# Ethan Ooi | Establishing the Data

## Setup

In terminal:

git clone https://github.com/vys5hb/Design-Final.git

```
pip install kagglehub -q
```

## Import Libraries

In [1]:

```
import pandas as pd
import kagglehub
```

## Download Data

https://www.kaggle.com/datasets/ehallmar/nba-historical-stats-and-betting-data/data

This data was produced from

In [ ]:

```
path = kagglehub.dataset_download("ehallmar/nba-historical-stats-and-betting
print("Path to dataset files:", path)
```

```
Warning: Looks like you're using an outdated `kagglehub` version (installed:
0.3.6), please consider upgrading to the latest version (0.3.13).
Downloading from https://www.kaggle.com/api/v1/datasets/download/ehallmar/nb
a-historical-stats-and-betting-data?dataset_version_number=1...
100%|██████████| 36.5M/36.5M [00:01<00:00, 21.6MB/s]
Extracting files...
Path to dataset files: /Users/ethanooi/.cache/kagglehub/datasets/ehallmar/nb
a-historical-stats-and-betting-data/versions/1
```

## Read Data

In [3]:

```
bets = pd.read_csv(f'{path}/nba_betting_spread.csv')
games = pd.read_csv(f'{path}/nba_games_all.csv')
lines = pd.read_csv(f'{path}/nba_betting_money_line.csv')
```

In [4]:

```
bets.head()
```

Out[4]:

| | game_id | book_name | book_id | team_id | a_team_id | spread1 | spread2 | price1 |
|---|---|---|---|---|---|---|---|---|
| **0** | 21000358 | Pinnacle Sports | 238 | 1610612749 | 1610612742 | 7.5 | -7.5 | -106.0 |
| **1** | 21000358 | 5Dimes | 19 | 1610612749 | 1610612742 | 7.5 | -7.5 | -110.0 |
| **2** | 21000358 | Bookmaker | 93 | 1610612749 | 1610612742 | 7.5 | -7.5 | -110.0 |
| **3** | 21000358 | BetOnline | 1096 | 1610612749 | 1610612742 | 7.5 | -7.5 | -110.0 |
| **4** | 21000358 | Bovada | 999996 | 1610612749 | 1610612742 | 8.0 | -8.0 | -115.0 |

```
In [5]: games.head()
```

Out[5]:

| | game_id | game_date | matchup | team_id | is_home | wl | w | l | w_pct | min |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 20800741 | 2009-02-06 | SAC vs. UTA | 1610612762 | f | W | 29.0 | 22.0 | 0.569 | 240 |
| **1** | 20800701 | 2009-01-31 | POR vs. UTA | 1610612762 | f | L | 26.0 | 22.0 | 0.542 | 240 |
| **2** | 20800584 | 2009-01-16 | MEM vs. UTA | 1610612762 | f | W | 24.0 | 16.0 | 0.600 | 240 |
| **3** | 20800558 | 2009-01-12 | IND @ UTA | 1610612762 | t | W | 23.0 | 15.0 | 0.605 | 240 |
| **4** | 20800440 | 2008-12-27 | HOU vs. UTA | 1610612762 | f | L | 18.0 | 14.0 | 0.563 | 290 |

5 rows × 32 columns

```
In [6]: lines.head()
```

Out[6]:

| | game_id | book_name | book_id | team_id | a_team_id | price1 | price2 |
|---|---|---|---|---|---|---|---|
| **0** | 41100314 | Pinnacle Sports | 238 | 1610612759 | 1610612760 | 165.0 | -183.0 |
| **1** | 41100314 | 5Dimes | 19 | 1610612759 | 1610612760 | 165.0 | -175.0 |
| **2** | 41100314 | Bookmaker | 93 | 1610612759 | 1610612760 | 160.0 | -190.0 |
| **3** | 41100314 | BetOnline | 1096 | 1610612759 | 1610612760 | 165.0 | -190.0 |
| **4** | 41100314 | Bovada | 999996 | 1610612759 | 1610612760 | 155.0 | -175.0 |

# How to get the Data?

1. You can go to this link, https://www.kaggle.com/datasets/ehallmar/nba-historical-stats-and-betting-data/data, and download the data from there.

2. You can use the code at the top of this notebook and it will automatically download the data for you, the data reads works directly with this code.

## Who produced the data? And how?

This data comes from Kaggle by the user Evan Hallmark, username: ehallmar. This data hasn't been updated in 7 years, but it's a great resource for historical NBA data and has a large collection of data. I believe the data was bought from another source which regularly updates this data file. The original data link can be found here: https://www.scottfreellc.com/shop/p/nba-historical-odds-data.

## COLS Tables

In [7]:

```
bet_info = pd.DataFrame({
    "Count": bets.count(),
    "Types": bets.dtypes,
    "Nulls": bets.isnull().sum()
}).reset_index()
bet_info
```

Out[7]:

| | index | Count | Types | Nulls |
|---|---|---|---|---|
| **0** | game_id | 131690 | int64 | 0 |
| **1** | book_name | 131690 | object | 0 |
| **2** | book_id | 131690 | int64 | 0 |
| **3** | team_id | 131690 | int64 | 0 |
| **4** | a_team_id | 131690 | int64 | 0 |
| **5** | spread1 | 131690 | float64 | 0 |
| **6** | spread2 | 131690 | float64 | 0 |
| **7** | price1 | 131690 | float64 | 0 |
| **8** | price2 | 131690 | float64 | 0 |

In [8]:

```
games_info = pd.DataFrame({
    "Count": games.count(),
    "Types": games.dtypes,
    "Nulls": games.isnull().sum()
}).reset_index()
games_info
```

Out[8]:

| | index | Count | Types | Nulls |
|---|---|---|---|---|
| **0** | game_id | 125624 | int64 | 0 |
| **1** | game_date | 119376 | object | 6248 |
| **2** | matchup | 125624 | object | 0 |
| **3** | team_id | 125624 | int64 | 0 |
| **4** | is_home | 125624 | object | 0 |
| **5** | wl | 125614 | object | 10 |
| **6** | w | 41000 | float64 | 84624 |
| **7** | l | 41000 | float64 | 84624 |
| **8** | w_pct | 41000 | float64 | 84624 |
| **9** | min | 125624 | int64 | 0 |
| **10** | fgm | 125607 | float64 | 17 |
| **11** | fga | 90894 | float64 | 34730 |
| **12** | fg_pct | 90871 | float64 | 34753 |
| **13** | fg3m | 95248 | float64 | 30376 |
| **14** | fg3a | 84316 | float64 | 41308 |
| **15** | fg3_pct | 80062 | float64 | 45562 |
| **16** | ftm | 125605 | float64 | 19 |
| **17** | fta | 120011 | float64 | 5613 |
| **18** | ft_pct | 119960 | float64 | 5664 |
| **19** | oreb | 83786 | float64 | 41838 |
| **20** | dreb | 83639 | float64 | 41985 |
| **21** | reb | 125624 | int64 | 0 |
| **22** | ast | 90068 | float64 | 35556 |
| **23** | stl | 83960 | float64 | 41664 |
| **24** | blk | 84407 | float64 | 41217 |
| **25** | tov | 84293 | float64 | 41331 |
| **26** | pf | 121742 | float64 | 3882 |
| **27** | pts | 125624 | int64 | 0 |
| **28** | a_team_id | 125624 | int64 | 0 |
| **29** | season_year | 125624 | int64 | 0 |
| **30** | season_type | 125624 | object | 0 |
| **31** | season | 125624 | object | 0 |

In [9]:

```
lines_info = pd.DataFrame({
    "Count": lines.count(),
    "Types": lines.dtypes,
    "Nulls": lines.isnull().sum()
}).reset_index()
lines_info
```

Out[9]:

| | index | Count | Types | Nulls |
|---|---|---|---|---|
| **0** | game_id | 125286 | int64 | 0 |
| **1** | book_name | 125286 | object | 0 |
| **2** | book_id | 125286 | int64 | 0 |
| **3** | team_id | 125286 | int64 | 0 |
| **4** | a_team_id | 125286 | int64 | 0 |
| **5** | price1 | 125286 | float64 | 0 |
| **6** | price2 | 125286 | float64 | 0 |