




Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II.

doi: 10.18637/jss.v000.i00

Analisi e valutazione del benessere della società: caso studio sugli indici di vita

Carmela Pia Senatore 
Università degli studi di Salerno

1. Descrizione del dataset

Il **Better Life Index** è un database globale che raccoglie dati provenienti da circa 40 Stati in tutto il mondo. L'obiettivo principale del database è coinvolgere e rendere consapevoli i cittadini in un dibattito significativo sulla valutazione del benessere della società e consentire di partecipare attivamente al processo di formulazione delle politiche che influenzano la qualità della vita di tutti. Questo strumento si propone di rendere le persone più informate e partecipi nell'ambito della governance.

Il Better Life Index è suddiviso in 11 sottotemi o macroaree, ognuna delle quali è supportata da uno a tre indicatori. All'interno di ciascuna delle macroaree, viene calcolata una media degli indicatori, ciascuno dei quali ha lo stesso peso nella valutazione. La scelta degli indicatori si basa su criteri statistici rigorosi, come la pertinenza (ovvero quanto un indicatore rifletta in modo adeguato il concetto che si intende misurare), la profondità (la capacità di catturare aspetti rilevanti del benessere), e la rilevanza politica (cioè quanto l'indicatore è utile per le decisioni politiche). Inoltre, la qualità dei dati è un criterio fondamentale nella selezione degli indicatori. Questo include la validità predittiva (la capacità di un indicatore di prevedere il benessere futuro), la copertura (quanto ampiamente l'indicatore può essere applicato a diverse realtà), la tempestività (la disponibilità dei dati in tempo reale o con aggiornamenti regolari), e la comparabilità tra paesi (la possibilità di confrontare dati tra nazioni diverse). Tutti questi criteri sono stabiliti in consultazione con l'Organizzazione per la Cooperazione e lo Sviluppo Economico (OCSE) e i paesi membri di questa organizzazione. Gli indicatori selezionati sono quindi considerati buone misure dei concetti di benessere, specialmente quando si tratta di condurre un confronto tra paesi. Il Better Life Index mira quindi a fornire un quadro informativo completo e affidabile per valutare il benessere delle società a livello globale, incoraggiando un coinvolgimento significativo dei cittadini nel processo decisionale.

E', quindi, un dataset multivariato che contiene più osservazioni, ciascuna con un set completo di valori per le variabili coinvolte. Questo significa che il dataset è costituito da una matrice di dati in cui le righe rappresentano le osservazioni e le colonne rappresentano le variabili. Sono presenti dati prevalentemente di tipo quantitativi.

Il dataset include le seguenti macroaree e relative variabili:

- **Housing**

- **Dwellings without basic facilities:** Questo indicatore si riferisce alla percentuale della popolazione che vive in una residenza senza un water destinato esclusivamente all'uso delle loro famiglie. I water situati all'esterno della residenza non sono da considerare in questa categoria. Inoltre, vengono conteggiati anche i water situati in una stanza in cui è presente anche un'unità doccia o una vasca da bagno. Essenzialmente esprime la percentuale di famiglie che vive senza utilità di tipo primarie.
- **Housing expenditure:** Questo indicatore considera la spesa delle famiglie per l'abitazione e la manutenzione della casa, come definito nel Sistema delle Nazioni Unite (P31CP040: Alloggio, acqua, elettricità, gas e altri combustibili; P31CP050: Arredi, attrezzature per la casa e manutenzione ordinaria della casa). Essa comprende l'affitto effettivo e imputato per l'alloggio, la spesa per la manutenzione e la riparazione dell'abitazione (compresi servizi vari), l'approvvigionamento idrico,

l'energia elettrica, il gas e altri combustibili, beni e servizi per la manutenzione ordinaria della casa come percentuale del reddito lordo disponibile delle famiglie.

- **Rooms per person:** si riferisce al numero di stanze (escludendo cucina, dispensa, bagno, gabinetto, garage, studi medici, ufficio, negozio) in un'abitazione diviso per il numero di persone che vivono nell'abitazione.

- **Income**

- **Household net adjusted disposable income:** l'importo massimo che una famiglia può permettersi di consumare senza dover ridurre il proprio patrimonio o aumentare i propri debiti. Si ottiene sommando al reddito lordo delle persone (guadagni, reddito da lavoro autonomo) i trasferimenti sociali in natura che le famiglie ricevono dai governi, e quindi sottraendo le tasse sul reddito e sulla ricchezza.
- **Household net financial wealth:** la ricchezza finanziaria netta composta da: oro monetario, valuta e depositi, titoli diversi dalle azioni, prestiti, azioni e altri strumenti patrimoniali (inclusi i titoli emessi dai fondi d'investimento).

- **Jobs**

- **Employment rate:** Si tratta del numero di persone occupate di età compresa tra 15 e 64 anni rispetto alla popolazione della stessa fascia d'età. Le persone occupate sono coloro che hanno 15 anni o più e che dichiarano di aver lavorato in un'occupazione retribuita.
- **Job security:** l'indicatore rappresenta il numero di dipendenti occupati con una anzianità lavorativa inferiore a 6 mesi rispetto all'occupazione dipendente totale. L'anzianità lavorativa è misurata in base alla durata del tempo durante il quale i lavoratori hanno svolto il loro attuale o principale lavoro o sono stati impiegati presso il loro attuale datore di lavoro.
- **Long-term unemployment rate:** questo indicatore si riferisce al numero di persone che sono state disoccupate per un anno o più, espresso come percentuale della forza lavoro (la somma delle persone occupate e delle persone disoccupate).
- **Personal earnings:** l'indicatore si riferisce alla retribuzione annua media per equivalente a tempo pieno di un dipendente, ottenuta dividendo il totale dei salari basati sui conti nazionali per il numero medio di dipendenti nell'economia complessiva, moltiplicato poi per il rapporto tra le ore medie settimanali solitamente lavorate da un dipendente a tempo pieno e le ore settimanali solitamente lavorate da tutti i dipendenti.

- **Community:**

- **Quality of support network:** esprime una misura del supporto percepito all'interno della rete sociale. Questo indicatore si basa sulla domanda: "Se avessi dei problemi, hai parenti o amici su cui puoi contare per aiutarti ogni volta che ne hai bisogno o no?" e tiene conto delle risposte positive da parte dei partecipanti. In altre parole, misura la percezione delle persone sulla disponibilità di parenti o amici pronti ad aiutarle in caso di necessità.

- **Education**

- **Educational attainment:** si tratta del numero di individui che hanno completato almeno con successo l'istruzione di livello secondario superiore rispetto alla popolazione compresa tra i 25 e i 64 anni.
- **Student skills:** Il punteggio medio degli studenti nei test di lettura, matematica e scienze, valutato dal Programma per la Valutazione Internazionale degli Studenti (PISA) dell'OCSE. PISA è un programma di valutazione globale che misura le competenze degli studenti in queste tre aree chiave ed è utilizzato per confrontare le performance degli studenti in tutto il mondo. Il punteggio medio riflette le capacità medie degli studenti in queste materie in un determinato paese o regione.
- **Years in education:** durata media degli studi in anni.

- **Environment**

- **Air pollution:** si tratta di una misura della qualità dell'aria nelle aree urbane, con un'enfasi sulle città di dimensioni maggiori, in cui le concentrazioni di PM10 vengono ponderate in base alla popolazione residente. Questo indicatore fornisce una stima della contaminazione atmosferica e del potenziale impatto sulla salute dei residenti nelle aree urbane.
- **Water quality:** misura il grado di soddisfazione delle persone riguardo alla qualità dell'acqua nell'area in cui vivono, basandosi sulle loro risposte soggettive.

- **Civic engagement**

- **Consultation on rule-making:** L'indicatore è una media ponderata delle risposte sì/no a varie domande sulla presenza di consultazioni legali da parte dei cittadini e sulla presenza di procedure formali che consentano al pubblico in generale di influenzare la regolamentazione e le azioni governative. L'indicatore descrive l'entità in cui processi formali di consultazione sono integrati in fasi chiave della progettazione di proposte di regolamentazione e quali meccanismi esistono affinché l'esito di tali consultazioni influenzi la preparazione di leggi primarie e regolamenti subordinati. Le domande su cui si basa l'indicatore riguardano l'esistenza di procedure formali che permettono al pubblico in generale, alle imprese e alle organizzazioni della società civile di influenzare la regolamentazione e le azioni governative, nonché se le opinioni dei cittadini su tali procedure di consultazione siano rese pubbliche.
- **Voter turnout:** la partecipazione elettorale è definita come il rapporto tra il numero di individui che hanno votato durante un'elezione e la popolazione registrata per votare. Poiché le caratteristiche istituzionali dei sistemi di voto variano notevolmente tra i paesi e tra i tipi di elezioni, l'indicatore si riferisce alle elezioni che hanno attratto il maggior numero di elettori in ciascun paese.

- **Health**

- **Life expectancy:** l'aspettativa di vita misura quanti anni in media le persone possono attendersi di vivere in base ai tassi di mortalità specifici per età. Questa misura si riferisce alle persone nate oggi e viene calcolata come una media ponderata dell'aspettativa di vita per gli uomini e le donne.
- **Self-reported health:** si riferisce alla percentuale della popolazione di età superiore a 15 anni che dichiara di avere uno stato di salute "buono". L'Organizzazione Mondiale della Sanità (OMS) raccomanda di utilizzare un sondaggio standard di intervista sulla salute per misurare questo indicatore, formulando la domanda come

“Come valuti la tua salute in generale?” con una scala di risposta che include “Molto buona/ Buona/ Sufficiente/ Cattiva/ Molto cattiva”.

- **Life Satisfaction:** misura il livello di soddisfazione e benessere generale delle persone in base alla loro percezione della loro situazione di vita rispetto ai loro migliori e peggiori scenari possibili.
- **Safety**
 - **Assault rate:** la variabile misura la percentuale di persone che ha subito un’aggressione o una rapina entro l’arco dei 12 mesi precedenti all’indagine.
 - **Homicide rate:** numero annuale di omicidi intenzionali registrati dalla polizia.
- **Work-life balance**
 - **Employees working very long hours:** proporzione dei dipendenti lavoratori il cui contratto di lavoro prevede 50 ore o più.
 - **Time devoted to leisure and personal care:** quantità di minuti (o ore) al giorno che, in media, le persone impiegate a tempo pieno dedicano al tempo libero e alle attività di cura personale.

2. Definizione degli obiettivi

La seguente analisi si pone degli obiettivi tramite l’utilizzo del database OECD, con particolare attenzione su tre aspetti principali.

- In primo luogo, nella parte precedente, è stata condotta un’analisi delle serie storiche basata su osservazioni annuali per comprendere l’evoluzione di specifiche variabili nel tempo. Questo ha consentito di acquisire una prospettiva longitudinale e identificare eventuali tendenze o cambiamenti significativi.
- In seguito, in questa parte, ci si concentra su un anno specifico (il 2017), in cui verrà condotta un’analisi esplorativa e di clustering. Quest’ultima permetterà di ottenere una visione più dettagliata dei pattern e delle relazioni presenti nei dati di quell’anno in particolare.
- Infine, verrà scelta una variabile aleatoria sulla quale svolgere inferenza.

L’obiettivo principale è capire quali elementi hanno un impatto significativo sul livello di benessere e qualità della vita.

3. DATA WRANGLING

Il “data wrangling,” spesso chiamato anche “data munging” rappresenta un passaggio cruciale nell’analisi dei dati. Molto spesso si hanno a disposizione un insieme di dati grezzi provenienti da diverse fonti. Questi dati possono essere disorganizzati, contenere errori, dati mancanti o informazioni in formati diversi. Senza una preparazione adeguata, l’analisi e la modellazione dei dati sarebbero difficili, se non impossibili.

Il data wrangling è il processo di trasformazione dei dati disordinati e sporchi in un formato coerente e adatto all’analisi. Il passaggio comporta molteplici attività, tra cui la pulizia dei

dati per correggere errori e rimuovere duplicati, la standardizzazione delle unità di misura, la trasformazione dei dati categorici in forme numeriche comprensibili, e la creazione di nuove variabili o caratteristiche quando necessario.

L'obiettivo finale del data wrangling è creare un dataset pulito, coerente e pronto per essere analizzato, riducendo così i potenziali errori e garantendo che i risultati dell'analisi siano accurati e significativi.

3.1. Raccolta dei dati e importazione del dataset

Il primo passo è quello di caricare in memoria il dataset per l'analisi. Poiché è un dataset di tipo excel è necessario installare la libreria apposita contenente funzioni necessarie al caricamento dei dati.

Import del dataset

Viene quindi caricato il set di dati in memoria. Il dataset ,contenuto nel file csv/excel etc , viene letto in un dataframe. Il dataframe è la struttura dati utilizzata da R per memorizzare una matrice di dati.

Il dataset viene letto in una struttura dati detta *tibble*, che è una forma moderna di dataframe, in cui vengono mantenute le caratteristiche base del dataframe.

La classe del dataframe è appunto tibble.

```
[1] "tbl_df"      "tbl"        "data.frame"
```

Panoramica dei dati

Si utilizza il comando `head()` per visualizzare le prime unità statistiche.

```
# A tibble: 6 x 25
  Stato                Dwellings without ba~1 'Housing expenditure' 'Rooms per person'
  <chr>                <dbl>                <dbl>                <dbl>
1 Australia            1.1                20                2.3
2 Austria              1                21                1.6
3 Belgium              2.3                21                2.2
4 Canada               0.2                22                2.5
5 Chile                9.4                18                1.9
6 Czech Republic       0.6                24                1.4
# i abbreviated name: 1: 'Dwellings without basic facilities'
# i 21 more variables: 'Household net adjusted disposable income' <dbl>,
# 'Household net financial wealth' <dbl>, 'Labour market insecurity' <dbl>,
# 'Employment rate' <dbl>, 'Long-term unemployment rate' <dbl>,
# 'Personal earnings' <dbl>, 'Quality of support network' <dbl>,
# 'Educational attainment' <dbl>, 'Student skills' <dbl>,
# 'Years in education' <dbl>, 'Air pollution' <dbl>, ...
```

Invece, si utilizza il comando `tail()` per visualizzare le ultime osservazioni del dataset.

```
# A tibble: 6 x 25
  Stato          Dwellings without ba~1 'Housing expenditure' 'Rooms per person'
  <chr>                <dbl>                <dbl>                <dbl>
1 Türkiye          6.5                20                1
2 United Kingdom   0.4                24                2
3 United States    0.1                18               2.4
4 Brazil           6.7                20               0.8
5 Russia          13.8                19                1
6 South Africa     37                18               0.7
# i abbreviated name: 1: 'Dwellings without basic facilities'
# i 21 more variables: 'Household net adjusted disposable income' <dbl>,
# 'Household net financial wealth' <dbl>, 'Labour market insecurity' <dbl>,
# 'Employment rate' <dbl>, 'Long-term unemployment rate' <dbl>,
# 'Personal earnings' <dbl>, 'Quality of support network' <dbl>,
# 'Educational attainment' <dbl>, 'Student skills' <dbl>,
# 'Years in education' <dbl>, 'Air pollution' <dbl>, ...
```

Sono quindi presenti 38 osservazioni relative agli stati dell'indagine e 25 variabili di descrizione.

[1] 38 25

3.2. Pulizia dei dati (data cleaning)

I dati grezzi sono spesso disordinati e formattati male. Inoltre, potrebbero mancare definizioni appropriate che tengano conto della scala di misurazione utilizzata.

Per cui la pulizia dei dati consiste nel procedimento mediante il quale si esaminano e si migliorano i dati contenuti nel dataset, con l'obiettivo di assicurare che siano di alta qualità e validi per l'analisi statistica. Le procedure che caratterizzano questo passaggio sono le seguenti:

- **Analisi dei missing values.** Si verifica la presenza/assenza dei valori mancanti (NA) che può notevolmente influire sull'analisi. Si procede, quindi, decidendo di eliminarli (*na.omit*) sostituendoli con valori reali (esempio: imputazione di media o mediana).
- **Individuazione di valori anomali.** Si esamina attentamente il dataset per individuare eventuali valori inesatti o insoliti, e si prendono decisioni su come trattarli. Queste decisioni possono includere l'eliminazione dei valori problematici o la loro sostituzione con valori appropriati.
- **Trasformazione delle variabili.** Si verifica che tutte le variabili abbiano la classe appropriata. (Double o Integer per i numeri, factor per le variabili categoriali, ordered per le variabili categoriali ordinate).
- **Implementazione di nuove variabili.** Sulla base delle variabili già presenti nel dataset si possono creare delle nuove variabili, ad esempio facendo operazioni matematiche tra due variabili o creando una variabile multilivello sulla base di una variabile numerica.

Analisi dei missing values

Si verifica la presenza degli NA nel dataset. Nel dataset non sono presenti NA per l'anno corrente.

[1] 0

Individuazione di valori anomali

Per l'individuazione di valori anomali o errati è possibile utilizzare la funzione `summary()`. La funzione fornisce un resoconto automatico delle statistiche di base per ciascuna variabile nel dataset. Queste statistiche comprendono il valore minimo, il valore massimo, la media, la mediana e la deviazione standard. Se la variabile è numerica, vengono calcolati anche i quartili e il range interquartile. Nel caso in cui la variabile sia di tipo carattere o un factor, la funzione restituirà il conteggio delle osservazioni per ciascun livello o valore univoco presente nella variabile stessa.

Dwellings without basic facilities	Housing expenditure	Rooms per person	
Min. : 0.000	Min. :15.00	Min. :0.700	
1st Qu.: 0.125	1st Qu.:20.00	1st Qu.:1.200	
Median : 0.600	Median :21.00	Median :1.750	
Mean : 3.450	Mean :20.89	Mean :1.642	
3rd Qu.: 4.275	3rd Qu.:23.00	3rd Qu.:1.900	
Max. :37.000	Max. :26.00	Max. :2.500	
Household net adjusted disposable income	Household net financial wealth		
Min. :10872	Min. : 2260		
1st Qu.:18725	1st Qu.: 18600		
Median :24903	Median : 39468		
Mean :25114	Mean : 49363		
3rd Qu.:30332	3rd Qu.: 71257		
Max. :44049	Max. :176076		
Labour market insecurity	Employment rate	Long-term unemployment rate	
Min. : 1.500	Min. :43.00	Min. : 0.030	
1st Qu.: 2.600	1st Qu.:65.00	1st Qu.: 1.260	
Median : 4.000	Median :69.00	Median : 1.980	
Mean : 5.476	Mean :67.74	Mean : 3.198	
3rd Qu.: 5.525	3rd Qu.:73.75	3rd Qu.: 3.917	
Max. :26.500	Max. :86.00	Max. :16.950	
Personal earnings	Quality of support network	Educational attainment	
Min. :11554	Min. :76.00	Min. :37.00	
1st Qu.:23924	1st Qu.:88.00	1st Qu.:75.50	
Median :38223	Median :90.50	Median :81.50	
Mean :37436	Mean :90.03	Mean :77.24	
3rd Qu.:49291	3rd Qu.:93.00	3rd Qu.:87.75	
Max. :62636	Max. :98.00	Max. :95.00	
Student skills	Years in education	Air pollution	Water quality
Min. :391.0	Min. :14.80	Min. : 3.00	Min. :54.00
1st Qu.:481.5	1st Qu.:16.43	1st Qu.: 9.25	1st Qu.:73.75

Median :496.0	Median :17.30	Median :14.00	Median :84.00
Mean :486.8	Mean :17.38	Mean :13.39	Mean :82.26
3rd Qu.:506.0	3rd Qu.:18.10	3rd Qu.:17.50	3rd Qu.:91.75
Max. :529.0	Max. :21.20	Max. :28.00	Max. :99.00
Stakeholder engagement for developing regulations			Voter turnout
Min. :0.80			Min. :49.00
1st Qu.:1.50			1st Qu.:62.25
Median :2.10			Median :69.50
Mean :2.05			Mean :70.03
3rd Qu.:2.60			3rd Qu.:77.75
Max. :3.50			Max. :91.00
Life expectancy		Self-reported health	Life satisfaction
Min. :57.40	Min. :33.00		Min. :4.800
1st Qu.:78.17	1st Qu.:62.00		1st Qu.:5.900
Median :81.15	Median :70.00		Median :6.650
Mean :79.54	Mean :67.45		Mean :6.529
3rd Qu.:82.25	3rd Qu.:76.00		3rd Qu.:7.275
Max. :83.90	Max. :88.00		Max. :7.500
Feeling safe walking alone at night			Homicide rate
Min. :36.10			Min. : 0.200
1st Qu.:60.98			1st Qu.: 0.600
Median :70.40			Median : 1.000
Mean :68.63			Mean : 2.934
3rd Qu.:79.88			3rd Qu.: 1.625
Max. :87.70			Max. :27.600
Employees working very long hours		Time devoted to leisure and personal care	
Min. : 0.160		Min. :12.59	
1st Qu.: 3.795		1st Qu.:14.47	
Median : 6.225		Median :14.89	
Mean : 8.723		Mean :14.83	
3rd Qu.:12.373		3rd Qu.:15.16	
Max. :33.770		Max. :16.36	

Nello studio dei dati, vengono quindi esaminati una serie di indicatori che forniscono una visione approfondita della qualità della vita e del benessere. Partendo dall'indicatore "Abitazioni senza servizi di base," si osserva che circa il 50% degli stati esprime un valore medio prossimo all'1, il che significa che in media l'un per cento delle famiglie potrebbero ancora affrontare sfide legate all'accesso a servizi essenziali come l'acqua corrente e i servizi igienici. Le "Spese per l'abitazione" costituiscono una parte significativa della spesa familiare, con una media di circa 20,87 dollari. Media e mediana coincidono esprimendo una possibilità di distribuzione pressochè simmetrica, la questione verrà approfondita successivamente. Il numero medio di "Stanze per persona" è di circa 1,65, ma con variazioni significative. In termini di reddito, il "Reddito familiare netto disponibile aggiustato" è di circa 25.254 dollari. Tuttavia, essendoci una presenza di stati con ricchezza territoriale e lavorative diverse, questo fa sì che ci si confronti con patrimoni diversi fino a un massimo di 44049. Lo stesso problema si riflette in maniera più evidente sul "Patrimonio finanziario netto delle famiglie" con valore medio di circa 50.419 dollari. Per questa variabile la differenza tra il terzo e quarto quartile è di oltre

100000 unità monetario dando evidenza di probabili outlier o valori anomali. Il “Tasso di occupazione” medio è del 67,72%, il che rappresenta la percentuale di persone in età lavorativa attualmente impiegate, il minimo è prossimo al 44% per il Sud Africa mentre il massimo all’88% per l’Islanda. L’“Aspettativa di vita” media di circa 79,55 anni suggerisce una buona prospettiva di salute e longevità. Il paese con aspettativa di vita più alta è il Giappone con 84 anni mentre il più basso è del Sud Africa con aspettativa di vita prossima ai 57 anni.

Questi indicatori forniscono una panoramica completa delle condizioni di vita tenendo conto, tuttavia, delle differenze che caratterizzano ciascun paese. L’analisi di questi dati costituisce un passo importante per identificare i fattori positivi e negativi al fine di aumentare il benessere generale. L’analisi dettagliata di ciascuna variabile si terrà in seguito con il relativo approfondimento su valori anomali.

3.3. Descrizione del dataset

Dalla descrizione delle variabili del dataset “better life index” mostrata in tabella, è possibile comprendere che, ad esempio: - al tipo *chr* (*carattere*), ovvero una stringa di testo, appartengono le variabili relative allo stato come unità statistica; - al tipo *numeric* (*numerico*), ovvero i numeri con la virgola mobile, fanno parte le variabili che esprimono una percentuale, come: tasso di occupazione, tasso di omicidio, tempo dedicato alla cura personale etc.; - al tipo *integer* (*numeri interi*), ovvero i numeri interi senza decimali, appartengono le variabili: qualità dell’aria e dell’acqua, spese per la casa, reddito etc.;

tibble [38 x 25] (S3: tbl_df/tbl/data.frame)

\$ Stato	:	chr	[1:38]	"Australia"	"Austria"	"Be
\$ Dwellings without basic facilities	:	num	[1:38]	1.1	1	2.3
\$ Housing expenditure	:	num	[1:38]	20	21	21
\$ Rooms per person	:	num	[1:38]	2.3	1.6	2.2
\$ Household net adjusted disposable income	:	num	[1:38]	33417	32544	29968
\$ Household net financial wealth	:	num	[1:38]	57462	59574	104084
\$ Labour market insecurity	:	num	[1:38]	4.3	2.7	4.8
\$ Employment rate	:	num	[1:38]	72	72	62
\$ Long-term unemployment rate	:	num	[1:38]	1.36	1.94	3.98
\$ Personal earnings	:	num	[1:38]	52063	48295	49587
\$ Quality of support network	:	num	[1:38]	94	92	92
\$ Educational attainment	:	num	[1:38]	80	85	75
\$ Student skills	:	num	[1:38]	502	492	503
\$ Years in education	:	num	[1:38]	21.2	17.1	18.2
\$ Air pollution	:	num	[1:38]	5	16	15
\$ Water quality	:	num	[1:38]	92	93	84
\$ Stakeholder engagement for developing regulations:	:	num	[1:38]	2.7	1.3	2.2
\$ Voter turnout	:	num	[1:38]	91	75	89
\$ Life expectancy	:	num	[1:38]	82.5	81.3	81.1
\$ Self-reported health	:	num	[1:38]	85	70	75
\$ Life satisfaction	:	num	[1:38]	7.3	7	6.9
\$ Feeling safe walking alone at night	:	num	[1:38]	63.6	80.7	70.7
\$ Homicide rate	:	num	[1:38]	1	0.4	1
\$ Employees working very long hours	:	num	[1:38]	13.2	6.78	4.31

```
$ Time devoted to leisure and personal care      : num [1:38] 14.3 14.6 15.8 14.4 14.9
```

4. EDA (Exploratory data analysis)

L'EDA (Analisi Esplorativa dei Dati) è un approccio all'analisi del set di dati per riassumere le loro caratteristiche principali, spesso con metodi di visualizzazione e grafici. Questo passaggio viene svolto preventivamente all'applicazione di modelli statistici poichè serve a capire cosa effettivamente i dati sono in grado di comunicare. Gli obiettivi principali dell'EDA sono:

- **Massimizzare l'intuizione in un set di dati.** L'EDA aiuta ottenere una comprensione approfondita dei dati. Consente di esaminare le caratteristiche principali, le distribuzioni delle variabili, le relazioni tra di esse e i possibili valori anomali.
- **Scoprire la struttura sottostante.** Durante l'EDA, vengono creati grafici e visualizzazioni che consentono di esplorare la distribuzione dei dati. Queste visualizzazioni possono includere istogrammi, grafici a dispersione (scatter plot), diagrammi a barre e etc.. La visualizzazione dei dati fornisce un'immagine visiva della struttura dei dati, evidenziando modelli o tendenze che potrebbero non emergere in modo evidente dai dati grezzi.
- **Effettuare confronti.** I grafici permettono di confrontare diverse categorie o gruppi di dati, ad esempio analizzando e confrontando le distribuzioni o le frequenze di diverse variabili.
- **Condividere i risultati.** l'utilizzo di grafici consente di esporre in modo efficace i risultati di un'analisi a un pubblico più vasto, comprese le persone che non hanno una conoscenza approfondita di statistica o analisi dei dati. Ad esempio, è possibile includere un grafico all'interno di un report o di una presentazione per presentare chiaramente e in modo sintetico i risultati dell'analisi.

In letteratura l'analisi dei dati distingue in maniera cruciale tre sezioni:

- **Analisi univariata** , i metodi appartenenti a questa sezione esaminano una variabile (colonna di dati alla volta);
- **Analisi bivariata** in cui l'esplorazione viene svolta su coppie di variabili (sia dello stesso tipo che di tipo diverso);
- **Analisi multivariata** in cui vengono coinvolti metodi in cui vengono esaminate tre o più variabili alla volta per esplorare le relazioni.

Nella strutturazione dell'analisi delle variabili del progetto, si terrà conto della distinzione appena presentata. Questi tre approcci consentono di esplorare le caratteristiche delle variabili da diverse prospettive, permettendo di avere una comprensione più completa dei dati e delle relazioni tra le variabili coinvolte nello studio statistico.

5. ANALISI UNIVARIATA

L'analisi univariata permette di esplorare una variabile alla volta. Le Statistiche descrittive sono strumenti cruciali per riassumere un gruppo di osservazioni nel modo più semplice possibile.

Per le variabili categoriali: 1. si calcolano le tabelle di frequenza 2. diagramma a barre 3. pie chart

Per le variabili quantitative: 1. si calcolano le misure di posizione o tendenza centrale come media, mediana e quartili 2. misure di dispersione come la deviazione standard, MAD, varianza etc.. 3. misure di forma come asimmetria e curtosi 4. boxplots 5. stime di densità Kernel 6. Normal probability plots

Proseguendo con l'analisi delle variabili qualitative:

5.1. Tabelle di frequenza (per le variabili categoriali)

Per esplorare le variabili categoriali le tabelle di frequenza sono uno degli strumenti più efficaci e semplici. Si utilizza il comando `table(datasetvariabile)` per ottenere il conteggio delle osservazioni per ogni categoria di una variabile. Nel caso in cui fossero presenti NA, la funzione è utile perchè questi ultimi vengono conteggiati e segnalati nella tabella di frequenza.

Poichè all'interno del dataset non sono presenti variabili categoriali, è possibile trasformare variabili numeriche in categoriali sotto un criterio specifico. In tal caso, si è deciso di distinguere, per il puro scopo esplorativo, due variabili in particolare: **Air pollution** e **Life Satisfaction**. Il criterio logico che giustifica la scelta è il seguente: per ciascuna variabile è possibile individuare un valore soglia (media o mediana della variabile numerica) che riesca a distinguere le unità statistiche (o i paesi) in due gruppi, gli stati che hanno valori al di sotto e al di sopra del valore soglia tale da costruire una separazione netta.

5.2. Air pollution (Variabile 1)

Tabella di frequenza

La variabile **AirPollution** rappresenta la contaminazione dell'aria. Questa vale 1 per i paesi il cui valore è maggiore della mediana mentre vale 0 altrimenti. E' una variabile categorica nominale (factor), in quanto i valori possibili sono limitati e non possono essere ordinati in base alla loro scala di valore.

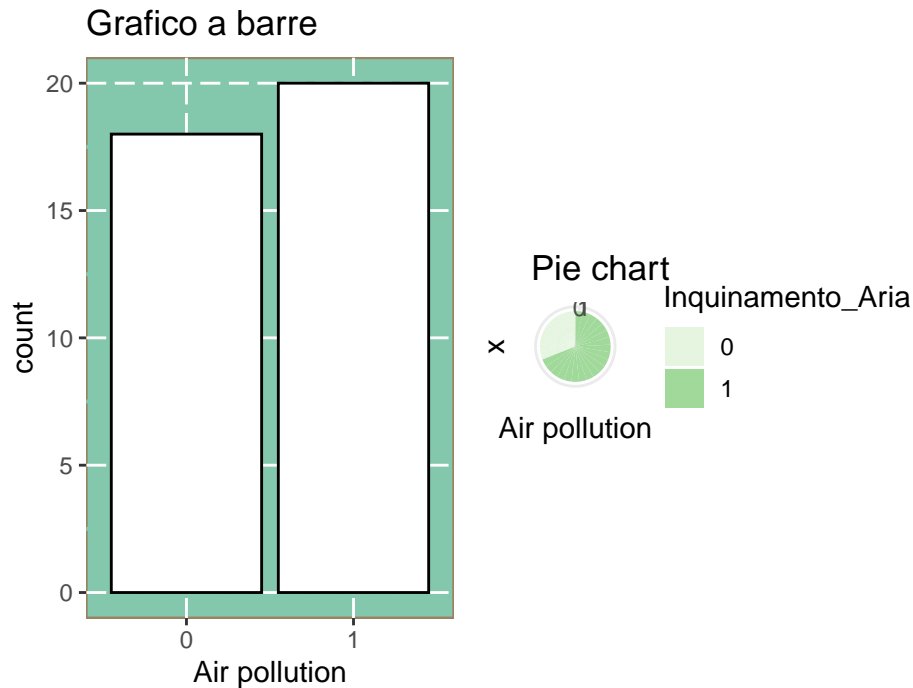
	Frequenza	Frequenza_percentuale
0	18	47.36842
1	20	52.63158

Sono presenti 18 osservazioni con valore di `airpollution` al di sotto della mediana e 21 con valori maggiori della mediana.

Per visualizzare la variabile **AirPollution** si può utilizzare un grafico a barre e un grafico a torta.

Grafico a barre e a torta

Nel grafico a barre ogni barra rappresenta il numero di osservazioni per il livello della variabile. Il grafico a torta è un tipo di grafico utilizzato per visualizzare le frequenze di una variabile di tipo carattere o factor. In particolare, ogni settore della torta rappresenta la percentuale di osservazioni per un determinato livello o valore univoco della variabile.



5.3. Life Satisfaction (Variabile 2)

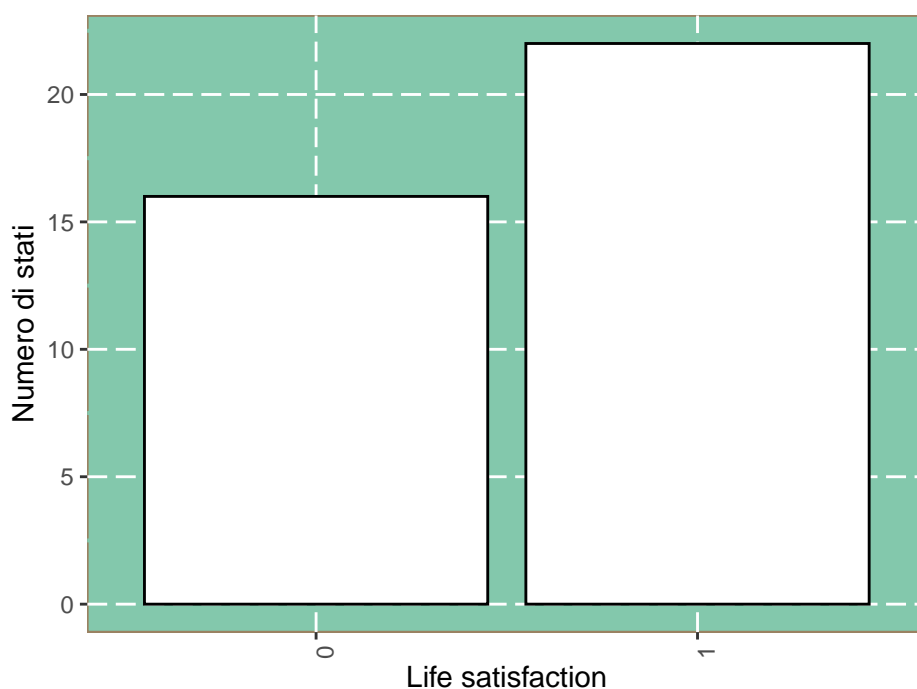
Tabella di frequenza

La variabile **life satisfaction** rappresenta la soddisfazione media degli individui riguardo la propria vita. La variabile factor è costruita a partire dalla variabile numerica, dopo aver individuato il valore medio, ciascuno stato in corrispondenza della feature assume valore 0 se si trova al di sotto della media di soddisfazione di vita 1 altrimenti. Dato il campione ristretto, il numero di osservazioni si riduce al 43% degli stati con soddisfazione di vita al di sotto della media mentre il restante 57% al di sopra.

	Frequenza	Frequenza_percentuale
0	16	42.10526
1	22	57.89474

Grafico a barre

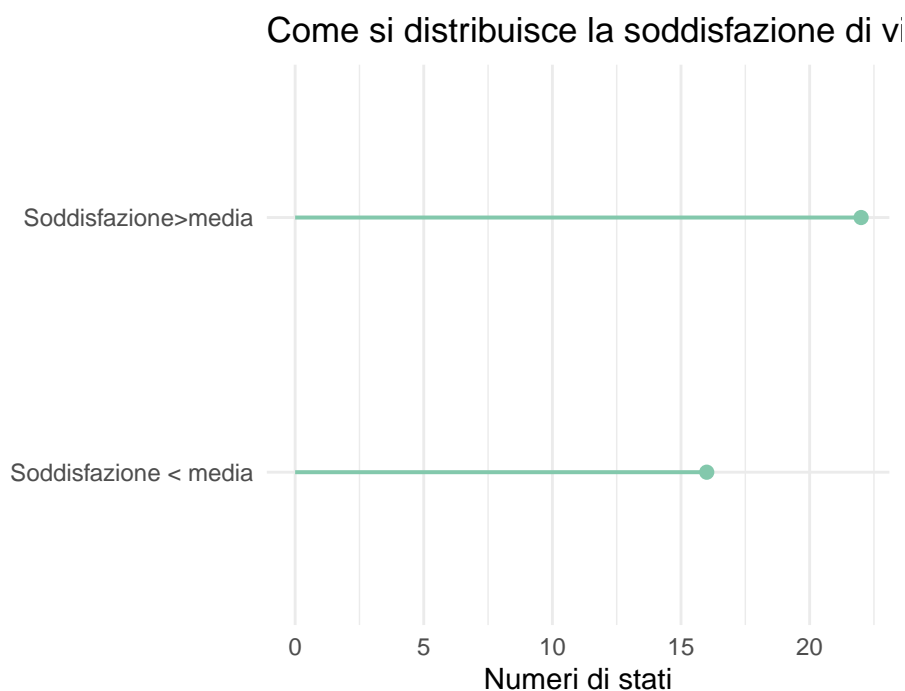
Per la rappresentazione grafica, essendo una variabile factor, è possibile rappresentare la distribuzione tramite il grafico a barre. La disparità tra le due classi è più evidente della variabile precedente. L'asse delle y rappresenta il numero di stati mentre l'asse delle x è espressione delle due entità della feature.



Lollipop Chart

Un metodo alternativo per la visualizzazione di una variabile categoriale è il lollipop chart. Un *Lollipop Chart* è un grafico che presenta punti di dati come cerchi o dischi (“lollipops”) posti su un asse orizzontale, che rappresenta una variabile indipendente o categoria. Ogni lollipop rappresenta un singolo punto dati e la sua posizione sull’asse orizzontale indica il valore di quella variabile. Il “lollipop” è collegato a una linea verticale o “astina” che si estende verso sinistra da ciascun punto dato fino a una seconda scala. La seconda scala rappresenta la variabile dipendente o un valore di riferimento, come una media o una soglia.

Per cui, 16 stati hanno valore medio della soddisfazione dell’individuo al di sotto della media complessiva, mentre circa 22 stati hanno valore medio della soddisfazione al di sopra della media.

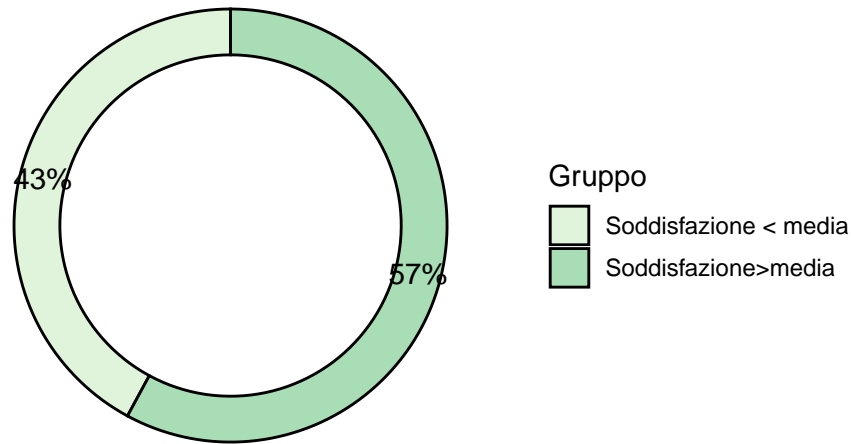


Donut chart

Il “Donut Chart” è un grafico circolare diviso in sezioni, o “fette,” che rappresentano le diverse categorie di dati. Ogni fetta corrisponde a una categoria specifica e la dimensione di ciascuna fetta è proporzionale alla percentuale che quella categoria rappresenta rispetto al totale. E’ un’alternativa al “Pie Chart” distinguendosi per il foro al centro, che crea un anello vuoto. Il “Donut Chart” è utile per visualizzare chiaramente come le diverse categorie contribuiscono a un totale, e la dimensione dell’anello interno rappresenta il totale complessivo.

Ancora una volta la lettura delle informazioni è in percentuale, il 57% degli stati ha valore di soddisfazione al di sopra della media mentre il 43% al di sotto.

Soddisfazione di vita



5.4. Indici di sintesi

Gli indici di sintesi, detti anche statistiche, sono utili a descrivere i dati numerici. Si fa riferimento a media, mediana, moda, varianza, deviazione standard. La media, moda e mediana sono misure di centralità, mentre la varianza e deviazione standard misurano la dispersione dei dati.

Media campionaria

La **media campionaria** è la media aritmetica dei valori in corrispondenza di ciascuna osservazione. Si consideri ad esempio la variabile della ricchezza finanziaria netta, **Household net financial wealth**. Il valore della media campionaria è:

[1] 49362.79

Nel contesto della ricchezza finanziaria, una media di 49362.79 dollari indica che, se si considerano tutte le famiglie nella popolazione, la quantità media di ricchezza finanziaria posseduta è di circa 49362.79 dollari.

Definizione.. Dati i livelli osservati x_1, x_2, \dots, x_n per una variabile X , si definisce la media come:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

La media campionaria è importantissima poichè utilizza tutti i dati ed è influenzata in maniera sensibile dai valori eccezionalmente alti o bassi. L'importanza della media è dovuta, tuttavia,

alla proprietà che soddisfa: la *somma degli scarti è nulla*, per cui la media costituisce il baricentro di una distribuzione di frequenza.

Definiti gli *scarti della media* come $(x_i - \bar{x})$, $\forall i = 1, 2, \dots, n$, si ottiene:

$$\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n (x_i) - \sum_{i=1}^n (\bar{x}) = \sum_{i=1}^n x_i - n\bar{x} = \sum_{i=1}^n (x_i) - n(\sum_{i=1}^n (x_i)/n) = \sum_{i=1}^n (x_i) - \sum_{i=1}^n (x_i) = 0$$

per cui, implica che la media \bar{x} , essendo interna ai valori delle modalità di X , presenterà alcuni scarti positivi, eventualmente alcuni nulli ed altri negativi. La media \bar{x} è il valore che bilancia esattamente positivi e negativi in quanto baricentro di una distribuzione di masse.

Mediana campionaria

La **mediana campionaria** è la modalità dell'unità statistica che occupa il posto centrale nella distribuzione ordinata delle osservazioni. La media della ricchezza finanziaria è:

[1] 39468

La mediana è significativamente diversa dalla media, per cui potrebbe indicare la presenza di valori estremi che influenzano la media.

Definizione.. Dati i livelli osservati x_1, x_2, \dots, x_n per una variabile X , si definisce la mediana, indicata con $\text{Med}(X)$, il quantile a livello $\alpha = 50$. La mediana campionaria dipende da solo uno o due valori centrali dei dati e non risente dei valori estremi. La mediana è una posizione sia di posizione, in quanto individua un punto nel range dei dati, sia di centralità perchè individua un punto che divide il range in parti contenenti uguale massa di dati.

Moda campionaria

La **moda** è la modalità a cui corrisponde la massima frequenza, assoluta o relativa.

Definizione.. Dati i livelli osservati x_1, x_2, \dots, x_n per una variabile X , si definisce la moda, indicata con $\text{Mod}(X)$:

- il livello osservato più frequentemente nei dati *non continui*;
- il livello di massima densità nei dati *continui*.

A differenza di *media* e *mediana* può essere calcolata per qualsiasi tipo di variabile. Non esiste in R una funzione per estrarre la moda o la classe modale di una distribuzione di dati poiché è facilmente ricavabile osservando il grafico delle frequenze assolute o l'istogramma delle frequenze relative, ma è possibile dedurlo manualmente. Vale 10000 per la variabile in analisi.

[1] 10000

Quartili, decili, percentili

Anche i **quantili** possono essere visti come misure di posizione. Tuttavia, seguono un *principio di localizzazione* diverso.

Definizione.. Per serie di quantili intendiamo i quantili calcolati su una griglia di valori $\alpha_1, \alpha_2, \dots, \alpha_n$ equispaziata.

I Quantili si ottengono dividendo l'insieme di dati ordinati in 4 parti uguali. Nel caso studio, questi sono i quantili della ricchezza patrimoniale:

- **Primo Quantile (25%)**: Il primo quantile rappresenta il 25% inferiore della distribuzione dei dati sulla ricchezza patrimoniale. Questo valore indica che il 25% delle famiglie degli stati presi in considerazione ha una ricchezza patrimoniale inferiore a 19082,5 dollari.
- **Mediana (50%)**: la mediana della ricchezza patrimoniale è di 43493,0 dollari. Ciò significa che il 50% della popolazione ha una ricchezza patrimoniale inferiore a questo valore e il restante 50% ha una ricchezza patrimoniale superiore a questo valore.
- **Terzo Quantile (75%)**: il 25% delle famiglie degli stati ha una ricchezza patrimoniale superiore a 73842,0 dollari.

25%	50%	75%
18599.75	39468.00	71256.75

I Decili , invece, dividono la distribuzione in 10 parti uguali.

0%	10%	20%	30%	40%	50%	60%	70%
2260.0	9722.8	17067.2	20453.2	27229.2	39468.0	57865.4	63797.6
80%	90%	100%					
79699.4	92774.1	176076.0					

I Percentili dividono la distribuzione in 100 parti uguali.

0%	1%	2%	3%	4%	5%	6%	7%
2260.00	3062.53	3865.06	4464.31	4583.08	4701.85	5267.44	6137.68
8%	9%	10%	11%	12%	13%	14%	15%
7007.92	8337.52	9722.80	11136.57	12672.44	14208.31	15351.60	16080.50
16%	17%	18%	19%	20%	21%	22%	23%
16809.40	16988.75	17016.50	17043.89	17067.20	17090.51	17246.68	17621.12
24%	25%	26%	27%	28%	29%	30%	31%
17995.56	18599.75	19314.22	20028.69	20155.64	20266.27	20453.20	20846.14
32%	33%	34%	35%	36%	37%	38%	39%
21239.08	21803.80	22499.40	23195.00	23599.08	23957.61	24480.84	25855.02
40%	41%	42%	43%	44%	45%	46%	47%
27229.20	28635.85	30080.70	31525.55	32330.04	32928.70	33533.96	34254.72
48%	49%	50%	51%	52%	53%	54%	55%
34975.48	36489.50	39468.00	42446.50	45707.00	49120.25	52533.50	54342.00
56%	57%	58%	59%	60%	61%	62%	63%

56058.80	57367.36	57405.84	57444.32	57865.40	58611.69	59357.98	59508.45
64%	65%	66%	67%	68%	69%	70%	71%
59543.60	59685.55	60511.02	61336.49	62159.24	62978.42	63797.60	64121.33
72%	73%	74%	75%	76%	77%	78%	79%
64261.56	64489.45	67873.10	71256.75	73614.76	73836.02	74057.28	76271.72
80%	81%	82%	83%	84%	85%	86%	87%
79699.40	83127.08	84205.02	85075.63	86097.52	87667.80	89238.08	90136.14
88%	89%	90%	91%	92%	93%	94%	95%
90397.36	90658.58	92774.10	95322.29	97854.56	100255.49	102656.42	107733.65
96%	97%	98%	99%	100%			
116736.12	125738.59	140806.86	158441.43	176076.00			

Varianza e deviazione standard

L'indice più importante per misurare la variabilità di una distribuzione espresso dalla media degli scarti al quadrato è la **varianza**, dove per variabilità si intende la dispersione rispetto al centro della distribuzione. La **varianza campionaria** è così definita.

Definizione.. Sia x_1, x_2, \dots, x_n un campione osservato con media \bar{x} . Si definisce Varianza campionaria la quantità:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

con

$$(n = 2, 3, \dots)$$

- s^2 sarà tanto più grande se la dispersione rispetto a \bar{x} è grande e sarà tanto più piccola se la dispersione rispetto a \bar{x} è piccola.
- s^2 è tanto più piccola quanto più \bar{x} è rappresentativa/centrale per la distribuzione.

Per la ricchezza patrimoniale la varianza è pari a:

[1] 1498325115

Sulla quale è possibile svolgere delle considerazioni:

- 1. **Quantità di Variabilità:** una varianza di 1502434845 implica che i dati nella variabile sono piuttosto dispersi o variabili per cui ci sono differenze significative tra i singoli valori osservati nella variabile.
- 2. **Scostamenti Quadrati:** la varianza viene calcolata utilizzando gli scostamenti quadrati tra ciascun dato e la media dei dati. Questo significa che le differenze positive e negative sono elevate al quadrato, il che può amplificare l'effetto delle deviazioni dai valori medi.
- 3. **Unità di Misura Quadrata:** La varianza ha unità di misura quadrata. Ad esempio, la variabile rappresenta la ricchezza patrimoniale e ha dollari come moneta, la varianza avrà come indicatore dollari al quadrato, il che potrebbe non essere

facilmente interpretabile in termini reali. Per ottenere una misura della dispersione nella stessa unità di misura della variabile, si utilizza la radice quadrata della varianza, chiamata deviazione standard.

Definizione.. Sia x_1, x_2, \dots, x_n un campione osservato con media \bar{x} e varianza s^2 . Si definisce Deviazione standard campionaria la radice quadrata della varianza:

$$s = \sqrt{(s^2)}$$

[1] 38708.2

I dati nella variabile sono distribuiti in modo tale che, in media, ciascun dato si discosta dalla media della variabile in questione di circa 38708,25 dollari.

Osservazioni sulla varianza.

- sia la varianza che la deviazione standard sono misure non negative;
- var e sd sono uguali a 0 **se e solo se** tutte le x_i sono uguali alla media \bar{x} ;
- la varianza non ha un massimo quindi non è possibile fare valutazioni assolute;
- è possibile confrontare varianze di una stessa variabile misurata su campioni diversi, posto che siano espresse nella stessa unità di misura.

Coefficiente di variazione

Poichè varianza e scarto quadratico medio sono indici assoluti è bene parlare di indici relativi.

Definizione.. Si definisce Coefficiente di variazione come rapporto della varianza campionaria e il valore assoluto della media campionaria. E' un coefficiente privo di unità di misura, ovvero un numero puro, che esprime la variazione media del fenomeno in rapporto alla sua media utile per confrontare la variabilità di un fenomeno in circostanze differenti.

$$CV = \frac{s}{|\bar{x}|}$$

Per la variabile di benessere economico il coefficiente di variazione vale :

[1] 0.7841576

Il risultato indica che la deviazione standard è pari al 76.88% della media dei dati, suggerendo che la variabilità relativa dei dati è abbastanza elevata, poiché la deviazione standard è significativamente più grande della media. Il risultato è una possibile indicazione di distribuzione dei dati con valori estremi o una variazione significativa tra le osservazioni. (Basti osservare la differenza tra terzo quartile e valore massimo per questa osservazione, si distanziano di circa 100 mila dollari).

Coefficiente di simmetria

Nella descrizione di una distribuzione risulta fondamentale la definizione di **asimmetria**. Se la media supera la mediana, significa che la distribuzione si *attarda* verso valori alti e quindi superiori alla mediana, per cui si è in presenza di una distribuzione che presenta una coda più

pesante verso il semiasse positivo delle x , in tal caso si parla di **asimmetria positiva**. Per contro, se la media è inferiore alla mediana, si è in presenza di una distribuzione che presenta una coda più pesante verso sinistra, e quindi, si parla di **asimmetria negativa**.

$$\gamma = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1}$$

- Per $\gamma = 0$ la distribuzione di frequenze è simmetrica;
- Per $\gamma \geq 0$ la distribuzione di frequenze è asimmetrica positiva (la coda di destra è più lunga)
- Per $\gamma < 0$ la distribuzione di frequenze è asimmetrica negativa (la coda di sinistra è più lunga)

L'indice di simmetria per la variabile del benessere economico è positivo per cui è espressione di asimmetria positiva nella distribuzione, esprimendo code più lunghe verso la fine della distribuzione.

25%
0.2073893

Coefficiente di curtosi

La curtosi è una proprietà relativa alla velocità con cui la densità dei dati diminuisce man a mano che ci allontaniamo dal centro andando verso i valori estremi. Si definisce **curtosi campionaria** il valore:

$$\gamma_2 = \beta_2 - 3$$

dove $\beta_2 = \text{frac}m_4m_2/3$ è l'indice di pearson adimensionale. Gli indici β_2 e γ_2 permettono di confrontare la distribuzione di frequenze dei dati con una densità di probabilità normale standard, caratterizzata da $\beta_2 = 3$ e indice di curtosi $\gamma_2 = 0$. Se risulta:

- $\beta_2 < 3$ ($\gamma_2 < 0$): la distribuzione di definisce **platicurtica**, ossia la distribuzione è più piatta di una normale (code meno pesanti della n.);
- $\beta_2 > 3$ ($\gamma_2 > 0$): la distribuzione di definisce **leptocurtica**, ossia la distribuzione è più a punta di una normale (code più pesanti della n.);
- $\beta_2 = 3$ ($\gamma_2 = 0$): la distribuzione di definisce **normocurtica**, ossia la distribuzione piatta come una normale (code spesse come quelle della n.);

Il calcolo della curtosi campionaria è significativo soprattutto quando si analizzano distribuzioni di frequenze unimodali, in quanto questo indice viene confrontato con la curtosi di una distribuzione normale standard.

Per la variabile relativa al benessere finanziario l'indice di curtosi è positivo espressione di distribuzione platicurtica.

Funzione di distribuzione empirica

L'ECDF, che sta per “**Empirical Cumulative Distribution Function**” (“Funzione di Distribuzione Cumulativa Empirica”), è una funzione statistica utilizzata per rappresentare la distribuzione cumulativa empirica di un insieme di dati, ovvero mostra quanto siano distribuiti i dati rispetto alla loro frequenza cumulativa.

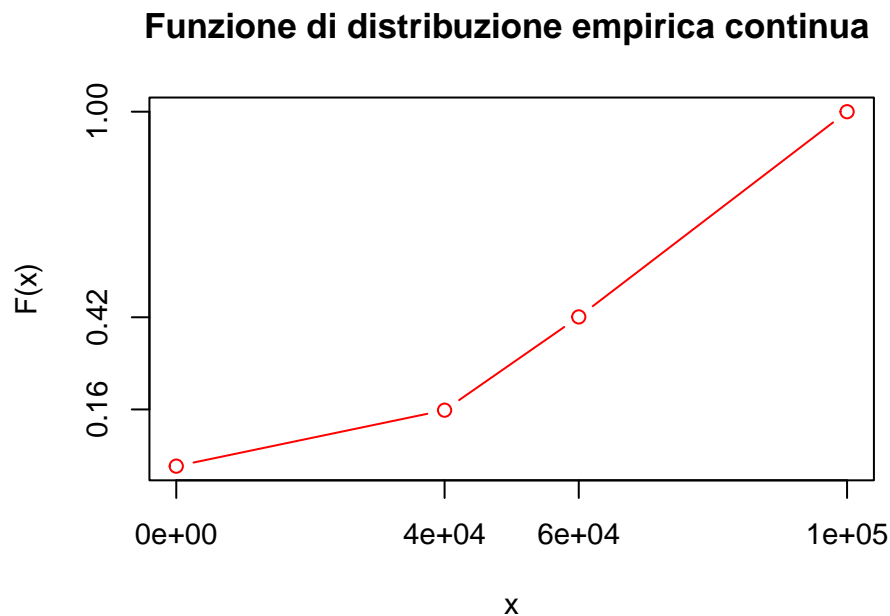
Definizione.. Si definisce Funzione di distribuzione empirica nel punto t , data una variabile X quantitativa e x_1, x_2, \dots, x_n un campione osservato:

$$F(t) = (\text{proporzione delle osservazioni } \leq t) = \frac{1}{n} \sum_{i=1}^n (1(x_i \leq t))$$

Funziona in questo modo:

1. **Creazione dell'ECDF:** Per costruire l'ECDF, si ordinano i dati in ordine crescente e si assegna a ciascun dato una probabilità cumulativa basata sulla sua posizione nell'ordine. Semplicemente si calcola la percentuale di dati che sono inferiori o uguali a ciascun valore nei dati.
2. **Rappresentazione grafica:** L'ECDF viene solitamente rappresentata graficamente attraverso un diagramma a scala di probabilità. Sull'asse delle ascisse (orizzontale) sono posizionati i valori dei dati, mentre sull'asse delle ordinate (verticale) si trova la probabilità cumulativa corrispondente. L'ECDF inizia a zero e raggiunge uno quando tutti i dati sono stati considerati.
3. **Interpretazione:** Guardando il plot, è possibile determinare facilmente la percentuale di dati al di sotto o al di sopra di un valore specifico. Ad esempio, se l'ECDF raggiunge il valore 0,75 a un certo punto, significa che il 75% dei dati è inferiore o uguale a quel valore.

Si ottiene in questo modo il grafico della funzione di distribuzione empirica continua introducendo delle classi fittizie: $c1=[0;40000]$, $c2=[40000;60000]$, $c3=[60000, 100000]$ e $c4=[100000, 176000]$ per la variabile di ricchezza finanziaria.



5.5. Servizi sanitari (variabile 1)

Statistiche

Nell'esaminare la situazione abitativa è necessario prendere in considerazione le condizioni di vita, come il numero medio di vani a persona e la presenza di dotazioni di base.

In media, si è osservato che la percentuale di famiglie che non possiedono servizi sanitari primari a disposizione è circa 3,42. La maggior parte delle famiglie, invece, ha accesso a un ai servizi sanitari primari. La mediana è di 0,6%, per cui la metà degli stati ha in percentuale 0,6% di famiglie senza servizi igienici e l'altra metà ha un numero in percentuale superiore. Si approfondisce quali sono gli stati. Il risultato suggerisce una distribuzione leggermente asimmetrica.

```
# A tibble: 18 x 1
```

```
  Stato
  <chr>
1 Australia
2 Austria
3 Belgium
4 Chile
5 Estonia
6 Hungary
7 Israel
8 Japan
9 Korea
```

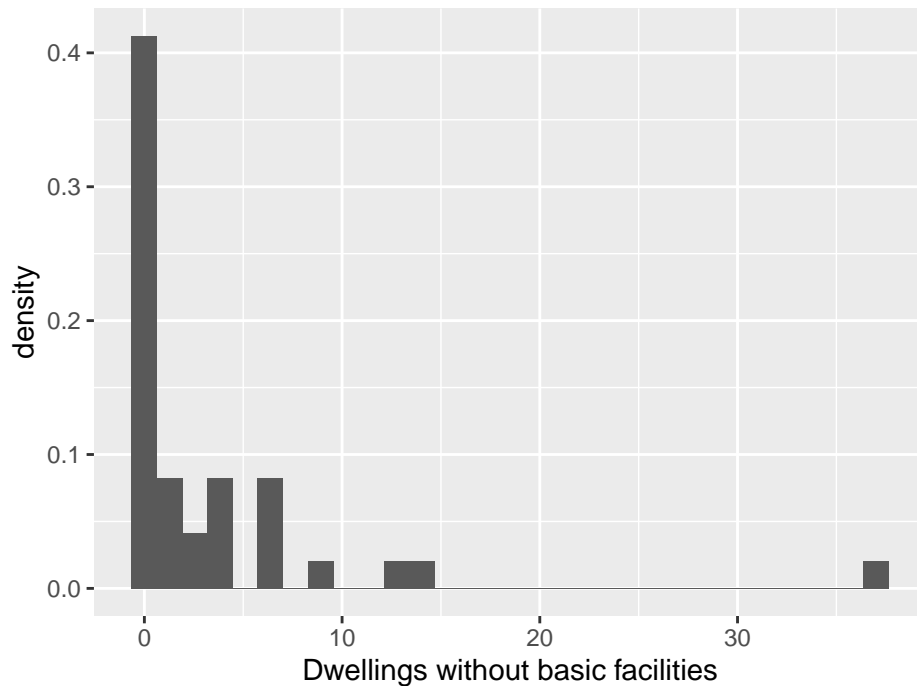
- 10 Latvia
- 11 Mexico
- 12 Poland
- 13 Portugal
- 14 Slovak Republic
- 15 Türkiye
- 16 Brazil
- 17 Russia
- 18 South Africa

La percentuale di 2,5% di famiglie senza sanitari è il numero riportato con maggiore frequenza tra gli stati. Rispetto ai quartili, il primo quartile (Q1) è di 0,15%, il secondo quartile (Q2, che coincide con la mediana) è di 0,6 % e il terzo quartile (Q3) è di 4.25 %. E' presente una chiara variazione significativa tra le abitazioni dei diversi stati. La varianza è di circa 42.96, suggerendo una notevole variabilità. La deviazione standard è di circa 6.55% esprimendo quanto i dati si discostino in media dalla media. La simmetria è positiva, con un valore di circa 0.78, per cui la distribuzione tende ad essere leggermente spostata verso destra, indicando una maggiore concentrazione di stati con un numero inferiore di famiglie con bagni rispetto a stati con famiglie con un numero superiore. La curtosi, con un valore di circa 15.00, indica una distribuzione con code pesanti. Ciò suggerisce che potrebbero essere presenti stati con un numero di percentuali estremamente alto o basso che contribuiscono a una maggiore curtosi.

	Statistiche	Valori
1	Media	3.4500000
2	Mediana	0.6000000
3	Moda	2.5000000
4	Q1	0.1250000
5	Q2	0.6000000
6	Q3	4.2750000
7	Varianza	44.0744595
8	Deviazione standard	6.6388598
9	Simmetria	0.7710843
10	Curtosi	14.4809954

Istogramma

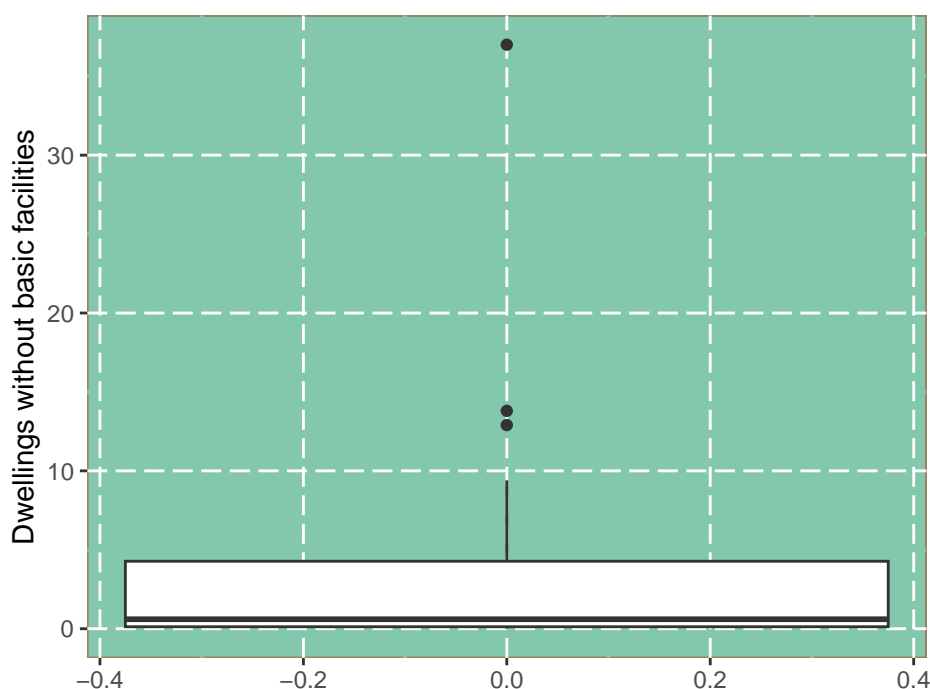
L'istogramma è una rappresentazione grafica per la visualizzazione della variabile numerica. Dal grafico non c'è evidenza di bimodalità ma forte presenza di asimmetria positiva. In particolar modo la maggior parte delle osservazioni si concentrano tra 0 e 0.6, come evidenziato dalla mediana.



Boxplot

Il boxplot è composto da una scatola che si estende da un quartile all'altro, cioè dal primo quartile (Q1) al terzo quartile (Q3). La lunghezza della scatola rappresenta l'intervallo interquartile (IQR) e indica dove si concentra la maggior parte dei dati. All'interno della scatola c'è una linea che rappresenta la mediana dei dati, che è il valore centrale nella distribuzione. I baffi si estendono dalla scatola verso l'alto e verso il basso. Possono rappresentare la dispersione dei dati e indicare quanto lontano si estendono i dati al di fuori dell'intervallo interquartile. I punti che cadono al di fuori dei baffi sono spesso considerati valori anomali o estremi. Questi punti rappresentano dati che si discostano significativamente dalla maggior parte delle osservazioni.

La mediana è posizionata molto in basso rispetto osservazioni. Data la lunghezza della scatola è evidente che vi è concentrazione dei dati a sinistra della mediana e dispersione sul lato sinistro. Vista la lunghezza del baffo si percepisce asimmetria positiva. In particolare, sono stati individuati circa tre valori anomali, rappresentati dagli stati di Lettonia e Russia, con una frequenza approssimativa intorno al 13-14%. Inoltre, il Sudafrica si distingue con una percentuale massima, pari al 37% di popolazione che non dispone di servizi igienici adeguati. Si parla di valori anomali rispetto alla popolazione di riferimento ma conoscendo il background degli stati le percentuali non sono al di fuori della norma. In genere, i valori anomali potrebbero essere riconosciuti nei casi in cui si aggregano individui o entità pressochè simili.



5.6. Spese per la casa (variabile 2)

Statistiche

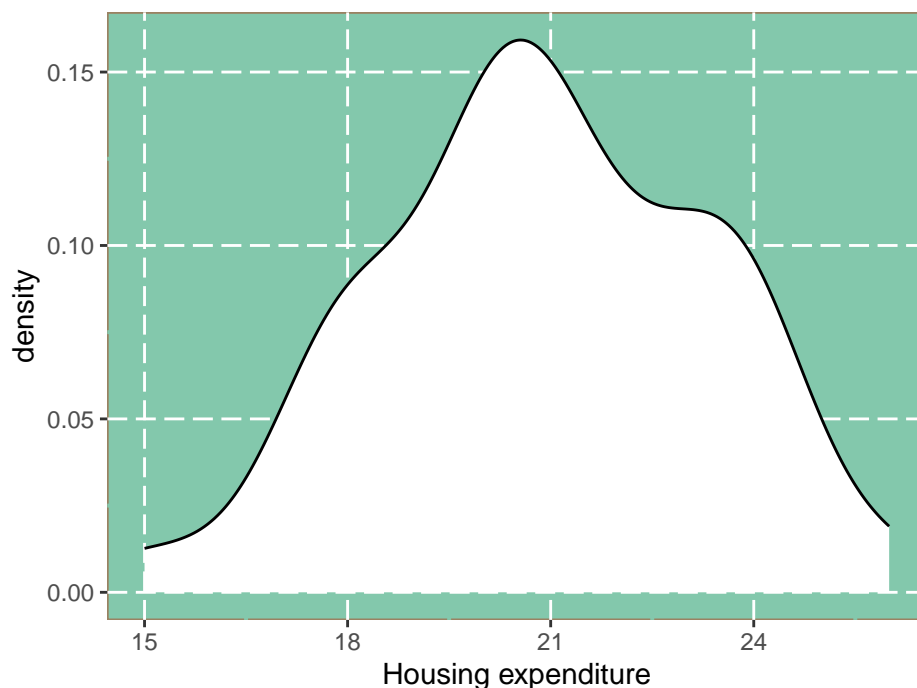
La spesa abitativa riveste un ruolo centrale nel bilancio familiare e rappresenta la principale voce di spesa per molti individui e famiglie se si prendono in considerazione affitto, gas, elettricità, acqua, mobilio e riparazioni. In media, le famiglie degli stati spendono circa il 20% del reddito disponibile lordo corretto per il mantenimento delle loro case. L'incidenza della spesa abitativa sul bilancio familiare varia da oltre il 26% nella Nuova Zelanda a meno del 15% in Corea. La mediana è di 21,00%, suggerisce una distribuzione relativamente bilanciata. La moda è anch'essa di 21,00%. Rispetto ai quartili, il primo quartile (Q1) è di 20,00%, il secondo quartile (Q2, che coincide con la mediana) è di 21,00% e il terzo quartile (Q3) è di 23,00 %. Questi valori quartili indicano che la maggior parte delle famiglie ha spese contenute, mentre alcune famiglie spendono di più (Q3) o di meno (Q1). In media le spese si discostano di 2,41% dalla media. La simmetria è positiva, con un valore di 0,33 con distribuzione leggermente spostata verso destra. La curtosi è negativa, con un valore di -0,52. Questo indica una distribuzione leggermente appiattita rispetto a una distribuzione normale, suggerendo che ci siano meno valori estremi o "code" rispetto a una distribuzione più appuntita.

	Statistiche	Valori
1	Media	20.8947368
2	Mediana	21.0000000
3	Moda	21.0000000
4	Q1	20.0000000
5	Q2	21.0000000

```
6           Q3 23.0000000
7           Varianza  5.8264580
8  Deviazione standard  2.4138057
9           Simmetria  0.3333333
10          Curtosi -0.5216590
```

Density

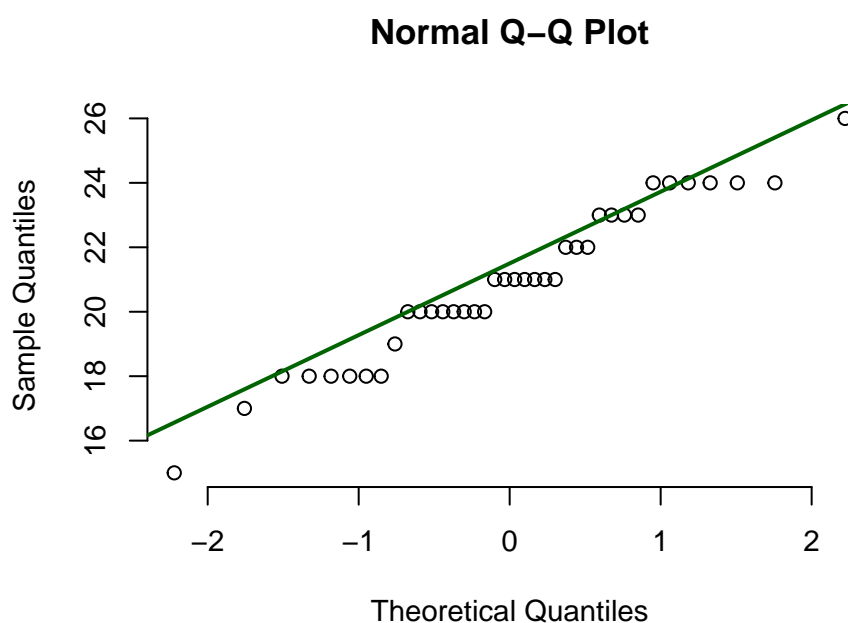
Il grafico di densità è un tipo di grafico che permette di visualizzare la forma della distribuzione di una variabile continua. Questa visualizzazione è particolarmente utile per comprendere la concentrazione e la variabilità dei dati e per individuare eventuali modalità o tendenze all'interno della distribuzione. Nel grafico di densità, sull'asse delle ordinate è rappresentata la densità di probabilità, che rappresenta la probabilità di trovare un'osservazione in una determinata posizione lungo l'asse delle ascisse. Più la curva è alta in un punto, maggiore è la probabilità che le osservazioni cadano in quella posizione. La distribuzione mostra una curva un po' più appiattita di una normale, il risultato è dovuta a punto con massa di probabilità che potrebbero far pensare a una bimodalità, in tal caso sarebbe meglio utilizzare l'indice della moda come riferimento. L'asimmetria è positiva, anche se in presenza di più punti di maggiore densità non ha senso parlare di asimmetria.



Normal qqplot

Un QQ-plot rappresenta una comparazione tra i quantili di due distribuzioni: la distribuzione dei dati osservati e una distribuzione teorica di riferimento (in questo caso la distribuzione normale). L'asse orizzontale rappresenta i quantili teorici della normale. L'asse verticale

rappresenta i quantili osservati delle spese per la casa rispetto alla normale. I punti sul grafico rappresentano le coppie di quantili. Ogni punto corrisponde a un valore nei dati osservati (asse y) e al suo corrispondente quantile teorico (asse x) secondo la normale. I dati osservati non seguono esattamente la distribuzione della normale infatti pochi punti si allineano sulla retta di riferimento evidenziando discrepanze dalla distribuzione. Poichè, molte osservazioni si trovano al di sotto della retta per cui le code dei dati osservati sono più leggere rispetto alla distribuzione di riferimento.



5.7. Camere per persona (variabile 3)

Per misurare la condizione di sovraffollamento si divide il numero di locali presenti nell'abitazione per il numero di persone che vi abitano. Un alloggio sovraffollato, infatti, può avere un'incidenza negativa sulla salute fisica e mentale, sui rapporti con gli altri e sullo sviluppo dei bambini. Il sovraffollamento, inoltre, comporta spesso servizi carenti in materia di fornitura idrica e fognature.

Statistiche

Nell'area dell'OCSE (degli stati considerati), le abitazioni in media contengono 1,7 vani a persona. Per quanto riguarda le dotazioni di base, il 97% delle abitazioni nei Paesi dispone di un accesso privato ai servizi igienici interni con scarico. La mediana, invece, è circa 1,75. Il risultato permette di comprendere che la metà degli stati ha un numero di camere per persona superiore a questo valore e l'altra metà ha un numero inferiore. Il valore più frequentemente presente è 1,9.

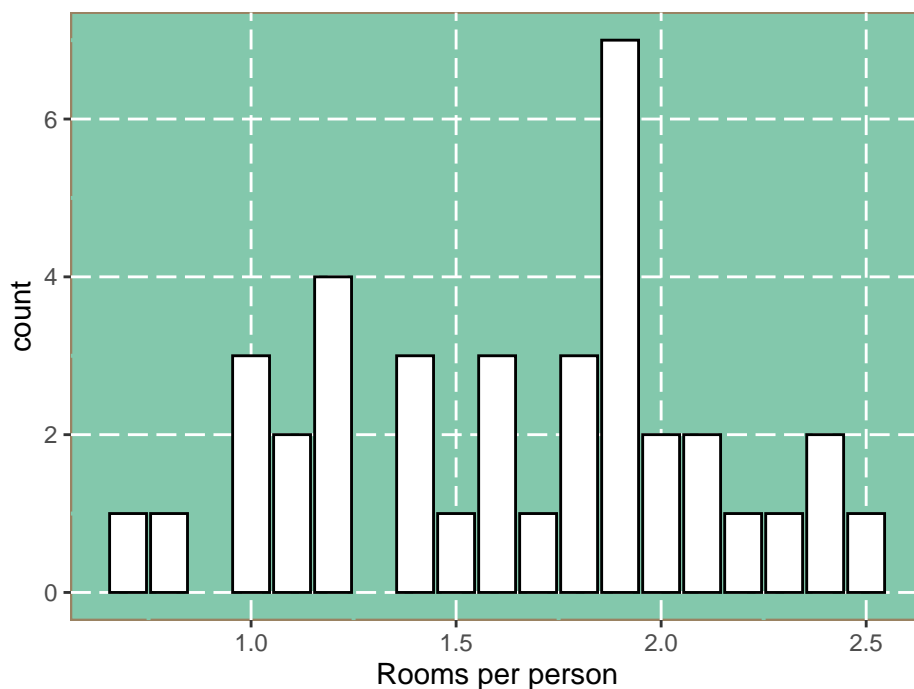
Il primo quartile (Q1) è 1,2 camere, il secondo quartile (Q2, che è anche la mediana) è 1,75 camere e il terzo quartile (Q3) è 1,9 camere per persona. La varianza è di circa 0,22, il che

indica che i dati sono relativamente vicini alla media. La deviazione standard è 0,47 indicando che la dispersione dei dati è moderata. Il valore della simmetria è -0,57 esprimendo una leggera tendenza verso la sinistra nella distribuzione, cioè alcuni stati potrebbero avere un numero inferiore di camere rispetto alla media. La curtosi misura quanto le code della distribuzione differiscono da una distribuzione normale. Il valore di -1,00 indica che la distribuzione ha code meno pesanti di una distribuzione normale.

	Statistiche	Valori
1	Media	1.6421053
2	Mediana	1.7500000
3	Moda	1.9000000
4	Q1	1.2000000
5	Q2	1.7500000
6	Q3	1.9000000
7	Varianza	0.2230441
8	Deviazione standard	0.4722754
9	Simmetria	-0.5714286
10	Curtosi	-1.0024714

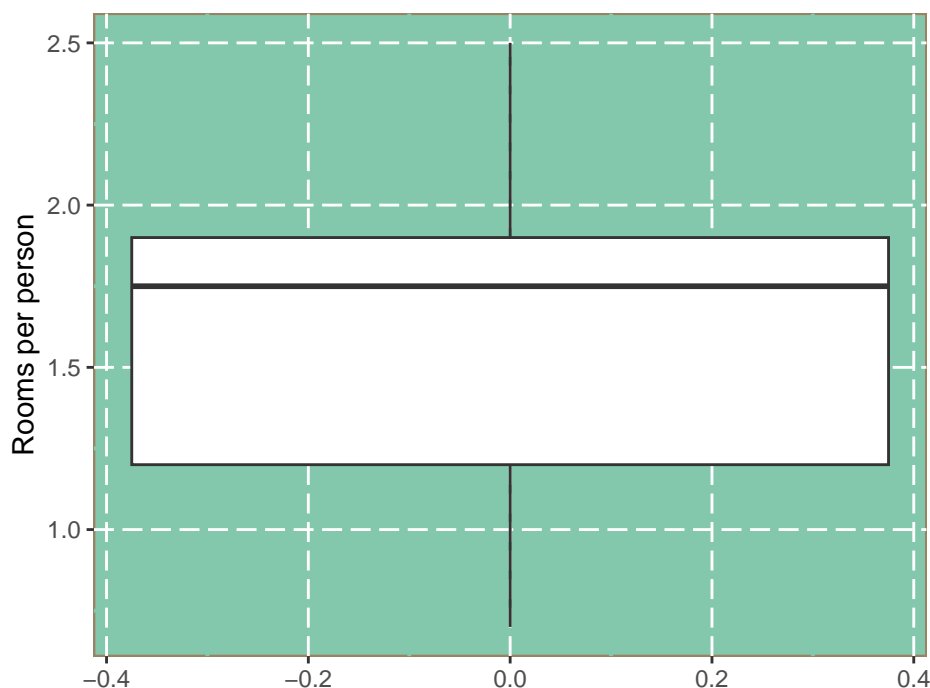
Istogramma

Il valore delle camere per persona viene racchiuso nel range [0.7; 2.5]. Come già evidenziato dalla moda la maggior frequenza delle osservazioni si presenta nel valore di 1.9, questo si riflette nel grafico con presenza di barra più alta. Non c'è evidenza di bimodalità ma presenza di leggera asimmetria negativa.



Boxplot

Il boxplot evidenzia l'asimmetria negativa nei dati. Non sono presenti valori anomali.

**5.8. Tasso di occupazione (variabile 4)***Statistiche*

All'interno dell'area dell'OCSE, circa il 67% della popolazione in età lavorativa, compresa tra 15 e 64 anni, ha un lavoro retribuito. I livelli occupazionali più elevati si registrano in Islanda (86%), Svizzera (80%) e Svezia (78%); mentre i più bassi si registrano in Sudafrica (43%), Turchia (51%), Grecia (52%). La mediana è del 69%: la metà degli stati ha un tasso di occupazione inferiore al 69% e l'altra metà ha un tasso superiore. Inoltre, il tasso di occupazione di molte aree è concentrato intorno al 72,5%. Osservando i quartili, si nota che il primo quartile (Q1) è del 65%, il secondo quartile (Q2, equivalente alla mediana) è del 69%, e il terzo quartile (Q3) è del 73,75%. La dispersione intorno alla media è circa 8,21%. La simmetria, con un valore vicino a zero (0,086), suggerisce che la distribuzione dei tassi di occupazione è approssimativamente simmetrica, senza una forte tendenza verso destra o sinistra. La curtosi, con un valore positivo di 1,02, indica una distribuzione con code leggermente più pesanti rispetto a una distribuzione normale.

	Statistiche	Valori
1	Media	67.73684211
2	Mediana	69.00000000
3	Moda	72.50000000
4	Q1	65.00000000

```

5           Q2 69.00000000
6           Q3 73.75000000
7           Varianza 67.44238976
8  Deviazione standard 8.21233157
9           Simmetria 0.08571429
10          Curtosi 1.01589713

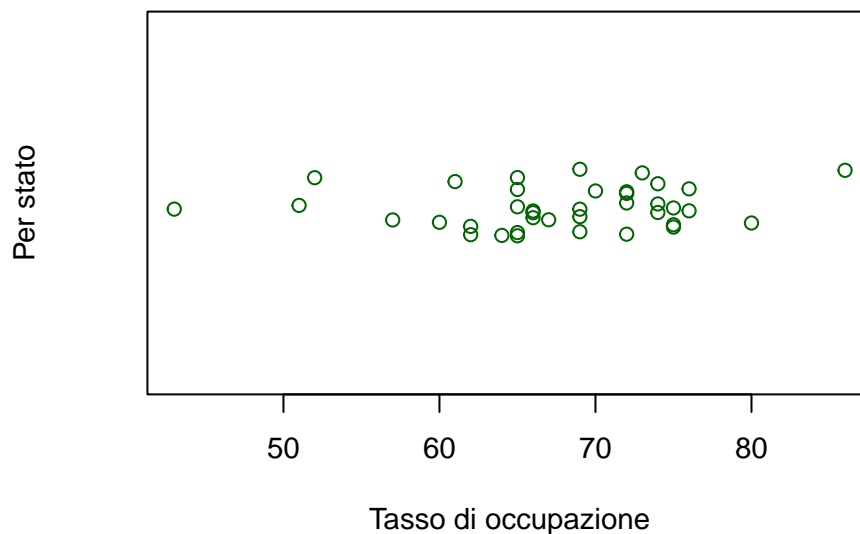
```

Stripchart

Nel grafico a stripchart, i punti dati vengono disposti lungo un'unica linea orizzontale o verticale, in modo che ciascun punto rappresenti un'osservazione o un valore specifico. I punti possono sovrapporsi se ci sono molte osservazioni con lo stesso valore. Questo tipo di grafico è spesso utilizzato per evidenziare la distribuzione dei dati, la concentrazione dei punti in determinate regioni e possibili outliers o valori anomali.

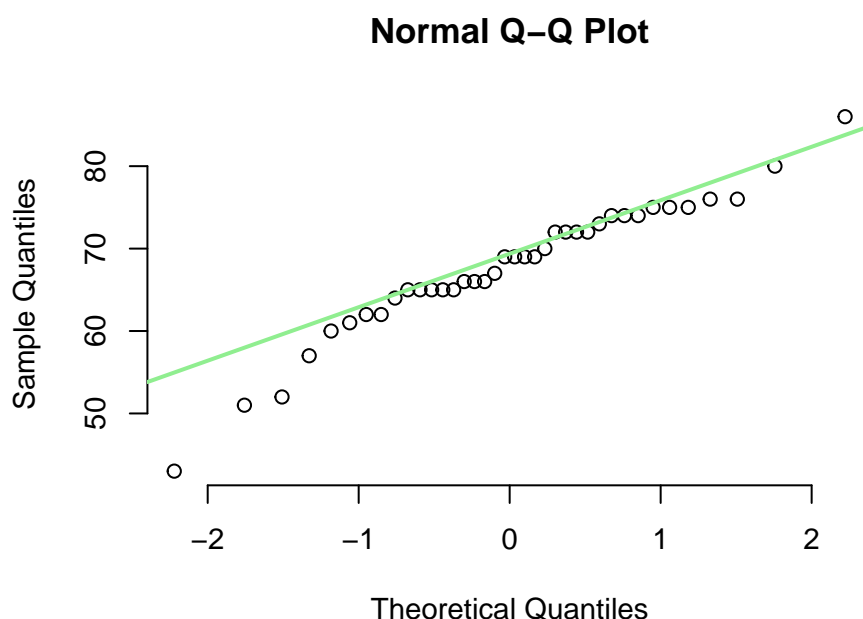
Per la variabile del tasso di occupazione i dati si concentrano per lo più tra 60 e 77 con 2 valori che cadono fuori l'intervallo dell'80% e 3 valori sotto il 50% (già analizzati precedentemente). In condizioni di osservazioni normali si potrebbe pensare a valori anomali ma in questo caso i valori giustificano l'appartenenza dello stato.

Percentuale dei tassi di occupazione nei diversi stati



Normal Qq plot

Per la variabile relativa al tasso di occupazione la similarità della distribuzione con quella della variabile casuale normale è più evidente rispetto alla variabile analizzata precedente. I dati osservati seguono esattamente la distribuzione teorica della normale, i punti, infatti, si allineano quasi perfettamente alla retta di 45 gradi tranne per i valori che sono stati già ritenuti “outliers” per la distribuzione ma non anomali per gli stati.



5.9. Reddito personale (Variabile 5)

Statistiche

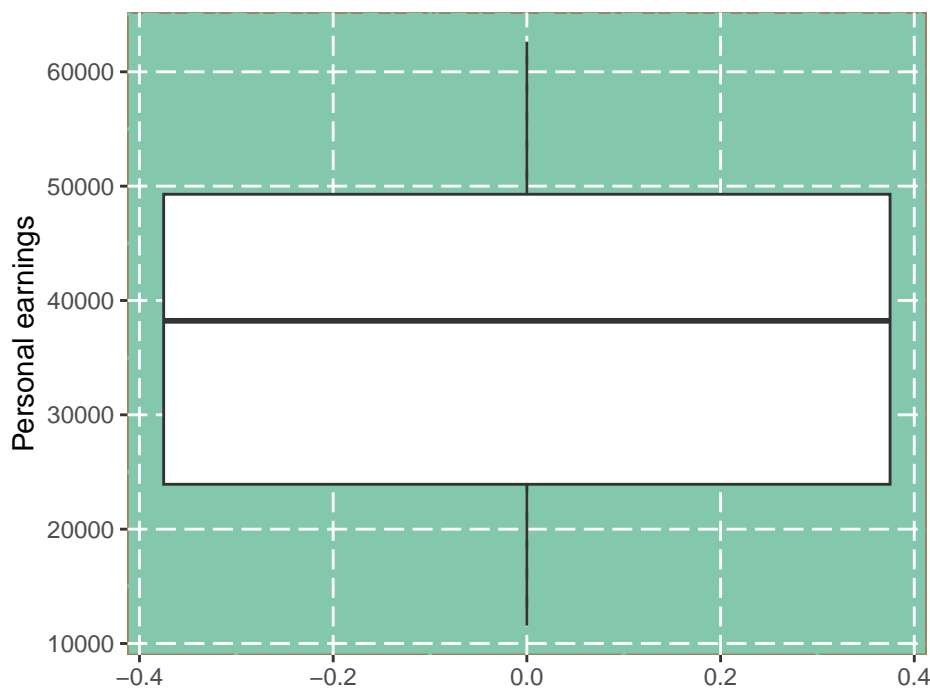
Nei Paesi dell'OCSE le persone guadagnano in media 37.436 USD all'anno, ma il reddito medio varia in maniera significativa da un Paese all'altro. Negli Stati Uniti, nel Lussemburgo e in Svizzera il reddito da lavoro medio è più del doppio di quello registrato in Europa orientale, Cile, Grecia, Ungheria, Messico e Portogallo. Tuttavia, è fondamentale notare che la mediana, che è di 38.223, è leggermente superiore alla media. Questo suggerisce che la distribuzione del reddito ha una leggera tendenza verso il lato superiore, con la maggior parte delle persone che guadagnano vicino o sopra la mediana. La moda, che è di 25.000, indica che ci sono molte persone con un reddito vicino a questa cifra. È interessante notare che la moda è significativamente inferiore alla media, suggerendo la presenza di un gruppo significativo di stati con redditi più bassi. Esaminando i quartili, si può vedere che il primo quartile è di 23.923, il secondo quartile, che coincide con la mediana, è di 38.223, e il terzo quartile è di 49.291. La varianza, che è di 203.261.600.000, suggerisce una notevole variabilità nei redditi, mentre la deviazione standard, di circa 14.256, indica quanto i redditi si discostino dalla media. La simmetria, con un valore leggermente negativo di -0,127, suggerisce una leggera asimmetria verso sinistra nella distribuzione del reddito, cioè alcuni stati potrebbero avere una proporzione di persone con redditi inferiori alla media più elevata. Infine, la curtosi, con un valore di -1,239, indica che la distribuzione ha code leggermente più pesanti rispetto a una distribuzione normale.

	Statistiche	Valori
1	Media	3.743595e+04

2	Mediana	3.822300e+04
3	Moda	2.500000e+04
4	Q1	2.392375e+04
5	Q2	3.822300e+04
6	Q3	4.929100e+04
7	Varianza	2.032616e+08
8	Deviazione standard	1.425699e+04
9	Simmetria	-1.273788e-01
10	Curtosi	-1.239225e+00

Boxplot

Il boxplot evidenzia una leggera asimmetria negativa data la grandezza della scatola. Non sono presenti valori anomali rispetto alla distribuzione. Lo stato il cui reddito medio percepito è più alto è lo stato del lussemburgo, mentre il più basso è il South Africa.



5.10. Student Skill (Variabile 6)

Statistiche

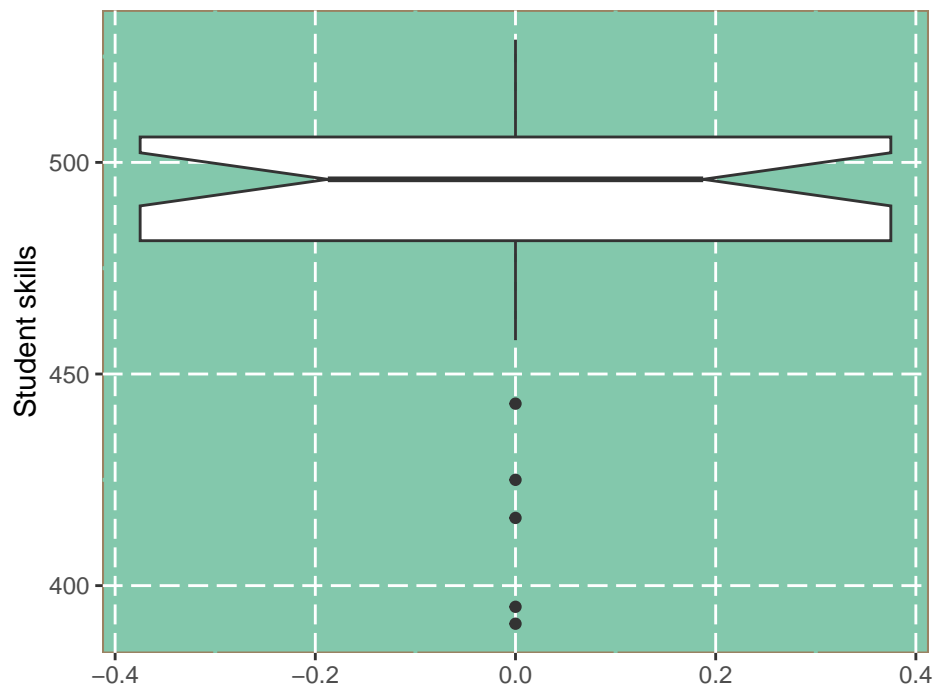
La media delle abilità degli studenti negli stati è di circa 486,76. Tuttavia la mediana, che è di 496, è superiore alla media. Ciò suggerisce che la maggior parte degli studenti si colloca al di sopra del valore medio, con alcune prestazioni eccezionali che spingono la mediana verso l'alto. La moda, che è di 490, indica che ci sono molti studenti con abilità vicine a questo valore. La moda è leggermente inferiore alla media ma comunque piuttosto vicina. Esaminando i

quartili, si vede che il primo quartile (Q1) è di 481,5, il secondo quartile (Q2), che coincide con la mediana, è di 496, e il terzo quartile (Q3) è di 506. La varianza è 1122,94 e la deviazione standard è di 33,51. La simmetria, con un valore leggermente negativo di -0,18, suggerisce una leggera asimmetria verso sinistra nella distribuzione delle abilità. Infine, la curtosi, con un valore di 1,40, indica che la distribuzione ha code più pesanti rispetto a una distribuzione normale. Il Giappone è in testa alla classifica, con un punteggio medio PISA di 529 punti. Seguono l'Estonia e la Finlandia con 524 e 523 punti. Il Paese con i risultati più bassi è il South Africa, con un punteggio medio di 391 punti. Ciò significa che il divario tra i Paesi con il più alto e il più basso rendimento è di 138 punti.

	Statistiche	Valori
1	Media	486.7631579
2	Mediana	496.0000000
3	Moda	490.0000000
4	Q1	481.5000000
5	Q2	496.0000000
6	Q3	506.0000000
7	Varianza	1122.9423898
8	Deviazione standard	33.5103326
9	Simmetria	-0.1836735
10	Curtosi	1.3955682

Box a Intaglio

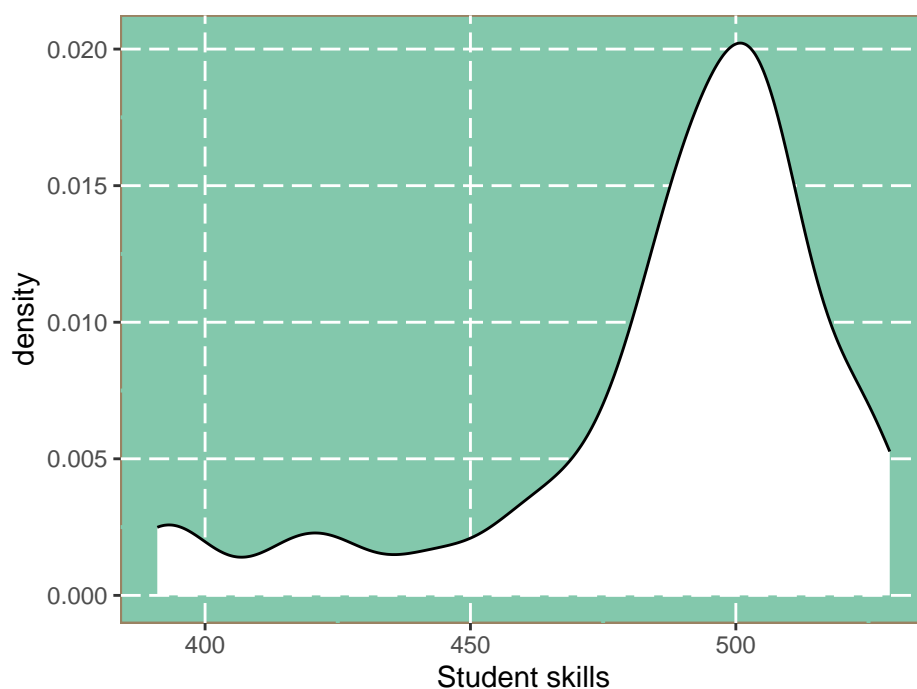
Si possono utilizzare anche i boxplot ad intaglio i quali sono una rappresentazione grafica dei boxplot ma con l'aggiunta dell'intervallo di confidenza. Con un grado di fiducia del 95%, l'intervallo di confidenza approssimato per la mediana delle skills degli studenti è (489.76; 502.23). Dal grafico vengono evidenziati possibili valori anomali, per il South Africa e il Brasile il punteggio medio è molto al di sotto della mediana, addirittura al di sotto del 400.



50%	50%
489.7602	502.2398

Density

Dalla visualizzazione della distribuzione è chiaramente possibile assimilare l'asimmetria negativa della distribuzione con code più lunghe a destra che a sinistra. Le code più pesanti fanno sì che la distribuzione si mostri leptocurtica, per cui più appuntita di una normale.



5.11. Aspettativa di vita (Variabile 7)

Statistiche

In media, la speranza di vita alla nascita raggiunge 79,54 anni, rappresentando una stima del periodo di vita medio nei paesi. Tuttavia la mediana è di 81,15 anni, supera la media. In molte nazioni l'aspettativa di vita si avvicina a 82,5. Esaminando i quartili, si vede che il primo quartile (Q1) è di 78,18 anni, il secondo quartile (Q2), coincidente con la mediana, è di 81,15 anni, e il terzo quartile (Q3) è di 82,25 anni. La deviazione standard, di circa 4,69, misura quanto le aspettative di vita si discostino dalla media, molto poco considerando la differenza interquartile. La simmetria, con un valore negativo di -0,46, è espressione di leggera asimmetria verso sinistra nella distribuzione delle aspettative di vita. Infine, la curtosi, con un valore notevolmente alto di 10,49, indica che la distribuzione ha code estremamente pesanti, il che suggerisce la presenza di alcuni paesi con aspettative di vita molto al di fuori della norma. Il Paese in cui la speranza di vita è più elevata è il Giappone, dove la media è di 84 anni; all'altra estremità della scala, la speranza di vita più bassa tra i Paesi si registra in Brasile a 74 anni, a 71 anni nella Federazione Russa e a 57 anni in Sudafrica.

	Statistiche	Valori
1	Media	79.5394737
2	Mediana	81.1500000
3	Moda	82.5000000
4	Q1	78.1750000
5	Q2	81.1500000
6	Q3	82.2500000

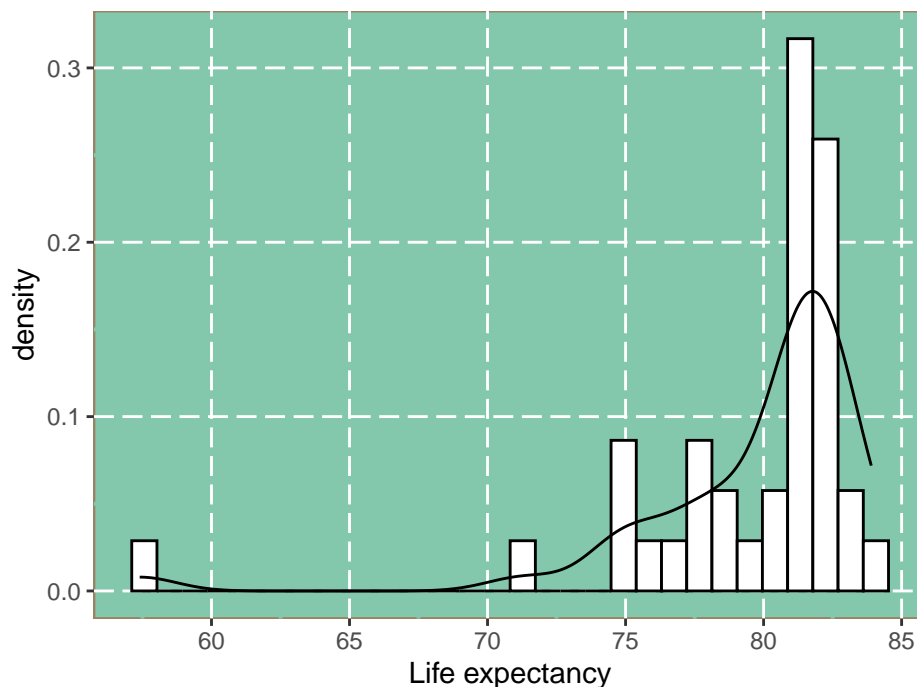
```

7          Varianza 21.9965078
8  Deviazione standard  4.6900435
9          Simmetria -0.4601227
10         Curtosi 10.4942969

```

Istogramma e Kernel Density

L'asimmetria negativa e lo spostamento dell'andamento della curva è pressochè dovuta alla presenza di valori anomali per la distribuzione (faccio riferimento all'Africa che si attesta valore medio di aspettativa per gli abitanti relativamente bassa, oltre che la Russia). Nonostante non sia presente bimodalità, la distribuzione è notevolmente più appuntita di una normale giustificando il risultato della statistica calcolata.



5.12. Stato di salute (Variabile 8)

Statistiche

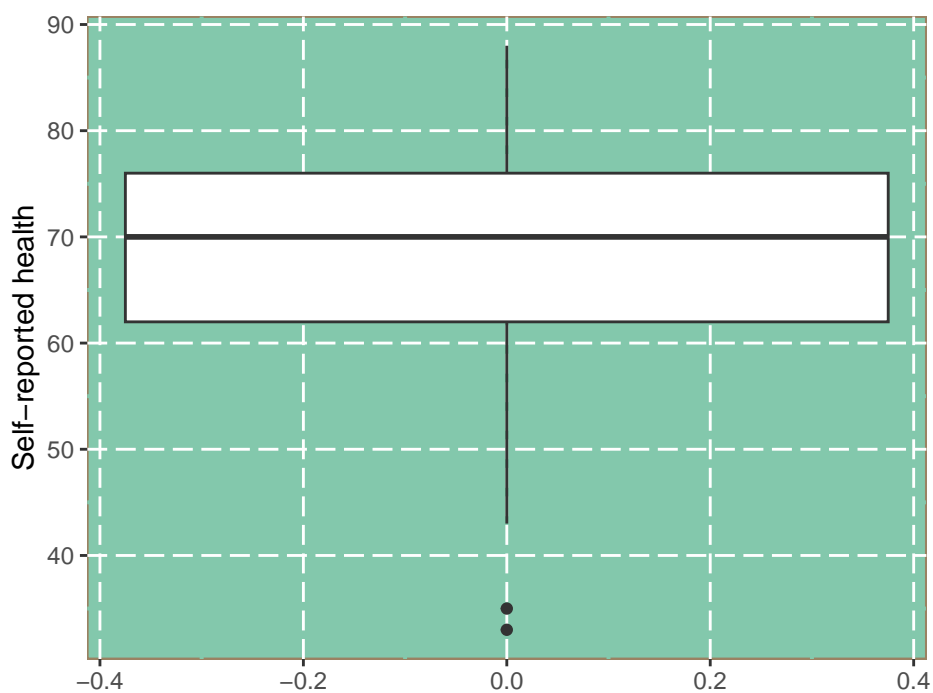
In tutta l'area ricoperta dallo studio, circa il 68% della popolazione adulta dichiara di essere in “buona” o “ottima” salute. In Canada, Nuova Zelanda e negli Stati Uniti, l'88% degli adulti dichiara di essere in buona salute, mentre in Giappone e Corea, nonostante le aspettative di vita molto alte, meno del 40% delle persone dichiara di essere in “buona” o “ottima” salute. La mediana, che si colloca a 70, suggerisce che la stragrande maggioranza delle regioni si trova al di sopra della media in termini di benessere salutare. La modalità, ossia il valore più frequente, è 65. Guardando i quartili, si nota che il primo quartile (Q1) è di 62, il secondo quartile (Q2) è di 70, e il terzo quartile (Q3) è di 76. Non intercorrono differenze

abissali nella massa di distribuzione. Le valutazioni discostano dalla media di circa 13,98, in termini percentuali. Il valore di simmetria è leggermente negativo -0.14 , suggerendo che la distribuzione delle valutazioni è leggermente inclinata verso sinistra. Infine, la curtosi ha valore pressochè nullo.

	Statistiche	Valori
1	Media	67.44736842
2	Mediana	70.00000000
3	Moda	65.00000000
4	Q1	62.00000000
5	Q2	70.00000000
6	Q3	76.00000000
7	Varianza	195.33499289
8	Deviazione standard	13.97622957
9	Simmetria	-0.14285714
10	Curtosi	-0.07705805

Boxplot

Il boxplot relativo alla percezione dello stato di salute dei cittadini dei diversi stati rende visibile una leggera asimmetria calcolata dall'indice precedente e una pressochè nulla curtosi. Vengono individuati due valori anomali al di sotto del valore 40 imputabili a Giappone e Korea. Il 35% dei giapponesi e il 33% dei coreani ha una percezione positiva del proprio stato di salute.



5.13. Tasso di Omicidi (Variabile 9)

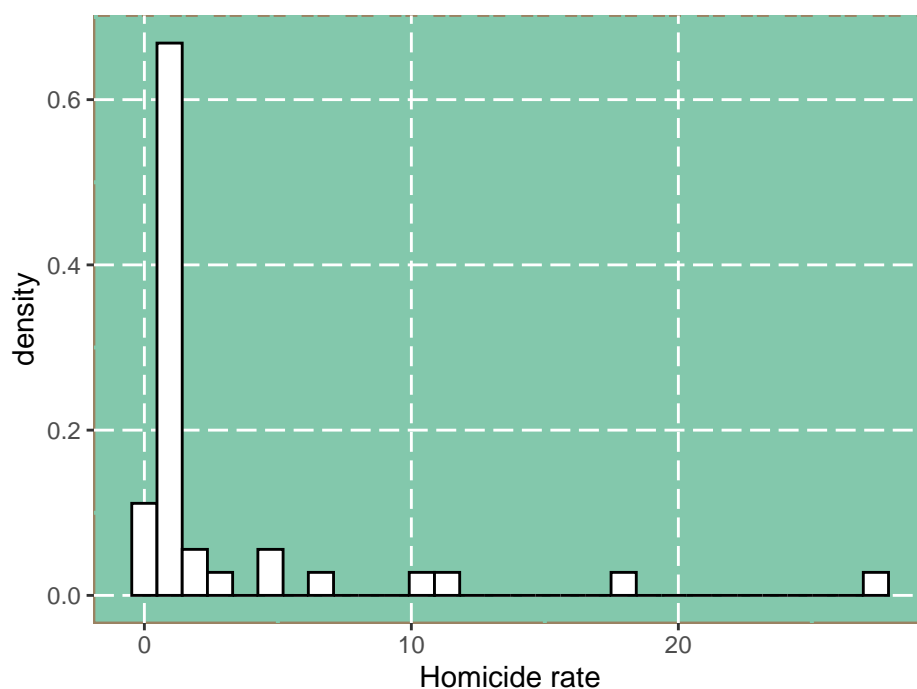
Statistiche

Il tasso di omicidi (numero di omicidi ogni 100 000 abitanti) si riferisce solo alla forma più estrema di violenza sulla persona e, di conseguenza, non fornisce informazioni sulle condizioni di sicurezza più rappresentative. Secondo i dati, il tasso medio di omicidi nell'area OCSE è pari a 2.93 omicidi ogni 100 000 abitanti. Il maggior tasso di omicidi è registrato in Brasile, mentre il tasso minore di omicidi in Gran Bretagna. La moda, ovvero il valore che compare più frequentemente, è 2.5. Il primo quartile (Q1) è 0.6, il secondo quartile (Q2) è 1, e il terzo quartile (Q3) è 1.625. Questi quartili indicano che la maggior parte dei paesi ha un tasso di omicidi relativamente basso, ma ci sono alcune eccezioni con tassi più elevati. La varianza è abbastanza alta, pari a circa 29.9, il che suggerisce una notevole variabilità nei tassi di omicidi tra i paesi. La deviazione standard implica che il valore intorno alla media si discosta in media di 5.47. La simmetria è positiva, con un valore di 0.22, suggerendo che la distribuzione è leggermente spostata verso destra, ma in generale, è approssimativamente simmetrica. La curtosi è piuttosto alta, a 9.62, il che suggerisce che la distribuzione ha code pesanti, cioè ci possono essere paesi con tassi di omicidi molto elevati rispetto alla media.

	Statistiche	Valori
1	Media	2.9342105
2	Mediana	1.0000000
3	Moda	2.5000000
4	Q1	0.6000000
5	Q2	1.0000000
6	Q3	1.6250000
7	Varianza	29.8985277
8	Deviazione standard	5.4679546
9	Simmetria	0.2195122
10	Curtosi	9.6166836

Istogramma

I valori calcolati dalle statistiche per descriverne la distribuzione non hanno effettivamente senso per questa variabile. Il valore del terzo quartile fissato a 1.62, indica che circa il 75% delle osservazioni rientrano prima della soglia fissata mentre il restante 25% è compreso tra la soglia e il valore massimo osservato per il Brasile pari al 27%. Per cui, parlare di curtosi e asimmetria non risulta essere del tutto corretto. In casi generici, si potrebbe pensare che i valori potrebbero essere anomali, ma siccome si sta facendo aggregazione su gruppi di stati che potrebbero avere indici differenti, si comprende la non anomalia dei rispettivi dati.



5.14. Attenzione sull'italia:

L'Italia ottiene risultati positivi in diverse dimensioni del benessere rispetto ad altri paesi nel Better Life Index. L'Italia supera la media equilibrio tra lavoro e vita privata e coinvolgimento civico, qualità ambientale e aspettativa di vita. Tuttavia, si colloca al di sotto della media in termini di salute, reddito, occupazione, istruzione, relazioni sociali e soddisfazione nella vita.

In Italia, il reddito disponibile netto medio per persona è di 35.397 dollari all'anno, leggermente inferiore alla media dei paesi di 37435 dollari all'anno.

Per quanto riguarda l'occupazione, circa il 57% delle persone tra i 15 e i 64 anni in Italia ha un lavoro retribuito, al di sotto della media totale del 67%. Il 66% degli uomini lavora, rispetto al 48% delle donne. In Italia, il 4% dei dipendenti lavora molte ore in modo retribuito, al di sotto della media totale del 8.72%, con il 5% degli uomini che lavorano molte ore in modo retribuito rispetto al 2% delle donne.

Nell'ambito educativo, il 60% degli adulti tra i 25 e i 64 anni ha completato l'istruzione secondaria superiore, inferiore alla media totale degli stati del 77%. Tuttavia, il completamento varia tra uomini e donne, con il 58% degli uomini che hanno completato con successo il liceo rispetto al 62% delle donne. Per quanto riguarda la qualità del sistema educativo, il punteggio medio degli studenti è di 485 in lettura, matematica e scienze nel Programma per la Valutazione Internazionale degli Studenti (PISA) dell'OCSE. Questo punteggio è leggermente inferiore alla media complessiva di 486. Le ragazze ottengono risultati migliori dei ragazzi di 4 punti, per cui 3 punti al di sopra della media, molto al di sotto del divario medio di 2 punti.

Per quanto riguarda la salute, l'aspettativa di vita alla nascita in Italia è di circa 82 anni, tre anni in più rispetto alla media totale di 79 anni. L'aspettativa di vita per le donne è di 85 anni, rispetto a 80 anni per gli uomini. Il livello di PM2,5 atmosferico, piccole particelle inquinanti dell'aria abbastanza piccole da penetrare e danneggiare i polmoni, è di 18 microgrammi per

metro cubo, superiore alla media di 13 microgrammi per metro cubo. In Italia, il 71% delle persone dichiara di essere soddisfatto della qualità dell'acqua, inferiore alla media dell'82%.

Per quanto riguarda la sfera pubblica, in Italia c'è un senso moderato di comunità e un alto livello di partecipazione civica, dove il 90% delle persone crede di poter contare su qualcuno in caso di necessità, inferiore alla media pari a 91%. Il tasso di affluenza alle urne, una misura della partecipazione dei cittadini al processo politico, è stato del 75% nelle elezioni del 2017, superiore alla media registrata come 70%. Lo status sociale ed economico può influenzare i tassi di voto.

Quando si chiede loro di valutare la loro soddisfazione generale nella vita su una scala da 0 a 10, gli italiani le danno una media di 5.9, inferiore alla media totale di 6.53.

6. ANALISI BIVARIATA

L'analisi bivariata permette di lavorare su ogni unità statistica e rilevare *congiuntamente* i due caratteri statistici X e Y , generando la rilevazione doppia (X, Y) . Si può trattare di due caratteri qualitativi, due caratteri quantitativi o un carattere qualitativo e uno quantitativo. Prima di analizzare le relazioni fra variabili si rende necessaria un'analisi preliminare degli indici e misure che caratterizzano tali rapporti.

6.1. 1) DUE VARIABILI QUANTITATIVE

6.2. COVARIANZA

Si è resa necessaria un'analisi preliminare delle variabili numeriche per comprendere quali fra queste hanno un legame di associazione elevato tale da poter essere considerato *statisticamente significativo*. E' possibile analizzare la matrice di varianza e covarianza per studiarne l'associazione. La **covarianza** permette di misurare la presenza di legame lineare tra due variabili numeriche ma non la forza di tale legame.

Il coefficiente è il seguente:

$$\text{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n ((x_i - \bar{x})(y_i - \bar{y}))$$

- se $\text{cov} = 0$, X e Y sono incorrelate, non esiste alcun legame lineare fra di loro;
- se $\text{cov} > 0$, X e Y sono correlate positivamente; a variazioni positive (negative) di una variabile corrispondono variazioni positive(negative) dell'altra variabile;
- se $\text{cov} < 0$, X e Y sono correlate negativamente; a variazioni positive di una variabile, corrispondono in media, variazioni negative dell'altra variabile e viceversa;

Nonostante la grandezza del dataset, si è cercato di cogliere i tratti essenziali dalla matrice di covarianza:

1. La variabile "Dwelling" con "Housing Exp.", e "Rooms per Person" mostra coefficiente leggermente negativo o vicino a zero, indicando che non c'è una forte associazione tra queste variabili.

2. Le variabili legate al reddito e alla ricchezza familiare (“Household Income” e “Household Wealth”) hanno covarianze positive molto elevate con altre variabili finanziarie come “Earnings” e “Support Net.”.
3. La variabile “Empl. Rate” è positivamente correlata con “Long-term Un.” e “Homicide Rate”, suggerendo che in alcune situazioni, un tasso di occupazione più basso potrebbe essere correlato a tassi più alti di disoccupazione a lungo termine e omicidi.
4. Alcune variabili, come “Air Pollution” e “Water Quality,” sono leggermente negative o vicine a zero con “Stakeholder Engagement” e “Voter Turnout”, indicando che non c’è una forte associazione tra i fattori.
5. Le variabili relative alla salute, come “Life Expectancy” e “Self-Reported Health” mostrano una correlazione positiva tra loro, suggerendo che paesi con un’aspettativa di vita più alta tendono ad avere una migliore autopercezione della salute.
6. La variabile “Safety at Night” ha una covarianza positiva significativa con “Homicide Rate”, indicando che paesi con un tasso di omicidi più elevato potrebbero percepire una minore sicurezza notturna.
7. “Education Att.” e “Student Skills” mostrano una forte correlazione positiva, il che suggerisce che i paesi con un maggior accesso all’istruzione tendono ad avere una migliore preparazione degli studenti.
8. Le variabili “Long Working Hours” e “Leisure Time” hanno una covarianza negativa, suggerendo che paesi con lunghe ore di lavoro potrebbero avere meno tempo libero.

	Dwelling	Housing Exp.	Rooms per Person
Dwelling	4.407446e+01	-4.85675676	-1.694595e+00
Housing Exp.	-4.856757e+00	5.82645804	1.126600e-01
Rooms per Person	-1.694595e+00	0.11266003	2.230441e-01
Household Income	-2.996674e+04	311.14509246	2.895658e+03
Household Wealth	-9.010458e+04	5333.67994310	1.217756e+04
Job Security	2.067419e+01	-0.28904694	-9.524893e-01
Empl. Rate	-2.811351e+01	3.59317212	1.949218e+00
Long-term Un.	1.008120e+01	1.01109531	-5.980057e-01
Earnings	-5.357245e+04	1956.42674253	5.443810e+03
Support Net.	-9.166216e+00	4.40825036	1.161024e+00
Education Att.	-3.375811e+01	2.51209104	2.592461e+00
Student Skills	-1.220824e+02	8.62304410	1.036159e+01
Years in Education	-3.435811e+00	0.51635846	3.050782e-01
Air Pollution	1.285811e+01	-3.30867710	-1.746799e+00
Water Quality	-3.751081e+01	5.24466572	3.315647e+00
Stakeholder Engagement	-8.358108e-01	0.22972973	4.162162e-02
Voter Turnout	-8.514865e+00	-1.15931721	1.161024e+00
Life Expectancy	-2.790500e+01	2.49615932	1.405050e+00
Self-Reported Health	-3.140676e+01	8.04836415	3.042817e+00
Life Satisfaction	-2.700676e+00	0.37069701	2.309104e-01
Safety at Night	-5.748257e+01	5.81991465	4.053926e+00
Homicide Rate	1.602608e+01	-2.32873400	-1.307966e+00
Long Working Hours	1.304803e+01	-1.37560455	-8.811906e-01
Leisure Time	-1.134541e+00	0.02315789	1.558179e-01
	Household Income	Household Wealth	Job Security

Dwelling	-2.996674e+04	-9.010458e+04	2.067419e+01
Housing Exp.	3.111451e+02	5.333680e+03	-2.890469e-01
Rooms per Person	2.895658e+03	1.217756e+04	-9.524893e-01
Household Income	6.441730e+07	2.466280e+08	-1.948626e+04
Household Wealth	2.466280e+08	1.498325e+09	-6.473552e+04
Job Security	-1.948626e+04	-6.473552e+04	2.615159e+01
Empl. Rate	3.583582e+04	1.375832e+05	-3.321721e+01
Long-term Un.	-1.157200e+04	-3.870139e+04	1.608694e+01
Earnings	1.073686e+08	4.211290e+08	-3.412451e+04
Support Net.	1.982465e+04	6.446363e+04	-5.339900e+00
Education Att.	5.492928e+04	1.954363e+05	-4.866721e+01
Student Skills	1.491813e+05	5.259050e+05	-1.018031e+02
Years in Education	3.843693e+03	1.016295e+04	-1.551927e+00
Air Pollution	-2.042469e+04	-6.081351e+04	8.166358e+00
Water Quality	5.898739e+04	2.055516e+05	-2.750982e+01
Stakeholder Engagement	3.422243e+02	2.484627e+03	-4.009459e-01
Voter Turnout	3.385865e+04	9.151636e+04	7.736131e-01
Life Expectancy	2.293420e+04	7.750696e+04	-1.354390e+01
Self-Reported Health	5.319610e+04	2.346073e+05	5.027738e-01
Life Satisfaction	4.168224e+03	1.500092e+04	-2.302809e+00
Safety at Night	7.481129e+04	2.498113e+05	-2.986917e+01
Homicide Rate	-2.185663e+04	-7.112191e+04	4.337589e+00
Long Working Hours	-1.388678e+04	-2.847870e+04	8.850617e+00
Leisure Time	2.248988e+03	7.733752e+03	-4.466003e-01

	Empl. Rate	Long-term Un.	Earnings	Support Net.
Dwelling	-2.811351e+01	1.008120e+01	-5.357245e+04	-9.1662162
Housing Exp.	3.593172e+00	1.011095e+00	1.956427e+03	4.4082504
Rooms per Person	1.949218e+00	-5.980057e-01	5.443810e+03	1.1610242
Household Income	3.583582e+04	-1.157200e+04	1.073686e+08	19824.6529161
Household Wealth	1.375832e+05	-3.870139e+04	4.211290e+08	64463.6273115
Job Security	-3.321721e+01	1.608694e+01	-3.412451e+04	-5.3399004
Empl. Rate	6.744239e+01	-2.179821e+01	6.721226e+04	18.5476529
Long-term Un.	-2.179821e+01	1.383058e+01	-1.916707e+04	-2.5488549
Earnings	6.721226e+04	-1.916707e+04	2.032616e+08	40112.4338549
Support Net.	1.854765e+01	-2.548855e+00	4.011243e+04	21.1073969
Education Att.	8.311807e+01	-2.360375e+01	9.211920e+04	19.5071124
Student Skills	1.728279e+02	-4.773760e+01	2.752719e+05	58.3036984
Years in Education	4.520626e+00	-1.020625e+00	9.565804e+03	2.9303698
Air Pollution	-2.562304e+01	5.583933e+00	-4.173679e+04	-18.7133713
Water Quality	6.326031e+01	-1.412125e+01	1.082977e+05	29.3982930
Stakeholder Engagement	5.486486e-01	-5.402297e-01	-2.546027e+02	-0.1878378
Voter Turnout	2.412518e+00	-4.907774e+00	6.308892e+04	13.5668563
Life Expectancy	2.002688e+01	-6.964661e+00	4.291126e+04	6.1259602
Self-Reported Health	1.752632e+01	-4.272902e-01	1.069867e+05	34.7716927
Life Satisfaction	4.361878e+00	-1.683362e+00	8.233202e+03	2.1127312
Safety at Night	6.608762e+01	-1.452567e+01	1.450722e+05	35.1909673
Homicide Rate	-1.190697e+01	8.632994e-01	-4.080084e+04	-6.7738976

Long Working Hours	-1.903671e+01	-2.322192e+00	-2.840446e+04	-14.5168421
Leisure Time	1.356543e+00	4.216292e-01	4.331126e+03	1.3482646
	Education Att.	Student Skills	Years in Education	
Dwelling	-3.375811e+01	-1.220824e+02	-3.435811e+00	
Housing Exp.	2.512091e+00	8.623044e+00	5.163585e-01	
Rooms per Person	2.592461e+00	1.036159e+01	3.050782e-01	
Household Income	5.492928e+04	1.491813e+05	3.843693e+03	
Household Wealth	1.954363e+05	5.259050e+05	1.016295e+04	
Job Security	-4.866721e+01	-1.018031e+02	-1.551927e+00	
Empl. Rate	8.311807e+01	1.728279e+02	4.520626e+00	
Long-term Un.	-2.360375e+01	-4.773760e+01	-1.020625e+00	
Earnings	9.211920e+04	2.752719e+05	9.565804e+03	
Support Net.	1.950711e+01	5.830370e+01	2.930370e+00	
Education Att.	2.558613e+02	3.928684e+02	4.921977e+00	
Student Skills	3.928684e+02	1.122942e+03	2.049424e+01	
Years in Education	4.921977e+00	2.049424e+01	1.939154e+00	
Air Pollution	-8.582504e+00	-6.295804e+01	-3.817425e+00	
Water Quality	7.401707e+01	2.189289e+02	8.614509e+00	
Stakeholder Engagement	9.040541e-01	6.391892e-01	2.094595e-02	
Voter Turnout	-3.349289e+01	-2.218279e+01	4.887127e+00	
Life Expectancy	2.502283e+01	1.017853e+02	2.741771e+00	
Self-Reported Health	-7.730441e+00	-2.029659e+01	5.743314e+00	
Life Satisfaction	3.984851e+00	8.996230e+00	5.255690e-01	
Safety at Night	9.984573e+01	3.276137e+02	9.737048e+00	
Homicide Rate	-4.062454e+01	-1.253565e+02	-3.155925e+00	
Long Working Hours	-6.086374e+01	-1.120649e+02	-2.117653e+00	
Leisure Time	4.028165e+00	1.195454e+01	2.344808e-01	
	Air Pollution	Water Quality	Stakeholder Engagement	
Dwelling	1.285811e+01	-37.510811	-0.83581081	
Housing Exp.	-3.308677e+00	5.244666	0.22972973	
Rooms per Person	-1.746799e+00	3.315647	0.04162162	
Household Income	-2.042469e+04	58987.394026	342.22432432	
Household Wealth	-6.081351e+04	205551.624467	2484.62702703	
Job Security	8.166358e+00	-27.509815	-0.40094595	
Empl. Rate	-2.562304e+01	63.260313	0.54864865	
Long-term Un.	5.583933e+00	-14.121252	-0.54022973	
Earnings	-4.173679e+04	108297.743954	-254.60270270	
Support Net.	-1.871337e+01	29.398293	-0.18783784	
Education Att.	-8.582504e+00	74.017070	0.90405405	
Student Skills	-6.295804e+01	218.928876	0.63918919	
Years in Education	-3.817425e+00	8.614509	0.02094595	
Air Pollution	3.478592e+01	-34.485064	-0.25540541	
Water Quality	-3.448506e+01	118.091038	1.67027027	
Stakeholder Engagement	-2.554054e-01	1.670270	0.48472973	
Voter Turnout	-1.847013e+01	21.425320	-0.72027027	
Life Expectancy	-9.167354e+00	26.146088	-0.01554054	
Self-Reported Health	-3.674893e+01	52.608819	1.69864865	

Life Satisfaction	-2.557681e+00	5.113798	0.05986486
Safety at Night	-3.386522e+01	106.425889	0.19608108
Homicide Rate	6.834282e-01	-28.757895	0.48797297
Long Working Hours	1.328169e+01	-26.502205	0.96597297
Leisure Time	-8.673826e-01	3.247511	-0.13735135
	Voter Turnout	Life Expectancy	Self-Reported Health
Dwelling	-8.5148649	-2.790500e+01	-3.140676e+01
Housing Exp.	-1.1593172	2.496159e+00	8.048364e+00
Rooms per Person	1.1610242	1.405050e+00	3.042817e+00
Household Income	33858.6529161	2.293420e+04	5.319610e+04
Household Wealth	91516.3570413	7.750696e+04	2.346073e+05
Job Security	0.7736131	-1.354390e+01	5.027738e-01
Empl. Rate	2.4125178	2.002688e+01	1.752632e+01
Long-term Un.	-4.9077738	-6.964661e+00	-4.272902e-01
Earnings	63088.9203414	4.291126e+04	1.069867e+05
Support Net.	13.5668563	6.125960e+00	3.477169e+01
Education Att.	-33.4928876	2.502283e+01	-7.730441e+00
Student Skills	-22.1827881	1.017853e+02	-2.029659e+01
Years in Education	4.8871266	2.741771e+00	5.743314e+00
Air Pollution	-18.4701280	-9.167354e+00	-3.674893e+01
Water Quality	21.4253201	2.614609e+01	5.260882e+01
Stakeholder Engagement	-0.7202703	-1.554054e-02	1.698649e+00
Voter Turnout	136.2965861	7.071906e+00	6.704196e+01
Life Expectancy	7.0719061	2.199651e+01	1.497105e+01
Self-Reported Health	67.0419630	1.497105e+01	1.953350e+02
Life Satisfaction	3.3586771	1.894232e+00	6.897511e+00
Safety at Night	13.9801565	4.405645e+01	6.097888e+01
Homicide Rate	-1.0441679	-1.453760e+01	-8.185989e+00
Long Working Hours	3.7026174	-5.559642e+00	-1.164632e+01
Leisure Time	1.0020484	9.119374e-01	7.199573e-01
	Life Satisfaction	Safety at Night	Homicide Rate
Dwelling	-2.700676e+00	-5.748257e+01	1.602608e+01
Housing Exp.	3.706970e-01	5.819915e+00	-2.328734e+00
Rooms per Person	2.309104e-01	4.053926e+00	-1.307966e+00
Household Income	4.168224e+03	7.481129e+04	-2.185663e+04
Household Wealth	1.500092e+04	2.498113e+05	-7.112191e+04
Job Security	-2.302809e+00	-2.986917e+01	4.337589e+00
Empl. Rate	4.361878e+00	6.608762e+01	-1.190697e+01
Long-term Un.	-1.683362e+00	-1.452567e+01	8.632994e-01
Earnings	8.233202e+03	1.450722e+05	-4.080084e+04
Support Net.	2.112731e+00	3.519097e+01	-6.773898e+00
Education Att.	3.984851e+00	9.984573e+01	-4.062454e+01
Student Skills	8.996230e+00	3.276137e+02	-1.253565e+02
Years in Education	5.255690e-01	9.737048e+00	-3.155925e+00
Air Pollution	-2.557681e+00	-3.386522e+01	6.834282e-01
Water Quality	5.113798e+00	1.064259e+02	-2.875789e+01
Stakeholder Engagement	5.986486e-02	1.960811e-01	4.879730e-01

Voter Turnout	3.358677e+00	1.398016e+01	-1.044168e+00
Life Expectancy	1.894232e+00	4.405645e+01	-1.453760e+01
Self-Reported Health	6.897511e+00	6.097888e+01	-8.185989e+00
Life Satisfaction	6.026529e-01	5.832226e+00	-6.077738e-01
Safety at Night	5.832226e+00	1.741250e+02	-4.933120e+01
Homicide Rate	-6.077738e-01	-4.933120e+01	2.989853e+01
Long Working Hours	-1.272391e+00	-3.320284e+01	8.243159e+00
Leisure Time	1.314424e-01	3.817203e+00	-1.556229e+00
	Long Working Hours	Leisure Time	
Dwelling	1.304803e+01	-1.13454054	
Housing Exp.	-1.375605e+00	0.02315789	
Rooms per Person	-8.811906e-01	0.15581792	
Household Income	-1.388678e+04	2248.98770982	
Household Wealth	-2.847870e+04	7733.75226174	
Job Security	8.850617e+00	-0.44660028	
Empl. Rate	-1.903671e+01	1.35654339	
Long-term Un.	-2.322192e+00	0.42162916	
Earnings	-2.840446e+04	4331.12563300	
Support Net.	-1.451684e+01	1.34826458	
Education Att.	-6.086374e+01	4.02816501	
Student Skills	-1.120649e+02	11.95453770	
Years in Education	-2.117653e+00	0.23448080	
Air Pollution	1.328169e+01	-0.86738265	
Water Quality	-2.650220e+01	3.24751067	
Stakeholder Engagement	9.659730e-01	-0.13735135	
Voter Turnout	3.702617e+00	1.00204836	
Life Expectancy	-5.559642e+00	0.91193741	
Self-Reported Health	-1.164632e+01	0.71995733	
Life Satisfaction	-1.272391e+00	0.13144239	
Safety at Night	-3.320284e+01	3.81720341	
Homicide Rate	8.243159e+00	-1.55622902	
Long Working Hours	6.079845e+01	-3.99852176	
Leisure Time	-3.998522e+00	0.57101963	

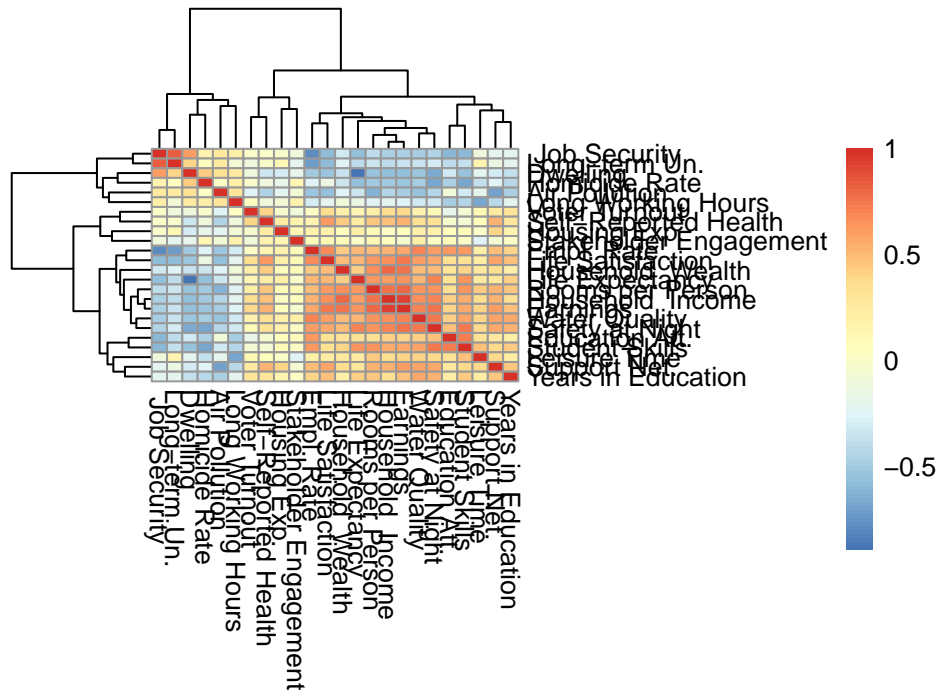
6.3. HEATMAP

L'evidenza dell'associazione fra variabili è possibile misurarla attraverso la **correlazione**. La correlazione è , appunto, una misura della relazione tra due variabili. Questa descrive come le variazioni di una variabile sono associate alle variazioni dell'altra variabile.

$$cor(X, Y) = \frac{S_X Y}{S_X S_Y}$$

La correlazione può essere positiva, quando le variazioni delle due variabili vanno nella stessa direzione, o negativa, quando le variazioni delle due variabili vanno in direzioni opposte. L'espressione del coefficiente di correlazione avviene attraverso un valore, che varia da -1 a +1. Un coefficiente di correlazione di 0 indica l'assenza di correlazione, mentre un valore di

+1 o -1 indica una correlazione perfetta tra le due variabili positivamente o negativamente. E' possibile quindi procedere alla visualizzazione del suddetto coefficiente attraverso un grafico heatmap, utilizzato per creare mappe di calore, ovvero grafici in cui le celle di una tabella sono colorate in base al loro valore, al fine di evidenziare modelli o tendenze nei dati.



Le righe e le colonne del grafico rappresentano le variabili numeriche, come prima descritte. La scala di valori rappresentata dalla legenda posizionata a destra è espressione del grado di correlazione fra variabili. Per cui, variabili che presentano un grado di associazione molto elevato, positivo, avranno in corrispondenza del match riga-colonna un colore tendente al rosso/arancione. Una casella colorata di giallo indica un'assenza di correlazione; una casella colorata celeste, tendente al blu scuro, indica una debole o forte associazione negativa. Risulta essenziale denotare che è una matrice simmetrica, la diagonale principale è colorata di rosso scuro poiché riflette un'ovvia associazione pari a 1 della variabile con sé stessa, per cui tutto ciò che si ripete nella parte superiore della diagonale si ripete nella parte inferiore della diagonale. Dalla lettura dell'heatmap della correlazione tra variabili, emergono diverse relazioni tra le variabili:

Forte correlazione negativa tra

1. **Dwelling e Life Expectancy:** paesi con condizioni abitative di bassa qualità (Dwelling) tendono a avere aspettative di vita più basse (Life Expectancy). Ciò potrebbe riflettere il fatto che le condizioni abitative influenzano la salute e il benessere della popolazione.
2. **Insecurity e Employment Rate:** La correlazione negativa tra Insecurity e Employment Rate suggerisce che in paesi con alti livelli di insicurezza nel mondo del lavoro, potrebbe esserci una minore occupazione. L'insicurezza del mondo del lavoro potrebbe influire negativamente sull'occupazione e la stabilità economica.

3. **Air Pollution e Support Net:** in aree con alti livelli di inquinamento dell'aria (Air Pollution), i sistemi di supporto sociale (Support Net) possono essere meno sviluppati. L'inquinamento dell'aria potrebbe avere un impatto negativo sulla qualità della vita e sul bisogno di reti di supporto.

Forte correlazione positiva tra:

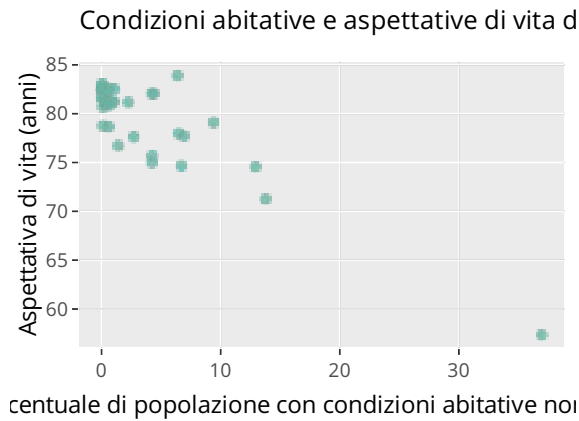
1. **Insecurity e Long-term Unemployment:** paesi con alti livelli di insicurezza lavorativa potrebbero avere anche tassi di disoccupazione a lungo termine più elevati.
2. **Household Income e Earnings:** I redditi familiari più elevati (Household Income) sono correlati positivamente ai guadagni (Earnings) a livello individuale. Questo suggerisce che in famiglie con redditi più alti, è probabile che ci siano anche guadagni individuali più alti.
3. **Household Wealth e Household Income:** Una correlazione positiva tra ricchezza familiare (Household Wealth) e reddito familiare (Household Income) suggerisce che famiglie più ricche tendono ad avere redditi più elevati.
4. **Employment Rate e Water Quality:** L'occupazione (Employment Rate) è correlata positivamente alla qualità dell'acqua (Water Quality), suggerendo che in paesi con alti tassi di occupazione, la qualità dell'acqua potrebbe essere migliore.
5. **Life Satisfaction e Earnings:** Una correlazione positiva tra la soddisfazione nella vita (Life Satisfaction) e i guadagni (Earnings) indica che in paesi con redditi più alti, le persone potrebbero essere generalmente più soddisfatte della propria vita.
6. **Safety at Night e Life Expectancy:** La sicurezza notturna (Safety at Night) è correlata positivamente all'aspettativa di vita (Life Expectancy), suggerendo che paesi in cui le persone si sentono più sicure durante la notte tendono ad avere aspettative di vita più lunghe.

Si procede all'analisi esplorativa bivariata tenendo conto delle informazioni precedenti.

6.4. Analisi Dwelling e Life Expectancy

Scatter plot. L'applicativo plotly permette di rendere interattivo il grafico e ottenere informazioni su ciascuno stato (se letto dal file html).

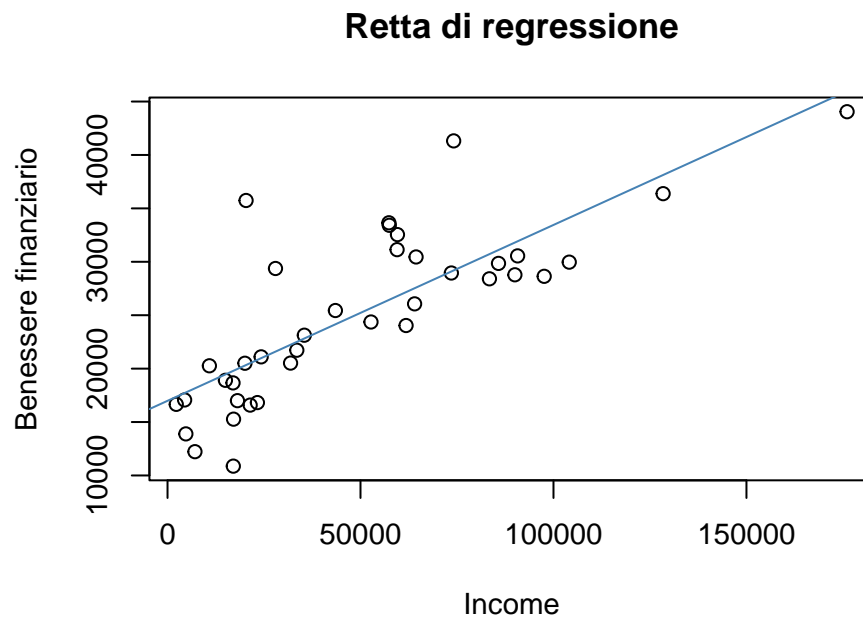
All'aumentare della qualità delle abitazioni, ci si aspetta che l'aspettativa di vita aumenti, e viceversa. Un miglioramento delle condizioni abitative, ad esempio attraverso una migliore o istituzione di servizi sanitari, un accesso più agevole all'acqua potabile o un ambiente più sicuro, può influenzare positivamente la salute delle persone, aumentando la loro aspettativa di vita.





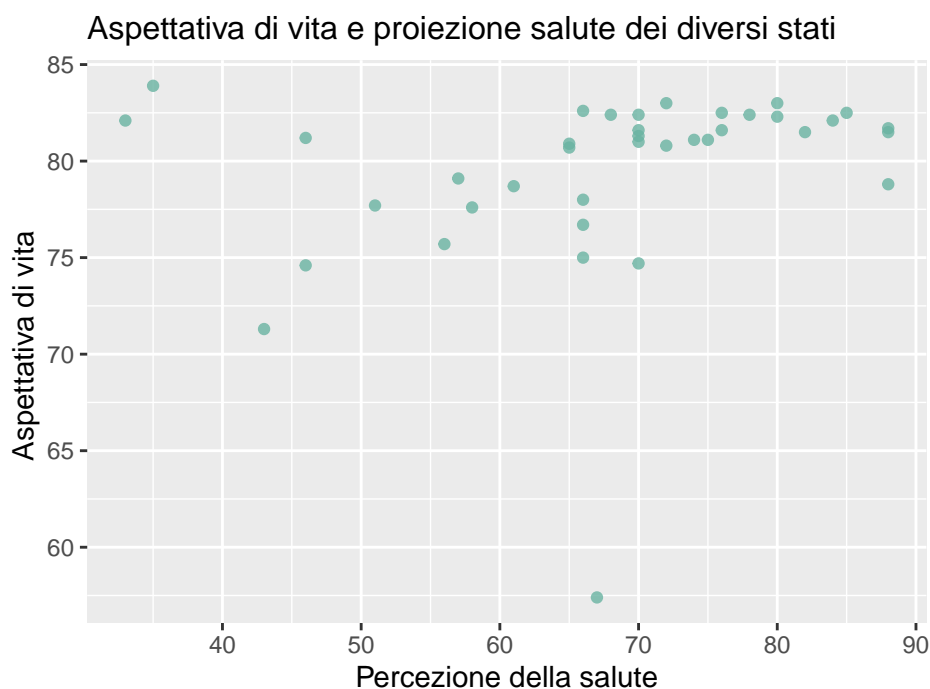
6.6. Analisi Household Income e Household Wealth

Retta di regressione. Una relazione positiva tra “Household Income” (reddito familiare) e “Household Wealth” (patrimonio familiare) suggerisce che, in generale, famiglie con redditi più elevati tendono ad avere anche una maggiore ricchezza complessiva. In modo logico, le famiglie con redditi più elevati hanno una maggiore capacità di risparmiare. Possono mettere da parte una parte più consistente dei loro redditi, contribuendo così a incrementare il loro patrimonio, oltre che permettersi di investire in strumenti finanziari o proprietà che generano rendimenti, contribuendo così a far crescere il loro patrimonio. Dal grafico è evidente che le due variabili sono moderatamente legate linearmente, infatti lo scostamento delle osservazioni dalla retta di regressione lineare (retta interpolante ascendente) è discreto.



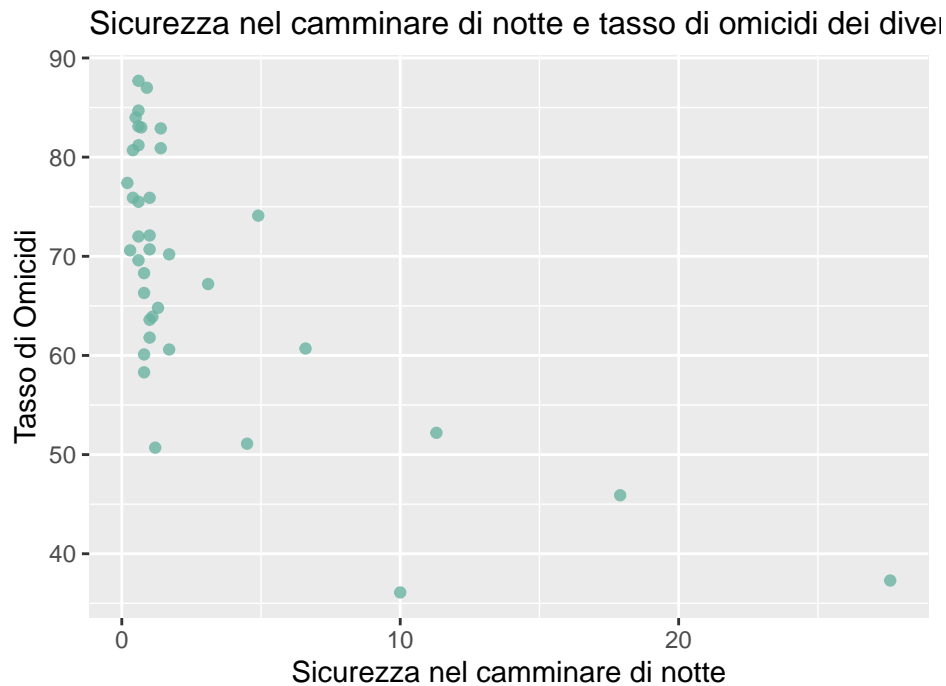
6.7. Analisi Life Expectancy e Self-Reported Health

Scatter plot Tra le due variabili è presente una correlazione prossima allo 0.22. C'è un legame positivo ma non perfettamente lineare. In linea di massima, la relazione positiva suggerirebbe che le persone che valutano la propria salute come migliore tendono ad avere aspettative di vita più lunghe.



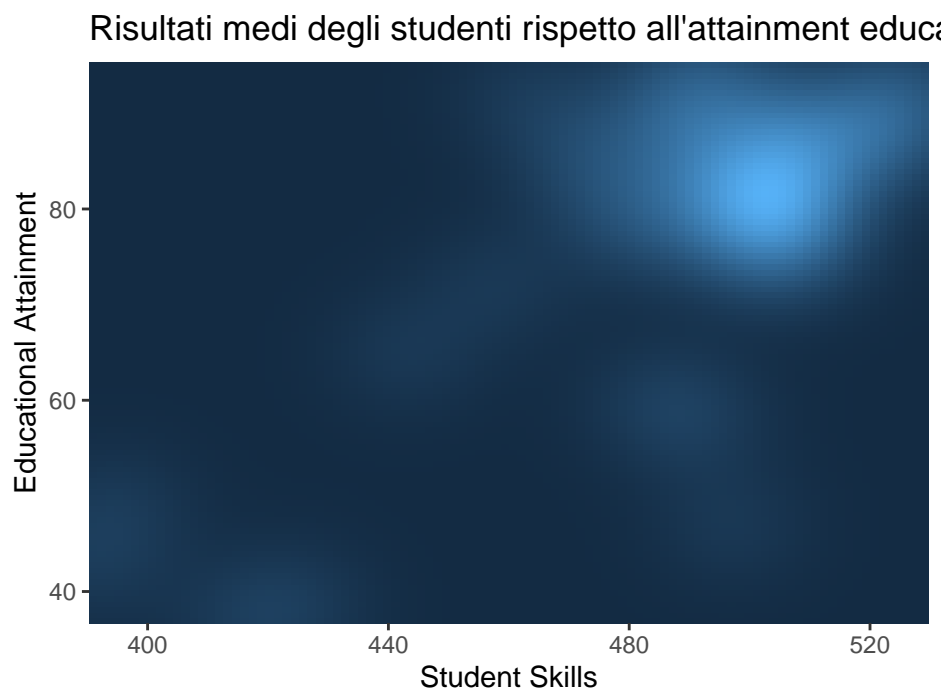
6.8. Analisi Safety at night e Homicide Rate

Scatter plot Una correlazione di -0.68 tra sicurezza durante la notte e tasso di omicidi indica una correlazione negativa moderatamente forte tra queste due variabili. Il valore negativo della correlazione suggerisce che quando il livello di sicurezza nel camminare di notte di soli aumenta (indicando una maggiore percezione di sicurezza durante la notte), il tasso di omicidi tende a diminuire. Viceversa, se la “Safety at night” diminuisce (indicando una minore percezione di sicurezza durante la notte), il tasso di omicidi tende a aumentare. Chiaramente, il risultato suggerisce che una maggiore sicurezza percepita durante la notte potrebbe essere associata a una minore incidenza di omicidi in una determinata area. E’ tuttavia importante notare che “Safety at night” è una misurazione soggettiva basata sulla percezione delle persone riguardo alla propria sicurezza. Questa percezione può essere influenzata da vari fattori, tra cui la presenza di polizia, l’illuminazione pubblica, il livello di criminalità effettiva e l’educazione sulla sicurezza; si aggiunge che il valore presente all’interno del dataset è una media di tutte le osservazioni per ciascuno stato ma non sul singolo individuo. È importante tenere presente che la correlazione non implica causalità diretta. Altri fattori possono influenzare sia la percezione della sicurezza durante la notte che il tasso di omicidi. Ad esempio, un tasso di omicidi più elevato può influenzare negativamente la percezione della sicurezza, ma ci possono essere altri fattori, come la povertà, la disoccupazione e la demografia, che possono confondere la relazione.



6.9. Analisi Education Att. e Student Skills

2D Density plot La correlazione tra livello di istruzione conseguito e abilità degli studenti è prossima al 0.73. Il valore positivo della correlazione suggerisce che quando il livello di “Education Attainment” aumenta (cioè, quando le persone hanno completato livelli di istruzione più elevati), le “Student Skills” tendono ad aumentare. Questa correlazione indica che un maggiore livello di istruzione è generalmente associato a un miglioramento delle abilità degli studenti. L’istruzione formale può fornire alle persone le conoscenze e le competenze necessarie per avere successo in vari settori della vita, sia accademici che professionali. L’istruzione non riguarda solo l’acquisizione di conoscenze, ma anche lo sviluppo di abilità cognitive, analitiche e di problem-solving.



6.10. 2) DUE VARIABILI QUALITATIVE

6.11. Analisi Air pollution e Life satisfaction

Distribuzione di frequenza congiunta

Per sintetizzare i dati di una distribuzione doppia si può utilizzare una tabella di contingenza in cui vengono riportate le **frequenze congiunte** delle modalità di X, Y .

In tal caso si sta confrontando la distribuzione congiunta di stati rispetto a soddisfazione della vita e inquinamento dell'aria. In particolare: 9 osservazioni sono nel gruppo di unità con valore di inquinamento 1 e soddisfazione 1, 5 unità nel gruppo di soddisfazione 0 e inquinamento 0, 13 inquinamento 0 e soddisfazione 1, 11 inquinamento 1 e soddisfazione 0.

	0	1
0	5	13
1	11	9

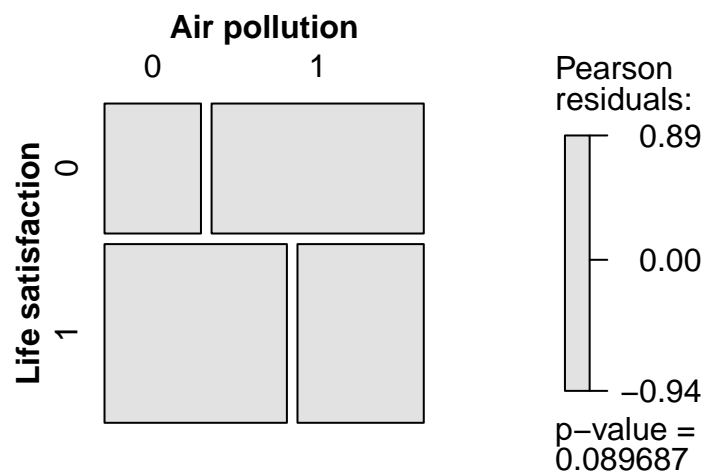
Mosaic plot e grouped bar plot

Il test V di Cramer permette di fornire informazioni circa il grado di associazione tra due variabili qualitative. Tanto più il valore è prossimo a 1, tanto più è presente dipendenza tra i caratteri. Il valore di Cramer's V intorno a 0.275 suggerisce un'associazione moderata tra le variabili. Per valutarne la significatività viene analizzato il mosaic plot e i relativi residui di

pearson. La proporzione maggiore di osservazioni è visualizzata sulla base della grandezza dei rettangoli; in tal caso pollution 0 e soddisfazione 1 ha la priorità. Ci sono effettive differenze tra i livelli di inquinamento dell'aria. La soddisfazione è più alta in condizioni inquinamento dell'aria basso.

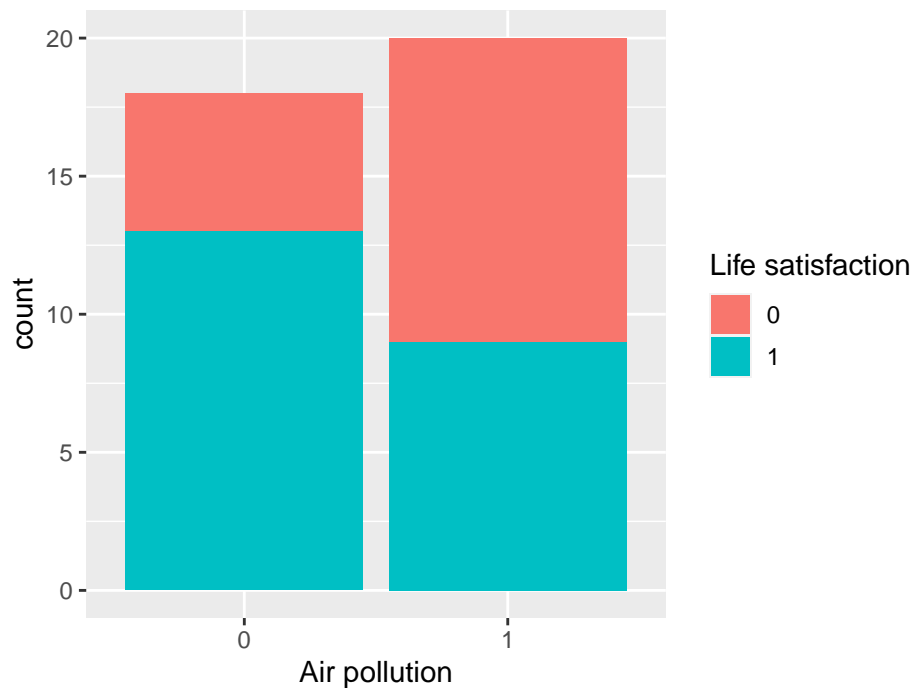
Un p-value di 0.08 è al di sopra di molti livelli comuni di significatività (0.05). Per cui, un p-value superiore a 0.05 potrebbe suggerisce che non vi è evidenza sufficiente per respingere l'ipotesi nulla di assenza di associazione.

[1] 0.2752983



Gruped Bar plot

E' possibile trarre le stesse informazioni osservando anche questo grafico.



6.12. 3) UNA VARIABILE QUALITATIVA E UNA QUANTITATIVA

Infine, è anche possibile rappresentare relazioni presenti tra una variabile quantitativa e una qualitativa.

Si utilizza quindi le variabili che sono state costruite nell'analisi della variabile categoriale per svolgere osservazioni su gruppi differenti. Si ricorda che:

- Air pollution (factor) è stata creata classificando 1 i paesi la cui media di inquinamento dell'aria è maggiore della mediana totale 0 altrimenti;
- Life Satisfaction (factor) è stata creata classificando con 1 i paesi la cui media della soddisfazione personale è maggiore della media totale 0 altrimenti;

6.13. Analisi Life Satisfaction e Personal Earnings

Analisi numerica

Life satisfaction	Personal earnings
0:16	Min. :11554
1:22	1st Qu.:23924
	Median :38223
	Mean :37436
	3rd Qu.:49291
	Max. :62636

Svolgendo analisi sulle medie dei gruppi, tenendo conto che:

- **Life Satisfaction 1:** è il gruppo che include tutti i paesi il cui valore di soddisfazione personale è superiore alla media.
- **Life Satisfaction 0:** è il gruppo che include tutti i paesi il cui valore di soddisfazione personale è al di sotto della media.

e la variabile “Media_guadagno” rappresenta il guadagno medio per ciascuno di questi due gruppi. E’ possibile distinguere la popolazione in :

1. **Life Satisfaction 1 (Soddisfatti):** gruppo con i paesi in cui le persone sono più soddisfatte della loro vita rispetto alla media globale. Il valore medio del guadagno in questi paesi è di 44,411.86 .
2. **Life Satisfaction 0 (Insoddisfatti):** gruppo con i paesi in cui le persone sono meno soddisfatte della loro vita rispetto alla media globale. Il valore medio del guadagno in questi paesi è di 27,844.06.

La differenza tra i due gruppi è notevole. Nei paesi in cui le persone sono più soddisfatte della loro vita (Life Satisfaction 1), il guadagno medio è significativamente superiore rispetto ai paesi in cui le persone sono meno soddisfatte (Life Satisfaction 0). Il risultato è espressione di correlazione positiva tra la soddisfazione nella vita e il guadagno medio, indicando che nei paesi con livelli di soddisfazione personale più elevati, è più probabile che le persone abbiano guadagni medi più alti.

```
# A tibble: 2 x 2
  'Life satisfaction' media_guadagno
  <fct>                <dbl>
1 0                    27844.
2 1                    44412.
```

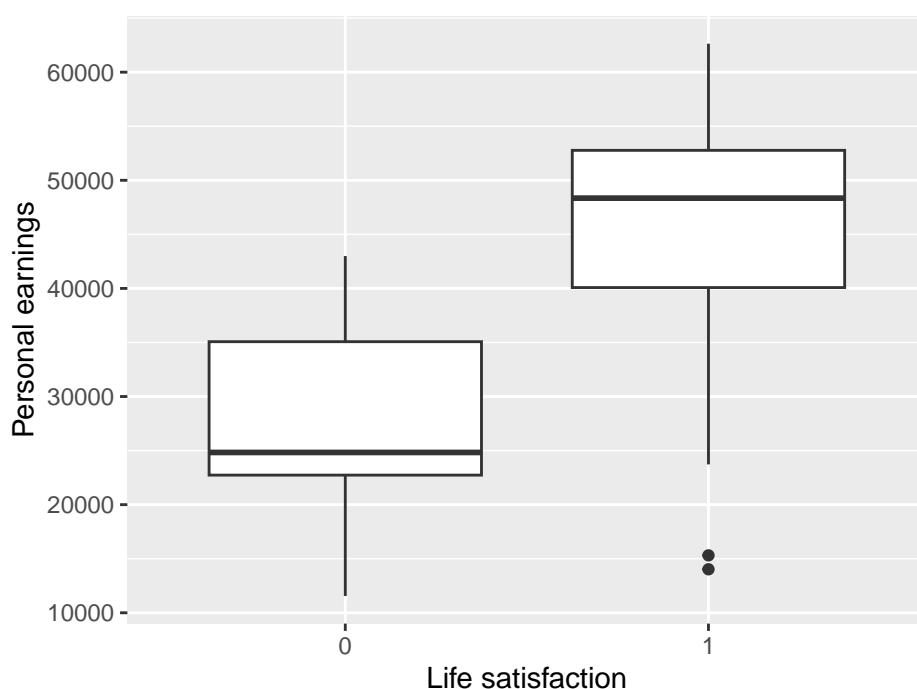
Confrontando i gruppi sulle mediana, in definitiva i paesi che si trovano nel sottoinsieme 1 per soddisfazione della vita guadagnano molto di più rispetto a quelli classificati come 0, ma la differenza in termini monetari è maggiore di circa 7000 unità.

```
# A tibble: 2 x 2
  'Life satisfaction' mediana_guadagno
  <fct>                <dbl>
1 0                    24826.
2 1                    48349
```

Boxplot paralleli

I boxplot paralleli, noti anche come boxplot a confronto, sono uno strumento grafico utile per visualizzare e confrontare la distribuzione di più gruppi o categorie in base a una variabile. Nel caso specifico sono utili per confrontare i guadagni medi (Media_guadagno) in base ai

due gruppi di soddisfazione nella vita (Life Satisfaction 0 e Life Satisfaction 1). L'asse delle ordinate rappresenta il valore della variabile del guadagno personale. L'asse delle ascisse rappresenta le categorie: Life Satisfaction 0 e Life Satisfaction 1. Per ciascun gruppo è stato costruito il boxplot. Sulla base della rappresentazione grafica, risulta chiaramente evidente la distinzione delle mediane dei due gruppi, più bassa per il gruppo di life satisfaction 0. La grandezza dei box è piuttosto simile, per cui in termini di variabilità si potrebbe aspettare gli stessi termini, tuttavia ciò che le differenzia è una asimmetria positiva per il gruppo 0 e negativa per il gruppo 1. Sono presenti due valori anomali, per cui l'aspettativa di vita è al di sopra della mediana ma il salario personale è piuttosto basso: Brasile e Messico.



Densità Kernel

Dal plot della densità risulta essere più chiara la distribuzione. Sebbene le mediane dei gruppi sono differenti, entrambi le classi mostrano asimmetrie rispettivamente positiva per il gruppo 0 e negativa per il gruppo 1.



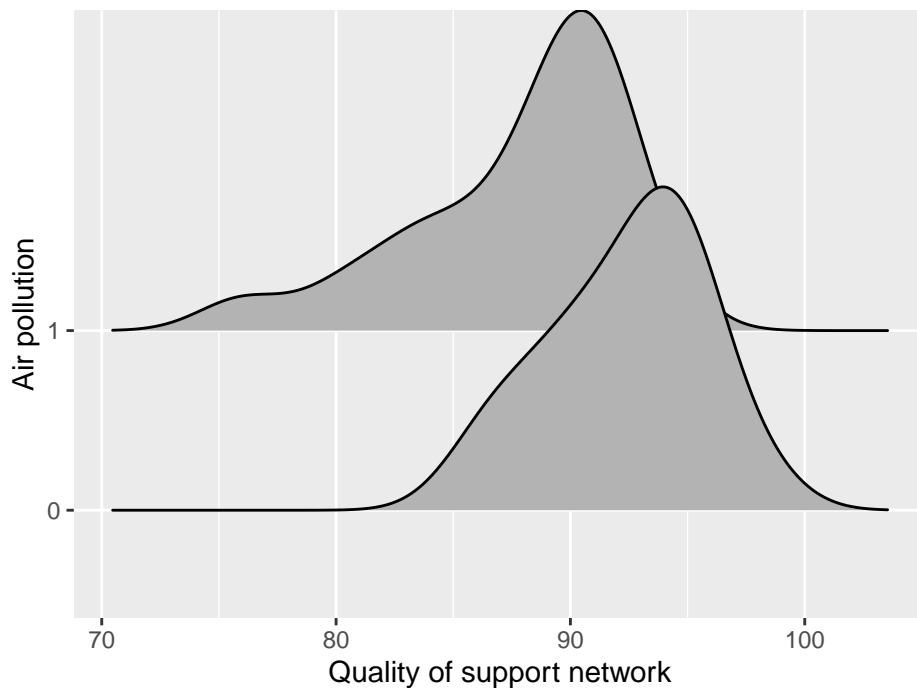
6.14. Analisi Air pollution e Support Network

Analisi numerica

Tenendo conto che la variabile pollution assume 1 per tutti gli stati il cui livello di inquinamento dell'aria è superiore alla media, 0 altrimenti, è possibile confrontare le medie dei due gruppi rispetto alla qualità della rete di supporto. E' evidente che il gruppo "Air pollution 0" ha una media della qualità dell'aria più elevata (92.38889) rispetto al gruppo "Air pollution 1" (87.90000). La differenza nella media della qualità dell'aria tra i due gruppi suggerisce che gli stati con livelli inferiori di inquinamento dell'aria tendono a presentare una qualità dell'aria media più elevata rispetto agli stati con livelli superiori di inquinamento.

```
# A tibble: 2 x 2
  'Air pollution' media_qualità_supporto
  <fct>          <dbl>
1 0              92.4
2 1              87.9
```

Densità Kernel

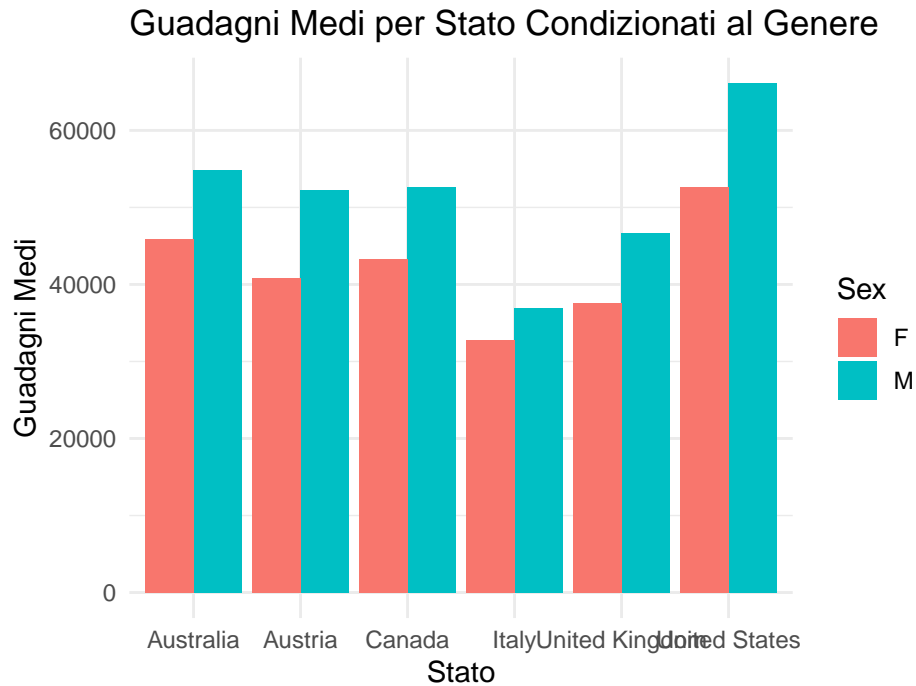


7. ANALISI MULTIVARIATA

Si può infine, esplorare la relazione fra tre e più variabili.

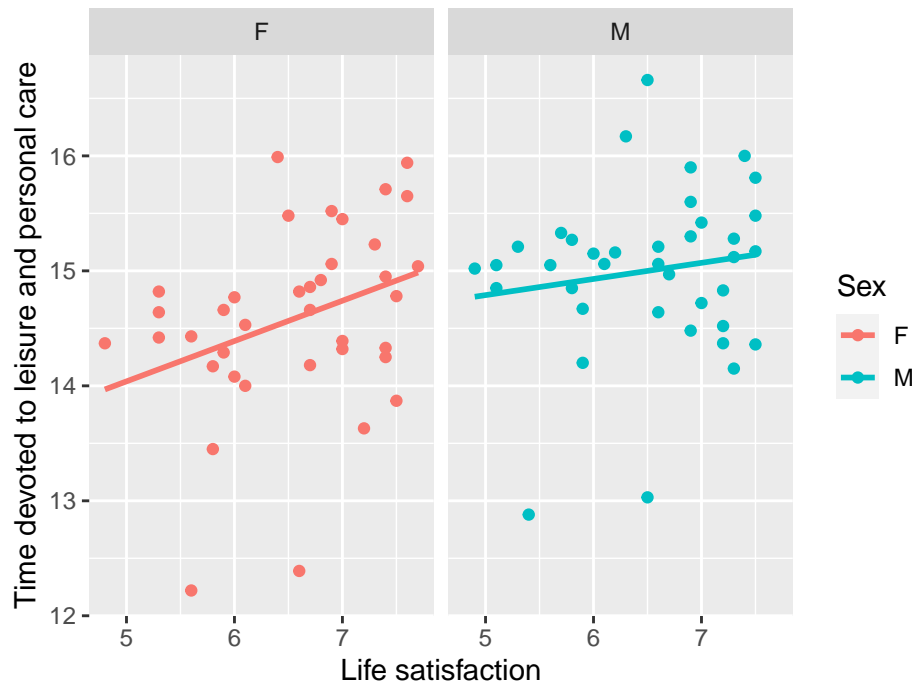
7.1. RELAZIONE: Stato, Salario e genere

Il better index dataset permette di splittare il dataset originale in osservazioni in merito al sesso biologico per poter effettuare analisi di discriminazioni. La prima relazione multivariata è tra gli stati (ne ho considerate solo 6 per una migliore visualizzazione visiva), il reddito personale percepito medio e il genere. E' chiaro che, in linea generale, in tutti gli stati il salario è più alto per gli uomini che per le donne. La maggiore differenza tra i due sessi in termini monetari viene percepita negli Stati Uniti, che per altro possiede anche i valori più alti. La minore differenza di salario tra i due sessi è registrata in Italia, che per altro possiede i salari minimi più bassi tra i paesi confrontati.



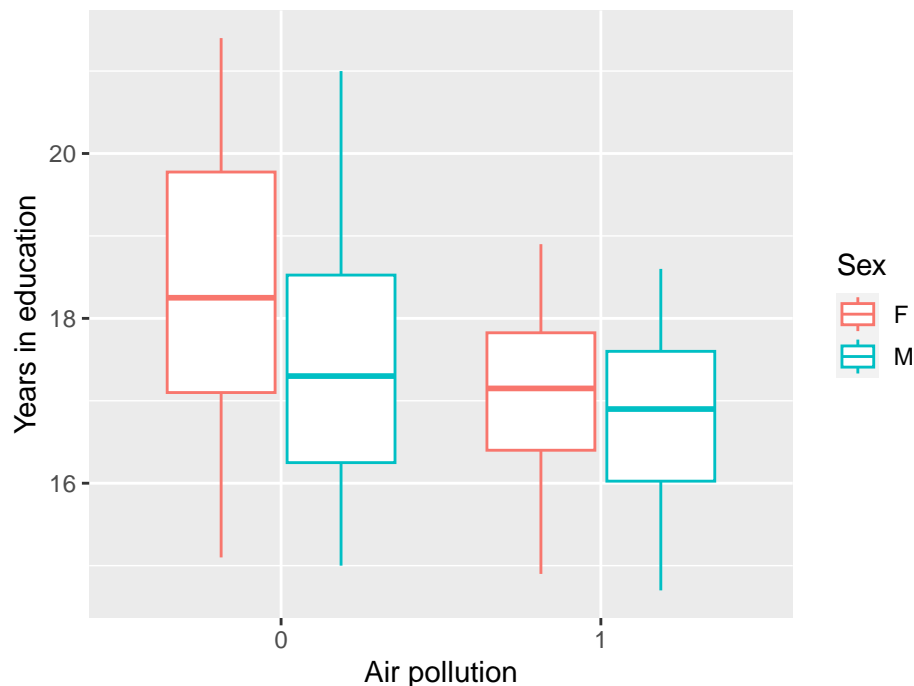
7.2. RELAZIONE: Life Satisfaction, Work-life balance e Gender

La relazione tra “Life Satisfaction” e “Work-Life Balance” condizionata al genere viene esaminata tramite lo scatter plot condizionato al fine di ottenere informazioni rilevanti e capire l’interazione. In generale, tutti i punti tendenzialmente tendono a salire da sinistra a destra, il che significa che le persone con un migliore equilibrio tra lavoro e vita tendono ad essere più soddisfatte nella vita. Le osservazioni per gli uomini sono leggermente spostate verso l’alto rispetto a quelle delle donne, con valori di soddisfazione della vita in media più alti. Inoltre, i punti delle femmine sono più dispersi, il che potrebbe indicare una maggiore variabilità nelle risposte ottenute dai diversi stati. Due valori potrebbero essere considerati pressochè anomali, per entrambi work-life balance e life satisfaction esprimono quotazioni basi sia nella distribuzione maschile che femminile, per questo sono stati approfonditi: si sta facendo riferimento a Turchia e Messico.



7.3. RELAZIONE: Air pollution, gender e Years in education

Si è deciso, in seguito, di studiare la relazione presente tra gli anni di educazione per genere condizionato all'indice relativo l'inquinamento dell'aria dei diversi paesi. Tenendo conto che a seconda dell'indice di inquinamento dell'aria, si sono distinte due classi: gruppo 0 a cui appartengono gli stati che hanno indice di inquinamento minore della mediana, gli altri al gruppo 1. Visualizzando la relazione tramite i boxplot, si osserva che, in generale, le donne sembrano avere più anni di istruzione in mediana rispetto agli uomini. Tuttavia, è presente una discrepanza interessante: nei paesi con Air Pollution = 1, il numero di anni di istruzione è inferiore rispetto a quelli con Air Pollution = 0. Questo potrebbe essere dovuto a diverse ragioni. Una spiegazione potrebbe essere che l'inquinamento dell'aria è associato a condizioni socioeconomiche o geografiche che influenzano l'accesso all'istruzione. Ad esempio, persone in aree altamente inquinate potrebbero affrontare sfide economiche o sociali che limitano la loro opportunità di istruzione. Allo stesso tempo, potrebbe esserci una distribuzione geografica non uniforme dei dati, con le regioni altamente inquinate che influenzano i risultati.



8. Modelli di regressione

Il dataset potrebbe essere adatto per condurre un'analisi di regressione lineare, a seconda delle questioni di ricerca o delle ipotesi che si intendono esaminare. La regressione lineare multipla è un modello statistico che consente di esplorare le relazioni tra più variabili indipendenti, o variabili di *input*, e una variabile dipendente, comunemente chiamata *risposta*.

Per implementare un modello di regressione multipla è necessario che siano presenti due o più predittori e che ciascuna osservazione in presenza di queste variabili abbiano un dato registrato. Nel dataset better life index le variabili *Life expectancy* e *Life satisfaction* potrebbero essere utilizzate come variabile dipendente mentre le altre come covariate.

Poichè lo scopo è **prevedere l'aspettativa di vita o soddisfazione di vita in funzione degli altri indici del dataset** si considera l'utilizzo del modello di regressione. Esistono diversi tipi di modelli di regressione: lineare semplice, lineare multiplo, etc.. . Poichè il dataset permette di lavorare con molte variabili, è possibile confrontare il modello di regressione lineare semplice con quello multiplo.

8.1. Split in train set e test set

Lo split dei dati in un set di addestramento e un set di test si rende necessario per valutare l'accuratezza del modello. Il *train set* è quel set di dati utilizzato per l'addestramento del modello mentre il *test set* viene utilizzato per prevedere la variabile dipendente e tramite metriche statistiche valutare l'accuratezza del modello.

L'obiettivo principale dei modelli è ottenere delle stime significative e valide che contribuiscano alla comprensione delle relazioni causa-effetto specifiche nella popolazione. La perseguibilità

dell'obiettivo tiene conto dell'individuazione di risultati non necessariamente generalizzabili ma piuttosto validi per un sottoinsieme ristretto della popolazione. Per tanto, è stata presa la decisione di analizzare tutti i paesi a eccezione di quelli non presenti nell'aria OECD; questi ultimi nelle analisi precedenti hanno riportato problemi di non compatibilità e anomalia con il resto dei paesi dell'area altamente sviluppati, faccio riferimento a: South Africa, Brasile e Russia.

8.2. Modello di regressione lineare semplice

Il modello di regressione lineare semplice è specificato dalla relazione:

$$Y = \beta_0 + \beta_1 x_1 + \epsilon$$

Dove Y è la variabile dipendente, X_1 la covariata, ϵ l'errore residuo, β_0 e β_1 sono i parametri di intercetta e coefficiente angolare della retta di regressione. L'obiettivo geometrico è quello di esplicitare la variabilità delle ordinate tramite quella delle ascisse, ovvero trovare la retta che passi tra i punti nel modo migliore e sintetizzando l'andamento complessivo medio che si percepisce osservando i dati. Poichè si cerca di spiegare Y tramite X_1 la qualità statistica della retta dovrà essere giudicata sulla base della sua capacità di adattarsi ai valori delle Y osservate. Pertanto, una buona retta di regressione sarà quella che minimizzerà la distanza verticale tra i valori osservati di y e quello teorici $y(\beta_0, \beta_1)$. Per cui i coefficienti vengono stimati minimizzando la somma dei quadrati degli errori residui.

Implementazione

I modelli di regressione implementati per il caso studio sono stati scelti tenendo conto della rilevanza degli aspetti statistici nel contesto del caso di studio. Come per la serie storica precedente, si ispeziona il variare dell'aspettativa di vita in relazione alla percentuale di famiglie che vivono senza l'accesso ai servizi sanitari primari. I risultati sono espressi nella tabella del summary. Per la valutazione dell'importanza e della significatività dei coefficienti stimati del modello viene effettuato il test di Wald sui singoli parametri, confrontando l'ipotesi H_0 : il coefficiente associato a una variabile sia statisticamente diverso da 0, contro l'alternativa H_1 : il coefficiente è statisticamente diverso da 0. Il valore del test viene quindi espresso dal p-value, fissata una soglia critica di livello $\alpha = 0.05$, risultano essere statisticamente significative sia il valore di intercetta che la variabile relativa a dwelling.

L'output del modello e i valori dei coefficienti esprimono l'effetto delle variabili sul valore di aspettativa di vita. Per valori positivi dei coefficienti, l'aspettativa di vita aumenta; per valori negativi dei coefficienti l'aspettativa di vita diminuisce. Il valore dell'intercetta è uguale a 81.89, indicando che se tutte le covariate fossero uguali a zero l'aspettativa di vita della popolazione mediamente è 81.89 anni. L'interpretazione dei coefficienti si sviluppa in questo modo: per ogni incremento unitario di percentuali di famiglie che non hanno a disposizione l'accesso ai servizi sanitari primari, l'aspettativa di vita diminuisce di 0.63 anni. Per quanto riguarda invece il test sui parametri contemporaneamente, su tutti i parametri del modello, si fa affidamento alla statistica F in basso e il corrispondente p-value. Il valore di F è 140.5, il suo p-value è molto basso, per cui prossimo allo 0, rifiutiamo H_0 secondo la quale tutti i parametri sono uguali a 0, quindi accettiamo l'ipotesi H_1 che esiste almeno un parametro b_j diverso da 0.


```

Call:
lm(formula = 'Life expectancy' ~ 'Dwellings without basic facilities',
    data = dat_train)

Residuals:
    Min       1Q   Median       3Q      Max
-4.3575 -0.3323  0.2439  1.1192  2.8424

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   81.4557     0.4211 193.426 < 2e-16 ***
'Dwellings without basic facilities' -0.4996     0.1424  -3.508  0.00149 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.9 on 29 degrees of freedom
Multiple R-squared:  0.2979,    Adjusted R-squared:  0.2737
F-statistic: 12.31 on 1 and 29 DF,  p-value: 0.001492

```

Previsioni

Avendo stimato e addestrato il modello di regressione è possibile quindi fare previsione sul test set (non utilizzato per l'addestramento).

R-quadro

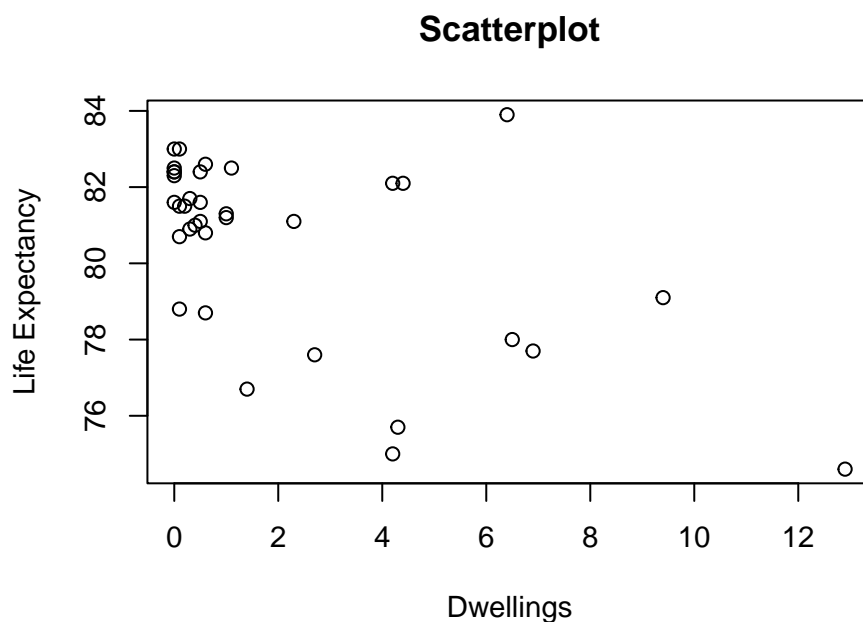
L' R^2 è un coefficiente di determinazione: misura la bontà di adattamento del modello ai dati. Se il valore dell' R^2 è prossimo all' 1, significa che i regressori predicono bene il valore della variabile dipendente nel campione; mentre se è uguale a 0, significa che non lo fanno.

Ad ogni modo il coefficiente non è in grado di stabilire la significatività delle variabili o se è stato scelto il gruppo di regressori più appropriato. L'indice dell' R^2 corretto è prossimo allo 0.29, per cui il modello così costruito esprime il 29% della variabilità dei dati.

```
[1] 0.2979438
```

Visualizzazione legame variabili

Tuttavia, confrontando praticamente non sembra che tra le due variabili sia presente un legame lineare. Per cui, il modello di regressione lineare non è in grado di approssimare bene questa relazione.



8.3. Modello di regressione lineare multiplo

Il modello di regressione lineare multipla è una tecnica statistica utilizzata per prevedere una variabile dipendente, conosciuta anche come variabile risposta o target, sulla base della sua relazione con più variabili indipendenti, chiamate anche variabili esplicative o predictor. Questo modello presuppone che esista una relazione lineare tra la variabile dipendente e le variabili indipendenti.

Implementazione

La regressione lineare multipla è comunemente impiegata quando ci sono più di una variabile indipendente che potrebbe influenzare la variabile dipendente. Ad esempio, si potrebbe utilizzare il modello di regressione lineare multipla per prevedere l'aspettativa di vita (variabile dipendente) in base a variabili indipendenti come non solo dwellings, numero di vani per persona, reddito netto disponibile, ricchezza finanziaria, qualità del supporto di rete, qualità dell'aria, qualità dell'acqua e il tasso di omicidi. Questo modello consente di comprendere come ciascuna di queste variabili indipendenti contribuisce alla variazione nella variabile dipendente e di stimare il loro impatto combinato sul risultato.

La tabella delle stime mostra come solo la soddisfazione di e il tasso di omicidi ha un p-value statisticamente significativo fissato alpha prossimo a 0.05. E' tuttavia necessario sottolineare che l'importanza previsiva di una variabile non deve essere necessariamente valutata rispetto alla singola ma anche rispetto alla remota possibilità che da sola può non avere potere previsivo ma questo potrebbe aumentare se congiunta con un'altra variabile. Di solito si procede valutando il p-value e costruendo un modello con le sole variabili significative, ed è quello che viene fatta alla prossima implementazione, ma esistono dei metodi di feature selection che permettono di scegliere le variabili che hanno effettivamente effetto sulla variabile dipendente.

Verrà analizzata nella terza implementazione.

Call:

```
lm(formula = model1, data = dat_train)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.1163	-0.3620	-0.1172	0.3217	1.9862

Coefficients:

	Estimate	Std. Error
(Intercept)	8.434e+01	2.739e+01
'Dwellings without basic facilities'	-3.282e-01	2.938e-01
'Housing expenditure'	-3.512e-01	2.650e-01
'Rooms per person'	-1.769e+00	2.382e+00
'Household net adjusted disposable income'	7.296e-06	1.697e-04
'Household net financial wealth'	1.103e-05	2.162e-05
'Labour market insecurity'	-2.406e-02	2.041e-01
'Employment rate'	-7.501e-03	1.747e-01
'Long-term unemployment rate'	4.860e-01	2.831e-01
'Personal earnings'	-1.310e-05	9.064e-05
'Quality of support network'	-8.816e-02	1.741e-01
'Educational attainment'	-7.180e-02	5.220e-02
'Student skills'	4.977e-02	2.886e-02
'Years in education'	-4.054e-01	4.052e-01
'Air pollution'	-1.876e-01	1.779e-01
'Water quality'	-7.213e-02	1.202e-01
'Stakeholder engagement for developing regulations'	3.402e-01	7.988e-01
'Voter turnout'	-1.469e-02	3.799e-02
'Self-reported health'	-8.371e-02	6.980e-02
'Life satisfaction'	4.674e+00	1.784e+00
'Feeling safe walking alone at night'	-4.527e-02	6.134e-02
'Homicide rate'	-7.035e-01	2.865e-01
'Employees working very long hours'	7.897e-02	1.048e-01
'Time devoted to leisure and personal care'	-6.782e-01	7.591e-01
	t value	Pr(> t)
(Intercept)	3.079	0.0178 *
'Dwellings without basic facilities'	-1.117	0.3009
'Housing expenditure'	-1.325	0.2267
'Rooms per person'	-0.743	0.4819
'Household net adjusted disposable income'	0.043	0.9669
'Household net financial wealth'	0.510	0.6258
'Labour market insecurity'	-0.118	0.9095
'Employment rate'	-0.043	0.9670
'Long-term unemployment rate'	1.717	0.1297
'Personal earnings'	-0.145	0.8892

'Quality of support network'	-0.506	0.6282
'Educational attainment'	-1.376	0.2113
'Student skills'	1.725	0.1282
'Years in education'	-1.000	0.3504
'Air pollution'	-1.055	0.3267
'Water quality'	-0.600	0.5675
'Stakeholder engagement for developing regulations'	0.426	0.6830
'Voter turnout'	-0.387	0.7105
'Self-reported health'	-1.199	0.2695
'Life satisfaction'	2.619	0.0344 *
'Feeling safe walking alone at night'	-0.738	0.4846
'Homicide rate'	-2.455	0.0438 *
'Employees working very long hours'	0.754	0.4757
'Time devoted to leisure and personal care'	-0.893	0.4013

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.335 on 7 degrees of freedom

Multiple R-squared: 0.9164, Adjusted R-squared: 0.6418

F-statistic: 3.337 on 23 and 7 DF, p-value: 0.05319

Implementazione 2

Poichè molte delle variabili risultano non statisticamente significative, si è pensato di creare un modello con le sole significative.

Il modello così implementato mostra valore negativo per il tasso di omicidi e positivo per la soddisfazione di vita. In maniera logica, qualora si vivesse in uno stato in cui il tasso di omicidi è elevato l'aspettativa di vita diminuisce: in particolar modo, per ogni incremento unitario della percentuale di omicidi l'aspettativa di vita diminuisce di 0.38 anni. La soddisfazione della vita personale, invece, contribuisce in maniera positiva all'aumento di vita: per ogni incremento unitario di soddisfazione l'aspettativa di vita aumenta di 0.13 anni. Il valore della statistica R-quadro è diminuita.

Call:

```
lm(formula = 'Life expectancy' ~ 'Life satisfaction' + 'Homicide rate',
    data = dat_train)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.6053	-0.7803	0.0222	0.8330	2.5552

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	72.66448	2.63637	27.562	< 2e-16 ***
'Life satisfaction'	1.30259	0.39597	3.290	0.002711 **

```

'Homicide rate'      -0.38125      0.09339   -4.082 0.000337 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.629 on 28 degrees of freedom
Multiple R-squared:  0.5018,    Adjusted R-squared:  0.4662
F-statistic:  14.1 on 2 and 28 DF,  p-value: 5.799e-05

```

Implementazione 3

Si utilizza il metodo della STEPWISE come variable selection. La stepwise è identificata come un metodo per la selezione del modello. Nei casi in cui un dataset contiene un numero elevato di variabili potrebbe essere utile implementare una procedura che permetta di selezionare le variabili rilevanti per il problema oggetto di studio. La stepwise sviluppa una sequenza di modelli di regressione in cui a ogni step viene eliminata o aggiunta una variabile, tra quelle considerate. Il criterio di aggiunta o eliminazione si basa in termini di somma degli errori delle riduzioni al quadrato, coefficiente di correlazione, statistica t, statistica F. Il risultato finale identifica un unico modello di regressione come il “migliore”. Esistono tre tipi di possibili selection: Forward Stepwise Selection, Backward Elimination e Stepwise selection. La differenza principale tra questi approcci sta nel modo in cui le variabili vengono aggiunte o rimosse dal modello. Forward e backward sono più semplici, poiché il primo inizia la selezione partendo dal modello con la sola intercetta; mentre il secondo inizia la selezione delle variabili partendo dal modello completo e andando a ritroso. Tuttavia, possono produrre modelli subottimali se le variabili significative sono fortemente correlate tra loro. La stepwise forward selection fonda la selezione del modello migliore sulla base degli indici per la bontà di adattamento e criteri di informazione. 1) Individua un modello di regressione semplice con la sola intercetta per ciascuna delle potenziali variabili di X , per ciascuna calcola la statistica test per valutare se l'intercetta è statisticamente significativa o meno. La variabile con la statistica t più elevata e il p-value minore rispetto alla soglia stabilità α viene aggiunta al modello; 2) Si assume che la variabile x_i è la variabile scelta allo step 1. La stepwise regression adatta tutti i modelli di regressione con due X , in cui una delle due è x_i . Per ogni modello individuato, la variabile x_k con p-value più piccolo viene selezionata e utilizzata per lo step successivo e analogamente si sviluppano i passi successivi esaminando quale variabile x è la migliore candidata valutando se c'è una variabile da dover eliminare. 3) Quando aggiungere o eliminare una variabile non cambia il risultato, la ricerca si conclude. Secondo questa logica è stato quindi individuato il modello migliore con le migliori variabili che hanno statisticamente un impatto sulla variabile dipendente.

Il modello così individuato permette di tener conto di un numero maggiore di variabili determinate come statisticamente significative: guadagno personale, student skills, soddisfazione della vita, lavoro a lungo termine hanno un impatto positivo sulla variabile di aspettativa di vita mentre le variabili del tasso di omicidi, educational attainment, qualità di supporto di rete sociale, tempo dedicato alla persona hanno un impatto negativo.

Call:

```
lm(formula = Step5.reg$formula, data = dat_train, x = TRUE)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.92463	-0.63430	-0.05198	0.66721	1.76387

Coefficients:

	Estimate	Std. Error	t value
(Intercept)	7.500e+01	7.206e+00	10.408
'Personal earnings'	4.206e-05	2.631e-05	1.598
'Homicide rate'	-4.645e-01	1.018e-01	-4.561
'Educational attainment'	-8.760e-02	1.959e-02	-4.472
'Student skills'	5.064e-02	1.295e-02	3.910
'Life satisfaction'	2.506e+00	5.871e-01	4.269
'Long-term unemployment rate'	2.820e-01	8.292e-02	3.401
'Quality of support network'	-1.829e-01	6.129e-02	-2.985
'Time devoted to leisure and personal care'	-6.471e-01	3.563e-01	-1.816
'Years in education'	-2.628e-01	1.825e-01	-1.441
	Pr(> t)		
(Intercept)	9.57e-10	***	
'Personal earnings'	0.124929		
'Homicide rate'	0.000170	***	
'Educational attainment'	0.000210	***	
'Student skills'	0.000805	***	
'Life satisfaction'	0.000341	***	
'Long-term unemployment rate'	0.002690	**	
'Quality of support network'	0.007063	**	
'Time devoted to leisure and personal care'	0.083622	.	
'Years in education'	0.164469		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.058 on 21 degrees of freedom

Multiple R-squared: 0.8425, Adjusted R-squared: 0.775

F-statistic: 12.48 on 9 and 21 DF, p-value: 1.364e-06

Previsioni

Dopo aver stimato il modello di regressione si possono prevedere i valori della variabile dipendente su un set di dati che non è stato usato per addestrare il modello.

Confronto fra modelli

Per effettuare un confronto accurato, sono stati presi in considerazione diversi criteri, tra cui l'adeguatezza del modello alle specifiche della popolazione di studio, la precisione delle stime dei parametri, la bontà di adattamento e la capacità predittiva tramite l'utilizzo di indici.

Si può tener conto di :

Errore quadratico medio (RMSE) L' RMSE (errore quadratico medio) è definito come :

$$RMSE = \sqrt{(\frac{1}{n} \sum_{i=1}^N (y_i - \hat{y})^2)}$$

dove: y_i è il valore reale di y ; $\hat{y} = x_i\hat{\beta}$ è il valore previsto dal modello; n è il numero di osservazioni del test set.

E' una misura di errore assoluta in cui le deviazioni vengono elevate al quadrato per evitare che valori positivi e negativi possano annullarsi l'uno con l'altro. E' chiaramente preferibile il modello che presenta indice rmse più basso.

Errore assoluto medio (MAE) Il mae è invece definito come :

$$MAE = (\frac{1}{n} \sum_{i=1}^N |y_i - \hat{y}|)$$

Un valore piccolo del MAE per il modello è sinonimo di previsioni accurate mentre un valore grande del MAE è espressione di previsioni del modello meno accurate.

AIC E BIC I due criteri di informazione più comunemente utilizzati per i modelli di regressione sono AIC e BIC. Sia L la verosimiglianza del modello; k il numero di parametri del modello; n il numero di osservazioni nel campione, i criteri si definiscono come: Il criterio di informazione di **Akaike (AIC)**: tiene conto del grado di adattamento del modello ai dati e del numero di parametri del modello stesso. L'indice è calcolato come:

$$AIC = -2\log(L) + 2k$$

L'**AIC** assegna un punteggio ai diversi modelli e il modello con il punteggio più basso è considerato il modello migliore. Il criterio di **informazione bayesiano (BIC)**: questo criterio è simile all'AIC, ma penalizza in modo più severo i modelli che hanno un numero elevato di parametri. L'indice è calcolato come: $BIC = -2\log(L) + k\log(n)$.

Il **BIC** assegna un punteggio ai diversi modelli e il modello con il punteggio più basso è considerato il modello migliore. Per entrambe le formule, il primo termine ($-2\log(L)$) è una misura della devianza del modello, ovvero una misura della differenza tra i valori osservati e quelli predetti dal modello. Il secondo termine è una penalizzazione per il numero di parametri del modello, in modo che modelli più complessi vengano penalizzati rispetto a quelli più semplici. Entrambi questi criteri possono essere utilizzati per confrontare diversi modelli di regressione e selezionare quello migliore. In generale, il modello con il valore più basso di AIC o BIC è considerato il migliore tra i modelli candidati.

La valutazione dell'aspettativa di vita rispetto alle singole feature di tassi di omicidi e aspettativa di vita nel modello di regressione ha portato a una diminuzione del chi-quadrato nel set di dati ma risultati migliori in termini di previsioni, indicando che le variabili hanno un valore predittivo statisticamente significativo. Nel complesso, queste le prestazioni dei modelli:

- Il modello contenente solo la variabili relativa alla percentuale di famiglie degli stati che non hanno accesso ai servizi sanitari primari mostra un livello basso di R-quadro,

prossimo al 44% e mostra valori elevati per AIC e BIC, oltre che previsioni piuttosto pessime sul test set (RMSE E MAE);

- Il modello completo con tutte le variabili mostra un decremento dell'adattamento ai dati con R-quadro basso oltre che BIC pressochè identico al modello precedente penalizzando appunto il fatto che siano aumentate il numero di features ma non sono effettivamente in grado di spiegare l'andamento dei dati, infatti l'AIC sensibile a un incremento artificiale al crescere del numero di variabili è minore. Tuttavia, mostra valori di previsione nettamente inferiori rispetto al modello con la sola variabile;
- Il modello aggiornato che contiene le sole variabili di homicide rate e life satisfaction (poichè le uniche variabili significative nel modello precedente) registra un adattamento ai dati da parte del modello piuttosto bassa per l'R-quadro (91%) ma AIC e BIC migliori.
- In termini previsivi il migliore è quello con le variabili scelte dalla stepwise.

Modello	R-quadro	MSE	MAE	AIC	BIC
semplice	44%	9.656	2.015	142.86	147.26
completo	97%	3.68	1.699	109.73	145.58
aggiornato	91%	6.21	2.03	123.07	128.81
con stepwise	98%	4.92	1.65	101.37	117.15

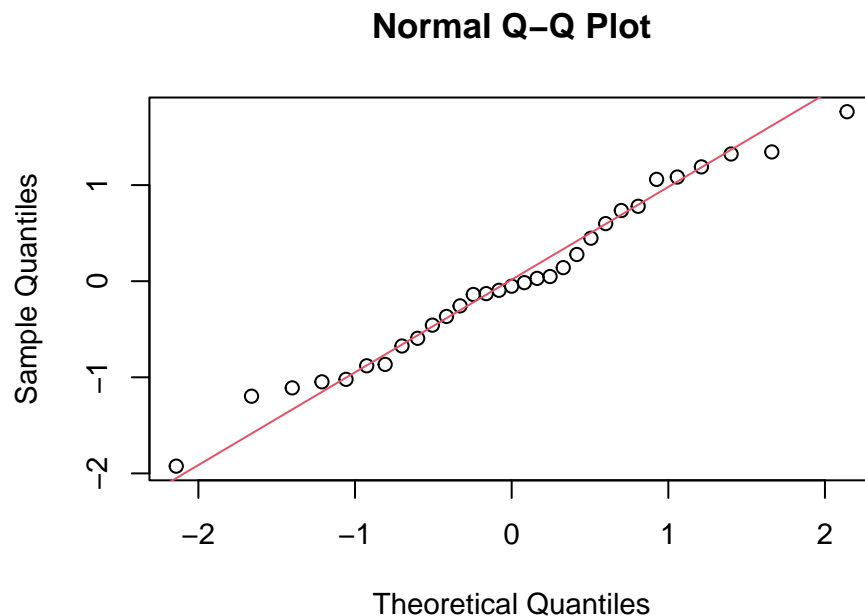
8.4. Analisi dei residui

Un'ulteriore approfondimento sull'accuratezza del modello può essere svolto attraverso lo scatter plot dei **valori osservati e quelli stimati dal modello**.

I residui sono definiti come la differenza puntuale tra i veri valori della y e quelli previsti dal modello. I residui sono così definiti:

$$E_i = y_i - \hat{y}_i = y_i - (\alpha + \alpha x_i)$$

L'analisi dei residui consiste nel verificare se i residui, seguono una distribuzione normale e se sono distribuiti in modo casuale. Ciò è importante perché una distribuzione normale e casuale dei residui indica che il modello è adeguato alla relazione tra le variabili dipendenti e indipendenti. Per eseguire un'analisi dei residui, è possibile utilizzare alcuni strumenti come un grafico di istogramma dei residui, un grafico Q-Q plot e un grafico di dispersione dei residui sui valori previsti. Nella realtà, spesso i dati non seguono perfettamente una distribuzione normale, ma è importante valutare se l'adesione al modello non sia troppo lontana dai valori attesi. Il grafico Q-Q plot mostra che i residui seguono una distribuzione approssimativamente normale, in caso contrario i punti non si troverebbero sulla linea di regressione (escluse alcune osservazioni che discostano dalla linea, giustificati dal fatto che si sta aggregando per paesi che hanno quadri culturali, economici e sociali diversi). La maggior parte dei punti non sono direttamente sulla retta ma intorno per cui questo indica che il modello può essere sicuramente migliorato. Ad ogni modo sul training set ha un ottimo fitting.



La verifica della Normalità è possibile effettuarla utilizzando alcuni test di Normalità. Il Test di **Shapiro-Wilk** : Questo test verifica se i dati o i residui seguono una distribuzione normale. L'ipotesi nulla (H_0) del test è che i dati seguono una distribuzione normale. Il p-value, prossimo allo 0.85, ottenuto dal test è maggiore del livello di significatività specifico (0,05), allora non si può rifiutare l'ipotesi nulla e si può concludere che i residui possono essere considerati approssimativamente normalmente distribuiti.

Shapiro-Wilk normality test

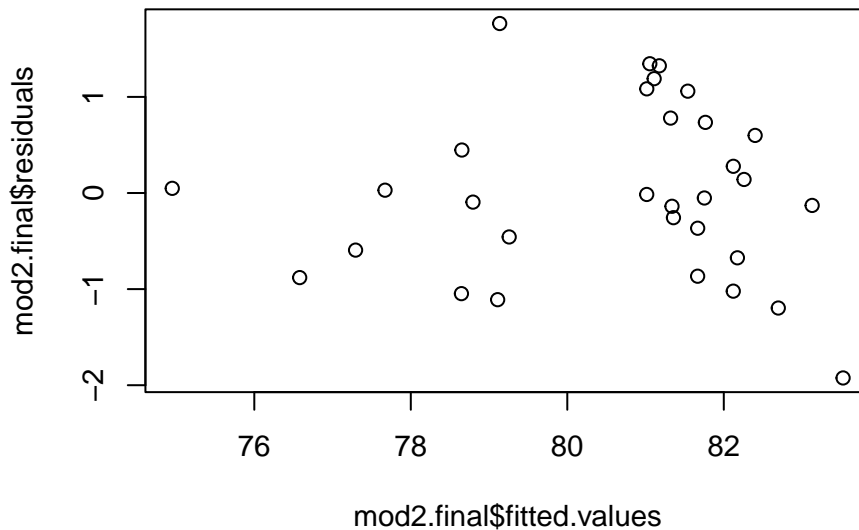
```
data: mod2.final$residuals  
W = 0.98155, p-value = 0.8542
```

Test di **Jarque-Bera** : Questo test è specifico per verificare la normalità dei residui. Calcola la statistica di Jarque-Bera, che è basata sulla curtosi (misura della forma della distribuzione) e sull'asimmetria (misura della simmetria della distribuzione) dei residui. L'ipotesi nulla (H_0) di questo test è che i residui seguono una distribuzione normale. Il p-value, prossimo allo 0.78, ottenuto dal test è superiore a un determinato livello di significatività (0.05), l'ipotesi nulla viene mantenuta, indicando che i residui possono essere considerati normalmente distribuiti.

Jarque Bera Test

```
data: mod2.final$residuals  
X-squared = 0.48811, df = 2, p-value = 0.7834
```

Infine il grafico di dispersione dei residui sui valori previsti mostra una distribuzione casuale, ovvero senza alcuno schema visibile, per cui espressione di non dipendenza.



Per la valutazione delle proprietà dei residui:

1. $\sum(e) = 0$: Questa espressione rappresenta la somma dei residui e_i su tutte le osservazioni i nel modello di regressione. Un requisito importante per un buon modello di regressione è che la somma dei residui sia il più vicina possibile a zero così che il modello dovrebbe essere in grado di prevedere correttamente il valore medio della variabile dipendente. Il risultato di somma dei residui pari a zero indica che, in media, il modello non sovrastima né sottostima le previsioni.
2. $\sum(x * e) = 0$: Questa espressione rappresenta la somma del prodotto tra le variabili indipendenti x_i e i residui e_i corrispondenti su tutte le osservazioni i . Questa condizione indica che la covarianza tra le variabili indipendenti e i residui è zero. Ciò implica che le variabili indipendenti stanno spiegando adeguatamente la variazione nella variabile dipendente. Il risultato ottenuto non è quello atteso.
3. $\sum(y * e) = 0$: Questa espressione rappresenta la somma del prodotto tra la variabile dipendente y_i e i residui e_i corrispondenti su tutte le osservazioni i . Questa condizione indica che la covarianza tra la variabile dipendente e i residui è zero. Questo suggerisce che il modello di regressione stia catturando in modo adeguato la relazione tra le variabili indipendenti e la variabile dipendente. Il risultato ottenuto non è quello atteso.

[1] 7.494005e-16

[1] -49247.07

[1] 227654.6

9. Clustering

Per clustering si intende un'insieme di tecniche statistiche volte alla ricerca di sottogruppi (*clusters*) omogenei all'interno di una popolazione di riferimento. Fa parte dei task di learning *unsupervised* e si caratterizza per la ricerca di pattern nascosti, di struttura significative all'interno di dati unlabeled, ovvero di dataset per i quali non si hanno informazioni a disposizione circa la reale associazione esistente tra le istanze in essi contenuti ed in gruppi. Nei metodi di clustering l'obiettivo ultimo si configura nell'utilizzo della distribuzione osservata delle variabili al fine di ricercare gruppi all'interno del data set in input. Tale processo di individuazione viene eseguito facendo in modo che le osservazioni appartenenti allo stesso cluster siano quanto più possibile simili l'una all'altra (*omogeneità within*), mentre le istanze associate a gruppi diversi siano abbastanza dissimili tra di loro (*eterogeneità between*).

Nell'analisi di cluster molto spesso ci si pone dinanzi al problema della *standardizzazione delle variabili*. La standardizzazione delle variabili è essenziale in quanto permette di eliminare le differenze di scala, in quanto le variabili in un set di dati possono avere unità di misura diverse o range di valori molto diversi. Senza standardizzazione, le differenze di scala possono influire in modo significativo sui risultati del clustering, poiché le variabili con valori più elevati avranno un peso maggiore nella distanza tra le osservazioni. La standardizzazione rende tutte le variabili comparabili, garantendo che ciascuna abbia la stessa importanza nel processo di clustering. Inoltre migliora le prestazioni degli algoritmi basati sulla distanza tra le osservazioni, senza standardizzazione, le variabili con scale diverse influiranno in modo sbilanciato sulla distanza. La standardizzazione garantisce che le variabili contribuiscano in modo equo alla misura della distanza, migliorando le prestazioni degli algoritmi di clustering. Sui risultati del clustering influisce, in maniera significativa, la scelta del numero di variabili da utilizzare oltre che quali variabili effettivamente sono utili a scindere i diversi gruppi, per farlo si utilizza la **PCA**.

Il metodi di cluster si distinguono in cluster **gerarchico** e cluster **non gerarchico**.

Clustering Gerarchico Il clustering gerarchico è un metodo di raggruppamento dei dati che organizza le osservazioni in una struttura gerarchica a forma di albero, comunemente rappresentata come un dendrogramma. Questo tipo di clustering è intrinsecamente gerarchico, il che significa che i cluster possono essere ulteriormente suddivisi in sotto-cluster, creando una struttura ad albero. Nel clustering gerarchico, ci sono principalmente due approcci:

Agglomerativo: l'approccio inizia considerando ogni punto dati come un cluster separato e successivamente unisce iterativamente i cluster più vicini fino a ottenere un unico cluster contenente tutti i punti. Questo crea una struttura gerarchica dalla base verso l'alto.

Divisivo: Al contrario, l'approccio divisivo inizia con tutti i punti dati in un unico grande cluster e quindi lo suddivide iterativamente in cluster più piccoli. È come "dividere" il cluster in modo progressivo. Questo crea una struttura gerarchica dall'alto verso il basso.

Clustering Non Gerarchico (o Partizionamento): Il clustering non gerarchico, non crea una struttura gerarchica a forma di albero. Invece, assegna direttamente ogni punto

dati a uno dei cluster, senza creare sottoclassi. Questo tipo di clustering è meno orientato alla gerarchia ed è più adatto quando si desidera suddividere i dati in gruppi distinti senza ulteriori suddivisioni interne.

Un esempio comune di clustering non gerarchico è il K-Means, in cui si specifica un numero prefissato di cluster (K) e l'algoritmo assegna ogni osservazione a uno dei K cluster basandosi sulla loro somiglianza.

Si analizzeranno quindi:

- Algoritmi di **clustering Gerarchico** ;
- Algoritmi di **clustering Non-Gerarchici: K-Means**.

9.1. Clustering gerarchico

Il clustering gerarchico produce raggruppamenti di item seguendo un processo gerarchico: parte da tutti gli oggetti separati in gruppi individuali e quindi procede iterativamente con l'aggregazione di coppie di gruppi “che si somigliano di più”, arrivando al termine a produrre un unico gruppo contenente tutti gli item. Il punto di partenza di ciascun metodo di Clustering Gerarchico (HC) è il calcolo della dissimilarità di ciascun individuo rispetto a tutti gli altri. I passaggi sono i seguenti:

1. **Scelta del criterio di omogeneità:** gli elementi all'interno di uno stesso gruppo si somigliano;
2. **Funzione obiettivo:** misura la qualità del clustering in base al livello conseguito di omogeneità/eterogeneità.
3. **Algoritmo:** la procedura numerica per massimizzare la funzione obiettivo.

Similarità e dissimilarità

Si consideri un oggetto i come una riga della matrice dei dati espresso da p features rappresentate nel vettore x_i . Molto spesso gli oggetti sono di tipo euclideo mentre altre volte potrebbero non esserlo (nel caso specifico...).

Le **misure di similarità** permettono di quantificare la somiglianza tra due oggetti o individui. L'obiettivo è assegnare un valore numerico che rappresenti quanto due oggetti siano simili tra loro, con 0 che indica assoluta assenza di somiglianza e 1 che indica massima somiglianza. Queste misure sono spesso utilizzate in contesti in cui la somiglianza è un concetto chiave, come il clustering o il riconoscimento di pattern. Alcuni esempi di misure di similarità includono la similarità coseno, la similarità di Jaccard, e la similarità di Pearson.

Le **metriche di somiglianza** si basano principalmente sul concetto di funzioni distanza tra i vettori delle caratteristiche degli oggetti. Una funzione a valori reali $d(X_i, X_j)$ è considerata una “funzione distanza” se e solo se soddisfa alcune condizioni specifiche. Queste condizioni sono solitamente:

- a. **Positività:** La distanza deve essere sempre non negativa, ovvero $d(X_i, X_j) \geq 0$ per tutti i X_i, X_j .

- b. **Identità dei non-simili:** La distanza tra oggetti identici deve essere zero, ovvero $d(X_i, X_i) = 0$.
- c. **Simmetria:** La distanza tra X_i e X_j deve essere uguale a quella tra X_j e X_i , ovvero $d(X_i, X_j) = d(X_j, X_i)$.
- d. **Disuguaglianza del triangolo:** La distanza tra X_i e X_j attraverso un punto intermedio X_k deve essere inferiore o uguale alla somma delle distanze dirette, ovvero $d(X_i, X_j) \leq d(X_i, X_k) + d(X_k, X_j)$.

Alcune delle funzioni distanza più comuni includono la distanza euclidea, la distanza di Manhattan, la distanza di Mahalanobis e la distanza di Hamming.

Il calcolo delle distanze avviene tramite il comando *dist*. Innanzitutto si crea una partizione selezionando il 70% delle osservazioni del campione totale:

```
# A tibble: 28 x 25
  Stato      Dwellings without basic ~1 'Housing expenditure' 'Rooms per person'
  <chr>                <dbl>                <dbl>                <dbl>
1 Australia            1.1                    20                    2.3
2 Austria              1                    21                    1.6
3 Belgium             2.3                    21                    2.2
4 Canada              0.2                    22                    2.5
5 Finland             0.5                    23                    1.9
6 France              0.5                    21                    1.8
7 Germany             0.1                    20                    1.8
8 Greece              0.5                    24                    1.2
9 Hungary             4.3                    18                    1.2
10 Ireland            0.1                    21                    2.1
# i 18 more rows
# i abbreviated name: 1: 'Dwellings without basic facilities'
# i 21 more variables: 'Household net adjusted disposable income' <dbl>,
# 'Household net financial wealth' <dbl>, 'Labour market insecurity' <dbl>,
# 'Employment rate' <dbl>, 'Long-term unemployment rate' <dbl>,
# 'Personal earnings' <dbl>, 'Quality of support network' <dbl>,
# 'Educational attainment' <dbl>, 'Student skills' <dbl>, ...
```

Si procede con lo scale e standardizzazione dei dati per far sì che si abbia la possibilità di lavorare con dati omogenei dal punto di vista della varianza campionaria.

Segue il calcolo delle principali metriche di distanza

Metriche di distanza

La distanza **euclidea** è fornita di default per il metodo *dist*. Questa in genere è calcolata come, presi due punti a e b :

$$d(a, b) = \sqrt{(b_1 - a_1)^2 + (b_2 - a_2)^2 + (b_n - a_n)^2}$$

La distanza di **Manhattan**, è calcolata come la somma delle differenze assolute tra le coppie di valori tra due punti. Questa distanza è utile quando le dimensioni sono indipendenti e il percorso tra i punti deve seguire una griglia rettangolare. È chiamata “Manhattan” perché

rappresenta la distanza tra due punti su una griglia stradale di una città, dove puoi muoverti solo lungo strade parallele o perpendicolari.

Qui la sua formulazione:

$$d_1(X_i, X_j) = \sum_{k=1}^p |x_{ik} - x_{jk}|$$

La distanza di **Chebyshev** è calcolata come la massima differenza tra i valori corrispondenti delle coppie tra due punti. Questa distanza coinvolge anche una procedura di ordinamento per calcolare la massima differenza. È utile quando si desidera misurare la massima differenza tra le dimensioni dei punti.

Qui la sua formulazione:

$$d_\infty(X_i, X_j) = \max_{k=1,2,\dots,p} |x_{ik} - x_{jk}|$$

La distanza di **Minkowski** è una metrica più generale che include sia la distanza di Manhattan (quando il parametro “p” è uguale a 1) che la distanza euclidea (quando il parametro “p” è uguale a 2) come casi particolari. Il parametro “p” consente di regolare il comportamento della distanza in base alle caratteristiche specifiche dei dati. È utile quando le dimensioni sono importanti e le relazioni tra di esse sono complesse. Quando “p” è diverso da 1 o 2, la distanza di Minkowski può essere utilizzata per adattarsi a situazioni diverse.

Qui la sua formulazione:

$$d_r(X_i, X_j) = \left[\sum_{k=1}^p |x_{ik} - x_{jk}|^r \right]^{1/r}$$

Misure di non omogeneità statistica

Le misure di non omogeneità statistica sono metriche utilizzate nelle applicazioni statistiche. Si utilizza di solito la matrice di varianza e covarianza.

Per mezzo della matrice di covarianza è possibile calcolare la SSM, statistical scatter matrix moltiplicata per (n-1). Sommando la diagonale principale della matrice si ottiene la traccia. (non viene stampata perché è molto grande e occupa spazio)

Questa misura riguarda l'intera popolazione e viene affiancata da due altre misure: una che valuta la differenza all'interno dei gruppi (denominata “within”) e un'altra che valuta la differenza tra i gruppi stessi (denominata “between”). L'obiettivo di queste misure è determinare se il metodo utilizzato per raggruppare individui ha effettivamente creato gruppi con individui simili all'interno di ciascun gruppo, ma diversi tra gruppi diversi.

Metriche di similarità

Una funzione di similarità, o coefficiente di similarità, è una funzione che genera un valore numerico compreso tra 0 e 1 che misura la somiglianza o la differenza tra due punti all'interno di un dataset. Per essere considerata una funzione di similarità, deve soddisfare tre importanti proprietà:

Unitarietà: La funzione di similarità deve restituire il valore massimo (1) quando si confrontano due punti che sono identici. In altre parole, la massima somiglianza deve essere assegnata quando i punti sono gli stessi.

range tra 0 e 1: Il valore restituito dalla funzione di similarità deve essere compreso tra 0 e 1. Questo implica che la somiglianza tra due punti non può essere inferiore a 0 (assoluta assenza di somiglianza) né superiore a 1 (massima somiglianza).

Simmetria: La funzione di similarità deve essere simmetrica. Questo significa che la similarità tra i punti X_1 e X_2 deve essere identica alla similarità tra i punti X_2 e X_1 .

La funzione di similarità viene spesso rappresentata attraverso una matrice S , in cui l'elemento s_{ij} rappresenta la similarità tra il punto i -esimo e il punto j -esimo all'interno del dataset. Inoltre:

1. Ogni distanza è una dissimilarità ma non è vero viceversa;
2. Una dissimilarità non deve necessariamente soddisfare la disuguaglianza triangolare;

Le metriche di similarità vengono spesso utilizzate per validare la partizione. La validazione è una problematica concettuale poichè non c'è una soluzione oggettiva dei dati. Servono principalmente a stabilire quanto due partizioni concordano.

Prese due collezioni A e B , si definisce il **coefficiente Jaccard** (espressione della somiglianza tra due collezioni):

$$s(A, B) = |A \cap B| / |A \cup B|$$

L'indice assume valore compreso tra 0 e 1, 0 indica assenza di somiglianza mentre 1 espressione di massima somiglianza.

Per la definizione del miglior numero di gruppi si utilizza SILHOUETTE WIDTH (spiegato successivamente).

I metodi gerarchici si dividono in due gruppi di algoritmi clustering:

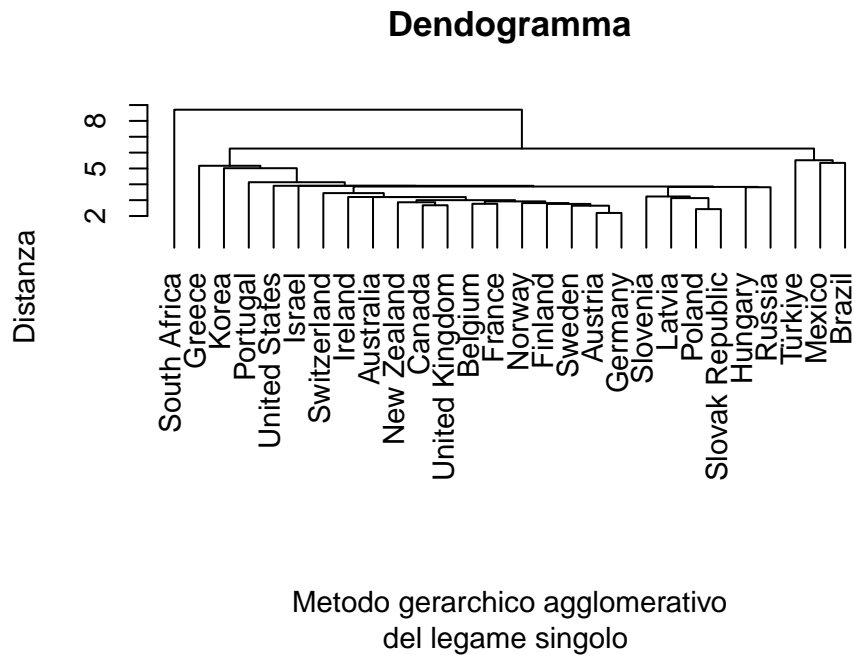
1. AGGLOMERATIVI,

- si parte da una partizione in cui ogni osservazione appartiene a un cluster;
- in ogni step si fondono i 2 cluster più simili in un unico cluster;
- lo step precedente viene ripetuto sino a ottenere un unico cluster.

2. DIVISIVI

- si parte da una partizione contenente tutte le osservazioni;
- in ogni step il cluster più grande viene diviso in due cluster dissimili fra loro;
- lo step precedente viene ripetuto sino a ottenere un unico cluster

Il processo di fusione gerarchico viene sintetizzato in un **grafo** chiamato *Dendogramma*. Nel dendogramma ogni nodo rappresenta un gruppo; alla base ci sono i *leaf nodes* ovvero gli n gruppi formati dagli n oggetti. In testa è presente il *root node*, contenente tutto il dataset. Ogni nodo ha al suo interno due *children nodes*, ovvero i gruppi della fusione dai quali si ottiene il cluster rappresentato dal parent node.



In R per analizzare i metodi di clustering gerarchici si utilizza il comando *hclust*. I passi sono i seguenti:

1. Si inizia con un set di dati rappresentati dalla matrice X , e si decide se standardizzare questi dati. Si sceglie il metodo per calcolare le distanze tra gli elementi e si crea una matrice delle distanze D o una matrice di similarità S .
2. Si trova la coppia di elementi con la distanza più breve nella matrice D e si raggruppano in un cluster. Poi, si ricalcola la matrice delle distanze considerando il nuovo cluster.
3. Si crea una nuova matrice delle distanze con una riga e una colonna in meno, escludendo i dati del cluster appena creato.
4. Si ripete il passo 2 sulla nuova matrice finché non si ottiene una matrice 2×2 , eseguendo $n-1$ iterazioni in totale.
5. Si rappresenta l'intero processo mediante il dendrogramma, che mostra come gli elementi sono stati raggruppati in cluster.

SINGLE LINKAGE

Il termine “**linkage**” si riferisce alla misura di dissimilarità o di similarità tra due gruppi di dati all'interno di un algoritmo di clustering gerarchico. L'obiettivo di utilizzare diverse misure di linkage è di determinare come i dati dovrebbero essere aggregati in cluster in base alle relazioni di dissimilarità o similarità tra di essi.

Esistono diverse misure di linkage, tra cui:

1. **Single Linkage:** Questo metodo calcola la distanza tra i punti più vicini dei due cluster. In altre parole, misura la dissimilarità tra i due punti più vicini, uno da ciascun cluster. Questo può portare a cluster molto allungati e vulnerabili a errori dovuti a punti anomali.
2. **Complete Linkage:** In questo caso, la misura di linkage si basa sulla distanza tra i punti più lontani dei due cluster. Ciò significa che la dissimilarità è calcolata tra i punti più distanti nei due cluster. Questo metodo tende a creare cluster più compatti, ma può anche soffrire di effetto di “crowding.”
3. **Average Linkage:** Questo metodo calcola la media delle distanze tra tutti i punti nei due cluster. Questo approccio può essere considerato una via di mezzo tra il single e complete linkage e tende a produrre cluster di dimensioni più uniformi.

Cambiando il tipo di linkage utilizzato, si otterranno diversi risultati nella struttura gerarchica dei cluster. La scelta del linkage dipenderà dal tipo di dati e dall’obiettivo del clustering. Ad esempio, se si desidera identificare cluster compatti e ben separati, si potrebbe optare per il complete linkage, mentre se si desidera sensibilità alle differenze sottili tra i dati, si potrebbe scegliere l’average linkage.

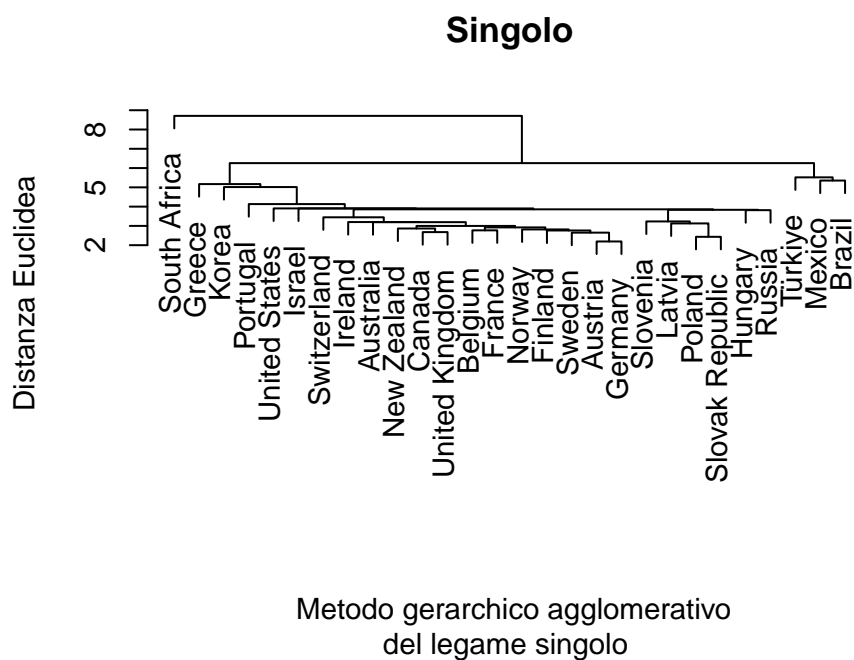
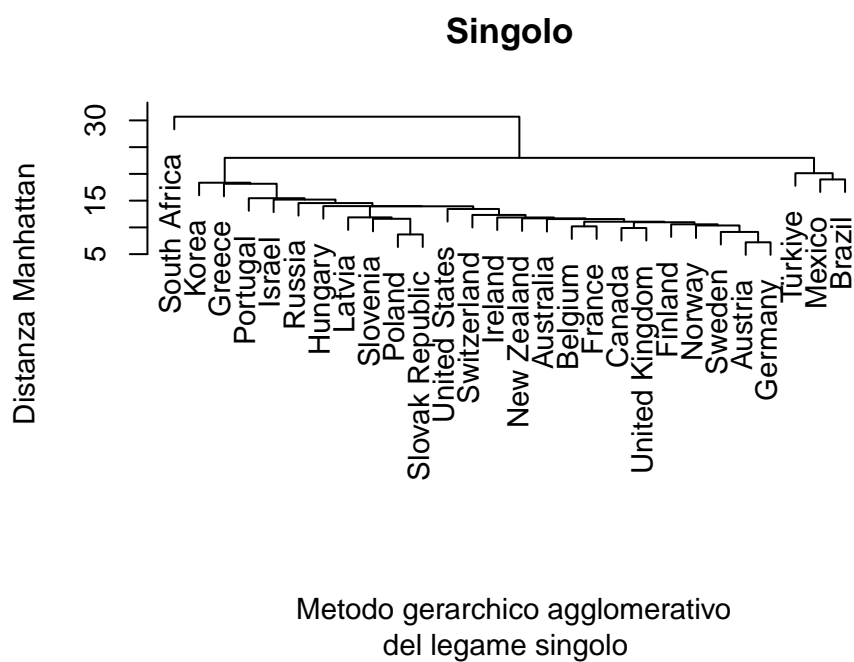
Ci si concentra in primo luogo sul single linkage. Questp è definito come, dati due gruppi A e B , la minima dissimilarità tra due oggetti appartenenti ai due gruppi:

$$d_S L(A, B) = \min_{a \in A, b \in B} d(a, b)$$

Si procede confrontando distanza euclidea e distanza di manhattan e con la visualizzazione del plot.

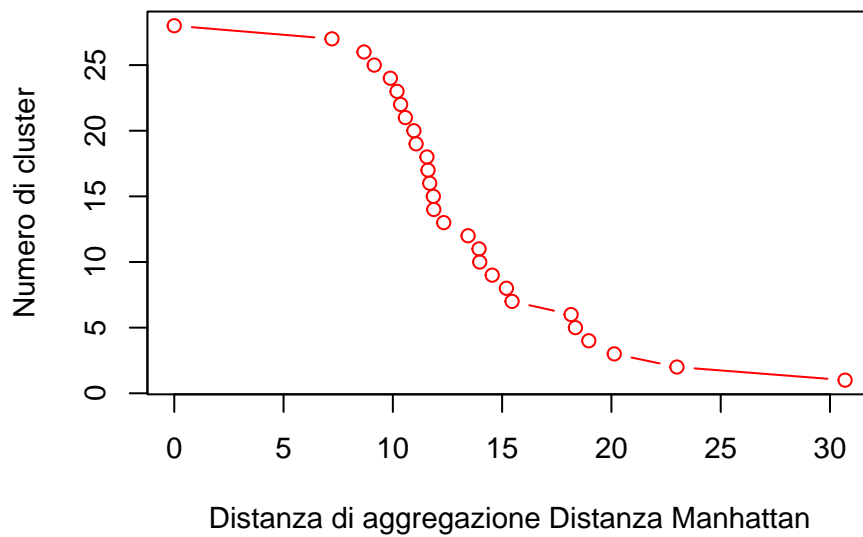
Nel primo caso, ho eseguito un’analisi di clustering gerarchico utilizzando la distanza di Manhattan. Questo metodo misura la somma delle differenze assolute tra le variabili tra i punti del dataset. Il risultato è un dendrogramma che mostra come gli stati sono stati raggruppati in base alle loro somiglianze. Nel grafico risultante, si osserva che alcuni stati con valori di indici di vita simili si sono uniti in cluster, mentre altri rimangono isolati.

Nel secondo caso, ho eseguito un’analisi simile utilizzando la distanza euclidea. Questa misura calcola la distanza geometrica tra i punti del dataset. I cluster sono diversi rispetto al caso della distanza di Manhattan che tende a creare cluster più allungati, a differenza della distanza euclidea che forma cluster più compatti. Tuttavia, sono presenti delle similarità. Ad esempio, alcuni stati si raggruppano nello stesso modo in entrambi i grafici (South Africa da solo; Turchia messico e Brasile), suggerendo che la struttura dei dati è stabile rispetto alle due misure di distanza.

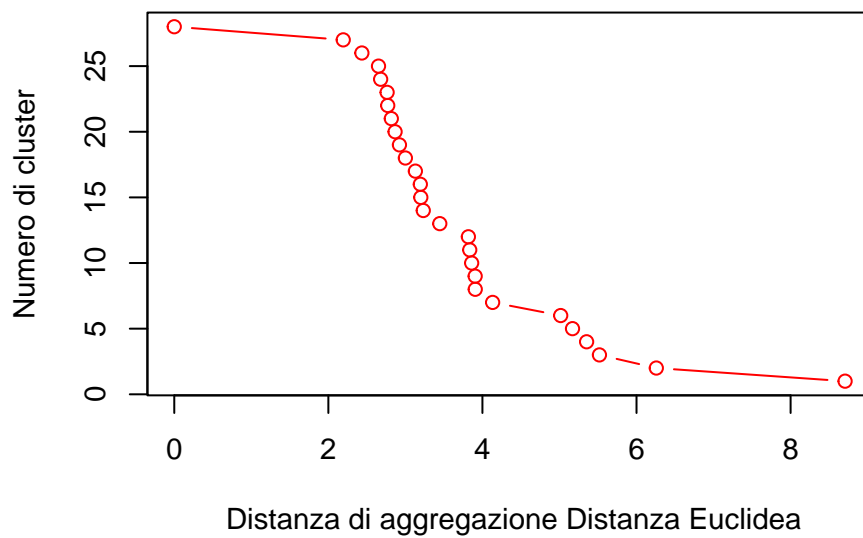


Per entrambe le distanze viene riportato lo screeplot, seppur considerazioni sulla validazione dei cluster verranno effettuate successivamente.

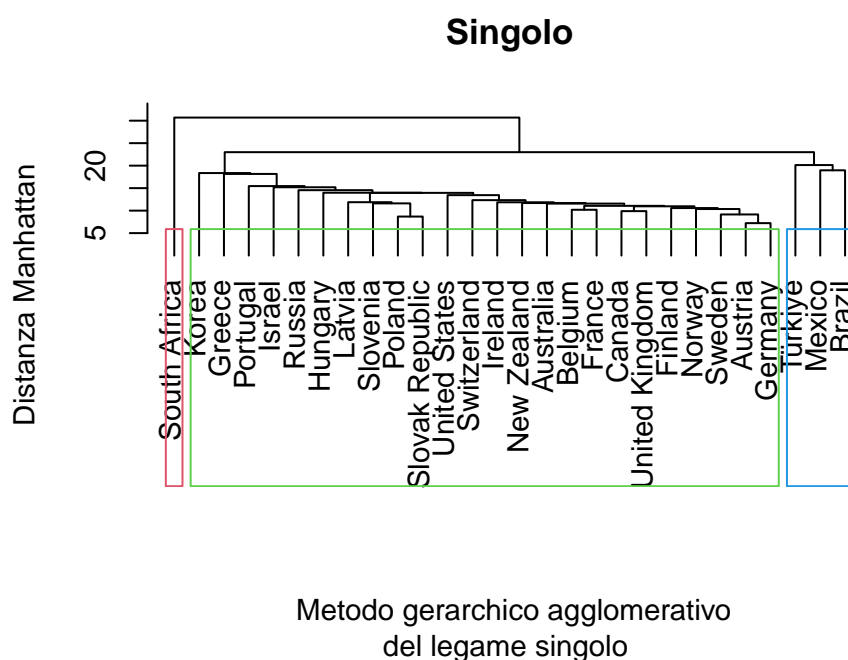
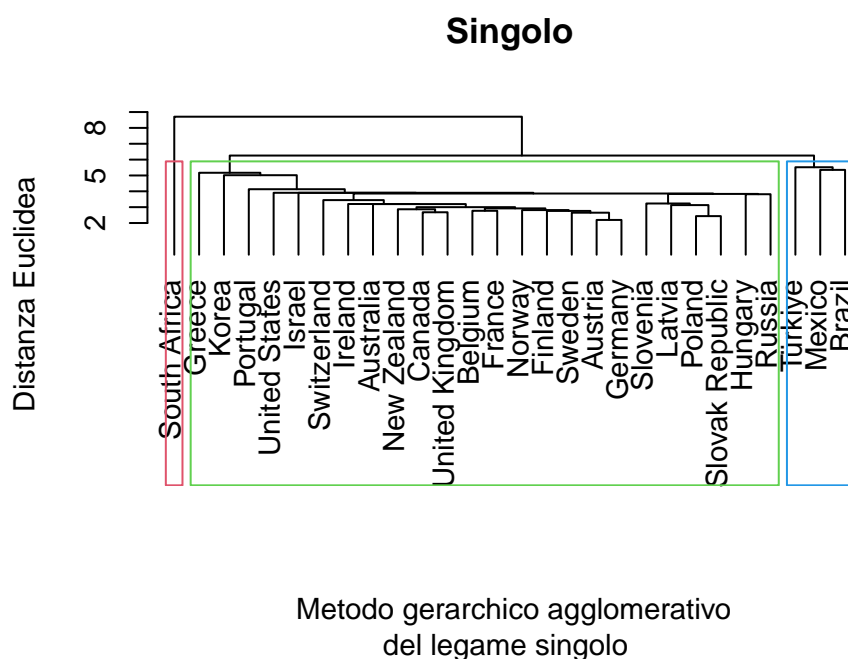
Screeplot



Screeplot



Sulla base del dendrogramma si ritiene che il taglio ad altezza di 3 cluster potrebbe essere ottimale. Si visualizzano i gruppi scelti sul dendrogramma. All'altezza 3, per entrambe le distanze si ottengono gli stessi sottogruppi, nonostante la riduzione della similarità è differente.



Per valutare l'omogeneità o la non omogeneità dei tre cluster, prima si calcola una misura globale della loro omogeneità. Questo permette di ottenere una comprensione generale della similarità o differenza tra i cluster stessi. La misura di non omogeneità totale è 648.

Successivamente, si procede al calcolo dell'omogeneità all'interno di ciascun cluster, ossia quanto gli elementi all'interno di ogni cluster siano simili tra loro. Questa misura interna ai cluster è comunemente denominata “within-cluster heterogeneity” o “non omogeneità interna ai cluster”. I risultati sono i seguenti: - nel primo cluster la misura di non omogeneità statistica è la più alta poichè è il gruppo che contiene più elementi al suo interno; - nel secondo cluster (con presenza di tre stati) il valore è nettamente diminuito; - nel terzo cluster il risultato è 0 perchè è presente una sola osservazione (South Africa)

```

      trHs1      trHs2 trHs3
[1,] 380.121 37.00194      0

```

Il valore della metrica within è significativamente più grande dell'indice between si potrebbe dedurre che la suddivisione in tre cluster potrebbe non essere soddisfacente. Questo suggerisce che i punti all'interno dei cluster sono molto simili tra loro, ma ci potrebbero essere notevoli differenze tra i cluster stessi.

Una situazione in cui la varianza all'interno dei cluster è molto più grande della varianza tra i cluster potrebbe indicare che i cluster non stanno realmente catturando strutture significative nei dati o che la suddivisione dei dati in cluster non è stata efficace.

Si potrebbero considerare delle alternative come:

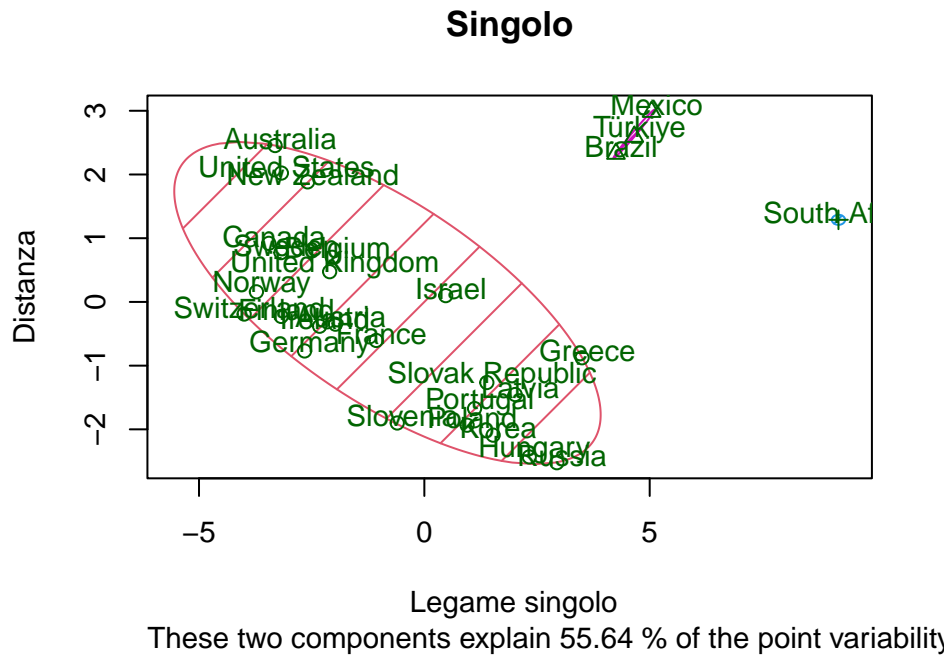
- Aumentare il numero di cluster per cercare di catturare strutture più dettagliate nei dati.
- Utilizzare un'altra tecnica di clustering che potrebbe essere più adatta.

Within Total	Between Total
417.123	230.877

Il rapporto tra la misura di non omogeneità tra cluster between e la traccia della matrice di non omogeneità è del 35.63%, dunque per nulla alta.

E' possibile in fine visualizzare i risultati in un diagramma di Venn i tre cluster.

Dal diagramma si nota che è presente un singolo grande gruppo e gli altri 2 invece sono più piccoli, questo potrebbe essere un effetto collaterale del metodo a legame singolo, proprio perché viene a crearsi un effetto a catena che lega elementi dissimili perché si guarda solo alla distanza minima. Il south Africa costituisce un cluster a sè, oltre che Turchia, Brasile e Messico.

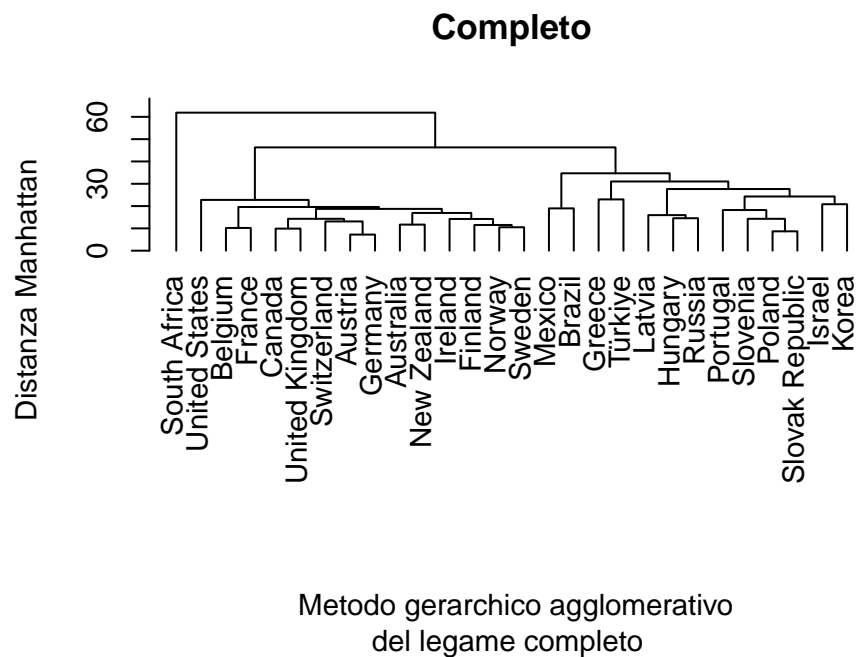
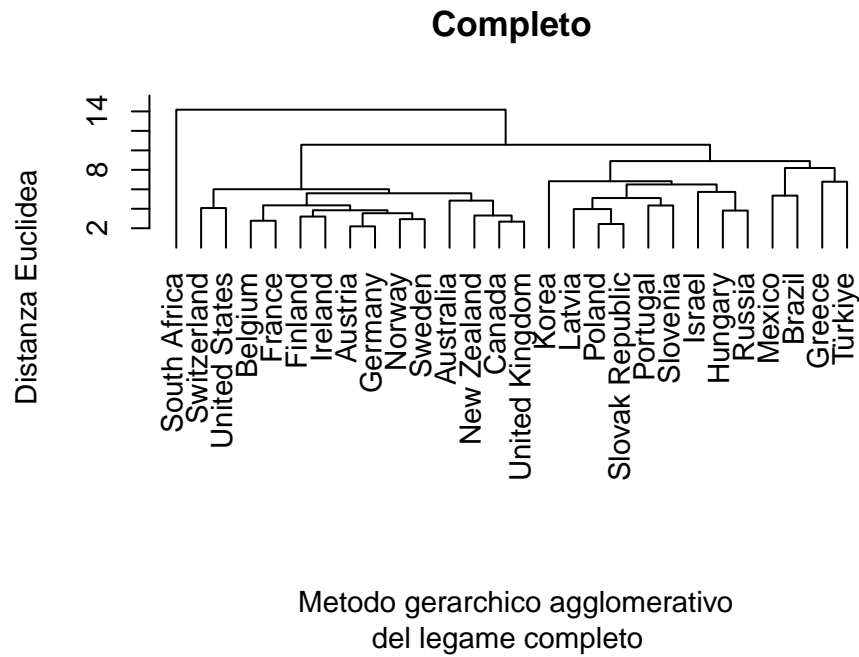


COMPLETE LINKAGE

Il complete linkage tra due gruppi A e B è definito come la massima dissimilarità tra due oggetti appartenenti ai due gruppi:

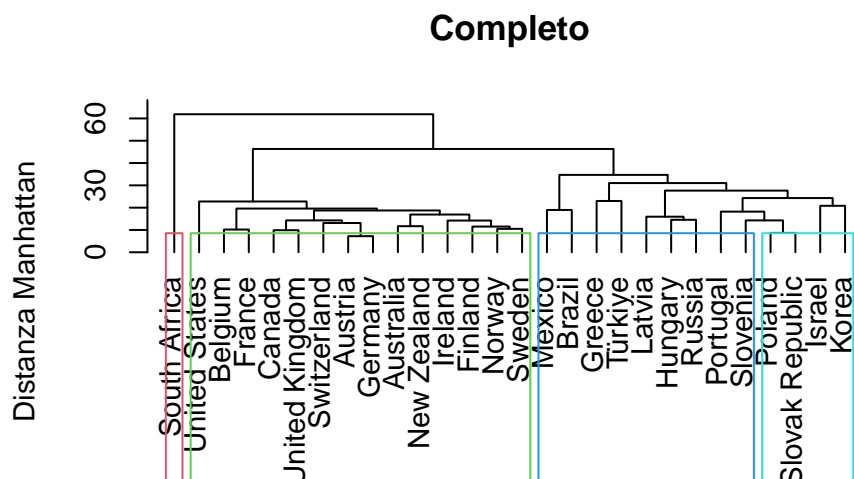
$$d_{CL}(A, B) = \max d(a, b) : a \in A, b \in B$$

Il dendrogramma ottenuto con il legame completo ha rami più lunghi rispetto al dendrogramma ottenuto con il legame singolo, poiché i gruppi si formano a livelli di distanza maggiori. Infatti la scala delle ordinate va da 2 a 12 per la distanza euclidea e non da 2 a 8 come nel metodo del legame singolo. La differenza nei dendrogrammi per le due distanze è evidente con gruppi più compatti per la distanza di Manhattan e meno compatta per quella euclidea, oltre che sottogruppi di Stati diversi. Rispetto al legame singolo non sembra esserci presente la distinzione di Turchia, Messico e Brasile, tuttavia South Africa continua a non essere associato a nessuna osservazione.

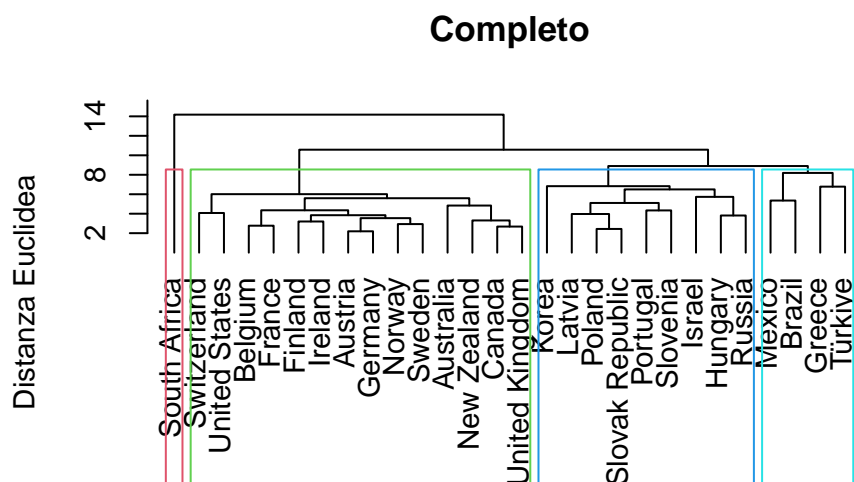


In corrispondenza di altezze alte si riduce di molto la dissimilarità. Dalla rappresentazione grafica sono presenti situazioni in cui la dissimilarità sembra ridursi molto a un taglio di cluster pari a 4, in tal caso vengono uniti gruppi piuttosto differenti sulla base della misura. Ora considerare la distanza di manhattan o euclidea rende significativa la diversa associazione. South Africa continua a rimanere in un cluster a parte; E' presente un cluster di 4 che si

distingue nei paesi Messico, Brasile, Grecia e Turchia per la distanza euclidea mentre Polonia, Slovacchia, Israele e Korea per la distanza di manhattan.



Metodo gerarchico agglomerativo
del legame completo



Metodo gerarchico agglomerativo
del legame completo

Per entrambe le distanze si calcolano le metriche utili alla descrizione. La misura di non omogeneità è 648.

[1] 648

La tabella presenta quattro gruppi di cluster con misurazioni di eterogeneità basate su distanza euclidea e di Manhattan. Considerando la distanza euclidea, il Gruppo 1 mostra un alto grado di eterogeneità (112.078), seguito da Gruppo 3 (104.69), mentre Gruppo 2 è meno eterogeneo (71.544). Gruppo 4 è l'unico con un'eterogeneità euclidea pari a zero, poichè è presente un'unica osservazione.

Nel caso dell'eterogeneità di Manhattan, i risultati sono simili, ma con valori leggermente più bassi in generale. Gruppo 1 è ancora il più eterogeneo (77.593), seguito da Gruppo 3 (14.322) e Gruppo 2 (50.427). Gruppo 4 ha ancora un'eterogeneità di Manhattan pari a zero, sempre poichè possiede un'unica osservazione.

Gruppo	Within Eterogenity Euclidea	Within Eterogenity Manhattan
Gruppo 1	112.078	77.593
Gruppo 2	71.544	50.427
Gruppo 3	104.69	14.322
Gruppo 4	0	0

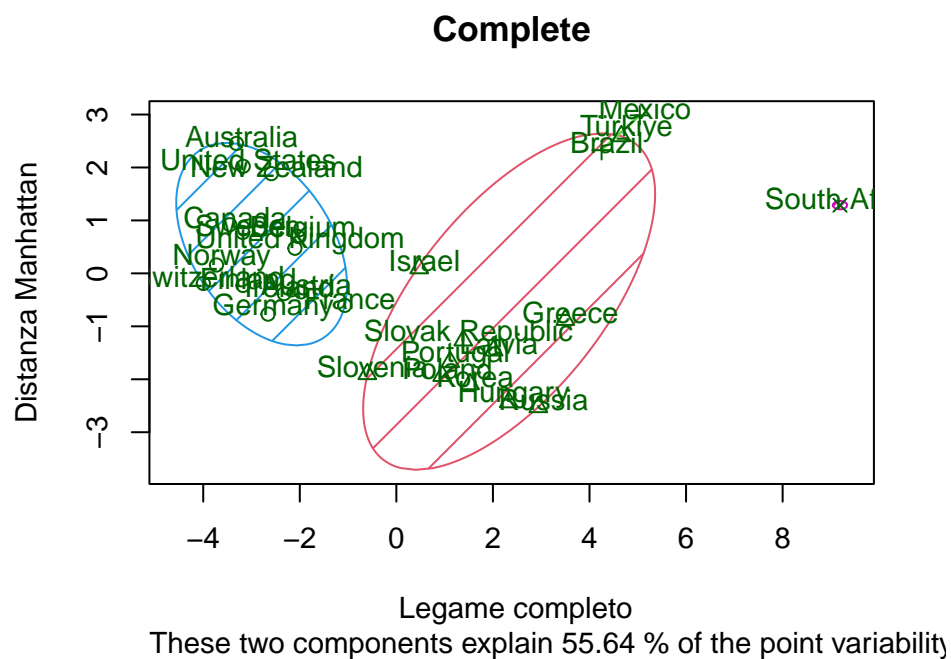
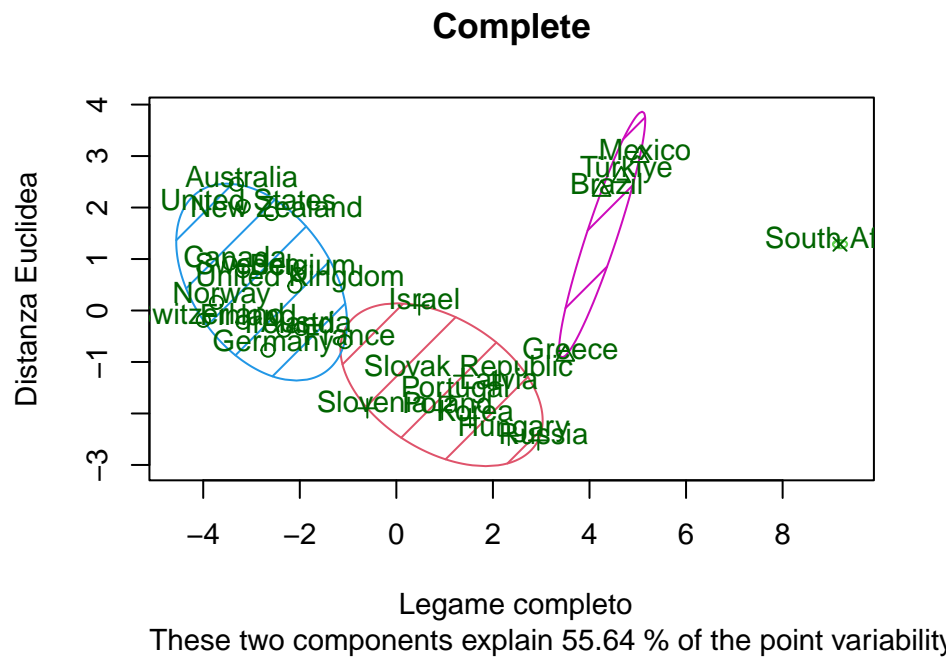
La distanza euclidea mostra un valore di within maggiore mentre la distanza di manhattan mostra un valore maggiore per la betwee. In entrambi i casi, qualunque delle due distanze si consideri, la “Between Total” è maggiore della “Within Total”, suggerendo che c'è una maggiore differenza tra i cluster rispetto alla variabilità all'interno dei cluster. Per questo motivo, la suddivisione dei dati in cluster sembra essere in grado di catturare una notevole differenza tra i cluster stessi.

Distanza	Within Total	Between Total
Euclidea	288.316	359.683
Manhattan	142.343	505.656

Per la distanza euclidea, il rapporto tra “Between Total” e “Total” è del **55,55%**, il che significa che circa il 55,55% della varianza totale è dovuta alle differenze tra i cluster (between), mentre il restante 44,45% è dovuto alla variabilità all'interno dei cluster (within). Così, il metodo completo (utilizzato con la distanza euclidea) ha una maggiore capacità di catturare le differenze tra i cluster rispetto al legame singolo con circa il 17% in più. Nel caso in cui si consideri la distanza di Manhattan, il rapporto tra “Between Total” e “Total” cresce significativamente fino al **78,03%**. Per cui, il metodo completo utilizzato con la distanza di Manhattan è particolarmente efficace nel catturare le differenze tra i cluster, con una percentuale molto maggiore rispetto alla distanza euclidea.

Euclidea	Manhattan
55.55%	78.03%

Anche se la distanza di manhattan mostra metriche migliori per i gruppi, la rappresentazione grafica rende evidente come vengono aggregate osservazioni anche molto dissimili fra loro.

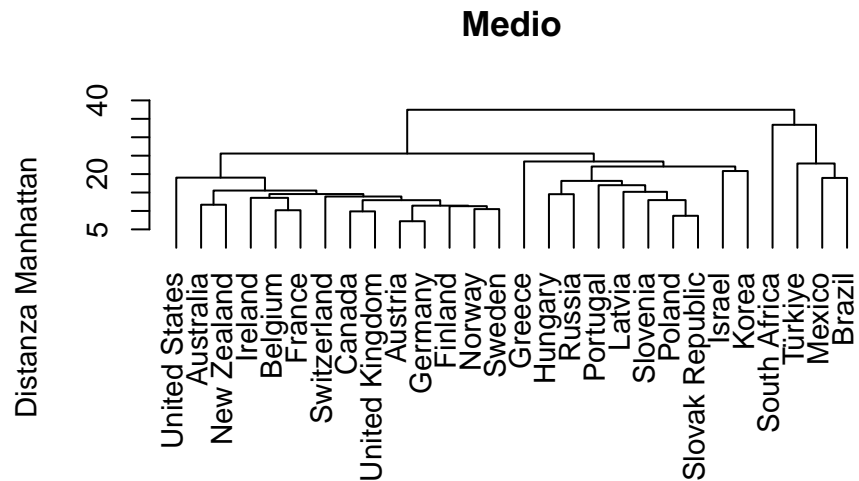
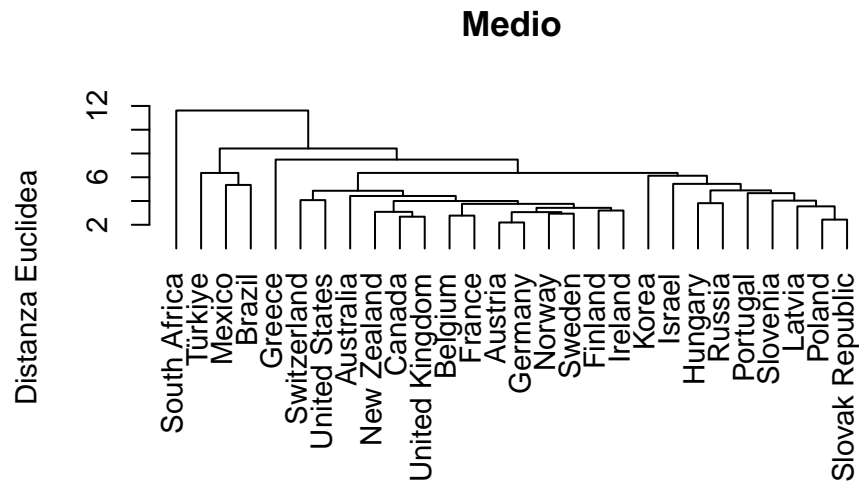


AVERAGE LINKAGE

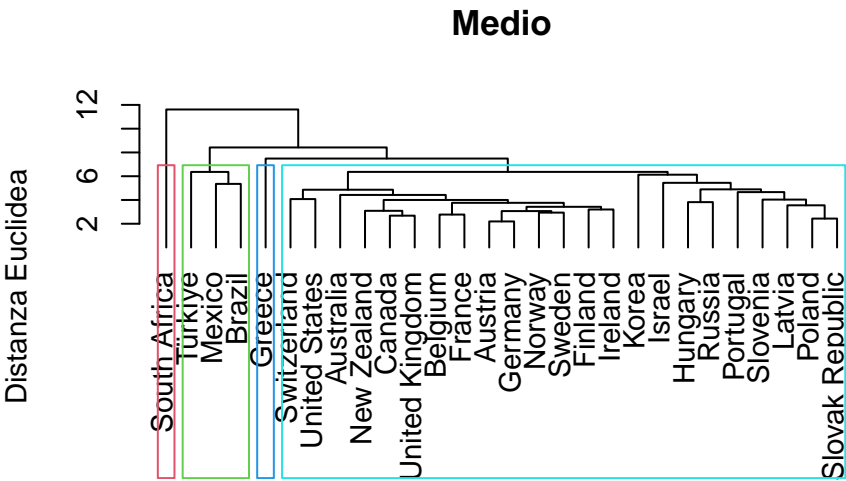
L'average linkage tra i gruppi A e B è definito come la dissimilarità media tra tutte le coppie di punti appartenenti a A e B .

$$d_{(AL)}(A, B) = \frac{1}{n_A n_B} \sum_{a \in A} \sum_{b \in B} d(a, b)$$

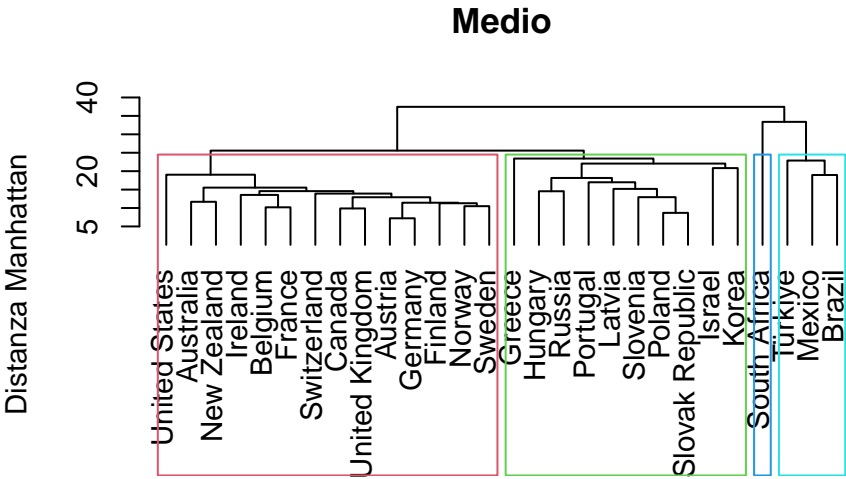
Il dendrogramma ottenuto utilizzando il legame medio ha rami di lunghezza intermedia rispetto al dendrogramma ottenuto con il legame singolo e a quelli ottenuti con il legame completo. Questo perché i gruppi si formano a livelli di distanza media tra i punti dati, il che significa che i cluster nel dendrogramma medio sono più coesi rispetto a quelli nel dendrogramma completo ma meno coesi rispetto a quelli nel dendrogramma singolo.



Un taglio ad altezza 4 mi sembra ragionevole in quanto tagliare ad altezza 2 significherebbe scremare troppo poco gli stati. Ancora una volta l'utilizzo della metrica individua cluster completamente diversi: sulla base della distanza euclidea, si ottengono metriche pressocchè simili al legame singolo ad eccezione di un cluster unico con la Grecia.



Metodo gerarchico agglomerativo
del legame medio



Metodo gerarchico agglomerativo
del legame medio

Si confrontano i 4 cluster ottenuti dall'average linkage sulla base delle distanze euclidee e di Manhattan.

Gruppo	Within Eterogenity Euclidea	Within Eterogenity Manhattan
Gruppo 1	339.050	77.593
Gruppo 2	0	0
Gruppo 3	37.001	37.001
Gruppo 4	0	0

La distanza Euclidea è sensibile a dispersioni su tutte le dimensioni, mentre la distanza di Manhattan è più robusta contro dispersioni su una singola dimensione.

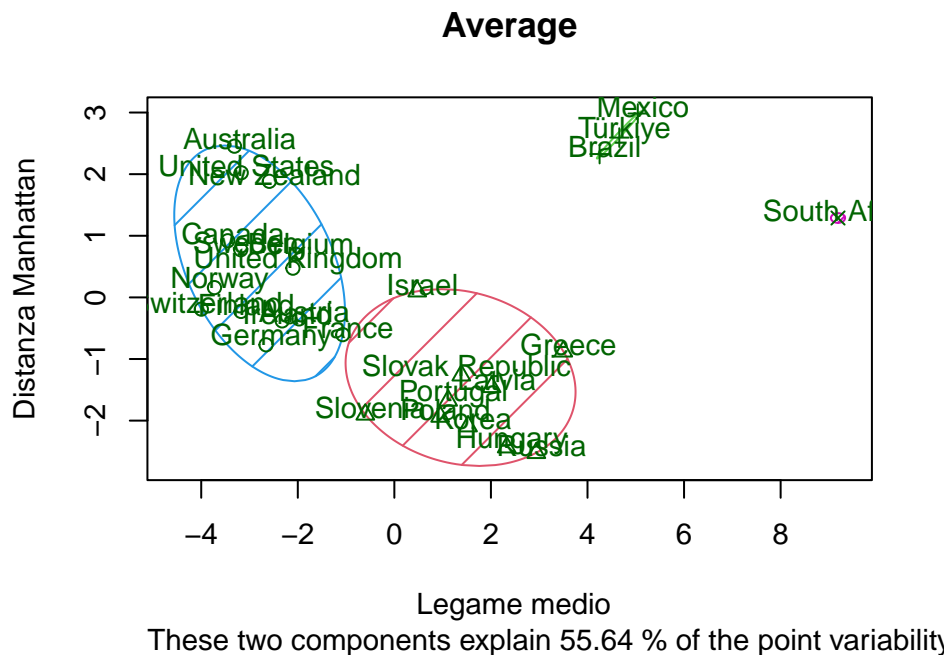
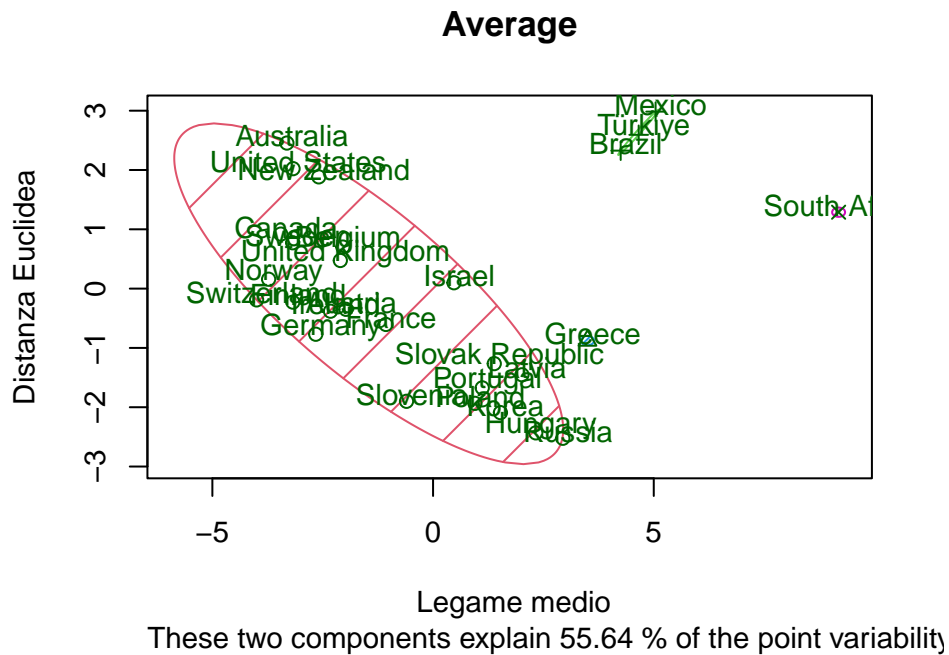
Distanza	Within Total	Between Total
Euclidea	376.052	271.94
Manhattan	114.595	533.405

Per la distanza euclidea, il rapporto tra “Between” e “Total” è del **41.97%**, il che significa che circa il 41.97% della varianza totale è dovuta alle differenze tra i cluster (between), mentre il restante è dovuto alla variabilità all'interno dei cluster (within). Così, il metodo dell'average (utilizzato con la distanza euclidea) ha una capacità maggiore di catturare le differenze tra i cluster rispetto al legame singolo ma inferiore rispetto al legame completo.. Nel caso in cui si consideri la distanza di Manhattan, il rapporto tra “Between” e “Total” cresce significativamente fino al **82.32%**. Il legame dell'average sulla distanza di Manhattan mostra risultati migliori della distanza euclidea e di tutti i linkage in generale.

Euclidea	Manhattan
41.97%	82.32%

[1] "82.32%"

Anche graficamente i cluster basati sulla distanza di Manhattan , sul legame medio, sono più comprensibili e compatti.

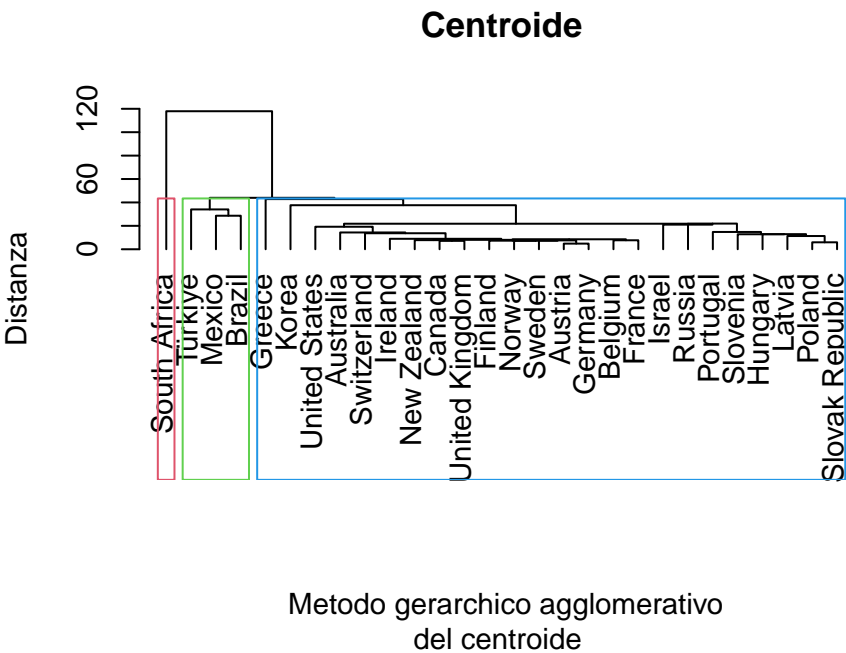


Il grafico che sembra più chiaro è quello basato sul linkage average, si riesce a visualizzare meglio e a comprendere a quali altezza viene effettuata la fusione. Questo perché il linkage average permette di non avere problemi né di crowding né di chaining, ovvero fondere rispetto alla massima o alla minima distanza fra coppie evitando tutte le altre.

Metodo del centroide

Nel contesto del metodo del centroide e del metodo della mediana, è necessario adoperare con la distanza euclidea. Il metodo del centroide, un algoritmo di clustering gerarchico, determina la distanza tra due cluster valutando la distanza tra i rispettivi centroidi, che rappresentano le medie campionarie dei dati nei due cluster.

$$d_{((i,j),k)} = \frac{1}{2}(d_{(i,k)}^2 + d_{(j,k)}^2) - \frac{1}{4}d_{i,j}^2$$



La partizione creata è pressocchè la stessa del legame singolo e medio.
La traccia ha sempre lo stesso valore, poichè si considerano lo stesso numero di unità statistiche.

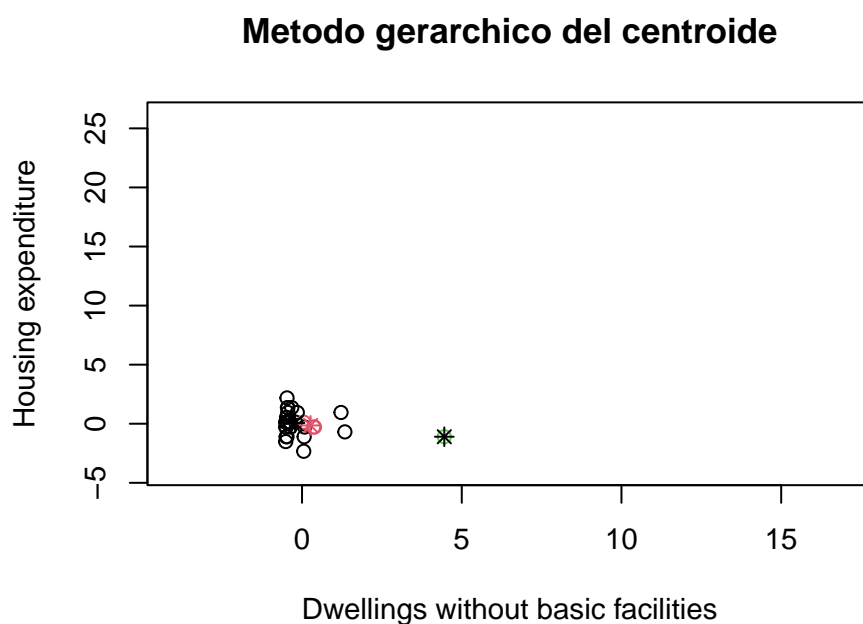
[1] 648

Anche i valori di eterogeneità interna, within totale e between totale sono gli stessi del metodo gerarchico con il linkage medio.

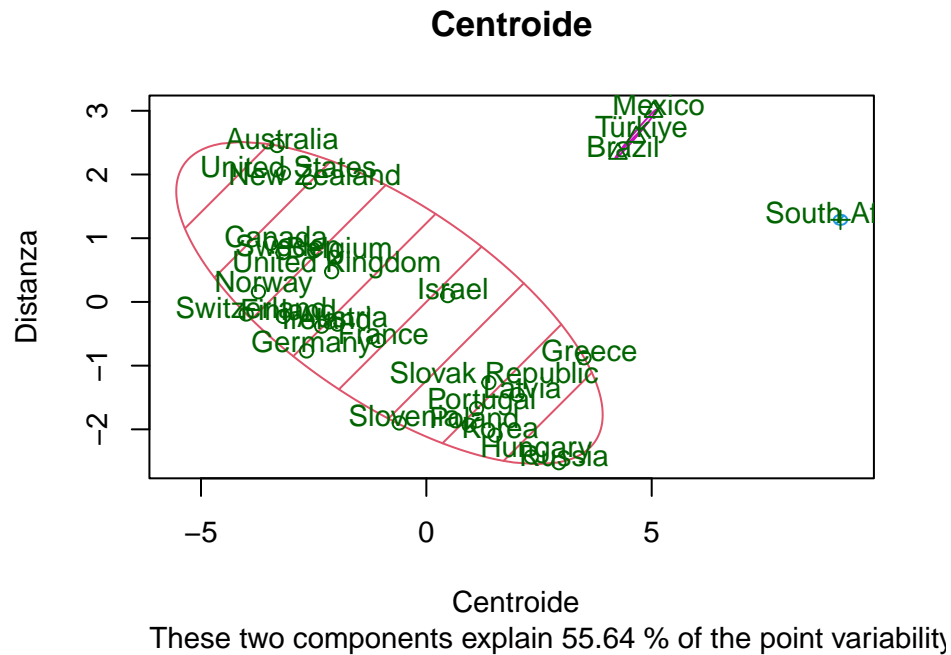
Gruppo	Within Eterogenity Euclidea
Gruppo 1	380.050
Gruppo 2	0
Gruppo 3	37.001

Distanza	Within Total	Between Total
Euclidea	417.123	230.87

Con questo grafico riusciamo a vedere come si distribuiscono i punti dei vari gruppi, e possiamo vedere che la partizione a tre va bene in quanto si nota il distacco dai gruppi.



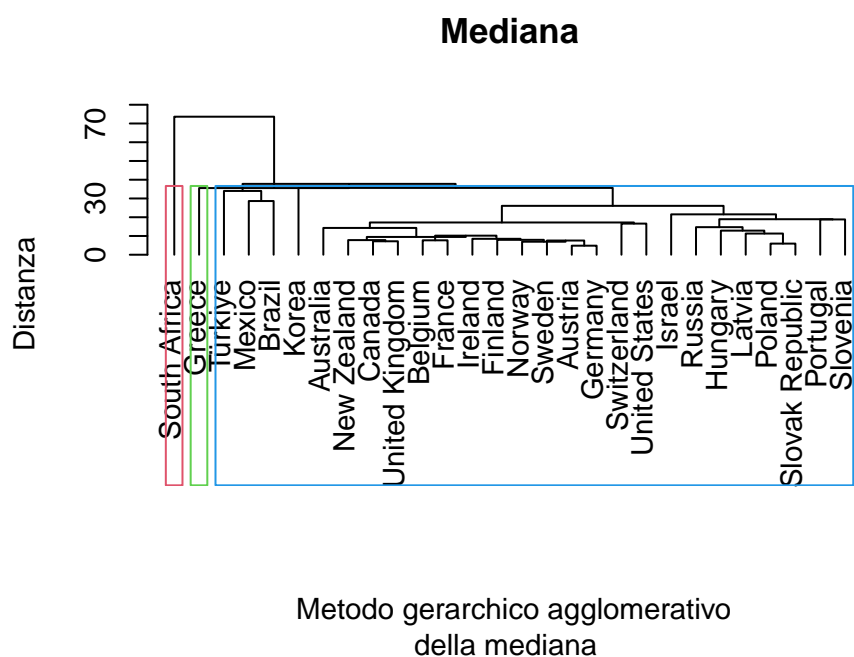
Proprio come il metodo del legame singolo e del legame medio si è formato un fenomeno gravitazionale che ha unito degli elementi dissimili tra loro in un unico cluster e ha creato cluster con un'unica informazione.



Metodo della mediana

Il procedimento della mediana presenta somiglianze con quello del centroide, ma si distingue per il fatto che la sua esecuzione è indipendente dal numero di cluster. Quando due gruppi si fondono, il nuovo centroide viene calcolato come la media dei due centroidi precedenti.

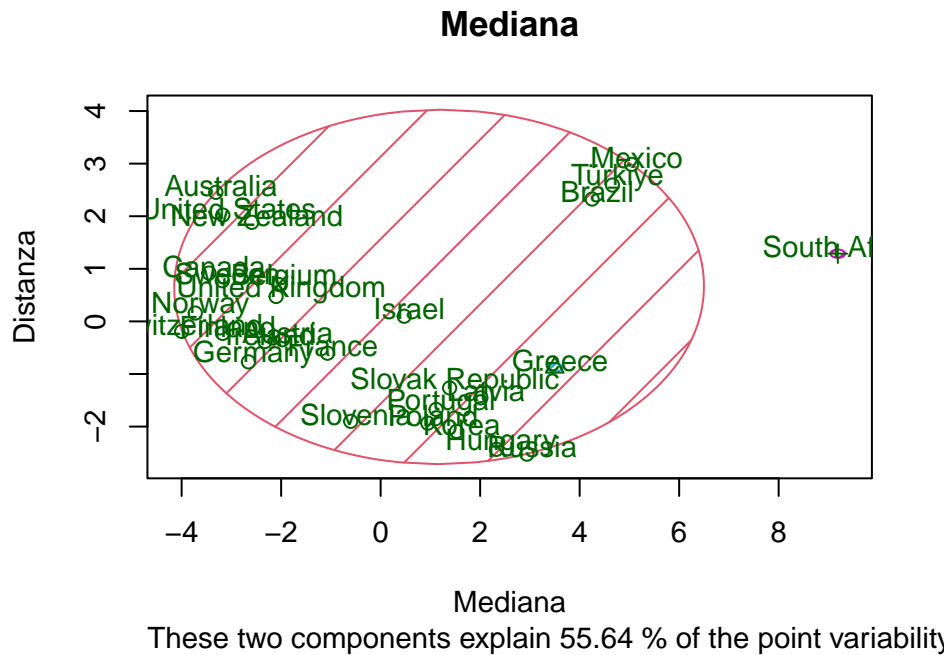
Dal grafico, a un plot di taglio pari a $k=3$ cluster si generano situazioni ancora peggiori di quelle precedenti: sono presenti 2 cluster con un'unica osservazione e un terzo cluster che possiede il resto degli stati.



Le metriche di rappresentazione sono piuttosto pessime. Il gruppo 1 omogeneizza troppo contenendo tutte le osservazioni meno 2 (Grecia e South Africa) con valori chiaramente elevati per la eterogeneità interna.

Gruppo	Within Eterogeneity Euclidean
Gruppo 1	497.160
Gruppo 2	0
Gruppo 3	0

Distanza	Within Total	Between Total
Euclidean	497.16	150.83



9.2. Validazione e scelta di K:

Il clustering è un task unsupervised. Ex-post non si possono confrontare i risultati con il **ground-truth**. Tuttavia, si ha bisogno di strumenti che in qualche modo danno una misura della qualità del clustering.

Scelta del numero di gruppi K

In alcuni metodi viene visto come un problema di stima, questo presuppone l'esistenza di un modello che definisce un "vero" K. In molte applicazioni i diversi valori di K potrebbero dare una "buona" rappresentazione della popolazione, per cui la scelta di K è un problema di validazione piuttosto che di stima.

I metodi più utilizzati per misurare qualità e accuratezza si basano su una statistica.

Silhouette Width

Il metodo prende in input un vettore con i cluster label e la matrice di dissimilarità. L'idea di base è che in un buon clustering ogni punto è ben connesso al suo cluster, mentre è scarsamente connesso agli altri clusters.

La dissimilarità media tra l'oggetto i del cluster C e tutti i punti del suo cluster è data da, misurando la connessione di i al suo cluster:

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

dove: C_i è il cluster al quale appartiene il punto $d(i, j)$ è la distanza tra il punto i e il punto j . La dissimilarità media tra i e il suo cluster più prossimo è il minimo di, misurando la connessione tra i e gli altri cluster a cui i non appartiene:

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j)$$

Il **silhouette width di i** è espresso come:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Questa metrica esprime quanto bene un punto è ancorato al suo cluster e quanto meno ancorato al cluster simile. Se il valore è prossimo a 1, allora il punto è ben accomodato nel suo cluster; se il valore è prossimo a -1 allora starebbe meglio nel cluster simile; se il valore è prossimo allo 0 allora il punto si trova in una regione di transizione tra i due clusters.

Infine, l'**average Silhouette Width**

$$ASW = \frac{1}{N} \sum_{i=1}^N S(i)$$

La situazione ottimale è volere un clustering con il più alto ASW possibile.

Criterio di Calinski-Harabasz

Il metodo prende in input i labels del clustering e la matrice di dissimilarità. L'idea di base è che in un buon clustering si ha un alto rapporto tra dissimilarità between e dissimilarità within.

Nella dissimilarità within-cluster si valuta la media delle distanze (o dissimilarità) tra tutti i punti all'interno di uno stesso cluster, misurando quanto i membri di un cluster sono simili tra loro. Nella dissimilarità between-clusters si calcola la media delle distanze tra i centroidi (o altri rappresentanti) dei diversi cluster, fornendo un'indicazione di quanto i cluster siano separati tra loro.

L'indice del criterio di Calinski è rappresentato da:

$$CHC = \frac{\text{Varianza tra cluster}}{\text{Varianza all'interno del cluster}} \frac{(Numero di osservazioni) - (Numero di cluster)}{(Numero di cluster) - 1}$$

Aumentando il numero di cluster generalmente W diminuisce e B aumenta, tuttavia l'**effetto sul fattore di correzione** è inverso, e per alti valori di K questo produce una forte deflazione del rapporto correggendo l'effetto di overfitting.

Dunn Index

L'indice Dunn è comunemente utilizzato per identificare cluster compatti e ben separati tra di loro, con una bassa variabilità all'interno di ciascun gruppo. Tale concetto è formalizzato dalla seguente espressione:

$$D = \frac{\min_{i \neq j} d(C_i, C_j)}{\max_i \max_{a, b \in C_i} d(a, b)}$$

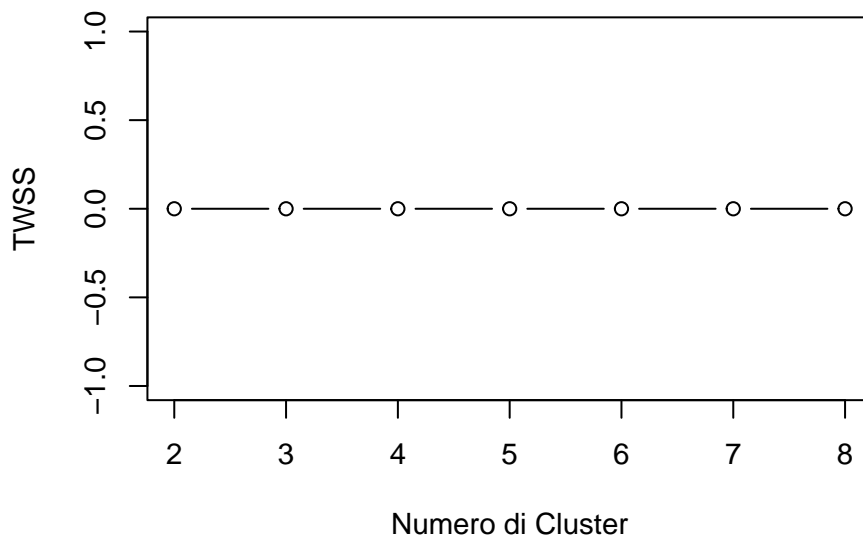
o meglio:

$$D = \frac{(\text{Minima distanza tra centroidi})}{(\text{Massima dimensione interna a un cluster})}$$

dove: C_i rappresenta la distanza tra i centroidi dei cluster ; $d(C_i, C_j)$ rappresenta la distanza tra i centroidi dei cluster i e j ; $d(a, b)$ rappresenta la distanza tra gli oggetti a e b nello stesso cluster.

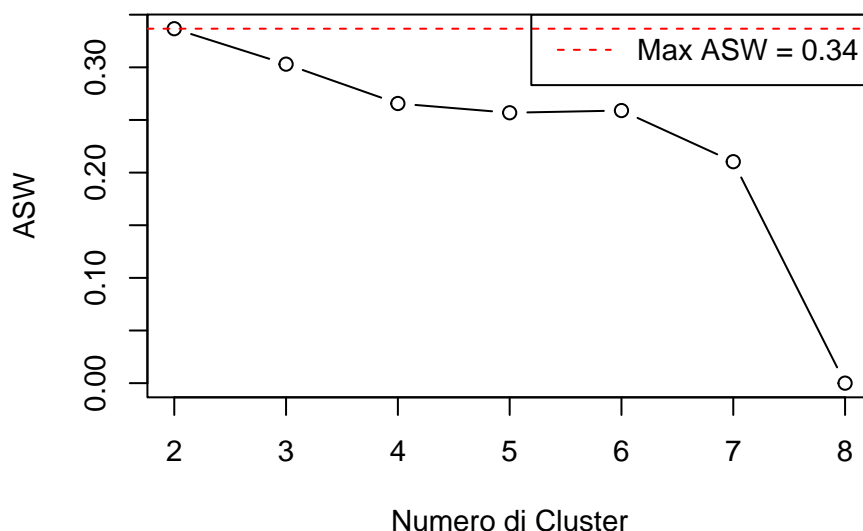
TWSS vs k: scelta del numero di gruppi

In base ai dati e agli obiettivi dell'analisi, la combinazione della distanza di Manhattan e il linkage "average" ha prodotto cluster che erano più omogenei tra i loro membri rispetto ad altre opzioni. Questa configurazione ha mostrato una maggiore coesione all'interno dei cluster, che è l'obiettivo desiderato nell'analisi di clustering, soprattutto se si cercano gruppi con pattern simili o comportamenti analoghi. La TWSS resta costante qualunque sia il taglio di cluster.



ASW vs K: scelta del numero di gruppi

Il plot mostra come varia il valore dell'ASW in corrispondenza di gruppi che vanno da $k=2$ a $k=8$. L'ASW diminuisce drasticamente, indicando che aggiungere ulteriori cluster non migliora significativamente la coerenza dei gruppi. Si osserva un repentino cambiamento nella pendenza dell'ASW da $k=7$ in poi, $k=7$ rappresenta il punto in cui l'aggiunta di cluster non migliora la coerenza e potrebbe addirittura peggiorare la qualità del clustering (portando l'ASW prossimo allo 0). Il numero ottimale di cluster è rappresentato dal picco massimo di ASW in $k=2$, indicando che due cluster potrebbero essere l'opzione ottimale. Questo punto riflette una buona coerenza all'interno dei cluster, mentre aggiungere ulteriori cluster sembra non portare a miglioramenti significativi.



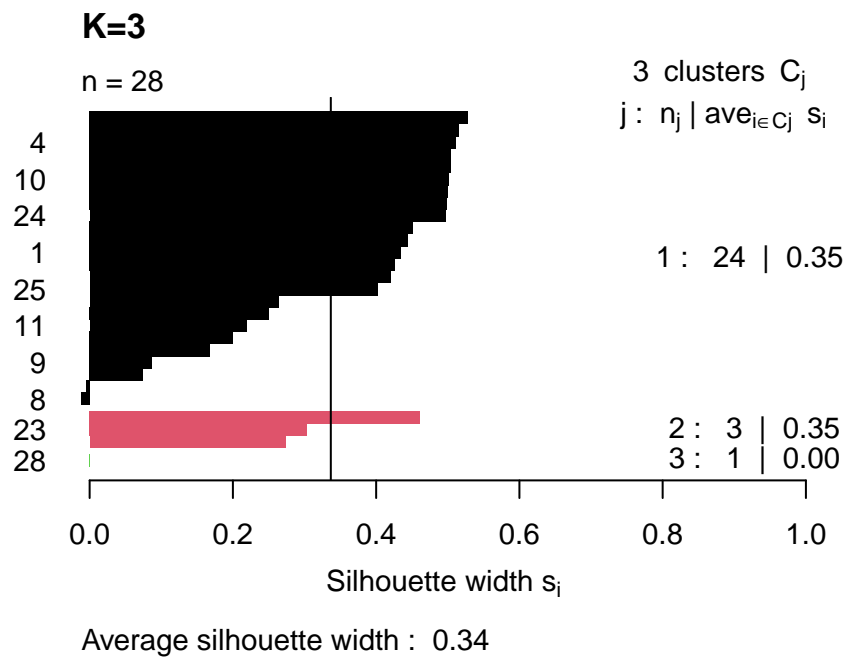
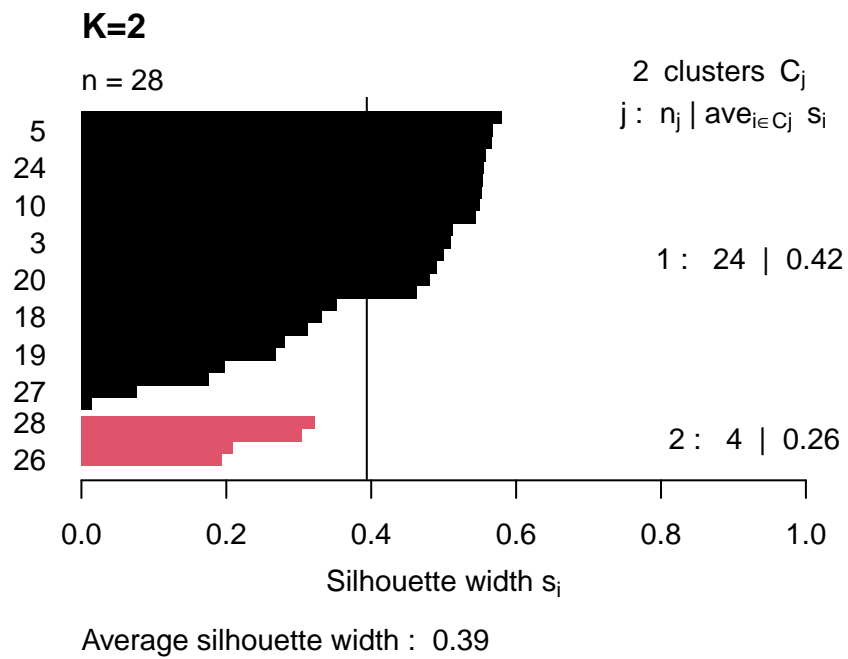
Analisi dell'ASW Migliore

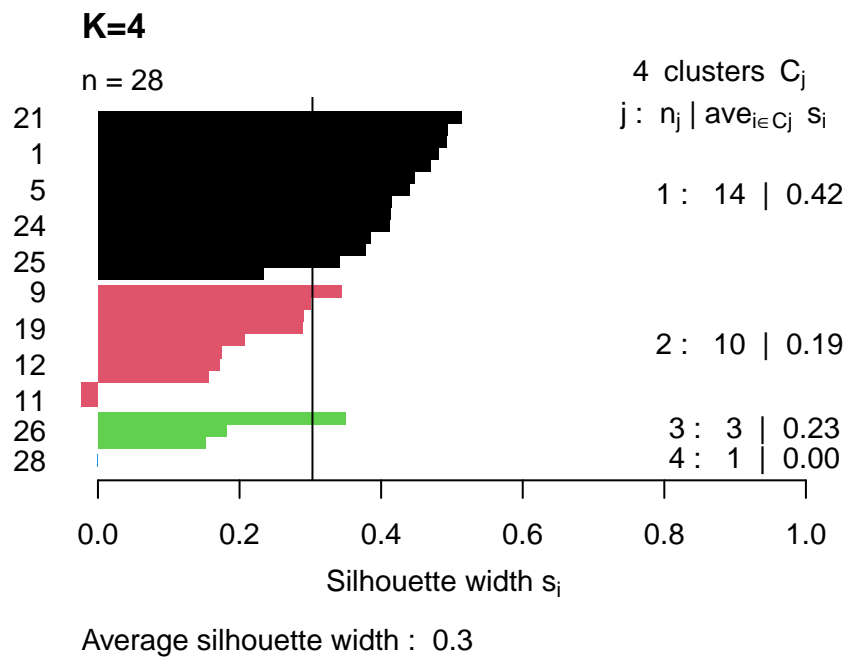
Il numero di gruppi migliori per l'ASW è 2, tuttavia, si confrontano graficamente k da 2 a 4. In dettaglio, si analizzano i silhouette plot delle partizioni più utili per effettuare valutazioni ulteriori circa la qualità del partizionamento ottenuto. A ciascuna osservazione presente nel campione è associata una barra orizzontale, di lunghezza pari al suo silhouette. Sarebbe auspicabile che, per ciascun gruppo, il maggior numero possibile di silhouette superasse la linea verticale posta in corrispondenza dell'ASW; tuttavia non è esattamente questo il caso specifico.

Per $k=2$, 24 osservazioni si trovano nel primo cluster e 4 nel restante. Il primo gruppo è abbastanza compatto: ad un buon numero di istanze è associato un silhouette maggiore della media, ed essi decrescono seguendo un andamento smooth e regolare, il primo cluster mostra ASW discreto (0.42). Il secondo cluster mostra valore più basso prossimo allo 0.26. Nonostante ciò il valore medio complessivo è il migliore 0.39.

Per un numero di gruppi pari a 3, la situazione è evidentemente peggiore. Per quanto concerne il primo gruppo, non solo la decrescita dei silhouette - da quelli delle unità più centrali al cluster a quelle meno ancorato ad esso è meno lenta e più frastagliata, ma sono anche presenti delle unità per cui $sw < 0$. Ciò implica che tali osservazioni sono state con ogni probabilità mis-classificate, in quanto risultano in media più dissimili dalle istanze presenti all'interno del proprio gruppo che da quelle appartenenti al cluster ad esse più prossimo. Queste, così come le numerose osservazioni per cui sw prossimo allo 0 che si trovano lungo la regione di transizione tra due gruppi - risultando in una scarsa connessione ad entrambi - portano ad un abbassamento dell'ASW totale misurato. Questa partizione individua anche un cluster con un'unica osservazione. L'asw totale diminuisce sino a 0.30 rispetto alla partizione precedente.

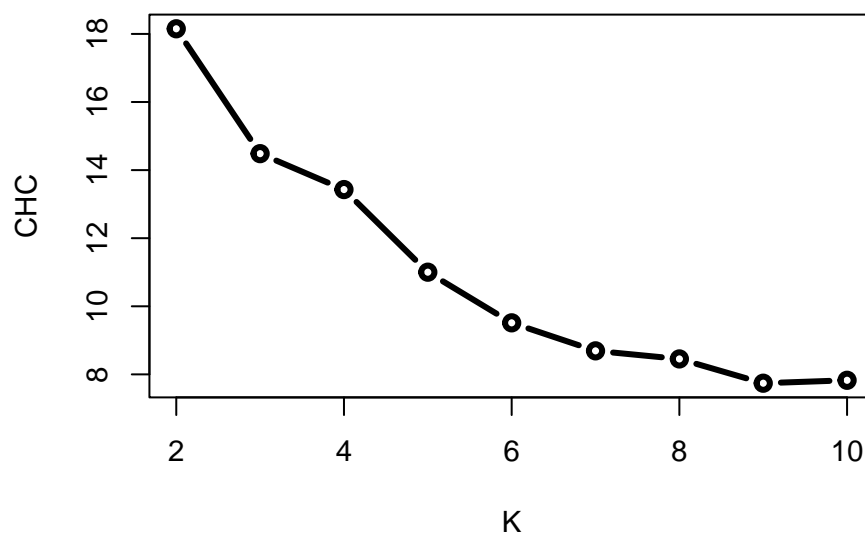
Per $k=4$, si creano cluster sempre più piccoli e più omogenei. Oltre che, al crescere di k insorgono sempre un numero maggiore di punti di frontiera.





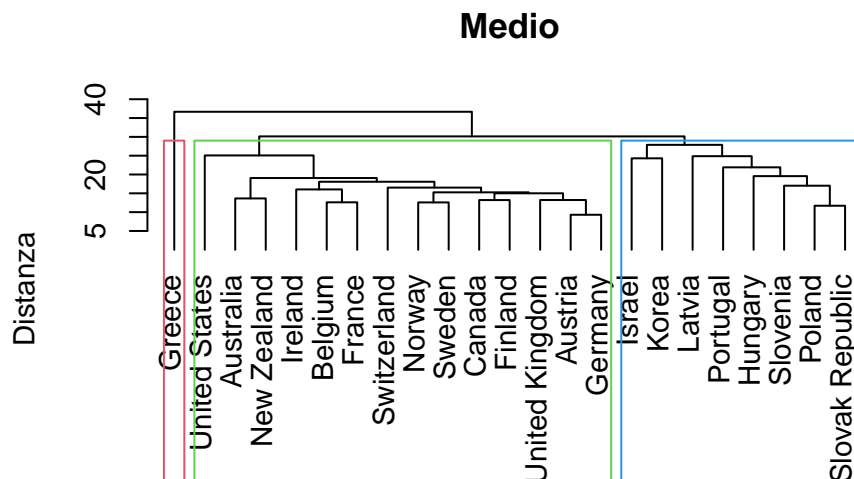
CHC vs K: scelta del numero di gruppi

Come l'ASW il numero di gruppi che massimizza il CHC sono 2. L'aumentare del numero di gruppi non assicura in alcun modo un incremento dell'indice, piuttosto corrisponde una ripida discesa.

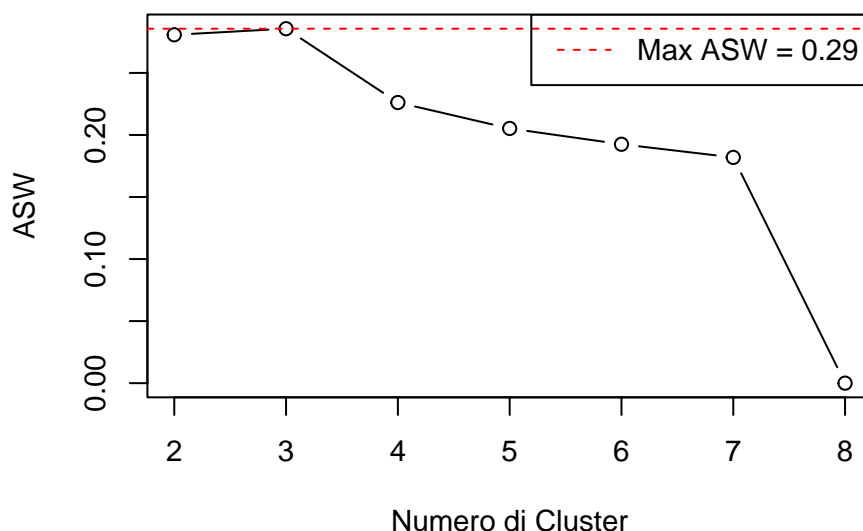


9.3. APPRONDIMENTO CON CLUSTER SENZA VALORI ANOMALI

Si decide, visti i risultati precedenti, di lavorare con le osservazioni selezionate appartenenti ai paesi aderenti all' OECD, così che i paesi estranei a quest'area non influiscano in maniera notevole per via delle differenze registrate piuttosto sensibili. Nonostante ciò, la differenza dovuta ad aggregazioni tra nazioni con background diversi permane facendo sì che continuino a essere presenti cluster con 1 e valori per l'ASW ancora più bassi del caso precedente.



Metodo gerarchico agglomerativo
del legame medio



9.4. Clustering non gerarchico

Dato un numero preassegnato K di gruppi, i metodi di clustering non gerarchico (NHC) cercano la partizione dei dati (“righe” o item) in K cluster in maniera tale da ottenere un buon clustering, espresso come partizione dove gli item entro ciascun cluster o gruppo siano quanto possibile simili tra loro, mentre gli item appartenenti a cluster diversi siano quanto possibile diversi. Tra i molti algoritmi di clustering non gerarchici sviluppati finora, il più popolare è il **K-means**. Per questo algoritmo la disomogeneità tra due oggetti è determinata sulla base della distanza Euclidea. Un possibile approccio per ottenere questo risultato potrebbe passare attraverso l’elencazione di tutti i possibili raggruppamenti in K gruppi costruibili con gli item e quindi scegliere come migliore soluzione il raggruppamento che ottimizza un qualche criterio predefinito. Sfortunatamente, un tale approccio diventerebbe rapidamente inapplicabile, specialmente per grandi dataset, poiché richiederebbe una quantità enorme di tempo macchina e di spazio di memoria. Infatti, il problema del K-means è **NP-hard**: il calcolo alla soluzione del problema della funzione obiettivo è computazionalmente intrattabile, poiché bisognerebbe verificare questa condizione per tutte le possibili assegnazioni di n oggetti a K liste. Di conseguenza tutte le tecniche di clustering disponibili sono iterative e operano solo su un numero molto ristretto di ripetizioni. Nella sua implementazione fondamentale, l’algoritmo K-means inizia assegnando gli item a uno dei K cluster predeterminati e quindi calcolando i K centroidi di gruppo, oppure pre-specificando i K centroidi di gruppo. I centroidi pre-specificati possono essere item selezionati casualmente oppure possono essere ottenuti “tagliando” un dendrogramma ad una altezza appropriata. In seguito, tramite una *procedura iterativa*, l’algoritmo cerca di minimizzare la somma dei quadrati entro gruppi (**WGSS**) su tutte le variabili, riassegnando (“spostando”) gli item sui diversi cluster. Si riallocano i punti per i quali i centroidi sono più vicini e si aggiornano i centroidi e i gruppi fin quando non si ottengono miglioramenti di TWSS con ulteriori riassegnamenti (per cui le

osservazioni non hanno necessità di cambiare gruppo).

- E' dimostrato che alla fine di ogni step si ottiene un decremento delle somme dei quadrati delle distanze dei punti osservati dai prototipi di clusters. Il K-means presenta alcuni vantaggi, come ad esempio la sua semplicità di implementazione e il suo tempo di esecuzione relativamente breve. Tuttavia, ha anche alcuni svantaggi, come la dipendenza dalla scelta dei centroidi iniziali e la difficoltà nel gestire cluster di forma irregolare o di densità variabile. Inoltre, K-means è sensibile alle outlier (punti anomali) e non è adatto a dati non linearmente separabili. Si visualizza la partizione ottenuta utilizzando il k-means con la scelta casuale di due centroidi. Il metodo quindi individua due partizioni dello spazio: un cluster con 8 elementi e un secondo cluster con 14 elementi. La misura di non omogeneità statistica within è circa 185, mentre quella tra cluster, between è 165.

```
1  2
8 15
```

K-means clustering with 2 clusters of sizes 8, 15

Cluster means:

	Dwellings without basic facilities	Housing expenditure	Rooms per person
1	0.7966213	0.03354579	-1.0879173
2	-0.4248647	-0.01789109	0.5802226
	Household net adjusted disposable income	Household net financial wealth	
1	-1.0305640		-0.7085442
2	0.5496341		0.3778902
	Labour market insecurity	Employment rate	Long-term unemployment rate
1	0.5930987	-0.7069324	0.4746378
2	-0.3163193	0.3770306	-0.2531402
	Personal earnings	Quality of support network	Educational attainment
1	-1.1492026	-1.0190430	-0.16139723
2	0.6129081	0.5434896	0.08607852
	Student skills	Years in education	Air pollution
1	-0.7555615	-0.6087262	0.8806068
2	0.4029662	0.3246540	-0.4696570
	Water quality		
1			-1.0034114
2			0.5351528
	Stakeholder engagement for developing regulations	Voter turnout	
1		-0.3373830	-0.5247148
2		0.1799376	0.2798479
	Life expectancy	Self-reported health	Life satisfaction
1	-0.7453341	-0.8380393	-0.9872247
2	0.3975115	0.4469543	0.5265198
	Feeling safe walking alone at night	Homicide rate	
1		-0.9214246	0.3168194
2		0.4914264	-0.1689703
	Employees working very long hours	Time devoted to leisure and personal care	
1		0.2111839	-0.5788629

2 -0.1126314 0.3087269

Clustering vector:

Australia	Austria	Belgium	Canada	Finland
2	2	2	2	2
France	Germany	Greece	Hungary	Ireland
2	2	1	1	2
Israel	Korea	Latvia	New Zealand	Norway
1	1	1	2	2
Poland	Portugal	Slovak Republic	Slovenia	Sweden
1	1	1	2	2
Switzerland	United Kingdom	United States		
2	2	2		

Within cluster sum of squares by cluster:

[1] 185.0413 177.1164

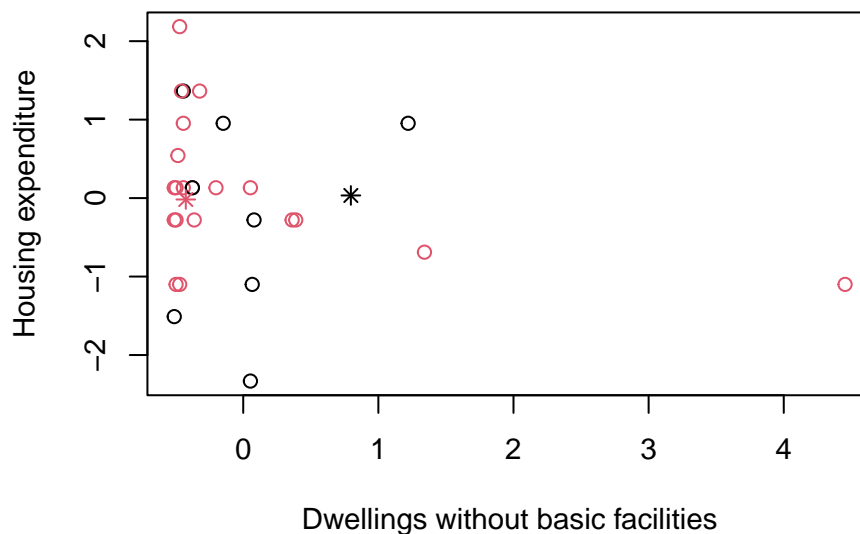
(between_SS / total_SS = 31.4 %)

Available components:

[1] "cluster"	"centers"	"totss"	"withinss"	"tot.withinss"
[6] "betweenss"	"size"	"iter"	"ifault"	

Il grafico mostra dove sono posti i centroidi, e la distribuzione dei cluster dei due gruppi rispetto alle prime due features.

Metodo non gerarchico del k-means con 2 centroidi



Si vede adesso il k-means con 3 centroidi. Oltre a cambiare il numero dei centroidi è stato

cambiato anche nstart, ossia si richiede che l'algoritmo di agglomerazione venga ripetuta 10 volte, assicurandoci di minimizzare la misura within e massimizzare la between. Anche in questo caso la misura within vale circa 112.07891, 132.03 e 91.55 per ciascun cluster, quindi diminuita e la between invece è cresciuta sino a 312.32. Dunque la partizione in 3 gruppi restituisce delle misure di non omogeneità statistica migliori, infatti il rapporto between e total è prossimo al 48.2 % (risultato migliore del metodo precedente).

```
1  2  3
14 4 10
```

K-means clustering with 3 clusters of sizes 14, 4, 10

Cluster means:

	Dwellings without basic facilities	Housing expenditure	Rooms per person	
1	-0.4470614	0.16132853	0.8532484	
2	1.3143509	-0.38132198	-1.3705722	
3	0.1001456	-0.07333115	-0.6463188	
	Household net adjusted disposable income	Household net financial wealth		
1	0.8452833		0.6674148	
2	-1.3255072		-0.9251258	
3	-0.6531938		-0.5643304	
	Labour market insecurity	Employment rate	Long-term unemployment rate	
1	-0.40125994	0.5717775	-0.3283610	
2	1.26216959	-1.5158751	0.5046717	
3	0.05689608	-0.1941384	0.2578367	
	Personal earnings	Quality of support network	Educational attainment	
1	0.8714302	0.6937037	0.3431146	
2	-1.4057273	-0.7384588	-2.0482904	
3	-0.6577114	-0.6758017	0.3389557	
	Student skills	Years in education	Air pollution	Water quality
1	0.54606661	0.5091394	-0.6535928	0.7230810
2	-2.13793950	-1.0156812	0.5274609	-1.2422977
3	0.09068254	-0.3065227	0.7040456	-0.5153943
	Stakeholder engagement for developing regulations	Voter turnout		
1		0.1517618		0.3571774
2		0.2790459		0.4816979
3		-0.3240849		-0.6927275
	Life expectancy	Self-reported health	Life satisfaction	
1	0.5020060	0.6178419		0.7704731
2	-1.4670340	-0.1100609		-0.7572650
3	-0.1159949	-0.8209543		-0.7757563
	Feeling safe walking alone at night	Homicide rate		
1		0.6292865		-0.3889589
2		-1.6365103		1.7262912
3		-0.2263970		-0.1459740
	Employees working very long hours	Time devoted to leisure and personal care		

1	-0.2736866	0.4885065
2	1.5828390	-1.3860691
3	-0.2499744	-0.1294815

Clustering vector:

Australia	Austria	Belgium	Canada	Finland
1	1	1	1	1
France	Germany	Greece	Hungary	Ireland
1	1	3	3	1
Israel	Korea	Latvia	Mexico	New Zealand
3	3	3	2	1
Norway	Poland	Portugal	Slovak Republic	Slovenia
1	3	3	3	3
Sweden	Switzerland	Türkiye	United Kingdom	United States
1	1	2	1	1
Brazil	Russia	South Africa		
2	3	2		

Within cluster sum of squares by cluster:

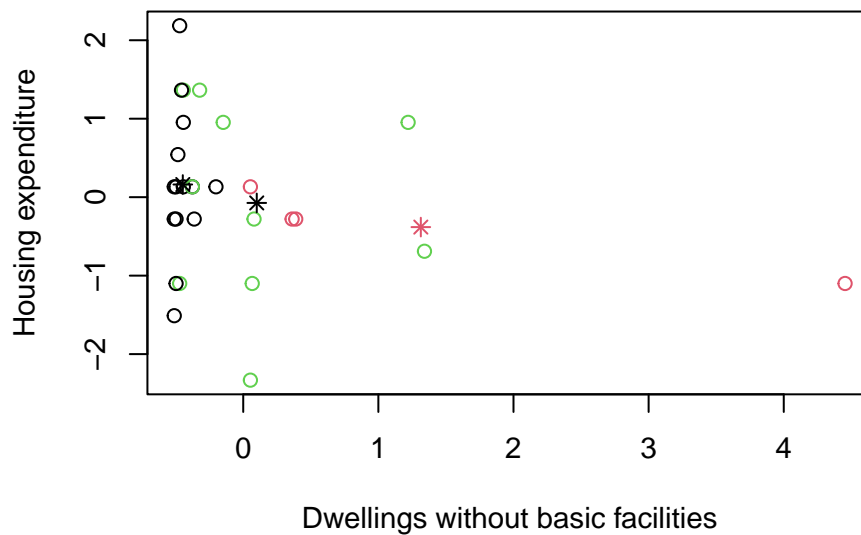
```
[1] 112.07891 91.55773 132.03889
(between_SS / total_SS = 48.2 %)
```

Available components:

```
[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

Sempre rispetto alle prime due variabili, si ottiene il plot rappresentativo.

Metodo non gerarchico del k-means con 3 centroidi



Si procede all'esplorazione del k-means utilizzando le misure gerarchiche. I tre cluster così individuati tenendo conto della distanza euclidea contengono ciascuno 15, 12 e 1 Paesi. La within è aumentata per il secondo cluster poichè contiene un numero di elementi maggiori rispetto al caso precedente mentre è uguale a 0 per il terzo cluster poichè è presente una sola osservazione.

K-means clustering with 3 clusters of sizes 15, 12, 1

Cluster means:

	Dwellings without basic facilities	Housing expenditure	Rooms per person	
1	-0.4485947	0.077242146	0.7833761	
2	0.1895898	-0.004888743	-0.8375719	
3	4.4538430	-1.099967264	-1.6997782	
	Household net adjusted disposable income	Household net financial wealth		
1	0.7559616		0.5799681	
2	-0.8083202		-0.6652657	
3	-1.6395816		-0.7163337	
	Labour market insecurity	Employment rate	Long-term unemployment rate	
1	-0.3949231	0.5256807	-0.2903262	
2	0.1627219	-0.4092567	0.1082095	
3	3.9711834	-2.9741292	3.0563785	
	Personal earnings	Quality of support network	Educational attainment	
1	0.8080426	0.6678452	0.3583642	
2	-0.8674693	-0.8080778	-0.2821165	
3	-1.7110075	-0.3207447	-1.9900649	
	Student skills	Years in education	Air pollution	Water quality

1	0.5552663	0.5105249	-0.5855580	0.7184091
2	-0.4796975	-0.5126030	0.6211043	-0.8037772
3	-2.5726243	-1.5066371	1.3301187	-1.1308095
	Stakeholder engagement for developing regulations Voter turnout			
1			0.2013700	0.2247924
2			-0.1892943	-0.3058400
3			-0.7490181	0.2981940
	Life expectancy Self-reported health Life satisfaction			
1	0.4936165	0.5588092	0.6624601	
2	-0.2733448	-0.6878807	-0.6545352	
3	-4.1241096	-0.1275706	-2.0824787	
	Feeling safe walking alone at night Homicide rate			
1		0.6717783	-0.3942213	
2		-0.6492970	0.4063095	
3		-2.2851112	1.0376051	
	Employees working very long hours Time devoted to leisure and personal care			
1		-0.2951505		0.45866343
2		0.2736283		-0.57461563
3		1.1437173		0.01543608

Clustering vector:

Australia	Austria	Belgium	Canada	Finland
1	1	1	1	1
France	Germany	Greece	Hungary	Ireland
1	1	2	2	1
Israel	Korea	Latvia	Mexico	New Zealand
2	2	2	2	1
Norway	Poland	Portugal	Slovak Republic	Slovenia
1	2	2	2	1
Sweden	Switzerland	Türkiye	United Kingdom	United States
1	1	2	1	1
Brazil	Russia	South Africa		
2	2	3		

Within cluster sum of squares by cluster:

```
[1] 128.7197 220.5248 0.0000
(between_SS / total_SS = 46.1 %)
```

Available components:

```
[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

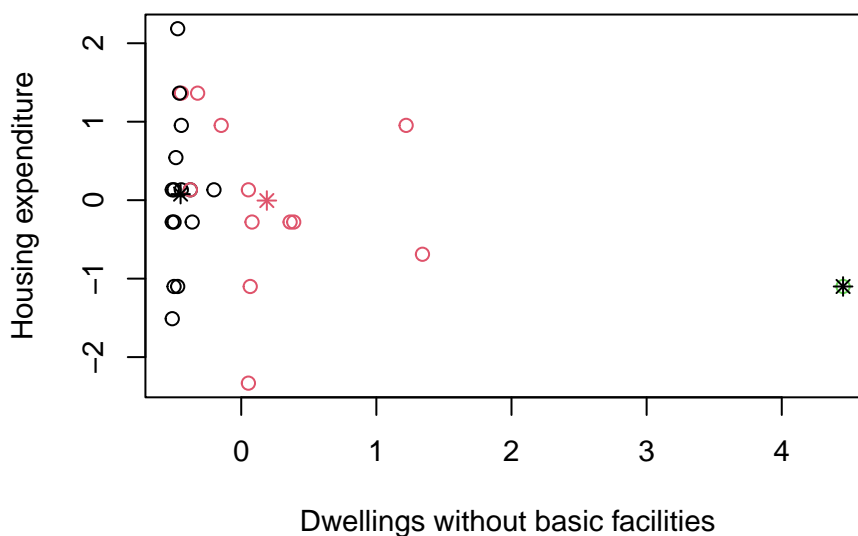
	Dwellings without basic facilities	Housing expenditure	Rooms per person
1	-0.4485947	0.077242146	0.7833761
2	0.1895898	-0.004888743	-0.8375719
3	4.4538430	-1.099967264	-1.6997782

	Household net adjusted disposable income		Household net financial wealth	
1		0.7559616		0.5799681
2		-0.8083202		-0.6652657
3		-1.6395816		-0.7163337
	Labour market insecurity		Employment rate	
1	-0.3949231	0.5256807		-0.2903262
2	0.1627219	-0.4092567		0.1082095
3	3.9711834	-2.9741292		3.0563785
	Personal earnings		Quality of support network	
1	0.8080426	0.6678452		0.3583642
2	-0.8674693	-0.8080778		-0.2821165
3	-1.7110075	-0.3207447		-1.9900649
	Student skills		Years in education	
1	0.5552663	0.5105249	-0.5855580	0.7184091
2	-0.4796975	-0.5126030	0.6211043	-0.8037772
3	-2.5726243	-1.5066371	1.3301187	-1.1308095
	Air pollution		Water quality	
1				
2				
3				
	Stakeholder engagement for developing regulations		Voter turnout	
1		0.2013700		0.2247924
2		-0.1892943		-0.3058400
3		-0.7490181		0.2981940
	Life expectancy		Self-reported health	
1	0.4936165	0.5588092		0.6624601
2	-0.2733448	-0.6878807		-0.6545352
3	-4.1241096	-0.1275706		-2.0824787
	Life satisfaction			
1				
2				
3				
	Feeling safe walking alone at night		Homicide rate	
1		0.6717783		-0.3942213
2		-0.6492970		0.4063095
3		-2.2851112		1.0376051
	Employees working very long hours		Time devoted to leisure and personal care	
1		-0.2951505		0.45866343
2		0.2736283		-0.57461563
3		1.1437173		0.01543608

[1] 128.7197 220.5248 0.0000

[1] 298.7556

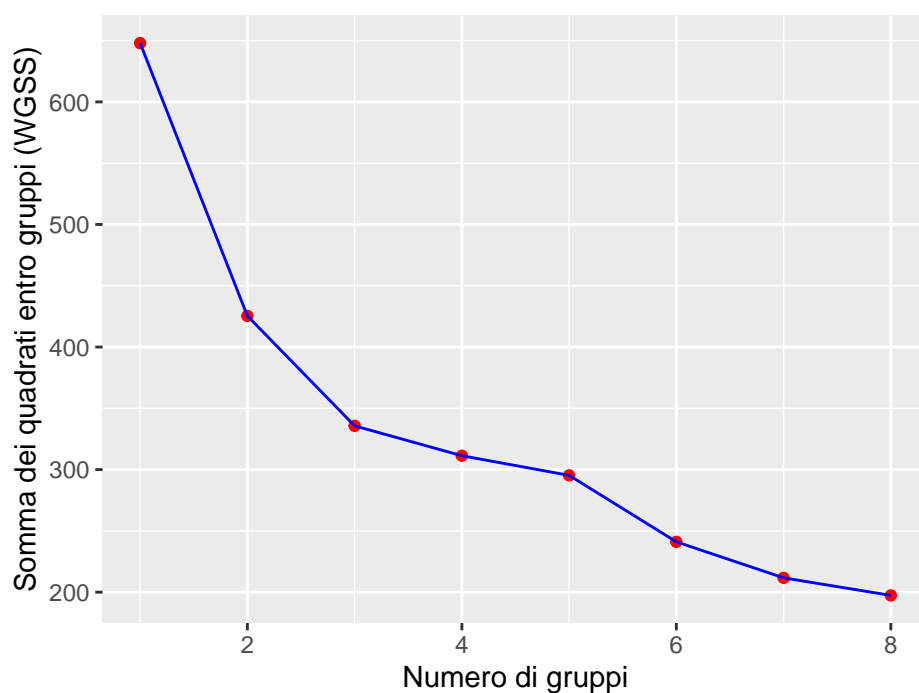
Metodo non gerarchico del k-means con 3 centroidi



9.5. Validazione dei cluster

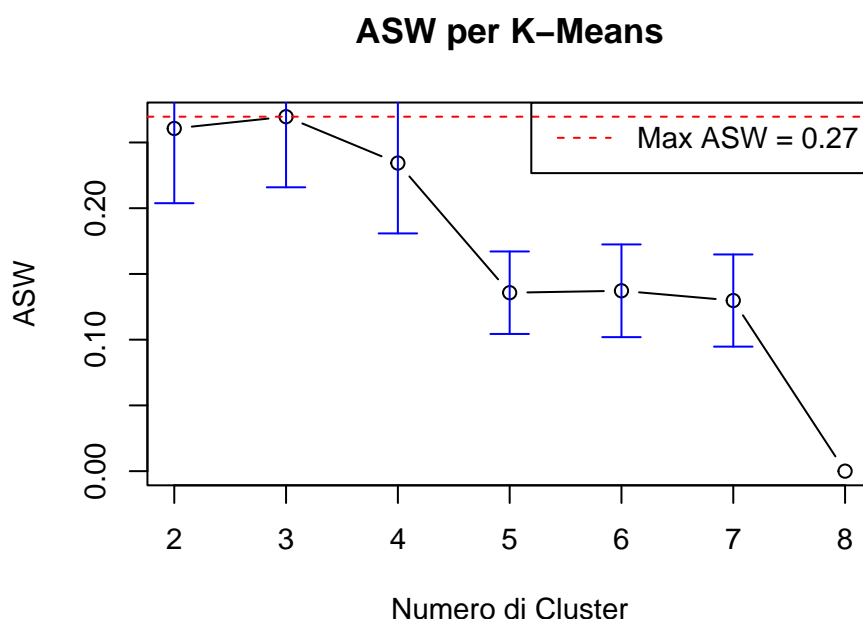
WGSS vs K: scelto del numero di cluster

Dapprima tracciamo i valori di WGSS per soluzioni di raggruppamento con un numero di cluster che varia tra 1 e 8 per vedere se possiamo ottenere indicazioni sul numero di gruppi (che usualmente non è noto a priori). Valori di $K > 8$ non vengono considerati in quanto, oltre che comportare una forte instabilità di convergenza, si ritiene che conducano ad un partizionamento eccessivamente granulare dei dati considerata la natura del problema: i cluster derivanti risulterebbero infatti di complessa interpretazione. Il grafico può essere ottenuto come segue: Mano a mano che il numero di gruppi aumenta, la somma dei quadrati necessariamente si riduce; Secondo il criterio dell'elbow, il numero di gruppi per cui optare è quello in corrispondenza del quale la curva osservata cambia pendenza: un ulteriore incremento di K renderebbe sempre più complessa la struttura con la quale i dati vengono descritti, senza tuttavia condurre ad un significativo miglioramento della qualità del clustering. Seppure non in maniera marcata, il plot mostra un elbow in corrispondenza di $K = 3$; per valori maggiori, la decrescita della WGSS è più smooth e meno ripida.



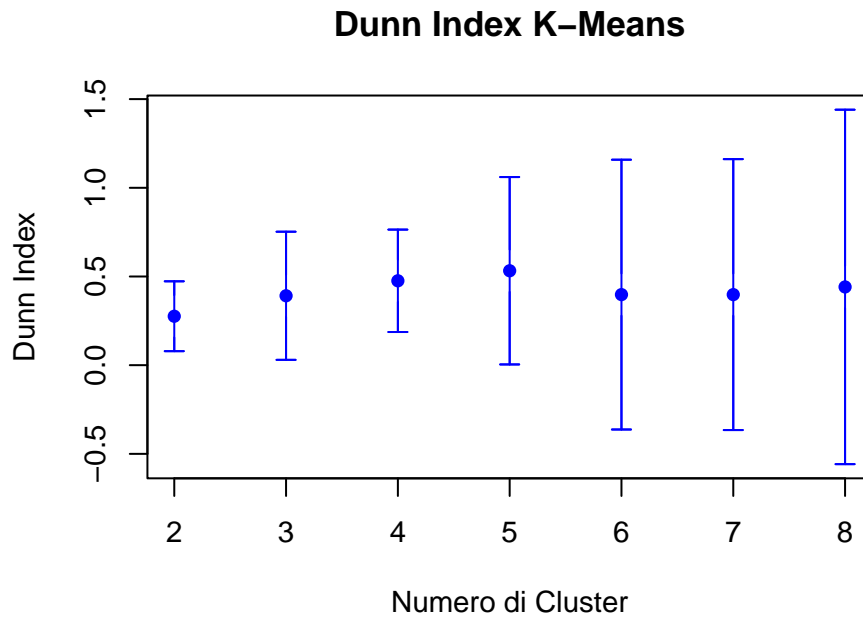
ASW vs k: scelta del numero di gruppi

In Figura è possibile visualizzare l'andamento dell'average silhouette width al variare di K ; un picco è chiaramente individuabile in corrispondenza di $K = 3$, per cui in media l'ASW calcolato è di circa 0.27; lo standard error misurato è inoltre alquanto ridotto, per cui è plausibile attendersi un discreto ASW in corrispondenza di tale numero di cluster. All'aumentare di K non solo in media il valore dei silhouette tende a decrescere drasticamente, ma aumenta altresì, in maniera generale, lo standard error; ne consegue che i risultati ottenuti sono più instabili e meno affidabili.



Dunn Index vs K: scelta del numero di gruppi

Il Dunn index si presta ad un'interpretazione poco netta e immediata; si evidenzia l'ampiezza delle bande di confidenza rappresentate, per cui si osservano valori calcolati del DI piuttosto variabili sulle diverse partizioni C_k . Ad ogni modo, lo standard error misurato non rimane inalterato al variare di K , al crescere del numero di gruppi questo aumenta, motivo per cui si pone quale elemento di giustificazione nella selezione di un valore anziché di un altro. Il DI assume valori maggiori in corrispondenza di $K = 3$: per tale scelta è ragionevole attendersi cluster compatti e ben separati, in cui in media la distanza minima tra osservazioni non all'interno dello stesso gruppo è maggiore della massima differenza intra-cluster. Al contrario delle metriche precedentemente consultate, tuttavia, il Dunn Index suggerisce un numero di gruppi $K = 3$ in maniera meno decisa, l'instabilità dei risultati generati implica che, su taluni campioni casuali, anche per $K > 3$ sia possibile ottenere valori del DI migliori rispetto a quelli ottenuti in corrispondenza del K ottimale.



Dunn Index per diversi numeri di cluster: 0.2758069 0.3912878 0.4757175 0.5322949 0.397844

9.6. Conclusioni clustering e Interpretazione dei risultati

L'analisi di clustering ha reso evidente come clustering non gerarchico con linkage average e distanza di manhattan con 2 cluster si presta meglio a spiegare il problema in analisi.

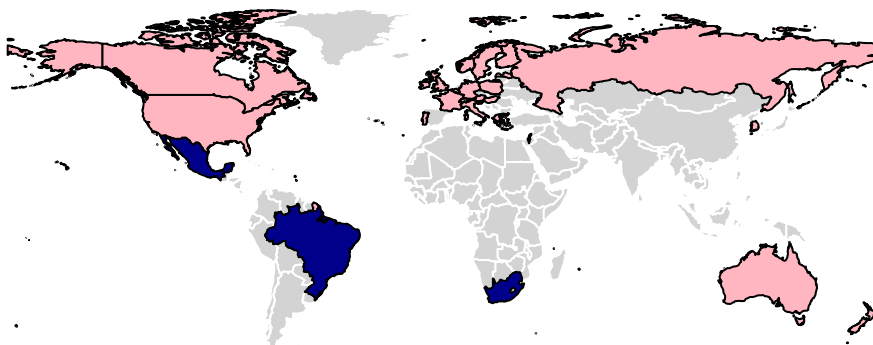
Gruppo 1. Il primo gruppo è quello più numeroso, comprendendo l'85% del campione preso in esame. Si tratta di paesi che mostrano una situazione abitativa con accessi a servizi sanitari piuttosto buona (il 2% della popolazione in media non possiede l'accesso), una buona densità di stanze per persona e un reddito disponibile medio-alto. La ricchezza finanziaria delle famiglie è elevata, così come il tasso di occupazione e il livello di istruzione. Nonostante la qualità della vita apparentemente buona, l'inquinamento dell'aria è significativo. La durata della vita è relativamente alta, e c'è un alto grado di soddisfazione nella vita.

Gruppo 2. Il secondo gruppo è quello che contiene un numero ridotto di osservazioni (solo 4) facente parte del restante 15% del dataset. I paesi mostrano una situazione abitativa con un elevato numero di strutture di base mancanti, una bassa densità di stanze per persona e un reddito disponibile basso. La ricchezza finanziaria delle famiglie è inferiore rispetto al primo gruppo, così come il tasso di occupazione e il livello di istruzione. L'inquinamento dell'aria è più elevato, e la durata della vita e la soddisfazione nella vita sono leggermente inferiori rispetto al primo gruppo.

Variabili	Gruppo 1	Gruppo 2
Dwellings without basic facilities	2.17	13.600
Rooms per person	1.72	0.875

Variabili	Gruppo 1	Gruppo 2
Household net adjusted disposable income	26524.041	13514.250
Household net financial wealth	53362.00	8330.750
Employment rate	69.00	54.750
Educational attainment	83.04	42.000
Student skills	497.12	406.750
Air pollution	13.16	17.000
Life expectancy	80.21	71.275
Life satisfaction	6.59	5.875

Infine, nella mappa geografica si evidenziano i due cluster analizzati.



10. Conclusioni e limiti nelle analisi

La vita si mostra finanziariamente precaria in molte famiglie dei paesi. Mentre il 12% della popolazione vive in povertà relativa al reddito, la percentuale di coloro che segnalano difficoltà a far fronte alle spese nelle nazioni europee dell'OCSE è quasi il doppio, arrivando al 21%. La ricchezza media delle famiglie è però aumentata, in media, nei paesi in cui esistono dati. Una famiglia su cinque a reddito spende una parte considerevole del proprio reddito disponibile (in media il 20%) per le spese abitative, lasciando poco per le esigenze essenziali della vita. Molte delle famiglie in alcuni stati mostrano difficoltà nell'accesso ai servizi sanitari primari e vivono in condizioni pietose (alcuni stati registrano il 37% della popolazione). La qualità della vita riguarda anche le relazioni. Nei paesi, le persone trascorrono circa 14 ore a settimana interagendo con amici e familiari e alla cura della propria persona. Inoltre, una persona su

10 afferma di non avere parenti o amici su cui contare in caso di bisogno (10% vs 90%). Inoltre, sebbene la soddisfazione nella vita sembra essere migliorata in media dal 2013, una considerevole parte della popolazione nei paesi riporta livelli molto bassi di soddisfazione nella vita.

Tuttavia, è importante considerare i seguenti limiti nella valutazione dei risultati e delle conclusioni di questo progetto:

1. *Limitazioni dei dati*: È possibile che i dati utilizzati per l'analisi siano limitati in termini di dimensione del campione, copertura temporale o completezza delle informazioni. Queste limitazioni possono influire sulla generalizzabilità dei risultati e sulla precisione delle stime ottenute.
2. *Manca di variabili rilevanti*: È possibile che alcune variabili potenzialmente influenti sul problema non siano state incluse nell'analisi a causa di limitazioni dei dati o di altre ragioni. La mancanza di queste variabili potrebbe limitare la capacità dei modelli di spiegare completamente i fattori che contribuiscono a una vita migliore.
3. *Potenziale confondimento*: Nonostante gli sforzi per controllare le variabili confondenti, potrebbe essere presente un potenziale confondimento non misurato o non considerato. Ciò potrebbe influire sugli effetti stimati delle variabili di interesse e sulla validità delle conclusioni.
4. *Limitazioni dell'interpretazione causale*: Nonostante gli sforzi per controllare le variabili confondenti, gli studi osservazionali come questo possono presentare limitazioni nella determinazione di relazioni causali tra le variabili di interesse. La natura osservazionale dello studio potrebbe rendere difficile stabilire relazioni di causa-effetto definitive tra gli indici di vita e il benessere generale.

Fine prima parte

Affiliation:

Carmela Pia Senatore

Università degli studi di Salerno

Matricola: 0522501721

E-mail: c.senatore50@studenti.unisa.it