Analyzing Salary and Peformance Metrics of the New York Yankees

STAT 184 Fall 2024 Final Project

Kyle Gilbert, Cole Kvasnak, Dylan Van Berkel

2024-12-18

Overview

When Michael Lewis (Lewis) published his 2003 book *Moneyball*, he changed the game of baseball forever, bringing the power of statistical analysis into the spotlight. Suddenly, it wasn't just about home runs or batting averages – it was about uncovering hidden values through data.

Building on this legacy, our project dives into the intersection of money and performance, analyzing how salary allocations align with player performance metrics in America's pastime. Focusing exclusively on 21st century New York Yankees players - the cornerstone of Major League Baseball's (MLB) most valuable franchise - we aim to uncover insights behind the numbers that define success for the Bronx Bombers.

Our research intends to answer the following questions:

- What does the distribution of salaries for the New York Yankees look like?

- Which offensive and defensive metrics correlate the most with a player's salary?

- What is the salary breakdown across different positions? In other words, which positions get paid the most and least?

- Which players provided the most and least value per dollar they were paid? In other words, who were the top over performers and top under performers?

Datasets

For our analysis, we will be using the *Baseball Databank* dataset found on Kaggle (Sports). This dataset has been adapted from the Sean Lahman Baseball Dataset (Lahman). Sean Lahman is a well-respected journalist and data analyst with a passion for baseball history.

The Sean Lahman Baseball Dataset is a collection of historical baseball data, dating back to professional baseball's inception. The *Baseball Databank* covers the years 1871 - 2015. Sean Lahman has compiled the dataset to provide baseball researchers, analysts, and fans with a comprehensive and easily accessible database for historical baseball data (Lahman).

The data provided in *Baseball Databank* abides by the FAIR principles. FAIR ensures findability: the ability to navigate to data of interest easily. The *Baseball Databank* consistently employs the camelcase naming convention for its twenty-two files. The initial word in a filename indicates the main category its data belongs to, and subsequent words describe further classifications. As a result, parsing the databank for a specific data subset is straightforward. In accordance with FAIR principles, the *Baseball Databank* dataset has enhanced ease of access. Readily available on Kaggle (Sports) via a shareable link, this dataset may be shared or adapted according to its Creative Commons Attribution-ShareAlike 3.0 Unported license.

Baseball Databank stores its data in the customary CSV format type – satisfying the interoperability FAIR principle. The corresponding Kaggle Data Card page for this dataset contains a thorough background for the data collection, a description of the tables, and an explanation of each column within the tables. These inclusions improve comprehension of the data, which is a key attribute of the reusability FAIR principle.

In our analysis, we will be using the following tables from the *Baseball Databank*:

- Master: player names, DOB, biographical info. Cases are individual players.

- Batting: batting statistics. Cases are individual players on a certain team in a given year.

- Pitching: pitching statistics. Cases are individual players on a certain team in a given year.

- Fielding: fielding statistics and player positions. Cases are individual players on a certain team in a given year.

- Salaries: player salaries. Cases are individual players on a certain team in a given year.

- Teams: team statistics and ballpark data. Cases are individual franchises in a given year.

To supplement the *Baseball Databank* dataset, we will also be using a smaller *Consumer Price Index* dataset. This dataset provides historical consumer price index (CPI) data.

The *Consumer Price Index* dataset is sourced from the Bureau of Labor Statistics (Labor Statistics). The dataset can be accessed directly from the Bureau of Labor Statistics website. For our specific case, we have downloaded the CPI data for the years 2000 - 2015. The Bureau of Labor Statistics provides public CPI data to ensure citizens, organizations, and policymakers have access to a reliable measure of economic conditions.

The CPI can be used to gauge inflation, which we use to adjust player salaries for inflation. The details of adjusting the players' salaries for inflation are discussed under Methodology.

The *Consumer Price Index* for All Urban Consumers (CPI-U) data also complies with the FAIR principles. Regarding findability, the *Consumer Price Index* dataset has variable file names that specify the date and time of data extraction. Moreover, the dataset is openly accessible on a government platform through a shareable link. The *Consumer Price Index* data translates well to R Studio and other statistical software because its file type is comma-separated values (CSV). Researchers, statisticians, and others can understand and reuse this data with help from the provided series title, area, item, and base period descriptions.

In our analysis, we will be using the following table from the *Consumer Price Index* dataset:

- CPI: consumer price index. Cases are represented by a specific year.

Background

Baseball statistics provide a way to quantify an individual player's or team's success by comparing them to other players and teams in the league. In this section, we will go over the statistics we have chosen to focus our analysis on, and provide the mathematical definition for each.

There are a variety of statistics used in analysis of professional baseball, ranging from as simple as H and HR to advanced sabermetrics like $wOBA$. Our analysis will be focusing on a few of the statistics we find most effective in quantifying a player's success, broken up into two primary categories: batting statistics and pitching statistics.

Batting Statistics

For batting statistics, our analysis will focus on three metrics we consider to be the most indicative of a player's success at the plate: wins above replacement WAR , batting average BA, and normalized on base plus slugging $OPS+$.

WAR is an advanced metric that essentially measures how many wins an individual player contributes to their team with respect to a replacement level player.

Our calculations are based on Baseball Reference's definitions ("WAR Explained") for WAR (aka $bWAR$). Metrics such as WAR can be calculated differently depending on the use case.

According to Baseball Reference's definition of $bWAR$ for a position player ("Position WAR Calculations and Details"):

$$bWAR = \frac{RC - RC_r}{a}$$

where RC denotes runs created, RC_r denotes replacement level runs created, and a denotes a constant for the average number of runs per win. Baseball Reference approximates $= 10$ in their calculations.

RC can be computed using the following formula:

$$RC = \frac{(H + BB) \times TB}{AB + BB}$$

where H denotes hits, BB denotes walks, TB denotes total bases, and AB denotes at bats.

TB is calculated using the following formula:

$$TB = H + 2 \times X2B + 3 \times X3B + 4 \times HR$$

where $X2B$ denotes doubles, $X3B$ denotes triples, and HR denotes home runs.

Replacement level runs created, RC_r can be computed using the following formula:

$$RC_r = b \times RC_l$$

where RC_l denotes the league average runs created and $\boldsymbol{\beta}$ denotes a constant for what replacement level is considered to be. Baseball Reference approximates $b = 0.8$, meaning that a replacement level player produces roughly 80% of the league average runs created.

Batting average measures the proportion of at bats, AB, where a player will get a hit, H. Batting average, BA can be calculated using the following formula:

$$BA = \frac{H}{AB}$$

On base plus slugging, OPS is the sum of two other baseball statistics:

$$OPS = OBP + SLG$$

where OBP denotes on base percentage and SLG denotes slugging percentage. These metrics can be calculated using the following formulas:

$$OBP = \frac{H + BB + HBP}{AB + BB + HBP + SF}$$

$$SLG = \frac{TB}{AB}$$

where HBP denotes hit by pitches and SF denotes sacrifice flies.

Normalized on base plus slugging, $OPS+$, is essentially OPS normalized to the league average. An $OPS+$ of 100 denotes the league average. $OPS+$ can be computed with the following formula:

$$OPS+ = \frac{OPS}{OPS_l} \times 100$$

where OPS_l denotes the league average OPS.

Pitching Statistics

For pitching statistics, our analysis will focus on the three metrics we consider to be the most indicative of a player's success at the mound: wins above replacement (WAR), earned run average (ERA), and walks and hits per inning pitched (WHIP).

WAR is calculated differently for pitchers than it is position players. According to Baseball Reference's definition of WAR for a pitcher ("Pitcher WAR Calculations and Details"):

$$bWAR = \frac{(RA9_r - RA9_a) \times \frac{IP}{9}}{a}$$

where $RA9_a$ is adjusted runs allowed per 9 innings, $RA9_r$ is replacement level runs allowed per 9 innings, IP is innings pitched, and a denotes a constant for the average number of runs per win. As previously mentioned, Baseball Reference assumes $a = 10$ for their calculations.

Note: Baseball Reference's official definition ("Pitcher WAR Calculations and Details") of $bWAR$ for a pitcher also includes a leverage factor in the computation. The leverage factor essentially scales a pitcher's WAR differently depending on if they are primarily a relief pitcher or a starting pitcher. Due to the lack of that information in our dataset, and for overall simplicity, our computation will be leaving this factor out.

$RA9_a$ can be computed using the following formula:

$$RA9_a = \frac{RA}{IP} \times 9 \times PPF$$

where RA denotes the number of runs a pitcher allows and PPF denotes the player's home ballpark pitching park factor. The pitching park factor essentially weighs some ballparks differently, due to the inconsistent size of field of play among MLB ballparks.

$RA9_r$, or replacement level $RA9$, can be calculated using the following formula:

$$RA9_r = c \times RA9_l$$

where $RA9_l$ is the league average $RA9$ and c denotes a constant for what replacement level is considered to be. Baseball reference approximates $c = 1.2$, meaning that a replacement level pitcher allows roughly 20% more runs per 9 innings than the league average.

Earned run average, ERA is the average number of earned runs a pitcher will allow in 9 innings. ERA can be calculated using the following formula:

$$ERA = \frac{ER}{IP} \times 9$$

where ER denotes earned runs.

Walks and hits per inning pitched, $WHIP$ is average number of walks and hits a pitcher will allow in an inning. $WHIP$ can be computed using the following formula:

$$WHIP = \frac{BB + H}{IP}$$

where BB is total walks allowed and H is total hits allowed.

For more information on baseball statistics, please refer to Baseball Reference ("WAR Explained").

Methodology

In this section, we outline the methodology used to transform raw data into feature-rich dataframes suitable for analysis. This process involved data cleaning, wrangling, and deriving new features from the datasets.

The data cleaning process started by taking our raw data tables directly from the CSV files. From these raw tables, we filtered out only the cases that we would need for our analysis: 21st century Yankees players.

Next, we needed to take the raw data, which was in the form of stats that can be recorded directly from a game (H, BB, HR) and transform it into statistics that are more indicative of a player's success (WAR, BA, $OPS+$). This was the majority of the data wrangling process due to the complexity of some of the formulas, as mentioned in the previous section.

The other major form of data wrangling involved adjusting player salaries for inflation. Adjusting player salaries for inflation ensures that . Since we will be comparing salaries to player performance metrics, we do not want players from later years to be unfairly evaluated, simply because of the depreciating value of the US dollar.

Player salaries were scaled using the Consumer Price Index, CPI, a metric used to gauge inflation, based on the following formula:

$$S_a = S \times \frac{CPI_b}{CPI}$$

where S_a denotes the salary adjusted for inflation, S denotes the raw salary, CPI denotes the Consumer Price Index in the year of the salary, and CPI_b denotes a constant for the baseline CPI. CPI_b essentially is the CPI for the year in which all salaries will be scaled to. For our analysis, CPI_b will be set to the CPI value in 2015, the most recent year of our dataset.

Throughout this process, tables were joined together using the yearIDs, teamIDs, and playerIDs.

Data Analysis

With our data now cleaned and organized, we can move forward with the analysis. To address our research questions effectively, we have developed a series of visualizations. This section is structured around each research question, with a focused analysis for each.

What does the distribution of salaries for the New York Yankees look like?

This research question aims to uncover the salary range, salary spread, and the skewness of the adjusted annual salary distribution.

To explore this question, Figure 1 shows a density plot for the adjusted annual salaries.

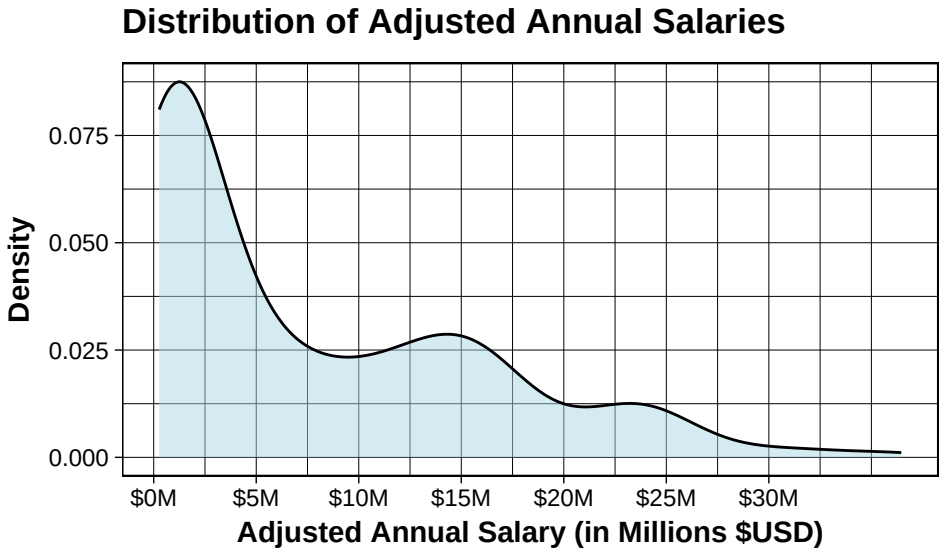

Figure 1: Density plot showing salary distribution

We can see that the distribution of salaries is clearly skewed to the right. This makes sense as the vast majority of MLB players make close to the league minimum salary, while only the top players make top dollar.

There is clear clustering around the league minimum salary, which for 2000 - 2015 ranged from $200,000 - $507,500.

We can see that there is some clustering around the $15 million mark. This could be the result of one contract that spans several years where each year is paid the same amount of money. Since each case is a yearly salary, all of these salaries would "stack" on top of one another, creating this cluster in the distribution.

From this analysis of the salary distributions, we have taken away that most cases cluster towards the low end of the salary range. This information prompts us to use a logarithmic

scale for adjusted salary in future plots, to prevent so much clustering of salaries at the lower end of the range.

Which offensive and defensive metrics correlate the most with a player's salary?

This research question aims to uncover which performance metrics correlate the most with a player's annual earnings. This section will be divided into three parts, Offensive Metrics, Defensive Metrics, and a Summary

Offensive Metrics

Figure 2 will look at the correlation between WAR and adjusted annual salary, plotted on a logarithmic scale.

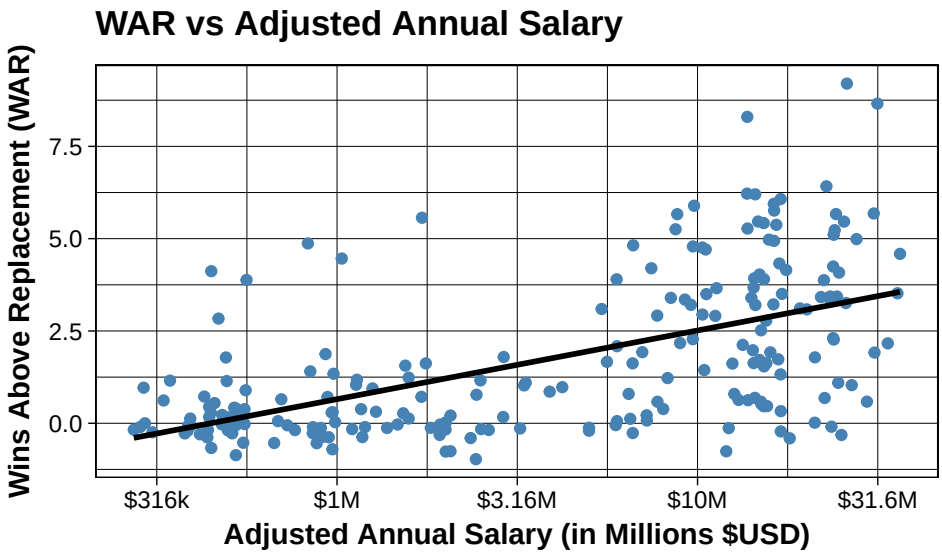

Figure 2: Scatterplot of WAR vs Salary

When analyzing the plot between WAR and annual salary, it can be seen that there is a very strong positive correlation between WAR and what a player is being paid. It is evident that players who have higher WAR values contribute more to their teams success, therefore they have higher earnings. The majority of the players with a WAR greater than 5 are paid above $10 million dollars annually showing that they are paid well on the offensive side.

There is also extremely high variability when WAR gets lower, some players with extremely low WAR are still getting paid millions and millions of dollars. Why is this? Some of these factors include long term contracts, or the fact that they are very experienced in their specific position. Some players who have WAR values over 3 are getting paid under a million dollars, showing those players may be undervalued.

As will be further explored in summary, when looking at the relationships between the offensive stats analyzed, WAR has the highest positive correlation out of the three, showing that players who can win for their team will generally get paid more.

Figure 3 will look at the correlation between BA and adjusted annual salary, plotted on a logarithmic scale.

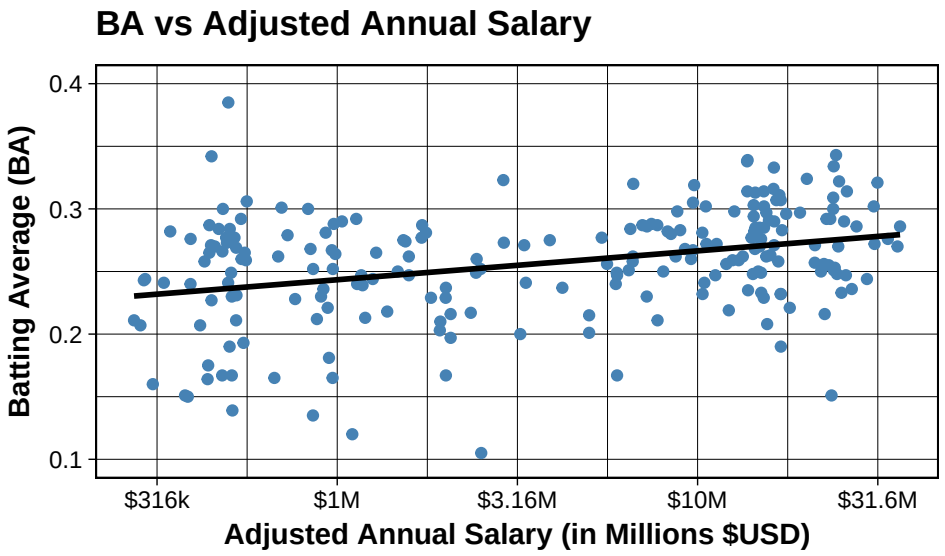

Figure 3: Scatterplot of BA vs Salary

Looking at this plot it is evident that there is a positive correlation between batting average and annual salary, but not a clear correlation. Players who have a higher batting average generally do get paid more, but the correlation isn't overly strong.

There is also lots of variability between players and their batting averages. There are many outliers that when looking at just batting average, it does not justify what a player is being paid. For example, one player has a batting average of about .400 and is only getting paid about $500,000, yet another player with a batting average at around .150 is getting paid roughly $25,000,000. These outliers, among others, show that other factors must be having a stronger effect on salary.

Figure 4 will look at the correlation between $OPS+$ and adjusted annual salary, plotted on a logarithmic scale.

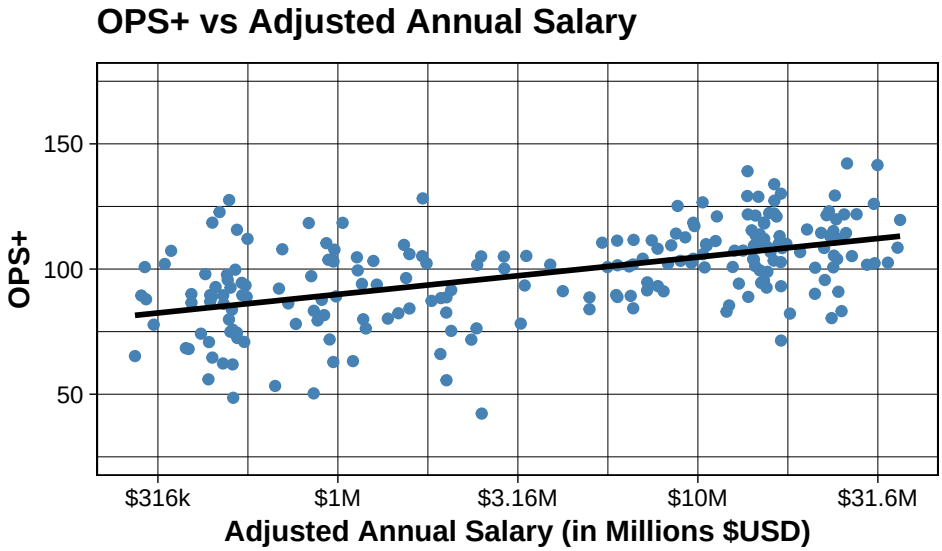

Figure 4: Scatterplot of OPS+ vs Salary

When looking at the correlation between $OPS+$ and adjusted annual salary, it can be observed that there is a positive correlation. As expected, players with higher $OPS+$ values generally do get paid more than with lower $OPS+$ values. We observe that many players with $OPS+$ values higher than 100, which is the league average, make between \$10M and \$31.6M dollars, showing that those players are rewarded.

There also seems to be lots of undervaluation, for $OPS+$, compared to other metrics: many players close to 100 or even just above 100 are not getting paid nearly as much as their counterparts, when having very similar stats.

Defensive Metrics

Figure 5 will look at the correlation between WAR and adjusted annual salary, plotted on a logarithmic scale.

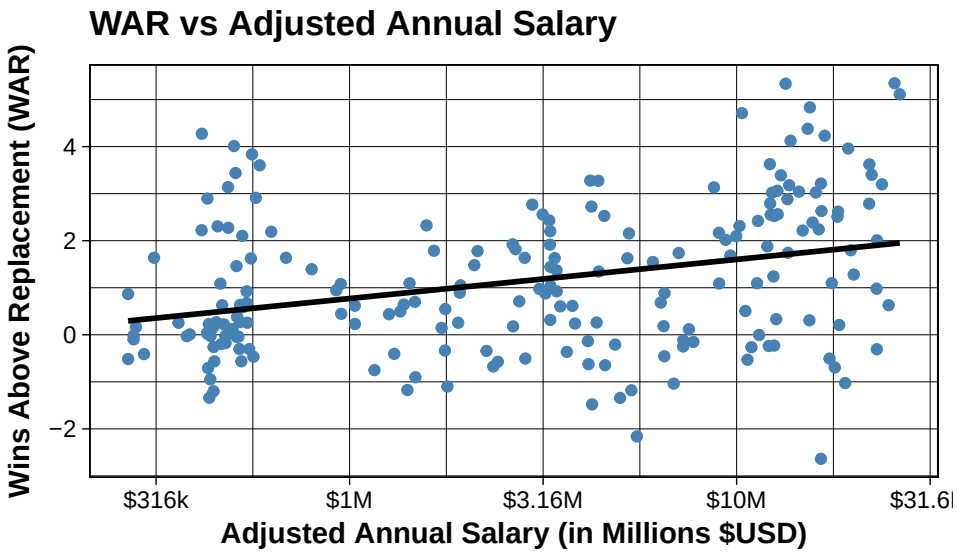

Figure 5: Scatterplot of WAR vs Salary

When looking at WAR for pitchers, there definitely is a positive correlation, but there much more spread between values, as compared to WAR for position players. As WAR increases, generally so does salary for pitchers.

Pitchers with a WAR over 2.5 generally make the most, showing that elite players are properly compensated. Pitchers with WAR close to 0 have an extremely wide variety of salaries ranging all the way from the lowest to even the highest salaries on the graph. This exemplifies that pitchers get paid for many other reasons, other than just WAR.

In other words, higher WAR for pitchers generally means higher compensation, but lower WAR does not necessarily mean lower compensation.

Figure 6 will look at the correlation between ERA and adjusted annual salary, plotted on a logarithmic scale.

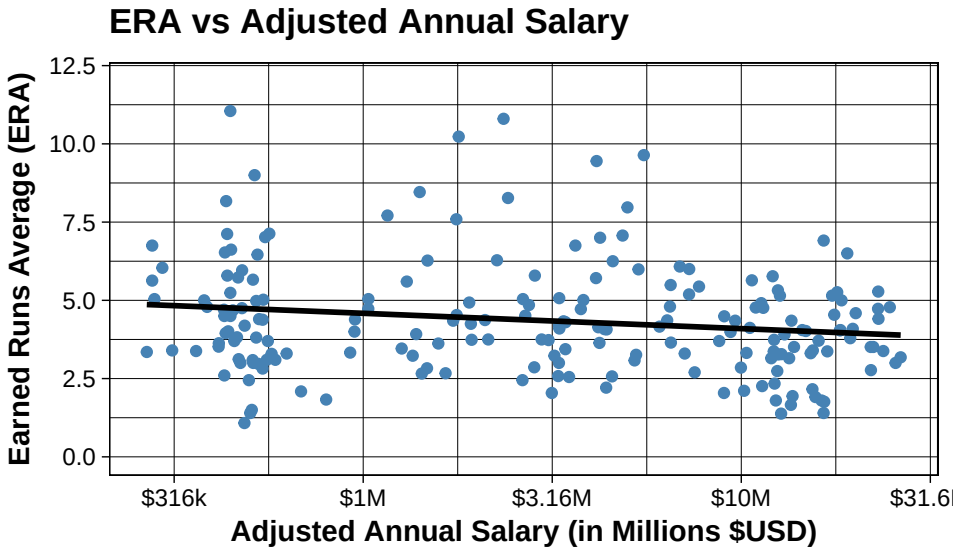

Figure 6: Scatterplot of ERA vs Salary

The first thing that is evident from this plot is the fact that ERA and adjusted annual salary actually has a negative correlation. This is due to the fact that a lower ERA value means that a pitcher allows less runs, and as a result, they make more money.

However, there are still many players that make millions even though their ERA is over 5, possibly due to the fact that these players have much more experience, better past performance, or contracts which influence what the pitcher is being paid.

There is also lots of variability in this plot, which may account for other factors, so the next the plot will be evaluated to see if more findings can be uncovered.

Figure 7 will look at the correlation between $WHIP$ and adjusted annual salary, plotted on a logarithmic scale.

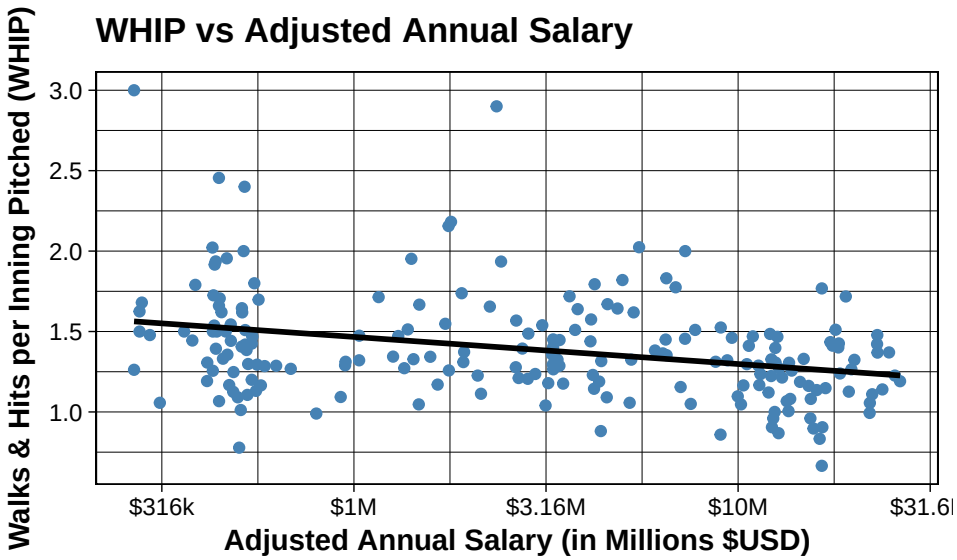

Figure 7: Scatterplot of WHIP vs Salary

Similar to ERA, it is also evident that the correlation between $WHIP$ and adjusted annual salary plot is negative. Lower $WHIP$ indicates less walks and hits, and therefore fewer baserunners allowed.

For pitchers with $WHIP$ between 1.3 and 1.8, salaries vary drastically. This indicates that $WHIP$ is not a sole predictor of a pitcher's salary. Some pitchers with $WHIP$ over 2.0 (which is not good) still get paid very well. Why could this be the case? Some factors are that the pitchers may have small sample size, a large contract spanning years, or the fact that they are a veteran pitcher.

Most pitchers with a $WHIP$ below 1.0 are paid extremely well showing great compensation for these pitchers skillset.

Summary

This section will summarize the correlations found between the various performance metrics and average annual salary.

Figure 8 is a heat map showing the correlations between the various batting performance metrics.

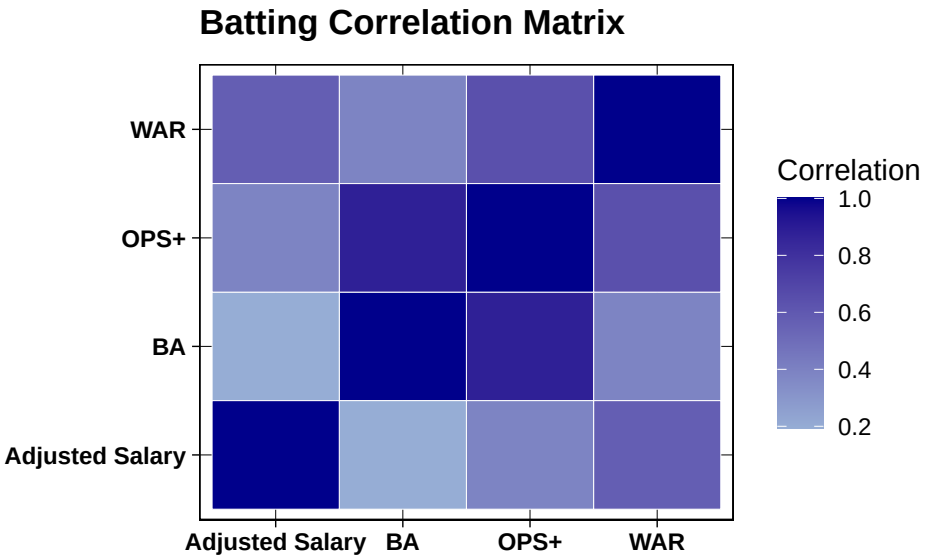

Figure 8: Heat Map of Batting Metrics

The batting correlation matrix provides a visual summary of relationships between offensive performance and adjusted annual salary.

The strongest correlation with adjusted annual salary, as alluded to previously, is with WAR, showing that WAR is a good measure of a players contributions and value to their team.

$OPS+$ also shows a positive correlation with adjusted annual salary, showing that players with higher offensive production are often compensated for their contributions.

BA has the weakest correlation with average annual salary among the group, showing that teams now prioritize WAR and $OPS+$ over how well a player can hit. When looking at the shift in modern baseball analytics, it can be seen that WAR and $OPS+$ are the preferred indicators of a players overall value.

Figure 9 is a heat map showing the correlations between the various pitching performance metrics.

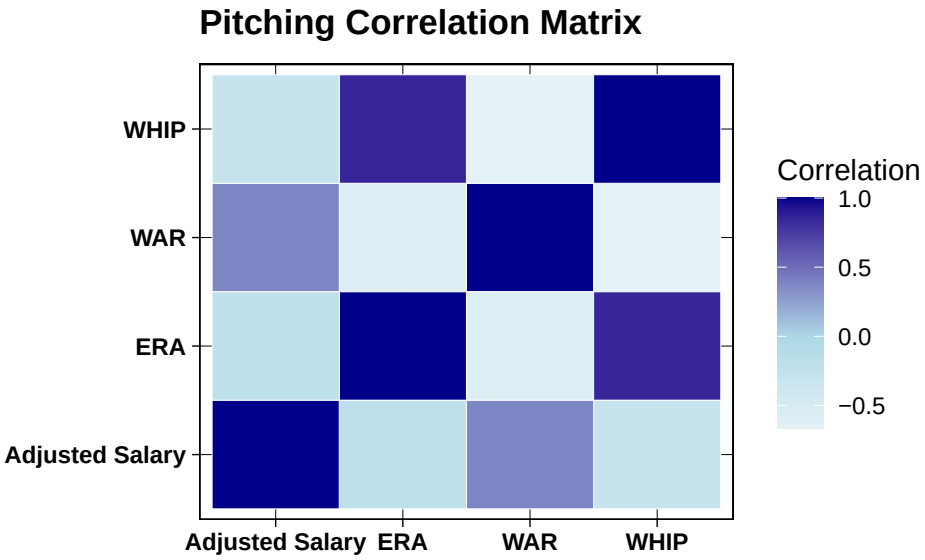

Figure 9: Heat Map of Pitching Metrics

The pitching correlation matrix provides a visual summary of relationships between all pitching factors and adjusted salary. Darker blue indicates a strong correlation whereas lighter blue indicates weaker or negative correlations.

It is evident that there is a positive correlation (around 0.5) between WAR and adjusted annual salary, showing that higher WAR values contribute more to higher payouts and team success. Both ERA and $WHIP$ have negative correlations to adjusted annual salary, which makes sense due to the fact that as a pitcher, when you lower the amount of runners on bases, you are doing a better job. While ERA is important, having a weaker correlation shows that ERA is a less reliable factor when looking at a pitcher's contributions.

From all of this analysis, we can conclude that WAR is the best predictor of a player's annual adjusted salary.

What is the salary breakdown across different positions?

This section seeks to uncover which positions the New York Yankees pay the most and least.

Figure 10 shows the average annual adjusted salary for players of a certain position group.

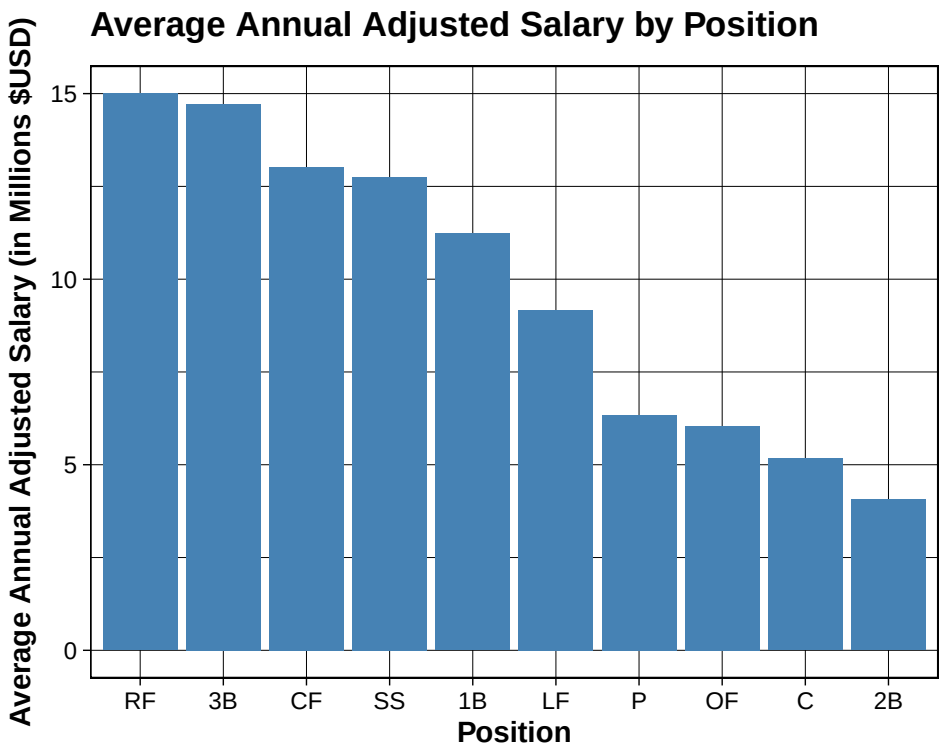

Figure 10: Bar Graph of Salary by Position

When looking at a player's average salary based on position, it can be seen that right field (RF) and third base (3B) have the highest average salary at around $15 million for both positions. This may be because these positions are usually filled with players who contribute heavily to a teams offense: home run hitters and run producers.

Analyzing the positions with mid-tier salaries, it can be seen that positions like center field (CF), shortstop (SS), and first base (1B) get paid between $11 and $13 million dollars a year. These positions may be valued highly due to shortstops having a duel role providing strong defense and contributing offensively, and center fielders being very athletic, having good fielding abilities, and speed.

The positions with lower tier salaries is where this analysis gets interesting. Seeing pitchers (P) with much lower average salaries is eye opening, considering pitching is crucial to a team's success. Players with positions such as catcher (C) and 2nd baseman (2B), have more defensive contributions rather than offensive, which may be why they are paid less.

Which players provided the most and least value per dollar they were paid?

In this section, we hope to uncover who the top 10 over and under performers on the New York Yankees from 2000 - 2015. In previous analysis, we concluded that WAR was the best predictor of performance for both a hitter and pitcher, so to evaluate these players will use WAR per million dollars.

Over Performers

Table 1 shows the top 10 New York Yankees players in WAR per million dollars from 2000 - 2015.

Table 1: Top 10 Over Performers (Lowest WAR per Million Dollars)

Year	Player	Position	WAR	Salary ($USD)	WAR / $M
2006	Chien-Ming Wang	P	4.27	415238	10.29
2006	Robinson Cano	2B	4.12	447952	9.20
2014	Dellin Betances	P	4.01	502696	7.98
2007	Robinson Cano	2B	3.88	561043	6.92
2007	Chien-Ming Wang	P	3.84	559557	6.86
2015	Dellin Betances	P	3.44	507500	6.78
2008	Joba Chamberlain	P	2.90	429334	6.74
2011	David Robertson	P	3.14	485173	6.47
2013	Ivan Nova	P	3.60	585631	6.15
2003	Nick Johnson	1B	2.84	469117	6.04

Table showing top 10 players for WAR / $M

When looking at the top 10 over performers list of Yankees from 2000-2015, it is evident that pitchers dominate the list. 7 out of the top 10 over performers are pitchers. This may suggest that underdeveloped or young pitchers out perform what is expected of them during their rookie contracts or before free agency.

It can be observed that Chien-Ming Wang was a huge over performer having back to back entries on this where in 2006 his WAR per million dollars was 10.29, and in 2007 his WAR per million dollars was 6.86. Robinson Cano and Dellin Betances were also great over performers both appearing twice on this list.

All players on this list made less than $590,000 dollars and over performed greatly all having WAR per million dollars values over 6.00. This highlights the value of players in their earlier contracts when starting out and shows how teams have a huge financial advantage during the pre-arbitration years. This also shows how teams can pay players much less to have them develop skills that will take their game to the next level.

Under Performers

Table 2 shows the bottom 10 New York Yankees players in WAR per million dollars from 2000 - 2015.

Table 2: Top 10 Under Performers (Lowest WAR per Million Dollars)

Year	Player	Position	WAR	Salary ($USD)	WAR / $M
2008	Ian Kennedy	P	-1.34	434040	-3.08
2007	Jeff Karstens	P	-1.20	445239	-2.70
2007	Sean Henn	P	-0.95	436726	-2.17
2001	Christian Parker	P	-0.52	267715	-1.93
2013	Chris Stewart	C	-0.86	524077	-1.65
2008	Ross Ohlendorf	P	-0.71	430902	-1.64
2010	Ramiro Pena	3B	-0.67	447935	-1.49
2002	Randy Choate	P	-0.41	294303	-1.40
2008	Phil Hughes	P	-0.56	447333	-1.26
2012	Cory Wade	P	-0.56	525379	-1.07

Table showing bottom 10 players for WAR / $M

All players on this list have negative WAR which is an awful performance. It can be observed that all players only appear once on this list, likely because when underperforming at this level, a player will likely be released from the team after that season.

Pitchers also dominate this list with 8 out of the 10 players being pitchers. This is because pitchers are more likely prone to under perform due to small sample sizes, injuries, or poor outings by them which impact their WAR. There is only one catcher on this list which means if his WAR was that poor that his defensive contributions negatively affected the team. From looking at the players salaries these players were likely early in their careers and still under performed significantly. One player, Phil Hughes, was a great prospect and failed to meet his expectations, which are the risks a team may have when paying someone so young when they arent fully developed.

Conclusion

This section caps off our analysis of the New York Yankees' salary and performance data and reveals the following trends in player valuation and resource allocation:

The salary distribution of the New York Yankees from 2000-2015 was skewed to the right, with most salaries being close to the league minimum.

WAR has the best positive correlation with the adjusted salary of the players, this means that the players who produce the most WAR to the team receive the largest pay. The players with high WAR values and especially those with WAR above 5, earn more than 10 million dollars a year, and contrasts with the lower WAR values which are more variable with the salary, which may be due to factors such as long term contracts or experience.

Other metrics such as $OPS+$ also correlate well with adjusted annual salary, as they show the increasing focus on the advanced statistical data. BA, on the other hand, has a low correlation coefficient, as it shows that there is the decreasing emphasis on the conventional statistics.

As for the pitchers, their ERA and $WHIP$ are negatively correlated with the salary; this means that those with better control, and who do not allow many baserunners to the opposition (lower $WHIP$) earn more, on average. Nevertheless, the differences in the salaries of pitchers with equal or close to equal $WHIP$ and ERA indicate that there are other factors, including the experience of the pitcher or the terms of the contract. The correlation matrices also agree with this notion that WAR is the best predictor of a player's salary for hitters, while $WHIP$ is slightly better for pitchers.

When the players are classified by their positions, then offensive minded players who play in right field (RF) and third base (3B) earn the maximum amount of money whereas the defensive players who are catchers (C) or second basemen (2B) earn relatively less. It was observed that pitchers (P) had lower average salaries even though they are considered to be very vital in the team, which may be as a result of the wide variation in the salaries of the pitching staff or the abundance of arms on a team.

The list of top overperformers presents the value of young players on the team's control: Chien-Ming Wang, Betances, and Cano. Underperformers (who mainly were pitchers) show how in risky the unproven early players stage or those their careers who were affected by injuries are.

In general, the Yankees effectively reward the best players, according to WAR, while at the same time, utilizing cheap young players they have secured on rookie deals. However, underperformance shows that the problem of player assessment on small sample sizes and then signing long-term contracts is still present.

References

Labor Statistics, Bureau of. *Consumer Price Index for All Urban Consumers (CPI-U)*. U.S. Department of Labor, 2024, https://data.bls.gov/timeseries/CUUR0000SA0?years_option=all_years.

Lahman, Sean. *Sean Lahman Baseball Database*. SeanLahman.com, 2023, http://seanlahman.com.

Lewis, Michael. *Moneyball: The Art of Winning an Unfair Game*. W.W. Norton & Company, 2003.

"Pitcher WAR Calculations and Details." *Baseball Reference*, https://www.baseball-reference.com/about/war_explained_pitch.shtml. Accessed 18 Dec. 2024.

"Position WAR Calculations and Details." *Baseball Reference*, https://www.baseball-reference.com/about/war_explained_position.shtml. Accessed 18 Dec. 2024.

Sports, Open Source. *Baseball Databank*. Kaggle, 2019, https://www.kaggle.com/datasets/open-source-sports/baseball-databank.

"WAR Explained." *Baseball Reference*, https://www.baseball-reference.com/about/war_explained.shtml. Accessed 18 Dec. 2024.

Code Appendix

```
# This code will be using the tidyverse style guide

# Load necessary packages
library(tidyverse)
library(knitr)
library(kableExtra)
# Set working directory to path to repo (uncomment if needed)
#setwd("~/path/to/repo")

# Read in data ------------------------------------
master_raw <- read.csv("data/Master.csv")
salaries_raw <- read.csv("data/Salaries.csv")
batting_raw <- read.csv("data/Batting.csv")
pitching_raw <- read.csv("data/Pitching.csv")
fielding_raw <- read.csv("data/Fielding.csv")
teams_raw <- read.csv("data/Teams.csv")
cpi_raw <- read.csv("data/CPI_data.csv")
### Computing wins above replacement (WAR) -------------------------------------

## Position players

# Using Baseball Reference's definition of WAR (bWAR):
# - https://www.baseball-reference.com/about/war_explained.shtml
# - https://www.baseball-reference.com/about/war_explained_position.shtml
# - https://www.baseball-reference.com/about/war_explained_pitch.shtml

# Constant for average number of runs per win, Baseball Reference assumes this is 10
AVG_RUNS_PER_WIN <- 10

# Filter teams table to 21st century
teams <- teams_raw %>%
  filter(yearID > 1999)

# Select park factors
teams_bpf <- teams %>%
  select(yearID, teamID, BPF) # Batting park factor

teams_ppf <- teams %>%
  select(yearID, teamID, PPF) # Pitching park factor

# Join park factors to batting and pitching tables by year and team
batting_with_bpf <- left_join(batting_raw, teams_bpf, by = c("yearID", "teamID"))
pitching_with_ppf <- left_join(pitching_raw, teams_ppf, by = c("yearID", "teamID"))
```

```r
# Compute the league average runs per at-bat
avg_runs_per_ab <- mean(teams$R / teams$AB)
# Replacement level is defined as scoring 80% of the league average runs per at-bat
replacement_level_runs_per_ab <- avg_runs_per_ab * 0.8

# Compute bWAR for 21st century Yankees hitters
nyy_batting <- batting_with_bpf %>%
  # Filter cases to 21st century Yankees players
  filter(yearID > 1999 & teamID == "NYA") %>%
  # Add columns for the WAR computation
  mutate(
    TB = H + X2B + 2 * X3B + 3 * HR, # total bases (TB)
    RC = ifelse((AB + BB) > 0, (H + BB) * TB / (AB + BB), 0), # runs created (RC)
    replacement_level_runs = replacement_level_runs_per_ab * AB, # replacement runs
    runs_above_replacement = RC - replacement_level_runs, # runs above replacement
    WAR = runs_above_replacement / AVG_RUNS_PER_WIN # bWAR
  ) %>%
  # Sort by descending WAR
  arrange(desc(WAR))


## Pitchers

# Compute league average runs allowed per 9 innings
lg_RA9 <- (sum(teams$RA) / sum(teams$IPouts / 3)) * 9
# Replacement level is defined as allowing 20% more runs per 9 innings than the league avera
replacement_RA9 <- lg_RA9 * 1.2

# Compute bWAR for 21st century Yankees pitchers
nyy_pitching <- pitching_with_ppf %>%
  # Filter cases to 21st century Yankees players
  filter(yearID > 1999 & teamID == "NYA") %>%
  # Add columns for the WAR computation
  mutate(
    RA9 = (R / (IPouts / 3)) * 9, # runs allowed per 9 innings
    RA9_adj = RA9 * (PPF / 100), # runs allowed per 9 innings, adjusted for home ballpark
    WAR = (replacement_RA9 - RA9_adj) * ((IPouts / 3) / 9) / AVG_RUNS_PER_WIN, # bWAR
  ) %>%
  # Sort by descending WAR
  arrange(desc(WAR))

### Computing other common statistics ------------------------------------

## Position players
```

```r
# Compute league on-base percentage (OBP)
lg_obp <- (sum(teams$H) + sum(teams$BB) + sum(teams$HBP)) / (sum(teams$AB) + sum(teams$BB) +
# Compute league slugging percentage (SLG)
lg_slg <- ((sum(teams$H) - (sum(teams$X2B) + sum(teams$X3B) + sum(teams$HR))) + (2 * sum(tea
# Compute league on-base plus slugging (OPS)
lg_ops <- lg_obp + lg_slg

# Add statistics to batting dataframe
nyy_batting <- nyy_batting %>%
  mutate(
    BA = H / AB, # batting average (BA)
    OBP = (H + BB + HBP) / (AB + BB + HBP + SF), # on-base percentage (OBP)
    SLG = ((H - (X2B + X3B + HR)) + (2 * X2B) + (3 * X3B) + (4 * HR)) / AB, # slugging perce
    OPS = OBP + SLG, # on-base plus slugging (OPS)
    OPS_plus = (100 * (OPS / lg_ops) / (BPF / 100)) # OPS+ adjusted for league
  ) %>%
  # Round columns to 3 decimal places
  mutate(
    WAR = round(WAR, 3),
    BA = round(BA, 3),
    OBP = round(OBP, 3),
    SLG = round(SLG, 3),
    OPS = round(OPS, 3),
    OPS_plus = round(OPS_plus, 3)
  ) %>%
  select(playerID, yearID, teamID, WAR, BA, OBP, SLG, OPS, OPS_plus)

## Pitchers

# Add extra statistics to pitching dataframe
nyy_pitching <- nyy_pitching %>%
  mutate(
    WHIP = (BB + H) / (IPouts / 3), # walks and hits per inning pitched (WHIP)
    K_per_9 = (SO * 9) / (IPouts / 3), # strikeouts per 9 innings (K9)
    BB_per_9 = (BB * 9) / (IPouts / 3) # walks per 9 innings (BB9)
  ) %>%
  # Round columns to 3 decimal places
  mutate(
    WAR = round(WAR, 3),
    WHIP = round(WHIP, 3),
    K_per_9 = round(K_per_9, 3),
    BB_per_9 = round(BB_per_9, 3)
  ) %>%
  select(playerID, yearID, teamID, WAR, ERA, WHIP, K_per_9, BB_per_9)
```

```r
### Joining tables to create the main dataframes ------------------------------------

## Add salary column

# Adjust salaries for inflation using CPI
cpi_data <- cpi_raw %>%
  mutate(
    cpi = (Jan + Feb + Mar + Apr + May + Jun + Jul + Aug + Sep + Oct + Nov + Dec) / 12
  ) %>%
  select(Year, cpi)

# Base CPI is for the year 2015
base_cpi <- cpi_data$cpi[cpi_data$Year == 2015]

# Clean and adjust salaries for inflation
salaries <- salaries_raw %>%
  filter(yearID > 1999 & teamID == "NYA") %>%
  left_join(cpi_data, by = c("yearID" = "Year")) %>%
  mutate(
    salary_adj = round(salary * (base_cpi / cpi))
  ) %>%
  select(-lgID, -cpi)

# Add salary and adjusted salary columns to batting and pitching dataframes
nyy_batting <- inner_join(nyy_batting, salaries, by = c("yearID", "teamID", "playerID"))
nyy_pitching <- inner_join(nyy_pitching, salaries, by = c("yearID", "teamID", "playerID"))

## Add position column

# Clean the fielding dataframe
fielding <- fielding_raw %>%
  select(yearID, teamID, playerID, POS, G) %>%
  group_by(yearID, teamID, playerID) %>%
  filter(G == max(G)) %>%
  slice(1) %>%
  ungroup() %>%
  select(-G) %>%
  filter(yearID > 1999 & teamID == "NYA")

# Add position column to batting and pitching dataframes
nyy_batting <- inner_join(nyy_batting, fielding, by = c("yearID", "teamID", "playerID"))
nyy_pitching <- inner_join(nyy_pitching, fielding, by = c("yearID", "teamID", "playerID"))
```

```r
# Filter out non-pitchers from the pitching dataframe
nyy_pitching <- nyy_pitching %>%
  filter(POS == "P")

## Add name column

# Clean the master dataset
master <- master_raw %>%
  mutate(name = paste(nameFirst, nameLast, sep = " ")) %>%
  select(playerID, name)

# Add name column to batting and pitching dataframes
nyy_batting <- inner_join(nyy_batting, master, by = "playerID")
nyy_pitching <- inner_join(nyy_pitching, master, by = c("playerID" = "playerID"))

# Filter out pitchers from batting dataframe
nyy_batting <- nyy_batting %>%
  filter(POS != "P") %>%
  select(yearID, teamID, playerID, name, POS, WAR, BA, OBP, SLG, OPS, OPS_plus, salary, sala

nyy_pitching <- nyy_pitching %>%
  select(yearID, teamID, playerID, name, POS, WAR, ERA, WHIP, K_per_9, BB_per_9, salary, sal

# Create a team dataframe for easier data aggregation across the team for data analysis
nyy_team <- full_join(nyy_batting, nyy_pitching, by = c("yearID", "teamID", "playerID", "nam
## Individual salaries vs individual performance ------------------------------------

# Density plot of adjusted salary
plot1 <- nyy_team %>%
  ggplot(aes(x = salary_adj / 1e6)) +
  geom_density(fill = "lightblue", alpha = 0.5) +
  labs(title = expression(bold("Distribution of Adjusted Annual Salaries")),
       x = expression(bold("Adjusted Annual Salary (in Millions $USD)")),
       y = expression(bold("Density"))) +
  scale_x_continuous(
    breaks = c(0, 5, 10, 15, 20, 25, 30),
    labels = c("$0M", "$5M", "$10M", "$15M", "$20M", "$25M", "$30M")
  ) +
  theme_linedraw()

## Analyze correlation between salary and batting statistics ------------------------------

# Salary_adj plotted on log scale

# WAR vs Adjusted Salary
```

```r
plot2 <- nyy_batting %>%
  ggplot(aes(x = log10(salary_adj / 1e6), y = WAR)) +
  geom_point(color = "steelblue") +
  geom_smooth(method = "lm", se = FALSE, color = "black", linetype = "solid") +
  labs(title = expression(bold("WAR vs Adjusted Annual Salary")),
       x = expression(bold("Adjusted Annual Salary (in Millions $USD)")),
       y = expression(bold("Wins Above Replacement (WAR)"))) +
  scale_x_continuous(
    breaks = c(-0.5, 0, 0.5, 1, 1.5),
    labels = c("$316k", "$1M", "$3.16M", "$10M", "$31.6M")
  ) +
  theme_linedraw()

# BA vs Adjusted Salary
plot3 <- nyy_batting %>%
  ggplot(aes(x = log10(salary_adj / 1e6), y = BA)) +
  geom_point(color = "steelblue") +
  geom_smooth(method = "lm", se = FALSE, color = "black", linetype = "solid") +
  labs(title = expression(bold("BA vs Adjusted Annual Salary")),
       x = expression(bold("Adjusted Annual Salary (in Millions $USD)")),
       y = expression(bold("Batting Average (BA)"))) +
  scale_x_continuous(
    breaks = c(-0.5, 0, 0.5, 1, 1.5),
    labels = c("$316k", "$1M", "$3.16M", "$10M", "$31.6M")
  ) +
  scale_y_continuous(limits = c(0.1, 0.4)) +
  theme_linedraw()

# OPS+ vs Adjusted Salary
plot4 <- nyy_batting %>%
  ggplot(aes(x = log10(salary_adj / 1e6), y = OPS_plus)) +
  geom_point(color = "steelblue") +
  geom_smooth(method = "lm", se = FALSE, color = "black", linetype = "solid") +
  labs(title = expression(bold("OPS+ vs Adjusted Annual Salary")),
       x = expression(bold("Adjusted Annual Salary (in Millions $USD)")),
       y = expression(bold("OPS+"))) +
  scale_x_continuous(
    breaks = c(-0.5, 0, 0.5, 1, 1.5),
    labels = c("$316k", "$1M", "$3.16M", "$10M", "$31.6M")
  ) +
  scale_y_continuous(limits = c(25, 175)) +
  theme_linedraw()

## Analyze correlation between salary and pitching statistics ------------------------------
```

```r
# Salary_adj plotted on log scale

# WAR vs Adjusted Salary
plot5 <- nyy_pitching %>%
  ggplot(aes(x = log10(salary_adj / 1e6), y = WAR)) +
  geom_point(color = "steelblue") +
  geom_smooth(method = "lm", se = FALSE, color = "black", linetype = "solid") +
  labs(title = expression(bold("WAR vs Adjusted Annual Salary")),
       x = expression(bold("Adjusted Annual Salary (in Millions $USD)")),
       y = expression(bold("Wins Above Replacement (WAR)"))) +
  scale_x_continuous(
    breaks = c(-0.5, 0, 0.5, 1, 1.5),
    labels = c("$316k", "$1M", "$3.16M", "$10M", "$31.6M")
  ) +
  theme_linedraw()

# ERA vs Adjusted Salary
plot6 <- nyy_pitching %>%
  ggplot(aes(x = log10(salary_adj / 1e6), y = ERA)) +
  geom_point(color = "steelblue") +
  geom_smooth(method = "lm", se = FALSE, color = "black", linetype = "solid") +
  labs(title = expression(bold("ERA vs Adjusted Annual Salary")),
       x = expression(bold("Adjusted Annual Salary (in Millions $USD)")),
       y = expression(bold("Earned Runs Average (ERA)"))) +
  scale_x_continuous(
    breaks = c(-0.5, 0, 0.5, 1, 1.5),
    labels = c("$316k", "$1M", "$3.16M", "$10M", "$31.6M")
  ) +
  scale_y_continuous(limits = c(0, 12)) +
  theme_linedraw()

# WHIP vs Adjusted Salary
plot7 <- nyy_pitching %>%
  ggplot(aes(x = log10(salary_adj / 1e6), y = WHIP)) +
  geom_point(color = "steelblue") +
  geom_smooth(method = "lm", se = FALSE, color = "black", linetype = "solid") +
  labs(title = expression(bold("WHIP vs Adjusted Annual Salary")),
       x = expression(bold("Adjusted Annual Salary (in Millions $USD)")),
       y = expression(bold("Walks & Hits per Inning Pitched (WHIP)"))) +
  scale_x_continuous(
    breaks = c(-0.5, 0, 0.5, 1, 1.5),
    labels = c("$316k", "$1M", "$3.16M", "$10M", "$31.6M")
  ) +
  theme_linedraw()
```

```r
## Analyze correlations between variables -----------------------------------

# Create correlation matrix between metrics and salaries for batters and pitchers

# Batters
nyy_batting_renamed <- nyy_batting %>%
  rename(
    "OPS+" = "OPS_plus",
    "Adjusted Salary" = "salary_adj"
  )
cor_batting_df <- nyy_batting_renamed %>%
  select(WAR, BA, `OPS+`, `Adjusted Salary`) %>%
  cor(use = "complete.obs") %>%
  as.data.frame() %>%
  mutate(Variable1 = rownames(.)) %>%
  pivot_longer(
    cols = -Variable1,
    names_to = "Variable2",
    values_to = "Correlation"
  )

# Pitchers
nyy_pitching_renamed <- nyy_pitching %>%
  rename(
    "Adjusted Salary" = "salary_adj"
  )
cor_pitching_df <- nyy_pitching_renamed %>%
  select(WAR, ERA, WHIP, `Adjusted Salary`) %>%
  cor(use = "complete.obs") %>%
  as.data.frame() %>%
  mutate(Variable1 = rownames(.)) %>%
  pivot_longer(
    cols = -Variable1,
    names_to = "Variable2",
    values_to = "Correlation"
  )

## Create heatmaps to show correlations between variables ---------------------------------

# Batters
plot8 <- cor_batting_df %>%
  ggplot(aes(x = Variable1, y = Variable2, fill = Correlation)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "white", mid = "lightblue", high = "darkblue", midpoint = 0) +
  labs(title = expression(bold("Batting Correlation Matrix")), fill = "Correlation") +
```

```
  theme_linedraw() +
  theme(axis.text.x = element_text(face = "bold"),
        axis.text.y = element_text(face = "bold"),
        axis.title.x = element_blank(),
        axis.title.y = element_blank())

# Pitchers
plot9 <- cor_pitching_df %>%
  ggplot(aes(x = Variable1, y = Variable2, fill = Correlation)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "white", mid = "lightblue", high = "darkblue", midpoint = 0) +
  labs(title = expression(bold("Pitching Correlation Matrix")), fill = "Correlation") +
  theme_linedraw() +
  theme(axis.text.x = element_text(face = "bold"),
        axis.text.y = element_text(face = "bold"),
        axis.title.x = element_blank(),
        axis.title.y = element_blank())

## Average Salary by Position -----------------------------------

# Get dataframe of salary by position
salary_by_position <- nyy_team %>%
  group_by(POS) %>%
  summarize(avg_salary = mean(salary_adj, na.rm = TRUE), .groups = "drop") %>%
  mutate(avg_salary = avg_salary / 1e6) %>%
  arrange(desc(avg_salary))

# Create a bar chart plotting the average salaries
plot10 <- salary_by_position %>%
  ggplot(aes(x = reorder(POS, -avg_salary), y = avg_salary, fill = POS)) +
  geom_bar(stat = "identity", fill = "steelblue", show.legend = FALSE) +
  labs(title = expression(bold("Average Annual Adjusted Salary by Position")),
       x = expression(bold("Position")),
       y = expression(bold("Average Annual Adjusted Salary (in Millions $USD)"))) +
  theme_linedraw() +
  theme(axis.text.x = element_text())

## Over and under performers: WAR / $ ---------------------------------

# Over performers
top_10_over_performers <- nyy_team %>%
  mutate(WAR_per_dollar = (WAR * 1e6) / salary_adj) %>%
  arrange(desc(WAR_per_dollar)) %>%
  slice_head(n = 10)
```

```r
# Under performers
top_10_under_performers <- nyy_team %>%
  mutate(WAR_per_dollar = (WAR * 1e6) / salary_adj) %>%
  arrange(WAR_per_dollar) %>%
  slice_head(n = 10)

# Over performers table
top_10_over_table <- top_10_over_performers %>%
  select(yearID, name, POS, WAR, salary_adj, WAR_per_dollar) %>%
  kable(col.names = c("Year", "Player", "Position", "WAR", "Salary ($USD)", "WAR / $M"),
        caption = "Top 10 Over Performers (Lowest WAR per Million Dollars)",
        digits = c(0, 0, 0, 2, 0, 2)) %>%
  kable_styling(bootstrap_options = c("striped", "condensed", "bordered"))

# Under performers table
top_10_under_table <- top_10_under_performers %>%
  select(yearID, name, POS, WAR, salary_adj, WAR_per_dollar) %>%
  kable(col.names = c("Year", "Player", "Position", "WAR", "Salary ($USD)", "WAR / $M"),
        caption = "Top 10 Under Performers (Lowest WAR per Million Dollars)",
        digits = c(0, 0, 0, 2, 0, 2)) %>%
  kable_styling(bootstrap_options = c("striped", "condensed", "bordered"))
plot1
plot2
plot3
plot4
plot5
plot6
plot7
plot8
plot9
plot10
top_10_over_table
top_10_under_table
```