



Big Data Analysis for the Relationship between a Catcher's Batting Performance and Average Daily Temperature in Baseball Games

***Semin Kim**, *Adjunct Professor, Dept. of Computer Education, Jeonju National University of Education, 50 Seohakro Wansangu Jeonju, 55101, Republic of Korea*

Abstract. Background/Objectives: This study utilizes big data analysis to investigate whether the average daily temperature affects the batting performance of catchers.

Methods/Statistical analysis: The targets of the study are the batting records, by date, of catchers who participated in the 2019 KBO League, and the average daily temperature in the region where baseball games are played. To this end, data was collected from the official website of the KBO League and the Weather Data Open Portal. The data suitable for this study was then pre-processed, and then output as visualization data for scouting reports through the analysis.

Findings: As a result of this study, a higher average daily temperature was found to affect the batting performance of catchers. However, above 15 degrees, both batting power and hitting productivity began to decrease, and above 20 degrees, hitting accuracy began to decrease. The reason for this is that a player in the position of catcher has a lot of physical strength due to a lot of equipment and hard play. Through this study, we can see that the myths and stories told in the baseball field can be proven through data. Through the data and implications derived from this study, it is expected that the scouting reports will help the players, coaching staff, and the front of the baseball club's players to operate the games and the season.

Improvements/Applications: In less than thirty words, this research makes it possible to write about the results of the improved research and other applications.

Keywords: Big Data Analysis, KBO League, Batting Performance, Average Daily Temperature, SEMMA Model.

Received: 09.12.2020

Accepted: 10.01.2021

Published: 03.02.2021

INTRODUCTION

Recently in the field of sports, efforts are being made to improve players' performance by utilizing data and analyzing images in line with the development of information, communication, and media technology. The cumulative records and ratio records that occur during the game are called classical records or primary records, and in baseball, these records were defined as secondary records by using scientific statistical methods. Sabermetrics is being actively used not only in the MLB (Major League Baseball) but also in the KBO (Korea Baseball Organization) League. In addition, game videos have been used to analyze and utilize the trajectory, number of spins, and the angle and distance of balls struck by the batter [1-4]. Baseball games have more kinds of records than other sports, more players participating in the games, more games are played over a longer period of time, and many games are played for a long time. Therefore, baseball games are the most data-driven sport, and the research and industry used to analyze them are also active [5].

Baseball is a representative outdoor sport. Dome stadiums are being built in recent years, but they account for only a small number of stadiums, and most games are still being played in outdoor stadiums. Therefore, baseball cannot ignore the variables given by the natural environment. For example, small variables may affect the results due to wind, rainfall, temperature, humidity, and so on. For example, on a humid day, a pitcher who throws a breaking ball is known to be more advantageous against the batter. This is because in high humidity, the characteristics of baseball made of leather can absorb moisture and cause friction between the baseball and pitcher's fingers, so the number of ball rotations can increase whenever a breaking ball is used. Fastballs also have a high number of ball rotations, which in turn boost the ball's pitch, helping the team win with powerful pitches [6].

Additionally, in baseball, each position has its own characteristics, so in Sabermetrics, each position is weighted and data is reprocessed. The catcher is recognized as the toughest position in baseball. They wear the most and heaviest equipment among all players, so they will struggle even more in the hot and humid summer of the Korean Peninsula [7-10]. Therefore, in this study, we try to analyze the difference in the batting performance of catchers according to the daily average temperature. For this purpose, data

provided by the official website of KBO League was collected, processed, and visualized for analysis.

2.1. The subject of study

The subjects of this study are records of all catchers participating in the KBO League in 2019 and the average daily temperature in areas with baseball stadiums in South Korea. The catcher's batting performance was based on batting average, on-base percentage, and slugging percentage, and was collected with Python from the KBO League's official website (<http://www.koreabaseball.com>) and converted into an Excel file as shown in Figure 2. In addition, the average daily temperature in the area where the baseball field is located in Korea was collected by downloading a CSV file from the Weather Data Open Portal (<https://data.kma.go.kr/cmmn/main.do>). The intervals for which the daily average temperature was set were divided into 14.9 degrees Celsius or less, 15 to 19.9 degrees Celsius, 20 to 24.9 degrees Celsius, and 25 degrees or more. The reason for this is that although the number of samples of the total data must be matched to some extent, the interval must be constant for each section. Table 1 shows the stadiums and regions used by the 10 baseball clubs in the KBO League.

Table 1: Information of Baseball Parks for KBO League's 10 Baseball-Clubs

Local	Baseball Park	Baseball Club	Local	Baseball Park	Baseball Club
Seoul	Seoul Jamsil Baseball Stadium	Doosan Bears	Busan	Sajik Baseball Stadium	Lotte Giants
		LG Twins	Ulsan	Munsu Baseball Stadium	
	Gocheok Sky Dome	Kiwoom Heros	Daegu	Daegu Samsung Lions Park	Samsung Ions
Incheon	Incheon SK Happy Dream Park	SK Wiverns	Pohang	Pohang Baseball Stadium	
Daejeon	Daejeon Hanwha Life Eagles Park	Hanhwa Eagles	Suwon	Suwon KT Wiz Park	KT Wiz
Cheongju	Cheongju Baseball Stadium		Changwon	Changwon NC Park	NC Dinos

2.2. Data Analysis Methods

The data analysis method used for this study was appropriately modified for this study by referring to the SEMMA (Sampling Exploration Modification Modeling Assessment) model proposed by SAS (Statistical Analysis System), as shown in Figure 1. On the official website of the KBO League, all catchers' records by date were collected. On the official website of the KBO League, data was collected using Python as shown in Table 2. The official website of the Weather Data Open Portal provided the daily average temperature by region in a CSV file as shown in Figure 2. Based on the collected data, data suitable for this study were rearranged as shown in Figure 2 and pre-processed.

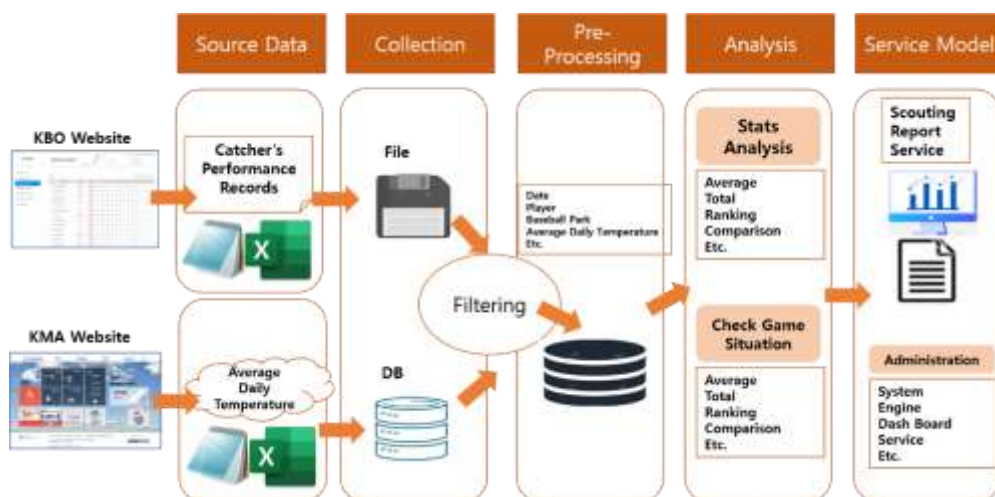


Figure 1. Data Analysis Model

	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	HBP	SO	GDP	SF	Temp	Sorting
2019-03-23	3	1	1	0	0	1	2	0	0	1	0	0	0	0	5.7	1
2019-03-24	4	0	0	0	0	0	0	0	0	0	0	1	0	0	5.7	1
2019-03-26	5	2	1	0	0	1	1	0	0	1	0	2	1	0	12.4	1
2019-03-27	3	1	3	0	0	1	2	1	0	0	0	0	0	0	13.1	1
2019-03-28	4	1	1	0	0	0	0	0	0	0	0	0	1	0	12.1	1
2019-03-30	2	0	2	1	0	0	2	0	0	1	0	0	0	0	6.1	1
2019-03-31	2	1	1	0	0	1	1	0	0	2	0	0	0	0	5.5	1
2019-04-02	4	1	1	1	0	0	0	0	0	1	1	1	1	0	8.3	1
2019-04-03	4	1	1	1	0	0	0	0	0	0	0	3	0	0	9	1
2019-04-04	3	0	1	0	0	0	0	0	0	0	0	1	0	0	11.1	1
2019-04-05	3	1	1	1	0	0	0	0	0	1	0	0	0	0	11.4	1
2019-04-07	4	1	2	1	0	0	2	0	0	1	0	0	0	0	12.6	1
2019-04-10	3	1	1	0	0	0	0	0	0	1	0	0	0	0	7.9	1
2019-04-11	4	0	2	0	0	0	2	0	0	0	0	0	0	0	10.1	1
2019-04-14	3	2	1	0	0	1	1	0	0	1	0	0	0	0	12.5	1
2019-04-16	5	0	1	0	0	0	0	0	0	0	0	1	1	0	14.8	1
2019-04-17	4	1	1	1	0	0	1	0	0	0	1	0	0	0	16.3	2
2019-04-19	4	0	2	1	0	0	1	0	0	1	0	1	0	0	13.6	1
2019-04-20	4	1	3	0	0	1	1	0	0	0	0	0	0	0	13.4	1
2019-04-21	3	0	1	1	0	0	1	0	1	0	1	1	1	0	12.5	1
2019-04-23	4	1	1	0	0	0	0	0	0	0	0	0	0	0	19.1	2
2019-04-24	3	0	0	0	0	0	1	0	0	0	0	1	0	1	18.2	2
2019-04-27	4	1	1	0	0	1	3	0	0	0	0	1	0	0	11.5	1
2019-04-28	4	1	2	0	0	0	1	0	0	0	1	1	0	0	12.1	1
2019-04-30	3	0	1	0	0	0	0	0	1	0	0	0	0	0	15.6	2
2019-05-01	3	0	1	0	0	0	0	0	0	2	1	1	0	0	17	2
2019-05-03	2	1	1	0	0	1	2	0	0	2	0	0	0	0	16.9	2
2019-05-04	3	0	0	0	0	0	0	0	0	1	0	0	0	0	17.9	2
2019-05-07	1	0	0	0	0	0	0	0	0	0	0	0	0	0	15.6	2
2019-05-08	3	0	1	0	0	0	0	0	1	0	0	0	0	0	17.8	2
2019-05-09	4	1	2	0	0	0	0	0	1	0	0	0	0	0	20.1	3
2019-05-10	4	1	0	0	0	0	0	0	1	0	0	0	0	0	18.3	2

Figure 2. Excel File of Catcher's Batting Performance

In addition, since the batting average, on-base rate, slugging rate, and so on have been downloaded by date, they had to be recalculated by creating a separate formula on the spreadsheet as shown below. Based on this complete data, the game was analyzed and a scouting report was made through visualization.

$$AVG = H / AB$$

$$OPB = (H + BB + HBP) / (AB + BB + HBP + SF)$$

$$TB = (H + 2B \times 2 + 3B \times 3 + HR \times 4)$$

$$SLG = TB / AB$$

Results and Discussion

3.1. Batting Average

The batting average of the catchers participating in the 2019 KBO League is shown in Figure 3, and in games that took place on a day when the daily average temperature was 14.9 degrees Celsius or less, their average hit .270. At 15 to 19.9 degrees Celsius, .266 was recorded, at 20 to 24.9 degrees Celsius, .236 was recorded, and at 25 degrees Celsius or higher, .240 was recorded. In general, as the days get hotter, the catchers' batting average tends to decrease. Generally, batting averages are high below 19.9 degrees Celsius and batting averages lower above 20 degrees Celsius.

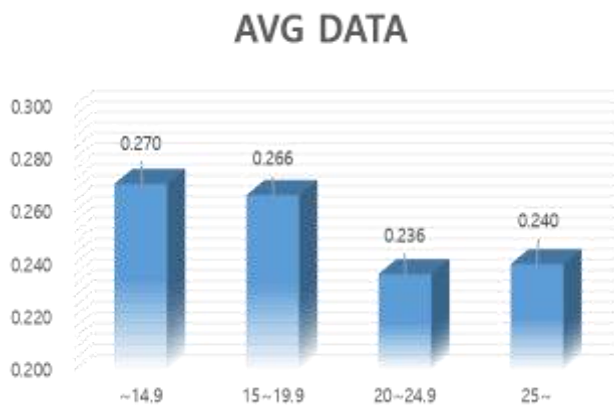


Figure 3. Batting Average Data

3.2. On-Base Percentage

The on-base percentage of catchers participating in the 2019 KBO League is shown in Figure 4, and in games that took place on the day when the average daily temperature was 14.9 degrees Celsius or less, an on-base percentage of .341 was recorded. At 15 to 19.9 degrees Celsius, it was .332, at 20 to 24.9 degrees Celsius, it was .310, and at 25 degrees Celsius or higher, it was recorded at .313. As with batting average, the catchers' on-base percentage tends to decrease as the weather gets hotter. Generally, it can be seen that the on-base percentage is high below 19.9 degrees Celsius and the on-base percentage is low above 20 degrees Celsius. However, the record for the difference between on-base percentage and batting average (OPB-AVG) was .070 when the average daily temperature was below 14.9 degrees Celsius. At 15 to 19.9 degrees Celsius, it was 0.066; at 20 to 24.9 degrees Celsius, it was 0.074; and above 25 degrees Celsius, it was recorded at .073. Rather, it can be noted that in the weather above 20 degrees Celsius, more walks were given.

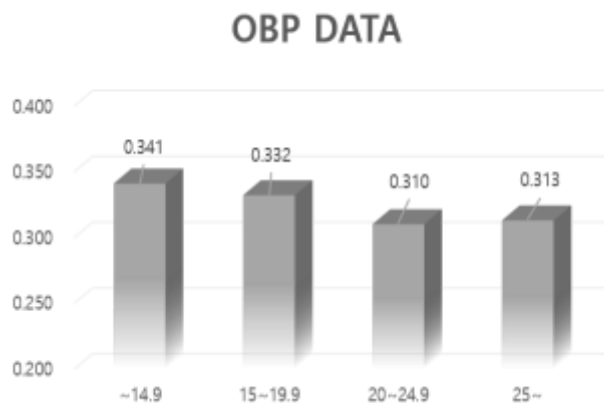


Figure 4. On-base Percentage Data

3.3. Slugging Percentage

The slugging percentage of the catchers participating in the 2019 KBO League is shown in Figure 5, and in games that took place on the day when the average daily temperature was 14.9 degrees Celsius or less, the slugging percentage was recorded .430. At 15 to 19.9 degrees Celsius, it was .349, at 20 to 24.9 degrees Celsius, it was .338, and at 25 degrees Celsius or higher, it was recorded at .342. In terms of slugging percentage, unlike batting average and on-base percentage, it was seen to be high only below 14.9 degrees Celsius, and low above 15 degrees Celsius.



Figure 5. Slugging Percentage Data

3.4. On-base Plus Slugging

OPS is a record of adding on-base percentage and slugging percentage, and it was created because it is not possible to determine the scoring productivity of batters by batting average alone. Since the maximum value was obtained with two different data, it may be thought that it may be somewhat less

probable, but it has the advantage of being easy to calculate and produces a surprisingly reasonable evaluation when viewed by Sabermetrics. Therefore, it is treated as important data in Sabermetrics. The OPS of the catchers who participated in the 2019 KBO League is shown in Figure 6, and in games that took place on the day when the daily average temperature was 14.9 degrees Celsius or less, it was recorded at .771. At 15 to 19.9 degrees Celsius, it was .681, at 20 to 24.9 degrees Celsius, it was .648, and at 25 degrees Celsius or higher, it was recorded at .655. Since OPS is a record that is greatly affected by slugging percentage, it can be seen that OPS is high only below 14.9 degrees Celsius and low above 15 degrees Celsius. Unlike batting average and on-base percentage, however, the trend of falling OPS is relatively mild as daily average temperatures rise.

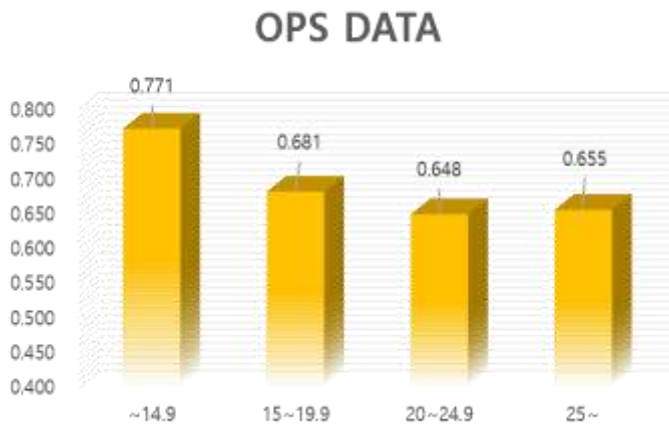


Figure 6. On-base Plus Slugging Data

3.5. Discussion

As a result of examining the batting performance of catchers through this study, we could derive the following implications.

First, the catchers results may suggest that batting performance declines as the day gets hotter. The reason for this is that the equipment worn by catchers makes them vulnerable to heat and exhausted quickly.

Second, the catchers' batting average and on-base percentage fell sharply at 20 degrees Celsius. This may suggest that the accuracy of the catchers' hitting is based on 20 degrees Celsius. However, considering that the gap between on-base percentage and batting average is maintained at around .070, it can be suggested that a sharp batter's eye was maintained.

Third, the catchers' slugging percentage and OPS declined sharply based on 15 degrees Celsius. This can suggest that the catchers' slugging power and hitting productivity are based on 15 degrees Celsius.

The implications derived are the basic sentences that can be written in the scouting report. Considering the implications, it can be seen that coaching staff and club fronts need to maintain batting performance by managing the physical strength of catchers according to the daily average temperature.

1. Conclusion

In this study, the relationship between how the batting performance of KBO League catchers in 2019 is related to the average daily temperature was studied through big data analysis. Accordingly, the SEMMA model was appropriately used for this study, and data was collected on the official website of the KBO League and the official website of the Weather Data Opening Portal, and various implications were derived so that a scouting report could be prepared by visualization through the pre-processing procedure.

As a result of the study, it can be seen that the batting performance of catchers participating in the 2019 KBO League generally decreased as the daily average temperature increased. The data proved that the reason catchers should manage their physical strength is because they are exhausted faster by more equipment and hard play compared to other players. In addition, it was proven that there is a need to manage the squad through various perspectives by analyzing data through factors outside the competition such as daily average temperature. Following Classic Stat, which was the primary record that focused on ratio records and cumulative records, and Sabermetrics, which is the second record that shows productivity through scientific statistical methods, we can expect to develop into a third record that

analyzes competition records by intervening in the external environment of the competition. Therefore, the results of this study are expected to play a new role in the sports data industry.

The limitation of this study is that it cannot be compared with pitchers, infielders, and outfielders, since only catchers were studied.

Finally, a future research task is to increase the value of the third record by comparing the data of catchers with that of other players.

References

- [1] Kang, Seungae. "Utilization and Prospect of Big Data Analysis of Sports Contents." *Journal of Information and Security* 19, no.1 (2019): 121-126.
<http://www.earticle.net/Article/A354948>
- [2] Oh, Jaewon, Woo Hyun Ahn, and Taegong Kim. "Automatic Extraction of Dependencies between Web Components and Database Resources in Java Web Applications." *Journal of Information and Communication Convergence Engineering* 17, no. 2 (June 30, 2019): 149-60. doi:10.6109/JICCE.2019.17.2.149.
- [3] Lee, Seung-Hun, and Hyong-Jun Choi. "The Analysis of Pitching Result According to the Velocity and Pitch of Pitcher in That Case of Full-Counting on Major League Baseball(MLB)." *Korean Journal of Sports Science* 28, no. 3 (June 30, 2019): 973-981. doi:10.35159/kjss.2019.06.28.3.973.
- [4] Kim, Semin. "The Effect of Daily Average Temperature on the Batter's Performance in Baseball Game : focused on Big Data Analysis." Master's Thesis, The Graduate School of Hoseo University, 2020.
http://www.riss.kr/search/detail/DetailView.do?p_mat_type=be54d9b8bc7cdb09&control_no=c0f5954edc45830cffe0bdc3ef48d419
- [5] Kim, Semin, and You, Kangsoo. "The Effect of Daily Average Humidity on Pitcher's Stats of Strike-Out." *Journal of Industrial Convergence* 18, no. 1 (February 28, 2020): 65-71. doi:10.22678/jic.2020.18.1.065.
- [6] "So That's It!, Which of the Rainy Season Pitchers and Batters is More Advantageous" *Maeil Business Newspaper*. last modified Jul 20. 2015, accessed Jun 10. 2020, <https://www.mk.co.kr/news/it/view/2015/07/695136/>.
- [7] Jeon, Young Nam. "A Study of the Athletic Satisfaction and Stress Factors of High School Baseball Players in the Case of Different Positions." Master's Thesis, The Graduate School of Kyungwon University, 2010.
http://www.riss.kr/search/detail/DetailView.do?p_mat_type=be54d9b8bc7cdb09&control_no=84c842ae1786e3b9ffe0bdc3ef48d419
- [8] Kim, Hyeon-Gyu. "Suggestion of Batting and Starting Pitcher Ability Index in Korea Professional Baseball - focusing on Sabermetrics Statistics WAR." Master's Thesis, The Graduate School of Youngnam University, 2017.
http://www.riss.kr/search/detail/DetailView.do?p_mat_type=be54d9b8bc7cdb09&control_no=98e8f18e6b0fe77bffe0bdc3ef48d419
- [9] Yun, Hyo-Jun. "A Real-time Players Evaluation Model Development based on Social Big Data in Korea Professional Baseball: Sentiment Analysis Using Machine Learning." Dissertation, The Graduate School of Korea National Sport University, 2018.
http://www.riss.kr/search/detail/DetailView.do?p_mat_type=be54d9b8bc7cdb09&control_no=a8c991936d9cfd80ffe0bdc3ef48d419
- [10] Lee, Jang Taek. "Long Term Trends in the Korean Professional Baseball." *Journal of the Korean Data and Information Science Society* 26, no. 1 (January 31, 2015): 1-10. doi:10.7465/jkdi.2015.26.1.1.