mention specific papers & discuss

have found no significant difference in pay between white and minority players. However, these studies often rely on traditional performance statistics, which may not capture the full extent of a player's contribution to their team. By using Statcast data, this paper aims to provide a more accurate assessment of the impact of race on salary differentials in MLB.

Overall, this paper seeks to contribute to the ongoing debate on salary discrimination in professional sports and provide insight into the role of player performance data in understanding the factors that contribute to pay differentials.

2 Data Section

The data starts with collecting players' value data from Baseball Reference. This data source included each players team, WAR, and salary for that year. For this study the years 2015-2022 were chosen because Statcast first started tracking stats in 2015. The year 2020 is not included since it was only a 60 game season due to COVID-19 and normally seasons are 162 games. This time frame is more recent that any other study on this topic and covers a large enough period that would allow some players to sign multiple contracts. ✓

WAR is a relatively new performance metric that is considered to be one of the most robust measurements of a player's productivity. WAR measures how many more games a player's team wins than the team would have won if it had the league-average player in the stead of the observed player. Because it takes into account both defensive and offensive production, it will serve as an exceptionally strong indicator of a player's value.

relationship between these entities?

The next source of data is Statcast which is powered by Google Cloud. From Baseball Savant we are able to add variables to each observation which include traditional stats adn Statcast-powered stats. The differences between these two categories can be seen in table 1. To learn more about how these are calculated go to Baseball Savant's website where they have clearly defined every statistic. The Statcast data is mainly powered by player tracking data which allows models to calculate a player's expected chance of making an individual play. Which can then be used to calculate expected runs created for offensive plays.

reword for clarity

Does Baseball Savant just publish Statcast figures online?

We are then able to add robustness to the data by collecting demographic data for all players. This includes birth country and the date of when the player debuted in the major leagues. This allows us to create our international dummy variable and service time variable. The service time variable simply subtracts the year the player debuted from the year the stats were produced.

has very particular meaning in econometrics -- maybe instead say something like "able to supplement player statistics with ..."

The variable indicating a player's race was created by the author. Race was determined using each player's roster picture as well as referring to his name. In the final dataset, players are classified as white or non-white.

Unlike previous studies, one of the main independent variables in the study will be Service Time. Previous studies have used age or contract length to control for 'experience'. I will discuss more in the results section, but in reference to my data, the use of service time vastly improved the strength of my model. I also chose to use the per year value of a contract in contrast to the average annual value of a contract because some contracts will be structured differently than others. For example, a five-year contract worth $5 million may be divided as follows: Year 1, $1.1 million; Year 2, $0.9 million; Year 3, $1 million; Year 4, $1.05 million; Year 5, $0.95 million. I chose to incorporate this structure of salary representation in order to control for teams that create irregular values throughout the length of the contract. However, like previous studies, the dependent variable will be the natural logarithm of a player's annual salary in a given year. This will be done to track percentage changes associated with changes in the independent variable instead of dollar changes. It is more logical to track percentage changes because dollar changes are more significant at lower salary levels than at higher salary levels, whereas percentage changes are constant.