2016 Chicago Cubs: In-Depth Season Analysis

Visuals Created in Tableau

OMIS 473 Final Report

Report & Visuals created by:
John Acton, Marcus Diaz, Sean Hastings, Daniel Wilson

Table of Contents

Purpose

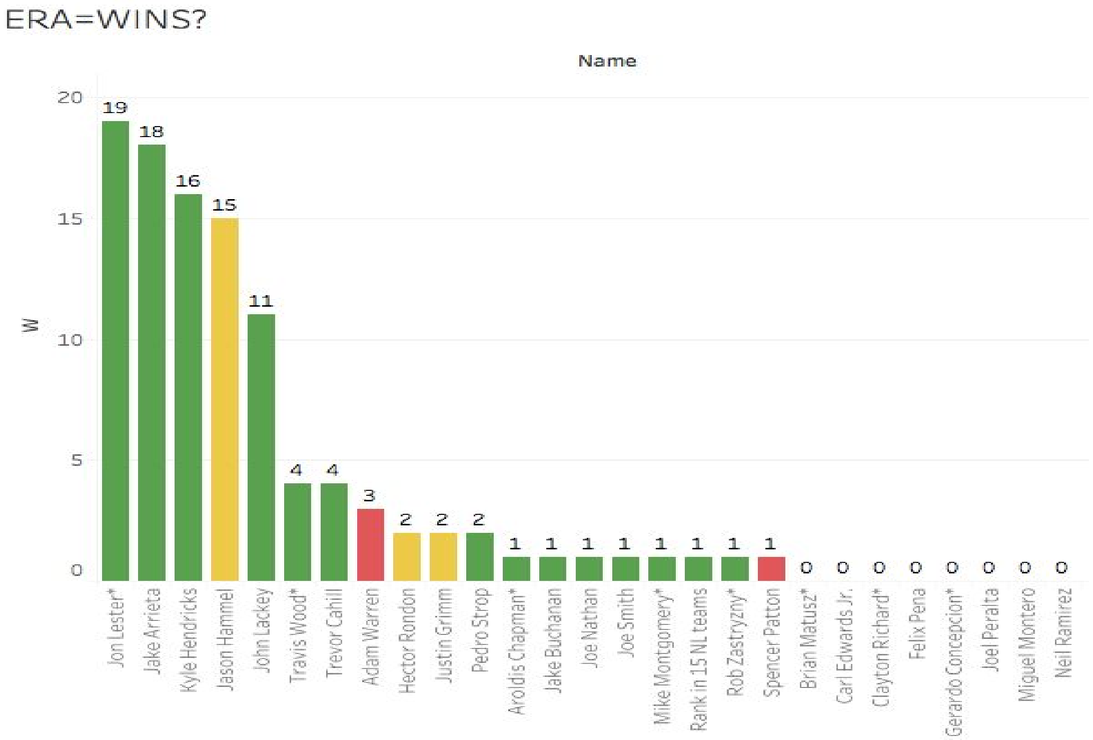
During their 2016 MLB season, the Chicago Cubs won the World Series for the first time in over 100 years. The purpose of this report is to provide a deeper look into the numbers and statistics behind the Chicago Cubs 2016 MLB season to discover what made them so dominant, as well as what they can do in the future to remain that way. To aid in accomplishing this, our team used Tableau to analyze our selected data set and create meaningful visualizations. These visualizations gave us the means to quantify the value of variables within the data set, such as night or day games, attendance number, and fielding percentage. In doing so, we were able to gauge the individual degrees of impact these variables had on determining the outcome of the 2016 season. Not only does the information gleaned from this data set tell the story of the 2016 Chicago Cubs, but it also provides measurable organizational value to decision makers on the team itself, as well as other teams competing against the Cubs in the MLB. In utilizing the visualizations within this report, decisions makers and individuals of vested interest can see vulnerable areas within the team's lineup and pursue players to help fill the weak spots, or strategize on how to capitalize on them.

Data Set Description

The game of baseball can be very complex and may be analyzed using countless different variables and methodologies. To create a useful analysis that describes how the Chicago Cubs stood out as MLB World Champions in 2016, our team came to the conclusion that we would need to find a data set that provides a certain requisite amount of differing variables and statistics for their 2016 season. Ultimately, we found a data set titled “2016 Chicago Cubs Schedule,” which contains hundreds of different standard-practice baseball statistics from the Cubs’ 2016 season. These statistics are used within the data to gauge performance across three separate categories, including team pitching, batting, and fielding. Additionally, the data set includes two other highly useful components. One component is a full team roster listing each player on the team along with their respective season statistics, and the other is the Cubs’ full 2016 game schedule that includes the results of each game played. To our satisfaction, the data set incorporates categorical and numeric data types, has date/time components, geography components, and has both average and sum aggregations present in the data. The source from which we obtained this data set was baseball-reference.com, which is a well-established, accurate data hub website for baseball statistics.

Visuals: Descriptions & Justifications

Visual 1

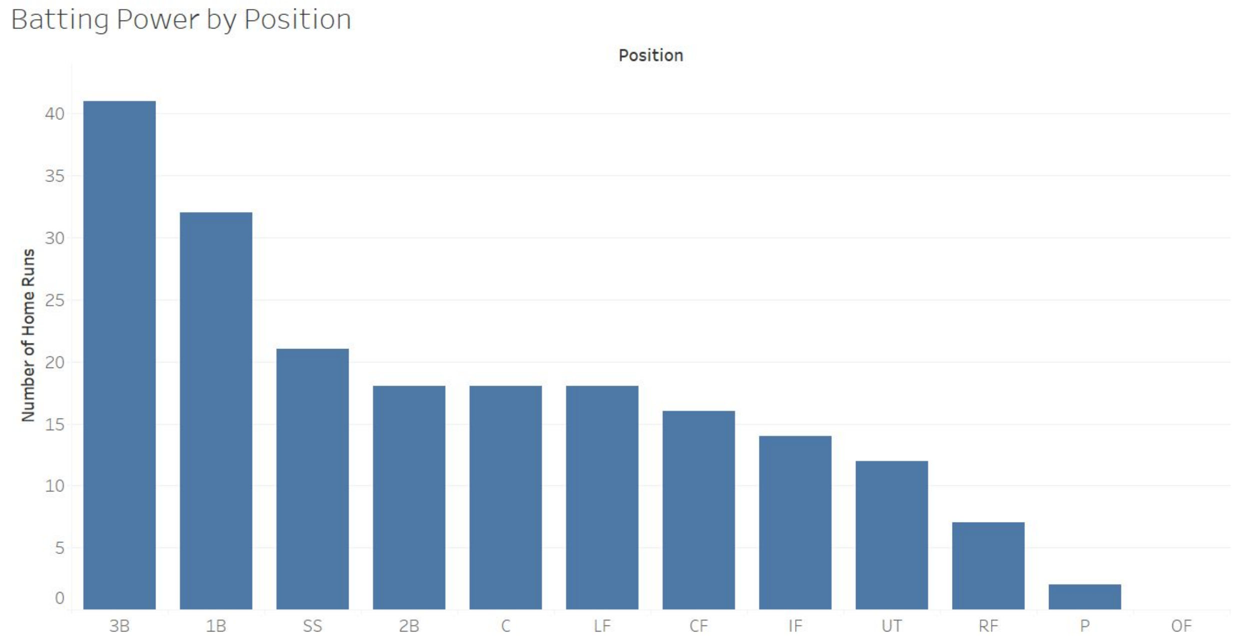


(V-1) Our first visualization shows how each pitcher's ERA equates to their total wins for the season. ERA, or Earned Run Average, is a statistic that provides the average number of runs that a pitcher gives up per nine innings of baseball played. ERA tracks a pitcher's performance over nine innings because that is the number of innings that make up one complete game. So for instance, if a pitcher were to pitch nine innings and gives up two runs, his ERA would come out to 9/2, or 4.5. As can be logically deduced, the lower the ERA the pitcher has, the better he has performed throughout the season. This visual was built using a column chart displaying a list of Cubs pitchers along the x-axis, while the y-axis represents the number of games won. The ERA data for each pitcher has been split into three categories that make up number ranges based on each pitcher's performance. This was accomplished by creating a calculated field to break up

ERA ratings into ranges. If a pitcher's ERA is below 3.5, that is considered to be above average and is shown in green. If a pitchers ERA is between 3.5 and 4.5, that is considered to be average and is shown in yellow. If a pitchers ERA is above 4.5, it is considered to be below average and is shown in red.

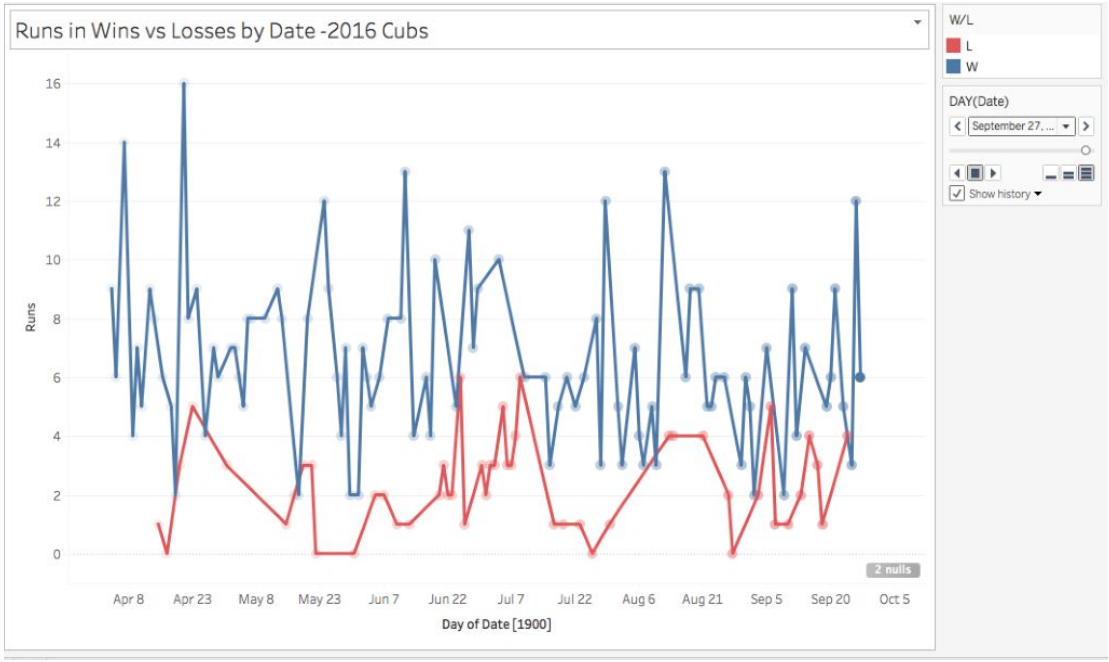
Glancing over the visual, the first thing that can be observed is that the vast majority of pitchers in the Cub's rotation have above average, or green ERAs. Only five pitchers in the rotation have either average or below average ERAs, which designates good performance across the board in terms of pitching. In addition to this, we found that the Cubs' top five pitchers with the most wins all have either a good or average ERA. Looking deeper into this, the top 5 pitchers with the highest number of wins are a combination made up of both starting pitchers, (pitchers that begin each game), and relief pitchers, (pitchers who come in to finish off the game for starting pitchers). This occurrence was unexpected to say the least. More often than not in baseball, as a team gets lower in its rotation (pitching moves from starters to relief), the ERA goes up. This was not exactly the case for the 2016 Cubs, and may be part of the underlying reason as to why they were so successful.

Visual 2



(V-2) Visual two was constructed in the form of a column chart to display the number of home runs that were hit from each position on the team. The x-axis designates the player positions, while the y-axis represents the number of home runs that were hit. It is important to note that this visual draws no type of comparison to other teams' home run performances in the league, and exists solely as an internal means of tracking home run performance only for the Cubs. This said, the chart serves one internal purpose that is essential to the team. It serves as a useful means for the Cubs to track which positions on the team are lacking in the homerun category. This will allow the cubs to pursue certain players in those positions who are better batters, which in turn, will boost the overall number of team home runs. As a general rule of thumb in the game of baseball, when a team is composed of good power hitters who hit a lot of home runs, it puts fear into the opposing pitchers. This fear from pitchers tends to lead a team to more opportunities to hit homeruns, get more walks, and score more runs overall to provide the best chances of winning. In essence, the purpose of this visual is to provide future direction for the Cubs in terms of decision making, rather than looking back as to why they performed the way they did in 2016.

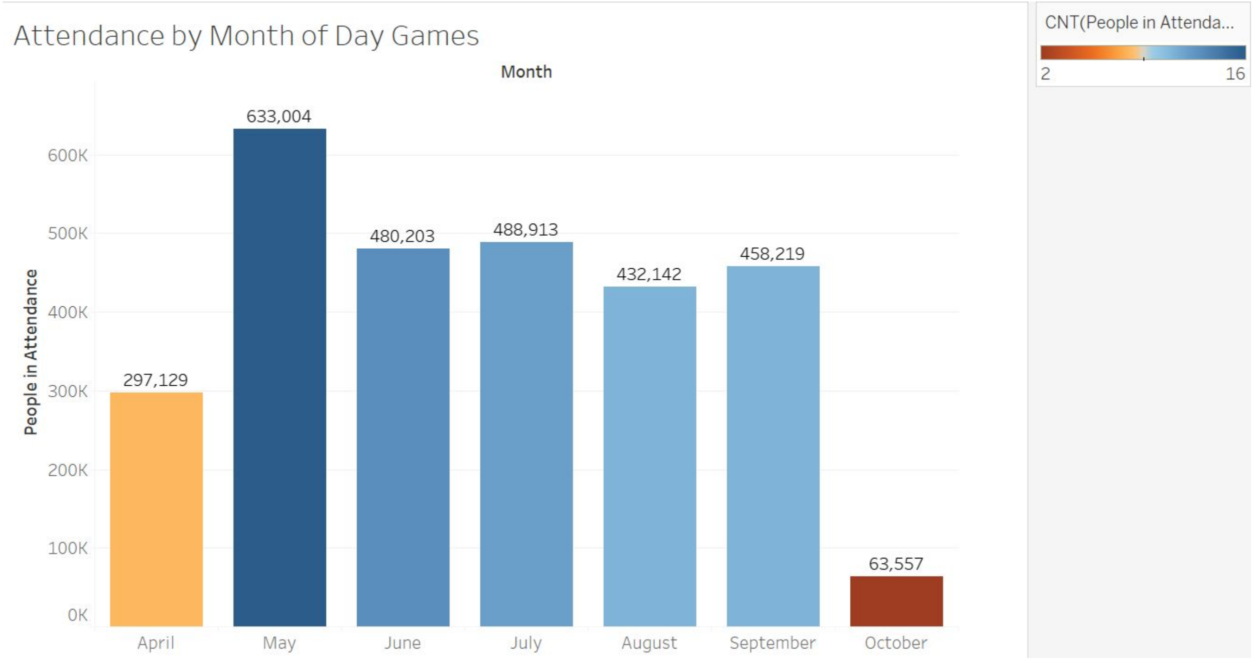
Visual 3



(V-3) Our third visualization shows the relationship between the number of runs scored by the Cubs with respect to their number of wins and losses recorded over time during the 2016 season. It is constructed in the form of a line chart, with game date making up the x-axis. The x-axis was then formatted to show the date dimension in days so that fluctuations in the data would be more clearly visible. The y-axis, in turn, represents the number of runs scored in each game. The blue line represents the number of wins, while the red represents the total number of losses. Evidently, the season high for runs scored during the year by the Cubs was 16, and took place on April 21, 2016.

Delving deeper into the visual, the average number of runs scored for each winning game came out to be 4.98, while the average number of runs scored during games in which the team lost lost was 3.84. We consider these findings to make logical sense due to the fact that as the number of runs scored increases, the likelihood of winning the game increases in turn. The information drawn from this visual could potentially be used by team management and other decision makers as a benchmark for how many runs the Cubs need to score on average per game to win. This would help them construct their batting lineup accordingly. For instance, if the team were to undergo a series of games in which they won, but only scored 1-3 runs per game, it would provide some insight toward what they were doing right in terms of fielding and pitching.

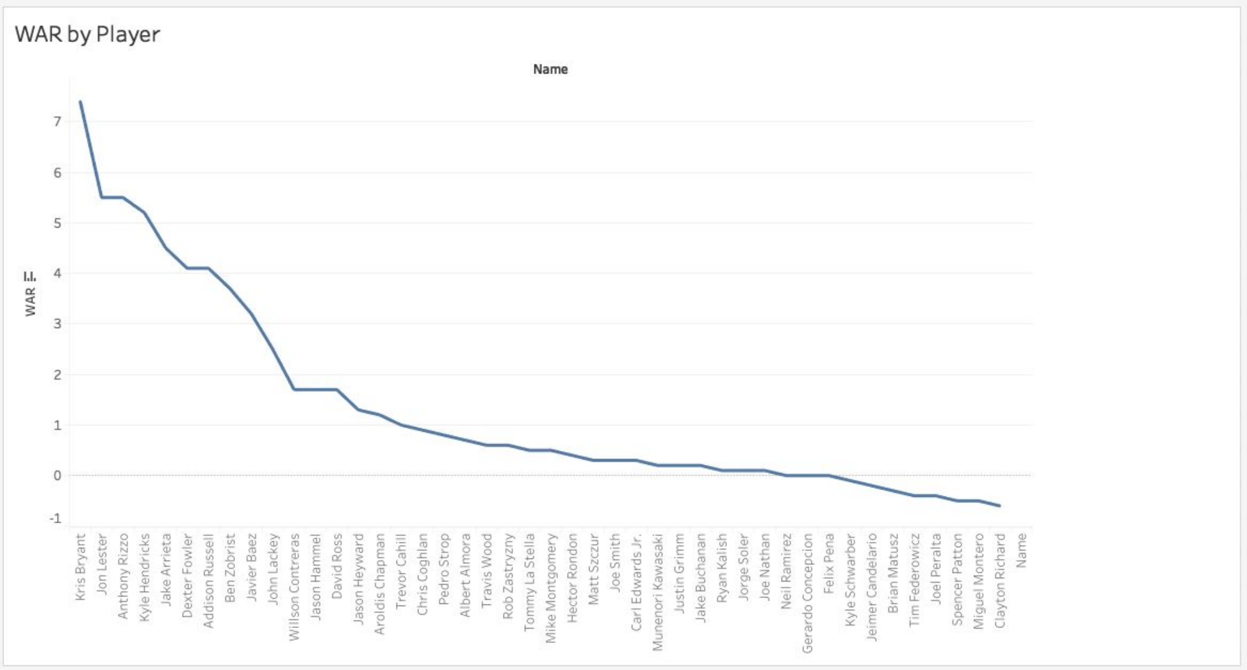
Visual 4



(V-4) Visualization four provides a glimpse at Cubs fans total monthly attendance of day games over the regular season. Please note that only two games were played in October 2016, which is the reason why attendance appears to be much lower in this month compared to the others. The core purpose behind creating this column chart was to provide ourselves with a means of gauging the degree of audience interest and participation in games as the season progressed. This column chart visual was created by positioning the number of day games of each month into the x-axis, and the number of fans in attendance into the y-axis. The low rate of fan attendance during the month of April could be attributed to a few different factors: children still being in school, the weather being unseasonably cold, or fans not knowing whether the team will be worth watching or not. When you take these factors into consideration, it is not surprising the total attendance for April is low. Next, during the month of May, one can see that attendance totals almost double from the previous month. We theorize this to have occurred because more kids were out of school for summer break, and the weather began to get nicer. However, this then begs the question of: Why didn't attendance stay at such high levels moving past May?

To answer this, we looked back at the Cubs' record and statistics during the months of April and May, and realized attendance levels coincided with how the Cubs were performing during that time. The Cubs went 17-5 in April, which made it clear to fans that the team was a force to be reckoned with and were well worth following. The initial excitement drove fans to the ballpark in flocks, but then attendance began to level off in June and July as the regular season progressed. During this time, attendance was still above the league average but was nowhere near what it was in May. It is well known and widely accepted that the month of August is considered to be the "doldrums" of baseball. It is a hot month where children are waiting to go back to school, and there is no end in sight for the season. Additionally, it is important to note that the Cubs had a sizeable lead on their division at this point as they started the month 9 games up. The excitement returned in September when it was clear the Cubs were destined for the top seed in the playoffs. Fundamentally, this visualization helped our team paint a broad, conceptual picture of the entire scope of the Cubs winning season.

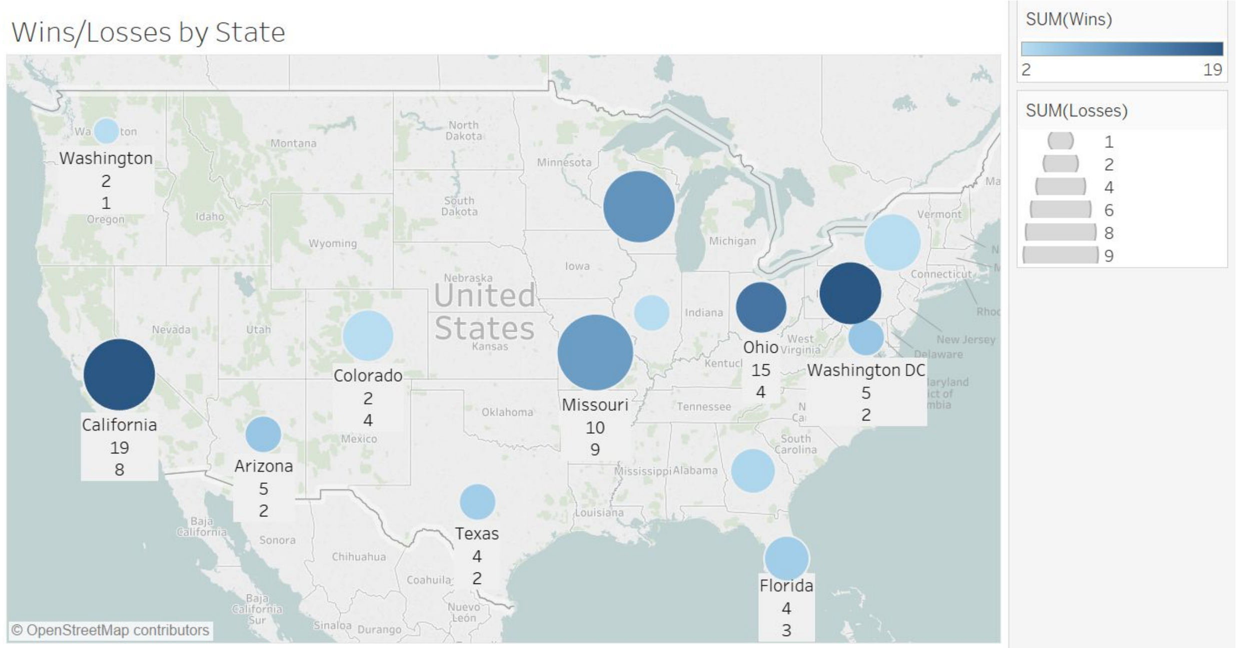
Visual 5



(V-5) Our fifth visual, in the form of a line chart, was created to show each player's total WAR (Wins Above Replacement). WAR is a highly useful statistic in baseball that incorporates the use of six components which summarize a player's total contribution to runs generated for his team. The components that make up a player's WAR include: the player's runs batted in, runs scored from baserunning, runs added or lost due to grounding into double plays, fielding runs, positional adjustment runs, and replacement level runs. A player's WAR is also dependent on what position he plays, with more value being attributed to weaker hitting positions, such as catcher, than positions with stronger hitting on average. When a player is awarded a high WAR value, it often reflects successful performance, a large quantity of playing time, or some combination of both. So in essence, WAR is a statistic that measures each player's individual contribution toward their team's success.

This line chart was created simply by adding the Cubs roster to the x-axis, and adding players' WAR values to the y-axis. We then sorted the WAR values in descending order by player name in order to make the visual understandable and appealing to viewers. Our group elected to use a line chart for this statistic because it allows the viewer a clear picture of the trend of WAR values from player to player. Through dissecting this visual, one would find that the players who contributed the most success toward the standings of the Chicago Cubs tend to have the highest WARs, which makes logical sense. Considering that the vast majority of the team have positive WAR values, it goes to shows just how talented and dominant the 2016 Chicago Cubs were during their winning season.

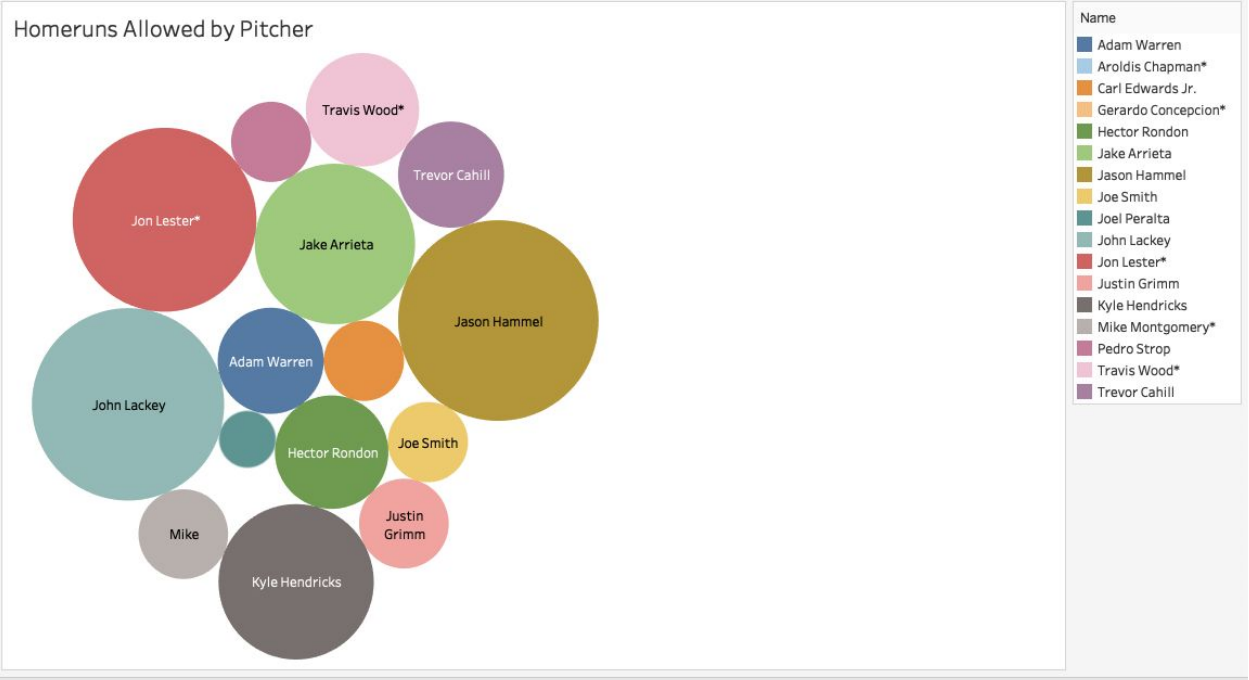
Visual 6



(V-6) Visual six was created in the form of a geo bubble map. It portrays the Cubs total wins and losses by each state in which they played teams from. The map was created by developing a geographic hierarchy for the opposing teams that the Cubs played during the season, which includes their respective states, regions, and cities. The color of each bubble on the map corresponds to the number of wins earned by the Cubs in that state, with darker colors representing more wins and lighter colors representing fewer wins. The physical size of each bubble resembles the number of losses acquired in each state, where larger bubble sizes designate higher incidences of losses, and smaller bubbles designate fewer losses. For additional clarification, the bubbles have been labeled with their corresponding state names, as well as the total number of wins and losses in that state, (wins on top, losses on bottom).

This visual was created for the purpose of determining the states and regions the Cubs excelled in, as well as the states and regions in which the team struggled in. The information garnered from this visual could be utilized for two distinct reasons. In one instance, this map could be used as a reference to compare with other visuals in order to connect what factors were present during a game that lead the team to a win or loss. These factors might include weather conditions, opposing team pitching, whether the team was home or away, team injuries, depth chart changes, and so on. On the other hand, this map also serves as a benchmark for determining how the team will perform in future years. For instance, during their 2016 season the Cubs won 19/25 games in the state of Pennsylvania, but only won 2/7 games in New York. In the following years, the coaching staff might consider allocating more preparation and practice toward competing against the teams in New York rather than teams in Pennsylvania.

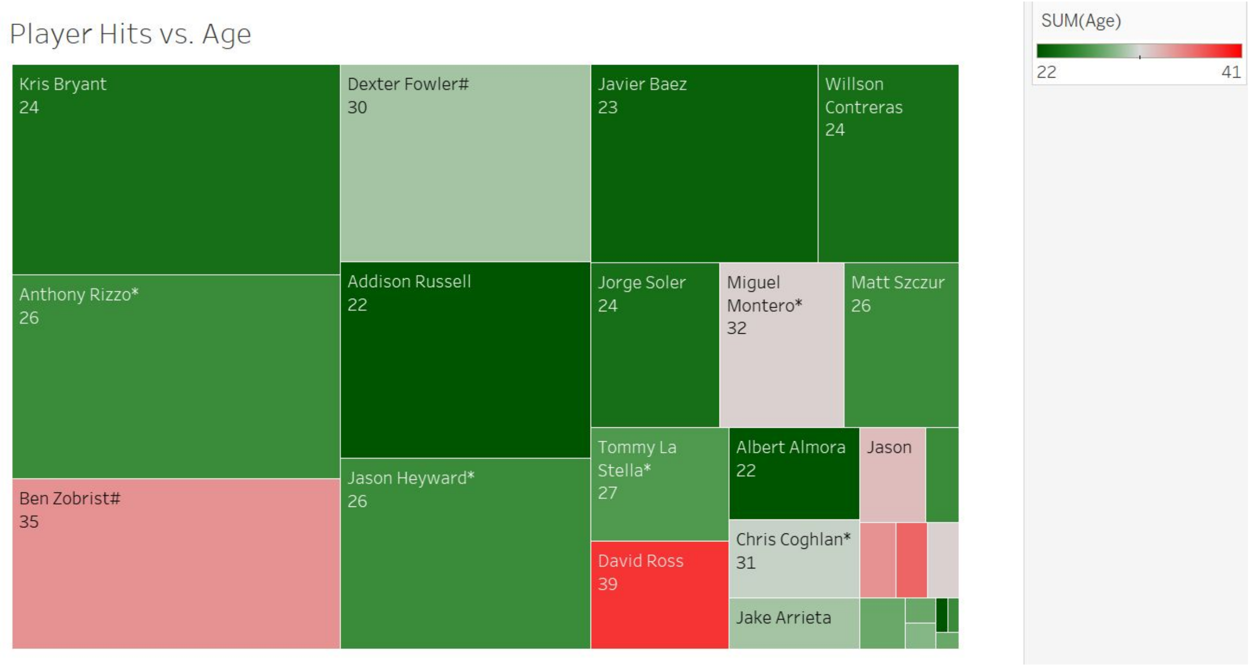
Visual 7



(V-7) Visual seven is a packed bubble chart which displays the individual number of home runs given up by each Cubs pitcher during the regular season. The visual was built by assigning each pitcher on the Cubs rotation a colored bubble that uniquely identifies them among all other pitchers. Each bubble was then labeled with the pitcher's name for clarification purposes. Lastly, we chose to designate the size of each bubble to represent the number of home runs given up by the pitcher. This said, The larger the physical size of the bubble, the greater the amount of home runs given up by that particular pitcher.

We believe this to be a powerful visualization for a couple of key reasons. At face value, the chart shows the past results of the Cubs' 2016 pitching in an easy-to-read format. One can plainly see that the top three pitchers that gave up the most home runs were Jason Hammel, John Lackey, and Jon Lester. On another more important note, this visual also provides an excellent means for future decision makers and coaching staff on the team to understand how to control pitching. Consider a future hypothetical situation in which the opposing team has a power hitter up to bat, and the Cubs pitcher currently on the mound has a tendency to get nervous and give up unnecessary home runs. By using the information obtained from this visual, Cubs decision makers can make calculated adjustments to their pitching staff so that the risk of the power hitter knocking a home run out of the park is minimized. This logic can also be used in reverse. If the pitcher on the mound of our hypothetical situation has been assigned a small bubble, then the risk associated with him pitching to a power hitter is minimized, an no adjustments would be necessary.

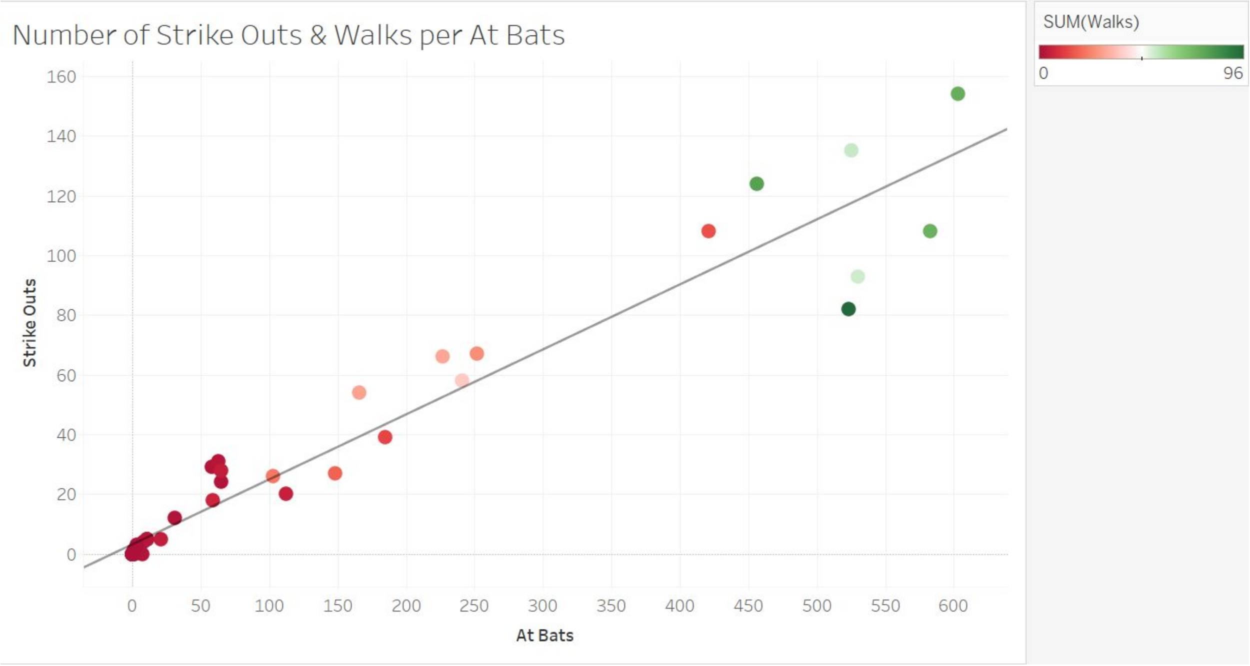
Visual 8



(V-8) For visual number eight, our team decided to construct a heat map to find if a connection exists between a player’s age and his batting abilities.We decided to use a Heat Map for this visual because it was the most efficient way to provide a comprehensive overview of the entire roster’s age and hitting performance. The first step carried out in the creation of this heat map was selecting the top batters on the team in terms of the number of hits they had during the season. The size of each box was made to represent the quantity of hits each batter had, with larger boxes equating to more hits. We then designated that the color of each box represents each batter's corresponding age, where green represents younger players and redder colors represent older players. The boxes were then labeled with each batter's respective age for additional clarification. After just seconds of viewing this chart, one can deduce who the oldest, middle aged, and youngest player on the team were, while also seeing who the best hitters were.

As is evident from the visual, the top 3 hitting players on the 2016 Cubs team were Kris Bryant, Anthony Rizzo, and Ben Zobrist. Followed by them were Dexter Fowler, Addison Russell, and Jason Heyward. Of those top six hitters, two of them have moved out of the green area, meaning they are getting older in age, and are most likely moving past their primes. Ultimately what can be ascertained from this visual is that the Chicago Cubs roster for 2016 was composed predominantly of good hitting, young, talented players along with a few older guys that were still making a difference. We believe this to be a substantial factor as to why they did so well in 2016. The information gleaned from this visual could be utilized by decision makers and coaching staff for discovering areas in the team roster that they need to be aware of when contract talks for recruiting new players begin.

Visual 9



(V-9) Our ninth and final visualization illustrates the relationship between each batter's total number of walks and strikeouts they have had throughout the season with respect to their total at-bats. This type of visual, known as a scatter plot, is an excellent analytical tool for uncovering correlations that exist naturally between certain target variables within a data set. Through exploring our data, we began to consider that there may be some correlation between the number of times a player strikes out, versus the amount of times they are walked, all in relation to the total number of times they have batted during the season. We tested this hypothesis by placing at-bats in the x-axis of the scatter plot, and strikeouts in the y-axis. Each color-coded point within the plot makes up an individual batter on the Cubs roster. The colors of these points range from red to green with red points depicting batters that have a low amount of walks, and green points representing batters that have a greater amount.

As it turns out, our team was correct in our hypothesis of a correlation existing between these variables. The shape of the scatter plot and slope of the trend line denote a positive correlation between a player's number of walks and strikeouts with respect to their total at-bats. This relationship confirms what we had already presumed to be true - the more times a player bats during the season, the more strikeouts and walks he is prone to accumulate. It is interesting to note that the correlation seems to be stronger near the origin, and disperse in accuracy as the number of at-bats increase. Almost every player with 400 or more at-bats tends to have a number of walks greater than 50, which is the range at which the points start to shade to green, while their amount of strikeouts tend to vary more substantially between each batter. We theorize this to be true because as players accumulate more and more at-bats throughout their season, they will continue to gain experience in being able to read pitches and react accordingly to increase the number of walks or hits they wrack up. This visual doubles as being both a useful tool for understanding the fundamental concepts of how batting works within the games of baseball, and as being a guide for Chicago Cubs coaches to look into players who are underperforming in some category and correct the problem.

Conclusion

As aforementioned, the purpose behind creating this report was to delve deeper into the underlying factors behind why the Cubs won the 2016 World series. In doing so, our team came to understand these contributing factors at a more profound level, and we see now why the Cubs had such success in 2016. Coupled with this newfound comprehension of the Cubs' 2016 success, we also now have a means of prescribing some of the steps they can take in the future to reassert that level of dominance. Using Tableau to analyze our selected data set, "2016 Chicago Cubs Schedule", our team was able to construct nine separate visualizations which all play an important role in depicting how the team succeeded, as well as suggesting what they might do in the future to relive that success.

The visuals that provided us with a glimpse at the Cubs success during their 2016 season were visuals one, four, five and eight. Our first visual, which quantified the relationship between a pitcher's number of wins versus his ascribed ERA, designates that a major contributor toward the Cubs' success came from the fact that the team's pitching rotation was predominantly composed of pitchers having above-average ERAs. Visual four came in the form of a column chart and displayed each month's individual contribution toward fan attendance of day games. This visual provided us with the understanding that as team performance improved, the number of fans in attendance of day games increased in turn. Generally speaking, the more fans there are in attendance of a game, the higher morale the team has during that game. This in turn encourages players to perform at a higher level than they would if they felt discouraged. Visual five, a line chart showing each player's individual WAR, proposed that another significant reason for the Cubs' 2016 success was the fact that the vast majority of players on the team had

WAR values above zero, which denotes exceptional performance from players across the board. Finally, visual eight, a heat map describing the association between a player's batting ability in reference to his age, provided us with the understanding that the 2016 cubs batting staff consisted largely of young, talented power hitters that were properly mentored by older players on the team still carrying their torch. Through creating and incorporating these visuals into our report, our team has surmised that the stand-alone reasons the cubs were so successful in 2016 were: its strong pitching staff that performed consistently well throughout the season, the high incidence of players with positive WARs, the team's young and talented batting rotation, and continued support from fans in attendance of games.

The next category of visualizations, which include visuals two and three, were created more for the purpose of developing strategies and methodologies to promote the Chicago Cubs success in future seasons. Visual two incorporated a column chart which displayed the number of home runs hit by each respective position on the team. Our group presumed this visual would be more useful in prescribing future courses of action rather than describing the way things were in 2016, because the visual serves more as a means for the Cubs' coaching staff to track internal home run performance in an effort to focus on players who are underperforming in their batting. Visual 3 was a line chart that displays the relationship between the Cub's total number of runs scored versus their total number of wins and losses recorded. This visual helped our team grasp the understanding that the Cubs have a significantly higher chance of winning a game if they score a minimum of at least four runs during the game. Because this statistic reigns true for the vast majority of all games played during the 2016 season, Cubs players and coaching staff alike would be inclined to shoot for obtaining this benchmark number of runs each game in order to increase the likelihood of winning future games. Evidently, these visuals serve purposes more along the lines of planning for future seasons and developing performance benchmarks to promote success in the future.

Visuals six, seven and nine can be interpreted several ways depending on how one views them. These visuals can be used for developing strategies to promote future success within the MLB, or can be used to look back at the team's 2016 performance. The geo bubble map created for visual six provides a useful glimpse into the Cubs 2016 season to see how they performed against other teams. In another light, this visual can provide some insight as to how players and coaches should prepare to face teams they played poorly against in previous years. Visual seven, on one hand, shows the shortcomings in pitching during the cubs 2016 season. On another note, the visual can also be utilized as a tool that allows coaches and staff to control pitching performance in future games. Much like the other two, visual nine doubles as both a means of understanding how batting works fundamentally within the MLB, and as a tool to be used by coaches to evaluate any given player's batting performance. Due to these visuals incorporating multiple uses, our group theorizes that they would likely provide the greatest value to interested parties among all the other visuals created

During their 2016 season, the Cubs obtained a record of 103-58, which is no simple feat for any professional baseball team. In all, we feel that we have learned a lot from creating this report. Not only did we use Tableau to analyze data from a real-world scenario, but we used it with information that we deemed exciting and interesting. In breaking down a data set into smaller components to building visuals, one will be awarded with a more profound, and encompassing understanding of the data than what could normally be gained at face value.

Works Cited: Data Source

"2016 Chicago Cubs Statistics." *Baseball Reference*, Sports Reference LLC, 2016, www.baseball-reference.com/teams/CHC/2016.shtml.