Final Project

PSTAT 100: Spring 2024 (Instructor: Ethan P. Marzban)

AARTI GARAYE (aartigaraye) SEAN REAGAN (screagan)
VARUN BASKARAN (vbaskaran) SAM GUIMTE (samuelguimte)

June 12, 2024

Overview

This is the final project of PSTAT 100: Introduction to Data Science. In this project we will be working and exploring the ClimateWatch historical emissions data: greenhouse gas emissions by U.S. state 1990-present. In order to work with the data, we will need to tidy and format the dataset a little using pivot and melt functions. The historical emissions data set shows the emission of all gasses in different countries from 1990 to 2019. It also includes a row with the total emission of the entire world.

This report outlines the methodology, findings, and implications of our research on historical emissions of different countries, offering actionable recommendations. We have also built an interactive app so that users can select countries and see the emissions plots and percentages of World's emissions. Through this report, we aim to highlight the challenges and opportunities in controlling global warming, supporting informed decision-making.

Introduction

In this project we will be doing some nontrivial **exploratory analysis**, **descriptive analysis**, and some **statistical modeling** on the data to answer a few questions about trends in emission. There will be visual representation of trends, detailed analysis and comparison of these trends, and regression analysis to identify factors associated with changes in emission. We will mainly focus on how have greenhouse gas emissions trends have changed over time across different countries around the world. We will explore which countries have the highest and lowest greenhouse gas emissions, and how have their emissions profiles changed over time, primary from 1990-2019? We shall also dive into what actions should be taken to improve the current situation to make Earth a better place.

The data is provided by the instructor. We will be working with ClimateWatch historical emissions data: greenhouse gas emissions by U.S. state 1990-present. The data has following columns:

Country: Name of the country. There are total of 195 countries.

Data Source: Where we got the data from for that specific country.

Sector: This is pretty self explanatory, all sectors are "Total including LUCF."

Gas: Includes all greenhouse gases.

Years: Each year is a different column. It goes from 1990-2019.

Required Packages

Before moving on, let's load the necessary packages.

```
Attaching package: 'dplyr'


The following objects are masked from 'package:stats':

    filter, lag


The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union


-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v forcats   1.0.0     v readr     2.1.4
v ggplot2   3.5.0     v stringr   1.5.1
v lubridate 1.9.3     v tibble    3.2.1
v purrr     1.0.2     v tidyr     1.3.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become e
Registered S3 method overwritten by 'quantmod':
  method            from
  as.zoo.data.frame zoo


Attaching package: 'zoo'


The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

Tiding up the data

Before moving on, we need to tidy and format our data using pivot and melt to be made usable for our analysis.

```
emissions <- read_csv("/home/jovyan/100-sp24/Final_Proj/data/historical_emissions/historical_
```

```
Rows: 195 Columns: 35
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (6): Country, Data source, Sector, Gas, Unit, 1990
dbl (29): 2019, 2018, 2017, 2016, 2015, 2014, 2013, 2012, 2011, 2010, 2009, ...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
sources <- read_csv("/home/jovyan/100-sp24/Final_Proj/data/historical_emissions/sources.csv")
```

```
Rows: 0 Columns: 3
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (3): Source, Property, Value

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
emissions_tidy = emissions %>%
  mutate(`1990` = as.numeric(`1990`)) %>%
  rename_with(~ gsub("^X", "", .))

emissions_pivot = emissions_tidy %>%
  pivot_longer(
    cols = -c(Country, `Data source`, Sector, Gas, Unit),
    names_to = "Year",
    values_to = "Amount"
  )


emissions_final = emissions_pivot %>%
  select(!`Data source` & !Sector & !Gas & !Unit)
unique(emissions_tidy$Unit)
```

```
[1] "MtCO e"
```

```r
world_emission = emissions_final %>%
  filter(Country == "World" & Year == '1990') %>%
  pull(Amount)

country_emission = emissions_final %>%
  filter(Country == "India" & Year == '1990') %>%
```

```
  pull(Amount)

world_emission
```

```
[1] 32523.58
```

```
country_emission
```

```
[1] 1002.56
```

```
emission_ratio = country_emission / world_emission
emission_ratio
```

```
[1] 0.03082563
```

In the above chunk of code, we pivoted and melted the data so that our years are not a separate column. There are three columns now: Country, Year, and Amount. Country and Year are pretty self explanatory variables, Amount, on the other hand, is the total amount of emission that country observed during that particular year. In the last part of the code, we just tried out the emission ratio of India in 1990. To do so we pulled the total amount of gas emitted by the entire world in 1990 and stored it in world_emission. We then pulled the total amount emitted by India in 1990 and stored it in country_emission. We them found the ratio of it by dividing the amount of gas India emitted by the amount of gas the World emitted and stored it in emission_ratio.

Analysis

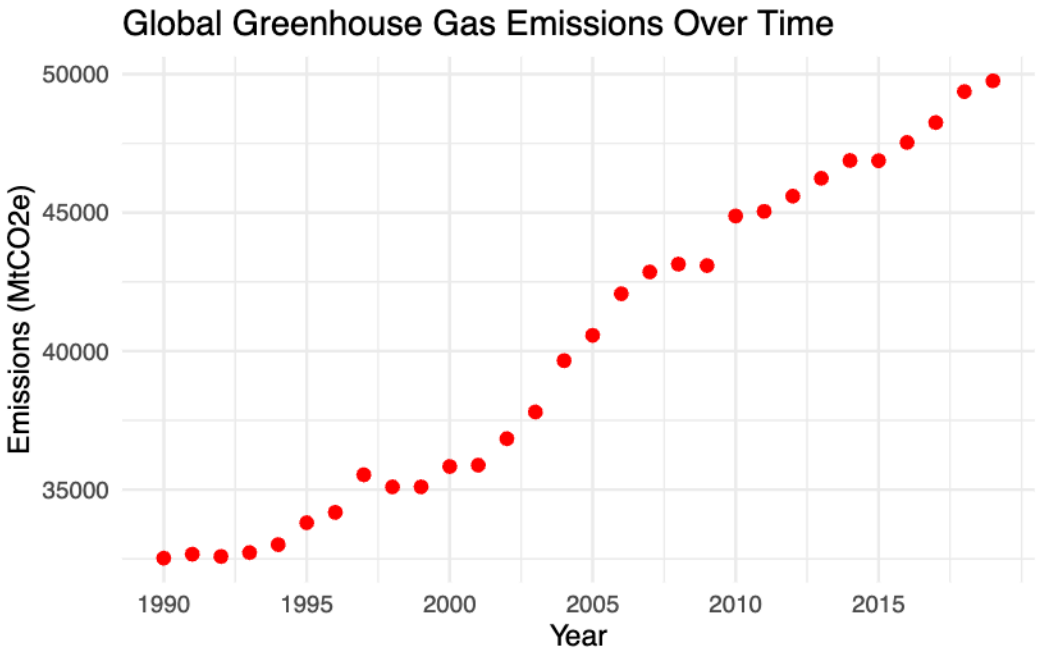
Exploratory Data Analysis (EDA)

Overall Global Trend

In this section of the report we will visualize the overall trends of greenhouse gas emissions and comparing them across time and different countries. Let's start by visualizing the global trend of greenhouse emission from 1990-2019.

```
emissions_final <- emissions_final %>%
  mutate(Year = as.numeric(Year))

emissions_final <- emissions_final %>%
  mutate(Amount = as.numeric(Amount))

global_emission <- emissions_final %>%
  filter(Country == "World")
```

```r
ggplot(global_emission, aes(x = `Year`, y = `Amount`)) +
  geom_point(color = "red", size = 2) +
  labs(title = "Global Greenhouse Gas Emissions Over Time",
       x = "Year",
       y = "Emissions (MtCO2e)") +
  scale_x_continuous(breaks = seq(min(global_emission$Year, na.rm = TRUE),
                                  max(global_emission$Year, na.rm = TRUE),
                                  by = 5)) +
  theme_minimal()
```

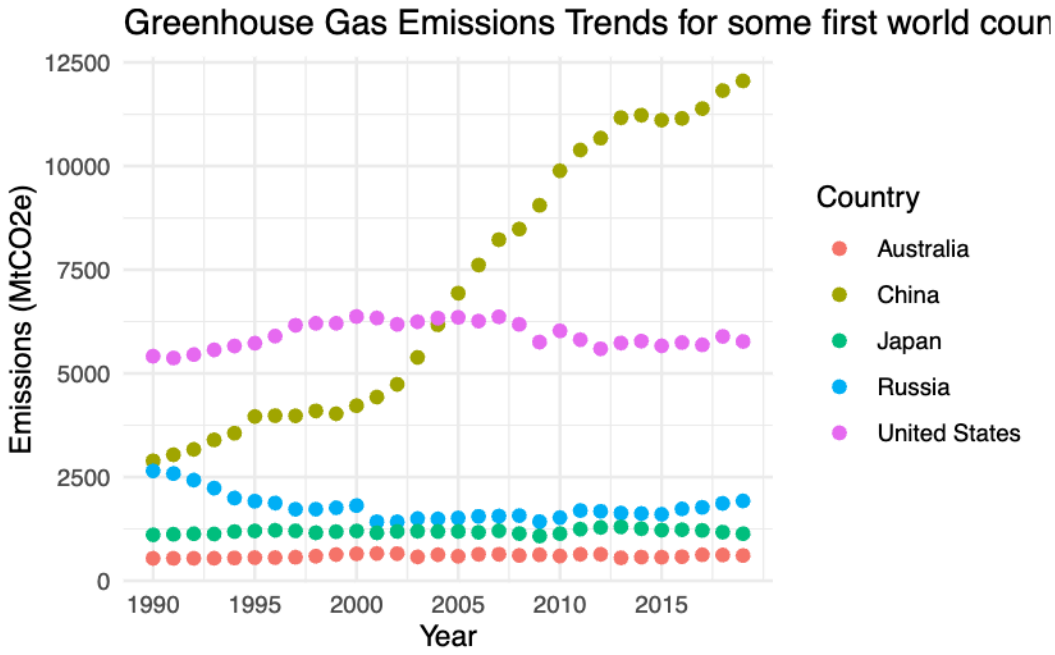


The graph above shows a concerning overall trend of positive constant growth in emission of greenhouse gas. It seems like we were doing kind of a good job in mid 2007 to 2010 and then it jumped in 2010. One of the major concerns for this could be because of the increase in production and technology we saw in the past decade. Although, some may argue that we see a more passive growth in the past decades, looking at the slopes, and that may be due to the shift in our focus on sustainable energy. Introduction and popularity gain of solar power, geothermal energy, and electric cars might be some answers to the decrease in the slope of the growth.

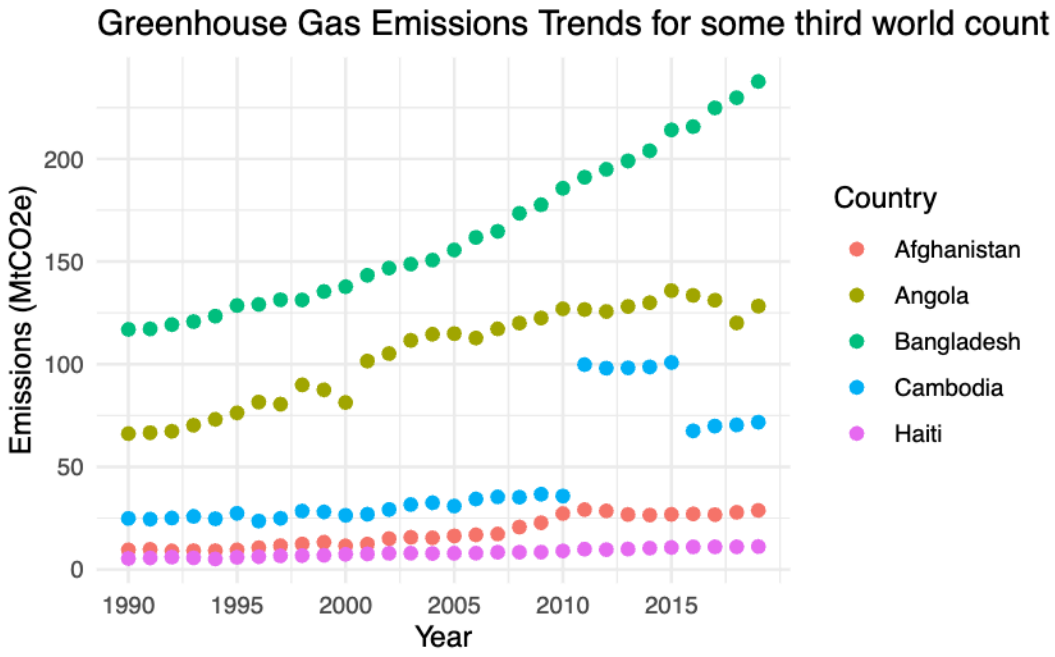
Trend of Different Countries

Now lets look at trends from what we normally consider the first world countries, and have side by side box plots for these and then compare these trends to what are considered the third world countries. This will provide us with more knowledge on which types countries should focus more on controlling their greenhouse emission. We have chosen five countries for each categories across different continents with somewhat a similar population distribution based on the total area a country has.

```r
first_countries <- c("United States", "China", "Russia", "Japan", "Australia")
third_countries <- c("Angola", "Afghanistan", "Haiti", "Bangladesh", "Cambodia")

first_country_emissions <- emissions_final %>%
  filter(Country %in% first_countries)

third_country_emissions <- emissions_final %>%
  filter(Country %in% third_countries)

first_country_emissions <- first_country_emissions %>%
  mutate(Year = as.numeric(Year))

first_country_emissions <- first_country_emissions %>%
  mutate(Amount = as.numeric(Amount))

third_country_emissions <- third_country_emissions %>%
  mutate(Year = as.numeric(Year))

third_country_emissions <- third_country_emissions %>%
  mutate(Amount = as.numeric(Amount))

ggplot(first_country_emissions, aes(x = `Year`, y = `Amount`, color = `Country`)) +
  geom_point(size = 2) +
  labs(title = "Greenhouse Gas Emissions Trends for some first world countries",
       x = "Year",
       y = "Emissions (MtCO2e)") +
  scale_x_continuous(breaks = seq(min(first_country_emissions$Year, na.rm = TRUE),
                                  max(first_country_emissions$Year, na.rm = TRUE),
                                  by = 5)) +
  theme_minimal()
```




```r
ggplot(third_country_emissions, aes(x = `Year`, y = `Amount`, color = `Country`)) +
  geom_point(size = 2) +
  labs(title = "Greenhouse Gas Emissions Trends for some third world countries",
       x = "Year",
       y = "Emissions (MtCO2e)") +
  scale_x_continuous(breaks = seq(min(third_country_emissions$Year, na.rm = TRUE),
                                  max(third_country_emissions$Year, na.rm = TRUE),
                                  by = 5)) +
  theme_minimal()
```



From the two graphs above we can clearly see the distinction in the scale of y-axis for the first world and third world countries. We can also see some pattern between the different countries as well. The United States of American is very clearly emitting a lot of greenhouse gas compared to other first world countries. Some would say this might be because we clearly have more area and population compared to the other countries, but it is also important to note the magnitude of difference. Japan, for example, have 125 million population and US has 333 million, so for 208 million more people we are emitting 4500 MtCO2e. Another country to note is China and the almost exponential growth it is showing in the greenhouse gas emission. Around 2004 China surpassed the US and became the highest country to emit greenhouse gasses. Some might say it is because of China's vast population, 1.4 billion, and it might be good answer, however, it is vital to find a solution to make Earth sustainable.

In the third world countries, it seems that Bangladesh is consistently increasing it's greenhouse gas emission. It is a developing country and compared to other countries on that graph, Bangladesh seem to be the country to have the most economic growth and capitalistic development. Another thing to note is the weird jump from 2010 to 2015 and from 2015 to 2019 in Cambodia. Google suggests that Cambodia had an unusual year in 2010 for it's industry, the gap shows that there might be something to investigate. Instead of focusing on one country, we are going to move on to the big picture.

Discriptive Analysis

In this section we will focus on which country has the highest and the lowest greenhouse gas emission rates. We will also explore total emission rates for each country and the average annual emission rates as well. From our exploratory data analysis, we can safely assume that the country that would have the highest gas emission rate would probably be a country from the first world. It is also not a bad idea to bet that it will be China because we can see from the graph that as if 2019 it has reached 12500 MtCO2e. And the country to have the lowest greenhouse gas emission rates would be a third or a fourth world country. Also, it is important to keep in mind that population seem to be playing a key role in determining this, so probably a less developed country with low population would be a country with the lowest rates.

Highest Total Emission Rates

Let's first determine which country would have the highest greenhouse gas emission rates. We can do this by using the groupby() and summarize functions. The second row of the new table would be the country with the highest gas emission rates, because the first one would be the World as it is every country combined.

```
total_emission_by_country <- emissions_final %>%
  group_by(Country) %>%
  summarise(total = sum(Amount, na.rm = T)) %>%
  arrange(desc(total))

head(total_emission_by_country, 3)
```

```
# A tibble: 3 x 2
  Country          total
  <chr>            <dbl>
```

```
1 World         1205415.
2 China          212227.
3 United States  177781.
```

As we predicted, the highest greenhouse emitting country is China. Please note that China has a very large population so it makes sense why it would be the country to have the highest greenhouse emission rates. Let's see how far is the runner up country, the US. So the difference in the total emission between US and China is $212227.0 - 177781.2 = 34445.8$ China produces 34445.8 MtCO2e more than the US, and China has approximately one billion seventy-eight million seven hundred thousand more people than the US.

Lowest Total Emission Rates

We can do the similar thing for finding the country with the lowest emission rate, we just would have to arrange it in an ascending order, which is the default for arrange. Before we begin, it is safe to say that countries with low population and less industrial and production development would have a really low emission rate.

```r
total_emission_by_country <- emissions_final %>%
  group_by(Country) %>%
  summarise(total = sum(Amount, na.rm = T)) %>%
  arrange(total)

head(total_emission_by_country, 3)
```

```
# A tibble: 3 x 2
  Country   total
  <chr>     <dbl>
1 Bhutan  -118.
2 Fiji     -12.9
3 Niue       0.26
```

As we can see till 2019, Bhutan had the lowest greenhouse gas emission rates totaling to -118.05 MtCO2e. It makes sense according to our prediction. Furthermore, Bhutan's majority of population identify themselves as Mahayana Buddhism, which is practiced by between 73–85% of the population. The remaining population is mainly Hindu, with some Christians and Muslims. Since the majority is Buddhism and Hinduism, it is common practice in these religions to be vegetarian. If the country's most of the population is vegetarian it makes sense that the carbon footprint of people living in Bhutan would be significantly low.

Highest Average Annual Emission Rates

It would be interesting to see if the average annual emission rates are different and if it would change the country who would have the highest emission rates. We will follow a similar procedure, however, this time intead of using sum() in our summarise function, we would use mean() to get the average. We would also have to arrange it in a descending order to get the country with the highest average annual emission rates.

```r
total_emission_by_country <- emissions_final %>%
  group_by(Country) %>%
  summarise(Average = mean(Amount, na.rm = T)) %>%
  arrange(desc(Average))

head(total_emission_by_country, 3)
```

```
# A tibble: 3 x 2
  Country       Average
  <chr>           <dbl>
1 World          40181.
2 China           7074.
3 United States   5926.
```

China still has the highest annual average emission rates with the US as the runner up.

Lowest Average Annual Emission Rates

Again we would do exactly the same thing as we did above, however, we would have nothing in the argument of arrange, as ascending is the default. What are your predictions, which country would have the lowest average annual emission rates?

```r
total_emission_by_country <- emissions_final %>%
  group_by(Country) %>%
  summarise(Average = mean(Amount, na.rm = T)) %>%
  arrange(Average)

head(total_emission_by_country, 3)
```

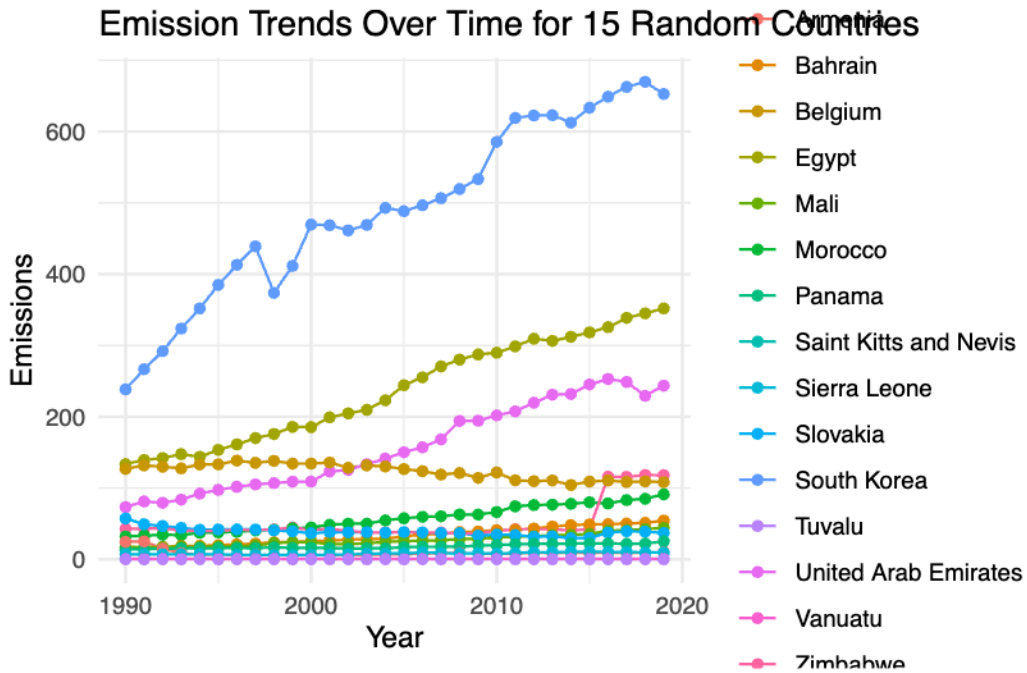
```
# A tibble: 3 x 2
  Country  Average
  <chr>      <dbl>
1 Bhutan  -3.94
2 Fiji    -0.431
3 Niue     0.00867
```

Bhutan has the lowest average emission rates similar to the one we saw previously. Bhutan is also known for it's peaceful nature and monasteries making it one of the most serene country with very less greenhouse gas emission.

Statistical Modeling

In this section of Analysis, we will be doing some modeling using various machine learning techniques and statistical analysis methods to identify and assess some factors associated with the emission changes.

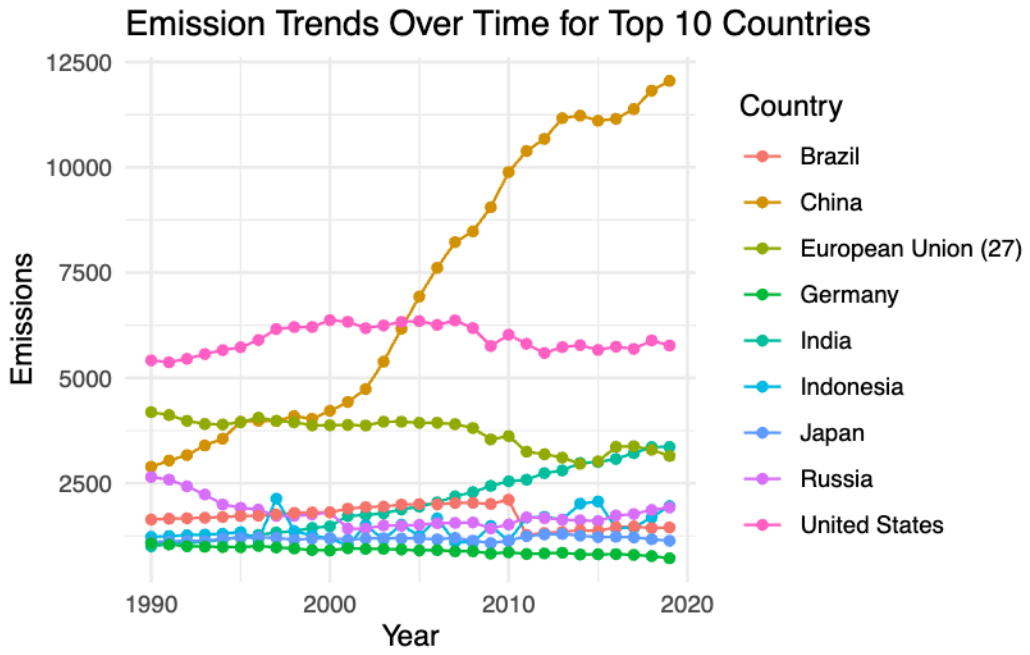
```r
set.seed(77)

random_countries <- sample(unique(emissions_final$Country), 15) #sampling 15 random countries
data_subset <- emissions_final %>% filter(Country %in% random_countries)
data_subset %>% ggplot(aes(x=Year,y=Amount,color=Country)) +
  geom_line() +
  geom_point() +
  labs(title = "Emission Trends Over Time for 15 Random Countries", x = "Year", y = "Emission
  theme_minimal()
```



The above code displays time-series data for 15 different random countries every time it is run. For display and analytical purposes, we have a time series graph displayed for the 15 countries listed to the right of the visual in the legend denoting which color represents which country. There are no apparent trends showing significant decreases in emission amounts among any country. The only noticeable trends are positive ones from South Korea, Egypt, and the UAE. This seems to be the case every time this simulation is done; the majority of the countries represented are flat at the bottom but the rest are well above the others steadily increasing in emission amounts. There are a few exceptions like Kazakhstan that fluctuate up and down more than they show a pattern. Nonetheless, the general positive trend remains for every 15 different random countries plotted.

```r
highest_total_emission_by_country <- emissions_final %>%
  group_by(Country) %>%
  summarise(Average = mean(Amount, na.rm = T)) %>%
  arrange(desc(Average))

top_emission_countries <- head(highest_total_emission_by_country, 10)
data_subset2 <- emissions_final %>% filter(Country %in% top_emission_countries$Country) %>% f
```

```r
data_subset2 %>% ggplot(aes(x=Year,y=Amount,color=Country)) +
  geom_line() +
  geom_point() +
  labs(title = "Emission Trends Over Time for Top 10 Countries", x = "Year", y = "Emissions") +
  theme_minimal()
```



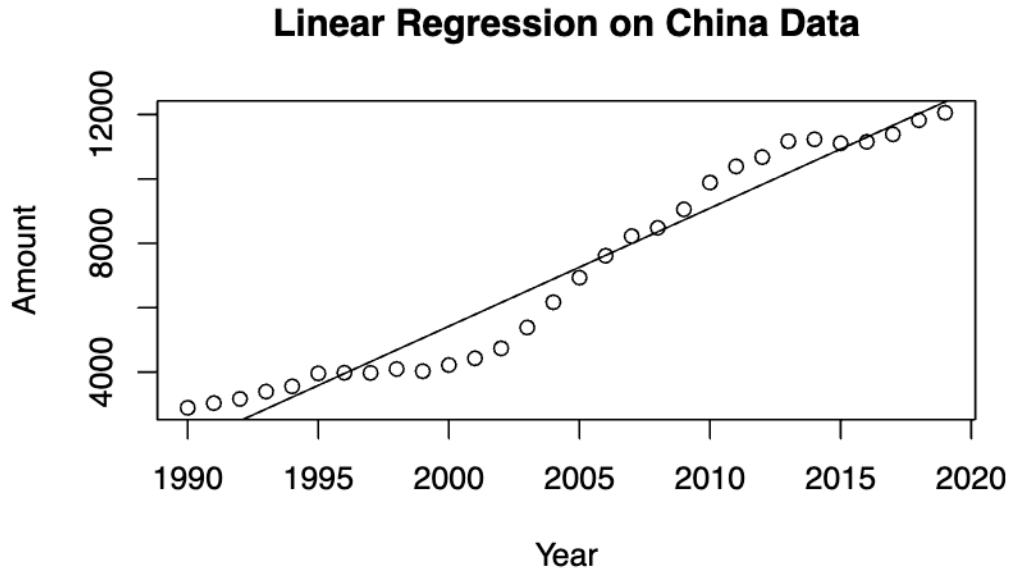
If we examine just the top 10 countries (excluding "World") with the most emissions, we see something different than what we might expect. Only the top country China shows a strong and steady growing emission rate among the top 10 countries while the rest have plateaued or are decreasing like the U.S. or the E.U. Inspecting the time series data for each of these countries has allowed for some interesting findings, but we can take a step further by fitting a model to specific countries' data.

```r
china_data <- emissions_final %>% filter(Country == "China")
model <- lm(Amount ~ Year, data = china_data)
summary(model)
```

```
Call:
lm(formula = Amount ~ Year, data = china_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1420.19  -343.47    74.16   637.52  1136.63

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -728249.24   31938.66  -22.80   <2e-16 ***
Year            366.84      15.93   23.02   <2e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 755.4 on 28 degrees of freedom
Multiple R-squared:  0.9498,    Adjusted R-squared:  0.948
F-statistic: 530.1 on 1 and 28 DF,  p-value: < 2.2e-16
```

```r
plot(Amount ~ Year, data = china_data, main="Linear Regression on China Data")
abline(model)
```

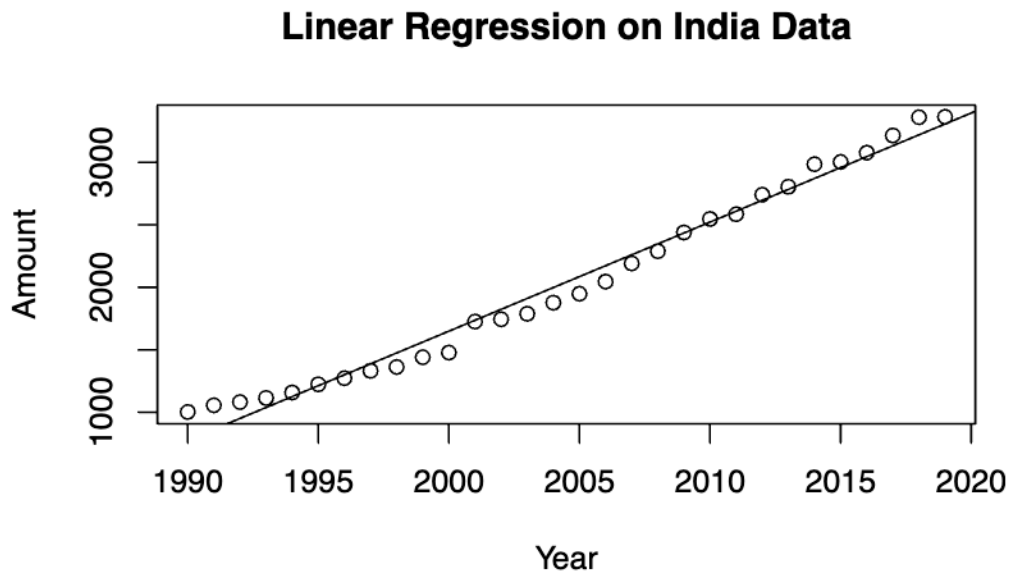



```r
india_data <- emissions_final %>% filter(Country == "India")
model2 <- lm(Amount ~ Year, data = india_data)
summary(model2)
```

```
Call:
lm(formula = Amount ~ Year, data = india_data)

Residuals:
     Min       1Q   Median       3Q      Max
-171.071  -76.473    7.815   53.768  226.262

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.729e+05  4.387e+03  -39.41   <2e-16 ***
Year         8.726e+01  2.188e+00   39.88   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 103.7 on 28 degrees of freedom
```

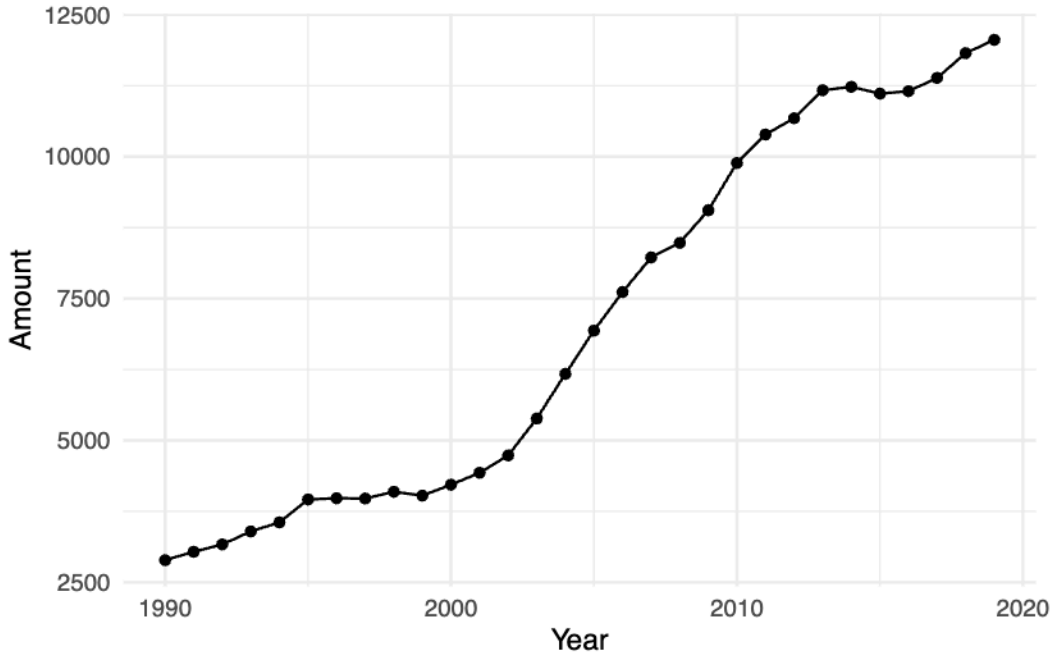
```
Multiple R-squared:  0.9827,    Adjusted R-squared:  0.9821
F-statistic:  1590 on 1 and 28 DF,  p-value: < 2.2e-16
```

```r
plot(Amount ~ Year, data = india_data, main="Linear Regression on India Data")
abline(model2)
```



For example, let's take a look at 2 of the top 10 emission countries: China and India. Just from observing their time series data alone, we can observe strong positive linear patterns for both sets of data. After fitting a simple linear regression model for both of the countries with "Year" as the predictor and "Amount" as the response variable, we come out with very convincing results. From the above plots, we can be confident that linear regression models fit very well to the existing data and could have certainly be used to predict emission amounts for at least these two countries. With respective adjusted R-squared values of 0.948 and 0.9821, the countries of China and India show strong linear relationships that have potential for predictive purposes in future years.

```r
emissions_final %>% filter(Country=="China") %>%
  ggplot(aes(x=Year,y=Amount)) +
  geom_point() +
  geom_line() +
  theme_minimal()
```



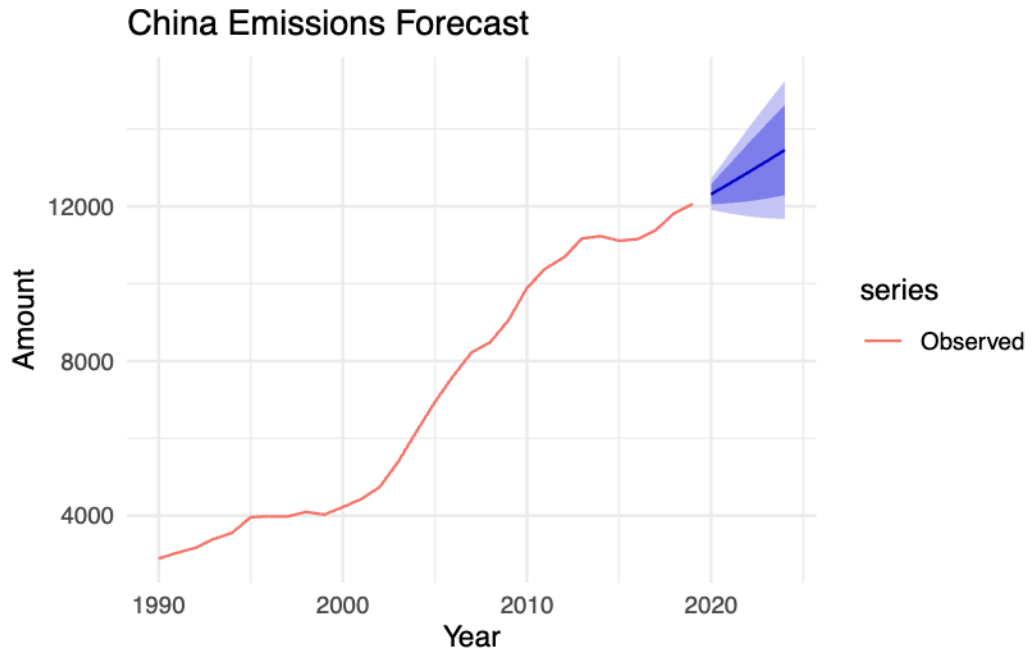

```
china_ts <- ts(china_data$Amount, start = min(china_data$Year), end = max(china_data$Year), f

time_index <- time(china_ts)
reversed_values <- rev(coredata(china_ts))
rev_china_ts <- ts(reversed_values, start = start(china_ts), end = end(china_ts), frequency =

china_ts_model <- auto.arima(rev_china_ts)
summary(china_ts_model) #ARIMA (1,1,0)
```

```
Series: rev_china_ts
ARIMA(1,1,0) with drift

Coefficients:
         ar1     drift
      0.6465  301.9411
s.e.  0.1350   98.1045

sigma^2 = 41942:  log likelihood = -194.72
AIC=395.44   AICc=396.4   BIC=399.55

Training set error measures:
                    ME     RMSE      MAE        MPE     MAPE      MASE
Training set 4.637932 194.2885 161.1409 -0.1093553 2.555237 0.4895018
                    ACF1
Training set 0.04666967
```

```
china_forecast <- forecast(china_ts_model, h = 5)
china_forecast
```

```
     Point Forecast   Lo 80    Hi 80    Lo 95    Hi 95
2020       12313.26 12050.8 12575.72 11911.87 12714.66
2021       12586.70 12081.1 13092.31 11813.44 13359.96
2022       12870.21 12129.1 13611.32 11736.78 14003.64
2023       13160.24 12198.6 14121.88 11689.54 14630.94
2024       13454.48 12288.9 14620.05 11671.89 15237.07
```

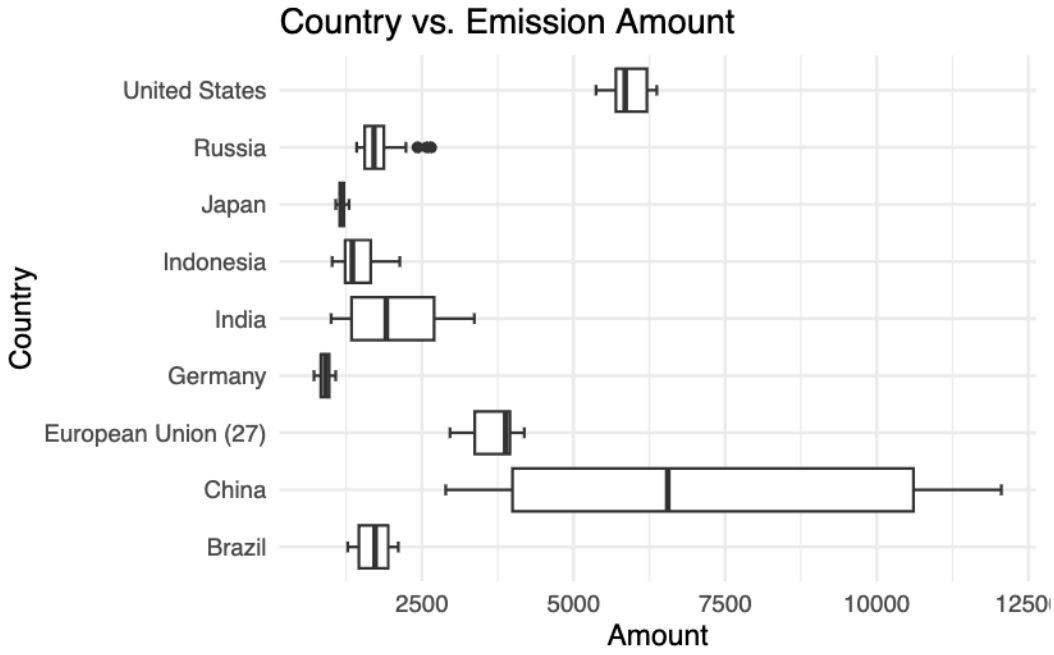
```
autoplot(china_forecast) +
  autolayer(rev_china_ts, series = "Observed") +
  labs(title = "China Emissions Forecast",
       x = "Year",
       y = "Amount") +
  theme_minimal()
```



This brings us to using a time series model to try to forecast future values for the next 5 years for the emission data for China. Using auto.arima an our time-series converted data for China resulted in a ARIMA (1,1,0) model. Thus, the model has one autoregressive term, no moving average term, and it was differenced once. The forecasted values are plotted above in blue with a band of confidence surrounding them; the values themselves are 12313.26, 12586.70, 12870.21, 13160.24, 13454.48 for 2020-2024 respectively.

```
data_subset2 %>%
  ggplot(aes(x=Amount, y=Country)) +
```

```
  geom_boxplot(staplewidth=0.25) +
  theme_minimal() +
  ggtitle("Country vs. Emission Amount")
```




```
aov(Amount ~ Country, emissions_pivot) %>% summary() #check for statistically significant dif
```

```
              Df    Sum Sq   Mean Sq F value Pr(>F)
Country      194 5.095e+10 262607867    1086 <2e-16 ***
Residuals   5654 1.367e+09    241799
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
1 observation deleted due to missingness
```

We can lastly use a boxplot and conduct an ANOVA to test whether there is a statistically significant difference in average emission amounts across the top 10 emission countries extracted earlier in this analysis. From the boxplot and our ANOVA results, we can confirm this statistically significant difference with confidence.

Conclusion

In conclusion, we have explored the historical greenhouse gas emissions across various U.S. states from the year 1990 to the present, focusing on trends, disparities, and potential strategies for mitigation. Throughout our analysis, we used statistical tools to identify emission patterns across different countries and attempted to understand the underlying factors through exploratory analysis and statistical modeling. Our findings illustrate a consistent rise in global emissions over the years. There

are significant variations when comparing developed and developing regions as well. Industrialized nations such as the US and China are among the highest emitters, driven by their economic activity and large populations. In contrast, countries with smaller economies and populations, like Bhutan, show lower emission figures, which reflects their different industrial and development profiles. This highlights the uneven contribution to global emissions but also underscores the need for differentiated responsibilities and strategies in tackling climate change as a whole. By using analytical techniques such as time-series analysis and linear regression models, we have gained insights into the dynamics of greenhouse gas emissions. These models have confirmed strong linear relationships in the data, particularly for major emitters like China and India, suggesting their potential utility in forecasting future emissions and policy decisions made by other nations. Our project underscores the critical role of data science in environmental studies and emphasizes the urgency for international cooperation in environmental management and policy creation. The goal is clear: to foster a sustainable future through informed, data-driven strategies that address the complexities of global emissions and their impacts on climate change.