World Bank - data vizualization key insights about the internet use and economical aspects

Abstract

My project on data vizualization focuses on the correlation between multiple different aspects of coutntries. I believe that big data gathered from different countries may shed light of different mechanisms that link aspects of our daily lives with macro indicators. Data vizualization is currently widely use to create an image which can give researchers from different domain to talk abou important problems that our society is facing, to find, define and evaluate the problems and to create solution. The main focus of this work is around the literacy levels and the internet use: their correlation, connection with other economical aspects and socail implications of the findings.

Introduction

The work done in this field along with the rich datasets collected each year by the World Bank give researchers important information on which they can build solutions. In order to create a better society we need to identify, separate and define the existing problems. This requires interdisciplinary teams that can communicate with each other and for that is mandatory to have data vizualization. Without means to express impact, damage, sizes and numbers the teams of specialists with different backgrouns can not have a clear overview on the current situation thus being unable to take a stand.

Dataset

The World Bank dataset is an organization that coordinates statistical and data work and maintains a number of macro, financial and sector databases. Working closely with the Bank's regions and Global Practices, the group is guided by professional standards in the collection, compilation and dissemination of data to ensure that all data users can have confidence in the quality and integrity of the data produced. Each year, they provide multiple datasets that are then used to evaluate and analyze different aspects from an economical and cultural perspective that are general and influence day to day lives.

Existing work

Directions of interests can be analyzed using the macro dataset. For example, in the context of climate change and concerns about fossil fuels, territories around the world are being remapped for their renewable energy generation potential. The World Bank and other

institutions dominated by the global North are urging countries, especially in the global South, to undertake such mapping. The resulting maps are shown to potential investors in efforts to accelerate and direct the rapidly growing flow of capital into the renewable energy sector[1]. These representations of territory and the new patterns of investment and land use they facilitate and foreshadow engage core concerns of political ecology: who claims, uses, and controls rural lands and resources; how are competing claims contested and legitimated; and who benefits or suffers as new visions of development, accumulation, and 'sustainability' are inscribed upon the land and new aspects of 'nature' are drawn into circuits of capital[1].

The dramatic expansion of renewable energy production from abiotic sources is an important but under-researched component of the global land rush, one that differs in key analytical ways from the agricultural and extractive sectors examined by most scholarship in that domain. Second, since powerful new visualizations are central to this expansion, there is an urgent need for closer engagement between political ecology and critical data studies to analyze their production, deployment, and effects. [1].

Taking other directions, the dataset from the World Bank can be mixed with other datasets to expand the areas in which these data can be used. For example researchers combined two dataset in order to obtain key information about to analyze the role of governmental, trade, and competitiveness considerations in the formation of official COVID-19 reports[2].

Besides the economical aspects that can derive from this dataset, key personal biases can be observed. Researchers have also analyzed life expectancy and births attended by skilled health staff [3], how iare childbirth deaths related to education and what is the impact of unemployment on a society [3].

The work done in this field along with the rich datasets collected each year by the World Bank give researchers important information on which they can build solutions. In order to create a better society we need to identify, separate and define the existing problems. This requires interdisciplinary teams that can communicate with each other and for that is mandatory to have data vizualization. Without means to express impact, damage, sizes and numbers the teams of specialists with different backgrouns can not have a clear overview on the current situation thus being unable to take a stand.

Expected outcomes and objectives of the current work

In the current work I intend to analyze the correlation between the number of people who use the internet and different economical aspects of the country. I expect that the internet use is strongly correlated with the wealth of a country denoted by GDP and exports. Also I think that the level of urbanization is impacting the number of internet users from a specific region. Last but not least I am interested in how the level o literacy is connected with the internet use: is the internet indeed mostly used for entertainment and thus the level of literacy is not related to that at all. Or contradictory, the level of literacy is correlated since the people have access to multiple information from all over the world.

Process

In [ ]:
```
import matplotlib.pyplot as plt # plotting
import numpy as np # linear algebra
import os # accessing directory structure
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import seaborn as sns
```

Reading the dataframe

For reading the dataframe I used pandas

In [45]:
```
df = pd.read_csv('data.csv', thousands=',')
```

Visualizing the first values from the dataset.

In [54]:
```
df.head()
```

Out[54]:

	Country Name	Region Code	Country Code	GDP, PPP (current international $)	Population, total	Population CGR 1960-2015	Internet users (per 100 people)	Popltn Largest City % of Urban Pop	2014 Lif expectanc at birtl tota (years
0	Aruba	MA	ABW	NaN	103889.0	1.19%	88.7	NaN	75.
1	Andorra	EU	AND	NaN	70473.0	3.06%	96.9	NaN	Na
2	Afghanistan	ME	AFG	6.291267e+10	32526562.0	2.36%	8.3	53.4%	60.
3	Angola	AF	AGO	1.844377e+11	25021974.0	2.87%	12.4	50.0%	52.
4	Albania	EU	ALB	3.266324e+10	2889167.0	1.07%	63.3	27.3%	77.

Printing the columns from the dataset to have their complete name

In [61]:
```
df.columns
```

Out[61]:
```
Index(['Country Name', 'Region Code', 'Country Code',
       'GDP, PPP (current international $)', ' Population, total ',
       'Population CGR 1960-2015', 'Internet users (per 100 people)',
       'Popltn Largest City % of Urban Pop',
       '2014 Life expectancy at birth, total (years)',
       'Literacy rate, adult female (% of females ages 15 and above)',
       'Exports of goods and services (% of GDP)'],
      dtype='object')
```

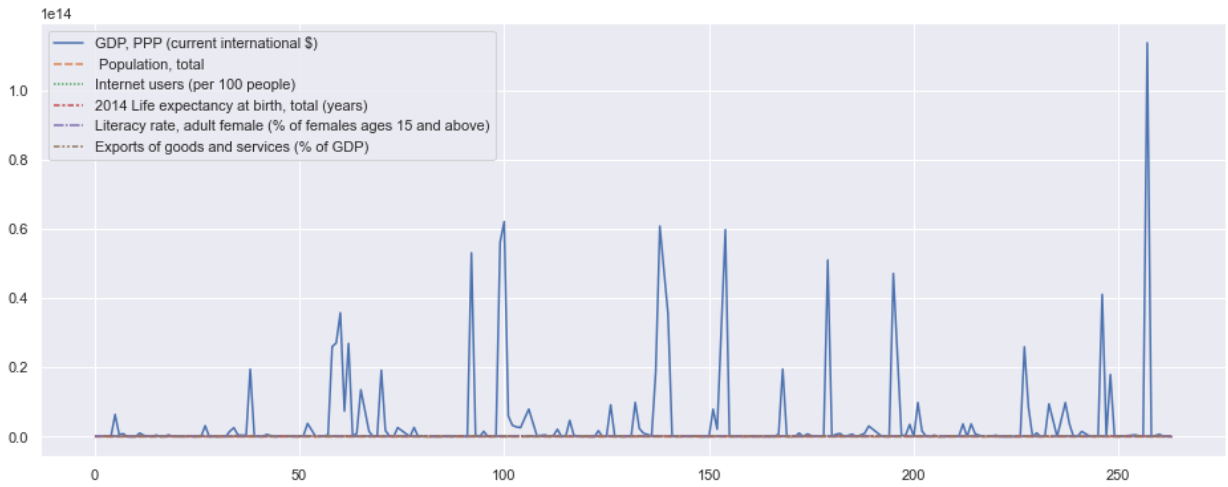
Overall look and conclusions on the dataframe

Plotting all the data to see the range of values in different columns. As can be seen from below the GDP column hav very high values, making the graph unreadable for the other values.

The GDP ranges from 0 to 1e14

In [57]:
```
plt.figure(figsize=(16,6))
sns.lineplot(data=df)
```

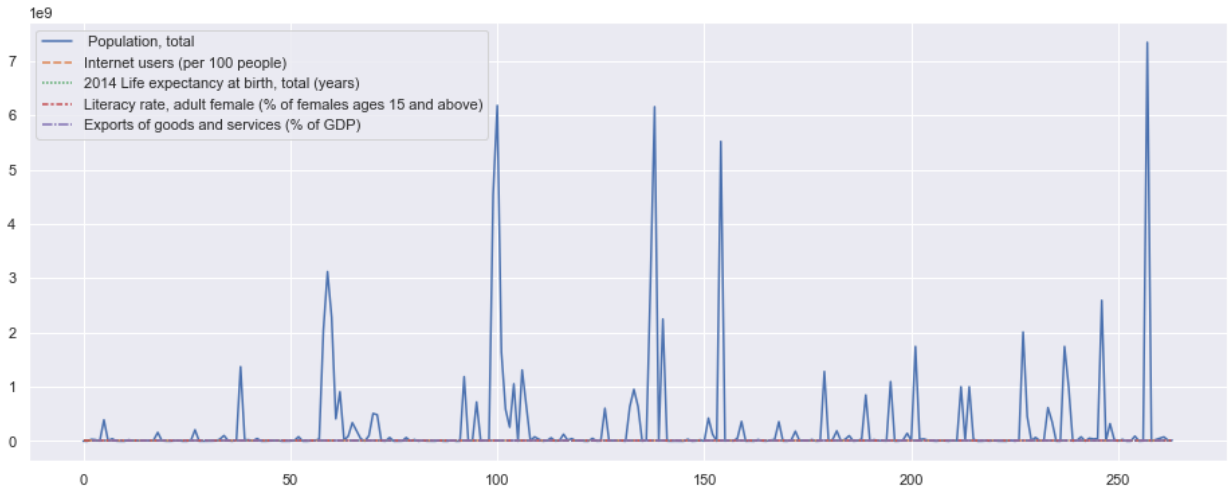
Out[57]:
```
<AxesSubplot:>
```



In order to have a more accurate vizualization of the data I plotted all the values except the GDP column mentioned above. It can be seen that again, there is a column that takes the scene which is the population total column, with values in range 0 to 7e9

In [59]:
```
plt.figure(figsize=(16,6))
sns.lineplot(data=df.drop(columns=["GDP, PPP (current international $)"]))
```

Out[59]:
```
<AxesSubplot:>
```

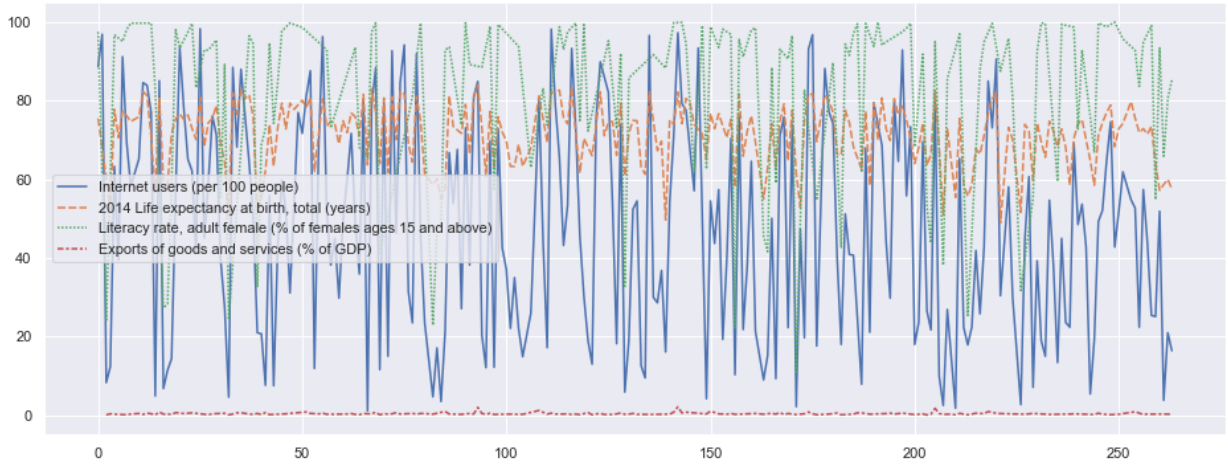


The rest of the numerical values are between 0 and 100. Each of them varies in different ranges. The column that varies the most is the internet users.

The only column not readable from the is the exports of goods and services.

In [62]:
```
plt.figure(figsize=(16,6))
sns.lineplot(data=df.drop(columns=["GDP, PPP (current international $)", ' Population,
```

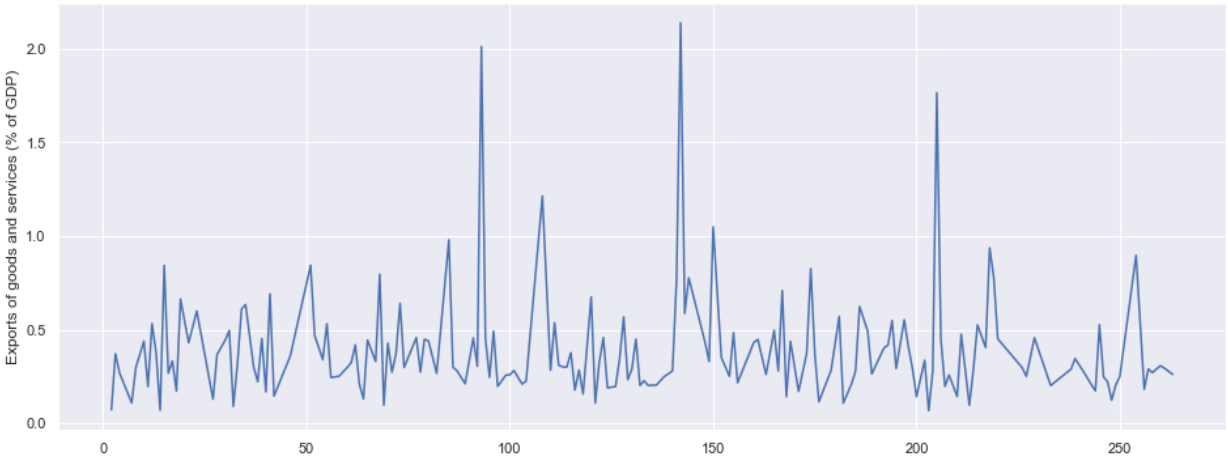
Out[62]:
```
<AxesSubplot:>
```



The exports of good and services in this dataframe varies between 0 and 2.5

In [63]:
```
plt.figure(figsize=(16,6))
sns.lineplot(data=df['Exports of goods and services (% of GDP)'])
```

Out[63]:
```
<AxesSubplot:ylabel='Exports of goods and services (% of GDP)'>
```

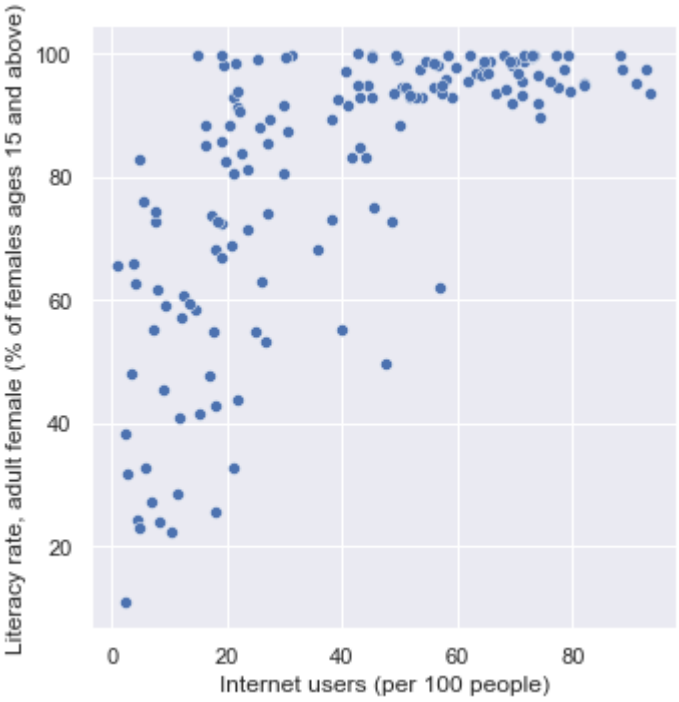


Dive into the data

Find correlations between the literacy rates in adult femals and other indicators for a certain country

Below I looked at the correlation between internet users and literacy rates per 100 people. As can be seen from the graph below, the literacy rate is strongly correlated with the number of internet users. It seems that the more people are using internet the higher the literacy rate in a country

In [69]:
```
sns.relplot(x="Internet users (per 100 people)", y="Literacy rate, adult female (% of
```

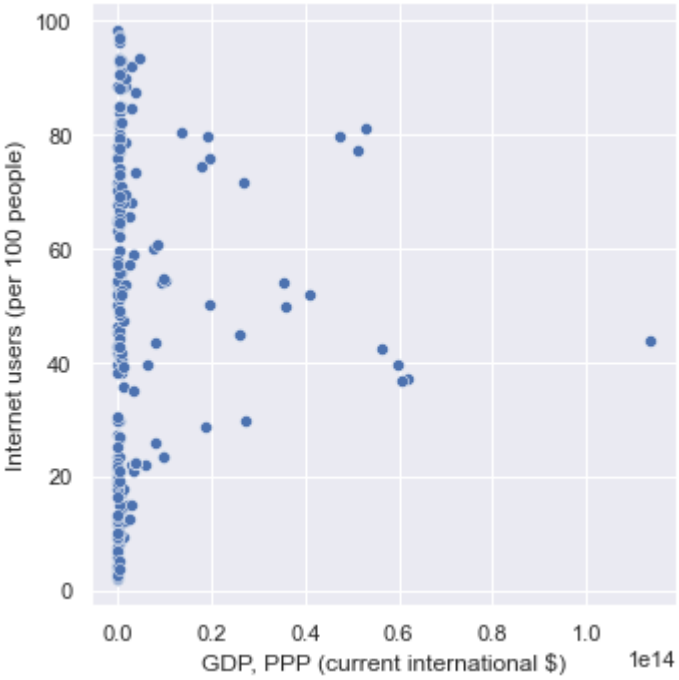


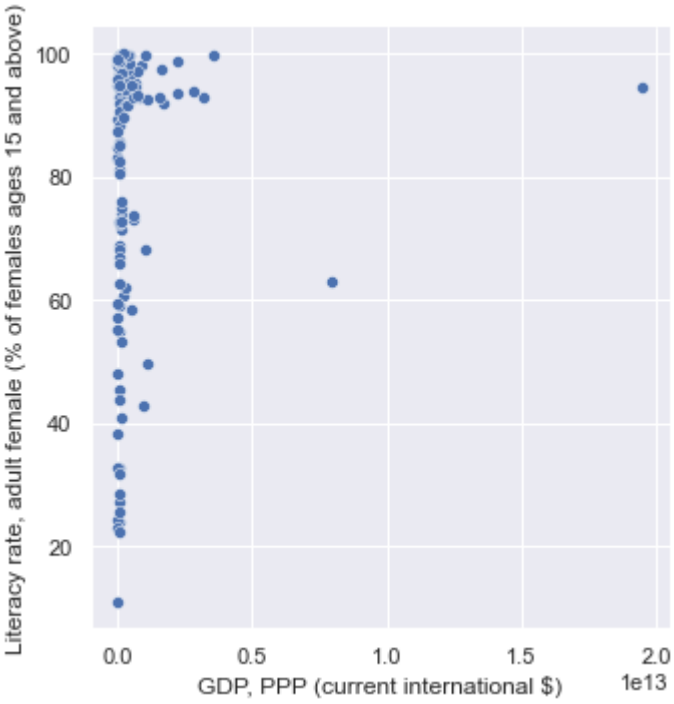
But the graph from above may be a coincidence and actually the literacy rate is correlated with how rich a country is and the internet users is also correlated with the wealth of the country. I decided to evaluate a country richness based on the GDP. Below I plotted the graph between internet use and GDP and literacy rates and gdp

In [75]:
```
plt.figure(figsize=(16,6))
sns.relplot(x="GDP, PPP (current international $)", y="Internet users (per 100 people)
sns.relplot(x="GDP, PPP (current international $)", y="Literacy rate, adult female (%
```

Out[75]:
```
<seaborn.axisgrid.FacetGrid at 0x1c4c4cc2438>

<Figure size 1152x432 with 0 Axes>
```





The figures do not say to much becuase most of the values are at low GDP in terms of bothe Internet users and literacy rate.

Create more graphics in order to zoom on smaller numbers of GDP correlated with internet users.

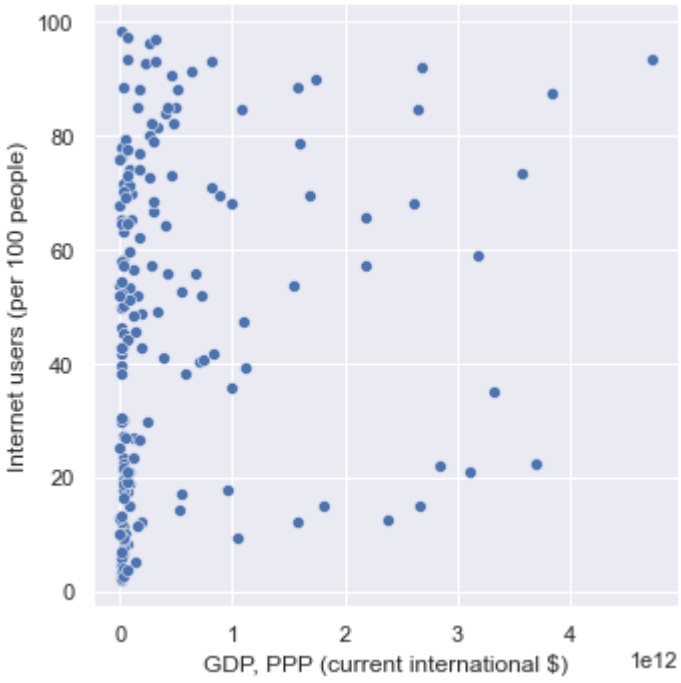
As can bee seen from the graphs below there is no correlation (or a very low correlation) between the internet use and the GDP

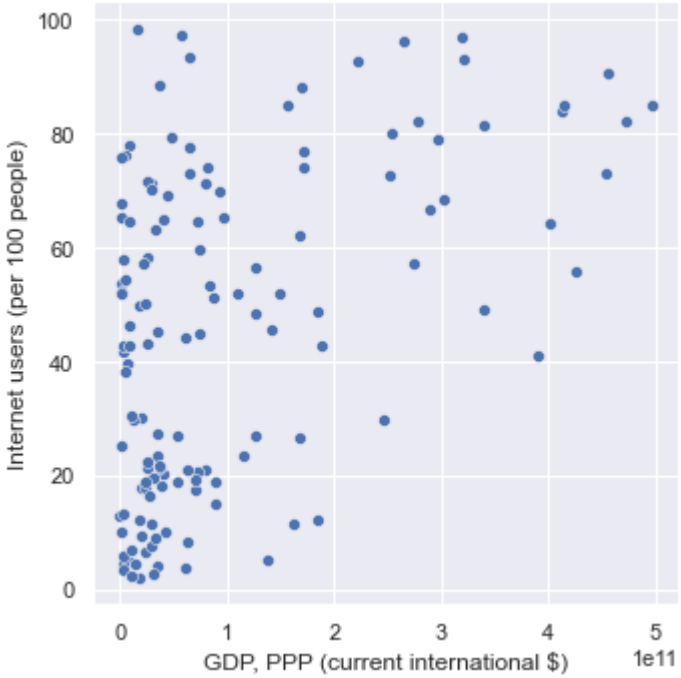
In [78]:
```
gdp = df[df["GDP, PPP (current international $)"] < 5000000000000]

sns.relplot( x="GDP, PPP (current international $)", y="Internet users (per 100 people

gdp = df[df["GDP, PPP (current international $)"] < 500000000000]

sns.relplot( x="GDP, PPP (current international $)", y="Internet users (per 100 people
```





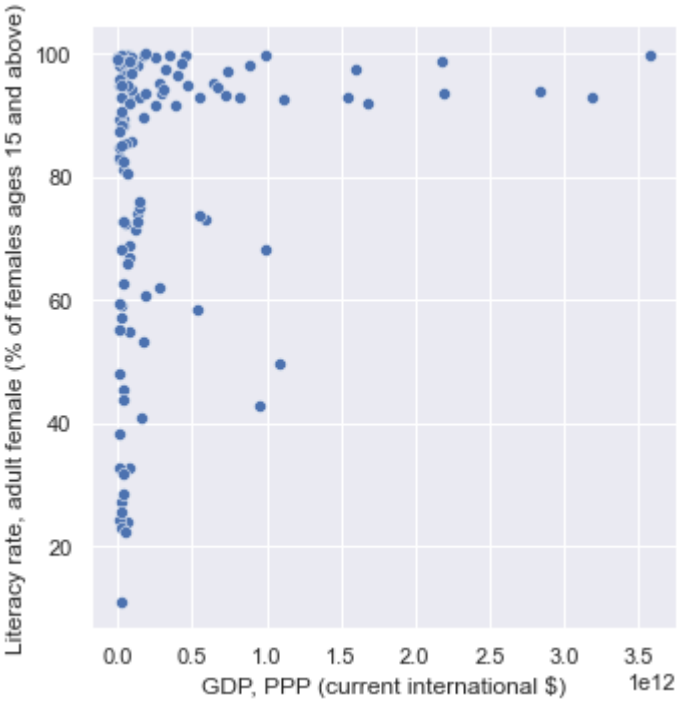
Never the less the literacy in women is correlated also with the wealth of the country as can be seen in the graphs below.

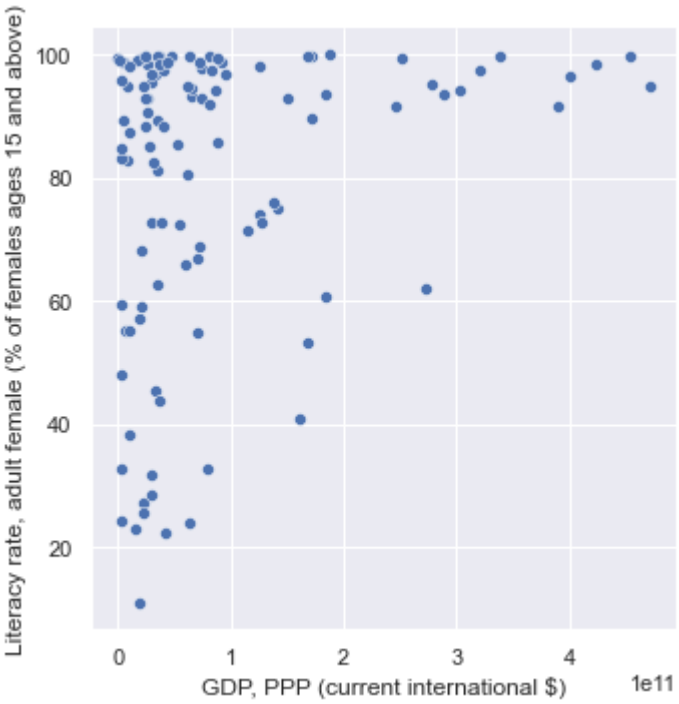
In [84]:
```
gdp = df[df["GDP, PPP (current international $)"] < 4000000000000]

sns.relplot( x="GDP, PPP (current international $)", y="Literacy rate, adult female (%

gdp = df[df["GDP, PPP (current international $)"] < 500000000000]

sns.relplot( x="GDP, PPP (current international $)", y="Literacy rate, adult female (%
```





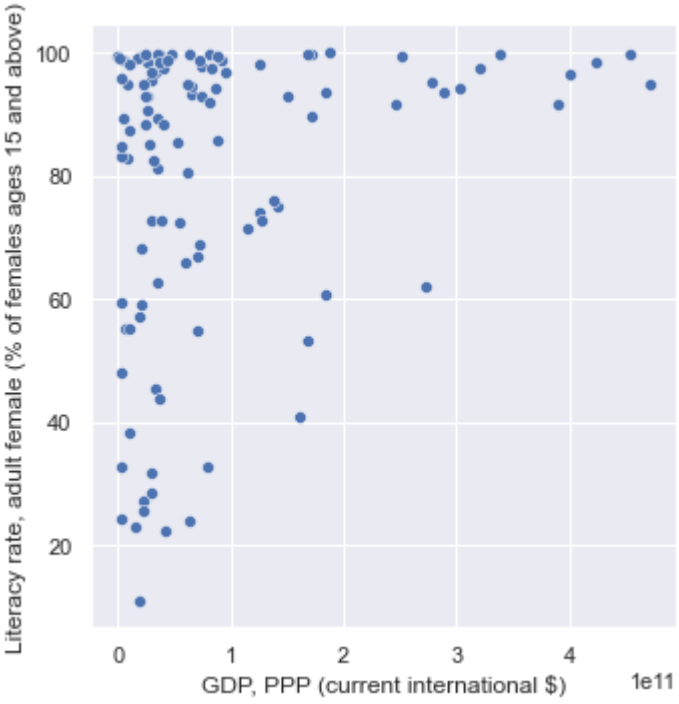
Conclusions: the literacy in the young females from the dataset is correlated with two independent factors: the wealth of the country denoted by the GDP and the internet use / 100 persons

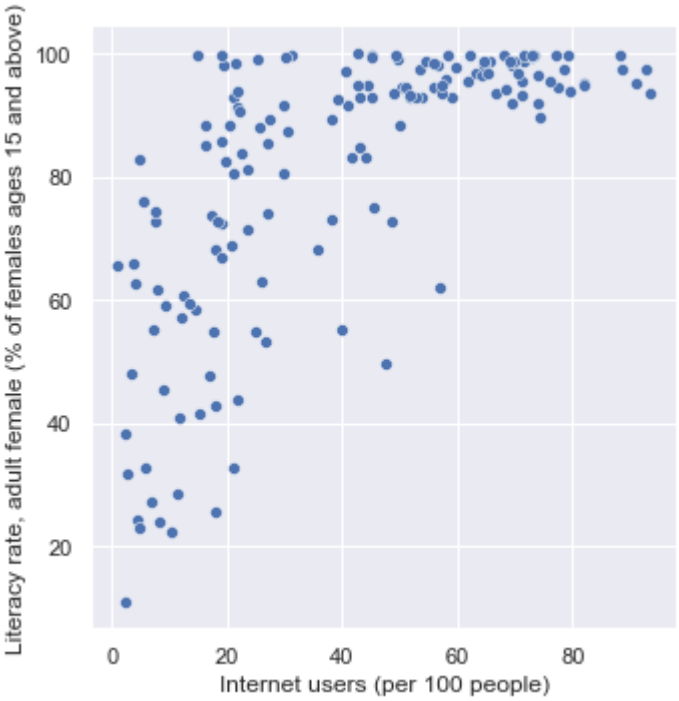
The relevant graphics of the data are presented below. It can be seen that the internet use is more relevant than the GDP when correlating with the literacy rate. Literay rate is direct proportional with both factors.

In [85]:
```
gdp = df[df["GDP, PPP (current international $)"] < 500000000000]

sns.relplot( x="GDP, PPP (current international $)", y="Literacy rate, adult female (%
```

```
sns.relplot(x="Internet users (per 100 people)", y="Literacy rate, adult female (% of
```



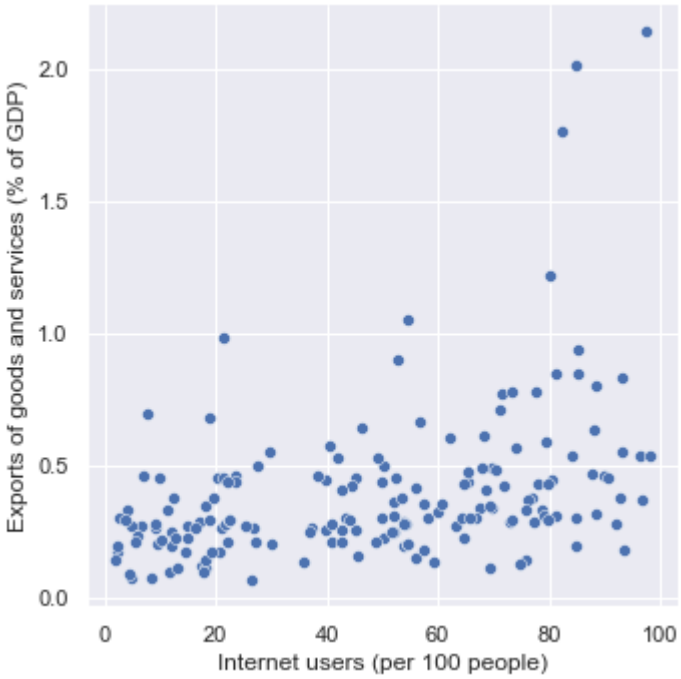


Find correlations between the internet users and other indicators for a certain country

From the above experiments I found interesting the correlation between literacy and internet users. My question became which indicators may be correlated with internat use. As can be seen above, the internet use seems to have no correlation with the wealth of a country as measured in GDP.

Below I plotted a graph between the Exports and the internet ue. There is again no visible strong correlation between the two.

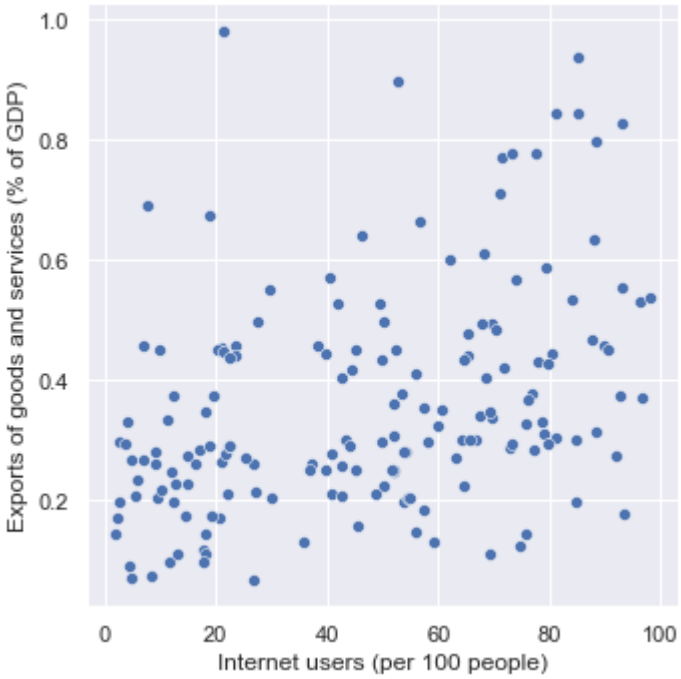
In [87]:
```
sns.relplot(x="Internet users (per 100 people)", y="Exports of goods and services (% c
```



We need to remove the outliers from the graph above in order to have a more clear view on the data

In [88]:
```
gdp = df[df["Exports of goods and services (% of GDP)"] < 1]

sns.relplot(x="Internet users (per 100 people)", y="Exports of goods and services (% c
```



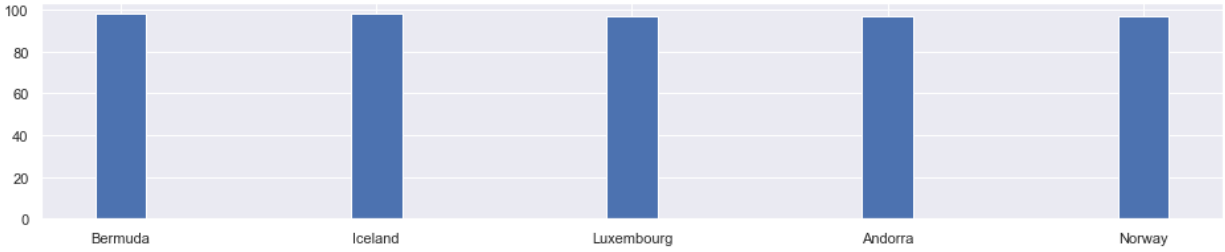
Plot comparative the number of internet users (blue) and the number of exports in goods and

services (green)

In [119…
```
df.nlargest(5, ['Internet users (per 100 people)'])
plt.figure(figsize=(16,3))
ax = plt.subplot(111)
x = x_axis = np.arange(5)
x_ticks = df.nlargest(5, ['Internet users (per 100 people)'])["Country Name"]
y =  df.nlargest(5, ['Internet users (per 100 people)'])["Internet users (per 100 peop
print(y)
z =  df.nlargest(5, ['Internet users (per 100 people)'])["Exports of goods and service
print(z)
ax.bar(x, y, width=0.2, color='b', align='center')
plt.xticks(x_axis, x_ticks)
```

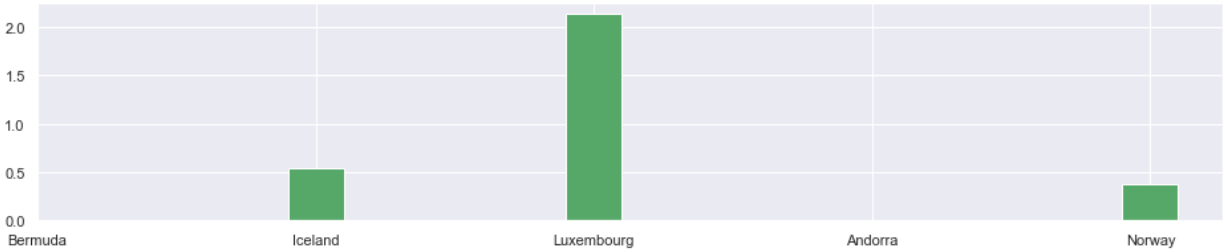
```
[98.3, 98.2, 97.3, 96.9, 96.8]
[nan, 0.537353880890662, 2.13849844087311, nan, 0.370544057103514]
```

Out[119]:
```
([<matplotlib.axis.XTick at 0x1c4cbead5f8>,
  <matplotlib.axis.XTick at 0x1c4cbeaa978>,
  <matplotlib.axis.XTick at 0x1c4cbeaa748>,
  <matplotlib.axis.XTick at 0x1c4cc03eb70>,
  <matplotlib.axis.XTick at 0x1c4cc04e1d0>],
 [Text(0, 0, 'Bermuda'),
  Text(1, 0, 'Iceland'),
  Text(2, 0, 'Luxembourg'),
  Text(3, 0, 'Andorra'),
  Text(4, 0, 'Norway')])
```



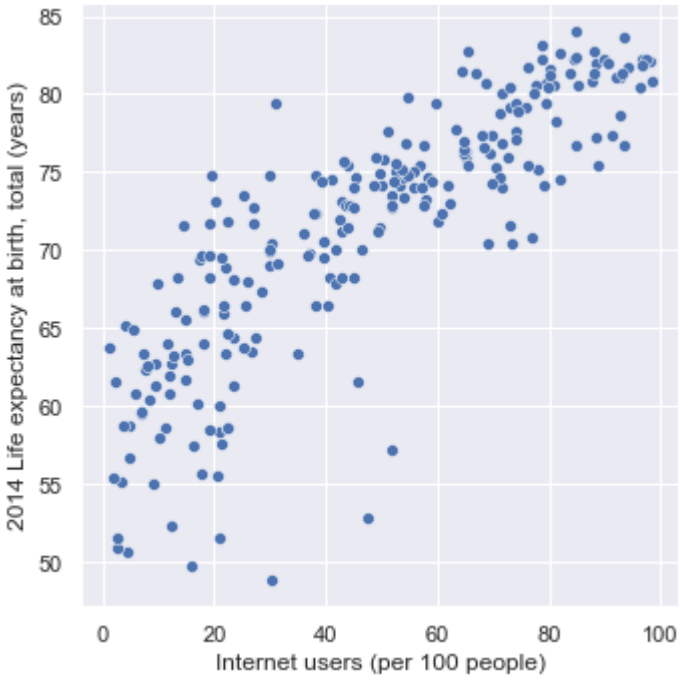
In [120…
```
plt.figure(figsize=(16,3))
ax = plt.subplot(111)
x = x_axis = np.arange(5)
x_ticks = df.nlargest(5, ['Internet users (per 100 people)'])["Country Name"]
z =  df.nlargest(5, ['Internet users (per 100 people)'])["Exports of goods and service
ax.bar(x, z, width=0.2, color='g', align='center')
plt.xticks(x_axis, x_ticks)
```

Out[120]:
```
([<matplotlib.axis.XTick at 0x1c4cc237320>,
  <matplotlib.axis.XTick at 0x1c4cc22fac8>,
  <matplotlib.axis.XTick at 0x1c4cc22f470>,
  <matplotlib.axis.XTick at 0x1c4cc26b7b8>,
  <matplotlib.axis.XTick at 0x1c4cc26b908>],
 [Text(0, 0, 'Bermuda'),
  Text(1, 0, 'Iceland'),
  Text(2, 0, 'Luxembourg'),
  Text(3, 0, 'Andorra'),
  Text(4, 0, 'Norway')])
```



Maybe the correlation with the literacy was just a coincidence and the use of internet is not relevant to test the well being in a country. To test this hypotesis I plotted the correlation between life expectancy and the use of internet in different countries.

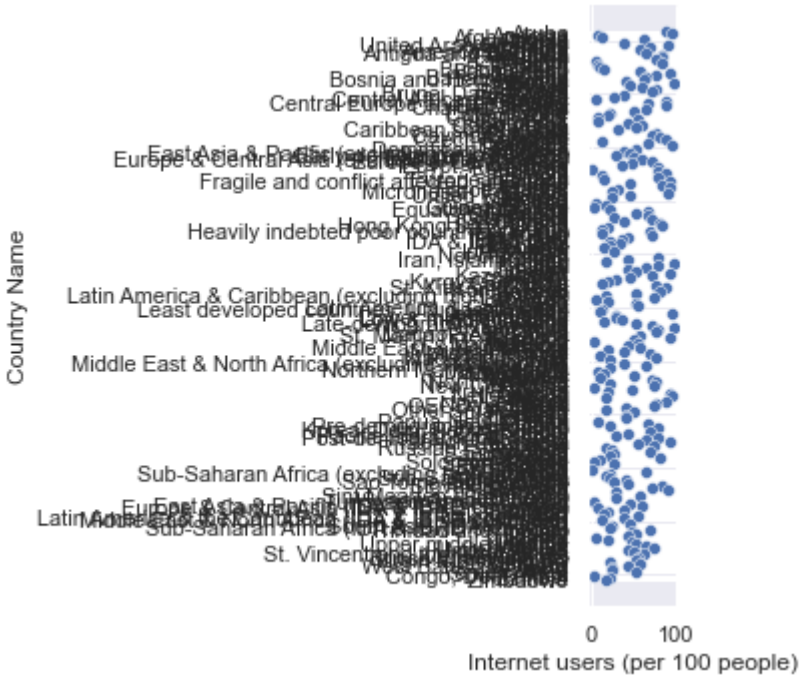
In [89]:
```
sns.relplot(x="Internet users (per 100 people)", y="2014 Life expectancy at birth, tot
```

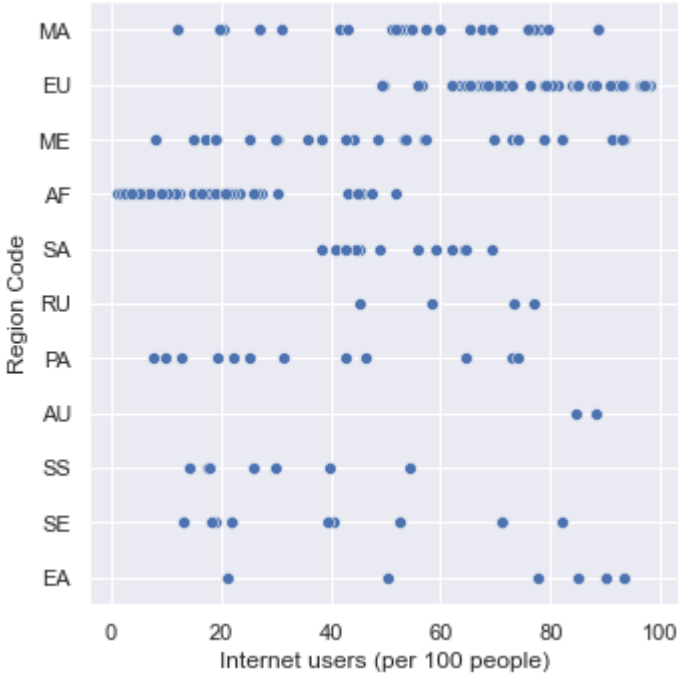


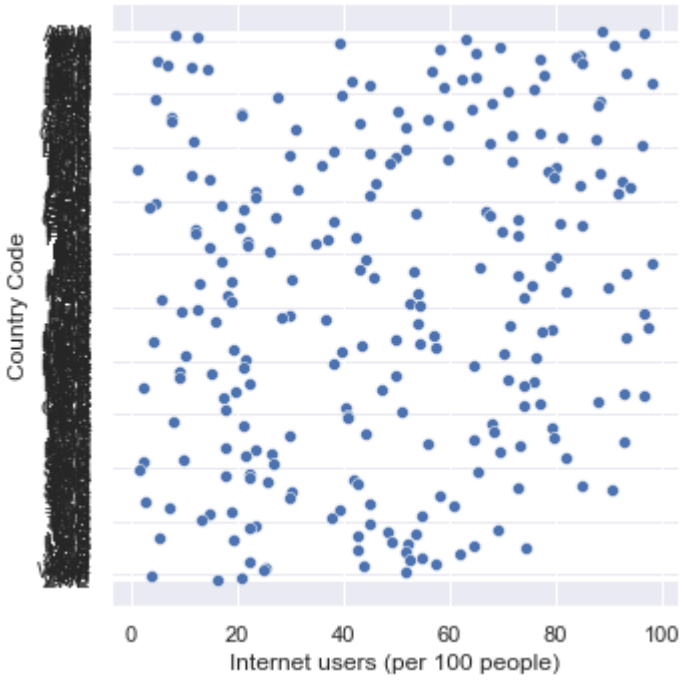
As can be seen above There seems to be again a strong correlation.

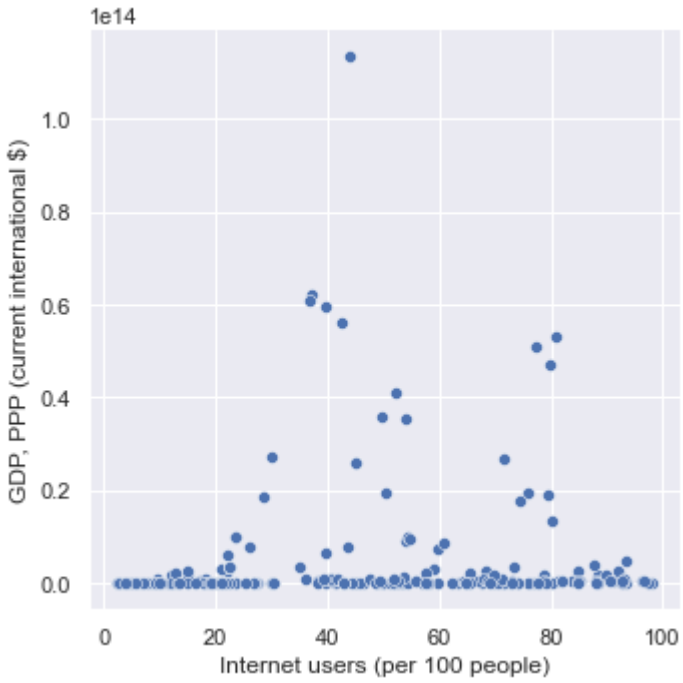
I decided to plot every value against the internet use

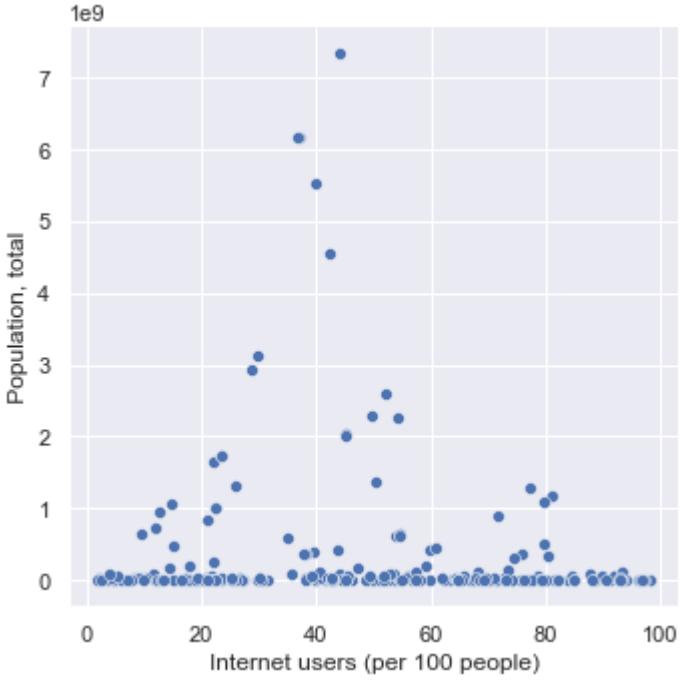
In [91]:
```
for column in df:
    sns.relplot(x="Internet users (per 100 people)", y=column, data=df);
```

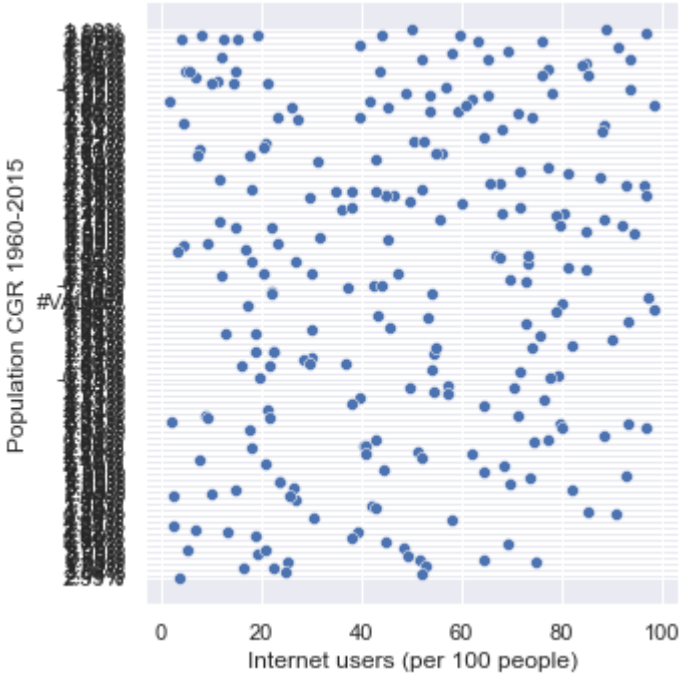

Country Name
Internet users (per 100 people)
0
100

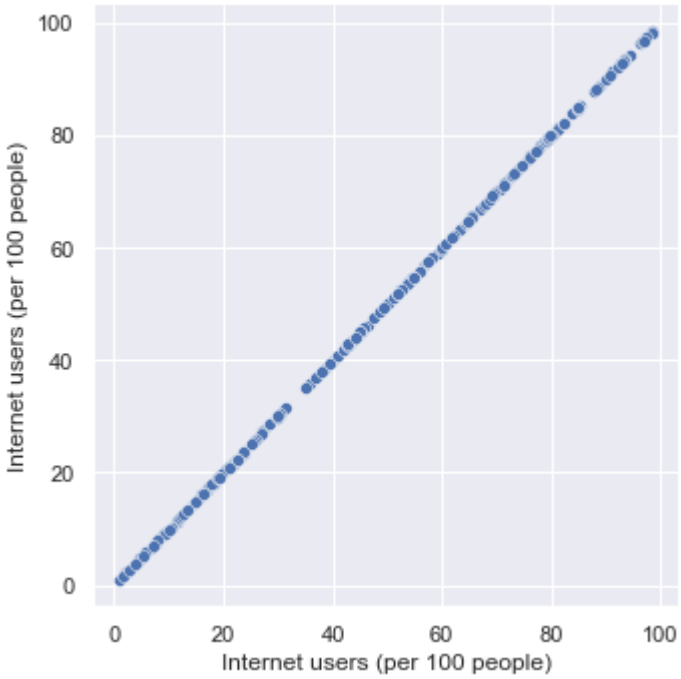

Region Code
MA
EU
ME
AF
SA
RU
PA
AU
SS
SE
EA
0
20
40
60
80
100
Internet users (per 100 people)

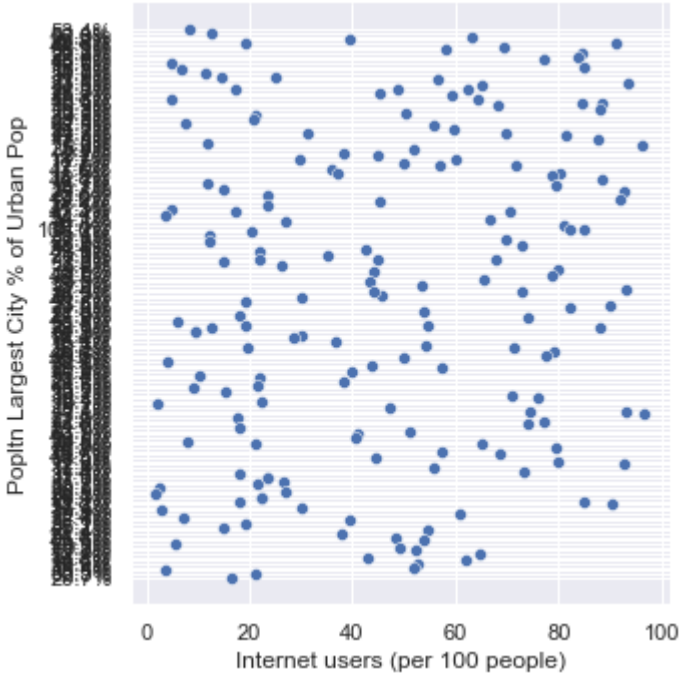

Country Code
0
20
40
60
80
100
Internet users (per 100 people)

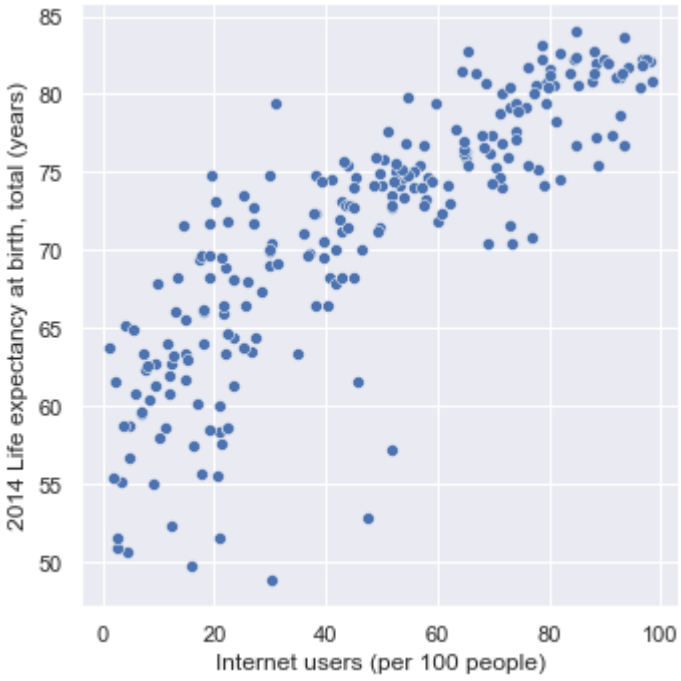

1e14
1.0
0.8
0.6
0.4
0.2
0.0
GDP, PPP (current international $)
0
20
40
60
80
100
Internet users (per 100 people)

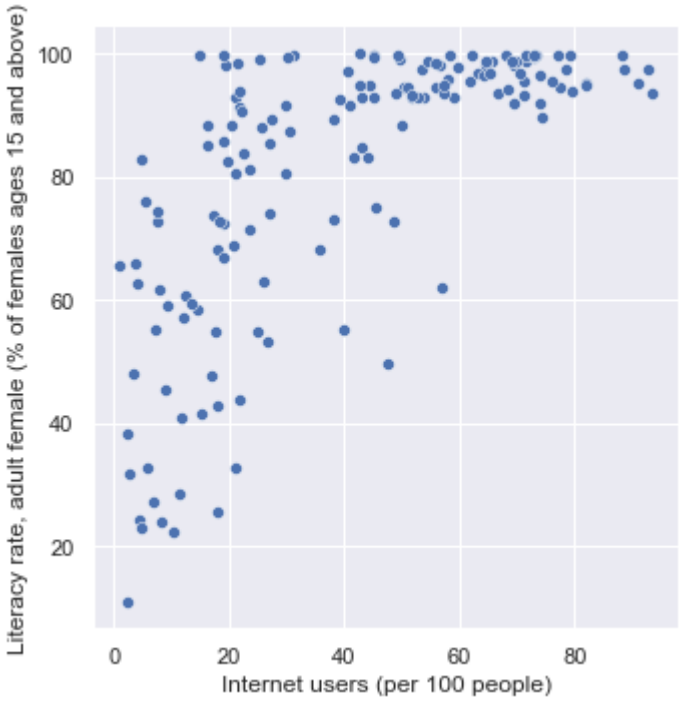

1e9
7
6
5
4
3
2
1
0
Population, total
0
20
40
60
80
100
Internet users (per 100 people)

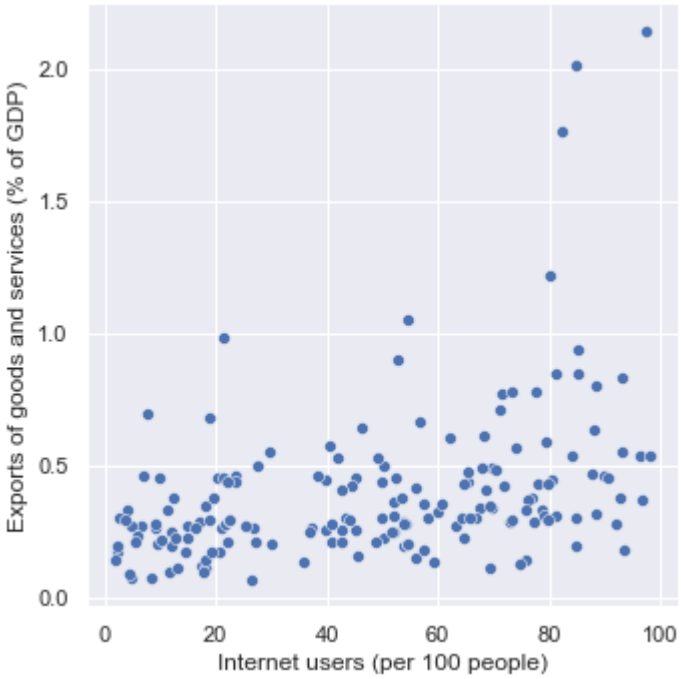

Population CGR 1960-2015
#V
0
20
40
60
80
100
Internet users (per 100 people)


Internet users (per 100 people)
0
20
40
60
80
100
Internet users (per 100 people)


PopItn Largest City % of Urban Pop
0
20
40
60
80
100
Internet users (per 100 people)


2014 Life expectancy at birth, total (years)
85
80
75
70
65
60
55
50
0
20
40
60
80
100
Internet users (per 100 people)


Literacy rate, adult female (% of females ages 15 and above)
100
80
60
40
20
0
20
40
60
80
Internet users (per 100 people)



As seen from the graphs: the correlation exists between: life expectancy, literacy rate and internet use, without any clear connection with other indicators

Results and insights

As showned in the above charts during the study I found that the literacy in the young females from the dataset is correlated with two independent factors: the wealth of the country denoted by the GDP, the internet use / 100 persons and the life expectancy. In a peculiar way, the most important indicator of the literacy was the internet use which is not correlated with the wealth of the country. From this results it will seem that even if a country is poorer, by creating a context in which the people can access information without a big investment is a big step in havin a raised level of litercay. Also, as from the results, the internet is not only used as means of entertainment but is also a powerfull culturalization tool.

Also regarding the internet use, I found interesting the lack of connection between wealth and the use of internet. It will seem that the internet became an accessible tool, and right now, there is no link between the use of the internet and the wealth.

Conclusions

In my opinion the study showed relevant aspects that influence the literacy levels and the internet use in different countries. By using different methods of data vizualization, the link between economical and cultural aspects of the countries can be seen as a whole, and conclusions can be drawn.

The study may be expanded by research on multiple cultural dataset, in order to have a broather overview of the economical impact of different factors and the level of education.

References

[1] McCarthy, James, and Jim Thatcher. "Visualizing new political ecologies: A critical data studies analysis of the World Bank's renewable energy resource mapping initiative." Geoforum 102 (2019): 242-254.

[2] Kurbucz, Marcell Tamás. "A joint dataset of official COVID-19 reports and the governance, trade and competitiveness indicators of World Bank group platforms." Data in brief 31 (2020): 105881.

[3] Tabbara, Yasmina. "Data Visualization."

In [ ]:

In [ ]:

In [ ]:

In [ ]: