Data Warehousing and Business Intelligence Project

on

Energy Consumption

Sumit Jadhav

X18129633

MSc Data Analytics – 2019/20

Submitted to: Sean Henney

National College of Ireland
Project Submission Sheet – 2019/20
School of Computing

Student Name:	Sumit Jadhav
Student ID:	X18129633
Programme:	MSc Data Analytics
Year:	2019-20
Module:	Data Warehousing and Business Intelligence
Lecturer:	Sean Henney
Submission Due Date:	12/04/2019
Project Title:	Energy Consumption

I hereby certify that the information contained in this (my submission) is information pertaining to my own individual work that I conducted for this project. All information other than my own contribution is fully and appropriately referenced and listed in the relevant bibliography section. I assert that I have not referred to any work(s) other than those listed. I also include my TurnItIn report with this submission.

ALL materials used must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is an act of plagiarism and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	12/04/2019

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Table 1: Mark sheet – do not edit

Criteria	Mark Awarded	Comment(s)
Objectives	of 5	
Related Work	of 10	
Data	of 25	
ETL	of 20	
Application	of 30	
Video	of 10	
Presentation	of 10	
Total	of 100	

Project Check List

This section capture the core requirements that the project entails represented as a check list for convenience.

☒ Used LaTeX template

□ Three Business Requirements listed in introduction

□ At least one structured data source

□ At least one unstructured data source

□ At least three sources of data

□ Described all sources of data

□ All sources of data are less than one year old, i.e. released after 17/09/2017

□ Inserted and discussed star schema

□ Completed logical data map

□ Discussed the high level ETL strategy

□ Provided 3 BI queries

□ Detailed the sources of data used in each query

□ Discussed the implications of results in each query

□ Reviewed at least 5-10 appropriate papers on topic of your DWBI project

Energy Consumption

Sumit Jadhav
1234567

12/04/2019

Abstract

Motivation behind this project was to study the global level of energy consumption in terms of the countries and how it is affected to the environment as well as to the economic condition of countries. The emissions produced by the burning of in creating energy has a deep impact on our society which lead to numerous health problems and issues, this project helps us to determine what regulations and laws are applied by the different government which helps to reduce the amount of these gases such as nitrous oxide, carbon monoxide, Sulphur dioxide. Energy in terms as fuel whether it is renewable sources or non-renewable the value of these fuel keeps on changing in terms of demand and supply and also based on the market price and the government regulation rules. To explain all these relations this project gives us the idea in terms of the relation between different entities like production, consumption and the market price of the fuel in various countries.

1 Introduction

In this project the various constraints were considered in terms of business which helps us to take a considerate decision based on the real facts of the various data. Mainly the focus of the project is to determine the underlying cause of the increasing price within the limited countries and what changes were observed throughout the years of usage and production. n 6.

(Req-1) What is the impact of energy production on the environment?

(Req-2) What are the top 10 countries which has the direct dependencies on energy production and being a developed nation?

(Req-3) What is the impact on natural gas consumption in India in domestic use?

2 Data Sources

For this project I have used 5 different types of data sources which includes information regarding the worldwide usage of energy consumption and production and the emission caused by it. We gain an insight in the interrelation between these entities which helps us to understand the difference in gross domestic product of countries based on the energy intensity.

Source	Type	Brief Summary
Enerdata	Structured	This data-set provides the information regarding the energy consumption of electricity, gasoline and natural gas as well as the total energy intensity of GDP for the different countries
OECD.Stat	Structured	This Data-set provides the information related the worldwide emission of harmful gases.
Statista	Structured	From this source i have taken two data-sets which provides the information regarding the production of fuels like Gasoline, Natural gas and as well as the production of electricity.
GlobalPetrolPrices	Unstructured	From this source i have scrapped the data regarding the fuel prices in various countries such as Gasoline, Natural gas and electricity.

Table 2: Summary of sources of data used in the project

2.1 Source 1: Enerdata

Enerdata gives a very detailed information regarding the total consumption of energy components such as fuel like gasoline, natural gas, coal, crude oil and electricity. This source also proved the information regarding the energy intensity of GDP across various countries. It provides the year range of 1990 to 2017 and country wise distribution of the data which embassies on the consumption and production with measuring metrics like billion cubic meters for natural gas, Million tons for oil production, terawatt hour for electricity and on based on dollars at constant exchange rate, price and purchasing power parities of the year 2015 per kilo of oil equivalent.

However, relevant to this project this data was cleaned with R and the relevant factor were considered such as production of gasoline, electricity, natural gas and the energy intensity of GDP on various countries from which it was filtered down to 52 countries.

Link : https://yearbook.enerdata.net

Source 2: OECDStat OECDStat gives us the information regarding the emission occurring world wide for harmful gases like sulphur oxide, nitrogen oxide, carbon monoxide, particulates (P10), particulates (P2.5), non methane volatile organic compounds. It provides us the information year range of 1990 to 2016 and various countries. All these emission are classified under man made emission for this data-set.

As in relevance of this project there is a direct relationship between the consumption, production and the emission of gasses and the impact of emission to the environment.

Link :https://stats.oecd.org/Index.aspx?DataSetCode=AIR$_EMISSIONS$

2.2 Source 3: Statista

From this source 2 data-sets were used to study the production rate for non renewable fuels like natural gas, petroleum/gasoline, coal and uranium, and electricity from year 1990 to 2016.

For this project this data is important to study the relation between demand and supply chain and examine countries dependencies on the production of energy. Energy production can show us the possible affection on the countries growth whether it is a developed nation or a developing nation.

Link 1(Electricity Production): https://www.statista.com/statistics/270281/electricity-generation-worldwide/

Link 2(Non Renewable Energy Production): https://www.statista.com/statistics/263232/lobal-production-of-non-renewable-energy-resources/

2.3 Source 4: GlobalPetrolPrices

This is the unstructured data-set from which the scrapping of data is done. This data-set provides the global fuel prices from over all the countries. The unit conversion for this source was kilo watt hour per US dollar.

However relevant to this project, prices of the fuels like gasoline, natural gas and electricity could be used to determine the relationship between production and prices from the given countries.

3 Data Model

Creating this dataware house bottom up approach was implemented as noted by Ralph Kimball.

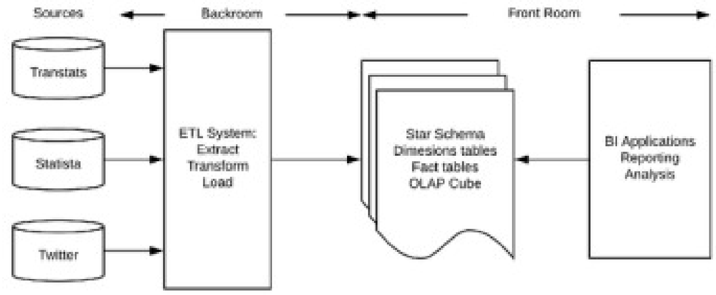

Figure 1: Architecture(Using Kimball's Approach)

Figure 1: Star Schema

In this project star schema model was followed in which there were two dimensions such as Dim$_{Y}ear, Dim_{C}ountryandonefacttableasFactTable.$

3.1 Source 3: Dimensions

In this data model primary key was assigned to the two dimension which are Dim$_{Y}ear$ $andDim_{C}ountrybywhichtheforeignkeywasappliedtothemeasurepresentedinitwhichisYear_{I}DandCou$

Dim$_{Y}ear : ThisdimensionhasYear_{I}DandhasYearwhichconsistsoftheyearspanof2007to20016$
Dim$_{C}outry : ThisdimensionhasCountry_{I}Dandhas53countrieslisted.$

3.2 Source 3: Fact Table

FactTable : This table has 13 facts which as are follows:

CO
= Carbon Monoxide Provides the information related to emission of carbon monoxide in the atmosphere from year 2007 to 2016

SO2
= Sulphur Dioxide Provides the information related to emission of Sulphur Dioxide in the atmosphere from year 2007 to 2016

NO2
= Nitrogen Oxide Provides the information related to emission of Nitrogen Oxide in the atmosphere from year 2007 to 2016
[E Con] = Electricity Consumption Provides the electricity consumption worldwide from year 2007 to 2016
[Electricity Production] = Electricity Production
[Gasoline Price] = Gasoline Price
[Electricity Price] = Electricity Price
[Natural Gas Price] = Natural Gas Price
[NGCons] = Natural Gas Consumption
[Oilcons] = Oil Consumption
[Oilprod = Oil Production
[Gasoline] = Gasoline Production
[Natural Gas] = Natural Gas Production

As per the business requirement stated in section one business relation can be mentioned as follows:

Emission: As there is a decrease in emission even though the increase in production of electricity, it states that the pollution regulating rules are effectively working and new techniques are being invented to produce energy efficiently.

GDP and Energy Production: As the developed nation has more dependencies on oil production such as America, China and Middle East.

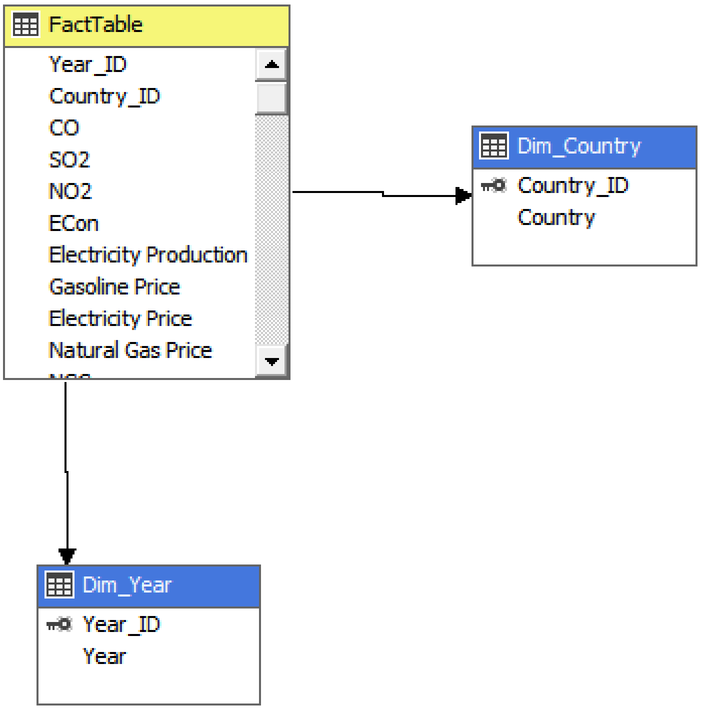

Figure 2: Star Schema

4 Logical Data Map

Table 3: Logical Data Map describing all transformations, sources and destinations for all components of the data model illustrated in Figure 3

Source	Column	Destination	Column	Type	Transformation
1	1990-2017	Dim$_Year$	Year	Dimension	Year taken from 2007 to 2016 by removing columns
1	Countries	Dim$_Country$	Dimension	Out of 60 countries only 53 countries were	
2	CO	FactTable	CO	Fact	Columns removed before and after the year of 2007 and 2016
2	NO2	FactTable	NO2	Fact	Columns removed before and after the year of 2007 and 2016 $
2	SO2	FactTable	So2	Fact	Columns removed before and after the year of 2007 and 2016
3	Electricity generation worldwide from 1990 to 2016 (in terawatt hours)	Dim$_Year$	Year	Dimension	No transformation were done $

Continued on next page

Table 3 – *Continued from previous page*

Source	**Column**	**Destination**	**Column**	**Type**	**Transformation**
3	Electricity generation in terawatt hours	fact	Energy Consumption	Fact	No tranformation Required
3.2	Distribution of selected energy carriers as a share of non-renewable energy production worldwide from 2007 to 2016	Dim$_{Y}ear$	Year	Dimension	No transformation required $
3.2	petroleum	Genre	Dimension	Fact	Gasoline
3.2	Hardcoal	Removed	removed	-	Not Required
3.2	Natural gas	Fact	Natural Gas	Not Required	
3.2	Uranium	Removed	removed	-	Not Required $
3.2	Removed	removed	-	Not Required	
4	Country	Fact Table	Country	Fact	Melt function used in R $
4	Price	Fact Table	Electricity Price	Fact	Melt Function used
4	Country	Fact Table	Country	Fact	Melt function used in R $
4	Price	Fact Table	LPG Price	Fact	Melt Function used
4	Country	Fact Table	Country	Fact	Melt function used in R $
4	Price	Fact Table	Gasoline Price	Fact	Melt Function used

5 ETL Process

Extraction transformation and loading of the dare are the basic principle of data ware house and business intelligence process. Collecting data from diff rent structured and unstructured data and gaining some useful information is the main objective here. In data gathering process the main task is to have a clean data which does not have any duplicates of null values which could impact out future result.

5.1 Extraction

The data related to the Energy consumption and production was fetched by 4 different data sources. Some of the data was structured and the unstructured was only extracted and transformed transformed using R library tidyverse().

5.2 Cleaning

While constructing a data ware house it is important that the data should be consistent, should not have any anomaly. To achieve this data cleaning was done to make sure data is available and clean and does not contain any redundancy.

5.2.1 Cleaning Source 1

This source contained multiple sheets of data so, with the help of R all the sheets were read one by one and the required data was stored in a .CSV file.

5.2.2 Cleaning Source 2

From this source three data files were taken in xlsx format and were merged with the melt function in R. The NA values were omitted using na.omit() function. Inquired columns and rows were also removed in this process.

5.2.3 Cleaning Source 3

In this source the 2 data sets were taken for the required energy production values. Only the column headers were changed to Electricity production.

5.2.4 Cleaning Source 4

This was the unstructured data source in which the columns were converted to character value and the combined with the help of cbind function.

5.3 Transformation

After making sure that data has no noise or inconsistency the data was safely transformed into CSV format. Data integration as well as data aggregation was performed in this stage.

5.4 Load

In this stage we use the final clean data. Here we used Microsoft's SSIS tool to integrate with the Mcrosoft SQL server management studio(SSMS). In this process we have used the flat file as a raw data and loaded it to destination. In the staging area 9 raw files were used. These files will be used to populate our fact and dimension table. The most important part in this step is connectivity with the MOLAP server.

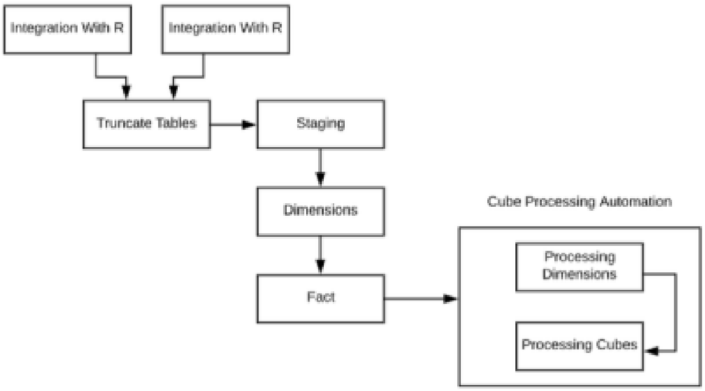

Figure 3: Overview of ETL

Figure 3: Star Schema

In integration with R we have to integrate the R script in the SSIS as it can map the data to the SSMS which will help us to populate the facts and dimensions. The CSV file generated by the R script is loaded to the SSMS by which we can proceed to the further step.

Here for all the data set the R script was wit ten for cleaning purpose and then the raw files were stored locally. After the staging area we are now proceeding to populate the dimensions and fact table.

Now in this stage, first we have to truncate the tables to make sure there are no duplicate entries in the database. As this is the automated process we need to make sure this step.

After the staging we have to populate the dimension. When creating the dimension table we must assign a primary key to it. The challenge faced in this stage was mismatch in the data type. It was further solved by matching and manipulating data types in staging table.

After this process we focus on the fact table. one thing we must make sure of the primary key and the foreign key relationship. By doing this we can easily populate the fact table as the mapping of the data is made sure.

After this step we proceed to deploying the cube with the help of analysis services present in SSIS. With the help of SSIS and SSMS we can deploy the cube and a source view can be generated.

5.5 Degree Of Automation Process

The degreee of automation of ETL process was achieved on just a single click. When we click on the start process tfollowing process were automated.

Data Extraction
Cleaning
Transformation
Loading the data in SSIS
Populating dimensions and facts
Cube Deployment

6 Application

subsectionBI Query 1 : What is the impact of energy production on the environment?

As shown in the image the contradiction in increase in the production of electricity the emission has decreased which directly states that the new techniques of production and law of regulation on the emission of gasses are visible.

subsectionBI Query 2 : What are the top 10 countries which has the direct dependencies on energy production and being a developed nation?

As shown in the visualization the direct co relation of oil production and the increase in GDP could be seen. America, China and middle east tend to have more dependency on oil as they are already developed nation.

subsectionBI Query 3 : What is the impact on natural gas consumption in India in domestic use? As the year span between the 2007 to 2016 it can be observed in following image that as the awareness increased in the natural gas consumption for domestic use the prices were affected as the more efficient options were made available a slight decreased can be observed.

6.1 Discussion

By the reference of all the BI queries we have satisfied the business requirement as discussed in Section 1.

Considering the first BI query the facts and the contradiction in the production of electricity and the emission. We can relate this query to the laws confirming to the regulating emissions as well as the new techniques being developed for the production in electricity.

For the second query we have discussed the relation between the Top 10 countries based on their GDP and oil production and we can clearly say that the most developed nation has more dependency on oil production.

For the third query, in India the domestic usage of natural gas has increased over the years but as the options were made available we can say that more efficient ways could be explored even after the prices are dropped for the natural gas.

7 Conclusion and Future Work

Energy Consumption plays a vital role in different possible ways which can be in terms of GDP or the emission which can affect the environment in global level.

Built data ware house was able to answer all the queries and made easy to understand the underlying cause. Although the built data ware house was able to perform queries based on energy consumption and production of three most important fuel such as Gasoline, Electricity and Natural Gas. To find more level of granularity we need more data to be processed and could be helpful for attaining more queries which could answer them all on a single platform.

8 References

Ralph Kimball Ross M (2014) THE DATA WAREHOUSE ETL TOOLKIT

L. Moody, D. Kortnik, M.(2000) From Enterprise model to dimensional models

References

Appendix

R code

```
library(tidyverse)
library(dplyr)
library(openxlsx)
library(ggplot2)

carbon <- read.xlsx("C:\\Users\\MOLAP\\Desktop\\DWBI\\carbon.xlsx",colNames

#remove Cols
carbon <- carbon[,-c(2:18,19,20,31)]

#remove rows
carbon <- carbon[-c(1,2,3,4,9,20,23,27,44,45,46,47,48),]
#rename header
colnames(carbon) <- c("Country","2007","2008","2009","2010","2011","2012","2

df<-carbon
library(reshape2)

p<-melt(df,id.vars = "Country" , variable.name = "Year" , value.name = "CO")
```

```
write.csv (p,"C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\Carbon.csv", ro


###############ECONS

library(openxlsx)
library(tidyverse)
library(dplyr)
library(ggplot2)

Econs <- read.xlsx("C:\\Enerdata.xlsx", sheet = "Electricity␣domestic␣consum

#omitting NA
Econs <- na.omit(Econs)

#remove rows
Econs <- Econs[-c(1,2,3,4,5,6),]

#remove Cols
Econs <- Econs[,-c(2:18,30,31)]

#rename header
colnames(Econs) <- c("Country","2007","2008","2009","2010","2011","2012","20

df<-Econs
library(reshape2)

x<-melt(df,id.vars = "Country" , variable.name = "Year" , value.name = "Econ

write.csv(x,"C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\ECons.csv", row.


##########EPROD

library(openxlsx)
library(tidyverse)
library(dplyr)
library(ggplot2)

EProd <- read.xlsx("C:\\Users\\MOLAP\\Desktop\\DWBI\\Statista\\E.xlsx", shee

#omitting NA
EProd <- na.omit(EProd)

#remove rows
EProd <- EProd[-c(1,2,3,4,5),]

#rename header
colnames(EProd) <- c("Year","Electricity_Production")
```

```
write.csv(EProd,"C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\EProdStatist


########################GASOLINEPRICE

library(tidyverse)
library(rvest)
library(stringr)

urlgasoline <- "https://www.globalpetrolprices.com/gasoline_prices/"
readgasolineurl <- read_html(urlgasoline)

gasolinecountrynames <- readgasolineurl %>%
  html_nodes(xpath = '//div[contains(@id,"outsideLinks")]//a') %>%
  html_text()


gaspriceimage <- readgasolineurl %>%
  html_nodes(xpath = '//div[contains(@id,"graphic")]//img') %>%
  html_attr('src')

gaspriceimage
gaspricevalue <- strsplit(gaspriceimage,",")
gaspricevalue <- gaspricevalue[[1]]


gasoline <- cbind(gasolinecountrynames,gaspricevalue)
write.csv(gasoline,"C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\gasolineB


#Cleaning
uncleangasoline <- read.csv("C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\
venezuela <- uncleangasoline$gaspricevalue[1]
venezuela <- str_sub(venezuela,51,54)

uncleangasoline$gaspricevalue <- as.character(uncleangasoline$gaspricevalue)
uncleangasoline$gaspricevalue[1] <- venezuela

Zimbabwe <- uncleangasoline$gaspricevalue[164]
Zimbabwe <- substr(Zimbabwe,1,4)
uncleangasoline$gaspricevalue[164] <- Zimbabwe

#removestar
uncleangasoline$gasolinecountrynames <- gsub("*", "", uncleangasoline$gasoli


uncleangasoline$gasolinecountrynames <- gsub("*", "", uncleangasoline$gasoli

#remove Cols
uncleangasoline <- uncleangasoline[,-c(1)]
```

```
#rename header
colnames(uncleangasoline) <- c("Country","Gasoline␣Price")


write.csv(uncleangasoline,"C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\Ga


###############INDUSTRY

library(openxlsx)
library(tidyverse)
library(dplyr)
library(ggplot2)

Industry <- read.xlsx("C:\\Users\\MOLAP\\Desktop\\DWBI\\Dataset\\Major␣Datas


#omitting NA
Industry <- na.omit(Industry)

#remove Cols
Industry <- Industry[,-c(1,2,5:21)]

#remove rows
Industry <- Industry[-c(1,3868,3869),]
Industry <- Industry[apply(Industry!=0, 1, all),]

#rename header
colnames(Industry) <- c("Industry","Fuel␣Used","2007","2008","2009","2010","


write.csv(Industry,"C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\Industry.


###################LPGCLEANING

library(tidyverse)
library(rvest)
library(stringr)

urllpg <- "https://www.globalpetrolprices.com/lpg_prices/"
readlpgurl <- read_html(urllpg)

lpgcountrynames <- readlpgurl %>%
  html_nodes(xpath = '//div[contains(@id,"outsideLinks")]//a') %>%
  html_text()


lpgpriceimage <- readlpgurl %>%
  html_nodes(xpath = '//div[contains(@id,"graphic")]//img') %>%
  html_attr('src')

lpgpriceimage
lpgpricevalue <- strsplit(lpgpriceimage,",")
lpgpricevalue <- lpgpricevalue[[1]]

lpg <- cbind(lpgcountrynames,lpgpricevalue)
write.csv(lpg,"C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\LpgBeforeClean

#Cleaning
uncleanlpg <- read.csv("C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\Lpgbe
Algeria <- uncleanlpg$lpgpricevalue[1]
Algeria <-  substring(Algeria,regexpr("=",Algeria)+1)

uncleanlpg$lpgpricevalue <- as.character(uncleanlpg$lpgpricevalue)
uncleanlpg$lpgpricevalue[1] <- Algeria

Sweden <- uncleanlpg$lpgpricevalue[53]
Sweden <- substr(Sweden,1,4)
uncleanlpg$lpgpricevalue[53] <- Sweden

uncleanlpg$lpgcountrynames <- gsub("*", "", uncleanlpg$lpgcountrynames, fixe

#remove Cols
uncleanlpg <- uncleanlpg[,-c(1)]

#rename header
colnames(uncleanlpg) <- c("Country","Natural␣Gas␣Price")


write.csv(uncleanlpg,"C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\LpgAfte


#####################NGCONS

library(openxlsx)
library(tidyverse)
library(dplyr)
library(ggplot2)

ngcons <- read.xlsx("C:\\Enerdata.xlsx", sheet = "Natural␣gas␣domestic␣consu

#omitting NA
ngcons <- na.omit(ngcons)

#remove rows
ngcons <- ngcons[-c(1,2,3,4,5,6),]

#remove Cols
ngcons <- ngcons[,-c(2:18,30,31)]
```

```
#rename header
colnames(ngcons) <- c("Country","2007","2008","2009","2010","2011","2012","2
df<-ngcons
library(reshape2)

y<-melt(df,id.vars = "Country" , variable.name = "Year" , value.name = "NGCo

write.csv(y,"C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\NGCons.csv", row


#############NITROGEN

library(tidyverse)
library(dplyr)
library(ggplot2)
library(openxlsx)

Nitrogen <- read.xlsx("C:\\Users\\MOLAP\\Desktop\\DWBI\\Nitrogen␣Oxide.xlsx"

#remove Cols
#remove Cols
Nitrogen <- Nitrogen[,-c(2:20,31)]

#remove rows
Nitrogen <- Nitrogen[-c(1,2,3,4,9,20,23,27,44,45,46,47,48),]


#rename header
colnames(Nitrogen) <- c("Country","2007","2008","2009","2010","2011","2012",
df<-Nitrogen
library(reshape2)

d<-melt(df,id.vars = "Country" , variable.name = "Year" , value.name = "NO2"


write.csv (d,"C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\Nitrogen.csv",


################OILCONS

library(openxlsx)
library(tidyverse)
library(dplyr)
library(ggplot2)

oilcons <- read.xlsx("C:\\Enerdata.xlsx", sheet = "Oil␣products␣domestic␣con

#remove rows
oilcons <- oilcons[-c(1:8),]
```

```
#remove Cols
oilcons <- oilcons[,-c(2:18,30,31)]

#omitting NA
oilcons <- na.omit(oilcons)

#rename header
colnames(oilcons) <- c("Country","2007","2008","2009","2010","2011","2012","
df<-oilcons
library(reshape2)

z<-melt(df,id.vars = "Country" , variable.name = "Year" , value.name = "Oilc


write.csv(z,"C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\OilCons.csv", ro

###############OILPROD

library(openxlsx)
library(tidyverse)
library(dplyr)
library(ggplot2)

oilprod <- read.xlsx("C:\\Enerdata.xlsx", sheet = "Refined␣oil␣products␣prod

#omitting NA
oilprod <- na.omit(oilprod)

#remove rows
oilprod <- oilprod[-c(1:6),]

#remove Cols
oilprod <- oilprod[,-c(2:18,30,31)]

#rename header
colnames(oilprod) <- c("Country","2007","2008","2009","2010","2011","2012","

df<-oilprod
library(reshape2)

a<-melt(df,id.vars = "Country" , variable.name = "Year" , value.name = "Oilp


write.csv(a,"C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\Oiprod.csv", row


###################PGPROD

PGPROD <- read.xlsx("C:\\Users\\MOLAP\\Desktop\\DWBI\\Statista\\PG.xlsx", sh
#remove rows
PGPROD <- PGPROD[-c(1,2),]

#remove Cols
PGPROD <- PGPROD[,-c(3,5,6,7)]

#rename header
colnames(PGPROD) <- c("Year","Gasoline","Natural␣Gas")


write.csv(PGPROD,"C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\PGPRODStati


###################Sulphur

Sulphur <- read.xlsx("C:\\Users\\MOLAP\\Desktop\\DWBI\\sulphur.xlsx",colName

#remove Cols
Sulphur <- Sulphur[,-c(2:20,31)]

#remove rows
Sulphur <- Sulphur[-c(1,2,3,4,9,20,23,27,44:48),]


#rename header
colnames(Sulphur) <- c("Country","2007","2008","2009","2010","2011","2012","

df<-Sulphur
library(reshape2)

c<-melt(df,id.vars = "Country" , variable.name = "Year" , value.name = "SO2"


write.csv (c,"C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\Sulphur.csv", r


####################GDP

totalenergygdp <- read.xlsx("C:\\Enerdata.xlsx", sheet = "Energy␣intensity␣o

#rename header
colnames(totalenergygdp) <- c("Country","1990","1991","1992","1993","1994","

#remove rows
totalenergygdp <- totalenergygdp[-c(1:8),]
#remove Cols
totalenergygdp <- totalenergygdp[,-c(2:18,29:31)]


#omitting NA
totalenergygdp <- na.omit(totalenergygdp)
```

```
df<-oilcons
library(reshape2)

b<-melt(df,id.vars = "Country" , variable.name = "Year" , value.name = "GDP"


write.csv(b,"C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\totalenergyGDP.c


#############ELEPRICE

library(tidyverse)
library(rvest)
library(stringr)

url <- "https://www.globalpetrolprices.com/electricity_prices/"
readurl <- read_html(url)

countrynames <- readurl %>%
  html_nodes(xpath = ’//div[contains(@id,"outsideLinks")]//a’) %>%
  html_text()

priceimage <- readurl %>%
  html_nodes(xpath = ’//div[contains(@id,"graphic")]//img’) %>%
  html_attr(’src’)

priceimage
pricevalue <- strsplit(priceimage,",")
pricevalue <- pricevalue[[1]]

electricity <- cbind(countrynames,pricevalue)
write.csv(electricity,"C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\Electr

#Cleaning
unclean <- read.csv("C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\Electric
burma <- unclean$pricevalue[1]
burma <- substring(burma,regexpr("=",burma)+1)
unclean$pricevalue <- as.character(unclean$pricevalue)

unclean$pricevalue <- as.character(unclean$pricevalue)

unclean$pricevalue[1] <- burma

denmark <- unclean$pricevalue[94]
denmark <- substr(denmark,1,4)
str(unclean)
a <- sub(’"’,"",denmark)
unclean$pricevalue[94] <- as.factor(a)

unclean$countrynames <- gsub("*", "", unclean$countrynames, fixed = TRUE)
```

```
#remove Cols
unclean <- unclean[,-c(1)]

#rename header
colnames(unclean) <- c("Country","Electricity␣Price")


class(pricevalue)
write.csv(unclean,"C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\Electricit

################Emission combine

carbon.df = read.csv("C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\Carbon.
Nitrogen.df = read.csv("C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\Nitro
#merge files
merge(carbon.df,Nitrogen.df, all = TRUE)
write.csv(merge(carbon.df,Nitrogen.df, all = TRUE),"C:\\Users\\MOLAP\\Deskto

emission.df = read.csv("C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\Emiss
Sulphur.df = read.csv("C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\Sulphu
emission.df<- merge(emission.df,Sulphur.df, all = TRUE)

emission.df <- na.omit(emission.df)

emission.df <- emission.df[,-c(3)]

write.csv(emission.df,"C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\Emissi

##############FUELPRICE_COMBINE

gasoline.df = read.csv("C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\Gasol
electricity.df = read.csv("C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\El
#merge files
fuelprice <- merge(gasoline.df,electricity.df, all = TRUE)

fuelprice <- na.omit(fuelprice)
write.csv(merge(gasoline.df,electricity.df, all = TRUE),"C:\\Users\\MOLAP\\D

fuelprice.df = read.csv("C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\fuel
LPG.df = read.csv("C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\LpgAfterCl
fuelprice<- merge(fuelprice.df,LPG.df, all = TRUE)

fuelprice <- na.omit(fuelprice)

fuelprice <- fuelprice[,-c(2)]

colnames(fuelprice) <- c("Country","Gasoline␣Price","Electricity␣Price","Nat

write.csv(fuelprice,"C:\\Users\\MOLAP\\Desktop\\DWBI\\Cleaned␣Data\\fuelpric
```