UNIVERSITÀ DEGLI STUDI DELL'INSUBRIA

Corso di Laurea Triennale In Economia e Management

L'IMPORTANZA DEI RICERCATORI PER LO SVILUPPO SCIENTIFICO E TECNOLOGICO

Lavoro svolto da: -Popaj Jessica 745293

jpopaj@studenti.uninsubria.it

-Vilei Syria 744924
svilei@studenti.uninsubria.it

-Premoli Tommaso 744425
tpremoli@studenti.uninsubria.it

CONTENUTI

ABSTRACT

L'analisi presentata si propone di esaminare l'influenza di alcune variabili selezionate sul numero di ricercatori in diversi Stati dei Paesi appartenenti all'OCSE.

Questo tema è di grande attualità, nelle imprese la ricerca è importantissima per le società commerciali, organizzazioni e aziende.

Analizzando alcune variabili tra cui l'investimento ponderato rispetto al PIL, la domanda di brevetti, la spesa per l'educazione terziaria, il tasso di disoccupazione, l'educazione e il PIL pro capite, sono emerse alcune tendenze significative.

Nella conclusione, si è stabilito che le variabili significative per il numero di ricercatori per Paese sono l'investimento in R&S, la spesa in educazione e il PIL pro-capite. L'analisi mostra che esistono macro-zone più predisposte alla Ricerca e Sviluppo come l'Asia-Pacifico, l'Europa Occidentale e Settentrionale e l'America del Nord. Invece, altre macro-zone come l'America Centrale e Meridionale, l'Europa mediterranea e dei Paesi baltici sono meno pronte ad un futuro più innovativo dal punto di vista scientifico. I Paesi asiatici e nordici si distinguono per l'impegno nella ricerca, mentre altre regioni affrontano sfide economiche e scientifiche.

È necessario promuovere una distribuzione equa delle risorse e una cooperazione internazionale inclusiva per ridurre le disuguaglianze socioeconomiche tra i Paesi, sostenere gli sforzi dei Paesi meno sviluppati, affrontare le sfide globali condivise e favorire l'innovazione scientifica. Questo crea un ambiente globale in cui tutti i Paesi hanno l'opportunità di partecipare, contribuire e beneficiare dello sviluppo scientifico e tecnologico.

Infine, i risultati dell'analisi sono Stati confrontati con il Global Innovation Index, un indice che classifica le economie più innovative del mondo. Paesi come Svizzera, USA, Svezia e Regno Unito si sono posizionati in alto nella classifica grazie all'investimento in R&S grazie alla collaborazione tra università e aziende e ad una solida ricerca scientifica. L'Italia si trova al 28º posto.

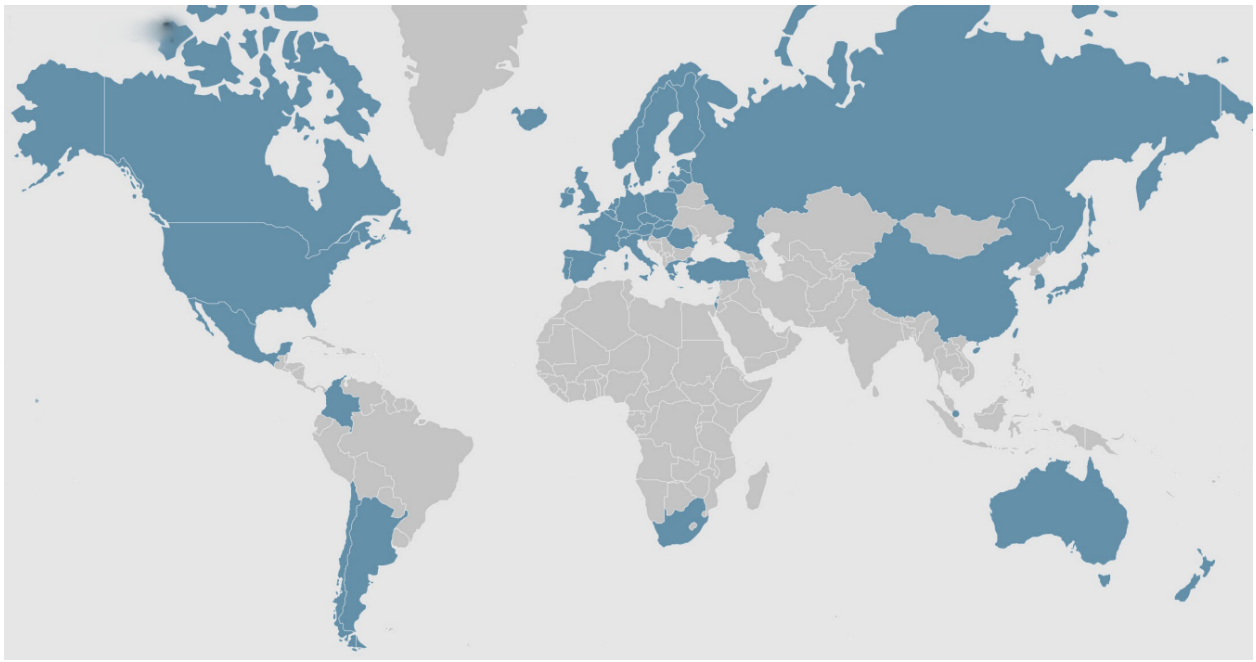
La Cina, il Giappone, Corea del Sud e USA sono i Paesi in cui si concentra un'alta attività e un'importante concentrazione di risorse, competenze e innovazione in diversi settori; infatti, rappresentano le principali aree geografiche in cui si sviluppano e prosperano le industrie, le istituzioni di ricerca e le iniziative imprenditoriali legate all'innovazione.

1. INTRODUZIONE E MOTIVAZIONE

L'obiettivo dell'analisi che verrà presentata è di comprendere come alcune variabili selezionate influenzino il numero di ricercatori in alcuni Stati, tra cui quelli che fanno parte dell'OCSE (Organizzazione per la Cooperazione e lo Sviluppo Economico). L'OCSE è un'organizzazione internazionale composta da 38 Paesi membri impegnati a promuovere politiche economiche e sociali volte a migliorare il benessere dei cittadini e a incentivare la crescita economica sostenibile. Per raggiungere l'obiettivo dell'analisi, sono Stati selezionati 44 Paesi, di cui 38 appartenenti all'OCSE e 6 Paesi importanti dal punto di vista economico, sociale e geopolitico, tra cui Cina, Taiwan, Romania, Russia, Singapore e Sud Africa.

La nostra analisi si basa sui dati del 2020, anno in cui i Paesi membri dell'OCSE erano 37. La Costa Rica, essendo stata ammessa nell'OCSE solo nel 2021, non era inclusa nell'analisi perché i dati relativi a questo Paese non erano disponibili sul sito dell'OCSE al momento dell'analisi.

Di seguito la mappa con i Paesi selezionati:



Sono Stati selezionati i Paesi dell'OCSE per l'analisi perché l'Organizzazione per la Cooperazione e lo Sviluppo Economico è un'organizzazione internazionale che riunisce i Paesi più sviluppati del mondo, impegnati a promuovere politiche economiche e sociali mirate ad incrementare la ricchezza di un territorio. Di conseguenza, questi Paesi rappresentano una base di confronto ideale per studiare come la ricerca e lo sviluppo influenzano l'economia e la società, grazie al loro alto livello di sviluppo e innovazione. Inoltre, l'OCSE è un'organizzazione che raccoglie e pubblica molti dati sulle attività di ricerca e sviluppo dei suoi membri, il quale ha fornito una vasta gamma di informazioni utili per l'analisi elaborata.

I ricercatori sono professionisti altamente qualificati che lavorano in vari ambiti, tra cui l'industria, l'accademia, il governo e le organizzazioni non profit. La loro attività consiste nell'effettuare ricerche e studi per acquisire nuove conoscenze e sviluppare nuove tecnologie, prodotti e servizi. La ricerca può essere di natura scientifica, tecnologica, medica, sociale o umanistica, e può avere

applicazioni in vari settori, come ad esempio la medicina, l'energia, l'informatica, l'ambiente, l'agricoltura e molti altri.

Le variabili indipendenti selezionate per questa analisi sono: investimento ponderato rispetto al PIL, domanda di brevetto, spesa per l'educazione terziaria, tasso di disoccupazione, educazione, GDP per Capita. L'obiettivo definito inizialmente potrebbe, in aggiunta, consentire di individuare le possibili leve di intervento per favorire la crescita e lo sviluppo della ricerca scientifica e tecnologica.

Attraverso l'analisi di queste variabili, si potrebbe identificare quali fattori sono più importanti per attrarre ricercatori in un determinato Paese e, di conseguenza, sviluppare una strategia per promuovere la ricerca scientifica e tecnologica nelle aree con carenza di ricercatori. Così facendo, si contribuirebbe al miglioramento dell'innovazione, della crescita economica e della competitività del Paese.

2. ANALISI DEI DATI

L'analisi verrà svolta utilizzando un set di dati ricavati a seguito di alcune ricerche (sitografia presente in seguito, anno di riferimento:2020), ognuno dei quali svolge un ruolo fondamentale nella determinazione dell'obbiettivo finale, ovvero, cogliere quali sono le forze che hanno l'impatto più significativo nella determinazione del numero di ricercatori.
La ricerca, oggi, rappresenta uno degli strumenti più importanti della sfera economica, poiché grazie ad essa è possibile raggiungere livelli di benessere elevati, dei quali può beneficiare l'intera società. In aggiunta, lo sviluppo contribuisce a concretizzare le nuove scoperte e innovazioni. Inoltre, aiuta a preservare una posizione differenziata e soprattutto di leader in una sfida comune verso la crescita. Se si dovesse fare un focus dettagliato sui dati raccolti potremmo così riassumere:

Numero di ricercatori:
I *ricercatori* sono soggetti qualificati che si occupano di una o più branche della scienza, impegnati continuamente nella scoperta di nuove particolarità, idee e metodologie. Scoprire nuove cose è sinonimo di cambiamento nonché condurre uno studio strutturato da diversi passi, il quale ha come unico fine trovare un output che produca stupore, innovazione e cambiamento. Questo indicatore è misurato per 1000 persone occupate, ad esempio in Argentina nell'anno 2020 ci sono Stati 3,287x1000 persone occupate nel settore della ricerca.

Spesa interna lorda per R&S:
La *Spesa interna lorda* rappresenta l'insieme degli investimenti a breve e a lungo termine che un Paese pone di fare per crescere. Nell'analisi presentata l'area geografica di riferimento è costituita da alcuni dei componenti dell'OCSE, i quali si articolano a loro volta in innumerevoli attori della ricerca come: università, imprese, enti specializzati e istituzioni pubbliche che possono scegliere o meno di investire i loro risparmi nel ramo considerato. Si tratta quindi di eseguire un insieme di lavori sistematici, al fine di accrescere l'insieme delle conoscenze ma anche per tramutare queste conoscenze in invenzioni pratiche. Questo indicatore è misurato a prezzi costanti in USD utilizzando l'anno base 2015 e a parità di potere d'acquisto (PPP) come percentuale del PIL, ad esempio, in Argentina nell'anno 2020 la spesa interna in R&S è stata rappresentata dal 0,52% del PIL.

Domande di brevetto:
Il *brevetto* è sinonimo di punto di forza grazie al quale il titolare ottiene il diritto di monopolio per un periodo di tempo limitato, consistente nel diritto esclusivo di realizzarlo, disporne e farne un uso commerciale, escludendo tutti gli altri soggetti dal suo utilizzo in quanto non autorizzati. La domanda di brevetto è una procedura online che consente in caso di esito positivo il beneficio di poter godere di esso e quindi in termini pratici di poter beneficiare di differenziabilità ed esclusività sul mercato. In questo caso vengono contate le unità di brevetti posseduti nell'anno 2020 sul PIL espresso in miliardi, per esempio le imprese argentine nel 2020 possedevano brevetti per un totale di 930/389,064 =2,3904.

Spesa per l'educazione terziaria:

La *spesa per l'educazione terziaria* comprende la spesa totale per il più alto livello di istruzione; quindi, include la spesa per le scuole, università e altri istituti di formazione di carattere privato. A livello terziario questi tipi di istituti vengono finanziati con fondi pubblici, anche se non manca il contributo dell'ambito privato. In questo caso si esprime come percentuale della spesa totale nell'istruzione. Ad esempio, in Spagna emerge che nell'anno 2020 l'0,835% della spesa per l'istruzione è stata di carattere terziario.

Educazione:
Una delle misure più significative dello sviluppo di una nazione è il suo sistema educativo. Nelle nazioni sviluppate, le opportunità educative sono abbondanti e la maggior parte della popolazione adulta è alfabetizzata e possiede almeno un'istruzione superiore di base. Questa condizione può avere a che fare con il numero di ricercatori in un Paese, poiché il ricercatore è una figura che possiede un livello di istruzione elevato. In aggiunta, nelle nazioni ancora in via di sviluppo, i tassi di alfabetizzazione e il numero di persone che hanno completato la scuola superiore tendono entrambi ad essere inferiori. Questo indicatore viene misurato in percentuale dei cittadini che hanno completato un qualsiasi ciclo di educazione terziaria. Ad esempio, in Colombia il 24,6% dei cittadini possiede un livello di istruzione di tipo terziario.

Tasso di disoccupazione:
Il *tasso di disoccupazione* è un indicatore Statistico del mercato del lavoro che quantifica l'incidenza della popolazione che ha un'occupazione sul totale della popolazione in un Paese e si calcola come rapporto percentuale tra il numero di persone occupate e la popolazione. Ad esempio, l'Italia è un Paese che nell'anno 2020 ha registrato un tasso di disoccupazione pari a 9,3%.

GDP per capita:
Il *GDP per capita*, chiamato anche Prodotto Interno Lordo è una misura macroeconomica che indica il valore aggregato, di tutti i beni e servizi finali realizzati all' interno di uno Sato, in un determinato arco temporale, nonché nel caso studiato l'anno solare. Ad esempio, in Colombia, nell'anno 2020 si è registrato un GDP pari a 5362,062 euro.

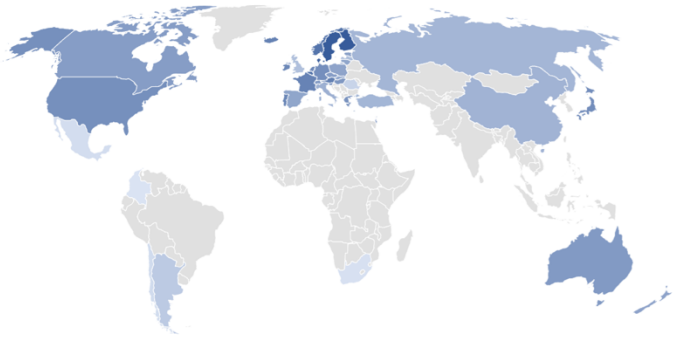
3. STATISTICA DESCRITTIVA

Il set di dati preso in esame è Stato raccolto dal sito ufficiale dell'OCSE (Organizzazione per la Cooperazione e lo Sviluppo Economico) e dal sito ufficiale della World Bank Data. Come affermato precedentemente, tutti i dati si riferiscono all'anno 2020.
Il dataset finale che è Stato studiato in questa analisi è allegato.
Il primo step che viene svolto in questa indagine è lo studio e l'approfondimento dell'analisi univariata e delle Statistiche descrittive.
Per questo motivo viene mostrata una tabella con tutti i fattori più rilevanti dell'analisi e le variabili prese in esame.

	Media	Mediana	Minimo	Massimo	Range	IQR	Deviazione Standard	Varianza	Coefficiente di Gini
Researchers	8,591	8,80	0,500	16,605	16.105	4,90425	4,096145	16,7784	0,2660657
Invest_on_GDP	2,006	1,847	0,289	5,706	5,417	1,79025	1,187326	1,409743	0,322032
Patent_App	11,0789	2,7125	0,1763	109,7341	109,5578	3,43106	25,0	623,8605	0,7350632
Edu_Spending	0,9809	0,904	0,418	2,2	1,782	0,46075	0,4064175	0,165175	0,2233691
Education	39,85	40,350	15,8	69,9	54,10	18,875	12,07229	145,7402	0,1699172
Unemployment_ Rate	7,375	6,213	2,55	28,7	26,15	3,6045	4,771077	22,76318	0,2978976
GDP_percapita	36.494	31.757	5.363	118.084	112.721	32700,57	24.640,26	607.142.278	0,3631954

Tramite l'analisi della Statistica descrittiva è possibile iniziare a elaborare delle considerazioni basate sui risultati ottenuti.

La prima analisi viene rivolta al numero di ricercatori ogni mille abitanti. La media è di 9,591 che è un valore simile alla mediana; dunque, non esistono outlier.
I Paesi con meno ricercatori ogni mille abitanti sono la Colombia (0,5), Sud Africa, Messico e Cile. Dall'altra parte, i Paesi con la maggior densità di ricercatori sono la Korea del Sud con 16,605 ricercatori ogni 1000 persone, Svezia, Taiwan, Danimarca, Finlandia e Norvegia.

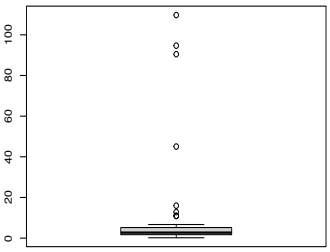


Osservando la mappa, la prima conclusione a cui si può giungere immediatamente è che il numero di ricercatori nel mondo non è distribuito uniformemente. Infatti, i Paesi del blocco scandinavo hanno un numero di ricercatori significativamente maggiore rispetto a Paesi economicamente meno avanzati come i Paesi del Centro, Sud America e dell'Africa.

L'analisi passa alla spesa interna lorda in Ricerca e Sviluppo. Anche per questa variabile la media che è pari a 2,006 è simile alla mediana.

I Paesi che investono la minor parte del proprio PIL in R&S sono la Colombia con lo 0,289% di spesa, Messico, Cile, Romania e Argentina. Invece, i Paesi che investono di più per il proprio progresso scientifico sono Israele (5,706%), Korea, Taiwan, Svezia e gli Stati Uniti.
Israele è dal 2016 nella prima posizione per quanto riguarda il maggiore impegno in R&S. Lo Stato israeliano è caratterizzato dal fatto di avere una visione strategica e imprenditoriale e un'innata capacità di innovazione soprattutto grazie a università al di sopra della media mondiale. Gli Stati Uniti si trovano al top per quanto riguarda la spesa assoluta, ma per percentuale del PIL si trovano al quinto posto; perciò, ci sono ancora ampi margini di miglioramento. Una menzione si può fare anche alla Cina, che con gli USA, rappresentano il 63% dell'aumento totale della spesa per il progresso scientifico negli ultimi anni.
Come per il numero di ricercatori, in Sud America ed in Messico sono stati ottenuti risultati notevolmente bassi. Per fare un esempio, la Corea del Sud investe il 4,8% in R&S contro lo 0,3% della Colombia, del Messico e del Cile e lo 0,5% dell'Argentina.

Ora verrà realizzata l'analisi delle domande di brevetto in rapporto del PIL che genera ogni Paese. La media che è uguale a 11,08 si distacca significativamente dalla mediana, perciò siamo in presenza di outlier. Ciò significa che la distribuzione dei dati è asimmetrica verso destra e che ci sono valori elevati che spingono la media verso l'alto.



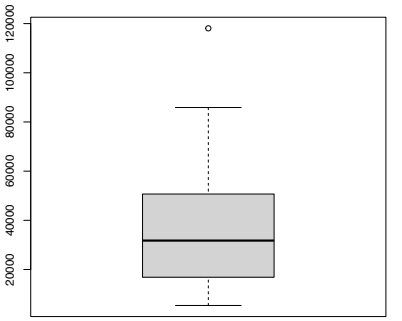
I Paesi con il "*rapporto brevetti/PIL*" minore sono l'Irlanda, Regno Unito, Estonia, Messico e Spagna. Questi Paesi registrano un basso livello di innovazione e una probabile maggiore dipendenza da tecnologie esterne. I Paesi con il rapporto più alto, invece, sono la Corea con un valore di 109,73, Taiwan, Cina e Giappone.
Il range tra i valori è estremamente elevato e, per questo motivo, si allega il grafico del box plot. Inoltre, la varianza assume un valore molto elevato, pari a 623,86 e, quindi, si assiste ad una grande dispersione dei dati.
Dall'analisi è possibile osservare un predominio asiatico per questa variabile. Infatti, nel 2021 la *World Intellectual Property Organization* ha registrato 3,4 milioni di richieste di brevetto, di cui due terzi arrivano dall'Asia. La Cina si trova al top con 1,6 milioni di domande.
Per quanto riguarda il 2020 sono state depositate numerose domande di brevetto per nuove tecnologie volte a contrastare la pandemia di COVID-19. Cina, Stati Uniti, Corea del Sud, Germania e Regno Unito sono stati i principali innovatori.
Attualmente, una questione molto rilevante nel mercato globale è quella dei microchip. La Cina nel 2022 ha superato Washington nella richiesta di ottenimento di proprietà intellettuale per i semiconduttori. Nonostante ciò, la nazione in cui viene prodotta la quantità maggiore di microchip è Taiwan con 4.739 brevetti depositati. Non a caso Pechino adotta un atteggiamento minaccioso e potenzialmente bellicoso nei confronti dell'isola di Taiwan, che non viene riconosciuta come uno Stato indipendente dal governo cinese.

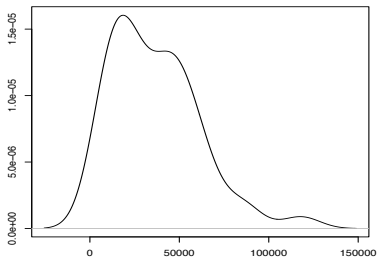
Dopodiché ci si appresta ad analizzare la spesa nell'educazione in percentuale del PIL. In media gli Stati investono lo 0,98% del PIL nella formazione terziaria. I Paesi che investono meno sono Lussemburgo con solamente lo 0,418%, Colombia, Giappone, Regno Unito e Italia. Dall'altra parte, Singapore con il 2,2%, Cina, Norvegia, Taiwan e Austria promuovono l'educazione al di sopra della media globale.

Collegato a ciò, viene analizzato il livello di educazione terziaria per Stato che può essere descritto come uno dei migliori indici per valutare i piani di sviluppo di una nazione. In media ogni Paese ha il 39,85% di laureati. I Paesi più “istruiti” sono Singapore con il 69,9% di cittadini con una laurea o un diploma equivalente, Canada, Russia e Taiwan. Al contrario, il Sud Africa, Messico e Italia sono le nazioni con la minor percentuale di laureati.
Un'opinione pressoché unanime tra gli esperti è che investire nell'istruzione terziaria sia uno dei fattori più importanti per la crescita economica e lo sviluppo a lungo termine di un Paese. I dati dell’OCSE confermano questa tesi.

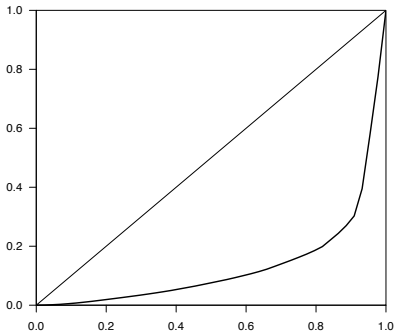
Altre due variabili che sono state prese come oggetto di studio sono il tasso di disoccupazione e il PIL pro capite che consentono di valutare la competitività economica di una nazione.
I 44 Paesi in media hanno il 7,37% di tasso di disoccupazione. Le nazioni con il maggior tasso sono il Sud Africa con il 28,7%, Grecia, Colombia e Spagna. I Paesi con una quantità di occupati maggiori, invece, sono la Repubblica Ceca, Giappone, Singapore, Polonia e Germania.
Esaminando il grafico, si può notare come alcune zone come l’America Latina, il Sud Africa e l’Europa Mediterranea siano colpiti particolarmente da questo problema.

Per concludere, l’analisi viene svolta per il PIL pro capite. I Paesi in media hanno un PIL per persona di 36.494 dollari. Il Lussemburgo si trova in cima a questa classifica seguito da Svizzera e Irlanda. Le nazioni in fondo alla graduatoria sono Colombia, Sud Africa e Messico. La maggior parte dei Paesi più ricchi del mondo si concentra principalmente in Europa e in America del Nord, dove il PIL pro-capite raggiunge rispettivamente i 34.500 e 59.000 dollari. In questa indagine consideriamo il Lussemburgo come un outlier in quanto il valore del PIL pro-capite si distacca nettamente dagli altri, come si può notare dal box plot.
Inoltre, è possibile notare che i dati sono altamente dispersi tra di loro e, di conseguenza, si ha una varianza e una deviazione standard elevate. Pertanto, si allega la curva di densità che rappresenta la distribuzione di probabilità di una variabile continua.





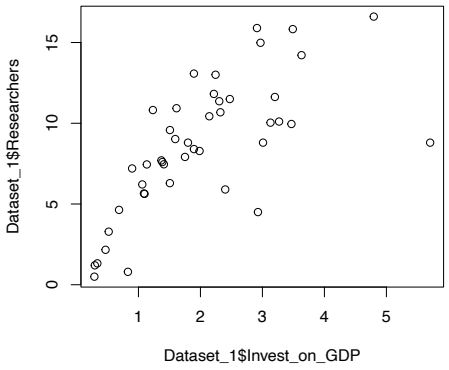
Infine, per quanto riguarda il *Coefficiente di Gini*, si può immediatamente notare come mediamente tutti i dati presentino una distribuzione abbastanza omogenea in quanto compresi tra 0 e 0,35. L'unica variabile con un coefficiente elevato riguarda la domanda di brevetti, pari a 0,735. Ciò significa che pochi Paesi hanno presentato una grande percentuale di domande di brevetto, mentre gli altri ne hanno presentati una quantità relativamente piccola. Pertanto, si allega il grafico della Curva di Lorenz.



Dopo aver analizzato singolarmente ogni variabile, lo studio verrà svolto per analizzare le eventuali dipendenze tra due variabili. Ai fini della ricerca viene riportata la tabella con gli indici di correlazione di Pearson.

Coefficienti di Correlazione	Researchers	Invest_on_GDP	Patent_App	Edu_Spending	Education	Unemployment_Rate	GDP_percapita
Researchers	1	0,6956306	0,2848655	0,4069525	0,4814344	-0,4158348	0,494242
Invest_on_GDP		1	0,4405414	0,2527862	0,493081	-0,3945079	0,3903168
Patent_App			1	0,1766262	0,3263952	-0,2236709	-0,1367245
Edu_Spending				1	0,314405	-0,1879829	0,165877
Education					1	-0,4123959	0,5303689
Unemployment_Rate						1	-0,3322000
GDP_percapita							1

L'indice di correlazione più elevato e più significativo coincide con la variabile "Researchers" e "Invest_on_GDP". L'indice di correlazione è pari a ρxy=0,6956. Ciò significa che c'è un'intensa relazione lineare diretta e, perciò, si ha una correlazione positiva tra queste due variabili.
Pertanto, si può affermare che i Paesi che investono maggiormente in R&S, tendenzialmente possiedono più ricercatori.
Viene riportato il grafico di dispersione che rappresenta la relazione tra le due variabili quantitative.

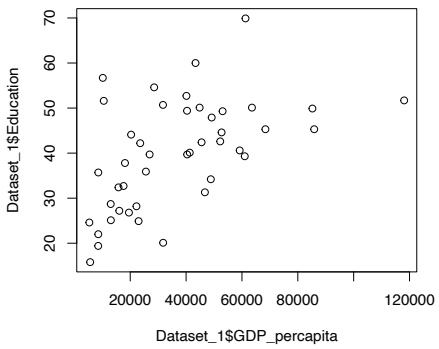


Per confermare questa conclusione, si allega il risultato ottenuto dell'analisi della regressione lineare semplice utilizzando come variabile indipendente la spesa in ricerca e sviluppo.

```
Call:
lm(formula = Dataset_1$Researchers ~ Dataset_1$Invest_on_GDP)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6695 -1.5705  0.2347  1.7413  5.1317

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             3.7760     0.8889   4.248 0.000117 ***
Dataset_1$Invest_on_GDP   2.3998     0.3824   6.275 1.59e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.977 on 42 degrees of freedom
Multiple R-squared:  0.4839,    Adjusted R-squared:  0.4716
F-statistic: 39.38 on 1 and 42 DF,  p-value: 1.595e-07
```

Il modello spiega il 48,39% della variabilità totale come indicato dal coefficiente di determinazione. Ciò rappresenta una capacità di spiegare e predire i valori osservati abbastanza buona. Inoltre, dato che il p-value è estremamente inferiore al livello minore di significatività pari a 0,1%, è possibile dichiarare che la variabile “Invest_on_GDP” è statisticamente significativa. Di conseguenza, come già affermato dal coefficiente di correlazione di Pearson, esiste una relazione significativa tra le due variabili. Infine, da quanto mostrato dall’output, si può asserire che per ogni incremento unitario della spesa in R&S in percentuale del PIL, ci si aspetta un aumento di 2,4 ricercatori ogni mille abitanti.

Un altro indice di correlazione considerevole corrisponde con le variabili “Education” e “GDP_percapita”. L’indice è pari a ρxy=0,5304 e quindi si ha una correlazione buona. Questo dato significa che le nazioni con un PIL pro capite superiore hanno maggiori probabilità di avere un livello di educazione più elevato.
Il grafico di dispersione, infatti, rappresenta la relazione lineare diretta e dimostra che all’aumentare per PIL pro-capite, il livello di educazione tende ad aumentare.

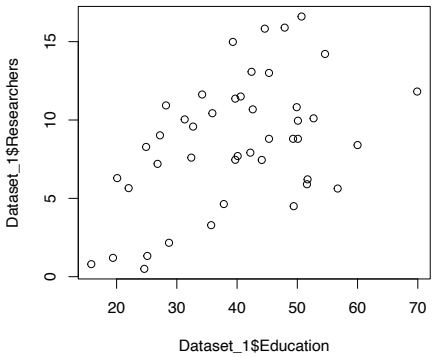


Anche per questo caso si analizza il modello lineare semplice per verificare che la variabile indipendente del PIL pro-capite abbia un impatto rilevante sulla variabile che riguarda il livello di educazione di ogni Paese.

```
Call:
lm(formula = Dataset_1$Education ~ Dataset_1$GDP_percapita)

Residuals:
    Min      1Q  Median      3Q     Max
-18.524  -7.519  -1.247   5.286  23.690

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             3.036e+01  2.812e+00  10.798 1.08e-13 ***
Dataset_1$GDP_percapita 2.599e-04  6.409e-05   4.054 0.000213 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.36 on 42 degrees of freedom
Multiple R-squared:  0.2813,    Adjusted R-squared:  0.2642
F-statistic: 16.44 on 1 and 42 DF,  p-value: 0.000213
```

Questo modello spiega debolmente la variabilità totale perché il coefficiente di determinazione è pari al 28,13%. Il p-value, come nel caso precedente, è statisticamente significativo in quanto il suo valore è inferiore ad ogni usuale livello di significatività. Prendendo solamente "GDP_percapita" come variabile esplicativa, si può affermare che per ogni aumento unitario del PIL pro capite, la percentuale di cittadini che hanno completato un qualsiasi ciclo di educazione terziaria aumenta di 0,0476.

Ai fini dell'analisi, si analizza anche la correlazione tra il livello di educazione del Paese e i numeri di ricercatori ogni mille abitanti. Il coefficiente è pari a 0,4814 che indica un livello di connessione moderata. Ciò significa che i Paesi con un livello di educazione più elevato tendono ad avere un maggior numero di ricercatori ogni mille abitanti.

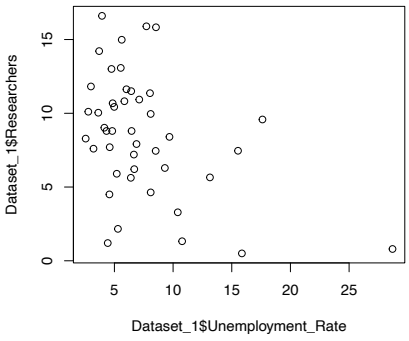



```
Call:
lm(formula = Dataset_1$Researchers ~ Dataset_1$Education)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7174 -3.5168 -0.1662  2.7891  6.4829

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.08134    1.90880    1.09 0.281757
Dataset_1$Education 0.16335    0.04589    3.56 0.000938 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.633 on 42 degrees of freedom
Multiple R-squared:  0.2318,    Adjusted R-squared:  0.2135
F-statistic: 12.67 on 1 and 42 DF,  p-value: 0.0009379
```

Il modello, anche per questo output, indica che la variabilità totale è pari a 23,18% spiegandola debolmente. Il p-value è comunque altamente significativo in quanto è minore di ogni usuale livello di significatività. Si può affermare che per ogni aumento unitario della percentuale di cittadini con educazione terziaria, il numero di ricercatori aumentano in media di 0,16.

Per concludere dalla tabella dei coefficienti è possibile notare che esistono altre correlazioni moderate positive tra la spesa nell'educazione terziaria ed il PIL pro-capite con il numero di ricercatori; il livello di educazione e la spesa in R&S in percentuale del PIL.

È doveroso menzionare che esiste una correlazione moderata negativa tra il tasso di disoccupazione e il numero di ricercatori ogni mille abitanti, pari a -0,4158. Ciò significa che quando il tasso di disoccupazione aumenta, il numero di ricercatori ogni mille abitanti tende a diminuire, e viceversa in quanto la relazione lineare è inversa.



```
Call:
lm(formula = Dataset_1$Researchers ~ Dataset_1$Unemployment_Rate)

Residuals:
    Min      1Q  Median      3Q     Max
-8.4391 -2.1415 -0.1519  2.3245  7.6562

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                   11.2234     1.0547  10.642  1.7e-13 ***
Dataset_1$Unemployment_Rate  -0.3570     0.1205  -2.963    0.005 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.769 on 42 degrees of freedom
Multiple R-squared:  0.1729,    Adjusted R-squared:  0.1532
F-statistic: 8.781 on 1 and 42 DF,  p-value: 0.004996
```

In conclusione, si può affermare che questo modello spiega solamente il 17,29% della variabilità totale. Il p-value è pari all’1%, perciò è statisticamente significativo. Inoltre, è possibile affermare che per ogni aumento unitario della percentuale del tasso di disoccupazione, il numero di ricercatori diminuisce di 0,357. In altre parole, se aumenta il numero di disoccupati all’interno di una nazione, il numero di ricercatori diminuisce.

4. MODELLO DI REGRESSIONE

Si prosegue l'analisi elaborando un modello di regressione lineare multipla, tenendo in considerazione il numero di ricercatori come variabile dipendente, definita come y, la quale viene influenzata da una serie di variabili indipendenti, con una numerosità pari a K=6, ciascuna delle quali attribuisce un'influenza diversa rivelata successivamente.
Il modello di regressione si può esprimere come segue:

$$Reserchers = \beta_0 + \beta_1 Invest\_on\_GDP + \beta_2 Patent\_App + \beta_3 Edu\_Spending + \beta_4 Education + \beta_5 Unemployment\_Rate + \beta_6 GDP\_percapita.$$

Primo caso

```
Call:
lm(formula = Dataset_1_2_$Researchers ~ Dataset_1_2_$Invest_on_GDP +
    Dataset_1_2_$Patent_App + Dataset_1_2_$Edu_Spending + Dataset_1_2_$Education +
    Dataset_1_2_$Unemployment_Rate + Dataset_1_2_$GDP_percapita,
    data = Dataset_1_2_)

Residuals:
    Min      1Q  Median      3Q     Max
-6.0621 -1.5130  0.1183  1.8581  4.5419

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  2.226e+00  2.169e+00   1.026  0.31143
Dataset_1_2_$Invest_on_GDP   1.701e+00  4.763e-01   3.572  0.00101 **
Dataset_1_2_$Patent_App      7.842e-03  2.198e-02   0.357  0.72329
Dataset_1_2_$Edu_Spending    2.221e+00  1.113e+00   1.995  0.05343 .
Dataset_1_2_$Education      -6.050e-03  4.926e-02  -0.123  0.90292
Dataset_1_2_$Unemployment_Rate -8.019e-02  1.020e-01  -0.786  0.43685
Dataset_1_2_$GDP_percapita   4.159e-05  2.385e-05   1.744  0.08951 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.792 on 37 degrees of freedom
Multiple R-squared:  0.6003,    Adjusted R-squared:  0.5354
F-statistic:  9.26 on 6 and 37 DF,  p-value: 3.31e-06
```

Dal primo studio effettuato, che prende in considerazione tutte le variabili, emerge che ad incidere significativamente sono tre variabili, ovvero: Investment_on_GDP, Education_Spending, e, infine, GDP per capita. Tuttavia, questa affermazione la si può individuare osservando la presenza di '**' accanto al p-value della variabile Investment_on_GDP, la quale afferma che la variabile è significativa ad un livello pari a 0,01.
In seguito, accanto al p-value di Education_Spending e GDP per capita è presente questo simbologia '.' che suggerisce un livello di significatività delle sue variabili pari a 0.1. Le altre componenti dell'analisi non hanno un impatto importante sul numero di ricercatori pertanto non saranno protagoniste rilevanti del modello di regressione. Ad esempio, se si dovesse interpretare il coefficiente della variabile Education si nota subito che ha un valore negativo, pari a -0,3, pertanto, incrementando di un punto percentuale l'educazione, in questo caso di tipo terziario, si produrrebbe una decrescita dei ricercatori mediamente intorno a 0,3 (soggetti), a parità di tutte le altre condizioni. Questo dato non è rilevante ai fine della ricerca ma è una considerazione molto particolare poiché non è immediata ed intuitiva.

Analogamente, lo stesso ragionamento si potrebbe fare per la variabile Unemployment_Rate, in quanto è stimato un coefficiente pari a -1,08, quindi incrementando di un punto percentuale il tasso di disoccupazione si produrrebbe una riduzione di ricercatori mediamente pari a 1(-1,08), a parità di tutte le altre condizioni. In questo caso, è logico pensare che, se il numero di disoccupati in un Paese aumenta allora il numero di ricercatori impiegati di conseguenza diminuisce.
Inoltre, la stima del coefficiente della variabile Investement_on_GDP è pari a 1.701 la quale indica che per ogni incremento unitario della variabile Investment_on_GDP, la variabile researchers aumenta, in media, di 1.701 unità, a parità di ogni altra condizione. Inoltre, analoghe osservazioni si possono fare per la variabile Education_Spending, la cui stima del coefficiente è pari a 2.221, che indica che per ogni aumento unitario della variabile Edu_Spendig, si produce un aumento dei ricercatori in media pari a 2.221, *ceteris paribus*. In ultimo luogo, analogamente a quanto si ravvisa nelle precedenti situazioni, la stima del coefficiente della varibile GDP_percapita è pari a: 0,028 che indica che per un aumento unitario di quest'ultima variabile, si produce un aumento in media di 4.159e-05 unità di ricercatori, a parità di ogni altra condizione.
Per una valutazione del modello nel suo complesso si può prendere come punto di riferimento l'indice R^2,il quale è un indicatore della bontà del modello di regressione, che assume valori compresi tra 0 e 1 la cui formula equivale al rapporto tra la varianza spiegata e la varianza totale campionaria. Nel modello l'indice $R^2 = 0{,}6005$, pertanto il 60% della variazione nel numero di ricercatori è spiegata dal set di variabili che sono state selezionate.
Tuttavia, un ulteriore indicatore che è più allineato al modello di regressione lineare multipla assume il simbolo di $\bar{R}^2$, in quanto consente una qualità migliore del confronto tra modelli di regressione lineare multipla con un numero maggiore di variabili indipendenti. Nell' analisi questo indice assume il valore di 0,5354, il quale è inferiore ad R^2, ed indica che il 53,54% della variazione nel numero di ricercatori impegnati è spiegata dalla variazione delle altre variabili indipendenti, tenendo conto della dimensione campionaria N=44 e del numero di variabili indipendenti K=6.
Successivamente si può calcolare l'F test, per esprimere quanto il modello è significativo ad un livello complessivo, infatti dall' analisi emerge che F=9,26, dove k=6 e i gradi di libertà sono pari a 37. La regola di decisione afferma che si rifiuta l'ipotesi nulla se il p-value del test F è minore degli usuali livelli di significatività. Nel caso analizzato, il p-value del test F pari a 3.31e-06, che è quindi minore di tutti gli usuali livelli di significatività (0.1, 0.05, 0.01, 0.001), di conse guenza si rifiuta l'ipotesi nulla.

```
Analysis of Variance Table

Response: Dataset_1_2_$Researchers
                           Df Sum Sq Mean Sq F value    Pr(>F)
Dataset_1_2_$Invest_on_GDP  1 349.12  349.12 44.7905 7.305e-08 ***
Dataset_1_2_$Patent_App     1   0.42    0.42  0.0535   0.81832
Dataset_1_2_$Edu_Spending   1  42.03   42.03  5.3918   0.02584 *
Dataset_1_2_$Education      1   9.42    9.42  1.2092   0.27860
Dataset_1_2_$Unemployment_Rate  1   8.38    8.38  1.0754   0.30645
Dataset_1_2_$GDP_percapita  1  23.70   23.70  3.0407   0.08951 .
Residuals                  37 288.40    7.79
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tra le misure di variazione compare la 'Somma dei quadrati Totale', che si articola in Somma de i quadrati della regressione, in questo caso pari a 433,07, ed in Somma dei quadrati degli errori pari a 288,40. Infatti, SST=SSR+SSE=433,07+288,40=721,47. Questo studio permette di comprendere che la variazione si compone di una parte che si riferisce alla relazione lineare tra la variabile dipendente e le variabili indipendenti, la quale è misurata dalla SSR, inoltre da una parte che si riferisce a diversi fattori presenti nella relazione lineare tra le stesse variabili. L'indice R^2,spiegato in precedenza, è dato inoltre calcolando SSR/SST=433,07/721,47=0.6003.

Secondo caso

In seguito per poter approfondire l'analisi si potrebbe ricostruire il modello di regressione multipla, includendo, però, solo ed esclusivamente le variabili ritenute significate nel modello precedente, configurando la stessa dimensione campionaria N=44 ma 40 gradi di libertà, poiché K si è ridotto a 3.
Da questo studio emerge un livello di significatività più robusto per ciascuna variabile, questo perché si ottiene un modello più snello, non influenzato da un numero eccessivo di componenti, in quanto viene escluso l'utilizzo di variabili indipendenti non importanti.

```
Call:
lm(formula = Dataset_1_2_$Researchers ~ Dataset_1_2_$Invest_on_GDP +
    Dataset_1_2_$Edu_Spending + Dataset_1_2_$GDP_percapita, data = Dataset_1_2_)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5334 -1.5937  0.3286  1.8402  4.3071

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.087e+00  1.214e+00   0.896   0.3758
Dataset_1_2_$Invest_on_GDP 1.871e+00  3.871e-01   4.832 2.02e-05 ***
Dataset_1_2_$Edu_Spending  2.311e+00  1.056e+00   2.189   0.0345 *
Dataset_1_2_$GDP_percapita 4.066e-05  1.830e-05   2.222   0.0320 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.715 on 40 degrees of freedom
Multiple R-squared:  0.5914,    Adjusted R-squared:  0.5607
F-statistic:  19.3 on 3 and 40 DF,  p-value: 6.767e-08
```

Nello specifico la variabile Investment on GDP, è significativa al livello 0.001 questo è denotato dai tre asterischi '***' posizionati accanto al p-value, infatti quest'ultimo assume valori piccoli poiché la variabile è decisamente significativa. La variabile Edu_spending e la variabile GDP per capita sono significative al livello 0,05 poiché accanto al loro p-value compare questo simbolo '*'.
Tuttavia, come il livello di significatività anche le stime dei coefficienti variano. La stima del coefficiente della prima variabile indipendente è pari a 1.871 questo significa che per un aumento unitario degli investimenti sul PIL i ricercatori aumentano, in media, di 1.871 unità, a parità di tutte le altre condizioni. La stima del coefficiente della variabile Edu_Spending è di 2,3311, che indica che per un incremento unitario nella spesa per l'educazione, i ricercatori aumentano in media di 2,3311 unità, a parità di tutte le altre condizioni. Infine, l'ultima variabile, ovvero il GDP per capita, assume un coefficiente pari a 0,027, che indica che per un incremento unitario GDP per capita allora i ricercatori saranno, in media 0,027 unità in più, a parità di tutte

le altre condizioni. In seguito, l'indice R^2 del modello è pari a 0,5914, molto simile a quello precedente, configurandosi un'altra volta in un modello di regressione multipla '*abbastanza buono'*, poiché indica che il 59,14% della variazione nel numero di ricercatori impegnati è spiegata dalla variazione delle altre 3 variabili indipendenti. Il suo valore è leggermente inferiore dato che le variabili esplicative sono ridotte rispetto al modello precedente. Tuttavia, l'indice $\bar{R}^2$, che penalizza l'uso di variabili indipendenti non importanti, che in questo esempio sono state escluse, assume valori leggermente superiori a quello precedente infatti è pari a 0,5607.
Nel secondo caso analizzato, il p-value del test F pari a 6,676e-08, che è quindi minore di tutti gli usuali livelli di significatività (0.1, 0.05, 0.01, 0.001), di conseguenza si rifiuta l'ipotesi nulla. C'è sufficiente evidenza empirica per affermare che almeno una variabile indipendente del modello influenzi la variabile y, nell'ultimo esempio tutte.

```
Analysis of Variance Table

Response: Dataset_1_2_$Researchers
                          Df Sum Sq Mean Sq F value    Pr(>F)
Dataset_1_2_$Invest_on_GDP  1 349.12  349.12 47.3688 2.734e-08 ***
Dataset_1_2_$GDP_percapita  1  42.22   42.22  5.7287   0.02147 *
Dataset_1_2_$Edu_Spending   1  35.32   35.32  4.7917   0.03449 *
Residuals                  40 294.81    7.37
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

L' analisi della variazione si compone della SSR pari a 349.12+42,22+35,32=426,66, che è leggermente più piccola di quella del modello precedente, in quanto il numero di variabili esplicative sono di meno e quindi la porzione spiegata si riduce, e della SSE pari a 294,81, che leggermente più grande poiché inserendo più variabili la probabilità di sbagliare è più piccola, ma ciò che è fondamentale è tenere in considerazione non la numerosità delle variabili, bensì *la qualità* delle stesse. Infatti, l'indicatore di confronto più importante in questo caso è $\overline{R^2}$, poiché penalizza l'utilizzo di variabili indipendenti non rilevanti; infatti, nel secondo caso il suo valore è maggiore, come conferma di validità e correttezza dell'analisi.

In conclusione, sono stati analizzati due modelli, il primo che includeva tutte le variabili del Dataset, mentre, il secondo che prendeva in considerazione solo le variabili ritenute significative nel primo modello, con lo scopo di capire se esistesse o meno un legame tra le variabili esplicative e la variabile dipendente e di che tipologia di legame si trattasse.

5. CLUSTER ANALYSIS

La cluster analysis, nota anche come analisi dei cluster o analisi di clustering, è una tecnica di analisi dei dati utilizzata per scoprire pattern o strutture nascoste all'interno di un insieme di dati non etichettati. Consiste nel raggruppare gli oggetti o le osservazioni simili tra loro in gruppi chiamati cluster. L'obiettivo principale della cluster analysis è quello di suddividere un insieme di dati in modo tale che gli oggetti all'interno di ciascun cluster siano più simili tra loro rispetto agli oggetti in altri cluster.

La cluster analysis è spesso utilizzata in diversi ambiti, come l'analisi dei dati, la bioinformatica, la ricerca di mercato e molte altre discipline. Di conseguenza alcuni dei suoi utilizzi principali:

-Esplorazione dei dati: la cluster analysis permette di esplorare un insieme di dati per individuare strutture e pattern nascosti. Può aiutare a identificare gruppi omogenei o sottogruppi all'interno dei dati, fornendo una panoramica delle relazioni tra le osservazioni.

-Segmentazione dei clienti: nell'ambito del marketing e della ricerca di mercato, la cluster analysis può essere utilizzata per suddividere i clienti in segmenti omogenei in base a comportamenti di acquisto, preferenze o caratteristiche demografiche. Questo aiuta le aziende a personalizzare le strategie di marketing e a fornire prodotti o servizi mirati ai diversi segmenti.

-Ricerca scientifica: in ambito scientifico, quest'ultima può essere utilizzata per analizzare i dati ottenuti da esperimenti o studi, per identificare gruppi di pazienti con caratteristiche simili o per individuare pattern di comportamento in dati complessi.

-Analisi di sequenze: nella bioinformatica, può essere impiegata per analizzare sequenze di DNA o proteine, al fine di individuare similitudini o relazioni evolutive tra di esse.

Per effettuare la cluster analysis, vengono utilizzati diversi algoritmi, come il k-means e il clustering gerarchico e molti altri. Ogni algoritmo ha il proprio approccio per definire la similarità tra gli oggetti e per assegnarli ai cluster appropriati.

5.1. LA SILHOUETTE

La silhouette è un numero, che viene assegnato ad ogni osservazione, nonché una misura che dipende dal cluster analizzato, sulla base del dataset a disposizione. Innanzitutto, esiste un comando che ci permette di generare la 'silhouette', il quale mostra per ogni osservazione quale è il cluster di appartenenza, quale è il cluster più vicino e, infine, quantifica la silhouette per ogni singola osservazione.

Pertanto, alla luce dello studio bisogna affermare che questo strumento permette di capire in quanti cluster è opportuno suddividere il set di dati presi in analisi. Nel caso analizzato è necessario definire una partizione degli Stati, preferibilmente in aree territoriali, le quali si accomunano per il possesso di determinate caratteristiche, che verranno successivamente messe in evidenza.

Infatti, la silhouette è definita dalla seguente espressione: $S(i) = \frac{b(i)-a(i)}{MAX(a(1);b(i))}$.

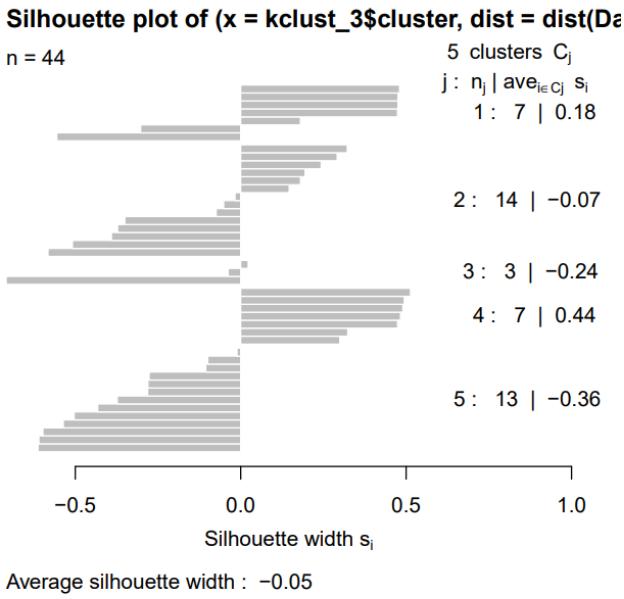
Tuttavia, si può approfondire la sua composizione spiegando i seguenti elementi:

- a(i)=si definisce come la media tra un'osservazione qualsiasi del dataset (osservazionei_esima) e il resto delle osservazioni presenti nello stesso cluster.
- b(i)= si definisce come la distanza tra l'osservazionei_esima ed il vicino più vicino.

A tale proposito l'obbiettivo è quello di partizionare le osservazioni, ovvero i Paesi appartenenti all'OCSE ed altri Paesi importanti selezionati, in gruppi che presentano caratteristiche omogenee, cercando di rendergli più distanti possibili tra di loro, quindi più diversi, ed omogenei al loro interno. La determinazione del numero di cluster è quindi un processo che contempla l'utilizzo della silhouette, come già affermato in precedenza.

L'ideale sarebbe cercare un numero di cluster che permette di generare il più alto valore di questa misura, tuttavia deve esserci un filo conduttore, questo significa indentificare un numero di aree congruo ed opportuno per cogliere delle conclusioni importanti.

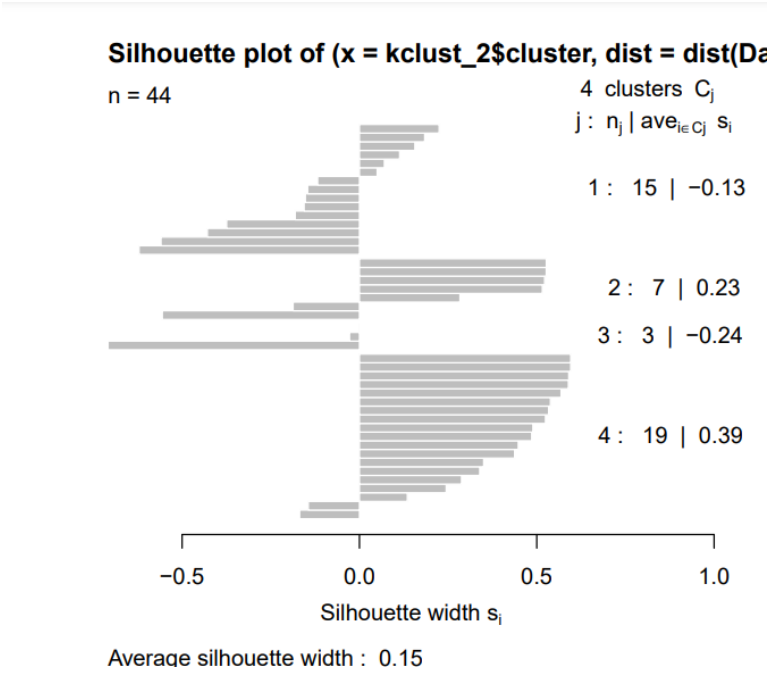
Il primo grafico che si presenta è il seguente:



Il grafico presenta una ripartizione dei dati in cinque cluster aventi numerosità differente. Se si dovesse fare una valutazione individuale del cluster si potrebbero immediatamente identificare tre casi particolari. Il cluster numero 2 possiede una numerosità pari a 14 elementi ed ha una silhouette negativa pari a -0,07, in questo caso la distanza intra cluster è leggermente maggiore della distanza tra cluster, ed è per questo che assume valori negativi.

Analoghe considerazioni possono riferirsi ai cluster numero 3 e 5, con un rilievo differente. Tuttavia, questi dati empirici sono tutti confermati dalla composizione del grafico; di fatti una buona parte delle osservazioni contenute nei vari cluster hanno silhouette negativa, dato che le barre grigie sono posizionare a sinistra del grafico. Pertanto, questo non è un buon risultato dato che essa deve assumere valori più alti possibili per confermare la ragionevolezza del processo di clustering. In conclusione, la media ponderata dell'indicatore è pari -0,05, un valore negativo che suggerisce di riformulare il numero dei cluster.

A tal proposito si procede con il seguente grafico:

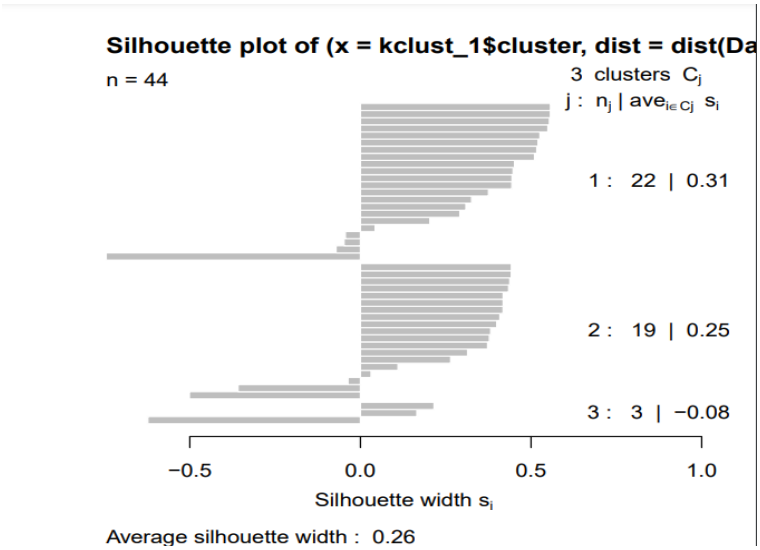


Il grafico presenta i dati raggruppati in quattro cluster. Il primo ha una numerosità pari a 15 con una silhouette negativa equivalente a -0,13. Il secondo, contenente 7 osservazioni, possiede una silhouette pari a 0,23. Il gruppo successivo conta solo 3 osservazioni e una silhouette pari a -0,24. In conclusione, il quarto cluster ha una numerosità di 19 osservazione e mostra una straordinaria silhouette di 0,39. Per concludere la silhouette media ponderata equivale a 0,15 un valore che non è eccellente ma sicuramente migliore rispetto al caso precedente. Infine, anche se in alcuni casi la distanza tra le osservazioni del medesimo cluster supera la distanza tra cluster, il valore finale è comunque positivo. Si è nel caso in cui si potrebbe procedere con un'analisi di clustering, ma per essere più scrupolosi e precisi si prosegue con un ulteriore tentativo.

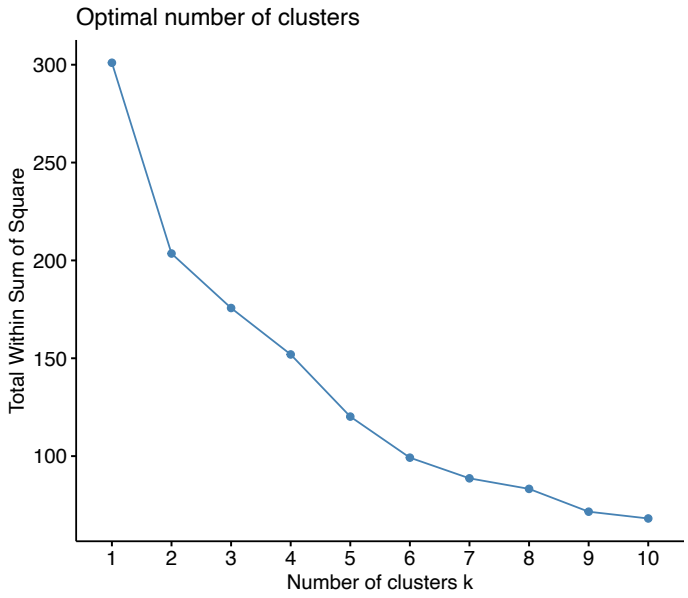
L'ultimo grafico che si illustra in basso ci permette di concludere la selezione del numero di cluster, con la quale condurre l'analisi. Sono Stati elaborati tre cluster, tra i quali i primi due contengono una buona parte delle osservazioni, contrariamente all' ultimo cluster che conta, invece, di solo tre osservazioni. La silhouette è positiva nei primi due casi mentre assume valori negativi nell' ultimo cluster. Tuttavia, queste considerazioni potrebbero alludere alla presenza di tre aree geografiche, che per fattori di diversa natura, come ad esempio fattori economici o politici, possiedono caratteristiche molto simili, ed è per questo che vengono ripartite come segue. Tali considerazioni troveranno modo di essere confermate nei successivi procedimenti.

Alla luce di quanto detto possiamo concludere che il set di dati verrà ripartito in tre cluster, con una numerosità differente, i quali presentano una silhouette media ponderata pari a 0,26. Tale valore è certamente il più alto tra quelli che si ottengono ripartendo il set di dati in 3 o 5 gruppi, quindi il migliore.

Il grafico decisivo è il seguente:



A conferma di quanto osservato dallo studio del metodo della Silhouette, si riporta il grafico dell'Elbow Method. In particolare, il punto rilevante è individuato dalla creazione di un gomito nel grafico. Tale punto rappresenta il numero ottimale di cluster per il dataset analizzato, poiché corrisponde al punto in cui l'incremento della varianza spiegata dai cluster successivi diventa meno significativo. In questa analisi il punto rilevante viene identificato nel valore 3.



5.2. K-MEANS CLUSTERING

L'analisi procede con lo studio e l'interpretazione dei risultati ottenuti tramite il processo di clustering. Lo scopo di questo step è stato quello di riscontrare similarità tra le osservazioni per trovare una struttura intrinseca ai dati. Si presume, infatti, che le osservazioni possano essere raggruppate in categorie naturali, cioè si desidera verificare l'ipotesi che tra le entità esistano sottogruppi differenziati e separati.

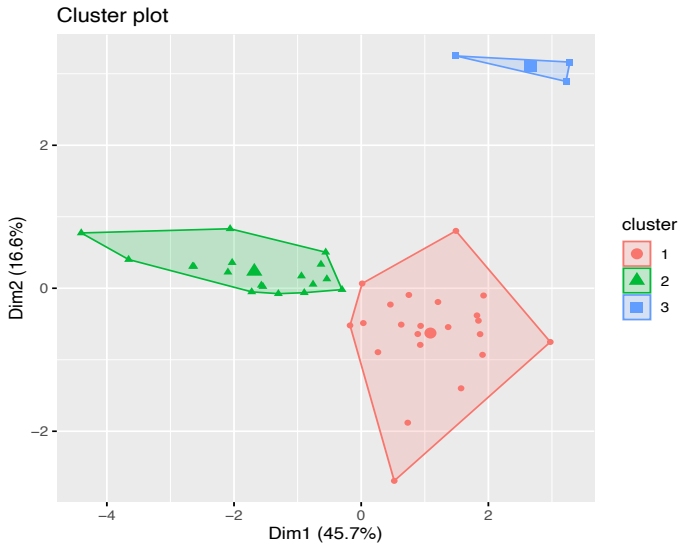
La prima tipologia di clustering che verrà analizzata è il K-means clustering, uno dei più diffusi e utilizzati. Prima di eseguire l'analisi dei dati, è stato impiegato il comando set.seed (54) per inizializzare il generatore di numeri casuali. Ciò ha garantito che i risultati fossero riproducibili,

consentendo di ottenere gli stessi risultati quando il codice veniva eseguito nuovamente. Si procede ad applicare l'algoritmo di K-means alle 44 osservazioni specificando il numero di cluster desiderato. Come osservato precedentemente nel processo della silhouette, per questa analisi il numero di partizioni ideale deve essere di 3. Ogni Paese viene associato al cluster collegato al centroide più vicino. Si va ad ottenere il seguente output.

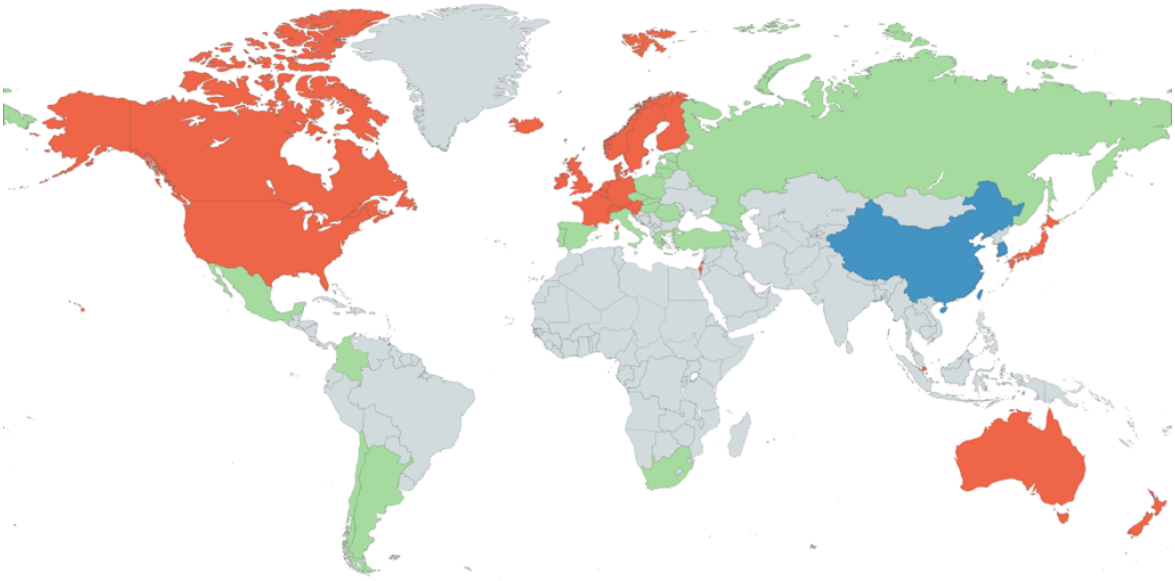
```
K-means clustering with 3 clusters of sizes 22, 19, 3

Cluster means:
  Researchers Invest_on_GDP Patent_App Edu_Spending  Education Unemployment_Rate GDP_percapita
1   0.5032282     0.4940436 -0.2182525    0.2571806  0.5107488        -0.3356589     0.7823030
2  -0.7233751    -0.7851988 -0.2986530   -0.4450752 -0.7542580         0.4910530    -0.8232534
3   0.8910355     1.3499392  3.4919871    0.9328181  1.0314755        -0.6485035    -0.5229508

Clustering vector:
 [1] 2 1 1 1 1 2 3 3 2 2 1 2 1 1 1 2 2 1 1 1 2 1 3 2 2 1 2 1 1 1 2 2 2 2 1 2 1 2 2 1 1 2 1 1

Within cluster sum of squares by cluster:
[1] 81.14595 64.55862 12.47190
 (between_SS / total_SS =  47.4 %)

Available components:

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

Osservando dall'output, sono stati ottenuti 3 cluster con una quantità di osservazioni pari rispettivamente a 3, 22 e 19. Utilizzando il clustering vector, è possibile determinare l'assegnazione di ciascun dato al cluster di appartenenza.

Inoltre, viene riportato il diagramma che rappresenta graficamente la posizione di ciascuna osservazione e dei tre centroidi selezionati al momento della convergenza dell'algoritmo, cioè quando i centroidi si sono stabilizzati e le osservazioni non hanno più cambiato partizione.



Le osservazioni dei cluster sono colorate per simboleggiare l'assegnazione al proprio gruppo di appartenenza. In questo modo è possibile osservare la distribuzione e la relazione tra i punti in questo sistema di coordinate. Questo cluster plot si rivela estremamente utile per facilitare la visualizzazione e l'interpretazione del clustering ottenuto. Per una migliore comprensione e approfondimento del risultato finale del K-means clustering, si allega la cartina con le nazioni suddivise per cluster, in modo tale da indagare riguardo a macro-zone geografiche che possono avere caratteristiche simili.

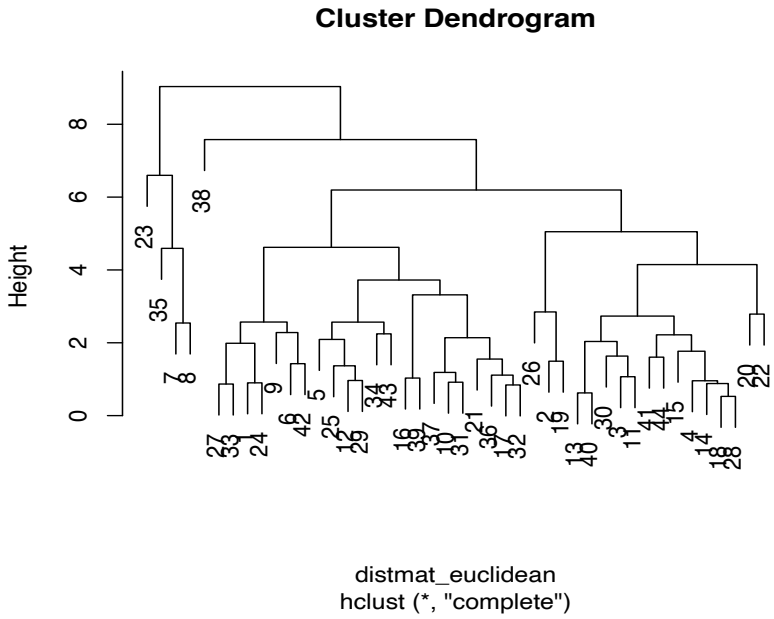


Il clustering ha evidenziato delle differenze tra raggruppamenti di Stati.
In rosso sono rappresentati i Paesi appartenenti al primo cluster con 22 osservazioni. Ne fanno parte macro-zone come il Nord America, Europa Occidentale e Settentrionale, Australia, Nuova Zelanda e Giappone.
Le nazioni in verde fanno parte del secondo cluster con 19 osservazioni. I Paesi si possono raccogliere nelle seguenti zone geografiche: America Centrale e Meridionale, Africa del Sud, Europa Meridionale, Orientale e la Russia.
Infine, i 3 Paesi in blu, quali Cina, Taiwan e Korea, appartengono al terzo cluster in quanto considerati più simili rispetto agli altri.
Nel capitolo conclusivo, saranno presentate le conclusioni relative ai raggruppamenti di Stati che condividono caratteristiche simili come, ad esempio, innovazione e sviluppo.

5.3. CLUSTERING GERARCHICO

Dopo aver analizzato il metodo del k-means clustering, si passa allo svolgimento del clustering gerarchico. A differenza del K-means clustering, per quello gerarchico non è richiesto predeterminare il numero di cluster. Per lo svolgimento del processo, è stato deciso di selezionare la misura Complete Linkage (rispetto al metodo Single o Average Linkage), cioè che per valutare la distanza di due cluster si prende la distanza massima tra tutti i punti dei due gruppi. La decisione è stata presa principalmente sulla base della migliore rappresentazione visiva del dendrogramma con cluster ben definiti e distinti. Il dendrogramma rappresenta graficamente la struttura dei cluster nel clustering gerarchico. Si tratta di un diagramma a forma di albero che mostra la gerarchia dei cluster generata durante il processo di clustering.
Nel dendrogramma, gli oggetti da clusterizzare sono rappresentati come foglie dell'albero, mentre i cluster sono rappresentati dai nodi interni. La lunghezza dei rami nel dendrogramma riflette la distanza o la similarità tra i cluster o le osservazioni. Attraverso un processo iterativo, i cluster simili vengono uniti fino a formare cluster più grandi, ripetendo questo procedimento finché tutti gli oggetti o i cluster sono raggruppati in un unico cluster radice.

La visualizzazione del grafico fornisce una panoramica della struttura dei cluster e delle relazioni tra di essi. Questo strumento può aiutare nell'identificazione di gruppi omogenei di oggetti o nel rilevamento di sottostrutture all'interno dei cluster. Inoltre, la struttura a livelli del dendrogramma permette di selezionare il numero desiderato di cluster, tagliando il diagramma a un determinato livello di similarità.
In conclusione, si può definire come uno strumento utile per interpretare e comunicare i risultati del clustering gerarchico in modo intuitivo. Il grafico viene riportato.



Come suggerito dal metodo della silhouette, il grafico mostra che sono presenti 3 gruppi principali in cui i Paesi si distinguono. Tagliando il dendrogramma ad un'altezza pari circa a 7, si evidenzia la creazione di 3 cluster. Quello a sinistra formato dai Paesi dell'Asia Orientale, quello centrale costituito solamente dall'Africa del Sud (osservazione 38) e, infine, quello a destra include tutti gli altri Paesi rimanenti.

Non è assolutamente anomalo osservare che il Sud Africa sia incluso in un cluster da solo. Infatti, come è stato possibile notare nei paragrafi precedenti, mostra una situazione in termini di indici di innovazione, sviluppo e ricerca scientifica estremamente gravi.
Questo clustering gerarchico presenta dei limiti quando si desidera dividere il dendrogramma in tre gruppi, in quanto i Paesi europei, soprattutto quelli del nord che sono molto avanzati economicamente e scientificamente, e i Paesi dell'America Centrale e Latina, che sono meno sviluppati sotto questi punti di vista, vengono assegnati allo stesso cluster.

6. CONCLUSIONE

Il seguente capitolo ha l'obiettivo di analizzare e riassumere i risultati ottenuti tramite i diversi metodi statistici.

La prima fase della conclusione è rivolta a capire quali sono le variabili che influiscono maggiormente al numero di ricercatori per Paese. Tramite lo svolgimento della regressione multipla, è stato osservato che le tre variabili significative sono l'investimento in Ricerca e Sviluppo in percentuale del PIL, la spesa in educazione in percentuale del PIL e PIL pro-capite. In primo luogo, l'investimento in R&S e formazione rappresenta l'investimento di un Paese nello sviluppo tecnologico. Investimenti più elevati indicano la priorità della ricerca e dell'innovazione, con risorse dedicate al finanziamento di progetti di ricerca, infrastrutture e borse di studio per i ricercatori. Ciò crea un ambiente favorevole per attrarre e trattenere i talenti della ricerca, incoraggiando la formazione di più ricercatori nel Paese. Successivamente, il PIL pro-capite è una misura della ricchezza e dello sviluppo economico di un Paese. I Paesi con un PIL pro capite più elevato hanno le risorse finanziarie disponibili per investire in R&S, compresa l'assunzione di più ricercatori. Un'economia più forte fornisce un solido sostegno finanziario per sostenere la formazione e l'occupazione dei professionisti della ricerca.
In sintesi, la spesa in R&S in percentuale del PIL, la spesa in istruzione in percentuale del PIL e il PIL pro-capite sono correlati al numero di ricercatori in ciascun Paese perché riflettono l'impegno di un Paese per lo sviluppo scientifico e tecnologico, la disponibilità di risorse finanziarie sostenere la formazione dei ricercatori, l'occupazione, l'accesso a infrastrutture e risorse avanzate, l'attrazione di talenti internazionali e le priorità politiche per la ricerca e l'innovazione. L'aumento degli investimenti in questi fattori crea un ambiente favorevole alla crescita del numero di ricercatori, all'innovazione, al progresso scientifico e al progresso sociale ed economico.

La seconda ed ultima fase del seguente capitolo ha lo scopo di analizzare dal punto di vista geografico quali sono le macro-zone in cui il numero di ricercatori ogni mille abitanti è significativamente rilevante tenendo conto delle sue variabili correlate. Per raggiungere il risultato finale verrà utilizzata l'analisi univariata e il metodo di clustering. Per quanto riguarda il clustering, per la valutazione verrà adoperato il metodo K-means clustering in quanto i Paesi vengono suddivisi in una maniera migliore con una menzione specifica al solo Paese del Sud Africa che, con il metodo del clustering gerarchico, rappresenta un cluster con una sola partizione.
Sin da subito, l'indagine univariata ha rivelato una persistente differenza tra i Paesi del mondo per quanto riguarda aspetti economici, scientifici e educativi. Si è osservato un dualismo regionale tra tre macrogruppi di Paesi. Un gruppo formato dalle nazioni dell'Asia Pacifica come Cina, Taiwan, Corea del Sud e il Giappone. L'altra partizione costituita dalle nazioni dell'Europa Occidentale e Settentrionale e America del Nord. L'ultimo gruppo include i Paesi dell'America Centrale e Meridionale e l'Europa Mediterranea e dell'ex blocco sovietico ed il Medio Oriente. Infine, come suggerito dal cluster gerarchico, il Sud Africa viene considerato come un cluster a sé.
Questo dualismo si riferisce alle disparità e alle disuguaglianze nello sviluppo economico e nel progresso scientifico tra le regioni del mondo. Dal punto di vista economico, alcune nazioni sono diventate motori della crescita globale, attirando investimenti e capitali, creando prosperità

economica e posti di lavoro. Queste regioni sono generalmente concentrate in Paesi ad alto reddito o in località privilegiate, che beneficiano di infrastrutture avanzate, istituzioni forti, reti commerciali ben sviluppate e abbondanti risorse naturali. Il risultato è stato un aumento della produzione di beni e servizi, industrie innovative che investono molto nel progresso scientifico e una crescita economica sostenuta.
Da un punto di vista economico, il primo gruppo di Paesi è economicamente sviluppato, altamente industrializzato, con un'ampia gamma di settori economici sviluppati e prosperità economica. I Paesi dell'Asia-Pacifico come la Cina, il Giappone e Singapore hanno sperimentato una rapida crescita economica e la loro capacità di attrarre investimenti internazionali è notevole. Questi Paesi stanno investendo molto nella ricerca e nello sviluppo di nuove tecnologie. Inoltre, la regione ha saputo sfruttare le opportunità offerte dalla globalizzazione per diventare un importante centro di produzione e commercio internazionale. Un'altra chiave del successo è la crescente connettività tra le economie della regione, che ha permesso di costruire una solida rete di collegamenti commerciali.
Per quanto riguarda il secondo macrogruppo, l'Europa Occidentale ha stabilito un'economia altamente sviluppata e un solido sistema finanziario con Paesi come Germania, Francia e Regno Unito. Difatti, l'Unione Europea ha creato un ambiente favorevole per la cooperazione commerciale e scientifica tra i suoi membri, contribuendo a promuovere la crescita economica e lo scambio di conoscenze scientifiche, portando alla creazione di una forte e sentita *industrial atmosphere*. L'America del Nord, rappresentato principalmente da Stati Uniti e Canada, è nota per la sua forza economica, l'innovazione tecnologica e la presenza di aziende leader a livello mondiale. Gli Stati Uniti e il Canada hanno investito molto nell'istruzione superiore con università specializzate che attirano studenti da tutto il mondo. I due Paesi dedicano ogni anno ingenti fondi pubblici e privati alla ricerca scientifica e promuovono la cooperazione tra ricercatori grazie alla loro avanzata infrastruttura di ricerca. I due Paesi citati sono caratterizzati da una cultura che promuove l'innovazione e l'imprenditorialità con la costituzione di nuove startup.
D'altra parte, il terzo gruppo di Paesi deve affrontare sfide economiche significative. I Paesi dell'America centrale e meridionale, come Argentina, Messico e Colombia, spesso affrontano instabilità economica, uso inefficiente delle risorse, disuguaglianza sociale e aumento dell'inflazione. Inoltre, altri problemi estremamente gravi che affliggono questi Paesi sono la forte corruzione e la cattiva e instabile gestione governativa che porta ad ostacolare lo sviluppo economico e scientifico. L'Europa mediterranea, compresi Paesi come Spagna, Italia e Grecia, ha subito le conseguenze della crisi economica del 2010 e stenta ancora a riprendersi. I Paesi dell'ex blocco sovietico, come la Russia, hanno lottato per passare a un'economia di mercato e hanno lottato per affrontare la corruzione e l'inefficienza.
Nel campo della scienza, il primo e il secondo Paese del gruppo hanno istituti di ricerca all'avanguardia, prestigiose università e un ambiente favorevole alla ricerca e all'innovazione. Ciò si traduce in alti livelli di produzione scientifica, importanti scoperte e partecipazione attiva alla comunità scientifica internazionale. Questi Paesi investono ingenti risorse nella ricerca scientifica e nello sviluppo per promuovere l'innovazione tecnologica e il progresso scientifico. Nel terzo gruppo di Paesi, tuttavia, il settore scientifico deve affrontare delle sfide. Le risorse per la ricerca scientifica possono essere limitate e l'accesso a infrastrutture e risorse avanzate può essere ancora più limitato. Ciò può comportare una minore coerenza della produzione scientifica e una partecipazione meno attiva della comunità scientifica internazionale.

Per quanto riguarda il Sudafrica, si ritiene che il Paese abbia una delle economie più avanzate e diversificate dell'Africa. La base industriale è solida e il settore dei servizi è sviluppato. Tuttavia, l'analisi univariata mostra che il Sudafrica deve ancora affrontare sfide importanti, come l'elevata disoccupazione, la disuguaglianza socioeconomica e gli alti livelli di povertà. Inoltre, la crescita economica è stata lenta negli ultimi anni.
Nel complesso, si può affermare che i Paesi dell'Asia-Pacifico e dei Paesi nordici si distinguono per numero di ricercatori, dimostrando un notevole impegno nell'investire nelle risorse umane e nel sostenere la ricerca scientifica. Questa attenzione all'innovazione e alla formazione di una forza lavoro altamente qualificata ha aiutato queste regioni a diventare leader nella produzione di conoscenza e nel progresso scientifico. Tuttavia, è importante promuovere una distribuzione più equa delle risorse, sostenere gli sforzi di altri Paesi per espandere la base di ricercatori e promuovere una cooperazione internazionale più inclusiva per affrontare le sfide globali condivise.

Infine, a conferma delle considerazioni fatte, si è voluto confrontare i risultati ottenuti con il *Global Innovation Index* (GII), un indice di aggiornamento globale che rivela le economie più innovative del mondo classificando le performance dei Paesi. L'edizione 2022 vede la Svizzera in testa alla classifica per il 12° anno consecutivo, seguita da Stati Uniti, Svezia, Regno Unito e Paesi Bassi, mentre la Cina è più vicina alla top 10. Grazie all'elevata qualità della vita, la Svizzera attrae i migliori ricercatori e ci sono strette collaborazioni tra università e aziende. Gli Usa non stupiscono nessuno, anche perché lì si trovano le 4 società con la maggiore spesa in R&S, ovvero Amazon, Alphabet, Microsoft e Apple. Molti Paesi dell'Europa Occidentale e del Nord si trovano nella top 10, quali Svezia (3), Regno Unito (4), Paesi Bassi (5), Germania (8), Finlandia (9) e Danimarca (10). La Corea del Sud, in cui si ha un elevato grado di collaborazione tra imprese e università, si trova al sesto posto. L'Italia, invece, si trova solamente al ventottesimo posto.
L'indice mostra anche i maggiori cluster di innovazione scientifica e tecnologica del mondo. Tra questi i migliori centri, anche definiti hub, si trovano a Tokyo-Yokohama (Giappone), seguito da Shenzhen-Hong Kong-Guangzhou (Cina e Hong Kong, Cina), Pechino (Cina), Seoul (Repubblica di Corea) e San Josè-San Francisco (Stati Uniti).
Confrontando questo indice ai risultati ottenuti da questa analisi è possibile affermare che è stato raggiunto un buonissimo livello di affidabilità, correttezza e attendibilità.

SITOGRAFIA

I dati che sono Stati utilizzati per svolgere l'analisi sono Stati estratti dai seguenti siti:

Anno di riferimento:2020

Researchers	https://data.oecd.org/rd/researchers.htm#indicator-chart
Investment on GDP	https://data.oecd.org/rd/gross-domestic-spending-on-r-d.htm
Patents	https://data.worldbank.org/indicator/IP.PAT.RESD
Education spending (tertiary)	https://data.oecd.org/eduresource/education-spending.htm#indicator-chart
Education	https://worldpopulationreview.com/country-rankings/most-educated-countries
Unemployment rate	https://data.oecd.org/unemp/unemployment-rate.htm
GDP percapita	https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

https://sciencebusiness.net/news/number-scientists-worldwide-reaches-88m-global-research-spending-grows-faster-economy#:~:text=Number%20of%20scientists%20worldwide%20reaches,than%20the%20economy%20%7C%20Science%7CBusiness

https://www.infodata.ilsole24ore.com/2019/01/10/investire-ricerca-sviluppo-israele-campione-del-mondo/?refresh_ce=1

https://innovitalia.esteri.it/pagina_paese/Israele

https://www.cespi.it/it/eventi-attualita/dibattiti/america-latina-que-pasa/la-struttura-produttiva-latinoamericana-causa

https://www.thewatcherpost.it/innovazione/chip-la-cina-sorpassa-gli-usa-il-boom-di-brevetti-dellinnovazione/

https://worldpopulationreview.com/country-rankings/most-educated-countries

https://www.almalaurea.it/sites/almalaurea.it/files/docs/universita/profilo/profilo2021/almalaurea_profilo_rapporto2021_01_contesto_di_riferimento.pdf

https://www.infodata.ilsole24ore.com/2023/04/08/i-piu-ricchi-e-i-piu-poveri-per-pil-pro-capite-le-due-mappe-a-confronto/
https://www.creditnews.it/paesi-piu-innovazione-2022/