STAT 223

Laboratory Activity #1 by Akbota Serikkyzy

MADfun

The function MADfun calculates the Mean Absolute Deviation (MAD) of a given vector. The R code for the function:

```
MADfun <- function(vec) {
  if (length(vec) == 0) {
    stop("Input should not be empty. Please try again!")
  }
  if (!is.numeric(vec)) {
    stop("Input should be number. Please try again!")
  }

  # Calculation
  mean = mean(vec)  # Mean of vector
  mad <- mean(abs(vec - mean))  # 1st absolute diff from mean, 2st mean of absolute diff
  return(mad)
}
```

First, this function ensures that the input vector is not empty and numeric. After this, it calculates the mean of the dataset and finds absolute differences between each data point and the mean. Finally, MAD is calculated by meaning absolute difference.

The function worked well during testing, returning accurate results for vectors and catching errors for wrong inputs.

About MLBTeams.csv

MLBTeams.csv is a collection of data about MLB from 1995 to 2021. The following are my summary about the dataset.

Decade	Mean Win%	Mean HR	Mean BA
1995-2004	0.500	174	0.267
2005-2014	0.500	161	0.260
2015-2021	0.500	177	0.250

Table 1: MLB Team Performance Metrics by Decade (1995-2021)

The winning percentage (Win %) accoring to table 1 has remained consistent throughout all decades, indicating that teams have maintained a stable level of success.

When looking at home runs (HR), we see a slight drop between 2005 and 2014, followed by an increase from 2015 to 2021. This suggests that teams may have adjusted their strategies during that time, possibly focusing less on power hits for a period before returning to it later.

Meanwhile, the batting average (BA) has steadily declined over the decades. This could indicate a shift in offensive strategy, where teams rely less on pure hitting and more on other factors, such as home runs. Another possible reason for this decline is that pitching and defensive strategies have improved, making it harder for batters to maintain high averages.

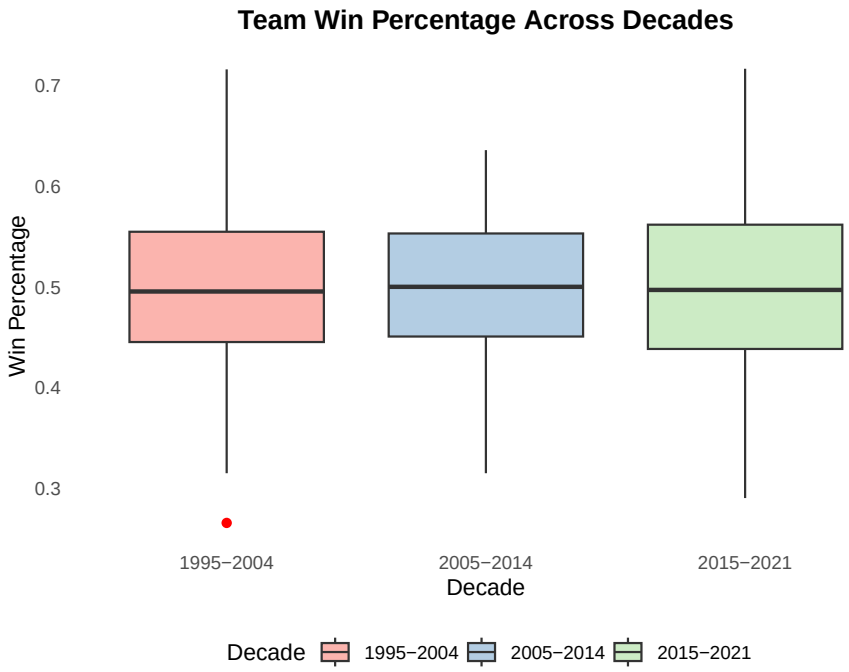

Figure 1: Team Win Percentage Across Decades

From the box plot 1, we can see that the median percentage of wins has remained stable over the decades, indicating that the overall performance of the teams has not changed significantly. There is a slight variation in the spread of the win percentages, but no drastic changes are found.

Between 1995 and 2004, the red dot represents an outlier, meaning that certain teams performed much worse than the others during that period. However, from 2005 onward, there are no noticeable outliers, suggesting that teams have become more consistent in their performance.

In general, these trends reflect how the game of baseball has evolved over time, with teams adapting their strategies based on changes in player performance, analytics, and league-wide trends. The box plot shows that while the median win percentage has remained stable over the decades, the reduction in outliers starting from 2005 suggests that the league has become more competitive, with fewer teams performing exceptionally poorly.

BAStd

BAStd (Standardized Batting Average) is a way to adjust a team's batting performance to compare it more fairly with other teams in the league. Positive values in BAStd show which teams are performing better than the average league performance. Negative values show which teams that are falling behind the league's overall performance. Zero indicates teams that are performing at the league's average.

In order to find the BAStd, we subtract the mean of BA from each team's BA, then divide that by the standard deviation of BA.

The average BAStd is 1.608, which means that teams perform better than the league average in terms of batting.

DivWin

DivWin is a variable that is 1 if the team won their division that season and 0 if not.

Statistics	Value
Min.	0.0000
Median	0.0000
Mean	0.2015
Max.	1.0000

Table 2: Statistics for Division Winner

DivWin Value	Count
0 (Non-Winner)	642
1 (Winner)	162

Table 3: Division Winner Counts

From this table 2, we can notice that the median value is 0, while the mean is 0.2015. This suggests that most teams did not win their division, but there is a small proportion of teams that win their division. The fact that the mean is higher than the median indicates that it is largely right skewed.

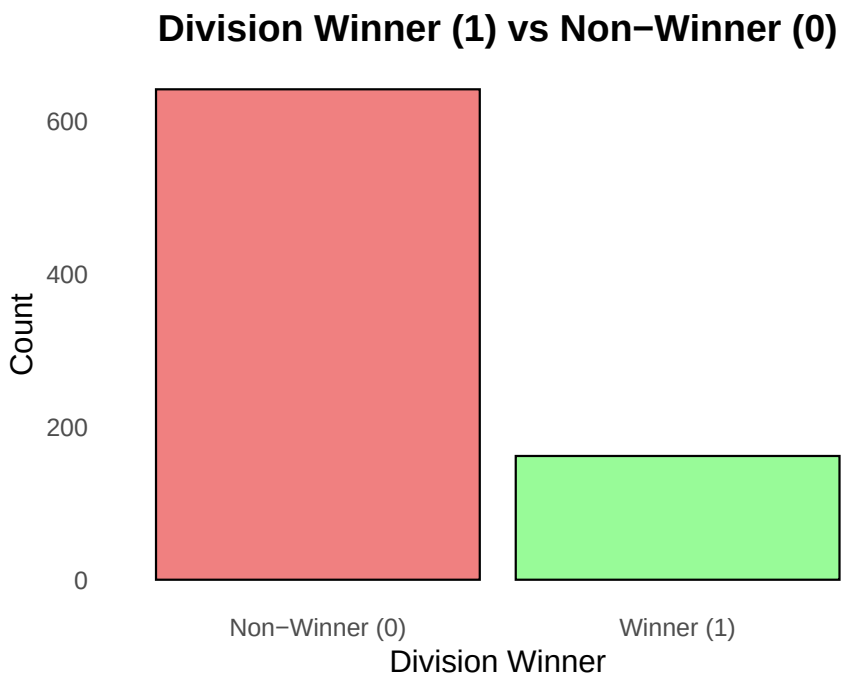

Figure 2: Division Winner vs Non-Winner

Furthermore, as shown in table 3 and figure 2, we observe that nonwinners more than winners approximately four times, which indicates that winning teams are much less.

WAR

Wins Above Replacement is an advanced metric that combines different information about player performance, with the intent to quantify how many wins a team gets starting that player over a "replacement level" player.

Statistics	Value
Min.	-4.50
Median	18.20
Mean	18.48
Max.	44.40
Std	8.70

Table 4: Statistics for Wins Above Replacement

As shown in table 4, the median and the mean are quite similar, Wins Above Replacement values is stable. However, when we look at the standard deviation, it is relatively high, meaning that there is a significant amount of variation in the WAR.

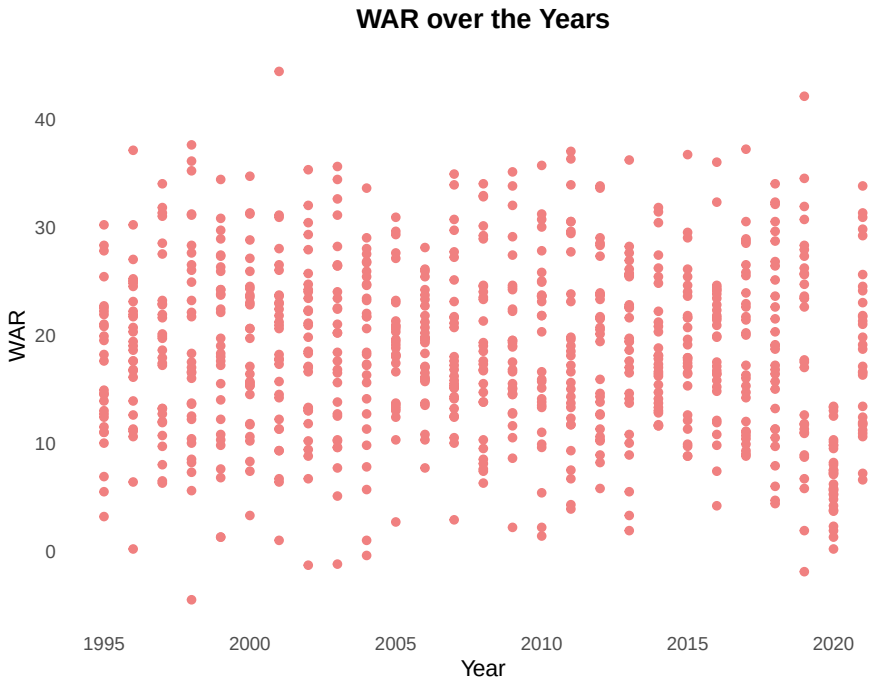

Figure 3: WAR over the years

Looking at Figure 3, the scatter points appear to be similar over each year, indicating that there is not much variation over the years. We can observe that the minimum WAR value occurred between 1995 and 2000, while the maximum value was seen between 2000 and 2005. Another notable observation is the year 2020, where the WAR values are relatively low, with the highest being around 15. Considering that this was during the COVID-19 period, it is possible that the pandemic influenced these results, or there may have been other factors influenced.

WinPct

WinPct is the team's winning percentage that season, and Division is the division that team plays in. There are six divisions in Major League Baseball across the two leagues: AL East, AL Central, AL West, NL East, NL Central and NL West.

Division	Mean	Median	Std	Variance	Min	Max	Q1	Q3
AL Central	0.4850	0.4877	0.0766	0.0059	0.2654	0.6944	0.4259	0.5432
AL East	0.5122	0.5247	0.0813	0.0066	0.2901	0.7037	0.4537	0.5741
AL West	0.5108	0.5185	0.0714	0.0051	0.3148	0.7160	0.4630	0.5617
NL Central	0.4926	0.4877	0.0697	0.0049	0.3167	0.6481	0.4444	0.5432
NL East	0.4984	0.4938	0.0695	0.0048	0.3333	0.6543	0.4414	0.5480
NL West	0.5028	0.5062	0.0700	0.0049	0.3148	0.7167	0.4552	0.5556

Table 5: Summary Statistics of WinPct by Division

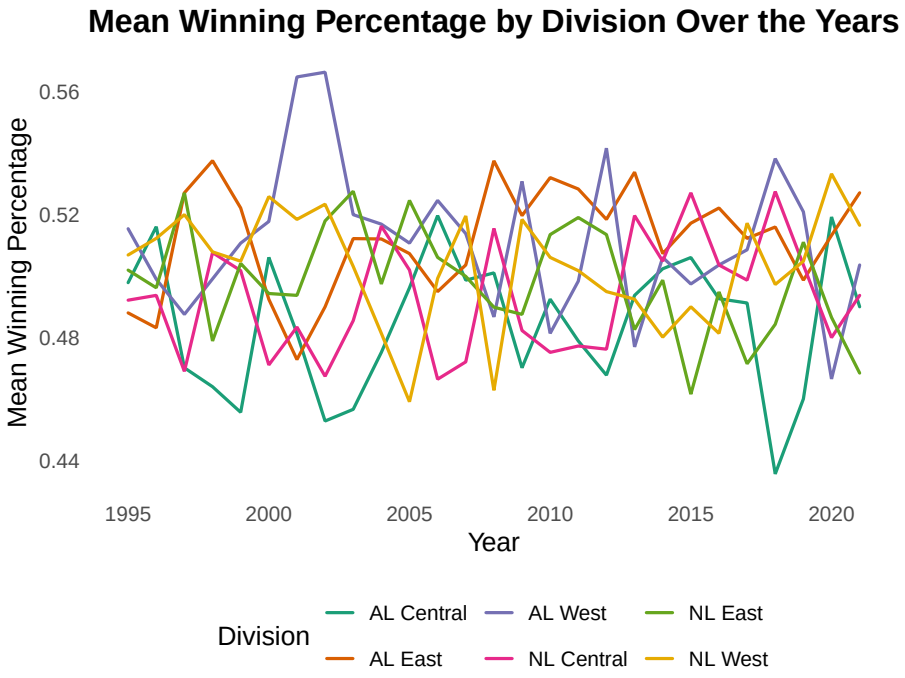

Figure 4: Mean Winning Percentage by Division Over the Years

In Table 5, it is visible that the best performing divisions are AL East and AL West, with means of 0.5122 and 0.5108. In contrast, AL Central has the lowest average winning percentage at 0.4850, suggesting relatively weaker performance.

Additionally, the standard deviation is highest for AL East with 0.0813, indicating greater variability in team performance. In comparison, NL Central and NL East have lower Stds around 0.0695–0.0697, suggesting more consistent performance.

Looking at the minimum and maximum values, AL Central has the lowest minimum winning percentage with 0.2654, while AL West and NL West share the highest maximum (0.7160 and 0.7167). This implies that while AL Central has had some of the weakest performing teams, AL West and NL West divisions seem to have strong performing teams.

When comparing quartiles, AL East and AL West also have relatively higher Q3 values (0.5741 and 0.5617), meaning the 75th percentile of teams often perform well. In contrast, AL Central's and NL Cenral's Q3 is 0.5432, which is lower compared to other divisions.

Overall, this summary suggests that while all divisions show some range in performance, AL East and AL West tend to be stronger overall, and AL Central is generally weaker.

In Figure 4, we observe that the purple line, representing the AL West, shows the best performance overall. While AL West has demonstrated strong results over the decades, it also faces variance. This inconsistency highlights the unpredictability in the division. Despite these fluctuations, AL West remains one of the top performing divisions.

However, it is important to note that none of the divisions show consistent performance over time. While AL West delivers the best overall results, its variance suggests that performance is not steady.

On the other hand, AL Central shows the lowest overall performance, as we saw earlier in the table 5 statistics. This division consistently records some of the weakest results, but like AL West, it also faces variance.

In fact, across all divisions, the variance is mostly between 0.48 and 0.54, indicating that the differences in performance between the divisions are not drastic, but still significant. The similar levels of variance across divisions suggest that while no division shows perfect consistency, the fluctuations in performance are relatively uniform across MLB divisions.

HR

HR is the number of home runs the team hit that season, and HRA is the number of home runs allowed by the team, the number of home runs that opposing teams hit while playing the team. Teams better at hitting will have high HR and teams better at pitching will have high HRA.

Statistics	Value
Min.	51.0
Median	168.0
Mean	169.9
Max.	307.0
Std	40.8

Table 6: Statistics for Home Runs Hit

Statistics	Value
Min.	62.0
Median	170.0
Mean	169.9
Max.	305.0
Std	34.2

Table 7: Statistics for Home Runs Allowed

Looking at Tables 6 and 7, we can observe that the mean for both HR and HRA is nearly identical, at 169.9. This suggests that teams hit and allow a similar number of home runs each season.

However, when examining the standard deviation, I notice both HR and HRA have high variability. The range between the min and max values for both statistics is quite large, with the min values being approximately five times less than the max values. This indicates that while most teams have moderate home run totals, there are some teams that either hit or allow an exceptionally high number of home runs compared to the majority.

Notably, HR shows even greater variance with a standard deviation of 40.8, which is higher than the standard deviation of 34.2 observed for HRA. This suggests that teams' offensive capabilities in terms of home runs are more spread out compared to home runs allowed.

In Figure 5, it is noticeable that there is a positive correlation between home runs hit and home runs allowed. This means that teams that are better at hitting home runs tend to also allow more home runs. In other words, teams that are good at hitting also good at pitching.

However, it is important to keep in mind that this correlation does not necessarily mean that hitting more home runs causes teams to allow more home runs. There are many other factors that influence both HR and HRA, such as pitching quality, team defense, and stadium size. So, while there is a relationship between HR and HRA, it might not be a direct cause-and-effect connection.

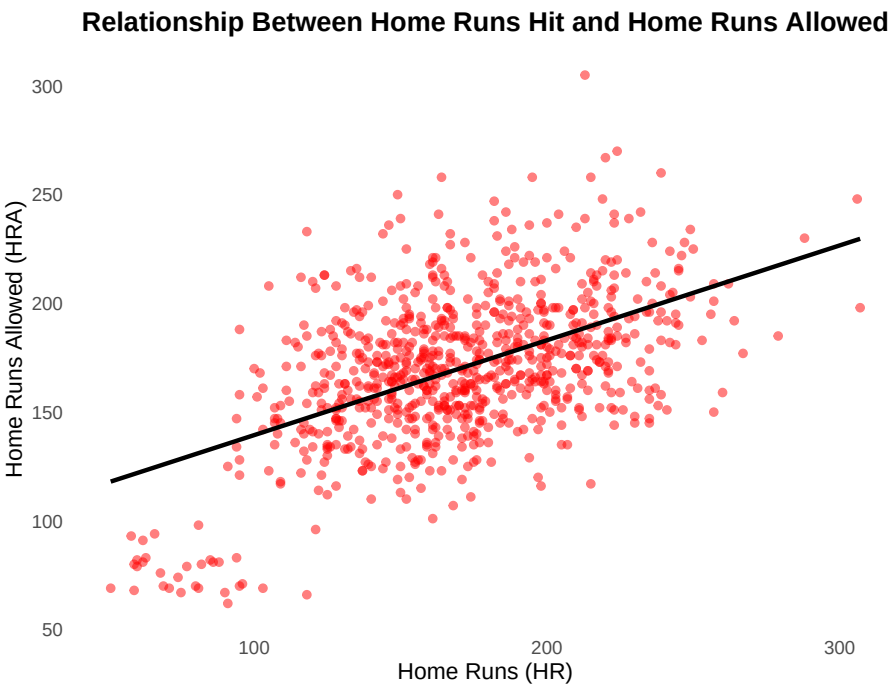

Figure 5: Relationship Between Home Runs Hit and Home Runs Allowed