Using Gaussian Mixture Models to Segment Starting Pitchers Based on Performance in Different Temperatures and Humidity

Cory Boyko

Harrisburg University of Science and Technology

Author Note

This research was done in conjunction with GRAD 699 Applied Projects in Analytics at Harrisburg University of Science and Technology and serves as the culmination of the Master's degree in analytics. The work was done under the advising of Arnie Miles.

Correspondence regarding this article should be addressed to Cory Boyko. E-mail: Cboyko95@gmail.com

Abstract

This paper shows that there are groups of starting major league pitchers where temperature and humidity affect offensive production statistics like ERAd (a variation of ERA), WHIP, and XBH differently. Temperature and humidity also affect the groups' home runs, walks, and strikeouts differently. Groups did not see any differences between each other for the style of balls hit into play (ground, fly, or line) and command (percentage of pitches that are balls). Major League Baseball has regular season games from April to September and games are played in all parts of the United States. The natural question that arises is if the weather affects players and if it affects them all the same. This work studied whether temperature or humidity affected the performance of starting pitchers by clustering starting pitchers' changes in performance from average to high and low temperature and humidity. Pitchers were sampled from 1995 – 2019 and included if they started at least 10 games in average temperature and humidity and 5 games in high and low temperature and humidity. Temperature (N = 582 pitchers) and humidity (N = 614 pitchers) were clustered separately and together (N = 537) and by using different groups of statistics (offensive production, style of balls hit in play, 3 true outcomes, and command). The groups were found using Gaussian mixture models and the final model was selected based on BIC and interpretability.

Keywords: Baseball, pitchers, MLB, weather, clustering, Gaussian mixture model, unsupervised learning, machine learning

Using Gaussian Mixture Models to Segment Starting Pitchers Based on Performance in Different Temperatures and Humidity

Major League Baseball is a unique sport where it is played throughout multiple seasons of the year and the majority of the stadiums do not have coverings or domes. The regular season itself is played from the beginning of April to the end of September with the playoffs going into October and sometimes November. With that in mind, players are playing through all of the elements in all parts of the country from cold and rainy spring days to sweltering hot summer days to cool fall nights. Naturally, weather could impact a pitcher's performance. The question was though, does it affect all pitchers and if it does affect all pitchers, does it affect them all the same? Do pitchers fit to certain profiles based on how they handle different weather situations?

The current world of sabermetrics spends a lot of time analyzing a current player and their performance. They focus on everything from simple statistics like batting average (BA) and earned run average (ERA) to abstract statistics like wins above replacement (WAR) and fielding independent pitching (FIP). These are statistics that are generally controllable. If a player has a low average, then they might not play. If a pitcher has trouble pitching against left-handed hitters, then they're most likely only going to play against right-handed hitters. One thing they cannot control is weather. Teams should want to know how their players preform in different weather scenarios. Not only can they use this information to decide who plays certain days, but they can also use this information to decide who to sign. Baseball teams spend tens if not hundreds of millions of dollars on their roster every year, so they need to know as much information as they can in order to make the best decision for their teams. Based on what is currently publicly available though, there does not appear to be a lot of research on baseball and meteorology. However, with more and more data becoming publicly available in both fields, this

can now be done. Though this research will be focused on baseball, similar questions can be asked and answered about other outdoor sports, such as soccer and football, through similar means.

Though there was not much research, there were some interesting articles that acted as a starting place. The study by Kent and Sheridan (2011) focused on how cloud cover affects pitching and hitting statistics in major league baseball. They had often heard of sunlight affecting a player's sight, but there was no other empirical research. Kent and Sheridan (2011) preformed their analysis with 35,000 games between 1987 and 2002. They received their baseball data from STATS Inc and their weather data from the National Climatic Data Center. Through their analysis, they were able to demonstrate that "offensive production generally declines during clearer-sky daytime games" (Kent and Sheridan, 2011, pp. 14). Since pitching is the opposite of hitting, pitching statistics tended to increase during clearer days.

Koch and Panorska (2013) looked to build on the research by Kent and Sheridan (2013) by analyzing the effects of temperature on Major League Baseball hitters. To do this they analyzed 29,150 games from 2000-2011 They retrieved their baseball data from Retrosheet and their weather data from the National Climatic Data Center. They then grouped the games into "cold", "average", and "warm" based on their temperatures (Koch & Panorska, 2013, pp.360). Their results show that hitting statistics increase in warm weather compared to colder weather including batting average and home runs. They also noted that teams in the American League had larger effects than those in the National League.

This research is immensely important to the field, but there was more to be learned. There is more to weather than temperature and cloud cover. Other everyday weather attributes include wind speed, humidity, and the combinations of them all. Both Koch and Panorska (2013) and

Kent and Sheridan (2011) focused on overall player statistics. In an effort to advance the field, this research focused on slightly differing questions surrounding weather and baseball. First, the analysis was done to investigate the effects of temperature and humidity and how those affect pitching. To dig deeper than past research, individual pitcher performances were investigated to see how players handled pitching through the different weather scenarios instead of team statistics.

Literature Review

Physical Effects of Weather on Baseballs

To understand if weather could affect a pitcher, the effects on the baseball itself should be understood. Faber & Smith (2011) studied the effect that temperature and humidity would have on both baseballs and softballs. Due to the balls' composition, they can be affected by both temperature and humidity, though there is not much quantitative research on this. Some teams in drier climates even keep balls in controlled environments so they do not affect the results and will be consistent to other climates. Smith and Faber studied these effects by measuring CCOR, a measure of the ratio of the speed of an object bouncing off a cylindrical surface to the incoming speed of the same object, as well as weight and stiffness. The baseballs they used were standard MLB and NCAA balls while the softballs were standard ASA balls. They first started by putting all the balls into an environment that had a relative humidity of 50% for four months which was considered the control. The balls were then divided into groups to be placed in other environments of ranging humidity from 11% to 97% for another six weeks. Through this they were also able to show how the balls gain weight and that both baseballs and softballs gain similar amounts. As expected, the higher the humidity the higher the weight gain. Each ball also experienced a decrease in CCOR, but MLB balls experienced the greatest decrease. Similar

trends existed for stiffness of the balls as well. To test temperature, the balls were placed in an environment set at 72 degrees and 50% humidity for at least two weeks. Like the humidity test, the balls were then split into groups and put into coolers or ovens based on the desired temperature for another twenty-four hours. As temperature increased the CCOR did as well with the effect being most pronounced on the MLB baseball. As the temperature increased, the stiffness generally decreased though the MLB ball increased slightly. Faber and Smith concluded that generally stiffness decreased with an increase of temperature and humidity while the CCOR decreased with humidity and increased with temperature.

Moreover, Bahill, Baldwin, and Ramberg (2009) focused their research on how altitude and other meteorological conditions can affect a baseball when thrown by a pitcher and hit by a batter. Instead of taking a purely statistical approach like the previous papers, they looked at the problem through as a physical phenomenon and used applied physics to answer multiple questions. Their research included multiple types of pitches, the physical attributes of a baseball, and the attributes when thrown such as speed and spin rate. Through their equations they were able to demonstrate how distance has an inverse relationship with density and density has an inverse relationship with altitude, temperature, and humidity. Density also has a direct relationship to air pressure. From these relationships, they simulated observations and performed a regression analysis on them to see what affects air density the most. The results of the analysis showed that altitude affected it the most, accounting for 80% of the variation followed by temperature (13%), pressure (4%), and lastly humidity (3%). They found that a 10% decrease in air density would increase a fastball by 1 mile per hour. They also concluded that a reduction in air density leads to an increase in home run distance.

In addition to studies in a lab, there has also been studies applied directly to Major League Baseball. Kraft & Skeeter (1995) researched how weather conditions affect the flight of a baseball when hit. Their hypothesis is that a ball would travel further in warmer weather, since objects can move more freely in less dense air. Besides temperature, Kraft and Skeeter also included humidity and wind speed with a similar hypothesis that less dense air leads to a ball traveling further. Their data consisted of 4000 games between 1991 and 1992 and only fly balls were used in their analysis. The final dataset had 50,000 observations of fly balls. To analyze temperature, they split the data multiple times. First, they analyzed fly ball distances for games above and below 70 degrees. Second, they split the temperatures into the following categories: less than 50 degrees, 50-59 degrees, 60-69 degrees, 70-79 degrees, 80-89 degrees, and greater than 90 degrees. Wind speeds blowing out towards the outfield were left unchanged, speeds blowing from left to right or right to left was given a value of 0, and speeds blowing in towards home plate were given negative values. The data was also categorized based on the previously mentioned wind speeds. The analysis was done by comparing the means by using nonparametric tests. Kraft and Skeeter also used regression analysis with distance as the dependent variable and temperature, wind speed, wind direction, humidity, altitude as the independent variables. Ultimately, the results showed that the balls flew further as temperatures increased. Wind is overall not significant to fly ball distance, but when Kraft and Skeeter looked into individual cities, wind was significant for a portion. Humidity was insignificant as well. Kraft and Skeeter concluded that temperature is the most important to fly ball distance, though the total R^2 of the model was only 0.062.

In addition, Chambers, Page, and Zaidins (2003) studied to see whether baseballs have a higher flyball distance in Colorado, a widely held belief since Denver is one of the highest cities

in the country. Their goal of their study was to answer 2 questions, if a ball travels 10% further in Denver and if they can attribute it to the air density. Their data consisted of 100,000 observations of flyballs between 1995 and 1998 with about 8,000 from each teams' stadium. To obtain their weather data, they set up actual meteorological equipment in the stadium during the 1997 season. They found out that the average fly ball distance at Coors Field (where the Colorado Rockies play) was 302.8 feet while the rest of the league had an average distance of 284.5 feet, for a difference of 6%. Their multiple regression analysis of the weather conditions on fly ball distance showed that the east to west wind was the only significant variable. Chambers, Page, and Zaidins conclude that the belief that balls travel 10% further in Denver is false.

Physical Effects of Weather on People

The effects of weather on people should also be understood. For instance, Elattrache et al. (2011) studied how the strength, range of motion, and playing time differs in high school baseball pitchers who are in warm and cold climates. Their hypothesis was that those who played in warmer weather climates would have an increase in strength, range of motion, and playing time. Their reason for researching this is because of the popular opinion that their hypothesis holds true, without proper research done. Their subjects consisted of 100 male high school pitchers who have not reported injuries; fifty from warm weather climates and fifty from cold weather climates. Other criteria included being between the ages of 14 and 18 and having played competitive baseball for at least the past three years. The cold weather pitchers were from Minnesota while the warm weather putchers were from California and Arizona. The subjects then underwent range of motion and strength tests. To analyze their data Elattrache et al. used t-tests to find differences. The only statistical difference they found in range of motion was the rotation of the throwing limb where pitchers from warm weather climates preformed 8 degrees

better compared to those from cold weather climates and internal rotation with a 5 degree betterment for warm weather pitchers than cold weather pitchers. The combination of these two was also statistically significant. Cold climate pitchers had statistically significant higher dominant arm rotational strength (3% greater) as well as statistically significant higher ratios of external to internal rotational strength (18% greater). Warm weather pitchers pitched an average of 9 months of the year, while cold weather pitchers only pitched 6 months, also statistically significant. They ultimately concluded that high school pitchers from warm weather climates do in fact pitch more and there is an inverse relationship between the amount of pitching and their rotational strength. They state that, “these athletes are a previously unrecognized, vulnerable population in terms of injury risk” (Elattrache et al., 2011, pp. 327).

Similarly, the research by Bahner, Netrer, and Rammsayer (1995) was about the effect of cold temperatures on a person’s response time and how they process information. They note that prior to this research, there has been very little work done on the topic. To study this, they collected 30 males between the ages of 20 and 32 years old. These subjects were also made sure to not have certain allergies, chronic drug use, heart disease, etc. They were then split into two groups with the control group being in an environment with a constant temperature of 28 degrees Celsius. The other group was placed in an environment with a temperature of 5 degrees Celsius that varied to get to the desired temperature safely. To test their response times, the subjects had to look at a picture on a computer and decide the position of an X among dots or among stars. They also had two types of answers. In the first type they had to answer if the X was on the left or right side of the screen and in the second type, they had to click which position (out of four) the X was. These prompts occurred in two groups of forty with a one-minute break in between. To analyze the data, they used three-way ANCOVA. The independent variables were the group

they were in, the prompt they were reacting to (the dots or stars) and the complexity of the response (left or right or the four options). The baselines for the subjects were also used as covariates. After the ANCOVA they used Tukey's HSD to further find the differences. Mean reaction times for the control and experimental group were statistically significant with the mean of the control being 538 milliseconds and the mean of the treatment group being 549 milliseconds. Though the different prompts and response complexities were not significant. They concluded that a lower temperature does impact reaction time, but more research needs to be done to fully describe the effects.

Furthermore, Bush-Joseph et al. (2014) focused their research on if pitchers that grew up in warmer weather had more elbow injuries that required Tommy John surgery, a relatively common but serious injury that could take a pitcher a full season to recover from. Their hypothesis was that those in warmer areas would be more susceptible to those kinds of injury possibly because of the environment allowing them to pitch longer throughout the year. The data included pitchers who played in at least one major league baseball game before needing the surgery between the year of 1974 and 2014. State and country information was used to classify the type of region the pitcher originated from. This left them with 247 total observations, with 139 from warm weather areas and 108 from cold weather areas. To analyze the data the counts were compared by using a Chi-Square test. The test was significant and they concluded that major league pitchers from warm weather regions do in fact require Tommy John surgery more often than those from colder regions.

Previous Research on the Effects of Weather on Baseball

As previously mentioned, there have been studies directly related to the topic of this proposal. The study by Konda and Yamamoto (2019) was about how temperature affects the

Nippon Professional Baseball league, the Japanese equivalent to Major League Baseball. Like the MLB, NPB is also played from March to October, thus also experiencing similar seasonal effects. In addition, they also go through what the Japanese call the rainy season and the typhoon season. Their baseball data consisted of 201,819 observations of plate appearances between 2016 and 2018. Their weather data consisted of temperature, wind speed, and wind direction at ten-minute intervals for those games. After taking missing temperature data into consideration, they were left with 195,542 complete observations. Konda and Yanamoto classified the games as warm if they were at least one standard deviation above the mean, cold if more than one standard deviation below the mean, average-high if between the mean and one standard deviation above the mean, and average-low if between the mean and one standard deviation below the mean. For other variables, they focused on batting statistics such as batting average and on-base percentage. To analyze the data, they used the Wilcox Rank Sum test to compare means. They found that more differences were significant when they compared just warm against cold, specifically batting average, slugging percentage, home run percentage, weighted on base percentage, isolated power, and batting average on balls in play. They concluded that their results match what other studies about Major League Baseball have concluded, that warmer temperatures tend to increase offensive statistics.

Additionally, Skeeter (2009) researched what would be the best timeframe for baseball to be played. This paper was meant to be used as an update to Skeeter's previous work on the same topic. The reason for his updates is that there are new teams, teams have moved, teams have new stadiums, and the playoff structure is different. His belief is that though baseball is played in cold weather at times, it is not ideal. Therefore, the maximum number of games as possible should be in warmer weather. His data consisted of 14 cities where major league baseball is

played with the others being excluded due to not experiencing low seasonal temperature shifts or if they had a dome. Average daily temperatures for the remaining cities were then obtained. Skeeter then proposes two schedules based on whether you want to include the playoffs or not, both with a focus on centering around the warmest days and months. For the season without playoffs, he proposes starting between April 15 and April 23 (approximately 3 weeks later than what is used currently) and ending between October 17 and October 24th. The dates of the season with playoffs shift slightly, so the season should start between April 3 and April 11 and finish between October 30th and November 7th, resulting a shift of about a week. To build on that, Skeeter also collected precipitation data to show that April often has more rain and snow compared to October. Based on the evidence he concluded the season should be shifted 1 to 3 weeks.

One aspect of this proposal is a clustering on pitchers to create groups of those that pitch similarly in different weather conditions. Though there are not any academic quantitative studies on this, there is a qualitative study. Ahrens (2019) studied how one Cub's player, Carl Lundgren, played exceptionally well in cold weather through a historical deep dive. Ahrens wrote that Lundgren, an Illinois native, first began receiving recognition for pitching well in the cold while in college. After pitching well throughout college, he eventually signed with the Chicago Colts (later to be renamed the Cubs), where he earned the name the "Human Icicle" because of his "uncanny effectiveness in wintery conditions" (Ahrens, 2019, pp.93). Though he was gaining recognition as a great pitcher in the cold, he often struggled in the heat of the summer. For example, in the 1905 season he pitched very little in June and July. Ahrens then goes into detail how Lundgren struggled in the summer games throughout the years while still being effective in the early and later months of the season. Even though he would always pitch well in March and

April, he would ultimately end up being taken out of the rotation as the temperature heated up. Though he struggled in the summer months, Lundgren still ended up in the baseball hall of fame, because of his great games in the cold weather of the spring and the fall.

Methods

The research that was conducted was with secondary, quantitative data only. It was done as an exploratory case study which can lead to further studies. The data were collected between January and March of 2021 with the analysis conducted in April of 2021. The data were collected online through a variety of resources. The research consisted of starting pitchers in Major League Baseball.

Participants

A starting pitcher in Major League Baseball is defined as the pitcher who begins the game. These pitchers often pitch the majority of the game as well. In a 162 game season, a starting pitcher starts on average 32 games if they remain in the rotation and do not get injured through the year. The sample was collected as a convenience sample and consisted of players who started at least one game between the years of 1995 and 2019. These years were selected as they were the most recent and provided enough of a sample to conduct the analysis. Data for 2020 was available, but due to the coronavirus pandemic, the 60 game shortened season, the season's late start, the schedule, and the rule changes made for the 2020 season only, a decision was made to exclude that data.

The initial sample consisted of 2,049 starting pitchers. All stadiums in Major League Baseball were built to the team's preference, so some games were played in a dome and therefore closed to the outside weather conditions. Others were built with a retractable roof that are able to open and close the stadium. Most were built being always open to the weather. Games that were played in a dome were automatically removed since they were not affected by the weather conditions. Games that were played in a stadium with a retractable roof were removed as well since there was nothing in any of the datasets that let one know whether the roof was open or not.

Games played at any stadium that was completely open were included. A table of the teams whose games were included and excluded is included in the appendix as table 1. The status of the stadiums' roofs was found on the respective team and stadiums' page from ballparksofbaseball.com. Games where the weather data were not available were removed as they created missing data and could not be analyzed. Games that were played as double headers had both games removed as there was no variable designating which game was which in the play by play logs, so the games could not have the proper weather data attributed to them.

After these cuts, the data consisted of 1,937 starting pitchers. Each of the games were designated as a high, low, and average temperature and humidity game. To qualify for the analysis, starting pitchers had to start at least 5 games in the high and low categories and at least 10 games in the average categories. This number of players in the temperature dataset was 582. The number of players in the humidity dataset was 614. The number of players in both datasets was 537. These were the final counts used in the analysis.

Procedures

Being all secondary data, each compilation of datasets came from a different source. All of the baseball data was pulled from Retrosheet.com, an organization that digitized box scores (game summaries) and play by play files, as well as other baseball data that was not used for this research (Retrosheet, Inc). They collected their data by digitally copying scorebooks, written accounts of baseball games, received directly from Major League teams as well as other personnel closely related to the team (Retrosheet, Inc). The entire process is done through volunteers and the data goes through a lengthy process of proofing. All data provided through them is free as well. Though data from 1970s and before may have some missing data, data from more recent games is all complete and there were no missing data in the 1995 – 2019 data used

for this research (Retrosheet, Inc). From this, there were two sets of data that were created that were used for this research. The first set contained play by play information. The play by play set contained data for every at bat (Retrosheet, Inc). These sets were accessed using an R script from the book "Analyzing Baseball with R" by Max Marchi, Jim Albert, and Benjamin S. Baumer. The script itself was downloaded from Benjamin Baumer's github repository. The script contained a group of functions that accessed Retrosheet's site and downloaded the data year by year (Baumer, 2019). This was done using R (Version 4.0.3; R Core Team, 2020). The other sets that were used contained a summary of each game. These game logs were available on Retrosheet as csv files, so they were simply downloaded.

All weather data for this research came from the National Climatic Data Center. This is a government organization that provides free data and "is responsible for hosting and providing access to one of the most significant archives on Earth, with comprehensive oceanic, atmospheric, and geophysical data." ("About Us"). The data used specifically in this research was the "Local Climatological Data" which is a database of weather data from airports and other weather stations ("U.S. Local Climatological Data (LCD)", 2018). These sets contained hour by hour (and sometimes other increments) local weather data. To access this data, a request was made on their website and then the data was emailed. To find the closest station to the individual baseball stadiums, google maps was used. The list of airports used can be found as table 2 in the appendix. The data was requested in 10 year intervals, the maximum allowed.

Measures

All data transforming and cleaning was done using R (Version 4.0.3; R Core Team, 2020). Each of the sets contained many variables, but not all were used. The weather sets needed very little processing and their measures were able to be immediately used. Each city was given

a flag that matched the home team flag found in the Retrosheet sets. The data was then split into day games and night games and given a flag for which time of day the data represented. Since there were no consistent starting times of each game, the observations between 12 o'clock and 2 o'clock pm local time were used for the day games and the observations between 6 o'clock and 8 o'clock pm local time were used for night games. There were two weather measures that were used from this set. The first was what is called dry bulb temperature. Dry bulb temperature is what is generally referred to when people refer to temperature and was measured in Fahrenheit ("Dry Bulb, Wet Bulb and Dew Point Temperatures"). The other weather measure was relative humidity, or just humidity, and can be between 0% and 100%. Every day had a value of temperature and humidity where the day value is the average of the observations between the day times and the night value is the average of the observations between the night times. An example of an observation of a single day from the final weather set can be found in table 3 of the appendix.

The game logs were used simply to get the time of day the game occurred. The logs had a variable that designated the game as a day game ("D") or as a night game ("N"). These sets were imported into R and combined as one file. They were later used to merge the type of game into the play by play sets.

The play by play needed the most restructuring to go from the raw measures to the measures used for analysis. The initial files had a row for every event that occurred during an at bat. An event was when a play occurred when a player reached base, advanced a base, or made an out. These files contained all information surround the at bat prior to the event occurring such as the pitcher, the defensive players and their positioning, as well as all information after the event that occurred such as whether a run was scored and which base the offensive players ended

up on. The first step was to create a subset of only starters. This was done by creating a set of all pitchers who had their ID in an observation where the inning count was set to 1, namely the first inning. All observation that had a different pitcher than one of the starting pitchers were deleted.

17 variables were used from this set that were used to merge with the other sets or were used to calculate other statistics. There were three date related variables, namely the year, the month, and the day which were used for merging. The flag for the home team was used for merging as well as each game's unique set of variables consisted of the year, month, day, and home team. Pitcher ID was used as the identifier for each pitcher. Pitching hand contained whether the pitcher threw with their left or right hand. Pitching sequence contained the sequence of every pitch thrown and other actions performed by a pitcher such as a pickoff attempt. Event code contained a numeric code of the events. Table 4 of the appendix shows the meaning of each code (Turocy, 2019). Hit flag signified the type of hit, if there was one, which had the value of 0 for no hit,1 for a single, 2 for a double, 3 for a triple, and 4 for a home run. Event outs count held the value of the number of outs a play resulted in and could have been 0, 1, 2, or 3. Batted Ball code contained the code of the type of batted ball such as "G" for ground ball, "F" for fly ball, and "L" for line drive. The final variables contained the destination of the player who batted and the players who were on base after the play, and most importantly if they scored.

From the above variables, new variables were created for each observation. The variables that were created were: outs, hit, home run, extra base hit, walk, strikeout, number of strikes, number of balls, runs, ground, fly, and line. Outs was the same as event outs and was simply renamed. Hits was defined as a 1 if hit flag was not 0 else 0. Home Runs was defined as 1 if hit flag was equal to 4 else 0. Extra base hit (abbreviated as XBH) was defined as a 1 if the hit flag was equal to a 2, 3, or 4 and was a 0 otherwise. Walk was defined as a 1 if the event code

was a 14 or 15 and 0 otherwise. Strikeout was defined as a 1 if the event code was a 3 and 0 otherwise. Runs was defined as the number of players who had a 4 (signifying an earned run reaching home) as their destination variable. Ground was defined as a 1 if batted ball code was a "G" and 0 otherwise. Line was defined as a 1 if batted ball code was an "L" and 0 otherwise. Fly was defined as a 1 if batted ball code was an "F" and 0 otherwise. To get the number of strikes, balls, and pitches more processing had to be done. First, the sequences had to have all symbols removed that did not indicate a ball or strike. A list of what was considered a ball and strike can be found in appendix as table 5 (Retrosheet, Inc). The other issue was that sequences were duplicated if an event occurred that did not involve the batter. For example, if a player was thrown out while stealing a base there were two observations with the sequence before the event and with the sequence before and after the event. To make sure those balls and strikes were not counted twice, the first observation was removed and only the sequence with the entire at bat sequence was kept. The number of balls and strikes was then found by getting the length of the sequences with only symbols dictating a ball and strike respectively.

The data was then summarized using the Dplyr package (Version 1.0.2; Wickham, Francois, Henry, & Müller, 2020). The first round of summarization was to create the stats on a per start basis. Each of the previously mentioned variables were still used but were summed across the entire start. At this point, the starts were spot checked with Baseball Reference, a site that is widely used also uses Retrosheet data for their statistics (MLB Stats, Scores, History, & Records). One limitation of the data was getting the total number of runs. Based off of the rules of baseball, the number of runs for a pitcher in any given start may have been understated. If a pitcher leaves the game with a runner on base and the new pitcher allows the runner to score, the run is attributed to the previous pitcher. There is nothing in the data that can directly account for

this, so though it is not often, the runs can possibly be undercounted. In the summarized sets the number of runs or earned run average was called direct runs and direct earned run average so as not to confuse with the normal definition of earned run and earned run average.

The second and final round of summarization was to get the overall statistics in the low, average, and high temperature and humidity categories. The final set included 11 variables: strikeouts per 9 innings (KK), walks per 9 innings (BB), home runs per 9 innings (HR), balls thrown percentage (B), strikes thrown percentage (K), ground balls per 9 innings (Grnd), fly balls per 9 innings (Fly), line drives per 9 innings (Line), direct earned run average (ERAd), walks and hits per innings pitched (WHIP), and extra base hits per 9 innings. Strikeouts per 9 innings was calculated by summing the number of strikeouts divided by the number of innings thrown (total number of outs divided by 3 since there are 3 outs in an inning) and multiplied by 9. The same was done for walks, home runs, ground balls, fly balls, line drives, extra base hits, and direct earned run average (runs). Presenting stats in the scale of per 9 innings is a common way of scaling baseball statistics. Strikes and balls percentage were calculated by summing the number of balls and strikes divided by the total number of pitches. WHIP was calculated by summing walks and hits and dividing by the number of innings pitched. These are the final measures that were used in the analysis. Each of these are scaled naturally where the higher means more and lower means less. The following statistics are ones that a pitcher would want to be lower: walks, home runs, percentage of balls thrown, ERAd, WHIP, and extra base hits. Strikeouts and percentage of strikes thrown are better for the pitcher when they are lower. The amount of ground balls, fly balls, and line drives generally do not have a preference to a pitcher, though more fly balls can mean more home runs, but some pitchers tend to be “fly ball pitchers”

meaning their style allows for more fly balls. An example of the final data can be found in table 6 of the appendix.

Analysis

The first step that was done was to create the cutoff for what was considered high and low temperature and humidity. Based on the histograms (figures 1 and 2 of the appendix) and the previous work done by Konda and Yanamoto (2019), the cutoff was chosen to be 1 standard deviation away from the mean on either side for high and low temperature and humidity. The high cutoff for temperature was 82.62 degrees Fahrenheit and the cutoff for the low temperature was 60.14 degrees Fahrenheit. The high cutoff for humidity was 78.63% and the low cutoff was 42.48%. The histograms were created using the ggplot2 package (version 3.3.3 ; Wickham, 2016). Once the players' starts were categorized into the low, average, and high categories, the data was summarized to get their total stats for each category. To measure how weather affects a pitcher, new sets of variables were created by subtracting the average weather start statistics from the high and low weather start statistics, respectively.

There were four different sets of statistics that were used to perform the clustering on. Separate sets were used to ensure that clear groups were found for desired statistics. The first set consisted of ERAd, WHIP, and XBH. This set was an overall measure of offense against a pitcher as ERAd measured the number of runs scored, WHIP measured the number of players reaching base, and XBH measured the mount of power an offense was producing against them. The second set consisted of ground ball percentage, fly ball percentage, and line drive percentage. This set was used to get see if their style of play changes in different weather conditions. Often a player is referred to as a ground ball or fly ball pitcher, so this was used to see if it changed with the weather condition. The third set consisted of walks, home runs, and

strikeouts. These statistics are commonly referred to as the “three true outcomes” of baseball (Firstman, 2018). The final set consisted of percentage balls thrown. This was used simply to get an idea of how a pitcher’s command changes in weather conditions. The data was then centered and scaled before the clustering.

To find the different classes of pitchers that had changes in performance in different weather conditions, Gaussian Mixed Methods were used. Gaussian Mixed Methods was the choice of clustering algorithm, because it can allow for clusters of different shapes, sizes, and densities as well as returning a probability for each cluster while still using well known clustering algorithms like K-Means and EM (Géron, 2019). The clustering was done on temperature and humidity separately and then combined. This was done to find well separated classes and allowed the effects of temperature of humidity to be identified alone and combined. For each model, multiple clusters were checked as well since there is no true way to find the real number of clusters. The best clusters were identified by BIC, how separated they were graphically, and how interpretable their centroids were. The models were run using the ClusterR package (version 1.2.2; Mouselimis, 2020). Other packages using in the analysis were base (version 4.0.3 ; R Core Team, 2020), stats (version 4.0.3; R Core Team 2020), ggplot2 (version 3.3.3 ; Wickham, 2016), and ggpubr (version 0.4.0; Kassambara, 2020).

Results

To find out if temperature and humidity affects players differently, each group of variables and weather group had two or three cluster solutions. The full contents of the clusters are in the appendix with high level descriptions in this section. The plots of the solutions were made using principal component analysis to reduce the data to two dimensions to make it plottable. The points themselves were color coded based on their respective clusters.

Temperature – ERAd, WHIP, and XBH

The first set was the offensive output statistics. The BIC plot (figure 3) shows that the number of clusters with the lowest BIC would be 9 clusters. However, 9 clusters produced some clusters with very small sample sizes and makes them tough to interpret. For this 6, 5, and 4 cluster solutions were used. The scatter plots of the solutions can be found as figure 4 of the appendix. The centroids for the solutions can be found as tables 7, 8, and 9.

In the 6 cluster solution, cluster 0 had a decrease in ERAd, WHIP, and XBH for both high and low temperatures, but had bigger decreases in high temperature games. Cluster 1 had an increase in high temperature games and a decrease in low temperature games. Cluster 2 had an increase in each of the statistics in both high and low temperature games but had bigger increases in low temperature games. Cluster 3 had decreases in everything besides an increase in XBH in high temperature games. Cluster 4 had an increase in all stats across high and low temperature games, similar to cluster 2, but larger increases. Cluster 5 had increases in high temperature games and decreases in low temperature games, similar to cluster 1, but generally with smaller increases and decreases.

In the 5 cluster solution, cluster 0 had decreases in each statistic in both high and low temperatures. Cluster 1 had increases in high temperatures and decreases in low temperatures for

each statistic. Cluster 2 had increases in each statistic in both high and low temperatures. Cluster 3 had increases in each statistic in high temperatures and decreases in each statistic in low temperatures, similar to cluster 1, but with smaller increases and decreases. Cluster 4 had increases in all statistics in both high and low temperatures similar to cluster 2, but with larger increases.

In the 4 cluster solution, cluster 0 had increases across all statistics in both high and low temperatures, besides XBH which had a decrease in low temperatures. However, most of these are close to 0, besides ERAd in high temperatures. Cluster 1 had increases in each statistic in high temperatures and decreases in each statistic in low temperatures. Cluster 2 had increases in all statistics in both high and low temperatures. Cluster 3 had decreases in all statistics in both high and low temperatures.

Temperature – Ground Ball, Fly Ball, and Line Drive

The next set was the set with the outcomes of batted balls which consisted of ground balls, fly balls and line drives. Based off of the BIC plot (figure 5), 4 and 6 cluster solutions have the lowest BIC. The scatter plots of the solutions can be found as figure 6 in the appendix. The centroids can be found as tables 10 and 11.

In the 6 cluster solution, cluster 0 had decreases in ground balls in high temperature and fly balls in both high and low temperatures. Ground balls increased in low temperatures. Line drives increased in both high and low temperatures. Cluster 1 had increases in all statistics in both high and low temperature. Cluster 2 had increases in ground balls in high temperatures, ground balls in low temperatures, and line drives in high temperatures. Cluster 2 had decreases in fly balls in low temperatures, fly balls in high temperatures, and line drives in low temperatures. Cluster 3 had decreases in ground balls and line drives in both temperatures with increases in fly

balls in both temperatures. Cluster 4 had decreases in ground balls in both high and low temperatures with increases in both fly balls and line drives in both high and low temperatures. Cluster 5 had increases in fly balls in both high and low temperatures and increases in line drives in high temperatures. Cluster 5 had decreases in ground balls in high temperatures, ground balls in low temperatures, and line drives in low temperatures.

In the 4 cluster solution, cluster 0 had increases in ground balls and line drives in both high and low temperatures and decreases in fly balls in both high and low temperatures. Cluster 1 had decreases in ground ball in both high and low temperatures as well as a decrease in line drives in low temperature. Cluster 1 had increases in fly balls in both high and low temperatures and in line drives in high temperatures. Cluster 2 had decreases in all statistics in both high and low temperatures, besides fly balls in high temperatures. Cluster 3 had decreases in ground balls and fly balls in both temperatures and increases in line drives in both temperatures.

Temperature – Walks, Strikeouts, and Home Runs

The third set consisted of the three true outcome variables. Based off of the BIC plot (figure 7), 4 and 6 cluster solutions had the lowest BIC. The scatter plots can be found as figure 8 of the appendix and the centroids can be found as tables 12 and 13.

In the 6 cluster solution, cluster 0 had increases in home runs and walks in both high and low temperatures with decreases in strikeouts in both high and low temperatures. Cluster 1 also had increases in home runs and walks in both high and low temperatures with decreases in strikeouts in both high and low temperatures. Cluster 2 had decreases in all statistics besides walks in low temperatures which increased. Cluster 3 had increases in all statistics besides strikeouts in high temperatures. Cluster 4 had increases in home runs, strikeouts, and walks in high temperatures with decreases in home runs, strikeouts, and walks in low temperatures.

Cluster 5 had increases in home runs in high temperatures, strikeouts in both high and low temperatures, and walks in low temperatures. Cluster 5 had decreases in home runs in low temperatures and walks in high temperatures.

In the 4 cluster solution, cluster 0 had increases in home runs and walks in both temperatures and decreases in strikeouts in both temperatures. Cluster 1 had increases in home runs in high temperatures and walks in both high and low temperatures. Cluster 1 had decreases in home runs in low temperatures and strikeouts in both high and low temperatures. Cluster 2 had increases in home runs in high temperatures and strikeouts in both high and low temperatures. Cluster 2 had decreases in home runs in low temperatures and walks in both high and low temperatures. Cluster 3 had increases in home runs in high temperature, strikeouts in both high and low temperatures, and walks in low temperatures. Cluster 3 had decreases in home runs in home runs and walks in low temperatures.

Temperature – Ball Percentage

The fourth consisted just of percentage of thrown pitches being balls. Based off of the BIC line plot (figure 9), the solutions that were investigated were 2 and 3 cluster solutions. The scatter plots (figure 10) and cluster centroids (tables 14 and 15) are in the appendix.

In the 3 cluster solution, cluster 0 had increases in balls in both high and low temperatures. Cluster 1 also had increases in balls in both high and low temperatures, though of a larger size compared to cluster 0. Cluster 2 had decreases in both high and low temperatures.

In the 2 cluster solution, cluster 0 had a decrease in balls in high temperatures and an increase in low temperatures. Cluster 1 had increases in balls in both high and low temperatures.

Humidity – ERAd, WHIP, and XBH

The BIC plot (figure 11) shows that the number of clusters with the lowest BIC would be 10 clusters. However, 10 would be too high to be interpreted easily and would create clusters with a very small size. For this 4, 5, and 6 cluster solutions were checked. The scatter plots of the solutions can be found as figure 12 of the appendix. The centroids for the solutions can be found as tables 16, 17, and 18.

In the 6 cluster solution, cluster 0 had decreases in all statistics in both high and low humidity. Cluster 1 had decreases in all statistics in both high and low humidity similar to cluster 0, but with higher magnitudes. Cluster 2 had increases in all statistics in both high and low humidity. Cluster 3 had increases in ERAd, WHIP, and XBHs in low humidity and decreases in high humidity. Cluster 4 had a decrease in ERAd in high humidity, no change in WHIP in high humidity, and increases in ERAd in low humidity, WHIP in low humidity, and XBH in high and low humidity. Cluster 5 had decreases in each statistic in low humidity and increases in each statistic in high humidity.

In the 5 cluster solution, cluster 0 had decreases in all statistics in both high and low humidity. Cluster 1 also had decreases in all statistics in both high and low humidity, but of larger magnitudes than cluster 0. Cluster 2 had increases in ERAd in high humidity, WHIP in high humidity, XBH in high and low humidity, a decrease in WHIP in low humidity, and no change in ERAd in low humidity. Cluster 3 had increases in all statistics in high and low temperature besides XBH in high humidity which had no change from average. Cluster 4 had decreases in all statistics in high humidity and increases in all statistics in low humidity.

In the 4 cluster solution, cluster 0 had decreases in all statistics in high humidity and increases in all statistics in low humidity. Cluster 1 had decreases in all statistics in both high and

low humidity. Cluster 2 had increases in all statistics in both high and low humidity. Cluster 3 had decreases in all statistics in high humidity and increases in all statistics in low humidity.

Humidity – Ground Ball, Fly Ball, and Line Drive

Based off of the BIC plot (figure 13), 4, 5, and 6 cluster solutions have the lowest BIC. The scatter plots of the solutions can be found as figure 14 in the appendix. The centroids can be found as tables 19, 20 and 21.

In the 6 cluster solution, cluster 0 had decreases in ground balls in both high and low humidity and increases in fly balls and line drives in both high and low humidity. Cluster 1 had increases in ground balls and line drives in both high and low humidity and decreases in fly balls in high and low humidity. Cluster 2 had increases in ground balls in high humidity, ground balls in low humidity, and fly balls in low humidity. Cluster 2 had decreases in fly balls in high humidity, line drives in low humidity, and line drives in high humidity. Cluster 3 had decreases in ground balls in high and low humidity as well as line drives in high humidity. Cluster 3 had increases in fly balls in both high and low humidity and line drives in low humidity. Cluster 4 had decreases in ground balls in high humidity, fly balls in high humidity, line drives in high and low humidity with increases in ground balls and fly balls in low humidity. Cluster 5 had increases in ground balls in high humidity, fly balls in low humidity, and line drives in both high and low humidity. Cluster 5 had decreases in ground balls in low humidity and fly balls in high humidity.

In the 5 cluster solution, cluster 0 had decreases in ground balls in high humidity, ground balls in low humidity, and line drives in low humidity with increases in fly balls in low humidity, fly balls in high humidity, and line drives in high humidity. Cluster 1 had increases in ground balls and line drives in both high and low humidity and decreases in in fly balls in both high and

low humidity. Cluster 2 had increases in ground balls in high and low humidity and decreases in fly balls and line drives in both high and low humidity. Cluster 3 had decreases in ground balls in high and low humidity and line drives in low humidity. Cluster 3 had increases in fly balls in both high and low humidity and line drives in low humidity. Cluster 4 had decreases in ground balls in high and low humidity, fly balls in high humidity, and line drives in drives in high humidity. Cluster 4 had increases in fly balls in low humidity and line drives in low humidity.

In the 4 cluster solution, cluster 0 had decreases in ground balls in high and low humidity and line drives in low humidity with increases in fly balls in low and high humidity and line drives in low humidity. Cluster 1 had increases in ground balls and line drives in both high and low humidity and decreases in fly balls in both high and low humidity. Cluster 2 had decreases in ground balls in high and low humidity and increases in fly balls in high and low humidity. Cluster 3 had decreases in ground balls in high humidity, ground balls in low humidity, fly balls in high humidity, and line drives in high humidity with increases in fly balls in low humidity and line drives in low humidity.

Humidity – Walks, Strikeouts, and Home Runs

Based off of the BIC plot (figure 15), 3, 4, and 5 cluster solutions had the lowest BIC. The scatter plots can be found as figure 16 of the appendix and the centroids can be found as tables 22, 23, and 24.

In the 5 cluster solution, cluster 0 had decreases in all statistics in high and low humidity except for home runs in high humidity which had an increase. Cluster 1 decreases in all statistics in high and low humidity except strikeouts in low humidity which increased. Cluster 2 had increases in all statistics in high and low humidity except for home runs in high humidity, similar to cluster 0, but with higher magnitude increases a lower decrease. Cluster 3 had increases in

home runs and walks in both high and low humidity and decreases in strikeouts in high and low humidity. Cluster 4 had increases in home runs and walks in low humidity, decreases in strikeouts in high humidity, strikeouts in low humidity, and walks in high humidity. Cluster 4 had no change in home runs in high humidity.

In the 4 cluster solution, cluster 0 had increases in home runs in low humidity, strikeouts in high humidity, strikeouts in low humidity, and walks in low humidity. Cluster 0 had a decrease in home runs in high humidity and no change in walks in high humidity. Cluster 1 had decreases in home runs and walks and increases in strikeouts in both high and low humidity. Cluster 2 had increases in home runs and walks in both high and low humidity, a decrease in strikeouts in low humidity, and no change in strikeouts in high humidity. Cluster 3 had increases home runs in high humidity, home runs in low humidity, and walks in high humidity. Cluster 3 had decreases in strikeouts in high humidity, strikeouts in low humidity, and walks in low humidity.

In the 3 cluster solution, cluster 0 had decreases in home runs and walks and increases in strikeouts in high and low humidity. Cluster 1 had decreases in home runs in high humidity, strikeouts in high humidity, and strikeouts in low humidity with increases in home runs in low humidity, walks in high humidity, and walks in low humidity.

Humidity – Ball Percentage

Based off of the BIC line (figure 17), the solutions that were investigated were 3 and 4 cluster solutions. The scatter plots (figure 18) and cluster centroids (tables 25 and 26) are in the appendix.

In the 4 cluster solution, cluster 0 had a decrease in high humidity and an increase in low humidity. Cluster 1 had decreases in both high and low humidity. Cluster 2 had increases in both

high and low humidity. Cluster 3 also had increases in both high and low humidity, but of greater magnitudes.

In the 3 cluster solution, cluster 0 had increases in both high and low humidity. Cluster 1 had decreases in both high and low humidity. Cluster 2 had a decrease in high humidity and an increase in low humidity.

Temperature & Humidity – ERAd, WHIP, and XBH

The BIC plot (figure 19) shows that the number of clusters with the lowest BIC would be 10 clusters. However, 10 clusters can be hard to interpret and clusters might have small sizes. For this 4, 5, and 6 cluster solutions were checked. The scatter plots of the solutions can be found as figure 20 of the appendix. The centroids for the solutions can be found as tables 27, 28, and 29.

In the 6 cluster solution, cluster 0 had decreases in ERAd, WHIP, and XBH in high and low temperatures and humidity. Cluster 1 had increases in: ERAd in high temperatures, WHIP in high temperatures, WHIP in low temperatures, XBH in high temperatures, and WHIP in high humidity with decreases in: ERA in low temperatures, XBH in low temperatures, ERA in high humidity, ERA in low humidity, XBH in high humidity, and XBH in low humidity. Cluster 1 had no change in WHIP in low humidity. Cluster 2 had increases in all of the statistics in high and low temperatures and decreases in all of the statistics in high and low humidity. Cluster 3 had decreases in ERAd in high temperature and WHIP in high temperature and increases in all other statistics in high and low temperatures and humidity. Cluster 4 had decreases in ERAd in high humidity, WHIP in high humidity, and XBH in high humidity with increases in all other statistics in high and low temperature and humidity. Cluster 5 had increases in: ERAd in high temperatures, WHIP in high temperatures, XBHs in high temperatures, ERAd in low humidity, WHIP in low humidity, and XBH in low humidity and decreases in: ERAd in low temperatures,

WHIP in low temperatures, XBH in low temperatures, ERAd in high humidity, WHIP in high humidity, and XBH in high humidity.

In the 5 cluster solution, cluster 0 had increases in: ERAd in high temperatures, WHIP in high temperatures, XBH in high temperature, ERAd in low humidity, WHIP in low humidity, and XBH in low humidity with decreases in: ERAd in low temperatures, WHIP in low temperatures, XBH in low temperature, ERAd in high humidity, WHIP in high humidity, and XBH in high humidity. Cluster 1 had increases in all statistics in both high and low temperatures and humidity. Cluster 2 had increases in: ERAd in high temperatures, WHIP in high temperatures, and XBH in high temperatures with decreases in: ERAd in low temperature, WHIP in low temperature, XBH in low temperature, and all statistics in high and low humidity. Cluster 3 had decreases in all statistics in high and low temperatures and increases in all statistics in high and low humidity. Cluster 4 had decreases in ERAd in high humidity, WHIP in high humidity, and XBH in high humidity with increases in ERAd in in low humidity, WHIP in low humidity, XBH in low humidity, and all statistics in high and low temperature.

In the 4 cluster solution, cluster 0 had decreases in all statistics in high and low temperatures and humidity. Cluster 1 had increases in: ERAd in high temperatures, WHIP in high temperatures, ERAd in low humidity, WHIP in low humidity, and XBH in low humidity with decreases in ERAd in low temperatures, XBH in low temperatures, ERAd in high humidity, and XBH in high humidity. Cluster 1 had no changes in WHIP in low temperatures and WHIP in high humidity. Cluster 2 had decreases in: ERAd in high humidity, WHIP in high humidity, and XBH in high humidity with increases in: ERAd in low humidity, WHIP in low humidity, XBH in low humidity, and all statistics in high and low temperatures. Cluster 3 had increases in all statistics in high and low temperatures and humidity.

Temperature & Humidity – Ground Ball, Fly Ball, and Line Drive

Based off of the BIC plot (figure 21), 3 and 4 cluster solutions have the lowest BIC. The scatter plots of the solutions can be found as figure 22 in the appendix. The centroids can be found as tables 30 and 31.

In the 4 cluster solution, cluster 0 had decreases in ground balls in high temperatures, ground balls in low temperatures, fly balls in low temperatures, line drives in low temperatures, flay balls in high humidity, and line drives in high humidity with increases in: fly balls in high temperatures, line drives in high temperatures, ground balls in high humidity, ground balls in low humidity, fly balls in low humidity, and line drives in low humidity. Cluster 1 had decreases in: ground balls in high humidity, ground balls in low humidity, fly balls in high humidity, and line drives in high humidity with increases in: fly balls in low humidity, line drives in low humidity and all statistics in high and low temperatures. Cluster 2 had increases in: ground balls in high temperatures, fly balls in high temperatures, fly balls in low temperatures, line drives in high temperatures, fly balls in high humidity, fly balls in low humidity, and line drives in high humidity with decreases in: ground balls in low temperatures, line drives in low temperatures, ground balls in high humidity, ground balls in low humidity, and line drives in low humidity. Cluster 3 had increases in: ground balls in high temperatures, ground balls in low temperatures, line drives in high temperatures, ground balls in high humidity, ground balls in low humidity, and line drives in low humidity with deceases in: fly balls in high temperatures, fly balls in low temperatures, line drives in low temperatures, fly balls in high humidity, fly balls in low humidity, and line drives in high humidity.

In the 3 cluster solution, cluster 0 had increases in: ground balls in high temperatures, line drives in high temperatures, and ground balls in low humidity with decreases in ground balls in

low temperatures, fly balls in high temperatures, fly balls in low temperatures, line drives in low temperatures, ground balls in high humidity, fly balls in high humidity, fly balls in low humidity, line drives in high humidity, and line drives in low humidity. Cluster 1 had decreases in: line drives in low temperatures, ground balls in high humidity, ground balls in low humidity, fly balls in high humidity, and line drives in high humidity with increases in: ground balls in high temperatures, ground balls in low temperatures, fly balls in high temperatures, fly balls in low temperatures, line drives in high temperature, fly balls in low humidity, and line drives in low humidity. Cluster 2 had decreases in: ground balls in high temperatures, ground balls in low temperatures, line drives in low temperatures, ground balls in low humidity, fly balls in high humidity, and line drives in high humidity with increases in: fly balls in high temperatures, fly balls in low temperatures, line drives in high temperatures, ground balls in high humidity, fly balls in low humidity, and line drives in low humidity.

Temperature & Humidity – Walks, Strikeouts, and Home Runs

Based off of the BIC plot (figure 23), 3 and 4 cluster solutions had the lowest BIC. The scatter plots can be found as figure 24 of the appendix and the centroids can be found as tables 32 and 33.

In the 4 cluster solution, cluster 0 had increases in: home runs in high temperatures, strikeouts in high temperatures, walks in high temperatures, walks in low temperatures, home runs in low humidity, strikeouts in low humidity, and walks in low humidity with decreases in: home runs in low temperatures, strikeouts in low temperatures, home runs in high humidity, strikeouts in low humidity, and walks in low humidity. Cluster 1 had increases in all statistics besides strikeouts in low temperatures and strikeouts in high humidity. Cluster 2 had increases in: home runs in high temperatures, strikeouts in high temperatures, home runs in low

temperatures, strikeouts in low humidity, and walks in low humidity with decreases in: home runs in low temperatures, strikeouts in low temperatures, walks in high temperatures, walks in low temperatures, home runs in high humidity, strikeouts in high humidity, and walks in high humidity. Cluster 3 had increases in: home runs in high temperatures, strikeouts in low temperatures, walks in low temperatures, home runs in low humidity, strikeouts in high humidity, walks in low humidity and walks in high humidity with decreases in: home runs in low temperatures, strikeouts in low temperatures, walks in high temperatures, and strikeouts in low humidity. Cluster 3 had no change in home runs in high humidity.

In the 3 cluster solution, cluster 0 had increases in: home runs in high temperatures, strikeouts in low temperatures, walks in low temperatures, home runs in low humidity, strikeouts in high humidity, walks in high humidity, and walks in low humidity with decreases in: home runs in low temperatures, strikeouts in high temperatures, walks in low temperatures, and strikeouts in low humidity. Cluster 0 had no change in home runs in high humidity. Cluster 1 had increases in: home runs in high temperatures, strikeouts in high temperatures, walks in high temperatures, walks in low temperatures, home runs in low humidity, strikeouts in low humidity, and walks in low humidity with decreases in: home runs in low temperatures, strikeouts in low temperatures, home runs in high humidity, strikeouts in high humidity, and walks in high humidity. Cluster 2 had increases in: home runs in high temperatures, strikeouts in high temperatures, walks in low temperatures, home runs in low humidity, strikeouts in low humidity, walks in high humidity, and walks in low humidity with decreases in: home runs in low temperatures, strikeouts in low temperatures, home runs in high humidity, and strikeouts in high humidity. Cluster 2 had no change in walks in high temperatures.

Temperature & Humidity – Ball Percentage

Based off of the BIC line (figure 25), the solutions that were investigated were 4 and 6 cluster solutions. The scatter plots (figure 26) and cluster centroids (tables 34 and 35) are in the appendix.

In the 6 cluster solution, cluster 0 had increases in both high and low temperatures and decreases in both high and low humidity. Cluster 1 had increases in balls in low temperatures, high humidity, and low humidity with a decrease in balls in high temperatures. Cluster 2 had increases in balls in low temperatures and low humidity and decreases in high temperatures and high humidity. Cluster 3 had increases in balls in low temperatures, high humidity, and low humidity with a decrease in balls in high temperatures. Cluster 4 had decreases in both high and low temperatures and humidity. Cluster 5 had increases in balls in high and low temperatures and decreases in balls in high and low humidity.

Ine the 4 cluster solution, cluster 0 had increases in both high and low temperatures and humidity. Cluster 1 had increases in balls in low temperatures, high humidity, and low humidity with a decrease in high temperature. Cluster 2 had decreases in both high and low temperatures and humidity. Cluster 3 had increases in balls in low temperatures, high humidity, and low humidity and a decrease in high temperatures.

Discussion

Temperature – ERAd, WHIP, and XBH

The 4 cluster solution was the most interpretable and had clusters all of sufficient size which made it the best solution. The majority of the sample do not see that large of a difference in high and low temperatures with their changes mostly around 0. The average pitcher would then have the same outcomes regardless of the temperature which was expected. There was a cluster that consisted of starting pitchers who performed better in colder temperatures than higher temperatures, representing the Carl Lundgren-like group. Cluster 1 showed that while the extreme parity of Carl Lundgren may be an anomaly, the idea that some pitchers pitch better in cold weather and worse in warm weather is supported by these results (Ahrens, 2019). Other patterns found were a cluster that consisted of players worse in both warmer and colder temperatures, one that performed better in both warmer and colder temperatures, and one that consisted of players who performed better in both high and low temperatures. Notably, there is not a cluster that preforms strictly worse in low temperatures and better in high temperatures, dispelling a common belief in baseball. It did not exist in the other cluster solutions either showing that it might not exist at all. Nonetheless, there are 4 distinct groups found in the data. This solution showed that there are clear groups of pitchers that are each affected by temperature in their own way.

The research done by Koch and Panorska (2013) and Konda and Yamamoto (2019) both showed that runs increased as temperatures increased and a similar effect is seen here. Cluster 0, 1, and 2 all have higher ERAd in high temperatures, but there is a substantial subset that pitches better in warmer temperatures in cluster 3 showing that there is more to the effects of temperature effects than strictly saying runs increase as temperatures increase.

Temperature – Ground Ball, Fly Ball, and Line Drive

The 4 cluster solution was the more optimal solution, because the scatter plot had less overlap and all of the cluster sizes were of sufficient size. The clusters didn't have a lot of clear patterns like the offensive statistics. The clearest pattern was that when there were increases in fly balls in both temperatures and decreases in ground balls and line drives in both temperatures (and the opposite), showing that as one increases the others decrease. This makes sense from a practical standpoint since if balls are being hit more in one style, the other(s) would have to decrease and vice versa. Overall, though separate groups were found, they weren't entirely separable and the changes in each are very minor. The statistics are each reported in a unit of per 9 innings so an extra line drive or fly ball per start wouldn't affect the outcome of the game. Temperature does not appear to affect the style of balls hit in play differently for pitchers.

Temperature – Walks, Strikeouts, and Home Runs

The 4 cluster solution was the better solution even though the plot had overlap, but the cluster sizes were all of adequate size. Though the plot didn't look as clear, the cluster means showed that there are differences between each. One cluster contains pitchers with extreme changes, showing that there's a subset of pitchers that can't pitch as well outside of average temperatures. One cluster showed that again, a large portion of the pitchers are immune to temperature. Another cluster is one that's objectively better than the others, showing that there is a group who handles pitching in non-average temperatures very well. Cluster 2 had the best statistics in high and low temperatures where they are close to their statistics in average temperatures or better, never worse. Though there are not clear patterns such as statistics increasing in high temperatures and decreasing in low temperatures or something of that nature in every cluster, looking at each statistic individually across each of the clusters showed that

there was clear differentiation between the clusters. Each of the groups are unique in their own way and each would have a different effect on the outcome of a game. Taking all of that into account, temperature does affect the true outcomes differently for each group.

Across each of the clusters, home runs increased in high temperatures which is consistent with Koch and Panorska's (2013) and Konda and Yamamoto's (2019) work. Home runs decreased in low temperatures similar to their work, but there was a cluster that gave up more home runs in low temperatures. This also shows that there are more insights to be found in the data other than only saying that home runs increase as temperature does. Both of their works also showed that strikeouts and walks were relatively flat across the temperature changes, but these results show that there is more to the underlying structure of data. Overall, strikeouts and walks might not change with temperature, but within the total there are groups where they do change.

Temperature – Ball Percentage

Neither the 2 and 3 cluster solutions are the optimal solutions as they aren't very separable and sizable clusters weren't always able to be discovered. More importantly, neither of the solutions' clusters showed that big of a change between high and low temperatures. If a starting pitcher threw 100 pitches with 40% of them being strikes, a decrease of 1.74% (the largest magnitude change) would be 38 balls. This change would have no impact on the game overall. These clusters show that temperature does not have an effect on a pitcher's command nor does it affect pitchers differently.

Humidity – ERAd, WHIP, and XBH

The 4 cluster was the optimal solution since each of the cluster sizes were of substantial size. Each of the solutions had their own merits, but the patterns generally hold for each solution as well. Similar to the temperature solution, the majority of pitchers have very little change in

high and low humidity, though they did have a more moderate increase in ERAd in low humidity. Other clusters consisted of those whose stats increased in both high and low humidity, decreased in both high in low humidity, and a cluster that decreased in high humidity and increased in low humidity. Each of these represent a clearly defined group of pitchers that preform differently based on the humidity. Notably, no cluster had the pattern where they had increases in high humidity and decreases in low humidity. This pattern existed in the 6 cluster solution, but the decreases in low humidity are close to 0, showing that this pattern does not exist strongly within the sample. Each of the clusters also had a sizable change (besides cluster 0) in each statistic, showing that the effects are impactful on the game, though WHIP had the smallest effects. Overall, this solution showed that humidity does affect a pitcher's offensive statistics differently, though the effects on WHIP are small.

Humidity – Ground Ball, Fly Ball, and Line Drive

The 4 cluster solution was the best solution since the cluster sizes were of substantial size and was the most separable. Similar to the temperature solution, there wasn't a clear pattern throughout. Like the solutions before, the majority of pitchers had very little change in high and low humidity. Generally, if ground balls and line drives changed, fly balls changed in the opposite direction. However, this was not universally true across all clusters, which made it more difficult to interpret. Even though the clusters look to be separable, the changes are not large and would have very little effect on the game, showing that humidity does not have a large effect on the balls hit in play and generally affects all pitchers the same way.

Humidity – Walks, Strikeouts, and Home Runs

The 3 cluster solution was the optimal solution as it had clusters all of substantial size and was the most separable. Similar to what had been seen in the other solutions, the majority of

pitchers had very small to insignificant changes. One aspect that was in this solution and not seen in any of the others was that was a clear separation of clusters where one is much more desirable than the other. Cluster 0 is the more desirable cluster to be in as home runs and walks decreased while strikeouts increased. Cluster 2 had the less desirable changes with increases in home runs and walks and decreases in strikeouts. With each cluster in mind, there was essentially a linear hierarchy where one cluster had worse statistics, one had no change, and one had better statistics. The changes overall aren't as high as the temperature solution, but they could still impact a game. One thing to note is that there doesn't appear to be groups that had an increase or decrease in high humidity and the opposite change in low humidity which was seen in the temperature solutions. This does make them harder to interpret, but the groups are still clearly separate from each other. Overall, humidity has a mild effect on the three "true" outcomes which affects pitchers differently.

Humidity – Ball Percentage

The 3 cluster solution was the best solution due to having adequate sample size and was the most separable. The changes in other clusters were larger than those that were in the temperature solution, but again they were still of little consequence. Practically, the average number of balls would only be increasing or decreasing by one or two balls. Overall, this showed that humidity also does not affect pitchers nor does it affect them differently.

Temperature & Humidity – ERAd, WHIP, and XBH

The 4 cluster solution was the best solution as it was the most separable and had clusters all with sufficient sample. The solutions for temperature and humidity alone also had 4 cluster solutions, so this made it easier to compare. Essentially, the patterns that exist in the combined solution have analogous solutions in the individual cluster solutions and are consistent with the

individual solutions. No new patterns were discovered by looking at temperature and humidity together and no new conclusions can be made.

Temperature & Humidity – Ground Ball, Fly Ball, and Line Drive

The 4 cluster solution was the most separable and both of the individual solutions were 4 cluster solutions, so that was the one that was considered the best solution. The combined solution did have some differences than the individual solutions. Overall, the magnitudes are smaller in the combined solution than in the additional solution, but similar patterns hold. Being of smaller magnitude, the changes are even less significant. This solution reinforces the result that temperature and humidity do not affect pitchers differently.

Temperature & Humidity – Walks, Strikeouts, and Home Runs

Both solutions appeared to be good solutions, but the 3 cluster solution plot was more separable, making it the optimal solution. Similar to the humidity only solution, most pitchers see little to no change in statistics in high and low humidity in the combined solution. The other clusters have fewer clear patterns in the combined solution than the individual humidity solution. For example, the individual solution showed that there was a cluster where strikeouts increased when home tuns and walks decreased. That pattern was meaningful and impactful, but it didn't clearly exist in table 33. The patterns for temperature statistics in the combined solution are of less magnitude similar to the other combined solutions. The clusters are also less separable from each other. Taking all of that into account the combined solution patterns are less meaningful. Looking at temperature and humidity together for the three outcomes did not produce new patterns, but actually made the solutions less clear and less meaningful.

Temperature & Humidity – Ball Percentage

The 4 cluster solution had less overlap which made it more separable overall, so it was the better solution. The patterns here are again analogous to the individual solutions. Additionally, similar to the individual solutions, different patterns are found, but the overall magnitude of the changes makes them meaningless. The conclusion that temperature and humidity do not affect command still stand.

Limitations and Future Research

Like all studies, this one had limitations as well. First, the data consisted of the years 1995-2019, which includes part of the steroid era of baseball, a time where anabolic steroids were used throughout the league which influenced offense. This may have influenced the results here as well. There is no clear documentation of which players used steroids, how widespread use was, or the true effects on the game, though it is believed that offense decreased after they were banned in 2005. However, this is something all studies that use baseball data will have to accept as part of the data. As previously mentioned, ERAd is very similar to ERA, but ERA is the more commonly used statistic and would be found in other literature. Though the effect is minimal, ERA would be more consistent. A lot of data had to be excluded from the research due to playing in a stadium with a retractable dome. None of the datasets had any way of signifying if the stadium was open or not. Though there was no way to avoid it, this could have biased the data against players who play most of their careers on a team that played at a home stadium with a dome or retractable dome. Those players most likely did not have enough games played in each of the categories to be included in the research. More and more teams are building new stadiums with domes or retractable domes, so this will continue to be a limitation in all studies on baseball and weather.

Future research can build off of this as well. Temperature and humidity were analyzed separately, but with a larger sample size the combinations can also be analyzed. For example, the statistics of starting pitchers in high temperature and high humidity games, low temperature and high humidity, or any other combination. This research covered a variety of statistics, but there are many other advanced statistics that can be studied. For example, the spin rate of pitches is an emerging topic in baseball research. Future work can be done on how weather affects spin rates, if it affects spin rates for pitchers differently, and if there is a subsequent effect on other statistics. Over the past couple of years pitching performance has been steadily increasing, so future work can see if the patterns found here hold in future years. Past years can be checked as well to see if these patterns are the same for each era of baseball. Besides temperature and humidity, there are many other weather conditions to be studied such as wind speeds and direction. This work adds to the current knowledge of effects of weather on major league baseball, but there is much more to be researched.

References

About us. (n.d.). Retrieved from https://www.ncdc.noaa.gov/about

Ahrens, A. (2019). Carl Lundgren: The Cubs' Cold-Weather King. *The Baeball Research Journal*, *48*(2), 91–101. Retrieved from http://eds.a.ebscohost.com.proxy-harrisburg.klnpa.org

Bahill, A. T., Baldwin, D. G., & Ramberg, J. S. (2009). Effects of Altitude and Atmospheric Conditions on the Flight of a Baseball. *International Journal of Sports Science and Engineering*, *3*(2), 109–128. Retrieved from https://www.researchgate.net/profile/David_Baldwin8

Bahner, E., Netrer, P., & Rammsayer, T. H. (1995). Effects of Cold on Human Information Processing : Application o f a Reaction Time Paradigm. *Integrative Psychological and Behavioral Science*, *30*(1), 34–45. https://doi.org/10.1007/BF02691388

Ballparks of Baseball - Your Guide to Major League Baseball Stadiums. (n.d.). from https://www.ballparksofbaseball.com/

Baumer, B. (2019, March 15). Beanumber/baseball_r. Retrieved February 28, 2021, from https://github.com/beanumber/baseball_R

Bush-Joseph, C., Cohen, M., Erickson, B. J., Md, Harris, J. D., Romeo, A. A., & Tetreault, M. (2014). Is Tommy John Surgery Performed More Frequently in Major League Baseball Pitchers From Warm Weather Areas? *Orthopaedic Journal of Sports Medicine*, *4*(10), 1–6. https://doi.org/10.1177/2325967114553916

Chambers, F., Page, B., & Zaidins, C. (2003). Atmosphere, weather and baseball: How much farther do baseballs really fly at Denver's Coors Field? *Professional Geographer*, *55*(4), 491–504. https://doi.org/10.1111/0033-0124.5504007

Dry Bulb, Wet Bulb and Dew Point Temperatures. (n.d.). Retrieved from https://www.engineeringtoolbox.com/dry-wet-bulb-dew-point-air-d_682.html

Elattrache, N. S., Hurd, W. J., Kaplan, K. M., Jobe, F. W., Kaufman, K. R., & Morrey, B. F. (2011). Comparison of Shoulder Range of Motion, Strength, and Playing Time in Uninjured High School Baseball Pitchers Who Reside in Warm- and Cold-Weather Climates. *The American Journal of Sports Medicine*, *39*(2), 320–328. https://doi.org/10.1177/0363546510382230

Faber, W., & Smith, L. (2011). The Effect of Temperature and Humidity on the Performance of Baseballs and Softballs. *Procedia Engineering*, 13(1), 200–206. https://doi.org/10.1016/j.proeng.2011.05.073

Firstman, D. (2018). The Growth of 'Three True Outcomes': From Usenet Joke to Baseball Flashpoint. *Baseball Research Jounal,* (Spring). Retrieved from https://sabr.org/journal/article/the-growth-of-three-true-outcomes-from-usenet-joke-to-baseball-flashpoint/

Géron, A. (2019). *Hands-On Machine Learning With Scikit-Learn, Keras, And TensorFlow: Concepts, Tools, And Techniques* (2nd ed.). Sebastopol, CA: O'Reilly Media.

Google. (n.d.). Google Maps. Retrieved from https://www.google.com/maps

Kassambara, Alboukadel. (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr

Kent, W. P., & Sheridan, S. C. (2011). The Impact of Cloud Cover on Major League Baseball. *Weather, Climate, and Society,* 3(1), 7–15. https://doi.org/10.1175/2011WCAS1093.1

Koch, B. L. D., & Panorska, A. K. (2013). The impact of temperature on Major League Baseball. In Weather, Climate, and Society, 5(4), 359–366. https://doi.org/10.1175/WCAS-D-13-00002.1

Konda, K., & Yamamoto, Y. (2019). Analysis of the Impact of Temperature on Nippon Professional Baseball. 17th International Conference on ICT and Knowledge Engineering. https://doi.org/10.1109/ICTKE47035.2019.8966878

Kraft, M. D., & Skeeter, B. R. (1995). The effect of meteorological conditions on fly ball distances in North American major league baseball games. *Geographical Bulletin - Gamma Theta Upsilon*, *37*(1), 40–48. Retrieved from https://www.researchgate.net/profile/Brent_Skeeter

Marchi, M., Albert, J., & Baumer, B. S. (2019). *Analyzing baseball data with R*. Boca Raton: CRC Press.

MLB Stats, Scores, History, & Records. (n.d.). Retrieved from https://www.baseball-reference.com/

Mouselimis, L. (2020). ClusterR: Gaussian Mixture Models, K-Means, Mini-Batch-Kmeans, K-Medoids and Affinity Propagation Clustering. R package version 1.2.2. https://CRAN.R-project.org/package=ClusterR

National Centers for Environmental Information. Data Tools: Local Climatological Data (LCD). Local Climatological Data (LCD) | Data Tools | Climate Data Online (CDO) | National Climatic Data Center (NCDC). https://www.ncdc.noaa.gov/cdo-web/datatools/lcd.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Retrosheet, Inc. Retrosheet. https://www.retrosheet.org/.

Skeeter, B. R. (2009). A revised climatically optimal Major League Baseball season in North America. *Geographical Bulletin - Gamma Theta Upsilon*, *50*(2), 83–91. Retrieved from http://eds.a.ebscohost.com.proxy-harrisburg.klnpa.org

Turocy, T. (2019). Cwevent: Expanded event descriptor. Retrieved from http://chadwick.sourceforge.net/doc/cwevent.html#cwtools-cwevent-eventtype

U.S. Local Climatological Data (LCD). (2018, May 11). Retrieved from https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc:C00684/html

Wickham, H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Wickham, H., François, R., Henry, L., & Müller, K. (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr

Appendix

Table 1. Status of Stadium Roofs

Team	Open or Closed
Arizona Diamondbacks	Closed
Atlanta Braves	Open
Baltimore Orioles	Open
Boston Red Sox	Open
Chicago Cubs	Open
Chicago White Sox	Open
Cincinnati Reds	Open
Cleveland Indians	Open
Colorado Rockies	Open
Detroit Tigers	Open
Houston Astros	Closed
Kansas City Royals	Open
Los Angeles Angels	Open
Los Angeles Dodges	Open
Miami Marlins	Closed since 2012
Milwaukee Brewers	Closed since 2001
Minnesota Twins	Closed until 2010
New York Mets	Open
New York Yankees	Open
Oakland Athletics	Open
Philadelphia Phillies	Open
Pittsburgh Pirates	Open
San Diego Padres	Open
San Francisco Giants	Open
Seattle Mariners	Closed
St. Louis Cardinals	Open
Tampa Bay Rays	Closed
Texas Rangers	Open
Toronto Blue Jays	Closed
Washington Nationals/Montreal Expos	Closed until 2005

Table 2. Airports Used to Approximate Stadium Weather Conditions

Team	Airport
Atlanta Braves	Hartsfield-Jackson Atlanta International Airport
Baltimore Orioles	Baltimore/Washington International Thurgood Marshall Airport
Boston Red Sox	Boston Logan International Airport
Chicago Cubs	O'Hare International Airport
Chicago White Sox	O'Hare International Airport
Cincinnati Reds	Cincinnati/Northern Kentucky International Airport
Cleveland Indians	Cleveland Hopkins International Airport
Colorado Rockies	Denver International Airport
Detroit Tigers	Detroit Metropolitan Wayne County Airport
Kansas City Royals	Kansas City International Airport
Los Angeles Angels	Los Angeles International Airport
Los Angeles Dodgers	Hollywood Burbank Airport
Miami Marlins	Miami International Airport
Milwaukee Brewers	Timmerman Airport
Minnesota Twins	Minneapolis−Saint Paul International
New York Mets	LaGuardia Airport
New York Yankees	LaGuardia Airport
Oakland Athletics	Oakland International Airport
Philadelphia Phillies	Philadelphia International Airport
Pittsburgh Pirates	Pittsburgh International Airport
San Diego Padres	San Diego International Airport
San Francisco Giants	San Francisco International Airport
St. Louis Cardinals	St. Louis Lambert International Airport
Texas Rangers	Dallas/Fort Worth International Airport
Washington Nationals	Ronald Reagan Washington National Airport

Table 3. Example of a Day's Data in the Weather Set

Year	Month	Day	DayNight	Temperature	Humidity	HomeTeam
2006	5	21	D	83	51.5	ATL
2006	5	21	N	78.25	45	ATL

Table 4. Event Code Meanings

Code	Event
2	Generic Out
3	Strikeout
4	Stolen Base
5	Defensive Indifference
6	Caught Stealing
8	Pickoff
9	Wild Pitch
10	Passed Ball
11	Balk
12	Other Advance/Out Advancing
13	Foul Error
14	Walk
15	Intentional Walk
16	Hit by Pitch
17	Interference
18	Error
19	Fielder's Choice
20	Single
21	Double
22	Triple
23	Home Run

Table 5. Codes for Balls and Strikes

Strikes		Balls	
Code	Meaning	Code	Meaning
C	Called strike	B	Ball
F	Foul	H	Hit Batter
K	Strike	I	Intentional Ball
L	Foul Bunt	P	Pitchout
M	Missed Bunt	V	Called Ball, Pitcher Went to his Mouth
O	Foul Tip on Bunt		
Q	Swinging on Pitchout		
R	Foul on Pitchout		
S	Swinging Strike		
T	Foul Tip		
X	Ball put Into Play		
Y	Ball put Into Play on Pitchout		

Table 6. Example of Final data (below is average temperature starts)

Pit_ID	Hand	HR	XBH	KK	BB	Grnd	Line	Fly	Ks	Bs	ERAd	WHIP
Hallr001	R	0.779	2.36	7.27	2.06	14.71	5.07	5.54	66	34	3.52	1.22

Figure 1. Histogram of Temperature

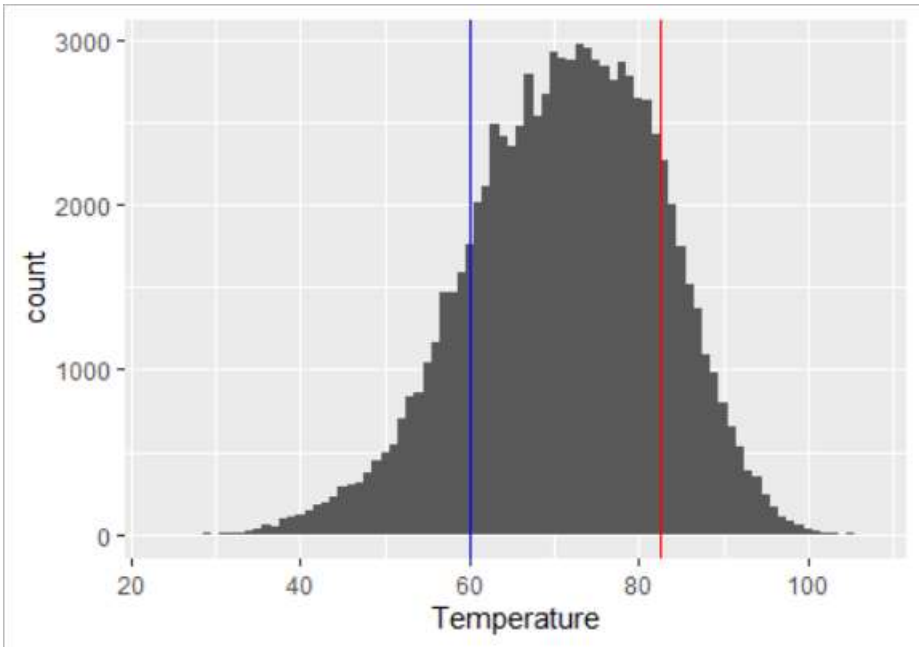

Figure 2. Histogram of Humidity

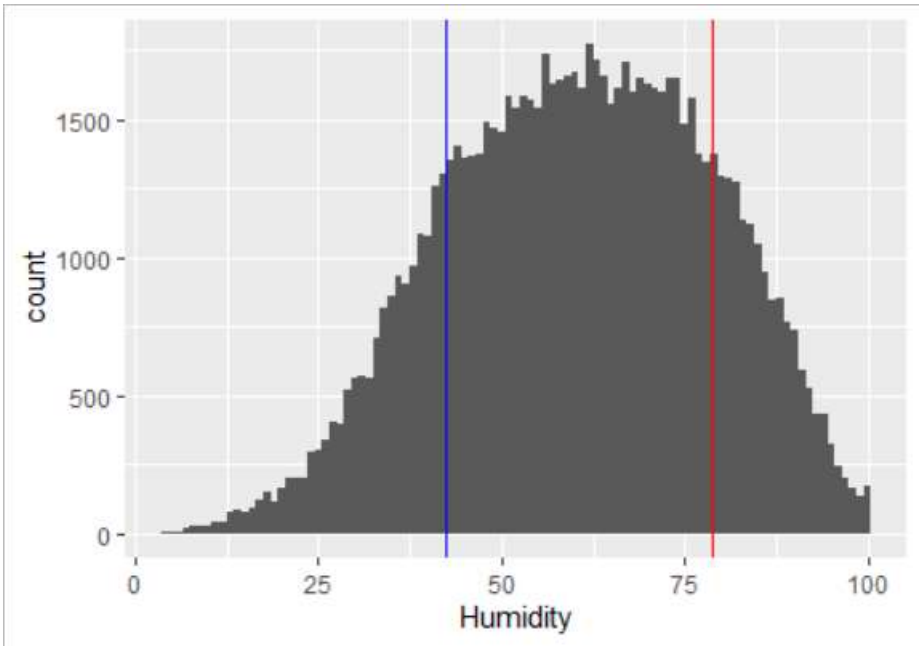

Figure 3. BIC Line Plot for Temperature - ERAd, WHIP, and XBH

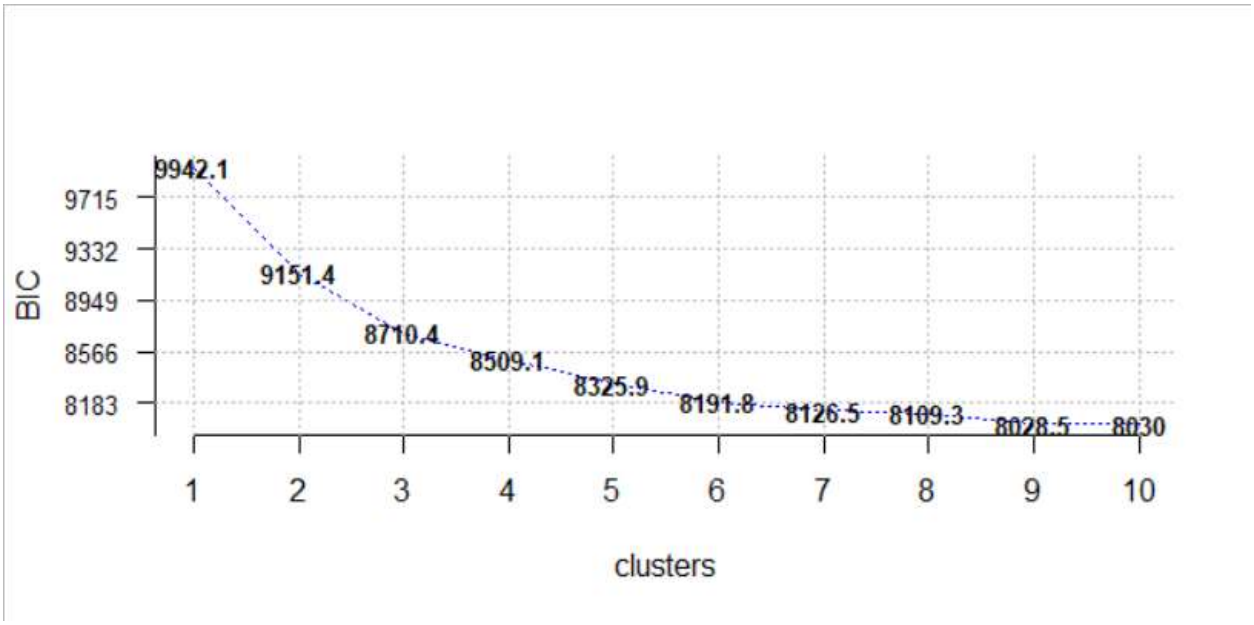

Figure 4. Scatter Plots for Temperature - ERAd, WHIP, and XBH

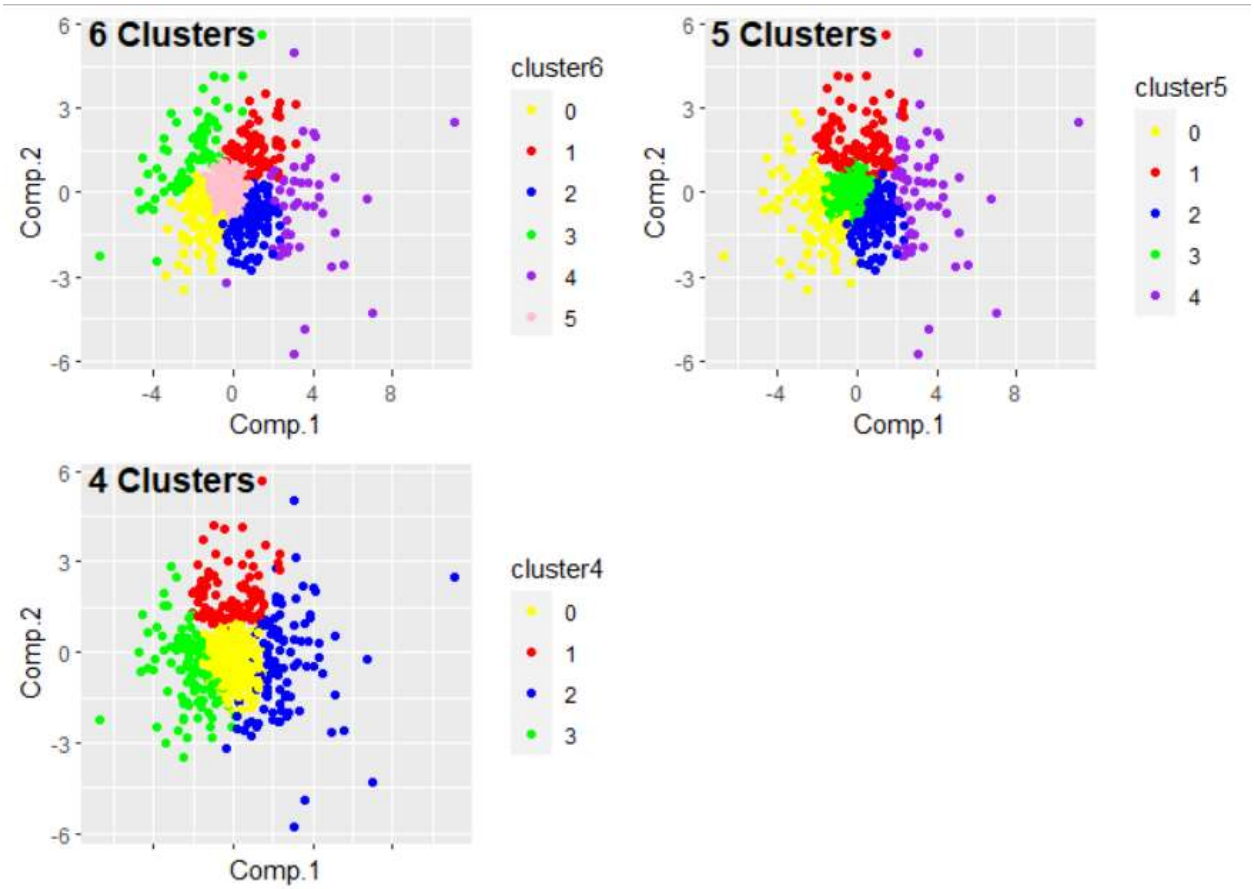

Table 7. 6 Cluster Solution for Temperature – ERAd, WHIP, and XBH

Cluster	Size	ERAd_high	ERAd_low	WHIP_high	WHIP_low	XBH_high	XBH_low
0	80	-1.13	-0.34	-0.19	-0.07	-0.52	-0.30
1	70	2.02	-0.36	0.28	-0.03	1.15	-0.21
2	122	0.34	0.88	0.05	0.15	0.22	0.45
3	70	-0.02	-1.90	-0.03	-0.28	0.09	-1.15
4	48	2.07	1.97	0.33	0.34	1.23	1.01
5	192	0.30	-0.30	0.04	-0.02	0.22	-0.24

Table 8. 5 Cluster Solution for Temperature – ERAd, WHIP, and XBH

Cluster	Size	ERAd high	ERAd low	WHIP high	WHIP low	XBH high	XBH low
0	99	-1.26	-0.83	-0.20	-0.12	-0.57	-0.63
1	102	1.42	-1.01	0.20	-0.15	0.86	-0.64
2	131	0.37	0.80	0.06	0.14	0.23	0.44
3	189	0.20	-0.31	0.02	-0.02	0.18	-0.22
4	61	2.29	1.59	0.35	0.28	1.27	0.85

Table 9. 4 Cluster Solution for Temperature – ERAd, WHIP, and XBH

Cluster	Size	ERAd high	ERAd low	WHIP high	WHIP low	XBH high	XBH low
0	282	0.27	0.04	0.04	0.03	0.18	-0.06
1	93	1.39	-1.11	0.19	-0.17	0.89	-0.70
2	102	1.67	1.37	0.25	0.24	0.93	0.81
3	105	-1.21	-0.84	-0.20	-0.13	-0.54	-0.59

Figure 5. BIC Line Plot for Temperature - Ground Ball, Fly Ball, and Line Drive

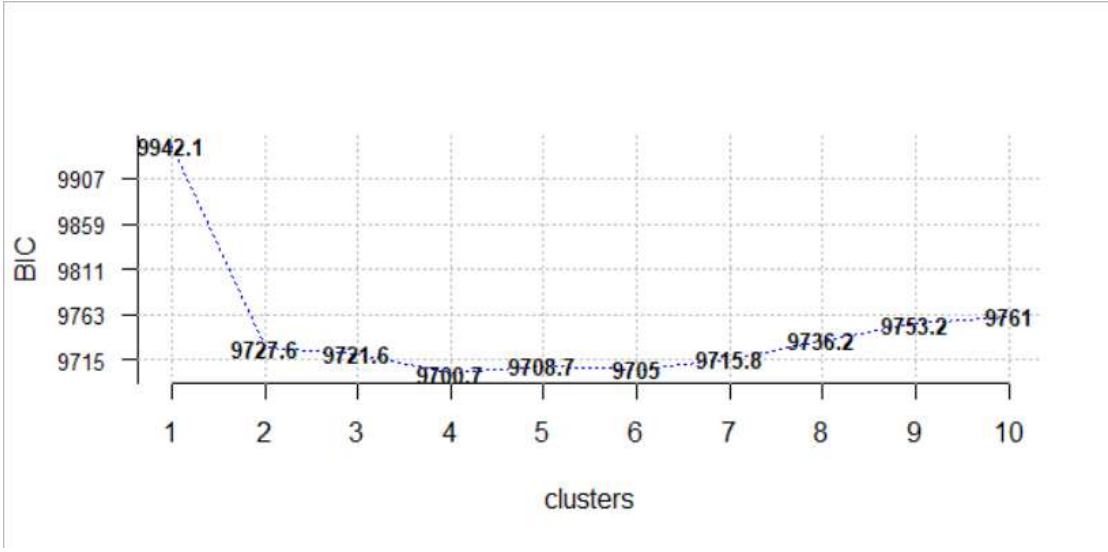

Figure 6. Scatter Plots for Temperature - Ground Ball, Fly Ball, and Line Drive

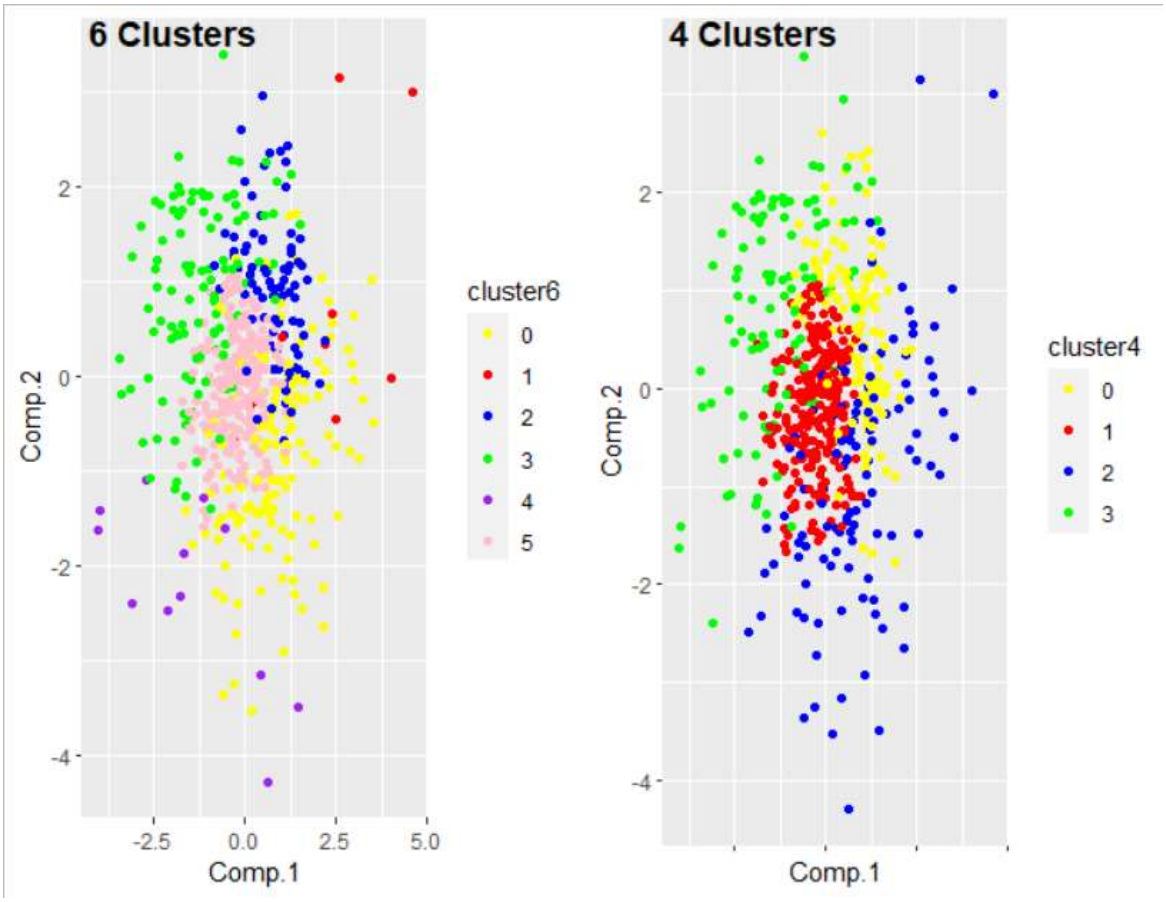

Table 10. 6 Cluster Solution for Temperature – Ground Ball, Fly Ball, and Line Drive

Cluster	Size	Grnd high	Grnd low	Fly high	Fly low	Line high	Line low
0	143	-0.27	0.02	-0.14	-0.46	0.70	0.97
1	9	2.89	3.68	0.62	0.19	1.24	0.74
2	82	1.21	0.61	-0.44	-0.52	0.36	-0.66
3	106	-0.16	-0.35	0.45	0.43	-0.71	-1.30
4	12	-0.37	-1.95	1.83	2.25	0.62	1.22
5	230	-0.19	-0.22	0.39	0.08	0.28	-0.22

Table 11. 4 Cluster Solution for Temperature – Ground Ball, Fly Ball, and Line Drive

Cluster	Size	Grnd high	Grnd low	Fly high	Fly low	Line high	Line low
0	124	0.74	0.47	-0.46	-0.61	0.36	-0.29
1	243	-0.22	-0.29	0.44	0.12	0.29	-0.19
2	115	0.12	0.24	0.07	-0.12	0.91	1.10
3	100	-0.29	-0.41	0.53	0.45	-0.81	-1.26

Figure 7. BIC Line Plot for Temperature – Walks, Strikeouts, and Home Runs

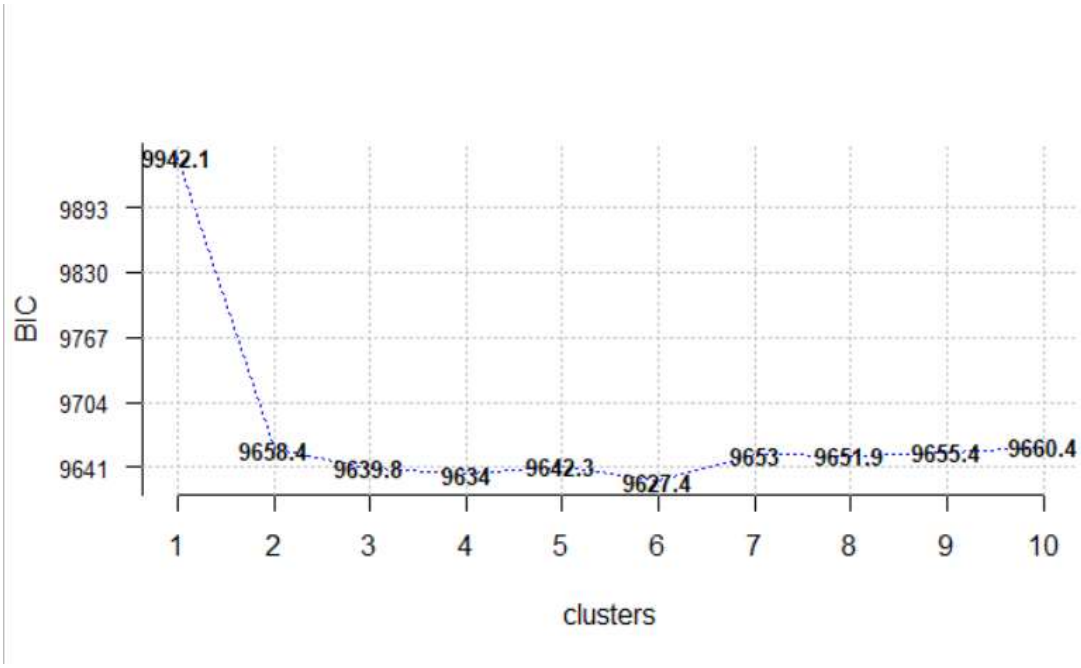

Figure 8. Scatter Plots for Temperature – Walks, Strikeouts, and Home Runs

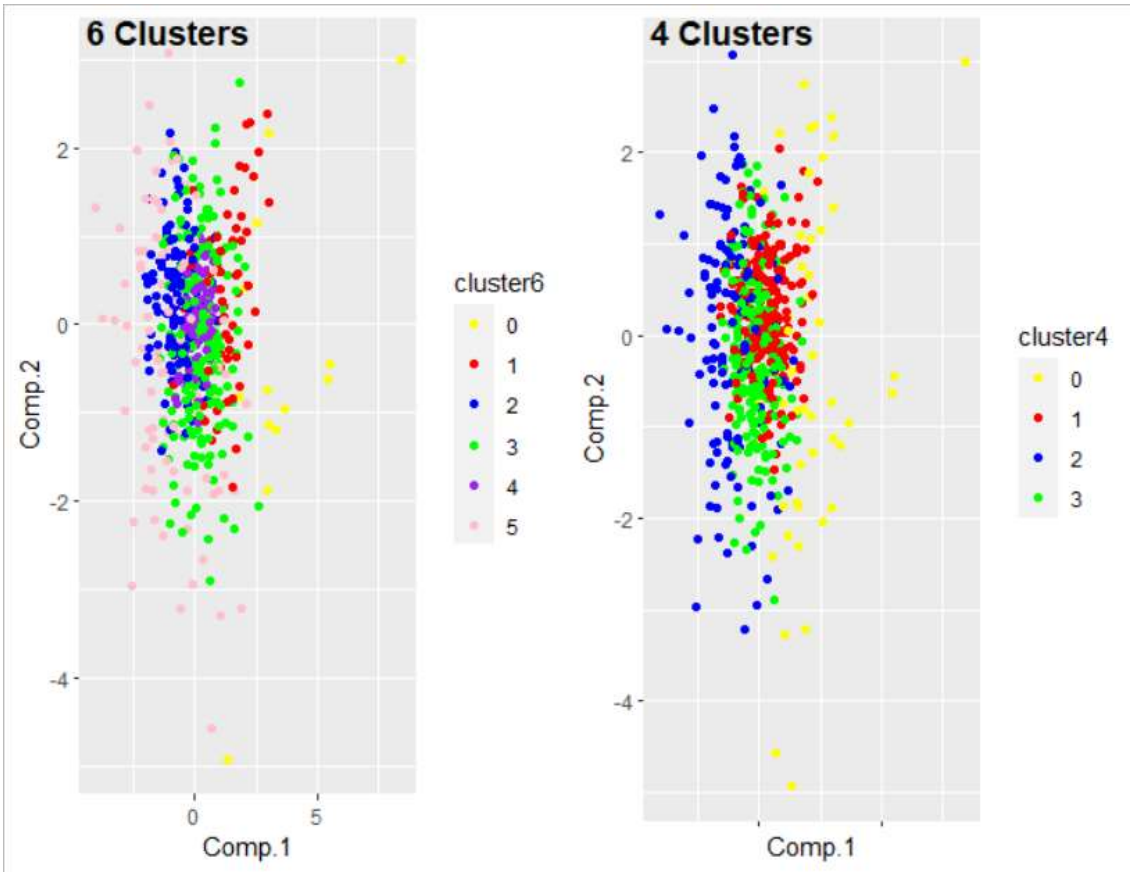

Table 12. 6 Cluster Solution for Temperature - Walks, Strikeouts, and Home Runs

Cluster	Size	HR high	HR low	KK high	KK low	BB high	BB low
0	13	0.95	0.89	-0.53	-0.16	1.67	1.01
1	64	0.75	0.05	-0.36	-0.42	0.39	0.42
2	120	-0.18	-0.31	-0.04	-0.23	-0.31	0.21
3	190	0.17	0.02	-0.09	0.27	0.00	0.40
4	98	0.22	-0.10	0.11	-0.07	0.07	0.31
5	97	0.11	-0.48	0.84	0.01	-0.18	0.09

Table 13. 4 Cluster Solution for Temperature – Walks, Strikeouts, and Home Runs

Cluster	Size	HR high	HR low	KK high	KK low	BB high	BB low
0	48	0.70	0.51	-0.15	-0.25	0.72	1.02
1	202	0.25	-0.11	-0.21	-0.27	0.16	0.33
2	135	0.08	-0.45	0.54	0.03	-0.40	-0.04
3	197	0.04	-0.08	0.10	0.26	-0.06	0.36

Figure 9. BIC Line Plot for Temperature – Ball Percentage

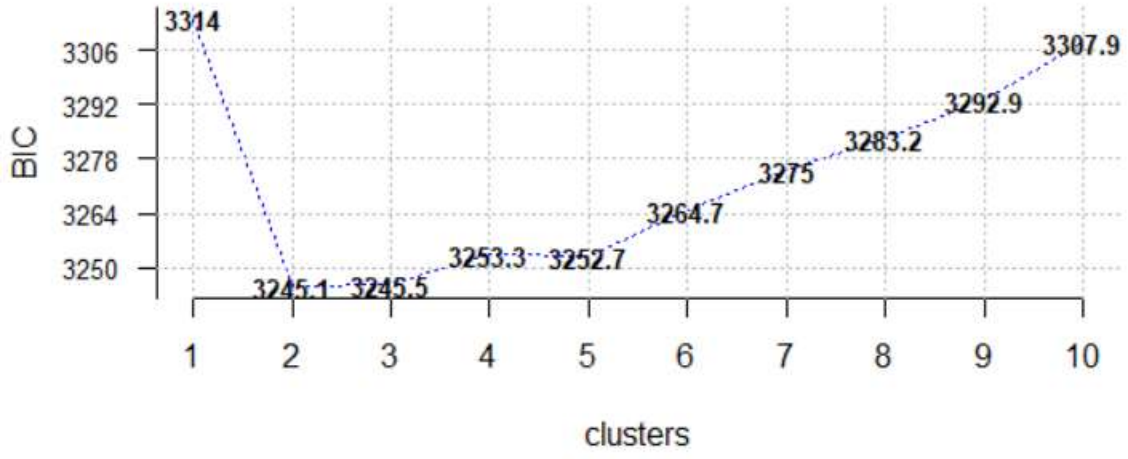

Figure 10. Scatter Plots for Temperature – Ball Percentage

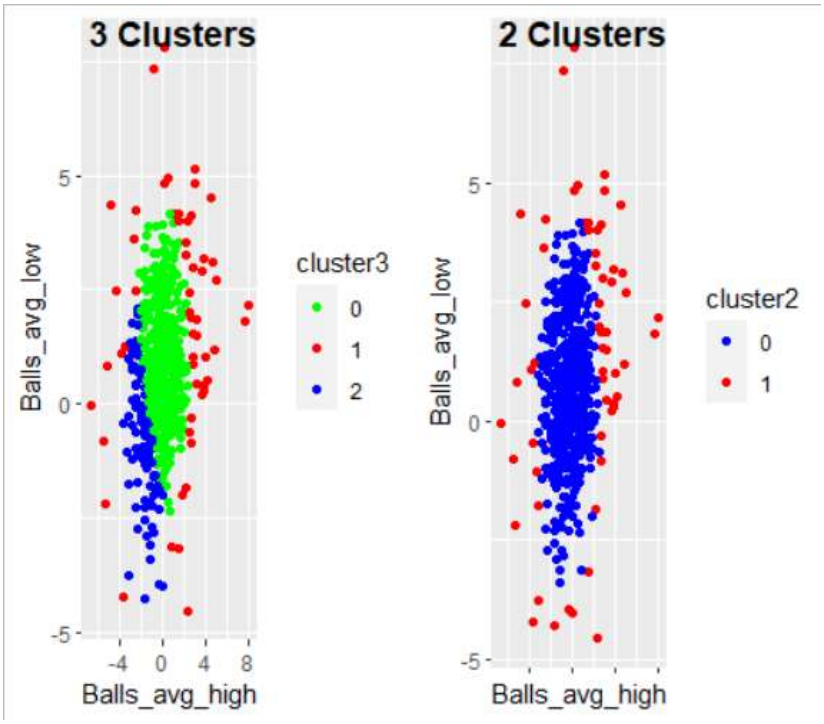

Table 14. 3 Cluster Solution for Temperature – Ball Percentage

Cluster	Size	Balls high	Balls low
0	419	0.03	1.01
1	57	1.37	1.71
2	106	-1.74	-0.77

Table 15. 2 Cluster Solution for Temperature – Ball Percentage

Cluster	Size	Balls high	Balls low
0	523	-0.29	0.68
1	59	0.99	1.41

Figure 11. BIC Line Plot for Humidity - ERAd, WHIP, and XBH

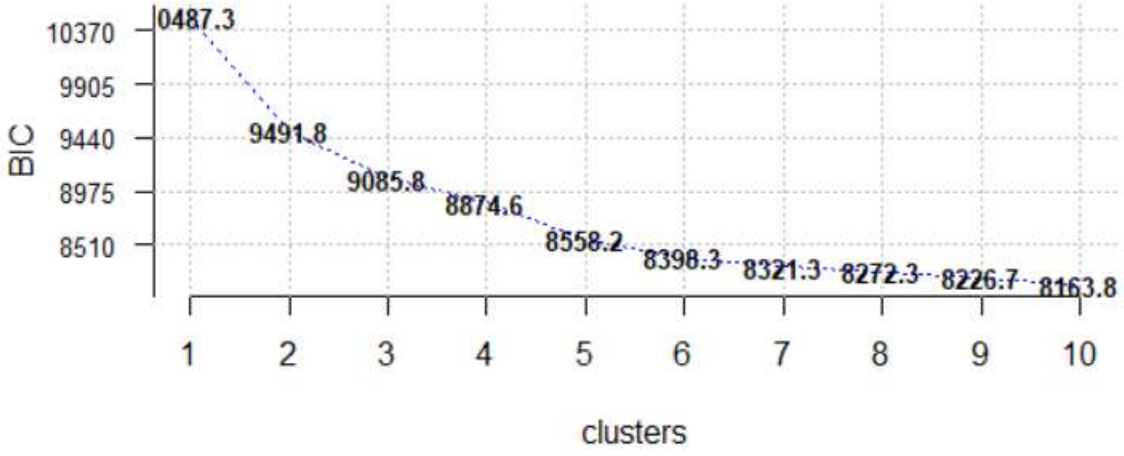

Figure 12. Scatter Plots for Humidity - ERAd, WHIP, and XBH

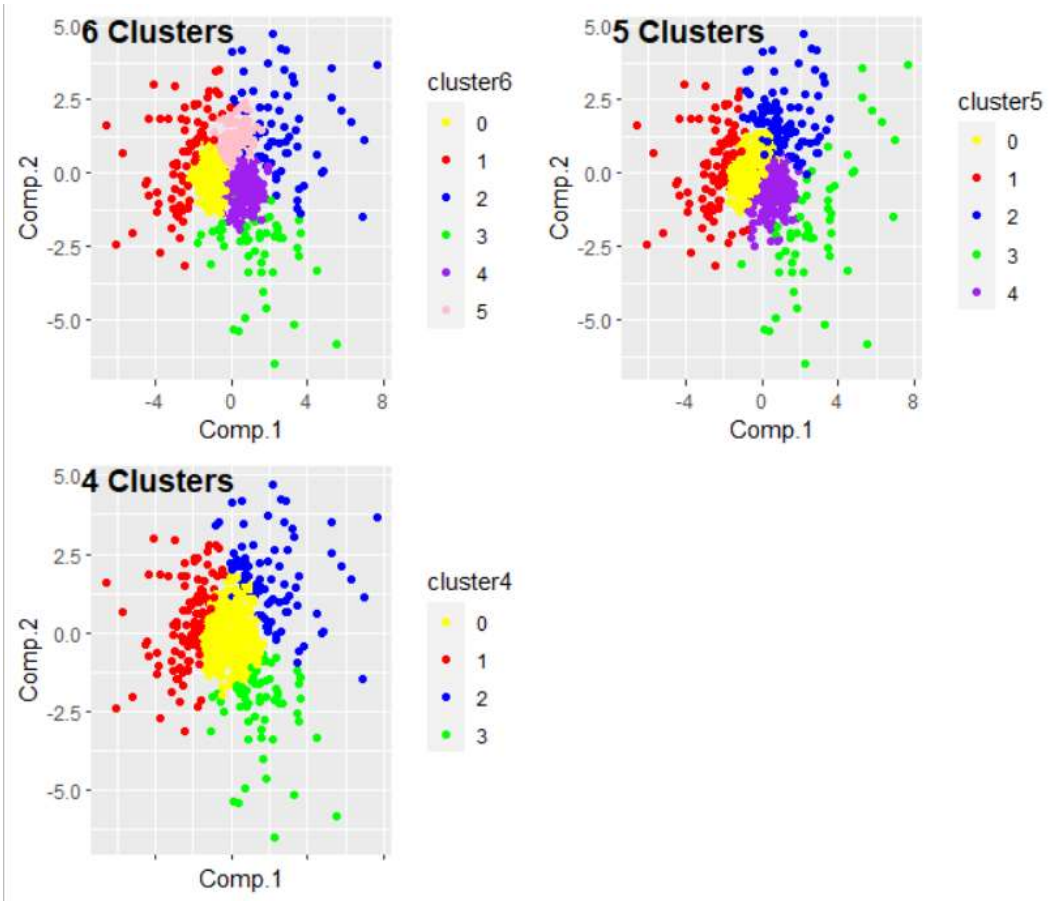

Table 16. 6 Cluster Solution for Humidity – ERAd, WHIP, and XBH

Cluster	Size	ERAd_high	ERAd_low	WHIP_high	WHIP_low	XBH_high	XBH_low
0	148	-0.69	-0.25	-0.10	-0.02	-0.39	-0.10
1	80	-1.07	-1.61	-0.15	-0.02	-0.59	-0.91
2	67	2.32	1.18	0.36	0.17	1.09	0.62
3	54	-0.76	2.59	-0.09	0.40	-0.64	1.54
4	152	-0.75	0.96	0.00	0.15	0.01	0.61
5	113	0.58	-0.19	0.11	-0.03	0.39	-0.08

Table 17. 5 Cluster Solution for Humidity – ERAd, WHIP, and XBH

Cluster	Size	ERAd high	ERAd low	WHIP high	WHIP low	XBH high	XBH low
0	205	-0.24	-0.13	-0.03	-0.01	-0.15	-0.07
1	98	-1.26	-1.40	-0.17	-0.17	-0.68	-0.74
2	97	1.57	0.00	0.25	-0.01	0.88	0.08
3	54	0.60	3.30	0.13	0.49	0.00	1.65
4	160	-0.20	1.02	-0.02	0.16	-0.05	0.69

Table 18. 4 Cluster Solution for Humidity – ERAd, WHIP, and XBH

Cluster	Size	ERAd high	ERAd low	WHIP high	WHIP low	XBH high	XBH low
0	350	-0.13	0.28	-0.01	0.05	-0.04	0.18
1	110	-1.10	-1.38	-0.14	-0.18	-0.62	-0.76
2	87	2.10	0.71	0.32	0.09	1.07	0.42
3	67	-0.53	2.52	-0.05	0.39	-0.48	1.52

Figure 13. BIC Line Plot for Humidity - Ground Ball, Fly Ball, and Line Drive

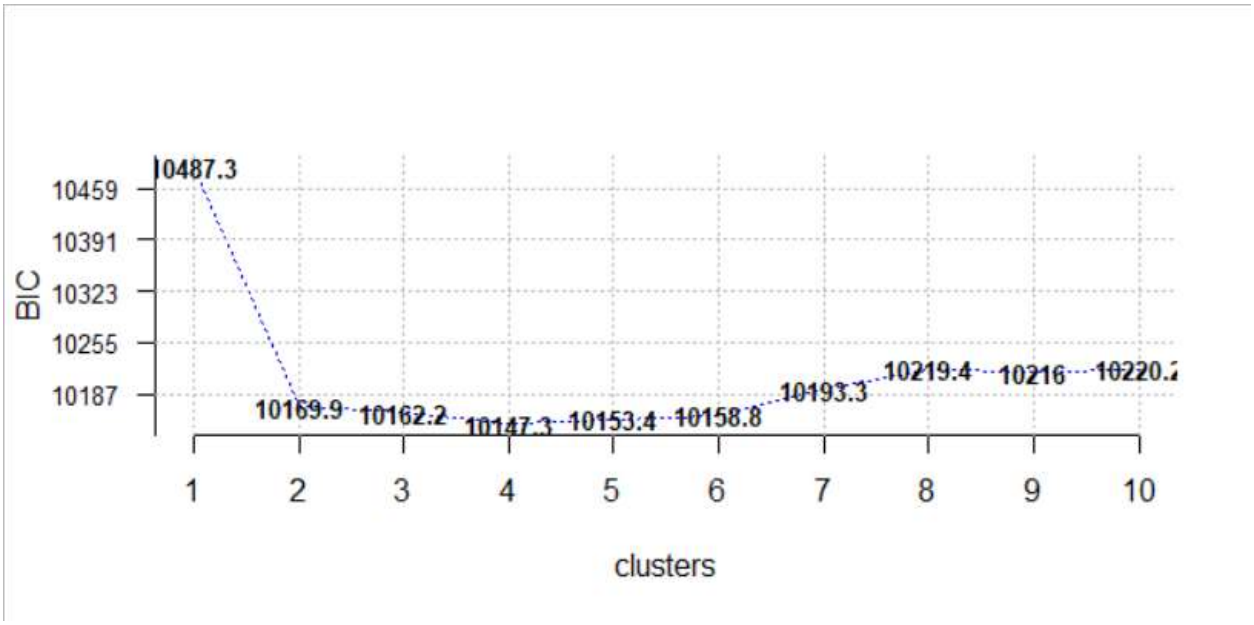

Figure 14. Scatter Plots for Humidity - Ground Ball, Fly Ball, and Line Drive

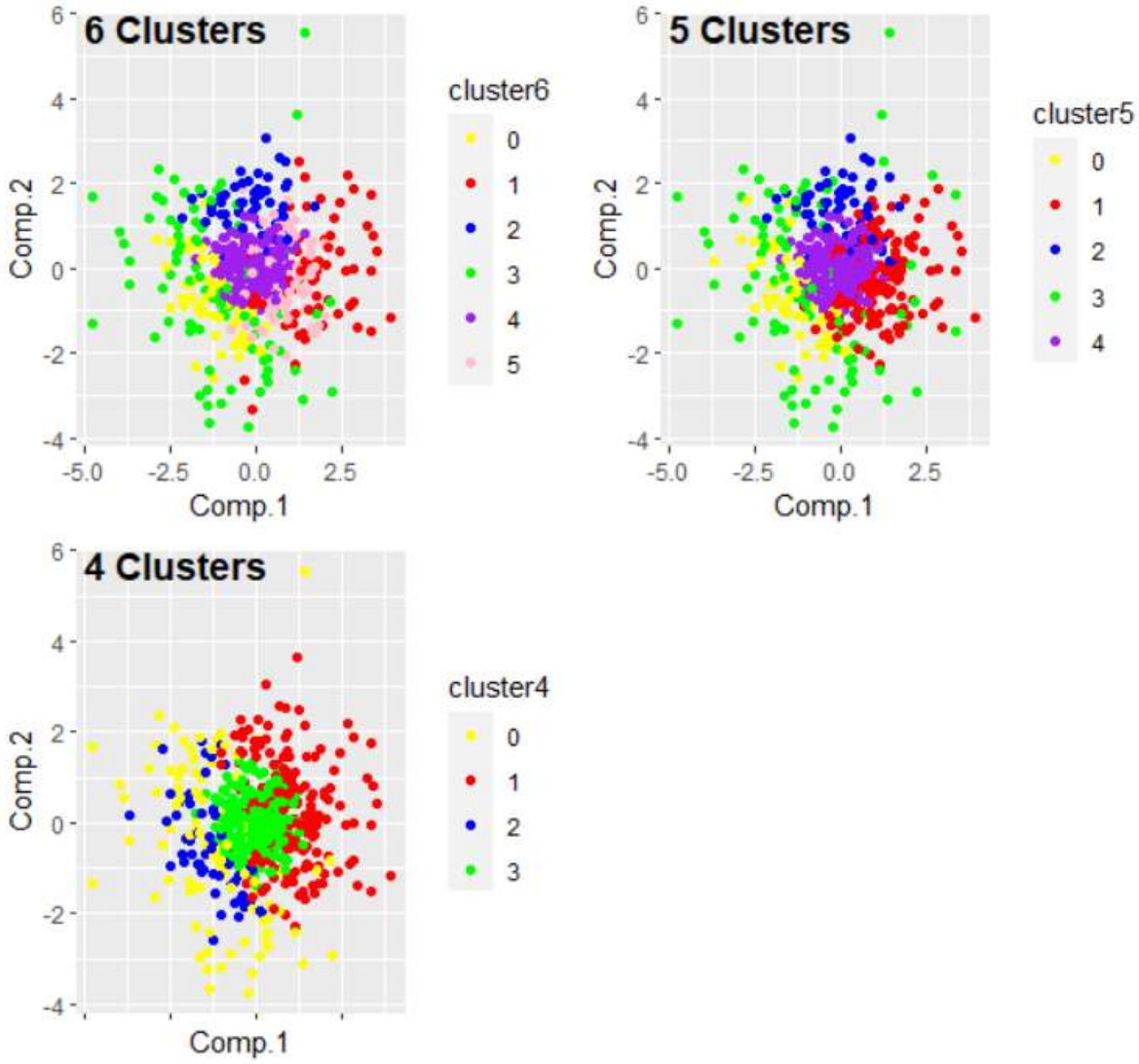

Table 19. 6 Cluster Solution for Humidity – Ground Ball, Fly Ball, and Line Drive

Cluster	Size	Grnd high	Grnd low	Fly high	Fly low	Line high	Line low
0	56	-1.28	-0.45	1.13	0.58	0.03	0.08
1	76	0.55	1.32	-0.59	-1.20	0.26	0.40
2	55	0.06	0.31	-0.49	0.24	-0.82	-1.61
3	89	-0.71	-1.25	0.51	0.93	-0.05	0.20
4	235	-0.11	0.02	-0.01	0.13	-0.25	-0.05
5	103	0.83	-0.22	-0.60	0.28	0.28	0.65

Table 20. 5 Cluster Solution for Humidity – Ground Ball, Fly Ball, and Line Drive

Cluster	Size	Grnd high	Grnd low	Fly high	Fly low	Line high	Line low
0	56	-1.43	-0.51	1.09	0.72	0.05	-0.12
1	143	0.69	0.56	-0.63	-0.35	0.33	0.69
2	63	0.13	0.42	-0.51	-0.04	-0.62	-1.48
3	102	-0.46	-0.93	0.41	0.61	-0.08	0.20
4	250	-0.04	-0.08	-0.04	0.18	-0.24	0.00

Table 21. 4 Cluster Solution for Humidity – Ground Ball, Fly Ball, and Line Drive

Cluster	Size	Grnd high	Grnd low	Fly high	Fly low	Line high	Line low
0	99	-0.77	-1.06	0.57	0.90	-0.06	0.18
1	181	0.72	0.58	-0.69	-0.51	0.10	0.07
2	59	-1.34	-0.47	1.02	0.68	-0.01	-0.33
3	275	-0.02	-0.03	-0.08	0.22	-0.19	0.03

Figure 15. BIC Line Plot for Humidity – Walks, Strikeouts, and Home Runs

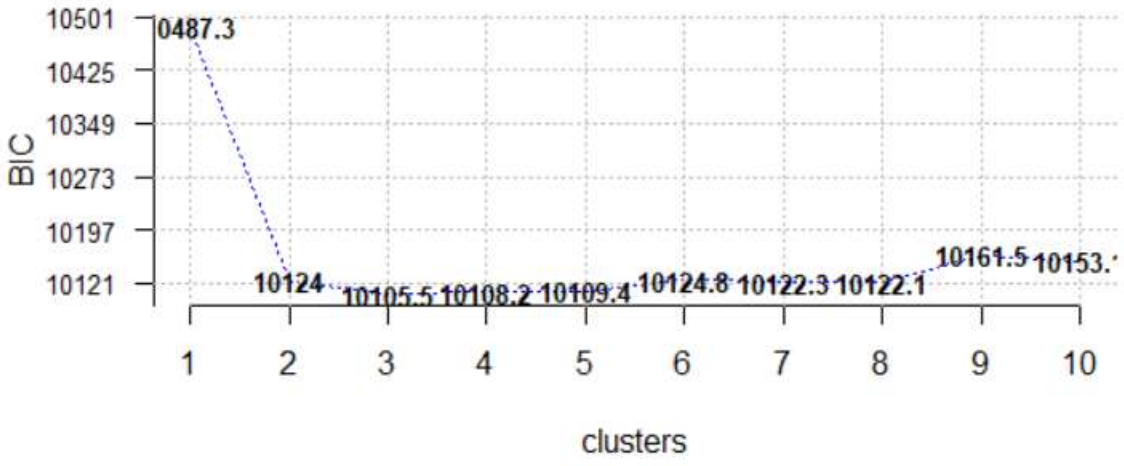

Figure 16. Scatter Plots for Humidity – Walks, Strikeouts, and Home Runs

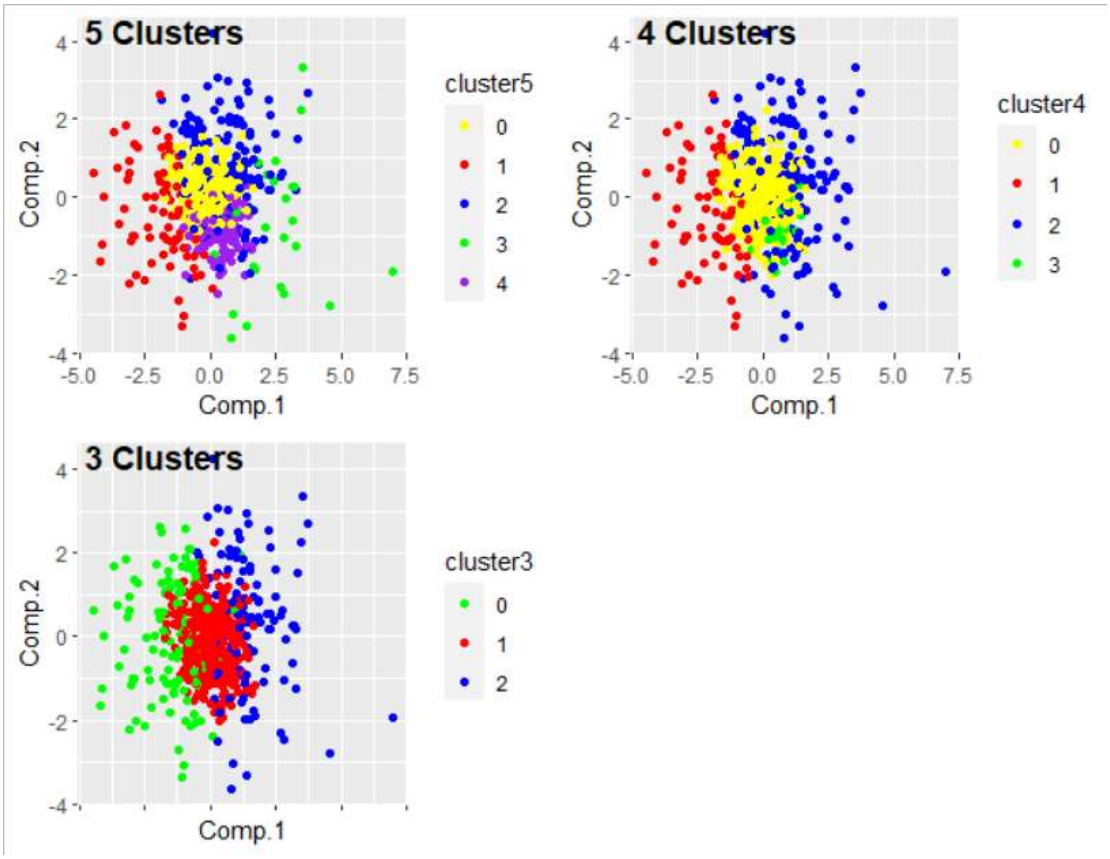

Table 22. 5 Cluster Solution for Humidity - Walks, Strikeouts, and Home Runs

Cluster	Size	HR high	HR low	KK high	KK low	BB high	BB low
0	241	-0.04	0.04	0.26	0.08	0.03	0.14
1	83	-0.58	-0.45	-0.06	0.41	-0.47	-0.49
2	150	-0.01	0.28	0.34	0.17	0.52	0.64
3	24	0.66	0.64	-1.37	-0.63	0.97	0.44
4	116	0.00	0.13	-0.79	-0.39	-0.07	0.17

Table 23. 4 Cluster Solution for Humidity - Walks, Strikeouts, and Home Runs

Cluster	Size	HR high	HR low	KK high	KK low	BB high	BB low
0	348	-0.05	0.04	0.02	0.04	0.00	0.18
1	73	-0.62	-0.46	0.01	0.46	-0.53	-0.50
2	160	0.10	0.34	0.00	-0.06	0.60	0.60
3	33	0.07	0.33	-0.81	-0.52	0.14	-0.03

Table 24. 3 Cluster Solution for Humidity - Walks, Strikeouts, and Home Runs

Cluster	Size	HR high	HR low	KK high	KK low	BB high	BB low
0	115	-0.49	-0.32	0.15	0.51	-0.33	-0.25
1	381	-0.03	0.07	-0.07	-0.04	0.03	0.16
2	118	0.20	0.48	-0.05	-0.21	0.74	0.74

Figure 17. BIC Line Plot for Humidity – Ball Percentage

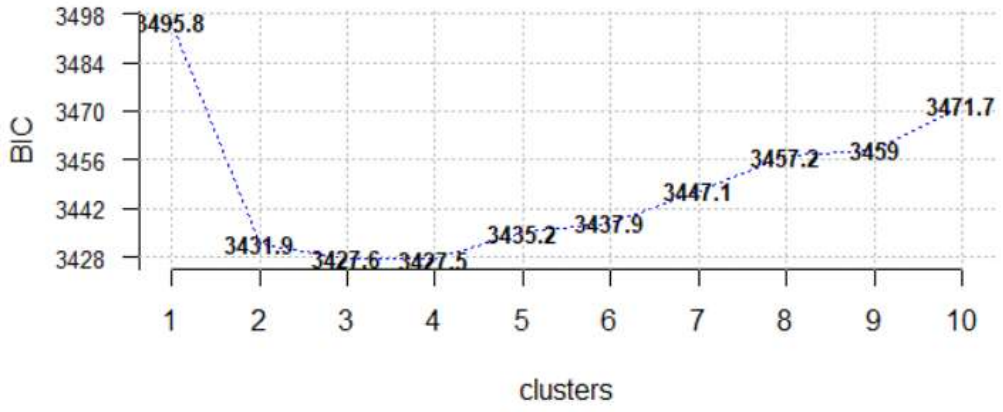

Figure 18. Scatter Plots for Humidity – Ball Percentage

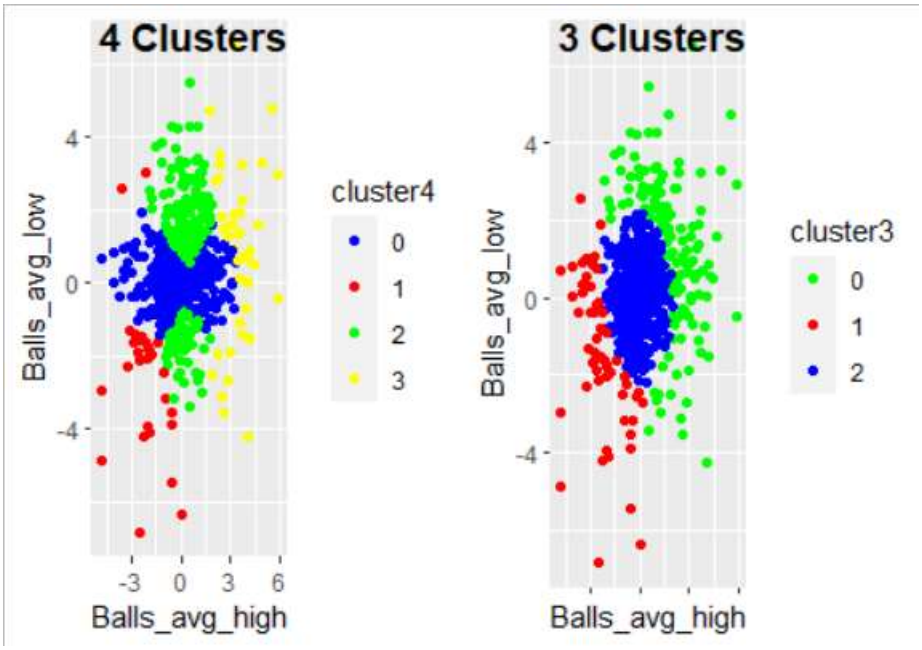

Table 25. 4 Cluster Solution for Humidity – Ball Percentage

Cluster	Size	Balls high	Balls low
0	332	-0.10	0.07
1	30	-2.19	-2.47
2	212	0.27	0.79
3	40	3.47	0.84

Table 26. 3 Cluster Solution for Humidity – Ball Percentage

Cluster	Size	Balls high	Balls low
0	134	1.86	1.31
1	63	-2.36	-1.40
2	417	-0.01	0.15

Figure 19. BIC Line Plot for Temperature & Humidity - ERAd, WHIP, and XBH

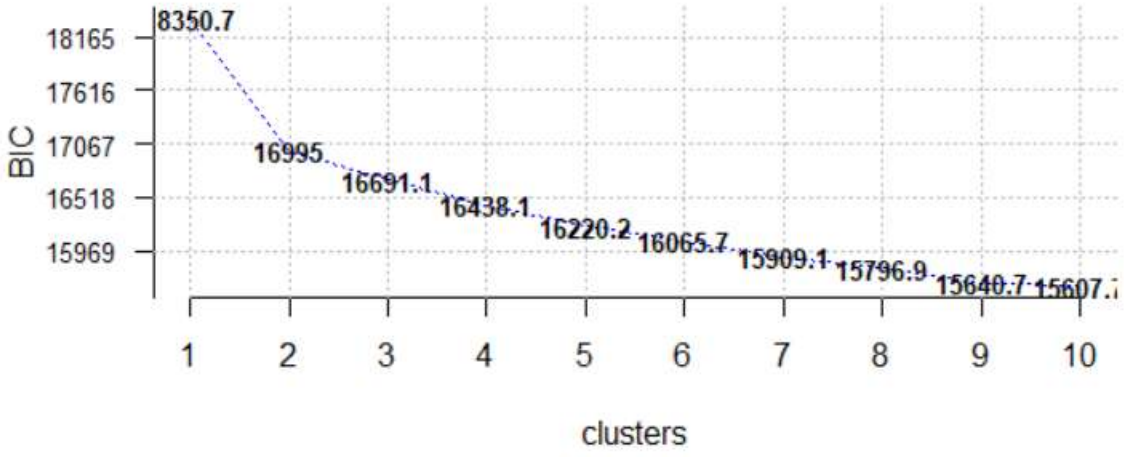

Figure 20. Scatter Plots for Temperature & Humidity - ERAd, WHIP, and XBH

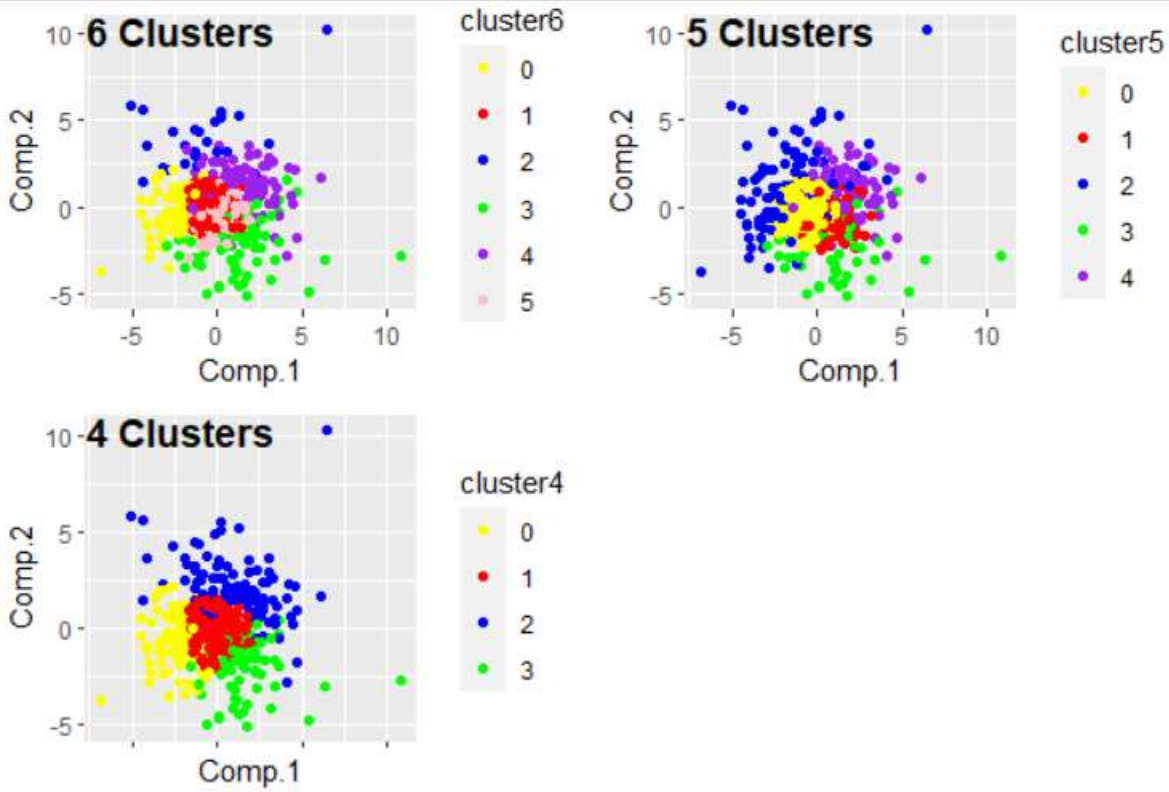

Table 27. 6 Cluster Solution for Temperature & Humidity – ERAd, WHIP, and XBH

Temperature							
Cluster	Size	ERAd_high	ERAd_low	WHIP_high	WHIP_low	XBH_high	XBH_low
0	72	-0.56	-0.81	-0.11	-0.14	-0.31	-0.70
1	171	0.27	-0.02	0.04	0.02	0.18	-0.05
2	25	1.92	0.25	0.25	0.06	1.09	0.22
3	69	-0.01	0.26	-0.01	0.05	0.07	0.19
4	93	1.52	0.45	0.23	0.08	1.03	0.19
5	107	0.28	-0.39	0.03	-0.04	0.19	-0.32
Humidity							
Cluster	Size	ERAd_high	ERAd_low	WHIP_high	WHIP_low	XBH_high	XBH_low
0	72	-0.53	-1.02	-0.14	-0.30	-0.30	-0.56
1	171	-0.06	-0.08	0.00	-0.01	-0.01	-0.01
2	25	-2.22	-0.98	-0.14	-1.41	-1.41	-0.70
3	69	1.71	0.87	0.15	0.87	0.87	0.58
4	93	-0.46	1.17	0.19	-0.43	-0.43	0.74
5	107	-0.11	1.07	0.16	0.16	-0.01	0.69

Table 28. 5 Cluster Solution for Temperature & Humidity – ERAd, WHIP, and XBH

Temperature							
Cluster	Size	ERAd_high	ERAd_low	WHIP_high	WHIP_low	XBH_high	XBH_low
0	206	0.14	-0.44	0.01	-0.05	0.14	-0.35
1	109	0.27	0.62	0.04	0.13	0.19	0.38
2	96	0.44	-0.20	0.05	-0.05	0.25	-0.29
3	41	-0.08	-0.25	-0.03	-0.03	-0.06	-0.18
4	85	1.50	0.20	0.22	0.06	1.01	0.11
Humidity							
Cluster	Size	ERAd_high	ERAd_low	WHIP_high	WHIP_low	XBH_high	XBH_low
0	206	-0.27	0.32	-0.03	0.05	-0.13	0.21
1	109	0.56	0.18	0.09	0.05	0.35	0.14
2	96	-0.89	-1.15	-0.13	-0.15	-0.55	-0.64
3	41	1.96	1.36	0.30	0.21	0.95	0.82
4	85	-0.51	1.67	0.05	0.26	-0.44	1.08

Table 29. 4 Cluster Solution for Temperature & Humidity – ERAd, WHIP, and XBH

Temperature							
Cluster	Size	ERAd_high	ERAd_low	WHIP_high	WHIP_low	XBH_high	XBH_low
0	89	-0.65	-0.92	-0.13	-0.15	-0.38	-0.77
1	264	0.32	-0.13	0.04	0.00	0.21	-0.14
2	112	1.67	0.44	0.25	0.09	1.10	0.22
3	72	0.17	0.39	0.01	0.07	0.15	0.30
Humidity							
Cluster	Size	ERAd_high	ERAd_low	WHIP_high	WHIP_low	XBH_high	XBH_low
0	89	-0.42	-0.66	-0.06	-0.08	-0.24	-0.37
1	264	-0.11	0.34	0.00	0.06	-0.03	0.24
2	112	-0.84	0.68	-0.12	0.11	-0.67	0.44
3	72	1.66	0.90	0.25	0.15	0.90	0.60

Figure 21. BIC Line Plot for Temperature & Humidity - Ground Ball, Fly Ball, and Line Drive

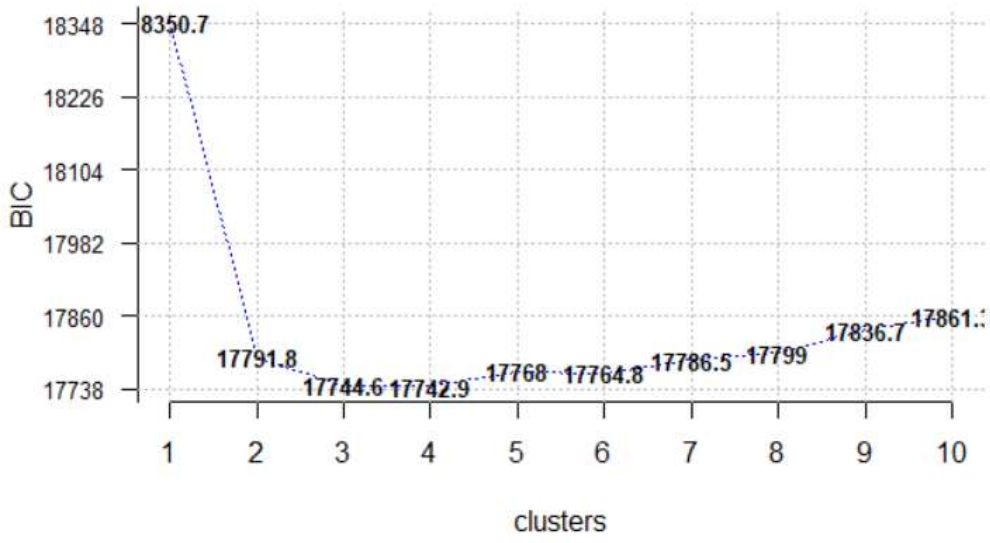

Figure 22. Scatter Plots for Temperature & Humidity - Ground Ball, Fly Ball, and Line Drive

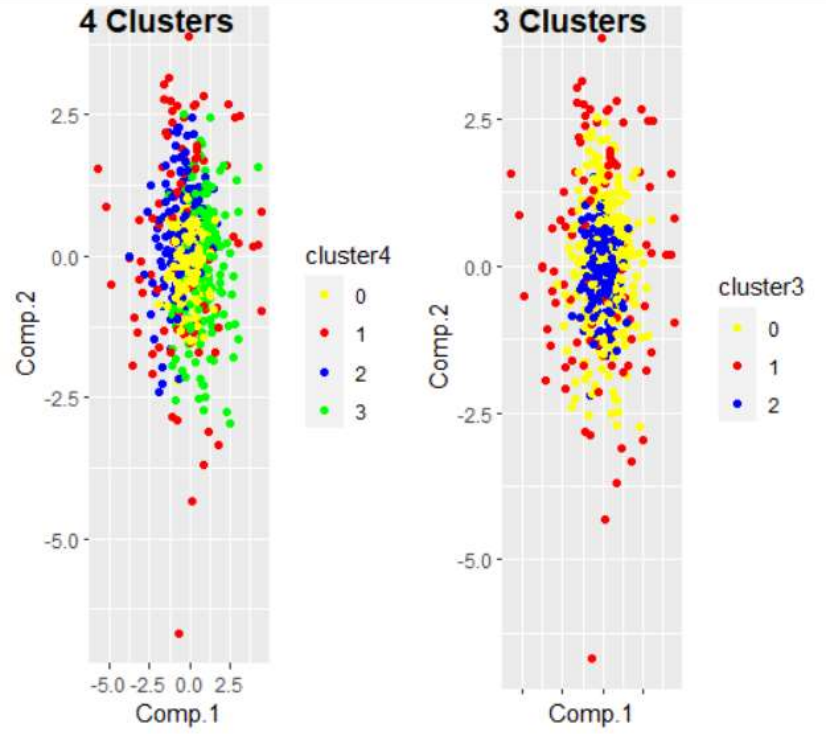

Table 30. 4 Cluster Solution for Temperature & Humidity - Ground Ball, Fly Ball, and Line Drive

Temperature							
Cluster	Size	Grnd high	Grnd low	Fly high	Fly low	Line high	Line low
0	163	-0.19	-0.04	0.46	-0.01	0.20	-0.22
1	96	0.21	0.10	0.61	0.07	0.02	0.03
2	127	0.03	-0.46	0.18	0.50	0.18	-0.16
3	151	0.22	0.29	-0.27	-0.49	0.38	-0.26
Humidity							
Cluster	Size	Grnd high	Grnd low	Fly high	Fly low	Line high	Line low
0	163	0.11	0.03	-0.07	0.21	-0.17	0.16
1	96	-0.23	-0.39	-0.14	0.59	-0.12	0.22
2	127	-0.59	-0.29	0.38	0.21	0.01	-0.15
3	151	0.38	0.39	-0.45	-0.24	-0.13	0.05

Table 31. 3 Cluster Solution for Temperature & Humidity Ground Ball, Fly Ball, and Line Drive

Temperature							
Cluster	Size	Grnd high	Grnd low	Fly high	Fly low	Line high	Line low
0	264	0.09	-0.05	-0.02	-0.05	0.28	-0.15
1	101	0.28	0.15	0.43	0.04	0.08	-0.13
2	172	-0.16	-0.08	0.44	0.02	0.20	-0.22
Humidity							
Cluster	Size	Grnd high	Grnd low	Fly high	Fly low	Line high	Line low
0	264	-0.06	0.12	-0.07	-0.03	-0.07	-0.07
1	101	-0.16	-0.37	-0.22	0.47	-0.10	0.27
2	172	0.05	-0.03	-0.03	0.24	-0.16	0.15

Figure 23. BIC Line Plot for Temperature & Humidity – Walks, Strikeouts, and Home Runs

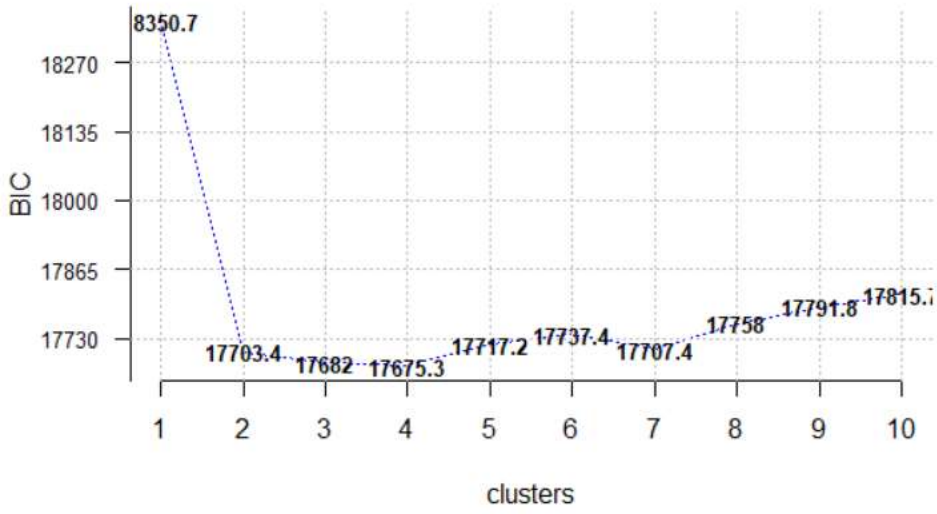

Figure 24. Scatter Plots for Temperature & Humidity – Walks, Strikeouts, and Home Runs

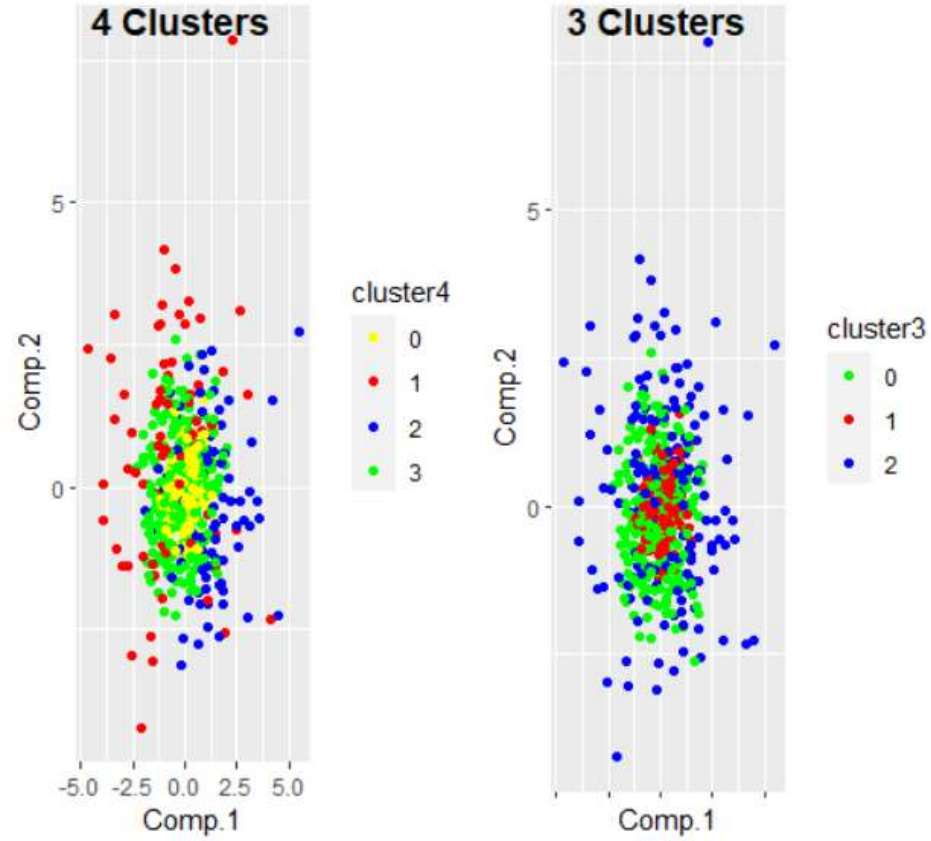

Table 32. 4 Cluster Solution for Temperature & Humidity - Walks, Strikeouts, and Home Runs

Temperature							
Cluster	Size	HR high	HR low	KK high	KK low	BB high	BB low
0	130	0.18	-0.12	0.13	-0.10	0.04	0.31
1	74	0.28	0.02	0.19	-0.24	0.16	0.39
2	86	0.36	-0.33	0.16	-0.19	-0.12	-0.07
3	247	0.06	-0.11	-0.11	0.10	-0.06	0.39
Humidity							
Cluster	Size	HR high	HR low	KK high	KK low	BB high	BB low
0	130	-0.08	0.13	-0.21	0.09	-0.02	0.06
1	74	0.07	0.30	-0.15	0.06	0.44	0.52
2	86	-0.45	0.01	-0.08	0.25	-0.36	0.08
3	247	0.00	0.03	0.13	-0.07	0.13	0.24

Table 33. 3 Cluster Solution for Temperature & Humidity - Walks, Strikeouts, and Home Runs

Temperature							
Cluster	Size	HR high	HR low	KK high	KK low	BB high	BB low
0	254	0.04	-0.13	-0.09	0.07	-0.06	0.32
1	122	0.21	-0.12	0.12	-0.10	0.06	0.30
2	161	0.34	-0.13	0.16	-0.17	0.00	0.26
Humidity							
Cluster	Size	HR high	HR low	KK high	KK low	BB high	BB low
0	254	0.00	0.03	0.09	-0.05	0.08	0.24
1	122	-0.08	0.14	-0.20	0.09	-0.03	0.06
2	161	-0.21	0.15	-0.07	0.15	0.08	0.27

Figure 25. BIC Line Plot for Temperature & Humidity – Ball Percentage

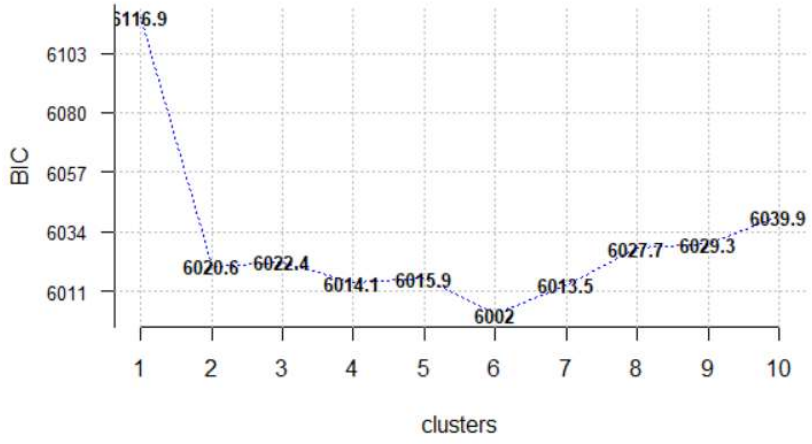

Figure 26. Scatter Plots for Temperature & Humidity – Ball Percentage

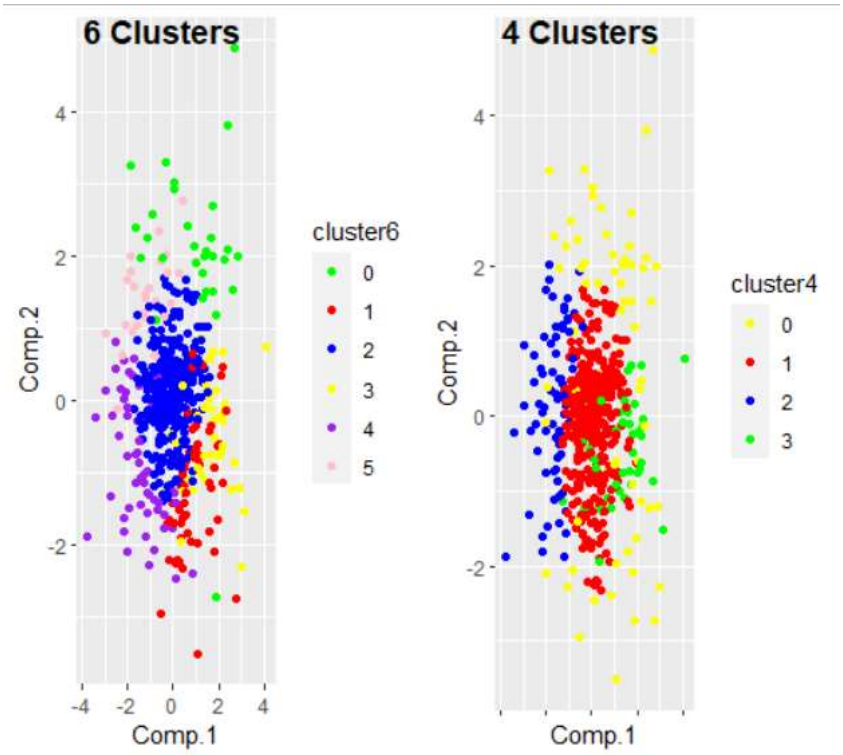

Table 34. 6 Cluster Solution for Temperature & Humidity – Ball Percentage

Temperature			
Cluster	Size	Balls high	Balls Low
0	30	2.70	2.47
1	60	-0.52	1.22
2	316	-0.06	0.77
3	44	-0.24	1.51
4	60	-1.82	-0.72
5	27	0.20	0.54
Humidity			
Cluster	Size	Balls high	Balls Low
0	30	-0.71	-0.12
1	60	2.32	0.71
2	316	-0.01	0.11
3	44	0.51	3.04
4	60	-0.14	-0.61
5	27	-2.67	-0.30

Table 35. 4 Cluster Solution for Temperature & Humidity – Ball Percentage

Temperature			
Cluster	Size	Balls high	Balls Low
0	60	0.86	2.04
1	352	-0.18	0.84
2	76	-0.91	-0.80
3	40	-0.24	1.32
Humidity			
Cluster	Size	Balls high	Balls Low
0	60	0.44	0.18
1	352	0.32	0.26
2	76	-1.32	-0.75
3	40	0.42	2.89