**Data Management**

**Section – 001**

***Homework 5***

Palash Siddamsettiwar

M10676557

# Dataset Description:

Sustainable Energy for All (SE4ALL) is a global initiative to promote sustainable energies, improve access to energy for all and improve efficiencies of energy production. This dataset on SE4All provided by World Bank, enlists various metrics related to energy consumption, share of renewable energy, access to electricity and non-solid fuel, and energy intensity rate of improvement. We can find time series data of these various metrics for all the countries in the world.

# Overview of the variables and values in the dataset:

**Main**

**Countries:** Since a lot of countries were present in the data (251)- many with incomplete data- it was thought best to keep only the top 20 major economies in the world in the consideration:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| USA | China | Japan | Germany | UK | France | India | Italy | Brazil | Canada |
| South Korea | Russia | Australia | Spain | Mexico | Indonesia | Netherlands | Turkey | Switzerland | Saudi Arabia |

**Years:** As was present in the data: we looked at the 21-year time period between 1990 and 2010.

**Yearly Metrics:** While 54 metrics were present in thedata, to keep the data relevant and to avoid missing values, only 20 metrics were finally considered:

|  |  |  |
| --- | --- | --- |
| **Metric** | **Metric Description** | **Topic** |
| 1.1\_ACCESS.ELECTRICITY.TOT | Access to electricity (% of total population) | Access to electricity |
| 1.2\_ACCESS.ELECTRICITY.RURAL | Access to electricity (% of rural population) | Access to electricity |
| 1.3\_ACCESS.ELECTRICITY.URBAN | Access to electricity (% of urban population) | Access to electricity |
| 8.1.2\_FINAL.ENERGY.INTENSITY | Energy intensity level of final energy (MJ/$2005 PPP) | Energy efficiency |
| 8.1.1\_FINAL.ENERGY.CONSUMPTION | Total final consumption (TJ) | Energy efficiency |
| 4.1\_SHARE.RE.IN.ELECTRICITY | Renewable electricity (% in total electricity output) | Renewable energy production |
| 5.1.1\_TOTAL.CAPACITY | Total installed generation capacity (GW) | Renewable energy production |
| 5.1.2\_RE.CAPACITY | Renewable energy installed capacity (GW) | Renewable energy production |
| 1.1\_TOTAL.FINAL.ENERGY.CONSUM | Total final energy consumption (TFEC) | Renewable energy consumption |
| 2.1.8\_SHARE.WASTE | Waste energy consumption (% in TFEC) | Renewable energy consumption |
| 2.1.5\_SHARE.WIND | Wind energy consumption (% in TFEC) | Renewable energy consumption |
| 2.1.9\_SHARE.BIOGAS | Biogas consumption (% in TFEC) | Renewable energy consumption |
| 2.1.7\_SHARE.GEOTHERMAL | Geothermal energy consumption (% in TFEC) | Renewable energy consumption |
| 2.1.3\_SHARE.HYDRO | Hydro energy consumption (% in TFEC) | Renewable energy consumption |
| 2.1.4\_SHARE.BIOFUELS | Liquid biofuels consumption (% in TFEC) | Renewable energy consumption |
| 2.1.10\_SHARE.MARINE | Marine energy consumption (% in TFEC) | Renewable energy consumption |
| 2.1.2\_SHARE.MODERNBIO | Modern biomass consumption (% in TFEC) | Renewable energy consumption |
| 2.1.1\_SHARE.TRADBIO | Traditional biomass consumption (% in TFEC) | Renewable energy consumption |
| 2.1\_SHARE.TOTAL.RE.IN.TFEC | Renewable energy consumption (% in TFEC) | Renewable energy consumption |
| 2.1.6\_SHARE.SOLAR | Solar energy consumption (% in TFEC) | Renewable energy consumption |

We also have a country dataset that enlists each country’s information such as income level for countries and the regions they fall in.

# Database normalization:

To normalize the data, we need to carry out a few tasks:

1. Shift the years to a column rather than having one column for each year (wide to long)
2. Shift the metrics from one column to multiple columns as they represent different variables
3. Check if resultant data has 1 row for each ‘country, year’ pair and 20 columns for the 20 metrics
4. Separate out metric descriptions into a separate table
5. Divide the data into different tables based on logical differentiation of metrics into the following categories-

|  |
| --- |
| **Topic** |
| Access to electricity |
| Energy efficiency |
| Renewable energy production |
| Renewable energy consumption |

In the end, we have 5 tables:

1. **country\_details**: columns of Country\_name, CountryCode, Income Group and Region
2. **energy\_access**: Country\_Name, Year, [ACCESS.ELECTRICITY.TOT], [ACCESS.ELECTRICITY.RURAL], [ACCESS.ELECTRICITY.URBAN]
3. **energy\_efficiency**: Country\_Name, Year, [FINAL.ENERGY.INTENSITY], [FINAL.ENERGY.CONSUMPTION]
4. **renewable\_consumption**: Country\_Name, Year, [TOTAL.FINAL.ENERGY.CONSUM], [SHARE.WASTE], [SHARE.WIND], [SHARE.BIOGAS], [SHARE.GEOTHERMAL], [SHARE.HYDRO], [SHARE.BIOFUELS], [SHARE.MARINE], [SHARE.MODERNBIO], [SHARE.TRADBIO], [SHARE.SOLAR], [SHARE.TOTAL.RE.IN.TFEC]
5. **renewable\_production**: Country\_Name, Year, [SHARE.RE.IN.ELECTRICITY], [TOTAL.CAPACITY], [RE.CAPACITY]

# Problems in data:

1. We can observe that the data contained a lot of smaller countries which contained a lot of NULL values.
2. Some metrics, such as those in energy\_access, are only present for certain years (1990, 2000, 2010 for )
3. A lot of columns originally are only small changes on other columns- such as conversion of raw values into percentages and subtractions of some columns from others
4. A couple columns are also all NULL values
5. All column names for years were badly named and changed to reflect the true values
6. All column names for metrics after pivoting) had section numbers at the start which had to be removed

# General statistics:

**Total number of records in each table:**

**TableName**   **RowCounts**

--------------------------------------------------------------------------------------------

country\_details 20

energy\_access 60

energy\_efficiency 420

renewable\_consumption 420

renewable\_production 420

**Total number of countries in consideration by Region:**

**Region Number of Countries**

-------------------------------------------------- -------------------

Europe 7

Eastern Asia 3

Western Asia 2

Latin America and Caribbean 2

Northern America 2

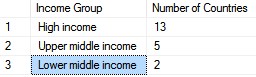
Oceania 1

South Eastern Asia 1

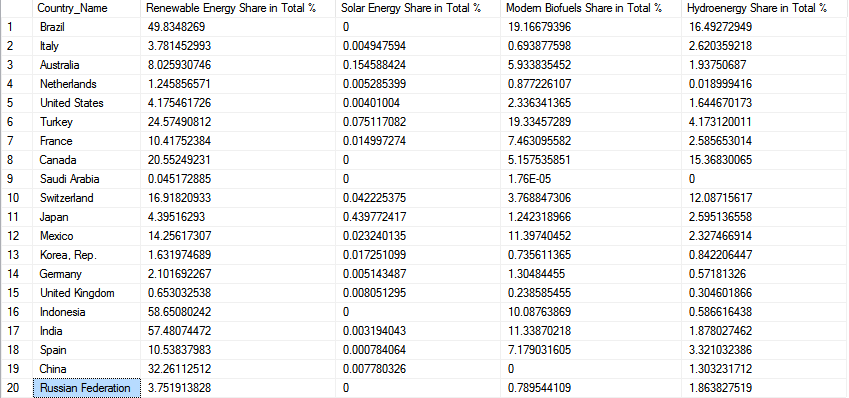
Southern Asia 1

Eastern Europe 1

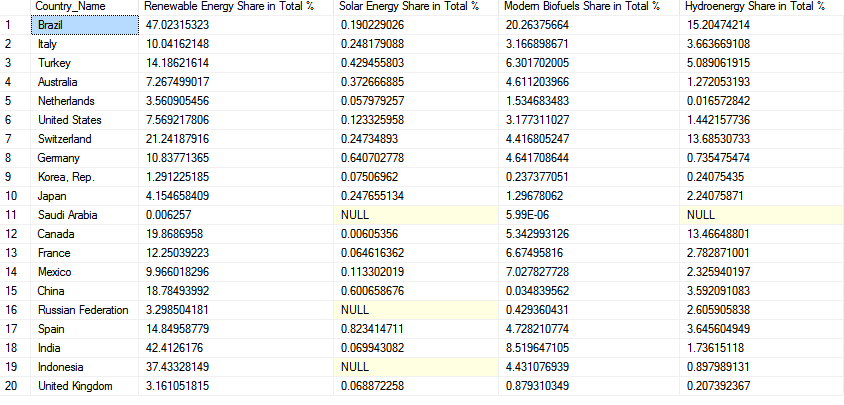
**Distribution of countries by income level:**



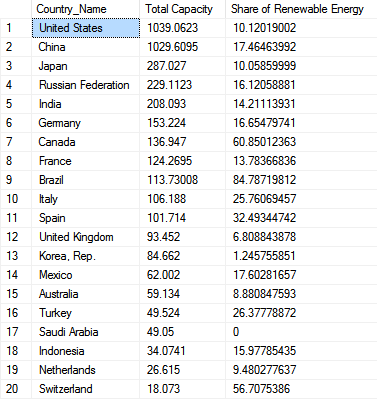
**Average share of renewable energy consumption by type of renewable energy in 1990:**



**Average share of renewable energy consumption by type of renewable energy in 2010:**



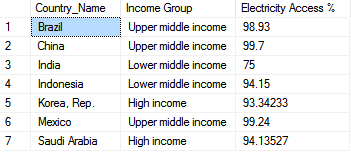
**Order countries by their total electricity capacity in 2010:**



**Maximum final energy intensity for each country:**



**Countries in 2010 with access to electricity less than 100% overall and their access % values:**



# R Analysis:

In R, I performed analysis about correlations between a lot of important variables and tried to find a few insights. A subsetted view of these correlations are provided below:

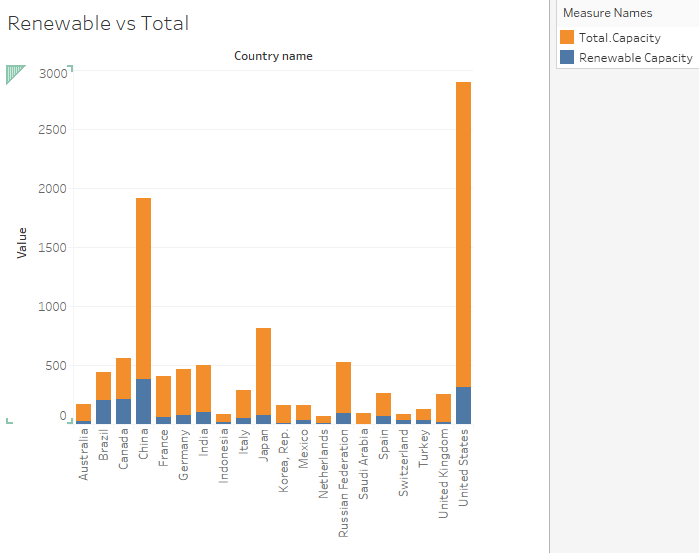
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **SHARE.WASTE** | **SHARE.BIOGAS** | **SHARE.MODERNBIO** | **SHARE.TOTAL.RE.IN.TFEC** | **ACCESS.ELECTRICITY.TOT** |
| **SHARE.WASTE** | 1.00 | 0.50 | -0.14 | -0.11 | 0.20 |
| **SHARE.BIOGAS** | 0.50 | 1.00 | -0.24 | -0.18 | 0.23 |
| **SHARE.MODERNBIO** | -0.14 | -0.24 | 1.00 | 0.70 | -0.26 |
| **SHARE.TOTAL.RE.IN.TFEC** | -0.11 | -0.18 | 0.70 | 1.00 | -0.64 |
| **ACCESS.ELECTRICITY.TOT** | 0.20 | 0.23 | -0.26 | -0.64 | 1.00 |
| **FINAL.ENERGY.INTENSITY** | -0.28 | -0.23 | -0.17 | 0.27 | -0.26 |
| **FINAL.ENERGY.CONSUMPTION** | -0.16 | 0.19 | -0.22 | -0.02 | 0.06 |
| **SHARE.RE.IN.ELECTRICITY** | 0.15 | -0.09 | 0.65 | 0.56 | 0.06 |
| **TOTAL.CAPACITY** | -0.11 | 0.27 | -0.20 | -0.08 | 0.13 |
| **RE.CAPACITY** | -0.13 | 0.31 | 0.04 | 0.19 | 0.11 |

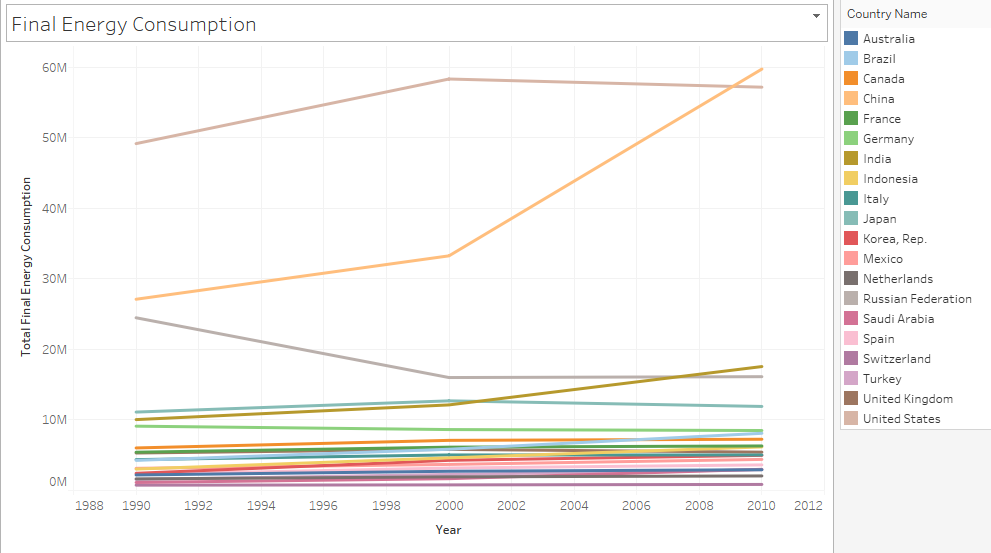
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **FINAL.ENERGY.**  **INTENSITY** | **FINAL.ENERGY.**  **CONSUMPTION** | **SHARE.RE.IN.**  **ELECTRICITY** | **TOTAL.**  **CAPACITY** | **RE.**  **CAPACITY** |
| **SHARE.WASTE** | -0.28 | -0.16 | 0.15 | -0.11 | -0.13 |
| **SHARE.BIOGAS** | -0.23 | 0.19 | -0.09 | 0.27 | 0.31 |
| **SHARE.MODERNBIO** | -0.17 | -0.22 | 0.65 | -0.20 | 0.04 |
| **SHARE.TOTAL.RE.IN.TFEC** | 0.27 | -0.02 | 0.56 | -0.08 | 0.19 |
| **ACCESS.ELECTRICITY.TOT** | -0.26 | 0.06 | 0.06 | 0.13 | 0.11 |
| **FINAL.ENERGY.INTENSITY** | 1.00 | 0.33 | -0.04 | 0.10 | 0.15 |
| **FINAL.ENERGY.CONSUMPTION** | 0.33 | 1.00 | -0.15 | 0.95 | 0.78 |
| **SHARE.RE.IN.ELECTRICITY** | -0.04 | -0.15 | 1.00 | -0.13 | 0.26 |
| **TOTAL.CAPACITY** | 0.10 | 0.95 | -0.13 | 1.00 | 0.84 |
| **RE.CAPACITY** | 0.15 | 0.78 | 0.26 | 0.84 | 1.00 |

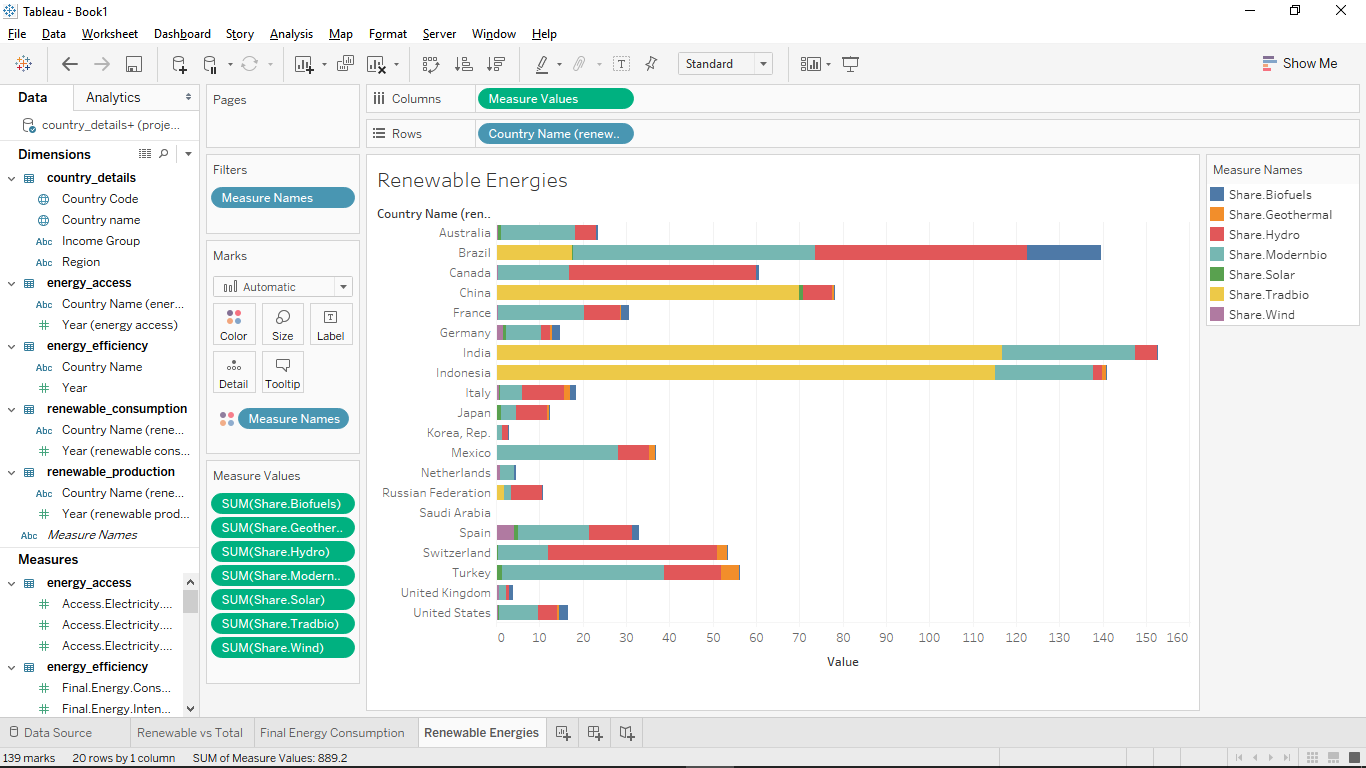
We can see that:

* Access to electricity is weakly and inversely correlated to the total share of renewable energy
* Modern biofuels’ share is positively correlated to total share of renewable energy in consumption
* Renewable energy capacity is highly correlated to final energy consumption

# Tableau Visualizations:







**ER Diagram**



# Summary:

In this analysis, we looked at the top 20 economies in the world and looked at their existing energy capacities and their existing renewable energy methods.

From this analysis, we could see that renewable energy as a practical means of energy sustenance for any country depends on their development, income levels and their size. We could make various insights about each country’s contribution to making this place safer to live by understanding how much energy is being consumed by then and which are the best types of renewable energies we can find in each country.

We could see that:

* Most of the top economies in the world are European High income countries
* We could see that developing countries have higher share in renewable energies but as they develop more and more, they shift more towards non-renewable sources of energy- as evident from India, China and Indonesia’s data
* US and China have the highest energy needs but only 10 and 17% of their needs are met by renewable energies
* Canada and Brazil have excellent renewable energy setups
* India is one of the only top economies with less than 90% energy access
* Access to electricity is weakly and inversely correlated to the total share of renewable energy
* Modern biofuels’ share is positively correlated to total share of renewable energy in consumption
* Renewable energy capacity is highly correlated to final energy consumption
* China’s growth of energy consumption has been astronomical while USA and Russian Federation have plateaued or even reduced in their energy needs
* US, China, Canada and Brazil have the highest renewable energy productions in that order
* A large part (India and Indonesia) of developing economies’ renewable energies come from traditional biofuels
* Saudi Arabia has next to none renewable energies production

# Challenges Faced:

* Data Cleaning and understanding the data took a major chunk of the time- this included understanding what each metric means and to know which ones to keep and which ones to remove. The other challenge was to clean the data and to remove the columns with high number of NULL values. Data normalization was also a huge challenge since it involved understanding how best to keep the data

**References:**

1. Data source: <http://data.worldbank.org/data-catalog/sustainable-energy-for-all>
2. Top Economies in the world: IMF data: <http://goo.gl/dIIzqK>

**Tools Used:**

1. Microsoft SQL Server
2. Tableau 10
3. R and RStudio

APPENDIX A (SQL Codes)

use project;

/\* main.csv and country.csv imported from import wizard\*/

drop table dbo.country\_details;

select distinct

Country\_name,

CountryCode,

case

when [Income Group] = 'High income: nonOECD' then 'High income'

when [Income Group] = 'High income: OECD' then 'High income'

when [Income Group] = 'Upper middle income' then 'Upper middle income'

when [Income Group] = 'Lower middle income' then 'Lower middle income'

end as [Income Group],

Region

into dbo.country\_details

from

dbo.country a

inner join

dbo.main b

on b.Country\_Code= a.CountryCode

where Country\_Name in ('United States', 'China', 'Japan', 'Germany', 'United Kingdom', 'France', 'India', 'Italy', 'Brazil', 'Canada', 'Korea, Rep.', 'Russian Federation', 'Australia', 'Spain', 'Mexico', 'Indonesia', 'Netherlands', 'Turkey', 'Switzerland', 'Saudi Arabia')

;

drop table dbo.energy\_long;

select a.Country\_name, b.Indicator\_Code, Year, Value

into dbo.energy\_long

from

dbo.country\_details a

left join

dbo.main b

on a.Country\_Name = b.Country\_Name

CROSS APPLY

(

VALUES

(1990, F5),

(1991, F6),

(1992, F7),

(1993, F8),

(1994, F9),

(1995, F10),

(1996, F11),

(1997, F12),

(1998, F13),

(1999, F14),

(2000, F15),

(2001, F16),

(2002, F17),

(2003, F18),

(2004, F19),

(2005, F20),

(2006, F21),

(2007, F22),

(2008, F23),

(2009, F24),

(2010, F25)

) x (Year, Value)

;

drop table dbo.energy\_long\_subset;

select Country\_name, Indicator\_Code, Year, Value

into dbo.energy\_long\_subset

from

dbo.energy\_long a

where Indicator\_Code in (

'1.1\_ACCESS.ELECTRICITY.TOT',

'1.2\_ACCESS.ELECTRICITY.RURAL',

'1.3\_ACCESS.ELECTRICITY.URBAN',

'8.1.2\_FINAL.ENERGY.INTENSITY',

'8.1.1\_FINAL.ENERGY.CONSUMPTION',

'4.1\_SHARE.RE.IN.ELECTRICITY',

'5.1.1\_TOTAL.CAPACITY',

'5.1.2\_RE.CAPACITY',

'1.1\_TOTAL.FINAL.ENERGY.CONSUM',

'2.1.8\_SHARE.WASTE',

'2.1.5\_SHARE.WIND',

'2.1.9\_SHARE.BIOGAS',

'2.1.7\_SHARE.GEOTHERMAL',

'2.1.3\_SHARE.HYDRO',

'2.1.4\_SHARE.BIOFUELS',

'2.1.10\_SHARE.MARINE',

'2.1.2\_SHARE.MODERNBIO',

'2.1.1\_SHARE.TRADBIO',

'2.1\_SHARE.TOTAL.RE.IN.TFEC',

'2.1.6\_SHARE.SOLAR'

)

;

drop table dbo.energy\_wide;

select

a.Country\_Name,

a.Year,

max(CASE WHEN Indicator\_Code='1.1\_ACCESS.ELECTRICITY.TOT' THEN Value END) AS 'ACCESS.ELECTRICITY.TOT',

max(CASE WHEN Indicator\_Code='1.2\_ACCESS.ELECTRICITY.RURAL' THEN Value END) AS 'ACCESS.ELECTRICITY.RURAL',

max(CASE WHEN Indicator\_Code='1.3\_ACCESS.ELECTRICITY.URBAN' THEN Value END) AS 'ACCESS.ELECTRICITY.URBAN',

max(CASE WHEN Indicator\_Code='8.1.2\_FINAL.ENERGY.INTENSITY' THEN Value END) AS 'FINAL.ENERGY.INTENSITY',

max(CASE WHEN Indicator\_Code='8.1.1\_FINAL.ENERGY.CONSUMPTION' THEN Value END) AS 'FINAL.ENERGY.CONSUMPTION',

max(CASE WHEN Indicator\_Code='4.1\_SHARE.RE.IN.ELECTRICITY' THEN Value END) AS 'SHARE.RE.IN.ELECTRICITY',

max(CASE WHEN Indicator\_Code='5.1.1\_TOTAL.CAPACITY' THEN Value END) AS 'TOTAL.CAPACITY',

max(CASE WHEN Indicator\_Code='5.1.2\_RE.CAPACITY' THEN Value END) AS 'RE.CAPACITY',

max(CASE WHEN Indicator\_Code='1.1\_TOTAL.FINAL.ENERGY.CONSUM' THEN Value END) AS 'TOTAL.FINAL.ENERGY.CONSUM',

max(CASE WHEN Indicator\_Code='2.1.8\_SHARE.WASTE' THEN Value END) AS 'SHARE.WASTE',

max(CASE WHEN Indicator\_Code='2.1.5\_SHARE.WIND' THEN Value END) AS 'SHARE.WIND',

max(CASE WHEN Indicator\_Code='2.1.9\_SHARE.BIOGAS' THEN Value END) AS 'SHARE.BIOGAS',

max(CASE WHEN Indicator\_Code='2.1.7\_SHARE.GEOTHERMAL' THEN Value END) AS 'SHARE.GEOTHERMAL',

max(CASE WHEN Indicator\_Code='2.1.3\_SHARE.HYDRO' THEN Value END) AS 'SHARE.HYDRO',

max(CASE WHEN Indicator\_Code='2.1.4\_SHARE.BIOFUELS' THEN Value END) AS 'SHARE.BIOFUELS',

max(CASE WHEN Indicator\_Code='2.1.10\_SHARE.MARINE' THEN Value END) AS 'SHARE.MARINE',

max(CASE WHEN Indicator\_Code='2.1.2\_SHARE.MODERNBIO' THEN Value END) AS 'SHARE.MODERNBIO',

max(CASE WHEN Indicator\_Code='2.1.1\_SHARE.TRADBIO' THEN Value END) AS 'SHARE.TRADBIO',

max(CASE WHEN Indicator\_Code='2.1\_SHARE.TOTAL.RE.IN.TFEC' THEN Value END) AS 'SHARE.TOTAL.RE.IN.TFEC',

max(CASE WHEN Indicator\_Code='2.1.6\_SHARE.SOLAR' THEN Value END) AS 'SHARE.SOLAR'

into dbo.energy\_wide

from dbo.energy\_long\_subset a

group by Country\_Name, Year

;

select

Country\_Name,

Year,

[ACCESS.ELECTRICITY.TOT],

[ACCESS.ELECTRICITY.RURAL],

[ACCESS.ELECTRICITY.URBAN]

into

dbo.energy\_access

from

dbo.energy\_wide

where YEAR in (1990, 2000, 2010);

select

Country\_Name,

Year,

[FINAL.ENERGY.INTENSITY],

[FINAL.ENERGY.CONSUMPTION]

into

dbo.energy\_efficiency

from

dbo.energy\_wide

;

select

Country\_Name,

Year,

[SHARE.RE.IN.ELECTRICITY],

[TOTAL.CAPACITY],

[RE.CAPACITY]

into

dbo.renewable\_production

from

dbo.energy\_wide

;

select

Country\_Name,

Year,

[TOTAL.FINAL.ENERGY.CONSUM],

[SHARE.WASTE],

[SHARE.WIND],

[SHARE.BIOGAS],

[SHARE.GEOTHERMAL],

[SHARE.HYDRO],

[SHARE.BIOFUELS],

[SHARE.MARINE],

[SHARE.MODERNBIO],

[SHARE.TRADBIO],

[SHARE.TOTAL.RE.IN.TFEC],

[SHARE.SOLAR]

into

dbo.renewable\_consumption

from

dbo.energy\_wide

;

drop table dbo.country;

drop table dbo.energy\_long;

drop table dbo.energy\_wide;

drop table dbo.energy\_long\_subset;

drop table dbo.main;

/\* Total number of records in each table: \*/

SELECT

TableName = t.NAME,

RowCounts = p.rows

FROM

sys.tables t

INNER JOIN

sys.indexes i ON t.OBJECT\_ID = i.object\_id

INNER JOIN

sys.partitions p ON i.object\_id = p.OBJECT\_ID AND i.index\_id = p.index\_id

WHERE

t.is\_ms\_shipped = 0

GROUP BY

t.NAME, p.Rows

ORDER BY

t.Name

;

/\* Total number of countries in consideration by Region: \*/

select

Region,

count(Country\_name) as 'Number of Countries'

from

dbo.country\_details a

group by Region

order by count(Country\_name) desc;

;

/\* Distribution of countries by income level: \*/

select

[Income Group],

count(Country\_name) as 'Number of Countries'

from

dbo.country\_details a

group by [Income Group]

order by count(Country\_name) desc;

;

/\* Average share of renewable energy consumption by type of renewable energy in 1990: \*/

select Country\_Name,

[SHARE.TOTAL.RE.IN.TFEC] as 'Renewable Energy Share in Total %',

[SHARE.SOLAR] as 'Solar Energy Share in Total %',

[SHARE.MODERNBIO] as 'Modern Biofuels Share in Total %',

[SHARE.HYDRO] as 'Hydroenergy Share in Total %'

from

dbo.renewable\_consumption

where

Year=1990;

/\* Average share of renewable energy consumption by type of renewable energy in 2010: \*/

select Country\_Name,

[SHARE.TOTAL.RE.IN.TFEC] as 'Renewable Energy Share in Total %',

[SHARE.SOLAR] as 'Solar Energy Share in Total %',

[SHARE.MODERNBIO] as 'Modern Biofuels Share in Total %',

[SHARE.HYDRO] as 'Hydroenergy Share in Total %'

from

dbo.renewable\_consumption

where

Year=2010;

/\* Order countries by their total electricity capacity in 2010: \*/

select

Country\_Name,

[TOTAL.CAPACITY] as 'Total Capacity',

[SHARE.RE.IN.ELECTRICITY] as 'Share of Renewable Energy'

from

dbo.renewable\_production

where

Year='2010'

order by [TOTAL.CAPACITY] desc

;

/\* Maximum final energy intensity for each country: \*/

select

Country\_Name,

max([FINAL.ENERGY.INTENSITY]) as 'Maximum Energy Intensity'

from

dbo.energy\_efficiency

group by Country\_Name

order by max([FINAL.ENERGY.INTENSITY]) desc;

/\* Countries in 2010 with access to electricity less than 100% overall and their access % values: \*/

select

a.Country\_Name,

[Income Group],

[ACCESS.ELECTRICITY.TOT] as 'Electricity Access %'

from

dbo.energy\_access a

inner join

dbo.country\_details b

on a.Country\_Name = b.Country\_Name

where Year='2010' and [ACCESS.ELECTRICITY.TOT]<>100;

APPENDIX B (R Codes)

library(RODBC)

library(ggplot2)

library(xtable)

Local <- odbcConnect("Test", uid = "", pwd = "")

country\_details <- sqlQuery(Local, "

select \* from project.dbo.country\_details;")

energy\_access <- sqlQuery(Local, "

select \* from project.dbo.energy\_access;")

energy\_efficiency <- sqlQuery(Local, "

select \* from project.dbo.energy\_efficiency;")

renewable\_consumption <- sqlQuery(Local, "

select \* from project.dbo.renewable\_consumption;")

renewable\_production <- sqlQuery(Local, "

select \* from project.dbo.renewable\_production;")

correlation\_analysis <- sqlQuery(Local, "

select

\*

from

project.dbo.renewable\_consumption a

inner join

project.dbo.energy\_access b

on

a.Country\_Name = b.Country\_Name

and

a.Year=b.Year

inner join

project.dbo.energy\_efficiency c

on

a.Country\_Name = c.Country\_Name

and

a.Year=c.Year

inner join

project.dbo.renewable\_production e

on

a.Country\_Name = e.Country\_Name

and

a.Year=e.Year;"

)

correlation\_analysis <- correlation\_analysis[-c(15,16,20,21,24,25)]

write.csv(as.data.frame(cor(correlation\_analysis[-c(1,2,3,5, 7, 8, 9, 10, 12, 14)])), "corr.csv")