Allen Primack

Data Analytics 401

Dr. Lavin

5/10/2021

Final Draft

The effect of the rise of baseball analytics on MLB free agency and player valuation

Abstract

In this paper, contract and performance data for Major League Baseball free-agent hitters from 1999-2019 are used to explore if Moneyball and the analytics movement in baseball changed the way free agent players are valued. Michael Lewis’ book Moneyball, released in 2003, brought awareness to the role of analytics in baseball, and influenced teams and fans to think about the game differently. I hypothesized that free agent players' statistics, specifically their more advanced statistics, have been considered more heavily by teams since Moneyball, and have become better predictors of their pay due to the increasing importance and presence of analytics in the sport. To test my hypothesis, I used python to conduct exploratory data analysis, regression analysis, and machine learning analysis, stata to generate multiple regression models, and tableau to create data visualizations. My results revealed that Moneyball and the analytics movement in baseball have changed the way free agent hitters are valued, and have significantly impacted which offensive statistics are the most valued by teams and players. Through rolling samples of regressions, I discovered that over time, free agents past statistics have become better predictors of their pay through regression models. Additionally, their advanced performance statistics are the strongest determinants of their pay, while their more traditional performance statistics were weaker determinants. By better-understanding how the determinants of contracts have changed over time, and which past performance statistics are the most correlated to players future performance, teams strategies of valuing a player’s advanced statistics over their traditional “baseball card” statistics can be validated. This analysis could also allow baseball fans to have a better understanding of the recent history of the dynamics of the free agent market for hitters.

Introduction

There is an abundance of existing literature on the topic of the relationship between pay and performance in major league baseball. A pattern in the literature is that long-term free agent contracts tend to have a lower value than short and medium length contracts, yet major league baseball teams are becoming more and more aggressive in pursuing free agents to improve their team. This leads many to believe that over time, teams have considered a player’s advanced statistics such as on base percentage more heavily when determining how much money to offer them. The release of Michael Lewis’ Moneyball in 2003 has popularized the use of less traditional statistics in evaluating player performance. The book highlighted the success of the Oakland Athletics, despite having the lowest payroll in the major leagues, through valuing players differently than other teams through the use of analytics. It also revealed how more traditional offensive statistics such as home runs and batting average have been over-valued by team’s decision makers and fans, while a player’s on base percentage is actually more important to their contributions to their team. The landscape of major league baseball free agency has changed drastically over the years with the rise of data analytics in the game as teams' decision makers employ different strategies in free agency to try to gain an upper-hand on their competitors. Due to this, it’s more difficult than ever before for teams to gain a competitive advantage over each other. My main research question is if Moneyball and the analytics movement in baseball have changed the way free agent hitters are valued.

Literature Review

An aspect of this topic that has been covered in-depth by scholars is which player performance statistics tend to be the most correlated to how much MLB free agent players are paid. Barnes, S.L. and Bjarnadóttir, M.V. (2016) studied which performance statistics contribute the most to how a player is valued. The authors use these statistics to build a model to assess whether or not free agent players live up to their contracts after signing. However, this study is limited to the most commonly-used statistics such as batting average, and fails to take into account more modern, advanced statistics. Their main focus is on the statistic WAR (wins above replacement), which is perhaps the most widely-accepted statistic to measure player performance. Barnes, S.L. and Bjarnadóttir, M.V. (2016) discovered that hitters and pitchers with the highest WAR, runs created, win probability added, rates of quality starts, and strikeouts were paid the highest. I conducted a similar analysis by considering similar performance statistics over a more recent time period. Bradbury, J. C. (2007) looked into which statistics are the most commonly used to value pitchers performance and argues that raw run-prevention statistics such as earned run average are overrated, and don’t reveal the true value of pitchers. Despite this, he concluded that teams favor signing pitchers who have the best run-prevention statistics, while overlooking the impact that their team’s fielders defense has on their performance. This reveals a blind-spot that MLB teams could have in not controlling for the pitcher’s defense, as how good it is impacts the pitcher’s ability to prevent runs. In my analysis, I only considered free agent hitters, but I will be using a similar approach to Bradbury, J. C. (2007) to understand which statistics are the most commonly used to value a hitter’s performance. Ehrlich, J.A. and Potter, J.M. (2020) concluded in their study that teams tend to significantly favor signing players with high offensive production over players with high defensive production. In doing so, they argued that teams are irrational in their approach, and should consider players defensive statistics more when valuing them. Analyzing which statistics teams consider the most when determining the value of free agents is highly important, as it can influence decision making. I will uniquely contribute to the existing literature through employing a greater quantity of more advanced performance statistics, for more recent free agency periods from 1999-2019. Additionally, my focus is specifically on hitters, and does not include pitchers. This analysis can be applied to other professional sports beyond baseball, such as the NBA and NFL.

Another aspect of this topic that has been covered in-depth by scholars is the definition of a good vs. bad free agent contract. A common theme in existing literature is that long-term contracts tend to be worse for teams, as a player’s productivity decreases over the length of a contract as they age. Although the short-term return of a long-term contract may be a good value for the team, players often fail to live up to expectations near the end of their contract. Barnes, S.L. and Bjarnadóttir, M.V. (2016) identified through statistical models that teams like the New York Yankees overpay to sign star-players such as Alex Rodriguez and Derek Jeter to long-term contracts. Solow, J. L., & Krautmann, A. C., & Oppenheimer, M. (2020) compare the value of notable free-agent contracts through a cost-benefit analysis for the teams that signed the players, concluding through an analysis of 152 long-term contracts that the benefits these players bring to their teams is typically less than the costs to their team in terms of their pay. Specifically, the authors conclude that teams overpay on contracts 3 years and longer, and the longer the length of the contract, the higher the likelihood that it ends up turning out poorly. My analysis will also challenge the argument of Solow, J. L., & Krautmann, A. C. (2020) that the value of a free agent’s contract should be determined based on their expected performance as opposed to their actual performance after signing a contract. They argue that,”the relevant economic question is not whether a player’s ex post performance justifies his salary, but whether his ex ante (expected) performance did so at the time the contract was signed” (Solow, J. L., & Krautmann, A. C., 2020, 3). I disagree with this, as I believed that it is fair to grade the contract based on the free agent’s performance after signing the contract as opposed to forecasted performance. I also believed that it was important to analyze players' performance in the years after signing their contract to truly understand if the contract they were awarded was good or bad, as free agents contracts cannot be adjusted mid-way through- the team is stuck with paying the player the agreed-upon amount regardless of their performance. I did this by analyzing the free agents’ WAR values in the years after signing their contract while being under contract in my exploratory analysis. My research adds to the existing literature by looking at the guaranteed money in free agent’s contracts and comparing it to the players performance/contributions to their team after signing it across various advanced statistics. The type of cost-benefit analysis employed in past studies and my study can relate to other industries such as investment banking and retail, where underperformance/overperformance and cost-benefit analysis are essential, along with other professional sports.

Another aspect of this topic that has been covered in-depth by scholars is changes in the market drivers of major league baseball free agency over time as MLB teams have been utilizing data analytics more. Barnes, S.L. and Bjarnadóttir, M.V. (2016) identified that more traditional player performance statistics such as batting average are being considered less and less over time as new, more meaningful performance metrics have been developed by Sabermetricians. This is what I studied- using more advanced statistics such as runs above replacement and weighted runs created in determining which ones are the most correlated with large free agent contracts. Brown, D. T., Link, C. R., & Rubin, S. L. (2017) research how major league baseball free agent contracts have adjusted since the release of Moneyball. Their results reveal that in the “post-Moneyball era” or after 2003, players with high on-base percentage have been more valued as free agents than in years past. They determine that players with high on-base percentages with less impressive batting averages and slugging percentages have been the most affected by Moneyball. The Oakland Athletics success as a team with less money in a smaller market undoubtedly inspired teams to employ similar strategies when signing players. Hakes, Jahn, K., and Raymond D. Sauer (2006) determined that after Moneyball’s publication, players who were overlooked in the past saw an increase in their pay proportional to their on-base percentage. It is clear through existing literature that teams’ free-agency strategies have changed. I build off of this by looking at trends in more recent years, comparing free agents’ statistics that are the most correlated with their contracts between the early 2000s and more recent years. I also examined if players with higher on base percentages received larger contracts after the release of Moneyball compared to before. I hypothesized that more advanced statistics, including on-base percentage, have been valued more heavily in the most recent years due to teams investing in data analytics at an increasing rate. This work relates to the field of economics broadly in terms of supply and demand.

Another aspect of this topic that has been covered in-depth by scholars is qualitative factors such as how successful free agents prior teams were that are related to the amount of money they are paid. The literature shows that a player’s on-field performance is not the only factor that determines the size of the contract they’ll receive in free agency. Ryan P. Terry, Jeffrey E. McGee, Malcolm J. Kass (2018) examined relationships between the amount of money free-agents were paid and qualitative factors such as attributes of the player’s previous team. The authors' results conclude that free agents that played for successful teams in their contract year tended to get paid more than players who came from poor-performing teams. Ryan P. Terry, Jeffrey E. McGee, Malcolm J. Kass (2018) filled a gap in the existing literature- that only quantitative factors were assessed as determinants of how much free agents got paid, by shedding additional light on how factors peripheral to a player’s on-field performance influence the free agent market. This work made me consider analyzing variables in my project that have not been considered in existing research, although I ultimately decided not to as this was outside the scope of my research question. The authors also state that their study is,”based on ex ante performance data, and future researchers should incorporate ex post data to more accurately assess the true return on investment in free agents (Ryan P. Terry, Jeffrey E. McGee, Malcolm J. Kass, 2018, pg. 12).” I aimed to do just that, incorporating players data in seasons after signing their free agent contract. Turner, Chad & Hakes, Jahn, 2007 analyzed the difference in free agent contracts between players of different ages, concluding that younger players get paid less than older players with similar statistics. I build off of this study by looking further into how age affects the amount of money a free agent player is paid in my exploratory analysis and regression/machine learning models. These articles helped me to consider potential qualitative factors determining how much a free agent player is paid aside from just quantitative ones. Qualitative analysis such as this can also be applied to other professional sports, along with investment banking, by considering factors that influence pay/prices beyond available quantitative statistics.

From reviewing existing literature, it is clear that there are a variety of factors that determine the size of the contract a major league baseball free agent receives. Over time, these factors have changed, as more major league baseball teams are using data analytics to find players that are hidden gems. This is a trend that has taken off since the release of Moneyball, but whereas in the early 2000s teams like the Oakland Athletics who had relatively larger analytics departments had a competitive advantage in terms of making smart decisions, it is now more difficult as all teams have invested more into data analytics. The definition of what can be considered a good free agent contract has been consistent- players whose performance statistics such as their Wins Above Replacement (WAR) after signing their contract exceed their earnings relative to other players. Also consistent is the fact that longer-term contracts tend to be less valuable to the team than shorter- or medium-term contracts, due to players performance declining after a certain age/playing a certain number of seasons. These takeaways and more make the study of baseball player valuation so interesting, and valuable to professional baseball team’s decision-makers. As the use of baseball technology increases, there will be more performance statistics that could be analyzed to determine player value. My research represents a next step to gaining scholarly knowledge in this area, as it addresses many of these topics for future discussion that are not as present in the current literature. The substantial number of player performance variables I analyze against the amount free-agent players get paid is unique and extends beyond just the commonly used wins above replacement statistics. The models I develop to predict contract amounts of free agents are unique in the variety of statistics used and time periods considered. Prior research on this topic does not include rolling sample regressions, as I do. Lastly, I analyze free agent contract and performance data from before and after the release of Moneyball, to test if the book influenced teams and players to value statistics such as on-base percentage significantly more in free agents. My hope is that professionals working for MLB teams can learn from my analysis to make smarter free-agent signings going forward. By understanding how the determinants of contracts have changed over time, and which past performance statistics are the most correlated to players future performance, teams’ decision makers can benefit greatly. I am excited for the potential that my analysis could offer to help improve professional baseball teams’ strategies on the free-agent market.

Ethical Considerations

The domain of sports analytics is less controversial than other domains, and naturally has less ethical considerations. Given that the data I use is publicly available, there are no main ethical considerations with the data itself. All contract and performance data are accessible to the public. One ethical consideration is that I do not work for a professional baseball team, and do not have the same training as employees of teams. Teams may have more insight into the character/make-up of free agent players, and some of the free agents whose information I analyze may have done bad things off the field that have negatively impacted society. For example, labeling a free agent player who performs well yet has done bad things off the field as a good player. Teams have more information on free agents than I do as a fan, and the way I interpret my results may not be suitable for team decision makers. The same could be said with my methods. If my findings are taken too seriously by team decision makers, without being validated with their insights, it could result in poor decisions being made, as my analysis has its limitations. Scouts play an important role in evaluating players beyond their statistics, which is why teams employ them. Teams do not rely solely on numbers in their decision making. Thus, my conclusions should be taken with a grain of salt and encourage teams to conduct further research rather than automatically accepting my results.

Data

I analyzed 884 free agent position players from the 1999-2019 MLB free agency periods, 492 of which are unique. I used two datasets for my analysis to create one comprehensive dataset. The first dataset I worked with contains historical MLB free agent contract data from 1999 to the present. It comes from a website called “Cot’s Baseball Contracts” (*Cot's baseball CONTRACTS*). This is publicly available data with no permissions. I cleaned the data to only include the years 1999-2019, along with changing the headings to be the contract information variables. As Moneyball was released in 2003, the chosen years allowed me to analyze changes in the determinants of players contracts before and after its release. I also filtered out all values with missing data, such as free agent players that never received contracts. All free agents who signed contracts for at least 1 season were included. The original format of the data was a google sheet, and I exported it to an excel file.

The second dataset I worked with contains player performance statistics from 1999-2019 for all of the free agent hitters from the other dataset. They come from the website “Fangraphs”, in their Leaders section (*FanGraphs baseball: Baseball statistics and analysis*). I created custom reports to only include the statistics that I deemed to be relevant for my analysis, and exported player performance data across different seasons. This is also publicly available data with no permissions. I was able to easily export the data based on my report, and did separate exports for each individual season. The variables in this data that I considered are a mix of traditional performance statistics and more advanced performance statistics. I initially considered all of the available statistics on fangraphs in my analysis, but decided to narrow them down to only those that I deemed as the most relevant. I decided to only include a player’s performance data from their past 2 seasons played from the date of their free agent contract. I believed that considering each player’s past two seasons allowed for a more holistic picture of their performance compared to just one year, as players who get injured the season before becoming a free agent and miss many games are often judged by their most recent season when healthy. I considered any performance statistics before or after more than 2 years from their free agent year as irrelevant to the contract amount. Thus, for free agents who had more than 2 years of playing experience, only statistics from their 2 seasons prior to them signing were included.

I combined the data from both sources by using the player’s first and last name to connect them. Once I cleaned the free agent contract data, I filtered the fangraphs data. Fangraphs has a feature that allowed me to easily filter out the free agent players, and display their relevant statistics for their 2 seasons prior to signing. I repeated this process for each year of free agency. For example, for the first year of free agents I analyzed, 1999, I exported the players performance statistics from the 1997 and 1998 seasons. I repeated this for each free agent class. This allowed me to pair the players’s statistics with their contract information and effectively combine both datasets. In addition to including all of the player’s relevant past performance statistics, I also included the players WAR statistics for the seasons where they were under contract. For example, for a player signed in 1999 that was under contract for 4 seasons from the 1999-2002 seasons, their WAR after signing statistics reflected their combined WAR for the 1999, 2000, 2001, and 2002 seasons. I repeated this for each player, looking at the term variable to identify which seasons each player was under contract for. This allowed me to not only analyze the players performance in the 2 seasons leading up to signing their contract, but also their performance in the years after signing their contract.

The graph below displays the number of free agent hitters and the average free agent contract amount for each free agency period analyzed.

Figure 1- Number of free agent hitters and the average contract amount by year



Between 2004-2017, there were between 40-60 free agent hitters each year, with the years prior to and following having less. After 2017 has significantly less, as I only considered players who play out the duration of their contract, thus free agents that signed longer contracts during these years are excluded. The average contract amount values over time reflect the large role that money plays in MLB free agency, as for most years the average contract amount falls between $10 million and $15 million. It is clear that the pool of free agent hitters remains consistent over time.

The key variables in my data that I considered in my analysis are explained in the following tables:

Figure 2- Definitions of Key Variables

|  |  |
| --- | --- |
| Variable Name | Definition |
| Pos’n | The player’s position. |
| Age | This is the age of the player at the time of them signing their free agent contract. |
| Years | The number of years in the player’s contract. |
| Guarantee | The total amount of money in the player’s contract. This is one of the key dependent variables I consider. |
| Term | The specific years when the player is under contract. |
| Player Agent | This is the player’s agent who helps them get their contract. |
| Club Owner | The owner of the team/new club that signed the free-agent player. |
| Baseball ops head/club gm | The person responsible for the team’s decision to sign the free agent player. |
| Year | This is the year the player signed their contract. This variable is a unique identifier to tell players apart by the year they signed their contract. |
| Future WAR | This is the player’s WAR in the years after signing their contract, taking into account all years during the contract. This is the other key dependent variable that I consider. |
| Performance Statistics: Past WAR-Clutch | These are the player’s past performance variables in the 2 years before signing their contract that I mentioned. These will be key independent variables. I list the key statistics below. A full list and description of the performance statistics for hitters that I consider can be seen through this link: https://library.fangraphs.com/offense/offensive-statistics-list/ (Weinberg, *Complete list (offense)* |

Figure 3- Definitions of Key Baseball Performance Statistics (\* denotes an advanced statistic)

|  |  |
| --- | --- |
| Past WAR\* | Wins above replacement- an advanced statistic that estimates how many more wins a player has been worth to his team compared to a replacement-level player at his same position |
| OBP\* | On-base percentage- the “Moneyball” statistic- the rate at which the hitter reaches base safely |
| AVG | Batting Average- The player’s number of hits divided by their plate appearances |
| HR | Home Runs- The player’s number of home runs- hits that go into the stands and stay fair |
| K% | Strikeout Percentage- How often a hitter strikes out- strikeouts divided by plate appearances |
| SB | Stolen Bases- The number of stolen bases a player has |
| BB | Walks- How often a hitter walks- walks divided by plate appearances |
| IBB | Intentional Walks- The number of times the player was intentionally walked |
| WRc\* | Weighted Runs Created- The number of runs a player has generated for his team as a result of his playing time |
| RAR\* | Runs Above Replacement- How many more runs a player contributes his team compared to a replacement-level player |
| WPA\* | Win probability added- the total impact a batter’s plate appearances have on his team’s win probability compared to league average |

Methods

I conducted my quantitative analysis for this project using jupyter notebooks with the programming language python, along with stata and tableau software. I began by importing and preprocessing my data in python. In my analysis of the free agent contract information for players, I identified the impact of each independent variable on the size of the player’s contract, along with their future WAR. After completing exploratory data analysis, I exported my results to csv files, then imported them into Tableau. With tableau, I created data visualizations that allowed me to identify relationships between free agents pay and the independent variables I considered over time.

After analyzing the variables from the contracts dataset, I analyzed the variables from the performance datasets alongside them. My objective was to identify which performance statistics were the most correlated with free-agent contract values (the Guarantee dependent variable) and future performance (the future WAR dependent variable) over time. I did this by analyzing the correlations between key independent variables and these two dependent variables. This allowed me to see if some statistics became more/less correlated with contract amount/future WAR over time. I did a separate correlation analysis for separate seasons to track the change in how correlated each variable is with the contract amount over time. I then exported the correlation data to Tableau to produce visualizations showing trends in the correlations between the statistics over time.

Next, I developed OLS regression models in python to not only predict the value of a free agent’s contract based on their past performance statistics, but also to understand changes in the performance of the models and significance of the variables over time. I discovered that multicollinearity was a large issue when running my initial regressions. This is unsurprising, as many offensive baseball statistics measure the same thing. For example, the variable OPS (on-base percentage plus slugging), is its own variable, but the variables on-base percentage and slugging percentage are included in it. Thus, multicollinearity exists between these three variables. There were many instances of multicollinearity, and I eliminated problematic variables by generating VIF values for each. Variables with VIF values above 10 were identified and isolated from my updated models, as this is a common technique to fix multicollinearity (*Variance inflation factor* 2020). After this, I had a more narrow set of performance statistics to include in regressions. I next removed variables that were not statistically significant in the models at the conventional levels, along with those that did not have any economic significance. From there, I one-hot encoded categorical variables to convert them into dummy variables for the model, as this is a common way to solve the problem of categorical data not working in a machine learning model (Brownlee, 2020). I split the data into a training and testing set in order to evaluate the predictive power of the models using the train\_test\_split function. This allowed me to split my data into subsets that minimize the potential for bias in my evaluation and validation process (Real Python, 2021). I then decided to try normalizing the data, as the free agent contracts may not be normally distributed. I checked the model’s quality with and without normalization, and the results were similar. Due to this, I decided to avoid normalization. I compared initial models and determined which ones were the best performing, and which combination of performance statistics best-predicted a free agent’s contract amount. This regression analysis laid the groundwork for a more rigorous analysis in Stata.

I decided to implement my regression models in Stata. I made this decision due to having a higher comfort level of completing regression analysis in stata. This was a matter of personal preference, as I find Stata's software to allow for tables of regression results to be created and shared more easily, along with the code being easier to replicate in my opinion.

Once I finalized my base regression model, I ran two sets of OLS regressions for two different time periods- before and after the release of Moneyball. This allowed me to analyze if there were significant changes in the influence of certain independent variables on a player’s contract amount before and after Moneyball, along with the model's predictive power. I defined the pre-Moneyball time period as the years 1999-2004, and the post-Moneyball period as the years between 2005-2010, as Moneyball was released in 2003. This accounted for the lagged effect of teams incorporating Moneyball methodologies in their free agency strategy.

I also ran twelve additional regressions that were rolling samples, to demonstrate changes over time more accurately than comparing results from 1999-2004 to 2005-2010. To see if free agents' past performance statistics were becoming more predictive over time, I decided to create models over 7 year increments. These included the time periods 1999-2006, 2000-2007, 2001-2008,..., 2010-2017, etc. I compared the adjusted r-squared values for each model to identify changes in the models predictive power over time. I then used tableau to display the adjusted r-squared value for each of the 12 models.

After interpreting my OLS regression results, I tested out different machine learning models in pycaret, which is commonly used in machine learning projects (Brownlee, 2020). This enabled me to easily compare and evaluate different machine learning models such as K Nearest Neighbors, Random Forest Regressor, and Gradient Boosting Regressor for days from both the pre-and-post Moneyball periods. The two strongest machine learning models in predicting contract amount were the Gradient Boosting Regressor and Random Forest Regressor. I originally chose to use the Gradient Boosting Regressor, as it is known to be one of the best machine learning techniques developed, and it usually provides more robust results with less effect of overfitting than the machine learning models (Brownlee, 2020). However, I ultimately decided to use the Random Forest Regressor, as it had a significantly higher R-squared value than the other models. After this, I generated feature importance plots from the models with contract amount as the dependent variable for three different time periods: the pre-Moneyball, post-Moneyball years from before, and the full time period.

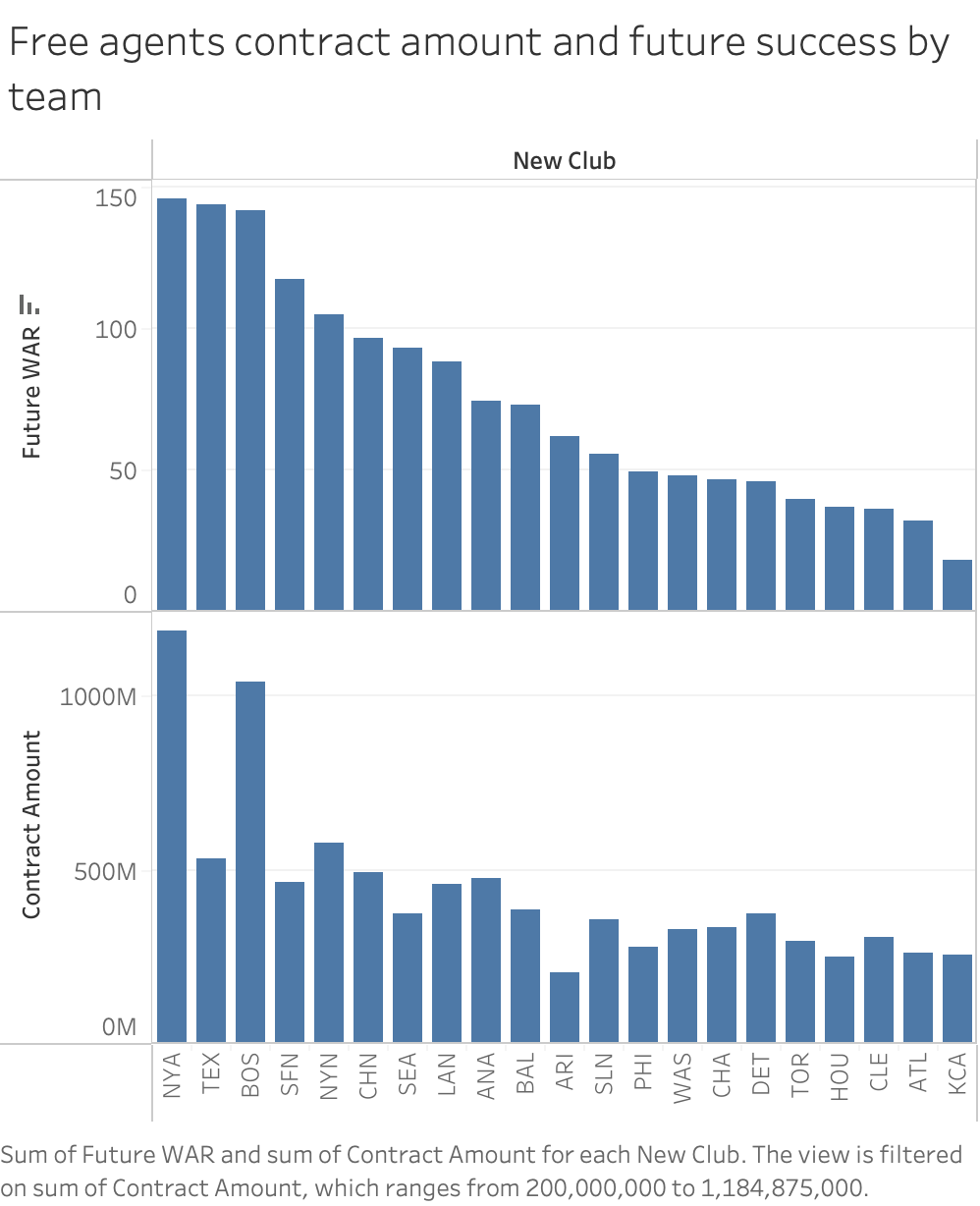
I also tested out regression and machine learning models with free agents' future WAR as the key dependent variable. However, I discovered that the free agents' past performance statistics considered were poor predictors of their future WAR. The R-squared values were either very small or negative, which evidenced a lack of correlation between future WAR and the other statistics. This may have been due to the past performance metrics being more of a short-term calculation than a long-term aggregation. With that, a player’s physical and emotional conditions along with luck could have had more of an impact on their future success than their past performance statistics, which is why they can vary widely across different seasons. Due to these issues, I decided to only include the regression models predicting free agents contract amount in my results.

Results and Interpretation

Exploratory Data Analysis

To begin my analysis, I conducted exploratory data analysis. I began with analyzing teams spending on free agents, and their subsequent production from 1999-2019.

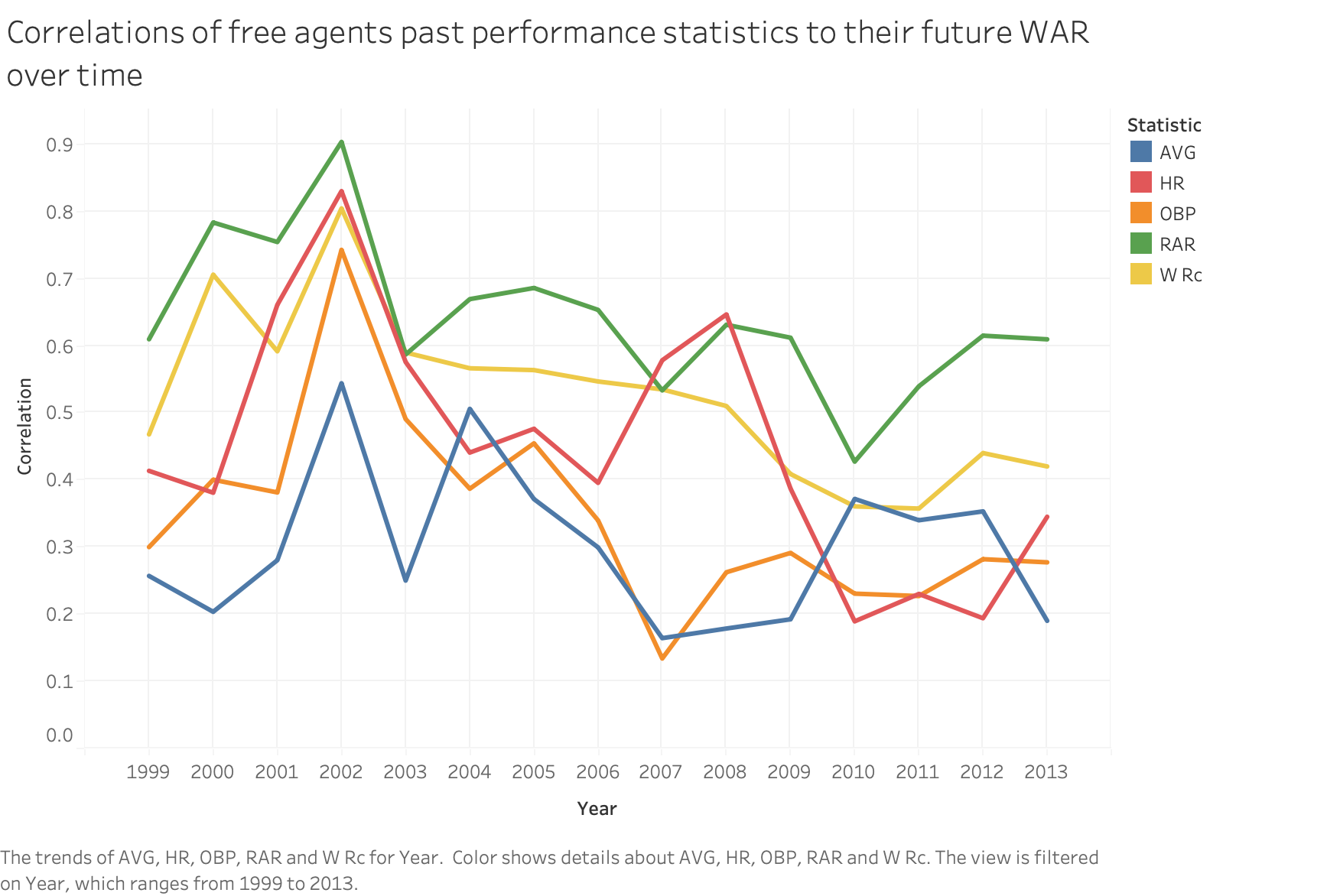
Figure 4- Teams spending on free agents and their production from 1999-2019



It’s clear from the graph below that some teams have historically spent more money on free agent players than others. Teams in larger markets such as the New York Yankees and Boston Red Sox have spent the most money on free agent players than other teams. This is unsurprising, as these teams play in bigger cities and have some of the largest payrolls in baseball. Free agents signed by small market teams such as the Texas Rangers and San Francisco Giants have exceeded expectations by performing above their pay compared to teams that paid more money to free agents. This also reveals the gap between the teams in largest markets, and smaller-market teams like the Oakland Athletics, who aren’t even on this list, evidencing a key message from Moneyball- that teams in smaller markets need to correctly identify undervalued players on the free agent market in order to succeed. This could also mean that players could see these teams as the most appealing due to their name recognition and market size, along with usually having the ability to receive more money from them, allowing them to attract more free agent talent.

I then analyzed the correlations of free agents past performance statistics with their pay over time.

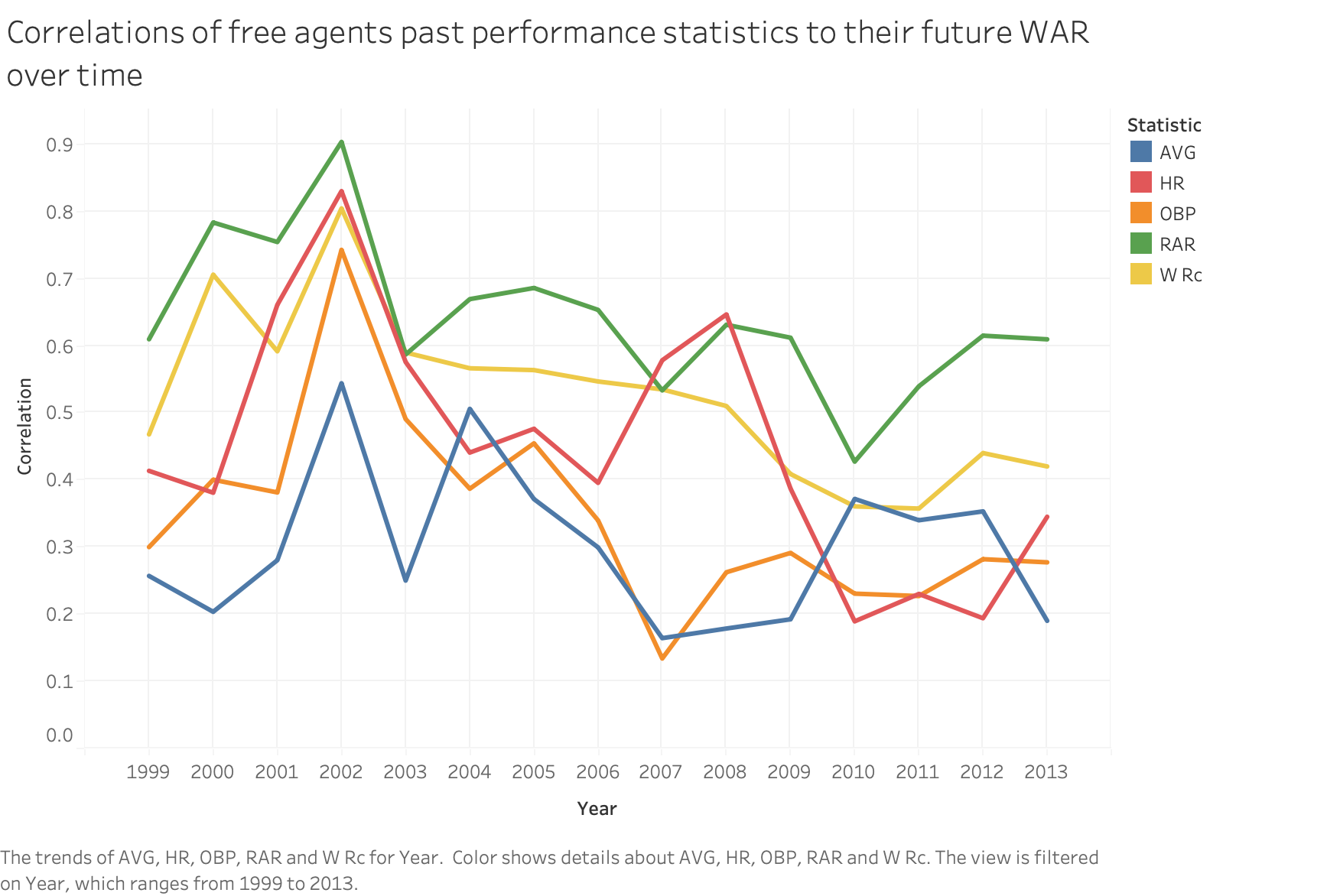
Figure 5- Correlations of free agents past performance statistics to their pay from 1999-2013



This graph shows that free agents past advanced performance statistics such as RAR and wRC have more correlation with the size of their contracts than traditional ones. Surprisingly, directly after the release of Moneyball, the correlation of OBP decreases. There is no clear trend in the correlations between past statistics over the full time period considered. However, for some statistics, like home runs, there are clearer differences between years. This could reveal the variation in how players are valued. For example, in some free agency years teams may have valued defense more, with defensive statistics becoming more correlated to contracts than offensive statistics. I believe that these results show that a year-by-year analysis of changes in the significance of free agents performance statistics in determining contracts amounts was not the best approach, due to the variation and inconsistency over time. Additionally, with there only being around 40 free agent hitters for each year, this was a very small sample to base conclusions off of.

After this, I analyzed the correlations of free agents past performance statistics to their future performance over time, to see if the results were similar.

Figure 6- Correlations of free agents past performance statistics to their future performance from 1999-2013

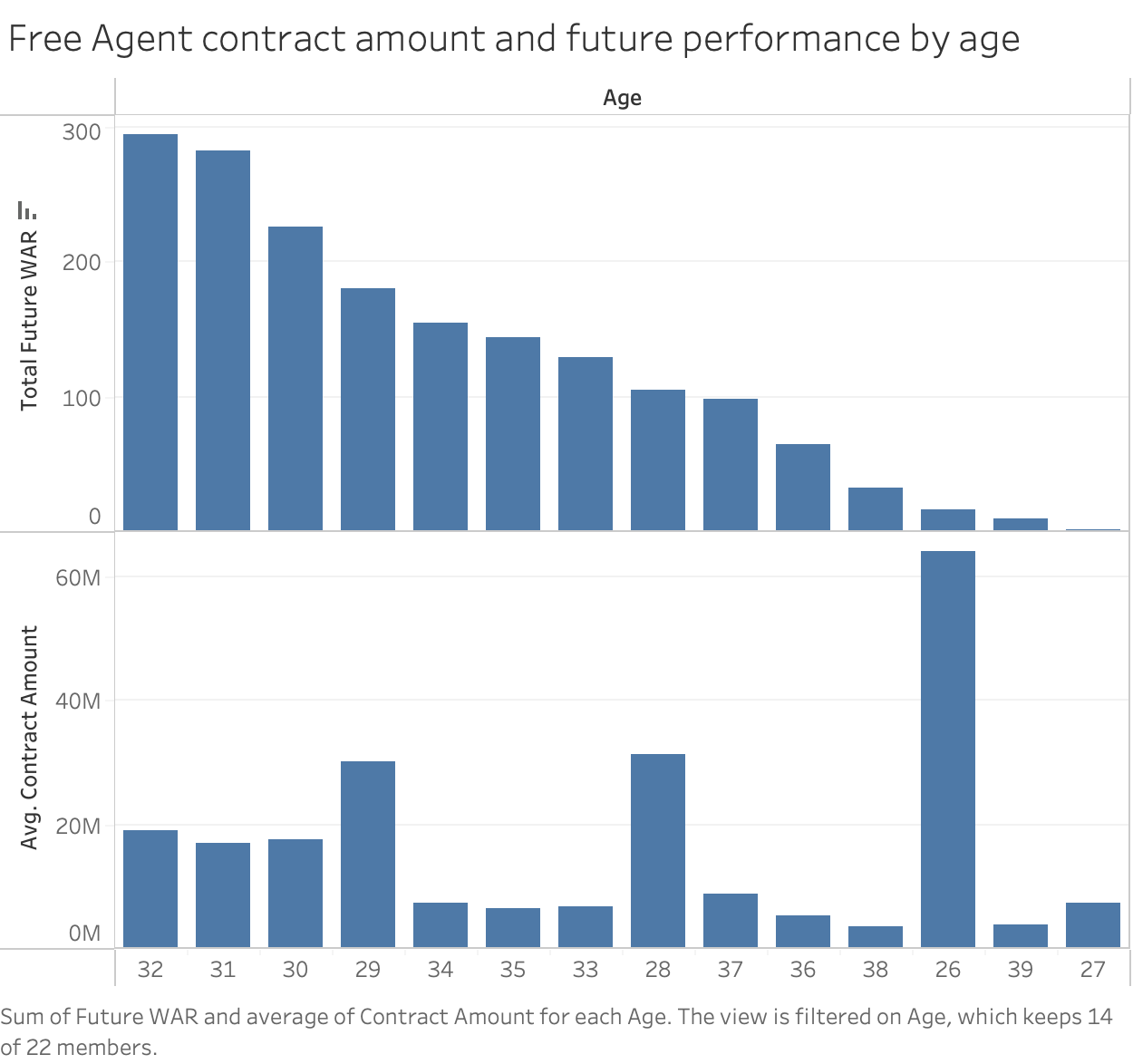


This graph reveals that the correlations of free agents past performance statistics to their future performance are similar to the correlations to their pay. It’s clear that the more advanced statistics such as RAR and wRC are consistently the most correlated statistics to players' future performance. As with the previous graph, it’s observed that there is no clear increase/decrease in the correlations over time.

These correlation graphs reveal that a regression analysis of the past performance statistics determining free agents pay and future performance is more appropriate, with larger samples of players compared over time.

I then analyzed free agents pay and future performance by age from 1999-2011.

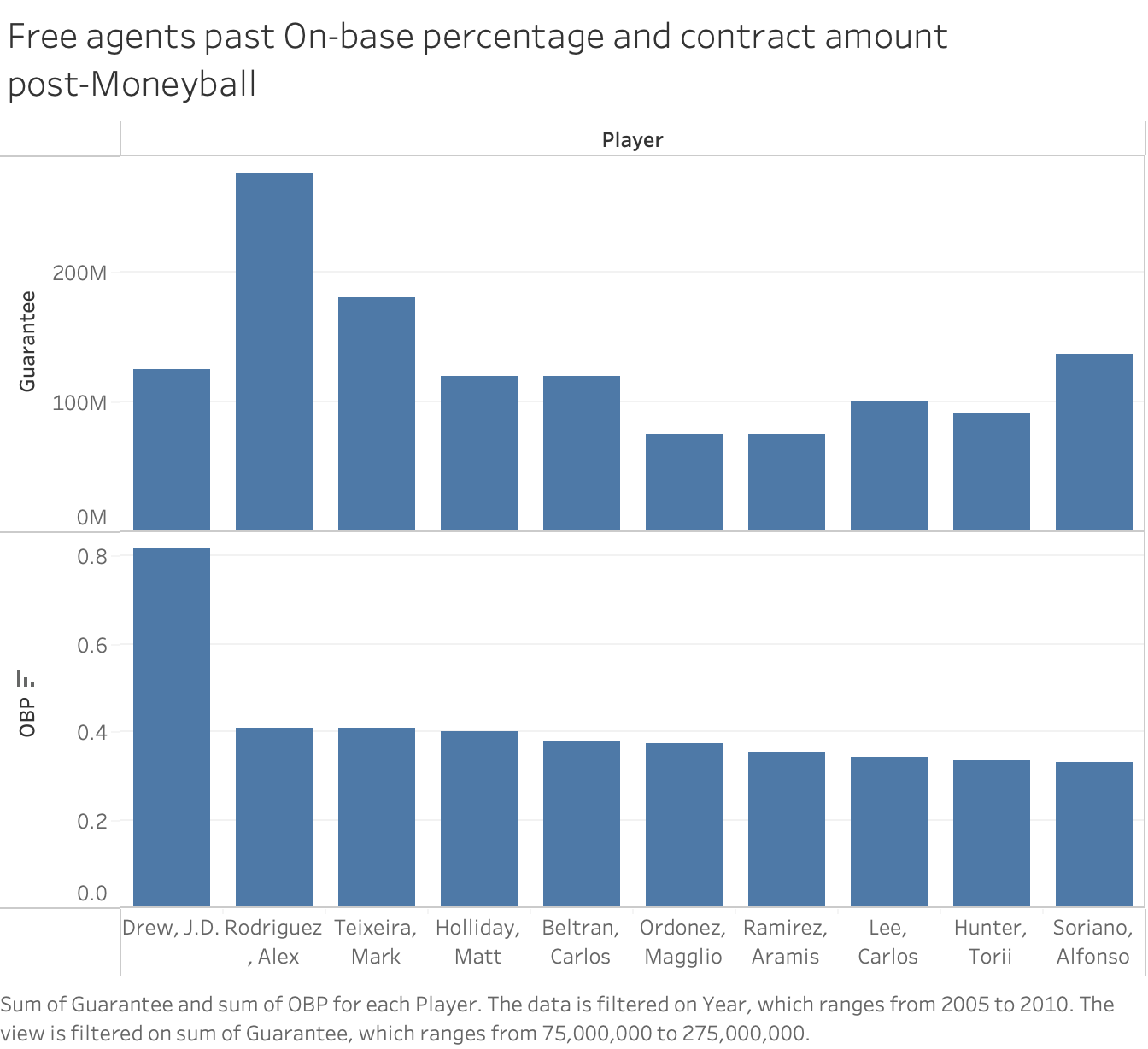
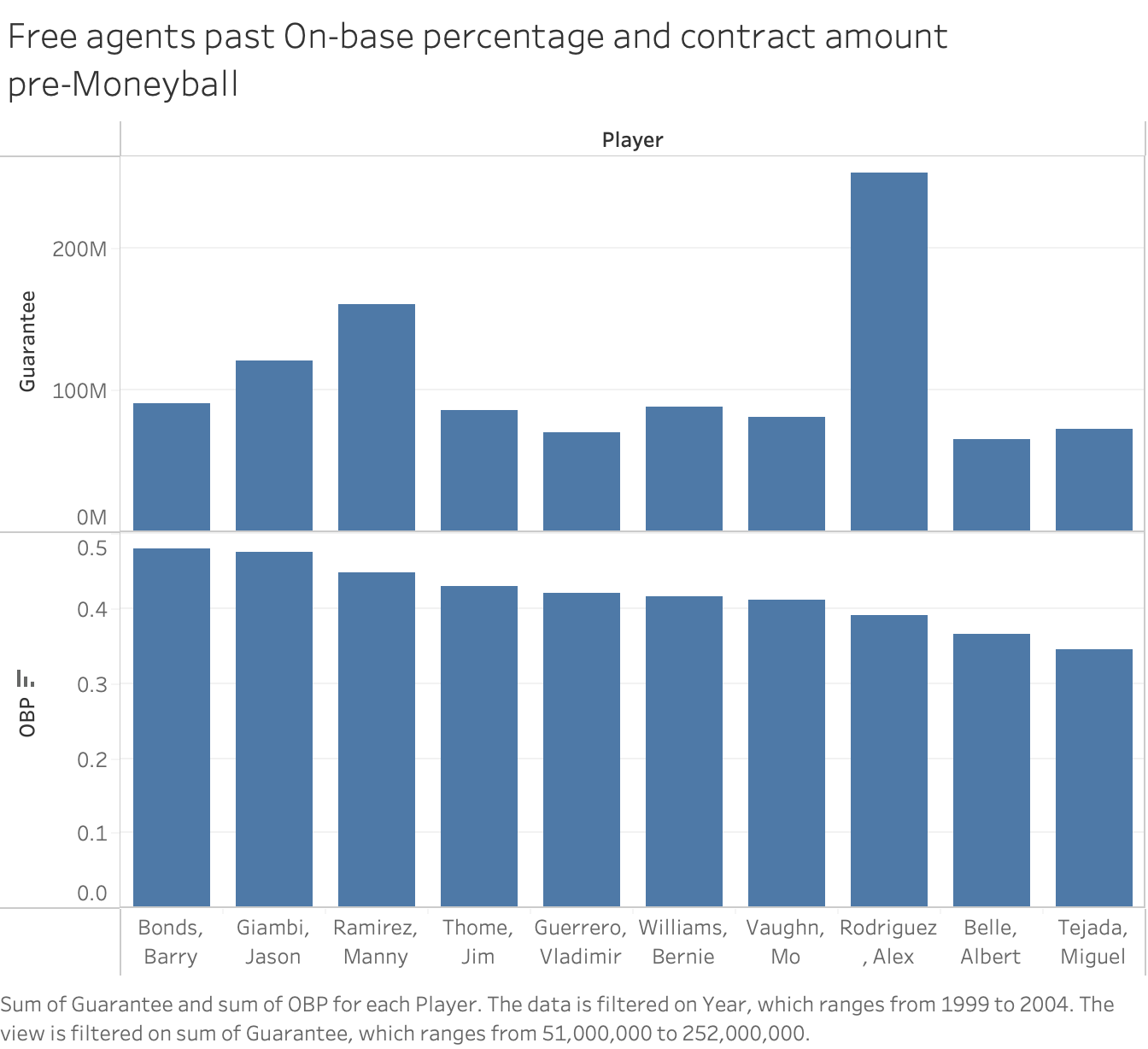
Figure 7- Free agents pay and future performance by age from 1999-2018



This graph reveals that free agents between 29-32 years-old at the time of signing their contract had the best future performance. It’s clear that younger free agents tend to receive larger contracts, while older free agent players receive smaller contracts. This makes theoretical sense, as younger players have less games played and less wear and tear on their bodies, enabling them to play more seasons after singing than older players. However, players younger than 29 years old have a much lower total future WAR as it’s far less common for younger players to be on the free agent market at that point in their careers.

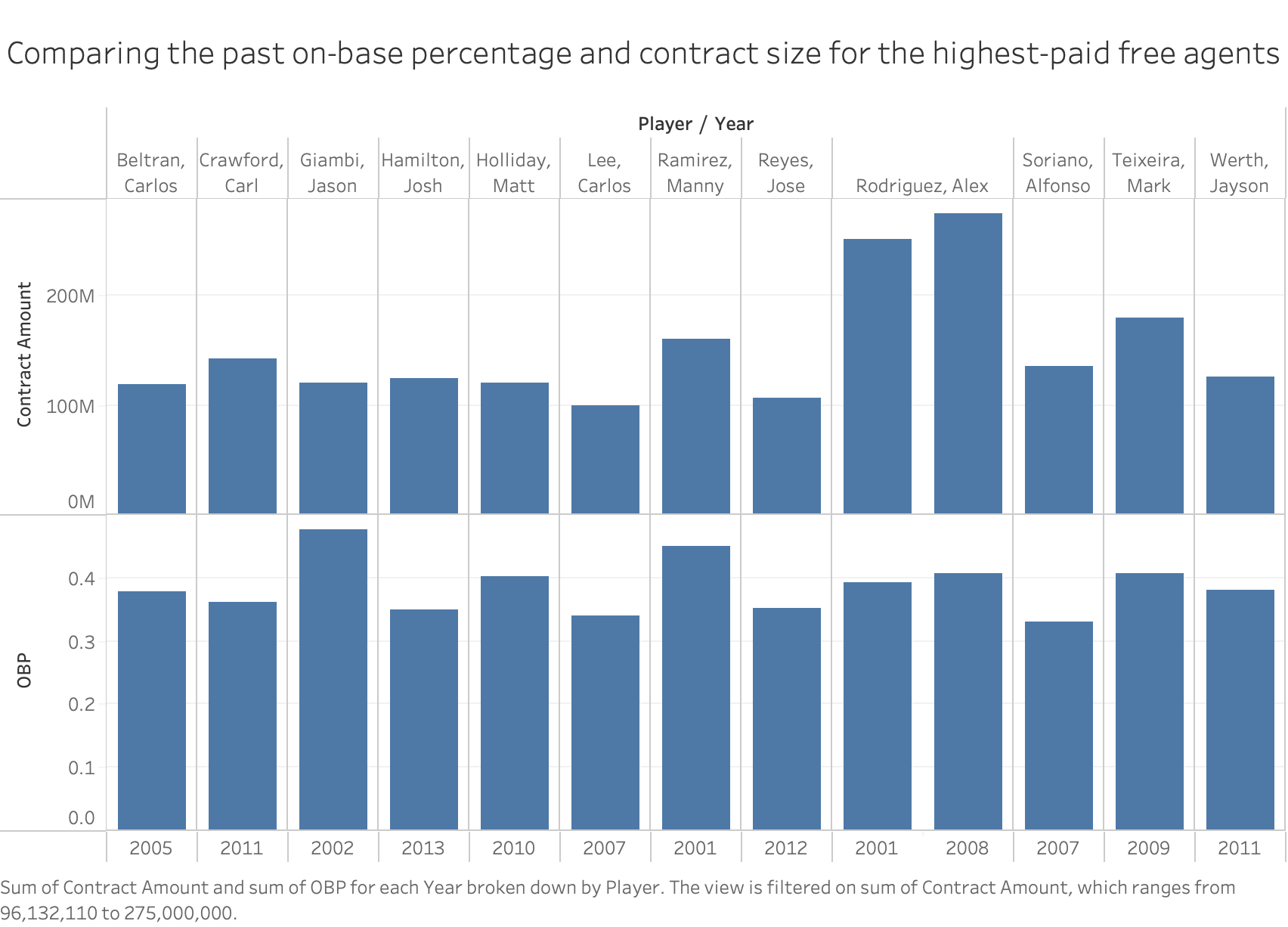
I then analyzed the past on-base percentage of the highest paid free agents pre-and-post Moneyball.

Figures 8 and 9- The past on-base percentage of the highest paid free agents pre-and-post Moneyball

These results reveal that free agents’ past on-base percentage factored more into their pay post-Moneyball than pre-Moneyball. Players with the highest on-base percentage had the highest contract amounts for the most part post-Moneyball, while in pre-Moneyball the players with the highest on-base percentages often received a smaller contract. For example, Manny Ramirez and Alex Rodriguez received the largest contract of all of the pre-Moneyball free agents despite having lower on-base percentages than the top two, while post-Moneyball, Alex Rodriguez and Mark Teixeria had the 2nd and 3rd highest pass on-base percentages and also received the largest contracts. This could reveal that teams placed more emphasis on players on-base percentage than other statistics when determining their value post-Moneyball, conforming to my hypothesis.

I explore this further by analyzing the past-on base percentages of the twelve highest free agents between 1999-2018.

Figure 10- The past on-base percentage and contract amount for the highest-paid free agent hitters from 1999-2018



It’s evident that in more recent years, players with the highest on-base percentages were also likely to be among the highest-paid, compared to older years. For example, Jason Giambi, a 2002 free-agent, had the highest past on-base percentage, but received a significantly smaller contract than Manny Ramirez’s in 2001. However, in more recent years, players seemed to receive a contract proportional to their on-base percentage, as seen by Mark Texiera in 2009, Carl Crawford in 2011, and Matt Holliday in 2013.

To conclude my exploratory data analysis, I identified the free agents who proved to be the most valuable in comparing their WAR after signing their contract, to their pay.

Figure 11- The highest-value free agent contracts from 1999-2018



It’s clear that all of the highest-value free agent contracts were one or two year deals, seen through the length of the players term values. This supports findings in existing literature that short-term contracts tend to be better values for teams than long-term contracts. When analyzing the full sample of players, free agents signed to longer-term contracts had far lower values on average.

Regression Analysis

After conducting exploratory data analysis, I generated regression models. As mentioned in my method sections, I removed all variables causing multicollinearity in my models, and removed statistically insignificant variables.

I first ran two separate regression models, for my defined pre-Moneyball and post-Moneyball time periods. This allowed me to compare both models to demonstrate changes in their predictive power over time.

Figure 12- Regression Results Pre-and-Post Moneyball (Contract Amount is the Dependent Variable)

|  |  |  |
| --- | --- | --- |
| Time Period | Pre-Moneyball | Post-Moneyball |
| \_cons | -1.045e+09 | -1.626e+09 |
|  | (6.505e+08) | (1.158e+09) |
| Year | 504360.267 | 804430.806 |
|  | (323249.959) | (576133.789) |
| Age | 51124.148 | 198765.328 |
|  | (252400.563) | (271334.292) |
| Years | 12915576.118 | 17036232.199 |
|  | (708668.876)\*\*\* | (835686.351)\*\*\* |
| pastWAR | 556190.039 | 941468.858 |
|  | (405023.004) | (520510.977)\* |
| RBI | 7164.083 | 32723.649 |
|  | (21411.390) | (25722.588) |
| IBB | -23847.558 | 417991.603 |
|  | (128174.247) | (157744.954)\*\*\* |
| SB | -28490.487 | -73654.237 |
|  | (40356.299) | (58671.423) |
| AVG | 35019833.117 | -6.509e+07 |
|  | (38003749.304) | (45220077.587) |
| BB | 31833809.084 | -2.568e+07 |
|  | (27741110.316) | (33226205.078) |
| K | 6289518.138 | -4718437.609 |
|  | (19292646.662) | (20976254.760) |
| ISO | 22191356.635 | -3000253.416 |
|  | (21315169.806) | (23358064.319) |
| WPA | 473727.152 | 814339.576 |
|  | (439229.381) | (506251.676) |
| N | 292 | 227 |
| Adj R-Squared | 0.789 | 0.843 |
| SEE | 11380579.170 | 11180644.796 |
| F-ratio | 91.512 | 102.119 |
| SSR | 3.614e+16 | 2.675e+16 |

These results reveal that the post-Moneyball model is more predictive of free agents than the pre-Moneyball model, with an adjusted r-squared value of 0.843 compared to 0.789. The Past WAR variable is statistically significant at the 10% significance level in the post-Moneyball model, while not being statistically significant in the first model. The intentional walks variable also becomes statistically significant in the more recent model. Interestingly, in the post-Moneyball model, the coefficients of the more advanced variables are more positive, and the coefficients of the more traditional statistics are more negative than in the pre-Moneyball model. For example, the coefficient of the WPA variable goes from 473,727.152 in the pre-Moneyball model to 814,339.576 in the post-Moneyball model, while the coefficient of the batting average variable goes from 35,019,833.117 to -6.509e+07. The only exceptions to this are the RBI and IBB variables. This supports my hypothesis that more advanced performance statistics have become stronger determinants of free agents' pay after the release of Moneyball, and that a mix of statistics become better predictors of their pay over time.

I then decided to compare the results of rolling samples of regressions for different time periods throughout the 2000s over 7-year intervals. This also allowed me to identify if the statistics I consider become more predictive of free agents' pay over time. A different mix of variables was considered than in the previous models, as I completed a new multicollinearity analysis for these time periods, and some of the variables removed that were removed differed from the previous models.

I compared results from twelve 7-year intervals in total. The results for each of the time periods are displayed in the appendix section at the end of this paper. This table presents the regression results from the first and last time periods considered.

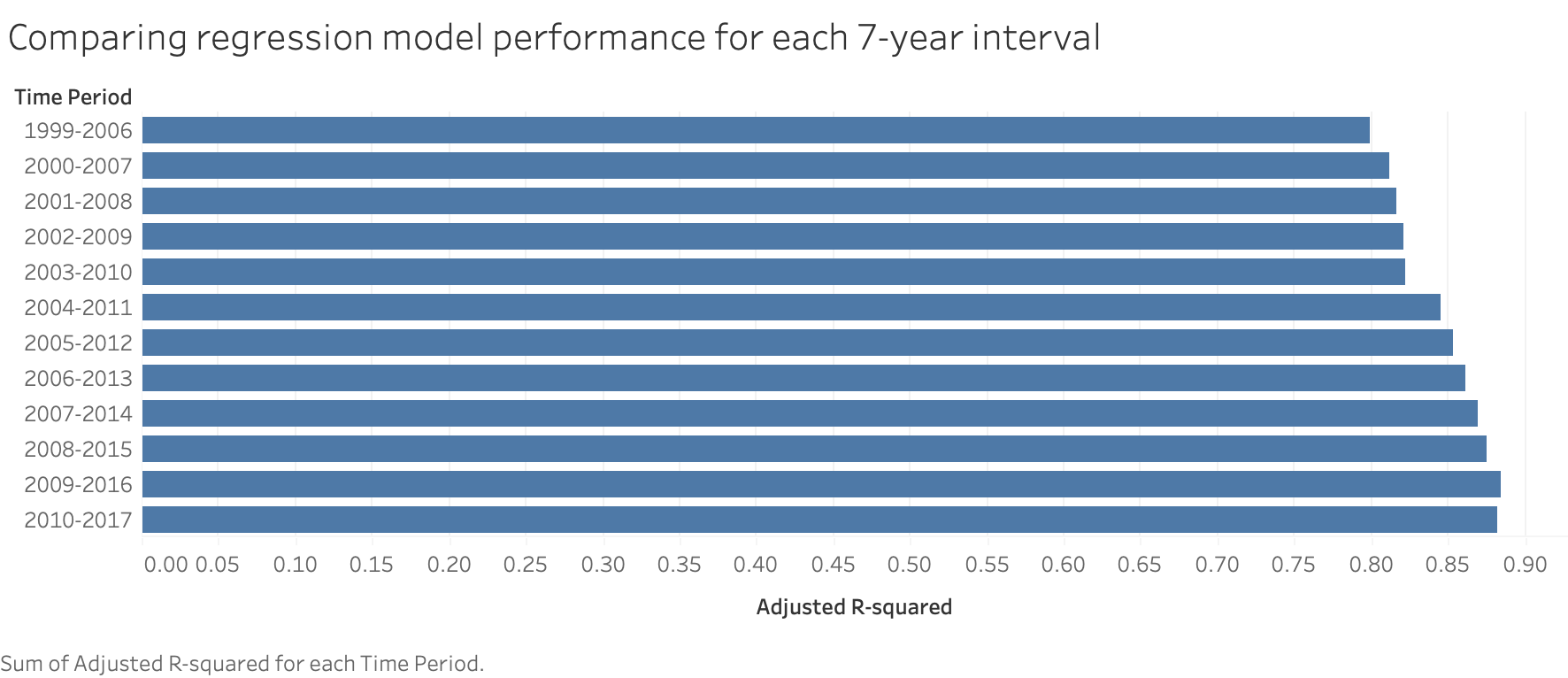
Figure 13- Regression results from 1999-2006 compared to 2009-2016 (Contract Amount is the Dependent Variable)

|  |  |  |
| --- | --- | --- |
| Time Period | 1999-2006 | 2009-2016 |
| Constant | -1.136e+09 | -1.028e+09 |
|  | (4.979e+08)\*\* | (3.965e+08)\*\*\* |
| Year | 549626.626 | 510196.565 |
|  | (247753.765)\*\* | (195993.564)\*\*\* |
| Years | 12545989.234 | 15623135.293 |
|  | (560536.320)\*\*\* | (487383.724)\*\*\* |
| HR | 284181.963 | 109401.946 |
|  | (88113.066)\*\*\* | (63572.848)\* |
| IBB | -72780.790 | 394860.831 |
|  | (114126.823) | (97510.765)\*\*\* |
| OBP | 89097823.538 | -4.841e+07 |
|  | (34584801.404)\*\* | (26532626.346)\* |
| SLG | -2.239e+07 | -5883210.397 |
|  | (19904711.718) | (13255868.989) |
| wRC | -69607.790 | -21497.805 |
|  | (26811.467)\*\*\* | (20385.325) |
| RAR | 83144.778 | 135841.886 |
|  | (36717.489)\*\* | (26536.655)\*\*\* |
| WPA | 335468.922 | 948185.969 |
|  | (375907.531) | (283743.881)\*\*\* |
| N | 336 | 361 |
| Adj R-Squared | 0.799 | 0.884 |
| SEE | 10578098.799 | 7793406.919 |
| F-ratio | 149.148 | 307.184 |
| SSR | 3.648e+16 | 2.132e+16 |
| Standard errors in parentheses | \* p<0.10, \*\* p<0.05, \*\*\* p<0.01 |  |

Similar to the previous models, these results show that the past performance statistics of free agents considered become stronger determinants of their pay over time, with the model from the years 1999-2006 having an adjusted r-squared value of 0.799 compared to 0.884 in 2009-2016. More advanced performance statistics that were not statistically significant in the earlier model become statistically significant in the more recent model. For example, in the model considering 2009-2016 RAR is statistically significant at the 1% significance level compared to 5% in the earlier model, while WPA is statistically significant at the 1% level in the more recent model but not statistically significant in the earlier one. This could reveal that free agents advanced performance statistics have indeed become stronger determinants of their pay over time. From the coefficients and statistical significance of each variable, it’s evident that the more advanced performance statistics become more of a factor in the prediction of contract amount over time. The home run statistic becomes less statistically significant over time, and as with the previous models, the coefficients of the more advanced statistics become more positive over time while the coefficients of the more traditional statistics become more negative. The statistics that become more predictive over time are intentional walks, RAR, and WPA. RAR is the most advanced statistic in my model and is also the most predictive. I believe that these results demonstrate a clear Moneyball effect over time, supporting my hypothesis.

I then display the adjusted r-squared values from each of the 12 models.

Figure 14- Comparing the predictive performance of each rolling sample regression model



Interestingly, each model sees an increase in its adjusted r-squared value as the time periods become more recent. Each rolling sample sees an increase, aside from the last time period considered, evidencing a gradual increase in the predictiveness of the statistics over time.

My regression results clearly support my hypothesis, and prove that the findings from the correlation analysis are not as relevant, as free agents' past performance statistics become more predictive of their contract amount over time according to these models.

Machine Learning Analysis

As mentioned in my methods section, I decided to utilize the Random Forest Regressor for more advanced predictions to me. Compared to other models, this method resulted in the lowest mean squared error and most consistent predictive power. The resulting mean R-squared was 0.3693, which is reasonable given the nature of the data. Clear correlations between variables in the data are observed, but accurate predictions aren’t guaranteed. The past performance statistics considered helped to explain the variation in players contract amounts, but not all of them. There are many other factors that are considered significant that are outside the scope of my research, which may cause omitted variable bias. I generated variable importance plots to show which statistics were the most significant in my model for three separate time periods- pre-Moneyball (1999-2004), post-Moneyball (2005-2010), and over the full time period (1999-2019)

Figure 14- Feature Importance plot for the pre-Moneyball time period

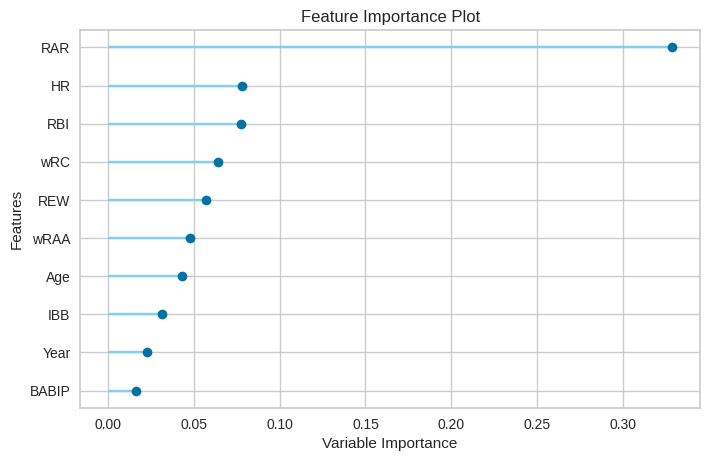


Figure 15- Feature Importance plot for the post-Moneyball time period

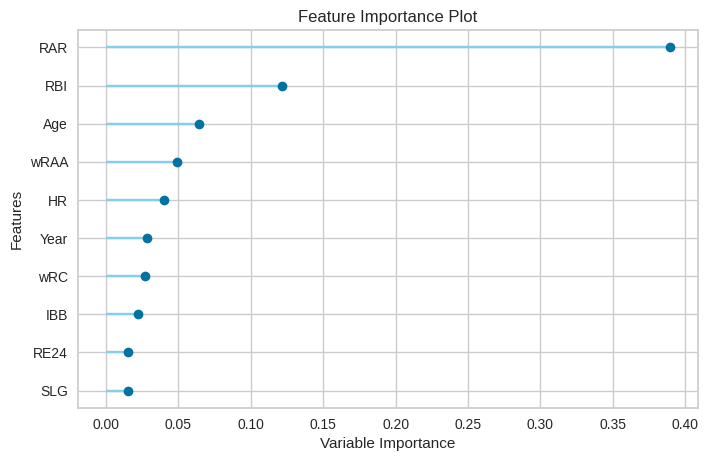
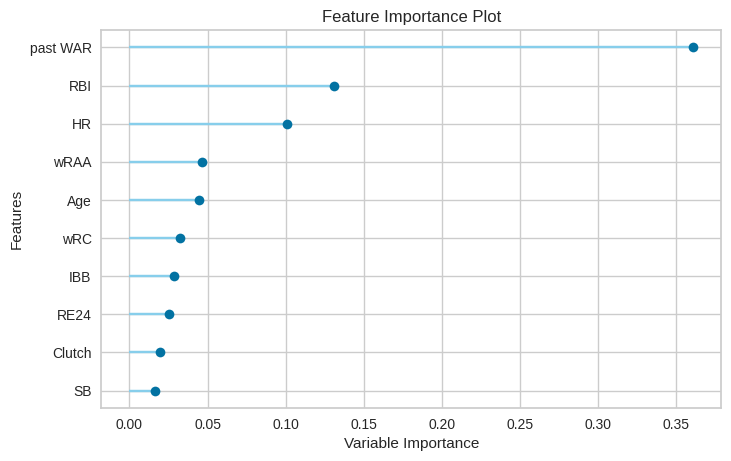


Figure 16- Feature Importance plot for the full time period



These graphs reveal that the most important variable in the model predicting free agents contract amount is their past RAR- runs above replacement. This is true for both the pre-and-post Moneyball time periods. However, in the post-Moneyball time period, the variable importance score of this variable is higher- ~0.35 vs. ~0.38. This could reveal that this variable has become more significant in a free agent’s contract amount after the release of Moneyball. The second most important variable in the pre-Moneyball model was home runs at ~0.07, while in the pre-Moneyball model it’s the 5th most important at ~0.03. This could reveal that over time, home runs have been less of a determinant of a free agent’s contract. It’s important to note the significant difference between the first and second most important variables in the models, as Runs Above Replacement is by far the most important. Interestingly, the second most important variable in the post-Moneyball model is RBI, at 0.12 compared to 0.07. The Age variable is the third most important in the post-Moneyball model, which could reveal that age has become a more significant determinant of free agents contract amounts over time. Aside from this, there aren’t any significant differences in the importance between the traditional and advanced statistics. When looking at the full data, it’s unsurprising the free agents past WAR is the most important variable. These results are interesting and conform to my hypothesis of free agents advanced performance statistics being stronger determinants of their contract amounts. However, there are no drastic differences as seen through the feature importance plots, aside from RAR being more important and home runs being less important post-Moneyball.

Discussion

There are many similarities between the results from my analysis and those in existing literature. The discovery of Barnes, S.L. and Bjarnadóttir, M.V. (2016) that hitters with the highest WAR, runs created, and win probability were paid the highest align with my findings. It’s clear that free agents' more advanced statistics have played a larger role in their valuation on the free agency market. This is seen through the changes in the coefficients and statistical significance of the advanced statistics in my regression models over time, and feature importance plots from my machine learning models. Additionally, Barnes, S.L. and Bjarnadóttir, M.V. (2016)’s conclusion that more traditional player performance statistics such as batting average are being considered less and less over time is validated through my results, as free agents past batting average has proven to be less correlated with and less of a determinant of their contract amount than their more advanced statistics. My exploratory analysis of the performance of free agents after signing contracts revealed that players between 29-32 years-old proved to be more valuable than older players, which aligns with existing literature such as the findings of both Barnes, S.L. and Bjarnadóttir, M.V. (2016) and Solow, J. L., & Krautmann, A. C., & Oppenheimer, M. (2020), that players' productivity decreases as they become older, and older players signing longer-term contracts turn out to be worse signings than younger players. It’s clear that the free agents who proved to be the best value and exceeded expectations the most were those signed to shorter contracts, validating these authors' conclusions. Additionally, the findings of Turner, Chad & Hakes, Jahn, 2007 that the youngest free agents are paid less than oldest ones are proven, through my exploratory analysis and random forest model revealing that age is one of the most important variables in predicting free agents' pay in the post-Moneyball time period. My comparisons of the relationships between the most noteworthy free agents past on-base percentage and pay over time align with the findings of both Brown, D. T., Link, C. R., & Rubin, S. L. (2017) and Hakes, Jahn, K., and Raymond D. Sauer (2006), that players with high on-base percentage have been more valued as free agents and have benefited the most from the publication of Moneyball. Examples of the relationship between free agents past on-base percentage and pay before and after Moneyball reveal that players with higher on-base percentages were rewarded more post-Moneyball. Free agent players such as J.D. Drew with high on-base percentages who in turn received larger contracts, despite having less-impressive batting average and home-run numbers, seem to have been the most affected by Moneyball. This is also reflected in my rolling sample regression results as over time, free agents past statistics have been stronger predictors of their future performance. Players have been paid more according to their statistics since the release of Moneyball. All in all, my results align with findings in existing literature and support my hypothesis.

Conclusion

As seen through my results, free agents past performance statistics, especially their advanced ones, have become better predictors of their pay over time, demonstrating the Moneyball effect that has been covered in existing literature and providing my hypothesis to be correct. It’s clear from this research that Moneyball and the analytics movement in baseball have changed the way free agent hitters are valued, and have significantly impacted which offensive statistics are the most valued by teams and players. My exploratory data analysis paired with regression and machine learning analysis may reveal that more teams have adopted the Moneyball philosophy. Free agents past advanced performance statistics are more correlated with their pay and future performance, while more traditional performance statistics are less correlated. However, there is no clear trend over time, even though it’s clear that some of the performance variables were more highly correlated with pay for some years than others. I originally hypothesized that in the earlier years in the data, traditional statistics like AVG, HR, RBI were more correlated with the contract amount than in later years. I also hypothesized that over time, more advanced statistics like OBP, wRC, RAR have become more correlated to contract amounts than other statistics. This could be due to the belief of teams emphasizing using more advanced statistics to measure player performance than traditional statistics in valuing free agents as time has gone on. The results from my correlation analysis surprisingly do not conform to this original hypothesis.

However, my regression results did conform to my hypothesis, as free agents' past performance statistics have become better predictors of their contract amount over time. The inconsistencies in the correlation results, and the larger sample size in the regression models analyzed de-value the findings from the correlation analysis. My rolling sample regressions reveal that free agents' past performance statistics have become better predictors of their pay over time. The results show that free agents past statistics, especially advanced statistics, have become stronger predictors of their pay over time, as the models considering twelve 7-year time intervals from 1999-2006 to 2010-2017 with key performance statistics as the independent variables and contract amount as the dependent variable have performed significantly better over time according to their adjusted r-squared values.

Additionally, machine learning models show that free agents' more advanced statistics have proven to be the most important in determining their pay. By using Pycaret I saw the perspectives of different machine learning models, in this case a random forest, and the variables that were identified as the most important in the model’s prediction of free agents contract amount across all years were RAR and RBIs, with advanced statistics proving to be more important than traditional ones in the post-Moneyball time period.

I believe that both professional baseball teams and impending free agents could benefit from this analysis. By better-understanding how the determinants of contracts have changed over time, and which past performance statistics have been the strangers determinants of players pay, teams will be better-informed in making free agency decisions going forward. The contract amount of future free agents hitters can be predicted using the regression models developed, which become more accurate over time, which could help teams to value them properly. MLB players approaching free agency can also learn from this analysis the importance of performing well in terms of advanced statistics, in order to receive the largest possible contract. By realizing the important role that these statistics play in how they are valued, they can adjust their hitting strategies to focus more on excelling in those areas. Additionally, they can be incentivized to work as hard as they can to perform well in the two seasons prior to their free agency. This work can also allow professional baseball fans to have a better understanding of the recent history of the dynamics of the free agent market for hitters. They would also be able to see why certain free agents may be paid more than others, and how their past statistics play the largest role in their pay, no matter how well-known they are. Lastly, teams' strategies of valuing a player’s advanced statistics over their traditional “baseball card” statistics can be validated.

Limitations and Future Work

There are limitations to this analysis worth mentioning. I do not consider free agent pitchers, and it could be possible that there are clearer trends in the correlations of pitching performance variables over time than with hitters. Future work could compare the changes over time in the determinants of pay between free-agent hitters and pitchers, and share significant differences. Another limitation is that my analysis did not include more advanced hitting statistics that have been recently introduced. There are more advanced offensive statistics such as barrel % (the percentage of times a hitter makes contact with the baseball on the barrel of their bat), hard hit % (the percentage of times a hitter has a hard hit), and exit velocity (how hard the player is hitting the ball). These statistics were not tracked prior to 2015, so I decided to not include them in my analysis due to free agents prior to 2015 not having values for these variables. Doing so would have resulted in an incredibly small sample. I hypothesize that these statcast statistics have become more correlated with a player’s pay over time, as they can reveal more about a player’s offensive talent. Future research could look at the correlations between these statcast statistics, and contract amount and future performance over time from the time they were first introduced to the present day. As more of this data becomes available, this analysis will be very interesting and feasible. Additionally, my data did not separate players with larger, multi-year contracts from players with smaller, single-year contracts. There could very well be differences in the results for players signing long-term vs. short-term contracts. Future work could focus on one or the other, or compare the results between players signed to long vs. short-term contracts. Another limitation of my work is not factoring in players’ injuries. Future work could identify players who suffered significant injuries before or after signing their contract and filter them out. Players with injuries will have lower performance statistics due to playing less games while they recover, and keeping them in the data could be skewing the results. Another limitation is the sample size of my data- 884. This could be considered too small of a sample to get valid correlation and regression results over time, with there being only approximately 40 free agents per year. There could also be additional work to predict a player's future performance more effectively. The results of future WAR were not explained by the performance statistics in my machine learning models. Future work could investigate performance statistics in the short term instead of the full horizon (in other words, instead of looking at past/present performance metrics and future WAR, a study could use performance statistics of year X and WAR of year X+1). This may show a stronger relationship between the variables, but would not mean that the identified correlation would be necessarily bigger. This would require much more extensive work, as the analysis would have to be replicated for many years to see if any identified correlation in one of the years stood strong in the other years as well. I believe that future performance could be predicted more effectively with using more variables and in the short-term, and any conclusions that stand for the long-term could be considered as noise. Lastly, I did not consider qualitative variables in my final analysis such as the winning percentage of the free agents former team, as Ryan P. Terry, Jeffrey E. McGee, Malcolm J. Kass (2018) did. I believed that including these variables was outside the scope of my research question, and data containing this information was not easily accessible. Future research could include more qualitative factors to paint a bigger picture of the dynamics of major league baseball free agency.

References

Barnes, S.L. and Bjarnadóttir, M.V. (2016), Great expectations: An analysis of major league baseball free agent performance. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 9:* 295-309. https://doi.org/10.1002/sam.11311

Bradbury, J. C. (2007). Does the Baseball Labor Market Properly Value Pitchers? *Journal of Sports Economics, 8(6),* 616–632. https://doi.org/10.1177/1527002506296366

Brown, D. T., Link, C. R., & Rubin, S. L. (2017). Moneyball After 10 Years: How Have Major League Baseball Salaries Adjusted? *Journal of Sports Economics, 18(8),* 771–786. <https://doi.org/10.1177/1527002515609665>

Brownlee, J. (2020, June 30). Why one-hot encode data in machine learning? Retrieved March 13, 2021, from <https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/>

Brownlee, J. (2020, August 14). A gentle introduction to the gradient boosting algorithm for machine learning. Retrieved May 09, 2021, from https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/

Brownlee, J. (2020, November 15). A Gentle Introduction to PyCaret for Machine Learning. Machine Learning Mastery. https://machinelearningmastery.com/pycaret-for-machine-learning/.

Cot's baseball CONTRACTS. (n.d.). Retrieved March 13, 2021, from https://legacy.baseballprospectus.com/compensation/cots/

Ehrlich, J.A. and Potter, J.M. (2020), "Is offense worth more than defense and pitching? Marginal revenue product and revenue sharing in major league baseball", *Managerial Finance*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/MF-04-2020-0210

Ehrlich, J., Perline, J., Potter, J. and Sanders, S. (2020), “Does a salary premium exist for offensive output in Major League Baseball?”, *Managerial Finance*, Vol. 47 No. 3, pp. 326-335. https://doi.org/10.1108/MF-04-2020-0186

FanGraphs baseball: Baseball statistics and analysis. (n.d.). Retrieved March 13, 2021, from https://www.fangraphs.com/

Hakes, Jahn, K., and Raymond D. Sauer. 2006. "An Economic Evaluation of the *Moneyball* Hypothesis." *Journal of Economic Perspectives*, 20 (3): 173-186.

Turner, Chad & Hakes, Jahn, 2007. "Pay, productivity and aging in Major League Baseball," *MPRA Paper 4326, University Library of Munich, Germany*.

Krautmann, A. C., & Oppenheimer, M. (2002). Contract Length and the Return to Performance in Major League Baseball. *Journal of Sports Economics*, *3*(1), 6–17. https://doi.org/10.1177/1527002502003001002

Real Python. (2021, February 13). *Split Your Dataset With scikit-learn's train\_test\_split()*. Real Python. https://realpython.com/train-test-split-python-data/.

Ryan P. Terry, Jeffrey E. McGee & Malcolm J. Kass (2018) The not-so-free agent: Non-performance factors that contribute to free agent compensation premiums, *Sport Management Review, 21:2,* 189-201, DOI: 10.1016/j.smr.2017.06.006

Solow, J. L., & Krautmann, A. C. (2020). Do You Get What You Pay for? Salary and Ex Ante Player Value in Major League Baseball. *Journal of Sports Economics*, *21*(7), 705–722. https://doi.org/10.1177/1527002520930259

Variance inflation factor. (2020, December 16). Retrieved May 06, 2021, from https://www.statisticshowto.com/variance-inflation-factor/

Appendices

Link to the Github Repository containing the code and data behind this research paper: https://github.com/JAllenPrimack/PrimackDA401

Tables of Regression Results over 7-year intervals

|  |  |  |  |
| --- | --- | --- | --- |
| Time Period | 1999-2006 | 2000-2007 | 2001-2008 |
| Constant | -1.136e+09 | -1.113e+09 | -1.892e+09 |
|  | (4.979e+08)\*\* | (5.515e+08)\*\* | (6.080e+08)\*\*\* |
| Year | 549626.626 | 540424.256 | 927570.577 |
| ( | 247753.765)\*\* | (274204.688)\*\* | (302239.812)\*\*\* |
| Years | 12545989.234 | 13602948.510 | 15323402.022 |
|  | (560536.320)\*\*\* | (549609.456)\*\*\* | (599785.344)\*\*\* |
| HR | 284181.963 | 337855.174 | 569760.952 |
|  | (88113.066)\*\*\* | (92818.369)\*\*\* | (105409.825)\*\*\* |
| IBB | -72780.790 | -44880.781 | -150531.263 |
|  | (114126.823) | (111719.790) | (124657.668) |
| OBP | 89097823.538 | 76661726.857 | 1.321e+08 |
|  | (34584801.404)\*\* | (36939825.906)\*\* | (41311277.536)\*\*\* |
| SLG | -2.239e+07 | -2.932e+07 | -7.399e+07 |
|  | (19904711.718) | (21161511.846) | (23795561.857)\*\*\* |
| wRC | -69607.790 | -71660.473 | -109083.850 |
|  | (26811.467)\*\*\* | (27735.686)\*\* | (30177.678)\*\*\* |
| RAR | 83144.778 | 60319.302 | 62150.787 |
|  | (36717.489)\*\* | (37685.891) | (41899.450) |
| WPA | 335468.922 | 425960.772 | 292707.830 |
|  | (375907.531) | (394585.109) | (433271.593) |
| N | 336 | 335 | 357 |
| Adj R-Squared | 0.799 | 0.812 | 0.816 |
| SEE | 10578098.799 | 10650647.440 | 12118004.666 |
| F-ratio | 149.148 | 160.937 | 175.986 |
| SSR | 3.648e+16 | 3.687e+16 | 5.096e+16 |

|  |  |  |  |
| --- | --- | --- | --- |
| Time Period | 2002-2009 | 2003-2010 | 2004-2011 |
| Constant | -2.445e+09 | -2.084e+09 | -1.820e+09 |
|  | (5.327e+08)\*\*\* | (5.050e+08)\*\*\* | (4.656e+08)\*\*\* |
| Year | 1208225.435 | 1028380.172 | 894573.608 |
|  | (265398.132)\*\*\* | (251526.261)\*\*\* | (231715.195)\*\*\* |
| Years | 14579902.591 | 15184320.826 | 16057330.766 |
|  | (555351.333)\*\*\* | (542278.184)\*\*\* | (532622.864)\*\*\* |
| HR | 381745.460 | 324650.304 | 306043.497 |
|  | (91748.797)\*\*\* | (88106.237)\*\*\* | (87030.853)\*\*\* |
| IBB | 66723.489 | 91654.163 | 144684.464 |
|  | (106871.457) | (105855.954) | (107097.060) |
| OBP | 74197865.086 | 58862515.810 | 69927737.897 |
|  | (36228859.816)\*\* | (34445941.210)\* | (34648482.755)\*\* |
| SLG | -5.325e+07 | -4.243e+07 | -4.688e+07 |
|  | (20538255.335)\*\*\* | (19804844.466)\*\* | (20022073.191)\*\* |
| wRC | -66236.930 | -53228.056 | -58393.093 |
|  | (26321.219)\*\* | (25865.823)\*\* | (25733.824)\*\* |
| RAR | 51238.064 | 25375.934 | 58762.155 |
|  | (35553.931) | (34539.491) | (33785.878)\* |
| WPA | 559693.569 | 640698.071 | 497984.215 |
|  | (377975.287) | (363552.293)\* | (350883.290) |
| N | 366 | 373 | 385 |
| Adj R-Squared | 0.821 | 0.822 | 0.845 |
| SEE | 10728011.312 | 10515908.701 | 10343380.232 |
| F-ratio | 186.749 | 192.319 | 234.413 |
| SSR | 4.097e+16 | 4.014e+16 | 4.012e+16 |

|  |  |  |  |
| --- | --- | --- | --- |
| Time Period | 2005-2012 | 2006-2013 | 2007-2014 |
| Constant | -1.127e+09 | -1.561e+08 | 82100077.714 |
|  | (4.525e+08)\*\* | (4.534e+08) | (4.411e+08) |
| Year | 552723.914 | 73429.974 | -42223.037 |
|  | (224568.050)\*\* | (224598.151) | (218233.049) |
| Years | 16161840.209 | 16384495.620 | 16584287.488 |
|  | (523215.602)\*\*\* | (536502.187)\*\*\* | (524941.557)\*\*\* |
| HR | 250811.717 | 294928.618 | 341546.398 |
|  | (86395.365)\*\*\* | (79987.563)\*\*\* | (76659.507)\*\*\* |
| IBB | 252359.869 | 453967.326 | 462184.884 |
|  | (112171.405)\*\* | (111323.938)\*\*\* | (109659.776)\*\*\* |
| OBP | 35364005.536 | 12794846.121 | 5413838.804 |
|  | (34126130.941) | (32282624.604) | (31467234.163) |
| SLG | -3.733e+07 | -4.296e+07 | -5.213e+07 |
|  | (19295706.651)\* | (18070035.858)\*\* | (16945128.488)\*\*\* |
| wRC | -46119.492 | -64734.803 | -70197.615 |
|  | (25590.712)\* | (24358.759)\*\*\* | (23652.125)\*\*\* |
| RAR | 79930.185 | 130859.505 | 120681.171 |
|  | (32952.791)\*\* | (33844.323)\*\*\* | (32834.235)\*\*\* |
| WPA | 710050.035 | 669217.587 | 962272.028 |
|  | (351380.991)\*\* | (344897.528)\* | (339703.978)\*\*\* |
| N | 384 | 378 | 383 |
| Adj R-Squared | 0.853 | 0.861 | 0.870 |
| SEE | 10119829.058 | 9852003.101 | 9549046.057 |
| F-ratio | 248.364 | 259.474 | 284.798 |
| SSR | 3.830e+16 | 3.572e+16 | 3.401e+16 |

|  |  |  |  |
| --- | --- | --- | --- |
| Time Period | 2008-2015 | 2009-2016 | 2010-2017 |
| Constant | 39345904.598 | -1.028e+09 | -1.323e+09 |
|  | (4.638e+08) | (3.965e+08)\*\*\* | (3.712e+08)\*\*\* |
| Year | -25067.006 | 510196.565 | 655639.548 |
|  | (229486.807) | (195993.564)\*\*\* | (183757.342)\*\*\* |
| Years | 17567706.483 | 15623135.293 | 15020016.114 |
|  | (561203.865)\*\*\* | (487383.724)\*\*\* | (478080.901)\*\*\* |
| HR | 287551.604 | 109401.946 | 121207.015 |
|  | (77486.773)\*\*\* | (63572.848)\* | (59237.945)\*\* |
| IBB | 353559.118 | 394860.831 | 327717.095 |
|  | (116209.616)\*\*\* | (97510.765)\*\*\* | (100348.763)\*\*\* |
| OBP | 7868592.581 | -4.841e+07 | -4.003e+07 |
|  | (32378217.896) | (26532626.346)\* | (25256571.448) |
| SLG | -3.305e+07 | -5883210.397 | -7099268.153 |
|  | (16592411.107)\*\* | (13255868.989) | (12709595.796) |
| wRC | -60559.884 | -21497.805 | -11974.532 |
|  | (24118.306)\*\* | (20385.325) | (20132.881) |
| RAR | 108064.332 | 135841.886 | 131126.994 |
|  | (32667.563)\*\*\* | (26536.655)\*\*\* | (25769.212)\*\*\* |
| WPA | 745188.144 | 948185.969 | 933291.246 |
|  | (345249.463)\*\* | (283743.881)\*\*\* | (274111.227)\*\*\* |
| N | 366 | 361 | 354 |
| Adj R-Squared | 0.875 | 0.884 | 0.882 |
| SEE | 9458738.626 | 7793406.919 | 7342077.077 |
| F-ratio | 284.120 | 307.184 | 293.093 |
| SSR | 3.185e+16 | 2.132e+16 | 1.854e+16 |