Linear Regression and Exploratory Data Analysis(EDA)

Sreeja Dacharla   
S20200020256  
UG3,ECE,IIITSmail:sreeja.d20@iiits.inDharani Pokuri  
S20200020294  
UG3,ECE,IIITSmail:dharani.p20@iiits.in

***Abstract*—This article discusses linear regression and exploratory data analysis of a given dataset. A brief introduction to each of the topics is given. The dataset is explained thoroughly.The article consults about the models that would be able to fit the dataset we used.**

***Keywords—Linear regression, k-fold cross-validation, Factor analysis, Structural equation modeling(SEM).***

# **1.Introduction**

Exploratory data analysis is used by analysts to deduct information from datasets. This is often paired with visualization for easy understanding. EDA methodology can vary from dataset to dataset. Many EDA techniques can be applied to a given dataset, but the method which maximizes the information and the accuracy of the information we get from the data is considered to be the most efficient.

In the progress of this article, we begin by applying linear regression to our dataset, then linear regression with cross-validation, followed by factor analysis and finally applying Structural equation modeling (SEM).

# **2.Dataset Description**

The dataset consists of a total of 9568 data points with 5 columns. The dataset is collected over a span of 6 years(from 2006 to 2011) from a combined cycle power plant(CCPP). The power plant was set to work with a full load when the data is collected. The columns of the dataset are Temperature(T), Ambient Pressure(P), Relative Humidity(RH), Exhaust Vacuum(V), and Energy output(EP) which are to be predicted by the other four.

The combined cycle power plant(CCPP) is composed of gas turbines(GT) and stream turbines(ST). The electricity is produced by both of these turbines combined in one circle. The Exhaust Vacuum(V) affects the steam turbine(ST) while the gas turbine(GT) is affected by Temperature(T), Ambient Pressure(P), and Relative Humidity(RH).

The dataset variables histograms(Fig.1) and boxplot(Fig.2) are displayed.The correlation between the variables is plotted using seaborn library and heatmaps(Fig.3). The dataset does not have any null values.The dataset does not contain any outliers (all the data points are within 1.5 times of interquartile range from Q1 and Q3).

# **3.Problem Statement**

The dataset is to be analyzed for different deductions. The primary task is to complete linear regression on the dataset with 2-fold cross-validation.

The EDA on this dataset includes Factor analysis and Structural Equation modeling. Different visualization for the dataset and the results are to be presented.

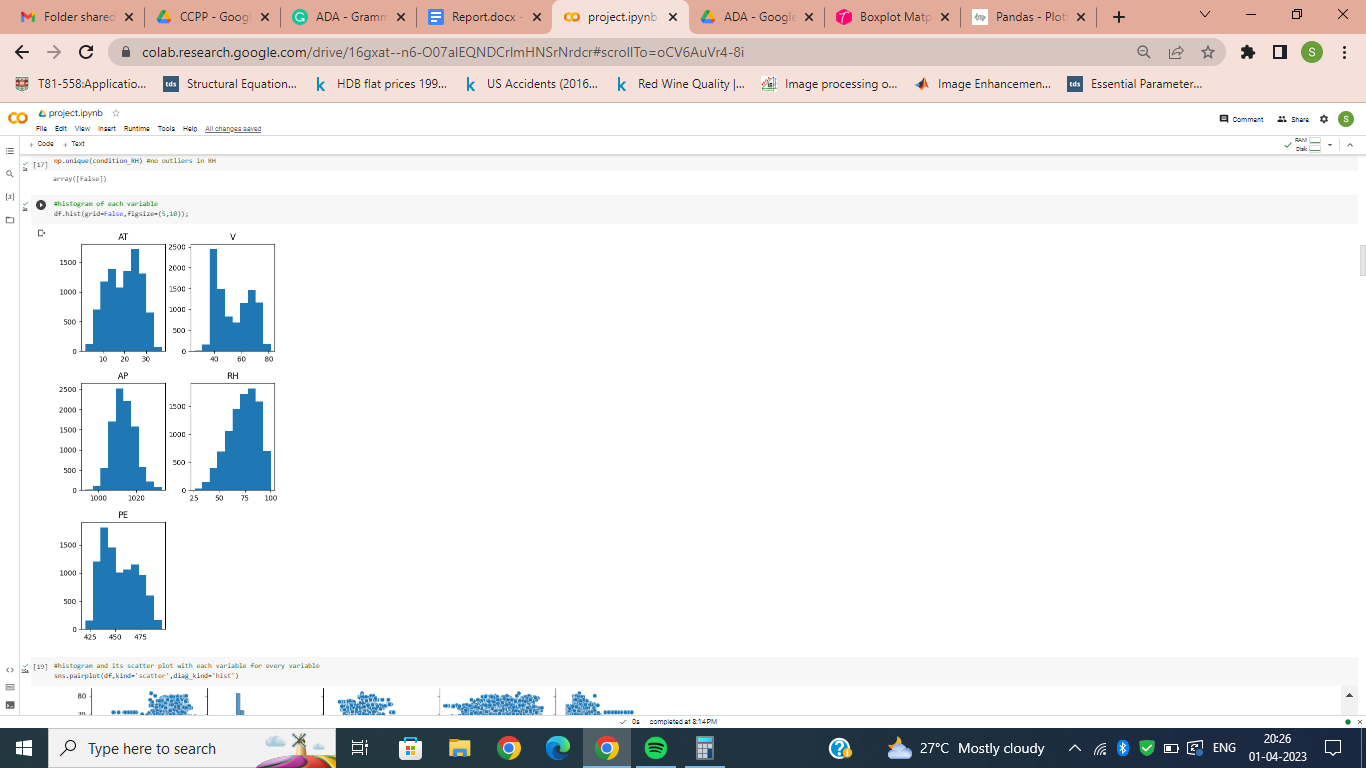


Fig.1.Histogram of the variables

# **4.Methodology**

Initially Linear Regression model is applied to the dataset, followed by factor analysis and concluding with Structural Equation modeling.

For each of these models we tested the assumptions and goodness of fits along with visualizations.

# **5.Linear regression with cross-validation**

The Linear Regression model is:

Y=Z+

The assumptions we have in this model are that the error follows N(0,) .i.e; the mean of error is zero and the error follows normal distribution.

The final solution which estimates is:

=Z`Y

and

est(Y)=Z.est(),

est( = Y-est(Y).

Cross-validation is a technique to avoid overfitting and evaluate the performance of a model. The dataset is shuffled five times and a 2-fold cross validation is applied on each of these five datasets.

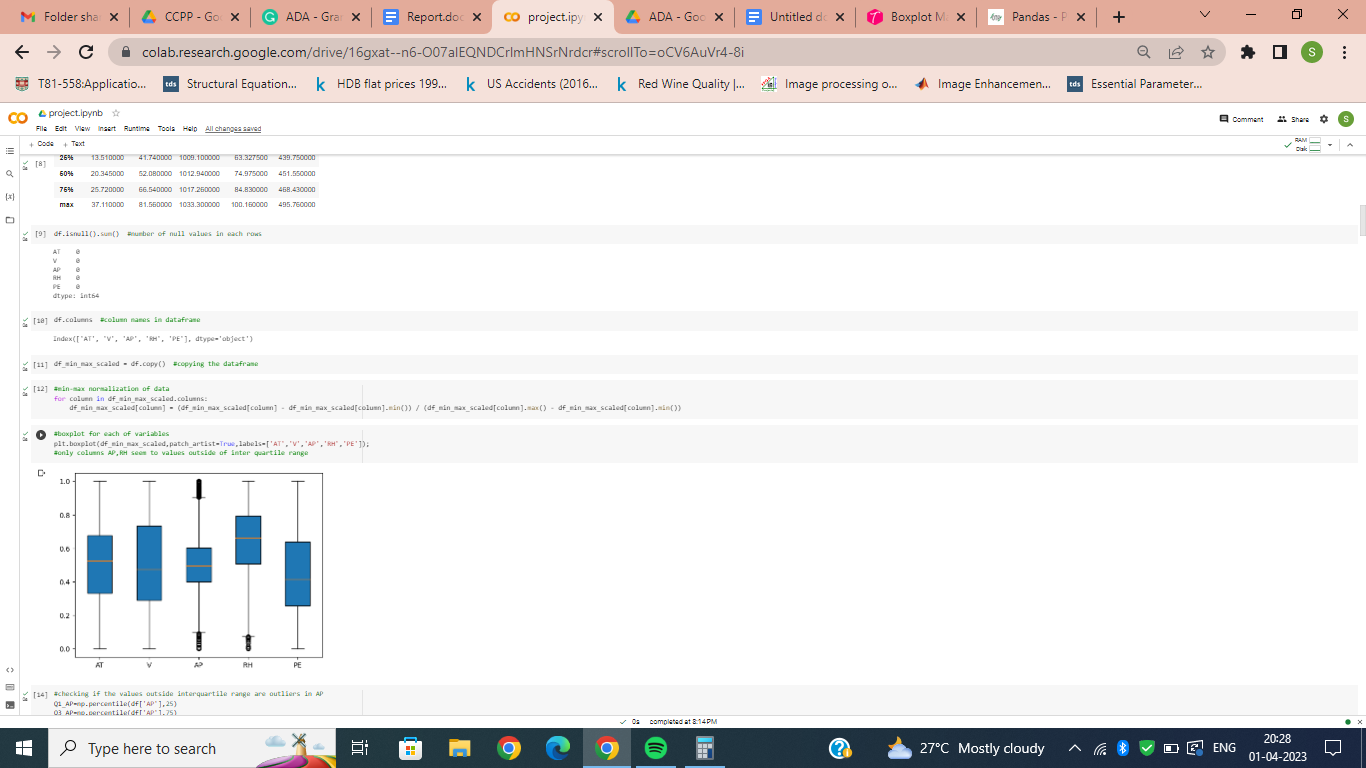


Fig.2.Box plot of variables



Fig.3.Heatmap of correlation of variables

## **5.1 Error Normality**

The Linear Regression model takes the normality of the error function as a prerequisite for the estimation.Since the true Error values can not be estimated by us, the best estimate of errors is taken as our error, i.e; Y-est(Y).

A QQ-plot is plotted for the estimate of errors to show that our true errors follow normal distribution.The QQ-plot of errors is shown in Fig.4.

## **5.2 Assumption Tests**

In Linear Regression a series of assumptions are taken for the final outcome. We can test each of these assumptions using different tests to test the adequacy of the final model.

i)Homoscedasticity

This is an assumption which states that all the errors have the same variance. It can be tested by plotting a graph between est( and est(Y)

The outcome graph should not be a funnel or opposite funnel, if the errors all have the same variance. The graph plotted between est( and est(Y) is shown in Fig.5.

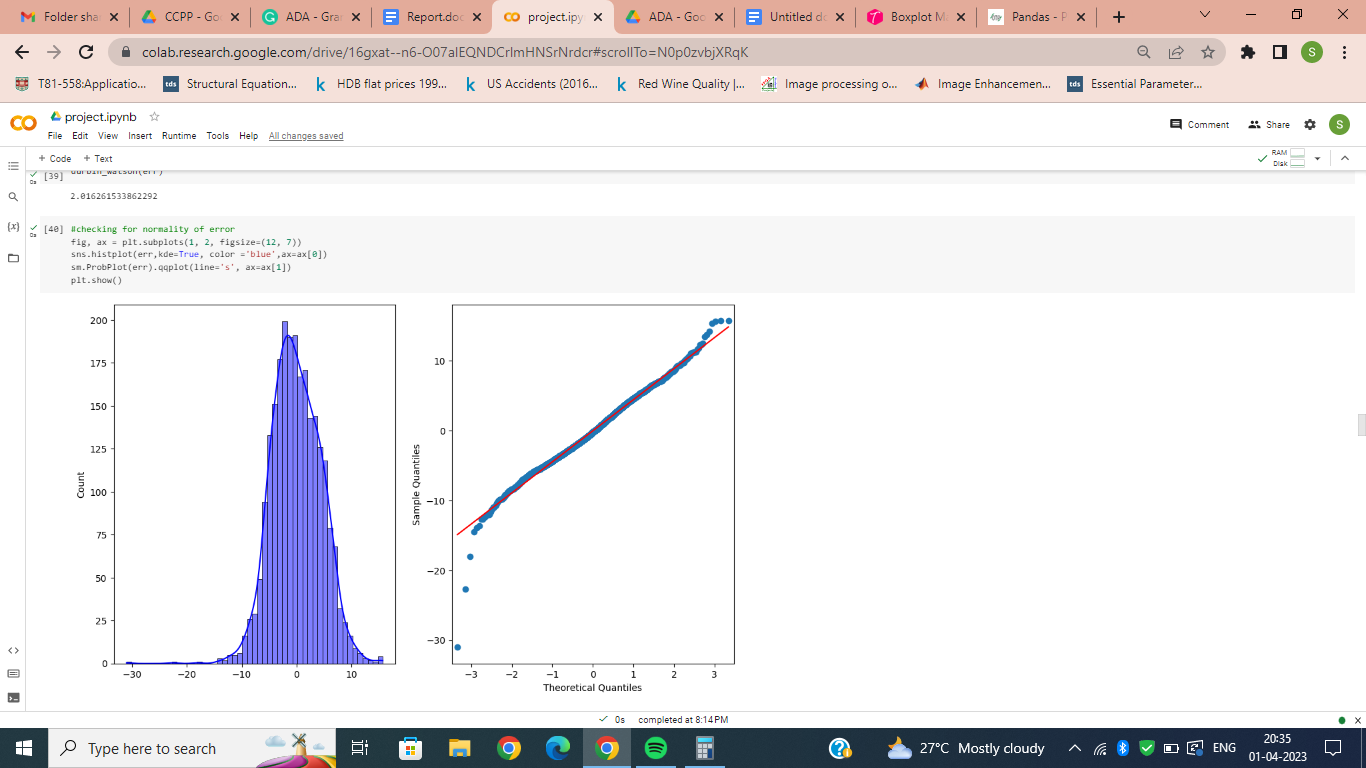


Fig.4.QQ-plot of estimate of errors

ii)Multicollinearity

One of the assumptions in Linear Regression is that the variables of Regression are independent. This is tested by the Variance Inflation Factor(VIF) of each of the variables.

The VIF of each variable is calculated by R2 score of regression with that variable as Y and remaining variables as Z. The relation between VIF and R2 score is as follows:

=1/(1-

The VIF is generally supposed to be less than 10, i.e; R2 score should be less than 0.9. The VIF for each of the variables in the dataset are greater than 10 but the variables are not reduced since the variables in this dataset are very less in number.

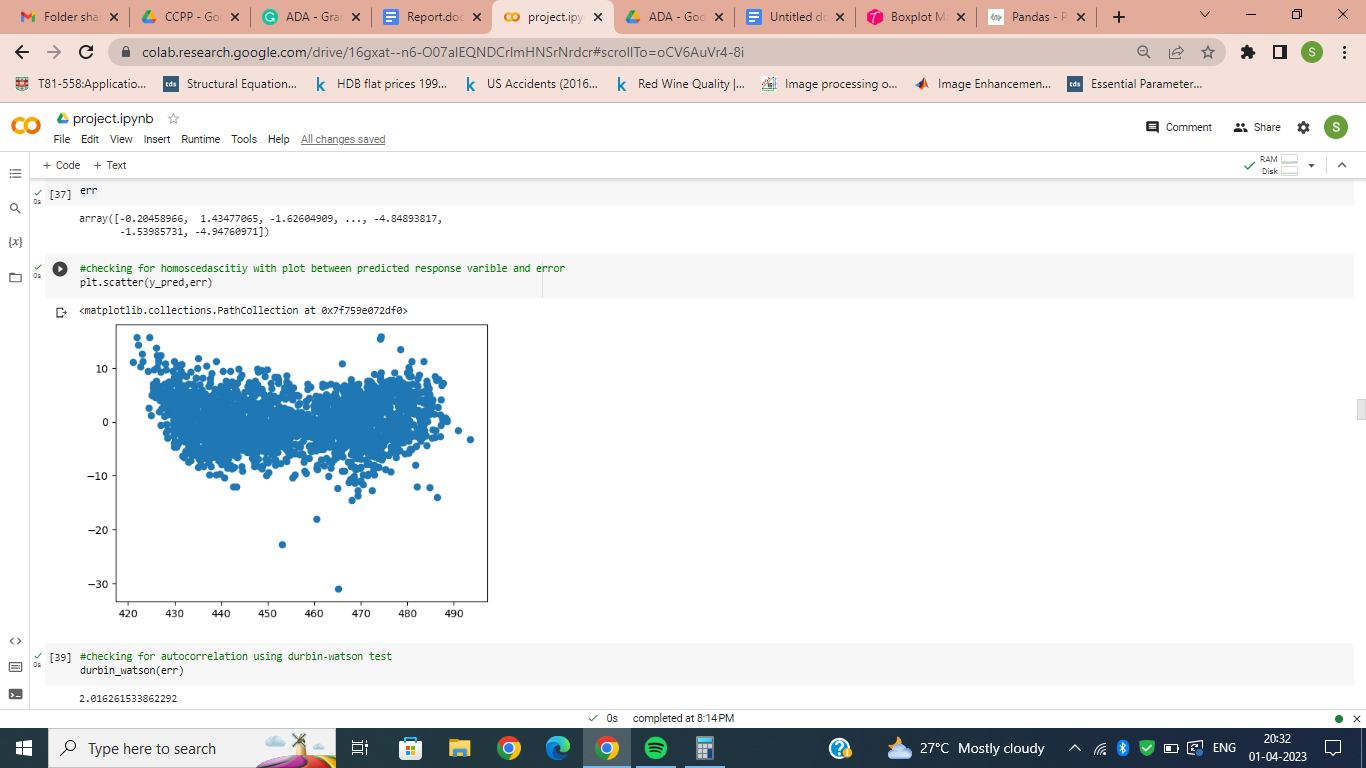


Fig.5.plot for Homoscedasticity

iii)Autocorrelation

This tests the assumption that the random errors are not correlated with each other. The test for this is the Durbin-Watson Test (DW Test).

DW=2(1-r)

r =

If r is more then autocorrelation is present. The DW value is as near as possible to 2. The DW value for the error estimates was nearly 2 for the linear regression.

## **5.3 Linear regression**

Linear regression is accomplished on the dataset with AT, V, AP and RH as regressors and PE as response variable.

First the dataset is divided into train and test sets with a 3:1 ratio. Then a linear Regression model is instantiated from sklearn library.The model is fitted with the training dataset and the regressor part of the testing dataset is given as input for the model to predict the outcomes. Finally the response variable part of the test dataset and predicted values are tested for accuracy of the regression model.

The scatter plot between the response variable part of the test dataset and predicted values is shown in Fig.6.

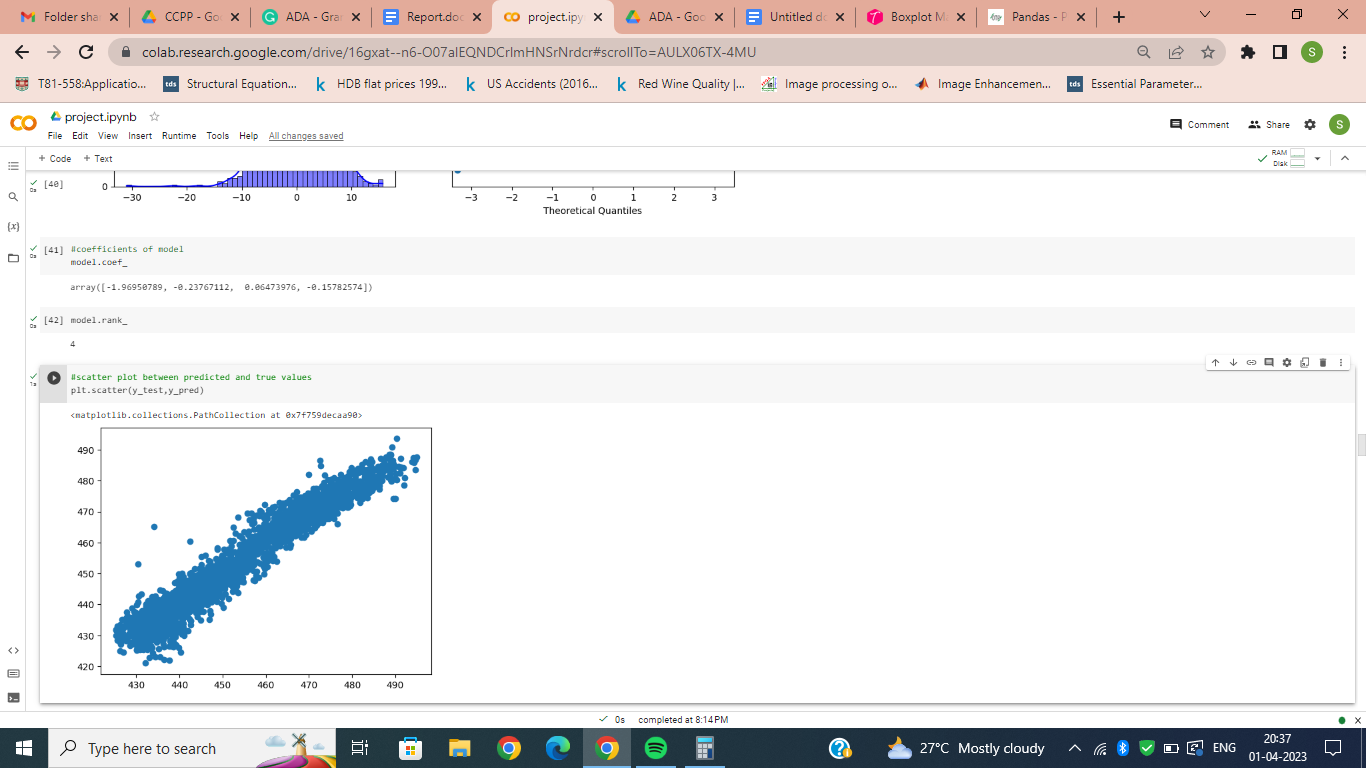
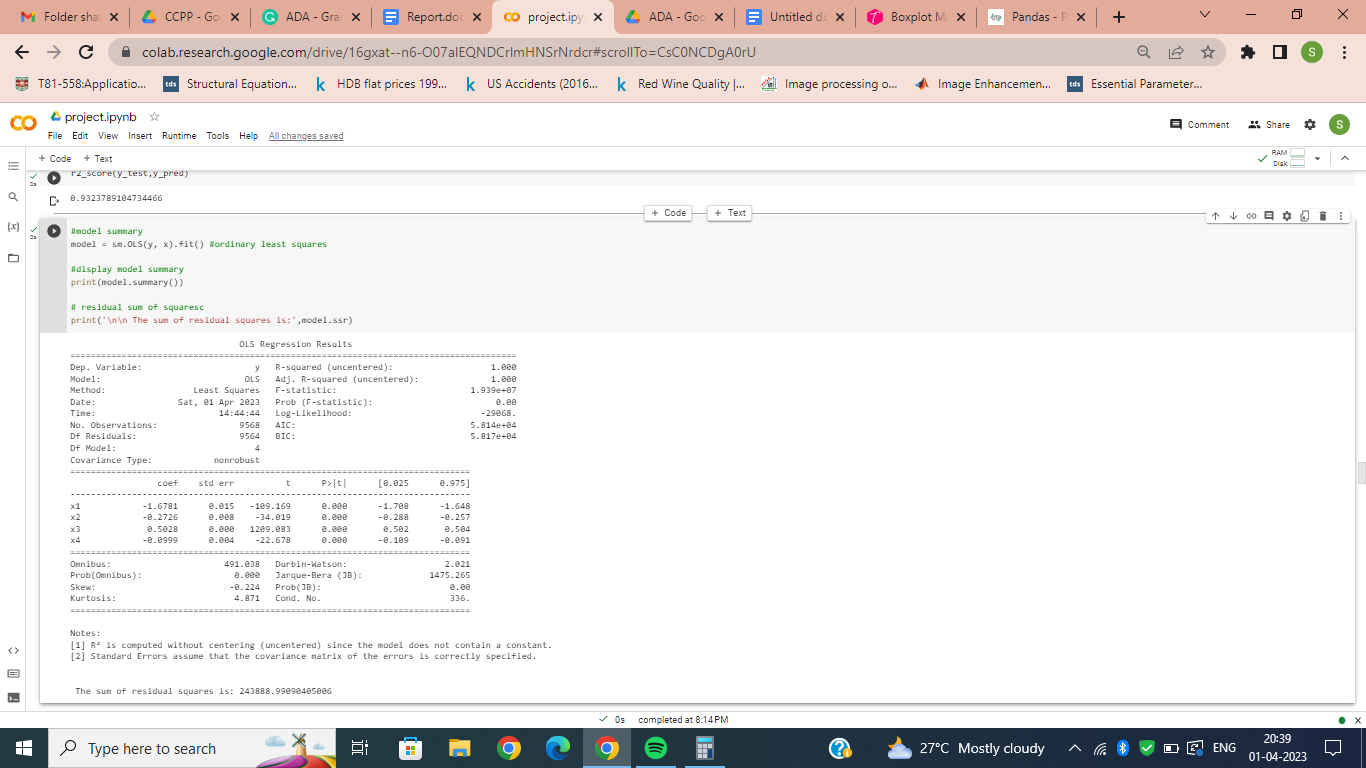


Fig.6.scatter plot between predicted and true values



## **5.4 Metrics**

The metrics for linear regression include R2 score, adjusted R2(coefficient of determination) score, sum of residual squares and so on.

The R2 score of this linear regression is greater than 0.9 which states the fit of the model is good. The OLS(ordinary least squares) is a statistical method used in linear regression analysis to estimate the parameters of a linear model by minimizing the sum of squared residuals.The summary of this model is given in Fig. 7.

## **5.5 Cross-validation**

The dataset is shuffled five times and each of the shuffled dataset is put through a 2-fold cross validation and then linear regression is applied.

K-fold Cross validation divides the dataset into K parts. The model is applied taking the i’th dataset as testing dataset and the remaining K-1 datasets as training datasets. This checks if the model is overfitting to the dataset.

## **5.6 Metrics for cross validation**

The R2 score after the cross validation is not very less than when no cross validation is used. This shows that the linear regression model for this dataset is not over-fitted and gives a good accuracy.

The R2 score for the model after cross validation is still greater than 0.9 which states that the model is a good fit for the dataset.

# **6.Factor analysis**

The Factor Analysis equation is:

X-u = LF +

The assumptions we have in this model are that the error has zero mean and diagonal correlation. The factors F have mean zero and correction matrix close to an identity matrix.

## **6.1 Bartlet’s Sphericity test**

Bartlet’s Sphericity test shows that if the data requires dimensionality reduction. If the original data has independent variables then performing factor analysis would

be an inadequate task. This test checks for the above condition.

The p-value value for the test should be less than 0.05 for us to continue the factor analysis.Our dataset has p less than 0.05.

## **6.2 Assumption tests**

The final factors after analysis should follow certain assumptions for the goodness of the model.

The factors mean should be nearly zero and the correlation matrix should be nearly Identity matrix(Fig.8).These assumptions are followed by the final factors in our analysis

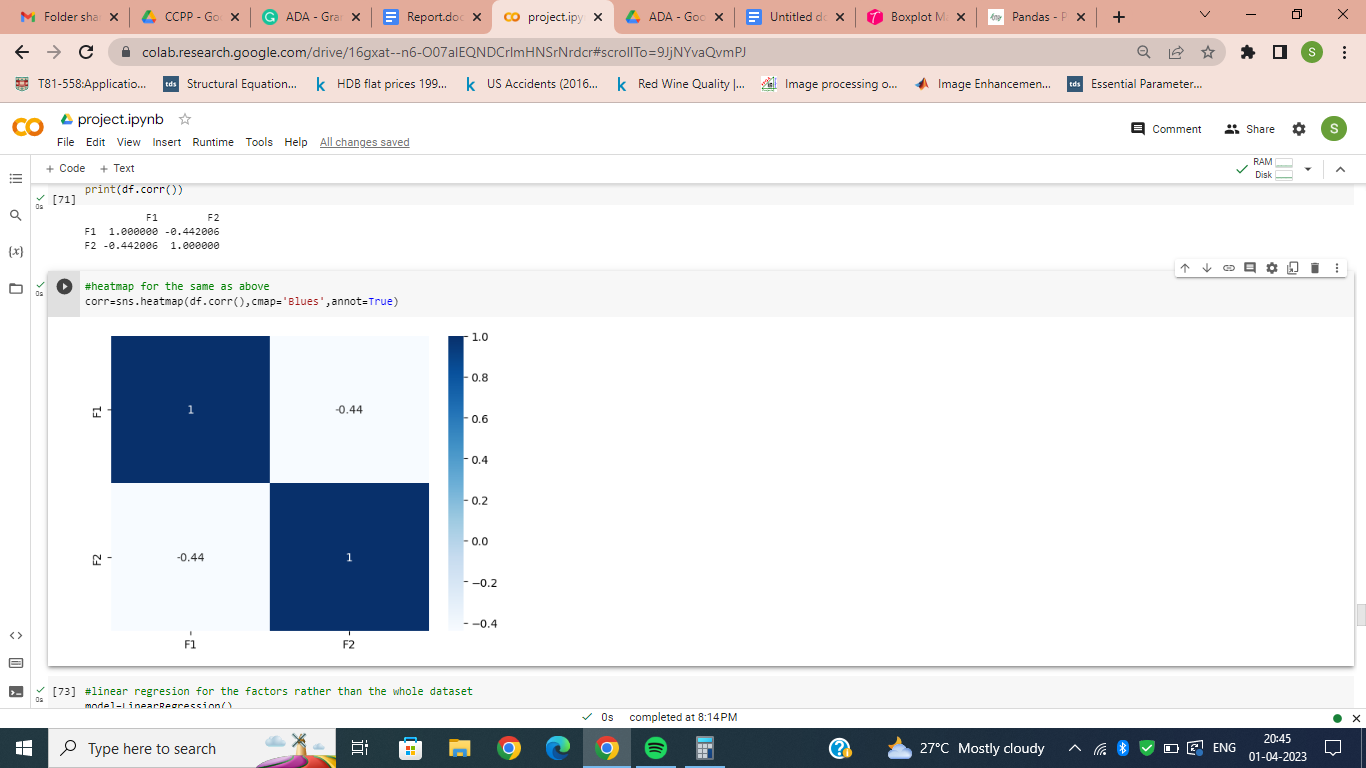


Fig.8.Heatmap of correlation of factors

## **6.3 Scree plot**

There should be tests which conclude how many final factors we should take.One of these tests is Scree plot or elbow plot(Fig.9).

It is a plot of eigen values of correlation matrix of original data in decreasing order.The point where the graph falls drastically is the number of factors we should consider.The number of factors for this dataset are two.

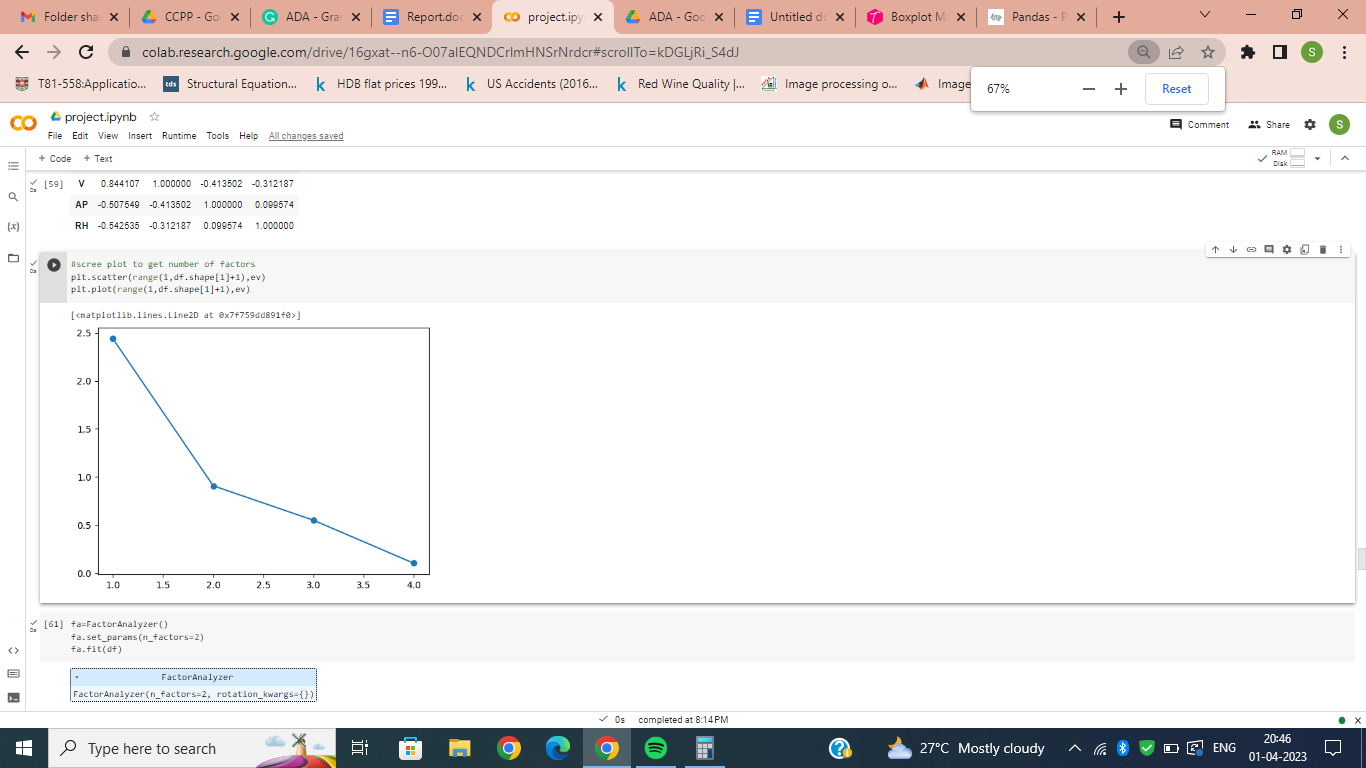


Fig.9.scree plot for number of factors

## **6.4 Factor analysis**

The final factor analysis is done taking two as the final number of factors.

The loadings and communalities are also calculated.A new dataset is synthesized from the loading and the previous original dataset.(this dataset column wise mean is zero and correlation matrix is nearly identity matrix)

Factor variance is taken and both factors cumulatively explain a total of 72% of the data.

## **6.5 Linear regression with final factors**

The data frame with final factors as columns is used for linear regression. When this is done the R2 score is decreased by 5%, which is less.

The number of columns is reduced by half but the accuracy is only reduced by 5%.

## **6.6 Linear regression using cross-validation with final factors**

Linear regression is performed using 2-fold cross validation using the data frame which contains the final factors.The r2-score for this 4% percent reduced which is an insignificant decrease considering our data is halved.

# **7.Structural equation modeling**

Structural Equation modeling is a multivariate analysis technique which combines both Multilinear regression and factor analysis.It is used to analyze the latent variables using manifest variables.

## **7.1 Parameter estimation**

The given dataset is modeling using a SEM technique. The variables Temperature(T), Ambient Pressure(P), and Relative Humidity(RH) affect the gas turbine(GT) while the exhaust Vacuum(V) affects the steam turbine(ST).The final response variable(PE) is taken as CCPP(combined cycle power plant).

All the parameters i.e; the correlation coefficients, casualties are predicted using a model. All the parameters are estimated to minimize an equation of difference of log likelihood functions.(Fig.10)

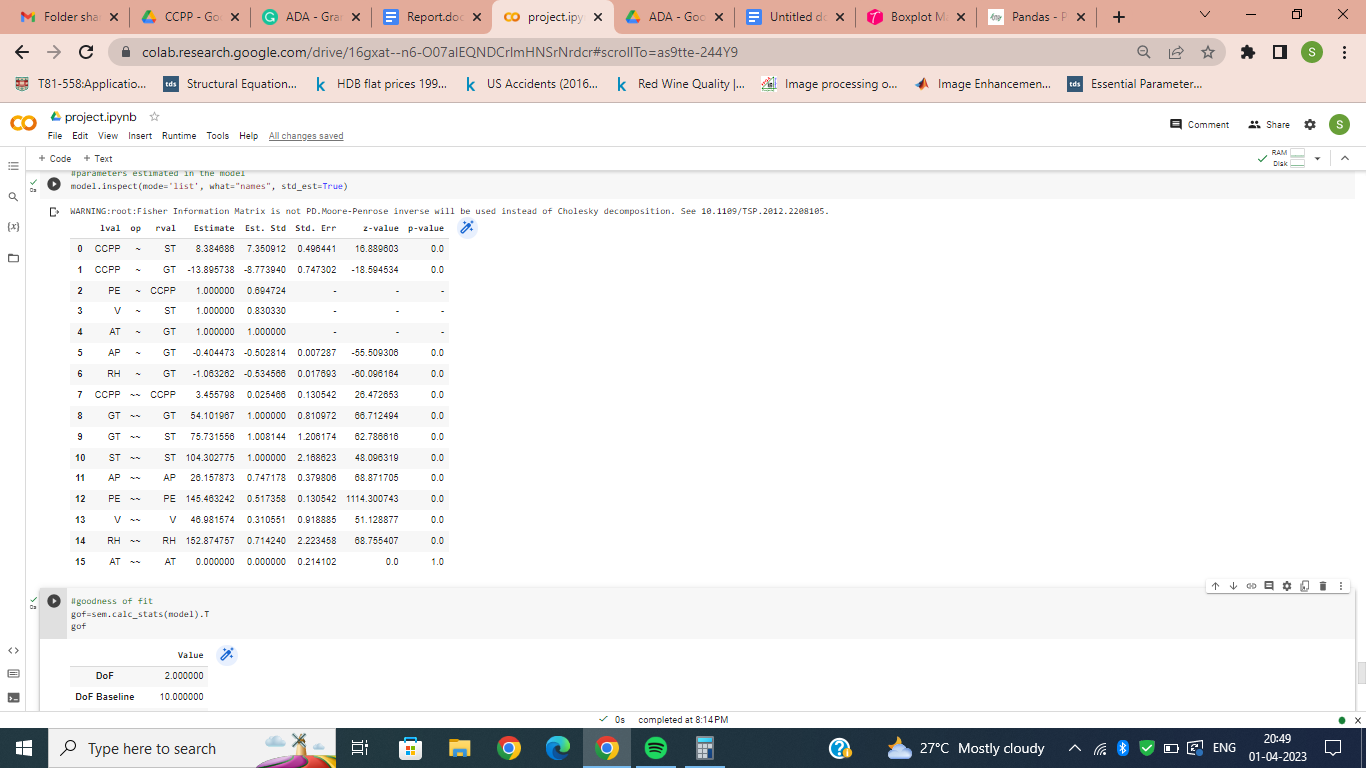


Fig.10.Estimated Parameters

## **7.2 Goodness of fit**

The goodness of fit is decided by different tests and measures. Some of them are R2, GFI, , RMR, RMSEA in absolute fit indices

Some of the relative fit indices are NFI, CFI, AGFI and PNFI.All these values are given by a summary table for SEM.The table is shown in Fig.11.All values indicate that the fit is good.

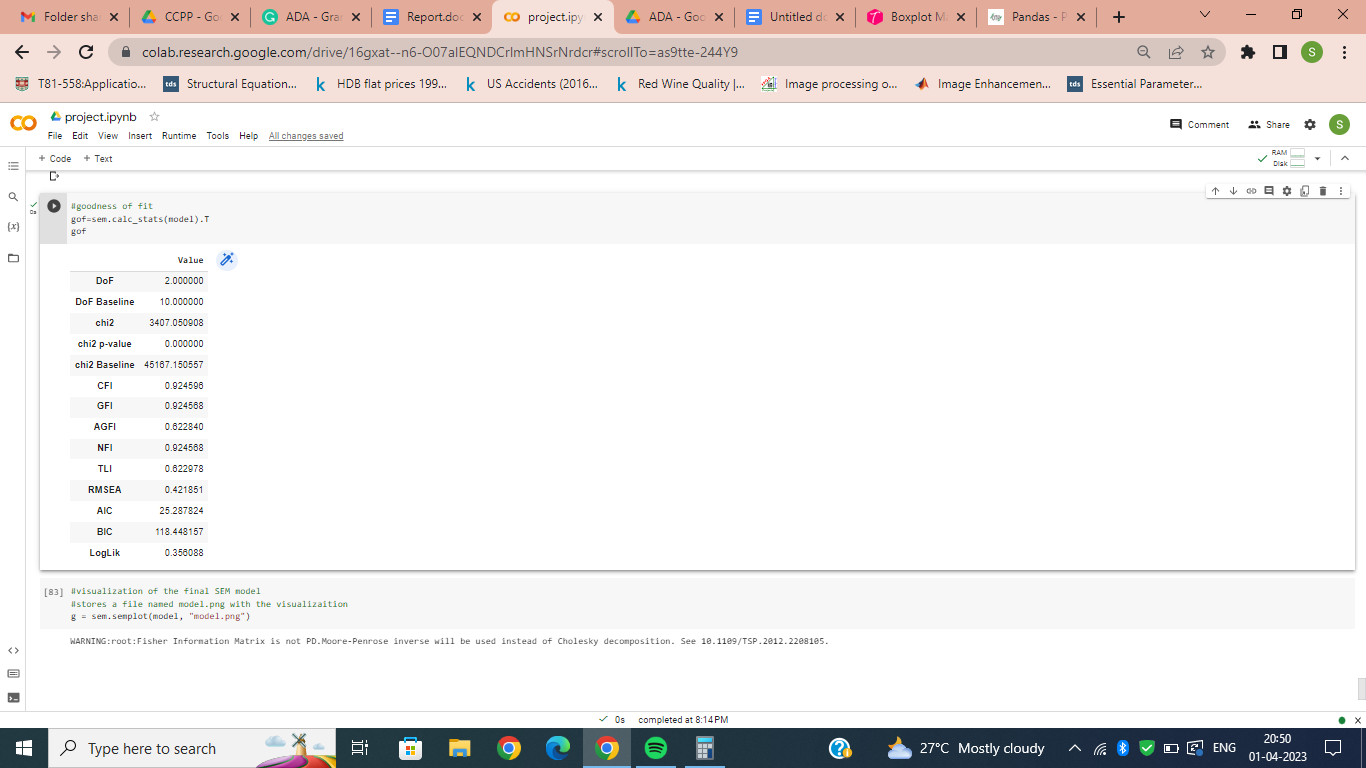


Fig.11.Goodness of fit

## **7.3 Visualization**

The model visualization is as important as the goodness of it. The model will be inferior if we are not able to visualize the data, even if the fit is good.

The visualization of the SEM model for this dataset is shown in Fig.12.

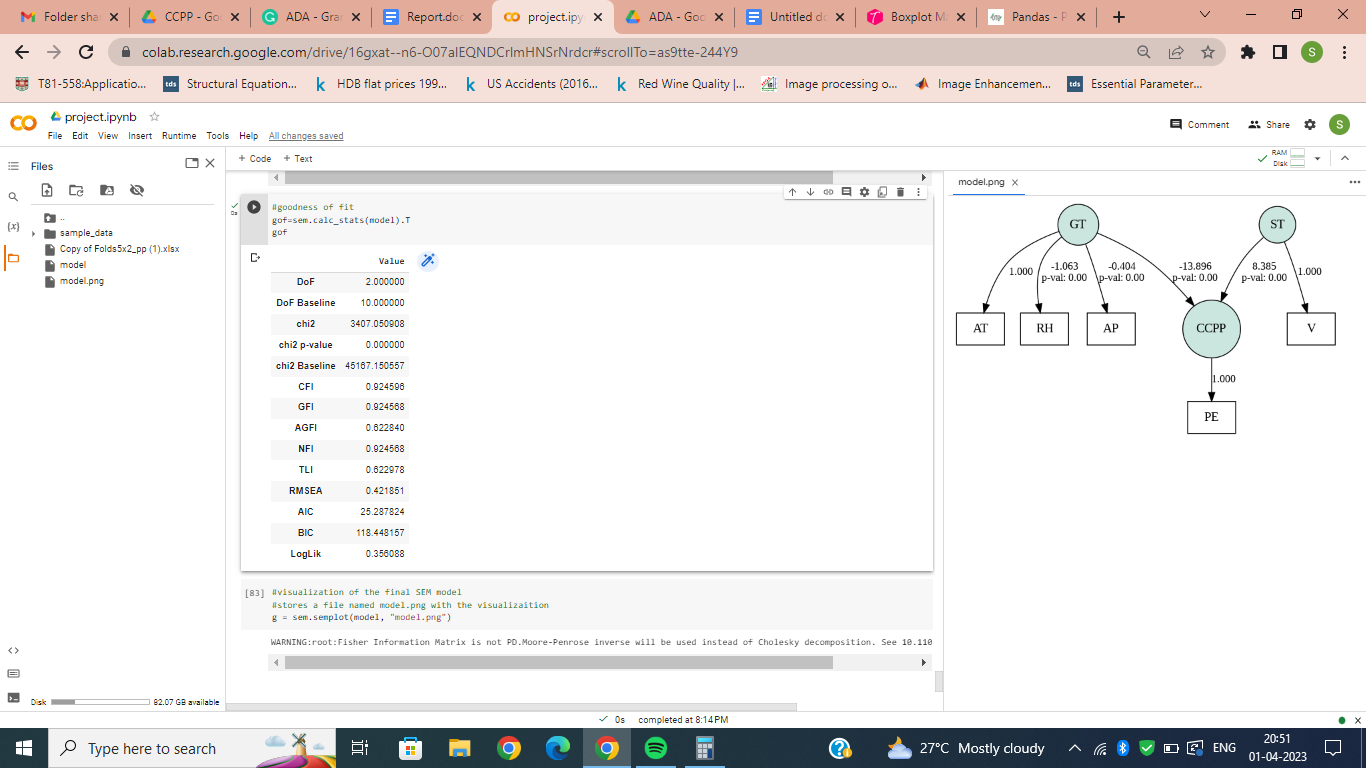


Fig.12.Visualization of SEM

**8.Conclusion**

The dataset is subjected to different models throughout the analysis. Starting with Linear regression, followed by factor analysis and finally Structural Equation modeling.

After the factor analysis, the data size is reduced to half which reduces the storage space while the accuracy drop is only by 5% which is acceptable.

SEM provides us an overall analysis rather than a window of the analysis like linear regression or Factor analysis does. In a way SEM includes the analysis done by both linear regression and factor analysis.