During the data cleaning phase, we removed unnecessary values, countries and inconsistent data. We were prepared to load the data into R to start our analysis. But we first needed to categorize the countries into categories like Developed, Developing and Transition economies.. We referred the data from the United Nations Conference On Trade And Development website to obtain the list of countries for our categorization.

Then when we tried to run basic descriptive statistics like mean, median etc, we encountered a problem. Our datasets had the “years” as the columns and countries as rows. This made even basic statistics and running commands grouped by countries difficult. So we used the transpose function to transpose the .csv files so that the rows and columns were interchanged and were easier to use.

> c=read.csv("Countries classification\_consumption.csv")

> t(c)

This t() function help us transpose the data frame. We later saved it as file for later use.

Then we had to remove empty rows because the data was classified by type with spaces in between.

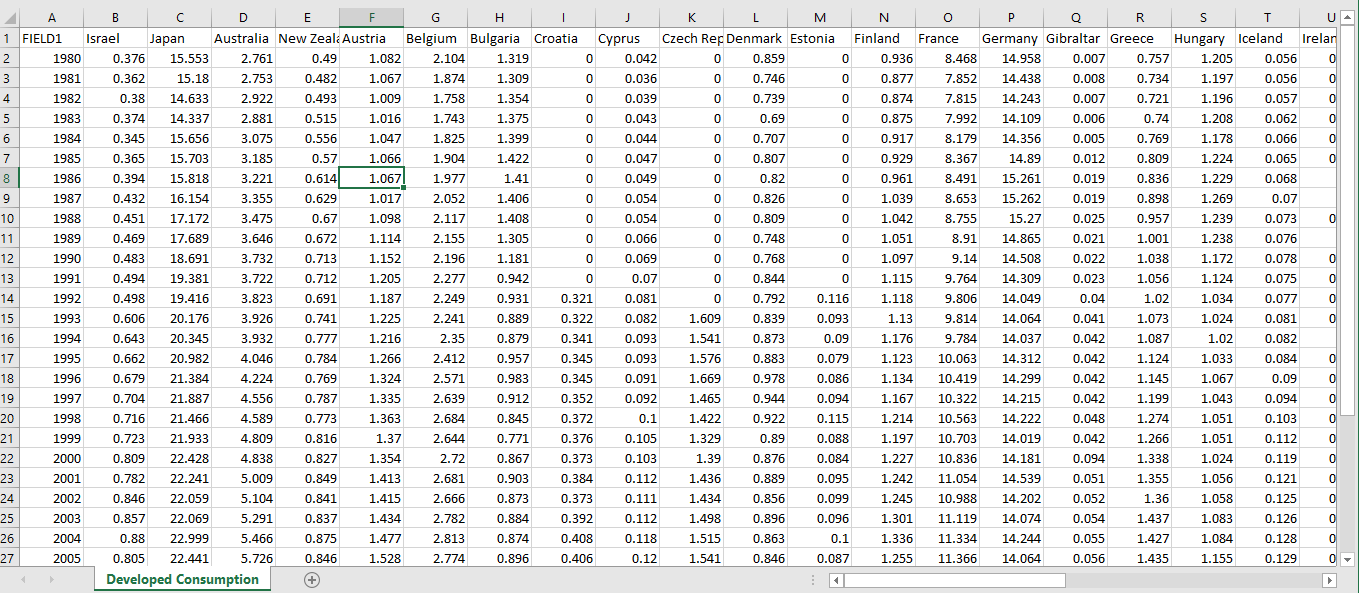
> c=read.csv("Countries classification\_consumption.csv")

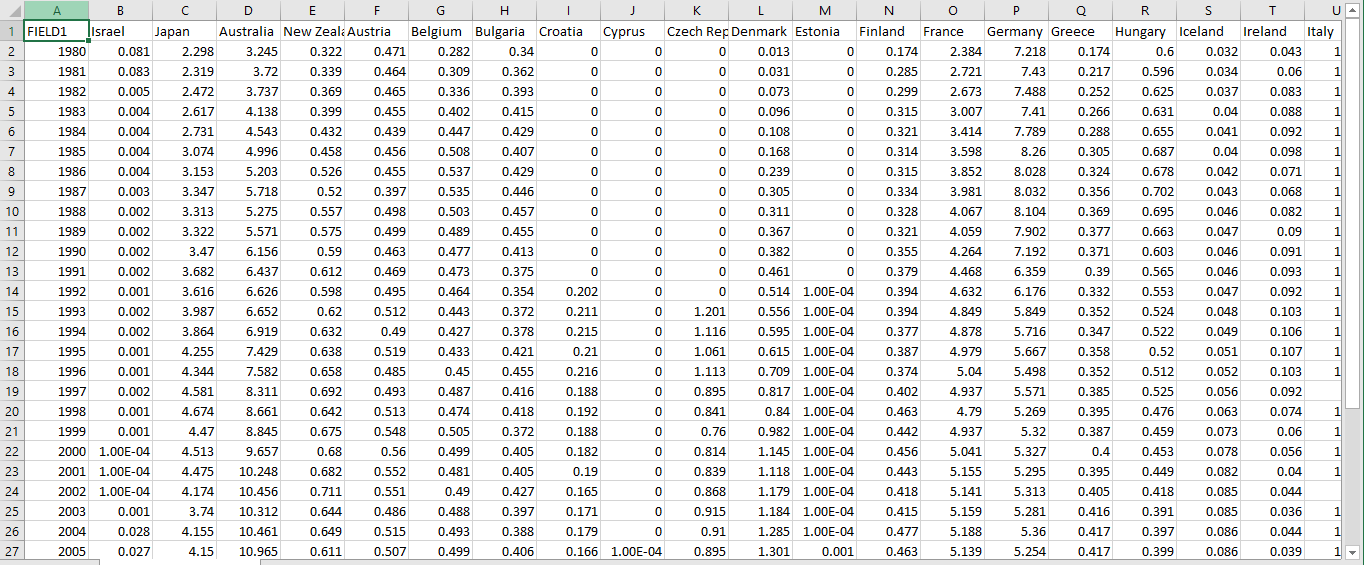
> mean(c$X1980)

[1] NA

We could have used the na.rm parameter to exclude empty rows but we are planning to perform analyses for different types of countries (developed etc.) and not as a whole. So we split the consumption and production files further into multiple files. So according to the type of economy (developed, developing, transition) the 2 production and consumption files were split further. So we now have 6 files in total. These files allow us to conduct analyses easier.

Then we used the R function cor() to obtain the correlation between 2 sets of data.





> dc=read.csv("Developed Consumption.csv")

> dp=read.csv("Developed Production.csv")

So we load these two files into 2 separate objects.

Now we can run cor() function for unique countries from these two files to find their correlation coefficients.

For example

> cor(dc$Japan,dp$Japan)

[1] 0.6795989

> cor(dc$Canada,dp$Canada)

[1] 0.9854069

> cor(dc$Israel,dp$Israel)

[1] 0.3606118

> cor(dc$United.States,dp$United.States)

[1] 0.6631786

So by just running these commands for few countries we could observe that some countries like Israel have below a relatively weak correlation between the energy they consume and produce. Countries like United States and Japan have a moderate (~0.66) correlation coefficient. Some countries like Canada have a strong correlation (0.98). We are trying to match these correlation coefficients with some other external reports (like Energy Independence) so that it can validate our future analyses more strongly.

We are also currently trying to populate our data frame entries into a special type of matrix. This will allow us to obtain data without having to enter commands with the names of different countries manually. This matrix will only have data where the countries match each other. This will be clearer when we see the below example.

> cor(dc,dp)

FIELD1 Israel Japan Australia New.Zealand Austria

FIELD1 1.0000000 0.41748184 0.27464696 0.9610542 0.8406056 0.7973781

Israel 0.9785999 0.36061178 0.35508462 0.9621230 0.8376815 0.8128381

Japan 0.8395256 0.04429691 0.67959892 0.8671845 0.8908334 0.7095396

Australia 0.9861119 0.37140525 0.33914818 0.9798709 0.8196956 0.7738356

New.Zealand 0.9266873 0.14849617 0.54274352 0.9183079 0.9469326 0.7496230

Austria 0.9458793 0.31265254 0.44518443 0.9484485 0.8071142 0.7842260

Belgium 0.8994621 0.20821158 0.57242437 0.9229862 0.8619257 0.7300080

Bulgaria -0.8643528 -0.22477844 -0.49001559 -0.8503704 -0.8435981 -0.7869485

Croatia 0.8622473 0.18333986 0.52806545 0.8570949 0.8059699 0.7614682

Cyprus 0.9688158 0.29828474 0.44920531 0.9566848 0.8657499 0.7899305

Czech.Republic 0.8647786 0.25090440 0.44606004 0.8419778 0.7871021 0.7419410

Denmark 0.2499567 -0.29391460 0.76960604 0.3255719 0.4731928 0.1767836

Estonia 0.7860284 0.10162943 0.53284489 0.7800692 0.7770991 0.7145977

Finland 0.8931436 0.13005048 0.57450525 0.9142103 0.8847278 0.6936281

France 0.9169153 0.17893253 0.55988917 0.9295209 0.8728380 0.7472428

Germany -0.7002779 -0.36586885 -0.06359277 -0.6554757 -0.5139644 -0.6202318

Gibraltar 0.8412964 0.28653079 0.23366228 0.8602445 0.7708288 0.7115632

Greece 0.8913410 0.10224121 0.60827809 0.9416893 0.8602142 0.7080719

Hungary -0.7178102 -0.27151607 -0.23328509 -0.6276231 -0.7087275 -0.7524475

Iceland 0.9310961 0.57989543 0.04662857 0.8908684 0.6855979 0.7409029

Ireland 0.9180018 0.27102338 0.45878534 0.9527590 0.7641015 0.7520999

Italy 0.8594683 0.08955479 0.62004363 0.9115983 0.8358241 0.6735180

Latvia 0.8016175 0.13985101 0.51150881 0.7928220 0.7768429 0.7081431

Lithuania 0.7286605 0.06266936 0.55952202 0.7214944 0.7426432 0.6536109

Luxembourg 0.9632786 0.36868947 0.35929576 0.9621068 0.7910690 0.7532228

Malta 0.8352142 0.58188037 -0.03904549 0.7768906 0.6465610 0.6778728

Netherlands 0.9468450 0.25873238 0.47889515 0.9608783 0.8697834 0.7332281

Norway 0.8905123 0.08359241 0.52089847 0.9070682 0.8917583 0.7939643

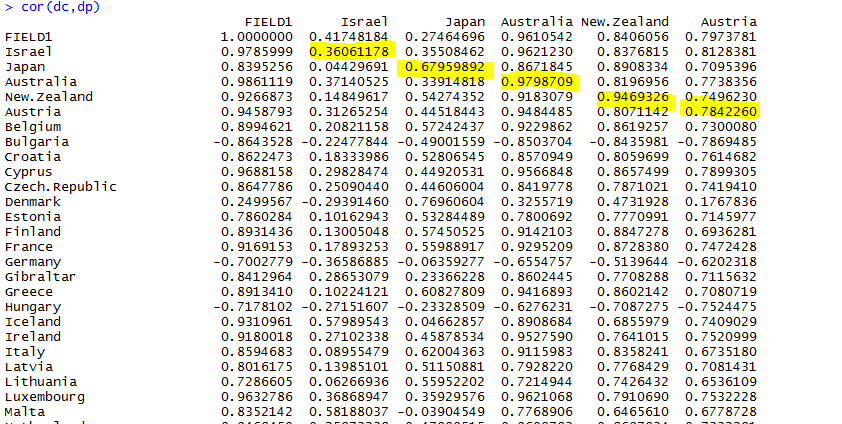
Poland -0.6183964 0.10570219 -0.60285670 -0.6589509 -0.7056461 -0.5999859

Portugal 0.9327621 0.15673160 0.55135957 0.9462019 0.8932659 0.7601564

Romania -0.9113010 -0.33121472 -0.38796229 -0.8917510 -0.7943980 -0.8091604

Slovakia -0.8496220 -0.19340482 -0.50257784 -0.8371734 -0.7984439 -0.7603155

So as you can see it runs the test for all the countries from the consumption table paired along with all the countries the production table. But we need only the entries that are paired with the same countries and other entries can be ignored i.e. the we only need the highlighted values shown below. So this is where our specialized table comes into play. That table will not have values paired with different countries. It will only have values that are populated if both the row and the column match by country name.



Another approach we are also trying to take is to try and store the data obtained by running the above command into an excel file. Then we can perform data cleaning and manipulation on Excel, so that we obtain the correlation coefficients of different countries easily. This also saves us time, but costs us time in data cleaning. So we are exploring both the matrix/table approach and the manual manipulation of data in excel side by side.

Also we ran functions like ncol() and nrow() to check if the rows and columns are matching.

> nrow(dc)

[1] 34

> nrow(dp)

[1] 34

> ncol(dc)

[1] 43

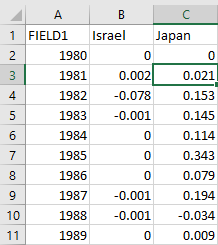
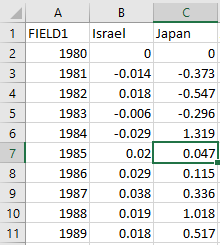
> ncol(dp)

[1] 39

We found that the number of columns did not match. This made us go back to the datasets and compare the countries. We observed that some countries that were present in a dataset were missing from the other. So we performed some more data cleaning activities on our datasets to sort out these discrepancies.

We are also working on finding the correlation between the year-on-year increase or decrease in production and the corresponding year’s year-on-year increase or decrease in consumption.

Below is an illustration of the year-on-year change in production and consumption.



> dc\_inc=read.csv("Developed Consumption Deficit.csv")

> dp\_inc=read.csv("Developed Production Deficit.csv")

> cor(dc\_inc$Israel,dp\_inc$Israel)

[1] -0.200774

> cor(dc\_inc$Japan, dp\_inc$Japan)

[1] 0.4902785

We are running cor() for finding the correlation of these two trends. Also, we will try to store these values in a table/matrix like we mentioned earlier and try to run cor() functions for them as a whole, on one single object instead of writing separate commands for individual countries.

We also ran the summary() function to get some basic values of developed countries’ energy consumption and production.

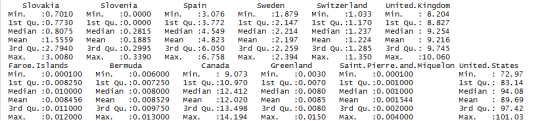
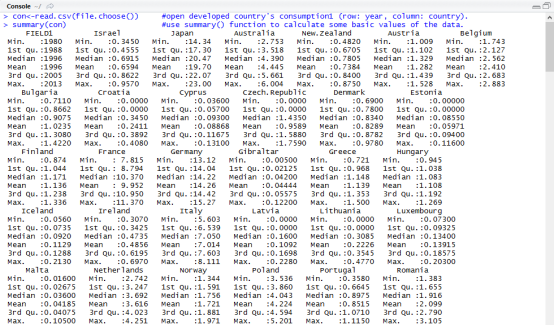
#open developed country's consumption1 (row: year, column: country).

con<-read.csv(file.choose())

#use summary() function to calculate some basic values of the data.

summary(con)

Executing the above code allowed us to see each developed country's minimum consumption and the maximum consumption, median consumption, 1st Qu and 3rd Qu consumption during period from 1980 to 2013.



Also we will try to explore few more research questions one of which is “How production and consumption trends vary in countries in times of wars”. We will conduct similar correlation tests and find the trends and also try to split the years during wartime and peacetime, to observe if there are any significant trends along these lines.