Ingoude Company

DATA MINING

FINAL PROJECT

TABLE OF CONTENT

Medium granularity

column location	column name	description	unit	required	type	column location
emissions_medium_granularity.csv	year	The year of the data point	-	yes	integer	emissions_medium_granularity.csv
emissions_medium_granularity.csv	parent_entity	The entity to whom the emissions are traced to	-	yes	string	emissions_medium_granularity.csv
emissions_medium_granularity.csv	parent_type	The type of the parent_entity. Can be one of: investor-owned company, state-owned entity, nation state.	-	yes	string	emissions_medium_granularity.csv
emissions_medium_granularity.csv	commodity	Specifies which commodity the production refers to: Oil & NGL, Natural Gas, Anthracite Coal, Bituminous Coal, Lignite Coal, Metallurgical Coal, Sub-Bituminous Coal, Thermal Coal, or Cement.	-	yes	string	emissions_medium_granularity.csv
emissions_medium_granularity.csv	production_value	The quantity of production	-	yes	float	emissions_medium_granularity.csv
emissions_medium_granularity.csv	production_unit	The unit of production (Oil & NGL - million barrels, Natural Gas - billion cubic feet, Coal - million tonnes, Cement - million tonnes CO2 (see methodology for explanation))	Billion cubic feet per year (Bcf/yr), Million barrels per year (Million bbl/yr), or Million tonnes per year	yes	string	emissions_medium_granularity.csv

Greenhouse gas giants

Historical production data from the largest fossil fuel producers

Carbon Majors Data has the following features:

- Open Source: The data is available for download as CSV files for non-commercial use. InfluenceMap's Terms and Conditions apply.
- Annual Updates: The data is updated annually in November, and the downloads represent the latest available data.

Levels of Data Granularity:

1. **Low Granularity**: Includes year, entity, entity type, and total emissions.
2. **Medium Granularity**: Includes year, entity, entity type, commodity, commodity production, commodity unit, and total emissions.
3. **High Granularity**: Includes the same fields as the medium granularity file, as well as the reporting entity, data point source, product emissions, and four different operational emissions: flaring, venting, own fuel use, and fugitive methane.

	year	parent_entity	parent_type	commodity	production_value	production_unit	total_emissions_MtCO2e
0	1962	Abu Dhabi National Oil Company	State-owned Entity	Oil & NGL	0.91250	Million bbl/yr	0.363885
1	1962	Abu Dhabi National Oil Company	State-owned Entity	Natural Gas	1.84325	Bcf/yr	0.134355
2	1963	Abu Dhabi National Oil Company	State-owned Entity	Oil & NGL	1.82500	Million bbl/yr	0.727770
3	1963	Abu Dhabi National Oil Company	State-owned Entity	Natural Gas	4.42380	Bcf/yr	0.322453
4	1964	Abu Dhabi National Oil Company	State-owned Entity	Oil & NGL	7.30000	Million bbl/yr	2.911079

Table 1. First five rows in dataset

The dataset has **12551 examples** and **7 features,** with **no null values.**

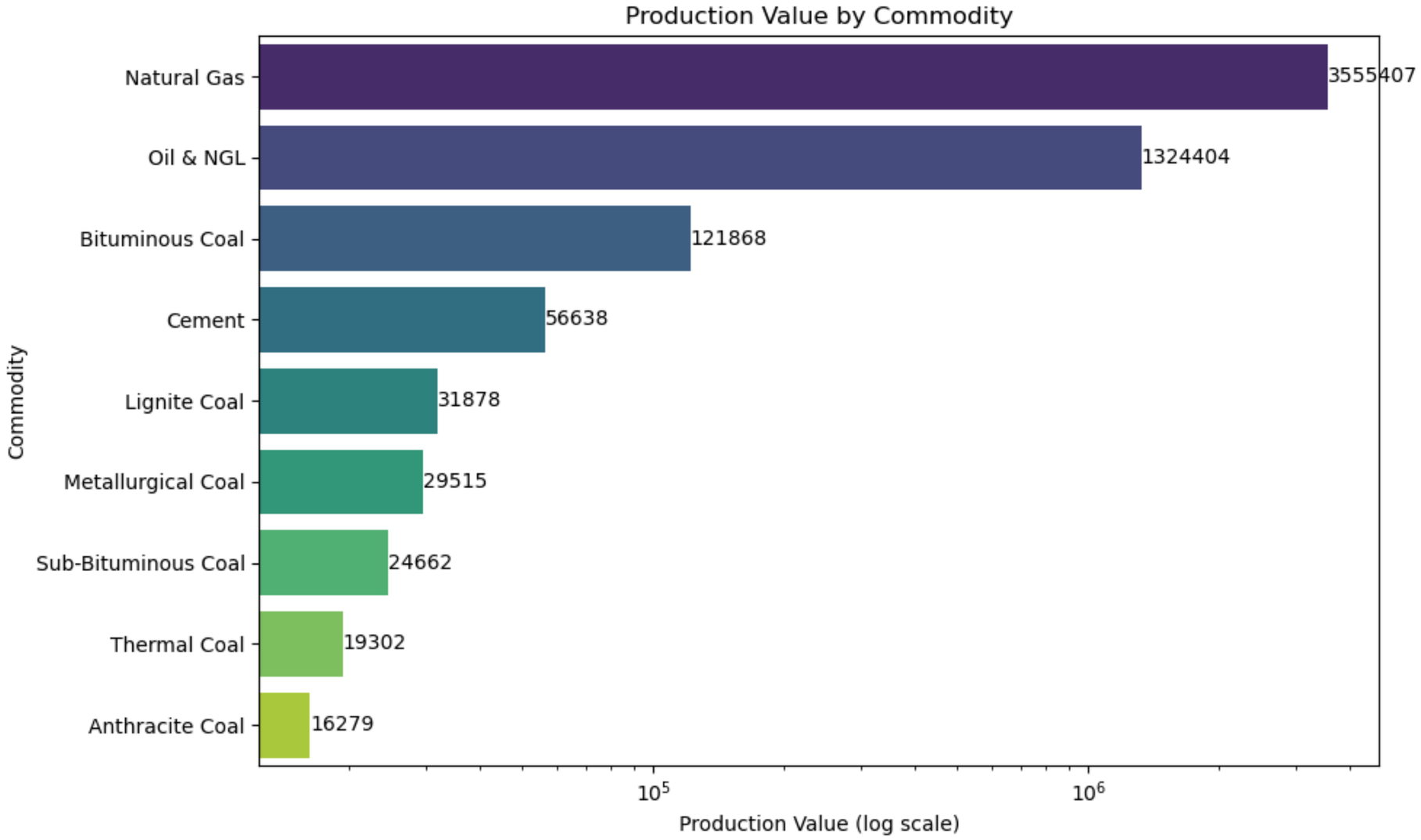
Explain features:

- **Year**: The year of the data point
- **Parent_entity:** The entity to which the emissions are traced (fuel distributor, fuel company,...)
- **Parent_type**: Investor-owned company, state-owned entity, nation-state.
- **Commodity**
- **Production_value**: The quantity of production
- **Production_unit**: The unit of production
- **Total_emissions_MtCO2e**: The total emissions

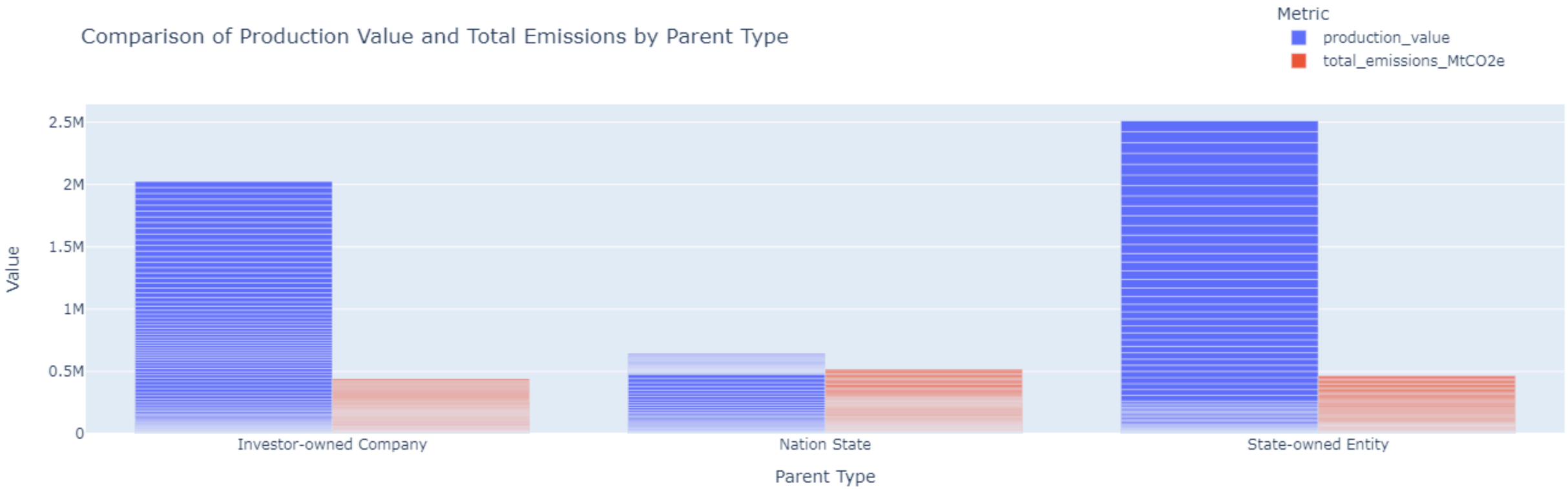
=> 169 years, 122 companies with 3 different types, 9 commodities and 4 production units

- **Commodity**: 'Oil & NGL' 'Natural Gas' 'Sub-Bituminous Coal' 'Metallurgical Coal' 'Bituminous Coal' 'Thermal Coal' 'Anthracite Coal' 'Cement' 'Lignite Coal'
- **Parent_type**: 'State-owned Entity', 'Investor-owned Company', 'Nation State

The quantity of Production based on Commodity

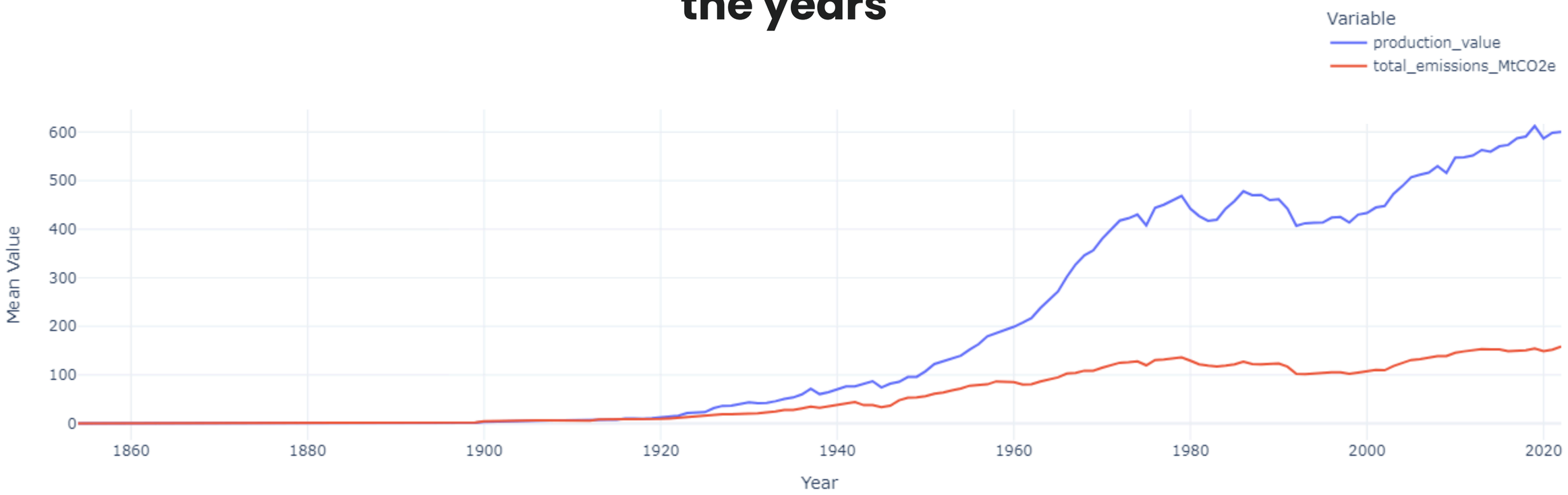


Total Production Value by Year for Each Parent Entity



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

Avg. global Fossil Fuels and Mineral production & CO2 emissions over the years



=> The total production of Fossil Fuels and Mineral 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 increases rapidly over time, while CO2 emissions increase but at a much slower pace. There could be several different reasons

Results of EDA

Natural Gas and Oil & NGL have had such high production values over time

- After World War II, many countries experienced unprecedented economic growth, characterized by the rise of automobile culture, suburbanization, and heavy industrial activities, all heavily dependent on oil and natural gas.
- Natural gas burns cleaner than coal and oil, producing less CO2. This makes it a more environmentally friendly option in the global transition towards cleaner energy.

Several factors influence the annual variation in total production value for each parent entity:

- Historical events like the Soviet industrial policy's rise in oil production and the Soviet Union's collapse.
- Fluctuations in production values for entities like ExxonMobil, Shell, BP, and Chevron due to oil embargoes, wars (e.g., the Gulf War), and financial crises
- Advances in mining technology, such as fracking, driving increased output.

Fossil fuel and mineral production has increased significantly, but CO2 emissions have risen more slowly:

- Technological advancements and the adoption of cleaner energy have improved production efficiency and reduced the carbon footprint.
- Strict environmental regulations and the implementation of carbon capture technologies have helped curb the growth of CO2 emissions.

Post-EDA Applications

Identifying Production Values

- Determine which goods have the highest and lowest production values.
- Supports economic, investment, and resource management decisions.

Monitoring Production Fluctuations

- Monitor production fluctuations over the years.
- Identify trends and predict the future, providing information for economic development strategies.

Relationship Between Production and Pollution

- Understand the relationship between production and environmental pollution.
- Supports the development of policies to reduce emissions and protect the environment.

Machine Learning

Using different models to predict the **Total emission of CO2** of the parent entities.

Models:

- Linear Regression
- XGboost
- Lightboost
- Random Forest
- Catboost
- Decision tree
- Polynomial Regression
- SVR

Testing **3 methods** to find the best parameters for the model:

- Catboost with RandomizedSearchCV
- Polynomial Regression with Bayesian Optimization
- SVR with Grid Search

Using the **Cross Validation** method to avoid overfitting

Machine Learning

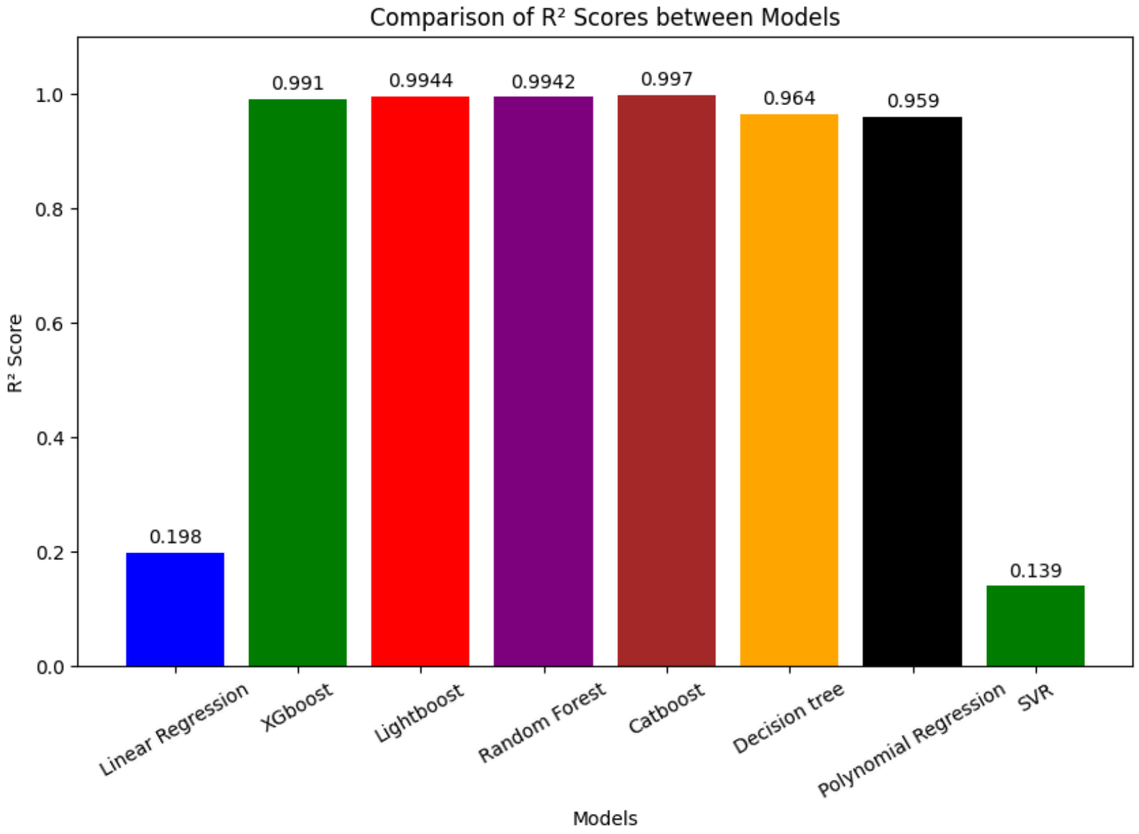
Data preprocessing

- Step 1: Handling missing values
- Step 2: Delete unnecessary column (production_unit)
- Step 3: Label Encoder with data columns of object type ('parent_type', 'parent_entity', 'commodity')
- Step 4: Determine Y, determine target X for prediction, divide the train and test set at a ratio of 80:20

Modeling

- Using cross-validation is 10 and metric is R2 Score and then computing the average r2 score.
- The average training time for models is **15-20 minutes**
- 2 search methods:
  + Bayesian Optimization has a run time of **35 minutes**
  + Randomized Search has a running time of **20 minutes**
- Grid Search has a running time of up to 6 hours but still has no results. So SVR try different optimizers and achieve an R2 score of only 0.11 -> 0.14.

Machine Learning



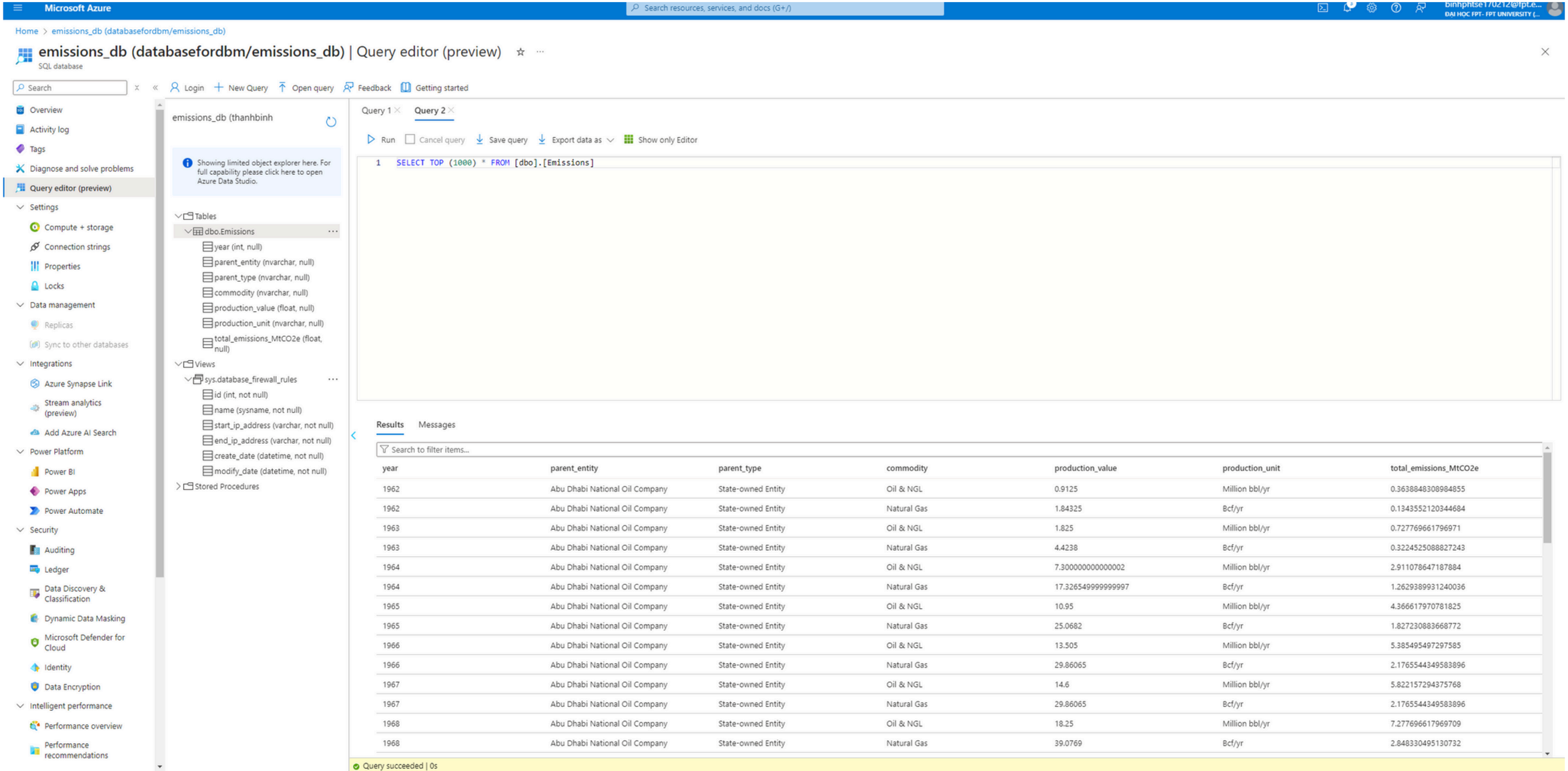
- R^2 = 1: Perfect model, accurately predicts all actual values.
- R^2 = 0: The model does not explain any variation in the actual data.
- R^2 < 0: The prediction model is worse than the average prediction model.

Evaluate

- XGBoost, Lightboost, Catboost, Random Forest, and Decision Tree, Poly achieved very high R^2 indexes of 0.991, 0.9944, and 0.997, 0.9942, 0.964, and 0.959 respectively, showing that these models effectively explain the variation in the data.
- SVR and Linear Regression showed the lowest R^2 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.

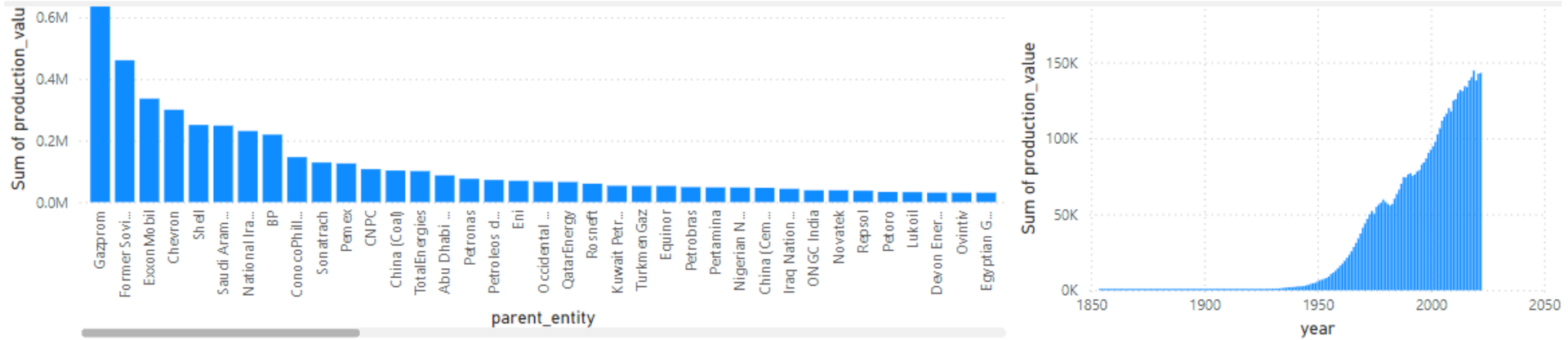
==> Decision tree-based models such as XGBoost, Lightboost, Random Forest, and Catboost show excellent performance, possibly due to their ability to handle complex and nonlinear data.
SVR and Linear Regression may not be suitable for a particular dataset or require reconfiguration of parameters to improve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

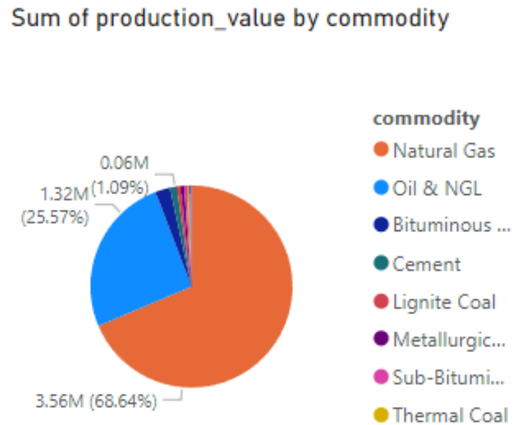
DATABASE

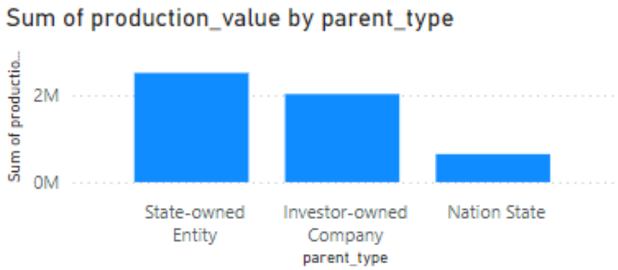

Microsoft Azure
Search resources, services, and docs (G+/)
Home > emissions_db (databasefordbm/emissions_db)
emissions_db (databasefordbm/emissions_db) | Query editor (preview)
SQL database
Search
Login
New Query
Open query
Feedback
Getting started
Overview
Activity log
Tags
Diagnose and solve problems
Query editor (preview)
Settings
Compute + storage
Connection strings
Properties
Locks
Data management
Replicas
Sync to other databases
Integrations
Azure Synapse Link
Stream analytics (preview)
Add Azure AI Search
Power Platform
Power BI
Power Apps
Power Automate
Security
Auditing
Ledger
Data Discovery & Classification
Dynamic Data Masking
Microsoft Defender for Cloud
Identity
Data Encryption
Intelligent performance
Performance overview
Performance recommendations
emissions_db (thanhbinh
Showing limited object explorer here. For full capability please click here to open Azure Data Studio.
Tables
dbo.Emissions
year (int, null)
parent_entity (nvarchar, null)
parent_type (nvarchar, null)
commodity (nvarchar, null)
production_value (float, null)
production_unit (nvarchar, null)
total_emissions_MtCO2e (float, null)
Views
sys.database_firewall_rules
id (int, not null)
name (sysname, not null)
start_ip_address (varchar, not null)
end_ip_address (varchar, not null)
create_date (datetime, not null)
modify_date (datetime, not null)
Stored Procedures
Query 1
Query 2
Run
Cancel query
Save query
Export data as
Show only Editor
1 SELECT TOP (1000) * FROM [dbo].[Emissions]
Results
Messages
Search to filter items...
year
parent_entity
parent_type
commodity
production_value
production_unit
total_emissions_MtCO2e
1962 Abu Dhabi National Oil Company State-owned Entity Oil & NGL 0.9125 Million bbl/yr 0.3638848308984855
1962 Abu Dhabi National Oil Company State-owned Entity Natural Gas 1.84325 Bcf/yr 0.1343552120344684
1963 Abu Dhabi National Oil Company State-owned Entity Oil & NGL 1.825 Million bbl/yr 0.727769661796971
1963 Abu Dhabi National Oil Company State-owned Entity Natural Gas 4.4238 Bcf/yr 0.3224525088827243
1964 Abu Dhabi National Oil Company State-owned Entity Oil & NGL 7.300000000000002 Million bbl/yr 2.911078647187884
1964 Abu Dhabi National Oil Company State-owned Entity Natural Gas 17.326549999999997 Bcf/yr 1.2629389931240036
1965 Abu Dhabi National Oil Company State-owned Entity Oil & NGL 10.95 Million bbl/yr 4.366617970781825
1965 Abu Dhabi National Oil Company State-owned Entity Natural Gas 25.0682 Bcf/yr 1.827230883668772
1966 Abu Dhabi National Oil Company State-owned Entity Oil & NGL 13.505 Million bbl/yr 5.385495497297585
1966 Abu Dhabi National Oil Company State-owned Entity Natural Gas 29.86065 Bcf/yr 2.1765544349583896
1967 Abu Dhabi National Oil Company State-owned Entity Oil & NGL 14.6 Million bbl/yr 5.822157294375768
1967 Abu Dhabi National Oil Company State-owned Entity Natural Gas 29.86065 Bcf/yr 2.1765544349583896
1968 Abu Dhabi National Oil Company State-owned Entity Oil & NGL 18.25 Million bbl/yr 7.277696617969709
1968 Abu Dhabi National Oil Company State-owned Entity Natural Gas 39.0769 Bcf/yr 2.848330495130732
Query succeeded | 0s

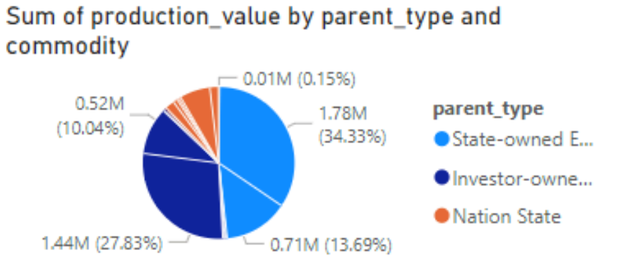
Deploy website (streamlit)

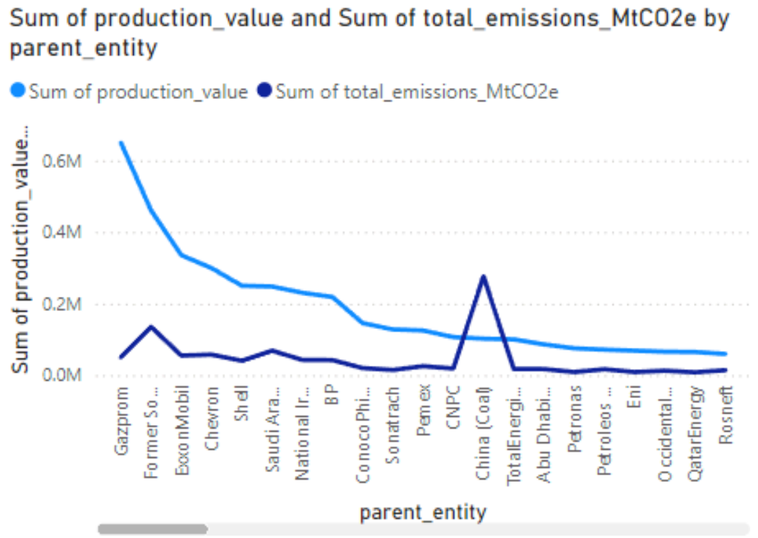
Power BI



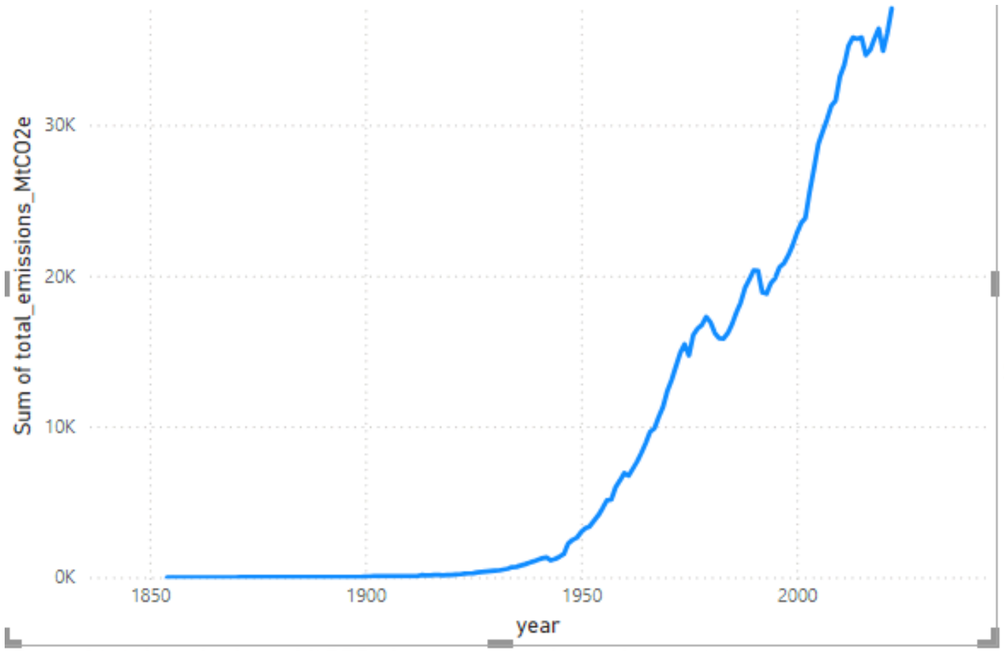


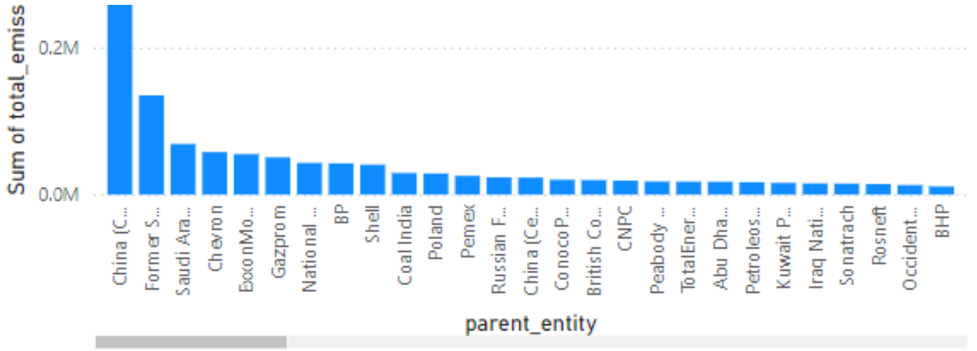


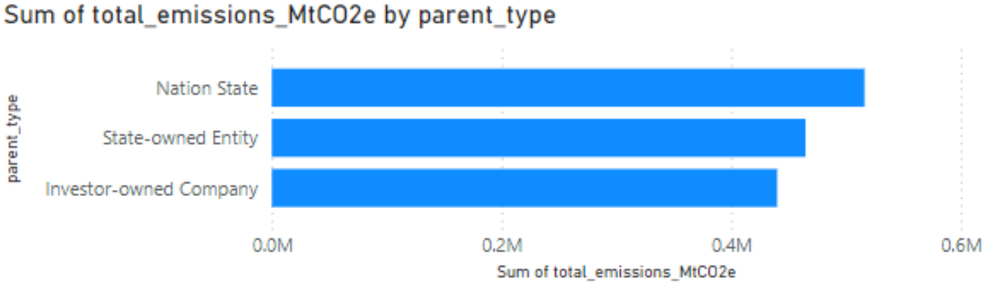


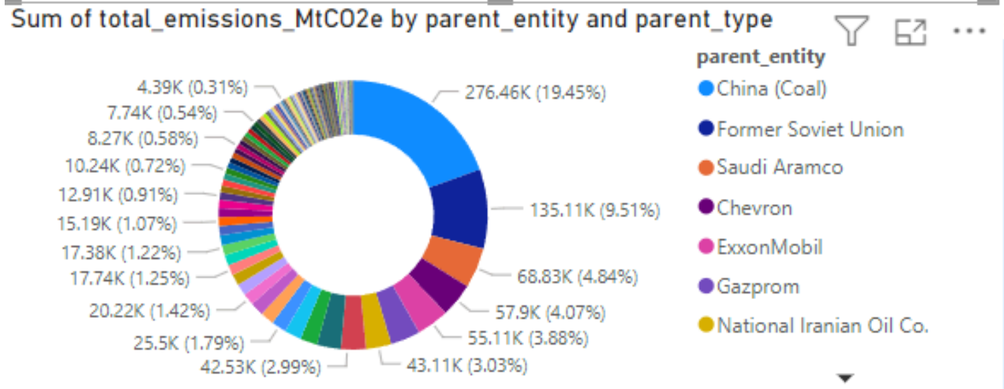


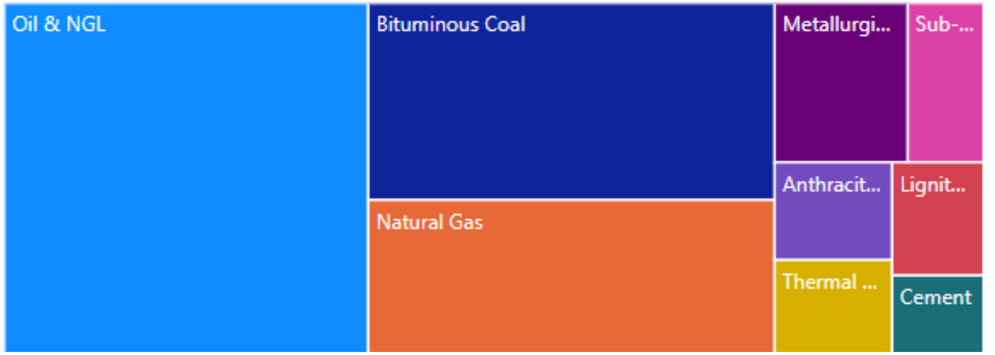
Report on Analysis of Production Value and CO2 Emissions by Type of Organization and Goods











Total CO2 Emission Analysis Report by Year, Organization Type and Commodity Type

REPORT

THANK YOU