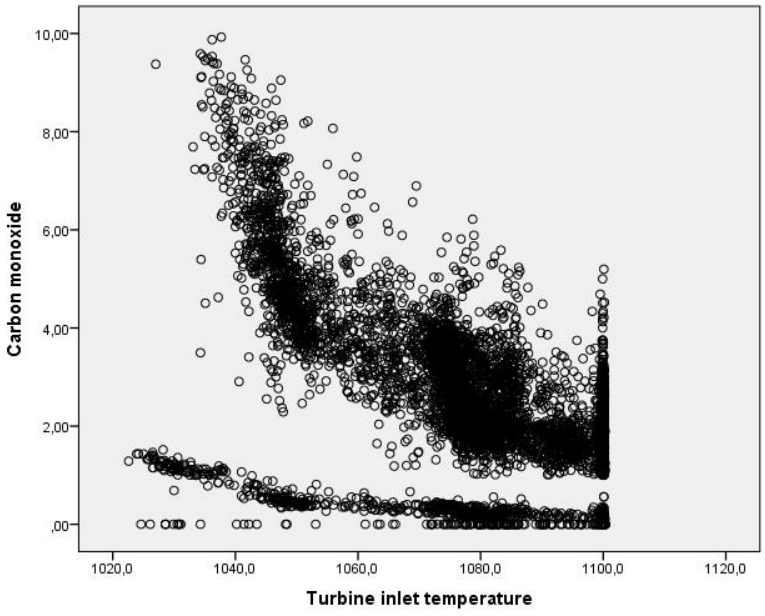

Figure 6. Scatterplot between CO and turbine inlet temperature

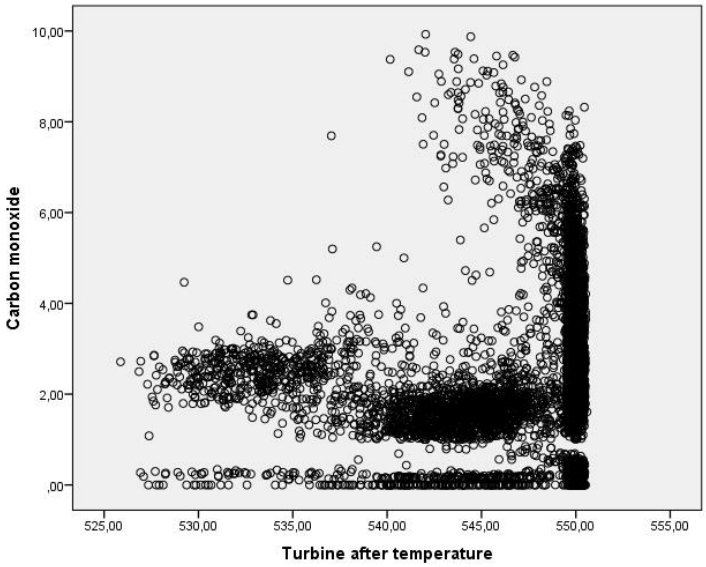

Figure 7. Scatterplot between CO and turbine after temperature

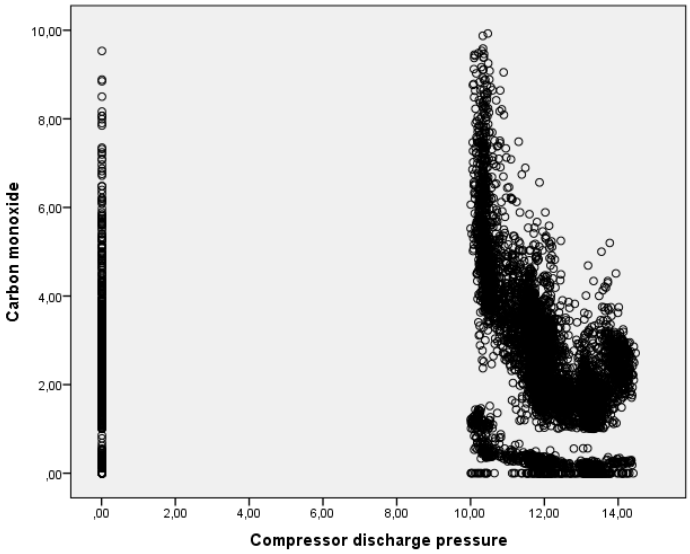

Figure 8. Scatterplot between CO and compressor discharge pressure

2.3 Correlation analysis techniques

In this section will analyze for the first dataset the correlation within the variables. A commonly definition of correlation is the following. We can describe the degree which of two variables are linearly related. This is an important step in bi-variate data analysis. In the broadest sense correlation is actually any statistical relationship, whether causal or not, between two random variables in bivariate data. Moreover, the correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A correlation of -1.0 shows a perfect negative correlation, while a correlation of 1.0 shows a perfect positive correlation. A correlation of 0.0 shows no linear relationship between the movement of the two variables. From the summary review on the class we saw that the Pearson correlation [5] can evaluate only a linear relationship between two continuous variables. A relationship is linear only when a change in one variable is associated with a proportional change in the other variable. However, the Spearman correlation [6] can evaluate a monotonic relationship between two variables where can be continuous or ordinal and it is based on the ranked values for each variable rather than the raw data. The Spearman's rho index was used to investigate the relationships among the variables. This nonparametric test is more suitable when it does not know whether the normality assumption is met and due to the fact that the linearity assumption is not met according to figures 1 – 8 that we have seen above [7]. It can be seen that the dependent variable, carbon monoxide is correlated negatively or positively with the dependent variables. The strongest correlation is with the turbine inlet temperature (rho = -.553, $p < .01$) and the weakest is with ambient temperature (rho = -.063, $p < .01$). Also, it can be seen that there is a statistically significant positive correlation of strong intensity between turbine inlet temperature and the turbine energy yield (rho = .958, $p < .01$). Because the correlation is very strong, we have excluded the turbine energy yield variable from the regression analysis in order to avoid a multicollinearity problem.

	Ambient temperature (AT)	Ambient pressure (AP)	Ambient humidity (AH)	Air filter difference pressure	Gas turbine exhaust pressure	Turbine inlet temperature	Turbine after temperature	Compressor discharge pressure	Carbon monoxide	Turbine energy yield
Ambient temperature (AT)	1,000	-,023*	-,085**	,159**	,079**	,152**	-,139**	,101**	-,063**	,127**
Ambient pressure (AP)	-,023*	1,000	,063**	-,108**	-,059**	-,099**	-,107**	-,033**	,148**	-,016

Ambient humidity (AH)	-,085**	,063**	1,000	-,150**	-,212**	-,228**	,057**	-,162**	,130**	-,170**
Air filter difference pressure	,159**	-,108**	-,150**	1,000	,568**	,764**	-,420**	,593**	-,467**	,723**
Gas turbine exhaust pressure	,079**	-,059**	-,212**	,568**	1,000	,758**	-,439**	,624**	-,388**	,771**
Turbine inlet temperature	,152**	-,099**	-,228**	,764**	,758**	1,000	-,511**	,773**	-,553**	**,958****
Turbine after temperature	-,139**	-,107**	,057**	-,420**	-,439**	-,511**	1,000	-,457**	,257**	-,562**
Compressor discharge pressure	,101**	-,033**	-,162**	,593**	,624**	,773**	-,457**	1,000	-,422**	,778**
Carbon monoxide	-,063**	,148**	,130**	-,467**	-,388**	-,553**	,257**	-,422**	1,000	-,494**
Turbine energy yield	,127**	-,016	-,170**	,723**	,771**	,958**	-,562**	,778**	-,494**	1,000

Table 1. Correlation table

2.4 Regression model first execution

In table 3 it can be seen the adjusted R square of each model where the stepwise method was used in order to result to the most suitable model and the Durbin Watson index [8] for the last one. As it can be seen the adjusted coefficient of determination is equal to 0.373 which means that the independent variables explain 37.3% of the variability of the dependent variable. Also, it can be seen that the Durbin Watson index is equal to 1.495 as acceptable values from Andy's Field model can be between 1 to 3 [9].

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson

1	,596[a]	,355	,355	1,40119	
2	,607[b]	,369	,369	1,38600	
3	,610[c]	,372	,372	1,38295	
4	,611[d]	,373	,373	1,38178	
5	,611[e]	,374	,373	1,38109	1,495

Table 2. Regression Model Summary

2.5 Anova

Our next statistical indicator of the dataset will be the ANOVA. Where ANOVA is Analysis of variance and it is a statistical method that separates observed variance data into different components to use for additional tests. A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables. In table 4 it can be seen that the regression model is statistically significant, F (5, 7378) = 880.469, p =.000.

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regres-sion	7976,671	1	7976,671	4062,806	,000[b]
	Residual	14493,377	7382	1,963		
	Total	22470,049	7383			
2	Regres-sion	8291,191	2	4145,595	2158,047	,000[c]
	Residual	14178,858	7381	1,921		
	Total	22470,049	7383			
3	Regres-sion	8355,363	3	2785,121	1456,227	,000[d]
	Residual	14114,686	7380	1,913		
	Total	22470,049	7383			
4	Regres-sion	8381,298	4	2095,324	1097,429	,000[e]

	Residual	14088,751	7379	1,909		
	Total	22470,049	7383			
5	Regres-sion	8397,116	5	1679,423	880,469	,000[f]
	Residual	14072,933	7378	1,907		
	Total	22470,049	7383			

Table 3. Anova

2.6 Coefficient

In table 5 it can be seen that the predictor variables are all statistical significant, Turbine inlet temperature (b = -.054, p < .01), Ambient temperature (b = .009, p < .01), Turbine after temperature (b = -.013, p < .01), Ambient pressure (b = .010, p < .01) and compressor discharge pressure (b = .008, p < .01). Furthermore, there is not a multicollinearity problem since all VIFs < 10. In addition, in figures 9 and 10 it can be seen that the normality assumption is not met while the homoscedasticity assumption is partially met (Figure 11).

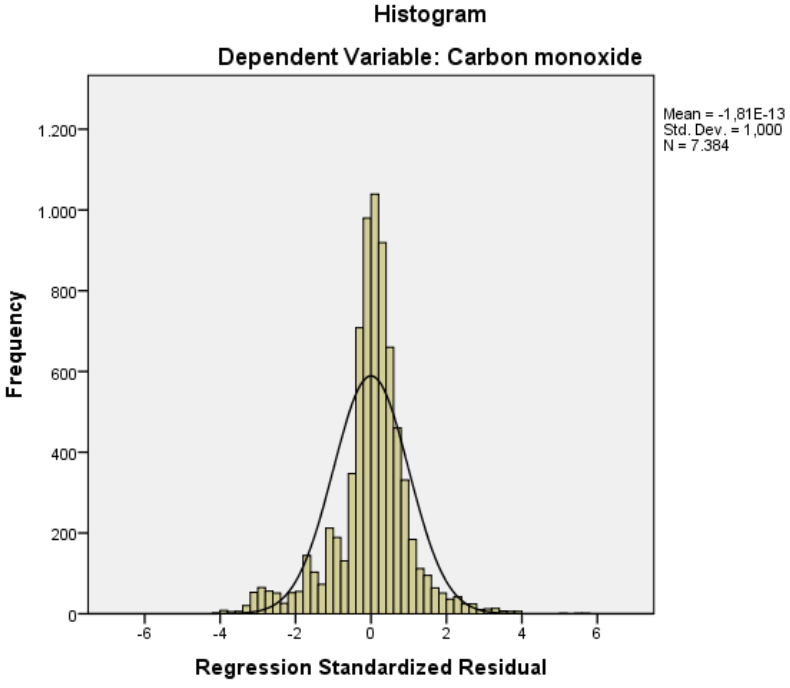

Figure 9. Histogram

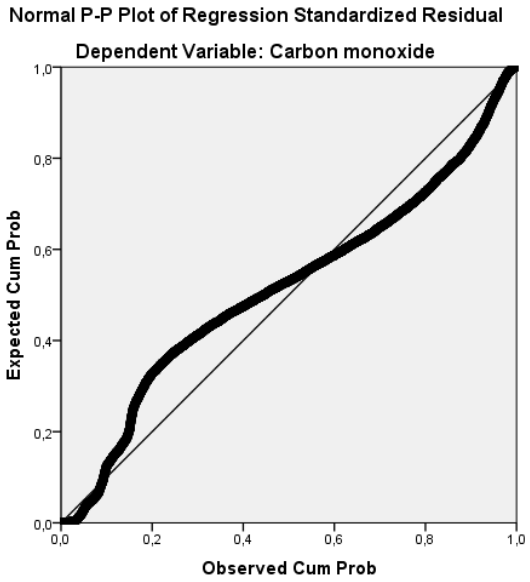

Figure 10. Normality test

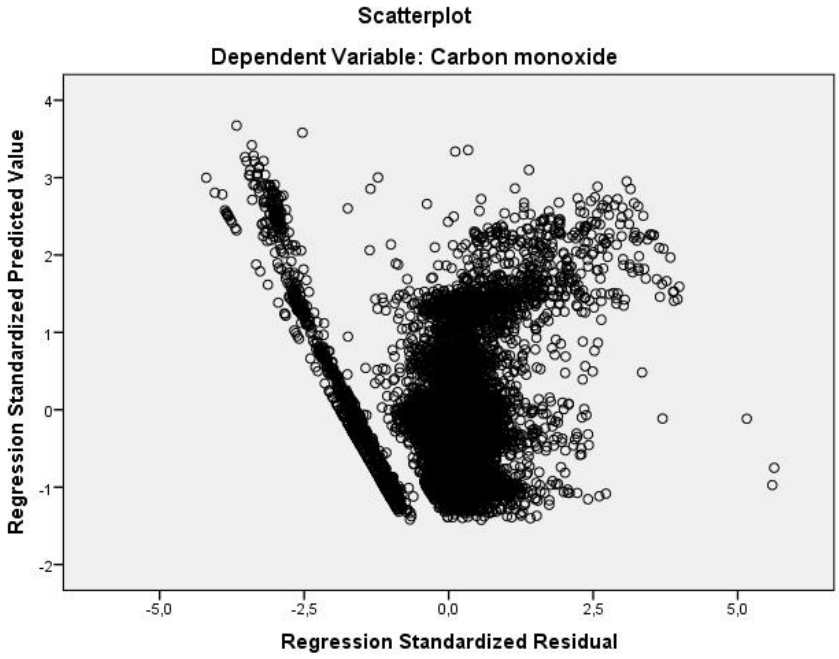

Figure 11. Heteroscedasticity test

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	59,347	,890		66,646	,000		
	Turbine inlet temperature	-,053	,001	-,596	-63,740	,000	1,000	1,000
2	(Constant)	58,577	,883		66,348	,000		

	Turbine inlet temperature	-,052	,001	-,590	-63,753	,000	,998	1,002
	Ambient temperature (AT)	,009	,001	,118	12,796	,000	,998	1,002
3	(Constant)	70,926	2,307		30,748	,000		
	Turbine inlet temperature	-,054	,001	-,613	-60,911	,000	,839	1,191
	Ambient temperature (AT)	,009	,001	,115	12,470	,000	,994	1,006
	Turbine after temperature	-,019	,003	-,058	-5,792	,000	,840	1,190
4	(Constant)	58,343	4,119		14,164	,000		
	Turbine inlet temperature	-,053	,001	-,605	-58,700	,000	,800	1,251
	Ambient temperature (AT)	,008	,001	,112	11,990	,000	,982	1,018
	Turbine after temperature	-,014	,003	-,045	-4,149	,000	,738	1,355
	Ambient pressure (AP)	,009	,003	,037	3,686	,000	,859	1,164
5	(Constant)	57,638	4,124		13,975	,000		

Turbine inlet temperature	-,054	,001	-,606	-58,793	,000	,799	1,252
Ambient temperature	,009	,001	,112	12,075	,000	,981	1,019
Turbine after temperature	-,013	,003	-,042	-3,891	,000	,733	1,365
Ambient pressure (AP)	,010	,003	,038	3,784	,000	,858	1,165
Compressor discharge pressure	,008	,003	,027	2,880	,004	,985	1,015

Table 4. Coefficients Dependent Variable: Carbon monoxide

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	1,0789	6,5142	2,5968	1,06647	7384
Std. Predicted Value	-1,423	3,673	,000	1,000	7384
Standard Error of Predicted Value	,017	,253	,036	,017	7384
Adjusted Predicted Value	1,0780	6,5408	2,5968	1,06675	7384
Residual	-5,79314	7,77733	,00000	1,38063	7384
Std. Residual	-4,195	5,631	,000	1,000	7384
Stud. Residual	-4,201	5,634	,000	1,001	7384
Deleted Residual	-5,81162	7,78559	,00000	1,38299	7384

Stud. Deleted Residual	-4,206	5,646	,000	1,001	7384
Mahal. Distance	,097	245,988	4,999	12,659	7384
Cook's Distance	,000	,073	,000	,002	7384
Centered Leverage Value	,000	,033	,001	,002	7384

Table 5. Residuals statistics a. Dependent Variable: Carbon monoxide

2.7 Regression model second execution

In table 7 it can be seen that the cook's distance and the leverage are below 1. Thus, based on these criteria there is no need to exclude from the analysis observations. However, they are residuals greater than the absolute value of 3.29. In addition, from the M distance we have found the multivariate outliers based on the values of the new probability variable (Probability variable =1- (CDF.CHISQ (MAH_1,5)) which are less than .001 (Identifying Multivariate Outliers in SPSS - Statistics Solutions, 2021). Thus, we have implemented the following filter (Probability.MAH_1 <= .001 & -3.29 < ZRE_1 <3.29) and 99 outliers were excluded from the analysis. The remaining observations are 7285 and we repeated the analysis.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,610[a]	,373	,373	1,37133	
2	,624[b]	,389	,389	1,35297	
3	,629[c]	,395	,395	1,34635	
4	,629[d]	,396	,396	1,34585	1,525

Table 6. Model summary

As it can be seen in table 7 the adjusted coefficient of determination is equal to .396 which means that the independent variables explain 39.6% of the variability of the dependent variable. Also, it can be seen that the Durbin Watson index is equal to 1.525 (acceptable values between 1 – 3, Field, 2005).

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8136,937	1	8136,937	4326,922	,000[b]
	Residual	13695,951	7283	1,881		
	Total	21832,889	7284			
2	Regression	8502,891	2	4251,446	2322,508	,000[c]
	Residual	13329,997	7282	1,831		
	Total	21832,889	7284			
3	Regression	8634,831	3	2878,277	1587,865	,000[d]
	Residual	13198,058	7281	1,813		
	Total	21832,889	7284			
4	Regression	8646,453	4	2161,613	1193,389	,000[e]
	Residual	13186,435	7280	1,811		
	Total	21832,889	7284			

Table 7. Anova

In table 5 it can be seen that the regression model is statistical significant, F(4, 7280) = 1193.389, p =.000.

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF

1	(Constant)	61,612	,897		68,658	,000		
	Turbine inlet temperature	-,055	,001	-,610	-65,779	,000	1,000	1,000
2	(Constant)	60,886	,887		68,654	,000		
	Turbine inlet temperature	-,054	,001	-,606	-66,100	,000	,999	1,001
	Ambient temperature (AT)	,010	,001	,130	14,139	,000	,999	1,001
3	(Constant)	80,957	2,513		32,220	,000		
	Turbine inlet temperature	-,058	,001	-,645	-63,078	,000	,794	1,260
	Ambient temperature (AT)	,009	,001	,118	12,862	,000	,979	1,022
	Turbine after temperature	-,030	,003	-,088	-8,532	,000	,785	1,274
4	(Constant)	72,460	4,190		17,292	,000		
	Turbine inlet temperature	-,057	,001	-,639	-61,014	,000	,755	1,324
	Ambient temperature (AT)	,009	,001	,115	12,338	,000	,956	1,045

Turbine after temperature	-,027	,004	-,080	-7,406	,000	,717	1,395
Ambient pressure (AP)	,006	,003	,025	2,533	,011	,873	1,145

Table 8. Coefficients a. Dependent Variable: Carbon monoxide

In table 9 it can be seen that the predictor variables are all statistical significant (reduced by one in comparison the previous model), Turbine inlet temperature (b = -.057, p < .01), Ambient temperature (b = .009, p < .01), Turbine after temperature (b = -.027, p < .01) and ambient pressure (b = .006, p < .05) . Furthermore, there is not a multicollinearity problem since all VIFs < 10. In addition, in figure 12 it can be seen that the normality assumption is not met while the homoscedasticity assumption is partially met (Figure 13). A potential solution to this is the implementation of a bootstrap regression analysis. However, this was not possible even though we reduced the number of samples (100). The following message was printed from the SPSS22.0: "Available memory was exhausted while compiling output. All the output for this command has been deleted". Table 10 Residuals Statistics.

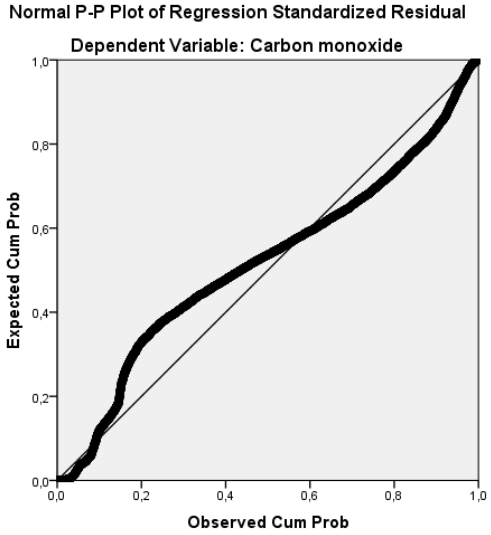

Figure 12.Normality test

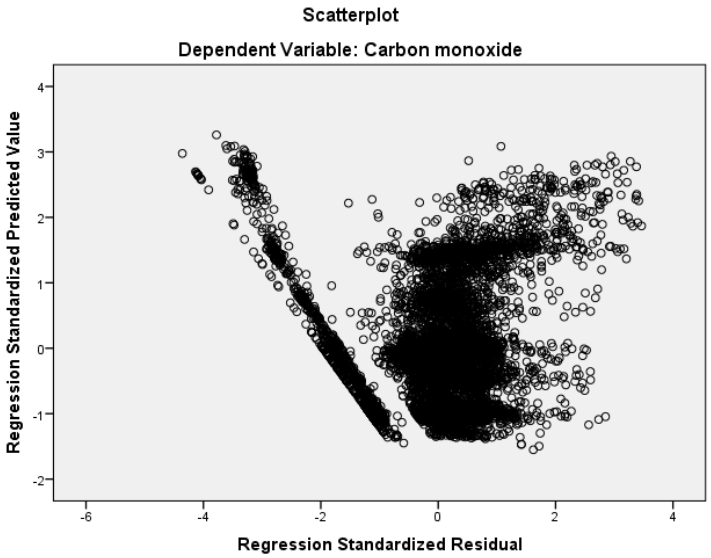

Figure 13. Heteroscedasticity test

Residuals Statistics

	Mini-mum	Maxi-mum	Mean	Std. Devia-tion	N
Predicted Value	,9275	6,0134	2,5676	1,05717	7259
Std. Predicted Value	-1,551	3,259	,000	1,000	7259
Standard Error of Predicted Value	,016	,079	,032	,012	7259
Adjusted Predicted Value	,9226	6,0287	2,5676	1,05736	7259
Residual	-5,71335	4,5397 6	,00000	1,31030	7259
Std. Residual	-4,359	3,464	,000	1,000	7259
Stud. Residual	-4,366	3,466	,000	1,000	7259
Deleted Residual	-5,73096	4,5466 9	,00000	1,31175	7259

Stud. Deleted Residual	-4,371	3,469	,000	1,001	7259
Mahal. Distance	,053	25,390	3,999	3,980	7259
Cook's Distance	,000	,012	,000	,001	7259
Centered Leverage Value	,000	,003	,001	,001	7259

Table 10. Residuals Statistics Dependent Variable: Carbon monoxide

2.8 Parametric and non-parametric models

In this section we present the implementation of the parametric Anova, t – test and non-parametric tests chi square on a data file retrieved from the website UCI Machine Learning Repository: AI4I 2020 Predictive Maintenance Dataset Data Set [4]. The dataset consists of 10000 observations. We have used the following variables: the product quality variants (low, medium, high) and machine failures (yes / no) to implement a chi square test, the product quality variants (low, medium, high) and rotational speed [rpm] to implement an ANOVA test, machine failures (yes /no) and rotational speed [rpm] to implement a t – test and tool wear failure (yes / no) and rotational speed [rpm] to implement a t – test. The variables from the dataset that we will analyze in this section are product I is consisting of a letter L, M, or H. For low (50% of all products), medium (30%) and high (20%) as product quality variants and a variant-specific serial number. The air temperature K generated using a random walk process later normalized to a standard deviation of 2 K around 300 K. The process temperature K generated using a random walk process normalized to a standard deviation of 1 K, added to the air temperature plus 10 K. The rotational speed rpm calculated from a power of 2860 W, overlaid with a normally distributed noise. The torque Nm torque values are normally distributed around 40 Nm with a Ïf = 10 Nm and no negative values. The tool wear min quality variants H/M/L add 5/3/2 minutes of tool wear to the used tool in the process. And last one a machine failure label that indicates whether the machine has failed in this particular datapoint for any of the following failure modes are true.

			Machine failure		
			No	Yes	Total
Type	H	N	982_a	21_b	1003
		%	97,9%	2,1%	100,0%
		Std. Residual	,4	-2,2	

	L	N	5765[a]	235[b]	6000
		%	96,1%	3,9%	100,0%
		Std. Residual	-,4	2,2	
	M	N	2914[a]	83[b]	2997
		%	97,2%	2,8%	100,0%
		Std. Residual	,3	-1,8	
Total		N	9661	339	10000
		%	96,6%	3,4%	100,0%

Table 9. Crosstabulation. Each subscript letter denotes a subset of Machine failure categories whose column proportions do not differ significantly from each other at the ,05 level.

In table 11 it can be seen that 2.1% of the high variant quality products presented machine failure compared to 3.9% of the low variant quality products and the 2.8% quality of the medium variant. This difference is statistical significant according to chi square test, X2(2) = 13.752, p = .001 (Table 12) (The Fisher test results to the sample conclusion even though the assumptions of the chi square test are met, non-zero cell and less than 20% of the cells with frequency below 5). However, the intensity of the relationship is very low according to Cramer's V index (Table 13).

				Monte Carlo Sig. (2-sided)		
					99% Confidence Interval	
	Value	df	Asymp. Sig. (2-sided)	Sig.	Lower Bound	Upper Bound
Pearson Chi-Square	13,752[a]	2	,001	,001[b]	,000	,002
Likelihood Ratio	14,536	2	,001	,001[b]	,000	,001
Fisher's Exact Test	14,021			,001[b]	,000	,002
N of Valid Cases	10000					

Table 12. Chi square test

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 34,00.
b. Based on 10000 sampled tables with starting seed 299883525.

					Monte Carlo Sig.		
						99% Confidence Interval	
			Value	Approx. Sig.	Sig.	Lower Bound	Upper Bound
Nominal by Nominal	Phi		,037	,001	,001[c]	,000	,002
	Cramer's V		,037	,001	,001[c]	,000	,002
N of Valid Cases			1000				
			0				

Table 13. Cramer's V

c. Based on 10000 sampled tables with starting seed 299883525.

2.9 Normality test

In table 14 it can be seen that the rotational speed does not follow the normal distribution in each level of the variant quality product, High, Statistic(1003) = .110, p =.000, Low, Statistic (6000) = .107, p =.000 and medium, Statistic(2997) = .101, p =.000. In this case the Kruskal Wallis test was used.

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Type1	Statistic	df	Sig.	Statistic	df	Sig.
Rotational speed [rpm]	H	,110	1003	,000	,863	1003	,000
	L	,107	6000	,000			

	M	,101	2997	,000	,871	2997	,000

Table 14. Normality test

a. Lilliefors Significance Correction

In table 15 it can be seen that there is not a statistically significant difference among the three variant quality products in relation to the rotational speed, X2(2) = .248, p = .883.

			Rotational speed [rpm]
Chi-Square			,248
df			2
Asymp. Sig.			,883
Monte Carlo Sig.	Sig.		,884[c]
	99% Confidence Interval	Lower Bound	,876
		Upper Bound	,892

Table 15. Kruskal Wallis test

a. Kruskal Wallis Test

b. Grouping Variable: Type1

c. Based on 10000 sampled tables with starting seed 1335104164.

2.10 First T-Test

In table 16 it can be seen that the rotational speed does not follow the normal distribution in each level of the machine failure, yes, Statistic (339) = .339, p =.000 and no, Statistic(9661) = .096, p =.000. In this case the bootstrap method must be used. However, because it is very difficult to present the bootstrap method (calculation problems) we present the t – test.

		Kolmogorov-Smirnov[a]	Shapiro-Wilk

	Machine failure	Statistic	df	Sig.	Statistic	df	Sig.
Rotational speed [rpm]	No	,096	9661	,000			
	Yes	,339	339	,000	,548	339	,000

Table 16. Normality test

a. Lilliefors Significance Correction

In table 17 it can be seen that homogeneity assumption is not met, F = 295.620, p = .001. It can be seen that in the case of mechanical failure in comparison to the case of no mechanical failure the average of rotational speed is higher, t (342.500) = 2.087, p = .038.

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Rotational speed [rpm]	Equal variances assumed	295,620	,000	4,423	9998	,000	43,773	9,898	24,372	63,175

	Equal variances not assumed	2,087	342,500	,038	43,773	20,977	2,514	85,032

Table 10. t – test

2.11 Second T-Test

In table 18 it can be seen that the rotational speed does not follow the normal distribution in each level of the tool wear failure, yes, Statistic (46) = .156, p =.007 and no, Statistic(9954) = .104, p =.000. In this case the bootstrap method could be used. However, because it is very difficult to present the bootstrap method, we present the t – test.

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	TWF	Statistic	df	Sig.	Statistic	df	Sig.
Rotational speed [rpm]	No	,104	9954	,000			
	Yes	,156	46	,007	,852	46	,000

Table 11. Normality test

a. Lilliefors Significance Correction

In table 19 it can be seen that the homogeneity assumption is met, F = 1.507, p = .200. It can be seen that in the case of tool wear mechanical failure in comparison to the case of no mechanical failure the average of rotational speed is higher, t(9998) = .220, p = .299.

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Rotational speed [rpm]	Equal variances assumed	1,507	,220	-1,039	9998	,299	-27,524	26,495	-79,460	24,411
	Equal variances not assumed			-,906	45,316	,370	-27,524	30,392	-88,725	33,676

Table 19. T - test

3 Conclusion and purpose of the research

The statistical analysis of these two datasets prove that we can have a better understanding of what models prove. For the first model where the dataset can be well used for predicting turbine energy yield (TEY) using ambient variables as features. We saw that the initial target which initially aimed to predict CO and NOx emissions from gas turbines had strong relation of the carbon monoxide and the ambient temperature. However the data as they were collected in an operating range of the turbine between partial load (75%) and full load (100%) can give us a strong indicator as it was used almost

in the pick of the usage as it a=was greater than 75% of its load. The second dataset of our analysis was about the machine failure where it consists of five independent failure modes. The tool wear failure (TWF) the tool will be replaced of fail at a randomly selected tool wear time between 200 and 240 mins. At this point in time, the tool is replaced 69 times, and fails 51 times we also saw that the heat dissipation causes a process failure, if the difference between air- and process temperature is below 8.6 K and the tools rotational speed is below 1380 rpm. This is the case for 115 data points.
power failure (PWF). The product of torque and rotational speed (in rad/s) equals the power required for the process. If this power is below 3500 W or above 9000 W, the process fails, which is the case 95 times in our dataset. Finally, the random failures (RNF) of each process has a chance of 0,1 % to fail regardless of its process parameters. This is the case for only 5 datapoints, less than could be expected for 10,000 datapoints in the dataset.

4 References

1. Author, Ilias Kalemis. A portfolio of work in Statistical Techniques on SPSS software.
2. SPSS. Software https://www.ibm.com/analytics/spss-statistics-software , last accessed on 2021/5/20.
3. 1st dataset from the paper: Heysem Kaya, PÄ±nar TÃ¼fekci and ErdinÃ§ Uzun. 'Predicting CO and NOx emissions from gas turbines: novel data and a benchmark PEMS', Turkish Journal of Electrical Engineering & Computer Sciences, vol. 27, 2019, pp. 4783-4796. Heysem Kaya, Department of Information and Computing Sciences, Utrecht University, 3584 CC, Utrecht, The Netherlands Email: h.kaya'@' uu.nl and PÄ±nar TÃ¼fekci, Ã‡orlu Faculty of Engineering, NamÄ±k Kemal University, TR-59860 Ã‡orlu, TekirdaÄŸ, Turkey and Email: ptufekci '@' nku.edu.tr , link: https://archive.ics.uci.edu/ml/datasets/Gas+Turbine+CO+and+NOx+Emission+Data+Set , last accessed on 2021/5/20.
4. 2nd dataset from the paper: Stephan Matzka, 'Explainable Artificial Intelligence for Predictive Maintenance Applications', Third International Conference on Artificial Intelligence for Industries (AI4I 2020), 2020 (in press). Stephan Matzka, School of Engineering - Technology and Life, Hochschule fÃ¼r Technik und Wirtschaft Berlin, 12459 Berlin, Germany, stephan.matzka '@' htw-berlin.de https://archive.ics.uci.edu/ml/datasets/AI4I+2020+Predictive+Maintenance+Dataset , last accessed on 2021/5/20.
5. Pearson correlation https://libguides.library.kent.edu/SPSS/PearsonCorr , last accessed on 2021/5/20.
6. Spearman correlation Spearman C. (1904). "The proof and measurement of association between two things". American Journal of Psychology. 15 (1): 72–101. doi:10.2307/1412159. JSTOR 1412159, last accessed on 2021/5/20.
7. Mavuto Mukaka (2012). Statistics corner: A guide to appropriate use of correlation coefficient in medical research. Malawi medical journal : the journal of Medical Association.
8. Durbin, J.; Watson, G. S. (1950). "Testing for Serial Correlation in Least Squares Regression, I". Biometrika. 37 (3–4): 409–428. doi:10.1093/biomet/37.3-4.409. JSTOR 2332391.
9. Discovering Statistics Using IBM SPSS Statistics, Field, A. (2005). Discovering statistics using SPSS (2nd ed.). Sage Publications, Inc

5 Appendices

```
Regression & Correlation
DATASET ACTIVATE DataSet1.
GRAPH
  /SCATTERPLOT(BIVAR)=AT WITH CO
  /MISSING=LISTWISE.
GRAPH
  /SCATTERPLOT(BIVAR)=AP WITH CO
  /MISSING=LISTWISE.
GRAPH
  /SCATTERPLOT(BIVAR)=AH WITH CO
  /MISSING=LISTWISE.
GRAPH
```

```
  /SCATTERPLOT(BIVAR)=AFDP WITH CO
  /MISSING=LISTWISE.
GRAPH
  /SCATTERPLOT(BIVAR)=GTEP WITH CO
  /MISSING=LISTWISE.


GRAPH
  /SCATTERPLOT(BIVAR)=TIT WITH CO
  /MISSING=LISTWISE.
GRAPH
  /SCATTERPLOT(BIVAR)=TAT WITH CO
  /MISSING=LISTWISE.

GRAPH
  /SCATTERPLOT(BIVAR)=CDP WITH CO
  /MISSING=LISTWISE.
NONPAR CORR
  /VARIABLES=AT AP AH AFDP GTEP TIT TAT CDP CO TEY
  /PRINT=SPEARMAN TWOTAIL NOSIG
  /MISSING=PAIRWISE.
DESCRIPTIVES VARIABLES=AT AP AH AFDP GTEP TIT TAT CDP CO
TEY
  /STATISTICS=MEAN STDDEV MIN MAX.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA COLLIN TOL
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT CO
  /METHOD=STEPWISE AT AP AH AFDP GTEP TIT TAT CDP
  /SCATTERPLOT=(*ZPRED ,*ZRESID)
  /RESIDUALS DURBIN HISTOGRAM(ZRESID) NORMPROB(ZRESID)
  /CASEWISE PLOT(ZRESID) OUTLIERS(3)
  /SAVE MAHAL COOK LEVER ZRESID SDBETA SDFIT.

COMPUTE Probability.MAH_1=1- (CDF.CHISQ (MAH_1,5)).
EXECUTE.
USE ALL.
COMPUTE filter_$=(Probability.MAH_1 > .001 &  -3.29<
ZRE_1 <3.29).
VARIABLE LABELS filter_$ 'Probability.MAH_1 > .001 &  -
3.29< ZRE_1 <3.29 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
```

```
FILTER BY filter_$.
EXECUTE.
Chi square


CROSSTABS
  /TABLES=Type BY Machinefailure
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ PHI
  /CELLS=COUNT ROW SRESID BPROP
  /COUNT ROUND CELL
  /METHOD=MC CIN(99) SAMPLES(10000).

Anova
EXAMINE VARIABLES=Rotationalspeedrpm BY Type1
  /PLOT BOXPLOT STEMLEAF NPPLOT
  /COMPARE GROUPS
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.

NPAR TESTS
  /K-W=Rotationalspeedrpm BY Type1(1 3)
  /MISSING ANALYSIS
  /METHOD=MC CIN(99) SAMPLES(10000).

T - test (1st)
EXAMINE VARIABLES=Rotationalspeedrpm BY Machinefailure
  /PLOT BOXPLOT STEMLEAF NPPLOT
  /COMPARE GROUPS
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.

BOOTSTRAP
  /SAMPLING METHOD=SIMPLE
  /VARIABLES TARGET=Rotationalspeedrpm
INPUT=Machinefailure
  /CRITERIA CILEVEL=95 CITYPE=BCA  NSAMPLES=5000
  /MISSING USERMISSING=EXCLUDE.
T-TEST GROUPS=Machinefailure(0 1)
  /MISSING=ANALYSIS
  /VARIABLES=Rotationalspeedrpm
```

```
  /CRITERIA=CI(.95).

T – test (2nd)

EXAMINE VARIABLES=Rotationalspeedrpm BY TWF
  /PLOT BOXPLOT STEMLEAF NPPLOT
  /COMPARE GROUPS
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.

T-TEST GROUPS=TWF(0 1)
  /MISSING=ANALYSIS
  /VARIABLES=Rotationalspeedrpm
  /CRITERIA=CI(.95).
```

6 Datasets

1) Gas Turbine CO and NOx Emission Data Set Data Set
WebSource:
https://archive.ics.uci.edu/ml/datasets/Gas+Tu
bine+CO+and+NOx+Emission+Data+Set#

2) AI4I 2020 Predictive Maintenance Dataset Data Set
WebSource:
https://archive.ics.uci.edu/ml/datasets/AI4I+2020+Predictive+Maintenance+Dataset#