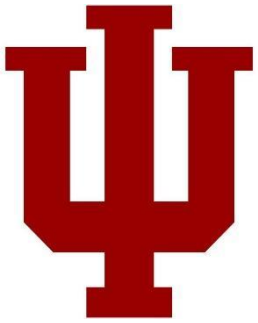

Major League Baseball Pitcher Scouting Visualization Report

Joel Klein

DSCI 590: Data Visualization

Indiana University - Bloomington

July 14, 2022

Abstract:

Major League Baseball (MLB) is the second highest grossing major sports league, reaching 10.7 billion USD in revenue in 2019 [1]. Batting in the MLB is a very difficult task with the league average on-base percentage in 2021 merely .317. Most MLB teams attempt to prepare their hitters prior to games with analysis on the opposing pitcher, typically including information on pitch locations, velocity, spin rates, and tendencies. Starting in 2015, video technology and cloud services help track pitch-level game events further enabling both in-game and out-of-game baseball analytics. Such tracking capabilities capture every pitch including pitch velocity, location, spin rate, and the result of the pitch (i.e., strike, ball, out, single, home run). Using the publicly available MLB data, this effort intends to overcome limitations of common MLB pitch tracking visuals while building an interactive dashboard preparing pre-game scouting reports on MLB pitchers.

Introduction:

Motivation

The goal to gain competitive advantages for hitters in Major League Baseball through intel dates back historically to the game's inception. Today, MLB teams prepare detailed scouting packets and reports prior to every game capturing opposing pitchers' tendencies attempting to provide their hitters an advantage while they are at the plate.

Throughout an at-bat, hitters need to be ready to face several pitch types moving in multiple directions, with different movement patterns, thrown at different velocities and directions. This is not an easy task which is why the big leaguers are paid substantial multi-million-dollar salaries to hit a round white ball. For instance, in the case of a 90 mile per hour fastball, the pitch only takes .4 seconds to reach home plate upon release from the pitcher's hand. According to researchers at UC Berkeley, hitters only have about .25 second to see the ball as it is moving towards them [2].

Only 24.4% of pitches thrown are hit in the 2021 season [3]. Any intel enabling a hitter to better guess what pitch is coming next serves a potential advantage against the opposing team.

Background & Objectives

Although Major League Baseball team's scouting reports are proprietary and not made available publicly, there are several pitch visualizations and reports created by broadcast outlets such as

ESPN, Fox Sports, and MLB Network as well as *Baseball Savant*, the MLB data provider, in their *Pitcher Visualization Report* capturing pitcher tendencies [4].
The Baseball Savant Pitcher Visualization Report is riddled with very poor charts to display pitcher statistics and basic tendencies. The series of charts lists the attempt to visualize key statistics, a critique, and an objective for an improved visual solution.

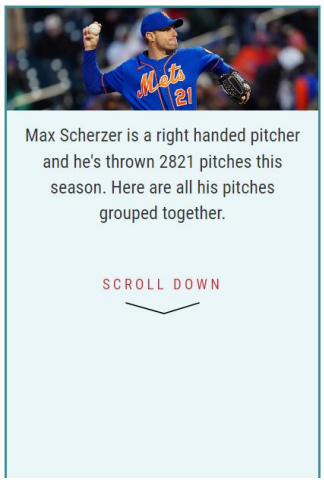

Figure 1: Scherzer 2021 Pitches (Baseball Savant)

Attempt: Illustrate the magnitude of the number of pitches thrown.

Critique: There is no relative comparison to other pitchers.

Objective: Plot a bar chart to show the rank order of Max Scherzer's single game average pitch count compared to other MLB pitchers.

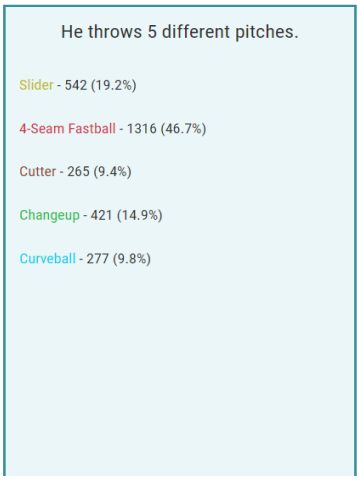

Figure 2: Scherzer 2021 Pitches by Pitch Type (Baseball Savant)

Attempt: Illustrate the number and percentage of pitches thrown for each pitch type.

Critique: Readers can interpret length better than area. The circles are not sorted.

Objective: Plot a bar chart to show the percentages of each pitch type thrown sorted by most common to least.

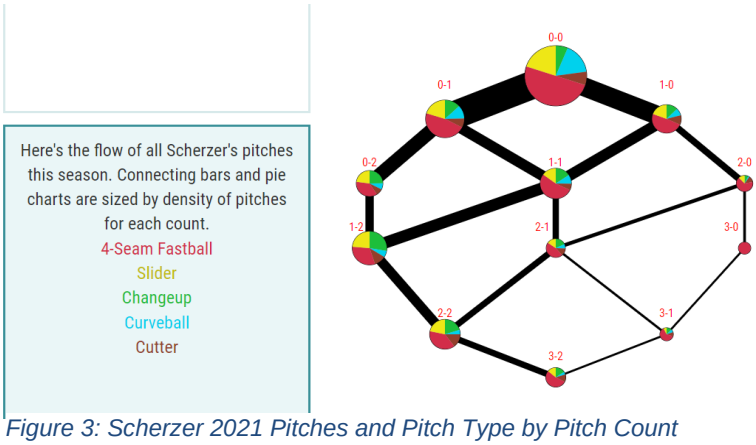

Figure 3: Scherzer 2021 Pitches and Pitch Type by Pitch Count (Baseball Savant)

Attempt: Illustrate the pitch type breakdown and number of pitches thrown by pitch count situation.

Critique: Pie charts are difficult to interpret and very small to view. Differing sizes introduce difficulty interpreting area pitch selection.

Objective: Split into two charts showing pitch selection and number of pitches thrown by count situation.

One of the most common charts shown while viewing MLB games on television is a form of visualization which shows pitch location. Several common visualization charts exist but fail to capture the true density of pitch locations, easily identifiable to the reader. The images from left to right respectively all struggle in unique ways.

One common visual (left chart) partitions the strike zone into 9 equal size boxes and highlights each box by the frequency a pitch is thrown in that vicinity [5].

This chart fails to show enough granularity of the true areas of the strike zone where pitches are thrown. Another common form shows both the trajectory of the pitch, and the location of the pitch in the strike zone (center chart) [6].

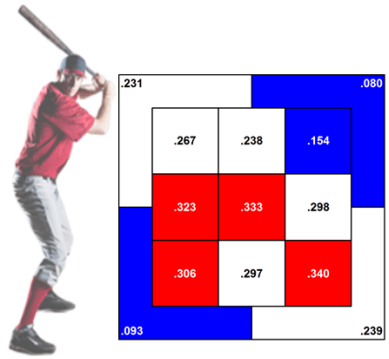

Figure 4: Partitioned Strike Zone Heat Map (The Athletic)

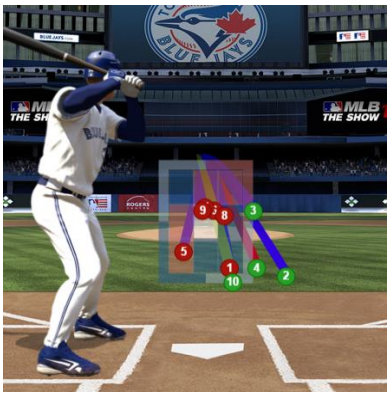

Figure 5: At-bat Strike Zone Graph (Royal Review)

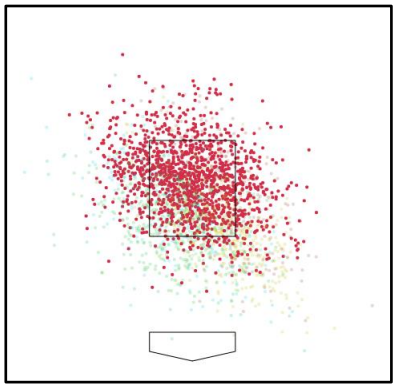

Figure 6: Scherzer 2021 4-Seam Scatterplot (Baseball Savant)

This visual works well in scenarios of single at-bats with few pitches but wouldn't be effective with more data. The third visual (right chart) from *Baseball Savant's Pitcher Visualization Report* shows a scatterplot of all Max Scherzer's four-seam fastballs in red during the 2021 season, with all other pitches faded in the background. This visual runs into the issue where too many points are plotted, and the true density cannot be detected (because Max Scherzer throws a lot of pitches).

The objective or solution to solving the pitch location challenge is to chart a 2D density plot (assuming there are enough pitches thrown in the data) to be able to highlight an appropriate level of granularity to prospective hitters where pitches might be thrown in an at-bat.

Data:

The data sets used in this study are sourced from publicly available data, courtesy of Major League Baseball. During each game, several state-of-the-art tracking cameras, known as Hawkeye, record every pitch from multiple angles. The sensors track the movement of every player on the field, the flight of the pitch as it flies toward the batter, and the trajectory of batted balls [7]. Much of this data, including all information about pitch statistics such as location, velocity, spin rate, and movement, is made publicly available on a per-pitch level on MLB's data website, *Baseball Savant*, dating back to the 2015 season [8].

Baseball Savant data is a tabular and relational database filled with rich information capable of illustrating pitcher tendencies. There are several main features within the data set which will be used to visualize pitch level data: pitcher id, pitch type (i.e. four-seam fastball, two-seam fastball, slider, changeup, etc.), pitch release speed, pitch release position, pitcher-handedness, pitch ball/strike result, pitch count, outs, score, innings, pitcher game pitch count, pitch velocity, pitch movement, and pitch position when crossing the plate.

The analysis is scoped to pitchers who started a game in 2021 having over 100 pitches registered in the 2021 season. In addition, only the first 9 innings are included, no postseason data is included, and rare pitch types such as knuckleball, eephus, and screwballs are filtered out of the data set.

Methods & Process:

Statcast data summary pitch-level tracking statistics enable MLB hitters to analyze pitcher tendencies and performance. The interactive scouting report is split into these two sections.

Pitcher Tendencies

The most informative pitcher tendencies to prepare hitters for in-game at bats are pitch types thrown (both overall and by situation such as count), pitch location, pitch movement, and pitch release point.

Pitch Type Selection

Most pitchers mix the pitch types they throw substantially during at-bats to get hitter out. Often pitchers alter their pitch selection by in-game scenarios, performance, feel, and sometimes fatigue. A simple visual to assess pitch type selection is a simple bar chart with relative pitch frequencies. End users can quickly interpret the bar length comparing relative frequencies amongst pitch types.

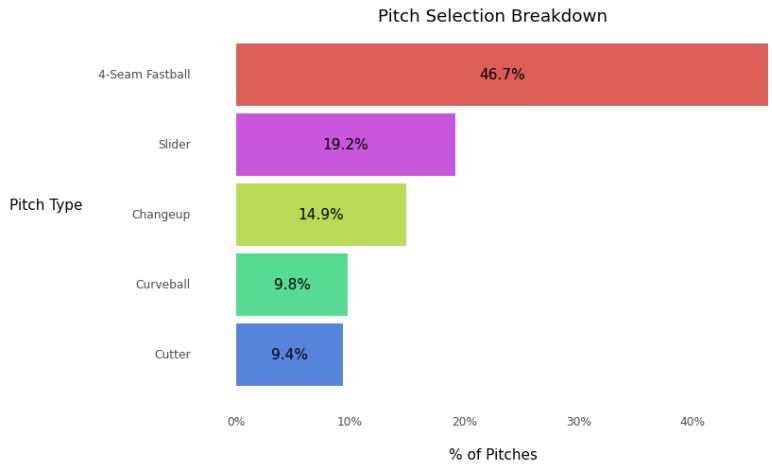

Figure 7: Pitch Selection Design

Pitch selection is heavily dependent on the pitch count. Multiple visuals analyzed together breaks down the overall pitch selection most effectively. Overall pitch selection is based on the number of pitches thrown in each count and the pitches selected for the respective counts. A Sankey diagram not only enables the end user to assess the volume of pitches by count, but also the pitch count flow based on pitch results. An end user can relatively determine the number and percentage of pitches with a 0-0 count which result in either a 0-1 count or a 1-0 count. Also in this scenario, the ordering and location of the pitch counts can be preset to eliminate overlap in the Sankey chart.

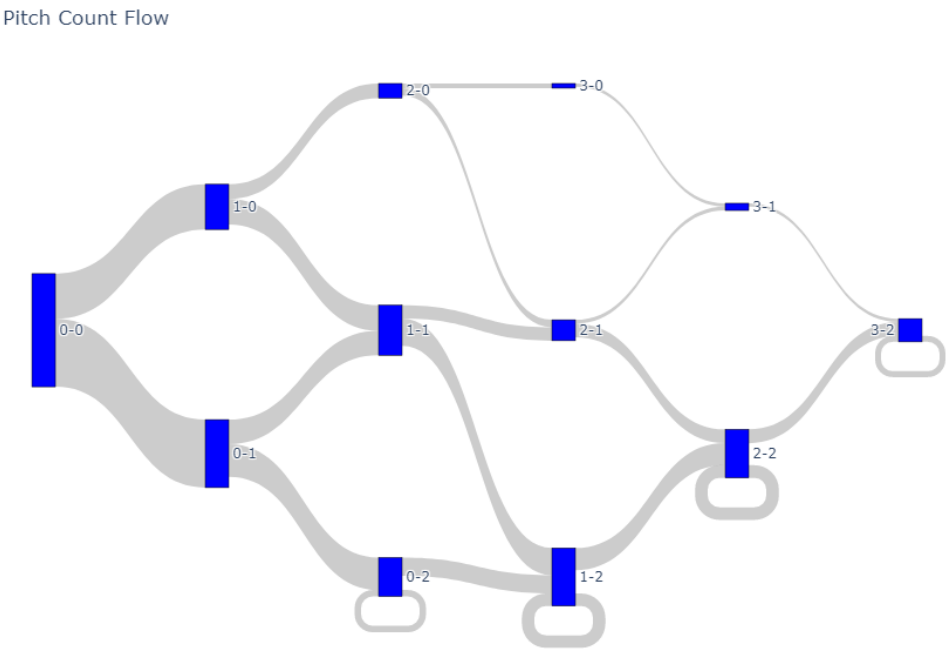

Figure 8: Pitch Count Flow Design

A simple bar chart isn't sufficient to represent the pitch selection by count due to parts of a whole within the count category. A stacked bar chart with the relative frequencies of pitch type by count effectively illustrates the change in pitch selection by count. It is important to note that the colors for the pitch types are the same as the bar chart above. The coloring scheme allows consistency across multiple visuals to quickly enable the end user to interpret subsequent breakdowns of pitch type.

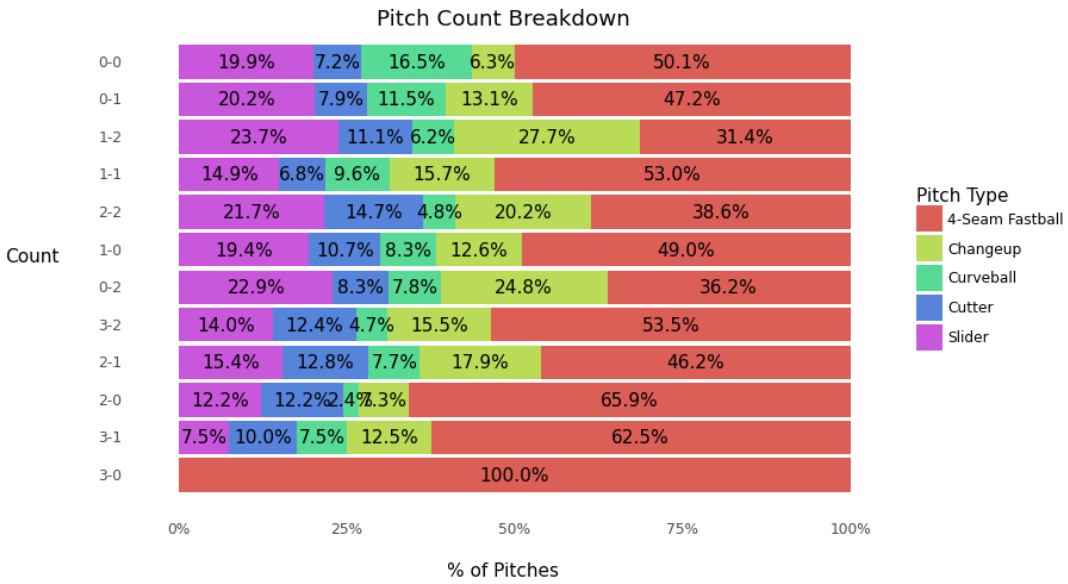

Figure 9: Pitch Selection by Count Design

Humans often struggle to interpret charts with more than 7 categories. There are twelve pitch count possibilities. It is easier to summarize the pitch selection based on groupings of the count situation. Most pitchers' tendencies are similar based on their advantage in the count against a hitter. A pitcher either has a count advantage (ahead), a disadvantage (behind), or neither an advantage or disadvantage (even). Often pitchers choose to throw their "strikeout" or best pitch with 2 strikes. Naturally we split the ahead advantage into two categories: ahead w/ less than 2 strikes and ahead with 2 strikes. Count categories are defined in table 1. The grouped counts generate only four categories making the pitch selection trends more evident.

Even counts	1-0, 2-1, 3-2
Behind counts	2-0, 3-0, 3-1
Ahead (<2 strikes) counts	0-0, 0-1, 1-1
Ahead (2 strikes) counts	0-2, 1-2, 2-2

Table 1: Pitch Count Advantage/Disadvantage Details

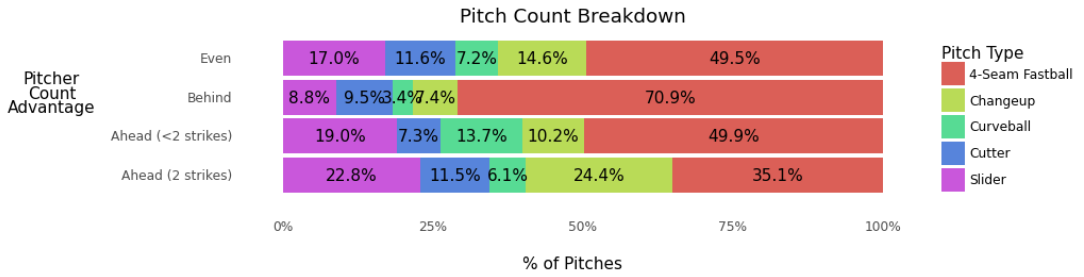

Figure 10: Pitch Selection by Count Design

Pitch Location

Knowing where a pitch is likely to be thrown is highly important to increasing the hitter's chances of reaching base. Common visuals described in the motivation section lack the ability to truly show a granular level where most pitches are thrown. A two-dimensional density plot with the Viridis color mapping effectively illustrates the highest portions of the zone with the most pitches (figure 11). The shaded yellow regions represent higher densities of pitches thrown relative to the strike zone (white box). Pitch location is heavily influenced by the pitch type thrown, the batter's stance, the pitch count, and other in-game scenarios. The interactive dashboard enables the end user to add any of these factors as breakdowns to the 2d density plot.

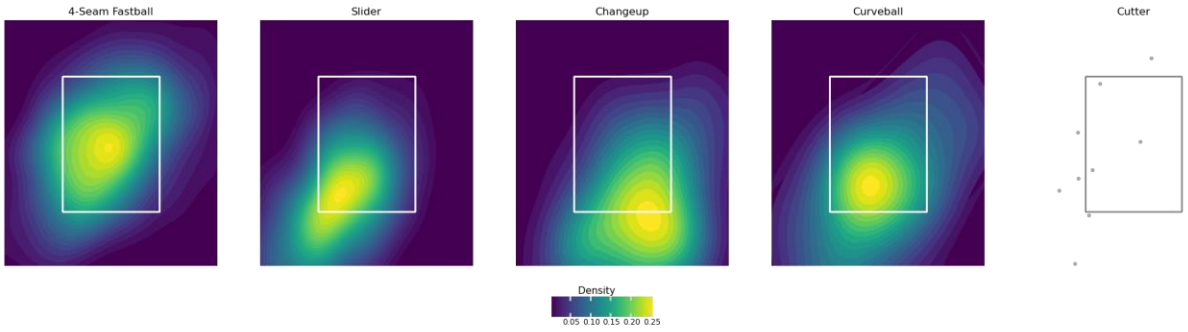

Figure 11: Pitch Location Design

This plot required several iterations to adjust for nuances in plotting pitch location appropriately. First, there are specific breakdowns with very few pitches. 2D density plots often struggle to identify the appropriate contours and plots show random noise where there is limited data. An adjustment to change the plot type to a simple scatterplot where there are less than 50 pitches registered mitigates this risk. With 50 or less points, the end-user can assess via a simple scatterplot where clusters of points are located for the pitch.

A major issue requiring attention is normalizing the strike zone before plotting the results. There are differing heights among batters which adjusts the strike zone and effects where pitchers will locate their pitches (see image below). By adjusting the pitch height feature to be normalized to league average strike zone based on the top and bottom of the strike zone, there is less noise in the data based on hitter height.

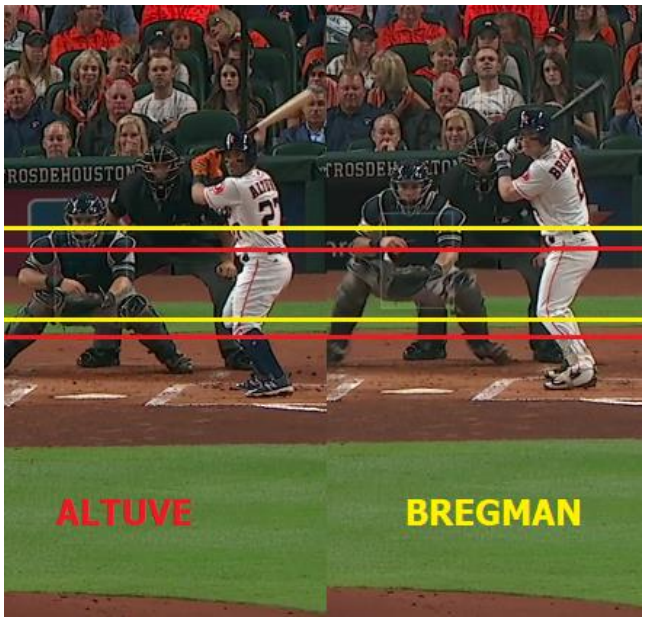

Figure 12: Differences in Strike Zone by Hitter Height

Pitch Movement

Differing pitches vary in spin rate and movement. Movement of pitches varies greatly by the pitch thrown. Pitchers with more movement variation in their pitches tend to perform quite well compared to pitches whose pitches move similarly by pitch type. A scatterplot of vertical and horizontal movement is a simple and effective approach to visualizing pitch movement by pitch type. There are typically clearly visible point clusters for each pitch type.

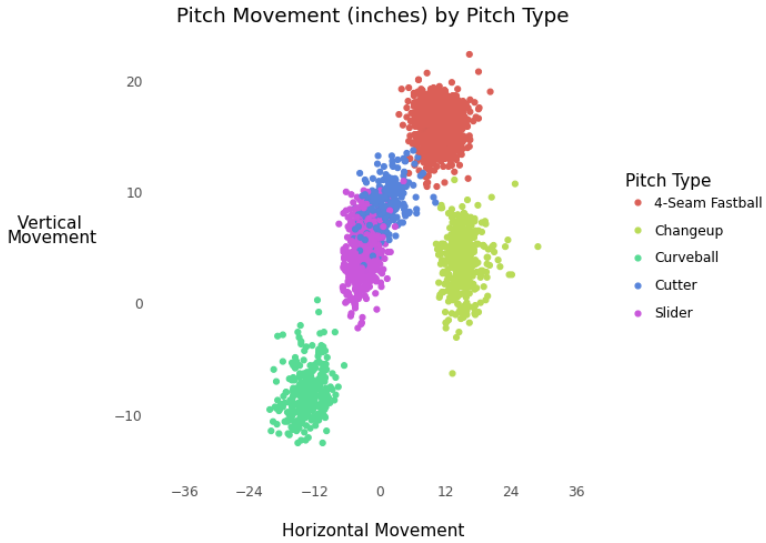

Figure 13: Pitch Movement Design

Pitch Release Point

Where a pitcher releases pitches also sometimes varies greatly by the pitch type thrown. Pitcher's releasing different pitch types in the same relative arm height and angle tend to perform quite well compared to pitches whose arm height and angle differ greatly by pitch type. Most major league hitters perform well enough to identify pitch type based on the pitcher's arm angle if there are differences in release points. Again, the location of the arm angle visualized with a scatterplot colored by pitch type is an effective form diagnosing release point variation. Although the plot sometimes appears to be a blob of points, this represents that the pitcher doesn't vary their release point based on the pitch he throws.

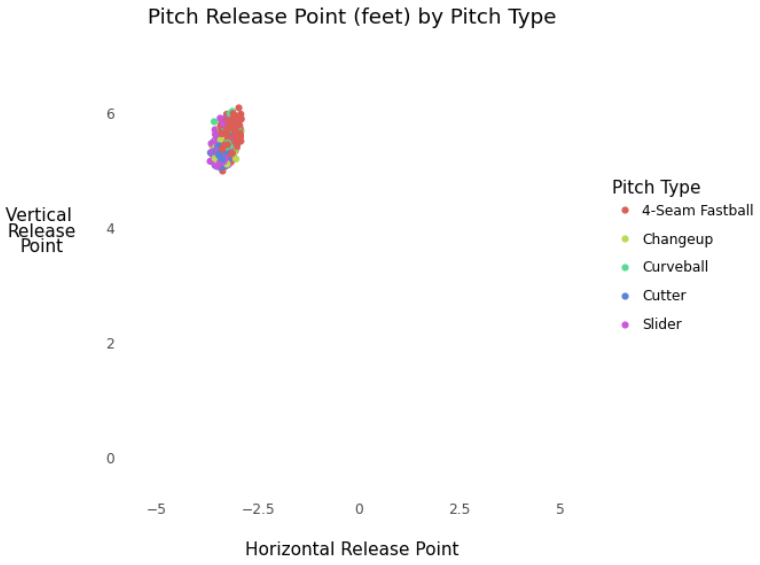

Figure 14: Pitch Release Location Design

Pitcher Performance

In addition to adjusting their approach to pitcher tendencies, hitters also adjust based on likely at-bat results. For instance, if a right-handed pitcher typically gives up weak base hits to right field and it is unlikely a hitter will generate hard contact to hit an extra base hit, the hitter will likely adjust his approach to try hitting the ball to right field. Pitcher at-bat results, contact types, and batted ball locations enable hitters to adjust based on historical results. Hitters also adjust their approach by comparing a pitcher to other MLB pitchers based on historical performance. Peer-norming (determining player percentile ranks amongst other players) is a common technique MLB analysts assess to evaluate pitchers.

At-bat Results

One way to evaluate a pitcher's ability is to analyze the results of their at-bats. Often pitchers are evaluated by their strikeout and walk rates. A simple bar chart provides such information in addition to home-run rate and other possible at-bat results.

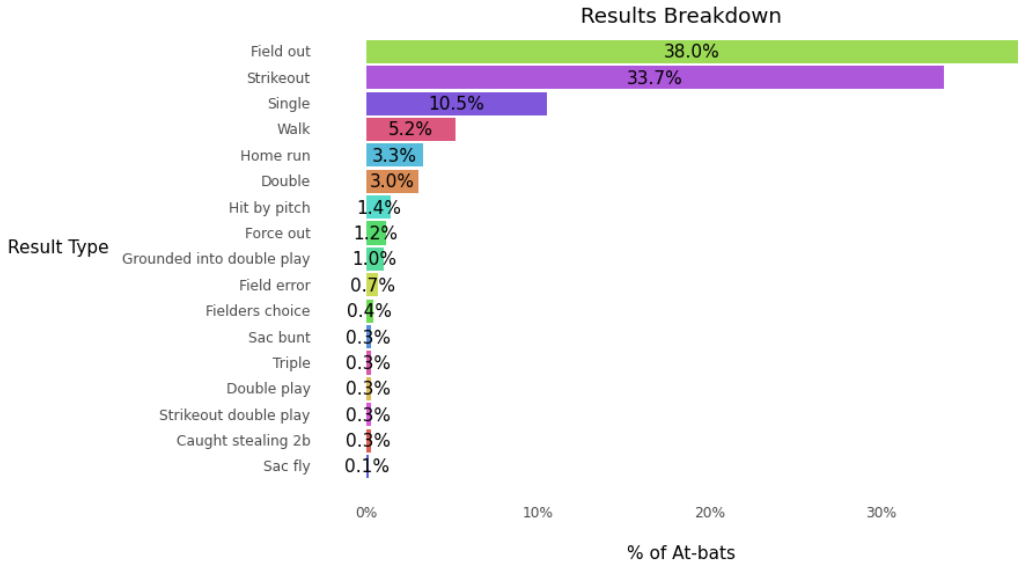

Figure 15: At-bat Results Design

Batted Balls (Contact Type & Location)

A more advanced performance measure is determined based on hitter contact type for his respective batted balls. There are several types of contact defined by the MLB. Typically, high performing pitchers incur lower barrel and solid contact rates.

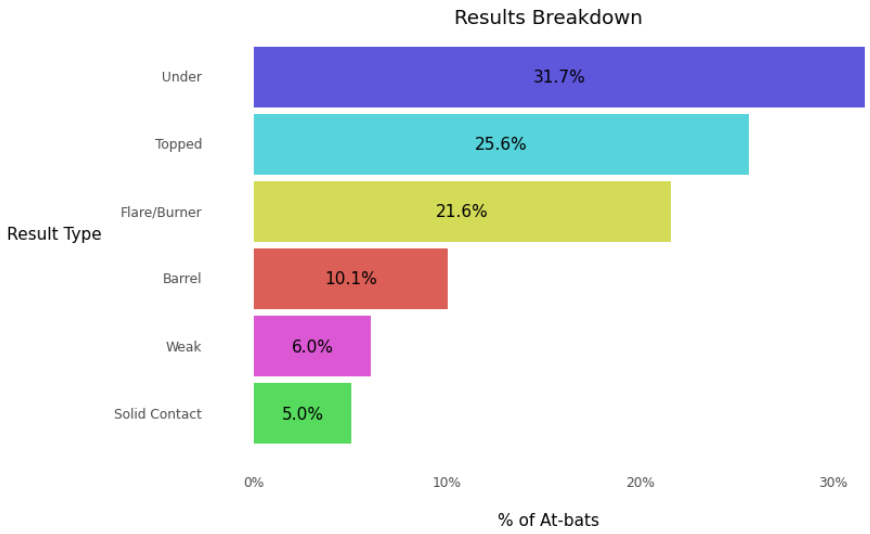

Figure 16: Batted Ball Contact Type Design

In addition to results and contact types, pitcher's batted ball location may determine how defenses are going to position themselves against certain hitters. A hitter can also adjust their approach based on similar hitter results. For example, figure 17 highlights batted balls for right-handed hitters against Max Scherzer in 2021. There are a significant number of batted balls to the left side of the infield, while the actual hits are scattered evenly across the outfield. A basic scatterplot works well with this data as there are typically less than 200 batted ball data points to display.

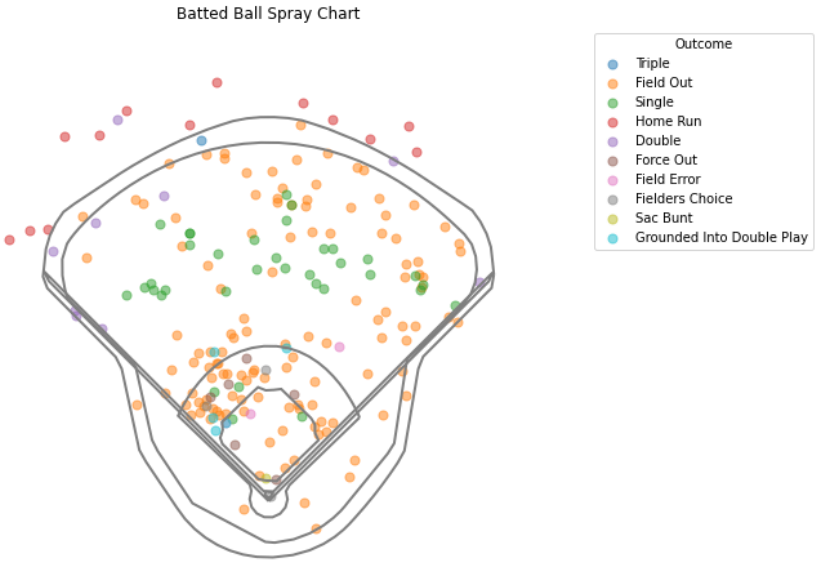

Figure 17: Batted Ball Spray Chart Design

A challenge for producing this visual is selecting the stadium layout for the outline. No two stadiums are alike, so the dimensions of the field for example could be the difference of a deep fly ball being a double vs. a home run.

MLB Pitcher Comparisons:

We can compare each pitcher to the rest of the league by analyzing their percentile rank amongst other MLB pitchers across modern day statistics used to assess pitcher performance. The higher the percentile, the better the pitcher performs in the respective statistical category against other MLB pitchers.

Strike %:	The percentage of pitches thrown deemed as strikes.
Whiff %:	The percentage of pitches thrown which resulted in a swing and miss.
wOBA	A version of on-base percentage accounting for how a player reached base. The value for each method of reaching base is determined by how much that event is worth in relation to projected runs scored (example: a double is worth more than a single)
Exit Velocity	How fast, in miles per hour, a ball was hit by a batter (on average).
Spin Rate	How much spin, in revolutions per minute, a pitch was thrown with (on average).

Table 2: Pitcher Comparison Statistics

A typical visual for player assessment is a radar chart which displays values across multiple features or variables for a single observation. In this case, the radar chart displays a relative comparison of the selected pitcher against all other MLB starting pitchers across each statistical category. The chart depicts which statistic(s) the player performs exceptional (closer to the edge) and/or poorly (closer to the center). In the case of Max Scherzer, in 2021 he performed amongst the top pitchers across each of these major statistical categories.

	Strike %	Whiff %	wOBA	Avg. Exit Velocity	Avg. Spin Rate
Max Scherzer	67.9	37.6	0.249	81.2	2290.1
Rest of MLB	64.1	25.5	0.329	82.4	2224.0

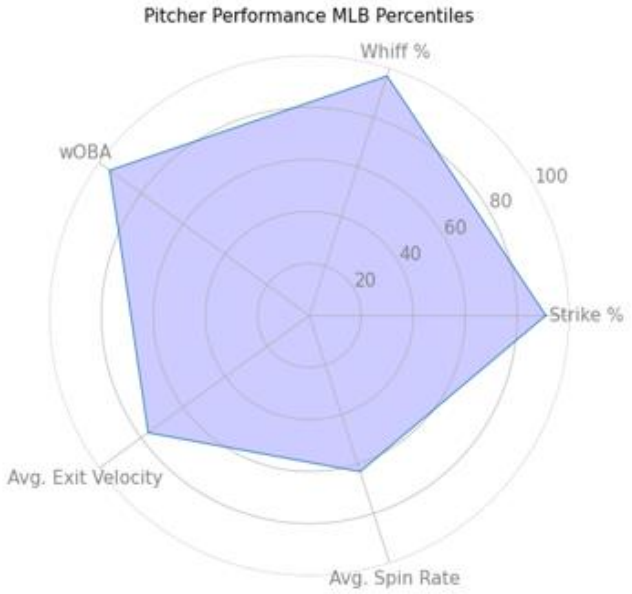

Figure 18: Max Scherzer MLB Pitcher Comparison

To supplement the radar chart, a simple table with the actuals enables the end user to compare the raw statistics for the pitcher to assess how much better the selected pitcher is against league average.

Interactivity

Pitcher tendencies and performance are often heavily influenced by in-game situations. More granular situation breakdowns bring out the most value for hitters versus just overall summary statistics when preparing for plate appearances. A main value-add in the visualization strategy results from plotting pitcher tendencies across in-game situations such as pitch count, outs, inning, score, etc. The end report enables interactive filtering by game situations and offers side-

Pitcher Filters	Batter Filters	Game Situation Filters
• Pitcher Name • Pitch type	• Batter Stance • Batter Name	• Count • Outs • Inning • Runners on base • Score (run differential: pitching team - batting team)

Table 3: Interactive Report Filters

by-side comparison charts comparing scenarios for certain visuals such as pitch type selection and pitch location. Several filters in table 3 are made available to the end user to breakdown the above visuals.

Results:

The purpose of the tool is to prepare for a pitcher prior to facing him in a game. To speak effectively to the endless amounts of insights the tool offers, let's simulate a scenario where 2021 World Series champion short stop Dansby Swanson is scheduled to hit against one of the elite pitchers in the MLB Max Scherzer (who in 2021 finished third in voting for the National League Cy Young Award). Swanson is a consistent hitting right-handed batter whose slash in 2021 was .248/.311/.449 with 27 home runs and 88 RBIs. Max Scherzer is a right-handed pitcher who throws five different pitch types all amongst the top in the league. For Swanson to prepare for Scherzer he must have a detailed understanding for which pitches he will face, when, where they will be in the zone and where he should be targeting to hit the ball to reach base.

To best prepare, the end user can filter to Max Scherzer against right-handed hitters in 2021 and then filter to get more detailed results on previous match ups between Dansby and Max.

Pitch Selection

Dansby Swanson has only faced Max Scherzer for a total of 21 pitches in 2021 so it is difficult to rely on this limited data to estimate the pitcher tendencies. To effectively prepare in this simple report, let's filter to Scherzer's history with right-handed hitters.

Scherzer predominantly alternates throwing fastballs and sliders (pitches with very different actions). When the count is even or he is ahead with less than 2 strikes, he is rarely throwing a pitch outside these options. As he falls behind, he tends to pitch the 4-seam fastball at a significantly higher rate of 72%. The biggest changes in Scherzer's approach occurs when he is ahead in the count with 2 strikes. He throws his changeup at a significantly higher rate (19.4%) compared to roughly a mere 5% in other count situations.

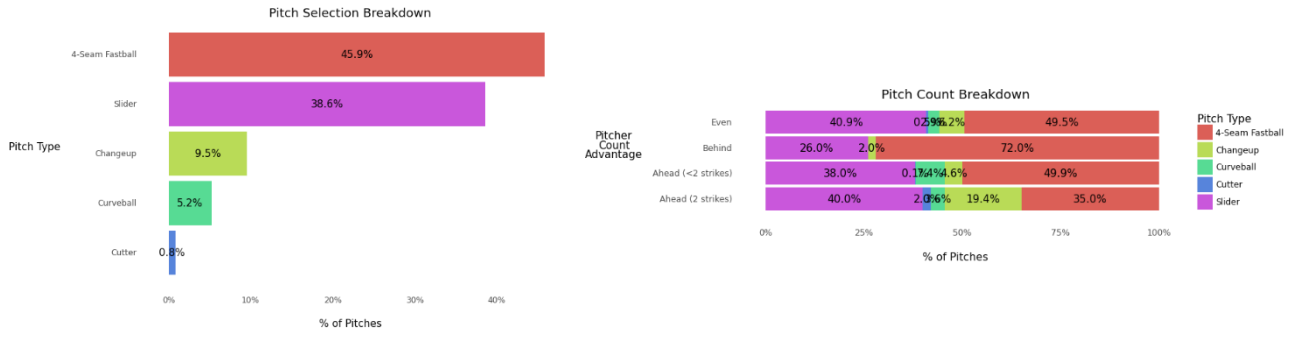

Figure 19: Max Scherzer Pitch Selection v. RH Batters

Pitch Location

The location of Max's pitches is predominantly affected by pitch type and count situation. The predominant locations in the zone are anticipated based on the pitch type against right-handed hitters. The 4-seam fastball is typically thrown over the plate, slider low and away, changeup low and in, and curveball low and central/away. There is less consistency in the 4-seam fastball location precision compared to the other pitchers, making the 4-seam more challenging to hit. The other pitches are also difficult to hit because of the inability for hitters to recognize the pitch movement and release point (highlighted in next sections).

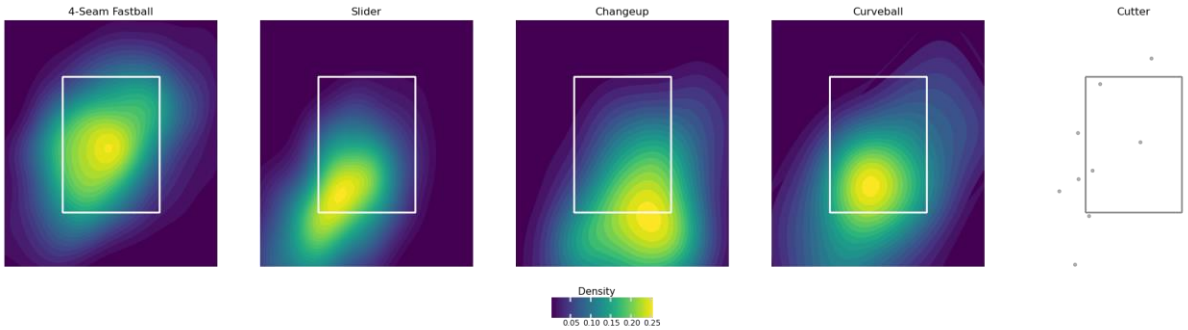

Figure 20: Max Scherzer Pitch Location by Pitch Type v. RH Batters

Scherzer, along with other pitchers, are significantly more dominant when provided a count advantage. When Scherzer is ahead with less than 2 strikes, he is locating pitches low and away. With 2 strikes, the same mentality persists, except he is locating pitches even lower to get hitters to chase and swing at pitches out of the strike zone. In even counts, he still is pitching low in the zone but isn't targeting as far outside while attacking hitters. There is a significant shift in location when he is behind in the count. There is more variation in location throughout the zone and most pitches thrown are over the middle of the zone. Hitters' success chances increase significantly if they find themselves in this situation. However, this situation is rare as there are only very thin lines of pitches in the pitch flow for the behind counts.

Figure 21: Max Scherzer Pitch Location by Count v. RH Batters

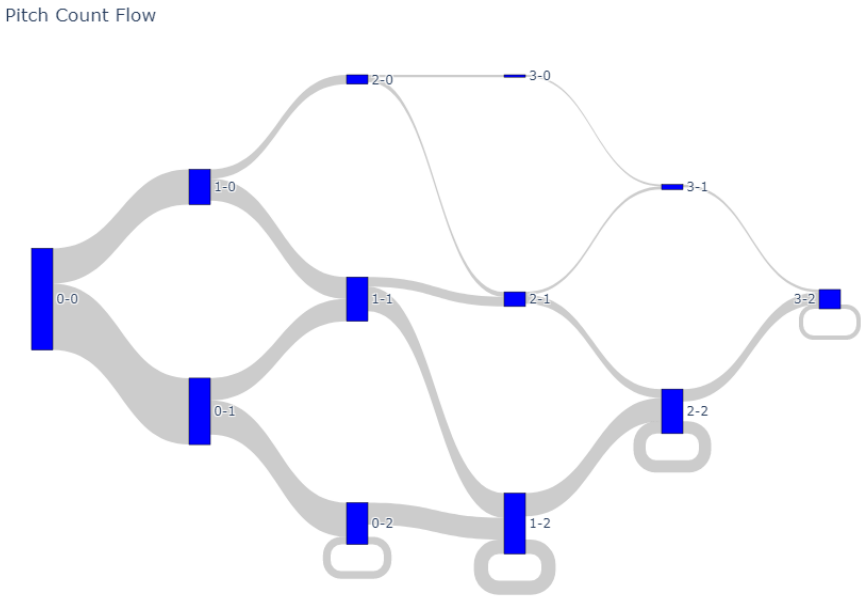

Figure 22: Max Scherzer Pitch Count Flow v. RH Batters

Pitch Movement & Release Point

Max Scherzer is so dominant not just because of his arsenal of pitches, his command, and his ability to locate them, but because he also disguises his pitches effectively. Disguising pitches is the act of making hitters perceive pitches with an expectation which is different than reality. Disguising pitches effectively results from two factors: pitch movement differences and pitch release point similarity by pitch type.

Scherzer effectively differentiates his pitches via drastic changes in movement. Specifically, his 4-seam fastball moves from left to right by about 10 inches, while his slider moves from right to left slightly, with a significant vertical drop. His curveball and changeup also significantly differ in movement and breaks downward substantially in different directions.

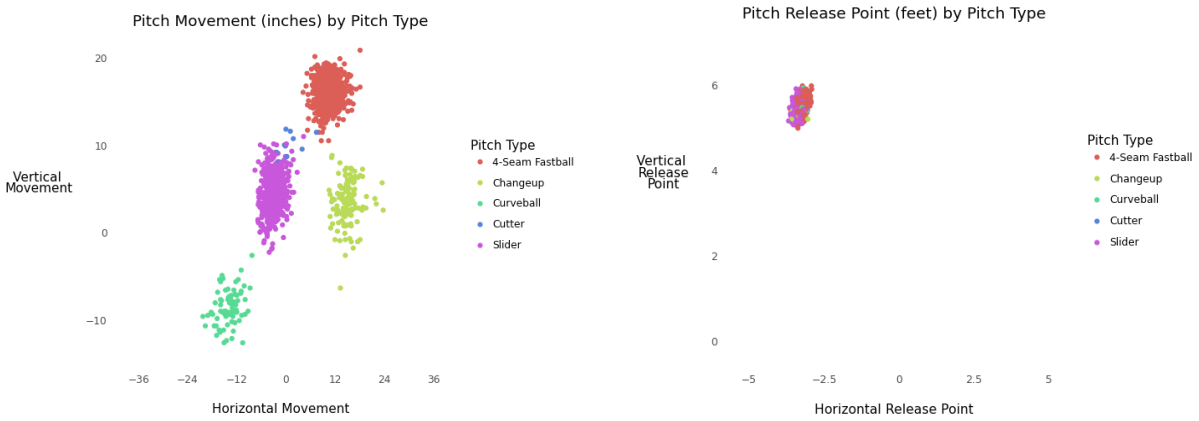

Figure 23: Max Scherzer Pitch Movement & Pitch Release Point by Pitch Type v. RH Batters

It is fascinating how Scherzer varies his pitches drastically while maintaining the same arm angle and release point. This is the ultimate factor resulting in Scherzer's elite status. Numerous videos illustrate this difficulty by overlaying pitches in at-bats with varying movements after the pitches start in the same location. Check out this example of Scherzer against Swanson in 2021.

At-bat Results

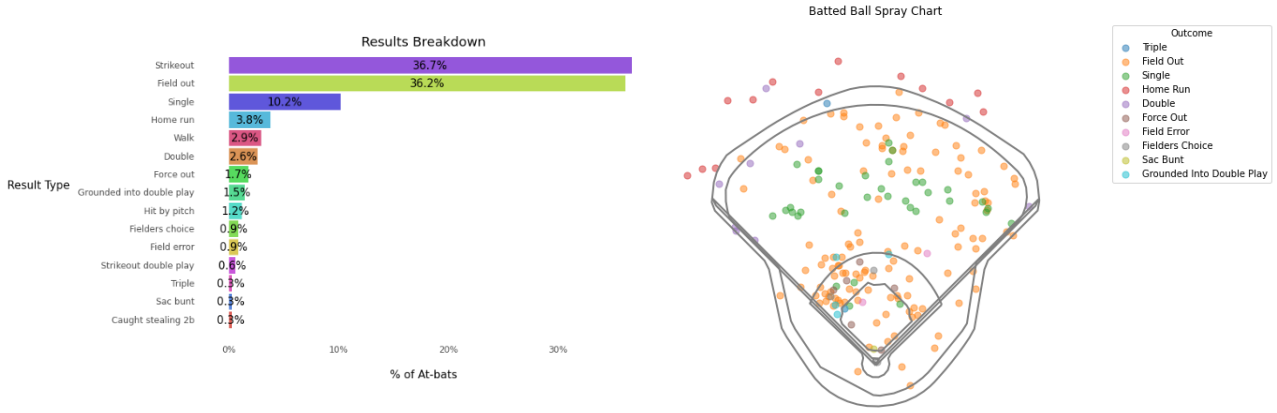

Figure 24: Max Scherzer At-bat Results v. RH Batters

Scherzer has one of the highest strikeout rates in the MLB, striking out over 36% of hitters faced. It is fairly rare to see a hitter reach base against Scherzer (21%) and extremely rare a hitter bats an extra base hit (6.7%). A right-handed hitter's approach against Scherzer shouldn't be to hit the ball deep especially since most of his deeply hit balls are made outs and only 10% of batted balls are what the MLB considers to be 'barreled'. Most of Scherzer's outs against right-handed hitters come from the balls hit to the left side of the infield. Swanson is notoriously a pull hitter, and it is very likely that the defensive team behind Scherzer would place a shift on within the infield to shift defenders to the left side.

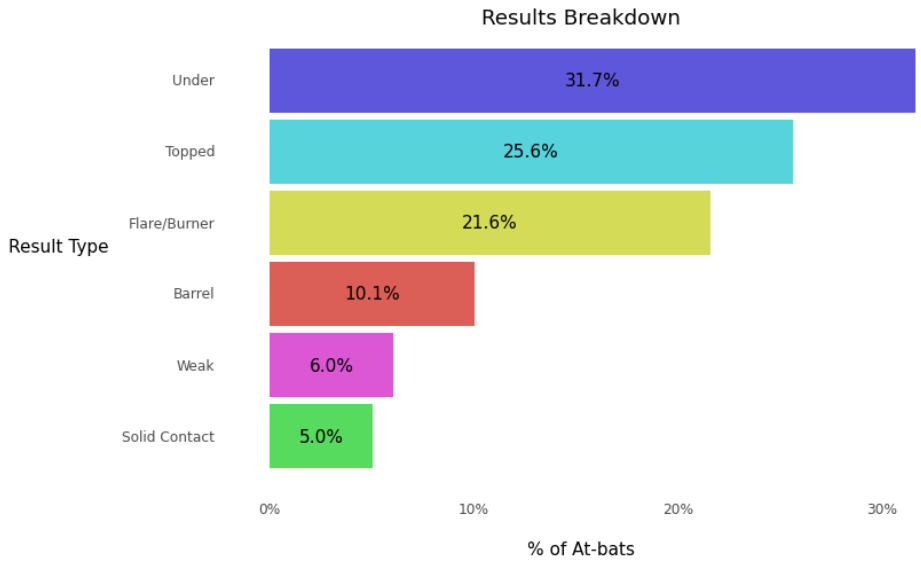

Figure 25: Max Scherzer Batted Ball Contact Type v. RH Batters

In 2021, Swanson faced Scherzer five times resulting in 3 strikeouts, a pop-up out to right field, and surprisingly a homerun to right field. It is possible that Swanson adjusted his approach to attempt and hit the ball to the opposite, but there are so few data points available reducing the influence of only Swansons history with Scherzer to the approach.

Discussion, Conclusion, and Future Work:

Dansby Swanson's chances (or really any hitter's chances) for success against Max Scherzer is limited. Through a series of data visualization, any MLB hitter could quick within 10 minutes prepare for their matchup not just against Scherzer, but of any MLB starting pitcher in 2021. Compared to *Baseball Savant's* interactive pitcher report, more simple and effective visual choices such as bar charts and scatterplots in combination with slightly more advanced charts like 2D density, spray, and radar charts effectively empower hitters to not just consume the historical data but do so quickly and in a simple enough form to reduce cognitive effort memorizing the insights.

Although this report specifically focused on a single matchup between pitcher and hitter, hitters may also use the report to compare pitchers against others recently faced. For example, a hitter may have recently faced Jose Berrios, another above average right-handed pitcher. Berrios pitch selection, location, and action are much different than Scherzer's. For instance, Berrios specializes in his sinker and curveball (different pitch types with significantly different movements).

Although the new interactive details significant information to prepare hitters, there remains substantial opportunities continuing to add more visuals detailing pitcher tendencies. An additional area of interest is determining pitcher pitch selection after throwing a given pitch type the pitch before. For example, if a pitcher throws a 4-seam fastball the previous pitch, are they likely to throw that again?

Another fast follow to this project is optimizing the user experience and hosting the application via a friendly URL increasing public awareness and accessibility versus relying solely on *Baseball Savant's* poor visualizations attempting to detail the same information.

References:

[1] M. Brown. “MLB Sees Record $10.7 Billion In Revenues For 2019” *Forbes*, December 2019.

[2] L. Sommer. “How Can Anyone Hit a 90 mph Fastball? Science Explains!” *KDEQ*, May 2013.

[3] Fangraphs.com, “Leaderboards”, 2021.

[4] Baseballsavant.mlb.com, “Pitcher Visualization Report: Max Scherzer”, 2021.

[5] M. Simon. “Four factors that could shape Francisco Lindor's postseason performance” *The Athletic*, October 2018.

[6] S. Newkirk. “A primer on using strike zone graphs” *Royals Review*, October 2015.

[7] B. Jedlovec. “Introducing Statcast 2020: Hawk-Eye and Google Cloud.” *Medium*, MLB Technology Blog, July 2020.

[8] MLB.com, “Statcast”, 2021.