

Università degli studi di Milano - Bicocca
Corso di Laurea Magistrale in Data Science



I dati nel mondo del calcio

Data Visualization, appello di Gennaio 2023

Andrea Lucini Paioni, n° matricola 826578

Lorenzo Lecce, n° matricola 830881

Indice

1. Il Calcio in Europa	Pag. 3
1.1 Introduzione	Pag. 3
1.2 I top 5 campionati Europei	Pag. 3
1.3 Expected Goals	Pag. 6
2. Descrizione Database	Pag. 9
3. Le infografiche	Pag. 11
3.1 Realizzazione delle infografiche	Pag. 11
3.2 L'importanza degli expected Goals	Pag. 11
3.3 Confronto prestazioni difensive e prestazioni offensive	Pag. 12
3.4 L'importanza del possesso palla	Pag. 14
3.5 Gli effetti del Covid 19	Pag. 16
4. Valutazione delle infografiche	Pag. 18
4.1 Tecniche utilizzate	Pag. 18
4.1.1 Valutazione euristica	Pag. 18
4.1.2 Test utente	Pag. 20
4.1.3 Questionario psicometrico	Pag. 22
5. Sviluppi futuri e conclusioni	Pag. 26
6. Fonti	Pag. 27

1 Il Calcio in Europa

1.1 Introduzione

Nel corso degli ultimi anni, il mondo del calcio ha vissuto importanti cambiamenti sotto diversi punti di vista, da quello economico a quello tecnico-tattico. Col passare del tempo, è cresciuta l'importanza che hanno i dati e l'analisi di partite, allenamenti, profili dei giocatori nel migliorare i risultati, con squadre che utilizzano sempre più sofisticati sistemi di raccolta e analisi dei dati per ottenere un vantaggio competitivo, per ridurre al minimo gli infortuni, rendere al massimo delle proprie potenzialità, ricercare profili mirati in base alle esigenze della squadra. Le squadre sono sempre più attente alla preparazione fisica e all'alimentazione dei giocatori, vengono inserite nelle società sempre più figure altamente specializzate (come data analyst e scout), e utilizzano sofisticati sistemi di scouting per individuare in anticipo i talenti del futuro, così da poter avere un vantaggio economico sui competitor.

Inoltre, l'avvento del Covid-19 ha avuto un impatto sulla gestione tecnica e finanziaria delle società (essendo pur sempre aziende, esse hanno risentito non poco della crisi che ha portato il Covid in tutti i settori), costringendo le federazioni a dover adottare misure per garantire la sicurezza di spettatori e calciatori.

In questo report prenderemo in considerazione questi cambiamenti e l'influenza che hanno avuto sul mondo del calcio di questi ultimi anni. In particolare, ci concentreremo sull'impatto che può avere lo stile di gioco di ciascuna squadra sul suo rendimento, e sull'effetto causato dal Covid-19 per la presenza degli spettatori nei vari stadi dei top 5 campionati europei...

1.2 I top 5 campionati europei

Nel corso degli ultimi decenni, sono cinque i campionati europei di calcio che, anche grazie a budget superiori agli altri competitor, si sono imposti come i migliori e si sono spartiti i trofei internazionali più importanti, investendo cifre significative per l'acquisizione dei migliori talenti a livello mondiale, per il miglioramento di strutture e

metodi di preparazione delle partite e delle stagioni, per lo sviluppo dei loro brand a livello nazionale e internazionale...

Si tratta della Premier League (la massima serie Inglese, campionato fondato nel 1992 come successore della First Division nata nel 1888/89), della Serie A (la massima serie Italiana, la cui prima edizione risale al 1898/99), della Bundesliga (o 1. Fußball-Bundesliga, ovvero la massima serie professionistica del campionato Tedesco di calcio, fondata nel 1962), della Primera Division (meglio nota come LaLiga Santander o semplicemente LaLiga, ovvero la massima serie professionistica del campionato di calcio Spagnolo, che nasce nel 1929) e della Ligue 1 (la massima serie professionistica del calcio Francese, nata nel 1932).

Sono competizioni a cui partecipano ogni stagione 20 squadre (tranne per la Bundesliga, che è l'unico di questi campionati ad essere a 18 squadre), che ogni anno si contendono non solo la conquista del titolo nazionale, ma anche l'accesso alle prestigiose competizioni internazionali europee, ovvero la UEFA Champions League (che ha preso il posto nel 1992 della vecchia Coppa dei Campioni), la UEFA Europa League (che ha sostituito la vecchia Coppa Uefa dal 2009 in avanti) e la recentissima UEFA Conference League, nata solamente nel 2021 ed ora alla seconda edizione di sempre.

Nel corso delle 38 partite che compongono una stagione (34 per la Bundesliga), le squadre lottano non solo per le alte posizioni in classifica, ma anche per evitare la retrocessione in categorie minori, in caso di arrivo agli ultimi posti. In particolare, la formula è identica per Premier League, Serie A e LaLiga:

- Dal 1° al 4° posto in classifica: le squadre hanno accesso diretto alla UEFA Champions League;
- 5° e 6° posto: accesso diretto alla UEFA Europa League;
- 7° posto: accesso ai preliminari di UEFA Conference League;
- Dal 18° al 20° posto: retrocessione diretta nella lega minore (Championship per l'Inghilterra, Serie B per l'Italia, Segunda Division per la Spagna), sostituite da tre squadre provenienti da essa.

Per la Bundesliga, le prime 7 classificate hanno lo stesso trattamento delle squadre inglesi, spagnole e italiane; tuttavia, la situazione cambia nella zona retrocessione, anche per il minor numero di squadre presenti: la 17° e la 18° squadra sono retrocesse direttamente in Bundesliga 2, e la 16° affronta uno spareggio andata/ritorno con la terza classificata della seconda lega tedesca, per sancire la permanenza in Bundesliga della 16° arrivata o l'eventuale discesa della stessa nella lega minore.

Diverso, invece, è il caso della Ligue 1 che, a causa di risultati più altalenanti delle squadre francesi nelle competizioni europee nell'ultimo decennio, ha perso punti nel ranking UEFA, un punteggio, basato sul rendimento internazionale delle diverse squadre, che sancisce il numero di slot disponibili per le competizioni UEFA di ogni nazione:

- 1° e 2° posto in classifica: le squadre hanno accesso diretto alla UEFA Champions League;
- 3° posto: accesso ai preliminari di Champions League;
- 4° posto: accesso diretto alla UEFA Europa League
- 5° posto: accesso ai preliminari di UEFA Conference League
- 18° posto: spareggio andata/ritorno con la terza classificata della seconda lega francese (formula identica allo spareggio presente in Bundesliga per la 16° in classifica).
- 19° e 20° posto: retrocessione diretta nella lega minore (Ligue 2), sostituite da due squadre provenienti da essa.

Tenendo conto del livello di competitività di tutti questi campionati nazionali, ogni squadra cerca sempre di preparare al meglio una stagione in base agli obiettivi stabiliti, puntando sempre al miglior risultato possibile: un buon piazzamento in campionato è questione di prestigio (per attirare giocatori, sponsor, investimenti), di ambizione (per presidente, azionisti, tifosi, giocatori, allenatori...) e, perchè no, di guadagno, considerando quanto comportano a livello di profitti gli sponsor, i diritti tv, il merchandising, la partecipazione alle coppe europee, il player-trading legato allo sviluppo dei giocatori...

Per questi motivi, in uno sport come il calcio - tra i più seguiti al mondo anche perchè è quello dove più spesso capita che una squadra meno forte possa vincere contro una squadra sulla carta migliore -, le società che investono decine, centinaia di milioni ogni stagione non lasciano più nulla al caso, e prestano molta attenzione ai particolari nella preparazione delle stagioni, sempre più con occhio di riguardo verso le statistiche e i numeri in modo da avvicinarsi più facilmente ai propri obiettivi. E tra le statistiche più importanti, gli expected goals negli anni stanno assumendo un ruolo sempre più fondamentale.

1.3 Expected Goals

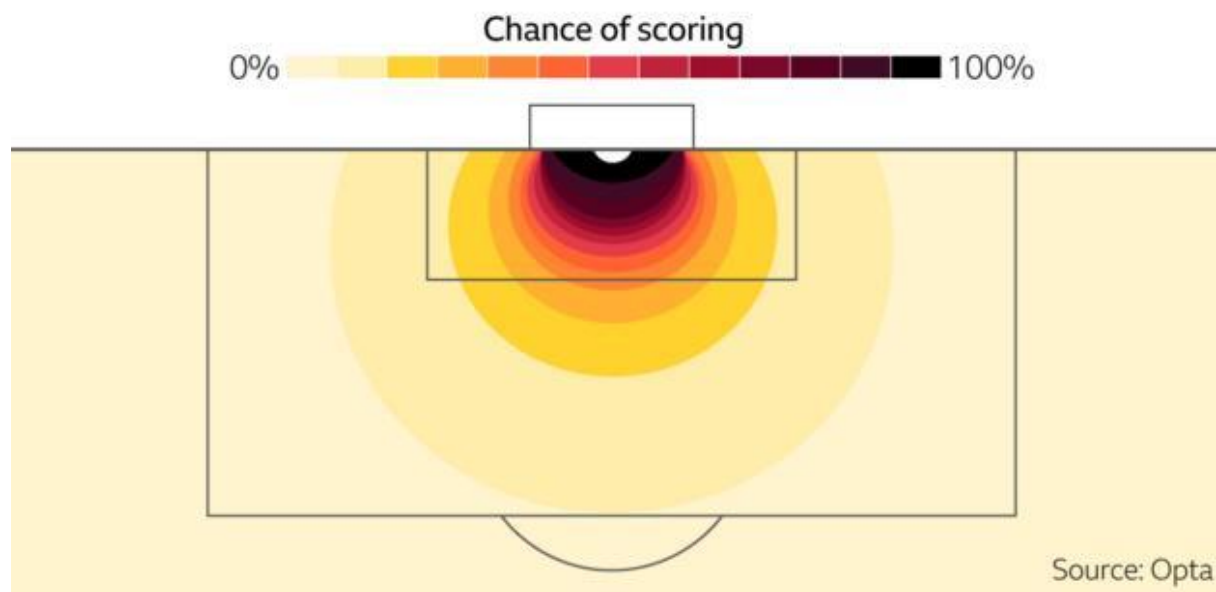
Gli expected goals (spesso definiti col diminutivo xG) sono una misura statistica che può essere utilizzata per valutare la qualità delle occasioni da rete che una squadra riesce a creare durante una partita di calcio. L'idea che sta alla base degli expected goals è la seguente: ad ogni azione offensiva che termina con una conclusione viene assegnato un valore numerico, pari alla probabilità che quest'azione si trasformi in un goal. La probabilità dipende da diversi fattori, tra cui:

- la posizione in cui viene effettuato il tiro (che quindi tiene conto sia dell'angolo di tiro che della distanza dalla porta);
- la posizione del portiere e dei difensori (due tiri dalla stessa posizione, uno a porta vuota, l'altro con diversi giocatori tra pallone e rete, hanno ovviamente valori differenti;
- la parte del corpo con cui viene effettuata la conclusione (colpo di testa, piede, petto, ginocchio...)
- la tipologia di assist ricevuto;
- l'altezza a cui viene impattato il pallone (un pallone rasoterra è più semplice da colpire di uno a mezz'aria).
- la situazione di gioco in cui si svolge l'azione che porta alla conclusione (punizione, rigore, azione manovrata...)

Nel concreto, un giocatore può effettuare una conclusione da lunga distanza a cui viene assegnato un xG di 0.1, ovvero con una probabilità del 10% che il tiro si

tramuti in rete. Viceversa, un tiro ravvicinato, che può avere un xG di 0.9, è sinonimo di alta probabilità (del 90% esattamente) che vada a segno.

How likely is a goal from different positions?



1.1 Esempio di modello utilizzato per calcolare gli expected goals; frutto dell'azienda Opta, tra le società leader a livello mondiale nella creazione di dati sportivi. In questo caso si tiene conto solamente della posizione da cui parte la conclusione, al momento senza ancora considerare altri fattori come la posizione di portiere e difensori, la parte del corpo con cui viene colpita la palla o l'assist ricevuto.

Gli xG vengono dunque utilizzati per valutare non solo le prestazioni offensive di una squadra, ma anche quelle di un singolo giocatore, e il tutto può essere circoscritto ad una singola partita o addirittura ad un'intera stagione. Ad esempio, una squadra avente xG totali di 4.0 in una partita ma che riesce a segnare solo 1 goal non è riuscita a sfruttare affatto le occasioni create; allo stesso modo, se un giocatore ha un xG di 0.4 in una partita ma riesce a terminare la partita con 2 goal effettuati, vuol dire che il giocatore ha sfruttato al meglio le poche opportunità che ha avuto a sua disposizione, e ha avuto una performance al di sopra della media.

Oltre agli expected goals, un'altra metrica utile strettamente legata ad essi sono gli expected goals against (sigla xGA): si tratta di una misura statistica, opposta rispetto agli expected goals, usata per la valutazione delle opportunità di segnare che una squadra di calcio concede all'avversario all'interno di una partita. Gli xGA sono calcolati allo stesso modo degli xG: ad una valore di 1.8 xG creati da una squadra A

corrisponde un valore di 1.8 xGA che la squadra avversaria B ha concesso alla squadra A.

Gli xG, alla pari degli xGA, sono metriche utili per valutare la prestazione di una squadra o di un giocatore, sia in una singola partita, che per un'intera competizione o per analizzare l'andamento di una squadra in momenti diversi della stagione; è bene ricordare che, comunque, si tratta solo di stime della probabilità di segnare/subire una rete, e non garantiscono il risultato finale di una partita: una squadra può terminare una partita senza segnare neanche una rete pur avendo valori molto elevati di xG, così come una squadra può subire diversi goal nonostante abbia concesso poche occasioni con valori bassi di xGA.

Tuttavia, si tratta di valori sempre più studiati ed analizzati dalle società sportive, e dagli analyst presenti nello staff delle squadre, perché possono dare indicazioni su cui poter contare: possono indicare situazioni di gioco dove una squadra tende a concedere più gol del normale (calci piazzati, inserimenti senza palla...), possono mostrare giocatori particolarmente efficaci o altri che spesso tendono a sprecare molte occasioni prima di segnare un gol, possono dare indicazioni sul dove poter migliorare la squadra con nuovi acquisti, sia a livello difensivo che a livello offensivo.

2. Descrizione Database

Il database utilizzato per questa analisi è stato preso dal sito <https://fbref.com/en/>, e ricavato dalla fusione di cinque dataset differenti; ognuno di essi esprimeva i dati di ciascuna squadra dei 5 campionati sopra descritti, per le seguenti stagioni: 2017/18, 2018/2019, 2019/2020, 2020/2021 e 2021/2022.

Per ciascun dataset sono dunque presenti 98 osservazioni, pari al numero di squadre rappresentate (le 20 squadre italiane, inglesi, francesi e spagnole e le 18 tedesche), e per ciascuna di esse vi sono 35 colonne. Inoltre, per ogni stagione, sono stati integrati dei dati più specifici relativi alle caratteristiche di gioco di ciascuna squadra. Per l'esattezza sono state aggiunte 10 colonne di dati per ogni squadra di ogni campionato.

Dunque, complessivamente il dataset è composto da 490 osservazioni, per 35 variabili. Per ogni squadra sono riportate le seguenti informazioni:

- **Rk:** E' un indice generato su tutti i campionati per ordinare le squadre, si basa sui punti ottenuti rispetto alle partite giocate.
- **Squad:** Indica il nome della squadra.
- **Country:** Indica il paese della squadra
- **League Rk:** Indica il posizionamento finale in classifica..
- **Match Played:** Indica il numero di partite giocate.
- **Win:** Indica il numero di partite vinte.
- **Draw:** Indica il numero di partite pareggiate.
- **Loss:** Indica il numero di partite perse.
- **Goals Scored:** Indica il numero di gol fatti.
- **Goals Against:** Indica il numero di gol subiti.
- **Goals Difference:** Indica la differenza tra gol fatti e gol subiti.
- **Points:** Indica il totale di punti fatti
- **Points/MP:** Indica il rapporto tra punti fatti e partite giocate.
- **XG:** Indica il totale di expected goal fatti.
- **XGA:** Indica il totale di expected goal subiti
- **XGD:** Indica la differenza totale di expected goal.

- **XGD/90:** Indica la differenza di expected goal per ciascuna partita
- **Attendance:** Indica la media di tifosi presenti allo stadio nelle partite casalinghe di una stagione intera.
- **Top Team Scorer:** Indica il nome del miglior marcatore della squadra.
- **Goalkeeper:** Indica il nome del portiere titolare della squadra.
- **Year:** Indica l'anno calcistico a cui sono associati i dati
- **Players Used:** Indica il numero totale di giocatori impiegati durante l'anno.
- **Possession:** Indica la percentuale di possesso palla medio durante l'anno.
- **Touches:** Indica il totale di tocchi di palla da parte dei giocatori della squadra.
- **Def Pen:** Indica il totale di tocchi di palla nella propria area di rigore.
- **Def 3rd:** Indica il totale di tocchi di palla nel terzo di campo vicino alla propria porta.
- **Mid 3rd:** Indica il totale di tocchi di palla nel terzo di campo centrale.
- **Att 3rd:** Indica il totale di tocchi di palla nel terzo di campo vicino alla porta avversaria.
- **Att pen:** Indica il totale di tocchi di palla nell'area di rigore avversaria.
- **Live Touches:** Indica il totale di tocchi di palla in movimento (esclude i calci da fermo).
- **Dribbling Completed:** Indica il totale di dribbling completati.
- **Dribbling Attempted:** Indica il totale di dribbling provati.
- **Dribbling %:** Indica la percentuale di dribbling riusciti.
- **Pass Received:** Indica il totale di passaggi ricevuti correttamente.
- **Progressive Pass:** Indica il totale di passaggi effettuati verso l'area di rigore avversaria.

3. Le infografiche

3.1 Realizzazione delle infografiche

L'infografica che vogliamo proporre è composta da più grafici, disposti su più pagine consecutive.

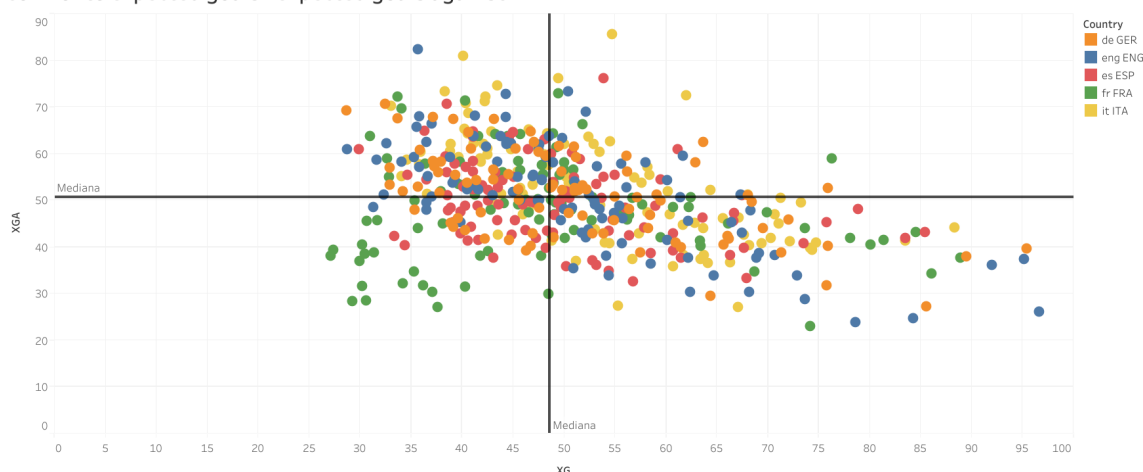
Gli obiettivi che ci siamo posti sono quelli di analizzare alcune delle metriche più utilizzate nel mondo del calcio negli ultimi anni, analizzando quanto influiscono valori come possesso palla, expected goals e differenza reti contribuiscono ai risultati delle squadre nei principali campionati di calcio europei, e vedere se alcune di queste sono più significative per il raggiungimento di obiettivi di squadra importanti.

Il target a cui è indirizzata la seguente infografica è composto da persone con un interesse per il mondo del calcio: appassionati, tifosi, eventualmente anche esperti del settore, comunque non persone che si affacciano per la prima volta sul mondo del pallone. Infatti andiamo ad analizzare aspetti tattici e metriche non scontate, per cui serve una conoscenza di base dello sport.

3.2 L'importanza degli expected goals

La prima visualizzazione che presentiamo propone un confronto tra due metriche fondamentali, ovvero gli expected goals (variabile **XG**, sull'asse delle ascisse) e gli expected goals against (variabile **XGA**, sull'asse delle ordinate). Il colore mostra i dettagli relativi al campionato (variabile **Country**), come indicato dalla legenda a lato. Abbiamo deciso di dividere in quadranti il grafico, col calcolo delle due mediane, per individuare squadre con valori simili per xG e xGA. Inoltre è possibile vedere dalla visualizzazione su Tableau il posizionamento finale in classifica (dato dalla variabile **League RK**).

Confronto expected goals - expected goals against



3.1 Scatterplot xG-xGA

Dando una rapida occhiata a questa prima visualizzazione, e interagendo col grafico, si può dedurre che le squadre presentano numero di expected goals fatti più alti, ed expected goals subiti più bassi, sono quelle che solitamente si posizionano meglio in classifica (quelle del quadrante in basso a destra, tra cui spiccano Manchester City, Liverpool, Bayern Monaco e PSG). Al contrario, quelle con numero di expected goals fatti più bassi, ma un numero elevato di expected goals subiti, tendono a posizionarsi nelle zone basse della classifica (sono le squadre presenti nel quadrante in alto a sinistra).

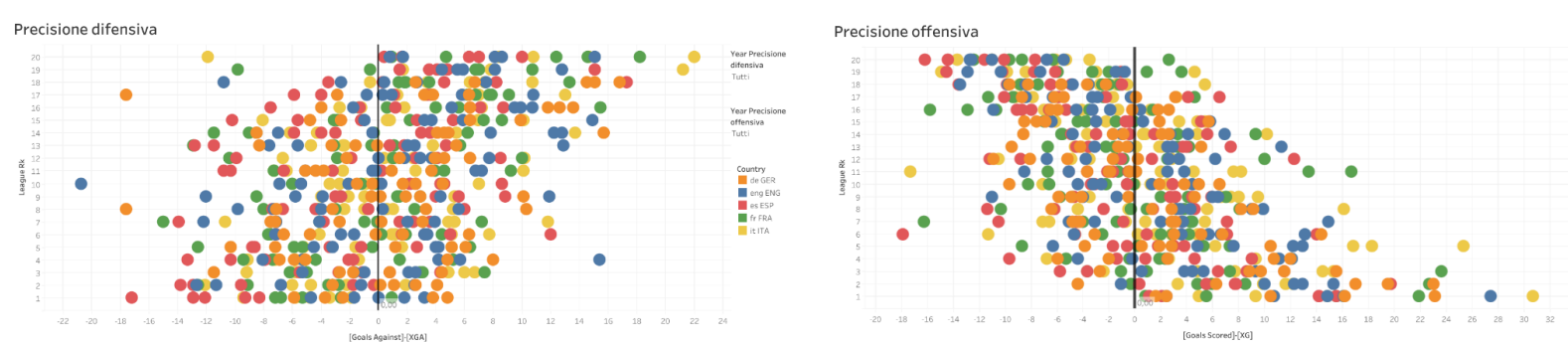
3.3 Confronto prestazioni difensive e prestazioni offensive

Un altro confronto interessante è quello tra gli expected goals, che rimangono comunque una stima della pericolosità offensiva di una squadra, e il numero effettivo di gol segnati (e, allo stesso modo, anche il confronto tra expected goals against e gol subiti), indicati rispettivamente con le differenze **Goals Scored - XG** (grafico sulla destra) e **Goals Against - XGA** (grafico sulla sinistra).

Infatti questi valori possono essere significativi nel mostrare, rispetto alle occasioni da rete create, le effettive capacità finalizzative dei giocatori (giocatori che segnano più della media rispetto alle occasioni avute), e allo stesso tempo l'abilità difensiva dei componenti della rosa (a parità di occasioni, squadre che subiscono meno gol rispetto a quanti dovrebbero).

I seguenti grafici permettono un confronto simile, con indicazioni relative alla posizione in classifica (variabile **League Rk**, sull'asse delle ordinate) rispetto alle variabili sulle ascisse indicate sopra **Goals Scored - XG** (grafico sulla destra) e **Goals Against - XGA** (grafico sulla sinistra).

Inoltre, abbiamo deciso di indicare le differenze tra i vari campionati (variabile **Country**, definita dal colore come indicato nella legenda), e lasciando la possibilità di scegliere sia una stagione in particolare con i due filtri (variabile **Year**, distinta per i due grafici), sia una vista generale delle cinque stagioni dalla 2017/18 alla 2021/22.



3.2 Scatterplot (Goals Against - xGA) - League RK sulla sinistra, scatterplot (Goals Scored - xG) - League RK sulla destra.

Dalla visualizzazione 3.2 si può osservare che, tendenzialmente, le squadre classificate tra le prime posizioni, presentano una differenza alta tra **Goal scored** ed **Expected Goals**, ciò sta ad indicare una buona capacità realizzativa da parte delle prime squadre, ciò potrebbe essere dovuto ad una qualità di giocatori migliori (come gli esempi della Juventus, del Manchester City, del Bayern Monaco, della Lazio); lo stesso discorso si può effettuare per i **gol subiti**, dove le squadre che si sono posizionate meglio hanno subito meno gol rispetto a quanto ci si potrebbe aspettare (come il Real Madrid, il Barcellona, l'Atletico di Madrid...).

Viceversa, le squadre posizionate nella parte bassa della classifica hanno effettuato meno gol di quanto ci si potrebbe aspettare e ne hanno subiti di più rispetto alle aspettative, sinonimo di giocatori di qualità inferiore...

3.4 L'importanza del possesso palla

Abbiamo deciso poi di considerare il possesso palla in quanto spesso, negli ultimi anni, con l'evoluzione tattica che sta avendo il gioco del calcio, ci sono scuole di pensiero molto contrastanti relativamente a questa metrica, da sempre tra le più considerate ed osservate nel mondo del pallone, e ai benefici/svantaggi che esso porta in base ai valori osservati. Infatti il possesso palla non è da solo sinonimo di successo, e ogni squadra ed allenatore devono riuscire a trovare il giusto equilibrio tra possesso palla e capacità di creare opportunità di gol, a seconda delle proprie caratteristiche, dello stile di gioco adottato e persino degli avversari che affrontano.

Alcuni allenatori (come Guardiola o Sarri) e alcune società (il Barcellona in primis, dove si è sviluppato il cosiddetto tiki-taka a partire dal 2008 con l'arrivo di Guardiola in panchina, ma anche altre società come il Bayern Monaco, il Manchester City, il PSG e il Real Madrid in alcune stagioni) sono tra i principali sostenitori di un calcio in cui il possesso palla fa da padrone, ed è imprescindibile proprio come filosofia di gioco. Anche lo sviluppo delle giovanili, i giocatori acquistati, gli allenatori ingaggiati, sono tutte decisioni in cui lo stile di gioco è tra i fattori principalmente considerati.

Al contrario, altri allenatori si affidano ad un gioco più diretto, non necessariamente dominando la partita con numeri elevati per quanto riguarda il possesso palla, ma invece cercando di sfruttare gli spazi in avanti, con pressing alto efficace, o di sorprendere l'avversario con le transizioni rapide rese possibili spesso dalle grandi doti atletiche dei loro interpreti, come le squadre di Klopp (Borussia Dortmund prima, ora il Liverpool), il Milan di Pioli, le squadre in orbita Redbull (il Lipsia in Germania e il Salisburgo in Austria)...

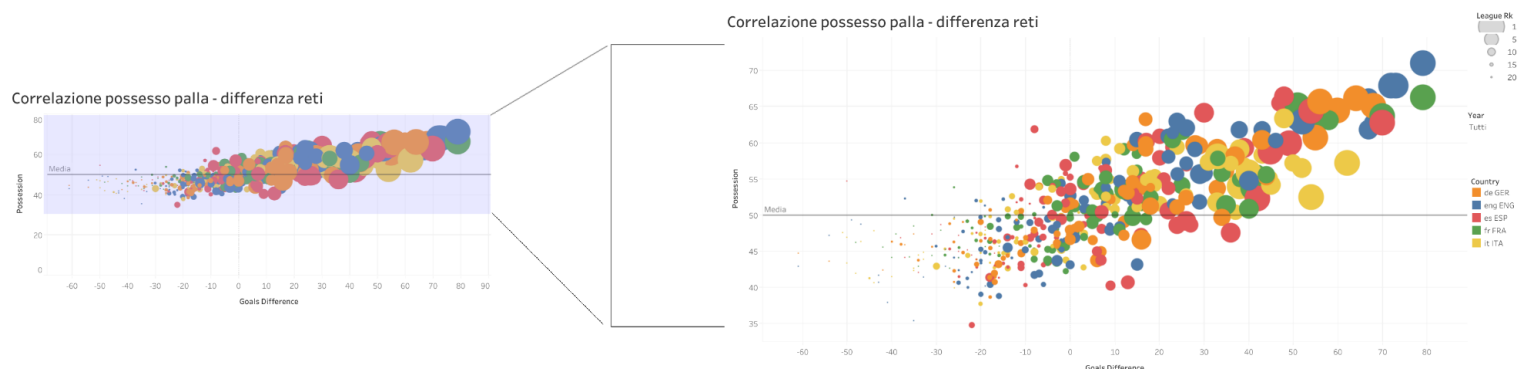
Sono scuole di pensiero contrapposte, che hanno portato a risultati positivi e risultati negativi in entrambi i casi, ma in generale c'è un trend che va maggiormente in una direzione ben definita, come si evince dalla seguente visualizzazione.

Abbiamo di seguito un confronto tra la differenza reti (variabile **Goals Difference**, sull'asse delle ascisse) e il dato relativo al possesso palla medio delle squadre in una singola stagione (variabile **Possession**, sull'asse delle ordinate); inoltre, viene

indicata per ogni squadra in tutte le stagioni la posizione finale in classifica (variabile **League Rk**) attraverso la dimensione, con cerchi più grandi che corrispondono alle posizioni alte in classifica mentre figure più piccole per le posizioni di bassa classifica.

La visualizzazione che consideriamo (quella a destra) è frutto di un troncamento dell'asse delle ordinate, avente valori compresi tra il 35 e l'80% circa di possesso palla: questo per non lasciarci ingannare da bias visivi legati al considerare solo una parte del grafico, quella veramente significativa.

Come per gli altri grafici, viene sempre indicata la nazione (variabile **Country**) come indicato in legenda, e viene data la possibilità, col filtro sulla sinistra (variabile **Year**) sia di avere la vista completa su tutte le cinque stagioni considerate, sia di visualizzare una stagione per volta.



3.3 Scatterplot Goals Difference - Possession: versione originale sulla sinistra, con asse delle ordinate troncato sulla destra per valori compresi tra 35 e 80% di possesso palla.

In generale si nota che al crescere dei valori della differenza reti, cresce anche il possesso palla medio delle squadre, abbiamo quindi una correlazione positiva in questo senso; inoltre, si nota che molte delle squadre in posizioni basse della classifica sono anche quelle con dati del possesso palla mediamente bassi e differenza reti negative, con solamente alcune eccezioni.

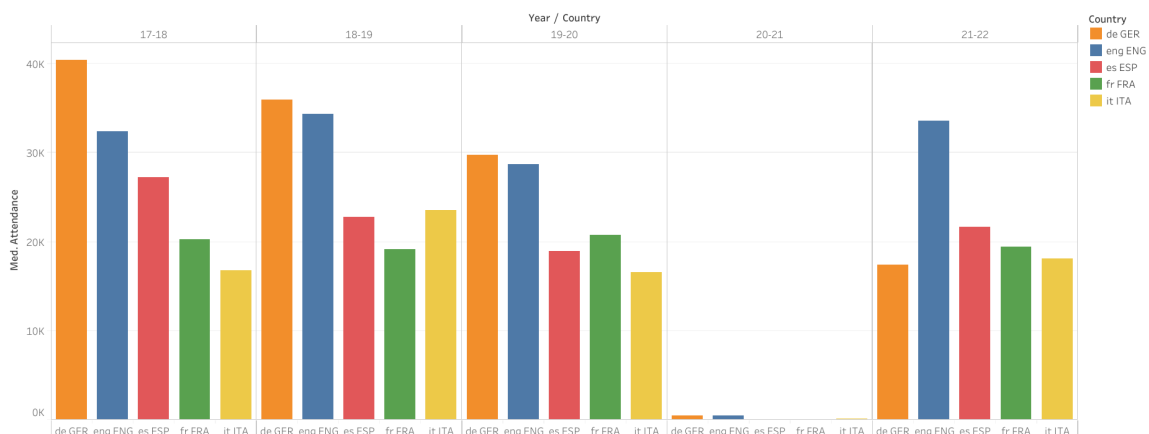
Allo stesso modo le squadre delle zone alte della classifica sono anche quelle con possesso palla mediamente alto (quasi tutti superiore al 50%, tranne alcune eccezioni come l'Atletico di Madrid, il Valencia, il Lipsia e lo Schalke 04 che hanno avuto piazzamenti importanti con dati di poco inferiori alla soglia del 50%) e con

differenza reti positiva e con valori elevati, fino ad arrivare ai picchi di rendimento di squadre come Bayern Monaco, Barcellona, PSG, Liverpool e Manchester City che hanno avuto i valori più elevati per entrambe le metriche.

3.5 Gli effetti del Covid-19

L'ultimo aspetto su cui abbiamo posto la nostra attenzione è stato l'andamento del tifo prima durante e dopo la pandemia che ci ha colpito nel 2020. Nel seguente grafico verranno illustrati gli spettatori medi (calcolato dalla media della variabile **Attendance**) presenti allo stadio per ciascuno dei 5 campionati (definito dalla variabile **Country** come in legenda) per ogni anno (variabile **Year**) calcistico presente nel nostro studio.

The Covid-effect



3.4 Bar Chart Spettatori - Grafico rappresentante degli spettatori per ciascuna nazione in tutti gli anni presenti nello studio.

Dalla visualizzazione 3.4 si può subito notare come la pandemia abbia avuto un effetto drastico sull'affluenza del pubblico allo stadio (come nel resto delle attività quotidiane). Infatti, nell'anno 2020/2021 non vi sono stati spettatori allo stadio (e ciò ha comportato grandi perdite nelle casse delle società), mentre si può anche notare che nell'anno 2021/2022 vi sono dei valori che mostrano differenze evidenti rispetto agli anni antecedenti al Covid: sembrerebbe infatti che la Premier League sia diventato il campionato con il maggior numero di spettatori allo stadio, ma tale dato è

distorto dalle normative che erano differenti per ciascun paese (infatti, il governo Inglese ha imposto meno limiti rispetto al resto dei paesi).

In generale, vista anche la capienza degli stadi, la Germania è la nazione che presenta più spettatori medi negli stadi, e subito dopo vi è l'Inghilterra, nota per la passione dei suoi tifosi. Invece, Spagna, Francia e Italia hanno un andamento molto simile, e presentano meno spettatori rispetto alle altre due nazioni.

4 Valutazione delle infografiche

4.1 Tecniche utilizzate

Una delle fasi più importanti della progettazione di una visualizzazione risiede nella valutazione della stessa da parte di utenti esterni. Questa permette di rendersi conto dell'efficacia dell'infografica, dal punto di vista della chiarezza, della leggibilità, della facilità nel trarre conclusioni interessanti e dell'utilità per l'utente finale, e di poter dunque ragionare relativamente a soluzioni da adottare per migliorarla e sviluppi futuri legati alla visualizzazione.

Per valutare la qualità della data visualization sono state utilizzate le seguenti tre tecniche:

- **La valutazione euristica:** una sorta di intervista, finalizzata ad individuare eventuali problemi presenti nelle nostre visualizzazioni, in cui viene richiesto ad utenti esterni di commentare liberamente una visualizzazione;
- **Il test utente:** con lo scopo di valutare l'efficacia - in termini di tempo impiegato dagli utenti - e l'efficienza - valutata tramite dei task assegnati e completati/non completati dagli utenti - della visualizzazione.
- **Il questionario psicometrico:** utile per valutare la qualità dell'infografica, ponendo una serie di domande relative ad essa che valutano fattori come utilità, chiarezza, bellezza, potere informativo e facilità di comprensione, oltre ad un giudizio generale;

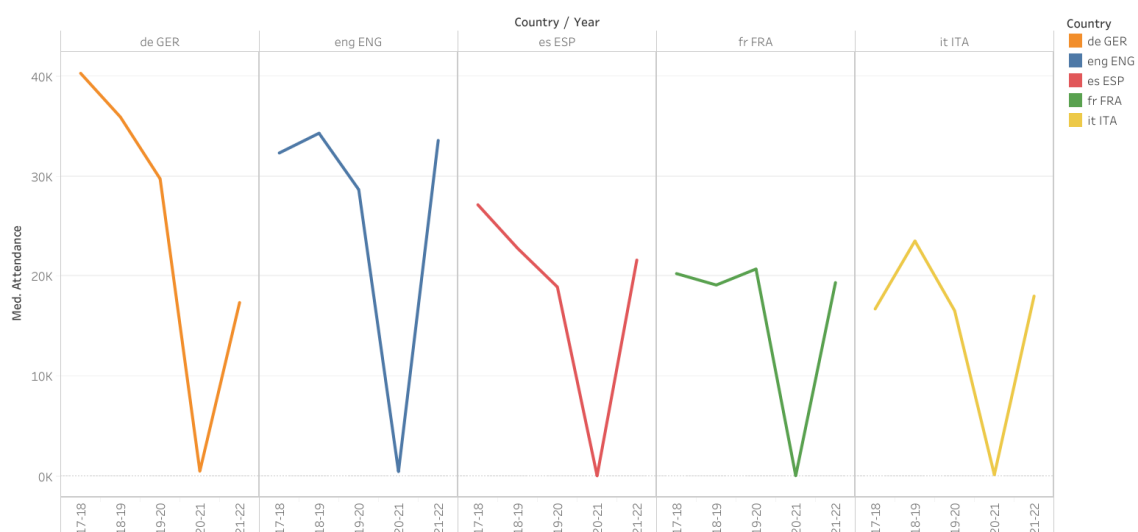
4.1.1 Valutazione euristica

La valutazione euristica è una procedura che consente di ricevere un feedback utile relativo ad una o più infografiche, da parte di utenti esterni: consiste nel presentare la visualizzazione ad un numero predefinito di utenti (nel nostro caso, abbiamo coinvolto 4 utenti) a cui abbiamo chiesto di commentare la data visualization interagendo con essa per qualche minuto.

Da questa fase, è possibile trarre indicazioni utili per risolvere eventuali problematiche relative alla nostra data visualization, dal punto di vista della chiarezza, dell'utilità, della facilità di navigazione. Queste sono le indicazioni tratte dai 4 utenti intervistati:

- Utente n°1: l'utente in generale ha mostrato interesse e soddisfazione per le infografiche mostrate; l'unica che l'ha convinto un po' meno è stata quella della correlazione possesso palla - differenza reti, dove avrebbe preferito avere meno informazioni contemporaneamente per renderla di più semplice interpretazione.
- Utente n°2: l'Utente 2, relativamente all'infografica su prestazioni difensive e offensive, ha chiesto se fosse possibile avere entrambi i grafici sulla schermata contemporaneamente, per poter fare confronti più rapidi; tuttavia, mostrata questa soluzione (che avevamo scartato poichè i grafici risultavano molto piccoli e dunque difficili da leggere), ha confermato la nostra idea iniziale.
- Utente n°3: l'utente 3 ha suggerito una modifica del grafico relativo agli spettatori, con una distinzione non per anno ma bensì per nazione (osservando di seguito tutte le stagioni consecutive per ogni singola nazione); tuttavia, dopo avergli presentato il grafico come suggerito da lui, ha affermato di preferire ancora la nostra visualizzazione originale, definendola più chiara per confrontare le singole stagioni. Relativamente alle altre visualizzazioni non ha avuto niente da aggiungere.

The Covid-effect



4.1 Modifica al grafico della media di spettatori, suddiviso in base alla nazione e non in base alla stagione come in precedenza.

- Utente n°4: L'utente 4 ha suggerito l'utilizzo di grafici con meno informazioni, ma più dettagliate. In particolare, suggeriva la visione di grafici contenenti solo poche squadre ma con più informazioni, come il numero di passaggi riusciti e il numero di dribbling riusciti. In linea generale, ha comunque ritenuto esaustive le visualizzazioni.

4.1.2 Test utente

La seconda fase di valutazione, ovvero il test utente, ci ha permesso di approfondire con gli utenti della visualizzazione relativamente ad eventuali problematiche riscontrate a livello di usabilità e comprensione delle infografiche.

In questa fase, dove abbiamo coinvolto 12 soggetti, sono stati assegnati 3 task (abbastanza complessi, non di risoluzione immediata) ad ognuno di loro; per ognuno di essi abbiamo misurato il tempo di esecuzione di ogni compito, tenendo ovviamente conto del numero di task portati a termine da ognuno.

Per il test utente, sono stati richiesti agli utenti i seguenti task:

- Individuare le prime squadre classificate per ciascun campionato nella stagione 2018/19 nella visualizzazione "Confronto expected goals - expected goals against" (3.1)
- Osservando l'infografica "Precisione difensiva", quali sono alcune delle squadre che, nonostante siano arrivate tra le ultime posizioni in classifica, hanno comunque avuto un rendimento buono relativamente agli xGA e i gol effettivamente subiti?
- Sempre osservando l'infografica "Precisione difensiva, precisione offensiva", come interpretate i valori pari a 0 per le differenze (Goals Against - xGA) e (Goals Scored - xG)?

Queste, invece, sono le indicazioni in termini di tempo che abbiamo ottenuto cronometrando gli utenti durante l'esecuzione dei task:

Utente	Task 1	Task 2	Task 3
1	68	87	45
2	78	68	31
3	70	69	26
4	45	73	38
5	42	58	61
6	77	88	43
7	56	90	33
8	57	43	67
9	47	35	45
10	60	54	24
11	83	74	69
12	49	67	37

Tabella 1: Tempo trascorso da ciascun utente su ciascun task ed eventuale successo/insuccesso.

	Task 1	Task 2	Task 3	Totale
Media	61.0	67.2	43.3	57.2
Mediana	58.5	68.5	40.5	55.8

Tabella 2: Media e mediana del tempo trascorso su ciascun task. .

Questi sono i risultati ottenuti, osserviamo che la task che ha richiesto più tempo (che si è rivelata essere anche la più complessa) è la seconda, con un tempo medio di 67.2 secondi; al contrario la terza è stata la più veloce, con una media di 43.3 secondi per utente. Possiamo osservare queste informazioni anche dal grafico seguente, che contiene sia il boxplot che lo scatterplot dei differenti task.

Inoltre, si può notare dalla tabella che non tutti i task sono stati completati con successo senza un nostro suggerimento (infatti, quelli colorati di verde sono stati completati in autonomia, mentre quelli rossi hanno richiesto un nostro intervento), e in particolare abbiamo avuto percentuali di successo più elevate per il primo e il terzo task (75% di tasso di completamento in autonomia), mentre una situazione di quasi equilibrio per il secondo task (7 successi, 5 necessità di intervento)

	Task 1	Task 2	Task 3	Totale
Successo	9	7	9	25
Aiuto necessario	3	5	3	11

Tabella 3 Conteggio successi ed insuccessi del test utente per ciascun task.

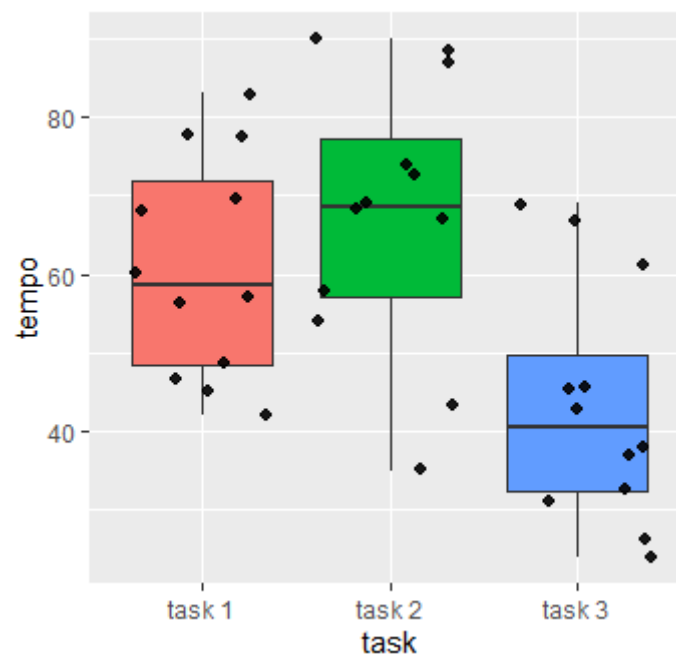


Figura 4.2: Box plot relativi al tempo trascorso su ciascun task..

4.1.3 Questionario psicometrico

Dopo aver sottoposto alcuni utenti alla valutazione euristica e al test utente, siamo arrivati alla terza fase, ovvero il questionario psicometrico, col fine di valutare la qualità della data visualization.

Abbiamo deciso di utilizzando la Cabitza-Locoro Scale: l'utente può valutare la qualità della visualizzazione attribuendo un valore che può andare da 1 (pochissimo) a 6 (moltissimo) relativamente a diversi fattori, ovvero:

- Utilità dell'infografica
- Chiarezza dell'infografica
- Informatività dell'infografica
- Bellezza dell'infografica
- Intuitività dell'infografica
- Valore complessivo dell'infografica

A questa fase di compilazione del questionario hanno preso parte 24 soggetti, che hanno risposto nel seguente modo:

Voti	Utile	Chiara	Informativ a	Bella	Intuitiva	Valutazion e generale
1	2	0	0	0	0	0
2	4	2	2	3	5	3
3	6	7	5	3	2	4
4	5	9	8	4	7	3
5	2	3	5	4	7	7
6	5	3	4	10	3	7

Tabella 4 Conteggio voti per il test psicometrico.

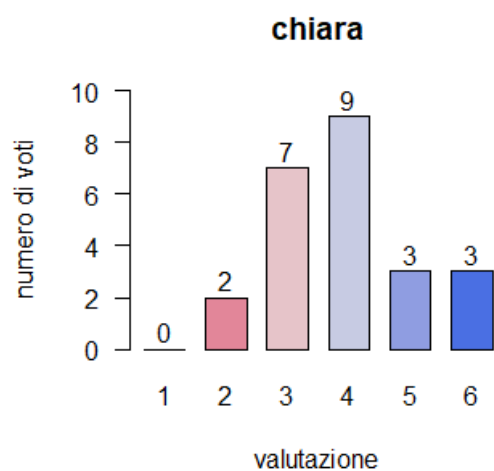
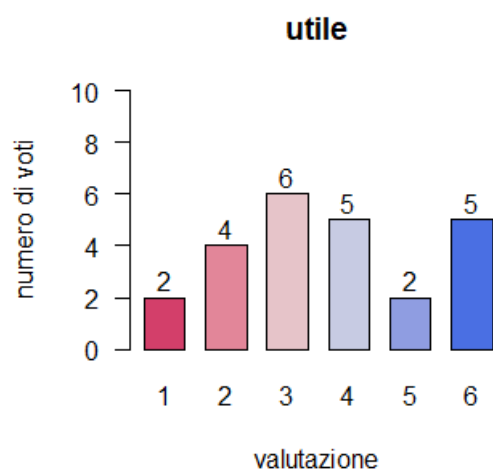


Figura 4.3: Barchart per il conteggio dei voti per “utile”. **Figura 4.4:** Barchart per il conteggio dei voti per “chiara”.

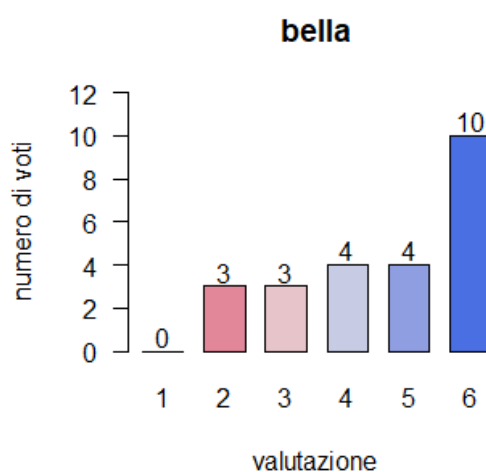
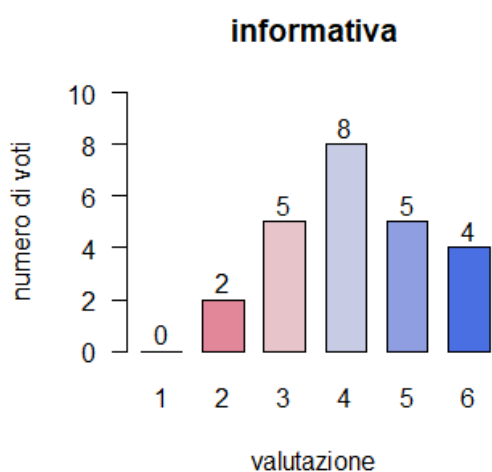


Figura 4.5: Barchart del conteggio voti per “informativa”. **Figura 4.6:** Barchart del conteggio voti per “Bella”.

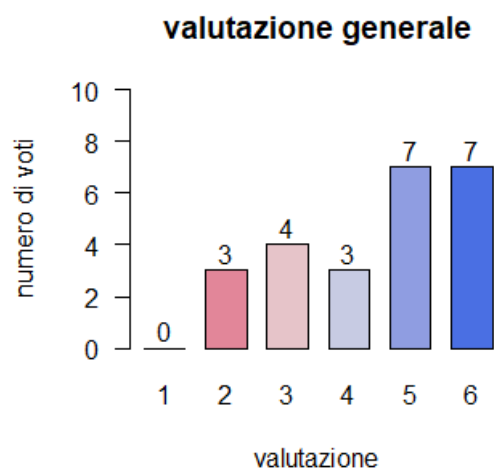
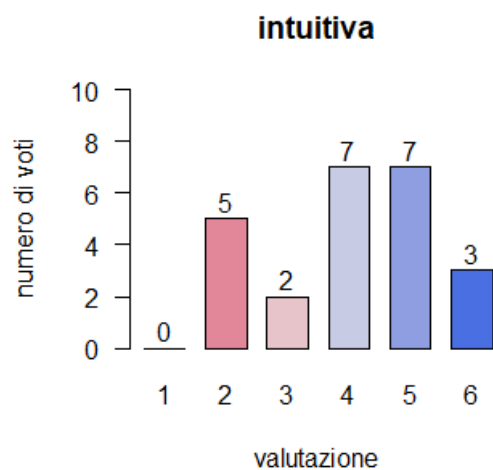


Figura 4.7:Barchart per il conteggio dei voti per “Intuitiva”. **Figura 4.8:**Barchart per il conteggio dei voti per “VG”.

Osserviamo dalla tabella, e dai grafici a barre relativi ai singoli campi, che abbiamo avuto voti generalmente medio/alti. In particolare, solo alla voce “utile” abbiamo ricevuto due valutazioni col voto minimo, mentre per tutte le altre abbiamo ottenuto voti da 2 punti a crescere.

Inoltre, nelle valutazioni generali delle infografiche, abbiamo ottenuto voti mediamente elevati, con un 58% di voti tra il 5 e il 6.

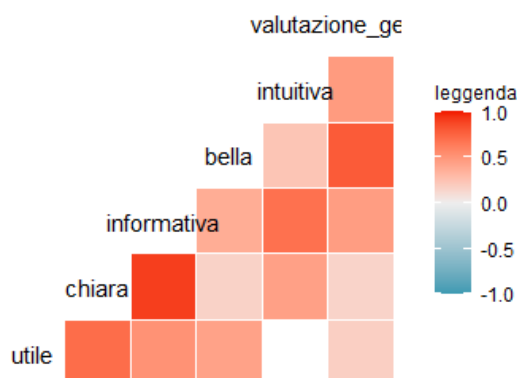


Figura 4.9: Matrice di correlazione tra i conteggi delle risposte..

Infine, osservando la matrice di correlazione, si osserva una forte correlazione positiva tra gli aggettivi “chiara” e “informativa” (0.91), tra “bella” e la valutazione generale” (0.79), ed una buona correlazione tra “chiara” e “utile” (0.71) e tra “intuitiva” e “informativa” (0.69).

Le altre correlazioni hanno valori minori, fino ad arrivare a “utile” e “intuitiva” in cui la correlazione è praticamente nulla.

5. Sviluppi futuri e conclusioni

L'obiettivo di questo report era di porre l'attenzione sull'importanza che sempre più hanno i dati e i numeri nel mondo dello sport, e in particolare nel calcio. Sempre più squadre e allenatori si affidano ad essi per poter ambire ai trofei più prestigiosi, e in uno sport dove il calcio, che ha sofferto (come tutti i settori) con l'arrivo del Covid ma che attira sempre un interesse incredibile da tutto il mondo, è normale pensare che si faranno sempre più utilizzi di questi mezzi.

Pensando anche solo allo scouting, dove arrivare uno o due anni prima ad acquistare le future star dello sport può portare a benefici economici e di rendimento pazzeschi, ai metodi per monitorare la qualità degli allenamenti e il recupero dagli infortuni, ai dati relativi ai giocatori migliori da acquistare in un determinato contesto, all'utilizzo della tecnologia legato a innovazioni come il Var o la Goal Line Technology. Sono tutti contesti dove i dati possono fare la differenza.

E anche a noi, semplici appassionati dello sport, piace spesso analizzare aspetti simili, come abbiamo provato a fare all'interno di questo report, e sappiamo che ci possono essere aspetti che non abbiamo ancora toccato con questo lavoro: come un'analisi della fase di pressing, sempre più fondamentale nel contesto tattico delle squadre; un'analisi più approfondita relativa a quanto hanno veramente perso le società col Covid, tra sponsor, diritti tv e stadi chiusi; o l'integrazione dei dataset già utilizzati con altre informazioni utili ed interessanti, che possono essere significative per il successo o no di una determinata squadra (come l'età media dei giocatori, o il valore di mercato medio, o il budget degli stipendi delle singole società); o, infine, anche solo verificare se i trend che abbiamo avuto negli ultimi anni si confermeranno anche per le prossime stagioni.

6. Fonti

Slides lezioni del corso di Data Visualization (a.a. 2022/23)

<https://coolors.co/>

<https://understat.com/>

<https://footballxg.com/>

<https://fbref.com/en/>

https://it.wikipedia.org/wiki/Pagina_principale

<https://www.statsperform.com/opta-analytics/>

<https://theanalyst.com/eu/2021/07/what-are-expected-goals-xg/>