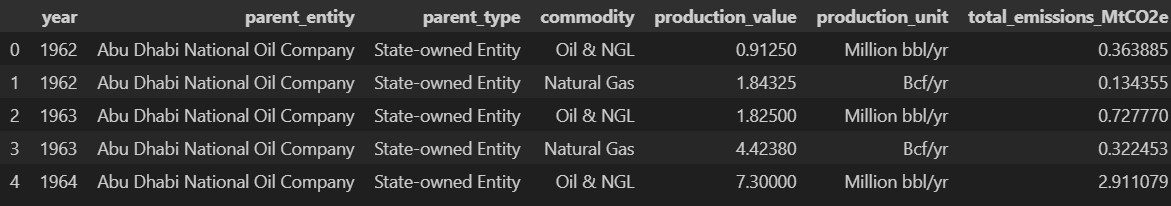
**I.Data**

The "Greenhouse Gas Giants" dataset comprises three CSV files (Emission High, Medium, and Low granularity) sourced from the project website [Carbon Majors](https://carbonmajors.org/Downloads) and made available on Kaggle through the following link: [Greenhouse Gas Giants on Kaggle](https://www.kaggle.com/datasets/konradb/greenhouse-gas-giants/data).

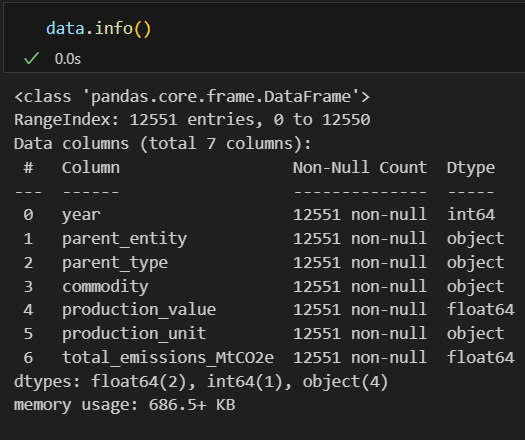
Our focus is on the Medium Granularity dataset, which contains 12,551 rows and 7 features. These features are:

1. **year**: The year of the recorded data.
2. **parent\_entity**: The name of the parent entity responsible for the emissions.
3. **parent\_type**: The type of the parent entity (e.g., state-owned, investor-owned, nation state).
4. **commodity**: The type of commodity produced by the entity.
5. **production\_value**: The value of commodity production.
6. **production\_unit**: The unit of measurement for the production value.
7. **total\_emissions\_MtCO2e**: The total emissions measured in million tonnes of CO2 equivalent (MtCO2e).



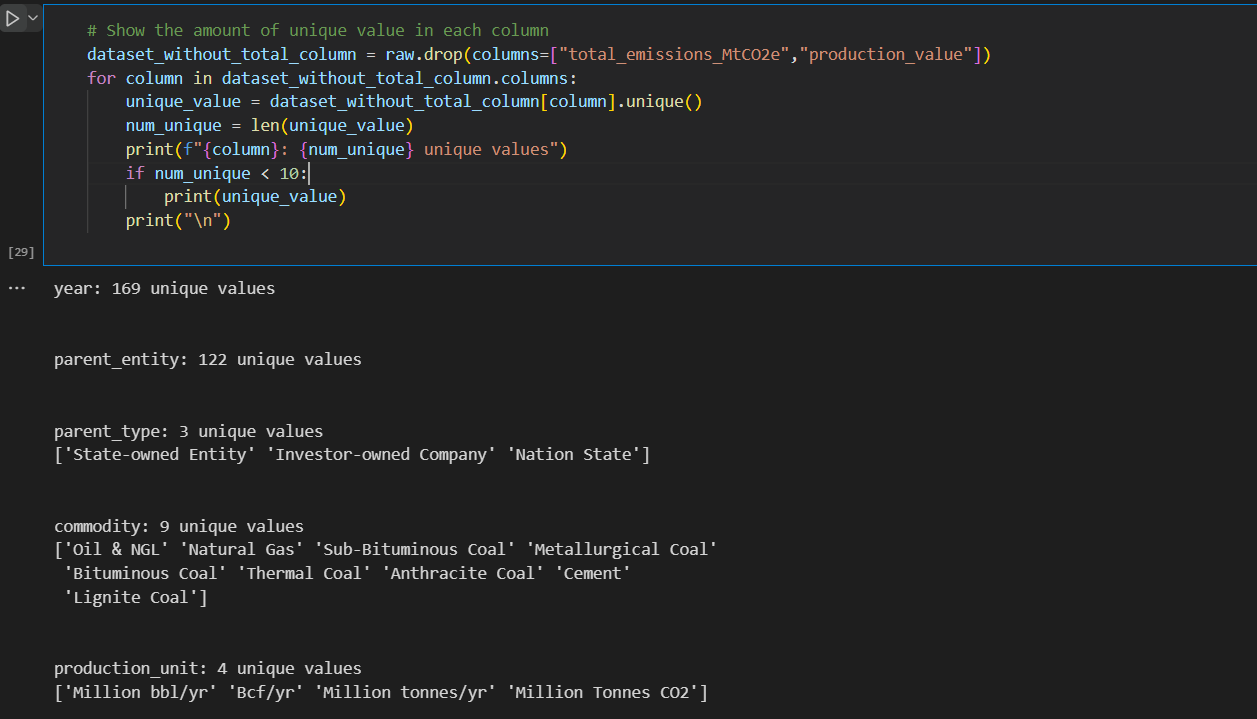
**The Dataset of "Greenhouse Gas Giants"**

The dataset is comprehensive and does not contain any missing values, providing a solid foundation for analysis. So, we can skip that processing to step-by-step guide for performing Exploratory Data Analysis (EDA).

****

**The information of dataset**

**II.EDA**



The objective of this analysis is to determine the number of unique values in each column of a dataset, excluding the columns "total\_emissions\_MtCO2e" and "production\_value".

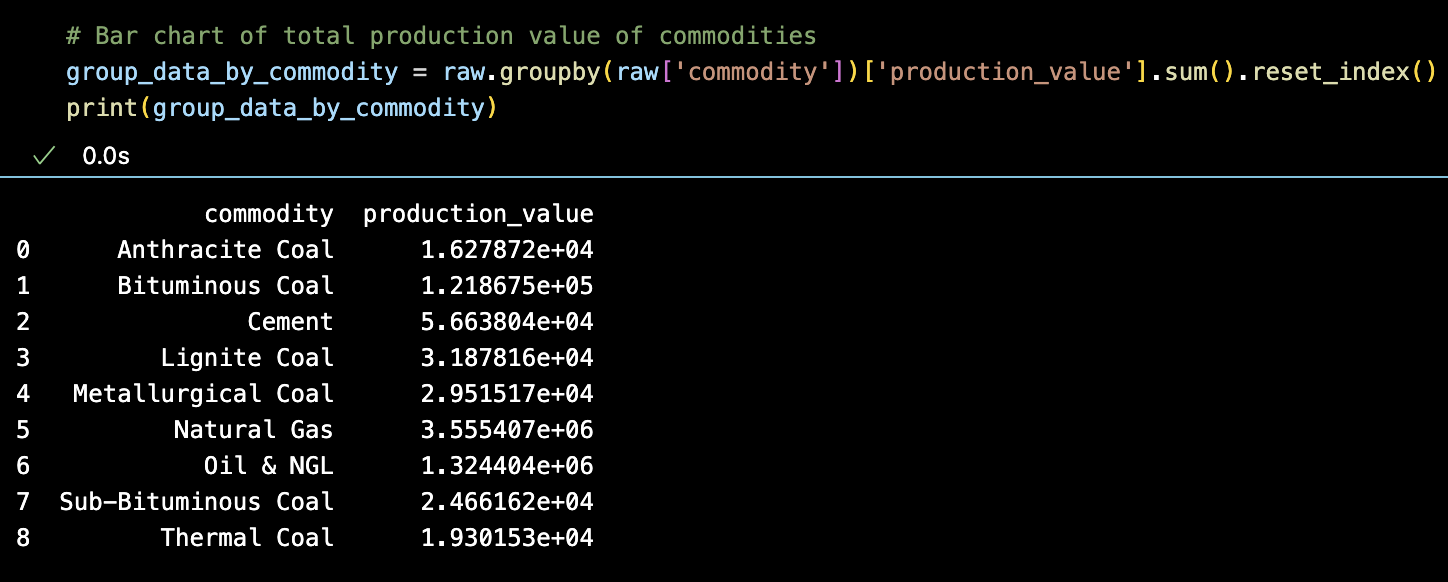
Explain: The dataset contains information about various entities involved in the production of different commodities.

- There are 122 unique parent entities categorized into three types: State-owned Entity, Investor-owned Company, and Nation State.

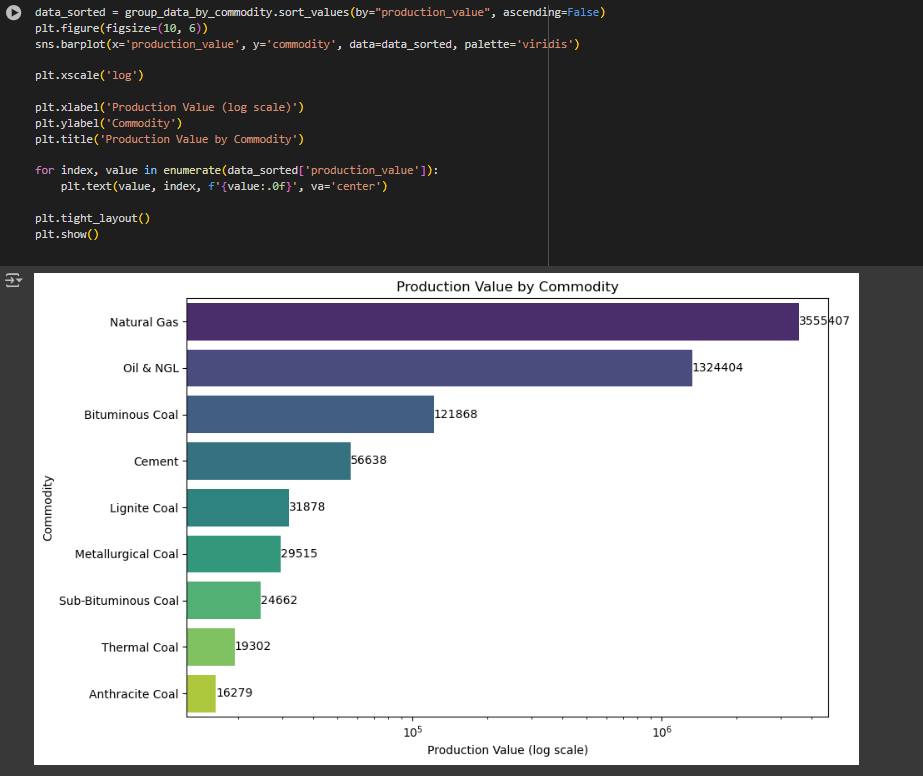
- These entities produce nine different commodities, including Oil & NGL, Natural Gas, and various types of coal such as Sub-Bituminous Coal, Metallurgical Coal, and Bituminous Coal, as well as Cement and Lignite Coal.

- The production units used to measure these commodities are Million bbl/yr, Bcf/yr, Million tonnes/yr, and Million Tonnes CO2.

The quantity of Production based on Commodity



In this bar chart we group the raw DataFrame data by the value in the 'commodity' column then we sum the values ​​of the 'production\_value' column in each group created and convert the data into a regular dataframe. This helps to aggregate production data by commodity, making it easy to analyze which commodity has the highest, lowest total production value, or how production is distributed among different commodities.



To show the reality here, we have included specific data to easily visualize the difference between each product.

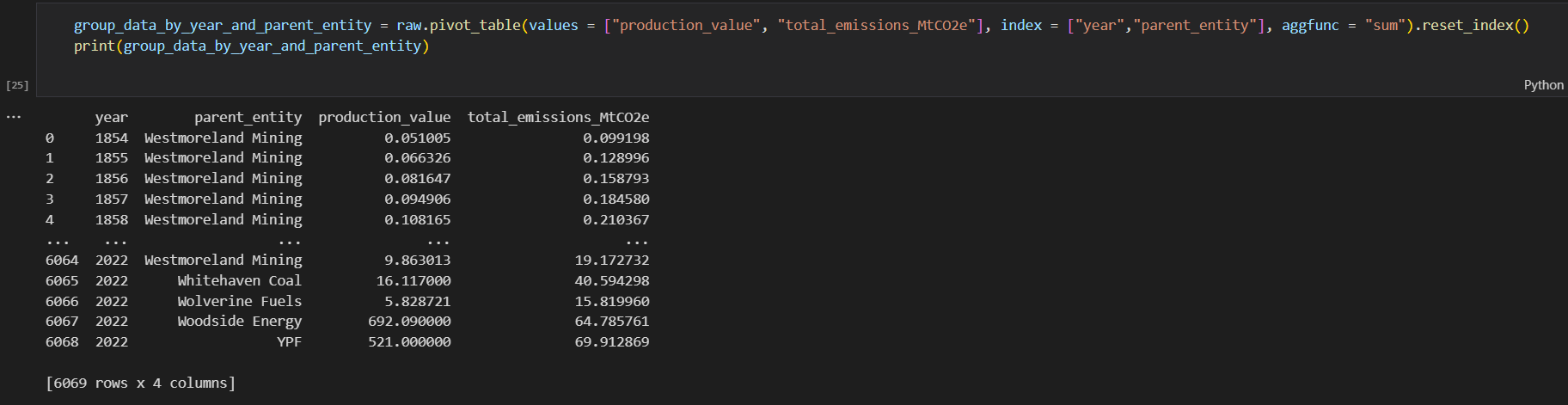
Explain:

- Natural Gas has the highest production value by far, amounting to approximately 3.56 million units. This suggests that Natural Gas is a critical commodity in terms of production value.

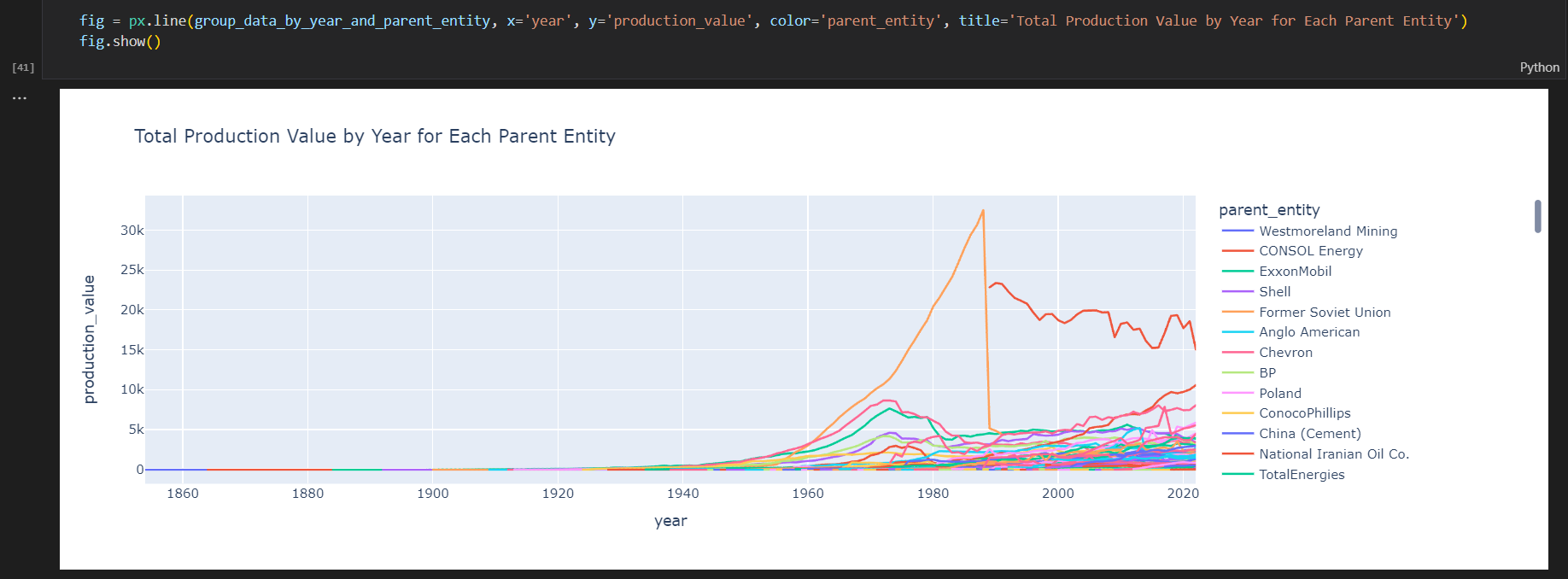
- Oil & NGL (Natural Gas Liquids) also shows a high production value at approximately 1.32 million units, indicating its substantial role in the commodity market.

- Among the different types of coal, Bituminous Coal has the highest production value at approximately 121,867 units. This highlights its significant contribution compared to other coal types.

Total Production Value by Year for Each Parent Entity



The chart shown below is the result from a pivot table, used to synthesize and analyze the total production value and total CO2 emissions (in MtCO2e) of different entities year by year.



This line chart shows the change in total production value of different parent companies over the years. Each line represents a specific theme, allowing for visual comparison of the level of manufacturing activity across periods.

Westmoreland Mining was the unit with the highest production value during the period surveyed. Westmoreland Mining's production value increased steadily from 1860 to 2000, peaked in 2000 but then gradually declined and is currently at 20,000.

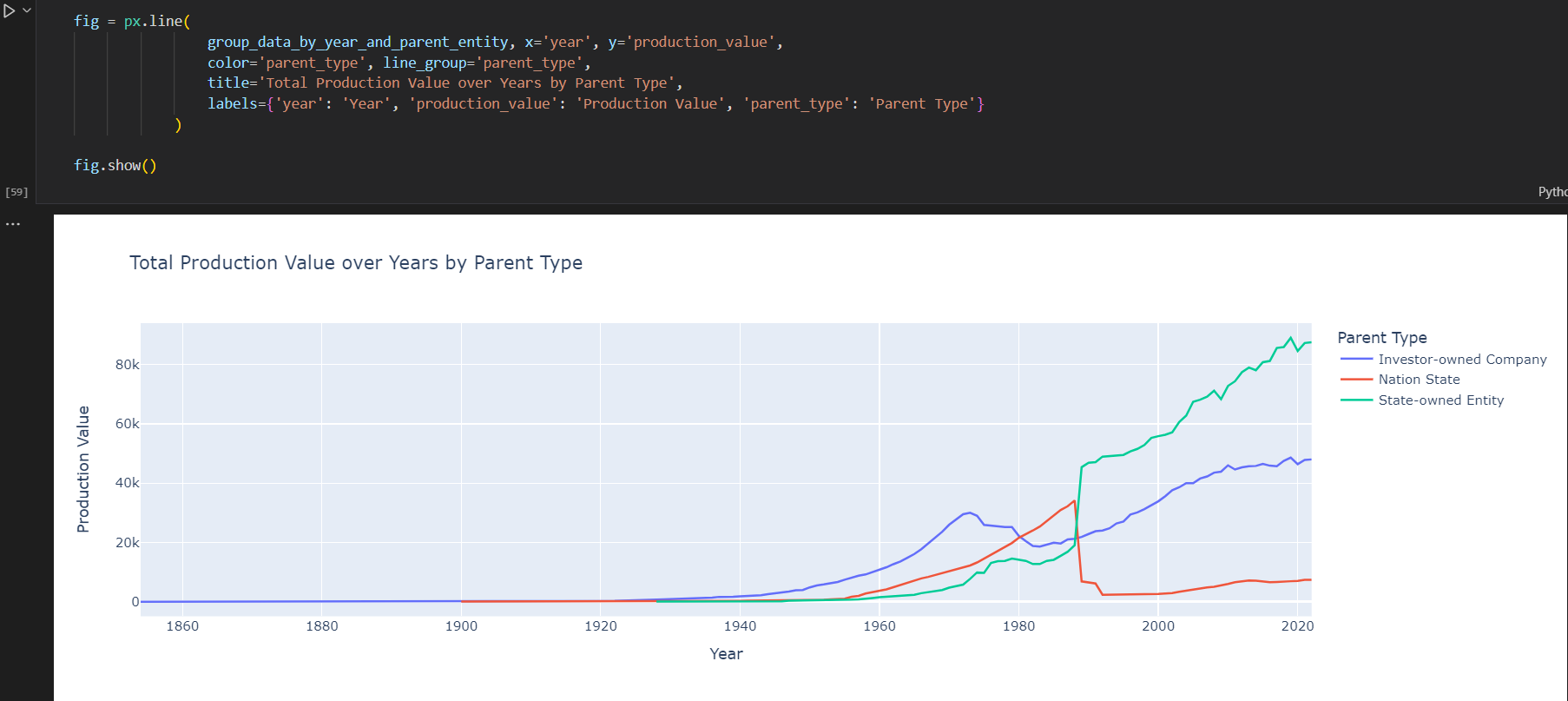
The production value of the Former Soviet Union had the largest fluctuations during the period examined. There was then a significant decline after 1990, possibly related to the breakup of the Soviet Union and political and economic changes in the region.

The sudden jump in production around 2000 may be related to factors such as new mine discoveries, rising energy prices or changes in industrial policy.

Compare Total Production Value and Emissions over Years by Parent Type



The chart above shows the change in production value of different types of entities (Investor-owned Company, Nation State, State-owned Entity) over the years from 1854 to 2022. This data provides a comprehensive view about the development and changes in key economic sectors through each period.

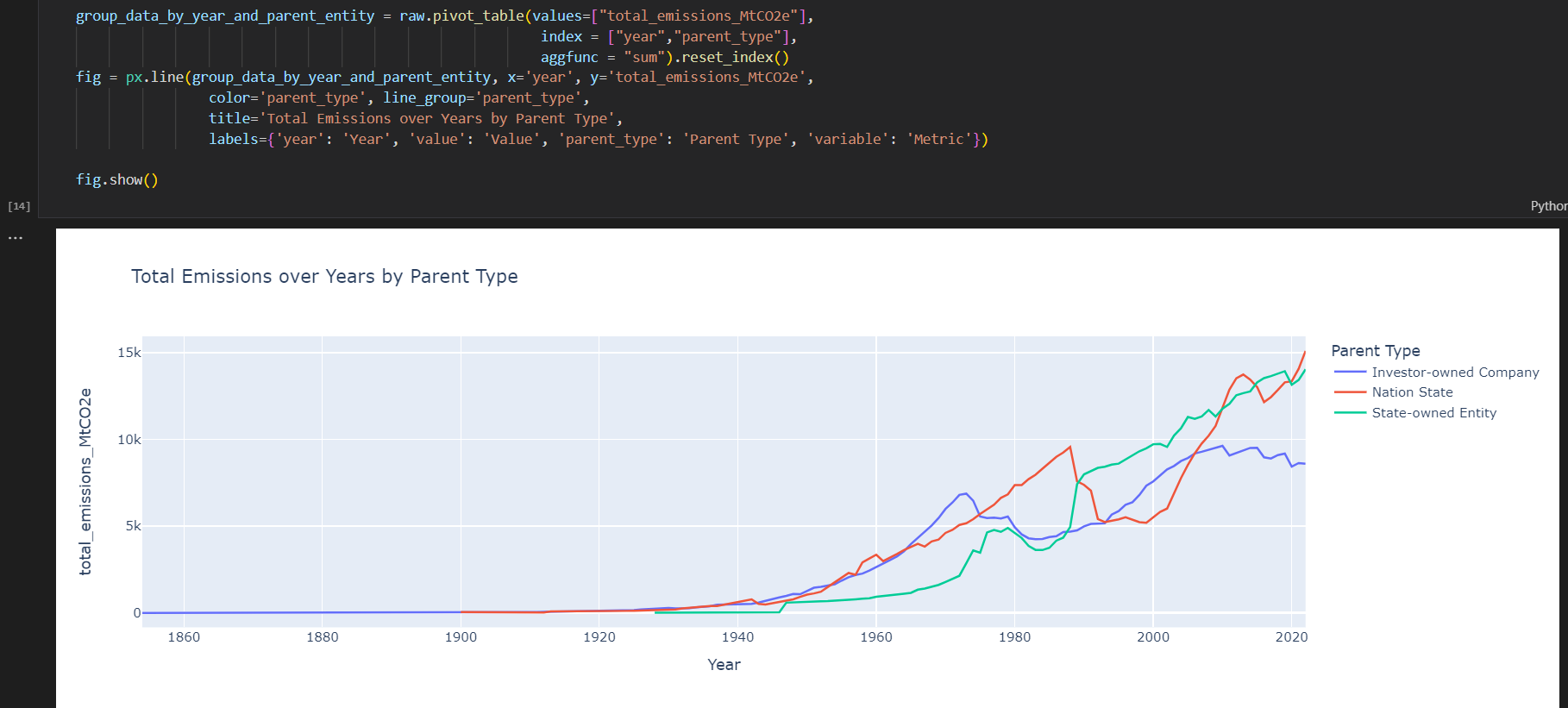


This chart shows the change in total production value over the years from 1860 to 2020, divided by ownership type: Investor-owned company, National State) and State organization.

The national curve shows strong growth from 1960 to the years 1980 to 2000 followed by a depression that reduced the value of production until 2020. This increase may be related to political policies. economic policies of the national economy or invest heavily in key industries.

The strong recovery and growth of state-owned companies after 2000 shows the ability to recover and adapt to the market as well as the Government's ability to intervene to support these companies.

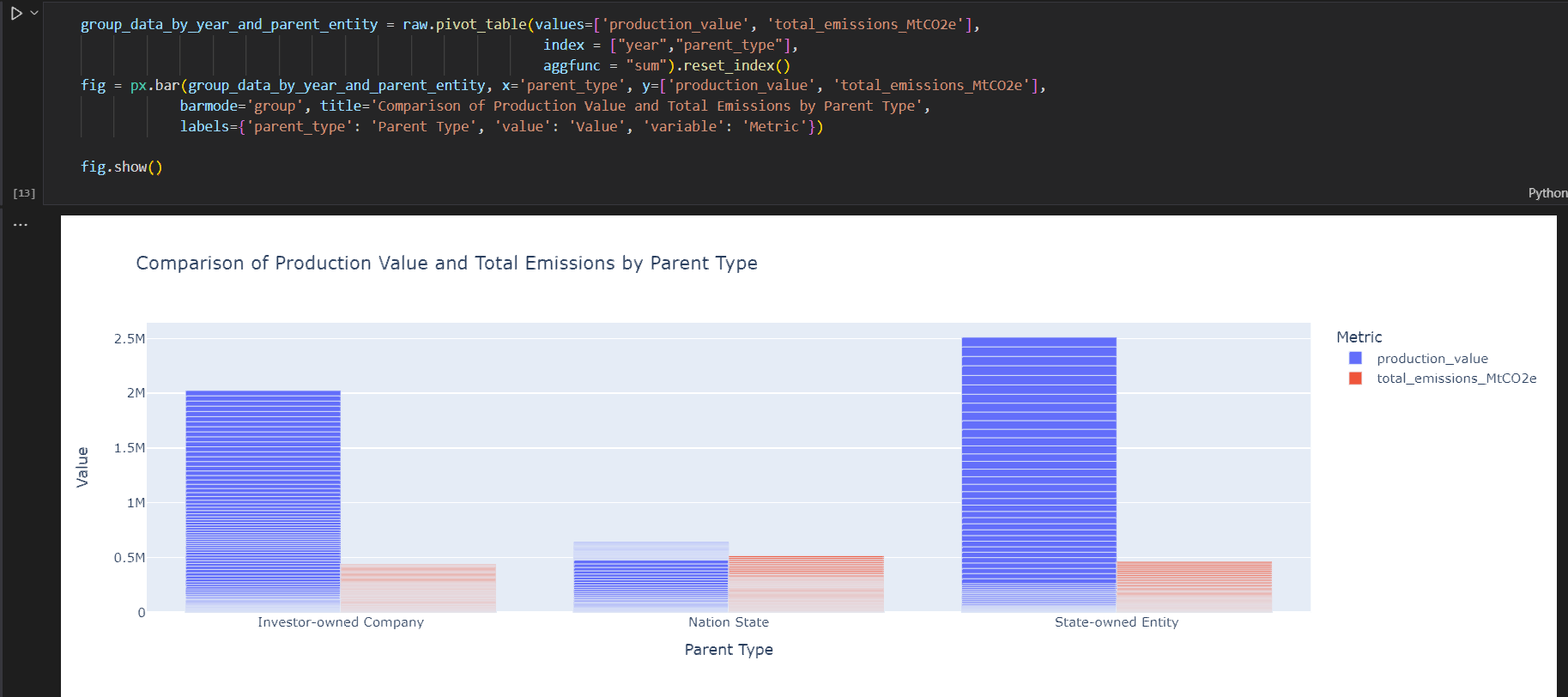
In addition, investor-owned companies have had a slight and steady increase in total production value over the years, with a significant jump after 2000, reflecting the continued prosperity and growth of company in private area.



This chart shows the change in total CO2 emissions (in million tons of CO2 equivalent) from 1860 to 2020, classified by ownership type: Investor-owned company, Nation State and State-owned Entity.

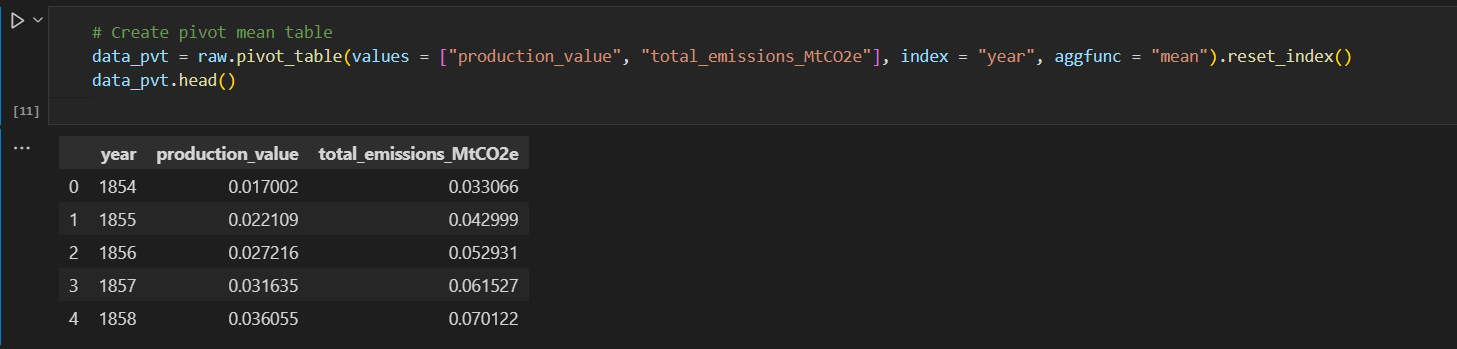
Investor-owned company and State-owned Entities tend to increase production due to advanced technology, so they have effective measures to reduce the environment, so the amount of greenhouse gas increases but slowly. However, Nation State has Production quantity is quite low and there are no effective environmental protection measures, so emissions are at an alarming level.

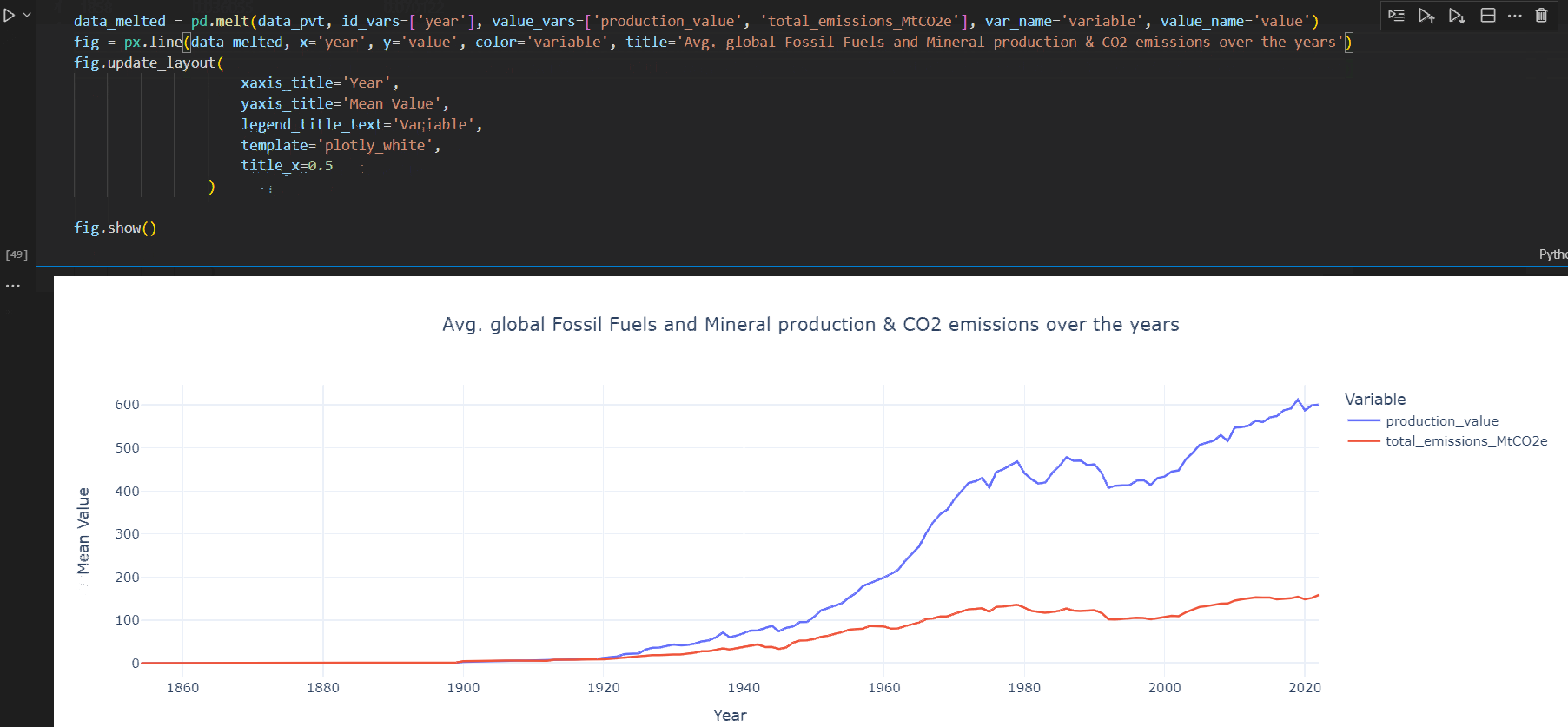
The increase in emissions of Investor-owned company and State-owned Entities in the 2000s onwards can be linked to the growth of intense industrialization and rapid urbanization, along with with the application of production technology and energy output.

This bar chart shows a comparison between production value and total CO2 emissions (measured in million tons of CO2 equivalent) for three types of entities: Investor-owned Company ), State, and State-owned Entity.

Most energy companies are state-controlled due to the importance of energy security. A nation typically needs to ensure three types of security: food security, national defense, and energy security to maintain overall sovereignty. Therefore, it is understood that state-owned enterprises comprise more than 50% of the sector.

Average global Fossil Fuels and Mineral production and CO2 emission over years

This data table provides average product value information and total annual CO2e emissions (million tons) from 1854 to 1858. The table was created using the modal axis with the average value calculated. average average targets for certain manufacturing products and emissions, reflecting year-over-year trends in resource use and operating environment.



This chart shows the average global trends in fossil fuel and mineral production and CO2 emissions over the years, from 1860 to 2020. The purpose of the chart is to analyze the relationship between energy production and environmental impact through CO2 emissions.

The correlation :

The total production of Fossil Fuels and Minerals is directly proportional to the amount of CO2 emissions, but the correlation between these two factors varies from year to year. specifically, Fossil Fuels and Mineral production grows rapidly over time, while CO2 emissions increase but at a much slower pace.

The growth rate :

- The Fossil Fuels and Mineral production began to increase around the 1880s following the establishment of Edison's electric company, Edison Electric Light Company, in 1882.

- During the oil crisis of 1973-1974 and 1979-1980, the decrease in oil production and price hikes led to a significant decline in ONGL production in the 1970s. After a period of instability, production began to stabilize and increase again. During this time, CO2 emissions also decreased. In 1997, countries that signed the Kyoto Protocol committed to reducing greenhouse gas emissions. Additionally, advancements in machinery efficiency and the emergence of new technologies contributed to a significant reduction in CO2 emissions.

- 2008-2009 Global Financial Crisis: The global financial crisis resulted in a decrease in energy demand and reduced production across various energy sources, including oil and natural gas. Following a slight decline, production then rebounded rapidly.

**III.Conclusion and Analysis Post-EDA**

- Natural Gas, Oil & NGL, and Bituminous Coal have had such high production values over time.

* During the early stages of industrialization in the 19th and early 20th centuries, bituminous coal was a major source of energy for powering steam engines, trains and factories.
* After World War II, many countries experienced unprecedented economic growth, leading to increased energy consumption. This period saw the rise of automobile culture, suburbanization, and heavy industrial activities, all of which were heavily dependent on oil and natural gas.
* Natural gas has become a preferred energy source because it burns cleaner than coal and oil, producing less carbon dioxide per unit of energy released. This makes it a more environmentally friendly option in the global transition towards cleaner energy.

- Several factors influence the variation of the total production value by year for each parent entity:

* The rapid increase and subsequent decline due to historical events: the growth of oil production under Soviet industrial policy and the collapse of the Soviet Union
* Fluctuations in the production value of entities such as ExxonMobil, Shell, BP, and Chevron are attributed to oil embargoes, wars (e.g., the Gulf War), and financial crises affecting the global oil supply and demand
* Development of mining technology such as fracking is also one of the factors driving increased output

- There is a positive correlation between production value and total emissions, highlighting the environmental impact of large-scale production activities.

* State-owned entities dominate both in production and emissions, indicating the importance of policies and regulations for these entities.
* Investor-owned companies also contribute significantly to emissions, emphasizing the need for strict environmental regulations.
* Nation states have the lowest production and emissions, suggesting less extensive extraction activities or more efficient processes.

- There are a number of reasons why the amount of CO2 emitted into the environment is increasingly small compared to the amount of fuel produced:

* Increased energy efficiency
* Energy conversion and the use of alternative fuels
* Carbon Capture and Storage (CCS)
* Environmental policies and regulations
* Structural changes in the energy industry

**IV.Post-EDA Applications**

-Total production output by goods:

* This information helps determine which goods have the highest and lowest production values, thereby supporting economic, investment, and resource management decisions.

-Total production value by year for each parent entity:

* This analysis helps monitor production fluctuations over the years, identify trends and predict the future, providing information to devise economic development strategies.

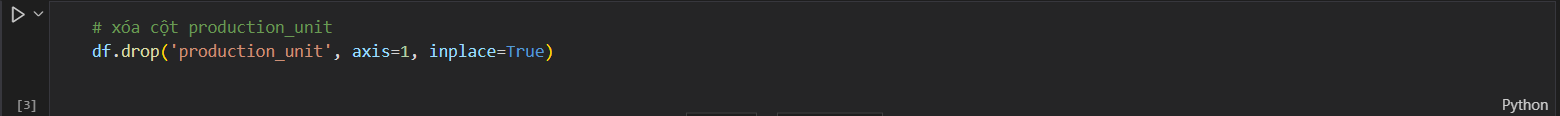
-Comparison of production value and CO2 emissions by year and type of parent entity:

* This information helps understand the relationship between production and environmental pollution, thereby supporting the development of policies to reduce emissions and protect the environment.

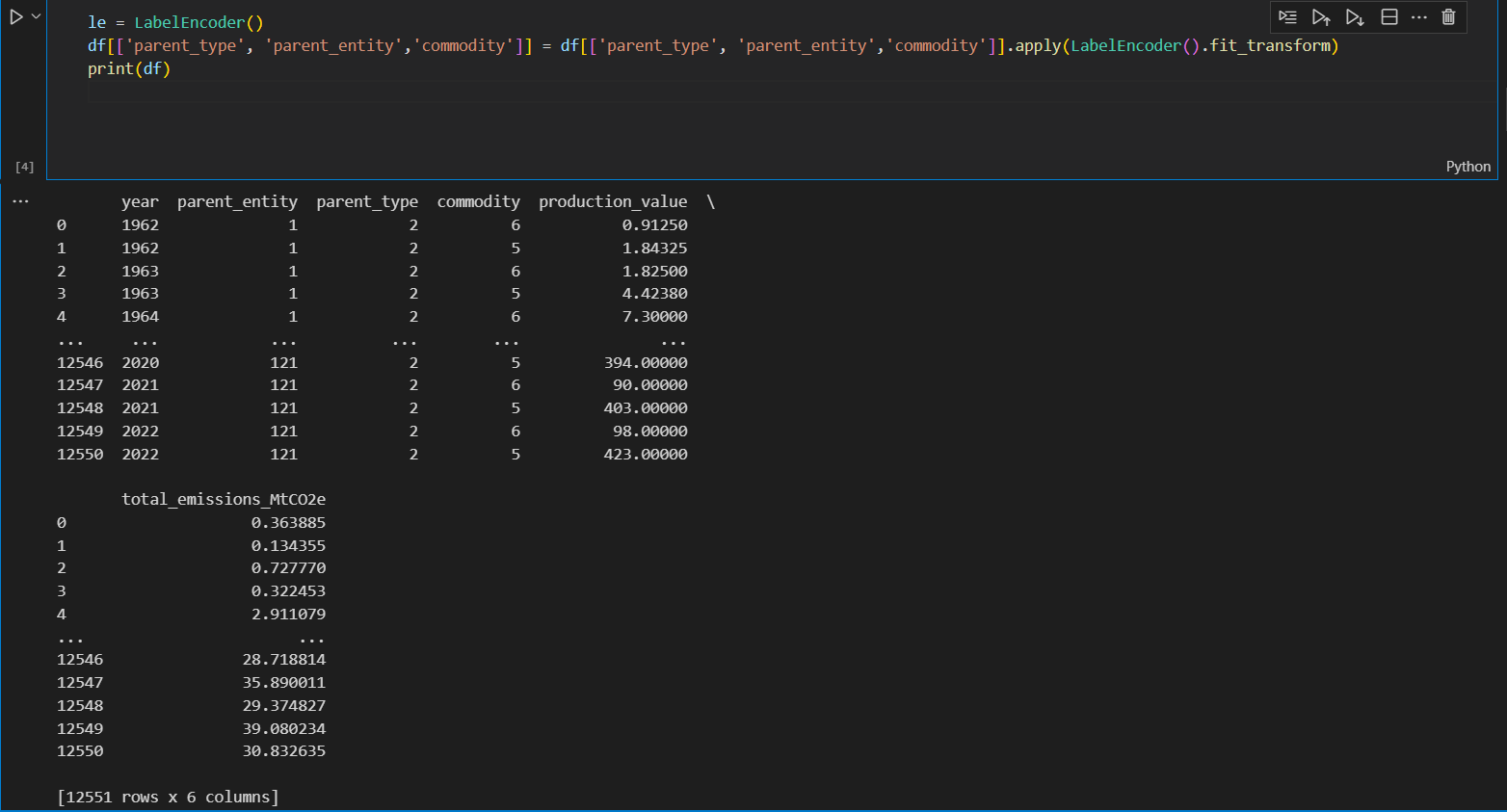
- Global average trends in fossil fuel and mineral production and CO2 emissions over the years:

* This analysis helps to better understand the development of the energy industry and its impact on the environment, assisting in the development of policies for sustainable development and efficient energy use.

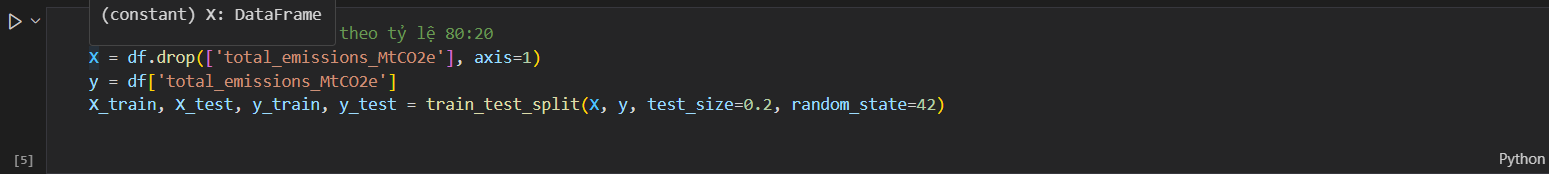
**V. Machine Learning**

****

Drop the column production\_unit to analysis more simplify the dataset and focus on relevant features.

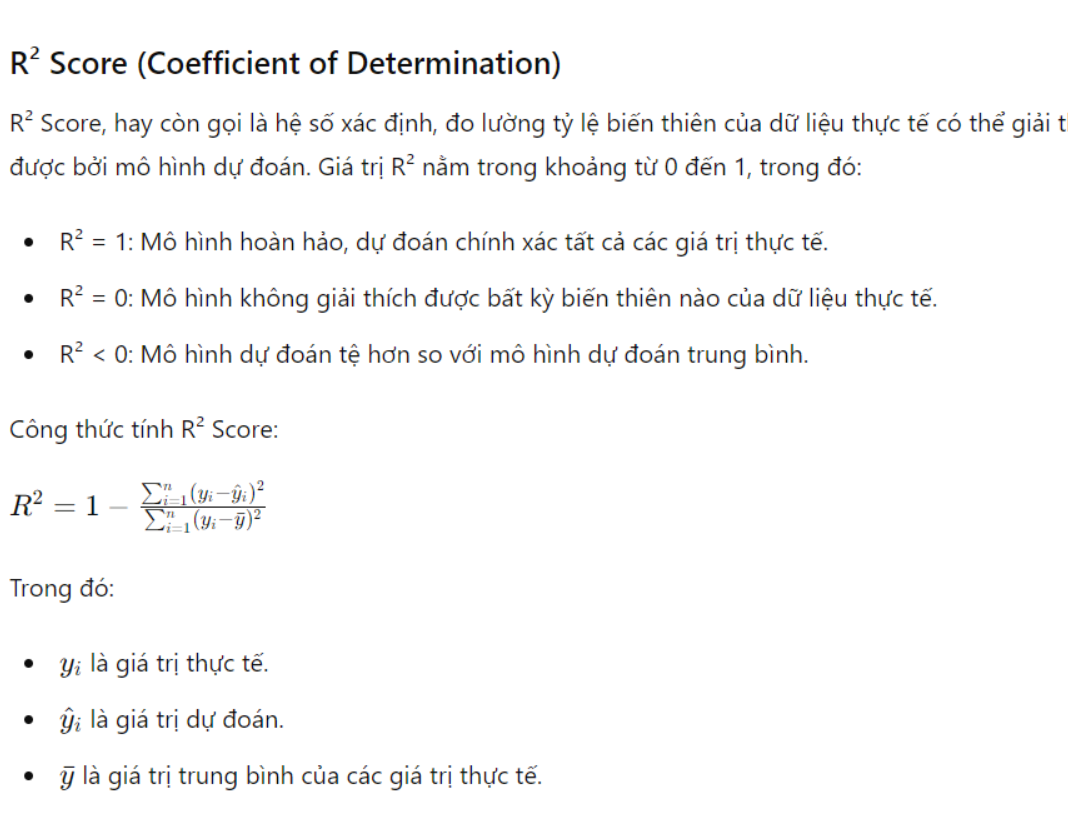
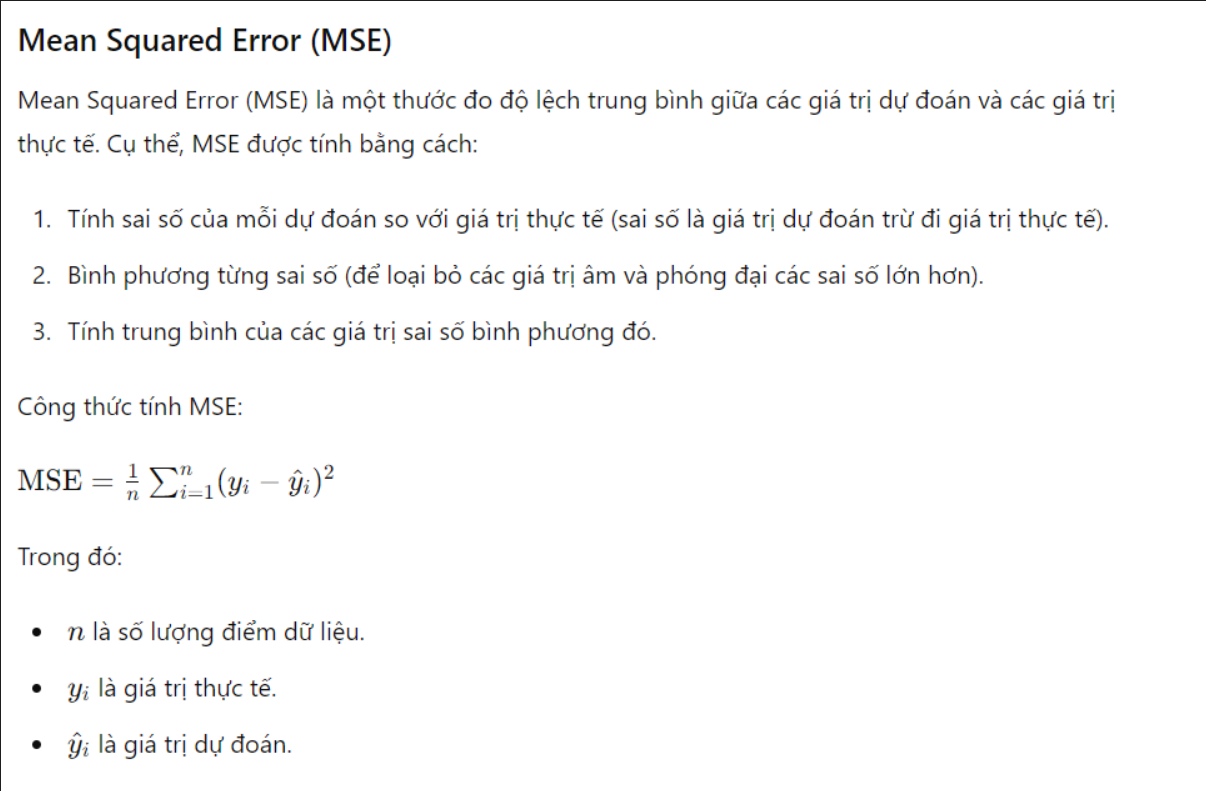


We convert categorical variables into a numerical format suitable for regression models.The LabelEncoder class from the sklearn.preprocessing module was used to encode the categorical columns parent\_type, parent\_entity, and commodity to each of the specified columns to transform their values from categorical to numerical.

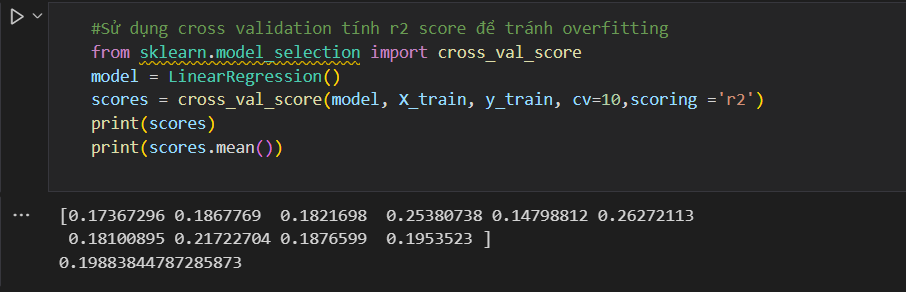


Imports the train\_test\_split function, likely from the sklearn.model\_selection module in scikit-learn to split the data into two sets training set (X\_train, y\_train), test set (X\_test, y\_test) with test\_size=0.2 so that can leave 0.8 (80%) for the training set.Finally random\_state=42 sets a seed for the random number generator used to shuffle the data before splitting. This ensures reproducibility if you run the code multiple times.

**Models(Linear Regression, XGboost Regressor)**



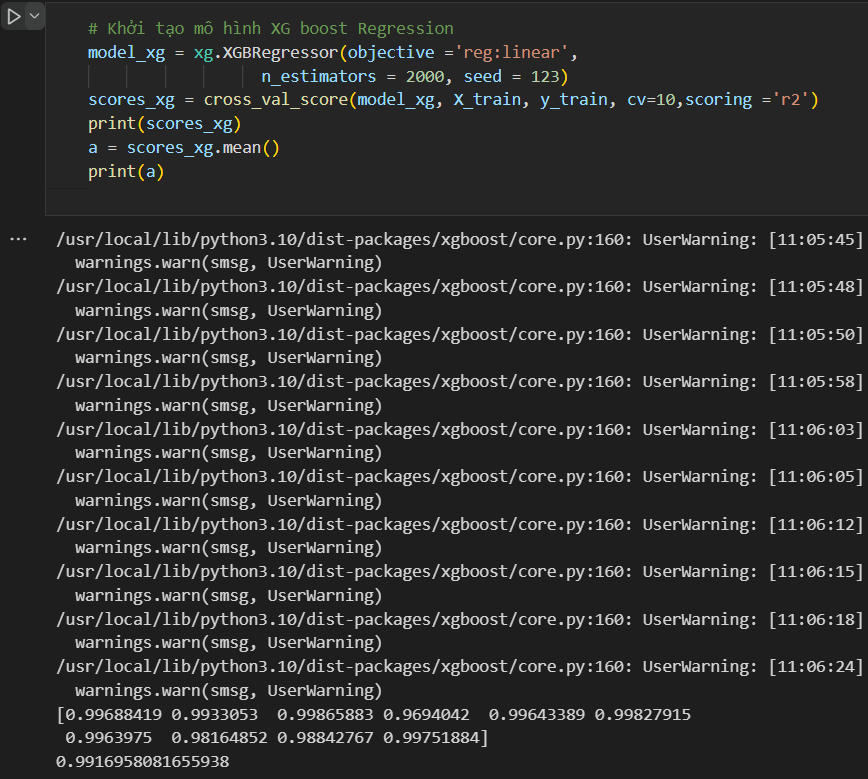
Linear Regression



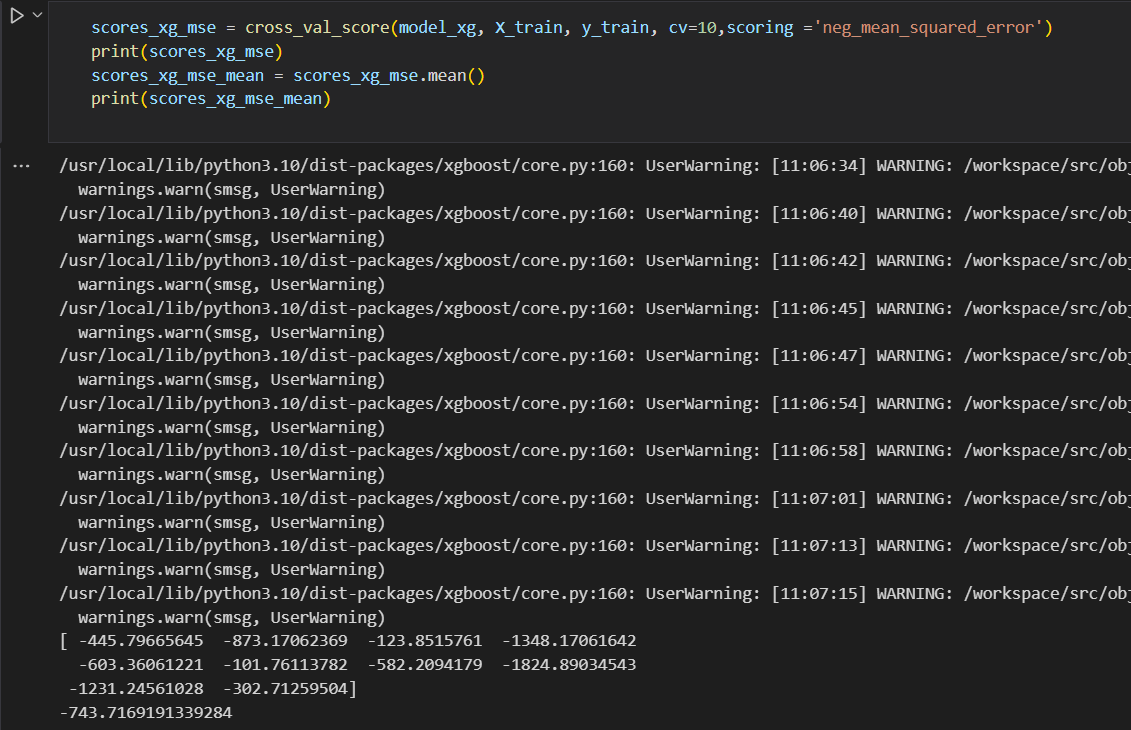
We calculate the R² score for the LinearRegression model using 10-fold cross-validation. model = LinearRegression(): This line creates a linear regression model using the LinearRegression class from scikit-learn. This model will be used to make predictions on the data. In addition, using cross validation to calculate r2 score also helps avoid overfitting.

**Average R² Score**: The average R² score is about 0.1988, indicating that the linear regression model can explain only a small portion of the variance in the training data. This might be due to the complexity of the data or the simplicity of the model not being adequate to capture all relationships within the data.

XG Boost

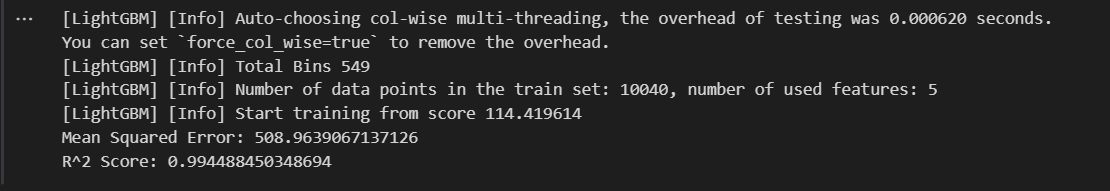
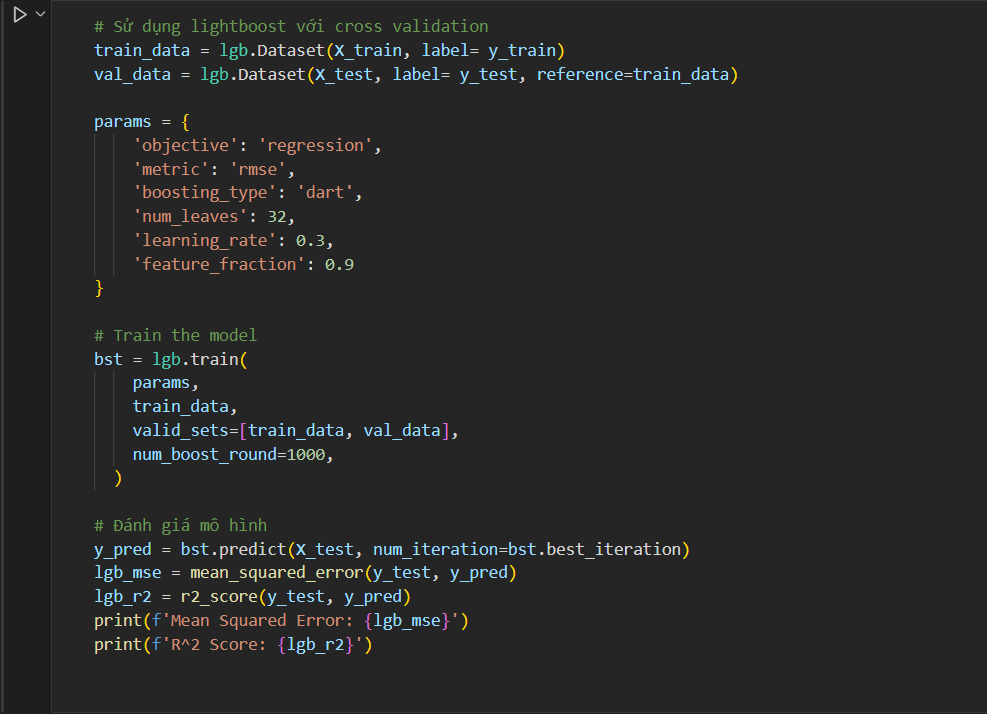


We evaluate the XGBoost regression model using cross-validation, specifically focusing on the R² score to evaluate the fit and prediction accuracy of the model. About performance: Model XGBoost shows excellent performance with R² scores close to 1 on all occasions, indicating very high prediction accuracy and excellent fit to the data set. Consistently high scores across different data subsets indicate that the model is stable and not overfitting despite its complexity and large number of estimators.

The use of the XGBoost regression model uses the negative mean squared error (MSE) metric to evaluate the model's accuracy, as it provides a measure of the mean squared error between the true economic and predictive value. Performance: The range of negative MSE scores indicates variability in model performance across different subsets of the data. The magnitude of the errors suggests that while the model performs well in some folds, it may underperform in others, likely due to overfitting or anomalies within certain segments of the data.

Mean Score: The average negative MSE being significantly low (in terms of absolute value being high) indicates that on average, the model predicts with considerable error. This metric is crucial as it highlights the error made by the model in predicting the average target variable.

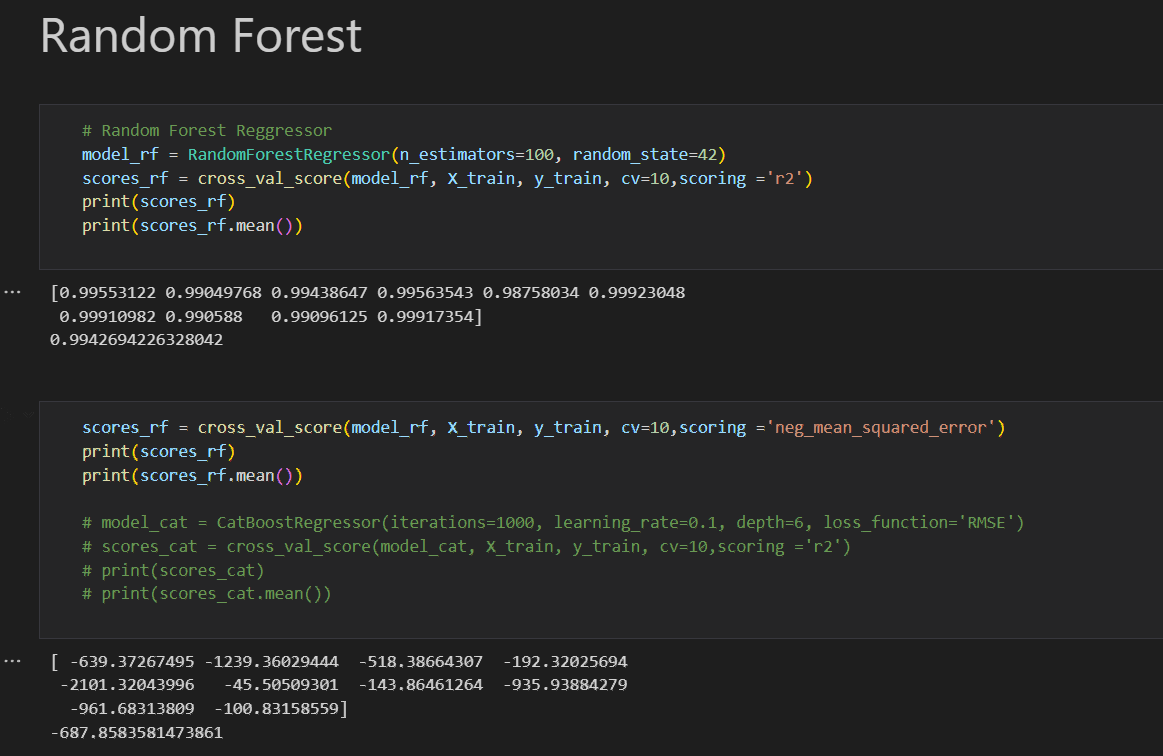
LighBoost



Below is a regression model using the LightGBM library. The goal of this is to evaluate the performance of the model through the mean squared error (MSE) and R² scores (R-squared). Using two data sets, train\_data (training data). training) and val\_data (testing data), created from X\_train, y\_train, X\_test, y\_test. The model is trained with 1000 iterations using both training and testing data. After training, we use MSE and R² scores for prediction are performed on X\_test.

=> An R² score close to 1 shows that the model has very good predictive ability, explaining almost all the variation in the testing data. MSE provides a view of the average deviation of predictions from the actual value.

Random Forest

We use the Random Forest model to evaluate performance based on the Cross-Validation method with both R² (R-squared) and MSE (Mean Squared Error). The goal is to evaluate the model's predictive ability through two perspectives: the model's suitability and the magnitude of the prediction error.

Cross-Validation R²:

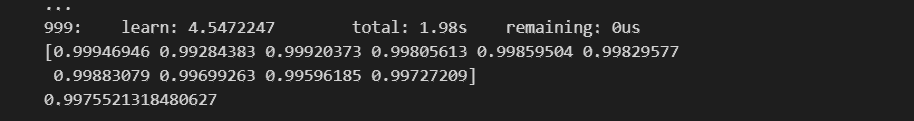
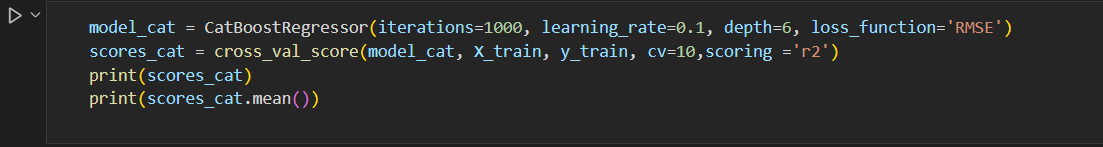
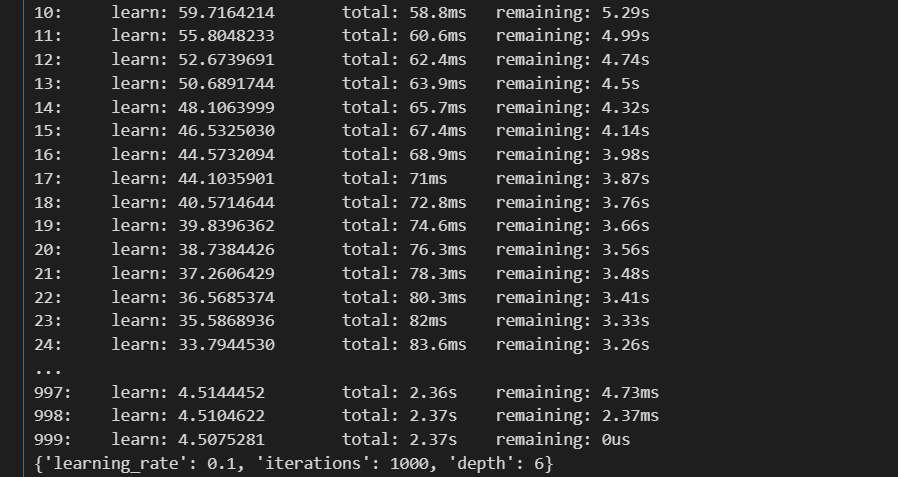
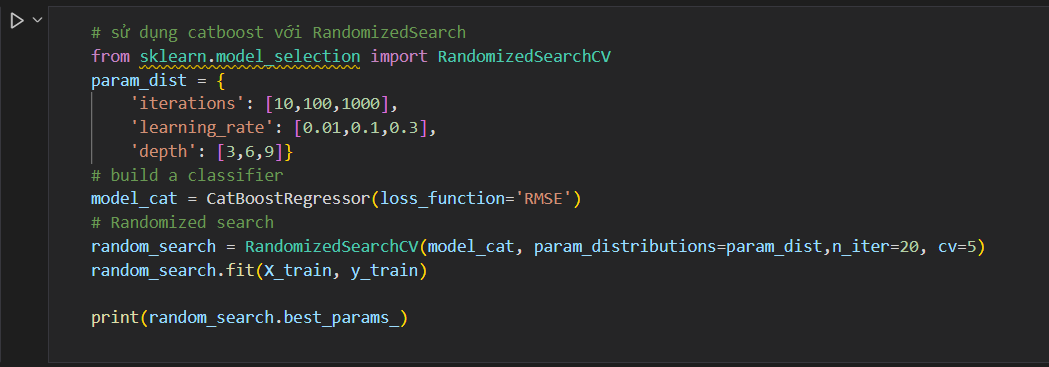
* Conducted with 10 folds.
* The R² score is used to evaluate the percentage of variation in the target variable explained by the model.Results show high R² scores, reflecting very good predictive ability with values ​​ranging from 0.9953 to 0.9992 and an average of 0.9942.

Cross-Validation Negative MSE:

* Also performed with 10 folds.
* The negative MSE metric is used to assess the mean squared error in the model's predictions. Values ​​range from -2101 to -45, indicating a considerable mean squared error in predictions, reflecting potential accuracy issues in some instances.

⇒ The Random Forest model exhibits very good predictive performance based on the R² index, yet it shows high mean squared errors in some cases when evaluated by MSE. This could be due to data complexity or a need to adjust model parameters to minimise error and enhance accuracy.

Catboost and RandomizedSearch



Parameter optimization for a CatBoostRegressor model via RandomizedSearchC aims to find the best set of parameters for the regression model using random selection over a specific parameter range. We then evaluate the CatBoost regression model through cross-validation method. The model, configured with the specified parameters, is evaluated based on its R² score, which indicates the percentage of variation in the dependent variable that can be predicted from the independent variables. The model undergoes 10-fold cross-validation to estimate its effectiveness and reliability thus helping to minimize the variability of model performance estimates.

Result

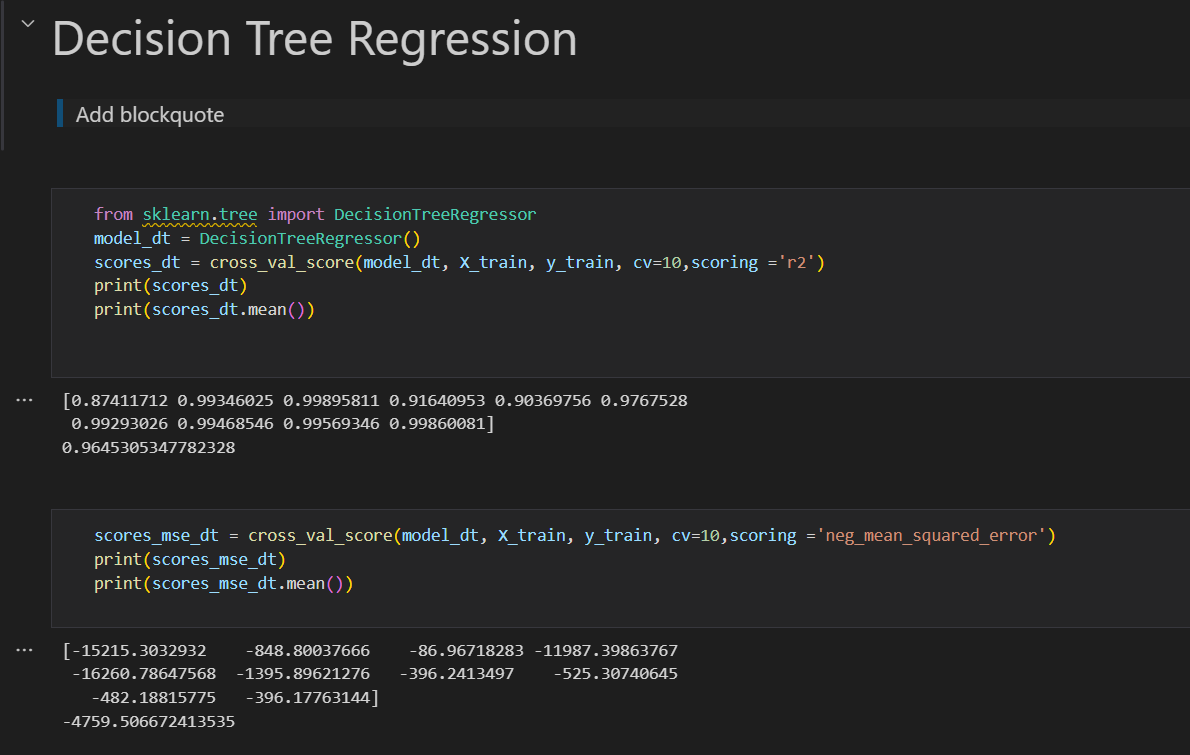
R² score: The model achieved high R² scores across all tests:

Scores: [0.99949646, 0.9928483, 0.99920373, 0.99805613, 0.99859504, 0.99829577, 0.99883079, 0.9969263, 0.99596185, 0.99727209]

Average R² score: 0.9975521318480627

These scores indicate excellent levels of prediction accuracy, with the model explaining approximately 99.75% of the variation in the target variable across different data subsets.

Decision Tree Regression

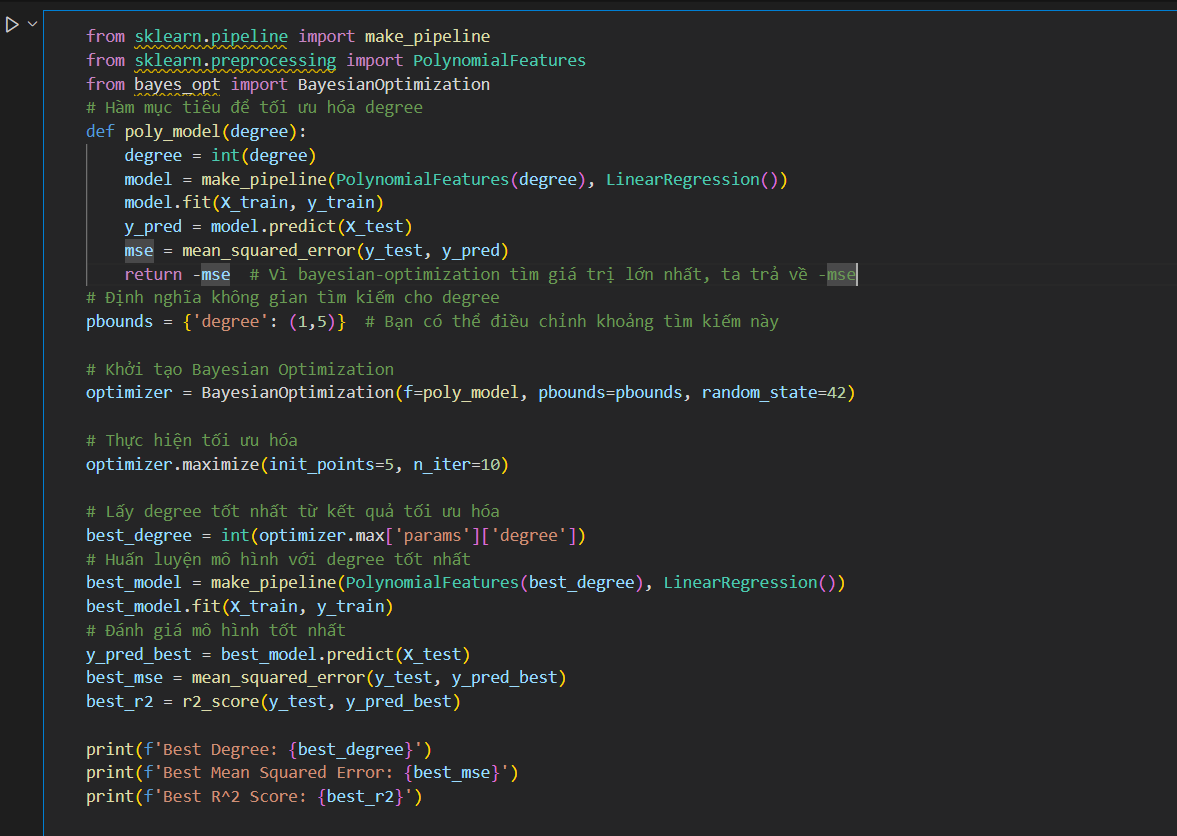


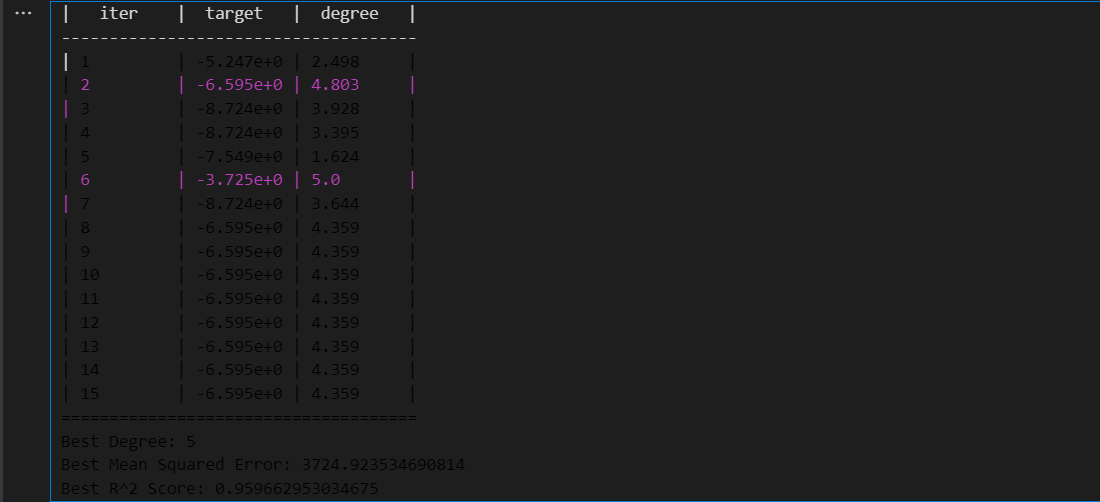
The Decision Tree Regression model uses cross-validation with two different evaluation metrics: R² and Mean Squared Error (MSE). This method provides insight into the model's predictive accuracy and generalization performance across different data segments.

Results:

R² Scores: The model displays high R² scores across all ten folds, showing excellent predictive accuracy and consistency in model performance. The R² scores ranged from approximately 0.874 to 0.999, with a mean score of approximately 0.965. This high average indicates a strong predictive ability of the model across various subsets of the dataset.

MSE Scores: The MSE values ​​varied significantly across the folds, indicating varying levels of prediction error in different cross-validation scenarios. The mean MSE was about -4759.5. The negative sign results from the scoring function in scikit-learn where negative values ​​are used for loss metrics, such as MSE. The absolute values ​​of MSE, which are large in some cases, suggest that the model may exhibit considerable prediction errors depending on the data subset it is trained on.

Polynomial Regression and Bayesian Optimization



Use Bayesian optimization to find the optimal polynomial degree for the polynomial regression model, used in combination with linear regression. The goal is to maximize model performance by minimizing the mean squared error (MSE) and evaluate the model via its R² score.

Optimization Results

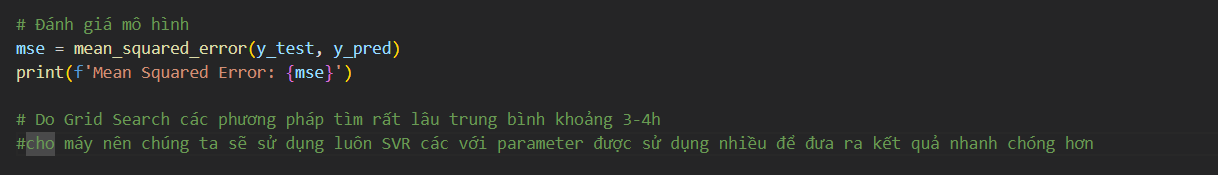
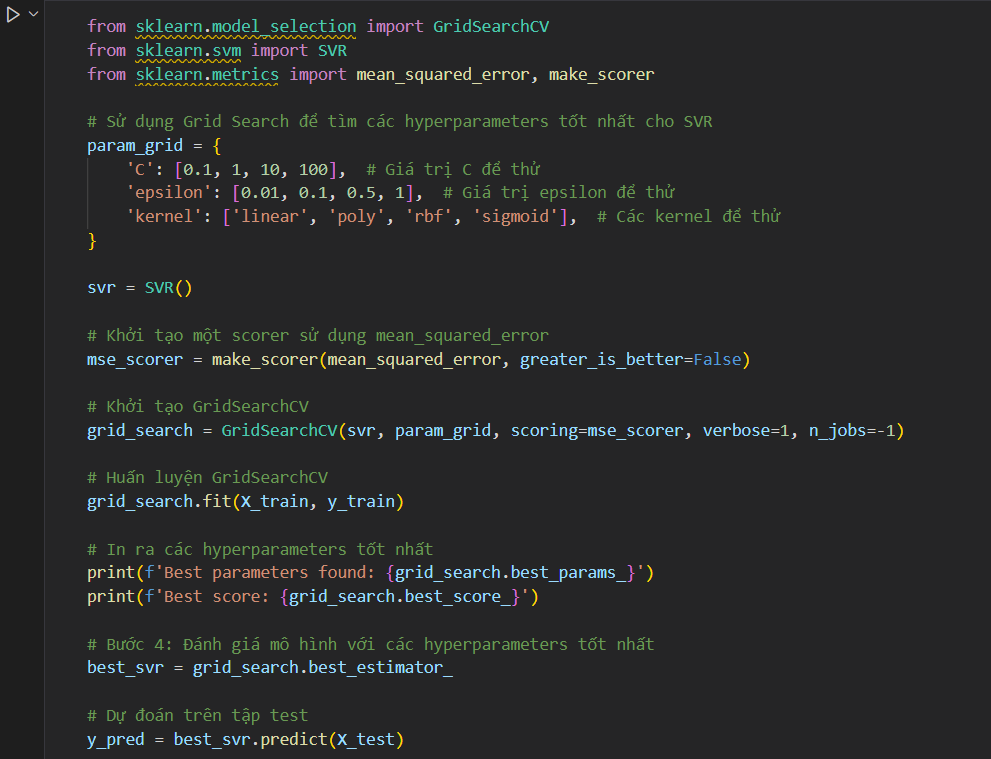
The optimization ran 15 iterations, with degree values ​​from 1 to 5 tested.The optimal polynomial degree was found to be 5, indicating that the model has the best performance when the degree of the polynomial is the highest within the allowable range.

Results:

MSE: 3724.9235, reflecting the model's error level, is still relatively high, showing that although optimized, the model still has room for improvement.

The R² score: 0.9597, shows that the model explains about 96% of the variation in the data, which is a good result, reflecting the model has good predictive ability.

Support Vector Regressor combined with Grid Search method



The Support Vector Regression (SVR) model is a powerful regression method, often used in problems with complex and high-dimensional data. This report details the process of finding optimal hyperparameters for the SVR model using the Grid Search method, an effective tool for systematically exploring the parameter space.

Search Results

Best Parameter:

C: 10 epsilon: 0.1 kernel: 'rbf'

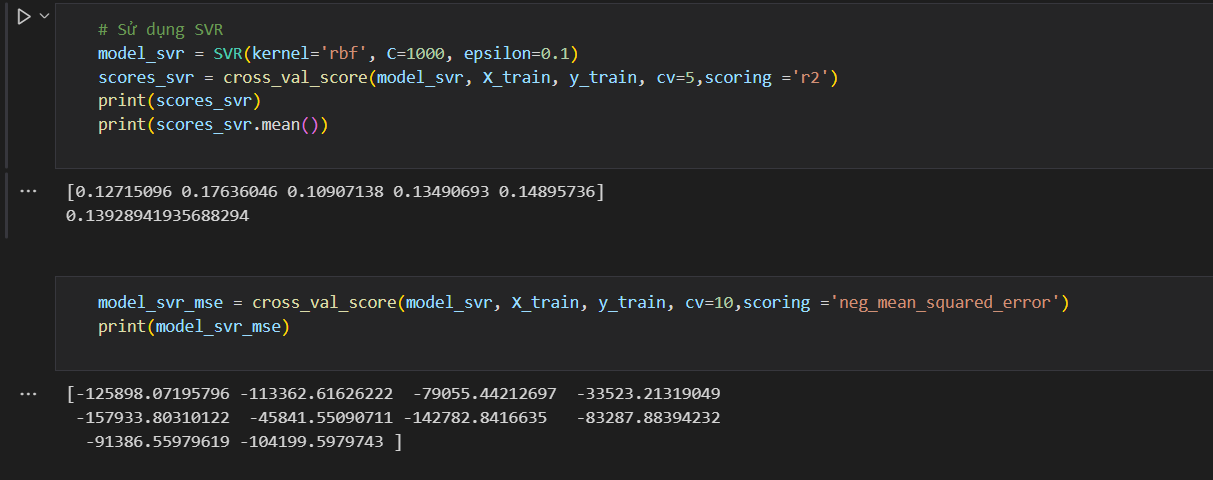
Best score: The best MSE score found via Grid Search is -320 (negative value due to make\_scorer scoring).

Model Evaluation

Optimal Model: The SVR model with optimal parameters was retrained and evaluated on the test set.

Result:

Mean Square Error (MSE): 3724, evaluates the model's effectiveness in predicting compared to the actual value.

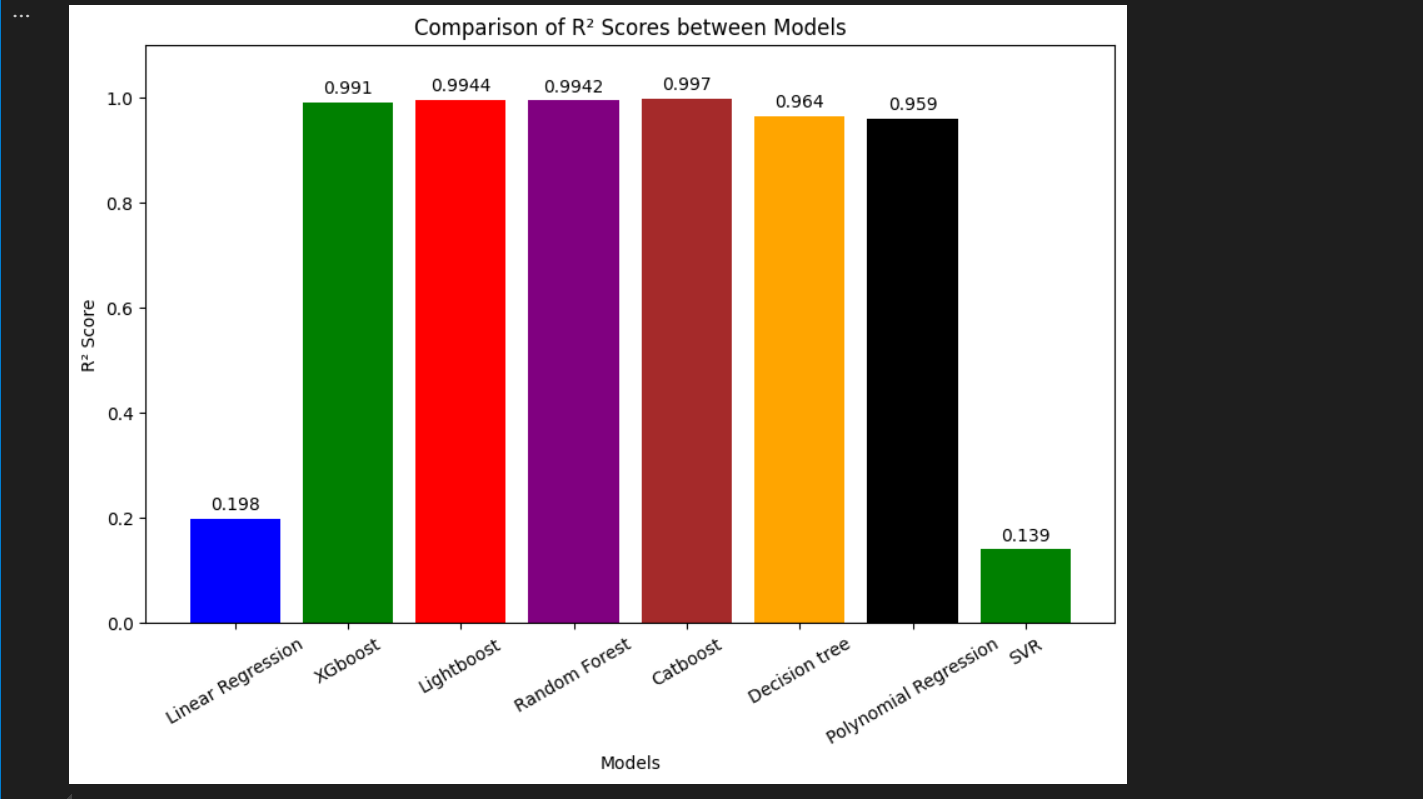
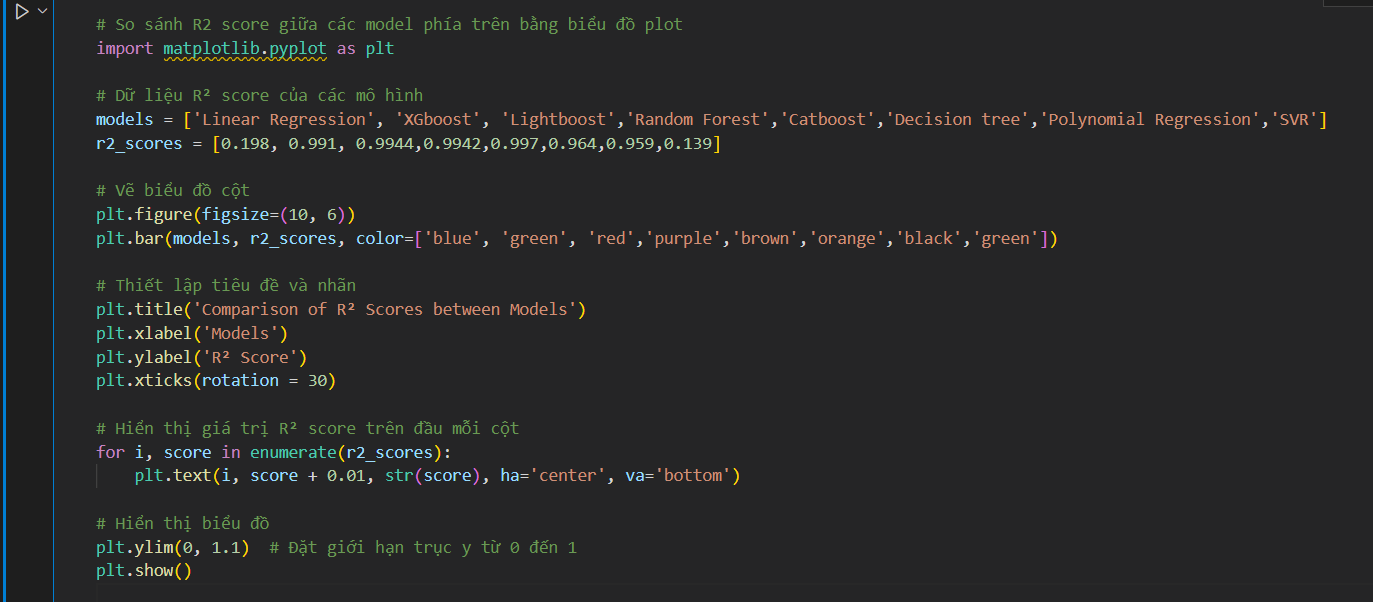


This report provides an overview of evaluating SVR (Support Vector Regression) models using cross-validation. The model was configured with the parameters selected for optimization, including the rbf kernel, the calibration value C=1000, and the epsilon value=0.1.

Result

R² Scores: The scores obtained through cross-validation show a fairly even distribution, with a mean value of about 0.1393. This score is quite low, showing that the model's ability to explain the variation in the data is not high.

MSE Scores: The MSE values ​​obtained from cross-validation are very large, with very high and negative values ​​due to the greater\_is\_better=False setting in the evaluation function. The average value of MSE is -4759.5066, showing that the model has high prediction error on the training set.



This report provides a detailed view of the performance comparison of different regression models based on R² index, through a bar chart. The goal is to evaluate and compare the ability of each model to explain variation in the data. We use the R² Index and Column Chart methods.

Result R² Scores:

* Linear Regression: 0.198
* XGBoost: 0.991
* Lightboost: 0.994
* Random Forest: 0.992
* Catboost: 0.997
* Decision Tree: 0.964
* Polynomial Regression: 0.959
* SVR: 0.139

Analysis

XGBoost, Lightboost, and Catboost achieved very high R² indexes of 0.991, 0.9944, and 0.997, respectively, showing that these models are very effective in explaining the variation in the data.

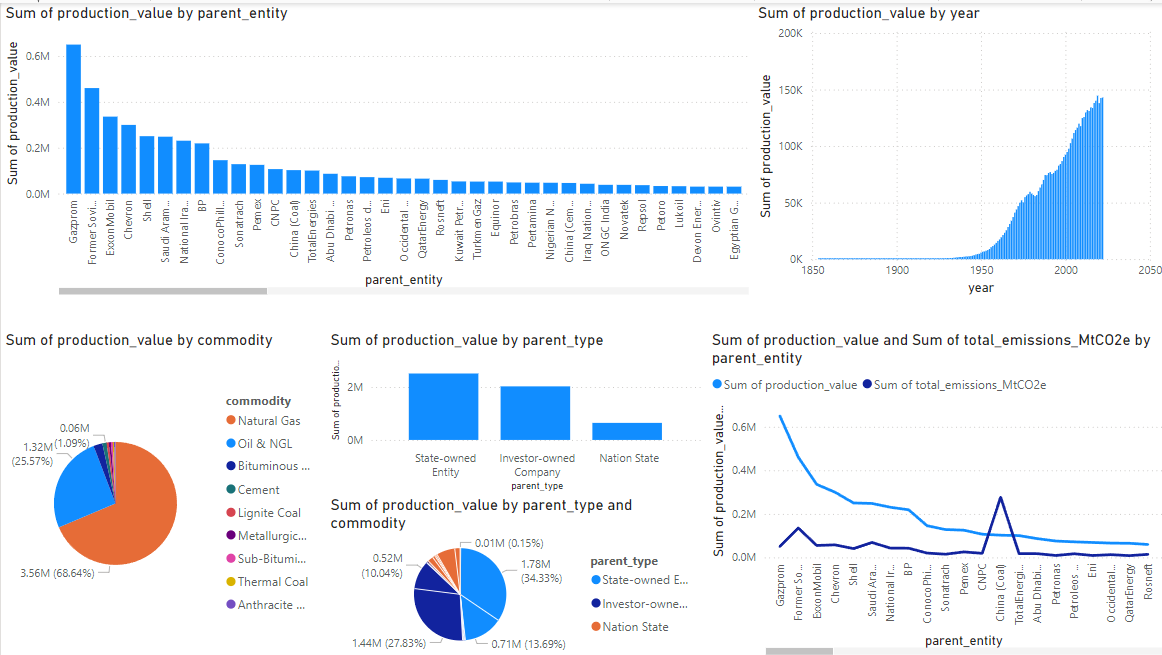
Random Forest and Decision Tree also showed good performance with R² index of 0.9942 and 0.964 respectively.

Polynomial Regression has an R² index of 0.959, indicating decent performance.

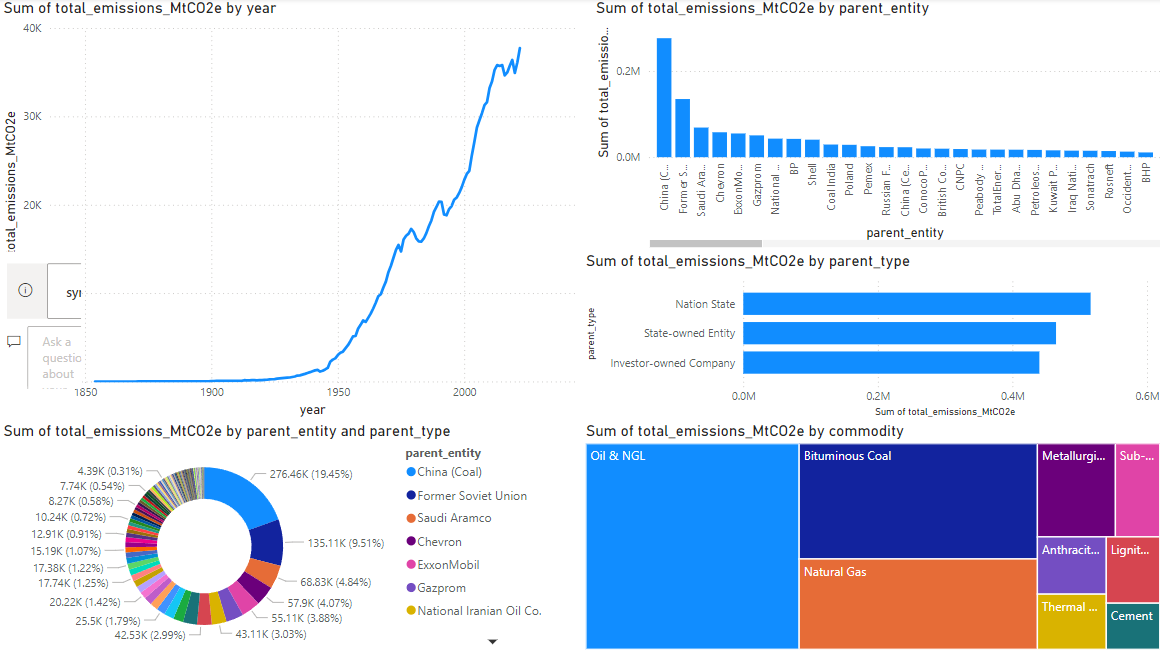
SVR and Linear Regression showed the lowest R² of the group, 0.139 and 0.198, respectively, indicating that these models were not as effective in predicting or explaining the data compared to other models.

=> Decision tree-based models such as XGBoost, Lightboost, Random Forest, and Catboost show excellent performance, possibly due to their ability to handle complex and nonlinear data.SVR and Linear Regression may not be suitable for a particular dataset or require reconfiguration of parameters to improve performance.

**VI.Power BI**

****

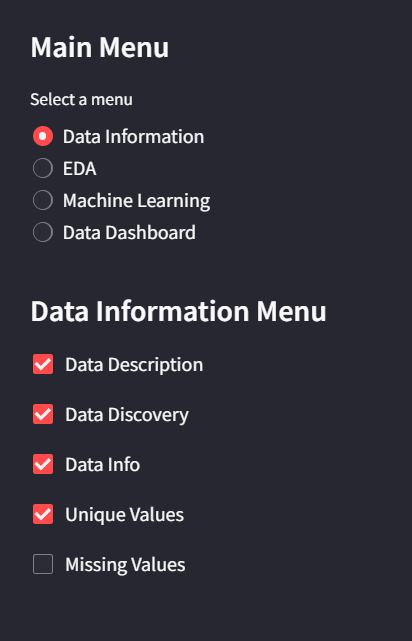
The presented chart provides a comprehensive view of the production value and total CO2 emissions of different production entities along with analysis by commodity and entity type over the years.



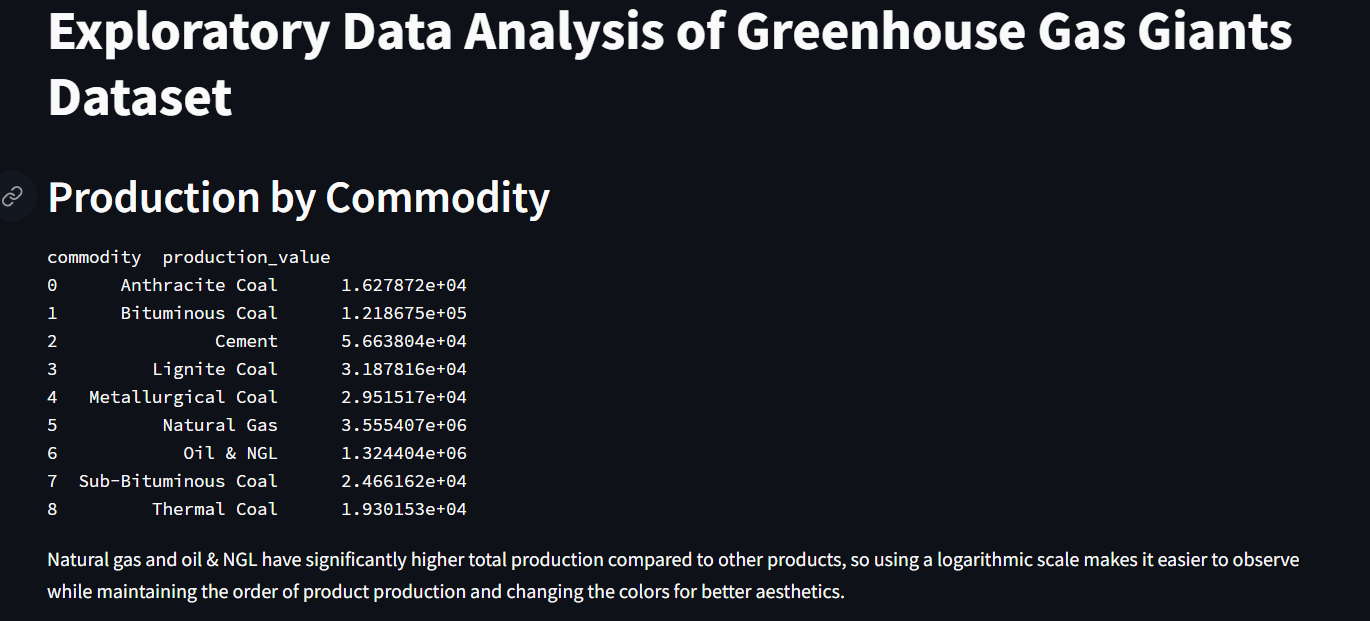
The chart below provides a detailed view of total CO2 emissions (in MtCO2e) over the years, broken down by subject and commodity type. This analysis helps understand the origin and distribution of emissions, thereby providing appropriate mitigation solutions.

**VII. App**

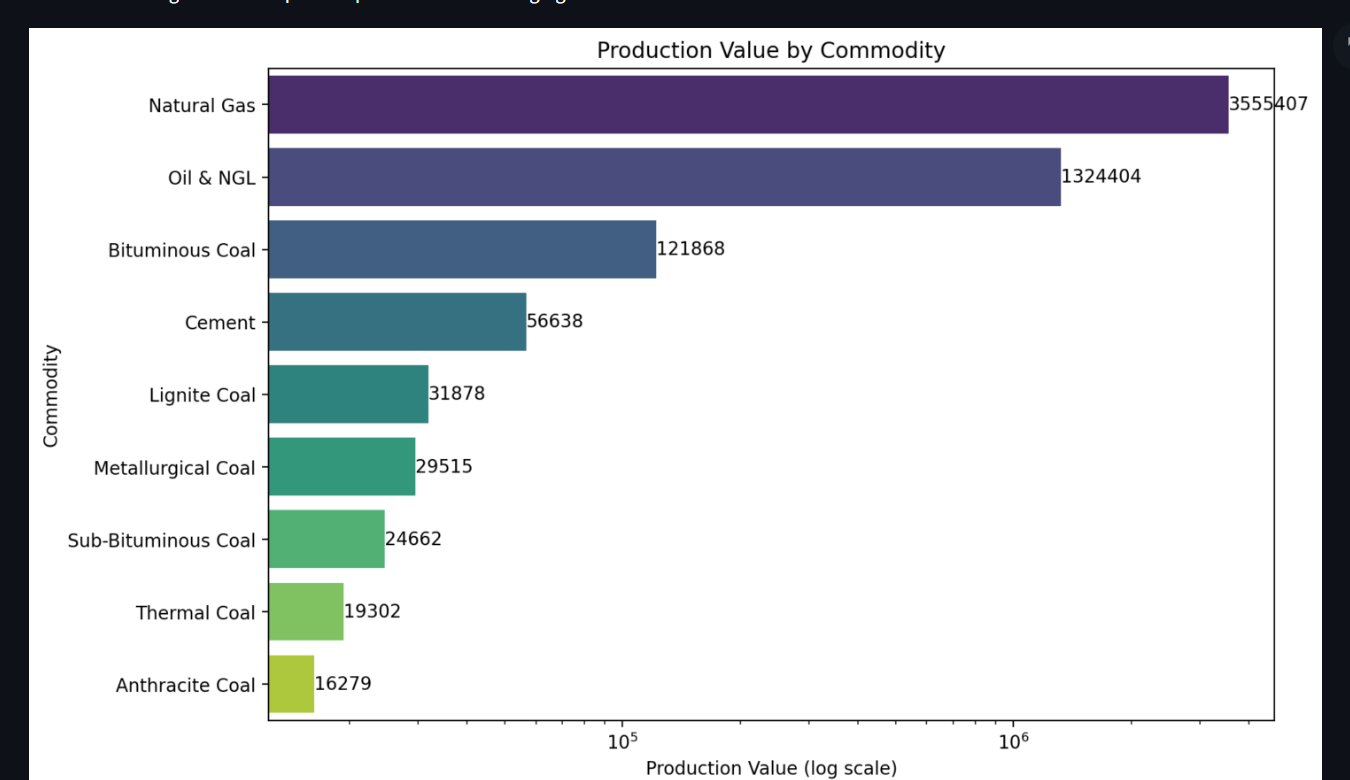
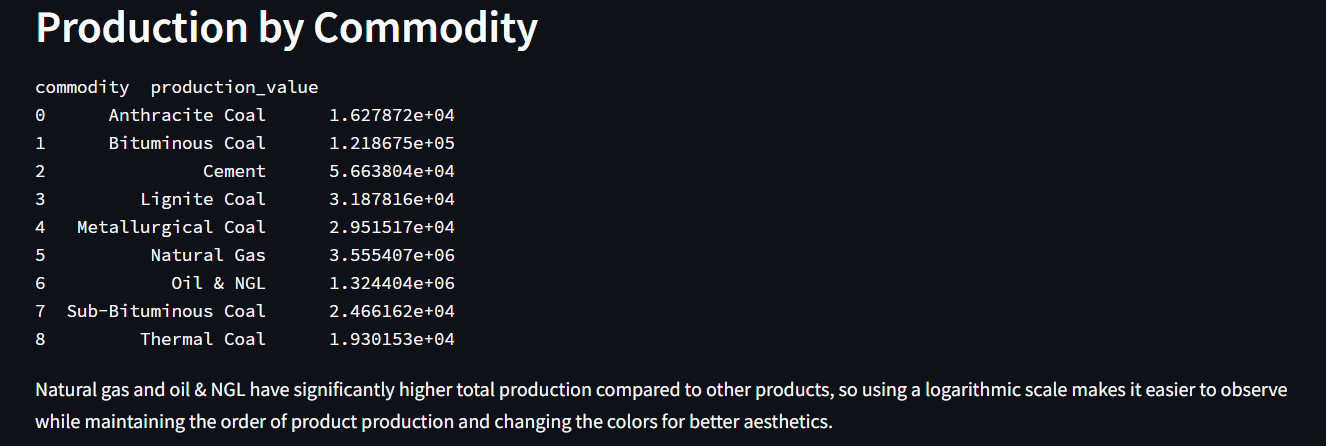
Functional interface



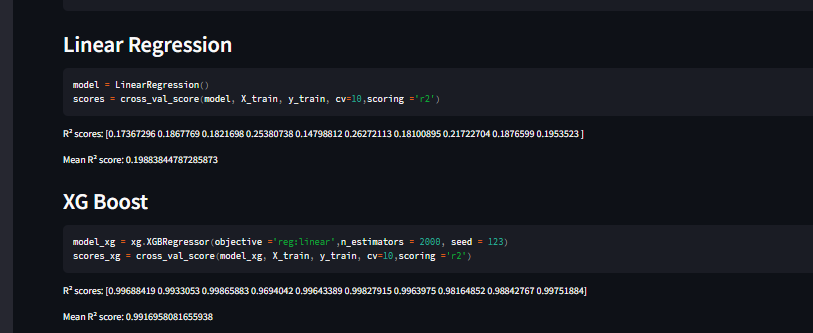
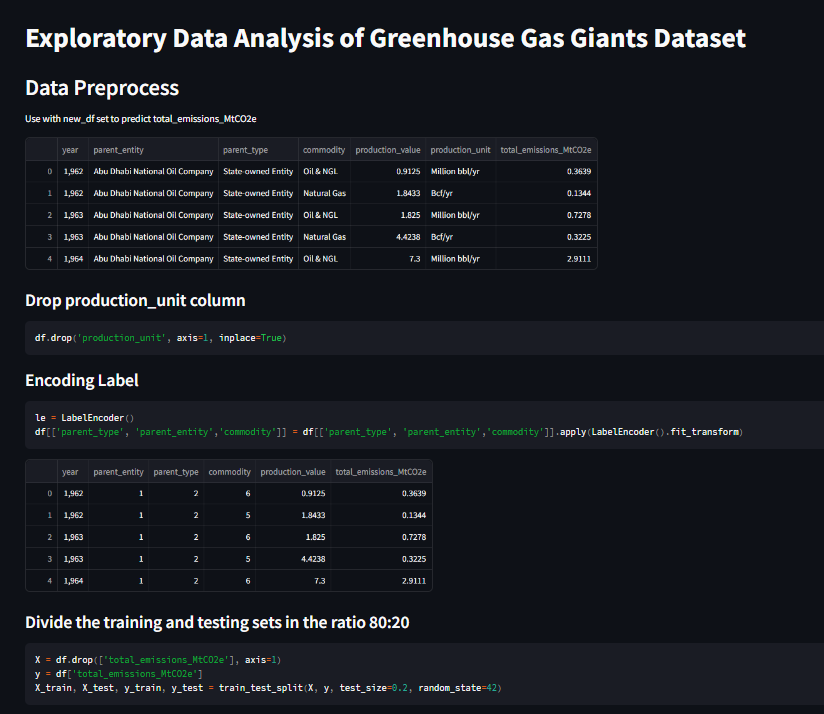
DATA INFORMATION



EDA



MACHINE LEARNING



DATA DASHBOARD

