Università Politecnica delle Marche

Facoltà di Ingegneria

Dipartimento di Ingegneria dell'Informazione


# Football Analysis mediante i tools Qlik, Tableau e PowerBI

**Docenti:**
Domenico Ursino
Luca Virgili

**Studenti:**
Lorenzo Fratini
Federico Miscia
Andrea Pinciaroli

**Anno Accademico 2021/2022**

# Indice

# Capitolo 1

# Introduzione

Il seguente elaborato prende in considerazione un dataset contenente i dati relativi ai top 5 campionati europei di calcio (Premier League, Serie A, Bundesliga, La Liga e Ligue 1) dalla stagione 2014/2015 alla stagione 2020/2021. Tale dataset è reperibile al seguente link:

`https://www.kaggle.com/technika148/football-database`

L'obiettivo del progetto è quello di acquisire conoscenza dai dati presi in considerazione. Nello specifico, l'intento è quello di svolgere un'analisi sotto tre principali punti di vista, con un livello di specificità crescente: i campionati, le singole squadre e i giocatori. Per fare ciò, si è fatto uso dei tre principali tool leader di mercato per la Data Analytics e Business Intelligence, secondo quanto riportato dal Magic Quadrant di Gartner come viene mostrato in Figura 1.1.
I software utilizzati sono: **Qlik**, **Tableau** e **Power BI**.


Figura 1.1: Magic Quadrant Gartner

## 1.1 ETL

Nella fase preliminare del progetto è stato realizzato il processo di Extract-Transform-Load (ETL) sul dataset scelto, al fine di ripulirlo, risolvere eventuali incoerenze rilevate ed, in generale, prepararlo per le successive analisi tramite i tool di Business Intelligence.

### 1.1.1 Dataset

Come già accennato in precedenza, il dataset di riferimento raccoglie i dati relativi ai top 5 campionati europei di calcio (Premier League, Serie A, Bundesliga, La Liga e Ligue 1) dalla stagione 2014/2015 alla stagione 2020/2021. Tale dataset, nella sua versione originaria, è composto da sette file in formato CSV che sono:

- **appearances.csv**: contiene tutte le presenze che i giocatori hanno fatto nei vari campionati europei
- **games.csv**: contiene tutte le partite che sono state giocate
- **leagues.csv**: contiene i top 5 campionati europei
- **players.csv**: contiene tutti i giocatori che hanno giocato almeno una volta nei vari campionati europei
- **shots.csv**: contiene tutti i tiri effettuati dai giocatori nei vari campionati europei
- **team.csv**: contiene tutti i team che hanno partecipato almeno una volta ai vari campionati europei
- **teamstats.csv**: contiene tutte le statistiche dei vari team nelle partite che hanno giocato

Di seguito sono riportate le varie tabelle con gli attributi che caratterizzano ciascuna di esse e la relativa descrizione, questo per fornire una migliore comprensione di quelle che saranno poi gli attributi utilizzati nelle successive fasi di analisi.

| **Leagues** | |
|---|---|
| **Attributo** | **Descrizione** |
| leagueID | ID del campionato |
| name | Nome esteso del campionato |

Tabella 1.1: leagues.csv

| Teams | |
|---|---|
| **Attributo** | **Descrizione** |
| teamID | ID del team |
| name | Nome esteso del team |

Tabella 1.2: team.csv

| Players | |
|---|---|
| **Attributo** | **Descrizione** |
| playerID | ID del giocatore |
| name | Nome esteso del giocatore |

Tabella 1.3: players.csv

| Games | |
|---|---|
| **Attributo** | **Descrizione** |
| gameID | ID del match |
| leagueID | ID del campionato |
| season | Stagione calcistica |
| date | Data in cui si è tenuto il match |
| homeTeamID | ID del team che gioca in casa |
| awayTeamID | ID del team che gioca fuori casa |
| homeGoals | Goals fatti dalla squadra in casa |
| awayGoals | Goal fatti dalla squadra ospite |
| homeProbability | Probabilità di vincita della squadra in casa |
| drawProbability | Probabilità di pareggio |
| awayProbability | Probabilità di vincita della squadra ospite |
| homeGoalsHalfTime | Goal fatti dalla squadra in casa nel primo tempo |
| awayGoalsHalfTime | Goal fatti dalla squadra in casa nel secondo tempo |

Tabella 1.4: games.csv

| TeamStats | |
|---|---|
| **Attributo** | **Descrizione** |
| gameID | ID del match |
| teamID | ID del team |
| season | Stagione calcistica |
| date | Data in cui si è tenuto il match |
| location | Luogo in cui il team ha giocato la partita (casa o fuori casa) |
| goals | Numero di goal segnati dal team |
| xGoals | Numero di goal attesi dal team nella partita |
| shots | Numero di tiri effettuati dal team nella partita |
| shotsOnTarget | Numero di tiri effettuati nella porta avversaria |
| deep | Numero di passaggi completati negli ultimi 18 metri di campo |
| ppda | Rapporto fra il numero di passaggi effettuati dalla squadra che imposta e il numero di azioni difensive (tackle, intercetti, falli) compiute dalla squadra che aggredisce senza palla. |
| fouls | Numero di falli commessi dalla squadra |
| corners | Numero di corner della squadra |
| yellowCards | Numero di cartellini gialli ottenuti dal team nella partita |
| redCards | Numero di cartellini rossi ottenuti dal team nella partita |
| result | Risultato della partita per il team (W: vittoria, D: pareggio, L: sconfitta) |

Tabella 1.5: teamstats.csv

| Appearences | |
|---|---|
| **Attributo** | **Descrizione** |
| gameID | ID del match |
| playerID | ID del giocatore |
| goals | Goal realizzati dal giocatore nel match |
| ownGoals | Autogol realizzati dal giocatore nel match |
| shots | Tiri effettuati dal giocatore nel match |
| xGoals | Goal previsti per il giocatore nel match |
| assists | Assist realizzati dal giocatore nel match |
| keyPasses | Passaggi chiave realizzati dal giocatore nel match |
| xAssists | Assist previsti per il giocatore nel match |
| position | Ruolo del giocatore |
| yellowCard | Numero di cartellini gialli ricevuti dal giocatore nella partita |
| redCard | Numero (max. 1) di cartellini rossi ricevuti dal giocatore nella partita |
| time | Minuti giocati |
| substituteIn | Minuto in cui il giocatore ha fatto il proprio ingresso in campo |
| substituteOut | Minuto in cui il giocatore è stato sostituto (0 se non è stato sostituito) |
| leagueID | ID del campionato |

Tabella 1.6: appearences.csv

| Shots | |
|---|---|
| **Attributo** | **Descrizione** |
| gameID | ID del match in cui è avvenuto il tiro |
| shooterID | ID del giocatore che ha effettuato il tiro |
| assisterID | ID del giocatore che ha effettuato l'assist per il tiro |
| minute | Minuto in cui è avvenuto il tiro |
| situation | Situazione di gioco in cui è avvenuto il tiro |
| lastAction | Ultima azione prima del tiro |
| shotType | Modalità con cui è avvenuto il tiro (piede dx/sx, di testa) |
| shotResult | Esito del tiro (Goal, Bloccato, Deviato, Terminato fuori ...) |
| xGoal | Probabilità di segnare |
| positionX | Posizione in cui è stato effettuato il tiro in relazione al lato lungo del campo |
| positionY | Posizione in cui è stato effettuato il tiro in relazione al lato corto del campo |

Tabella 1.7: shots.csv

### 1.1.2 Operazioni eseguite

Nel dataset appena presentato sono stati preventivamente rimossi degli attributi ritenuti non di interesse per le analisi desiderate, quali ad esempio tutta una serie di indicatori relativi alle quote di betting per le partite. In seguito, a causa della struttura del dataset,

è emerso un problema legato all'impossibilità di ricostruire la rosa delle squadre nel corso delle stagioni, aspetto che avrebbe limitato notevolmente le possibilità di analisi. Alla luce di ciò, la soluzione adottata ha previsto l'accorpamento delle tabelle *Players* e *Teams* all'interno di *Appearences* in modo tale che le informazioni relative ad ogni presenza di un giocatore fossero complete del nome del giocatore e della sua appartenenza ad una certa squadra. Dopodiché, a causa del fatto che un giocatore può aver cambiato squadra nel corso degli anni, per tenere traccia di ciò, è stato inserito all'interno della medesima tabella l'attributo relativo alla stagione in modo da identificare univocamente la presenza di un giocatore in relazione alla squadra di appartenenza in una determinata annata ed evitare, di fatto, righe duplicate. In ultima istanza, all'interno delle tabelle *Appearences* e *Teamstats*, è stato creato un attributo concatenando 'gameID' e 'teamID', con lo scopo di agevolare la relazione tra le due. Tali manipolazioni sono state svolte in *Python*, per mezzo della libreria *Pandas*. Di seguito è riportata la tabella *Appearences* risultante:

| **Appearences** | |
|---|---|
| **Attributo** | **Descrizione** |
| gameID - teamID | Attributo creato in qualità di chiave esterna |
| gameID | ID del match |
| playerID | ID del giocatore |
| goals | Goal realizzati dal giocatore nel match |
| ownGoals | Autogol realizzati dal giocatore nel match |
| shots | Tiri effettuati dal giocatore nel match |
| xGoals | Goal previsti per il giocatore nel match |
| assists | Assist realizzati dal giocatore nel match |
| keyPasses | Passaggi chiave realizzati dal giocatore nel match |
| xAssists | Assist previsti per il giocatore nel match |
| position | Ruolo del giocatore |
| yellowCard | Numero di cartellini gialli ricevuti dal giocatore nella partita |
| redCard | Numero (max. 1) di cartellini rossi ricevuti dal giocatore nella partita |
| time | Minuti giocati |
| substituteIn | Minuto in cui il giocatore ha fatto il proprio ingresso in campo |
| substituteOut | Minuto in cui il giocatore è stato sostituto (0 se non è stato sostituito) |
| leagueID | ID del campionato |
| season | Stagione calcistica in cui è registrata la presenza |
| teamID | ID della squadra cui il giocatore fa riferimento |
| teamName | Nome esteso della squadra |
| playerName | Nome esteso del giocatore |

Tabella 1.8: Nuova versione di appearences.csv

# Capitolo 2

# Qlik

Qlik è un software di visualizzazione e business intelligence che permette uno sviluppo rapido di dashboard completamente personalizzabili attraverso cui è possibile mostrare informazioni utili sui dati a disposizione. Qlik permette di svolgere prettamente analisi di tipo descrittivo e diagnostico: obiettivo di una prima fase di analisi in qualsiasi campagna di Data Science, infatti, è quello di presentare i dati in maniera intuitiva e facilmente comprensibile. Una caratteristica molto interessante, inoltre, è la possibilità di cambiare dinamicamente le visualizzazioni sulla base dei filtri e delle selezioni effettuate dagli utenti. La versione utilizzata per il progetto è Qlik Sense

Figura 2.1: Logo Qlik

## 2.1 Caricamento dati

Creata una nuova applicazione, sono state caricate le tabelle tramite l'interfaccia grafica che Qlik mette a disposizione. Al termine del caricamento, il software ha riconosciuto la presenza dei campi comuni fra le diverse tabelle e questo ha semplificato la fase di creazione delle associazioni. Inoltre, per favorire la leggibilità del modello sono state eseguite delle operazioni di join fra la tabella *games* e *league* (tramite il campo *leagueID*), fra *apperances* e *team* (tramite il campo *teamID* e fra ciò che si ottiene e la tabella *player* (tramite il campo *player*). Il modello che è stato ottenuto è mostrato in Figura 2.2


Figura 2.2: Modello Qlik

## 2.2 Panoramica Generale

### 2.2.1 Overview Campionato

La prima analisi effettuata sul dataset a disposizione riguarda l'estrazione della classifica finale di un campionato, ottenuta mediante la costruzione di una tabella contenente, per ogni squadra in competizione per il titolo, un riepilogo riguardo le partite giocate (M), le vittore (W), i pareggi (D), le sconfitte (L), i goal fatti (GF), i goal subiti (GA), la differenza reti (GD) e i punti.
Nella Figura 2.3 è visibile una dashboard contenente la classifica del campionato di Serie A nella stagione 2020/2021.

## 2.2. PANORAMICA GENERALE

**Serie A - 2020**

| n° | Team | M | W | D | L | GF | GA | GD | Pts |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Inter | 38 | 28 | 7 | 3 | 89 | 35 | 54 | 91 |
| 2 | AC Milan | 38 | 24 | 7 | 7 | 74 | 41 | 33 | 79 |
| 3 | Atalanta | 38 | 23 | 9 | 6 | 90 | 47 | 43 | 78 |
| | Juventus | 38 | 23 | 9 | 6 | 77 | 38 | 39 | 78 |
| 5 | Napoli | 38 | 24 | 5 | 9 | 86 | 41 | 45 | 77 |
| 6 | Lazio | 38 | 21 | 5 | 12 | 61 | 55 | 6 | 68 |
| 7 | Roma | 38 | 18 | 9 | 11 | 68 | 55 | 13 | 63 |
| 8 | Sassuolo | 38 | 17 | 11 | 10 | 64 | 56 | 8 | 62 |
| 9 | Sampdoria | 38 | 15 | 7 | 16 | 52 | 54 | -2 | 52 |
| 10 | Verona | 38 | 10 | 13 | 15 | 43 | 48 | -5 | 43 |
| 11 | Genoa | 38 | 10 | 12 | 16 | 47 | 58 | -11 | 42 |
| 12 | Bologna | 38 | 10 | 11 | 17 | 51 | 65 | -14 | 41 |
| 13 | Fiorentina | 38 | 9 | 13 | 16 | 47 | 59 | -12 | 40 |
| | Udinese | 38 | 10 | 10 | 18 | 42 | 58 | -16 | 40 |
| 15 | Spezia | 38 | 9 | 12 | 17 | 52 | 72 | -20 | 39 |
| 16 | Cagliari | 38 | 9 | 10 | 19 | 43 | 59 | -16 | 37 |
| | Torino | 38 | 7 | 16 | 15 | 50 | 69 | -19 | 37 |
| 18 | Benevento | 38 | 7 | 12 | 19 | 40 | 75 | -35 | 33 |
| 19 | Crotone | 38 | 6 | 5 | 27 | 45 | 92 | -47 | 23 |
| 20 | Parma Calcio 1913 | 38 | 3 | 11 | 24 | 39 | 83 | -44 | 20 |

Figura 2.3: Classifica Serie A 2020/2021

Sempre nell'ambito di una panoramica generale inerente al campionato di Serie A, è stata creata un'altra dashboard contenente le informazioni circa i goal segnati dai top 10 giocatori più prolifici e gli assist effettuati dai 10 migliori assistman nella stagione 2020/2021. Il tutto è accompagnato nella parte destra della dashboard, da uno scatter plot che permette di visualizzare chi sono i giocatori che hanno reso meglio rispetto le attese previste. La dashboard discussa è riportata in Figura 2.4.

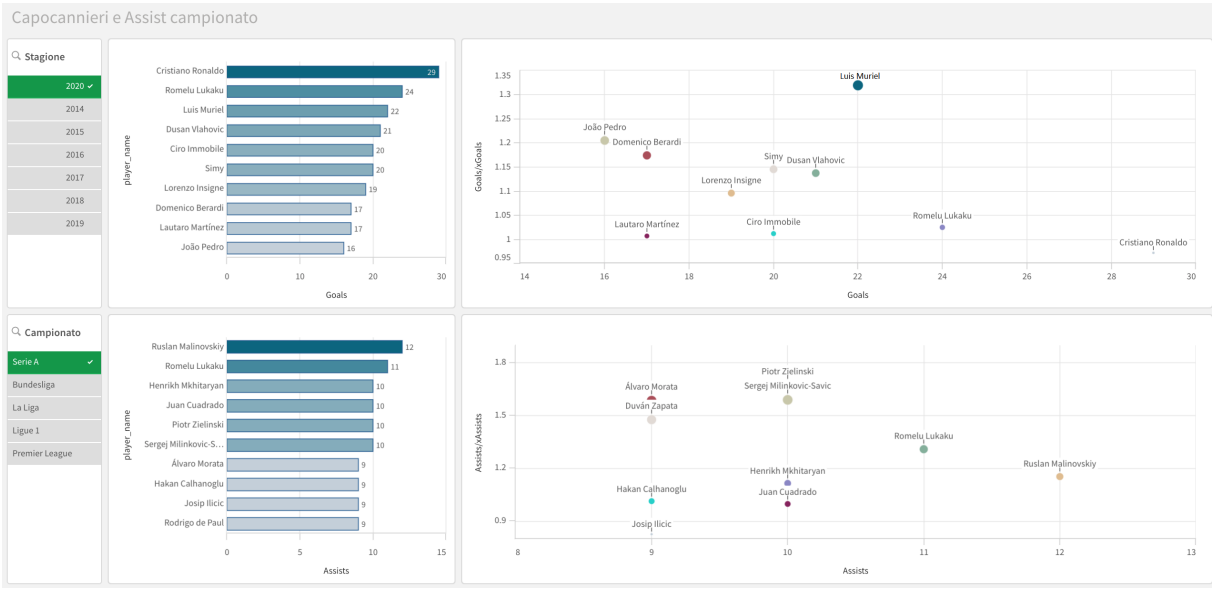


Figura 2.4: Dashboard giocatori Serie A 2020/2021

Da tale dashboard si osserva che più il giocatore si trova nella parte in alto dello scatterplot e più il rapporto tra Goals e xGoals è alto (rispettivamente Assists e xAssists) e ciò significa che il giocatore ha reso meglio rispetto quelle che erano le sue attese, ovvero si può considerare come una sorpresa. Il caso del giocatore Luis Muriel è abbastanza esplicativo, in quanto ha segnato 22 reti nella stagione 2020/2021 contro i 17 goal attesi; un altro elemento estremo è il giocatore Cristiano Ronaldo il quale ha siglato 29 reti e si trova nella parte bassa dello scatter plot, in quanto ha rispecchiato quelli che erano i gol attesi e quindi ha confermato la previsione. Invece, guardando lo scatterplot relativo agli assist si può notare come il numero di giocatori che hanno fatto meglio delle attese sia più alto confrontandolo con quello dei marcatori.

### 2.2.2 Overview Team

La seconda analisi effettuata ha focalizzato l'attenzione su un singolo team di un campionato. Ciò ha permesso non solo di confermare quanto già analizzato precedentemente, ma ha dato anche la possibilità di far emergere delle statistiche più dettagliate.
Nella Figura 2.5 viene mostrata la dashboard d'interesse considerando come esempio la squadra Juventus nella stagione 2020/2021.

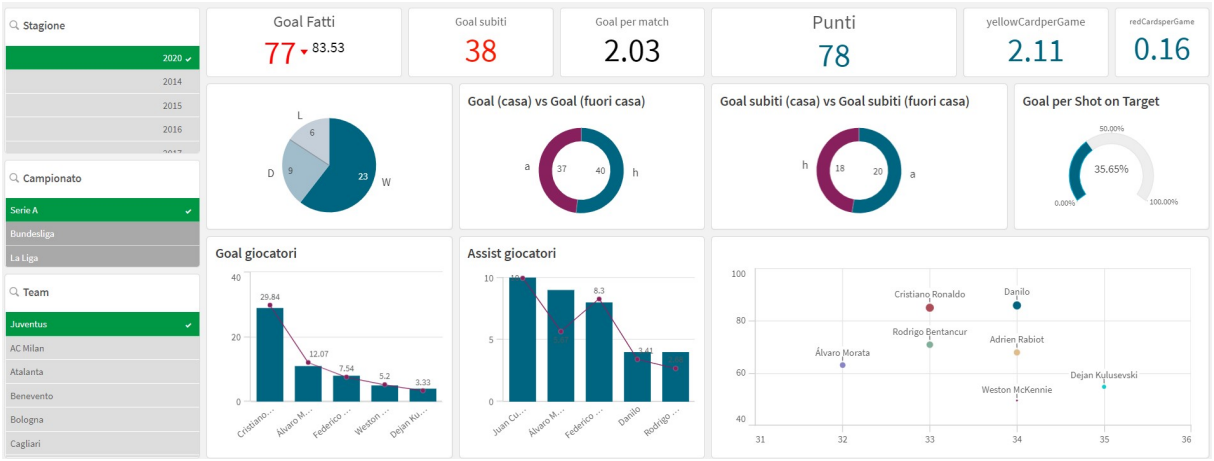


Figura 2.5: Dashboard Juventus 2020/2021

Come si può notare, la parte sinistra della dashboard comprende una serie di filtri che permettono di selezionare la stagione, il campionato e il team di interesse. Nella parte destra si possono notare tre sezioni. La prima, quella più in alto, contiene una serie di KPI che riportano, rispettivamente, i goal fatti (il cui colore è in relazione ai goal attesi), i goal subiti, i goal per partita, i punti nella stagione in questione, i cartellini gialli e rossi per partita. In particolar modo, riguardo il primo KPI si osserva che è stato accompagnato anche da un'indicazione di quelli che sono i goal attesi dalla squadra, questo per permettere di capire fin da subito se sono state soddisfatte le previsioni circa i goal segnati dal team.
La fascia centrale della dashboard contiene tre grafici a torta che riportano, rispettivamente, la distribuzione delle vittore e delle sconfitte, i goal fatti/subiti in casa/fuori casa e un indice di performance per mostrare la percentuale dei goal fatti in funzione dei tiri in porta effettuati dal team.

Infine, nella sezione più in basso sono stati riportati dei dati più focalizzati sui giocatori del team. I due diagrammi a barre riportano i primi 5 giocatori con più goal e più assist confrontati con quelli attesi. Da questo si osserva che il giocatore Alvaro Morata ha effettuato molti più assist rispetto le attese, mentre i restanti giocatori hanno rispecchiato più o meno le previsioni. Invece, il grafico più a destra riporta uno scatterplot dove si riportano i giocatori in funzione delle partite giocatori e dei minuti per partita. Per una maggiore comprensione, tale grafico viene riportato singolarmente nella Figura 2.6.

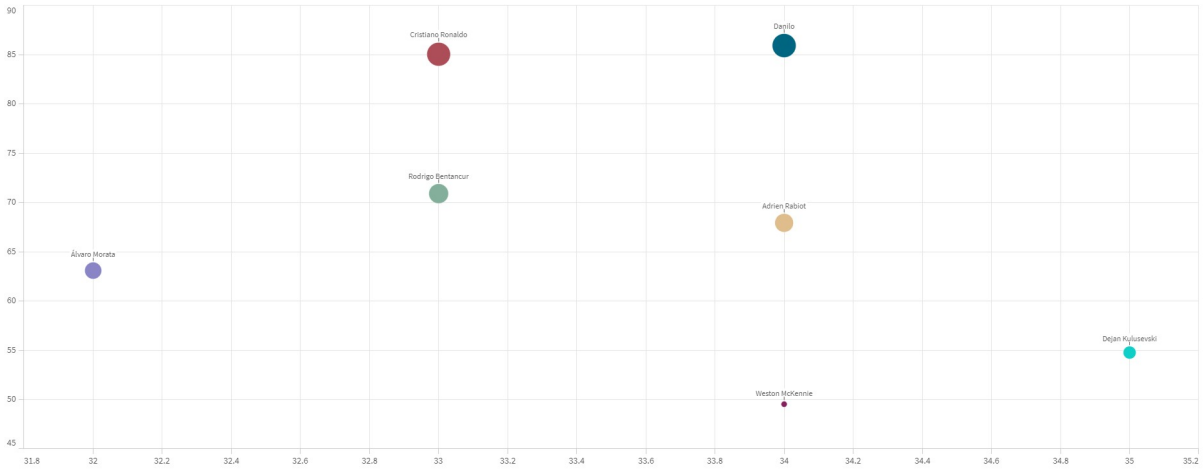


Figura 2.6: Scatterplot giocatori Juventus

Grazie a questo grafico emergono quali sono stati i giocatori più utilizzati dal team selezionato e anche quanto hanno giocato, infatti la grandezza della bolla indica la media dei minuti per ogni partita. Nell'esempio in questione questo significa che i giocatori Cristiano Ronaldo e Danilo sono stati quelli con un minutaggio maggiore, indice della loro centralità nel team, mentre Dejan Kulusevski è stato quello maggiormente utilizzato (con 35 partite), ma con un minutaggio leggermente inferiore, indice del fatto che spesso è subentrato a partita in corso oppure è stato sostituito.

# Capitolo 3

# Tableau

Tableau è uno dei software di Buisness Intelligence tra i più affermati ed utilizzati nel mercato, infatti si trova subito dietro a Power BI nel Magic Quadrant di Gartner. Tableau può essere considerato come un tool di analisi visiva che facilita l'utente nella gestione ed esplorazione dei dati dai quali estrarre e condividere conoscenza. Tali informazioni ricavate possono essere poi evidenziate in maniera efficace tramite la forte componente di rappresentazione visiva che Tableau mette a disposizione. Gli utenti, infatti, possono, creare e condividere dashboard interattive contenenti gli insight o i trend ricavati dall'analisi dei dati. Ulteriore punto di forza di Tableau è dato dalla possibilità di connettersi a dataset di diverse tipologie (file di testo, file excel, file JSON, etc.) e a diverse sorgenti, siano queste da macchina locale piuttosto che dal cloud.

Figura 3.1: Logo Tabeau

## 3.1 Caricamento dati

In maniera analoga a quanto fatto con Qlik, anche con Tableau la prima fase è stata quella di caricamento dei dati mediante l'interfaccia messa a disposizione dal tool stesso. In questo caso il modello dati ottenuto è rappresentato in Figura 3.2

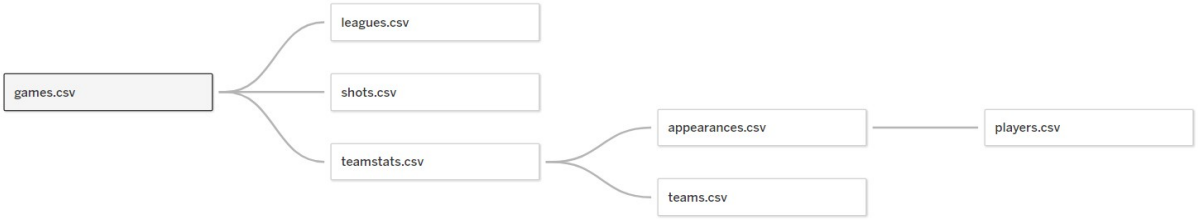


Figura 3.2: Modello Tabeau

## 3.2 Analisi dei campionati

In maniera simile a Qlik, anche con Tableau si è partiti dall'analisi dei campionati. Anche in questo caso ci si è soffermati sull'analisi di una singola competizione, nell'esempio la Serie A, tuttavia senza fare riferimento ad una specifica stagione, come avvenuto per Qlik, ma si è considerato l'intero arco temporale dalla stagione 2014/2015 a quella del 2020/2021.
L'obiettivo della prima analisi effettuata è stato quello di mostrare dei trend e delle previsioni, nonchè estrarre informazioni d'interesse circa le performance dei vari team e giocatori nelle differenti stagioni.
In Figura 3.3 è mostrata la dashboard che contiene tutte le informazioni sopra citate.

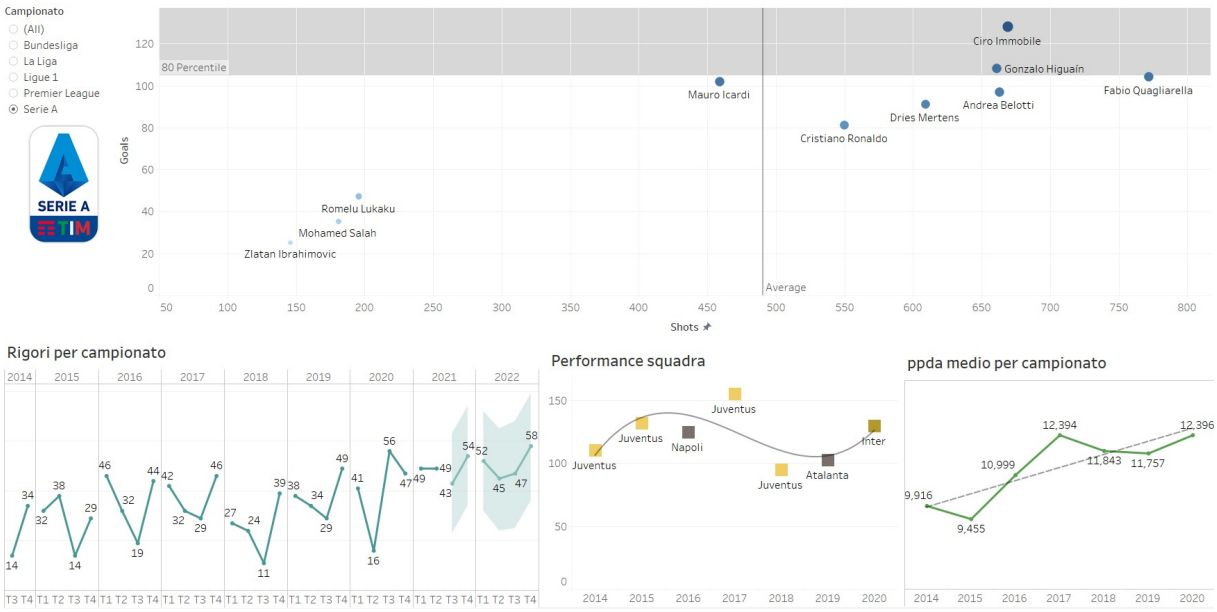


Figura 3.3: Overview campionati

Il primo grafico in alto prende in considerazione i 10 migliori giocatori della Serie A per numero di tiri e gol segnati, con l'obiettivo di individuare quali sono stati gli attaccanti più profilici nell'arco temporale a disposizione nel dataset. In tale grafico si è poi tracciata l'area corrispondente al percentile 80 dei gol segnati e si osserva che fanno parte di tale area gli unici due giocatori che negli ultimi anni hanno eguagliato il record di 36 gol in stagione, ovvero Gonzalo Higuain nella stagione 2015/2016 e Ciro Immobile nella stagione 2019/2020. Da tale scatterplot è interessante osservare la posizione di Mauro Icardi (rimasto in serie A fino alla stagione 2018/2019 per poi essersi trasferito al Paris Saint Germain) il quale si posiziona al di sotto della media dei tiri effettuati dai giocatori, ma spicca il numero di gol che ha effettuato, ovvero di poco sotto al percentile 80. Questo risultato è indice di quanto a tale attaccante bastino pochi tiri per poter portare la propria squadra in goal.
Invece, nella parte in basso della dashboard vengono riportate delle analisi più legate ai team di Serie A. Infatti, il secondo grafico contiene il numero dei rigori assegnati in Serie A nei vari anni, suddivisi per trimestri. Si è considerata tale analisi interessante soprattutto

per vedere l'andamento dei rigori a seguito dell'introduzione della tecnologia VAR (Video Assistance Referee) e per la continua evoluzione della regolamentazione che gestisce le varie azioni di gioco, come ad esempio i tocchi di mano in area. Da tale grafico emerge che il numero di rigori è distribuito nei vari trimestri in modo quasi uniforme, anche se nel caso del T3 del 2020 si osserva un picco. Questo si tratta di un caso anomalo perchè tale stagione fa riferimento a quella della pandemia dove il campionato era stato interrotto e poi ripreso durante la stagione estiva. Inoltre, è stata anche introdotta una previsione di quelli che saranno i rigori nelle nei vari trimestri fino alla fine del 2022, dove si osserva che il trend è in aumento, prevedendo anche di superare il picco avvenuto nel terzo trimestre del 2020.
La terza analisi misura le performance delle squadre della serie A nelle varie stagioni. In questo caso come indice si è considerato il prodotto tra i punti realizzati dalla squadra e la relativa differenza reti nella stagione e di tale indice si è poi presa la squadra migliore. Da tale analisi si osserva che molto spesso il team che compare coincide anche con il vincitore del campionato nella stagione, tuttavia compaiono anche due outlier. Infatti, nella stagione 2016/2017 compare il Napoli che in quell'anno è stata la squadra che ha avuto la migliore differenza reti nel campionato (55 reti contro le 50 della Juventus vincitrice del campionato), ma è arrivata terza in stagione con 4 punti di distacco dalla prima e tale situazione compare di nuovo nella stagione 2019/2020 dove l'Atalanta ha avuto la migliore differenza reti (50 reti), ma è arrivata terza a 5 punti di distacco dalla Juventus vincitrice.
Infine, nell'ultimo grafico è riportato l'andamento del ppda medio della Serie A nelle diverse stagioni. Il **ppda** è un indice definito come il rapporto tra il numero di passaggi effettuati dalla squadra che imposta e il numero di azioni difensive (tackle, intercetti e falli) compiute dalla squadra che aggredisce senza palla. Più basso è il valore di questo indice, più alta sarà la pressione applicata dalla squadra oggetto di valutazione sull'avversario in possesso.
Da tali andamenti si osserva come nel corso delle stagioni tale indice è mediamente aumentato simbolo del fatto che le squadre pressano meno gli avversari che sono in possesso della palla. Per comprendere meglio tale risultato si può fare riferimento al grafico in Figura 3.3 dove nella stagione 2014/2015 il ppda indicava che la squadra in possesso della palla effettuava quasi 10 passaggi prima che gli avversari interrompessero tale possesso, mentre nell'ultima stagione dal valore del ppda si legge che il numero dei passaggi medi effettuati dalla squadra in possesso è salito a poco più di 12.

## 3.3 Analisi del Team: Juventus

Dopo aver presentato in *Qlik* una dashboard generale relativa alla squadra Juventus nell'arco della stagione 2020/21, se ne fornisce ora un'analisi più approfondita, sotto diversi aspetti, sfruttando le potenzialità del software *Tableau*.

### 3.3.1 Mappa dei goal

Una tipologia di grafico finora non chiamata in causa, ed esplorata dunque in *Tableau*, è la mappa. Nello specifico, l'obiettivo era quello di sfruttare gli attributi *positionX* e *positionY*, presenti all'interno del file *shots.csv*, per rappresentare in modo appropriato

la distribuzione territoriale dei tiri effettuati dal team, in particolare i tiri trasformati in goal.
Al fine di adattare la visualizzazione al contesto calcistico, però, non è stato possibile utilizzare il grafico di default 'mappa' offerto da *Tableau*, bensì è stato necessario creare una mappa personalizzata che utilizzasse come sfondo un campo da gioco per poi mappare i valori delle coordinate dei tiri nelle dimensioni in pixel dell'immagine del campo da gioco. A tal proposito sono stati creati dei campi calcolati opportuni in modo da automatizzare l'operazione.

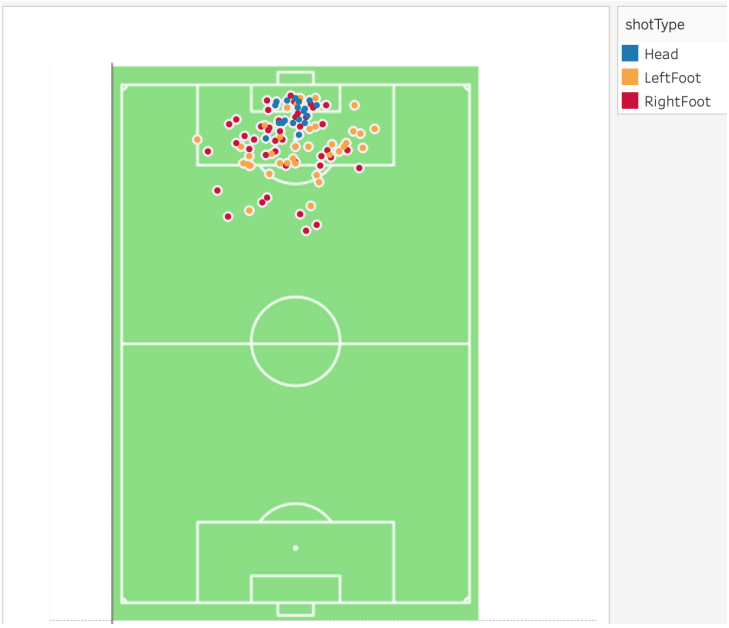


Figura 3.4: Distribuzione territoriale goal Juventus stagione 20/21

La Figura 3.4 riporta la distribuzione territoriale dei tiri effettuati dai giocatori della Juventus nella stagione 2020/21, realizzando una marcatura.
Come si poteva prevedere, è possibile notare come la maggior parte dei goal sia dovuta a tiri provenienti dall'interno dell'area di rigore mentre solo una modesta quantità si deve a tiri da fuori area; da ciò si può evincere che la squadra non disponesse di numerosi tiratori dalle grandi distanze.
Oltre a ciò, è presente una suddivisione su base colore della tipologia dei tiri, in relazione al fatto che questi siano stati effettuati di testa, con il piede sinistro o con il piede destro. In tal caso è da sottolineare che i goal di testa sono una minoranza rispetto al totale mentre c'è un'equa ripartizione tra i goal realizzati con il piede sinistro e con il destro.

### 3.3.2 Profilo Stagionale

La successiva analisi ha preso in considerazione l'andamento stagionale del team in termini di vittorie, pareggi e sconfitte nel tentativo di individuare eventuali peculiarità distintive. Più in dettaglio, nella Figura 3.5 è riportato il profilo dei risultati del team Juventus nel corso delle stagioni 2017/18, sopra, e 2020/21, in basso. La scelta di effettuare il confronto con la stagione 2017/18 non è casuale: questa, infatti, corrisponde alla stagione in cui la Juventus ha conquistato il proprio record di punti (95) e si è posizionata al di sopra della

linea di trend del grafico delle performance dei team riportato in Figura 3.3.
Inoltre, dato che nel file *teamstats.csv* il campo 'result' contiene dati categorici espressi sotto forma di stringhe in cui 'W' sta per vittoria, 'D' sta per pareggio e 'L' sta per sconfitta, si è scelto di associare a tali simboli rispettivamente i valori 1,0,-1 in modo da poter graficare l'andamento.

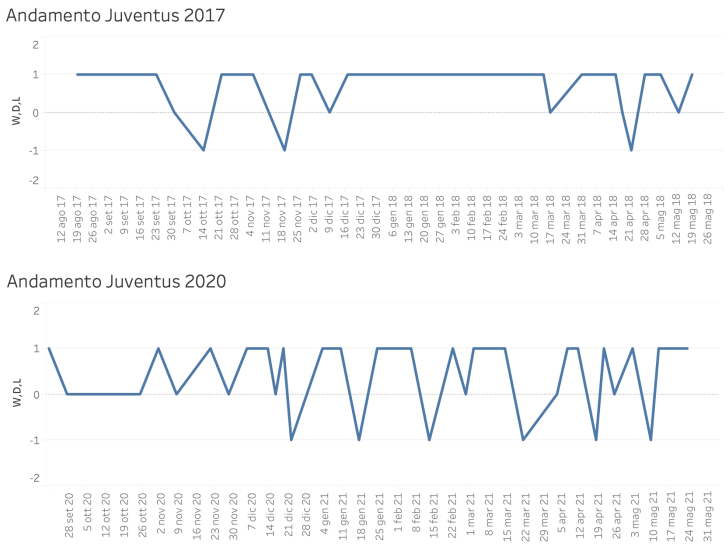


Figura 3.5: Andamento Stagionale Juventus

La Figura 3.5 presenta in ascissa le date in cui si sono svolte le partite nel corso della stagione (da Agosto/Settembre a Maggio) e in ordinata il valore 1 in caso di vittoria, 0 in caso di pareggio e -1 in caso di sconfitta.
Nella stagione 2017/18, annata in cui il team ha vinto il campionato italiano con 95 punti, si nota come siano state inanellate diverse vittorie consecutive dando luogo ad ampi tratti costanti con valore 1, in particolar modo nella fase decisiva della stagione, tra gennaio e marzo.
Nella stagione 2020/21, invece, il team ha chiuso il proprio campionato al quarto posto con 78 punti ed infatti il grafico mostra un andamento del tutto incostante, a denotare le difficoltà riscontrate nell'ottenere serie di risultati utili consecutivi.

## 3.4 Overview espulsioni e ammonizioni

Nel capitolo precedente è stata presentata una panoramica delle informazioni di carattere generale estrapolate dal dataset concentrando l'attenzione sul campionato di Serie A e su di una squadra in particolare. L'analisi prosegue andando a considerare un altro fattore che incide notevolmente sui risultati di una squadra nel corso del campionato ovvero l'inclinazione che ha un team nel ricevere cartellini rossi e gialli. A tal proposito è stata realizzata una dashboard, rappresentata in Figura 3.6, che prende in analisi il campionato di Serie A nella stagione 2020/2021.

Nelle due barre in basso a sinistra viene mostrato il numero totale di espulsioni e ammonizioni nella stagione in esame, suddivisi poi rispettivamente nei due istogrammi per ogni

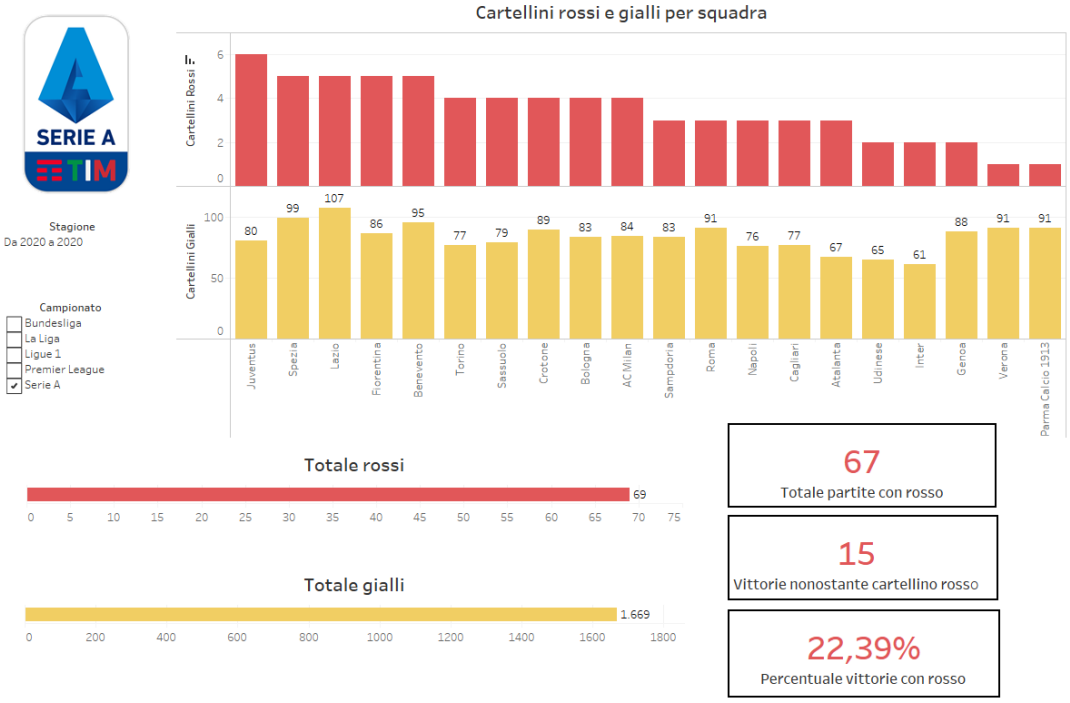


Figura 3.6: Dashboard espulsioni e ammonizioni

squadra del campionato.
Essi mostrano la somma totale dei cartellini rossi e dei cartellini gialli per la singola squadra nel corso della stagione.
Facendo un confronto tra i due grafici la Lazio sembra essere la squadra con un atteggiamento piuttosto falloso giustificato proprio dal totale delle ammonizioni raccolte che la vedono primeggiare nella classifica. Ulteriore riscontro di questo atteggiamento si ha analizzando il numero di espulsioni ricevuto (cinque) che la vedono seconda solo alla Juventus che ha raccolto ben sei rossi. Infine, osservando la parte finale degli istogrammi si nota come nonostante Verona e Parma Calcio 1913 abbiano ottenuto un numero sopra la media di ammonizioni, sono state le squadre che hanno ricevuto solo un espulsione nell'intera stagione.
Infine, i due indicatori nella parte in basso a destra della dashboard rappresentano effettivamente l'incidenza che i cartellini gialli e rossi hanno nei risultati delle partite. Nella stagione 2020/2021 si sono registrate ben sessantasette partite in cui c'è stata almeno un espulsione, di queste partite solo quindici sono state vinte dalla squadra che ha subito la perdita di almeno un giocatore. Da questi risultati si è ricavato quindi che, una volta ricevuta un espulsione, la probabilità di vittoria per la squadra penalizzata diminuisce fino al 22,39 %.

# Capitolo 4

# Power BI

Power BI è un software di business intelligence (BI) sviluppato da Microsoft che permette di effettuare delle analisi dati di tipo descrittivo, diagnostico e predittivo. La forza di questo software risiede nella capacità di poter integrare script in R e Python, linguaggi alla base della data science che permettono, tra le altre cose, di aumentare le funzionalità del tool.
Come i suoi competitor, Power BI offre una serie di strumenti in grado di rappresentare i dati in maniera grafica ed intuitiva attraverso l'utilizzo dei report, concettualmente simili ai fogli di Qlik ed alle dashbord di Tableau.
Rispetto ai due tool precedentemente illustrati, comunque, Power BI offre una serie di strumenti aggiuntivi per la manipolazione dei dati attraverso l'editor di Power Query, da un lato, e l'uso di un linguaggio simile a quello utilizzato per Microsoft Excel, dall'altro. Queste funzionalità gli permettono di essere un buon alleato anche durante la fase di ETL.


Figura 4.1: Logo PowerBI

## 4.1 Caricamento dei dati

Durante la fase di caricamento dei dati sono state eseguite delle operazioni di ETL di pulizia e di riorganizzazione dei dati, per facilitare alcune analisi successive. Oltre ad aver opportunamente cambiato il formato dei dati, è stato eseguito un merge tra la tabella *games* e *teamstats* in modo da poter riportare in quest'ultima non solo i goal fatti dal team, ma anche quelli subiti. Alla luce di ciò, sono state aggiunte le seguenti colonne:

- **Tabella Games**
  - *risultato*: si riporta il risultato finale della partita come combinazione dei goal fatti in casa e fuori casa

- **Tabella Teamstats**
  - *gameID-teamID*: combinazione della colonna *gameID* e della colonna *teamID* che viene utilizzato per legare la tabella con *appearances*
  - *punti*: contiene il valore 3 se la squadra nella partita vinto, 1 se la squadra ha pareggiato e 0 se ha perso
  - *score*: identico a *risultato*
  - *goal subiti*: campo ottenuto grazie al merge di tale tabella con la tabella *games*. Il valore della colonna è stato ottenuto come la somma dei gol fatti dalla squadra in casa e da quella fuori casa e a tale quantità si sono tolti i gol fatti dalla squadra corrispondente alla riga della tabella *teamstats*
- **Tabella Apperances**
  - *gameID-teamID*: analogo al medesimo campo in *teamstats*

Oltre ai campi appena riportati, è stata aggiunta anche la tabella *leagues_ URL* che contiene come campi solo il codice del campionato e un URL che fa riferimento al logo del campionato. Tale tabella verrà utilizzata solo per mostrare dinamicamente i rispettivi logo all'atto della selezione del campionato tra i filtri.
Terminata la descrizione delle operazioni che sono state effettuate durante il caricamento dei dati in Figura 4.2 è mostrato il modello che si è ottenuto.

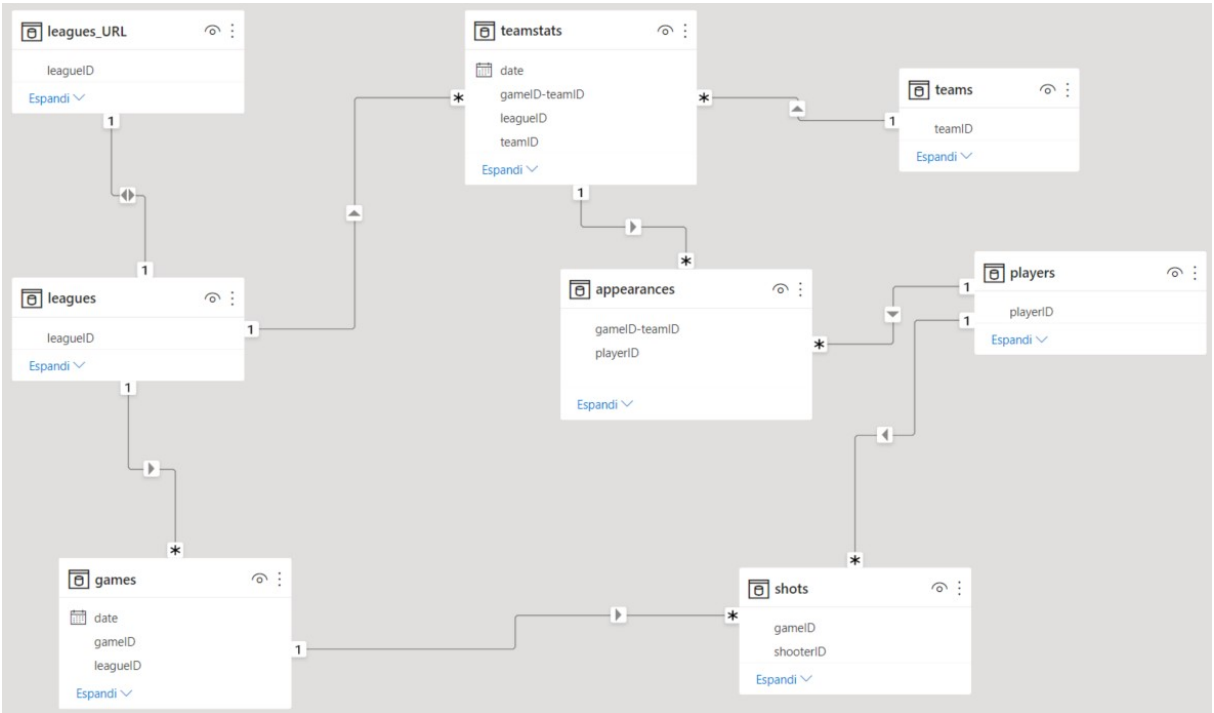


Figura 4.2: Modello Power BI

## 4.2 Creazione Report ed Analisi

Come detto in precedenza, i fogli di Microsoft Power BI sono molto simili a quelli in Qlik ed alle dashboard di Tableau, nei quali è possibile organizzare una serie di grafici al fine di analizzare un determinato fenomeno. I fogli realizzati sono i seguenti, ognuno con un determinato focus di analisi:

1. Panoramica Serie A
2. Analisi Juventus
3. Previsioni Juventus

### 4.2.1 Panoramica Serie A

In questo report si fornisce un'ulteriore panoramica del campionato Serie A, nell'intento di completare le analisi effettuate con i due tool in precedenza, ovviamente sotto aspetti differenti.
In Figura 4.3 si può osservare ciò che è stato creato.

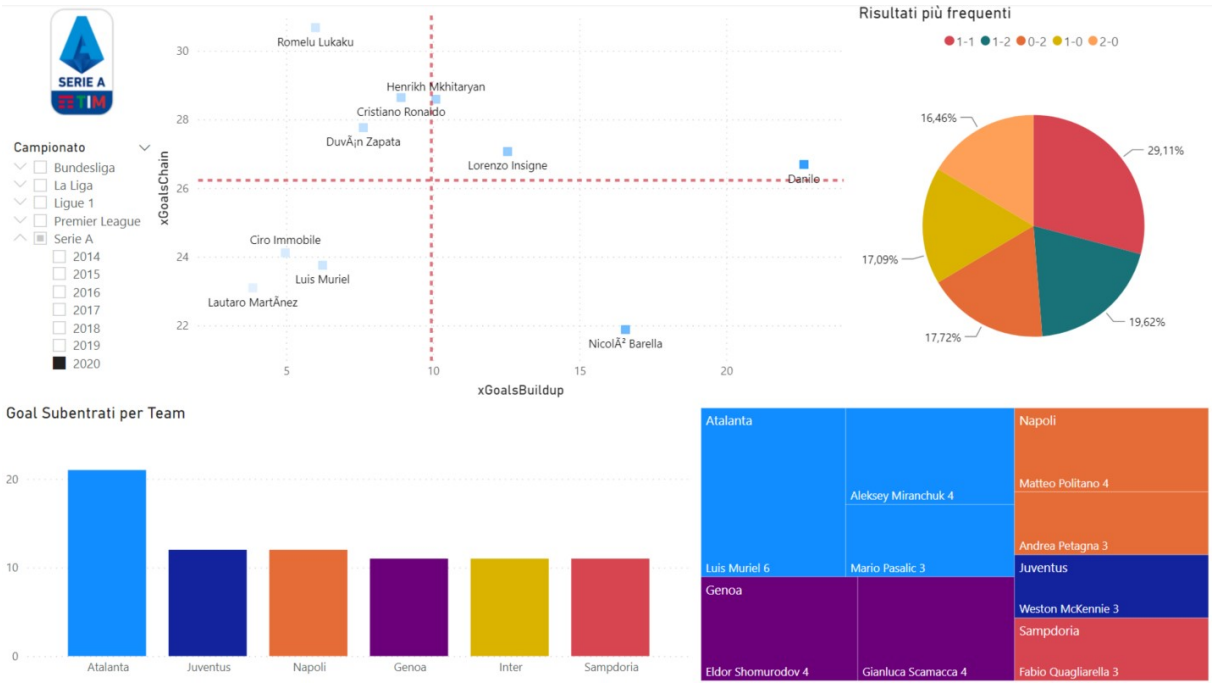


Figura 4.3: Panoramica Serie A

Con tale panoramica sono state mostrate delle analisi che tenessero in considerazione tutte le squadre della Serie A, con i rispettivi giocatori, sempre nella stagione 2020/2021. Una overview di tale stagione calcistica era stata già fornita con Qlik, tuttavia in questo caso sono sono stati riportati dei dati più puntuali.
Dopo aver mostrato con Qlik i giocatori più inattesi in termini di gol e assist realizzati e dopo aver mostrato i giocatori più prolifici in termini di tiri effettuati mediante Tableau, con Power BI si è voluto mostrato quali fossero quei giocatori che in termini di gol fatti

o assist non occupano i vertici delle classifiche, ma sono comunque centrali per quanto riguarda un'azione da gol. Per fare ciò sono stati considerati due indici:

- **xGoalsChain**: rappresenta gli xG ogni volta che il giocatore è coinvolto in un possesso
- **xGoalsBuildUp**: identico agli xGC, ma senza tener conto dei gol e degli assist

Da tale analisi si osserva come Danilo (Juventus) e Nicolò Barella (Inter) siano i giocatori che saltano all'occhio. In particolar modo, il primo occupa il quadrante in alto a sinistra, mentre il secondo quello sottostante. Ciò significa che quando Danilo è in possesso palla sono attesi in tutto più di 26 gol, mentre nel caso di Barella il valore scende a poco meno di 22. Questo dimostra la loro centralità nella squadra e nelle azioni del team, infatti non è un caso che Danilo, nell'analisi del team mediante Qlik (Figura 2.5), sia stato nella stagione 2020/2021 fra i giocatori più utilizzati.
Affianco a tale scatterplot è riportato un grafico a torta dove si possono osservare i 5 risultati più frequenti nelle partite di Serie A ed è interessante osservare come il risultato più frequente sia il pareggio per 1-1: esso, infatti, occupa poco meno di un terzo di tutti i risultati della stagione 2020/2021.
Infine, nella parte inferiore della dashboard, si è posta l'attenzione sui giocatori subentranti. Si è considerata tale analisi interessante in relazione al cambiamento del regolamento avvenuto a partire dalla stagione 2019/2020, attraverso cui il numero di cambi è passato da 3 a 5, con la possibilità, dunque, che questo possa diventare fondamentale per svoltare l'andamento delle partite in corso. Mediante tale analisi si può osservare come l'Atalanta sia stata la squadra che ha segnato di più attraverso giocatori subentrati e questo lo ha fatto con ampio distacco dalle altre, come si può notare dall'istogramma.
In secondo luogo, mediante la mappa ad albero a fianco, i risultati ottenuti sono stati dettagliati per visualizzare meglio quali fossero i giocatori subentranti che sono andati di più in gol. Da ciò si osserva come Muriel sia stato il giocatore più incisivo dalla panchina in tutto il campionato, attraverso i suoi 6 gol realizzati nonostante non sia stato schierato come titolare.

### 4.2.2 Analisi Juventus

Un primo aspetto analizzato in questa sede è stato la distribuzione dei goal tra i vari reparti della squadra Juventus, nel corso di ciascuna stagione presente nel dataset, al fine di evidenziare eventuali punti di forza o criticità. A tal proposito, è stato utilizzato un *Radar Chart*, oggetto visivo importato dalla community di Power BI, organizzato nel modo seguente: ogni "vertice" presenta come etichetta la stagione di riferimento, dopodiché sugli assi sono riportate le diverse quote goal.
Facendo riferimento al dataset, l'attributo *position* della tabella *Appearences* è di tipo categorico e presenta la sigla del ruolo ricoperto da ciascun giocatore, secondo un livello di dettaglio trascurabile per l'analisi in questione. Per questo motivo, per poter studiare la distribuzione dei goal per ruolo, è stato necessario creare delle misure *ad hoc* aggregando i ruoli per macro-categorie: attaccanti, centrocampisti e difensori.

Osservando il *Radar Chart* mostrato nella Figura 4.4, si possono trarre alcune considerazioni.

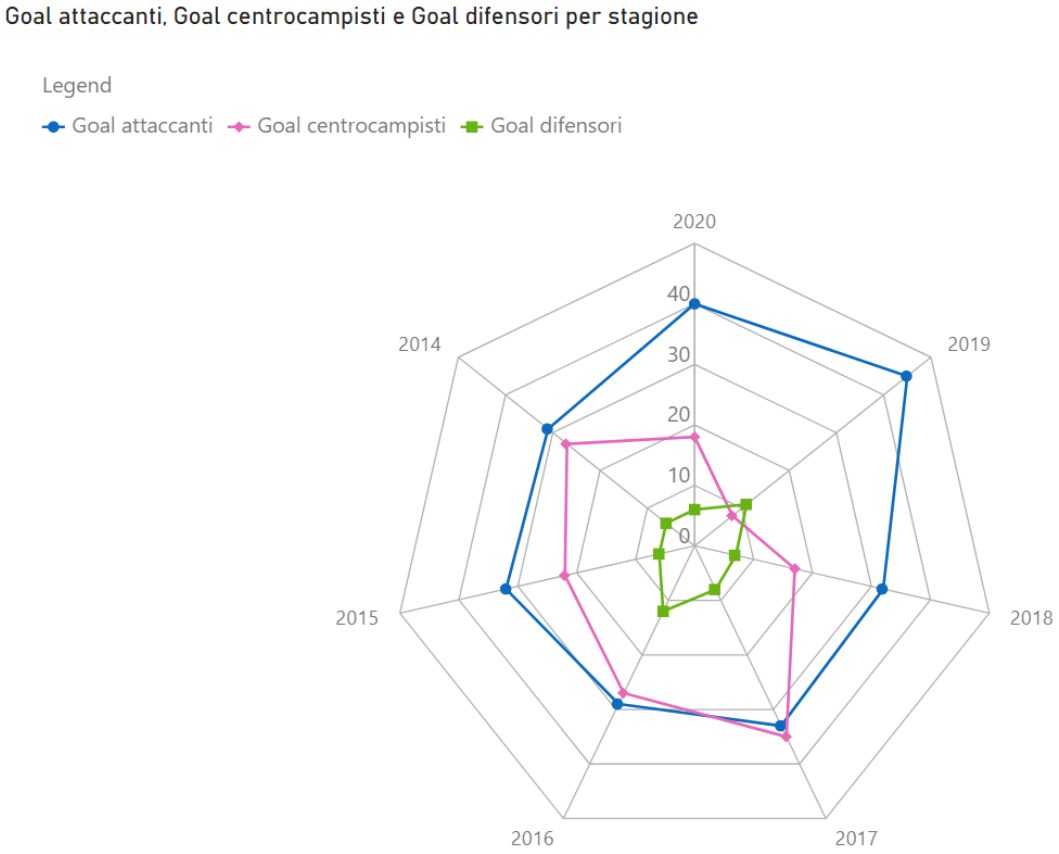


Figura 4.4: Distribuzione goal Juventus per ruolo

In primo luogo, come ci si aspettava, il reparto che mediamente predomina, in quanto a goal realizzati, è l'attacco. Ci sono, però, alcune particolarità da evidenziare.
Il periodo di tempo che va dalla stagione 2014/15 alla stagione 2017/18 ha visto solamente una lieve differenza in termini di marcature tra attaccanti e centrocampisti, a riprova del fatto che la squadra ha puntato molto sugli inserimenti e le capacità realizzative dei propri centrocampisti. Emblema di ciò è la stagione 2017/18, annata in cui i goal realizzati dal reparto di centrocampo hanno addirittura superato i goal messi a segno dagli attaccanti. Da tale stagione in poi, il divario realizzativo tra i due reparti è aumentato considerevolmente, indice del fatto che la squadra ha investito maggiormente sull'attacco e, parallelamente, il sistema di gioco adottato ha messo in luce le capacità realizzative degli attaccanti, a scapito del centrocampo.
Particolarmente significativa è la stagione 2019/20, in cui si è registrato un record negativo per i centrocampisti: le marcature realizzate dai centrocampisti, infatti, sono state meno di 10 ma soprattutto meno dei goal dei difensori. Di contro, il reparto di attacco ha controbilanciato questo trend negativo dei centrocampisti, infatti i goal messi a segno dagli attaccanti sono aumentati notevolmente in valore assoluto nel corso delle ultime stagioni fino a raggiungere e superare la quota dei 40 goal.

Successivamente è stata analizzata la ripartizione dei goal realizzati dalla Juventus, tra primi e secondi tempi, nel corso della stagione 2020/21. Per realizzare tale analisi

è stato utilizzato il linguaggio *DAX* di *Power BI* al fine di creare due misure, *Goal1T* e *Goal2T*, che distinguessero la somma dei goal effettuati rispettivamente nei due tempi della partita, informazione che nel dataset di partenza non era disgiunta.

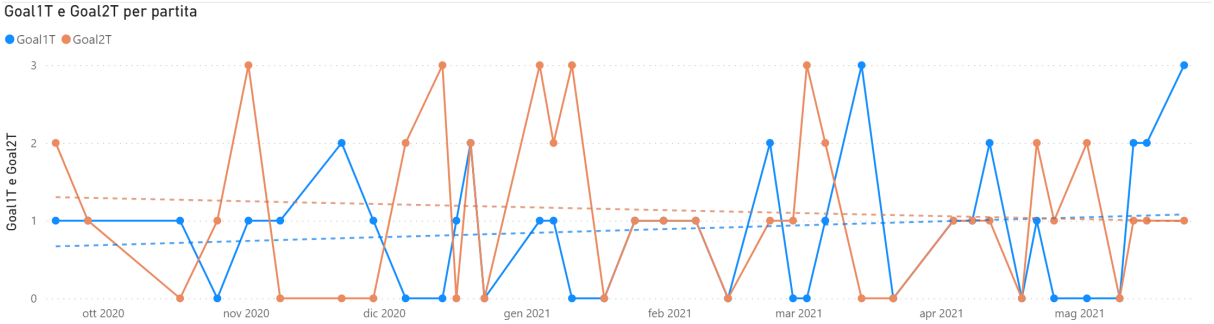


Figura 4.5: Goal primo tempo VS goal secondo tempo, stagione 2020/21

Il grafico risultante, mostrato nella figura 4.5, è un grafico a linee: sull'asse delle ascisse è riportato il periodo della stagione, sulle ordinate il numero di goal realizzati; i puntini indicano i dati relativi alla singola partita.
E' possibile notare come, nella prima parte della stagione, la squadra abbia realizzato mediamente più goal nei secondi tempi rispetto ai primi tempi. Dopodiché, progressivamente, si è registrata un'inversione in tale comportamento: sul finale della stagione, infatti, la squadra ha cercato di risolvere le partite nei primi tempi tant'è che il numero di goal realizzati nei primi 45 minuti di gioco è stato spesso superiore rispetto ai goal segnati nel secondo tempo. Questa conclusione è confermata anche dai trend riportati in sovrimpressione sul grafico a linee, i quali mostrano una progressiva diminuzione dei goal realizzati nel secondo tempo e, contemporaneamente, un aumento dei goal nel primo tempo. Le due linee di tendenza si intersecano in corrispondenza della fine del mese di aprile 2021.

### 4.2.3 Previsioni Juventus

La possibilità di effettuare delle previsioni è stata utilizzata per svolgere le ultime due analisi, la prima delle quali è rappresentata in Figura 4.6 e riguarda la somma dei gol realizzati dalla Juventus nel corso di tutte le stagioni disponibili nel dataset, suddividendo l'ammontare delle reti per trimestre. Nel grafico sono stati evidenziati i trimestri con dei marker di forma circolare di colore rosso. Il primo marker più a sinistra indica il trimestre di inizio del dataset (2014/T3) dal quale, procedendo in maniera sequenziale, si ottengono tutti i trimestri successivi. In generale si può osservare che la maggior parte dei gol sono stati realizzati nei trimestri T1 e T4, giustificato dal fatto che sono quelli i periodi in cui la squadra disputa un numero maggiore di partite. I picchi più in basso infatti rappresentano i trimestri T3 cioè i mesi di Luglio, Agosto e Settembre dove le squadre generalmente sono a riposo o comunque giocano un numero di partite ufficiali notevolmente minore. Unica eccezione a quanto detto è visibile nel T3 della stagione 2019/2020 dove la Juventus ha realizzato ben ventidue reti dovuto al recupero di tutte le partite perse del periodo di lockdown nazionale.
La linea centrale di color arancione indica la media corrispondente a diciannove goal mentre la linea tratteggiata di colore nero indica il trend che, nonostante le prestazioni meno

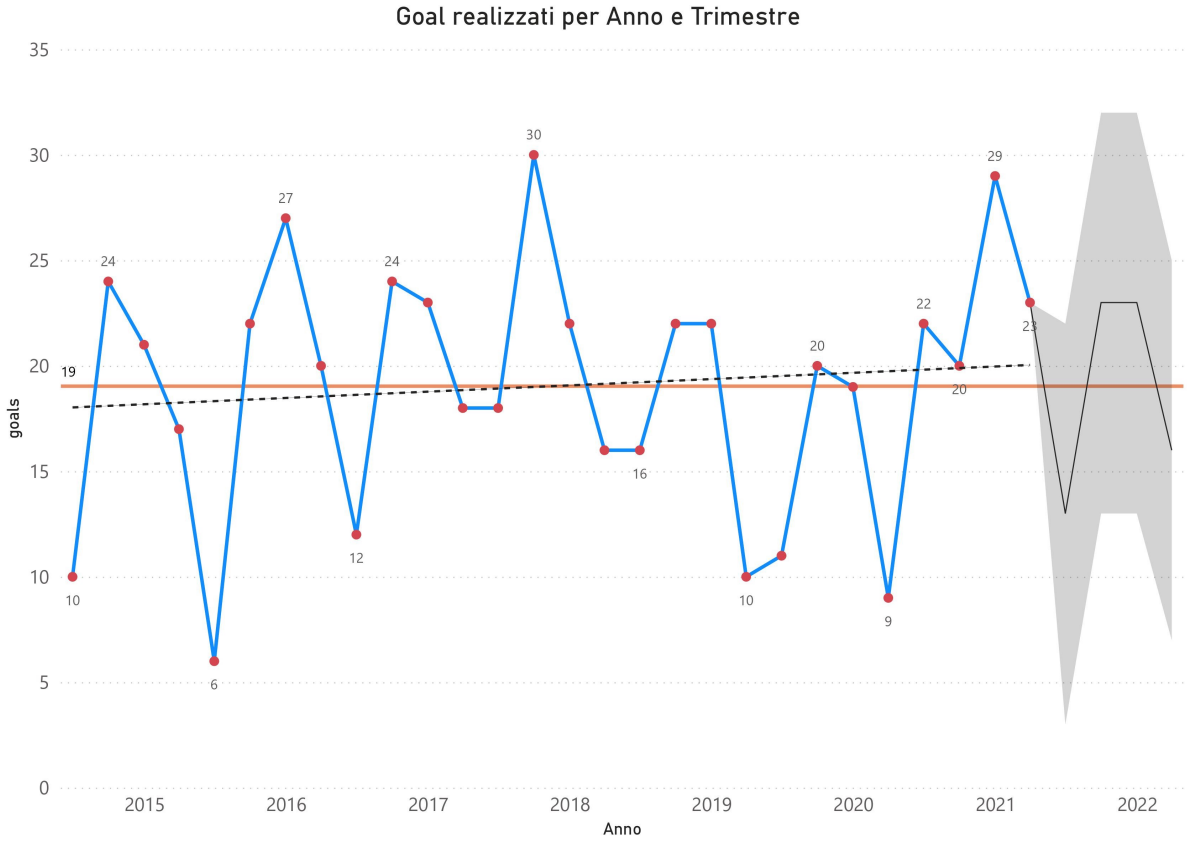


Figura 4.6: Previsione goal realizzati Juventus

brillanti dell'ultimo anno, è in aumento.
La parte finale del grafico mostra una previsione di lunghezza pari a quattro trimestri, calcolata con un intervallo di confidenza del 95%, che quindi si conclude al termine della stagione 2022. I valori della previsione sembrano essere in linea con i valori dei trimestri precedenti con ventitre goal previsti per il T4 del 2021 e il T1 del 2022. L'area grigia attorno alla previsione sta invece ad indicare il limite superiore e inferiore che forniscono l'intervallo in cui al 95% sono contenuti i valori corretti della previsione.

Infine, l'ultima analisi che è stata effettuata è quella relativa ai punti raccolti dalla Juventus nei vari mesi che sono disponibili nel dataset, il cui andamento è visibile in Figura 4.7.

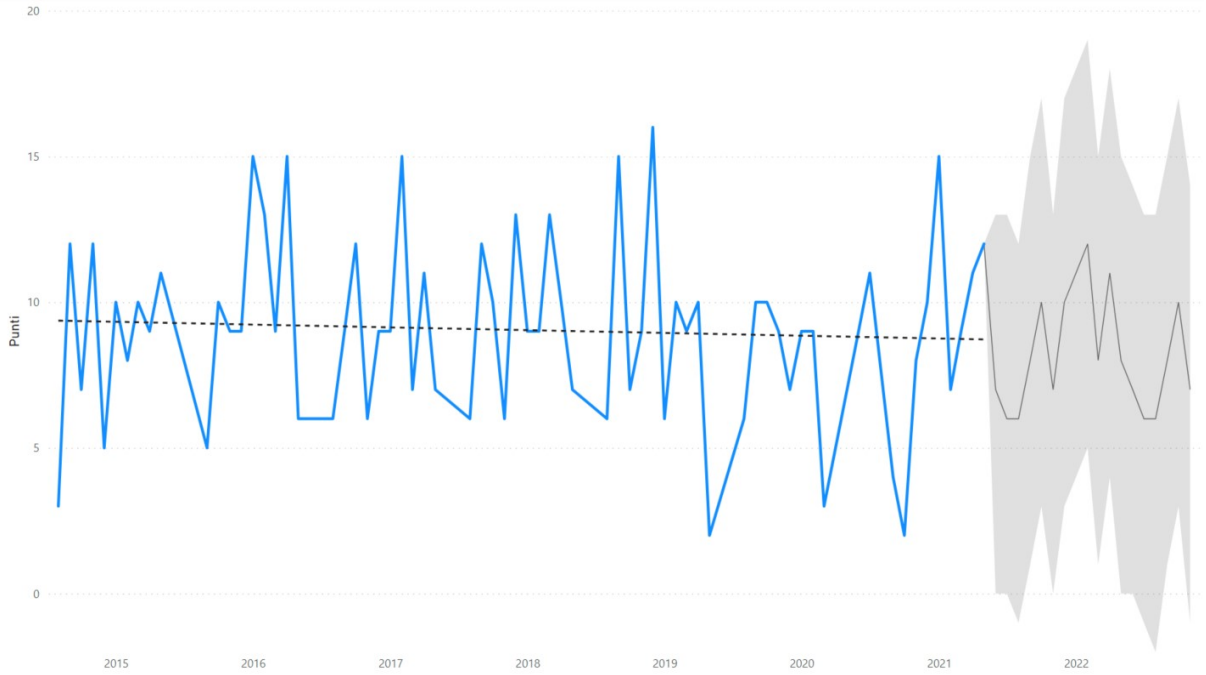


Figura 4.7: Previsione punti Juventus

Da questo grafico si può osservare come il trend dei punti conquistati dalla Juventus sia in calo e infatti questo è riscontrabile anche con ciò che è successo alla squadra nelle ultime due stagioni dove ha riscontrato qualche difficoltà di troppo, rispetto a quelle passate. Ovviamente anche in questo grafico i picchi più in basso sono spesso associabili al mese di agosto dove la squadra prendeva parte a poche partite ufficiali.
In maniera analoga a quanto fatto nell'analisi precedente, anche in questo è riportato una previsione che, questa volta, riguarda i punti ottenuti nei successivi 18 mesi rispetto alla fine del dataset. Come si può osservare i risultati non sono molto promettenti, infatti sembrerebbe che il trend in calo sia ulteriormente rispettato e soprattutto i picchi che si hanno non sono ai livelli delle stagioni migliori e questo fa pensare come la squadra si debba trovare in difficoltà anche nelle stagioni future.