UNIVERSITÀ DEGLI STUDI DI NAPOLI

FEDERICO II

HomeWork 1. PySpark applicato su database Calcistico

FRANCESCO PANARIELLO M63001433
GIOVANNI RICCARDI M63001480
RICCARDO ROMANO M63001489

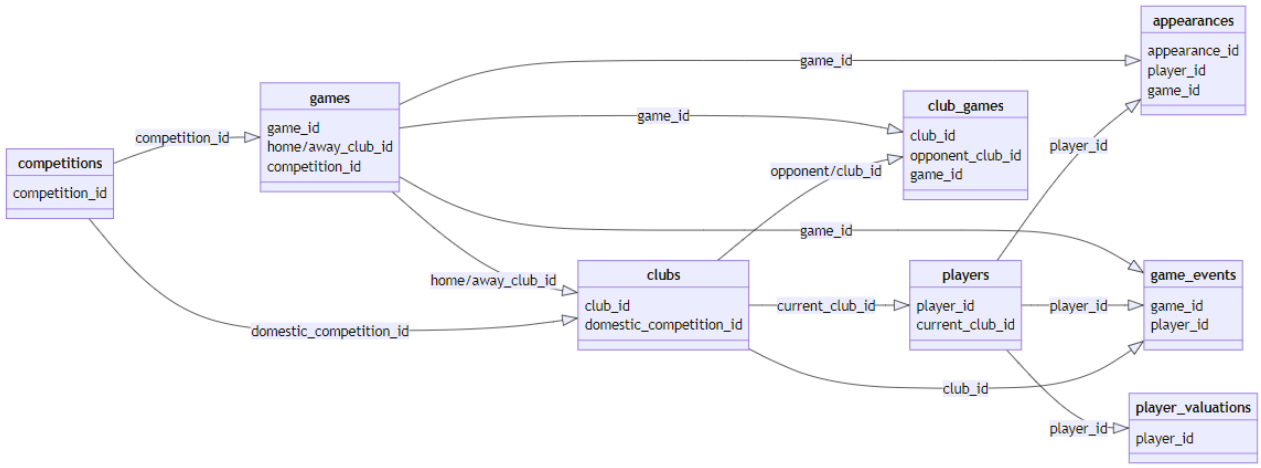
1. Scelta del Dataset

Il dataset scelto, disponibile su Kaggle al link https://www.kaggle.com/datasets/davidcariboo/player-scores , è un insieme di dati relativi al calcio, che include:

- Oltre 60.000 partite da molte stagioni in tutte le competizioni principali.
- Più di 400 club provenienti da tali competizioni.
- Oltre 30.000 giocatori appartenenti a tali club.
- Oltre 400.000 record storici delle valutazioni di mercato dei giocatori.
- Più di 1.200.000 record delle presenze dei giocatori in tutte le partite.

I dati sono puliti, strutturati e aggiornati automaticamente, provenienti da Transfermarkt, una fonte affidabile per le informazioni sul calcio.

Il dataset è composto da diversi file CSV con informazioni su competizioni, partite, club, giocatori e presenze, che vengono aggiornati automaticamente una volta alla settimana. Ogni file contiene gli attributi dell'entità e gli ID che possono essere utilizzati per unirli insieme.



2.Operazioni preliminari

Come ambiente d sviluppo abbiamo deciso di optare per l'utilizzo di Google Colab. In primis, abbiamo installato le librerie necessarie per utilizzare PySpark. Per semplificare l'importazione dei dataset, abbiamo utilizzato una libreria di Kaggle, consentendo l'importazione automatica dei dati senza doverli inserire manualmente ad ogni esecuzione. Infine, abbiamo organizzato tutte le tabelle importate in un dizionario per una gestione più efficiente dei dati.

3.Queries

Abbiamo effettuato varie analisi sullo stile di gioco di allenatori, giocatori e arbitri con anche un focus su strategie di gioco e moduli.

Ogni query è accompagnata da grafici interattivi presenti su Colab, ma non in questa presentazione.

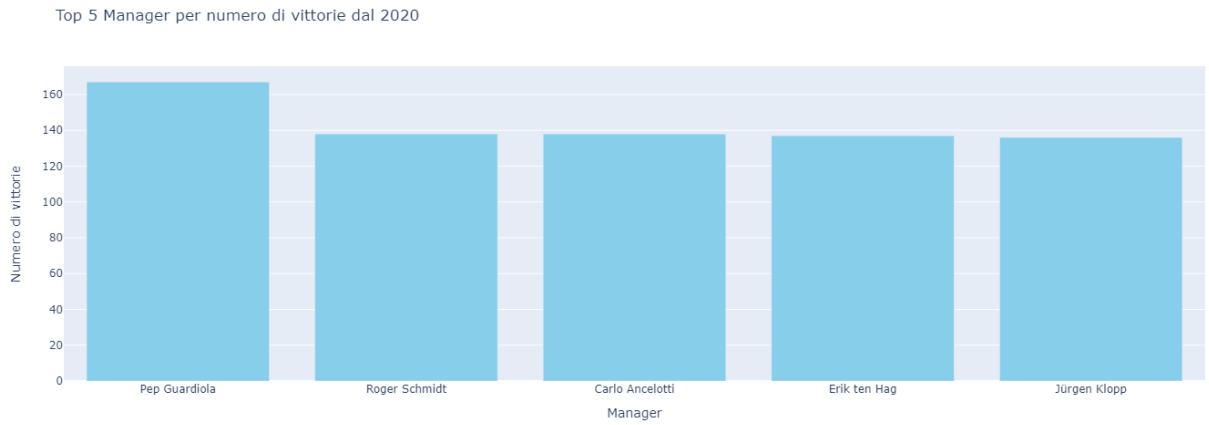
3.1 Allenatori

3.1.1 Allenatori che hanno vinto più partite dalla stagione 2020 fino ad oggi

Quest'analisi è stata scelta per individuare i manager con il maggior numero di vittorie a partire dal 2020. L'obiettivo è identificare i manager più vincenti nel periodo recente.

Cosa è stato fatto: Due DataFrame, "*club_games*" e "*games*", sono stati uniti utilizzando il campo "*game_id*" come chiave di join. Successivamente, i risultati sono stati filtrati per includere solo le partite vinte dai manager (con valore "*is_win*" uguale a 1) e che si sono svolte a partire dal 2020. Dopodiché, i risultati sono stati raggruppati per "*own_manager_name*" (rinominato "*manager*") e calcolato il conteggio delle vittorie per ogni manager. Infine, i risultati sono stati ordinati in ordine decrescente in base al numero di vittorie.

Come si può notare dal grafico sottostante, il Manager più vincente, a partire dall'anno 2020, è stato Pep Guardiola.



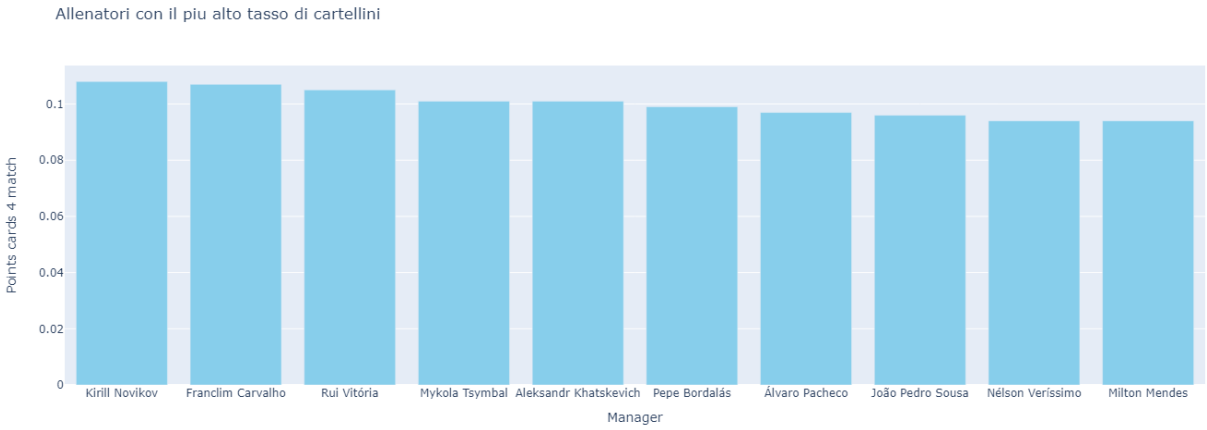
3.1.2 Allenatori che hanno il più alto tasso di cartellini ricevuti su partite giocate

È stata creata una subquery per calcolare le scorrettezze dei calciatori allenati da uno specifico allenatore, punite con cartellini. L'obiettivo è valutare l'aggressività di gioco degli allenatori, considerando il numero di cartellini gialli e rossi ricevuti nelle partite a partire dal 2020.

Cosa è stato fatto: Sono stati creati DataFrame temporanei unendo i dati relativi alle partite, alle presenze e ai cartellini. Successivamente, i dati sono stati filtrati per includere solo le partite a partire dal 2020 e sono stati calcolati i punteggi dei cartellini per ciascun giocatore dello specifico allenatore.

I punteggi sono stati determinati assegnando 1 punto per ogni cartellino rosso e 0.5 punti per ogni cartellino giallo. I risultati sono stati aggregati per ciascun allenatore e sono stati esclusi quelli con meno di 500 partite.

Come si può notare dal grafico, l'allenatore che ha un gioco presumibilmente più "aggressivo" è Kirill Novikov.

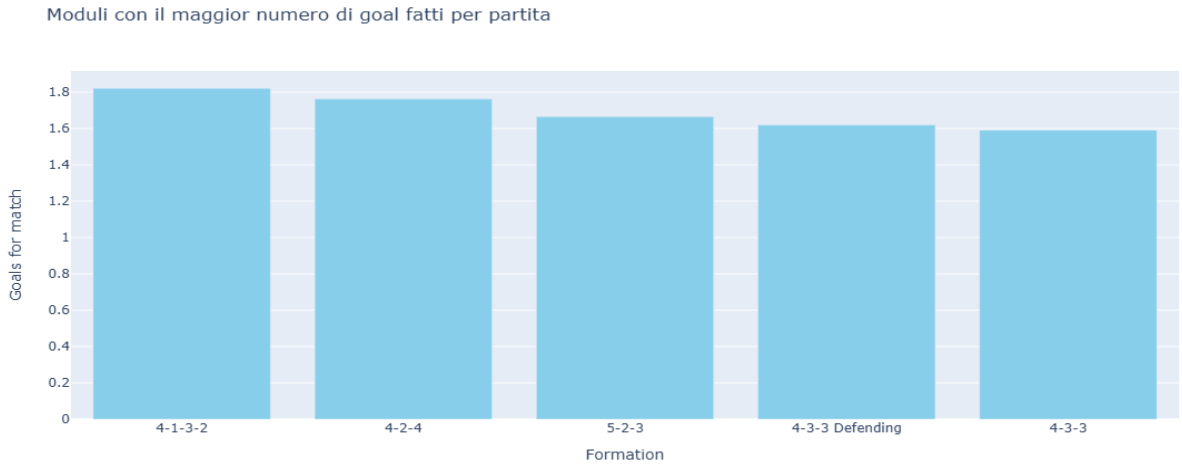


3.2 Moduli

3.2.1 Moduli con il maggior numero di goal effettuati su partite giocate

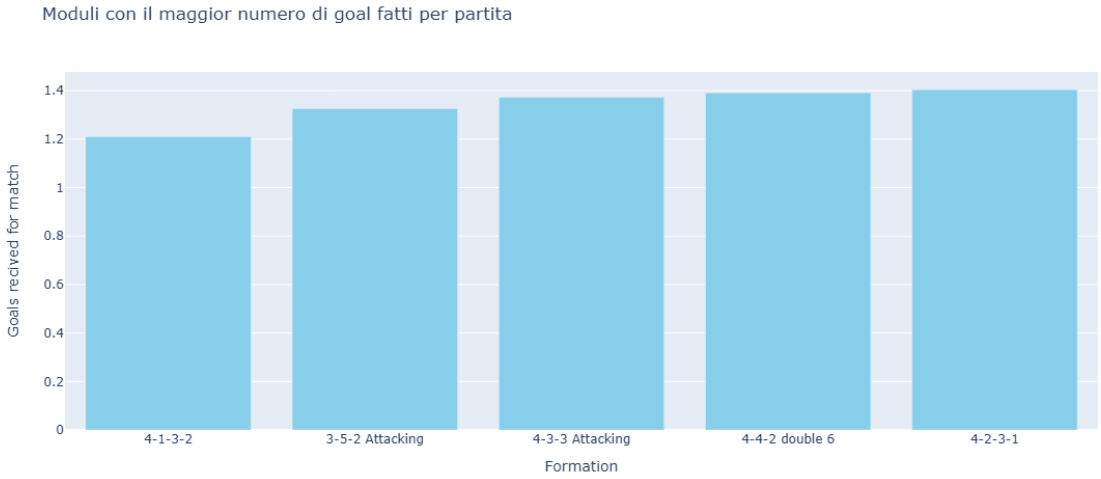
Sono state calcolate le statistiche per l'analisi dei vari moduli utilizzabili. L'obiettivo è valutare il rendimento delle squadre in base alla formazione utilizzata, considerando il numero di goal segnati per partita.

Cosa è stato fatto: Sono stati creati due DataFrame distinti per la squadra di casa e in trasferta. I dati sono stati raggruppati in base alla formazione utilizzata e sono stati calcolati il numero totale di goal segnati e il numero totale di partite disputate. I due DataFrame sono stati poi uniti sulla base della formazione. Infine, sono state calcolate le statistiche finali mediante il rapporto tra goal effettuati su partite giocate ed i risultati sono stati filtrati per includere solo le formazioni con almeno 150 partite giocate in modo da non rendere "falsata" l'analisi.



3.2.2 Moduli che hanno subito il minor numero di goal su partite giocate

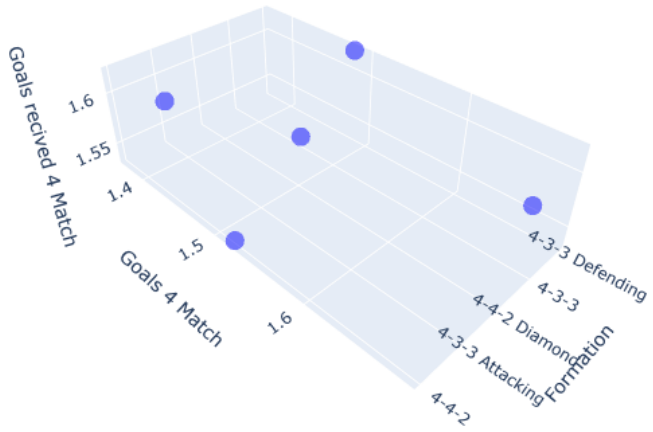
Cosa è stato fatto: Lo sviluppo di quest'analisi è pressoché identico al precedente. L'unica modifica sostanziale è stata quella di andare a considerare la colonna dei goal subiti (sia quelli della squadra in casa che quella in trasferta) e sommarli per il risultato finale.



3.2.3 Statistica modulo migliore rispetto a goal fatti e subiti

Sono state calcolate le statistiche per l'analisi dei vari moduli utilizzabili. L'obiettivo è valutare il rendimento delle squadre in base alla formazione utilizzata, considerando il numero medio di goal segnati e subiti per partita in modo da poter ottenere il miglior modulo da utilizzare.

Cosa è stato fatto: Due DataFrame sono stati creati per le squadre di casa e in trasferta, aggregando i dati in base alla formazione utilizzata e calcolando il numero totale di goal segnati e subiti, nonché il numero totale di partite disputate. I due DataFrame sono stati poi uniti sulla base della formazione. Infine, sono state calcolate le statistiche finali mediante il rapporto dei goal subiti e goal effettuati sulle partite giocate e i risultati sono stati filtrati per includere solo le formazioni con almeno **500** partite giocate. Si noti che sono state aumentate il numero minimo di partite giocate per poter ottenere un modulo valido e testato molte volte.

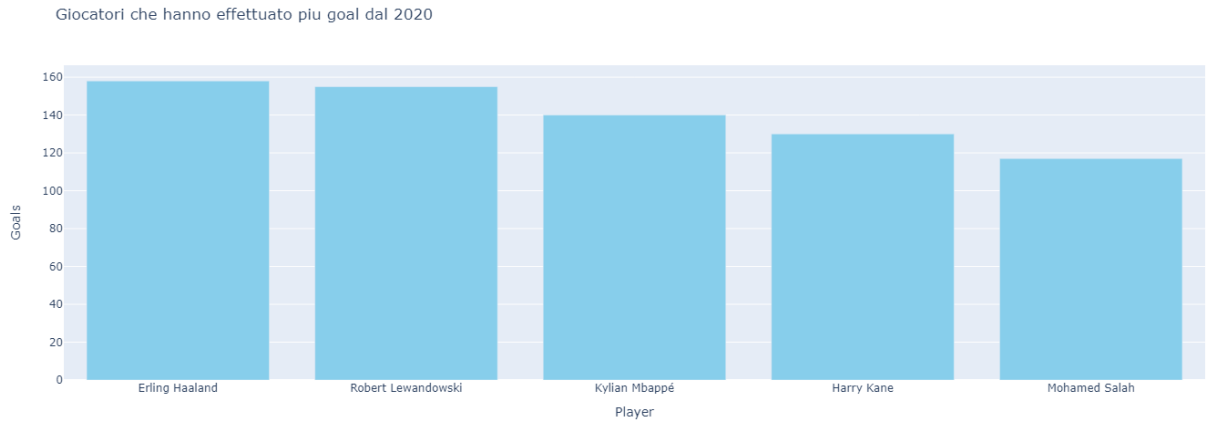


3.3 Giocatori

3.3.1 Giocatori che hanno effettuato più goal dalla stagione 2020 fino ad oggi

I dati delle partite sono stati caricati dalla tabella "*games*", i dati degli eventi delle partite dalla tabella "*game_events*" e i dati dei giocatori dalla tabella "*players*". Sono state eseguite le join tra questi DataFrame per ottenere informazioni sui goal segnati da ciascun giocatore.

Cosa è stato fatto: I dati delle partite sono stati filtrati per includere solo quelle dalla stagione 2020 e sono state eseguite le join con i dati degli eventi e dei giocatori. Successivamente, i dati sono stati raggruppati per il nome del giocatore e sono stati contati i goal segnati da ciascuno. Infine, i risultati sono stati ordinati in ordine decrescente in base al numero di goal e visualizzati i primi 5 giocatori con il maggior numero di goal segnati.



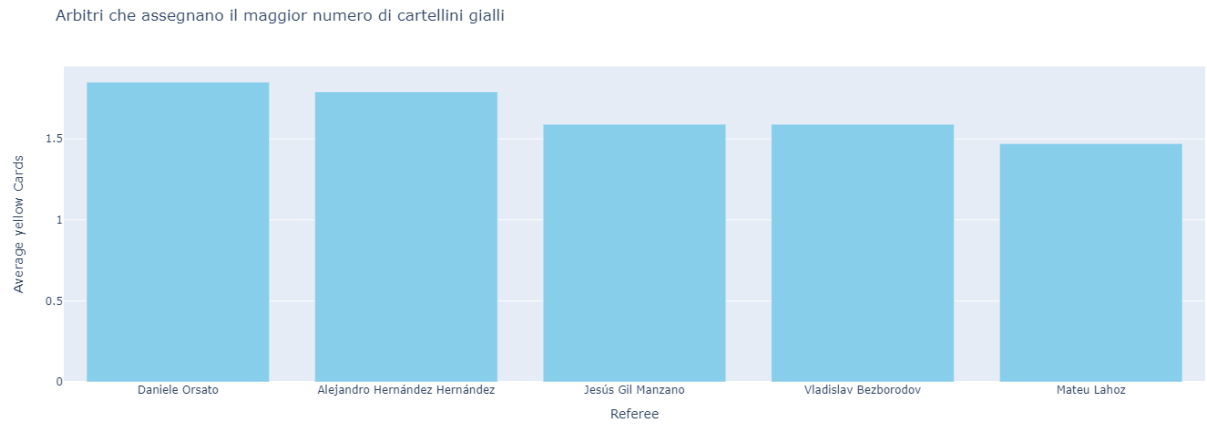
3.4 Arbitri

3.4.1 Arbitri che assegnano il maggior numero di cartellini gialli

Sono stati analizzati i cartellini gialli assegnati dagli arbitri durante le partite. L'obiettivo è valutare il rapporto tra il numero di cartellini gialli assegnati da ciascun arbitro e il numero totale di partite arbitrate da loro in modo da analizzare quale sia l'arbitro più "severo".

Cosa è stato fatto: È stato creato un DataFrame filtrando la tabella degli eventi per individuare gli eventi di tipo "*Cards*" con la descrizione che contiene la parola "*Yellow*". Successivamente, sono stati aggregati i cartellini gialli per arbitro, unendoli con i dati delle partite per ottenere informazioni sugli arbitri di ciascuna partita. In seguito, è stato conteggiato il numero totale di partite arbitrate da ciascun arbitro.

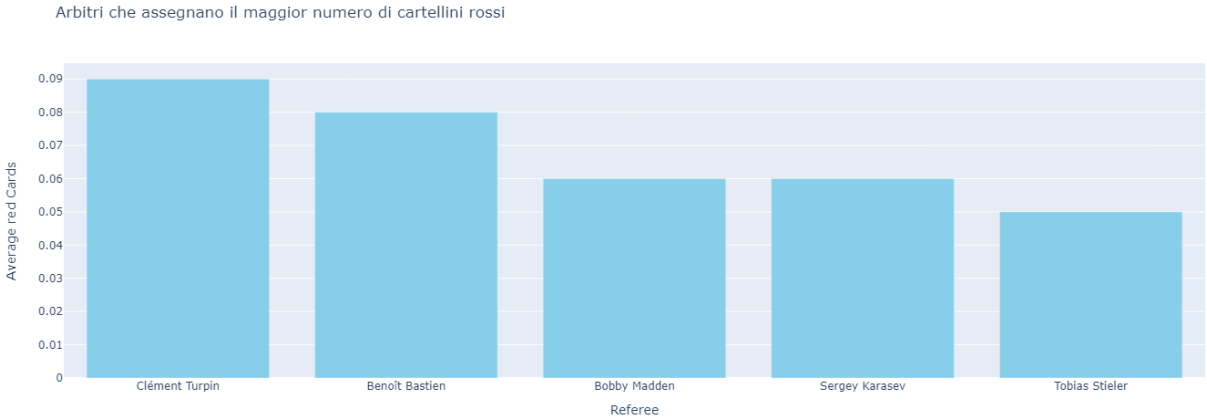
I due DataFrame sono stati uniti ed è stato calcolato il rapporto tra il numero di cartellini gialli e il numero totale di partite arbitrate per ciascun arbitro. Infine, i risultati sono stati ordinati in ordine decrescente in base al rapporto dei cartellini gialli.



3.4.2 Arbitri che assegnano il maggior numero di cartellini rossi

Sono stati analizzati i cartellini rossi assegnati dagli arbitri durante le partite. L'obiettivo è valutare il rapporto tra il numero di cartellini rossi assegnati da ciascun arbitro e il numero totale di partite arbitrate da loro.

Cosa è stato fatto: Lo sviluppo di quest'analisi è pressoché identico al precedente con la differenza che il filtraggio è stato effettuato sugli eventi di tipo "*Cards*" che contenessero la descrizione "*Red*".



3.5 Curiosità

3.5.1 Quante partite ha vinto il Napoli con in campo Mario Rui dal 1°minuto?

Sono state effettuate delle analisi per ottenere informazioni specifiche sul giocatore "*Mário Rui*" del club "*Napoli*" nelle partite vinte. L'obiettivo è ottenere il numero di partite vinte con in campo dal 1° minuto il giocatore, il tutto a causa delle ingenti critiche rivolte al suddetto in modo da poter valutare, oggettivamente, la sua prestazione.

Cosa è stato fatto: Sono state eseguite delle join tra i DataFrame "*club_games*", "*clubs*" e "*game_lineups*" utilizzando le colonne "*club_id*" e "*game_id*". Sono state applicate delle condizioni di filtraggio per selezionare solo le partite vinte dal Napoli, in cui il giocatore "*Mário Rui*" è stato titolare. Successivamente, i dati sono stati raggruppati per il nome del giocatore e conteggiate le partite vinte. Infine, i risultati sono stati mostrati, evidenziando il numero di partite vinte da "*Mário Rui*" con il Napoli.

Player_name	Winning_match
Mário Rui	108

3.5.2 Il Napoli vince più spesso con in campo Mario Rui o Mathías Olivera dal 1°minuto?

Dato che la precedente analisi offriva un risultato molto generico, è stato preferito analizzare se la squadra calcistica Napoli, vincesse più spesso con in campo Mario Rui dal 1° minuto oppure con Mathías Olivera in quanto giocatori dello stesso ruolo.

Cosa è stato fatto: Sono state eseguite le stesse join tra i DataFrame precedenti.

Sono state applicate delle condizioni di filtro per selezionare solo le partite del Napoli in cui Mário Rui o Mathías Olivera sono titolari. Successivamente, i dati sono stati raggruppati per il nome del giocatore e sono stati conteggiati il numero di partite giocate e il numero di partite vinte. Infine, è stato calcolato il rapporto tra il numero di partite vinte e il numero di partite giocate per i calciatori.

Come si può notare, la squadra calcistica Napoli, con la presenza dal 1° minuto in campo di Mário Rui, ha ottenuto un ratio di vittorie pari quasi al 60%, leggermente più alto di quello con in campo invece Mathías Olivera.

Player_name	Winning_match	Total_matches	Winning_ratio
Mário Rui	108	181	**0.597**
Mathías Olivera	14	27	**0.519**

!!! Si noti inoltre che mediante la cella di codice implementata su Colab, è possibile visualizzare le partite vinte ed il corrispettivo "*winning_ratio*" di qualsiasi giocatore appartenente a qualsiasi altra squadra modificando il nome del calciatore da analizzare ed il nome della squadra.