**CINF-5931 Big Data in Enterprise**

**Fall 2017**

****

“**Impact of increasing growth rate of Renewable Energy Resource on different countries and environment”**

Working under Dr. Abukmail

Submitted By:

Shishir Jindal(1538348)

Ramanjit Kaur(1595607)

**Outline:**

1. Purpose
2. Weekly Progress:
3. Week1
4. Week 2
5. Week 3
6. Week 4
7. Week 5
8. Week 6
9. Background Research
10. Data Sets (1-8)
11. Data Visualization
12. Null Values
13. Outlier
14. Data Processing
15. Class 1
16. Class 2
17. Conclusion
18. References

**Purpose:**

The Project aims at showing that how the consumption of non-renewable energy resources is changing due to drift of several countries in world towards using renewable energy resources. This in future will not only cause a positive impact on environment but will also have a big impact on industries and sectors especially dealing with non-renewable energy and renewable energy resources and this will depicted via detailed analysis of several Data Sets with different attributes.

**Weekly Progress:**

**Week 1:**

In the first week, we came up with the project proposal which had an abstract view of what we planned to work on the project. We revised all our big data concepts and decided how we would apply them into our project to get desired output.

**Week 2:**

Started doing background research to understand and familiarize with concepts of law of demand and supply and gathered information necessary for our project. We also collected multiple Data Sets and finalized the Data Sets that were relevant to our project and ruled out the ones which did not contained required information and the ones which did not contain relevant information.

**Week 3:**

Familiarized with tools like R-Studio and studied R Language and revised to work on Octave and Excel. We then decided to use two machine learning techniques-Clustering and Linear Regression. Also during this week we studied about different clustering algorithms and revised the concepts of Linear Regression. We started working with Class 1 Problem using application of Linear Regression which will be described later.

**Week 4:**

Completed working on Class 1 Problem and derived the result of our data analysis. Started working on Class 2 Problem which uses application of clustering using K means algorithm and evaluation algorithms and techniques like Dendograms, Silhouette, Chi Square etc.

**Week 5:**

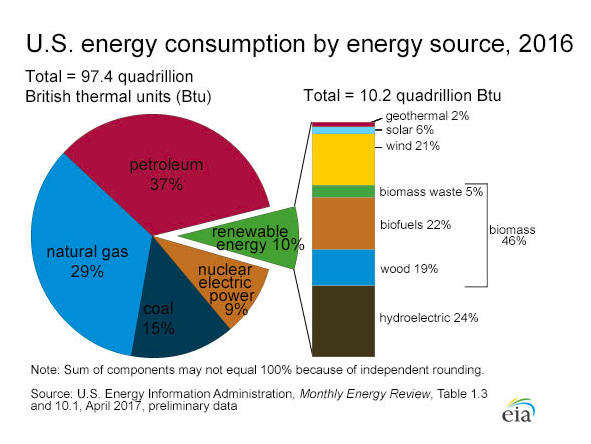
In this week, we finished Class 2 problem. Performed the various evaluation measures and also tried the different visualization measures to evaluate the quality of the clusters formed. Drew out the various scenarios to be able to understand whether or not the results were covering all possible cases and the satisfactory output was available at the end of the evaluation.

**Week 6:**

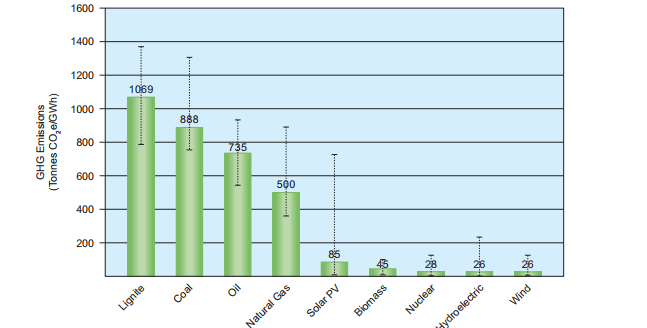
During the last week, we structured all our analysis and think about future work and prepared final presentation.

**Background Research**

**Renewable resources** are the resource which are abundant in nature and can be used again and again. They are important as per the aspect of sustainability. According to the U.S. Energy Information Administration, the most frequently used renewable resources are biomass, water, geothermal, wind and solar. A renewable resource improves public health and environmental quality. Generating electricity from renewable energy rather than fossil fuels offers significant public health benefits. The air and water pollution emitted by coal and natural gas plants is linked to breathing problems, neurological damage, heart attacks, and cancer. Today, the use of renewable energy sources is increasing, especially bio-fuels, solar, and wind. In 2016, about 10% of total U.S. energy consumption was from renewable energy sources (or about 10.2 quadrillion British thermal units (Btu)—1 quadrillion is the number 1 followed by 15 zeros). About 55% of U.S. renewable energy use is by the electric power sector for producing electricity, and about 15% of [U.S. electricity generation](https://www.eia.gov/energyexplained/index.cfm?page=electricity_in_the_united_states) was from renewable energy sources in 2016. This rate of producing energy from the renewable resources is increasing year by year.

Most developing countries have abundant [renewable energy](https://en.wikipedia.org/wiki/Renewable_energy) resources, including [solar energy](https://en.wikipedia.org/wiki/Solar_energy), [wind power](https://en.wikipedia.org/wiki/Wind_power), [geothermal energy](https://en.wikipedia.org/wiki/Geothermal_energy), and [biomass](https://en.wikipedia.org/wiki/Biomass), as well as the ability to manufacture the relatively labor-intensive systems that harness these. By developing such energy sources developing countries can reduce their dependence on oil and natural gas, creating energy portfolios that are less vulnerable to price rises. In many circumstances, these investments can be less expensive than fossil fuel energy systems.

**Non-renewable energy** comes from sources that will run out or will not be replenished in our lifetimes—or even in many, many lifetimes. Most non-renewable energy sources are fossil fuels: coal, petroleum, and natural gas. Carbon is the main element in fossil fuels. Increasing the reliance on renewable energy sources, such as solar, wind and geothermal generators, can also help conserve the dwindling supplies of fossil fuels that remain in the ground.

*Figure: WNA report about CO2 emission*

As per WNA report, Coal fired power plants have the highest GHG emission intensities on a life cycle basis. Although natural gas, and to some degree oil, had noticeably lower GHG emissions, biomass, nuclear, hydroelectric, wind, and solar photo voltaic all had life cycle. GHG emission intensities that are significantly lower than fossil fuel based generation. The use of fossil fuel resources, which are non-renewable, contributes to the [rise of global warming](https://www.c2es.org/publications/multi-gas-contributors-global-climate-change) due to the release of carbon dioxide (among other gases).The [BBC reported](http://www.bbc.co.uk/schools/gcsebitesize/geography/energy_resources/energy_rev1.shtml) in 2014 that at the rates of non-renewable uses in that year, the world would run out oil in 40 years, gas in 50 years and coal in 250 years.

**Regression**

Regression is widely used for prediction and forecasting in field of machine learning. Focus of regression is on the relationship between dependent and one or more independent variables. The “dependent variable” represents the output or effect, or is tested to see if it is the effect. The “independent variables” represent the inputs or causes, or are tested to see if they are the cause. Regression analysis helps to understand how the value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are kept unchanged. In the regression, dependent variable is estimated as function of independent variables which is called regression function. Regression model involves following variables.

1. Independent variables X.
2. Dependent variable Y
3. Unknown parameter θ

## Linear regression

In the Linear regression, dependent variable(Y) is the linear combination of the independent variables(X). Here regression function is known as hypothesis which is defined as below.

Hθ (X) = f(X, θ)

Suppose we have only one independent variable(x), then our hypothesis is defined as below.



The goal is to find some values of θ (known as coefficients), so we can minimize the difference between real and predicted values of dependent variable(y). If we take the values of all θ are zeros, then our predicted value will be zero. Cost function is used as measurement factor of linear regression model and it calculates average squared error for m observations. Cost function is denoted by J (θ) and defined as below.



As we can see from the above formula, if cost is large then, predicted value is far from the real value and if cost is small then, predicted value is nearer to real value. Therefore, we have to minimize cost to meet more accurate prediction.

**Clustering**

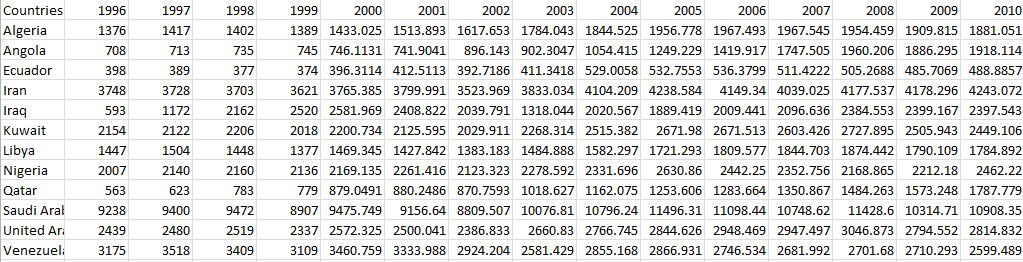
Clustering is the process of making a group of abstract objects into classes of similar objects. Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing. Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.

**Data Sets:**

We have used following data sets for getting desired output:

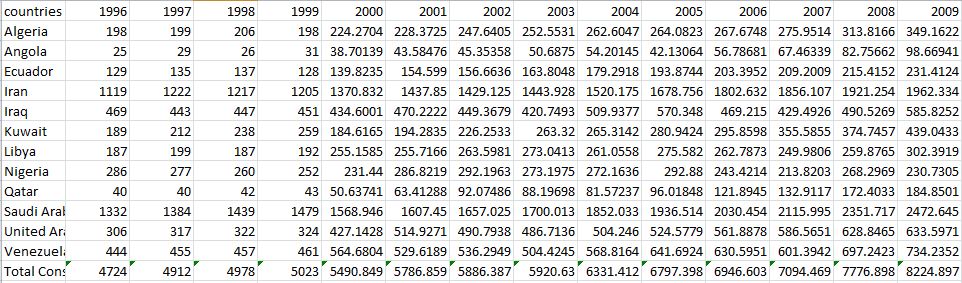
**Data Set 1**

**Production of Petroleum by Organization of the Petroleum Exporting Countries (OPEC) Countries** –

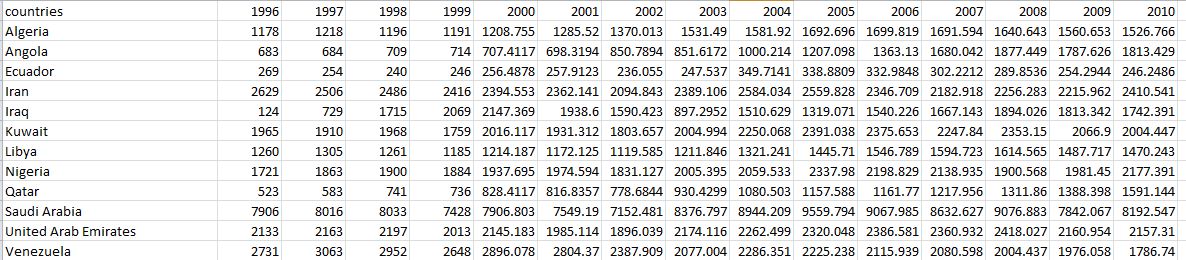


**Data Set 2**

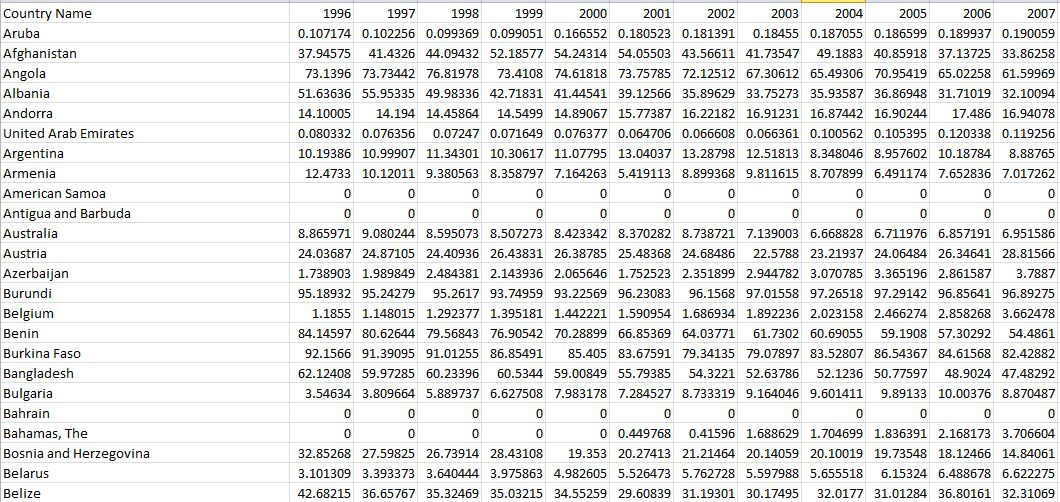
**Consumption of Petroleum by Organization of the Petroleum Exporting Countries (OPEC) Countries** -



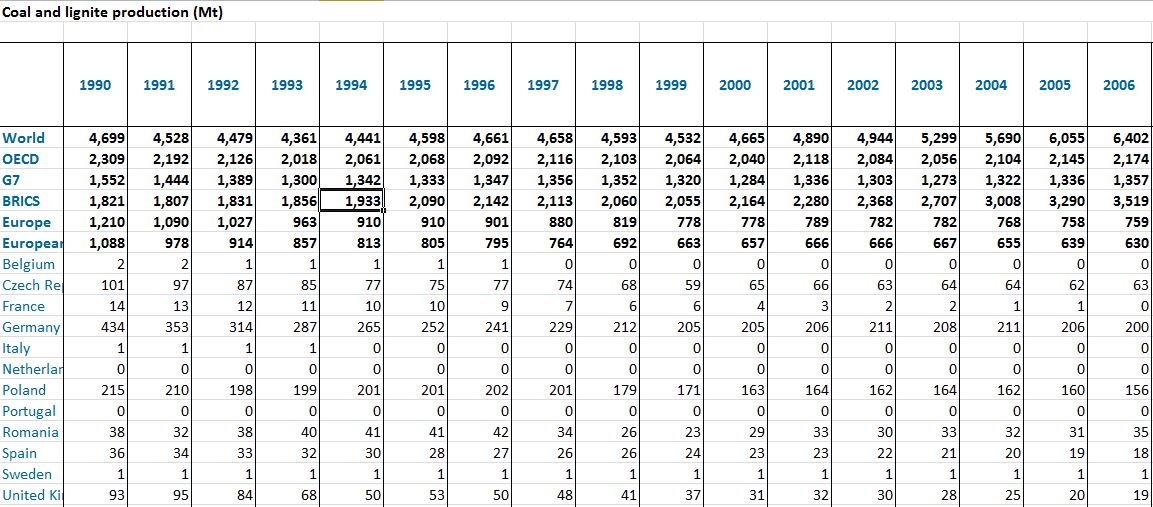
**Data Set 3 (Computed) – Difference between Production and Consumption of Petroleum by OPEC countries-**



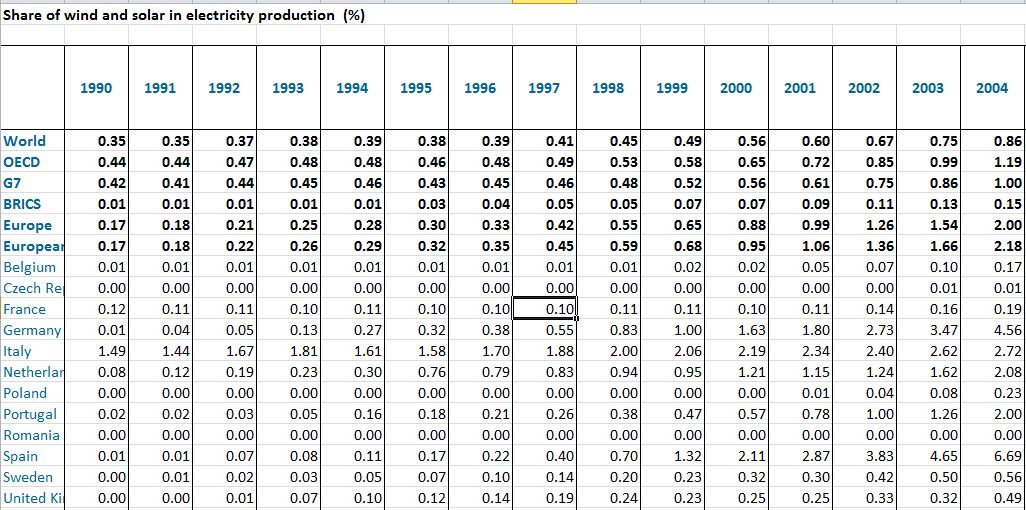
**Data Set 4 - Renewable Energy Consumption by different countries in the world –**



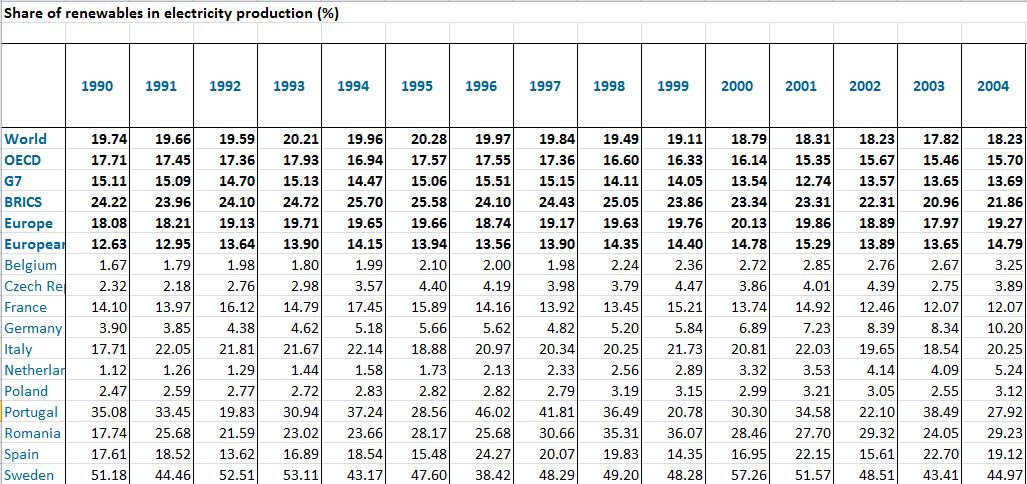
**Data Set 5 - Coal and lignite production**



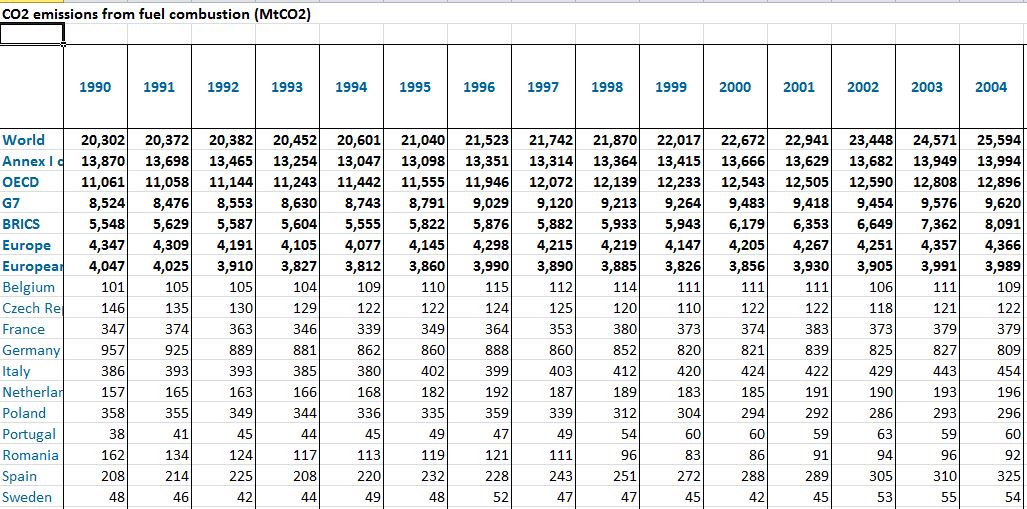
**Data Set 6 – Share of Wind and Solar in Electricity Generation**



**Data Set 7 – Share of Renewable Energy Resources in Electricity Generation**



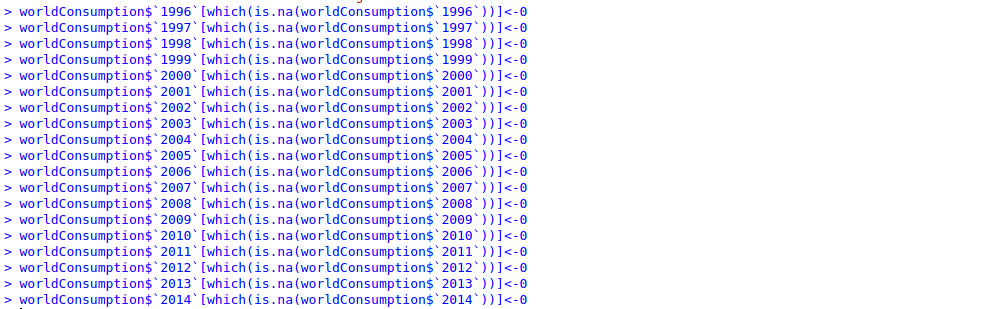
**Data Set 8 – CO2 emissions from fuel consumption**

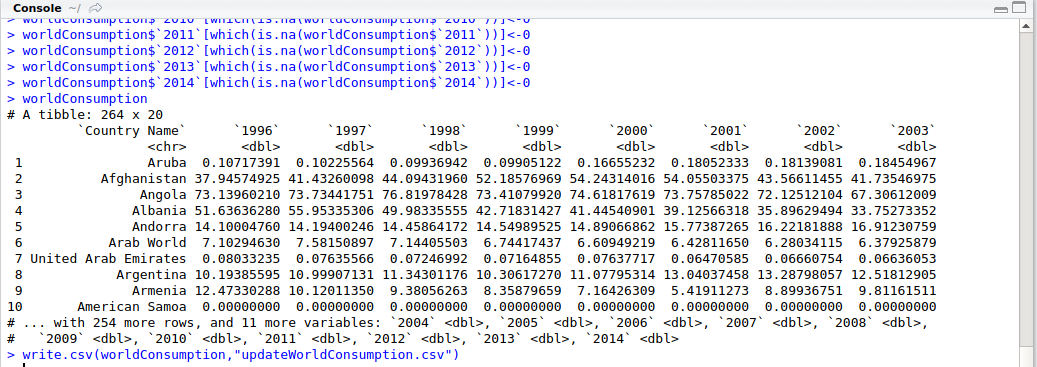


**Data Visualization:**

Different Visualization Techniques such as Histograms, Line Chart, Scatter Plot, and Pie Chart have been used to analyze data. These helped in determining outliers and noisy data.

We removed **null valued** attributes by using following code -





**Outliers -**

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations. We looked into all our Data Sets and removed all the Outliers. This step was really important because the presence of outliers can result in erroneous results.

**Data Processing:**

The entire data Processing is done by dividing the entire project into 2 classes of problems – Class 1 and Class 2.

**Class 1 Problem -**

**How the consumption of Non-Renewable Energy Resources is changing due to increase in demands of Renewable Energy Resources in major countries.**

After detailed background research, we found that that major exporters of non-renewable energy resources are OPEC countries. Organization of the Petroleum Exporting Countries (OPEC) is an [intergovernmental organization](https://en.wikipedia.org/wiki/Intergovernmental_organization) of 14 nations as of May 2017, founded in 1960 in [Baghdad](https://en.wikipedia.org/wiki/Baghdad) by the first five members ([Iran](https://en.wikipedia.org/wiki/Iran), [Iraq](https://en.wikipedia.org/wiki/Iraq), [Kuwait](https://en.wikipedia.org/wiki/Kuwait), [Saudi Arabia](https://en.wikipedia.org/wiki/Saudi_Arabia), [Venezuela](https://en.wikipedia.org/wiki/Venezuela)), and headquartered since 1965 in [Vienna](https://en.wikipedia.org/wiki/Vienna). As of 2016, the 14 countries accounted for an estimated 44 percent of [global oil production](https://en.wikipedia.org/wiki/List_of_countries_by_oil_production) and 73 percent of [the world's "proven" oil reserves](https://en.wikipedia.org/wiki/List_of_countries_by_proven_oil_reserves), giving OPEC a major influence on [global oil prices](https://en.wikipedia.org/wiki/Price_of_oil) that were previously determined by American-dominated [multinational oil companies](https://en.wikipedia.org/wiki/Seven_Sisters_(oil_companies)). So after lot of research, we landed at two Data Sets which depict the production and consumption of Petroleum (or Renewable Energy Resources).

Below is the pattern of data that was observed via application of different charts.

**Production in OPEC countries -** This graph shows how the production of non-renewable energy resources by OPEC countries to different countries.

**Consumption in OPEC countries –** This graph shows export of Petroleum by OPEC countries to different countries.

**Difference between Production and Consumption of Petroleum -**

The following graph is based on the difference of production and consumption of petroleum.

So, according to the law of demand and supply when the price of product in the market increases its demand decreases. Similarly, when the oil prices get higher the demand of petroleum decreases. Therefore, export of Petroleum products by OPEC countries determine the oil prices in the world. So using this concept of law of demand and supply as soon as the difference between production and consumption of increases it clearly implies that the demand (consumption of Petroleum by countries) decreases which is also shown by the following graph. From the graph it can be seen the difference has been increasing in the recent years which implies that demand of Petroleum has been decreasing in the market.

This result was also verified according to news report on following website

<https://www.eia.gov/energyexplained/index.cfm?page=oil_imports>

**U.S. reliance on petroleum imports has declined in recent years**

U.S. petroleum imports peaked in 2005 and generally declined up until 2015. This trend was the result of many factors, including a decline in consumption, increased use of domestic biofuels (ethanol in gasoline and biodiesel in diesel fuel), and increased domestic production of crude oil and hydrocarbon gas liquids. The economic downturn following the financial crisis of 2008, improvements in vehicle fuel economy, and changes in consumer behavior contributed to the decline in U.S. petroleum consumption. Imports increased in 2015 and 2016 along with consumption.

**Renewable Energy Resources Production and Consumption –**

In recent years, there has been an increased growth in the demand or consumption of Renewable Energy Resources. This has caused a very severe impact on market and will continue to grow according to our analysis. After analyzing the data set, we are showing how the leading countries of the world are showing a significant increase in the usage of Renewable Energy Resources using Data Set 4.

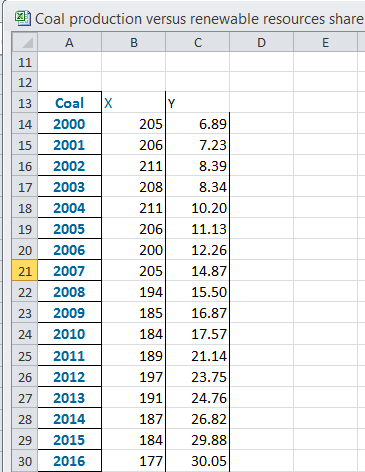
**Application of Linear Regression –**

Our next study of this same case problem uses Linear Regression –

This study shows how coal production is affecting the share of Renewable Energy Resources in production of Electricity by major countries in world. We used Data Set 5 and Data Set 7 for this comparative analysis. We used machine learning technique- Linear Regression for this analysis. Linear Regression helped us to predict future values on the basis of computed Regression Line between coal Production(being an independent variable X) and Share of Renewable Energy Resources in production of Electricity(being a dependent variable Y). Below are different plots which show analysis of above study problem.

**Following Procedure is used –**

1. Extracted the row of Germany from Data Set 5(Coal Production) and made it as X and extracted row of Germany from Data Set 7(Share of Renewable Energy Resources in production of Electricity) and made it as Y.



1. Used R code to compute Linear Regression –

# Center predictors.

coal.c = scale (coal\_renewable$coal, center=TRUE, scale=FALSE)

renewable.c = scale (coal\_renewable$renewable, center=TRUE, scale=FALSE)

# bind these new variables into newdata and display a summary.

new.c.vars = cbind(coal.c, renewable.c)

newdata = cbind(newdata, new.c.vars)

names(newdata)[5:7] = c("coal.c", "renewable.c")

summary(newdata)

# fit a linear model and run a summary of its results.

mod1 = lm (rateofincrease ~ coal.c + renewable.c , data=newdata)

summary(mod1)

1. Finally following analysis was obtained-
2. Checked if the coefficient matched the Y and X intercepts.
3. Analyzed R square value –

**R Square** - R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determinations for multiple regression.

The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by a linear model. Or:

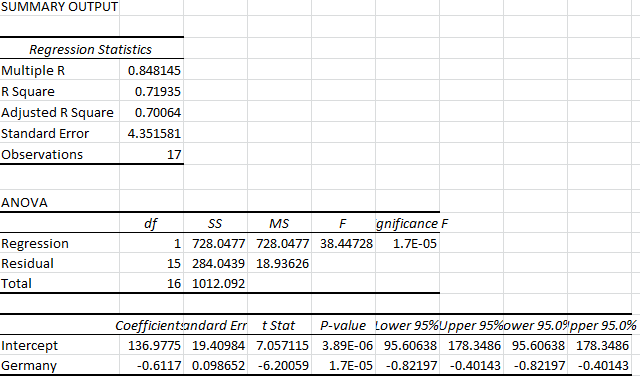
R-squared = Explained variation / Total variation

R-squared is always between 0 and 100%:

0% indicates that the model explains none of the variability of the response data around its mean.

100% indicates that the model explains all the variability of the response data around its mean.

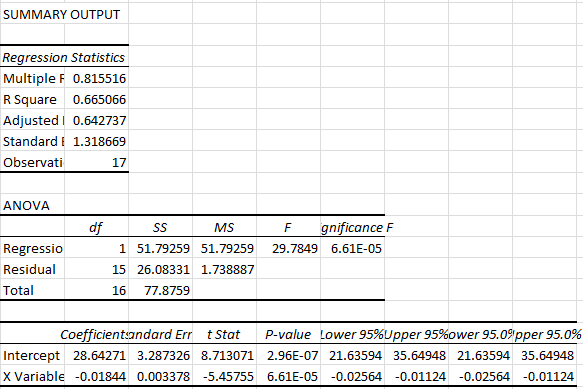
So since R2 in this case is



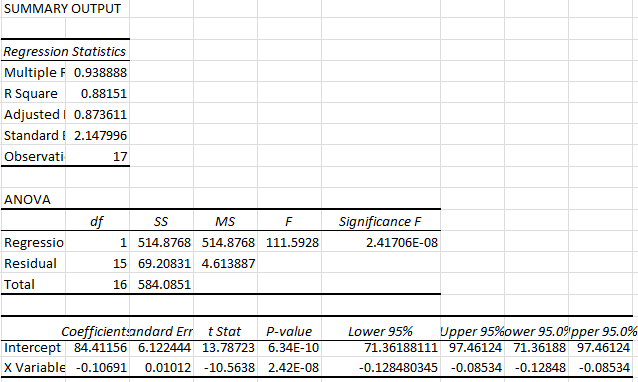
Finally following graph obtained with the regression line which can be used to predict future values –

**Based upon above process Graphs and Regression Analysis was computed following countries and regions –**

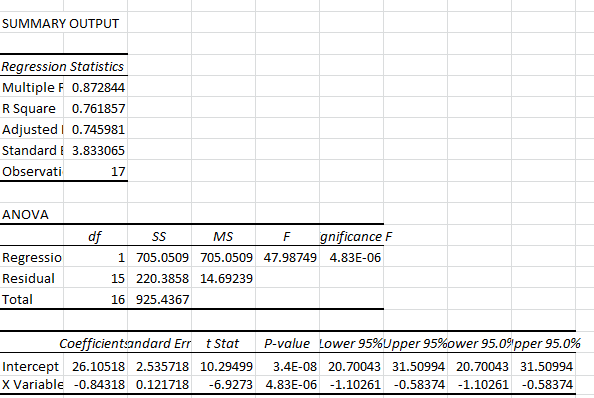
**United States -**



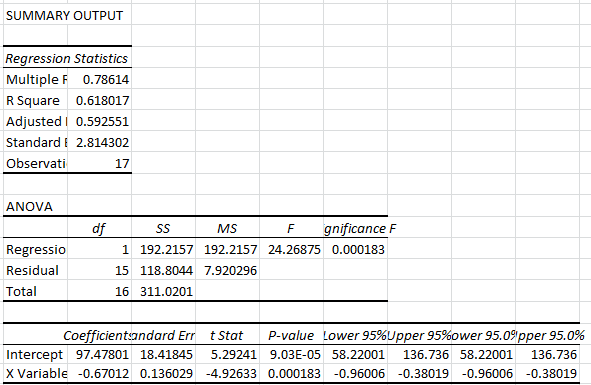
**European Union-**



**United Kingdom -**



**Poland –**



**According to following 2 report on following website it clearly verifies our result that rate of renewable energy production and consumption is expected significantly in coming years –**

1. https://www.clickenergy.com.au/news-blog/12-countries-leading-the-way-in-renewable-energy/

A recent study by Stanford University researchers predicted that the world could be powered entirely by renewable energy in just 20 to 40 years from now. And given that we already have the technology, it’s not that hard to imagine. Almost 50 countries that would be adversely affected by climate change have agreed to make their energy production 100% renewable by the year 2050 and countries all over the world are actively embracing solar, wind, and geothermal energy.

1. <https://cleantechnica.com/2016/02/04/how-11-countries-are-leading-the-shift-to-renewable-energy/>

Who’s embracing wind? Solar? Geothermal? These countries could provide blueprints for the worldwide shift to renewable energy. This December, almost 200 countries from every corner of the world signed the Paris Agreement, committing to decrease greenhouse gas emissions and – dare we say – save the world! The question on everyone’s mind: How? The truth is, we don’t have to wait on scientists to invent some newfangled contraption. The solutions are already here! We simply need to ramp up renewable energy generation, and fast.

**After this finally rate of Change of consumption of Non Renewable Energy Resources versus Rate of consumption of change of Renewable Energy was analyzed –**

Below is the tabular table that we generated on the basis of Dataset 2 used. This has been shown graphically as well. Using this average rate of change the future values can also be predicted using the graph. The average rate of change of consumption of Petroleum from 1996 – 2014 is found to be **0.032874**

**Tabular format of Rate of change in consumption**

|  |  |
| --- | --- |
| YEARS | RATE OF CHANGE |
| 1996-1997 | 0.038274 |
| 1998-1999 | 0.008959 |
| 2000-2001 | 0.051152 |
| 2002-2003 | 0.005784 |
| 2004-2005 | 0.068554 |
| 2006-2007 | 0.020842 |
| 2008-2009 | 0.054469 |
| 2010-2011 | -0.00173 |
| 2012-2014 | 0.049566 |
| AVERAGE | 0.032874 |

**Graphical Representation of Rate of change in consumption-**

**Result of Case 1 Problem -** This clearly shows that rate of change of consumption of renewable energy resource has increased and will increase in future as well and rate of consumption of non-renewable energy has been decreasing and will continue to decrease in future as well. The reason is that countries have started consuming these renewable energy resources which are abundant in nature and causes less harm on environment. On the other hand, non-renewable resources are exhaustible and cause severe damage to environment and this has been proving to be a triggering factor according to world study.

**Case 2 Problem –**

In this, we form different clusters of countries on basis renewable energy consumption showing effect on environment and analysis of non-renewable energy demands followed by checking the quality of clusters.

**Solution –**

Following Algorithms are being used –

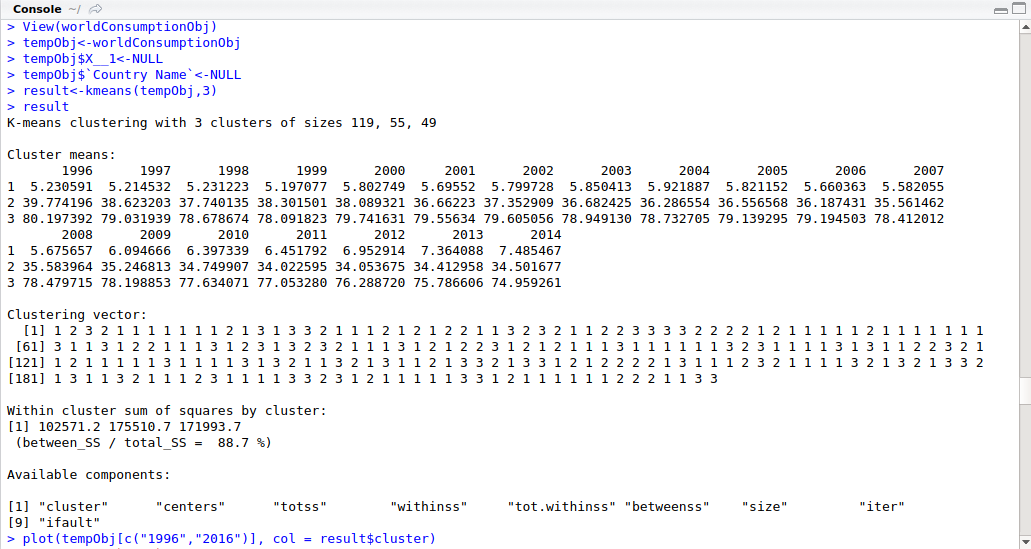
K means - k-means clustering aims to partition of n observations into k clusters, in which, each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

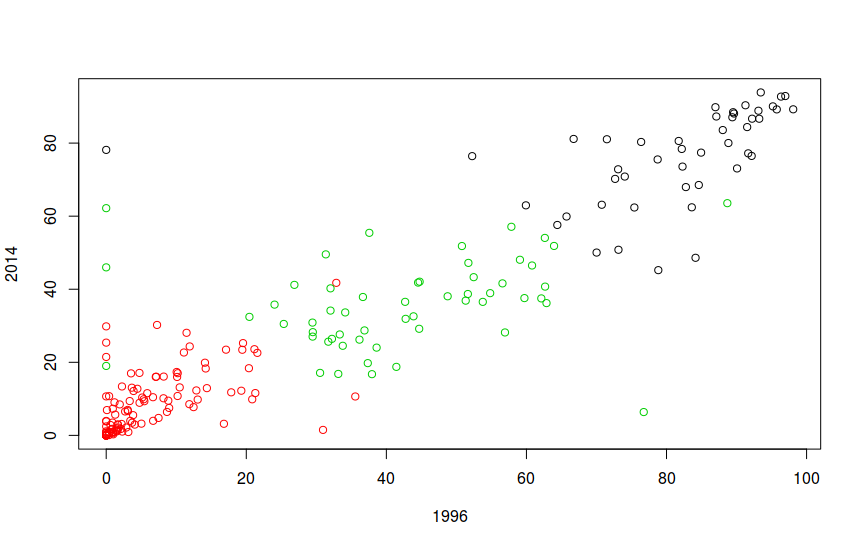
Dendogram – A dendrogram (from Greek dendro "tree" and gramma "drawing") is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering

Silhouette analysis - Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to access parameters like number of clusters visually.

Chi Square Analysis - Chi-Square Test is used for testing independence. The test is applied when you have two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables.

**Step1** – Applied K means Algorithm on Consumption of Renewable Energy Resource Consumption and following clusters were obtained –





We got 3 clusters –

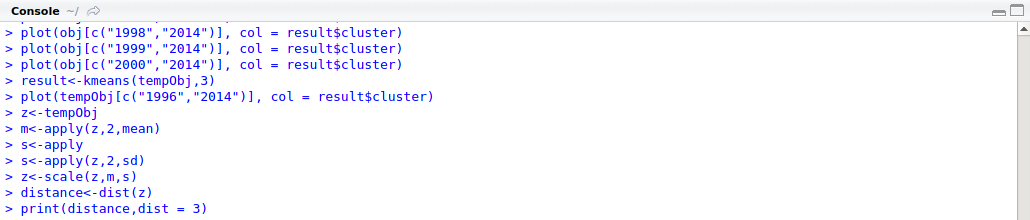
Cluster 1(Red) – This represents those countries whose renewable resources consumption is less than consumption of non-renewable energy resources. So, these countries will cause more harm to environment. So they should move more towards renewable resources production and consumption.

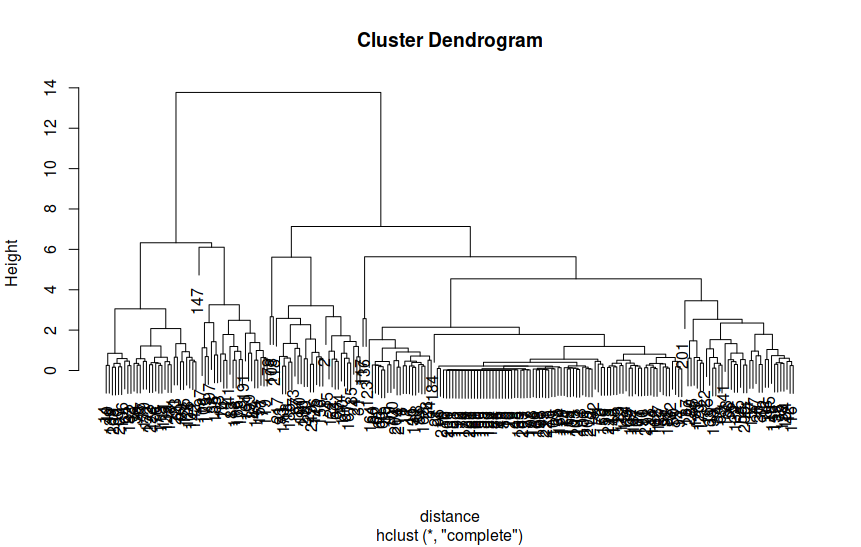
Cluster 2(Green) – This represents those countries whose renewable energy resources consumption as well as non-renewable energy resources consumption is moderate so these countries will not cause much harm to environment when compared to countries belonging cluster 1. So they can consider aligning a little towards renewable resources production and consuming it.

Cluster 1(Black) – This represents those countries whose renewable energy resources consumption is less conversely non-renewable energy resources consumption is high so, these countries will cause more harm to environment. So they should move more towards renewable resources production and consume it.

**Step 2** - Dendogram :

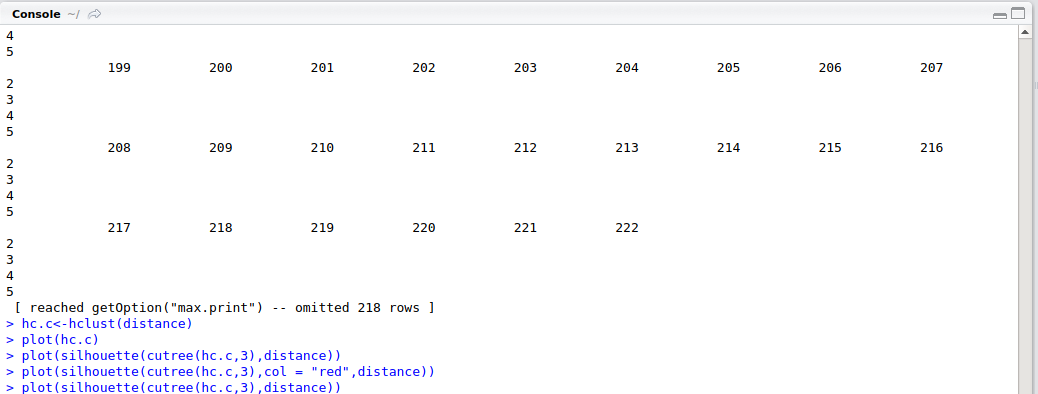
Below is the hierarchical structure of the Data Set via Dendogram –

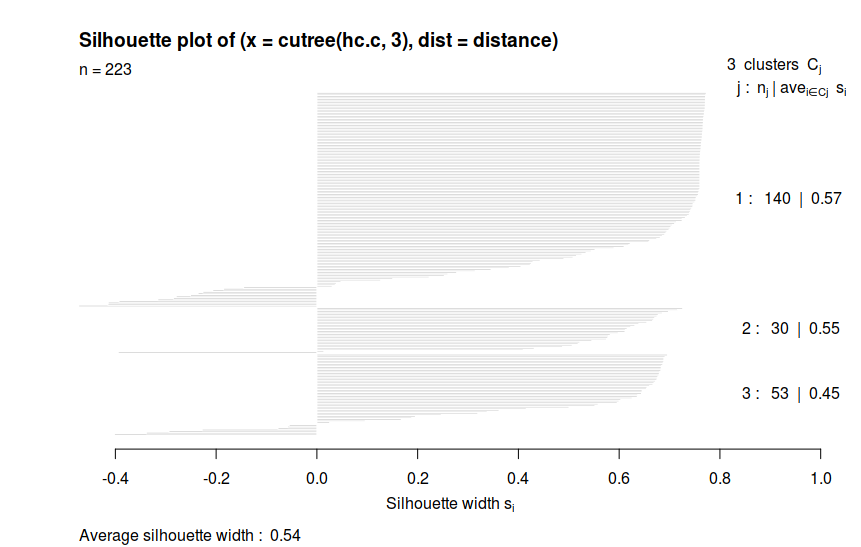




Silhouette Analysis

**Step 3**: Now we need to check the quality of the clusters and see how they are different from each other. This will done via Silhouette Analysis.

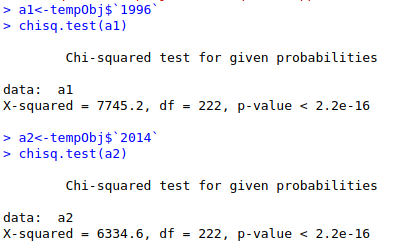




As we can see the average Silhouette Width is 0.54. This clearly shows that the quality of the clusters formed is really good.

**Step 4** - Chi Square Analysis:

This will test the correlation between the entities and decide about null hypothesis.



As we see that the p-value is 2.2e – 16 which is very less than 0.05. So this clearly shows that it rejects the null hypothesis.

Finally, it concludes that the quality of the clusters formed is really good.

**Conclusion**

The non renewable energy resource consumption has been decreasing significantly and renewable energy resources consumption is increasing. So there would be a time when renewable energy resource production and consumption will be very low. As we know, OPEC countries have their economy strongly dependent on exporrt of petroleum and other non-renewable energy resources will be adversaly effected and on the other hand industries like Planet Biogastechnik, Nordex, Okobit and which uses renewable energy will have a huge increase in profit in Market and countries like Germany, USA, Denmark, United Kingdom, Sweden, etc will have large economical growth.

**References:**

**DataSets:**

<https://data.worldbank.org/indicator>

<https://www.eia.gov/beta/international/data/browser/#/?pa=000000000000000000000000000000000000000000000020000007u&f=M&c=00000000000000000000000000000000001a1&ct=2&tl_id=5-M&vs=INTL.78-3-OPEC-TBPD.M&vo=0&v=H&start=198201&end=201605&vid=7>

<https://www.energy-xprt.com/companies/keyword-ethanol-production-5821>

<https://www.google.com/search?ei=f04kWqygMoqYjwSDx7ewDQ&q=ethanol+in+transportation+producing+comapnies&oq=ethanol+in+transportation+producing+comapnies&gs_l=psy-ab.3..33i21k1.8221.44621.0.44763.58.44.1.1.1.0.146.4534.14j28.43.0....0...1c.1.64.psy-ab..18.37.4021.6..0j0i22i30k1j35i39k1j0i131k1j0i67k1j0i131i46k1j46i131k1j0i20i264k1j0i131i20i264k1j0i10k1j0i20i263k1j0i7i30k1j0i8i7i30k1j33i22i29i30k1j33i160k1.180.xgNWnHfqOR4>

<https://www.investopedia.com/markets/stocks/bep/>

<http://seananderson.ca/2013/10/19/reshape.html>

<https://www.investopedia.com/articles/markets-economy/091716/top-5-hydroelectric-power-stocks-2016-bep-ida.asp?lgl=af-top-textlink-baseline>

**Error:**

<https://stackoverflow.com/questions/12320198/r-package-reshape-function-melt-error-id-variables-not-found-in-data-when-worki>

<https://stackoverflow.com/questions/29830189/how-to-plot-multiple-columns-in-r-for-the-same-x-axis-value>

<https://stackoverflow.com/questions/33745567/how-to-generate-a-sequence-based-on-two-columns-in-r>

<https://stackoverflow.com/questions/29830189/how-to-plot-multiple-columns-in-r-for-the-same-x-axis-value>

<https://stackoverflow.com/questions/33049273/plot-multiple-plots-corresponding-to-multiple-columns-specified-by-index-on-1>

**BLOG:**

<https://cleantechnica.com/2016/02/04/how-11-countries-are-leading-the-shift-to-renewable-energy/>

<http://libguides.gatech.edu/c.php?g=54014&p=349002>

<https://asb.opec.org/index.php/interactive-charts/oil-trade>

<https://www.worldenergy.org/data/resources/resource/hydropower/>

<https://www.eia.gov/beta/international/data/browser/#/?pa=0000000000000000000000000000000000vg&c=ruvvvvvfvtvvvv1vvvvvvfvvvvvvfvvvsu20evvvvvvvvvvvfuvg&ct=0&tl_id=5-A&vs=INTL.53-1-AFG-TBPD.A&cy=2015&vo=0&v=H&end=2016&vid=1>

<https://www.eia.gov/beta/international/data/browser/#/?pa=000000000000000000000000000000000000000000000020000007u&f=M&c=00000000000000000000000000000000001a1&ct=2&tl_id=5-M&vs=INTL.78-3-OPEC-TBPD.M&vo=0&v=H&start=198201&end=201605&vid=7>