# Mini Project

## Dataset Selection

In this project I will be using the **'World Development Indicators'** dataset.
This dataset was previously used in this course and it was obtained from the **World Bank**. It contains more than a thousand annual indicators of economic development from hundreds of countries around the world.

## Dataset Initial Exploration

I start here my initial dataset exploration. My purpose is to understand the dataset in order to work with it.

In [1]:

```
# We import the necessary libraries to start our data exploration.
import pandas as pd
import numpy as np
import os
```

In [2]:

```
# Let's check my current working directory path.
os.getcwd()
```

Out[2]:

```
'/Users/riccardobellio/Desktop'
```

In [3]:

```
# Let's check if the file Indicators.csv (which contains the dataset)
# is available in my current working directory.
os.listdir('/Users/riccardobellio/Desktop/SAN DIEGO - EDX/5 - VISUALISATIONS')
```

Out[3]:

```
['MATPLOTILB - DATA VISUALISATIONS.ipynb',
 '.DS_Store',
 'vis_3d.png',
 'TYPE OF PLOTS - COMMANDS.ipynb',
 'bar_chart.svg',
 'plot_data.html',
 'Indicators.csv',
 '05a_Matplotlib_Notebook.ipynb',
 'NEW PLOTS - INDICATORS ACROSS COUNTRIES.ipynb',
 'DATA VISUALISATION.ipynb',
 'vis_boxplot.png',
 'IMPORTANT VISUALISATION LIBRARIES.ipynb',
 'FOLIUM-NOTEBOOK-INTERACTIVE-MAP.ipynb',
 '.ipynb_checkpoints',
 'TYPES OF CHART.rtfd',
 'plot_data2.html',
 'world-countries.txt',
 'MINI_PROJECT.ipynb',
 'vis_bubbleplot.png']
```

In [4]:

In [4]:

```
# Let's read the .csv file into a Pandas dataframe.
my_df = pd.read_csv('/Users/riccardobellio/Desktop/SAN DIEGO - EDX/5 -
VISUALISATIONS/Indicators.csv')
```

In [5]:

```
# Let's check the length of the dataframe and visualise its first 5 rows.
print(len(my_df))
my_df.head(5)
```

```
5656458
```

Out[5]:

| | CountryName | CountryCode | IndicatorName | IndicatorCode | Year | Value |
|---|---|---|---|---|---|---|
| 0 | Arab World | ARB | Adolescent fertility rate (births per 1,000 wo... | SP.ADO.TFRT | 1960 | 1.335609e+02 |
| 1 | Arab World | ARB | Age dependency ratio (% of working-age populat... | SP.POP.DPND | 1960 | 8.779760e+01 |
| 2 | Arab World | ARB | Age dependency ratio, old (% of working-age po... | SP.POP.DPND.OL | 1960 | 6.634579e+00 |
| 3 | Arab World | ARB | Age dependency ratio, young (% of working-age ... | SP.POP.DPND.YG | 1960 | 8.102333e+01 |
| 4 | Arab World | ARB | Arms exports (SIPRI trend indicator values) | MS.MIL.XPRT.KD | 1960 | 3.000000e+06 |

As we can see, it is a big dataset which contains 5656458 rows and 4 main dimensions:

- Country (name and code).
- Indicator (name and code).
- Year.
- Value.

In [6]:

```
# Let's check if the dataset contains any NULL value.
my_df.isnull().any()
```

Out[6]:

```
CountryName      False
CountryCode      False
IndicatorName    False
IndicatorCode    False
Year             False
Value            False
dtype: bool
```

No, there is not a single NULL value, which is good.
We need to explore our indicators to see if we can find anything that can attract our interest.
We also need to know how many countries are included in our dataset and the overall year range.

In [7]:

```
# How many country names? Let's visualise them.
country_name = my_df['CountryName'].unique().tolist()
print(len(country_name))
# I won't execute the country_name command or the whole list will be visualised on the pdf file
# country_name
```

```
247
```

So we have 247 country names.
Some country names include individual countries, whilst other names include world regions or group of countries with similar feaures or pertaining to a similar area. It is important to take it into account of it in our research.

In [8]:

```
# Let's check the year range.
print('MIN YEAR VALUE:\t', my_df['Year'].min(), '\tMAX YEAR VALUE:\t', my_df['Year'].max())
```

```
MIN YEAR VALUE:	 1960	MAX YEAR VALUE:	 2015
```

The dataset as a whole has values ranging from 1960 to 2015.

In [9]:

```
# Let's check how many indicators are available and visualise them.
indicator_name = my_df['IndicatorName'].unique().tolist()
print(len(indicator_name))
# I won't execute the indicator_name command or the whole list will be visualised on the pdf file
# indicator_name
```

```
1344
```

Wow, 1344 different categories! I will go through them and choose the ones that attract my attention.

# Fertility Rate in Italy

I have chosen to explore the fertility rate in Italy.
Italy is a beautiful developed country which has one of the **lowest fertility rates** on the planet. What is causing this decrease in fertility rate? Is it a global trend?
Is it caused by a decreased economical prosperity for the country?
In my opinion this is a paricularly strange trend Italian people have a family culture, they tend to spend a lot of time with their relatives and live together in the same area. Thus, you would expect a high fertiltiy rate in this country.
I am relly interesting in getting a better understanding of this phenomenon.
So, let's first have a look at the fertility rate in Italy.

In [10]:

```
# We need to create some masks for our dataframe to get the data
# relating to 'Fertility rate' for 'Italy'.
mask_1 = my_df['IndicatorName'].str.contains('Fertility rate')
mask_2 = my_df['CountryName'].str.contains('Italy')

fertility_italy = my_df[mask_1 & mask_2]
print(len(fertility_italy))
fertility_italy.head(5)
```

```
54
```

Out[10]:

| | CountryName | CountryCode | IndicatorName | IndicatorCode | Year | Value |
|---|---|---|---|---|---|---|
| 12569 | Italy | ITA | Fertility rate, total (births per woman) | SP.DYN.TFRT.IN | 1960 | 2.37 |
| 37633 | Italy | ITA | Fertility rate, total (births per woman) | SP.DYN.TFRT.IN | 1961 | 2.42 |
| 65217 | Italy | ITA | Fertility rate, total (births per woman) | SP.DYN.TFRT.IN | 1962 | 2.44 |
| 93711 | Italy | ITA | Fertility rate, total (births per woman) | SP.DYN.TFRT.IN | 1963 | 2.50 |
| 122510 | Italy | ITA | Fertility rate, total (births per woman) | SP.DYN.TFRT.IN | 1964 | 2.65 |

In [11]:

```
# We can crete a lineplot to visualise the trend.
import matplotlib.pyplot as plt
from matplotlib import style
```

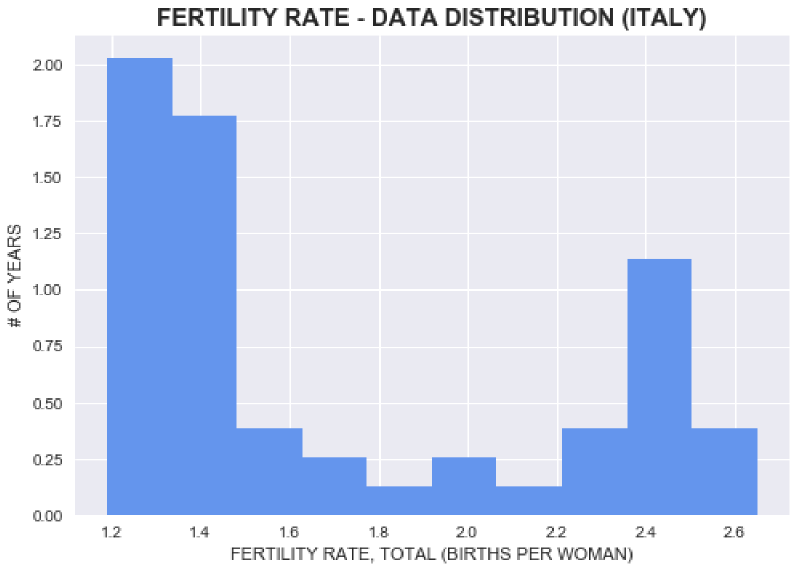
```
style.use('seaborn')
plt.plot(fertility_italy['Year'].values, fertility_italy['Value'].values, color = 'cornflowerblue',
linewidth=3)

plt.xlabel('YEAR')
plt.ylabel(fertility_italy['IndicatorName'].iloc[0].upper())
plt.title('FERTILITY RATE - BIRTHS PER WOMAN (ITALY)', fontweight="bold", fontsize= 15)
# We adjust the axis to provide a correct data visualisation
plt.axis([1959, 2015,0,2.8])
plt.grid(True)
plt.show()
```

```
<Figure size 800x550 with 1 Axes>
```

Wow, as we can see, in the 60's, the average woman was having 2.7 babies a year, and that value dropped to 1.2 babies a year in 1993... that is a remarkable drop! Since then, it seems that the rate has remained approximately stable. What happened in Italy to justify it? Let's plot an histogram of the fertility rate to double check its data distribution.

In [12]:

```
style.use('seaborn')
data_histogram = fertility_italy['Value'].values
plt.hist(data_histogram, 10, density=True, facecolor='cornflowerblue')

plt.xlabel(fertility_italy['IndicatorName'].iloc[0].upper())
plt.ylabel(' # OF YEARS')
plt.title('FERTILITY RATE - DATA DISTRIBUTION (ITALY)',fontweight="bold", fontsize= 15)
plt.grid(True)
plt.show()
```



The data distributon confirms the previous result.
Now, it is important to check **how Italy relates to other countries** and visualise where italy sits in terms of fertility on the planet.

We can visualise it for the year 1965 (peak value for Italy) and 1995 (bottom value).

In [13]:

```
mask_3 = my_df['IndicatorName'].str.contains('Fertility rate')
mask_4 = my_df['Year'].isin([1965])

global_1965 = my_df[mask_3 & mask_4]
print(len(global_1965))
global_1965.tail()
```

```
223
```

Out[13]:

|  | CountryName | CountryCode | IndicatorName | IndicatorCode | Year | Value |
|---|---|---|---|---|---|---|
| 167424 | Vietnam | VNM | Fertility rate, total (births per woman) | SP.DYN.TFRT.IN | 1965 | 6.475 |
| 167507 | Virgin Islands (U.S.) | VIR | Fertility rate, total (births per woman) | SP.DYN.TFRT.IN | 1965 | 5.679 |
| 167588 | Yemen, Rep. | YEM | Fertility rate, total (births per woman) | SP.DYN.TFRT.IN | 1965 | 7.566 |
| 167687 | Zambia | ZMB | Fertility rate, total (births per woman) | SP.DYN.TFRT.IN | 1965 | 7.292 |
| 167899 | Zimbabwe | ZWE | Fertility rate, total (births per woman) | SP.DYN.TFRT.IN | 1965 | 7.373 |

In [14]:

```
mask_5 = my_df['Year'].isin([1995])

global_1995 = my_df[mask_3 & mask_5]
print(len(global_1995))
```

```
230
```

In [15]:

```
global_1965[global_1965['CountryName'].str.contains('Italy')]
```

Out[15]:

|  | CountryName | CountryCode | IndicatorName | IndicatorCode | Year | Value |
|---|---|---|---|---|---|---|
| 153240 | Italy | ITA | Fertility rate, total (births per woman) | SP.DYN.TFRT.IN | 1965 | 2.59 |

In [16]:

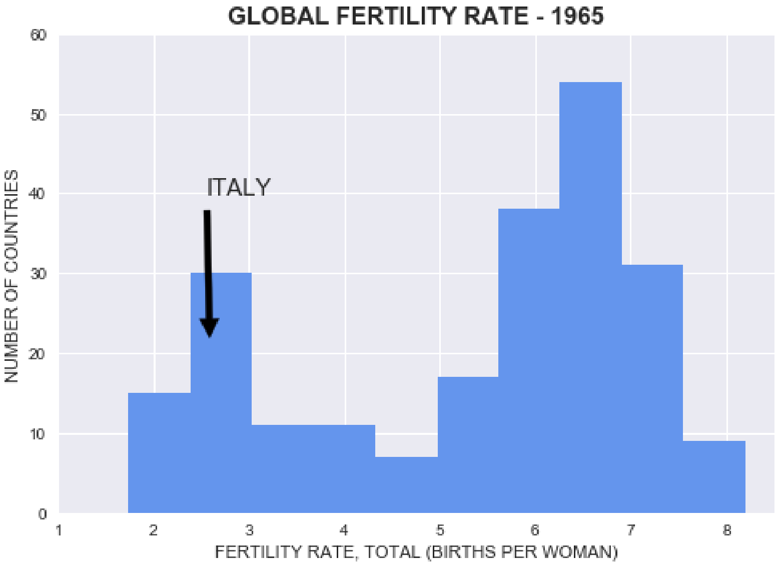
```
global_1995[global_1995['CountryName'].str.contains('Italy')]
```

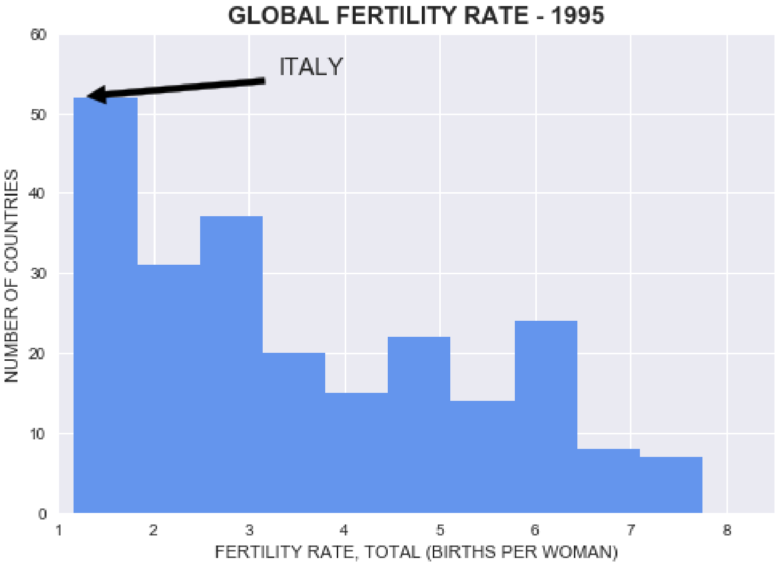
Out[16]:

|  | CountryName | CountryCode | IndicatorName | IndicatorCode | Year | Value |
|---|---|---|---|---|---|---|
| 2554388 | Italy | ITA | Fertility rate, total (births per woman) | SP.DYN.TFRT.IN | 1995 | 1.19 |

In [17]:

```
fig, ax = plt.subplots()
plt.hist(global_1965['Value'], 10, density=False, facecolor='cornflowerblue')

ax.annotate('ITALY', size = 15,
            xy=(2.59, 21),  xycoords='data',
            xytext=(0.3, 0.7), textcoords='axes fraction',
            arrowprops=dict(facecolor='black', shrink=0.05),
            horizontalalignment='right', verticalalignment='top',
            )

plt.xlabel(global_1965['IndicatorName'].iloc[0].upper())
plt.ylabel('NUMBER OF COUNTRIES')
plt.title('GLOBAL FERTILITY RATE - 1965', fontweight="bold", fontsize= 15)
plt.axis([1, 8.5, 0, 60])

plt.grid(True)
plt.show()

fig, ax = plt.subplots()
plt.hist(global_1995['Value'], 10, density=False, facecolor='cornflowerblue')

ax.annotate('ITALY', size = 15,
            xy=(1.19, 52),  xycoords='data',
            xytext=(0.4, 0.95), textcoords='axes fraction',
            xytext=(0.4, 0.95), textcoords='axes fraction',
            arrowprops=dict(facecolor='black', shrink=0.05),
            horizontalalignment='right', verticalalignment='top',
            )

plt.xlabel(global_1995['IndicatorName'].iloc[0].upper())
plt.ylabel('NUMBER OF COUNTRIES')
plt.title('GLOBAL FERTILITY RATE - 1995', fontweight="bold", fontsize= 15)
plt.axis([1, 8.5, 0, 60])

plt.grid(True)
plt.show()
```





As we can see:

- The **trend on a planetary scale sees a reduction in fertility rate**: I deduct that there may be some **socio-economical factors** on a global scale that must have push female fertility on a downtrend. An increase in female employment to seek independence? Is it a byproduct of globalisation?

- **Italy sees a reduction in its fertility rate** and is **one of the countries with the lowest rate**. So, if this is a global trend, why Italy, and not other countries are experiencing it in such a bad way?

I must admit that this global downtren in fertility is a big surpirse for me! I would like to visualise it on an **interactive map** using Folium, which will allow me to see **how different countries** have shifted from 1965 to 1995, and will provide an insight on the matter.

In [18]:

```
# Here we remove the rows on the Dataframes containing feritily rates in 1965 and 1995.
# We will keep only the countries containing individual country names.
modified_1965 = global_1965.drop([136019, 136129, 136238, 136339, 136487, 136655, 136799, 136927, 1
37053, 137179, 137285, 137447, 137581, 137701, 137855, 138041, 138217, 138342, 138485, 138601, 13876
8, 138886, 139033, 139207, 139374, 139515, 139612, 139702, 139818, 140013, 140150, 140305, 140463])
modified_1995 = global_1995.drop([2485195, 2485584, 2486037, 2486499, 2487001, 2487625, 2488141, 24
88698, 2489319, 2489817, 2490183, 2490627, 2491118, 2491574, 2492071, 2492610, 2493237, 2493699, 24
94329, 2494951, 2495547, 2495994, 2496644, 2497257, 2497707, 2498197, 2498567, 2498907, 2499362, 24
99986, 2500477, 2501131, 2501734])
```

In [19]:

```
# We double check that the cut provided the desired result.
visualise_country_list = modified_1965['CountryName'].to_list()
# I won't execute the visualise_country_list command or the whole list will be visualised on the p
df file.
# visualise_country_list
```

In [20]:

```
visualise_country_list2 = modified_1995['CountryName'].to_list()
# I won't execute the visualise_country_list2 command or the whole list will be visualised on the
pdf file.
# visualise_country_list2
```

In [28]:

```
# Let's install and then import the Folium library, which will allow us to create out interactive
maps.
!pip install folium
import folium
```

```
Requirement already satisfied: folium in /Users/riccardobellio/opt/anaconda3/lib/python3.7/site-pa
ckages (0.11.0)
Requirement already satisfied: branca>=0.3.0 in
/Users/riccardobellio/opt/anaconda3/lib/python3.7/site-packages (from folium) (0.4.1)
Requirement already satisfied: jinja2>=2.9 in
/Users/riccardobellio/opt/anaconda3/lib/python3.7/site-packages (from folium) (2.10.3)
Requirement already satisfied: numpy in /Users/riccardobellio/opt/anaconda3/lib/python3.7/site-
packages (from folium) (1.17.2)
Requirement already satisfied: requests in /Users/riccardobellio/opt/anaconda3/lib/python3.7/site-
packages (from folium) (2.22.0)
Requirement already satisfied: MarkupSafe>=0.23 in
/Users/riccardobellio/opt/anaconda3/lib/python3.7/site-packages (from jinja2>=2.9->folium) (1.1.1)
Requirement already satisfied: chardet<3.1.0,>=3.0.2 in
/Users/riccardobellio/opt/anaconda3/lib/python3.7/site-packages (from requests->folium) (3.0.4)
Requirement already satisfied: idna<2.9,>=2.5 in
/Users/riccardobellio/opt/anaconda3/lib/python3.7/site-packages (from requests->folium) (2.8)
Requirement already satisfied: certifi>=2017.4.17 in
/Users/riccardobellio/opt/anaconda3/lib/python3.7/site-packages (from requests->folium)
(2019.9.11)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in
/Users/riccardobellio/opt/anaconda3/lib/python3.7/site-packages (from requests->folium) (1.24.2)
```

In [29]:

```
#We create two dataframes (1965 & 1995) containing the country codes and values for our maps.
plot_data_1965 = modified_1965[['CountryCode','Value']]
plot_data_1995 = modified_1995[['CountryCode','Value']]
```

In [61]:

```
# We read the file that we will use to create the map.
country_geo = '/Users/riccardobellio/Desktop/SAN DIEGO - EDX/5 - VISUALISATIONS/world-
countries.txt'

# Let's set up the zoom and intial location for the map.
map = folium.Map(location=[45, 20], zoom_start=1.3)

# We use choropleth maps to visualize data combinations.
map.choropleth(geo_data=country_geo, data=plot_data_1965,
               columns=['CountryCode', 'Value'],
               key_on='feature.id',
               fill_color='YlOrRd', fill_opacity=0.8, line_opacity=0.4,
                legend_name='CHILDREN PER WOMAN - 1965')
# We create the Folium plot.
map.save('plot_data.html')

# And fianlly import the Folium interactive html file.
from IPython.display import HTML
HTML('<iframe src=plot_data.html width=540 height=350></iframe>')
```

Out[61]:

ot found.

In [62]:

```
map = folium.Map(location=[45, 20], zoom_start=1.3)
map.choropleth(geo_data=country_geo, data=plot_data_1995, columns=['CountryCode', 'Value'], key_on=
'feature.id', fill_color='YlOrRd', fill_opacity=0.8, line_opacity=0.4, legend_name='CHILDREN PER
WOMAN - 1995')
map.save('plot_data2.html')
HTML('<iframe src=plot_data2.html width=540 height=350></iframe>')
```

Out[62]:

ot found.

Wow! Here we can see how North and South America, China, Australia, North and South Africa as well as other countires in Europe have experienced this decrease in feritility. I find comparing these two maps very interesting!

These maps prove that the majority of European countries have seen a similar trend, and I am wondering if it would be interesting to compare the Italian and European trends. Let's do it.

In [34]:

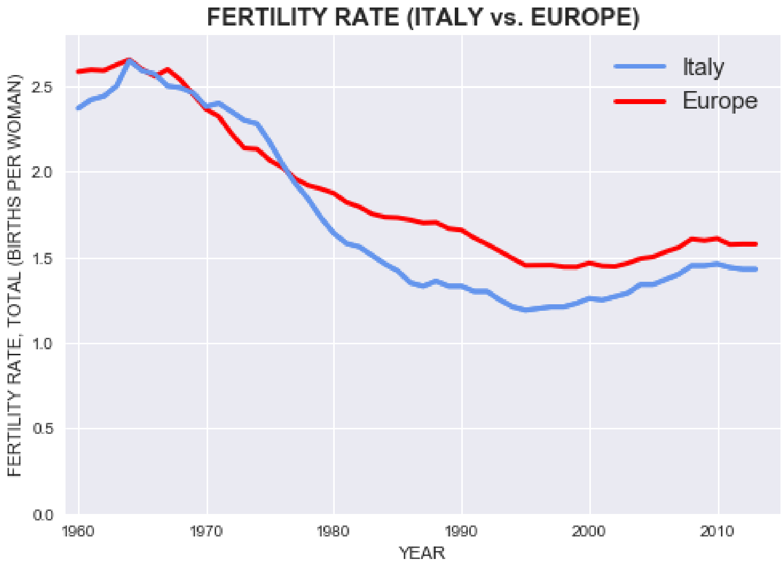
```
mask_europe = my_df['CountryName'].str.contains('European Union')
```

```
fertility_europe = my_df[mask_1 & mask_europe]
print(len(fertility_europe))
print(len(fertility_italy))
```

```
54
54
```

In [35]:

```
style.use('seaborn')
plt.plot(fertility_italy['Year'].values, fertility_italy['Value'].values, color = 'cornflowerblue',
linewidth=3)
plt.plot(fertility_europe['Year'].values, fertility_europe['Value'].values, color = 'red', linewidt
h=3)

plt.xlabel('YEAR')
plt.ylabel(fertility_italy['IndicatorName'].iloc[0].upper())

plt.title('FERTILITY RATE (ITALY vs. EUROPE)', fontweight="bold", fontsize=15)
# We adjust the axis to provide a correct data visualisation
plt.axis([1959, 2015,0,2.8])

plt.plot(fertility_italy['Year'].values, fertility_italy['Value'].values, color = 'cornflowerblue',
linewidth=3)

plt.gca().legend(('Italy','Europe'), fontsize= 15)
plt.grid(True)
plt.show()
```

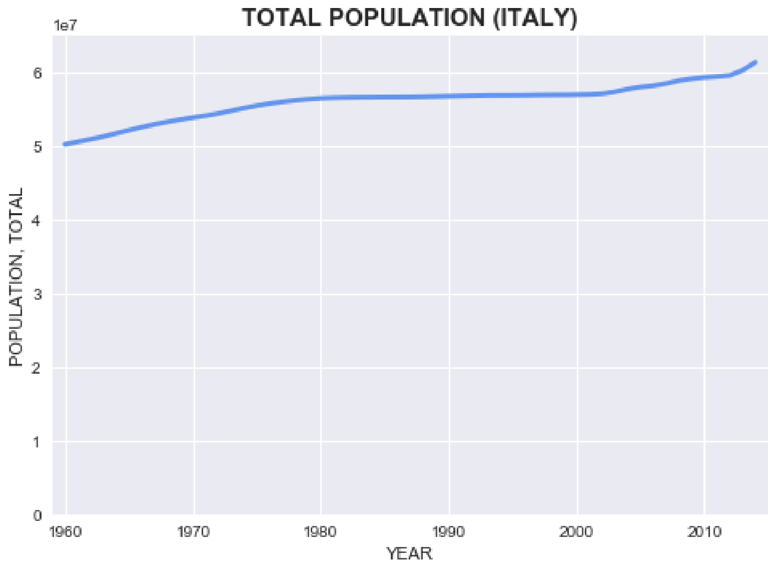


As we can see, these trends are very similar but Italy is underperforming Europe.

Has this decrease in fertility had an impact the Italian population growth?

In [52]:

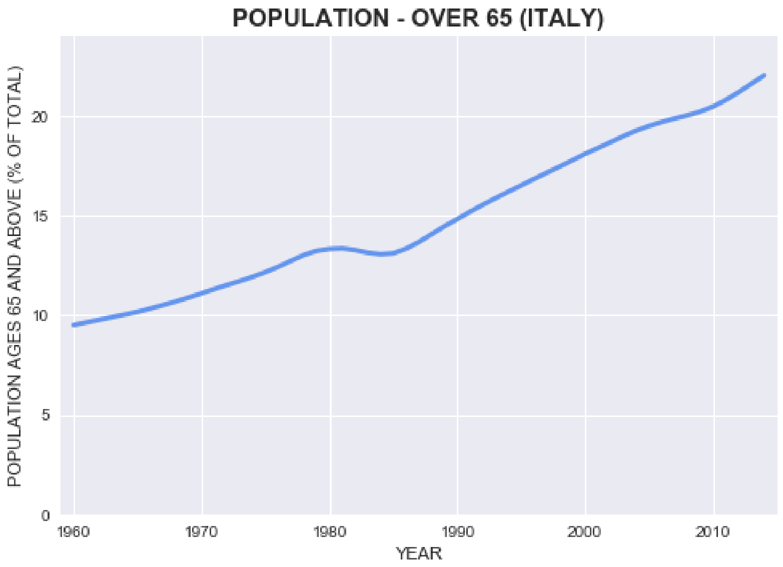
```
mask_6 = my_df['IndicatorName'].str.contains('Population, total')
population = my_df[mask_2 & mask_6]

style.use('seaborn')
plt.plot(population['Year'].values, population['Value'].values, color = 'cornflowerblue', linewidth
=3)

plt.xlabel('YEAR')

plt.ylabel(population['IndicatorName'].iloc[0].upper())
plt.title('TOTAL POPULATION (ITALY)', fontweight="bold", fontsize = 15)

plt.axis([1959, 2015,0,65000000])
plt.grid(True)

plt.show()
```



The population is growing, at a low rate. I expected a decrease. This can only be explained by an increase in immigration or an increase in the aged population (over 65). Here I will check if the older population in Italy is increasing.

In [37]:

```
mask_7 = my_df['IndicatorName'].str.contains('Population ages 65 and above \(% o')
population = my_df[mask_2 & mask_7]

style.use('seaborn')
plt.plot(population['Year'].values, population['Value'].values, color = 'cornflowerblue', linewidth
=3)
plt.xlabel('YEAR')
plt.ylabel(population['IndicatorName'].iloc[0].upper())

plt.title('POPULATION - OVER 65 (ITALY)', fontweight="bold", fontsize=15)
plt.axis([1959, 2015,0,24])
plt.grid(True)
plt.show()
```



As we can see the population over 65 keeps increasing.

# Research question

As we have seen so far, Italy has experienced a fall in fertility, my research question is: **how can we explain this phenomenon?** In order to answer this question we will explore other indicators and try to see if this trend is due to:

- **Economical factors**, such as a higher unemployment rate or a lower GDP. In fact, a loss of economical power can be correlated to the decision of not having a child or having less children.
- The **entry of women into the extra-domestic labor force**, a global trend that can have an impact on women's deciosions in Italy.

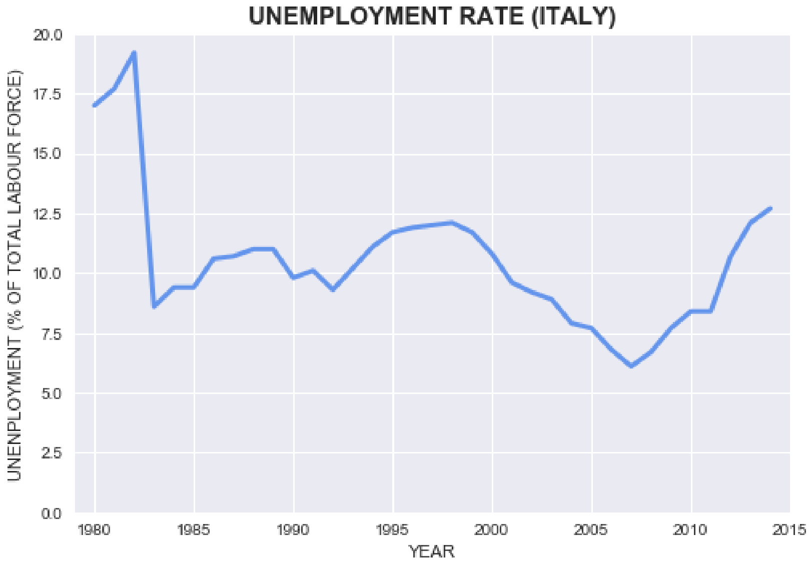
We will check the relationship between the **fertility rate** and the **unemployment** and **grdp rate**.
For this purpose I will use use **scatter plots** to visualise this relationship and **statistical correlation** to measure it.

# Fertility & Unemployment Rate correlation

Is it possible that the Italian unemployemnt rate is correlated to a low fertility trend?

In [38]:

```
mask_8 = my_df['IndicatorName'].str.contains('Unemployment, total (% of total labor force) (nationa
l estimate)', regex = False)
unemployment = my_df[mask_8 & mask_2]

style.use('seaborn')
plt.plot(unemployment['Year'].values, unemployment['Value'].values, color = 'cornflowerblue', linew
idth=3)

plt.xlabel('YEAR')
plt.ylabel('UNENPLOYMENT (% OF TOTAL LABOUR FORCE)')
plt.title('UNEMPLOYMENT RATE (ITALY)', fontweight="bold", fontsize=15)
plt.axis([1979, 2015,0,20])
plt.grid(True)
plt.show()
```



The rate drops and then has some ups and downs.
The year range for this plot is different compared with the fertility plot. We will check this difference and adjust it so that we can draw a scattaerplot. In fact **scatterplots need equal length arrays to compare**.

In [39]:

```
# Current year range?
print("FERTILITY:\tMIN YEAR:", fertility_italy['Year'].min(), "\tMAX YEAR:", fertility_italy['Year'
].max())
print("UNEMPLOYMENT:\tMIN YEAR:", unemployment['Year'].min(), "\tMAX YEAR:", unemployment['Year'].m
ax())
```

```
FERTILITY: MIN YEAR: 1960  MAX YEAR: 2013
UNEMPLOYMENT: MIN YEAR: 1980  MAX YEAR: 2014
```

In [40]:

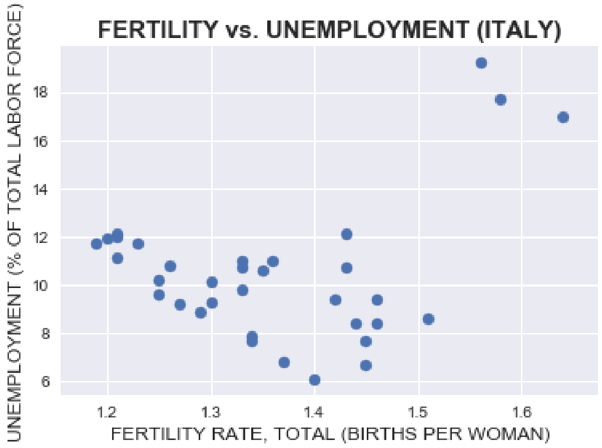
```
# We need to cut 19 years (1960-1979) of fertility data, and 1 year (2014) for unemployment.
```

```
unemployment_cut = unemployment[unemployment['Year'] < 2014]
fertility_cut = fertility_italy[fertility_italy['Year'] > 1979]

# Let's check the result.
print("FERTILITY:\tMIN YEAR:", fertility_cut['Year'].min(), "\tMAX YEAR:", fertility_cut['Year'].ma
x())
print("UNEMPLOYMENT:\tMIN YEAR:", unemployment_cut['Year'].min(), "\tMAX YEAR:", unemployment_cut['
Year'].max())

# Let's also check if we have the same amount of data for the scatterplot.
print(len(fertility_cut))
print(len(unemployment_cut))
```

```
FERTILITY: MIN YEAR: 1980  MAX YEAR: 2013
UNEMPLOYMENT: MIN YEAR: 1980  MAX YEAR: 2013
34
34
```

In [41]:

```
%matplotlib inline
fig, axis = plt.subplots()

axis.yaxis.grid(True)
axis.set_title('FERTILITY vs. UNEMPLOYMENT (ITALY)',fontweight = 'bold', fontsize=15)
axis.set_xlabel(fertility_cut['IndicatorName'].iloc[0].upper(),fontsize=12)
axis.set_ylabel('UNEMPLOYMENT (% OF TOTAL LABOR FORCE)',fontsize=12)
x = fertility_cut['Value']
y = unemployment_cut['Value']

axis.scatter(x, y)
plt.show()
```



It looks like these indicators are not correlated. I will test it by checking their statistical correlation.

In [42]:

```
np.corrcoef(fertility_cut['Value'],unemployment_cut['Value'])
```

Out[42]:

```
array([[1.        , 0.27296815],
       [0.27296815, 1.        ]])
```

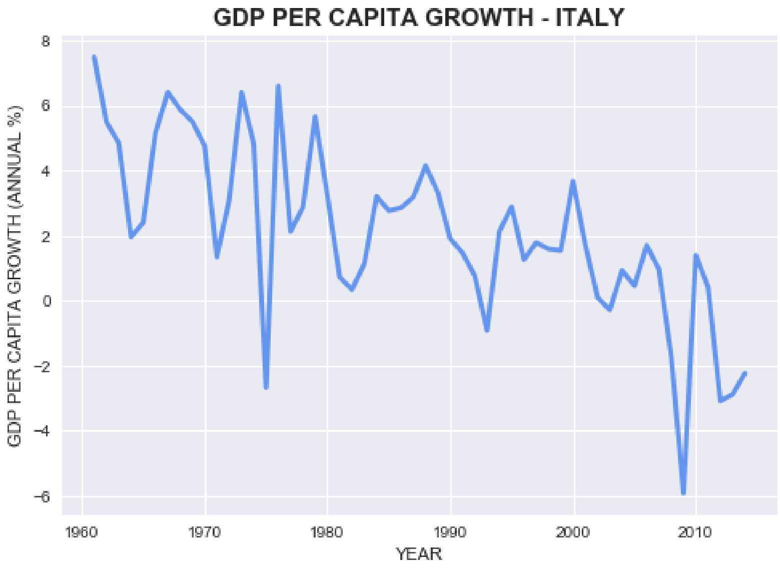
As we can see **0.27** is a **weak positive linear relationship**. So, we can conclude the path of the italian unemployment rate is not associated with the women's decision of having babies.
What about the Gdp ratio? Is it possible that the country is experiencing a decrese in Gdp, which is the result of a fragile economy, and this is having an impact on womens' decisions? Let's check it out.

# Fertitlity and Gdp per capita

We follow the same procedure as before.

In [43]:

```
mask_9 = my_df['IndicatorName'].str.contains('GDP per capita growth \(ann')
gdp = my_df[mask_2 & mask_9]

style.use('seaborn')
plt.plot(gdp['Year'].values, gdp['Value'].values, color = 'cornflowerblue', linewidth=3)

plt.xlabel('YEAR')
plt.ylabel(gdp['IndicatorName'].iloc[0].upper())

plt.title('GDP PER CAPITA GROWTH - ITALY', fontweight="bold", fontsize = 15)

plt.grid(True)
plt.show()
```



As we can see the Italian Gdp was growing at 8% in 1960, and since then it has followed a downtrend that has taken it to -2% in 2015. Let's check the correlation between the Gdp downtrend (constantly pointing downwards) and the fertility downtrend (a quick drop followed by a steady path).

In [44]:

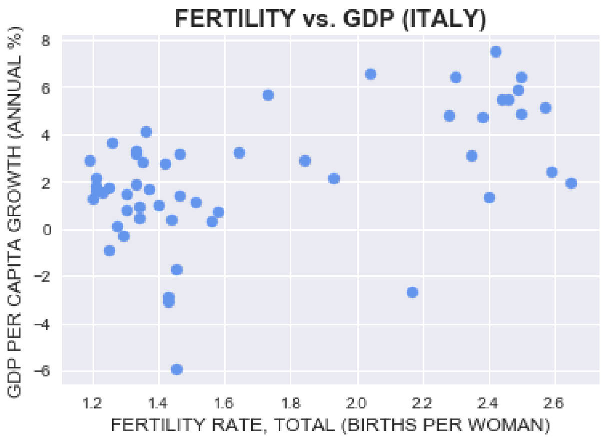
```
# Year range?
print('BEFORE')
print("FERTILITY:\tMIN YEAR:", fertility_italy['Year'].min(), "\tMAX YEAR:", fertility_italy['Year'
].max())
print("GDP:\t\tMIN YEAR:", gdp['Year'].min(), "\tMAX YEAR:", gdp['Year'].max())

# We need to cut the year 1960 from fertility and 2014 from gdp.
fertility_cut2 = fertility_italy[fertility_italy['Year'] > 1960]
gdp_cut = gdp[gdp['Year'] < 2014]
print()
print('AFTER')
print("FERTILITY:\tMIN YEAR:", fertility_cut2['Year'].min(), "\tMAX YEAR:", fertility_cut2['Year'].
max())
print("GDP:\t\tMIN YEAR:", gdp_cut['Year'].min(), "\tMAX YEAR:", gdp_cut['Year'].max())
print()

# Data Length.
print(len(fertility_cut2))
print(len(gdp_cut))
```

```
BEFORE
FERTILITY: MIN YEAR: 1960  MAX YEAR: 2013
GDP:  MIN YEAR: 1961  MAX YEAR: 2014

AFTER
FERTILITY: MIN YEAR: 1961  MAX YEAR: 2013
GDP:  MIN YEAR: 1961  MAX YEAR: 2013
```

```
GDP:  MIN YEAR: 1961  MAX YEAR: 2013

53
53
```

In [45]:

```
%matplotlib inline
fig, axis = plt.subplots()

axis.yaxis.grid(True)
axis.set_title('FERTILITY vs. GDP (ITALY)',fontweight = 'bold', fontsize=15)
axis.set_xlabel(fertility_cut2['IndicatorName'].iloc[0].upper(),fontsize=12)
axis.set_ylabel(gdp_cut['IndicatorName'].iloc[0].upper(),fontsize=12)

x = fertility_cut2['Value']
y = gdp_cut['Value']

axis.scatter(x, y, color='cornflowerblue')
plt.show()
```



In [46]:

```
np.corrcoef(fertility_cut2['Value'],gdp_cut['Value'])
```

Out[46]:

```
array([[1.       , 0.5221686],
       [0.5221686, 1.       ]])
```

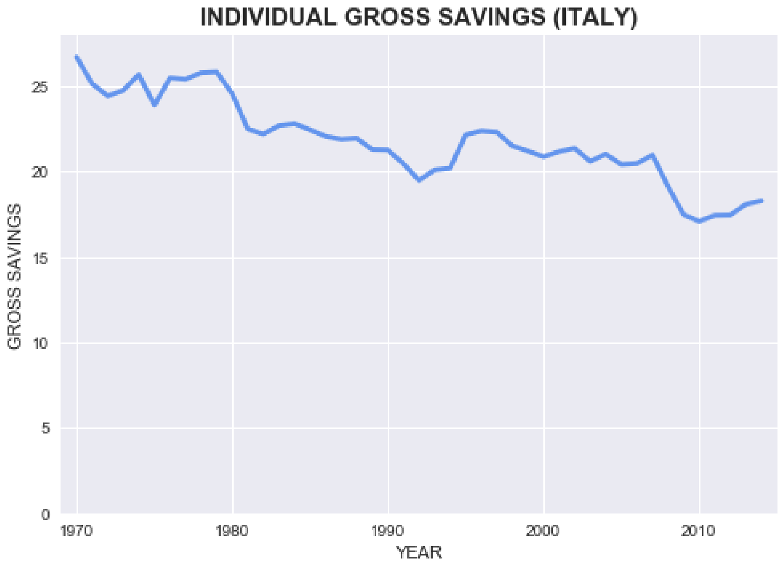
Once again, both the scatter plot and the statistical analysis **do not show any correlation between the Gdp and the fertility rate** . I am questioning if the Gdp Growth is a valid indicator to understand if an individual is in a solid financial situation, which would facilitate the decision of having babies. Probably not. In fact, other factors, such as the inflation rate, the average individual salary (etc...) are essential to better understand it.

For this reason, I think that we can check a better indicator, such as the **gross savings**, which will tell us how much individuals can save once that all the cost of living have been discounted.

# Fertitlity and Gross Savings

In [47]:

```
mask_10 = my_df['IndicatorName'].str.contains('Gross savings \(% of GN')
gross_savings = my_df[mask_2 & mask_10]

style.use('seaborn')
plt.plot(gross_savings['Year'].values, gross_savings['Value'].values, color = 'cornflowerblue', lin
ewidth=3)

plt.xlabel('YEAR')
plt.ylabel('GROSS SAVINGS')
plt.axis([1969, 2015,0,28])
plt.title('INDIVIDUAL GROSS SAVINGS (ITALY)', fontweight="bold", fontsize = 15)

plt.grid(True)
plt.show()
```



In [48]:

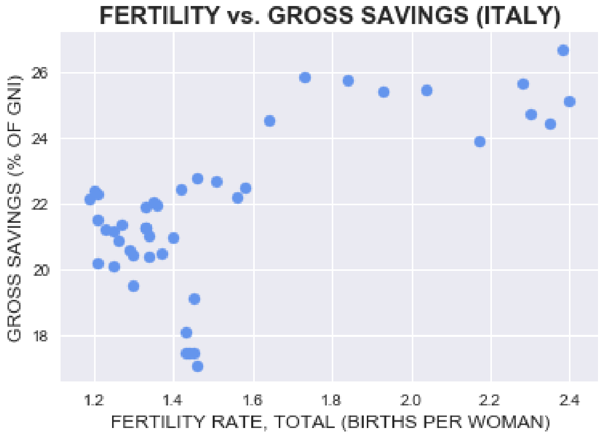
```
# Year range?
print('BEFORE')
print("FERTILITY:\tMIN YEAR:", fertility_italy['Year'].min(), "\tMAX YEAR:", fertility_italy['Year'
].max())
print("SAVINGS:\tMIN YEAR:", gross_savings['Year'].min(), "\tMAX YEAR:", gross_savings['Year'].max(
))

# We need to cut 9 years (1960 - 1969) from fertility and 2014 from savings
fertility_cut3 = fertility_italy[fertility_italy['Year'] > 1969]
gross_savings_cut = gross_savings[gross_savings['Year'] < 2014]
print()
print('AFTER')
print("FERTILITY:\tMIN YEAR:", fertility_cut3['Year'].min(), "\tMAX YEAR:", fertility_cut3['Year'].
max())
print("SAVINGS:\tMIN YEAR:", gross_savings_cut['Year'].min(), "\tMAX YEAR:",
gross_savings_cut['Year'].max())
print()

# Data length
print(len(fertility_cut3))
print(len(gross_savings_cut))
```

```
BEFORE
FERTILITY: MIN YEAR: 1960   MAX YEAR: 2013
SAVINGS: MIN YEAR: 1970   MAX YEAR: 2014

AFTER
FERTILITY: MIN YEAR: 1970   MAX YEAR: 2013
SAVINGS: MIN YEAR: 1970   MAX YEAR: 2013

44
44
```

In [45]:

```
%matplotlib inline
fig, axis = plt.subplots()

axis.yaxis.grid(True)
axis.set_title('FERTILITY vs. GROSS SAVINGS (ITALY)',fontsize=15, fontweight = 'bold')
axis.set_xlabel(fertility_cut3['IndicatorName'].iloc[0].upper(),fontsize=12)
axis.set_ylabel(gross_savings_cut['IndicatorName'].iloc[0].upper(),fontsize=12)

x = fertility_cut3['Value']
y = gross_savings_cut['Value']
```

```
y = gross_savings_cut['Value']

axis.scatter(x, y, color='cornflowerblue')
plt.show()
```



In [55]:

```
np.corrcoef(fertility_cut3['Value'],gross_savings_cut['Value'])
```

Out[55]:

```
array([[1.        , 0.69003177],
       [0.69003177, 1.        ]])
```

A coefficient of correlation of approximately 0.70 indicates that **there is an overall correlation between the fertility rate and the downtrend in saving money in Italy**. This correlation **is not perfect**, but it is **strong enough to be acknowledged** and **researched further**.

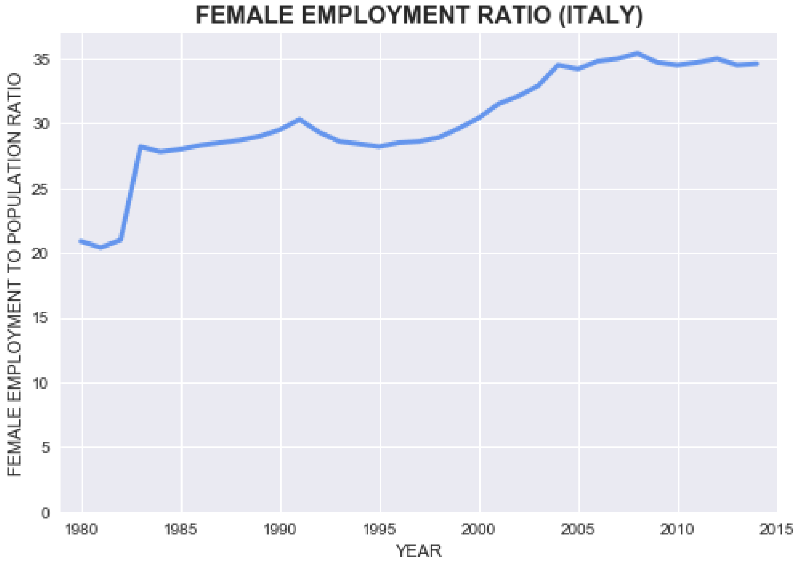
As we can see after a couple of researches we are going on the right direction.
When we look at the ability of saving money in Italy, we can see that people are saving less and less money. This can be due to many factors: increased cost of living, increased property price, decreased salary, more precarious contracts, etc..
This sense of financial instability can potentially force women to postpone the decision of having a child, at least until they have achieved a better financial situation. This could be a feasable conclusion, but, as we said, further and more detailed investigations are required.

Finally, let's check if the global trend of an increase female labour force is correlated with a lower fertility.

# d) Fertitlity and female labor force rate

In [56]:

```
# In this case I will use the indicator code for female employment to population ratio.

mask_11 = my_df['IndicatorCode'].str.contains('SL.EMP.TOTL.SP.FE.NE.ZS')

labour = my_df[mask_2 & mask_11]

style.use('seaborn')

plt.plot(labour['Year'].values, labour['Value'].values, color = 'cornflowerblue', linewidth=3)

plt.xlabel('YEAR')

plt.ylabel('FEMALE EMPLOYMENT TO POPULATION RATIO')

plt.axis([1979, 2015,0,37])

plt.title('FEMALE EMPLOYMENT RATIO (ITALY)', fontsize = 15, fontweight="bold")

plt.grid(True)

plt.show()
```



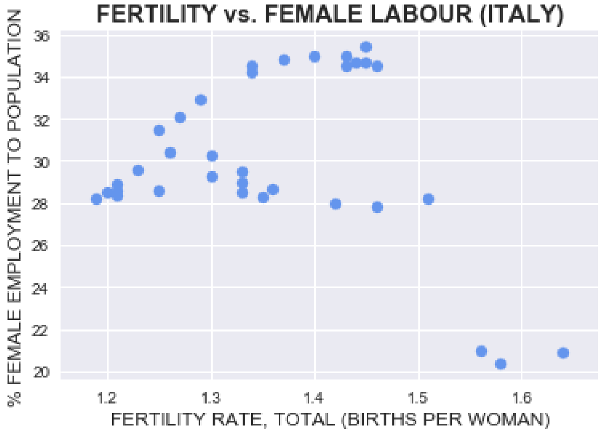
In [57]:

```
# Year range?
print('BEFORE')
print("FERTILITY:\tMIN YEAR:", fertility_italy['Year'].min(), "\tMAX YEAR:", fertility_italy['Year'
].max())
print("LABOUR:\t\tMIN YEAR:", labour['Year'].min(), "\tMAX YEAR:", labour['Year'].max())

# We need to cut 19 years (1960 - 1979) from fertility and 2014 from labour.
fertility_cut4 = fertility_italy[fertility_italy['Year'] > 1979]
labour_cut = labour[labour['Year'] < 2014]
print()
print('AFTER')
print("FERTILITY:\tMIN YEAR:", fertility_cut4['Year'].min(), "\tMAX YEAR:", fertility_cut4['Year'].
max())
print("SAVINGS:\tMIN YEAR:", labour_cut['Year'].min(), "\tMAX YEAR:", labour_cut['Year'].max())
print()

# Data length
print(len(fertility_cut4))
print(len(labour_cut))
```

```
BEFORE
FERTILITY: MIN YEAR: 1960  MAX YEAR: 2013
LABOUR:   MIN YEAR: 1980  MAX YEAR: 2014

AFTER
FERTILITY: MIN YEAR: 1980  MAX YEAR: 2013
SAVINGS: MIN YEAR: 1980  MAX YEAR: 2013

34
34
```

In [58]:

```
%matplotlib inline
fig, axis = plt.subplots()

axis.yaxis.grid(True)

axis.set_title('FERTILITY vs. FEMALE LABOUR (ITALY)',fontsize=15, fontweight = 'bold')

axis.set_xlabel(fertility_cut4['IndicatorName'].iloc[0].upper(),fontsize=12)

axis.set_ylabel('% FEMALE EMPLOYMENT TO POPULATION'.upper(),fontsize=12)

x = fertility_cut4['Value']
y = labour_cut['Value']

axis.scatter(x, y, color='cornflowerblue')

plt.show()
```



```
In [59]:
```

```
np.corrcoef(fertility_cut4['Value'],labour_cut['Value'])
```

```
Out[59]:
```

```
array([[ 1.        , -0.24026599],
       [-0.24026599,  1.        ]])
```

There is not a statistical correlation between the increase in female employment and a decrease in female fertitlity.

# Conclusion

My research has allowed me to come to an understanding that **a low fertitily rate**, is a phenomenon which is more spread out on the planet than I expected.
In order to analyse the Italian situation, I had to compare it to the global fertility trend, and this provided an insight on the world scenario.
Since the 60s, Europe has been one of the areas with the lowest rate, and later on China, North and South America, North and South Africa, Australia and other countries followed this trend.
Italy has one of the lowest fertility rate on the planet.
Thanks to the exploration of our dataset we came to the conclusiopn that:

- The Italian unemployment rate is not correlated to the decision of postponing a maternity.
- The Gdp growth is not a good indicator because it is too generalised.
- A reduced ability to save money is correlated with a lower fertility rate and should be further investigated. Italy present great regional differences, and it would be interesting to outsurce a more detailed dataset to investigate this trend in different regions.
- An increase in female employment is surprisingly not correlated with the decision or not of not having babies.

This research has shown the path to follow to further understand why a lower fertility rate is affected by the italian economy. In fact, this could be caused by:

- A rigid labor market.
- A limited set of employment contracts.
- An inadequate supply of public childcare.
- An increase in divorce.

Basically, the next step will be going beyond aggregate country comparisons and research more carefully changing fertility behavior within Italy.