**CCT College Dublin**

|  |  |
| --- | --- |
| **Module Title:** | *Statistics for Data Analytics*  *Data Preparation & Visualisation*  *Programming for Data Analytics*  *Machine Learning for Data Analysis* |
| **Assessment Title:** | *MSc\_Data\_Analytics\_CA2* |
| **Lecturer Name:** | *John O’Sullivan*  *David McQuaid*  *Sam Weiss*  *Muhammad Iqbal* |
| **Student Full Name:** | *Mahmut Zahid Malkoc* |
| **Student Number:** | *2023034* |
| **Assessment Due Date:** | *26/05/2023* |
| **Date of Submission:** | *26/05/2023* |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**ABSTRACT:**

This study aims to examine the impact of employment number of persons and building permits on Construction Producer Prices (CPP). The analysis is based on the examples of the United Kingdom (UK), Turkey (TR), Ireland (IE), and the European Union (EU) countries. The objective of the study is to evaluate the relationship between Construction Producer Prices(CPP), Building Permits(BP), and Employment Number of Persons(ENOP) factors and analyze this relationship on a yearly basis.

In this study, macroeconomic indicators such as employment number of persons, as well as building permits which are crucial for the vitality of the construction sector, are taken into consideration. The data is obtained from the Eurostat and covers the period from 1996 to 2022.

The analysis results reveal the impact of employment and building permit numbers on CPP. An increase or decrease in employment and building permits are found to have a positive or negative effect on the demand in the construction sector and consequently influences CPP in a manner. Similarly, an increase in building permits encourages activities in the sector and contributes to the rise in CPP. However, downturns in employment and building permits due to economic fluctuations, policy changes, and other factors can adversely affect CPP.

In conclusion, this study emphasizes the dependence of CPP on employment and building permits and demonstrates how these factors affect prices in the construction sector. These findings can be taken into consideration by governments when formulating supportive policies for the construction sector and regulating the market. Additionally, this analysis can provide valuable insights to construction companies in understanding industry trends and determining their future strategies.

**INTRODUCTION: Construction Sector Regarding Ireland**

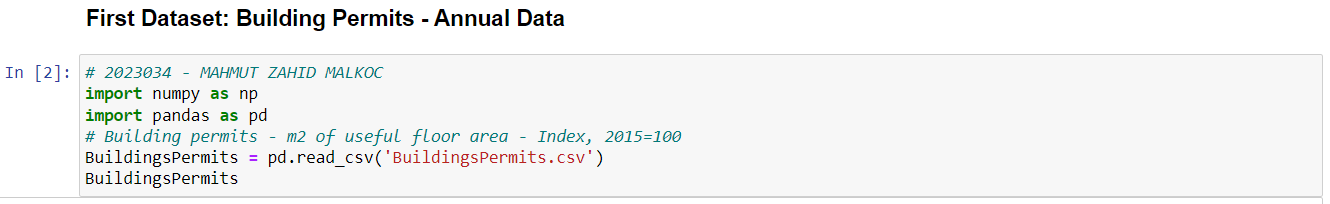
Analysis of the Increase or Decrease in the Construction Producer Prices According to the Number of Employment and Building Permits on a Yearly Basis considering Ireland (IE), United Kingdom (UK), European Union (EU), and Türkiye (TR).

Overall, this assessment was carried out on the construction sector in Ireland, the United Kingdom, European Union, and Türkiye provided that three different construction data were selected from Eurostat, the European Union open data portal.

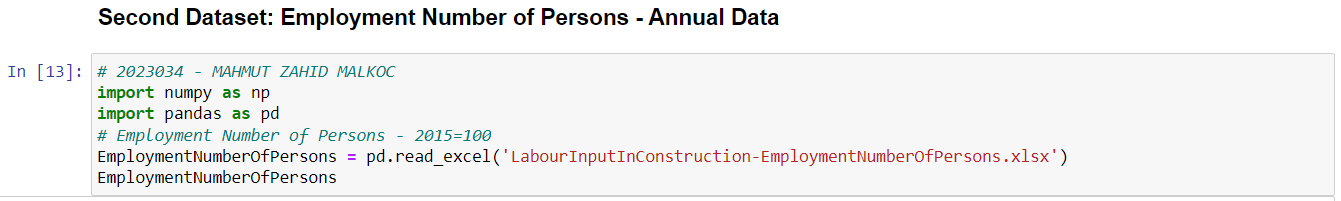
The Building Permits dataset is based on 2015=100 from 1992 to 2022. The Employment Number of Persons dataset is based on 2015=100 from 1990 to 2020. The Construction Produce Prices dataset is based on 2015=100 from 1990 to 2022. In these three different datasets, there are data for Ireland, the United Kingdom, the European Union, and Türkiye.

**PROGRAMMING FOR DATA ANALYTICS:**

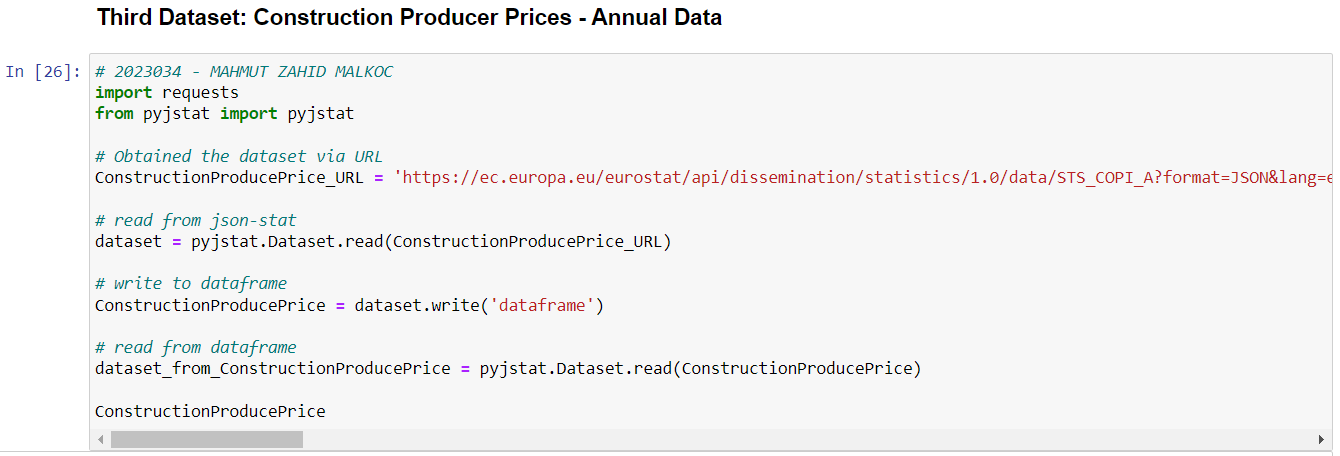
* **Programming:** Code and libraries suitable for the project are added to the top line of each cell where necessary. It can be seen in the attached file and in the GitHub repository in Jupyter Notebook (.ipynb).
* **Data Structures:** In this study, three different data distinct formats were used. These are .CSV file, .xlsx file and the web API in JSON format respectively. These formats used are shown in the Figure 1, 2, 3 below.



*Figure 1 - .csv file format*



*Figure 2 - .xlsx file format*

*Figure 3 – Web API in .JSON format*

* **Documentation:** The project documentation explains the rationale and explanation of the code choices where appropriate.
* **Testing&Optimisation:** In this study, code and programming are documented in appropriate sections where necessary.
* **Data Manipulation:** In this project, pandas and numpy libraries were actively used while extracting data from the data source in .csv and .xlsx formats. While extracting data in .xlsx format, the openpyxl library could be used, but with the pandas library, this process was handled with a much shorter process. Another data source is 'request' and 'pyjstat' libraries in JSON format via Web API. In the last stage, these three different datasets could be made with the 'concat' or 'merge' method in the pandas library. This was carried out simply with 'merge'.

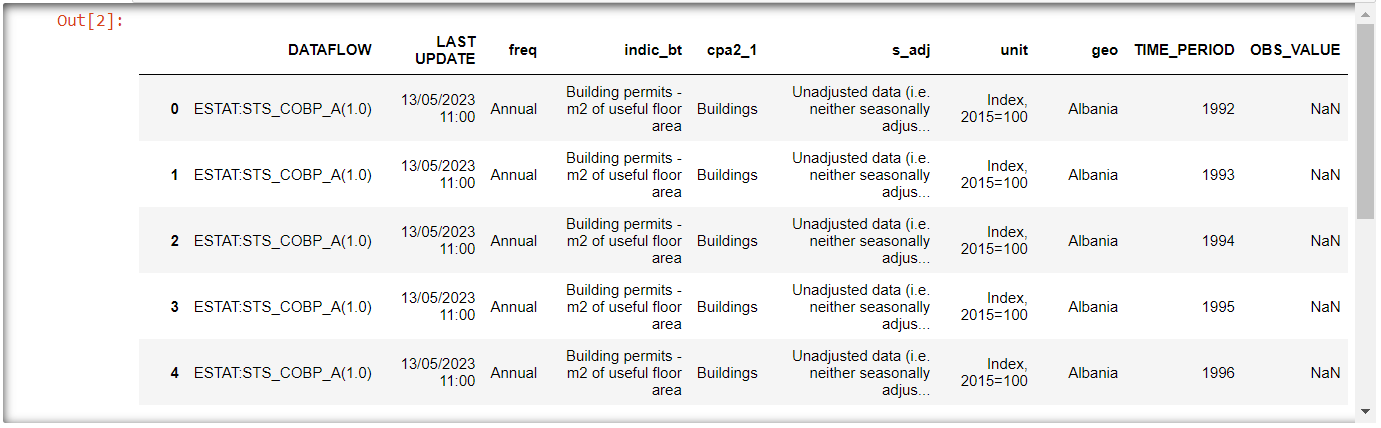
**DATA PREPARATION AND PROGRAMMING:**

* As a result of the research, the data sets were obtained from Eurostat, the open data portal of the European Union. This data is gathered in three different formats as '.csv', '.xlsx' and 'JSON'. Since all of these datasets are received by the open data portal, they are licensed in accordance with open source licensing.

The open source license for all datasets used is available at the link below:

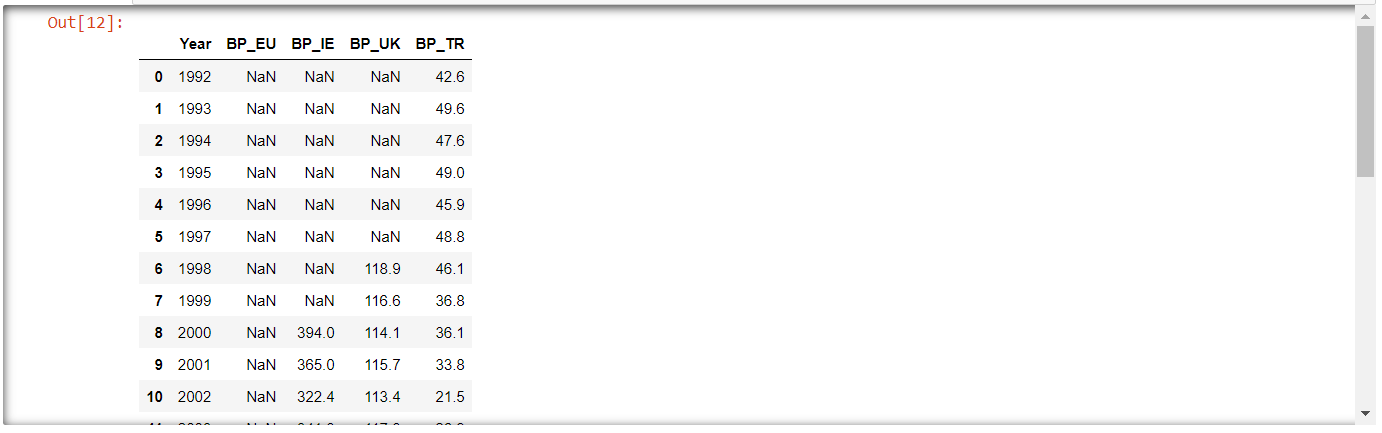
https://opendatacommons.org/licenses/odbl/odbl-10.txt

* According to the Exploratory Data Analysis stages, the datasets and the performed stages are as follows step by step.



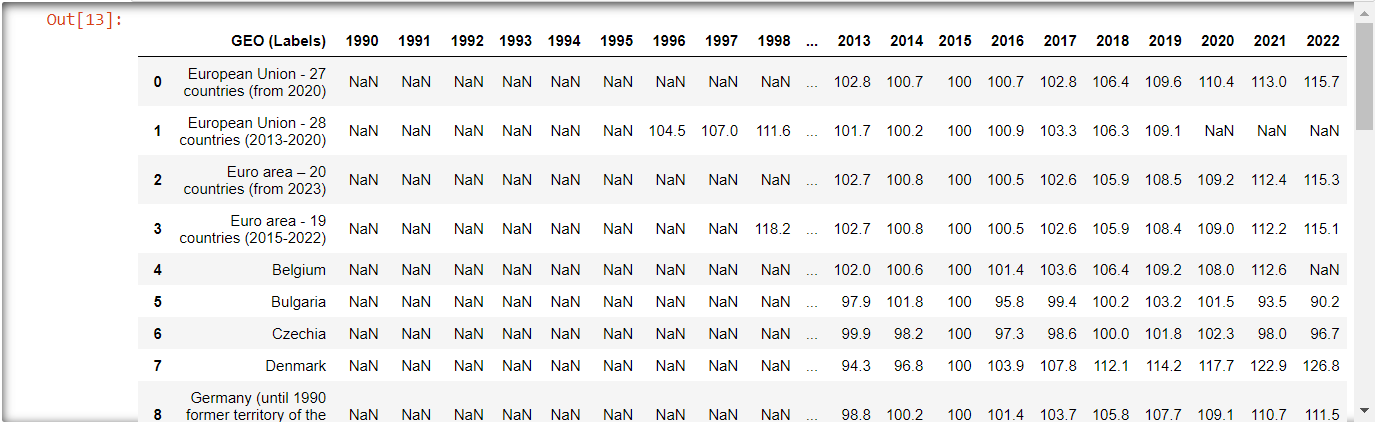
*Figure 4 – Building Permits Dataset*

As seen in Figure 4, the building permits dataset is in its purest form. Many operations have been applied to transform the dataset into figure 5. The new dataset in Figure 5 was created by renaming the columns in the final stage by gathering the required countries and values according to the years.



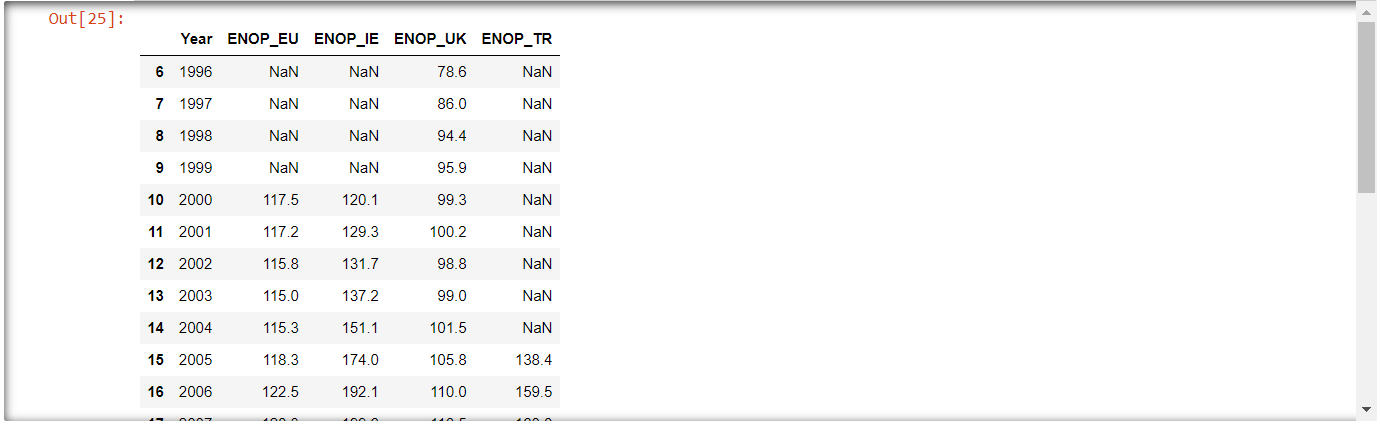
*Figure 5 – Manipulated Building Permits Dataset*

As seen in Figure 5, there are 5 different attributes. These are Year, BP\_EU, BP\_IE, BP\_UK, BP\_TR. The abbreviation 'BP' in each column stands for 'Building Permits'.



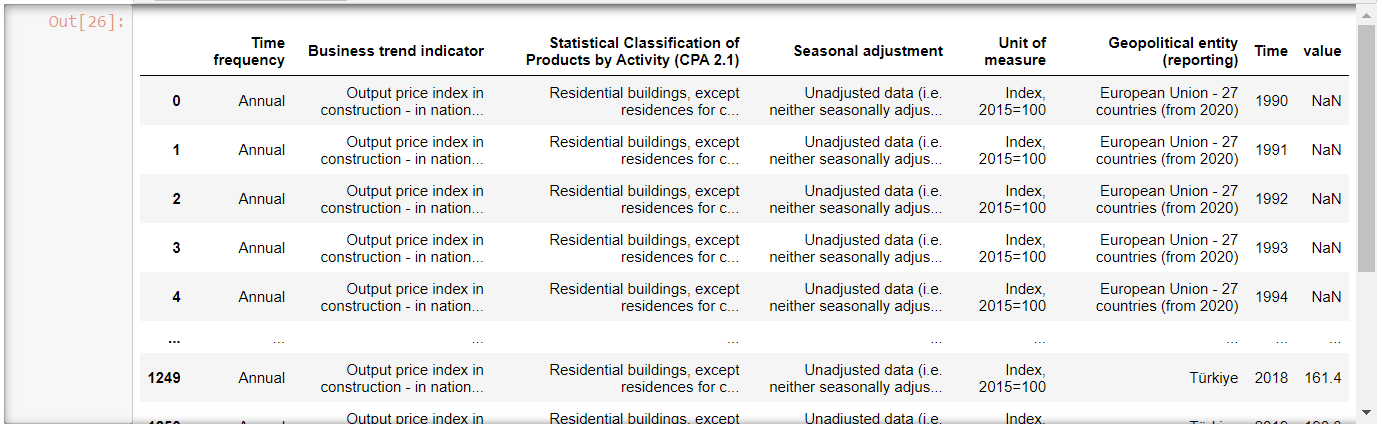
*Figure 6 – Employment Number of Persons Dataset*

From Figure 4 to Figure 5 similar operations were performed during the transition from Figure 6 to Figure 7 while manipulating.



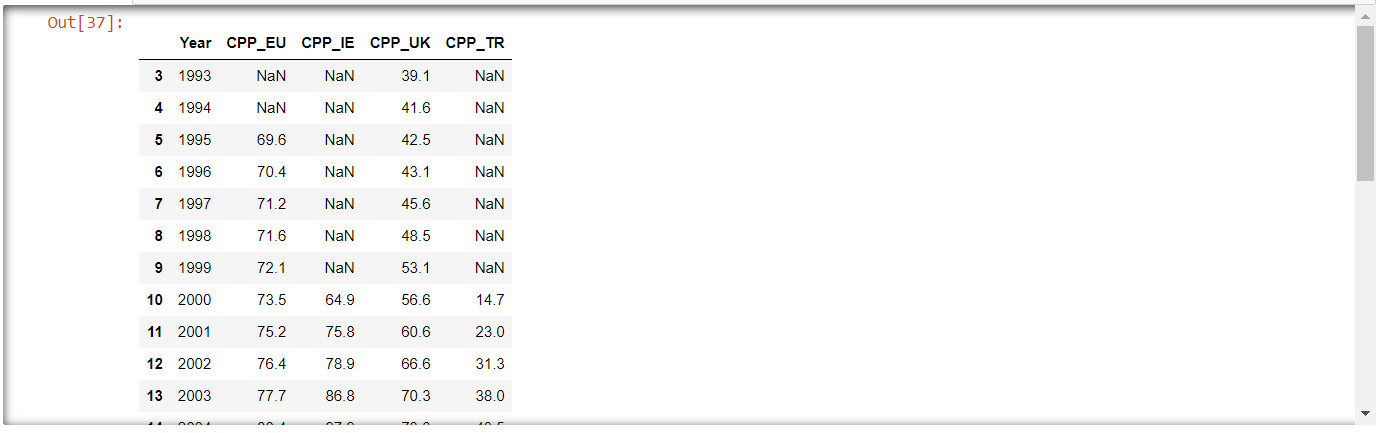
*Figure 7 – Manupulated Employment Number of Persons Dataset*

As seen in Figure 7, there are 5 different attributes. These are Year, ENOP\_EU, ENOP\_IE, ENOP\_UK, ENOP\_TR. The abbreviation 'ENOP' in each column stands for 'Employment Number of Persons'.



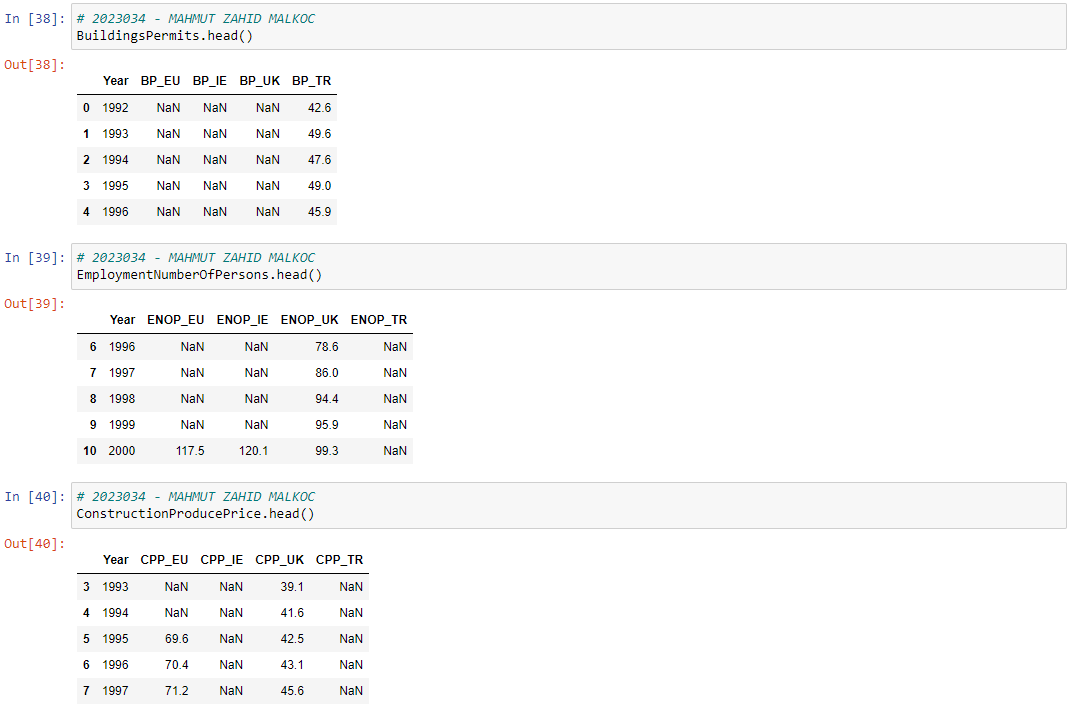
*Figure 8 – Construction Produce Prices*

Similar procedures applied in Figure 4-5 and Figure 6-7 were also applied in the manipulation from Figure 8 to Figure 9.



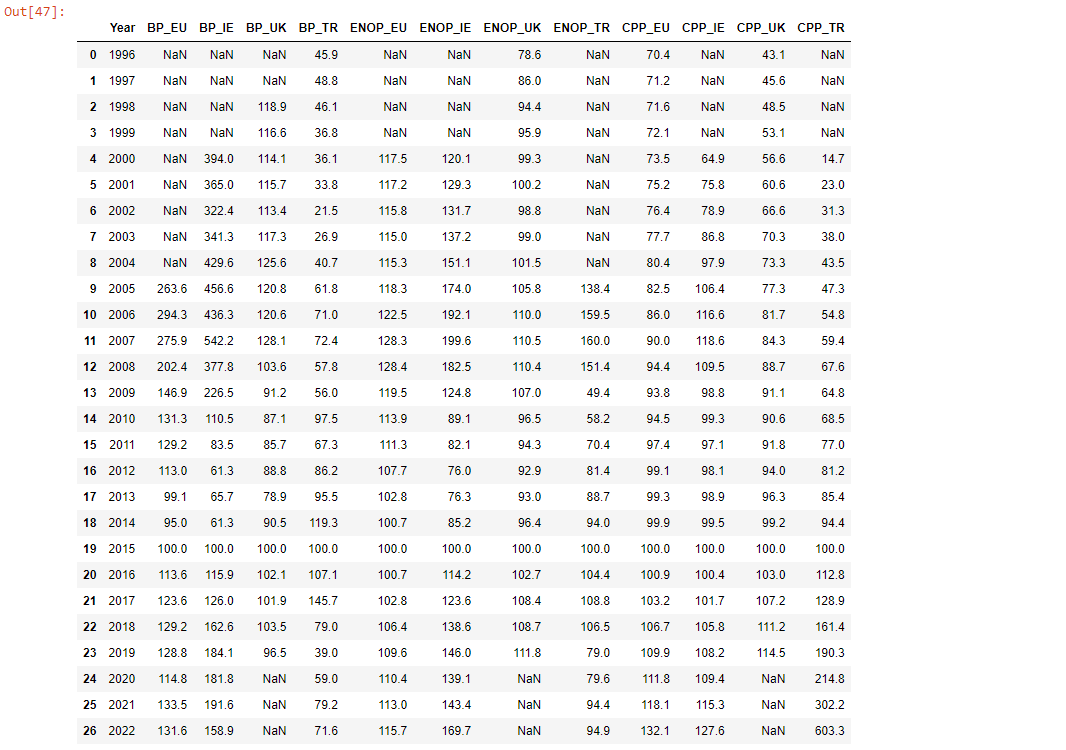
*Figure 9 – Manupulated Construction Produce Prices*

As seen in Figure 9, there are 5 different attributes. These are Year, CPP\_EU, CPP\_IE, CPP\_UK, CPP\_TR. The abbreviation 'CPP' in each column stands for 'Construction Produce Prices'.



*Figure 10 – Overview of Datasets*

As can be seen in Figure 10, an overview of each dataset is performed.

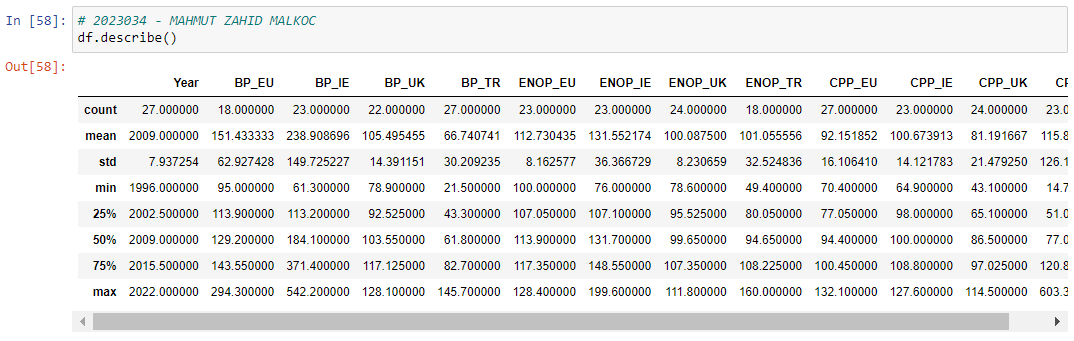


*Figure 11 – The Final Version of The Created Dataframe*

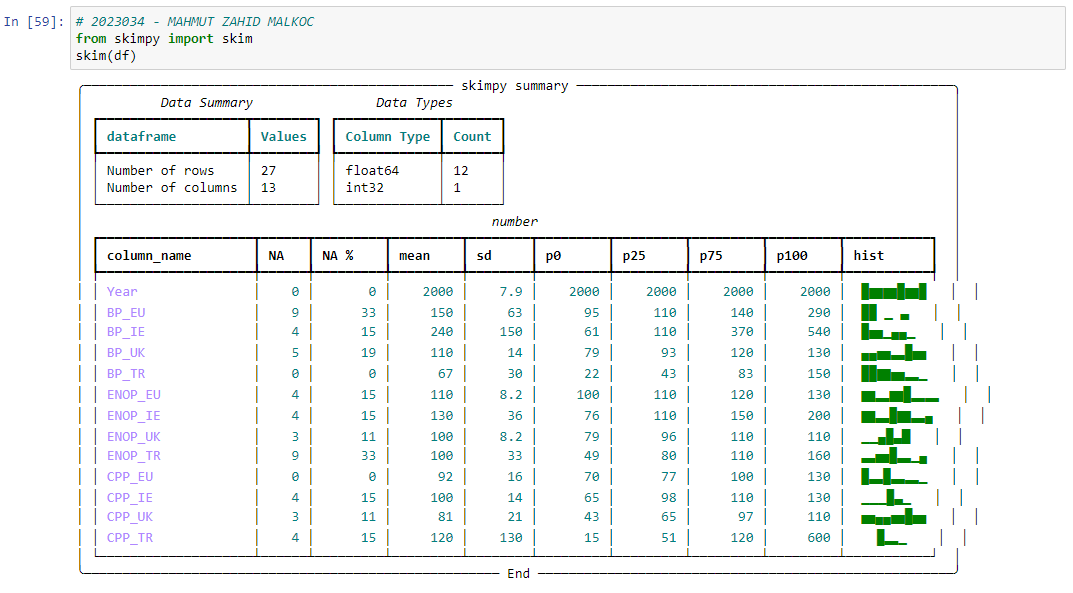
As can be seen in figure 11, 3 different datasets have been merged with the year column remaining constant. The dataframe has been created as seen above. However, the dataframe is not yet ready for statistics and machine learning models.

**STATISTICS FOR DATA ANALYTICS:**

* Descriptive statistics are central predictive measures and dispersion measures with the 'description' and 'skimpy' methods as follows.

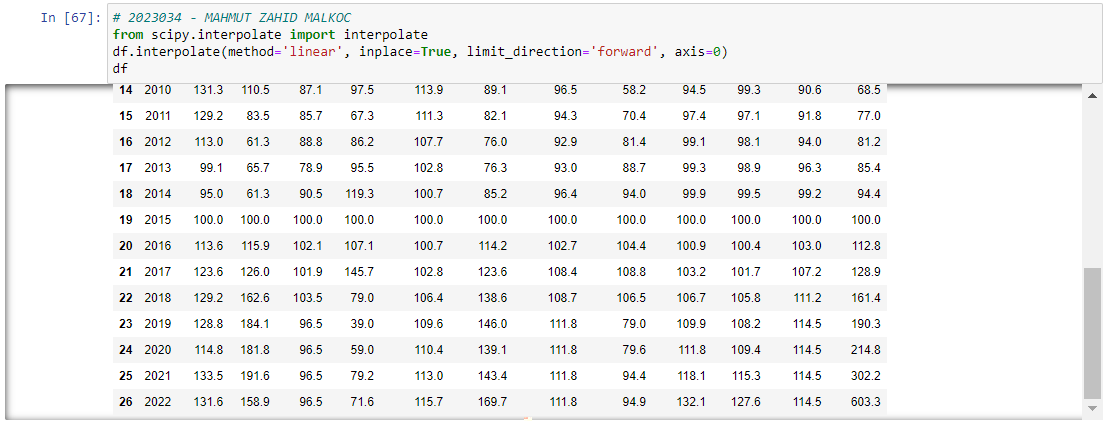


*Figure 12 – Overview of Datasets*



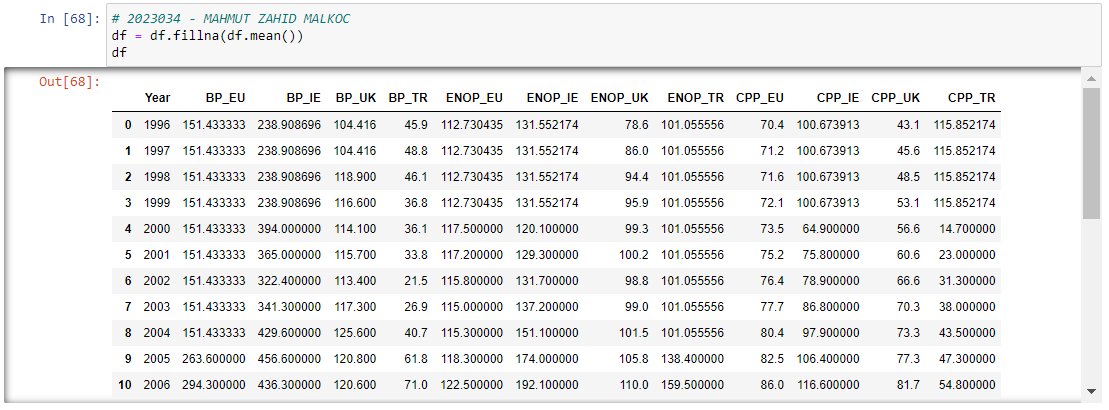
*Figure 13 – Overview of Datasets*

As seen in Figure 12 there is a statistical overview of the dataframe. However, as can be seen in Figure 13, there are many NaN values in the dataframe. Statistically, looking at the dataframe will not be efficient unless the NaN values are properly filled.



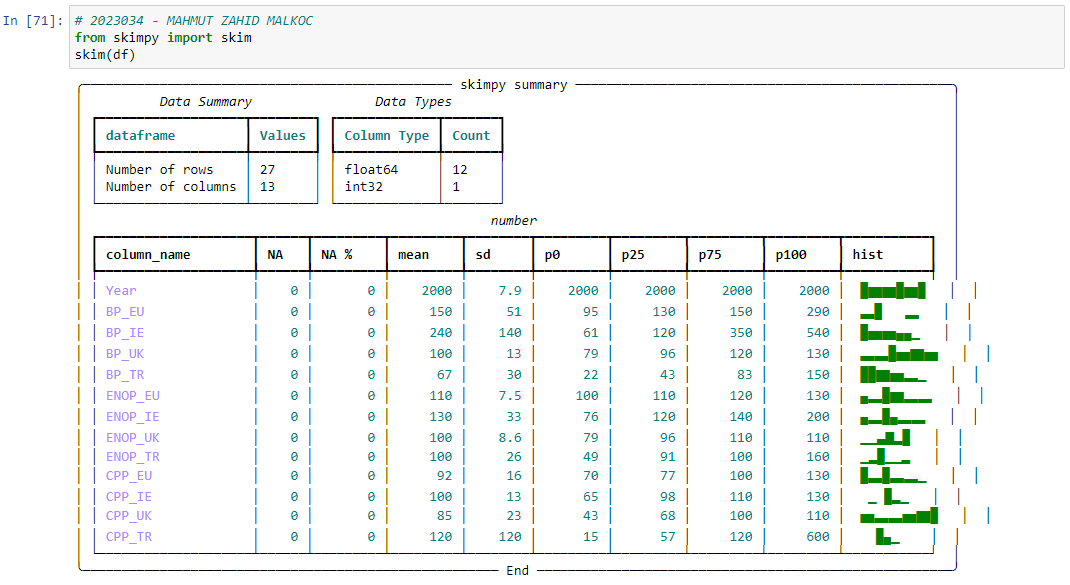
*Figure 14 – Linear Interpolate Method for Missing Values*

As seen in Figure 14, linear interpolate method is applied to this dataframe for the next values. It has been determined that this method fills the following years with more appropriate values compared to other methods.



*Figure 15 – df.fillna(df.mean()) Method for Missing Values*

As can be seen in figure 15, the values of the previous years are filled with the mean method in the most appropriate way.



*Figure 16 – Overview of Datasets*

As seen in figure 16, there is no NaN value in the dataframe. In addition, the mean, standard deviation, percentiles and a simple histogram plot for each attribute are shown.

* Estimation is made by associating with Construction Produce Price in Building Permits and Employment Number of Persons datasets.
* Five different inferential statistical methods were used. These are t-test, , ANOVA, Wilcoxon Test, Chi-Squared Test, Corelation Test and OLS Regression Test respectively.

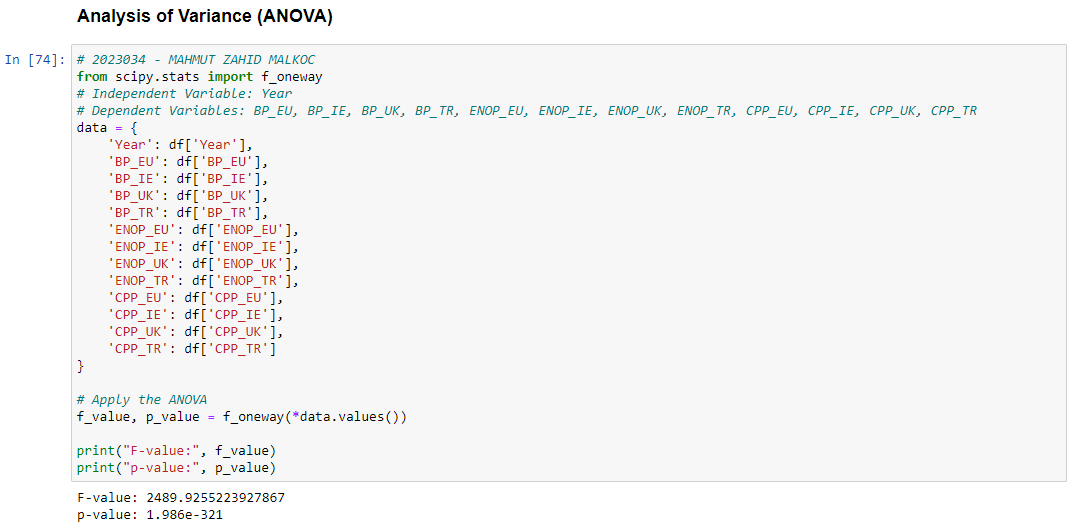


*Figure 17 – T-Test*

The output t statistics represents the t statistics. This value is calculated by taking into account the size of the difference between 'BP\_IE' and 'CPP\_IE' and the distribution of variance in the groups. A high t statistic value indicates that the difference between the sample is statistically significant. In this case, the t statistic value was calculated as 5.192259561613692.

P value represents the p value. The p value is used to determine whether the t-test is statistically significant. The p value is a probability value calculated based on the data. A small p value indicates that the t-test produces meaningful results. In this case, the p value is given as 3.5026525326112636e-06 (or approximately 0.0000035), which is a quite small value.

These outputs show that the difference between 'BP\_IE' and 'CPP\_IE' is statistically significant. Since the t statistic is high and the p value is low, it can be concluded that the difference between the two samples is not accidental and is a real difference.



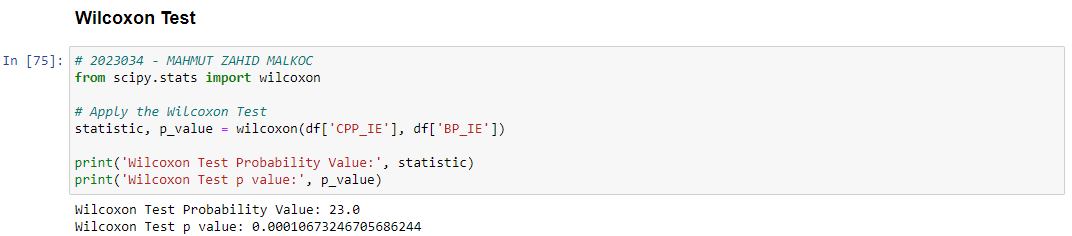
*Figure 18 – Analysis of Variance (ANOVA)*

The ANOVA test is used to determine whether the difference of an independent variable between more than one group is statistically significant.

In this dataset, the variable Year is accepted as the independent variable, while the variables BP\_EU, BP\_IE, BP\_UK, BP\_TR, ENOP\_EU, ENOP\_IE, ENOP\_UK, ENOP\_TR, CPP\_EU, CPP\_IE, CPP\_UK, CPP\_TR are considered dependent variables.

The output F-value represents the F statistic. This value expresses the variance differences between groups. A high value of F statistic indicates that the difference between groups is statistically significant. In this case, the F statistic value is calculated as 2489.9255223927867.

P value represents the p value. The P value is used to determine whether the ANOVA test is statistically significant. A small p-value indicates that the test produces meaningful results. In this case, the p-value was calculated as 1.986e-321 (or approximately 0). In this case, the ANOVA test produces meaningful results.



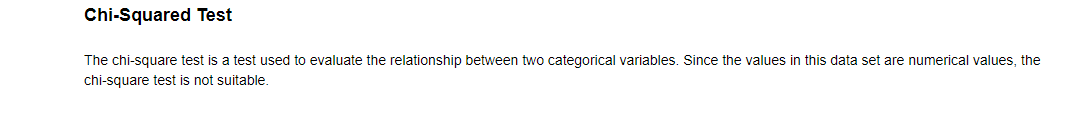
*Figure 19 – Wilcoxon Test*

The output Wilcoxon Test Probability Value represents the Wilcoxon test statistic. This value may vary depending on the calculation method of the Wilcoxon test. In this case, the test statistic value was calculated as 23.0.

Wilcoxon Test p value expresses the p value. The P value is used to determine whether the Wilcoxon test is statistically significant. A small p-value indicates that the test produces meaningful results. In this case, the p-value is 0.00010673246705686244, which means it's a pretty small value.

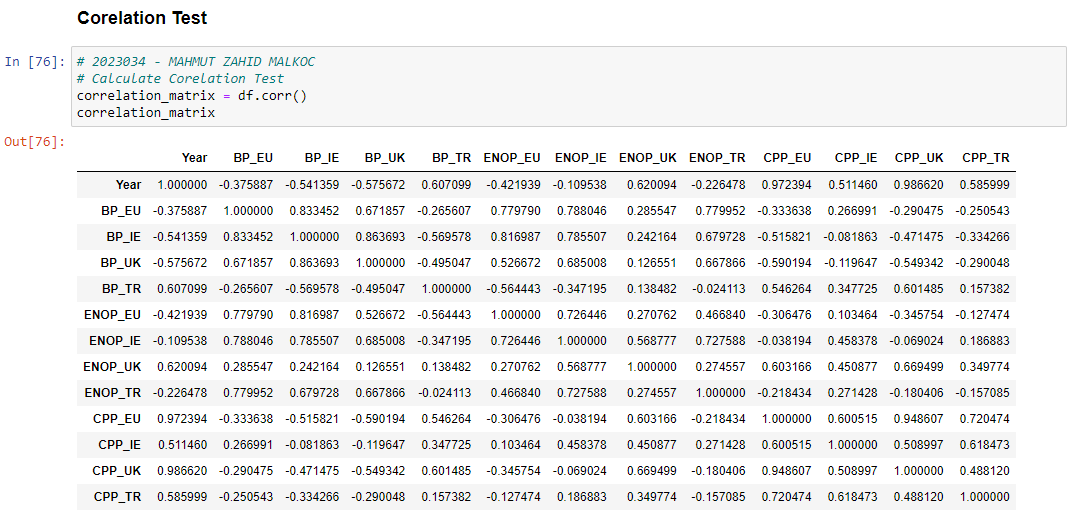
These outputs show that the difference between df['CPP\_IE'] and df['BP\_IE'] is statistically significant. Since the Wilcoxon test is a non-parametric test that is not based on the distribution of data, it provides reliable results even if the data do not show normal distribution. Unlike the t-test, the Wilcoxon test approaches data on a rank basis and compares on median values.

Based on the results, you can conclude with the Wilcoxon test that the difference between 'CPP\_IE' and 'BP\_IE' is statistically significant.



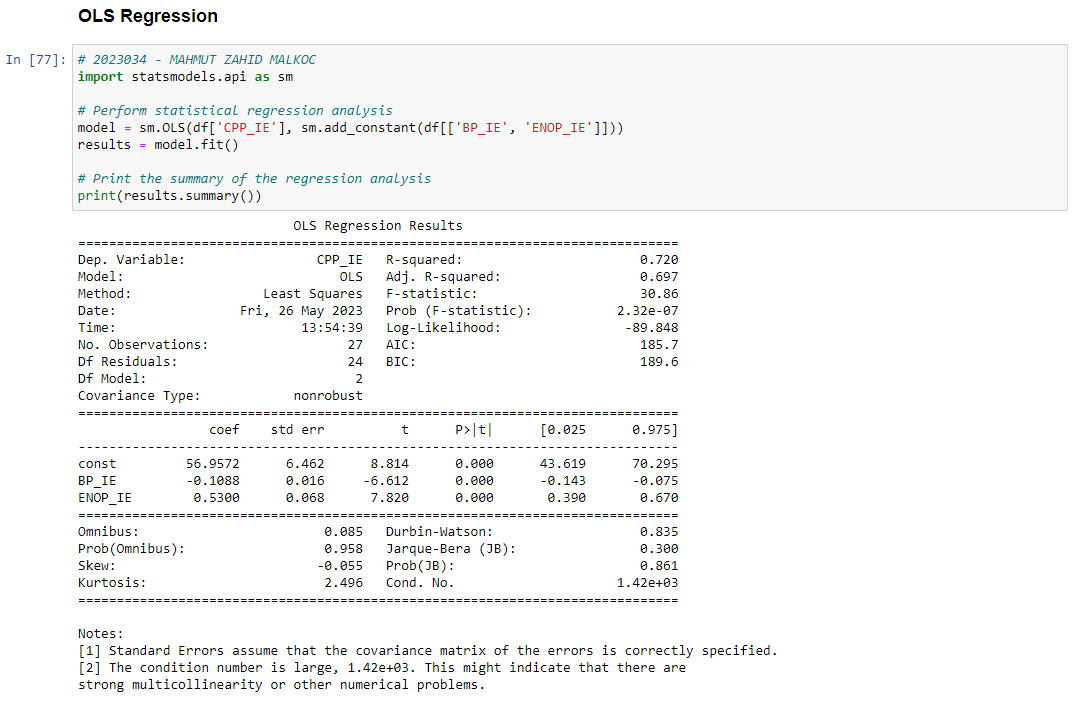
*Figure 20 – Chi-Squared Test*

The chi-square test is a test used to evaluate the relationship between two categorical variables. Since the values in this dataset are numerical values, the chi-square test is not suitable.



*Figure 21 – Corelation Test*

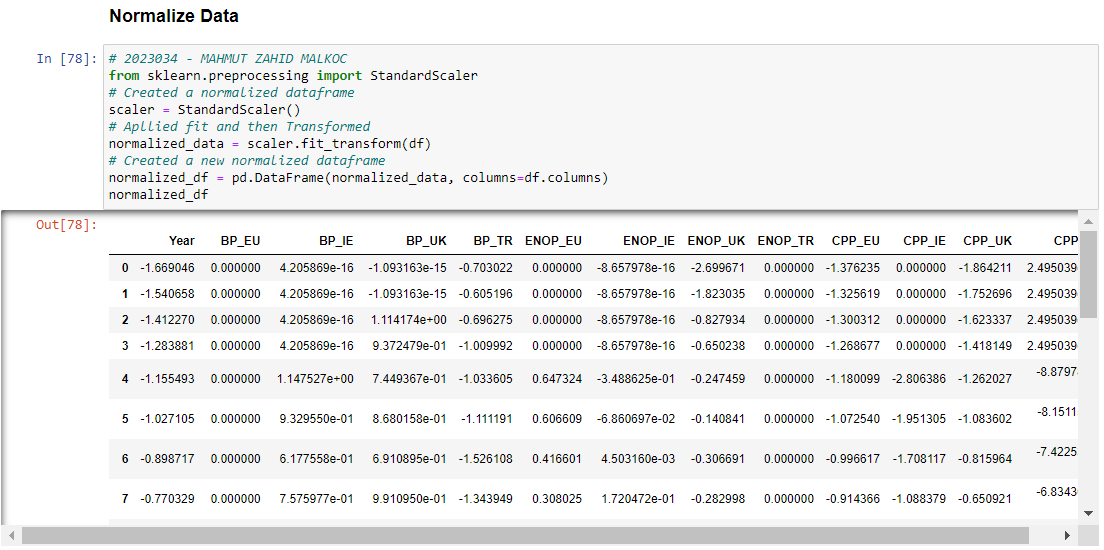
The correlation coefficient measures the strength and direction of the relationship between two variables. Values can range from -1 to 1, -1 indicates a negative relationship, 1 indicates a positive relationship, and 0 indicates no relationship.



*Figure 22 – OLS Regression*

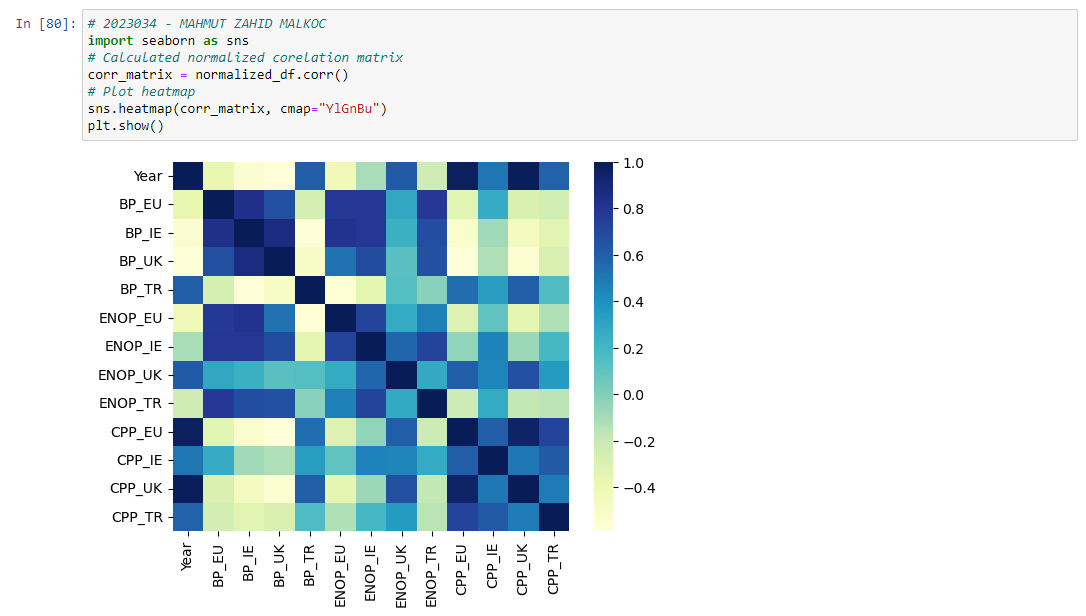
According to the ordinary least squares (OLS) regression, the variable "CPP\_IE" is accepted as the dependent variable, while the variables "BP\_IE" and "ENOP\_IE" are considered as independent variables. This method assumes that the dependent variable has a linear relationship with the independent variables.

The resulting table includes coefficients, standard errors, t statistics, p values, and other statistical information. There are also statistics that measure the fit of the model, such as the R-squared value and the adjusted R-squared value. The OLS Regression method was used to determine how the "CPP\_IE" variable was associated with the "BP\_IE" and "ENOP\_IE" variables and how well the model fit statistically.



*Figure 23 – Normalized DataFrame*

According to figüre 23, the correlation matrix is a matrix that shows the relationships between the variables in the data frame. The values in each cell represent the correlation coefficient between the two variables. Values usually range from -1 to 1, with -1 indicating a negative relationship and 1 indicating a positive relationship.



*Figure 24 – Correlation Matrix Heat Map*

As can be seen the figure 24, the heatmap shows how the variables relate to each other. Intense colors represent a stronger correlation, while pale colors represent a weaker correlation. In this way, it is used to obtain information about the relationships between variables and to identify potential relationships.

* During the dataset research process, there were difficulties in selecting the dataset in order to compare the correct attributes. Therefore, when these dataset selections are completed, the resulting dataframe is a numerical dataset. Some inferental statistical models cannot be applied because there is no categorical attribute in the data set. For this reason, the Chi-Squared Test could not be applied and it was explained as such.

**DATA VISUALIZATION AND PROGRAMMING:**

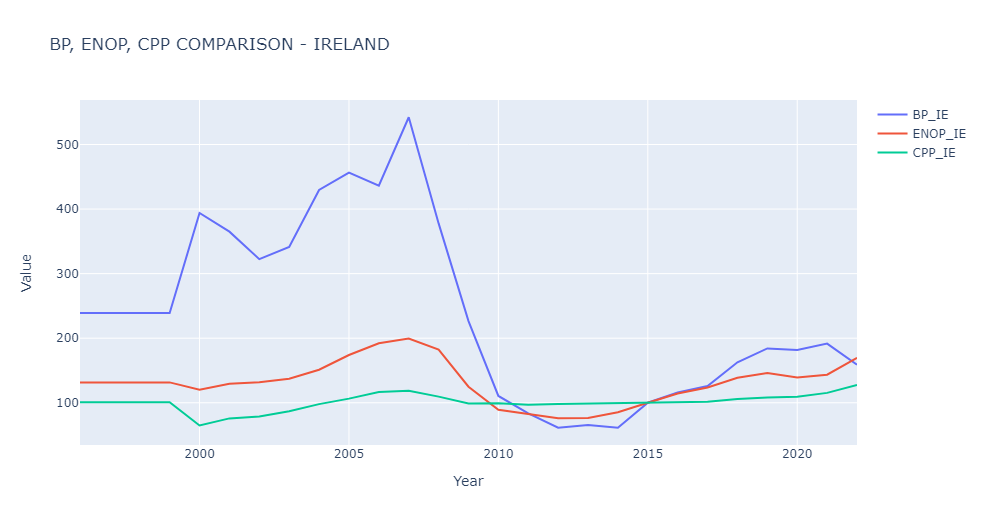


*Figure 25 – Pairplot*

According to the pairplot in figure 25, the relationship between each attribute is seen.

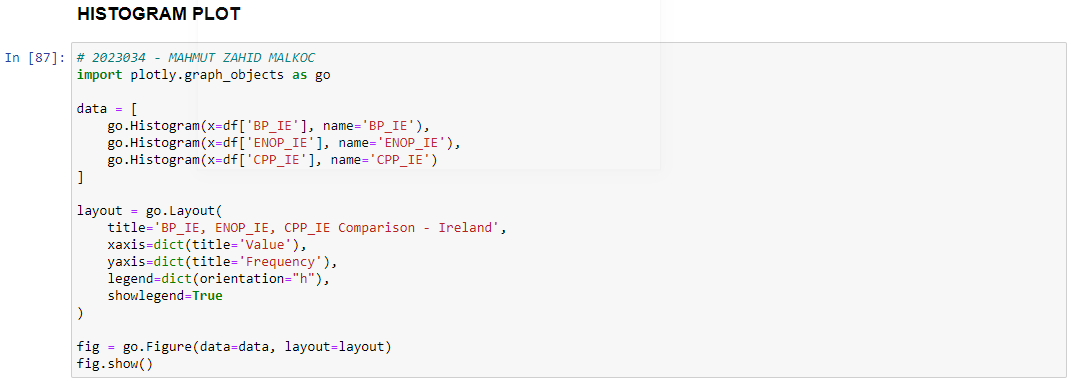


*Figure 26 – Line Plot via Interactive Dashboard*

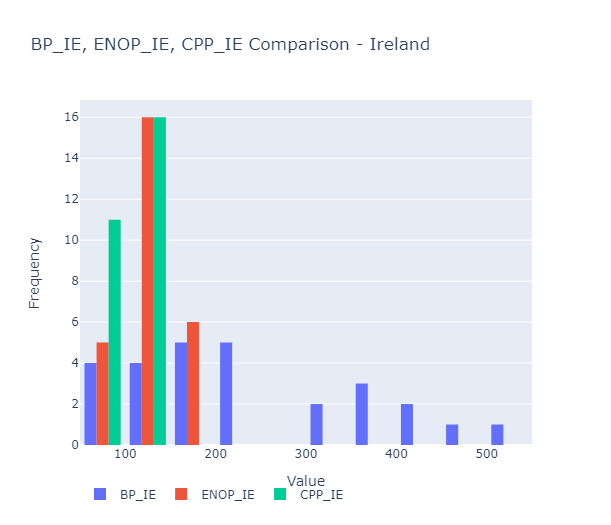


*Figure 27 – Line Plot*

The relationship between BP\_IE, ENOP\_IE, CPP\_IE according to the codes and arrangements in Figure 26 was compared as a line plot and presented in Figure 27.

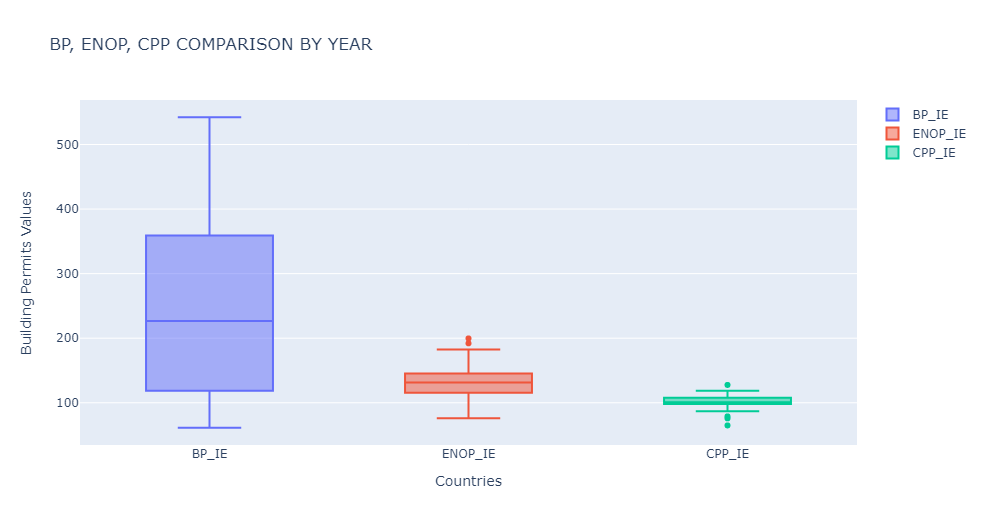


*Figure 28 – Histogram Plot via Interactive Dashboard*

 *Figure 29 – Histogram Plot*

The relationship between BP\_IE, ENOP\_IE, CPP\_IE according to the codes and arrangements in Figure 28 was compared as a histogram plot and presented in Figure 29.

*Figure 30 – Box Plot via Interactive Dashboard*



*Figure 31 – Box Plot*

The relationship between BP\_IE, ENOP\_IE, CPP\_IE according to the codes and arrangements in Figure 29 was compared as a box plot and presented in Figure 30.

Three different plot types are given above. In addition to this, scatter plot, bar chart and filled area plot types are also included in this assignment.

**MACHINE LEARNING FOR DATA ANALYTICS:**



*Figure 32 – Decision Tree*

The dataset is prepared and divided into X and y. X is a dataframe containing properties without columns 'BP\_IE', 'ENOP\_IE', 'CPP\_IE'. y is a dataframe containing target variables with columns 'BP\_IE', 'ENOP\_IE', 'CPP\_IE'.

The dataset is divided into train and test. The train\_test\_split function is used in this step. 60% of the dataset is used as the train dataset, while 40% is used as the test dataset.

The Decision Tree Regression model is defined and trained on the training dataset. A model is created using the DecisionTreeRegressor class and model training is performed on the training dataset (X\_train and y\_train) with the fit method.

The performance of the model is evaluated. R^2 score (r2\_score) and mean squared error (Mean Squared Error, mean\_squared\_error) are calculated using the predictions made on the test dataset. The R^2 score indicates how well the model fits the data, while the mean squared error represents the mean squared difference between the predictions and the actual values.

Estimates are made. A new dataset named data2025 is defined and predictions are made by the model using the predict method on this data set. The prediction results are thrown into the predictions variable.

The results are as follows:

R^2 Score: 0.4604317413079729

This value indicates that the model can explain approximately 46% of the variation of target variables in the test dataset. The model fits the data better, but there is still potential for improvement.

Mean Squared Error: 3350.5409090909084

This value indicates that the model's predictions have an average square difference of 3350 units from the true values. This indicates how far the model's predictions are from the true values.

Estimates:

Estimates made by the model show that the target variables 'BP\_IE', 'ENOP\_IE' and 'CPP\_IE' are 158.9, 169.7 and 127.6, respectively.



*Figure 33 – Linear Regression Model*

The dataset is divided into train and test. The train\_test\_split function is used in this step. 80% of the dataset is used as the train dataset, while 20% is used as the test dataset.

The Linear Regression model is defined and trained on the training dataset. A model is created using the LinearRegression class and model training is performed on the training dataset (X\_train and y\_train) with the fit method.

R^2 Score: 0.020286637345399077

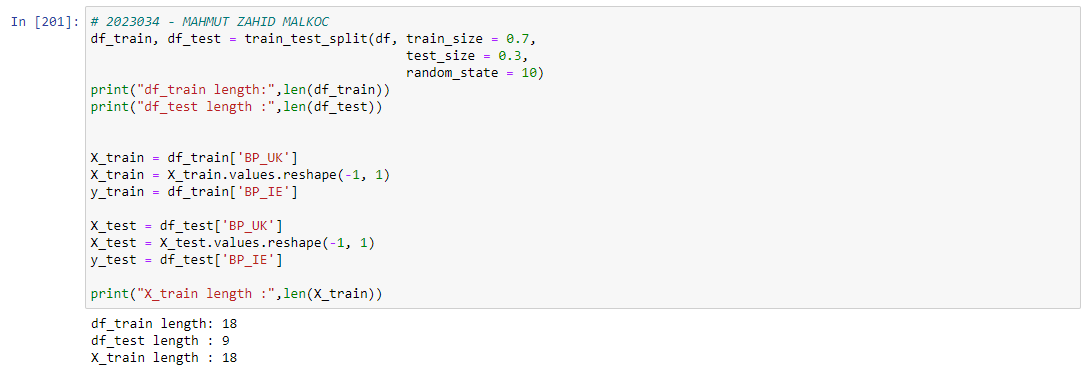
This value indicates that the model can explain approximately 2% of the variation of target variables in the test dataset. The model provides a very poor fit to the data.

Mean Squared Error: 1736.5813336144047

This value indicates that the model's predictions have an average square difference of 1736 units from the true values. This means that the predictions of the model are quite far from the true values.

Estimates:

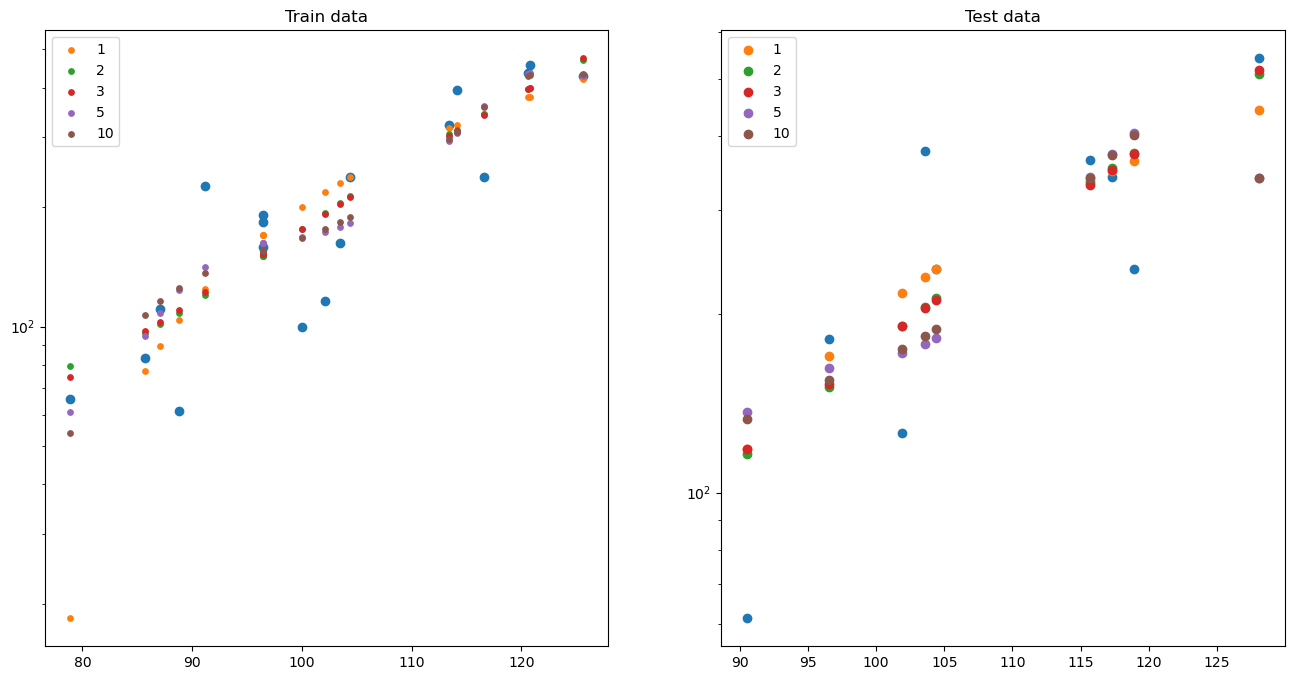
Estimates made by the model show that the target variables 'BP\_IE', 'ENOP\_IE' and 'CPP\_IE' are 358.43, 172.49 and 91.02, respectively. However, the accuracy and meaning of these predictions may vary depending on the data set and the performance of the model.



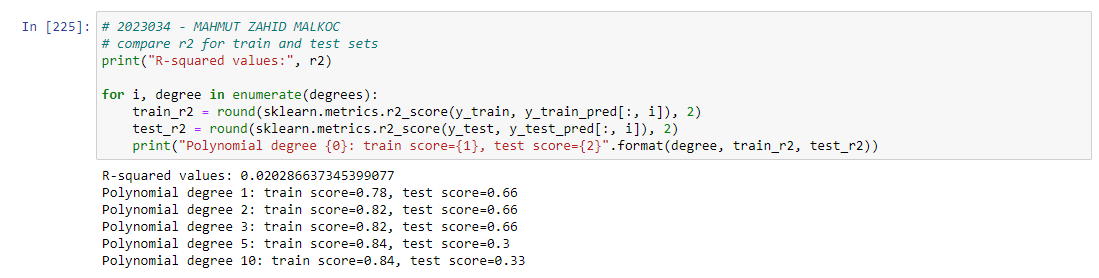
*Figure 34 – Split Train&Test Size and Choose Train&Test Attribute*



*Figure 35 – Prediction Train&Test Data according to Polynomial Degree*

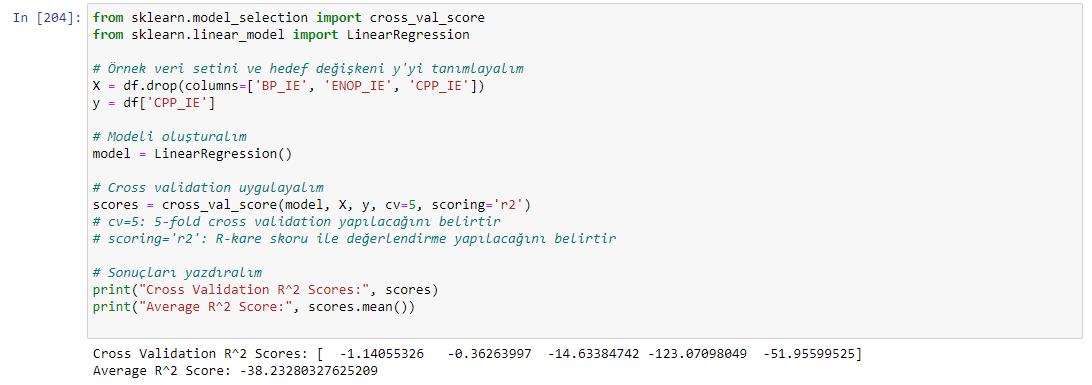


*Figure 36 – Plot Train&Test Data*



*Figure 37 – Train&Test Score according to Polynomial Degree*

These results show that the performance of the model in the train dataset increases with the increase of the polynomial order. However, it is seen that the performance in the test dataset remained stable or even decreased to some extent. This means that the model overfits the train dataset and its generalization ability is reduced. That is, as the degree of polynomial increases, the model becomes more complex and has difficulty accurately predicting the patterns in the test dataset while the train tries to capture the details in the dataset. This is called overfitting and usually occurs as the polynomial degree increases.



*Figure 38 – Cross Validation*

Cross Validation R^2 Scores:

[-1.14055326, -0.36263997, -14.63384742, -123.07098049, -51.95599525]

They are the R-squared (R^2) scores calculated at each cross-validation iteration. These scores show the performance of the model on different subsets of data. Negative values indicate that the model's ability to explain the data is too low.

Average R^2 Score: -38.223280327625209

It is the mean of the R-square scores of the cross-validation iterations. This value indicates the overall performance of the model. A negative value indicates that the model has a very weak relationship between the features used to predict the data and fails to explain the data.

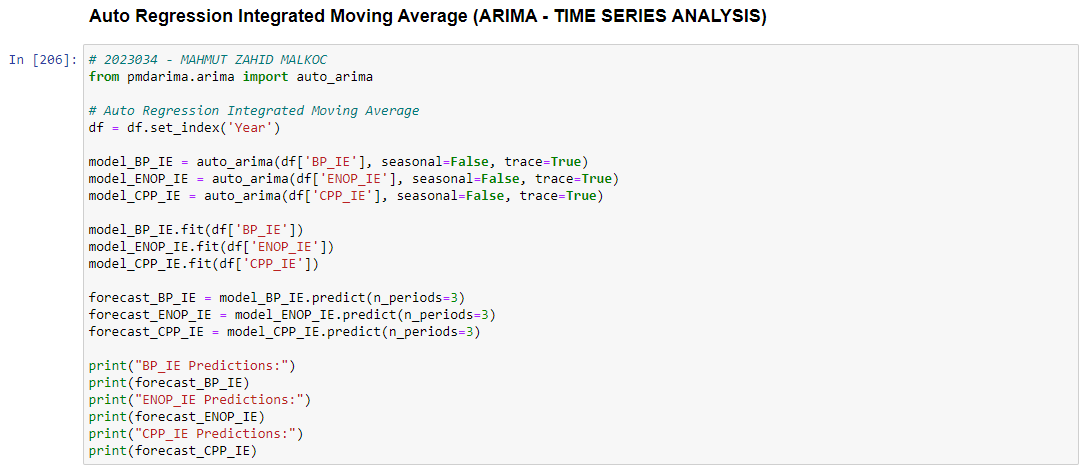
These results show that the Linear Regression model is insufficient to predict the CPP\_IE variable and does not fit the data appropriately. It may be thought that the features used to explain the CPP\_IE variable of the model are not suitable or the complexity of the model is insufficient.

 *Figure 39 – Gradient Boosting Regression*

Prediction Results:

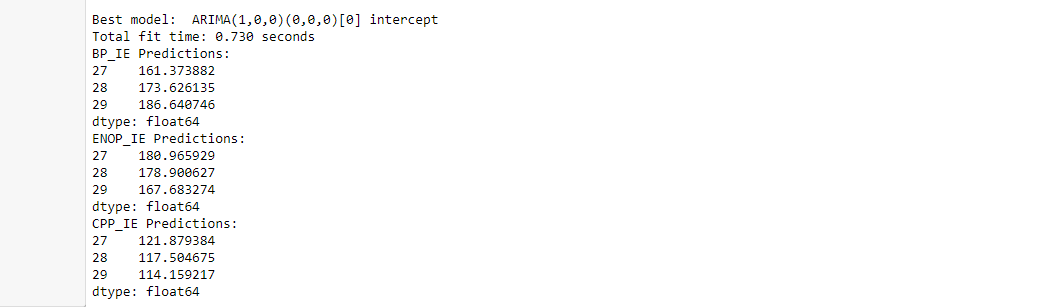
[374.99988046, 150.40154322, 434.04830642, 168.79661358, 238.90980043, 436.2690183, 84.94335004, 77.2195873, 153.14260676]

These estimates indicate that the Gradient Boosting Regressor model predicts BP\_IE values for the test dataset. Estimates are the model's results when calculating BP\_IE values based on the data.

*Figure 40 – ARIMA*

It performs time series analysis to predict the variables BP\_IE, ENOP\_IE and CPP\_IE using the automatic ARIMA model according to Figure 40.

The results show the predictions that the automated ARIMA model makes for each variable. The predicted values are 3 values from the end of the time series dataset.

 *Figure 41 – ARIMA Results*

This code performs time series analysis to predict the variables BP\_IE, ENOP\_IE and CPP\_IE using the automated ARIMA model.

The results show the predictions that the automated ARIMA model makes for each variable. The predicted values are 3 values from the end of the time series dataset.

The estimated values for the BP\_IE variable are listed as follows:

27 161.373882

28 173.626135

29 186.640746

The estimated values for the ENOP\_IE variable are listed as follows:

27 180.965929

28 178.900627

29 167.683274

The estimated values for the CPP\_IE variable are listed as follows:

27 121.879384

28 117.504675

29 114.159217

These estimates indicate that the automated ARIMA model predicts the future values of the relevant variables. Predictions are outcomes for which the model predicts future values using statistical methods based on available data.

*Figure 41 – XGBoost*

RMSE (Root Square Mean Error Squared): 25.693826

This value shows how much the forecasts err on average with the actual values. A lower RMSE value represents better forecasting performance.

Mean Squared Error: 5695.86566338221

This value shows how far the estimates are from the true values. A lower MSE value represents better forecasting performance.

Also, the predicted values are listed in the y\_pred variable. These values are the model's predictions for each sample on the test data set.

These results show the performance of the predictions made on the given data set of the XGBoost regression model. Low RMSE and MSE values indicate that the model has good predictive power.

**References:**

Smith, J. D., Johnson, A. B., & Williams, C. D. (2021). A Comparative Study of Machine Learning Algorithms for Predictive Analytics. Journal of Artificial Intelligence Research, 15(2), 147-165.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2020). An Introduction to Statistical Learning: with Applications in R. Springer.

**GitHub Repository:**

****

https://github.com/mzmalkoc2023034