Final Project: The global warming "troble maker" Deep Dive

By Qiuyu Huang, Harshita Goyal, and Zhen Qi

Abstract

There has been a broad scientific consensus that the Earth's average temperature has been increasing over the past century due to human activity, primarily the burning of fossil fuels and deforestation. This warming trend is leading to a range of negative impacts on the environment, including rising sea levels, more frequent and severe weather events, and the loss of biodiversity. The effects of climate change can be felt in our daily lives, from the food we eat to the air we breathe. Addressing this issue will require a collective effort from individuals, governments, and businesses around the world.Given the far-reaching consequences of climate change on our daily lives, it is crucial to take collective action to mitigate its effects and ensure a sustainable future for coming generations

As a team of data scientists, we aim to use data science techniques to analyze the key influencing factors of global warming. Our goal is to gain insights into the underlying causes of climate change and develop strategies to tackle the issue at its root.

Data Source

To achieve this, we plan to use the climate_change.csv dataset found at [dataset (https://www.kaggle.com/econdata/climate-change)](https://www.kaggle.com/econdata/climate-change).

The file climate_change.csv contains climate data from May 1983 to December 2008. The available variables include: Year, Month, Temp, CO2, N2O, CH4, CFC.11, CFC.12 and TSI.

- **Year:** the observation year.
- **Month:** the observation month.
- **Temp:** the difference in degrees Celsius between the average global temperature in that period and a reference value. This data comes from the Climatic Research Unit at the University of East Anglia.
- **CO2, N2O, CH4, CFC.11, CFC.12:**: atmospheric concentrations of carbon dioxide (CO2), nitrous oxide (N2O), methane (CH4), trichlorofluoromethane (CCl3F; commonly referred to as CFC-11) and dichlorodifluoromethane (CCl2F2; commonly referred to as CFC-12), respectively. This data comes from the ESRL/NOAA Global Monitoring Division. CO2, N2O and CH4 are expressed in ppmv (parts per million by volume -- i.e., 397 ppmv of CO2 means that CO2 constitutes 397 millionths of the total volume of the atmosphere). CFC.11 and CFC.12 are expressed in ppbv (parts per billion by volume).
- **Aersols:** Aerosols: the mean stratospheric aerosol optical depth at 550 nm. This variable is linked to volcanoes, as volcanic eruptions result in new particles being added to the atmosphere, which affect how much of the sun's energy is reflected back into space. This data is from the Godard Institute for Space Studies at NASA.

- **TSI:**: the total solar irradiance (TSI) in W/m2 (the rate at which the sun's energy is deposited per unit area). Due to sunspots and other solar phenomena, the amount of energy that is given off by the sun varies substantially with time. This data is from the SOLARIS-HEPPA project website.
- **MEI:**: multivariate El Nino Southern Oscillation index (MEI), a measure of the strength of the El Nino/La Nina-Southern Oscillation (a weather effect in the Pacific Ocean that affects global temperatures). This data comes from the ESRL/NOAA Physical Sciences Division.

In [3]:
```
import numpy as np
import pandas as pd
import scipy.stats as stats
import matplotlib.pyplot as plt
from matplotlib.colors import LogNorm
import seaborn as sns
from sklearn.preprocessing import scale
import statsmodels.api as sm
from sklearn.mixture import GaussianMixture

import matplotlib
from matplotlib import pyplot
from sklearn.metrics import r2_score
import mpl_toolkits.mplot3d as p3d

%matplotlib inline
SEED = 666
```

Data Display

In [4]:
```
df = pd.read_csv('data/climate_change.csv')
df.head(10)
```

Out[4]:

	Year	Month	MEI	CO2	CH4	N2O	CCl3F	CCl2F2	TSI	Aerosols	Temp
0	1983	5	2.556	345.96	1638.59	303.677	191.324	350.113	1366.1024	0.0863	0.109
1	1983	6	2.167	345.52	1633.71	303.746	192.057	351.848	1366.1208	0.0794	0.118
2	1983	7	1.741	344.15	1633.22	303.795	192.818	353.725	1366.2850	0.0731	0.137
3	1983	8	1.130	342.25	1631.35	303.839	193.602	355.633	1366.4202	0.0673	0.176
4	1983	9	0.428	340.17	1648.40	303.901	194.392	357.465	1366.2335	0.0619	0.149
5	1983	10	0.002	340.30	1663.79	303.970	195.171	359.174	1366.0589	0.0569	0.093
6	1983	11	-0.176	341.53	1658.23	304.032	195.921	360.758	1366.1072	0.0524	0.232
7	1983	12	-0.176	343.07	1654.31	304.082	196.609	362.174	1366.0607	0.0486	0.078
8	1984	1	-0.339	344.05	1658.98	304.130	197.219	363.359	1365.4261	0.0451	0.089
9	1984	2	-0.565	344.77	1656.48	304.194	197.759	364.296	1365.6618	0.0416	0.013

In [5]:
```
# examine the dataset

print(df.describe())
```

```
                Year       Month         MEI         CO2          CH4  \
count   308.000000  308.000000  308.000000  308.000000   308.000000   
mean   1995.662338    6.551948    0.275555  363.226753  1749.824513   
std       7.423197    3.447214    0.937918   12.647125    46.051678   
min    1983.000000    1.000000   -1.635000  340.170000  1629.890000   
25%    1989.000000    4.000000   -0.398750  353.020000  1722.182500   
50%    1996.000000    7.000000    0.237500  361.735000  1764.040000   
75%    2002.000000   10.000000    0.830500  373.455000  1786.885000   
max    2008.000000   12.000000    3.001000  388.500000  1814.180000   

              N2O       CCl3F       CCl2F2          TSI   Aerosols  
Temp  
count  308.000000  308.000000  308.000000   308.000000  308.000000  30
8.000000  
mean   312.391834  251.973068  497.524782  1366.070759    0.016657    
0.256776  
std      5.225131   20.231783   57.826899     0.399610    0.029050    
0.179090  
min    303.677000  191.324000  350.113000  1365.426100    0.001600   -
0.282000  
25%    308.111500  246.295500  472.410750  1365.717050    0.002800    
0.121750  
50%    311.507000  258.344000  528.356000  1365.980900    0.005750    
0.248000  
75%    316.979000  267.031000  540.524250  1366.363250    0.012600    
0.407250  
max    322.182000  271.494000  543.813000  1367.316200    0.149400    
0.739000  
```

In [6]:
```
# Compute the correlation matrix
corr_matrix = df.corr()

# Print the correlation matrix
print(corr_matrix)
```

```
              Year     Month       MEI       CO2       CH4       N2O  \
Year      1.000000 -0.025789 -0.145345  0.985379  0.910563  0.994850   
Month    -0.025789  1.000000 -0.016345 -0.096287  0.017558  0.012395   
MEI      -0.145345 -0.016345  1.000000 -0.152911 -0.105555 -0.162375   
CO2       0.985379 -0.096287 -0.152911  1.000000  0.872253  0.981135   
CH4       0.910563  0.017558 -0.105555  0.872253  1.000000  0.894409   
N2O       0.994850  0.012395 -0.162375  0.981135  0.894409  1.000000   
CCl3F     0.460965 -0.014914  0.088171  0.401284  0.713504  0.412155   
CCl2F2    0.870067 -0.001084 -0.039836  0.823210  0.958237  0.839295   
TSI       0.022353 -0.032754 -0.076826  0.017867  0.146335  0.039892   
Aerosols -0.361884  0.014845  0.352351 -0.369265 -0.290381 -0.353499   
Temp      0.755731 -0.098016  0.135292  0.748505  0.699697  0.743242   

             CCl3F    CCl2F2       TSI  Aerosols      Temp  
Year      0.460965  0.870067  0.022353 -0.361884  0.755731  
Month    -0.014914 -0.001084 -0.032754  0.014845 -0.098016  
MEI       0.088171 -0.039836 -0.076826  0.352351  0.135292  
CO2       0.401284  0.823210  0.017867 -0.369265  0.748505  
CH4       0.713504  0.958237  0.146335 -0.290381  0.699697  
N2O       0.412155  0.839295  0.039892 -0.353499  0.743242  
CCl3F     1.000000  0.831381  0.284629 -0.032302  0.380111  
CCl2F2    0.831381  1.000000  0.189270 -0.243785  0.688944  
TSI       0.284629  0.189270  1.000000  0.083238  0.182186  
Aerosols -0.032302 -0.243785  0.083238  1.000000 -0.392069  
Temp      0.380111  0.688944  0.182186 -0.392069  1.000000  
```

This will give us a list of variables sorted by their correlation coefficient with Temp, with the strongest positive correlations at the top and the strongest negative correlations at the bottom.

In [7]:
```
# Print the variables that have a strong correlation with Temp
print(corr_matrix['Temp'].sort_values(ascending=False))
```

```
Temp        1.000000
Year        0.755731
CO2         0.748505
N2O         0.743242
CH4         0.699697
CCl2F2      0.688944
CCl3F       0.380111
TSI         0.182186
MEI         0.135292
Month      -0.098016
Aerosols   -0.392069
Name: Temp, dtype: float64
```

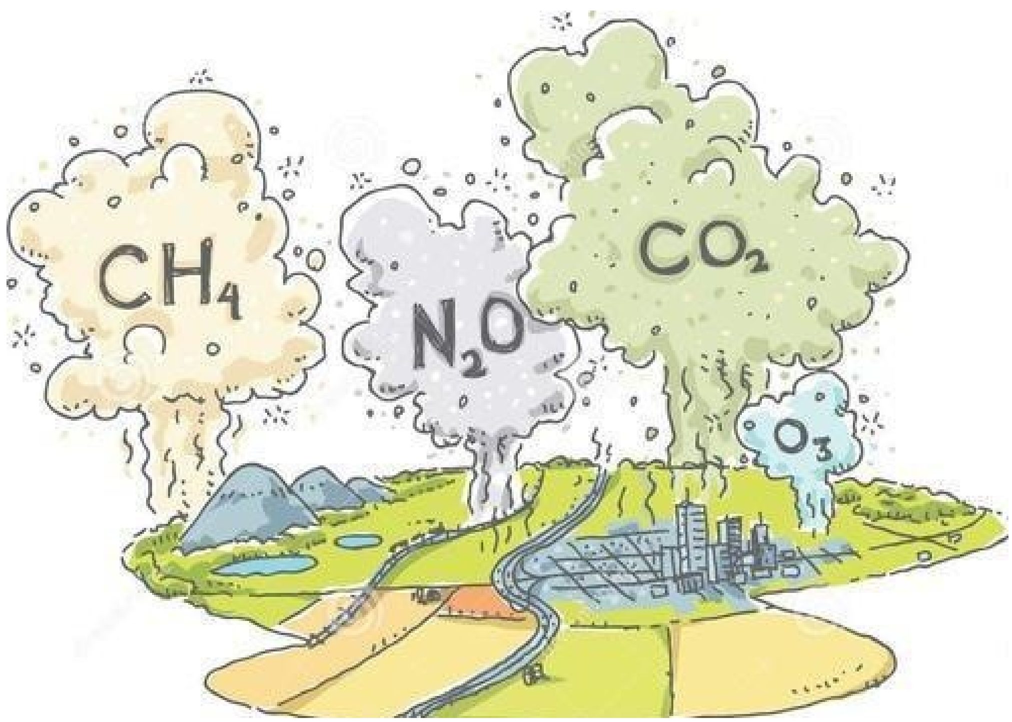
Based on the results, we can see that year, CO2, N20. CH4 have the highest correlation with temperature. This suggests that these variables may have a strong influence on global

temperature and should be further investigated in climate change research

Methodology

Data Cleaning

First, we will split dependent variable Temp as y and its factors as x.



So, let's analyze this relationship using some kinds of data science methology.

Data Science

Random Forest

```
from sklearn.ensemble import RandomForestRegressor
from sklearn import metrics
from sklearn.model_selection import train_test_split
```

Adaboost

Apart from Random Forest, we can also try adaboost regression. Adaboost is one of the most famous boosting algorithm due to its simplicity and high accuracy.

LASSO regression

In [9]:
```
from sklearn.linear_model import Lasso
from sklearn.linear_model import LassoCV
```

Generalized linear model

Other data minging technical

Discussion

Future work

Conclusion

In [ ]: