Do We Need More Babies to Save the Planet?

A Study on Greenhouse Gas Emissions & Population Growth

MGT 6203 Final Report - Team 64

Melissa Gibson, Stephen Kim, Nitya Nandagopal, and Jessica Peterson

Table of Contents

1. Introduction

The past century has been marked by transformative demographic and environmental changes all over the world. Until the industrial revolution, average life expectancy was short, due to disease, lack of technology, and society at the time. The industrial revolution brought a shift with medical and technological advancements, social changes such as worker's rights, feminism and the changing labor market. Life expectancy increased exponentially, and with that came increased population growth and production. Per Professor Scott Galloway of NYU, "in the past two hundred years, per capita GDP has grown 15x; we now live twice as long as our great-grandparents, and our population is up eightfold — from 1 billion to 8 billion" (Galloway, 2023). A larger, longer-living population has yielded bigger problems to address, such as increased income inequality, global warming, and climate change, all of which have widespread implications for the futures of the planet and humanity.

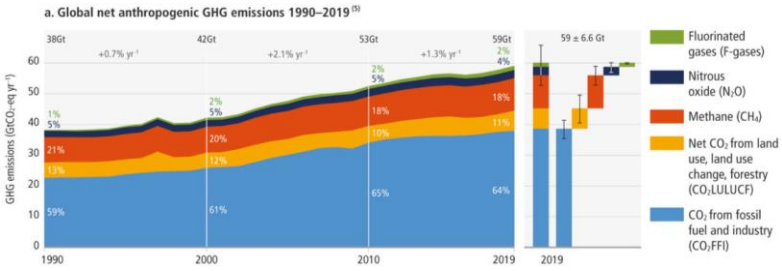

Figure A: Rise of GHG over Time

The increase in greenhouse gases (GHG) since the industrial revolution has undeniably impacted the environment detrimentally, with carbon emissions being the majority of the GHG emissions released into the Earth's atmosphere. Over the last 40 years, "CO2 emissions have increased by about 90%, with emissions from fossil fuel combustion and industrial processes contributing about 78% of the total [GHG] emissions increase from 1970 to 2011," (EPA, 2023) (**Figure A**). Agriculture, deforestation, and other land-use changes are the second-largest contributor to GHG emissions (EPA, 2023).

Although the world's population has increased rapidly over the past century, the overall population growth rate is currently slowing. Population growth rates have mainly declined due to the sharp decline in the global total fertility rate, from 5.3 in 1963 to 1.8 in 2022 (Galloway, 2023). A reflection of this can be seen with America's population. Over the last 100 years, America's population grew from 77 million in 1901 to over 330 million in 2020 – a total growth of 330% over 119 years (Tallungs, 2022). Although the overall population has grown, America's population growth rate was only 0.4% in 2022, the lowest recorded population growth rate in the country's history (Galloway, 2023). Most regions of the world are also experiencing slowing population growth as nations advance technologically, change child labor laws, and see a change in social values, the average family sizes are shrinking.

Population growth is commonly thought to negatively impact our global environment due to the strain on resources and increased production necessary to support a larger population. These pressures of population growth resulted in population control measures imposed by governments over the past 50 to 100 years; however, due to these policies and other changing values of modern society, many countries are facing the threat of population decline by the end of the century. With the global population soon to be entering a decline, its effects on GHG emissions are still unknown. Solely blaming population growth for climate change might not be fair, as there are many factors that contribute to GHG emissions. Increased innovations and technology could reduce GHG emissions to improve the health of the planet.

A question arises as to whether population significantly contributes to GHG emissions. If the world population shrinks considerably, and younger generations must support the larger, aging population, who will continue to improve the future health of the planet? Without a robust population to carry on the work and innovations of prior generations, how will the world continue to develop? The purpose of our investigation and analysis is to examine if and how population growth contributes to GHG emissions.

2. Literature Review

Although researchers have explored the relationship between climate change and population growth, there are conflicting opinions about it. Many scientists have used the "Kaya identity," a formula that states emissions are a product of population, per capita GDP, energy used per unit of GDP, and CO2 generated per unit of energy, to explain the increase in carbon emissions (D'Souza, 2022). Studies that adhere to this formula stipulate a positive correlation between population and emissions, stating that nations with larger populations yield higher levels of GHG emissions. However, a positive relationship between them does not necessarily prove causation. In fact, many scientists argue that "suggesting a straightforward positive relationship between population and climate change ignores the influence of income inequality on the interaction between population and emission levels" (D'Souza, 2023). To ignore this complexity would be counterproductive to finding a solution to the climate change problem.

To try and capture the influence of income inequality, researchers have used another formula, known as the "IPAT" equation: Environmental Impact = Population x Affluence x Technology (Kaplan, 2021). Affluence is defined as GDP per capita, and technology is a measure of the amount of resources required to produce a unit of GDP. Although this equation also relates population and emissions positively, some researchers argue that affluence and technology hold greater weights than population. U.N. reports show that global resource use has primarily been driven by increases in affluence, not population (Kaplan, 2021). A small minority of wealthy people produce far more than their share of carbon emissions per capita, but the poor are directly affected by climate change. The richest 10% are responsible for up to 50% of the global GHG emissions, while the poorest 50% are responsible for only 10% (Oxfam, 2015).

Climate change is happening now, and its victims around the world are being displaced from homes due to flooding, drought, and wildfires. In 2021, more than 7 million acres in the United States burned in wildfires (The Center for Biological Diversity). Between 2000 and 2015, the global population at risk of flooding increased by 20-24 percent (Zurich, 2021). Furthermore, 700 million people are at risk of being displaced because of drought by 2030, according to the WHO (Zurich, 2021). These effects have widespread implications for food production, livestock/crops, as well as the availability of habitable land.

Other observations imply that population growth is not the direct cause of GHG emissions. According to Our World in Data, the United States is the leading nation in total GHG emissions, despite being the third largest country by population (Ritchie & Roser, 2020). Regions that have more people do not necessarily emit more carbon. For example, Kenya has 55 million people, 95 times more than the population of the state of Wyoming, but Wyoming emits 3.7 times the amount of carbon dioxide as Kenya (Chang, 2022). Between 1900 and 2000, carbon emissions increased up to fifteen times, while population increased less than four times (Cohen, 2010). This indicates that there may be other factors at play affecting the levels of GHG emissions released into the atmosphere. The world economy grew up to sixteen times over this same period, and with that came massive increases in burning fossil fuels to support production and transportation needs (Cohen, 2010). As such, we would like to explore the relationship between GHG emissions and other factors in addition to population growth.

3. Hypothesis

It is commonly thought that population growth negatively contributes to climate change and other environmental problems. The purpose of our analysis is to examine if and how population growth contributes to GHG emissions. Our hypothesis is that population growth does not significantly contribute to increased GHG emissions. In fact, we do not expect population to be the largest contributing factor to the increase in GHG emissions.

4. Overview of Datasets

The primary datasets used by the team are Greenhouse Gases Emissions Data (1990 – 2019), which lists GHG emissions per country, and World Population Growth Data (1962 – 2021), which lists population data per country. Since these are the two main variables we are trying to compare in our analysis, these are the two primary datasets we are using. We built our working dataset upon this foundational data for our project, where each observation in the dataset is the applicable data per country for a given year.

To perform regression on GHG emissions as the dependent variable using population growth as one of the independent variables, we searched for additional independent variables to enhance our current working dataset. Since our dataset looks at emissions on a per country level, we considered other factors that also had that level of granularity. For example, industrial production within a country might affect its GHG emissions. To effectively measure production, we could look at a country's GDP, employment rate, land usage, exports of goods and services, and manufacturing volume. Another factor we considered was a country's technological advancement. We theorized that technologically advanced countries could better combat the negative effects of GHG emissions. To try and capture this data, we looked at a variety of possible factors like the volume of patent applications of a given country and their education rates.

Additional datasets we collected included information on the following: Age dependency ratio, agricultural land (as % of land area), air transport (freight), air transport (passengers), birth rate (per 1000 people), life expectancy at birth, fertility rate, exports of goods and services, GDP in US$, GDP growth rate, manufacturing value added, patent applications, natural resources rents, unemployment rate, labor force, energy consumption per capita, government expenditure on education, new vehicle registrations by type, researchers in R&D per million people, and death rate from air pollution. All these additional datasets showed year over year data of each factor for the countries listed in our primary dataset.

5. Data Cleaning Process

Our first step was to merge world population and GHG emissions datasets by country and year as the primary key. Since these datasets only provided values for one factor for this key, only a few data manipulations were required to merge them. First, we converted the years to dates in the GHG emissions dataset. Since the population dataset had the years as columns, we used the 'melt' function in Python to transform the column years into rows per country. Then, we merged the two datasets on Country Code, Country Name, and Year as the primary key to create our working combined dataset. After our main datasets were merged, we merged the additional independent variable datasets with our main dataset with the same primary key.

During our exploratory data analysis, we noticed that the initial GHG dataset contained some region values that were a combination of country values, such as the "Rest of the World" or "South East Asia Region". While these are interesting geographical breakdowns, these combined regions are outliers and were not incorporated in our regression (**Figure B**). Removing these combinations decreased the number of "countries" in our dataset from 350 to 270. It should be noted that this dataset includes some territories that are not fully recognized as independent countries.

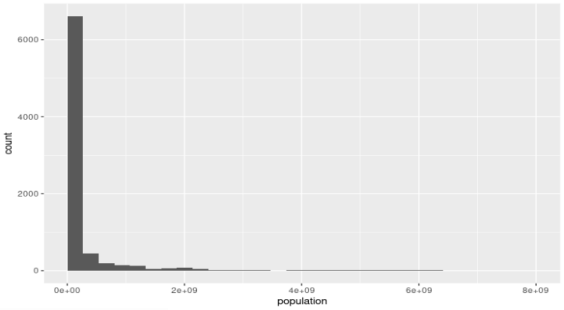

Figure B: Population Histogram

Additionally, we noticed that there were several countries that had a lot of missing values. Certain countries had "NA" values for entire variables or were missing more than 20 years of data. As such, we disregarded these countries with insufficient data for analysis. After removing these countries, we were

left with 147 countries and about a total of 55 remaining "NA" values in the dataset. As these missing values were less than 10% of our total dataset, we imputed them by country based on varying methods. For example, the country Georgia had some "NA" values for primary energy consumption per capita and natural resources rents. To impute missing primary energy consumption per capita, we used the growth rates of surrounding years as we could see a trending decline in the surrounding years. For natural resources rents, we used an average of the surrounding years. Turkmenistan is another country with missing data, specifically for labor force. Since the labor force data for the country varied over the 29 years of data, we imputed based on an average of the total for that variable. We utilized these same imputation methods of averages or growth rates for Luxembourg, Belarus, Tajikistan, and Armenia.

In addition to finding countries that did not meet the team's requirements, some factors we considered only addressed a small fraction of the countries we wanted to analyze. We totaled the number of blank values per factor to identify what should be excluded due to insufficient data. The original master dataset had twenty-one factors (excluding country-identifying information and GHG and population data). Out of these factors, eleven had data for less than half of the countries. Specifically, vehicle-related data (e.g., the number of electric vehicles), yielded about 290 valid data points out of the total combined dataset value of around 7900. This is only 3% of valid data out of the whole dataset. As such, removing these factors was necessary as it did not address the global reach of our data.

For the final dataset, we removed the air transport, air travel, birth rate, mortality rate, vehicle information, researchers per million, patent applications, and unemployment data from our dataset due to insufficient data. In addition, we removed GDP growth, manufacturing and exports as these variables were highly correlated with a correlation coefficient of 0.87. Please refer to the **Appendix** for a summary and explanation of variables used and see **Figure C** below for a snippet of the final combined dataset.

Index	Country_Cod	Country_Name	Year	age_dependency_ratio	agricultural_	fertility_rate	GDP	labor_force	life_expectancy	natural_resources_rent	energy_consumption_per_capita	population	GHG_with_LUCF	deaths_air_pollute	GHG_without_LUCF
1	TJK	Tajikistan	1/1/90	88.6	32.07	5.34	2631768953	1372150	61.88	0.78	17274.67	5417860	17899999.62	192.5	17870000.84
2	TJK	Tajikistan	1/1/91	89.76	32.07	5.18	1352000000	1399115	61.38	0.65	15902.46	5556306	17010000.23	195.85	16969999.31
3	ARM	Armenia	1/1/91	55.7	41.1	2.65	2069870130	1645616	68.64	0	19601.42	3617631	25139999.39	134.21	25010000.23
4	GEO	Georgia	1/1/90	51.15	46.47	2.31	7753501868	2371540	68.39	0.48	24620.21	4802000	31399999.62	160.64	49759998.32
5	GEO	Georgia	1/1/92	52.86	46.47	2.13	3690328964	2443064	67.78	0.34	17979.8	4873500	14390000.34	161.89	32650001.53
6	LUX	Luxembourg	1/1/90	44.23	49.72	1.6	12778792854	157920	75.44	0.04	97439.53	381850	11550000.19	40.16	12189999.58
7	LUX	Luxembourg	1/1/91	44.82	49.72	1.6	13834219728	165345	75.46	0.02	103333.91	387000	12039999.96	38.6	12680000.31
8	LUX	Luxembourg	1/1/92	45.47	49.72	1.64	15518702635	172615	75.77	0.02	102274.26	392175	11739999.77	36.65	12380000.11

Figure C: Snapshot of Final Dataset

6. Exploratory Data Analysis (EDA)

To get a better understanding of the GHG emissions data and population data, we created preliminary histograms that showed each of the respective variables against their counts (**Figure D**). We also created an exploratory variable GHG/population to calculate the GHG per capita. Through this analysis, we noticed that the initial GHG emissions dataset we pulled had negative GHG emissions values. After doing further research, we realized that this dataset takes land use and forestry changes (LUCF) into account. Countries with negative GHG emissions value during certain years may have reallocated the usage of land or replenished forests, resulting in negative net emissions as the benefit exceeded the costs. We were able to find GHG emissions data without LUCF from the same data source to incorporate in our dataset. Per a literature review, there is debate among experts about the accuracy and precision of LUCF data before 2010. As we were unsure about LUCF's importance, we created two models, using GHG with and without LUCF as the dependent variable, to determine the model with a better fit.

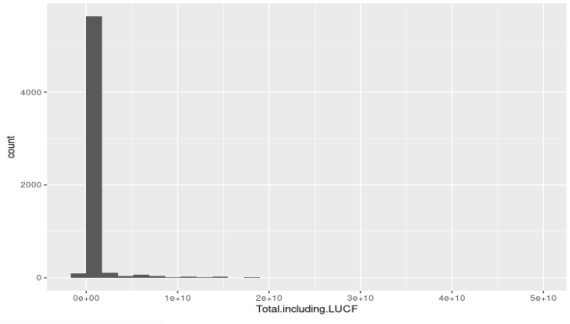

Figure D: GHG with LUCF Histogram

To gain a better understanding in general between population, GHG, and other potentially significant factors, the team created a correlation plot of all the variables used in our models (**Figure E**). This included the dependent variables: GHG with and without LUCF, and the independent variables: GDP, population, labor force, agricultural land, fertility rate, age dependency ratio, deaths by air pollution, natural resources rents, energy consumption per capita, and life expectancy. Based on the correlation plot, it appears as though population, GDP and labor force are highly correlated with GHG. Meanwhile, fertility rate, age dependency, and deaths from air pollution are slightly negatively correlated with GHG.

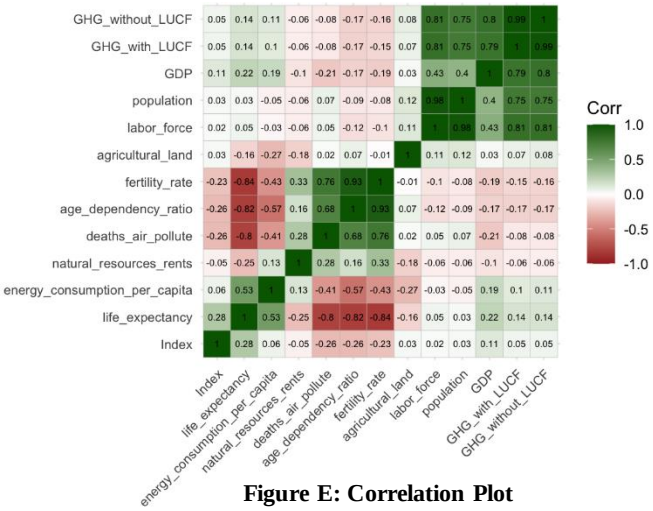

Figure E: Correlation Plot

Since our hypothesis primarily focused on analyzing the relationship between population growth and GHG emissions, we wanted to quickly visualize how these two factors relate to each other as they change year over year. To do so, we created a random sample of 15 countries and tracked the changes in GHG and emissions for them from 1990 to 2018. A quick animation was created in R that showed all 15 countries and where they fell on a GHG Emissions (without LUCF) vs Population graph and how their positioning changed over the years. **Figure F** below shows a snapshot of the animation of one such random sample of countries in 1990. **Figure G** shows a snapshot of that same random sample of countries and where they fell in 2018.

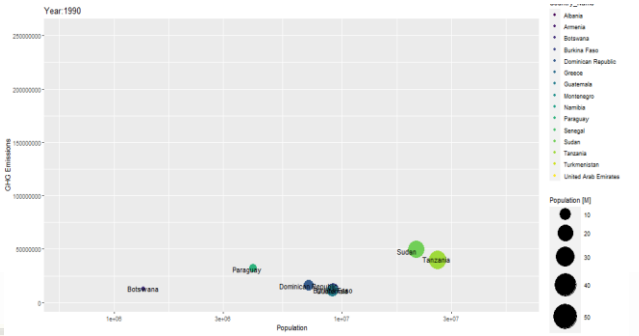

Figure F: GHG vs Population (1990)

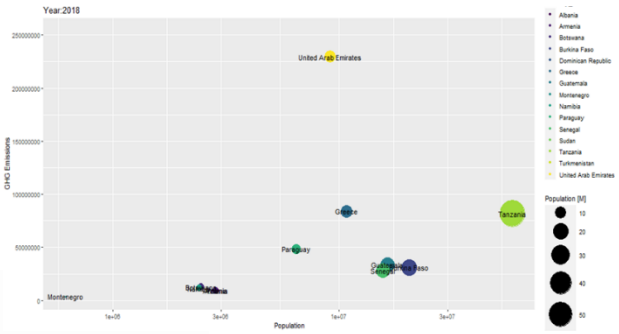

Figure G: GHG vs Population (2018)

These plots show that while an increase in population can relate to an increase in GHG, there are also instances where a country's population remains relatively constant, but the emissions significantly increase over a two-to-five-year span, like Tanzania and UAE, respectively, in the two figures above.

Taking our analysis of the relationship between population and GHG to the next level, we ran preliminary scatterplots plotting population versus GHG. This was done for both GHG with LUCF and GHG without LUCF to see if there was a difference between them (**Figure H**). The scatterplots were nearly visually identical, although the scatterplot with LUCF shows negative GHG values and more variation in the spread of the data. Summary statistics on both GHG with and without LUCF show that the spread of GHG with LUCF was much larger as it contained higher maximum and negative minimum values.

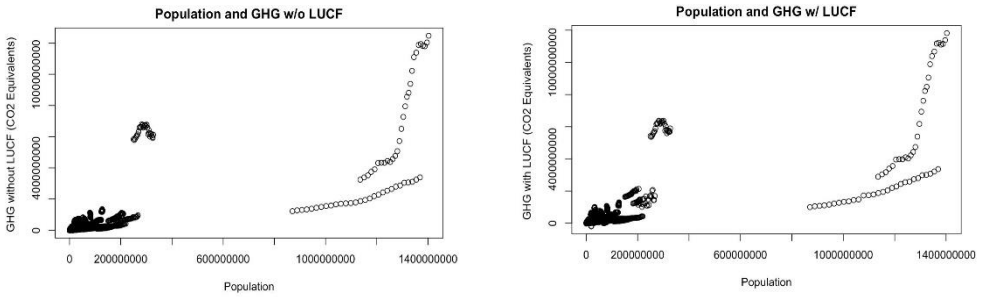

Figure H: Scatterplots of GHG (without and with LUCF) vs Population

For a final analysis, we calculated the average GHG emissions without and with LUCF by country to identify the top 10 emitting countries. As expected, the top three countries were China, the United States and India respectively (**Figure I**). The analysis resulted in the same top 10 emitters regardless of LUCF as a factor; however, with LUCF, the ranks of Brazil and Indonesia are swapped with Japan and Germany's.

Top 10 Emitters

	Country Name	Avg GHG without LUCF
1	China	7,338,392,418
2	United States	6,323,145,154
3	India	2,088,687,938
4	Japan	1,260,978,958
5	Germany	945,082,412
6	Brazil	855,693,105
7	Indonesia	704,072,415
8	Canada	648,563,104
9	United Kingdom	638,088,269
10	Australia	568,306,210

	Country Name	Avg GHG with LUCF
1	China	6,902,467,319
2	United States	5,931,351,041
3	India	1,996,045,858
4	Brazil	1,724,079,657
5	Indonesia	1,427,456,547
6	Japan	1,186,609,653
7	Germany	911,357,931
8	Canada	805,756,893
9	United Kingdom	626,850,344
10	Australia	596,173,449

Figure I: Top 10 GHG Emitters Table

As shown in the plots below, GHG emissions with and without LUCF have similar distributions, but there is more variability in **Figure J**, which shows GHG with LUCF. All three plots show China, India, and the United States as the top three in GHG emissions with LUCF, GHG emissions without LUCF, and population (**Figures J**, **K**, and **L**). Based on the visuals of these plots, population rates in each country have increased at a relatively consistent rate, but GHG emissions have increased at a much higher rate.

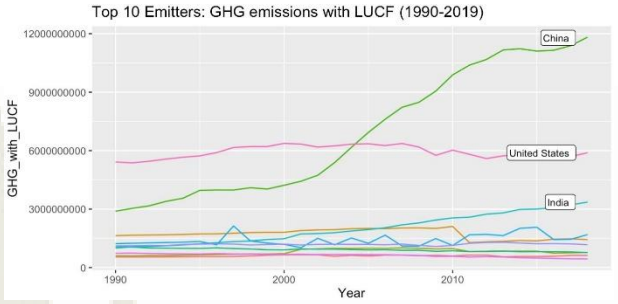

Figure J: Top 10 Emitters – GHG with LUCF

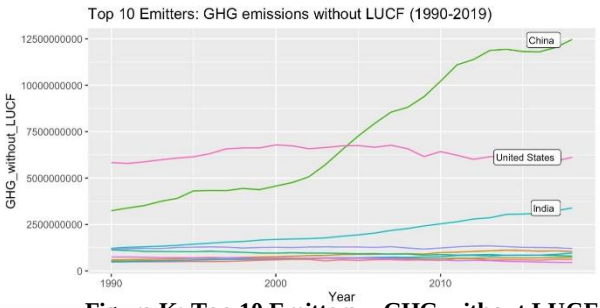

Figure K: Top 10 Emitters – GHG without LUCF

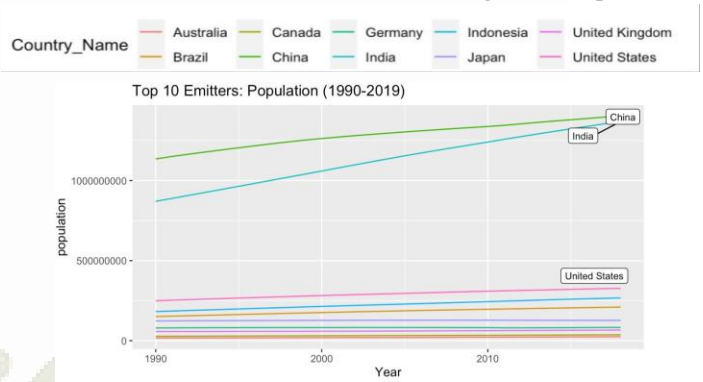

Figure L: Top 10 Emitters – Population

7. Methodology

To examine the relationship between GHG emissions and population, we used multi-linear regression. Several different models were built and compared using their performance metrics to determine which model had the better fit. The dependent variables used in our models were GHG emissions with and without LUCF. The independent variables include country, population, age dependency ratio, agricultural land area, fertility rate, GDP, labor force, life expectancy, natural resources rents, energy consumption per capita, and deaths from air pollution. Prior to running the regression, the numerical independent variables were scaled, so the regression coefficients could be compared on the same level.

Our models were trained on a randomized 80% of our data, and we used 10-fold cross-validation to validate these models. PCA and lasso regression were also used to try and reduce the number of variables involved in our model, as we had many independent variables to assess. A variety of different regression models were assessed to determine the best fit for the data without introducing any additional biases.

8. Results

We built the following multiple linear regression models to examine the relationships between our dependent and independent variables.

Table 1: All Models Created for Analysis

Model	Dependent Variable	Independent Variables	Cross-Validated R^2
Model 1	GHG_without_LUCF	Country name, population, age dependency ratio, agricultural land area, fertility rate, GDP, labor force, life expectancy, natural resources rents, energy consumption per capita, and deaths from air pollution	0.9642891
Model 1a (removed labor force)	GHG_without_LUCF	Country name, population, age dependency ratio, agricultural land area, fertility rate, GDP, life expectancy, natural resources rents, energy consumption per capita, and deaths from air pollution	0.9629279
Model 1b (removed Country)	GHG_without_LUCF	population, age dependency ratio, agricultural land area, fertility rate, GDP, labor force, life expectancy, natural resources rents, energy consumption per capita, and deaths from air pollution	0.9256678
Model 2	GHG_with_LUCF	Country name, population, age dependency ratio, agricultural land area, fertility rate, GDP, labor force, life expectancy, natural resources rents, energy consumption per capita, and deaths from air pollution	0.9467334
Model 3	GHG_without_LUCF	population	0.5662694
Model 4	GHG_without_LUCF	GDP	0.6784612

Model 5 (Removal of highly correlated variables GVIF>5)	GHG_without_LUCF	Country name, population, age dependency ratio, life expectancy, GDP, natural resources rents	0.9626004
Model 6	Global_GHG	Global population	0.9751
Model 7	GHG_without_LUCF	GDP, population	0.8577862
Model 8	GHG_without_LUCF	Country name	0.9397507
Model 9 (Removal of highly correlated variables VIF>5)	GHG_without_LUCF	GDP, population, agricultural_land, natural resources rents, energy consumption per capita, deaths from air pollution	0.8602951

The first step to figuring out the best model was to create two baseline models for the dataset, one that had GHG without LUCF as a dependent variable, and another with GHG with LUCF as a dependent variable. In these baseline models, Country_Name was used as a categorical variable, and every single independent variable was included. These models are listed as Model 1 and Model 2 in Table 1 above.

Upon comparing GHG_without_LUCF and GHG_with_LUCF, the model for GHG_without_LUCF had higher R^2 and Adjusted R^2 values. As such, we moved forward with GHG_without_LUCF as our dependent variable. The summary of Model 1 showed that population was not a significant variable, but as the multiple linear regression models trained on 80% of the dataset were producing extremely high R^2 values (~0.95), we theorized that there could be multicollinearity and overfitting. Using tools, like VIF, we confirmed that several of our factors were highly correlated (**Figure M**). Based on the VIF results for Model 1, labor force and population were highly correlated factors (GVIF>10). Other factors such as age dependency, agricultural land, fertility rate, life expectancy, energy consumption per capita and deaths from air pollution are also correlated (GVIF >5). As a result, we cross-validated several models (Models 1a, 5) without these highly correlated variables, but left population in our models as it is the primary variable of investigative interest.

	GVIF	Df	GVIF^(1/(2*Df))
Country_Name	2.480086e+11	146	1.094012
age_dependency_ratio	2.537748e+01	1	5.037607
agricultural_land	5.624852e+01	1	7.499901
fertility_rate	4.816827e+01	1	6.940336
GDP	8.430021e+00	1	2.903450
labor_force	1.192534e+03	1	34.533080
life_expectancy	2.766264e+01	1	5.259528
natural_resources_rents	6.153199e+00	1	2.480564
energy_consumption_per_capita	2.804072e+01	1	5.295349
population	1.029402e+03	1	32.084293
deaths_air_pollute	4.895036e+01	1	6.996453

Figure M: VIF for Model 1

Furthermore, after exploring regressions with and without Country_Name, we theorized that the Country_Name variable was reporting some independent variables to be more significant at a country level. For example, Model 1a showed all our independent variables to be significant except natural_resources_rents. However, Model 1b, where we removed Country_Name, showed that several variables were not significant. We, therefore, validated several models without Country_Name. We also performed a VIF analysis on Model 1b without Country_Name, which showed that labor force, age dependency ratio, fertility rate and population were highly correlated (**Figure N**). Regression Model 8 with only Country_Name against our dependent variable showed certain countries are highly significant to overall GHG emissions levels. For example, the UK, the United States, Japan, India, Indonesia, Germany, Brazil,

labor_force	age_dependency_ratio	agricultural_land
20.648364	11.278887	1.214345
fertility_rate	GDP	life_expectancy
13.479727	1.391672	5.430618
natural_resources_rents	energy_consumption_per_capita	population
1.372050	2.048128	19.853572
deaths_air_pollute		
3.314927		

Figure N: VIF (Without Country_Name)

Canada, and China had highly significant p-values of less than 2e^-16, which aligns with our earlier EDA analysis of the top 10 emitting countries. Although Country_Name provides interesting insight at the country level, it was not necessary to include it in our model, as we are examining the relationship between population and GHG emissions without categorical distinction.

For the sake of comparison, the team decided to run three regression models that looked at GHG_without_LUCF against only one independent variable to serve as baselines. The first of these models, Model 3, had population as the sole independent variable. The R^2 value for Model 3 was 0.5662694. This showed that GHG and population do have some kind of relationship, but population might not be the sole factor driving the high R^2 values of the previous models with more factors. We also ran a regression model, Model 6, that had the overall global population as the sole independent variable to see how that compared to Model 3, which used raw population data. Model 6's R^2 value was 0.9751, which was extremely high. This could be due to the considerably smaller dataset over which the regression was run, resulting in overfitting. This model didn't really provide a whole lot of value in terms of figuring out whether our hypothesis was correct or not, although interesting. The last baseline model the team experimented with was using GDP as the sole independent variable (Model 4). The R^2 value for this model was 0.6784612, which was higher than the R^2 value for Model 3, showing that GDP might be more significant than population when relating to GHG emissions.

Based on independent variable selection after cross-validating the above models, we selected Model 9 as the model to move forward with. This model had a final performance R^2 of 0.8789 when evaluated on the 20% test data. See **Figure O** for a summary of the selected model fitted to the full dataset.

```
Call:
lm(formula = GHG_without_LUCF ~ GDP + population + agricultural_land +
    natural_resources_rents + energy_consumption_per_capita +
    deaths_air_pollute, data = GHGdata)

Residuals:
       Min         1Q     Median         3Q        Max
-2.952e+09 -1.491e+07  2.803e+07  5.127e+07  3.983e+09

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  219700932    4992936  44.002  < 2e-16 ***
GDP                          517322472    5691882  90.888  < 2e-16 ***
population                   450238546    5566840  80.879  < 2e-16 ***
agricultural_land             13324538    5287284   2.520   0.0118 *
natural_resources_rents       25582415    5481482   4.667 3.15e-06 ***
energy_consumption_per_capita 24674711    5947461   4.149 3.41e-05 ***
deaths_air_pollute             5075882    5989010   0.848   0.3967
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 323700000 on 4196 degrees of freedom
Multiple R-squared:  0.8625,    Adjusted R-squared:  0.8623
F-statistic:  4386 on 6 and 4196 DF,  p-value: < 2.2e-16
```

Figure O: Model 9 Regression

We also performed PCA and lasso regression on our dataset to try and reduce the amount of variables we are including. The lasso regression model was cross-validated to find the best lambda value. After we applied the best lambda value to the lasso regression, the model selected all independent variables and produced an R^2 value of 0.92 on the test data (**Figure P**). Since the lasso regression selected all variables, there was no reduction of variables. However, PCA selected three primary components, and its performance on the test data resulted in an R^2 value of 0.82 (**Figure Q** and **Figure R**).

```
12 x 1 sparse Matrix of class "dgCMatrix"
                                      s0
(Intercept)                  212525886.7
Country_Name                    113257.1
age_dependency_ratio         -47667683.2
agricultural_land             20151033.8
fertility_rate                37948354.5
GDP                          460799518.7
labor_force                 1024781462.3
life_expectancy              -34962624.3
natural_resources_rents       15500396.3
energy_consumption_per_capita 23571768.2
population                  -519893136.0
deaths_air_pollute            -9704662.3
          RMSE          Rsquared               MAE
293805095.6283337 0.9211111 105058855.5864614
```

Figure P: Lasso

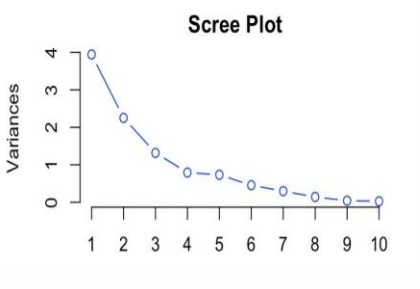

Figure Q: PCA Scree

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 209754867    6698217   31.32   <2e-16 ***
PC1         112710418    3373420   33.41   <2e-16 ***
PC2         449100203    4464204  100.60   <2e-16 ***
PC3         130560265    5835524   22.37   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 388400000 on 3359 degrees of
freedom
Multiple R-squared:  0.7775,    Adjusted R-squared:  0.7773
F-statistic:  3912 on 3 and 3359 DF,  p-value: < 2.2e-16


          RMSE          Rsquared               MAE
441784037.9930761 0.8219102 145069602.9210091
```

Figure R: PCA

When we combined the PC values with the dependent variable and created a regression model, we noted that the performance R^2 value was 0.8219, which was reasonable given the reduction to three PCs.

However, since we do not plan to use these models for prediction or forecasting, we are using Model 9 to draw conclusions about our data.

9. Conclusion

In the end, our hypothesis was correct: population is not the most significant variable regarding a country's GHG emissions, but it is very highly correlated. GDP was the only variable more significant in GHG emissions. Regressing population and GDP individually against GHG emissions showed a 0.52 and 0.67 R2 respectively. Consistently across our models, including the final linear regression model, GDP had a much higher coefficient (517,322,472) than population (450,238,546), which indicates that it has a larger impact on GHG emissions. GDP also had a higher t-value and lower p-value than population, which indicates that consumption and production have a much larger impact on emissions than population growth. In other words, GHG emissions may be more influenced by the way people are consuming goods and services rather than the number of people living on the planet. It logically follows that increased population rates could lead to higher GDP, but GDP and population had a correlation coefficient of only 0.4. As such, population could be highly correlated with GHG emissions because of the energy and consumption systems the world currently runs on, and today's culture of mass production and overconsumption; however, further research would have to be conducted to conclusively determine this.

9.1. Limitations and Unexpected Challenges

Since our dataset required subjective judgment to decide what variables to include or exclude, there was a chance we left out data that would greatly affect our model. In addition, with the limited data available, we may simply have lacked the breadth of data needed to fully consider the issue from all angles. We initially tried to incorporate up to twenty additional datasets as independent variables, however, there were many datasets that had insufficient values for analysis and were ultimately removed. There was also the potential for biases in selecting additional data to enhance our dataset. Since the decision to include or exclude specific datasets was subjective and depended upon what information was available, there was a chance that the variables we chose were more likely to create a specific outcome in analysis. Furthermore, our analysis showed that there was some multicollinearity between highly correlated dependent variables, such as population and labor force. This could be problematic in performing linear regression, which requires the assumption that variables are independent.

Another unexpected challenge was how high our R^2 values were for the models we built containing multiple variables. High R^2 values tend to indicate overfitting, but the values remained just as high after removing variables that were insignificant in the initial models. This could also be due to high multicollinearity between many of our variables and is ultimately why we selected a model with a lower R^2 value than the others.

9.2. Future Improvement

As years go by and data collection improves, it will be possible to deepen the analysis to include some of the pertinent variables that we were forced to remove for the sake of data cleanliness. Additionally, with further years of data reporting, a summarized global dataset could be used to interpret GHG emissions correlation with these variables across the world.

10. Works Cited

Center for Biological Diversity. *Population Pressure and the Climate Crisis.* Center for Biological Diversity. Retrieved March 29, 2023, from https://www.biologicaldiversity.org/programs/population_and_sustainability/climate/

Chang, Nicole Lin. (2022, November 15). *Is population growth fuelling climate change? It's not that simple, say experts*. Euronews.green. Retrieved March 29, 2023 from https://www.euronews.com/green/2022/11/15/is-population-growth-fuelling-climate-change-its-not-that-simple-say-experts

Cohen, Joel. (2010, June). *Population and climate change*. National Library of Medicine. Retrieved March 29, 2023 from https://pubmed.ncbi.nlm.nih.gov/21553595/

D'Souza, Renita. (2022, July 11). *Population drives climate change: A myth or reality?* Observer Research Foundation. Retrieved March 29, 2023, from https://www.orfonline.org/expert-speak/population-drives-climate-change/

Galloway, Scott. (2023, January 27). *More Babies*. No Mercy/No Malice. Retrieved March 28, 2023, from https://www.profgalloway.com/more-babies/

Kaplan, Sarah. (2021, May 25). *It's wrong to blame 'overpopulation' for climate change*. The Washington Post. Retrieved March 29, 2023, from https://www.washingtonpost.com/climate-solutions/2021/05/25/slowing-population-growth-environment/

Tallungs, Kaj. (2022, November 16). *America's demographics over 100+ years*. Visual Capitalist. Retrieved March 25, 2023, from https://www.visualcapitalist.com/cp/animated-americas-demographics-over-100-years/

Union of International Associations (UIA). (2023). *Depopulation*. The Encyclopedia of World Problems & Human Potential. Retrieved March 25, 2023, from

http://encyclopedia.uia.org/en/problem/depopulation

United Nations (UN). (2022, April 4). U*N climate report: It's 'now or never' to limit global warming to 1.5 degrees*. United Nations. Retrieved March 25, 2023, from https://news.un.org/en/story/2022/04/1115452

United States Environmental Protection Agency (EPA). (2023, February 15). *Global Greenhouse Gas Emissions Data*. United States Environmental Protection Agency. Retrieved March 25, 2023, from https://www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data#:~:text=Global%20carbon%20emissions%20from%20fossil,increase%20from%201970%20to%202011

Oxfam. (2015, December 2). *Extreme Carbon Inequality*. Oxfam.org. Retrieved March 29, 2023 from https://oxfamilibrary.openrepository.com/bitstream/handle/10546/582545/mb-extreme-carbon-inequality-021215-en.pdf

Quinson, Tim. (2022, December 13). *ESG Market in US Significantly Smaller Than Earlier Thought*. Bloomberg.com. Retrieved March 25, 2023, from https://news.bloomberglaw.com/esg/esg-market-in-us-significantly-smaller-than-earlier-thought-1

Ritchie, Hannah and Roser, Max. (2020). *CO_2 and Greenhouse Gas Emissions.* Our World in Data. Retrieved March 29, 2023 from https://ourworldindata.org/greenhouse-gas-emissions

Zurich. (2021, November 16). *Here's how climate change will impact businesses everywhere.* Zurich.com. Retrieved March 29, 2023, from https://www.zurich.com/en/knowledge/topics/climate-change/how-climate-change-will-impact-business-everywhere

11. Appendix:

11.1. Explanation of Variables

Variable	Data Type	Description	Source
Country_Code	Character	Unique identifier code for each country	Our World in Data
Country_Name	Character	Country name	Our World in Data
Year	Date	Year	Our World in Data
GHG_with_LUCF	Numerical	Total greenhouse gas emissions from 1990-2019, including land use and forestry. Emissions are measured in carbon dioxide-equivalents	Our World in Data: https://ourworldindata.org/greenhouse-gas-emissions
population	Numerical	World population data from 1962-2021	WorldBank: https://data.worldbank.org/indicator/SP.POP.TOTL
age_dependency_ratio	Numerical	Age dependency ratio of dependents - people younger than 15 or older than 64 to the working-age population	WorldBank: https://data.worldbank.org/indicator/SP.POP.DPND
agricultural_land	Numerical	Agricultural land (% of land area)	WorldBank: https://data.worldbank.org/indicator/AG.LND.AGRI.ZS
fertility_rate	Numerical	Fertility rate - births per woman	WorldBank: https://data.worldbank.org/indicator/SP.DYN.TFRT.IN
life_expectancy	Numerical	Life expectancy at birth: the number of years a newborn infant would presumably live.	WorldBank: https://data.worldbank.org/indicator/SP.DYN.LE00.IN
GDP	Numerical	Gross Domestic Product in current US$	WorldBank: https://data.worldbank.org/indicator/NY.GDP.MKTP.CD
natural_resources_rents	Numerical	Total natural resources rents are the sum of oil rents, natural gas rents, coal rents (hard and soft), mineral rents, and forest rents, in % of GDP	WorldBank: https://data.worldbank.org/indicator/NY.GDP.TOTL.RT.ZS
labor_force	Numerical	Labor force comprises people ages 15 and older who supply labor for the production of goods and services. It includes people who are currently employed and people who are unemployed but seeking work as well as first-time job-seekers	WorldBank: https://data.worldbank.org/indicator/SL.TLF.TOTL.IN

energy_consumption_per_capita	Numerical	Energy use per person measured in kilowatt-hours per capita. Energy use per person Energy use not only includes electricity, but also other areas of consumption including transport, heating and cooking.	Our World in Data: https://ourworldindata.org/grapher/per-capita-energy-use?tab=table
deaths_air_pollute	Numerical	Death rate from air pollution, 1990 to 2019: Estimated annual number of deaths attributed to air pollution per 100,000 people.	Our World in Data: https://ourworldindata.org/grapher/death-rate-from-air-pollution-per-100000?tab=table
GHG_without_LUCF	Numerical	Total greenhouse gas emissions from 1990-2019, excluding land use and forestry. Emissions are measured in carbon dioxide-equivalents.	Our World in Data: https://ourworldindata.org/grapher/total-ghg-emissions-excluding-lufc?tab=chart&country=~ROU