Our datasets come from Kaggle Competition named “NBA games data”. There are 5 CSV files in total, and three of them were used which are “games.csv”, “games\_details.csv” and “players.csv”. Aiming to analyze for our research questions, we did some data cleaning and preprocessing on the datasets. The Fig1 and Fig2 are the screenshots of our annotated dataset. The marked columns are added columns which are necessary for our analysis. The START\_POSTION is the significant factor in our research questions and WIN is the only response variable through the whole project.

There are no previous-published solutions to this problem. There are several published articles on making predictions on NBA games. However, these articles are using different predictors in different aspects to predict win-lose of NBA games. For example, the performance of the whole team is considered in one article while in our project, the performance of the single player is the factor we are focusing on. The predictions on NBA games are also common in the gambling industry while different factors are analyzed.

We did not have certain expectations going into the project since the results of sports games could be considered as one of the “most unpredictive” things. The unexpected winners were not rare to see in the history of sports. The predictors we used in our project are just a small part of all the factors on the results of group sports games, especially we are focusing on the benches which might be less impacts on the other factors like starters.