THE CURIOSITY CUP

A Global SAS® Student Competition

Top Contributors to Climate Crisis

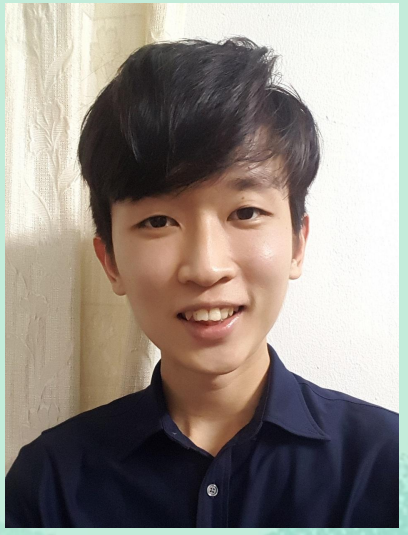
Team SASpicious 2 

MMU®
MULTIMEDIA UNIVERSITY

Team SASpicious 2

Ts. Dr. Ng Kok Why

Advisor



Liew Kuan Yung

Team Leader,
Data Scientist

Koh Han Yi

Data Analyst

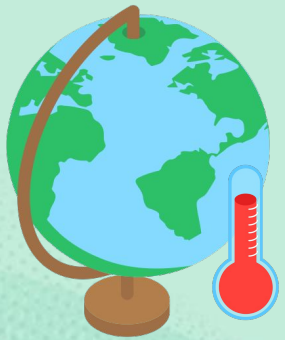
Soh Jing Guan

Data Scientist

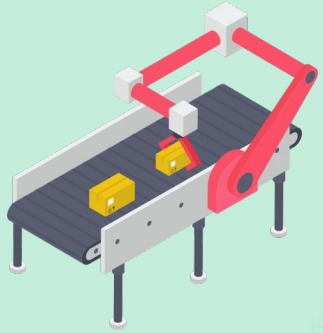
Chua Bing Quan

Data Engineer

Overview

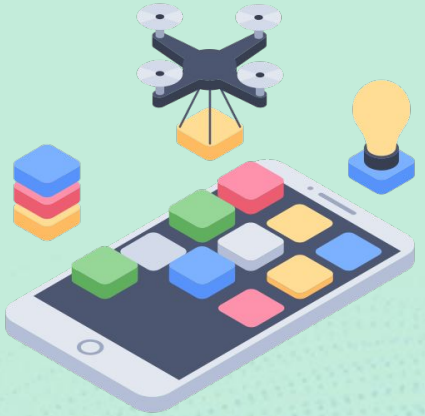


Background



Data Preprocessing

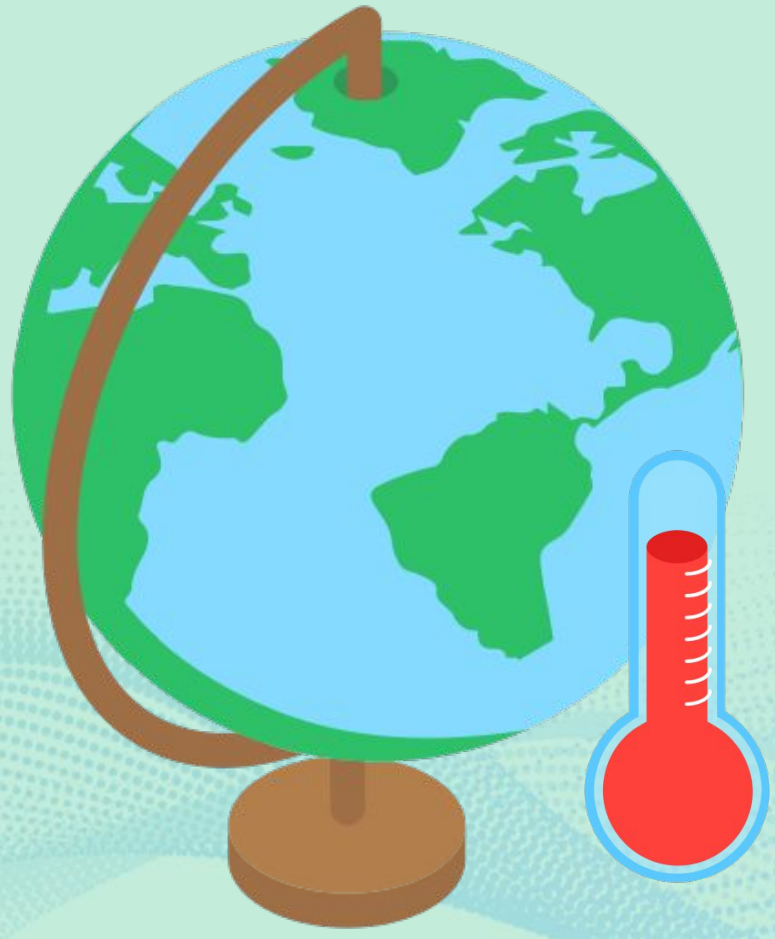
Data Analysis



Recommendation

01
Background

The current climate situation, motivations, and problem statements

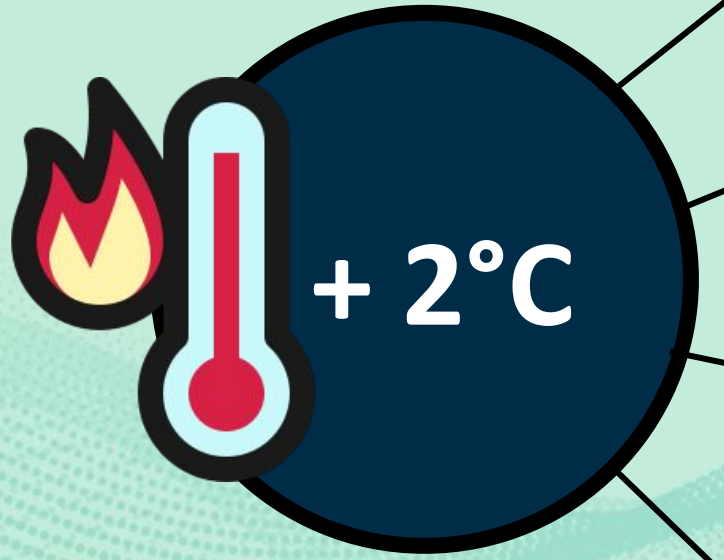


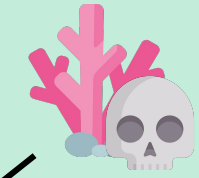
Background

United Nations Climate Change Conference in Glasgow (COP26)

- Most countries agreed climate change issues is serious and should be highly valued.
- 151 countries reaffirmed their commitment to limit the temperature rise below 2°C

What happens when the Earth increases 2°C?





All reefs will be lost

Habitat loss for all species would double or triple

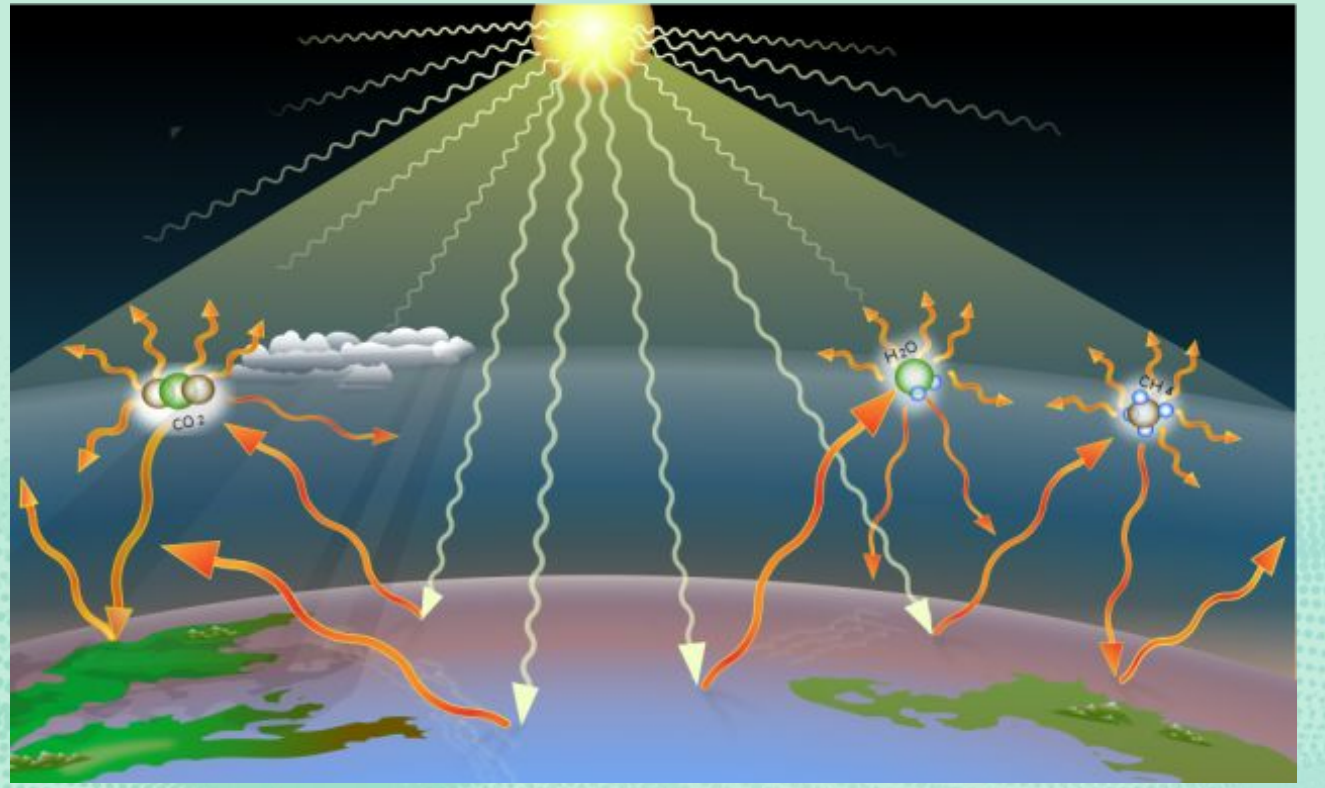
37% of the global population exposed to severe heat every 1 in 5 years

36% of land to extreme rainfall and cause average rainfall to rise 4%

Anomaly Temperature (1990 - 2019)

- Temperature anomaly is the difference from an average, or baseline temperature. The baseline temperature is computed by averaging 30 years of temperature data, here we are based on data between 1961-1990.
- The temperature anomaly has been increasing since the year 1990.

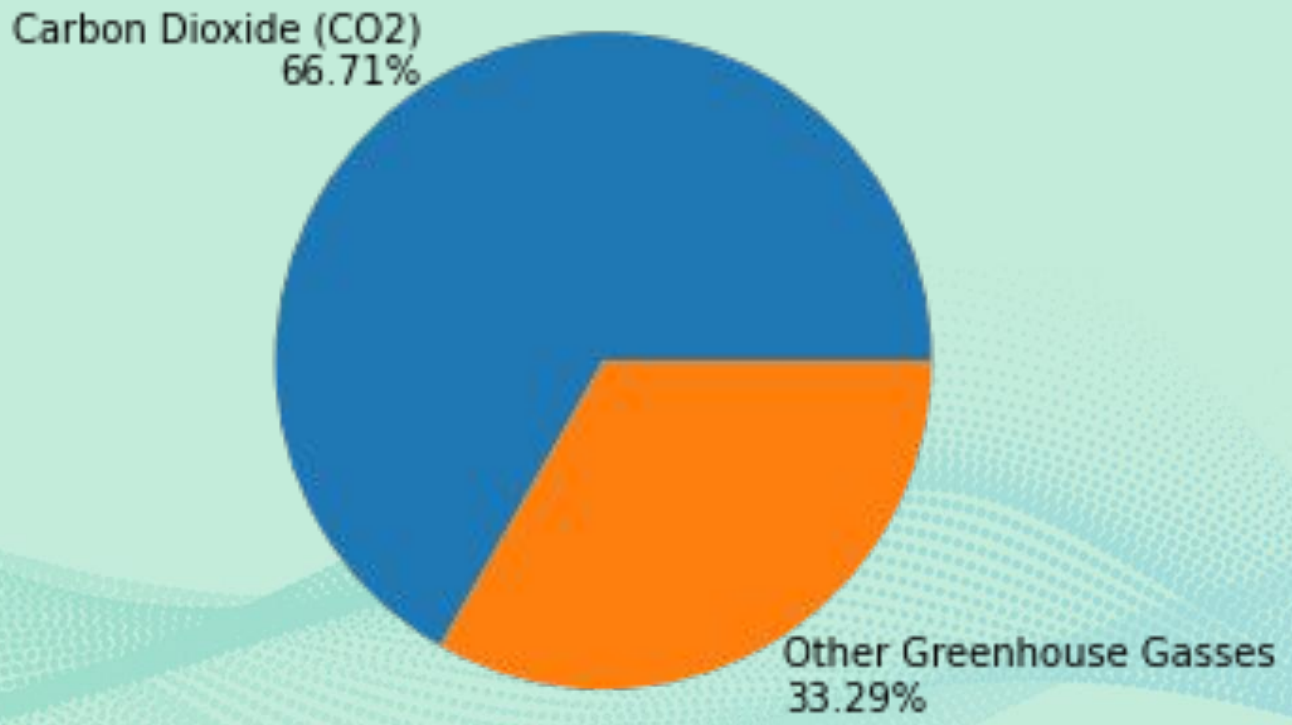
What causes Climate Change?



- Greenhouse gas emissions by each of the countries
- Example:
  Carbon Dioxide (CO2),
  Methane (CH4),
  Nitrous Oxide (NO)
- They absorb solar energy and keep heat close to Earth's surface

Source: https://en.wikipedia.org/wiki/Greenhouse_gas (Wikipedia)

Ratio of Carbon Dioxide (CO_2) in Total Greenhouse Gas Emissions from 1980 to 2020



Problem Statement

- Droughts (moisture deficits) and floods (moisture surpluses) are becoming more common as a result of climate change, which have a negative impact on crop growth and yields.

- The majority of people are unaware of the gravity of the climate situation. Even former US President had a number of tweets claiming that cold weather disproves climate change.

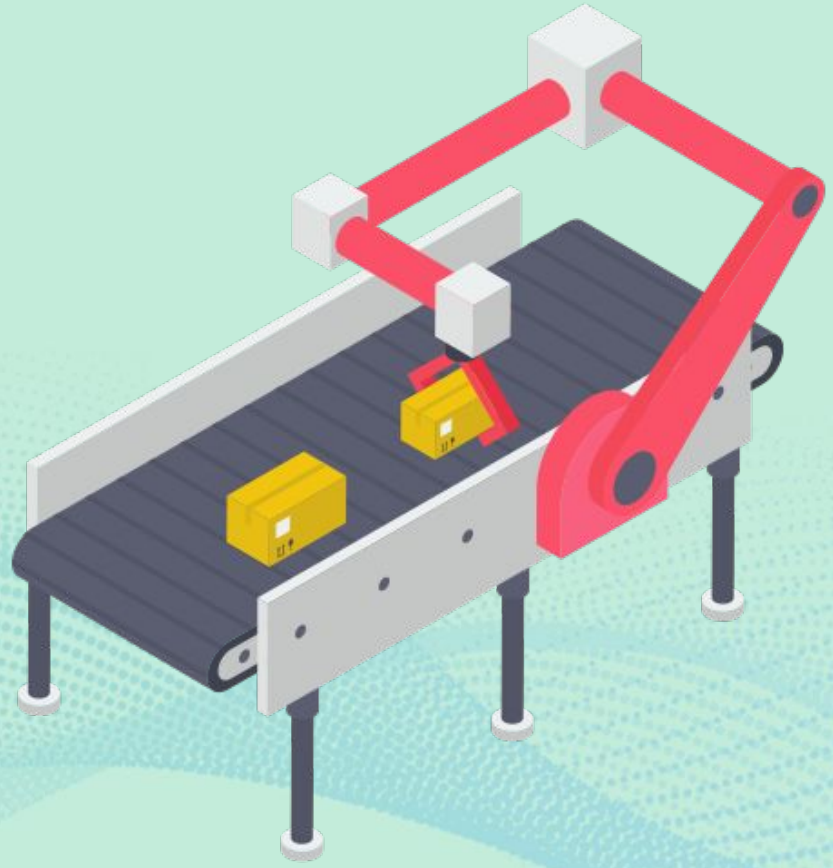
Data Sources

Primary Data Source

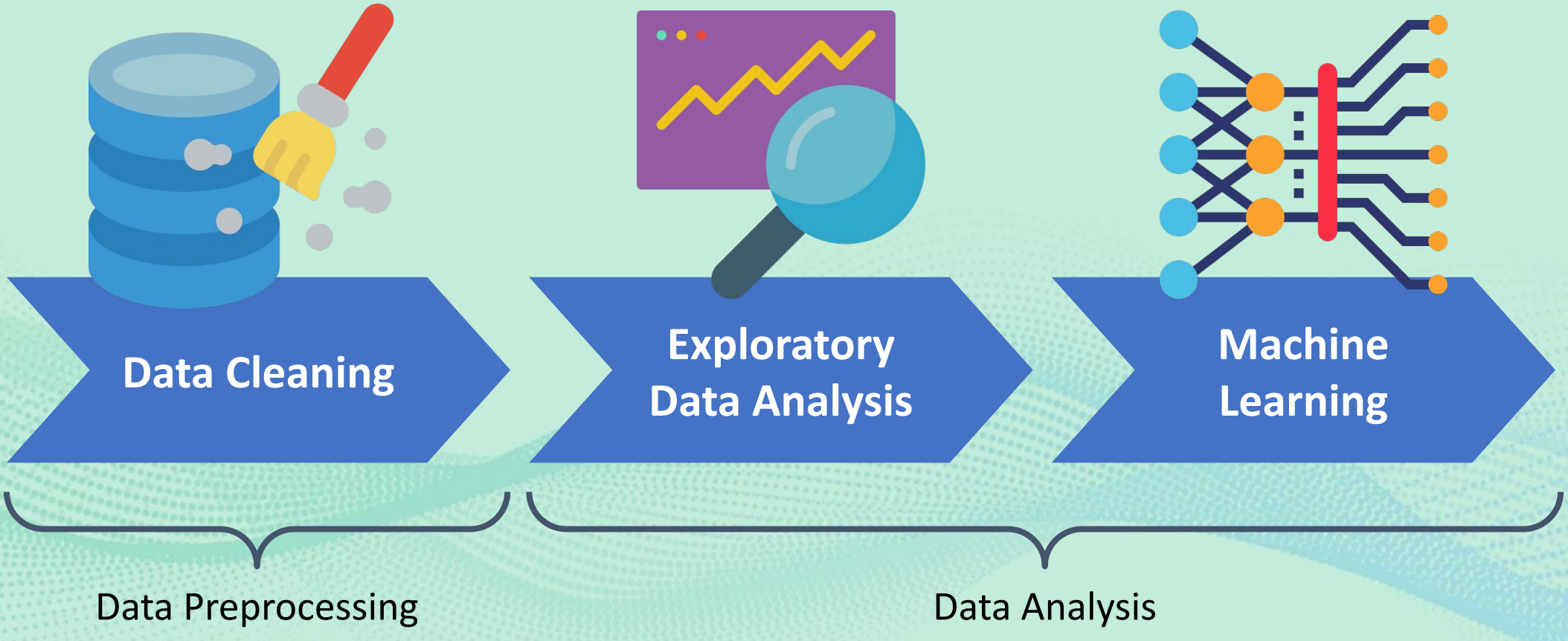
Secondary Data Source

02
Data Preprocessing

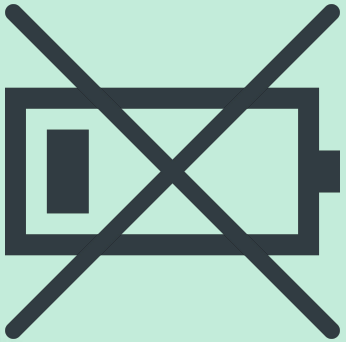
The way we prepare our data for further analysis



Data Pipeline



Data Cleaning Phase 1



Remove countries with low data

Remove variables with more than 50% of missing values

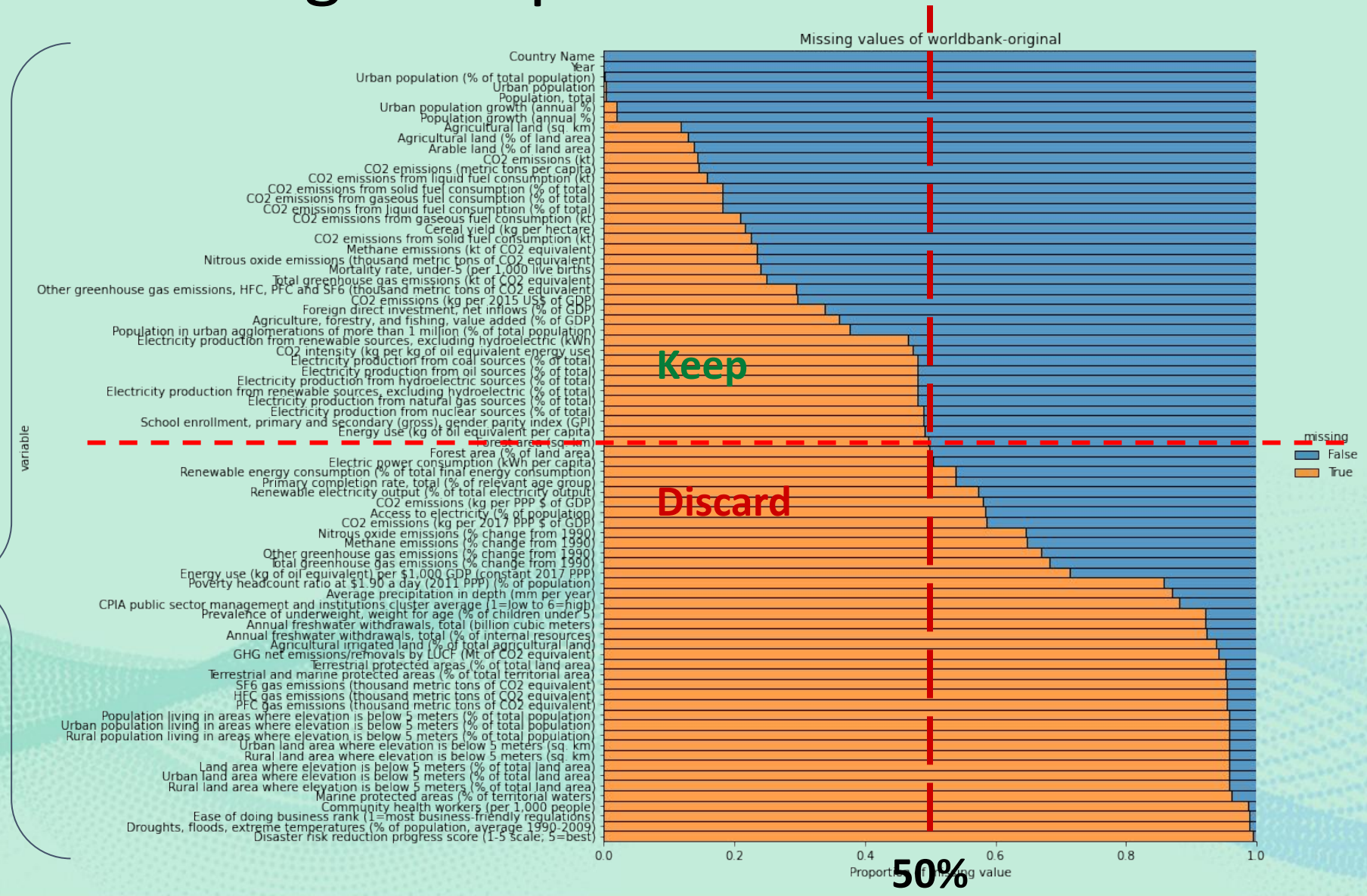
Scoped at year 1980 - 2020

Percentage of Missing Data per Variable

Missing

Non-Missing

World Bank Data Variables



Data Cleaning Phase 2

Feature selection to select best indicators

Multivariate imputation

How can Feature Selection benefits us?

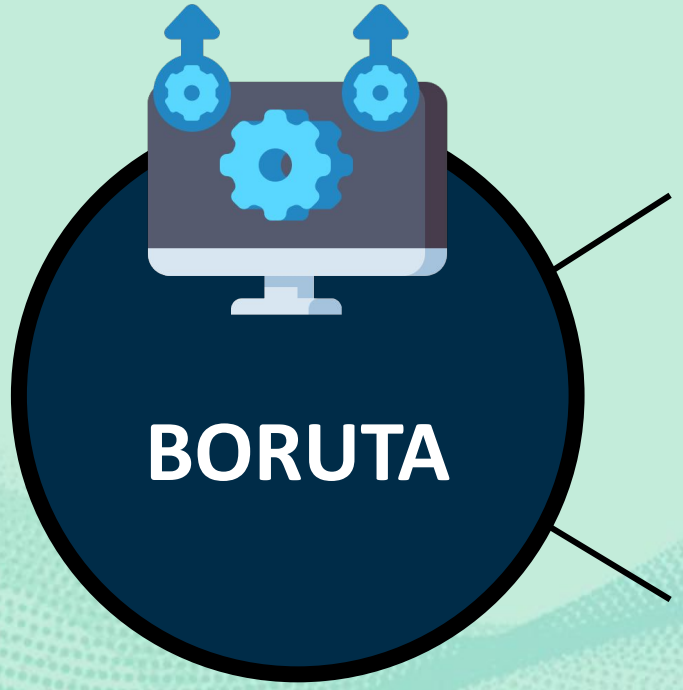
Simpler, straightforward imputation model

Shorter training time

Increases imputation model's performance

Avoids the model from overfitting

Why specifically Boruta?



Takes into account
multi-variable relationships

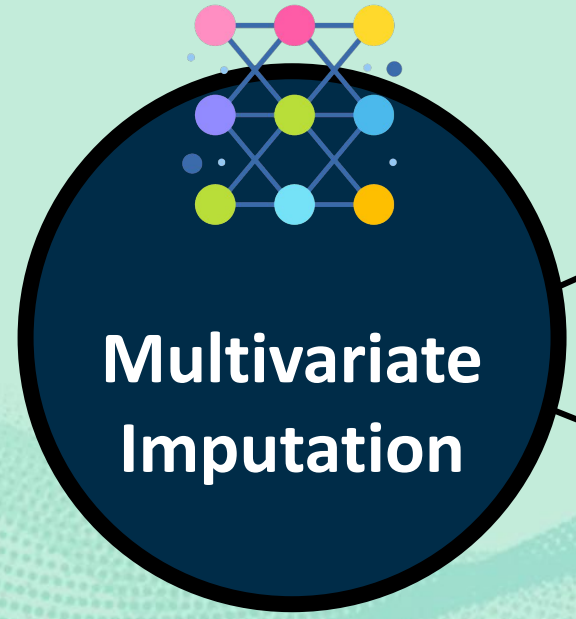
Improvement on random forest
variable importance measure

Feature Selection Results

A list of variables that are highly correlated with
CO2 emissions (metric tons per capita) identified by **BORUTA**:

Year	**Energy use (kg of oil equivalent per capita)**
Urban population (% of total population)	**Total greenhouse gas emissions (kt of CO2 equivalent)**
Population growth (annual %)	**Other greenhouse gas emissions, HFC, PFC and SF6 (thousand metric tons of CO2 equivalent)**
CO2 intensity (kg per kg of oil equivalent energy use)	**Methane emissions (kt of CO2 equivalent)**

How is Multivariate Imputation superior?



Takes account of uncertainty in missing values

Considers the relationships between variables

Multivariate Imputation

Iterative Imputer

- estimates each variables from other existing variables by iterating through them
- `max_iter: 20`
- `initial_strategy: “median”`

Multivariate Imputation

Percentage of imputed data of each variable.

Year	0.00%	**Energy use (kg of oil equivalent per capita)**	40.52%
Urban population (% of total population)	0.09%	**Total greenhouse gas emissions (kt of CO2 equivalent)**	11.06%
Population growth (annual %)	0.24%	**Other greenhouse gas emissions, HFC, PFC and SF6 (thousand metric tons of CO2 equivalent)**	17.68%
CO2 intensity (kg per kg of oil equivalent energy use)	39.59%	**Methane emissions (kt of CO2 equivalent)**	10.10%
CO2 emissions (metric tons per capita)	13.36%	**CO2 emissions (kt)**	13.33%

Data Cleaning Phase 2

Feature selection to select best indicators

Multivariate imputation

Verify and correct imputed data

Preprocessed Dataset Overview

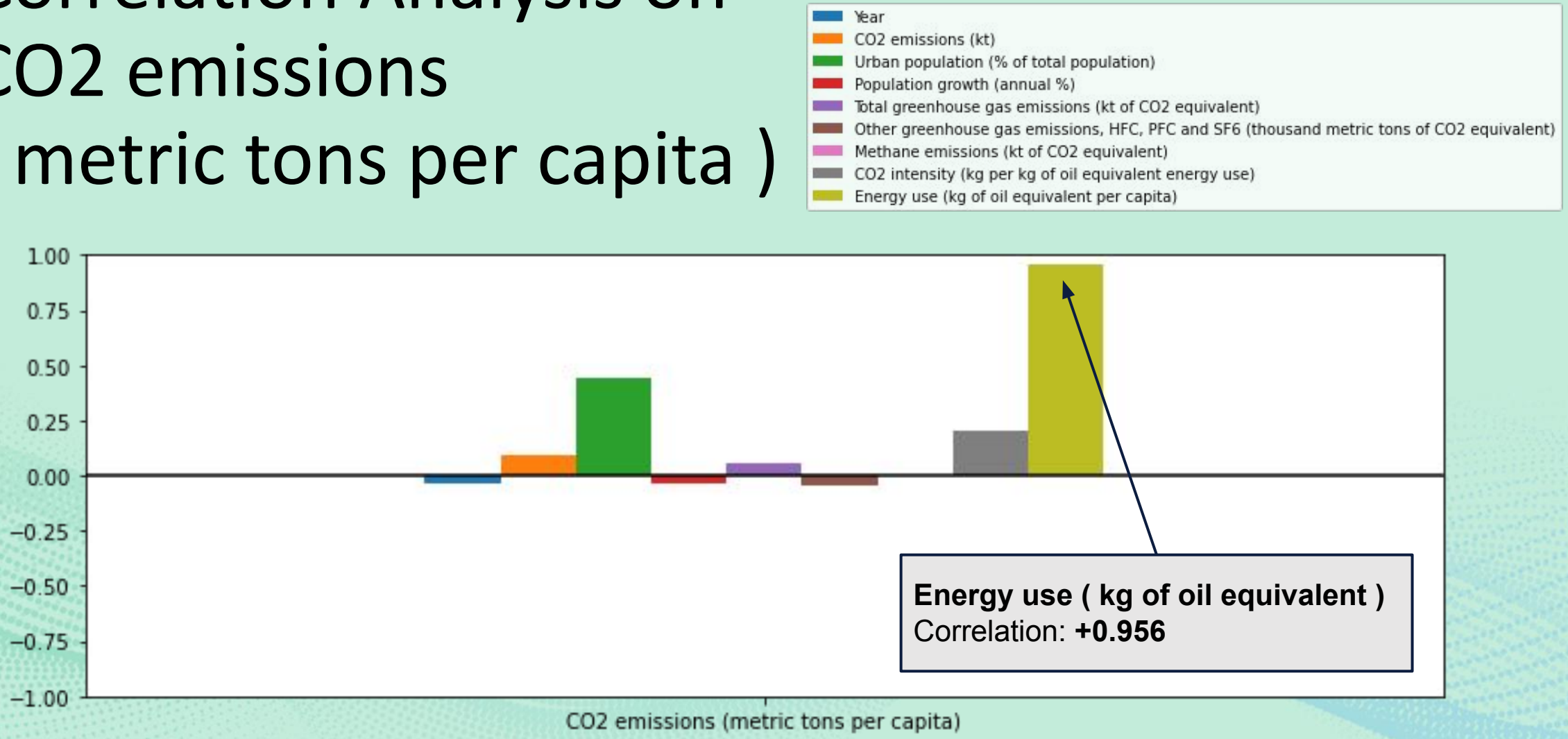
Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
1	Country Name	Char	5	$5.	$5.
2	Year	Num	8	BEST12.	BEST32.
3	CO2 emissions (metric tons per capita)	Num	8	BEST12.	BEST32.
4	CO2 emissions (kt)	Num	8	BEST12.	BEST32.
5	Urban population (% of total population)	Num	8	BEST12.	BEST32.
6	Population growth (annual %)	Num	8	BEST12.	BEST32.
7	Total greenhouse gas emissions (kt of CO2 equivalent)	Num	8	BEST12.	BEST32.
8	Other greenhouse gas emissions,HFC, PFC and SF6 (thousand metric tons of CO2 equivalent)	Num	8	BEST12.	BEST32.
9	Methane emissions (kt of CO2 equivalent)	Num	8	BEST12.	BEST32.
10	CO2 intensity (kg per kg oil equivalent)	Num	8	BEST12.	BEST32.
11	Energy use (kg of oil equivalent)	Num	8	BEST12.	BEST32.

11 Variables
10, 455 Observations
No missing values

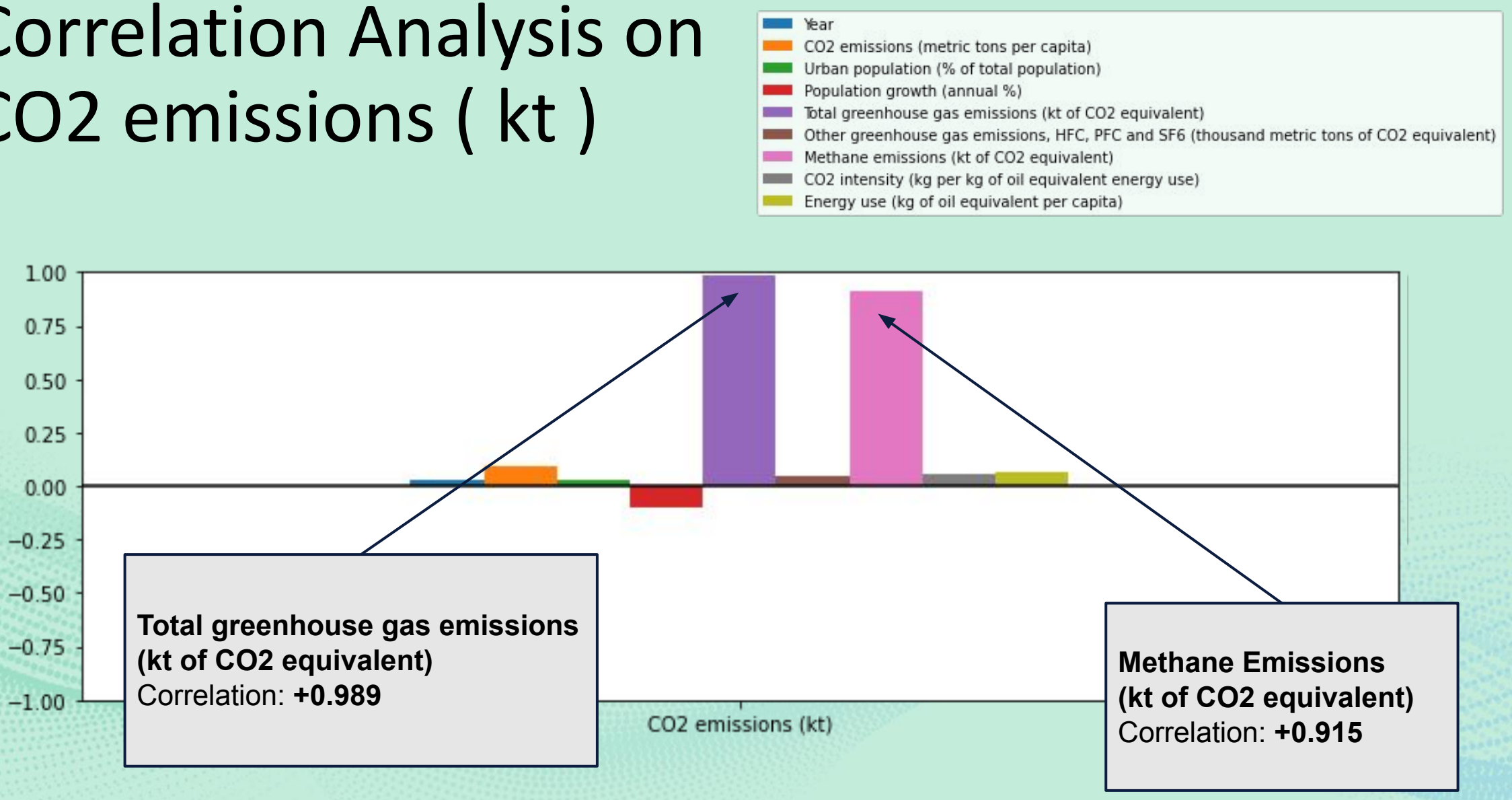
03
Data Analysis

Analyse our imputed data using machine learning algorithms

Correlation Analysis on CO2 emissions (metric tons per capita)

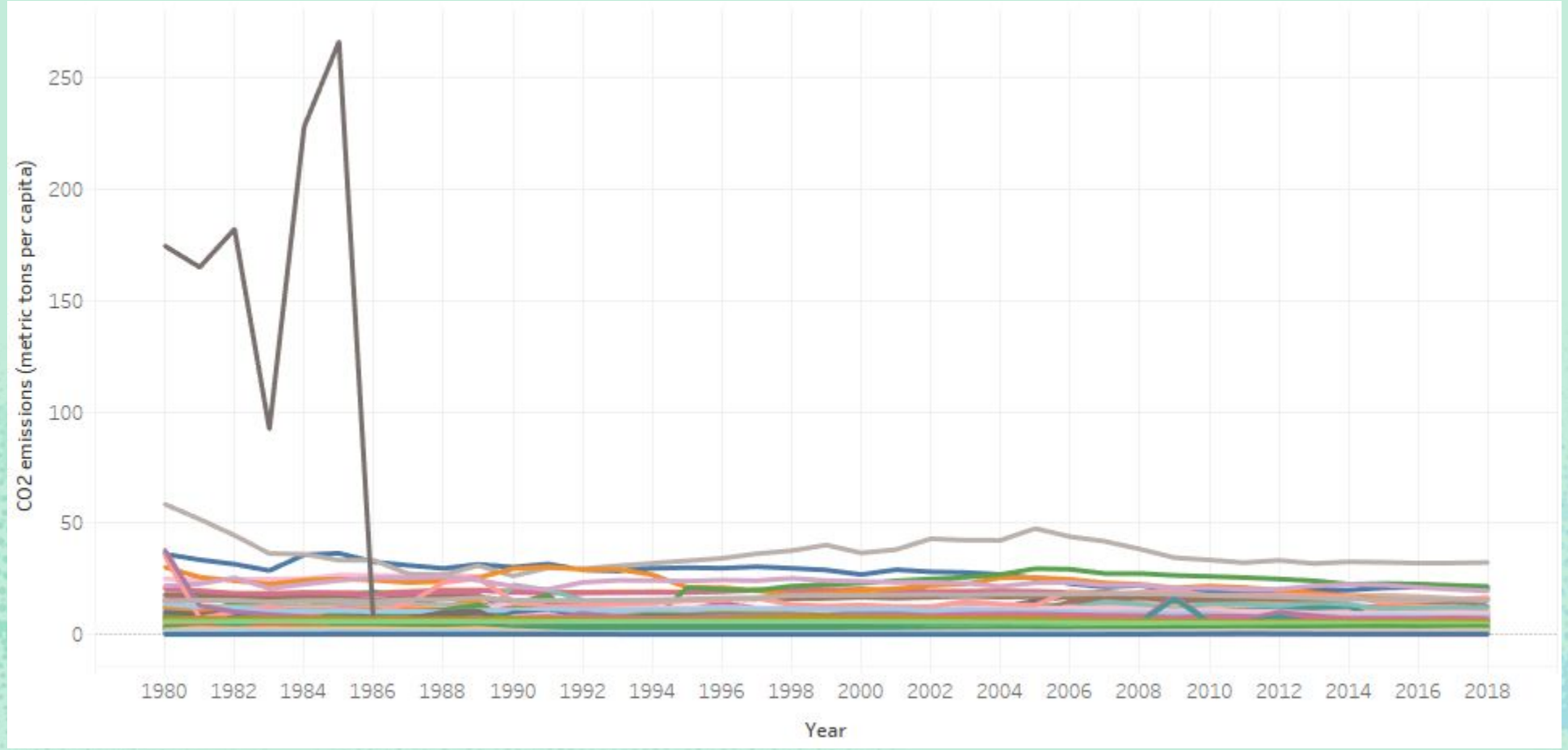


Correlation Analysis on CO2 emissions (kt)



Time Series Clustering

for **CO2 Emissions (metric tons per capita)**



Time Series Clustering

for **CO2 Emissions (metric tons per capita)**

Data preprocessing

13.36% of data is imputed using dimensions selected by feature selection, Boruta

Years 1980 - 2018

Total 200 countries (6 countries are excluded)

Aruba, Qatar, Russian Federation, United States, China, India

Model Comparison

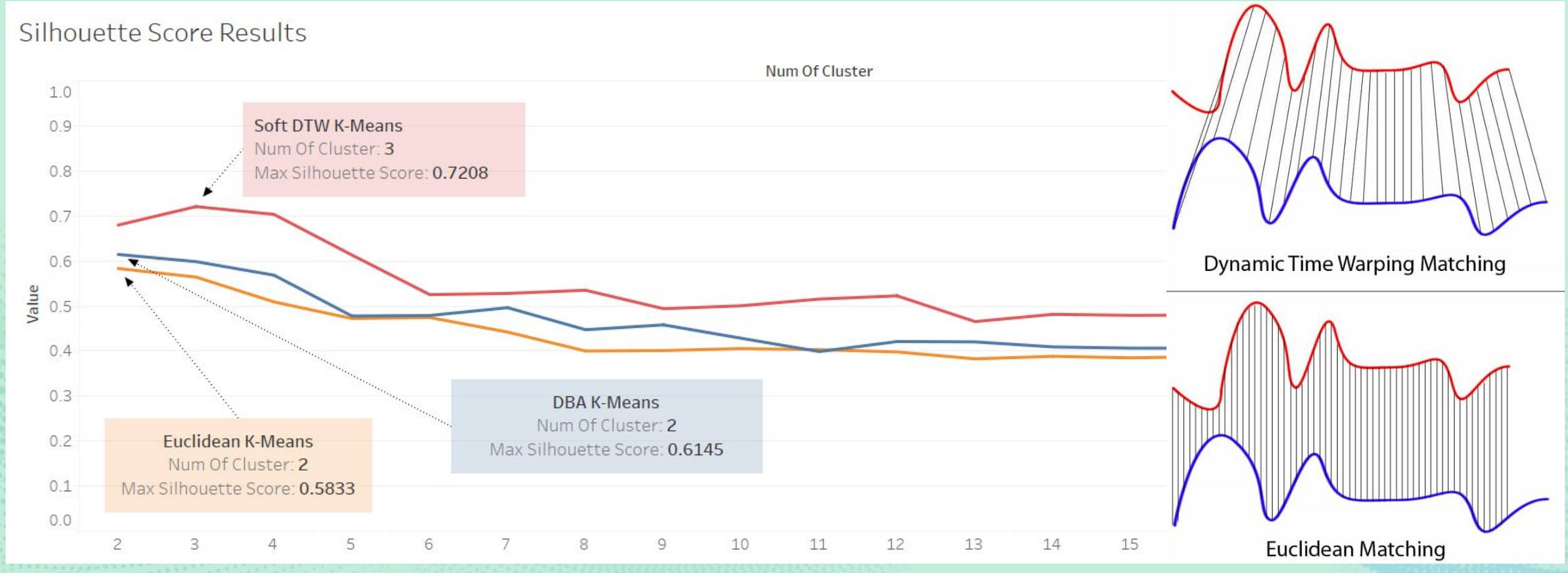
Select best number of cluster, k using silhouette score.

Silhouette score will indicate the best algorithm to be used as well.

Clustering and analysis

Start time series clustering and plot their results.

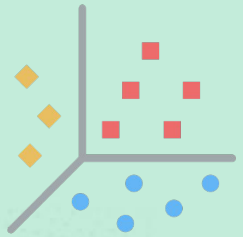
Silhouette Scores



Silhouette Scores

Best k-means Algorithm

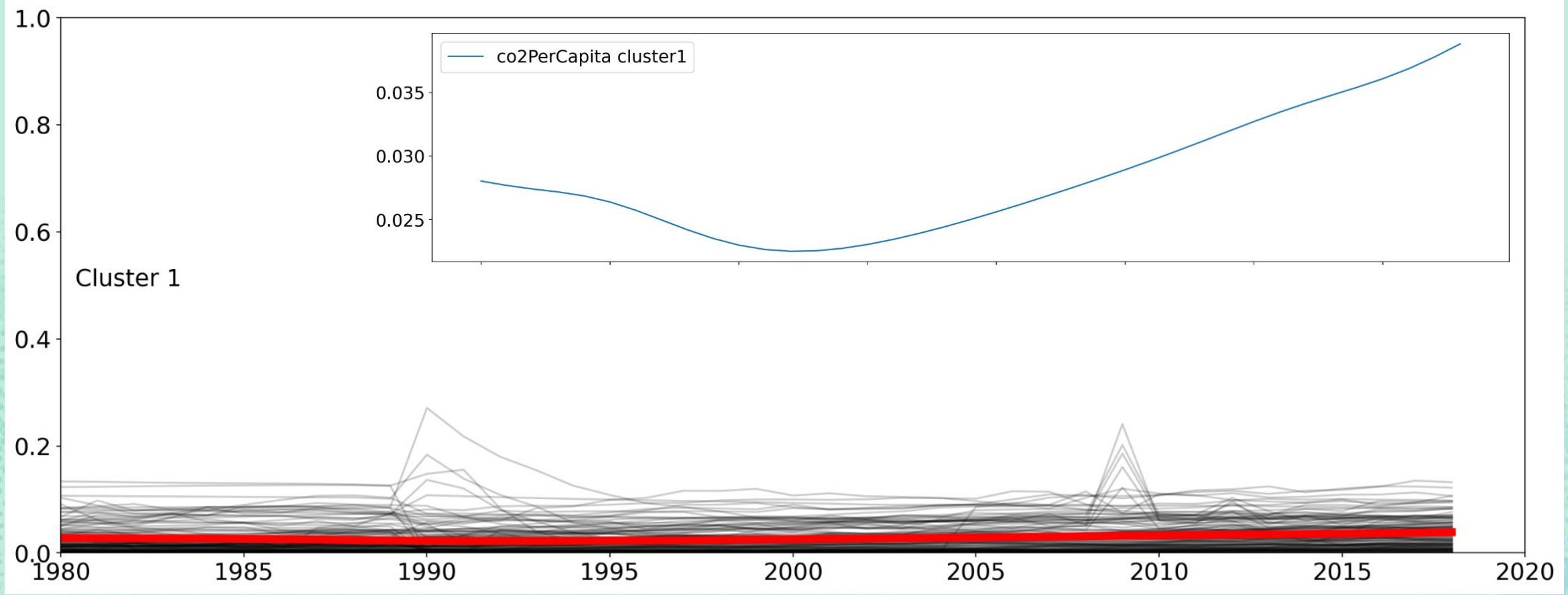
Soft DTW K-means



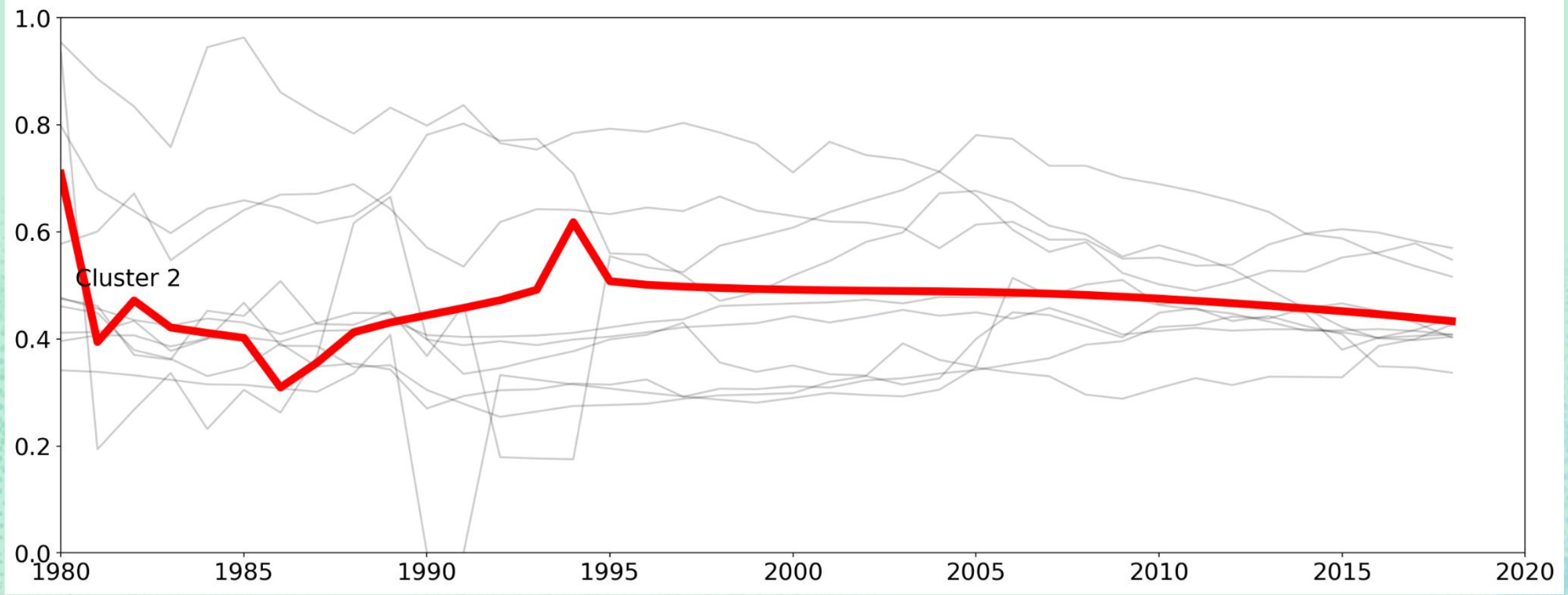
Optimal Number of Cluster

k = 3

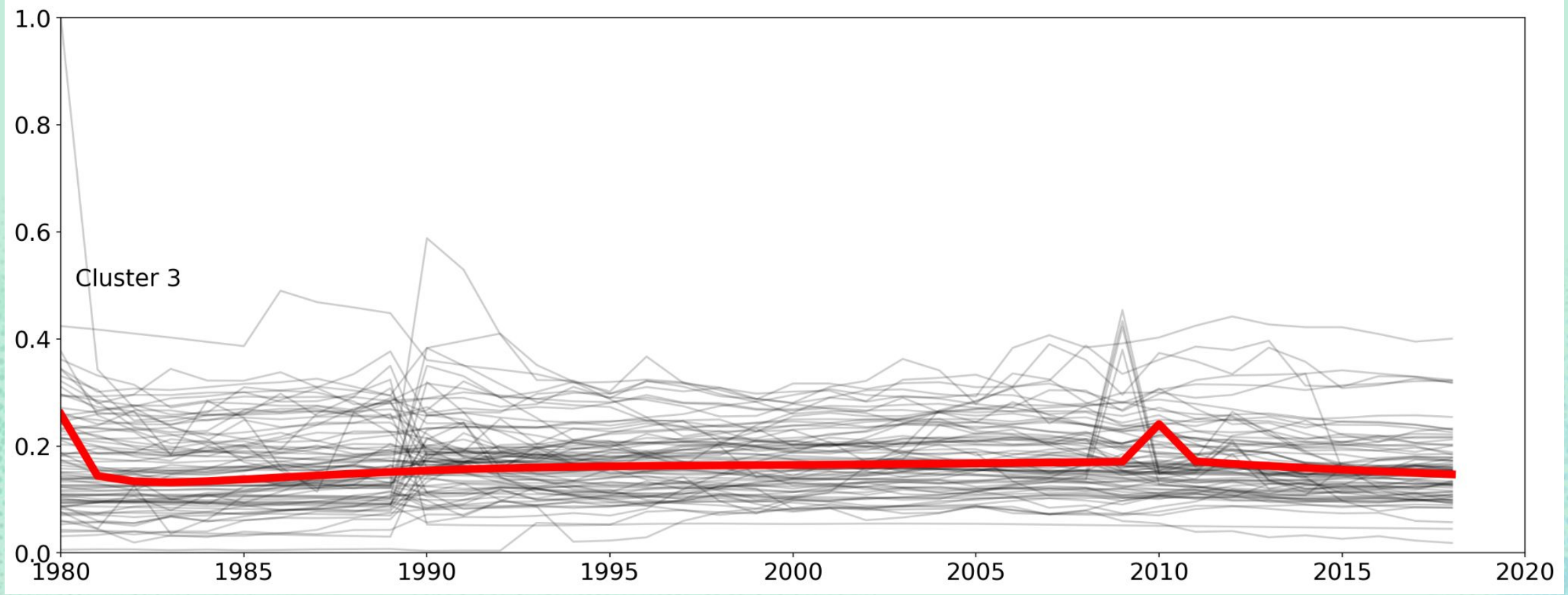
Clustering Result of CO2 emissions (metric tons per capita)



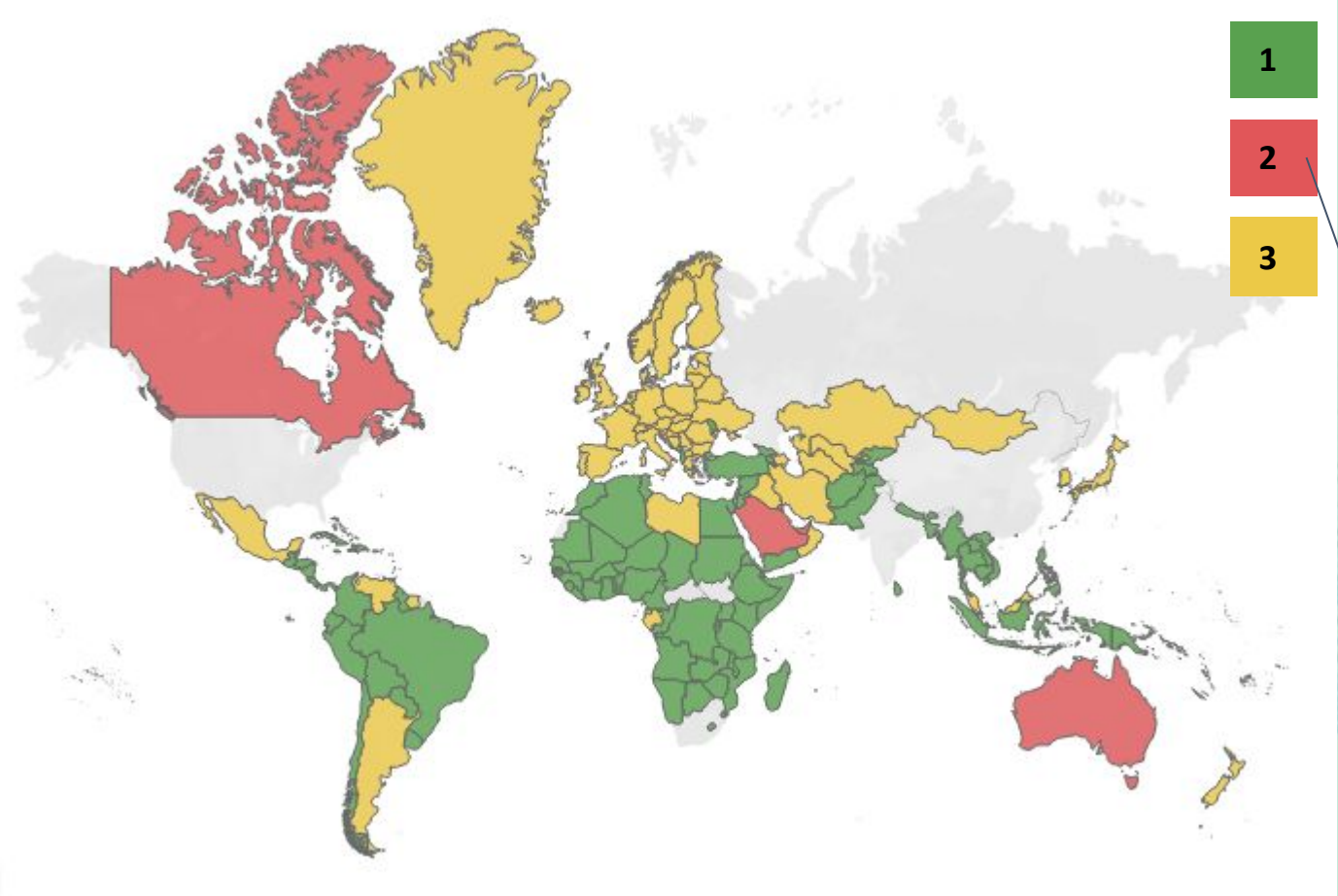
Clustering Result of CO2 emissions (metric tons per capita)



Clustering Result of CO2 emissions (metric tons per capita)



Clustering Result Choropleth



Cluster 2

1. United Arab Emirates
2. Australia
3. Bahrain
4. Brunei Darussalam
5. Canada
6. Kuwait
7. Luxembourg
8. Palau
9. Saudi Arabia
10. Trinidad and Tobago

Clustering Result Country Listing - Cluster 1

Afghanistan	Congo, Rep.	Haiti	Mozambique	Sudan
Albania	Costa Rica	Honduras	Myanmar	Syrian Arab Republic
Algeria	Cote d'Ivoire	Indonesia	Namibia	Tajikistan
Angola	Cuba	Jamaica	Nepal	Tanzania
Armenia	Djibouti	Jordan	Nicaragua	Thailand
Bangladesh	Dominica	Kenya	Niger	Timor-Leste
Belize	Dominican Republic	Kiribati	Nigeria	Togo
Benin	Ecuador	Kyrgyz Republic	Pakistan	Tonga
Bhutan	Egypt, Arab Rep.	Lao PDR	Panama	Tunisia
Bolivia	El Salvador	Lebanon	Papua New Guinea	Turkey
Botswana	Eritrea	Lesotho	Paraguay	Tuvalu
Brazil	Eswatini	Liberia	Peru	Uganda
British Virgin Islands	Ethiopia	Madagascar	Philippines	Uruguay
Burkina Faso	Fiji	Malawi	Rwanda	Vanuatu
Burundi	French Polynesia	Maldives	Samoa	Vietnam
Cabo Verde	Gambia, The	Mali	Sao Tome and Principe	Yemen, Rep.
Cambodia	Georgia	Marshall Islands	Senegal	Zambia
Cameroon	Ghana	Mauritania	Sierra Leone	Zimbabwe
Chad	Grenada	Mauritius	Solomon Islands	
Chile	Guatemala	Micronesia, Fed. Sts.	Somalia	
Colombia	Guinea	Moldova	Sri Lanka	
Comoros	Guinea-Bissau	Montenegro	St. Lucia	
Congo, Dem. Rep.	Guyana	Morocco	St. Vincent and the Grenadines	

Clustering Result Country Listing - Cluster 3

American Samoa	France	Macao SAR, China	Suriname
Andorra	Gabon	Malaysia	Sweden
Antigua and Barbuda	Germany	Malta	Switzerland
Argentina	Gibraltar	Mexico	Turkmenistan
Austria	Greece	Mongolia	Turks and Caicos Islands
Azerbaijan	Greenland	Nauru	Ukraine
Bahamas, The	Guam	Netherlands	United Kingdom
Barbados	Hong Kong SAR, China	New Caledonia	Uzbekistan
Belarus	Hungary	New Zealand	Venezuela, RB
Belgium	Iceland	North Macedonia	Virgin Islands (U.S.)
Bermuda	Iran, Islamic Rep.	Norway	West Bank and Gaza
Bosnia and Herzegovina	Iraq	Oman	
Bulgaria	Ireland	Poland	
Caribbean small states	Israel	Portugal	
Cayman Islands	Italy	Puerto Rico	
Croatia	Japan	Romania	
Cyprus	Kazakhstan	Serbia	
Czech Republic	Korea, Dem. People's Rep.	Seychelles	
Denmark	Korea, Rep.	Singapore	
Equatorial Guinea	Latvia	Slovak Republic	
Estonia	Libya	Slovenia	
Faroe Islands	Liechtenstein	Spain	
Finland	Lithuania	St. Kitts and Nevis	

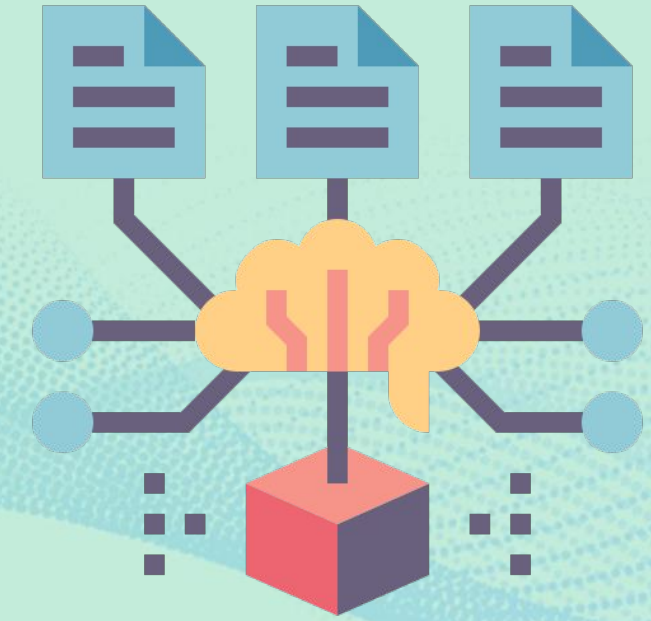
Time Series Forecasting Model

Algorithm or Techniques Used:

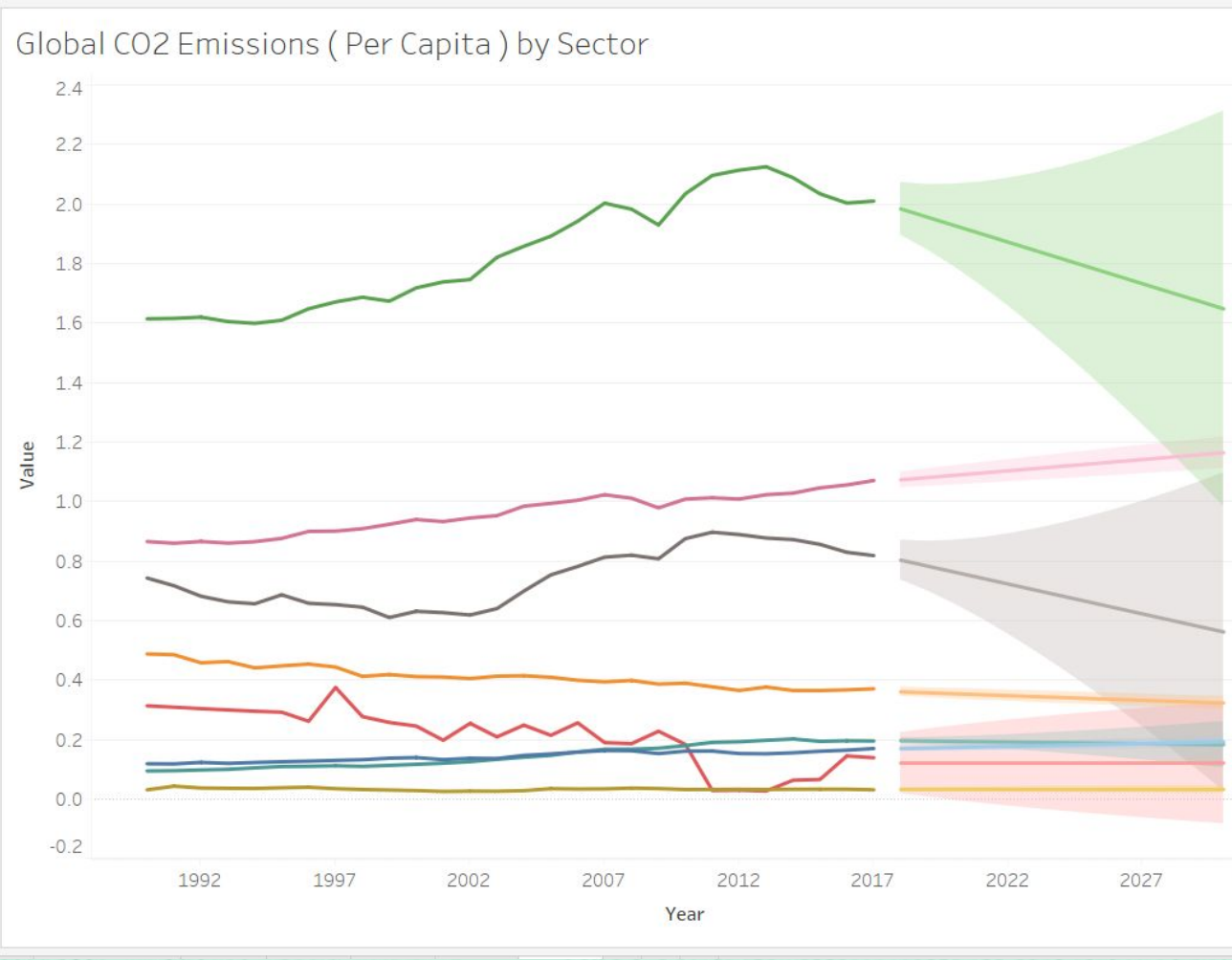
- We used a technique known as exponential smoothing.
- Find a regular pattern in target variables that can be continued into the future.
- Recent data are given more weights.

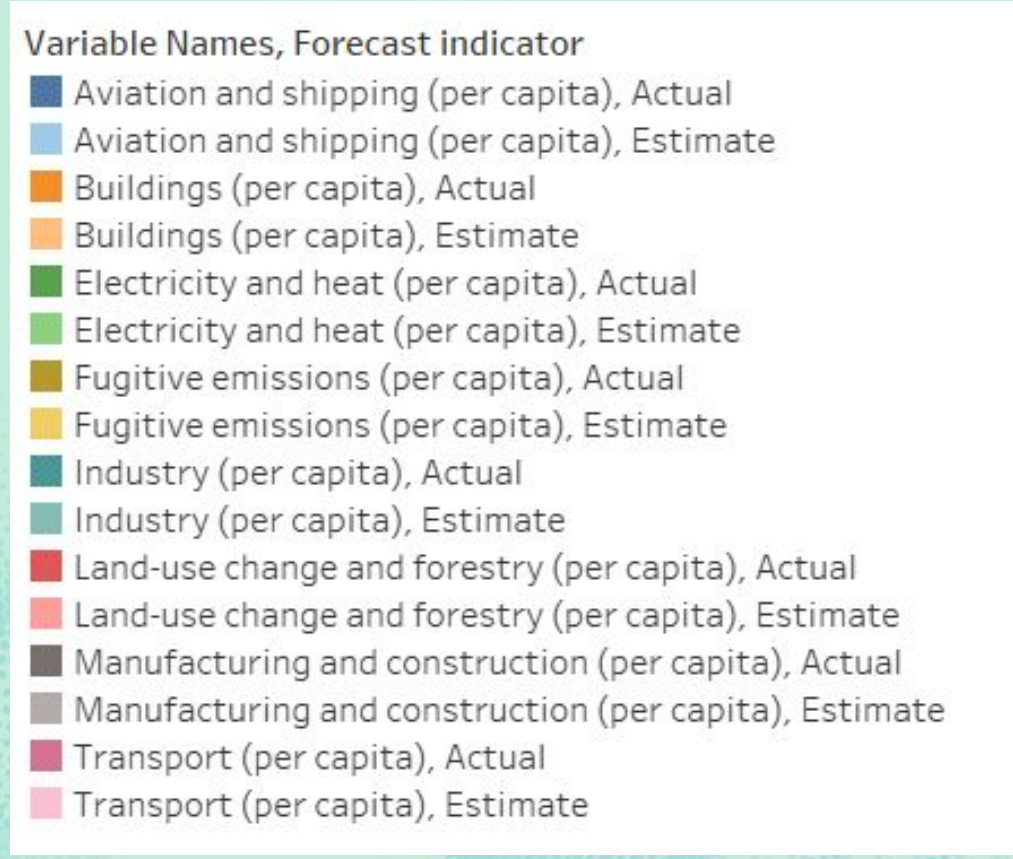
Basic Decomposition Model:

- **Additive: = Trend + Seasonal + Random**
- The additive model is useful when the seasonal variation is relatively constant over time while the absolute values are growing.

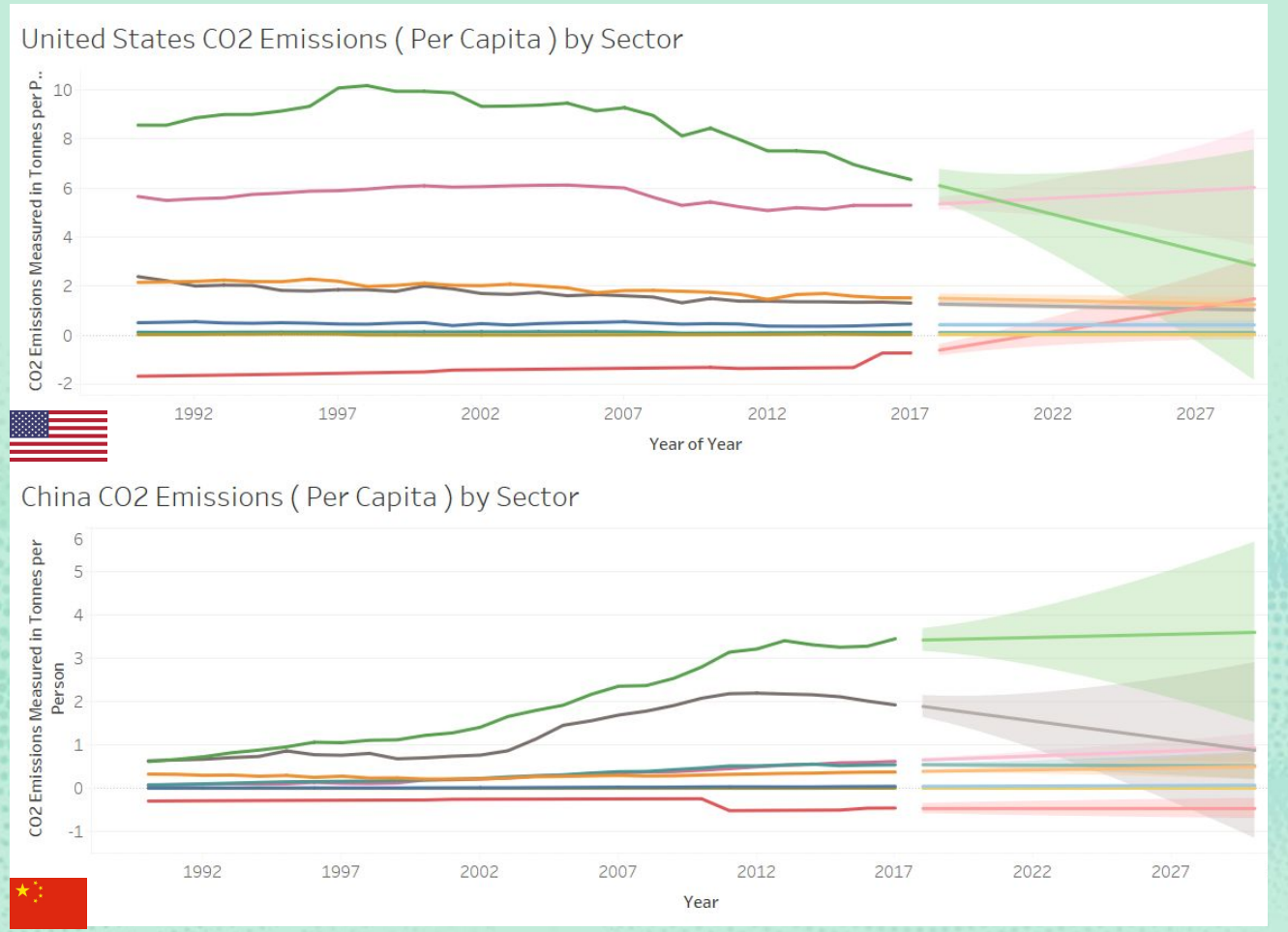


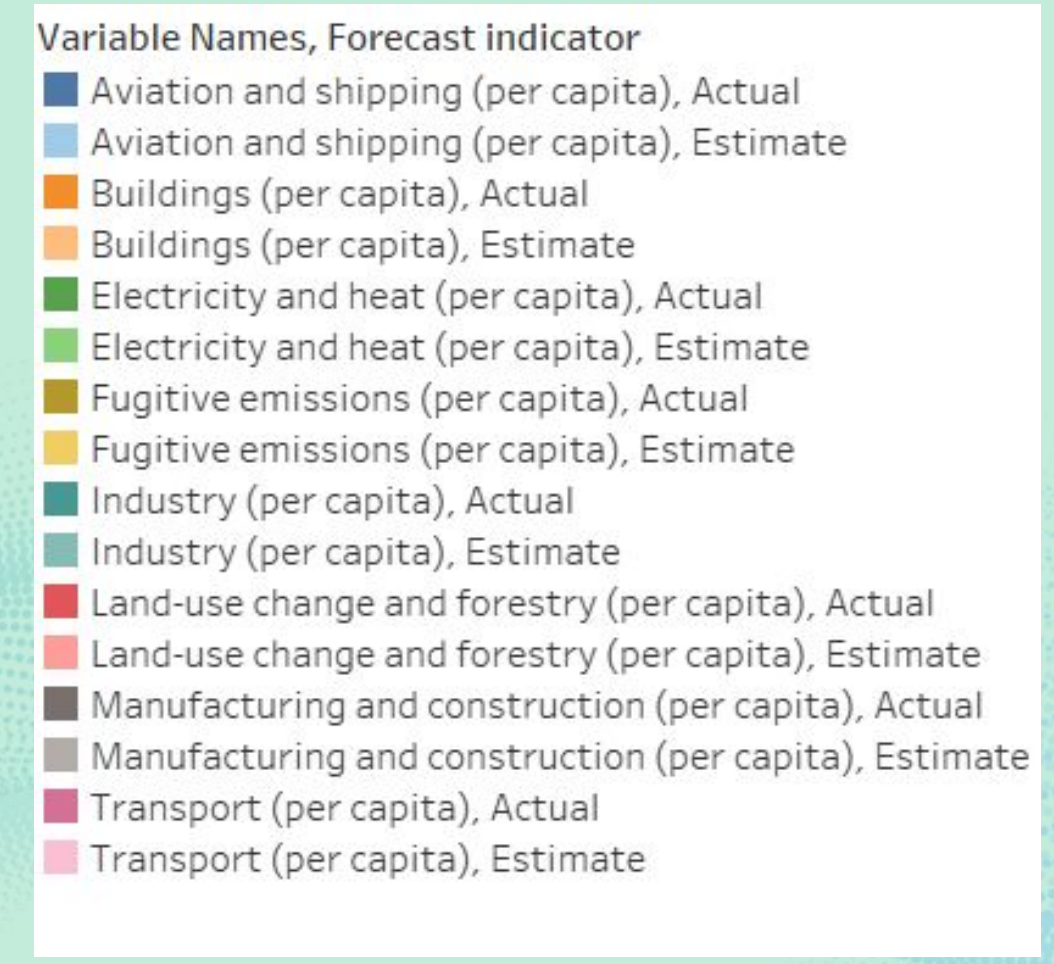
Global CO2 Emissions Trend Per Capita (tons) by Sector



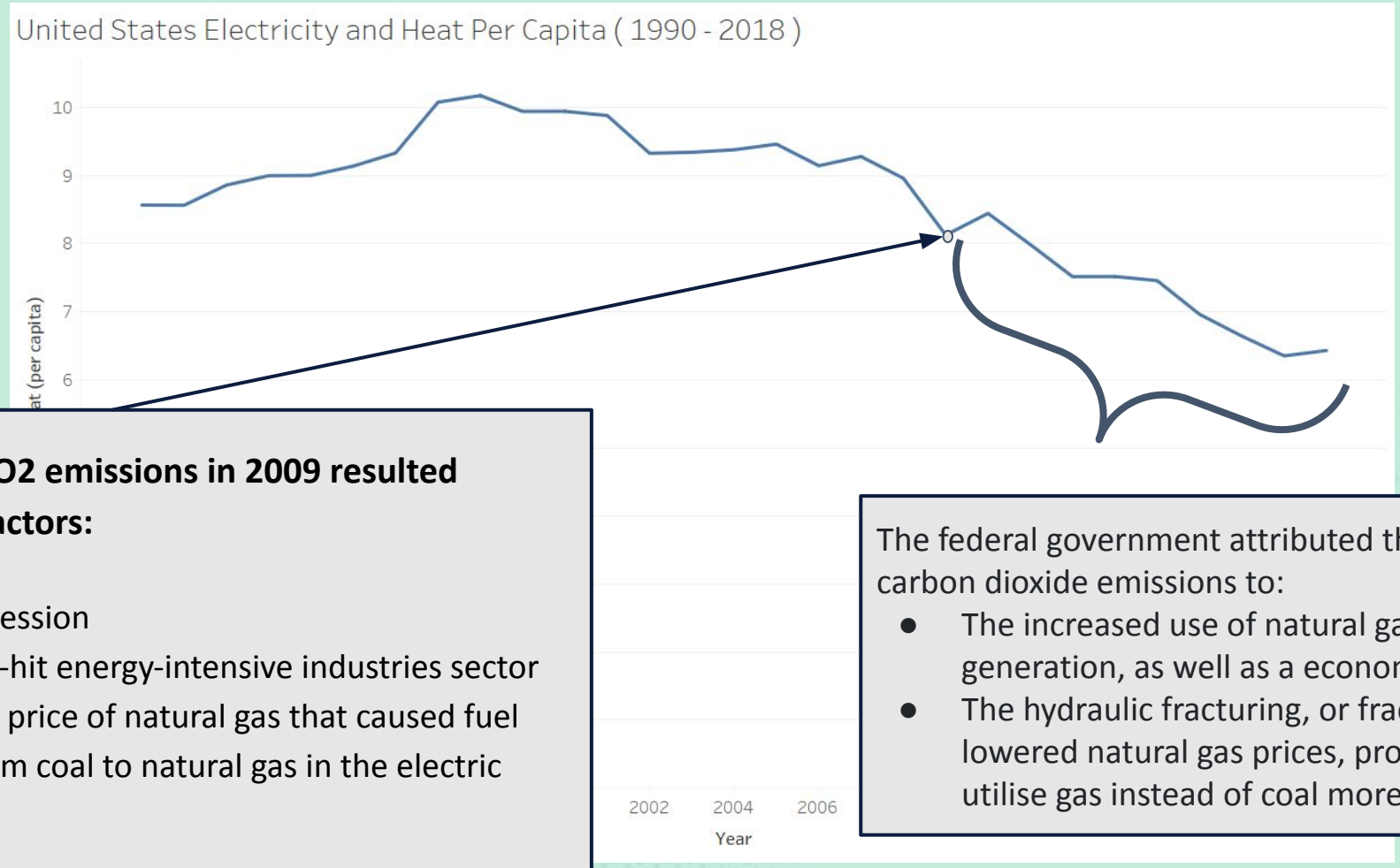


CO2 Emissions Per Capita (tons) by Sector between USA & China





USA Electricity and Haat CO2 Emissions Per Capita (tons)



The decrease in U.S. CO2 emissions in 2009 resulted primarily from three factors:

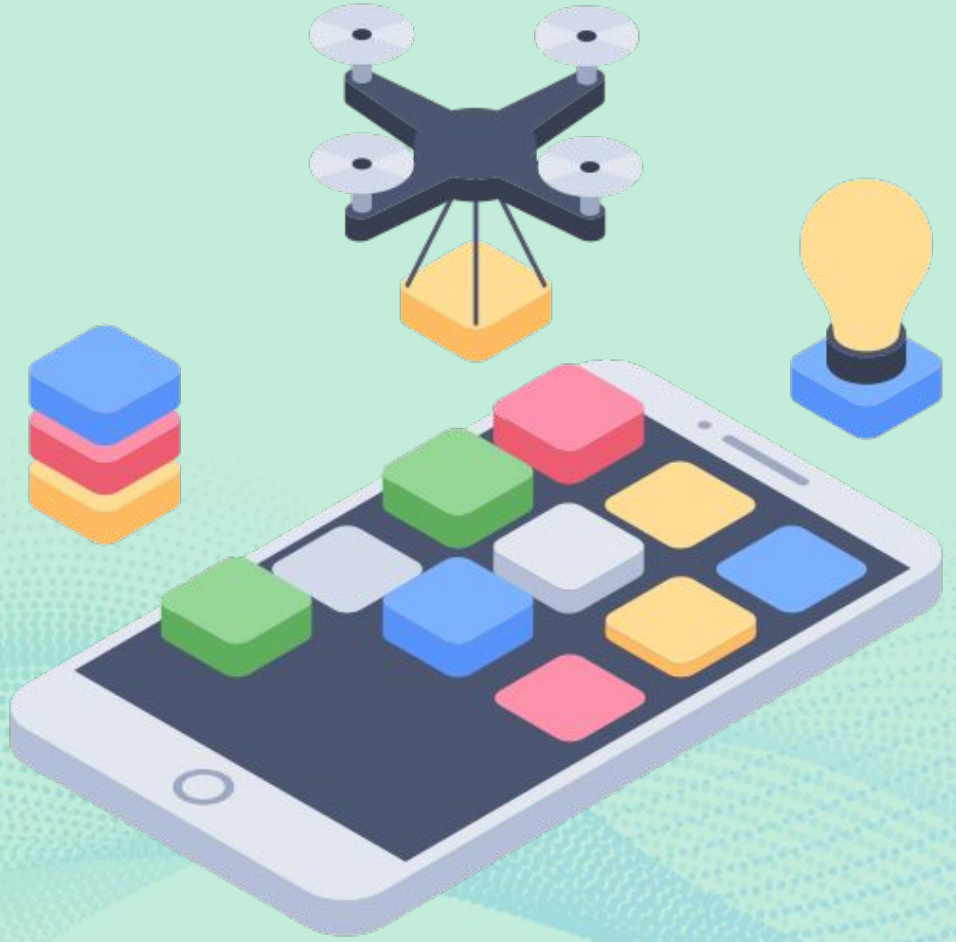
1. An economy in recession
2. A particularly hard-hit energy-intensive industries sector
3. A large drop in the price of natural gas that caused fuel switching away from coal to natural gas in the electric power sector.

The federal government attributed the decrease in carbon dioxide emissions to:

- The increased use of natural gas for energy generation, as well as a economy recession.
- The hydraulic fracturing, or fracking, boom has lowered natural gas prices, prompting utilities to utilise gas instead of coal more frequently.

04
Recommendation

Provides conclusions to our findings and suggest solutions to the problem

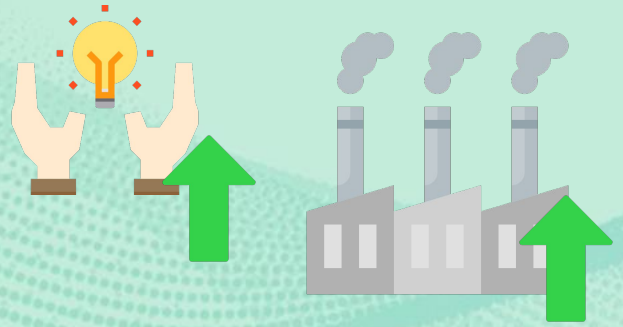


Recommendations

- Takes more fields into research, link them with the co2 to understand the co2 emissions impact to different areas.
- Get more data to increase the accuracy / performance of the machine learning model.
- Request review from professionals.

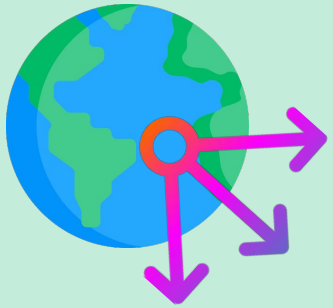
Conclusion

Carbon dioxide and **methane** emissions are key factors to the rise of **greenhouse gas** emissions

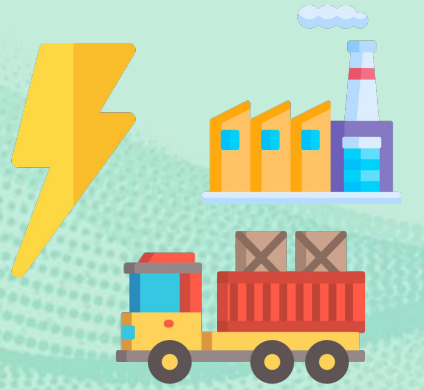


Energy use per capita also highly correlates with the **CO2 emissions per capita**.

Conclusion



Grouped countries into **three main clusters** according to their **CO2 emission pattern**



Energy, Transportation and Manufacturing sector should be prioritised to prevent exponential growth in CO2 emissions along with the population growth.

Reference

Article in conference proceedings Alcántara, V., Duarte, R., & Obis Artal, M. 2006. “Regional decomposition of CO2 emissions in the world: a cluster analysis.” *Working papers (Universitat Autònoma de Barcelona. Departament d'Economia Aplicada)*
Available at https://ddd.uab.cat/pub/estudis/2006/hdl_2072_2097/wpdea0306.pdf

Journal Article Cuturi, M. & Blondel, M.. 2017. “Soft-DTW: a Differentiable Loss Function for Time-Series.” *Proceedings of the 34th International Conference on Machine Learning, in Proceedings of Machine Learning Research*, 70:894-903.
Available at https://proceedings.mlr.press/v70/cuturi17a.html

Thank you!❤️

Contact Information
liewkuanyung@gmail.com