Global CO_2 Emissions

1. Introduction

1.1 Problem Statement

The growth of CO2 emissions is a major contributing factor to the speed of climate change. Global carbon emissions from fossil fuels have significantly increased since 1900. Since 1970, CO2 emissions have increased by about 90%, with emissions from fossil fuel combustion and industrial processes contributing about 78% of the total greenhouse gas emissions increase from 1970 to 2011. Agriculture, deforestation, and other land-use changes have been the second-largest contributors. Emissions of non-CO2 greenhouse gases have also increased significantly since 1900. In order to reduce CO2 emissions, governments and industries must play an active role to curb energy consumption activities most related to emissions growth. In the current energy consumption landscape, a lot of emphases is also placed on the growing consumption of renewable energy sources (e.g., wind, solar). In order to prioritize which initiative has the highest impact on reducing CO2 emissions, governments and industries must be equipped with high-performing, predictive tools for future emissions.

1.2 Key Stakeholders

Potential parties that could be interested in this project include:

1) Politics and policies: ministries, departments, agencies, and directions of national governments;
2) Research and education: universities, institutes, research centers, laboratories;
3) Supply and demand: industrial companies related to energy, food, air, equipment manufacturing, etc.;
4) Organizations, societies, and influencers related to energy, environment, health, etc.

2. Data Preprocessing

2.1 Data Overview

Source data obtained for this project contains information on different kinds of greenhouse gas emissions, energy consumption, agriculture, and food production. The CO_2 and Greenhouse

Gas Emissions dataset is a collection of key metrics maintained by Our World in Data. It is updated regularly and includes data on CO2 emissions (annual, per capita, cumulative and consumption-based), other greenhouse gases, energy mix, and other relevant metrics of different countries from the year 1750 - 2019. The data set of agriculture and food production are sourced from UNDATA containing the information on agricultural land use and beef production of different countries from the year 1750 - 2019.

The features and corresponding information contained in the raw CO_2 emission data set is shown in the following figures:

```
co2_raw_data.columns
```

```
Index(['iso_code', 'country', 'year', 'annual_co2_prod_Megaton',
       'co2_growth_prct', 'co2_growth_abs', 'consumption_co2', 'trade_co2',
       'trade_co2_share', 'co2_per_capita', 'consumption_co2_per_capita',
       'share_global_co2', 'cumulative_co2', 'share_global_cumulative_co2',
       'co2_per_gdp', 'consumption_co2_per_gdp', 'co2_per_unit_energy',
       'cement_co2', 'coal_co2', 'flaring_co2', 'gas_co2', 'oil_co2',
       'other_industry_co2', 'cement_co2_per_capita', 'coal_co2_per_capita',
       'flaring_co2_per_capita', 'gas_co2_per_capita', 'oil_co2_per_capita',
       'other_co2_per_capita', 'share_global_coal_co2', 'share_global_oil_co2',
       'share_global_gas_co2', 'share_global_flaring_co2',
       'share_global_cement_co2', 'cumulative_coal_co2', 'cumulative_oil_co2',
       'cumulative_gas_co2', 'cumulative_flaring_co2', 'cumulative_cement_co2',
       'share_global_cumulative_coal_co2', 'share_global_cumulative_oil_co2',
       'share_global_cumulative_gas_co2',
       'share_global_cumulative_flaring_co2',
       'share_global_cumulative_cement_co2', 'total_ghg', 'ghg_per_capita',
       'methane', 'methane_per_capita', 'nitrous_oxide',
       'nitrous_oxide_per_capita', 'primary_energy_consumption_10Gwh',
       'energy_per_capita', 'energy_per_gdp', 'population', 'gdp'],
      dtype='object')
```

```
co2_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23708 entries, 0 to 23707
Data columns (total 7 columns):
 #   Column                            Non-Null Count  Dtype  
---  ------                            --------------  -----  
 0   iso_code                          20930 non-null  object 
 1   country                           23708 non-null  object 
 2   year                              23708 non-null  int64  
 3   annual_co2_prod_Megaton           23170 non-null  float64
 4   primary_energy_consumption_10Gwh  6044 non-null   float64
 5   population                        21071 non-null  float64
 6   gdp                               13002 non-null  float64
dtypes: float64(4), int64(1), object(2)
memory usage: 1.3+ MB
```

The features and corresponding information contained in the raw agricultural land use data set is shown in the following figures:

```
agri_land_raw_data.columns
```

```
Index(['Country or Area', 'Element', 'Year', 'Unit', 'Value_agri_1000hectare',
       'Value Footnotes'],
      dtype='object')
```

```
agri_land_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14378 entries, 0 to 14377
Data columns (total 4 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Country or Area         14377 non-null  object
 1   Year                    14369 non-null  float64
 2   Unit                    14369 non-null  object
 3   Value_agri_1000hectare  14369 non-null  float64
dtypes: float64(2), object(2)
memory usage: 449.4+ KB
```

The features and corresponding information contained in the raw beef production data set is shown in the following figures:

```
beef_prod_raw_data.columns
```

```
Index(['Country or Area', 'Element', 'Year', 'Unit', 'Value_beef_tonnes',
       'Value Footnotes'],
      dtype='object')
```

```
beef_prod_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13197 entries, 0 to 13196
Data columns (total 4 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Country or Area    13196 non-null  object
 1   Year               13194 non-null  float64
 2   Unit               13194 non-null  object
 3   Value_beef_tonnes  13194 non-null  float64
dtypes: float64(2), object(2)
memory usage: 412.5+ KB
```

As shown in the above figures, two important considerations can be proposed and need to be handled using the data cleaning method before building machine learning models upon that:

1) The CO_2 data set contains excessive features (columns). Which ones are important key features? And which one is the target feature?
2) It seems many data are missing. How to deal with the missing data?

2.2 Data processing

In the last section, two important considerations are proposed and need to be addressed.

Firstly, the CO_2 data set includes CO2 emissions by annual, per capita, cumulative, and consumption-based, and other greenhouse gases, energy mix, and other relevant metrics of different countries from the year 1750 - 2019. The objective of this project is to use machine learning methods to predict annual CO_2 production ("annual_co2_prod_Megaton"), which is the target feature. The features of primary energy consumption, population, GDP contained in this dataset are relevant and crucial for predicting CO_2 emissions. Accordingly, by joining the data sets of CO_2 emissions, agricultural land use, and beef production, the new CO_2 emission data set are shown in the following figures:

```
co2_data.columns
```

```
Index(['iso_code', 'country', 'year', 'annual_co2_prod_Megaton',
       'primary_energy_consumption_10Gwh', 'population', 'gdp', 'Unit_argi',
       'Value_agri_1000hectare', 'Unit_beef', 'Value_beef_tonnes',
       'energy_isnan', 'gdp_isnan', 'population_isnan', 'argi_isnan',
       'beef_isnan'],
      dtype='object')
```

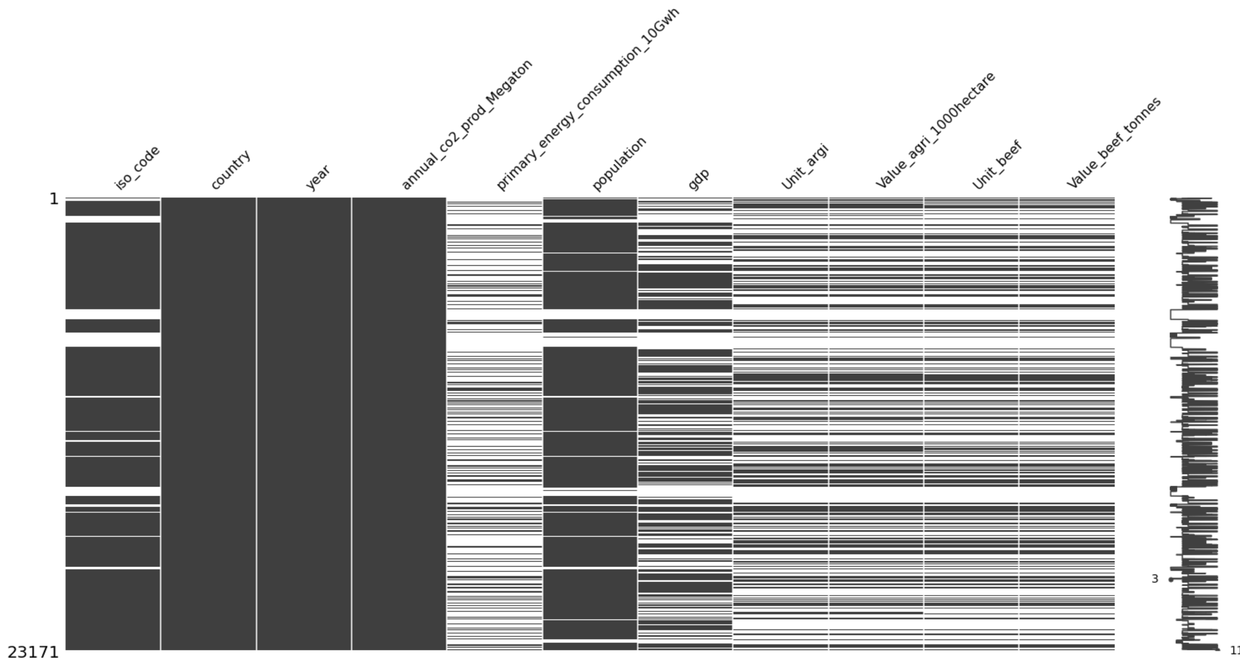
```
co2_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 23171 entries, 0 to 23708
Data columns (total 16 columns):
iso_code                            20440 non-null object
country                             23171 non-null object
year                                23171 non-null datetime64[ns]
annual_co2_prod_Megaton             23171 non-null float64
primary_energy_consumption_10Gwh    6045 non-null float64
population                          20583 non-null float64
gdp                                 12973 non-null float64
Unit_argi                           9818 non-null object
Value_agri_1000hectare              9818 non-null float64
Unit_beef                           9377 non-null object
Value_beef_tonnes                   9377 non-null float64
energy_isnan                        23171 non-null bool
gdp_isnan                           23171 non-null bool
population_isnan                    23171 non-null bool
argi_isnan                          23171 non-null bool
beef_isnan                          23171 non-null bool
dtypes: bool(5), datetime64[ns](1), float64(6), object(4)
memory usage: 2.2+ MB
```

```
co2_data.describe()
```

	annual_co2_prod_Megaton	primary_energy_consumption_10Gwh	population	gdp	Value_agri_1000hectare	Value_beef_tonnes
count	23171.000000	6045.000000	2.058300e+04	1.297300e+04	9.818000e+03	9.377000e+03
mean	270.234760	1638.034068	6.053309e+07	4.405589e+11	7.341125e+04	7.789382e+05
std	1509.880287	9665.709679	3.773372e+08	3.670729e+12	3.935006e+05	4.531116e+06
min	-1.165000	0.208000	1.000000e+03	6.378000e+07	3.000000e-01	0.000000e+00
25%	0.546000	46.326000	1.433000e+06	8.911988e+09	3.340000e+02	3.233000e+03
50%	5.170000	148.688000	5.004000e+06	2.946853e+10	3.495500e+03	4.080000e+04
75%	44.785000	518.789000	1.632450e+07	1.220000e+11	2.052425e+04	1.705510e+05
max	36441.388000	153848.433000	7.713468e+09	1.065610e+14	4.882180e+06	7.160131e+07

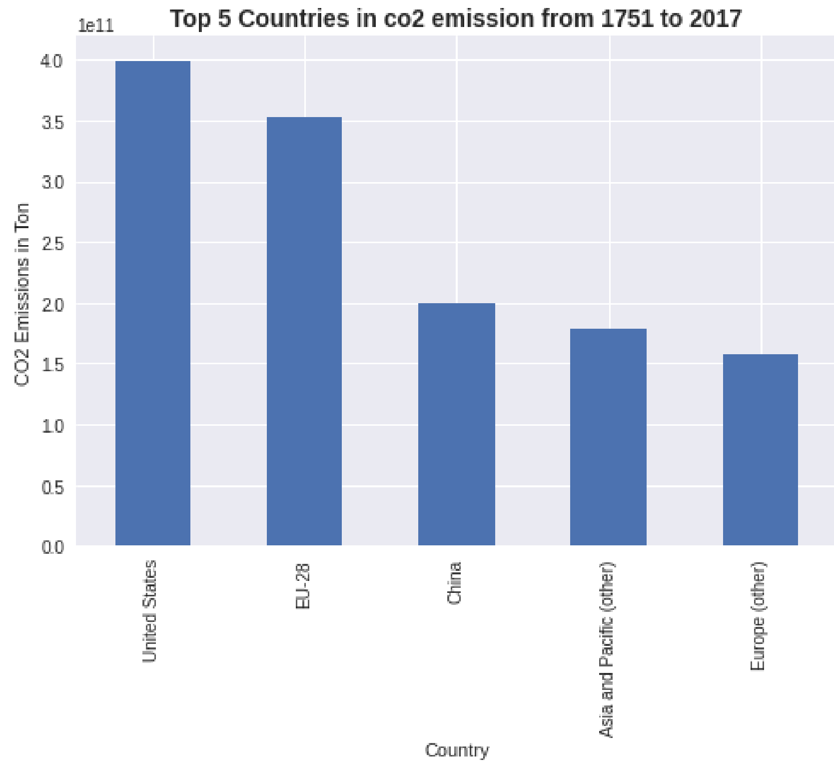
Secondarily, it seems there are a lot of missing values. To visualize the missing data, the package of "missingno" is imported and utilized. The results are shown in the following figure:



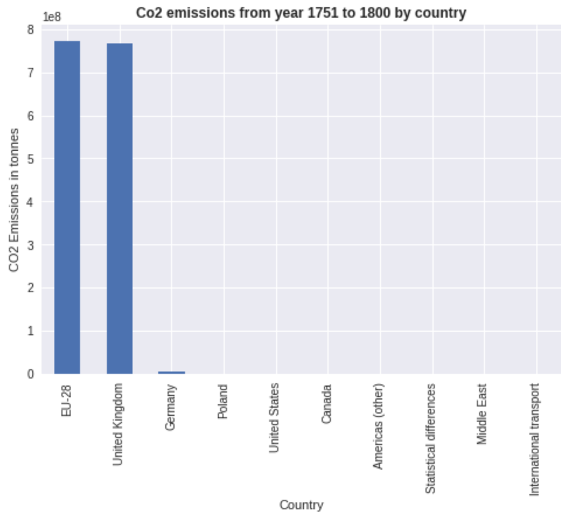
The results show that, compared with the columns of country, year, annual_co2_prod_Megaton, and population, there is a significant amount of data is missing in the rest columns. The missing data mostly belongs to the early time data of different countries due to the lack of recording intentions and techniques. The method to deal with NaN is elaborated in the following section.

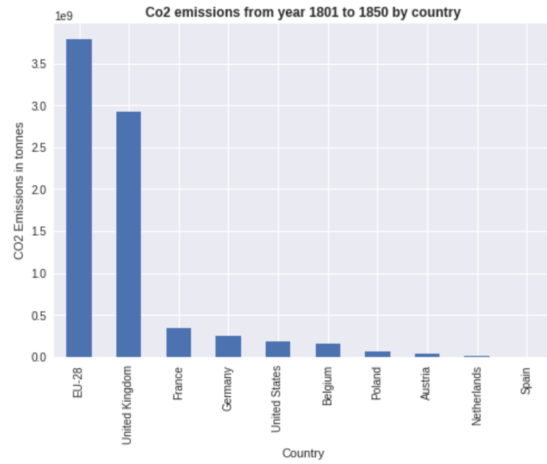
2.3 Data Story

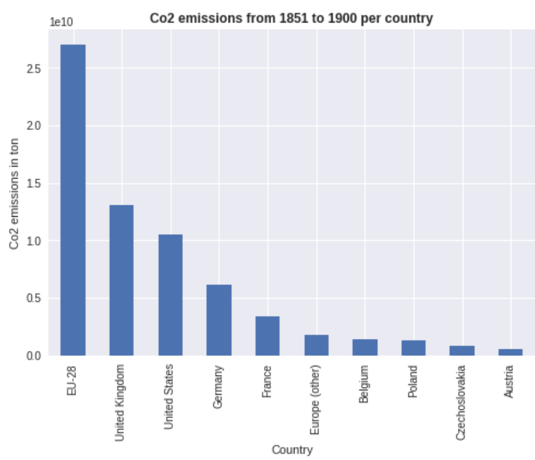
With the data preprocessing finished, exploratory data analysis (EDA) can be utilized for us to better understand the data. The figure below shows the top 5 countries in cumulative CO_2 emissions from 1751 to 2017. The results show the US has the highest cumulative CO_2 emission among all the countries and regions.

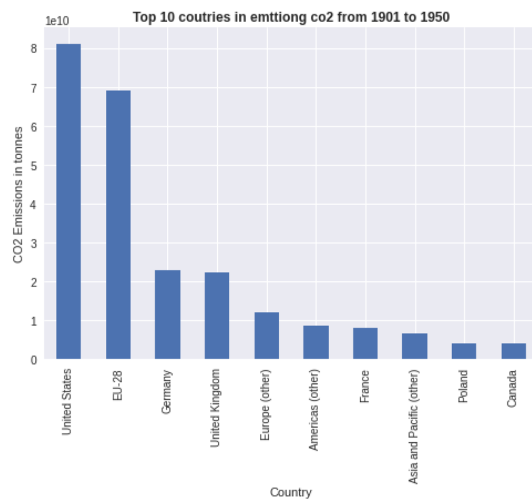


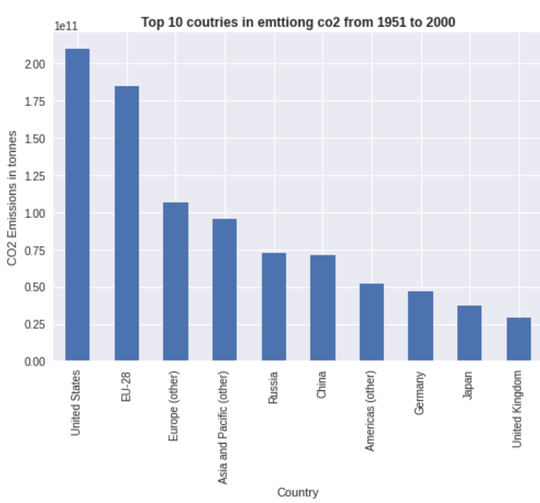
The figures below show the top 10 countries in cumulative CO_2 emissions within a 50-year period from 1751 to 2017. The result shows during the pre-industrial stage (1750 - 1850), only the UK had significant CO_2 emissions. With the start of the industrial revolution (the 1850s), the CO_2 emission of the US and Germany increased rapidly and exceeded the UK during 1901 – 1950. Starting from late 20th, China and India began their first industrial revolutions and appeared on the list of top 10 countries after 1951 and 2000, respectively, while others, such as the United States and western Europe, began undergoing “second” industrial revolutions by the late 19th century. In 21st, China exceeded the US and became the No. 1 CO_2 emission countries.

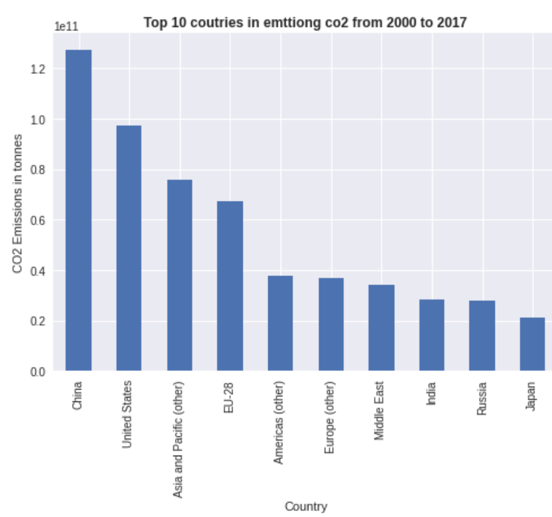

Co2 emissions from year 1751 to 1800 by country
1e8
CO2 Emissions in tonnes
EU-28
United Kingdom
Germany
Poland
United States
Canada
Americas (other)
Statistical differences
Middle East
International transport
Country


Co2 emissions from year 1801 to 1850 by country
1e9
CO2 Emissions in tonnes
EU-28
United Kingdom
France
Germany
United States
Belgium
Poland
Austria
Netherlands
Spain
Country

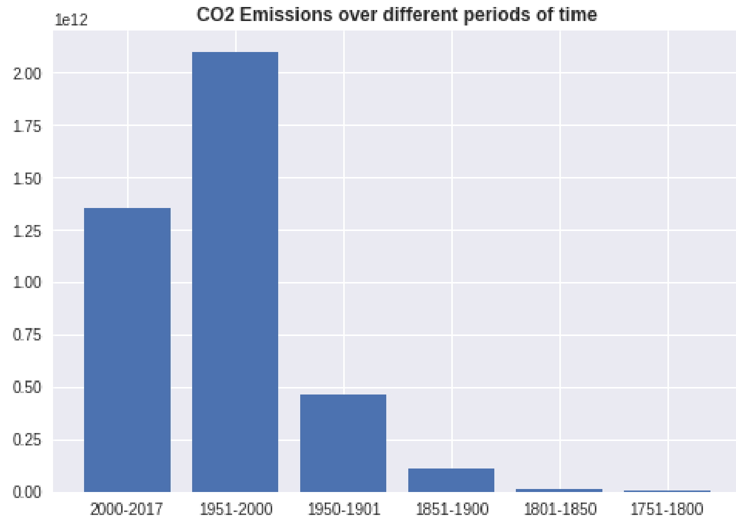

Co2 emissions from 1851 to 1900 per country
1e10
Co2 emissions in ton
EU-28
United Kingdom
United States
Germany
France
Europe (other)
Belgium
Poland
Czechoslovakia
Austria
Country


Top 10 coutries in emttiong co2 from 1901 to 1950
1e10
CO2 Emissions in tonnes
United States
EU-28
Germany
United Kingdom
Europe (other)
Americas (other)
France
Asia and Pacific (other)
Poland
Canada
Country


Top 10 coutries in emttiong co2 from 1951 to 2000
1e11
CO2 Emissions in tonnes
United States
EU-28
Europe (other)
Asia and Pacific (other)
Russia
China
Americas (other)
Germany
Japan
United Kingdom
Country


Top 10 coutries in emttiong co2 from 2000 to 2017
1e11
CO2 Emissions in tonnes
China
United States
Asia and Pacific (other)
EU-28
Americas (other)
Europe (other)
Middle East
India
Russia
Japan
Country

In addition, the figure below compares the total CO2 emissions within different periods from 1751 to 2017. We can observe the exponential increase of CO_2 emission with time. It is noted that the CO_2 emission of 2000 – 2017 has reached 60% of 1951-2000 within only 17 years.



2.4 Feature Engineering

3. Modeling

4. Results and Discussions

5 Conclusions