ANALISI DELL'INFLUENZA DEGLI SPETTATORI SUI RISULTATI DELLE PARTITE DI NBA DALLA STAGIONE 2000-01

Corso di laurea magistrale in Data Science

Bonera Michele 890256
Demurtas Federica 884755
Garofoli Damiano 837589

INDICE

1. INTRODUZIONE

Il basket nasce negli Stati Uniti alla fine del 1800 da un'idea di James Naismith, un insegnante di educazione fisica, e dall'inizio del 1900 divenne subito uno degli sport più seguiti soprattutto nel Paese fondante, ma anche all'estero. La NBA, o National Basketball Association, nacque nel 1946 ed è la lega professionistica del Nord America che odiernamente include 30 squadre.
Vista la fama della NBA, che negli ultimi decenni è cresciuta in modo esponenziale, lo studio ha voluto creare un dataset in grado di rispondere a varie domande di ricerca, tutte incentrate sul pubblico presente alle partite. Le principali domande di ricerca si incentrano sulla significatività delle persone presenti in una partita rispetto al risultato di esse o se vi sono differenze fra le diverse formule (regular season o play-off).

2. BUSINESS UNDERSTANDING

La domanda di ricerca prevista per lo studio riguarda l'incidenza del pubblico nelle partite di NBA. L'obiettivo è quello di capire se le partite con maggiore affluenza del pubblico della squadra in casa abbiano un impatto non solo sui risultati finali ma anche su altre variabili come la percentuale dei tiri segnati, gli assist o i rimbalzi catturati.
Inoltre è interessante studiare quali siano le squadre più influenzate dal pubblico e quali invece non abbiano grandi variazioni di risultati con elevata affluenza al palazzetto o meno. Raggruppando invece per partite in casa di una determinata squadra e per partite in trasferta, si può studiare l'andamento delle due e notare le differenze; è infatti risaputo e fortemente creduto che le squadre che giocano in casa abbiano un senso di sicurezza in loro maggiore e quindi che tendano ad ottenere risultati migliori.
Il dataset che si otterrà avrà poi una divisione fra regular season e play-off; durante i play-off solitamente vi è più pubblico in quanto sono partite che possono determinare la vittoria finale del campionato, perciò si potrebbe osservare se questa differenza è significativa tra le due tipologie di partite
Infine, potrebbe essere interessante notare come il pubblico di una determinata squadra, in base alle stagioni precedenti, cambi affluenza del pubblico a seconda che essa abbia avuto una stagione migliore o peggiore delle precedenti. Ad esempio, se, nella stagione regolare, una squadra da quinta passa terza alla stagione successiva e poi seconda a quella dopo ancora, ci si aspetta che nella stagione successiva abbia un'alta affluenza e viceversa.
Le domande di ricerca sono svariate e possono essere ancora ampliate, quindi il dataset finale potrebbe essere molto rilevante dal punto di vista statistico ed economico.

3. DATA PREPARATION

L'analisi prevista necessita dell'acquisizione di varie fonti dati, in particolare due: un dataset riguardanti le partite di NBA e uno riguardante la frequenza del pubblico alle partite. Infine è stato scaricato un terzo dataset riguardante le informazioni sui palazzetti in cui sono state svolte le partite.

3.1 DATA ACQUISITION

DATASET NBA

Attraverso tecniche di API, è stato possibile ottenere i dati relativi alle partite di NBA dalla stagione 2000/2001.
Le API (Application Programming interface) permettono a diverse applicazioni di comunicare fra loro per facilitare la programmazione e l'integrazione.
Attraverso la libreria nba_api di Python, è stato possibile ottenere i dati riguardanti 22 stagioni di nba, di tutte le leghe e squadre.
Il dataset iniziale contiene 2 righe per ogni partita, in modo da avere in una riga i dati di una squadra e nella sottostante i dati dell'altra; da un dataset verticale, esso è stato trasformato in orizzontale, unendo le righe a due a due, passando da 56130 righe a 27657 e da 30 variabili a quasi il doppio. In seguito, anche per sviluppi futuri, si è deciso di tenere solo le variabili più significative per le partite.
In questa fase è avvenuto anche il data cleansing, ovvero la pulizia dei dati per migliorarne la qualità, rendendoli più completi e consistenti. Difatti, si è filtrato per le squadre dell'nba, si sono controllate modifiche e si sono create nuove variabili:

- Tipo partita: divisione fra regular season e play-off
- Home/Away: prima di unire le righe due a due, per indicare se la partita è stata svolta in casa o in trasferta, verrà poi eliminata

Le variabili che sono state mantenute sono le seguenti:

- TEAM_ID_home: numero identificativo della squadra in casa
- TEAM_NAME_home: nome della squadra in casa
- GAME_ID: numero identificativo della partita
- GAME_DATE: data della partita
- PTS_home, FG_PCT_home, FG3_PCT_home, FT_PCT_home, REB_home, AST_home: statistiche riguardanti la partita rispetto alla squadra in casa
- Season: stagione corrente
- Tipo_partita: regular season o play-off

- TEAM_ID_away: numero identificativo della squadra in trasferta
- TEAM_NAME_away: nome della squadra in trasferta
- PTS_away, FG_PCT_away, FG3_PCT_away, FT_PCT_away, REB_away, AST_away: statistiche riguardanti la partita rispetto alla squadra in trasferta
- win_lose: se ha vinto la squadra in casa: "home", altrimenti "away"

È importante sottolineare che una squadra nel corso degli anni ha cambiato nome. Se questo non fosse stato preso in considerazione, numerosi nulla values sarebbero stati presenti nel dataset, rendendolo inconsistente e non aggiornato. La soluzione a questo problema è l'aggiornamento di tutti i record contenenti il nome precedente con quello corrente. In particolare, la squadra ha cambiato nome da 'LA Clippers' a 'Los Angeles Clippers'. In questo modo, si ha un dataset consistente e temporalmente aggiornato per gli scopi dello studio.

DATASET ATTENDANCE

Per l'ottenimento del secondo dataset sono state utilizzate le tecniche di scraping. Lo scraping consiste nell'estrazione di dati attraverso programmi e codici. La metodologia è simile per tutti i siti, ciò che cambia è l'url del sito a cui ci stiamo riferendo. Questa tecnica è capace di trasformare i dati non strutturati di un sito in un database strutturato. Esso inoltre simula la navigazione umana in ogni parte del sito richiesto e crea una tabella strutturata per contenere i dati che si incontrano.
Attraverso questa tecnica e il sito https://www.basketball-reference.com è stato possibile ottenere i dati riguardanti il pubblico presente ad ogni partita.
Le variabili presenti sono le seguenti:

- Data: indica data in cui la partita si è svolta
- TeamHome: indica la squadra che gioca in casa
- TeamAway: indica la squadra che viene ospitata
- Attendance: indica numero di persone presenti alla partita

DATASET PALAZZETTI

Il terzo dataset riguarda la capienza massima di ogni palazzetto ed è stato scaricato da GitHub. I palazzetti totali dove vengono giocate tutte le partite sono 30. Inizialmente le variabili presenti sono 14, ma 5 di esse sono state eliminate poiché non necessarie ai fini del progetto.
In seguito, sono state svolte alcune righe di processing e arricchimento dei dati per migliorarne la qualità, come aggiornamenti e cambi di nomi di variabili. Inoltre è stato necessario aggiungere dati relativi ai palazzetti in cui le squadre nba giocavano precedentemente, in quanto nel dataset iniziale erano presenti soltanto i palazzetti attuali. Successivamente sono stati indicati anche i palazzetti delle squadre che non sono più presenti nella lega ma che giocavano ancora nelle partite presenti nel range di anni considerati per l'analisi.
Infine, anche questo dataset è stato caricato su MongoDB.

Le variabili mantenute al termine della fase di preprocessing sono:

- TEAM_ID: chiave primaria del dataset, indica il codice che identifica la squadra
- MIN_YEAR: valore che indica l'anno in cui il palazzetto è stato utilizzato per la prima volta per una partita della squadra
- MAX_YEAR: valore che indica l'anno in cui il palazzetto è stato utilizzato per l'ultima volta per una partita della squadra
- ABBREVIATION: abbreviazione del nome della squadra
- CITY: città in cui è presente il palazzetto relativo alla squadra
- ARENA: nome del palazzetto della squadra
- ARENACAPACITY: capacità massima del palazzetto

Una parte di data quality è stata appunto svolta in questo dataset, poiché alcuni record non erano aggiornati temporalmente. Per migliorare la qualità di questi dati, è stata effettuata una ricerca e si sono modificate le capienze, aggiornandole al 2022.

3.2 DATA STORAGE

Per la memorizzazione dei dati, si è deciso di utilizzare MongoDB. Esso è un document-based management system, salva i documenti in formato JSON ed è un metodo veloce e dinamico per memorizzare grandi quantità di dati, anche non strutturati. Attraverso la libreria di PyMongo e l'applicazione di MongoDB è stato possibile caricare i dati sul sistema come dei documenti, ed in seguito richiamarli per unire i dataset.
MongoDB è conosciuto per la velocità dell'interrogazione dei dati e per la struttura flessibile; inoltre, esso si adatta facilmente allo scraping e alle API, che sono stati i due metodi prevalenti in questo progetto, quindi MongoDB è risultato essere lo strumento più utile.

4. INTEGRATION

I due dataset principali, nba e attendance, sono stati uniti attraverso le variabili ‘squadra in casa’, ‘squadra in trasferta’ e ‘data della partita’, in modo da individuare con queste tre variabili solo una partita. Successivamente, è stato integrato un terzo dataset attraverso il TEAM_ID, attributo chiave sia del dataset dei palazzetti che del dataset creato con l’unione precedentemente citata. Aggiungendo tale dataset è stata anche introdotta una nuova variabile volta ad indagare la percentuale di pubblico presente alla partita sul totale degli spettatori che possono essere contenuti nel palazzetto.. Il primo step viene comunemente chiamato integrazione, mentre il secondo step può essere categorizzato come arricchimento del dataset.

5. DATA QUALITY

Terminata la fase di integration dei dataset si è verificata la qualità dei dati presenti in ciascuno di essi, questo al fine di evitare criticità nella fase di data exploration che avrebbero portato a dati distorti.
Inizialmente è stato necessario verificare la presenza di valori nulli, i quali potrebbero comportare o la mancanza di alcune informazioni o problematiche generate dall’operazione di integration. All’interno di tutti i dataset non sono però stati identificati valori nulli.
Successivamente sono stati analizzati gli outliers, al fine di comprendere se vi fossero alcune variabili contenenti valori estremamente anomali (come un numero molto basso di punti realizzati da una squadra in una partita). In questo caso sono stati rilevati alcuni outliers critici, che tuttavia non influenzano le analisi svolte nella data exploration in quanto relativi a partite esterne da quelle oggetto di studio, poiché svolte in off-season.

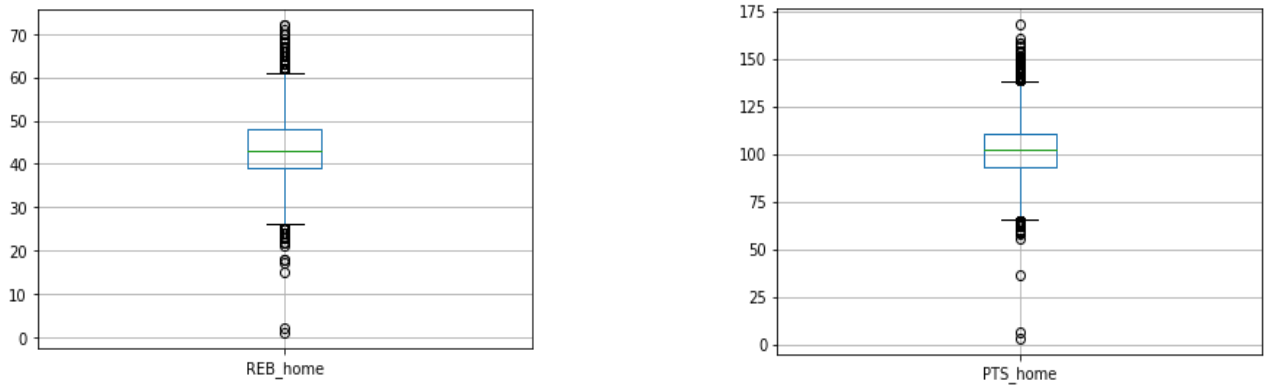

Figura 5.1 Box plot relativi ai Punti segnati dalla squadra in casa e dei rimbalzi fatti dalla squadra in casa (contengono i valori di tutte le squadre)

Infine è stata verificata la coincidenza dei valori del TEAM_ID del dataset dei palazzetti, con i valori del TEAM_ID_home contenuti nel dataset delle partite svolte. Questa operazione è risultata necessaria in quanto nella fase di preprocessing sono state introdotte delle modifiche all’ID delle squadre per poterli uniformare ai diversi palazzetti che sono stati utilizzati dalle squadre negli anni.

In questo modo è stata quindi verificata la consistenza dei dati.
Infine c'è da precisare che ulteriori operazioni di data quality sono state svolte nella fase di preprocessing per ciascun dataset.

6. DATA EXPLORATION

La parte di data exploration si è focalizzata sulle correlazioni, su grafici per mostrare differenze fra variabili in squadre o stagioni differenti e su istogrammi delle variabili principali.

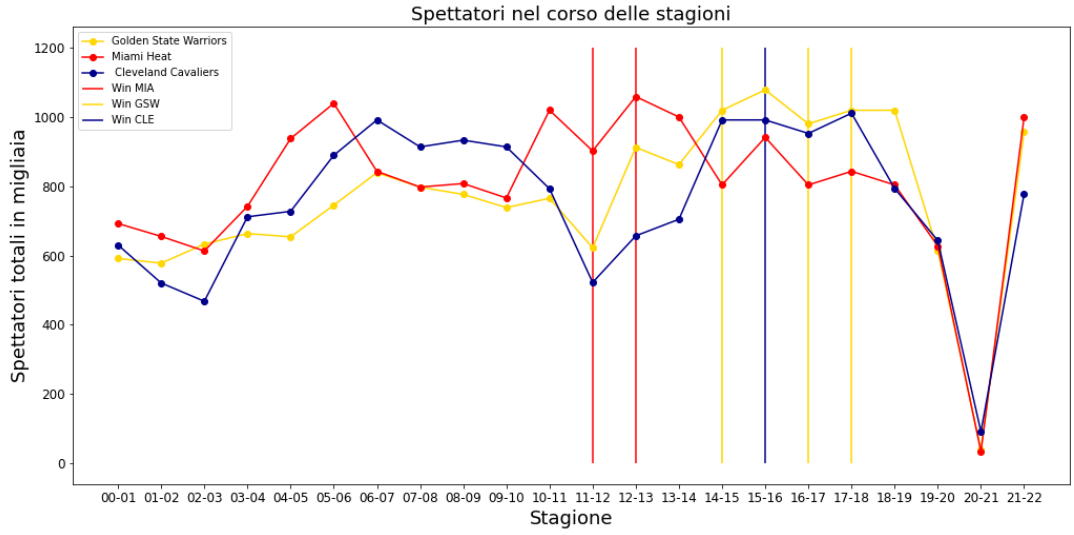

Fig 6.1: Spettatori nel corso delle stagioni

Nella Fig. 6.1 vengono mostrati il numero di spettatori in migliaia nel corso delle stagioni; come si può notare nella stagione 2020-21 si ha un crollo drastico, dovuto dalla pandemia di Covid-19. Le linee verticali indicano le stagioni in cui rispettivamente hanno vinto le squadre nella legenda. Ad esempio, nelle stagioni 2011-12 e 2012-13 i Miami Heat hanno vinto il campionato. Dal 2011-12 al 2012-13 vi è stato un aumento degli spettatori di questa squadra. Stesso andamento può essere notato per i Golden State Warriors tra le stagioni 2016-17, 2017-18 e anche nella stagione successiva, forse dovuto al fatto che nelle due precedenti hanno ottenuto due vittorie.

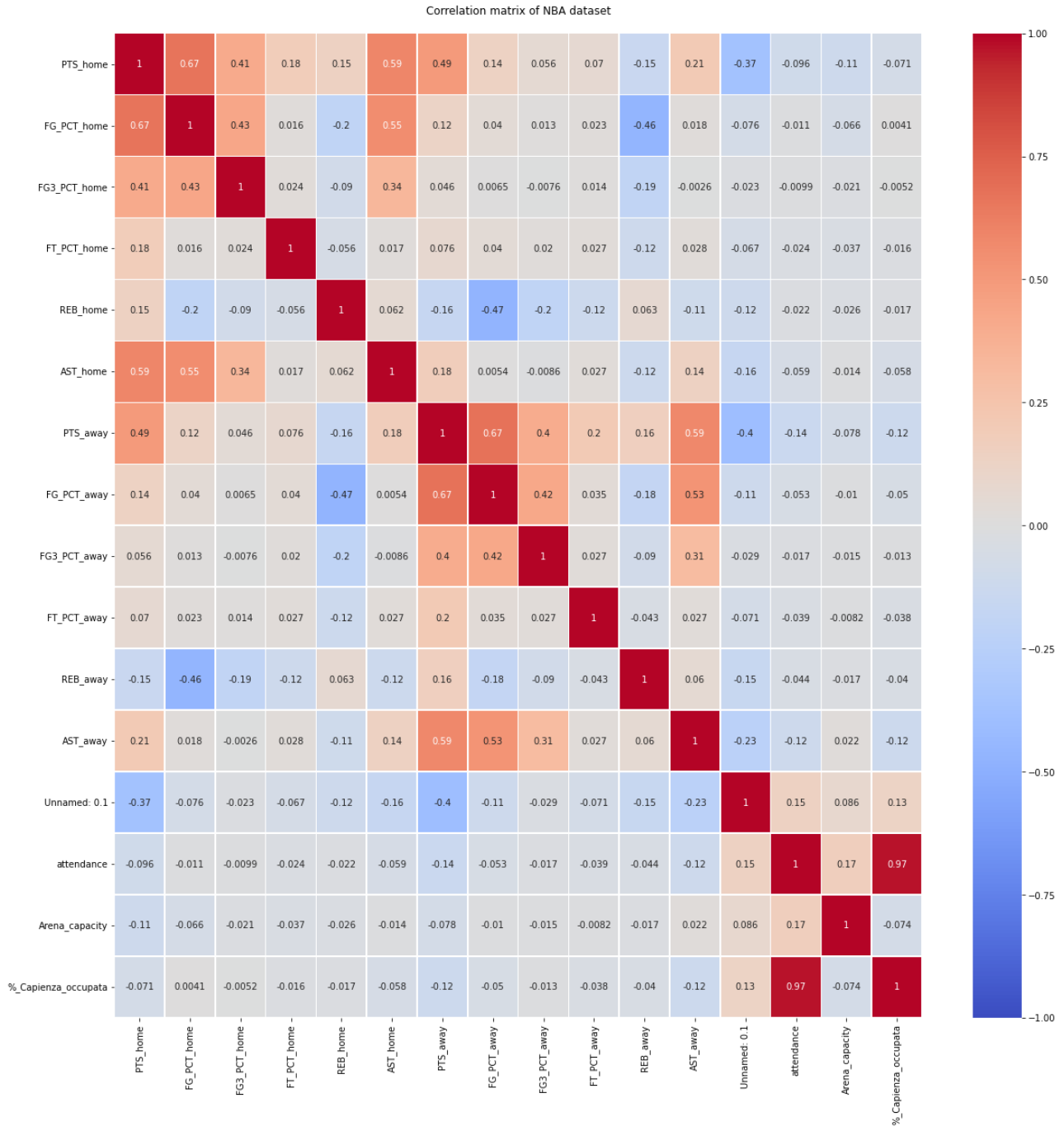

Fig. 6.2: Correlazioni fra variabili numeriche

Lo step successivo è stato lo studio delle correlazioni (Fig. 6.2), ma nessuna di esse è risultata essere abbastanza alta (eccetto per attendance e percentuale di capienza occupata poiché la seconda è una combinazione lineare della prima).

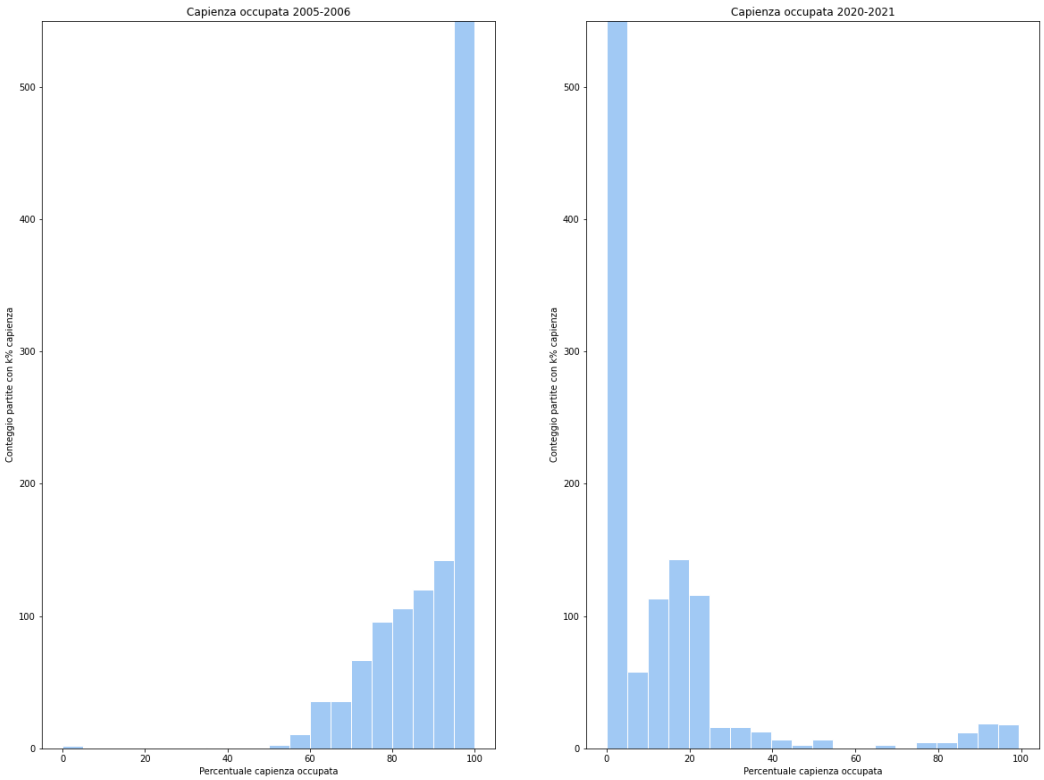

Fig. 6.3: Differenza fra capienze nella stagione 2005-06 e 2020-21.

È interessante come si possa notare una differenza considerevole tra la distribuzione della capienza nella stagione 2005-06 e la stagione 2020-21. Ciò è quasi sicuramente stato causato dalla pandemia e dalle innumerevoli restrizioni che essa ha causato.

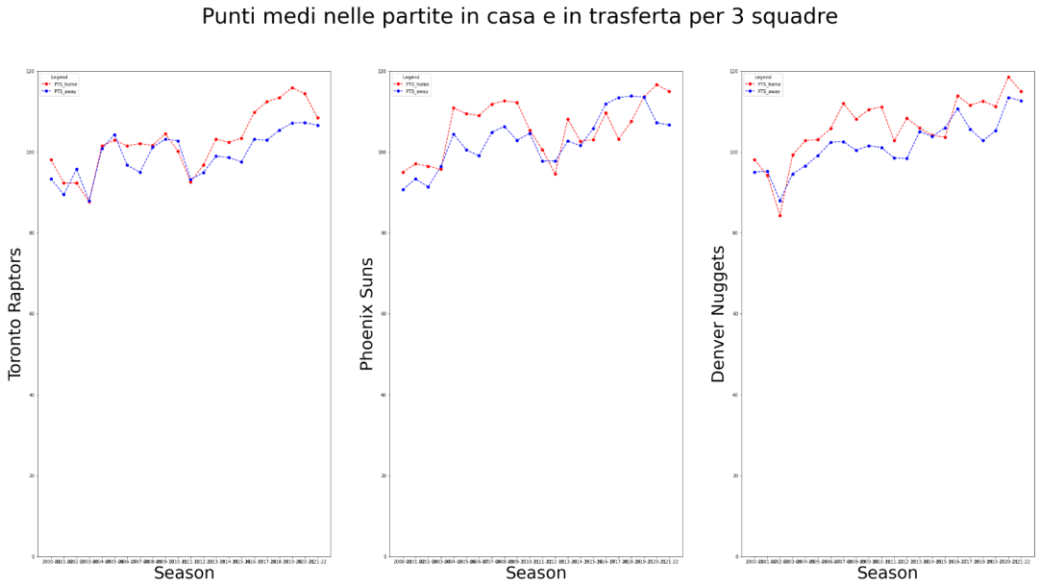

Fig 6.4: Punti medi nelle partite in casa e in trasferta per 3 squadre

La figura 6.4 mostra l'andamento dei punti medi per le partite in casa (rosso) e in trasferta (blu) di 3 squadre diverse, i Toronto Raptors, i Phoenix Suns e i Denver Nuggets nelle diverse stagioni. In generale la prima e la terza squadra mostrano un andamento crescente negli anni, mentre la seconda ha più sbalzi. La linea rossa, ovvero quella riferentesi a partite giocate in casa, in generale sovrasta quella delle partite in trasferta, indicando che durante le partite in casa è più usuale fare punti.

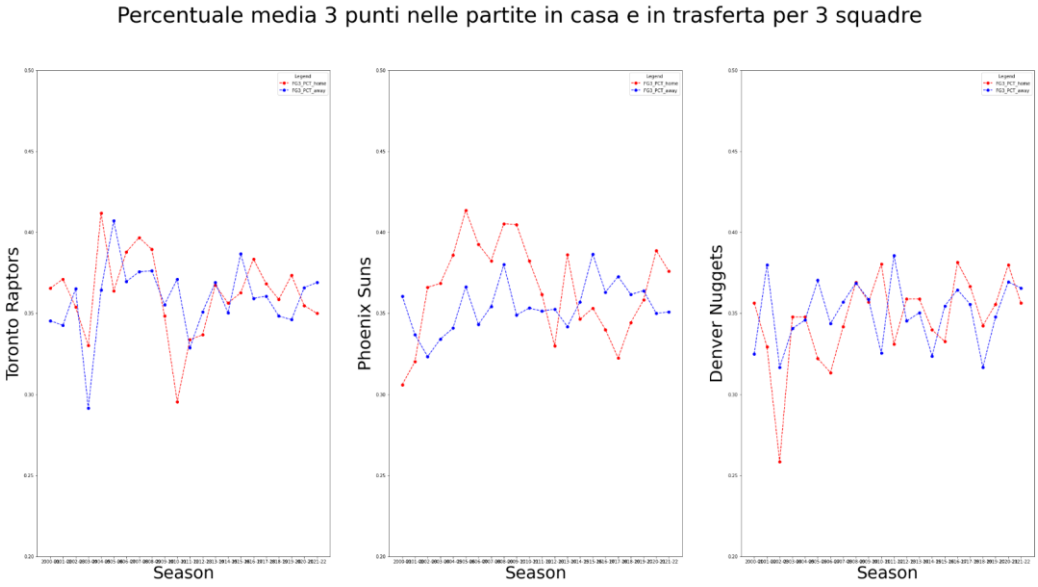

Fig. 6.5 Percentuale media 3 punti per 3 squadre

La figura 6.5 mostra l'andamento della percentuale di punti da 3 per le partite in casa (rosso) e in trasferta (blu) rispettivamente per i Toronto Raptors, i Phoenix Suns e i Denver Nuggets. A differenza del grafico precedente, non ci sono andamenti che indichino un miglioramento o un peggioramento nel corso degli anni e non ci sono differenze visibili tra partite in casa e in trasferta.

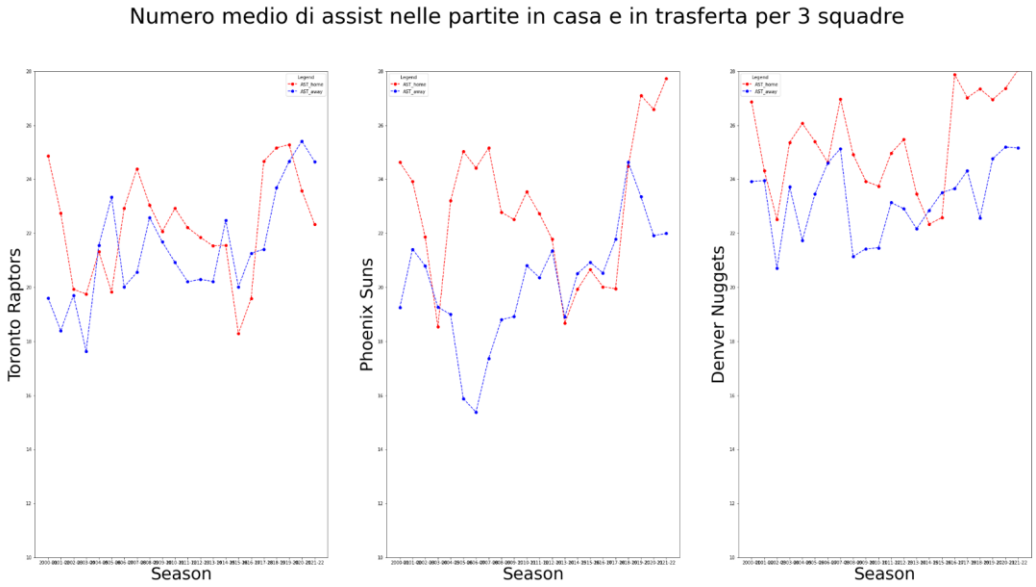

Fig. 6.6: Numero medio di assit nelle partite in casa e in trasferta per 3 squadre

Nell'ultimo grafico riguardante la comparazione fra le tre squadre si è studiato il numero medio di assist nelle partite in casa e in trasferta durante le varie stagioni. Come nel primo grafico, si può notare come le partite in casa delle tre squadre ottengano un numero medio di assist maggiori rispetto alle partite in trasferta.

7. CONCLUSIONI

Le tecniche utilizzate per l'acquisizione dei dati si sono rivelate particolarmente utili agli obiettivi di ricerca, in quanto hanno permesso di usufruire dei dataset con la semplice e corretta applicazione di esse, senza complicazioni particolari nel processo di acquisizione. Lo storage dei dati, ricorrendo all'utilizzo di MongoDB, si è rivelata una scelta adeguata in quanto ha permesso di salvare, modificare e accedere ai dati in maniera rapida ed efficiente.

I dati acquisiti tramite download diretto da Kaggle e GitHub, API_NBA e web scraping su 'basketball reference' erano già fruibili ma necessitavano di diversi accorgimenti per renderli disponibili per l'implementazione.

L'obiettivo del progetto era quello di analizzare l'influenza degli spettatori rispetto a determinate statistiche di gioco e relativamente all'andamento delle partite nel corso degli anni considerati. Da questa analisi è stato possibile verificare se le ipotesi definite nella fase di business understanding fossero corrette, lasciando spazio ad eventuali sviluppi futuri.

I risultati ottenuti dallo studio hanno dimostrato che:

- in generale le statistiche delle partite riferite alle squadre in casa ottengono risultati migliori
- il numero di spettatori nel corso degli anni aveva un trend crescente, con un unico tracollo nella stagione 2020-21 dovuta al Covid, sarebbe quindi interessante vedere se nei prossimi anni il trend riprenda a crescere
- le variabili non sono correlate fra loro in modo significativo, quindi il dataset sarebbe utilizzabile anche per la creazione di modelli statistici

8. IMPLEMENTAZIONI FUTURE

Il dataset creato può essere utilizzato dalle squadre e dai venditori per applicare scontistiche nel caso in cui ci siano periodi in cui l'affluenza del pubblico non sia molto alta. In questo modo, si potrebbe ottenere un'affluenza e un guadagno maggiori.
Inoltre, le squadre e i coach possono trarre beneficio dal dataset osservando gli andamenti delle statistiche (quali punti o assist) in casa e in trasferta; qualora le statistiche risultassero molto minori nelle partite in trasferta potrebbe essere utile trovare nuovi metodi per motivare la squadra che non sia il pubblico.