EUROLEAGUE BASKETBALL ANALYSIS AND GAME PREDICTOR

Data Analytics Project
Raúl Fuente - Oscar Colom

MOTIVATION OF THE PROJECT AND PROBLEM STATEMENT

The aim of the project is to analyze, visualize and predict basketball statistics from datasets containing Euroleague data. The purpose are (1) to give an interactive interface where the user can see a clear and clean visualization on the stat she/he wants, (2) to identify the key stats in order to win and give a prediction on a game given 2 teams, and (3) identify and show important trends in Euroleague via a history with some interactive dashboards.

The problem we wish to solve is to give a visual analytics tool for a better understanding of this basketball league, providing key insights on what is needed to win. Also, this tool serves as a way to view Euroleague history in an easy and playable way for curious users, including multiple ways to view data such as line charts, scatter plots, maps, etc.

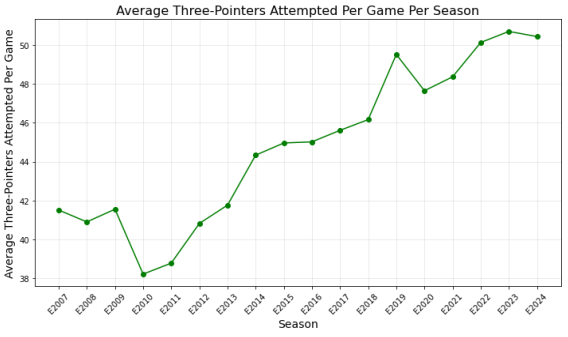
DATASET OVERVIEW

To obtain data for the project, we searched for datasets on Kaggle webpage, we managed to find a very complete set of datasets containing pretty much all the information we wanted (https://www.kaggle.com/datasets/babissamothrakis/euroleague-datasets), and were kept up to date, which was important as we wanted to make predictions on the current season.

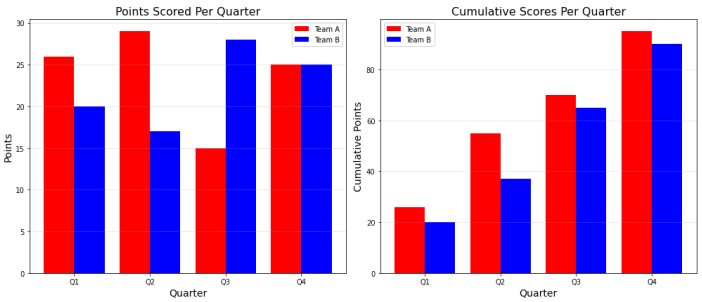
- **eurolague_header**: The rows in this dataset represent each game and include game-related statistics such as date, teams, and stadium.
- **eurolague_players**: This dataset contains statistics per game and season totals for each player, divided by seasons.
- **eurolague_points, eurocup_points**: The rows in this dataset represent each basket (scored or missed), indicating the game, player, type of basket, etc.
- **eurolague_teams**: This dataset contains statistics per game and season totals for each team (similar to **players** but for the entire team.)
- **eurolague_box_score**: Contains performance statistics for each player in each game played, such as points, rebounds, assists, and more."
- **df_merging_teams_cities, teams_updated**: Created by us to enhance visualizations.

Some Exploratory Data Analysis which allowed us to start the project and get ideas. This EDA is the first part of the notebook (EDA_predictive_models) and some dashboards of the Storytelling are:

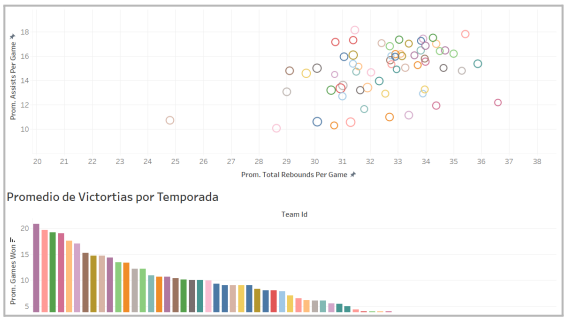
Current trend of the league to shoot more three-pointers:



Analysis per quarters of how each team scored in a game:



Correlation between assists and rebounds along with their impact in wins:



BUSINESS QUESTIONS AND OBJECTIVES

In this project, we aim to address two main questions:

1. **Understanding the evolution of basketball in the EuroLeague:**
   We want to explore how basketball has evolved in the EuroLeague in recent years, identify which actions in a game are most crucial in determining victories, and uncover the key factors that contribute to becoming one of the best players.
2. **Predicting team outcomes with precision:**
   We aim to investigate whether it's possible to predict the result of a game with accuracy based on the data available and gain a better understanding of how these predictions work.

Additionally, we want to provide an interactive interface where users can easily explore and visualize the statistics they are interested in, offering clear and clean visualizations that enhance understanding.

METHODOLOGY

We decided to build some ML models on the prediction of a game and present results on Streamlit, alongside with the explainability of our models, and some data analysis. We also developed a story in Tableau, including trends in the league and a geographical analysis. Deeper explanation in the following sections:

- **EDA and ML Models in Python Notebook**: We did some EDA to get familiar with data, and extract the most correlated features with a team winning a game. More specifically, we created a data frame which was the one that was given to the predictive models. These predictive models are of two types: simple (predict which team wins) and multioutput (predict the scores of each team in the game). For the simple type, we developed a logistic regression model and a random forest model. For the multioutput models, we developed a linear regressor, support vector regressor and gradient boosting. Final part of the notebook include a season simulator (using the logistic regressor), local/global explainability.

Best models (error for points would be around 6.5 = sqrt(43) for each team):

```
Logistic Regression Performance:
              precision    recall  f1-score   support

           0       0.93      0.94      0.93      1014
           1       0.89      0.87      0.88       580

    accuracy                           0.91      1594
   macro avg       0.91      0.90      0.90      1594
weighted avg       0.91      0.91      0.91      1594

Accuracy: 0.911543287327478
MSE: 0.08845671267252196
```

```
Linear Regression - MSE for points_a: 43.962110095673 62
Linear Regression - MSE for points_b: 42.632291905835025
```

➢ **Web-App in Streamlit**: The app consists of three main sections:

- **Season Analysis**: Explore how players and teams perform during a given season. You can also view detailed statistics for specific games.
- **Tops**: Allows search for the top players or teams based on your chosen statistic.
- **Predictions**: View the predictions generated by different models, along with explanations of their local and global interpretability.

➢ **Storytelling in Tableau**: Consists of four dashboards designed to explore the data, uncover insights, and gain a deeper understanding of the EuroLeague.

- **Historical Scoring Trends**
- **Keys to Winning**
- **Informative Map**
- **MVP Candidate**

CONCLUSIONS

We will divide the conclusions into our 2 main sections. The first one, which is “Understanding the evolution of basketball in the EuroLeague”, we can conclude that the most important stat, or at least, the one that teams need to look at it the most, is the defensive rebounds and assists. That is, a team needs its offense to be generous (moving the ball with many passes) and its defense to be focused on getting the rebound to not allow more shots of the opponent. This would be a general analysis on the sport itself, if we look deeper into the trends of the Euroleague, we can see that currently teams are scoring more points than ever, shooting year by year more shots, more thee-pointers, and therefore, playing more possesions per game. This last remark is important to see other trends, like the increase on the speed of the game, and it being less stationary. Finally, conclusions come easy looking at players or teams tops, e.g. Mike James is currently the top scorer but did not need many games to do so, which is impressive.

The ML section of the predictors, we saw that the simple models had high accuracy because it may be easier to predict a binary target (who wins), while we could not reduce the MSE for the multioutput models from 42-43. The explainability is pretty visual and self-explanatory.