Final Project EDA
Seth Johnson, Noor Dhaliwal, Swayam Chidrawar
11/12/25

Our Project

We are developing a regression model to predict the amount of points an NBA player will score in an upcoming game.

Our Dataset

We are using the nba_api python package as our primary data source. This is an unofficial API that provides easier access to NBA stats than the existing NBA.com APIs. It is entirely up-to-date with the original website.

It is divided into two main modules. The static module contains statistics that don't change season to season, like player and team biographical information. The endpoints module contains the live stats about every player in every game for every season since the 1946-47 season. Some notable endpoints include:

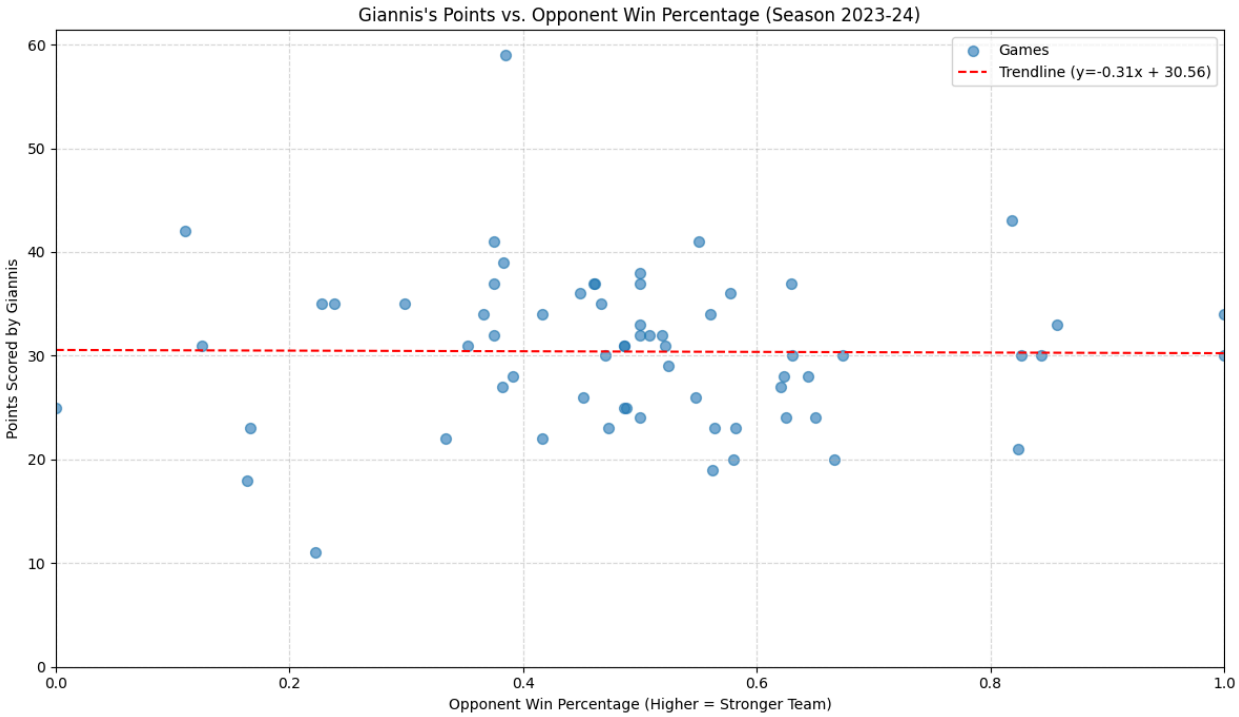
- playercareerstats: Season by season stats for a given player
- playergamelog: Game by game stats for a given player in a given season
- boxscoresummaryv2: The box score for a given game
- boxscoreadvancedv2: The advanced stats box score for a given game

This dataset is remarkably thorough because it is drawn directly from the official NBA website. However, there do exist some null values. This is primarily dependent on the time that the stat was recorded. First of all really old data, about 1990 and earlier, is pretty sparse outside of just the box score. Second, super recent data that maybe hasn't been uploaded or calculated yet is not always present.

Exploratory Data Analysis

This dataset contains an immense amount of information. Most of which is useless to us when we are specifically looking for points. This allows us to greatly narrow our focus to a limited scope. When predicting the amount of points a player will score, the variables we use will fall into just a few categories such as their past performance, their opponents defensive prowess, their teammate's performances, etc. Most of this information is found in the playergamelog endpoint, so we will primarily focus on that.

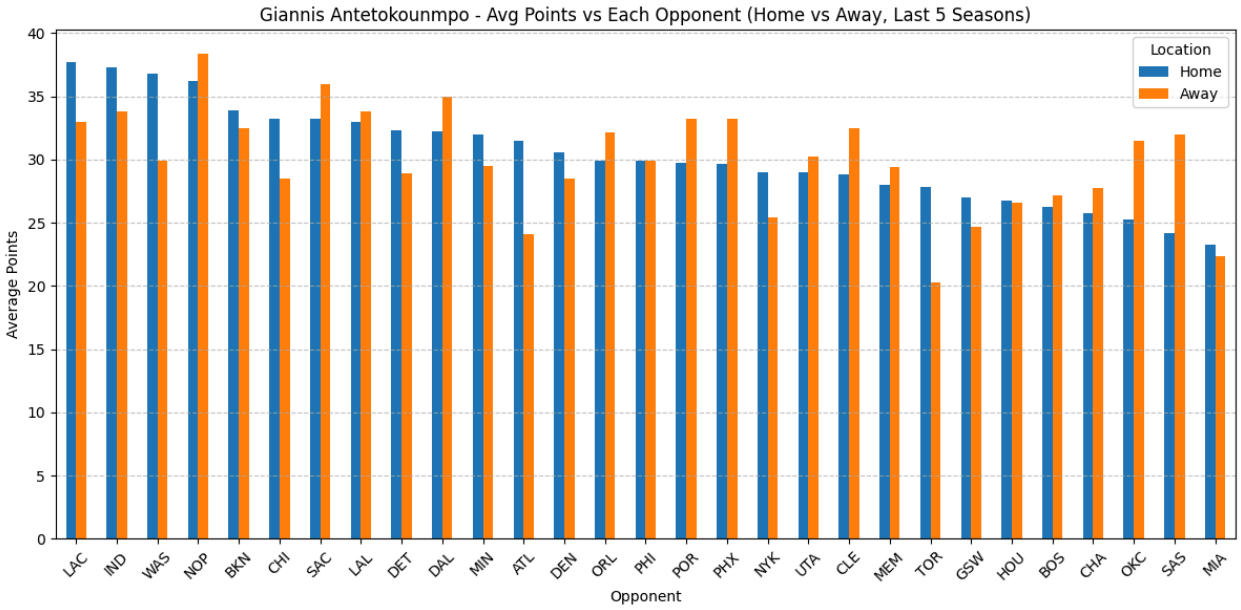
Based on our initial analysis here are some visualizations that we developed as we explored different features:



One feature we were concerned with was opponent difficulty. We chose to represent this with opponent win percentage, but it appears it will not be a good predictor. This will likely be

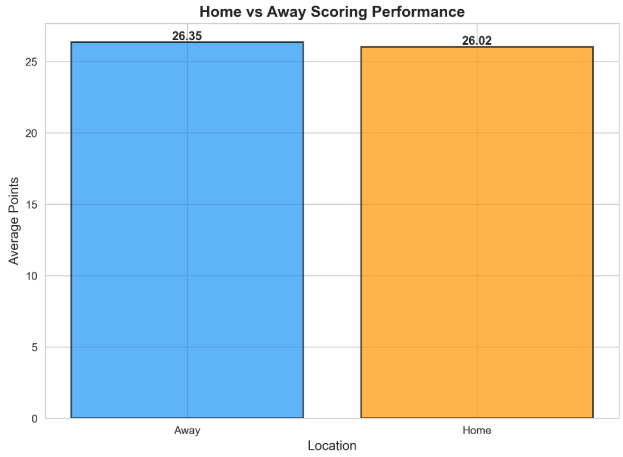
factored into a more compound opponent difficulty feature that we develop as we work on our model.

Another feature we examined was how Giannis performed against each team at home compared to his performance when he played the same team but on their homecourt:

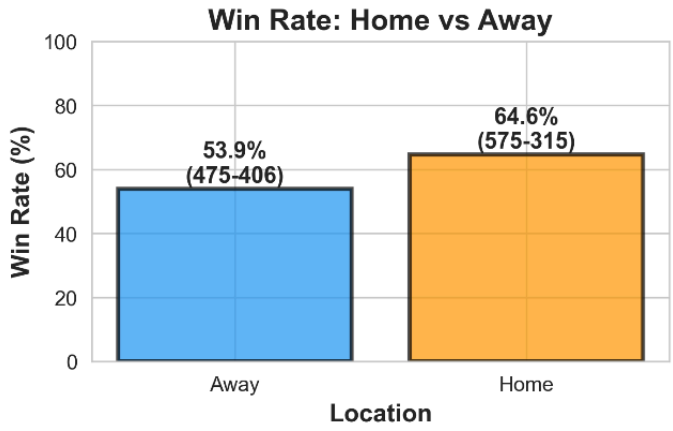


In a lot of the cases, as expected, Giannis performed better against the same team on average at home versus playing the same team away. However, against a surprising number of teams, Giannis performed better on average while being away. Based on this information, it may take some more consideration as to whether or not this should be a highly weighted / considered feature in our training. If a player historically plays better against a certain team when they are away, that could indeed be a powerful feature.

We analyzed 1,771 games (from the 23-24 season) from top scorers to examine home court advantage. Surprisingly, individual player scoring was nearly identical at home (26.02 PPG) versus away (26.35 PPG), a negligible 0.33 point difference.



However, teams won significantly more at home (64.6% win rate) compared to away (53.9%), a +10.7% advantage.



This paradox reveals that home court advantage comes from defensive performance rather than offensive output. Home teams win by allowing fewer opponent points, not by their stars scoring more. Away wins actually require higher scoring (27.72 PPG) than home wins (27.04 PPG), suggesting teams must compensate offensively when lacking defensive advantages on the road. For our prediction model, this indicates home/away location will be a low-priority feature for predicting individual player points, as elite scorers maintain consistent output regardless of venue.