UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Informatica

CORSO DI LAUREA MAGISTRALE IN INFORMATICA

DATA SCIENCE E MACHINE LEARNING

PROGETTO DI STATISTICA E ANALISI DEI DATI

Violenza di genere durante la pandemia: aggravamento o attenuazione dei numeri?

DOCENTE

Prof. Amelia Giuseppina Nobile

STUDENTI

Maria Natale, matricola: 0522500967

Gaetano Casillo, matricola: 0522501057

ANNO ACCADEMICO 2020-2021

SOMMARIO

1 INTRODUZIONE

Si sente spesso parlare di violenza di genere, ma che cosa vuol dire? Le Nazioni Unite hanno definito la violenza di genere come "*ogni atto legato alla differenza di sesso che provochi o possa provocare un danno fisico, sessuale, psicologico o una sofferenza della donna, compresa la minaccia di tali atti, la coercizione o l'arbitraria privazione della libertà sia nella vita pubblica che nella vita privata*". Per supportare le vittime delle violenze di genere è stato attivato il numero verde 1522 attivo 24 ore su 24 offrendo accoglienza in italiano, inglese, francese, spagnolo e arabo.

1.1 CASO DI STUDIO

Nel 2020 è stato vissuto il lockdown per 3 mesi, in questo periodo molto si è parlato del lato economico, della scuola, ma poco si è discusso del lato sociale di questo evento. Si è pensato pertanto di analizzare le chiamate e i messaggi effettuate al 1522, numero verde contro lo stalking e la violenza, confrontandole con lo stesso periodo (marzo/giugno) degli anni precedenti. In particolare, nell'analisi statistica effettuata si fa riferimento agli utenti e alle vittime per regione di provenienza e anno. Secondo quanto rilevato dall'Istat, che ha analizzato i dati messi a disposizione dal numero antiviolenza 1522, tra marzo e giugno 2020 le telefonate e le comunicazioni via chat con il centralino sono più che raddoppiate rispetto allo stesso periodo dell'anno precedente.

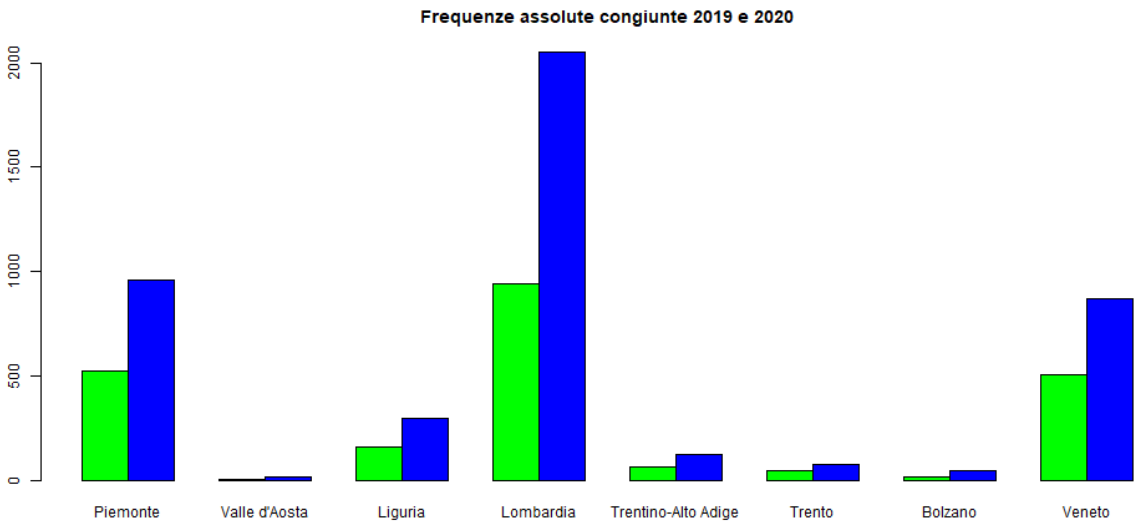
Per l'analisi del fenomeno in esame si considerano i dati relativi agli utenti del numero antiviolenza 1522 effettuate nei mesi di marzo-giugno suddivisi per regione ed anno (2013-2020). In particolare, nell'analisi statistica univariata, verranno esaminate nei dettagli le curve relativi ai dati della regione Campania e la media delle chiamate degli utenti e delle vittime effettuate sull'intero territorio nazionale.

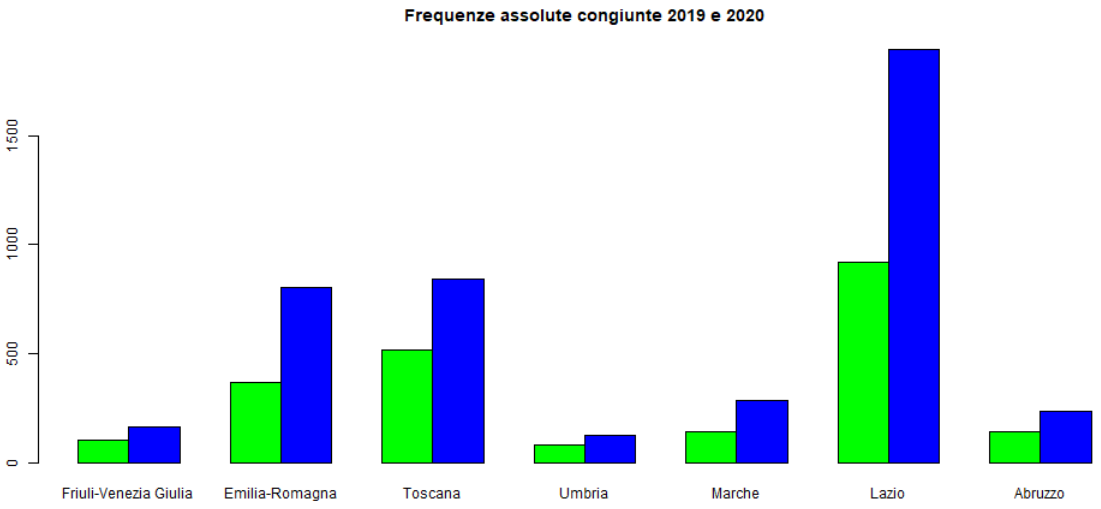
Nella seguente tabella vengono mostrati i dati relativi agli utenti del numero 1522 suddivisi per regione ed anno.

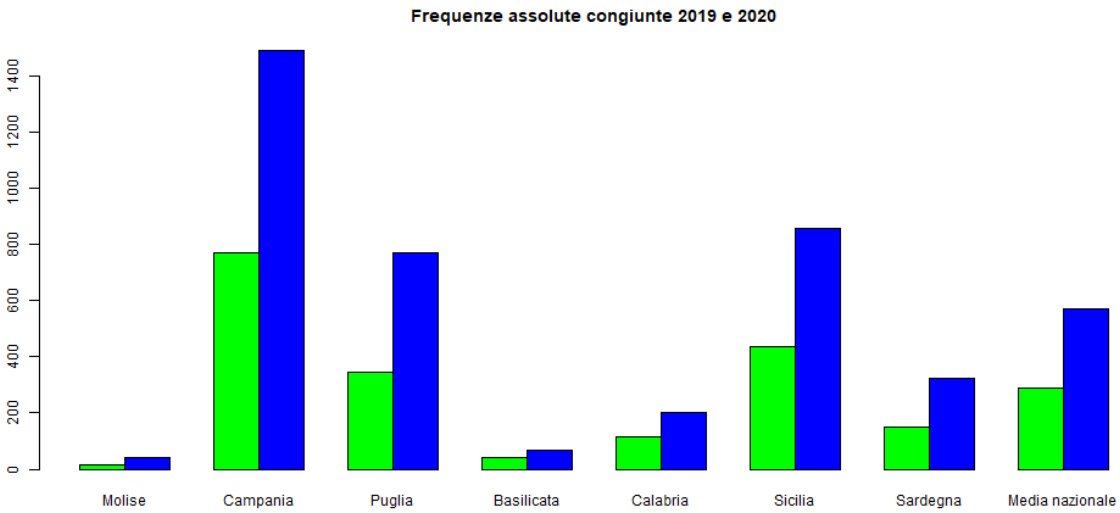
Regioni	2013	2014	2015	2016	2017	2018	2019	2020
Piemonte	667	681	543	441	408	511	524	958
Valle d'Aosta	17	15	8	9	4	6	6	14
Liguria	298	259	200	184	106	180	160	296
Lombardia	1.488	1.174	886	878	822	1.034	939	2.055
Trentino-Alto Adige	66	53	48	54	22	49	63	121
Trento	45	41	41	42	18	36	47	74
Bolzano	20	12	7	12	4	13	14	43
Veneto	751	636	397	306	273	429	506	868
Friuli-Venezia Giulia	154	124	99	70	65	93	103	163
Emilia-Romagna	654	492	341	311	216	352	365	804
Toscana	594	484	298	273	311	398	514	841
Umbria	161	121	79	76	44	89	83	125
Marche	258	164	174	154	114	137	141	286
Lazio	1.306	1.304	838	833	593	959	919	1.898
Abruzzo	261	186	176	128	118	184	144	235
Molise	46	31	30	27	18	13	16	43
Campania	1.316	1.059	815	635	537	823	772	1.492
Puglia	768	677	391	350	246	450	344	771
Basilicata	61	62	40	25	31	39	42	67
Calabria	210	135	122	99	60	125	115	200
Sicilia	922	701	446	481	390	492	438	859
Sardegna	356	296	226	184	172	157	149	322
Media nazionale	474	396	282	253	208	299	291	570

1.2 RAPPRESENTAZIONE GRAFICA

Osservando la tabella dei dati ci si accorge subito che nel 2020 i casi sono quasi raddoppiati rispetto al 2019. Nei seguenti grafici infatti vengono mostrate le frequenze assolute delle chiamate effettuate nelle varie regioni, mostrando in verde quelle effettuate nel 2019 e in blu quelle effettuate nel 2020. In percentuale si è avuto un aumento medio del 95.9%. Per alcune regioni, tuttavia si è avuto un aumento ancora più significativo, ad esempio la Lombardia nel 2020 ha visto aumentare il numero di chiamate del 118.8% rispetto all'anno precedente. Altre regioni, ad esempio, la Toscana hanno visto un aumento meno significativo rispetto alle altre regioni, essa infatti ha registrato un aumento del 63.6%. La Campania, invece, con un aumento del 92.1% è molto vicina all'aumento medio registrato.


Frequenze assolute congiunte 2019 e 2020
0
500
1000
1500
2000
Piemonte
Valle d'Aosta
Liguria
Lombardia
Trentino-Alto Adige
Trento
Bolzano
Veneto


Frequenze assolute congiunte 2019 e 2020
0
500
1000
1500
Friuli-Venezia Giulia
Emilia-Romagna
Toscana
Umbria
Marche
Lazio
Abruzzo

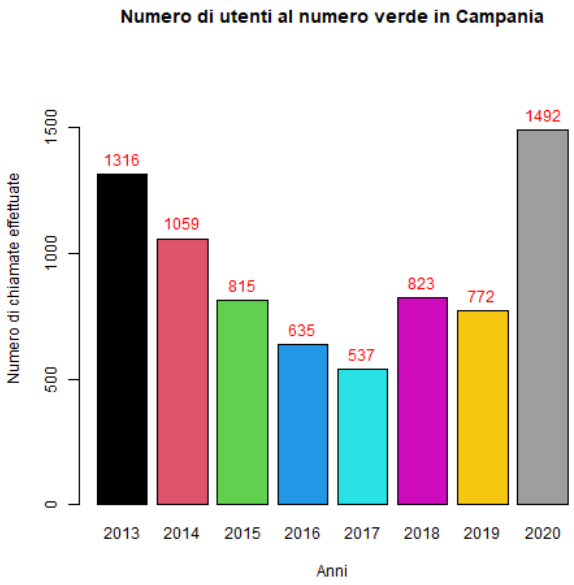

Frequenze assolute congiunte 2019 e 2020
0
200
400
600
800
1000
1200
1400
Molise
Campania
Puglia
Basilicata
Calabria
Sicilia
Sardegna
Media nazionale

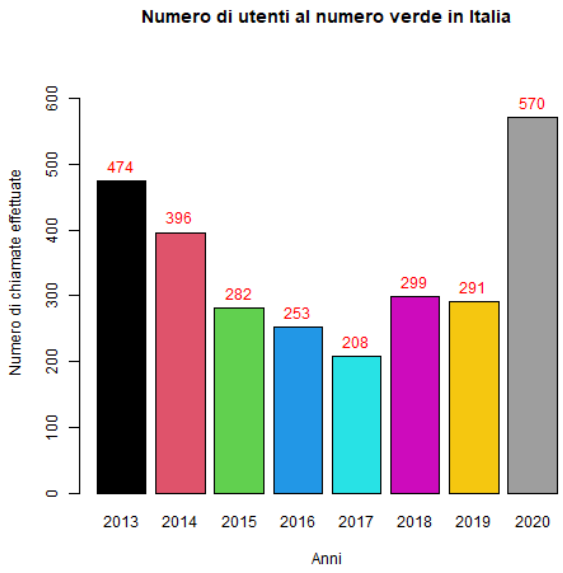
Ci si proporrà nel seguito di analizzare la curva negli anni della regione Campania confrontandola con ciò che è avvenuto sull'intero territorio nazionale (considerando quindi la media di tutti gli anni). È stato già mostrato che per quanto riguarda l'ultimo anno l'incremento di casi registrato in Campania si attesa ad un livello assai vicino alla media nazionale. Nel seguito si cercherà di capire se anche gli altri hanno si sono registrati livelli di incrementi e decrementi simili o se si è verificata qualche anomalia in Campania rispetto ai numeri registrati nel resto del paese.

Di seguito vengono mostrati i due barplot relativi ai dati della Campania e della media sull'intero territorio nazionale per quanto riguarda la tabella Utenti.

Il codice per realizzare il barplot della Campania:

```
png("grafici/chiamateEffettuateUtentiCampaniaFrequenza.png")
x<-barplot(utenti_campania, xlab="Anni", ylab="Numero di chiamate effettuate",
ylim=c(0,1800), col=1:9,
        names.arg = colnames, main ="Numero di utenti al numero verde in
Campania")
text(x, y=utenti_campania, pos = 3, labels = utenti_campania, col="red")
dev.off()
```





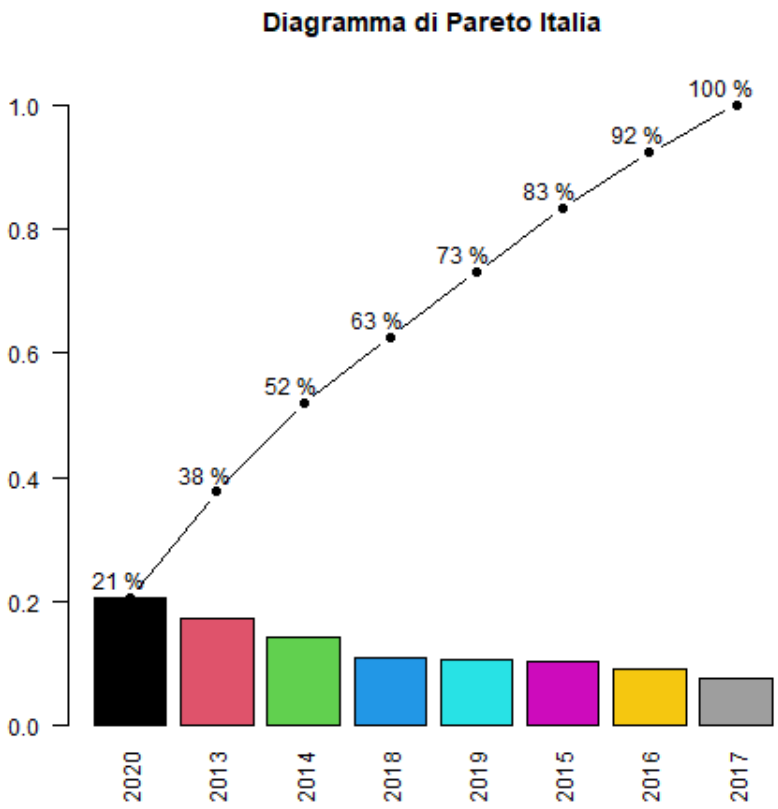
In entrambi i casi si può notare che la modalità a cui è associata la frequenza più alta è il 2020.

Il **diagramma di Pareto** è utile per analizzare un insieme di dati in modo da determinare le poche variabili che influenzano in modo significativo i risultati finali. Il diagramma è composto da barre che indicano l'incidenza percentuale sul fenomeno in esame dei singoli elementi. Le barre più alte

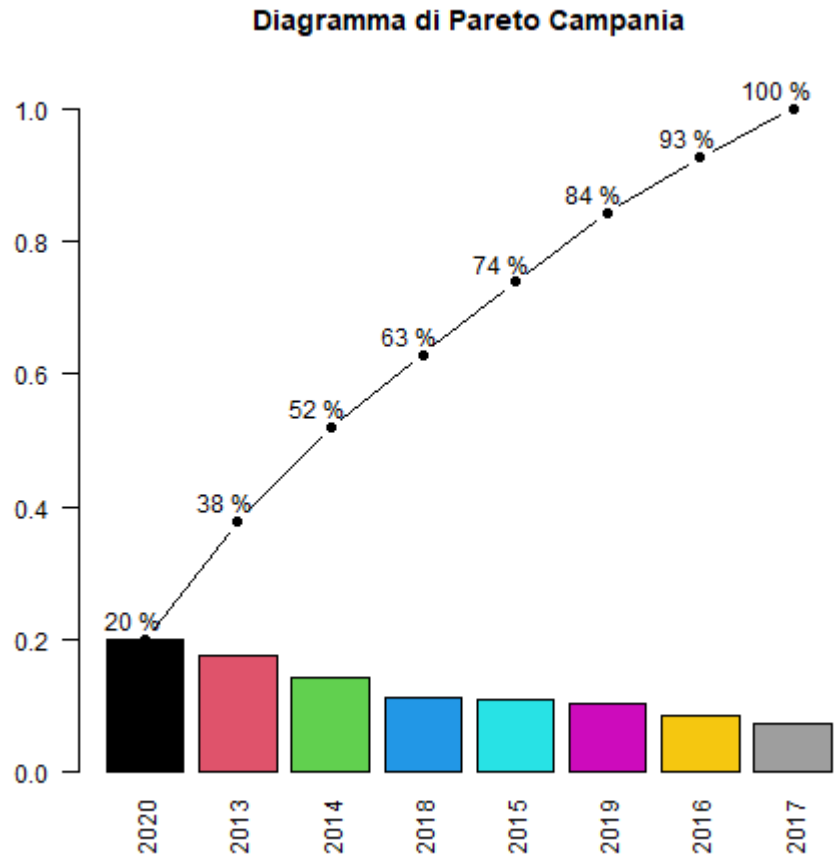
corrispondono agli elementi che incidono maggiormente sul fenomeno. Nel diagramma di Pareto è inoltre presente una linea che rappresenta le incidenze degli elementi sommate l'una all'altra.

Codice per generare il diagramma di Pareto per gli utenti al numero 1522 per il campione media nazionale.

```
tableNaz<-table(c(rep("2013", utenti_nazione[1]),
rep("2014",utenti_nazione[2]), rep("2015",utenti_nazione[3]),
                  rep("2016",utenti_nazione[4]),
                  rep("2017",utenti_nazione[5]),
                  rep("2018",utenti_nazione[6]),
                  rep("2019",utenti_nazione[7]),
                  rep("2020",utenti_nazione[8])))
ord<-sort(tableNaz, decreasing = TRUE)
propOrd <- prop.table (ord)
x <- barplot (propOrd , ylim = c(0, 1.05) , main = "Diagramma di Pareto
Italia", col =1:8 , las =2)
lines(x, cumsum(propOrd), type = "b", pch = 16)
text(x - 0.2, cumsum (propOrd) + 0.03 , paste (format(cumsum(propOrd) * 100,
digits = 2) , "%"))
```



Considerando gli ultimi 8 anni, il diagramma di Pareto mostra che il solo anno 2020 incide per il 21% sul totale delle chiamate registrate sulla media nazionale. Si tratta di una percentuale abbastanza alta in quanto un numero equo di chiamate per anno corrisponderebbe al 12.5%.



Anche nel caso della Campania, il 2020 incide per il 20% sul totale delle chiamate.

Confrontando i due diagrammi di Pareto ottenuti si nota, innanzitutto che nell'anno 2020 si è avuto un aumento significativo dei casi di violenza registrati rispetto ad altri anni. Inoltre, effettuando un'analisi più approfondita è possibile notare che c'è una leggera differenza tra l'incidenza degli anni 2019 e 2015 per la Campania e la media nazionale. In Campania l'anno 2015 risulta incidere maggiormente rispetto all'anno 2019 mentre per la media nazionale si nota che l'anno 2019 incide maggiormente rispetto al 2015. Tuttavia, si tratta di una differenza minima.

2 STATISTICA DESCRITTIVA UNIVARIATA

In questo capitolo verranno mostrati i risultati relativi all'analisi statistica univariata. In particolare, verrà mostrata la funzione di distribuzione empirica continua, i valori degli indici di sintesi, i quartili calcolati con i differenti algoritmi di R e gli indici di dispersione. Infine, verrà analizzata la forma della distribuzione di frequenze attraverso il calcolo della skewness campionaria e della curtosi campionaria. Le varie analisi verranno effettuate prendendo in esame i dati della Campania e della media nazionale negli anni 2013-2020, analizzando la tabella Utenti.

2.1 FUNZIONE DI DISTRIBUZIONE EMPIRICA CONTINUA

La funzione di distribuzione empirica continua viene utilizzata nel caso di dati continui che vengono strutturati in classi. Ad esempio, se si vuole considerare k classi distinte, le classi saranno così caratterizzate: $C_1 = [z_0, z_1)$, $C_2 = [z_1, z_2)$, … $C_k=[z_{k-1}, z_k]$ con $z_0 < z_1 < \ldots < z_{k-1} < z_k$, dove z_0 corrisponde al minimo delle osservazioni e z_k corrisponde al massimo delle osservazioni. La funzione di distribuzione empirica continua viene calcolata a partire dalle frequenze relative cumulative associate alle varie classi.

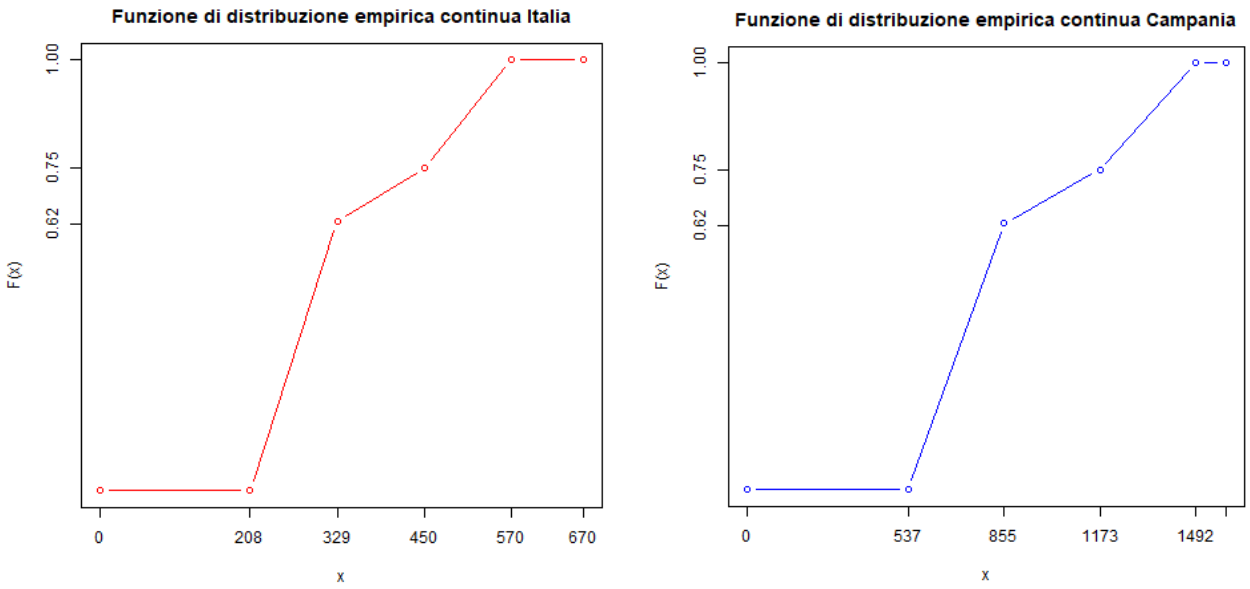
Per calcolare la funzione di distribuzione continua relativa alla tabella Utenti le osservazioni sono state suddivise in tre classi. Per quanto riguarda la media nazionale le classi individuate sono le seguenti: $C_1 = [208, 329)$, $C_2 = [329, 450)$, $C_3 = [450, 570]$. Per quanto riguarda la Campania le classi individuate sono le seguenti: $C_1 = [537, 855)$, $C_2 = [855, 1173)$, $C_3 = [1173, 1492]$.

Il seguente codice mostra come sono state calcolate le frequenze relative cumulative per le classi individuate per quanto riguarda la Campania.

```
minOsservazione = min(utenti_campania)
maxOsservazione = max(utenti_campania)
frequenza<-table(utenti_campania)/length(utenti_campania)
lung<-length(frequenza)
classe<-round((maxOsservazione-minOsservazione)/3, digits=0)
classi<-c(minOsservazione, minOsservazione+classe, minOsservazione+2*classe,
maxOsservazione)
frelclassi <-table (cut (utenti_campania, breaks = classi,right = FALSE ))/
length (utenti_campania)
Fcum <-cumsum (frelclassi)
Fcum[3]<-Fcum[3]+frequenza[lung]
```

Dopo aver calcolato le frequenze relative cumulative. sono stati quindi creati i grafici che mostrano le frequenze di distribuzione continue della Campania e dell'intera nazione. Si mostra il codice per generare il grafico della funzione di distribuzione empirica continua della Campania.

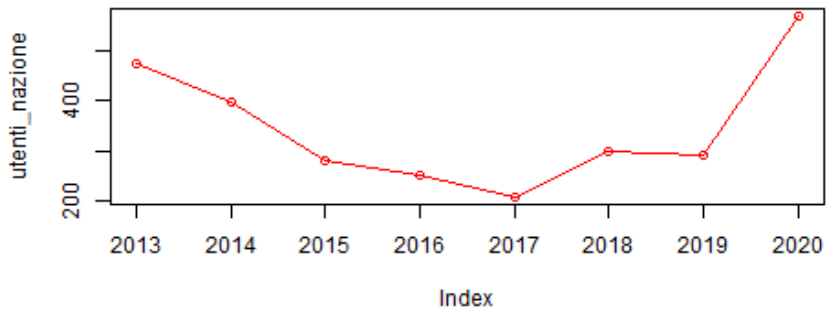
```
ascisse<-c(0, classi, maxOsservazione+100)
ordinate <-c(0, 0, FcumI [1:3] ,1)
plot(ascisse , ordinate , type = "b", axes = FALSE , main = "
Funzione di distribuzione empirica continua Campania", col =" red ",ylim=c(0
,1) ,xlab="x",ylab="F(x)")
axis (1, format(ascisse, digits=2))
axis (2, format(FcumI, digits=2))
box()
```

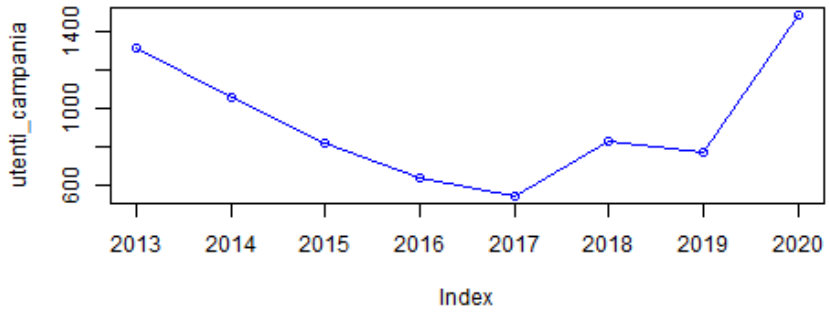


Dai grafici si può notare che avendo diviso i dati in 3 classi, le classi di entrambi i campioni di dati presentano le stesse frequenze relative. In particolare, la prima classe ha una frequenza relativa di 0.625, la seconda di 0.125 e la terza di 0.250.

2.2 INDICI DI SINTESI

Nel grafico seguente vengono mostrate le due curve relative ai dati che si stanno analizzando.





Entrambe le curve mostrano una distribuzione di frequenze non simmetrica, in particolare inizialmente sono decrescenti fino ad arrivare ad un valore minimo nel 2017 e successivamente hanno un picco massimo nell'ultimo anno 2020. Le due curve sono tra loro abbastanza simili.

Alcuni indici di sintesi utili a descrivere i dati sono media, mediana, moda, varianza, deviazione standard e coefficiente di variazione. Le prime tre sono misure di centralità dei dati mentre le altre misurano la loro dispersione.

Supponiamo di avere un insieme, x_1 , x_2 , …, x_n di n valori numerici. Si definisce **media campionaria** $\bar{x}$ la quantità:

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

Le medie campionarie dei due campioni di dati negli anni risultano essere:

```
> mean(utenti_campania)
[1] 931.125
> mean(utenti_nazione)
[1] 346.625
```

Pertanto, è possibile vedere quali sono gli anni in cui ci sono state più chiamate rispetto alla media e gli anni in cui ci sono state meno chiamate.

Media nazionale	
2013	474
2014	396
2015	282
2016	253
2017	208
2018	299
2019	291
2020	570

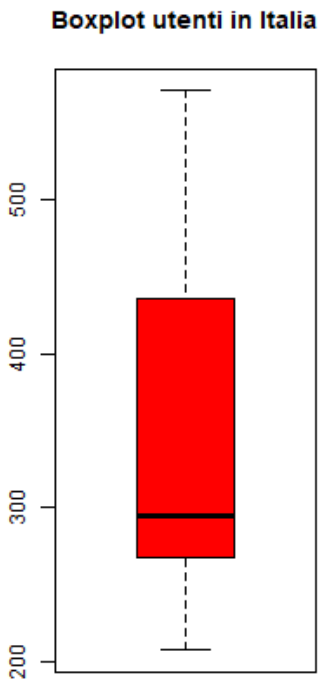
Campania	
2013	1316
2014	1059
2015	815
2016	635
2017	537
2018	823
2019	772
2020	1492

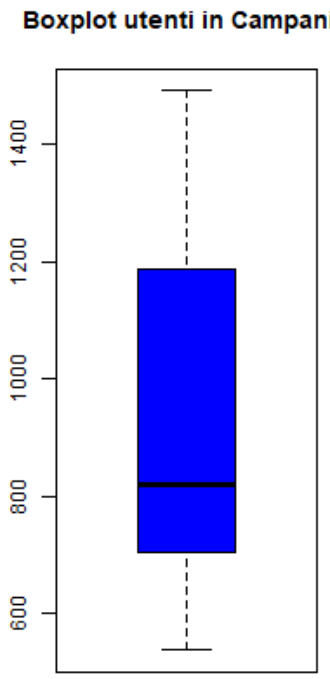
Sia per la Campania che per la media nazionale gli anni in cui ci sono state più chiamate rispetto alla loro media sono 2013, 2014 e 2020.

Prima di illustrare i dati attraverso un boxplot è utile ricordare i concetti di quantili e di mediana. Dato un campione di dati ordinato in maniera crescente, si definisce la **mediana** (o **valore mediano**) come il valore/modalità assunto dalle unità statistiche che si trovano nel mezzo della distribuzione. Se n è dispari, la mediana sarà il valore in posizione (n+1)/2; se n è pari la mediana sarà la media aritmetica dei valori in posizione n/2 e n/2+1. La mediana, quindi è quel valore che divide a metà l'insieme dei dati ordinati. Oltre a questo indice si possono considerare altri indici di posizione detti quantili che consentono di suddividere l'insieme dei dati ordinati in un fissato numeri di parti uguali. In particolare, verranno considerati i quartili che consentono di dividere l'insieme dei dati ordinati in quattro parti uguali.

Il grafico seguente mostra, invece, i boxplot di entrambi i campioni di dati per illustrare alcune caratteristiche della distribuzione di frequenza come centralità, dispersione, forma e la presenza di eventuali valori anomali. Il boxplot, detto anche "scatola con i baffi", rappresenta una scatola i cui estremi sono Q_1 (primo quartile) e Q_3 (terzo quartile) tagliata da una linea orizzontale in corrispondenza di Q_2 (secondo quartile). Sono inoltre presenti due ulteriori linee che rappresentano i baffi in alto e in basso. Il baffo inferiore corrisponde al valore più piccolo tra le osservazioni che risulta maggiore o uguale a $Q_1 - 1.5 * (Q_3 - Q_1)$, mentre il baffo superiore corrisponde al valore più grande delle osservazioni che risulta minore o uguale a $Q_3 + 1.5 * (Q_3 - Q_1)$. Se tutti i dati

rientrano nell'intervallo $\left(Q_1 - 1.5 * (Q_3 - Q_1), Q_3 + 1.5 * (Q_3 - Q_1)\right)$, i baffi risultano essere posti in corrispondenza del minimo e del massimo dei dati del campione. I valori anomali al di fuori di tale intervallo vengono visualizzati sotto forma di punti nel grafico.





Entrambi i boxplot rivelano la presenza di asimmetria nei dati in quanto le distanze tra primo e terzo quartile dalla linea della mediana sono molto diverse tra loro. Si può intuire che le curve hanno una coda più allungata a destra e ciò verrà confermato attraverso il calcolo della skewness campionaria.

Utilizzando la funzione summary in R è possibile calcolare minimo, massimo, media, mediana, primo e terzo quartile.

```
summary(utenti_nazione)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
208.0   274.8   295.0   346.6   415.5   570.0

summary(utenti_campania)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
537.0   737.8   819.0   931.1  1123.2  1492.0
```

Avendo ottenuto il valore dei quartili, è possibile calcolare il valore dei baffi del boxplot della Campania.

$$\left(Q_1 - 1.5 * (Q_3 - Q_1)\right) = 737.8 - 1.5 * (1123.2 - 737.8) = 159.7$$

$$(Q_3 + 1.5 * (Q_3 - Q_1)) = 1123.2 + 1.5 * (1123.2 - 737.8) = 1701.3$$

Tutti i dati rientrano nell'intervallo (159.7, 1701.3) pertanto i baffi sono posti in corrispondenza del minimo e del massimo delle osservazioni. (537, 1492)

Valore dei baffi nel boxplot della media nazionale:

$\left(Q_1 - 1.5 * (Q_3 - Q_1)\right) = 274.8 - 1.5 * (415.5 - 274.8) = 63.75$ quindi il baffo inferiore è posto in corrispondenza del valore 208.

$(Q_3 + 1.5 * (Q_3 - Q_1)) = 415.5 + 1.5 * (415.5 - 274.8) = 626.55$ quindi il baffo superiore è posto in corrispondenza del valore 570.

Tutti i dati rientrano nell'intervallo (63.75, 626.55) pertanto i baffi sono posti in corrispondenza del minimo e del massimo delle osservazioni. (208, 570)

Nei due boxplot non risultano esserci valori anomali.

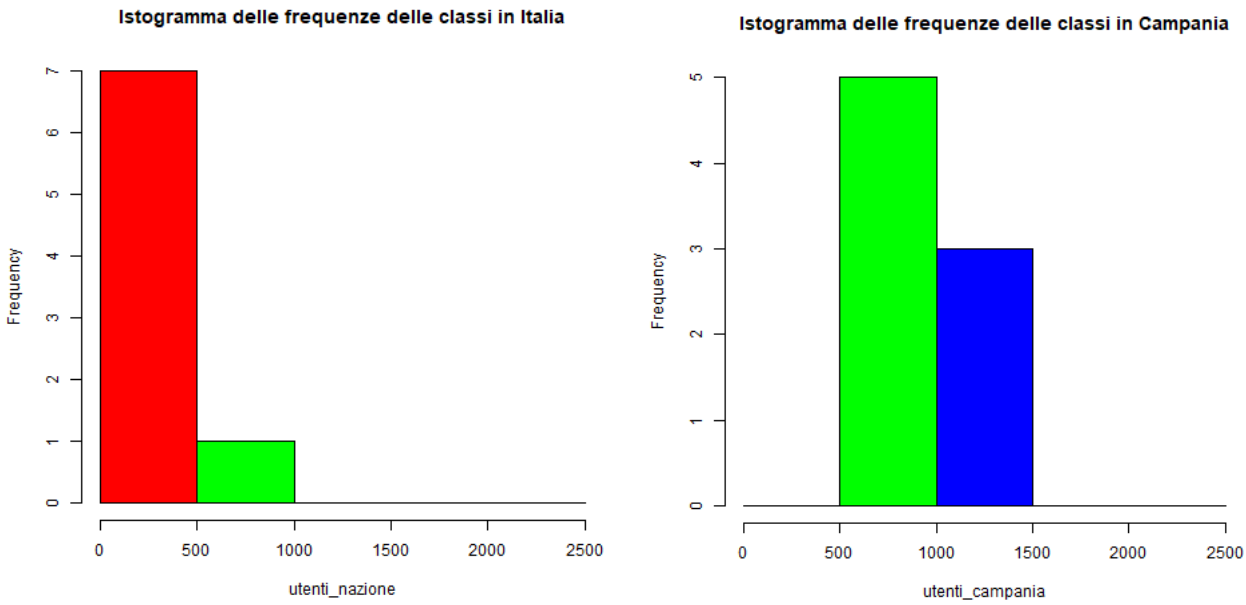
La **moda campionaria** di un insieme di dati è il valore a cui è associata la frequenza più elevata, non è obbligatorio che la moda esista in ogni insieme di dati e se esiste, è possibile che ne esista più di una; in questo caso, ogni valore è detto "valore modale". Se si hanno insieme di dati raggruppati in classi, la classe a cui è associata la frequenza più alta viene detta classe modale.

Per individuare la moda si considerano gli istogrammi delle frequenze dei dati considerando la loro suddivisione nelle seguenti cinque classi: C_1 = [0, 500), C_2 = [500, 1000), C_3 = [1000, 1500) C_4= [1500, 2000), C_5= [2000, 2500].

```
#calcolo delle frequenze associate alle classi
classi<-c(0, 500, 1000, 1500, 2000, 2500)
fclassiCampania <-table (cut (utenti_campania, breaks = classi,right = FALSE,
dig.lab = 10))
for (i in 1:length(utenti_campania)){
  if(utenti_campania[i]==2500)
    fclassiCampania[3]<-fclassiCampania[3]+1
}
fclassiItalia <-table (cut (utenti_nazione, breaks = classi,right = FALSE,
dig.lab=10))
for (i in 1:length(utenti_nazione)){
  if(utenti_nazione[i]==2500)
    fclassiItalia[3]<-fclassiItalia[3]+1
}


#creazione degli istogrammi per le classi
hist(utenti_campania, breaks=classi, col=rainbow(3), main="Istogramma delle
frequenze delle classi in Campania")
hist(utenti_nazione, breaks=classi, col=rainbow(3), main="Istogramma delle
frequenze delle classi in Italia")
```

La classe modale per l'Italia è C_1 = [0, 500), per la Campania invece la classe modale risulta essere C_2 = [500, 1000).



Dopo aver considerato gli indici di posizione sono stati considerati gli indici di dispersione.

Avendo un insieme di dati numerici $(x_1 x_2 x_3 x_4 \dots \dots x_n)$, si definisce **varianza campionaria** e si indica con s^2, la quantità:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \ \ (n = 2,3 \dots)$$

Si definisce **deviazione standard campionaria** la radice quadrata della varianza ossia:

$$s \ \ = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \quad con \ \ (n = 2,3 \dots)$$

Assegnato un campione di dati numerici x_1 , x_2 , …, x_n , si definisce **coefficiente di variazione** il rapporto tra la deviazione standard campionaria e il modulo della media campionaria: $CV = \frac{s}{|\bar{x}|}$.

Il seguente codice permette di mostrare i valori della varianza, della deviazione standard e del coefficiente di variazione dei due campioni di dati.

```
> var(utenti_campania)
[1] 110369

> sd(utenti_campania)
[1] 332.2183

> coefficienteVariazioneCampania<-sd(utenti_campania)/abs(mean(utenti_campania))

> coefficienteVariazioneCampania
[1] 0.3567923

> var(utenti_nazione)
[1] 15154.27

> sd(utenti_nazione)
[1] 123.1027

> coefficienteVariazioneItalia<-sd(utenti_nazione)/abs(mean(utenti_nazione))

> coefficienteVariazioneItalia
[1] 0.3551465
```

La varianza e la deviazione standard di entrambi i campioni risultano essere dei valori grandi e da tali valori non si riesce ad avere una effettiva misura della dispersione, pertanto si considera il coefficiente di variazione. Inoltre, non è possibile da queste due misure effettuare un confronto delle dispersioni dei due campioni in quanto la media nazionale risulta avere valori numerici molto più bassi rispetto alla sola regione Campania.

Il coefficiente di variazione del campione di dati della Campania è circa **0.3567**, mentre quello della media nazionale è circa **0.3551**. I due coefficienti sono tra loro molto vicini, indicano quindi una dispersione dei dati attorno alla media molto simile. Il coefficiente di variazione di entrambi è abbastanza vicino allo 0 quindi la media risulta essere abbastanza attendibile e i singoli valori non risultano essere molto distanziati da essa.

2.3 FORMA DELLA DISTRIBUZIONE DI FREQUENZE

In questo paragrafo verranno descritti gli indici statistici che permettono di analizzare la forma della distribuzione di frequenze misurando se essa presenta asimmetrie (positive o negative) o se essa è più o meno piccata rispetto ad una distribuzione di frequenze normale standard. Prima di definire tali indici è utile introdurre il concetto di momento campionario e di momento centrato.

Assegnato un insieme di dati numerici x_1, x_2, …, x_n, si definisce **momento campionario** di ordine j la quantità:

$$M_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j$$

Assegnato un insieme di dati numerici x_1 , x_2 , …, x_n , si definisce **momento campionario centrato** attorno alla media di ordine j la quantità:

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

La skewness campionaria permette di misurare la simmetria di una distribuzione di frequenze. Assegnato un insieme di dati numerici x_1 , x_2 , …, x_n , si definisce **skewness campionaria** il valore:

$$\gamma_1 = \frac{m_3}{m_2^{3/2}}$$

Se la distribuzione è simmetrica il valore γ_1 è nullo, $\gamma_1 > 0$ se la distribuzione ha un'asimmetria positiva (ovvero una coda a destra più allungata), $\gamma_1 < 0$ se la distribuzione ha un'asimmetria negativa (ovvero una coda a sinistra più allungata).

Il codice per calcolare la skewness campionaria in R è:

```
skw <-function (x){
  n<-length (x)
  m2 <-(n -1) *var (x)/n
  m3 <- (sum ( (x- mean(x))^3) )/n
  m3/(m2 ^1.5)
}
```

Da ciò che è stato dedotto graficamente dai boxplot ci si aspetta un'asimmetria positiva in entrambi i casi ed infatti, applicando tale funzione ai due campioni di dati, si ottengono i seguenti risultati.

```
> skw(utenti_campania)
[1] 0.5893432
> skw(utenti_nazione)
[1] 0.761168
```

Entrambe le distribuzioni di frequenze hanno un'asimmetria positiva, la distribuzione di frequenza ha quindi una coda più allungata a destra.

La **curtosi campionaria** è un indice che permette di misurare la densità dei dati intorno alla media.

Assegnati un insieme di dati numerici $(x_1 x_2 x_3 \ldots . x_n)$*si definisce curtosi campionaria il valore*

$$\gamma_2 = \beta_2 - 3$$

Dove $\beta_2 = \frac{m_4}{m_2^2}$ è l'indice di Pearson e $m_2 \, ed \, m_4$ sono rispettivamente il momento centrato campionario di ordine 2 ed ordine 4.

Da notare anche che β_2 è indipendente dall'unità di misura dei dati.

Gli indici $\gamma_2 e \, \beta_2$ permettono di confrontare la curva dei dati con una densità di probabilità normale standard

- Se $\beta_2 < 3 \, e \, quindi \, \gamma_2 < 0$ abbiamo una distribuzione di frequenze platicurtica, è quindi più piatta di una normale;
- Se $\beta_2 > 3 \, e \, quindi \, \gamma_2 > 0$ abbiamo una distribuzione di frequenze leptocurtica, è quindi più piccata di una normale;
- Se $\beta_2 = 3 \, e \, quindi \, \gamma_2 = 0$ abbiamo una distribuzione di frequenze normocurtica, è quindi piatta come una normale.

Il codice per calcolare la curtosi in R è:

```
curt <-function (x){
  n <-length (x)
  m2 <-(n -1) *var (x)/n
  m4 <- (sum ((x-mean(x))^4) )/n
  m4/(m2 ^2) -3
}
```

Applicando tale funzione ai due campioni di dati considerati si ottiene:

```
> curt(utenti_campania)
[1] -0.9366374
> curt(utenti_nazione)
[1] -0.6986567
```

Il valore di entrambe le curtosi campionarie è negativo quindi entrambe le distribuzioni di frequenza sono meno piccate di una distribuzione di frequenze normale standard.

Confrontando i valori ottenuti da questi due indici si ha un'ulteriore conferma del fatto che l'andamento negli anni delle due curve considerate risulta essere molto simile anche se i dati relativi all'intera nazione sono più bassi in quanto sono ottenuti dalla media di tutte le regioni, che viene fortemente influenzata dai valori bassi presenti in molte regioni con meno abitanti rispetto alla Campania.

3 STATISTICA DESCRITTIVA BIVARIATA

In questo capitolo verranno mostrate le analisi di regressione lineare semplice e di regressione lineare multipla calcolando il modello lineare, i residui e il coefficiente di determinazione. Le analisi verranno effettuate sulla tabella Utenti e si cercherà di individuare eventuali correlazioni lineari tra i vari anni considerati.

La statistica descrittiva bivariata si occupa dei metodi grafici e statistici atti a descrivere le relazioni che intercorrono tra due variabili X e Y.

3.1 REGRESSIONE LINEARE SEMPLICE

Il modello di regressione lineare semplice viene utilizzato per spiegare la relazione che esiste tra una variabile dipendente Y e una variabile indipendente X. In questa analisi verrà considerata come variabile indipendente l'anno 2019 e come variabile dipendente l'anno 2020.

Regioni	2019	2020
Piemonte	524	958
Valle d'Aosta	6	14
Liguria	160	296
Lombardia	939	2.055
Trentino-Alto Adige	63	121
Trento	47	74
Bolzano	14	43
Veneto	506	868
Friuli-Venezia Giulia	103	163
Emilia-Romagna	365	804
Toscana	514	841
Umbria	83	125
Marche	141	286
Lazio	919	1.898
Abruzzo	144	235
Molise	16	43
Campania	772	1.492
Puglia	344	771
Basilicata	42	67
Calabria	115	200
Sicilia	438	859
Sardegna	149	322

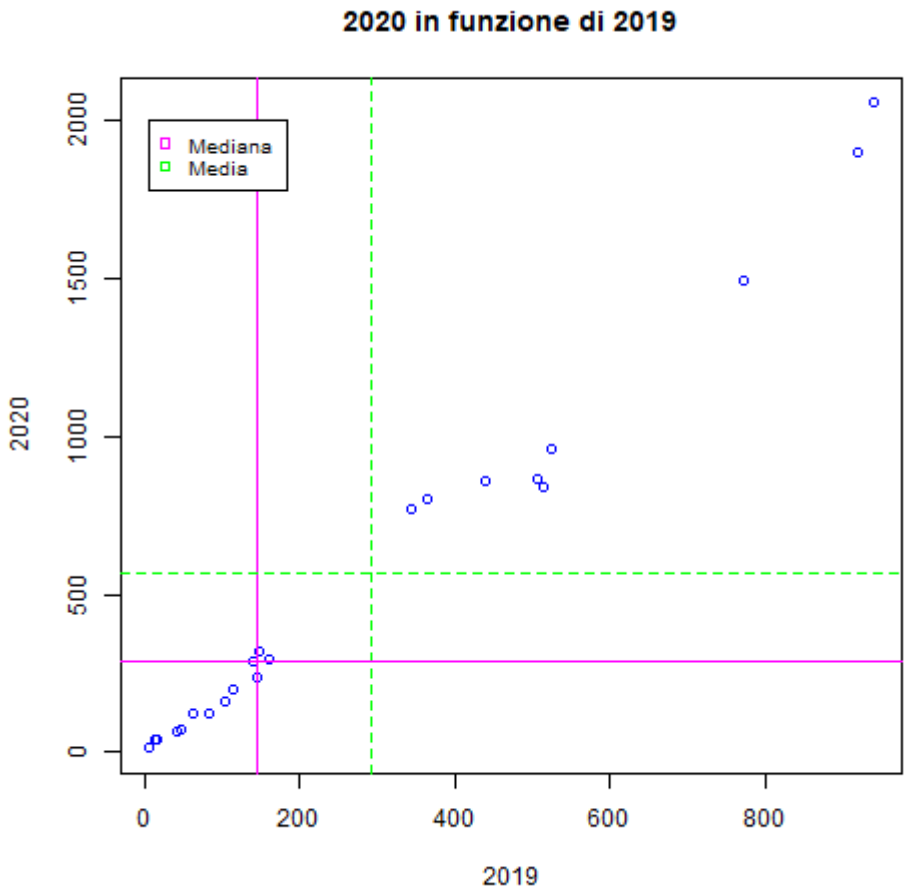
Si calcolano gli indici di posizione e di dispersione relativi alle due coppie di variabili.

```
> median(df$"2019")
[1] 146.5
> mean(df$"2019")
[1] 291.1
> sd(df$"2019")
[1] 294.7
> median(df$"2020")
[1] 291
> mean(df$"2020")
[1] 569.8
> sd(df$"2020")
[1] 605.8
```

Si nota che sia mediana, sia media che deviazione standard sono maggiori per la variabile Y.

Un primo passo per indagare l'eventuale dipendenza tra due variabili X e Y consiste nel disegnare il diagramma di dispersione o scatterplot. Il grafico che si ottiene mira ad evidenziare se le coppie di punti presentano qualche forma di regolarità.

Nello scatterplot si pone sull'asse delle ascisse la variabile indipendente 2019 e sulle ordinate la variabile dipendente 2020. Vengono poi tracciate delle linee orizzontali e verticali in corrispondenza delle mediane e delle medie delle due variabili.



Dallo scatterplot si nota che i dati (a parte qualche punto che si discosta un po' di più) sono posizionati lungo una retta ascendente quindi si può dedurre che esiste una correlazione lineare positiva tra le due variabili considerate.

Per ottenere una misura quantitativa della correlazione tra le variabili si calcola la covarianza campionaria, che è così definita:

Assegnato un campione bivariato $(x_1,y_1),(x_2,y_2), \ldots,(x_n,y_n)$ *di una variabile quantitativa bi-dimensionale (X,Y), siano* $\bar{x}$ *e* $\bar{y}$ *rispettivamente le medie campionarie di* $x_1,x_2, \ldots,x_n$ *e di* $y_1,y_2, \ldots,y_n$*. La covarianza campionaria tra le due variabili X e Y è così definita:*

$$C_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})\ (y_i - \bar{y})$$

Se $C_{xy} > 0$ le variabili sono correlate positivamente, se $C_{xy} < 0$ le variabili sono correlate negativamente, se $C_{xy} = 0$ le variabili non sono correlate.

```
> cov(df$"2019", df$"2020")
[1] 177155
```

La covarianza tra le due variabili risulta essere 177155, pertanto esiste una correlazione lineare positiva tra le due variabili come si poteva già intuite dal grafico dello scatterplot.

Per ottenere una misura quantitativa della correlazione tra le variabili si può anche considerare il coefficiente di correlazione campionario, che è così definito:

Assegnato un campione bivariato $(x_1,y_1),(x_2,y_2), \ldots,(x_n,y_n)$ *di una variabile quantitativa bidimensionale (X,Y), siano* $\bar{x}$ *e* s_x *la media campionaria e la deviazione standard di* $x_1,x_2, \ldots,x_n$ *ed inoltre siano* $\bar{y}$ *e* s_y *la media campionaria e la deviazione standard di* $y_1,y_2, \ldots,y_n$*. Il* ***coefficiente di correlazione campionario*** *tra le due variabili X e Y è cosi definito:*

$$r_{xy} = \frac{C_{xy}}{s_x s_y}$$

Il coefficiente di correlazione campionario r_{xy} misura la forza del legame di natura lineare esistente tra due variabili quantitative. In particolare, $-1 \leq r_{xy} \leq 1$ e il suo valore indica la direzione della retta interpolante.

- $r_{xy} = -1$: (correlazione perfetta negativa), tutti i punti sono allineati lungo una retta discendente;
- $-1 < r_{xy} < 0$ (correlazione negativa), i punti sono posizionati in una nuvola attorno ad una retta interpolante discendente;

- $r_{xy} = 0$: (nessuna correlazione), i punti sono completamente dispersi in una nuvola che non presenta alcuna evidente direzione di natura lineare;
- $0 < r_{xy} < 1$: (correlazione positiva), i punti sono posizionati in una nuvola attorno ad una retta interpolante ascendente;
- $r_{xy} = 1$: (correlazione perfetta positiva), tutti i punti sono allineati lungo una retta ascendente.

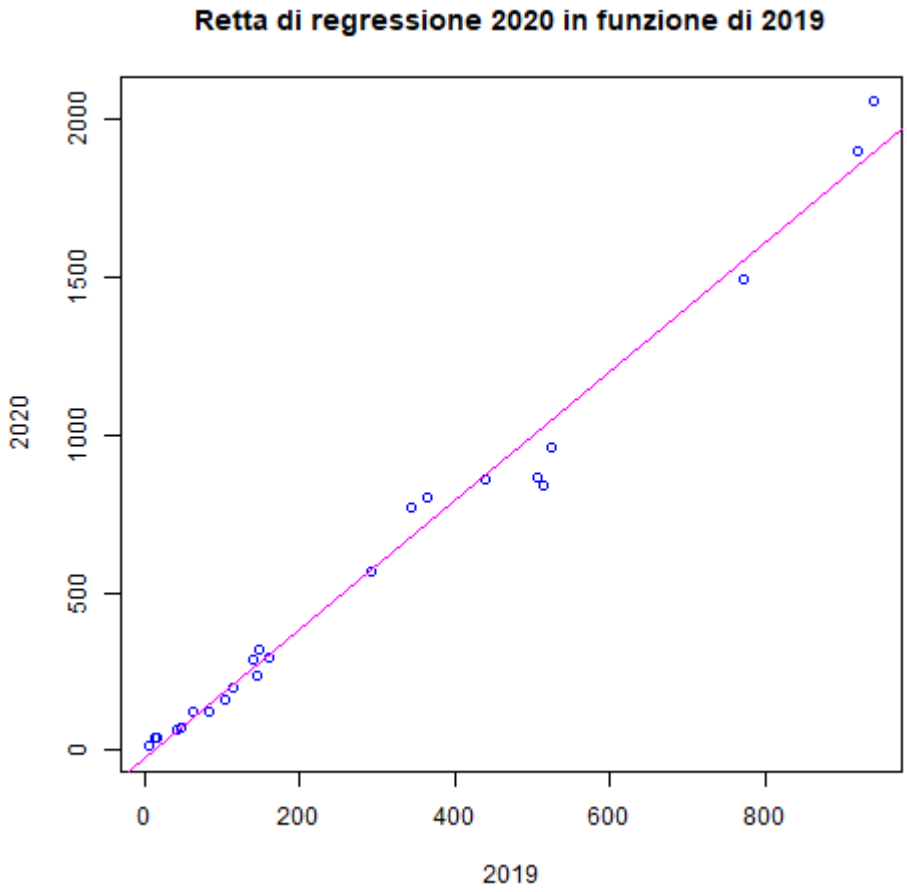
```
> cor(df$"2019", df$"2020")
[1] 0.9924
```

Siccome il coefficiente di correlazione è uguale a 0.9924 che è prossimo ad 1 i dati presentano un'altissima correlazione positiva. Per calcolare la retta interpolante di questi punti si utilizza il modello di regressione lineare semplice.

Il modello di regressione lineare semplice è esprimibile tramite l'equazione di una retta $Y = \alpha + \beta X$ che riesce ad interpolare la nuvola di punti dello scatterplot meglio di tutte le altre possibili rette, dove α è l'intercetta e β è il coefficiente angolare. Se β > 0 la retta di regressione è crescente, se β < 0 la retta di regressione è discendente, se β = 0 la retta è orizzontale. L'intercetta α corrisponde invece al punto di intersezione della retta interpolante con l'asse delle ordinate.

Il seguente codice permette di realizzare lo scatterplot relativo ai dati del 2019 e del 2020 con la retta interpolante stimata.

```
plot(df$"2019", df$"2020", main="Retta di regressione 2020 in funzione di
2019", col="blue",
     xlab="2019", ylab="2020")
abline(lm(df$"2020"~df$"2019"), col="magenta")
```



La funzione `lm(y~x)` permette di eseguire le analisi di regressione lineare della variabile dipendente y in funzione della variabile indipendente x.

```
> linearmodel<-lm(df$"2020"~df$"2019")
> linearmodel

Call:
lm(formula = df$"2020" ~ df$"2019")

Coefficients:
(Intercept)    df$"2019"
     -23.95         2.04
```

L'intercetta α vale -23.97, mentre il coefficiente angolare β vale 2.04. Siccome il coefficiente angolare ha segno positivo, la retta è ascendente. L'equazione della retta interpolante risulta pertanto:

$$Y = -23.97 + 2.04x$$

Dopo aver calcolato la retta interpolante, è possibile notare che esistono degli scostamenti tra i valori osservati del campione e i valori stimati attraverso la retta di regressione. Le differenze tra le ordinate dei punti dei valori osservati e le ordinate dei punti dei valori stimati prendono il nome di residui. Se si indica con y_i il valore osservato e con $\widehat{y_i}$ il valore stimato, i **residui** sono così definiti:

$$E_i = y_i - \widehat{y_i} = y_i - (\alpha + \beta x_i) \quad (i = 1,2,\dots,n)$$

La media campionaria dei residui è nulla.

Il codice seguente permette di visualizzare i valori stimati.

```
> linearmodel$fitted.values
       1        2        3        4        5        6        7        8        9       10       11       12
1044.843  -11.735  302.383 1891.330  104.530   71.894    4.583 1008.128  186.119  720.527 1024.446  145.324
      13       14       15       16       17       18       19       20       21       22
 263.628 1850.535  269.748    8.663 1550.695  677.693   61.696  210.596  869.427  279.946
```

Il seguente codice permette di visualizzare i residui, ossia di quanto le ordinate dei valori osservati si discostano dai valori stimati.

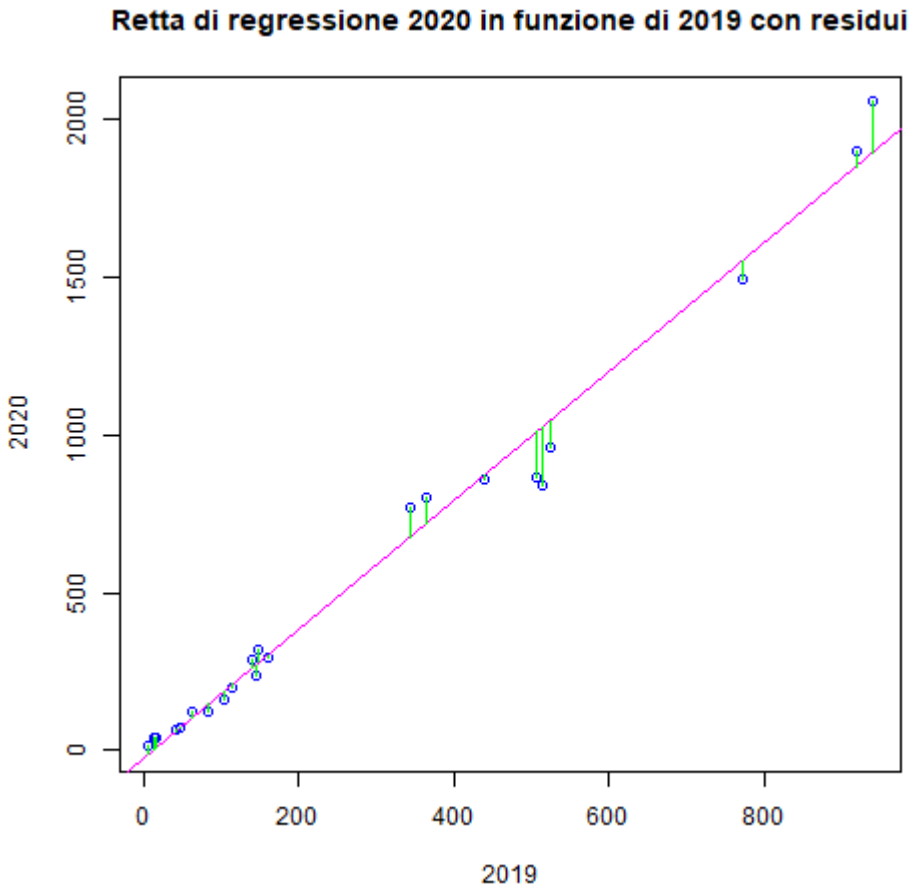
```
> linearmodel$residuals
       1        2        3        4        5        6        7        8        9       10       11
 -86.843   25.735   -6.383  163.670   16.470    2.106   38.417 -140.128  -23.119   83.473 -183.446
      12       13       14       15       16       17       18       19       20       21       22
 -20.324   22.372   47.465  -34.748   34.337  -58.695   93.307    5.304  -10.596  -10.427   42.054
>
```

La mediana, la varianza e la deviazione standard dei residui assumono i seguenti valori. Non è possibile calcolare il coefficiente di variazione in quanto la media campionaria dei residui è 0.

```
> median(linearmodel$residuals)
[1] 3.705
> var(linearmodel$residuals)
[1] 5586
> sd(linearmodel$residuals)
[1] 74.74
```

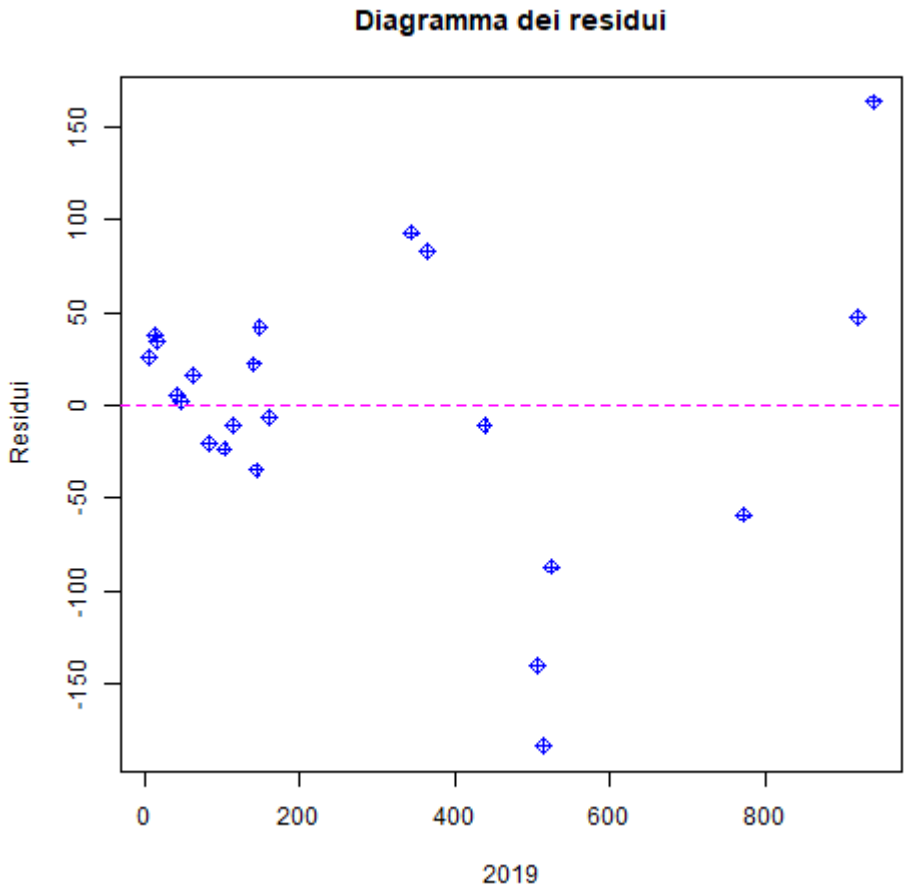
Di seguito viene mostrato il grafico che rappresenta lo scatterplot dei punti, la retta di regressione e i segmenti verticali che rappresentano i residui.

```
plot(df$"2019", df$"2020", main="Retta di regressione 2020 in funzione di 2019
con residui", col="blue",
     xlab="2019", ylab="2020")
abline(linearmodel, col="magenta")
segments (df$"2019", linearmodel$fitted.values, df$"2019", df$"2020"
,col="green")
```



Un esame più accurato del modo con cui la retta di regressione interpola i dati e di come i residui si dispongano intorno alla retta interpolante influenzandone la posizione, può essere ottenuto attraverso il diagramma dei residui che è un grafico in cui i valori dei residui sono posti sull'asse delle ordinate e quelli della variabile indipendente sull'asse delle ascisse.

```
plot(df$"2019", residui, main="Diagramma dei residui", xlab="2019",
ylab="Residui", col="blue", pch =9)
abline (h=0, col ="magenta",lty=2)
```



La linea tratteggiata è posizionata su 0 che indica la media campionaria dei residui. Si nota che i punti sono disposti casualmente attorno alla retta orizzontale e non si evidenzia nessun comportamento particolare nella distribuzione dei punti. La posizione della retta di regressione è fortemente influenzata dalla presenza di eventuali valori anomali che si discostano in modo significativo dagli altri. L'analisi dei residui aiuta ad individuare eventuali punti isolati (valori anomali) dovuti ad errori nella stima. Tali valori possono perturbare significativamente la stima dei parametri di regressione e influenzare l'interpretazione dei residui. Eliminando i valori anomali la varianza campionaria dei residui diminuisce.

Spesso è utile calcolare i residui standardizzati, così definiti:

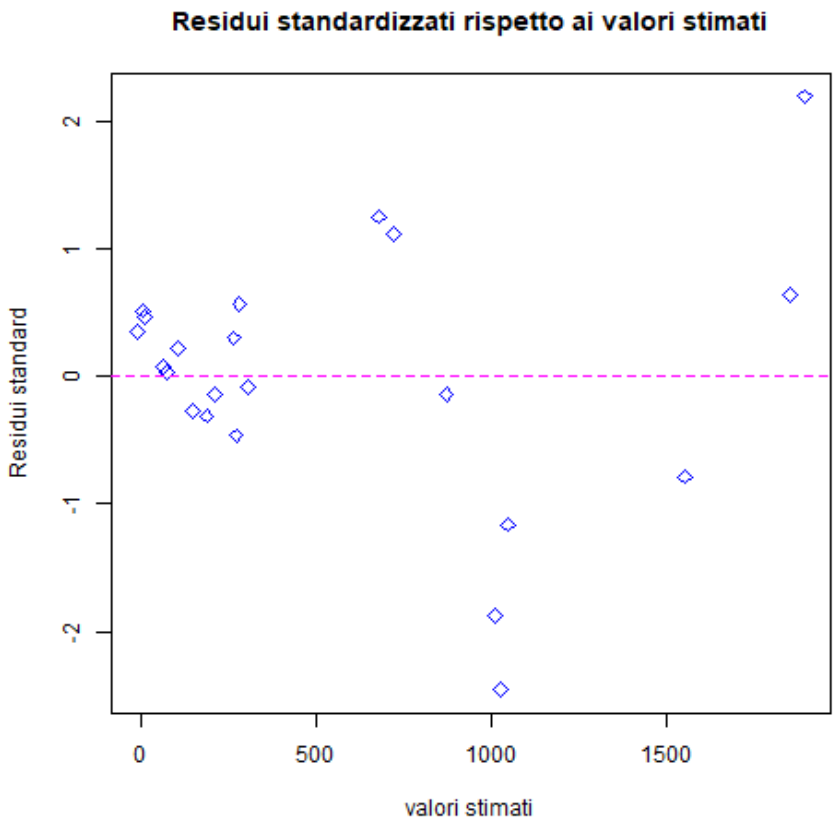
$$E_i^{(s)} = \frac{E_i - \bar{E}}{s_E} = \frac{E_i}{s_E}$$

I residui standardizzati sono caratterizzati da media nulla e varianza unitaria.

```
residui<-linearmodel$residuals
residuistandard<-residui/sd(residui)
```

```
> residuistandard
      1       2       3       4       5       6       7       8       9      10      11
-1.1619  0.3443 -0.0854  2.1899  0.2203  0.0281  0.5140 -1.8749 -0.3093  1.1168 -2.4545
     12      13      14      15      16      17      18      19      20      21      22
-0.2719  0.2993  0.6350 -0.4648  0.4593 -0.7852  1.2483  0.0709 -0.1417 -0.1394  0.5625
```

Successivamente è stato realizzato un grafico che mostra sulle ordinate i residui standardizzati e sulle ascisse i valori stimati.



La maggior parte dei residui standardizzati si concentra nell'intervallo [-1,1] e sono quei residui in corrispondenza di quei valori osservati che hanno lo stesso andamento dei valori attesi. Ci sono comunque valori che si discostano maggiormente dai propri valori attesi come Lombardia e Toscana. Per la Lombardia la differenza di ordinate tra valore osservato e valore stimato è circa 2.18 che è positivo, questo vuol dire che si è avuto un aumento di chiamate nel 2020 maggiore rispetto a quanto stimato dalla retta di regressione. Per la Toscana, invece, questa differenza vale circa -2.45 e si è quindi avuto un numero di chiamate nel 2020 più basso rispetto a quello stimato dalla retta di regressione (anche se è comunque un numero più alto rispetto all'anno precedente).

Per conoscere quanto la retta di regressione si adatta ai dati considerati si calcola il **coefficiente di determinazione** che viene definito come il rapporto tra la varianza dei valori osservati e la varianza dei valori stimati. Un coefficiente di determinazione prossimo ad 1 indica che tutti i punti tendono ad allinearsi lungo la retta di regressione, mentre un coefficiente di determinazione prossimo a 0 indica una completa incapacità della retta di rappresentare la distribuzione dei dati considerati. Per

il modello di regressione lineare semplice il coefficiente di determinazione corrisponde al quadrato del coefficiente di correlazione. È possibile ottenere il valore del coefficiente di determinazione in questo modo:

```
> summary(linearmodel)$r.square
[1] 0.9848
```

In questo caso il coefficiente di correlazione vale **0.9848**. Siccome è prossimo ad 1, significa che la retta descrive bene i dati considerati, infatti anche dai grafici visti precedentemente si nota che gli scostamenti dalla retta sono molto piccoli.

3.2 REGRESSIONE LINEARE MULTIPLA

Il modello di regressione lineare multipla viene utilizzato per spiegare la relazione tra una variabile quantitativa Y detta variabile dipendente e le variabili quantitative indipendenti X_1, X_2, …, X_p.

Il data frame considerato è il seguente e si utilizza il modello di regressione lineare multipla per spiegare la relazione le variabili indipendenti: 2013, 2014, 2015, 2016, 2017, 2018, 2019 e la variabile dipendente: 2020.

Regioni	2013	2014	2015	2016	2017	2018	2019	2020
Piemonte	667	681	543	441	408	511	524	958
Valle d'Aosta	17	15	8	9	4	6	6	14
Liguria	298	259	200	184	106	180	160	296
Lombardia	1.488	1.174	886	878	822	1.034	939	2.055
Trentino-Alto Adige	66	53	48	54	22	49	63	121
Trento	45	41	41	42	18	36	47	74
Bolzano	20	12	7	12	4	13	14	43
Veneto	751	636	397	306	273	429	506	868
Friuli-Venezia Giulia	154	124	99	70	65	93	103	163
Emilia-Romagna	654	492	341	311	216	352	365	804
Toscana	594	484	298	273	311	398	514	841
Umbria	161	121	79	76	44	89	83	125
Marche	258	164	174	154	114	137	141	286
Lazio	1.306	1.304	838	833	593	959	919	1.898
Abruzzo	261	186	176	128	118	184	144	235
Molise	46	31	30	27	18	13	16	43
Campania	1.316	1.059	815	635	537	823	772	1.492
Puglia	768	677	391	350	246	450	344	771
Basilicata	61	62	40	25	31	39	42	67
Calabria	210	135	122	99	60	125	115	200
Sicilia	922	701	446	481	390	492	438	859
Sardegna	356	296	226	184	172	157	149	322

Di seguito vengono mostrati i valori degli indici di posizione e di dispersione (mediana, media e deviazione standard) relativi alle variabili considerate.

```
> apply(df, 2, median)
 2013  2014  2015  2016  2017  2018  2019  2020
279.5 222.5 188.0 169.0 116.0 168.5 146.5 291.0
> round(apply(df, 2, mean),2)
 2013  2014  2015  2016  2017  2018  2019  2020
473.6 395.8 282.1 253.3 207.8 298.6 291.1 569.8
> round(apply(df, 2, sd),2)
 2013  2014  2015  2016  2017  2018  2019  2020
457.9 397.8 276.1 258.6 222.8 310.5 294.7 605.8
```

Media, mediana e deviazione standard sono maggiori per la variabile 2020.

Di seguito viene mostrata la matrice delle covarianze che contiene sulla diagonale principale la varianza delle singole colonne del dataframe, mentre gli altri elementi rappresentano le covarianze tra le coppie di variabili.

```
> cov(df)
       2013   2014   2015   2016   2017   2018   2019   2020
2013 209691 179943 124387 116136  98978 140379 131434 271793
2014 179943 158218 108475 101339  85271 122308 114911 237064
2015 124387 108475  76217  70507  60062  84984  79597 164413
2016 116136 101339  70507  66883  56614  79641  74193 154780
2017  98978  85271  60062  56614  49627  67978  63971 132424
2018 140379 122308  84984  79641  67978  96391  90467 187109
2019 131434 114911  79597  74193  63971  90467  86853 177155
2020 271793 237064 164413 154780 132424 187109 177155 366934
```

Da questa matrice si può notare come tutte le coppie di variabili siano tra di loro positivamente correlate. Tali numeri sono però elevati e non suggeriscono quanto sia forte il legame tra le variabili pertanto viene considerato il coefficiente di correlazione.

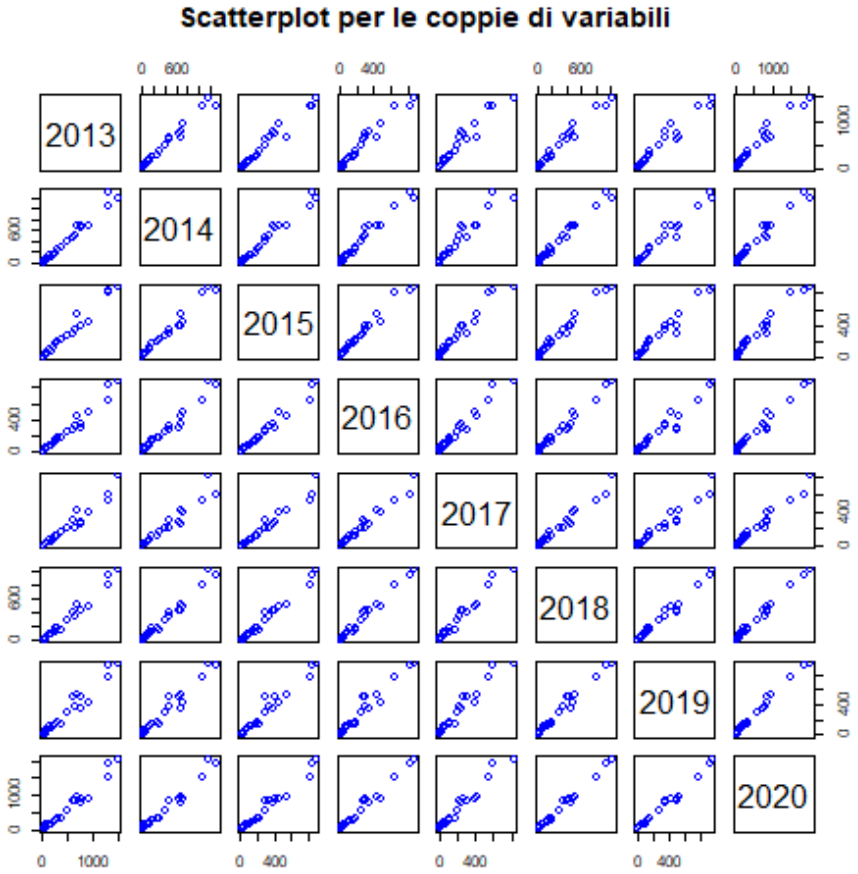
Di seguito viene quindi mostrata la matrice delle correlazioni che contiene tutte le correlazioni lineari tra le coppie di variabili, ossia misura la forza del legame di natura lineare esistente tra tutte le coppie di variabili quantitative. La matrice delle correlazioni contiene 1 sulla diagonale principale.

```
> cor(df)
       2013   2014   2015   2016   2017   2018   2019   2020
2013 1.0000 0.9879 0.9839 0.9807 0.9703 0.9874 0.9739 0.9798
2014 0.9879 1.0000 0.9878 0.9851 0.9623 0.9904 0.9803 0.9839
2015 0.9839 0.9878 1.0000 0.9875 0.9766 0.9915 0.9783 0.9831
2016 0.9807 0.9851 0.9875 1.0000 0.9827 0.9919 0.9735 0.9880
2017 0.9703 0.9623 0.9766 0.9827 1.0000 0.9829 0.9744 0.9813
2018 0.9874 0.9904 0.9915 0.9919 0.9829 1.0000 0.9887 0.9949
2019 0.9739 0.9803 0.9783 0.9735 0.9744 0.9887 1.0000 0.9924
2020 0.9798 0.9839 0.9831 0.9880 0.9813 0.9949 0.9924 1.0000
```

Si nota che esiste una forte correlazione lineare tra tutte le variabili considerate.

Il seguente grafico visualizza in un'unica finestra tutti gli scatterplot ottenuti mettendo in relazione le varie coppie di variabili.

```
pairs(df, main="Scatterplot per le coppie di variabili", col="blue")
```



Lo scatterplot permette di visualizzare graficamente la correlazione positiva esistente tra le varie coppie di variabili. Quasi tutti i punti, infatti, sono posizionati lungo una retta interpolante crescente.

Il modello di regressione lineare multipla con p variabili indipendenti è esprimibile attraverso l'equazione:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Dove:

- α è l'intercetta, ossia il valore di Y quando $X_1=X_2= \ldots = X_p=0$;
- $\beta_1, \beta_2, \ldots, \beta_p$ sono i regressori. In particolare, β_1 rappresenta l'inclinazione di Y rispetto alla variabile X_1 tenendo costanti le variabili X_2, X_3, …,X_p, …, β_p rappresenta l'inclinazione di Y rispetto alla variabile X_p tenendo costanti le variabili X_1, X_2, …,X_{p-1}.

Utilizzando il modello di regressione lineare multipla si ottiene:

```
multiplelinearmodel<-lm(df$"2020"~df$"2013" +df$"2014" + df$"2015" + df$"2016"
+ df$"2017" + df$"2018" + df$"2019")
```

```
Call:
lm(formula = df$"2020" ~ df$"2013" + df$"2014" + df$"2015" +
    df$"2016" + df$"2017" + df$"2018" + df$"2019")

Coefficients:
(Intercept)    df$"2013"    df$"2014"    df$"2015"    df$"2016"    df$"2017"    df$"2018"    df$"2019"
   -18.6585       0.0408      -0.2703      -0.3768       0.9871      -0.2231       0.8790       1.0864
```

Da cui si ricava che l'intercetta è -18.6585 e i regressori sono: 0.0408, -0.2703, -0.3768, 0.9871, -0.2231, 0.8790, 1.0864. Pertanto, il modello di regressione lineare multipla è descritto dall'equazione:

$$Y = -18.6585 + 0.0408X_1 - 0.2703X_2 - 0.3768X_3 + 0.9871X_4 - 0.2231X_5 + 0.8790X_6 + 1.0864X_7$$

I segni dei regressori β_1, β_4, β_6, β_7 sono positivi: questo indica che all'aumentare del numero di utenti nel 2013, 2016, 2018 e 2019 aumenta il numero di utenti nel 2020. Mentre i regressori β_2, β_3 , β_5 sono negativi quindi all'aumentare del numero di utenti nel 2014, 2015, 2017 diminuisce il numero di utenti nel 2020.

Il regressore β_1=0.0408 è prossimo allo zero, questo indica che il numero di utenti nel 2013 non incide in maniera significativa il numero di utenti nel 2020.

Il codice seguente permette di visualizzare i valori stimati rispetto al modello di regressione lineare multipla.

```
> stime<-multiplelinearmodel$fitted.values
> stime
      1       2       3       4       5       6       7       8       9      10      11
982.634  -5.249 338.165 2003.217 111.534  76.796  13.866 858.472 165.052 711.315 870.837
     12      13      14      15      16      17      18      19      20      21      22
159.048 282.167 1897.772 293.605  14.981 1510.863 742.257  49.674 226.604 957.550 273.841
```

Il seguente codice permette di visualizzare i residui.

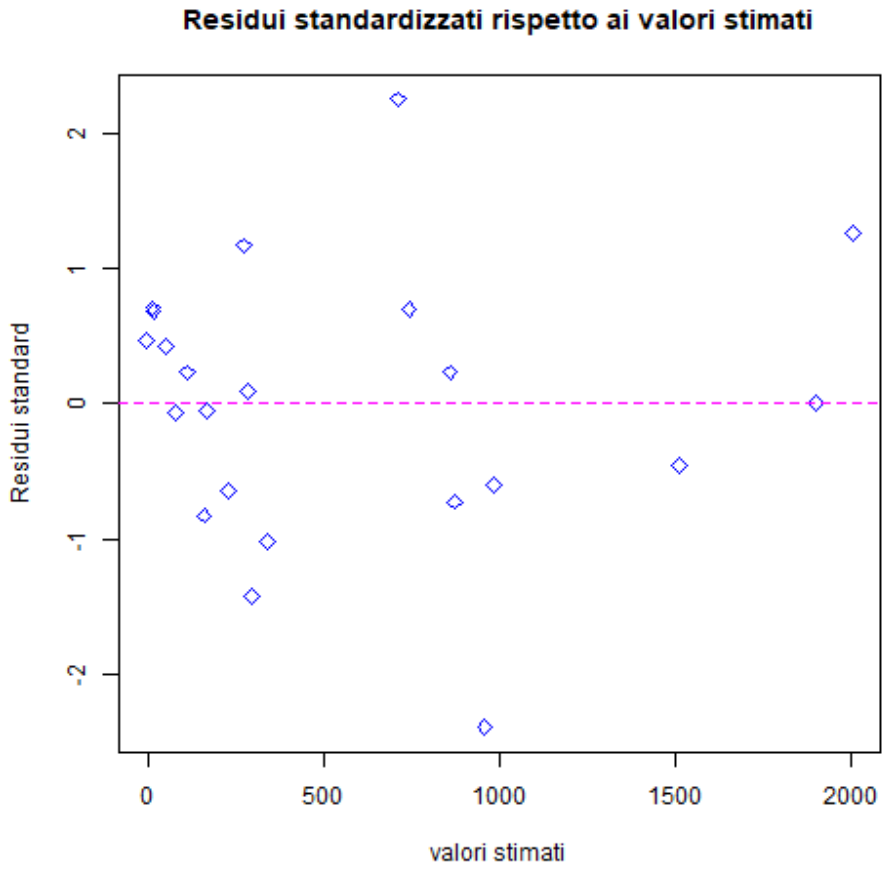
```
> residui<-multiplelinearmodel$residuals
> residui
      1       2       3       4       5       6       7       8       9      10      11
-24.6342 19.2492 -42.1648 51.7831  9.4656 -2.7958 29.1336  9.5281 -2.0521 92.6853 -29.8366
     12      13      14      15      16      17      18      19      20      21      22
-34.0481  3.8331  0.2282 -58.6050 28.0186 -18.8627 28.7428 17.3262 -26.6039 -98.5499 48.1591
```

Successivamente sono stati calcolati i residui standardizzati.

```
> residuistandard<-residui/sd(residui)
> residuistandard
         1          2          3          4          5          6          7          8          9         10
-0.598528  0.467691 -1.024462  1.258155  0.229982 -0.067928  0.707847  0.231501 -0.049858  2.251939
        11         12         13         14         15         16         17         18         19         20
-0.724928 -0.827254  0.093131  0.005545 -1.423903  0.680757 -0.458299  0.698353  0.420968 -0.646384
        21         22
-2.394428  1.170103
```

Di seguito viene mostrato il grafico che rappresenta i residui standardizzati in funzione dei valori stimati.

```
plot(stime, residuistandard, main="Residui standardizzati rispetto ai valori
stimati", xlab="valori stimati"
     , ylab="Residui standard", pch=5, col="blue")
abline (h=0, col ="magenta",lty =2)
```



La linea tratteggiata è posizionata su 0 che indica la media campionaria dei residui. Si nota che i punti sono disposti casualmente attorno alla retta orizzontale e non si evidenzia nessun comportamento particolare nella distribuzione dei punti. La maggior parte dei punti sono concentrati nell'intervallo [-1,1] pertanto gli scostamenti dei valori osservati rispetto ai valori stimati risultano essere molto bassi. Solo per qualche regione tali scostamenti sono più elevati come Emilia-Romagna e Sicilia.

Anche in questo caso il coefficiente di determinazione è prossimo ad 1, infatti vale **0.9954**. Il modello di regressione lineare multipla descrive bene i dati considerati.

```
> summary(multiplelinearmodel)$r.square
[1] 0.9954
```

4 ANALISI DEI CLUSTER

L'**analisi dei cluster** è una tecnica matematica usata in informatica e altre discipline, essa si basa sul considerare diversi tipi di dati (numerici, persone, misure) ed unirli in gruppi che contengono tutti elementi che hanno somiglianze tra di loro. La creazione dei cluster può essere effettuata con diversi metodi, ma tutte le tecniche hanno in comune lo scopo di rendere quanto più possibili omogenei gli elementi all'interno di un gruppo e rendere quanto più eterogenei gruppi diversi così che il grado di associazione sia alto tra membri dello stesso gruppo e basso tra membri di gruppi diversi.

Le tecniche di raggruppamento tendono ad unire quei dati che sono tra di loro simili e svolgono questo lavoro basandosi sul concetto che ogni elemento di un certo insieme di dati ha delle caratteristiche osservabili che possono essere il colore degli occhi per le persone, o possono essere le denunce al numero verde 1522 fatte di anno in anno per una regione.

Per effettuare il partizionamento in cluster occorre definire delle misure di distanza o similarità tra i vari individui in base alle caratteristiche che si vogliono considerare. Una funzione a valori reali $(X_i X_j)$ è detta funzione distanza se e solo se soddisfa le seguenti condizioni:

- $d(X_i X_j) = 0$ se e solo se $X_i = X_j$ in E_p;
- $d(X_i X_j) \geq 0$ per ogni $X_i \; e \; X_j$ in E_p;
- $d(X_i X_j) = d(X_j X_i)$ per ogni $X_i \; e \; X_j$ in E_p;
- $d(X_i X_j) \leq d(X_i, \; X_k) + d(X_k, X_j)$ per ogni $X_i, X_k \; e \; \; X_j$ in E_p. (disuguaglianza triangolare)

In generale, verrà definita una matrice D contenente le distanze tra tutte le possibili coppie di individui.

Una funzione a valori reali $s_{ij} = s(X_i X_j)$ è detta misura di similarità se e soltanto se soddisfa le seguenti condizioni:

- $s(X_i X_i) = 1$;
- $0 \leq s(X_i X_j) \leq 1$;
- $s(X_i X_j) = s(X_j X_i)$ per ogni $X_i \; e \; X_j$.

È sempre possibile trasformare una misura di distanza in una misura di similarità, ma non viceversa in quanto le misure di similarità non godono della proprietà di disuguaglianza triangolare di cui invece godono le misure di distanza.

Per effettuare il partizionamento in cluster è stata utilizzata una misura di distanza, in particolare è stata utilizzata la **metrica Euclidea** così definita:

$$d_2(X_i X_j) = [\sum_{k=1}^{p} (x_{ik} - x_{jk})^2]^{1/2}$$

Dove x_{ik} è il valore della k-esima caratteristica dell'individuo i.

Per effettuare il partizionamento in cluster un primo approccio a cui si potrebbe pensare è quello di considerare tutte le possibili suddivisioni. Tali metodi vengono detti metodi di enumerazione completa. Il numero totale di partizionare n individui in m cluster è dato dal numero di Stirling del secondo tipo così definito:

$$S(n,m) = \frac{1}{m!} \sum_{k=0}^{m} \binom{m}{k} (-1)^k (m-k)^n$$

Il codice per calcolare il numero di Stirling del secondo tipo in R è:

```
stirling2 <-function (n,m){
  s<-0
  if ((m >=1)&(m <=n)){
    for (k in seq (0,m)){
      s<-s+( choose (m,k)*(-1)^k*(m-k)^n/ factorial (m))}
    return (c(s))
  }
}
```

Se si volesse utilizzare tale metodo per gli individui considerati (22) per partizionarli in 2 cluster il numero di possibili partizionamenti sarebbe 2097151.

```
> nrow(Z)
[1] 22
>
> stirling2(nrow(Z),2)
[1] 2097151
```

Con 3 cluster il numero di possibili partizionamenti sarebbe ancora più elevato.

```
> stirling2(nrow(Z),3)
[1] 5228079450
```

Tali metodi risultano essere quindi molto onerosi, per questo vengono utilizzati i metodi non gerarchici e i metodi gerarchici.

- **Metodi gerarchici**: mirano a costruire gerarchie di cluster; si dividono in due tipologie: l'approccio agglomerativo è un approccio "bottom-up", si parte dall'inserire ogni elemento in un singolo cluster e si procede ad accorparli a due a due; l'approccio divisivo è un approccio "top-down" che parte da un singolo cluster che comprende tutti gli elementi e viene diviso in tanti sotto cluster. Tutti i metodi gerarchici producono una struttura ad albero chiamata "dendogramma". I metodi gerarchici hanno due vantaggi:
    - Forniscono una visione completa dell'insieme in termini di distanze;
    - Non comportano la scelta a priori del numero di cluster oppure la scelta a priori del numero di parametri da utilizzare per la determinazione automatica del loro numero.

  Uno svantaggio è che essi non consentono di riallocare gli individui che sono stati già classificati ad un livello precedente dell'analisi.
- **Metodi non gerarchici**: permettono di riposizionare elementi di un cluster qualora venga notato che un elemento piazzato in cluster conviene spostarlo in un altro, di questo metodo fa parte l'algoritmo k-means.

Per la suddivisione in cluster si è scelto inizialmente di considerare la suddivisione in 2 cluster. In seguito, si è deciso di effettuare un'ulteriore suddivisione in 3 cluster e di confrontare i risultati ottenuti. Tuttavia, al posto di considerare il data frame con i dati originali, si è scelto di scalarli sottraendo la media e dividendo per la deviazione standard, ottenendo dei dati standardizzati e più piccoli che risultano anche più semplici da gestire.

Il seguente codice permette di calcolare la matrice delle distanze euclidee a partire dal data frame Z scalato.

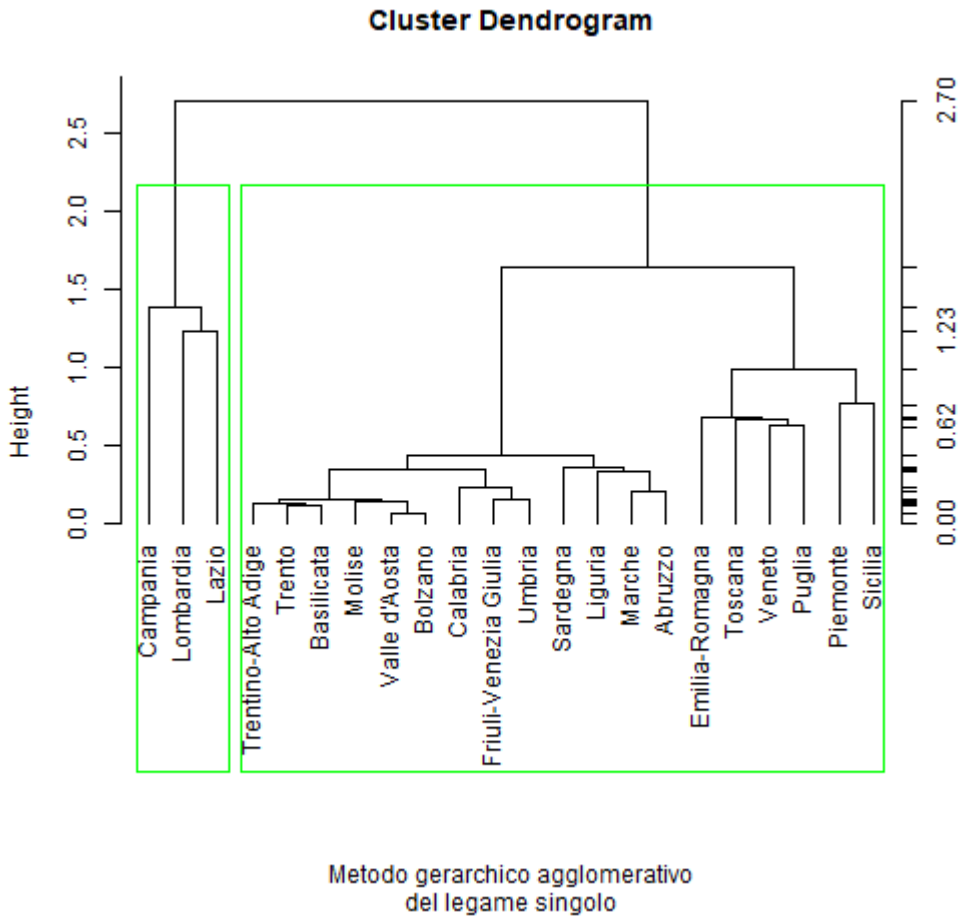
```
d<-dist(Z, method="euclidean", diag=TRUE, upper=TRUE)
```

4.1 METODI GERARCHICI

Metodo del legame singolo

In questo metodo la distanza tra i gruppi G1 (contenente n1 individui) e G2 (contenente n2 individui) è definita come la minima tra tutte le n1 n2 distanze che si possono calcolare tra ogni individuo di G1 e ogni individuo di G2.

```
hls<-hclust(d, method="single")
png("grafici/cluster/dendrogrammaUtenti_LegameSingolo.png")
plot(hls, hang=-1, xlab="Metodo gerarchico agglomerativo", sub="del legame
singolo")
rect.hclust(hls, k=2, border="green")
axis(side=4, at=round(c(0, hls$height),2))
dev.off()
```

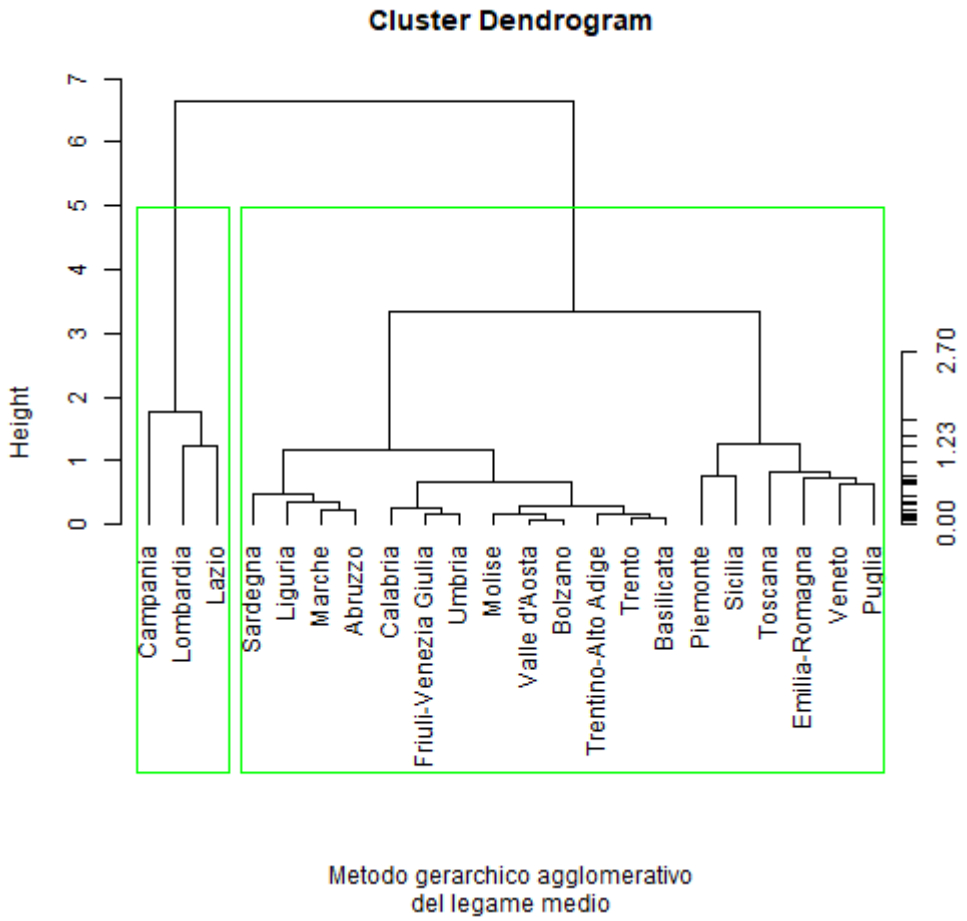


Occorre sottolineare che il metodo del legame singolo è in grado di individuare cluster di qualsiasi forma ma può dare origine alla formazione di una catena di individui.

Metodo del legame medio

Nel metodo del legame medio si considera, come distanza tra due gruppi, la media di tutte le distanze calcolate a due a due tra tutti gli elementi dei due gruppi.

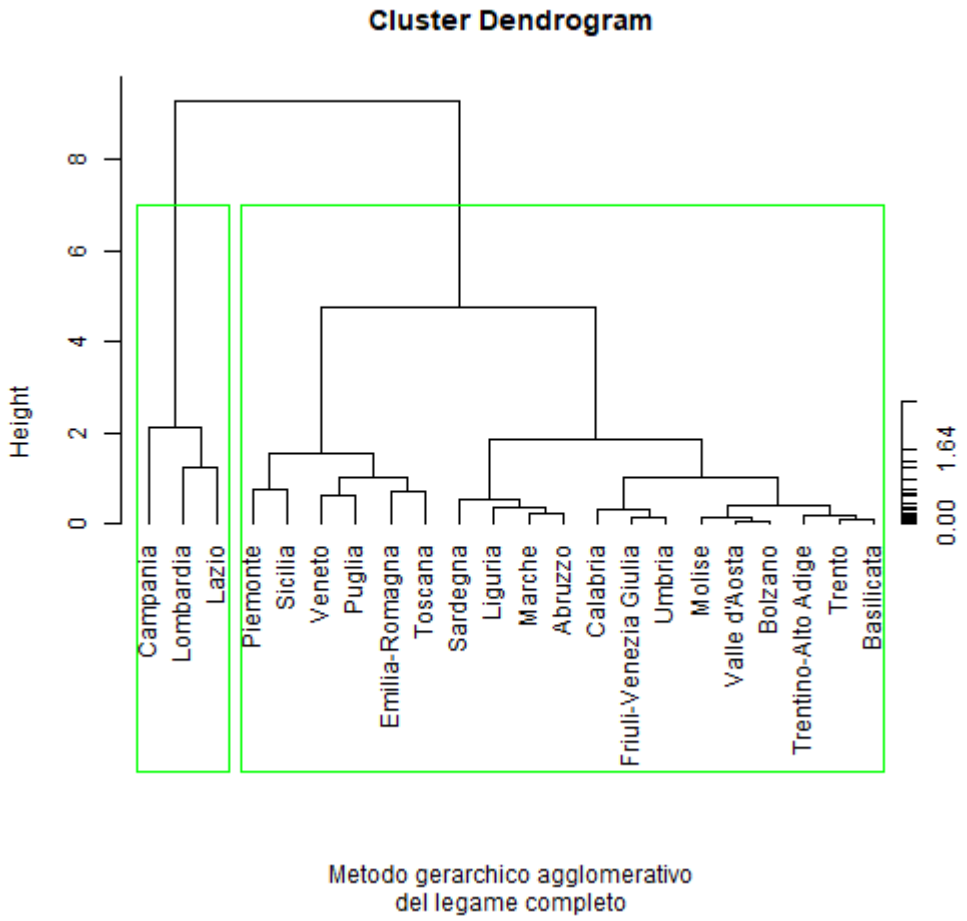
```
hlm<-hclust(d, method="average")
png("grafici/cluster/dendrogrammaUtenti_LegameMedio.png")
plot(hlm, hang=-1, xlab="Metodo gerarchico agglomerativo", sub="del legame
medio")
rect.hclust(hlm, k=2, border="green")
axis(side=4, at=round(c(0, hls$height),2))
dev.off()
```



Metodo del legame completo

La distanza tra due gruppi g1 e g2, con n1 e n2 individui, è definita come la massima tra tutte le distanze di g1 e g2, questo metodo privilegia la differenza tra i gruppi piuttosto che l'omogeneità del gruppo stesso.

```
hlc<-hclust(d, method="complete")
png("grafici/cluster/dendrogrammaUtenti_LegameCompleto.png")
plot(hlc, hang=-1, xlab="Metodo gerarchico agglomerativo", sub="del legame
completo")
rect.hclust(hlc, k=2, border="green")
axis(side=4, at=round(c(0, hls$height),2))
dev.off()
```

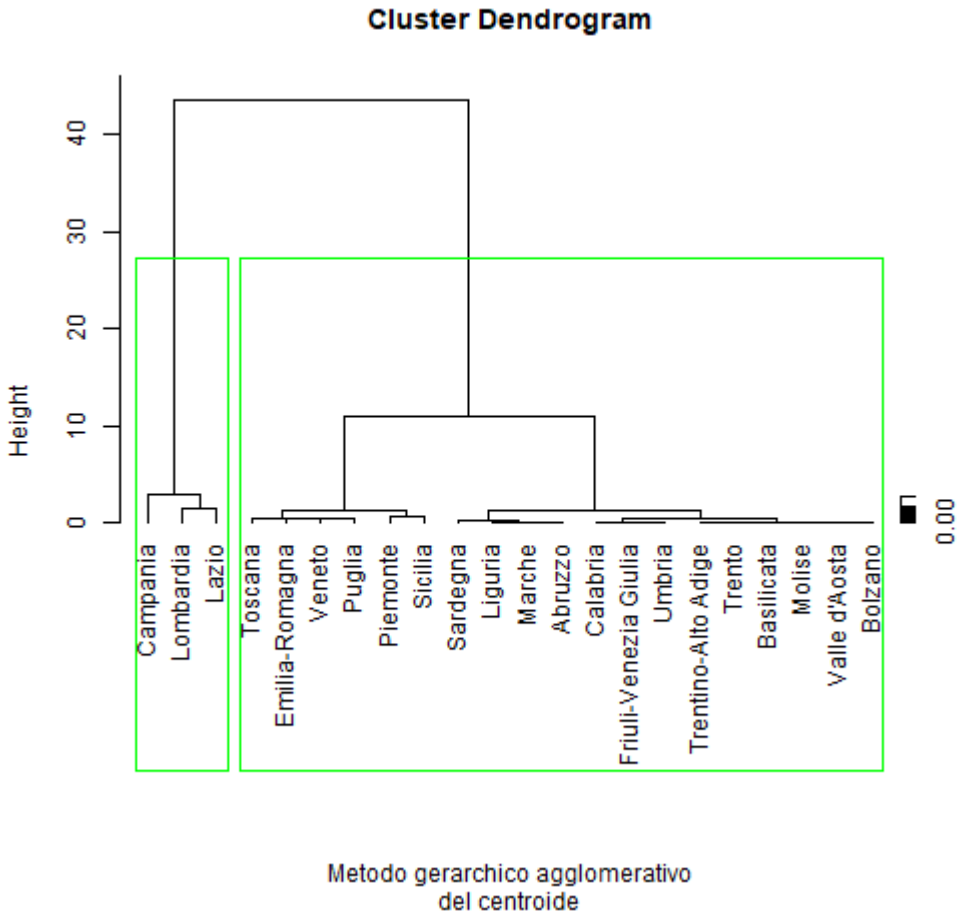


Con il metodo del legame completo i cluster sono sicuramente ben separati ma l'algoritmo privilegia l'omogeneità tra gli elementi interni ai vari gruppi.

Metodo del centroide

La distanza tra i gruppi g1 e g2 è calcolata sulle medie campionarie dei due gruppi. La particolarità di questo metodo è che tende ad avere un effetto gravitazionale: I gruppi più grandi tendono ad assorbire i gruppi più piccoli.

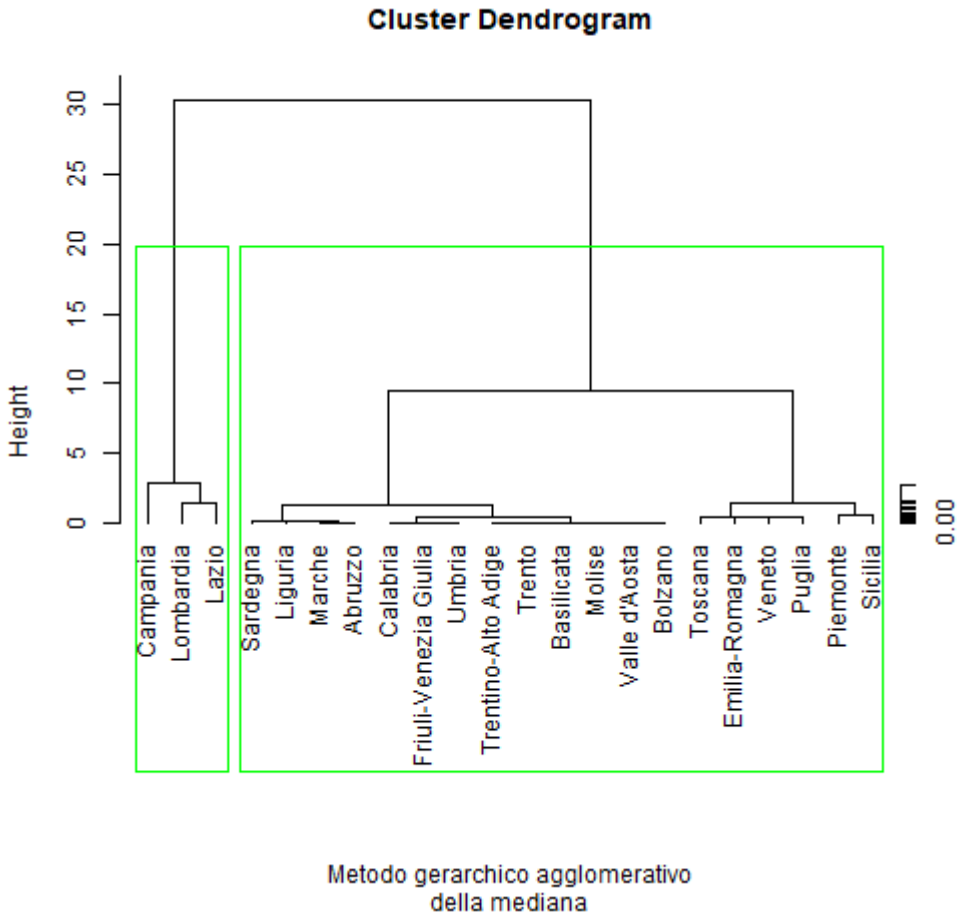
```
d2<-d^2
hc<-hclust(d2, method="centroid")
png("grafici/cluster/dendrogrammaUtenti_MetodoCentroide.png")
plot(hc, hang=-1, xlab="Metodo gerarchico agglomerativo", sub="del centroide")
rect.hclust(hc, k=2, border="green")
axis(side=4, at=round(c(0, hls$height),2))
dev.off()
```



Metodo della mediana

Il metodo della mediana è simile a quello del centroide, ma non è dipendente dalla numerosità del gruppo. Quando due gruppi si uniscono, il nuovo centroide è calcolato come la semisomma dei due gruppi precedenti.

```
hmed<-hclust(d2, method="median")
png("grafici/cluster/dendrogrammaUtenti_MetodoMediana.png")
plot(hmed, hang=-1, xlab="Metodo gerarchico agglomerativo", sub="della
mediana")
rect.hclust(hmed, k=2, border="green")
axis(side=4, at=round(c(0, hls$height),2))
dev.off()
```



Tutti i metodi gerarchici: legame singolo, legame medio, legame completo, metodo del centroide e metodo della mediana hanno fornito il seguente partizionamento in due cluster.

Primo cluster: 19 individui

Secondo cluster: 3 individui

Cluster 1	Piemonte, Valle d'Aosta, Liguria, Trentino-Alto Adige, Trento, Bolzano, Veneto, Friuli-Venezia Giulia, Emilia-Romagna, Toscana, Umbria, Marche, Abruzzo, Molise, Puglia, Basilicata, Calabria, Sicilia, Sardegna
Cluster 2	Lombardia, Lazio, Campania

Per valutare quanto questa suddivisione è "buona" si calcolano le misure di non omogeneità relative all'insieme totale degli individui (trT), ai singoli cluster ottenuti e alla somma delle loro misure di non omogeneità (trS) e alla misura di non omogeneità tra i cluster (trB).

$$trT = trS + trB$$

Poiché per ogni fissata matrice X dei dati si ha che la trT è fissata, i cluster dovrebbero essere individuati in modo da minimizzare la misura di non omogeneità statistica all'interno dei cluster (within) e massimizzare la misura di non omogeneità statistica tra i gruppi (between). Se, fissato il numero di cluster, due metodi conducono a due partizioni differenti occorre scegliere la partizione con la misura di non omogeneità statistica all'interno dei cluster più piccola. Si calcola quindi il rapporto tra la misura di non omogeneità tra i gruppi e la misura di non omogeneità totale. Verrà quindi scelta la suddivisione che massimizza tale rapporto.

Si mostra in R il codice per il calcolo delle misure di non omogeneità per i cluster ottenuti con il metodo del legame singolo. Siccome il partizionamento ottenuto è uguale anche per gli altri metodi i risultati saranno uguali.

```
n<-nrow(Z)
trT<-(n-1)*sum(apply(Z,2,var)) #misura di non omogenità totale
taglio<-cutree(hls, k=2)
num <-table (taglio) #numero di elementi dei gruppi
tagliolist<-list(taglio) #lista di indici per i gruppi
agvar <- aggregate (Z, tagliolist, var)[, -1]
trH1<-(num[[1]]-1)*sum(agvar [1, ]) #misura di non omogenità del primo gruppo
trH2<-(num[[2]]-1)*sum(agvar [2, ]) #misura di non omogenità del secondo gruppo
trB<-trT-trH1-trH2 #misura di non omogenità tra i cluster
rapportoLegameSingolo<-trB/trH
```

Prima di tutto viene calcolata la misura di non omogeneità totale all'interno del dataframe Z utilizzando la seguente istruzione: `trT<-(n-1)*sum(apply(Z,2,var))`. La funzione `apply` permette di applicare la funzione varianza alle colonne del dataframe Z. Per calcolare la misura di non omogeneità i valori delle varianze delle singole colonne vengono sommati e si moltiplica il tutto per il numero di individui nel dataframe (a cui si sottrae 1). Pertanto, la misura di non omogeneità totale risulta:

$$trT = (22 - 1) * 8 = 168$$

Applicando la funzione `cuttree` si ottiene un vettore contenente numeri interi positivi per indicare i cluster a cui sono stati associati gli individui. Successivamente si ricava il numero di elementi associati a ciascun cluster con l'istruzione `num<-table(taglio).` Il primo cluster contiene 19 individui, il secondo ne contiene 3.

```
taglio
 1  2
19  3
```

Si trasforma poi l'array ottenuto tramite `cuttree` in una lista di indici per i vari gruppi. La funzione `agvar<-aggregate(Z, tagliolist, var)` permette di aggregare le colonne del dataframe Z in base alla lista di indici passata che corrisponde quindi ai cluster. A tali gruppi viene applicata la funzione di varianza campionaria, avendo il seguente output.

```
> aggregate (Z, tagliolist , var)
  Group.1       2013      2014      2015      2016      2017      2018       2019      2020
1       1 0.42159888 0.4078818 0.3585118 0.3322855 0.3529968 0.3332359 0.39412156 0.3256009
2       2 0.04992107 0.0949638 0.0172183 0.2498850 0.4594330 0.1186864 0.09575231 0.2300386
```

Aggiungendo [,-1] dopo `aggregate(Z, tagliolist, var)` viene rimossa la prima colonna dall'output.

Per calcolare la misura di non omogeneità all'interno del primo cluster si utilizza l'istruzione `(num[[1]]-1)*sum(agvar [1, ])` che consente di sommare le colonne della prima riga della matrice agvar (ottenendo 2.926233) e successivamente si moltiplica tale valore per il numero di individui nel cluster -1. Quindi:

$$trH1 = (19 - 1) * 2.926233 = 52.6722$$

Per quanto riguarda il secondo cluster invece si ottiene:

$$trH2 = (3 - 1) * 1.315898 = 2.631797$$

Pertanto, la misura di non omogeneità tra i cluster risulta essere:

$$trB = trT - trH1 - trH2 = 168 - 52.6722 - 2.631797 = 112.696$$

Il rapporto risulta $\frac{trB}{trT} = \mathbf{0.6708096}$

La suddivisione ottenuta con i metodi gerarchici risulta essere abbastanza buona in quanto in termini percentuali è del 67%.

4.2 METODI NON GERARCHICI

Tra i metodi non gerarchici, il metodo usato nel progetto è stato "**k-means**", l'algoritmo funziona in diversi step:

1. Si fissa a priori il numero dei cluster e si scelgono m punti di riferimento iniziali che inducono una prima partizione provvisoria;
2. Si considerano tutti gli elementi e si attribuisce ognuno al cluster individuato dal punto di riferimento da cui ha la distanza minore;
3. Si ricalcolano i centroidi dei k gruppi costituendo i nuovi punti di riferimento per i cluster;
4. Si rivalutano le distanze per ogni unità rispetto ai centroidi dei vari cluster. Se un elemento x ha una distanza minore in corrispondenza di un altro centroide rispetto a quello del proprio cluster, si riposiziona l'elemento;
5. Si ricalcolano i centroidi;
6. Si ripete dallo step 4, se si arriva ad un punto in cui non ci sono stati spostamenti tra elementi dei cluster, l'algoritmo si conclude.

Il metodo non gerarchico K-means ha fornito il seguente partizionamento in due cluster.

Primo cluster: 9 individui

Secondo cluster: 13 individui

Cluster 1	Piemonte, Lombardia, Veneto, Emilia-Romagna, Toscana, Lazio, Campania, Puglia, Sicilia
Cluster 2	Valle d'Aosta, Liguria, Trentino-Alto Adige, Trento, Bolzano, Friuli-Venezia Giulia, Umbria, Marche, Abruzzo, Molise, Basilicata, Calabria, Sardegna

```
km <-kmeans (Z, centers=2, iter.max =10, nstart =1)
rapportoKMeans<-km$betweenss/km$totss
```

Il rapporto $\frac{trB}{trT} = \mathbf{0.7129243}$.

La suddivisione in cluster ottenuta con il metodo non gerarchico K-means risulta essere migliore in quanto supera il 70% mentre quella ottenuta con i metodi gerarchici era circa 67%.

4.3 SUDDIVISIONE CON 3 CLUSTER

Ma che cosa succederebbe se si volesse suddividere l'insieme degli individui in 3 cluster anziché 2?

Suddividendo in 3 cluster si è ottenuto, sia con i metodi gerarchici che con quelli non gerarchici, il seguente partizionamento:

Primo cluster: 3 individui

Secondo cluster: 13 individui

Terzo cluster: 6 individui

Cluster 1	Campania, Lombardia, Lazio
Cluster 2	Sardegna, Liguria, Marche, Abruzzo, Calabria, Friuli-Venezia Giulia, Umbria, Molise, Valle d'Aosta, Bolzano, Trentino-Alto Adige, Trento, Basilicata
Cluster 3	Piemonte, Sicilia, Veneto, Puglia, Emilia-Romagna, Toscana

Tabella riassuntiva del rapporto tra misura di non omogeneità tra i cluster e misura di non omogeneità totale $(\frac{trB}{trT})$

Metodo	Rapporto con 2 cluster	Rapporto con 3 cluster
Metodo del legame singolo	0.6708096	0.9375509
Metodo del legame completo	0.6708096	0.9375509
Metodo del legame medio	0.6708096	0.9375509
Medio del centroide	0.6708096	0.9375509
Metodo della mediana	0.6708096	0.9375509
Metodo k-means	0.7129243	0.9375509

Se si volesse suddividere l'insieme degli individui in 2 cluster, la suddivisione ottenuta con il metodo non gerarchico k-means risulta essere migliore. Se invece, si volesse suddividere l'insieme in 3 cluster tutti i metodi portano alla stessa suddivisione e, di conseguenza allo stesso rapporto di 0.9375509. La suddivisione in 3 cluster risulta essere migliore in quanto 0.9375509 > 0.7129243.