Documentation

Xinwen Xu, Jennifer Liu, Yuxuan Peng, David Kook, William Escobar

2022-11-15

Analysis and Prediction of Global CO2 Emissions

Data Acquisition:

The data file "CO2 Emission by countries Year wise (1750-2022)" downloaded from Kaggle.

Citation: Bhatti, M. A. (2022, September 14). CO2 emission by Countries Year Wise (1750-2022). Kaggle. Retrieved October 7, 2022, from https://www.kaggle.com/datasets/moazzimalibhatti/co2-emission-by-countries-year-wise-17502022

Data Cleaning

Overall the raw data is pretty cleaned, we randomly selected a few countries and verified the input value. But there are a lot of missing values. So we started examine the missing values in R.

Data Cleaning Part in R

Initial Exploration

We started with an overall view of the missing data. See chunk: {r Check for all the NA entries}
We checked that if one country has a missing value in one row, the missing value is consistent through all the years of of that country. (eg: if Congo is missing "Calling.Code", the "Calling.Code" is missing in all the 271 entries associated with Congo.)
Then we tried to find the the specific countries that had different missing data in each column, using function filter() and distinct() in dplyr. See chunk: {r NA Countries Within Columns}

Generate an overview table to better see which values are missing.

Since one country could 1 to 3 different columns that has NAs, we decided to create an overview table. See chunk: {r Missing Value table}
Since there are in total 41 countries that had different missing values, we need to find a threshold to determine which countries we want to keep the emission data and supplement the missing values and the others left blank.
We decided to use the lower quarter 45060000 as a threshold, any country have the 2020 CO2 emission lower than that we are gonna omit. See chunk: {r Find the threshold}

Data Cleaning Part in Excel Initially we examined the NA's in excel. But hard to find to formulate a overall table in Excel so we switched to R and done the above analysis.
But we did find in Excel that there are empty entries that was not read by R so we added the "na = c("","NA")" line when reading the data into R.
As for the data implementation, we just used sort and filter to filter out the missing rows and fill in the correct data accordingly.For the detailed cleaning process, see the meta data page in excel.

Analysis Process

Historical CO2 Data analysis and visualization

Aim 1.1 General overview of data, any visual or numerical differences

The aim of this analysis is to parse through the cleaned CO2 data and determine any trends in CO2 emission totals. NAs were omitted and the data was regrouped to find the earliest year where each country had a nonzero CO2 emission. This dataset was copied and renamed as "CO2_no0" to indicate that zero values were removed. Duplicates were also removed ("CO2_noduplicates") and the year and country data were plotted in a basic scatter plot. (Refer to Chunk {r General Overview} for code)

Finding total CO2 emissions per country & CO2 emissions per country per person

Total emissions per country were found by aggregating the CO2 emissions data from the CO2 data set with no zero values and saving it into a new dataset. Columns were renamed and the final dataset was joined with CO2_noduplicates to obtain population data. Running summary() on the new joined dataset provides the numerical summaries of average and median cumulative CO2 emissions. The results of `summary()` on the CO2total_byCountry data set indicate that the average total CO2 emissions in the data set is 3.024e+11 tons. (Refer to Chunk {r total emissions per country})

Aim 1.2A To visualize the cumulative CO2 emissions by country and continent

Continent data was added to the CO2 Country Totals data to visualize emissions by continent. A barplot was made, which indicates that Europe had the most emissions followed by Americas. (See Chunk 11)

Create a map visualizing the amount of CO2 emission by country

Creating a barplot &world map to display CO2 emissions

Making a world map just for 2019 data

The map was made using ggplot's map data. Coding of country names like "United States" and "Democratic Republic of Congo" had to match the way the world data coded those countries. The two datasets were then joined, and the CO2 emissions were visualized on the map using a scale with green being the least and red being the most emissions. (See Chunk 12-14)

Analyze the top 3 countries that emit the most CO2 by geographic region

Wanting to see which top 3 countries in each continent emitted the most CO2, we plotted this data by first creating separate datasets for each continent with the top 3 CO2 emitters for each one. Then these datasets were merged and plotted. United States is the most CO2 emitter out of the other countries around the world. (See Chunk 15)

Aim 1.2B CO2 Emissions by Area

Pearson's correlation was run to assess association between CO2 emissions and country area. The null hypothesis is that there is no association, and the alternate is that there is an association. This test was done because we hypothesized that a bigger area would allow the country to have more economic opportunities to build industries or for populations to increase, which would lead to more emissions. Pearson's r was 0.5 with a p-value < 0.05, indicating a mild positive association between area and the country's total CO2 emissions. (See Chunk {r CO2 Emissions by Area}).

CO2 Emissions in Developed vs. Non-Developed

Data from OurWorldinData cleaned and loaded into R. (See Chunk 17).

Finding CO2 per capita per country:

CO2 per capita data from OurWorldinData was used to visualize the country differences. The top 10 countries in that data set were plotted. As CO2 per capita is calculated by total emissions over total population, countries with smaller populations are more likely to have higher CO2 per capita rates, which is what the plot shows. (See Chunk 19).

CO2 by GDP

GDP data from OurWorldinData was merged with our CO2 data to obtain GDP reports. The country's total GDP was divided by the population to obtain GDP per capita. To classify a country as developed vs. non-developed, we set a threshold at $9,000 per capita. Originally, this was $12,000 based on Investopedia, but this was too conservative as countries like USA, Germany, and UK were being classified as developing. Results show that developed countries emit more CO2 than developing countries. (See Chunk 20).

Aim 2.1. Examine global CO2 emission in pre-pandemic and pandemic years.

Initial Analysis

First, to examine global CO2 emission in pre-pandemic and pandemic years, a subset of data about 2018, 2019, and 2020 CO2 emission in all countries is created. See chunk: {r subset pandemic years}

Then, we want to explore the general trend of changes in CO2 emission in all countries possibly associated with the pandemic. Thus, we presented the global cumulative CO2 emission during these three years through bar plot and separately analyzed the cumulative co2 emission of top5 emission countries. See chunk: {r cummulative co2 sum} and {r cummlative co2 in top5}

Further Analysis with new Data

Since we did not observed much difference as expected, we joined another data set from "Our World in Data" with information about absolute growth of CO2 and other information about CO2 per capita/GDP/unit energy to explore further on pandemic-CO2 emission relationship.
We read our new data set about "owid-co2-data" from github and filtered the wanted period. See chunk: {r new data set}

Next we joined the wanted variables with the current data set using sql in r. See chunk: {r sql join}

Before plotting the absolute growth of co2 emission during the pandemic in a global perspective, we first checked and corrected missing values in our interested data. See chunk: {r check NA}

Then, we visualized our results regarding absolute growth of CO2 emission in both global level and top 10 emission countries. We also grouped the results by co2 emission per capita as a comparison. See chunk: {r abs co2 visualization}

Heatmap

Finally, we used interactive heatmap to present our data in a broader view to see how features differ in the timescale. See chunk: {r heatmap top10 emission} and {r heatmap top10 populated}.

Aim 2.2 Time Series Prediction

Data Preparation

Generally, we decided to predict the CO2 emissions in the next 30 years by using the data from 1992 to 2019, and fit it into proper ARIMA models to do the prediction. We choose see the prediction of CO2 emissions in six countries and five continents. To see the data preparation before the time series analysis, please check the chunk: {r extract and prepare the data for forecasting}

Six Selected Countries

We chose to take a closer look at CO2 emissions in six countries including the U.S., the U.K., Germany, China, Russia, and France. We extracted the data for the six countries separately and set as time series variables using the `ts()` function to fit in time series models using the auto.arima() function from the `forecast` package. This function would directly return the best ARIMA model according to either AIC, AICc or BIC value of the data. Finally, use the `forecast()` function to generate the predicted values. See chunk: {r forecast on 6 countries}

Then, we plotted the predicted CO2 emissions for these six countries. We were expecting to see what's the increasing or decreasing trend of the CO2 emissions in the next 30 years. In general, all the CO2 emissions would keep increase, but if we look at the plots one by one, for example, we noticed that the growth of CO2 emissions in the U.K. is actually slowing down, while in China, it is exponentially increasing over the next 30 years. See chunk: {r plot the predictions of the six countries}

Thirdly, we plotted all the predictions based on the same scales, from 2.75e+10 to 6.5e+11, to compare the predicted CO2 emissions among the six countries. Compared to the U.S. and China, the rest of four

countries have linear but flat increasing trend which implies the growth is positive but slowly. In the current year, the U.S. has twice as much CO2 emissions as China, but according to our prediction, China will catch up and have similar amount of CO2 emissions by 2050, as it is growing exponentially. See chunk: {r plot the predictions based on the same scale (countries)}

Five Continents

To the forecast on CO2 emissions by continents, we filtered out the data for the five continents separately, including Americas, Asia, Africa, Europe, and Oceania, and set as time series variables using the `ts()` function to fit in time series models using the auto.arima() function. Finally, use the `forecast()` function to generate the predicted values. See chunk: {r extract and prepare the continent data} and {r time series model fits for five continents}

Then, we plotted the predicted CO2 emissions for these five continents. We were expecting to see what's the changing trend of CO2 emissions in the next 30 years. In general, all the CO2 emissions would keep increase, but if we look at the plots one by one, we noticed that only the growth of CO2 emissions in Asia is exponentially increasing over the next 30 years, while in other continents, the growth are all approximately linear positive. See chunk: {r plot predictions on CO2 emissions by continents}

Thirdly, we plotted all the predictions based on the same scales, from 1e+10 to 2e+12, to compare the predicted CO2 emissions among the five continents. Africa and Oceania have linear and flat increasing trends which implies the growth is very slow. The Americas and Europe both have similar linear growth trends, while Asia has an exponentially growth trend and will surpass the Americas and Europe by 2050. See chunk: {r plot the predictions based on the same scale (continents)}