DATA ANALYSIS EXAM

Pio Pasquale Trotta

2022-07-06

PRESENTAZIONE DEL DATATASET

Il dataset oggetto di analisi contiene numerose statistiche su ogni calciatore che ha preso parte al massimo campionato inglese di calcio, la **Premier League**, nella stagione 2021-2022. Ad esempio, per ogni calciatore possiamo vedere quale sia la squadra di appartenenza, la nazionalità, il numero di partite giocate o di goal segnati, ma vediamo le variabili più nel dettaglio.

1. Player : Nome del calciatore;
2. Team : Società di appartenenza del calciatore;
3. Nation : Nazinalità;
4. Pos : Posizione;
5. Age : Età del calciatore;
6. MP : Numero di Partite giocate;
7. Starts : Numero di partite da titolare;
8. Min : Numero di minuti giocati;
9. 90s : Minuti giocati divisi per 90;
10. Gls : Numero di Goal seganti;
11. Ast : Numero di assists;
12. G-PK : Goal segnati senza contare i rigori segnati;
13. PK : Numero di Rigori seganti;
14. PKatt : Rigori calciati;
15. CrdY : Numero di Cartellini gialli;
16. CrdR : Numero Cartellini rossi;
17. Gls : Numero di goal ogni 90 minuti;
18. Ast : Numero di assists ogni 90 minuti;
19. G+A : Goal e Assists ogni 90 minuti;
20. G-PK : Goal segnati senza contare i rigori segnati ogni 90 minuti;
21. G+A-PK: Goal segnati + assist ogni 90 min senza i rigori segnati.

Tabella di esempio del dataset

Player	Team	Nation	Pos	Age	MP	Starts	Min	90s
Bukayo Saka	Arsenal	eng<a0>ENG	FW,MF	19	38	36	2978	33.1
Gabriel Dos Santos	Arsenal	br<a0>BRA	DF	23	35	35	3063	34.0
Aaron Ramsdale	Arsenal	eng<a0>ENG	GK	23	34	34	3060	34.0
Ben White	Arsenal	eng<a0>ENG	DF	23	32	32	2880	32.0
Martin <d8>degaard	Arsenal	no<a0>NOR	MF	22	36	32	2785	30.9
Granit Xhaka	Arsenal	ch<a0>SUI	MF,DF	28	27	27	2327	25.9

La tabella continua alla pagina successiva

Gls	Ast	G-PK	PK	PKatt	CrdY	CrdR	Gls90	Ast90	G+A	G-PK90	G+A-PK
11	7	9	2	2	6	0	0.33	0.21	0.54	0.27	0.48
5	0	5	0	0	8	1	0.15	0.00	0.15	0.15	0.15
0	0	0	0	0	1	0	0.00	0.00	0.00	0.00	0.00
0	0	0	0	0	3	0	0.00	0.00	0.00	0.00	0.00
7	4	7	0	0	4	0	0.23	0.13	0.36	0.23	0.36
1	2	1	0	0	10	1	0.04	0.08	0.12	0.04	0.12

OBIETTIVI

Sfruttando tale dataset, andremo ad effettuare un'analisi descrittiva delle variabili, in questo modo ne capiremo le caratteristiche. Successivamente, potremo effettuare dei paragoni tra le diverse dimensioni per individuare somiglianze o marcate differenze. La dimensionalità del dataset verrà poi ridotta attraverso l'Analisi delle componenti principali (PCA). Vi sarà il raggruppamento delle osservazioni omogenee grazie alla Cluster Analysis. Infine cercheremo eventuali relazioni tra il numero di partite giocate e le altre variabili per capire se queste statistiche influenzano il tempo di impiego in campo di un calciatore.

ANALISI DESCRITTIVA

Grazie all'*analisi descrittiva* è possibile l'analisi e la sintesi dei dati osservati all'interno dei nostri campioni allo scopo di comunicare a terzi informazioni rilevanti sui dati.
Prima di calcolare e illustrare le varie informazioni sulle variabili di tipo numerico del dataset, è opportuno rimuovere tutte quelle osservazioni che assumono come valore **NA**. Una volta fatto ciò, è possibile calcolare alcune misure di tendenza centrale e di dispersione (o variabilità), nonché l'indice di asimmetria e l'indice di curtosi.
Per qanto concerne le misure di tendenza centrale e di dispersione (o variabilità), queste sono **media aritemtica, mediana, quartili, varianza, deviazione standard**.Parliamone in breve:

- La *media aritmetica ($\overline{x}$)* rappreenta quel valore rispetto al quale tendono a concentrarsi tutte le altre osservazioni del campione studiato. Per poterla calcolare:
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
- La *mediana* è una misura di posizione in grado di dividere in due parti uguali i livelli del nostro campione, più nello specifico, questa rappresenta quel valore che si trova esattamente al centro della distribuzione, facendo sì che sia rappresentato il 50% delle osservazioni da ambo i lati dello stesso. Dei quartili parleremo successivamente.
- La *varianza (σ^2)* indica la dispersione dei valori di un campione attorno alla media aritmetica, più la varianza è bassa più le osservazioni del campione si attesteranno vicino alla media, e questa potrà essere usata come misura rappresentativa della variabile; al contrario, più la varianza è alta, più tenderanno a discostarsi dalla media aritmetica. Si calcola nel seguente modo:
$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$
- La *deviazione standard (σ)* altro non è che la radice quadrata della varianza, avremo così la varianza nella stessa unità di misura della variabile considerata. Come per la varianza, in corrispondenza di valori alti, avremo una maggiore variabilità rispetto alla media e viceversa.
$$\sigma = \sqrt{\sigma^2}$$

Vediamo i risultati riportati in una tabella:

	Mean	Variance	St.Deviation	Min	First.Qu.	Median	Third.Qu.	Max
Age	2.545238e+01	1.914727e+01	4.3757593	16	22.000	25.00	29.00	39.00
MP	1.920330e+01	1.349586e+02	11.6171681	1	9.000	20.00	29.75	38.00
Starts	1.531136e+01	1.347873e+02	11.6097926	0	4.000	15.00	25.00	38.00
Min	1.375665e+03	1.025599e+06	1012.7186516	1	398.000	1328.00	2154.00	3420.00
90s	1.528388e+01	1.265971e+02	11.2515392	0	4.425	14.75	23.95	38.00
Gls	1.899267e+00	1.060818e+01	3.2570206	0	0.000	1.00	2.00	23.00
Ast	1.364469e+00	4.154992e+00	2.0383798	0	0.000	1.00	2.00	13.00
G-PK	1.745421e+00	8.619474e+00	2.9358941	0	0.000	1.00	2.00	23.00
PK	1.538462e-01	4.570219e-01	0.6760339	0	0.000	0.00	0.00	6.00
PKatt	1.886447e-01	6.193938e-01	0.7870157	0	0.000	0.00	0.00	7.00
CrdY	2.454212e+00	6.578634e+00	2.5648847	0	0.000	2.00	4.00	11.00
CrdR	7.875458e-02	8.002487e-02	0.2828867	0	0.000	0.00	0.00	2.00
Gls90	1.104396e-01	3.630036e-02	0.1905265	0	0.000	0.03	0.15	2.03
Ast90	1.002747e-01	2.470313e-01	0.4970224	0	0.000	0.03	0.12	11.25
G+A	2.106593e-01	2.894374e-01	0.5379938	0	0.000	0.10	0.29	11.25
G-PK90	1.032418e-01	3.383736e-02	0.1839493	0	0.000	0.03	0.14	2.03
G+A-PK	2.034432e-01	2.863155e-01	0.5350845	0	0.000	0.10	0.28	11.25

Informazioni estraibili dalla tabella

Grazie a queste analisi siamo in grado di estrarre alcune informazioni dalla tabella. Per esempio, adesso sappiamo che l'età dei calciatori in Premier League nella stagione 21/22 variava da un minimo di 16 anni ad un massimo di 39, con una media intorno ai 25. Possiamo vedere che la mediana di goal segnati per ogni giocatore è pari ad 1, mentre la media aritmetica è 1.899, la misura più affidabile in questo caso è la mediana, poiché sappiamo che nel corso di un campionato vi sono motli calciatori che non segnano affatto, o che addirittura non giocano, perciò sarebbe errato pensare che sia prossima a 2. Inoltre, per alcune variabili analizzate (*AGE, MP, STARTS, MIN, 90S, GLS,AST, G-PK*), il valore della varianza è maggiore di zero, quindi vorrà dire che, per quelle variabili (tra cui anche il numero di goal segnati), non sarà opportuno identificare la media come misura di posizione rappresentativa. Per le restanti, i valori sono molto piccoli e minori di zero, pertanto, si può considerare la media come rappresentativa.

Indici di asimmetria e di curtosi

In merito all'*indice di asimmetria* e al *indice di curtosi*, dobbiamo dire alcune cose.

L'**indice di asimmetria (γ_1)** ci permette di capire se la distribuzione di un campione attorno ad un valore x_0 è simmetrica o asimmetrica, nel qual caso distinguiamo asimmetria negativa e positiva. Come si calcola l'indice di asimmetria:

$$\gamma_1 = \frac{1}{n}\sum_{i=1}^{n}(\frac{x_i-\overline{x}}{\sigma})^3$$

- $\gamma_1 = 0$: Distribuzione *Simmetrica*, ciò vuol dire che la media aritmetica ($\overline{x}$), la mediana (Me) e la moda ($Mo(X)$) assumono lo stesso valore.

$$\overline{x} = Me = Mo(X)$$

- $\gamma_1 > 0$: *Asimmetria positiva* (o *a sinistra*): la moda assume un valore più piccolo delle mediana, la quale, a sua volta, è più piccola della media aritmetica. Da un punto di vista grafico, il "picco" della curva (corrispondente al valore della Moda) sarà spostato più a sinistra.

$$Mo(X) \leq Me \leq \overline{x}$$

- $\gamma_1 < 0$: *Asimmetria negativa* (o *a destra*): la moda assume un valore più grande delle mediana, la quale, questa volta, è maggiore della media aritmetica. Graficamente, il "picco" della curva sarà spostato a destra.

$$\overline{x} \leq Me \leq Mo(X)$$

Invece, l'**indice di curtosi** (γ_2) riguarda il grado di "appiattimento" della curva considerata, ovvero quanto velocemente le code tendono a zero.

$$\gamma_2 = |\frac{1}{n}\sum_{i=1}^{n}(\frac{x_i-\overline{x}}{\sigma})^4| - 3$$

- $\gamma_2 = 0$: *comportamento normale*, parleremo perciò di *curva Gaussiana* o *mesocurtica*, la curva è "piatta" come una normale.
- $\gamma_2 > 0$: *curva leptocurtica*, le code tendono a zero molto velocemente, maggiormente rispetto ad una normale.
- $\gamma_2 < 0$: *curva platicurtica*, le code tendono a zero lentamente, meno velocemente rispetto ad una normale.

	Simmetria	Curtosi
Age	0.1632672	2.498851
MP	-0.1384563	1.716491
Starts	0.2758292	1.839415
Min	0.2719936	1.880807
90s	0.2714974	1.880398
Gls	2.8528395	13.259539
Ast	2.2535623	9.198348
G-PK	2.8413945	13.648601
PK	5.4749584	35.762518
PKatt	5.4639525	37.113869
CrdY	1.1405915	3.726768
CrdR	3.6043808	15.890808
Gls90	3.9783127	29.689257
Ast90	20.8193369	465.747163
G+A	16.0944447	326.433335
G-PK90	4.3384555	34.263308
G+A-PK	16.3836037	334.452360

Osservando i valori presenti nella tabella, possiamo notare che nessuna distribuzione è perfettamente simmetrica oppure mesocurtica, infatti vediamo come tute le distribuzioni, con intensità differente, siano leptocurtiche ($curtosi > 0$, perciò le code delle curve che le rappresentano tendono a zero molto velocemente). Inoltre, solo un campione (MP) ha una distribuzione **asimmetrica negativa**, gli altri presentano tutti un'**asimmetria positiva**.

STRUMENTI GRAFICI PER L'ANALISI DESCRITTIVA

Boxplot - Spiegazione teorica

Il **Boxplot** è un tipo di grafico utilizzato per rappresentare distribuzioni di campioni attraverso indici di dispersione e di posizione $(X_{min}, Q_1, Me, Q_3, X_{Max})$. Detto anche "*Diagramma a scatola e baffi*", il boxplot è costituito da una parte centrale detta appunto "*scatola*", questa scatola è delimitata dal **Primo Qaurtile** (Q_1, lato sinistro della scatola) e dal **Terzo Quartile** (Q_3, lato destro).

- Q_1 : Il primo quartile ci fa comprendere che al di sotto di esso sono presenti il 25% dei valori osservati;
- Q_2 o Me : Il secondo quartile è rappresentato dalla mediana. Come già detto in precedenza, al di sotto e al di sopra di essa sono presenti il 50% delle osservazioni;
- Q_3 : Al di sotto del terzo quartile abbiamo il 75% dei valori.

Ovviamente, al di sopra di Q_3 avremo il restante 25% del campione. Quindi, ora sappiamo praticamente che all'interno della scatola è presente il 50% delle osservazioni. La scatola viene detta *Range Interquartile* (IQR), ed è la differenza tra il terzo e il primo quartile.

$$IQR = Q_3 - Q_1$$

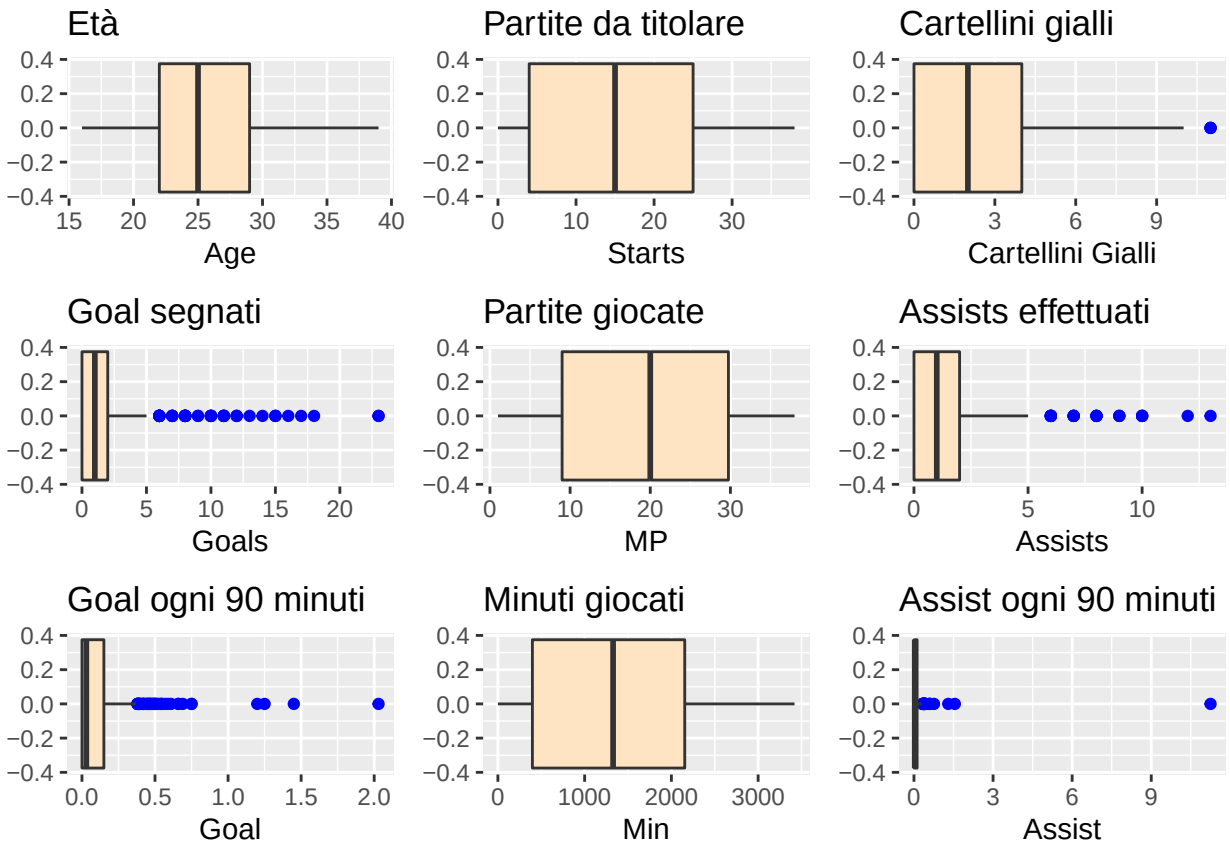
La linea all'interno della scatola rappresenta la Mediana Me. I segmenti che si diramano dai lati della scatola sono detti "**Baffi**", questi indicano la dispersione dei dati inferiori al primo quartile e maggiori al terzo quartile. Come si calcolano i baffi (h, H):

$$h = Q_1 - 1.5(IQR) = Q_1 - 1.5(Q_3 - Q_1)$$
$$H = Q_3 + 1.5(IQR) = Q_3 + 1.5(Q_3 - Q_1)$$

Se non vi sono valori anomali (**outliers**), allora X_{min} e X_{Max} saranno in corrispondenza di (h, H). Se, invece, sono presenti outliers:

$$X_{min} < h \text{ oppure } X_{Max} > H$$

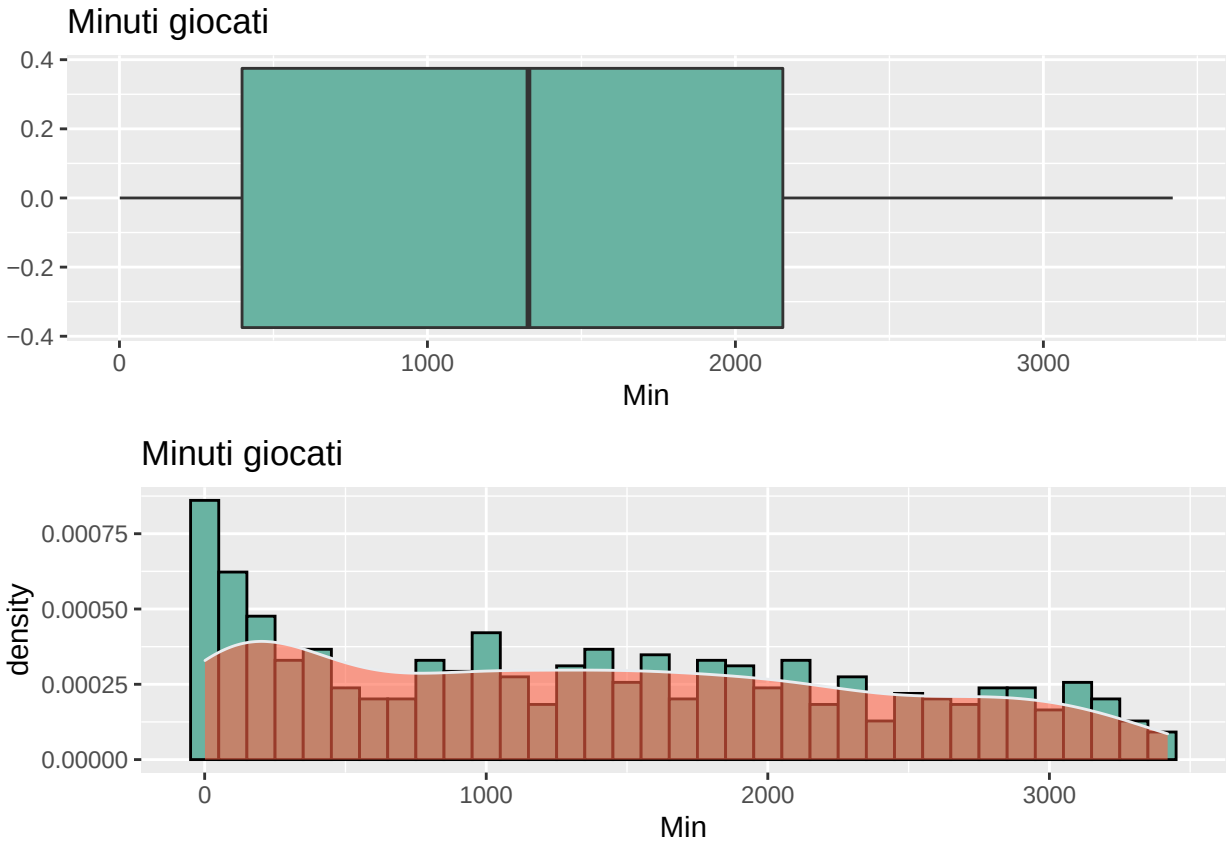
Boxplot - visualizzazione di alcune distribuzioni



Nella griglia sopra riportata sono presenti i boxplot di alcune distribuzioni del nostro dataset. Come possiamo vedere, alcuni di essi presentano dei punti blu, questi rappresentano gli outliers all'interno di quella specifica variabile. In questo caso, grazie a tali outliers, comprendiamo che vi sono alcuni pochi calciatori che hanno segnato o fornito degli assist in numero di gran lunga più elevato rispetto al resto della distribuzione. Vediamo come vi siano pochi calciatori ammoniti più degli altri. Possiamo notare che alcune distribuzioni (*Min, MP, Age, Starts*) non presentano outliers. Mediante i boxplot è possibile capire se una distribuzione è simmetrica o meno. Dai boxplot sopra riportati vediamo come siano tutte nettamente asimmetriche a sinistra, tranne *MP*, la quale è leggermente asimmetrica a destra (è possibile affermarlo osservando anche l'indice di asimmetria calcolato in precedenza).

Istogramma

L'*Istogramma* è un grafico utile per studiare la distribuzione di variabili numeriche (ma anche di variabili discrete se presenti molte osservazioi) e viene sfruttato per rappresentare le disribuizioni per classi. Sull'asse delle ascisse abbiamo i livelli della variabile considerata suddivisi in "classi", ogni classe può avere ampiezza diversa, ma nel nostro caso sono equiampie, quindi le barre del nostro istogramma avranno tutte la stessa ampiezza. Le barre sono tra loro adiacenti e la loro altezza è proporzionale alla *densità di frequenza* (rapporto tra la frequenza, assoluta o relativa, e l'ampiezza della classe), la quale ci dice come si distribuiscono le frequenze all'interno della classe.



Dal nostro istogramma relativo ai minuti giocati da tutti i calciatori, notiamo come la distribuzione di tale variabile sia leggermente asimmetrica a sinistra (osservando anche il boxplot per effetuare un confronto), quindi avremo una moda con un valore più piccolo rispetto alla mediana della distribuzione e notiamo anche la presenza di *multimodalità*.

Correlazione

Prima di parlare della correlazione, è bene dire cosa sia la **covarianza** ($Cov(X,Y)$ o s_{XY}). Questa rappresenta un indice di variabilità congiunta e si calcola nel seguente modo:

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

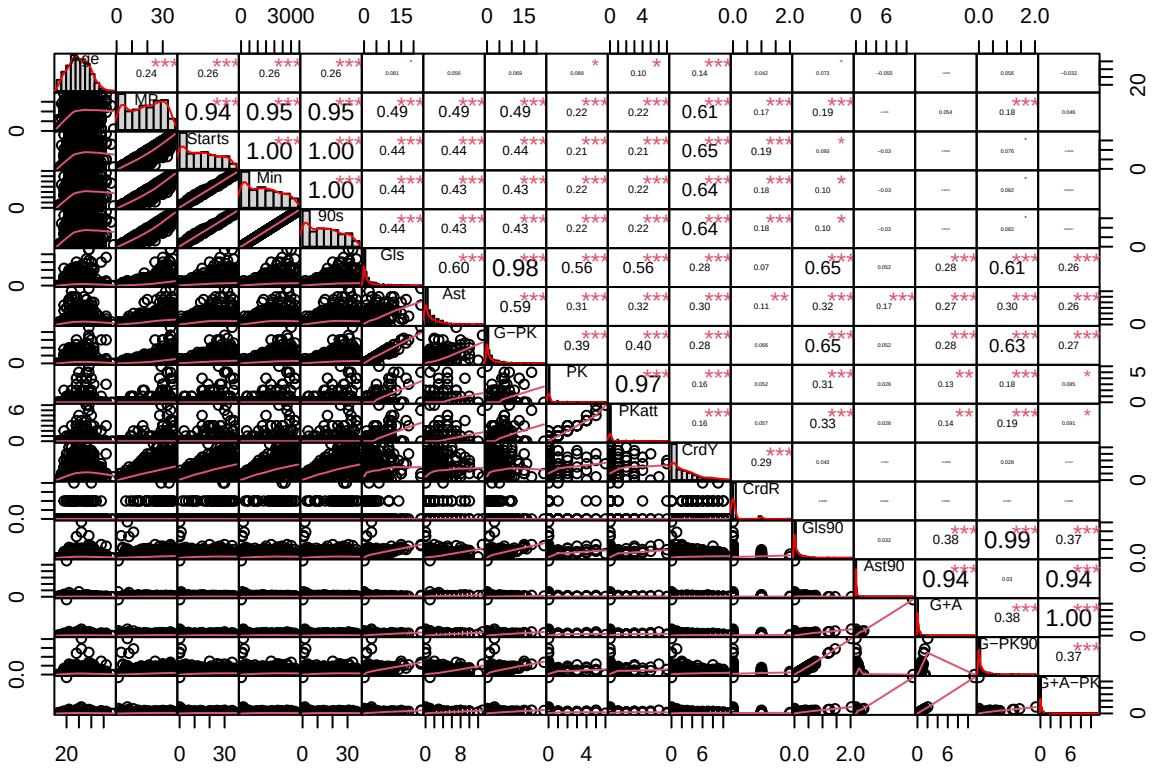
- $s_{XY} = 0$: assenza di legame lineare;
- $s_{XY} > 0$: dipendenza lineare positiva;
- $s_{XY} < 0$: dipendenza lineare negativa.

Vi è un problema, l'indice di covarianza, però, non ci dice quale sia l'intesità del legame lineare, perciò ci serviremo della correlazione.

La **correlazione** (r_{XY}) è un coefficiente in grado di rappresentare l'intesità del legame di dipendenza lineare tra due variabili, se presente. Il coefficiente di correlazione può assumere valori compresi tra $[-1, 1]$. Rappresenta il rapporto tra la covarianza (s_{XY}) e il prodotto delle deviazioni standard di X e Y.

$$r_{XY} = \frac{s_{XY}}{\sqrt{\sigma_X^2 \sigma_y^2}}$$

- $r_{XY} = 0$: Assenza di legame lineare;
- $0 < r_{XY} \leq 1$: Legame lineare positivo, ciò significa che ad incrementi in media della X avremo incrementi proporzionali della Y;
- $-1 \leq r_{XY} < 0$: Legame lineare negativo, ciò significa che ad incrementi (o decrementi) in media della X avremo decrementi (o incrementi) proporzionali della Y.



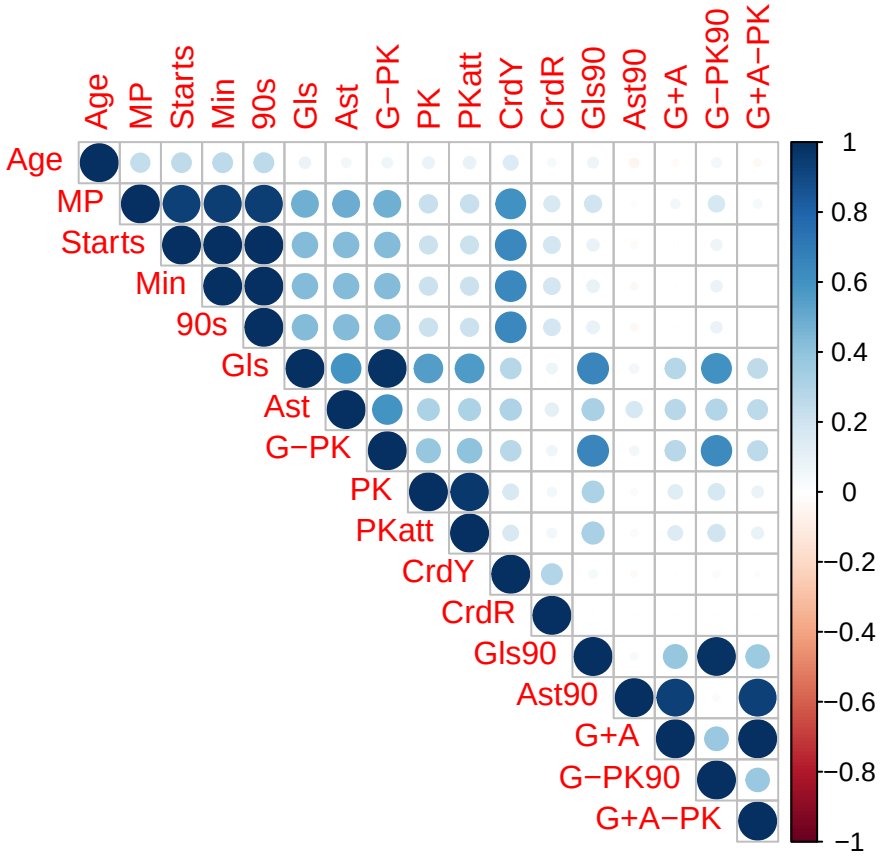
Il *correlation chart* qui riportato rappresenta le correlazioni tra ogni variabile numerica considerata del dataset. Lungo la diagonale principale sono presenti i nomi delle variabili considerate con i relativi istogrammi rappersentativi e la linea di tendenza; nella parte inferiore abbiamo gli scatter plot tra le variabili. Nella parte superiore, invece, sono rappresentati i valori dei coefficienti di correlazione, sulla base della grandezza del valore, il numero sarà graficamente proporzionto. Vicino a r_{XY} sono presenti degli asterischi, i quali indicano il **p-value**. Questo riguarda il *Test del coefficiente di correlazione*, dove l'ipotesi nulla è $H_0 = 0$, e lipotesi alternativa è $H_1 \neq 0$.

- p-values tra 0 e 0.001: ***
- p-values tra 0.001 e 0.05: **
- p-values tra 0.05 e 0.01: *
- p-values tra 0.01 e 0.1: .
- p-values tra 0.1 e 1: ' '

Ricordiamo che se il p-value è minore del livello di significatività (α), rifiuteremo l'ipotesi nulla.

Correlogramma

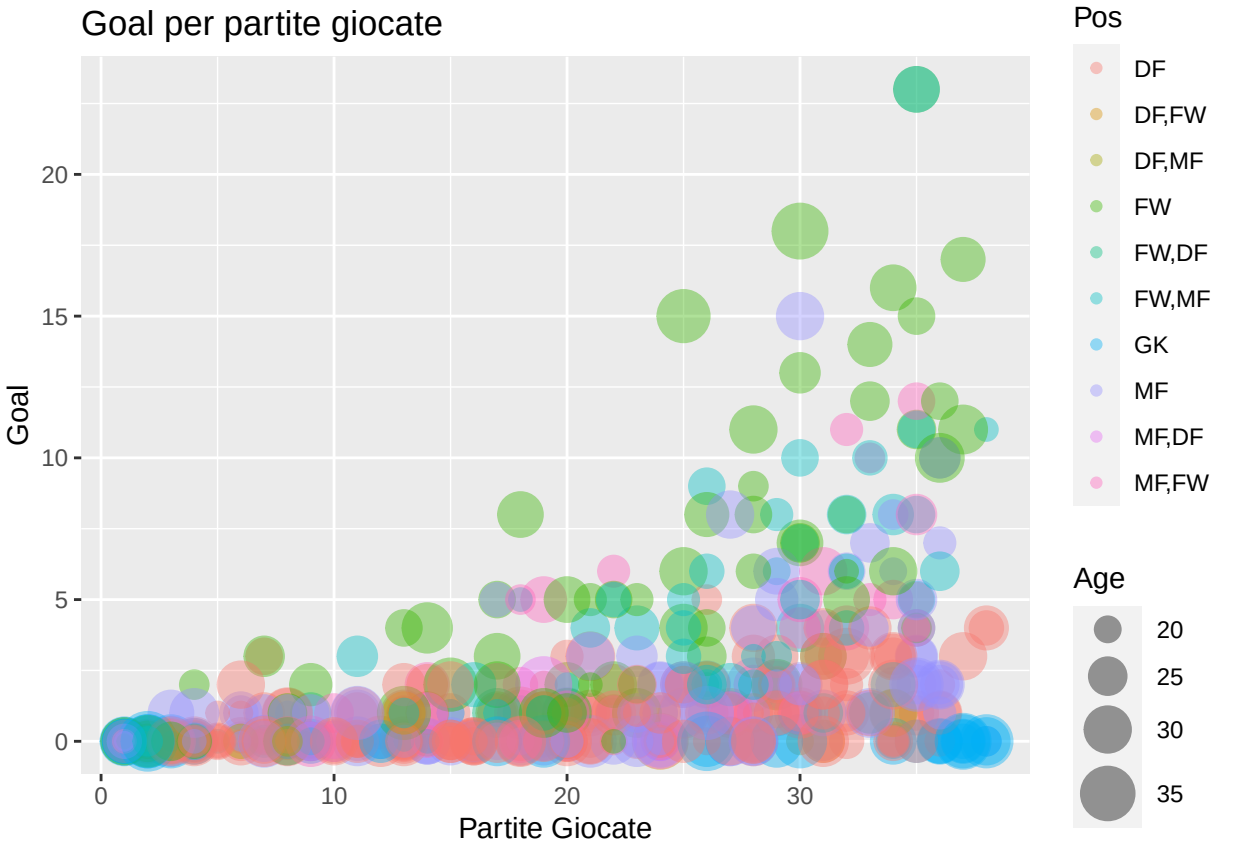
Il **Correlogramma** ci permette di capire il grado di correlazione tra due variabili dal grado del colore che visualizziamo, grado che è proporzionale al coefficiente di correlazione r_{XY}. Colore che sfuma dal blu più scuro (correlazione positiva), passando per il bianco (assenza di correlazione), fino a giungere al rosso (dipendenza lineare negativa).



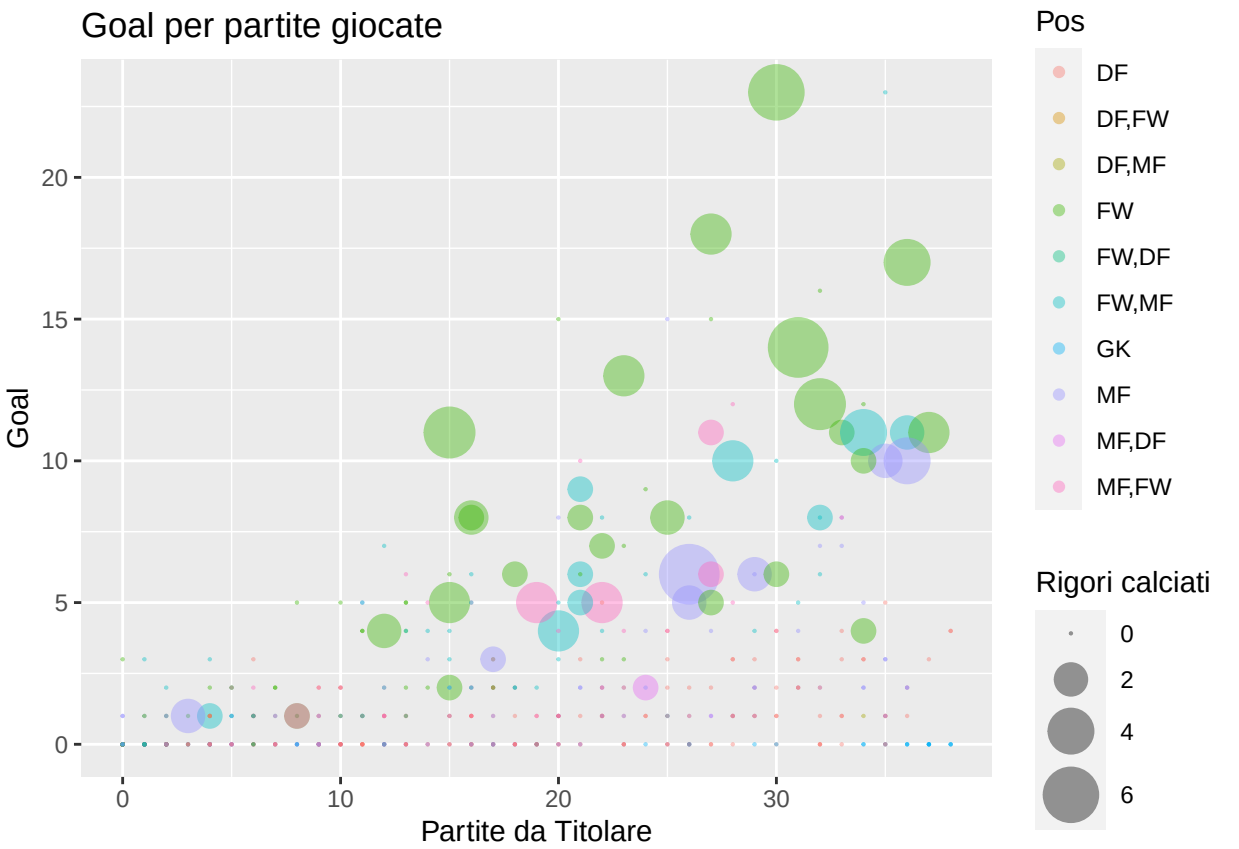
Dal correlogramma notiamo come vi sia solo correlazione positiva o assenza di correlazione tra le variabili, non è presente alcun tipo di correlazione negativa. Per esempio, vediamo come la correlazione positiva sia molto alta tra *MP* (partite giocate) e *Starts* (partite da titolare), meno alta, ma comunque presente, tra *CdrY* (cartellini gialli) e *CdrR* (cartellini rossi), e bassissima se non assente tra *PKatt* (rigori calciati) e *G+A* (rateo goal e assist ogni 90 minuti).

Bubble Plot

Il **Bubble Plot** è un tipo di *scatter plot* in grado di mettere in relazione più variabili, almeno tre, contemporaneamente e, quindi, di fornirci più informazioni. Le informazioni vengono estratte in base al tipo di relazione (se presente, osservarbile data la dispersione dei punti sul piano), nonché dalla forma e/o dal colore dei punti, i quali dipendono dalle variabili considerate.



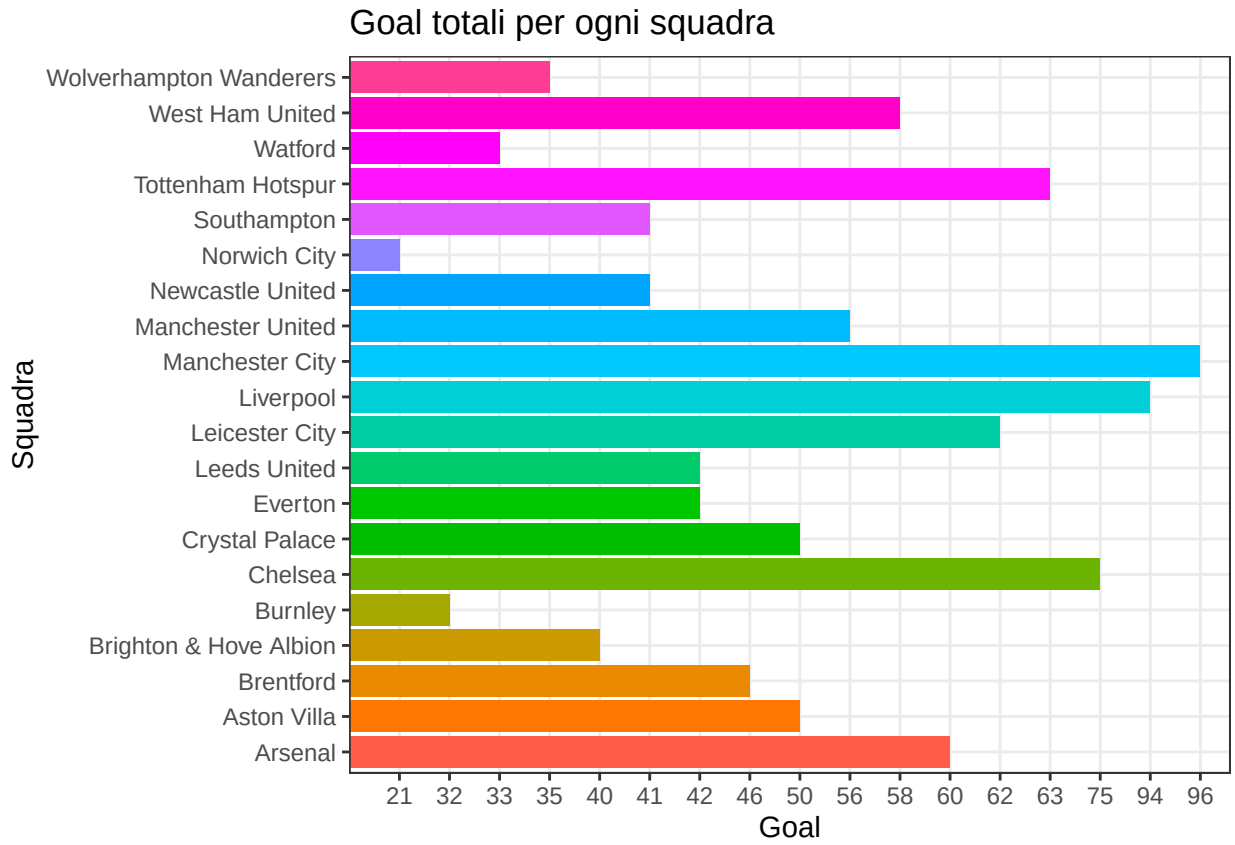
Il bubble plot appena visto rappresenta il numero di goal segnati in base al numero di partite giocate, evidenziando la posizione del calciatore (colore) e l'età (area bolla). Dalla dispersione dei vari punti è possibile capire che vi è dipendenza lineare positiva tra il numero di partite giocate e i goal segnati. Inoltre, grazie all'area della "bolla", si evince che l'età media dei giocatori è di 25 anni. Infine, vediamo come a segnare di più siano i calciatori che ricoprono ruoli offensivi (FW = *Forward*, attaccante, oppure FW = *Forward Wing*, ala offensiva), mentre i difensori o centrocampisti difensivi (*DF*, *DF,MF*) o i portieri (*GK*= Goal keeper) segnano pochissimo.



Questo secondo *Bubble plot* evidenzia come, linearmente, all'aumentare delle numero di partite da titolare, aumentino anche i goal segnati, Inoltre il numero di goal segnati è anche legato, in questo caso, al numero di rigori calciati (*PKatt*;il numero di rigori segnati *PK* non è considerato), poiché, come possiamo vedere, i punti hanno un'area maggiore in corrispondenza di un numero di goal maggiore. Infine, possiamo affermare che gran parte dei rigori sono calciati da giocatori che ricoprono ruoli offensivi (area più ampia e colore della *bolla* fanno sì che ciò si evinca).

Barplot - Diagramma a barre

Il *Barplot* (in italiano *diagramma a barre*) è un grafico in grado di illusttrare la frequenza assoluta dei livelli all'interno di un campione. Sull'asse delle ascisse poniamo le osservazioni del nostro campione, mentre sulle ordinate verrà posta la frequenza di ciascun livello. Le barre in corrispondenza di ogni osservazione avranno un'altezza proporzionale alla frequenza, mentre la base è sempre equiampia e non adiacente alla precedente/successiva.



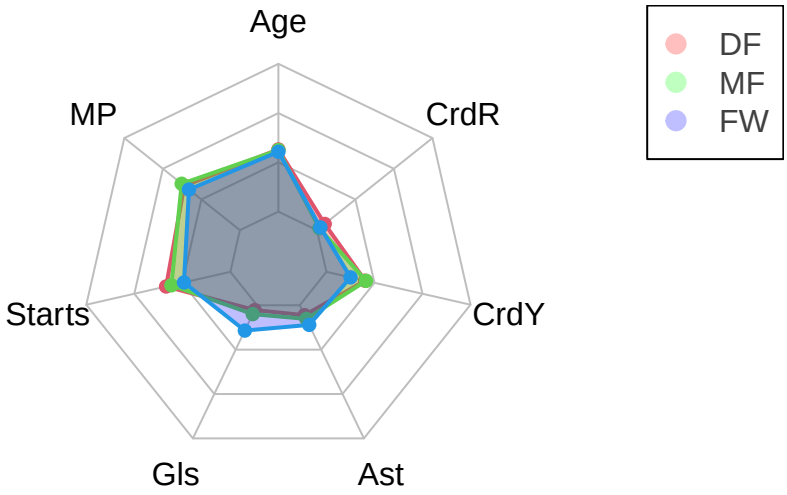
Per costruire il diagramma di cui sopra (in questo caso è rovesciato per una migliore visualizzazione), è stato necessario effettuare il **subsetting** (creazione di sottoinsiemi) dei calciatori rispetto alla propria squadra di appartenenza. Quindi, una volta ottenuto un sottoinsieme dei calciatori per ogni *Team*, sono stati considerati i goal di ognuno di essi e poi sommati, in modo tale da proddure il numero di goal segnati da ogni squadra nel corso del campionato. Vediamo come le squadre ad aver seganto di più siano il Manchester City e il Liverpool, seguite dal Chelsea; le squadre ad aver segnato meno goal sono il Burnley e il Norwich City (*nel grafico non è stato ritenuto necessario ordinare in modo crescente/decrescente le barre poiché i valori sono indicati in maniera precisa per ognuna*).

Esempio di subsetting effettuato

```
mancity_tot <- sum(subset(variabili_utilizzabili, select = Gls,
                          subset = Team == "Manchester City"))
```

Radarchart

Il *radarchart* è uno strumento grafico in grado di mostrare contemporaneamente più variabili e permetterne il confronto.



Il radarchart in figura mette a confronto quei calciatori che ricoprono soltanto un ruolo (quindi $DF=$ Difensori, $MF=$ centrocampisti, $FW=$attaccanti) e non più ruoli (FW,DF; DF,MF; etc... Questa scelta è stata fatta per avere una rappresentazione grafica non confusionale). Da tale grafico possiamo confrontare alcune statistiche dei giocatori sulla base della loro posizione. Vediamo come non ci siano enormi differenze, ma possiamo asserire con certezza che gli attaccanti (in blu) in media segnano di più dei giocatori negli altri ruoli e effettuano più assist. I centrocampisti sono quelli che ricevono più cartellini gialli e giocano più partite in assoluto, mentre i difensori sono coloro che ricevono più cartellini rossi in media e giocano più partite partendo da titolari.

ANALISI ESPLORATIVA

L'**Analisi Esplorativa** mira, mediante diverse tecninche, ad esplicitare le correlazioni tra i dati considerati. Abbiamo: l'*Analisi delle componenti principali* e l'*Analisi dei cluster*.

ANALISI DELLE COMPONENTI PRINCIPALI (PCA)

L'**Analisi delle componenti principali** (Principal Component Analysis) è una procedura matematica che punta a ridurre la dimensionalità di un dataset, semplificando quindi i dati di origine. Ciò che viene fatto è partire dal nostro numero di variabili per ottenere un numero di nuove variabili inferiore, ossia le **Componenti Principali** (PC). Queste manterranno gran parte della variabilità originaria.

Le nuove dimensioni che otteremo hanno diverse proprietà:

- Le PC non sono tra loro correlate;
- Hanno sempre media uguale a 0 ($E(Y_1) = 0, ..., E(Y_p) = 0$);
- Ogni autovalore (λ) corrisponde alla varianza della componente principale;

$$Var(Y_k) = \lambda_k$$

- Le componenti principali sono ordinate in ordine decrescente di varianza (autovalore)(quindi la prima PC avrà la varianza più alta);
- Le PC hanno correlazioni uguali a 0;
- Sommando tutti gli autovalori otteremo la varianza totale;

Come scegliere il numero di PC a cui ridurre il dataset

Per scegliere il numero di componenti principali esistono 3 diversi criteri:

1. Scegliere un numero di PC che mantenga l'80% della variabilità totale;
2. Scree Plot;
3. *Regola di Kaiser*: scegliere le PC che hanno autovalori maggiori di 1; .

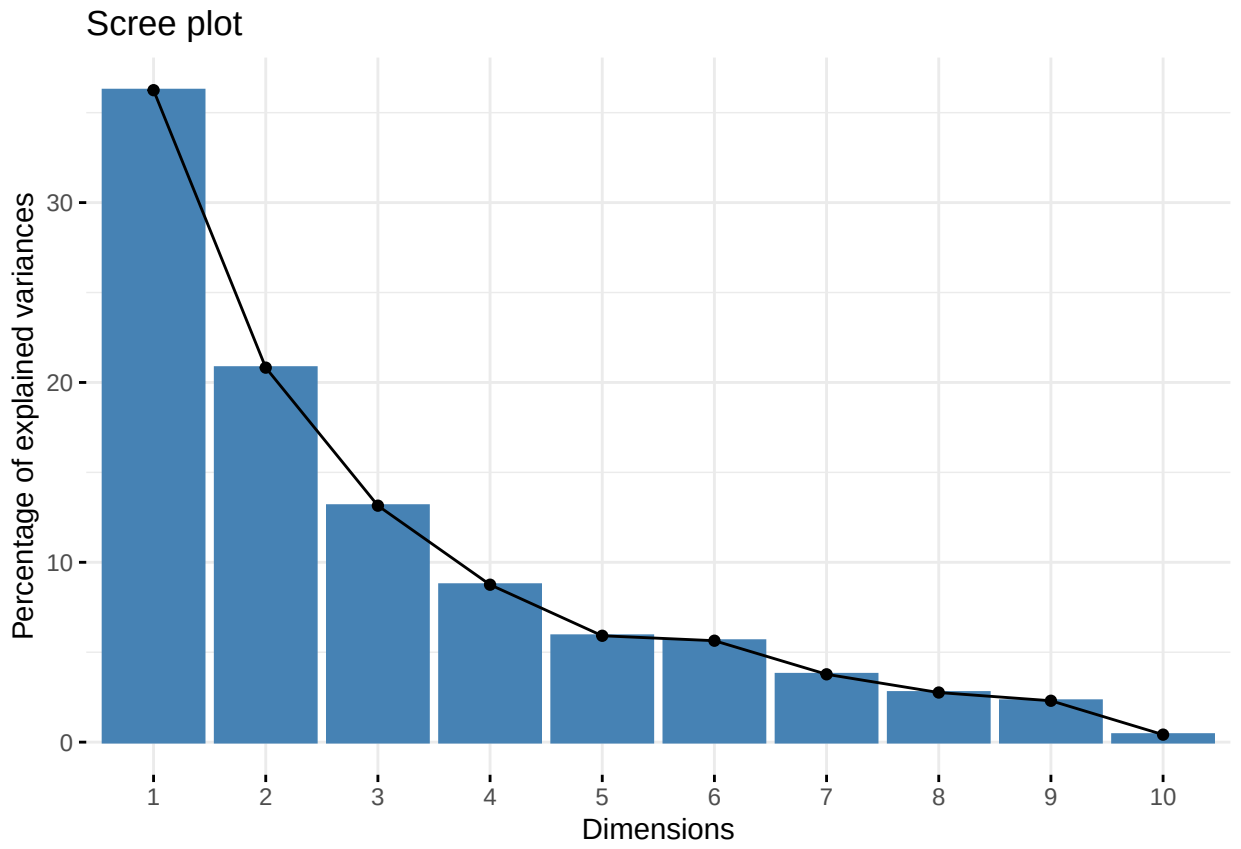
1. Percentuale di varianza spiegata

	eigenvalue	percentage.of.variance	cumulative.percentage.of.variance
comp 1	6.162503e+00	3.625002e+01	36.25002
comp 2	3.539863e+00	2.082272e+01	57.07274
comp 3	2.235127e+00	1.314780e+01	70.22054
comp 4	1.488130e+00	8.753707e+00	78.97425
comp 5	1.005544e+00	5.914962e+00	84.88921
comp 6	9.588751e-01	5.640442e+00	90.52965
comp 7	6.412864e-01	3.772273e+00	94.30193
comp 8	4.689572e-01	2.758572e+00	97.06050
comp 9	3.912510e-01	2.301477e+00	99.36198
comp 10	7.098248e-02	4.175440e-01	99.77952
comp 11	3.048220e-02	1.793071e-01	99.95883
comp 12	3.736730e-03	2.198076e-02	99.98081
comp 13	3.234222e-03	1.902484e-02	99.99983
comp 14	2.262465e-05	1.330862e-04	99.99996

comp 15	3.150949e-06	1.853500e-05	99.99998
comp 16	2.812825e-06	1.654603e-05	100.00000
comp 17	8.807106e-31	5.180651e-30	100.00000

Osservando la tabella, in cui sono riportati per ogni componente principale: gli autovalori, la percentuale di varianza spiegata e la percentuale cumulativa di varianza spiegata, secondo il *primo criterio*, sceglieremo le prime 4 PC, in quanto in corrispondenza della componente numero 4, avremo una varianza cumulativa pari all'78.97% (non è necessario ottenere un valore pari o superiore all'80% della variabilità se il numero di variabili di partenza è abbastanza elevato, in questo caso sono 17).

2. Scree plot



Anche osservando lo *scree plot*, siamo sicuri di scegliere solo le prime 4 componenti principali. Lo scree plot mette in relazione le nuove dimensioni con la percentuale (non cumulativa) di varianza spiegata e poiché questa decresce ad ogni PC successiva (come possiamo vedere anche nella tabella precedente), il grafico sarà una curva spezzata con pendenza negativa.

Per determinare quante componenti principali scegliere con tale strumento, dobbiamo osservare da quale componente in poi la curva tende ad appiattirsi, dobbiamo scegliere perciò le componenti principali prima di questo punto. Nel nostro caso, la curva tende ad appiattirsi a partire dalla componente numero 5, perciò sceglieremo solo 4 componenti principali.

3. Regola di Kaiser

Secondo la regola di Kaiser è necessario scegliere autovalori maggiori di 1. Per tale motivo, andando a guardare la tabella vista precedentemente, sceglieremo un numero di componenti pricipali pari a 4 ($\lambda_4 = 1.48130$), quindi maggiore di 1 ($\lambda_5 = 1.005544$).

CONTRIBUTO DELLE VARIABILI DEL DATASET ALLE COMPONENTI PRINCIPALI

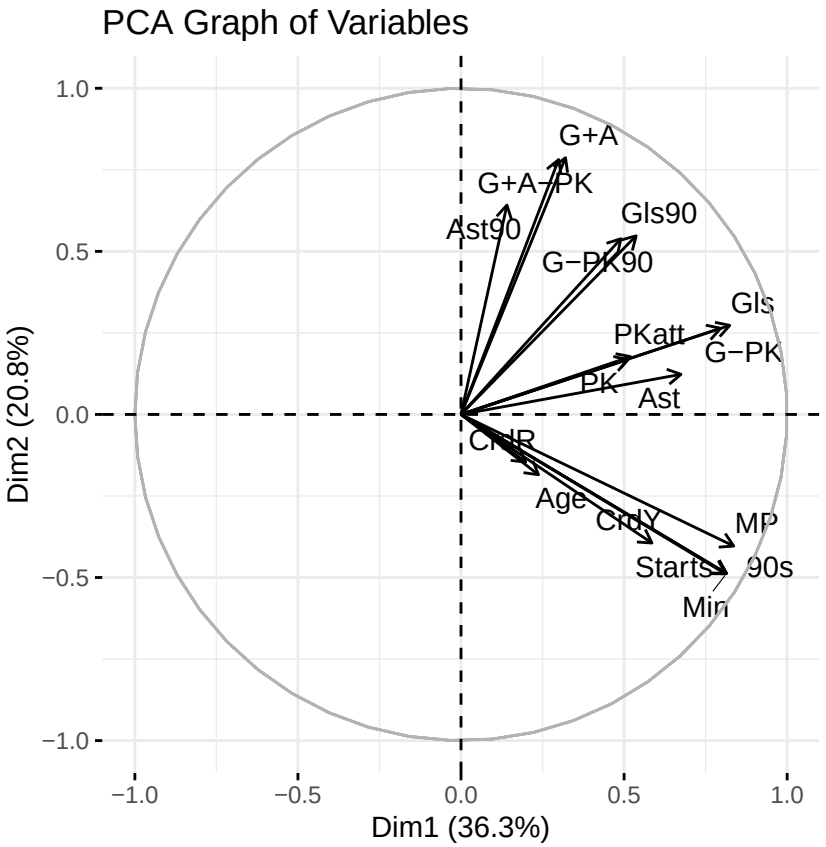
È possibile vedere quale sia il contributo dato dalla variabile alla nuova componente pricipale tramite la correlazione. Quest'ultima viene usata come coordinata della variabile nell'analisi delle PC. La rappresentazione delle variabili è differente dal grafico delle osservazioni, poiché:

- le osservazioni vengono rappresentate dalle loro proiezioni;
- le variabili vengono rappresentate dalle loro correlazioni.

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Age	0.9157464	0.9657142	0.022772153	0.004793133	3.122731e+01
MP	11.3475372	4.5981757	1.915834638	0.572373782	5.661181e-01
Starts	10.6836016	6.7530536	2.742126089	0.084403753	5.292492e-01
Min	10.7537939	6.7113739	2.701400487	0.088385505	6.592860e-01
90s	10.7551197	6.7104805	2.700883458	0.088293924	6.611364e-01
Gls	11.0184225	2.1070646	5.520291817	0.096897800	2.174951e-01
Ast	7.3801592	0.4269703	0.003758206	0.028573811	1.382825e+00
G-PK	10.2865562	1.9724954	3.737667452	2.743858539	3.597963e-01
PK	4.2590526	0.7994908	8.547756085	32.421537957	1.282296e-01
PKatt	4.3788315	0.8873120	8.773339467	31.405998088	1.333790e-01
CrdY	5.5620558	4.3911542	2.105457016	0.047051808	5.384836e+00
CrdR	0.6383439	0.6026120	0.639105651	0.224674045	5.842055e+01
Gls90	4.6785890	8.4729426	7.328545091	10.401022052	9.310401e-03
Ast90	0.3214658	11.6367180	22.747604604	3.659259020	1.127110e-01
G+A	1.6656022	17.5068782	11.868366727	0.394644200	1.176362e-01
G-PK90	3.8998949	8.1997316	5.554612949	17.626174765	6.192578e-04
G+A-PK	1.4552277	17.2578323	13.090478109	0.112057821	8.950590e-02

Relazioni tra variabili - Correlation Circle

La relazione tra variabili può essere mostrata grazie ad un *Correlation circle plot*, in cui le correlazioni tra le variabili di partenza e le PC sono rappresentate da coordinate. La lunghezza della linea e la prossimità alla circoferenza nel grafico ci dicono quanto bene è rappresentata la variabilità.



Intrepretazione del grafico

- Se l'ampiezza degli angoli tra le variabili è piccola vi è **correlazione positiva**;
- Se l'ampiezza degli angoli tra le variabili è prossima a 180°, **correlazione negativa**;
- Qualora dovessero formarsi angoli di 90°, allora vi è **assenza di correlazione**.

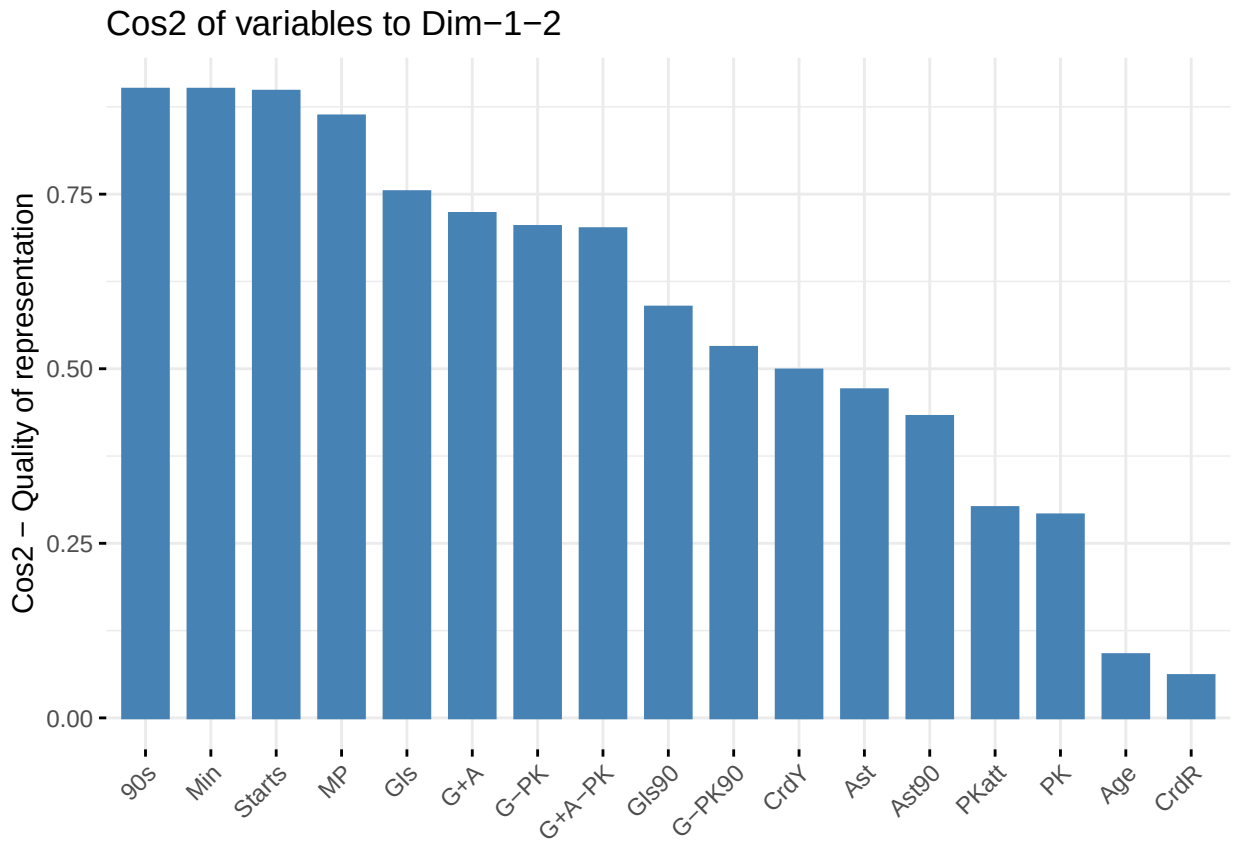
Dal grafico e sulla base di quanto appena detto, possiamo vedere come le prime due PC spieghino il 57.1% della variabilità totale (percentuali riportate ai lati del grafico). Andando ad analizzare gli angoli che si sono formati tra le diverse variabili vediamo come alcune siano positivamente correlate fra loro (ampiezza angolo bassa) ad esempio *Starts* e *Min* oppure *Ast90* e *G+A*. Altre invece mostrano assenza di correlazione formando angoli di quasi novanta gradi come *CrdY* e *Gls90* oppure *CrdR* e *G+A-PK*.

Qualità della rappresentazione delle variabili

È possibile illustrare quanto bene siano spiegate le variabili di partenza dalle nuove dimensioni mediante il cosiddetto cos^2 (*cosen quadrato*).

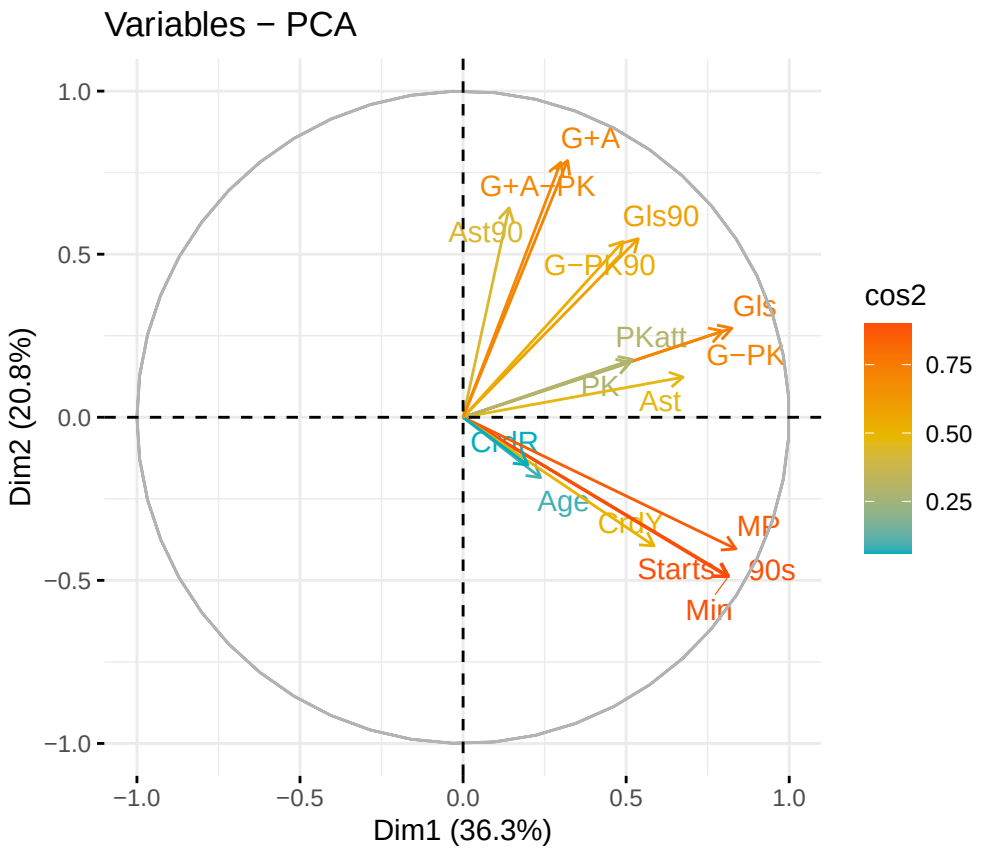
	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Age	0.05643290	0.03418496	5.089865e-04	7.132805e-05	3.140042e-01
MP	0.69929232	0.16276910	4.282133e-02	8.517667e-03	5.692564e-03
Starts	0.65837727	0.23904881	6.128999e-02	1.256038e-03	5.321831e-03
Min	0.66270287	0.23757341	6.037972e-02	1.315291e-03	6.629408e-03
90s	0.66278457	0.23754179	6.036817e-02	1.313929e-03	6.648014e-03
Gls	0.67901062	0.07458719	1.233855e-01	1.441965e-03	2.187007e-03
Ast	0.45480253	0.01511416	8.400068e-05	4.252155e-04	1.390491e-02
G-PK	0.63390934	0.06982363	8.354160e-02	4.083219e-02	3.617909e-03
PK	0.26246425	0.02830088	1.910532e-01	4.824747e-01	1.289404e-03
PKatt	0.26984563	0.03140962	1.960953e-01	4.673621e-01	1.341184e-03
CrdY	0.34276186	0.15544082	4.705963e-02	7.001922e-04	5.414687e-02
CrdR	0.03933796	0.02133164	1.428482e-02	3.343442e-03	5.874441e-01
Gls90	0.28831819	0.29993052	1.638023e-01	1.547807e-01	9.362014e-05
Ast90	0.01981034	0.41192382	5.084378e-01	5.445454e-02	1.133359e-03
G+A	0.10264279	0.61971942	2.652730e-01	5.872819e-03	1.182883e-03
G-PK90	0.24033114	0.29025922	1.241526e-01	2.623004e-01	6.226907e-06
G+A-PK	0.08967845	0.61090354	2.925888e-01	1.667566e-03	9.000208e-04

Nella tabella sono riportati i valori del cos^2 di ogni dimensione rispetto alle variabili di partenza. Una rappresentazione grafica permette di comprendere meglio il tutto.



Se il valore del cos^2 è alto, prossimo ad 1, allora la dimensione considerata rappresenta molto bene quella specifica variabile (ad esempio, le variabili *90s*, *Min*, *Starts* sono spiegate davvero bene dalle prime due PC, come vediamo nel grafico sopra riportato), e nel *correlation circle plot* la variabile si avvicina molto alla circonferenza. Invece, un valore di cos^2 basso ci dice che quella specifica variabile non è spiegata molto bene dalle PC considerate (si veda *Age* e *CdrR*), quindi nel *correlation circle plot* la variabile si avvicina molto al centro del grafico.

Per ogni variabile, la somma dei cos^2 sulle varie PC è pari ad 1. Alcune variabili necessitano di poche (una o due) componenti principali per essere rappresentate bene, altre invece hanno bisogno di più PC per avere un cos^2 accettabile.



- Le variabili con un cos^2 *alto* hanno un colore tendente all'arancione scuro;
- Le variabili con un cos^2 *basso* hanno un colore tendente all'azzurro/blu;
- Le variabili con un cos^2 *medio* hanno un colore tendente al giallo.

CLUSTER ANALYSIS

La **Cluster Analysis** è un insieme di tecniche le quali hanno il fine di raggruppare degli "oggetti" in base alle caratteristiche che possiedono. Quindi, permette il raggruppamneto di osservazioni, basato sulla *distanza*, tra loro simili per alcuni attributi.

Un *Cluster* è perciò un insieme di osservazioni tra loro molto simili e, se l'analisi dei cluster viene fatta correttamente, si avrà che le osservazioni all'interno dei singoli cluster hanno una somiglianza molto alta, invece i vari cluster saranno tra loro molto dissimili. La **somiglianza** è il grado di corrispondenza tra gli oggetti. Questa viene misurata con le *misure di correlazione* (poco usate) e le *misure di distanza* (molto usate). Per queste ultime, a valori elevati di distanza corrisponde poca somiglianza; distanza minore, invece, è uguale a somiglianza maggiore.

Distinguiamo l'*Analisi Gerarchica* e l'*Analisi non gerarchica*:

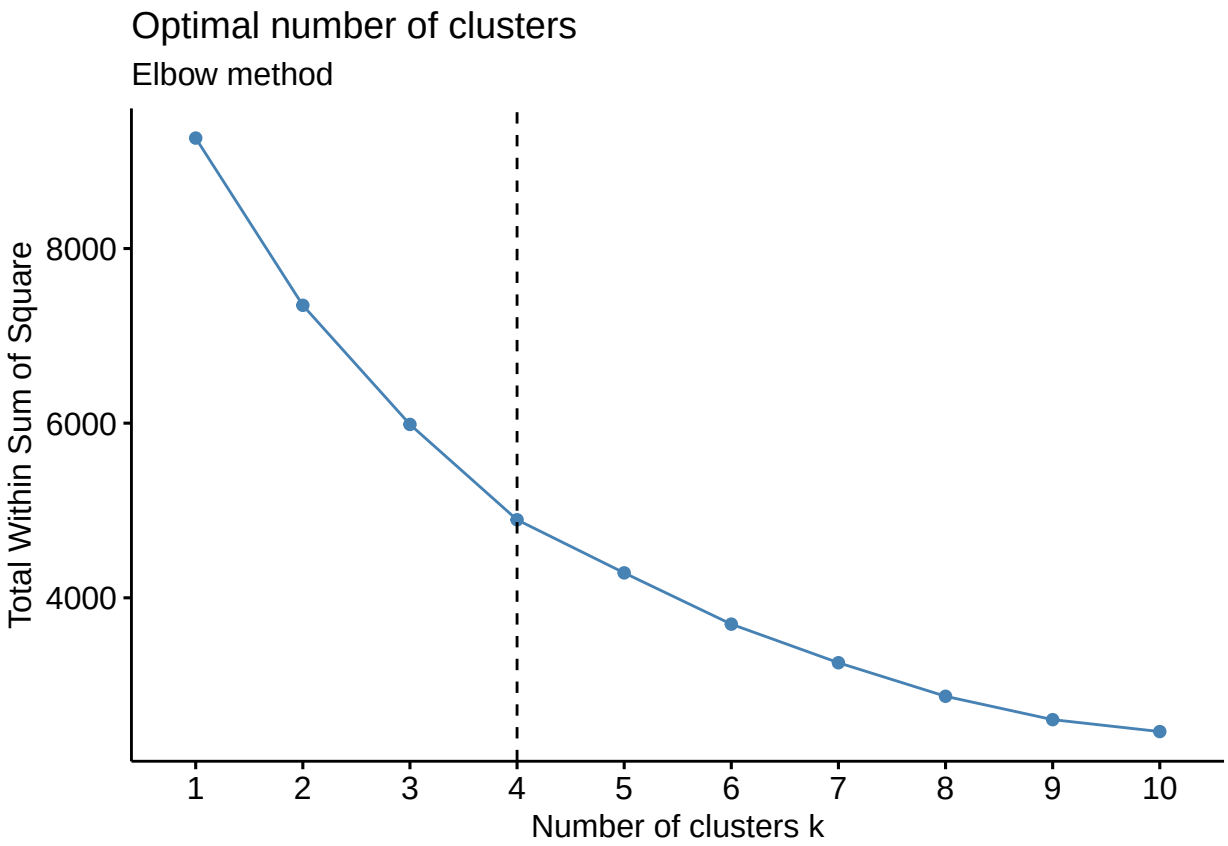
- **Analisi Gerarchica** : procedura step-by-step in cui vengono creati dei gruppi di record omogenei basandosi su specifiche caratteristiche, viene sfruttato un algoritmo (*Dendogramma*): *agglomerativo* (da n cluster ad 1) o *divisivo* (da 1 cluster ad n cluster).
- **Analisi non gerarchica** : questa non si serve del dendogramma e il tipo di metodo più usato è il *k-means*. Punta a suddividere n osservazioni in k cluster, ogni osservazione viene assegnata al cluster che ha una media vicina al valore dell'osservazione stessa.

Il modo più opportuno per fare un'ottima cluster analysis è utilizzare una combinazione delle due tipologie, dapprima usando l'analisi gerarchica e succesivamente la non gerarchica.

Analisi gerarchica (Hierarchical Clustering)

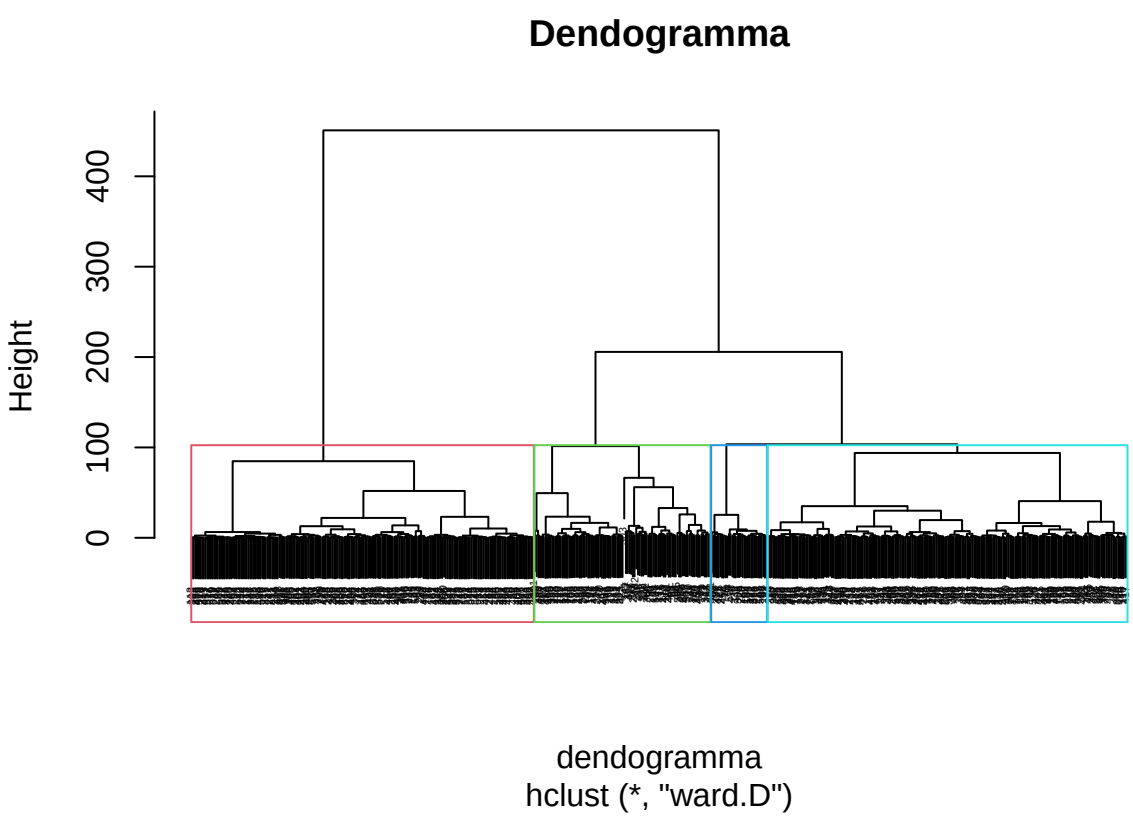
Come già detto in precedenza, l'anailisi gerarchica si basa su algoritmi divisivi e agglomerativi, i più utilizzati fra questi ultimi sono: il collegamento singolo, completo, medio, metodo del centroide e metodo di Ward. L'analisi gerarchica è utile per campioni non troppo grandi ed è, però, suscettibile agli outliers.

Per determinare il numero ottimale di cluster in cui raggruppare le nostre osservazioni possiamo tener conto di due grafici: l'**Elbow Method**,il quale mostra la percentuale di varianza spiegata in base del numero di cluster.



Grazie al grafico è possibile comprendere come i primi 4 cluster ($k = 4$) siano quelli in grado di conservare più varianza (a cui corrisponde il *gomito* del grafico); a partire dal cluster successivo la varianza spiegata da ogni cluster diminuisce, percò possiamo affermare che le nostre osservazioni potranno essere raggruppate in 4 cluster tramite l'analisi gerarchica.

Con l'analisi gerarchica è necessario servirsi di algorimti agglomerativi o divisi. Nel nostro caso ci serviamo di un algoritmo agglomerativo, ottenendo così la costruzione del **Dendogramma**, cioè un grafico ad "*albero*", dove ogni "*foglia*" rappresenta la singola osservazione (ognuna delle quali, all'inizio dell'analisi, rappresenta anche un singolo cluster). L'asse delle ascisse mostra la distanza tra i cluster (ossia lo strumento di cui ci serviamo per verificare la somiglianza tra i clusters); sull'asse delle ordinate è mostrato il livello gerarchico di aggregazione. L'algoritmo usato per creare il dendogramma è il *metodo di Ward ("ward.D")*, il quale si dimostra essere poco sensibile agli outliers.



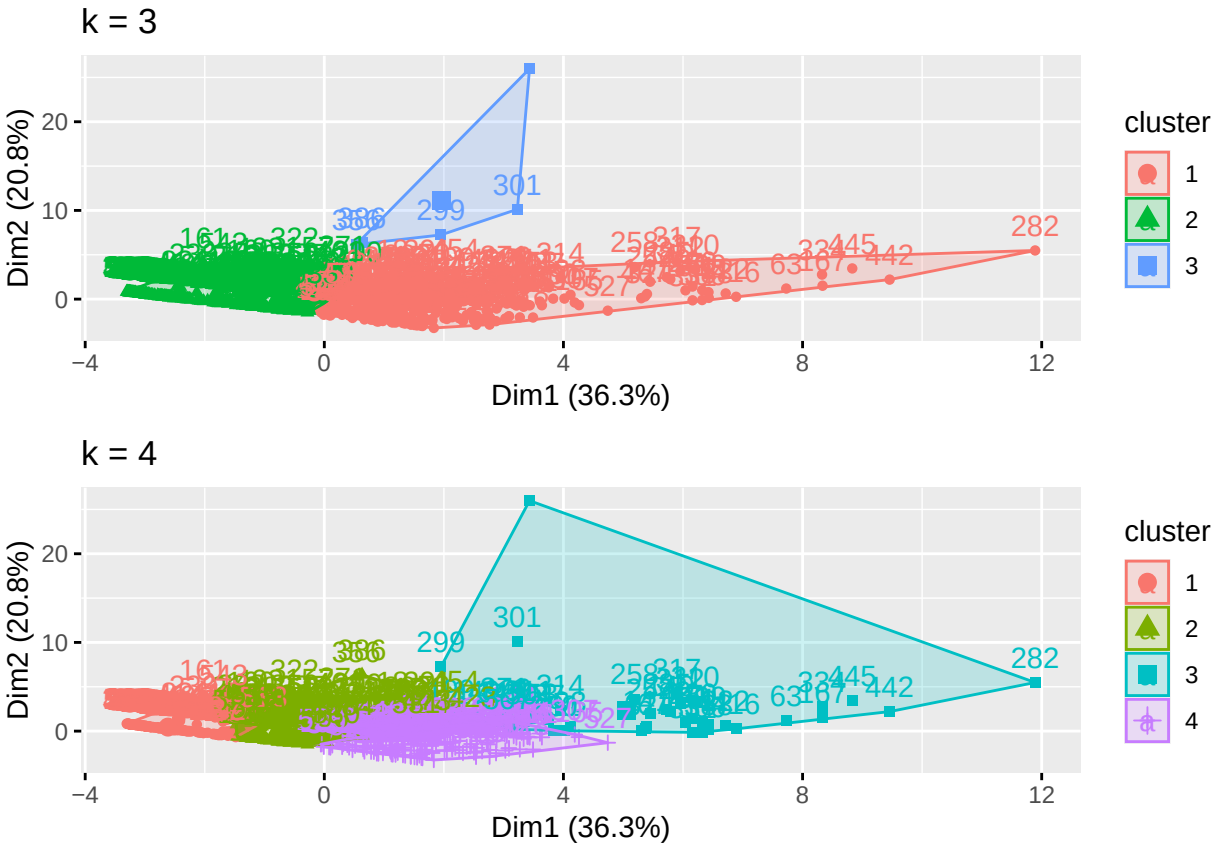
Per interpretare il grafico, dobbiamo dire che minore è la distanza (asse delle ascisse) e maggiore è il livello di somiglianza tra i cluster. Inoltre questo potrebbe darci la possibilità di conoscere il numero di cluster con cui effettuare l'analisi non gerarchica (informazione che abbiamo ottenuto grazie all'elbow method).

Analisi non gerarchica (k-means)

Una volta determinato il numero di cluster, possiamo effetuare l'analisi non gerarchica, la quale sfrutta il metodo *k-means*. Questo metodo punta ridurre la varianza intra-gruppo (ossia tra gli elementi facenti parte di uno stesso cluster), suddividendo n osservazioni in k cluster, ogni osservazione viene assegnata al cluster che ha un *centroide* (il quale di solito corrisponde alla media dei punti nel cluster) il cui valore è prossimo al valore dell'osservazione stessa.

Strumenti grafici k-means

Poiché conosciamo il numero di cluster k, procediamo con il raggruppamento delle osservazioni. Nel primo grafico poniamo un numero di clusters ($k = 3$) inferiore rispetto a quello "ottimale", nel secondo avviene il raggruppamento corretto in quattro clusters. Già grazie a questa occhiata al grafico vediamo come le osservazioni siano raggruppate decisamente meglio nel secondo grafico e non nel primo.


k = 3
Dim2 (20.8%)
Dim1 (36.3%)
cluster
1
2
3
k = 4
Dim2 (20.8%)
Dim1 (36.3%)
cluster
1
2
3
4

REGRESSIONE LINEARE

Il *modello di regressione lineare* è uno dei modelli più utilizzati per individuare le relazioni causa-effetto tra variabili.

Strumenti come lo scatter-plot, la covarianza e la correlazione ci permettono di capire soltanto se vi è una relazione di tipo lineare, e non funzionale.

Per costruire e usare il modello di regressione lineare è necessario seguire alcune fasi:

1. *Fase di specificazione*;
2. *Stima dei Parametri*;
3. *Verifica del modello (diagnostica).*

1. Fase di specificazione

Durante questa fase sono esplicitate le variabili che compongono il nostro modello.

$$Y = f(X_1, X_2, ..., X_p) + \epsilon$$

Dove Y è la *variabile dipendente*, mentre le X sono le *variabili indipendenti (o esplicative)*, le quali vanno ad influenzare Y, mentre ϵ rappresenta una *variabile residuale (o errore)* che racchiude l'insieme delle concause non note che influenzano la variabile dipendente Y e il suo comportamento deve essere imprevedibile.

Il modello si basa su 5 ipotesi differenti:

1. *Ipotesi di linearità* : la funzione è di tipo di lineare, linearità rispetto a β_j;
2. X è deterministica, ovvero non ha natura stocastica;
3. $E(\epsilon_i) = 0$;
4. *Ipotesi di omoschedacità* (varianza costante) : $Var(\epsilon_i) = \sigma^2$;
5. $Cov(\epsilon_i, \epsilon_j) = 0$

Il legame causa-effetto può essere di qualsiasi tipo, ma nella pratica si preferisce utilizzare una funzione lineare. Per questo motivo si parla di *regressione lineare multipla*, la quale ha la seguente formulazione:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_i X_i + \epsilon$$

Dove β_0 è detto termine noto, $\beta_1, \beta_2, \beta_i$ mentre sono detti coefficienti di regressione e, insieme alla varianza dell'errore, sono i parametri del modello da stimare sulla base delle osservazioni campionarie.

2. Stima dei Paramteri

Per stimare i parametri, in base alla situazione si potranno anche scegliere metodi diversi, ma in generale si sceglie il *metodo degli OLS* (minimi quadrati) minimizzando la somma dei quadrati degli stimatori.

3. Verifica del Modello di regressione (o fase di diagnostica)

Prima di essere utilizzato per effettuare delle prvisioni, è necessario verificare che il modello rispetti le ipotesi che abbiamo visto in precedenza. Per questo motivo si effettua la fase di diagnostica, in quanto qualora una o più ipotesi non dovessero essere verificate, la bontà del modello viene meno, ed è necessario ripartire dal primo step.

- Media degli errori non sia significamente diversa da zero : *Test t di Student*;

- Normalità della distribuzione degli errori : *Test di Shapiro Wilk*;
- Omoschedasticità dei residui : *Test di Breusch-Pagan*;
- Assenza di correlazione seriale (o autocorrelazione) : *Test di Durbin-Watson.*

Verranno poi considerati alcuni strumenti grafici nell'**analisi dei residui**, poi si verificherà la *bontà di accostamento del modello ai dati* tramite l'indice R^2.

Stima del modello di regressione del dataset

Consideriamo come osservazioni soltanto quelle in corrispndenza di calciatori che ricoprono la posizione di difensore. Quando il nostro modello di regressione viene costruito, l'indice R^2 è solitamente alto (esso può assumere valori tra 0 e 1). Un aumento di R^2 non significa necessariamente che la variabile aggiunta sia statisticamente significativa (la risposta a tale domanda passa per un test t).

Dal momento che R^2 è interpretabile comecompromesso tra bontà di adattamento e penalità dovuta al soprannumero di regressori "utili", una procedura ragionevole nella specificazione del modello consiste nel continuare ad includere regressori fino al momento in cui R^2 inizia a decrescere.

```
##
## Call:
## lm(formula = MP ~ Age + Starts + Min + `90s` + Gls + Ast + `G-PK` +
##     CrdY + PK + PKatt + CrdR + Gls90 + Ast90 + `G+A` + `G-PK90` +
##     `G+A-PK`)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7604 -1.3141 -0.2978  1.0208  6.8614
##
## Coefficients: (5 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.918564   0.873511   3.341  0.00102 **
## Age          -0.033971   0.033859  -1.003  0.31713
## Starts       -0.118060   0.171879  -0.687  0.49308
## Min           0.013441   0.052224   0.257  0.79720
## `90s`        -0.128850   4.701012  -0.027  0.97817
## Gls           0.007994   0.209388   0.038  0.96959
## Ast           0.147587   0.080037   1.844  0.06690 .
## `G-PK`              NA         NA      NA       NA
## CrdY          0.104299   0.064632   1.614  0.10841
## PK                  NA         NA      NA       NA
## PKatt               NA         NA      NA       NA
## CrdR          0.336809   0.379527   0.887  0.37607
## Gls90       -18.627065  42.344322  -0.440  0.66056
## Ast90       -21.341744  42.609453  -0.501  0.61710
## `G+A`        21.219101  42.611730   0.498  0.61914
## `G-PK90`            NA         NA      NA       NA
## `G+A-PK`            NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.71 on 173 degrees of freedom
## Multiple R-squared:  0.9764, Adjusted R-squared:  0.9749
## F-statistic: 649.7 on 11 and 173 DF,  p-value: < 2.2e-16
```

Miglioriamo il modello non considerando tutte quelle variabili aventi un coefficiente di regressione poco

significativo e sia tutte quelle che produrranno valori NA.

Usufruiamo quindi dell'algoritmo *Backward selection* per stimare i regressori.Tale algoritmo, considerando tutte le variabili del dataset e fissato un *livello di sgnificatività* ($\alpha = 0.05$), elimina la variabile con il coefficiente di regressione meno significativo, quindi le stime dei coefficienti delle variabili rimaste sono calcolate nuovamente e si ripete il procedimento sino a quando non vi sono più covariate che risultano non significative al livello prefissato. In questo modo, passiamo da un modello con 17 regressori iniziali, a soli 5 regressori.

```
##
## Call:
## lm(formula = MP ~ Min + Ast + CrdY + Ast90 + `G+A`)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5932 -1.1864 -0.3747  1.1403  6.8085
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.1015823  0.2436535   8.625 3.36e-15 ***
## Min          0.0106579  0.0001792  59.491  < 2e-16 ***
## Ast          0.1444512  0.0781574   1.848   0.0662 .
## CrdY         0.1149815  0.0622199   1.848   0.0663 .
## Ast90       -2.8353894  1.7538245  -1.617   0.1077
## `G+A`        2.7241850  1.7533191   1.554   0.1220
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.694 on 179 degrees of freedom
## Multiple R-squared:  0.976,  Adjusted R-squared:  0.9753
## F-statistic:  1455 on 5 and 179 DF,  p-value: < 2.2e-16
```

Perciò, il nostro miglior modello di regressione adesso sarà:

$$MP \sim Min + Ast + CrdY + Ast90 + (G + A)$$

Nota: (G+A) rappresenta, come sopra, un'unica variabile.

Test di specificità

Adesso verifichiamo che le ipotesi del modello siano verificate:

T Test

Il *T Test* ci permette di verificare che la media degli errori sia significativamente non diversa da zero.

$$H_0 : E(\epsilon_i) = 0$$
$$H_1 : E(\epsilon_i) \neq 0$$

```
##
##  One Sample t-test
##
## data:  residui
## t = -3.3105e-16, df = 184, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.2423937  0.2423937
## sample estimates:
##     mean of x
## -4.067223e-17
```

Rifiuterei il l'ipotesi nulla H_0 se $p - value < 0.05$. Dal test vediamo che il p-value è pari ad 1, perciò accettiamo l'ipotesi che la media dei residui non sia significativamente diversa da zero.

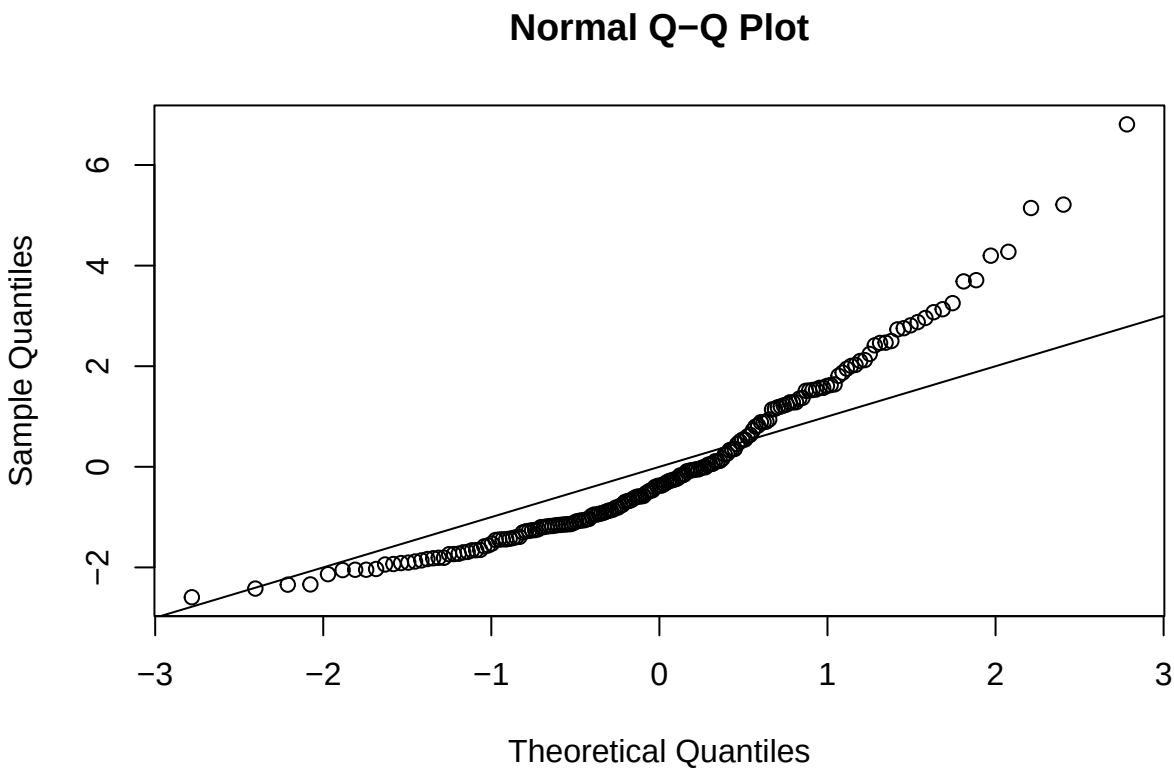
Shapiro-Wilk Test

Il *Test Shapiro-Wilk* ci permette di verificare, in questo caso, la normalità degli errori.

$$\epsilon_i \sim N(0, \sigma^2)$$

```
##
##  Shapiro-Wilk normality test
##
## data:  residui
## W = 0.92121, p-value = 1.984e-08
```

Il test dà come risultato un p-value di gran lunga inferiore al nostro livello di significatività (0.05), per questo motivo l'ipotesi nulla viene rifiutata, il che vuol dire che gli errori non si distribuiscono normalmente. Quindi l'ipotesi non è verificata. Si può notare anche dal *Normal Probability plot*, nel quale, se l'ipotesi fosse verificata, i punti si distribuirebbero molto vicini alla diagonale nel grafico e possiamo notare che non è così.



Procediamo con gli altri test.

Test Breusch-Pagan

Il *Test Breusch-Pagan* è utile per verificare l'ipotesi di omoschedasticità degli errori.

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 35.476, df = 5, p-value = 1.209e-06
```

Anche in questo caso, l'ipotesi non è verificata, quindi la varianza degli errori non è costante e si ha eteroschedasticità.

Test Durbin-Watson

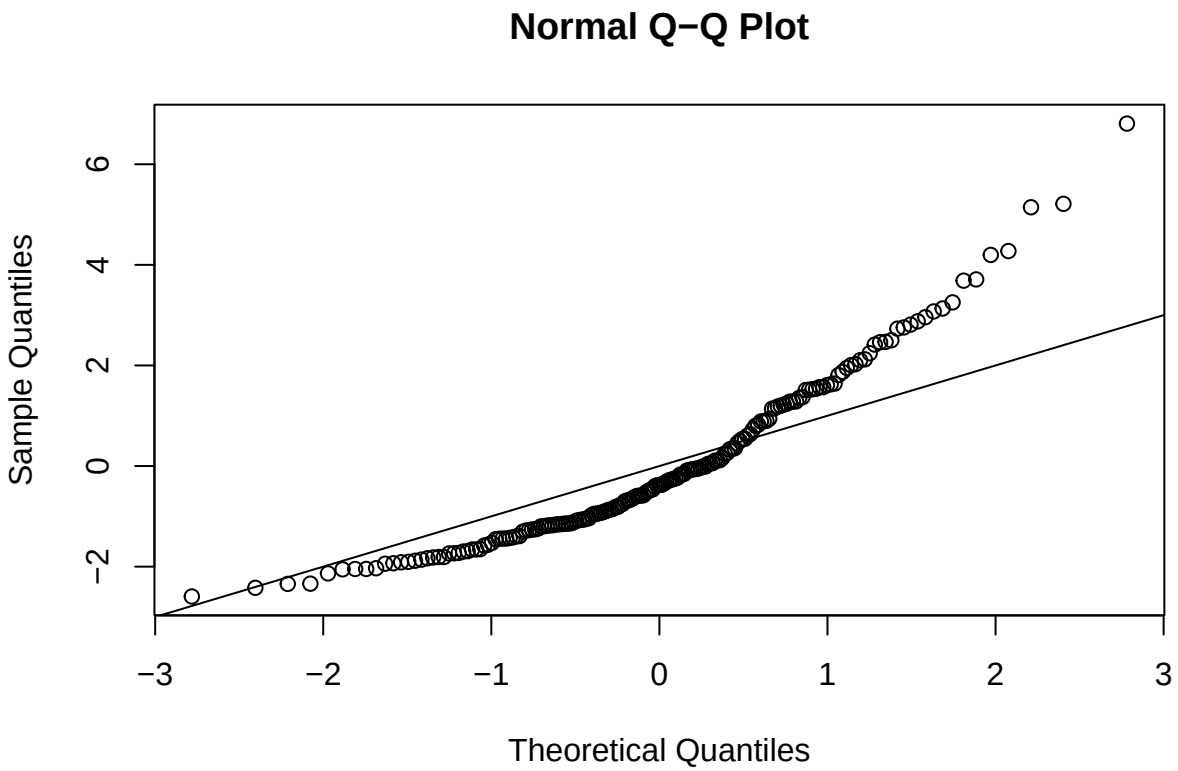
Tale test permette di verificare l'assenza di correlazione seriale degli errori, ovvero che gli errori non siano autocorrelati l'un l'altro.

```
##
##  Durbin-Watson test
##
## data:  model
## DW = 1.6401, p-value = 0.006137
## alternative hypothesis: true autocorrelation is greater than 0
```

Infine, anche nel caso di tale test, rifiuto l'ipotesi nulla in quanto il p-value è molto piccolo, perciò i residui sono tra loro autocorrelati.

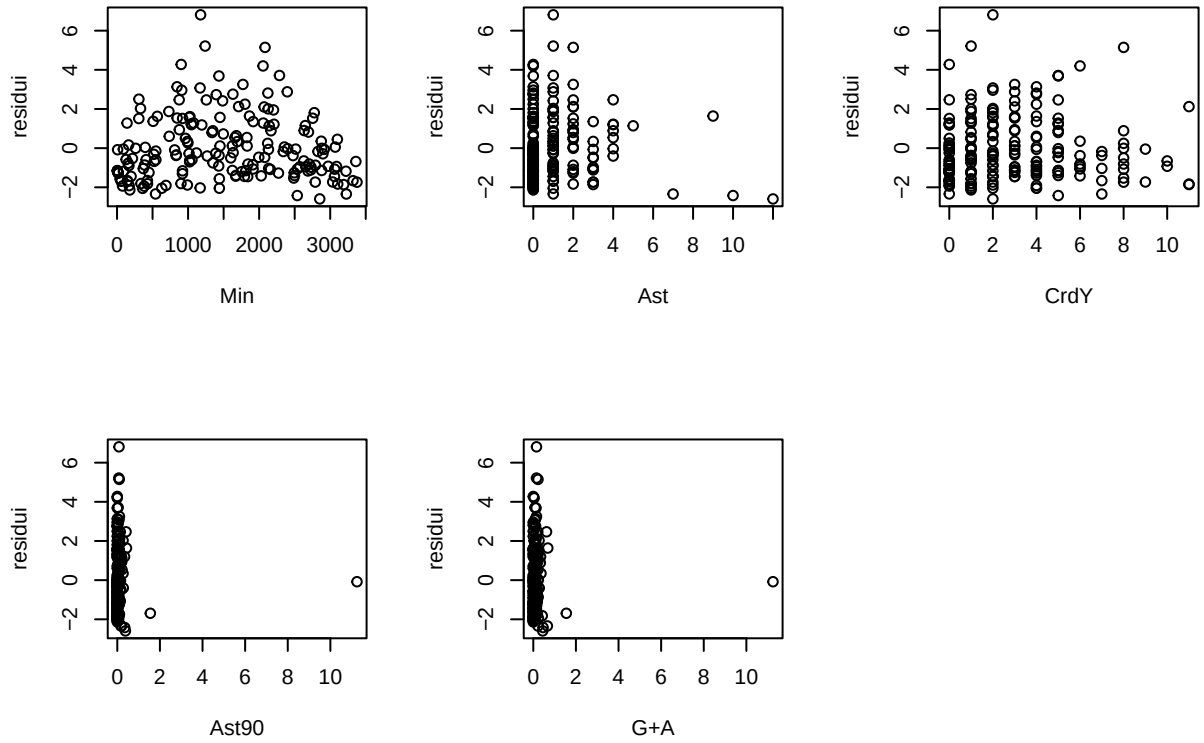
ANALISI DEI RESIDUI

Normal probability plot per verificare la normalità



All'interno del nostro *Normal Probability plot*, se l'ipotesi fosse verificata, i punti si distribuirebbero molto vicini alla diagonale nel grafico e possiamo notare che non è così, quindi gli errori non si distribuiscono normalmente.

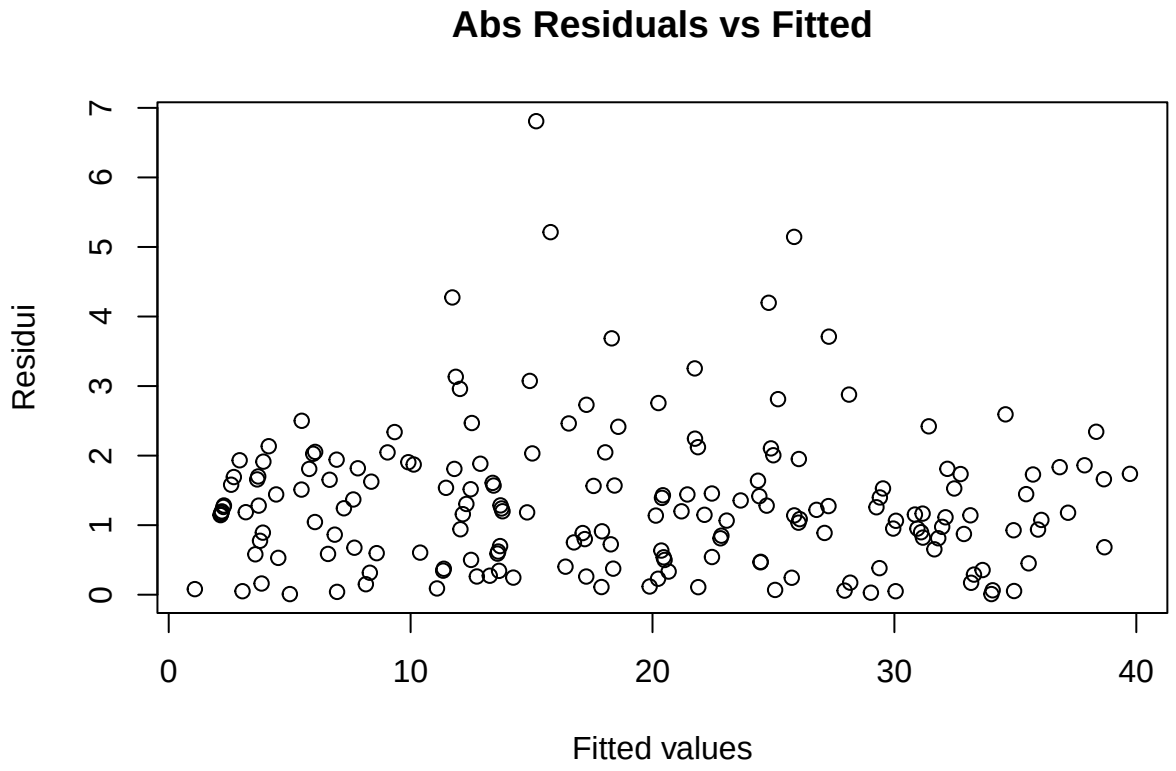
Verifica dell'incorrelazione tra le variabili esplicative e gli errori



La correlazione tra le variabili esplicative X e gli errori ϵ non deve esistere, in quanto, se così fosse, sarebbe possibile "prevedere" gli errori date le variabili, ma gli errori devono essere imprevedibili. L'ipotesi è confermata poiché dai grafici a dispersione sopra riportati non è possibile individuare alcun tipo di relazione lineare.

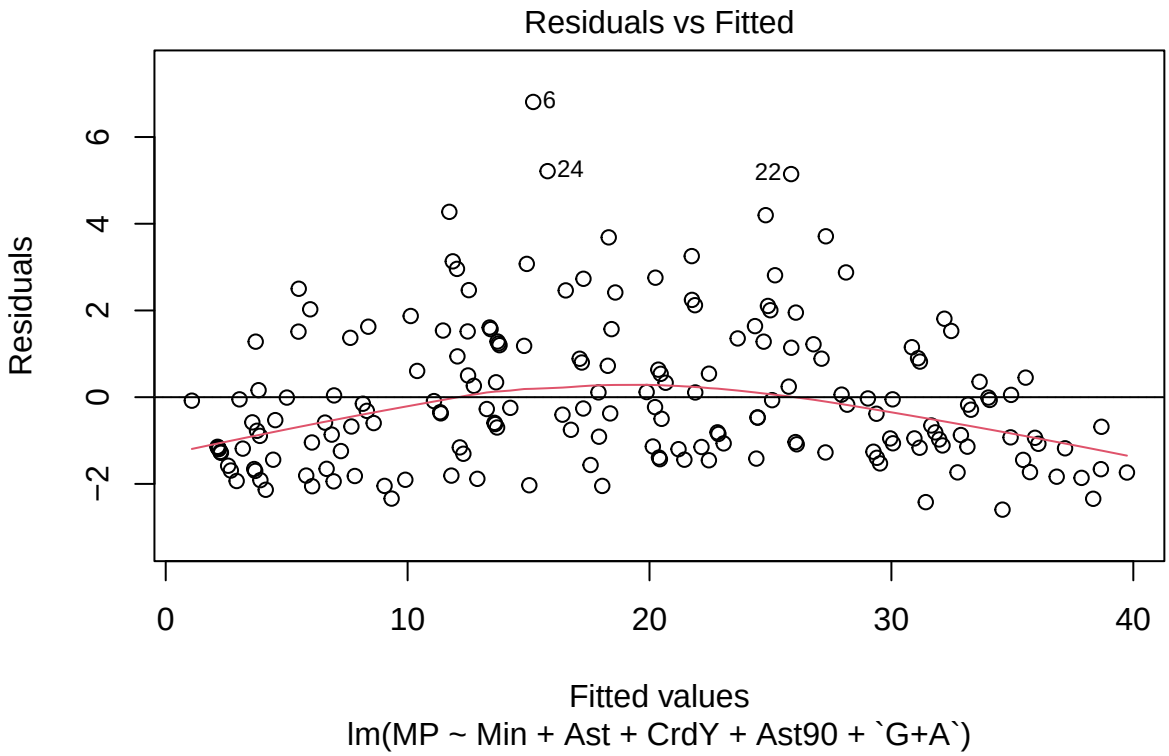
Ipotesi di omoschedasticità dei residui

Per verificare la presenza di omoschedasticità tramite l'analisi dei residui occorre tracciare il grafico dei residui in valore assoluto (ordinata) verso i valori stimati con il modello: la dispersione verticale dovrebbe essere approssimativamente costante. Nel grafico sottostante, questa dispersione verticale non si evidenzia, perciò vi è eteroschedasticità.



Ipotesi di distribuzione lineare dei residui

L'ipotesi deve essere verificata con lo stesso grafico *Residui vs Fitted values*. Nel grafico è andrà riportata una linea orizzontale in corrispondenza dei residui con media zero. Ricordiamo che i residui di un modello di regressione costruito con il metodo dei minimi quadrati (OLS) hanno per definizione sempre media zero. La linea rossa invece è una linea di tendenza, utile per verificare l'ipotesi. Se la linea rossa è abbastanza sovrapponibile alla linea tratteggiata, allora l'ipotesi di linearità è verificata. Secondo l'ipotesi di linearità, i dati devono infatti distribuirsi in modo casuale intorno allo 0. Se la dispersione non è casuale, ma assume un andamento preciso attorno allo 0, allora non vi è linearità nella distribuzione dei residui.



Vediamo che la dispersione dei dati visualizzabile grazie alla linea rossa non è *casuale*, quindi non vi è linearità nella distribuzione dei residui.

Verificare la presenza di outliers nel modello con l'analisi dei residui

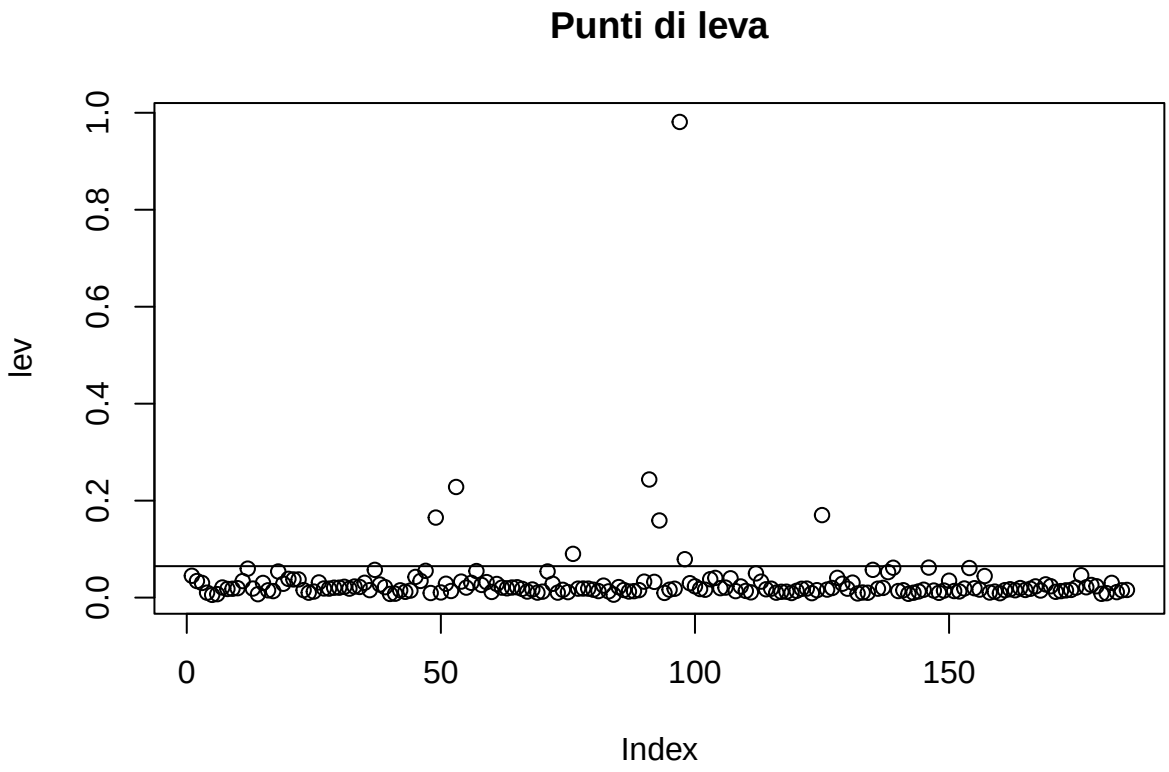
Il grafico utilizzato per verificare la linearità nella distribuzione degli errori, può essere sfruttato anche per individuare eventuali *outliers* nel modello. Semplicemente questi punti sono identificati da un numero e, nel grafico, sono più isolati rispetto agli altri punti. Rivedendo il grafico prcedente notiamo come siano presenti outliers al suo interno. Per verificare la presenza di outliers che influenzano la nostra retta di regressione, possiamo utilizzare anche:

- I **punteggi di leva** (*leverage points*) : sono compresi tra 0 ed 1. Un punteggio elevato di leva è quindi un valore vicino ad 1.
- La **distanza di Cook** : si considerano elevati i valori superiori ad 1.

Tutte queste tecniche utilizzano come approccio quello del togliere un'osservazione alla volta dal campione e vedere cosa cambia nei risultati. *Osservazioni che hanno valori elevati per tutte queste misure sono considerate un possibile problema.*

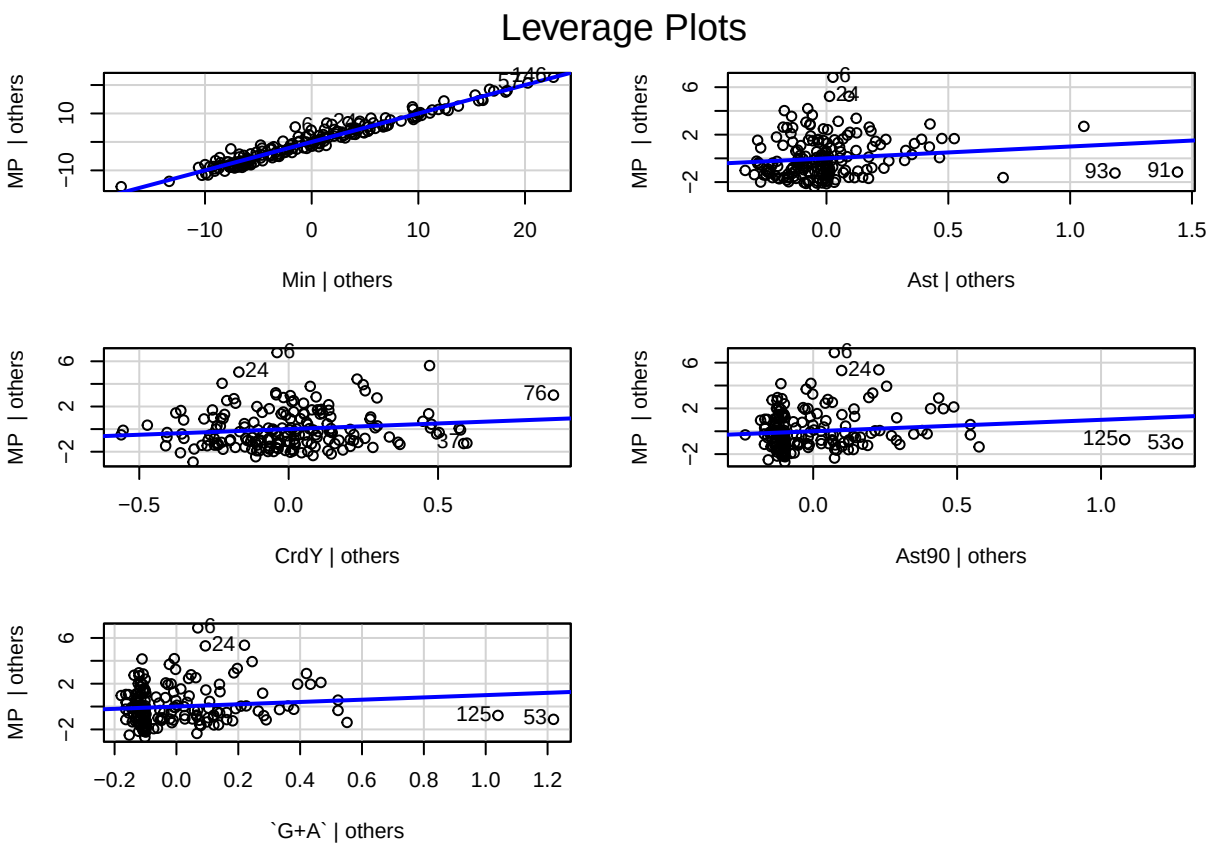
Punteggi di Leva

I punti con un elevato effetto di leva sono quei punti per cui si ha un altezza maggiore al rapporto, moltiplicato poi per due, tra la somma delle leve e la lunghezza di queste ultime (ovvero sono più distanti dalla retta orizzontale tracciata nel grafico). Possiamo vedere questi punti graficamente e identificare gli outliers.



Altro strumento utile per la diagnostica è il **partial leverage plots**. Quando le variabili esplicative sono più di una, la relazione tra i residui e una variabile esplicativa può essere influenzata per effetto degli altri regressori. Il partial leverage plots mette in evidenza queste relazioni.

Sull'asse delle ascisse sono rappresentati i residui della regressione della i.esima variabile esplicativa sui rimanenti k-1 regressori; sull'asse delle ordinate sono rappresentati i residui della regressione della variabile risposta su tutti i regressori escludento l'i.esimo. *Il partial leverage plots è usato per misurare il contributo della variabile indipendente al leverage di ciascuna osservazione, misura, cioè, come variano i punti di leva quando si aggiunge un regressore al modello.*

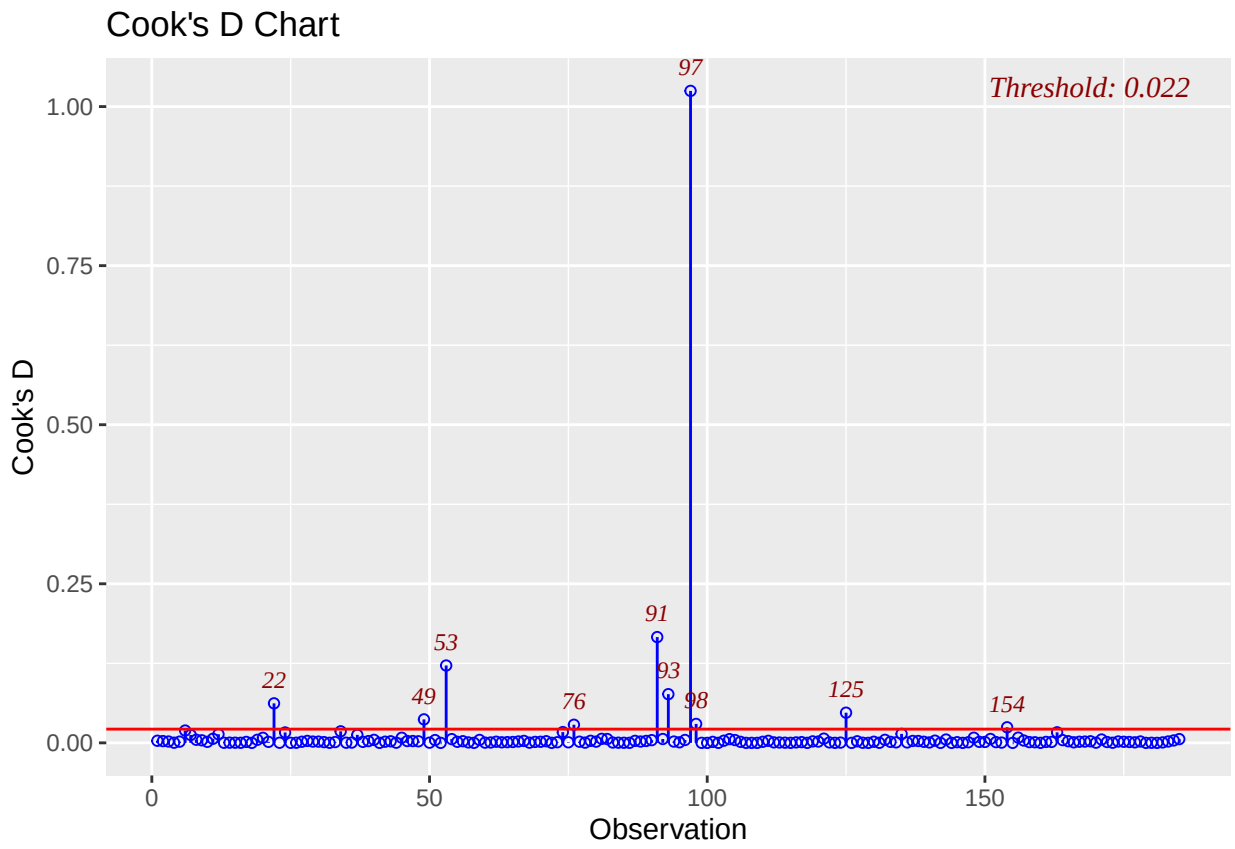


Ogni grafico ci mostra una linea di regressione con la pendenza corrispondente alla stima del parametro della variabile esplicativa presa in considerazione; mostrando inoltre una variazione quando vengono aggiunti al modello. Una variabile che risulta avere poca influenza sui punti di leva di ciascuna osservazione viene rappresentata vicino alla retta orizzontale $Y = 0$.

Distanza di Cook

La **Distanza di Cook** (D_i) ci permette di capire quale sia l'*influenza* di un punto in un modello di regressione cotruito col metodo dei minimi quadrati (OLS) e di comprendere come cambierebbe il nostro modello se uno specifico dato venisse rimosso. Punti con elevato residuo (outlier) o elevato leverage possono distorcere il risultato e l'accuratezza di un'analisi di regressione.

```
## $plot
```




```
##
## $outliers
##     observation cooks_distance
## 22           22     0.06210781
## 49           49     0.03691867
## 53           53     0.12158295
## 76           76     0.02847634
## 91           91     0.16627586
## 93           93     0.07654022
## 97           97     1.02464559
## 98           98     0.02976027
## 125         125     0.04742295
## 154         154     0.02446369
##
## $threshold
## [1] 0.02162162
```

Analizzando il nostro risultato, partiamo innanzitutto dal grafico. Sull'asse delle ascisse abbiamo le nostre osservazioni del modello (sono 185 osservazioni, poiché ricordiamo di aver costruito il modello solo per i

difensori); sull'asse delle ordinate abbiamo la distanza di Cook per ogni osservazione. Come vediamo varia da valori prossimi allo 0 ad un valore maggiore di 1. Per capire quanti outliers abbiamo vi sono 2 opinioni differenti:

1. L'osservazione rappresenta un outlier se assume un valore superiore ad 1;
2. L'osservazione rappresenta un outlier se assume un valore superiore ad una soglia (una distanza di cook) calcolata come di seguito e rappresentata mediante la linea rossa orizzontale nel grafico.

$$D_i = \frac{4}{n}$$

Nella formula n rappresenta il numero di osservazioni. Effetuando il calcolo otteremo che questo valore di soglia è pari a 0.02162162, valore che possiamo individuare nei risultati ottenuti al di sotto del grafico. Tra questi risultati sono presenti anche i nostri outliers (i quali sono 10 osservazioni e hanno una distanza di Cook superiore a questa soglia). Nel grafico, i punti indicati con un numero e che oltrepassano la linea rossa sono gli outliers del nostro modello.

Una volta che gli outliers sono stati individuati, devono essere rimossi per stimare il nuovo modello di regressione.

```
##
## Call:
## lm(formula = MP ~ Min + Ast + CrdY + Ast90 + `G+A`)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5932 -1.1864 -0.3747  1.1403  6.8085
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.1015823  0.2436535   8.625 3.36e-15 ***
## Min          0.0106579  0.0001792  59.491  < 2e-16 ***
## Ast          0.1444512  0.0781574   1.848   0.0662 .
## CrdY         0.1149815  0.0622199   1.848   0.0663 .
## Ast90       -2.8353894  1.7538245  -1.617   0.1077
## `G+A`        2.7241850  1.7533191   1.554   0.1220
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.694 on 179 degrees of freedom
## Multiple R-squared:  0.976,  Adjusted R-squared:  0.9753
## F-statistic:  1455 on 5 and 179 DF,  p-value: < 2.2e-16
```

CONCLUSIONE

Dopo aver effettuato le nostre analisi sul dataset considerato e aver estrapolato molte informazioni e aver costruito un modello di regressione lineare, è doveroso riportare che, in quest'ultima fase, sebbene gli outliers siano stati rimossi e i test di specificià necessari visti in precedenza siano stati ripetuti, i risultati ottenuti, in sostanza, non cambiano. In conclusione, i metodi utilizzati per stimare il nostro modello di regressione non sono ottimali, perciò sarebbe opportuno, nel nostro caso, fare affidamento ad altri metodi di stima dei parametri oppure utilizzare modelli di regressione di altra natura.