**ETL - Group 9**

**Proposal**

**Project Title:-** Global Renewable Energy

**Team Members:-** Adam Lever and Hideaki Kaneko

**Project Description/Outline:-**  Presenting the findings of global trends in renewable energy use to a global energy provider looking at opportunities to expand current and future business interests. Concentrating our analysis on Australia, United States of America, China and the countries with the highest and lowest usage of renewable energy by percentage and/or amount in the latest data available.

**EXTRACT. Data Sources:-**

**World Bank**

Renewable energy consumption (% of total final energy consumption)

* <https://data.worldbank.org/indicator/EG.FEC.RNEW.ZS>

GDP (current US$)

* <https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?end=2015&start=1990>

**International Energy Agency**

World Energy Balance

* <https://www.iea.org/data-and-statistics/data-product/world-energy-balances#data-sets>

**Weather API**

Average Temperature, Wind Speed, Sunshine Time

https://openweathermap.org/api

**TRANSFORM:-**

* Cleaning of data
* Filtering and selection of target data
* Calculating total amount of renewable energy usage
* Joining of data tables (likely inner join by country)

**LOAD:-**

* Postgres
* ERD (by Quick DBD)

**Breakdown of Tasks:-**

**Adam;**

Cleaning of data, filtering and selection of data. Report writing. Git Hub Repository

**Hideaki;**

Cleaning of data, ERD, Report writing.

**ETL - Group 9**

**Report**

**Motivation;**

Taking into account the trend of EX (Energy Transformation) in industries, identifying the trend of renewable energy consumption growth in each country by using the data from World Bank & International Energy Agency.

**Extract;**

Data was downloaded from the various websites described in the project's proposal in both .csv and the normal excel file type. These files were stored in a data lake in a local repository.

**Transform;**

It was found that an initial clean of the data, ready for reading into a jupyter notebook, was most easily accomplished in the datas current file format in either .csv or excel, with the excel file being cleaned and then the required data being then saved in .csv format.

All data was reduced to the time period of between 1990 to 2015 as these were the years covered within the ‘Renewable Energy Consumption’ data. This allowed for like to like comparison with no superfluous data. Excess columns with redundant data were also removed at this stage.

The ‘Total Energy Consumption’ worksheet was filtered to extract the ‘Total Final Consumption’ of energy for each country. Within this data, the country of China was stated given its long title of ‘People's Republic of China’, this did not match the other data and so was manually changed to ‘China’. The data then had to be copied and pasted into a new worksheet before being saved as a .csv.

All three .csv files were loaded into Jupyter Notebook and turned into dataframes using the pandas and csv modules with the read\_csv instruction. Once in Jupyter Notebook, the name of the column of the countries within the ‘Total final renewable energy consumption’ was added to match that of the other files as it was absent.

The Total Energy Consumption and Renewable Energy Consumption dataframes were merged on the Country Name column and the suffixes for the duplicated column names set. This new merged dataframe was then exported to a .csv.

Once in the .csv format additional cleaning of the data was undertaken by removing unwanted columns. Additional columns were then created and the equivalent tonne of oil of Renewable Energy Consumption for each year and country was calculated. This new .csv was then imported into Jupyter Notebook.

A data frame for Total Energy consumption and Energy Consumption per year was created using the .T function to transpose the index and columns. These data frames were then exported to a .csv and then subsequently cleaned in excel.

**Load;**

Code was written in Jupyter Notebook to extract and print column names from dataframe so that they could be entered into PostGres SQL.

A database was created to store our data with 5 tables being added to PostGres in total;

* Total and Renewable Energy Consumption by ktoe and with Renewable Energy Consumption as a percentage (tec\_rec)
* Total Energy Consumption (tec)
* Renewable Energy Consumption (rec)
* Total Energy Consumption per year (tec\_per\_year)
* Renewable Energy Consumption per year (rec\_per\_year)

First Table which is 'tec\_rec' is the table with that data "TEC (Total Energy Consumption) in ktoe (kilo ton oil equivalent)", "REC (Renewable Energy Consumption) Percentage in TOC (%)" and "REC in ktoe". The data has been sorted by each country. The other 2 tables, 'tec' & 'rec' are spin out from the 'tec\_rec' with the foregin key of 'country\_name'.

We've also created 2 tables for "TEC per year" & "REC per year" as mentioned in above. For these 2 tables, the data has been sorted by year (from 1990 to 2015). That allows us to compare the consumption of energy between countries in a specific year.

Relations between each table are described in ERD in the repository.

A table ‘etlproject’ was created in PostGres and the column names were added. The merged and subsequently cleaned .csv was then imported into the ‘etlproject’ table.