



# **Analiza podataka utakmica od osnivanja NBA lige do danas**

# Uvod

---

- NBA liga je naprednija i najpoznatija košarkaška liga na svetu, koja okuplja vrhunske igrače iz celog sveta
- Analiza podataka o utakmicama i igračima omogućava nam dublji uvid u dinamiku lige, kao i samog sporta
- Kroz analizu ovih podataka možemo otkriti obrasce, trendove i ključne faktore koji utiču na stil igre, performanse igrača i strategije timova

# Podaci

---

- NBA dataset je skup podataka preuzet sa sajta [Kaggle.com](https://www.kaggle.com) i sadrži podatke o 30 NBA timova, preko 4800 bivših i sadašnjih igrača, kao i 60+ hiljada utakmica
- Za svaku utakmicu postoje box-score podaci koji govore o statistikama timova koji igraju, poput broja poena, skokova, šutnutih trojki, faulova...
- Takođe, za svaku odigranu utakmicu od 1990. godine do danas postoji Play-by-Play kolekcija koji opisuje sve događaje koji su se desili na utakmici kao i vremenska odrednica kada se desilo (Na primer: Nikola Jokić trojka u 10. minutu 3. četvrtine). Ova kolekcija sadrži preko 13 miliona dokumenata

# Podaci

---

- Pored podataka o utakmicama, postoje još i podaci o draftovima održanim od 1950. godine do danas. Ova kolekcija sadrži preko 4500 dokumenata
- Takođe, postoji i kolekcija *News* koja sadrži novinske članke vezane za NBA timove i igrače, koja će se koristiti kako bi se analizirala popularnost pojedinih timova i igrača
- Sveukupno, ovaj dataset sadrži 17 kolekcija i preko 1.4 miliona dokumenata ukupne veličine približno 4GB

# Semantika

---

- U kolekciji *Game* koja predstavlja podatke o utakmicama odigranim tokom sezona, neke od najbitnih kolona za našu analizu su: **fg3a** – broj šutnutih trojki, **reb** – broj skokova jedne ekipe, **plus\_minus** – kvantitativna predstava efikasnosti ekipe...
- Za kolekciju *Draft* kolone od značaja su: **numberRound** i **numberRoundPick** – broj runda i pika u kojoj se igrač izabran, kao i **slugTeam** – koji predstavlja skraćenicu ekipe koja je izabrala igrača
- Kod kolekcije *News* – posmatraćemo kolonu **title** čijom analizom ćemo proveravati popularnost NBA tima ili igrača

# Prostiranje

---

Neki od primera prostiranja ključeva su:

- **id** u kolekciji *Game* koji je referenciran od strane *play\_by\_play* i *game\_summary* kolekcija u vidu kolone **game\_id**
- **slugTeam** u kolekciji *Draft* referencira **abbreviation** kolonu u kolekciji *Team*

# Kratak prikaz trenutne šeme podataka

## Kolekcija: Game

```
{
  "ast_away": 21.0,
  "ast_home": 22.0,
  "blk_away": 6.0,
  "blk_home": 9.0,
  "dreb_away": 33.0,
  "dreb_home": 26.0,
  "fg3_pct_away": 0.423,
  "fg3_pct_home": 0.25,
  "fg3a_away": 26.0,
  "fg3a_home": 24.0,
  "fg3m_away": 11.0,
  "fg3m_home": 6.0,
  "fg_pct_away": 0.43,
  "fg_pct_home": 0.417,
  "fga_away": 79.0,
  "fga_home": 84.0,
  "fgm_away": 34.0,
  "fgm_home": 35.0,
  "ft_pct_away": 1.0,
  "ft_pct_home": 0.727,
  "fta_away": 25.0,
  "fta_home": 22.0,
  "ftm_away": 22.0,
  "ftm_home": 25.0,
  "game_date": "2011-02-07 00:00:00",
  "game_id": "0021000760",
  "min": 240,
  "oreb_away": 11.0,
  "oreb_home": 17.0,
  "pf_away": 20.0,
  "pf_home": 24.0,
  "plus_minus_away": 12,
  "plus_minus_home": -12,
  "pts_away": 104,
  "pts_home": 92,
  "reb_away": 44.0,
  "reb_home": 43.0,
  "season_id": "22010",
  "stl_away": 6.0,
  "stl_home": 7.0,
  "team_abbreviation_away": "MIN",
  "team_abbreviation_home": "NOH",
  "team_id_away": "1610612750",
  "team_id_home": "1610612740",
  "team_name_away": "Minnesota Timberwolves",
  "team_name_home": "New Orleans Hornets",
  "tov_away": 15.0,
  "tov_home": 13.0,
  "wl_away": "W",
  "wl_home": "L"
}
```

## Kolekcija: News

```
{
  "author": "Lawrence Hurley",
  "authors": "Lawrence Hurley",
  "clean_url": "reuters.com",
  "country": "US",
  "html": "<!DOCTYPE html><html lang=\\"",
  "images": "{ 'https://static.reuters.co",
  "is_opinion": 0,
  "language": "en",
  "link": "https://www.reuters.com/artic",
  "media": "https://static.reuters.com/",
  "movies": "[]",
  "page_number": 1,
  "published_date": "2021-03-26 10:12:4",
  "published_date_precision": "full",
  "query": "Basketball",
  "querystring_from": "2021/03/26",
  "querystring_lang": "en",
  "querystring_page": 1.0,
  "querystring_page_size": 100.0,
  "querystring_q": "Basketball",
  "querystring_to": "2021/03/27",
  "rank": 88,
  "rights": "reuters.com",
  "scrape_date": "2021-04-26 19:37:31.8",
  "summary": "By Lawrence HurleyWASHINGTON",
  "text": "WASHINGTON (Reuters) -U.S. S",
  "title": "U.S. Supreme Court scrutiniz",
  "top_image": "https://static.reuters.co",
  "topic": "business",
  "twitter_account": "@Reuters"
}
```

## Kolekcija: Player

```
{
  "first_name": "Alaa",
  "full_name": "Alaa Abdelnaby",
  "id": "76001",
  "is_active": 0,
  "last_name": "Abdelnaby"
},
```

## Kolekcija: Draft

```
{
  "PLAYER_PROFILE_FLAG": 1.0,
  "cityTeam": "Minnesota",
  "idPlayer": 1630162.0,
  "idTeam": 1610612750.0,
  "locationOrganizationFrom": null,
  "nameOrganizationFrom": "Georgia",
  "namePlayer": "Anthony Edwards",
  "nameTeam": "Minnesota Timberwolves",
  "numberPickOverall": 1.0,
  "numberRound": 1.0,
  "numberRoundPick": 1.0,
  "slugOrganizationTypeFrom": "COL",
  "slugTeam": "MIN",
  "teamName": "Timberwolves",
  "typeOrganizationFrom": "College/University",
  "yearDraft": 2020.0
}
```

## Kolekcija: Team

```
{
  "abbreviation": "ATL",
  "city": "Atlanta",
  "full_name": "Atlanta Hawks",
  "id": "1610612737",
  "nickname": "Hawks",
  "state": "Georgia",
  "year_founded": 1949.0
},
```

# Predlog logičke sheme

---

Neki od primera prostiranja ključeva su:

- Podaci iz tabele `play_by_play` smešteni su u tabelu `game` jer se ovi podaci uvek koriste zajedno, preko referencirajućeg polja `game_id`
- Primenom šablona proširene reference, umesto kolone `team_id` u tabeli `game`, čuvaćemo objekat koji obuhvata `abbreviation` i `name`, i time izbeći joinovanje tabela prilikom svakog upita.
- Primenom šablona proračunavanja, napravićemo segregaciju i čuvaćemo za svaku sezonu kolekciju sa učinkom timova iz te sezone. Čuvaće se broj pobeda, poraza, prosečan broj poena po utakmici itd.





# **Predlog upita nad kolekcijama podataka**

## Odrediti kako se menjao prosečan broj šutnutih i pogođenih trojki po sezoni od 1985. godine do danas

---

- Košarka je igra koja stalno evoluirala. Igrači poput Stefa Karija i Kleja Tompsona su doprineli da se igra značajno ubrza i više orijentiše na šutiranje trojki
- Cilj je da uvidimo koliko se i kako igra kroz sezone menjala dok se dešavala tranzicija sa igre iz reketa na šutiranje trojki



## **Pronaći top 10 igrača po broju utakmica na kojima su postigli barem 40 poena**

---

- Kroz istoriju NBA lige pojavljivali su se izuzetni igrači koji su promenili dinamiku lige i doveli svoje timove do titula
- Cilj je da vidimo koliko se na ovom spisku nalazi igrača koji se smatraju za najveće ikada i da otkrijemo da li se ovde nalaze neki igrači koji su potcenjeni u tom razmatranju



## **Za svaki tim odrediti prosečan broj postignutih poena, asistencija i skokova po utakmici na domaćem terenu i u gostima**

---

- Za dosta ekipa se smatra da mnogo bolje igraju na domaćem terenu nego kada gostuju
- Naša želja je da procenimo koje to ekipe imaju velika odstupanja u ovom slučaju a na koje ne utiče pritisak gostovanja





## U prethodne tri sezone izračunati i uporediti efikasnost Denver Nagetsa i Filadelfija 76ersa

---

- U prethodne tri sezone se konstantno vodi polemika oko toga ko treba da bude MVP lige: Nikola Jokić ili Joel Embid
- Ideja da, izračunavanjem naprednih metrika, vidimo koliko su zapravo ekipe Denvera i Filadelfije efikasne pod vođstvom ova dva igrača i da utvrdimo da li je Embid zapravo zaslužio ovogodišnju nagradu
- Metrike koje će biti korišćene su: +/- ekipe, efektivni FG procenat, Turnover ratio, procenat ofanzivnih skokova i Free Throw attempt rate za koje NBA liga kaže da su najznačajnije kada se analizira efikasnost ekipe

## Za svaku ekipu izračunati ukupan broj koševa za pobedu po sezoni

---

- Neke ekipe dokazano imaju pobednički mentalitet i zato nam je cilj da uvidimo koje su to bili najsrčaniji timovi u različitim nba sezonama
- Analizom Play By Play kolekcije koja sadrži preko 1.3 miliona podataka o utakmicama utvrdićemo za svaku sezonu koji su timovi imali najviše smirenosti u odlučujućim trenucima



## **Odrediti uticaj prvog pika na draftu u odnosu na povećanje broja pobeda ekipe u sezoni**

---

- Prvi pik na draftu obično bude najtalentovaniji mladi igrač u jednoj klasi koji bi trebalo da donese prevagu ekipi koja ga izabere
- Naš cilj je da uvidimo koliko to često zapravo važi i koliko često ekipe zapravo dobiju na lutriji prilikom izbora prvog pika



## **Odrediti zastupljenost Lebrona Džejmsa u vestima i svih NBA ekipa, kao i od koliko ekipa je LeBron popularniji**

---

- LeBron Džejms je i pre nego što je došao u ligu bio jedan od najpopularnijih igrača, i tokom celom karijere bio je medijski najzastupljeniji igrač
- Neke od NBA franšiza, i pored svih uspeha, nisu previše medijski ispraćene zbog svoje lokacije, egzotičnosti, stila igre
- Ideja je da analizom novinskih članaka vidimo od koliko ekipa je LeBron Džejms, kao brend, popularniji i zastupljeniji u medijima





## **Odrediti procenat utakmica koje tim izgubi nakon što je vodio sa 10 ili više razlike na kraju treće četvrtine**

---

- U svetu košarke, kao i u drugim sportovima, igrači se suočavaju sa ogromnim pritiskom da postignu što bolje rezultate. Timovi koji poseduju mentalitet pobednika uspevaju da se adekvatno nose sa tim pritiskom i pronalaze načine da ga prevaziđu, dok neke to i ne mogu
- Cilj je da utvrdimo koje su to ekipe kojima nedostaje pobednički mentalitet i u kom procentu su zastupljene



## **Odrediti broj igrača izvan 10. pika koji su u prvoj sezoni imali bar 10 utakmica sa preko 20 poena**

---

- Danas je dobro poznato koliko je bitno da timovi u sportu posvete pažnju kvalitetnom skautingu talenata. Prepoznavanje adekvatnog potencijala u ranoj fazi omogućava timovima da maksimalno iskoriste svoj izbor na draftu kako bi ostali u koraku sa najboljim ekipama
- Ideja je da utvrdimo koliko je moguće da timovi sa niskim izborom na draftu pronađu odgovarajuće pojačanje koje može da napravi razliku



## Ispitati koliki broj utakmica u sezoni odigra 5 najbolje rangiranih igrača na prethodnom draftu

---

- Poznato je da nadolazeći mladi sportisti imaju danak neiskustvu i da nisu momentalno spremni da svoj potencijal sprovedu u delo
- Cilj je da utvrdimo koliko su najbolji mladi igrači spremni, kada budu bačeni u vatru i koliko bi vremena ekipe trebalo da im daju da sazru kako bi mogli da računaju na njih

