Data Mining Project

Data Analysis on World Development Indicators

Surajudeen Abdulrasaq
MLDM M1

Introduction: The World Development Indicators from the World Bank contain over a thousand annual indicators of economic social and all other aspects of human development from hundreds of countries around the world. It presents the most current and accurate global development data available, and includes national, regional and global estimates.

Data description: The data is presented in its raw format in a fragmented version about 5 different excels format containing different information therefore effort is required for proper understanding and merging of the data. This data is Available *at https://data.worldbank.org/data-catalog/world-development-indicators (World Bank),* in addition the data is also available (both the raw data from World Bank and the result of different integral analysis I perform in the **.CSV** format) on my GITHUB at surajrasaq.

Problem Understanding: In my quest to embark on these projects, I hope that at the end of this analysis questions like the following will be answered.

- What are the factors affecting global developments?
- What country or region is developing faster?
- What years are the country or region develop most
- At what rate is this developments taking place
- Can we compare region or countries together to know how they are faring
- Can we predict between some particular countries which one will develop faster in the near future.
- Is it possible to predict global development in 2018 using this data?

Data Understanding: Each excel file was merge and converted in to **.CSV** for easier understanding and analysis and the final file is named (WDIData.csv), first i check the structure and rename column for easy analysis and finally I have a data of *415800 observations and 63 variables*:

str(world_dev) #415800 obs. of 63 variables:

The Year Column *(1960...1961…. 1962.....)* is spread Across Rows, and this will be difficult to deal with So I Combine Multiple Columns of Data into a Single Column to create a single column I called *"Year"*. Then all Not available (NA) was Remove.

Then I regroup dataset using the *Indicator_Code & Indicator_Name* as an important factor so we can easily visualize countries affected by these indicators and the affected years with the number of year's group in another column, after this process a new *data-frame (world_indi)* was created with 1575 observation and 6 variables.

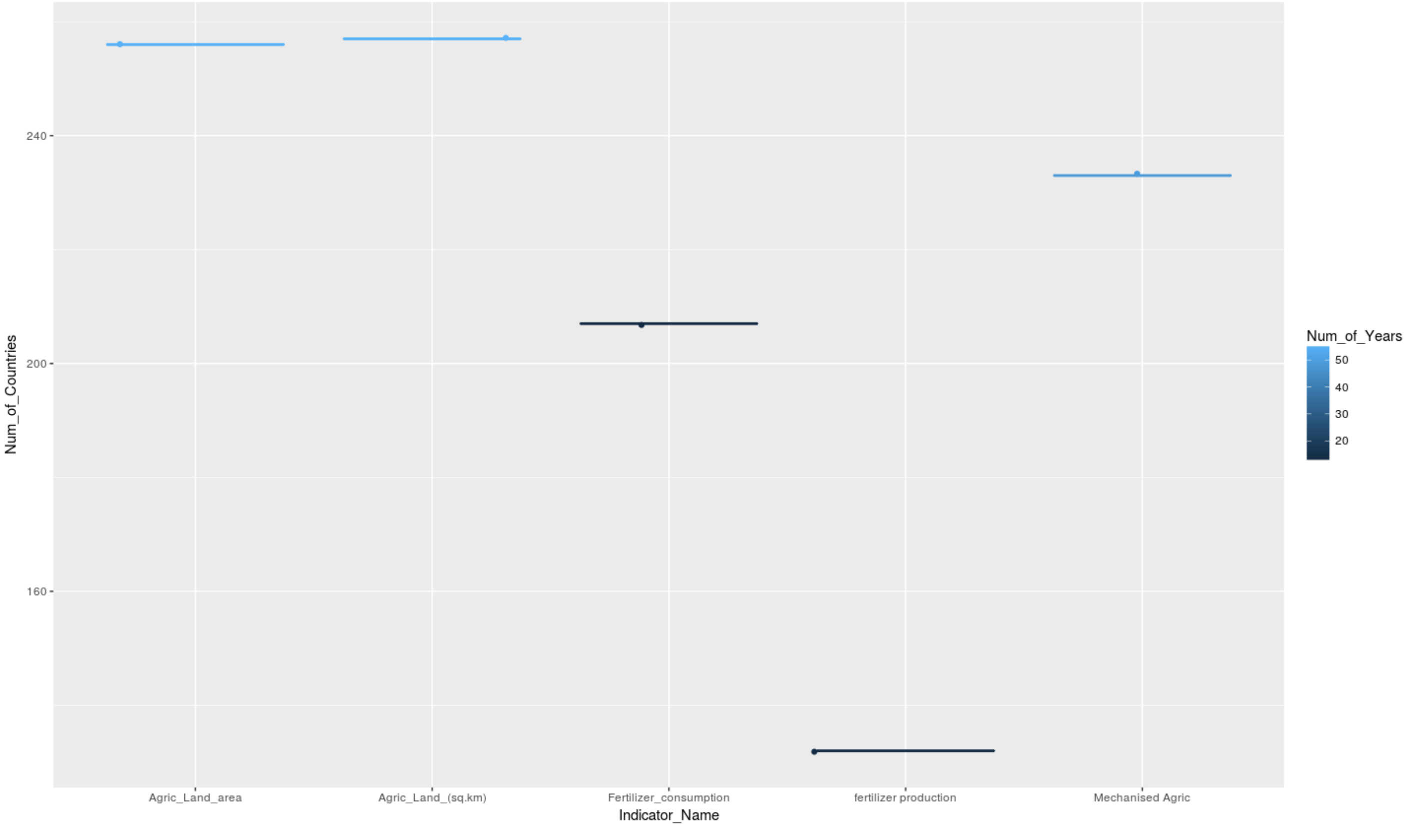
Data Preparation: Now we have a set of new set of transformed data (*world_indi*) which tell us the number of countries affected by particular indicators and The Number of years it affect them and from what year to what year this are spread out.

Now I examine Maximum Number of countries and the Maximum Number of years a particular indicator Affect any country, this enable us to understand and know some very important indicators in this study.

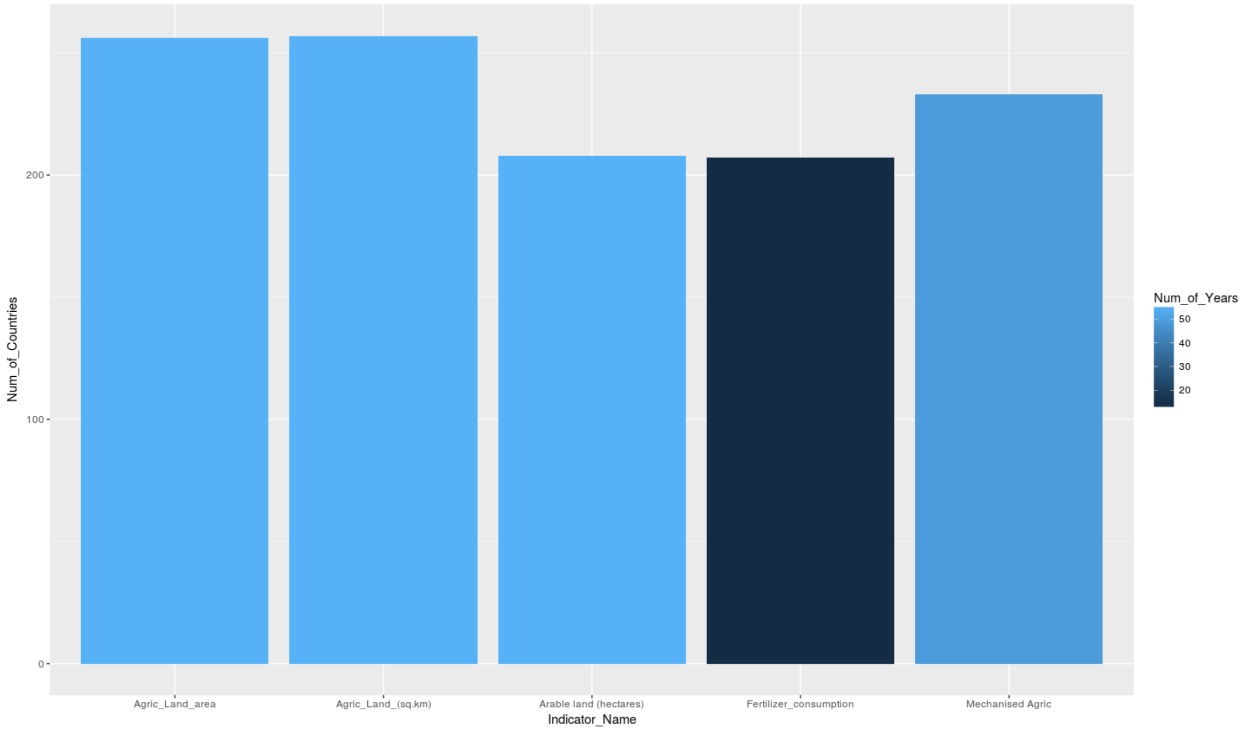
max(world_indi$Num_of_Countries) # Maximum No of country affected by A OR some particular indicator = 263.

max(world_indi$Num_of_Years)# Maximum No or years this indicator has affected the country = 58 YEARS

Visualizing some top 5 indicator: Agricultural land and mode of farming has been a major factor seem to be very important Indicator, But we can explore further to be certain by extracting more information from our data-set Since we know the biggest number's of affected countries is 263, i take a sample of this affect between 200 and 263 countries to get to know how this indicator affected them by creating a new data-frame called Important_Indicators to take the list of most important indicators.

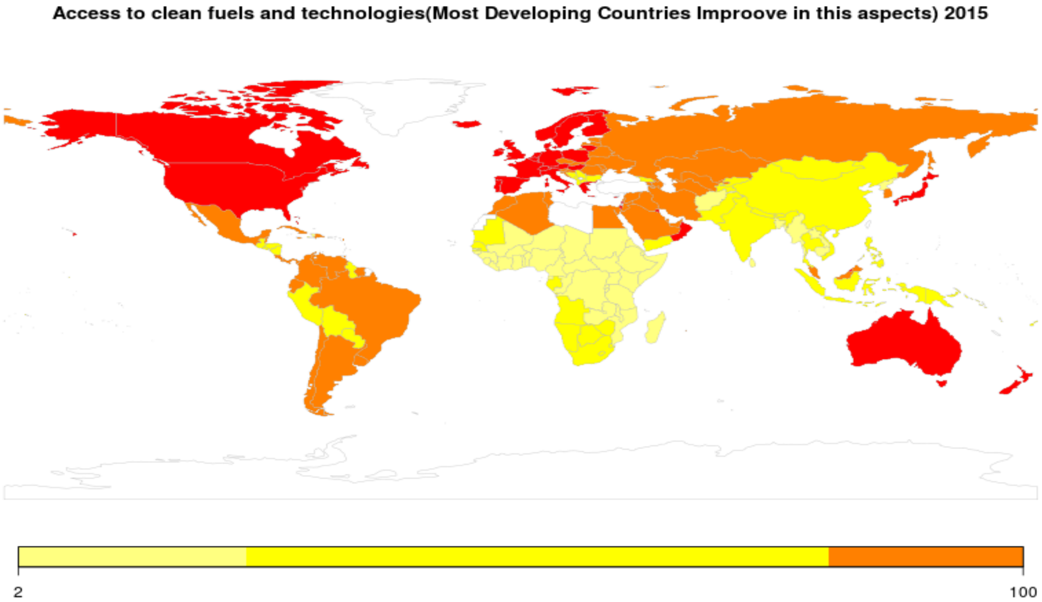


Agricultural land and mode of farming has been a major factor for the past 58 years. It’s much more obvious using bar plot below.

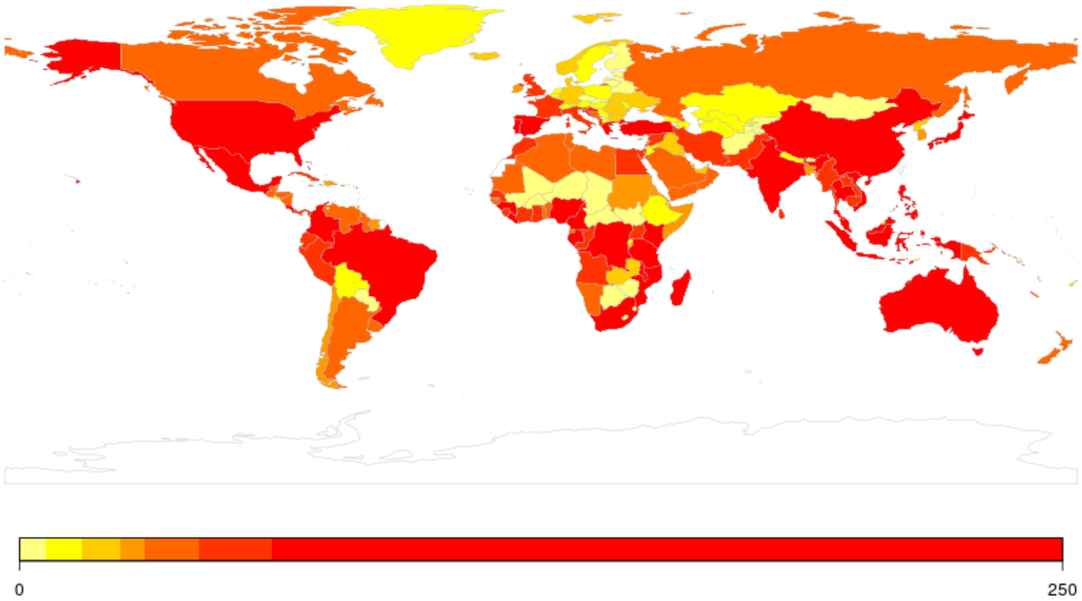


Now that we have different sets of data-sets *(world_dev and world_indi, important_indicators*) with different useful information I decided to merge the two dataset(world_indi and world_dev) in to a to single data-frame, taking out duplicate observation this is necessary to see if there can be any *surprising discovery,* the new data frame is *(data.combine with 5597260 obs. of 10 variables)* then we the correlation of some instances for ecample (*table(data.combine$Values) # Values is a very important factor)* then we continue to the visualize shows that *most developing countries improve in Access to clean fuels and technologies for cooking through the years.*

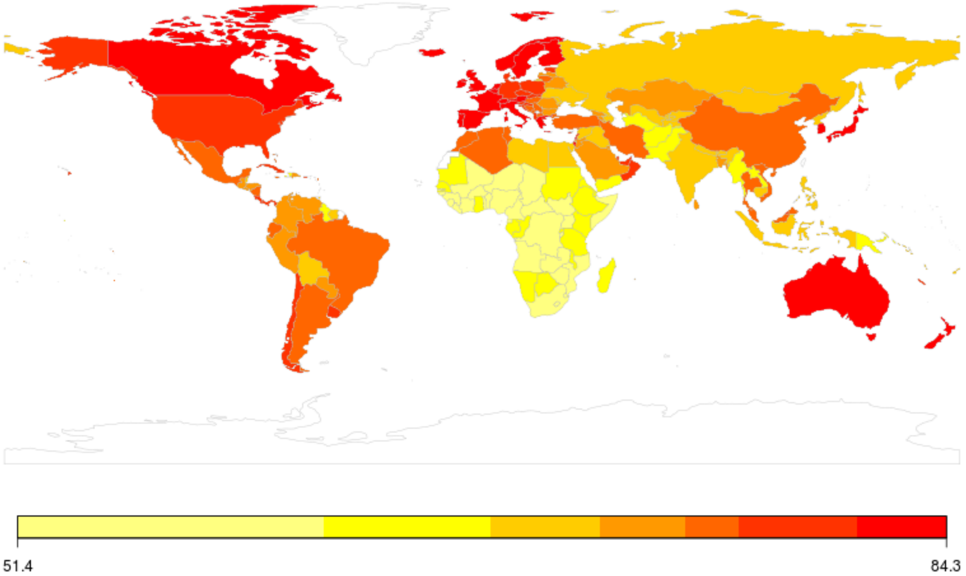
We randomly visualize some top most important indicators



Access to clean fuels and technologies(Most Developing Countries Improove in this aspects) 2015
2
100

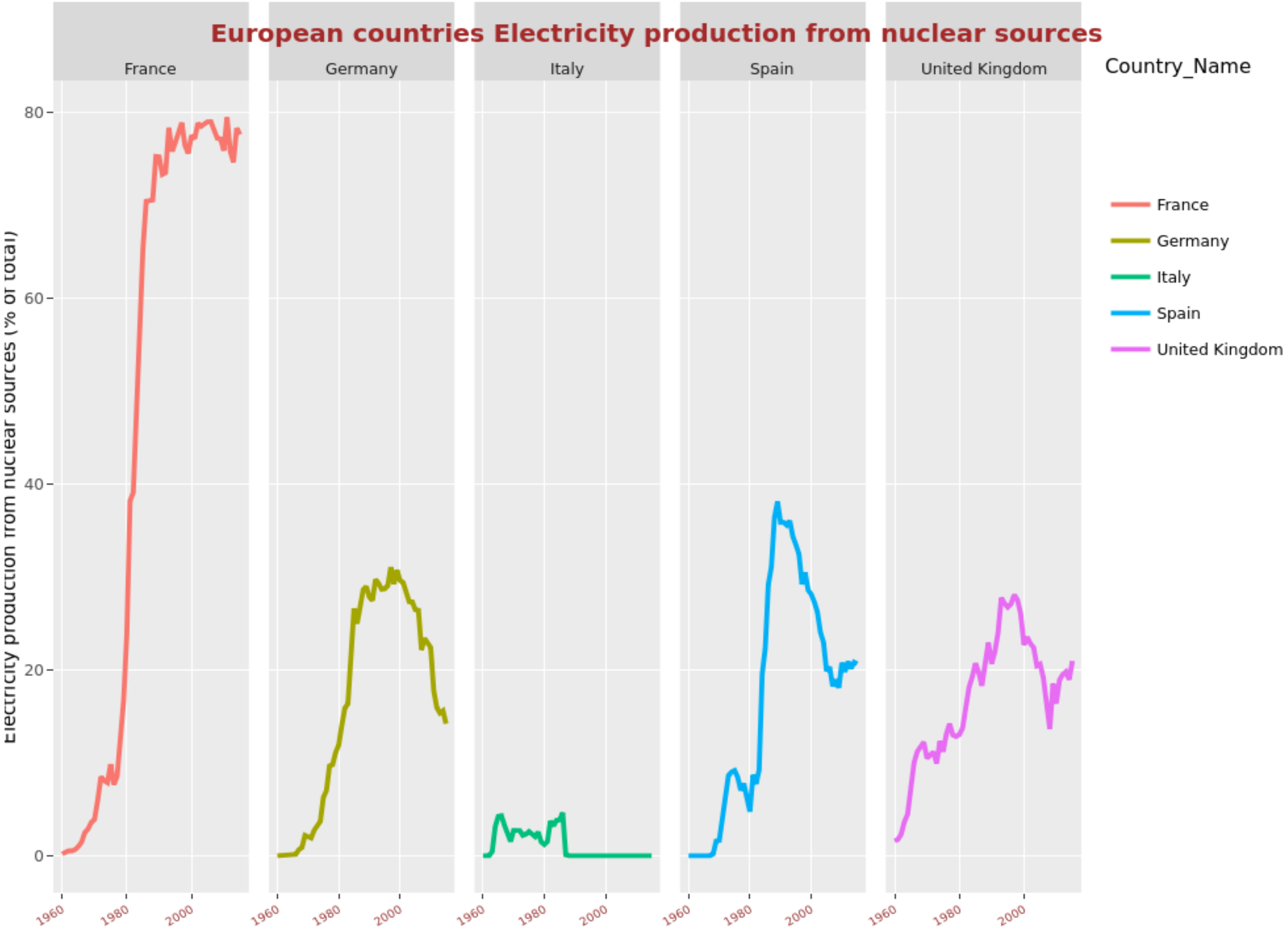


Fish species, threatened 2017

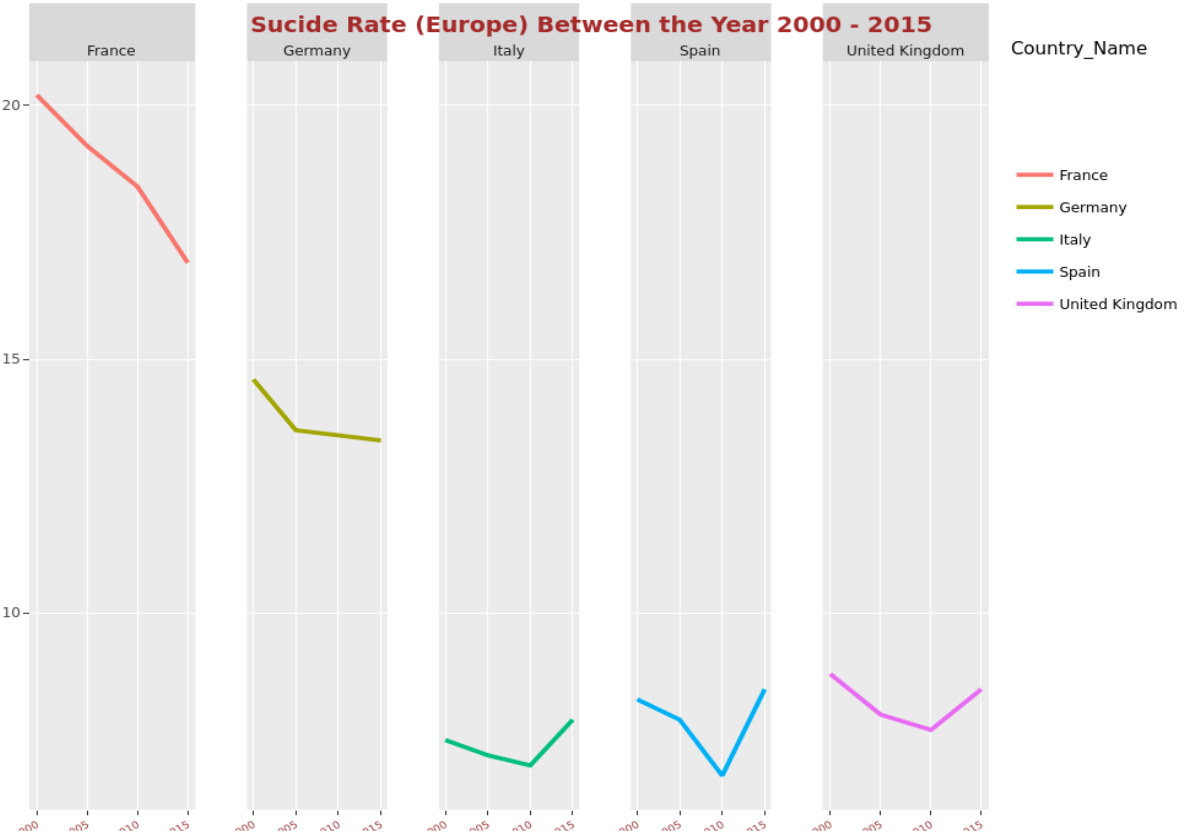


Pump price for gasoline (USdollar per liter) in year 2017

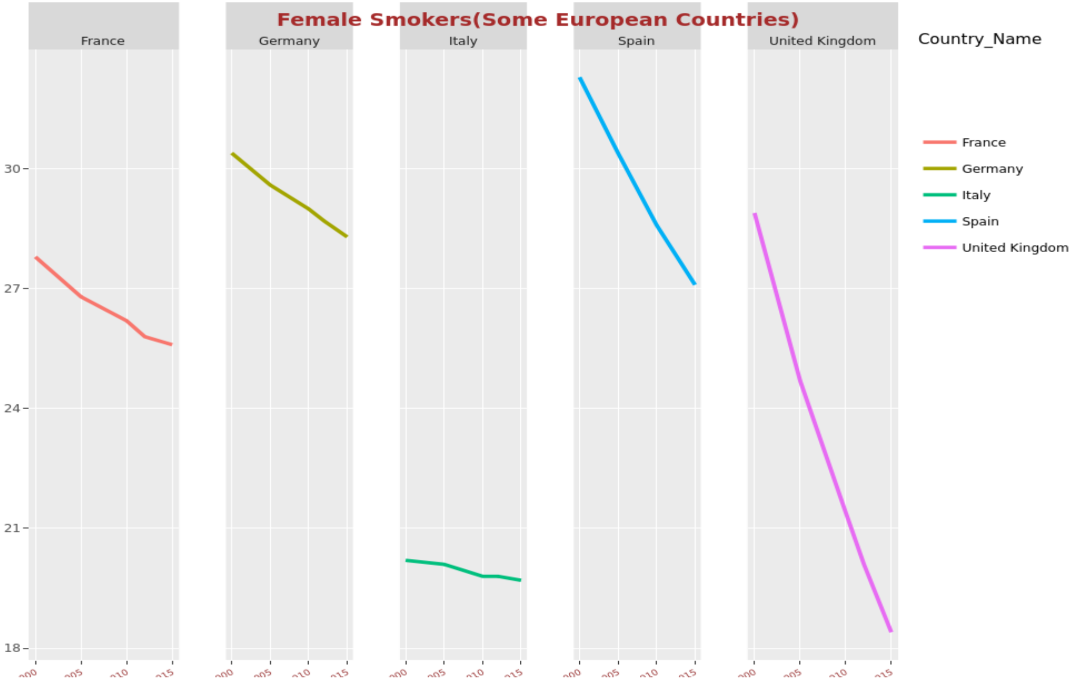
Exploratory Modeling: Due to the structure of this data I will to do exploring and modeling simultaneously, this enhance my understanding of the data and in the process surprise and more interesting facts may come up, also I choose France as a pivoting point in this study because this is where I leave for the moment and a place of interest, the exploration was done in such a way that we need to pick an indicator which will generate a new data frame and can be used to compare some countries of interest some of this is describe below, we can also compare two individual countries together based on certain indicator and determined how they fare.



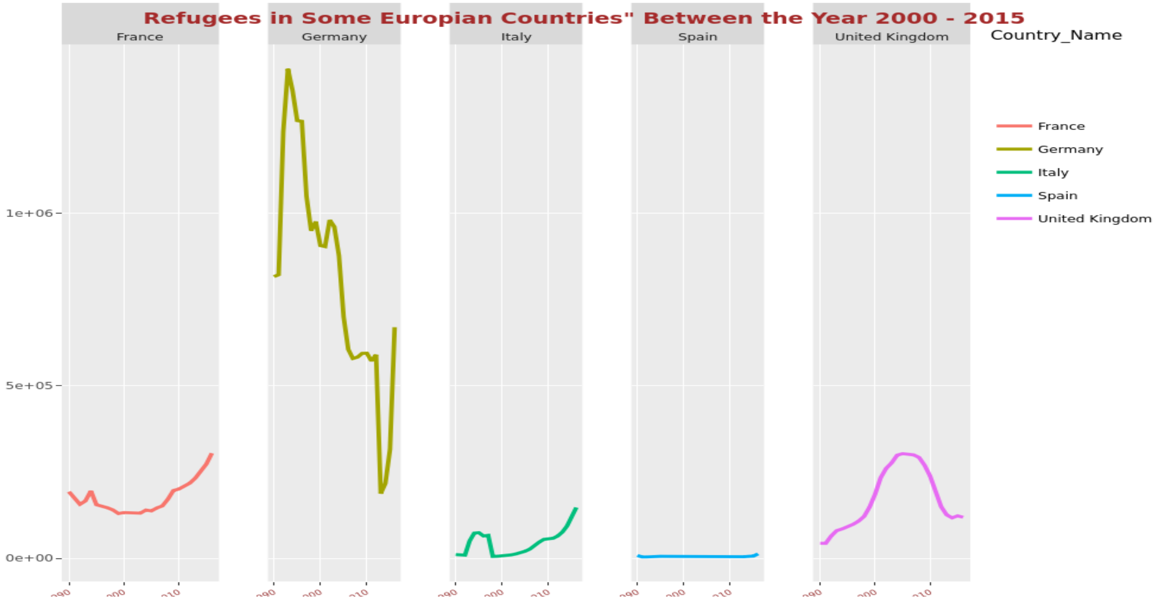
Electricity generation through nuclear source (France is leading compare to other big name in Europe)



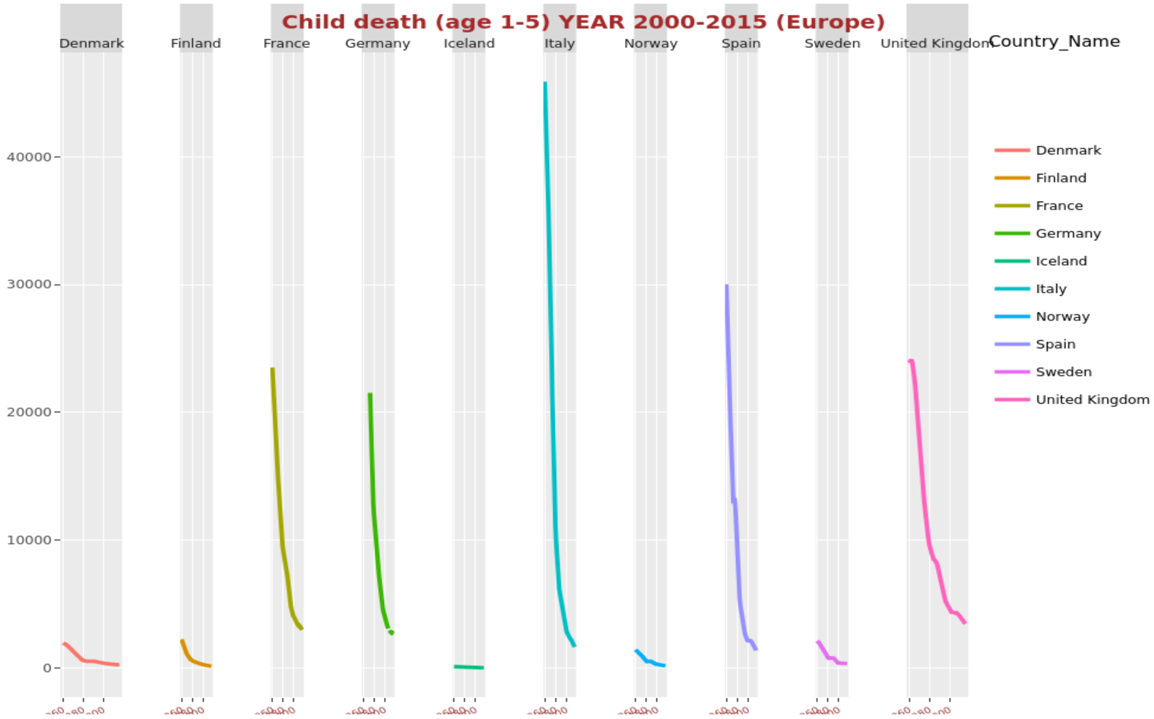
Suicide Rate in Europe (Shows France has the highest rate between year 2000 and 2015)



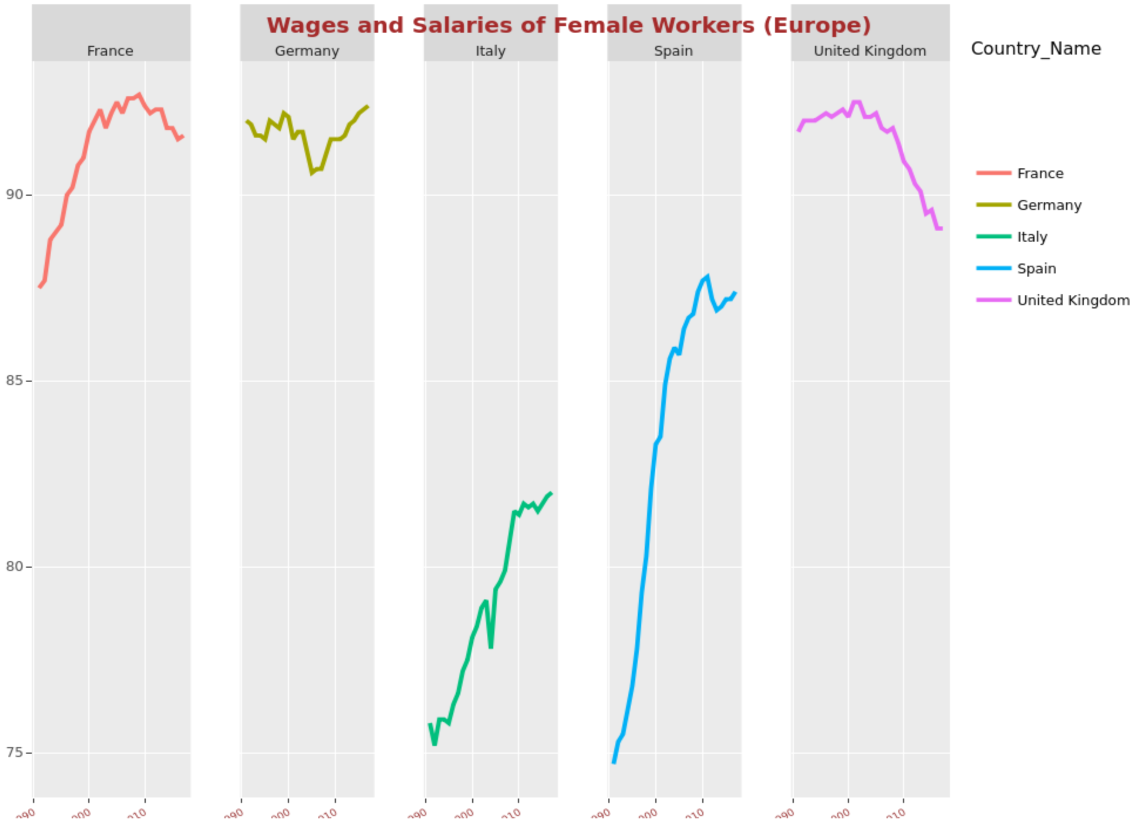
Female Smokers in some selected countries in Europe (Spain has more female smokers)



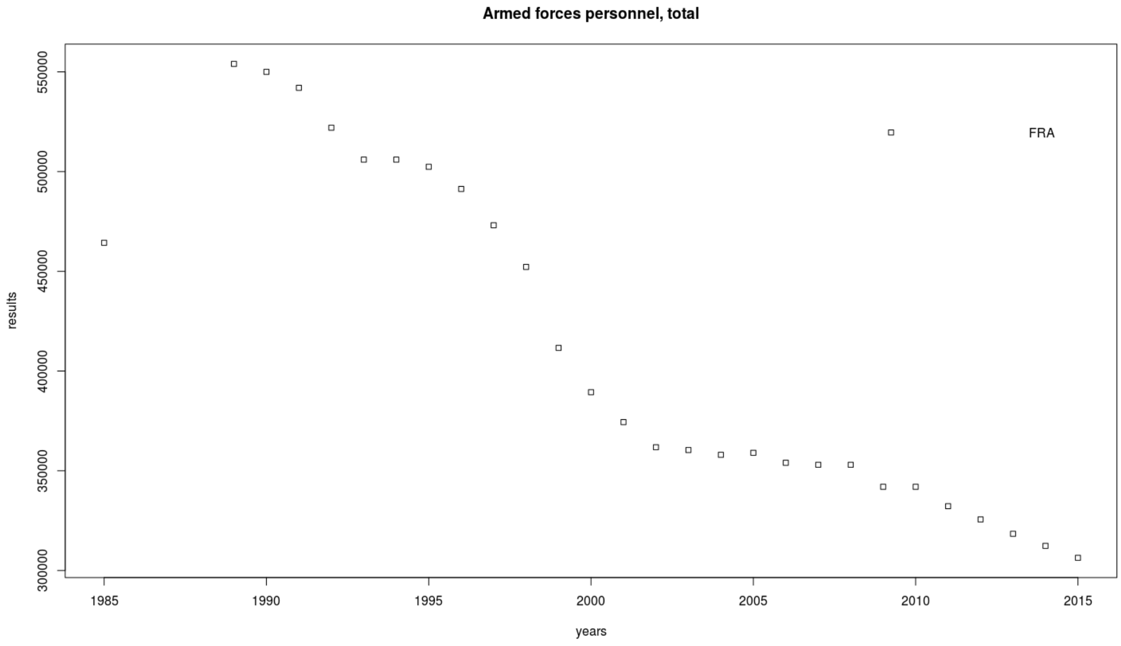
Refugees leaving in some European countries (Germany has the largest refugees from 2000-2015



Under-age child death, Italy is having a serious case follow by Spain

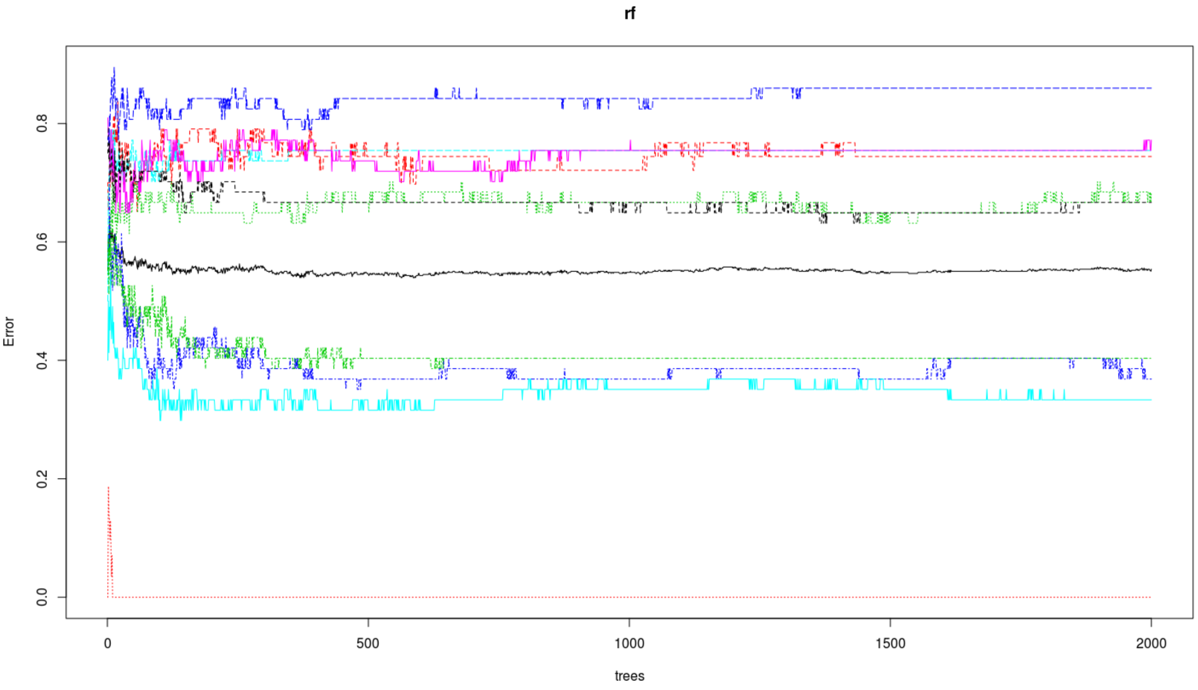


Wages and Salary of Female workers (France is good here comparing to other European Countries)



Total number of French armed force has been decline since 1990

Training the model using random forest: I choose to predict the prevalence of child death between the ages of 1-5, almost all the countries in Europe are included in this study, i set the number of trees to 2000.

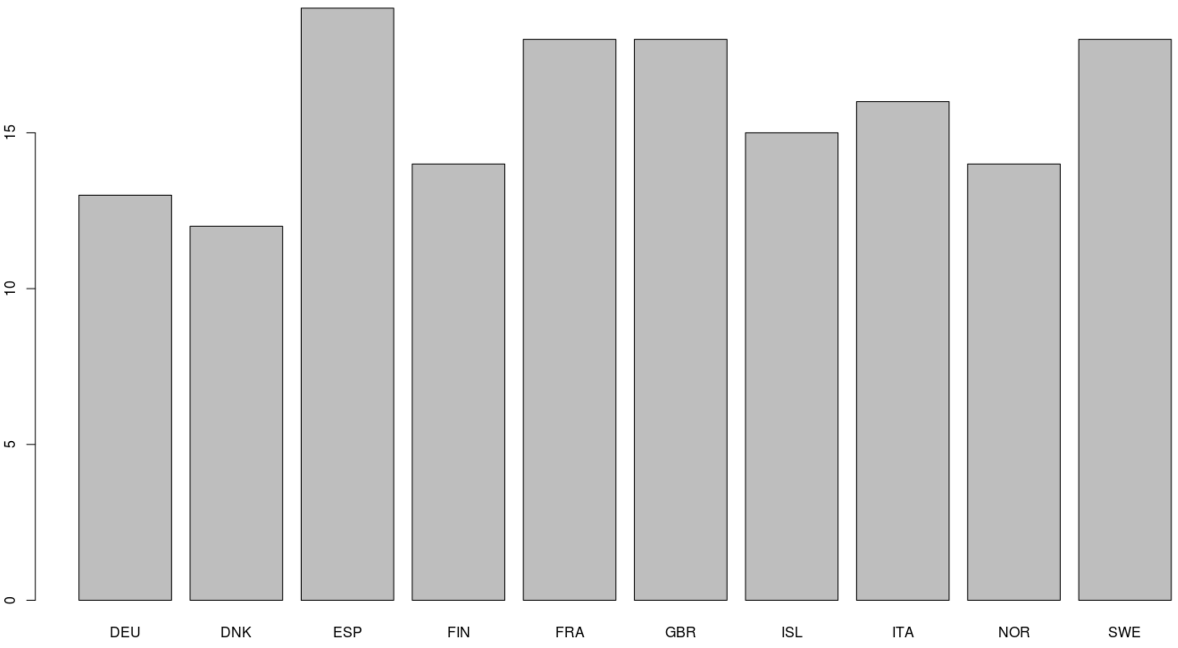


Plot of the random forest with the error rate

Evaluation: Now i do cross validation, this is necessary to avoid OVER-FITTING and help to evaluate our error rate against unseen data, after cross validation I test the model using the split data from the cross-validation, and the prediction shows Spain will be having more children death under the age of 5 years in the near future although France and Germany has high prevalence. Note that in our analysis Italy was having a serious case but our prediction for the future chooses Spain as the country of interest for under-age (1 -5 years) death, although the study initially shows Spain as second on the chart next to Italy but the model clearly predict Spain has having a higher case in the year 2018.

Table and plot of prediction below

```
DEU DNK ESP FIN FRA GBR ISL ITA NOR SWE
13  12  19  14  18  18  15  16  14  18
```



Conclusion: This study shows a very highly correlated events among the nations although I choose mainly to base the analysis on the European continent with France as a pivot study case, the data is quite revealing and can bring out more surprises if explore further, especially in the continent of Africa and Asia, I was able to extracts different information and save them as a *csv* file for future studies which can be based on region or continent and year intervals.