Stats 68 Spring 2024 Mini Project 2

Teddy Dong

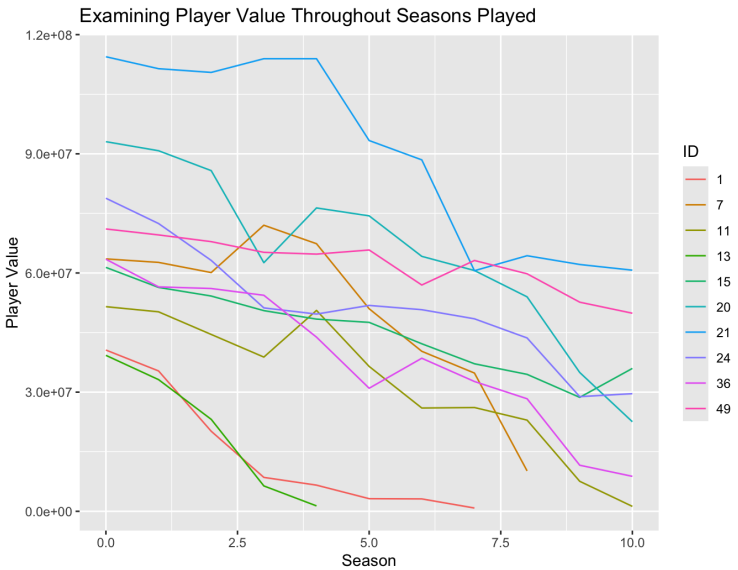
Due 6/12/2024

1 Executive Summary

- Out of the 11 seasons worth of data, the data tracks 50 different players providing information on 7 variables ranging from 'Player Value' to 'Sport' played.
- Missing values are present in the baseline data for the categories weekly training data (3 missing), age (3 missing), injuries (4 missing), location (1 missing), gender (3 missing), and sport which has one missing.
- The highest player value of 133,852,045 in the baseline data is associated with a 28-year-old Male Boxer from Germany (ID 38) and the lowest player value is associated with a 28-year-old Female Basketball player (ID 16) from the United Kingdom.
- There is visual evidence that at baseline there may be a positive relationship between number of injuries and player value, a slight positive relationship with weekly training hours, and a negative relationship with age.
- Based on ten randomly selected individuals in this dataset, player value is seen to decrease as more seasons are played.

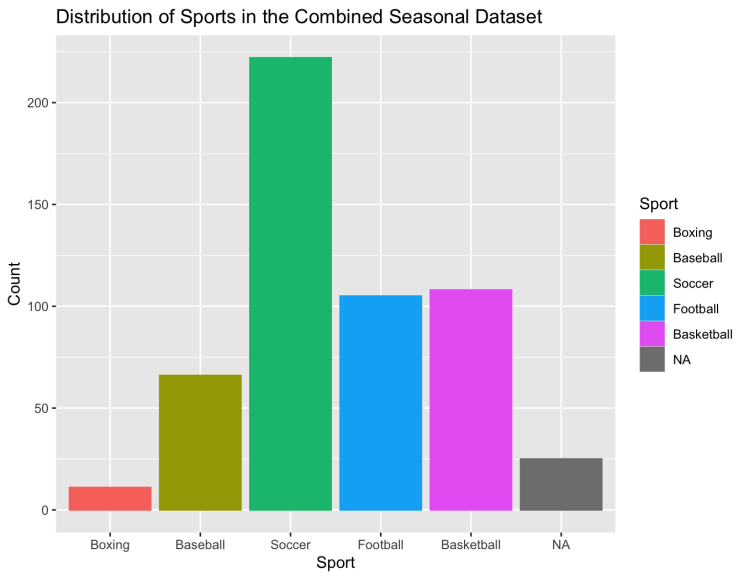
2 Figures

2.1 Figure 2.1



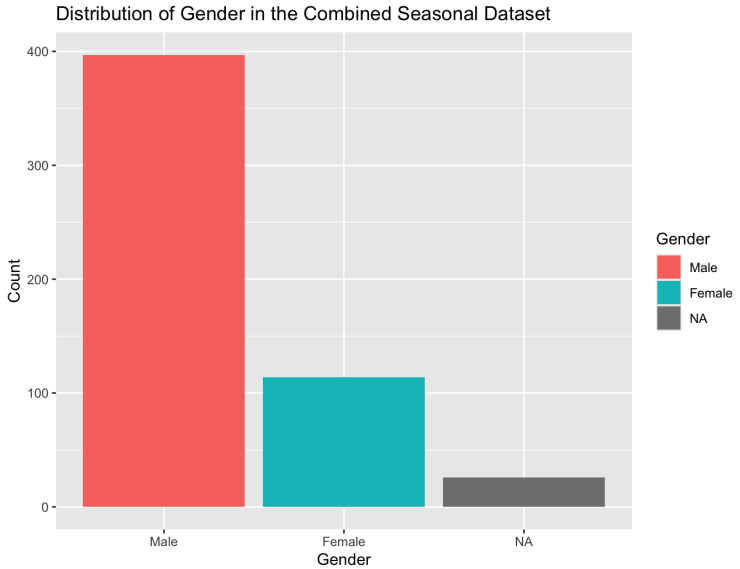
Season and player value were examined to see what effects longevity in a players respective sport would have on their value as a player. A downward trend was noted for all 10 randomly selected players indicating that an increase in seasons played may lead to decreased player value.

2.2 Figure 2.2



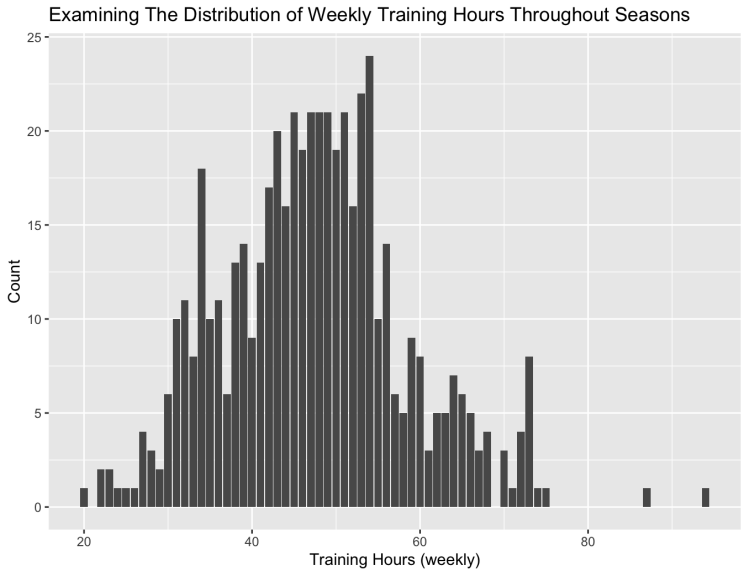
The distribution of sports played was examined to see which were most popular. Soccer has the highest count, with another level "Tennis" completely missing from the graph indicating there were no Tennis players in the dataset.

2.3 Figure 2.3



The distribution of gender was examined to see if there was going to be an equal of data from both male and female athletes, though that is evidently not the case as seen with this graph. The ratio between male and female athletes has stayed the same compared to 37 male, 10 female, and 3 NA makeup from the baseline data.

2.4 Figure 2.4



The distribution of training hours over all seasons was examined to see how many hours on average athletes are putting in each week to ensure that they are in their best shape for competition. The distribution is fairly uniform with the main peak around 50 hours of training per week, though there are outliers with some athletes training more than 80 hours per week.

3 Next Steps

One piece of information that would've made data analysis more efficient would've been to provide the exact values for Gender, Sport, and Location inside of the datasets, instead of using numbers as a key for different values in each category. Another piece of information that would've been interesting to look at is the years that each season is associated with. The major drops in player value could be concretely explained for if perhaps the later seasons occurred during the global pandemic, causing many sports to be put on hold and as a result, player values across all sports could drop. Data similar to this for other sports such as Badminton, Swimming, Rowing, etc. would also be interesting to examine.

4 Data

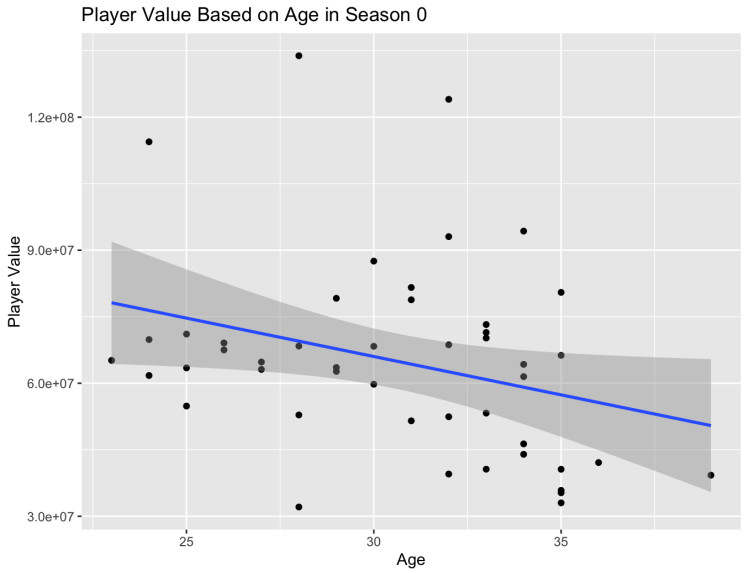
In this section you should describe the data. This can also include summary statistics or general overview of the data.

Out of the eight variables provided as part of the dataset, the variables pertaining to id, location, gender, and sport are categorical variables. The other four variables player value, training hours, age, injuries are numerical variables. As the seasons progressed, there were some cases where ID's stopped appearing under datasets for later seasons, indicating that they may have stopped competing in the sport that they played. Looking at the combined data set with information regarding all eleven seasons, the maximum player value

remained unchanged from the baseline data value of 133,852,045, but the minimum player value became 524,820 and the median player value was 43,846,469. Weekly training hours had a minimum of 20 hours, maximum of 94 hours, and median of 47 hours. The ages of participating athletes ranged from 23 to 39 (though it does have to be mentioned that these values are exactly the same as the baseline data minimums and maximums for age). There was a minimum of 2 injuries sustained, a median of 14, and a maximum of 30. Spain was the most prominent location and there also was surprisingly no data regarding Tennis players collected throughout the seasons, even though it was provided as a possible sport for the category.

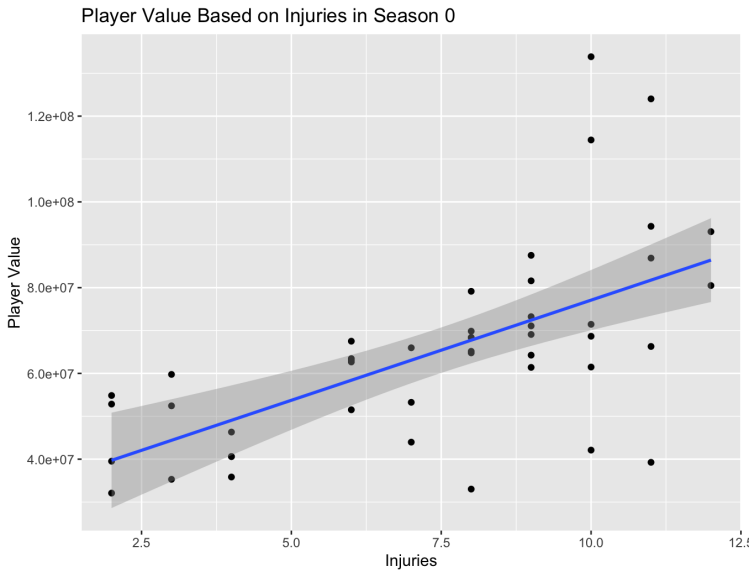
5 Appendix

5.1 Figure A.1



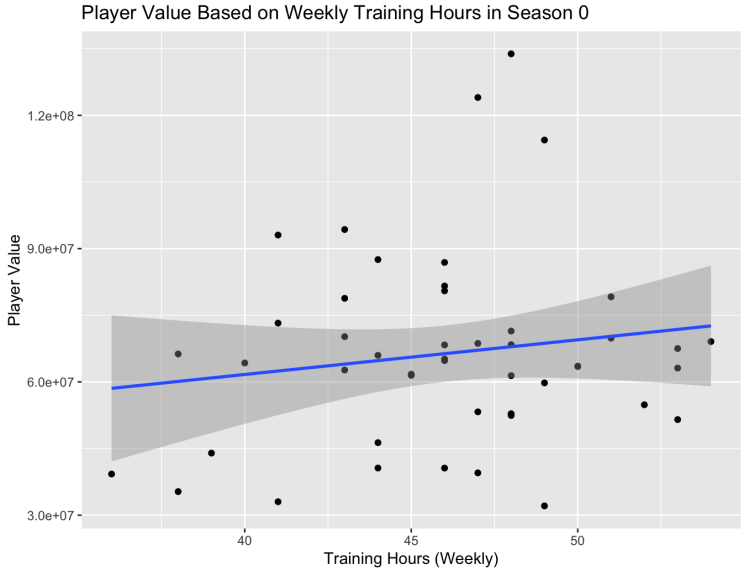
This graph was created to examine the impacts of age on player value. The least square regression line for this graph has a negative slope, showing that for the baseline data, older players were generally valued less.

5.2 Figure A.2



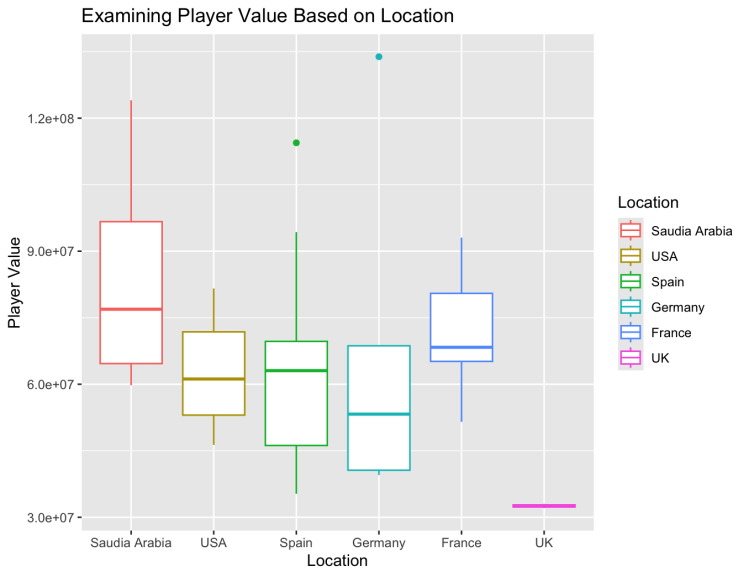
This graph was created to see if there is any relation between number of injuries sustained and player value. Oddly enough, the least squares regression line has a positive slope, indicating that at least from the baseline data, a player that has been injured more times has a higher player value.

5.3 Figure A.3



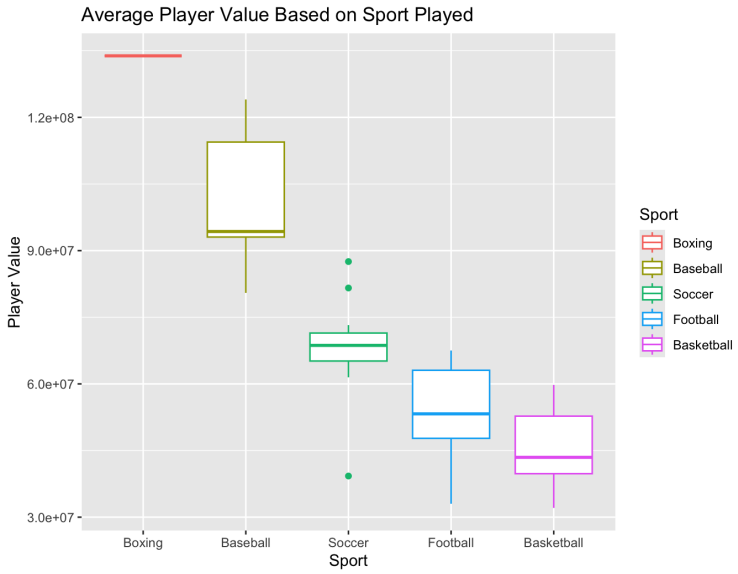
This graph was created to examine the effect that a player's number of weekly training hours had on their value. There is a very slight positive relationship, showing that there are benefits to spending more time training in one's respective sport to improve their value as a competitor.

5.4 Figure A.4



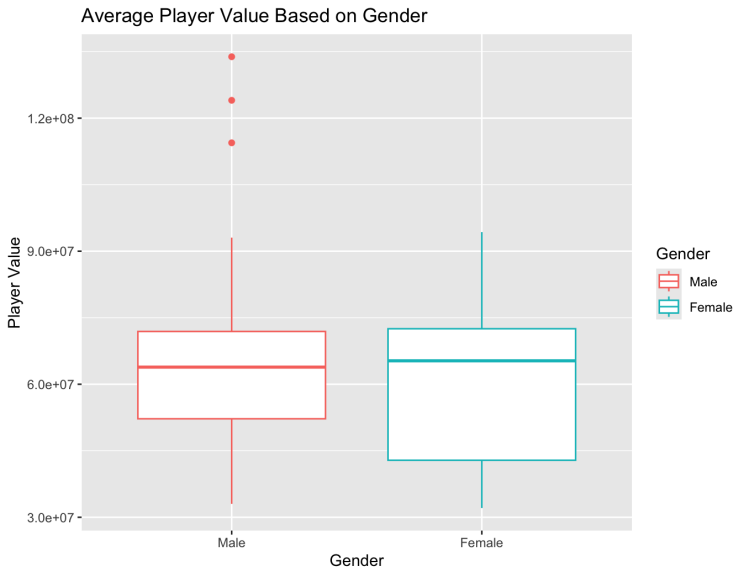
Player value based on location in season 0 was investigated to see which locations had the highest average player value. Based on the graph, we can see that Saudi Arabia generally has the highest player value averages, with the UK having the lowest average player value.

5.5 Figure A.5



Player value based on sport in season 0 was investigated to see which sports had the highest player values on average. As seen with the graph, boxing has the highest average player value while basketball has the lowest average player value. Baseball also has a fairly high Q3, but the Q1 value is much lower likely being the reason why the median is so low as well.

5.6 Figure A.6



This graph investigates player value based on gender to see if there are any major discrepancies between baseline player value in season 0. The median for both genders is fairly even, with the median for female being slightly higher, though females have a much lower Q1 player value and males have much higher maximum player values.