Investigating the relationship between technology and well-being

Eole Cervenka for Dr. David Oury (MA710)

March 29th, 2016

Contents

Introduction 1
Variable summary 7
Association rules 16
Clustering analysis 18
Conclusion 25

Introduction

This research studies the relationship between technology and well-being in countries. The data source used is the World Development Indicators data set created by the World Bank.

World Bank's World Development Indicators

The World Bank is a cooperative organization made up of 188 members countries and is governed by the minister of Finance or Development of member countries. The mission of the World Bank is to alleviate poverty and support development. The World Bank works of those fronts by providing both financial and technical assistance to countries.

One of the output of the World Bank efforts is the World Development Indicators (WDI) data set. The WDI presents the most current and accurate global development data available. The original data sources are the statistical systems of member countries and important international organizations. Because the quality of statistical systems varies from country to country and from organization to organization, the World Bank partners with the different national and international agencies to enforce a professional standard in the methodologies, definitions and classifications when sourcing the data.

There is a total to date of over 1300 indicators, open for non-commercial use. For each indicator, data is available by country/region with 248 countries/regions represented and by year from 1960 to present. The list of topics covered by the indicators include Agriculture & Rural Development, Aid Effectiveness, Climate Change, Economy & Growth, Education, Energy & Mining, External Debt, Financial Sector, Gender, Health, Science & Technology, Infrastructure, Labor & Social Protection, Poverty, Private & Public Sector, Social Development and Urban Development.

Missing data policy

Some indicators have been created more recently than others. Overall, recent data tend to be more complete than older data. Furthermore, the data is not equally available from one country to another. For each indicator there can be missing data for multiple countries and/or for multiple years. Too much missing data

can make an indicator irrelevant. I will assume that the critical amount of data for an indicator to be relevant is that it covers 80% of the total 248 countries/regions with at least one data point from 2010 to present. Any indicator that falls below this standard will not be used.

Indicators selection

I chose a subset of indicators data that would represent country's level technological development as well as their level of well-being. Well-being is a complicated concept that can be defined in various ways but for the purpose of this research, I have defined it as a function of `Safety`, `Gender equality`, `Education` and `Fertility`. These topics represent different layers of needs inspired by the 'Hierarchy of needs' as introduced by A. Maslow in his *Theory of Human Motivation*.

TOPIC	INDICATOR CODE	INDICATOR NAME
Technology	IT.NET.BBND.P2	Fixed broadband subscriptions (per 100 people)
Technology	IT.NET.USER.P2	Internet users (per 100 people)
Technology	IT.CEL.SETS.P2	Mobile cellular subscriptions (per 100 people)
Technology	IP.JRN.ARTC.SC	Scientific and technical journal articles
Technology	IT.NET.SECR.P6	Secure Internet servers (per 1 million people)
Infrastructure	BX.GSR.CCIS.CD	ICT service exports (BoP, current USD)
Infrastructure	BX.GSR.CCIS.ZS	ICT service exports (% of service exports, BoP)
Infrastructure	TX.VAL.ICTG.ZS.UN	ICT goods exports (% of total goods exports)
Infrastructure	IT.NET.BBND	Fixed broadband subscriptions
Infrastructure	IT.NET.SECR	Secure Internet servers
Physiology	SH.STA.ACSN	Improved sanitation facilities (% of population)
Safety	SP.DYN.LE00.IN	Life expectancy at birth, total (years)
Safety	VC.IHR.PSRC.P5	Intentional homicides (per 100,000 people)
Gender Equality	SG.GEN.PARL.ZS	% of seats held by women in national parliaments
Gender Equality	SL.TLF.TOTL.FE.ZS	Labor force, female (% of total labor force)
Education	SE.PRM.ENRL	Enrollment in primary education, both sexes (number)
Education	SE.SEC.ENRL.GC	Enrollment in secondary general, both sexes (number)
Fertility	SP.DYN.TFRT.IN	Fertility rate, total (births per woman)
Economy	NY.GDP.PCAP.CD	GDP per capita (current USD)
Economy	SL.UEM.TOTL.ZS	Unemployment, total (% of total labor force)

Research framework

The goal of this research is to uncover existing relationships between technological development and well-being in countries.

- My first objective is to define a metric to measure technological development at the country level. In order to study technological development in countries, I will analyze the countries distribution for the different technological indicators. This will enable me to study the relationship between economic performance and technological development in countries.
- The second objective is to understand the relationships between technological development and the different indicators of well-being (safety, equality, education, fertility, economy). For this purpose, I will use association rule learning to identify potential strong relationships existing between the variables selected.
- Ultimately, the reserach aims at defining the cases, if any, for which technological development has a relationship with the level of well-being in countries. I will perform cluster analysis on countries using different sets of variables to discover which are the variables that best define similarity between

countries in terms of technological development and in terms of well-being. This will enable me to study how naturally the groupings of countries by category of technological development overlap with the groupings of countries by category of level of well-being and reach a conclusion on the existence of a relationship between technology and well-being.

Data preparation

Use of packages

Various packages have been used support the extraction and manipulation of data, as well as the creation of visualizations. In order to collect the data from the World Bank internet website, I used the `WDI package`.

```
# Load installed packages

library('WDI')         # Pull data from worldbank website
library('countrycode')# Convert iso2c to iso3c country code
library("dplyr")       # Data manipulation functions
library("magrittr")    # Syntax with pipes '%>%'
library('ggplot2')     # Data visualization
library('ggrepel')     # Data visualization: Scatterplots label without overlap
library('gridExtra')   # Data visualization: Array of plots
library('grid')        # Data visualization
library('reshape2')    # Data visualization: 2 variables bar plot
library('arules')      # apriori function (association rules)
```

Data import

I initially put the indicators information in a list of vectors; each vector representing an indicator.

```
indicators.full.list = list(
  c("Technology","IT.NET.BBND.P2","internet_subscriptions_percent", ... ),
  c("Technology","IT.NET.USER.P2","internet_users_percent", ... ),
  ...
  c("Economy","SL.UEM.TOTL.ZS","unemployment_over_laborforce_percent", ... )
)
```

Then, I extracted the indicator information in different vectors using `sapply` over the elements of my indicator vectors list.

```
# Extract series codes to the `indicators.codes` variable
indicators.codes = sapply(indicators.full.list, function(x) x[2])

# Extract indicator user-friendly names to the `indicators.names` variable
indicators.names = sapply(indicators.full.list, function(x) x[3])

# Extract indicator full names to the `indicators.full.names` variable
indicators.full.names = sapply(indicators.full.list, function(x) x[4])

# Extract indicator topics to the `indicators.topics` variable
indicators.topics = sapply(indicators.full.list, function(x) x[1])

# Extract indicator topics to the `indicators.definitions` variable
indicators.definitions = sapply(indicators.full.list, function(x) x[5])
```

```
# Create indicators.df to store indicators info in a readable format
indicators.df =
  data.frame(indicators.codes, indicators.names, indicators.full.names, indicators.definitions, indicat
```

Finally, I used the `indicator.codes` vector as an input to the `WDI()` function of the `WDI package` to retrieve data for my variable set which I stored in the `data.frame`: `df`. I used the `indicator.names` vector to rename the variables in order to make them easily identifiable when manipulating `df`.

```
# Read the WDI data for the `indicator.codes` into the `df` dataframe
df <- WDI(indicator=indicators.codes, start = START_YEAR, end = END_YEAR,
          extra=TRUE # returns additional variables, see documentation
)

# Rename variables in `df` using user friendly indicator names
# Not using dplyr `rename` because changing variable set over time
for (i in 1:length(indicators.codes))
{
  colnames(df)[which(names(df) == indicators.codes[i])] <- indicators.names[i]
}
```

Filtering-off non-country entities

The original data set contains geographical entities that are not countries (e.g. : "Arab World"). Because I am interested in studying trends by country, I have to filter-out those entities from my data set. The non-country entities do have a non-official `iso2c` code in the original WDI dataset but there is no official `iso3c` code that it can be converted to.

Therefore, I have converted the `iso2c` codes of the dataset to `iso3c` codes using the `countrycode-package`.

```
# Convert iso2c to iso3c with `countrycode()` from `countrycode-package`
df$iso3c <- countrycode(df$iso2c, "iso2c", "iso3c", warn = FALSE)
```

The entities left without an `iso3c` code are those that are not countries.

```
# Regroup records from non-country entities in `aggregates.df`
aggregates.df <- subset(df, is.na(iso3c))

# Regroup records from country entities in `countries.df`
countries.df <- subset(df, !is.na(iso3c))
```

```
##  country
##  Arab World
##  World
##  East Asia & Pacific (developing only)
##  Europe & Central Asia (developing only)
##  South Asia
##  Central Europe and the Baltics
##  European Union
##  Fragile and conflict affected situations
##  Channel Islands
##  OECD members
##  Small states
```

```
##  Pacific island small states
##  Caribbean small states
##  Other small states
##  Euro area
##  High income
##  Heavily indebted poor countries (HIPC)
##  Latin America & Caribbean (developing only)
##  Kosovo
##  Least developed countries: UN classification
##  Low income
##  Lower middle income
##  Low & middle income
##  Middle income
##  Middle East & North Africa (developing only)
##  High income: nonOECD
##  High income: OECD
##  Upper middle income
##  North America
##  Not classified
##  East Asia & Pacific (all income levels)
##  Europe & Central Asia (all income levels)
##  Sub-Saharan Africa (developing only)
##  Sub-Saharan Africa (all income levels)
##  Latin America & Caribbean (all income levels)
##  Middle East & North Africa (all income levels)
```

Convert categorical variables to factor

In order to properly summarize categorical variables such as `year` and `country`, I have changed those variables data type to `factor`.

```
df %>%
  select(-capital,# Select out unwanted variables
         -longitude,
         -latitude,
         -lending,
         -iso2c) %>%
  mutate(country=factor(country), # Convert categorical variables to `factor`
         iso3c=factor(iso3c),
         year =factor(year),
         region =factor(region),
         income =factor(income)) %>%
         { . } -> df
```

Missing data

Because data points are missing by indicator and by year, I decided to create a new data set using the most recent data available for each indicator and each country entity. I will not be using data older than 2008 because I assume that for any country, the data of an indicator may have changed too much since 2008 to accurately describe a current situation.

Code:

```
# For each country, retrieve most recent indicator data available
# on period (START_YEAR - END_YEAR)

# Initialize final.df with distinct ISO3C code, country
countries.df %>%
  select(iso3c, country) %>%
  distinct() -> final.df

# Function(): get most recent data available in `df` for variable `var`
# @return: dafa.frame('iso3c','var')
get_recent_data <- function(df, var){
  df %>%
    mutate(year = as.numeric(levels(df$year))[df$year]) %>%
    select(iso3c, year, match(var, names(.))) %>%
    na.omit() %>%
    group_by(iso3c) %>%
    filter(year == max(year)) %>%
    select(iso3c, match(var, names(.))) -> recent.data.df

  return(recent.data.df)
}

# For each variable:
# Left join the most recent variable data with `df` on ISO3C
for (i in 1:length(indicators.names)){
  x <- get_recent_data(countries.df, indicators.names[i])
  final.df %>%
    left_join(x) -> final.df
}
```

Create quantiles for numerical variables

Categorization of numeric variables will come handy. I have created a function to translate a numerical value to its corresponding quantile within the variable data. By default, I will use quartiles.

Code:

```
# helper fct:
# @return `cutoff` vector
cutoff <- function(num_quantiles)
{
  i <- 1/num_quantiles
  cutoff = list()
  for (k in 1:(num_quantiles-1)){
    cutoff[k] <- k*i
  }
  cutoff_vec = unlist(cutoff)
  return(cutoff_vec)
}

# helper fct:
# @return quantile `labels` vector
labelize.it <- function(num_quantiles)
{
```

```
  labels = list()
  for (k in 1:num_quantiles){
    labels[k] <- paste(num_quantiles-k+1,'/',num_quantiles,'th',sep='')
  }
  labels_vec = unlist(labels)
  return(labels_vec)
}

# main function
# @return factor vector of quantiles for a given numeric vector variable
quantilize <- function(vec, num_quantiles)
{
  if(!is.numeric(vec)){
    stop('Please choose numeric vector')
  }

  quantile(vec,
           cutoff(num_quantiles),
           na.rm=TRUE
  ) %>%
  { c(-Inf, ., Inf) } %>%
    cut(vec,
        breaks = .,
        labels = labelize.it(num_quantiles)
    )
}


## EXAMPLE :
countries.df %>%
  mutate(
    GDP_per_capita_factor = quantilize(final.df$GDP_per_capita, 4)
  ) %>%
  { . } -> final.df
```

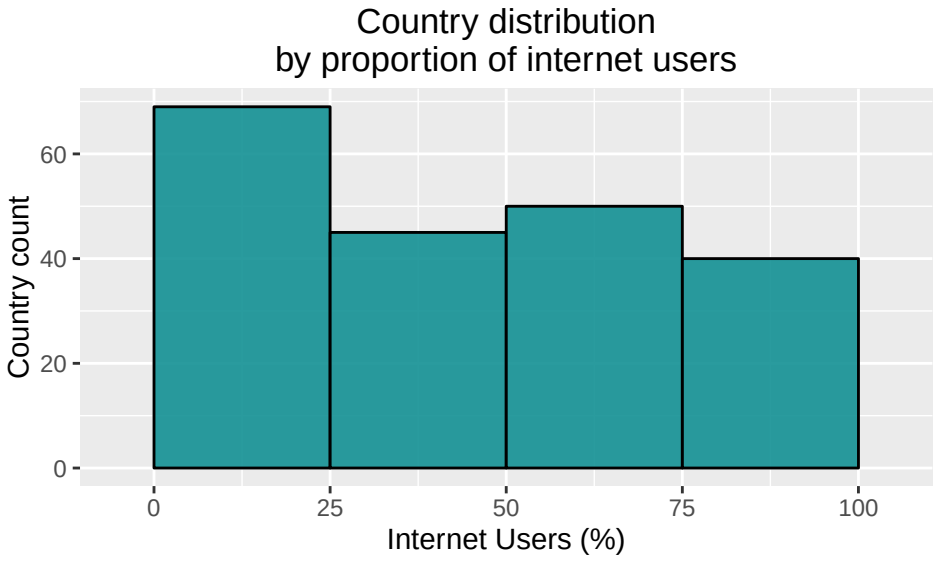
Variable summary

As a preliminary step, I will go through the variables selected and explore their distribution.

Internet users (per 100 people)

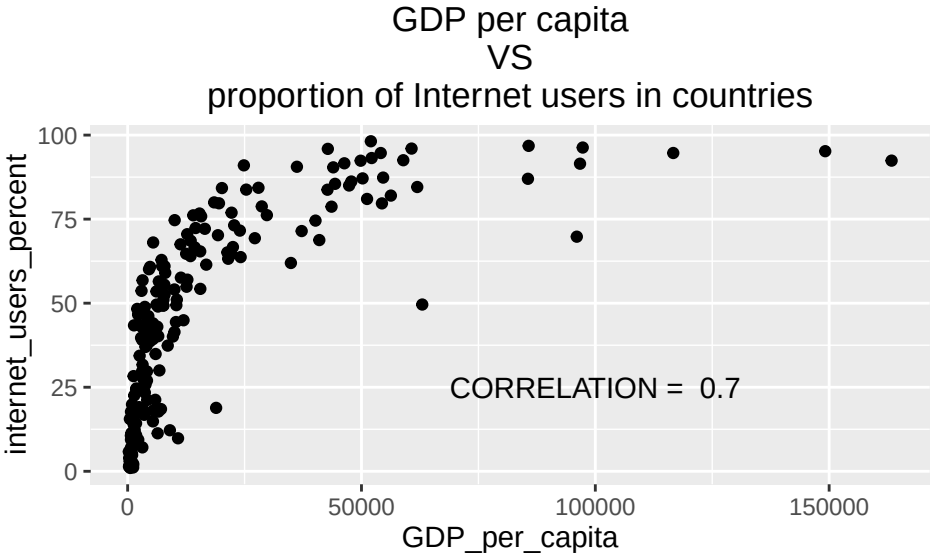
This variable represents the number of individuals who have used the Internet (from any location) in the last 12 months. Internet can be used via various devices including computer, mobile phone, personal digital assistant, games machine or digital TV.

The graph below explore the distribution of countries in terms of the proportion of internet users over the total national population.



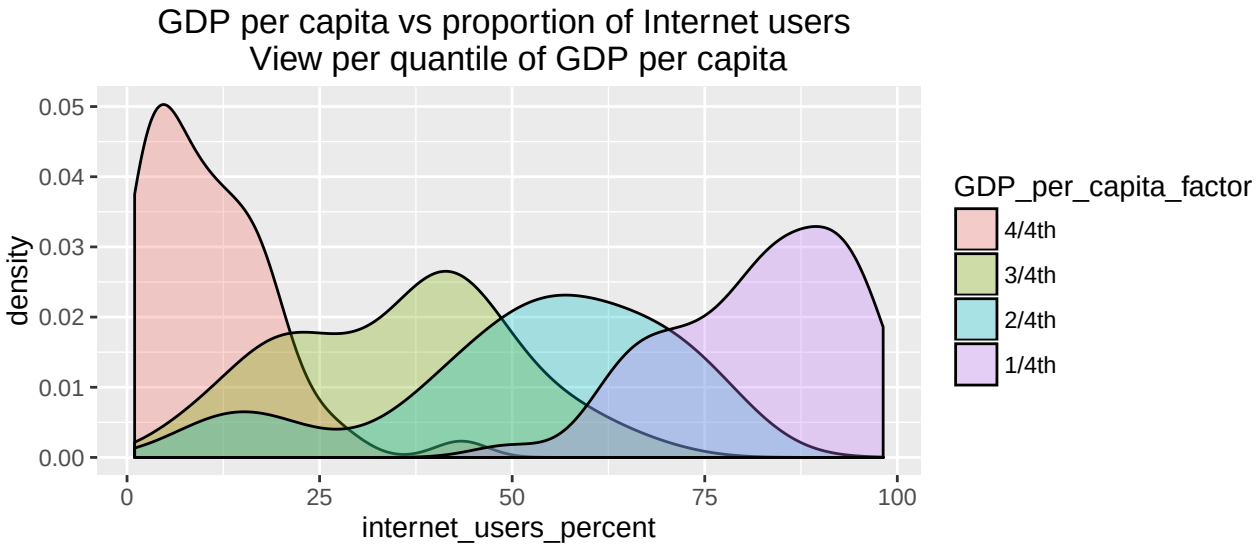
Based on the most recent data available, I have divided countries by the proportion of their population who are internet users. I observe that the largest category, composed of 70 countries, is that of countries where less than one individual out of four has used the internet in the last 12 months.

The scatter plot below puts in perspective the internet usage rates with the countries' relative economic performance and allows to visualize the correlation between economic performance and access to internet.



With a correlation of almost 0.7, the GDP per capita and the proportion of internet users in countries have a positive and moderately strong linear relationship. Looking at the lowest GDP per capita points, it is safe to assume the relationship between internet usage and GDP per capita is very strong for poor countries and becomes weaker as the GDP per capita increases.

To observe how the relationship between economic performance and internet usage changes across different levels of economic performance, I have categorized countries per quantile of GDP per capita.



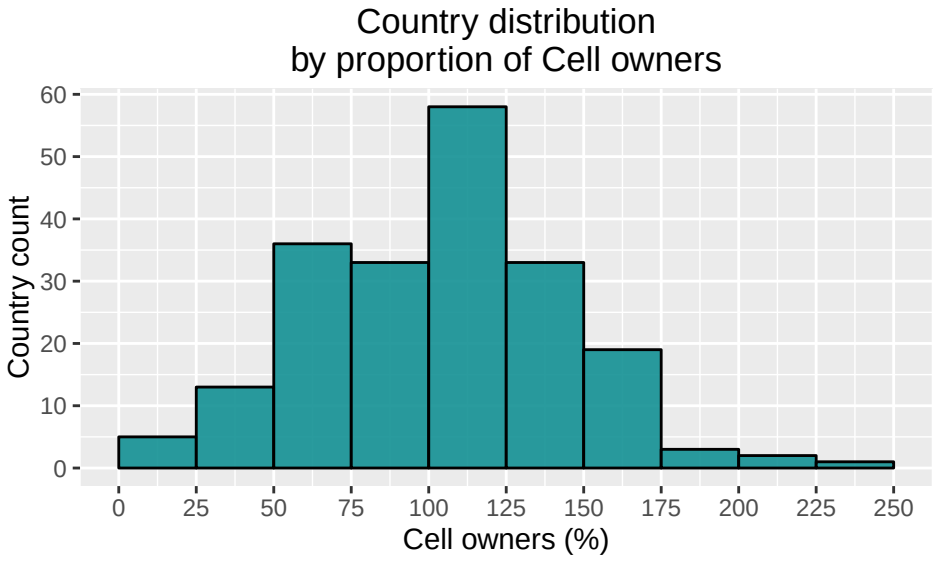
The 25% of countries with lowest GDP per capita all have very low internet usage. In those countries between 0% and 15% of the population only are internet users. Among the top 25% of countries in terms of economic performance, there are very few countries that have an internet usage lower than any of the 25% poorest countries. In fact, for almost all of the top25% countries in terms of GDP per capita, at least 50% of the country's population is an internet user.

As far as the countries in the 3/4th and 2/4th in terms of GDP per capita are concerned, interpretation of the density is more tedious. A relationship between economic performance and internet usage is still visible but the density curves would not be a relevant tool to analyze it.

Mobile cellular subscriptions (per 100 people)

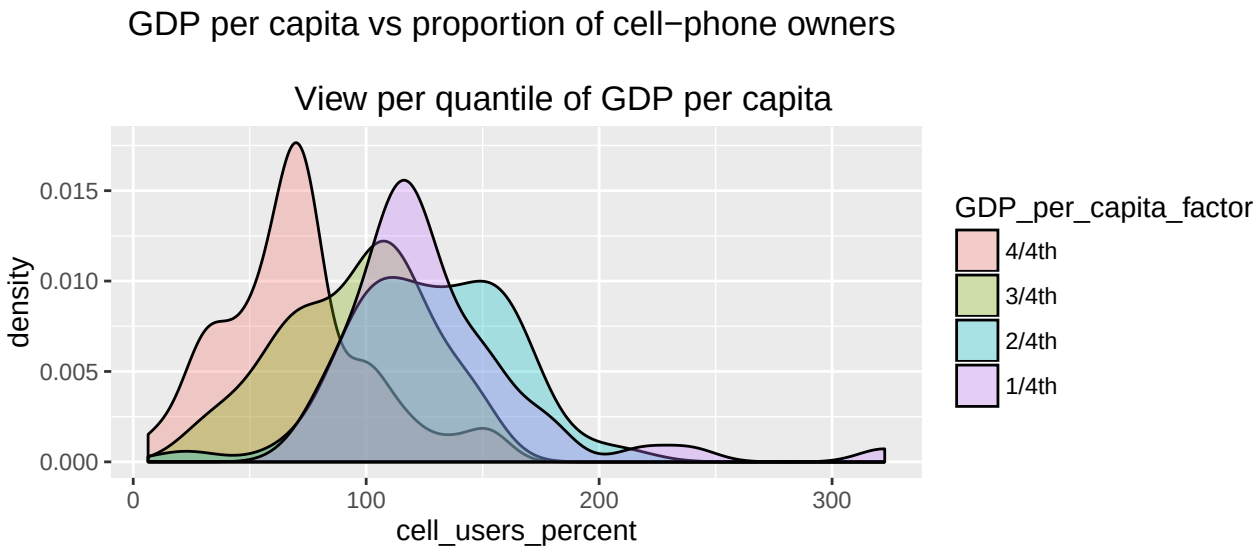
Mobile cellular telephone subscriptions are subscriptions to a public mobile telephone service that provide access to the PSTN using cellular technology. The indicator includes the number of postpaid subscriptions, and the number of active prepaid accounts (i.e. that have been used during the last three months). The indicator applies to all mobile cellular subscriptions that offer voice communications. It excludes subscriptions via data cards or USB modems, subscriptions to public mobile data services, private trunked mobile radio, telepoint, radio paging and telemetry services.

The graph below explore the density distribution of proportion of cell users in countries.



In almost 60 countries, the average mobile subscription rate is above one per individual. Overall, countries with more than one subscription per individual are more frequent than those with less than one subscription per individual.

To observe how the relationship between economic performance and cellular subscription rate changes across different levels of economic performance, I have categorized countries per quantile of GDP per capita. The density plot gives a quick visualization of the distribution of cellular subscription rate across different categories of economic performance.



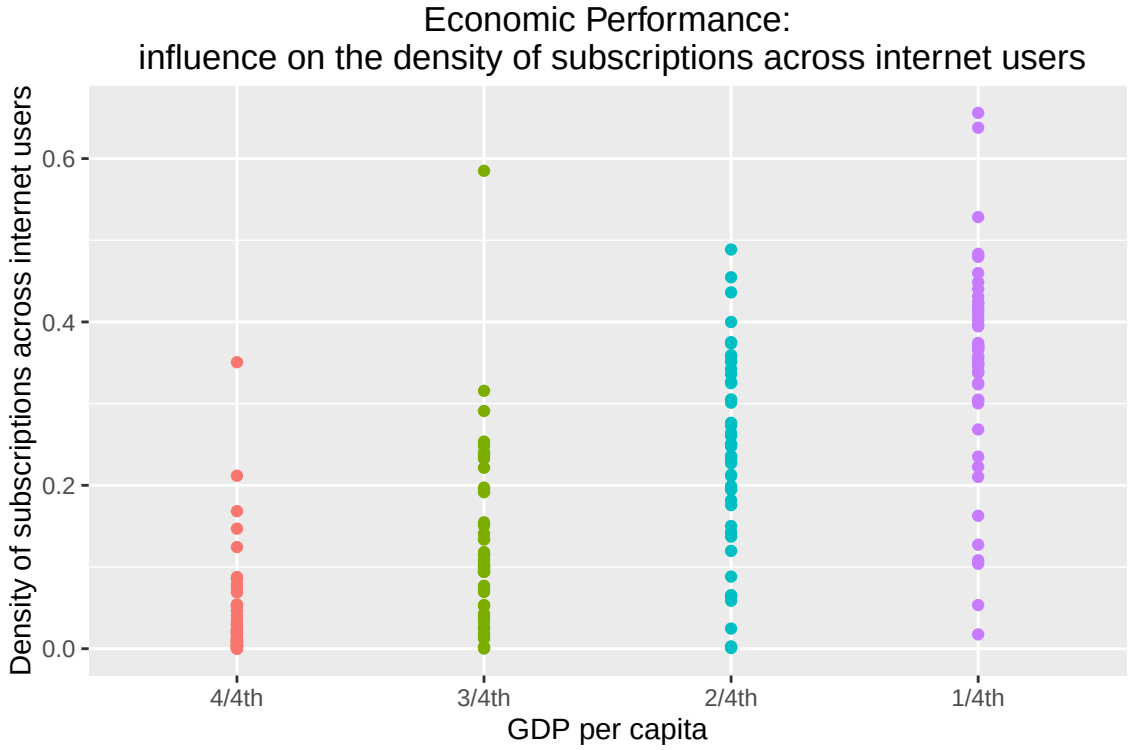
The only density curve that is not clearly overlapping with the others is that of the 25% of countries with lowest GDP per capita. While cellphone ownership is lower in the weakest economies, it is similar for the 75% rest of the countries. I can assume that better economic performance does not drive up the cell subscription rate past a certain point. That makes cell phone usage rate distribution much different from internet usage in that regard.

Fixed broadband subscriptions (per 100 people)

Fixed broadband subscriptions refers to fixed subscriptions to high-speed access to the public Internet (a TCP/IP connection), at downstream speeds equal to, or greater than, 256 kbit/s. This includes cable modem, DSL, fiber-to-the-home/building, other fixed (wired)-broadband subscriptions, satellite broadband and terrestrial fixed wireless broadband. This total is measured irrespective of the method of payment. It excludes subscriptions that have access to data communications (including the Internet) via mobile-cellular networks. It should include fixed WiMAX and any other fixed wireless technologies. It includes both residential subscriptions and subscriptions for organizations.

In order to understand better how different countries access internet according to their economic performance, I consider the internet subscription density given by:

subscription_density = internet_subscription(%) / internet_users(%)



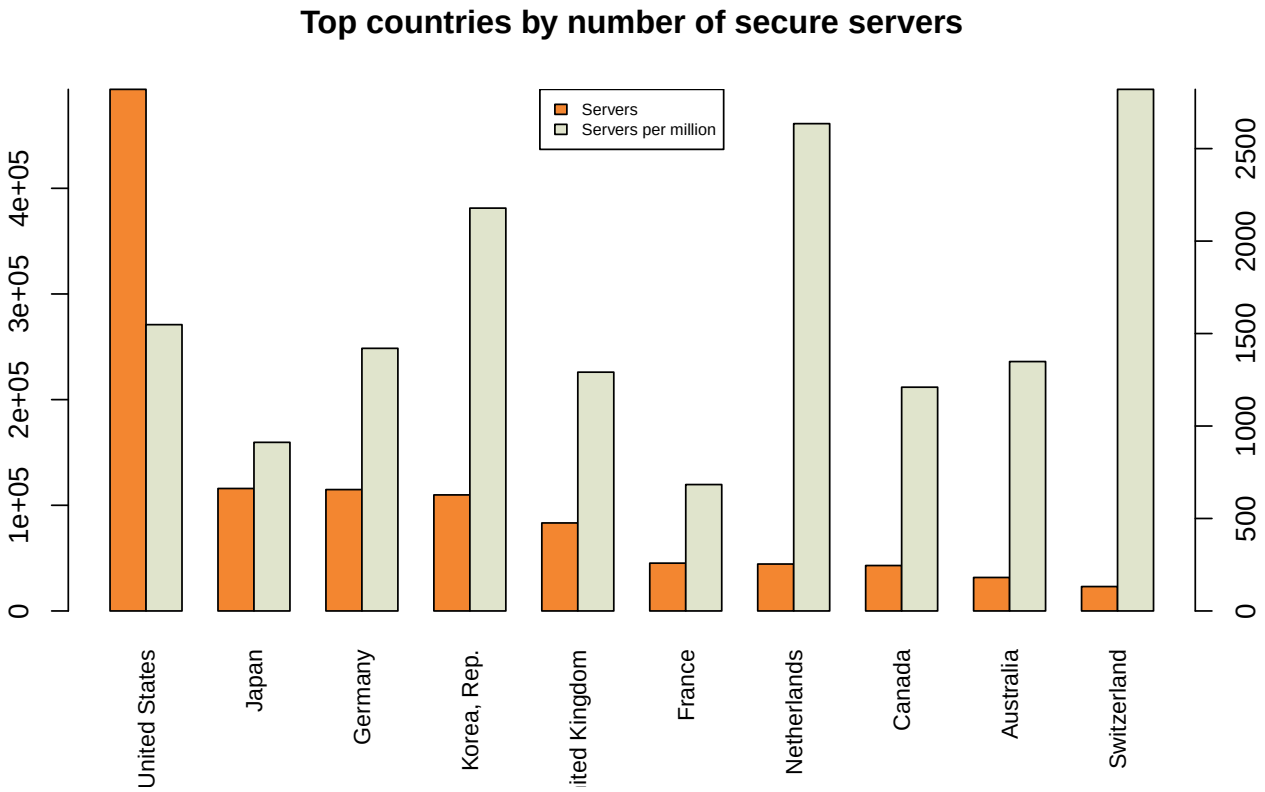
The richest the country, the more internet users subscribe to personal internet connection. Other practices that are apparently more common in less affluent countries could include, for example, sharing a private internet access with a community of people, using public internet access or surfing the internet on a mobile phone.

Secure Internet servers (per 1 million people)

E-commerce requires secure ways of conducting buying and selling transaction over the internet. E-commerce companies use secure servers (with encryption technology) for Internet transactions. The number of Secure Internet servers in countries provide one of the best indicators of the diffusion of e-commerce by country. Similarly, the volume of Secure Internet servers per million is a useful measure of the relative intensity of use of e-commerce. 'ICTs and the Information Economy' OECD.

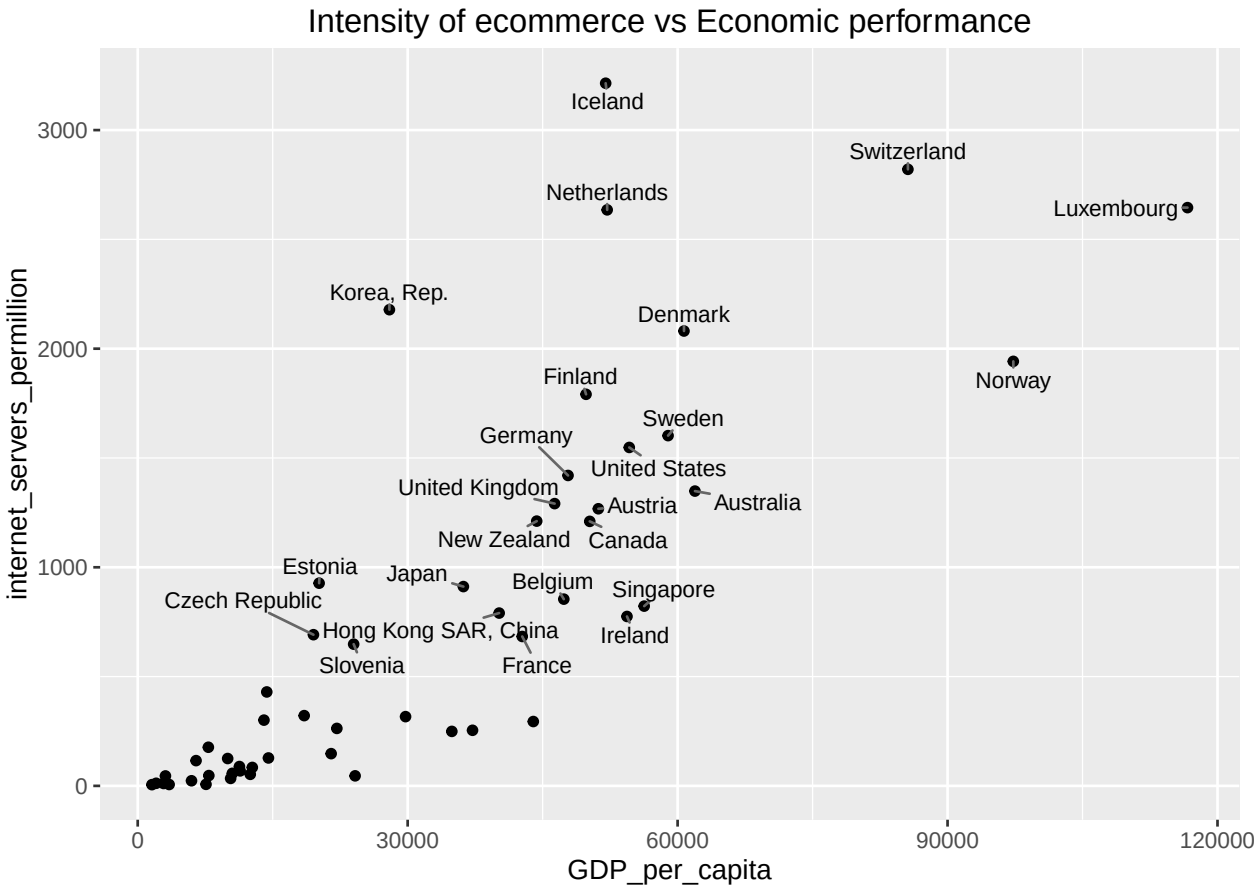
The volume of internet servers in a country is an indicator of its level of e-commerce. By the same token, the proportion of internet servers per million inhabitants in a country characterizes the intensity of e-commerce in a country.

This graph displays the top10 countries in terms of e-commerce activity (volume of servers) and their respective intensity.



The top players is by far the USA (5,000 servers), home to 5 times more servers than the second, Japan. However, in terms of e-commerce intensity, the US is actually in the average of the top10. South Korea, the Netherlands and Switzerland are e-commerce intensive countries with 2 to 2.5 servers per thousand inhabitants.

The following plot is a visualization of the relationship between economic performance in countries and the intensity of e-commerce.



This graph highlights those small countries who have heavily invested heavily in internet servers to be top internet players in spite of their small size.

To investigate further the nature of those small highly e-commerce oriented countries, I pulled the list of the top10 countries in terms of e-commerce intensity (number of servers per million inhabitants).

```
##    country        internet_servers internet_servers_permillion
## 1  Liechtenstein     364           9762.377
## 2  Bermuda           423           6489.621
## 3  Monaco            121           3216.118
## 4  Iceland          1053           3214.394
## 5  Switzerland     23100           2820.434
## 6  Luxembourg       1471           2645.331
## 7  Netherlands     44412           2635.073
## 8  Isle of Man       223           2559.482
## 9  Cayman Islands    140           2365.984
## 10 Korea, Rep.    109841           2178.350
```

Apart from South Korea ranking 10th, we only find countries that are to a certain extent defined as tax havens. Liechtenstein with almost 1 server per 100 inhabitants is considered by most economists the second major tax haven after Switzerland (Tolley's Tax Havens, 2000). Luxembourg and the Netherlands are primarily conduit tax havens (S. Markle and A. Shakelford, 2009). Other non-sovereign jurisdictions commonly labeled as tax havens include Bermuda, Isle of Man, the Cayman Islands.

Scientific and technical journal articles

Scientific and technical journal articles refer to the number of scientific and engineering articles published in the following fields: physics, biology, chemistry, mathematics, clinical medicine, biomedical research, engineering and technology, and earth and space sciences.

This indicator is particularly useful to demonstrate how advanced a country is in terms of both education and technology. Below is a summary of the top10 countries in terms of quantity of scientific and technical journal articles published in 2014.

```
##  country        scientific_publications
##  United States  208600
##  China           89894
##  Japan           47105
##  Germany         46258
##  United Kingdom  46035
##  France          31685
##  Canada          29016
##  Italy           26503
##  Korea, Rep.     25592
##  Spain           22910
```

The US is a world leader in the volume of scientific papers published per year with over two times more than the second, China. In fact, the US published almost as much as the rest of the top 5 countries for publication of scientific articles. This is indicative of a large country that has a high number of scientific institutions and is among the world leaders in technology.

ICT service exports

Information and communication technology service exports include computer and communications services (telecommunications and postal and courier services) and information services (computer data and news-related service transactions). Data are in current U.S. dollars.

ICT goods exports

Information and communication technology goods exports include telecommunications, audio and video, computer and related equipment; electronic components; and other information and communication technology goods. Software is excluded.

Improved sanitation facilities (% of population)

Access to improved sanitation facilities refers to the percentage of the population using improved sanitation facilities. Improved sanitation facilities are likely to ensure hygienic separation of human excreta from human contact. They include flush/pour flush (to piped sewer system, septic tank, pit latrine), ventilated improved pit (VIP) latrine, pit latrine with slab, and composting toilet.

Life expectancy at birth, total (years)

Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.

Intentional homicides (per 100,000 people)

Intentional homicides are estimates of unlawful homicides purposely inflicted as a result of domestic disputes, interpersonal violence, violent conflicts over land resources, intergang violence over turf or control, and

predatory violence and killing by armed groups. Intentional homicide does not include all intentional killing; the difference is usually in the organization of the killing. Individuals or small groups usually commit homicide, whereas killing in armed conflict is usually committed by fairly cohesive groups of up to several hundred members and is thus usually excluded.

Proportion of seats held by women in national parliaments (%)

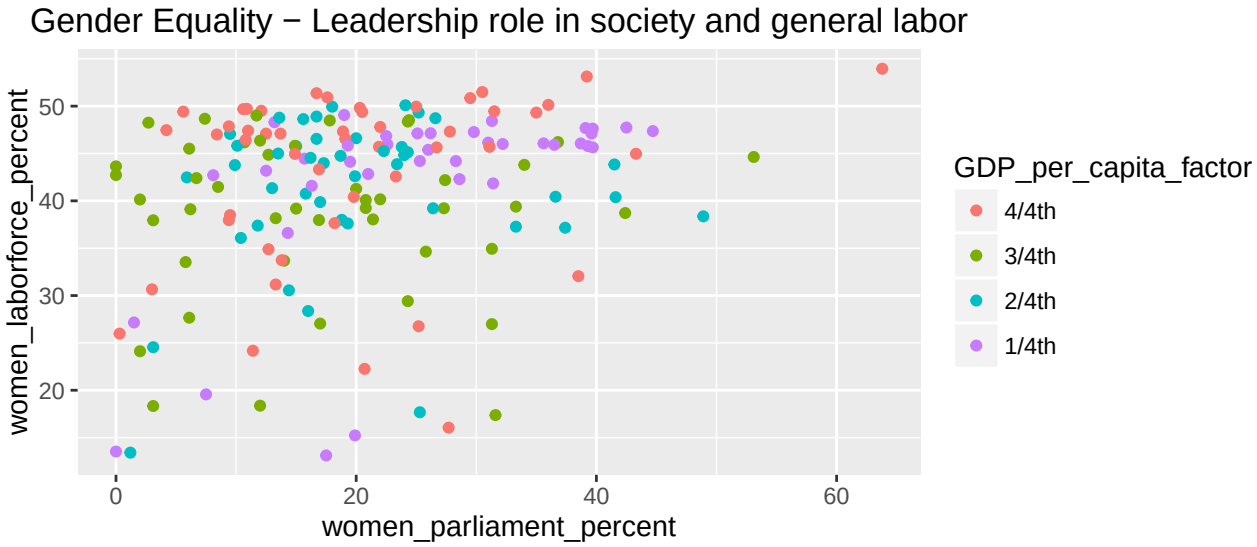
Women in parliaments are the percentage of parliamentary seats in a single or lower chamber held by women.

Labor force, female (% of total labor force)

Female labor force as a percentage of the total show the extent to which women are active in the labor force. Labor force comprises people ages 15 and older who meet the International Labour Organization's definition of the economically active population.

My assumption is that there is a possible relationship between the fact that a country has a balanced distribution of gender within its parliament and that it has a balanced distribution of gender within its labor force. Perhaps the distribution in both instances is also correlated with the relative economic performance of the county.

The graph below is a scatterplot mapping countries by the relative proportion of women in the parliament and relative proportion of women in the labor force. I added the categories of GDP per capita to observe if possible groupings on these gender variables correlated with the economic performance.



There is no observable relationship between economic performance and gender equality in parliament and general labor force composition in countries at this point.

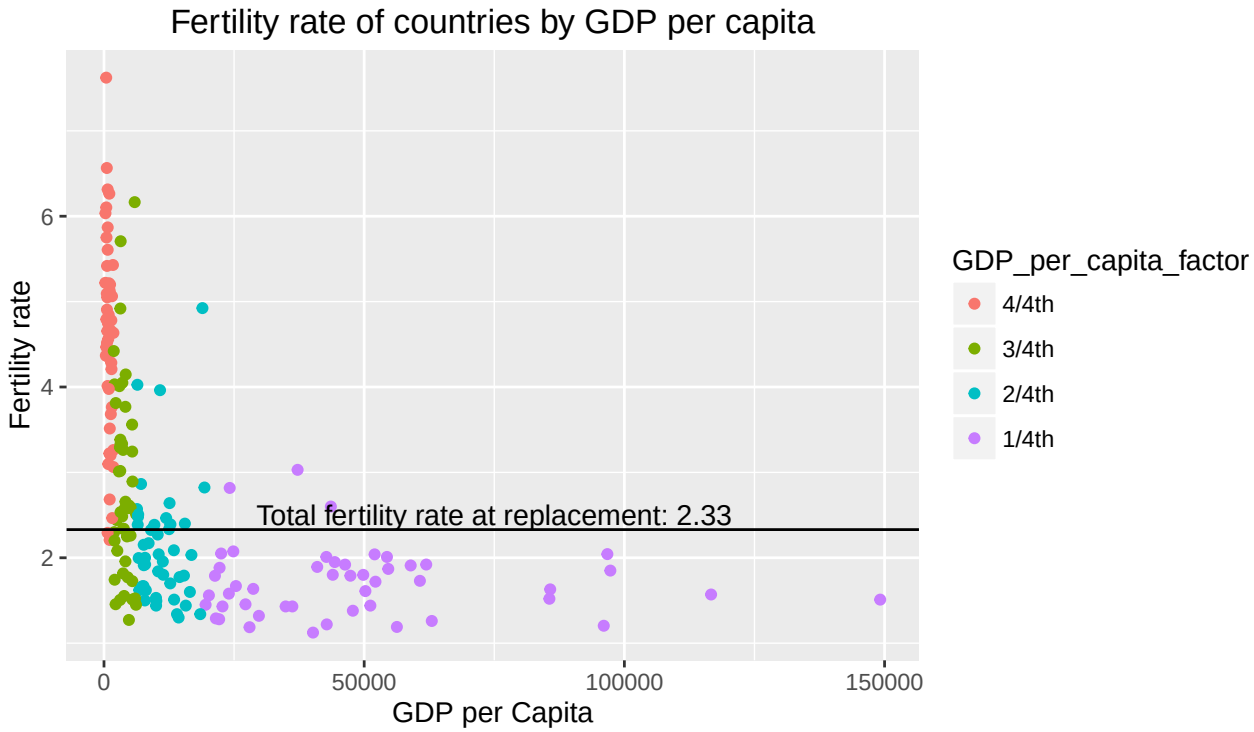
Enrolment in primary education, both sexes (number)

Total number of students enrolled in public and private primary education institutions regardless of age.

Fertility rate, total (births per woman)

Total fertility rate represents the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with current age-specific fertility rates.

I explored below a possible relationship between the fertility rate of countries against their economic performance. In order to have a general idea of the natural growth rate of countries' population, I drew the line of the fertility rate at replacement, which is 2.33 children per women in average.



The 25% poorest countries have a high to very high fertility rate, most of them above 4 children per woman. Almost all are above 2.33 children per child meaning none have a negative natural growth rate. The opposite observation can be made about the 25% richest countries since only a few reaches a fertility rate of 2 children per child. We can assume that low fertility rate represents an economic challenge for the richest economic countries. This challenge can perhaps be addressed with a positive net immigration and an increased mechanization/cybernation of labor.

GDP per capita (current US$)

GDP per capita is gross domestic product divided by midyear population. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in current U.S. dollars.

I use GDP per capita in this research as a single metric for economic performance in countries.

Unemployment, total (% of total labor force)

Unemployment refers to the share of the labor force that is without work but available for and seeking employment.

Association rules

Next I will investigate the association riles that I can generate with the data set.

First I need change the numeric variables of `countries.df` into categorical variables. The `quantilize()` function created earlier will come handy. The new data is stored in the `arules.df` data.frame specifically dedicated to association rules analysis.

```
# Setup a dataframe to be used by the `apriori` function
countries.df %>%
  mutate(
    internet_subscriptions_percent = quantilize(internet_subscriptions_percent,4),
    internet_users_percent = quantilize(internet_users_percent,4),
    cell_users_percent = quantilize(cell_users_percent,4),
    scientific_publications= quantilize(scientific_publications,4),
    internet_servers_permillion= quantilize(internet_servers_permillion,4),
    ict_service_exports= quantilize(ict_service_exports,4),
    ict_service_exports_percent= quantilize(ict_service_exports_percent,4),
    ict_good_exports = quantilize(ict_good_exports,4),
    ict_good_exports_percent = quantilize(ict_good_exports_percent,4),
    internet_subscriptions = quantilize(internet_subscriptions,4),
    internet_servers = quantilize(internet_servers,4),
    improved_sanitation_percent= quantilize(improved_sanitation_percent,4),
    life_expectancy= quantilize(life_expectancy,4),
    homicide_percent_1000= quantilize(homicide_percent_1000,4),
    women_parliament_percent = quantilize(women_parliament_percent,4),
    women_laborforce_percent = quantilize(women_laborforce_percent,4),
    enrolment_primary_education= quantilize(enrolment_primary_education,4),
    enrolment_secondary_education= quantilize(enrolment_secondary_education,4),
    fertility_rate = quantilize(fertility_rate,4),
    GDP_per_capita= quantilize(GDP_per_capita,4),
    unemployment_over_laborforce_percent = quantilize(unemployment_over_laborforce_percent,4)
  ) %>%
  { . } -> df.arules
```

Please note that `countries.df` already contains variables of datatype `factor` which will be left untouched, namely: `country`, `year`, `iso3c`, `region` and `income`.

The `apriori()` function of the `arules` package is ran to generate the association rules. I chose to discuss rules that have a support of at least 0.01 and a confidence of at least 0.01.

```
# Run the `apriori` function
rules = apriori(df.arules,
                parameter=list(maxlen=2,# Generate 2-variables rules
                               support=0.01,# Support is at least 0.01
                               confidence=0.01))# Confidence is at least 0.01
```

Below are the most noteworthy rules, in descending order of lift value:

```
##   lhs                            rhs                            support confidence    lift
## 1 {income=High income: OECD} => {ict_service_exports=1/4th} 0.1098383  0.7511521 4.53134
```

[1] *'A country from the top25% in terms of proportion of its total exports dedicated to ICT goods is also a top25% country in terms of absolute revenue generated from the exports of ICT good.'* The lift on this rule is 4.55 meaning that if a country belongs to the top25% of highest revenue countries in terms of ICT good exports, it is 4.55 times more likely to belongs to the top quartile of countries by share of exports dedicated to ICT goods. It is highly unlikely that a countries is a leader in ICT goods and yet do not dedicate a significant share of its exports to ICTs.

```
##   lhs                             rhs                                          support confidence   lift
## 1 {income=High income: OECD} => {internet_subscriptions_percent=1/4th} 0.1253369  0.8571429 4.4947
```

[2] *'An OECD country is in the top quartile of countries in terms of ICT service exports.'* The support of this rule is 0.1, so I know that 10% of records contained countries both in the OECD group and in the top quartile of countries in terms of ICT service exports. The confidence is 0.75 which means that 75% of OECD countries data for available years indicated that the country belonged to the top25% of countries in terms of ICT service exports. Finally the lift of 4.53 indicates that it is 4.53 times more likely that a country belongs to the top quartile in terms of ICT service exports if we know that it is an OECD country.

```
##   lhs                               rhs                           support confidence     lift
## 1 {ict_service_exports=1/4th} => {ict_good_exports=1/4th} 0.1381402  0.8333333 4.369847
```

[3] Similarly, an OECD country is very likely to belong to the top quartile in terms of proportion of the population using the internet. In fact, the confidence of 0.85 tells us that in 85% of OECD country records, data indicates that the OECD country is in the top quartile in proportion of internet users.

```
##   lhs                                                    rhs                                  sup
## 1 {region=Middle East & North Africa (all income levels)} => {women_laborforce_percent=4/4th} 0.0943
```

[4] *'If a country is in the region middle east, then the proportion of women in the labor force is among the 25% lowest worldwide.'* 9% of records provide cases of a country that is in the region 'Middle-East' and has a proportion of women in the labor force that is among the 25% lowest worldwide. The confidence for this rule is as high as 95% meaning that 95% records of countries in the Middle-East were in compliance with this rule. Finally, the lift is 4.36 meaning that it is 4.36 times more likely that the proportion of women in the labor force is among the 25% lowest worldwide if we were to pick a country at random in the Middle-East than if we were to pick any country worldwide.

Clustering analysis

Heat maps

The primary objective of my clustering analysis is to discover potential natural groupings formed by the countries across the different variables used in this analysis.

I consider to sets of variables: the 'tech variables' and the 'well-being variables'.

```
tech.indicators.names <-
  c(
    "internet_subscriptions_percent",
    "internet_users_percent",
    "cell_users_percent",
    "scientific_publications",
    "internet_servers_permillion",
    "ict_service_exports",
    "ict_good_exports",
    "ict_good_exports_percent",
    "internet_subscriptions",
    "internet_servers"
  )

wellbeing.indicators.names <-
```

```
  c(
    "improved_sanitation_percent",
    "life_expectancy",
    "homicide_percent_1000",
    "women_parliament_percent",
    "women_laborforce_percent",
    "enrolment_primary_education",
    "enrolment_secondary_education",
    "fertility_rate",
    "GDP_per_capita",
    "unemployment_over_laborforce_percent"
  )
```

The initial step consists in using PAM method to assign each country record to a cluster (I chose k = 4 simplicity). I performed 3 different sets of clusters. One using all variables available, one using only technology variables and one using only well-being variables

```
# GENERAL CLUSTERS (All variables included)
# Perform PAM clustering on entire data set
cluster.count = 4
pam.cluster.maker.df = pam(x=cluster.maker.df[,indicators.names], k=cluster.count)
pam.cluster.maker.df$cluster_all = factor(pam.cluster.maker.df$cluster)

# Concatenate Cluster columns to `cluster.maker.df`
cluster.maker.df <- cbind(cluster.maker.df, cluster_all = pam.cluster.maker.df$cluster_all)

# WELLBEING CLUSTERS (Only well being var included)
# Perform PAM clustering on well-being indicators only
cluster.count = 4
pam.cluster.maker.df = pam(x=cluster.maker.df[,wellbeing.indicators.names], k=cluster.count)
pam.cluster.maker.df$cluster_wellbeing = factor(pam.cluster.maker.df$cluster)

# Concatenate Cluster columns to `cluster.maker.df`
cluster.maker.df <- cbind(cluster.maker.df, cluster_wellbeing = pam.cluster.maker.df$cluster_wellbeing)

# TECH CLUSTERS (Only tech var included)
# Perform PAM clustering on tech indicators only
cluster.count = 4
pam.cluster.maker.df = pam(x=cluster.maker.df[,tech.indicators.names], k=cluster.count)
pam.cluster.maker.df$cluster_tech = factor(pam.cluster.maker.df$cluster)

# Concatenate Cluster columns to `cluster.maker.df`
cluster.maker.df <- cbind(cluster.maker.df, cluster_tech = pam.cluster.maker.df$cluster_tech)

# Convert to df
cluster.maker.df <- as.data.frame(cluster.maker.df)
```

Since my objective is to find natural groupings of countries, I need to discover which are the country attributes that overlap most with the country clusters. More specifically, I use the variables `GDP_per_capita`, `internet_users` and `fertility_rate` and study how well those variables allow me to form cluster of countries in terms of technological advancement and well-being.

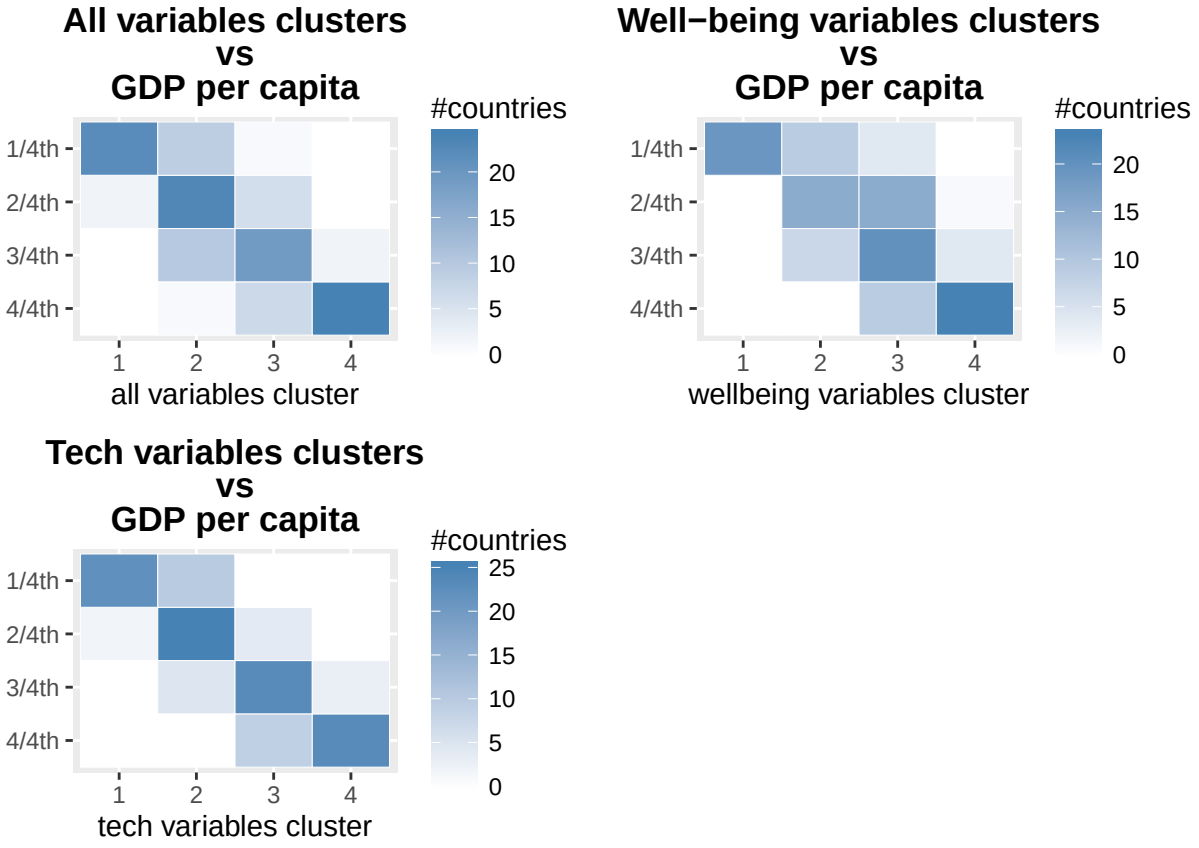
I need to classify each country in its respective quartile for the variables `GDP_per_capita`, `internet_users` and `fertility_rate`.

```
# Create factor var
cluster.maker.df %>%
  mutate(
    GDP_per_capita_factor = quantilize(GDP_per_capita, gdp.quantilize.count),
    internet_users_percent_factor = quantilize(internet_users_percent, internet_users.quantilize.count)
    fertility_rate_factor = quantilize(fertility_rate, fertility_rate.quantilize.count)
  ) %>%
  { . } -> cluster.maker.df
```

Finally, I can create the country distribution heat map that shows the how strong are the overlaps between clusters and selected variables.

Using GDP per capita to form natural groups

The first set of heat maps is an investigation of the relationship between the clusters and the GDP per capita.



The heat map is a graphical evidence of how well the GDP per capita overlaps with the natural clusters of countries for the selected set of variables.

The first heat map compares categories of countries by economic performance against clusters formed using the entire variable set. It is clear that each cluster overlap strongly with each group of countries for each quartile of GDP per capita.

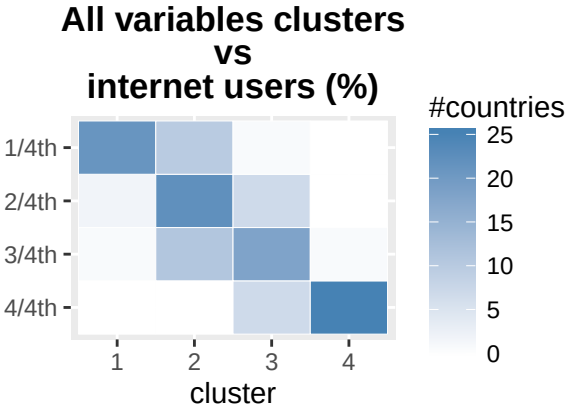
The heat map #2 only uses well-being variables to form clusters whereas heat map #3 only uses technology variables to form clusters. In both heat maps #2 and #3, the first and last quartiles of countries in terms of GDP per capita overlap completely with the cluster #1 and #4 which indicates that countries with high economic performance are the same countries that have high well-being and high technological development and the opposite is true for countries with low economic performance.

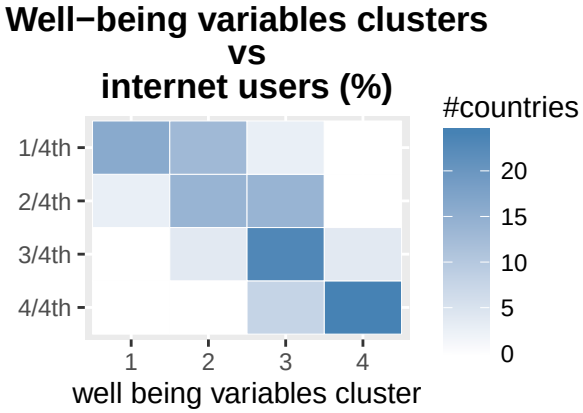
We can observe that the categories of economic performance overlap more clearly with the clusters formed using technological development indicators than with the clusters formed using well-being indicators. Indeed, the overlap is less clear for the 2nd and 3rd quartiles of countries in heatmap#2 than it is in heatmap#3.

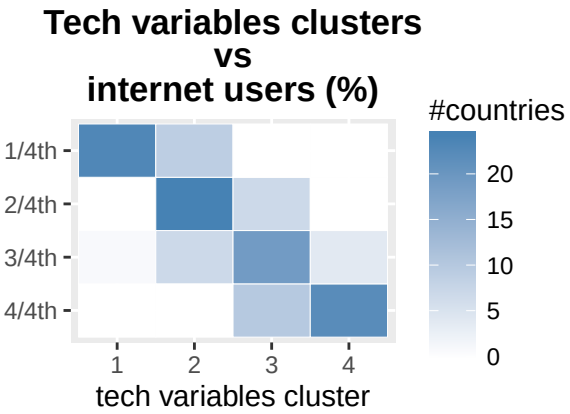
However, we can infer from this set of heat maps that GDP per capita is a key indicator that can be used to characterize fairly accurately groups of countries in terms of both technological development and/or level of well-being.

Using internet usage to form natural groups

The second set of heat maps is an investigation of the relationship between clusters and proportion of internet users in a country.





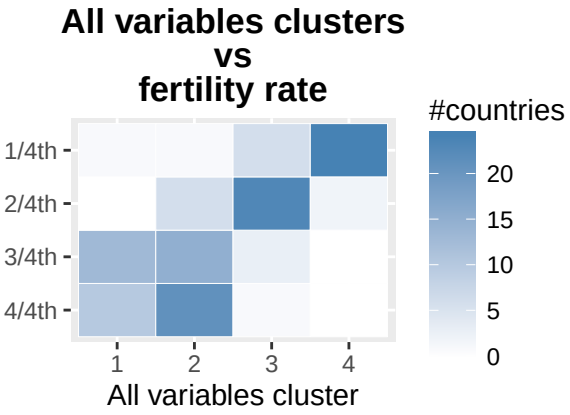


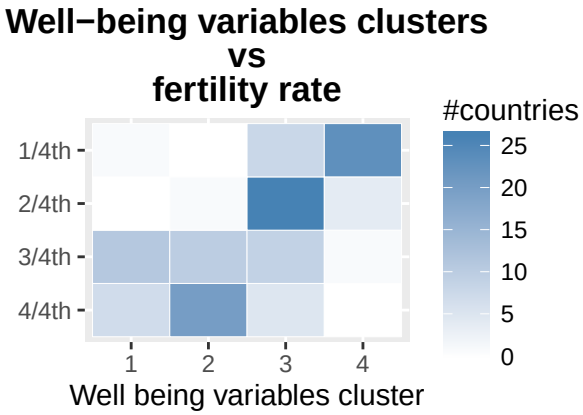
Similarly to the GDP per capita, the groups of countries in terms of their proportion of internet users overlap with the PAM clusters very well, especially for Cluster#1 and Cluster#4. Those heat maps support the claim that there is a distinct group of countries with high GDP per capita and high internet users percentage, and another group of countries with low GDP per capita and low internet users percentage.

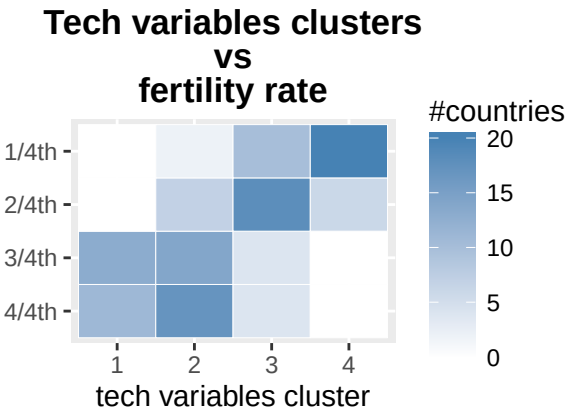
Not surprisingly, internet usage characterizes better the groupings of countries in terms of their technological development than it does groupings of countries in terms of their level of well-being. However, when looking at heat map #2, one cannot deny the relationship between well-being and internet usage.

Using fertility rate to form natural groups

The third heat map is an investigation of the relationship between cluster and fertility rate.



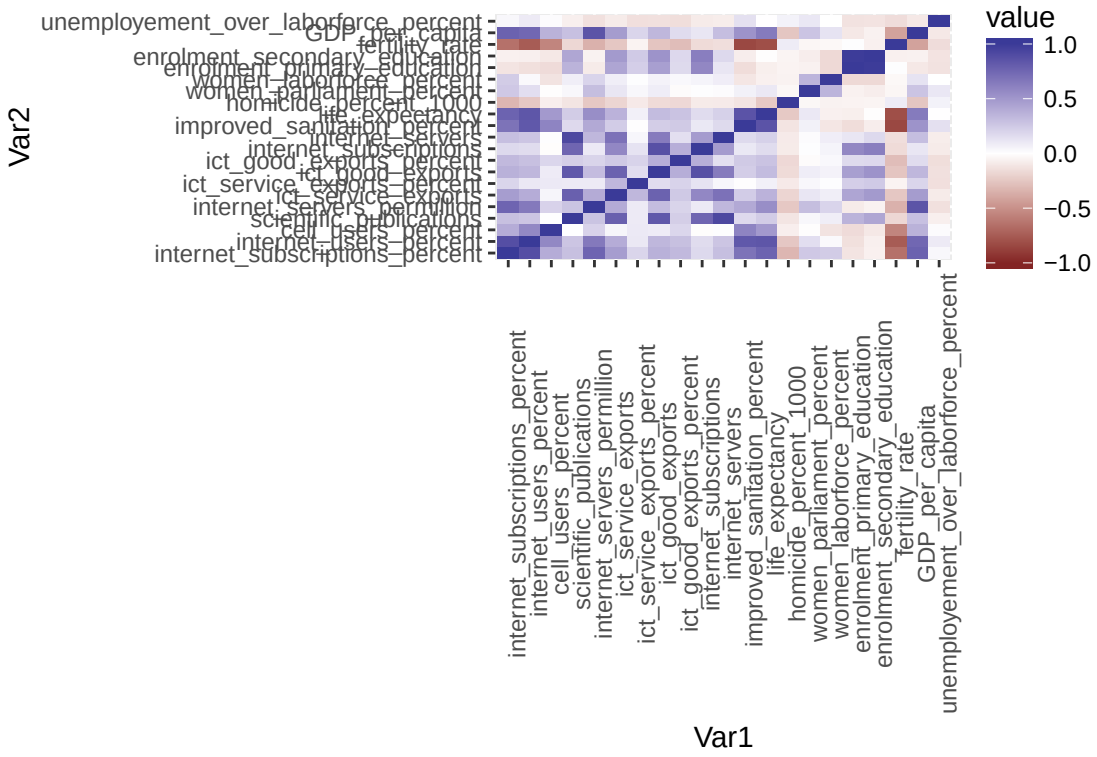




The fertility rate is not as precise an indicator of technological development and well being as GDP per capita or internet usage are, especially for the 50% of countries with lowest fertility rate. However,this set of heat maps, demonstrate a clear overlap between the top two quartiles of countries by fertility rate with cluster#3 and cluster#4. It is clear that the countries with lowest well being and lowest technological development are also the top countries in terms number of children per women.

Cluster plots

Another approach to analyzing the indicators set is to study their correlation with each other. Below is a heat map providing a visual representation of the correlation matrix for the indicators.



`life_expectancy` presents interesting correlation with other numeric variables. First, I observe that `life_expectancy` correlates positively with `sanitation_percent`, which is the percentage of the population with access to improved sanitation facilities. This suggests that the more a country's population has access to sanitation facilities, the highest is the average life expectancy in the country.

Second, it also correlates positively with `internet_users_percent` which is the percentage of the population with access to internet. This introduces the idea that the degree of adoption of information technology by the population provides information about the average life expectancy in the country.

Finally, `life_expectancy` correlates negatively with the `fertility_rate` which supports the idea that the lower a country's average life expectancy, the more children women have.

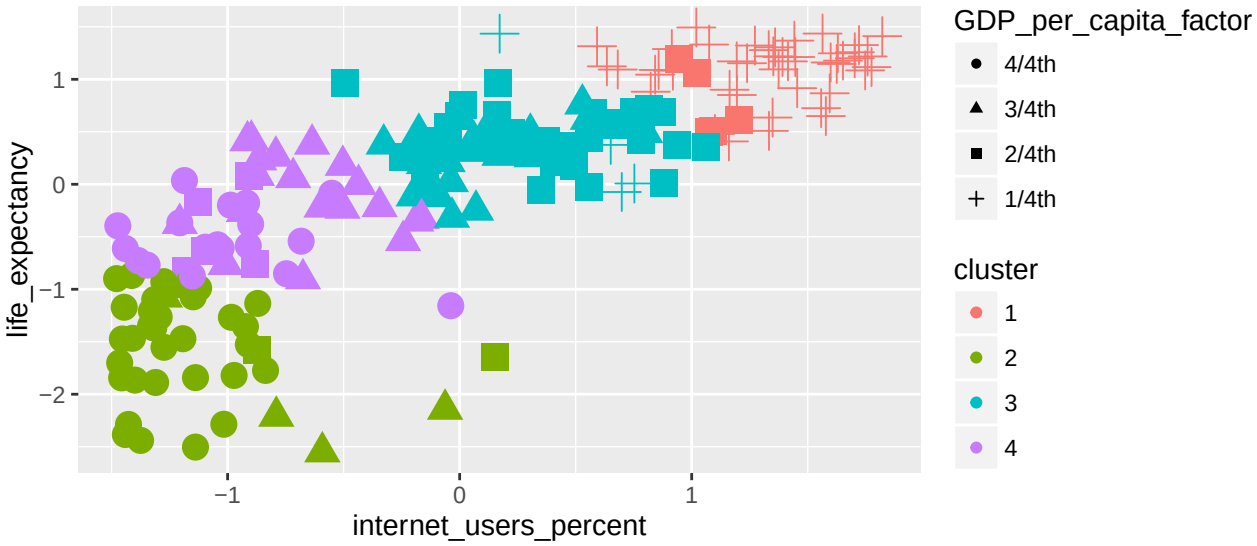
As a preliminary treatment to clustering, I normalized the variable data using the following function:

```
normalize1 = function(vec) (vec-mean(vec, na.rm=TRUE))/sd(vec, na.rm=TRUE)
```

The clustering technique used in the following subsection is PAM: partitioning around medoids.

Life expectancy versus Internet users

The first cluster analysis is based on the health variable `life_expectancy` and the technology variable `internet_users`.

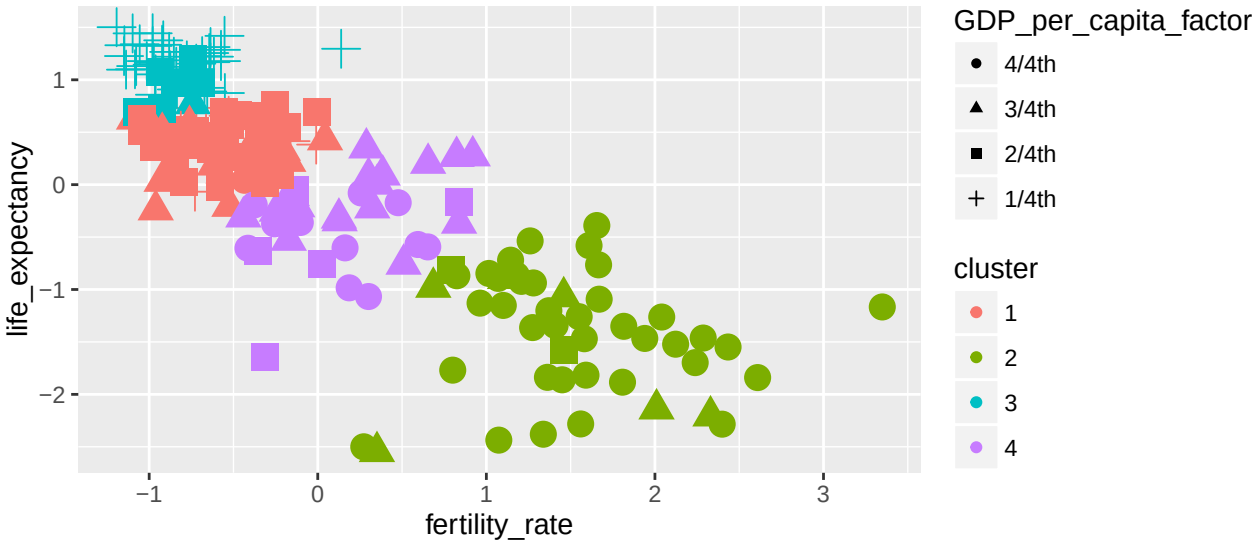


The groupings observed differentiate four groups of countries. The analysis suggests that GDP per capita is a driver of both internet adoption and life expectancy.

Cluster#1 consists of countries with high adoption of the internet and high life expectancy. We can note that this cluster strongly overlap with the group of countries that are in the top quartile in terms of GDP per capita. Cluster#2 represent countries with medium-high internet adoption and life expectancy. Those countries are in the 2nd and 3rd quartile in terms of economic performance. Cluster#3 represent countries with medium-low internet adoption and life expectancy. Although some of cluster#3 countries belong to the 2nd quartile in terms of economic performance, most of them are part of the 3rd or 4th quartile. Finally, cluster#4 groups the countries with lowest internet adoption and lowest life expectancy. The large majority of those countries is part of the 25% last countries in terms of GDP per capita.

Life expectancy versus fertility rate

The second cluster analysis is based on `life_expectancy` and `fertility_rate`. I am using the PAM clustering technique.



The distribution of points on this scatter plot demonstrates the negative linear correlation between fertility rate and life-expectancy. Cluster#3 consists almost exclusively of countries with highest GDP per capita. These countries all have low fertility rate and high life expectancy. Cluster#1 consists mostly of the 2nd quartile countries in terms of economic performance. Those countries have slightly higher fertility rate but lower life expectancy than cluster#3 countries. Cluster#2 is a grouping of countries characterized by their low life expectancy and high fertility rate. The countries in cluster#2 are mostly the 25% last in terms of economic performance.

Conclusion

The objective of this research was to find evidence of a relationship between technological development and well-being. I found that internet usage in countries is one of the most relevant indicators to characterize a country's level of technological development. This indicator also characterizes the level of well-being in a country. Therefore, there is an undeniable evidence of relationship between technological development and well-being.

However, I discovered that GDP per capita characterizes both the level of technological development and the level of well-being in countries in a way that is more accurate than when using internet usage. This suggests that the relationship between economic performance and well-being is more direct than that between well-being and technology.