# Women Fertility in World Countries

In this document we represent the an analysis for the Women Fertility dataset available in https://www.gapminder.org/data/ where it can be downloaded from.

The data set represent total fertility rate. The number of children that would be born to each women with prevailing age-fertility specific rates.

The data is represented in csv data file format that contains data about the fertility in terms of the number of children per a woman in 184 countries in the period between years 1800 and 2100. So the data table is 184 * 301 .

**The questions that I were interested in are:**

1. **Does the world overall woman fertility increase, decrease, or stay fixed with time?**
2. **The woman fertility in the world continents, how is it compared between each other?**
3. **Which country has the maximum woman fertility? And which one has the minimum fertility?**
4. **How is the fertility in the Arab countries compared to the mean fertility in the world?**
5. **I want to be able (as a user) to input a particular country name and see its mean woman fertility compared with the world mean fertility.**

## Assessment Issues in the Woman Fertility dataset:

After visual and Programmatic assessment we could see that:

- There is no NAN values.
- No duplicates.

But:

1) The country names are (in some cases) not compatible with the standard country names used in Python.
2) There is no a column for the continent in which the country exists
3) The organization of the data in 184 rows * 301 columns makes the process of analysis and visualization difficult

## Data Wrangling

In data wrangling stage we were interested in:

- Identifying the countries whose names are not standard.
- Modify the incorrect names to be compatible with the country names used in module pycountry_convert
- Create new column in the data set called 'Continent' that contains the continent in which each country exists.
- Saving the new data to a new dataset file of format 'csv'

We also needed to make some restructuring of the data to make it more readable and visualized in more meaningful form. So instead of representing the data in the following form clarified in the next figure:

```
df_c2.sample(5)
```

| | country | 1800 | 1801 | 1802 | 1803 | 1804 | 1805 | 1806 | 1807 | 1808 | ... | 2092 | 2093 | 2094 | 2095 | 2096 | 2097 | 2098 | 2099 | 2100 | Continent |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **58** | France | 4.41 | 4.36 | 4.31 | 4.26 | 4.21 | 4.16 | 4.10 | 4.05 | 4.00 | ... | 1.94 | 1.94 | 1.94 | 1.94 | 1.94 | 1.94 | 1.94 | 1.94 | 1.94 | Europe |
| **54** | Eswatini | 6.71 | 6.71 | 6.71 | 6.71 | 6.71 | 6.71 | 6.71 | 6.71 | 6.71 | ... | 1.79 | 1.79 | 1.79 | 1.79 | 1.79 | 1.78 | 1.78 | 1.78 | 1.78 | Africa |
| **13** | Barbados | 4.96 | 4.93 | 4.90 | 4.87 | 4.84 | 4.82 | 4.79 | 4.76 | 4.73 | ... | 1.85 | 1.85 | 1.85 | 1.85 | 1.85 | 1.85 | 1.85 | 1.85 | 1.85 | North America |
| **47** | Dominican Republic | 6.16 | 6.16 | 6.16 | 6.16 | 6.16 | 6.16 | 6.16 | 6.16 | 6.16 | ... | 1.77 | 1.77 | 1.77 | 1.77 | 1.77 | 1.78 | 1.78 | 1.78 | 1.78 | North America |
| **69** | Guyana | 5.01 | 5.01 | 5.01 | 5.01 | 5.01 | 5.01 | 5.01 | 5.01 | 5.01 | ... | 1.82 | 1.82 | 1.82 | 1.82 | 1.82 | 1.82 | 1.82 | 1.82 | 1.82 | South America |

Which has values in  184 rows * 301 columns.  This is can't be efficiently visualized using a figure.

For this reason I made aggregations of the years between 1800 to 2100 in periods of 30 years resulting in the following dataset structure:

```
df_c_grouped.sample(5)
```

| | country | Continent | 1801-1830 | 1831-1860 | 1861-1890 | 1891-1920 | 1921-1950 | 1951-1980 | 1981-2010 | 2011-2040 | 2041-2070 | 2071-2100 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **29** | Canada | North America | 5.720000 | 5.720000 | 4.797333 | 3.991000 | 3.107333 | 2.890667 | 1.608333 | 1.590000 | 1.691333 | 1.765000 |
| **139** | Saudi Arabia | Asia | 6.860000 | 6.860000 | 6.860000 | 6.860000 | 7.007333 | 7.242000 | 4.934333 | 2.249333 | 1.751000 | 1.737667 |
| **34** | China | Asia | 5.500000 | 5.500000 | 5.500000 | 5.500000 | 5.313000 | 4.967000 | 1.950667 | 1.671333 | 1.761667 | 1.795333 |
| **179** | Venezuela | South America | 5.685333 | 5.796333 | 5.912333 | 5.784333 | 5.629667 | 5.773333 | 3.151333 | 2.137333 | 1.818000 | 1.782333 |
| **169** | Turkey | Asia | 6.920000 | 6.920000 | 6.920000 | 6.879333 | 6.741667 | 5.835667 | 2.899333 | 1.923333 | 1.747333 | 1.766667 |

Which is 184 rows * 10 columns (considering only the numeric values) . This looks better and can be visualized easier.

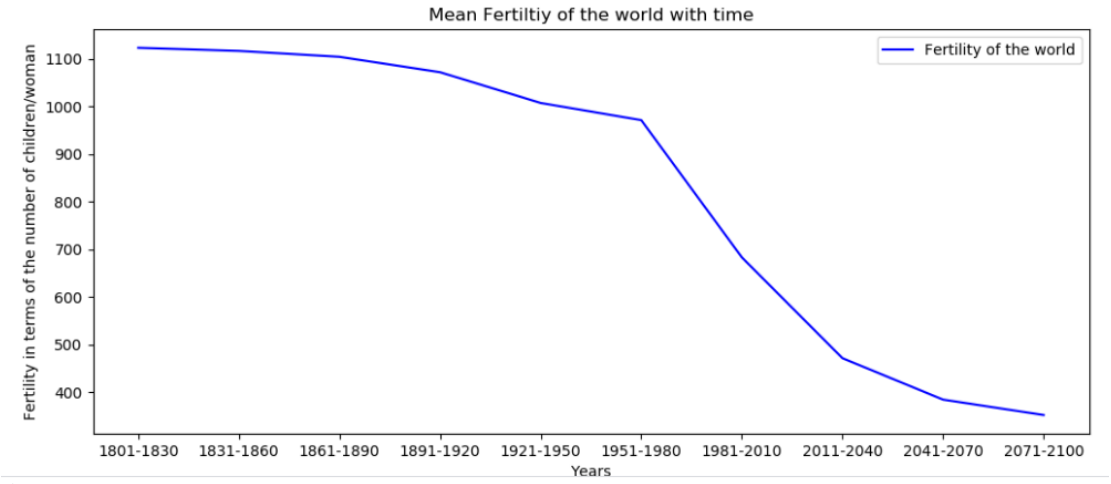
## Analysis and visualization

1) Statistics related to the fertility of the world:

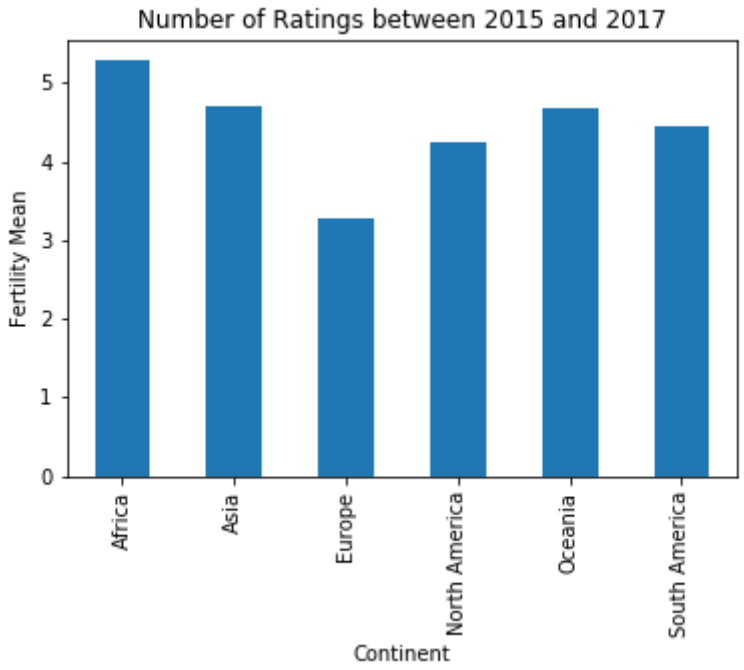
We see that:

- The mean fertility in the world= 4.508159396215514 with
  Standard deviation = 0.8878262022069234
- The country with the max fertility mean is: Niger with
  fertility mean = 6.303455
- The country with the min fertility mean is: France with
  mean fertility = 2.59794

- The mean fertility in the world decreases with time this is clear in the following figure:



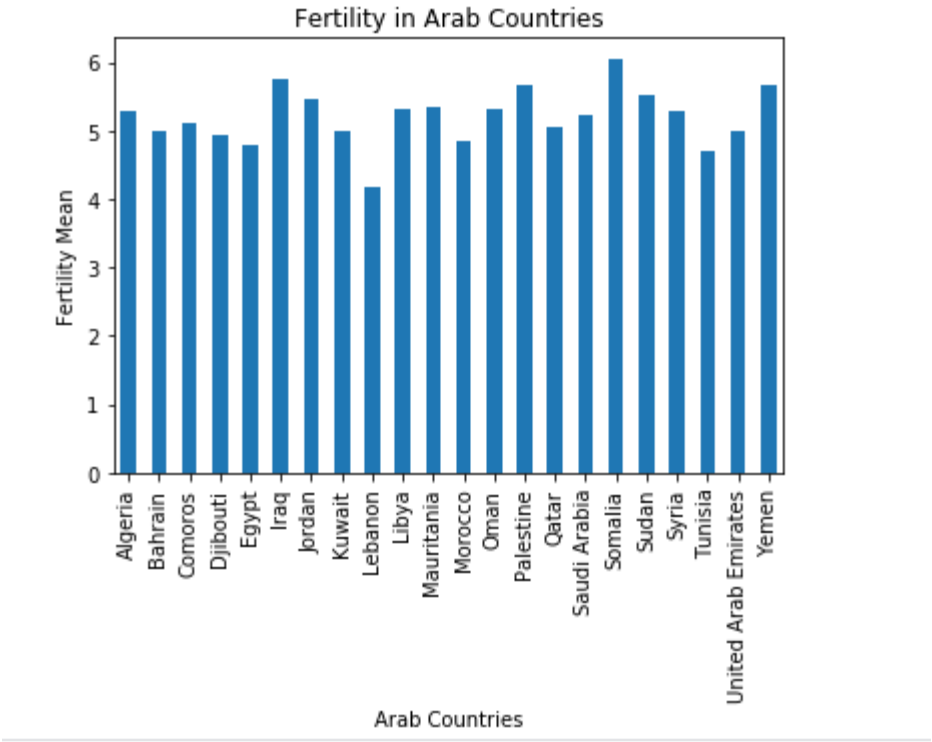

- The fertility in the world continents and how it is compared to each other, can be clarified using the following figure:



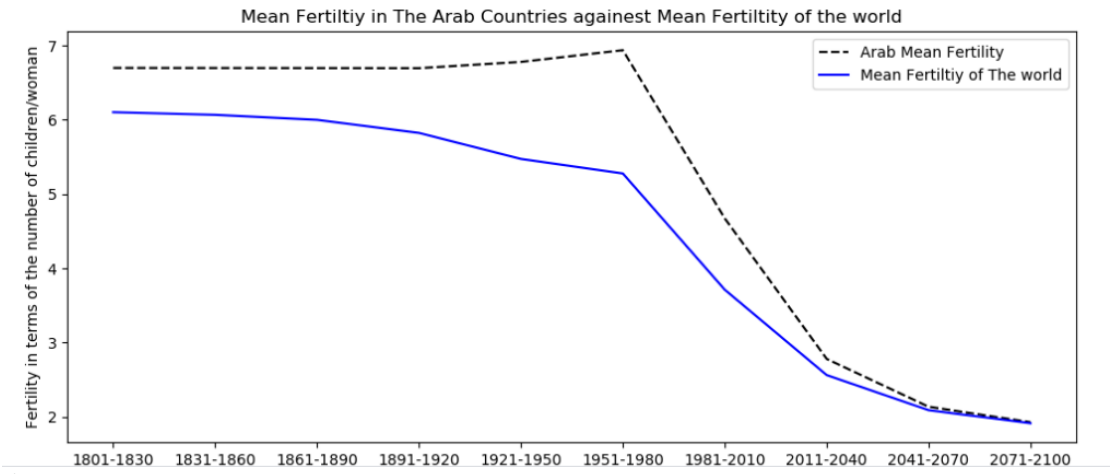

- Fertility in the arab countries. The arab countries are:

["Egypt", "Algeria" , "Sudan","Iraq", "Morocco" , "Saudi Arabia", "Yemen", "Jordan","United Arab Emirates","Syria" , "Lebanon", "Libya", "Palestine", "Oman", "Kuwait", "Somalia","Mauritania", "Qatar","Tunisia","Bahrain","Djibouti","Comoros"].
They are 22 countries.

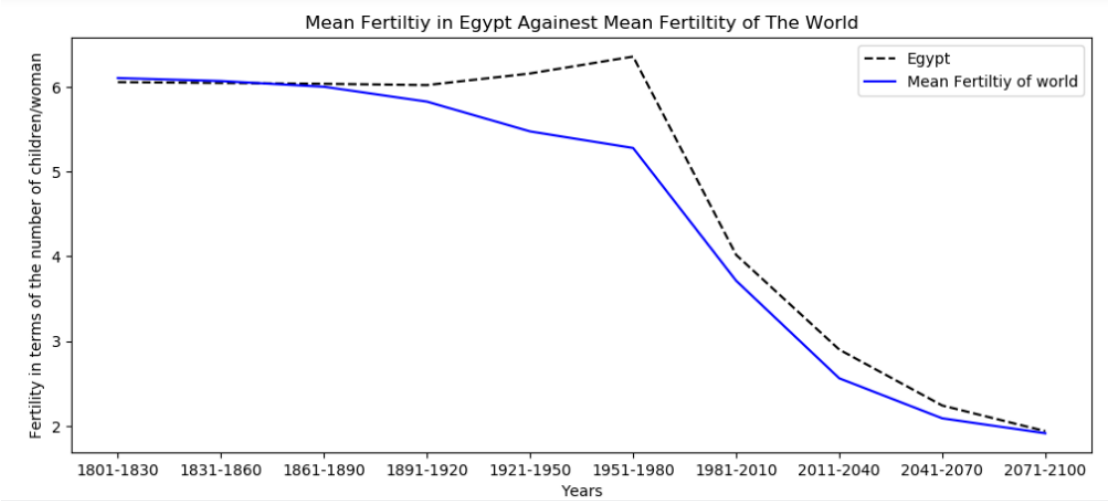
The fertility in them is shown in the following figure:




In comparison with the mean fertility of the world:




- We were interested to enable the user to input the name of the country that he wants to perform the analysis based on. As an example we input 'Egypt'. We obtained the following figure




## Conclusions and insights

According to the results of the wrangling and analysis of the "Babies per Women(total fertility)" dataset we can reach the following conclusions and answers for the questions stated in the start of this analytic study:

1. The overall world woman fertility decreases with time.
2. We can see that Africa has the highest woman fertility level, however Europe has the lowest.
3. Niger is the country with the highest fertility level, however France has the lowest level in terms of number of babies per woman.
4. Arab countries have higher fertility level than the world mean fertility. The fertility level in Arab countries decreases with time and it approaches the mean fertility level of the world
5. In general, Somalia has the largest fertility and Lebanon has the lowest level.

## Limitations and difficulties in the wrangling process:

The main difficulty I faced was "How can I reach the standard names of countries used in python?'

Unfortunately, I solved this problem by "try and error" and also by searching on the net. So there is a note that should be mentioned here is that there is 2 rows in the dataset with the same country name, "Congo".

The reason is that there are 2 countries that have the string "congo" in their names. One of them is called 'Congo' in python and the other is not. But I could not obtain its true standard name. So I did not delete its row but I gave it the name 'Congo' and preserved its data to be considered in the analysis. And fortunately this does not has a negative effect on the accuracy of the analysis as I think.

This problem also happened in with " Timor-Leste" country. I could not know its standard name. For  this reason I could not use the module 'pycountry_convert' and I had to arbitrarily set its continent to 'Asia' after search about its location.