

ANALISI DATASET “NATIONAL LIFE EXPECTANCIES”

Progetto d'esame conclusivo del corso di studi in Data Analytics anno 2023/2024 a cura di

Bortolussi Lorenzo
Bosco Francesco
Bredariol Francesco
Tonet Lorenzo

Introduzione

descrizione del dataset e definizione dell'obiettivo dell'analisi

Descrizione Dataset

Il nostro dataset, “National Life Expectancies”, è un dataset che contiene informazioni riguardo 185 Paesi nel Mondo. Queste informazioni passano da semplici descrizioni demografiche sino a dettagli geopolitici e la loro natura è quasi completamente numerica, poiché questi dati sono per lo più indici e percentuali.

La fonte dei dati è lo “United Nations Human Development Report”.

Obiettivo Analisi

Partendo dai dati in nostro possesso la nostra analisi ha l'obiettivo di trovare le relazioni più forti all'interno del dataset avendo come variabile target l'indice “Aspettativa di Vita”. Questa ricerca di relazioni sarà preceduta dall'analisi delle singole variabili per poter ottenere a priori informazioni che potrebbero poi rivelarsi importanti. Obiettivo ultimo sarà quello di raggruppare i paesi in cluster di “benessere di vita” sfruttando le informazioni desunte dal lavoro svolto in precedenza.

Variabili Dataset

descrizione variabili dataset e prime osservazioni

Descrizioni Variabili

Andiamo a descrivere le singoli variabili, visualizzandone nome, numero di osservazioni mancanti ed una breve descrizione

VARIABILE	OSSERVAZIONI MANCANTI	DESCRIZIONE
REGION	0	Variabile categoriale che indica una regione associata
COUNTRY	0	Nome del Paese
LIFEEXP	0	Aspettativa di vita dalla nascita misurata in anni
ILLITERATE	14	Tasso di analfabetismo percentuale negli adulti (età > 15 anni)
POP	1	Popolazione nel 2005 in milioni di abitanti
FERTILITY	4	Tasso di fertilità, nascite per donna
PRIVATEHEALTH	1	Spesa sanità 2004 misurata in %GDP
PUBLICEDUCATION	28	Spesa educazione misurata in %GDP
HEALTHEXPEND	5	Spesa sanitaria pro capite 2004 misura in USD
BIRTHATTEND	7	% Nascite assistite da personale specializzato
PHYSICIAN	3	Medici per 100,000 abitanti
SMOKING	88	% Fumatori maschi adulti
REASEARCHERS	95	Ricercatori in Ricerca e Sviluppo per Milioni di abitanti
GDP	7	Prodotto Interno Lordo in miliardi di USD
FEMALEBOSS	87	% di donne con posizioni di dirigenza

Osservazioni

Notiamo subito che sarà per noi necessaria la gestione delle osservazioni mancanti. Per esempio dobbiamo decidere come trattare la variabile “FEMALEBOSS” che ha un tasso di NA prossimo al 50%. Nel prossimo paragrafo ci soffermeremo dunque su questo aspetto e sulla scelta della fattorizzazione di alcune variabili quali potrebbero essere REGION.

Bonifica dei dati

gestione NA e fattorizzazione

La primissima operazione da compiere sarà il caricamento dei dati da file CSV. Per comodità abbiamo nominato il file "life expectation.csv". Nella lettura ci ricordiamo di settare header = TRUE. Eseguiamo subito una summary per poter trarre le prime conclusioni.

Prepariamo il nostro ambiente R importando le librerie necessarie e definendo dei colori utili alla visualizzazione dei grafici

```
library(patchwork)
library(ggplot2)
library(ggExtra)
library(moments)
library(sf)
```

```
## Linking to GEOS 3.11.2, GDAL 3.8.2, PROJ 9.3.1; sf_use_s2() is TRUE
```

```
library(rnaturalearth)
library(dplyr)
```

```
##
## Caricamento pacchetto: 'dplyr'
```

```
## I seguenti oggetti sono mascherati da 'package:stats':
##
##     filter, lag
```

```
## I seguenti oggetti sono mascherati da 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(MASS)
```

```
##
## Caricamento pacchetto: 'MASS'
```

```
## Il seguente oggetto è mascherato da 'package:dplyr':
##
##     select
```

```
## Il seguente oggetto è mascherato da 'package:patchwork':
##
##     area
```

```
set.seed(1235)
```

```
verde_acqua = "#64F9C7"
giallo = "#FFFF00"
arancione = "#FFB600"
verde_acqua_scurito = "#3da07f"
arancione_scurito = "#D69A01"
black ="#000000"

col_map = c("1" = "#64F9C7", "2" = "#FFFF00", "3" = "#FFB600", "4" = "#ee6055", "5" = "#57f054", "6" = "#c73596",
"7" = "#a26769", "8" = "#7ca5b8")
```

```
data = read.csv("life expectation.csv", header = TRUE)
original_data = data
summary(data)
```

```

##      REGION      COUNTRY      LIFEEXP      ILLITERATE
## Min.   :1.000  Length:185  Min.   :40.50  Min.   : 0.20
## 1st Qu.:3.000  Class  :character  1st Qu.:59.70  1st Qu.: 1.35
## Median :6.000  Mode   :character  Median :71.00  Median :10.10
## Mean    :4.692                           Mean   :67.05  Mean   :17.69
## 3rd Qu.:7.000                           3rd Qu.:75.10  3rd Qu.:27.50
## Max.    :8.000                           Max.   :82.30  Max.   :76.40
##                                     NA's   :14
##      POP        FERTILITY     PRIVATEHEALTH  PUBLICEDUCATION
## Min.   : 0.10  Min.   :0.90  Min.   :0.300  Min.   : 0.600
## 1st Qu.: 2.45  1st Qu.:1.80  1st Qu.:1.500  1st Qu.: 3.400
## Median : 7.80  Median :2.70  Median :2.400  Median : 4.600
## Mean   : 35.36  Mean   :3.19  Mean   :2.517  Mean   : 4.695
## 3rd Qu.: 23.60 3rd Qu.:4.30  3rd Qu.:3.300  3rd Qu.: 5.800
## Max.   :1313.00 Max.   :7.50  Max.   :8.500  Max.   :13.400
## NA's   :1       NA's   :4       NA's   :1       NA's   :28
##      HEALTHEXPEND  BIRTHATTEND      PHYSICIAN      SMOKING
## Min.   : 15.0  Min.   : 6.00  Min.   : 2.00  Min.   : 6.00
## 1st Qu.: 100.5 1st Qu.: 61.00  1st Qu.: 21.25  1st Qu.:24.00
## Median : 297.5  Median : 92.00  Median :107.50  Median :32.00
## Mean   : 718.0  Mean   : 78.25  Mean   :146.08  Mean   :35.09
## 3rd Qu.: 727.0 3rd Qu.: 99.00  3rd Qu.:247.00  3rd Qu.:47.00
## Max.   :6096.0  Max.   :100.00  Max.   :591.00  Max.   :68.00
## NA's   :5       NA's   :7       NA's   :3       NA's   :88
##      RESEARCHERS      GDP        FEMALEBOSS
## Min.   : 15.0  Min.   : 0.10  Min.   : 2.00
## 1st Qu.: 127.2 1st Qu.: 3.55  1st Qu.:23.25
## Median : 848.0  Median : 14.20  Median :30.00
## Mean   : 2034.7  Mean   : 247.55  Mean   :29.07
## 3rd Qu.: 2538.0 3rd Qu.: 109.28 3rd Qu.:36.75
## Max.   :45454.0  Max.   :12416.50  Max.   :58.00
## NA's   :95      NA's   :7       NA's   :87

```

Fattorizzazione

COUNTRY

Il summary ci evidenzia subito come la variabile COUNTRY sia un character e per questo motivo decidiamo di trattarla come factor. Dobbiamo sottolineare tuttavia che questa variabile non ha alcun valore duplicato e possiamo già dedurre che non è di rilievo ai fini della nostra analisi.

Potrebbe essere interessante trasformare questa variabile fattoriale in una variabile di tipo “posizione geografica”, tuttavia decidiamo di sfruttare a questo fine la variabile REGION.

```
data$COUNTRY = factor(data$COUNTRY)
```

REGION

REGION è una variabile che descrive l'appartenenza dei Paesi a cluster geografici ed è questo il motivo per cui decidiamo di convertirla a factor.

```
data$REGION = factor(data$REGION)
```

Gestione NA

Per quanto riguarda gli NA decidiamo di eseguire un approccio in base alla percentuale di quest'ultimi e al tipo di analisi da eseguire. Nei casi delle analisi univariate rimuoviamo gli NA; nei casi delle analisi bivariate valutiamo la percentuale di NA della seconda variabile (ricordiamo che la variabile target da noi considerata è sempre LIFEEXPECT): in caso questa sia troppo alta (>30%) non riterremo l'analisi attendibile e quindi la eviteremo, mentre se questa sarà sufficientemente bassa (<30%) rimuoveremo semplicemente le righe che presentano gli NA.

Il codice non verrà presentato ora poiché sarà nelle singole analisi che verrà effettuato questo passaggio di pulizia. Diamo comunque uno sguardo alla distribuzione degli NA, nel nostro dataset.

Ottieniamo una tabella in cui vengono indicati, in base al paese, quanti NA sono presenti per ogni variabile

```

country_na_counts = original_data %>%
  group_by(COUNTRY) %>%
  summarise(across(everything(), ~sum(is.na())))
  
country_na_counts

```

```

## # A tibble: 185 × 15
##   COUNTRY      REGION LIFEEXP ILLITERATE POP FERTILITY PRIVATEHEALTH
##   <chr>        <int>    <int>     <int> <int>    <int>       <int>
## 1 Afghanistan      0       0       0       0       0       0
## 2 Albania          0       0       0       0       0       0
## 3 Algeria          0       0       0       0       0       0
## 4 Angola           0       0       0       0       0       0
## 5 Antigua and Barbuda 0       0       0       0       1       0
## 6 Argentina         0       0       0       0       0       0
## 7 Armenia           0       0       0       0       0       0
## 8 Australia          0      10       0       0       0       0
## 9 Austria            0       0       0       0       0       0
## 10 Azerbaijan        0       0       0       0       0       0
## # i 175 more rows
## # i 8 more variables: PUBLICEDUCATION <int>, HEALTHEXPEND <int>,
## #   BIRTHATTEND <int>, PHYSICIAN <int>, SMOKING <int>, RESEARCHERS <int>,
## #   GDP <int>, FEMALEBOSS <int>

```

Ottieniamo la somma degli NA per paese e visualiziamone la testa della classifica.

```

country_na_counts = original_data %>%
  group_by(original_data$COUNTRY) %>%
  summarise(Total_NA = sum(rowSums(is.na(across(-COUNTRY)))))

country_na_counts = country_na_counts[order(country_na_counts$Total_NA, decreasing=TRUE), ]
country_na_counts

```

```

## # A tibble: 185 × 2
##   `original_data$COUNTRY`      Total_NA
##   <chr>                      <dbl>
## 1 Somalia                     7
## 2 Hong Kong, China (SAR)      5
## 3 Korea (Democratic People's Rep. of) 5
## 4 Occupied Palestinian Territories 5
## 5 Saint Kitts and Nevis       5
## 6 Afghanistan                 4
## 7 Comoros                     4
## 8 Congo (Democratic Republic of the) 4
## 9 Cuba                        4
## 10 Djibouti                   4
## # i 175 more rows

```

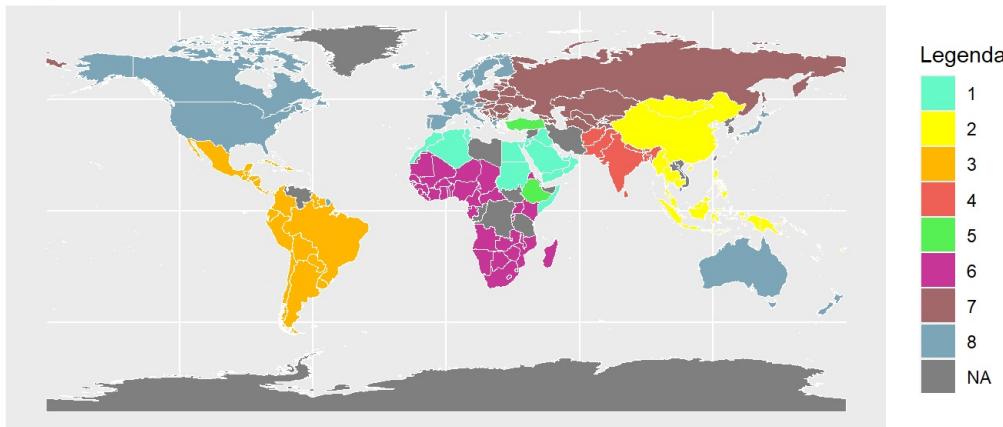
REGION

studio della suddivisione regionale

La prima analisi che andiamo ad effettuare, in maniera mirata, è l'analisi della composizione della variabile REGION. Il motivo di questa scelta è che, prima ancora di studiare la variabile, vogliamo comprendere come sono stati raggruppati i vari Paesi per poter comprendere se ci saranno comportamenti di gruppo giustificabili dalla REGION di appartenenza. Questo potrebbe poi portarci, nel caso, a studiare alcuni risultati a livello regionale anziché a livello statale, poiché questo può rendere più comprensibile il risultato dell'analisi se il comportamento statale presenta una variabilità troppo alta.

Vediamo una distribuzione non completamente uniforme delle Regioni, dunque decidiamo di visualizzare graficamente la cartina per capire meglio la suddivisione che è stata effettuata. Dobbiamo innanzitutto sistemare leggermente alcuni valori di COUNTRY per poter lavorare per poter abbinare loro a dei nomi validi per la .

Regioni



Scopriamo che non è stata effettuata una suddivisione totalmente geografica. In particolare sono le regioni 5 ed 8 ad essere non contigue, e immaginiamo dunque il motivo possa essere la struttura dell'organizzazione che raccoglie i dati. Comunque sia ora abbiamo uno sguardo più consci sulla struttura della variabile REGION e questo ci potrà sicuramente aiutare in futuro.

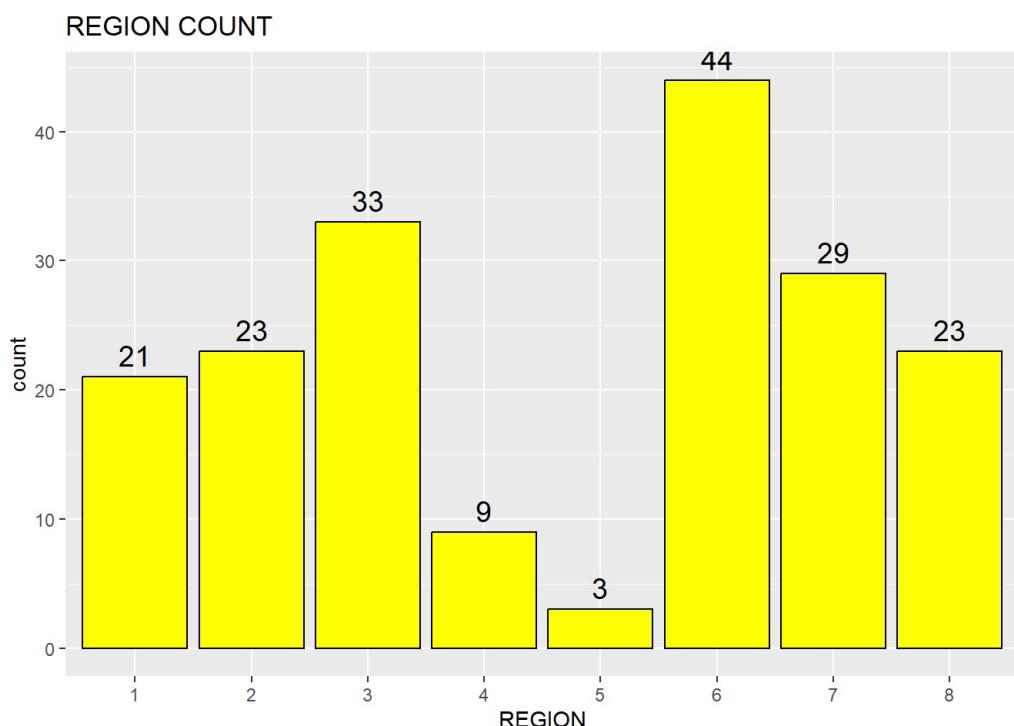
Analisi Univariata

delle singole variabili

In questo paragrafo affrontiamo lo studio univariato delle variabili. Abbiamo due tipi di variabile, le categoriali e le numeriche. Decidiamo tuttavia di non definire due tipi di analisi "prefissati" ma bensì ci lasciamo la libertà, variabile per variabile, di effettuare un analisi ad hoc, in cui in base ai risultati che troviamo decidiamo se esplorare con nuovi strumenti i dati.

REGION

La variabile REGION è una variabile categorica che indica la regione di appartenenza di un Paese. Le regioni in totale sono 8. Andiamo a visualizzare il conteggio dei paesi per regione.



Si può notare come il numero di paesi per regione sia molto variabile, per esempio la 5° ha solamente 3 misure, mentre la 6° ne ha 44. Principalmente le regioni meno popolate sono la 4° e la 5°. Sarà di interesse analizzare altre variabili condizionate alla regione per cercare eventuali relazioni.

COUNTRY

La variabile COUNTRY è anch'essa una variabile categorica, ma in questo caso indica il nome del paese quindi ha un'utilità più di "indice" che di effettiva variabile.

LIFEEXP

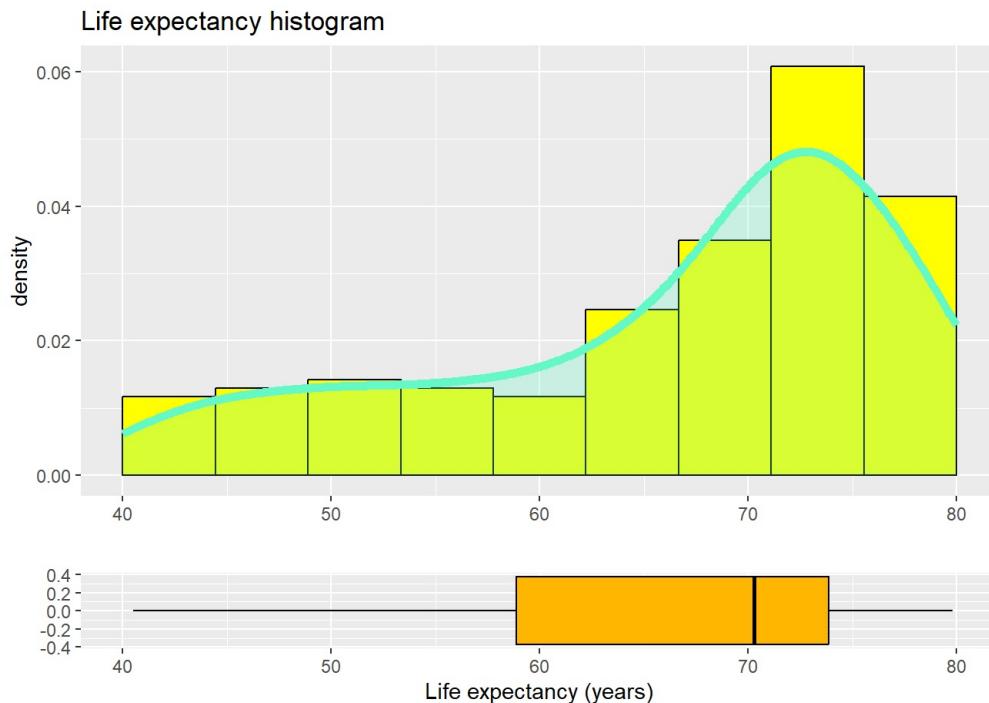
La variabile LIFEEXP indica l'aspettativa di vita di un determinato paese, misurata in anni. Andiamo a visualizzare il summary per avere un'idea della distribuzione.

```
summary(data$LIFEEXP)
```

```
##   Min. 1st Qu. Median   Mean 3rd Qu.   Max.  
##  40.50  59.70  71.00  67.05  75.10  82.30
```

Notiamo come lo spettro dei valori va da 40,5 a 82,3 quindi procediamo con un istogramma per visualizzare la distribuzione di tali dati. Come numero di colonne ci baseremo sul valore calcolato grazie alla formula di Sturges, facendo degli aggiustamenti dove necessario

```
sturges_k = nclass.Sturges(data$LIFEEXP)  
  
p1 = ggplot(data, aes(x = LIFEEXP, y = ..density...)) +  
  geom_histogram(bins = sturges_k + 1, fill = giallo, color = "black", boundary = 0) +  
  xlab("") +  
  ggtitle("Life expectancy histogram") +  
  geom_density(fill = verde_acqua, color = verde_acqua, alpha = 0.25, lwd = 2) +  
  xlim(c(40, 80))  
  
p2 = ggplot(data, aes(x = LIFEEXP)) +  
  geom_boxplot(fill = arancione, color = "black") +  
  xlab("Life expectancy (years)") +  
  xlim(c(40, 80))  
  
p1 + p2 + plot_layout(heights = c(6, 1))
```



Dal grafico e dal calcolo della misura dell'asimmetria si può notare come la distribuzione sia leggermente asimmetrica con la frequenza maggiore tra i 70 e gli 80 anni. La mediana della distribuzione è 71 anni, mentre la media è 67,05 anni. Il 50% della distribuzione si trova tra i 59,70 e i 75,1 anni. Per avere una misura dell'assimetria calcoliamo l'indice di skewness.

```
skewness(data$LIFEEXP)
```

```
## [1] -0.8408485
```

Skewness negativa indica che la coda della distribuzione è verso sinistra, come in questo caso.

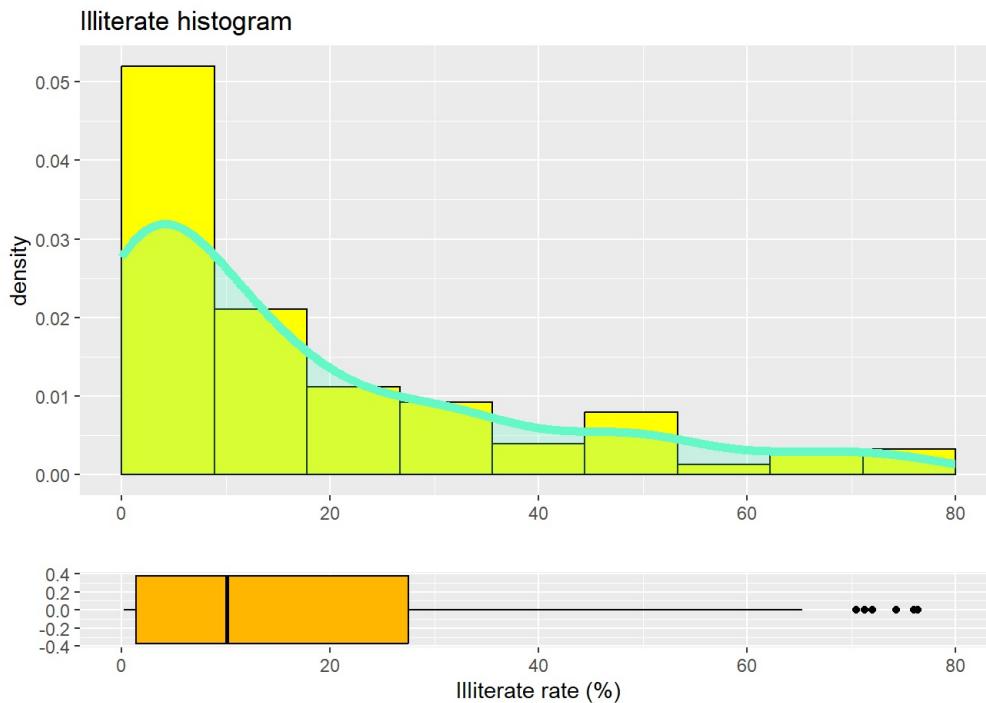
ILLITERATE

La variabile ILLITERATE indica la percentuale di analfabetismo negli adulti (età > 15 anni). Andiamo a visualizzare il summary.

```
summary(data$ILLITERATE)
```

```
##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.    NA's
##      0.20    1.35   10.10    17.69   27.50   76.40       14
```

Sono presenti 14 valori nulli, ovvero meno del 10% del totale. Procediamo con la visualizzazione della distribuzione.



Anche in questo caso notiamo una forte asimmetria, la maggior parte dei paesi ha un tasso di analfabetismo basso, inferiore al 30% mentre pochi paesi hanno un tasso molto alto. I paesi con valori considerati outlier sono (in ordine dal più alto): Burkina Faso, Mali, Chad, Afghanistan, Niger, Guinea. Notiamo come a parte per l'Afghanistan, tutti i paesi con tassi di analfabetismo molto alti facciano parte della regione 6.

```
skewness(data$ILLITERATE, na.rm = T)
```

```
## [1] 1.308789
```

La skewness positiva indica che la coda della distribuzione è verso destra, come in questo caso, il che non è un bene poiché indica un tasso di analfabetismo generalmente sopra la media.

POP

La variabile POP indica la popolazione nel 2004 in milioni di abitanti. Andiamo a visualizzare il summary.

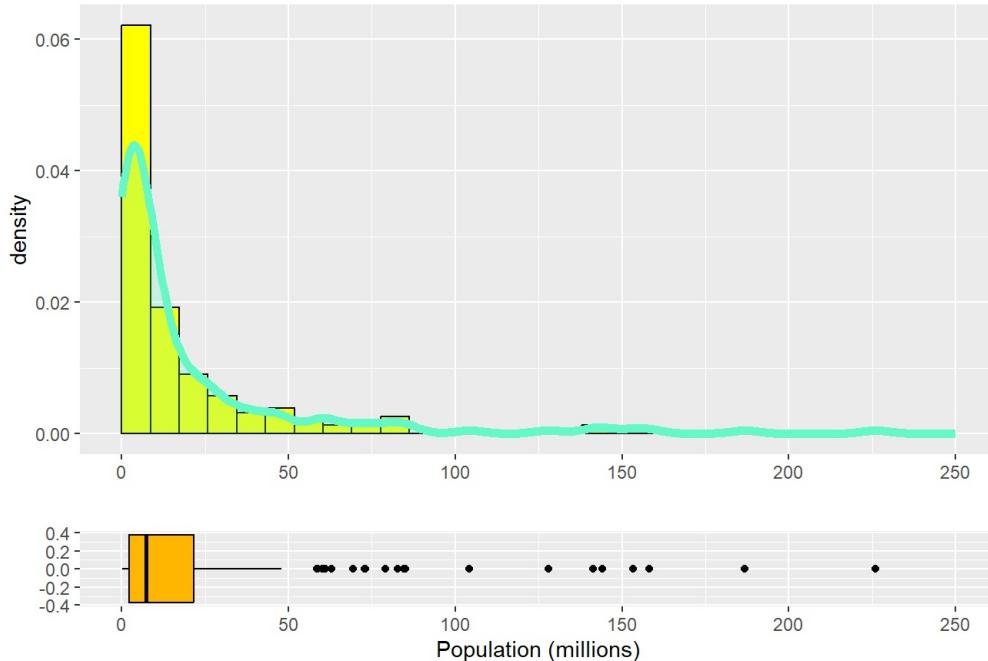
```
summary(data$POP)
```

```
##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.    NA's
##      0.10    2.45    7.80    35.36   23.60 1313.00       1
```

la presenza dell'unico valore mancante verrà ignorata nell'analisi. Procediamo con la visualizzazione della distribuzione.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Population histogram



A causa della fortissima dispersione delle misure di Cina e India, abbiamo deciso di tagliare il grafico per rendere leggibili i dati. La distribuzione resta comunque asimmetrica, tornando a evidenziare come la maggior parte dei paesi abbia una popolazione relativamente bassa mentre pochi paesi registrano numerosi abitanti

FERTILITY

La variabile FERTILITY indica il tasso di fertilità, ovvero il numero di nascite per donna. Andiamo a visualizzare il summary.

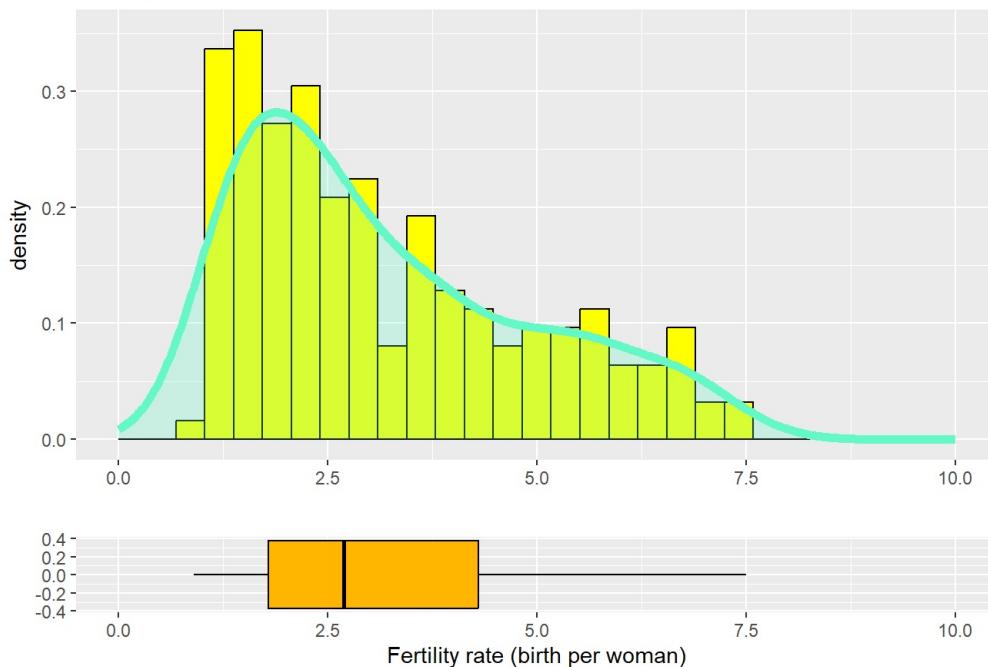
```
summary(data$FERTILITY)
```

```
##   Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
## 0.90  1.80  2.70  3.19  4.30  7.50  7.50     4
```

Procediamo con la visualizzazione.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Fertility histogram



Andiamo anche a calcolare l'indice di assimetria.

```
skewness(data$FERTILITY, na.rm = T)
```

```
## [1] 0.7742088
```

Anche in questo caso la distribuzione è asimmetrica, con la maggior parte dei paesi che hanno un tasso di fertilità basso, in media 3,15. La distribuzione è molto concentrata tra 1 e 5 figli per donna. La fertilità massima registrata è quella dell'Afghanistan con 7,5 figli per donna, mentre la minima è quella di Hong-Kong con 0,8 figli per donna.

PRIVATEHEALTH

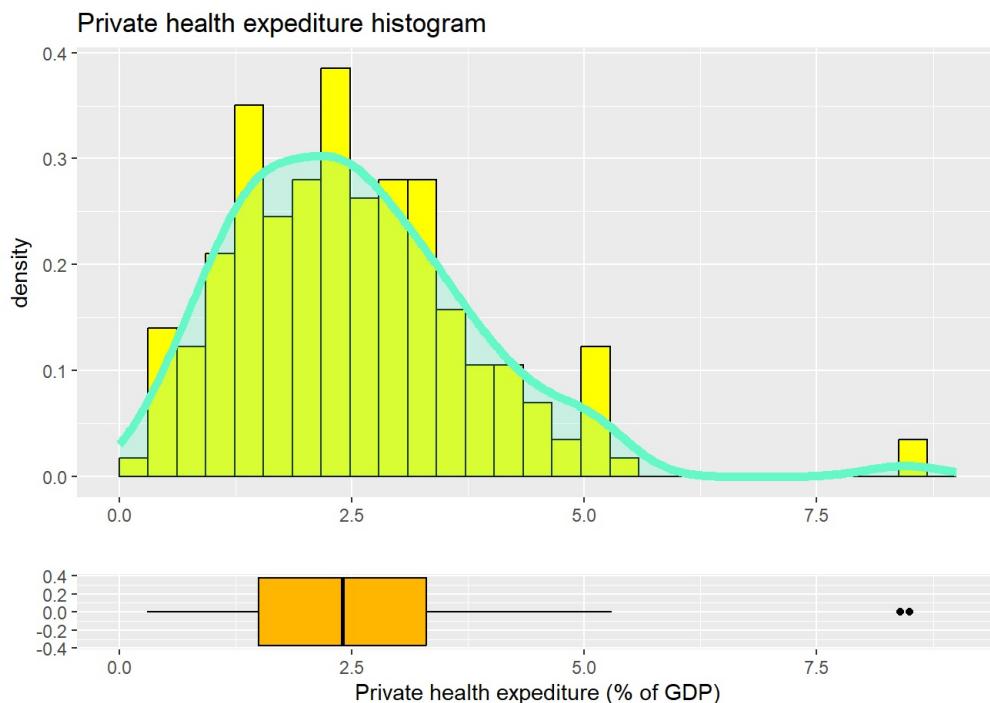
La variabile PRIVATEHEALTH indica la spesa per la sanità privata del 2004 misurata in %GDP. Andiamo a visualizzare il summary.

```
summary(data$PRIVATEHEALTH)
```

```
##   Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
## 0.300 1.500 2.400 2.517 3.300 8.500     1
```

la presenza dell'unico valore mancante verrà ignorata nell'analisi. Si può già notare come i valori non superino l'8,5% e che questo valore sia già abbastanza lontano dalla maggior parte dei dati guardando il primo e il terzo quartile. Procediamo con la visualizzazione della distribuzione per ricavare ulteriori informazioni.

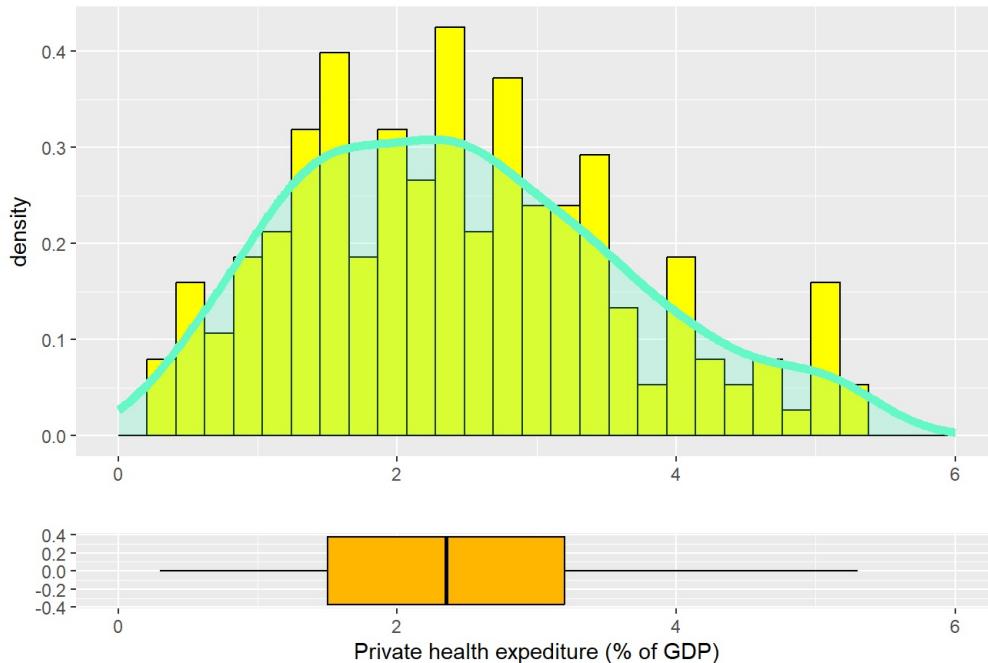
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Possiamo notare come la distribuzione sia concentrata principalmente tra 0 e 5% e come la maggior parte dei paesi abbia una spesa inferiore al 3% del GDP. La distribuzione è molto asimmetrica, con una coda verso destra. Vi sono presenti due outliers con i valori 8,5 e 8,4. Identificando questi due paesi si scoprono essere rispettivamente gli Stati Uniti e il Libano, ovvero due paesi in cui il sistema sanitario è altamente privatizzato e frammentato e i servizi medici gratuiti sono quasi inesistenti dunque la spesa privata è di conseguenza molto alta. Ora andiamo a vedere la distribuzione senza quei due valori.

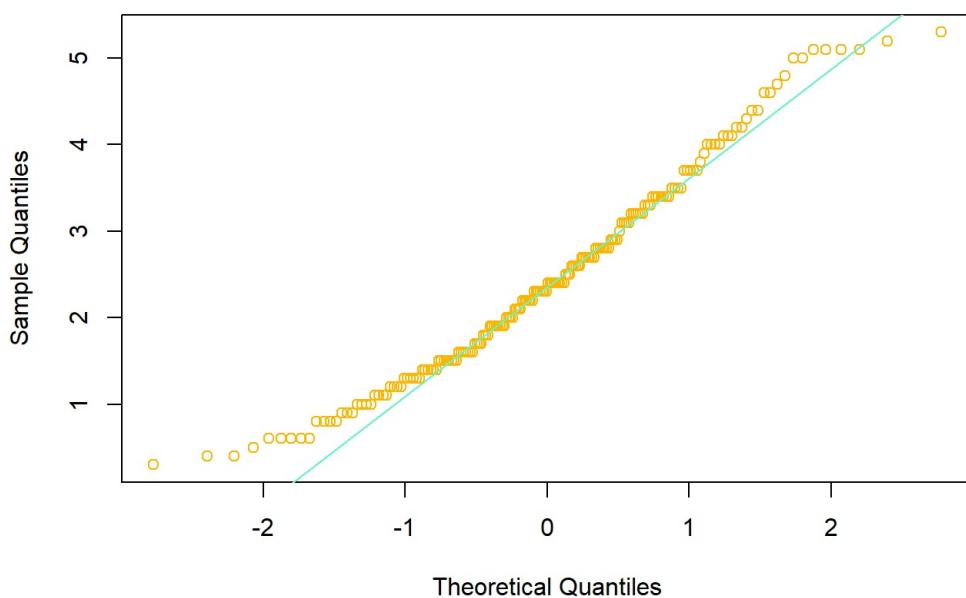
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Private health expenditure histogram

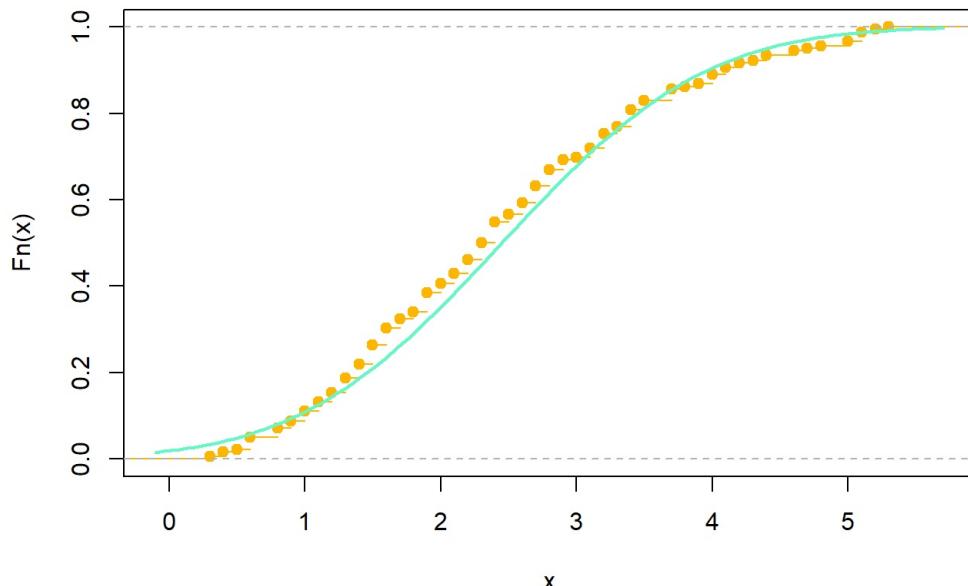


La distribuzione sembrerebbe seguire un andamento normale quindi si prova a confrontarla con una distribuzione normale teorica con media e deviazione standard calcolate sui dati. Per fare ciò si utilizza due grafici molto utili ovvero il quantile-quantile e la funzione di ripartizione empirica che ci permetterà di confrontare le distribuzioni.

Normal Q-Q Plot



Funzione di ripartizione empirica



Si nota come entrambi i grafici mostrino una forte aderenza alla distribuzione gaussiana, soprattutto nella funzione di ripartizione empirica. Questo ci fa pensare che la distribuzione sia effettivamente normale.

PUBLICEDUCATION

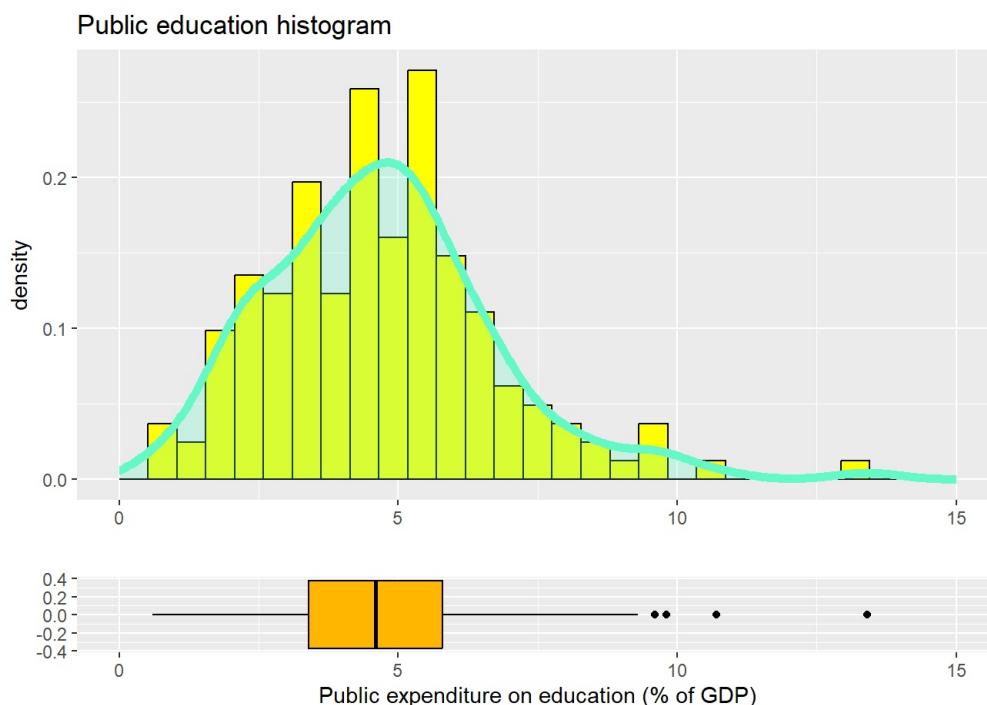
La variabile PUBLICEDUCATION indica la spesa per l'educazione pubblica del 2004 misurata in %GDP. Andiamo a visualizzare il summary.

```
summary(data$PUBLICEDUCATION)

##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.    NA's
## 0.600   3.400   4.600   4.695   5.800  13.400      28
```

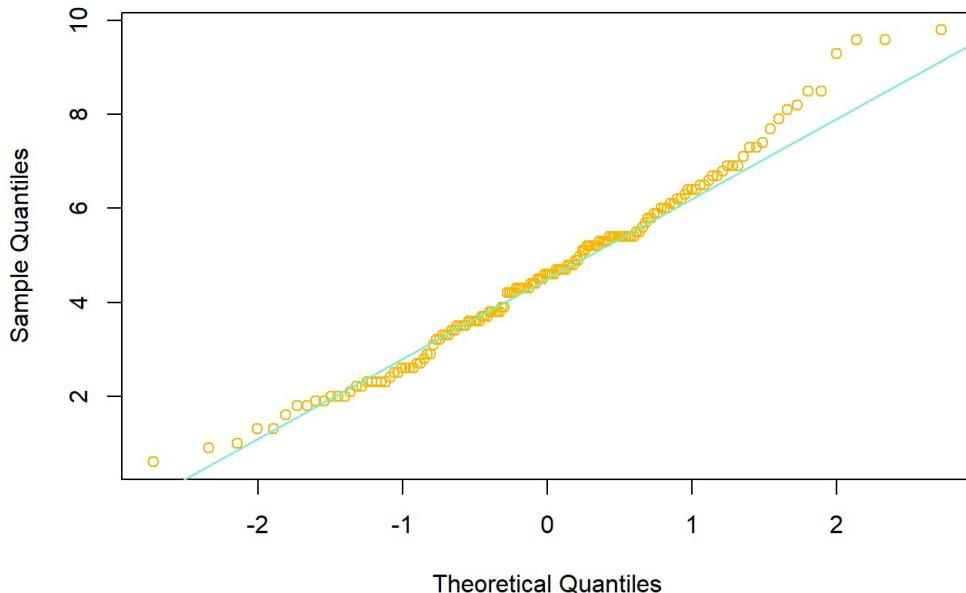
Si nota come in questa variabile la quantità relativa di valori nulli sia abbastanza alta (circa il 15% dei dati). Nonostante ciò si può vedere come la maggior parte dei dati presenti sia concentrata intorno ai valori di media e mediana, anch'essi molto vicini. Procediamo con la visualizzazione della distribuzione per ricavare ulteriori informazioni.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

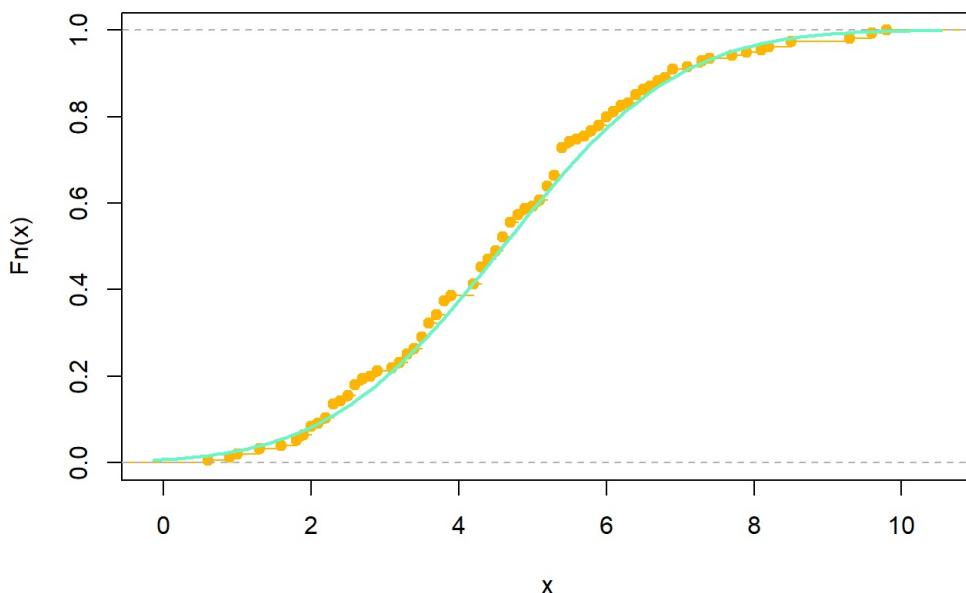


Anche per questa variabile, oltre alla presenza di 4 outliers si può notare la somiglianza ad una distribuzione normale, così come l'altra variabile che misura una percentuale di spesa rispetto al GDP. I 4 outliers sono in ordine Lesotho, Botswana, Cuba, Yemen quindi si può notare come i primi due valori più alti appartengano alla regione 6. Anche qui la distribuzione assomiglia abbastanza ad una normale, quindi si conducono i medesimi test rispetto alla variabile precedente.

Normal Q-Q Plot



Funzione di ripartizione empirica



Anche qui i grafici indicano una forte aderenza ad una distribuzione normale con media e deviazione standard calcolate sui dati (rimossi gli outlier).

HEALTHEXPEND

La variabile HEALTHEXPEND indica la spesa sanitaria pro capite del 2004 misurata in USD. Andiamo a visualizzare il summary.

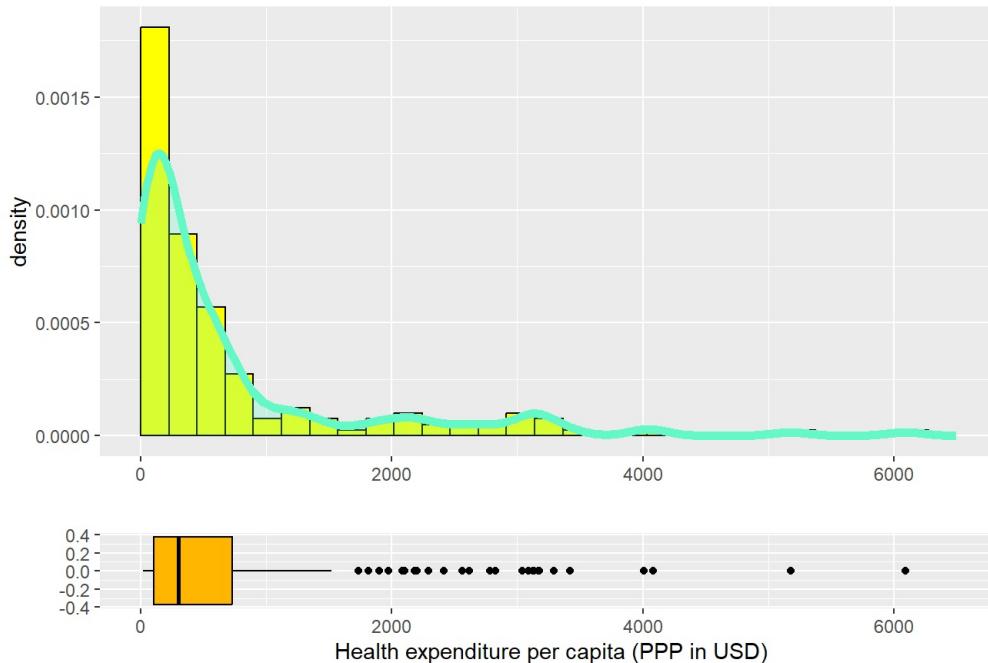
```
summary(data$HEALTHEXPEND)
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
##	15.0	100.5	297.5	718.0	727.0	6096.0	5

Già dal summary si può notare come la maggior parte dei dati sia tra i valori 100,5 e 727 nonostante il valore massimo osservato sia di 6096. Questo ci può già far intuire come la distribuzione sarà molto asimmetrica, probabilmente con una lunga coda a destra. Procediamo con la visualizzazione oppure in alternativa il massimo si tratterà di un outlier.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Health expenditure per capita histogram



Come avevamo previsto la maggior parte dei dati è concentrata sotto il valore 1000. Il valore più alto osservato corrisponde agli Stati Uniti come ci si può aspettare visto il sistema sanitario in questione seguito dal Lussemburgo e altri paesi europei. Calcoliamo l'indice di skewness per verificare l'assimmetria.

```
skewness(data$HEALTHEXPEND, na.rm = T)
```

```
## [1] 2.394425
```

Il valore fortemente positivo indica che la coda della distribuzione è verso destra, come previsto.

BIRTHATTEND

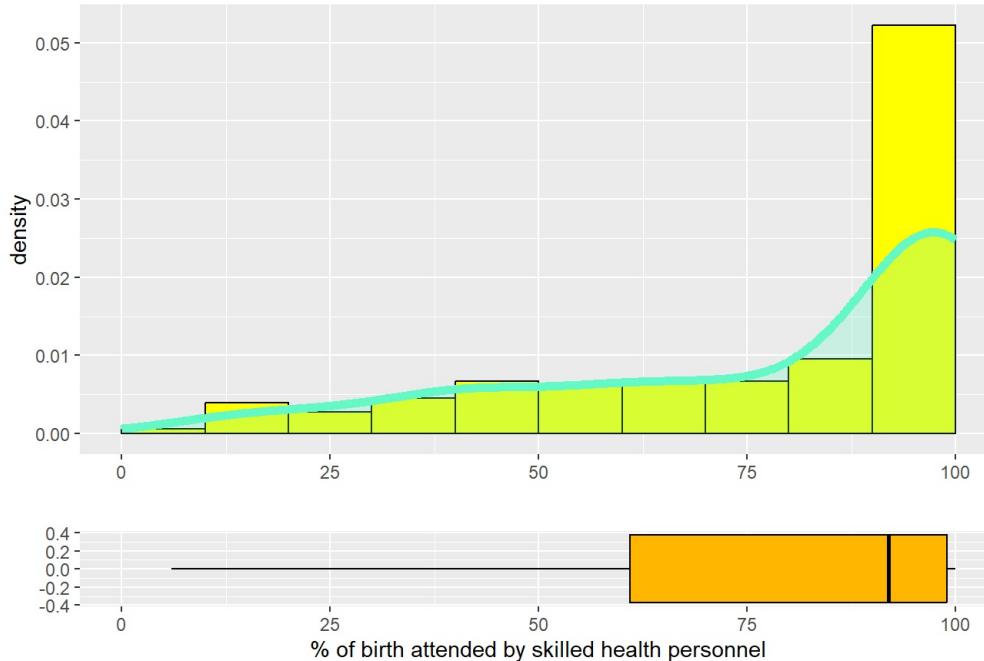
La variabile BIRTHATTEND indica la percentuale di nascite assistite da personale specializzato. Ci si aspetta che nella maggior parte dei paesi questa percentuale sia molto alta con pochi paesi in cui essa è molto bassa. Andiamo a visualizzare il summary.

```
summary(data$BIRTHATTEND)
```

```
##    Min. 1st Qu. Median    Mean 3rd Qu.    Max. NA's
##    6.00   61.00  92.00  78.25  99.00 100.00      7
```

Come ci si aspetta da una misura percentuale, la distribuzione è tra lo 0 e il 100% con una maggiore concentrazione nella parte alta della distribuzione ma con un valore minimo di 6, molto basso per questo tipo di variabile. Procediamo con la visualizzazione della distribuzione. Ci sono comunque 7 valori mancanti che verranno ignorati nell'analisi.

Birth attended by skilled health personnel histogram



I valori più bassi sono rispettivamente appartenenti alla regione 5 e tre appartengono alla regione 4, regioni che finora non sono mai spiccate particolarmente, sarà interessante cercare correlazioni aggiuntive. Si può dire che la distribuzione riaperti le ipotesi fatte in precedenza in quanto sia prevalentemente concentrata su valori vicini al 100.

PHYSICIAN

Questa è una variabile che crediamo abbia sicuramente legami con HEALTHEXPEND e PRIVATEHEALTH in quanto i medici presenti in un Paese sono uno degli aspetti principali che vengono considerati quando si parla di spesa medica (pubblica e non). Crediamo che possano esserci correlazioni con LIFEEXP in quanto, si sa, "un medico può salvarti la vita".

Andiamo a visualizzarne la percentuale di NA.

```
p_na = sum(is.na(data$PHYSICIAN))/length(data$PHYSICIAN)
p_na
```

```
## [1] 0.01621622
```

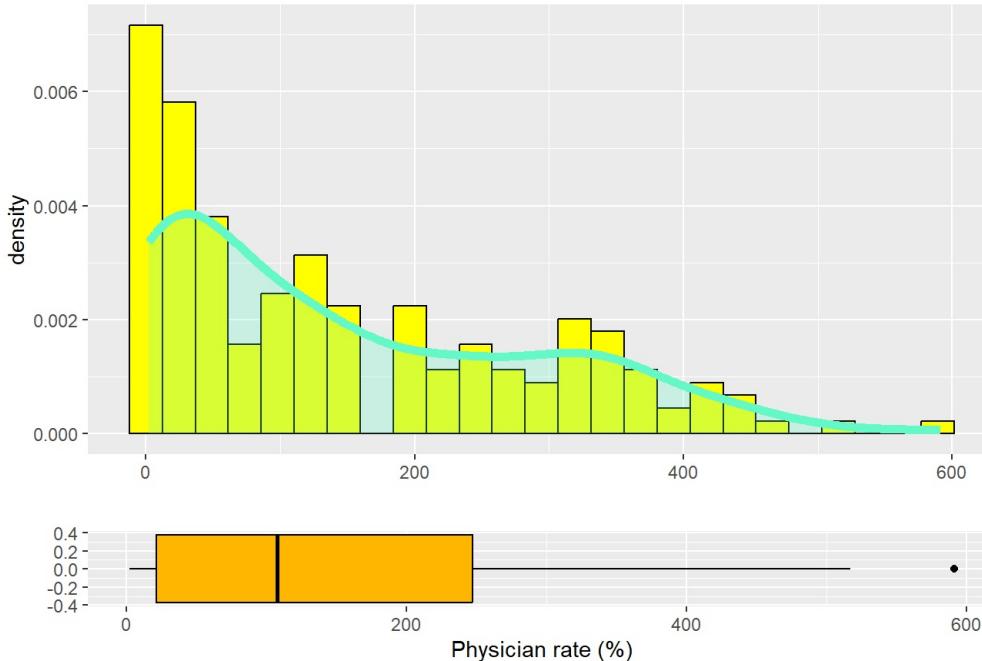
Ci troviamo davanti ad una delle variabili più "complete" del dataset, poiché da valore sia all'analisi univariate che alla bivariata che andremo poi ad eseguire con target LIFEEXP. Possiamo ripulire la variabile delle 3 misurazioni mancanti e cominciare l'analisi.

```
ph = na.omit(data$PHYSICIAN)
summary(ph)
```

```
##   Min. 1st Qu. Median    Mean 3rd Qu.    Max.
##   2.00  21.25 107.50 146.08 247.00 591.00
```

I valori sono abbastanza ampi e vediamo come il massimo sembri un outlier nei confronti del resto della distribuzione. Questo comportamento comunque non ci stupisce più molto, poiché come abbiamo già notato più volte in questo dataset la presenza di outlier verso destra è solo indice che alcuni Paesi hanno politiche molto più incentrate sulla variabile d'interessa rispetto agli altri. Visualizziamo i grafici

Physician histogram



Come previsto il massimo assume il significato di outlier destro per la nostra distribuzione anche se, ugualmente, abbiamo un range di valori molto densi che ricoprono bene l'intero dominio. Abbiamo comunque la maggiore concentrazione schiacciata verso il minimo, ma questa cosa non ci stupisce troppo dopo i risultati ottenuti da HEALTHEXPEND e PRIVATEHEALTH.

SMOKING

SMOKING potrebbe essere una variabile interessante, poiché, come è noto, il fumo è una delle cause di morte a livello globale, e dunque nell'analisi bivariata con target LIFEEXP ci possiamo aspettare qualcosa di interessante (tuttavia queste sono solo speculazioni a priori). Inoltre ricordiamo che SMOKING presenta il tasso di fumatori solo per la popolazione maschile e soprattutto è molto probabile che appaia come una distribuzione uniforme, poiché il vizio del fumo dovrebbe essere presente in egual modo in tutto il mondo.

Andiamo a visualizzarne la percentuale di NA.

```
p_na = sum(is.na(data$SMOKING))/length(data$SMOKING)
p_na
```

```
## [1] 0.4756757
```

La percentuale è davvero elevata e questo ci indica un'inaccuratezza della capacità descrittiva reale della distribuzione della variabile. Questo inoltre significa che difficilmente da un'analisi bivariata potremo ottenere risultati soddisfacenti.

```
sm = na.omit(data$SMOKING)
summary(sm)
```

```
##   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
##   6.00  24.00  32.00  35.09  47.00  68.00
```

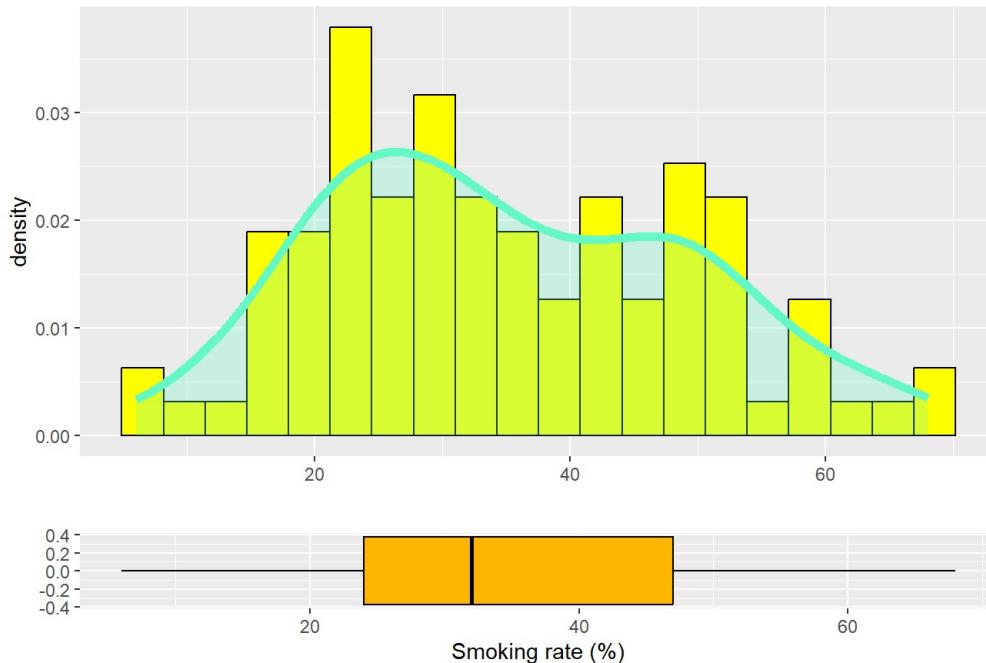
Dai dati sembra cadere l'ipotesi di uniformità della distribuzione, tuttavia media e mediana sembrano sufficientemente vicine e centrali per indicarci un'indice di simmetria vicino a 0(probabilmente leggermente spostato verso sinistra). Calcoliamolo.

```
idx = skewness(sm)
idx
```

```
## [1] 0.2857299
```

Nonostante la mediana alla sinistra della media l'indice di simmetria ci indica una leggera asimmetria verso destra. Dunque non ci resta che visualizzare il grafico per capire da cosa potrebbe essere dato questo comportamento.

Smoking histogram



Alla destra della media si presenta un leggero secondo picco che sbilancia la distribuzione e la rende leggermente asimmetrica, anche se si mantiene il picco principale tra media e mediana. I dati inoltre ci presentano una situazione abbastanza sensata, con un comportamento che tende verso la destra e non la sinistra, dunque ad una tendenza generale ad avere un tasso di fumatori più elevato. Crediamo che la mancanza del dato relativo alle donne sia abbastanza di rilievo, poiché LIFEEXP è indipendente dal sesso mentre questa variabile è fortemente dipendente e ciò rende molto più difficile riuscire a scovare correlazioni tra le due.

RESEARCHERS

RESEARCHERS è una variabile interessante, su cui non sappiamo granché esprimerci a priori. Possiamo ipotizzare valori abbastanza bassi e qualche outlier visto che nel mondo i fondi per la ricerca e lo sviluppo sono solitamente sottovalutati ad eccezione di alcuni Paesi che investono grandissimi quantità di denaro.

Valutiamo innanzitutto la percentuale di NA di RESEARCHERS

```
p_na = sum(is.na(data$RESEARCHERS))/length(data$RESEARCHERS)
p_na
```

```
## [1] 0.5135135
```

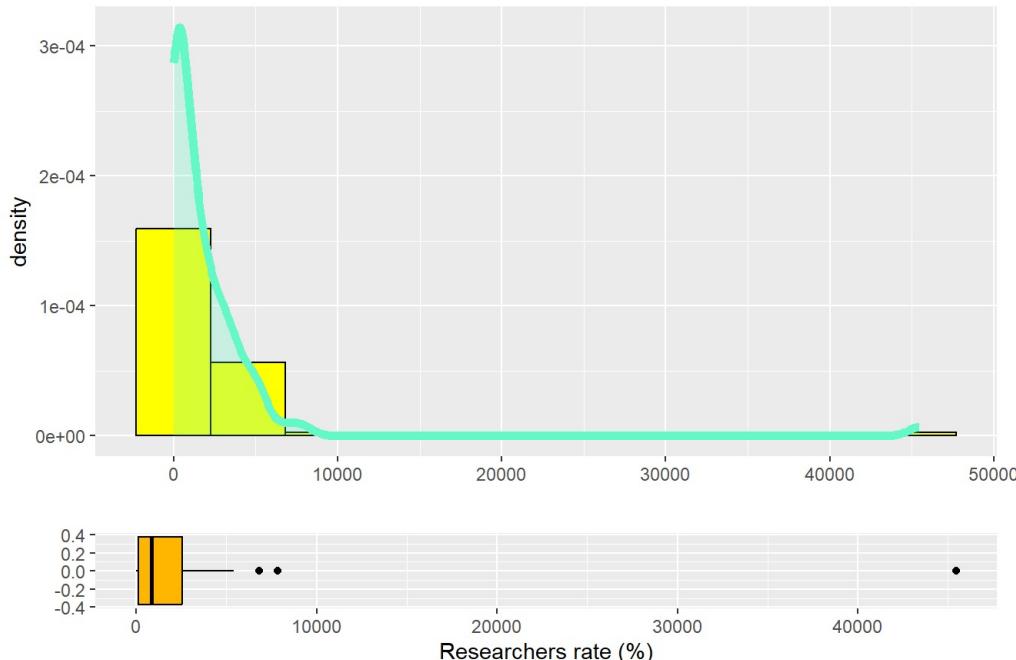
Si evince che la percentuale di NA è così elevata da coprire più della metà della variabile. Questo ci indica che l'analisi non sarà esaustiva del vero comportamento della variabile.

```
rs = na.omit(data$RESEARCHERS)
summary(rs)
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##      15.0   127.2  848.0  2034.7  2538.0 45454.0
```

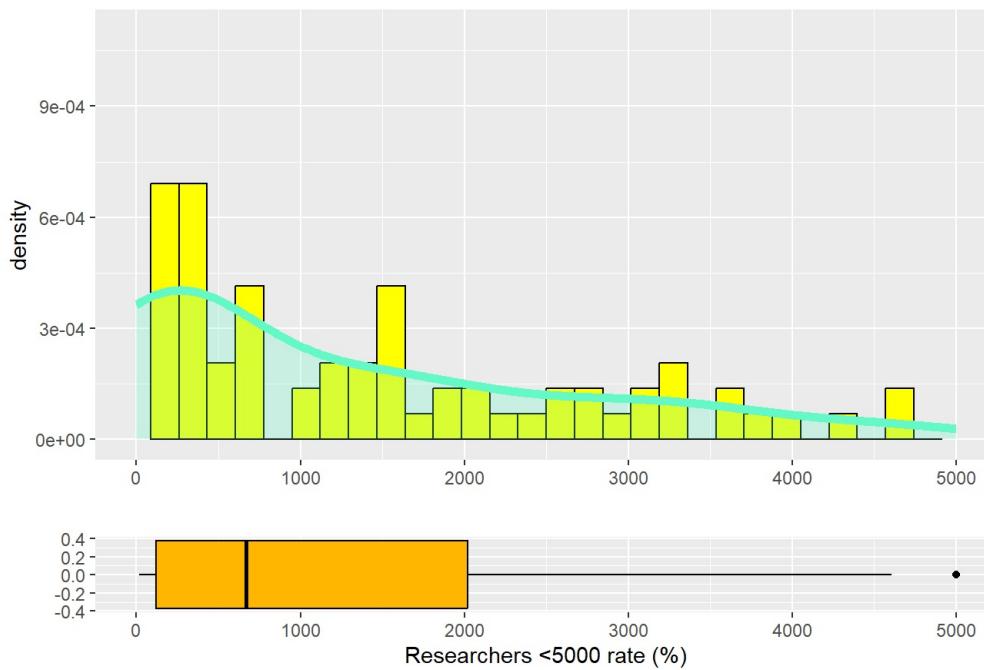
Media e Mediane sufficientemente lontane, range di valori molto vasto e massimo su tutta un'altra scala rispetto al resto. Probabilmente otterremo valori verso destra che ci andranno a rendere difficile la visualizzazione dei dati tramite grafici. Vedremo dunque poi se varrà la pena di circoscrivere l'analisi ad un range minore.

Researchers histogram



Come previsto, grafici pressoché senza valore per via di 3 outliers principali. Ci proponiamo allora di rivisualizzarli, con l'intento di trovare qualche comportamento interessante, studiando solo i valori minori di una certa soglia. Come soglia sceglieremo 5000, che dovrebbe tagliare fuori tutti gli outlier individuati precedentemente.

Researchers <5000 histogram



Il grafico ora risulta più comprensibile e (anche considerando gli outlier) si mantiene un comportamento abbastanza stabile di decrescenza. Si nota che intorno alla mediana abbiamo l'unico picco della distribuzione e la coda verso destra decresce ad una velocità pressoché costante. Questo è un comportamento che ci si poteva aspettare valutando le politiche globali sui temi della ricerca e lo sviluppo. Sarebbe interessante valutare questa variabile condizionatamente a REGION per vedere la distribuzione a livello regionale della ricerca e lo sviluppo. Ricordiamo infine nuovamente che la variabile presenta una percentuale di NA così elevata che dobbiamo "prendere con le pinze" le conclusioni che riusciamo ad ottenere.

GDP

GDP è la variabile che indica il prodotto interno lordo di un Paese, quindi a priori potrei aspettarmi una variabile che presenta dei grandi squilibri (ci sono paesi molto più ricchi e paesi molto più poveri) e sarebbe sicuramente interessante andare a valutare il GDP condizionato a REGION.

Andiamo a valutare la percentuale di NA presenti in GDP.

```
p_na = sum(is.na(data$GDP))/length(data$GDP)
p_na
```

```
## [1] 0.03783784
```

La percentuale di NA è davvero bassa, volendo verificare esattamente il numero di NA otteniamo:

```
c_na = sum(is.na(data$GDP))
```

```
c_na
```

```
## [1] 7
```

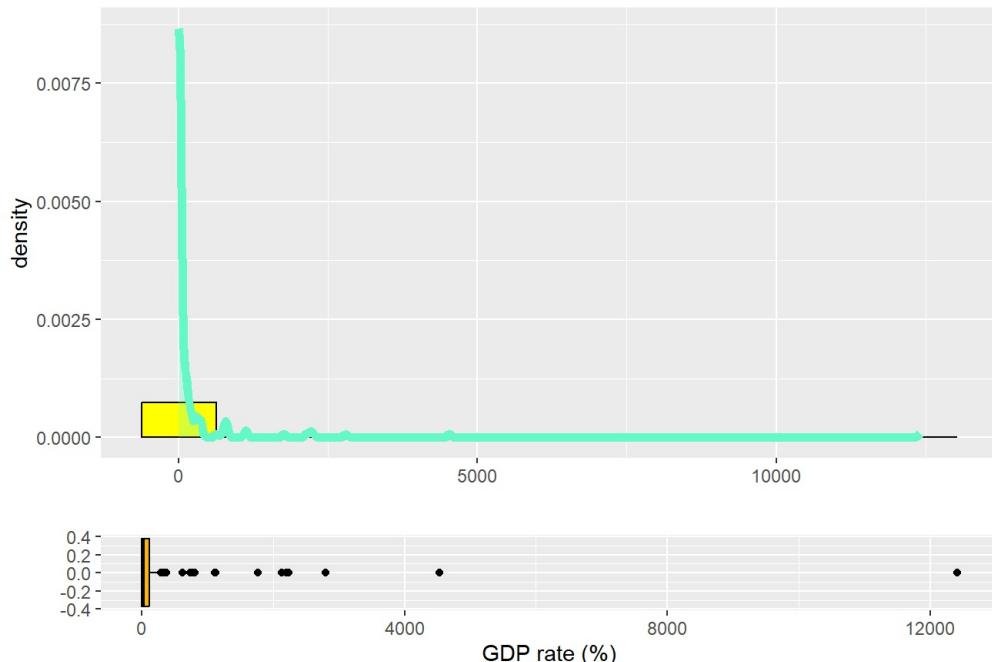
```
gdp = na.omit(data$GDP)
```

```
summary(gdp)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
## 0.10     3.55   14.20  247.55 109.28 12416.50
```

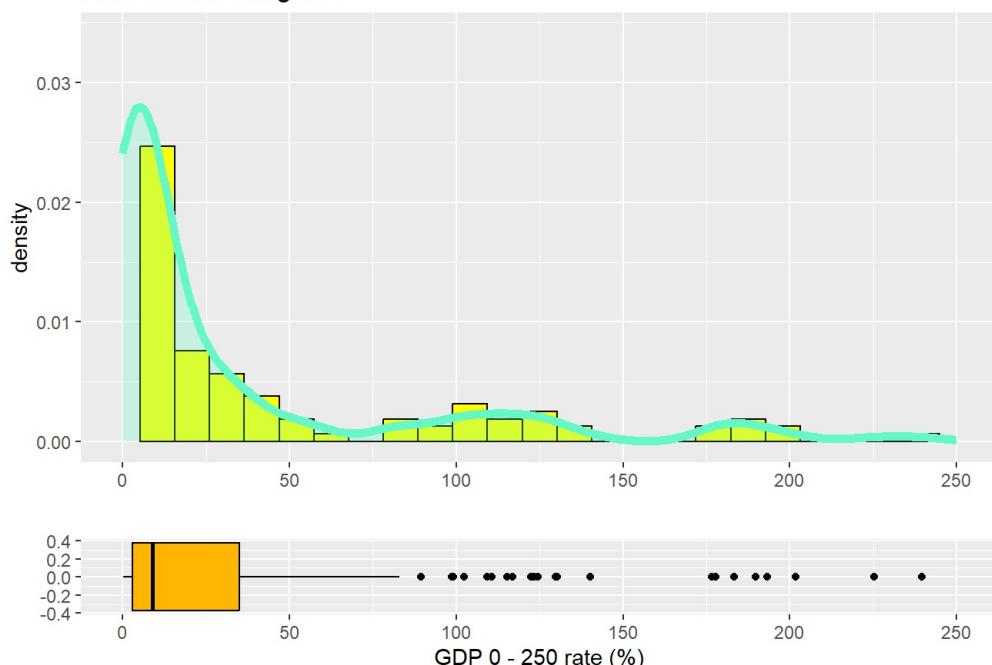
Come ci aspettavamo abbiamo un range di valori enormi, e probabilmente, già solo dal summary, possiamo aspettarci che verso il massimo siano presenti alcuni outlier che porteranno problemi nel momento in cui andremo a rappresentare i grafici. Media e Mediana sono completamente diverse e questo è quasi sicuramente dovuto alla presenza di outlier verso la coda destra della distribuzione.

GDP histogram



Come previsto la coda destra presenta degli outlier davvero significativi che rendono i nostri grafici pressoché privi di significato. Proviamo allora a rivisualizzarli questa volta considerando solo i dati nel range 0-250 (leggermente sopra la media).

GDP 0-250 histogram



Questa volta otteniamo un grafico già più comprensibile e da cui possiamo visualizzare informazioni aggiuntive. Comprendiamo infatti come gli outlier presentino dei cluster, probabilmente ad indicare che in realtà lo studio della singola variabile GDP non è molto utile e potrebbe essere di rilievo maggiore lo studio congiunto ad altri fattori, quali potrebbero essere REGION o PRIVATEHEALTH e simili (che sono valutati proprio in termini di %GDP).

FEMALEBOSS

FEMALEBOSS è la variabile che indica la percentuale di donne che ricoprono posizioni manageriali. A priori ci possiamo aspettare valori leggermente bassi per il retaggio culturale globale sull'argomento.

Andiamo a valutare la percentuale di NA presenti in FEMALEBOSS.

<

```
p_na = sum(is.na(data$FEMALEBOSS))/length(data$FEMALEBOSS)
p_na
```

```
## [1] 0.4702703
```

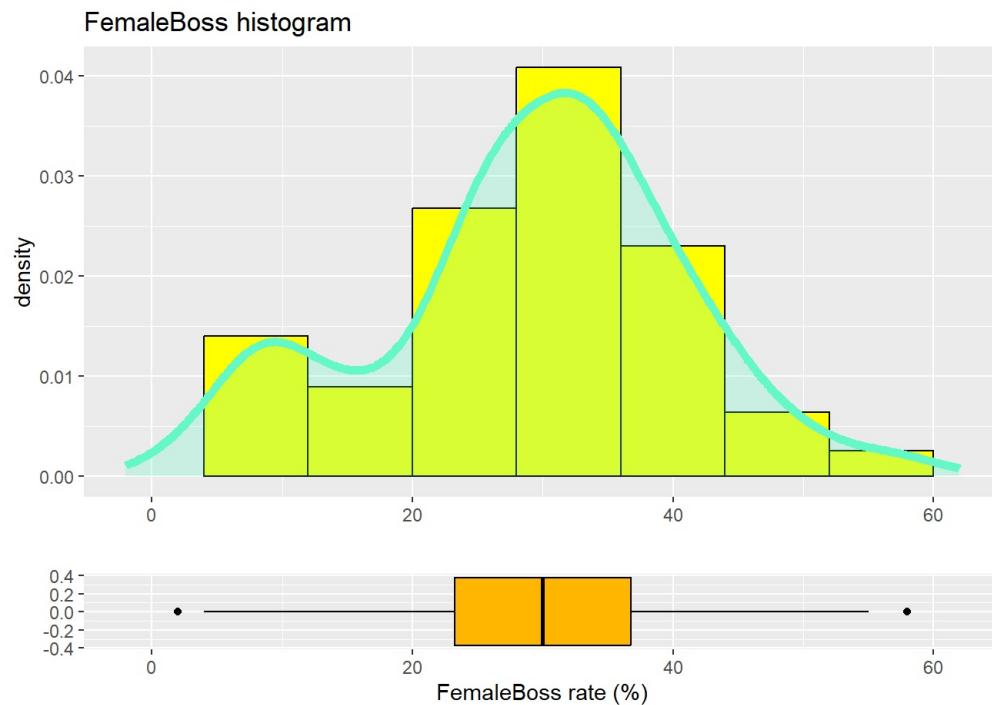
Notiamo che la percentuale è davvero elevata e questo ci indica che l'analisi che faremo sarà poco significativa, nonostante ciò procediamo con la pulizia della variabile e la successiva analisi vera e propria.

```
fb = na.omit(data$FEMALEBOSS)
summary(fb)
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##      2.00   23.25  30.00   29.07  36.75   58.00
```

I dati mostrano dei valori medi inferiori alla soglia del 50%, come ci aspettavamo, tuttavia vediamo massimi e terzo quartile sufficientemente spostati verso la soglia del 50%, segnale positivo. Notiamo che media e mediana sono molto vicine, e che sono anche molto vicine alla media aritmetica tra massimo e minimo, quindi possiamo aspettarci un indice di simmetria prossimo a 0 rispetto alla media. Andiamo a calcolarlo.

Esso è un po' più piccolo di 0 (anche se non eccessivamente), dunque non ci resta che visualizzare affiancati istogramma e grafico di densità empirica per vedere se ci abbiamo azzeccato.



Il grafico rivela un secondo picco oltre a quello della media verso il primo quartile che ci spiega l'indice di simmetria leggermente più piccolo di 0. Tuttavia dobbiamo ricordarci nuovamente come, per l'eccessiva presenza di NA, non ci è dato sapere se questo è il reale comportamento della variabile. Interessante sarebbe capire quali sono i paesi che presentano valore NA per cercare di comprendere la natura dell'assenza del dato.

Analisi Bivariata

analisi bivariata con target LIFEEXP

In questo paragrafo ci addentriamo all'interno dello studio bivariato con target LIFEEXP. Grazie all'analisi univariata precedentemente affrontata abbiamo a disposizione informazioni che andremo a sfruttare.

SELEZIONE DELLE VARIABILI

Come già anticipato all'inizio del documento per l'analisi bivariata non tutte le variabili verranno considerate. In particolare decidiamo di escludere dalla nostra analisi tutte le variabili con un'eccessiva percentuale di NA e la variabile COUNTRY, che risulta essere un factor con 185 livelli di tipo stringa diversi e dunque senza un'utilità per la previsione di LIFEEXP. Le variabili con eccessivi NA sono RESEARCHERS, SMOKING e

FEMALEBOSS. L'assenza di SMOKING è secondo noi una perdita abbastanza importante poiché il fumo è una delle principali cause di morte nel mondo, tuttavia davanti ad una percentuale di NA > 45% e alla presenza di dati che considerano solo gli uomini non possiamo che scartare la variabile.

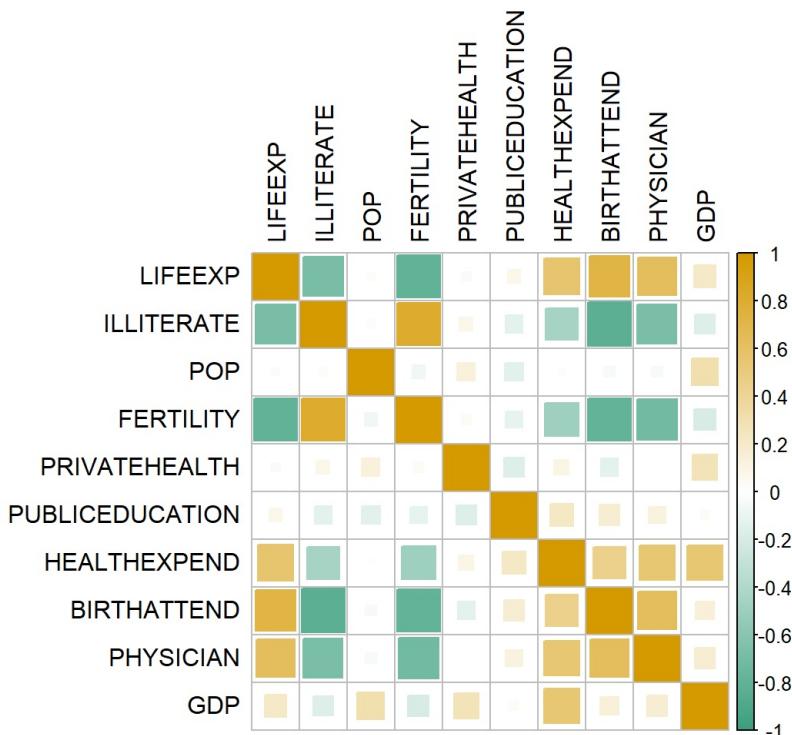
```
data = data[, !(names(data) %in% c("COUNTRY", "SMOKING", "RESEARCHERS", "FEMALEBOSS"))]
```

CALCOLO DELLE CORRELAZIONI

Procediamo ora a calcolare le correlazioni tra la variabile target ed il resto delle variabili. Dobbiamo ricordarci tuttavia che REGION è un fattore e dunque dovremo escluderla in questo processo. Dopo aver calcolato le correlazioni tra le variabili numeriche procediamo a visualizzarle in un grafico per individuare i valori più interessanti

```
numeric_vars = data[, sapply(data, is.numeric)]
cm = cor(numeric_vars, method = "pearson", use = "pairwise.complete.obs")

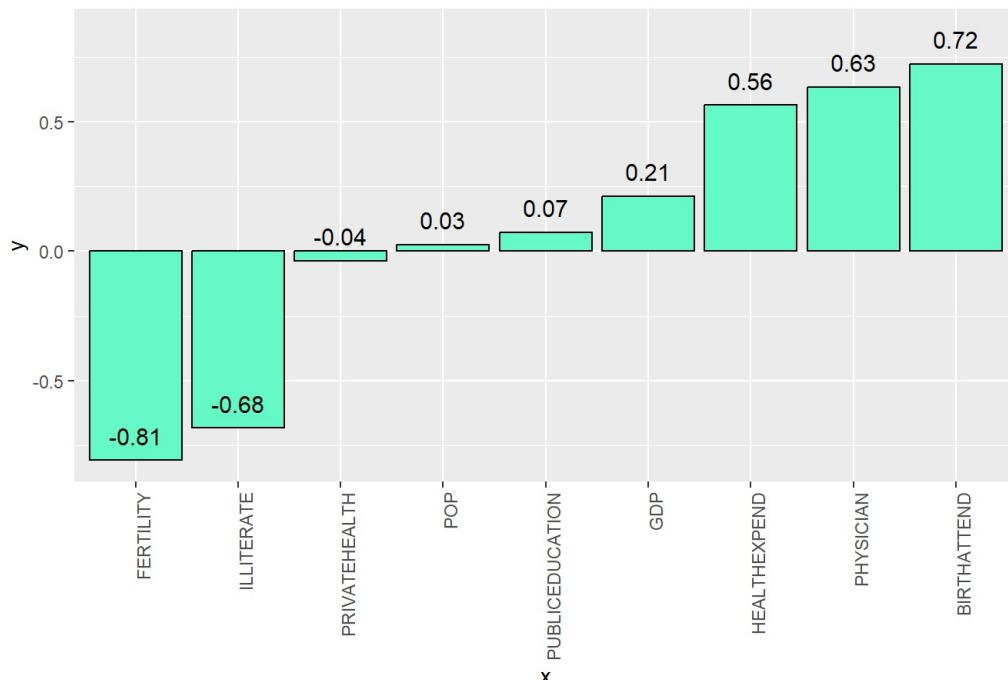
corrplot(cm, tl.col = 'black', method = 'square', col = colorRampPalette(c(verde_acqua_scurito, "white", arancione_scurito))(200))
```



```
correlations = sapply(numeric_vars, function(x) cor(numeric_vars$LIFEEXP, x, use = "pairwise.complete.obs", method = "pearson"))
```

Avendo scelto LIFEEXP come variabile target, per il momento ci interesseremo devi valori di correlazione con essa. Per visualizzarli meglio si decise di fare un grafico a barre in cui mettiamo in ordine per valore di correlazione le variabili relazionate alla target.

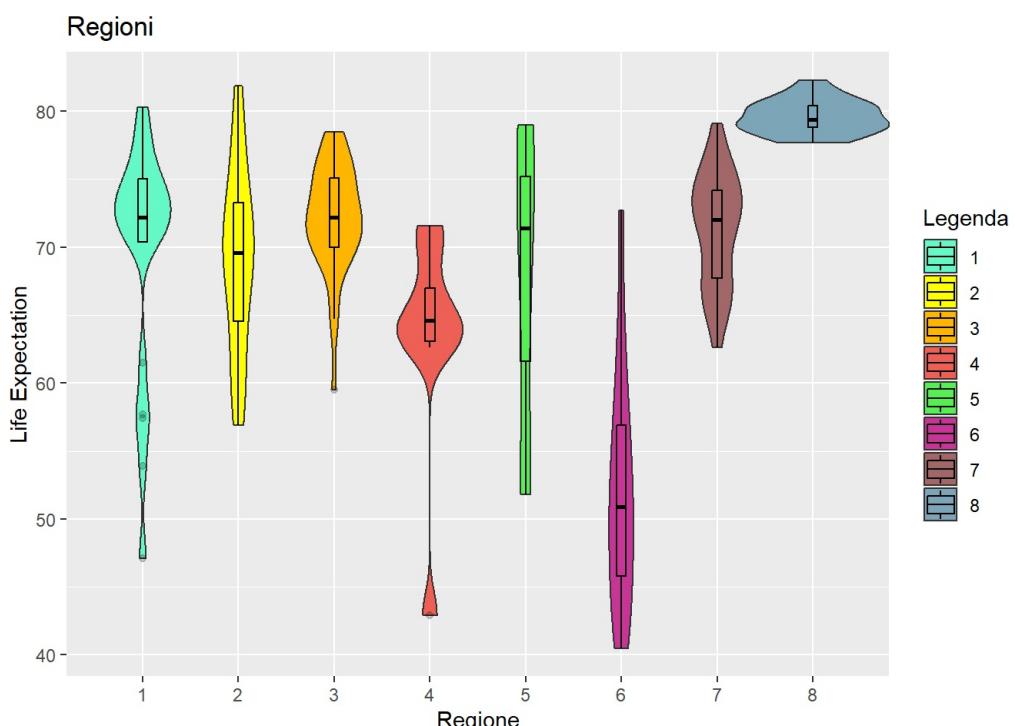
Correlation with life expectancy



I risultati sono soddisfacenti, poiché abbiamo abbastanza indici di Pearson superiori a 0.5 in modulo. Per quanto riguarda PRIVATEHEALTH, POP e PUBLICEDUCATION, che hanno indici prossimi a 0, decidiamo di non eseguire alcun tipo di analisi bivariata. Possiamo dunque cominciare a tutti gli effetti le nostre analisi bivariate.

REGION

Poiché la regione è un factor è necessario sviluppare un tipo di analisi completamente diversa rispetto alle altre. Per cominciare andiamo a visualizzare i violin boxplot affiancati.



Poiché non sembra esserci indipendenza, eseguiamo un test ANOVA.

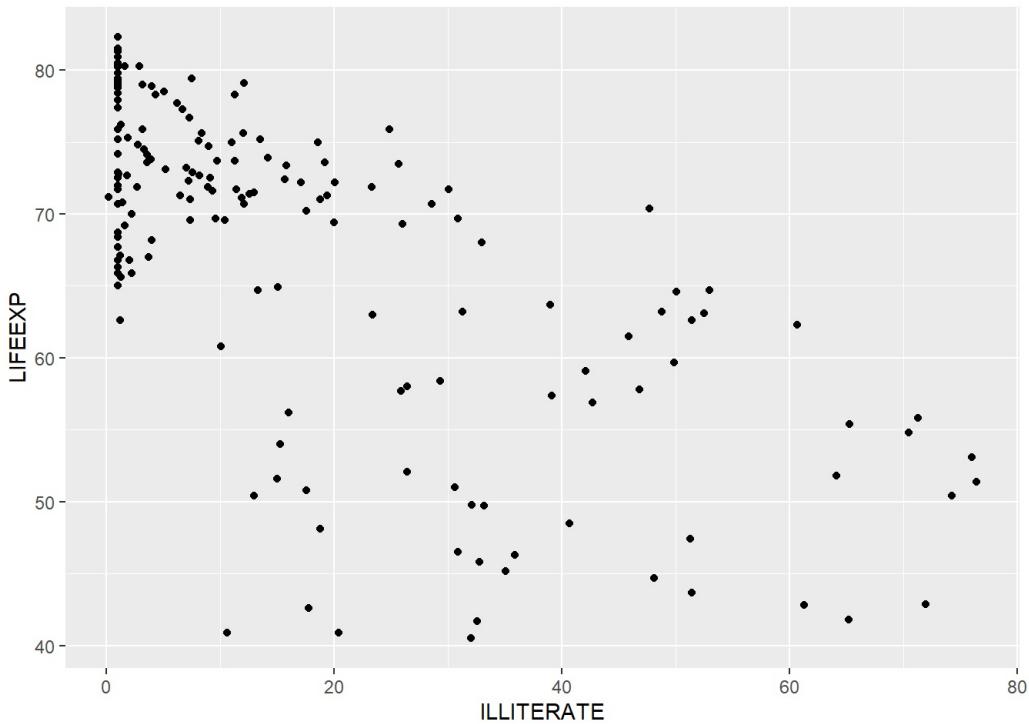
```
test = aov(data$LIFEEXP~rg)
summary(test)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)						
## rg	7	14948	2135.4	49.42	<2e-16 ***						
## Residuals	177	7649	43.2								
## ---											
## Signif. codes:	0	'****'	0.001	'**'	0.01	'*'	0.05	'. '	0.1	' '	1

Il test sull'indipendenza ha come risultato un p-value <2e-16 quindi abbastanza piccolo da non mostrare evidenze contro l'ipotesi di indipendenza. Risulta ragionevole pensare che vi è una relazione tra regione di appartenenza e aspettativa di vita.

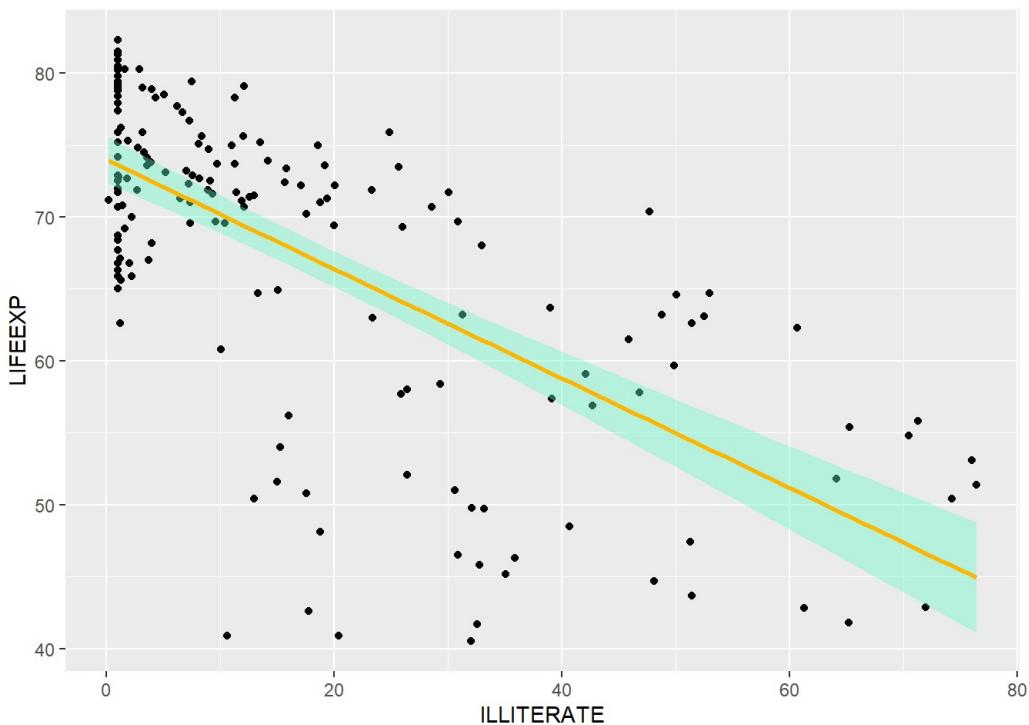
ILLITERATE

La variabile ILLITERATE a differenza di REGION è quantitativa quindi necessita un'analisi diversa rispetto a quest'ultima. Graficamente, visualizziamo la relazione tra le due variabili con uno scatter-plot. La variabile ILLITERATE presenta 14 valori mancanti



Nel grafico notiamo una prevalenza di due comportamenti delle osservazioni: una parte è concentrata su valori vicini allo 0 di ILLITERATE mentre gli altri sono distribuiti in maniera sparsa ma con una tendenza decrescente, a conferma del valore negativo dell'indice di correlazione calcolato in precedenza. Data questa tendenza ci aspettiamo una devianza dei residui abbastanza elevata. Essendo il modulo dell'indice di correlazione alto, proviamo ad analizzare la relazione tra ILLITERATE e LIFEEXP

```
## `geom_smooth()` using formula = 'y ~ x'
```



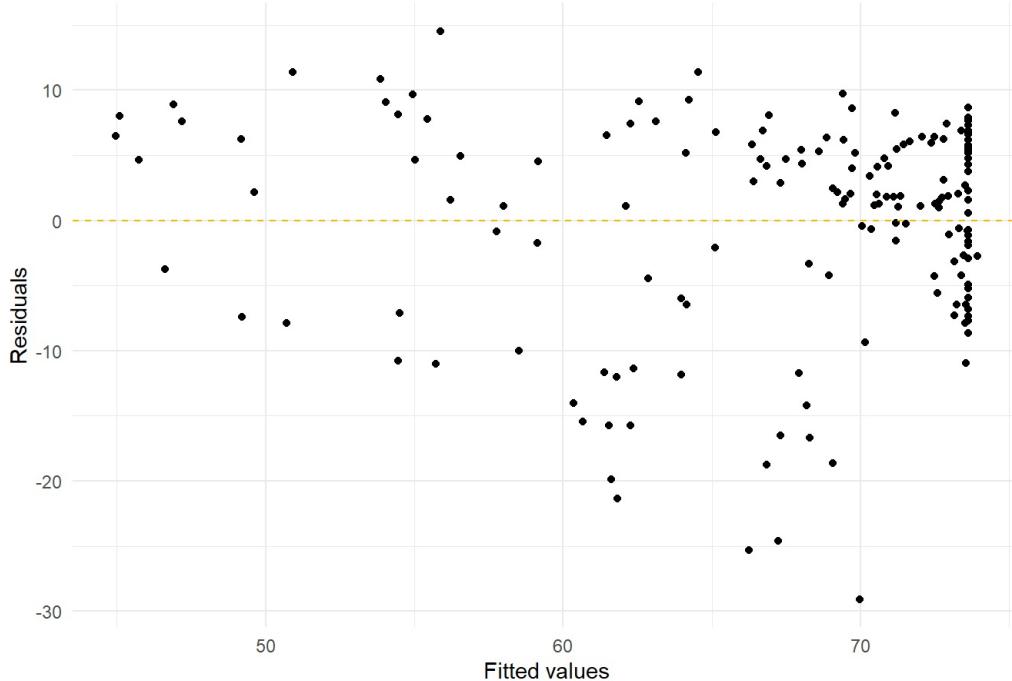
```
summary(model)
```

```

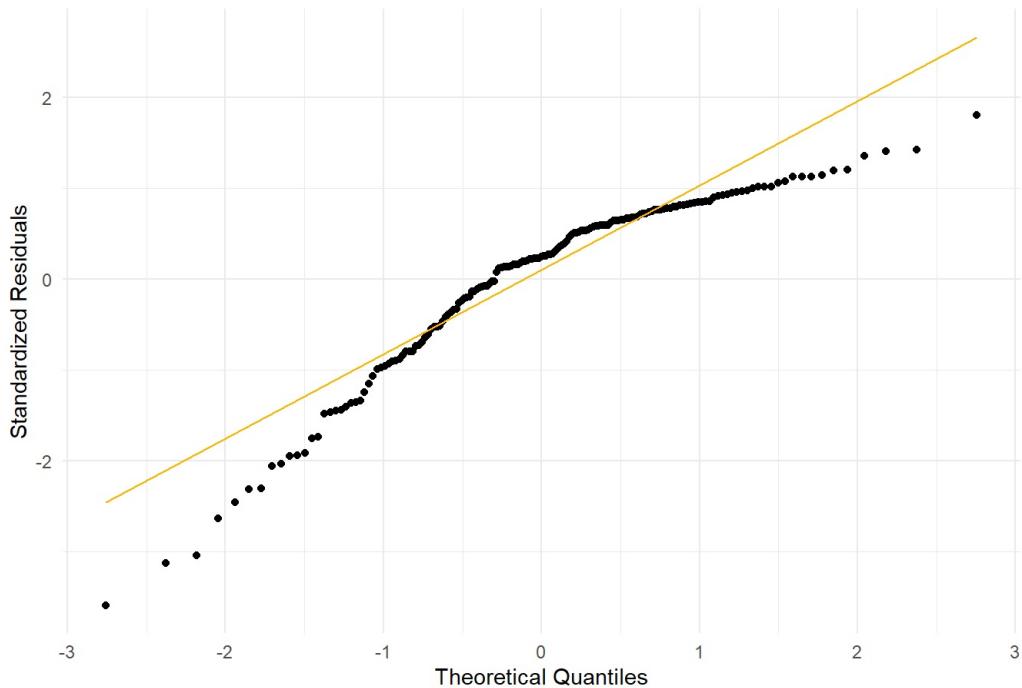
## 
## Call:
## lm(formula = LIFEEXP ~ ILLITERATE, data = data)
## 
## Residuals:
##    Min      1Q  Median      3Q     Max 
## -29.077 -4.268  1.953  5.884 14.532 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 74.00781   0.83264   88.88 <2e-16 ***
## ILLITERATE -0.38028   0.03136  -12.13 <2e-16 ***
## ---    
## Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 8.121 on 169 degrees of freedom
##   (14 osservazioni eliminate a causa di valori mancanti)
## Multiple R-squared:  0.4653, Adjusted R-squared:  0.4622 
## F-statistic: 147.1 on 1 and 169 DF,  p-value: < 2.2e-16

```

Residuals vs Fitted

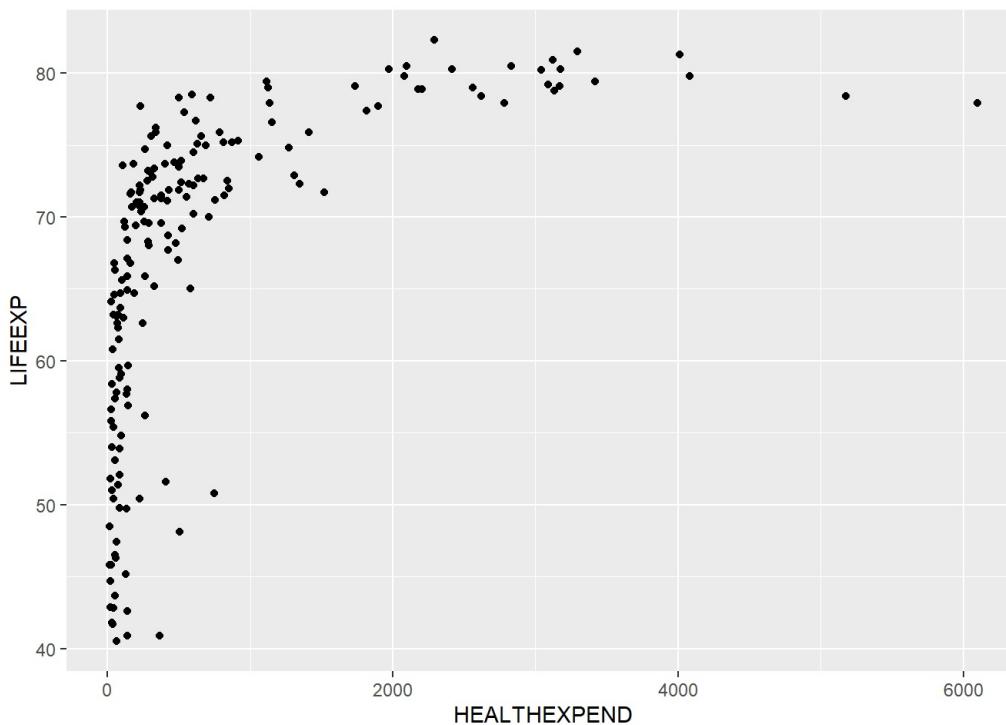


Q-Q Plot of Standardized Residuals

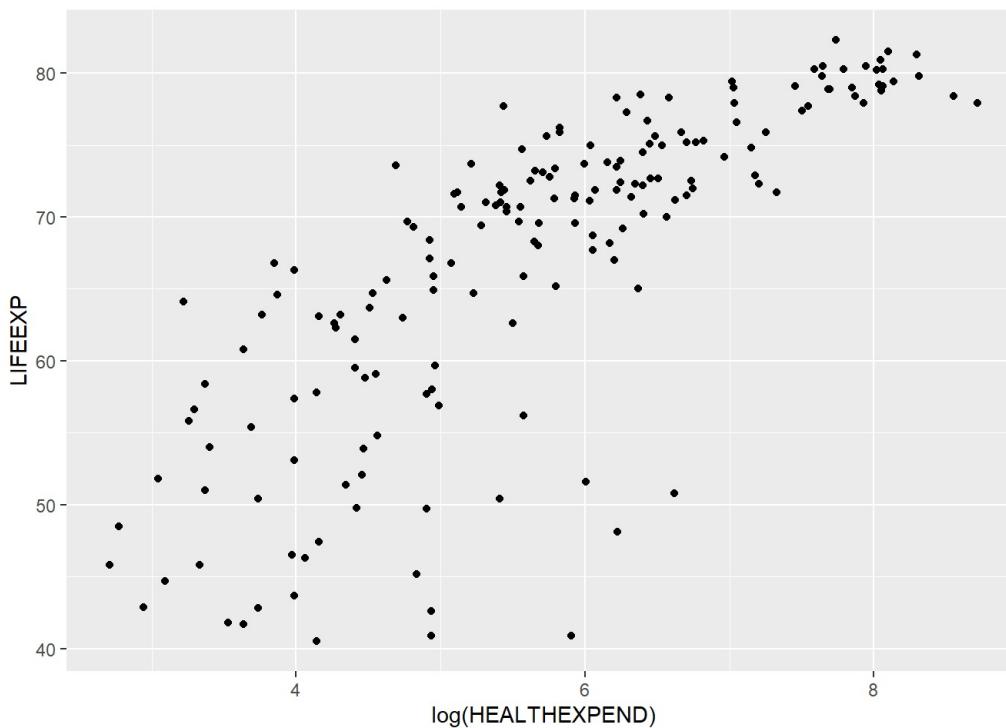


Il modello lineare mostra un indice R^2 di 0.4653, possiamo quindi dire che la variabilità dei dati viene rappresentata solo in parte dal modello lineare. Come ci aspettavamo, il grafico dei residui mostra una devianza abbastanza elevata, con molti valori che si discostano dalla retta orizzontale rossa. Il Q-Q plot mostra come i residui non oscillino in maniera casuale attorno ai quantili teorici, il che ci fa dubitare dell'efficacia di un modello lineare per predire l'aspettativa di vita attraverso l'analfabetismo.

HEALTHEXPEND



La distribuzione di questi dati è evidentemente non lineare, si può provare ad applicare la funzione logaritmo alla variabile sull'asse x per vedere se la correlazione lineare sale e se è possibile ricavare altre informazioni.

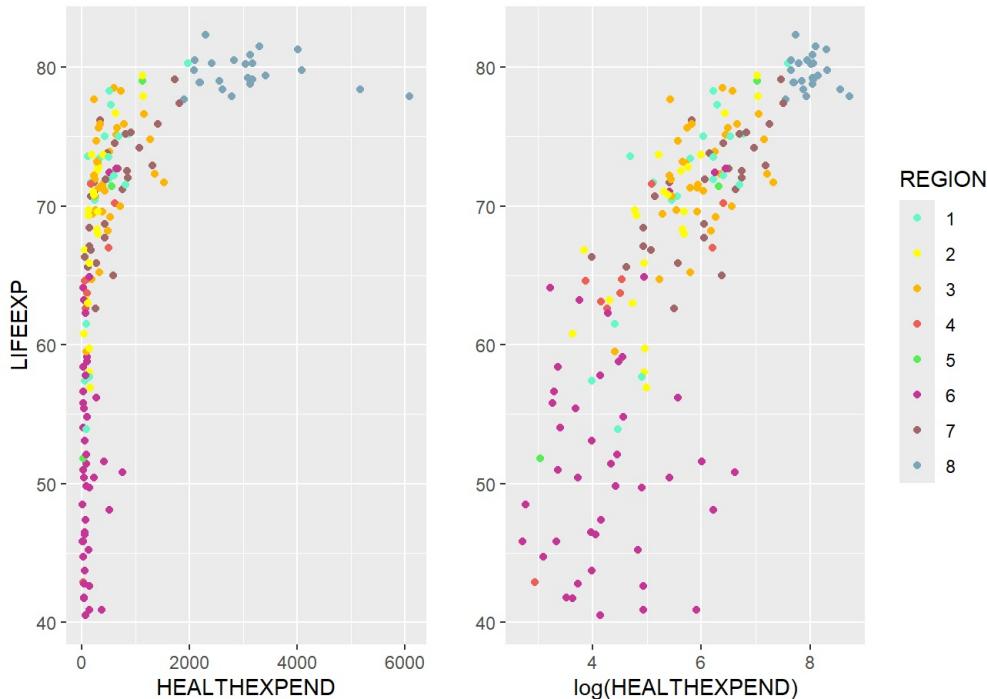


Dopo aver applicato il logaritmo alla variabile x si può notare come la distribuzione dei dati sia più simile ad una retta che taglia il primo quadrante. Si mettono a confronto i due indici di correlazione con la variabile base e quella a cui è stata applicata una trasformazione.

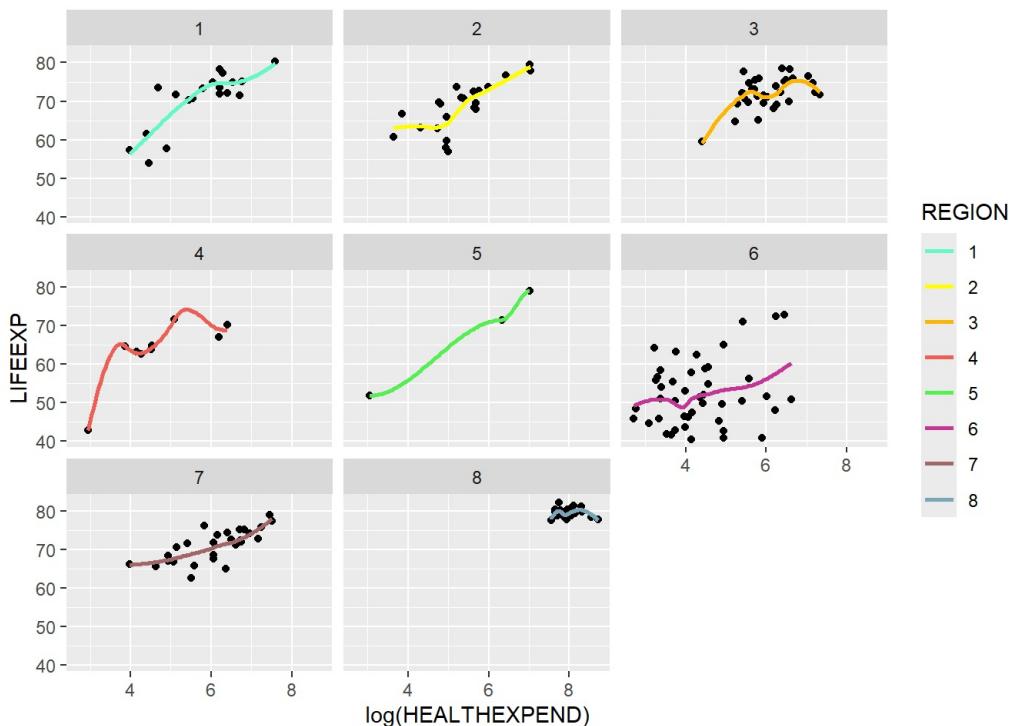
```
## [1] 0.5637716
```

```
## [1] 0.7801348
```

Applicando il logaritmo alla variabile HEALTHEXPEND il valore dell'indice di correlazione sale da 0,5638 a 0,7801. Un'altra cosa che possiamo facilmente visualizzare per aggiungere informazioni al grafico è la regione di appartenenza dei diversi punti.

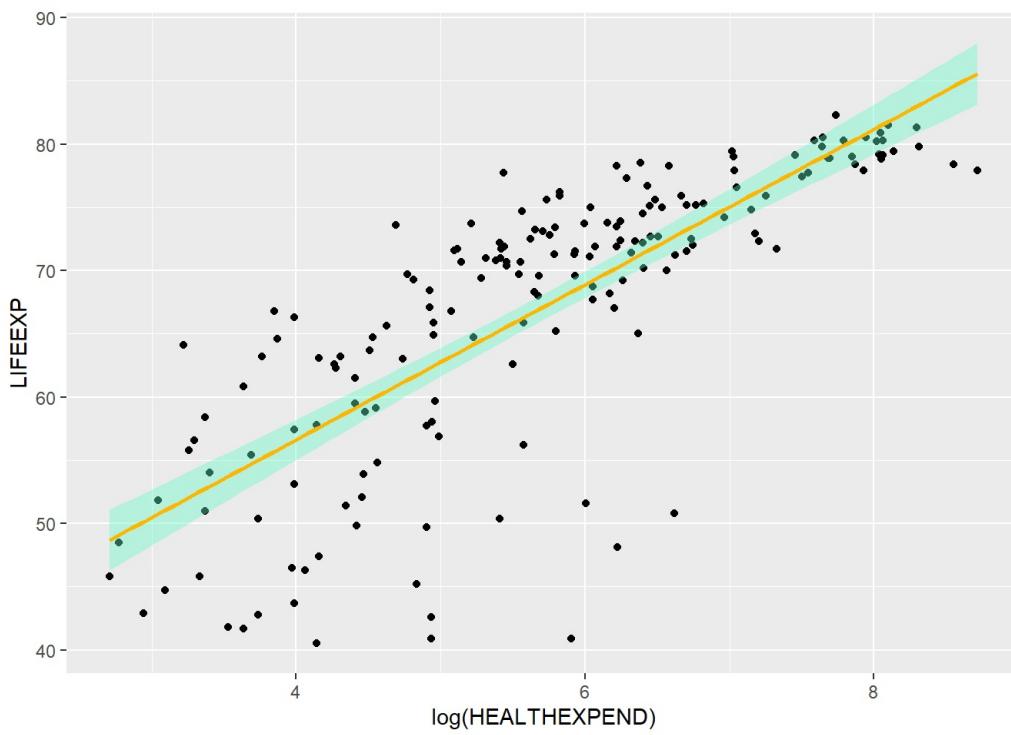


```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Si può facilmente notare la presenza di gruppi distinti sia prima che dopo aver applicato la trasformazione logaritmica. Essendo il coefficiente di correlazione lineare aumentato significativamente si ritiene opportuno costruire un modello lineare tra la variabile target e il logaritmo di HEALTHEXPEND.

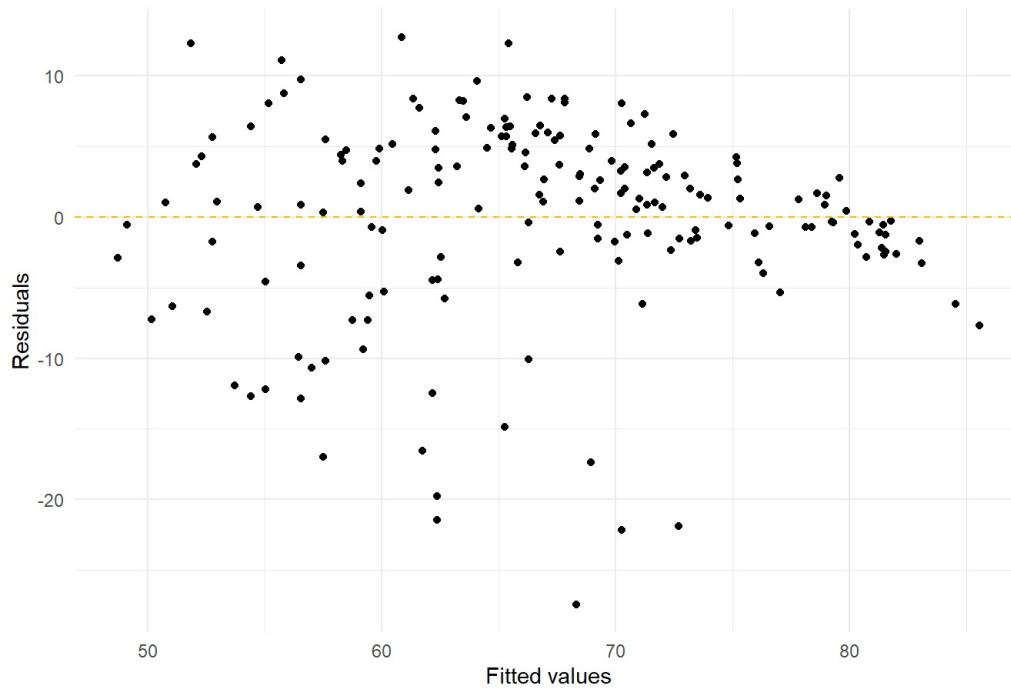
```
## `geom_smooth()` using formula = 'y ~ x'
```



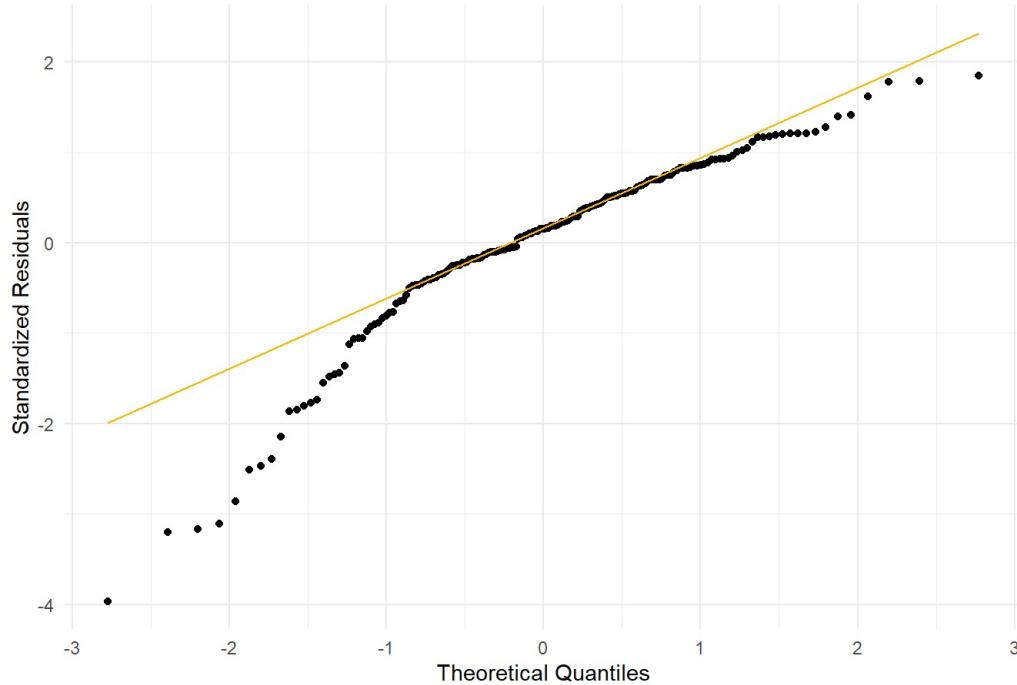
```
summary(model)
```

```
## 
## Call:
## lm(formula = LIFEEXP ~ log(HEALTHEXPEND), data = data)
## 
## Residuals:
##    Min     1Q   Median     3Q    Max 
## -27.412 -2.487  1.053  4.736 12.740 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 32.0659   2.1607  14.84 <2e-16 ***
## log(HEALTHEXPEND) 6.1377   0.3689  16.64 <2e-16 ***
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6.931 on 178 degrees of freedom
##   (5 osservazioni eliminate a causa di valori mancanti)
## Multiple R-squared:  0.6086, Adjusted R-squared:  0.6064 
## F-statistic: 276.8 on 1 and 178 DF,  p-value: < 2.2e-16
```

Residuals vs Fitted

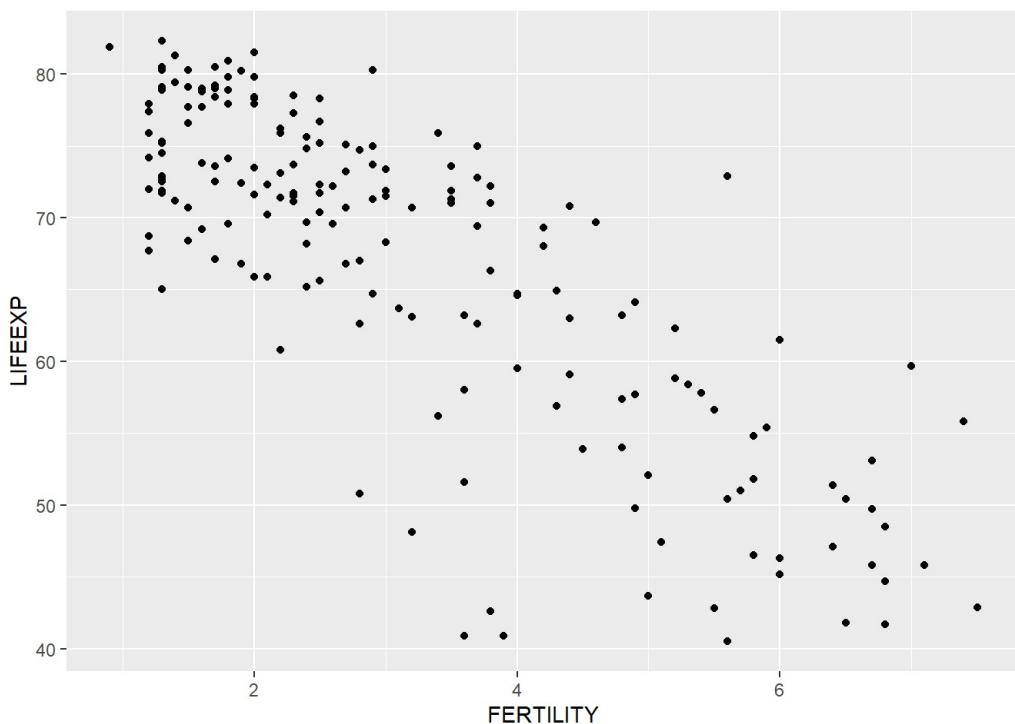


Q-Q Plot of Standardized Residuals

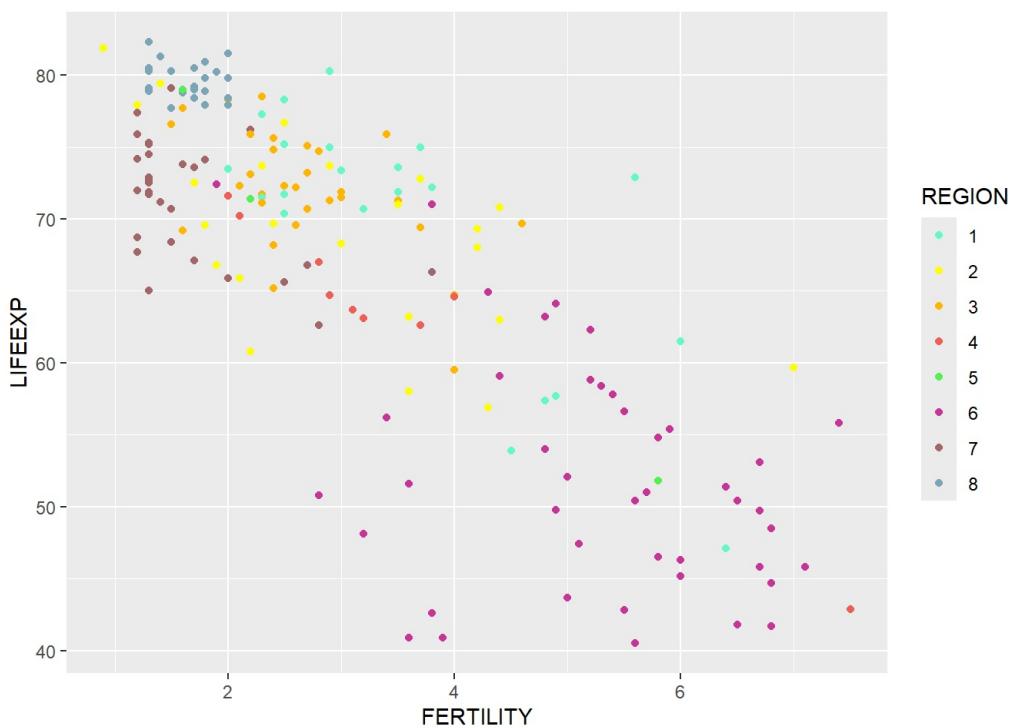


L'indice R^2 che indica la bontà del nostro modello lineare assume un valore di 0,6086, esprimendo come circa il 60% della variabilità dei dati può essere spiegata dal modello lineare, valore sufficientemente alto da dimostrare la bontà del modello. Anche in questo caso però i grafici dei residui non mostrano un comportamento desiderabile, conseguenza della distribuzione sparsa e asimmetrica delle osservazioni nello scatterplot iniziale.

FERTILITY

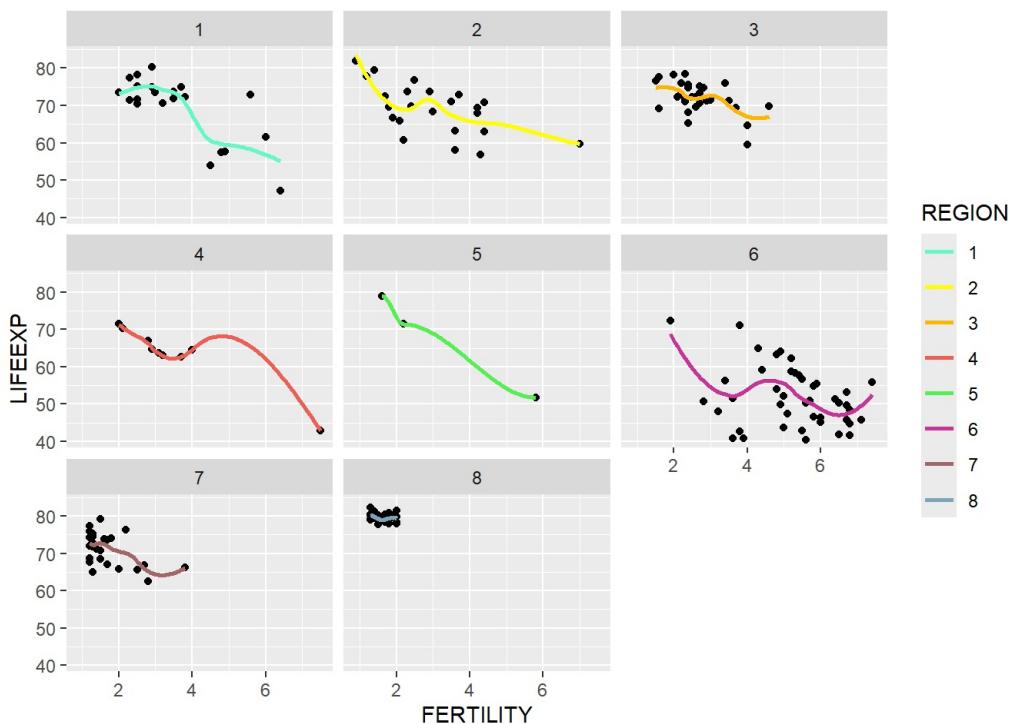


Un'altra cosa che possiamo facilmente visualizzare per aggiungere informazioni al grafico è la regione di appartenenza dei diversi punti.



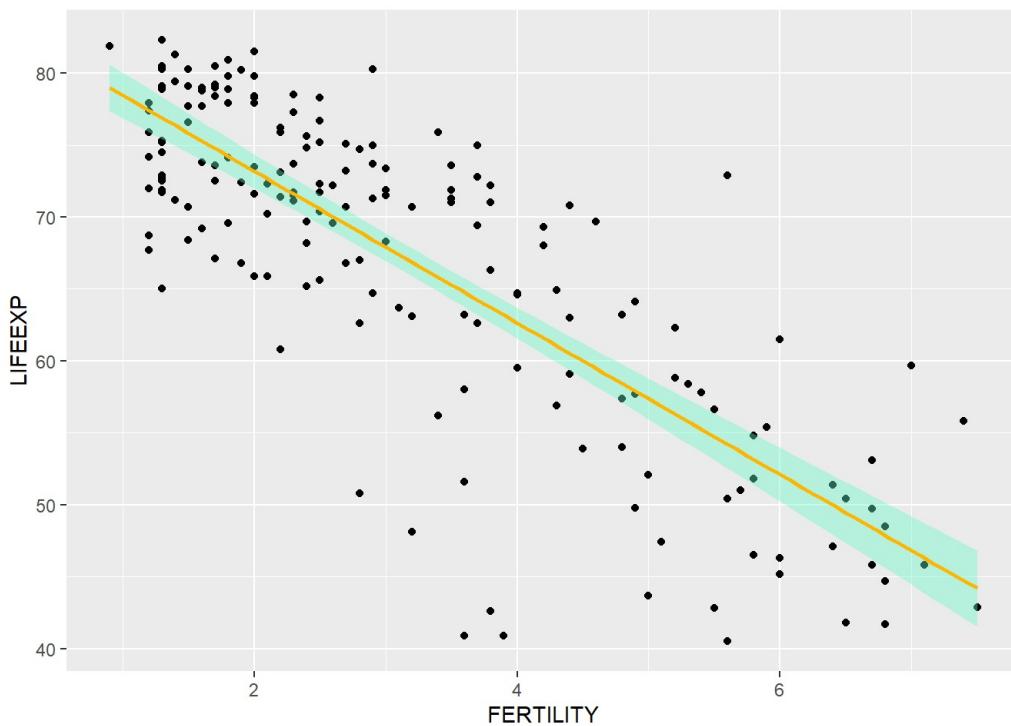
Nonostante anche in questo caso i punti siano piuttosto sparsi nello spazio, si riconosce, come per la variabile precedente, la vicinanza tra punti appartenenti alla stessa regione

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Notiamo come sia considerando separatamente i gruppi formati dalle regioni sia osservando il grafico generale, sembra essere presente una relazione lineare negativa tra le due variabili, come evidenziato dalla correlazione pari a -0,81. Ci aspettiamo quindi che una retta di regressione possa rappresentare bene il rapporto tra FERTILITY e LIFEEXP.

```
## `geom_smooth()` using formula = 'y ~ x'
```

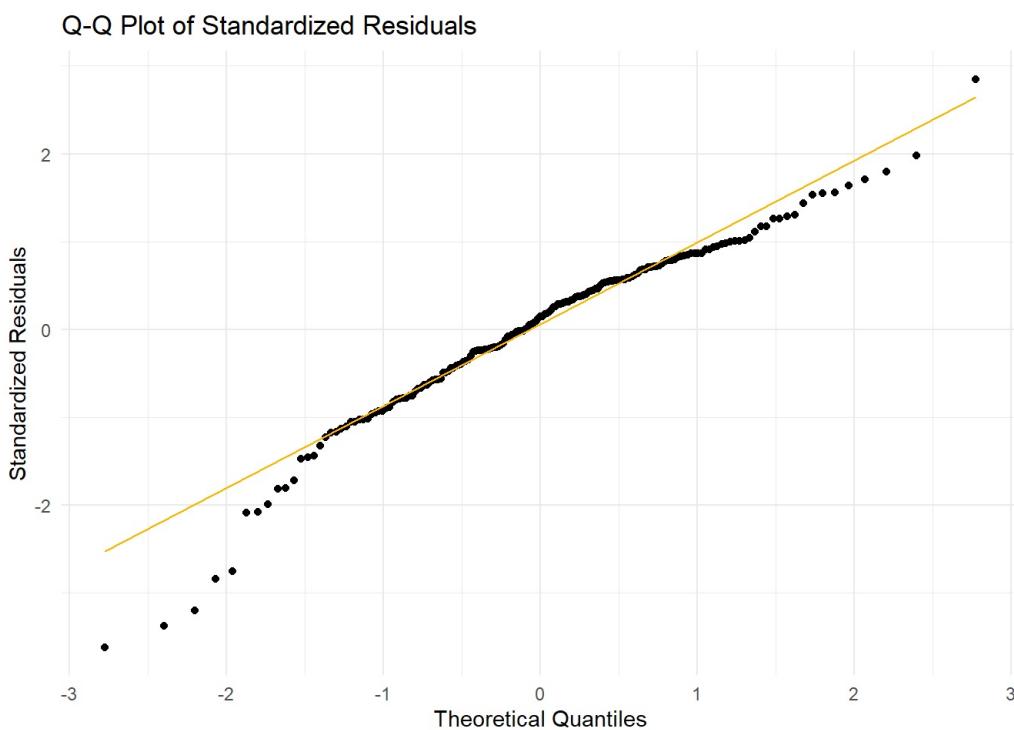
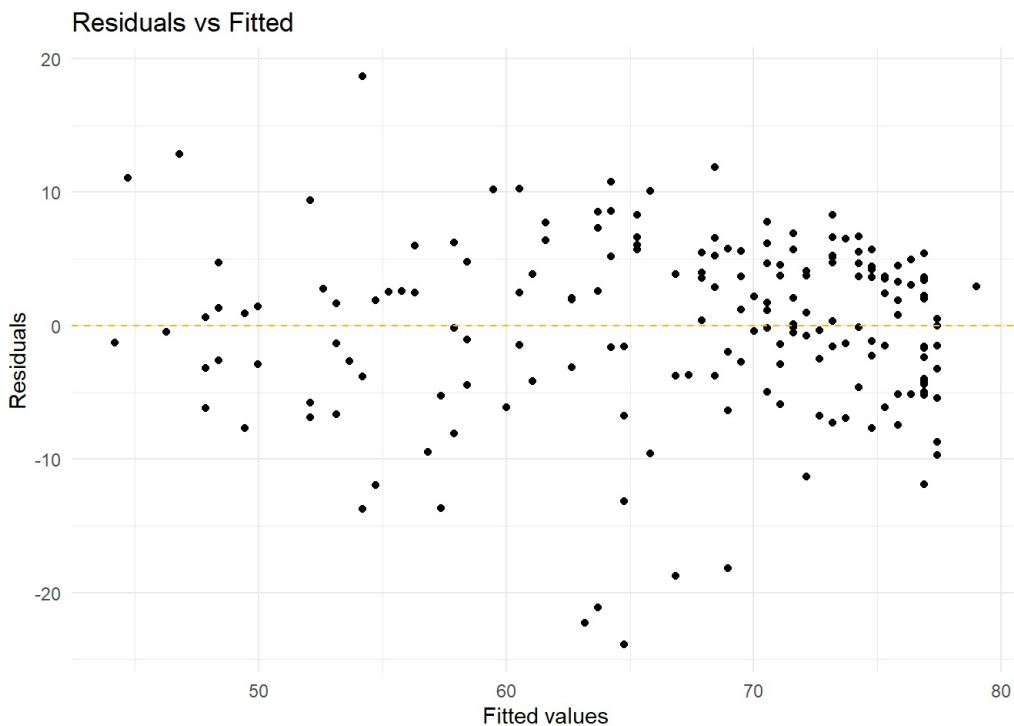


```
summary(model)
```

```

## 
## Call:
## lm(formula = LIFEEXP ~ FERTILITY, data = data)
## 
## Residuals:
##    Min      1Q  Median      3Q     Max 
## -23.8535 -3.7629  0.9397  4.5183 18.6935 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 83.7381    1.0439   80.22 <2e-16 ***
## FERTILITY   -5.2735    0.2887  -18.27 <2e-16 ***
## ---    
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6.615 on 179 degrees of freedom
## (4 osservazioni eliminate a causa di valori mancanti)
## Multiple R-squared:  0.6508, Adjusted R-squared:  0.6489 
## F-statistic: 333.7 on 1 and 179 DF,  p-value: < 2.2e-16

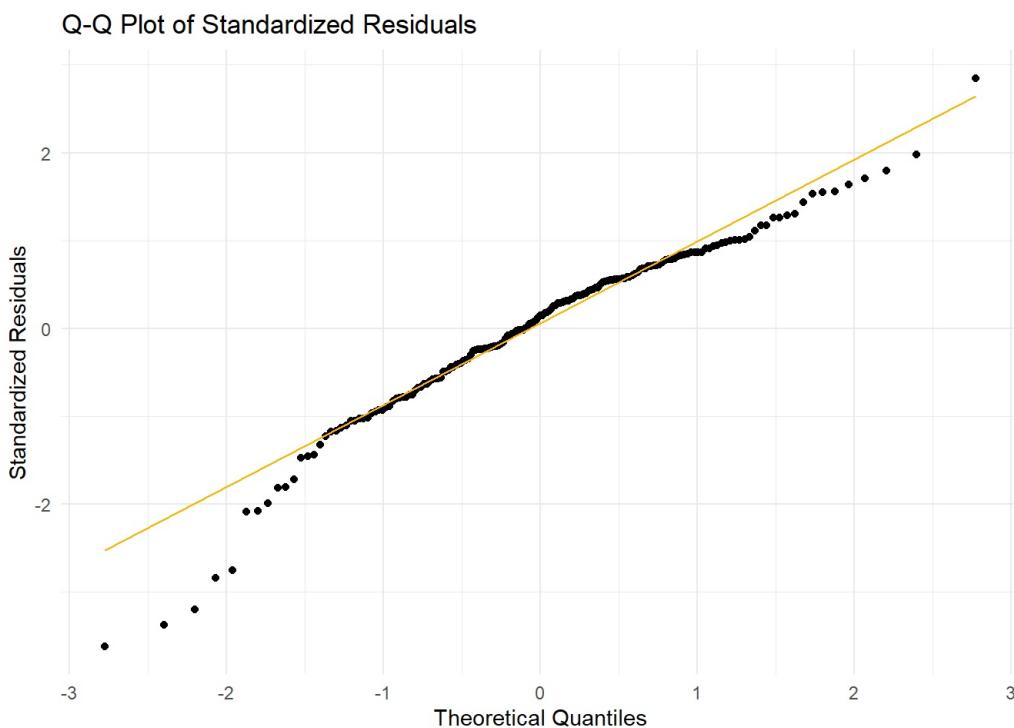
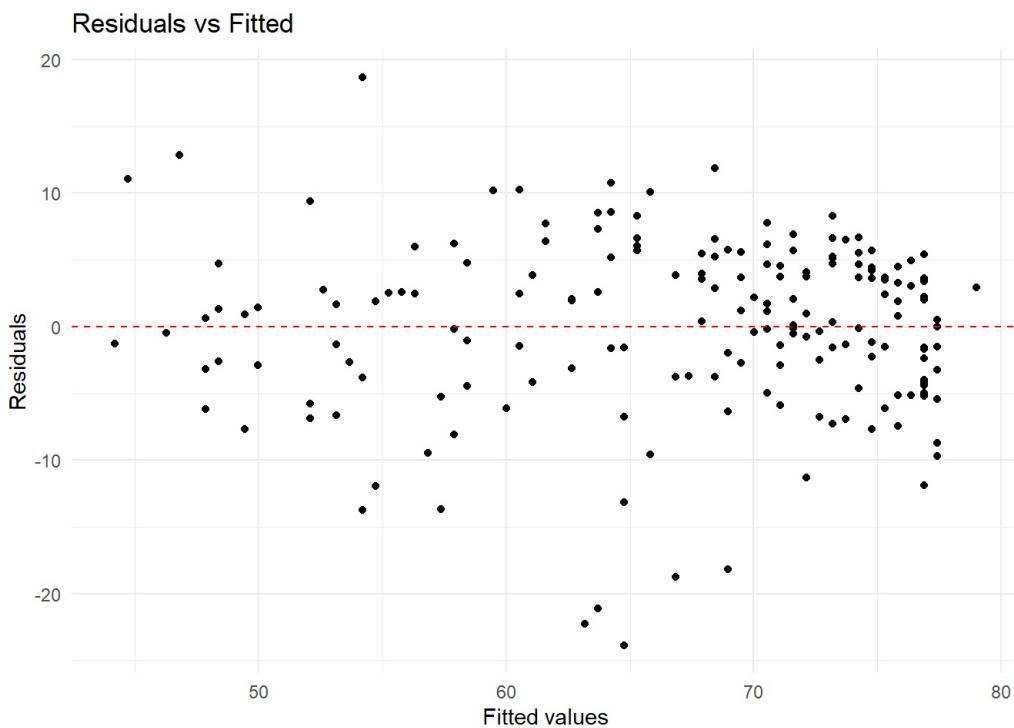
```



Analizzando questo modello di regressione lineare, troviamo un indice R^2 pari a 0.6508, valore che ci esprime come la retta svolga una buona approssimazione dei dati. Il comportamento dei residui è più casuale rispetto alle analisi precedenti, mentre rimane simile ai casi precedenti la varianza degli scarti rispetto al valore medio.

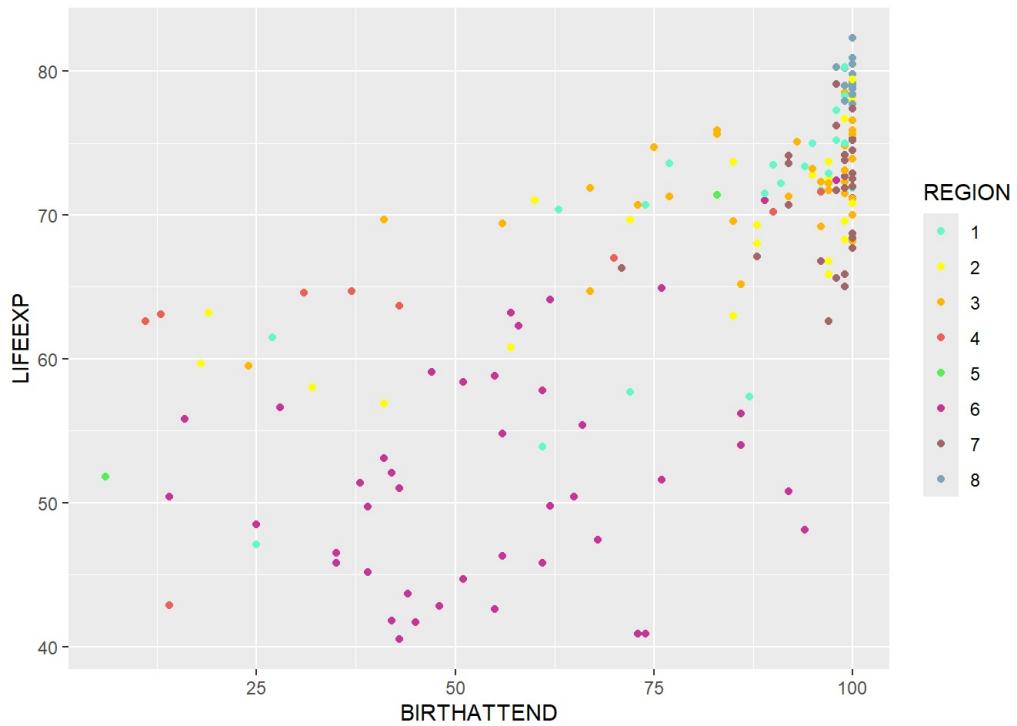
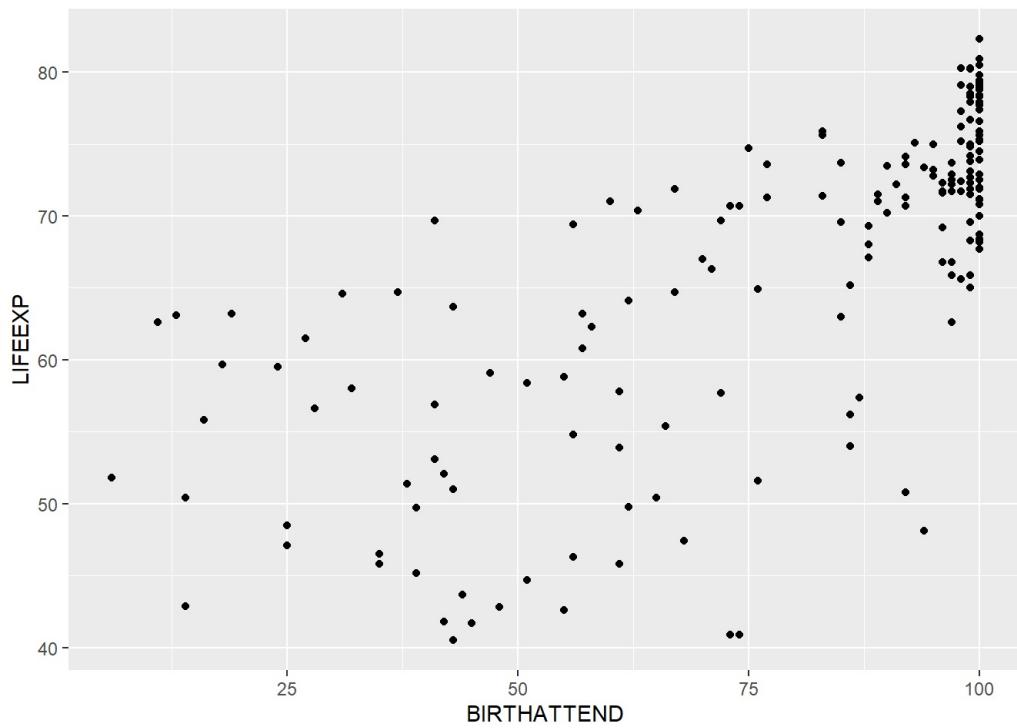
```
summary(model)
```

```
##  
## Call:  
## lm(formula = LIFEEXP ~ FERTILITY, data = data)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -23.8535 -3.7629  0.9397  4.5183 18.6935  
##  
## Coefficients:  
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 83.7381    1.0439   80.22 <2e-16 ***  
## FERTILITY   -5.2735    0.2887  -18.27 <2e-16 ***  
## ---  
## Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 6.615 on 179 degrees of freedom  
## (4 osservazioni eliminate a causa di valori mancanti)  
## Multiple R-squared:  0.6508, Adjusted R-squared:  0.6489  
## F-statistic: 333.7 on 1 and 179 DF, p-value: < 2.2e-16
```



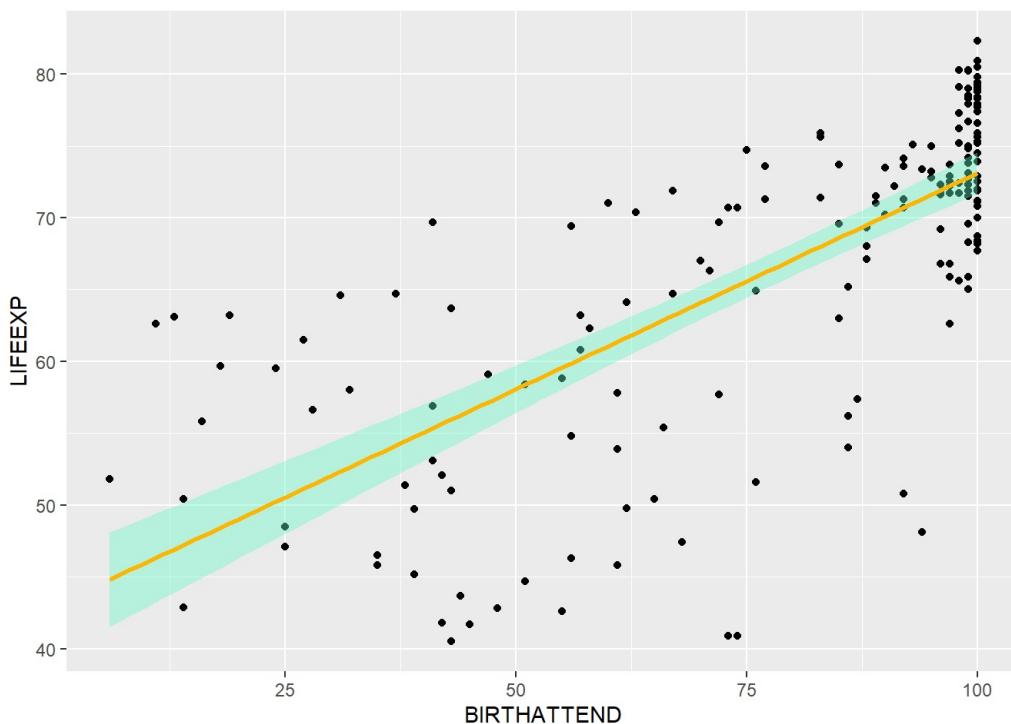
Analizzando questo modello di regressione lineare, troviamo un indice R² pari a 0.6508, valore che rappresenta una buona approssimazione dei dati da parte della retta. Il comportamento dei residui è più casuale rispetto alle analisi precedenti, ma rimane simile ai casi precedenti la varianza rispetto al valore medio.

BIRTHATTEND



Notiamo nuovamente due comportamenti distinti dei punti osservati, i quali si disperdonano nello spazio con una tendenza a crescere per poi concentrarsi in alto a destra nel grafico. Anche evidenziando le regioni, non sembra che i gruppi siano tanto distinguibili quanto nei casi precedenti. Ricordando come il coefficiente di correlazione lineare delle due variabili sia pari a 0,72 il modello di regressione lineare dovrebbe poter rappresentare bene il rapporto sotteso ai dati. Non dovrebbe stupire infatti che nei paesi in cui la presenza di personale specializzato ad assistere le nascite è capillare, la qualità della vita e di conseguenza la sua durata media sarà maggiore.

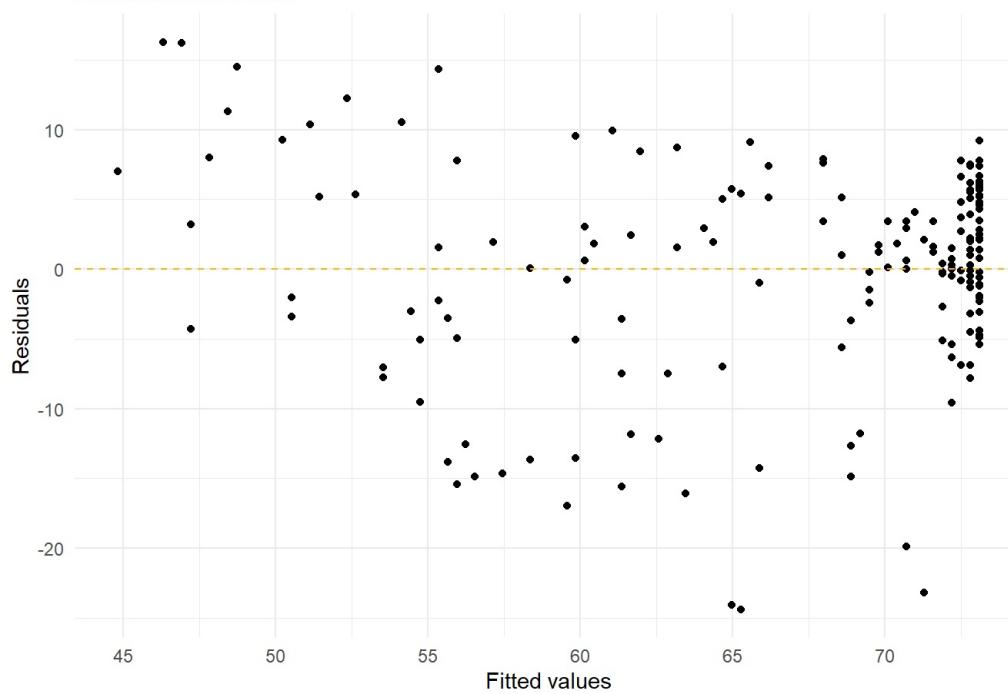
```
## `geom_smooth()` using formula = 'y ~ x'
```



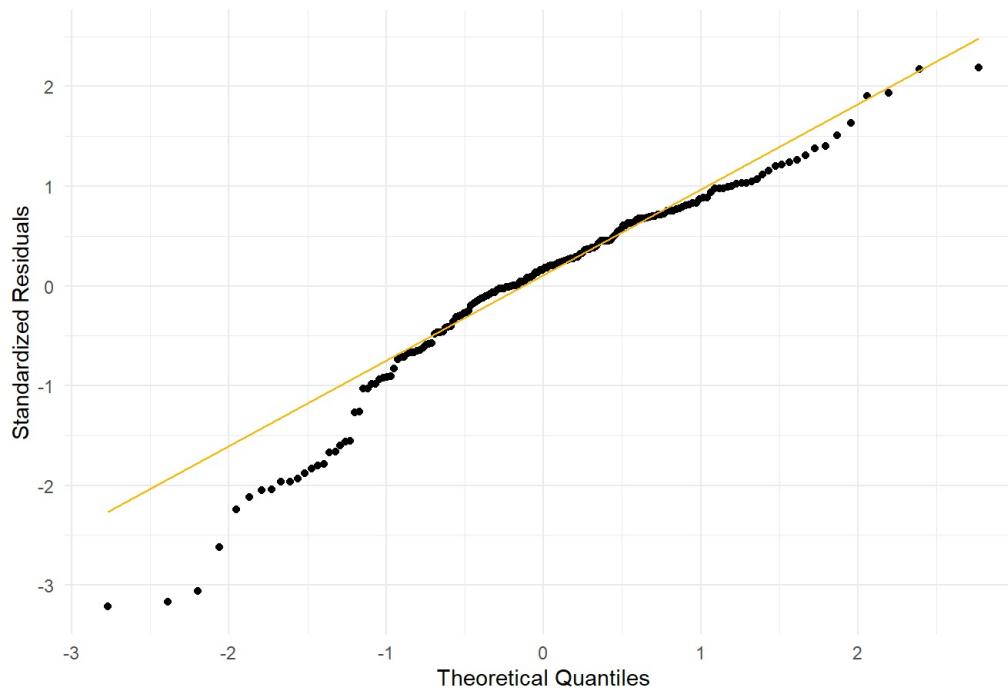
```
summary(model)
```

```
## 
## Call:
## lm(formula = LIFEEXP ~ BIRTHATTEND, data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -24.378 -3.560   1.302   5.189  16.286 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 43.00259   1.78803  24.05 <2e-16 ***
## BIRTHATTEND  0.30102   0.02166 13.90 <2e-16 ***
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 7.612 on 176 degrees of freedom
##   (7 osservazioni eliminate a causa di valori mancanti)
## Multiple R-squared:  0.5233, Adjusted R-squared:  0.5206 
## F-statistic: 193.2 on 1 and 176 DF,  p-value: < 2.2e-16
```

Residuals vs Fitted



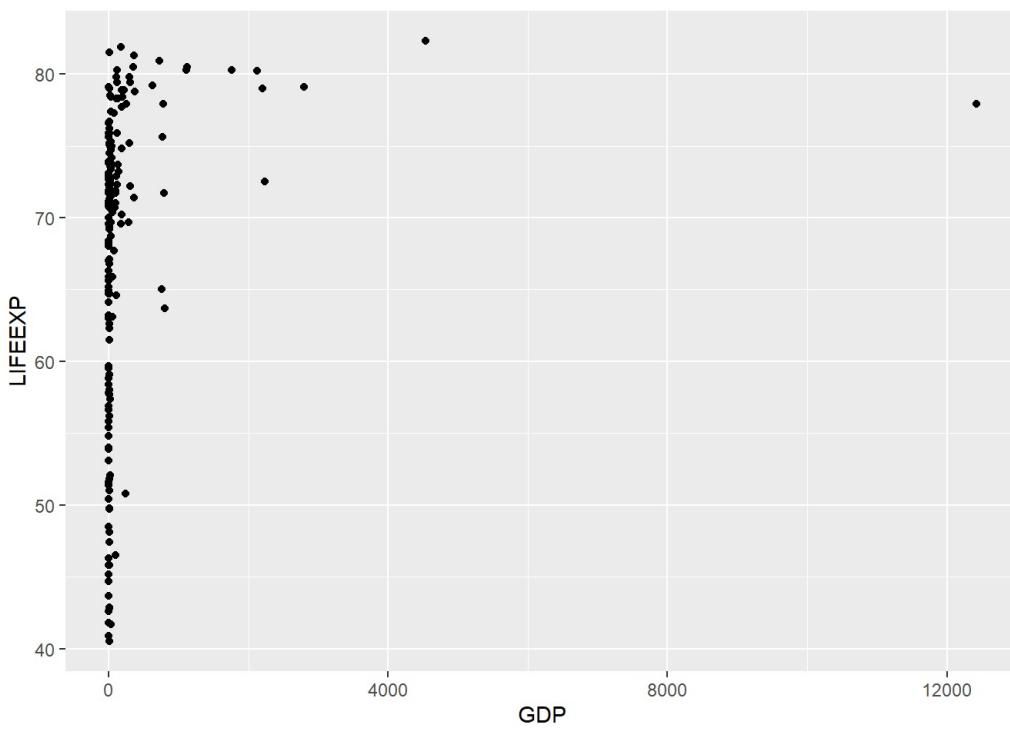
Q-Q Plot of Standardized Residuals



Il modello di regressione ha un R^2 pari a 0,5233, valore inferiore alle aspettative probabilmente a causa della solita varianza residua elevata dovuta alla dispersione dei punti nello spazio. Il grafico dei residui sembra mostrare una tendenza negativa degli scarti dalla retta, comportamento che ci porta a dubitare ulteriormente della bontà del modello realizzato.

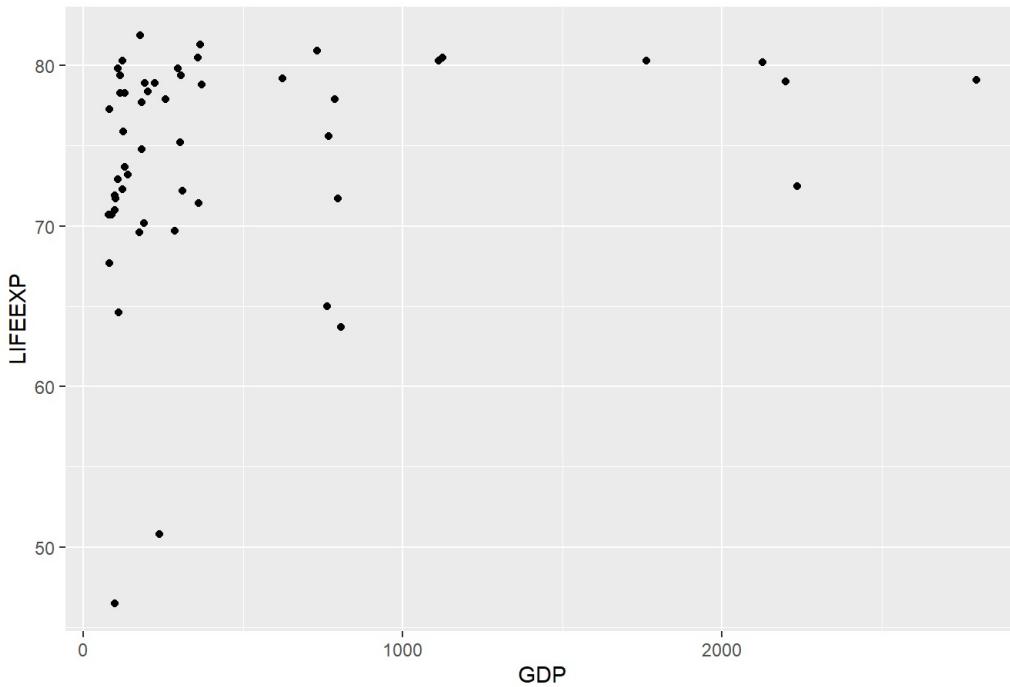
GDP

GDP non è una variabile con un indice di correlazione molto alto, tuttavia il GDP secondo noi dovrebbe essere un fattore correlato all'aspettativa di vita e decidiamo dunque di eseguire ugualmente un'analisi, nella speranza di ottenere risultati da sfruttare anche in un secondo momento.

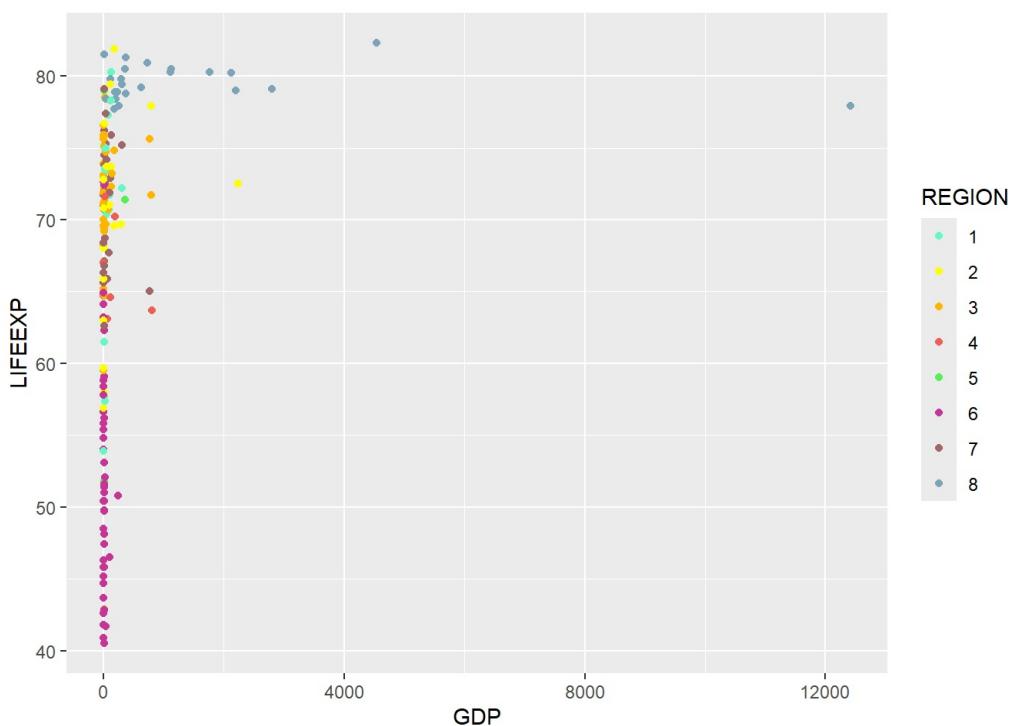


Come potevamo aspettarci dall'indice di Pearson è difficile anche solo visualizzare una retta (diversa da una retta verticale) che approssimi i dati in maniera efficace. Tuttavia potrebbe essere per noi interessante andare a vedere se, considerando solo i paesi con un GDP maggiore ad una certa soglia, otteniamo risultati migliori. Vengono rimossi anche i valori di Stati Uniti e Giappone in quanto costituiscono outliers.

GDP into LIFEEXP



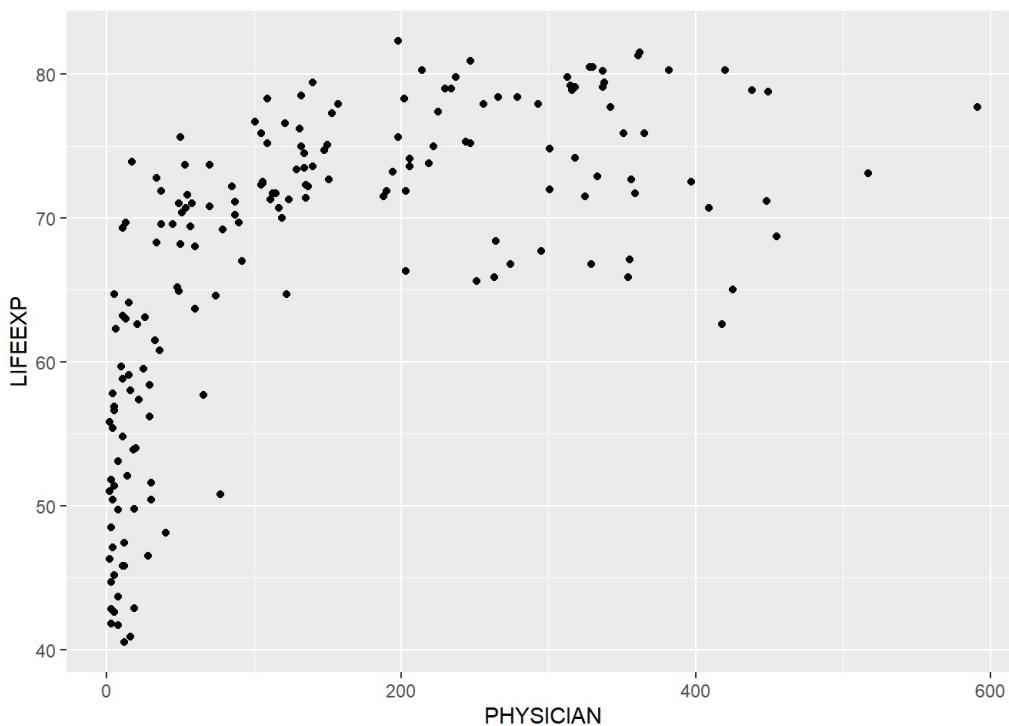
I risultati non sembrano migliorare. Considerando quindi il basso valore dell'indice di correlazione e la rappresentazione grafica ottenuta, riconosciamo la mancanza di una relazione tra GDP e LIFEEXP (anche solo localmente). Come ultimo test proviamo a visualizzare il grafico a dispersione evidenziando le regioni.



Concludiamo che nemmeno i comportamenti regionali sembrano assumere forma lineare, tuttavia potremmo dedurre che REGION definisce abbastanza bene alcuni gruppi in questa analisi, soprattutto per quanto riguarda la regione 6, la 3 e la 8.

PHYSICIAN

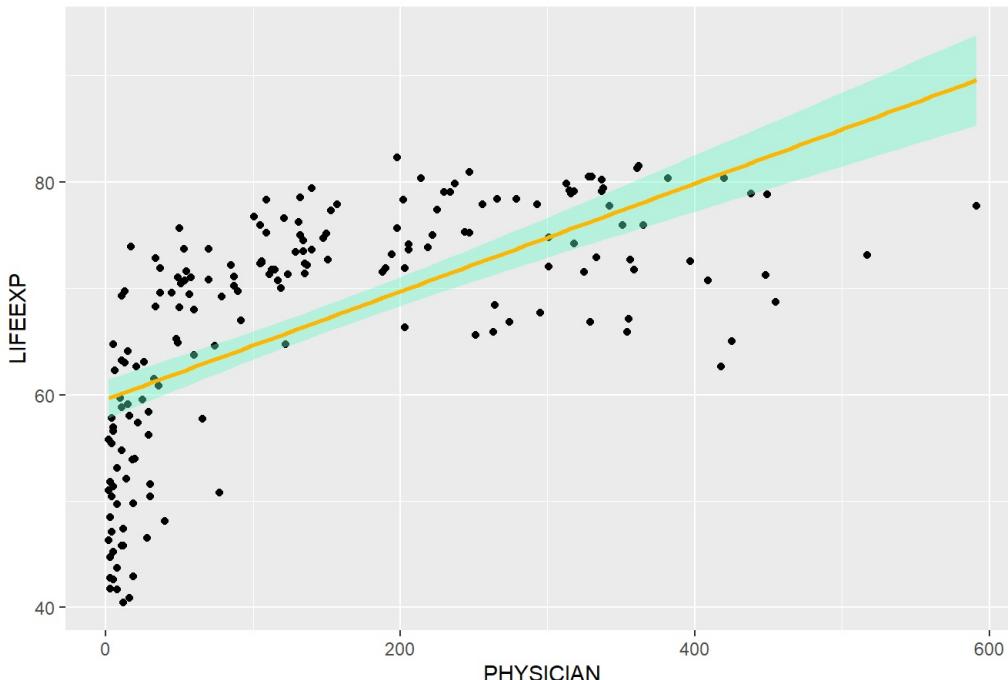
Come già commentato durante l'analisi univariata, abbiamo alte aspettative per la variabile physician di essere influente per LIFEEXP. Cominciamo subito visualizzando lo scatterplot.



Il grafico non fa che confermare i risultati ottenuti con il calcolo dell'indice di Pearson, e dobbiamo ammettere di essere molto soddisfatti da ciò. Proviamo allora a disegnare la retta dei minimi quadrati, visualizzandone anche l'intervallo di confidenza al 95%.

```
## `geom_smooth()` using formula = 'y ~ x'
```

y ~ x



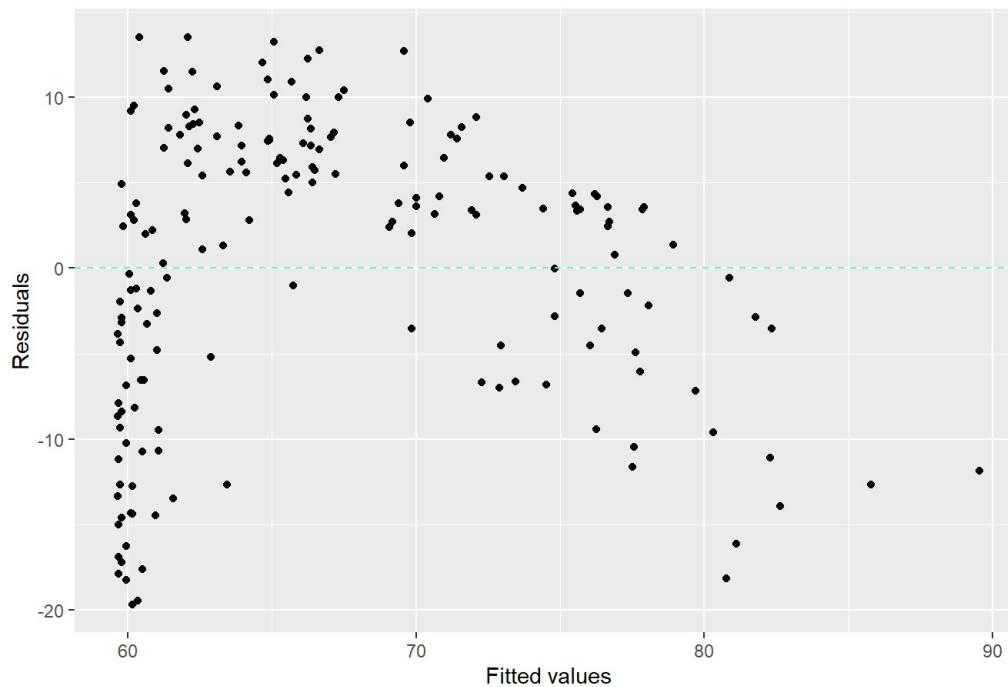
Il risultato ottenuto è davvero buono, e ci lascia soddisfatti. Proviamo ora a valutare la qualità della retta tramite il summary del modello ed ai grafici ResidualsVsFitted e QQS(standardized)Residuals.

```
model = lm(LIFEEXP~PHYSICIAN, data)
summary(model)
```

```
##
## Call:
## lm(formula = LIFEEXP ~ PHYSICIAN, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -19.655  -6.560   2.749   6.959  13.517 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 59.545693  0.929939  64.03 <2e-16 ***
## PHYSICIAN    0.050744  0.004625  10.97 <2e-16 ***
## ---      
## Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 8.621 on 180 degrees of freedom
##     (3 osservazioni eliminate a causa di valori mancanti)
## Multiple R-squared:  0.4008, Adjusted R-squared:  0.3974 
## F-statistic: 120.4 on 1 and 180 DF,  p-value: < 2.2e-16
```

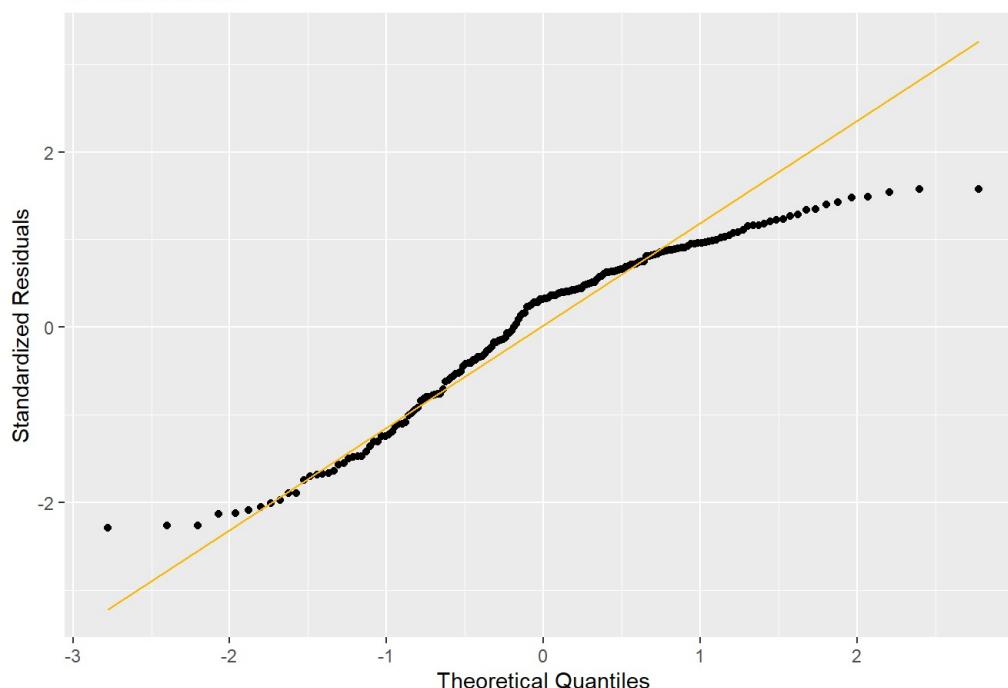
I risultati del summary mostrano un valore di R^2 di 0,3974, abbastanza basso da farci dubitare della relazione lineare semplice. Come potevamo aspettarci vengono individuati alcuni outlier, sia a sinistra che a destra, tuttavia non ci stupisce molto poiché già presupponevamo l'esistenza di questo fenomeno dall'analisi univariata.

Residuals vs Fitted

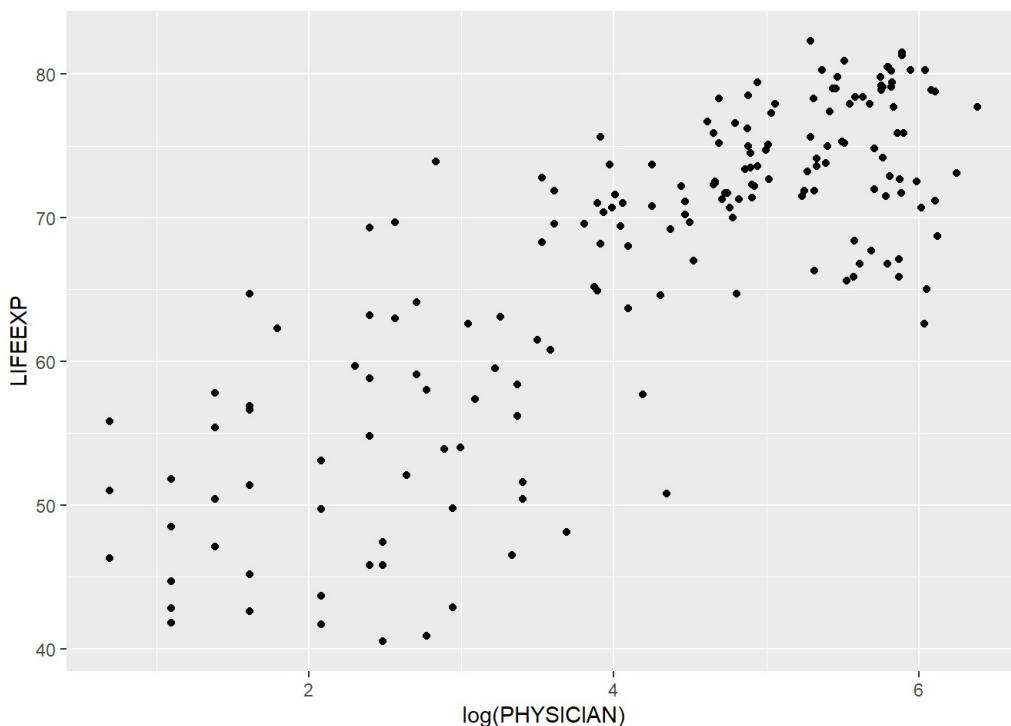


Abbiamo una forte eteroschedasticità e occorre quasi sicuramente pensare di trasformare i dati per ottenere risultati migliori. Vediamo prima di procedere cosa ci indica il grafico QQ SResiduals.

QQ SResiduals

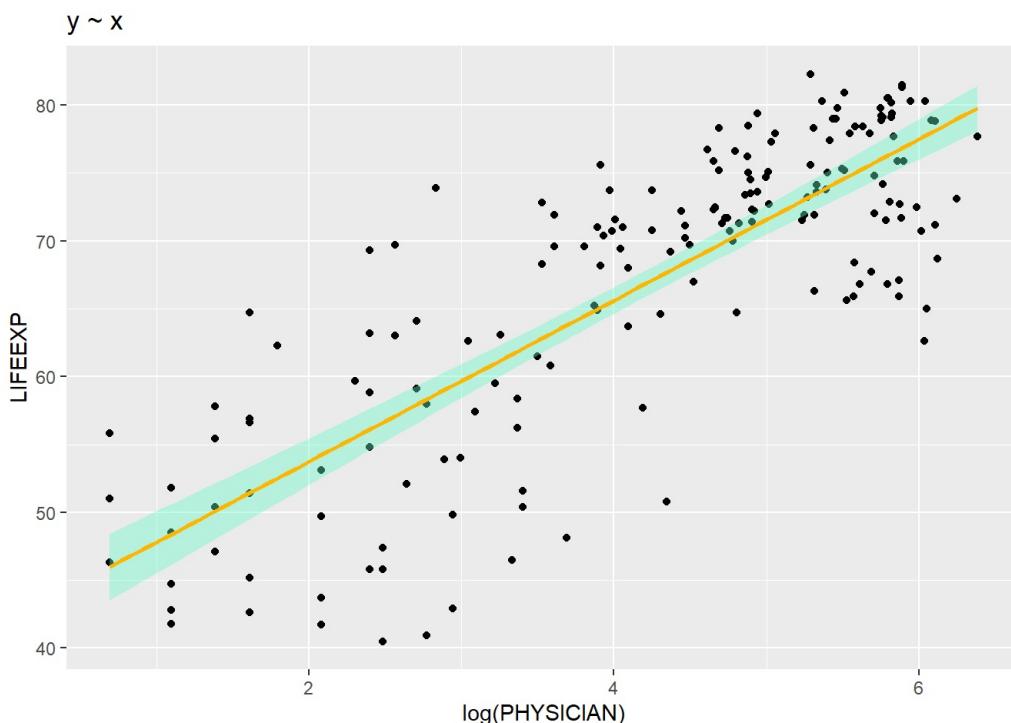


Sembra non esserci una distribuzione degli errori che segue distribuzione normale standard, e come precedentemente già valutato ci viene il dubbio che sia necessario effettuare una trasformazione dei dati per ottenere dei risultati soddisfacenti. Dunque proviamo a considerare il $\log(\text{PHYSICIAN})$.



Notiamo un netto miglioramento nella rappresentazione dei dati, dunque non ci resta che vedere il comportamento della retta di regressione lineare.

```
## `geom_smooth()` using formula = 'y ~ x'
```



Notiamo che la retta approssima sicuramente molto meglio rispetto a prima, ed infatti anche l'intervallo di confidenza si è ristretto rispetto.

Calcoliamo il miglioramento in termini di correlazione tra prima e dopo ed eseguiamo un summary sul modello lineare per visualizzare altri tipi di miglioramento.

```
res = cor(log(data$PHYSICIAN), data$LIFEEXP, method = "pearson", use = "pairwise.complete.obs")
res
```

```
## [1] 0.8065166
```

```
model = lm(LIFEEXP~log(PHYSICIAN) , data)
summary(model)
```

```

## 
## Call:
## lm(formula = LIFEEXP ~ log(PHYSICIAN), data = data)
## 
## Residuals:
##    Min      1Q  Median      3Q     Max 
## -17.426  -3.740   1.169   4.402  15.214 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 41.8539    1.4559   28.75 <2e-16 ***
## log(PHYSICIAN) 5.9411    0.3246   18.30 <2e-16 ***  
## --- 
## Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 6.584 on 180 degrees of freedom
##   (3 osservazioni eliminate a causa di valori mancanti)
## Multiple R-squared:  0.6505, Adjusted R-squared:  0.6485 
## F-statistic: 335 on 1 and 180 DF,  p-value: < 2.2e-16

```

Un miglioramento nella correlazione dell'11% (da 0.72 a 0.8) e un R^2 salito da circa 0,4 a un valore di 0,65. Per provare valutiamo l'ipotesi di usare un modello quadratico e valutiamo i risultati.

```

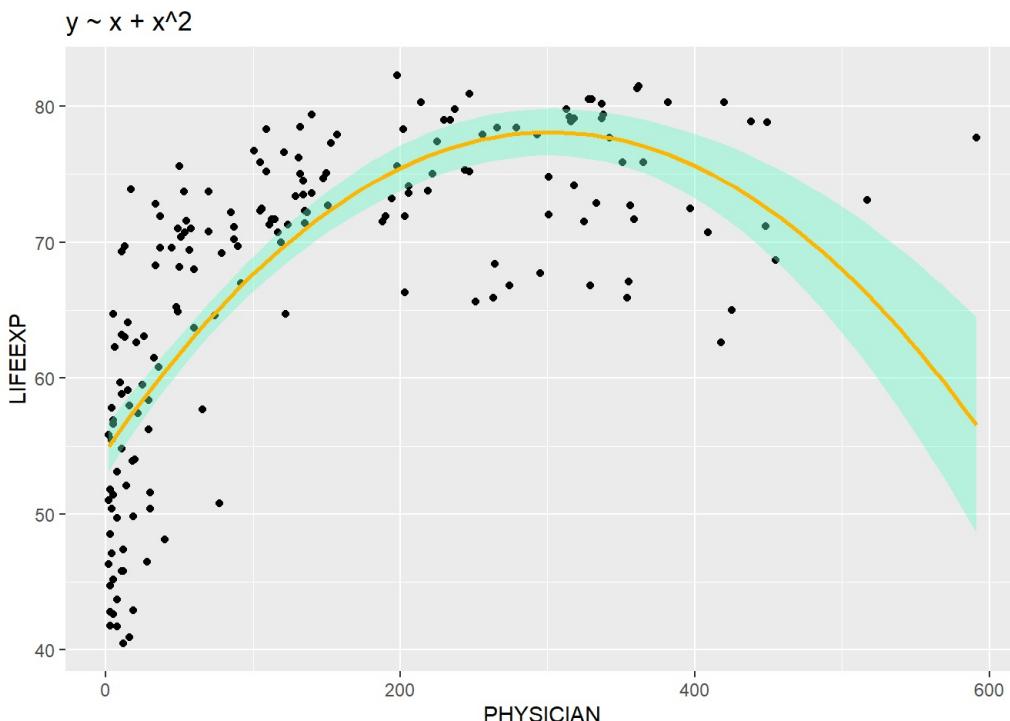
model = lm(LIFEEXP~PHYSICIAN + I(PHYSICIAN^2) , data)
summary(model)

```

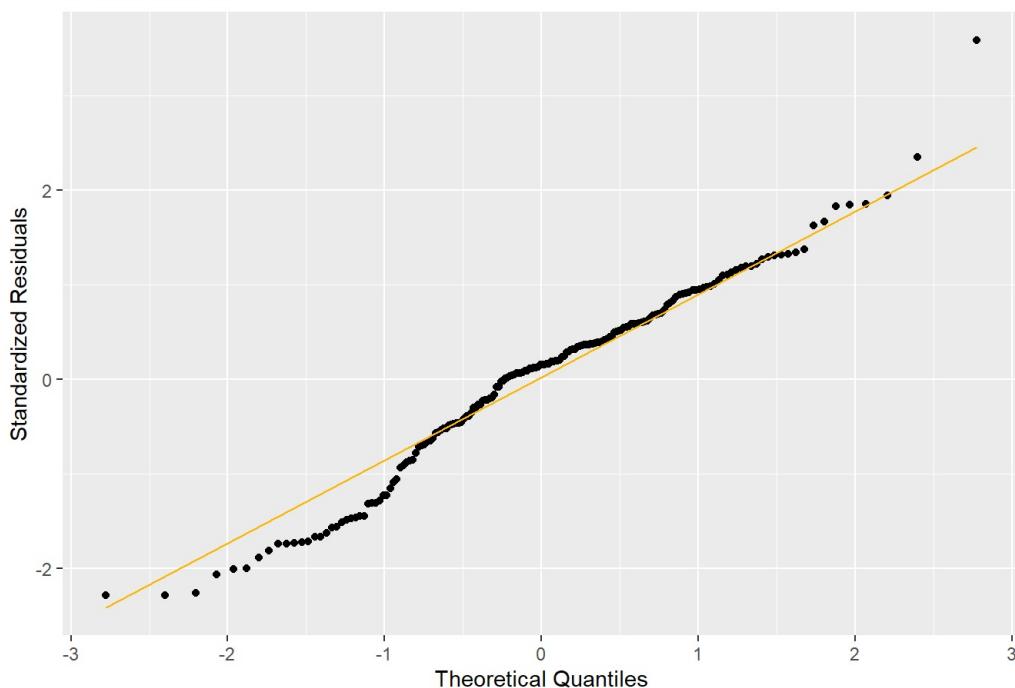
```

## 
## Call:
## lm(formula = LIFEEXP ~ PHYSICIAN + I(PHYSICIAN^2), data = data)
## 
## Residuals:
##    Min      1Q  Median      3Q     Max 
## -16.202  -4.039   1.073   4.330  21.138 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 5.468e+01 9.348e-01  58.500 <2e-16 ***
## PHYSICIAN   1.552e-01 1.203e-02 12.901 <2e-16 ***  
## I(PHYSICIAN^2) -2.573e-04 2.809e-05 -9.159 <2e-16 ***  
## --- 
## Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 7.134 on 179 degrees of freedom
##   (3 osservazioni eliminate a causa di valori mancanti)
## Multiple R-squared:  0.592, Adjusted R-squared:  0.5874 
## F-statistic: 129.9 on 2 and 179 DF,  p-value: < 2.2e-16

```



QQ SResiduals



Notiamo, grazie al QQSResiduals un miglioramento significativo per quanto riguarda le code della distribuzione. Concludiamo dunque la nostra analisi bivariata della variabile PHYSICIAN.

Analisi Multivariata

Analisi multivariata con target LIFEEXP

In questo paragrafo possiamo finalmente dedicarci all'analisi multivariata del nostro dataset con variabile target LIFEEXP. Ci approcceremo in 3 maniere diverse:

1. Aggiunta delle variabili più rilevanti trovate sino ad ora tramite le analisi precedentemente eseguite in un modello unico
2. Uso del metodo stepAIC della libreria MASS a partire dal modello nullo al modello che comprende tutte le variabili in maniera lineare
3. Combinazione dei risultati ottenuti dagli approcci 1 e 2 provando a trasformare qualora servissero le variabili considerate

APPROCCIO 1

Come primo approccio proviamo a prendere le variabili che hanno dato i risultati migliori nell'analisi bivariata per poi utilizzarle in un modello di regressione lineare multivariata. Si decide di selezionare le variabili che hanno mostrato un indice R^2 superiore a 0,6 ovvero FERTILITY, PHYSICIAN (il suo logaritmo), HEALTHEXPEND (anche in questo caso il logaritmo).

```
model <- lm(LIFEEXP ~ FERTILITY + log(PHYSICIAN) + log(HEALTHEXPEND), data = data)
summary(model)
```

```
##
## Call:
## lm(formula = LIFEEXP ~ FERTILITY + log(PHYSICIAN) + log(HEALTHEXPEND),
##      data = data)
##
## Residuals:
##       Min     1Q   Median     3Q    Max 
## -21.5130 -2.5545  0.8333  3.4719 11.4530 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 49.5135    4.7464 10.432 < 2e-16 ***
## FERTILITY   -1.9510    0.5338 -3.655 0.000342 ***
## log(PHYSICIAN) 2.4435    0.5792  4.219 3.98e-05 ***
## log(HEALTHEXPEND) 2.3256    0.5201  4.472 1.41e-05 ***
## ---      
## Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 5.753 on 171 degrees of freedom
##   (10 osservazioni eliminate a causa di valori mancanti)
## Multiple R-squared:  0.7388, Adjusted R-squared:  0.7343 
## F-statistic: 161.3 on 3 and 171 DF,  p-value: < 2.2e-16
```

Notiamo subito un miglioramento dell'indice R^2 rispetto alle analisi bivariate, risultato che ci aspettavamo considerando come sono state scelte le variabili predittive. Proviamo ora ad aggiungere la variabile categoriale REGION nel modello, dato che la ritengiamo importante nella determinazione delle aspettative di vita.

```
model <- lm(LIFEEXP ~ FERTILITY + log(PHYSICIAN) + log(HEALTHEXPEND) + REGION, data = data)
summary(model)
```

```
## 
## Call:
## lm(formula = LIFEEXP ~ FERTILITY + log(PHYSICIAN) + log(HEALTHEXPEND) +
##     REGION, data = data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -16.3015 -1.7300  0.1007  2.8625 12.7747 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 59.6481   5.3894 11.068 < 2e-16 ***
## FERTILITY   -1.6572   0.5242 -3.162 0.00187 **  
## log(PHYSICIAN) 1.6384   0.5903  2.776 0.00615 **  
## log(HEALTHEXPEND) 1.5311   0.5902  2.594 0.01033 *   
## REGION2    -0.1737   1.6753 -0.104 0.91754    
## REGION3    -0.1689   1.4732 -0.115 0.90886    
## REGION4    -3.5194   2.1326 -1.650 0.10080    
## REGION5    -1.5706   3.1145 -0.504 0.61474    
## REGION6    -9.5674   1.7172 -5.572 1.02e-07 ***
## REGION7    -4.4165   1.6461 -2.683 0.00804 **  
## REGION8     1.0742   1.8195  0.590 0.55574    
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 4.941 on 164 degrees of freedom
## (10 osservazioni eliminate a causa di valori mancanti)
## Multiple R-squared:  0.8152, Adjusted R-squared:  0.8039 
## F-statistic: 72.34 on 10 and 164 DF,  p-value: < 2.2e-16
```

```
AIC(model)
```

```
## [1] 1068.447
```

Osserviamo un aumento significativo dell'indice R^2, a conferma della nostra ipotesi. Abbiamo quindi ottenuto un buon modello predittivo utilizzando quattro variabili

APPROCCIO 2

Il metodo stepAIC() è una funzione di selezione del modello che utilizza il criterio di informazione di Akaike (AIC) per determinare quali variabili includere nel modello. Partendo dal modello più semplice possibile (il modello nullo) ed aggiungendo variabili (talvolta rimuovendole se serve) il metodo seleziona ad ogni passo il modello con il minor AIC fino ad arrivare ad un punto in cui non può più migliorare (il miglior modello raggiungibile).

```
clean_data = na.omit(data)
nrow(clean_data)
```

```
## [1] 131
```

```
null_model = lm(LIFEEXP ~ 1, clean_data)
full_model = lm(LIFEEXP ~ ., clean_data)

best_model = stepAIC(null_model, trace = 0, scope = list(lower = null_model, upper = full_model), direction = "both")

summary(best_model)
```

```

## 
## Call:
## lm(formula = LIFEEXP ~ REGION + BIRTHATTEND + PUBLICEDUCATION +
##     ILLITERATE + FERTILITY + HEALTHEXPEND, data = clean_data)
##
## Residuals:
##      Min    1Q   Median    3Q   Max 
## -12.5141 -2.2139  0.0434  2.4993 15.8932 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 6.579e+01 5.319e+00 12.369 < 2e-16 ***
## REGION2     -3.177e+00 1.951e+00 -1.629 0.106062  
## REGION3     -1.523e+00 1.769e+00 -0.861 0.391045  
## REGION4     -4.095e+00 2.728e+00 -1.501 0.136004  
## REGION5     -3.362e+00 3.106e+00 -1.082 0.281330  
## REGION6     -1.642e+01 1.810e+00 -9.072 3.13e-15 ***
## REGION7     -4.672e+00 1.910e+00 -2.446 0.015914 *  
## REGION8     -3.468e-02 2.773e+00 -0.013 0.990045  
## BIRTHATTEND 1.431e-01 3.808e-02 3.757 0.000269 *** 
## PUBLICEDUCATION -5.372e-01 2.186e-01 -2.457 0.015459 *  
## ILLITERATE   1.076e-01 4.143e-02 2.596 0.010624 *  
## FERTILITY    -1.531e+00 6.585e-01 -2.325 0.021758 *  
## HEALTHEXPEND 1.544e-03 8.421e-04 1.834 0.069174 .  
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.598 on 118 degrees of freedom
## Multiple R-squared:  0.8471, Adjusted R-squared:  0.8315 
## F-statistic: 54.46 on 12 and 118 DF, p-value: < 2.2e-16

```

```
AIC(best_model)
```

```
## [1] 785.7952
```

Il summary del best_model è soddisfacente sotto alcuni punti di vista e inaspettato sotto altri. Come si poteva prevedere, ancora a partire dalle analisi univariate, la regione 6 è una variabile molto incidente per la previsione di LIFEEXPECT. Il fatto che anche HEALTHEXPEND sia coinvolta nel modello conferma i ragionamenti che avevamo valutato durante l'analisi bivariata. Ci sorprendono invece BIRTHATTEND che, nonostante avesse un R^2 inferiori a 0.6 nell'analisi bivariata, ora diventa la seconda variabile per importanza di p-value. Anche PUBLICEDUCATION, che era stata scartata per un livello di correlazione troppo basso, ora ha un p-value significativo. Ci aspettavamo che PHYSICIAN rientrasse nel modello finale.

APPROCCIO 3

Ottenuti i risultati dall'approccio 1 e dall'approccio 2, proviamo ora a combinarli per ottenere qualcosa di ancora più efficace. Creiamo un modello con variabili indipendenti l'intersezione delle variabili inserite nell'approccio 1 e nell'approccio 2.

```

model = lm(LIFEEXP ~ REGION + BIRTHATTEND + PUBLICEDUCATION +
           ILLITERATE + FERTILITY + HEALTHEXPEND + log(PHYSICIAN) + log(HEALTHEXPEND), clean_data)

summary(model)

```

```

## 
## Call:
## lm(formula = LIFEEXP ~ REGION + BIRTHATTEND + PUBLICEDUCATION +
##     ILLITERATE + FERTILITY + HEALTHEXPEND + log(PHYSICIAN) +
##     log(HEALTHEXPEND), data = clean_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -12.7096 -2.1890 -0.2288  2.4051 15.1200 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 5.229e+01 7.639e+00 6.845 3.84e-10 ***
## REGION2     -9.217e-01 2.103e+00 -0.438 0.66196  
## REGION3     -7.708e-01 1.757e+00 -0.439 0.66163  
## REGION4     -3.261e+00 2.694e+00 -1.210 0.22860  
## REGION5     -2.075e+00 3.081e+00 -0.673 0.50202  
## REGION6     -1.378e+01 2.085e+00 -6.610 1.23e-09 ***
## REGION7     -3.949e+00 1.966e+00 -2.009 0.04686 *  
## REGION8      9.528e-01 2.772e+00 0.344 0.73169  
## BIRTHATTEND  1.016e-01 4.059e-02 2.502 0.01375 *  
## PUBLICEDUCATION -5.552e-01 2.224e-01 -2.496 0.01396 *  
## ILLITERATE    1.374e-01 4.237e-02 3.243 0.00154 ** 
## FERTILITY     -9.207e-01 7.010e-01 -1.313 0.19163  
## HEALTHEXPEND  5.656e-04 1.063e-03 0.532 0.59552  
## log(PHYSICIAN) 1.586e+00 7.711e-01 2.057 0.04197 *  
## log(HEALTHEXPEND) 1.221e+00 9.628e-01 1.269 0.20711 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.504 on 116 degrees of freedom
## Multiple R-squared:  0.8558, Adjusted R-squared:  0.8383 
## F-statistic: 49.16 on 14 and 116 DF,  p-value: < 2.2e-16

```

```
AIC(model)
```

```
## [1] 782.135
```

Siamo molto soddisfatti del risultato, poiché siamo riusciti a diminuire l'AIC del modello e a rendere PHYSICIAN rilevante ai fini delle previsioni, oltre ad aver aumentato leggermente R^2. Proviamo ora ad inserire dei termini al quadrato ed a rimuovere alcune variabili e verifichiamo che succede.

```

model = lm(LIFEEXP ~ REGION + BIRTHATTEND + PUBLICEDUCATION +
ILLITERATE + FERTILITY + log(PHYSICIAN) + log(HEALTHEXPEND) + I(PHYSICIAN^3), clean_data)

summary(model)

```

```

## 
## Call:
## lm(formula = LIFEEXP ~ REGION + BIRTHATTEND + PUBLICEDUCATION +
##      ILLITERATE + FERTILITY + log(PHYSICIAN) + log(HEALTHEXPEND) +
##      I(PHYSICIAN^3), data = clean_data)
##
## Residuals:
##       Min     1Q   Median     3Q    Max 
## -12.8320 -2.1922  0.0772  2.2127 14.6485 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 4.975e+01 7.110e+00 6.997 1.80e-10 ***
## REGION2     -6.693e-01 2.092e+00 -0.320 0.749576    
## REGION3     -5.441e-01 1.750e+00 -0.311 0.756415    
## REGION4     -3.352e+00 2.673e+00 -1.254 0.212415    
## REGION5     -1.826e+00 3.051e+00 -0.599 0.550641    
## REGION6     -1.326e+01 2.098e+00 -6.320 5.03e-09 ***
## REGION7     -3.049e+00 1.989e+00 -1.533 0.128106    
## REGION8     2.605e+00 2.083e+00 1.251 0.213557    
## BIRTHATTEND 9.427e-02 4.023e-02 2.343 0.020839 *  
## PUBLICEDUCATION -5.289e-01 2.211e-01 -2.392 0.018354 *  
## ILLITERATE 1.440e-01 4.223e-02 3.409 0.000896 *** 
## FERTILITY -8.818e-01 6.872e-01 -1.283 0.201983    
## log(PHYSICIAN) 2.021e+00 8.274e-01 2.443 0.016071 *  
## log(HEALTHEXPEND) 1.429e+00 7.477e-01 1.912 0.058399 . 
## I(PHYSICIAN^3) -3.362e-08 2.326e-08 -1.445 0.151016    
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.469 on 116 degrees of freedom
## Multiple R-squared:  0.858, Adjusted R-squared:  0.8408 
## F-statistic: 50.05 on 14 and 116 DF, p-value: < 2.2e-16

```

AIC(model)

```
## [1] 780.116
```

I risultati ottenuti sono davvero soddisfacenti. Abbiamo ottenuto risultati migliori sia in termini di AIC che di R². Questo ci aiuta a renderci conto di quanto, tutto il percorso di analisi, a partire dall'univariata in poi, sia stato utile per arrivare a questo punto. Ad evidenza di ciò, come ultima cosa, proviamo a vedere lo stepAIC cosa avrebbe generato se non avessimo, per esempio, scartato a priori le variabili con eccessivi NA.

```
t = read.csv("life expectation.csv", header = TRUE)

clean_t = na.omit(t)
nrow(clean_t)
```

```
## [1] 45
```

```
clean_t = clean_t[, !(names(clean_t) %in% c("COUNTRY"))]
null_model = lm(LIFEEXP ~ 1, clean_t)

full_model = lm(LIFEEXP ~ ., clean_t)

best_model = stepAIC(null_model, scope = list(lower = null_model, upper = full_model), direction = "both", trace = 0)

summary(best_model)
```

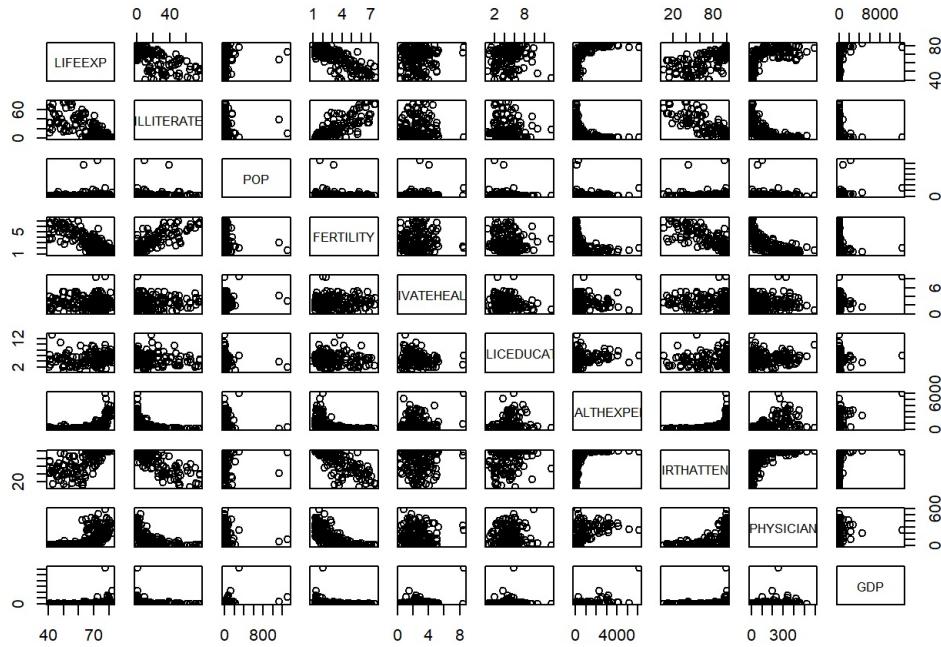
```

## 
## Call:
## lm(formula = LIFEEXP ~ HEALTHEXPEND + SMOKING + PRIVATEHEALTH +
##     REGION + FEMALEBOSS + ILLITERATE + PUBLICEDUCATION, data = clean_t)
##
## Residuals:
##   Min     1Q Median     3Q    Max 
## -4.6148 -1.2980  0.3245  1.0206  4.6715 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 88.4079954  3.2672898 27.059 < 2e-16 ***
## HEALTHEXPEND 0.0025751  0.0004974  5.177 8.16e-06 ***
## SMOKING     -0.1333163  0.0365972 -3.643 0.000822 *** 
## PRIVATEHEALTH -0.9399028  0.3258326 -2.885 0.006499 **  
## REGION      -0.4639067  0.2491147 -1.862 0.070528 .  
## FEMALEBOSS    -0.1197846  0.0419342 -2.856 0.006988 ** 
## ILLITERATE    -0.2064763  0.0408112 -5.059 1.17e-05 *** 
## PUBLICEDUCATION -0.5578627  0.3738535 -1.492 0.144126 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.316 on 37 degrees of freedom
## Multiple R-squared:  0.8095, Adjusted R-squared:  0.7735 
## F-statistic: 22.46 on 7 and 37 DF,  p-value: 1.603e-11

```

Possiamo vedere che, nonostante le variabili aggiuntive di cui stepAIC era a disposizione l'R^2 sia al di sotto della soglia precedentemente ottenuta ed, inoltre, ricordiamo che questo modello lavora su un dataset che all'incirca contiene il 30% delle osservazioni rispetto a prima.

Clustering e PCA



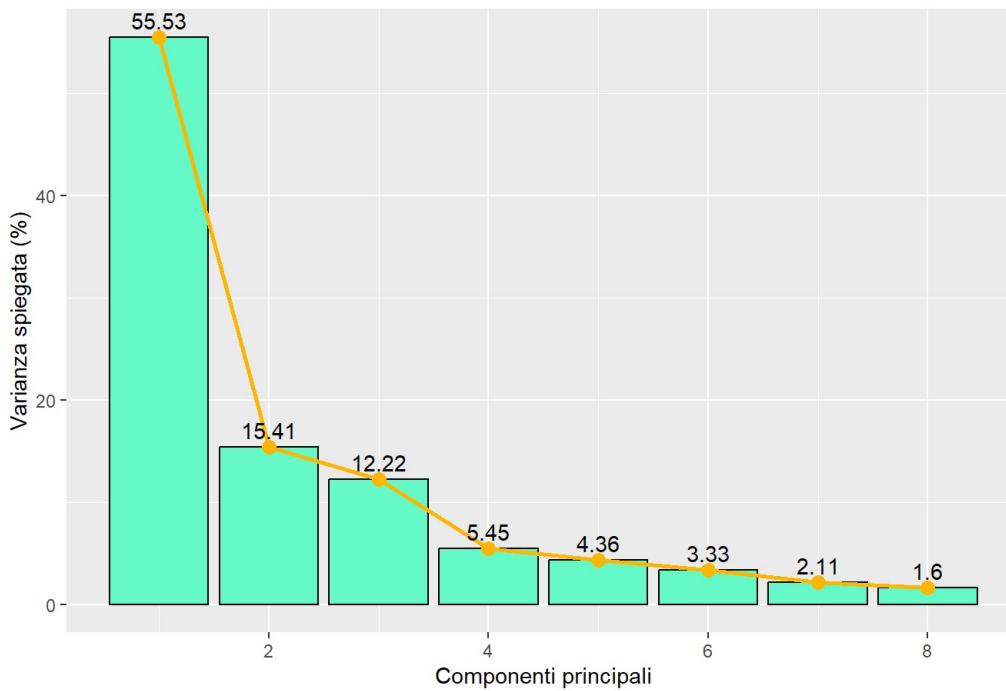
Si nota come i grafici a dispersione delle variabili prese a coppie non mostrino l'esistenza di gruppi separati su cui vale la pena eseguire un'operazione di categorizzazione grazie al clustering.

Ulteriori studi sul dataset

PCA

Non avendo informazioni su come siano state selezionate le regioni di partenza proviamo a ricavare 8 cluster basati sulle 8 variabili numeriche selezionate (senza quelle escluse) e confrontare i gruppi con quelli dati. Rieselezioniamo le variabili per poter eseguire una PCA sul dataset. Poi eseguiamo la PCA per provare ad eseguire clustering basandoci sulle componenti principali individuate.

Varianza spiegata dalle componenti principali



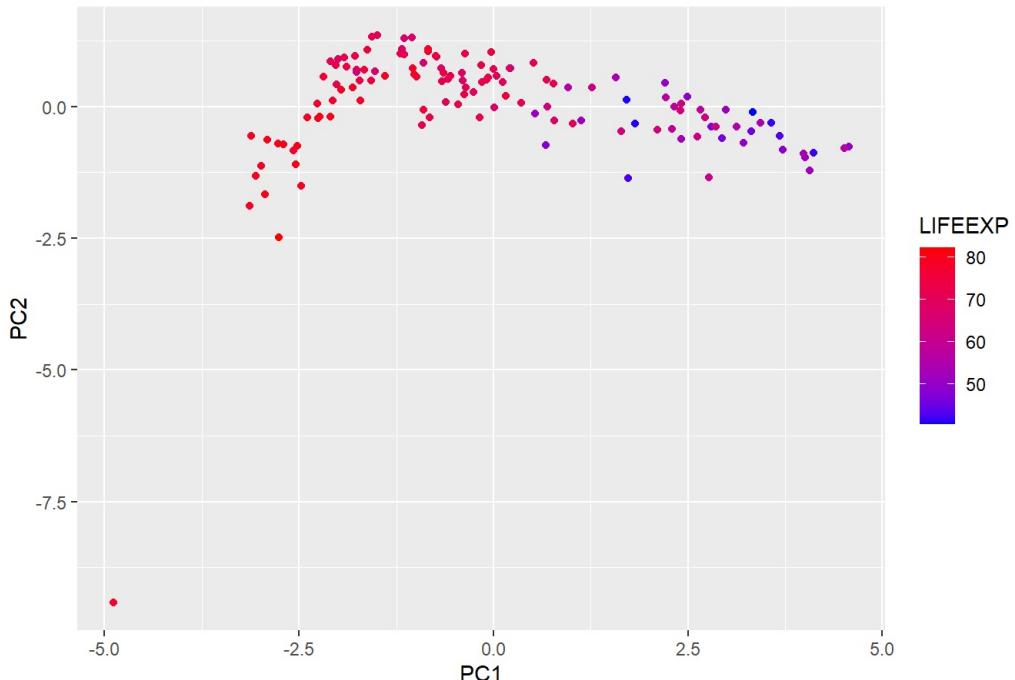
Si nota come risulta difficile ridurre la dimensionalità in quanto le prime 2 componenti principali siano in grado di spiegare solamente il 70% della varianza. Per arrivare ad un valore accettabile ($>90\%$) sarebbe necessario prendere le prime 5 componenti e considerando che si parte da 8 variabili, la riduzione non risulterebbe molto significativa.

pca

```
## Standard deviations (1, ..., p=8):
## [1] 2.1077075 1.1101527 0.9887782 0.6600603 0.5908420 0.5157935 0.4105515
## [8] 0.3578808
##
## Rotation (n x k) = (8 x 8):
##          PC1       PC2       PC3       PC4       PC5
## BIRTHATTEND -0.41857650  0.19280199 -0.039789008 -0.37478950 -0.01917665
## PUBLICEDUCATION -0.09909576 -0.28068350 -0.930934699 -0.06632341 -0.07083912
## ILLITERATE     0.41472095 -0.19510568  0.018065251  0.43360686 -0.22968532
## FERTILITY      0.43251212 -0.17092743  0.004079726  0.09021162  0.12334550
## HEALTHEXPEND   -0.33148590 -0.50868925  0.027401393  0.42975461 -0.01710392
## PHYSICIAN      -0.39515081  0.11815003  0.002520742  0.49124183  0.63975489
## GDP            -0.16075340 -0.73332143  0.335927714 -0.39919400  0.09232454
## LIFEEXP        -0.40502448  0.07922906  0.133528290  0.27132558 -0.71312126
##          PC6       PC7       PC8
## BIRTHATTEND   -0.2912095  0.67852169 -0.31624099
## PUBLICEDUCATION 0.1645921  0.03508452  0.08377914
## ILLITERATE     0.2843663  0.41334696 -0.54509063
## FERTILITY      -0.3422606  0.51037910  0.61864288
## HEALTHEXPEND   -0.6221856 -0.16908678 -0.17295848
## PHYSICIAN      0.3433012  0.21612114  0.12133084
## GDP            0.3743762  0.10378488  0.06886249
## LIFEEXP        0.2120653  0.14479248  0.40466287
```

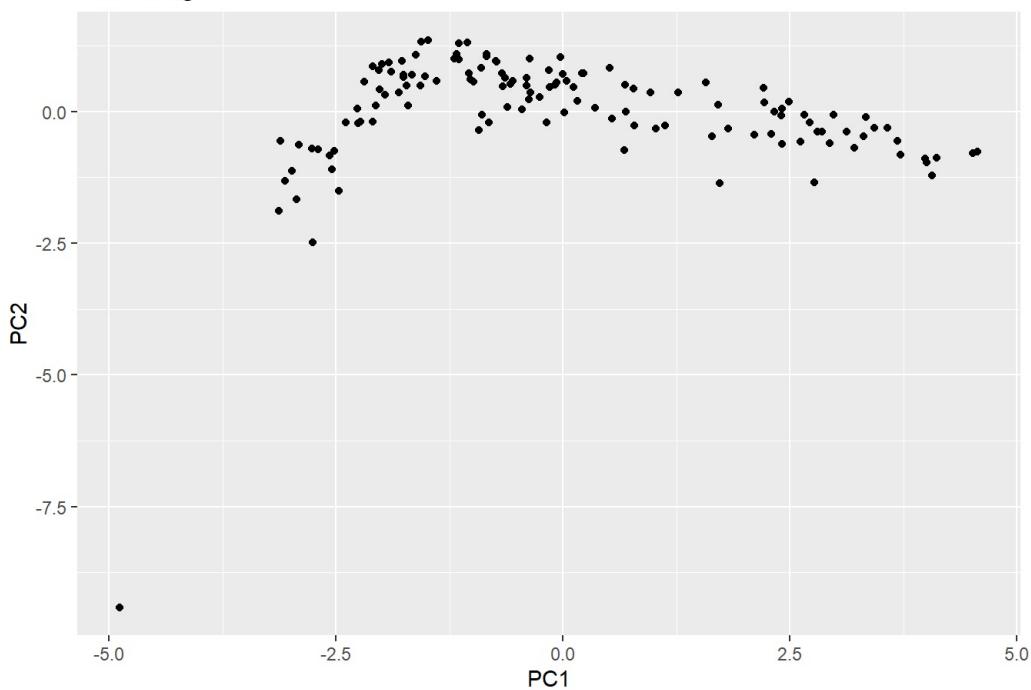
Plottiamo le due componenti principali.

PCA

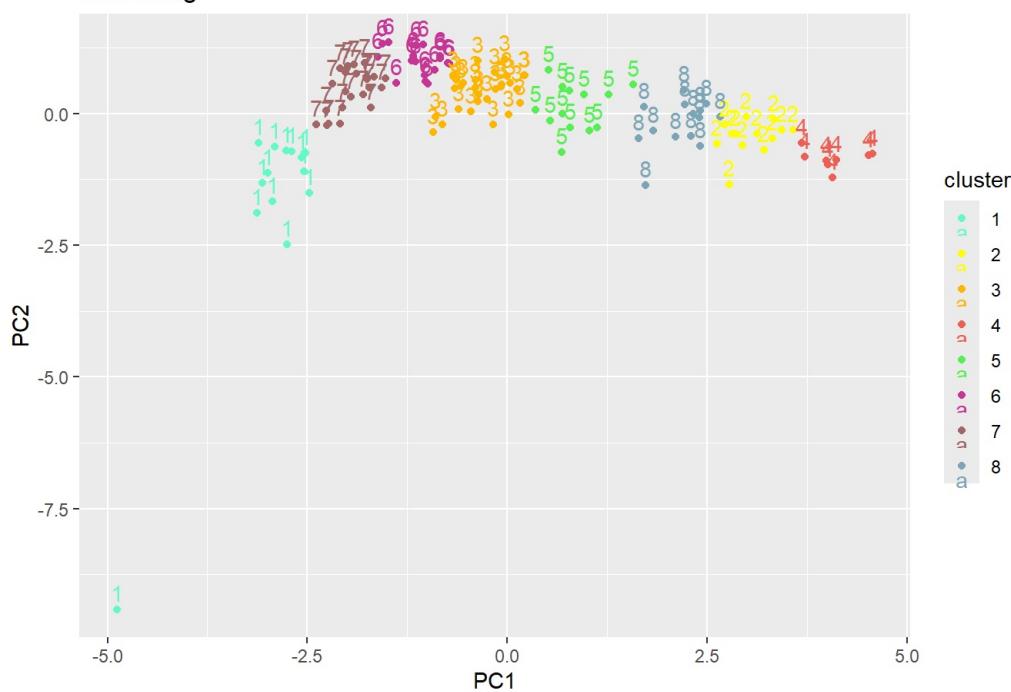


Eseguiamo un primo tentativo di clustering usando kmeans e 5 centroidi.

Clustering

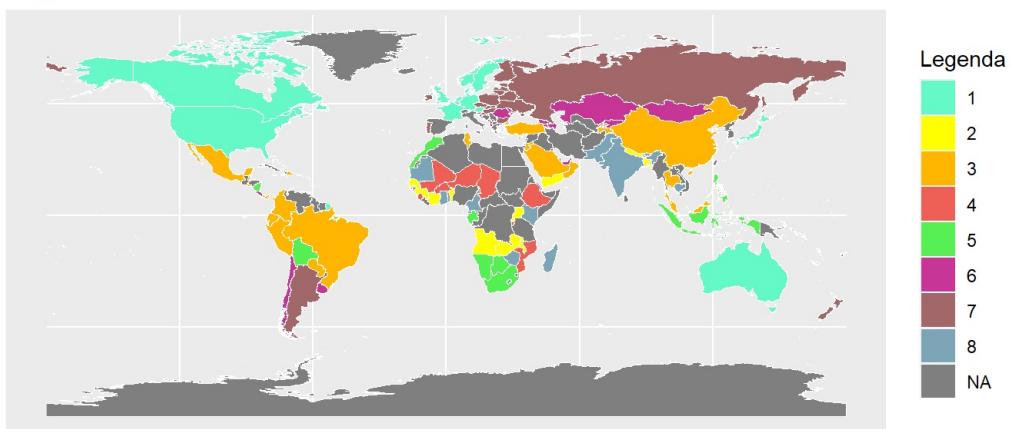


Clustering



Visualizziamo la mappa con le nuove regioni trovate.

Regioni



Si nota come i cluster trovati non siano molto significativi in quanto sono presenti moltissimi valori nulli. Nonostante ciò l'attuale gruppo 1 corrisponde abbastanza accuratamente alla regione 8 così come il gruppo 7 alla regione 7. La regione 6 (corrispondente all'Africa) in questo clustering è stata completamente divisa in altri gruppi. Alla regione 4 di partenza si riesce ad associare discretamente bene l'attuale gruppo 8 con l'aggiunta di altri paesi soprattutto appartenenti al continente africano. Riconoscete queste associazioni, concludiamo che i gruppi così trovati non forniscono intuizioni significative per interpretare la formazione dei gruppi presente nel dataset.