PROPOSAL VIDEO PRESENTATION

MENGYANG LIU
ID: 1006702739

INTRODUCTION TO MY RESEARCH

1. My research question and reason why I choose it.
2. Why it can be a interest to others.

1. MY RESEARCH QUESTION AND WHY.

My research question is related to a popular environmental problem mentioned these years which is the air pollution.

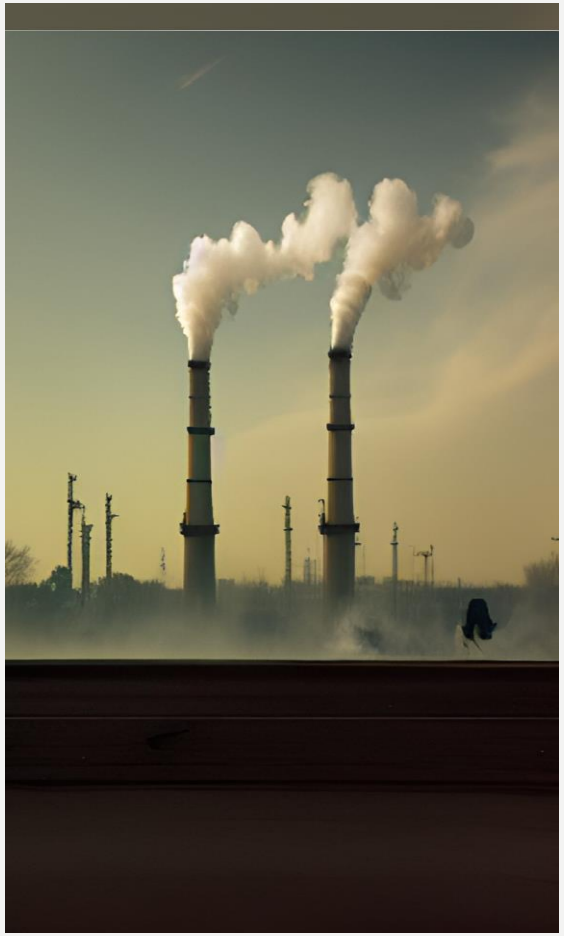
I'm interested to discover the relationship between emission amount of a few kinds of pollution gases and the effect that air pollution is making.

My research question is decided to be ***"Could the emission amount of polluted gases be used to predict the effect of air pollution and how can I do such prediction?***

2. WHY IT CAN BE A INTEREST TO OTHERS.

Water Pollution

Land Pollution



Air Pollution

As one of the three major types of pollution, air pollution is being seriously noticed by a lot of people.

Many countries have already taken action on reducing air pollution such as formulating related laws and controlling polluted gases emission.

My topic is aimed to find a relationship between polluted air and the pollution effect which could provide a direct view of how air pollution affects human beings' life.

PREVIOUS STUDIES

1. How popular is my topic?

2. Two examples of past academic papers related to my topic.

3. What are the differences and how can I incorporate these examples to my project.

1. HOW POPULAR IS MY TOPIC?

By searching the related papers of air pollution on U of T library website, I get **1664** results.

By searching on google scholar, I get **161,000** results and **520** of these results are mainly focused on air pollution exposure and effect

These results are not surprising since air pollution is a great topic and it also reflects that a lot of scholars are interested in this field.

2. THREE EXAMPLES OF ACADEMIC PAPER

The first example of academic study is researching Greenhouse gases and air pollution. It provides several kinds of GHG that are factors causing global warming and climate change.

The gases in the paper are the same as my topic variables. Picture 16 in the article shows that linear equations could be used as an estimation to deaths due to Outdoor Air Pollution. (Barbosa, 2022)

A surprising result from this report is "that are not absolute but have the intention to show a relationship between GHG and Deaths by Outdoor Air Pollution" (Barbosa, 2022) This suggests that in my topic, it might generate a weak relationship between air pollution effect and GHG.

2. THREE EXAMPLES OF ACADEMIC PAPER

The second example is an article about air pollution exposure and cardio-respiratory mortality. This article gives a report of the relationship between air pollution exposure and cardiovascular disease. (Hoek, 2013)

By recognizing cardiovascular disease as an example of the air pollution effect, this article has the same interest variables as my research topic. The air pollution exposure is also related to my predictor variables of polluted gases emission.

In addition, it is aimed to find some relation or causation between disease and air pollution. The result of this article is "Air pollution is an important public health issue, causing cardiovascular and pulmonary diseases worldwide." (Hoek, 2013) Therefore, it provides evidence for my question.

2. THREE EXAMPLES OF ACADEMIC PAPER

The last example is an article which explores the cardiovascular effect of air pollution focused on a more biological perspective of view.

The article discusses many biological pathways whereby PM may cause cardiovascular problems. CV disease could be seen an exchangeable variance of air pollution effect. Also, because VOC plays a significant role in the formation of ozone and PMs. (Brook, 2008) The variable of VOC value in my datasets will be significantly related to disease even death.

The result is that "PM air pollution is associated with an increased risk of CV morbidity and mortality." (Brook, 2008) which suggest that VOC emission and air pollution have relationship. A surprised result is that humans' health will still be threatened just in a short-term exposure to PM air. (Brook, 2008) This conclusion indicates that the VOC emission and air pollution effect may have a stronger relationship although the value of VOC is relatively small.

3. DIFFERENCES AND HOW TO INCORPORATE THEM WITH MY WORK

My research question is aimed to discover the linear relationship between the air pollution effect and factors which are constituent parts that cause air pollution. Mine is focusing on a comprehensive field rather than being specific to only one pollutant.

I will use elements mentioned in most of the article and combine them to generate a linear relationship between the effect and its most possible constituents of gases. I will adopt the results from the past academic paper to make my analysis stand reasonably.

INTRODUCTION TO MY DATASET

1. Where my data come from.
2. How I clean my dataset.
3. My dataset characteristics and limitation.

1. WHERE IS MY DATA FROM.

I find my datasets on OECD data provided by the link of the instruction document. There are two datasets that I'm going to use.

One is a dataset of air pollution effects from https://data.oecd.org/air/air-pollution-effects.htm#indicator-chart.

And another one is a dataset of air and GHG emissions from https://data.oecd.org/air/air-and-ghg-emissions.htm#indicator-chart.

These two datasets are from a research named "Environment at a Glance" being published in 2020. As they are in the same research, their origins are exactly the same.

2. HOW I CLEAN MY DATASET.

There are 7 datasets in total containing the variables that I'm interested in. Six of them are data of carbon dioxide, carbon monoxide, green house gas, nitrogen oxide, sulphur oxide and volatile organic compound with each kinds of gas as an independent dataset. The one left is a dataset of air pollution effect. There are 8 variables in every dataset which are location, indicator, subject, measure, frequency, time, value and flag code. All gas emission data are in the same unit of thousand tonnes except for CO2 with million tonnes.

Since I'm going to use data from separated datasets, I need to put them together as a whole to do further analysis. The original datasets have a lot of missing data such as missing years and missing values. This becomes a great obstacle when I was combining and cleaning these datasets.

2. HOW I CLEAN MY DATASET.

To solve the problems I mentioned in the last slides and create a cleaned and complete dataset, I have done a few steps to do so.

1. I only choose to have 10 countries for each datasets which contain no missing data from year 2008 to 2019 so that all these elements in my data are available.

2. I change the units of gas emission value from thousand tonnes to million tonnes by dividing by 1000 for each variable of value.

3. After then, I get 7 datasets with exactly the same formats and variables. Therefore, I could combine them together.

3. MY DATASET CHARACTERISTICS AND LIMITATION.

After doing the cleaning, process, I get a dataset that I'm going to use for linear regression. This datasets has **120** observations.

It has **9** variables which are location, time, the effect value and the emission value of carbon dioxide, carbon monoxide, green house gas, nitrogen oxide, sulphur oxide and volatile organic compound with each one as a single variable.

The limitation of my data is that the location variable of countries cannot be a representation of the whole world, so it may not lead to a accurate result for a global case. What's more, since the time I choose is only from 2008 to 2019, these 12 years interval can be biased and make some errors.

FITTING LINEAR MODEL BY MY DATASET

1. Using linear model to answer my research question.
2. Check assumptions of linear regression model and problems may be raised in my analysis.

1. USING LINEAR MODEL TO ANSWER MY RESEARCH QUESTION.

As a recap, my research question is ***"Could the emission amount of polluted gases be used to predict the effect of air pollution and how can I do such prediction?***

An idea is to see if a linear model can be formulated so that I could predict the air pollution effect when I'm given the emission value of several polluted gases. That is to say, I'm going to explore my research question by setting the air pollution effect to be my interest variable and those six types of polluted factors in my dataset be predictors for the linear model.

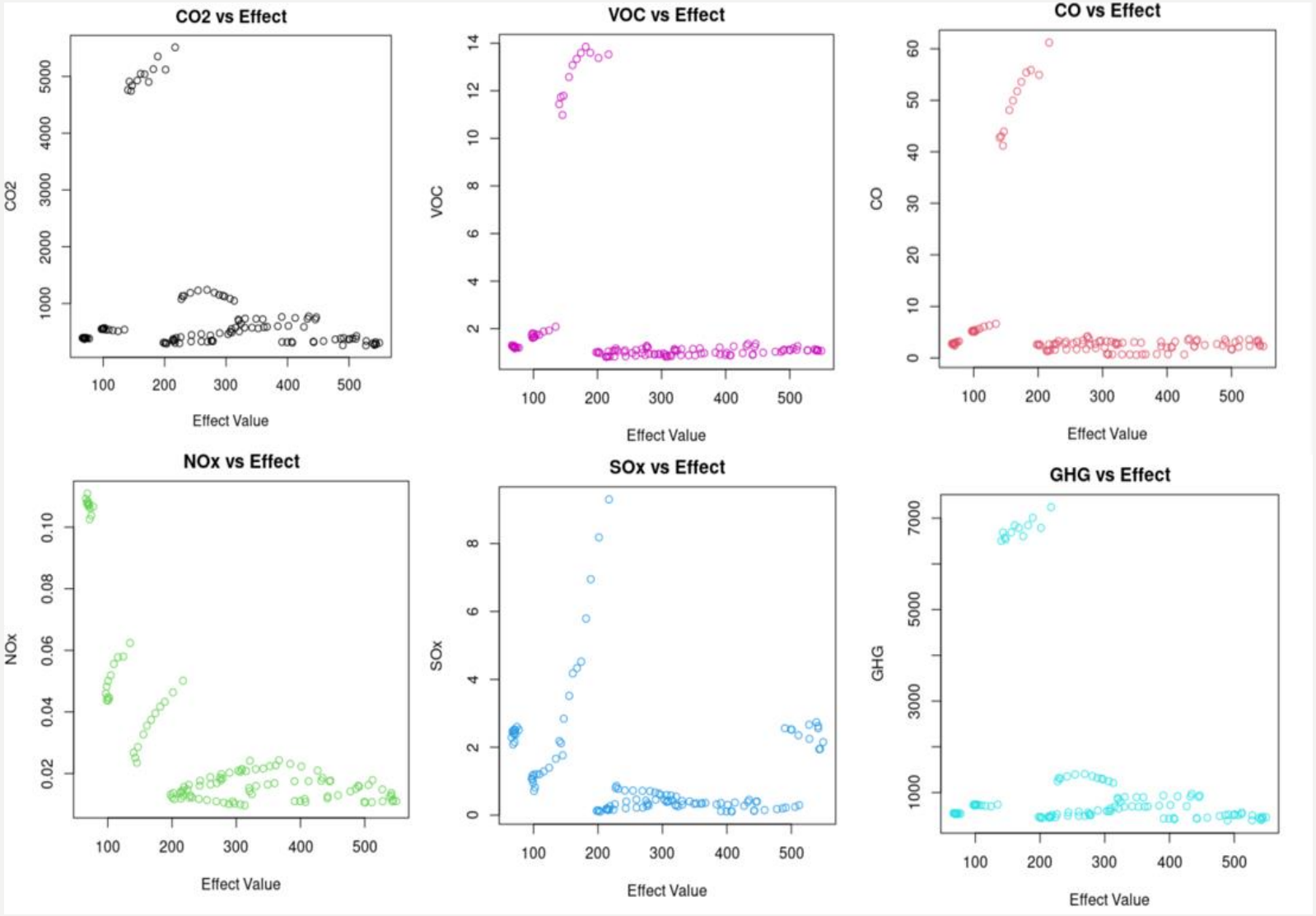
2. ASSUMPTIONS OF LINEAR REGRESSION

Are there uncorrelated errors or missing data?

Each observation is independent from others distinguished by the measuring times and countries. Also, as each variables are different gases emission value, they are not related. Therefore, it fitted the assumption of independence of linear regression.

By cleaning my data, there is no missing data and non-sense observations in the dataset I'm using for generating my linear model.

2. ASSUMPTIONS OF LINEAR REGRESSION

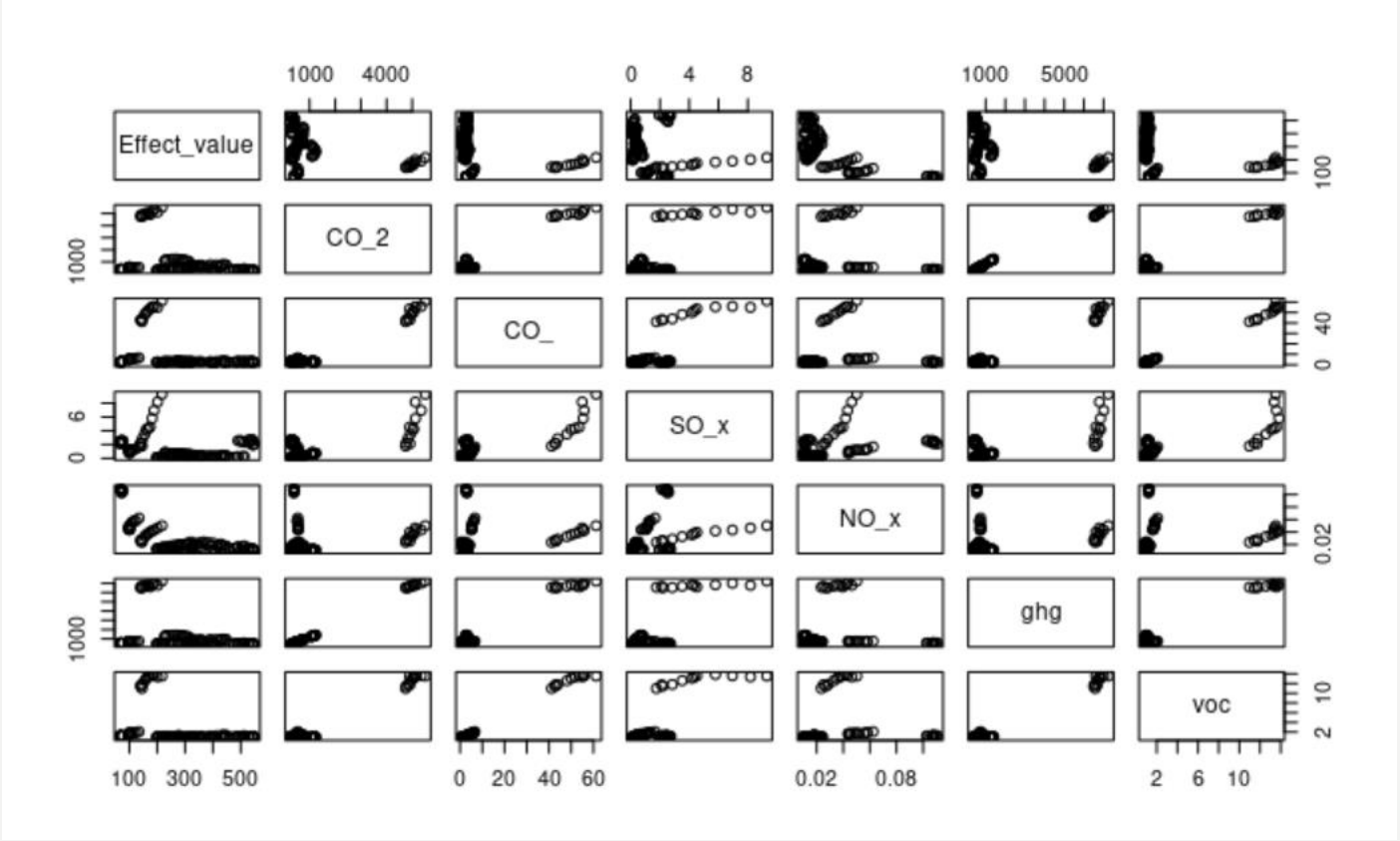


These graphs are used to check linearity between each predictor and our interest. Each variable appears to be linearly related with the response except nitrogen oxide and Sulphur oxide.

However, there are outliers when predict the air pollution effect values between the value from 100 to 200 with each variable and the non-linearity of nitrogen oxide and Sulphur oxide also happen in this interval.

Therefore, it might have some problems within my linear regression model in this prediction interval.

2. ASSUMPTIONS OF LINEAR REGRESSION

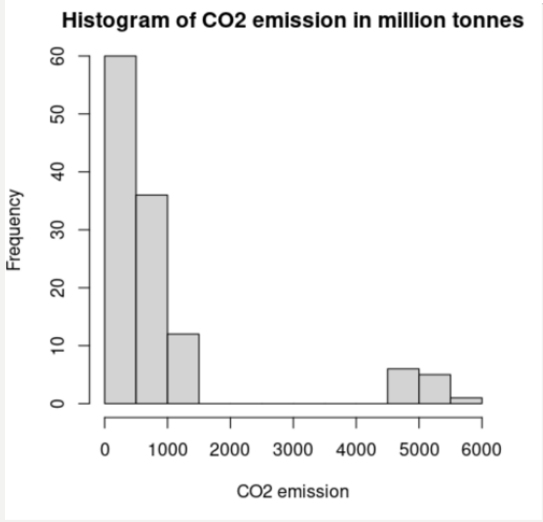


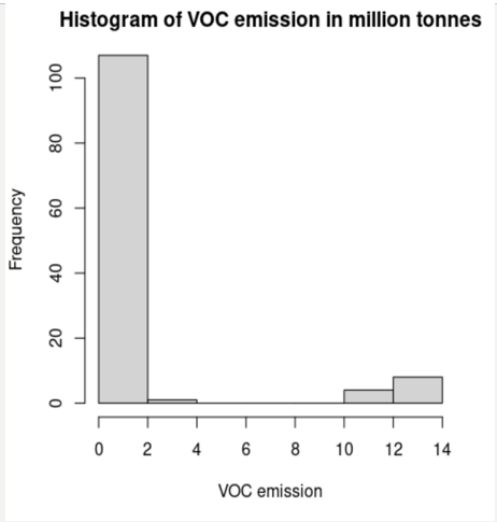
Then I need to check if my data fulfill the constant variance assumption of linear regression.

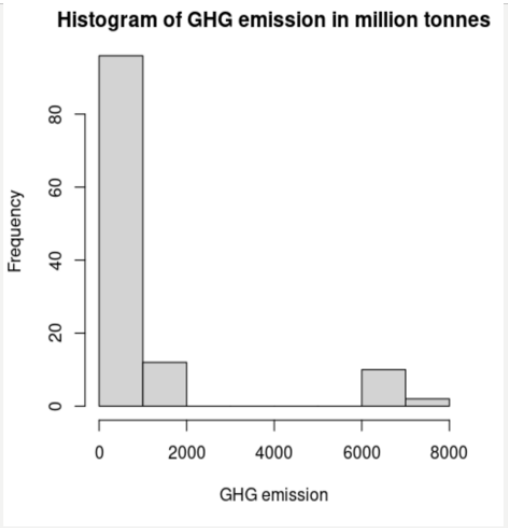
As the graph indicated, there is a constant variance when having effect value and each of gas emission value. However, it is obvious that the graphs of each gas emission values are having increasing variability.

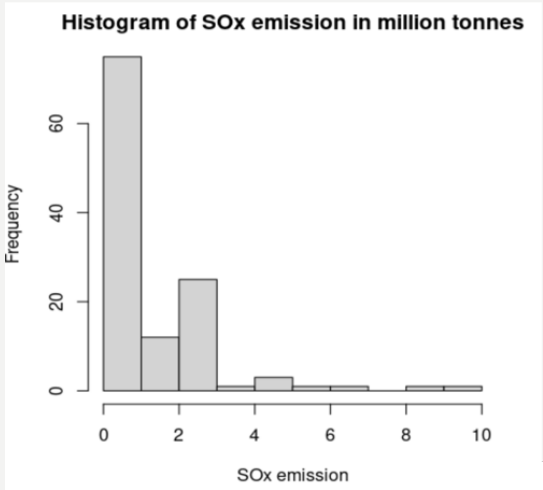
Therefore, the assumption of constant variance maybe violated and cause problem in further analysis

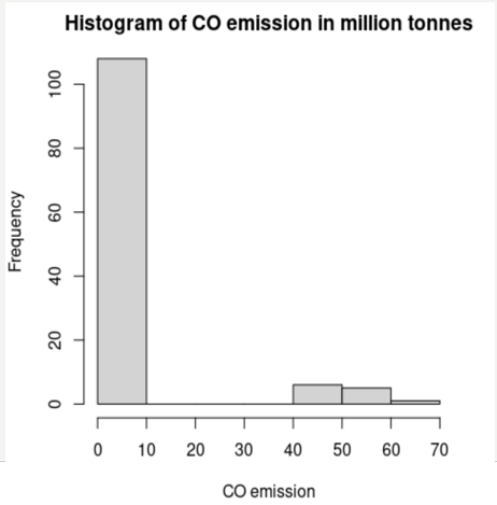
2. ASSUMPTIONS OF LINEAR REGRESSION

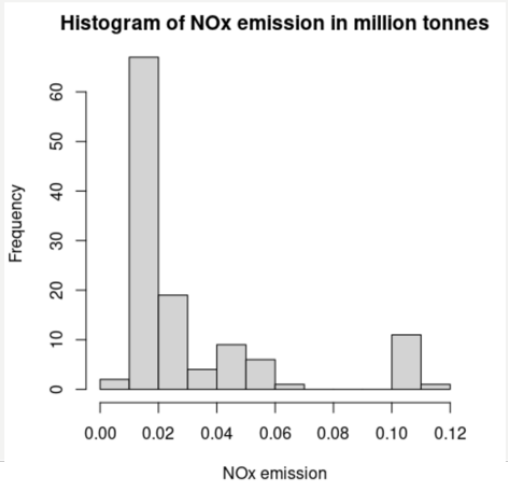












These graphs are used to check the normality assumption of linear model.

The distribution shape of all predictor variables are generally right-skewed. It mean that they are not normal distribution.

However, marginal distribution can be non-normal.

REFERENCE

Barbosa Lima, Leandro Jose & Hamzagic, Miroslava. (2022). Greenhouse gases and air pollution: commonalities and differentiators. Revista Científica Multidisciplinar Núcleo do Conhecimento. 6. 102-144. 10.32749/nucleodoconhecimento.com.br/environment/green-house-gases. https://www.researchgate.net/publication/363885205_GREENHOUSE_GASES_AND_AIR_POLLUTION_COMMONALITIES_AND_DIFFERENTIATORS

Hoek, G., Krishnan, R. M., Beelen, R., Peters, A., Ostro, B., Brunekreef, B., & Kaufman, J. D. (2013). Long-term air pollution exposure and cardio-respiratory mortality: a review. *Environmental health*, *12*(1), 1-16. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=air+pollution+effect+and+exposure&btnG=

Brook, R. D. (2008). Cardiovascular effects of air pollution. Clinical science, 115(6), 175-187. https://doi.org/10.1042/CS20070444

Pictures are from https://www.wombo.art/create

THANKS FOR WATCHING!