Data Engineering 1: Term 2 Project Report

Naida Dzigal

December 4, 2022

1 Project Goal

The goal of this project was to use Knime, an open source Analytics Platform software to build an analysis pipeline to gain first insights into an analytical question. The topic chosen was to review with which fuel sources countries around the world were satisfying their electricity production needs in the year 2010. The data for nuclear, oil, gas and coal electricity production was plotted on bar plots. Finally, due to the overwhelming monopoly oil, gas and coal seem to have in the electricity production of most countries around the world in 2010, a final plot of GDP-PPP versus population size was plotted for 2010 and compared to 1990 to review possible trends.

2 Executive Summary

In the past 60 years, the ability to produce nuclear energy has been associated to highly developed nations whereas oil, gas and coal are remnants of the industrial revolution. The claim that this project tries to give some insight to is that countries have prioritised cheap energy sources for curbing the needs of their growing populations, but also do not have a enough money to invest in more sustainable electricity production (such as nuclear).

Due to the visualization limitations of Knime, only two types of electricity production were looked at:

1. Nuclear electricity production
2. Oil, gas and coal electricity production

It was found that in 2010, nuclear consisted of a very low percentage of any country's total electricity production whereas over half of the countries worldwide satisfied more than 60% of their electrical energy needs by burning oil, gas and coal.

3 Knime Analytics

The Knime pipeline of this project consists of 4 sections:

1. Data loading from MySQL
2. Bar charts on electricity production from nuclear and oil, gas and coal sources for 1990 and 2010.
3. Data loading from APIs
4. Scatter plots of GDP-PPP versus population growth for the years 1990 and 2010.

Each of the above are described in the following sections in more detail.

3.1 Data Sets from MySQL

The data was initially downloaded using the World Bank APIs, accessed through the Postman software. Two datasets for 266 countries (data from 1960 - 2021) were acquired for the following variables:

1. Electricity production from oil gas and coal sources, indicator EG.ELC.FOSL.ZS linked here.
2. Electricity production from nuclear sources, indicator EG.ELC.NUCL.ZS linked here.

Once the data was downloaded, it was imported into the MySQL Workbench, first by creating a schema named wdi, and then a separate table for each of the datasets above. The tables were named: wdi_oilgascoal, wdi_nuclear. These tables were populated using csv files and it may be necessary to edit the file paths when reproducing this project. Each table was tested for import failures with the SELECT * FROM (table name) statement after importing, along with DESCRIBE (table name) to review the structure of imported tables. These tables made up the first operational data layer (see Figure 1.)

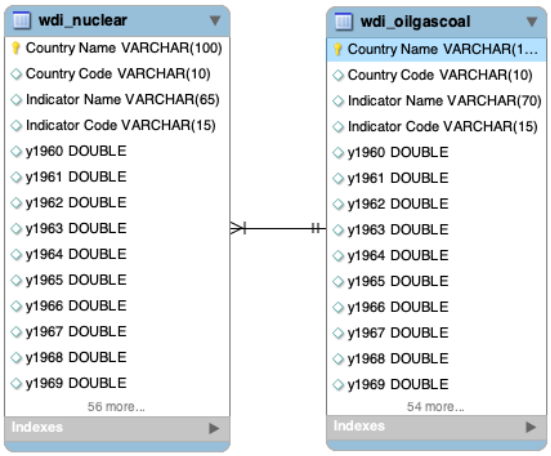

Figure 1: The EER diagram of the operational layer with two database tables in MySQL. Note that not all fields are shown (e.g. there are hidden year fields that start at y1960 and go to y2021.

3.2 Load data to Knime

Using the MySQL Connector node, the operational data layer described in the previous section, was loaded into a Knime workflow. The data were cleaned to only show nuclear (or oil, gas and coal) for the year 2010 (or 1990). The other two columns used were Country Name and Country Code.

Thus, each of the two tables wdi_nuclear and wdi_oilgascoal before joining consisted of 1 dimension for each fact (i.e. country). When joined (by Country Name), they contained a total of 2 dimensions (see Figures 2. and 3.)

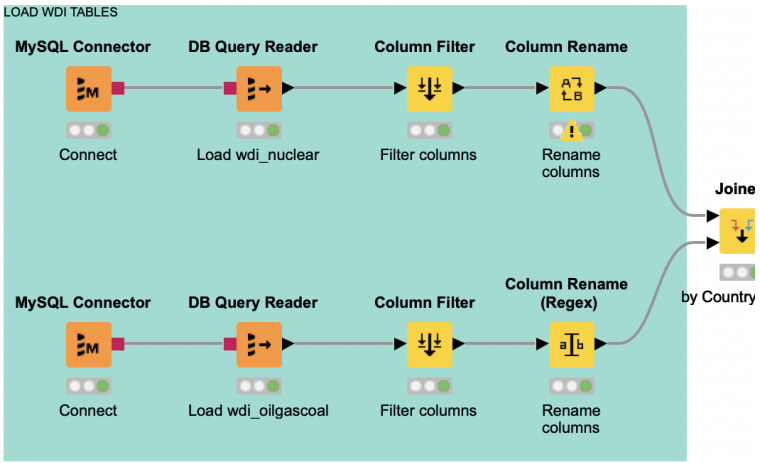

Figure 2: The ETL pipeline depicting the loading, transformation, cleaning and extraction of data from the WDI schema.

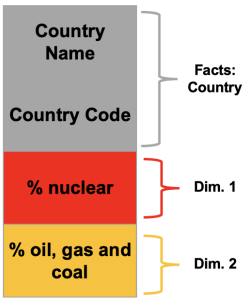

Figure 3: Dimensions of joined table from MySQL after filtering for a certain year.

3.3 Data Analytics

As stated above, one of the aims of this project is to look at how prevalent nuclear sources are for electricity production versus oil, gas and coal in the year 2010. To make a visualization for this purpose, two pipelines were set up for each electricity production source. After filtering rows for missing values, column filtering and renaming, the data were binned into 5 bins, with a single frequency count per country. See Figure 4 for the Knime pipeline for these bar charts and Figures 5, 6, 7 and 8 for the bar charts that show the following:

1. almost all countries have some electricity production from oil, gas and coal sources;
2. some countries have no nuclear electricity production;
3. Most countries that do have nuclear, only satisfy up to 16% of their electricity needs this way;
4. more than 60% of the countries are on a majority oil, gas and coal electricity production diet.
5. The same proportions for nuclear can be observed for the year 1990 (see Figure 6).
6. An increase in oil, gas and coal electricity production can be observed since the year 1990 (see Figure 8).

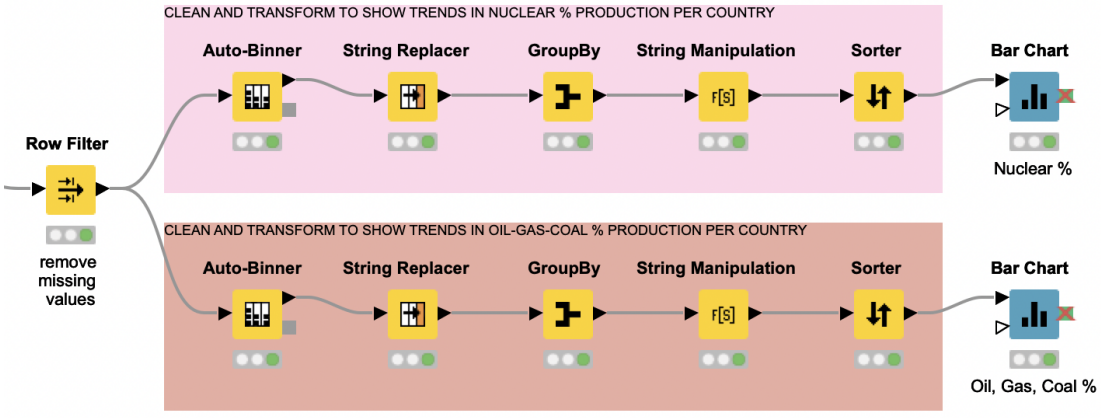

Figure 4: Knime pipeline for cleaning, transforming and plotting the data to produce two bar charts: (top) % nuclear and (bottom) % oil, gas and coal of total electricity production per country for the year 2010.

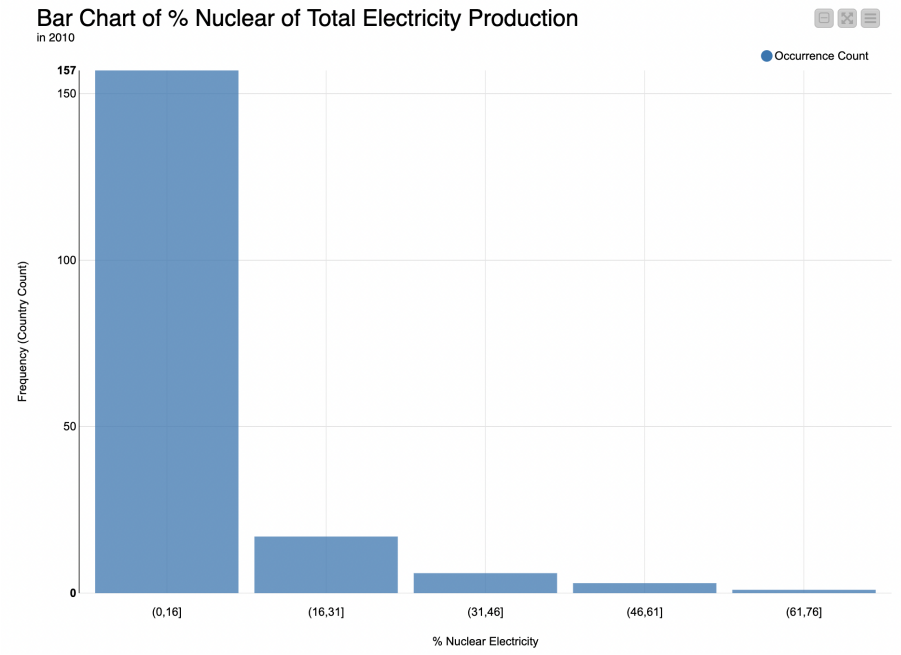

Figure 5: Bar chart for % nuclear of total electricity production per country for the year 2010.

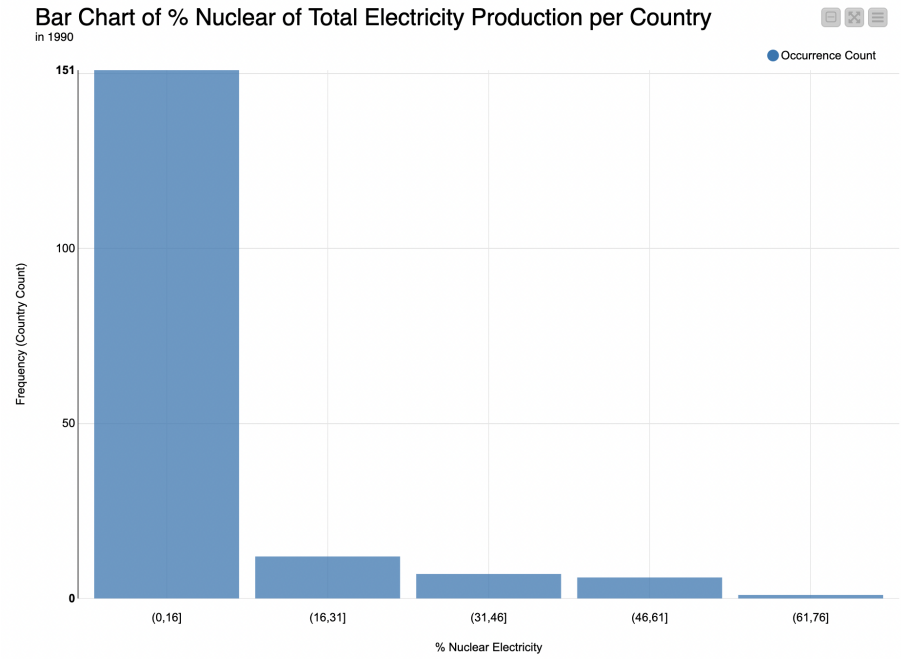

Figure 6: Bar chart of % nuclear in total electricity production by country for the year 1990. The same trend is observed as in Figure 5 (for the year 2010).

3.4 GDP-PPP versus population growth

As the bar charts only provide a snapshot of the trends in electricity production, it was decided to investigate the current state of population versus GDP-PPP in the year 2010 and compare it to a period 20 years ago (year 1990).

As such, GDP (Gross Domestic Product) is important because it gives information about the size of the economy and how an economy is performing. The growth rate of real GDP is often used as an indicator of the general health of the economy. In broad terms, an increase in real GDP is interpreted as a sign that the economy is doing well. Also, nations that have higher GDP are more likely to invest in sustainable policies and finance the growth of greener electricity production.

On the other hand, even though population growth should lead to a continuation of global GDP growth, the initial impact may be modest because growth in the working-age population is

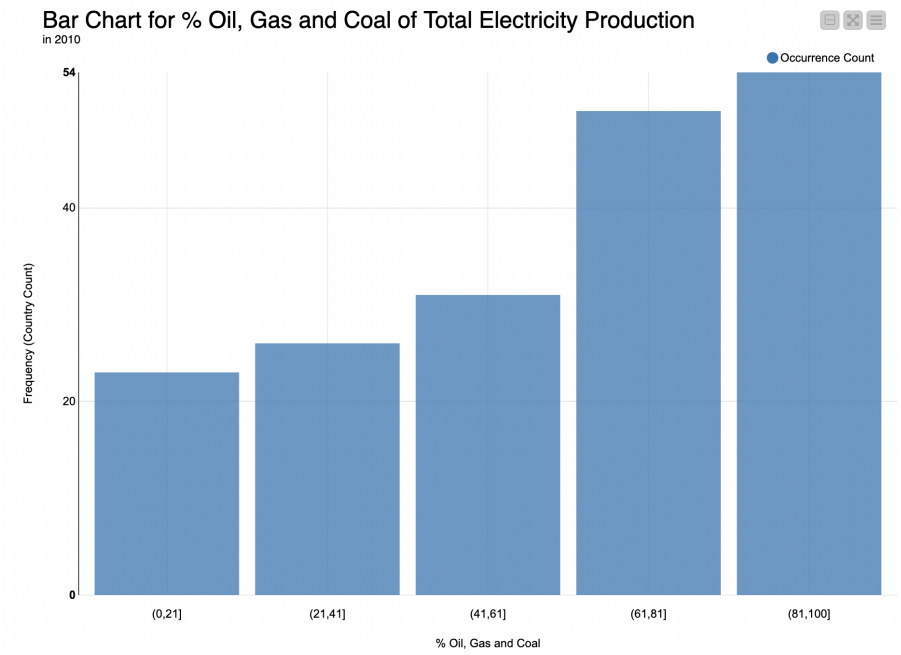

Figure 7: Bar chart for % oil, gas and coal of total electricity production per country for the year 2010.

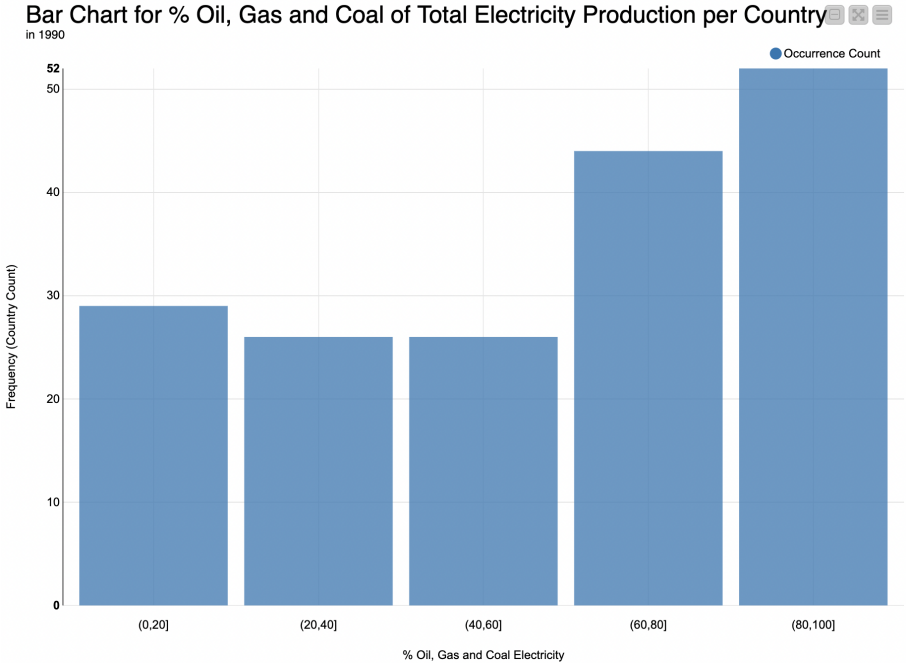

Figure 8: Bar chart of % oil, gas and coal in total electricity production by country for the year 1990. An increase in oil, gas and coal needs is observed compared to Figure 7 between 1990-2010.

concentrated in countries with low labor productivity. As a consequence, a sustained economic growth is manageable only for a few countries that can continue their economic growth at a higher rate than their population growth. By plotting GDP-PPP versus population for the years 1990 and 2010, it may be possible to observe general trends such as whether some countries are having a population growth with no increase in GDP, which would in turn justify the investment into fossil fuels.

The Knime pipeline to plot the scatterplots of the population versus GDP for the years 1990 and 2010 is shows in Figures 9 (API json data download) and 10 (cleaning and filtering the datasets, joining and plotting). Note that the indicators used from the World Development Indicators site are: Population, indicator SP.POP.TOTL and GDP, PPP (current international $), indicator NY.GDP.MKTP.PP.CD.

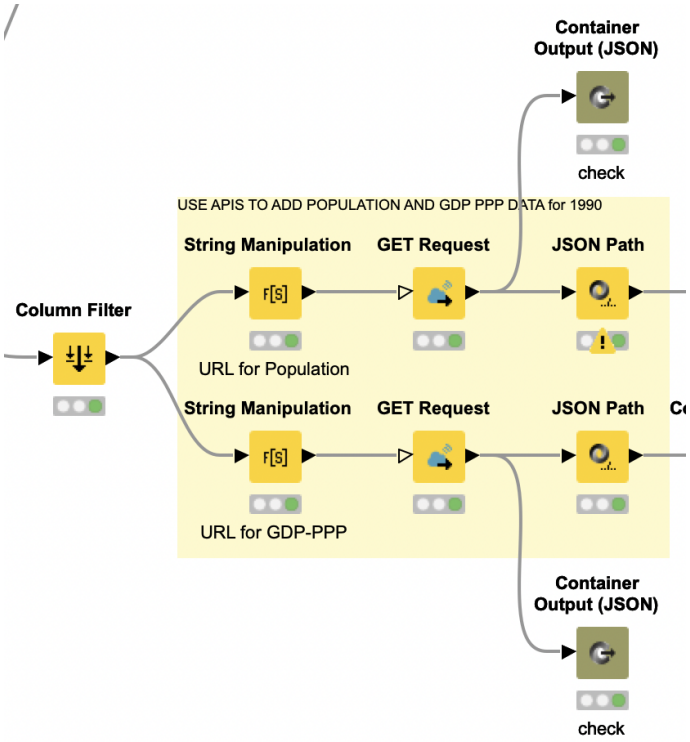

Figure 9: A series of nodes for querying the WDI API for obtaining data on the GDP-PPP and population per country for the year 1990. A second identical pipeline was constructed to plot the same variables for the year 2010.

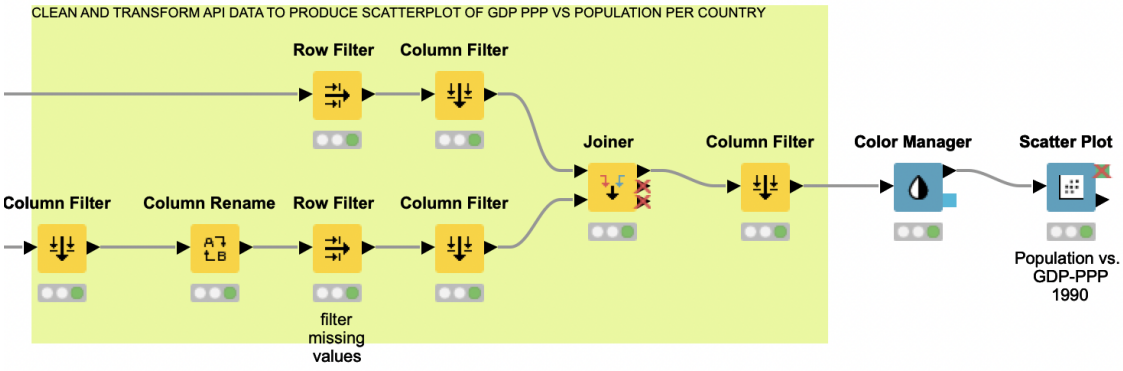

Figure 10: Removing missing values, renaming columns, filtering, adding colours and plotting a scatterplot for population versus GDP-PPP for the year 1990. A second identical pipeline was constructed to plot the same variables for the year 2010.

From figures 11 and 12 we can observe the following:

1. The scale on the population axis is roughly the same, whereas the scale on the y axis has increased almost 2.5 times. This indicates a general stretch in growth of the overall economy in the years between 1990 and 2010.
2. GDP, PPP (current international $) seems to have more than doubled for some countries (at roughly the same population).
3. There are more countries in the far right end of the population axis, indicating a high population growth at a smaller than average economic growth.

Overall, the analytics performed in this section are not conclusive. We were not able to assess that solely population growth was the reason why the world has been moving towards a higher oil, gas and coal diet. Moreover, the world's GDP close to tripled between 1990 and 2010, indicating

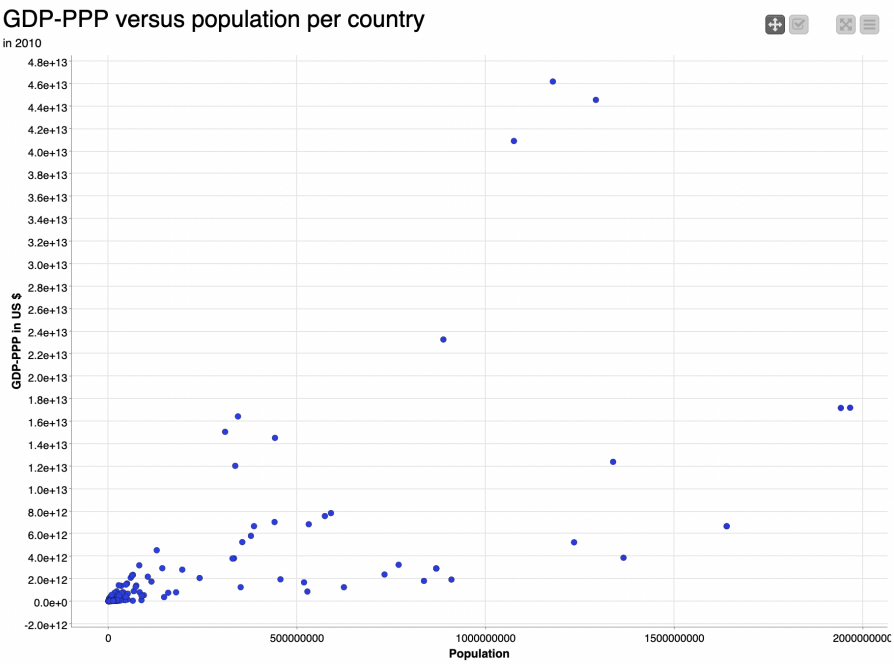

Figure 11: Scatterplot of GDP-PPP versus population numbers for year 2010.

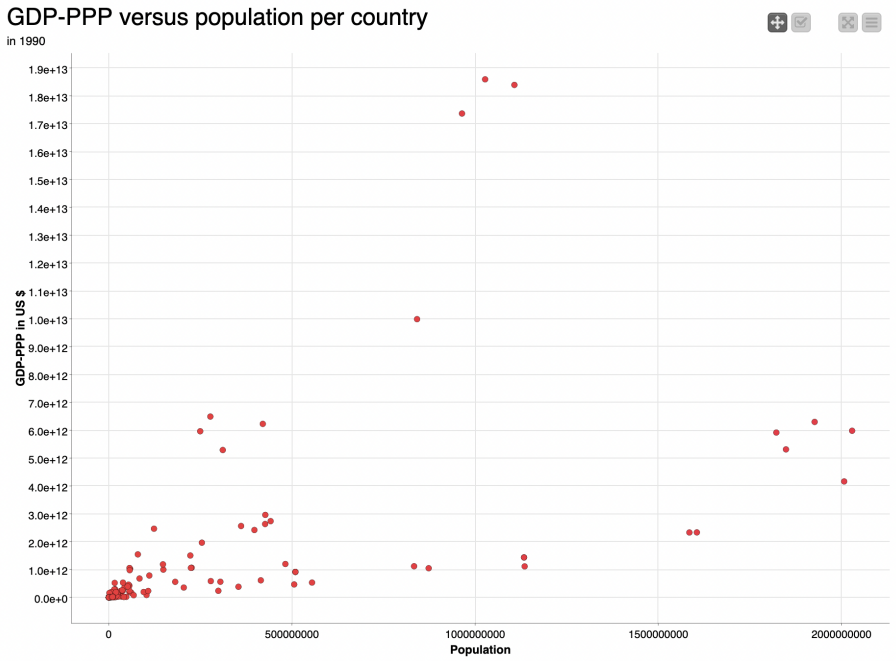

Figure 12: Scatterplot of GDP-PPP versus population numbers for year 1990.

more resources to invest in more sustainable energy solutions. The relatively stable population sizes (with few exceptions) were not indicative of a sudden rushed increase of need in energy that would prefer oil, gas and coal plants over longer-term investments such as nuclear.

It thus seems that the world between 1990 and 2010 went through an irresponsible economic boom and the consumption of fossil fuels was driven more by this boost in the economy than a major population growth.

4 Conclusion

The project aimed to look at the amount of electricity production from nuclear versus oil, gas and coal sources in the year 2010 and put that into the context of the economy by using GDP-PPP and population sizes as proxies for the health of the economy. It was found that a single snapshot in the year 2010 was not enough and a comparison to 20 years prior was helpful in providing context to the plots made for year 2010.

Secondly, it was found that the nuclear electricity production trends did not change between 1990-2010, but that there was an increase in oil, gas and coal electricity production during the same period. This push for cheaper electricity production by using oil, gas and coal versus more sustainable sources such as nuclear is indicative of an irresponsible growth trend and poor energy diet worldwide.

Still, the variables looked at and trends plotted are not sufficient to conclude that the use of oil, gas and coal was purely irresponsible as other reasons behind such investments may exist. It is proposed to investigate these trends for a longer period than 1990-2010 and to include a number of other variables such as country development index, purchasing power, political background, climate change trends, migration trends, etc. in the assessment.